Molecular dynamics simulation of inter-molecular interactions

Shamim Zahra Shadfar

A thesis
submitted to Massey University in Albany, Auckland
in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Computational Biochemistry

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Massey University Auckland
2019

Abstract

Many aspects of the operation of chemical and biological systems are based on intermolecular interactions. In this work, binding modes and interactions between molecules at a range of scales have been studied, using molecular dynamics (MD) simulations.

The first chapter provides an introduction of each of the different chemical and biological systems that are studied in this work. It also introduces MD and its role in the context of this research.

The second chapter corresponds to the study of host-guest interactions for cyclodextrin-bullvalene complexes. Bullvalene is a shapeshifter molecule, which interconverts between different isomers at room temperature. The goal of this chapter is to capture one favourable isomer of bullvalene (guest molecule) by binding it to cyclodextrin as a host molecule. This chapter consists of two smaller chapters (2i and 2ii). The former details the development and validation of a "host-guest binding potential energy profiling" (HGBPEP) method, which is a rotational interaction energy screening method designed for prediction of the most favourable orientation and position of bullvalene isomers with respect to cyclodextrin. The latter investigates the interaction of bullvalene isomers and cyclodextrin molecules, and finally binding free energy values of the complexes are calculated.

The third chapter describes KstR, a transcriptional repressor in Mycobacteria. KstR is required for *Mycobacterium tuberculosis* (Mtb) pathogenesis as well as regulating the initial steps in cholesterol degradation by controlling the expression of the enzymes that carry out the early stages of cholesterol catabolism. Therefore, this protein is of great interest for development of new tuberculosis treatments. In this chapter, the stability and conformational changes of KstR in its different states – apo, DNA-bound and ligand-bound –have been studied. The main goal is to investigate the binding mechanism of KstR to DNA, as well as the effect of DNA and ligand binding on the structure and dynamics of KstR more generally, using MD simulations.

In the fourth chapter, KstR2, another Mtb transcriptional repressor, is studied. KstR2 represses a 14-gene regulon involved in the later steps of cholesterol degradation. It is structurally similar to KstR, but has been proposed to act through a novel scissor-like mechanism. This chapter investigates two key questions regarding the mechanism of action of KstR2: first, the effect of mutating the key switch residue ARG170 to ALA, and second, the effect of ligand binding on its structure and motion.

The focus of the fifth and final chapter is phosphatidylinositide 3-kinases (PI3Ks), which are proteins that take part in signalling pathways regulating factors like cell growth, survival and proliferation, which in turn are involved in cancer. The interaction between PI3Kα and another protein, RAS, is very important in the formation, growth and maintenance of RAS-driven tumours. A model of PI3Kα (class IA PI3K) has therefore been built, as well as of RAS associated with a model cell membrane, and MD simulations used to investigate the process by which the two proteins interact with one another and with the lipid bilayer.

Altogether, this thesis uses MD simulations to provide insight into intermolecular interactions at a range of scales, with a particular focus on proteins involved in tuberculosis and in cancer.

Acknowledgements

In the beginning, I would like to express my deepest appreciation and gratitude to my supervisor A. Prof. Jane Allison for her continuous support, enthusiasm, wisdom, passion, and patience throughout the course of my PhD. Without her, it would not have been possible for me to achieve what I did over these three years. Thanks to her for believing in me.

I wish to show my gratitude to Dist. Prof. Peter Schwerdtfeger, director of the Centre for Theoretical Chemistry and Physics (CTCP) at Massey University, Auckland for his support, motivation, and kindness.

I would specially appreciate Dr Thomas Fallon and Dr Thomas Collier, who co-supervised me on the first two years of the course of my PhD, and assisted me with their priceless guidance and advice.

I would like to thank A. Prof. Shaun Lott and Dr Stephanie Dawes for their great help and support with Chapters 3 and 4. projects, as well as Dr Jack Flanagan for his wisdom and assistance with Chapter 5. I would also acknowledge Oussama Yahiaoui, who both helped me with Chapter 2, as my experimental collaborator in Dr Fallon's lab.

I would like to pay my special regards to my husband, Dr Morteza Bagheri, for all his love, care, understanding, and most of all the patience he afforded me in the completion of my thesis. Thanks to my parents Azim Shadfar and Shohreh Kabiri, and my brother M.J. Shadfar, who kept me going on and this work would not have been possible without their emotional support.

I especially acknowledge NZ Royal Society (Rutherford Discovery Fellowships) for providing me funding.

I wish to thank all my collaborators and colleagues whose assistance was a milestone in the completion of this project: Dr Ivan Welsh, Dr Ashar Malik, Dr William Irvine, Dr Eli Kobzev, Dr. Xavier Periole, Dr. Alex popinga, Dr Thomas Piggot, Jack Copping, and Aparajita Chakraborty.

I wish to show my appreciation and gratitude to Dr. Hossein Fallah-Bagher-Shaidaei, who supervised me throughout the course of my MSc, and for his continuous support always.

Last but by not least, I would like to thank the School of Natural and Computational Sciences (SNCS), and CTCP, and by extension, Massey University, and to New Zealand eScience Infrastructure (NeSi), School of Biological Sciences in the University of Auckland, that it was amazing to have the blessing to undertake my research their great facilities.

Table of Contents

Chapter 1. Introduction

1.1 Supramolecular and host-guest chemistry

If molecular chemistry is the chemistry of internal bonds, supramolecular chemistry can be defined as the chemistry of intermolecular interactions of two or more chemical species binding together (1). Examples of such systems range from small organic molecules through to biological macromolecules. In organic chemistry, the two species are often referred to as *host* and *guest*, whereas in sensing applications, biological chemistry and biochemistry, more common terminology includes *receptor* and *substrate* or *ligand*. The receptor or host is often the bigger partner, whereas the smaller would be the substrate or guest (2) (Figure 1. 1).

At all scales, the binding depends on the shape complementarity and strength of the interaction between the two molecules. The stability of the inclusion complex formed is determined by the fit of the "guest" molecules in the "host" cavity. The host is stereo-selective in the complexation of the guest (3).

Studying binding affinity is important to have a better understanding of the intermolecular interactions in chemical and biological processes, structural biology, and structure-function relationships. Also, it is of great importance in the drug discovery process for designing drugs that bind their receptors or targets (4). Therefore, in this work, binding modes and interactions between host and guest molecules at a range of scales have been studied, using computational methods, in order to investigate the binding affinity between the two systems.

Chapter 2 of this thesis describes the study of host-guest interactions between bullvalenes and cyclodextrins. These two molecules are introduced in more detail below.

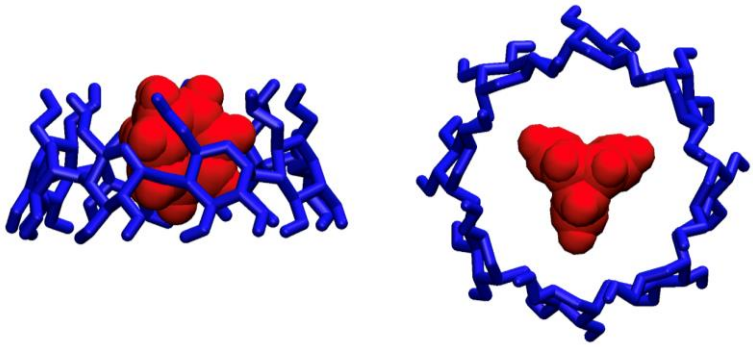


Figure 1. 1. An example of a host-guest system: bullvalene (red, VDW representation) encapsulated by γ-cyclodextrin (blue, liquorice representation).

1.1.1. Bullvalene

Bullvalene, $C_{10}H_{10}$, is an unusual organic cage molecule which spontaneously undergoes Cope rearrangements (5), generating more than 1.2 million (10!/3 = 1,209,600) degenerate (i.e. indistinguishable) isomers which interconvert (Figure 1. 2) (6, 7). The rearrangements involve bonds breaking and new bonds forming to make another isomer (8). In fact, bullvalene has no permanent carbon–carbon bonds. All ten carbon atoms are bonded to the same extent, and continuously exchange bonded configurations at room temperature (9). Such molecules are interesting potential pharmaceuticals, as they can explore a large number of shapes in order to find a perfect match with a biological target. Studies on bullvalene derivatives have shown that they have the potential to act as highly specific sensing molecules, able to differentiate structurally similar biomolecules from one another (10).

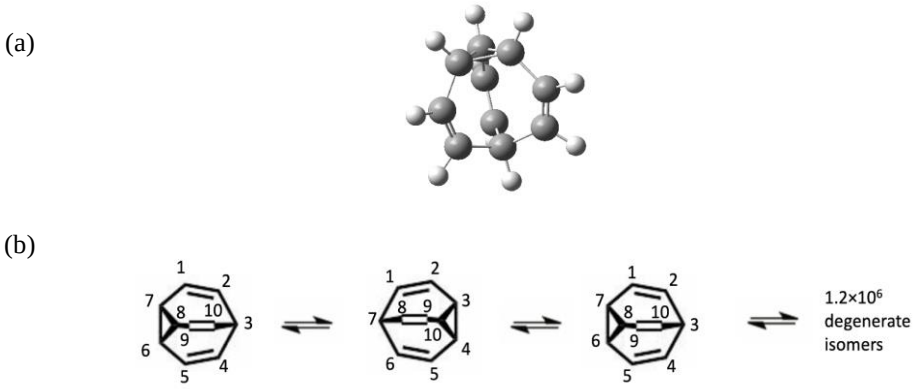


Figure 1. 2. (a) Bullvalene. Carbon atoms are shown as grey spheres and hydrogen atoms as white spheres. (b) Cope rearrangement and shape-shifting in bullvalene.

Experimentally, self-interconversion of bullvalene has been studied by nuclear magnetic resonance (NMR) in solution (11). The fluctionality of the bullvalene structure leads to interesting NMR results. Any substituted bullvalene generally represents a system of 1.2×10^6 valence isomers, but these isomers are not all structurally equivalent, unlike for unsubstituted bullvalene. Considering the three-fold symmetry of bullvalene, a monosubstituted bullvalene has four possible positional isomers, inconverting at room temperature (Figure 1. 3). Substituted bullvalenes, where different functional groups are attached, will isomerise to present these substituents in all possible arrangements, although some may be more

energetically favourable than others. Substituents on bullvalene also reduces the number of possible isomers, so that they can be distinguished.

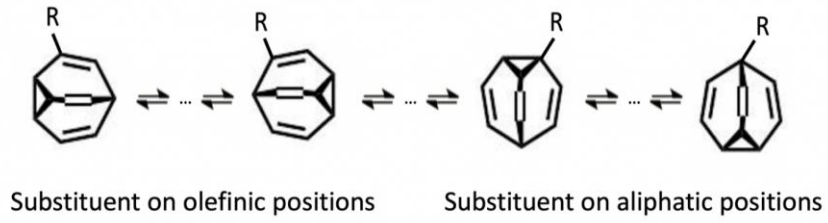


Figure 1. 3. The four distinguishable positional isomers of mono-substituted bullvalenes.

Both unsubstituted and substituted bullvalenes produce temperature-dependent NMR spectra (Figure 1. 4). The high-temperature spectra reflect the rapid shape-shifting and therefore a dynamic situation, whereas the low temperature spectra reflect the static situation in which a single isomer is populated (5, 7, 12). All protons are indistinguishable in both high and room temperature ^{1}H-NMR spectra. As the temperature goes up, there is an unusual increase in line width and decrease in intensity, which is related to bond isomerism in bullvalene. At very low temperature (−55°C) the spectrum relates to a single valence isomer and all hydrogens are distinguishable (11, 13).

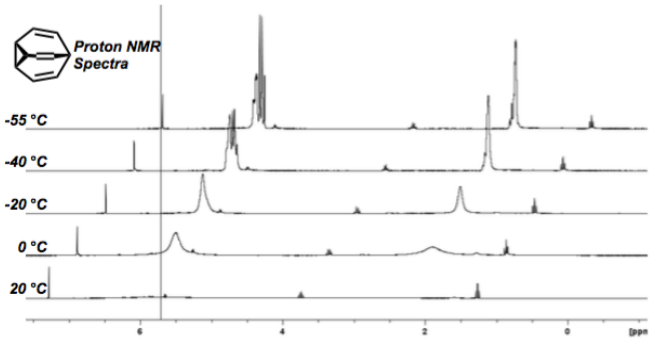
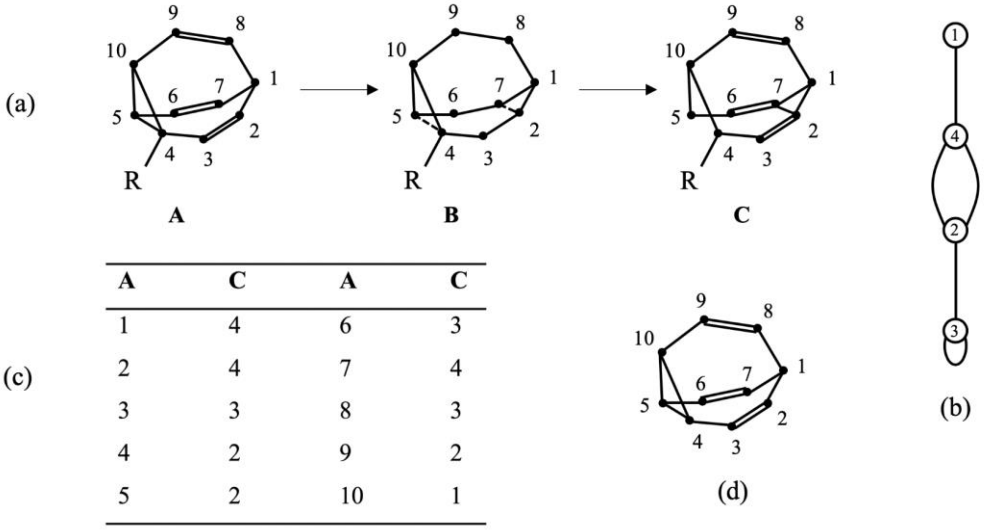


Figure 1. 4. Variable temperature ^{1}H-NMR spectra of unsubstituted bullvalene (14).

In mono-substituted bullvalenes, the substituent mostly prefers to occupy an olefinic position rather than an aliphatic one (7) (Figure 1. 3). Bullvalene has a triangular cyclopropane ring, in which the bond angles between carbon-carbon bonds are 60°, resulting in significant ring strain. However, substituents on a cyclopropane ring affect the ring bonds' stability. In a

cyclopropane ring with a π-acceptor substituent, the carbon-carbon bonds of the ring are strengthened and cyclopropane is more stable, whereas π-donor substituents weaken the same bonds and cause an unstable cyclopropane ring (15). This explains why substituents tend to occupy the olefinic position. For monosubstituted bullvalene with a donor -CH_2OH group, one of the least favourable isomers in energy is the one in which the substituent is in the aliphatic position on the cyclopropane ring. In contrast, the isomer with -CH_2OH in the olefinic position is the most populated one (15).

The position of each carbon atom in a bullvalene changes as a result of isomerization. For an unsubstituted bullvalene, this does not matter, as all carbons are chemically equivalent. However, for a substituted bullvalene, it will appear as if the substituent is shifting to be attached to a different carbon atom, whereas in fact it is still bonded to the same atom, but that that atom has a different position within the bullvalene.



A	C	A	C
1	4	6	3
2	4	7	4
3	3	8	3
4	2	9	2
5	2	10	1

Figure 1. 5. (a) Cope rearrangement in a substituted bullvalene, showing how the position (given by the numbers) of each atom in the bullvalene structure changes from isomer A to isomer C. (b) Change in the position of a substituent R during Cope rearrangement from isomer A to isomer C, using the atom position numbers labelled on isomer A in (a). Only the lowest valued of the symmetrically equivalent atom positions are given for isomer C. (c) Reaction graph for mono-substituted bullvalene, where the number denotes the position of the substituent in each isomer, using the atom position numbers labelled on isomer A in (a). (d) Structure of a labeled unsubstituted bullvalene (11).

Substituted bullvalenes can be thought of as a network or graph of interconverting isomers. Nodes denote isomers, and edges represent processes which connect one isomer to another (8, 16). Figure 1. 5 shows the interconversion network for four isomers of a monosubstituted bullvalene. According to Figure 1. 5 (a), there are 10 possible starting positions for a substituent at the beginning of isomerisation (A). If the substituent is on position 4 in A, it will end up

being on position 2 after the rearrangement (C) (Figure 1. 5 (b)). Since 4, 5 and 10 are cyclopropane carbon atoms in A, Cope rearrangement will convert two of them to sp^2 positions and one to sp^3 (in C). Figure 1. 5 (c) shows the connections between all four distinguishable isomers of a mono-substituted bullvalene. The number of each isomer is the lowest-valued position of the substituent in that isomer. If we start from isomer 1 at the top of the graph (in which the substituent is on position 1 in A, Figure 1. 5 (d)), a single Cope rearragement will form isomer 4, (with the substituent located at postion 4). Isomers with the substituent at positions 4, 5, and 10 (all equivalent to 4) in A will rearrange to form either isomer 2, with the substituent at position 2, or isomer 1, with the substituent returned to position 1. The two arcs between isomers 4 and 2 in Figure 1. 5 (c) represent two enantiomeric transition states between these isomers. Isomer 2 can also rearrange to form isomer 3, with the substituent on position 3. The loop at the bottom of the graph shows the degenerate rearrangement of isomer 3 into itself (8).

While such networks are useful for outlining the theoretically possible isomers and their transitions, to understand the true behaviour of a given bullvalene, it is necessary to also consider the energy of each isomer, which determines its likelihood of being formed. Therefore, our collaborator, Dr. Lukáš Pašteka (formerly Massey University, now Comenius University), has written a program that takes any group of substituents, and builds and optimizes the geometry of the structures of all possible isomers of a bullvalene with those substituents using quantum chemical calculations (B3LYP/6-311+G*) (14, 16). It determines the nature of the network, and the interconnectivity of all isomers and transition structures (Figure 1. 6). This program has been run for many bullvalene molecules, producing a library of the isomers of 54 mono-, 15 di- and one tri- substituted bullvalenes.

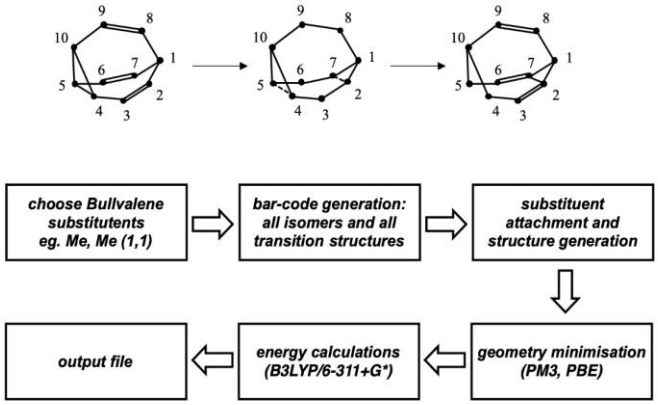


Figure 1. 6. Workflow for constructing and optimizing substituted bullvalene isomer structures (14, 16).

Importantly, the quantum chemical calculations also yield the potential energy of each isomer. From this, a network describing the interconnectivity of all the bullvalene isomers and their transition structures, as well as their energies, is built (the interconversion graph). An example of such a network is given in Figure 1. 7 for dimethylbullvalene. It is substantially more complicated than the mono-substituted bullvalene network shown in Figure 1. 5.

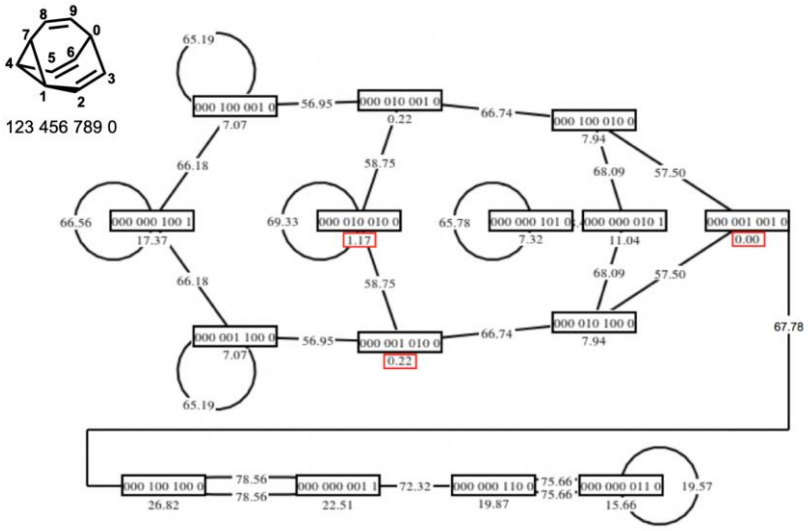


Figure 1. 7. The dimethylbullvalene network. Nodes (rectangular boxes) represent isomers, and the lines connecting them represent transition structures. The binary codes represent the positions of the two substituents (where 1 means that the substituent is on that carbon atom, while 0 means there is no substituent on that carbon atom). The numbers below the rectangles are the relative energies (relative to the lowest energy isomer) (kJ/mol) and the numbers between the lines are the relative transition state energies (relative to the lowest energy isomer) (kJ/mol) (14, 16).

The most complex example that Dr. Pašteka has explored so far is three-substituent bromo-fluoro-methylbullvalene. The network for this molecule has 138 isomers and 186 transition structures (Figure 1. 8) (14, 16).

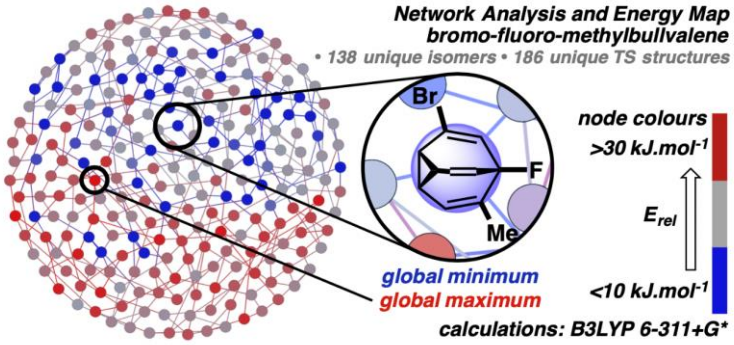


Figure 1. 8. Bromo-fluoro-methylbullvalene isomer network. The lowest energy nodes are shown in blue, and the global minimum and maximum are highlighted, with the global minimum energy structure drawn out in full (14, 16).

These interconversion networks are important for two reasons: they show both which isomers are most likely to occur, but also which transitions are possible at a given temperature. For the research described in this thesis, the most important aspect is that the most favourable isomer can be easily found. Additionally, the host-guest binding calculations, which are discussed in Chapter 2, all use the most stable isomer of a bullvalene as the guest molecule.

1.1.2. Cyclodextrin

Cyclodextrins are widely used in supramolecular chemistry. They are cyclic oligosaccharides naturally produced by bacteria from starch (17). Three main types of cyclodextrins are α-, β- and γ-cyclodextrins, which consist of six, seven and eight glucopyranose units, respectively (Figure 1. 9).

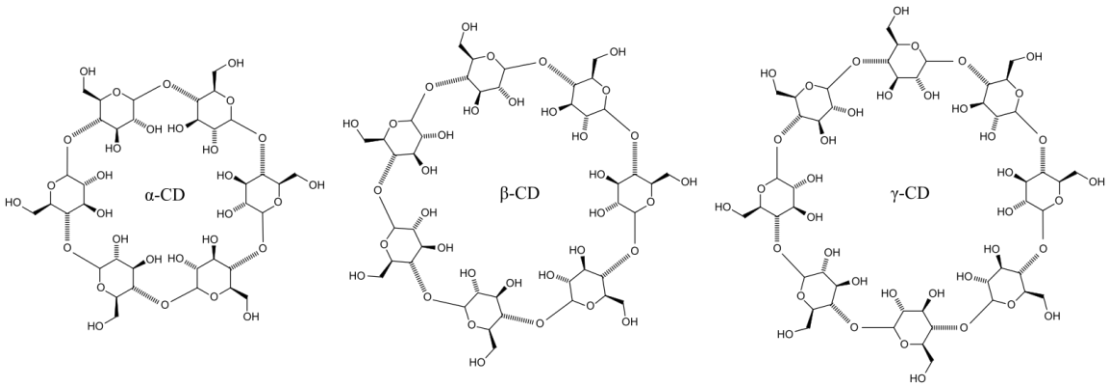


Figure 1. 9. The three main types of cyclodextrin, as labelled.

Cyclodextrins are used as active site analogues for enzymes (18), to aid the absorption of drugs in the body (drug delivery) (19), and to reduce volatility of insecticides in agriculture. They are utilized in the petroleum industry for separation of aromatic hydrocarbons (17). They can be used as chiral selectors, drug solubility reagents, and molecular reactors (20). Among the three types of cyclodextrins, β-cyclodextrin (βCD) is the most commercially available and generally useful, and the least expensive (21). βCD consists of seven α-D-glucopyranose units connected by α-1→4 glycosidic bonds (22).

The structure of cyclodextrin can be represented as a toroid with larger and smaller openings (truncated cone) (21, 22) (Figure 1. 10). The hydroxyl groups are located on the external surface, and circle the larger open rim, making the exterior sufficiently hydrophilic to make cyclodextrin soluble in polar solvents (3, 21). Due to this, cyclodextrins are considered as good candidates for encapsulating hydrophobic hosts, therefore generating a host-guest complex.

It is expected that as the number of oligosaccharides in cyclodextrin increases, so too does the solubility in water. However, it does not follow that trend (Table 1. 1) (21), which presents a puzzle (23).

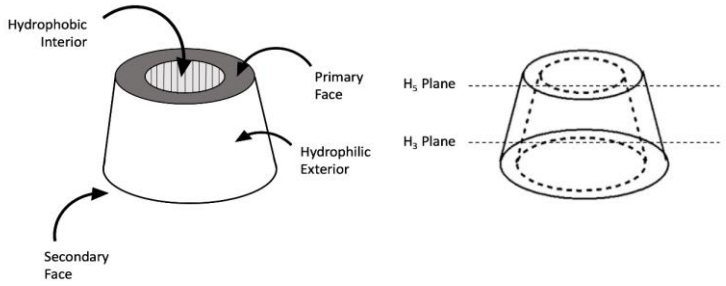


Figure 1. 10. Toroid structure of cyclodextrin (left) (Taken unaltered from (21)). In cyclodextrins, there are two series of (six in α-, seven in β-, and eight in γ-cyclodextrins) hydrogen atoms inside the cavity at each rim of cyclodextrin, called the H_5 and H_3 planes (right).

Table 1. 1. Properties of α-, β- and γ-cyclodextrin (24).

Cyclodextrin	Mass	Cavity Diameter (nm)		Solubility, g/kg H_2O (23)
		Smaller rim	Larger rim	
α, $(Glucose)_6$	972	0.45	0.53	129.5
β, $(Glucose)_7$	1134	0.60	0.65	18.4
γ, $(Glucose)_8$	1296	0.75	0.85	249.2

In aqueous solution, and without any guest molecule, the hydrophobic cavity of cyclodextrin contains about three (αCD), seven (βCD) or nine (γCD) water molecules, which are poorly held with low entropy and low density, and are ready to be displaced by guest molecules, so that the cavity can accommodate guest molecules (24). The complex formation and binding is driven by the release of these water molecules from the cavity, which are displaced by more hydrophobic molecules. In this way, the ring strain in cyclodextrin decreases, and brings about a more stable state with lower energy (21).

Among the three cyclodextrins, γCD is the most flexible, whereas αCD is the most rigid. In the presence of a guest molecule, γCD is more likely to take a symmetrical conformation, while in the absence of a guest inside the cavity, it assumes a less symmetrical conformation (25). It is interesting, therefore, to investigate cyclodextrin complexations at atomic level, using computational methods. In the work described in Chapter 2, β- and γCDs are explored as hosts for binding unsubstituted and substituted bullvalenes as guest molecules.

1.2. Proteins

Proteins are large biomolecules that carry out a vast array of functions required for life, such as cell structure maintenance, signal transduction, transporting molecules and catalysis. Amino acids are the building blocks of proteins, with the sequence of these amino acids controlled by the genes which encode them. The amino acid sequence of a protein determines its three-dimensional structure and shape, and thus also its function. The function of a protein is also dependent on its dynamics. Many diseases are caused by defects in protein synthesis, structure or function (26).

Based on proteins' different cellular functions, they are classified as structural proteins, transport proteins, defence proteins, enzymes and signalling proteins. Enzymes are specific catalysts for the majority of the chemical reactions that take place in the cell. Signalling proteins usually cooperate with effector molecules, which are typically small molecules such as hormones or metabolites. They also regulate physiological effects and biochemical pathways in the body.

There is a variety of different experimental methods to determine protein structures, like X-ray and neutron diffraction, solid- and solution-state NMR, transmission electron microscopy, atomic-force microscopy, and most recently, cryo-electron microscopy and X-ray free electron lasers (27). The structure is also determined by using molecular modelling to interpret the experimental data in terms of a 3D structure.

Computational methods such as molecular dynamics (MD) allow study of the physical movements of atoms and molecules due to their interactions, giving a view of the dynamic progress of the system during the simulation time. Due to advanced computer processing speeds, the viability of using MD to study biomolecular interactions as well as the motions of individual molecules has recently rapidly increased. Later, in Section 1.4. of this chapter, MD will be discussed in further detail.

1.2.1. Protein interactions

In a biological context, proteins interact with many types of molecules, such as other proteins, lipids, membranes, carbohydrates, and with deoxyribonucleic acid (DNA), as well as small molecules such as enzyme substrates and products, hormones and other signalling molecules, ligands and drugs. Therefore, they do not often act alone, and their functions tend to be regulated by their interactions with other molecules (26).

1.2.1.1. Protein-protein interactions

Proteins interact with each other to assemble into dimers, multi-protein complexes or longer chains (28). These may be homo-oligomers or hetero-oligomers. As many of the vital functions in organisms are dependent on protein-protein interactions, incorrect interactions can cause malfunction. The malfunction of proteins is in turn associated with a wide range of human diseases. Therefore, a variety of methods are being developed to prevent abnormal interactions (29) and also to generate inhibitors of protein–protein interactions that may form useful therapeutics for human disease. Consequently, such interactions are important to understand, and thus have been widely studied experimentally as well as theoretically.

1.2.1.2. Protein-DNA interactions

Proteins which have DNA-binding domains have an affinity for single- or double-stranded DNA (30–32). They interact with DNA in order to regulate its biological functions, mostly gene expression (33). In Section 2.1.4, an example of a DNA-binding protein that is the focus of Chapter 3 will be introduced and discussed.

1.2.1.3. Protein-ligand interactions

Proteins also interact with smaller molecules such as drugs and their cognate ligands. Like other protein interactions, this type of interaction has also an essential role in regulating biological processes as well as being the mechanism by which chemical reactions are carried out in living organisms (34)(35). Binding to small molecules can cause small- or large-scale conformational changes to the protein. In Sections 2.1.4 and 2.1.5. two examples of protein-ligand interactions, which are investigated in Chapters 3 and 4, are discussed.

1.2.1.4. Protein-membrane interactions

Biological membranes are lipid bilayers that surround cells and their organelles, making a physical boundary and a therefore a micro-environment, so that the cell or organelle can perform its biological processes (36). Cell membranes comprise a variety of different types of lipids, many of which contain hydrophobic fatty acid tails and a hydrophilic headgroup (37–39). A large number of proteins interact with the cell membrane, especially those that participate in signal transduction. Some human diseases linked to membrane proteins' misfunction are heart disease, Alzheimer's and cystic fibrosis (40). Therefore, it is essential to study and characterise the interactions between membrane proteins and the membrane itself in terms of their orientation, the position of potential or known drug binding sites, and specific

lipid interactions. There are different types of membrane proteins, of which the two main classes are: integral membrane proteins and peripheral membrane proteins (26) (Figure 1. 11). Integral membrane proteins are embedded in the membrane and are typically transmembrane, which means they stretch across the bilayer. Peripheral membrane proteins bind to only one surface of the membrane, and may have loops or lipid anchors that protrude into the membrane.

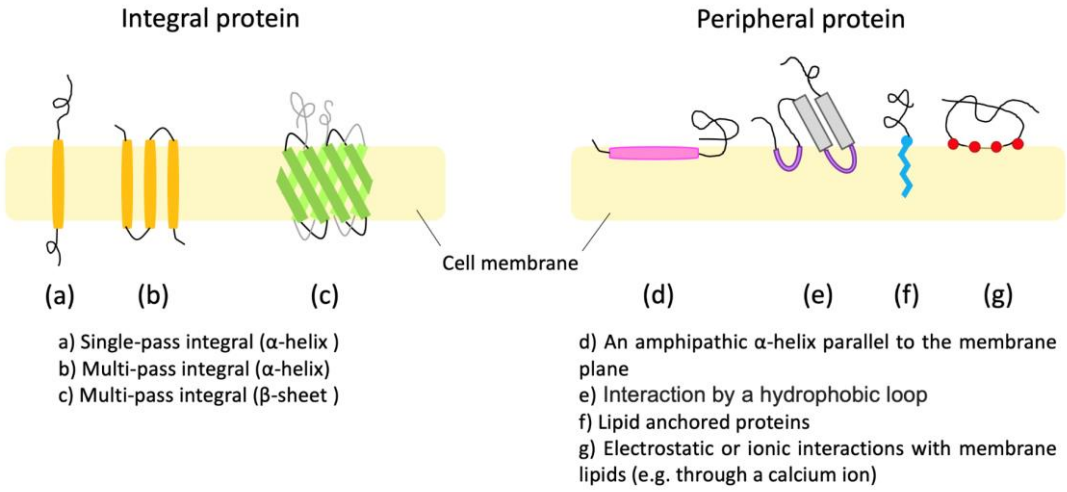


Figure 1. 11. Schematic simplified representation of integral (left) and peripheral (right) proteins, and their different types of interaction with the cell membrane (light yellow colour).

Membrane binding domains of peripheral membrane proteins may interact with the head groups of specific cell membrane lipids, largely through electrostatic interactions. These are typically mediated by the interaction of positively charged amino acids (ARG, LYS and HIS) with acidic (negatively-charged) phospholipids in the cell membrane. In many, but not all cases, there are also hydrophobic interactions between hydrophobic amino acids and the hydrophobic membrane core, which comprises the hydrocarbon lipid tails (41). Hydrophobic amino acids, such as TRP, PHE, TYR, LEU, ILE, and VAL, can therefore play an important role in membrane penetration by parts of the protein, with the effect of anchoring the protein to the membrane surface (42).

The process by which a peripheral membrane protein interacts with a cell membrane can be divided into two parts - membrane association, and membrane binding. Non-specific long range interactions, typically electrostatic interactions, drive membrane association, in which the protein approaches the membrane surface. The process of membrane association plays an important role in not only translocation of the protein to the membrane, but also

ensuring that membrane binding occurs with the protein in the proper orientation to attain its membrane-bound function (26).

1.2.2. Phosphatidylinositol 3-kinases (PI3Ks)

One of the most important types of protein-protein interactions are those between signalling molecules, which lead to signal transduction. Cellular functions are regulated by extracellular signals through these signalling pathways, and signal transduction has a vital role in various biological procedures and in many diseases, such as cancer (43).

An important family of intracellular signal transducer enzymes are the phosphatidylinositol 3-kinases (PI3Ks). PI3Ks were initially discovered in the 1980s (44, 45). They participate in signalling pathways regulating functions like cell growth, proliferation, and survival, as well as motility, differentiation, glucose transport (46) and intracellular trafficking, which are also involved in cancer. These lipid kinases are peripheral membrane proteins that catalyse the phosphorylation of the 3 position OH group of the inositol ring of phosphatidylinositol (PtdIns) (47), switching phosphatidylinositol 4-phosphate (PI(4)P) or phosphatidylinositol 4,5-bisphosphate ($PI(4,5)P_2$) to phosphatidyl 3,4-bisphosphate ($PI(3,4)P_2$) or phosphatidyl 3,4,5-triphosphate ($PI(3,4,5)P_3$), respectively (Figure 1. 12). These then cause increased downstream signalling of the Akt pathway, initiated by Akt, a serine/threonine kinase (48). Akt is fully activated through attraction to the membrane by interaction of its phosphoinositide docking sites with $PI(3,4)P_2$ or $PI(3,4,5)P_3$ (49). It then mediates downstream responses, including cell growth,, growth, migration and proliferation, by phosphorylating various intracellular proteins (50).

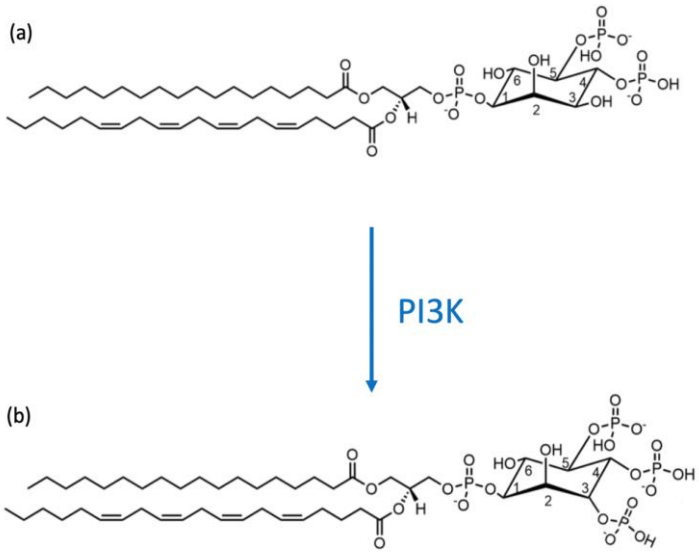


Figure 1. 12. Catalytic activity of PI3K, and structure of (a) 18:0-20:4 PI(4,5)P_2 and (b) 18:0-20:4 PI(3,4,5)P_3.

The PI3K family can be classified into three main classes: Class I, Class II, and Class III (Figure 1. 13). The divisions are on the basis of differences in structure, regulation mechanism, and substrate specificity (48). The three classes are described in more detail below, with a particular focus on Class I, the type of PI3K studied in this thesis.

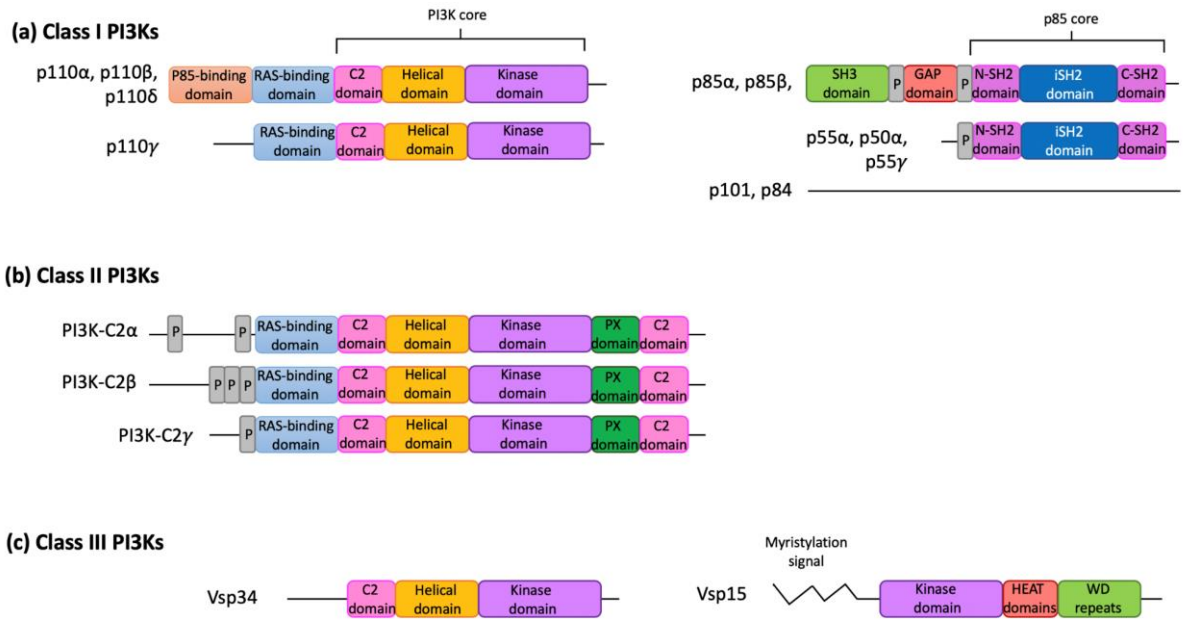


Figure 1. 13. Different domains of subunits in (a) class I, (b) class II, and (c) class III PI3Ks. Classes I and II have catalytic subunits that include a "core" structure, which has a C2 domain, a helical domain and a kinase domain, and class I also has a regulatory subunit with a core comprising two SH2 domains and an inter-SH2 domain.

1.2.2.1. Class I PI3Ks

The class I PI3Ks are in turn divided into four isoforms, PI3Kα, PI3Kβ, PI3Kδ and PI3Kγ. They can phosphorylate PtdIns, PI(4)P, and PI(4,5)P_2 *in vitro*; however, the preferred substrate of class I PI3Ks *in vivo* is PI(4,5)P_2, generating PI(3,4,5)P_3. The class I PI3Ks can be defined as heterodimers of a catalytic subunit and a regulatory subunit (Figure 1. 14) which are closely bound (51, 52). The catalytic subunit, called p110, has four variants in humans: p110α, p110β, p110δ and p110γ (52–55). There are two distinct families of regulatory subunits, which forms the basis of further subdivision into class IA and IB PI3Ks (56). The class IA PI3Ks have a p110α, p110β, or p110δ catalytic subunit, which binds to one of five p85 regulatory subunits: p85α (or its splice variants, p50α and p55α), p85β, or p55γ. On the other hand, the class IB PI3Ks are heterodimers of the p110 catalytic subunit and either a p101 or p84 regulatory subunit (52–55).

The catalytic and regulatory subunits of the class I PI3Ks are divided into different domains (Figure 1. 14). For class IA PI3Ks, the catalytic subunit has an N terminal adaptor binding domain (ABD), a RAS binding domain (RBD), a C2 domain, a helical domain, and finally the C terminal kinase domain (57). Conversely, the regulatory subunit in class IA PI3Ks is comprised of a p85 core structure containing two Src homology 2 (SH2) domains, nSH2 and cSH2, separated by a coil- coil domain known as the inter-SH2 domain (58). The larger p85 regulatory subunit isoforms, p85α and p85β, also have two proline-rich regions, an SH3

domain and a GAP domain with a homology to GTPase-activating proteins (GAPs) for Rho family G proteins (also termed the breakpoint cluster region homology [BH] domain) (59).

For class IB PI3Ks, the p110γ subunit has a C2 domain, helical domain, catalytic domain and a RBD (60). The p101 and p84 regulatory subunits do not have SH2 domains (with the rest of the structure remaining unclear) (61, 62).

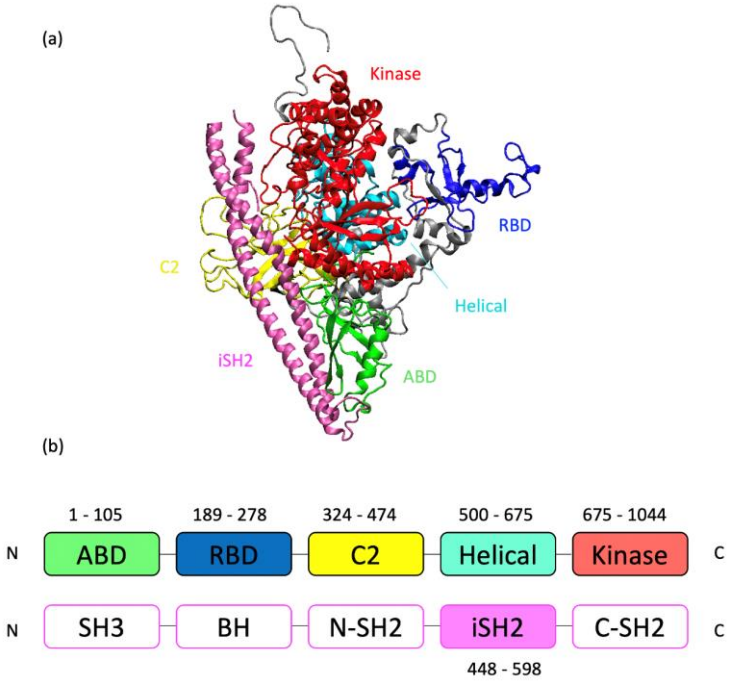


Figure 1. 14. a) Cartoon structure of the p110α catalytic subunit in complex with the iSH2 domain of the p85α regulatory subunit [PDB ID: 4OVV (49)], coloured according to domain. b) The domains present in (A) (shaded boxes) and their respective residue numbers comprising the p110α (top) and p85α (bottom) subunits.

1.2.2.2. Class II PI3Ks

The class II PI3Ks are large enzymes that can be categorised into three isoforms, PI3K-C2α, PI3K-C2β and PI3K-C2. The PI3K-C2α isoform is found in many human tissues, mostly in the heart, ovaries, and placenta. PI3K-C2β is found mainly in the thymus and placenta, and PI3K-C2 is expressed mostly in the liver and prostate. Like the class I PI3Ks, class II PI3Ks also have a RAS-binding domain, and the remainder of the p110 core structure (including C2, helical, and C-terminal kinase domains); however, they have no adaptor (p85) binding domain (60) (Figure 1. 13). A Phox homology (PX) domain and a second C2 domain are located at the C-terminus of all class II PI3Ks, while the N-terminal structures vary from one to another isoform. Even though PX domain exists in many eukaryotic signalling molecules, the function of this domain is still unknown (63). However, C2 domain is a phospholipid-binding element that can confer a calcium ion sensitivity (64).

1.2.2.3. Class III PI3Ks

This class is found in lower eukaryotes, plants and mammals; therefore, it can be considered as the earliest form of the PI3Ks (65). The class III PI3Ks of human tissues consist of one catalytic subunit, Vps34 (vacuolar protein sorting 34), and a regulatory subunit, Vps15 (Figure 1. 13). Class III PI3Ks are proposed to carry out essential functions in controlling intracellular vesicular trafficking and autophagy (60).

Class I domain function

In the p110α subunit, the ABD binds to the p85α subunit, whilst the C2 domain has been proposed to interact with the cell membrane. The RBD is similar to that of other GTPase-RAS-regulated proteins, and mediates attachment of all class I PI3Ks to RAS in a GTP-dependent manner (48, 66, 67). The function of the helical domain of PI3Kα is still unknown; nevertheless, it participates in some interactions of PI3Kα with other proteins, like phosphatase and tensin homolog (PTEN) (68). Lastly, the kinase domain carries out catalytic activity of PI3Kα, that is, the phosphorylation of $PI(4,5)P_2$ (48, 66).

In the p85α subunit, the nSH2 domain forms an inhibitory interaction with the p110α subunit (Figure 1. 14), directly mediating its activity through the binding and dissociation of phosphopeptides. The iSH2 domain acts only as a linker between the nSH2 and cSH2 domains, enabling their interaction with the p110α subunit, since the coiled helix itself does not inhibit the activity of the enzyme (48). The cSH2 domain contacts the RBD, preventing its activation by RAS (69); however, to be an efficient inhibitor, the cSH2 domain needs all other domains to be present. The roles of the rest of the p85α domains are not completely known yet (70).

Class I PI3K activation

There are different ways of activation of class I PI3Ks, which are outlined below.

Activation by receptor tyrosine kinases

Activation of Class IA PI3Ks can be triggered by receptor tyrosine kinases (RTKs) via the direct interaction of the SH2 domains of the regulatory subunit to phospho-YXXM sequences of the RTK (where X can be any amino acid). In the inactive state of the PI3K enzyme, the SH2 domains of the regulatory subunit keep the p110 subunit in a stable but low activity state through interactions with the p110 helical subunit (71, 72). It has been found that p110α only requires an nSH2-mediated inhibitory contact, whereas p110β and p110δ require inhibitory contacts with both the nSH2 and cSH2 domains (71, 73). In RTK-mediated activation, the

phospho-YXXM sequence of the RTK can bind to the SH2 domains of the PI3K and as a result reduce the inhibition of the p110 catalytic subunit (72).

Activation by RAS superfamily of G proteins

The activation of PI3K by RAS happens by direct interaction of the RBD of the p110 catalytic subunit with RAS (48, 67). This stimulation occurs in cooperation with the phospho-YXXM RTK sequence contacting the p85 regulatory subunit (74). According to hydrogen/deuterium exchange mass spectrometry (HDX-MS) and single-molecule total internal reflection fluorescence microscopy (TIRFM) experiments, cooperation of concurrent RTK and RAS stimuli increases the net PI3K activity due to improved membrane recruitment processes (75–77). Additional detail on RAS is provided in Section 1.2.3.

Activation by G protein-coupled receptors

Activation downstream of G protein-coupled receptors occurs in both the PI3Kβ and PI3Kγ isoenzymes, but not for PI3Kα, through direct interaction of the $G_{\beta\gamma}$ complex, a section of heterotrimeric G proteins (77, 78).

Class I PI3K membrane binding

Activation of the p110 catalytic subunit in turn triggers class I PI3K binding to the cell membrane, so that it can catalyse the conversion of its lipid substrate, $PI(4,5)P_2$, to $PI(3,4,5)P_3$. There are different zones of the p110 subunit that have been suggested to interact with the cell membrane. The key binding regions are the kα12 helix (Figure 1. 15 (a)) and the polybasic activation loop in the kinase domain (79). In p110α, deletion of the kα12 helix (residues 1050-1061), which is positioned in the C-terminal region of the p110 kinase domain, or even mutation of the hydrophobic motif TRP-ILE-PHE in that helix (residues 1057-1059) to ALA-ALA-ALA motif (WIF-AAA) both lead to the loss of lipid kinase activity and interaction with the membrane (79). Reduced membrane interaction and anchoring was also observed in MD simulations of the WIF-AAA mutant of PI3Kα (80). MD studies also showed that the position of the kα12 helix affects membrane binding of PI3Kα, with an open form conformation orientation resulting in optimal membrane interaction (80). Together, these results suggest that the kα12 helix, and in particular the WIF motif, are essential for PI3Kα membrane binding and catalysis. In the polybasic activation loop (residues 935-958 of p110α), the positively charged key residues LYS942 and ARG949 are particularly important for $PI(4,5)P_2$ identification by the class I PI3Ks, both experimentally (26, 79, 81) and in MD simulation studies (80, 82). Also,

it has been found that removing the kα12 helix removes the activity of the lipid kinase in p110δ and p110β (73).

There are also other regions and amino acids that affect membrane interaction in class I PI3K proteins. For instance, LYS776 in the p110α kinase domain is a key residue in PI(4,5)P_2 recognition, and together with HIS936 and HIS917, plays an important role in PI3Kα lipid kinase activity (83).

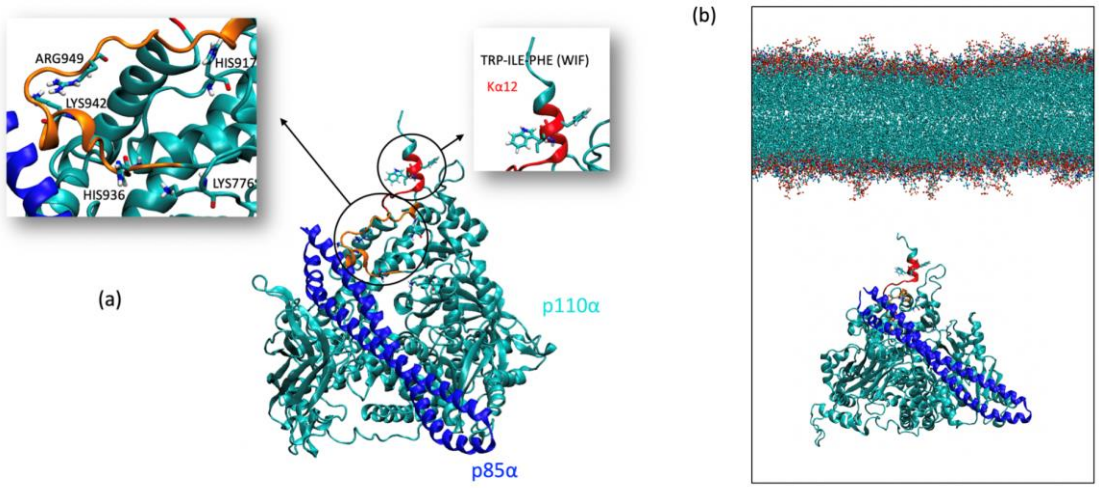


Figure 1. 15. (a) Regions and amino acids that affect membrane interaction in class I PI3K proteins. (b) PI3Kα and membrane; the protein is positioned in a way that the regions involved in membrane-binding are towards membrane. The p110α subunit is in cyan cartoon format and the p85α subunit is in blue cartoon format. The regions containing the key residues are highlighted in orange and red on the cartoon and the key residues are drawn in liquorice format and coloured according to atom type (cyan: carbon; white: hydrogen; red: oxygen and blue: nitrogen).

Class I PI3K signalling

Once PIP_3 is generated by phosphorylation of class I PI3Ks, it binds to a group of proteins with particular lipid-binding domains, such as PH domains, causing the temporary co-localisation of specific signalling proteins. There are two factors required for complete activation of Akt; PIP_3 binds the PH domain of the catalytically inactive Akt, causing a partially active conformation of the protein. Complete activation of Akt needs the phosphorylation of two of its amino acids: THR308, positioned in the activation loop (for the Akt1 isoform), by 3-phosphoinositide dependent protein kinase 1 (PDK1), and SER473 in the carboxyl-terminal hydrophobic motif, by mechanistic target of rapamycin complex 2 (mTORC2) (48). Once activated, Akt phosphorylates several further intracellular proteins, thus controlling a number of downstream signalling events related to the stimulation of cell survival, growth, and proliferation. Examples of such proteins are glycogen synthase kinase-3 (GSK3), mechanistic target of rapamycin (mTOR), the FoxO family of Forkhead transcription factors, Bcl-2-

associated death promoter (Bad), and IkB kinase (IKK) (48, 84–86) (Figure 1. 16). Following the stochastic detachment of $PI(3,4,5)P_3$, Akt takes on an autoinhibited, low-affinity substrate binding conformation, which is promptly dephosphorylated and thus fully inactivated by proteins such as PH domain and leucine-rich repeat protein phosphatases (PHLPPs) (48).

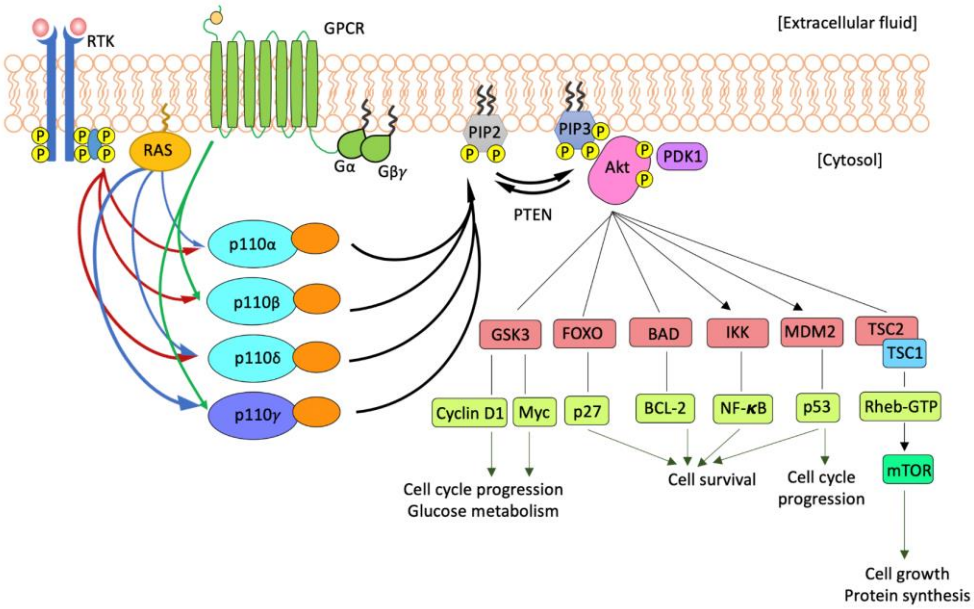


Figure 1. 16. Simplified representation of class I PI3K signalling pathway. Class I PI3Ks are activated by RTKs (with or without adaptor proteins), GPCRs and/or RAS. When they are activated, the class I PI3Ks phosphorylate $PI(4,5)P_2$ to produce $PI(3,4,5)P_3$. The reverse conversion is regulated by the phosphatase PTEN. $PI(3,4,5)P_3$ recruits and binds Akt, which in turn phosphorylates many proteins, causes the promotion of cell survival (anti-apoptosis), cell cycle progression, glucose metabolism, cell growth and protein synthesis. Lines represent inhibition, and pointed arrows indicate activation.

Apart from Akt signalling, the main downstream class I PI3K signalling pathway, there are also other PH-domain containing proteins, such as phospholipase Cγ (PLCγ) (87), Bruton's tyrosine kinase (Btk) (88, 89), and GEFs and GTPase-activating proteins (GAPs) for small GTPases of the RHO, RAS, RAC, and ADP-ribosylation factor (Arf) families (90, 91), which bind to $PI(3,4,5)P_3$. Many of these effectors can therefore join the cell membrane through the conversion of $PI(4,5)P_2$ to $PI(3,4,5)P_3$ by class I PI3Ks. This, eventually, causes a variety of responses, such as cell growth, differentiation, proliferation, survival and motility (45).

1.2.2.4. PI3K mutation and cancer

The PI3K pathway is one of the most commonly mutated signalling pathways in human cancers. Particularly, mutation of the *PIK3CA* gene (which encodes p110α) is the most often genetic alteration in this pathway, and causes cancer types, like colorectal cancer, endometrial cancer, breast cancer, and ovarian cancer (92–94). More than 1000 mutations of PIK3CA have

been found (95). These mutations mainly occur in several domains of p110α, with the three most common ones located in two "hotspots": the C-terminal lobe of the kinase domain (H1047R) and the helical domain (E542K and E545K) (94). These mutations result in increased catalytic activity of p110α, leading to enhanced downstream signalling of the Akt pathway mentioned above (96), making PI3Ks a key target for cancer therapy.

MD simulation studies comparing wild-type and the oncogenic mutant H1047R PI3Kα in solution (97) showed that this mutation causes exposure of a large number of basic residues on the membrane-binding surface of PI3Kα; subsequent simulations of wild-type and H1047R PI3Kα in the presence of a model cell membrane showed that this enhances its interactions with negatively-charged lipids such as POPS and thus facilitates membrane binding (80, 98). These investigations of PI3Kα membrane binding thus far have involved only PI3Kα itself, without key cellular partners such as RAS.

1.2.3. RAS

RAS proteins are defined as small GTPases, and are encoded by three RAS proto-oncogenes (HRAS, KRAS, and NRAS in *Homo sapiens*). There are various important cellular functions and signaling pathways that RAS genes are involved in modulating, including cell proliferation, growth, differentiation, cell cycle, and motility (99). The three RAS genes comprise the most commonly mutated oncogene family in human cancers, since the RAS genes are mutated in 33% of all cancers evaluated (100). Residues G12, G13, and Q61 are the three common spots for RAS mutational activation, which are involved in more than 95% of identified mutations (101). Therefore, developing therapeutic approaches for preventing unusual RAS function has been the target of considerable interest and attempts at developing cancer treatments (102).

RAS proteins play the role of binary switches between the active and inactive forms, RAS-GTP (or GNP) and RAS-GDP, respectively. Mg^{2+} makes a complex with the two phosphate groups of the ligand (GTP/GDP) closest to the nucleoside, and is needed for high affinity binding between the protein and the nucleotide (102). GNP, a non-hydrolysable analogue of GTP in which a nitrogen atom binds to the two phosphate groups instead of an oxygen (as in GTP), also binds tightly to RAS in the presence of the Mg^{2+} (67) (Figure 1. 17), and has been useful for studying RAS *in vitro*.

In the GTP-bound (or GNP-bound) or other words, the "ON state", RAS proteins interact with effectors, such as PI3Ks, an example of the role of protein-protein interactions in

cellular function. RAS-GTP binds directly with its effectors through two flexible regions named switch I and switch II (Figure 1. 17), which comprise residues 30-40 and 60-76, respectively, leading to activation of downstream signalling (67, 103–105).

Post-translational modifications with lipids, such as C-terminal prenylation and palmitoylation, are required for the function of RAS in a wide range of signal transduction processes (106). Lipids are added to a series of residues in the RAS C-terminal region, called the hypervariable linker domain (107) (Figure 1. 17). These lipids embed into the bilayer and so anchor RAS to the membrane, an example of protein-membrane interaction. Further details regarding the nature of these lipids and their attachment to RAS are provided in chapter 5, Section 5.2.2 and 5.3.1.3.

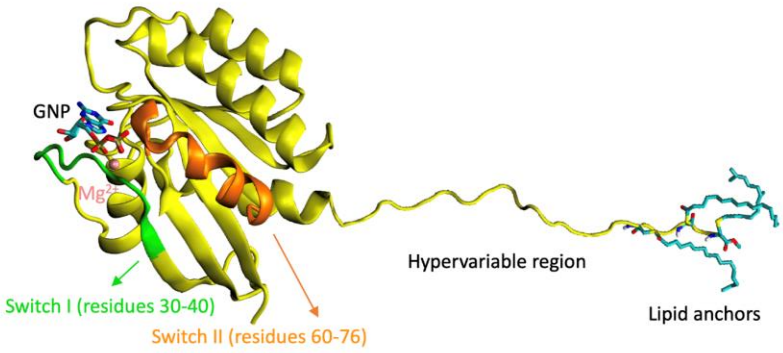


Figure 1. 17. Cartoon structure of GNP-RAS (with Mg^{2+}) with the post-translational modifications to the hypervariable linker domain. RAS is drawn in yellow cartoon format with the switch I and switch II regions coloured green and orange, respectively. GNP and the lipid anchors are drawn in liquorice format and coloured according to atom type (cyan: carbon; white: hydrogen, red: oxygen; blue: nitrogen).

1.2.4. KstR

1.2.4.1. Tuberculosis

Tuberculosis (TB) is a contagious bacterial infection which is a leading cause of death annually (108). It has been estimated that 10.4 million people suffered from TB in 2016, of which 6.3 million were diagnosed in 2016 (108, 109). There are approximately 300 cases of TB detected annually in New Zealand (110). TB is caused by *Mycobacterium tuberculosis* (Mtb), which is a pathogenic bacterial species in the Mycobacteriaceae family. The target of Mtb is mainly the respiratory system, as it thrives in the lungs, where it causes infection (111). Mtb cycles between activity and quiescence, and this cycling is vital for its persistence (112). TB has been considered as a difficult disease to treat, due to the antibiotic adaptations and drug resistance of Mtb. Many different antibiotic treatments have therefore been investigated (111).

Although there is a variety of drugs to combat TB, the multidrug-resistant nature of Mtb means that there is strong motivation to identify new strategies that could possibly overcome Mtb drug-resistance, or at least decelerate the spread of TB (113, 114). One avenue for the development of new drugs is the fact that the persistence of Mtb is highly dependent on lipid metabolism (115, 116), mostly utilising fatty acids and cholesterol as its energy and carbon source (117). In the course of infection, Mtb is dependent on cholesterol (118, 119), which is catabolised through the β-oxidative and energy production metabolic pathways. Fatty acids are catabolised by about 250 different enzymes via the methylcitrate cycle, methylmalonyl pathway, and tricarboxylic acid cycle (115, 118, 120). To achieve this, Mtb produces enzymes that stimulate the breakdown of the host membrane layers, controlling host cellular behaviour, the lipid cell envelope synthesis, and detoxification of oxidative stress (111, 121).

The expression of the cholesterol catabolism enzymes is principally regulated by KstR, a Tetracycline Receptor (TetR) protein that is therefore of great interest for development of new TB treatments, which will be discussed later in Section 1.2.4.3.

1.2.4.2. Overview of TetR family proteins

TetR is a family of bacterial proteins whose main function is regulation of the efflux of antibiotics (122, 123). At least one protein of the TetR family has been found in every sequenced bacterium, while bacteria which have larger genomes possess multiple homologues (124). Features of TetR proteins are found in other prokaryotic proteins, which are categorised as members of the Tetracyline Family of (Transcriptional) Regulators (TF(T)R) (125, 126). TFRs have a common structure, being mostly α-helical, and including a ligand binding domain (LBD) and a DNA binding domain (DBD) (127) (Figure 1. 21).

In addition to antibiotic efflux regulators like TetR comprising 25% of the TFR [39], different TFR members have other functions as: metabolic gene control, antibiotic production, and regulation of prokaryotic physiology (126, 128, 129). Some TFR members function as both activators or repressors, local or global regulators (130), and interact with ligands and proteins of various sizes, as well as binding to DNA (127). In spite of the vast range of research invested on TFRs, the dynamics and how the conformational motions of these protein contribute to their function remain elusive.

1.2.4.3. KstR

Bioinformatic analysis of microarray data led to the identification of KstR, which is a transcriptional repressor that regulates the expression of cholesterol degradation machinery in Mtb (131, 132). KstR activity depends on its homodimeric structure, like all other TFR proteins (133). Each monomer comprises a DNA binding domain (DBD) and a ligand binding domain (LBD) (134). The DBD interacts with particular motifs in the Mtb genome, recognising a specific curvature induced by the DNA sequence (135). Multiple ligands have been shown to bind to the ligand-binding pocket in the LBD, including early stage coenzyme-A (CoA) thioester metabolites of cholesterol, 3-oxo-4- cholestenoic-CoA (3OChA-CoA) and 3-oxo-23,24-bisnorchol-4-en-22-oyl-CoA (4-BNC-CoA) (136–138). This ligand binding inhibits DNA binding and thus allows transcription to occur (138).

KstR gene control

KstR is a transcriptional repressor, which controls the expression of the approximately 74 genes in the KstR regulon, including genes expressing fatty acid metabolising and cholesterol metabolising enzymes, as well as KstR itself (132, 139, 140). KstR binds to a short palindromic DNA sequence (119, 140, 141), namely [TnnAACnnGTTnnA] that is conserved among Mycobacterial species (132). Dimeric KstR was shown by electrophoretic mobility shift assays to bind specifically to this DNA sequence (132, 140).

KstR binding to the DNA inhibits ribonucleic acid (RNA) polymerase access to the genes' promoter region, therefore preventing gene expression (Figure 1. 18). The expression of these genes is stimulated by cholesterol and other lipids (132, 136). Therefore, KstR can be considered as a cholesterol indicator, since when there is a lack of cholesterol, the protein represses the transcription of genes involved in cholesterol catabolism, while in the presence of cholesterol, it induces cholesterol catabolism (138).

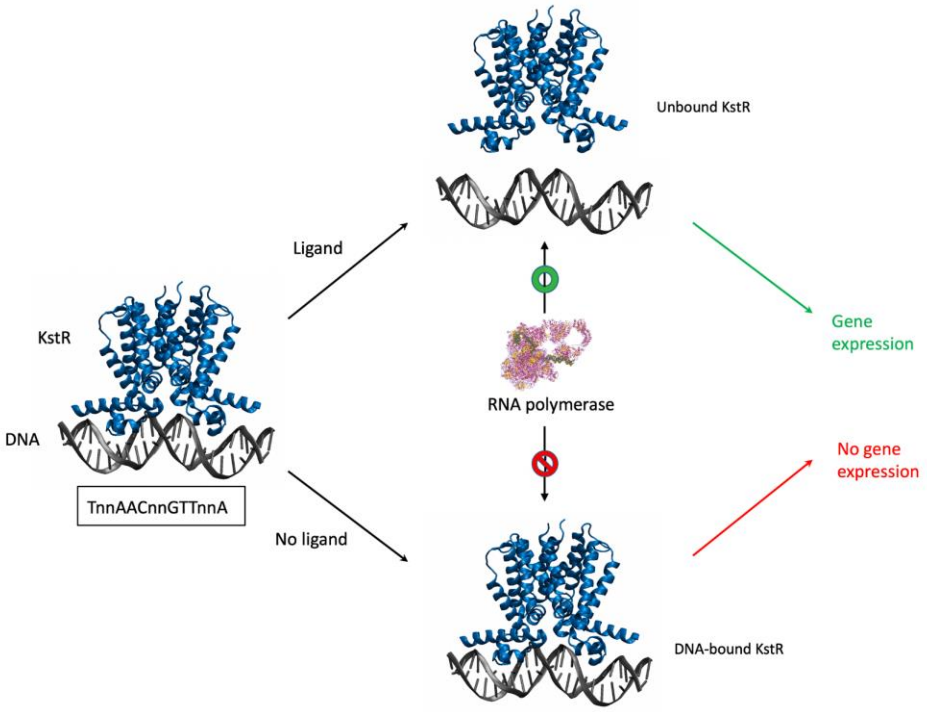


Figure 1. 18. Summary of inhibitory mechanism of KstR, and how the protein affects expression of genes in the presence and absence of ligand.

KstR structure

Structures of KstR have been solved by X-ray crystallography in three different states: bound to DNA, bound to ligands (3OChA-CoA and BNC-CoA, separately), and in its apo state (142). When bound to DNA, which inhibits transcription, KstR undertakes a conformation in which the two DBDs are close together, sitting in the major grooves of the DNA, while the ligand-bound structure takes on a conformation with the two DBDs further apart (143) . The apo conformation in between these (138, 139) (Figure 1. 19). Although the three states are well defined, the process by which KstR transitions between them is not known.

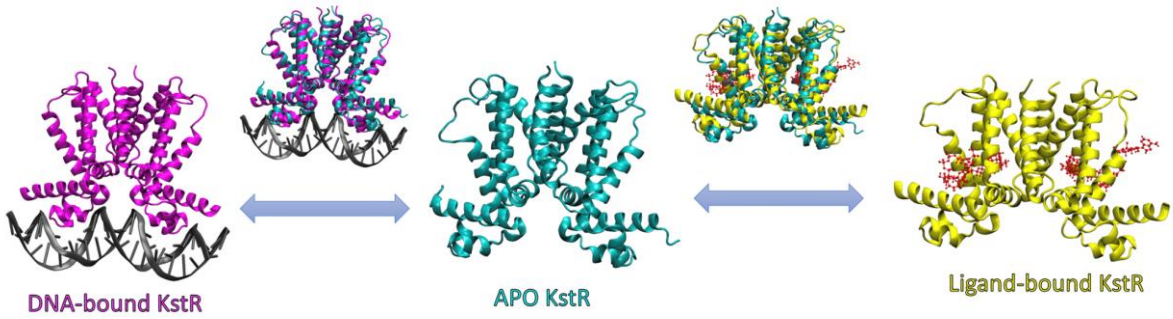


Figure 1. 19. Cartoon representation of proposed KstR transition model. DNA-bound KstR (PDB ID: 5UA2, magenta), apo KstR (PDB ID: 3MNL, cyan), and ligand(3OCh)-bound KstR (PDB ID: 5CW8, yellow). The superimpositions of the structures are shown on the arrows to indicate the transition between states.

The structure of KstR largely comprises α-helices, which form a clamp-like homodimeric complex (134, 144). Each monomer has nine α-helices, connected to each other by coiled loops. The DBD and LBD are connected by a linker sequence between the α4 and α5 helices. The DBD, which is located at the N-terminus of KstR, contains helices α1-α3. The α2 and α3 helices generate a helix-turn-helix (HTH) motif, representative of TFR proteins, and, more broadly, of DNA binding proteins (122, 138). The α4 helix acts as a channel between the two domains, anchoring the linker sequence between α5 and itself. The C-terminal LBD comprises helices α4-α9, which form a hydrophobic cavity into which the ligand binds. The ligand pocket is made up of the triangular groove between helices α5-α7, as well as the hydrophobic residues of helix α9 (Figure 1. 20) (138). The total surface area of each monomer of KstR is 9500 $Å^2$, of which 1099 $Å^2$ is the dimer interface, formed by the α6-α9 helices.

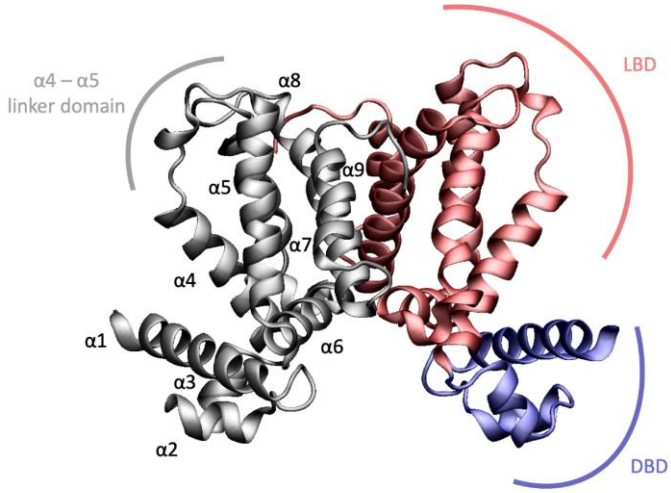


Figure 1. 21. Cartoon representation of KstR (PDB: 3MNL) dimer with α-helices annotated in one monomer (silver), and the DBD (blue) and LBD (red) regions indicated in the second monomer.

Apo KstR

In the crystal structure of the apo conformation of KstR (PDB ID: 3MNL), the distance between the two DBDs is about 3.8 nm, which is in between the inter-DBD distances of the DNA-bound and ligand-bound forms (explained in the following sections). It has been hypothesised that in this state, the DBDs should be quite mobile. The DBDs are smaller than the LBDs (3 α-helices *c.f.* 9), contain fewer hydrophobic amino acids and in turn fewer restraints by local chemical and structural forces, and are located towards the periphery of the structure and so more exposed to the aqueous solvent. Furthermore, the connection between the DBD and LBD

includes a long coiled linker domain between α4 and α5. Altogether, these factors suggest that the DBDs could be highly mobile when they are not bound to DNA (143).

DNA-bound KstR

In its DNA-bound state, each DBD of KstR binds to sequential major grooves in the DNA helix, but is unusual in that the majority of the interactions are with the backbone of the DNA rather than the bases (135, 142). The interactions between KstR and DNA are hydrogen bonds (H-bonds) and van der Waals (VDW) forces (Figure 1. 22). The α1 helix of the DBD and the N-terminal loop of α4 associate with the phosphate backbone of opposite sides of the DNA groove. The phosphate groups of a cytidine and a thymidine on one side of the DNA groove form hydrogen bonds with ARG17, THR52, and TYR56 (side chains) and ALA49 (backbone nitrogen) (Figure 1. 22). There are also hydrogen bonds between the side chains of residues GLN38, TYR54 and the side chain and backbone of LYS60 and the phosphate groups of a cytidine, thymidine, and guanidine, on the other side of the major groove. The only strong contacts between KstR and a DNA base are the two hydrogen bonds between the side chain of ARG40 and a guanidine. Otherwise, there is a weak hydrophobic interaction between the pyrimidine ring of a thymidine and the VAL50 side chain (142), and a potential CH···π interaction between the aromatic side chain of TYR54 and the methyl group on the pyrimidine ring of a thymidine.

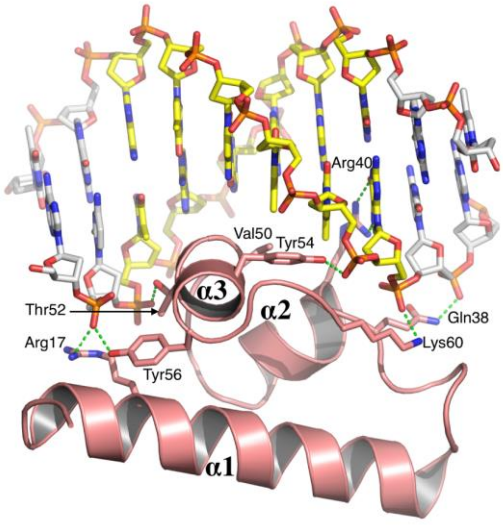


Figure 1. 22. Hydrogen bonds between KstR and DNA observed in the X-ray crystal structure (PDB ID: 5UA2, (135)). The protein is drawn in cartoon format and coloured pink, with residues making side-chain contact with the DNA shown in liquorice representation (note that ALA49 is hidden in this image). The DNA is drawn in liquorice representation, with one strand in yellow and the other in white. Oxygen atoms are in red, nitrogen atoms in blue and phosphate atoms in orange. Hydrogen bonds are shown as green dashed lines.

The dissociation constant (K_d) between KstR and DNA (10.5 ± 0.6 nM) suggests that there is tight binding between the protein and DNA, and that KstR has a high affinity for DNA without ligand present (142).

In DNA-bound KstR, the distance between the two DBDs of the monomers is about 3.2 nm, representing a compact protein conformation in comparison with the apo and ligand-bound states. This contracted conformation of the DNA-bound state is warped due to DNA pulling the DBDs of KstR together, resulting in a twist in the LBDs (138).

KstR ligands

Gene regulation by KstR is modulated by at least two metabolites generated early in cholesterol degradation, 3OChA-CoA and 4-BNC-CoA. It is thought that these compounds act as cholesterol catabolism signals instead of cholesterol itself because they are more soluble than cholesterol (137).

3OChA-CoA has the higher affinity for KstR (K_d =0.06 ± 0.02 µM), compared to K_d = 0.28 ± 0.05 µM for 4-BNC-CoA (138). It is even higher than the affinity of 3OChA, without the CoA moiety (K_d = 1.2 ± 0.4 µM).

3OChA-CoA is the product of the C3 cholesterol oxidation step early in the cholesterol degradation pathway, which is catalysed by the enzyme 3-β-hydroxysteroid dehydrogenase (3-β-HSD) (Figure 1. 23) (136, 137). 3-β-HSD is among the proteins for which gene expression is repressed by KstR, but a baseline expression persists in the absence of cholesterol (145), thus ensuring that some 3OChA-CoA is still produced (139). The enzyme that catalyses production of 4-BNC-CoA in the cholesterol degradation pathway is called 3-ketosteroid 9a-hydroxylase (KshAB) (146).

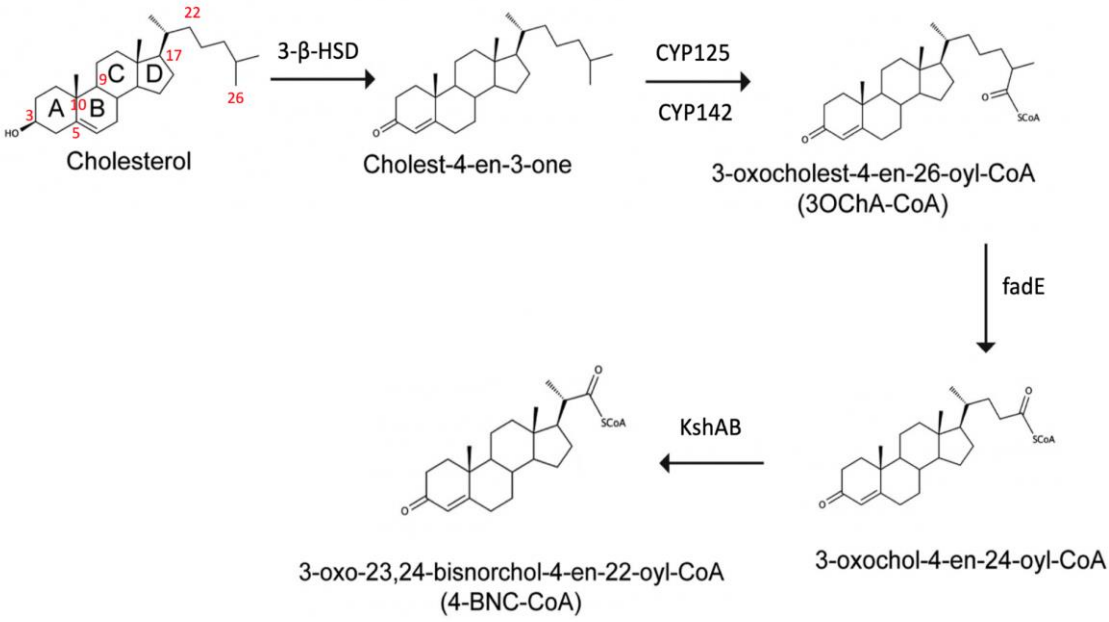


Figure 1. 23. Schematic of early stages of cholesterol degradation pathway towards 3OChA- CoA and 4-BNC-CoA in Mtb. [Edited from (138)].

Ligand-bound KstR

The ligand binding pocket of KstR is surrounded by helices α4, α5, α7 and α8 of the LBD, and lined with hydrophobic residues. The pocket entrance is bordered by the α4, α5, and part of the α9 helices, and by amino acids ARG102, ARG106, LYS99, ARG70, ARG74, and LYS78 (138). Depending on which ligand binds into the pocket, the ligand-interacting amino acids are different. However, generally theses residues include LEU68, GLN71, PHE72, ISO75, LEU100, ALA103, MET104, ASP107, LEU110, THR111, MET114, TYR118, ISO136, ARG158, SER161, TRP164, and LEU168, as well as the key ligand-binding residue ARG158, located at the end of the pocket. ARG158 interacts with the ketone oxygen of both 3OChA-CoA (PDB: 5CW8) and 4-BNC-CoA (PDB: 5CXI) (138).

Ligand binding brings about conformational changes across KstR, therefore, the LBD functions as an allosteric binding site (147, 148). These conformational changes include a rotation and a translational shift of the four-helix bundle of α6-α9, shifts in the α5-α7 group of helices, shifts in the α4 helix, and also the entire DBD.

1.2.4.4. Mechanism of action of KstR

Existing knowledge about the mechanism of action of KstR is mostly revealed from studying apo KstR (132). As with the classic categorization of ligand binding by proteins, there are two mechanisms by which apo KstR might bind to DNA: induced fit or conformational selection (Figure 1. 24) (149, 150). The induced fit model proposes that DNA stimulates the transition of KstR into its DNA-bound state via a two-step binding process: First, one DBD of one monomer of apo KstR interacts with and binds to the DNA, forming a half DNA-bound KstR system. Second, this partially bound complex brings the second KstR subunit into proximity with the DNA so that the second DBD also binds to the DNA (Figure 1. 24 (a)). On the other hand, conformational selection suggests that among the dynamic apo KstR dimers, some conformations are predisposed to bind to DNA, e.g. due to the DBDs being close together like in the DNA-bound state, and are then likely to bind to DNA if the two molecules (KstR and DNA) are close (Figure 1. 24 (b)) (151). Similarly, an induced fit mechanism for the reverse process implies that KstR's ligands can bind to KstR while it is bound to DNA and induce it to unbind from DNA, whereas a conformational selection model would involve KstR spontaneously unbinding from DNA in a stochastic manner and the ligands binding to KstR in its apo state and thus preventing rebinding to DNA. Both models predict fluctuation between DNA-bound, ligand-bound and apo KstR, but via a different mechanism. This can be difficult to distinguish experimentally, but may be possible to delineate using MD simulations in the presence and absence of DNA and KstR's ligands.

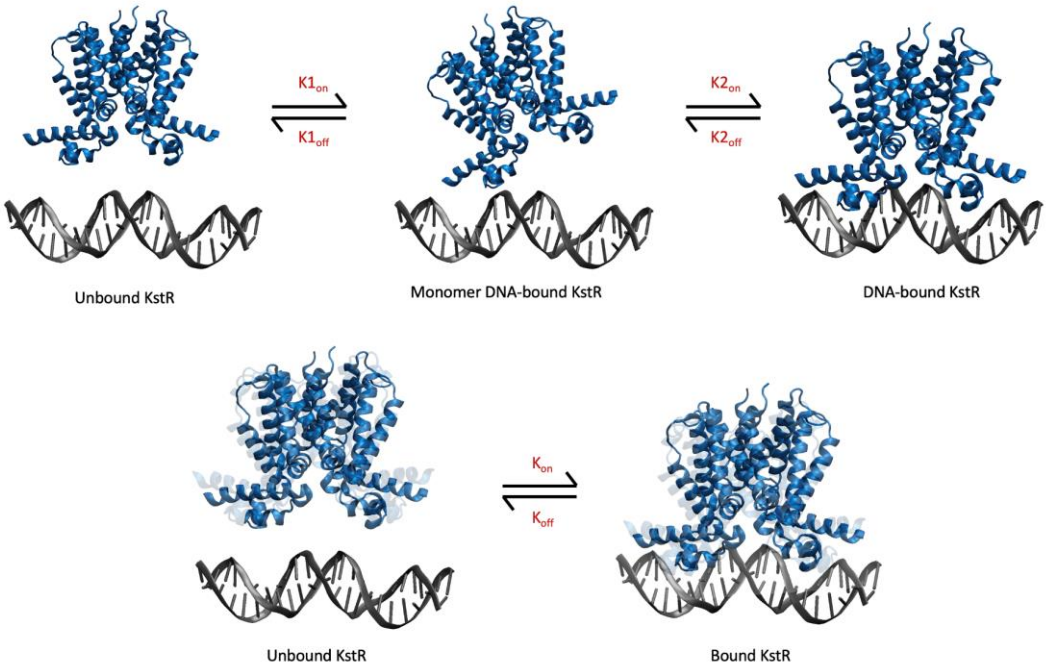


Figure 1. 24. Diagram of the two binding models: (top) two-step induced- fit binding model; (bottom) conformational selection binding model. Transparent structures indicate different possible conformational states.

1.2.5. KstR2

KstR2 is another TFR protein, which, like KstR, also regulates the cholesterol degradation metabolic pathway in Mtb. KstR is required for Mtb pathogenesis as well as regulating the initial steps in cholesterol degradation (152), showing that appropriate regulation of cholesterol metabolism is required for Mtb to cause disease. KstR2, in contrast, is not essential for *in vivo* survival of Mtb; nevertheless, genes in the KstR2 regulon are essential for persistence in a murine model of TB (152). KstR2 represses a 14-gene regulon involved in the later steps of cholesterol degradation, especially degradation of the C and D sterol rings (139, 153, 154).

Unlike KstR, which binds to DNA as a dimer, KstR2 binds to DNA as a dimer of dimers: the KstR2 operator sequence binds two KstR2 dimers to form a complex of two KstR2Mtb dimers and one DNA fragment (137). There are no crystal structures of KstR2 bound to DNA, however.

The entry compound to the KstR2-regulated pathway is 3aα-H-4α (3'-propanoate)-7aβ-methylhexahydro-1,5-indanedione (HIP). KstR2 binds to the CoA-thioester of HIP (S-[2-[3-[[(2R)-4-[[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl]oxy-oxidanyl-phosphoryl]oxy-3,3-dimethyl-2-oxidanyl-butanoyl]amino]propanoylamino]ethyl] 3-[(3aS,4S,7aS)-7a-methyl-1,5-bis(oxidanylidene)-2,3,3a,4,6,7-hexahydroinden-4-yl]propanethioate) or HIP-CoA, the product of the enzyme

FadD3 (Figure 1. 25) (128, 153, 154) (see Chapter 4, Section 4.2.2). Binding of KstR2 to HIP-CoA prevents it from binding to DNA and thus inhibiting transcription of the genes in its regulon.

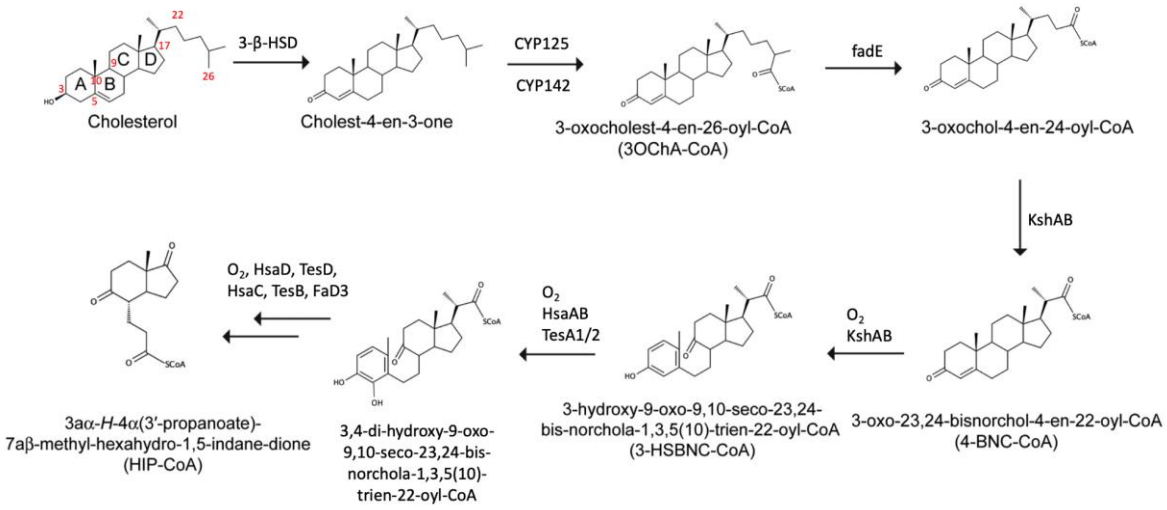


Figure 1. 25. Schematic of the cholesterol degradation process towards HIP-CoA in Mtb [edited from (155)].

1.2.5.1 Structure of KstR2

Crystal structures are available of KstR2 from Mtb in its apo state and bound to the HIP portion of HIP-CoA. There is also a ligand-free form of KstR2 from *Rhodococcus jostii* (strain RHA1) (Rjo), a Gram-positive polychlorinated biphenyl (PCB) degrader ,which shares 59% amino acid sequence identity with Mtb KstR2 (137, 156). The structure of KstR2 is similar overall to that of KstR and of other TFRs. Its all-helical structure is typical of TFRs in that it makes a horseshoe-like homo-dimer consisting of an N-terminal DBD, comprising the α1-α3 helices plus the N-terminus of helix α4, and a C-terminal LBD, comprising helices α5-α9 plus the C-terminus of helix α4. The LBD also provides two helices (α8, α9) to the dimer interface (157) (Figure 1. 26).

Each monomer of the KstR2 dimer has a binding pocket for HIP-CoA in its LBD. The hydrophobic HIP portion predominantly interacts with the hydrophobic residues such as PHE65, LEU66, LEU69, PHE70, TYR73, and VAL105 in the binding pocket, whereas the polar CoA portion, forms hydrogen bonds with residues ARG162, GLN109, ASP163, and ASN128 (137, 158).

Because there are no crystal structures of KstR2 bound to DNA, the details of how it does so are not yet known.

1.2.5.2 Mechanism of action of KstR2

The mechanism by which ligand and DNA binding are communicated, and whether KstR2 operates *via* an induced fit or conformational selection mechanism, are not known. It has been hypothesised (157), however, that KstR2 might act via a double scissor allosteric mechanism that, if correct, would be novel in the TetR family.

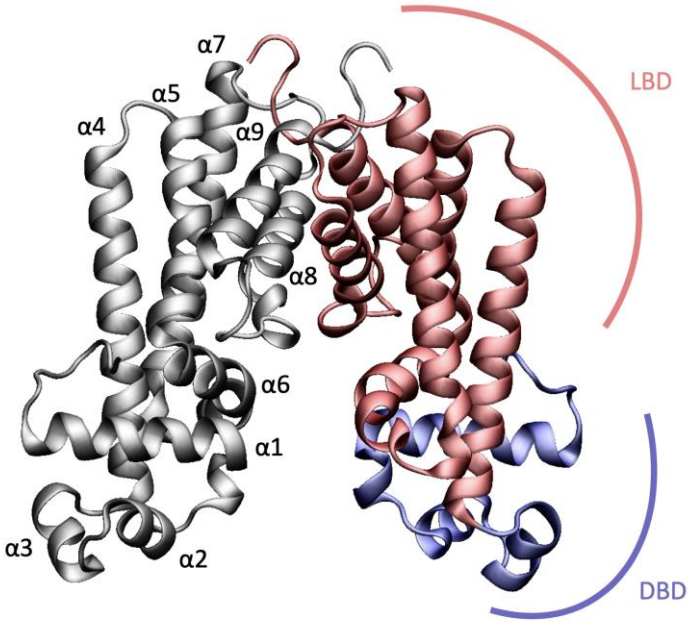


Figure 1. 26. Cartoon representation of the KstR2 (PDB: 4W1U) dimer with the α-helices annotated in one monomer (silver), and the DBD (blue) and LBD (red) regions indicated in the second monomer.

1.3. Advantages of Computational Representation

Current experimental techniques mainly lead to models of molecular structures that, while at atomic-level detail, are static, representing either a solid (crystalline) state or a single structure fitted to time- and ensemble-averaged data (159, 160). MD simulations initiated from such structures provide insight into conformational dynamics and, potentially, transitions between different structures and thus insight into molecular mechanisms.

For instance, for bullvalene-cyclodextrin complexes, MD simulations can explore binding modes of the host and guest molecules, as well as the most favourable binding orientation and position, as well as allowing calculation of binding free energies. For PI3Kα, MD may provide information and insights about how the interaction of PI3Kα with the cell membrane is affected by the presence of RAS, including aspects such as the orientation and any conformational changes of the protein complex as it approaches the lipid membrane. There

have been only a limited number of studies involving MD simulations of PI3Kα, either the wild type p110α or its mutants, so far (80, 97, 98, 161–163), none of which have involved its cellular partners such as RAS. MD simulations are also capable of investigating the conformational changes of KstR in the presence and absence of its ligands and DNA, in order to probe its allosteric mechanism, and the effect on KstR2 of mutating key residues.

For fluctional systems such as bullvalenes, MD provides an opportunity to study individual isomers that are impossible to separate experimentally, and both predict experimental results as well as aid the interpretation of experimental data to which the contribution of individual isomers would otherwise be indistinguishable. In the case of protein systems, the knowledge and insight provided by MD simulations extends what can be determined from the structures alone, and is able to both test hypotheses based on structural studies as well as propose additional hypotheses that can be tested experimentally.

1.4. Molecular Dynamics

Molecular dynamics (MD) is a classical computational simulation method used to simulate the time-dependent behaviour of particles. MD is often used to simulate the conformational changes of molecules by representing each atom with a particle. It does so by solving Newton's equations of motion for a system of particles by numerical integration of the equations to calculate the new positions of the particles after a certain time step (164)(165). This relies on having initial positions and (usually estimated) velocities, and a formula for computing the potential energy of the system, *U*, from the particle positions.

MD makes use of Newton's 2nd law of motion (Equation 1.1),

$$F_i = m_i a_i = m_i \frac{\partial^2 r_i}{\partial t^2} \quad ,$$

Equation 1.1

where F_i is the force on atom *i*, m_i is its mass and depends on atom type, a_i is the acceleration of particle *i* due to the force, r_i is the particle position and *t* is time.

The force on particle *i*, F_i, is given by the negative derivative of the potential energy function *U* with respect to the particle position r_i (Equation 1.2).

$$F_i = - \frac{\partial U}{\partial r_i}$$

Equation 1.2

By combining Equation 1.1 and Equation 1.2, we have Equation 1.3, which relates the acceleration of particle *i* with the derivative of the potential energy with respect to its current position.

$$- \frac{\delta U}{\delta r_i} = m_i \frac{\partial^2 r_i}{\partial t^2}$$

Equation 1.3

By integrating Equation 3 with respect to time, the change in velocity and thus the change in the position of the particle over time are calculated. The process for doing so uses numerical integration. A common form of numerical integrator utilised in MD simulations is *leapfrog* integration, which updates the positions *r* and velocities *v* at interleaved time points, so that they leap over one another. Given a time step Δt , positions r_i and velocities v_{i+1} , the positions at *i+1* are given by Equation 1.4, and the velocities are given by
.

$$r_{i+1} = r_i + v_{i+1/2} \Delta t$$

Equation 1.4

$$r_{i+1} = r_i + v_{i+1/2} \Delta t$$

Equation 1.5

The diagram in Figure 1. 27 summarises the process of a MD simulation. The major steps are described in more detail below in Sections 1.4.1 to 1.4.5, and the nature of the force field that gives rise to the potential energy *U* is outlined in Section 1.4.6.

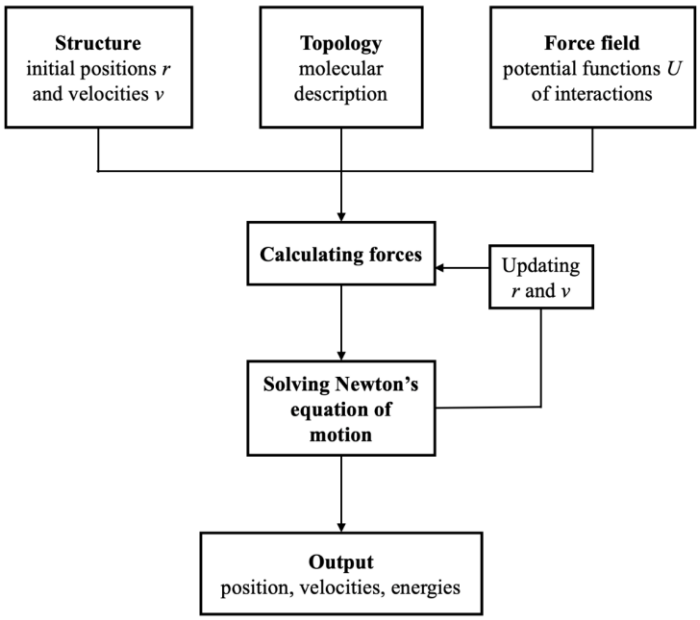


Figure 1. 27. The workflow of MD simulation.

1.4.1. Initial conditions

The initial coordinates for an MD simulation can be obtained in a number of ways. For small molecules, coordinates may be available from structural databases such as the Cambridge Crystallographic Database (166). Another option is to use the output from quantum mechanical structure optimisation. For macromolecules such as proteins, experimentally determined structures from databases such as the Protein DataBank (PDB) (167) are preferable. In their absence, it may be possible to predict the structure using homology modelling, but these should be interpreted with caution, particularly when homology to known structures is low.

Particle velocities are also required to initiate an MD simulation, as the integration algorithms rely on adjustment of existing velocities by the acceleration due to the calculated forces on the atoms due to their current positions. Because velocities are seldom known, they are typically assigned from a Boltzmann distribution at a low temperature. This ensures that contradictory velocities, for instance bonded atoms being assigned velocities in opposite directions, are small and thus unlikely to strongly perturb the system. The system is then equilibrated briefly at this low temperature to allow the velocities to become correlated before slowly heating it to the desired temperature. It is common to additionally (or sometimes, alternatively) restrain the atoms to their initial positions during this heating phase as a further means of avoiding anomalous atomic motions.

1.4.2. Solvation

The molecule or molecules whose conformational motion is to be followed using MD seldom exist in isolation. The amount of environment that can be included in a simulation is limited by the computational cost. Additionally, much of the environment is not of direct interest. Typically, these limitations are circumvented by surrounding the molecule(s) of interest with solvent molecules, often water. The amount of water to be added is a compromise between ensuring there is sufficient environment surrounding the molecule(s) of interest and the computational cost of simulating a larger system.

1.4.3. Periodic boundary conditions

A procedure that extends the effective solvation shell surrounding the molecule(s) of interest while reducing the computational cost is the use of periodic boundary conditions (PBC). PBCs are a set of boundary conditions that are applied to a finite sized system to approximate a large (infinite) system. This finite size system is called the unit cell, which can be thought of as a simulation 'box' in which the molecule(s) of interest are placed and then the remaining space filled with solvent molecules. The box must be of a shape that can be tessellated in three-dimensional space. Replication of the unit cell via translation in all three directions forms an infinite lattice, as shown in Figure 1. 28.

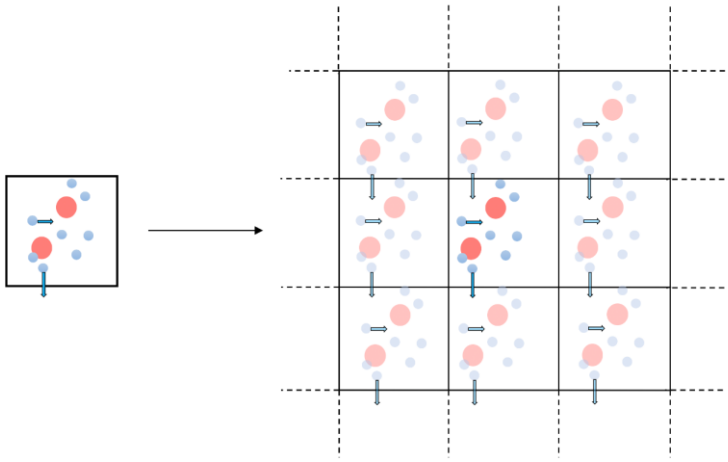


Figure 1. 28. Schematic representation of the idea of periodic boundary conditions used to mimic an infinite system. The image on the left shows a box of molecules with fixed boundary conditions. On the right, the same box is shown surrounded by identical images of itself (in two dimensions only), giving rise to periodic boundary conditions.

1.4.4. Calculation of interactions

The size of the unit cell is governed by the distance to which interactions between particles are calculated. Typically, not all interactions are calculated, but only those between particles within some cut-off distance, to reduce the computational effort. Outside of this, the interactions are

either not calculated, as their contribution is small (e.g. Lennard-Jones interactions, see Section 1.4.6.2), or are approximated using a continuum or reciprocal space method (Coulombic interactions, see Section 1.4.6.2). The distance from a particle to its periodic copy must be less than this cut-off distance so that particles do not interact with one another.

1.4.5. Thermodynamic ensembles

An MD simulation may be run in a number of different thermodynamic ensembles. In the microcanonical (NVE) ensemble, the number of particles (N), volume (V), and energy (E) are conserved. This is the 'default' ensemble for MD, and mimics situations such as an adiabatic process without any heat exchange. This ensemble is not representative of the conditions in which biological systems function, however. Therefore, other ensembles, either NVT or NPT, are frequently used to model biological systems.

In the canonical ensemble (NVT), the number of particles (N), volume (V) and temperature (T) are kept fixed. It is also sometimes called constant temperature molecular dynamics (CTMD). The temperature of the system is maintained using a thermostat, such as the Berendsen thermostat (168), Nosé-Hoover thermostat (169), Nosé-Hoover chains (170), Andersen thermostat and Langevin dynamics (171), (172).

The Berendsen thermostat, which is the simplest, is an algorithm to rescale the velocities of particles in MD simulations, which controls the simulation temperature due to the link between the kinetic energy, the particle velocities and the system temperature. Equation 1.6 shows the instantaneous temperature $\mathcal{T}$ of a system at time point t depends on the kinetic energy present within the system at that time.

$$\mathcal{T}(t) = \sum_{i=1}^{N} \frac{m_i v_i^2(t)}{k_B N_f}$$

Equation 1.6

In Equation 1.6, $v_i(t)$ is the velocity of particle i, N is the total number of particles in the system and N_f is the number of degrees of freedom. Changing the velocities is a natural way of modifying the temperature. The Berendsen thermostat mimics weak coupling with first order kinetics to an external heat bath with a given temperature T_0 (168). The deviation of the system temperature from T_0 is slowly corrected according to Equation 1.7,

$$\frac{d\mathcal{T}}{dt} = \frac{T_0 - \mathcal{T}}{\tau} \ ,$$

Equation 1.7

which means that a temperature deviation decays exponentially with a time constant τ.

In the isothermal–isobaric (NPT) ensemble, the number of particles (N), pressure (P) and temperature (T) are conserved. A barostat is therefore needed as well as a thermostat. The simulation will be much closer to laboratory conditions since both temperature and pressure are controlled. Some commonly used barostats are the Berendsen barostat, Andersen barostat, Nosé-Hoover barostat, Martyna-Tuckerman-Klein barostat, Nosé-Hoover Langevin piston barostat, and the Parrinello-Rahman barostat (173).

The Berendsen barostat, which is used in this study, rescales the system volume and (optionally) the atoms' coordinates within the simulation box every time step (168). The instantaneous pressure *P*, at time point *t* can be determined according to Equation 1.8.

$$P(t) = \frac{1}{V}\left(\frac{1}{3}\sum_{i=1}^{N} m_i v_i^2 + \mathrm{r}_i \mathrm{f}_i\right)$$

Equation 1.8

In this equation, r_i is the position of atom *i* and F_i is the force acting on it. The force is calculated as part of the integration of Newton's equations of motion (Equation 1.1). This dependence on position means that the pressure can be affected by scaling the box size and consequently the atomic position. For a given pressure change Δ*P*, the required change in volume is related by Equation 1.9, where k_T is the isothermal compressibility of the system.

$$\Delta P\,(t) = \frac{-\Delta V\,(t)}{\kappa_T V\,(t)}$$

Equation 1.9

1.4.6. Force fields

The potential energy of a system in a molecular dynamics simulation is calculated using a so-called "force field", whose name reflects the fact that the forces are calculated from the derivatives of the potential energy terms.

There are two main ways in which molecules are represented in atomic-level force fields: all-atom and united-atom. In all-atom force fields, interaction terms and parameters and equations are defined for every atom in the system. United-atom force fields only explicitly consider polar hydrogen atoms, thus reducing the number of particles for which energies and

forces must be calculated, particularly for hydrogen-heavy biological and organic molecules. Additionally, coarse-grained force fields can be used to model large systems over long time scales. These use a more abstract representation by grouping several atoms or even several molecules into larger beads (173).

The total potential energy $U(r)$ is calculated as a sum of bonded terms, for the interaction of atoms connected by covalent bonds (bond lengths, bond angles and dihedral angles), and non-bonded terms, which define the electrostatic (Coulombic) and van der Waals (Lennard-Jones) forces. Simply put, a force field applies rules to atoms to describe their interactions. The forms of the mathematical functions are largely similar between different force fields, reflecting the compromise between representing the underlying chemistry while retaining simplicity and transferability. Their parameters have been optimised over the years by making comparisons between MD simulations and experimental results.

1.4.6.1. Bonded terms

The bonded terms contributing to the potential energy of a system are given in Equation 1.10 and Figure 1. 29. These terms are most important for determining the geometry of a molecule.

$$U_{bonded} = U_b(r) + U_a(\theta) + U_d(\phi) + U_{id}(\phi)$$

Equation 1.10

Most atomic-level force fields provide four bonded terms: $U_b(r)$ represents the potential energy of stretching the bond between two atoms, $U_a(\theta)$ gives the potential energy of an angle among three atoms, $U_d(\phi)$ gives the potential energy of proper dihedrals (bond rotation), and $U_{id}(\phi)$ is the potential energy function of so-called ‘improper’ dihedrals (172). Each term is described in more detail below.

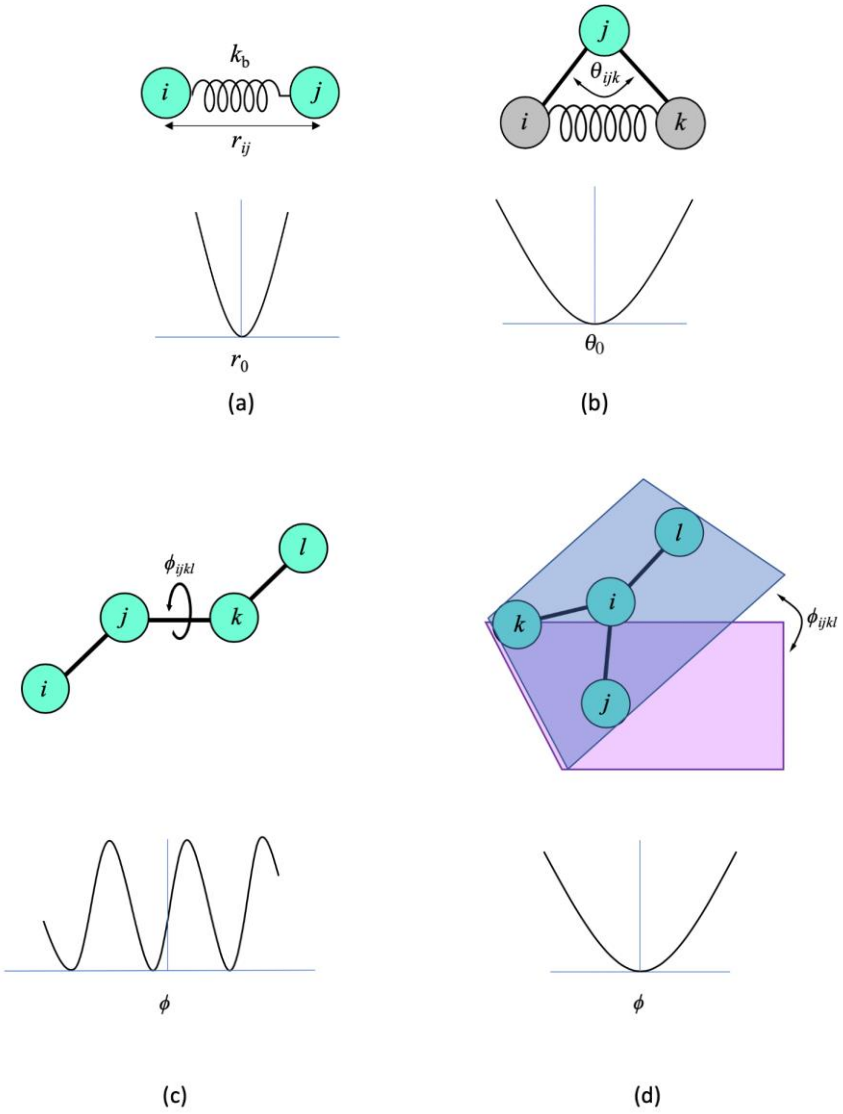


Figure 1. 29. Diagrammatic representation of force field terms. (a) displays the bond stretch term for the distance between two bonded atoms, (b) shows the angle bend term defined as the angle θ between two atoms bonded to a central atom, (c) shows the dihedral angles, and (d) the improper dihedral angles, and their potential energies. The dihedral term defined as the angle Φ between two bonds through rotation about a central bond.

Two-body bond stretching potential

The bond stretching potential energy function typically has the functional form of a harmonic potential (Figure 1. 29 (a)), and is written as:

$$U_b(r_{ij}) = \frac{1}{2} k_b (r_{ij} - r_0)^2$$

Equation 1.11

In Equation 1.11, r_{ij} is the current length of the bond between atoms i and j at time t, k_b is a force (spring) constant between the two bonded atoms and r_0 is the ideal bond length

(equilibrium bond length), both of which differ depending on the nature and chemical situation of atoms *i* and *j* (172, 174).

Three-body angle potential

The angle stretching potential maintains an ideal or equilibrium angle between three atoms, *i*, *j* and *k* (Figure 1. 29 (b)) (172):

$$U_a(\theta_{ijk}) = \frac{1}{2} k_a (\theta_{ijk} - \theta_0)^2$$

Equation 1.12

In Equation 1.12, θ_{ijk} is the current angle, k_a is the harmonic force (spring) constant and θ_0 is the ideal angle between the three linked atoms, with the latter two parameters depending on the nature and chemical situation of the atoms *i*, *j* and *k*.

Four-body proper dihedral angle potential

The torsion angle potential energy function (Equation 1.13), which depends on four consecutive bonded atoms, *i*, *j*, *k* and *l*, is used to limit the rotation around a bond (Figure 1. 29 (c)). It governs the angle between the planes defined by two sets of three atoms that have two atoms in common according to:

$$U_d(\phi_{ijkl}) = \sum \frac{1}{2} k_\phi [1 + \cos(n\phi_{ijkl} - \delta)]$$

Equation 1.13

In Equation 1.13, k_ϕ is a force constant, which is proportional to the barrier to rotation, *n* represents the number of minima and δ is the angle value by which the potential energy function is shifted (172, 175). Multiple dihedral angle terms can be combined for a single set of four atoms to provide complex rotational potential energy functions with multiple minima and maxima of different heights.

Four-body improper dihedral angle potential

Improper dihedral angle terms are used to maintain the chirality of a tetrahedral centre in united-atom force fields where one or more of the hydrogen atoms are omitted, or the planarity of a group of atoms. The improper term is based on a quartet of bonded atoms, *i*, *j*, *k* and *l*, which may be three atoms centred on a fourth atom, see Figure 1. 29 (d), or, in the case of

maintaining planarity, could be four sequential atoms like for the dihedral angle term. The improper potential is given by:

$$U_{id}(\phi_{ijkl}) = \frac{1}{2} k_{\phi} (\phi_{ijkl} - \phi_0)^2 \quad ,$$

Equation 1.14

where ϕ_{ijkl} is the current value of the dihedral angle, ϕ_0 is the idealized equilibrium dihedral angle value and k_{ϕ} is the harmonic force constant for the harmonic potential (172, 175).

1.4.6.2. Non-bonded terms

These terms describe the interactions between particles that are not directly bonded, based on their distance of separation. Non-bonded interactions involve two terms that describe van der Waals interactions and electrostatic interactions, respectively.

Van der Waals (Lennard-Jones) interactions

The van der Waals interactions are commonly described by the Lennard-Jones potential (Equation 1.15). This potential function is used because it is fast to compute, as the $1/r_{ij}$ terms can be computed directly from the square of the inter-atomic distance without taking (expensive) square roots, and provides a sufficiently good representation of the true chemical interactions.

$$U(r_{ij}) = 4\varepsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

Equation 1.15

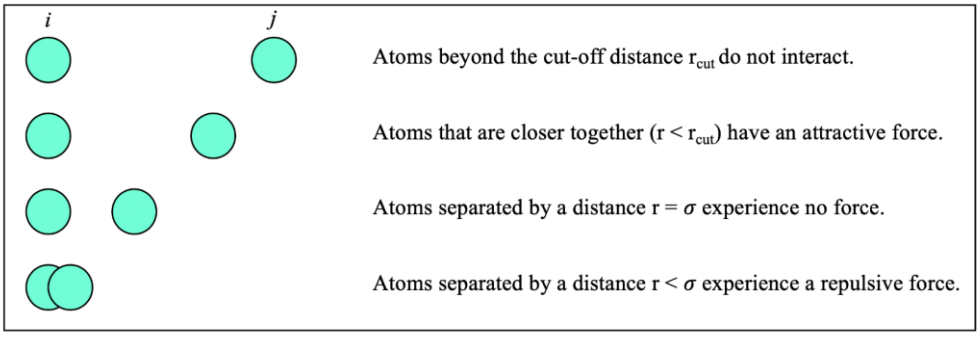


Figure 1. 30. Van der Waals interactions between two atoms as a function of their separation distance *r*.

In Equation 1.15, $U(r_{ij})$ is the intermolecular potential between two atoms i and j, r_{ij} is the distance between the atoms, ϵ_{ij} is the (favourable) potential energy at the ideal interatomic distance between the two atoms, and σ_{ij}, also known as the van der Waals radius, is the interatomic distance at which the potential energy function crosses from being negative (favourable) to positive (unfavourable), and thus determines how close two particles can get (Figure 1. 31). The term $1/r_{ij}^{12}$ accounts for the short-range repulsions, while $1/r_{ij}^{6}$ represents the attractive forces between two particles (172). The Lennard-Jones potential goes to zero at infinite separation, however the contributions are near zero even at relatively short distances. Due to this, as noted above, to save computing time, the Lennard-Jones potential is often truncated at a cut-off distance (Figure 1. 31), as the Lennard-Jones contributions at distances greater than this cut-off are negligible. For some force fields, the potential is 'shifted' or 'switched' upwards as it approaches the cut-off distances so that its value is zero at the cut-off.

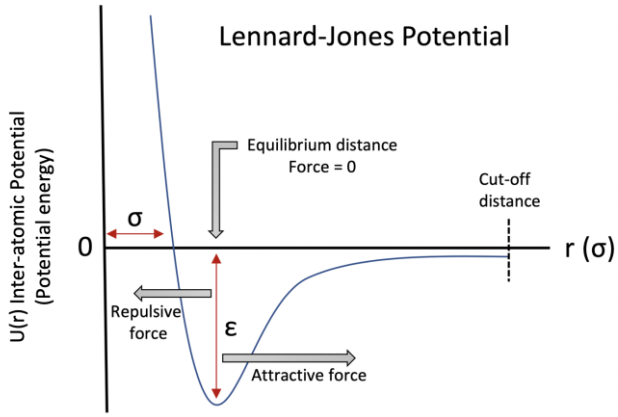


Figure 1. 31. Lennard-Jones potential. The deeper the well depth (ε) is, the stronger the attractive interaction. U(r) is zero when r is equal to σ.

Electrostatic (Coulombic) interactions

The electrostatic or Coulombic interaction energy between two non-bonded atoms is calculated by using Coulomb's law. This law describes the electrostatic force between two point charges:

$$U = k_e \frac{q_i q_j}{r_{ij}^2} \quad ,$$

Equation 1.16

where k_e is Coulomb's constant (k_e = 8.99×109 N m^2 C^{-2}), q_i and q_j are the partial charges of atoms i and j, respectively, and depend on the atom types, and r_{ij} represents the distance between them. The Coulombic potential energy is positive for interactions between particles of like charge and negative for interactions between oppositely charged particles, and increases in magnitude at small values of r_{ij} (Figure 1. 32).

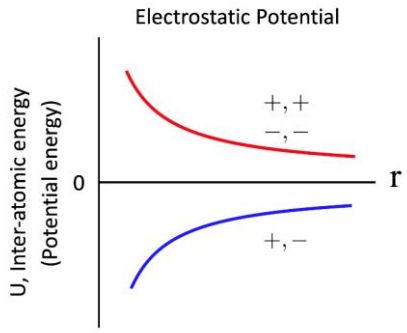


Figure 1. 32. Electrostatic (Coulombic) potential.

Unlike in quantum chemical calculations, in molecular dynamics the smallest explicit particle considered is the atom and therefore charge is not implicit. To overcome this, atoms are assigned point charges, which are usually in the range –1 to +1 and thus referred to as partial charges. The magnitudes of these charges are determined by fitting an electrostatic potential to the molecule and/or by trial and error adjustment to match experimental data describing molecular or solution properties.

As for the van der Waals interactions, the Coulombic interactions are also only calculated out to a certain inter-atomic distance, as calculation of the distance between pairs of atoms is one of the most expensive parts of an MD simulation. Beyond the cut-off distance, electrostatic effects are approximated by the scheme used to evaluate the long-range electrostatic interactions. These schemes include reaction-field (176, 177) , Ewald summation (178), and Particle-mesh Ewald summation (179).

1.4.7. Restraints

Sometimes, special potentials are used for applying restraints on the motion of the system. These restraints are not strictly part of the force field, but are included for reasons such as including knowledge from experimental data.

1.4.6.1. Position restraints

Position restraints are used to prevent rearrangements of critical and important parts of the system during the MD simulation. Selected particles of the system are restrained to fixed positions $\mathbf{R}_i$ according to (Figure 5. 5 32):

$$V_{pr}(\mathbf{r}_i) = \frac{1}{2} k_{pr} |\mathbf{r}_i - \mathbf{R}_i|^2 \quad , \qquad \textbf{Equation 1.17}$$

where $\mathbf{r}_i$ are the current coordinates of each of the i selected particles and k_{pr} is the harmonic restraint force constant that determines the strength of the restraints. Other functional forms are also possible (180). The position restraints can be switched on or off in each spatial dimension, which means atoms can be harmonically restrained to a line, a plane, or a point in three-dimensional space (181).

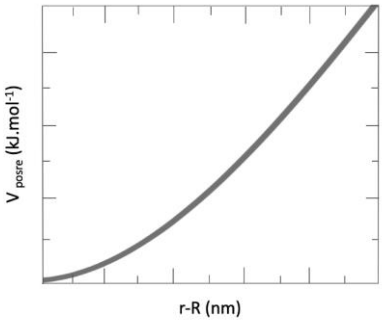


Figure 1. 32. Position restraint potential.

1.4.6.2. Distance restraints

Distance restraints are used to keep the distance between a pair or two groups of atoms below or above a specified distance. A standard approach is to use a flat-bottomed harmonic potential (180) (Figure 1. 33) V_{dr} that becomes linear at large distances (Equation 1.18. This means that the restraint potential is zero between distances of r_0 and r_1, is harmonic for distances below r_0 and above r_1, but becomes linear (resulting in a constant attractive force (181), Equation 1.19) for distances greater than r_2, to prevent extremely high forces at large distances.

$$V_{dr}(r_{ij}) = \begin{cases} \frac{1}{2}k_{dr}(r_{ij}-r_0)^2 & \text{for } r_{ij} < r_0 \\ 0 & \text{for } r_0 \le r_{ij} < r_1 \\ \frac{1}{2}k_{dr}(r_{ij}-r_1)^2 & \text{for } r_1 \le r_{ij} < r_2 \\ \frac{1}{2}k_{dr}(r_2-r_1)(2r_{ij}-r_2-r_1) & \text{for } r_2 \le r_{ij} \end{cases}$$

Equation 1.18

$$\boldsymbol{F}_i = \begin{cases} -k_{dr}(r_{ij} - r_0)\dfrac{\boldsymbol{r}_{ij}}{r_{ij}} & \text{for} \quad r_{ij} < r_0 \\ 0 & \text{for} \quad r_0 \leq r_{ij} < r_1 \\ -k_{dr}(r_{ij} - r_1)\dfrac{\boldsymbol{r}_{ij}}{r_{ij}} & \text{for} \quad r_1 \leq r_{ij} < r_2 \\ -k_{dr}(r_2 - r_1)\dfrac{\boldsymbol{r}_{ij}}{r_{ij}} & \text{for} \quad r_2 \leq r_{ij} \end{cases}$$

Equation 1.19

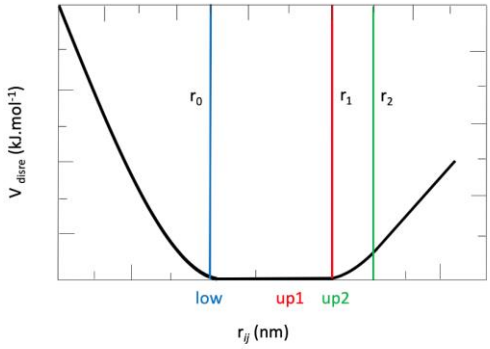


Figure 1. 33. Distance restraint potential. r_0, r_1 and r_2, respectively in equations 2 and 3 are indicated by coloured lines.

1.5. Free energy

The free energy is the energy in a system that can be converted to do work. It is considered as the most important thermodynamic quantity. Knowing the free energy change associated with a process allows its likelihood to be predicted, thus free energy calculations have a variety of uses in chemistry and biochemistry (182). Free energy calculations can determine the partitioning of a solute between immiscible liquids (partition coefficient calculations), solvation/hydration free energies, the stability of proteins, and receptor-drug, protein-protein and protein-DNA binding affinities (182). Binding and solvation free energies are connected; the former is often required for an accurate calculation of the latter. The solvation free energy of a small molecule provides information about its desolvation in the thermodynamic process of host-guest binding (183).

The Gibbs free energy (Equation 1.20), or free enthalpy, is the thermodynamic free energy at a constant temperature and pressure, and is therefore the relevant free energy for a simulation conducted under isothermal-isobaric conditions (NPT ensemble).

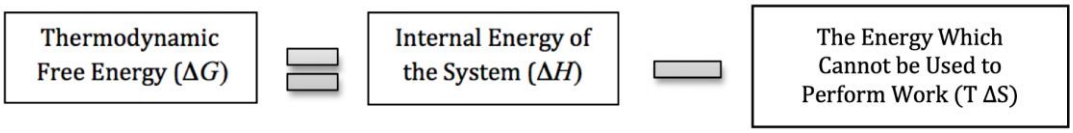
$$G(N,P,T) = U - TS + PV$$ **Equation 1.20**

$$U + PV = H$$ **Equation 1.21**

U is the internal energy (defined by Equations 1.18), T is the absolute temperature, S is the entropy, P is the pressure, and V is the volume, and in Equations 1.21, H is the enthalpy of the system. By combining Equations 1.20 and 1.21, we obtain $G = H - TS$, which describes the relative free energy difference, ΔG , between two equilibrium states (Equation 1.22).

$$\Delta G = \Delta H - T\Delta S$$ **Equation 1.22**

The Gibbs free energy change (ΔG) gives information about the energetics and spontaneity of the process. A large magnitude, negative value of ΔG indicates that the process is favourable and spontaneous (184), and a large positive value indicates an unfavourable process (Figure 1.).



$\Delta G = \Delta H - T\,\Delta S$	$\Delta H<0$	$\Delta H>0$
$\Delta S > 0$	Spontaneous $\Delta G < 0$	Spontaneous at high T
$\Delta S < 0$	Spontaneous at low T	Non-spontaneous $\Delta G > 0$

Figure 1. 33. Summary of the contributions to the Gibbs free energy ΔG (top) and categorisation of the spontaneity of a process according to the components of ΔG.

Experimentally, in a binding reaction A + B → AB, the binding constant is defined as K_b (equilibrium constant), which is calculated as

$$K_b = [AB]/([A][B]) \quad ,$$ **Equation 1.23**

where [AB], [A] and [B] are the equilibrium concentrations of bound complex, unbound host and unbound guest, respectively. The standard binding free energy and then be calculated as

$$\Delta G_b^\circ = -K_B T ln(C^\circ K_b) \ ,$$

Equation 1.24

where K_B is the Boltzmann constant, T is the absolute temperature and C° is a standard concentration (185).

1.5.1 Free energy calculation methods

There are many different methods for calculating free energies using MD simulations, most of which are computationally intensive and many of which are highly complex; here, just the one used in this study, thermodynamic integration, is described. This method takes advantage of the fact that free energy is a state function (186) and is independent of the reaction or process path. Therefore, it can be determined by integrating the potential energy between the starting and ending molecular states (187). The process path does not need to be a real chemical process, but can also follow an alchemical path. For instance, the actual process of the binding of one molecule to another is a chemical path, but this is very expensive to follow computationally, as many binding/unbinding transitions would need to be followed to obtain a reliable estimate of its probability of occurrence, i.e. ΔG_{bind}. On the other hand, the transformation of one molecule into another, or into an invisible 'dummy' molecule, is an alchemical transformation, which is impossible in real life but both possible and computationally less expensive to carry out in a simulation.

1.5.1.1. Thermodynamic integration

Thermodynamic integration (TI) is an example of an alchemical method. TI is based on the concept of a thermodynamic cycle, in which the sum of the free energy changes around the complete cycle is always zero (188). An example of a thermodynamic cycle for the case of calculating the solvation free energy is given in Figure 1. . Here, the chemical process for which we would like to calculate the free energy change is hydration, and we use TI to calculate two alchemical processes, $\Delta G_{\text{decoupling,vac}}$ and $\Delta G_{\text{decoupling,solv}}$, instead. In practise, the solute molecule transitions to a 'dummy' molecule in which the atoms have coordinates, masses, and bonded interactions, but no van der Waals or Coulombic interactions with the remainder of the system, so that the molecule is effectively invisible (189).

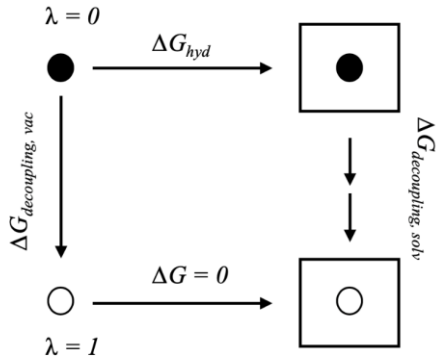


Figure 1. 34. Example of a thermodynamic cycle for the calculation of hydration free energy ΔG_{hyd}. In order to calculate the desired property, ΔG_{hyd}, the ΔG of "disappearing" or decoupling the solute in vacuum ($\Delta G_{decoupling,vac}$) and in solvent ($\Delta G_{decoupling,solv}$) are calculated. As the free energy of solvating a non-existent particle is equal to zero (bottom arrow), ΔG_{hyd} can be calculated by requiring the sum of ΔG around the thermodynamic cycle to sum to zero.

The progress along the alchemical path between two states of the cycle is controlled using a scaling variable λ. A series of simulations are run with values of λ between 0 and 1, where λ = 0 corresponds to the true molecule and λ = 1 to the dummy molecule. Δ*G*, the free energy difference between the two states of a system with λ = 0 and λ = 1, is calculated by integrating the ensemble average of the derivative of the potential energy with respect to λ along this path (190):

$$\frac{\partial G(\lambda)}{\partial \lambda} = \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_{\lambda}$$

Equation 1.25

$$\Delta G = \int_0^1 \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle d\lambda$$

Equation 1.26

In practice, the integral is evaluated numerically over a number of intermediate states, corresponding to simulations run at different λ values, using a method such as the Trapezoidal rule. This method is an algorithm for determining the difference in free energy between two states of a system (see Equation 1.26).

Binding free energy using thermodynamic integration

To calculate the binding free energy of two molecules, such as a guest to a host molecule, we need to compare the situation of the small guest being bound to the host to the situation where

both are separate in solution. The thermodynamic cycle for host-guest binding free energy calculations is illustrated in Figure 1. for a host-guest system where the host is a cyclodextrin. We assume that the free energy difference between the state with the guest and the host present, but not bound, and the state with only the host present, can be separated into the two parts, host → host, with $\Delta G = 0$, and the solvation free energy of the guest, ΔG_1. Therefore, by calculating ΔG_3, the free energy change for moving from the guest bound to the host to a dummy (i.e. invisible) guest bound to the host, and ΔG_1, the free energy change for moving from the guest in solution to a dummy (i.e. invisible) molecule in solution, ΔG_{bind} can be calculated as:

$$\Delta G_{\text{bind}} - \Delta G_3 + \Delta G_2 + \Delta G_1 = 0$$ **Equation 1.27**

$$\Delta G_{\text{bind}} = \Delta G_3 - \Delta G_1$$ **Equation 1.28**

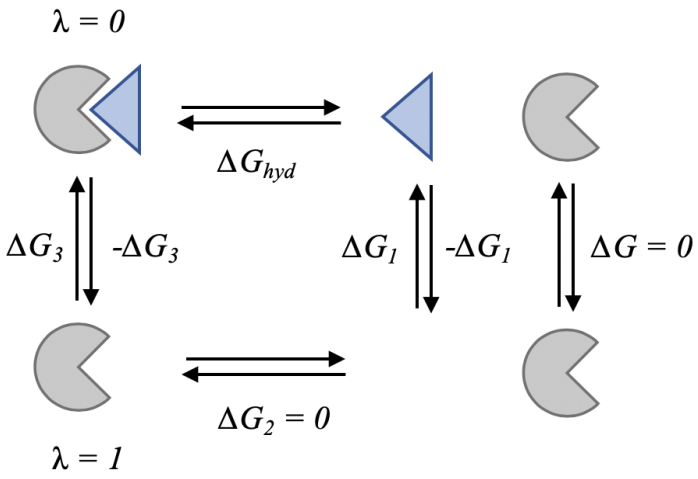


Figure 1. 35. Thermodynamic cycle used to calculate the binding free energy of a host-guest system via the thermodynamic integration technique. The host molecule is in grey and the guest is a blue triangle. Note that solvent is present in all four states, but not shown for clarity.

1.6. Aims and objectives

In this thesis, binding modes and intermolecular interactions between different molecules at a range of scales were studied using MD simulations. The systems that were investigated were complexes of the shapeshifter molecule bullvalene with cyclodextrin (Chapter 2), the transcriptional repressor KstR and its interactions with its ligand and DNA (Chapter 3), another

related transcriptional repressor KstR2 interacting with its ligand (Chapter 4), and the class IA phosphatidylinositide 3-kinase PI3Kα interacting with RAS and with a model cell membrane (Chapter 5).

In Chapter 2, the goal was to capture one isomer of the fluctional molecule bullvalene (guest molecule) by complexation with cyclodextrin (host molecule). Chapter 3 involved MD simulations of the apo, ligand-bound and DNA-bound states of KstR, an Mtb transcriptional repressor, in order to find whether DNA binding occurs via an induced fit or conformational selection mechanism, and to investigate possible allosteric communication pathways by which information about ligand and DNA binding are transmitted through the protein. Chapter 4 focused on KstR2, another Mtb transcriptional repressor that is structurally similar to KstR but has been proposed to act through a novel scissor-like mechanism. This chapter investigated two key questions regarding the mechanism of action of KstR2: first, the effect of mutating the key switch residue ARG170 to ALA, and second, the effect of ligand binding on its conformation and motion. The goal of Chapter 5 was how RAS affects the interaction of PI3Kα with the cell membrane. The majority of the work in this chapter involved painstaking building of a PI3Kα-RAS-membrane complex. Several alternative approaches had to be taken due to a number of unexpected problems that were encountered. The methods and results are discussed in each chapter in detail.

Chapter 2. Bullvalene-Cyclodextrin Complexation

The overarching goal of this chapter is to predict the binding mode and affinity of bullvalenes to cyclodextrins. The selection of the best solvent to use and of the best parameters and initial coordinates for bullvalenes and cyclodextrins are described in Appendix A. The first section of the chapter (2i), describes the testing and validation of the host-guest binding potential energy profiling (HGBPEP) method that was developed to predict optimal host-guest binding modes. The HGBPEP method was then used to predict the binding modes of bullvalene isomers to β- and to γCD, for which the results are described in detail in 2ii.

2i. Host-guest binding potential energy profiling (HGBPEP) method validation

2i.1. Energetic profiling of host-guest binding positions

In order to simulate host-guest binding and compute the binding free energy of the complex, it is first necessary to find the best position, including orientation, for binding of bullvalene to cyclodextrin from which to start the simulations. For this purpose, a "host-guest binding potential energy profiling" (HGBPEP) protocol was constructed (Figure 2i. 1).

2i.1.1. Overview of host-guest binding potential energy profiling procedure

The HGBPEP process performs a single-point energy calculation of the system (i.e. after one MD integration step) using the parameter sets chosen in Section 2.2.1.2 for each possible position and orientation of a guest molecule in a host cyclodextrin. The guest molecule is initially located at the centre of geometry of cyclodextrin (0) (Figure 2i. 2). It is rotated 360° around an axis parallel to the *z*-axis that projects through the centre of the host in 1° increments, then moved gradually up and down this axis in 0.01 nm increments, rotating through 360° after each translational step. The non-bonded (van der Waals and electrostatic) interaction energies between the βCD and the guest molecule were calculated for each orientational and translational position. The energies are plotted as 2D or 3D heat maps with respect to translational distance and orientation (Figure 2i. 3), allowing the most favourable binding position and orientation to be selected and utilised as the initial conditions for binding free energy calculations.

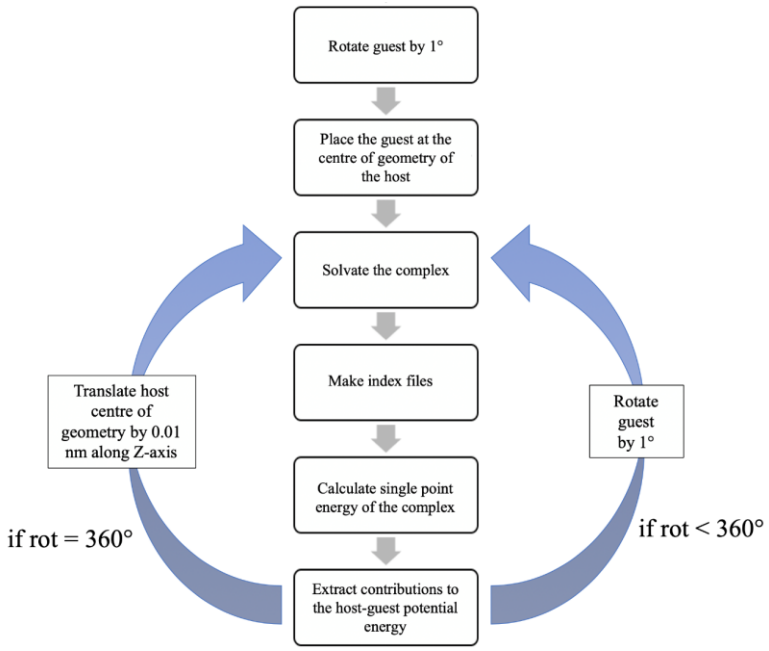


Figure 2i. 1. Diagram outlining the host-guest binding potential energy profiling (HGBPEP) process.

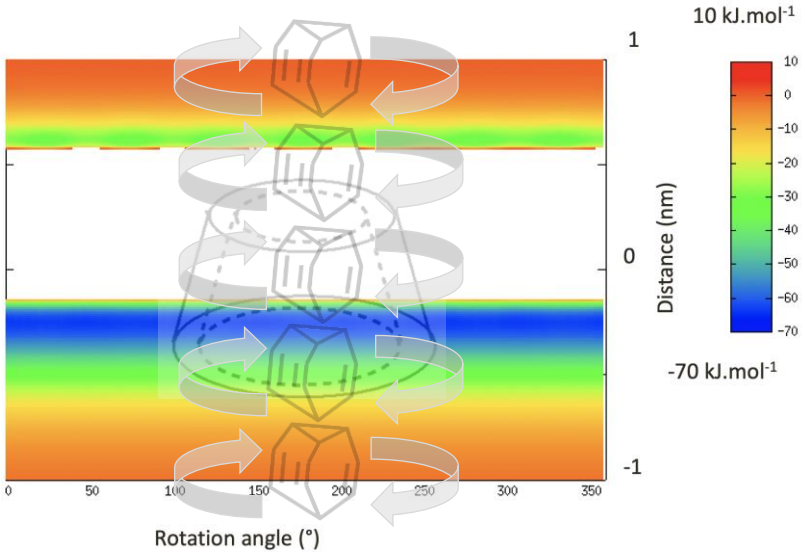


Figure 2i. 2. Example of HGBPEP 2D heat map. On the *x*-axis is the degree of rotation of bullvalene around the *z*-axis of the simulation box, and on the *y*-axis is the distance of bullvalene along the *z*-axis of the simulation box from the centre of geometry of cyclodextrin. The colour bar represents the non-bonded potential energy values. Dark blue and red colours indicate the most and least energetically favourable positions and orientations of bullvalene in complex to cyclodextrin, respectively. The white zone represents very large and positive potential energies, which are highly unfavourable and beyond the colour scale.

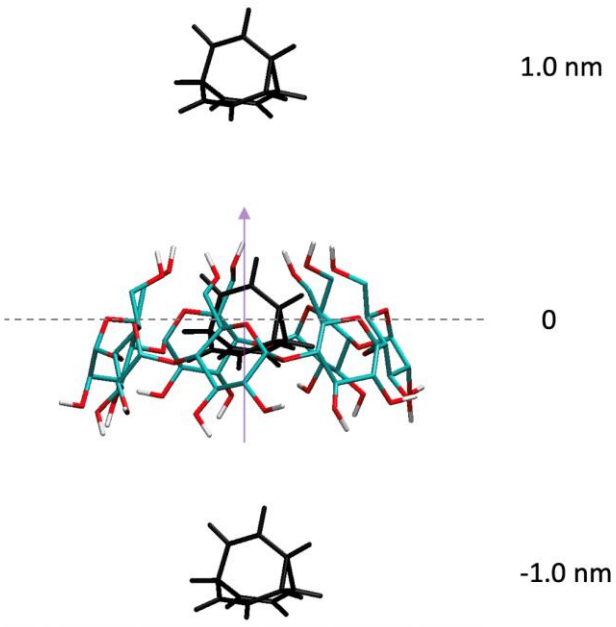


Figure 2i. 3. Position of guest molecule (here unsubstituted bullvalene) to the host (here βCD) at 0, 1.0 and -1.0 nm from the centre of geometry of cyclodextrin along the *z*-axis.

2i.1.2. Validation of HGBPEP

In order to validate whether the HGBPEP method is capable of predicting the optimal binding mode of a guest molecule to a host cyclodextrin molecule, and to build up an understanding of how different chemical groups might bind to βCD, tests were run using a series of molecules where the optimal binding mode is either easily predicted without running calculations or has been studied experimentally. The guest molecules, methanol, benzene, and para-chlorophenol (pCP), represent different chemical natures and sizes. pCP was chosen because there is previous experimental work on complexation of cyclodextrin and chlorophenol (191); methanol and benzene then represent the components of pCP. Their energy profiles help to explain how each feature of pCP contributes to its interactions with the βCD. The solvent was water in all cases.

2i.1.2.1. Methanol and βCD

To study the effect of the OH group of pCP on its interactions with βCD, HGBPEP was done for methanol as a guest molecule binding to βCD. The methanol was positioned with the OH group oriented either downwards (CH3OH-down) or upwards (CH3OH-up) along the central axis (Figure 2i. 4).

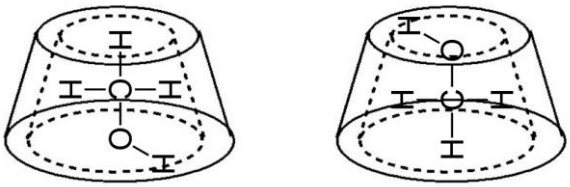


Figure 2i. 4. Schematic showing the complexes of methanol and βCD used in the HGBPEP calculations, showing the (left) CH3OH-down and (right) CH3OH-up orientations of methanol.

The most favourable complex of CH3OH-down forms when it lies inside the βCD cavity close to z = 0 nm (Figure 2i. 5, top). Both the Lennard-Jones and Coulombic interaction energy values gradually increase as methanol is moved out of the βCD in either direction. The most and least favourable Coulombic and Lennard-Jones interactions are found at the same distances and rotations in this complex. Slightly closer to the smaller rim, at z = 0.5 nm, the hydroxyl hydrogen of the CH3OH-down feels weakly unfavourable van der Waals interactions due to being too close to seven hydrogen atoms in the H5-plane in βCD. On the other hand, when the hydroxyl hydrogen is close to the OH groups of the bigger rim of βCD, the Coulombic potential energies were more favourable.

For CH3OH-up, the Lennard-Jones profile is similar to that of CH3OH-down, but has a more dominant favourable binding region around z = 0 nm (Figure 2i. 5, bottom). However, the Coulombic interaction energy profile of CH3OH-up is essentially the opposite to that of CH3OH-down. It shows only very small favourable energies when the CH3OH-up is outside the βCD, and large positive energies when it is inside the βCD.

Overall, these results show that the more energetically favourable case is when the OH group points downward, which is in good agreement with pCP results (see Section 2i.1.2.3) and the experimental work (191).

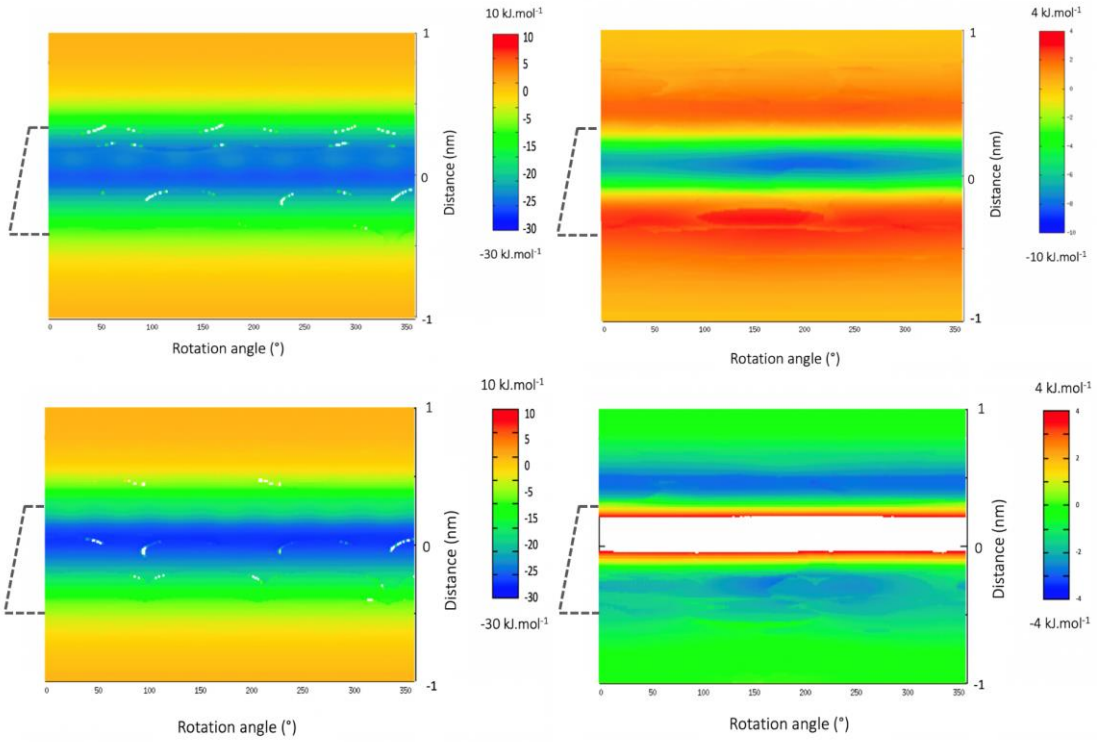


Figure 2i. 5. HGBPE profiles for complexes of (top) CH3OH-down and (bottom) CH3OH-up with βCD. The colour scale represents the van der Waals (left) and Coulombic (right) potential energies (kJ·mol^{-1}). The degree of rotation around the *z*-axis of the simulation box is on the horizontal axis (0-360°), and the translation distance along the *z*-axis is on the vertical axis (nm). The dashed lines show the position of cyclodextrin.

2i.1.2.2. Benzene and βCD

To study the effect of the overall size and shape of pCP on its interactions with βCD, HGBPEP was done for benzene as a guest molecule binding to βCD. The benzene was positioned with the plane of the ring parallel to the *z*-axis (Figure 2i. 6).

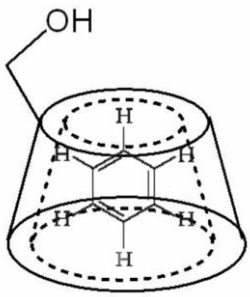


Figure 2i. 6. Schematic of the complex of benzene and βCD used in the HGBPEP calculations.

As Figure 2i. 7 (right) shows, the highest Coulombic energy values were obtained when the two ortho and para hydrogen atoms of benzene become close to the hydrogen atoms of the upper OH groups in βCD (E ≅ 3 kJ·mol^{-1}). The most favourable and lowest energy values were measured when benzene was located at the centre of the βCD cavity, slightly closer to the wider rim (E ≅ −6 kJ·mol^{-1}). In terms of the Lennard-Jones energies, there is a higher energy region when benzene is located close to the smaller rim of βCD (–0.1 to 0.0 nm along the *z*-axis) As the hydrogen atoms of the upper OH groups of the βCD become close to the lateral hydrogen atoms of βCD, around z = −0.1 to 0.4 nm, the van der Waals interactions tend to be less favourable, and when benzene lies within the narrower rim of βCD, the highest Lennard-Jones potential energies are recorded. These results show that the physical dimensions of a benzene ring, and therefore also of pCP, are too large to allow movement through the narrow rim of βCD without steric clashes occurring, thus binding of pCP to βCD will only occur if it approaches the central cavity through the wider rim.

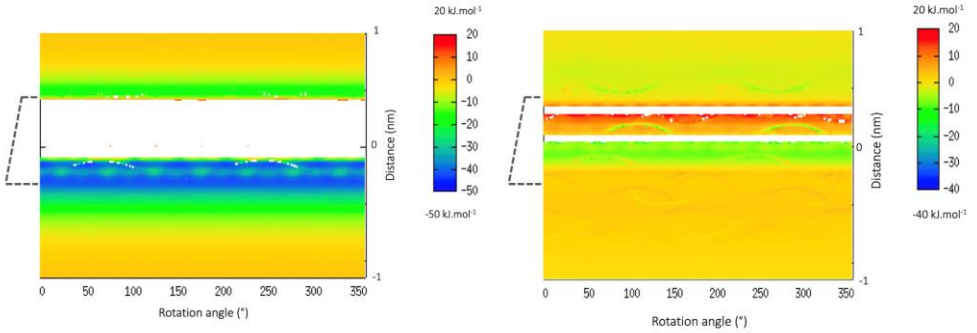


Figure 2i. 7. HGBPE profiles for complexes of benzene with βCD. The colour scale represents the van der Waals (left) and Coulombic (right) potential energies (kJ·mol^{-1}). The degree of rotation around the *z*-axis of the simulation box is on the horizontal axis (0-360°), and the translation distance along the *z*-axis is on the vertical axis (nm). Dashed lines show the position of cyclodextrin.

2i.1.2.3. Para-chlorophenol and βCD

Having understood the binding of methanol and benzene to βCD, HGBPEP was carried out for pCP. Previous experimental work done by Leyva *et. al.* (191) on complexation of cyclodextrin and pCP was ideal for validation of the HGBPEP results. Their UV-VIS and ^{1}H NMR results show that the stability of the complexes depends on host and guest structures, like the size of the guest molecule and the cavity of the host. It is also dependent on the polarity and other properties of the molecules. However, geometric complementarity is the most important issue on which stability of the complex is dependent. According to Leyva *et. al.*'s results, changes in chemical shifts of the H3 and H5 hydrogen atoms (Figure 2i. 8, right) inside the cavity show that the best complex forms between the smaller and larger rims of βCD (Figure 2i. 8, left).

There are cross peaks visible in their ^{1}H NMR spectrum that represent interactions between the H5 and H3 hydrogen atoms of βCD and the ortho- and meta-hydrogen atoms on pCP.

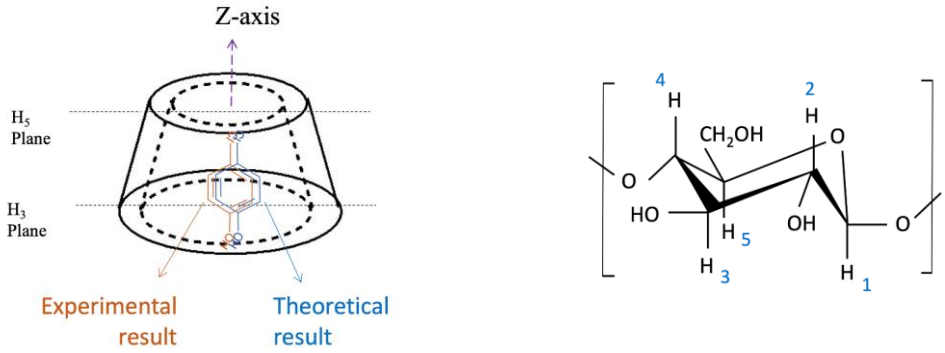


Figure 2i. 8. (Left) Schematic of the para-chlorophenol-cyclodextrin complex in aqueous solution predicted by Leyva et al. (191) vs the most favourable complex predicted using the host-guest binding potential energy profiling method. (Right) Hydrogen numbering according to (191) for each glucose molecule in cyclodextrin.

In the first instance, HGBPEP was carried out for pCP and βCD with the chloro group pointing “up”, i.e. towards the smaller opening of the βCD (pCP-up), and pointing “down", i.e. towards the larger opening (pCP-down). The Lennard-Jones and Coulombic energies are shown in Figure 2i. 9.

When comparing the magnitude of the total interaction potential energies for pCP-up and pCP-down, it can be seen that the former is more energetically favourable (minimum interaction potential energy of −100 kJ·mol^{-1}, the dark blue region in Figure 2i. 9, compared to the latter with −60 kJ·mol^{-1}). This matches Leyva *et. al*’s experimental work (191), which showed that the most favourable binding was achieved when the OH group of pCP is oriented downward and closer to the wider rim of βCD.

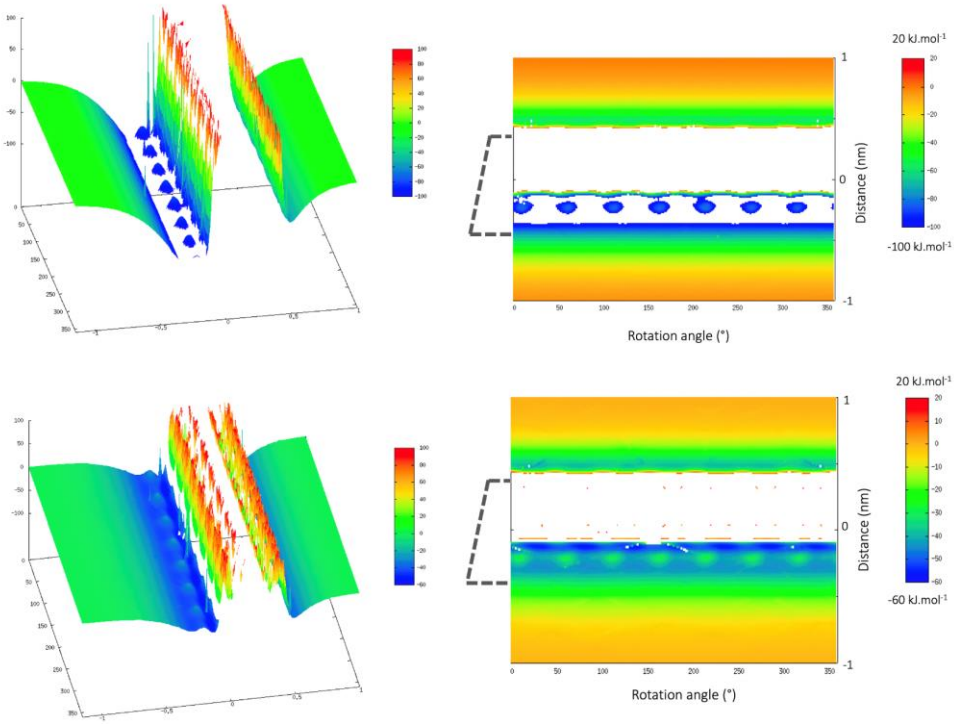


Figure 2i. 9. 3D (left) and 2D (right) representations of the HGBPE profile for complexes of (top) pCP-up and (bottom) pCP-down with βCD. The colour scale represents the total potential energy (kJ·mol^{-1}). The degree of rotation around the *z*-axis of the simulation box is on the horizontal axis (0-360°), and the translation distance along the *z*-axis is on the vertical axis (nm). Dashed lines show the position of cyclodextrin.

It would be energetically unfavourable for the pCP to enter βCD through the smaller, topmost opening, in either orientation, as shown by the discontinuity due to very high energy region between $z = -0.05$ and $z = +0.45$ that overlaps with the top-most opening. The size of the high-energy region in the *z*-direction indicates that pCP cannot enter the upper region of the βCD cavity at all. Instead, it binds in the lower region of the cavity, and favours specific rotation angles, indicated by the series of most favourable binding positions at about $z = -0.2$ to -0.25 that repeat periodically, once per sugar molecule (seven times).

To determine which interactions dominate the HGBPE, it was separated into van der Waals and Coulombic components (Figure 2i. 10). This showed that only for pCP-down were there binding modes with favourable Coulombic interactions, but that it is the van der Waals energies that dominate the total potential energy maps.

Overall, HGBPEP predicts the same energetically favourable location and orientation of pCP in βCD as the experimental results (191) (Figure 2i. 8, left).

The above simulations were conducted with the plane of pCP aligned to the *z*-axis. As this represents only one possible orientation in which pCP might bind, four deviations from strictly axial insertion were investigated. Specifically, pCP-up, which showed the most

favourable binding in the initial study of the axial orientation, was rotated by θ = 30, 45, 60 and 90° from the *z*-axis (Figure 2i. 11), and then subjected to HGBPEP (Figure 2i. 12-14).

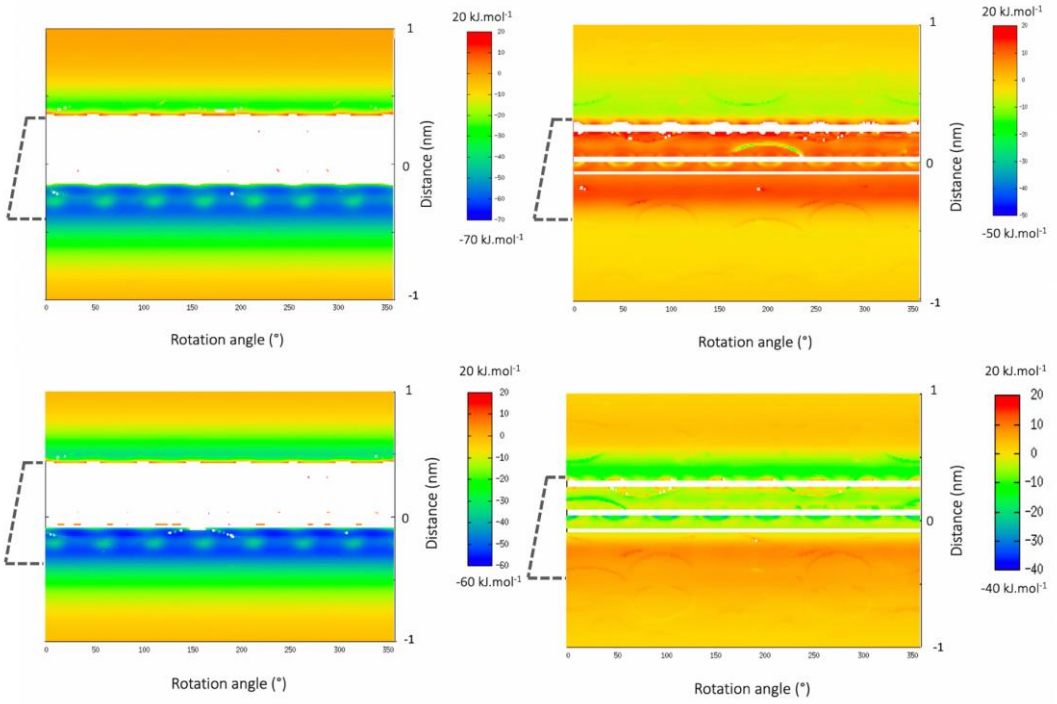


Figure 2i. 11. HGBPE profiles for complexes of (top) pCP-up and (bottom) pCP-down with βCD. The colour scale represents van der Waals (left) and Coulombic (right) potential energies (kJ·mol^{-1}). The degree of rotation around the *z*-axis of the simulation box is on the horizontal axis (0-360°), and the translation distance along the *z*-axis is on the vertical axis (nm). Dashed lines show the position of cyclodextrin.

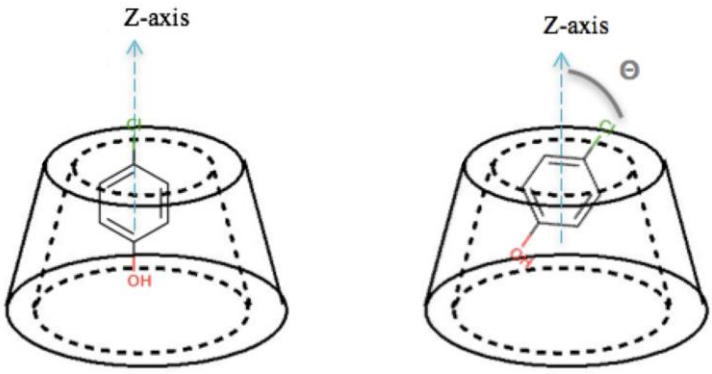


Figure 2i. 11. Schematic of complex of para-chlorophenol and βCD with axial orientation (left) and showing how the angle of deviation from axial insertion, θ, is defined (right).

In the case of θ = 45°**,** there are seven angles of rotation that have very high values of the Coulombic potential energy even at the centre of the βCD cavity (Figure 2i. 12, right), which were not observed for the axial orientation. The reason is that when pCP deviates by 45°

from the axial position, it has larger dimensions in the *x*-*y* plane, so that the chlorine atom becomes too close to the inside of the βCD cavity and has unfavourable interactions with seven hydrogen atoms in the H5-plane of βCD. The only favourable Coulombic interactions are found when the guest is located outside βCD, above the smaller rim.

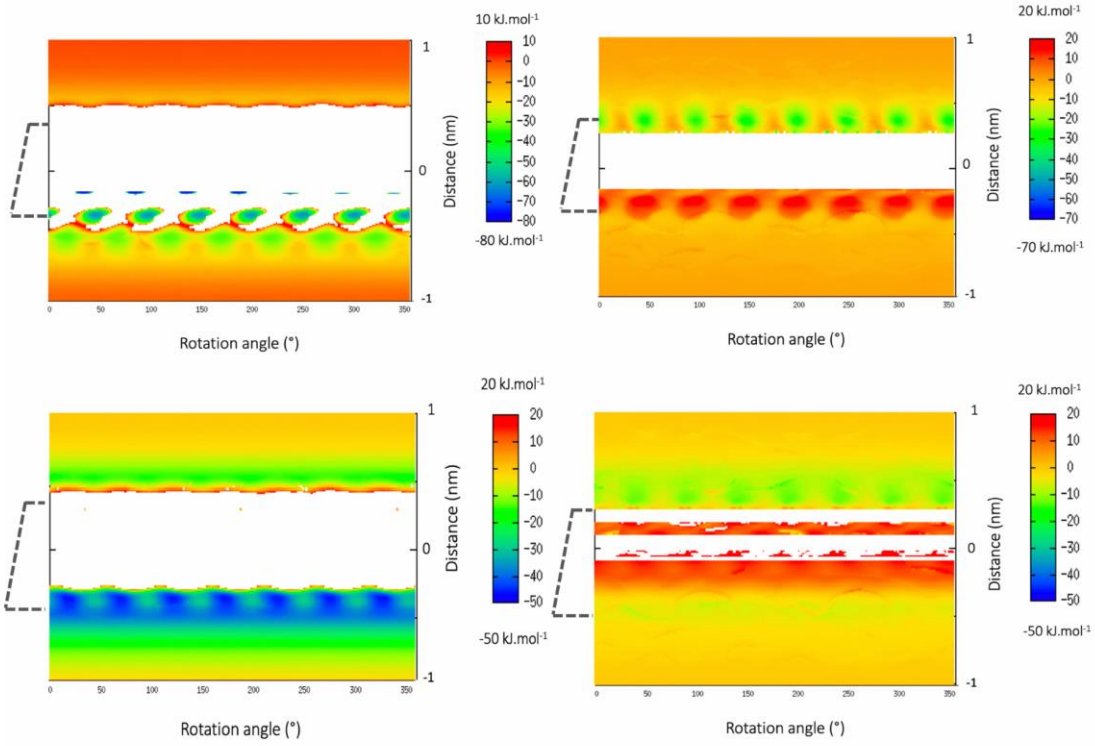


Figure 2i. 12. HGBPE profiles for complexes of (top) pCP-up-θ45° and (bottom) pCP-up-θ30° with βCD. The colour scale represents the van der Waals (left) and Coulombic (right) potential energies (kJ·mol^{-1}). The degree of rotation around the *z*-axis of the simulation box is on the horizontal axis (0-360°), and the translation distance along the *z*-axis is on the vertical axis (nm). Dashed lines show the position of cyclodextrin.

For pCP-up-θ30°**,** the blue line in the Lennard-Jones potential energy profile (Figure 2i. 12, left) shows better van der Waals interactions and also less unfavourable steric repulsion than for pCP-up-θ45°. When pCP-up-θ30° is above the βCD, we can see favourable Lennard-Jones interactions (which were present for pCP-up-θ45° as well). Seven higher-energy (green) spots appear in the low energy region around $z = -0.4$ nm, since the Cl atom becomes close to seven hydrogen atoms in the H5-plane of βCD. Considering Coulombic energies (Figure 2i. 12, right), the middle part of the profile, $z = -0.1$ nm to 0.25 nm, does not show totally unfavourable interactions (as occur for pCP-up-θ45°). The two white regions in the middle of the profile refer to when the hydrogen atoms of the aromatic ring in pCP are close to the OH groups of the smaller rim of βCD.

In the case of pCP-up-θ60°, there are more unfavourable steric repulsions in the centre of the βCD, from z = −0.4 to 0.5 nm, especially close to its smaller rim (Figure 2i. 13, left). The Lennard-Jones profile shows alternating favourable (green) and unfavourable (white/yellow) binding modes when pCP-up-θ60° is located at the bottom of βCD, z = −0.60 nm and the Cl atom interacts with seven H5-plane hydrogen atoms, getting close and being at an ideal distance, alternately. The seven small green curved lines at z = −0.35 nm could also be because of the OH atom of pCP interacting with the H3-plane hydrogen atoms, sometimes being too close (white) and other times at an ideal distance (green). There is a narrow strip of favourable van der Waals interactions (green) between two OH groups (one from pCP-up-θ60° and the other from the smaller rim of the βCD) around z = 0.5 nm. The 7-fold symmetric pattern of favourable Coulombic energies at z = 0.30 nm (green areas in Figure 2i. 13, right) are due to favourable electrostatic interactions between the hydroxyl group of pCP and the hydroxyl group of the narrow rim of βCD.

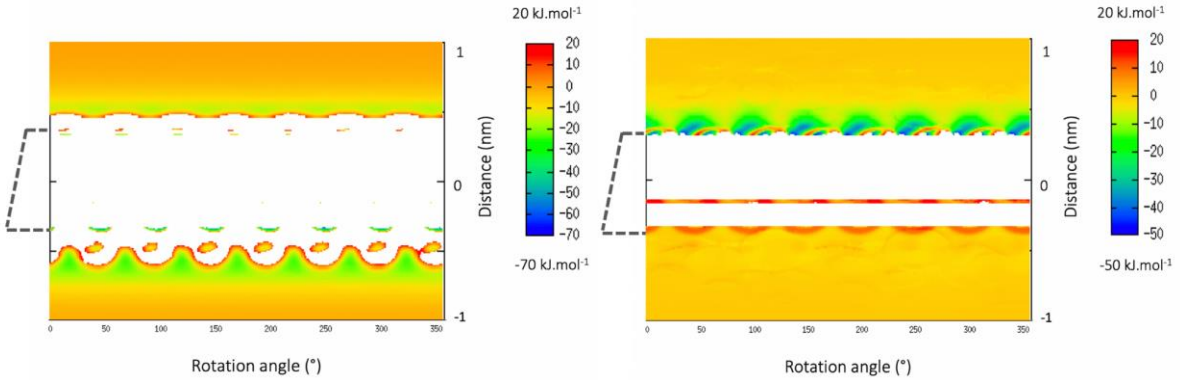


Figure 2i. 13. HGBPE profiles for complexes of pCP-up-θ60° with βCD. The colour scale represents the van der Waals (left) and Coulombic (right) potential energies (kJ·mol^{-1}). The degree of rotation around the *z*-axis of the simulation box is on the horizontal axis (0-360°), and the translation distance along the *z*-axis is on the vertical axis (nm). Dashed lines show the position of cyclodextrin.

When pCP lies in the horizontal plane, pCP-up-θ90°, there is a wide unfavourable region in the Lennard-Jones profile for z = −0.30 to 0.35 nm, which corresponds to positions of pCP between the smaller rim of βCD and the middle of the cage, as expected. At z = −0.32 nm, there is a seven-fold symmetric pattern of spots that are white in the middle, then blue, i.e. highly favourable, then decrease in energetic favourability and pass from green (favourable) through to red (unfavourable) on the outside (Figure 2i. 14, left). These colour changes represent changes in the Lennard-Jones potential energy from, for instance, 20,004.19 kJ·mol^{-1} at 301°, which is highly unfavourable (white centre) to −61.11 kJ·mol^{-1} at 302°, which is highly favourable (blue ring). Because it was necessary to truncate the range of positive

energies shown on the scale bar in order to show the details of how the energies change with rotation and translation, it is not easy to recognize how steep the Lennard-Jones energy profile is here. However, with only one degree of rotation, the Lennard-Jones energy value changes a large amount, suggesting that the preferred binding mode of pCP-up-θ90° is highly orientationally dependent. According to the Coulombic energy plot (Figure 2i. 14, right), at *z* = 0.30 nm, above the small rim of the βCD, there are favourable electrostatic interactions, similar to those observed for pCP-up-θ60°, between the hydroxyl groups on both species.

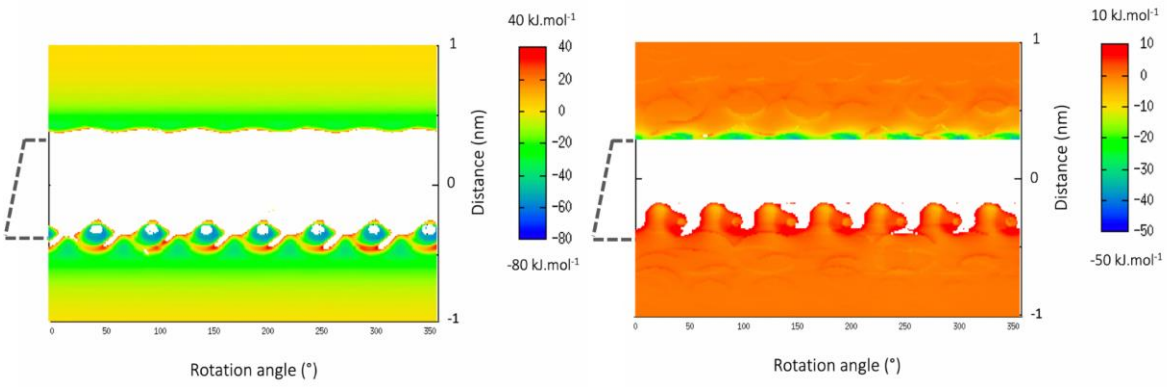


Figure 2i. 14. HGBPE profiles for complexes of pCP-up-θ90° with βCD. The colour scale represents the van der Waals (left) and Coulombic (right) potential energies (kJ·mol^{-1}). The degree of rotation around the *z*-axis of the simulation box is on the horizontal axis (0-360°), and the translation distance along the *z*-axis is on the vertical axis (nm). Dashed lines show the position of cyclodextrin.

Overall, only pCP-up and pCP-up-θ30° would be able to move from bulk solution into the βCD binding pocket, as in the other orientations, there is no continuous path along which a pCP molecule could approach and bind to the βCD, i.e. it would be unlikely to ever find the favourable binding positions. Of these two orientations, pCP-up is the most likely binding orientation due to the lower potential energy of the complex. This is in keeping with the experimental results (191), and thus validate HGBPEP as a method for predicting host-guest binding complexes.

2ii. Results and discussion

2ii.1. Overview

This chapter describes characterisation of the binding modes of host-guest complexes. The chosen acceptor molecule is the commonly-used host molecule cyclodextrin. The guest molecule is bullvalene, a fluctional molecule which undergoes interconverting isomerization (shape-shifting) at room temperature. The goal was to determine whether a particular bullvalene isomer can be selected and "frozen out" through binding to cyclodextrin as a host. A profiling method has been developed to determine the best binding mode of a host-guest complex. Subsequently, the binding free energy of the complex can be calculated. NMR experimental data and quantum mechanics (QM) calculation results (14) were used to better understand the isomer ratios of bullvalene molecules and thus to select which isomers to investigate. Subsequently, the developed profiling method was applied to determine the optimal host-guest complex structure. The stability of the complex was checked by running an unbiased MD simulation, followed by calculation of the binding free energy values. Ultimately, a comparison of the calculated and experimental binding free energies showed a good agreement. The overall procedure is summarised in Figure 2ii. 1.

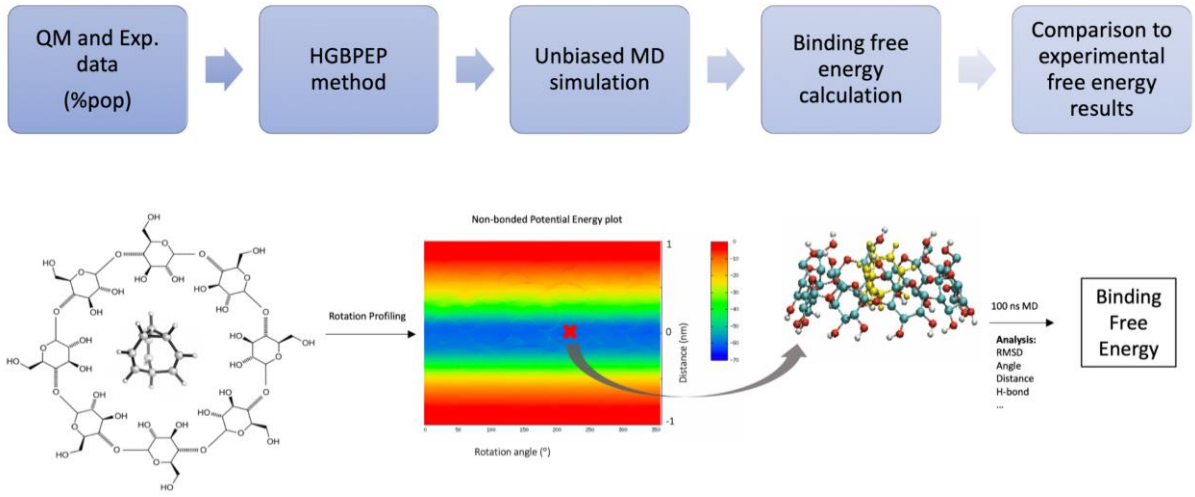


Figure 2ii. 1. Procedure for determining and evaluating the host-guest complex formed by a bullvalene and a cyclodextrin.

Initially, the simplest possible host-guest complex was investigated, that of unsubstituted bullvalene as the guest molecule binding to β-cyclodextrin (βCD) as the host molecule. Three different orientations of the bullvalene guest were studied in complex with βCD. Next, complexation of the mono-substituted (hydroxymethyl) bullvalene isomers were studied, showing that substitution dramatically alters the the number of possible binding modes and their stability. Finally, the procedure was applied on the most populated isomer of di-

substituted (hydroxymethyl) bullvalene. Bullvalene tends to bind at the edge of the cavity of βCD. Since γ-cyclodextrin (γCD) has a bigger cavity to encapsulate bullvalene, complexation of all bullvalene isomers mentioned above with γCD were also considered and studied. A complete list of all bullvalene-cyclodextrin complexes studied is provided in Table 2ii. 1. Figure 2ii. 2 also shows examples of bullvalene isomers and their code names that are used in Table 2ii. 1.

Table 2ii. 1. Overview of the studied bullvalene-cyclodextrin complexes. In the complex code names, the first two numbers refer to the number of substituents, the ΔUp/Down means the cyclopropane ring end of bullvalene is pointing upward/downward relative to the z-axis, Hori stands for the horizontal orientation of bullvalene relative to the z-axis, and β/γ indicates that the studied guest molecules were in complex with β- or γ-cyclodextrin, respectively. The cyclodextrin was always oriented such that its ring lay in the xy-plane and the z-axis ran through its centre. The letters A-D refer to the different isomers of mono-substituted (hydroxymethyl) bullvalene.

Complex code name	R-group 1	R-group 2	Isomer	Orientation	Cyclodextrin
00ΔUpβ/γ	-	-	-	Triangle upward	β or γ
00Horiβ/γ	-	-	-	Horizontal	β or γ
00ΔDownβ/γ	-	-	-	Triangle downward	β or γ
A01ΔUpβ/γ	CH_2OH	-	A	Triangle upward	β or γ
A01HoriRDownβ/γ	CH_2OH	-	A	Horizontal R-group upward	β or γ
A01HoriRUpβ/γ	CH_2OH	-	A	Horizontal R-group downward	β or γ
A01ΔDownβ/γ	CH_2OH	-	A	Triangle downward	β or γ
B01ΔUpβ/γ	CH_2OH	-	B	Triangle upward	β or γ
B01HoriRDownβ/γ	CH_2OH	-	B	Horizontal R-group upward	β or γ
B01HoriRUpβ/γ	CH_2OH	-	B	Horizontal R-group downward	β or γ
B01ΔDownβ/γ	CH_2OH	-	B	Triangle downward	β or γ
C01ΔUpβ/γ	CH_2OH	-	C	Triangle upward	β or γ
C01HoriRDownβ/γ	CH_2OH	-	C	Horizontal R-group upward	β or γ
C01HoriRUpβ/γ	CH_2OH	-	C	Horizontal R-group downward	β or γ
C01ΔDownβ/γ	CH_2OH	-	C	Triangle downward	β or γ
D01ΔUpβ/γ	CH_2OH	-	D	Triangle upward	β or γ
D01Horiβ/γ	CH_2OH	-	D	Horizontal	β or γ
D01ΔDownβ/γ	CH_2OH	-	D	Triangle downward	β or γ
11ΔUpβ/γ	CH_2OH	CH_2OH	-	Triangle upward	β or γ
11HoriRDownβ/γ	CH_2OH	CH_2OH	-	Horizontal R-groups downward	β or γ
11HoriUpβ/γ	CH_2OH	CH_2OH	-	Horizontal R-groups upward	β or γ
11ΔDownβ/γ	CH_2OH	CH_2OH	-	Triangle downward	β or γ

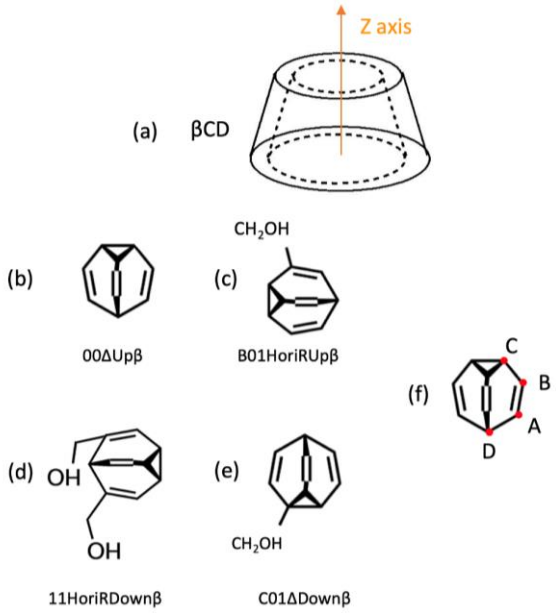


Figure 2ii. 2. (a) Schematic showing the orientation of βCD, (b-e) four examples of bullvalene isomers with their code names when in complex with βCD and (f) the substituent position letters used in the name codes.

2ii.2. Methods

2ii.2.1. Computational methods

2ii.2.1.1. Parameters

GROMOS 53A6_GLYC force field parameters were used for cyclodextrins (20, 192, 193), and GROMOS 54A7-compatible parameters as obtained from the ATB (194–196) were used for bullvalenes (173). The GROMOS force field used for all simulations in this chapter, as well as all other following chapters, is a united atom (UA) force field, with aliphatic hydrogen atoms absorbed into their parent carbon atoms. Choice and validation of force field parameters are described in details in Appendix Section.

2ii.2.1.2. Coordinates

For βCD initial coordinates, a crystal structure was taken from Cambridge Crystallographic Data Centre (166) (www.ccdc.cam.ac.uk , identifier: ARUXIU, and deposition number: 187189) and used for all subsequent simulations. For γCD, a crystal structure was taken from ChemSpider (CSID:10469499, http://www.chemspider.com/Chemical-Structure.10469499.html), and subsequently symmetrised using the molecular visualization program Visual Molecular Dynamics (VMD) (197).

All initial coordinates for substituted bullvalenes were taken from Dr. Pašteka's library (14) of QM-optimised isomer structures. For unsubstituted bullvalene, an unoptimised structure produced using Gaussview 5.0 (198) was submitted to ATB (194–196) and the

resulting QM-optimised coordinates were taken as well as the parameters. The choice and validation of the coordinates are explained in Appendix Section.

2ii.2.1.3. Simulation methods

Simulation software package

All MD simulations were prepared and performed with the GROMACS simulation package version 5.1.1, other than the free energy simulations, which were done using the GROMACS simulation package version 2016.3 (199–201). All visualisations were done using VMD. All graphs were plotted using QtGrace v 0.2.6 (202).

Standard MD simulation procedure

The following procedure was followed for all MD simulations unless otherwise stated. All simulations were initiated by filling the simulation box with solvent molecules (GROMOS methanol model (173) or SPC water (203)) and the entire system subjected to energy minimization (steepest descent minimization). Initial velocities were randomly generated from a Maxwell-Boltzmann distribution at 50 K and the system was heated from 50 to 298 K over 1 ns in the NVT ensemble, with temperature controlled using the modified Berendsen thermostat, V-rescale, with a temperature coupling constant (τ_T) of 0.1 ps (204), and periodic boundary conditions. The LINCS algorithm (205) was used with an order of 4 to constrain all bonds, allowing for an integration time step of 2 fs. Non-bonded interactions were calculated using a Verlet cut-off scheme (206), with searching of neighbouring grid cells. The non-bonded van der Waals interactions and electrostatic interactions within a cut-off distance of 1.4 nm were calculated explicitly. Outside the cut-off distance, electrostatic interactions were approximated using particle mesh Ewald (PME) (204), with a maximal spacing for the FFT grid of 0.16 nm and an interpolation order of 4. The relative dielectric permittivity was ε_{RF} = 62 (when the solvent was water) and 30 (when the solvent was methanol). Subsequently, the system was simulated for 10 ns under isotropic NpT conditions, with pressure controlled using the Berendsen barostat, with a pressure coupling constant (τ_p) of 0.5 ps, an isothermal compressibility of 7.4×10^{-4} bar^{-1}, and a reference pressure of 1 bar. All the other settings were the same as for the NVT ensemble.

Force field validation simulations

The simulations of β-cyclodextrin using the 2016H66 force field varied from the standard procedure in that after heating, the system was equilibrated for 100 ps in NVT followed by 100 ns of NpT simulation. Investigations using the GROMOS 53A6_GLYC varied from the standard procedure in that the final NpT simulations were only run for 50 ns.

HGBPEP simulations

The profiling HGBPEP method simulations lasted only a single integration step, with the purpose of calculating the instantaneous potential energies without the system undergoing dynamics. The rotation and translation of the guest molecule relative to the host were controlled by using a Python 2.7 script to make calls to GROMACS. In the HGBPEP calculations, long-range electrostatic interactions were computed using the particle mesh Ewald method (207) with an interpolation order of 4 and a maximal spacing for the FFT grid of 0.16 nm. The pressure coupling time for the Berendsen barostat was τ_p = 1.0 ps.

Bullvalene-cyclodextrin complex simulations

MD simulations of each bullvalene-cyclodextrin complex were initiated from the most energetically favourable coordinates identified using HGBPEP using the standard procedure, except that following the heating phase, the system was further equilibrated for 100 ps in the NpT ensemble prior to beginning the 100 ns production run. During all NpT simulations, long-range electrostatic interactions were computed using the particle mesh Ewald method (207) with an interpolation order of 4 and a maximal spacing for the FFT grid of 0.16 nm and a pressure coupling time for the Berendsen barostat, τ_p, of 1.0 ps.

Free energy calculations

Solvation and binding free energies were computed using thermodynamic integration (TI). The general protocol is outlined here and system-specific details are provided below. Twenty λ values from 0 to 1 were used to switch off the Coulombic and then van der Waals interactions. After solvating the system, for each λ value, the system underwent two steps of energy minimization (steepest descent minimization and limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) minimization), followed by heating from 50 to 298 K over 1 ns in the NVT ensemble, and equilibration for a further 1 ns under isotropic NpT conditions, following the standard procedure except that the particle mesh Ewald (PME) method was used

to compute the long-range electrostatic interactions with a maximal spacing for the FFT grid of 0.12 nm and an interpolation order of 6. The data production simulations were run for 10 ns at each λ value using stochastic dynamics (SD) with a Langevin thermostat, with τ_T, the inverse friction coefficient, of 0.1, a friction coefficient of mass/τ_T for each particle, and a pseudo-random seed for the random number generator for thermal noise. From these simulations, $\partial H/\partial \lambda$, and subsequently the free energy differences, were estimated using Bennett's acceptance ratio (BAR) as implemented in the GROMACS `gmx bar` program. To compute the binding free energy of methanol to βCD, a single methanol molecule was placed at the centre of the βCD cavity. The system was then solvated using SPC water (203) with a buffer region of 0.15 nm.

Dipole moment

The dipole moment values were calculated with Gaussian (208), using the basis set b3lyp/6-311+g(d,p) (209–211).

2ii.2.1.4. Analysis methods:

VMD was used to view the simulation trajectories and create structural images.

RMSD (Root Mean Square Deviation)

The root mean square deviation (RMSD) of specific groups of atoms in a molecule with respect to a reference structure, such as the first frame of the simulation trajectory, was calculated using the GROMACS `gmx rms` program as

$$\mathrm{RMSD}(t) = \left[\frac{1}{M} \sum_{i=1}^{N} m_i |\mathbf{r}_i(t) - \mathbf{r}_i^{\mathrm{ref}}|^2 \right]^{1/2},$$

Equation 2ii. 1

where $\mathbf{r}_i(t)$ is the position of atom i at time t after least squares fitting the structure to the reference structure and $\mathbf{r}_i^{\mathrm{ref}}$ is the position of atom i in the reference structure. N is the number of atoms, and $M=\Sigma_i m_i$ is the sum of the atomic masses m_i. The choice of atoms and reference structure depend on the situation and are therefore specified in each case.

Bullvalene-cyclodextrin interaction angle

The angle between the planes defined by three carbon atoms of bullvalene, and three oxygen atoms in cyclodextrins (Figure 2ii. 3) during the MD simulations was computed using the GROMACS `gmx gangle` program.

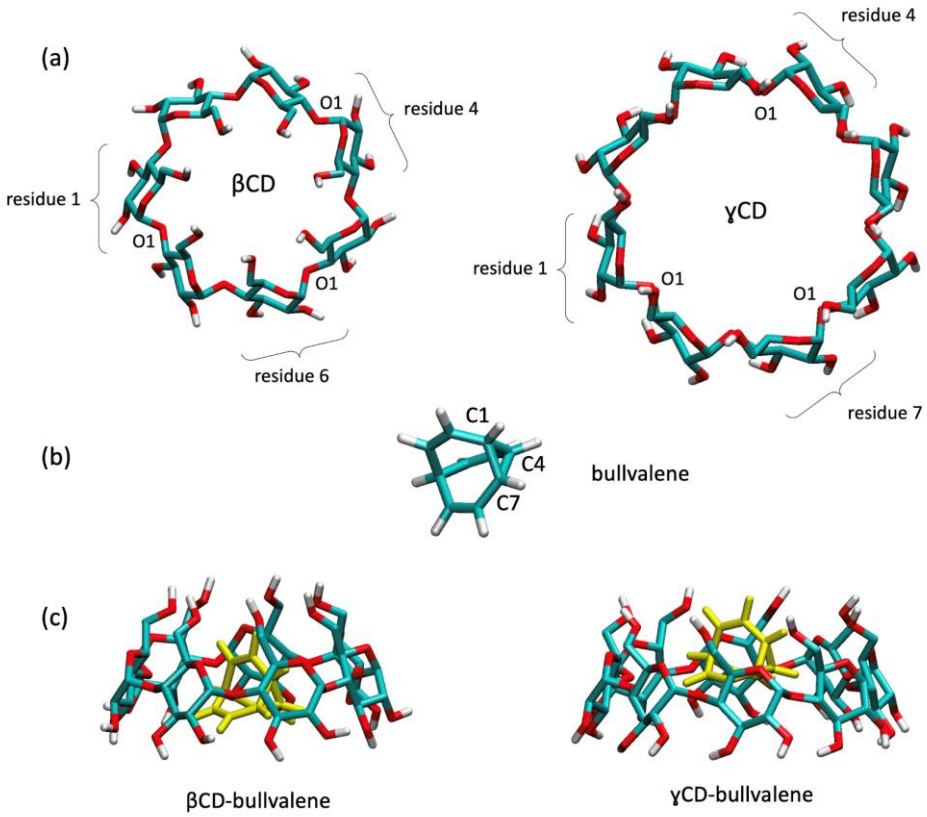


Figure 2ii. 3. Selected atoms which define the planes for angle analysis, (a) top view of βCD (left) and γCD (right) with the three oxygen atoms of three sugar units used to define the plane labelled, (b) the three chosen carbon atoms defining the plane in bullvalene, (c) side view of the bullvalene-cyclodextrin complexes. Bullvalene is coloured yellow for better visualisation.

Inter-atomic distances

The distance between pairs of atoms during an MD simulation was computed using the GROMACS `gmx mindist` program.

Hydrogen bonds

The formation of hydrogen bonds was analysed using in-house Python 2.7 scripts to call VMD.

Data visualisation

The 2D and 3D energy heat maps generated by the HGBPEP procedure and the hydrogen bond time-series were plotted using the Python 2.7 matplotlib package. The colour scales of the

energy heat maps are not always consistent because matplotlib does not allow the colour scale to go beyond the minimum and maximum values of the data.

2ii.2.2. Experimental methods

All experimental work was carried out by Oussama Yahiaoui and Dr Thomas Fallon. Unsubstituted bullvalene and its mono- and di-substituted methyl-hydroxy analogues were synthesised through cobalt-catalyzed [6+2] cycloaddition of cyclooctatetraene to alkynes, followed by photochemical di-π-methane rearrangement (14). Experimental isomer ratios of all studied bullvalenes were determined using low temperature NMR.

2ii. 3. Results and discussion

2ii.3.1. Unsubstituted bullvalene and β-cyclodextrin

Having shown that the HGBPEP method reproduces the experimental results for pCP, and gained insight into how geometry and charge affect the preferred binding mode, the HGBPEP method was used to determine the binding modes of the bullvalene molecules to β- and γCD, as laid out in Table 2ii. 1. Initially, the binding of unsubstituted bullvalene to βCD was studied.

2ii.3.1.1. Identification of most favourable modes for bullvalene binding to βCD

Three different orientations of unsubstituted bullvalene were used (Figure 2ii. 4): one in which its three-fold axis of symmetry is aligned with the *z*-axis and the cyclopropane ring is oriented upward towards the smaller rim of the cyclodextrin (00ΔUpβ), one where the ring is horizontal (00Horiβ), and the third in which the three-fold axis of symmetry is aligned with the *z*-axis and the cyclopropane ring is oriented downward, towards the larger rim of the cyclodextrin (00ΔDownβ). Figure 2ii. 5 shows 2D heat maps of the van der Waals, Coulombic and combined non-bonded potential energies for unsubstituted bullvalene in complex with βCD in the three different orientations produced using HGBPEP.

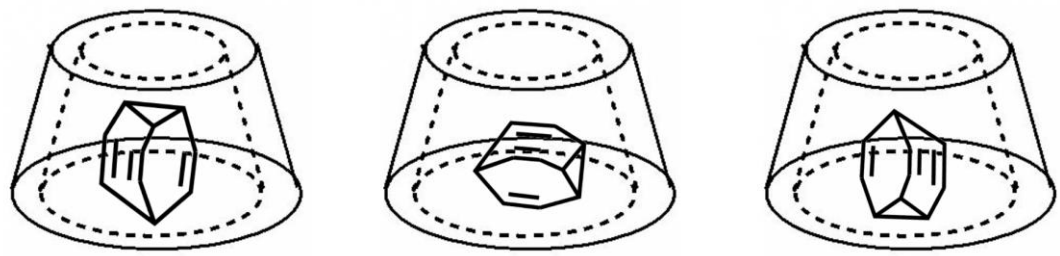


Figure 2ii. 4. Schematics of the three complexes of unsubstituted bullvalene and βCD, showing the (left) 00ΔUpβ, (middle) 00Horiβ and (right) 00ΔDownβ orientations of bullvalene.

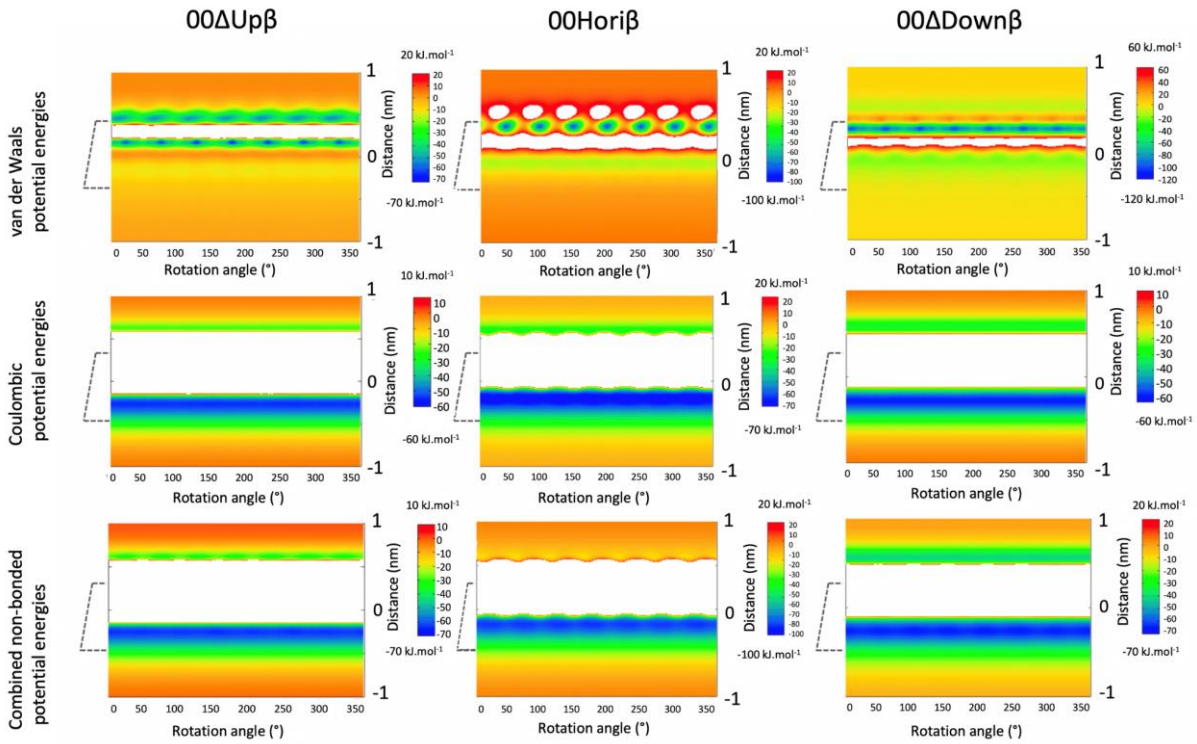


Figure 2ii. 5. HGBPE profiles for the (left) 00ΔUpβ, (centre) 00Horiβ and (right) 00ΔDownβ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

The first question is whether bullvalene fits into the cavity of βCD. According to the van der Waals profiles (Figure 2ii. 5, top row), there is only a small region with large unfavourable van der Waals energies (white horizontal band), situated between 0.20-0.40 nm for 00ΔUpβ, 0.10-0.30 nm for 00Horiβ and 0.10-0.20 nm for 00ΔDownβ. For 00ΔUpβ, bullvalene can bind favourably to βCD immediately above and below this zone, with both regions showing weak seven-fold symmetry, reflecting the symmetry of βCD. In contrast, for 00Horiβ and 00ΔDownβ there is a series of weakly favourable binding positions at about −0.10 nm and for 00Horiβ, a strongly seven-fold symmetric pattern of favourable binding positions at around 0.35 nm, and more weakly seven-fold symmetric pattern of favourable binding positions at around 0.30 nm

for 00ΔDownβ. The 00Horiβ complex also has a seven-fold symmetric series of unfavourable binding positions (white spots at z = 0.5 nm) of bullvalene at the upper rim of βCD. Therefore, based on van der Waals energies, it can be concluded that βCD can physically encapsulate bullvalene inside its cavity other than in a small region just inside the narrower upper rim, where there is a high energetic penalty to complexation due to atoms sterically clashing.

The Coulombic potential energy heat maps (Figure 2ii. 5, middle row), however, show that the region of highly unfavourable binding is much wider, spanning from –0.20 – 0.60 nm, i.e. from just below the centre of geometry of the βCD to its upper rim. The more energetically favourable region for bullvalene binding to βCD in any of the three orientations according to the Coulombic potential energy is between −0.20 and −0.25 nm below the centre of geometry of βCD, which is just at its wider rim. Only for 00Horiβ is there a hint of the seven-fold symmetry.

The separate van der Waals and Coulombic energy plots are useful for understanding the driving forces for complexation, but ultimately, successful complexation depends on the total potential energy (as well as the entropy, but that is not taken into account in HGBPEP). Since the white regions represent very unfavourable potential energies, the wider inaccessible binding regions in the Coulombic energy heat maps dominate in the total potential energy heat maps (Figure 2ii. 5, bottom row), obliterating the binding modes that are only favourable according to the van der Waals energies. Overall, therefore, the favourable binding region just below the centre of geometry of βCD is the most energetically favourable zone for unsubstituted bullvalene binding to βCD, with binding in this region only dependent on orientation for 00Horiβ, and even then only weakly.

2ii.3.1.2. Stability of favourable modes for bullvalene binding to βCD

The HGBPEP method is a useful way to rapidly screen different possible binding positions and orientations, but it does not allow either molecule or the solvent to relax (aside from energy minimisation). The most energetically favourable coordinates of the bullvalene-βCD complexes were therefore selected from the favourable binding regions in the HGBPEP potential energy heat maps and subjected to 100 ns unbiased MD simulations. In order to distinguish between the structures before and after MD simulations, the letters "MD" are added to the name codes of the complexes. Table 2ii. 2 shows the position along the z-axis and the rotation angles of the chosen coordinates from the plots, as well as their potential energies. First, the energetically most favourable z position was chosen according to the combined

energy plots. Then for that distance, a low energy rotation angle was picked. The coordinates at that combination of *z* position and rotation angle were used to initiate a 100 MD simulation.

Table 2ii. 2. Selected distances and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for 00ΔUpβ, 00Horiβ, and 00ΔDownβ.

Complex code name	Pos. (nm) *z*-axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
00ΔUpβ	-0.25	0	-64.02
00Horiβ	-0.17	0	-88.08
00ΔDownβ	-0.25	0	-64.69

Visualisation of the MD trajectory showed that bullvalene mostly stays below the centre of the cavity of βCD throughout the simulation, and the stability of the binding mode is supported by the consistently low RMSD values (Figure 2ii. 6, second row). To determine whether the orientation of the bullvalene changed during the MD simulations, the angle between two selected planes in the host and the guest molecules (Figure 2ii. 6, bottom row, right) was calculated during the simulation time.

Regardless of bullvalene's orientation at the start of MD simulations, during the simulation time, bullvalene mostly tends to orient horizontally with respect to βCD, with the angle between the two selected planes predominantly about 90-100° (Figure 2ii. 6, third row and bottom row, left)). This highlights the importance of running MD simulations after identifying favourable binding modes.

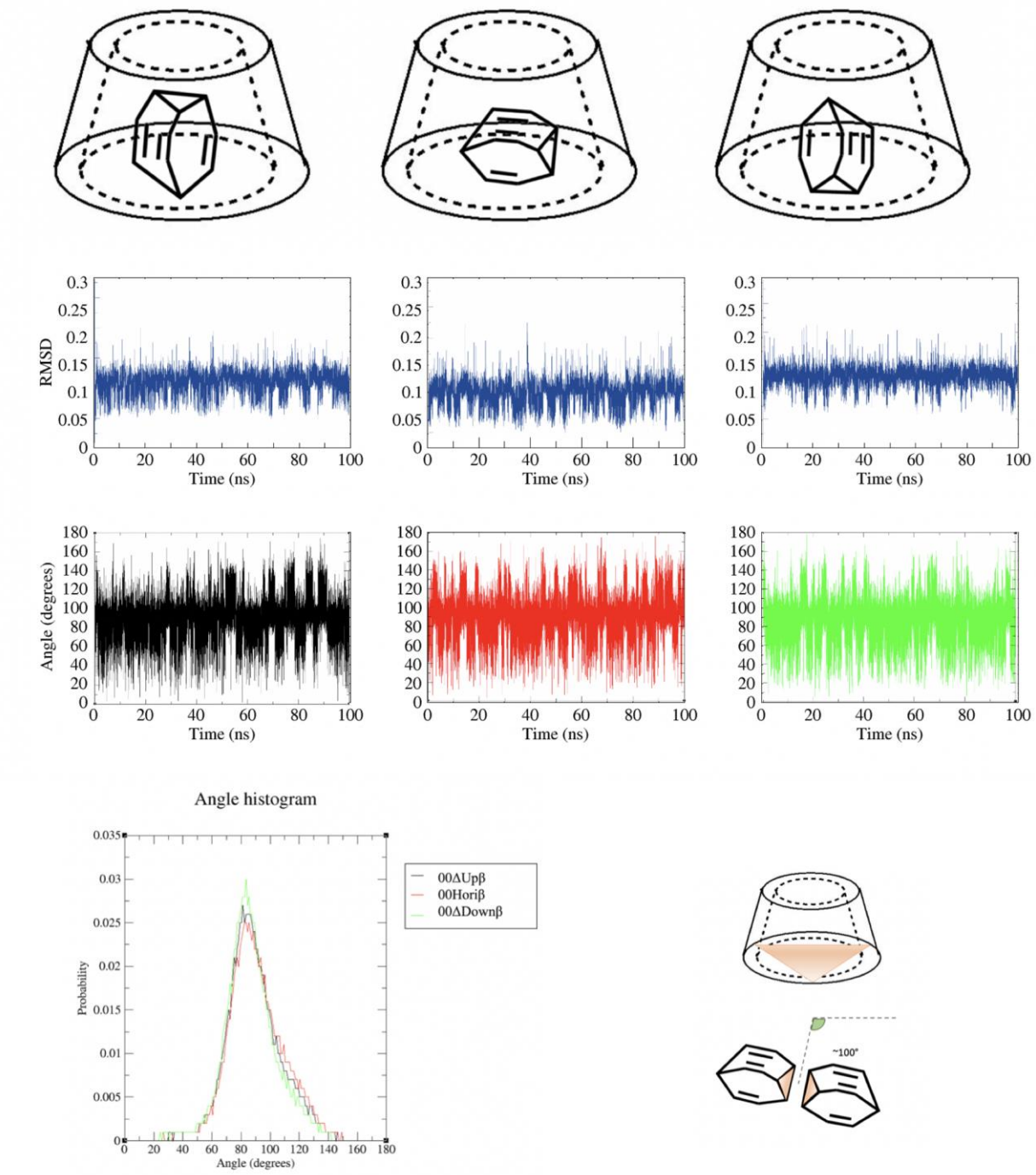


Figure 2ii. 6. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left) and illustration of the most populated the orientation angle (bottom row, right) of unsubstituted bullvalene in complex with βCD during 100 ns MD simulations initiated from three different orientations. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

2ii.3.1.3. Calculation of free energies for binding of unsubstituted bullvanene to βCD

While MD simulations incorporate flexibility and allow conformational motion, which was important for identifying that there is only one major mode of unsubstituted bullvalene binding to βCD, ultimately, binding affinity should be evaluated by computing the binding free energy, which takes into account entropic as well as enthalpic factors.

To choose favourable coordinates from the MD trajectory from which to initiate binding free energy calculations, the average value for the angle across the simulation was calculated. The distance between the centre of mass (COM) of bullvalene and the COM of βCD was measured over the simulation time (Table 2ii.3), and the average distance was also calculated. Subsequently, the time frame of the trajectory which has the two average values of angle and COM distance was taken. This procedure of choosing the coordinates from which to start free energy calculations is the same for all free energy calculations in this Chapter, and when it differs, this is stated explicitly in the relevant section.

Table 2ii. 3. Average values and standard deviation of the angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for 00ΔUpβMD, 00HoriβMD, and 00ΔDownβMD.

Complex code name	**Angle (°)** (between the two planes)	**Dist. (nm)** (COM of bullvalene and βCD)	**RMSD (nm)**
00ΔUpβMD	avg. 90.00 ± 19.27	0.144 ± 0.06	0.11 ± 0.02
00HoriβMD	avg. 90.47 ±19.85	0.141 ± 0.06	0.11 ± 0.02
00ΔDownβMD	avg. 86.73 ± 18.63	0.144 ± 0.06	0.12 ± 0.01

Based on the convergence of the MD simulations towards a horizontal binding mode, coordinates were selected using the above procedure from 00HoriβMD. The binding free energy of bullvalene to βCD was computed using TI, which also includes calculation of the solvation free energy of the bullvalene-βMD complex. Both values are favourable, indicating that unsubstituted bullvalene is likely to bind stably to βCD and the complex is soluble in water (Table 2ii. 4). Unfortunately, it is not possible to compare with experimental data as none are available.

In summary, a procedure for identifying favourable binding modes and evaluating their stability and binding affinity has been developed and used to study the complexation of unsubstituted bullvalene with βCD.

Table 2ii. 4. Calculated free energy values for the binding of unsubstituted bullvalene to βCD and for the solvation of the complex.

Complex Code	Binding Free Energy (kJ·mol^{-1})	Solvation Free Energy (kJ·mol^{-1})
00HoriβMD	-10.99 ± 0.51	-24.76 ± 0.17

2ii.3.2. Mono-substituted hydroxymethyl-bullvalene and β-cyclodextrin

While bullvalene is a fluctional molecule, the isomers of unsubstituted bullvalene are chemically indistinguishable. Substitution makes the isomers chemically distinguishable; in the case of mono-substitution, there are just four distinguishable isomers, but these isomers will interconvert constantly in solution. The goal of complexation of bullvalenes with βCD is therefore to capture a single isomer. The procedure developed in the previous section should allow prediction of possible binding modes and the binding free energy for each potential complex, and thus predict the likelihood of complexation and which isomer(s) are most likely to be captured.

After communication with our experimental collaborators, the mono-substituted bullvalene hydroxymethyl-bullvalene was chosen. It has four distinct isomers, which are shown in Figure 2ii. 7. The relative populations of these in aqueous solution have been determined experimentally, by NMR titration (14), and computationally, using QM calculations (14) (Table 2ii. 5). The procedure developed in Section 2ii.3.1 was applied to all four isomers with βCD.

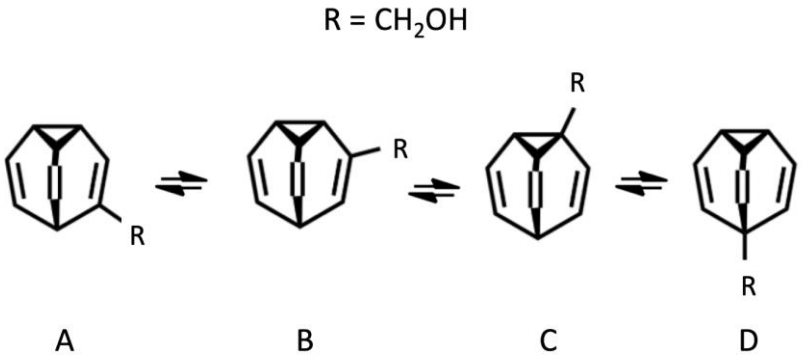


Figure 2ii. 7. The four distinct isomers of mono-substituted bullvalene.

Table 2ii. 5. Population of each isomer of hydroxymethyl-bullvalene determined experimentally, by NMR titration (14), and computationally, by QM calculations (212).

Isomer	Population (%)	
	Experimental	Computational
A	57	54
B	25	43
C	18	3
D	0	1

2ii.3.2.1. Identification of most favourable modes for hydroxymethyl-bullvalene binding to βCD

The HGBPEP method was applied for the four isomers of hydroxymethyl-bullvalene (A-D, which are also listed in Table 2ii. 1, on page 64) in complex with βCD to determine the most favourable binding mode(s). For each isomer, three different orientations of the cyclopropane ring relative to the central axis of the cyclodextrin cavity were considered as for unsubstituted bullvalene, ΔUp, ΔHori and ΔDown. For the horizontal orientation, there are additionally two different orientations of the hydroxymethyl substituent to consider for isomers A, B and C: oriented upwards, HoriRUp and downwards, HoriRDown. The complete abbreviated name for each orientation then begins with the isomer (A-D) and its binary code (01), and ends with the cyclodextrin used, β.

Isomer A

The HGBPE profiles for isomer A all show the seven-fold symmetry patterns representing the seven sugar units of βCD to a greater degree than for unsubstituted bullvalene, reflective of the increased size and asymmetry of the bullvalene due to addition of a substituent (Figure 2ii. 8). The CH_2OH substituent also results in more complicated van der Waals energy profiles compared to the unsubstituted bullvalene plots. The seven disallowed regions at z = 0.7 – 0.8 nm for A01ΔUpβ and A01HoriRDownβ are due to the substituent coming close to the OH groups of the upper rim of βCD, which does not occur for A01HoriRUpβ and A01ΔDownβ because the substituent is oriented upwards in these isomers. Similarly, A01HoriRDownβ has less favourable binding modes at the lower rim of βCD, indicating that its upward-oriented substituent has unfavourable interactions with the lower OH groups on the bigger rim of βCD. As for unsubstituted bullvalene, the Coulombic energy profiles exhibit a broad band of highly unfavourable binding from z = –0.3 nm to z = 0.7 nm, which has a dominant effect on the total potential energy. All four orientations have favourable Coulombic interactions with the lower rim of βCD, but only A01ΔUpβ and A01HoriRUpβ can bind favourably at the lower rim, due

to unfavourable interactions of the downward-oriented hydroxymethyl substituent of the other orientations with the upper rim of βCD. The coordinates for MD simulation were taken from the binding zones which were favourable in the total potential energy profiles for each of the four orientations, using the same method as Section 2ii.3.1 (Table 2ii. 6).

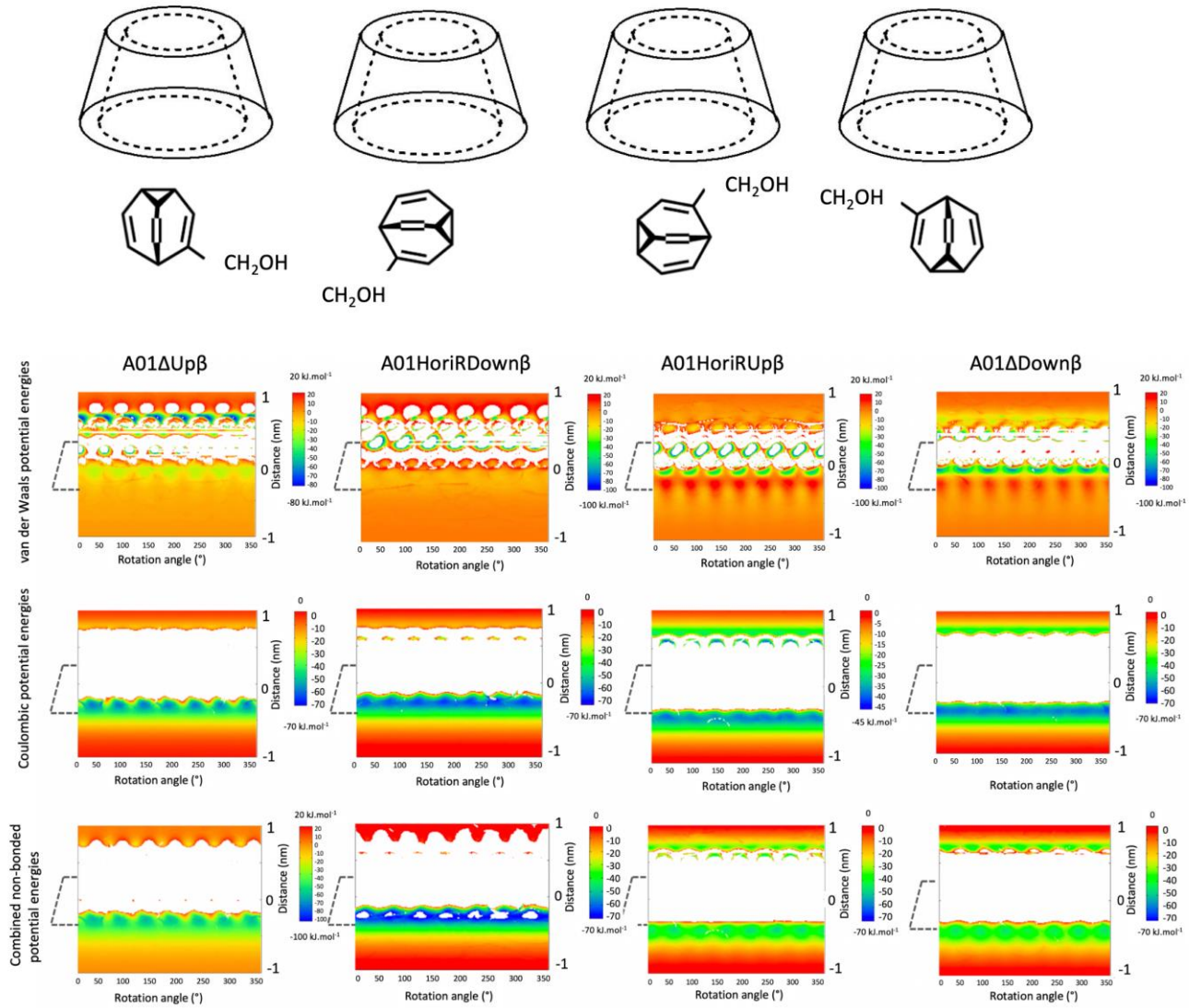


Figure 2ii. 8. HGBPE profiles for the (left) A01ΔUpβ, (centre left) A01HoriRDownβ, (centre right) A01HoriRUpβ and (right) A01ΔDownβ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

Table 2ii. 6. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for A01ΔUpβ, A01HoriRDownβ, A01HoriRUpβ, and A01ΔDownβ.

Complex code name	Pos. (nm) Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
A01ΔUpβ	-0.30	340	-67.19
A01HoriRDownβ	-0.25	255	-78.69
A01HoriRUpβ	-0.40	262	-49.25
A01ΔDownβ	-0.43	150	-42.83

Isomer B

Similar to A-βCD, the HGBPE profiles for isomer B-βCD show the more distinct seven-fold symmetry patterns and more complicated van der Waals energy profiles reflective of the increased size and asymmetry of the bullvalene due to addition of a substituent (Figure 2ii. 9). The van der Waals heat maps indicate that this isomer can physically fit into the lower parts of the cavity of βCD, but there is again a large region of highly unfavourable Coulombic energies that rule out binding within the cavity other than for B01ΔDownβ. In all cases, the Coulombic energy profiles show that B-βCD can bind favourably at the lower rim of βCD, and for B01ΔUp, B01HoriRUpβ and B01ΔDownβ there is an additional favourable binding zone at the upper rim of βCD at about z = 0.7 nm. This region is not favourable for binding in B01HoriRDownβ because in this orientation, the substituent is oriented downward, and while approaching βCD from the top, the partial charges on the substituent would be unfavourably close to the like charges on the upper OH groups of βCD. For all four orientations, the total potential energy profiles show favourable binding regions at or just within (B01ΔDownβ) the lower rim of βCD, from where coordinates were taken for MD simulation, using the method in Section 2ii.3.1. (Table 2ii. 7).

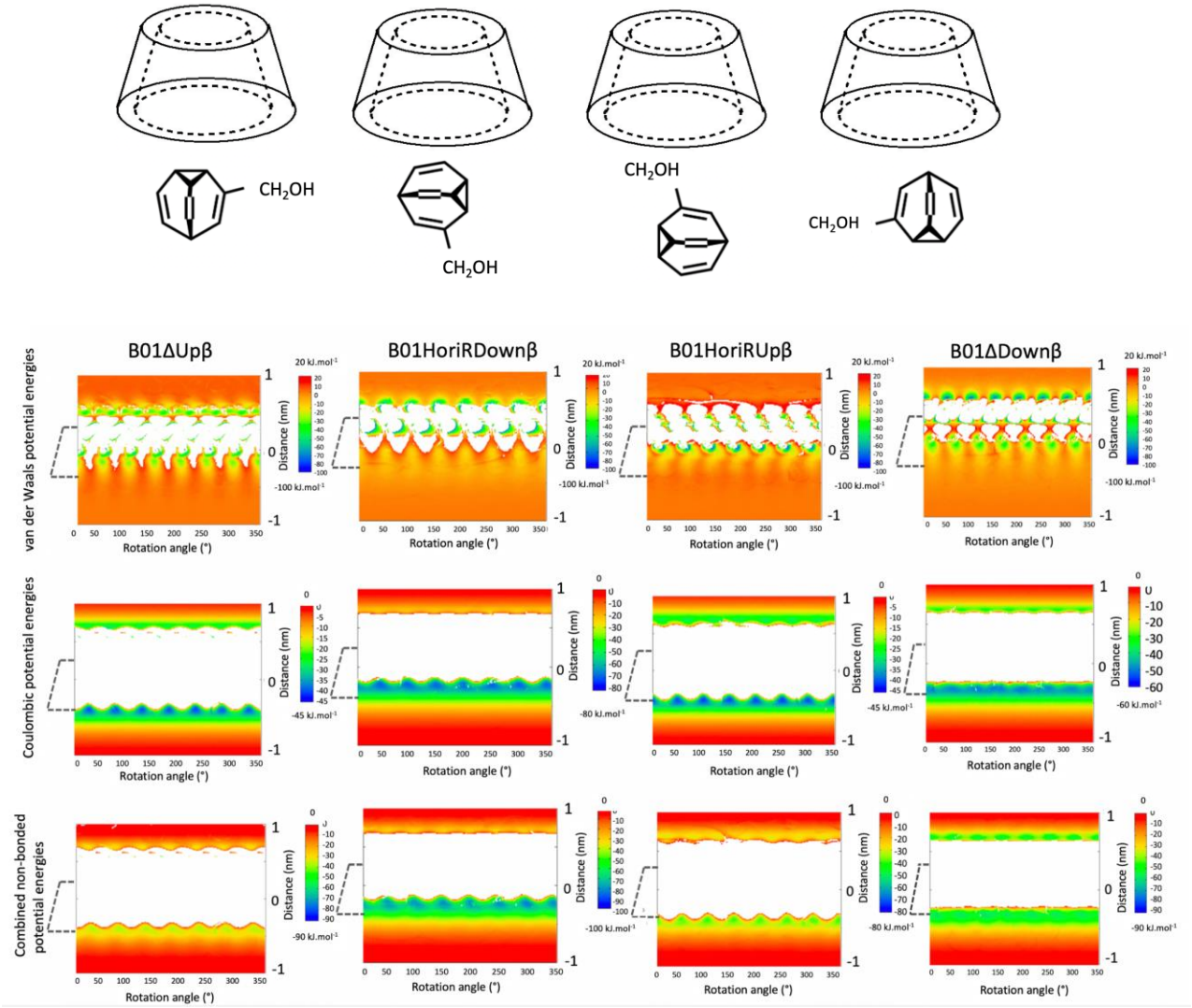


Figure 2ii. 9. HGBPE profiles for the (left) B01ΔUpβ, (centre left) B01HoriRDownβ, (centre right) B01HoriRUpβ and (right) B01ΔDownβ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

Table 2ii. 7. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for B01ΔUpβ, B01HoriRDownβ, B01HoriRUpβ, and B01ΔDownβ.

Complex code name	Pos. (nm) Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
B01ΔUpβ	-0.37	266	-42.06
B01HoriRDownβ	-0.25	203	-81.41
B01HoriRUpβ	-0.36	355	-37.73
B01ΔDownβ	-0.30	63	-58.00

Isomer C

The HGBPE profiles for C-βCD are again similar to those of A-βCD and B-βCD (Figure 2ii. 10). While it is sterically feasible for the hydroxymethyl-bullvalene to sit inside the βCD in specific orientations (both of the substituent and during the profiling) and locations, there is a wide region of unfavourable Coulombic interactions that largely prevent entry of hydroxymethyl-bullvalene into βCD. Ultimately, the total potential energy profiles show that isomer C can only bind to the lower rim of βCD, with a seven-fold symmetric pattern of the most favourable binding orientations in all cases. Although there are favourable binding modes within the cavity for the C01HoriRUpβ and C01ΔDownβ complexes, they lie within the unfavourable binding zone, and thus there is no path for bullvalene to access these binding modes. The coordinates for all four orientations were therefore again selected from the most favourable binding zones at the lower rim of βCD, using the method in Section 2ii.3.1 (Table 2ii. 8).

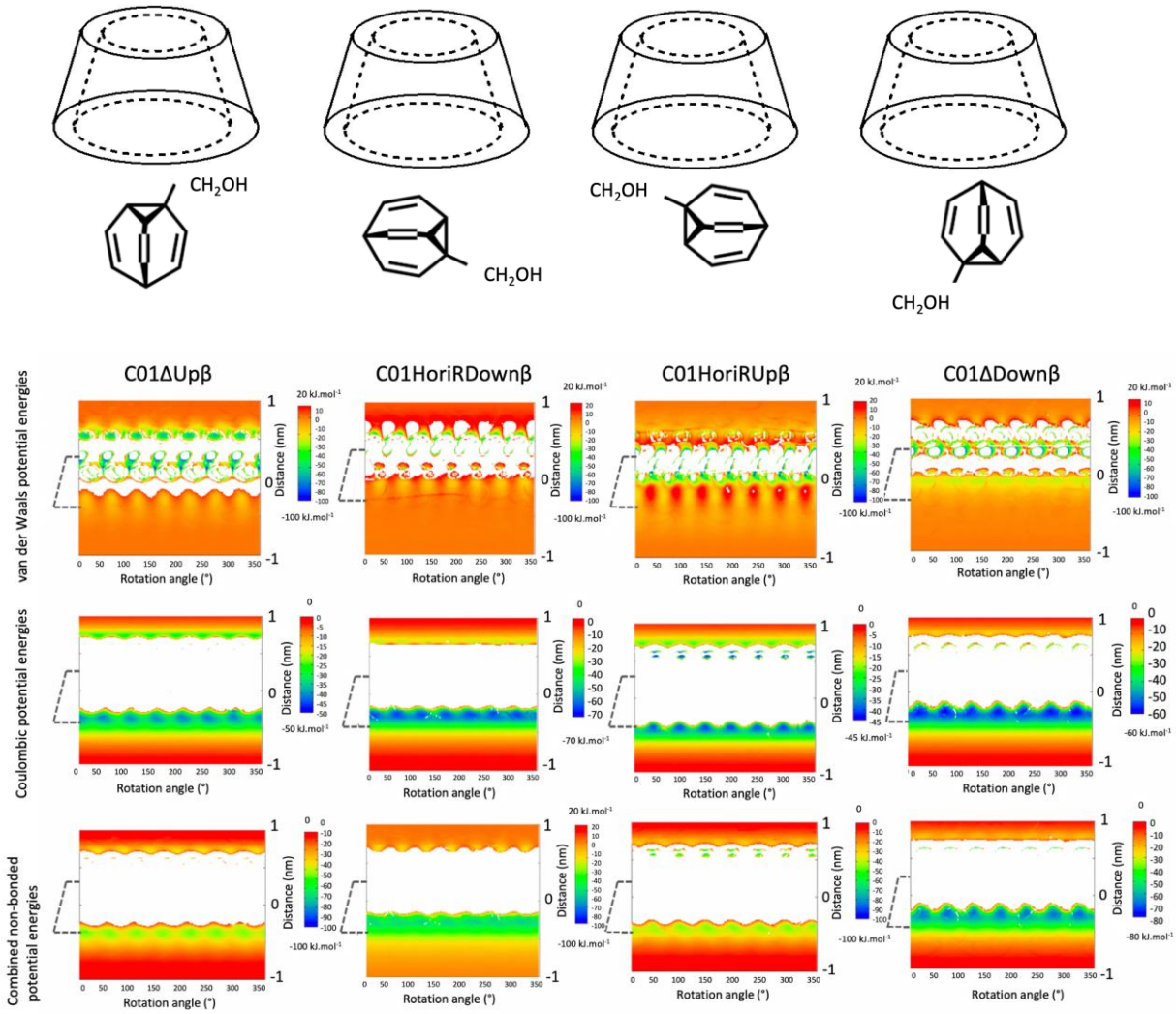


Figure 2ii. 10. HGBPE profiles for the (left) C01ΔUpβ, (centre left) C01HoriRDownβ, (centre right) C01HoriRUpβ and (right) C01ΔDownβ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

Table 2ii. 8. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for B01ΔUpβ, B01HoriRDownβ, B01HoriRUpβ, and B01ΔDownβ.

Complex code name	Pos. (nm) Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
B01ΔUpβ	-0.37	266	-42.06
B01HoriRDownβ	-0.25	203	-81.41
B01HoriRUpβ	-0.36	355	-37.73
B01ΔDownβ	-0.30	63	-58.00

Isomer D

There are again many similarities between the HGBPE profiles for D-βCD and those of the other isomers (Figure 2ii. 11). There are also some key differences, however. The van der

Waals profile for D01ΔDownβ has two highly unfavourable binding zones, one within the narrow part of the βCD cavity, and the other above it, which corresponds to the hydroxymethyl substituent pointing down and clashing with the βCD. The unfavourable binding zone for D01Horiβ is wider than for the other two orientations. This is because the elongated shape of isomer D, in its horizontal orientation, cannot fit within the narrow upper part of the βCD cavity. The Coulombic energy profiles again feature a wide unfavourable binding region, especially for D01Horiβ, due to its elongated shape, which also gives rise to more obvious seven-fold symmetry. The most favourable binding region here and in the total potential energy profiles is around z = –0.2 to –0.3 nm in D01ΔDownβ and D01ΔUpβ, and lower, between z = –0.5 to –0.3, for D01Horiβ, again due to its shape. The coordinates for MD simulation were therefore selected from these regions, using the same methods as in Section 2ii.3.1 (Table 2ii. 9).

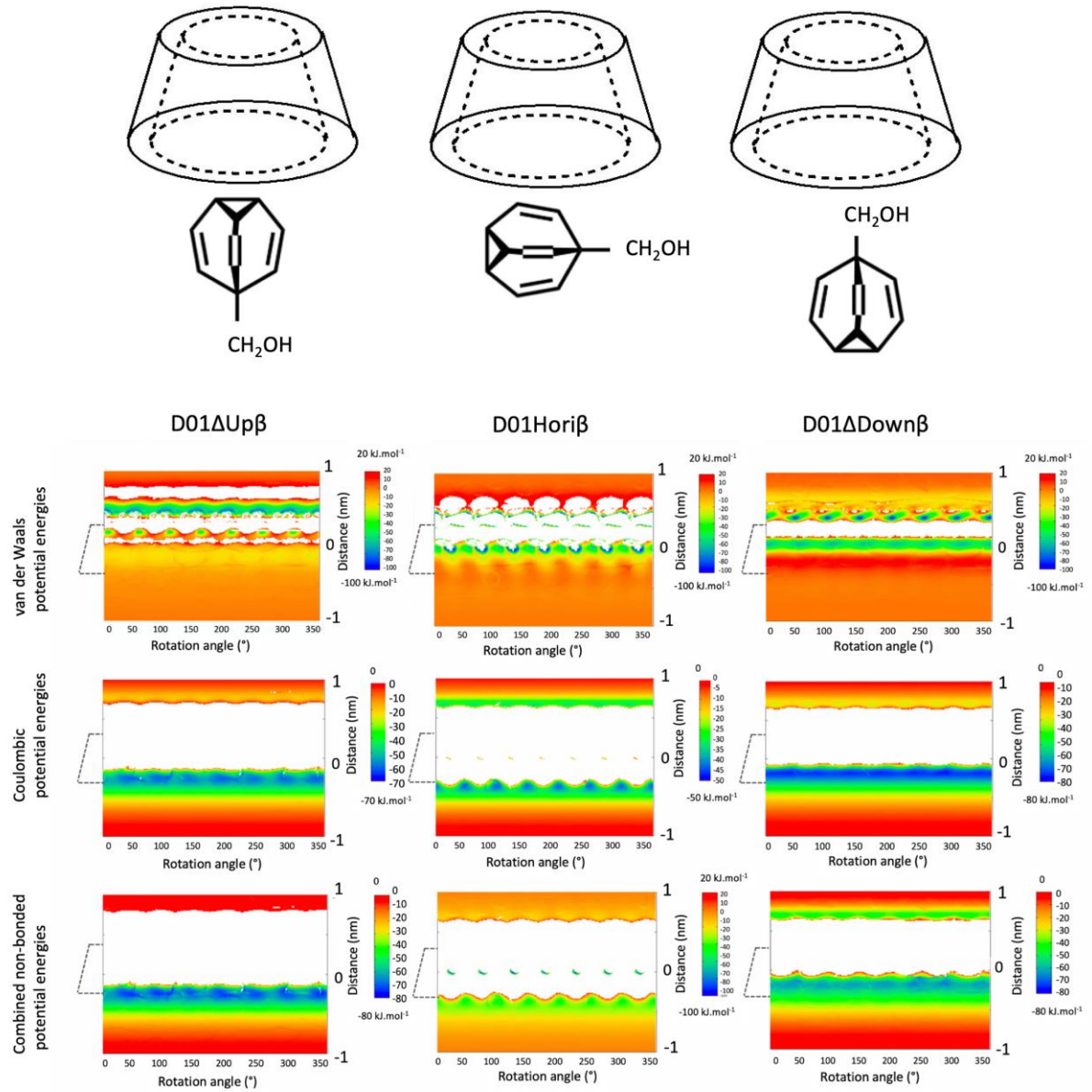


Figure 2ii. 11. HGBPE profiles for the (left) D01ΔUpβ, (centre) D01Horiβ, and (right) D01ΔDownβ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

Table 2ii. 9. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for D01ΔUpβ, D01Horiβ, and D01ΔDownβ.

Complex code name	Pos. (nm) Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
D01ΔUpβ	-0.25	90	-73.58
D01Horiβ	-0.35 -0.35	155 201	-40.85 -43.07
D01ΔDownβ	-0.20	289	-64.22

2ii.3.2.2. Stability of favourable modes for mono-substituted bullvalene binding to βCD

The most energetically favourable coordinates of the complexes of each orientation of each isomer of hydroxymethyl-bullvalene with βCD complexes were subjected to 100 ns unbiased MD simulations. The method of choosing these favourable coordinates was the same as in unsubstituted bullvalene, and the chosen coordinate sets are detailed in Table 2ii. 6-9.

Isomer A

The stability of the complex was monitored first by calculating the RMSD from the initial coordinates (Figure 2ii. 12, top row). All four complexes are either stable (A01HoriβRUpMD) or meta-stable, with A01UpβMD, A01HoriRDownβMD and A01DownβMD exhibiting step jumps in their RSMD time-series but remaining stable between these. Regardless of the orientation of the start point of each simulation, hydroxymethyl-bullvalene tends to lie at an angle of either 70° or 110° (Figure 2ii. 12, second row), with changes in the orientation angle corresponding to the jumps in the RMSD. In both of these orientations, the substituent mostly points toward the lower (70°) or upper (110°) rim of βCD, rather than the inside of the cavity, which makes sense given the hydrophobic interior of cyclodextrin and polar nature of the hydroxy group. A01HoriRUpβMD is more favourable than A01HoriRDownβMD, as it is present for a larger proportion of the simulations (Figure 2ii. 12, bottom row, left), and even in the simulation initiated from the A01HoriRDownβMD complex, the bullvalene rotated to spend the majority of the simulation with the substituent oriented upwards, with only a brief return to its initial configuration. Additionally, in the A01HoriRUpβMD simulation, the substituent remained pointing upward for the entire simulation.

It would have been helpful to extend the simulations as both the RMSD and angle values are not completely converged. However, as the complexes are meta-stable, two starting structures and orientations were chosen for binding free energy calculations, although free energy calculation is a computationally expensive method.

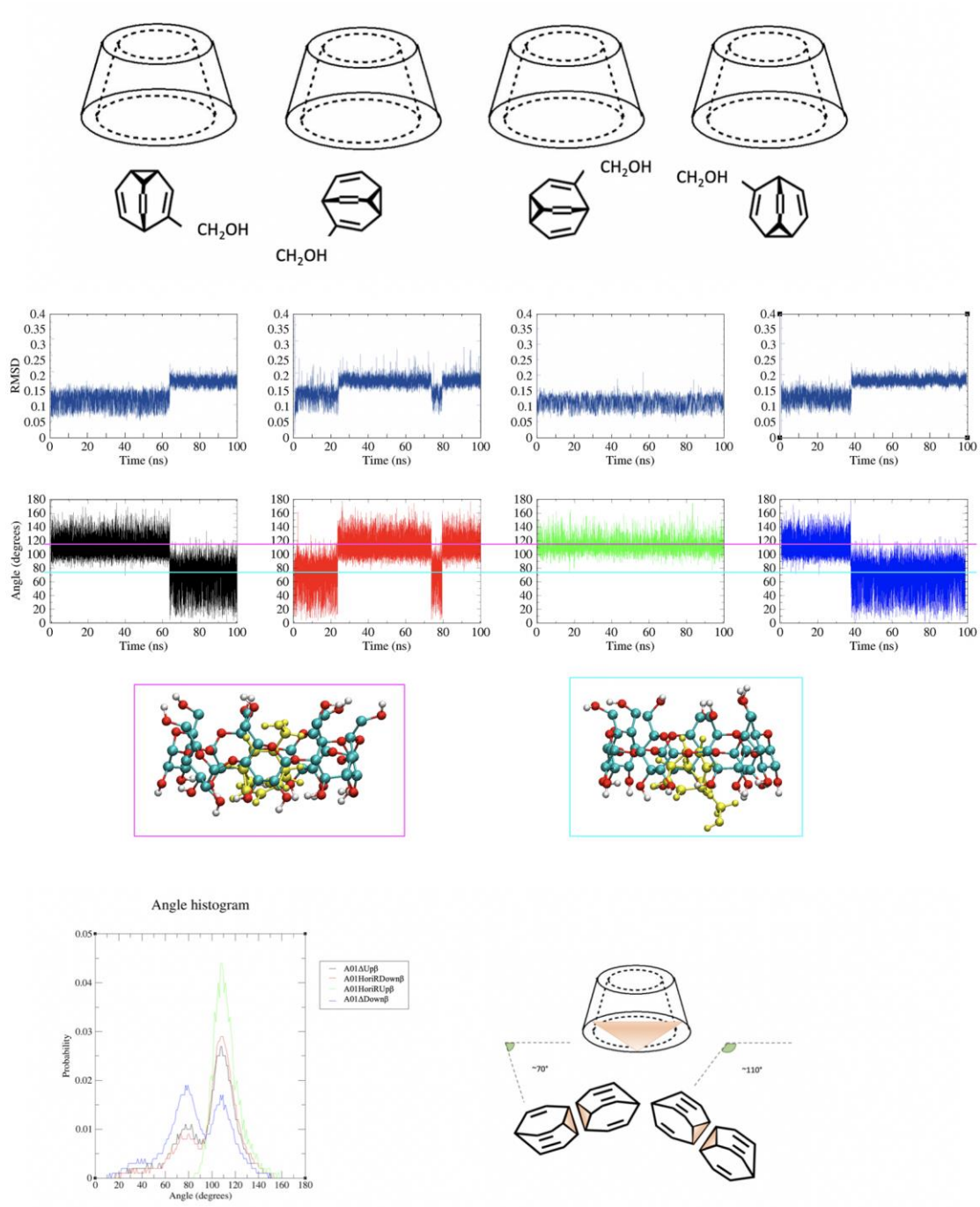


Figure 2ii. 12. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left) and most populated orientation angles (bottom row, right) of isomer A of hydroxymethyl-bullvalene in complex with βCD during 100 ns MD simulations initiated from four different orientations. The two structures in row four are of the two most populated binding modes of hydroxymethyl-bullvalene, indicated by horizontal lines in the angle time-series that match the border colours. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

To choose favourable coordinates from the MD trajectory for calculation of the binding free energy, the average value of the orientation angle was calculated for A01HoriRUpβMD and the modal angle value was calculated for the other three simulations (Table 2ii. 10). The transitions that occur during these simulations mean that the average angle would not represent

either sampled binding mode whereas the modal angle represents the most sampled orientation. The distance between the COM of bullvalene and of βCD was also calculated over the simulation time, and the average computed (Table 2ii. 10). Then, the time frame of the trajectory which has the two average (or modal) values of the orientation angle and COM distance was taken from the trajectory and used as the initial coordinates for the binding free energy calculation.

Table 2ii. 10. Average/modal values and standard deviation of the angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for A01ΔUpβMD, A01HoriRDownβMD, A01HoriRUpβMD, and A01ΔDownβMD.

Complex code name	**Angle (°)** (between the two planes)	**COM Dist. (nm)** (between COM of bullvalene and βCD)	**RMSD (nm)**
A01ΔUpβMD	mod. 105.65 ± 25.32	0.144 ± 0.08	0.13 ± 0.03
A01HoriRDownβMD	avg. 103.75 ± 23.68	0.136 ± 0.08	0.16 ± 0.03
A01HoriRUpβMD	avg. 106.93 ± 11.67	0.097 ± 0.05	0.11 ± 0.03
A01ΔDownβMD	mod. 109.08 ± 26.41	0.179 ± 0.08	0.16 ± 0.03

Isomer B

The RMSD time-series for isomer B, the second most populated isomer of hydroxymethyl-bullvalene, binding to βCD suggest that all four orientations are stable (Figure 2ii. 13, second row). The angle time-series and histogram, however, reveal that in all cases, the hydroxy-methyl bullvalene has rotated rapidly from its initial orientation to an essentially horizontal orientation with orientation angle of ~100° (Figure 2ii. 13, third row and bottom row, left). These most populated binding orientations are also shown in Figure 2ii. 13.

For B01ΔUpβMD, the bullvalene rotates so that the -CH_2OH group is mostly oriented upward, toward the smaller rim of βCD during the simulation time, so that bullvalene can stay deeper inside the cavity compared to the B01ΔUpβ initial orientation (with the cyclopropane ring upward). In B01HoriRUpβMD, the initial position and orientation of bullvalene was such that it stays slightly lower than the centre of βCD. During the MD simulation, the bullvalene turns such that it fits more into the cavity of the βCD, which is again a horizontal orientation, with the -CH_2OH group towards the bigger rim of βCD. However, for B01HoriRDownβMD, the bullvalene spends most time in a horizontal orientation with the substituent oriented upward. Lastly, in the simulation of B01ΔDownβMD, the bullvalene rotated slightly so that it remained horizontal but placed the substituent closer to the lower OH groups of βCD.

These reorientations of the bullvalene are likely due to the unfavourability of the hydrophilic hydroxymethyl substituent being inside the hydrophobic interior of βCD, which makes the -CH_2OH group tend to be drawn towards the outer edges of βCD, and the capacity for hydrogen bond formation between the -CH_2OH group and the OH groups of the upper and lower rims of βCD.

Because there were no step transitions in the angle time-series, the average value of the orientation angles and the average COM distance were used to select the most favourable coordinates for free energy calculations (Table 2ii. 11).

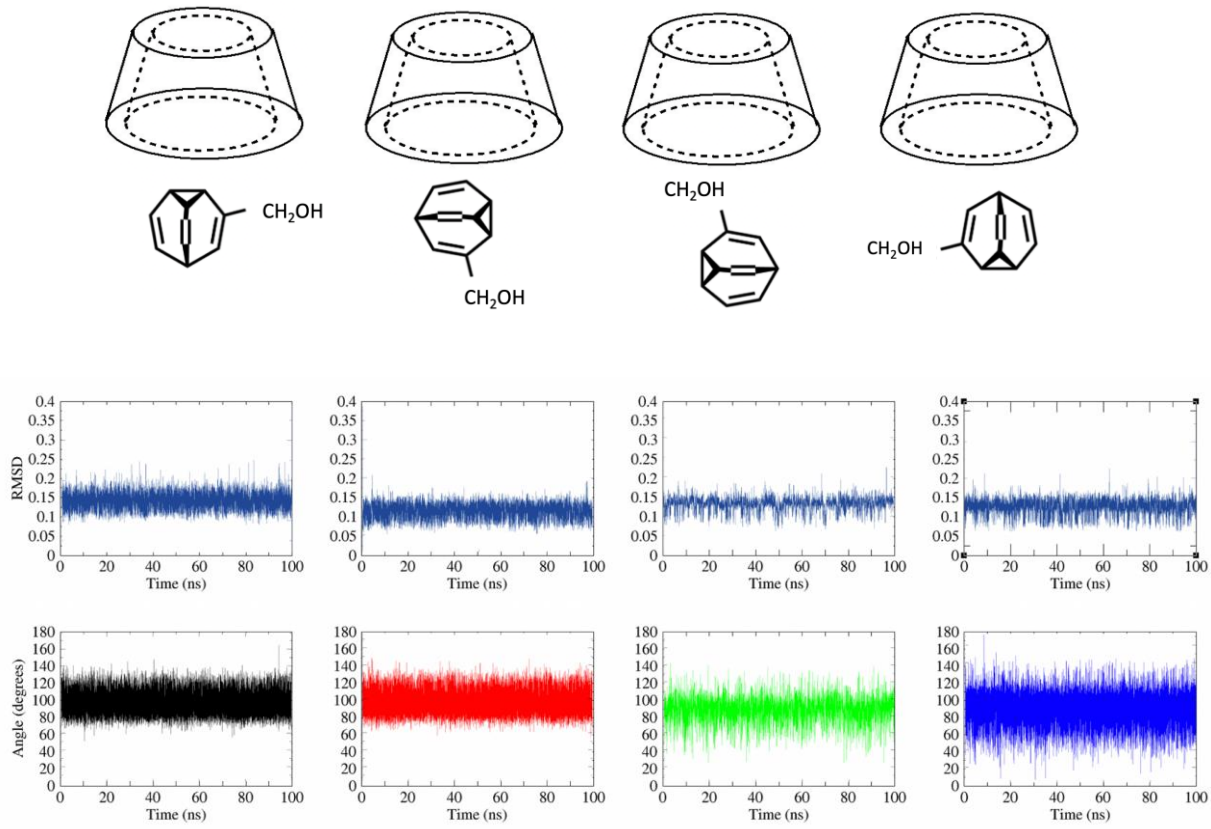


Figure 2ii. 13. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left) and most populated orientation angles (bottom row, right) of isomer B of hydroxymethyl-bullvalene in complex with βCD during 100 ns MD simulations initiated from four different orientations. The two structures in row four are of the two most populated binding modes of hydroxymethyl-bullvalene, indicated by horizontal lines in the angle time-series that match the border colours. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

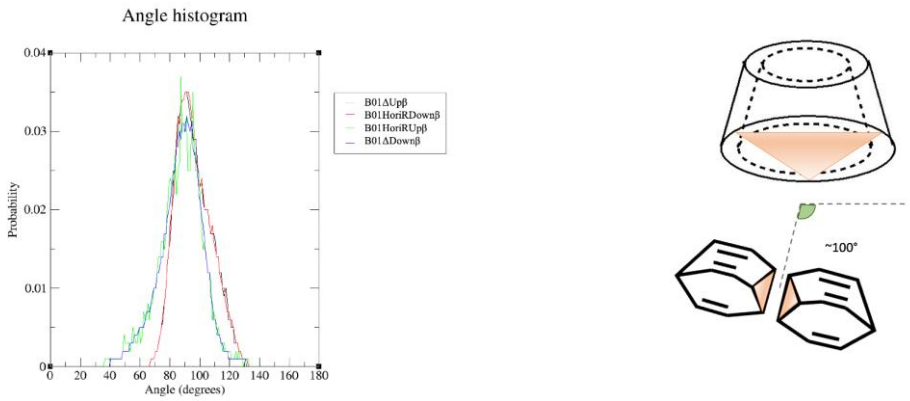


Figure 2ii. 13. (Continued). RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left) and most populated orientation angles (bottom row, right) of isomer B of hydroxymethyl-bullvalene in complex with βCD during 100 ns MD simulations initiated from four different orientations. The two structures in row four are of the two most populated binding modes of hydroxymethyl-bullvalene, indicated by horizontal lines in the angle time-series that match the border colours. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

Table 2ii. 11. Average values and standard deviations of the angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for B01ΔUpβMD, B01HoriRDownβMD, B01HoriRUpβMD, and B01ΔDownβMD.

Complex code name	**Angle (°)** (between the two planes)	**COM Dist. (nm)** (between COM of bullvalene and βCD)	**RMSD (nm)**
B01ΔUpβMD	avg. 96.03 ± 12.19	0.089 ± 0.04	0.15 ± 0.02
B01HoriRDownβMD	avg. 95.95 ± 12.20	0.088 ± 0.04	0.11 ± 0.02
B01HoriRUpβMD	avg. 87.43 ± 15.03	0.221 ± 0.05	0.13 ± 0.02
B01ΔDownβMD	avg. 87.49 ± 14.89	0.221 ± 0.05	0.12 ± 0.02

Isomer C

Isomer C behaves somewhat similarly to isomer A in complex with βCD, in that there are two major binding modes with orientation angles of 70° and 110/120° (Figure 2ii. 14). C01ΔUpβ rotates rapidly into an orientation similar to the stable initial orientation of C01HoriRDownβMD, with average orientation angles of ~110° and ~120°, respectively. In this binding mode, the substituent points towards the larger rim of βCD (Figure 2ii. 15), with the remainder of the bullvalene deep within the hydrophobic cavity. It also allows formation of hydrogen bonds between the -CH_2OH group of hydroxymethyl-bullvalene and the hydroxy groups in βCD.

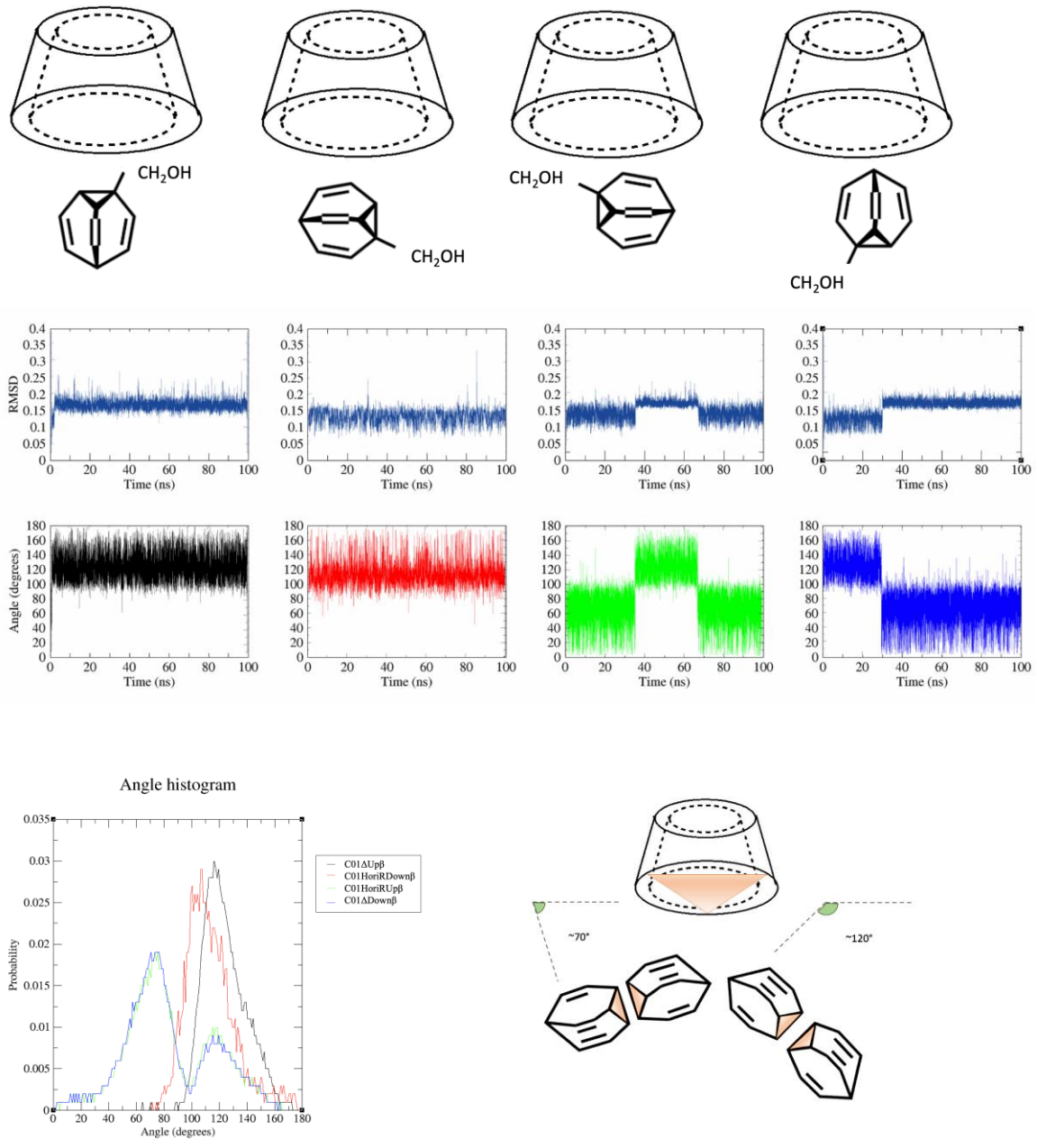


Figure 2ii. 14. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left) and most populated orientation angles (bottom row, right) of isomer C of hydroxymethyl-bullvalene in complex with βCD during 100 ns MD simulations initiated from four different orientations. The two structures in row four are of the two most populated binding modes of hydroxymethyl-bullvalene, indicated by horizontal lines in the angle time-series that match the border colours. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

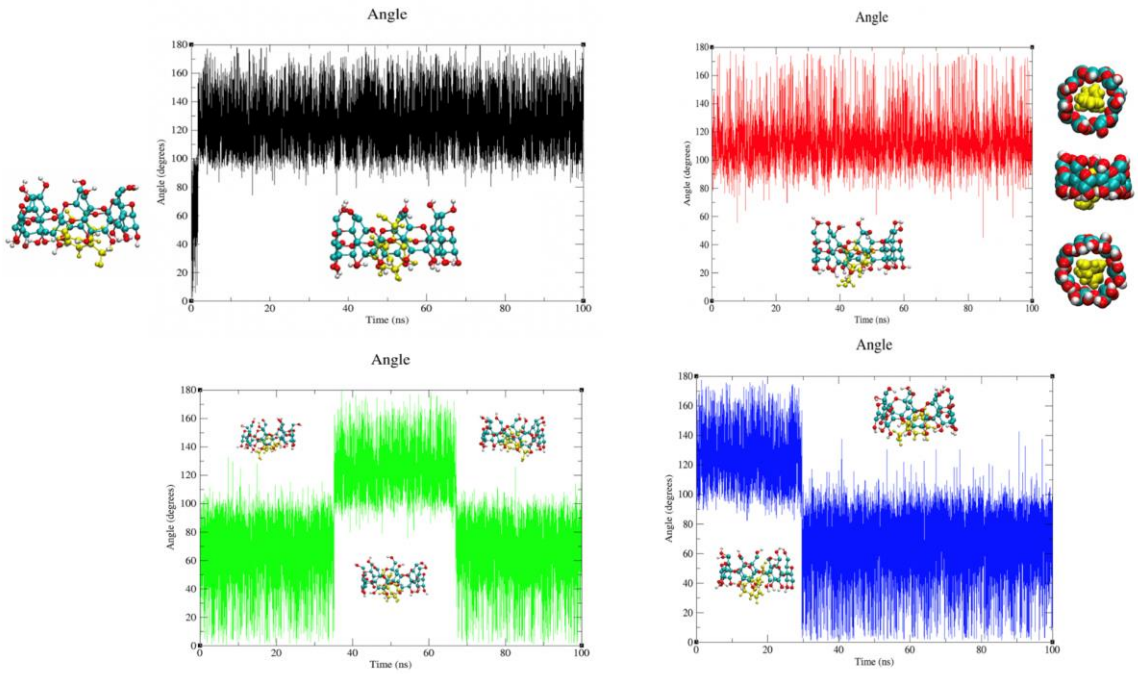


Figure 2ii. 15. Orientation angle time-series and snapshots of the hydroxymethyl-bullvalene-βCD complexes corresponding to the most populated orientation and (top left) the initial orientation during the 100 ns MD simulations initiated from the (black) C01ΔUpβ, (red) C01HoriβRDown, (green) C01HoriβRUp, and (blue) C01ΔDownβ orientations.

For both C01HoriRUpβMD and C01ΔUpβMD, the most stable binding mode had an orientation angle of 70° (Figure 2ii. 14, and Figure 2ii. 15), which corresponds to a horizontal orientation with the substituent pointing downwards towards the larger rim of βCD (Figure 2ii. 16), although both simulations also sampled the horizontal orientation with the substituent pointing upwards (similar to the initial configuration of C01ΔUpβMD). The advantage of the orientation with substituent pointing downwards could be that the hydrophobic part of hydroxymethyl-bullvalene fits inside the hydrophobic cavity of cyclodextrin and the -OH group of the substituent may form hydrogen bonds with the -OH groups of the larger rim of cyclodextrin.

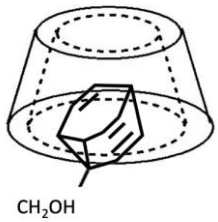


Figure 2ii. 16. The most populated binding mode for complexation in C01HoriRUpβMD and C01ΔUpβMD.

For C01ΔUpβMD and C01HoriRDownβMD, the average values of the orientation angle were used to select coordinates for free energy calculations, while for C01HoriRUpβMD and C01ΔDownβMD, the modal values were chosen; in all cases, the average of the COM distance was used (Table 2ii. 12).

Table 2ii. 12. Average/mode values and standard deviations of the angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for C01ΔUpβMD, C01HoriRDownβMD, C01HoriRUpβMD, and C01ΔDownβMD.

Complex code name	**Angle (°)** (between the two planes)	**COM Dist. (nm)** (between COM of bullvalene and βCD)	**RMSD (nm)**
C01ΔUpβMD	avg. 123.24 ± 16.90	0.106 ± 0.06	0.16 ± 0.01
C01HoriRDownβMD	avg. 113.76 ± 18.17	0.239 ± 0.06	0.13 ± 0.02
C01HoriRUpβMD	mod. 76.19 ± 32.07	0.196 ± 0.09	0.14 ± 0.02
C01ΔDownβMD	mod. 74.09 ± 32.17	0.199 ± 0.08	0.15 ± 0.03

Isomer D

For the least populated isomer of hydroxymethyl-bullvalene, isomer D, only three binding modes were studied due to the substituent being located on the apex (Figure 2ii. 17). There are again two major binding modes, with orientation angles of 25° and 140° degrees, which are offset from a horizontal orientation by 25° and 40°, respectively. D01UpβMD samples mostly the 140° orientation and C01HoriβMD samples mostly the 25°, but D01DownβMD switches between the two (Figure 2ii. 18). The 25° orientation directs the hydroxymethyl substituent towards the upper rim of βCD, where it forms hydrogen bonds (Figure 2ii. 19). The 140° orientation which directs the hydroxymethyl substituent towards the lower rim of βCD, is more favourable than it residing in the hydrophobic cavity but does not present hydrogen-bonding opportunities.

Modal values for angles were used to select the coordinates for free energy calculations, as well as the averages of the COM distances between bullvalene and βCD (Table 2ii. 13).

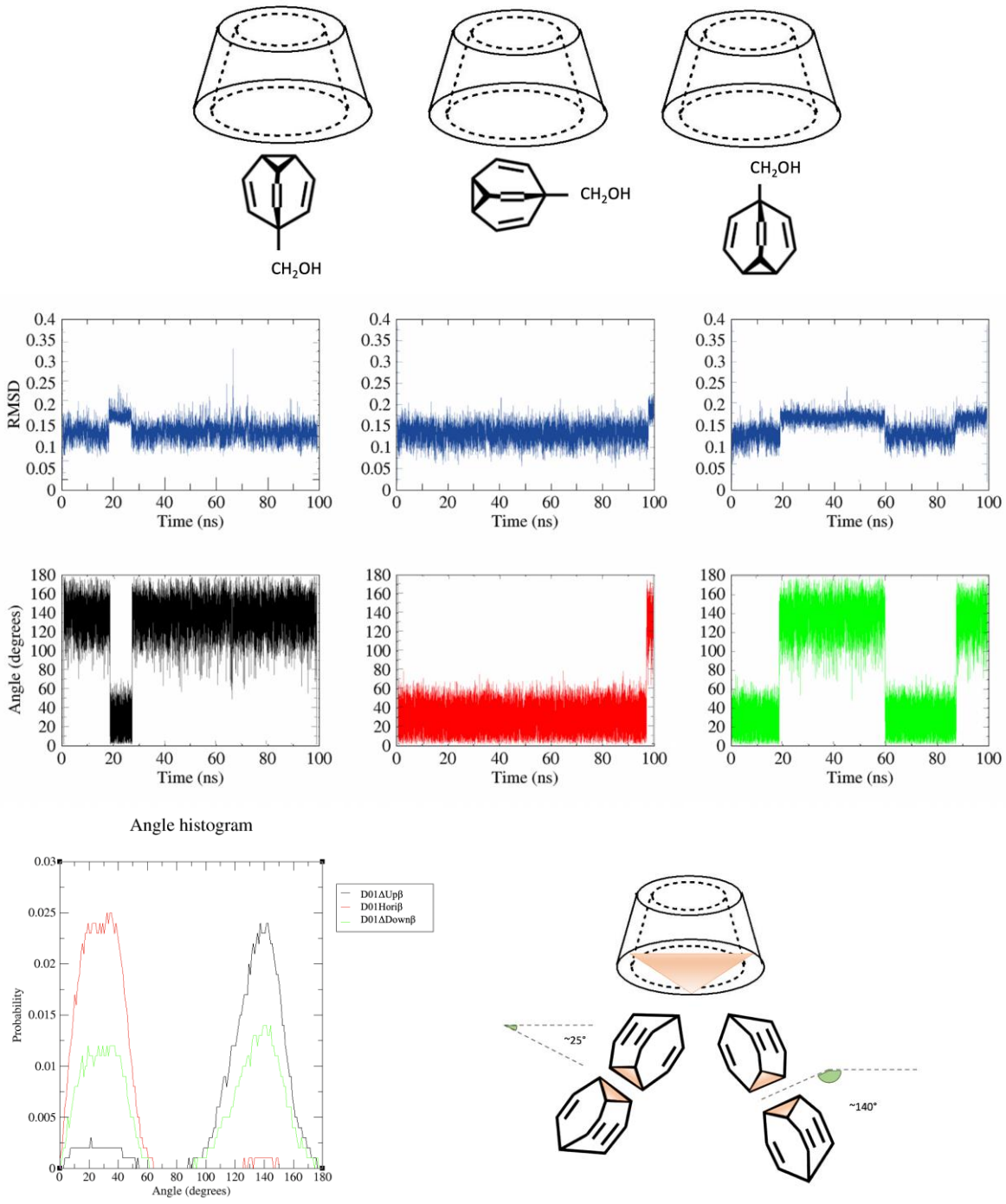


Figure 2ii. 17. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left) and most populated orientation angles (bottom row, right) of isomer D of hydroxymethyl-bullvalene in complex with βCD during 100 ns MD simulations initiated from four different orientations. The two structures in row four are of the two most populated binding modes of hydroxymethyl-bullvalene, indicated by horizontal lines in the angle time-series that match the border colours. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

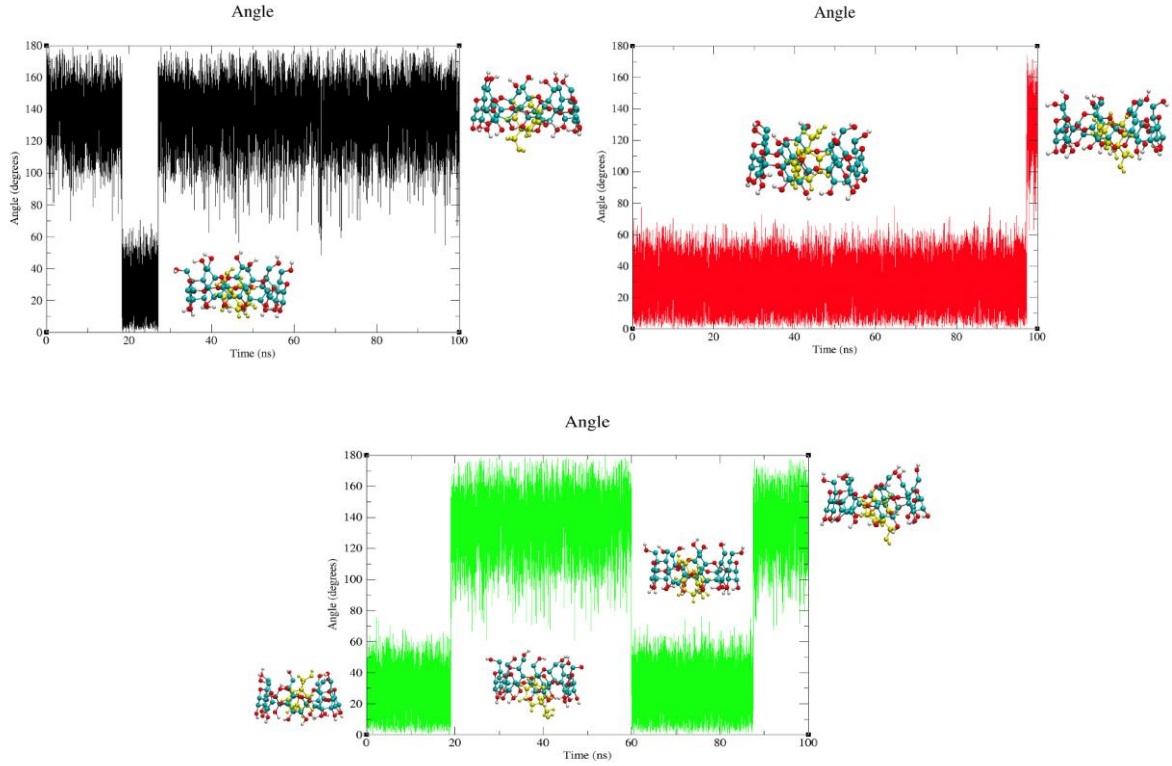


Figure 2ii. 18. Orientation angle time-series and snapshots of the hydroxymethyl-bullvalene-βCD complexes corresponding to the most populated orientation and (top left) the initial orientation during the 100 ns MD simulations initiated from the (black) D01ΔUpβ, (red) D01Horiβ and (green) D01ΔDownβ orientations.

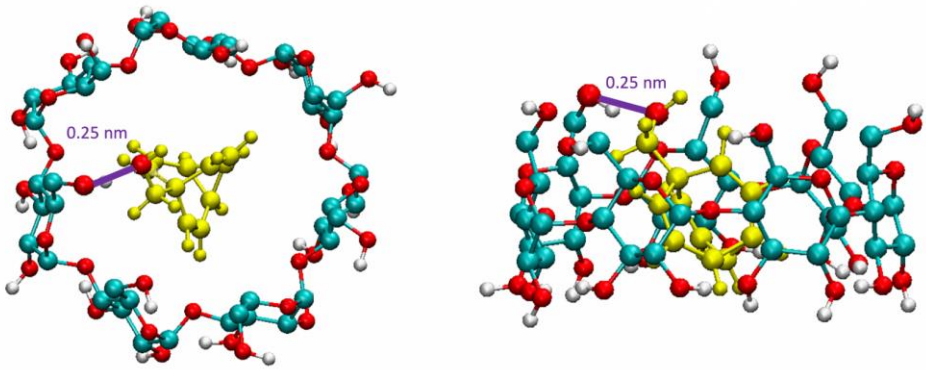


Figure 2ii. 19. One example of hydrogen bond formation (purple line) between the OH group of hydroxymethyl-bullvalene and the OH groups of the upper rim of βCD in the D01Horiβ complex.

Table 2ii. 13. Modal values and standard deviations of the angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for D01ΔUpβMD, D01HoriβMD, and D01ΔDownβMD.

Complex code name	**Angle (°)** (between the two planes)	**COM Dist. (nm)** (between COM of bullvalene and βCD)	**RMSD (nm)**
D01ΔUpβMD	mod. 137.78 ± 34.55	0.212 ± 0.08	0.12 ± 0.02
D01HoriβMD	mod. 24.43 ± 21.33	0.130 ± 0.06	0.13 ± 0.02
D01ΔDownβMD	mod. 134.46 ± 55.22	0.176 ± 0.08	0.14 ± 0.03

2ii.3.2.3. Calculation of free energies for binding of hydroxymethyl-bullvanene to βCD

The MD simulations revealed that, similarly to unsubstituted bullvalene, hydroxymethyl-bullvalene also tends to converge towards a limited number of stable binding modes. Unlike for unsubstituted bullvalene, however, there remain four different isomers to consider. Based on the results of the MD simulations, a set of seven binding modes were chosen for calculation of binding free energies: two for each of isomers A, B and C and one for the least favourable isomer, D. For isomers that sample multiple angles, the free energies for both angles were calculated to assess the effect of angle on the energies. However, in some cases, what appears to be two different angle values actually represent the same or very similar orientations of bullvalene with respect to cyclodextrin; in these cases, the free energy was only calculated for one orientation. For each selected orientation angle, those with inter-COM distances that matched the average distance were chosen for the free energy calculations. The calculated and experimental populations of each isomer in solution and the corresponding calculated solvation free energies and dipole moments are listed in Table 2ii. 14, and the selected binding modes and corresponding calculated binding free energies as well as the experimental binding free energy, which is a weighted average over all populated isomers, are given in Table 2ii. 15.

According to the calculated ΔG_{solv} values, isomer B is most soluble in water. The addition of the hydroxymethyl substituent increases the overall dipole moment, and isomer B has the largest dipole moment. This isomer is therefore likely to gain the least by binding to the hydrophobic interior of βCD. Interestingly, however, the calculated ΔG_{bind} that best matches the experimental value is that of isomer B, suggesting that it does indeed bind to βCD.

Moreover, while the experimental value may be a weighted average of the binding free energies of all four isomers, the extremely close correspondence between the calculated ΔG_{bind} value for isomer B, especially for the B01HoriRDownβMD orientation, and the experimental

value, as well as the lower ΔG_{bind} values for the other isomers, suggest that it is largely, if not entirely, isomer B that is captured by βCD.

Table 2ii. 14. Population of each isomer of hydroxymethyl-bullvalene determined experimentally, by NMR titration, and computationally, by QM calculations (14, 212); solvation free energy calculated using thermodynamic integration; dipole moment calculated from 100 ns unrestrained MD simulation in water.

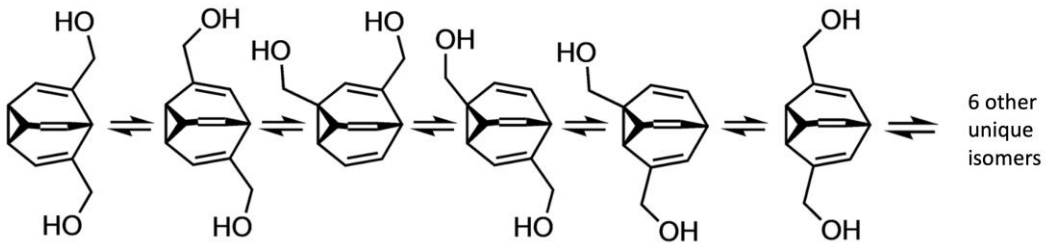
Isomer	Population (%)		ΔG_{solv} (kJ·mol^{-1})	Dipole moment (Debye)
	Experimental	Computational		
A	57	54	–34.04 ± 0.24	2.30
B	25	43	–40.96 ± 0.15	3.38
C	18	3	–34.59 ± 0.13	2.90
D	0	1	–36.14 ± 0.07	2.74

Table 2ii. 15. Calculated and experimental (O. Yahiaoui, Pers. Commun.) free energy values for binding of hydroxymethyl-bullvalene to βCD. The calculated value was determined at 298 K and the experimental value at 293.15 K.

Complex Code	ΔG_{bind}^{calc} (kJ·mol^{-1})	ΔG_{bind}^{exp} (kJ·mol^{-1})
A01HoriRDownβMD	-17.35 ± 0.43	-19.34 ± 1.74
A01HoriRUpβMD	-10.98 ± 2.32	
B01HoriRDownβMD	-19.46 ± 1.70	
B01ΔDownβMD	-19.94 ± 0.44	
C01HoriRDownβMD	-15.22 ± 1.26	
C01HoriRUpβMD	-15.26 ± 0.54	
D01HoriβMD	-14.01 ± 0.37	

2ii.3.3. Di-substituted dihydroxymethyl-bullvalene and β-cyclodextrin

Having gained a promising result for the prediction of the binding mode and the isomer of mono-substituted bullvalene to βCD, the procedure was applied to an even more complicated case, di-substituted dihydroxymethyl-bullvalene. Di-hydroxymethyl-bullvalene has 12 unique isomers, which six of them are shown in Figure 2ii. 20. Because of the much greater population of the most common isomer compared to the others, particularly when evaluated experimentally, and the combinatorial explosion of situations should all 12 isomers be studied, only the binding of the most populated isomer to βCD was investigated.



Calculated %	44	23	14	9	6	5
Experimental %	91	7	0	2	0	0

Figure 2ii. 20. The six most likely isomers of di-hydroxymethyl-bullvalene and their relative populations determined using QM calculations (14) and experimentally, by NMR titration (14).

2ii.3.3.1. Identification of most favourable modes for binding of dihydroxymethyl-bullvalene to β-cyclodextrin

The HGBPEP method was used to identify the most favourable binding modes of the most populated isomer of dihydroxymethyl-bullvalene in four different orientations (Figure 2ii. 21). The presence of two substituents make the patterns of more and less favourable binding modes more complicated and asymmetric again than for monosubstituted bullvalene, especially on the van der Waals heat maps, although overall seven-fold symmetry is retained. While the van der Waals energies suggest that dihydroxymethyl-bullvalene could fit inside the cavity of βCD cavity, the very broad region of unfavourable Coulombic energies rules this out. The binding locations with favourable Coulombic potential energy are within the lower part of the βCD cavity, and, for 11HoriRUpβ and 11ΔDownβ, at the outer edge of the narrower rim. For all four orientations, the combined non-bonded potential energy heat maps show that the most favourable binding location is between $z = -0.3$ to -0.4 nm, and thus complex coordinates were selected from here for MD simulation, using the same procedure as described in Section 2ii.3.1. (Table 2ii. 16).

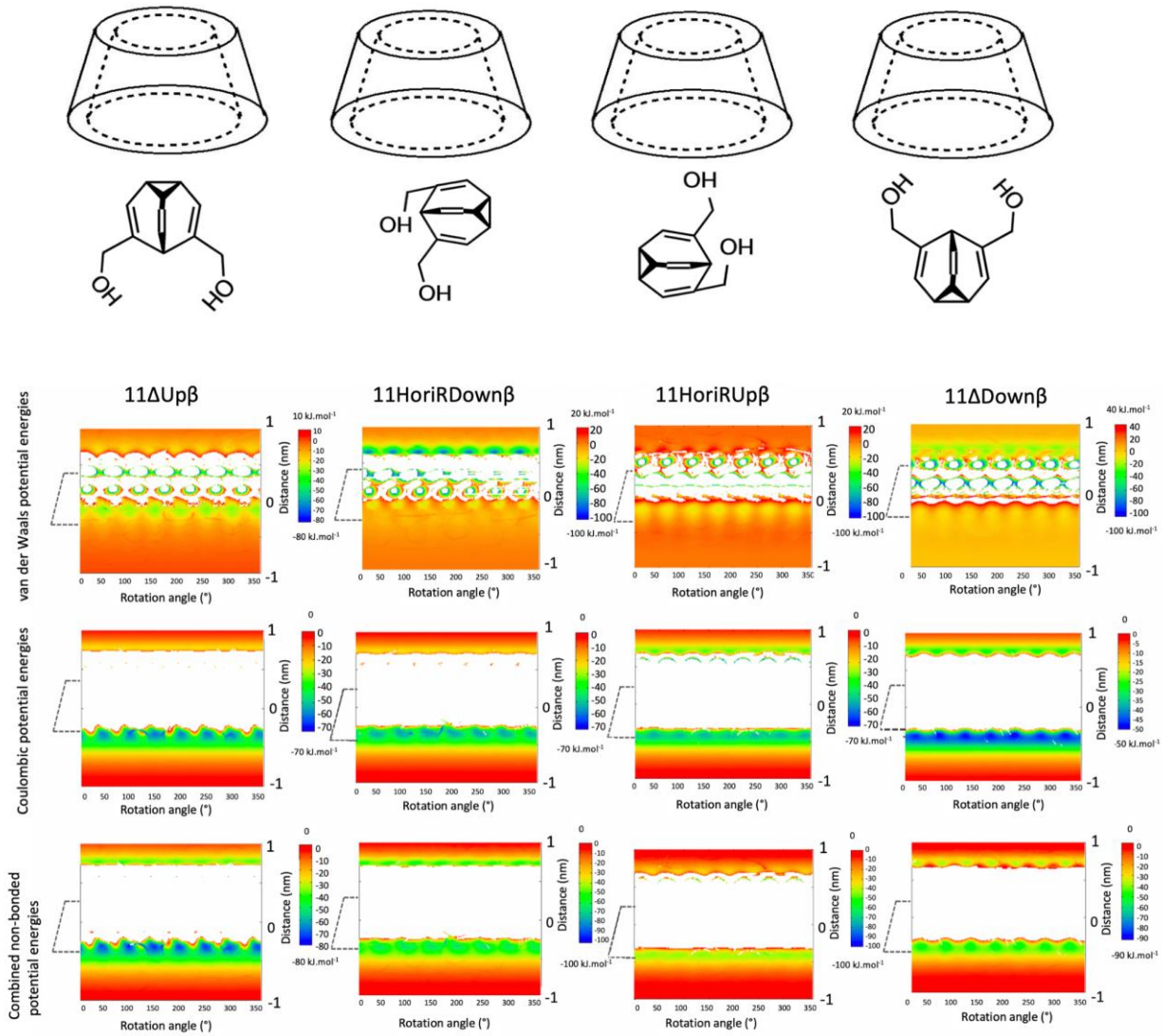


Figure 2ii. 21. HGBPE profiles for the (left) 11ΔUpβ, (centre left) 11HoriRDownβ, (centre right) 11HoriRUpβ and (right) 11ΔDownβ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

Table 2ii. 16. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for 11ΔUpβ, 11HoriRDownβ, 11HoriUpβ, and 11ΔDownβ.

Complex code name	Pos. (nm) Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
11ΔUpβ	-0.30	100	-79.34
11HoriRDownβ	-0.30	169	-76.22
11HoriUpβ	-0.41	267	-48.48
11ΔDownβ	-0.40	67	-51.96

2ii.3.3.2. Stability of favourable modes for binding of dihydroxymethyl-bullvalene to βCD

The results of the MD simulations initiated from each of the four different orientations (Figure 2ii. 22) show two dominant orientations for this isomer of di-hydroxymethyl-bullvalene, regardless of the initial orientation, with both lying between a vertical and horizontal orientation (angle ≅ 70° or 105°, (note that 180 − 70 = 110, if measuring from the other direction)). In one, the bullvalene sits in the lower part of the βCD cavity and the two substituents point downwards, towards the solvent (cyan box in Figure 2ii. 22), and in the other, one points towards the upper rim and the other towards the lower rim (pink box, Figure 2ii. 22). The former is the more populated orientation, and was therefore chosen for the binding free energy calculations. Coordinates for this binding mode were selected from 11HoriRDownβMD (Table 2ii. 17) by selecting the time frame with orientation angle and COM distance values that matched the averages of these properties. The binding free energy was also computed for coordinates taken from 11ΔUpβMD (where bullvalene was in the most favourable binding mode, according to the average value of the angles and COM distances, Table 2ii. 17), however, since the two substituent groups also point downward in this orientation.

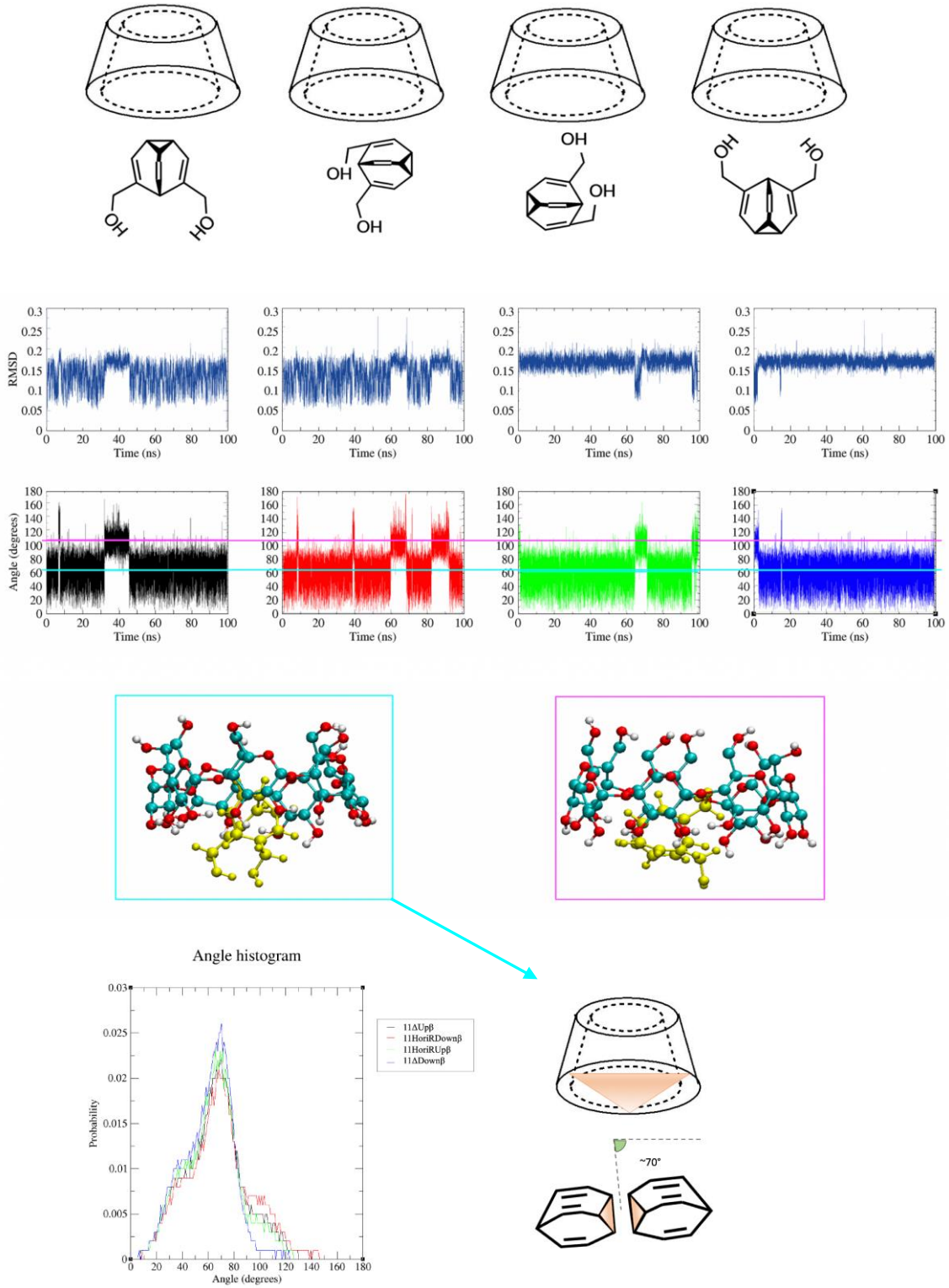


Figure 2ii. 22. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left) and most populated orientation angles (bottom row, right) of the most populated isomer of dihydroxymethyl-bullvalene in complex with βCD during 100 ns MD simulations initiated from four different orientations. The two structures in row four are of the two most populated binding modes of hydroxymethyl-bullvalene, indicated by horizontal lines in the angle time-series that match the border colours. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

Table 2ii. 17. Average values and standard deviations of angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for 11HoriRDownβMD, 11ΔupβMD, 11HoriUpβMD, and 11ΔDownβMD.

Complex code name	**Angle (°)** (between the two planes)	**COM Dist. (nm)** (between COM of bullvalene and βCD)	**RMSD (nm)**
11ΔUpβMD	avg. 65.73 ± 23.80	0.269 ± 0.05	0.14 ± 0.03
11HoriRDownβMD	avg. 68.94 ± 26.04	0.269 ± 0.06	0.13 ± 0.03
11HoriUpβMD	avg. 63.25 ± 22.31	0.271 ± 0.05	0.16 ± 0.02
11ΔDownβMD	avg. 60.16 ± 19.74	0.272 ± 0.05	0.17 ± 0.02

2ii.3.3.3. Calculation of free energies for binding of dihydroxymethyl-bullvanene to βCD

The solvation free energy of dihydroxymethyl-bullvalene is larger and more favourable than those of unsubstituted and monosubstituted bullvalene. This is reasonable given that it contains two polar hydroxyl groups compared to none or one. The 11ΔUpβMD binding mode has a slightly more favourable binding free energy, and appears to fit better in the cavity of βCD, compared to the 11HoriRDownβMD binding mode, but the difference between the two binding free energy values is not large given the standard deviation. Unfortunately experimental results were ultimately not provided by our collaborators, and so it is not possible to draw any further conclusions regarding whether the computational approach developed here was successful in the case of disubstituted bullvalene.

Table 2ii. 18. Solvation free energy and free energy of binding of dihydroxymethyl-bullvalene to βCD calculated using thermodynamic integration.

Complex Code	ΔG_{solv}^{calc} **(kJ·mol^{-1})**	ΔG_{bind}^{calc} **(kJ·mol^{-1})**
11ΔUpβMD	-45.35 ± 0.29	-15.04 ± 0.61
11HoriRDownβMD		-14.44 ± 0.70

2ii.3.4. Unsubstituted bullvalene and γ-cyclodextrin

While stable modes for binding of unsubstituted, monosubstituted and disubstituted bullvalene to βCD were identified, the increased size of monosubstituted and disubstituted bullvalene make it difficult for the bullvalene to fit inside the βCD cavity. To determine whether more favourable and stable complexation might occur with a larger cavity, the binding of all three bullvalenes to γCD was investigated. βCD is built from seven glucose subunits, whereas γCD is built from eight, increasing the mean diameter (between anomeric oxygen atoms) from 0.70 to 0.88 nm (24).

2ii.3.4.1. Identification of most favourable modes for bullvalene binding to γCD

The results of carrying out HGBPEP for unsubstituted bullvalene with γCD show that the larger cavity of γCD allows bullvalene to fit inside. According to the van der Waals energy maps (Figure 2ii. 23), 00Horiγ has the broadest favourable binding region. The Coulombic potential energy maps display a very low energy area in the middle of γCD for the all three orientations of bullvalene, however, and this dominates the smaller van der Waals energies so that the total potential energy maps also show favourable binding in the centre of γCD for all three orientations. There are some minor differences between the orientations, with the most negative overall energy value for host-guest complexation of –80 $kJ \cdot mol^{-1}$ occurring for the horizontal bullvalene interacting with γCD, whereas for the other two orientations it is –70 $kJ \cdot mol^{-1}$. There is no eight-fold symmetry visible, which is another result of the bigger size of the cavity of γCD. Coordinates for MD simulation were selected from the dark blue energetically favourable band from the combined potential energy plots at the orientation angle values given in Table 2ii. 19 according to the same procedure as explained in Section 2ii.3.1.

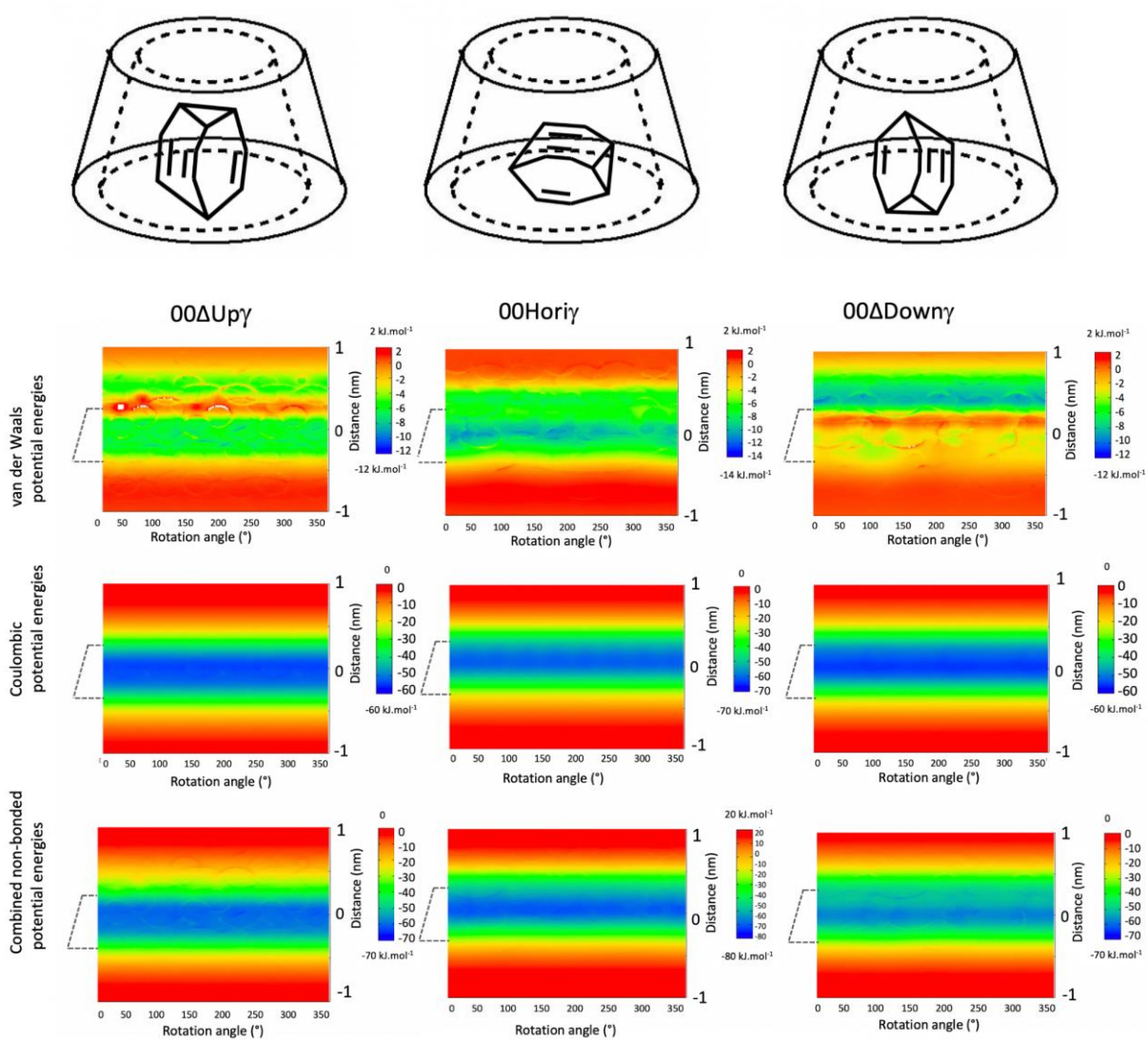


Figure 2ii. 23. HGBPE profiles for the (left) 00ΔUpγ, (centre) 00Horiγ, and (right) 00ΔDownγ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

Table 2ii. 19. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for 00ΔUpγ, 00Horiγ, and 00ΔDownγ.

Complex code name	Pos. Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
00ΔUpγ	0	215	-62.45
00Horiγ	0.10	165	-69.74
00ΔDownγ	0	120	-58.50

2ii.3.4.2. Stability of favourable modes for binding of unsubstituted bullvalene to γCD

The MD simulation results, in particular the orientation angle time-series and histogram, show that bullvalene has more freedom to move and rotate when in complex with γCD compared to with βCD (Figure 2ii. 24). Overall, it tends to bind in a more horizontal than vertical

orientation, preferring orientation angles of ~70° and ~120°. While the larger size of the γCD cavity means that the bullvalene can be entirely encapsulated, the additional space also means that γCD does not remain symmetrical during the MD simulation, at times collapsing around the bullvalene (Figure 2ii. 25).

To select coordinates from which to initiate binding free energy calculations, the modal values of the orientation angle and the average values of the distance between the COM of bullvalene and γCD were calculated, and the time-frames of the MD trajectories with these properties were selected (Table 2ii. 20).

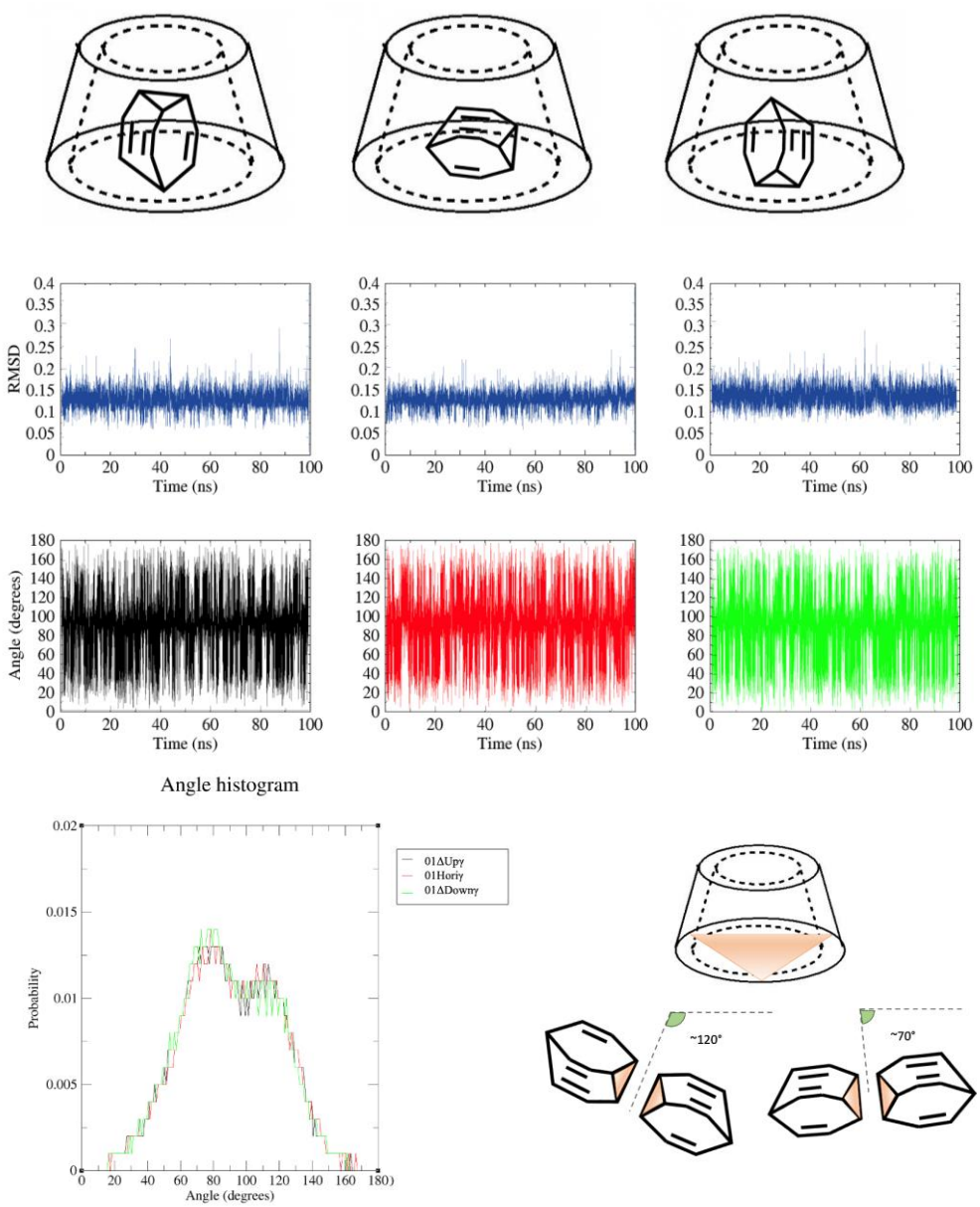


Figure 2ii. 24. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left) and most populated orientation angles (bottom row, right) of unsubstituted bullvalene in complex with γCD during 100 ns MD simulations initiated from three different orientations. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

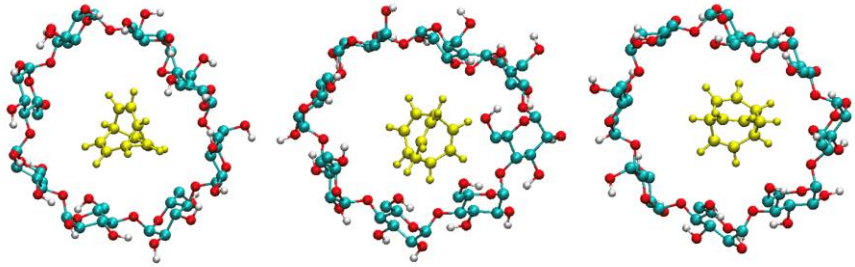


Figure 2ii. 25. Three snapshots from the MD simulation of the 00Horiγ. Bullvalene is in yellow and the atoms of γCD are coloured according to type: carbon is cyan; oxygen is red; and hydrogen is white.

Table 2ii. 20. Modal values and standard deviations of angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for 00ΔUpγMD, 00HoriγMD, and 00ΔDownγMD.

Complex code name	**Angle (°)** (between the two planes)	**Avg. Dist.** (COM of bullvalene and γCD)	**RMSD (nm)**
00ΔUpγMD	mod. 84.015 ± 29.38	0.114 ± 0.06	0.13 ± 0.02
00HoriγMD	mod. 77.991 ± 31.01	0.091 ± 1.23	0.12 ± 0.02
00ΔDownγMD	mod. 87.85 ± 33.16	1.430 ± 1.88	0.12 ± 0.03

2ii.3.4.3. Calculation of free energies for binding of unsubstituted bullvalene to γCD

The binding orientation identified as the most stable was 00HoriγMD, therefore the binding affinity for unsubstituted bullvalene in this orientation to γCD, was computed using the previously-calculated solvation free energy of unsubstituted bullvalene in water (Table 2ii. 21). Comparing the binding free energy values of unsubstituted bullvalene binding to βCD (−10.99 ± 0.51 kJ·mol^{-1}) and γCD (−11.00 ± 0.93 kJ·mol^{-1}), it can be concluded that unsubstituted bullvalene binds equally well to β- and γCD. Again, no experimental results were available for the binding of unsubstituted bullvalene to γCD.

Table 2ii. 21. Solvation free energy and free energy of binding of unsubstituted bullvalene to γCD calculated using thermodynamic integration.

Complex Code	**ΔG_{solv}^{solv} (kJ·mol^{-1})**	**ΔG_{bind}^{calc} (kJ·mol^{-1})**
00HoriγMD	−11.00 ± 0.93	−24.76 ± 0.17

2ii.3.5. Mono-substituted hydroxymethyl-bullvalene and γ-cyclodextrin

As for binding of monosubstituted bullvalene to βCD, the four distinguishable isomers of monosubstituted hydroxymethyl-bullvalene (A-D) in complexation with γCD were studied.

2ii.3.5.1. Identification of most favourable modes for hydroxymethyl-bullvalene binding to γCD

Isomer A

The increased size of monosubstituted bullvalene results in the potential energy maps reflecting the eight-fold symmetry of γCD (Figure 2ii. 26). It is still possible, however, for hydroxymethyl-bullvalene to bind to the interior of γCD; indeed, this is the most favourable binding zone according to the Coulombic energy heat maps. The overall potential energy heat maps reflect features from the van der Waals and Coulombic energy heat maps more than was observed for other bullvalene-cyclodextrin combinations. Interestingly, there are two series of favourable binding modes at different vertical positions in the combined energy plots of A01HoriRDownγ and A01ΔUpγ, both of which feature the substituent pointing downward. One of these binding modes positions the hydrophobic core of the bullvalene at the centre of the cavity with the hydrophilic OH group of the substituent interacting with solvent through the lower rim of the cyclodextrin. The other places the bullvalene just below (A01ΔUpγ) or at (A01HoriRDownγ) the narrow upper rim of the cyclodextrin. Coordinates for MD simulation were chosen from the blue zones of the combined non-bonded potential energy plots for each orientation, using the same method as in Section 2ii.3.1, for all isomers (Table 2ii. 22).

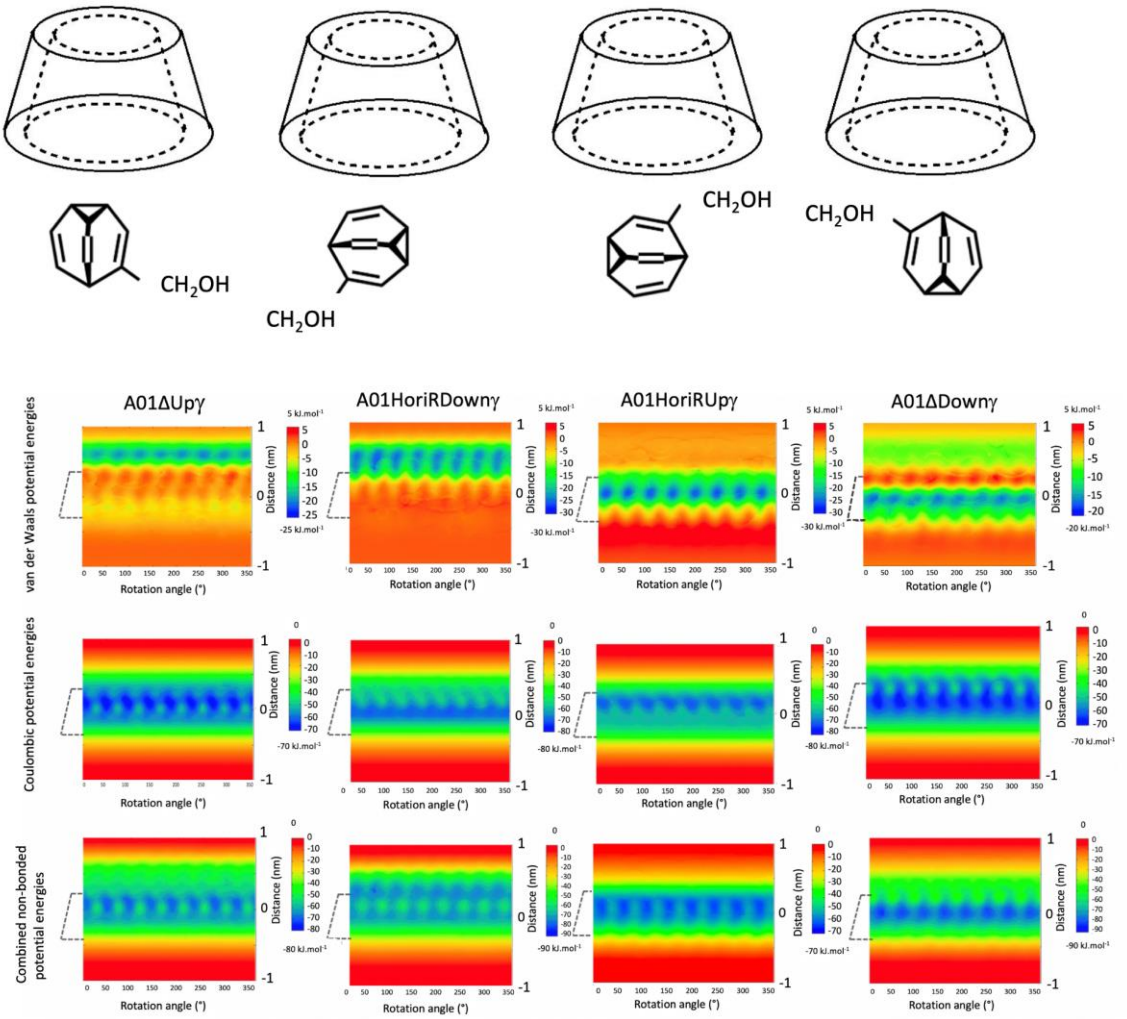


Figure 2ii. 26. HGBPE profiles for the (left) A01ΔUpγ, (centre left) A01HoriRDownγ, (centre right) A01HoriRUpγ and (right) A01ΔDownγ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

Table 2ii. 22. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for A01ΔUpγ, A01HoriRDownγ, A01HoriRUpγ, and A01ΔDownγ.

Complex code name	Pos. Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
A01ΔUpγ	0.10 0.60	100 260	-68.77 -43.36
A01HoriRDownγ	0.30 0	55 55	-73.51 -71.22
A01HoriRUpγ	0.10	105	-92.45
A01ΔDownγ	0	20	-80.88

Isomer B

For isomer B, not only is eight-fold symmetry again apparent, but there are now regions in which binding of hydroxymethyl-bullvalene to γCD is extremely unfavourable in terms of the Coulombic, and, for B01ΔUpγ and B01ΔDownγ, the van der Waals potential energy too (Figure 2ii. 27). The latter represents overlap between bullvalene and γCD atoms, whereas the latter is due to repulsion of like charges on bullvalene and γCD when the bullvalene is in the centre of the cavity. According to the overall non-bonded potential energy maps, the most favourable binding modes of B01ΔUpγ and B01HoriRUpγ occur when bullvalene is closer to the upper rim of γCD. This is because in these two orientations, the substituent points upward and while approaching γCD from the bottom, the CH_2OH group has unfavourable interactions with the OH groups of the lower rim of γCD. At the upper rim, in contrast, the bullvalene can be encapsulated in the hydrophobic core of γCD and the hydroxymethyl groups are directed towards the solvent with which they can form hydrogen bonds. Similarly, for B01HoriRDownγ and B01ΔDownγ, the most favourable binding positions are at about $z = –0.1$ to –0.2 nm, which means bullvalene is located inside the cavity near the bigger rim of γCD, and the hydrophilic CH_2OH group points downward, towards the solvent rather than the hydrophobic cavity of γCD. Since bullvalene's carbon cage is non-polar itself, it favourably sits inside γCD. Coordinates were selected from the most favourable binding region for each orientation for MD simulation, using the same method as in Section 2ii.3.1 (Table 2ii. 23).

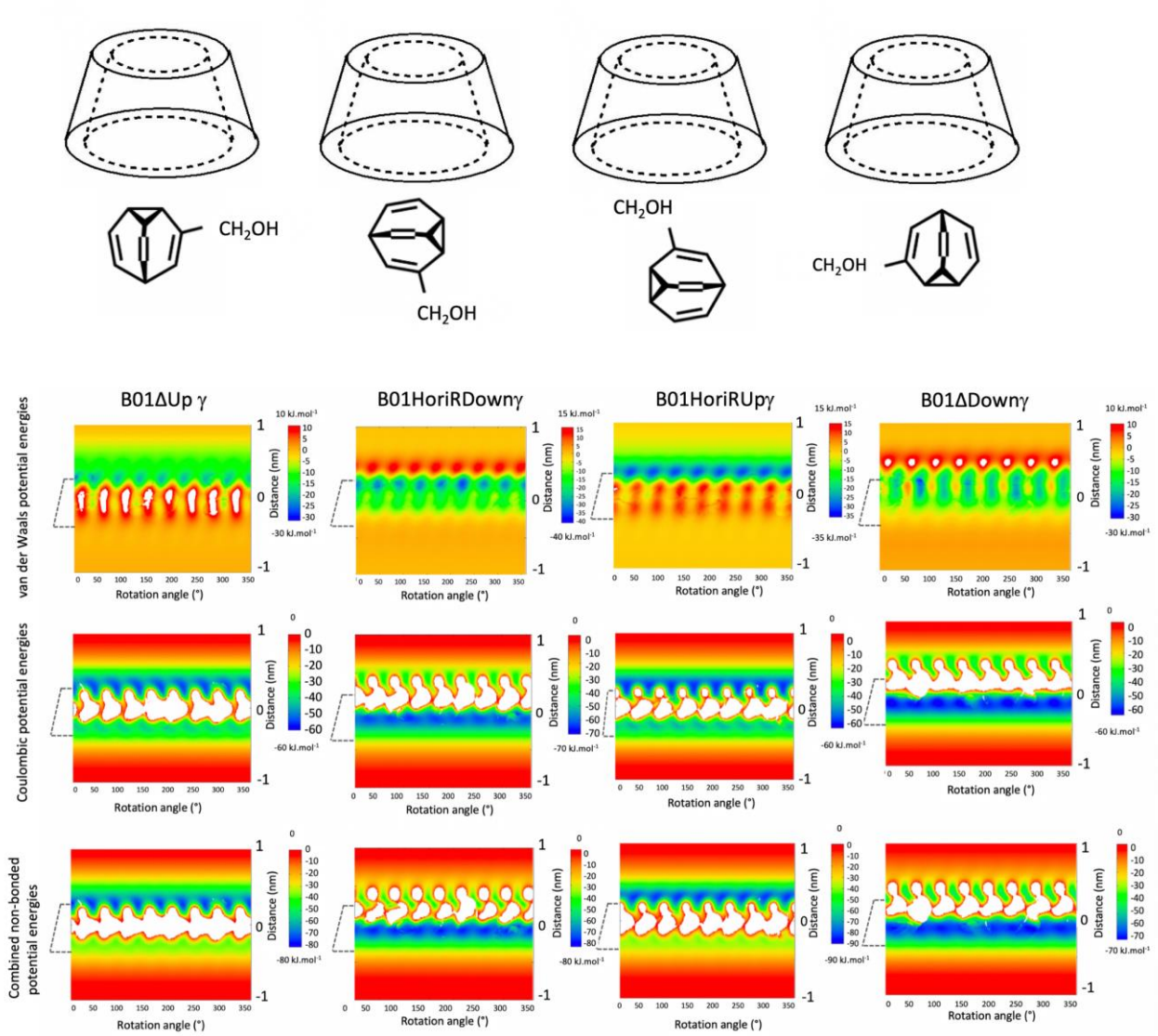


Figure 2ii. 27. HGBPE profiles for the (left) B01ΔUpγ, (centre left) B01HoriRDownγ, (centre right) B01HoriRUpγ and (right) B01ΔDownγ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

Table 2ii. 23. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for B01ΔUpγ, B01HoriRDownγ, B01HoriRUpγ, and B01ΔDownγ.

Complex code name	Pos. Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
B01ΔUpγ	0.40	354	-59.64
B01HoriRDownγ	-0.10	100	-75.08
B01HoriRUpγ	-0.30	73	-39.46
B01ΔDownγ	-0.12	273	-67.74

Isomer C

For this isomer, the HGBPEP results are a combination of those observed for isomers A and B (Figure 2ii. 28). The van der Waals energy maps show that for the C01ΔUpγ and C01HoriRUpγ complexes, the bullvalene binds most favourably just above and just below the centre of the γCD cavity, respectively, whereas for the C01HoriRDownγ and C01ΔDownγ complexes, the bullvalene binds most favourably at the upper rim of γCD. These preferred binding positions are largely driven by steric repulsion of the substituent, which points upwards for the first two complexes and downwards for the latter two complexes. The very unfavourable binding modes observed for C01ΔUpγ at z = –0.3 to –0.4 nm reflect positions and orientations in which the CH_2OH group of the bullvalene points upward and clashes with the OH groups of the lower rim of γCD. According to the Coulombic energy maps, it is favourable for the two vertically-orientated (C01ΔUpγ and C01ΔDownγ) bullvalenes to sit inside γCD. However, the horizontal orientations of this isomer have unfavourably high Coulombic potential energy values due to the like charges on the bullvalene substituent and the γCD getting too close while bullvalene is inside the cavity. In comparison to the βCD results, however, the unfavourable binding regions are not continuous, that is, specific binding orientations remain favourable. The eight-fold symmetry of γCD is also shown in the patterns in all plots. The overall combined potential energy plots confirm that the Coulombic energy is again dominant to the van der Waals energy. This leaves a large band of semi-continuously favourably binding positions for C01ΔUpγ and C01ΔDownγ, a narrow semi-continuous band of binding positions for C01HoriRUpγ, and two semi-continuous bands of favourable binding positions for C01HoriRDownγ. Coordinates were chosen from these regions to undergo further unbiased MD simulations, using the same method as Section 2ii.3.1 (Table 2ii. 24).

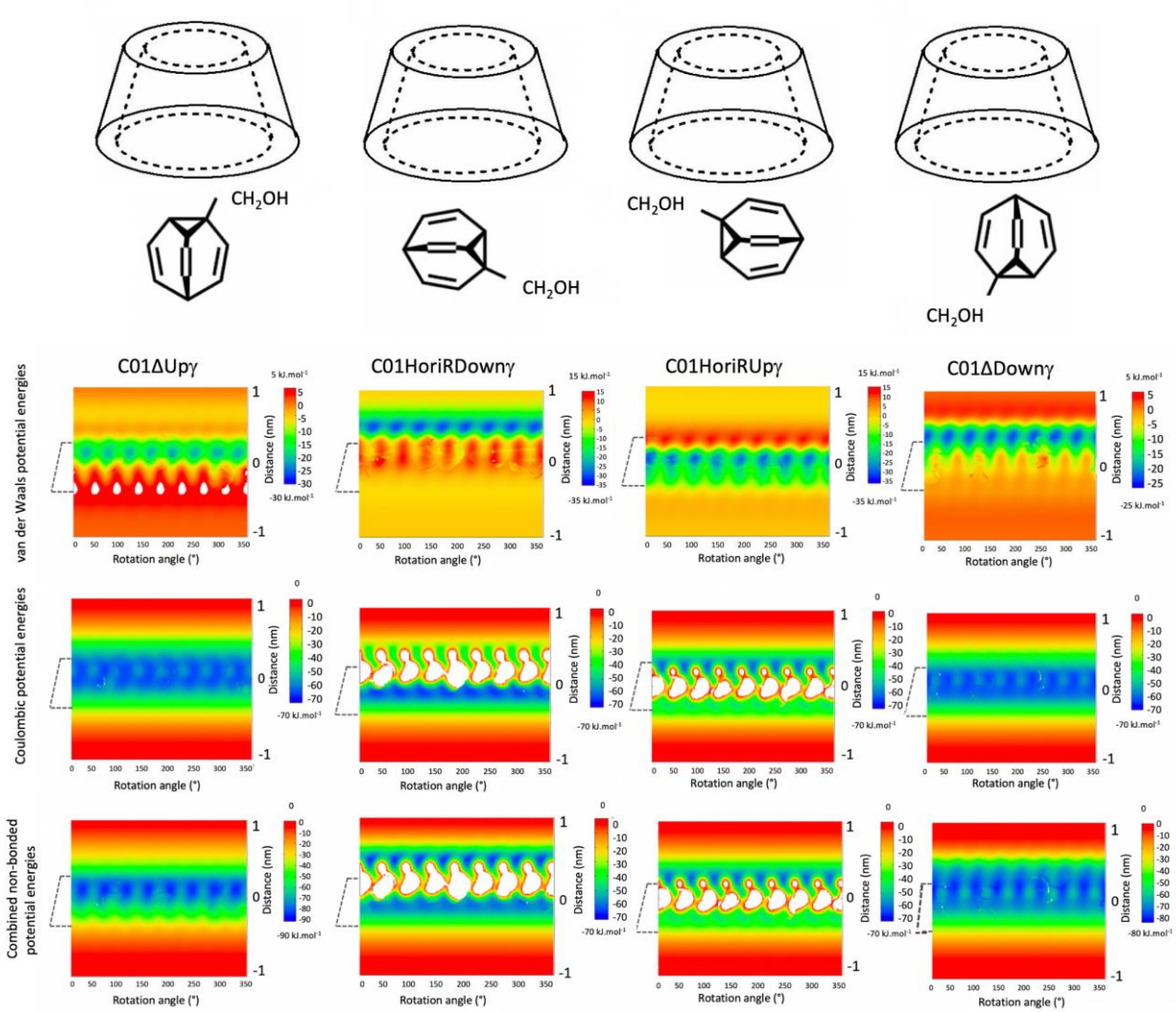


Figure 2ii. 28. HGBPE profiles for the (left) C01ΔUpγ, (centre left) C01HoriRDownγ, (centre right) C01HoriRUpγ and (right) C01ΔDownγ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

Table 2ii. 24. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for C01ΔUpγ, C01HoriRDownγ, 01HoriRUpγ, and C01ΔDownγ.

Complex code name	Pos. Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
C01ΔUpγ	0.10	335	-85.15
C01HoriRDownγ	-0.10	126	-60.81
C01HoriRUpγ	0.25	334	-55.76
C01ΔDownγ	0.20	136	-74.57

Isomer D

The least populated isomer of hydroxymethyl-bullvalene was also studied by HGBPE to predict its complexation with γCD (Figure 2ii. 29). Since the substituent is located on the apex

carbon in this isomer, the horizontal orientation is larger than the width of even the γCD cavity, and this is confirmed by the potential energy maps for D01Horiγ. The other two orientations can favourably sit inside the cavity, and their energy maps do not even show eight-fold symmetry, reminiscent of unsubstituted bullvalene. For these orientations, coordinates of bullvalene bound to the centre of the γCD cavity were selected for MD simulations. For D01Horiγ, there are two bands of favourable binding positions, a narrow and less favourable one at the lower rim and a more favourable one at the upper rim, from which coordinates were selected for MD simulation, the same way as Section 2ii.3.1 (Table 2ii. 25).

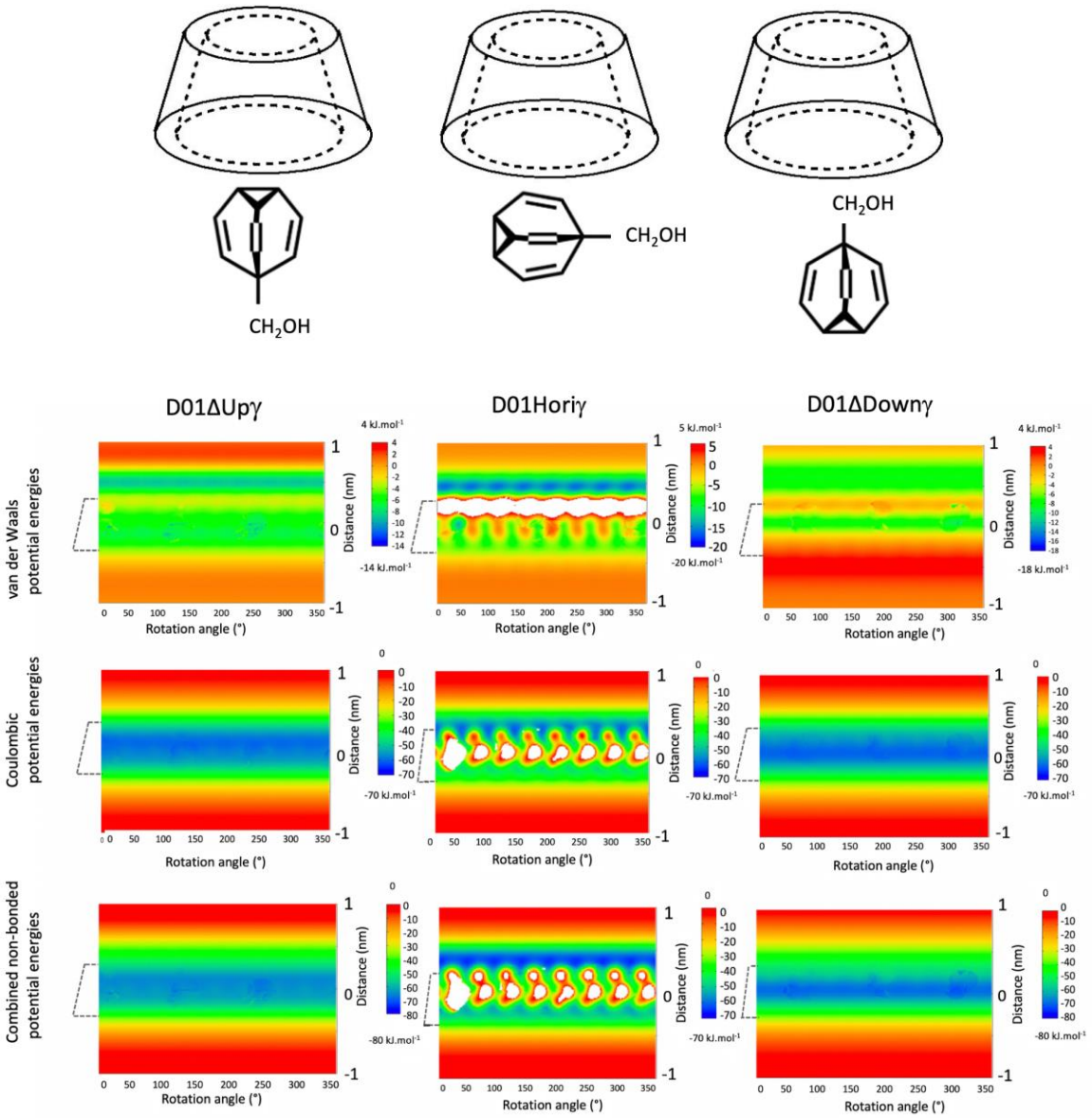


Figure 2ii. 29. HGBPE profiles for the (left) D01ΔUpγ, (centre) D01Horiγ and (right) D01ΔDownγ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies. Dashed lines show the position of cyclodextrin.

Table 2ii. 25. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for D01ΔUpγ, D01Horiγ, and D01ΔDownγ.

Complex code name	Pos. Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
D01ΔUpγ	-0.02	252	-68.02
D01Horiγ	0.40 -0.20	44 4	-64.15 -53.13
D01ΔDownγ	-0.02	49	-70.39

2ii.3.5.2. Stability of favourable modes for binding of hydroxymethyl-bullvalene to γCD

Isomer A

The MD simulations on each complex of isomer A with γCD show that regardless of the initial orientation, bullvalene mostly stays horizontal during all simulations, with an average orientation angle of 100° (Figure 2ii. 30). It does, however, undergo some rotation and movement, as was also observed for unsubstituted bullvalene in the larger γCD cavity. In the simulations of A01ΔUpγMD, A01HoriRUpγMD and A01ΔDownγMD, the substituent mostly points upward, while in A01HoriRDownγ, the substituent alternates between pointing towards the bigger and smaller rims of γCD, and the substituent is rarely inside the cavity. Although all isomers have the same angle distribution, this difference between the orientation of substituent may be because of the two possibilities of orientation for the one angle (Figure 2ii. 31). The average values of the orientation angle and the average values of the COM distance were used to select coordinates from each MD trajectory with favourable binding orientations from which to start the free energy calculations (Table 2ii. 26).

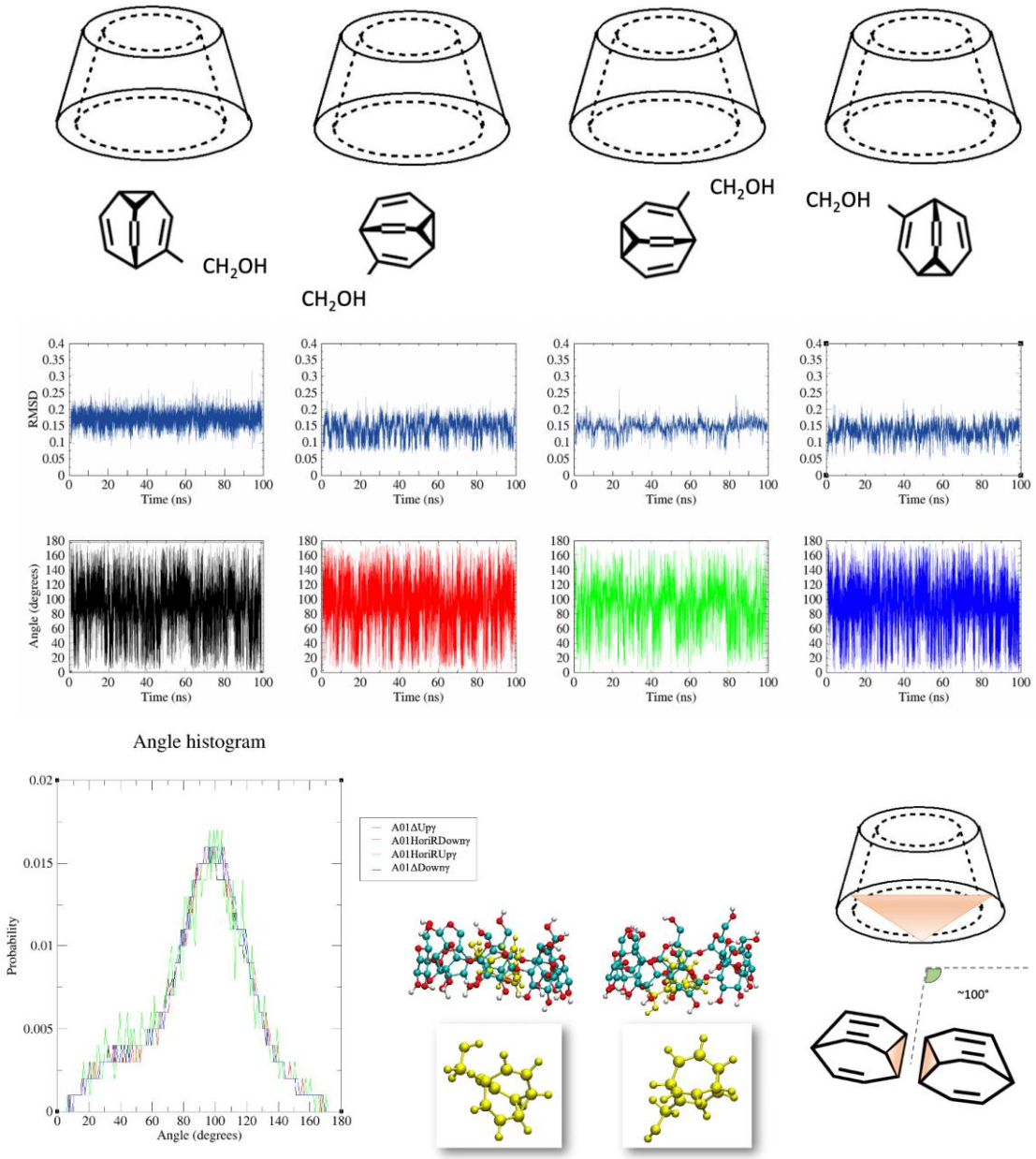


Figure 2ii. 30. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left), two example frames of when the plane angle is about 100°, and most populated orientation angles (bottom row, right) of isomer A of hydroxymethyl-bullvalene in complex with γCD during 100 ns MD simulations initiated from four different orientations. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

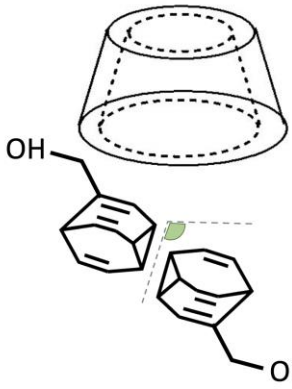


Figure 2ii. 31. The two possible orientation for one angle.

Table 2ii. 26. Average values and standard deviations of the angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for A01ΔUpγMD, A01HoriRDownγMD, A01HoriRUpγMD, and A01ΔDownγMD.

Complex code name	**Angle (°)** (between the two planes)	**COM Dist.** (between COM of bullvalene and γCD)	**RMSD (nm)**
A01ΔUpγMD	avg. 92.42 ± 31.23	0.110 ± 0.06	0.17 ± 0.02
A01HoriRDownγMD	avg. 93.086 ± 30.83	0.112 ± 0.06	0.14 ± 0.03
A01HoriRUpγMD	avg. 91.4897 ± 31.18	0.115 ± 0.06	0.15 ± 0.02
A01ΔDownγMD	avg. 92.437 ± 30.19	0.109 ± 0.06	0.14 ± 0.02

Isomer B

For isomer B in complex with γCD, the bullvalene is again highly mobile. In B01ΔUpγMD, B01HoriRDownγMD and B01ΔDownγMD, the -CH_2OH group predominantly points towards the γCD both rims, whereas in B01HoriRUpγMD, the substituent tends to point towards the hydroxyl groups of the upper rim of γCD (Figure 2ii. 32). Favourable binding coordinates were chosen to initiate binding free energy calculations based on the same procedure as for isomer A (Table 2ii. 27).

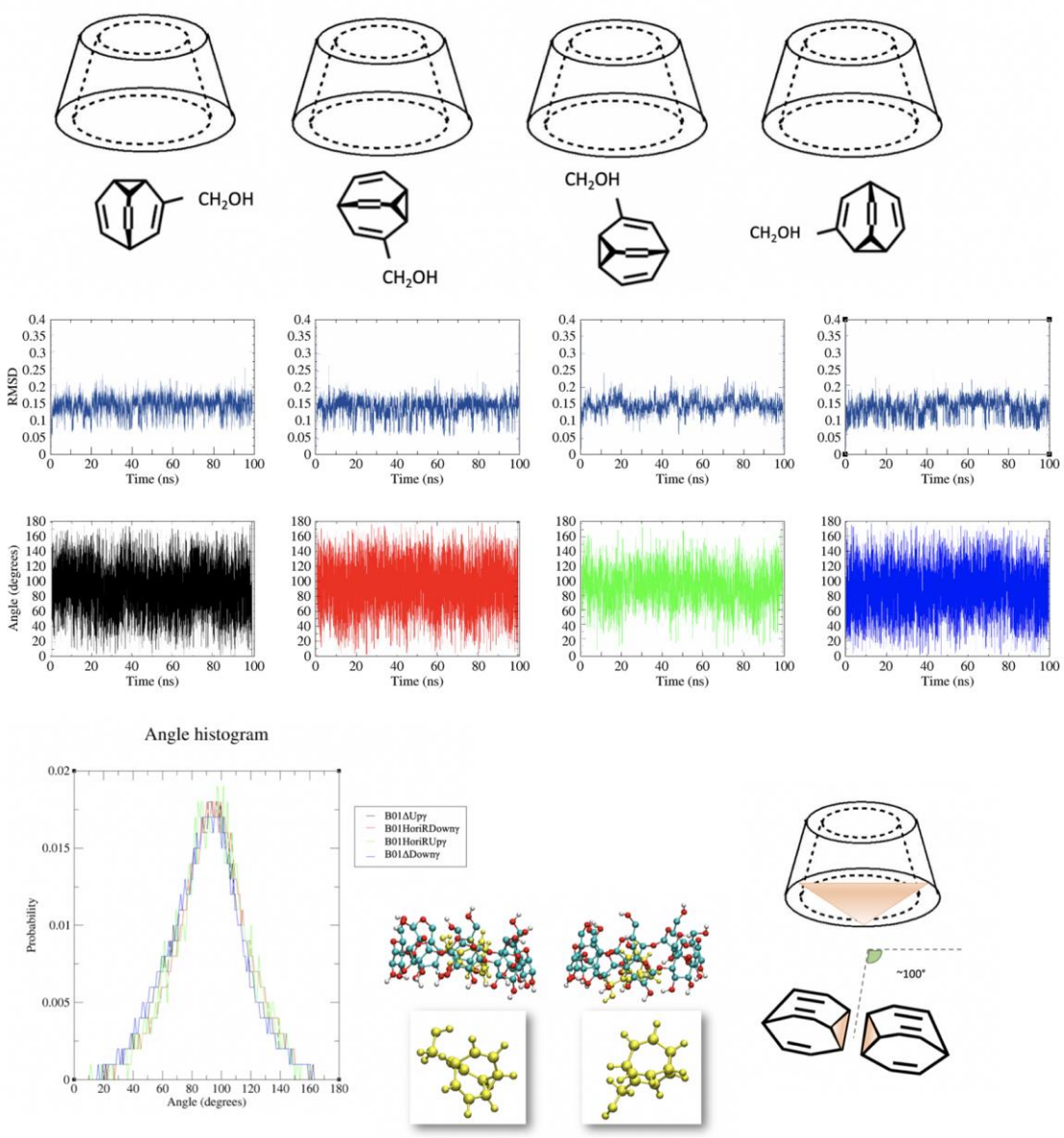


Figure 2ii. 32. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left), two example frames of when the plane angle is about 100°, and most populated orientation angles (bottom row, right) of isomer B of hydroxymethyl-bullvalene in complex with γCD during 100 ns MD simulations initiated from four different orientations. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

Table 2ii. 27. Average values and standard deviations of the angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for B01ΔUpγMD, B01HoriRDownγMD, B01HoriRUpγMD, and B01ΔDownγMD.

Complex code name	**Angle (°)** (between the two planes)	**COM Dist.** (between COM of bullvalene and γCD)	**RMSD (nm)**
B01ΔUpγMD	avg. 90.67 ± 25.80	0.120 ± 0.07	0.14 ± 0.02
B01HoriRDownγMD	avg. 93.95 ± 24.97	0.114 ± 0.06	0.14 ± 0.02
B01HoriRUpγMD	avg. 92.99 ± 25.23	0.114 ± 0..06	0.15 ± 0.02
B01ΔDownγMD	avg. 86.25 ± 26.08	0.123 ± 0.07	0.13 ± 0.03

Isomer C

Similar results were obtained for isomer C as for isomers A and B. This may be because of the large cavity of γCD, which allows rotation of bound hydroxymethyl-bullvalene around a preferred orientation of ~100° such that the substituent remains close to the two rims of γCD (Figure 2ii. 33). Therefore, again, the chosen orientation for the free energy calculation could be selected from the output coordinates of any of the four simulations, although for consistency, favourable coordinates were selected from each MD simulation trajectory according to the same procedure as for isomer A (Table 2ii. 28).

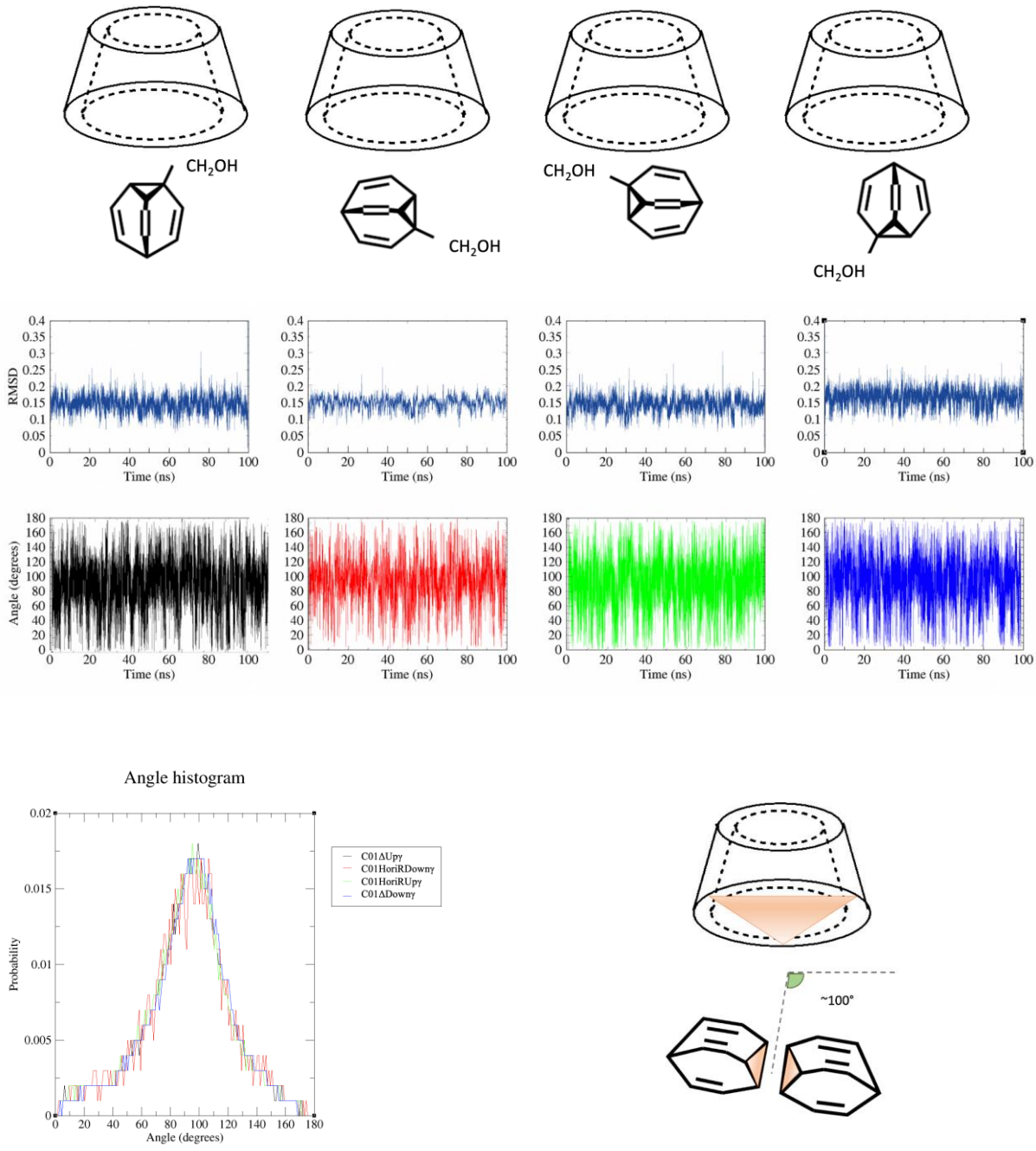


Figure 2ii. 33. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left), and most populated orientation angles (bottom row, right) of isomer B of hydroxymethyl-bullvalene in complex with γCD during 100 ns MD simulations initiated from four different orientations. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

Table 2ii. 28. Average values and standard deviations of the angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for C01ΔUpγMD, C01HoriRDownγMD, C01HoriRUpγMD, and C01ΔDownγMD.

Complex code name	**Angle (°)** (between the two planes)	**Avg. Dist.** (COM of bullvalene and γCD)	**RMSD (nm)**
C01ΔUpγMD	avg. 91.35 ± 30.83	0.113 ± 0.06	0.14 ± 0.02
C01HoriRDownγMD	avg. 91.33 ± 31.56	0.113 ± 0.06	0.15 ± 0.02
C01HoriRUpγ	avg. 90.80 ± 30.43	0.114 ± 0.07	0.14 ± 0.02
C01ΔDownγMD	avg. 92.77 ± 29.96	0.113 ± 0.06	0.16 ± 0.02

Isomer D

For the least populated isomer, D, the bullvalene leaves the γCD cavity for a brief period during the MD simulation of 11ΔUpγMD (Figure 2ii. 34). The bullvalene exits through the lower, larger rim of γCD and, after a brief period in the solvent where it moves around to the top of the γCD, it re-enters through the smaller, upper rim. This period of the simulation is illustrated in Figure 2ii. 35. Thus while it is favourable overall for the hydrophobic bullvalene to reside in the γCD cavity, the polar hydroxymethyl substituent means it can occasionally exist in solution. During the majority of the simulations when the bullvalene is within the γCD, it oscillates between two binding angles, ~50° and 140°. In both cases, the substituent points towards the upper rim of γCD and may form hydrogen bonds with the γCD OH groups. Favourable coordinates to start free energy calculations were chosen from the MD trajectories according to the modal values of the orientation angle and the average values of the COM distance between bullvalene and γCD (Table 2ii. 29).

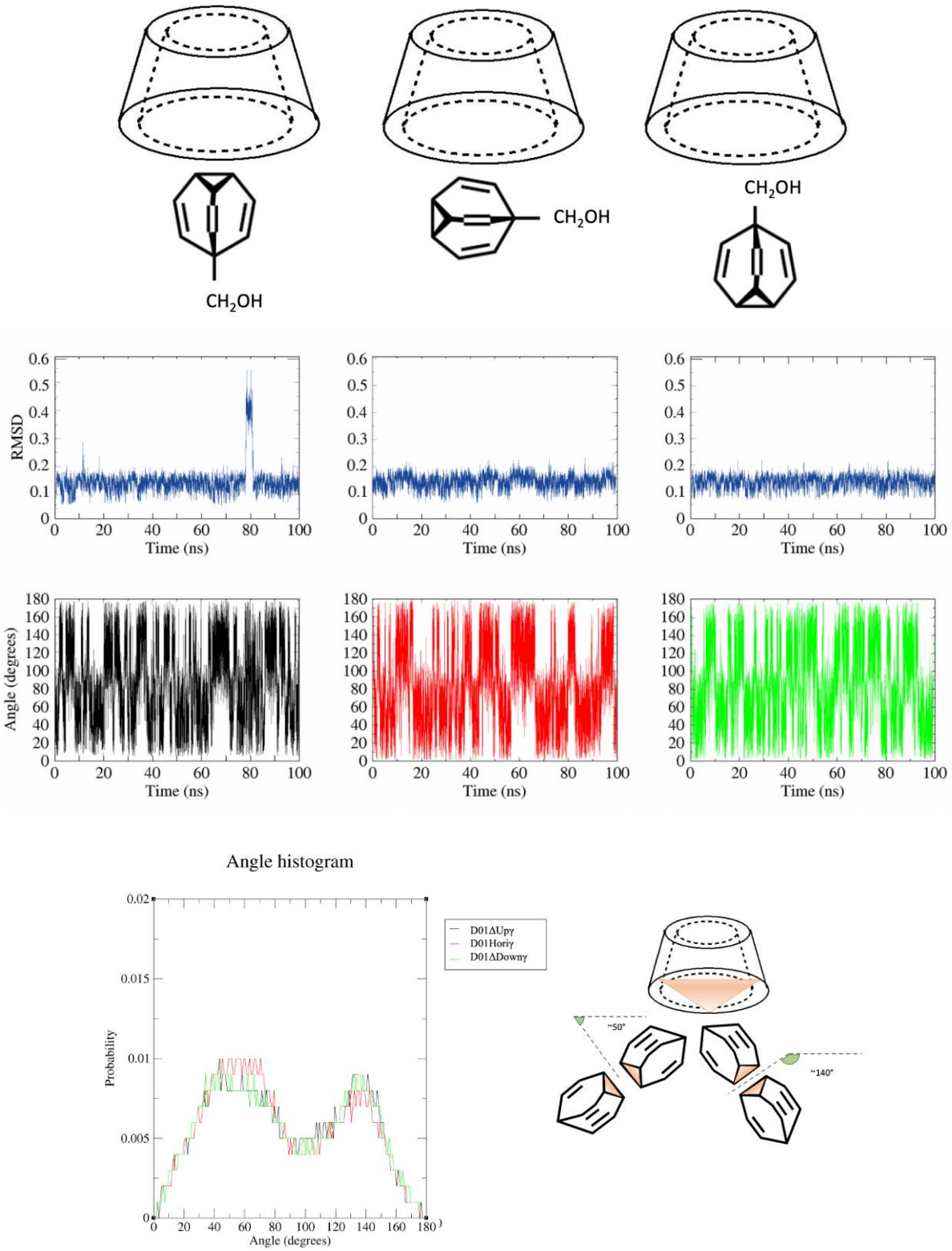


Figure 2ii. 34. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left), and most populated orientation angles (bottom row, right) of isomer B of hydroxymethyl-bullvalene in complex with γCD during 100 ns MD simulations initiated from four different orientations. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

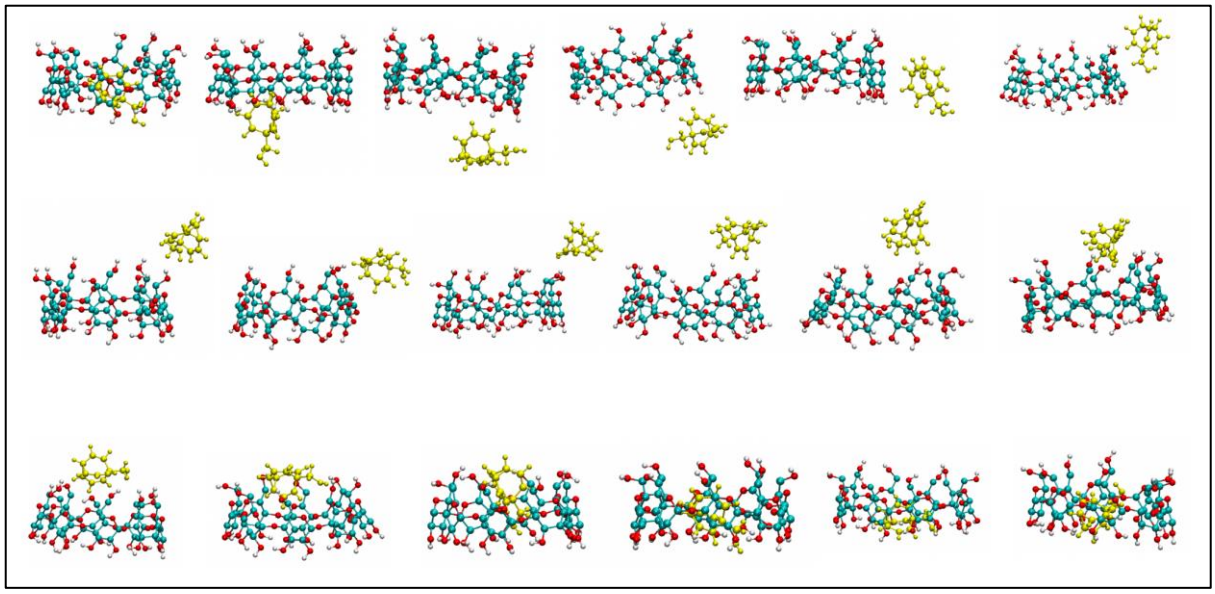


Figure 2ii. 35. Time-series of snapshots of hydroxymethyl-bullvalene and γCD, running from left to right across each row from top to bottom, of the period of the MD simulation of 11ΔUpγ around 80 ns during which the hydroxymethyl-bullvalene jumps out of the γCD cavity and comes back in again. Both molecules are drawn in CPK representation, with bullvalene in yellow and the atoms of γCD coloured according to type, with carbon in cyan, oxygen in red, and hydrogen in white.

Table 2ii. 29. Modal values and standard deviations of the angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for D01ΔUpγMD, D01HoriγMD, and D01ΔDownγMD.

Complex code name	**Angle (°)** (between the two planes)	**Avg. Dist.** (COM of bullvalene and γCD)	**RMSD (nm)**
D01ΔUpγMD	mod. 50.44 ± 43.69	0.142 ± 0.19	0.14 ± 0.05
D01HoriγMD	mod. 60.581 ± 42.95	0.109 ± 0.06	0.14 ± 0.02
D01ΔDownγMD	mod. 23.906 ± 44.27	0.105 ± 0.06	0.13 ± 0.02

2ii.3.5.3. Calculation of free energies for binding of hydroxymethyl-bullvalene to γCD

The solvation free energy values of hydroxymethyl-bullvalene in water are listed in Table 2ii. 30, as well as its binding free energy values in complex with γCD, for the orientations selected in the previous section. The negative values for ΔG_{bind} indicate that in all cases, complexation is favourable. Nevertheless, the numbers vary remarkably even for one particular isomer according to its binding orientation. The most negative binding free energy belongs to the most populated isomer in a horizontal orientation (A01HoriRDownγMD). However, in comparison with the ΔG_{bind} values for hydroxymethyl-bullvalene-βCD complexes, the γCD complexes have less negative binding free energies. It can be concluded that this mono-substituted bullvalene binds better to βCD than to γCD, despite the larger cavity of γCD allowing the

bullvalene to fit fully inside. This suggests that the larger size of the γCD cavity is in fact a hindrance, such that bullvalene is too small to exclude solvent molecules and thus has little advantage of being in the γCD cavity rather than in solution. This was confirmed by visual analysis which showed that there are indeed water molecules as well as bullvalene in the γCD cavity (Figure 2ii. 36).

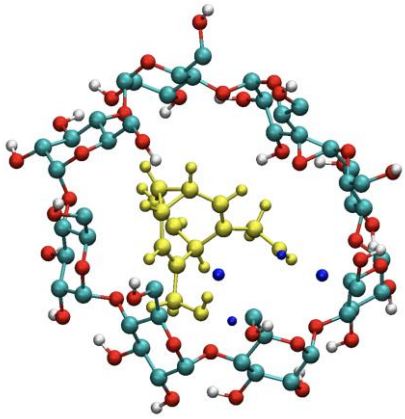


Figure 2ii. 36. Water molecules inside the cavity of γCD. 11ΔUpγMD was chosen as an example to show that even in the di-substituted bullvalene, which the guest is slightly bigger, still water molecules locate inside γCD during the simulation time.

Again, the experimental ΔG_{bind} values did n ot materialise. Comparing the calculated ΔG_{bind} values for the different isomers and orientations, the most favourable are for the A01HoriRDownγMD orientation of isomer A, and the B01ΔUpγMD and B01ΔDownγMD orientations of isomer B, followed by the D01ΔUpγMD and D01HoriγMD orientations of isomer D. This suggests that both isomer A, which is most populated in solution, and isomer B are likely to be bound by γCD, making it less useful for selecting a particular isomer. The much less favourable ΔG_{bind} value for isomer C suggests that this isomer does not bind to γCD. The greater variety of isomers and orientations with favourable binding free energies reflects the larger size of the γCD cavity, which makes it more permissive to binding substituted bullvalene but less selective. Additionally, the binding of monosubstituted bullvalene by γCD is weaker than βCD complexation.

Table 2ii. 30. Solvation free energy and free energy of binding of hydroxymethyl-bullvalene to γCD calculated using thermodynamic integration.

Complex Code	ΔG_{solv}^{calc} (kJ.mol^{-1})	ΔG_{bind}^{calc} (kJ.mol^{-1})
A01ΔUpγMD	-34.76 ± 0.24	-7.02 ± 0.93
A01HoriRDownγMD		-15.81 ± 1.68
B01ΔUpγMD	-40.96 ± 0.15	-12.80 ± 0.68
B01ΔDownγMD		-14.75 ± 1.43
C01HoriRDownγMD	-34.59 ± 0.13	-3.80 ± 2.42
D01ΔUpγMD	-36.14 ± 0.07	-10.72 ± 0.34
D01HoriγMD		-11.62 ± 1.31
D01ΔDownγMD		-7.6 ± 1.42

2ii.3.6. Disubstituted-bullvalene and γ-cyclodextrin

2ii.3.6.1. Identification of most favourable modes for dihydroxymethyl-bullvalene binding to γCD

The HGBPEP method was applied for four orientations of the most populated isomer of dihydroxymethyl-bullvalene in complex with γCD (Figure 2ii. 37). Interestingly, despite the large size of this disubstituted bullvalene, there are only a few highly unfavourable binding modes in the van der Waals energy maps. Only the 11HoriRDownγ van der Waals interaction energy map shows a narrow strip of unfavourable binding modes at the smaller upper rim (0.6 nm) of γCD, which may be because of the hydroxyl groups of dihydroxymethyl-bullvalene and the upper rim of γCD getting too close (Figure 2ii. 39 (b)). A broader, more complex region of unfavourable binding modes is present for 11HoriRUpγ when bullvalene is below the centre of γCD, again when the OH groups of the substituents are close to the upper rim (Figure 2ii. 39 (c)). This is consistent for the least favourable binding modes of 11ΔUpγ and 11ΔDownγ (Figure 2ii. 39 (a) and (d)). The overall favourability of binding is unlike the complexation of this isomer with βCD, and results from the larger cavity of γCD allowing dihydroxymethyl-bullvalene to sit inside the host (Figure 2ii. 38).

The Coulombic energy maps show an alternating eight-fold symmetric pattern of favourable binding positions for 11ΔUpγ and 11ΔDownγ, a single major strip of favourable binding positions for 11HoriRDownγ, and two favourable binding regions, either side of the centre of the γCD cavity, for 11HoriRUpγ. The most favourable binding regions occur when the OH groups of the substituents are close to both the upper and lower OH groups of γCD (Figure 2ii. 37).

In general, the Coulombic energy profiles dominate the combined potential energy profiles, other than the removal of the lower favourable binding zone for 11HoriRUpγ and the lower part of the favourable binding zone for 11ΔDown.

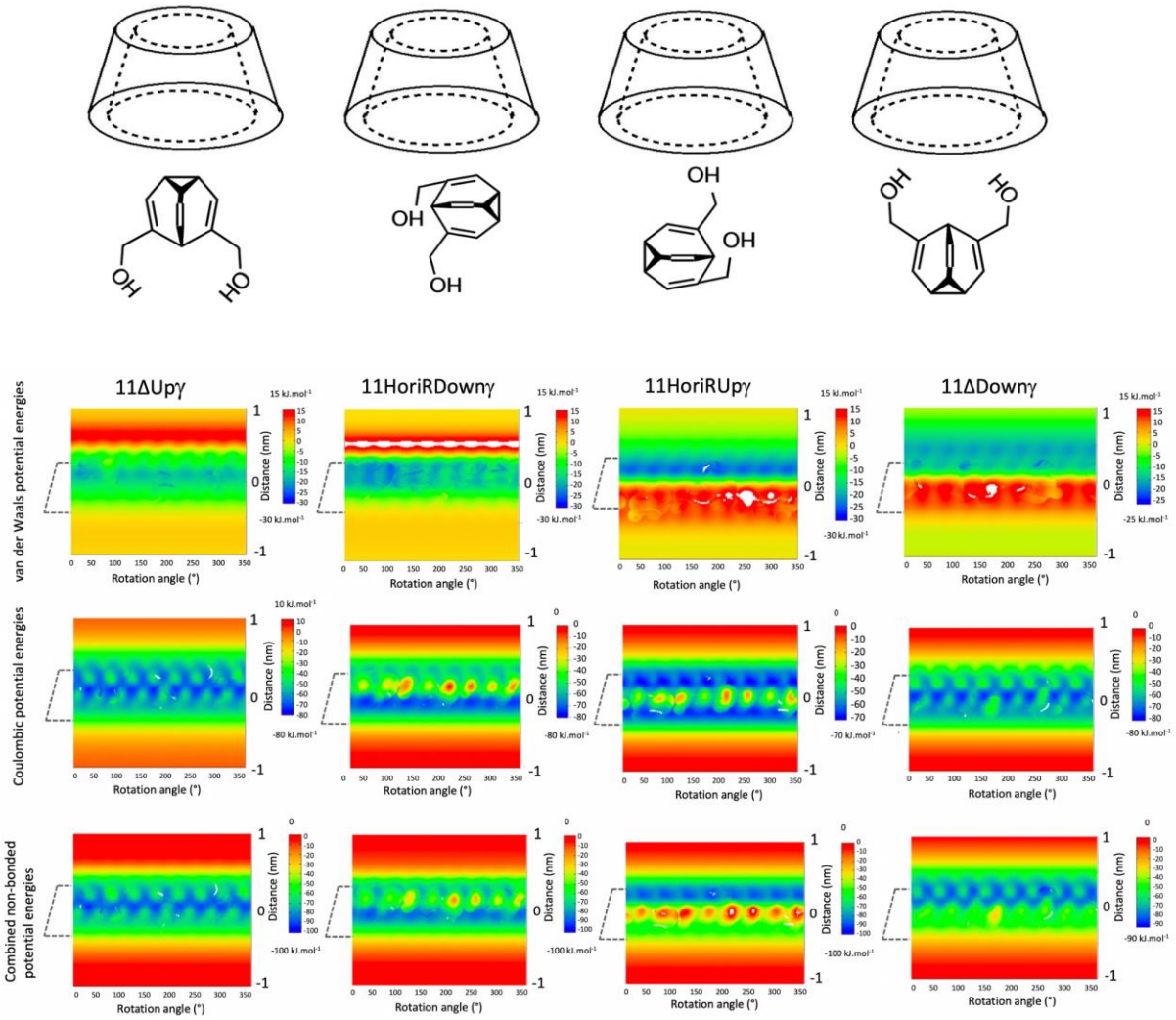


Figure 2ii. 37. HGBPE profiles for the (left) 11ΔUpγ, (centre left) 11HoriRDownγ, (centre right) 11HoriRUpγ and (right) 11ΔDownγ complexes. The first, second and third rows show the van der Waals, Coulombic, and combined non-bonded potential energies.

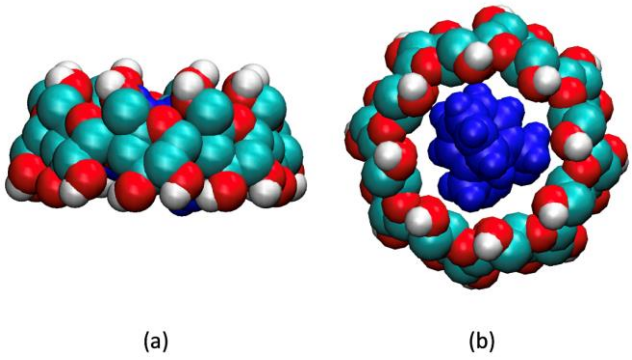


Figure 2ii. 38. VMD van der Waals representation of di-hyroxymethyl-bullvalene-γCD complex (11HoriRDownγ) (a) side view, (b) top view.

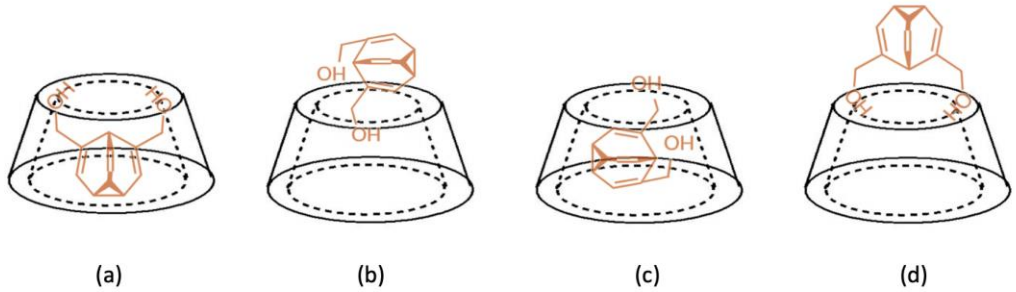


Figure 2ii. 39. Schematics of di-hyroxymethyl-bullvalene- γCD complexes for (a) 11ΔUpγ, (b) 11HoriRDownγ, (c) 11HoriRUpγ, and 11ΔDownγ (d) when the complexation is unfavourable in terms of van der Waals potential energies.

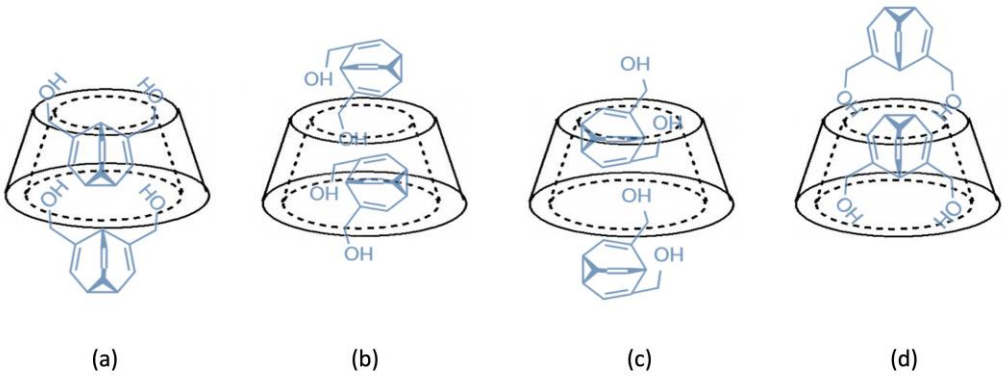


Figure 2ii. 40. Schematics of di-hyroxymethyl-bullvalene- γCD complexes for (a) 11ΔUpγ, (b) 11HoriRDownγ, (c) 11HoriRUpγ, and 11ΔDownγ (d) when the complexation is unfavourable in terms of Coulombic potential energies.

The low energy coordinates were taken from the overall potential energy plots for each orientation to undergo further MD simulations, using the same method as in Section 2ii.3.1 (Table 2ii. 31).

Table 2ii. 31. Selected *z* positions and rotation angles from the HGBPEP plots, with their combined nonbonded potential energy values, for 11ΔUpγ, 11HoriRDownγ, 11HoriRUpγ, and 11ΔDownγ.

Complex code name	Pos. Z axis	Rotation Angle (°)	Combined Pot. E ($kJ \cdot mol^{-1}$)
11ΔUpγ	0.04	33	-95.52
11HoriRDownγ	-0.10	153	-89.84
11HoriRUpγ	0.25	170	-90.62
11ΔDownγ	-0.40	67	-51.96

2ii.3.6.2. Stability of favourable modes for binding of dihydroxymethyl-bullvalene to γCD

In the MD simulations initiated from all four binding positions, bullvalene remains inside γCD. Even though di-hydroxymethyl-bullvalene includes two -CH_2OH groups and is slightly bigger than the mono-substituted bullvalene, it clearly moves and rotates during the 100 ns simulation time (Figure 2ii. 40). Nevertheless, orientations in which the substituents point towards either or both of the upper and lower rims of γCD were most populated (Figure 2ii. 41). The coordinates with favourable binding mode were taken from MD trajectories using the average values of the orientation angle and the average values of the COM distance (Table 2ii. 32).

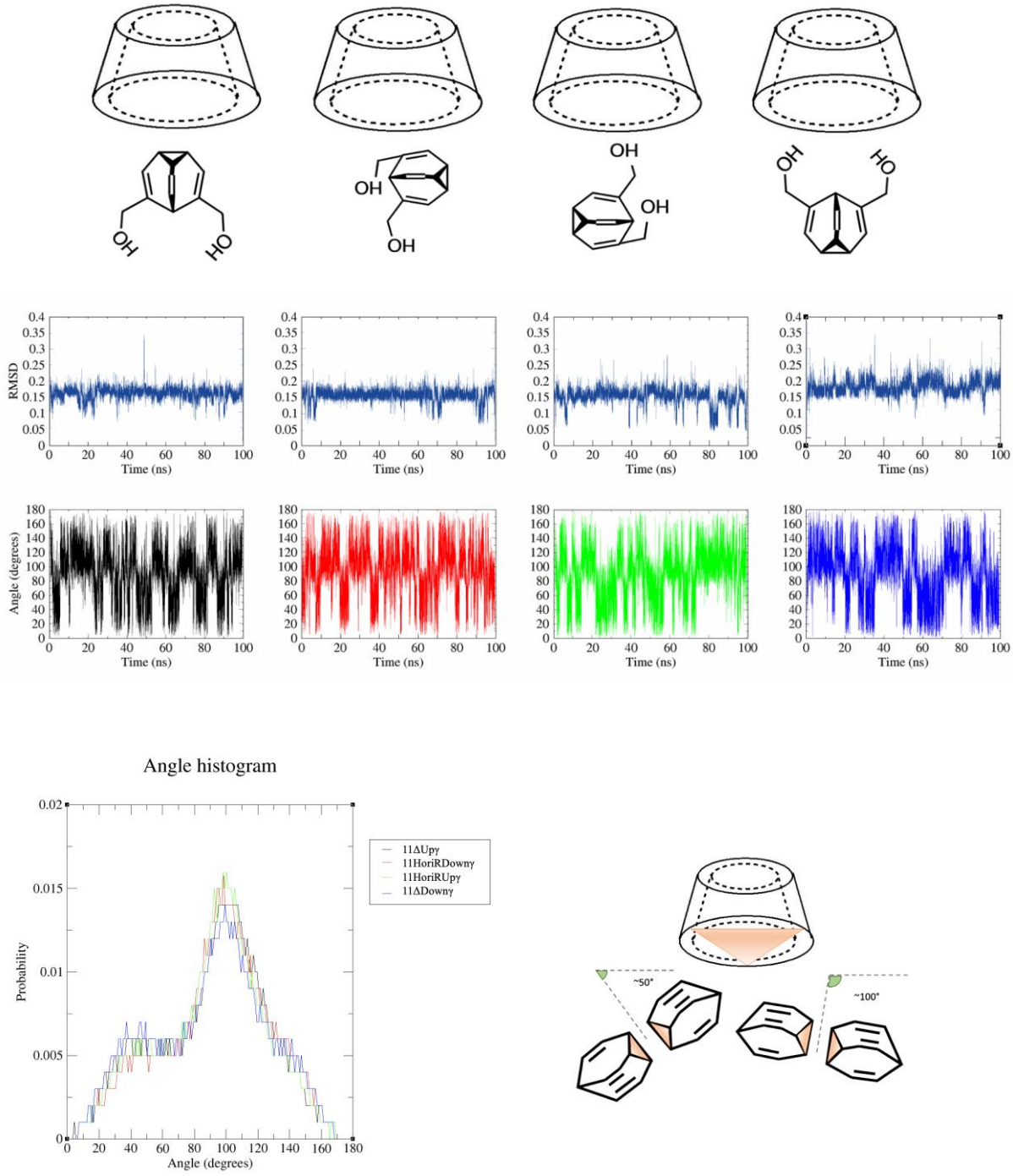


Figure 2ii. 40. RMSD (second row) and orientation angle (third row) time-series, histogram of the orientation angles (bottom row, left), and most populated orientation angles (bottom row, right) of isomer B of hydroxymethyl-bullvalene in complex with γCD during 100 ns MD simulations initiated from four different orientations. The superposition and calculation of the RMSD were both carried out for all atoms of bullvalene.

Table 2ii. 32. Average values and standard deviations of the angle between the two planes of bullvalene and cyclodextrin, the distance between the COM of bullvalene and cyclodextrin, and the RMSD, after each MD simulation, for 11ΔUpγMD, 11HoriRDownγMD, 11HoriRUpγMD, and 11ΔDownγMD.

Complex code name	**Angle (°)** (between the two planes)	**Dist.** (COM of bullvalene and γCD)	**RMSD (nm)**
11ΔUpγMD	avg. 63.25 ± 34.66	0.271 ± 0.07	0.16 ± 0.02
11HoriRDownγMD	avg. 93.62 ± 34.32	0.112 ± 0.06	0.16 ± 0.02
11HoriUpγMD	avg. 91.12 ± 34.14	0.111 ± 0.07	0.15 ± 0.03
11ΔDownγMD	avg. 89.41 ± 36.43	0.115 ± 0.07	0.17 ± 0.02

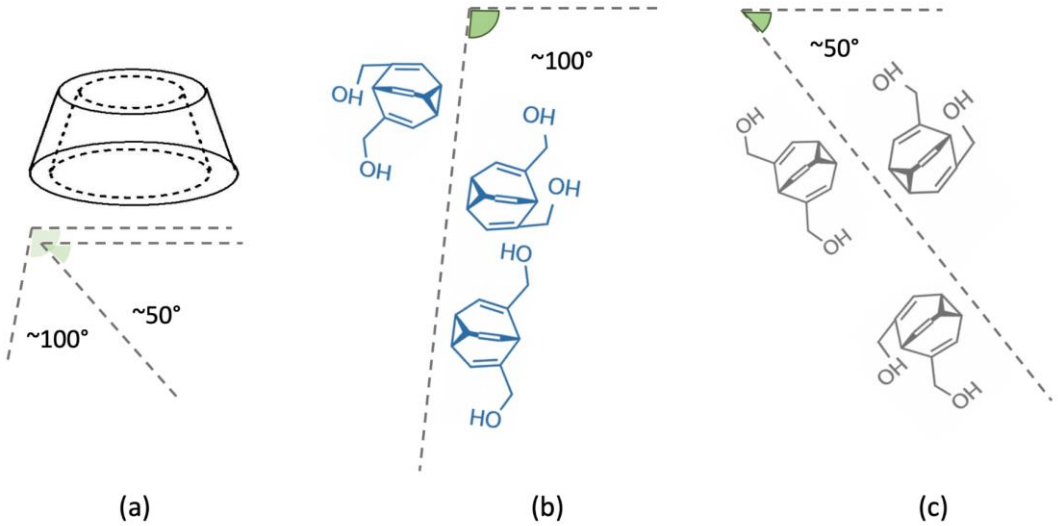


Figure 2ii. 41. (a) and (b) The first most common orientations of dihydroxymethyl-bullvalene in complex with γCD during the 100 ns MD simulations, (c) other probable orientations according to angle histogram (Figure 40).

There are three orientations for each populated angle, and bullvalene moves and rotates between all six orientations illustrated in Figure 2ii. 41 during the MD simulation. This may be because of the large cavity of γCD. Carrying out free energy calculations for all six would be computationally expensive, however, so just one representative each orientation angle was chosen for binding free energy calculations.

2ii.3.6.3. Calculation of free energies for binding of dihydroxymethyl-bullvalene to γCD

Binding free energy values were calculated for the two selected orientations of dihydroxymethyl-bullvalene with γCD using the solvation free energy calculated earlier (Table 2ii. 33). Although both chosen complexes have negative binding free energies, surprisingly, the 11ΔUpγMD complex has a very high negative value of ΔG_{bind}. Unfortunately, it is not

possible to validate these calculated binding free energy values because experimental values are not yet available for the dihydroxymethyl-bullvalene-γCD complex. However, the difference in value between the two orientations can be rationalised by considering the nature of each binding mode. The orientation chosen for 11ΔUpγMD was selected according to the second most populated orientation angle (~40°), while for 11HoriRDownγMD, the most populated orientation was selected (~100°). However, looking at Figure 2ii. 42 (a), when bullvalene has an orientation angle of ~40°, the OH groups of substituents are closer to the OH groups of the γCD rims, whereas at an orientation angle of ~100° (Figure 2ii. 42 (b)), bullvalene is more horizontal, and the OH groups of the substituents are oriented more vertically, with an increased distance to the OH groups of the γCD rims. This may explain why, in spite of both angles being populated, orientation angles of ~100 (93.62° for the coordinates selected from the MD simulation) have more favourable interactions and so more negative binding free energy than those with orientation angles of ~50° (63.25° for the coordinates selected from the MD simulation).

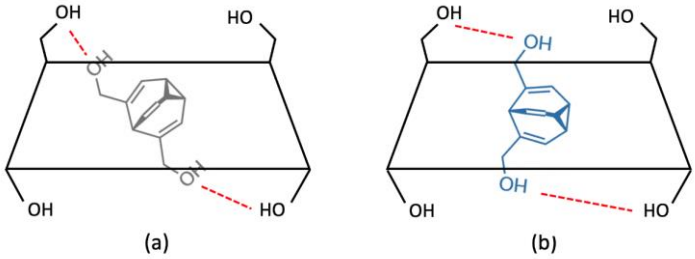


Figure 2ii. 42. The two selected orientations for free energy calculations. (a) 11HoriRDownγMD (orientation angle ~40°) and (b) 11ΔUpγMD (orientation angle ~100°). The red dashed lines represent the distances between the OH groups of dihydroxymethyl-bullvalene and γCD.

Table 2ii. 33. Solvation free energy and free energy of binding of dihydroxymethyl-bullvalene to γCD calculated using thermodynamic integration at 298 K.

Complex Code	ΔG_{solv}^{calc} (kJ·mol^{-1})	ΔG_{bind}^{calc} (kJ·mol^{-1})
11ΔUpγMD	-45.35 ± 0.29	-28.35 ± 1.43
11HoriRDownγMD		-13.06 ± 0.43

2ii.4. Conclusion

This chapter involved three major pieces of work. First, parameters and coordinates for bullvalene and β- and γCD were extensively tested, resulting in the choice of the GROMOS 54A7-compatible parameters from ATB (194–196) for bullvalenes and the 53A6_GLYC parameters for cyclodextrins (Appendix Section). Methanol was also investigated as a potential solvent but ultimately, after discussions with our experimental collaborator Dr Thomas Fallon, it was decided to use water.

Second, the HGBPEP method for rapid screening of possible binding modes for host-guest systems was developed and validated using p-CP and a series of small molecules representing its constituent parts (Chapter 2i). These investigations improved understanding of how the method works, and HGBPEP was shown to reproduce the binding mode identified experimentally for p-CP.

Third, the focus shifted to the complexation of the shape-shifter bullvalenes with cyclodextrins (Chapter 2ii). A pipeline of procedures which began with HGBPEP using 3-4 different orientations of bullvalene to identify the most favourable binding modes. These were used to initiate MD simulations to explore their stability. Lastly, the most stable binding modes from the MD simulations were used to calculate the binding free energy. In the case of monosubstituted bullvalene, these procedures were carried out for all four distinguishable isomers. For disubstituted bullvalene, only the most populated isomer, which is highly dominant, was studied.

The complete pipeline was initially applied to unsubstituted bullvalene binding to βCD. Although there are no experimental data to compare to for this complex, it provided a simple starting point for understanding how bullvalenes and cyclodextrins interact. Key points learnt from this included the seven-fold symmetric binding modes of bullvalene in certain orientations, and the fact that for specific orientations, the bullvalene cannot favourably bind to parts of the βCD cavity, due to a combination of unfavourable van der Waals and Coulombic interactions.

The complexation of a monosubstituted bullvalene, hydroxymethyl-bullvalene, with βCD was then studied. Adding one substituent (here hydroxymethyl group) to bullvalene, resulted in higher binding free energies, as well as higher solubility in water. However, as bullvalene has a larger non-polar group, it prefers the hydrophobic cavity of cyclodextrin. The most favourable binding position for mono-hydroxymethyl bullvalene is by the larger rim of βCD. Pleasingly, the calculated binding free energies were in excellent agreement with the

experimental binding constants (*Pers. Commun.*), and indicated that isomer B, although not the most populated in solution, is most like to bind to and be captured by βCD.

However, addition of two hydroxymethyl substituents (dihydroxymethyl bullvalene) brings about higher solvation free energies in water, however lower binding free energy to βCD. No experimental data is available to make a comparison.

Because of the difficulty in fully enclosing mono- and disubstituted bullvalene inside the βCD cavity, the complexation of each of the three bullvalenes with γCD was then characterised. All three bullvalenes fit inside the larger γCD cavity, with disallowed regions only occurring for specific orientations of mono- and di-substituted bullvalene. For unsubstituted bullvalene, the γCD cavity is in fact too large, such that the γCD collapses during MD simulation of the complexes and solvent is not entirely excluded from the cavity. This was also reflected in the binding free energies, which, other than for one orientation of dihydroxymethyl-bullvalene, were less favourable than for binding to βCD. Furthermore, experimental binding free energies are not available for bullvalene-γCD complexes as binding was not observed experimentally.

Overall, this chapter showed that it is indeed possible to bind bullvalenes with cyclodextrins, and the results with monosubstituted bullvalene and βCD, in particular, suggest that specific isomers can be captured, and that a tight fit is required between the host and guest molecules to ensure favourable, and likely also specific, binding.

Chapter 3. KstR

3.1. Overview

Tuberculosis (TB) has progressively become a more dangerous and difficult disease to treat due to its evolving antibiotic resistance, and propagating multi-drug resistant (MDR) strains. It is caused by *Mycobacterium tuberculosis* (Mtb), which is a pathogenic bacterial species in the Mycobacteriaceae family. It is very important to note that the persistence of Mtb is highly dependent on cholesterol metabolism (115, 116). As explained in section 1.2.31 (Chapter 1), the expression of the cholesterol catabolism enzymes is mostly regulated by KstR, which is therefore of great interest for development of new TB treatments (122, 123). KstR is a member of the large and important TetR family of one-component signal transduction systems. It controls cholesterol degradation as well as being essential for TB virulence.

The KstR structure (Figure 3. 1) (138) consists of nine α-helices. The DNA-binding domain (DBD) is composed of helices α1 to α3. The LBD is formed by helices α4 to α9. Links between helix α1 of the DBD and helices α4 and α6 of the LBD connect the two domains. The LBD includes two subdomains. Helices α5 to α7 make a triangle, while helices α8 and α9 form the dimer interface, making up a four-helix bundle with helices α8 and α9 of the other monomer (127).

Multiple ligands have been shown to bind to the hydrophobic ligand-binding pocket in the ligand binding domain (LBD), including early stage coenzyme-A (CoA) thioester metabolites of cholesterol, 3-oxo-4- cholestenoic-CoA (3OChA-CoA) (Figure 3.), and 4-BNC-CoA (213).

In the absence of cholesterol metabolites, KstR binds to [TnnAACnnGTTnnA] motifs at promoter regions of DNA (139), repressing the expression of the KstR regulon. Regardless of this repression, however, there is still a baseline expression of the occluded genes, like those encoding enzymes such as 3β-Hydroxysteroid dehydrogenase (3-β-HSD) and 3-Ketosteroid 9α-Hydroxylase (KshAB) (214). Consequently, in the presence of cholesterol, the CoA thioesters 3OChA-CoA and 4-BNC-CoA can still be produced. These then bind to KstR, which ceases to bind the DNA, thus releasing the repression by allowing RNA polymerase to bind to the previously occluded promoter regions of the KstR regulon.

Crystal structures of KstR are available in its apo state (PDB ID:3MNL) (213), bound to DNA (PDB ID: 5UA2) (142) and bound to its ligands, separately; 3OChA (PDB ID: 5CW8) and 4-BNC (PDB ID: 5CXI) (213). These structures provide insight into the conformational changes that underlie the mechanism of action of KstR. In particular, the distance between the

two DBDs is at its shortest (32 Å) in the DNA-bound state, which enables tight binding of KstR to the DNA, inhibiting ligand binding. In contrast, with either ligand bound, KstR forms a conformation with a wider distance between the DBDs (44 Å) (132, 137, 138, 215). In fact, ligand binding results in conformational changes throughout KstR, including a rotation and a translational shift of the four-helix bundle of α6-α9, shifts in the α5-α7 group of helices, and shifts in the α4 helix as well as the entire DBD. In the apo state, the distance is 38 Å, midway between its two bound states.

What the crystal structures cannot answer is the question of whether apo KstR is relatively stable in its apo conformation, and is only induced to form its bound states by the presence of ligand or DNA – an induced fit mechanism – or whether it has sufficient conformational flexibility to sample both the ligand-bound and DNA-bound states – a conformational selection mechanism. Additional questions are whether the assumptions, based on analysis of the crystal structures, that KstR binding to DNA inhibits ligand binding and *vice versa* are true. Initial insight into these questions was gained by the MD simulations of KstR carried out by Ali Razzak during his MSc (Table 3.1) (143). He ran unbiased 200 ns MD simulations initiated from the experimentally determined structures of DNA-bound KstR and ligand-bound KstR to explore their structural stability. He also ran simulations initiated from the DNA-bound and ligand-bound states but with the DNA or ligand removed to investigate transitions between states. This chapter seeks to build upon his work through additional analysis of these existing simulations, and by running and analysing new simulations, in order to determine the allosteric mechanism of KstR.

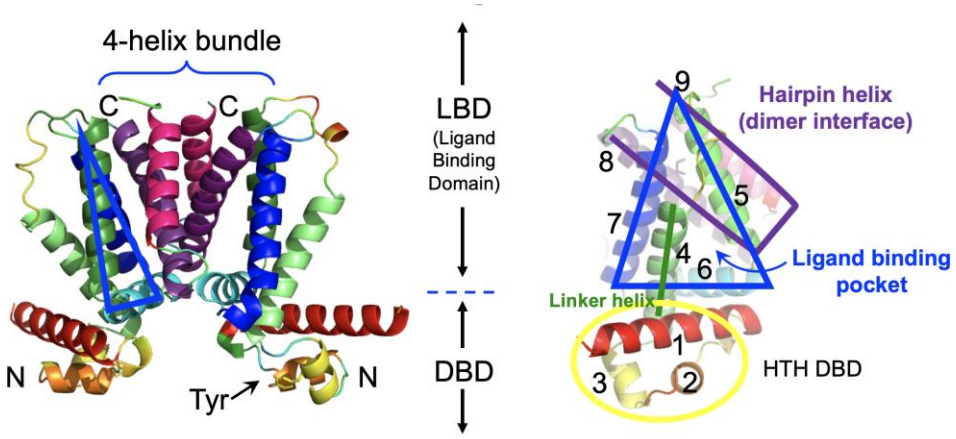


Figure 3. 1. KstR structure (PDB ID: 3MNL). The structure is drawn in cartoon format with each helix coloured differently but with corresponding colours for the equivalent helix in each monomer. Key features discussed in the text are labelled. The numbers refer to the α-helices.

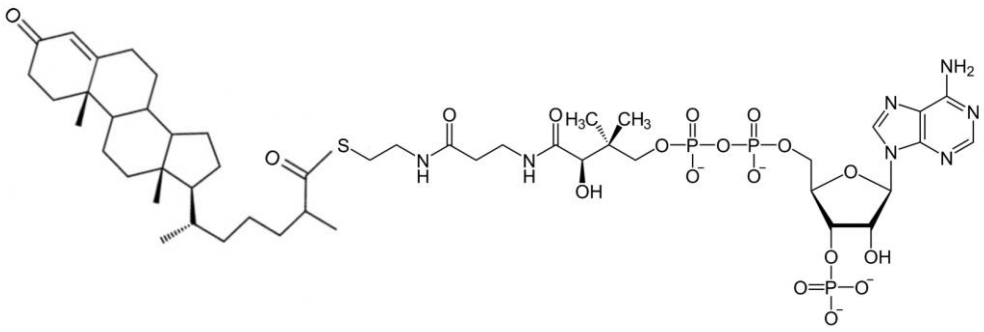


Figure 3. 2. Structure of 3-oxo-4-cholestenoic-CoA (3OChA-CoA). While crystal structures are available for KstR bound to both its ligands, only 3OChA-CoA was included in the MD simulations here.

3.2. Methods

3.2.1. Parameters

The protein and DNA were modelled using the CHARMM36 (216) and CHARMM27(217) parameter sets, respectively. Parameters for the ligands, 3OChA and 3OChA-CoA, were not available in the CHARMM27 or CHARMM36 force fields. Ligand parameters were generated by Ali Razzak (143) using CGENFF, a server that determines the parameters of a submitted PDB-format structure compatible with the CHARMM force field by comparison to already parameterised molecules (217, 218). Water was modelled using TIP3P (219).

3.2.2. Coordinates

All initial coordinates except for those of 3OChA-CoA, which was built by Ngoc Anh Thu Ho, were taken directly from the X-ray crystal structures from the PDB (3MNL, 5CW8, and 5UA2). Where residues were not resolved in the PDB structure, MODELLER was used to model them in the protein chain by Ali Razzak (143).

The coordinates of the 26 base pair (bp) DNA sequence were present in the DNA-bound KstR crystal structure (PDB: 5UA2). For the ligand-bound KstR structure (PDB: 5CW8), the acetyl-CoA tail of 3OChA-CoA is not resolved due to insufficient electron density. The complete ligand structure was therefore obtained from Ngoc Anh Thu Ho. The 3OChA ligand coordinates from the crystal structure (PDB: 5CW8) was used to orientate 3OChA-CoA by superimposing the two ligands' steroid ring atoms in Chimera (220) by Ali Razzak (143). Missing hydrogen atoms in ligand were built using Chimera and Pymol (221), also by Ali Razzak (143).

3.2.3. Computational methods

3.2.3.1. Simulation methods

All simulations for MD simulations in this work were prepared and performed with the GROMACS simulation package version 2016.3 (199–201). All MD simulations performed by Ali Razzak were done using GROMACS simulation package version 5.1.2 (143).

All MD simulations of KstR systems were initiated with the following scheme. The initial coordinates of the protein and, where appropriate, ligand and DNA were energy minimised. Then the system was solvated in TIP3P (219) water and ions added to the solvent to neutralise the system. The system was then energy-minimised again, to remove any steric clashes. The energy minimisations were done using the steepest descent minimisation algorithm for 0.1 ns until the force converged to within 1000 $kJ.mol^{-1}.nm^{-1}$.

Following the minimization, the initial velocities were generated from a Maxwell-Boltzmann distribution at 60 K. The system was then heated to the simulation temperature of 300 K over 3 ns in the NVT ensemble, with temperature controlled using the modified Berendsen thermostat, V-rescale, with a temperature coupling constant (τ_T) of 0.1 ps (204), and periodic boundary conditions. The LINCS algorithm (205) was used with an order of 4 to constrain hydrogen bonds, allowing for an integration time step of 2 fs. Non-bonded interactions were calculated using a Verlet cut-off scheme (206), with searching of neighbouring grid cells. The non-bonded Lennard-Jones and electrostatic interactions were calculated within a cut-off distance of 1.2 nm. Outside of this, electrostatic interactions were calculated using particle mesh Ewald (PME) summation (207).

The system was then further simulated for 200 ns at 300 K as outlined above but in the NPT ensemble, with pressure controlled using the Berendsen barostat in an isotropic

environment, with the pressure coupling constant (τ_p) of 0.5 ps, a reference pressure of 1 bar, and an isothermal compressibility of 4.5 x 10^{-5} bar^{-1}.

3.2.3.2. Analysis methods:

All analysis was carried out as part of this work (previous analysis by Ali Razzak was discarded) using GROMACS tools unless otherwise specified.

RMSD (Root Mean Square Deviation)

The RMSD provides a measure of the overall deviation of the molecule(s) from their initial coordinates over the course of the simulation. The Cα atoms of KstR were superimposed prior to calculation of the RMSD, again for the Cα atoms.

RMSF (Root Mean Square Fluctuation)

The root mean square fluctuation (RMSF, i.e. standard deviation) of the atomic positions from their average position during the simulation provides time-averaged information about molecular flexibility. The RMSF profiles were summarized per-residue to emphasise motion at a residue rather than atomic level.

Principal component analyses (PCA) and cartesian covariance

The functions of a protein are often closely related to its conformational dynamics. However, it is a challenge to extract the functionally relevant motions from a simulated trajectory. Correlated motions can be found by principal component analysis (PCA), which is also called covariance analysis or essential dynamics. This technique is able to represent the principal motion directions or, in other words, to understand the motions that are most fundamental to the activity of the protein by a set of eigenvectors. Eigenvectors are also called principal or essential modes, and are the projections of a simulated trajectory on the eigenvectors of the covariance matrix of the atomic coordinates. The simulation was projected onto the first five eigenvectors (data shown in plots) and the projection onto the first eigenvector was visualised in terms of the conformational changes the protein undergoes between the two extreme points of the projection. The correlated motions were calculated for the Cα atoms only, and, in cases where the termini were particularly dynamic, these were excluded.

Dynamical network analysis

NetworkView is a VMD plugin (222), which facilitates studying allostery and signalling through network models. It can be used to display representations of the networks projected onto the molecular structures. A network can be described as a series of nodes with linking edges. Typically, a node represents a group of atoms. Each atom could also be represented by a node, however, to make the analysis tractable, a coarse grained representation has been considered in this work. Therefore, each node represents an amino acid residue and is centred on the Cα atom. To define edges between pairs of nodes, one way is to link nodes that are within a specific distance of one another. Here, residues that are within a cut-off distance (4.5 Å) for at least 75% of an MD trajectory are considered linked (222).

The community substructure of the network was obtained by applying the Girvan-Newmann algorithm (223). Communities are subnetworks of the original network. Nodes which have more and stronger connectivity are partitioned as one community, that is, communities represent sets of residues that move mainly together. Nodes that connect communities so they lie in the interface between pairs of communities are called critical nodes.

The optimal and also suboptimal contact paths connecting pairs of residues were also computed from the initial dynamical network. An optimal path is the most effective pathway for allosteric communication between the two selected residues and the suboptimal paths are paths which are slightly longer than the optimal path. Each path passes each node only once; the optimal path is the shortest and has largest total edge weight.

The weight of an edge W_{ij} between the nodes i and j is the probability of information passing through that edge, and is computed as:

$$W_{ij} = -\log(|C_{ij}|) \ ,$$

Equation 3.1

where C_{ij} is the degree of correlation between two nodes:

$$C_{ij} = \frac{\langle \Delta\vec{r}_i(t) \cdot \Delta\vec{r}_j(t) \rangle}{(\langle \Delta\vec{r}_i(t)^2 \rangle \langle \ \Delta\vec{r}_j(t)^2 \rangle)^{1/2}}$$

Equation 3.2

and

$$\Delta\vec{r}_i(t) = \vec{r}_i(t) - \langle \vec{r}_i(t) \rangle$$

Equation 3.3.

$\vec{r}_i(t)$ represents the position of the i^{th} node at time t. If the two nodes (residues) i and j move in the same direction in most of the frames, the motion is considered to be correlated, and C_{ij} will be positive. If they move in opposite directions, the motion is considered anticorrelated, and C_{ij} will be negative. The motion is uncorrelated when C_{ij} is close to zero (222). As C_{ij} increases,

the weight between the two connected nodes decreases. This method for computing path weights does not differentiate between positive and negative correlation (positive and negative values of C_{ij}).

Hydrogen bond analysis

Hydrogen bond formation between the KstR monomers was computed using the VMD Hydrogen Bonds plugin with its default definition of hydrogen bond formation (donor-acceptor distance = 0.3 nm; hydrogen-donor-acceptor angle cutoff = 20°).

Hydrogen bond formation between the KstR and DNA or between KstR and ligands were computed using in-house Python scripts to call the VMD Hydrogen Bonds plugin with its default definition of hydrogen bond formation (donor-acceptor distance = 0.3 nm; hydrogen-donor-acceptor angle cutoff = 20°) and plotted using the Python matplotlib package (224).

3.3. Results and discussion:

The goal of this chapter is to provide insight into whether KstR operates via an induced fit mechanism or a conformational selection mechanism. Induced fit would mean that the apo state is only induced to form its bound states by the presence of ligand or DNA, whereas conformational selection would require the apo state to have sufficient conformational flexibility to sample both the ligand-bound and DNA-bound states. Additional questions to be answered are whether the assumptions, based on analysis of the crystal structures, that KstR binding to DNA inhibits ligand binding and *vice versa* are true.

To investigate these questions, the stability and conformational dynamics of each state of KstR – apo, ligand-bound, and DNA-bound – was investigated by running unbiased MD simulations of each (KstR:apo, KstR:3OC and KstR:DNA, respectively; Table 3. 1). These simulations were initiated from the crystal structures of KstR in its apo state (PDB ID: 3MNL, (213)), bound to 3OChA-CoA (PDB ID: 5CW8, (213)), and bound to a 26 bp DNA fragment (PDB ID: 5UA2, (142)). The initial structures of the ligand-bound and DNA-bound states were prepared by Ali Razzak (143); and the apo form was set up in this work. Each system was simulated for 200 ns (method as Section 3.2.3.1), with the apo state simulated twice to determine whether its limited conformational sampling was a true property or due to the initial conditions.

The transitions between states were also investigated by removing the ligand or DNA from the final coordinates of the MD simulation of each respective bound state and running unbiased 200 ns MD simulations (KstR:3OCremoved and KstR:DNAremoved, respectively; Table 1) to see whether and if so, how quickly, they reverted to the apo state.

Lastly, to investigate whether ligand binding forces KstR to unbind from DNA or *vice versa*, two more simulations were performed, one for KstR bound to DNA and ligand simultaneously (KstR:3OC:DNA), and the other one with the DNA removed, starting from the endpoint of the former simulation (KstR:3OC:DNAremoved). The initial structure of KstR:3OC:DNA was a hypothetical structure: while KstR has not been crystallised with both a ligand and DNA bound, the KstR:DNA structure (PDB ID: 5UA2) has a sufficiently large ligand-binding cavity to allow insertion of 3OChA. Preparation of the KstR:3OC:DNA structure was carried out by Ali Razzak (143), but running of the 200 ns MD simulation was carried out as part of this work. Subsequently, the DNA was removed and KstR:3OC:DNAremoved was run for 300 ns to determine whether the slow transition to an apo-like state would continue.

Table 3. 1. Key properties of the seven simulations of KstR

Simulation code	Description	Initial coordinates	Length	MD by
KstR:apo	KstR protein only	PDB ID: 3MNL	200 ns	ZS*
KstR:DNA	KstR protein bound to 26 bp DNA	PDB ID: 5UA2	200 ns	AR**
KstR:3CO	KstR protein bound to 3OChA ligand	PDB ID: 5CW8	200 ns	AR
KstR:DNAremoved	DNA removed from PDB ID: 5UA2	KstR:DNA	200 ns	AR
KstR:3OCremoved	3OChA removed from PDB ID: 5CW8	KstR:3OC	200 ns	AR
KstR:3OC:DNA	3OCh added to ligand binding pocket of PDB ID: 5UA2	PDB ID: 5UA2	200 ns	ZS
KstR:3OC:DNAremoved	DNA removed from endpoint of KstR:3OC:DNA	Endpoint of KstR:3OC:DNA	300 ns	ZS

* Zahra Shadfar

** Ali Razzak

3.3.1. Structural stability

The atom-positional root-mean-square deviation (RMSD) of Cα atoms from the initial conformation of KstR after solvation step at each point in time was calculated to determine how much the protein structure changes during each simulation (Figure 3.1). RMSD gives a single value quantifying the structural deviation of the protein from the reference structure at each time point in the simulation. The averages and standard deviations over the course of the simulations are provided in Table 3.2. Although the two KstR:apo simulations have similar inter-DBD distance average value, the RMSD averages were different. This could be possibly

because of the starting points, which are the last frame of the heating procedure were different from each other and the RMSD averages compared to starting coordinates were different in the two replicate simulations. However, the DBD- distances remained very similar in the two simulations.

3.3.1.1. RMSD, KstR:apo vs KstR:3OC vs KstR:DNA

The RMSD time-series for KstR:apo, KstR:DNA and KstR:3OC (Figure 3.1, top left) shows that after an initial increase, the RMSD plateaus and therefore the system seems to have equilibrated at approximately 50 ns in all three cases, although KstR3OC undergoes structural fluctuations during the first 50 ns than the other two systems and equilibrates to a higher RMSD value, suggesting greater movements away from the starting structures.

3.3.1.2. RMSD, DNAremoved vs 3OCremoved

KstR:DNAremoved and KstR:3OCremoved behave similarly to the aforementioned states, but if anything, equilibrate even faster, albeit to slightly higher RMSD values. While the rapid equilibration is somewhat surprising given that these simulations were intended to follow the transition between bound and unbound states, the higher RMSD values are expected; indeed, it is perhaps surprising that the RMSD values are not larger.

3.3.1.3. RMSD, KstR:3OC:DNA vs KstR:3OC:DNAremoved

KstR:3OC:DNA equilibrates quickly to a relatively low RMSD value of ~0.25 nm, with little fluctuation, suggesting that binding to both ligand and DNA renders the protein rigid. Upon removal of the DNA, the RMSD values only slowly increase but undergo more fluctuation, suggesting that escape from the rigid DNA- and ligand-bound state is slow. The simulation was then extended to 300 ns. There is even higher fluctuation of the RMSD in the last 100 ns of the simulation. These fluctuations are larger than in any of the other MD simulations (Table

3.2), suggesting that the process of escaping the rigid doubly-bound conformation and reverting to a ligand-bound state involves large-scale structural motions.

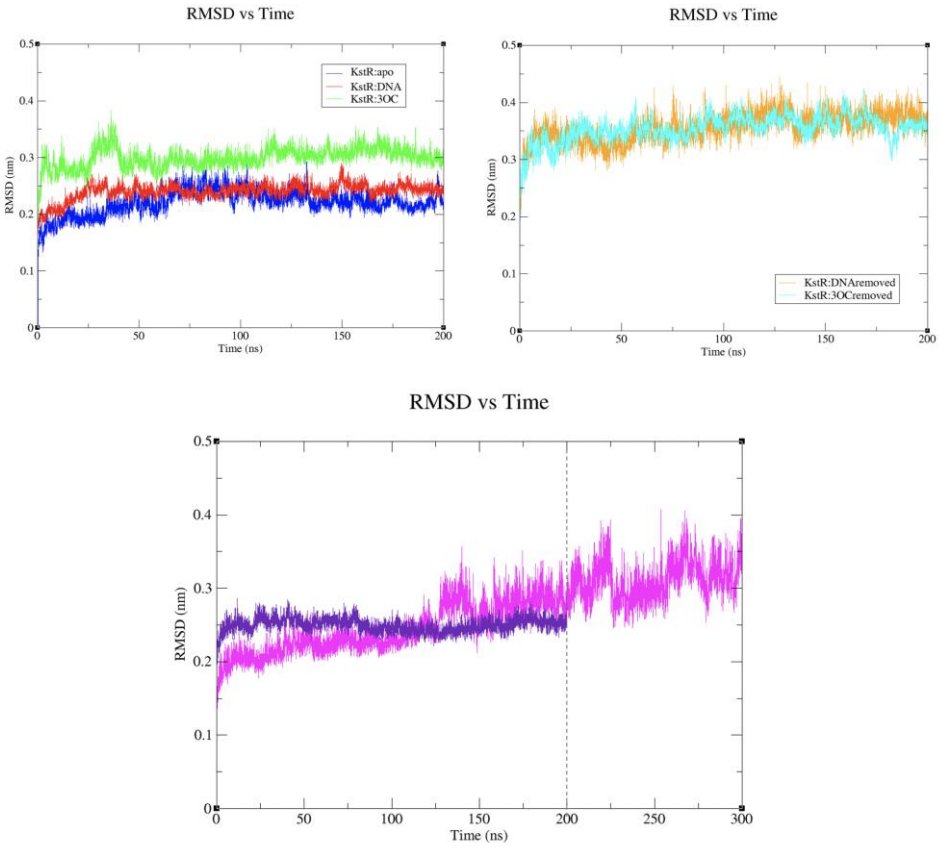


Figure 3. 1. RMSD of Cα atoms of KstR with respect to the starting structure (after solvation step) during the MD simulations of (top left) KstR:apo (blue), KstR:DNA (red) and KstR:3OC (green); (top right) KstR:DNAremoved (orange) and KstR:3OCremoved (cyan); and (bottom) KstR:3OC:DNA (purple) and KstR:3OC:DNAremoved (magenta).

Table 3. 2. Averages and standard deviations of the RMSD values and minimum distances between the TYR54 residue in each of the two monomers, excluding the first 50 ns of the simulation time.

Simulation code	RMSD (nm)	Minimum distance (nm) TYR54 (monomer A) - TYR54 (monomer B)
KstR:apo1	0.18 ± 0.02	4.17 ± 0.10
KstR:apo2	0.34 ± 0.02	4.11 ± 0.15
KstR:DNA	0.24 ± 0.01	3.22 ± 0.04
KstR:3OC	0.30 ± 0.01	4.60 ± 0.18
KstR:DNAremoved	0.36 ± 0.02	4.24 ± 0.15
KstR:3OCremoved	0.36 ± 0.02	4.39 ± 0.20
KstR:3OC:DNA	0.25 ± 0.01	3.34 ± 0.06
KstR:3OC:DNAremoved	0.20 ± 0.04	4.07 ± 0.31

While the RMSD measures the structural deviation of the entire protein averaged over all residues, the RMSF is the RMSD of each atom (or the group of atoms making up a residue) from its average position during the course of the simulation, averaged over time. It therefore

provides information on which residues undergo the most conformational changes during the simulation.

3.3.1.4. RMSF, KstR:apo vs KstR:3OC vs KstR:DNA

As is commonly the case for proteins undergoing MD simulation, the two terminals of each monomer exhibit high RMS fluctuations, appearing as high peaks on the middle, left and right end of the graph, during the simulations of KstR:apo, KstR:DNA and KstR:3OC (Figure 3. 2, red, blue and green ball models of atoms, respectively). The baseline RMSF values of KstR:apo and KstR:3OC are similar, but those of KstR:DNA are generally lower, suggesting again that this state is more rigid. The positions of the residues that undergo the largest fluctuations are also similar across the three simulations, but the degree of these fluctuations differs between simulations. KstR:3OC exhibits a high degree of fluctuation of the linker loop between α4 helix and α5 helix of each monomer (Figure 3. 2, yellow spheres). Although this motion exists in KstR:apo and KstR:DNA simulations, it is less than in the ligand bound simulation. KstR:DNA undergoes more fluctuation here than KstR:apo in monomer A, but less than KstR:apo for the equivalent residues in monomer B.

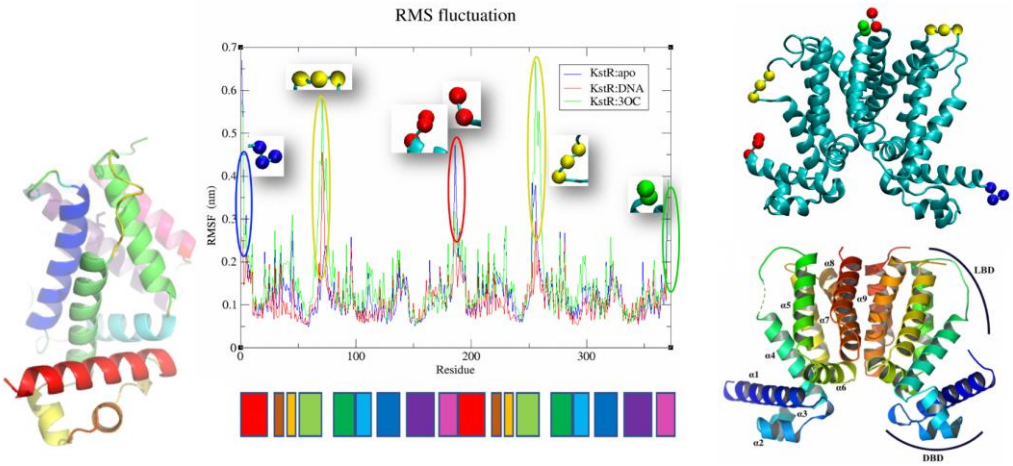


Figure 3. 2. RMSF summarised per-residue for the MD simulations of KstR:apo (blue), KstR:DNA (red) and KstR:3OC (green). The coloured panels underneath the graph indicate the residues that correspond to each helix of KstR as coloured in the structures shown at either side. The regions with the highest RMSF values are indicated by the inset coloured spheres that correspond to those on the structure at the top right.

3.3.1.5. RMSF, KstR:apo vs KstR:3OCremoved vs KstR:DNAremoved:

The RMSF values for KstR:3OCremoved and KstR:DNAremoved are largely similar to those of KstR:apo (Figure 3. 3). Interestingly, like for KstR:3OC, KstR:3OCremoved also exhibits

the most fluctuation, especially for the loops linking the α4 and α5 helices, despite removal of the ligand. Removing DNA, however, does change the flexibility, leading to a structure that is even more flexible structure than KstR:apo. These results suggest that the "apo" state sampled after removal of ligand or DNA is more flexible than the apo state sampled by KstR:apo, initiated from the crystal structure.

3.3.1.6. RMSF, KstR:3OC:DNA vs KstR:3OC:DNAremoved vs KstR:DNA vs KstR:3OC

Comparison of the RMSF values for the KstR:3OC:DNA and KstR:3OC:DNAremoved simulations (Figure 3. 4, left) shows that after removing DNA, there are slightly larger fluctuations of residues 35-60 on the plot (ARG45-ARG70) and 70-80 on the plot (ASP80-PHE90) in monomer A, and residues 220-245 on the plot (ARG45-ARG70) and 255-265 on the plot (ASP80-PHE90) in monomer B, which are highlighted in magenta in Figure 3. 4 (bottom left), and also shown with dashed lines on the graph (top left). Here, the numbers of the residues are according to the RMSF plots, which the residues are renumbered for RMSF calculations to start from 1, while in the crystal structure and the residue numbers provided in brackets, the first residue is number 11. Residues 35-60 and 220-245 are parts of the DBDs in both monomers, which play a key role in binding to DNA, and are therefore expected to become more mobile when DNA is removed. Comparison of the RMSF values of KstR:3OC:DNA with those of KstR:DNA (Figure 3. 4, right) shows the baseline RMSF values to be similarly low, but KstR:3OC:DNA exhibits higher RMSF values for residues about 110-130 for monomer A, suggesting that inclusion of 3OCh causes this part of the structure, which is in helices α5 and α6, in the LBD triangle, to move more. This could be due to the strain required to hold the ligand in the pocket while also binding DNA. These residues are highlighted in purple in Figure 3. 4 (bottom right), and also showed with dashed lines on the graph (top right). The RMSF values for KstR:3OC:DNAremoved are larger than those of KstR:3OC for residues 220-245, subunit B, but smaller for the loops linking the α4 and α5 helices. KstR:3OC:DNAremoved has therefore not reverted completely to a state equivalent to KstR:3OC during the 200 ns of the simulation analysed here (for consistency with the other simulations that were only run for 200 ns).

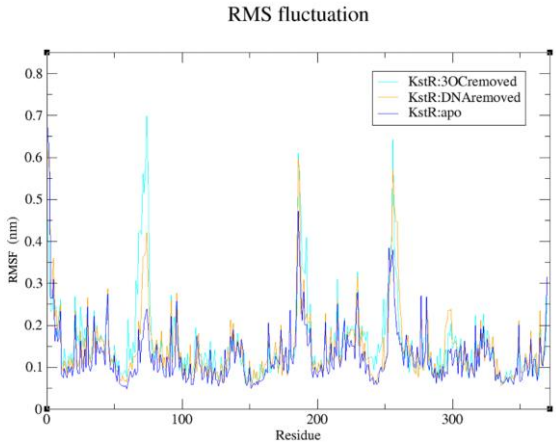


Figure 3. 3. RMSF summarised per-residue for the MD simulations of KstR:3OCremoved (cyan), KstR:DNAremoved (orange), and KstR:apo (blue) as a reference.

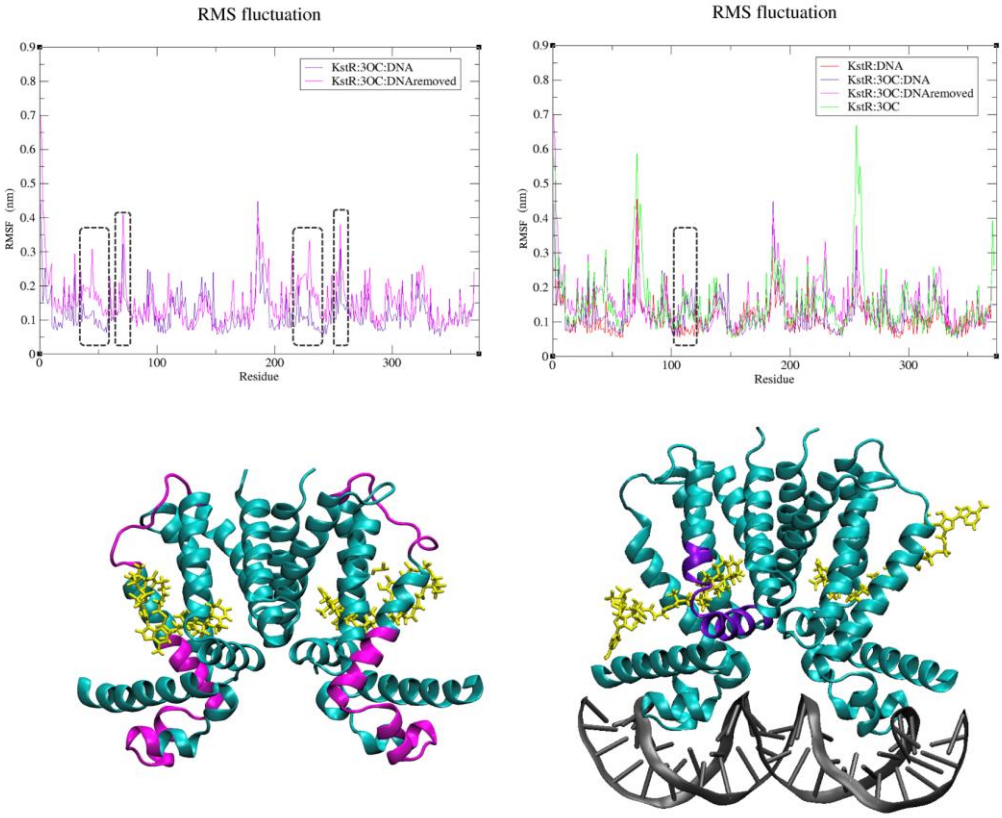


Figure 3. 4. RMSF summarised per-residue for the MD simulations of (left) KstR:3OC:DNA (purple) and KstR:3OC:DNAremoved (magenta) and (right) KstR:DNA (red), KstR:3OC:DNA (purple), KstR:3OC:DNAremoved (magenta), and KstR:3OC (green). (bottom, left) The residues that show greater fluctuations upon removing DNA are indicated in magenta on the cartoon structure of KstR, and are also shown with dashed lines on the left-hand graph. (bottom, right) The residues that show higher fluctuations in KstR:3OC:DNA compared to KstR:DNA are indicated in purple on the cartoon structure of KstR, and are also shown with dashed lines on the right-hand graph. In both cases, the 3OCh ligand is drawn in licorice format and coloured yellow.

3.3.2. Separation of DNA-binding domains

3.3.2.1. KstR:DNA vs KstR:apo vs KstR:3CO

A key factor in determining whether KstR is capable of binding DNA is the distance between the DBDs. A simple way to analyse this is to measure the distance between the Cα atoms of a key residue, TYR54, in the DBDs of each monomer. This TYR is one of the highly conserved residues that are found in almost all TetR family members, independent of the DNA sequence to which the protein binds, as they facilitate binding to DNA by inserting into the major groove rather than interacting directly with the DNA bases (Stephanie Dawes, *Pers. Commun.*). The role of TYR54 in DNA binding means that the distance between the TYR54 in each DBD is a useful proxy for whether that conformation of KstR could bind to DNA. In the crystal structures of apo, 3OCh-CoA-bound and DNA-bound KstR, these distances are 3.8 nm, 4.3 nm and 3.2 nm, respectively (Figure 3. 5, right). During the MD simulation of KstR:apo, the DBDs relax outward to sample distances between 4.0 and 4.25 nm, larger than their 3.8 nm separation in the crystal structure (Figure 3. 5, left). For KstR:DNA, however, the DBDs remain close, sampling distances around 3.25 nm during the MD simulation, which is in good agreement with the 3.2 nm distance measured for the crystal structure, and this distance fluctuates very little, suggesting tight and stable binding of the DNA. The inter-DBD distance for KstR:3OC was the most changeable, fluctuating between 4.0 and 5.25 nm during the simulation, a range of distances that encompasses the 4.3 nm separation observed for the crystal structure. Together, the simulations suggest that DNA binding stabilises the structure of KstR and keeps the DBDs close together, which is required for DNA binding, and that the apo and ligand-bound states are more flexible. The apo state does not sample the small inter-DBD distances required for DNA binding, which counts against a conformational selection mechanism.

A second MD simulation of KstR:apo was therefore run to check if it produced the same behaviour. The same analyses were carried out as are reported throughout this chapter, but only the inter-DBD distance (Figure 3. 5) and the average RMSD and inter-DBD distances (Table 3.2) are reported, as all results for the replicate simulation were essentially the same as for the first KstR:apo simulation, suggesting that it was not caught in a low-energy state. Therefore, the replicate results are not displayed on the other distance plots in the following sections.

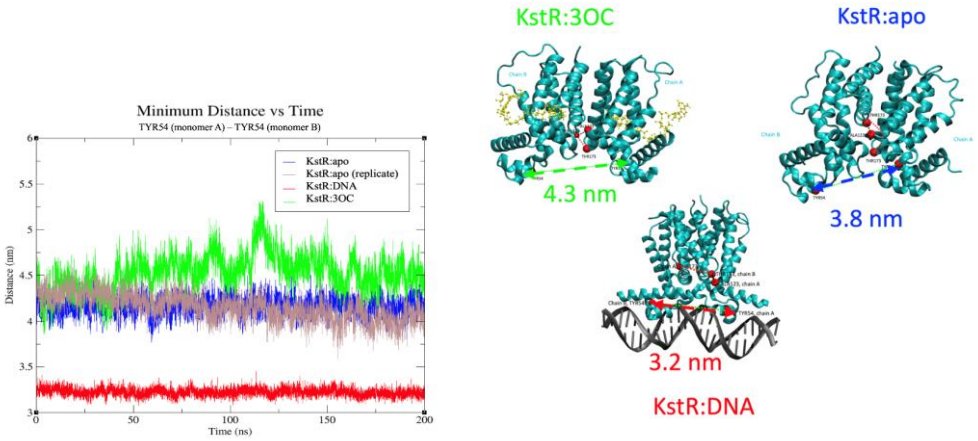


Figure 3. 5. (Left) Distance between the Cα atoms of TYR54 of each monomer of KstR during the simulations of KstR:apo (blue), KstR:apo-replicate (brown), KstR:DNA (red) and KstR:3OC (green). (Right) Crystal structures of each state in cartoon format, with the 3OCh-CoA ligand in yellow CPK representation and DNA in black cartoon format. Key residues between which distances were calculated are shown in red (note that not all distances are reported here) and the distance between the Cα atoms of TYR54 are indicated.

3.3.2.2. KstR:apo vs KstR:3COremoved vs KstR:DNAremoved vs KstR:DNA

Upon removal of DNA, the inter-DBD distance quickly reverted to a KstR:apo-like state, after only 1 ns (Figure 3. 6, left). In KstR:3OC, the inter-DBD distance was already wider and more flexible than that of KstR:apo (Figure 3. 5, Table 3.2), and removing the ligand did not noticeably tighten up the inter-DBD distance. These results and the RMSD and RMSF data presented in Section 3.3.1 caused concern that the KstR:apo simulation might be "stuck" in a low-energy state near the crystal structure, as KstR:3OCremoved and KstR:DNAremoved appear to be more flexible. Although the replicate simulation of KstR:apo gave similar results to the initial simulation, it remains possible that in both simulations KstR was unable to exit a low-energy state similar to the crystal structure. Extending these simulations or running a larger number of independent replicate simulations may be required to determine whether KstR:apo is indeed more rigid than KstR:3OCremoved and KstR:DNAremoved or whether this is a anomalous result due to some aspect of the initial coordinates .

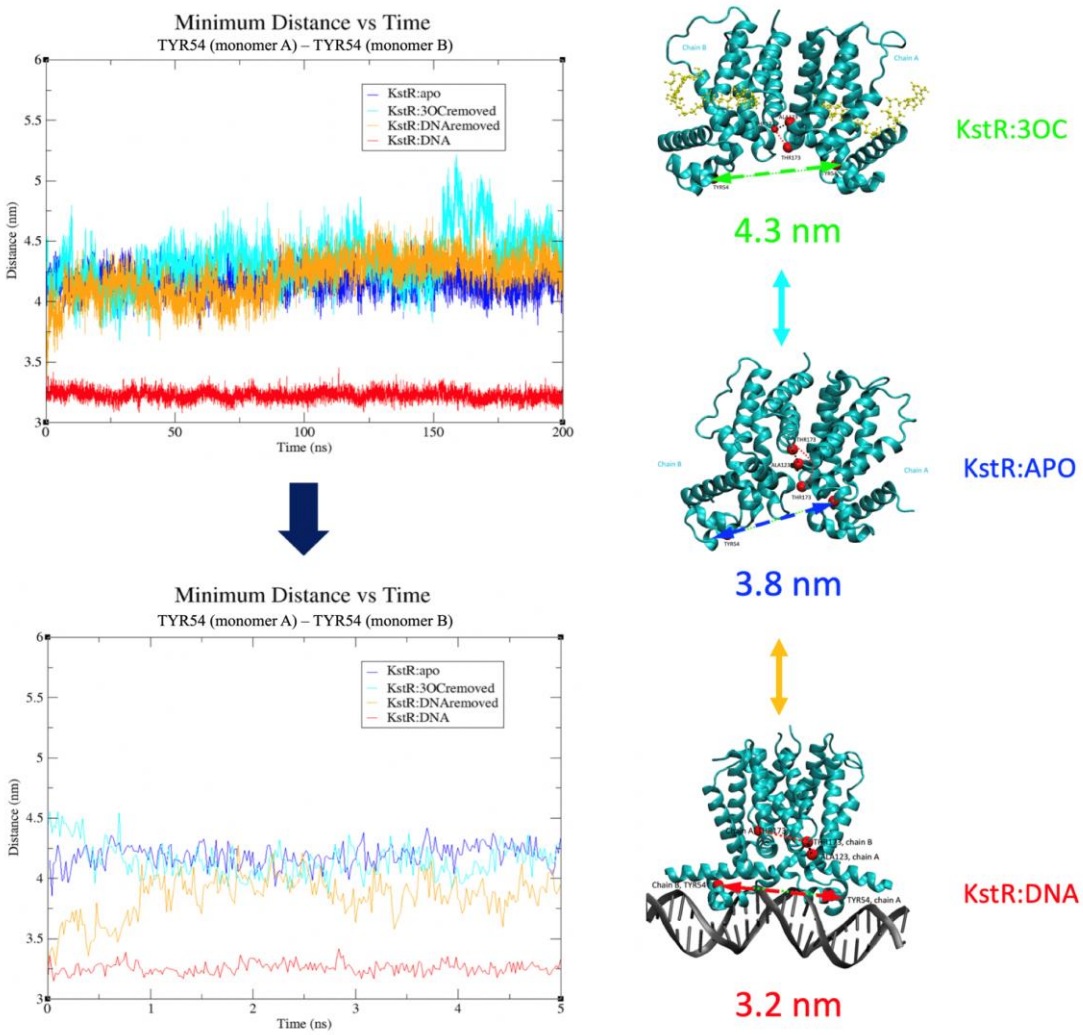


Figure 3. 6. (Top left) Distance between the Cα atoms of TYR54 of each monomer of KstR during the simulations of KstR:apo (blue), KstR:3OCremoved (cyan), KstR:DNAremoved (orange), and KstR:DNA (red) and (bottom left) for the first 5 ns of the simulation. (Right) Crystal structures of each state in cartoon format, with the 3OCh-CoA ligand in yellow CPK representation and DNA in black cartoon format. Key residues between which distances were calculated are shown in red (note that not all distances are reported here) and the distance between the Cα atoms of TYR54 are indicated.

3.3.2.3. KstR:3OC:DNA vs KstR:3CO:DNAremoved

The distance between the TYR54 residues of each DBD of KstR in KstR:DNA and KstR:3OC:DNA was similar throughout the simulations, although it is slightly larger for KstR:3OC:DNA, potentially signaling some stress in the protein structure when forced to bind DNA and 3OCh simultaneously (Figure 3. 7, Table 3.2). Binding of 3OCh to KstR along with DNA does not, however, cause rejection of KstR from DNA, at least on the time-scale of the simulation.

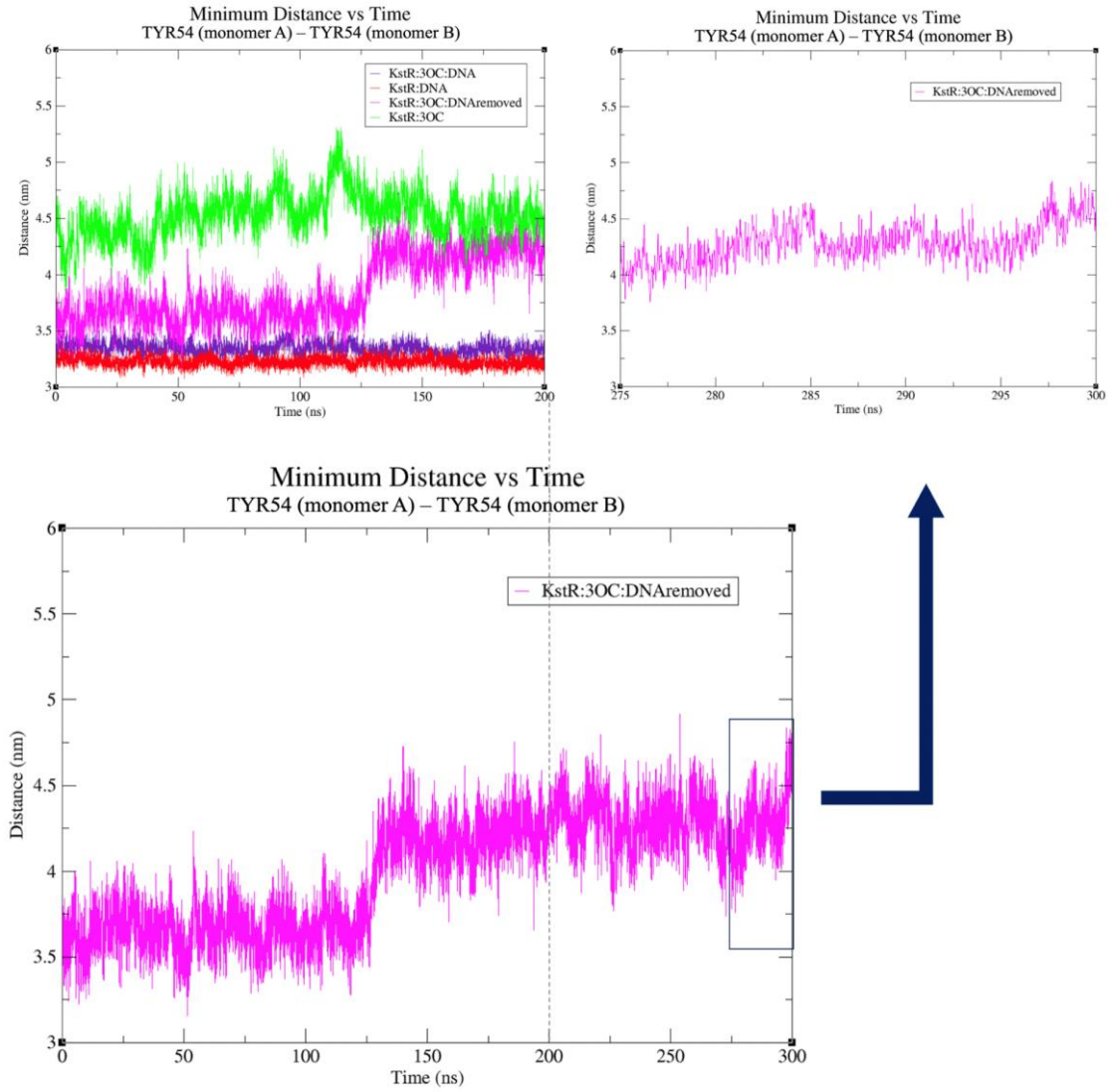


Figure 3. 7. (Top-left) Distance between the Cα atoms of TYR54 of each monomer of KstR during the simulations of KstR:3OC:DNA (purple), KstR:DNA (red), KstR:3OC:DNAremoved (magenta),and KstR:3OC (green) during the first 200 ns of MD simulation; (bottom left) for KstR:3OC:DNAremoved for the extended 300 ns MD simulation; and (top right) for the final 25 ns of the extended simulation of KstR:3OC:DNAremoved.

As soon as DNA was removed from this system, the inter-DBD distance only slowly reverted to be more like KstR:apo and KstR:3OC. For the first 125 ns, the DBDs move a little further apart and the inter-DBD distance fluctuates more, but only after this do they move further apart to > 4 nm separation. This simulation was therefore extended further to 300 ns. At the very end of this extended simulation, the TYR-TYR distance appeared to be increasing further to resemble KstR:3OC more, however, the simulation is not converged, and would need to be extended further to determine whether KstR:3OC:DNAremoved ultimately reverts to a state equivalent to KstR:3OC. Strikingly, the reversion to a KstR:3OC-like state was substantially slower here than the near-instantaneous reversion (~1 ns) to large DBD distances when DNA

was removed without ligand bound (Figure 3. 7). These results suggest that dual binding of DNA and ligand may "lock" the protein structure in some way, and that it takes some time for this effect to disappear.

3.3.3. Determination of concerted motions

To determine the major and concerted conformational motions that KstR undergoes during each simulation, cartesian covariance and principal component analysis (PCA) were carried out for each of the seven simulations. The entire simulation was analysed, except where indicated otherwise.

3.3.3.1. KstR:DNA vs KstR:apo vs KstR:3CO

Visualisation of the motion that KstR undergoes as it transitions between the two extremes of the first principle component revealed that KstR:DNA had the least motion, as expected based on the earlier analyses. The overall motion of KstR:DNA resembled a twisting of the LBDs relative to the DBDs, especially when viewed from above (Figure 3. 10, top). This leads to slight rotation of DBDs with respect to the LBDs, but does not result in an increase in the inter-DBD distance. KstR:apo behaves similarly, but with more obvious motions. KstR:3OC also shows some minor twisting of the two LBDs, but the more remarkable conformational change is the breathing motion of DBDs, which swing in and out. The videos of the extreme motion projected onto the first principal component for all simulations can be found on FigShare (https://figshare.com/s/3b59449529851d89dc17).

The projection of the simulation onto the first five eigenvectors (Figure 3. 8, bottom) supports the results above, in that the scale of the first eigenvector projection is larger for KstR:3OC than for the other two systems. KstR:apo in turn has less rigidity than KstR:DNA, as expected.

3.3.3.2. KstR:DNA vs KstR:DNAremoved vs KstR:apo

Upon removal of DNA, the KstR inter-DBD distance in KstR:DNAremoved rapidly reverted to KstR:apo-like state. Therefore, PCA was done separately for the first 10 ns of the simulation time, which is considered as the transition period, and also for the remainder of the simulation (during which the inter-DBD distance resembles that of KstR:3OC). In the first 10 ns of the simulation, the DBDs appear to twist with respect to the LBDs, like in the KstR:DNA

simulation (Figure 3. 9). For the remainder of the simulation, however, the motions were more like those of KstR:3OC – less twisty and more like a breathing of the DBDs.

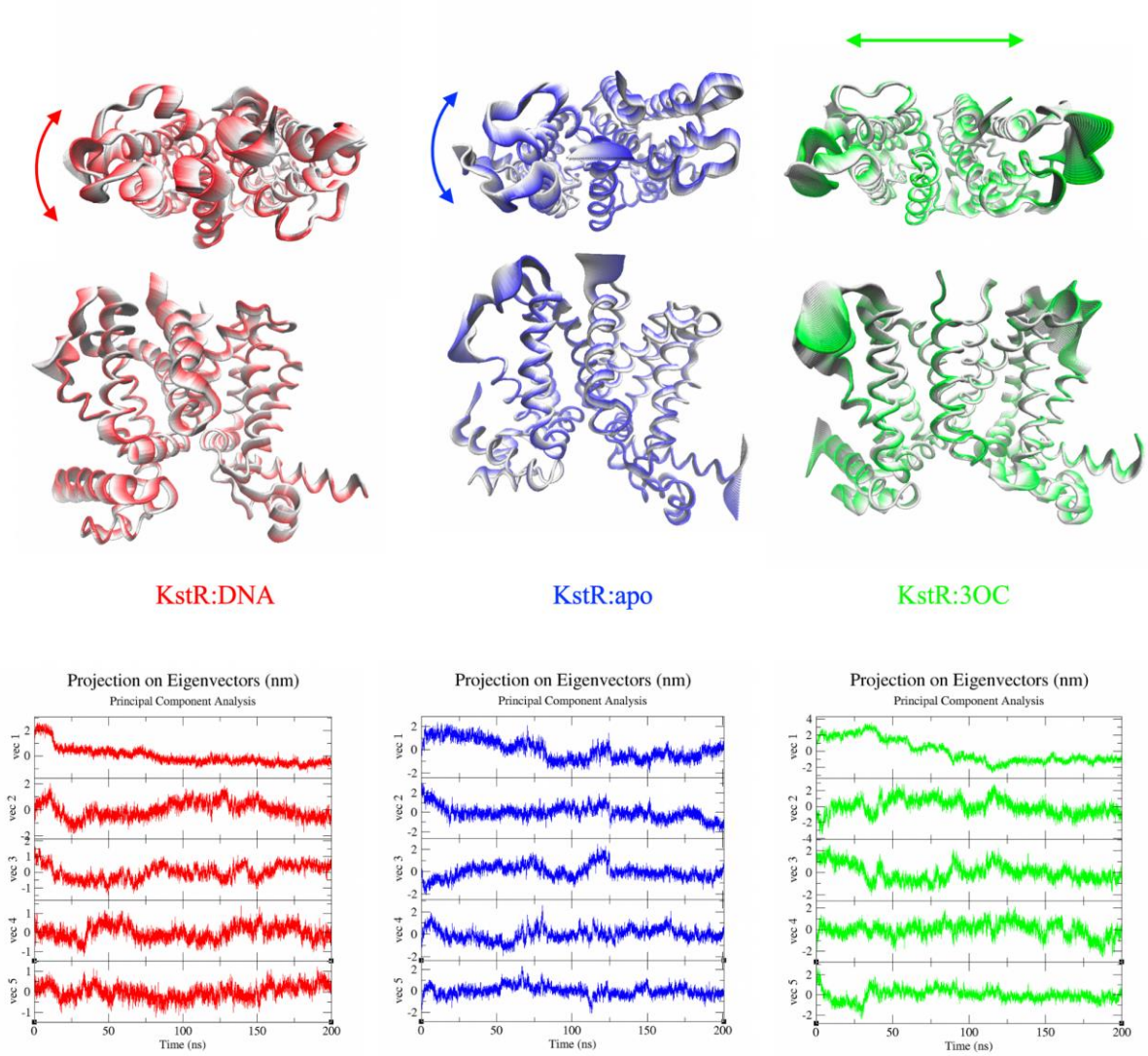


Figure 3. 8. (Top) Visualisation of the extreme motion projected onto the first principal component, where the two extremes are represented by white and by the colour (red) KstR:DNA, (blue) KstR:apo and (green) KstR:3OC. (Bottom) Projection onto the first five (largest) eigenvectors during the simulation time for (blue) KstR:apo, (red) KstR:DNA and (green) KstR:3OC.

3.3.3.3. KstR:3OC vs KstR:3OCremoved vs KstR:apo

After removing the ligand from KstR:3OC, KstR behaved similarly to the ligand-bound state, in terms of its overall motion. Considerable breathing motions of the DBDs occurred, whereas in KstR:apo there were constrained twisty motions (Figure 3. 10). As with previous results, these observations and those for KstR:DNAremoved suggest that the "apo" state sampled upon removal of DNA or ligand is quite different and substantially more flexible than that sampled in the KstR:apo simulation. Moreover, the nature of the conformational changes is different.

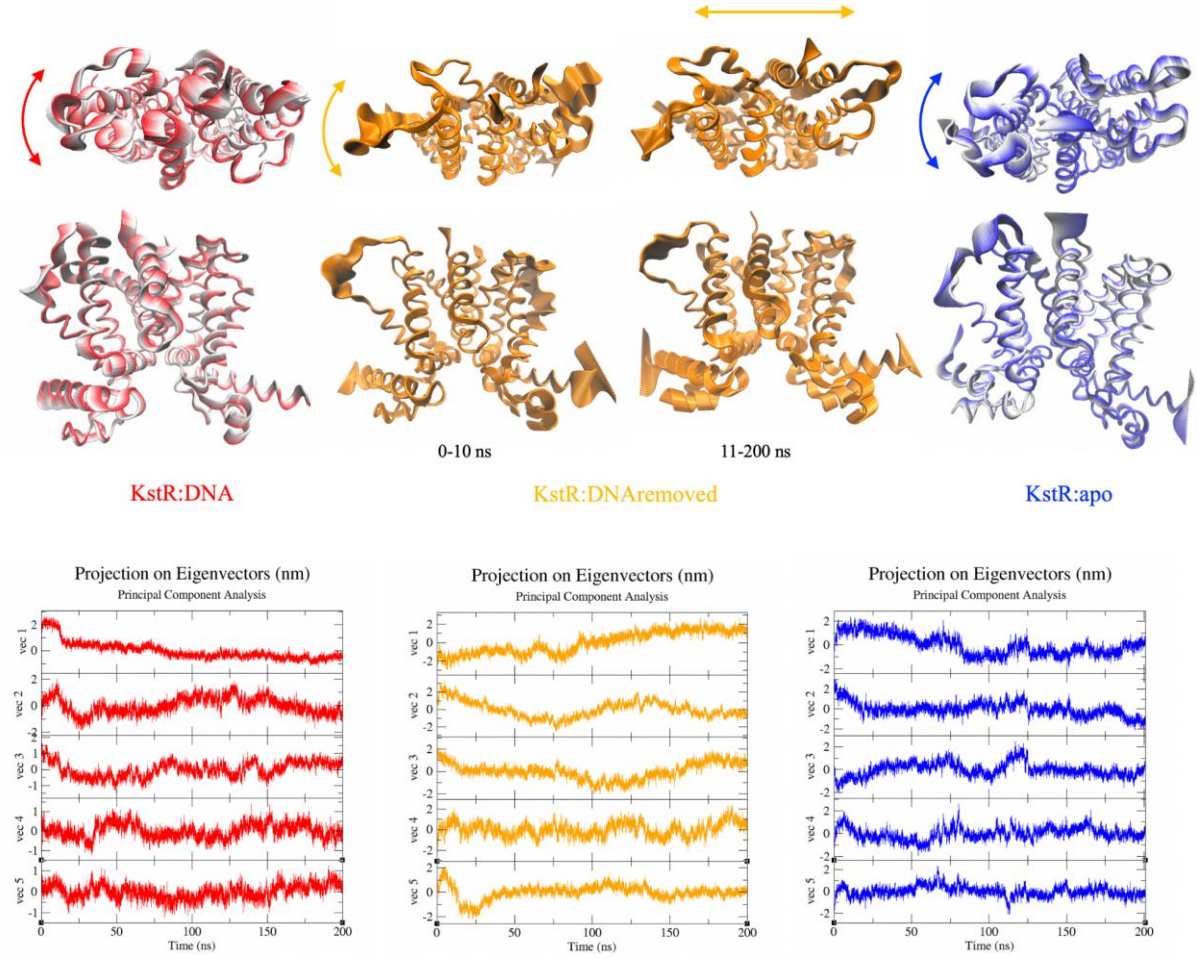


Figure 3. 9. (Top) Visualisation of the extreme motion projected onto the first principal component, where the two extremes are represented by white and by the colour (red) KstR:DNA, (orange) KstR:DNAremoved (0-10 ns and 10-200 ns analysed and displayed separately) and (blue) KstR:apo. (Bottom) Projection onto the first five (largest) eigenvectors during the simulation time for (red) KstR:DNA, (orange) KstR:DNAremoved and (blue) KstR:apo.

3.3.3.4. KstR:3OC vs KstR:3OC:DNA vs KstR:3OC:DNAremoved

KstR:3OC:DNA is highly constrained in terms of motions, as expected due to binding to DNA and as observed in the earlier analyses, and the nature of the motion is similar to that of KstR:DNA, comprising a twisting of the DBDs with respect to the LBDs (Figure 3. 11). However, after removal of DNA, the nature and scale of the motion became more like that of KstR:3OC.

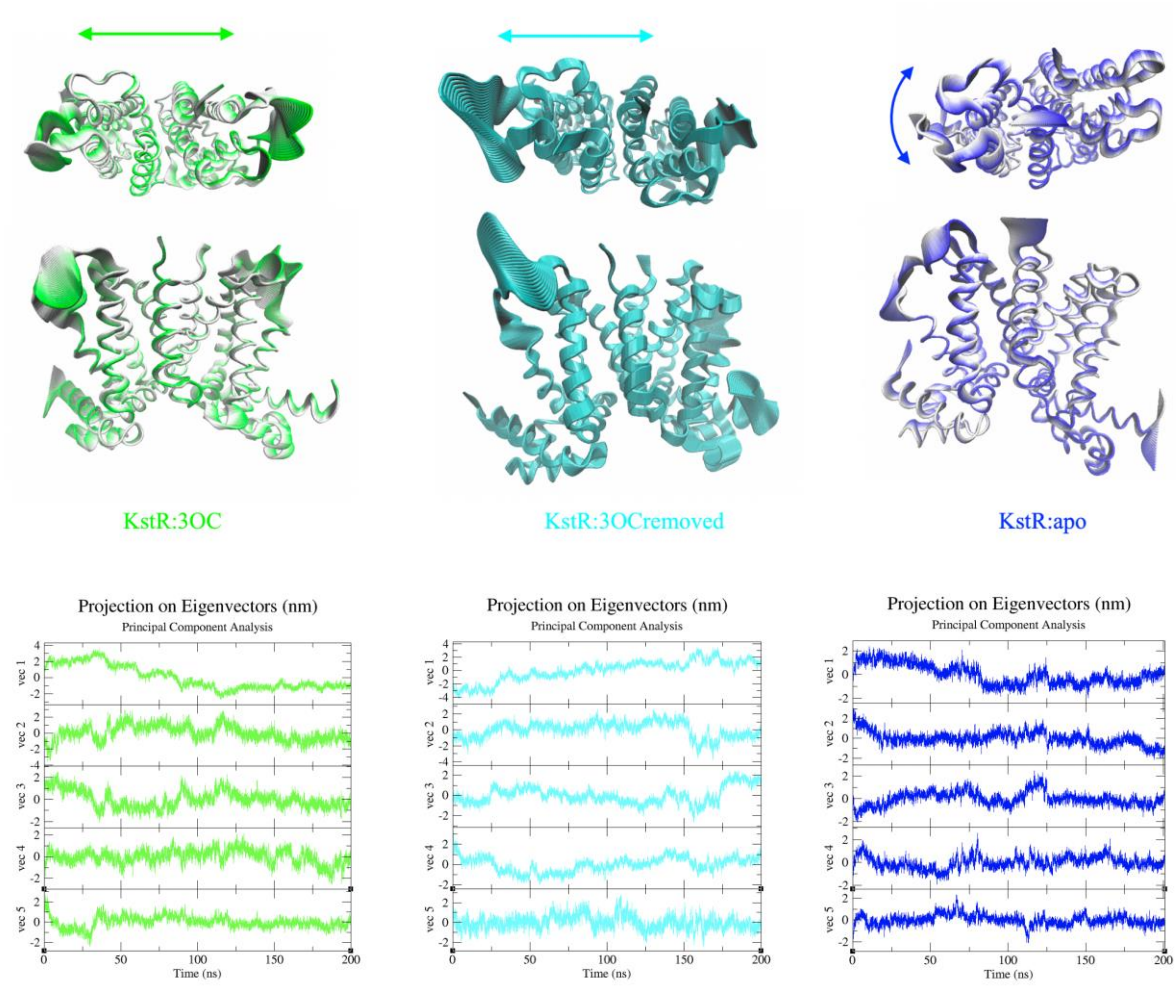


Figure 3. 10. (Top) Visualisation of the extreme motion projected onto the first principal component, where the two extremes are represented by white and by the colour (green) KstR:3OC, (cyan) KstR:3OCremoved and (blue) KstR:apo. (Bottom) Projection onto the first five (largest) eigenvectors during the simulation time for (green) KstR:3OC, (cyan) KstR:3OCremoved and (blue) KstR:apo.

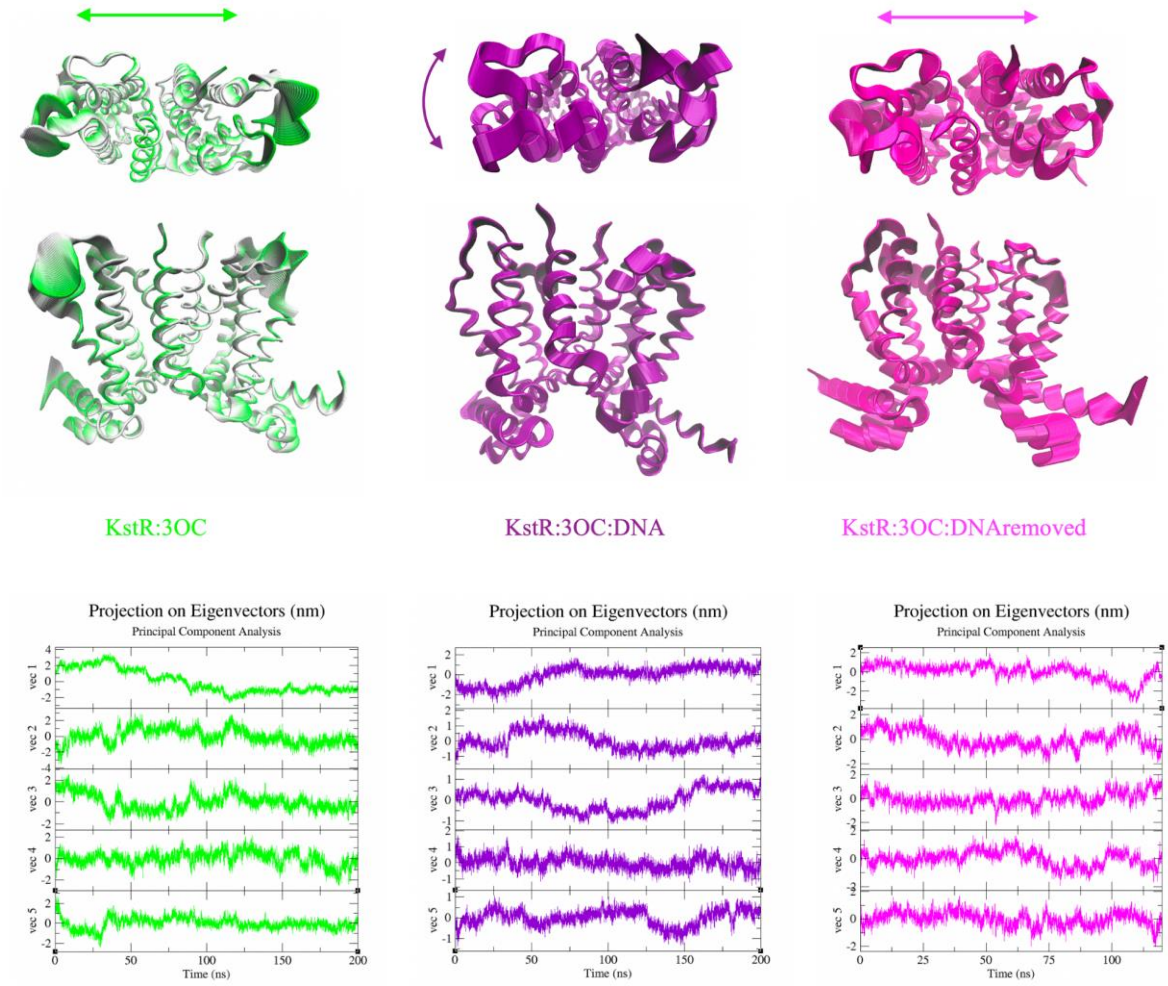


Figure 3. 11. (Top) Visualisation of the extreme motion projected onto the first principal component, where the two extremes are represented by white and by the colour (green) KstR:3OC, (purple) KstR:3OC:DNA and (magenta) KstR:3OC:DNAremoved. (Bottom) Projection onto the first five (largest) eigenvectors during the simulation time for (green) KstR:3OC, (purple) KstR:3OC:DNA and (magenta) KstR:3OC:DNAremoved.

When the inter-DBD distances were analysed for KstR:3OC:DNAremoved (Figure 3. 12, top), there was a noticeable transition point at which the distance between the two TYR54 residues of KstR changed from ~3.6 nm to ~4.2 nm. To investigate the conformational motions accompanying this transition, PCA was performed separately for four parts of the trajectory, 0-120 ns, 121-130 ns, 131-275 ns and 276-300 ns. Visualizing the extreme projection onto the first principal component, for the first part, the motion was a mix of two types, twisty rotation (more like KstR:apo and KstR:DNA) and in/out breathing (more like KstR:3OC). In 121-130 ns of the simulation time, which covers the main transition period in terms of inter-DBD distances, the motion was largely an in/out breathing of the DBDs, however, there was still some slight twisting motion as well. In the third section of the trajectory, the twisting motion was once again dominant, and finally, in the last part of the simulation, the in/out breathing was again the major motion. Generally, it can be assumed that after removing the DNA from

KstR:3OC:DNA, although the components were just KstR and 3OC, the protein still behaved like KstR:DNA. Gradually, there was more KstR:3OC-like behavior, but this was interspersed with a period showing a hybrid of the two types of motions. Since the starting structure (KstR:3OC:DNA) was an unusual form of KstR that may not actually exist outside of simulations, it was expected that it might exhibit some strange behaviour. It was surprising, however, that the transition to KstR:3OC upon removal of DNA took such a convoluted path.

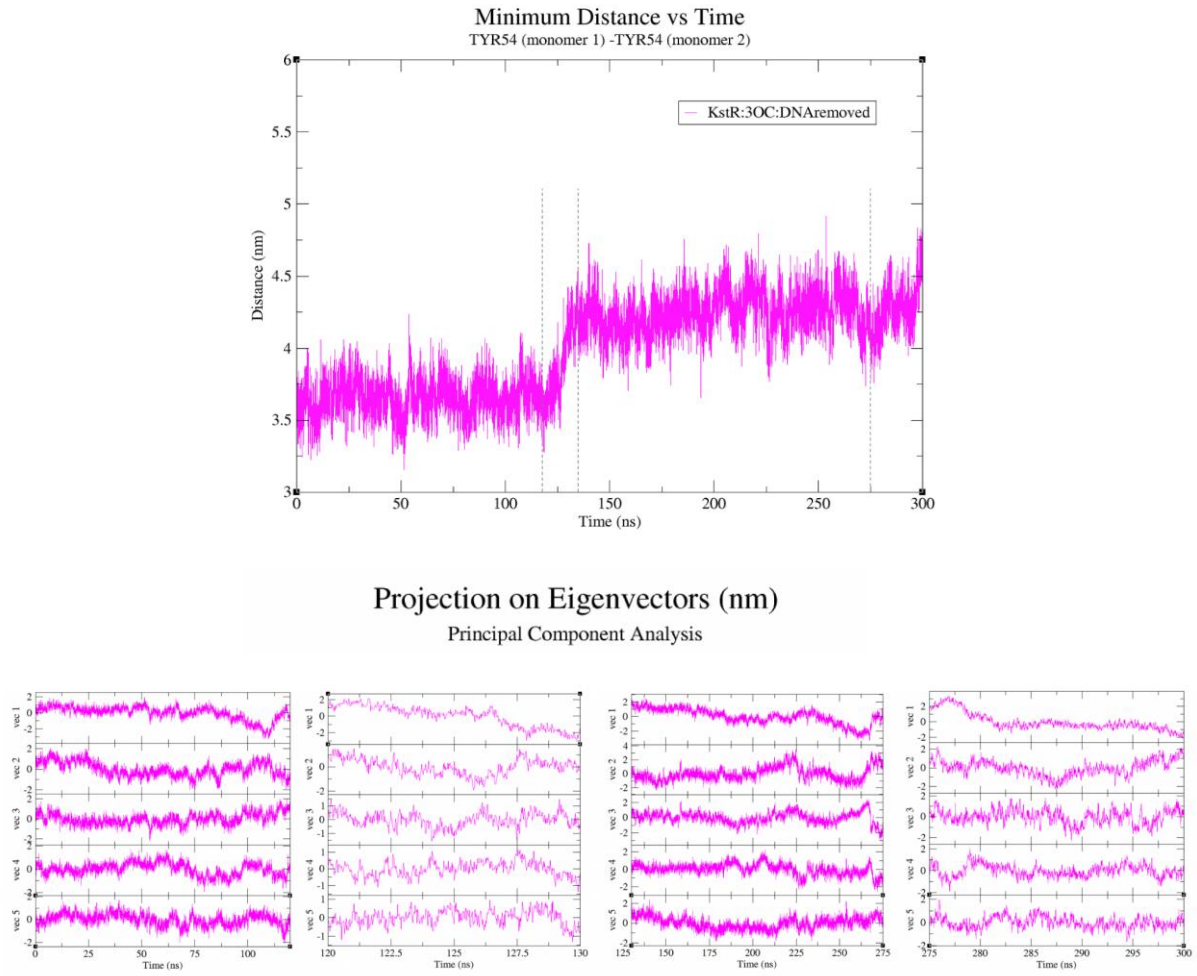


Figure 3. 12. (Top) Minimum distance plot of the alpha carbons of the two TYR54 residues in the two monomers of KstR (nm), for KstR:3OC:DNAremoved 300 ns, (bottom) same minimum distance plot for KstR:3OC:DNAremoved for 300 ns. (Bottom) Projection of the first five largest eigenvectors during the simulation time for KstR:3OC:DNAremoved, for 0-120, 121-130, 131-275 and 276-300 ns of the simulation time.

3.3.4. Dynamical network analysis

The PCA revealed interesting differences in the overall nature of the dynamics of different states of KstR, which can be loosely categorized as either “twisting” or “breathing”. To provide more detailed insight into the origin of these differences and how changes in state (e.g. removal of ligand or DNA) are communicated to other parts of the protein, the VMD NetworkView

plugin was used to compute cross-correlation matrices (Section 3.3.4.1), from which "communities" of residues that undergo correlated motion can be identified (Section 3.3.4.2). This plugin also allows calculation of communication pathways between user-specified pairs of residues (Section 3.3.4.3).

3.3.4.1. Cross correlation maps

The heat maps in Figure 3. 13 show the correlation of each residue to all other residues of KstR during the MD simulations of the seven different states. In all simulations, there are hardly any uncorrelated motions. In most cases, the DBD and, to a lesser extent, the LBD appear to be separate rigid bodies with anti-correlated motion, other than for the DNA-bound states, where this is much less evident. KstR:3OC has the most strongly correlated motions, both when simulated from the ligand-bound crystal structure and after removal of DNA from KstR:3OC:DNA, followed by KstR:DNAremoved and KstR:3OCremoved. Anti-correlated motions are most obvious for KstR:3OC:DNAremoved, followed by KstR:3OC and KstR:DNAremoved. These observations are in keeping with the larger-scale, "breathing" motions observed for these simulations, whereas those that undergo the smaller-scale "twisting" motion show a lower degree of correlated and anti-correlated motion.

3.3.4.2. Communities

Communities are groups of residues with strongly correlated (positively or negatively) motion. These are illustrated in Figure 3. 14. Each community is coloured separately to match as closely as possible the colours used to illustrate the different domains and helices of KstR (Figure 3. 14, top row, left), and connections between residues are weighted according to degree of correlated motion. Across all states and simulations, there are two major communities that are consistently present, DBD (red/orange), LBD (blue/metallic blue). The residues included in each of these vary, however. For instance, in KstR:DNA and KstR:3OC:DNA, and, to a lesser extent, in monomer A of KstR:3OC and KstR:3OCremoved, the DBD community extends partway up the linker helix. This helix, which connects the DBD to the LBD, has been hypothesised to be important in communicating ligand and DNA binding between the two domains (143).

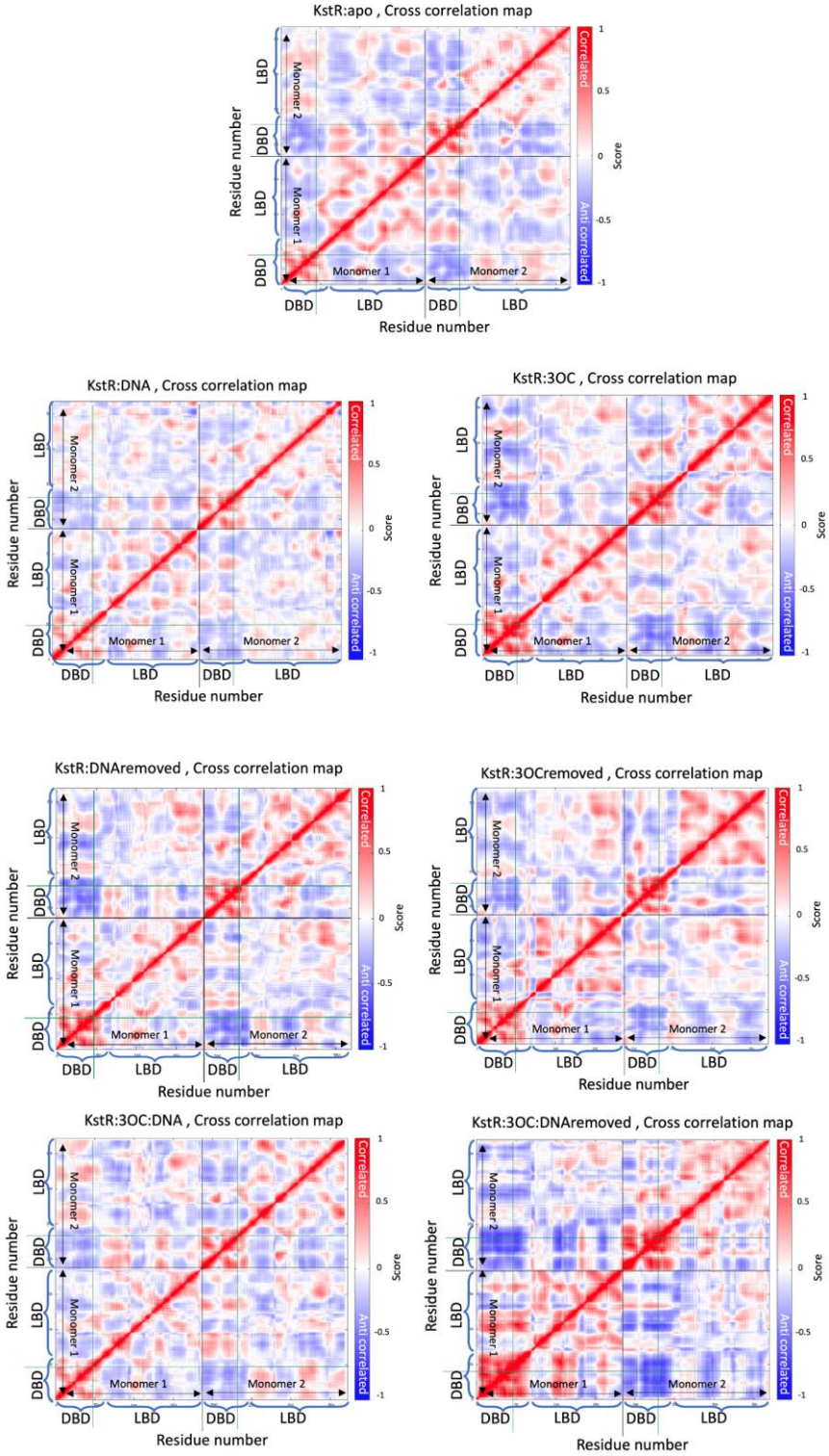


Figure 3. 13. Cross correlation maps for MD simulations of (top row) KstR:apo, (second row, left) KstR:DNA, (second row, right) KstR:3OC, (third row, left) KstR:DNAremoved, (third row, right) KstR:3OCremoved, (bottom row, left) KstR:3OC:DNA and (bottom row, right) KstR:3OC:DNAremoved). In the colour scale, red represents correlated motion and blue represents anti-correlated motion. The residues corresponding to each monomer and to the DBD and LBD are indicated on each heat map. Green lines indicate the position of TYR54 (note that residue numbering here starts from 1 but in crystal structure it starts from 15).

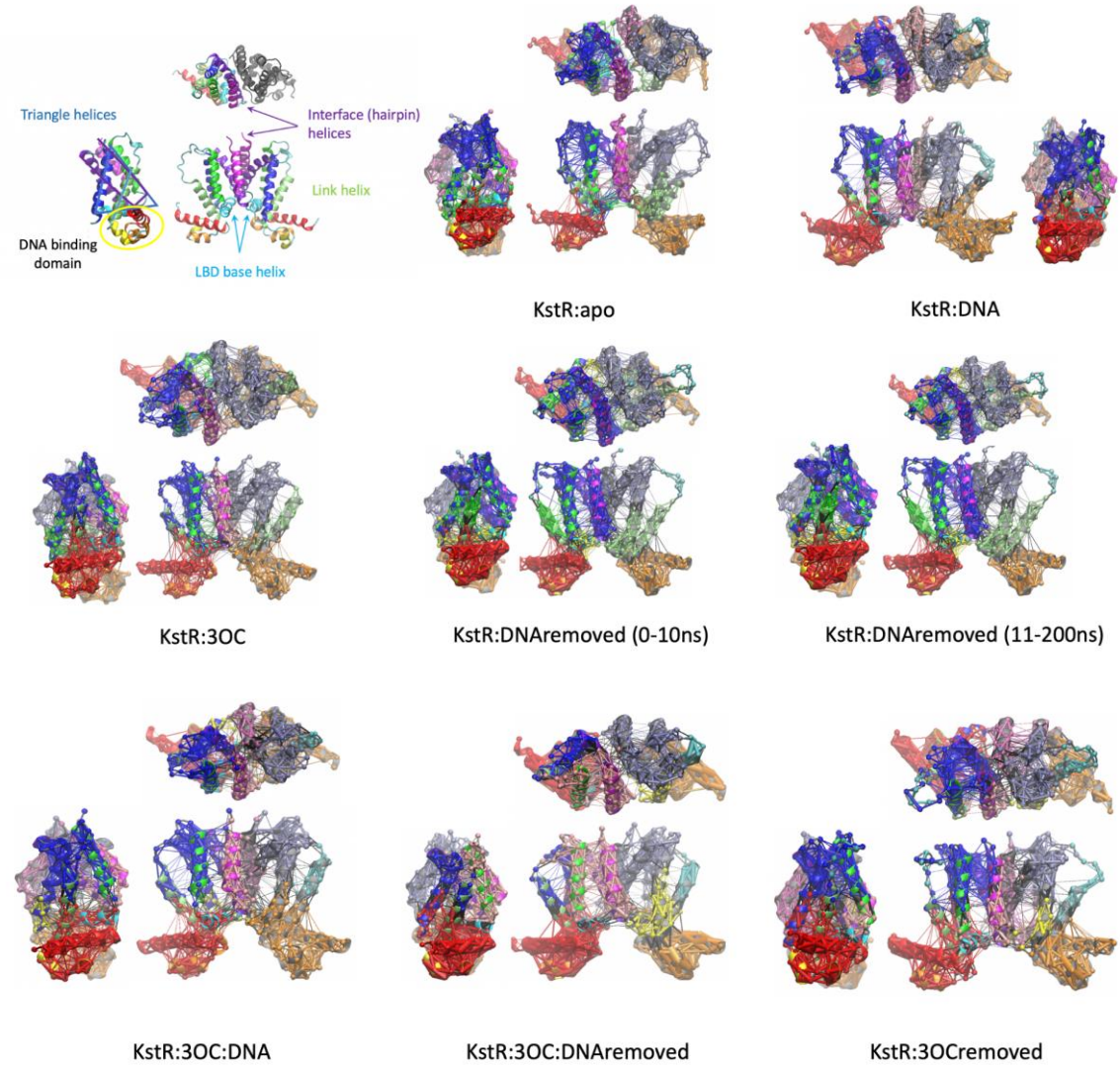


Figure 3. 14. (Top row, left) KstR structure (PDB ID: 3MNL) drawn in cartoon format with each helix coloured differently but with corresponding colours for the equivalent helix in each monomer. Key features discussed in the text are labelled. The remainder of the panels show the communities of residues that undergo correlated motions, viewed from the side, top and front, for (top row, centre) KstR:apo, (top row, right) KstR:DNA, (middle row, left) KstR:3OC, (middle row, centre) KstR:DNAremoved (0-10 ns), (middle row, right) KstR:DNAremoved (11-200), (bottom row, left) KstR:3OCremoved, (bottom row, centre) KstR:3OC:DNA, (bottom row, right) KstR:3OC:DNAremoved. Each community is coloured separately to match as closely as possible the colours used to illustrate the different domains and helices of KstR (as per the image at the top left), with brighter colours used for monomer A. The connections between residues are weighted according to the degree of correlated motion; the thicker the edges are, the stronger the connections and vice versa.

Here, it only forms a separate community for monomer B of KstR:apo, and for both monomers for KstR:DNAremoved. For KstR:3OC and KstR:3OC:DNA, the linker helix was split between DBD and LBD communities, whereas for KstR:DNA, it was coupled to the DBD, suggesting that adding 3OC to the KstR:DNA structure perturbed the dynamics of the DNA-bound state. The LBD base helix was expected to be part of the LBD community. However, the community analysis revealed this helix to be part of the DBD community, other than for

KstR:DNAremoved, when it moved separately as its own community. For KstR:3OC, the interface community (pink/salmon) is only present for one monomer and is absorbed into the LBD community for the other. For KstR:DNAremoved, this occurs for both monomers. Upon removing 3OC (KstR:3OCremoved), however, the interface reverts to being a separate community, and grew and even absorbed the residues in the upper parts of the LBD. The large loops connecting helix α4 to helix α5 are generally included in the LBD community, but for monomer B of KstR:DNA the loop forms its own, separate, community, and this separation is retained in KstR:DNAremoved, and also shows up in KstR:3OCremoved. This loop is the most flexible region of KstR according to the RMSF (Figure 3. 4) and PCA (Figure 3. 11) analysis.

Overall, a key take-home message from the community analysis is that the KstR communities are not always symmetric, that is, both the number of communities and the residues that comprise them are seldom the same (even at an approximate level) between the two monomers.

3.3.4.3. Communication pathways between DBD and LBD

Identifying pathways of residues through which events such as ligand or DNA binding are communicated are crucial to understanding allosteric mechanisms. One way to define two residues as communicating is if their motion is correlated, with a higher degree of correlation corresponding to greater communication. The shortest pathway between two residues is then a combination of it comprising few intervening residues and their motion being highly correlated. This is the definition used by the VMD NetworkView plugin (225). It was used to identify putative communication pathways between the ligand-binding and DNA-binding domains of KstR for each of the seven states simulated here. Residue ARG158 was chosen to represent ligand binding, as this is the residue in the LBD which binds to the –OH group of the cholesterol degradation products. TYR54 was selected to represent DNA binding, one of the key residues by which KstR binds to DNA. More details about the interaction of TYR54 and DNA are provided in the hydrogen bond analyses in Section 3.3.5.1.

KstR:apo

For KstR:apo, the DBD ends of the communication paths involve the same residues for both monomers, but the LBD ends of the paths are slightly different (Figure 3. 15). In both optimal paths, however, the positively charged ARG158 links to a negatively charged residue. For monomer A (Figure 3. 15, left), all of the communication pathways link to ARG158 via

ASP137. All the displayed paths then run down helix α7 (blue) towards the DBD, albeit through different residues. The stable secondary structure and vertical hydrogen bond patterns of helices mean that communication pathways do not need to pass through all residues in a helix. Instead, communication can pass along the hydrogen bond network and so rapidly traverse the length of a helix. At the base of the α7 helix, the optimal path passes through ALA125 to HIS62 in the α3 helix (green), along α3 to PRO58 (orange helix), and finally to TYR54 (α2 helix, orange), and the suboptimal paths follow a similar route. For monomer B, there is a greater diversity of communication pathways. While the optimal path crosses from ARG158 to GLU133, further along helix α7 towards the DBD than in monomer A, the best suboptimal path passes through ASP 137. There are additional suboptimal paths that cross to α7 even further from the DBD, as well as some that connect ARG158 directly to helix α6, at the base of the LBD. At the DBD there is less diversity, with the optimal path traversing through the same residues as for monomer A.

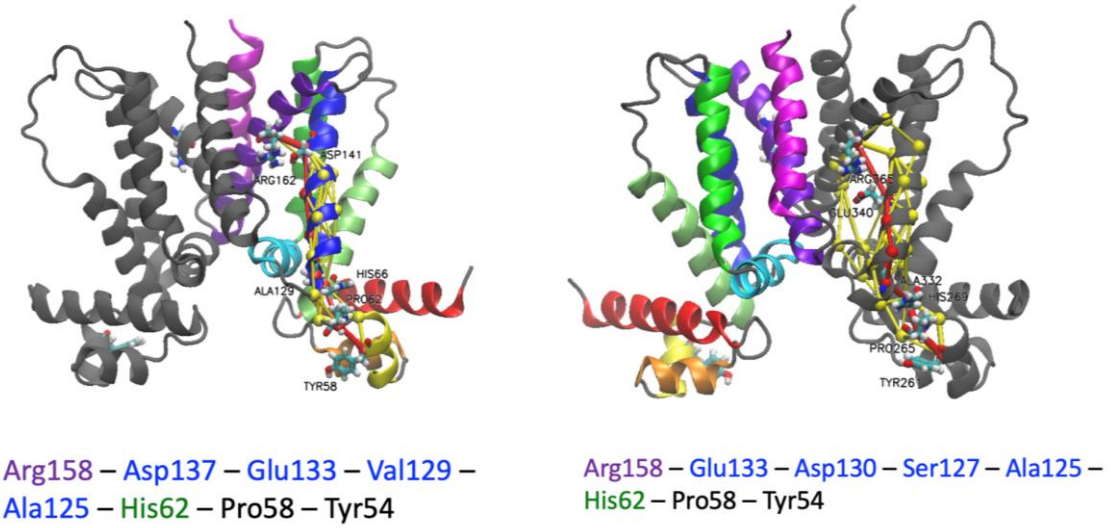


Figure 3. 15. Optimal (red) and suboptimal (yellow) communication pathways for KstR:apo. The coloured chain represents monomer A (colour coding of the helices the same as in Figure 17) and the grey chain is monomer B. Key residues that make connections between helices along these paths are drawn in licorice form, coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen) and numbered according to the coordinate file. These numbers are offset by 4 for monomer A and are continuous for the two monomers. The true residue numbers are provided in the list of key residues below the structure, which are coloured according to the helix in which they lie, other than helix α2 (orange) and α3 (yellow), for which the residues are coloured grey and black, respectively.

KstR:DNA

For monomer A of KstR:DNA (Figure 3. 16, left), the optimal pathway is similar to that of monomer A in KstR:apo, running from ARG158 to ASP137, down helix α7 via LEU135 to VAL132, and across to VAL65 and then HIS62 in helix α4. Subsequently, it goes down to SER59 in the loop and TYR54 in helix α2 of the DBD. Interestingly, the suboptimal paths largely follow the same route through the LBD, with redundancy only emerging in the DBD. This may reflect the constrained motion of KstR when bound to DNA. The optimal path in monomer B (Figure 3. 16, right) goes from ARG158 to SER161 then TRP162, also in helix α8, omitting helix α7 altogether. Subsequently, it steps across to PHE70, ARG68 and SER64 in the linker helix (α4, green), and finally down to PHE55 and TYR54 in helix α3 (yellow). The suboptimal paths do involve helix α7, however, and exhibit substantial redundancy in how they connect the LBD to the DBD. Overall, there is not much similarity between the communication pathways, and in particular between the two optimal paths, for the two monomers of KstR:DNA. This is somewhat surprising given that its motion should be constrained by it being bound to DNA, and suggests that DNA binding may enforce asymmetry.

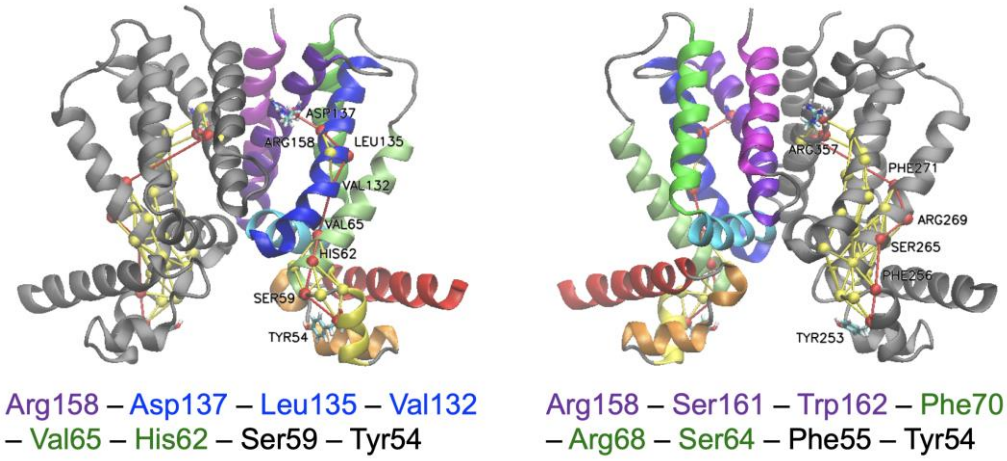


Figure 3. 16. Optimal (red) and suboptimal (yellow) communication pathways for KstR:DNA. The coloured chain represents monomer A (colour coding of the helices the same as in Figure 17) and the grey chain is monomer B. Key residues that make connections between helices along these paths are numbered, and the start and end residues are drawn in licorice form, coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen). The list of residues involved in the optimal path below each structure are coloured according to the helix in which they lie, other than helix α2 (orange) and α3 (yellow), for which the residues are coloured grey and black, respectively.

KstR:3OC

For KstR:3OC, the two monomers have almost identical optimal paths (Figure 3. 17). In both cases, the optimal path starts from ARG158 in helix α8 (purple), links to GLU133 in helix α7 (blue), then runs down to VAL129 within the blue helix, before linking across to VAL65 and then to HIS62 (both α4, green), and lastly running through PRO58 (monomer A) or SER59 in the loop (monomer B) to TYR54 (α3, yellow). For both monomers, there is some redundancy around the links from helix α8 to α7, the pathway down α7, and within the DBD. For monomer A, there is also a suboptimal path that runs down helix α8 to α6 (cyan).

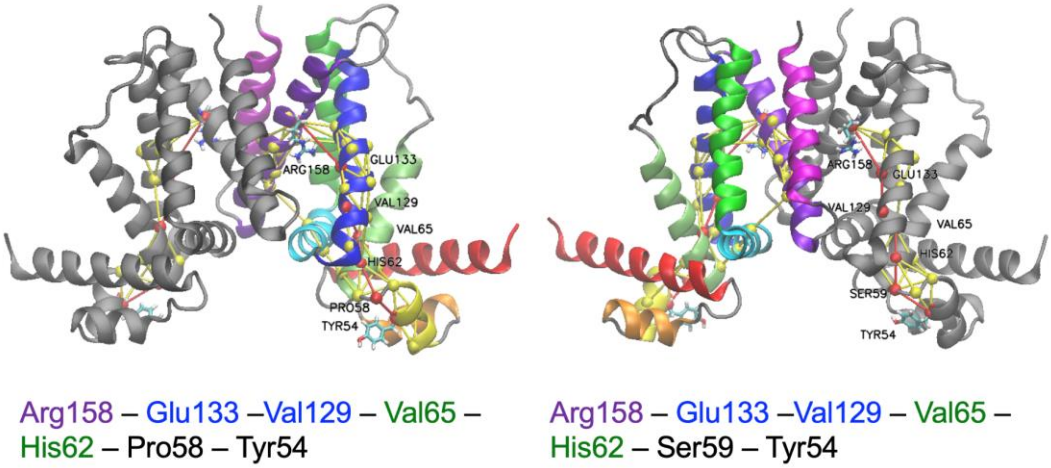


Figure 3. 17. Optimal (red) and suboptimal (yellow) communication pathways for KstR:3OC. The coloured chain represents monomer A (colour coding of the helices the same as in Figure 17) and the grey chain is monomer B. Key residues that make connections between helices along these paths are numbered, and the start and end residues are drawn in licorice form, coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen). The list of residues involved in the optimal path below each structure are coloured according to the helix in which they lie, other than helix α2 (orange) and α3 (yellow), for which the residues are coloured grey and black, respectively.

KstR:DNA vs KstR:apo vs KstR:3OC

Comparing the communication pathways of KstR:apo, KstR:DNA and KstR:3OC, KstR:DNA has more asymmetric pathways, and KstR:apo and KstR:3OC have more symmetric pathways, especially KstR:3OC, where the optimal paths in the two monomers are almost identical. It is difficult to explain this difference in terms of the other analyses carried out previously in this Chapter. When KstR is bound to DNA or ligand, the communication pathways run through the linker helix (α4), while in apo form of KstR, the paths jump the helix α7 directly across to the DBD. This suggests that the linker helix may indeed play a role in allosteric communication between the LBD and DBD, as has been hypothesized (143).

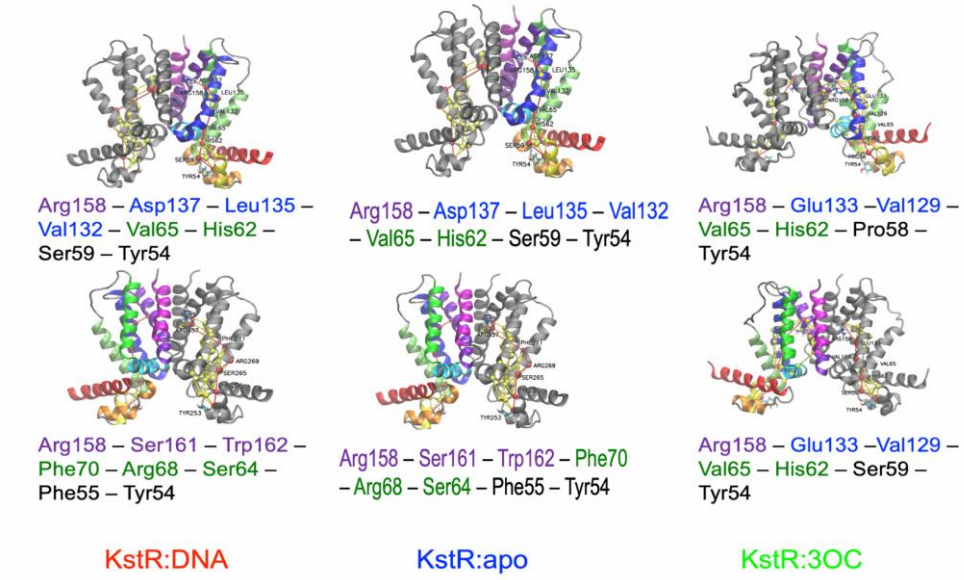


Figure 3. 18. Comparison of the optimal (red) and suboptimal (yellow) communication pathways for KstR:DNA, KstR:apo and KstR:3OC. Monomer A paths are shown in the top row and monomer B paths in the bottom row. See Figures 3. 17-19 for detailed descriptions.

KstR:DNAremoved

For KstR:DNAremoved, two sets of communication pathway analysis were performed, for 0-10 ns and 11-200 ns of the simulated trajectory, before and after the transition in TYR54-TYR54 distance plot (Figure 3. 7, Section 3.3.2.2). For the first 10 ns of the MD simulation, the optimal pathway for monomer A (Figure 3. 19, top left) goes from ARG158 in helix α8 to ASP137 of helix α7, runs down this helix to VAL132, then connects to VAL65 and subsequently to HIS62, both in the linker helix (α4), and lastly goes across to PHE47 and TYR54, both in helix α3. There is little redundancy in the communication pathways, other than one alternative connection to helix α7 and two alternative paths to α3. In monomer B (Figure 3. 19, top right), the optimal path is similar, running from ARG158 (α8) to LYS134 and VAL132 in helix α7 to GLU128 in the linker helix, LEU262 in helix α1 (red), then to PHE57 and at last to TYR54, both in helix α3. This monomer exhibits greater redundancy of communication pathways, largely at the lower end of α7 and in the DBD. The monomers' communication pathways are generally similar, and the asymmetry of the DNA-bound state appears to be rapidly lost.

For the 11-200 ns portion of the simulation, however, the two monomers revert to asymmetric optimal communication pathways, with only the final three residues the same. For monomer A (Figure 3. 19, bottom left), the optimal pathway resembles that of monomer B of KstR:DNA, running from ARG158 to SER161 and then to LEU165, all in helix α8, before connecting to TYR118 in the base helix of the LBD (α6). TYR118 in turn is linked to VAL65

and then to HIS62, both in helix α4, which are connected to PRO58 and finally to TYR54 in helix α3. In monomer B (Figure 3. 19, bottom right), the optimal path jumps directly from ARG158 to LYS134 in helix α7 and then runs through VAL132 (also α7) and HIS62 (α4) to PRO58 and TYR54 (α3), similar to the optimal paths observed for KstR:apo, KstR:3OC, and monomer A of KstR:DNA.

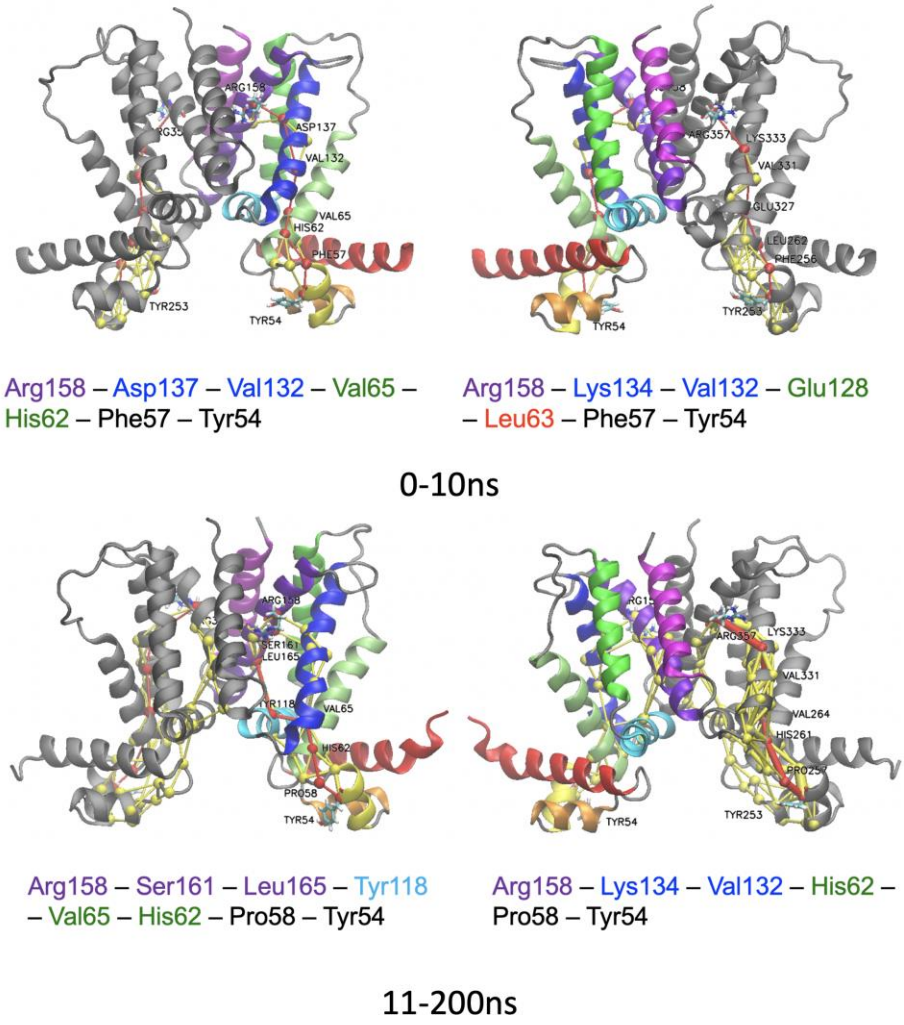


Figure 3. 19. Optimal (red) and suboptimal (yellow) communication pathways for KstR:DNAremoved: (top row) 0-10 ns and (bottom row) 11-200 ns. The coloured chain represents monomer A (colour coding of the helices the same as in Figure 17) and the grey chain is monomer B. Key residues that make connections between helices along these paths are numbered, and the start and end residues are drawn in licorice form, coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen). The list of residues involved in the optimal path below each structure are coloured according to the helix in which they lie, other than helix α2 (orange) and α3 (yellow), for which the residues are coloured grey and black, respectively.

KstR:DNA vs KstR:DNAremoved vs KstR:apo

In KstR:DNA, the upper parts of the optimal pathways communicating ligand and DNA binding are asymmetric. After removing the DNA, the pathways initially become more symmetric, but then revert to strong asymmetry. The pathway that misses out helix α7 switches from monomer B for KstR:DNA to monomer A for the 11-200 ns portion of KstR:DNAremoved (Figure 3. 20). These results suggest that after removal of DNA, the behavior of the protein has not fully reverted back to an apo-like state, even after 200 ns of simulation. This is in keeping with the quite different dynamics observed for KstR:DNAremoved compared to KstR:apo in the PCA.

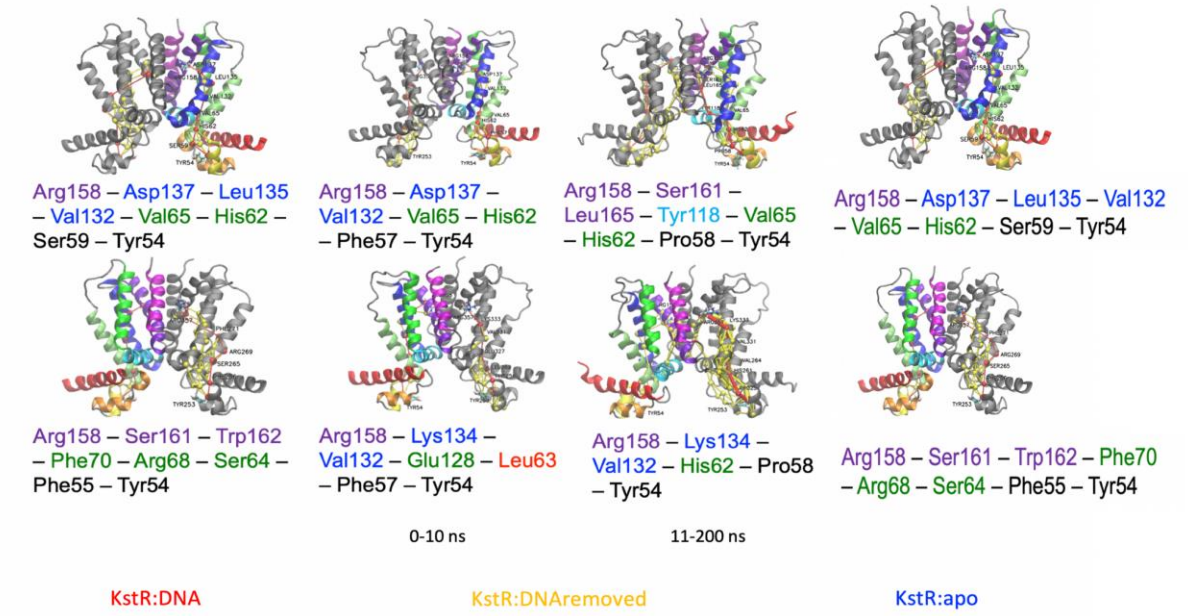


Figure 3. 20. Comparison of the optimal (red) and suboptimal (yellow) communication pathways for KstR:DNA, KstR:DNAremoved and KstR:apo. Monomer A paths are shown in the top row and monomer B paths in the bottom row. See Figures 3. 17, 18 and 21 for detailed descriptions.

KstR:3OCremoved

When ligand was removed from KstR:3OC, the symmetry of KstR:3OC is lost. The optimal pathway for monomer A (Figure 3. 21, left) runs fromARG158 to GLU133, then VAL129 and ALA125, all in helix α7, across to HIS62 in the linker helix (α4), and subsequently to PHE57 and ultimately TYR54, both in helix α3. The suboptimal paths follow a similar trajectory, following different routes down helix α7 and through the DBD to TYR54. In contrast, in monomer B (Figure 3. 21, right), the path jumps from ARG158 out to PHE72 and then GLY69 in the linker helix (α4), then straight down to PHE57 and then TYR54 in helix α3. There is again a range of suboptimal paths, all of which follow the same overall route but flow through

different residues in helices α3 and α4. The top part of the monomer B optimal path is unique among all the other simulated states of KstR, while that of monomer A is similar to those of KstR:apo, KstR:3OC, and monomer A of KstR:DNA.

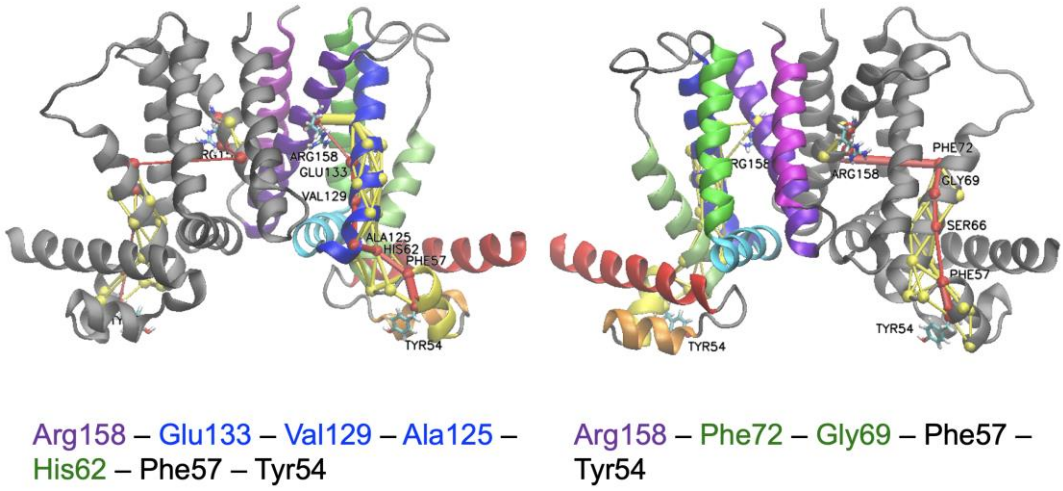


Figure 3. 21. Optimal (red) and suboptimal (yellow) communication pathways for KstR:3OCremoved. The coloured chain represents monomer A (colour coding of the helices the same as in Figure 17) and the grey chain is monomer B. Key residues that make connections between helices along these paths are numbered, and the start and end residues are drawn in licorice form, coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen). The list of residues involved in the optimal path below each structure are coloured according to the helix in which they lie, other than helix α2 (orange) and α3 (yellow), for which the residues are coloured grey and black, respectively.

KstR:3OC vs KstR:3OCremoved vs KstR:apo

Comparing the optimal pathways of the three states, after removing the ligand from KstR:3OC, the symmetry of the communication pathways was lost, with the optimal path for monomer A remaining the same while that of monomer B changed completely to a shorter path unique amongst the KstR systems studied here, although bearing some resemblance to monomer B of KstR:DNA. Like for KstR:DNAremoved, removal of 3OCh does not revert KstR to a state with communication pathways resembling those of KstR:apo.

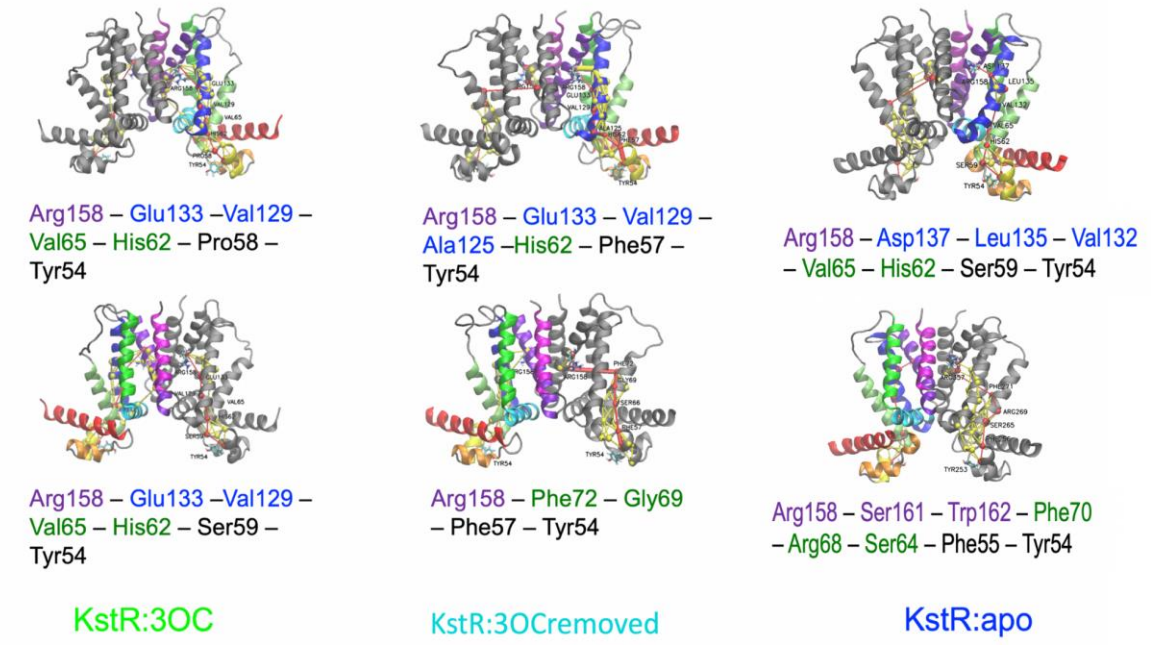


Figure 3. 22. Comparison of the optimal (red) and suboptimal (yellow) communication pathways for KstR:3OC, KstR:3OCremoved and KstR:apo. Monomer A paths are shown in the top row and monomer B paths in the bottom row. See Figures 3. 17, 19 and 23 for detailed descriptions.

KstR:3OC:DNA

When KstR is bound to DNA and ligand simultaneously, the optimal pathway of monomer A (Figure 3. 23, left) runs from ARG158 to SER161 and then to LEU165, all in helix α8, and then drops directly to TYR118 in the LBD base helix α6. The path then continues with VAl65 and then HIS 62 in the linker helix (α4), PHE57 and finally TYR54, both in helix α3. In monomer B (Figure 3. 23, right), the optimal path runs from ARG158 through ASP162 and LEU165, all in helix α8, down to the VAL119 and then to ALA117, both in the LBD base helix (α6). It continues with VAL61 in the linker helix (α4) linking to SER59 in the loop, then ends with TYR54 in helix α3. The paths in two monomers are highly similar, and resemble that seen for monomer A, KstR:DNAremoved 11-200ns.

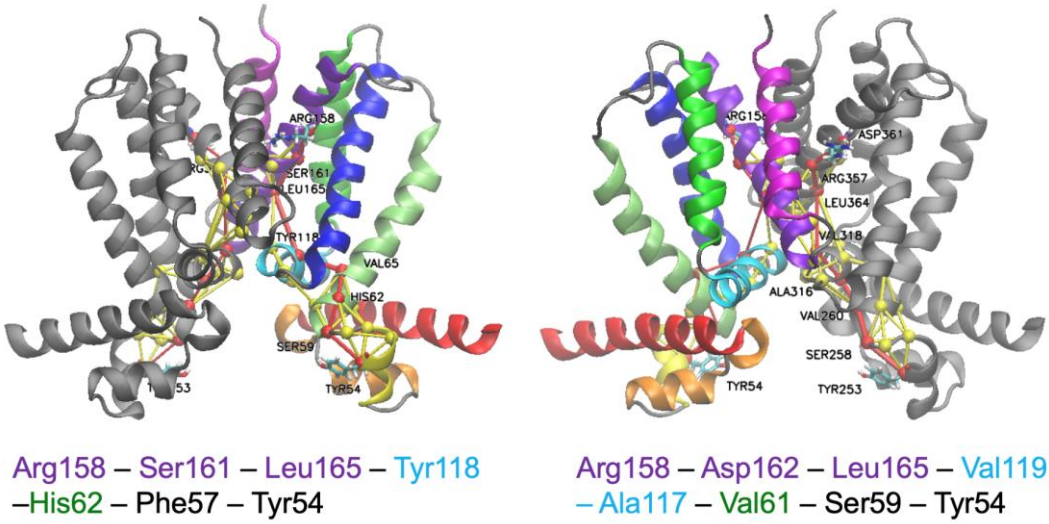


Figure 3. 23. Optimal (red) and suboptimal (yellow) communication pathways for KstR:3OC:DNA. The coloured chain represents monomer A (colour coding of the helices the same as in Figure 17) and the grey chain is monomer B. Key residues that make connections between helices along these paths are numbered, and the start and end residues are drawn in liquorice form, coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen). The list of residues involved in the optimal path below each structure are coloured according to the helix in which they lie, other than helix α2 (orange) and α3 (yellow), for which the residues are coloured grey and black, respectively.

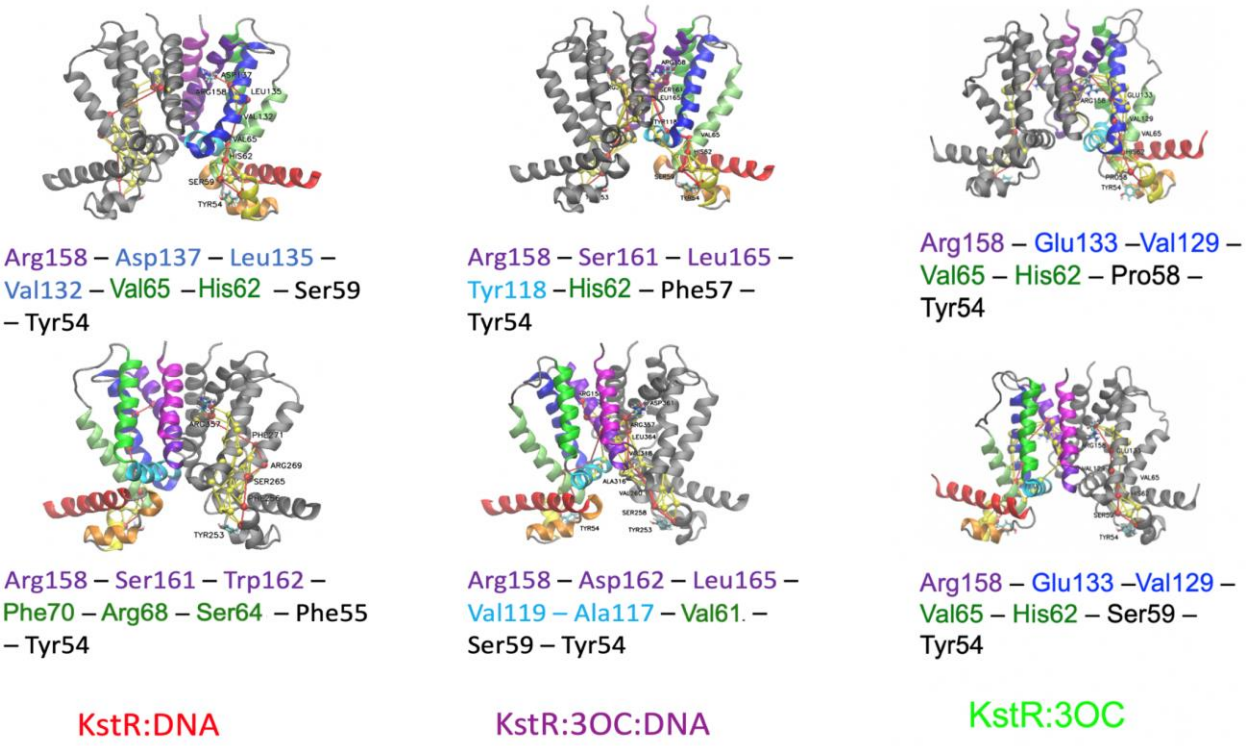


Figure 3. 24. Comparison of the optimal (red) and suboptimal (yellow) communication pathways for KstR:DNA, KstR:3OC:DNA and KstR:3OC. Monomer A paths are shown in the top row and monomer B paths in the bottom row. See Figures 3. 18, 19 and 25 for detailed descriptions.

KstR:DNA vs KstR:3OC:DNA vs KstR:3OC

Comparing KstR:3OC:DNA with KstR bound to DNA or to ligand separately (Figure 3. 24) shows that binding both of its ligands at once forces communication between the LBD and DBD of KstR to go via pathways that are different to those observed for either of the individual bound states. This is in keeping with the strained and unusual nature of KstR:3OC:DNA observed in the previous analyses.

KstR:3OC:DNAremoved

After removing the DNA from KstR:3OC:DNA, the most optimal path for monomer A (Figure 3. 25, left) runs from ARG158 to ASP137, LEU135 and VAL132, all in helix α7, then to VAL65 and then to LEU63, both in the linker helix, and ends at TYR54 (α3). This pathway resembles the "standard" pathway observed for KstR:3OC, KstR:apo and monomer A of KstR:DNA. In monomer B (Figure 3. 25, right), the optimal path is like those seen for KstR:3OC:DNA, running from ARG158 through SER161, TRP164 and LEU168, all in helix

α8, then jumps to THR115 and ALA117 in the LBD base helix (α6), and lastly passes through VAL61 in the linker helix and SER59 in the loop to end at TYR54.

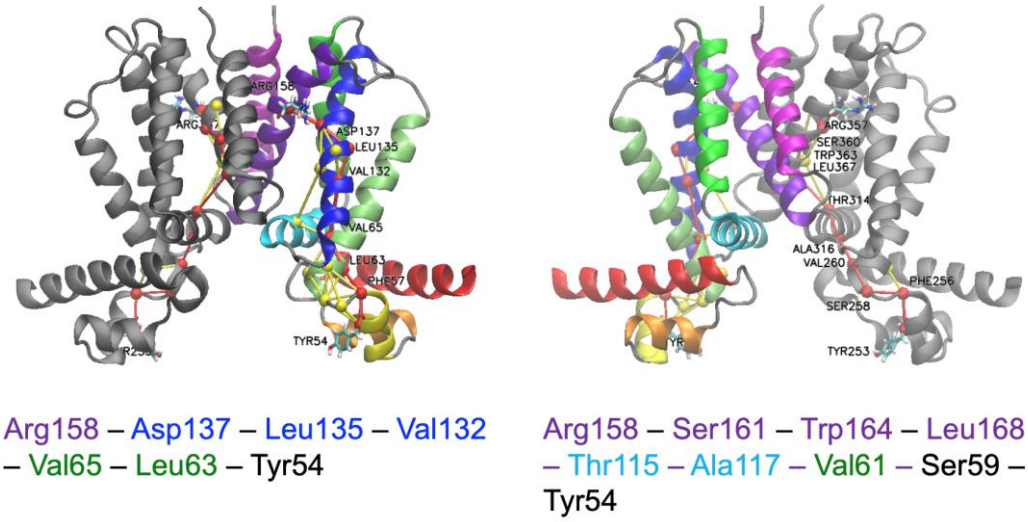


Figure 3. 25. Optimal (red) and suboptimal (yellow) communication pathways for KstR:3OC:DNAremoved. The coloured chain represents monomer A (colour coding of the helices the same as in Figure 17) and the grey chain is monomer B. Key residues that make connections between helices along these paths are numbered, and the start and end residues are drawn in liquorice form, coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen). The list of residues involved in the optimal path below each structure are coloured according to the helix in which they lie, other than helix α2 (orange) and α3 (yellow), for which the residues are coloured grey and black, respectively.

KstR:3OC vs KstR:3OC:DNA vs KstR:3OC:DNAremoved

When removing DNA from KstR:3OC:DNA, monomer A had a communication pathway similar to those of KstR:3OC, but monomer B still resembled the KstR:3OC:DNA pathways (Figure 3. 26).

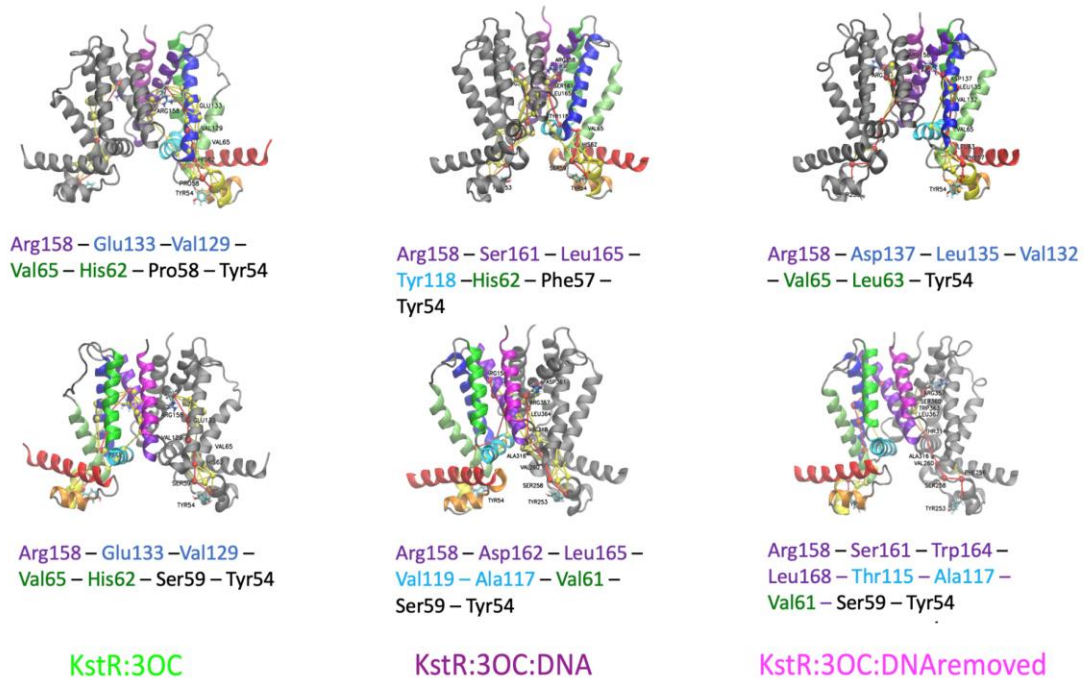


Figure 3. 26. Comparison of the optimal (red) and suboptimal (yellow) communication pathways for KstR:3OC, KstR:3OC:DNA and KstR:3OC:DNAremoved. Monomer A paths are shown in the top row and monomer B paths in the bottom row. See Figures 19, 25 and 27 for detailed descriptions.

Conclusion of communication pathways

In general, there are many similarities between the optimal pathways of different states of KstR. Three major types of pathways were identified, summarised here according to the helices they traverse: 1) α8 (purple) – α7 (blue) – α4 (green) –α3 (yellow); 2) α8 (purple) – α4 (green) – α3 (yellow); 3) α8 (purple) – α6 (cyan) – α3 (yellow). The key residues involved in each of the three types of pathways are largely conserved, but with some variation. The fact that the pathways differ between states suggests that communication of DNA and ligand binding is state-dependent. Interestingly, the optimal pathways were highly asymmetric across the KstR monomers when bound to DNA, but most symmetric when bound to 3OCh. Upon removal of DNA, the symmetry was initially restored, but after 10 ns, the pathways became asymmetric again. Upon removal of 3OCh, the pathways became asymmetric. KstR:apo, however, exhibits essentially symmetric communication pathways. Thus once again, in keeping with the earlier analyses, the apo states sampled after removal of DNA or 3OCh differ from the apo state initiated from the apo crystal structure. Helix α4 (green, linker) has been hypothesised to be involved in communication between the DBD and LBD. Here, only one type of pathway, which occurred in one monomer when DNA was bound to KstR or when DNA or 3OCh were

removed, involved more than one residue of this helix, suggesting that it may predominantly be involved in communicating changes of state.

3.3.5. Hydrogen bond formation

While the communication pathways, which comprise residues whose motion is strongly coupled, report on information transfer through the protein, the formation of hydrogen bonds is crucial for stable interactions. The hydrogen bonds formed during the simulations between monomers and between KstR and its ligands and DNA were therefore computed, as described below.

3.3.5.1. Hydrogen bonds between KstR and DNA

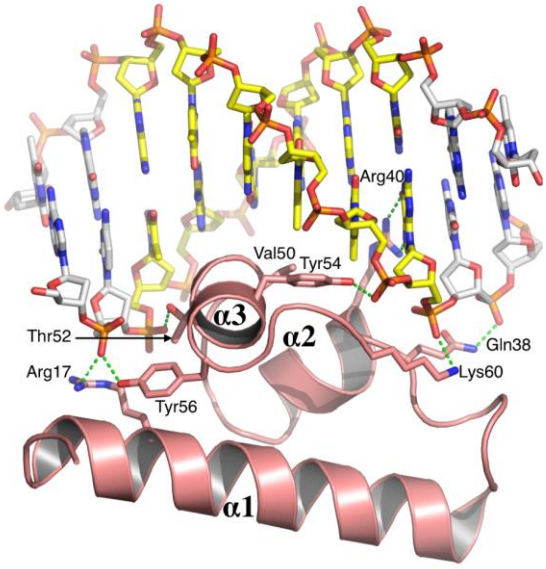


Figure 3. 27. Hydrogen bonds between KstR and DNA observed in the X-ray crystal structure (PDB ID: 5UA2, (135, 142)**). The protein is drawn in cartoon format and coloured pink, with residues making side-chain contact with the DNA shown in liquorice representation. The DNA is drawn in liquorice representation, with one strand in yellow and the other in white. Oxygen atoms are in red, nitrogen atoms in blue and phosphate atoms in orange. Hydrogen bonds are shown as green dashed lines. Note that this image is also provided in Figure 1. 22 in the Introduction, but is repeated here for easy reference.**

The hydrogen bonds formed between KstR and DNA have been studied in detail for the X-ray crystal structures (135, 142) (Figure 3. 27). KstR has been crystallized in complex with two different DNA duplexes with lengths of 18 bp and 26 bp. The structure in complex with the 26 bp DNA duplex was used for the simulations presented here. However, the first and the last bases were not resolved in the X-ray crystal structure and were not built prior to running the simulation as they are distant from KstR. Therefore, the residue numbering in each DNA strand starts from 2 and ends at 25. Figure 3. 27 shows the interaction of KstR with the central

recognition sequence of DNA. The base numbers for which experimental data is reported (142) refer to this recognition sequence, and are different from the base numbers in the crystal structure from which the simulations were initiated (PDB ID: 5UA2) and thus from the base numbers reported in the hydrogen bond analysis.

The interaction between KstR and DNA mainly involves residues in the α3 helix, some residues in the α2 helix, and the N-terminus of the α4 helix (135, 142). The side chains of residues ARG17, THR52, and TYR56 and backbone nitrogen of ALA49 form hydrogen bonds with the phosphate groups of C and T on one side of the major groove (Figure 3. 27). On the other side, hydrogen bonds are formed around phosphate groups C, G, and T with the side chains of GLN38, TYR54 and side chain and backbone of LYS60. The only hydrogen bonds between KstR and a DNA base are the two between the side chain of ARG40 and a guanidine. Presence of all of these hydrogen bonds during a simulation would indicate tight binding between KstR and DNA (142).

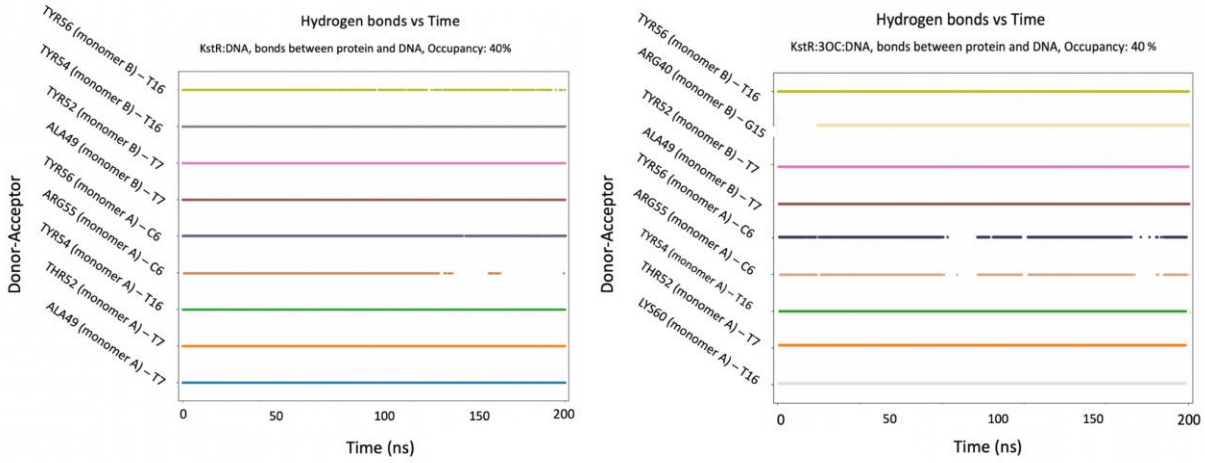


Figure 3. 28. Hydrogen bonds that form between KstR and DNA during the MD simulations of (left) KstR:DNA and (right) KstR:3OC:DNA. Only hydrogen bonds that are occupied for >40% of the simulation are shown.

To determine whether these hydrogen bonds are maintained during the simulations, and if any others are formed, the formation of hydrogen bonds between KstR and DNA during the KstR:DNA and KstR:3OC:DNA simulations was analysed (Figure 3. 28-32), and compared to those present in the crystal structure with DNA bound (Table 3. 3). These results are for hydrogen bonds that are occupied only for >40% of the simulation.

In KstR:DNA, residues TYR56, TYR54, TYR52, and ALA49 form hydrogen bonds with bases C6, T16, T7 and also T7, respectively, in monomer A and with T16, T16, T7 and T7, respectively, in monomer B. These hydrogen bonds remain formed during the entire simulation other than some very minor breaks to the TYR56-T16 hydrogen bond in monomer

B. ARG55 in monomer A also formed a hydrogen bond with C6 but this is mostly broken towards the end of the simulation.

For KstR:3OC:DNA, the hydrogen bonds formed by monomer A are the same as for KstR:DNA, except for an additional hydrogen bond between LYS60 and T16. The two hydrogen bonds between TYR56 and ARG55 and C6 are a bit unstable, however. For monomer B, the hydrogen bonds between TYR52 and T7 and ALA49 and T7 are formed, but the hydrogen bond between TYR54 and T16 is missing, replaced by a hydrogen bond between TYR56 and C6, and there is an additional hydrogen bond between ARG40 and G15 that forms after the first 15 ns of the simulation. It is interesting that this hydrogen bond, the only one that KstR forms directly with a base rather than the DNA backbone, is only formed in this MD simulation of a highly strained complex, and not in the more relaxed KstR:DNA simulation.

Overall, while not all of the hydrogen bonds in the KstR:DNA crystal structure remain formed during the KstR:DNA simulation, several key hydrogen bonds do remain formed and are present for at least 40% of the MD simulation, along with one additional hydrogen bond. When 3OCh is also bound to KstR (KstR:3OC:DNA), the main hydrogen bonds between KstR and DNA observed in the crystal structure are again still present, and another new hydrogen bond is formed. The interaction between KstR and DNA is therefore largely preserved during the MD simulations, and is only slightly perturbed by the addition of ligands to both monomers. This is in contradiction to the hypothesis that ligand binding causes KstR to unbind from DNA.

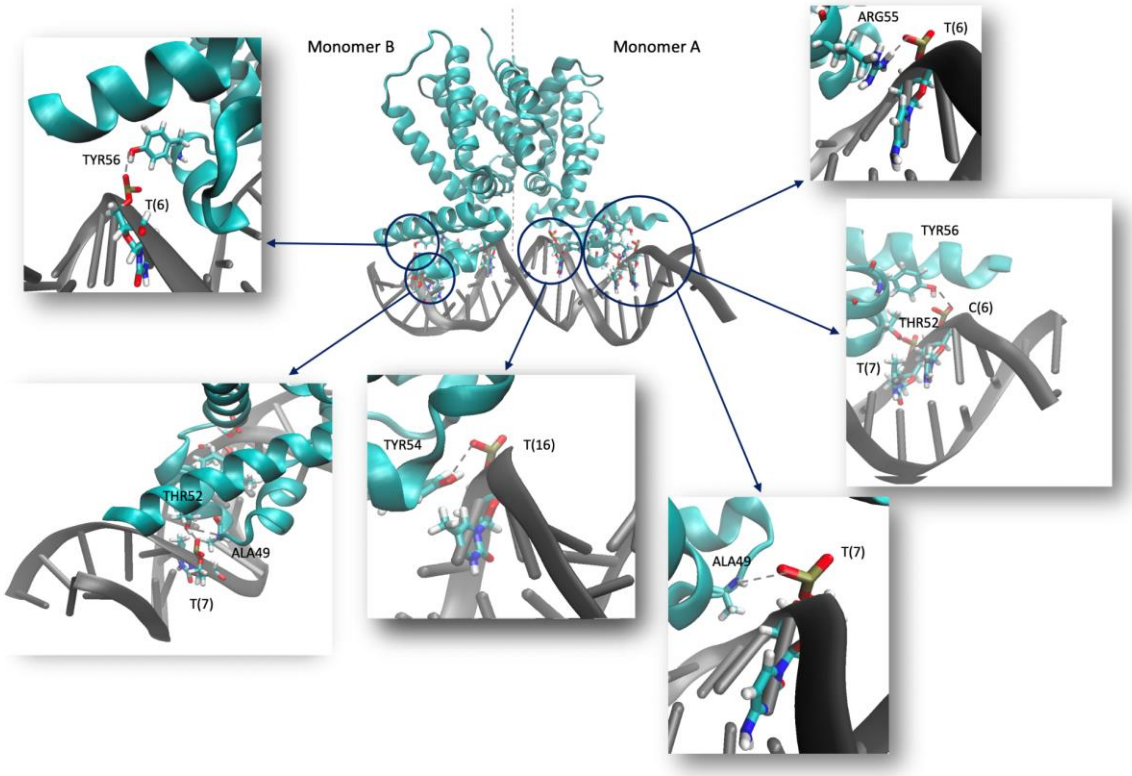


Figure 3. 29. Hydrogen bonds between KstR and DNA occupied for >40% of the MD simulation of KstR:DNA. KstR and DNA are drawn in cartoon format, with KstR in cyan and DNA in black. KstR residues and DNA bases that form hydrogen bonds are labelled and drawn in liquorice format and coloured according to atom type (carbon:cyan; oxygen:red; nitrogen:blue; hydrogen:white; phosphate:gold).

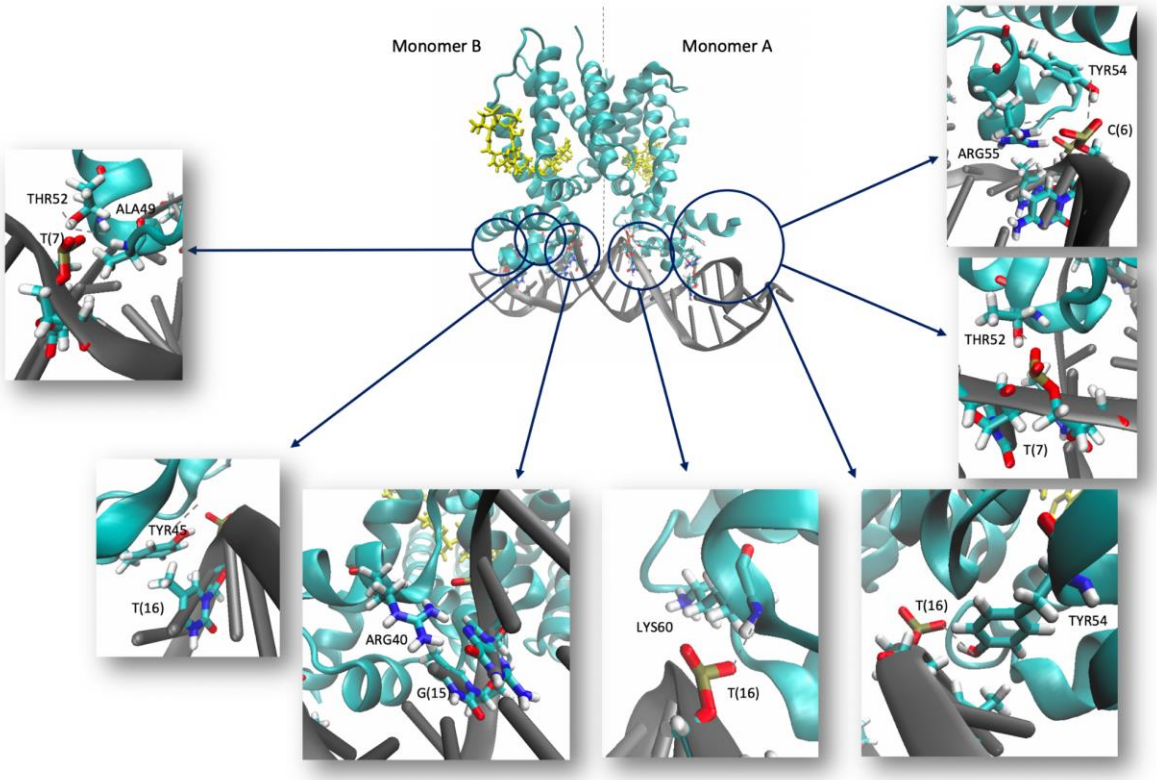


Figure 3. 30. Hydrogen bonds between KstR and DNA occupied for >40% of the MD simulation of KstR:3OC:DNA. KstR and DNA are drawn in cartoon format, with KstR in cyan and DNA in black, and 3OCh is drawn in liquorice format and coloured yellow. KstR residues and DNA bases that form hydrogen bonds are labelled and drawn in liquorice format and coloured according to atom type (carbon:cyan; oxygen:red; nitrogen:blue; hydrogen:white; phosphate:gold).

Table 3. 3. Comparison of hydrogen bonds between KstR and DNA in the KstR:DNA crystal structure 5UA2 and occupied for >40% of the MD simulations of KstR:DNA and KstR:3OC:DNA. The monomer is indicated in brackets. Hydrogen bonds that occur in the crystal structure and in both simulations are shaded grey.

PDB ID: 5UA2		KstR:DNA		KstR:3OC:DNA	
KstR residue	DNA base	KstR residue	DNA base	KstR residue	DNA base
ARG17 (A)	C				
ALA49 (A)	T	ALA49 (A)	T7		
THR52 (A)	T	THR52 (A)	T7	THR52 (A)	T7
TYR56 (A)	C	TYR56 (A)	C6	TYR56 (A)	C6
GLN38 (A)	C				
ARG40 (A)	G				
TYR54 (A)	T	TYR54 (A)	T16	TYR54 (A)	T16
		ARG55 (A)	C6	ARG55 (A)	C6
LYS60 (A)	C			LYS60 (A)	T16
ARG17 (B)	C				
ALA49 (B)	T	ALA49 (B)	T7	ALA49 (B)	T7
THR52 (B)	T	TYR52 (B)	T7	TYR52 (B)	T7
TYR56 (B)	C	TYR56 (B)	T16		
GLN38 (B)	C				
ARG40 (B)	G			ARG40 (B)	G15
TYR54 (B)	T	TYR54 (B)	T16	TYR54 (B)	T16
LYS60 (B)	C				

3.3.5.2. Hydrogen bonds between KstR and 3OCh

The hydrogen bonds formed between KstR and its ligand, 3OCh, have also been studied in detail in the corresponding X-ray crystal structure (135, 142). As the ligand pocket is hydrophobic and the hydrophobic steroid moiety of the ligand fits into the pocket, a hydrogen bond is only formed specifically with the 3-oxo group of the steroid moiety of the ligand (135, 142) (O71 in Figure 3. 31). The CoA tail may facilitate ligand binding as it increases the ligand solubility. However, the role of the CoA moiety in the KstR-ligand interaction has been unclear (135, 142). It has been shown that the CoA of the steroid ligands can increase the ligand affinity for KstR (135, 142). Also, whether the CoA tail interacts with the protein is an unresolved issue (135, 142).

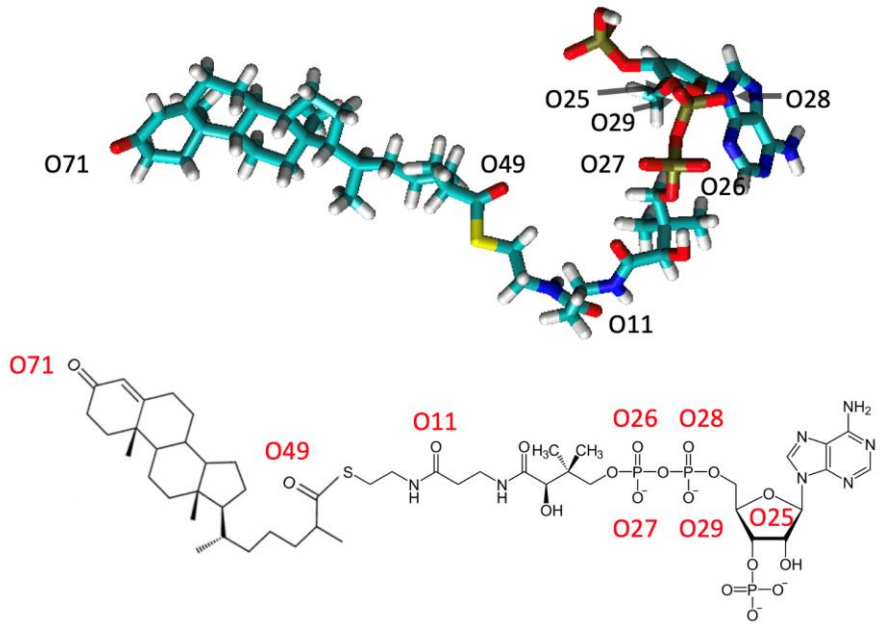


Figure 3. 31. (Top) Liquorice and (bottom) 2D representations of 3OC-CoA. The oxygen atoms involved in hydrogen bonds with KstR are labelled.

To determine whether these hydrogen bonds are maintained during the simulations, and if any others are formed, the formation of hydrogen bonds between KstR and 3OCh during the KstR:3OC, KstR:3OC:DNA and KstR:3OC:DNAremoved simulations was analysed (Figure 3. 32-8, and Table 3.). Here, hydrogen bonds occupied for more than 20% of the simulation for KstR:3OC and KstR:3OC:DNA, and 10% of the simulation for KstR:3OC:DNAremoved were considered as otherwise, only one hydrogen bond was found for KstR:3OC:DNA. For KstR:3OC, a hydrogen bond between ARG158 of monomer A and the (3-oxo, O71 in Figure 3. 31) hydroxyl group of 3OCh, which is also observed in the crystal structure (135, 142), is formed for the majority of the simulation, but disappears during the first 50 ns of the simulation for monomer B. In fact, there are fewer protein-ligand hydrogen bonds overall for monomer B, suggesting some asymmetry in the ligand binding. In monomer A, ARG106 also forms hydrogen bonds, with O29 and O49, as well as ARG158.

Interestingly, when KstR is bound to 3OCh and DNA at the same time, the number of hydrogen bonds between KstR and 3OCh is greatly decreased. The hydrogen bond between ARG158 and 3OCh was not formed in either monomer. Only three hydrogen bonds were occupied for more than 20% of the simulation for monomer B, between residues ARG106 and O25 and O28, and between SER161 and O71 (3-oxo), while for monomer A, 3OCh did not sit properly in the ligand-binding pocket and no persistent hydrogen bonds were formed.

After removing DNA from KstR:3OC:DNA, in the first 200 ns of the simulation, three hydrogen bonds were formed (Figure 3. 32, bottom left). One of these, between SER161 of monomer B and atom O71 of 3OCh, also occurred in the KstR:3OC:DNA simulation, but was lost ~100 ns after DNA was removed. The other two are formed between ARG106 of monomer A and atoms O25 and O28 of 3OCh. These hydrogen bonds are also found in monomer B for KstR:3OC:DNA. However, when extending the simulation to 300 ns, the number of hydrogen bonds increased, albeit mostly involving monomer A. In addition to the hydrogen bond between ARG106 of monomer A and atom O26 of 3OCh, two more hydrogen bonds were formed between this ARG and atoms O27 and O29 of the ligand (Figure 3. 32, bottom right). This hydrogen bond with O29 is also formed in monomer A in KstR:3OC. Additionally, two hydrogen bonds were formed between residues ARG102 and TRP164 of monomer A and atoms O27 and O71 of 3OCh, respectively. In monomer B, only one hydrogen bond forms between ARG102 and O29. These are not observed in any of the other simulations or in the crystal structure. The other hydrogen bonds do not show, when considering hydrogen bonds that are occupied for >20% of 300 ns simulation anymore. In general, atom O71 in the 3-oxo group of 3OCh is involved in hydrogen bonds in all three simulations, although with different residues in each case.

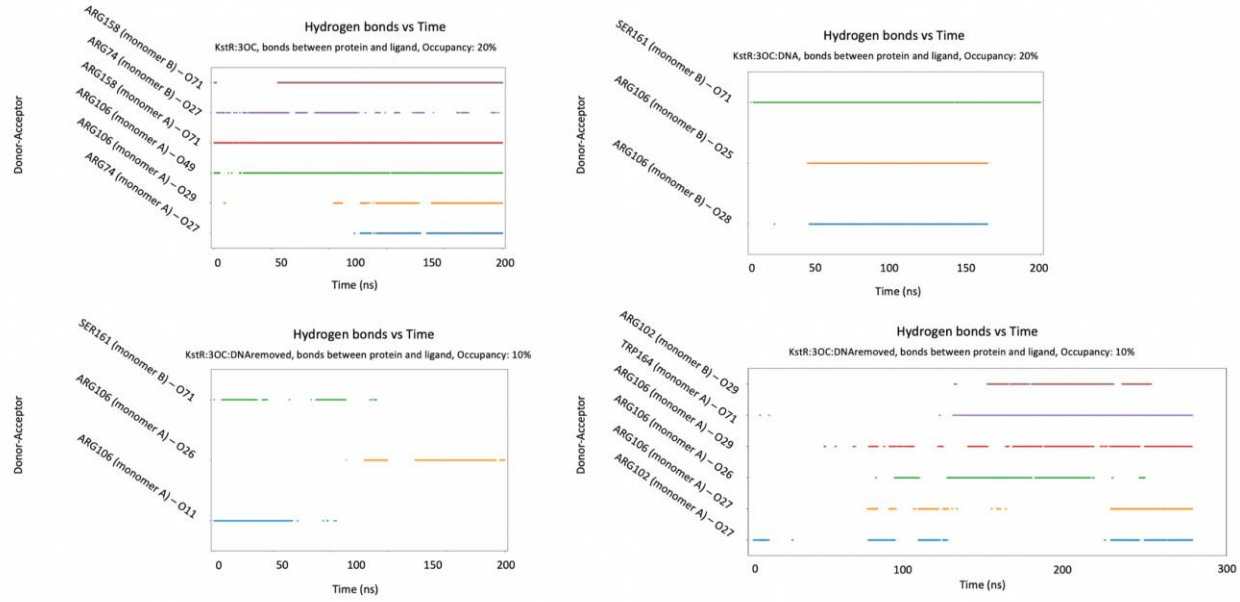


Figure 3. 32. Hydrogen bonds that form between KstR and 3OCh during the MD simulations of (left) KstR:3OC and (right) KstR:3OC:DNA. Only hydrogen bonds that are occupied for >20% (for KstR:3OC and KstR:3OC:DNA) and >10% (for KstR:3OC:DNAremoved) of the simulation are shown. Note that the same colours do not necessarily represent the same hydrogen bonds across the different graphs.

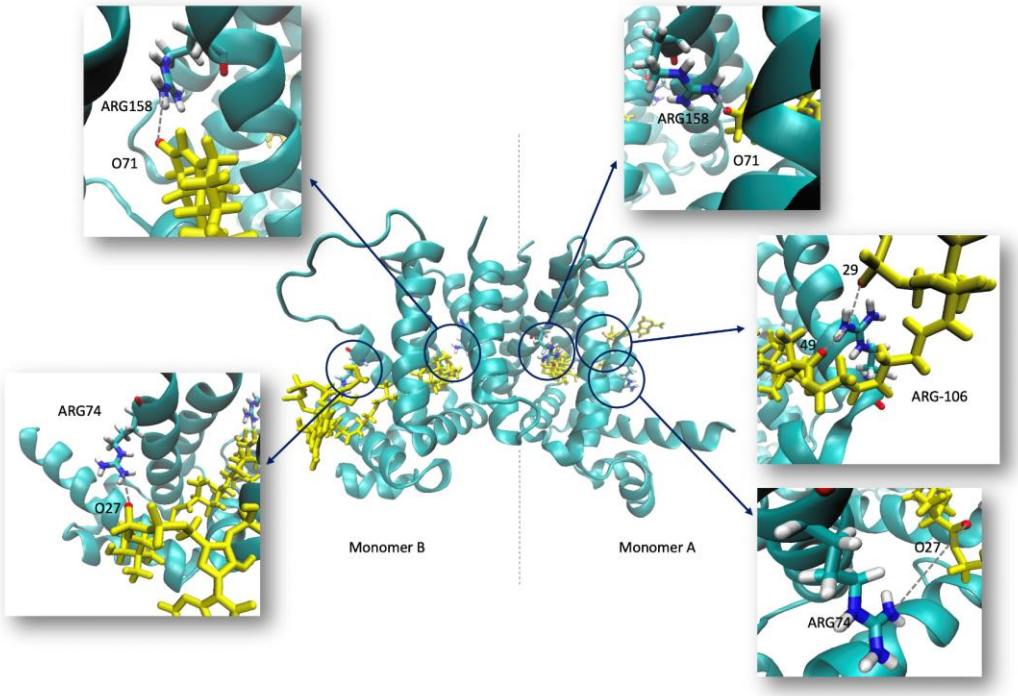


Figure 3. 33. Hydrogen bonds between KstR and 3OCh occupied for >20% of the MD simulation of KstR:3OC. KstR is drawn in cartoon format and coloured yellow, and 3OCh is drawn in liquorice format and coloured yellow. KstR residues that form hydrogen bonds are labelled and drawn in liquorice format and coloured according to carbon:cyan; oxygen:red; nitrogen:blue; hydrogen:white.

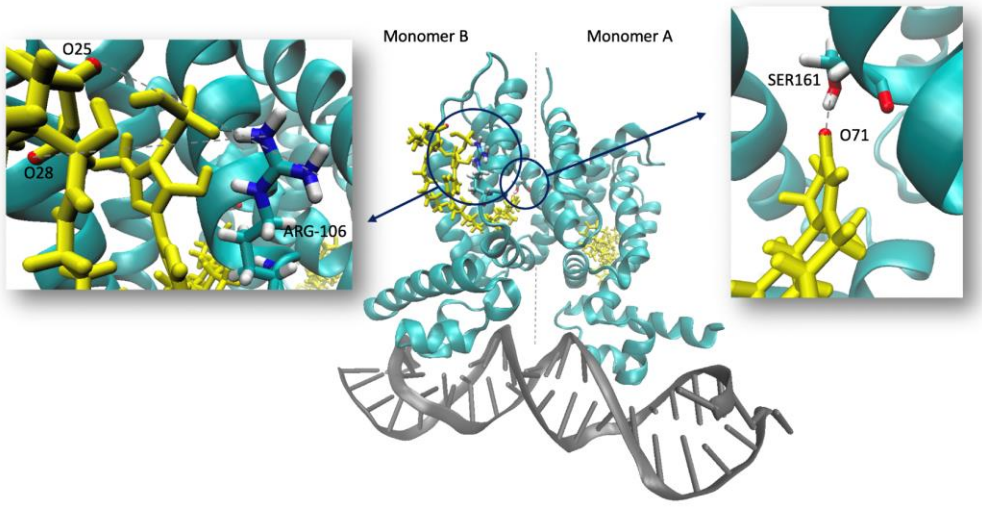


Figure 3. 34. Hydrogen bonds between KstR and 3OCh occupied for >20% of the MD simulation of KstR:3OC:DNA. KstR is drawn in cartoon format and coloured yellow, and 3OCh is drawn in liquorice format and coloured yellow. KstR residues that form hydrogen bonds are labelled and drawn in liquorice format and coloured according to carbon:cyan; oxygen:red; nitrogen:blue; hydrogen:white.

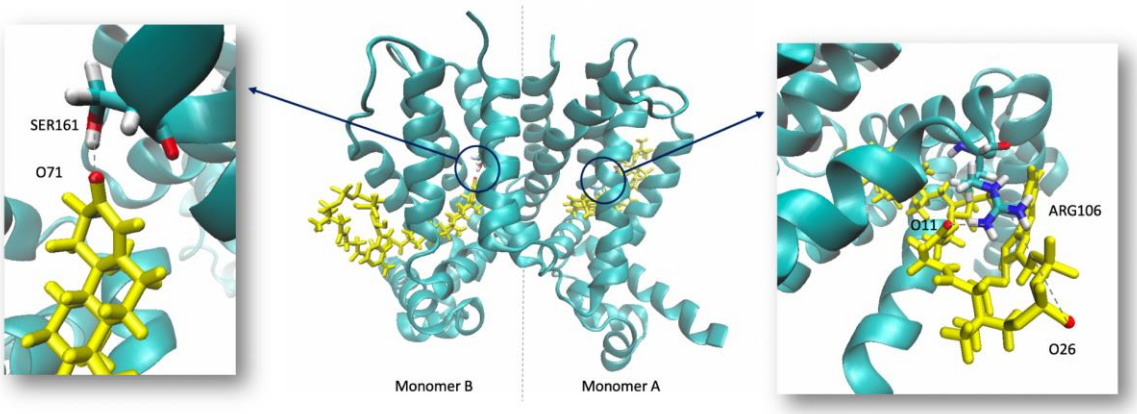


Figure 3. 35. Hydrogen bonds between KstR and 3OCh occupied for >20% of the first 200 ns of the MD simulation of KstR:3OC:DNAremoved. KstR is drawn in cartoon format and coloured yellow, and 3OCh is drawn in liquorice format and coloured yellow. KstR residues that form hydrogen bonds are labelled and drawn in liquorice format and coloured according to carbon:cyan; oxygen:red; nitrogen:blue; hydrogen:white.

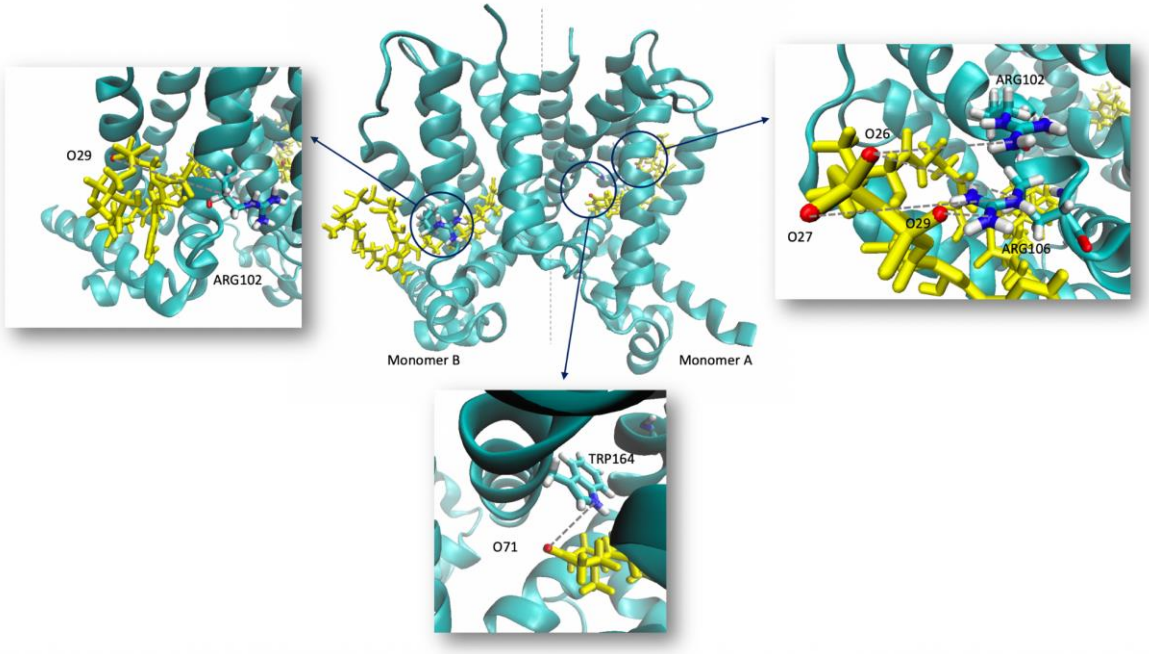


Figure 3. 36. Hydrogen bonds between KstR and 3OCh occupied for >20% of the complete 300 ns of the MD simulation of KstR:3OC:DNAremoved. KstR is drawn in cartoon format and coloured yellow, and 3OCh is drawn in liquorice format and coloured yellow. KstR residues that form hydrogen bonds are labelled and drawn in liquorice format and coloured according to carbon:cyan; oxygen:red; nitrogen:blue; hydrogen:white.

Table 3. 4. Comparison of hydrogen bonds between KstR and 3OCh for KstR:3OC, KstR:3OC:DNA and KstR:3OC:DNAremoved (first 200 ns and entire 300 ns). The monomer is indicated in brackets. Residues that are involved in forming hydrogen bonds in at least two simulations are shaded grey.

KstR:3OC occupied for >20%		KstR:3OC:DNA occupied for >20%		KstR:3OC:DNAremoved, occupied for >10%			
				200 ns		300 ns	
KstR residue	3OCh	KstR residue	3OCh	KstR residue	3OCh	KstR residue	3OCh
ARG74 (A)	O27						
						ARG102 (A)	O27
ARG106 (A)	O49			ARG106 (A)	O26	ARG106 (A)	O26
ARG106 (A)	O29			ARG106 (A)	O11	ARG106 (A)	O29
						ARG106 (A)	O27
ARG158 (A)	O71						
						TRP164 (A)	O71
ARG74 (B)	O27						
						ARG102 (B)	O29
		ARG106 (B)	O25				
		ARG106 (B)	O28				
ARG158 (B)	O71						
		SER161 (B)	O71	SER161 (B)	O71		

3.4. Conclusion

An important question in terms of determining the mechanism of KstR is whether it binds to DNA via conformational selection or induced fit (Figure 3. 37). According to the DBD distance analysis, during the MD simulation of KstR:apo, KstR never samples structures equivalent to its structure when bound to DNA. However, the motions of KstR:apo and KstR:DNA are similar, based on PCA analysis. In contrast, when bound to 3OCh, KstR is more mobile, even more than in its apo state. Similar conclusions can also be drawn from the RMSD and RMSF results. KstR:DNAremoved and KstR:3OCremoved behaved similarly to one another and to KstR:3OC but differently to KstR:apo, at least in terms of e.g. the PCA. In general, there is only limited overlap between the conformational states sampled in each simulation, even across the three apo-like simulations (KstR:apo, KstR:DNAremoved and KstR:3OCremoved). This is likely because more conformational sampling is required, and so is not necessarily evidence for or against a particular mechanism of action. However, since removing DNA from KstR:DNA also shows that KstR:APO never samples its DNA-bound form during the 200 ns simulation time, it seems that an induced fit mechanism is more likely than conformational selection (Figure 3. 37, right).

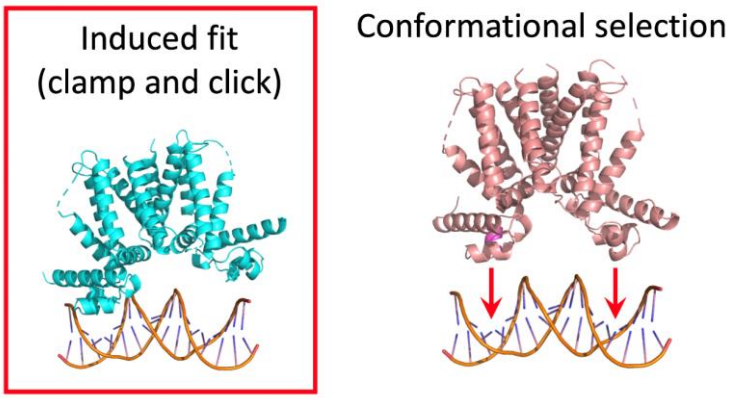


Figure 3. 37. The two possible binding models of KstR to DNA. According to the analysis of the MD simulations, KstR is more likely to bind to DNA by the induced fit method, but the release of KstR from DNA is more likely to follow a conformational selection mechanism.

A related question is whether ligand binding forces KstR off DNA (induced fit), or whether it has to first unbind from DNA before a ligand can bind (conformational selection). When KstR is bound to both DNA and 3OCh, although ligand is poorly bound to the protein, most likely due to being placed into the ligand binding pocket rather than present in a crystal structure, it does not force KstR to unbind from DNA. Moreover, the hydrogen bonds formed

between KstR and DNA are similar in KstR:DNA and in KstR:3OC:DNA. It seems, therefore, that KstR must first unbind from DNA before a ligand can bind and lock it into a state unable to bind to DNA (Figure 3. 37 andFigure 3. 38). This would appear to be an argument against an induced fit mechanism, at least in terms of the release of transcriptional repression by KstR.

After removing DNA from KstR:3OC:DNAremoved, there is a slow reversion towards a ligand-bound state like that of KstR:3OC, so that the simulation needed to be extended from 200 to 300 ns for it to occur. This again suggests that longer simulations are needed to fully examine the mechanism of KstR.

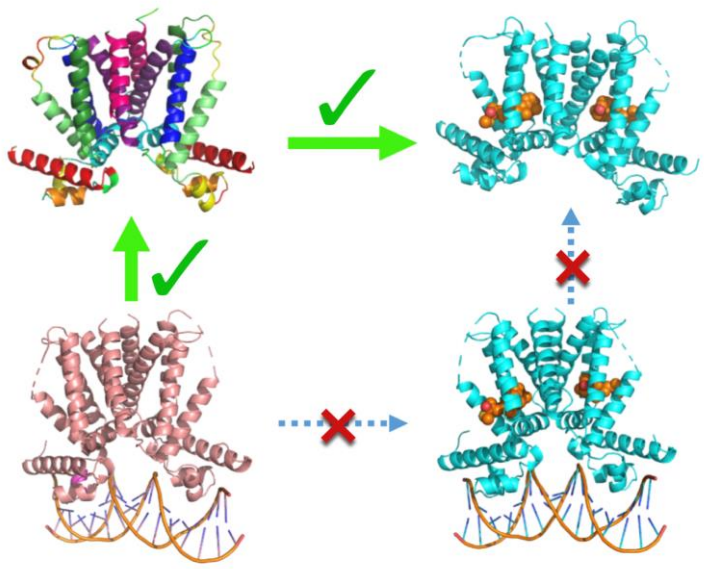


Figure 3. 38. The two possible pathways of KstR binding to ligand and DNA, starting from KstR:APO (bottom left corner in pink colour).

In both DNA-bound states (KstR:DNA, and KstR:3OC:DNA), the community analyses show that the DBD community is extended up to include the linker helix (α4) of DBD and LBD. This is consistent with the other results indicating that the DNA-bound states of KstR are more rigid. This helix has been hypothesised to be involved in communication between the DBD and LBD. Only the communication pathways that occurred in one monomer when DNA was bound to KstR or when DNA or 3OCh were removed involved more than one residue of this helix, suggesting that it may indeed be involved in communicating changes of state.

Furthermore, KstR communities are not always symmetric, that is, both the number of communities and the residues that comprise them are seldom the same (even at an approximate level) between the two monomers. There is also asymmetry in the hydrogen bonds and the communication pathways for the two monomers of KstR, and the RMSF plots show

asymmetric fluctuations of same residues in each monomer. It seems, therefore, that despite its apparent symmetry, KstR seldom acts as a symmetric dimer.

Overall, three major types of communication pathways between ARG158 (LBD) and TYR54 (DBD) were identified, summarised here according to the helices (Figure 3. 39) they traverse: 1) α8 (purple) – α7 (blue) – α4 (green) – α3 (yellow); 2) α8 (purple) – α4 (green) – α3 (yellow); 3) α8 (purple) – α6 (cyan) – α3 (yellow). The fact that the pathways differ between states suggests that communication of DNA and ligand binding is state-dependent.

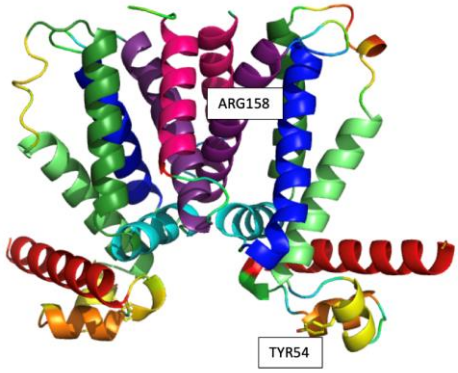


Figure 3. 39. Cartoon representation of KstR, with coloured helices, with showing the two selected residues for communication pathway analysis.

Overall, while it was not possible to comprehensively answer the questions regarding the mechanism of action of KstR, some insight was gained, while opening a number of new questions for investigation. However in addition to the limitation of incomplete sampling mentioned throughout, another reason why these results should be interpreted with caution is that only one replicate of each simulation was performed, except for KstR:apo. These replicates gave similar results, however, which is encouraging.

KstR is of particular interest due to its critical role in controlling cholesterol metabolism in Mtb, which is crucial for Mtb pathogenesis. The simulation of KstR:3OC:DNA creates a promising opportunity for drugs that target KstR and thus tackle TB. When 3OCh was bound as well as DNA, the 3OCh ligands appeared to be locked in the ligand pockets but did not form the typical interactions with the protein nor cause KstR to unbind from the DNA. This is likely in part due to the shape of pocket being different in the DNA-bound form of KstR to the apo or ligand-bound forms. Therefore, it might be possible to identify a drug that prefers to bind to the ligand pockets when KstR is bound to DNA and locks it onto the DNA, thus preventing transcription of the genes encoding the cholesterol metabolism machinery and also preventing 3OCh from binding. Such a drug would potentially reduce the pathogenesis of TB.

Chapter 4. KstR2

4.1. Overview

Cholesterol catabolism in *Mycobacterium tuberculosis* (*Mtb*) is influential for pathogenesis and therefore is a potential target for new anti-tubercular drugs and vaccines development. As mentioned in Chapters 1 and 3, two transcriptional repressors, KstR and KstR2, regulate cholesterol breakdown in *Mtb*. The initial steps in cholesterol degradation are regulated by KstR, which itself is required for pathogenesis (152), showing that appropriate regulation of cholesterol metabolism is essential for the bacterium to cause disease. KstR2 represses a 14-gene regulon involved in the later steps of cholesterol degradation, i.e. degradation of the C and D sterol rings (139, 153, 154).

KstR2, like KstR, is a homodimer (Figure 4. 1). Unlike KstR, however, KstR2 binds to DNA as a pair of dimers. There is no crystal structure of KstR2 bound to DNA yet (157), thus that aspect of its function was not investigated here.

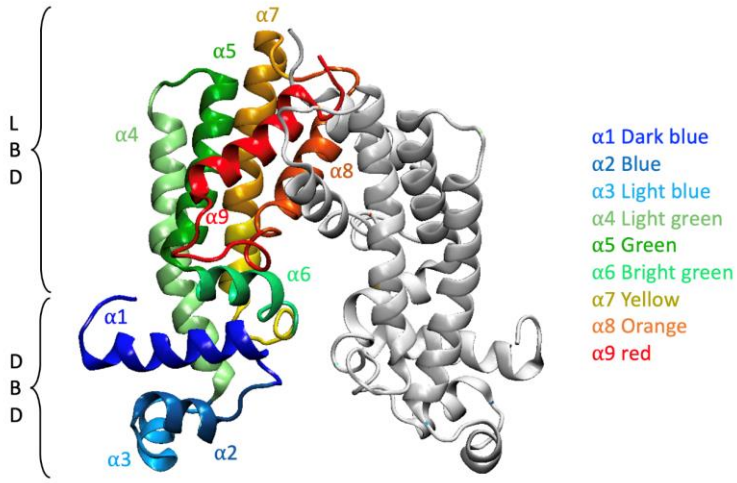


Figure 4. 1. Structure of KstR2. (Left) Side view of the homo-dimer, showing the DNA binding domain (DBD) consisting of helices *α*1-*α*3, and the ligand binding domain (LBD), consisting of helices *α*4-*α*9. The hairpin helices (*α*8 and *α*9) contributing to the four-helix bundle of the dimer interface are colored orange and red.

KstR2 is regulated by S-[2-[3-[[(2R)-4-[[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl]oxy-oxidanyl-phosphoryl]oxy-3,3-dimethyl-2-oxidanyl-butanoyl]amino]propanoylamino]ethyl] 3-[(3aS,4S,7aS)-7a-methyl-1,5-bis(oxidanylidene)-2,3,3a,4,6,7-hexahydroinden-4-yl]propanethioate (HIP-CoA), for which there is a binding pocket in each monomer. HIP-CoA represents the entry molecule of the section of the cholesterol degradation pathway for which

KstR2 regulates gene transcription. Without a DNA-bound KstR2 structure, however, it is not possible to investigate this at a structural level.

It is, however, possible to investigate the effect of ligand binding by KstR2. There is a crystal structure of apo Mtb-KstR2 (PDB ID: 4W1U) (157), and a ligand-bound structure (bound to HIP-CoA) (PDB ID: 4W97) (137). There is also a structure of apo KstR2 from *Rhodococcus jostii* (strain RHA1) (Rjo), a Gram-positive polychlorinated biphenyl (PCB) degrader. The Rjo KstR2 shares 59% amino acid sequence identity with Mtb KstR2, including 19 of the total 23 amino acids that interact with HIP-CoA ligand (PDB ID: 2IBD) (137, 156). Interestingly, the Rjo KstR2 structure has the DBDs positioned relatively close to one another, at a separation of ~4 nm, whereas the Mtb KstR2 structure has the DBDs positioned further apart, at a separation of ~ 5 nm. This latter state is more like the ligand-bound state of KstR2 (226).

While these structures have provided insight into how KstR2 binds to HIP-CoA, and the conformational changes associated with doing so, the mechanism by which ligand and DNA binding are communicated, and whether KstR2 operates *via* an induced fit or conformational selection mechanism, are not known. It has been hypothesised (157), however, that KstR2 might act via a double scissor allosteric mechanism (Figure 4. 2) that, if correct, would be novel among TFRs. This mechanism is based on the movements of four opposing sets of rigid bodies being driven by a centrally located arginine pair and is supported by biophysical and structure-guided mutagenesis experiments (157). The key arginine is ARG170, which sits at the intersection of the dimer interface and the ligand binding pocket and appears to be responsible for the different separation of the DBDs in Rjo KstR2 and Mtb KstR2 structures (226). The same experimental data also support a conformational selection mechanism of DNA binding by KstR2, although this was not directly investigated here.

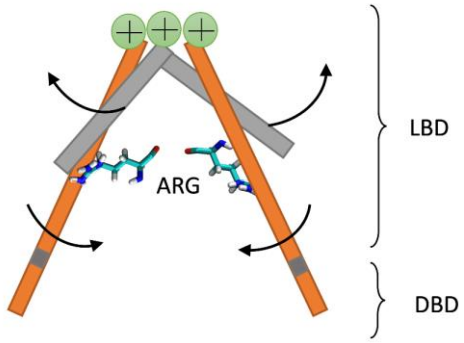


Figure 4. 2. *Rigid* body motions of KstR2. Four rigid bodies flexing about the dimer interface. The arginine residues are indicated. Each monomer is shown as a pair of sticks, one long orange stick representing the LBD and DBD, and a short grey stick representing one half of the dimer interface four-helix bundle. A green screw head denotes the axis of rotation and the arrows denote the general movement predicted to occur to move from ligand binding to DNA binding poses. The switch that modulates this scissor action, which is an arginine in the case of KstR2, is indicated.

In this chapter, the conformational changes of KstR2 in its apo state and with HIP-CoA bound, and its interactions with HIP-CoA, have been investigated using MD simulations. To investigate the hypothetical allosteric mechanism, simulations were also run of KstR2 variants with ARG170 mutated to alanine. Analysis of the simulations of WT and R170A Rjo and Mtb KstR2 in terms of properties such as residue-residue distances, dynamical communication networks, principal components and motions, and hydrogen bond formation support the scissor-like allosteric mechanism and the key role of ARG170 hypothesised by Dawes et al. (157), and indirectly support their hypothesis that KstR2 primarily uses a conformational selection mechanism of DNA binding (157). Subsequently, the same analysis carried out for the simulations of KstR2 with one or two HIP-CoA ligands bound, only the HIP portion bound, or the ligands removed, provided insight into how the CoA portion interacts with KstR2 and the effect that ligand binding has on the structure and dynamics of KstR2.

4.2. Methods

4.2.1. Parameters

KstR2 was modelled using the CHARMM36 force field (216). Since parameters for the HIP-CoA ligand were not available in the CHARMM36 force field, they were generated using CGENFF, which produces parameters for small molecules compatible with CHARMM36 (217, 218). Water was modelled using TIP3P (219).

4.2.2. Coordinates

Initial coordinates of KstR2 were taken from each of three crystal structures with PDB IDs 2IBD (156), 4W1U (157), and 4W97 (137), which are of apo Rj-KstR2, apo Mtb-KstR, and

ligand-bound Mtb-KstR2, respectively. Missing residues were built using Modeller by Ali Razzak (143) for all three KstR2 structures. Here, in order for the residue numbers to be consistent in the three structures, the numbers were modified accordingly after carrying out a structure-guided sequence alignment using Chimera. For simulations of wild-type KstR2, the original PDB coordinates were used after building coordinates for the missing residues. For simulations of mutated forms of KstR2, ARG170 was mutated to ALA in both monomers. 4W97 is a crystal structure of one monomer of KstR2 with a HIP-CoA molecule (Figure 4. 3) bound in the ligand binding pocket and four Cl^- ions (128). These ions were retained as their presence in the crystal structure suggest they may be important for ligand binding. Since the two monomers are supposed to be identical in KstR2, monomer A and the bound ligand were duplicated and rotated to form the dimer using the structure-guided superposition function in Chimera (227). For positioning the two monomers, KstR was used as a template due to there not being a ligand-bound structure of the KstR2 dimer available.

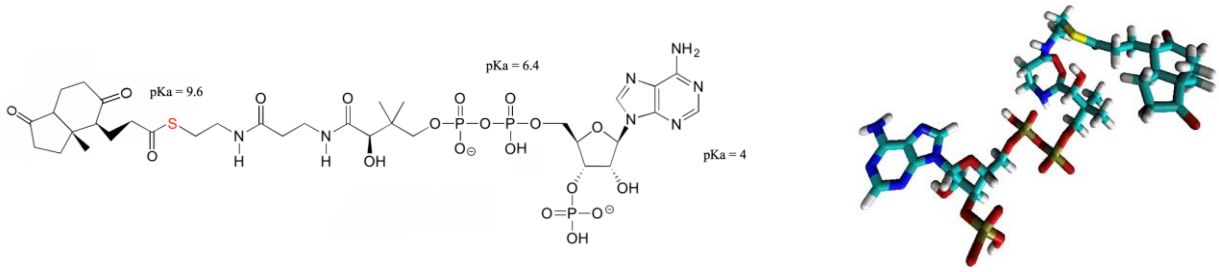


Figure 4. 3. Structure of HIP Co-enzyme A. Left: Chemical structure of HIP Co-enzyme A (full name: S-[2-[3-[[(2R)-4-[[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-4-oxidanyl-3-phosphonooxy-oxolan-2-yl]methoxy-oxidanyl-phosphoryl]oxy-oxidanyl-phosphoryl]oxy-3,3-dimethyl-2-oxidanyl-butanoyl]amino]propanoylamino]ethyl] 3-[(3aS,4S,7aS)-7a-methyl-1,5-bis(oxidanylidene)-2,3,3a,4,6,7-hexahydroinden-4-yl]propanethioate). Right: Liquorice representation of the 3D structure when bound to KstR2. Grey: carbon; red: oxygen; blue: nitrogen; yellow: sulfur; orange: phosphorous; white: hydrogen.

While HIP-CoA coordinates were available from the 4W97 crystal structure, they do not include coordinates for the hydrogen atoms, which are not identifiable using X-ray crystallography. The locations of the hydrogen atoms, in particular, those on phosphate groups, had to be carefully chosen in order to be chemically plausible. There is a salt bridge proposed to occur between ARG162 and one of the phosphates of HIP-CoA (Figure 4. 3), based on the crystal structure. This would require the arginine to be positively charged and the phosphate to be negatively charged. The charge of an amino acid depends on both its pKa and the pH. The pH of the crystallisation solution was 5.5, although if residues are in a hydrophobic environment (such as the core of the protein), then they can have full charges regardless of the solution pH. The pKa of the side chain guanidino group of arginine is 12.48 (228), so it would be expected to be protonated and therefore be positively charged at pH 5.5. The simulations

were run at pH 7.0 to better mimic a physiological environment; however, arginine is also positively charged at this pH. The pKa values in CoA are 9.6 (thiol), 6.4 (secondary phosphate) and 4.0 (adenine NH^{3+}). At pH 5.5, the secondary phosphate would be protonated, but at pH 7.0 as in the simulations, it should be deprotonated. Therefore, the secondary phosphate was deprotonated and ARG162 was protonated.

4.2.3. Computational methods

4.2.3.1. Simulation methods

The initial coordinates of the protein and ligand (where present) were energy minimised. Then the system was solvated in water, and ions added to the solvent to neutralise the system. The system was then energy minimised again, and heated in the NVT ensemble prior to MD simulation for 200 ns in the NpT ensemble. For Mtb-KstR2 (4W1U-WT), energy minimisation and the heating step were performed by Ali Razzak (143). The details of all simulation methods were as for KstR (Chapter 3, Section 3.2.3.1).

4.2.3.2. Analysis methods

RMSD, RMSF, key residue distances, principal component analysis, dynamical network and hydrogen bond analyses were carried out using the same procedures as for KstR (Chapter 3, Section 3.2.3.2).

4.3. Results and discussion

There are two major goals of the MD simulations described here. First, the structure and dynamics of the apo states of the two different KstR2s, Mtb and Rjo KstR2, were investigated, and the proposed scissor mechanism for allostery was investigated by simulating R170A mutants of these two apo KstR2 structures. The second goal was to investigate the conformational dynamics and stability of the protein-ligand interactions of HIP-CoA-bound Mtb KstR2. As part of these investigations, another ‘apo’ state of Mtb KstR2 was created by removing the HIP-CoA from the 4W97 crystal structure, analogous to the 3OCremoved simulations of KstR carried out in Chapter 3. Simulations were also carried out with both ligands bound, only one ligand bound, and with two ligands but with only the HIP portion. The CoA moiety of the ligand is hypothesized to be very important for activation in KstR2, by increasing the binding affinity and interaction of protein and ligand (Stephanie Dawes, *Pers. Commun.*). The simulations here of the complete HIP-CoA as well as only the HIP portion

bound to KstR2 therefore represent the first opportunity to investigate its mode of binding and effect on KstR2. To check if KstR can stably bind to only one ligand and whether this has any effects on the symmetry of the protein, a simulation was carried for KstR2 bound to only one HIP-CoA. Overall, eight unbiased MD simulations were performed for KstR2 (Table 4. 1).

In the remainder of this section, the effect of the R170A mutation on Rjo and Mtb KstR2 is described first, followed by the investigation of Mtb KstR2 with and without different numbers and types of ligands bound.

Table 4. 1. Key details for the eight MD simulations of KstR2

PDB ID	Type	Simulation Code	Description
2IBD	Rjo KstR2	Rjo-KstR2-WT	Wild-type, apo
		Rjo-KstR2-R170A	Mutated (R170A), apo
4W1U	Mtb KstR2	Mtb-KstR2-WT	Wild-type, apo
		Mtb-KstR2-WT	Mutated (R170A), apo
4W97	Mtb KstR2	Mtb-KstR-2Hip-CoA-removed	Ligand removed
		Mtb-KstR2-2HIP-CoA	Two HIP-CoA
		Mtb-KstR2-1HIP-CoA	One HIP-CoA (monomer A)
		Mtb-KstR2-2HIP	Two HIP

4.3.1. Effect of R170A mutation on Rjo and Mtb KstR2

4.3.1.1. Structural stability

To find out how much the protein structure changes during the simulation time, the atom-positional root-mean-square deviation (RMSD) of the Cα atoms of KstR2 from their initial positions (after solvation step) at each point in time was calculated. The results for key sets of simulations are compared below.

The R170A mutation does not result in a major change in the average RMSD of Rjo-KstR2 (Figure 4. 4, Table 4.2). Both systems undergo some rapid changes in the RMSD followed by periods where the overall change is less than the degree of fluctuation. While it is not clear if either system plateaus within the 200 ns timeframe of the simulation, there are also not ongoing increases in the RMSD, suggesting instead that the protein is sampling a range of semi-stable states. Snapshots of these states appear to show quite different distances between the key DNA-binding residue, TYR48, especially for Rjo-KstR2-WT (more details in Section 4.3.1.2).

Table 4. 2 Averages and standard deviations of the RMSD values and minimum distances between the TYR48 residue in each of the two monomers.

Simulation code	RMSD (nm)	Minimum distance (nm) TYR48 (monomer A) – TYR48 (monomer B)
Rjo-KstR2-WT	0.31 ± 0.04	4.23 ± 0.41
Rjo-KstR2-R170A	0.34 ± 0.02	3.94 ± 0.26
Mtb-KstR2-WT	0.25 ± 0.05	4.35 ± 0.52
Mtb-KstR2-R170A	0.24 ± 0.02	4.59 ± 0.31

The R170A mutation also has relatively little effect on the magnitude of the RMSD for Mtb-KstR2 (Figure 4. 5, Table 4.2). However, there is a jump in the RMSD value of Mtb-KstR2-WT at around 80 ns, which then reverts to a lower value after around 130 ns. Visualization of the simulation trajectory revealed that the larger RMSD corresponds to a state in which the DBDs are closer. The distances between the key residue for DNA binding (TYR48) during the simulation time is analysed in Section 4.3.3.

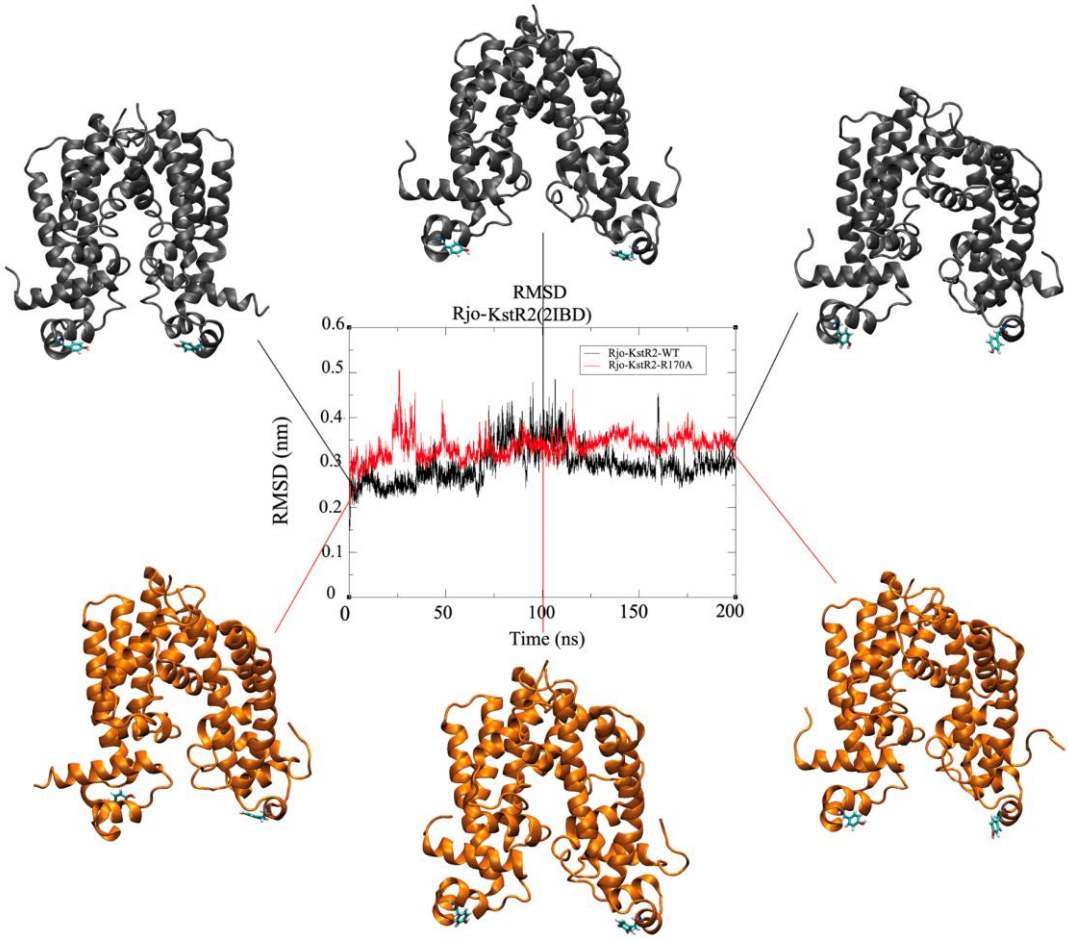


Figure 4. 4. RMSD time-series of the Rjo-KstR2-WT (black) and Rjo-KstR2-R170A (red) MD simulations of KstR2. Surrounding it are snapshots of the protein structure at the points in the simulation indicated, with KstR drawn in cartoon format and coloured according to the RMSD time-series, and TYR48 drawn in liquorice format and coloured according to atom type: carbon: cyan; oxygen: red; nitrogen: blue and hydrogen: white.

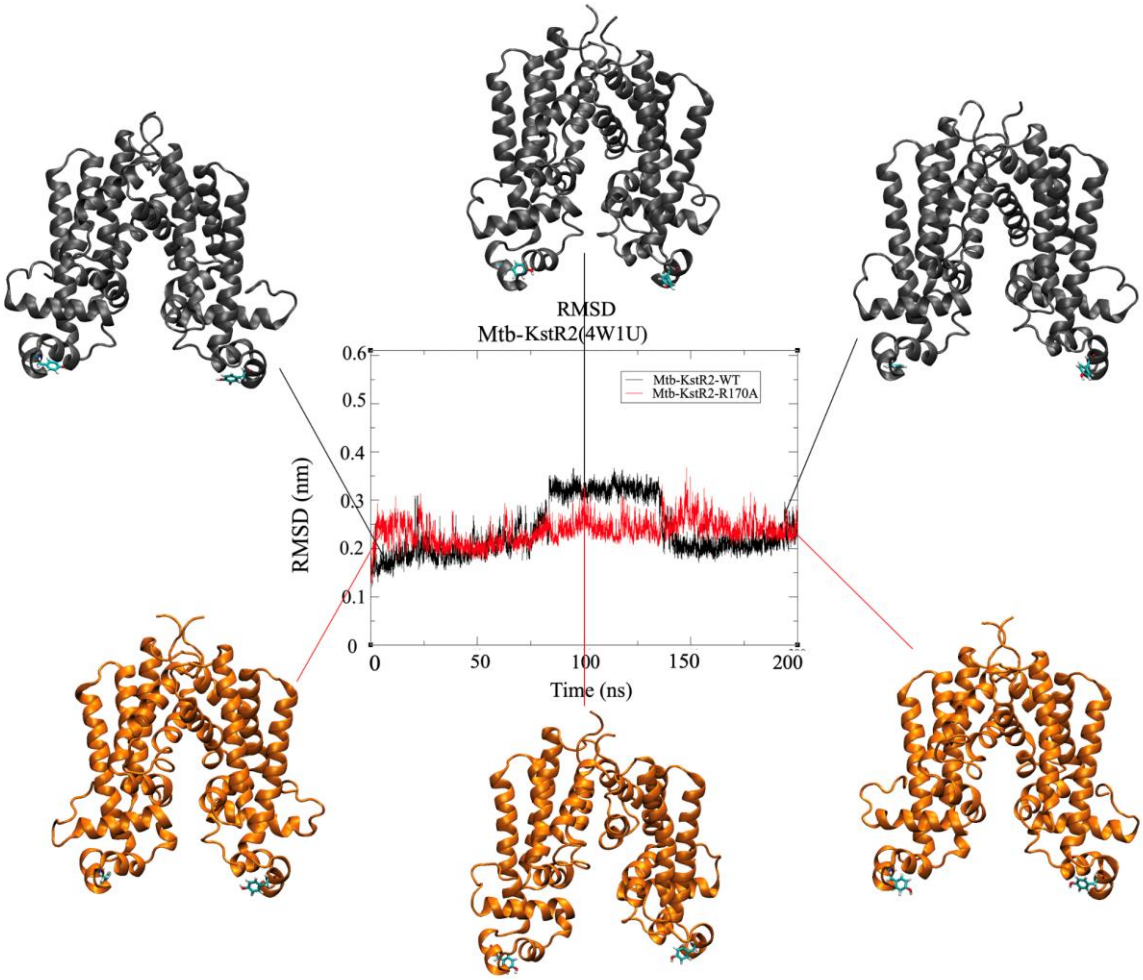


Figure 4. 5. RMSD time-series of the Mtb-KstR2-WT (black) and Mtb-KstR2-R170A (red) MD simulations of KstR2. Surrounding it are snapshots of the protein structure at the points in the simulation indicated, with KstR drawn in cartoon format and coloured according to the RMSD time-series, and TYR48 drawn in liquorice format and coloured according to atom type: carbon: cyan; oxygen: red; nitrogen: blue and hydrogen: white.

Whereas the RMSD gives a single value that summarises the deviation of the protein structure from its initial coordinates during the simulation, the RMSF shows which residues fluctuate the most on average across the simulation.

The mutation of ARG170 to ALA caused a slight jump in the RMSF of the mutated residue in monomer A of Rjo-KstR2 (Figure 4. 6, green). This is most likely due to the much smaller side chain of ALA compared to ARG, which means ALA has more space in which to move during the simulation. Although the increase in RMSF of the mutated residue occurs in both monomers, it is larger in monomer A, which could represent asymmetric behaviour of the protein. Otherwise, the RMSF values are largely similar between the two systems, other than residues 10-50, which correspond to the DBD, having slightly higher RMSF values for Rjo-KstR2-R170A. The N-termini are again more mobile (Figure 4. 6, yellow), but the C-terminal

residues are not particularly dynamic, most likely because they do not form such an extended loop as they do for Mtb-KstR2.

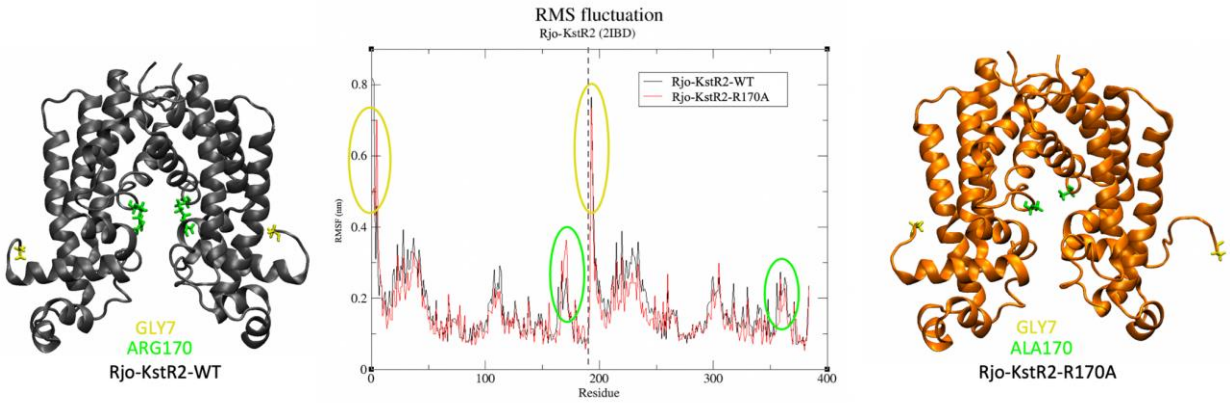


Figure 4. 6. RMSF of the Rjo-KstR2-WT (black) and Rjo-KstR2-R170A (red) MD simulations of KstR2, and snapshots from each simulation, with KstR2 drawn in cartoon format and coloured the same as the RMSF data. The residues in the regions circled in the RMSF plots are coloured accordingly in the snapshots. The residues were renumbered for this analysis in order to have a continuous plot (dashed line separates the two monomers).

The pattern of RMSF values for Mtb-KstR2-WT and Mtb-KstR2-R170A (Figure 4. 7) are similar to those for Rjo-KstR2-WT and Rjo-KstR2-R170A. The difference in the RMSF values of the DBD (residues 10-50) is even more pronounced, however, with the R170A mutation decreasing the dynamics of this region. The C-terminal residues are more mobile for Mtb-KstR2, most likely because the C-terminal is not truncated as it is for Rjo-KstR2.

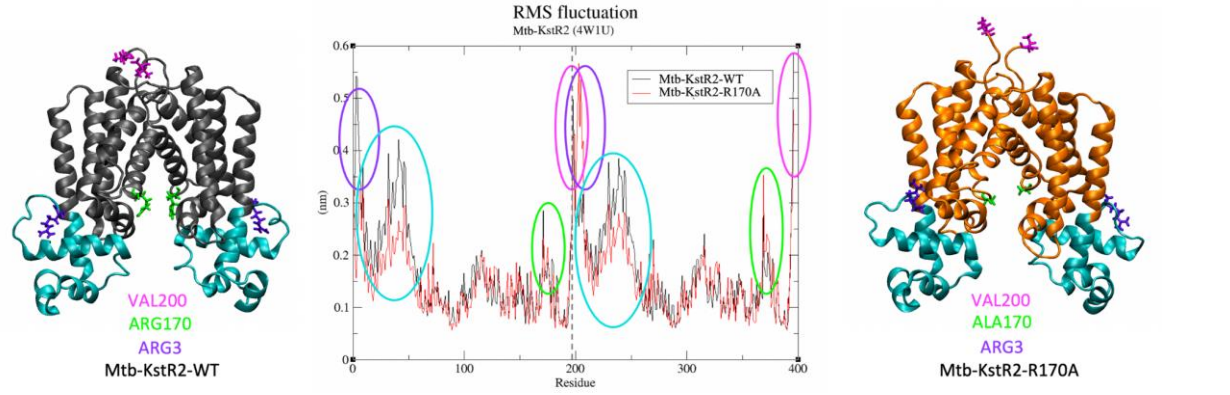


Figure 4. 7. RMSF of the Mtb-KstR2-WT (black) and Mtb-KstR2-R170A (red) MD simulations of KstR2, and snapshots from each simulation, with KstR2 drawn in cartoon format and coloured the same as the RMSF data. The residues in the regions circled in the RMSF plots are coloured accordingly in the snapshots. The residues were renumbered for this analysis in order to have a continuous plot (dashed line separates the two monomers).

4.3.1.2. Separation of DNA-binding domains

Similar to KstR (Chapter 3), KstR2 also has distinct DNA- and ligand-binding conformations. When binding to DNA (which has not been studied in this work), the two DBDs must be closer to each other in order to interact with the DNA, while in the ligand-bound state, they are further

apart (157). In order to investigate the DBD distance in the simulations, the Cα atoms of two TYR residues (TYR48 in both monomers) in the DBDs were chosen as reference points between which to calculate the inter-DBD distance.

The distance between the Cα atoms of the two TYR48 residues in crystal structure of Rjo KstR2 (PDB ID: 2IBD) (156) is about 4.3 nm (PDB ID: 2IBD), and both MD simulations of Rjo KstR2 start from inter-DBD distances of ~ 4 nm (Figure 4. 8). While Rjo-KstR2-R170A undergoes continuous modulation about an average distance of 3.94 nm (Table 4.2), similar to the initial distance, Rjo-KstR2-WT undergoes a sharp increase in the inter-DBD distance around 75 ns, then reverts to shorter inter-DBD distances around 120 ns, although the average distance is 4.23 nm, similar to what is observed in the crystal structure. The times of these transitions are consistent with the RMSD plots (Figure 4. 4), as is the lack of sharp changes for Rjo-KstR2-R170A. This breathing motion of the DBDs that occurs during the Rjo-KstR2-WT simulation therefore could explain the sharp increase in RMSD. However, a longer simulation is required to confirm this conclusion.

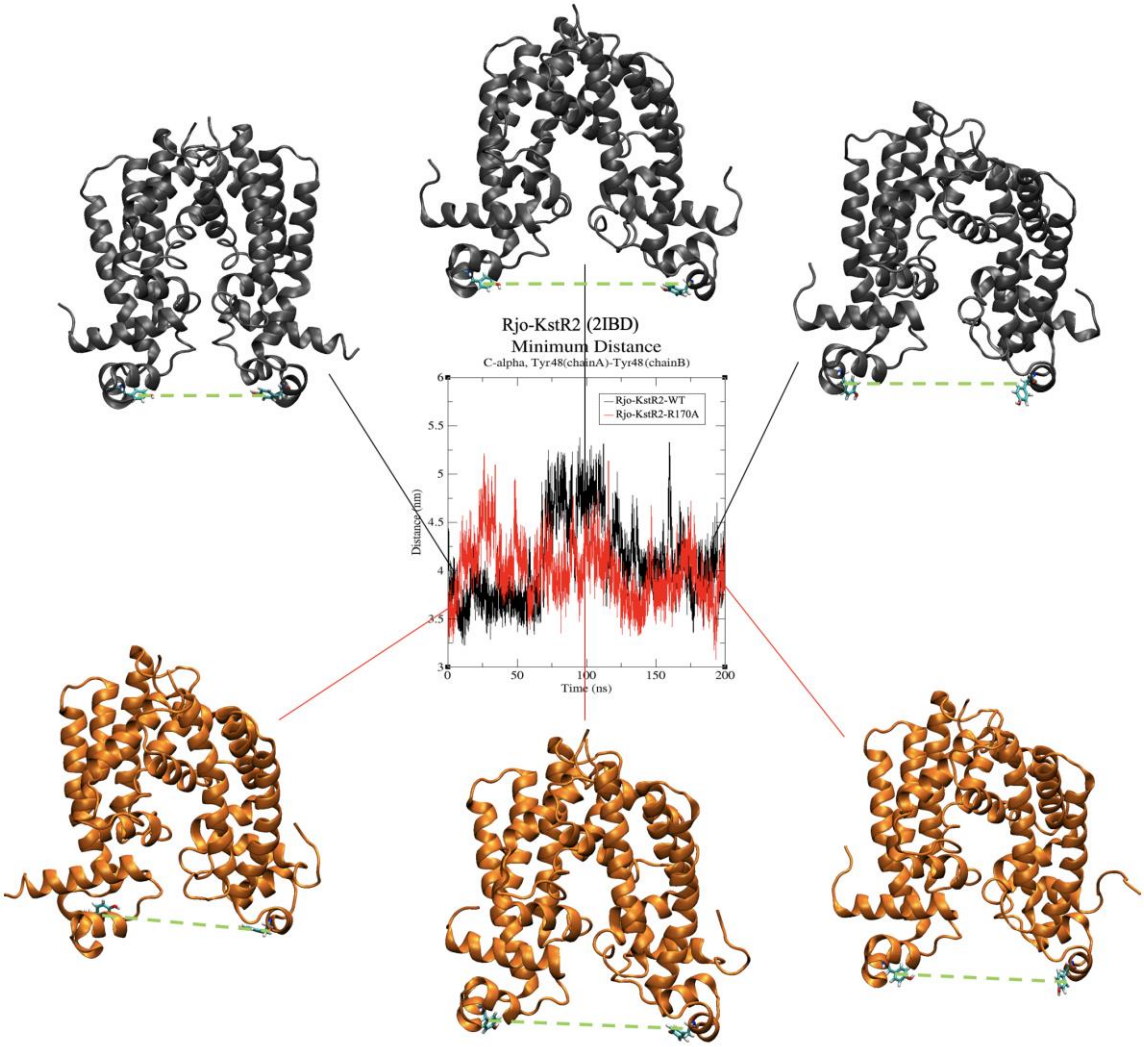


Figure 4. 8. Time-series of the distance between the Cα atoms of TYR54 of each monomer of KstR2 during the Mtb-KstR2-WT and Mtb-KstR2-R170A MD simulations, and snapshots from each simulation, with KstR2 drawn in cartoon format and coloured the same as the RMSF data. TYR48 is drawn in liquorice and coloured according to atom type (cyan, carbon; red, oxygen; white, hydrogen; blue, nitrogen). The green dashed lines connect the two TYR residues between which the distance was calculated.

The TYR48 distance in the Mtb KstR2 crystal structure (PDB ID: 4W1U) is about 5.0 nm. Mtb-KstR2-WT retains an inter-DBD distance close to this for some time, whereas in Mtb-KstR2-R170A the DBDs rapidly move closer together at the start of the simulation (Figure 4. 9). Like Rjo-KstR2-WT, Mtb-KstR2-WT also undergoes a sharp change in the inter-DBD distance during the simulation, although it does the opposite to Rjo-KstR2-WT: the inter-DBD distance decreases during the central portion of the simulation (80-130 ns), and then reverts to larger inter-DBD distances (Figure 4. 9). Mtb-KstR2-R170A, in contrast, undergoes a gradual increase in the inter-DBD distance after their initial collapse, and then settles with large inter-DBD distances of ~4.5 nm, although these remain shorter than that of the Mtb KstR2 crystal structure. Overall, however, the average inter-DBD distances are similar for both simulations (Table 4.2).

Therefore, in both Rjo KstR and Mtb KstR2, the R170A mutation prevents large-scale and rapid changes in the inter-DBD distance. This is in keeping with the reduced RMSF values for the DBD in Rjo-KstR2-R170A and Mtb-KstR2-R170A. These results support the hypothesis that ARG170 plays key role in directing conformational changes of KstR2.

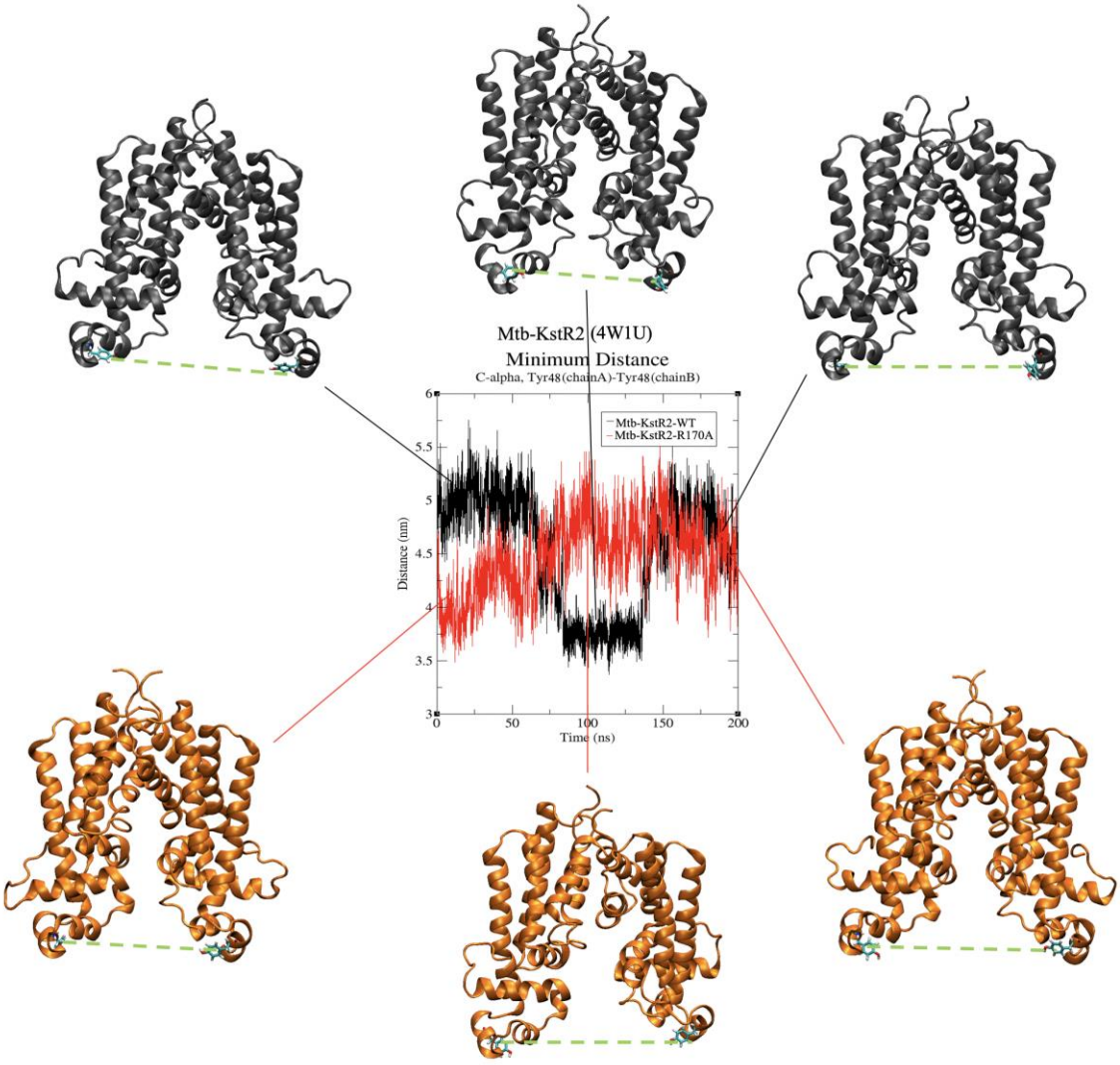


Figure 4. 9. Time-series of the distance between the Cα atoms of TYR54 of each monomer of KstR2 during the Mtb-KstR2-WT and Mtb-KstR2-R170A MD simulations, and snapshots from each simulation, with KstR2 drawn in cartoon format and coloured the same as the RMSF data. TYR48 is drawn in liquorice and coloured according to atom type (cyan, carbon; red, oxygen; white, hydrogen; blue, nitrogen). The green dashed lines are bolding the two TYR residues and the distance between them.

4.3.1.3. Determination of concerted motions

To determine the major and concerted conformational motions that KstR2 undergoes during each simulation, cartesian covariance and principal component analysis (PCA) were carried out for each of the simulations.

The results of PCA analysis for the WT and R170A forms of KstR2, both Rjo-KstR2 and Mtb-KstR2, are shown in Figure 4. 10. Visualisation of the projection of the extreme motion of Rjo-KstR2-WT shows that it undergoes remarkably large breathing motions of the DBDs. The motion is similar in nature for the Rjo-KstR2-R170A, but markedly smaller in scale. Absolute maximum values of the the projections onto the first eigenvector were compared. In both KstR2 proteins the values were decreased after the mutation (from 1.28 to

0.21 for Rjo-KstR2, and from 1.44 to 0.36 for Mtb-KstR2). In both cases, the LBD undergoes less motion than the DBDs, and the LBDs of Rjo-KstR2-R170A move less than those of Rjo-KstR2-WT. The greater range of motion for Rjo-KstR2-WT is supported by the projection onto the first eigenvector (Figure 4. 10, bottom row, left). Mtb-KstR2-WT and Mtb-KstR2-R170A also both have breathing motions, however, this motion is again more pronounced in the WT state, although to a lesser degree, with the range of values of the projection onto the first eigenvector similar in Mtb-KstR2-WT and Mtb-KstR2-R170A. The videos of the extreme motion projected onto the first principal component for all simulations can be found on FigShare (https://figshare.com/s/a3dd6b53d53211a8027b).

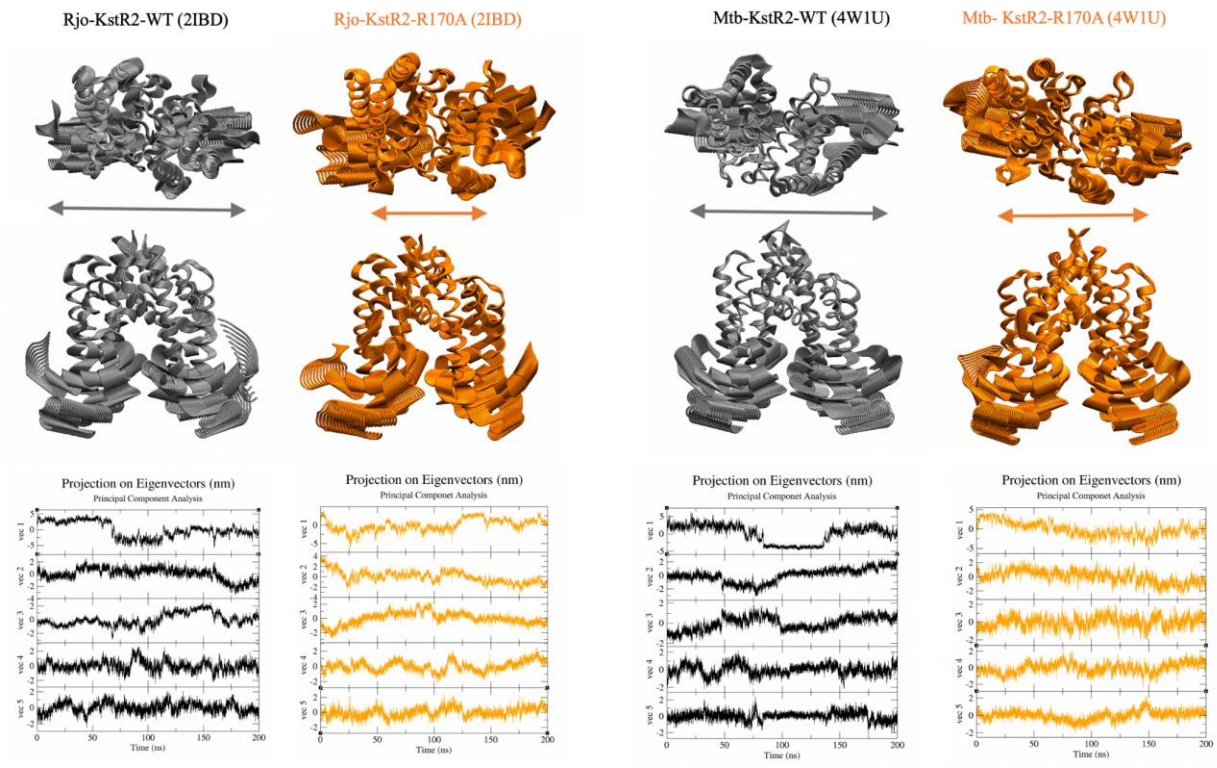


Figure 4. 10. (top row) Top view and (second row) front view of the extreme motion projected onto the first principal component and (bottom row) Comparison of the projections onto the first five (largest) eigenvectors for the simulations of (left) Rjo-KstR2-WT and Rjo-KstR2-R170A and (right) Mtb-KstR2-WT and Mtb-KstR2-R170A.

4.3.1.4. Dynamical network analysis

The PCA revealed the R170A mutation to not alter the nature of the conformational motions but to make them smaller in scale, in keeping with the RMSF and inter-DBD distance analysis. To provide more detailed insight into the origin of these differences, the VMD NetworkView plugin (225, 229) was used to compute cross-correlation matrices, from which “communities”

of residues that undergo correlated motion can be identified. This plugin also allows calculation of communication pathways between user-specified pairs of residues.

Cross correlation maps

It is noticeable that the R170A mutation in Rjo-KstR2 and Mtb-KstR2 results in less correlation (positive or negative). Additionally, Rjo-KstR2 exhibits areas with stronger positive correlations (darker red) than Mtb-KstR2 (Figure 4. 11).

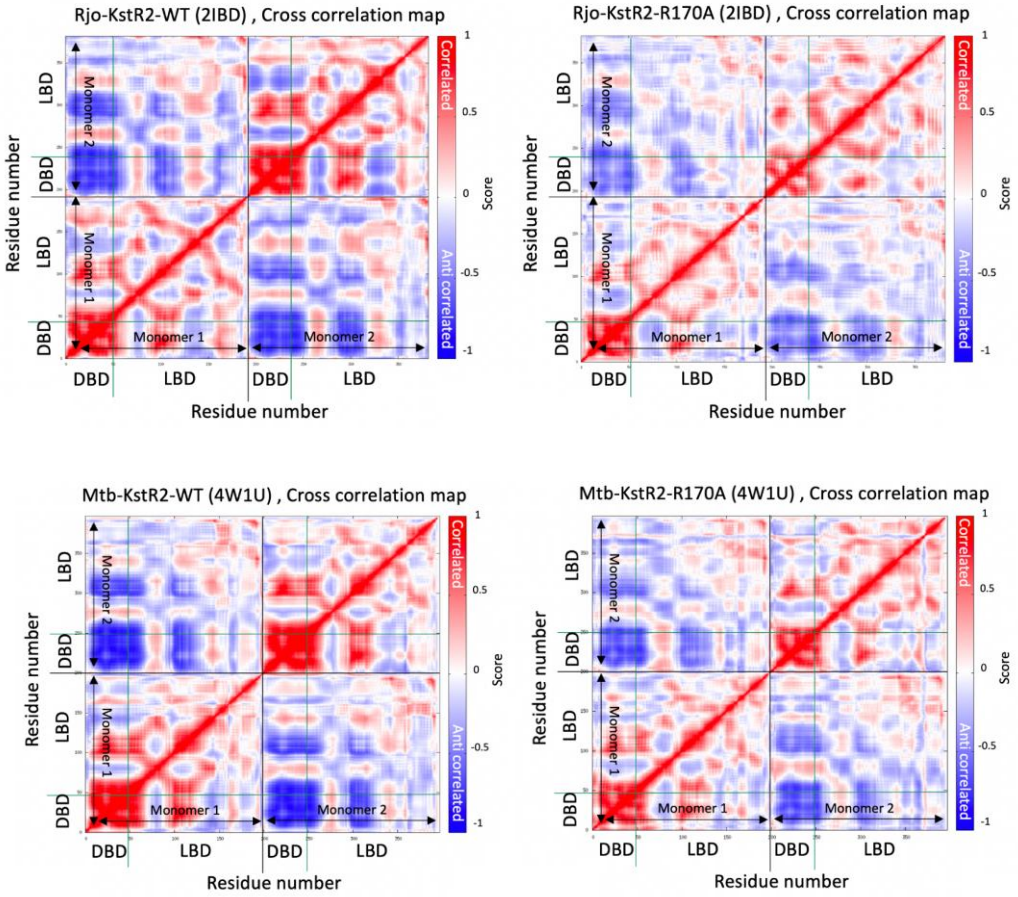


Figure 4. 11. Cross correlation maps for MD simulations of (top row, left) Rjo-KstR2-WT, (top row, right) Rjo-KstR2-R170A, (bottom row, left) Mtb-KstR2-WT, and (bottom row, right) Mtb-KstR2-R170A. In the colour scale, red represents correlated motion and blue represents anti-correlated motion. The residues corresponding to each monomer and to the DBD and LBD are indicated on each heat map. Cyan lines indicate the position of TYR54 (note that residue numbering here starts from 1 but in crystal structure it starts from 3 and 7 for Mtb-KstR2 (4W1U) and Rjo-KstR2 (2IBD), respectively).

While the overall pattern of correlations is generally similar between the WT and R170A states, Rjo-KstR2-R170A loses the positive intra-monomer correlations between residues 1-60 in the DBD (7-67 actual residue numbers) and 155-180 in LBD (162-187) and the positive inter-monomer correlations between residues 95-105 (102-112) of monomer A and 65-75 (72-82) in monomer B, both in the LBD, although these losses are not symmetric. In contrast, the regions

of positively correlated intra-monomer motion at the C-terminus of monomer B increase in size in Mtb-KstR2-R170A.

Network communities

The communities (groups of residues with strongly correlated (positively or negatively) motion) formed by KstR2 are displayed in Figure 4. 12 (Rjo-KstR2 and Mtb-KstR2). Edges are weighted according to the degree of correlated motion – the thicker an edge is, the stronger the connection, and *vice versa*. Each coloured region represents one community.

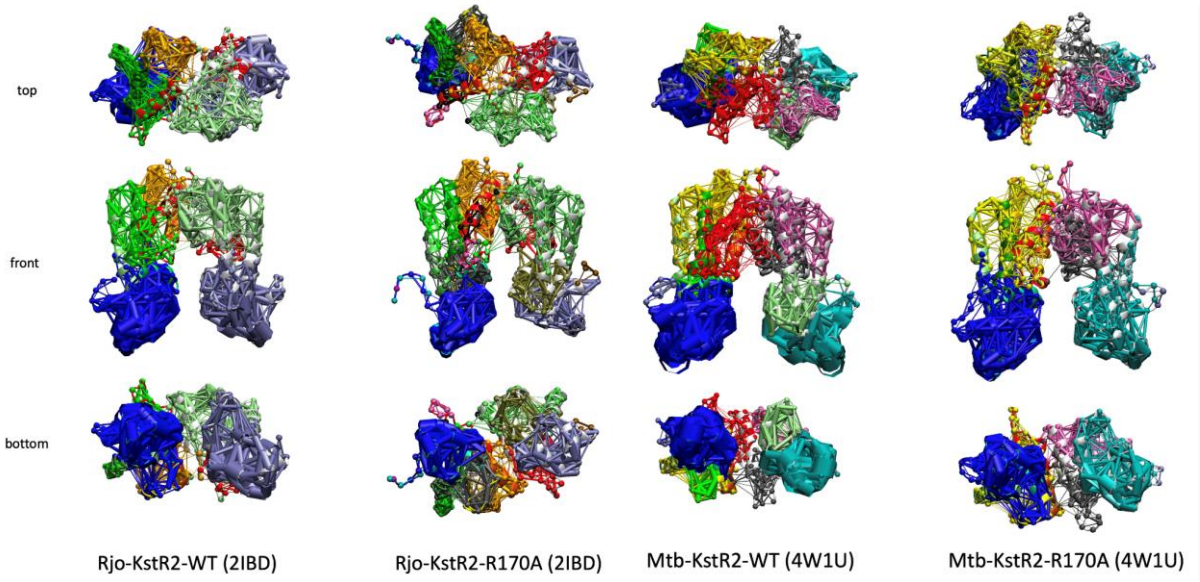


Figure 4. 12. Communities of residues that undergo correlated motions, viewed from the top, front and side, for (left) Rjo-KstR2-WT, (centre left) Rjo-KstR2-R170A, (centre right) Mtb-KstR2-WT, (right) Mtb-KstR2-R170A. Each community is coloured separately to match as closely as possible the colours used to illustrate the different domains and helices of KstR2 (Figure 4.1), with brighter colours used for monomer A. The connections between residues are weighted according to degree of correlated motion; the thicker the edges are, the stronger the connections and vice versa.

In Rjo-KstR2-WT, both DBDs move as rigid bodies and form their own communities (blue/pale blue), with very thick connections indicating strong correlations (Figure 4. 12). Monomer A is split into two communities, the LBD (green) and the interface (orange), whereas although monomer B also has two communities, these comprise the residues between the DBD and LBD (red) and the remainder of the LBD, including the interface (pale green). The R170A mutation results in the monomer B DBD splitting into two communities (pale blue/olive green), and a small interface community forming in the LBD (red). Therefore, it could be suggested

that Rjo-KstR2-WT has more coordinated motion, while in Rjo-KstR2-R170A form, the motion is less coordinated, which is likely to make message transmission harder.

For Mtb KstR2, however, Mtb-KstR2-WT has slightly more communities than Mtb-KstR2-R170A. In monomer B, the DBD is split into two communities, similar to Rjo-KstR2-R170A, although in monomer A, the DBD loop forms a separate (ice blue) community. Monomer A has three communities in the LBD (green, yellow, red) but monomer B has only two, one of which (pink) encompasses most of the LBD. This suggests greater mobility than Rjo-KstR2-WT. In Mtb-KstR2-R170A, each DBD is just one community, like Rjo-KstR2-WT. The LBD of monomer A forms just one community (yellow), but in monomer B, there are separate LBD (pink) and interface communities (grey). Thus, in Mtb-KstR2, the R170A mutation appears to decrease mobility, in agreement with the results of RMSD, RMSF, and PCA analyses.

Communication pathways between DBD and LBD

In order to investigate how information about ligand and DNA binding is transmitted through KstR2, the optimal and suboptimal communication pathways between selected key residues were calculated using the VMD NetworkView plugin (225, 229). This defines the shortest path between two residues as one that both comprises few residues and for which the motions of those residues are highly correlated. Here, putative communication pathways between ARG170 and the DBDs were calculated for simulation of different states of KstR2 to investigate the hypothesized scissor-like mechanism and arginine switch.

Residue ARG170 was chosen in both monomers as one end of the communication pathway because of its central position, at the intersection of the dimer interface and the ligand binding pocket, and in particular because it has been suggested to act as a switch that can change the conformation between a ligand-bound and apo state (226), thus causing the scissor-like motion (157). ARG170 appears to hold the DBDs closer together in Rjo KstR2, which resembles an apo state, and further apart in Mtb KstR2, which appears suitable for ligand binding (226). Residue TYR48 was chosen as the other residue in the communication pathway analysis, since it was chosen as a reference point between which to calculate the inter-DBD distance (in Section 4.3.1.2) and is a key determinant of the position of the DBDs.

The optimal and sub-optimal communication pathways for Rjo KstR2 and Mtb KstR2 in their WT and R170A states are shown in Figure 4. 13. For Rjo-KstR2-WT, the optimal paths are identical for the two monomers, running from ARG170 through ASP110 and ILE107 (helix

α6, bright green), PHE23 and THR31 (helix α1, dark blue) and VAL33 (helix α2, blue) to TYR48. The suboptimal pathways only differ within the DBD, close to TYR48. However, in Rjo-KstR2-R170, the paths are longer, and are no longer perfectly symmetrical. The optimal paths of both subunits now traverse helix α8 (orange) prior to helix α6 (bright green) and, in monomer B, the optimal path also returns to helix α9 (red). The lower part of both optimal paths is the same for both monomers and matches those observed for Rjo-KstR2-WT (Figure 4. 13).

For Mtb-KstR2-WT, similar to Rjo-KstR2-WT, the optimal paths in both monomers are identical. The series of residues is slightly different to Rjo-KstR2-WT, however, running from ARG170 through VAL106 (helix α6, bright green), ALA24 and MET22 (helix α1, dark blue), ASP35 and ALA37 (helix α2, blue), and LEU43 and GLY45 (helix α3, light blue) to TYR48. For Mtb-KstR2-R170A, the optimal pathways are different between monomers and to any of the other states. As for Rjo-KstR2-R170A, the paths include more residues in helix α9 (red), but the optimal path in monomer A includes only VAL106 in helix α6 (bright green), whereas in monomer B it includes three residues in this helix. The lower part of the monomer A path resembles those of 4W1U, other than the inclusion of SER44, but in monomer B, it includes LEU28 before resembling the lower parts of the pathways of Rjo-KstR2-R170A.

Together, the changes in the communication pathways upon mutation of ARG170 to ALA suggest that ARG170 plays an important role in transmitting information through KstR2.

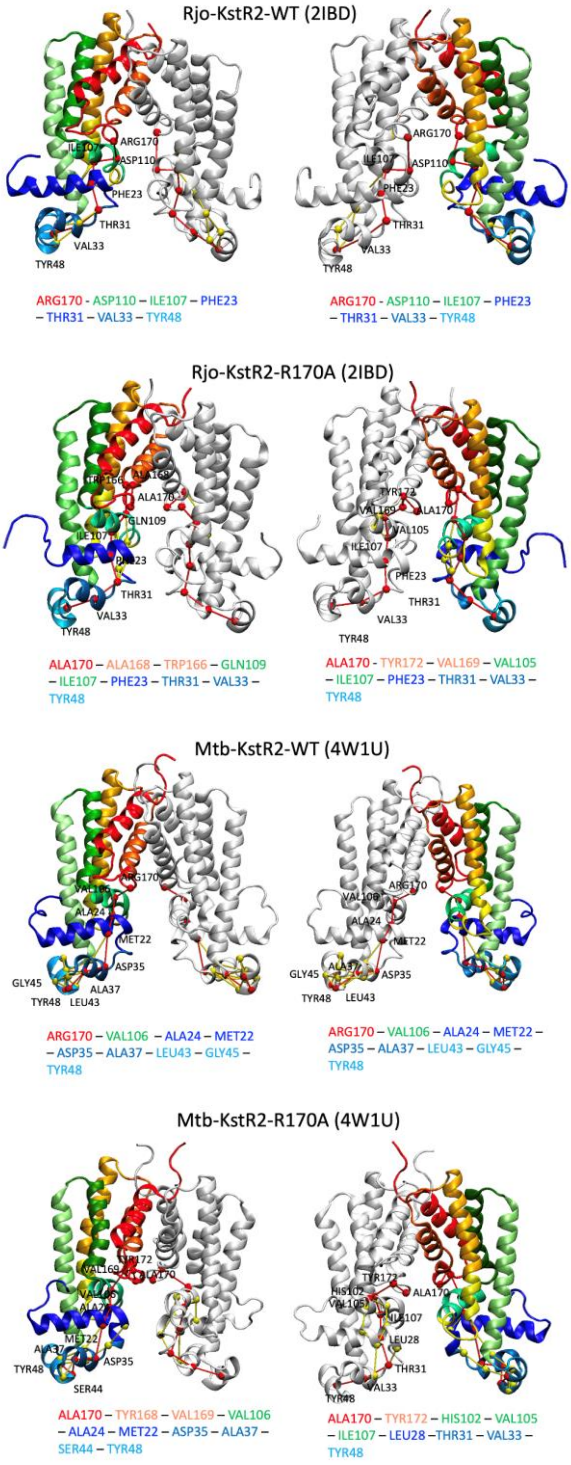


Figure 4. 13. Optimal (red) and suboptimal (yellow) communication pathways between ARG170 and TYR48 in both monomers of (top row) Rjo-KstR2-WT, (second row) Rjo-KstR2-R170A, (third row) Mtb-KstR2-WT and (bottom row) Mtb-KstR2-R170A. The coloured chain represents monomer A (the colour coding of the helices is the same as in Figure 4. 1) and the grey chain is monomer B. Key residues that make connections between helices along these paths are numbered according to the coordinate file.

4.3.2. Effect of ligand binding on KstR structure and dynamics

4.3.2.1. Structural stability

The atom-positional root-mean-square deviation (RMSD) of the Cα atoms of KstR2 from their initial positions (first frame of the simulation) at each point in time was again calculated to quantify how much the protein structure changes during the simulation time (Figure 4. 14, Table 4.3).

Mtb-KstR-2Hip-CoA-removed, Mtb-KstR-2HIP-CoA and Mtb-KstR-1HIP-CoA have a similar plateau RMSD of 0.3 nm and also have similar averages RMSD values (Table 4.3), although the RMSD of Mtb-KstR-2HIP-CoA increases in the last 20 ns of the simulation, suggesting that if the simulation were longer, it might sample higher RMSD states (Figure 4. 14, black, red and green). Other than this final jump, all three reach a stable RMSD value by approximately 75 ns. Therefore, the apo state does not appear to be more or less flexible than either of these ligand-bound states. However, for Mtb-KstR-2HIP, in which the CoA parts of the ligands were removed, although the RMSD again plateaued to around 0.3 nm initially, after 60 ns it underwent large fluctuations, including values up to 0.65 nm, other than a brief return to RMSD values around 0.3 nm at 150 ns (Figure 4. 14, blue), resulting in a much larger average RMSD value and standard deviation (Table 4.3). These large-scale conformational changes appear to be due to interactions between the first residues of the DBD (cyan coloured loop in Figure 4. 14) and the α1 helix (dark blue in Figure 4. 14) of each monomer. These motions are particularly evident in the principal component analysis (Section 4.3.2.3) and, to a lesser extent, in the inter-DBD distances (Section 4.3.2.2).

Table 4. 3 Averages and standard deviations of the RMSD values and the minimum distances between the TYR48 residue in each of the two monomers.

Simulation code	Average RMSD (nm)	Average minimum distance (nm) TYR48 (monomer A) – TYR48 (monomer B)
Mtb-KstR-2Hip-CoA-removed	0.30 ± 0.02	4.36 ± 0.26
Mtb-KstR-2HIP-CoA	0.32 ± 0.03	5.20 ± 0.22
Mtb-KstR-1HIP-CoA	0.29 ± 0.03	4.88 ± 0.23
Mtb-KstR-2HIP	0.43 ± 0.09	4.78 ± 0.28

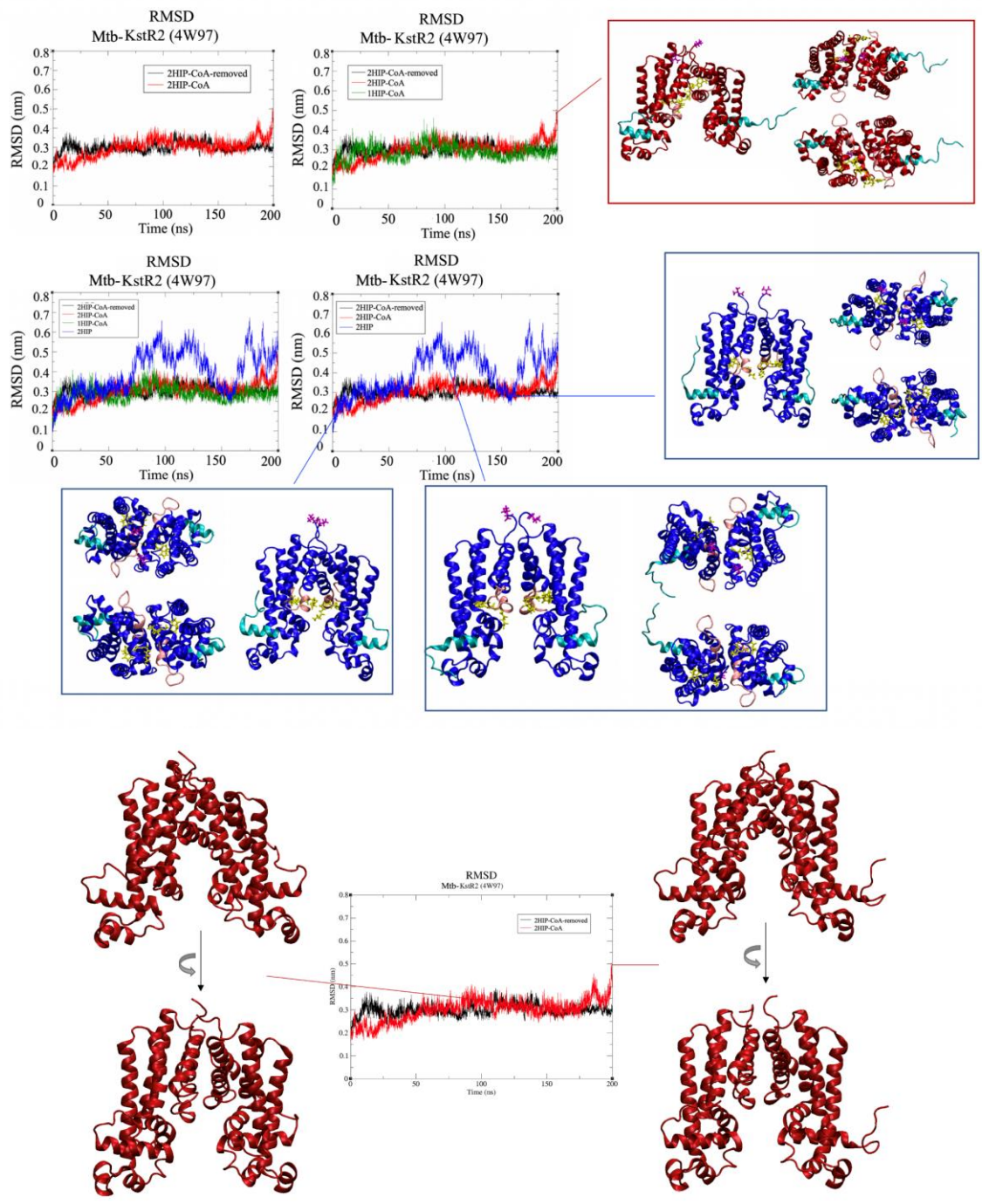


Figure 4. 14. RMSD time-series of the Mtb-KstR-2Hip-CoA-removed (black), Mtb-KstR-2HIP-CoA (red), Mtb-KstR-1HIP-CoA (green) and Mtb-KstR-2HIP (blue) MD simulations of KstR2. Surrounding the time-series are snapshots of the protein structure at the points in the simulation indicated, with KstR2 drawn in cartoon format and coloured according to the corresponding RMSD time-series, other than the pink, cyan and magenta regions, which represent the residues with the highest RMSF (see Section 4.3.2.3.). The ligands (HIP-CoA or HIP) and VAL200 in both monomers are drawn in liquorice format and coloured yellow and magenta, respectively.

The RMSF was computed to show which residues fluctuate the most on average across the simulation. The termini of KstR2 generally exhibit higher RMSF values, as is commonly see for the termini of proteins during MD simulations. The RMSF values of the N-terminal residues, which form the first half of helix α1 in the DBD (Figure 4. 15, cyan regions, residues

1-20, in both monomers), are low for Mtb-KstR-2Hip-CoA-removed in both monomers, low for Mtb-KstR-1HIP-CoA for monomer A (to which the HIP-CoA is not bound), moderately high for Mtb-KstR-2HIP-CoA for both monomers, and highest for Mtb-KstR-2HIP for both monomers. These results suggest that a monomer in an apo state exhibits reduced dynamics of the N-terminal residues, with dynamics increasing when ligand is bound to that monomer and highest when the ligand is missing CoA. For the C-terminal residues (Figure 4. 15, magenta regions, VAL200 in both monomers), which form loops at the non-DNA-binding end of KstR2, the RMSF values are similarly high in all systems.

The residues circled in pink in Figure 4. 15 (residues 166-179, in both monomers) have higher than baseline RMSF for Mtb-KstR-2Hip-CoA-removed in monomer A but for Mtb-KstR-2HIP-CoA in monomer B. Mtb-KstR-1HIP-CoA, which has HIP-CoA bound only to monomer A, has higher RMSF in monomer A (like Mtb-KstR-2Hip-CoA-removed) and in monomer B (like Mtb-KstR-2HIP-CoA). These residues are at the base of helix α8, which forms the top of the ligand binding pocket. This intriguing result suggests that with a ligand in each binding pocket, KstR2 in fact undergoes asymmetric dynamics of the ligand binding pocket, with more motion in monomer B. The Mtb-KstR-2Hip-CoA-removed state, which is also symmetric with respect to ligand binding, also undergoes asymmetric dynamics of the ligand binding pocket, but in the opposite manner, with more motion in monomer A. Mtb-KstR-1HIP-CoA, which is asymmetric with respect to ligand binding, undergoes symmetric dynamics of the ligand binding pocket. When CoA is removed from the ligands (Mtb-KstR-2HIP), the conformational changes of residues 166-179 show Mtb-KstR-2Hip-CoA-removed-like asymmetry, suggesting that the CoA portion in some way stimulates this dynamic response to ligand binding. More details in terms of the interactions between the protein and ligand associated with the RMSF differences are explained in Section 4.3.2.5.

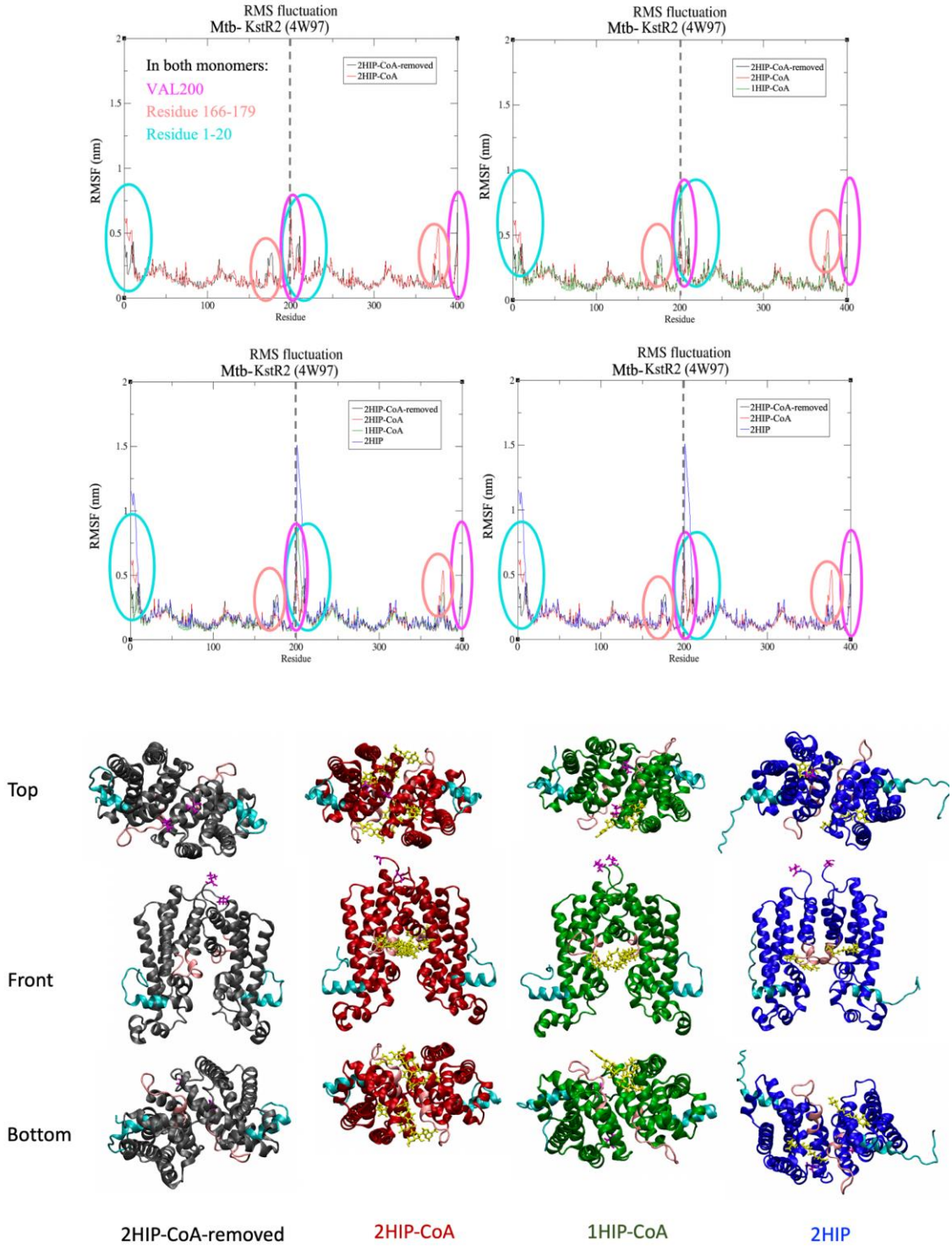


Figure 4. 15. (Top) RMSF of the Mtb-KstR2-2Hip-CoA-removed (black), Mtb-KstR2-2HIP-CoA (red), Mtb-KstR2-1HIP-CoA (green) and Mtb-KstR2-2HIP (blue) MD simulations of KstR2. The residues were renumbered for this analysis in order to have a continuous plot (1-400, instead of 1-200,1-200). (Bottom) Snapshots from each simulation showing three different views of KstR2, top, front and bottom, with KstR2 drawn in cartoon format and coloured the same as the RMSF data. The residues in the regions circled in the RMSF plots are coloured accordingly in the snapshots, as well as are written on the top left graph. The ligand is drawn in liquorice and coloured yellow. Dashed lines separate the two monomers.

4.3.2.2. Separation of DNA-binding domains

In order to investigate the DBD distance in the simulations, which is a key feature of the apo, ligand-bound and DNA-bound states of KstR2, the Cα atoms of two TYR residues (TYR48 in both monomers) in the DBDs were chosen as reference points between which to calculate the inter-DBD distance.

The distance between the DBDs increases from its initial value of ~4.8 nm during the MD simulation of Mtb-KstR2-2HIP-CoA to give an average distance of 5.20 nm, and decreases during the MD simulation of Mtb-KstR2-2Hip-CoA-removed to give an average distance of 4.36 nm (Figure 4. 16, Table 4.3). For Mtb-KstR2-1HIP-CoA and Mtb-KstR2-2HIP, there is no overall change in the distance but there are modulations larger than the fast time-scale fluctuations. Overall, more and larger ligands tend to increase the inter-DBD distance and *vice versa* (Table 4.3), suggesting that ligand binding may act to hold the DBDs apart. Given greater simulation time, it is possible that Mtb-KstR2-2Hip-CoA-removed might sample the short inter-DBD distances required for DNA binding.

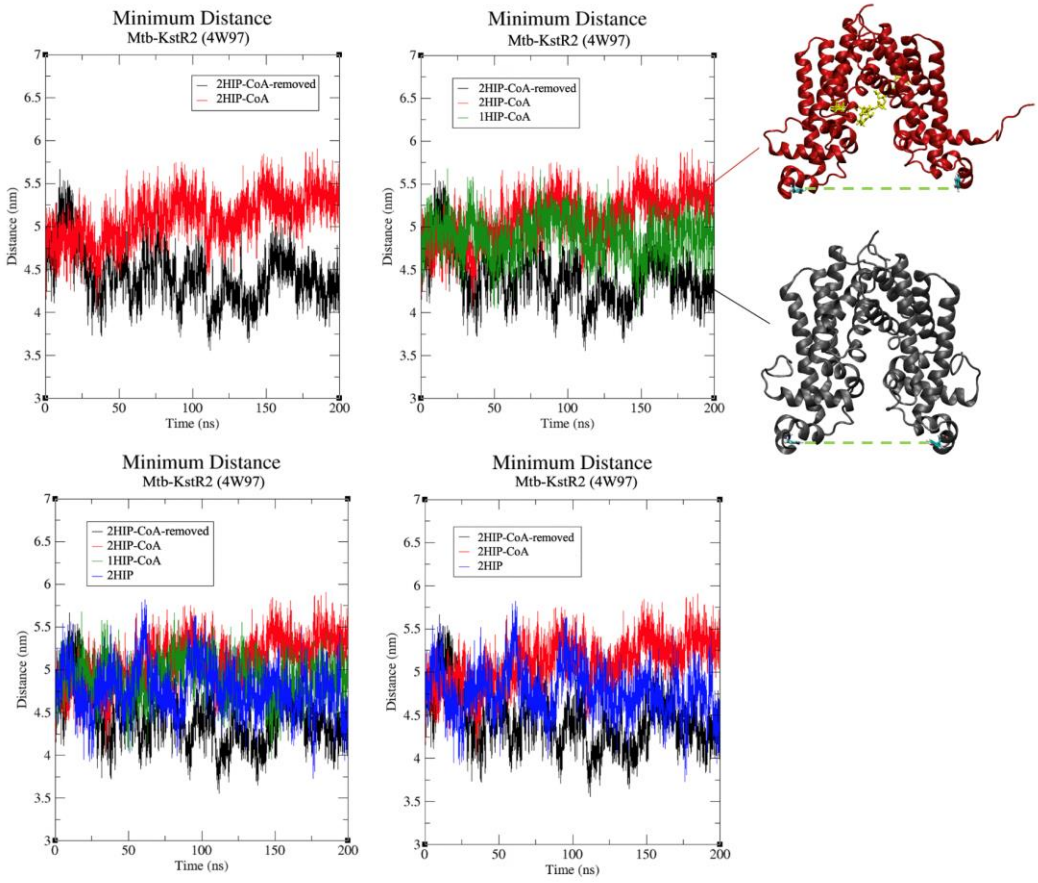


Figure 4. 16. Time-series of the distance between the Cα atoms of TYR54 of each monomer of KstR2 during the Mtb-KstR2-2Hip-CoA-removed, Mtb-KstR2-2HIP-CoA, Mtb-KstR2-1HIP-CoA and Mtb-KstR2-2HIP MD simulations, and snapshots from each simulation, with KstR2 drawn in cartoon format and coloured the same as the RMSF data. The ligand is drawn in liquorice and coloured yellow, TYR48 is drawn in liquorice and coloured according to atom type (cyan, carbon; red, oxygen; white, hydrogen; blue, nitrogen) and the distance that was measured is indicated with a green dashed line.

4.3.2.3. Determination of concerted motions

To determine the major and concerted conformational motions that KstR2 undergoes during each simulation, cartesian covariance and principal component analysis (PCA) were carried out for each. Visualisation of the extreme projection onto the first principal component showed Mtb-KstR2-2Hip-CoA-removed to undergo large breathing motions of the DBDs (Figure 4. 17, left). Similar motion was observed for apo Rjo KstR2 and Mtb KstR2, suggesting that this large-scale breathing motion is typical of apo states of KstR2. In contrast, the ligand-bound states all exhibited predominantly a twisting motion and only a minor breathing motion (Figure 4. 17, centre and right). This twisting is moderate for Mtb-KstR2-2HIP-CoA, but increases and becomes asymmetric upon removal of one HIP-CoA (Mtb-KstR2-1HIP-CoA). The monomer that has the HIP-CoA bound (Figure 4. 17, centre right, left-hand monomer) showed less movement than the monomer without the ligand (Figure 4. 17, centre right, right-hand

monomer), other than the N-terminus of the DBD. This region also showed very large motions in both monomers of Mtb-KstR2-2HIP. Removing the CoA from the ligand therefore appears to have the same effect on KstR2 dynamics as removing the ligand entirely. This also could be seen in the inter-DBD distance and RMSD analysis. The videos of the extreme motion projected onto the first principal component for all simulations can be found on FigShare (https://figshare.com/s/a3dd6b53d53211a8027b).

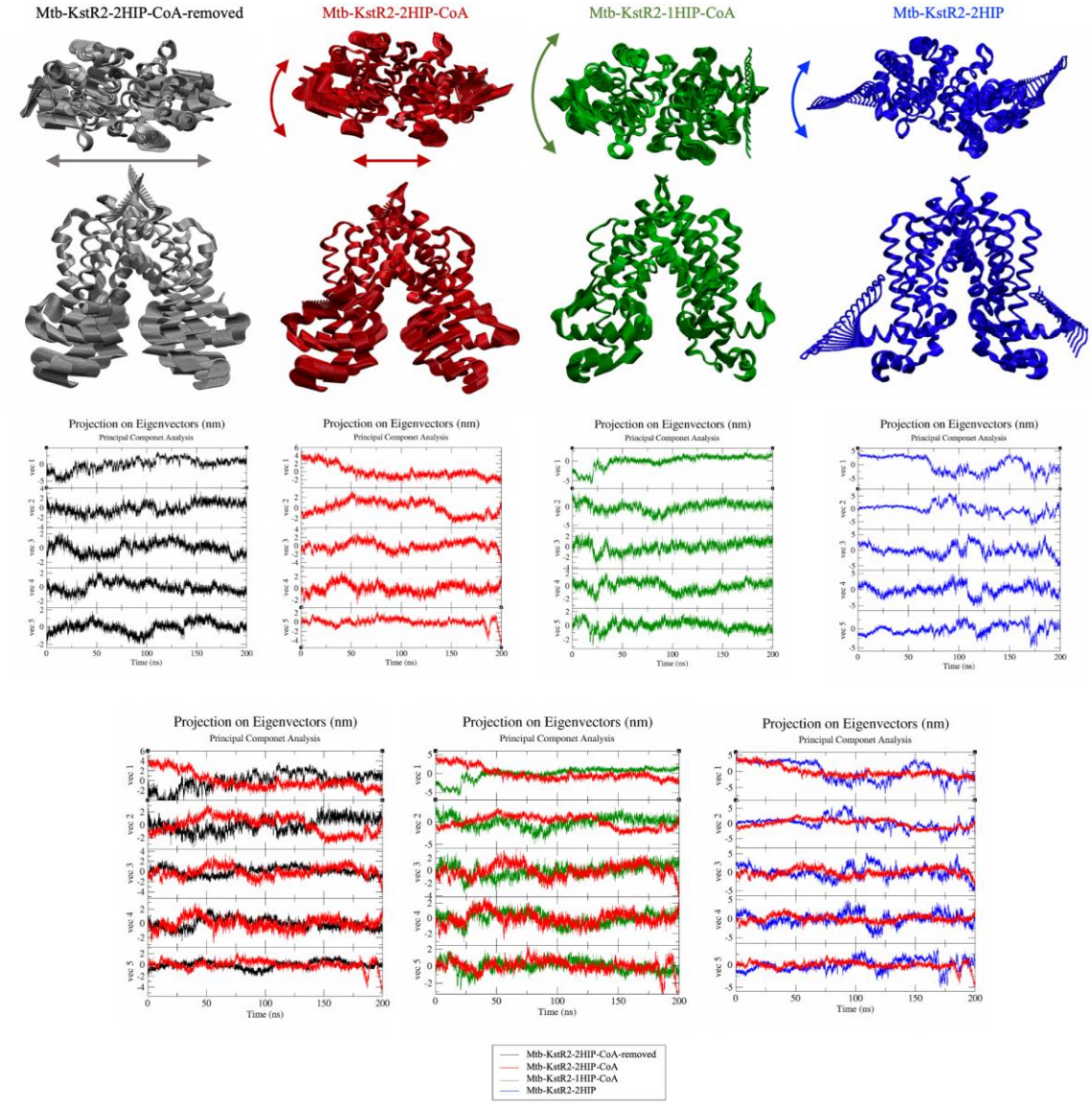


Figure 4. 17. (Top row) Top view and (second row) front view of the extreme motion projected onto the first principal component and (third row) projection onto the first five (largest) eigenvectors during the MD simulations of (black, left) Mtb-KstR2-2Hip-CoA-removed, (centre left, red) Mtb-KstR2-2HIP-CoA, (centre right, green), Mtb-KstR2-1HIP-CoA and (right, blue) Mtb-KstR2-2HIP. (Bottom row) Comparison of the projections onto the first five (largest) eigenvectors for the simulations of (left) Mtb-KstR2-2Hip-CoA-removed and Mtb-KstR2-2HIP-CoA, (centre) Mtb-KstR2-2HIP-CoA and Mtb-KstR2-1HIP-CoA and (right) Mtb-KstR2-2HIP-CoA and Mtb-KstR2-2HIP.

4.3.2.4. Dynamical network analysis

The PCA revealed the apo state to undergo a breathing motion of the DBDs, but the ligand-bound state to undergo more of a twisting of the DBDs with respect to the LBDs. Removing only the CoA part of the ligand has the same effect as removing the entire ligand. To provide more detailed insight into the origin of these differences, cross-correlation matrices were computed from which "communities" of residues that undergo correlated motion can be identified. Communication pathways between ARG170 and TYR54 were also computed.

Cross-correlation maps

In the cross-correlation maps for all simulations of Mtb KstR2, the DBD and LBD appear to be separate rigid bodies with anti-correlated motions, although the degree of this difference varies (Figure 4. 18). The presence of one or two ligands led to less correlation in general. This matches the PCA results showing that ligand binding makes the structure less mobile (Section 4.3.2.3). The greatest reduction in correlation occurs for Mtb-KstR2-1HIP-CoA. This also matches the RMSD results, as Mtb-KstR2-1HIP-CoA has the lowest RMSD values, although the difference between the RMSD values of different states is not great.

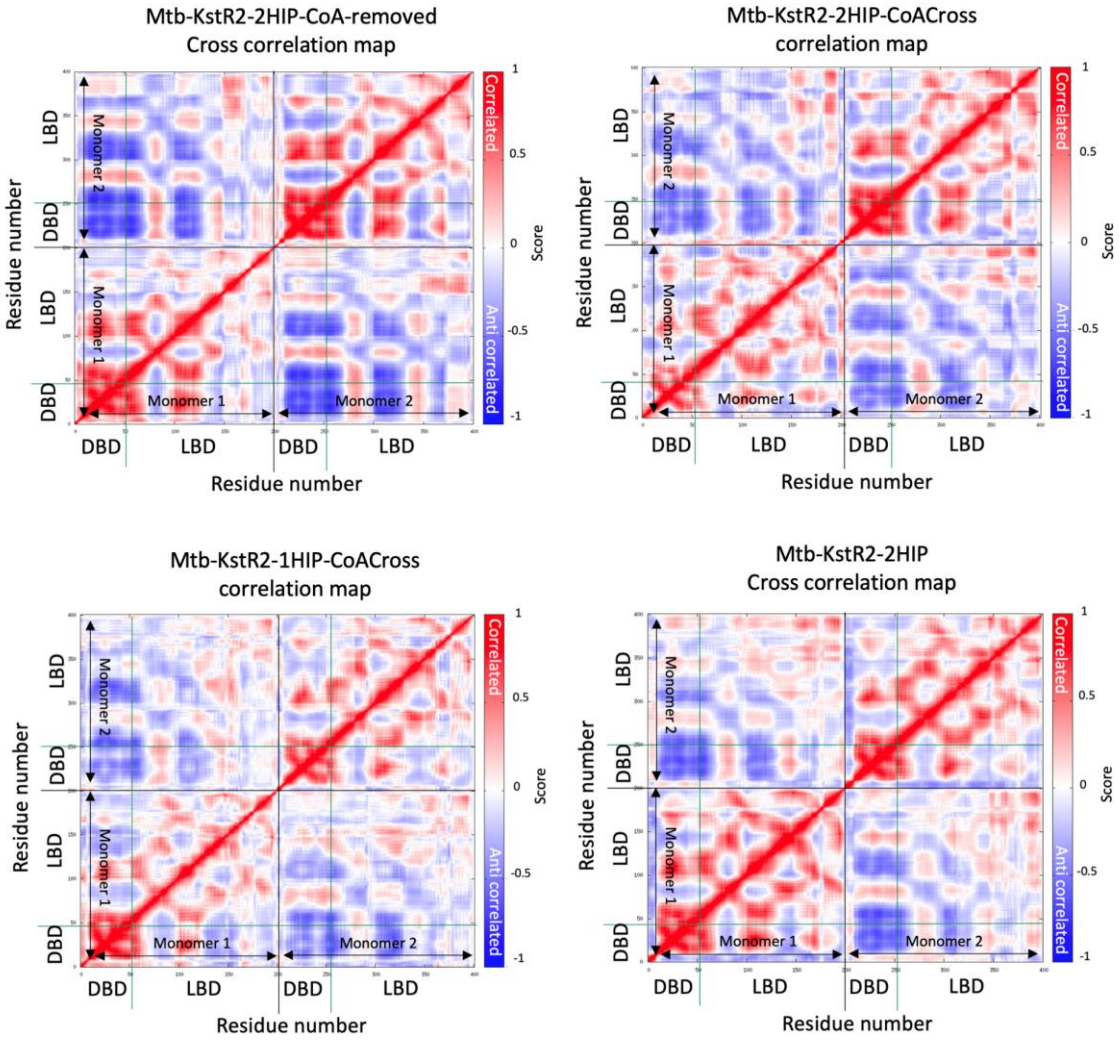


Figure 4. 18. (previous page) Cross correlation maps for MD simulations of (top row, left) Mtb-KstR2-2Hip-CoA-removed, (top row, right) Mtb-KstR2-2HIP-CoA, (bottom row, left) Mtb-KstR2-1HIP-CoA and (bottom row, right) Mtb-KstR2-2HIP. In the colour scale, red represents correlated motion and blue represents anti-correlated motion. The residues corresponding to each monomer and to the DBD and LBD are indicated on each heat map. Cyan lines indicate the position of TYR54.

For Mtb-KstR2-2HIP-CoA, it is mostly the positive correlation in monomer A that is reduced, with the DBD no longer moving as a single rigid body, and residues 95-130 (the lower half of the LBD 'triangle') also no longer moving in concert with the DBD. This asymmetry is in keeping with the RMSF results (Section 4.3.2.3), which showed that states with symmetric ligand binding exhibit asymmetric dynamics. Removing one ligand restores symmetry to the cross-correlation map, which is again in keeping with the RMSF results. Removing the CoA part from HIP (Mtb-KstR2-2HIP-CoA) predominantly results in less correlation (both positive

and negative) between the monomers, and reduces the correlation within monomer B, but increases the correlation within monomer A.

Network communities

The communities (groups of residues with strongly correlated (positively or negatively) motion) formed by Mtb KstR2 with and without different ligands bound are displayed in Figure 4. 19. Edges are weighted according to the degree of correlated motion – the thicker an edge is, the stronger the connection, and *vice versa*. Each coloured region represents one community. For Mtb KstR2 (Figure 4. 19), there are two communities per monomer that are always present, with the DBD forming one (blue/cyan in all images) and the LBD (yellow for monomer A in all images, and green or pink in monomer B). The colours of the communities are not consistent due to the different numbers of communities in different states. Interestingly, however, the LBD community of monomer B extends across the dimer interface to include the upper part of helix α9 in monomer A. This shows that the two monomers are tightly connected at the top, and also could relate to the scissor-like motion of KstR2. This link across the dimers is consistent in the communities found for all the KstR2 simulations.

For Mtb-KstR2-2Hip-CoA-removed, these, along with an interface community in each monomer (red/orange), are the only major communities present, other than two small communities formed by 1) the top of helix α4 in monomer B (black), and 2) the lower loop of helix α9 and lower part of helix α5 (ice blue).

With two HIP-CoA ligands present (Mtb-KstR2-2HIP-CoA), monomer A has an additional community that encompasses the residues between the DBD and LBD (Figure 4. 19). Interestingly, part of the loop extending out from the DBD of monomer A is included in the interface community (red), indicating that its motions are more correlated with the dimer interface than with the DBD to which it is directly attached. The links joining these two parts are only weak, however, and the remainder of the loop is in its own community (ice blue). Monomer B, however, only has three communities, having lost the interface community and absorbed the top of helix α4 into the LBD community (pink). This asymmetry is also seen in the RMSF (Section 4.3.2.1) and cross-correlation analysis, with the latter also showing monomer B to have better defined and larger rigid bodies.

For Mtb-KstR2-1HIP-CoA, which only has HIP-CoA bound to monomer A, the DBD loop does not extend outwards, and is split between the DBD (blue) and lower LBD (green) community. Monomer A has only these three communities, losing the interface (red)

community that is present in Mtb-KstR2-2Hip-CoA-removed and Mtb-KstR2-2HIP-CoA. Monomer B has four communities, but only two major ones – the DBD and the LBD. The DBD loop forms a separate community like it does for Mtb-KstR2-2Hip-CoA-removed (ice blue in both cases), and residues 72-79 (the linker loop between helices α8 and α9) also form a separate community, like they do in Mtb-KstR2-2HIP-CoA (black in both cases). The DBD communities in both monomers are similar to in Mtb-KstR2-2HIP-CoA, with that of monomer B also similar to in Mtb-KstR2-2Hip-CoA-removed monomer B. The low number of communities observed for Mtb-KstR2-1HIP-CoA is unexpected given the weak correlations observed in the cross-correlation maps. The asymmetry is in contrast to the symmetry observed in the RMSF results, but matches the asymmetric PCA motions.

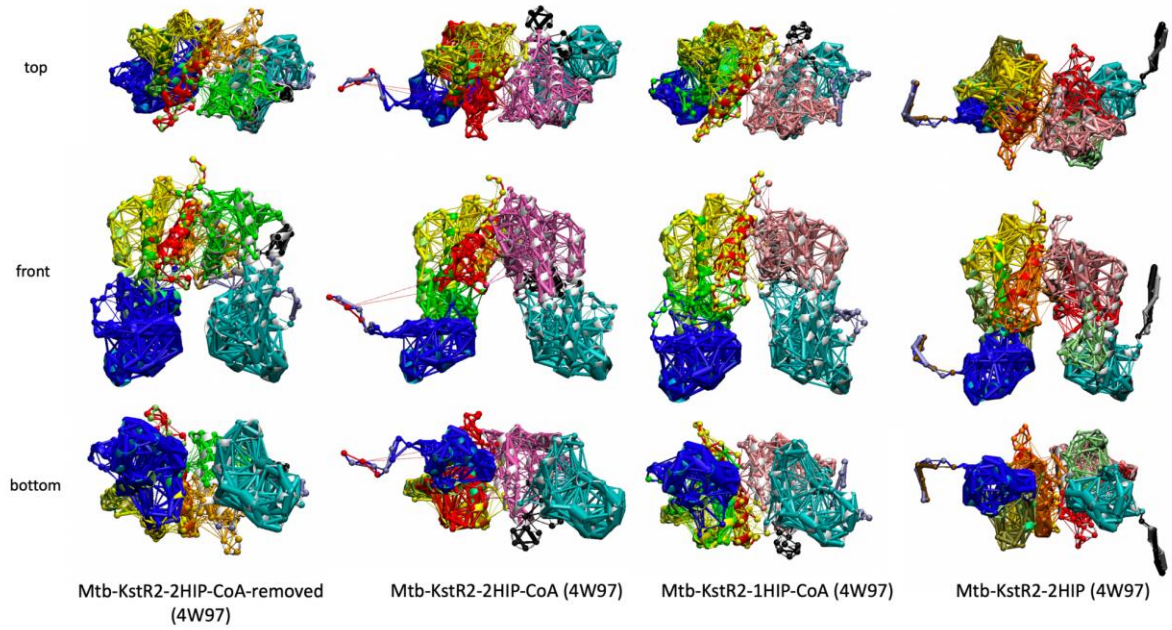


Figure 4. 19. Communities of residues that undergo correlated motions, viewed from the top, front and side, for (left) Mtb-KstR2-2Hip-CoA-removed, (centre left) Mtb-KstR2-2HIP-CoA, (centre right) Mtb-KstR2-1HIP-CoA, (right) Mtb-KstR2-2HIP. Each community is coloured separately to match as closely as possible the colours used to illustrate the different domains and helices of KstR2 (Figure 4.1.) with brighter colours used for monomer A. The connections between residues are weighted according to degree of correlated motion; the thicker the edges are, the stronger the connections and vice versa.

In Mtb-KstR2-2HIP, the DBD communities are small, and the DBD loops form separate communities, as is seen for monomer A of Mtb-KstR2-2HIP-CoA. In both monomers, the remainder of the DBD forms a separate community that extends up into the LBD (olive/pale green). The LBDs in both monomers are split into two communities (yellow/orange in monomer A, pink/red in monomer B), similar to the separate LBD and interface communities observed for Mtb-KstR2-2Hip-CoA-removed. Similar behaviour of these two states was also

seen in the PCA results. Overall, Mtb-KstR2-2HIP is more symmetric, which is in keeping with the RMSF results (Section 4.3.2.1).

Communication pathways between DBD and LBD

In order to investigate how information about ligand and DNA binding is transmitted through KstR2 and whether a scissor-like mechanism with ARG170 playing a key role as a switching residue is supported, the optimal and suboptimal communication pathways between ARG170 and TYR48 were computed. The optimal and suboptimal pathways for both monomers of KstR2 in a ligand-removed apo state and bound to one or two HIP-CoA molecules or two HIP molecules are shown in Figure 4. 20.

For Mtb-KstR2-2Hip-CoA-removed, the pathways are not symmetric across the monomers, although many of the sub-optimal paths in monomer A are similar to the optimal path in monomer B. The optimal pathway in monomer A travels from ARG170 through VAL106 (helix α6, bright green), ALA24 and ALA19 (helix α1, dark blue), MET57 (helix α4, light green) and PHE51 (helix α3, light blue) to TYR48. The optimal path in monomer B runs from ARG170 through ASP110 and ILE107 (helix α6, bright green) to VAL58 and GLU56 (helix α4, light green), and PHE51 (helix α3, light blue) to TYR48.

Addition of HIP-CoA into both ligand pockets results in both optimal paths changing and becoming longer, but they remain asymmetric, with the optimal path in monomer B now passing through helix α2 (blue) and that of monomer A passing predominantly through residues in helix α6 (bright green). Thus, the asymmetry observed for the Mtb-KstR2-2Hip-CoA-removed and Mtb-KstR2-2HIP-CoA states in the RMSF analysis is also reflected in their communication pathways.

Removing one ligand (from monomer B) makes the two paths perfectly symmetrical. This is in keeping with previous results that showed Mtb-KstR2-1HIP-CoA to undergo symmetric motion. The optimal paths travel from ARG170 through VAL106 (helix α6, bright green), ALA24 and MET22 (helix α1, dark blue), ASP35 and ALA37 (helix α2, blue) and SER44 (helix α3, light blue) to TYR48. This is the same as monomer B of Mtb-KstR2-2HIP-CoA, consistent with the results of RMSF and communities.

For Mtb-KstR2-2HIP, the paths are again asymmetric. In monomer A, the optimal pathway is the same as that of monomer B of Mtb-KstR2-2Hip-CoA-removed, and for monomer B, it is similar but slightly different to those of Mtb-KstR2-1HIP-CoA and monomer B of Mtb-KstR2-2HIP-CoA, similar to RMSF results.

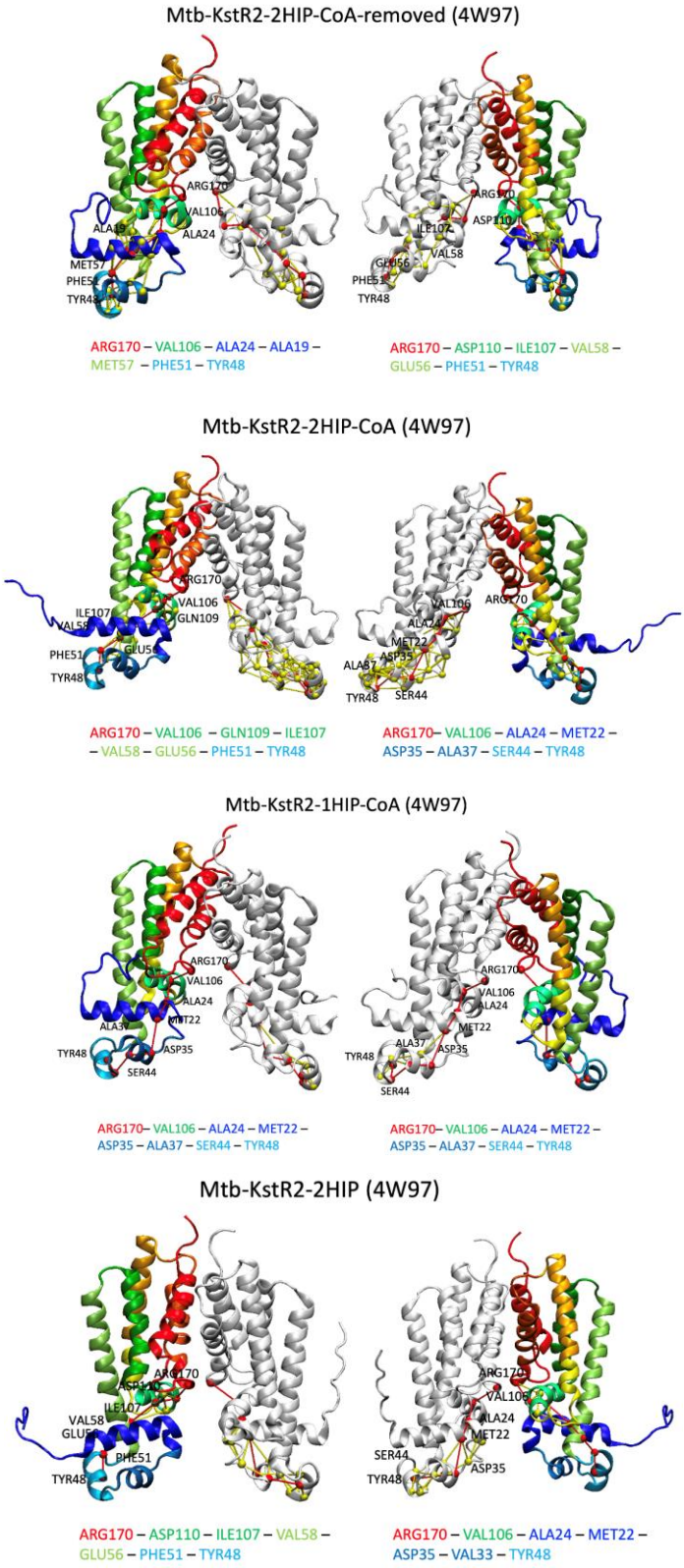


Figure 4. 20. Optimal (red) and suboptimal (yellow) communication pathways between ARG170 and TYR48 in both monomers of (top row) Mtb-KstR2-2Hip-CoA-removed, (second row) Mtb-KstR2-2HIP-CoA, (third row) Mtb-KstR2-1HIP-CoA and (bottom row) Mtb-KstR2-2HIP. The coloured chain represents monomer A (the colour coding of the helices is the same as in Figure 4. 1) and the grey chain is monomer B. Key residues that make connections between helices along these paths are numbered according to the coordinate file.

4.3.2.5. Hydrogen bond formation

While the communication pathways, which comprise residues whose motion is strongly coupled, report on information transfer through the protein, the formation of hydrogen bonds is crucial for stable interactions. The CoA of the steroid ligands has been shown to enhance the ligand affinity for KstR (135), and KstR2 (*Pers. Commun.* Stephanie Dawes, The University of Auckland). Like for the KstR ligand (3OCh), the CoA tail in HIP-CoA ligand may also increase ligand binding by improving the ligand solubility (135). However, the role of the CoA moiety in the KstR-ligand interaction has been an unresolved issue (135). To shed light on this, the hydrogen bonds formed during the simulations between monomers and between KstR2 and its ligands, which here were simulated both with and without the CoA moiety, were therefore computed for the Mtb KstR2 simulations with one or more ligands bound, in the first instance, focusing on hydrogen bonds present for at least 20% of the simulation time.

The greatest number of hydrogen bonds were formed for Mtb-KstR2-2HIP-CoA (Figure 4. 21), which is unsurprising, given the larger size of HIP-CoA relative to HIP and the fact that it has two ligands bound rather than just one. For Mtb-KstR2-2HIP-CoA, two hydrogen bonds were formed consistently during the simulation time: those between ARG173 (oxygen of the backbone carbonyl group) of monomer A and the CoA part of the ligand (nitrogen N64) (Figure 4. 22) in monomer B ligand pocket, and between GLN109 (side chain nitrogen) of monomer A and the HIP part of the ligand (5-carbonyl oxygen, O7) in monomer B (Figure 4. 21, red and blue, respectively). Only the hydrogen bond with GLN109 is present in the crystal structure of 4W97 (128) (Figure 4. 23). This hydrogen bond remains when either the ligand in monomer B (Figure HB3) or the CoA parts of the ligands are removed (Figure HB4), but that involving ARG173 disappears. TYR158, which does not form hydrogen bonds with the ligand in the crystal structure, forms a hydrogen bond with the CoA part of the ligand in the monomer A pocket in Mtb-KstR2-2HIP-CoA. These form at approximately the middle of the Mtb-KstR2-2HIP-CoA simulation and remain until the end (Figure 4. 21, orange and green, respectively), suggesting that they are able to form due to changes in conformation of the ligand and/or protein or in the ligand binding mode that occur during the simulation.

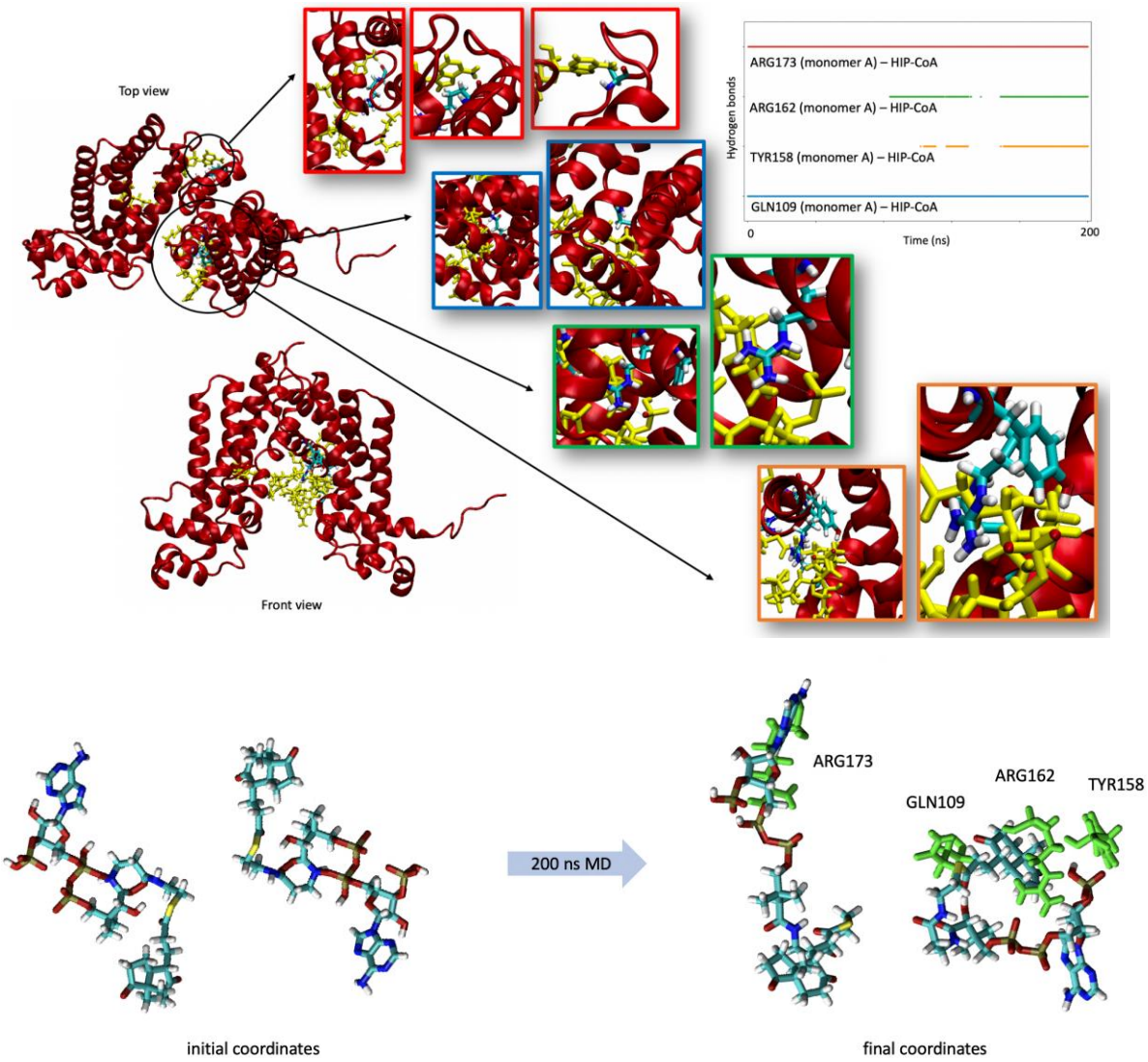


Figure 4. 21. Hydrogen bonds between KstR2 and HIP-CoA occupied for 20-100% of the Mtb-KstR2-2HIP-CoA simulation. Top panel: top and front view of KstR2 with insets showing the details of the interaction between each of the four residues that forms a hydrogen bond with HIP-CoA. The protein backbone is drawn in red cartoon format, the ligands are drawn in yellow liquorice format, and the residues with which they form hydrogen bonds are drawn in liquorice format and coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen; gold: phosphorous). The colours of the outlines of these insets match the colours of their respective lines in the time-series (left-hand side). Bottom panel: initial and final structures of the two HIP-CoA molecules and, for the final structures, the residues with which they form hydrogen bonds showed by green. The ligands and residues are drawn in liquorice format, with the ligands coloured according to atom type and the residues in green. The screenshots of the protein are from the endpoint of the 200 ns MD simulation.

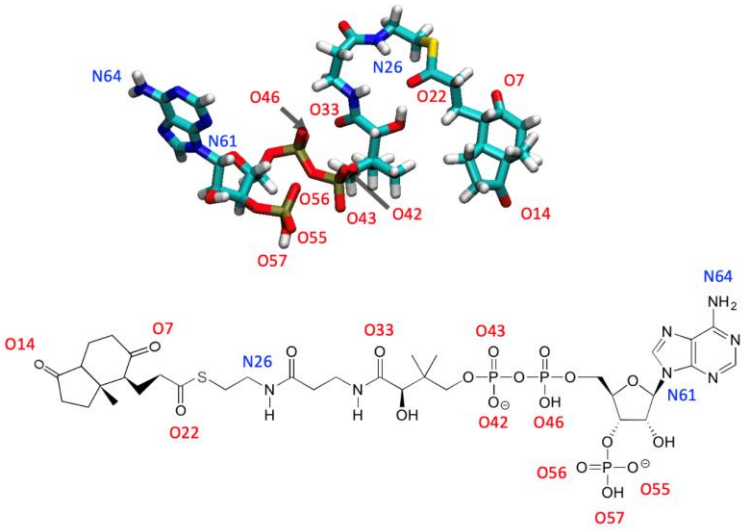


Figure 4. 22. (Top) Liquorice and (bottom) 2D representations of HIP-CoA ligand. The oxygen atoms involved in hydrogen bonds are labelled. The liquorice structure is coloured according to atom type (cyan: carbon; red: oxygen; blue: nitrogen; white: hydrogen).

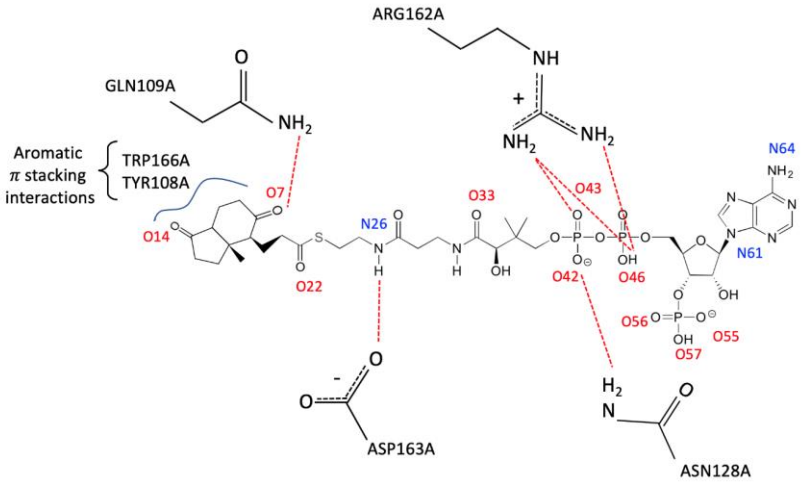


Figure 4. 23. 2D structure of HIP-CoA and its interactions (hydrogen bonds) with residues of KstR2 in the 4W97 crystal structure, according to the experimental data (137, 158). Two aromatic stacking interactions are also shown.

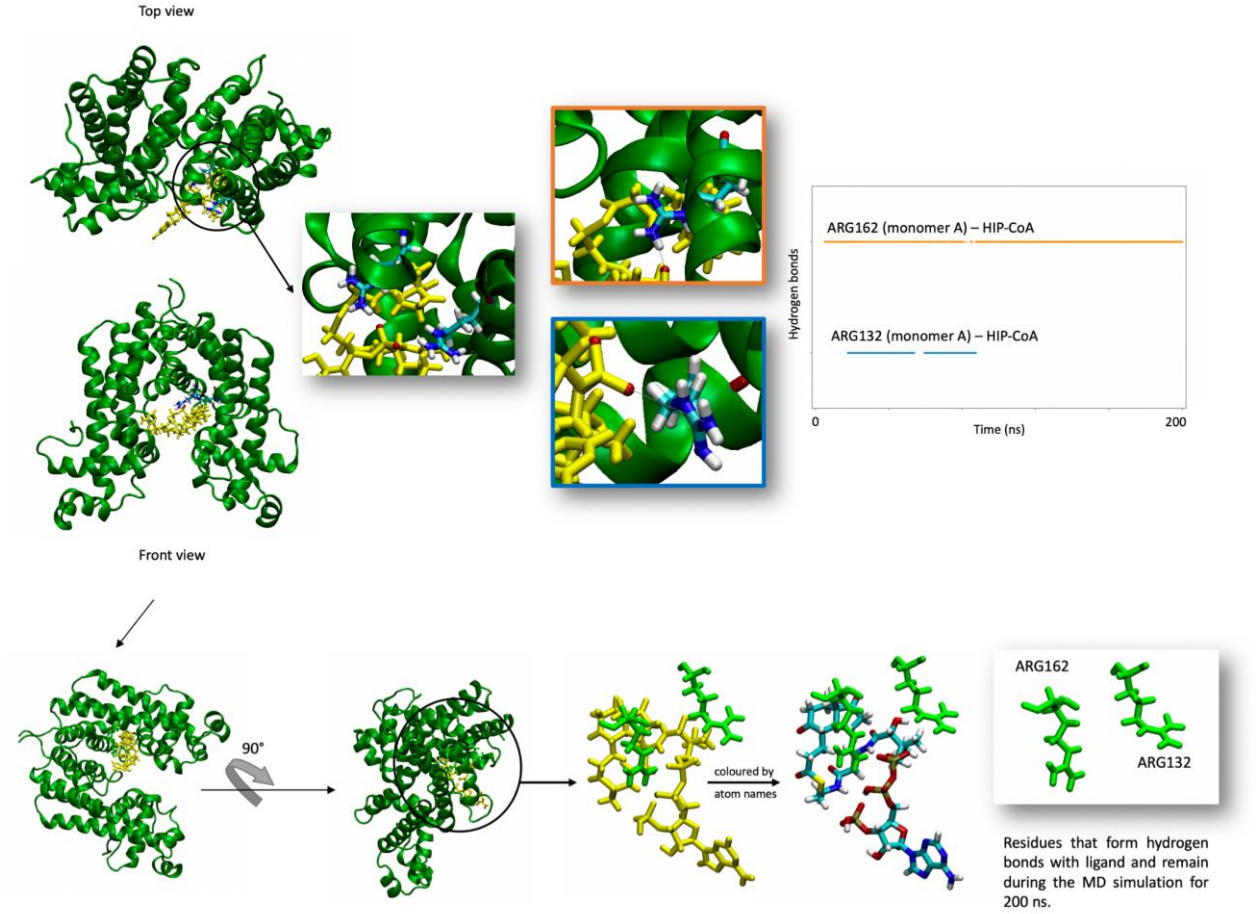


Figure 4. 24. Hydrogen bonds between KstR2 and HIP-CoA occupied for 20-100% of the Mtb-KstR2-1HIP-CoA simulation. Top panel: top and front view of KstR2 with insets showing the details of the interaction between each of the two residues that form a hydrogen bond with HIP-CoA. The protein backbone is drawn in green cartoon format, the ligands are drawn in yellow liquorice format, and the residues with which they form hydrogen bonds are drawn in liquorice format and coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen; gold: phosphorous). The colours of the outlines of these insets match the colours of their respective lines in the time-series (right-hand side). Bottom panel: Structures of the HIP-CoA molecule and the residues with which it forms hydrogen bonds. In the right-most structure, the ligand is coloured according to atom type and the residues that hydrogen bond to it are in green. The screenshots of the protein are from the endpoint of the 200 ns MD simulation.

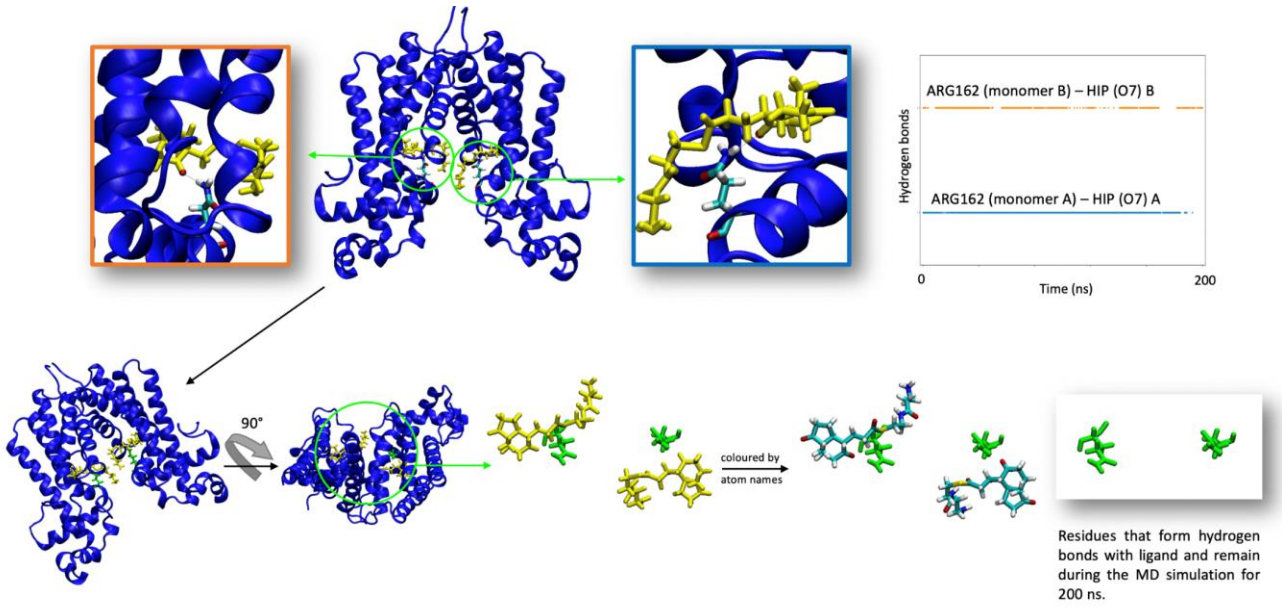


Figure 4. 25. Hydrogen bonds between KstR2 and HIP occupied for 20-100% of the Mtb-KstR2-2HIP simulation. Top panel: front view of KstR2 with insets showing the details of the interaction between the residues that form a hydrogen bond with HIP-CoA. The protein backbone is drawn in blue cartoon format, the ligands are drawn in yellow liquorice format, and the residues with which they form hydrogen bonds are drawn in liquorice format and coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen; gold: phosphorous). The colours of the outlines of these insets match the colours of their respective lines in the time-series (right-hand side). Bottom panel: Structures of the HIP-CoA molecule and the residues with which it forms hydrogen bonds. In the right-most structure, the ligand is coloured according to atom type and the residues that hydrogen bond to it are in green. The screenshots of the protein are from the endpoint of the 200 ns MD simulation.

No hydrogen bonds occupied for more than 20% of the Mtb-KstR2-2HIP-CoA simulation were formed between either monomer of KstR2 and the ligand bound to monomer B. To explain this, the initial and final coordinates of both ligands were visualized (Figure 4. 21, lower panel). After 200 ns MD simulation, the CoA part of the ligand bound to monomer A was folded, while the other HIP-CoA in monomer B remained extended and more similar to the initial coordinates. It could be interesting, therefore, to extend this simulation to examine whether the ligand in monomer A and/or B exits the binding pocket, and if this is related to its conformations and motion during the simulation.

To check if the other two hydrogen bonds observed in the crystal structure (128) (ASN128 and ASP163, Figure 4. 23) form at any point during the Mtb-KstR2-2HIP-CoA simulation, the hydrogen bonds were recalculated with a minimum occupancy of 5% of the simulation (Figure 4. 26). All the hydrogen bonds that form in the crystal structure now show up during the simulation (marked with grey arrows in Figure 4. 26). However, they form hydrogen bonds with different atoms of the ligand than in the crystal structure (Figure 4. 22Figure 4. 23). Additionally, many are present only fleetingly.

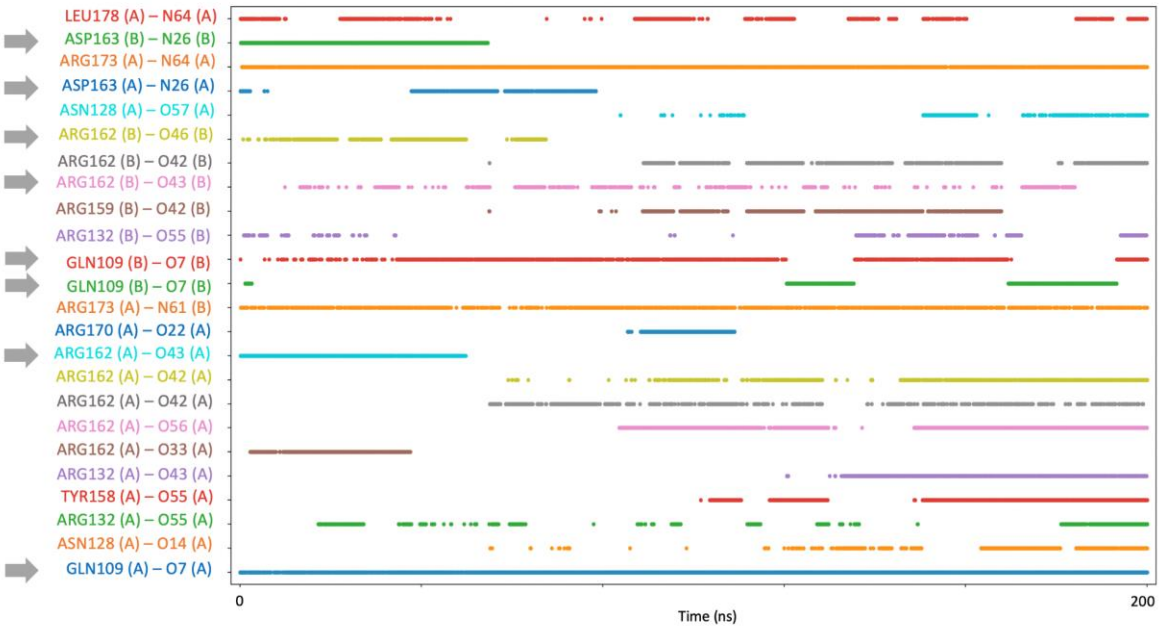


Figure 4. 26. Hydrogen bonds between KstR2 and ligand occupied for 5-100% of the Mtb-KstR2-2HIP-CoA simulation. Those marked with a grey arrow are also observed in the 4W97 crystal structure (137, 158).

Mtb-KstR2-1HIP-CoA only has a HIP-CoA molecule in monomer A. At a minimum of 20% occupancy, only two hydrogen bonds are formed, between ARG162 of monomer A (nitrogen in NH_2 group of the side chain) and the CoA part of the ligand (oxygen, O43), which remains formed for most of the simulation, and between ARG132 of monomer A (nitrogen in NH_2 group of the side chain) and the CoA part of the ligand (oxygen, O42), which only exists during the first half of the simulation. Two of the interactions of ARG162 with HIP-CoA observed in the crystal structure are therefore missing, or are formed very infrequently, along with the other hydrogen bonds involving ASN128, ASP163 and GLN109. The HIP-CoA twisted during the simulation (Figure 4. 27), similar to what was observed for the HIP-CoA bound to monomer A in the Mtb-KstR2-2HIP-CoA simulation, however to a lesser extent. When folding happens, the CoA moiety of the ligand twists back closer to the pocket, however mostly phosphate group approaching to the nitrogen, both in the ligand, may cause intramolecular interactions.

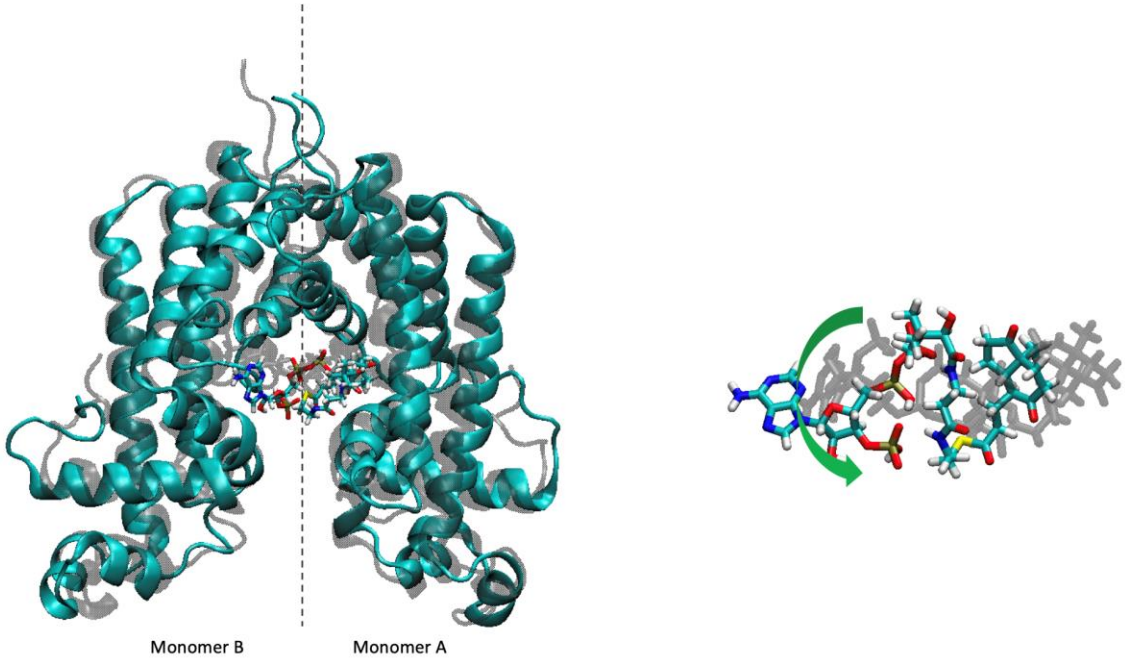


Figure 4. 27. The structure of (left) Mtb-KstR2-1HIP-CoA with one HIP-CoA ligand in ligand pocket of monomer A and (right) magnified view of the HIP-CoA. The green curly arrow points the twist. In the cartoon representation of KstR2, the grey shadow represents the conformation before starting the MD simulation and the cyan structure represents the conformation after 200 ns MD simulation. The HIP-CoA ligand is displayed in liquorice format for both conformations, as a grey shadow for the conformation before the MD simulation and coloured according to atom type (cyan: carbon; red: oxygen; white: hydrogen; blue: nitrogen; gold: phosphorous) for the conformation after the 200 ns MD simulation.

Analysis of the protein-ligand interactions in the Mtb-KstR2-2HIP simulation, in which only the HIP portion of the ligand is present provides an opportunity to examine the role of the CoA moiety in the KstR2-ligand interaction, which has been unclear (135).

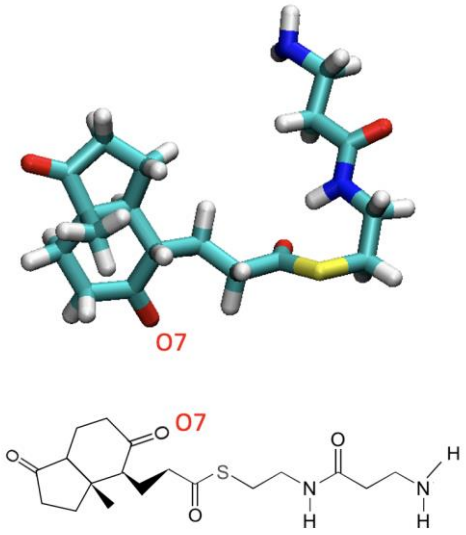


Figure 4. 28. (Top) Liquorice and (bottom) 2D representations of the HIP ligand. The oxygen atoms involved in hydrogen bonds are labelled. The liquorice structure is coloured according to atom type (cyan: carbon; red: oxygen; blue: nitrogen; white: hydrogen).

Since most of the hydrogen bonds formed in the two previous simulations were between KstR2 and the CoA parts of the ligands, removing the CoA led to fewer hydrogen bonds, as expected. Indeed, this may be one reason for the lower affinity of HIP to KstR2 compared to HIP-CoA. Only the ARG162 residues in both monomers of KstR2 (nitrogen in NH_2 group of the side chain) formed hydrogen bonds with the HIP parts left in the ligands (oxygen, O7) (Figure 4. 25Figure 4. 28). According to the experimental data (137), in HIP-CoA bound Mtb-KstR2, mutation of ARG162 noticeably decreases the affinity of KstR2 for HIP-CoA, thus leads to the loss of hydrogen bonds, which ARG162 forms with the ligand (137). ARG162 should interact with the diphosphate group of HIP-CoA, however, not the O7 oxygen atom, but the disphosphate group is no longer present in HIP. The smaller HIP ligand has therefore moved in the binding pocket to form this different hydrogen bond.

Many of the interactions between KstR2 and HIP visible in the crystal structure are not hydrogen bonds. For instance, there are also stacking interactions between the aromatic side chains of TYR166 (α8) and TYR108 (α6) and the cycloalkanone rings of the HIP moiety. Besides, the HIP part has several hydrophobic interactions with side chains of the residues around the pocket, such as PHE65, LEU66, LEU69, PHE70, TYR73, and VAL105. In order to determine whether these residues make hydrophobic interactions during the MD simulation, the minimum distance between the COM each residue and the COM of the ligand was calculated for the Mtb-KstR2-HIP simulation (Figure 4. 29).

PHE65, LEU66, LEU69 and VAL105 remained at distances of ~0.25 nm throughout the simulation, but PH70 and TYR73 did not stay in close contact with the HIP (Figure 4. 29). For TYR73, the distance to the HIP fluctuates between 0.25 and 0.6 nm, and the hydrophobic interactions reappear occasionally during the simulation. For PHE70, however, the distance is greater than 0.6 nm for the majority of the simulation. In monomer A, this interaction is lost after ~25 ns, whereas in monomer B, it is lost almost immediately but does reoccur twice either side of 100 ns of simulation.

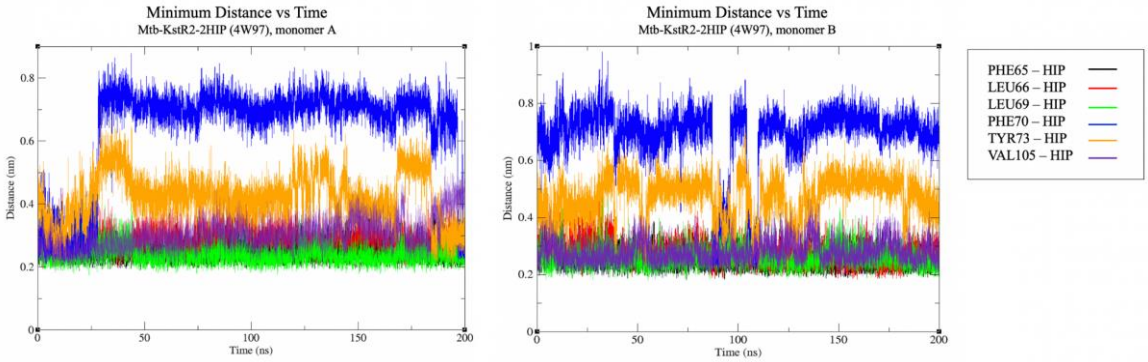


Figure 4. 29. Minimum distances between the COM of HIP and the COM of PHE65, LEU66, LEU69, PHE70, TYR73, and VAL105 in (left) monomer A, and (right) monomer B, during the MD simulation of Mtb-KstR2-2HIP.

4.4. Conclusions

This chapter investigated two key questions regarding the mechanism of action of KstR2: first, the effect of mutating residue ARG170 to ALA, and second, the effect of ligand binding on its structure and motion. To investigate the first question, two apo-state crystal structures of KstR2, 2IBD (Rjo KstR2) and 4W1U (Mtb KstR2), were simulated along with R170A mutations of each.

With regard to the role of ARG170, mutation of this residue to ALA resulted in a more rigid structure, with less fluctuations based on changes in RMSD values. According to the inter-DBD distances, in both Rjo KstR2 and Mtb KstR2, the R170A mutation removed the jump in fluctuation of the distances which were seen in the wild-type forms, resulting in more stable and less flexible positions of the DBDs. The RMSF analyses also supported this. However, the fluctuation patterns in the two monomers were asymmetric in both Rjo-KstR2-R170A and Mtb KstR2-R170A. The extreme projection onto the first principal component showed a notable reduction in the largest motion of the protein. However, the type of motion remained breathing-like for both Rjo-KstR2-R170A and Mtb KstR2-R170A. R170A causes less correlation and also alters the symmetry of the correlation patterns for both Rjo and Mtb

KstR2, according to the dynamical network analysis heatmaps, which are in good agreement with the other results, in terms of the rigidity of the mutated structures. However, the DBD and LBD still behaved as separate rigid bodies, with an anti-correlated motion, with the R170A mutation reducing the degree of anti-correlation. According to the results of the network community analyses, the LBD community in monomer B was extended to the top-right part of the LBD of monomer A. This shows a tight binding of the two monomers, and could be hypothesised to cause a scissor-like motion. This was a consistent result in all simulated states. The communication pathways between ARG170/ALA170 and TYR48 became longer in the Rjo-KstR2-R170A and Mtb-KstR2-R170A, and even extended to involve helix α8. This suggests that this mutation causes a remarkable change in the allosteric communications between the LBD and DBD. Overall, these results and the sometimes dramatic effect of the R170A mutation support the hypothesised scissor-like mechanism of action and in particular, the key switching role of ARG170.

The difference between the behavior of the two KstR2 structures was particularly clear in the inter-DBD distances. Both Rjo-KstR2 and Mtb-KstR2 exhibited large jumps in their inter- DBD distances, which matched the RMSD time-series. However, the jump was in the opposite direction in each case. For Rjo-KstR2, the two DBDs were close in the beginning (~3.75 nm), then moved further apart to ~5nm during the simulation, whereas the opposite occurred for Mtb-KstR2. It would be extremely interesting to extend these two simulations to see if these fluctuations repeat. If this occurs, it could be evidence in favour of the hypothesis that KstR2 has a conformational selection mechanism for binding to DNA, assuming that the KstR2 DNA-bound structure has a short inter-DBD distance, equivalent to other KstR-like proteins (137). However, full investigation of the DNA-binding mechanism also requires simulation of a DNA-bound structure of KstR2, for which a crystal structure of that state is required.

To investigate the second question, an apo-like state was created by removing both ligands from the 4W97 crystal structure of KstR2 bound to two molecules of HIP-CoA. An in-between state was created by removing the HIP-CoA only from monomer A, and a fourth state was created by removing the CoA moiety from both ligands. The CoA of the steroid ligands increases the ligand affinity for KstR2 (*Pers. Commun.* Stephanie Dawes, The University of Auckland). Since the role of the CoA moiety in the KstR-ligand interaction has been an unresolved issue (135), and as the ligand in KstR2 has the CoA moiety, one aim was to investigate whether and how the CoA moiety interacts with the protein or not. The simulations

were analysed in terms of their conformational motion, correlated motions, communication pathways and, for the ligand-bound states, the protein-ligand interactions.

Interestingly, states of KstR2 that are symmetric with respect to ligand binding exhibit asymmetric dynamics and communication pathways and *vice versa*.

Comparison of the Mtb-KstR2-2HIP-CoA simulation with the Mtb-KstR2-2Hip-CoA-removed simulation revealed that ligand binding pushes the two DBDs further apart. Furthermore, the extreme projection onto the first principal component showed a remarkable change in the type of the largest motion of the protein from a breathing motion to a more twist-like motion. Removing one ligand (Mtb-KstR2-1HIP-CoA) resulted in the type of the largest motion remaining twist-like. Overall, the motion of Mtb-KstR2-1HIP-CoA was symmetric, whereas that of Mtb-KstR2-2HIP-CoA showed asymmetric dynamics. Without the CoA moiety present (Mtb-KstR2-2HIP), the conformational changes of residues 166-179 (in the lower LBD) were asymmetric like the dynamics of Mtb-KstR2-2Hip-CoA-removed, suggesting that the CoA portion in some way stimulates the dynamic asymmetry in response to ligand binding.

In terms of protein-ligand interactions, the interactions with HIP-CoA were asymmetric for Mtb-KstR2-2HIP-CoA, but not for Mtb-KstR2-2HIP (at 20% minimum occupancy). The HIP-CoA bound to monomer A folded and twisted during both the Mtb-KstR2-2HIP-CoA and Mtb-KstR2-1HIP-CoA simulation, but the HIP moiety remained inside the pocket, and more, rather than fewer, hydrogen bonds were formed with the protein compared to monomer B of Mtb-KstR2-2HIP-CoA.

In general, few of the hydrogen bonds visible in the crystal structure (137) occurred with high occupancy in any of the simulations, but they were all observed at low (<5%) occupancy in the Mtb-KstR2-2HIP-CoA simulation. In the 2HIP-bound form, ARG162 forms hydrogen bonds with the 5-carbonyl oxygen (O7) atom of the HIP in both monomers, rather than with the diphosphate group of the CoA that is no longer present. For Mtb-KstR2-HIP, hydrophobic interactions were also computed as the HIP part of the ligand is mostly hydrophobic. Most of these interactions that are observed in the crystal structure were also present during the simulation, but some were lost, suggesting that without the CoA portion the ligand is less stably bound to KstR2.

The persistence of the hydrogen bond with ARG162 even when the CoA moiety was removed suggests that one future extension of this work could be mutation of this ARG to find if the ligand still binds to the protein or if it changes the conformation of LBD and the ligand

pocket. Further options include mutation of each of the charged residues in the pocket in turn to study and validate the role of CoA part of the ligand in binding to KstR2.

Chapter 5. Effect of RAS on PI3Kα-membrane interaction

5.1. Overview:

Phosphatidylinositide 3-kinases (PI3Ks) are proteins that take part in signalling pathways regulating factors like cell growth, survival and proliferation (230), which in turn are involved in cancer. The phosphorylation of the OH- group at the 3-position of the inositol ring of phosphatidylinositols (PIs) triggers the signalling pathways, by converting PI(4,5)P2 to PI(3,4,5)P3. This phosphorylation of the 3' position of the inositol ring is the reason behind naming the protein PI3K (47). When genes which encode the components of the pathway are mutated, amplified, or deleted, this can make PI3K activated in human cancers. Furthermore, mutations to proteins that activate PI3K such as RAS and activation of receptor tyrosine kinases can also cause the activation of PI3K (231).

RAS is a family of related proteins which belongs to a class of proteins called small GTPases, and participates in transmitting signals in cells. They can also be considered as binary molecular switches which control intracellular signalling pathways. When incoming signals switch RAS 'on', it in turn switches on other proteins, which turn on genes engaged in cell growth, differentiation and survival. Mutations in RAS genes can cause RAS proteins to remain continuously activated, which results in inappropriate and overactive signalling inside the cell, even without any incoming signals. Since these signals cause cell growth and division, overactive RAS signalling could, and in many cases does, ultimately result in cancer (232).

PI3Ks are one of the most important effectors of RAS signalling, interacting directly with RAS with no other proteins involved (233). The interaction between Class IA PI3Ks such as PI3Kα, the focus on this chapter, and RAS is very important in the formation, growth and maintenance of RAS-driven tumours (234).

Lipid phosphoinositides such as PIP_2, which are key mediators in many membrane signalling pathways (235), are located in the cell membrane. PI3K enzymes act on PIP_2, and require RAS for activation, and thus the interaction between RAS and PI3Kα also occurs on the lipid bilayer surface. Specific post-translational modifications are required for many biological functions of RAS. These modifications, which guide the localization and interaction of RAS with effectors such as PI3Kα are prenylation of the C-terminal region by farnesyl or geranylgeranyl isoprenoid lipids. These lipid anchors are embedded into the cell membrane and therefore promote membrane binding (102). Due to it occurring on the cell membrane surface, studying this interaction between PI3Kα and RAS at atomic level is experimentally challenging (234), but can be tackled computationally.

In this work, a model for the PI3Kα-RAS-membrane system was built, and the interactions between the proteins and of the protein complex with the lipid bilayer were investigated using molecular dynamics (MD) simulations.

5.2. Methods

5.2.1. Parameters

GROMOS54A7 force field parameters were used for PI3Kα, and for the protein component of RAS. The solvent for all simulations was SPC water (203).

To parameterize the RAS lipid anchors, the CYS residues 181, 184, and 186 to which two palmitoyl and one isoprenyl lipids are attached, respectively (Figure 5. 1), were parameterized according to MET in GROMOS54A7 (173) (see Section 5.2.2). For isoprenylation, the farnesyl group parameters were taken from ATB (194–196) (molID: 10209), which supplies GROMOS54A7-compatible parameters. ATB typically over-estimates partial charges, so following the standard GROMOS54A7 parameterisation scheme for hydrocarbon chains, all partial charges were set to zero. For the O-methylation of the farnesylated amino acid, GROMOS54A7-compatible parameters again were obtained from ATB (methyl-2-methylpropanoate molID: 26886). Parameters for palmitic acid were taken from the GROMOS54A7 DPPC parameter set (236). Parameters for the linker bond between the CYS residues and the fatty acid were also obtained from ATB (molID: 351504), GROMOS54A7 (Figure 5. 1).

GROMOS54A7-compatible parameters for the phosphoaminophosphonic acid guanylate ester ($C_{10}H_{17}N_6O_{13}P_3$, GNP) ligand bound to RAS were taken from ATB (molID: 337232, common name: 5'-Guanylylimidodiphosphate).

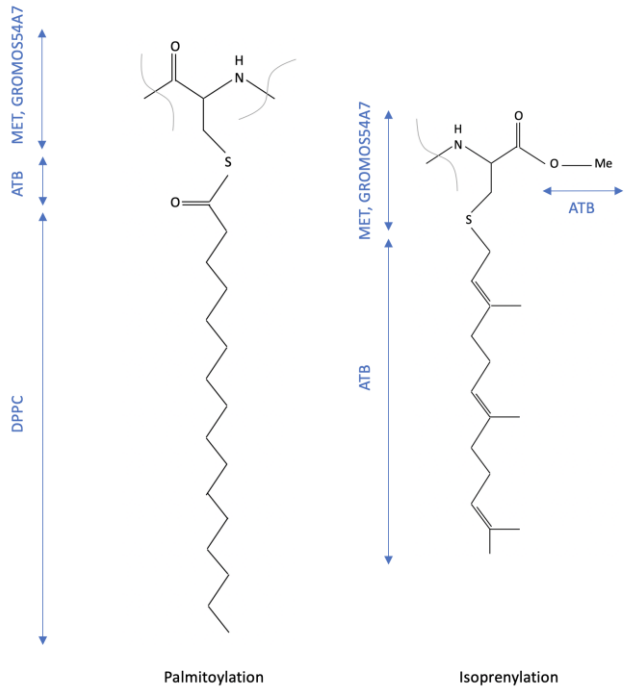


Figure 5. 1. Palmitoylation and isoprenylation of RAS. Blue arrows indicate the origin of each fragment's parameters.

A mixed lipid bilayer was previously built (82, 237) to mimic the lipid bilayers used in hydrogen-deuterium exchange experiments (238) , and subsequently updated to better match the lipids used in hydrogen-deuterium exchange mass spectrometry studies of the interaction of PI3K and RAS with lipid bilayers (76). 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphatidylcholine (16:0, 18:1; POPC) was retained, but 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphatidylethanolamine (POPE) and 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphatidylserine (POPS) were replaced by 1-stearoyl-2-oleoyl-sn-glycero-3-phosphatidylethanolamine (18:0, 18:1; SOPE) and 1-stearoyl-2-oleoyl-sn-glycero-3-phosphatidylserine (SOPS) (45) (Figure 5. 2), in a manner compatible with the GROMOS54A7 force field. Additionally, the non-sphingosine acyl chain of the sphingomyelin was changed from oleate (18:0) to nervonate (24:1), and the acyl chains of PIP_2 were changed from palmitate-oleate (16:0, 18:1) to stearate-arachidonate (24:4, 18:0). Here, the sphingomyelin parameters were slightly modified again to replace the non-standard bond and angle types surrounding the sphingosine double bond with standard GROMOS54A7 parameters (239) (as already used in the nervonate), and to add a previously-missing improper dihedral term across the trans double bond formed by atoms 16-19-20-21 of sphingosine (Figure 5. 3). Lastly, the atom type, bond, and bond angle parameters for cholesterol used previously (45, 82, 237) were changed here to standard GROMOS54A7

terms as used in the ATB cholesterol parameters (molID: 4202) and the unnecessary improper term on the cyclopentane ring was removed.

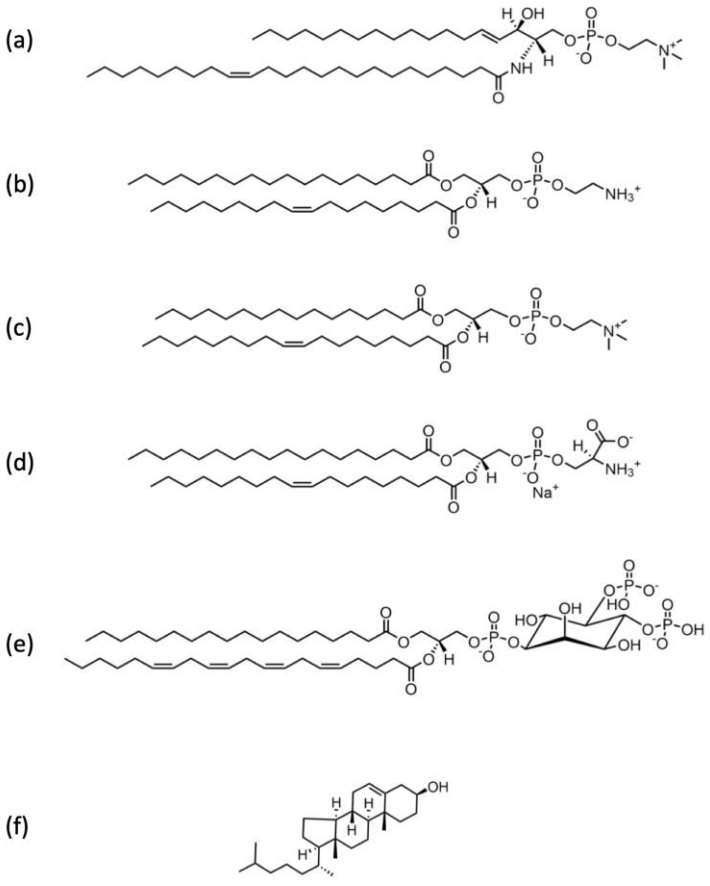


Figure 5. 2. Chemical structures of the lipid components in the simulated membrane bilayers. (a) 18:1-24:1 sphingomyelin; (b) 18:0-18:1 phosphatidylethanolamine (PE); (c) 16:0-18:1 phosphatidylcholine (PC); (d) 18:0-18:1 phosphatidylserine (PS); (e) 18:0-20:4 PI(4,5)P2; and (f) cholesterol [edited from (45)].

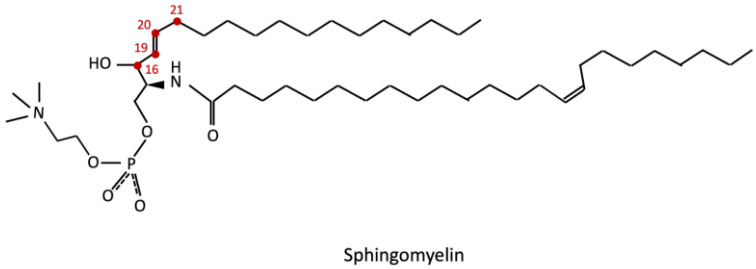


Figure 5. 3. Structure of sphingomyelin, in which the four red-coloured atoms form a trans double bond.

5.2.2. Coordinates

Initial coordinates of PI3Kα were taken from a crystal structure of the wild-type PI3Kα heterodimer, p110α in complex with the niSH2 (nSH2 and iSH2) subdomain of p85α and an analogue of the p110α lipid substrate, diC4-PIP_2 (PDB ID: 4OVV, refined to 3.4 Å resolution). The non-protein coordinates were removed from the structure, and coordinates for missing residues were predicted using Modeller 9.19 (240). For modelling the α3 helix of the p85α iSH2 domain, this process is described in more detail in Section 5.3.1.2, and for building the α-helix 'triangle' in the RBD the details are provided in Section 5.3.1.4.

For RAS, the coordinates were obtained from a crystal structure of the H-RAS G12V – PI3Kγ complex (PDB: 1HE8), in which RAS is bound to GNP and a Mg^{2+} ion. The mutation G12V has been identified as a common oncogenic mutation to RAS (241), which is not in the binding site of RAS and PI3Kα, and does not abolish binding of the two proteins (242). Coordinates for the residues in the C-terminal hypervariable linker domain (107) of RAS, which were not resolved in the crystal structure, were modelled using Modeller 9.19 (240). GNP is a non-hydrolysable analogue of GTP, in which the oxygen atom bridging the β and γ phosphates is replaced by a nitrogen atom. GNP binds tightly to RAS in the presence of a Mg^{2+} ion, locking RAS into in its active form (67). Both GNP and the Mg^{2+} ion were therefore included in the simulations. Hydrogen atoms were added to the ligand using GaussView 5.0 (198) to make the structure ready for parameterisation using ATB (see Section 5.2.1). The ATB-optimised united atom coordinates were used for the simulation.

Most RAS-related proteins are post-translationally altered through covalent binding of lipids. Such lipid modifications are vital and required for membrane association of these proteins, and also for their biological activity (243, 244). The RAS subfamily is prenylated, with farnesyl added to C-terminal CAAX motifs (245). Palmitoylation is another required lipidation of RAS proteins. Initial coordinates of palmitic acid and farnesyl groups were modelled using GaussView 5.0. The 23 residues of the RAS hypervariable linker domain, LYS167 to SER189 (KLRKLNPPDESGPGCMSCKCVLS), to which lipids are attached, are missing from the PI3Kγ-RAS complex crystal structure. Coordinates for these were built onto the C-terminus of RAS using Modeller. CYS residues 181, 184 and 186 were modified to add two palmitic acid and one farnesyl groups, respectively (Figure 5. 4). Subsequently, the C-terminal CAAX motif, which is CVLS here, was modified to C-OMe.

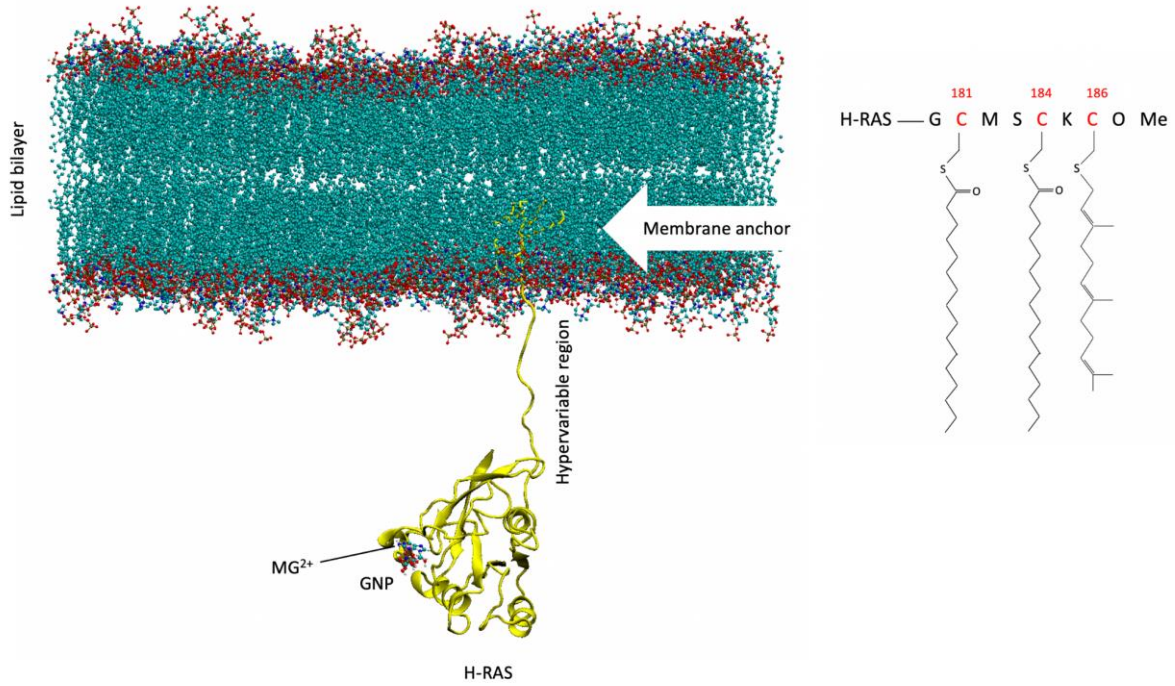


Figure 5. 4. (Left) Final coordinates of lipidated RAS with the lipids embedded in the mixed lipid bilayer, and (right) the lipidation pattern of the seven C-terminal residues of RAS. RAS is drawn in yellow, and the GNP ligand and the lipids are drawn in CPK representation with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen).

The initial coordinates for all lipids [9 SOPE (45%), 3 POPC (15%), 4 SOPS (20%), 2 cholesterol (10%), 1 sphingomyelin (5%), and 1 SA-PI[4,5]P_2 (5%)] were constructed by Emma McCafferty (45) in the form of an ordered 5 by 4 leaflet of 20 randomly distributed lipid molecules using the CHARMM GUI membrane builder (246). This membrane fragment was then translated 5 times in the *x* and 4 times in the *y* direction, and reflected in the *z* plane to generate the other leaflet of the bilayer. This resulted a complete lipid bilayer consisting of 800 lipids (360 SOPE, 120 POPC, 160 SOPS, 80 cholesterol, 40 sphingomyelin, and 40 SA-PI[4,5]P_2) (Figure 5. 2).

For embedding RAS lipid anchors into the lipid bilayer, two lipid molecules, one SOPE and one CHOL, were removed from the bilayer and three lipid tails of RAS were placed in the resulting gap.

5.2.3. Computational methods

5.2.3.1. Simulation methods:

All MD simulations in this work were prepared and performed with the GROMACS simulation package version 2016.3 (199–201), and visualised using the VMD molecular visualization

program (197). The solvent for all simulations was SPC water. For all membrane simulations, after solvating the system, water molecules between the two monolayers were removed.

Restraints

Restraints were used to control the position of RAS, in order to prevent the hypervariable region of RAS folding up (Section 5.3.1.10) and to encourage formation of interactions between PI3Kα and RAS (Section 5.3.2).

Position restraints

Position restraints were applied to the Cα atoms of RAS. The reference coordinates were its initial position, with the hypervariable domain extended, and the force constant was 1000 $kJ \cdot mol^{-1} \cdot nm^{-2}$ in each of the *x*, *y* and *z* dimensions.

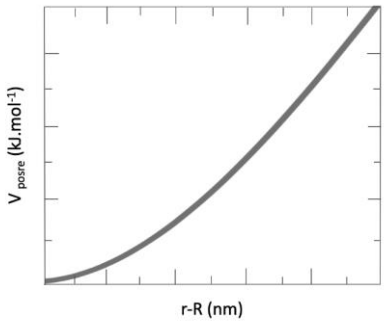


Figure 5. 5. Position restraint potential.

Distance restraints

Distance restraints were applied to encourage the formation of interactions between the Cα atoms of residues in p110α and RAS. The pairs of residues and distances between them are listed in Figures 5.12 and 5.13 and the force constant was 120 $kJ \cdot mol^{-1} \cdot nm^{-2}$.

Steered molecular dynamics

In order to extend the hypervariable domain (LYS167 (atom 1701) to PRO179 (atom 1825)) of RAS, a time-dependent external pulling force was applied to that part of the system using steered molecular dynamics (SMD) (247). By using this method, conformational changes are induced on time scales accessible to MD simulations (247–249). Residues in the globular portion of RAS (group 1, atoms 1-1700) were position restrained and the lipid anchors (group 2, atoms 1826-1932) were pulled away along the reaction coordinate (the *z*-axis) using a harmonic potential, which allows the force to vary based on the nature of the interactions of

group 1 with group 2 (249) (Figure 5. 6). It is important not applying very fast pulling rates or too strong force constants, which may deform the system (249). The distance between the COM of each group was increased at a rate of 0.01 $nm \cdot ps^{-1}$ (10 $nm \cdot ns^{-1}$) using a force constant of 1000 $kJ \cdot mol^{-1} \cdot nm^{-2}$. More information about this simulation can be found in Section 5.3.1.6.

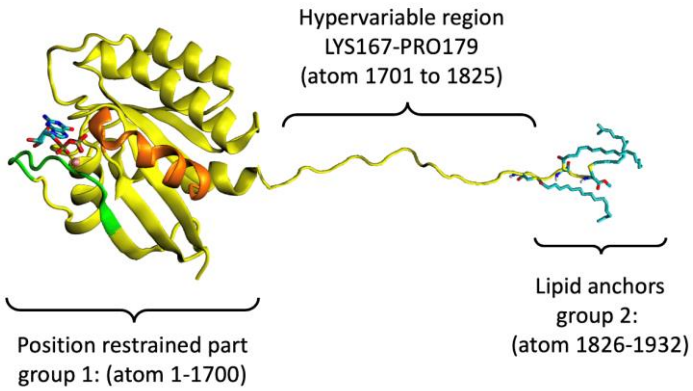


Figure 5. 6. Cartoon structure of GNP-RAS (with Mg^{2+}), with the posttranslational modifications to add lipid anchors to the hypervariable linker domain. The two sets of atoms used in the SMD, group 1 and group2, are indicated along with their atom numbers. The two groups are subject to a biasing potential, with group 1 being position restrained, and group 2 gently pulled away. RAS is coloured yellow other than the green and orange colour regions, which represent the switch I and switch II, respectively. GNP and the lipid anchors are drawn in liquorice format and coloured according to atom type (cyan: carbon; red: oxygen; blue: nitrogen).

Simulations without membrane

All simulations that did not include a membrane were initiated and run with the following equilibrium scheme. After two energy minimisations, steepest descent following by conjugate gradient, the initial velocities were generated from a Maxwell-Boltzmann distribution at 50 K in the NVT ensemble. The system was heated for 250 ps in NVT ensemble, with temperature controlled using the modified Berendsen thermostat, V-rescale, with a temperature coupling constant (τ_T) of 0.1 ps (204), and a reference temperature that increased linearly from 50 K to 298 K at a rate of 0.992 K/ps, and periodic boundary conditions. The lengths of all bonds involving hydrogen atoms were constrained using the LINCS algorithm (205) with an order of 4, allowing for an integration time step of 2 fs. The Verlet cut-off scheme (206) was used to calculate non-bonded interactions, with both Lennard-Jones and electrostatic interactions calculated explicitly within a cut-off distance of 1.4 nm. Outside of the cut-off distance, electrostatic interactions were calculated using reaction field. The system was subsequently equilibrated for 250 ns at 298 K in the NpT ensemble. Pressure was controlled using the Berendsen barostat, with a pressure coupling constant (τ_p) of 0.5 ps, a reference pressure of 1 bar, and an isothermal compressibility of 4.5 x 10^{-5} bar^{-1}. All the other procedures and values

are the same as for the NVT ensemble simulations. The system was then simulated for a 50 ns production MD run in the NpT ensemble.

Simulations including membrane

Simulations including a membrane were initiated and run with the following equilibrium scheme. The constructed lipid bilayer systems were simulated for 200 ns. The MD simulations were initiated through an equilibration scheme where the initial velocities were randomly generated from a Maxwell-Boltzmann distribution at 50 K, and then the system was heated to 300 K over the course of 500 ps equilibration in the NpT ensemble. The temperature and pressure were controlled using the Berendsen thermostat and barostat (168), respectively, in a semi-isotropic environment. The temperature coupling (τ_T) and pressure coupling (τ_p) constants were 0.1 and 1.0, respectively, under semi-isotropic periodic boundary conditions and with an isothermal compressibility of 4.5×10^{-5} bar^{-1}. The LINCS algorithm (205) was used to constrain all bond lengths with an order of 4, allowing for an integration time step of 2 fs. The Verlet cut-off scheme (206) was used to calculate non-bonded interactions, with both Lennard-Jones and electrostatic interactions calculated explicitly within a cut-off distance of 1.4 nm. Outside of the cut-off distance, electrostatic interactions were calculated using reaction field. The final equilibrated coordinates were used as the starting configuration for 50 ns MD simulations at 300 K, using NpT ensemble. In this production simulation, electrostatic interactions were calculated using particle mesh Ewald (PME) summation (207), with PME order of 4. All other procedures and values were same as the heating and equilibration simulation.

5.2.3.2. Analysis methods

For any RMSD (Root Mean Square Deviation), RMSF (Root Mean Square Fluctuation), inter-residue distance and hydrogen bond analysis done in this chapter, the analysis methods are the same as those described in Chapter 3. All superimpositions of the structures were performed using Chimera 1.13 (227). Visualization of the structures and trajectories was done using VMD 1.9.4 (197).

5.3. Results and discussion:

The goal of this research is to study the interaction and binding of a PI3Kα-RAS-membrane system. As there was no crystal structure of such a system, it needed to be assembled step by

step. The crystal structure of PI3Kγ-RAS (PDB ID: 1HE8) was used as a template to model PI3Kα-RAS, as this is the only available crystal structure of RAS bound to a PI3K protein. The RAS isoform in this crystal structure is HRAS, the isoform used in most structural and biochemical studies on RAS (242). The details of the whole process of modelling and all steps are following.

5.3.1. Building a model of the system (PI3Kα-RAS-membrane)

5.3.1.1. Preparation of PI3Kα

After personal communication with our experimental collaborator (Dr Jack Flanagan, The University of Auckland), the PI3Kα crystal structure PDB:4OVV was selected, because it has a complete RAS binding domain and activation loop. This structure includes coordinates for the p110α subunit and for the nSH2 and iSH2 domains of the regulatory p85α subunit. The HIS-tag (the first 28 residues) was removed, and all missing residues were built using Modeller (240). After each modification, the protein was equilibrated for 50 ns in SPC water (Section 5.2.3.1) to confirm that the modification had not adversely affected the behaviour of the protein.

5.3.1.2. Addition of α3 helix to p85α regulatory subunit

Because the goal is to simulate the interaction of PI3Kα with a model cell membrane, which is the prelude to phosphorylation of PIP_2 by PI3Kα to form PIP_3, an active-like state is desirable. There are multiple elements of inhibition of PI3Kα activity (250).

In the p85α subunit, binding of the two SH2 domains (N and C) to the p110α subunit restricts its lipid kinase activity (251, 252). The nSH2 domain makes an inhibitory contact with the p110α subunit (253). This is released when the two SH2 domains (N and C) interact with phosphorylated tyrosines in TYR-X-X-MET motifs in peptides. This causes dissociation of the SH2 domains away from the main PI3Kα structure, releasing the inhibitory contacts with the p110α subunit and thus allowing access to the kinase domain (70)(254) (72). The cSH2 domain contacts the RBD of the p110α subunit, preventing stimulation by RAS (255). These contacts are also released when the cSH2 domain binds to phosphorylated tyrosine motifs.

Therefore, removing nSH2 and cSH2 domains and their linkers relieves their inhibition and so can be thought of as forming an “active” state of PI3Kα. This leaves behind the iSH2 domain only, which has no effect on the PI3K activity and inhibition, although it closely interacts with the p110α subunit (253). Accordingly, the first 104 residues of p85α, which form

the nSH2 domain and its linker, were removed from the 4OVV structure (MET323 to LEU426), so that the protein structure starts from TYR427. It was not necessary to remove the cSH2 domain as it was not present in the crystal structure.

The 4OVV structure only contains coordinates for the antiparallel coiled coil of the iSH2 subunit, which consists of two helices, α1 and α2. However, the iSH2 domain also contains a short α3 helix (256)(252) as shown in Figure 5. 7. Coordinates for the α3 helix were built with Modeller (240) using the 2V1Y crystal structure, which does include this helix, as a template (12 residues, 589-600, in 2V1Y). The 2V1Y structure comprises a PI3Kα adaptor-binding domain (ABD) in a complex with the iSH2 domain from p85α (256).

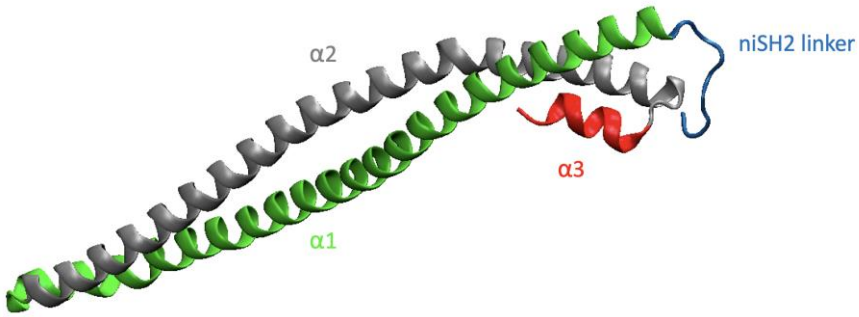


Figure 5. 7. Structure of the iSH2 domain of the p85α regulatory subunit, from the crystal structure PDB ID: 2V1Y. The α1 helix is green, the α2 helix grey, the α3 helix red, and the niSH2 linker is blue.

5.3.1.3. Preparation of PI3Kα-RAS complex

The initial goal was to obtain a reasonable structure of the PI3Kα-RAS complex before incorporating lipidation of RAS. Coordinates for RAS only (no ligand or lipids) were obtained from the crystal structure of the PI3Kγ-G12V RAS complex (PDB: 1HE8) and equilibrated for 50 ns. A putative PI3Kα-RAS complex was built by superimposing the coordinates of equilibrated PI3Kα (Section 5.3.1.2) and RAS onto those of PI3Kγ in the PI3Kγ-RAS complex (Figure 5. 8). The complex was then minimized in two steps, heated for 250 ps (NVT), followed by 250 ps NpT equilibration. Finally, a MD production was run for 50 ns. Each subsequent simulation in this chapter was run in the same way, unless otherwise stated. The simulation method is as described in Section 5.2.3.1.

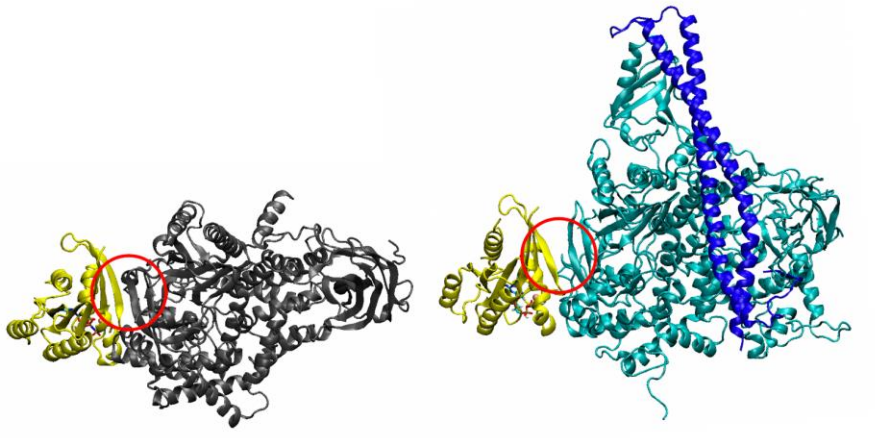


Figure 5. 8. Comparison of the structures of the complexes PI3Kγ-RAS (left), PDB ID: 1HE8, and PI3Kα-RAS (right), built from the PI3Kα coordinates from PDB ID: 4OVV and the RAS coordinates from 1HE8. PI3Kγ is in black, the p110α subunit is in cyan, p85α (α1 and α2 helices of iSH2 domain only) in blue, RAS in yellow, and the GNP ligand is drawn in liquorice representation with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen). The binding zones of RAS and PI3K are shown by red circles.

5.3.1.4. Remodelling of RBD α-helical 'triangle'

In the PI3Kγ-RAS complex, a β-strand in the RAS binding domain (RBD) of PI3Kγ is partially aligned with a β-strand in RAS such that the two strands are hydrogen-bonded and the β-sheets in which they reside become one continuous β-sheet. Visualisation of the MD simulation trajectory of the PI3Kα-RAS complex revealed that the two β-strands of RAS and the PI3Kα RBD did not remain aligned. A sequence of residues which form a triangle-like shape in the RBD, ASP214 to LEU252 (DCVPEQVIAEAIRKKTRSMLLSSEQLKLCVLEYQGKYIL), moved towards the two β-strands, and appeared to bring about their separation. In PI3Kγ, there is an equivalent triangle, in this case comprising three α-helices, in the RBD, formed by residues ASP238 to ARG273 (DDTPGAILQSFFTKMAKKKSLMDIPESQSEQDFVLR). Using the PI3Kγ triangle structure as a template, an equivalent triangle of α-helices was re-built for PI3Kα using the coordinates of PI3Kγ as a template with Modeller 9.19 (Figure 5. 9Figure 5. 10). The structure (binding to RAS) then was equilibrated in water for 50 ns.

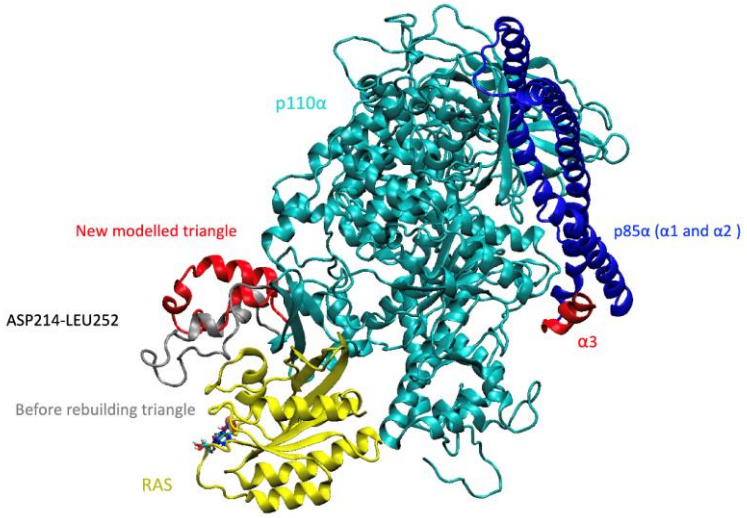


Figure 5. 9. Modelled helical portions of the PI3Kα-RAS complex. The p110α subunit is in cyan, p85α (α1 and α2 helices) in blue, RAS in yellow, and the newly-built helices (α3 and the α-helix triangle) are in red, and the triangle section as per the PI3Kα crystal structure (4OVV) is in grey. The GNP ligand is drawn in liquorice representation with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen).

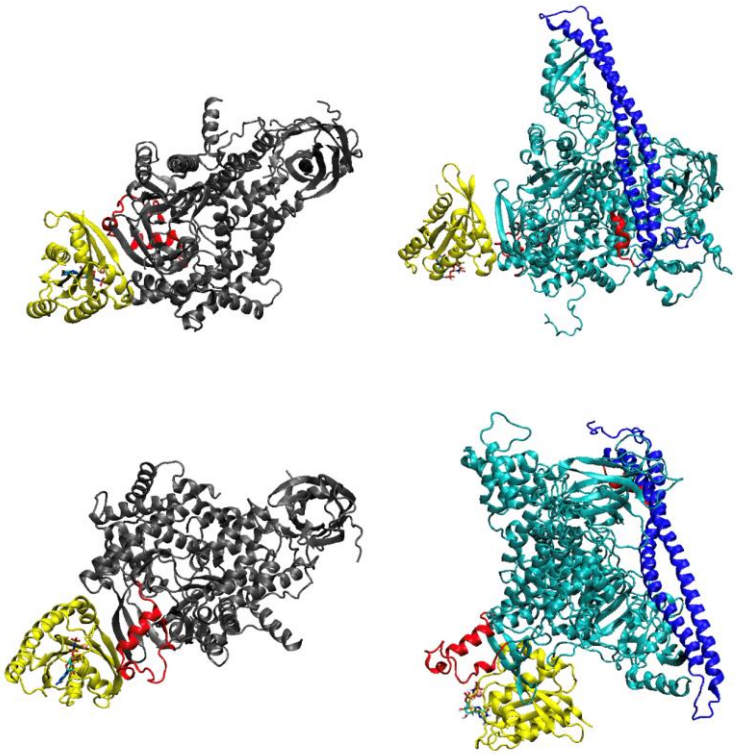


Figure 5. 10. Comparison of the triangle of α-helices in PI3Kγ (left) and as built for PI3Kα (right); front view (top), and back view (bottom). The p110α subunit is in cyan, p85α (α1 and α2 helices) in blue, RAS in yellow, and the newly-built helices (α3 and the α-helix triangle) are in red. The GNP ligand is drawn in liquorice representation with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen).

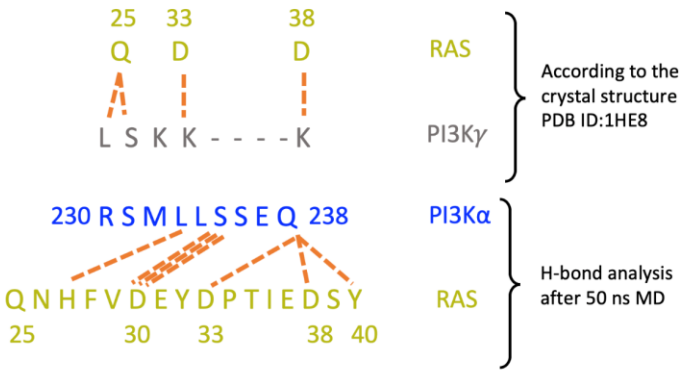


Figure 5. 11. Comparison of the pairs of residues (top) from PI3Kγ and RAS that interact in the PI3Kγ-RAS crystal structure (PDB: 1HE8) and (bottom) from PI3Kα and RAS that interact in the PI3Kα-RAS complex built here.

Stability of PI3Kα-RAS complex

In order to determine the nature of the PI3Kα-RAS interactions and whether they were maintained during the equilibration simulation after remodelling of the α-helical 'triangle', the distances between key residues and the formation of inter-protein hydrogen bonds were analysed.

First, the pairs of residues from PI3Kγ and RAS that interact in the PI3Kγ-RAS crystal structure were compared to the pairs of residues from PI3Kα and RAS that interact after the 50 ns MD simulation (Figure 5. 11). Comparing the binding sites of the two complexes, PI3Kα-RAS and PI3Kγ-RAS, the binding region is the same zone. Different residues of RAS interact with PI3Kα and -γ, because the sequence of the amino acids is completely different in the β-strand of PI3Kα and -γ with which RAS interacts. However, there are common interactions between ASP38 and ASP33 of RAS and both PI3Ks (Figure 5. 11).

The structure of the residues in RAS that interact with PI3Kα, which include the RAS switch I zone and the 7 residues preceding it, and of the residues in PI3Kα that they interact with in the PI3Kα-RAS complex, is shown in. The PI3Kα residues form parts of the newly-built triangle of α-helices in p110α. The distances between seven selected residues decrease during the simulation, indicating that RAS and PI3Kα are getting closer (Figure 5. 12).

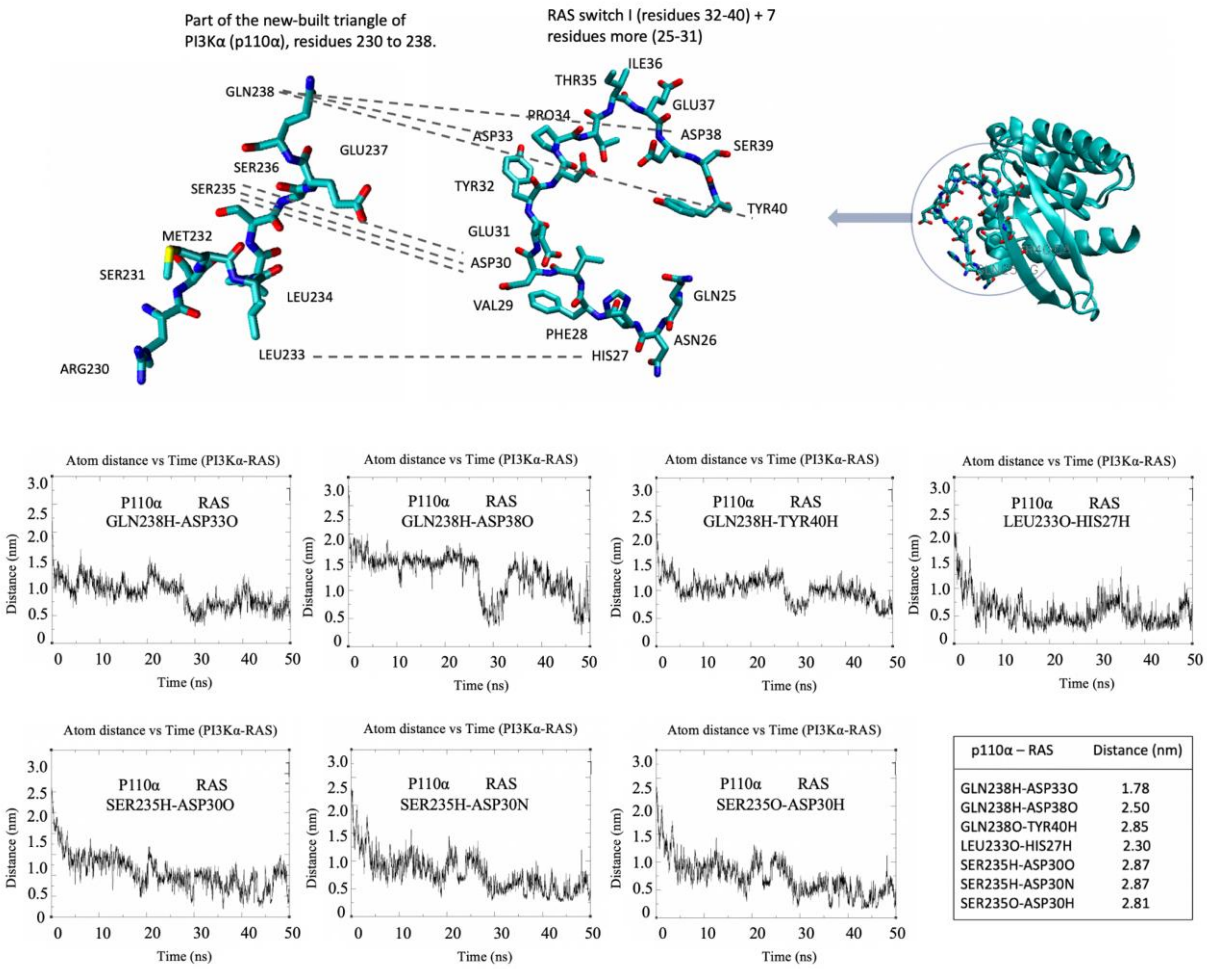


Figure 5. 12. Distances between key residues of PI3Kα and RAS in the PI3Kα-RAS complex. Top: Structure of the residues in RAS that interact with PI3Kα and of the residues in PI3Kα that they interact with in the PI3Kα-RAS complex, with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen). Dashed lines refer to the distances between key residue pairs prior to simulation. Bottom: Time-series of the distances between the key residue pairs during the 50 ns MD simulation, and the distances between these key residue pairs in the newly-built model before simulation.

All hydrogen bonds between PI3Kα and RAS that formed during the 50 ns MD simulation were calculated (Figure 5. 13 (a)). There are only two hydrogen bonds that are present for more than 70% of the simulation, between LEU244 of p110α and GLN25 of RAS, and between VAL243 of p110α and GLN25 of RAS. The time-series of the distance between the donor and acceptor atoms of the former (Figure 5. 13 (b)) confirms that the two chains are getting closer throughout the simulation, suggesting the proteins are slowly forming a tighter complex.

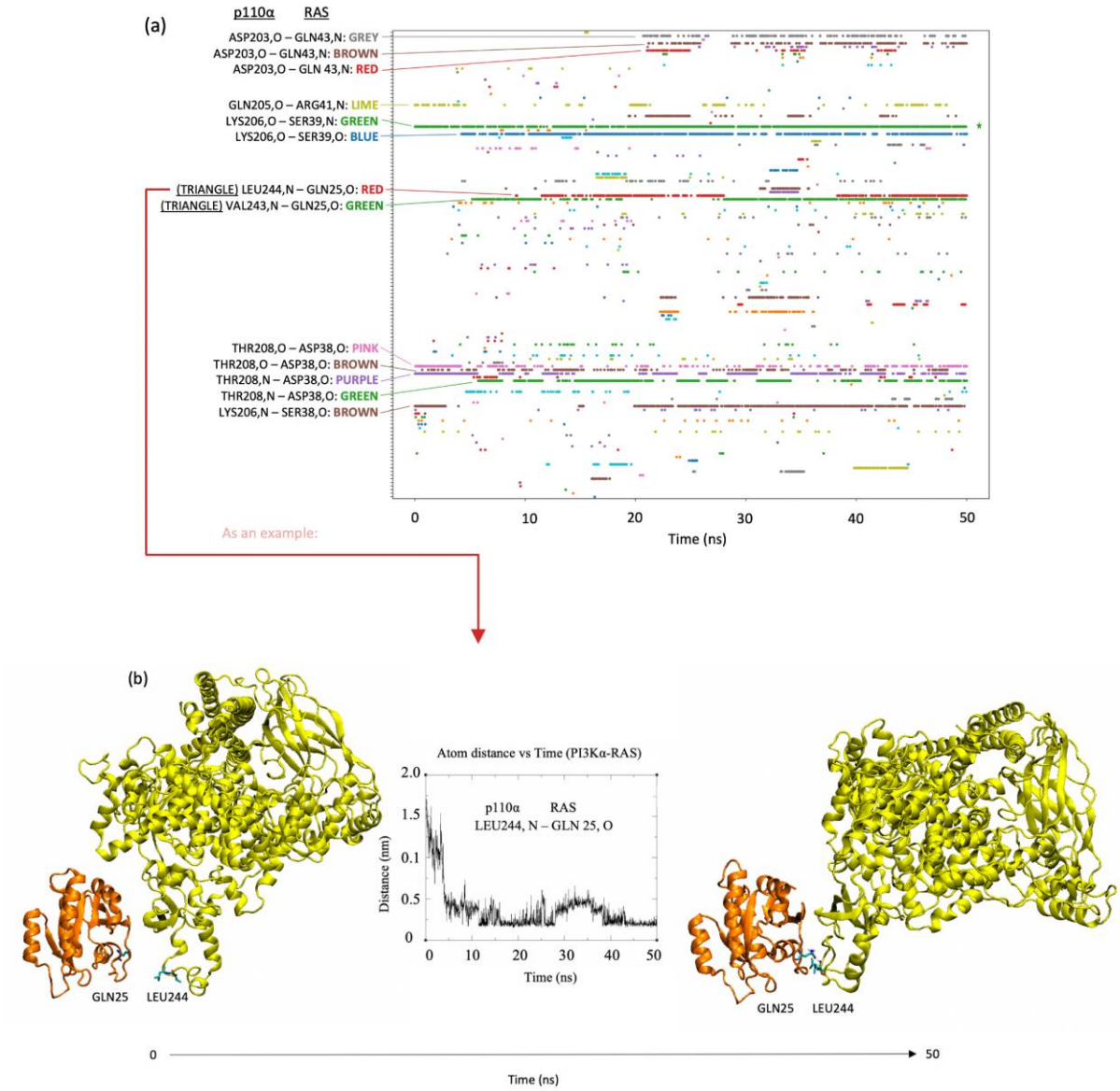


Figure 5. 13. Hydrogen bond formation between PI3Kα and RAS in the PI3Kα-RAS complex, with the α3 helix present. a) Time-series of hydrogen bond formation between residues in PI3Kα and RAS in the PI3Kα-RAS complex. The residues and atoms involved in the most frequently occurring hydrogen bonds are labelled, and those that involve the triangle of α-helices are underlined. The asterisk at the right-hand side of the graph shows the only hydrogen bonds which exist for more than 70% of the simulation time, and the inset to the left shows a typical time-series of the distance between the two atoms forming the hydrogen bond. b) Structures of the PI3Kα (yellow) - RAS (orange) complex at the (left) beginning and (right) end of the MD simulation, and (centre) time-series of the distance between LEU244 in PI3Kα and GLN25 in RAS, which are drawn in liquorice with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen).

5.3.1.5. Model excluding α3 helix of p85α

After visualising the simulation trajectory, it was noted that the α3 helix built in Section 5.3.1.1 was interfering with the activation loop of p110α during the simulation in a manner likely to prevent PI3Kα to its substrate, PIP_2. A new model of the complex was therefore built without the α3 helix, but retaining the triangle of α-helices built in Section 5.3.1.4.

The initial coordinates were taken directly from the Modeller output after building the triangle of α-helices in Section 5.3.1.4. Then α3 – the 12 residues at the end of p85α (589-600)

– was removed from the structure, which means p85α ends at GLY588, instead of ASN600. Finally, MD simulation was run to check the behaviour and interaction of the triangle of α-helices with RAS in the absence of the α3 helix.

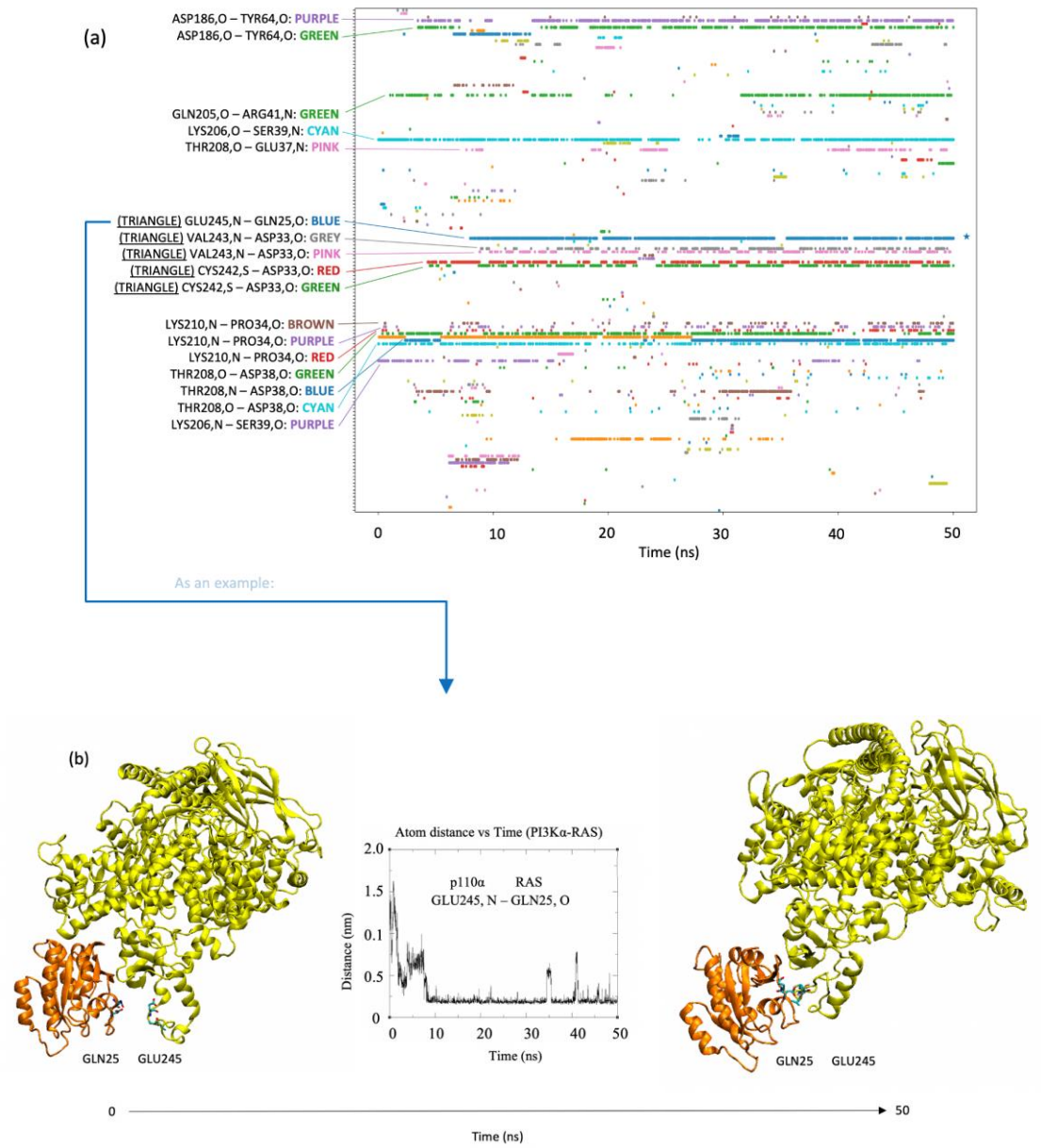


Figure 5. 14. a) Time-series of hydrogen bond formation between PI3Kα and RAS in the PI3Kα-RAS complex, without the α3 helix. The residues and atoms involved in the most frequently occurring hydrogen bonds are labelled, and those that involve the triangle of α-helices are underlined. The asterisk at the right-hand side of the graph shows the only hydrogen bonds which exist for more than 70% of the simulation time, and the inset to the left shows a typical time-series of the distance between the two atoms forming the hydrogen bond. b) Structures of the PI3Kα (yellow) - RAS (orange) complex without the α3 helix (left) before and (right) after the 50 ns simulation and (centre) time-series of the distance between GLU245 in PI3Kα and GLN25 in RAS, which are drawn in liquorice with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen).

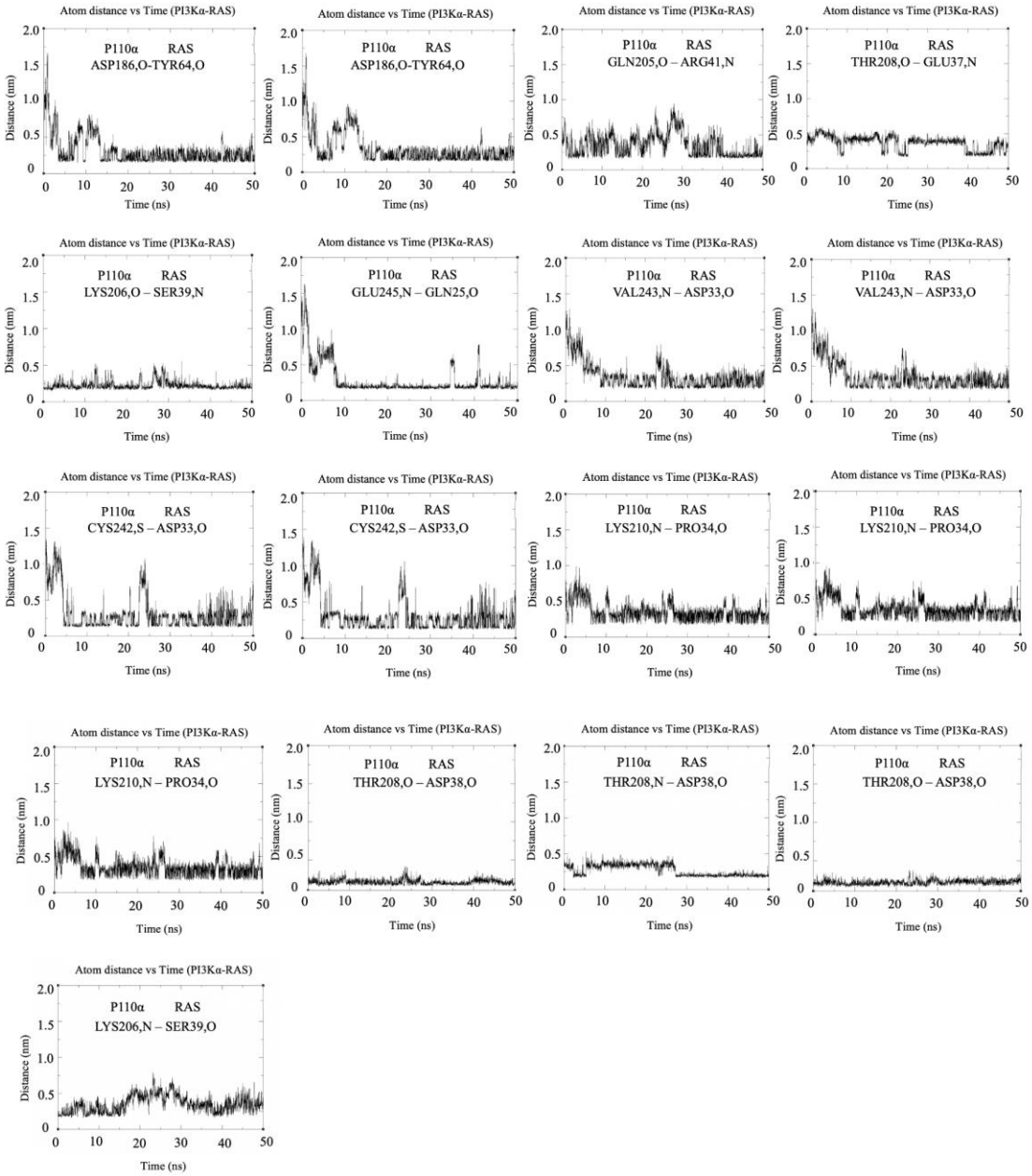


Figure 5. 15. Distance plots of all residue pairs that form persistent hydrogen bonds between PI3Kα and RAS in the PI3Kα-RAS complex built without the α3 helix over 50 ns simulation time.

Calculation of hydrogen bonds between PI3Kα and RAS during the simulation revealed that there are now five hydrogen bonds formed between the newly-built triangle of α-helices and the RAS switch I region. To have a better understanding of how the residues in each pair approach each other, the distances between all pair of residues identified as forming persistent hydrogen bonds in Figure 5. 14 (a) were plotted over the simulation time (Figure 5. 15). This showed that most of the bonds form either immediately or within the first 20 ns of the simulation and remain until the end of the simulation.

5.3.1.6. RAS-GNP-Mg^{2+} interactions

To determine whether the GNP ligand and Mg^{2+} ion remained stably bound by RAS, the interactions between GNP, RAS and the Mg^{2+} ion were analysed after the energy minimisation of the last simulation of the PI3Kα-RAS complex without the α3 helix (Figure 5. 16, and Figure 5. 18 (a)), to check if the initial coordinates of the simulation retain the interactions between GNP-Mg^{2+} and RAS. These interactions were consistent with those observed in the crystal structure of PI3Kγ-RAS (Figure 5. 17 (b), and Figure 5. 18) (67).

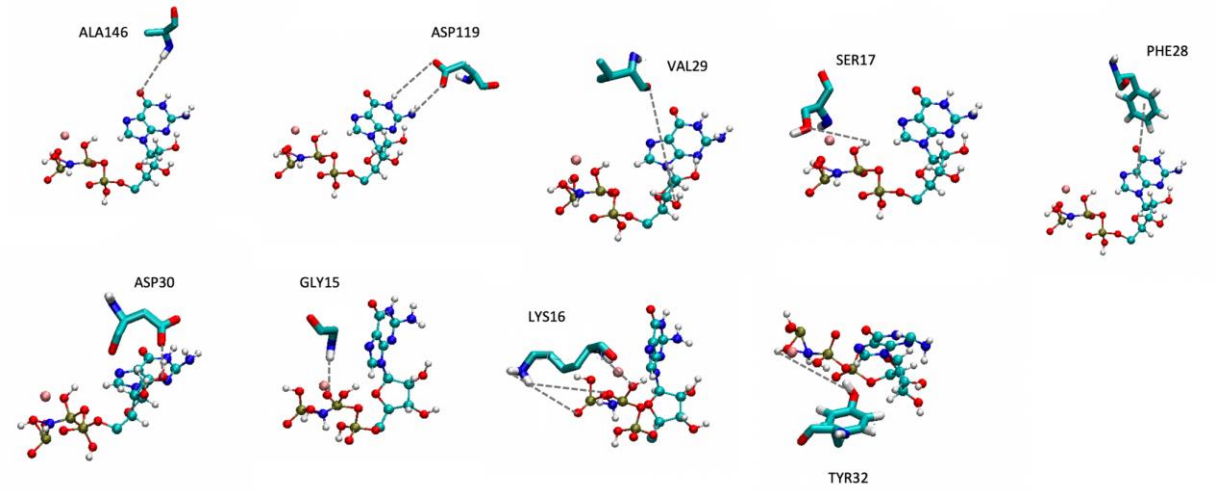


Figure 5. 16. Interactions between GNP and RAS residues in the PI3Kα-RAS complex without the α3 helix after energy minimisation. Dashed lines are drawn between atom pairs that form hydrogen bonds and the π-π interaction between PHE28 and the aromatic rings in GNP (top right)

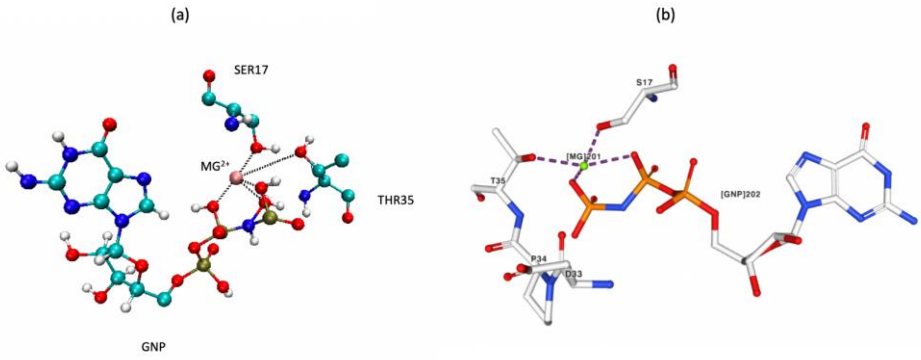


Figure 5. 17. Interactions between Mg^{2+} ion and RAS residues in (a) the PI3Kα-RAS complex without the α3 helix after energy minimisation, and (b) the PI3Kγ-RAS crystal structure (PDB ID: 1HE8 (67)). Right image from 1HE8 PDB website.

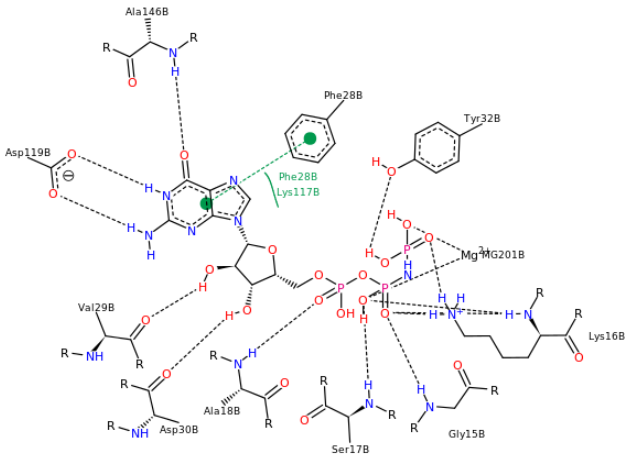


Figure 5. 18. Interactions between GNP and RAS residues in the PI3Kγ-RAS crystal structure (PDB ID: 1HE8 (67)). Image from 1HE8 PDB website.

In conclusion, removal of the α3 helix prevented it from blocking the activation loop of p110α, and there are also more hydrogen bonds formed between PI3Kα and RAS, and the interactions between RAS, its GNP ligand, and Mg^{2+} were consistent with those observed in the crystal structure. Therefore, the structure without the α3 helix was chosen to continue building the model of the PI3Kα-RAS complex.

5.3.1.7. Restraining of the PI3Kα-RAS interactions

The goal of building the triangle of α-helices was to prevent these residues disrupting the alignment of the β-strand in the RAS binding domain (RBD) of PI3Kα with a β-strand in RAS to form a continuous β-sheet. Unfortunately, at the end of the 50 ns simulation of the PI3Kα-RAS complex with the triangle of α-helices and without the α3 helix, despite the favourable properties reported above, the two β-strands did not remain aligned.

To test whether it is possible to keep the two β-strands aligned without disrupting either protein structure, p110α and RAS were position restrained and the system was again simulated for 50 ns. After visualisation of the simulated trajectory, it was found that the two β-strands were perfectly aligned as expected.

Position restraints are not a good solution to this problem, however, as the goal is to simulate the approach and interaction of the PI3Kα-RAS complex to the cell membrane. Distance restraints are independent of the position of the protein complex. Therefore, distance restraints between several pairs of residues of the two β-strands were applied. At first, the distances between the Cα atoms of 12 pairs of residues (Table 5. 1) were restrained, with the

target distances chosen according to the distances in the PI3Kγ-RAS complex (67). The system was heated in the NVT ensemble for 25 ps to test how the strands behaved. Unfortunately, after the short heating step, the RAS β-strand had become a loop (Figure 5. 19). This may be because the distance restraints were too strong or numerous and caused the β-strand to unfold. In order to avoid this, 5 pairs of residues were selected among the 12 to be distance restrained. After a short 25 ps NVT simulation, the two β-strands remained mostly β-strand and also stayed aligned (Figure 5. 20). Therefore, the system was simulated for 250 ps NVT, followed by 250 ps NpT equilibration. Favourably, the two β-strands retained their β-strand structure and remained aligned.

Table 5. 1. List of all 12 residues for which the distance between their Cα atoms was restrained. Shaded lines show the 5 chosen final residue pairs. r_1 and r_2 refer to the parameters of Equations 2 and 3.

Residues in p110α	Residues in RAS	r_1 (nm)	r_2 (nm)
ASP203	ARG41	0.80	0.85
LYS204	ARG41	0.80	0.85
LYS204	TYR40	0.80	0.85
GLN205	SER39	0.70	0.80
GLN205	TYR40	0.80	0.85
GLN205	ARG41	0.80	0.85
LYS206	SER39	0.70	0.80
LYS206	TYR40	0.80	0.85
TYR207	ASP38	0.70	0.80
TYR207	SER39	0.70	0.80
THR208	ASP38	0.70	0.80
LEU209	ASP38	0.70	0.80

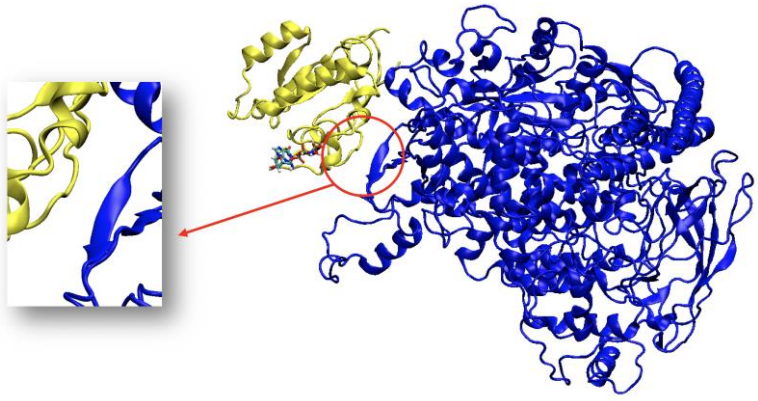


Figure 5. 19. Structure of the PI3Kα-RAS complex after a 25 ps NVT simulation with the distance between the Cα atoms of 12 pairs of residues (listed in Table 1) restrained. PI3Kα is blue, RAS is yellow, and the GNP ligand is drawn in liquorice style with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen). The RAS β-strand that became a loop is circled and shown in close-up.

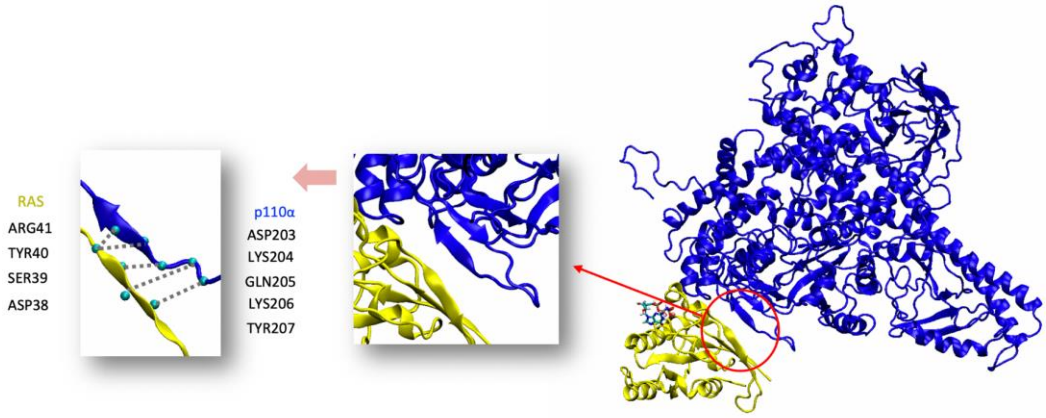


Figure 5. 20. Structure of the PI3Kα-RAS complex after a 25 ps NVT simulation with the distance between the Cα atoms of the final 5 pairs of residues (listed and highlighted in yellow in Table 1) restrained. PI3Kα is blue, RAS is yellow, and the GNP ligand is drawn in liquorice style with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen). The RAS β-strand that became a loop is circled and shown in close-up, the 5 pairs of residues are labelled and the distances between their Cα atoms (shown as cyan spheres) are drawn as dotted lines.

5.3.1.8. Adding cell membrane model

The goal of this research is to investigate the interaction of the PI3Kα-RAS complex with the cell membrane. A model of the cell membrane, in the form of a mixed lipid bilayer comprising SOPE, SOPS, POPC, cholesterol, sphingomyelin, and SA-PI[4,5]P_2) in the ratio 9:4:3:2:1:1, had been constructed and equilibrated previously (45). The coordinates of this mixed lipid bilayer were energy minimised and simulated for 50 ns in water to check how the membrane behaves with the adjusted parameters for sphingomyelin and cholesterol (as explained in Section 5.2.1). No anomalous behaviour was found (data not shown).

To simulate the approach of the PI3Kα-RAS complex to the model cell membrane, the starting position and orientation of the proteins relative to the lipid bilayer is very important. At first, the protein complex was superimposed with a protein configuration identified using rotational interaction energy profiling (RIEP) (82) to find the orientation with the optimal protein-bilayer interaction energies (45). However, in that work, the protein system consisted of just PI3Kα. With RAS added to the protein complex, superimposition resulted in some parts of the protein complex inserting into the membrane. Therefore, in the first instance, a 50 ns MD simulation was initiated from a starting protein configuration having the same orientation relative to the lipid bilayer but shifted further away. However, no movement towards membrane took place (data not shown).

An alternative initial orientation of the complex was sought by comparison with the Vps34 complex II from *S. cerevisiae* (257). The class III PI3K is known as Vps34 (vacuolar protein sorting 34, encoded by PIK3C3) and exists in all eukaryotes. The structure of Vps34

(for which there is a crystal structure, PDB ID: 5DFZ), contains some parts of p110α, and its interaction with the membrane has been predicted (257). The PI3Kα portion of the PI3Kα-RAS complex was therefore superimposed on the Vps34 complex II such that the orientation of the PI3Kα-RAS complex with respect to the membrane matched the predicted membrane interaction of the Vps34 complex II (257). This showed that this orientation was suitable for initiation of MD simulations to follow the approach and binding of PI3Kα-RAS to the model cell membrane.

After energy minimization followed by a 500 ps NpT equilibrium, it was found that not only did the PI3Kα-RAS complex not associate with the mixed lipid bilayer – rather, it moved away – the two β-strands did not remain aligned into a continuous β-sheet. A 100 ns MD simulation including distance restraints for the five residue pairs to link the two β-strands resulted in half of the RAS β-strand becoming a loop (Figure 5. 21Figure 5. 22).

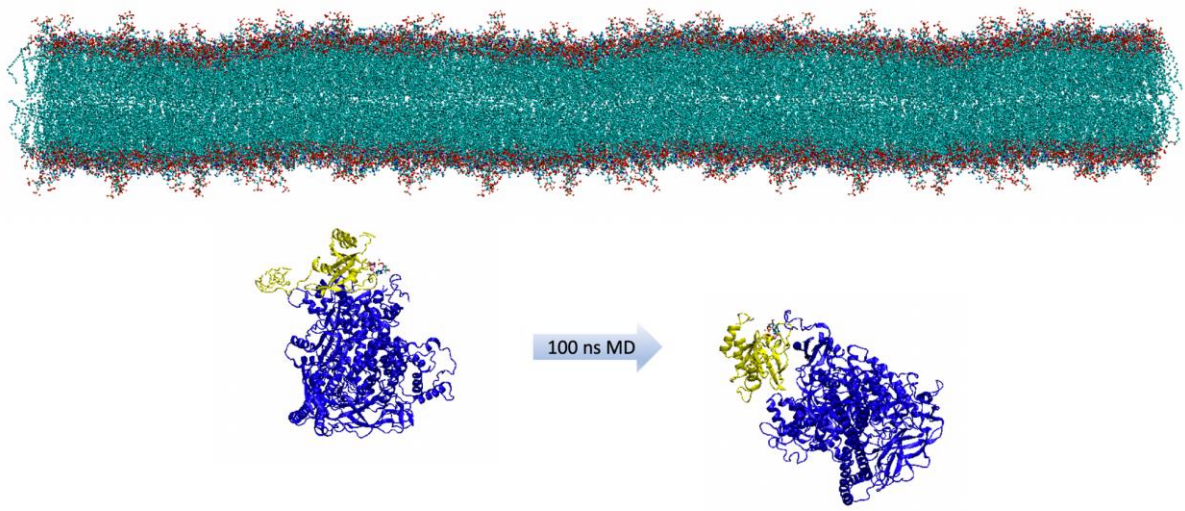


Figure 5. 21. Change in orientation and position of the PI3Kα-RAS complex during MD simulation in the presence of a mixed lipid bilayer. The position and orientation of the PI3Kα-RAS complex with respect to the mixed lipid bilayer (left) before and (right) after a 100 ns MD simulation. PI3Kα is blue, RAS is yellow, and the GNP ligand and lipids are drawn in liquorice and CPK styles, respectively, with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen). The mixed lipid bilayer was extended in the -X and +X directions and water molecules were removed for visualisation purposes.

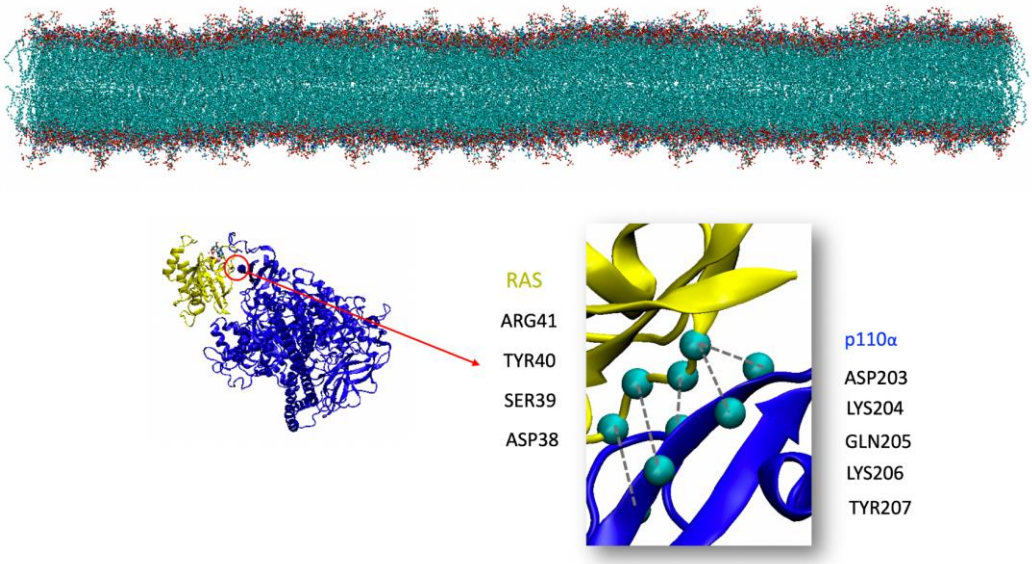


Figure 5. 22. Change in structure of the RAS β-strand after a 50 ns MD simulation of the PI3Kα-RAS complex in the presence of a mixed lipid bilayer with distance restraints between the five residue pairs (highlighted in yellow in Table 1). PI3Kα is blue, RAS is yellow, and the GNP ligand and lipids are drawn in liquorice and CPK styles, respectively, with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen). The RAS β-strand that became a loop is circled and shown in close-up, the 5 pairs of residues are labelled and the distances between their Cα atoms (shown as cyan spheres) are drawn as dotted lines. The mixed lipid bilayer was extended in the -X and +X directions and water molecules were removed for visualisation purposes.

5.3.1.9. Embedding RAS into membrane

Given the lack of success in simulating the approach of a PI3Kα-RAS to a cell membrane model, a different approach was required. Recent hydrogen-deuterium exchange mass spectrometry (HDX-MS) results from Siempelkamp et al. (234) suggested that membrane-resident H-RAS increases membrane association by PI3Kα. Additionally, soluble H-RAS causes a small decrease in the lipid kinase activity of PI3Kα. On the other hand, H-RAS coupled to the mixed lipid bilayer caused a major increase in PI3Kα lipid kinase activity. These results highlighted the importance of H-RAS being localized at the membrane in order to study the interactions of H-RAS-PI3Kα productively (234).

Having created a model of the PI3Kα-RAS complex in which p85α does not interfere with the activation loop of p110α, and the p110α-RAS interaction is stable and maintains the structure of the interacting β-strands, the next step was to lipidate RAS. Other studies have also shown that O-methylation of the carboxyl terminus of farnesylated proteins brings about an increase in binding of these proteins to the membrane by about 60% (258, 259). Therefore, RAS was lipidated as described in Section 5.2.2 and the lipid anchors were embedded into the lipid bilayer. This was achieved by placing lipidated RAS such that the lipids penetrated into the bilayer, then deleting the two lipid molecules (one CHOL and one PE) with which the RAS

lipids overlapped. After minimization of this system, with RAS anchored to the mixed lipid bilayer, PI3Kα was added, positioned such that the RBD faced towards the RAS switch I (residues 32-40) with which it interacts in the PI3Kγ-RAS crystal structure and in the PI3Kα-RAS model structures built earlier (Figure 5. 23).

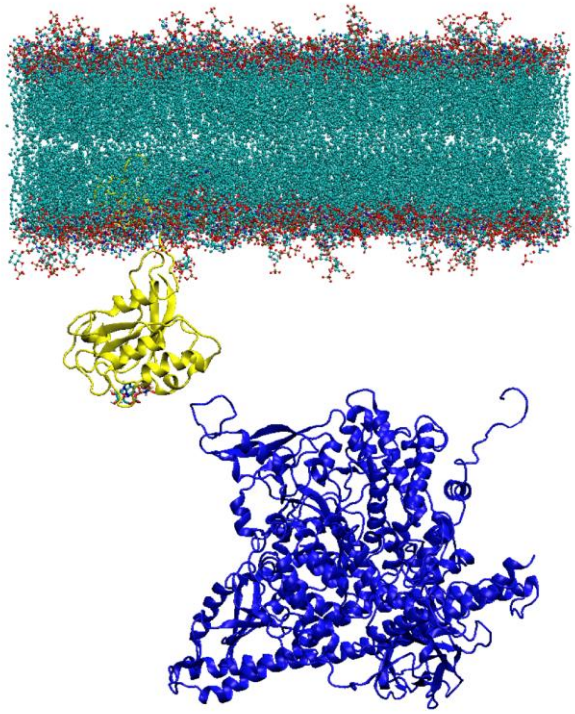


Figure 5. 23. PI3Kα-RAS and mixed lipid bilayer system with the RAS lipids embedded into the mixed lipid bilayer. PI3Kα is blue, RAS is yellow, and the GNP ligand and lipids are drawn in liquorice and CPK styles, respectively, with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen).

After energy minimisation, an attempt to equilibrate the system failed due to LINCS warnings. These typically indicate an underlying problem with the initial coordinates. After personal communication with our experimental collaborator (Jack Flanagan, University of Auckland), and a more detailed analysis of the Siempelkamp et al. (234) results, it was determined that the hypervariable domain of RAS should extend outwards from the bilayer surface so that the globular domain of RAS is positioned some distance away from the membrane.

Such a configuration was achieved by starting from the coordinates in which the RAS lipids are embedded in the mixed lipid bilayer, and applying a pulling force to RAS using steered MD (247–249) (see Section 5.2.3.1). At the end of the steered MD simulation, the hypervariable domain was extended and the minimum distance between RAS and the mixed lipid bilayer was 4.0 nm. These system coordinates were energy minimised and equilibrated for 500 ps. Finally, PI3Kα was positioned in the same way described previously, such that the RBD faced towards the RAS switch I region (residues 32-40) (Figure 5. 24).

The complete PI3Kα-RAS-membrane system was energy minimised, equilibrated, and simulated for 50 ns. Unfortunately, at the end of the MD simulation, the RAS hypervariable region was folded back against the mixed lipid bilayer (Figure 5. 25). While this drew the

PI3Kα towards the mixed lipid bilayer, the interactions between PI3Kα and RAS and between PI3Kα and the mixed lipid bilayer were not in agreement with experimental data (234, 238), with RAS associating with the opposite side of the PI3Kα-membrane binding site to what is expected (234).

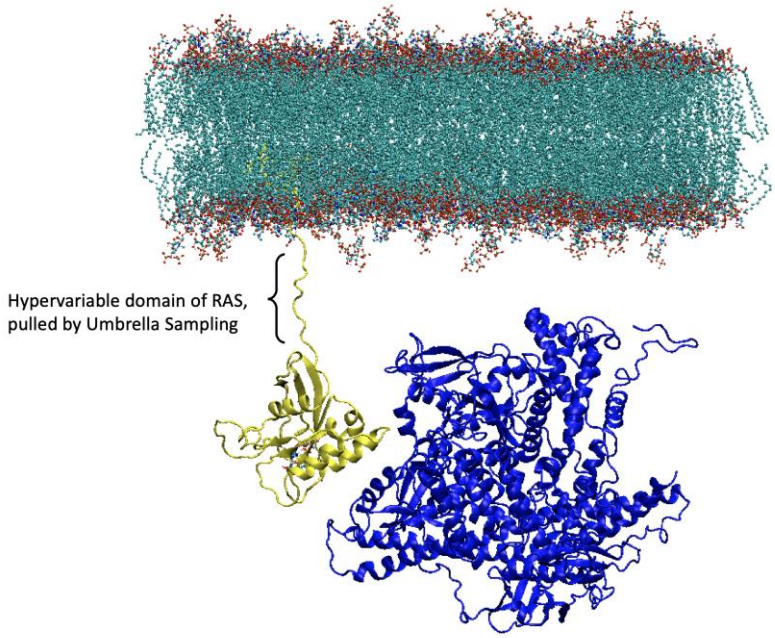


Figure 5. 24. PI3Kα-RAS and mixed lipid bilayer system with the RAS lipids embedded into the mixed lipid bilayer after extension of the hypervariable domain of RAS using steered MD. PI3Kα is blue, RAS is yellow, and the GNP ligand and lipids are drawn in liquorice and CPK styles, respectively, with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen).

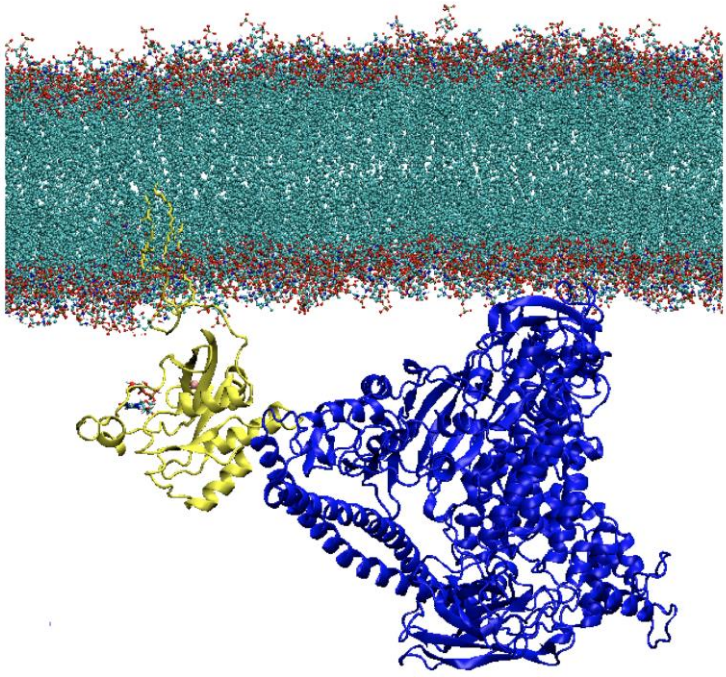


Figure 5. 25. PI3Kα-RAS and mixed lipid bilayer system with the RAS lipids embedded into the mixed lipid bilayer after 50 ns MD simulation. PI3Kα is blue, RAS is yellow, and the GNP ligand and lipids are drawn in liquorice and CPK styles, respectively, with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen).

5.3.1.10. Restraining RAS to its initial coordinates

To prevent this unfavourable folding of the hypervariable domain, RAS was position restrained to the initial coordinates, in which the hypervariable domain is extended. After a 50 ns simulation with position restraints, the hypervariable domain remained extended, suggesting that this approach would work. PI3Kα was therefore added as described previously. This time, after 50 ns MD simulation, no folding of the RAS hypervariable domain occurred, and PI3Kα was approaching RAS. During a subsequent additional 100 ns of MD simulation, PI3Kα continued to slowly approach RAS and rotate into an orientation suitable for RAS binding and interaction with the mixed lipid bilayer (Figure 5. 26).

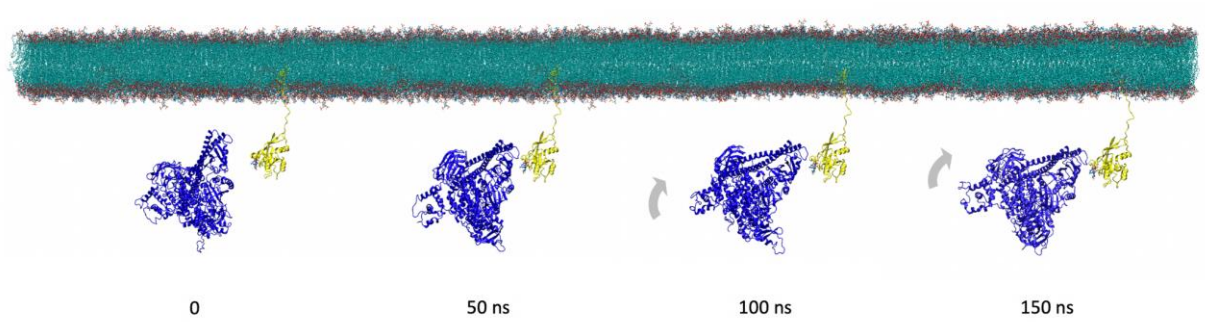


Figure 5. 26. PI3Kα-RAS and mixed lipid bilayer system with the RAS lipids embedded into the mixed lipid bilayer and RAS position-restrained so that the hypervariable domain remains extended. PI3Kα is blue, RAS is yellow, and the GNP ligand and lipids are drawn in liquorice and CPK styles, respectively, with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen). PI3Kα approaches RAS within the first 50 ns, and then slowly rotates into an orientation better suited to binding to RAS and for interaction with the mixed lipid bilayer

5.3.2. Development of PI3Kα-RAS interactions

To monitor the formation of interactions between PI3Kα and RAS, hydrogen bonds between PI3Kα and RAS were calculated during the first 100 ns of MD simulation (Figure 5. 27). Interestingly, all hydrogen bonds were between residues in the p85α subunit of PI3Kα and in RAS. Figure 5. 28 shows the five residue pairs which formed hydrogen bonds during the simulation. Looking at the hydrogen bonds plot (Figure 5. 27), after approximately 15 ns, two hydrogen bonds formed (red and green colours in Figure 5. 27).

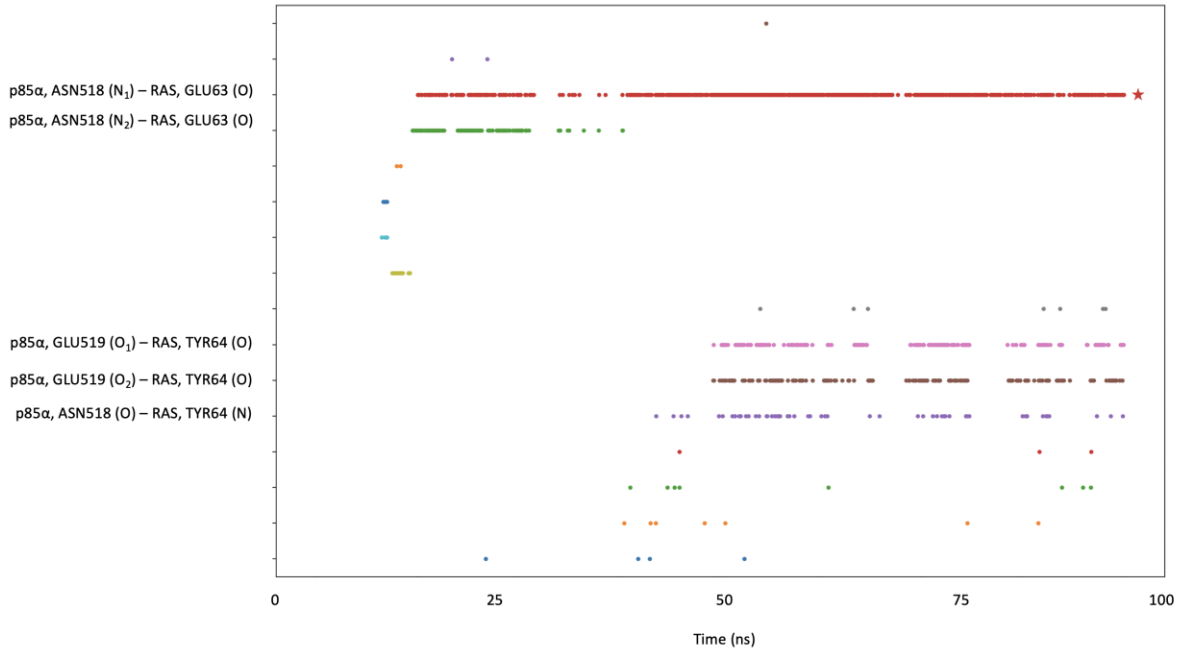


Figure 5. 27. Hydrogen bond formation between PI3Kα and RAS in the first 100 ns of the MD simulation of PI3Kα in the presence of RAS anchored to the mixed lipid bilayer and position restrained. The residues and atoms involved in the most frequently occurring hydrogen bonds are labelled, and the asterisk at the right-hand side of the graph shows the only hydrogen bond that exists for more than 90% of the simulation time.

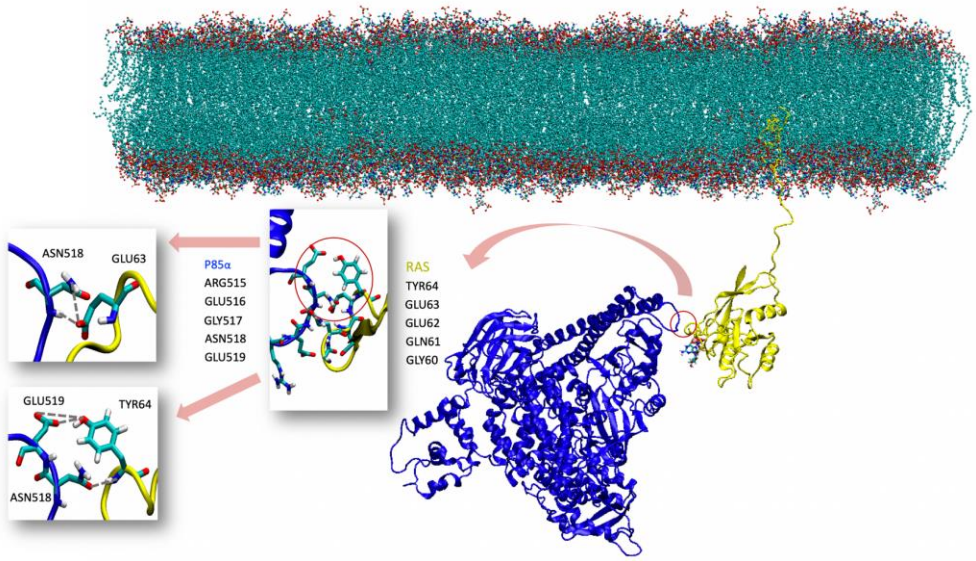


Figure 5. 28. The five hydrogen bonds between PI3Kα and RAS which formed during the 100 ns MD simulation of PI3Kα in the presence of RAS anchored to the mixed lipid bilayer and position restrained. PI3Kα is blue, RAS is yellow, and the GNP ligand and lipids are drawn in liquorice and CPK styles, respectively, with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen). The location of the hydrogen bonds is circled in red, the structures of the residues involved are drawn in liquorice representation and coloured according to type and the residue names and numbers are listed. In the insets to the left, the hydrogen bonds are indicated as grey dashed lines.

After 50 ns, three other bonds started forming between the two chains. However, one of the former bonds (green colour in Figure 5. 27) did not show up after that. It seems that the main hydrogen bond during the 100 ns MD simulation is between ASN518 (p85α) and GLU63

of RAS. This is unfortunately not in keeping with the expected interactions between the β-strands of PI3Kα and RAS.

The simulation was extended to 1 microsecond (1000 ns) in the hope that this would encourage PI3Kα and RAS to interact further, ideally resulting in interaction of the two β-strands. Surprisingly, even after 1 μs, the two β-strands were not aligned, and nor did PI3Kα move up to the membrane. However, PI3Kα remained interacting with RAS. The only noticeable, albeit slow, movement occurred between 405-715 ns (Figure 5. 29).

The hydrogen bonds between PI3Kα and RAS during the whole simulation are listed in Figure 5. 29, and illustrated in Figure 5. 30, according to the time frames in which they appeared. Before 405 ns, PI3Kα was interacting with the same zone of RAS, mostly ASN518 and GLU519 of p85α with GLU63 and TYR64 of RAS, as in the initial 100 ns. After 715 ns, new hydrogen bonds were formed between GLU516 of p85α and ASN86 of RAS. Hydrogen bonds were also formed between PI3Kα and GNP.

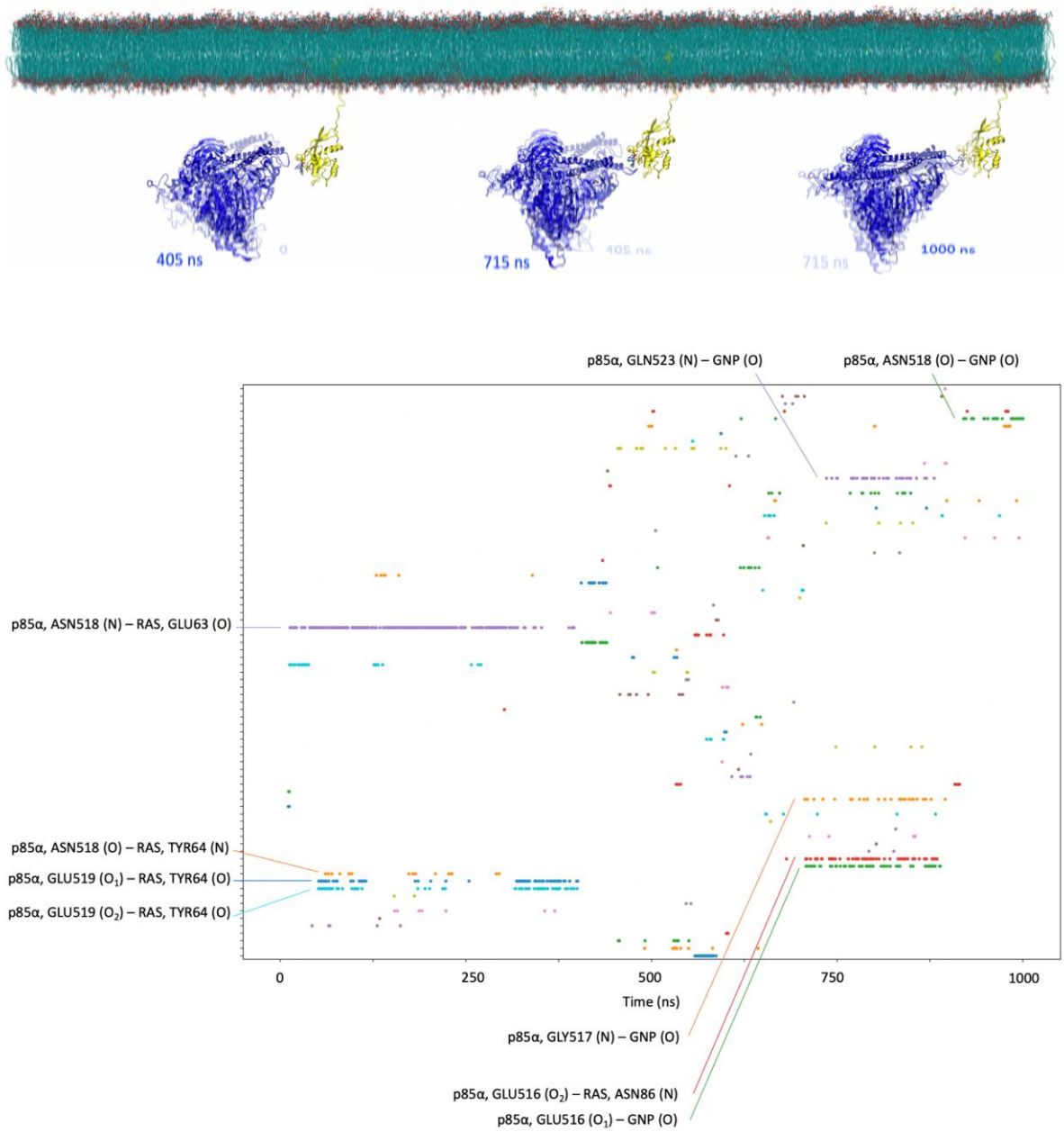


Figure 5. 29. (top) The major movements of PI3Kα relative to RAS and the membrane during the 1 μs MD simulation of PI3Kα in the presence of RAS anchored to the mixed lipid bilayer and position restrained. PI3Kα is blue, RAS is yellow, and the GNP ligand and lipids are drawn in liquorice and CPK styles, respectively, with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen). At each of the labelled time-points, the position of PI3Kα at the previous time-point is drawn as a shadow. (bottom) The hydrogen bonds between PI3Kα and RAS which formed during the 1 μs MD simulation. The residues and atoms that form hydrogen bonds present for more than a few ns are labelled.

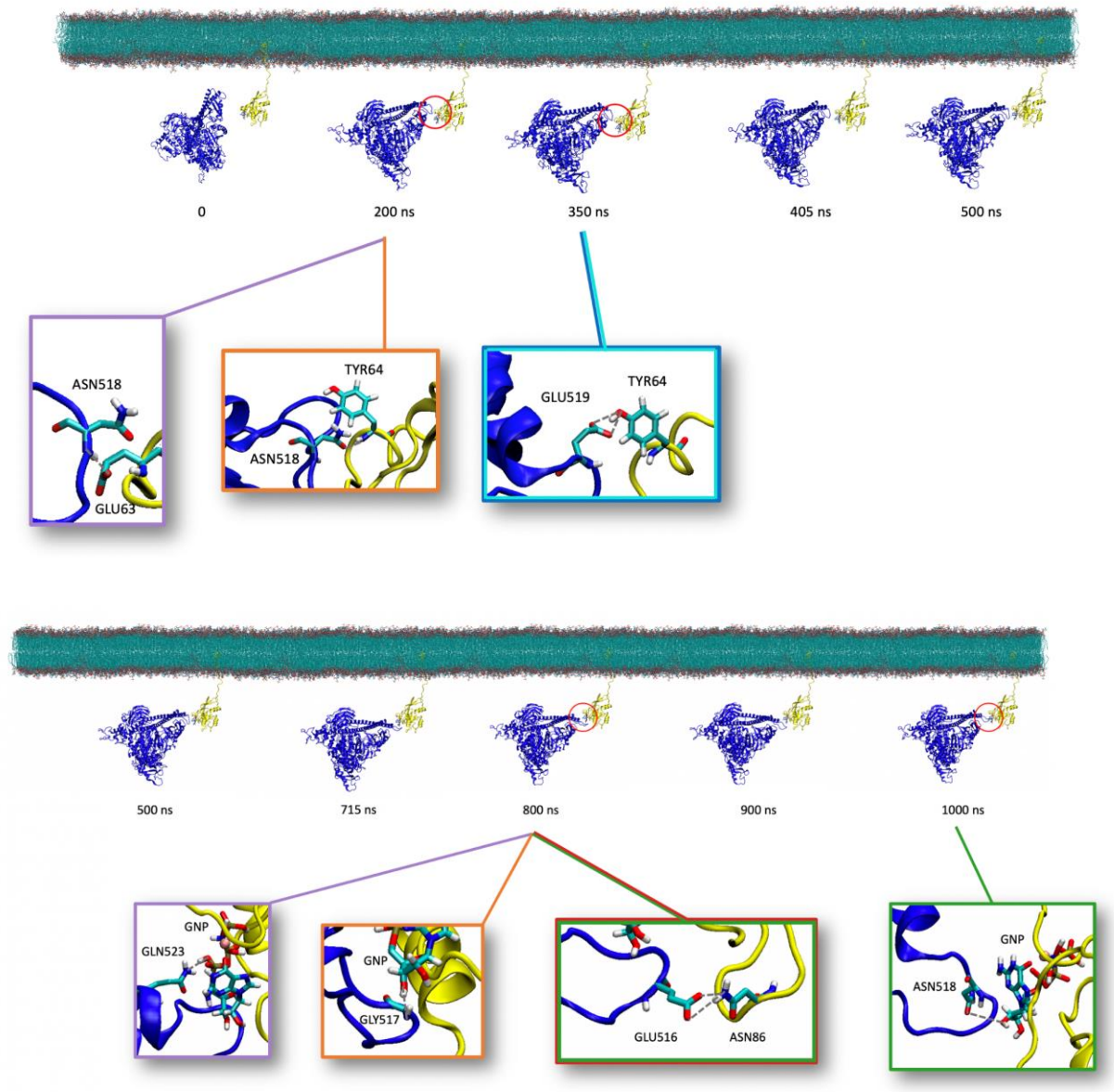


Figure 5. 30. Time-series of snapshots showing the progress of the interaction between PI3Kα and RAS during the 1 μs MD simulation of PI3Kα in the presence of RAS anchored to the mixed lipid bilayer and position restrained. PI3Kα is blue, RAS is yellow, and the GNP ligand and lipids are drawn in liquorice and CPK styles, respectively, with atoms coloured according to type (cyan: carbon; red: oxygen; blue: nitrogen). The location of the hydrogen bonds is circled in red, and in the insets, the structures of the residues involved are drawn in liquorice representation and coloured according to type, the residue names and numbers are listed, and the hydrogen bonds are indicated as grey dashed lines.

Although there are interactions between PI3Kα and RAS in this simulation, they do not match the expected interactions between the β-strands. Therefore, this orientation, which was outcome of superimposition of the two p110α domains in PI3Kα and the structure of Vps34 (PDB ID: 5DFZ), appears not to be a productive orientation and position for investigating the PI3Kα-RAS-membrane interaction. In future work, a closer homologue of PI3Kα interacting with membrane might be a better starting point.

5.4. Conclusion

In this chapter, the ultimate goal was to study the interaction of the PI3Kα-RAS complex with a cell membrane model, and thus understand how RAS affects this process. However, since there was no model for such a system, first, a complex of modified PI3Kα and lipidated RAS bound to GNP ligand and with its lipids embedded in a lipid bilayer was prepared in order to study the binding mode. Each component of the complex was prepared separately, starting with PI3Kα. Several modifications to PI3Kα were trialled, including removing the HIS-tag, remodelling the RBD α-helical 'triangle', and addition of α3 helix to p85α domain. The RAS models available did not include the hypervariable region and lipid anchors, therefore, a post-translational modified model of RAS was built. Then it was associated with the available lipid bilayer model (45) by embedding the lipids into the membrane. In order to add PI3Kα to the RAS-membrane system, a crystal structure of Vps34 complex II (PDB ID: 5DZF), which contains some parts of p110α and for which the interaction with the membrane has been predicted (257), was used as a template. The PI3Kα portion of the PI3Kα-RAS complex was superimposed on the Vps34 complex II such that the orientation of the PI3Kα-RAS complex with respect to the membrane matched the predicted membrane interaction of the Vps34 complex II (257). A 1 μs MD simulation, including position restraining of RAS, to prevent the hypervariable region folding up, was run for the final PI3Kα-RAS-membrane system. However, after analysis of features such as hydrogen bond formation and visualisation of the MD trajectory, it was found that the two β-strands (in RAS and PI3Kα) were not aligned. This alignment was expected according to the PI3Kγ-RAS interaction (67), and its absence suggests that the two proteins are not interacting as they should. Therefore, while PI3Kα did interact with RAS, it is unlikely to be the correct orientation and position of PI3Kα in complex with the RAS-membrane system. The fact that the PI3Kα-RAS complex did not move closer to the membrane is further evidence that the complex structure formed here is not productive for membrane interaction and binding.

A future approach to obtaining an initial orientation and position of the PI3Kα-RAS complex for simulation of its approach to the membrane surface could be by comparison with experimental HDX-MS results. Several such studies have been carried out for PI3Ks to determine the regions involved in membrane binding (76, 98, 238, 260), and one was carried out for a PI3Kα-RAS complex (76), making its results useful for informing the best initial orientation of the PI3Kα-RAS complex to use here. This would involve manual reorientation

and repositioning of PI3Kα and potentially also RAS (keeping the lipids buried in the lipid bilayer and the hypervariable region extended) such that both proteins are in an orientation in keeping with the HDX-MS data. Ideally, two simulations would be run, one with the PI3Kα-RAS complex already formed, to follow its approach to the membrane surface, and one with PI3Kα positioned close to and in the correct orientation for binding to RAS, such that formation of the complex can be studied.

Chapter 6: Conclusions

In this work, binding modes and intermolecular interactions between different molecules at a range of scales were studied using MD simulations. The systems that were studied were complexes of the fluctional molecule bullvalene with cyclodextrin (Chapter 2), the transcriptional repressor KstR and its interactions with its ligand and DNA (Chapter 3), another related transcriptional repressor KstR2 interacting with its ligand (Chapter 4), and the class IA PI3Kα (phosphatidylinositide 3-kinase) interacting with RAS and with a model cell membrane (Chapter 5).

The goal of Chapter 2 was to capture one isomer of the fluctional molecule bullvalene (guest molecule) by complexation with cyclodextrin (host molecule). In the first part of this chapter, 2i, a method for determining favourable binding modes of guest molecules in host molecules, named "host-guest binding potential energy profiling" (HGBPEP), was developed and validated. This method was able to reproduce the experimentally-determined binding mode of para-chlorophenol.

Chapter 2ii applied HGBPEP, as well as MD simulations and binding free energy calculations, to study the interaction of bullvalene isomers and cyclodextrin molecules. A methodological pipeline was developed that began with use of HGBPEP to identify favourable binding modes, MD simulations to evaluate their stability, and lastly binding free energy calculations to quantify the likelihood of each complex forming. This procedure was carried out for unsubstituted bullvalene, monosubstituted hydroxymethyl-bullvalene, and disubstituted dihydroxymethyl-bullvalene, with β-cyclodextrin and γ-cyclodextrin.

In all cases, bullvalene was likely to bind to cyclodextrin, but, for instance, some orientations of the most favourable isomer of the disubstituted bullvalene cannot fit inside the cavity of β-cyclodextrin and so instead binds at the wider rim of the cyclodextrin, whereas for unsubstituted bullvalene with γ-cyclodextrin, the cavity is too large and thus allows in water molecules as well as the bullvalene, making binding less favourable.

For hydroxymethyl-bullvalene with β-cyclodextrin and γ-cyclodextrin and dihydroxymethyl-bullvalene with β-cyclodextrin, experimental binding free energies have been determined using NMR titrations. The computational results agreed exceptionally well with the experimental results, and, for hydroxymethyl-bullvalene, were able to identify the most likely isomer captured by both β-cyclodextrin and γ-cyclodextrin to be isomer B, which is not the most populated isomer in solution. For dihydroxymethyl-bullvalene with β-cyclodextrin, the computational results also agreed almost perfectly with the experimental

results, and are able to identify the most likely binding mode of the most populated isomer. The advantage of using a computational approach to study this host-guest complexation is that each bullvalene isomer can be studied individually, whereas experimentally, the isomers constantly interconvert and cannot be distinguished, meaning that it is not possible to determine which isomer is bound by cyclodextrin.

Chapter 3 involved MD simulations of the apo, ligand-bound and DNA-bound states of KstR, an Mtb transcriptional repressor, in order to shed light on whether DNA binding occurs via an induced fit or conformational selection mechanism, and to investigate possible allosteric communication pathways by which information about ligand and DNA binding are transmitted through the protein.

The apo state was, somewhat surprisingly, found to be not particularly mobile, and did not sample a conformational capable of DNA binding during either of the two independent MD simulations that were run. The ligand-bound state was the most dynamic, with the DBDs moving in and out via a 'breathing' motion, which was noticeably different to the less extensive 'twisting' motion of the apo and DNA-bound states. Removal of the ligand or of DNA created an 'apo' state that behaved more like the ligand-bound state than the apo state simulation initiated from the apo crystal structure. However, neither of these simulations sampled a conformational capable of DNA binding either. These results suggested that an induced fit mechanism is the most likely way in which KstR binds to DNA.

A novel state of KstR was created by placing ligands into the ligand binding pockets while KstR was bound to DNA. It has been assumed that such a state is not feasible. The MD simulations of this doubly-bound state suggest that while it is possible, it is probably unlikely to form *in vivo*. The fact that ligand binding did not, as expected, force KstR to unbind from DNA, is an argument against an induced fit process for the unbinding of KstR from DNA. The doubly-bound state was highly strained, and underwent conformational dynamics similar to the DNA-bound state. Interestingly, upon removal of DNA, the protein remained constrained, and even after extension of the MD simulation, had not reverted to the highly mobile state observed for KstR with ligand bound.

Network analyses suggested that the α4 helix, which has been hypothesised to be involved in communication between the DBD and LBD, may indeed be involved in communicating changes of state. There was substantial asymmetry in the communities and communication pathways between the DBDs and LBDS for all states of KstR, suggesting that despite its apparent symmetry, it does not behave in a symmetric manner. Three major types

of communication pathways were observed, which differed between the different states of KstR, suggesting that communication of ligand and DNA binding depends on the state of KstR.

Chapter 4 focused on KstR2, another Mtb transcriptional repressor that is structurally similar to KstR but has been proposed to act through a novel scissor-like mechanism. This chapter investigated two key questions regarding the mechanism of action of KstR2: first, the effect of mutating the key switch residue ARG170 to ALA, and second, the effect of ligand binding on its structure and motion.

Mutation of ARG170 to ALA increased the rigidity of the KstR2 structure, but introduced asymmetry in the dynamics between the monomers. The overall nature of the dynamics remained 'breathing'-like, however. The communication pathways between ARG170/ALA170 and TYR48 became longer in the R170A state for both Rjo and Mtb KstR2, suggesting that this mutation causes a remarkable change in the allosteric communications between the LBD and DBD. Additionally, the inclusion of residues from the opposite monomer into the interface communities supported tight binding of the two monomers. This, and the evidence for a key role of ARG170 in communication, were in support of the proposed scissor-like mechanism.

The results of these simulations also provided insight into whether KstR2 operates via an induced fit or conformational selection mechanism. The apo state simulations of wild-type Rjo and Mtb KstR2 both showed jumps between widely-positioned DBDs, typical of an apo or ligand-bound structure, and conformations with the DBDs much closer together, reminiscent of a DNA-bound state. Without a crystal structure of DNA-bound KstR2, it is difficult to make more concrete conclusions from these results, but thus far, they support a conformational selection mechanism of DNA binding.

To investigate the effect of ligand binding on KstR2, simulations were carried out for KstR2 with one or two HIP-CoA ligands bound, only the HIP portion bound, or the ligands removed. Interestingly, states of KstR2 that are symmetric with respect to ligand binding exhibit asymmetric dynamics and communication pathways and *vice versa*. Ligand binding was shown to push the DBDs further apart, but, unlike for KstR, changed the motion from a relaxed breathing motion to a more twist-like motion that was retained with only one ligand bound. The dynamics of KstR2 were similarly asymmetric with one HIP-CoA or two HIP molecules bound, suggesting that the CoA portion in some way stimulates the dynamic asymmetry in response to ligand binding.

Interestingly, the complete HIP-CoA ligand in one monomer folded and twisted during the simulation, but this did not prevent formation of hydrogen bonds with KstR2. This mobility

may be why the CoA portion is not resolved in the crystal structure of ligand-bound KstR2. The protein-ligand interactions observed in the crystal structure mostly only occurred at low occupancy during the MD simulations, suggesting that ligand binding is dynamic. The higher binding affinity of the complete HIP-CoA ligand is most likely due to the ability of the CoA portion to form hydrogen bonds with KstR2. ARG162 was found to be a key residue for ligand binding, making mutational studies of this residue of interest for the future.

The focus Chapter 5 was how RAS affects the interaction of PI3Kα with the cell membrane. The majority of the work in this chapter involved painstaking building of a PI3Kα-RAS-membrane complex. Several alternative approaches had to be taken due to a number of unexpected problems that were encountered.

Ultimately, a 1 μs MD simulation was run starting from a complex in which RAS was anchored to the mixed lipid bilayer by the lipids added post-translationally to the hypervariable region, and PI3Kα was placed nearby in an orientation that both matched experimental results regarding the optimal orientation for binding to the membrane and appeared to position it in a way that might encourage interaction with RAS. Unfortunately, despite the long time-scale simulation, only minor changes in the position and orientation of PI3Kα were observed. The interactions that PI3Kα formed with RAS were not those expected based on the crystal structure of the related PI3Kγ-RAS complex. Time constraints meant that no further simulations could be run, but ideas on what to try next were proposed.

Bibliography

1. Lehn, J.-M. (1988) Supramolecular Chemistry-Scope and Perspectives Molecules, Supermolecules, and Molecular Devices(Nobel Lecture). *Angew. Chemie Int. Ed. English*. **27**, 89–112
2. Stoddart, J. F. (1988) Chapter 12. Host–guest chemistry. *Annu. Rep. Prog. Chem., Sect. B Org. Chem.* **85**, 353–386
3. Dotsikas, Y., and Loukas, Y. L. (2003) Efficient determination and evaluation of model cyclodextrin complex binding constants by electrospray mass spectrometry. *J. Am. Soc. Mass Spectrom.* **14**, 1123–1129
4. Malvern Panalytical l Ltd is a Spectris company [online] https://www.malvernpanalytical.com (Accessed November 6, 2019)
5. Poupko, R., Zimmermann, H., Muller, K., and Luz, Z. (1996) Dynamic NMR investigation of the Cope rearrangement in solutions of monosubstituted bullvalenes. *J. Am. Chem. Soc.* **118**, 7995–8005
6. He, M., and Bode, J. W. (2013) E pluribus unum: isolation, structure determination, network analysis and DFT studies of a single metastable structure from a shapeshifting mixture of 852 bullvalene structural isomers. *Org. Biomol. Chem.* **11**, 1306–17
7. Schröder, G., and Oth, J. F. M. (1967) Recent Chemistry of Bullvalene. *Angew. Chemie Int. Ed. English*. **6**, 414–423
8. Gimarc, B. M., and Brant, A. R. (1994) Bullvalene: Reaction Graphs Relating Polysubstituted Positional Isomers. *J. Chem. Inf. Model.* **34**, 1167–1173
9. Ault, A. (2001) The Bullvalene story. The conception of Bullvalene, a molecule that has no permanent structure. *J. Chem. Educ.* **78**, 924
10. Bismillah, A. N., Chapin, B. M., Hussein, B. A., and McGonigal, P. R. (2020) Shapeshifting molecules: The story so far and the shape of things to come. *Chem. Sci.* **11**, 324–332
11. A. C. Gossard, D. K. Hindermann, M. B. Robin N. A. Kuebler, T. H. G. (1967) *Crystal Structure of Bullvalene at 25 C*, Murray Hill, New Jersy
12. Roth, K. (1989) X-Ray, N.M.R., and Theoretical Studies of the Structures of (Ethy1thio)bullvalene
13. Luger, P., Ruble, J. R., and Jeffreysl, G. A. (1986) Single Crystal Neutron Diffraction Refinement of Bullvalene at 110 K
14. Yahiaoui, O., Pašteka, L. F., Judeel, B., and Fallon, T. (2018) Synthesis and Analysis of

Substituted Bullvalenes. *Angew. Chemie - Int. Ed.* **57**, 2570–2574

15. Hoffmann, R., and Stohrer, W. (1971) The Cope Rearrangement Revisited. *J. of'the Am. Chem. Soc.* **2511**, 6941–6948
16. Yahiaoui, O., Pašteka, L. F., Blake, C. J., Newton, C. G., and Fallon, T. (2019) Network Analysis of Substituted Bullvalenes. *Org. Lett.* 10.1021/acs.orglett.9b03737
17. Yorozu, T., Hoshino, M., and Imamura, M. (1982) Fluorescence studies of pyrene inclusion complexes with. alpha.-,. beta.-, and. gamma.-cyclodextrins in aqueous solutions. Evidence for formation of pyrene dimer in. *J. Phys.*
18. Furuki, T., Hosokawa, F., and Sakurai, M. (1993) Microscopic Medium Effects on a Chemical Reaction. A Theoretical Study of Decarboxylation Catalyzed by Cyclodextrins as an Enzyme Model. *J. Am. Chem. Soc.* **115**, 2903–2911
19. Rajeswari, C., Alka, A., Javed, A., and Kha, R. (2005) Cyclodextrins in drug delivery: an updated review. *AAPS PharmSciTech*. **6**, 329–357
20. Gebhardt, J., and Hansen, N. (2015) Calculation of binding affinities for linear alcohols to α-cyclodextrin by twin-system enveloping distribution sampling simulations. *Fluid Phase Equilib.* **422**, 1–17
21. Del Valle, E. M. M. (2004) Cyclodextrins and their uses: A review. *Process Biochem.* **39**, 1033–1046
22. Nutho, B., Khuntawee, W., Rungnim, C., Pongsawasdi, P., Wolschann, P., Karpfen, A., Kungwan, N., and Rungrotmongkol, T. (2014) Binding mode and free energy prediction of fisetin/β-cyclodextrin inclusion complexes. *Beilstein J. Org. Chem.* **10**, 2789–2799
23. Sabadini, E., Cosgrove, T., and Egídio, F. D. C. (2006) Solubility of cyclomaltooligosaccharides (cyclodextrins) in H2O and D2O: A comparative study. *Carbohydr. Res.* **341**, 270–274
24. Chaplin, M. Water Structure and Science. [online] http://www1.lsbu.ac.uk/water/ (Accessed November 7, 2019)
25. Harada, A., Li, J., Suzuki, S., and Kamachi, M. (1993) Complex Formation between Polyisobutylene and Cyclodextrins: Inversion of Chain-Length Selectivity between β-Cyclodextrin and γ-Cyclodextrin. *Macromolecules*. **26**, 5267–5268
26. Vaidyanathan, K., Vasudevan, D., S, S., and Vaidyanathan, K. (2017) Subcellular Organelles and Cell Membranes. in *Textbook of Biochemistry for Medical Students*, pp. 1–1, Jaypee Brothers
27. Chasman, D. I. (ed.) (2003) *Protein Structure: Determination, Analysis, and Applications for Drug Discovery*, Marcel Dekker, Inc.

28. Gonzalez, M. W., and Kann, M. G. (2012) Chapter 4: Protein Interactions and Disease. *PLoS Comput. Biol.* 10.1371/journal.pcbi.1002819
29. Ryan, D. P., and Matthews, J. M. (2005) Protein-protein interactions in human disease. *Curr. Opin. Struct. Biol.* **15**, 441–446
30. Travers, A. (1993) *DNA-protein interactions*, Springer, London
31. Pabo, C. O., and Sauer, R. T. (1984) PROTEIN-DNA RECOGNITION. *Annu. Rev. Biochem.* **53**, 293–321
32. Dickerson, R. E. (1983) The DNA helix and how it is read. *Sci. Am.* **249**, 94–111
33. Bewley, C. A., Gronenborn, A. M., and Clore, G. M. (1998) Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 105–131
34. Medina-Franco, J. L., Méndez-Lucio, O., and Martinez-Mayorga, K. (2014) Chapter One - The Interplay Between Molecular Modeling and Chemoinformatics to Characterize Protein–Ligand and Protein–Protein Interactions Landscapes for Drug Discovery. in *Biomolecular Modelling and Simulations* (Karabencheva-Christova, T. B. T.-A. in P. C. and S. B. ed), pp. 1–37, Academic Press, **96**, 1–37
35. Dunn, M. F. (2010) Protein–Ligand Interactions: General Description. *eLS*.
36. Korn, E. D. (1968) Structure and Function of the Plasma Membrane : A biochemical perspective. *J. Gen. Physiol.* **52**, 257—278
37. Lee, A. G. (2004) How lipids affect the activities of integral membrane proteins. *Biochim. Biophys. Acta - Biomembr.* **1666**, 62–87
38. McIntosh, T. J., and Simon, S. A. (2007) Bilayers as protein solvents: role of bilayer structure and elastic properties. *J. Gen. Physiol.* **130**, 225—227
39. Phillips, R., Ursell, T., Wiggins, P., and Sens, P. (2009) Emerging roles for lipids in shaping membrane-protein function. *Nature*. **459**, 379–385
40. Martin, J., and Sawyer, A. (2019) Elucidating the structure of membrane proteins. *Biotechniques*. **66**, 167–170
41. Mulgrew-Nesbitt, A., Diraviyam, K., Wang, J., Singh, S., Murray, P., Li, Z., Rogers, L., Mirkovic, N., and Murray, D. (2006) The role of electrostatics in protein-membrane interactions. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids*. **1761**, 812–826
42. Wimley W C, and White S H (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **3**, 842–848
43. Adjei, A. A., and Hidalgo, M. (2005) Intracellular Signal Transduction Pathway Proteins As Targets for Cancer Therapy. *J. Clin. Oncol.* **23**, 5386–5403

44. Whitman, M., Downes, C. P., Keeler, M., Keller, T., and Cantley, L. (1988) Type I phosphatidylinositol kinase makes a novel inositol phospholipid, phosphatidylinositol-3-phosphate. *Nature*. **332**, 644

45. Mccafferty, E., Flanagan, J., and Shepherd, P. (2018) *Influencing Factors of Phosphatidylinositol 3-Kinase Membrane Binding and Catalysis*. Ph.D. thesis, The University of Auckland

46. Hara, K., Yonezawa, K., Sakaue, H., Ando, A., Kotani, K., Kitamura, T., Kitamura, Y., Ueda, H., Stephenst, L., Jacksont, T. R., Hawkinst, P. T., Dhand, R., Clark, A. E., Holman, G. D., Waterfieldt, M. D., and Kasuga, M. (1994) -Phosphatidylinositol 3-kinase activity is required for insulin- stimulated glucose transport but not for RAS activation in CHO cells. **91**, 7415–7419

47. Foster, F. M. (2003) The phosphoinositide (PI) 3-kinase family. *J. Cell Sci.* **116**, 3037–3040

48. Okkenhaug, K., Turner, M., and Gold, M. R. (2015) PI3K Signalling, Frontiers

49. Osaki, M., Oshimura, M., and Ito, H. (2004) PI3K-Akt pathway: Its functions and alterations in human cancer. *Apoptosis*. **9**, 667–676

50. Manning, B. D., and Cantley, L. C. (2007) AKT/PKB Signaling: Navigating Downstream. *Cell*. **129**, 1261–1274

51. Carpenter, C. L., and Cantley, L. C. (1990) Phosphoinositide kinases. *Biochemistry*. **29**, 11147–11156

52. Hiles, I. D., Otsu, M., Volinia, S., Fry, M. J., Gout, I., Dhand, R., Panayotou, G., Ruiz-Larrea, F., Thompson, A., Totty, N. F., Hsuan, J. J., Courtneidge, S. A., Parker, P. J., and Waterfield, M. D. (1992) Phosphatidylinositol 3-kinase: Structure and expression of the 110 kd catalytic subunit. *Cell*. **70**, 419–429

53. Hu, P., Mondino, A., Skolnik, E. Y., and Schlessinger, J. (1993) Cloning of a novel, ubiquitously expressed human phosphatidylinositol 3-kinase and identification of its binding site on p85. *Mol. Cell. Biol.* **13**, 7677–7688

54. Stoyanov, B., Volinia, S., Hanck, T., Rubio, I., Loubtchenkov, M., Malek, D., Stoyanova, S. ;, Vanhaesebroeck, B., Dhand, R., Nurnberg, B., and Al., E. (1995) Cloning and characterization of a G protein-activated human phosphoinositide-3 kinase. *Science (80).* **269**, 690–693

55. Vanhaesebroeck, B., Welham, M. J., Kotani, K., Stein, R., Warne, P. H., Zvelebil, M. J., Higashi, K., Volinia, S., Downward, J., and Waterfield, M. D. (1997) P110δ, a Novel Phosphoinositide 3-Kinase in Leukocytes. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 4330–4335

56. Hawkins, P. T., Anderson, K. E., Davidson, K., Stephens, L. R., and Hawkins, P. (2006) Signalling through Class I PI3Ks in mammalian cells. *Biochem. Soc. Trans.*

57. Mario Amzel, L., and All, E. (2008) Structural Comparisons of Class I Phosphoinositide 3-kinases. *Nat Rev Cancer*. **8**, 665–669

58. Huang, C. H., Mandelker, D., Schmidt-Kittler, O., Samuels, Y., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Gabelli, S. B., and Amzel, L. M. (2007) The structure of a human p110α/p85α complex elucidates the effects of oncogenic PI3Kα mutations. *Science (80-.).* **318**, 1744–1748

59. Fruman, D. A. (2011) Regulatory Subunits of Class IA PI3K BT - Phosphoinositide 3-kinase in Health and Disease: Volume 1 (Rommel, C., Vanhaesebroeck, B., and Vogt, P. K. eds), pp. 225–244, Springer Berlin Heidelberg, Berlin, Heidelberg

60. Jean, S., and Kiger, A. A. (2014) Classes of phosphoinositide 3-kinases at a glance. *J. Cell Sci.* **127**, 923–928

61. Stephens, L. R., Eguinoa, A., Erdjument-Bromage, H., Lui, M., Cooke, F., Coadwell, J., Smrcka, A. S., Thelen, M., Cadwallader, K., Tempst, P., and Hawkins, P. T. (1997) The Gβγ Sensitivity of a PI3K Is Dependent upon a Tightly Associated Adaptor, p101. *Cell*. **89**, 105–114

62. Suire, S., Coadwell, J., Ferguson, G. J., Davidson, K., Hawkins, P., and Stephens, L. (2005) p84, a New Gβγ-Activated Regulatory Subunit of the Type IB Phosphoinositide 3-Kinase p110γ. *Curr. Biol.* **15**, 566–570

63. Ponting, C. P. (1996) Novel domains in NADPH oxidase subunits, sorting nexins, and PtdIns 3-kinases: Binding partners of SH3 domains? *Protein Sci.* **5**, 2353–2357

64. Rizo, J., and Sudhof, T. C. (1998) C2-domains, structure and function of a universal Ca2+-binding domain. *J. Biol. Chem.* **273**, 15879–15882

65. Lindmo, K. (2006) Regulation of membrane traffic by phosphoinositide 3-kinases. *J. Cell Sci.* **119**, 605–614

66. Backer, J. M. (2010) The Regulation of Class IA PI 3-Kinases by Inter-Subunit Interactions. in Phosphoinositide 3-kinase in Health and Disease. Current Topics in Microbiology and Immunology (Rommel C., Vanhaesebroeck B., V. P. ed), Springer, Berlin, Heidelberg

67. Pacold, M. E., Suire, S., Perisic, O., Lara-Gonzalez, S., Davis, C. T., Walker, E. H., Hawkins, P. T., Stephens, L., Eccleston, J. F., and Williams, R. L. (2000) Crystal Structure and Functional Analysis of Ras Binding to Its Effector Phosphoinositide 3-Kinase γ. *Cell*. **103**, 931–944

68. Walser, R., Burke, J. E., Gogvadze, E., Bohnacker, T., Zhang, X., Hess, D., Küenzi, P., Leitges, M., Hirsch, E., Williams, R. L., Laffargue, M., and Wymann, M. P. (2013) PKCβ Phosphorylates PI3Kγ to Activate It and Release It from GPCR Control. *PLOS Biol.* **11**, e1001587

69. Jiménez, C. ;, Hernández, C. ;, Pimentel, B. ;, and Carrera, A. C. (2002) The p85 regulatory subunit controls sequential activation of phosphoinositide 3- kinase by Tyr kinases and Ras. *J. Biol. Chem.* **277**, 41556–41562

70. Carpenter, C. L., Auger, K. R., Chanudhuri, M., Yoakim, M., Schaffhausen, B., Shoelson, S., and Cantley, L. C. (1993) Phosphoinositide 3-kinase is activated by phosphopeptides that bind to the SH2 domains of the 85-kDa subunit. *J. Biol. Chem.* **268**, 9478–9483

71. Burke, J. E., Vadas, O., Berndt, A., Finegan, T., Perisic, O., and Williams, R. L. (2011) Dynamics of the phosphoinositide 3-kinase p110δ interaction with p85α and membranes reveals aspects of regulation distinct from p110α. *Structure*. **19**, 1127–1137

72. Yu, J., Wjasow, C., and Backer, J. M. (1998) Regulation of the p85/p110a Phosphatidylinositol 3 ' -Kinase. **273**, 30199–30203

73. Vadas, O., Burke, J. E., Zhang, X., Berndt, A., and Williams, R. L. (2011) Structural biology structural basis for activation and inhibition of class I phosphoinositide 3-kinases. *Sci. Signal.* 10.1126/scisignal.2002165

74. Rodriguez-Viciana, P., Warne, P. H., Vanhaesebroeck, B., Waterfield, M. D., and Downward, J. (1996) Activation of phosphoinositide 3-kinase by interaction with Ras and by point mutation. *Eur. Mol. Biol. Organ. J.* **15**, 2442–2451

75. Buckles, T. C., Ziemba, B. P., Masson, G. R., Williams, R. L., and Falke, J. J. (2017) Single-Molecule Study Reveals How Receptor and Ras Synergistically Activate PI3Kα and PIP3 Signaling. *Biophys. J.* **113**, 2396–2405

76. Siempelkamp, B. D., Rathinaswamy, M. K., Jenkins, M. L., and Burke, J. E. (2017) Molecular mechanism of activation of class IA phosphoinositide 3-kinases (PI3Ks) by membrane-localized HRas. *J. Biol. Chem.* **292**, 12256–12266

77. Rommel, C., Vanhaesebroeck, B., and Vogt, P. K. . (eds.) (2015) *Phosphoinositide 3-kinases in health and disease*

78. Maier, U., Babich, A., and Nürnberg, B. (1999) Roles of non-catalytic subunits in Gβγ-induced activation of class I phosphoinositide 3-kinase isoforms β and γ. *J. Biol. Chem.* **274**, 29311–29317

79. Hon, W. C., Berndt, A., and Williams, R. L. (2012) Regulation of lipid binding underlies

the activation mechanism of class IA PI3-kinases. *Oncogene*. **31**, 3655–3666

80. Irvine, W. A. (2017) *Molecular Dynamics Simulations of Protein-Membrane Interactions Focusing on PI3K α and Its Oncogenic Mutants*. Ph.D. thesis, Massey University
81. Pirola, L., Zvelebil, M. J., Bulgarelli-leva, G., Obberghen, E. Van, Waterfield, M. D., and Wymann, M. P. (2001) Activation Loop Sequences Confer Substrate Specificity to. *Biochemistry*. **276**, 21544–21554
82. Irvine, W. A., Flanagan, J. U., and Allison, J. R. (2018) Computational prediction of amino acids governing protein-membrane interaction for the PIP3 cell signalling system. *Structure*. 10.1016/j.str.2018.10.014
83. Maheshwari, S., Miller, M. S., O'Meally, R., Cole, R. N., Amzel, L. M., and Gabelli, S. B. (2017) Kinetic and structural analyses reveal residues in phosphoinositide 3-kinase α that are critical for catalysis and substrate recognition. *J. Biol. Chem.* **292**, 13541–13550
84. Hawkins, P. T., and Stephens, L. R. (2015) PI3K signalling in inflammation. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids*. **1851**, 882–897
85. Mayer, I. A., and Arteaga, C. L. (2016) The PI3K/AKT Pathway as a Target for Cancer Treatment. *Annu. Rev. Med.* **67**, 11–28
86. Mukohara, T. (2015) Pi3K mutations in breast cancer: prognostic and therapeutic implications. *Dovpress Breast Cancer Targets Ther.* 10.2147/BCTT.S60696
87. Falasca, M., Logan, S. K., Lehto, V. P., Baccante, G., Lemmon, M. A., and Schlessinger, J. (1998) Activation of phospholipase Cγ by PI 3-kinase-induced PH domain-mediated membrane targeting. *EMBO J.* **17**, 414–422
88. Mohamed, A. J., and et al. (2009) Bruton's tyrosine kinase (Btk): function, regulation, and transformation with special emphasis on the PH domain. *Immunol. Rev.* **228**, 58–73
89. Salim1, K., Bottomley2, M. J., Querfurthl, E., Zvelebill, M. J., Gout1, I., Scaife3, R., Margolis3, R. L., Gigg4 ', R., Smith6, C. L. E., Driscoll2, P. C., Waterfieldl ', M. D., and Panayotou1 ', G. (1996) Distinct specificity in the recognition of phosphoinositides by the pleckstrin homology domains of dynamin and Bruton's tyrosine kinase. *EMBO J.* **1522**, 6241–6250
90. Di Paolo, G., and De Camilli, P. (2006) Phosphoinositides in cell regulation and membrane dynamics. *Nature*. **443**, 651–657
91. Vanhaesebroeck, B., Stephens, L., and Hawkins, P. (2012) PI3K signalling: The path to discovery and understanding. *Nat. Rev. Mol. Cell Biol.* **13**, 195–203
92. Malcolm, C., E., A. W., Carol, S., Bart, V., and R., T. E. (2004) Glucose-Potentiated

Chemotaxis in Human Vascular Smooth Muscle Is Dependent on Cross-Talk Between the PI3K and MAPK Signaling Pathways. *Circ. Res.* **95**, 380–388

93. Oda, K., Stokoe, D., Taketani, Y., and McCormick, F. (2005) High Frequency of Coexistent Mutations of PIK3CA and PTEN Genes in Endometrial Carcinoma. *Cancer Res.* **65**, 10669 LP – 10673
94. Samuels, Y., and Velculescu, V. E. (2004) Oncogenic Mutations of PIK3CA in Human Cancers. *Cell Cycle*. **3**, 1221–1224
95. Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U., and Campbell, P. J. (2015) COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811
96. Bader, A. G., Kang, S., and Vogt, P. K. (2006) Cancer-specific mutations in PIK3CA are oncogenic in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 1475–1479
97. Gkeka, P., Evangelidis, T., Pavlaki, M., Lazani, V., Christoforidis, S., Agianian, B., and Cournia, Z. (2014) Investigating the Structure and Dynamics of the PIK3CA Wild-Type and H1047R Oncogenic Mutant. *PLoS Comput. Biol.*
98. Irvine, W. A., Flanagan, J. U., and Allison, J. R. (2019) Computational Prediction of Amino Acids Governing Protein-Membrane Interaction for the PIP 3 Cell Signaling System. *Structure*. **27**, 371-380.e3
99. Tsuchida, N., Murugan, A. K., and Grieco, M. (2016) Kirsten Ras* oncogene: Significance of its discovery in human cancer research. *Oncotarget*. **7**, 46717–46733
100. Murugan, A. K., Grieco, M., and Tsuchida, N. (2019) RAS mutations in human cancers: Roles in precision medicine. *Semin. Cancer Biol.*
101. Cox, A. D., and Der, C. J. (2010) Ras history: The saga continues. *Small GTPases*. **1**, 2–27
102. Wittinghofer, A. (ed.) (2014) *Ras superfamily small G proteins: Biology and mechanisms 1: General features, signaling*
103. Huang, L., Hofer, F., Martin, G. S., and Kim, S.-H. (1998) Structural basis for the interaction of Ras with RaIGDS. *Nat. Struct. Biol.* **5**, 422–426
104. Nassar, N., Horn, G., Herrmann, C. A., Scherer, A., McCormick, F., and Wittinghofer, A. (1995) The 2.2 Å crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with RaplA and a GTP analogue. *Nature*. **375**, 554–560
105. Vetter, I. R., and Wittinghofer, A. (2001) The Guanine Nucleotide-Binding Switch in Three Dimensions. *Science (80-.).* **294**, 1299 LP – 1304

106. Colicelli, J. (2004) Human RAS superfamily proteins and related GTPases. *Sci. STKE*. **250**
107. Laude, A. J., and Prior, I. A. (2008) Palmitoylation and localisation of RAS isoforms are modulated by the hypervariable linker domain. *J. Cell Sci.* **121**, 421–427
108. World Health Organization (2017) *Global Tuberculosis Report*
109. Talaat, A. M., Lyons, R., Howard, S. T., and Johnston, S. A. (2004) The temporal expression profile of Mycobacterium tuberculosis infection in mice. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4602 – 4607
110. NZ Ministry of Health Tuberculosis disease
111. Bentrup, K. H. zu, and Russell, D. G. (2001) Mycobacterial persistence: Adaptation to a changing environment. *Trends Microbiol.* **9**, 597–605
112. Cambier, C. J., Falkow, S., and Ramakrishnan, L. (2014) Host evasion and exploitation schemes of Mycobacterium tuberculosis. *Cell*. **159**, 1497–1509
113. Liu, Y., Tan, S., Huang, L., Abramovitch, R. B., Rohde, K. H., Zimmerman, M. D., Chen, C., Dartois, V., VanderVen, B. C., and Russell, D. G. (2016) Immune activation of the host cell induces drug tolerance in Mycobacterium tuberculosis both in vitro and in vivo. *J. Exp. Med.* **213**, 809 LP – 825
114. Rodríguez, J. G., Hernández, A. C., Helguera-Repetto, C., Aguilar Ayala, D., Guadarrama-Medina, R., Anzóla, J. M., Bustos, J. R., Zambrano, M. M., González-y-Merchand, J., García, M. J., and Del Portillo, P. (2014) Global Adaptation to a Lipid Environment Triggers the Dormancy-Related Phenotype of Mycobacterium tuberculosis. *MBio*. **5**, e01125-14
115. Lee, W., VanderVen, B. C., Fahey, R. J., and Russell, D. G. (2013) Intracellular Mycobacterium tuberculosis exploits host-derived fatty acids to limit metabolic stress. *J. Biol. Chem.* **288**, 6788–6800
116. VanderVen, B. C., Fahey, R. J., Lee, W., Liu, Y., Abramovitch, R. B., Memmott, C., Crowe, A. M., Eltis, L. D., Perola, E., Deininger, D. D., Wang, T., Locher, C. P., and Russell, D. G. (2015) Novel Inhibitors of Cholesterol Degradation in Mycobacterium tuberculosis Reveal How the Bacterium's Metabolism Is Constrained by the Intracellular Environment. *PLOS Pathog.* **11**, e1004679
117. Pandey, A. K., and Sassetti, C. M. (2008) Mycobacterial persistence requires the utilization of host cholesterol. *Proc. Natl. Acad. Sci.* **105**, 4376 LP – 4380
118. Nesbitt, N. M., Yang, X., Fontán, P., Kolesnikova, I., Smith, I., Sampson, N. S., and Dubnau, E. (2010) A thiolase of Mycobacterium tuberculosis is required for virulence

and production of androstenedione and androstadienedione from cholesterol. *Infect. Immun.* **78**, 275–282

119. Lack, N. A., Kawamura, A., Fullam, E., Laurieri, N., Beard, S., Russell, A. J., Evangelopoulos, D., Westwood, I., and Sim, E. (2009) Temperature stability of proteins essential for the intracellular survival of Mycobacterium tuberculosis. *Biochem. J.* **418**, 369–378
120. Upton, A. M., and McKinney, J. D. (2007) Role of the methylcitrate cycle in propionate metabolism and detoxification in Mycobacterium smegmatis. *Microbiology*. **153**, 3973–3982
121. Parrish, N. M., Dick, J. D., and Bishai, W. R. (1998) Mechanisms of latency in Mycobacterium tuberculosis. *Trends Microbiol.* **6**, 107–112
122. Orth, P., Schnappinger, D., Hillen, W., Saenger, W., and Hinrichs, W. (2000) Structural basis of gene regulation by the tetracycline inducible Tet repressor-operator system. *Nat. Struct. Biol.* **7**, 215–219
123. Haberl, F., Lanig, H., and Clark, T. (2009) Induction of the tetracycline repressor: Characterization by molecular-dynamics simulations. *Proteins Struct. Funct. Bioinforma.* **77**, 857–866
124. Balhana, R. J. C., Singla, A., Sikder, M. H., Withers, M., and Kendall, S. L. (2015) Global analyses of TetR family transcriptional regulators in mycobacteria indicates conservation across species and diversity in regulated functions. *BMC Genomics*. **16**, 1–12
125. Le, T. B. K. ;, Schumacher, M. A. ., Lawson, D. M. ., and Brennan, R. G.;Buttner, M. J. (2011) The crystal structure of the TetR family transcriptional repressor SimR bound to DNA and the role of a flexible N-terminal extension in minor groove binding Tung. *Nucleic Acids Res.* **39**, 9433–9447
126. Le, T. B. K., Stevenson, C. E. M., Fiedler, H. P., Maxwell, A., Lawson, D. M., and Buttner, M. J. (2011) Structures of the tetr-like simocyclinone efflux pump repressor, SimR, and the mechanism of ligand-mediated derepression. *J. Mol. Biol.* **408**, 40–56
127. Cuthbertson, L., and Nodwell, J. R. (2013) The TetR Family of Regulators. *Microbiol. Mol. Biol. Rev.* **77**, 440–475
128. Crowe, A. M., Stogios, P. J., Casabon, I., Evdokimova, E., Savchenko, A., and Eltis, L. D. (2015) Structural and functional characterization of a ketosteroid transcriptional regulator of Mycobacterium tuberculosis. *J. Biol. Chem.* **290**, 872–882
129. Aldred, K. J., Kerns, R. J., and Osheroff, N. (2014) Mechanism of Quinolone Action

and Resistance. *Biochemistry*. **53**, 1565–1574

130. Tsuchiya, M., Selvarajoo, K., Piras, V., Tomita, M., and Giuliani, A. (2009) Local and Global responses in complex gene regulation networks. *Phys. A Stat. Theor. Phys.* **388**, 1738–1746
131. Jiang, J., Sun, X., Wu, W., Li, L., Wu, H., Zhang, L., Yu, G., and Li, Y. (2016) Construction and application of a co-expression network in Mycobacterium tuberculosis. *Sci. Rep.* **6**, 28422
132. Kendall, S. L., Withers, M., Soffair, C. N., Moreland, N. J., Gurcha, S., Sidders, B., Frita, R., Ten Bokum, A., Besra, G. S., Lott, J. S., and Stoker, N. G. (2007) A highly conserved transcriptional repressor controls a large regulon involved in lipid degradation in Mycobacterium smegmatis and Mycobacterium tuberculosis. *Mol. Microbiol.* **65**, 684–699
133. Ramos, J. L., Martı, M., Molina-henares, A. J., Tera, W., Brennan, R., and Tobes, R. (2005) TetR family of transcriptional. **69**, 1–31
134. Frénois, F., Engohang-Ndong, J., Locht, C., Baulard, A. R., and Villeret, V. (2004) Structure of EthR in a ligand bound conformation reveals therapeutic perspectives against tuberculosis. *Mol. Cell*. **16**, 301–307
135. Ho, N. A. T. (2017) *Structural basis for the function of the transcriptional regulator KstR from Mycobacterium tuberculosis*. Ph.D. thesis, University of Auckland
136. Ouellet, H., Johnston, J. B., and Montellano, P. R. O. de (2011) Cholesterol catabolism as a therapeutic target in Mycobacterium tuberculosis. *Trends Microbiol.* **19**, 530–539
137. Crowe, A. M., Stogios, P. J., Casabon, I., Evdokimova, E., Savchenko, A., and Eltis, L. D. (2015) Structural and functional characterization of a ketosteroid transcriptional regulator of Mycobacterium tuberculosis. *J. Biol. Chem.* **290**, 872–882
138. Ho, N. A. T., Dawes, S. S., Crowe, A. M., Casabon, I., Gao, C., Kendall, S. L., Baker, E. N., Eltis, L. D., and Lott, J. S. (2016) The structure of the transcriptional repressor KstR in complex with coa thioester cholesterol metabolites sheds light on the regulation of cholesterol catabolism in Mycobacterium tuberculosis. *J. Biol. Chem.* **291**, 7256–7266
139. Kendall, S. L., Burgess, P., Balhana, R., Withers, M., Ten Bokum, A., Lott, J. S., Gao, C., Uhia-Castro, I., and Stoker, N. G. (2010) Cholesterol utilization in mycobacteria is controlled by two TetR-type transcriptional regulators: kstR and kstR2. *Microbiology*. **156**, 1362–1371
140. Gomez, R. L., Jose, L., Ramachandran, R., Raghunandanan, S., Muralikrishnan, B.,

Johnson, J. B., Sivakumar, K. C., Mundayoor, S., and Kumar, R. A. (2016) The multiple stress responsive transcriptional regulator Rv3334 of Mycobacterium tuberculosis is an autorepressor and a positive regulator of kstR. *FEBS J.* **283**, 3056–3071

141. Ramos, J. L., Martínez-bueno, M., Antonio, J., Terán, W., Watanabe, K., Gallegos, M. T., Brennan, R., Ramos, J. L., Martı, M., Molina-henares, A. J., and Tera, W. (2005) The TetR family of transcriptional Repressors. *Microbiol Mol Biol Rev*. **69**, 326–356
142. Ho, N. A. T., Dawes, S. S. ., Allison, J. R. ., Baker, E. N. ., and Lott, J. S. The structure of KstR from Mycobacterium tuberculosis in complex with its target DNA demonstrates the role of indirect reading in DNA sequence recognition by a TetR-family transcriptional repressor. (In preparation)
143. Razzak, Ali; Lott, Shaun J.; Allison, J. R. (2018) *The Conformational Dynamics and Molecular Mechanism of KstR* . Ph.D. thesis, The University of Auckland
144. Willand, N., Dirié, B., Carette, X., Bifani, P., Singhal, A., Desroses, M., Leroux, F., Willery, E., Mathys, V., Déprez-Poulain, R., Delcroix, G., Frénois, F., Aumercier, M., Locht, C., Villeret, V., Déprez, B., and Baulard, A. R. (2009) Synthetic EthR inhibitors boost antituberculous activity of ethionamide. *Nat. Med.* **15**, 537–544
145. Capyk, J. K., Casabon, I., Gruninger, R., Strynadka, N. C., and Eltis, L. D. (2011) Activity of 3-ketosteroid 9α-hydroxylase (KshAB) indicates cholesterol side chain and ring degradation occur simultaneously in Mycobacterium tuberculosis. *J. Biol. Chem.* **286**, 40717–40724
146. Yang, M., Guja, K. E., Sampson, N. S., Wipperman, M. F., Lu, R., Garcia-Diaz, M., St. Clair, J. R., and Bonds, A. C. (2015) Unraveling Cholesterol Catabolism in Mycobacterium tuberculosis : ChsE4-ChsE5 α 2 β 2 Acyl-CoA Dehydrogenase Initiates β-Oxidation of 3-Oxo-cholest-4-en-26-oyl CoA. *ACS Infect. Dis.* **1**, 110–125
147. Nikiforov, P. O., Surade, S., Blaszczyk, M., Delorme, V., Brodin, P., Baulard, A. R., Blundell, T. L., and Abell, C. (2016) A fragment merging approach towards the development of small molecule inhibitors of Mycobacterium tuberculosis EthR for use as ethionamide boosters. *Org. Biomol. Chem.* **14**, 2318–2326
148. Reichheld, S. E., Yu, Z., and Davidson, A. R. (2009) The induction of folding cooperativity by ligand binding drives the allosteric response of tetracycline repressor. *Proc. Natl. Acad. Sci.* **106**, 22263–22268
149. Csermely, P., Palotai, R., and Nussinov, R. (2010) Induced fit, conformational selection and independent dynamic segments: An extended view of binding events. *Trends Biochem. Sci.* **35**, 539–546

150. Vogt, A. D., and Di Cera, E. (2012) Conformational selection or induced fit? A critical appraisal of the kinetic mechanism. *Biochemistry*. **51**, 5894–5902
151. Vogt, A. D., Pozzi, N., Chen, Z., and Di Cera, E. (2014) Essential role of conformational selection in ligand binding. *Biophys. Chem.* **186**, 13–21
152. Sassetti, C. M., and Rubin, E. J. (2003) Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12989–12994
153. Casabon, I., Zhu, S. H., Otani, H., Liu, J., Mohn, W. W., and Eltis, L. D. (2013) Regulation of the KstR2 regulon of Mycobacterium tuberculosis by a cholesterol catabolite. *Mol. Microbiol.* **89**, 1201–1212
154. Casabon, I., Crowe, A. M., Liu, J., and Eltis, L. D. (2013) FadD3 is an acyl-CoA synthetase that initiates catabolism of cholesterol rings C and D in actinobacteria. *Mol. Microbiol.* **87**, 269–283
155. Warnke, M., Jacoby, C., Jung, T., Agne, M., Mergelsberg, M., Starke, R., Jehmlich, N., von Bergen, M., Richnow, H. H., Brüls, T., and Boll, M. (2017) A patchwork pathway for oxygenase-independent degradation of side chain containing steroids. *Environ. Microbiol.* **19**, 4684–4699
156. Chang, C., Evdokimova, E., Kagan, O., Savchenko, A., Edwards, A. M., and Joachimiak, A. Crystal structure of Probable transcriptional regulatory protein RHA5900 (To be published)
157. Dawes, S. S., Keown, J. R., Clow, F., Kendall, S. L., Baker, E. N., and Lott, J. S. Paired Arginine Residues Control a Conformational Selection Mechanism of DNA Binding by KstR2, a TetR-Family Repressor Required for Complete Cholesterol Degradation in Mycobacterium tuberculosis. *To be Publ.*
158. Stierand, K., and Rarey, M. (2010) Drawing the PDB: Protein−Ligand Complexes in Two Dimensions. *ACS Med. Chem. Lett.* **1**, 540–545
159. Durrant, J. D; McCammon, J. A. (2011) Molecular dynamics simulations and drug discovery. *BMC Biol.* **9**, 1741–7007
160. Adcock, S. A., and McCammon, J. A. (2006) Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev.* **106**, 1589–1615
161. Gkeka, P., Papafotika, A., Christoforidis, S., and Cournia, Z. (2015) Exploring a Non-ATP pocket for potential allosteric modulation of pi3kα. *J. Phys. Chem. B.* **119**, 1002–1016
162. Gao, Y., Ma, Y., Yang, G., and Li, Y. (2016) Molecular dynamics simulations to investigate the binding mode of the natural product liphagal with phosphoinositide 3-

Kinase α. *Molecules*. **21**, 1–12

163. Han, M., and Zhang, J. Z. H. (2010) Class I Phospho-inositide-3-kinases (PI3Ks) Isoform-Specific Inhibition Study by the combination of docking and molecular dynamics simulation.
164. Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., Van Der Spoel, D., Hess, B., and Lindahl, E. (2013) GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*. **29**, 845–854
165. Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009) CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614
166. Groom, C. R., Bruno, I. J., Lightfoot, M. P., and Ward, S. C. (2016) The Cambridge Structural Database. *Acta Crystallogr. Sect. B*. **72**, 171–179
167. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242
168. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690
169. Evans, D. J., and Holian, B. L. (1985) The Nose–Hoover thermostat. *J. Chem. Phys.* **83**, 4069–4074
170. Martyna, G. J., Klein, M. L., and Tuckerman, M. (1992) Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* **97**, 2635–2643
171. Zhao, Y. (2011) *Brief introduction to the thermostats*
172. Collier, T. (2016) *Fully Atomistic Modelling of Collagen Cross- - -linking*. Ph.D. thesis, University College London
173. Schmid, N., Eichenberger, A. P., Choutko, A., Riniker, S., Winger, M., Mark, A. E., and van Gunsteren, W. F. (2011) Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.* **40**, 843–856
174. Lim, T.-C. (2003) Mathematical connections between bond-stretching potential

functions. *J. Math. Chem.* **33**, 29–37

175. Mayo, S. L., Olafson, B. D., and Goddard, W. A. (1990) DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909
176. Barker, J. A., and Watts, R. O. (1973) Monte carlo studies of the dielectric properties of water-like models. *Mol. Phys.* **26**, 789–792
177. Watts, R. O. (1974) Monte Carlo studies of liquid water. *Mol. Phys.* **28**, 1069–1083
178. Ewald, P. P. (1921) Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.* **369**, 253–287
179. Darden, T., York, D., and Pedersen, L. (1993) Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092
180. Gromacs User guide (2019) https://doi.org/10.5281/zenodo.3460414
181. Gromacs development team (2019) Gromacs manual. https://doi.org/10.5281/zenodo.2424486
182. Pohorille, A., Jarzynski, C., and Chipot, C. (2010) Good practices in free-energy calculations. *J. Phys. Chem. B.* **114**, 10235–10253
183. Martins, S. A., Sousa, S. F., Ramos, M. J., and Fernandes, P. A. (2014) Prediction of solvation free energies with thermodynamic integration using the general Amber force field. *J. Chem. Theory Comput.* **10**, 3570–3577
184. Liu, Y. (2009) Is the free energy change of adsorption correctly calculated? *J. Chem. Eng. Data*. **54**, 1981–1985
185. Deng, Y., and Roux, B. (2009) Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B*
186. Crooks, G. E. (1999) The Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences. **60**, 2721–2726
187. Mitchell, M. J., and McCammon, J. A. (1991) Free energy difference calculations by thermodynamic integration: Difficulties in obtaining a precise value. *J. Comput. Chem.* **12**, 271–275
188. Jorgensen, W. L., Buckner, J. K., Boudon, S., and Tirado-Rives, J. (1988) Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. *J. Chem. Phys.* **89**, 3742
189. Villa, A., and Mark, A. E. (2002) Calculation of the free energy of solvation for neutral analogs of amino acid side chains. *J. Comput. Chem.* **23**, 548–553
190. Straatsma, T. P., and Berendsen, H. J. C. (1988) Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by

molecular dynamics simulations. *J. Chem. Phys.* **89**, 5876–5886

191. Leyva, E., Moctezuma, E., Strouse, J., and García-Garibay, M. A. (2001) Spectrometric and 2D NMR studies on the complexation of chlorophenols with cyclodextrins. *J. Incl. Phenom.* **39**, 41–46
192. Hansen, N., and Van Gunsteren, W. F. (2014) Practical aspects of free-energy calculations: A review. *J. Chem. Theory Comput.* **10**, 2632–2647
193. Markthaler, D., Gebhardt, J., Jakobtorweihen, S., and Hansen, N. (2017) Calculation of thermodynamic properties for the system alpha-cyclodextrin / alcohol in aqueous solution using molecular simulations. *Chemie Ingenieur Technik.* **89**, (10), 1306-1314
194. Malde, A. K., Zuo, L., Breeze, M., Stroet, M., Poger, D., Nair, P. C., Oostenbrink, C., and Mark, A. E. (2011) An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J. Chem. Theory Comput.* **7**, 4026–4037
195. Canzar, S., El-Kebir, M., Pool, R., Elbassioni, K., Malde, A. K., Mark, A. E., Geerke, D. P., Stougie, L., and Klau, G. W. (2013) Charge Group Partitioning in Biomolecular Simulation. *J. Comput. Biol.* **20**, 188–198
196. Koziara, K. B., Stroet, M., Malde, A. K., and Mark, A. E. (2014) Testing and validation of the Automated Topology Builder (ATB) version 2.0: prediction of hydration free enthalpies. *J. Comput. Aided. Mol. Des.* **28**, 221–233
197. Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38
198. Dennington, Roy; Keith, Todd A.; Millam, J. M. (2016) GaussView, Version 5
199. Berendsen, H. J. C., van der Spoel, D., and van Drunen, R. (1995) GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56
200. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005) GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718
201. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. **1**–**2**, 19–25
202. Turner, P. J., Stambulchik, E., Winter, A., Engelson, V., and Kailasanathan, N. Grace-5.1.22/QtGrace v 0.2.6.
203. Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. van, and Hermans, J. (1981) Interaction Models for Water in Relation to Protein Hydration, in Intermolecular Forces. in *Intermolecular Forces* (Pullman, B. ed), pp. 331–342, Springer, Netherlands

204. Darden, T., York, D., and Pedersen, L. (1993) Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092
205. Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997) LINCS: A Linear Constraint Solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472
206. Páll, S., and Hess, B. (2013) A flexible algorithm for calculating pair interactions on SIMD architectures. *Comput. Phys. Commun.* **184**, 2641–2650
207. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593
208. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratch, J. B. (2016) Gaussian 09, Revision A02
209. Chong, D. P. (2011) Density functional theory study of the electron spectra of formamide vapor. *J. Electron Spectros. Relat. Phenomena*. **184**, 164–169
210. McLean, A. D., and Chandler, G. S. (1980) Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11–18. *J. Chem. Phys.* **72**, 5639–5648
211. Krishnan, R., Binkley, J. S., Seeger, R., and Pople, J. A. (1980) Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.* **72**, 650–654
212. Pasťeka, L., Yahiaoui, O., Blake, C. J., Newton, C. G., and Fallon, T. (2019) Network Analysis of Substituted Bullvalenes. *Org. Lett.* **21**, (23), 9574–9578
213. Ho, N. A. T., Dawes, S. S., Crowe, A. M., Casabon, I., Gao, C., Kendall, S. L., Baker, E. N., Eltis, L. D., and Lott, J. S. (2016) The structure of the transcriptional repressor KstR in complex with coa thioester cholesterol metabolites sheds light on the regulation of cholesterol catabolism in Mycobacterium tuberculosis. *J. Biol. Chem.* **291**, 7256–7266
214. Uhía, I., Galán, B., Medrano, F. J., and García, J. L. (2011) Characterization of the KstR-dependent promoter of the gene for the first step of the cholesterol degradative pathway in Mycobacterium smegmatis. *Microbiology*. **157**, 2670–2680
215. Aleksandrov, A., Schuldt, L., Hinrichs, W., and Simonson, T. (2009) Tetracycline-Tet repressor binding specificity: Insights from experiments and simulations. *Biophys. J.* **97**, 2829–2838
216. Huang, J., and MacKerell Jr, A. D. (2013) CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data Jing. *J Comput Chem.* **34**, 2135–

2145

217. Yu, W., He, X., Vanommeslaeghe, K., and MacKerell Jr, A. D. (2012) Extension of the CHARMM General Force Field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J. Comput. Chem.* **33**, 2451–2468
218. Vanommeslaeghe, K., Raman, E. P., and MacKerell, A. D. (2012) Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **52**, 3155–3168
219. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935
220. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612
221. Schrödinger, L. The PyMOL Molecular Graphics System, Version˜1.5.
222. Sethi, A., Eargle, J., Black, A. A., and Luthey-Schulten, Z. (2009) Dynamical networks in tRNA:protein complexes. *Proc. Natl. Acad. Sci.* **106**, 6620–6625
223. Girvan, M., and Newman, M. E. J. (2002) Community structure in social and biological networks. **99**, 7821–7826
224. Hunter, J. D. (2007) Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95
225. Developer, V. M. D., Stone, J., Eargle, J., and Eargle, J. (2012) Dynamical Network Analysis
226. Pednekar, D., Tendulkar, A., and Durani, S. (2009) Electrostatics-defying interaction between arginine termini as a thermodynamic driving force in protein–protein interaction. *Proteins Struct. Funct. Bioinforma.* **74**, 155–163
227. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612
228. McKee, T., and McKee, J. R. (2017) Amino Acids, Peptides, and Proteins. in *Biochemistry : the molecular basis of life*, Sixth, New York : Oxford University Press, 10.1002/9781118860588.ch3
229. Eargle, J., and Luthey-Schulten, Z. (2012) NetworkView: 3D display and analysis of protein{middle dot}RNA interaction networks. *Bioinformatics*. **28**, 3000–3001
230. Williams, R., Berndt, A., Miller, S., Hon, W.-C., and Zhang, X. (2009) Form and

flexibility in phosphoinositide 3-kinases. *Biochem. Soc. Trans.* **37**, 615–626

231. Cleary, J. M., and Shapiro, G. I. (2010) Development of phosphoinositide-3 kinase pathway inhibitors for advanced cancer. *Curr. Oncol. Rep.* **12**, 87–94
232. Goodsell, D. S. (1999) The Molecular Perspective: The ras Oncogene. *Oncol. 1999;4263-264*. **4**, 263–264
233. Rodriguez-Viciana, P., Warne, P. H., Dhand, R., Vanhaesebroeck, B., Gout, I., Fry, M. J., Waterfield, M. D., and Downward, J. (1994) Phosphatidylinositol-3-OH kinase direct target of Ras. *Nature*. **370**, 527–532
234. Siempelkamp, B. D., Rathinaswamy, M. K., Jenkins, M. L., and Burke, J. E. (2017) Molecular mechanism of activation of class IA phosphoinositide 3-kinases (PI3Ks) by membrane-localized HRas. *J. Biol. Chem.* **292**, 12256–12266
235. Balla, T. (2013) Phosphoinositides: Tiny Lipids With Giant Impact on Cell Regulation. *Physiol. Rev.* **93**, 1019–1137
236. Poger, D., Van Gunsteren, W. F., and Mark, A. E. (2010) A new force field for simulating phosphatidylcholine bilayers. *J. Comput. Chem.* **31**, 1117–1125
237. Irvine, W. ((2014) *Testing of parameters for a biologically accurate brain membrane and molecular dynamics simulations exploration of membrane interactions and conformational changes exhibited by p110α and its oncogenic mutants*. Ph.D. thesis, Massey University
238. Burke, J. E., Perisic, O., Masson, G. R., Vadas, O., and Williams, R. L. (2012) Oncogenic mutations mimic and enhance dynamic events in the natural activation of phosphoinositide 3-kinase p110α (PIK3CA). *Proc. Natl. Acad. Sci.*
239. Chen, R., Poger, D., and Mark, A. E. (2011) Effect of High Pressure on Fully Hydrated DPPC and POPC Bilayers. *J. Phys. Chem. B.* **115**, 1038–1044
240. Šali, A., and Blundell, T. L. (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **234**, 779–815
241. Hobbs, G. A., Der, C. J., and Rossman, K. L. (2016) RAS isoforms and mutations in cancer at a glance. *J. Cell Sci.* **129**, 1287–1292
242. Vetter, I. R. (2014) The structure of the G domain of the Ras superfamily. in *Ras Superfamily Small G Proteins: Biology and Mechanisms 1: General Features, Signaling* (Wittinghofer, A. ed), pp. 25–50, Springer Vienna, Vienna, 10.1007/978-3-7091-1806-1_2
243. Giannakouros, T., and Magee, A. I. (1992) Protein prenylation and associated modifications. in *Lipid modifications of proteins* (Schlesinger, M. J. ed), pp. 50–65,

CRC Press, Inc., Boca Raton, Fla.

244. Michaelson, D., Ali, W., Chiu, V. K., Bergo, M., Silletti, J., Wright, L., Young, S. G., and Philips, M. (2005) Postprenylation CAAX processing is required for proper localization of Ras but not Rho GTPases. *Mol. Biol. Cell*. **16**, 1606–1616

245. Paterson, H. (1994) N-Terminally Myristoylated Ras Proteins Require Palmitoylation or a Polybasic Domain for Plasma Membrane Localization. *Proteins*. **14**, 4722–4730

246. Jo, S., Lim, J. B., Klauda, J. B., and Im, W. (2009) CHARMM-GUI Membrane Builder for Mixed Bilayers and Its Application to Yeast Membranes. *Biophys. J.* **97**, 50–58

247. Izrailev, S., Stepaniants, S., Isralewitz, B., Kosztin, D., Lu, H., Molnar, F., Wriggers, W., and Schulten, K. (1999) Steered Molecular Dynamics BT - Computational Molecular Dynamics: Challenges, Methods, Ideas (Deuflhard, P., Hermans, J., Leimkuhler, B., Mark, A. E., Reich, S., and Skeel, R. D. eds), pp. 39–65, Springer Berlin Heidelberg, Berlin, Heidelberg

248. Frenkel, D., Smit, B., and Ratner, M. A. (1997) Understanding Molecular Simulation: From Algorithms to Applications. *Phys. Today*. **50**, 66–66

249. Lemkul, J. A., and Bevan, D. R. (2010) Assessing the Stability of Alzheimer's Amyloid Protofibrils Using Molecular Dynamics. *J. Phys. Chem. B.* **114**, 1652–1660

250. Engelman, J. A. (2009) Targeting PI3K signalling in cancer: opportunities, challenges and limitations. *Nat. Rev. Cancer*. **9**, 550–562

251. Quayle, S. N., Lee, J. Y., Cheung, L. W. T., Ding, L., Wiedemeyer, R., Dewan, R. W., Huang-Hobbs, E., Zhuang, L., Wilson, R. K., Ligon, K. L., Mills, G. B., Cantley, L. C., and Chin, L. (2012) Somatic Mutations of PIK3R1 Promote Gliomagenesis. *PLoS One*. **7**, e49466

252. Fu, Z., Aronoff-Spencer, E., Backer, J. M., and Gerfen, G. J. (2003) The structure of the inter-SH2 domain of class IA phosphoinositide 3-kinase determined by site-directed spin labeling EPR and homology modeling. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3275–3280

253. Miled, N., Yan, Y., Hon, W.-C., Perisic, O., Zvelebil, M., Inbar, Y., Schneidman-Duhovny, D., Wolfson, H. J., Backer, J. M., and Williams, R. L. (2007) Mechanism of Two Classes of Cancer Mutations in the Phosphoinositide 3-Kinase Catalytic Subunit. *Science (80-.).* **317**, 239 LP – 242

254. Yu, J., Zhang, Y., McIlroy, J., Rordorf-Nikolic, T., Orr, G. A., and Backer, J. M. (1998) Regulation of the p85/p110 phosphatidylinositol 3'-kinase: stabilization and inhibition of the p110alpha catalytic subunit by the p85 regulatory subunit. *Mol. Cell. Biol.* **18**,

1379–1387

255. Jiménez, C., Hernández, C., Pimentel, B., and Carrera, A. C. (2002) The p85 regulatory subunit controls sequential activation of phosphoinositide 3-kinase by Tyr kinases and Ras. *J. Biol. Chem.* **277**, 41556–41562

256. Miled, N., Yan, Y., Hon, W. C., Perisic, O., Zvelebil, M., Inbar, Y., Schneidman-Duhovny, D., Wolfson, H. J., Backer, J. M., and Williams, R. L. (2007) Mechanism of two classes of cancer mutations in the phosphoinositide 3-kinase catalytic subunit. *Science (80-.).* **317**, 239–242

257. Rostislavleva, K., Soler, N., Ohashi, Y., Zhang, L., Pardon, E., Burke, J. E., Masson, G. R., Johnson, C., Steyaert, J., Ktistakis, N. T., and Williams, R. L. (2015) Structure and flexibility of the endosomal Vps34 complex reveals the basis of its function on membranes. *Science (80-.).* **350**, 1–25

258. Silvius, J. R., and L'Heureux, F. (1994) Fluorimetric Evaluation of the Affinities of Isoprenylated Peptides for Lipid Bilayers. *Biochemistry*. **33**, 3014–3022

259. Brunsveld, L., Waldmann, H., and Huster, D. (2009) Membrane binding of lipidated Ras peptides and proteins - The structural point of view. *Biochim. Biophys. Acta - Biomembr.* **1788**, 273–288

260. Ziemba, B. P. ;, Swisher, G. H. ;, Masson, G., Burke, J. E. ., Williams, R. L. ., and Falke, J. J. (2016) Regulation of a Coupled MARCKS–PI3K Lipid Kinase Circuit by Calmodulin: Single-Molecule Analysis of a Membrane-Bound Signaling Module. **55**, 6395–6405

261. Malde, A. K., Zuo, L., Breeze, M., Stroet, M., Poger, D., Nair, P. C., Oostenbrink, C., and Mark, A. E. (2011) An Automated force field Topology Builder (ATB) and repository: Version 1.0. *J. Chem. Theory Comput.* **7**, 4026–4037

262. Manunza, B., Deiana, S., Pintore, M., and Gessa, C. (1997) Structure and internal motion of solvated beta-cyclodextrine: a molecular dynamics study. *J. Mol. Struct. THEOCHEM*. **419**, 133–137

263. Horta, B. A. C., Merz, P. T., Fuchs, P. F. J., Dolenc, J., Riniker, S., and Hünenberger, P. H. (2016) A GROMOS-Compatible Force Field for Small Organic Molecules in the Condensed Phase: The 2016H66 Parameter Set. *J. Chem. Theory Comput.* **12**, 3825–3850

264. Plazinski, W., Lonardi, A., and Hünenberger, P. H. (2015) Revision of the GROMOS 56A6CARBO force field: Improving the description of ring-conformational equilibria in hexopyranose-based carbohydrates chains. *J. Comput. Chem.* 10.1002/jcc.24229

265. Pol-Fachin, L., Rusu, V. H., Verli, H., and Lins, R. D. (2012) GROMOS 53A6GLYC, an Improved GROMOS Force Field for Hexopyranose-Based Carbohydrates. *J. Chem. Theory Comput.* **8**, 4681–4690

266. Wildman, S. A., and Crippen, G. M. (1999) Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873

267. Sholl, D., and Steckel, J. A. (2011) *Density Functional Theory A Practical Introduction*, NY John Wiley & Sons, New York

268. Bannan, C. C., Calabr, G., Kyu, D. Y., and Mobley, D. L. (2016) Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. *J. Chem. Theory Comput.* **12**, 4015–4024

269. Harada, A., Li, J., Suzuki, S., and Kamachi, M. (1993) Complex Formation between Polyisobutylene and Cyclodextrins: Inversion of Chain-Length Selectivity between β-Cyclodextrin and γ-Cyclodextrin. *Macromolecules*. **26**, 5267–5268

Appendix
(Chapter 2)

A.1. Choice and validation of parameters

The GROMOS 54A7 force field was chosen to simulate the host-guest complexation process because its parameterisation process includes a strong emphasis on reproduction of thermodynamic properties, which are what will predominantly drive the formation of a stable host-guest complex. The standard GROMOS 54A7 force field parameter set does not include parameters for cyclodextrins or bullvalene, however, so it was necessary to find parameters for both molecules compatible with the GROMOS 54A7 force field. The following sections describe first, validation of the GROMOS methanol solvent model, second, evaluation of the suitability of a number of different GROMOS-style parameter sets for cyclodextrins, and third, the parameterisation of bullvalenes. The simulation methods used during parameter validation are described in Chapter 2ii, Section 2ii.2.1.3.

A.1.1. Choice and validation of solvent parameters

Choosing the best solvent to use is an important step in any experiment as well as simulation. As cyclodextrin has a hydrophilic exterior and hydrophobic interior, a polar solvent would be a suitable choice to encourage complexation of a hydrophobic guest molecule. However, such guest molecules are not very soluble; this can make complexation very slow or sometimes not possible even if the complex ultimately formed would be stable. Therefore, an organic solvent could be used to dissolve the guest molecule. However, that solvent must not compete with the guest in complexation with the host. For experiments, it also should be evaporated easily. Two good example of such solvents could be ethanol and diethylether (21). Methanol was initially recommended as a good solvent to use by our experimental collaborators. Water is also a possible solvent, however, as it does not compete with the guest to bind to cyclodextrin (Dr. Thomas Fallon, Oussama Yahiaoui, *Pers. Commun*.).

All GROMOS force fields are parameterised to work with the SPC (2) water model, therefore it was also used for all simulations here with water as a solvent.

Methanol is available as a solvent as well as a solute in the GROMOS force field. To determine whether the GROMOS solvent model has been correctly implemented in the GROMACS software, MD simulations were performed for two different situations, each of which should reproduce the properties of liquid methanol. The first comprised a single solute methanol molecule in a box of methanol as solvent, and the second a box of solute methanol

molecules with no solvent (Figure A. 1). Each system was energy minimised, heated and equilibrated in the canonical (NVT) ensemble, and then simulated for 40 ps in the isobaric-isothermal (NpT) ensemble.

The mean square displacements (MSD) of atoms from their initial positions are roughly linear (Figure A. 2, A), meaning that the diffusion constant can be validly calculated using the Einstein relation. The diffusion constants calculated in this way are $1.34 \pm 0.36 \times 10^{-5}$ for the solute methanol in solvent methanol, and $1.37 \pm 0.44 \times 10^{-5}$ $cm^2 \cdot s^{-1}$ for solute methanol, respectively, which are equivalent within the error (the difference of the diffusion coefficients obtained from fits over the two halves of the fit interval (180)).

The radial distribution function (RDF) vs. distance describes structure in a solvent and so is a useful check of the liquid properties. The RDF for carbon-carbon, oxygen-oxygen, carbon-oxygen, and oxygen-carbon atoms was compared for the two cases (Figure A. 2, B(i) to B(iv)). The peaks are in the same position and of the same height for both methanol systems, suggesting similar properties.

The average value of the density over the course of the simulation is 820 $kg \cdot m^{-3}$ in both cases, close to the experimental value of 792 $kg \cdot m^{-3}$ (Figure A. 2, C). Additionally, the temperatures are stable over time, indicating that the system is well-equilibrated with respect to temperature (Figure A. 2, D).

These comparisons of the mean square displacement, RDFs, and density confirmed that the properties of liquid methanol are essentially the same for both systems, and thus that the GROMOS solvent model is correctly implemented in the GROMACS software.

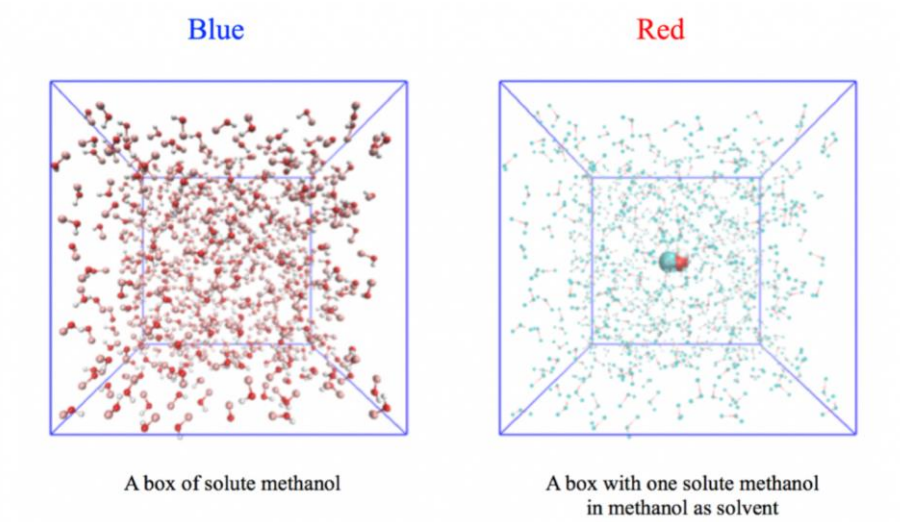


Figure A. 1. Snapshots of the solvated box environment of the two methanol systems. The cube represents the boundaries of the box. (left) A box of solute methanol molecules with no solvent (blue lines on graphs in Figure A.2.) and (right) solute methanol in a box of solvent methanol (red lines on graphs in Figure A.2.).

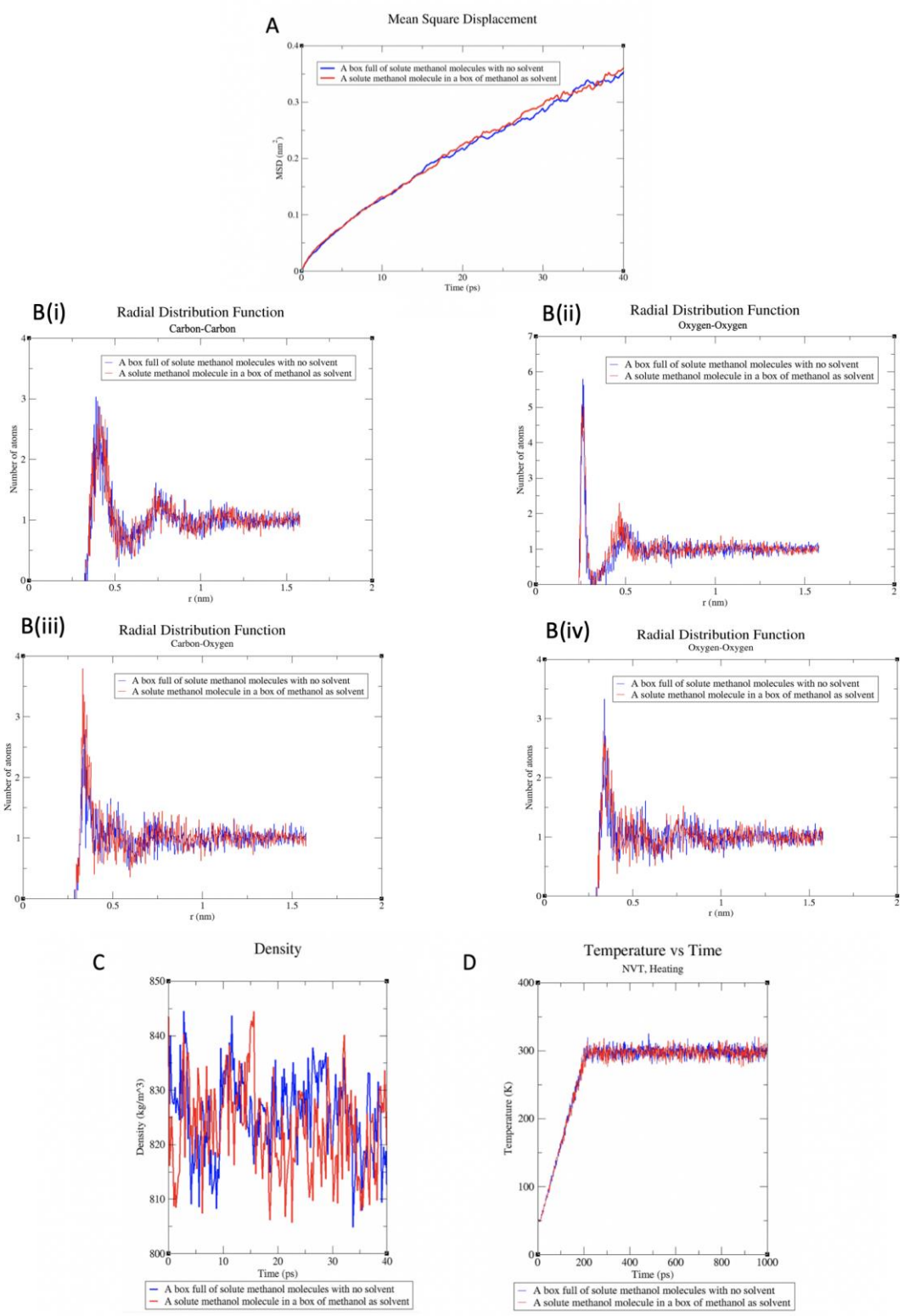


Figure A. 2. Key properties calculated from the 40 ps MD simulation of each methanol system illustrated in Figure A.1: (A) mean square displacement vs. time; (B) radial distribution function (RDF) vs. distance for (i) carbon-carbon, (ii) oxygen-oxygen, (iii) carbon-oxygen, (iv) oxygen-carbon; (C) density vs. time; and (D) temperature vs. time.

If cyclodextrin is to be used as a host molecule to bind to bullvalene in methanol as a solvent, then ideally cyclodextrin should prefer to bind to bullvalene rather than methanol. To find out, the binding free energy of a single methanol molecule to βCD (in water as a solvent), ΔG_{bind}, was calculated as per the thermodynamic cycle outlined in Figure A. 3. ΔG_1 and ΔG_3 were calculated using TI. Assuming that $\Delta G_2 = 0$, ΔG_{bind} can then be calculated according to:

$\Delta G_3 = -48.86 \pm 1.61\ kJ{\cdot}mol^{-1}$

$\Delta G_1 = -41.74 \pm 5.68\ kJ{\cdot}mol^{-1}$

$\Delta G_{Cycle} = 0 = \Delta G_{bind} - \Delta G_3 + \Delta G_2 + \Delta G_1$

$\Delta G_{bind} \cong -7.12\ kJ{\cdot}mol^{-1}$

The calculated negative binding free energy of methanol to βCD shows that methanol prefers the βCD cavity to bulk water. Therefore, in the case of bullvalene-βCD complexation in methanol as a solvent, there might be an unfavourable competition between bullvalene and methanol to bind to the host. As a result, although methanol is a good solvent for bullvalene, its favourable complexation with βCD means it would not be the best choice of solvent. Experimental results determined by Dr. Thomas Fallon (*Pers. Commun.*) also showed binding of methanol to βCD, supporting the theoretical conclusion. Therefore, as methanol may compete with the solute to bind to βCD, it was decided to use water as the solvent instead.

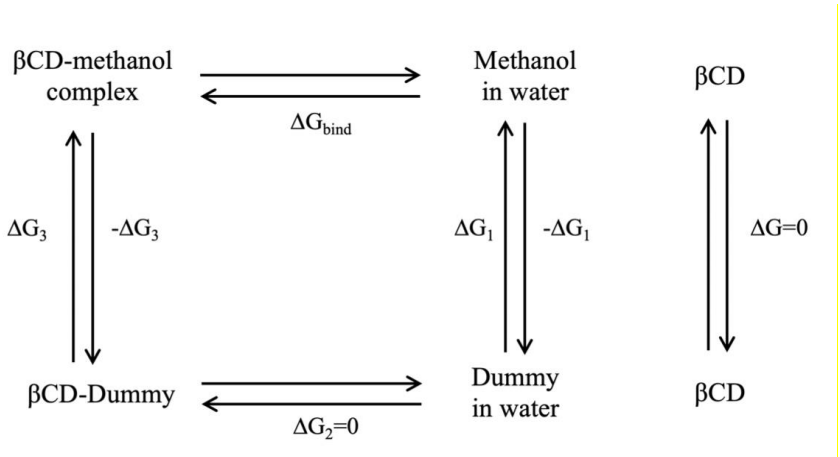


Figure A. 3. Thermodynamic cycle for computing the free energy of methanol-βCD complexation in water. $-\Delta G_1$ is the solvation free energy of methanol in water; $-\Delta G_3$ is the free energy cost of switching from a methanol molecule bound to βCD to a dummy atom bound to βCD, and 'Dummy' refers to a dummy atom that does not interact with other particles.

A.1.1.2. Choice and validation of cyclodextrin parameters

A.1.1.2.1. ATB-derived GROMOS 54A7 parameters for cyclodextrin

The Automated Topology Builder (ATB) (261) is a web server that serves as a repository for GROMOS-compatible parameter sets. It also allows users to submit new molecules for parameterisation, for which quantum mechanical (QM) calculations are carried out from which GROMOS-like parameter sets are derived. Seven parameter sets for βCD were already available on ATB (MolIDs 2698, 23854, 22007, 22072, 22069, 22120 and 19880). These parameter sets were tested to determine how well they describe βCD behaviour in water and in methanol. For each parameter set, the initial structure was the corresponding coordinates provided in ATB. The βCD was energy minimised, solvated in SPC water (203) or GROMOS solvent methanol (173)), heated and equilibrated in the NVT ensemble, and simulated for 100 ns in the NpT ensemble.

To quantify the structural stability, the atom-positional root-mean-square deviation (RMSD) of all atoms was calculated for the NpT trajectories relative to the first frame over the simulation time and also relative to a crystal structure of βCD (PDB ID: 4J3U). The average RMSD values relative to the crystal structure for the seven βCD parameter sets solvated in water are 0.9, 0.8, 0.8, 1.4, 1.1, 1.4 and 0.7 nm for MolIDs 2698, 23854, 22007, 22072, 22069, 22120 and 19880, respectively. Initially, the parameter set that gave the lowest RMSD relative to the crystal structure (MolID: 19880) was chosen among the seven and used in the calculations of the binding free energy of methanol to βCD (previous section). However, visual analysis showed βCD simulated with any of the ATB parameter sets to be highly rigid, whereas it is expected to exhibit a small degree of structural fluctuation (262). The ATB parameters may therefore be over-fitted to the QM-minimised structures. Additionally, as the ATB only carries out high-level QM calculations for small molecules (< 50 atoms), the ATB parameters for βCD have been derived from low-level (HF/STO-3G) calculations and so may not be particularly reliable. Therefore, after discussions with other simulation groups who have more experience in working with βCD (Dr Thomas Piggot, DSTL, UK, and Dr Niels Hansen, Universität Stuttgart, Germany, *Pers. Commun.*), it was decided to test alternative parameter sets.

A.1.1.2.2. 2016H66 force field parameters for cyclodextrin

A new GROMOS-style force field, 2016H66, was calibrated and validated for small organic molecules (263). It includes parameters for a range of different sugars, including those that

make up βCD, as well as coupling parameters to describe glycosidic linkages. This force field therefore seemed ideal for representing cyclodextrin.

To confirm that the 2016H66 force field performs well for cyclodextrin, and to generate equilibrated cyclodextrin structures for use in bullvalene binding free energy calculations, MD simulations with 2016H66 parameters were performed for βCD in SPC water. However, after visually analysing the energy minimization trajectory , it was found that the initial 'chair' structures of the sugar units in βCD were converted to 'boat' form (Figure A. 4), which is energetically less favourable (264) and should therefore not be formed during energy minimisation, and only occasionally occur during MD simulations. As well as being improbable, the chair-boat conversion resulted in βCD becoming unstable and the ring of sugar molecules collapsing inwards, reducing the size of the central cavity so that it is unsuitable for binding a guest molecule. Based on these results, it was concluded that 2016H66 may not be suitable for the simulation of bullvalene binding by βCD.

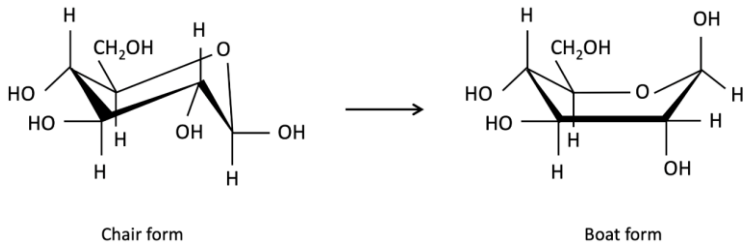


Figure A. 4. The unfavourable inversion from chair to boat form that occurred after energy minimization with the 2016H66 parameters.

A.1.1.2.3. GROMOS 53A6_GLYC parameters for cyclodextrin

GROMOS 53A6_GLYC is an improved parameter set for carbohydrate simulations that has recently been tested for cyclodextrins (20, 192, 193). All 16 possible aldohexopyranose-based monosaccharides are well described (265). This set has improved torsional potential parameters of sugar rings compared to the GROMOS 53A6 force field. Other parameters are taken directly from the GROMOS 53A6 force field. The proposed parameter modifications are compatible with the GROMOS 53A6 and 54A7 parameter sets (265). Therefore, this force field is a promising candidate for cyclodextrins. After energy minimization and a 50 ns MD simulation of βCD in SPC water, no inversion of the chair forms of the sugar units to boat forms occurred, and the system was stable. Thus, the GROMOS 53A6_GLYC was concluded to be the best choice for simulating bullvalene binding by cyclodextrins.

A.1.1.3. Choice and validation of bullvalene parameters

Unlike the solvents and cyclodextrin, bullvalene has never been studied using MD simulations before, thus there are no pre-existing force field parameters for bullvalene. One way to obtain force field parameters in a manner that should be consistent with the GROMOS 54A7 force field is by using ATB. Another possibility is to parameterise the bullvalenes manually by comparison to other chemically similar molecules for which the GROMOS force field already contains parameters. Both methods were explored, as outlined below

Validation of these parameters would then require comparison with the results of QM calculations, e.g. how similar the structures sampled in an MD simulation are to the lowest energy QM structure, and/or experimental data. Unfortunately, there is relatively little experimental data available for bullvalenes. One property that is relevant to the ultimate use of these parameters – calculating binding free energies in a particular solvent – is the solvation free energy, from which the partition coefficient can subsequently be derived. While experimental values for these properties have not yet been measured, approximate values have been calculated using the Crippen method (266).

A.1.1.3.1. ATB-derived GROMOS 54A7 parameters for bullvalene

The coordinates of the four distinct isomers of each of two monosubstituted bullvalenes (hydroxymethyl- and methyl-bullvalene), and the most favourable isomer of the disubstituted bullvalene (hydroxymethyl), as determined using QM calculations (14, 16) were submitted to ATB (194–196). Methyl-bullvalene was used for parameters validation, because it was initially the intended target of experiments by our experimental collaborators. The structure of unsubstituted bullvalene was also generated using Gaussview 5.0 (198) was also submitted to ATB. Because the number of atoms in all three bullvalenes is less than 50, the ATB-derived parameters were based on the highest level of QM calculations (Hessian, DFT, B3LYP/6-31G* (267)) provided by the ATB server. The resulting parameter sets remain available from the ATB as MolIDs 33737 for unsubstituted, 36408, 192857, 238996, and 239000 for the four isomers of mono-hydroxymethyl bullvalenes, 34907, 34908, 34909, and 34906 for the four isomers of methyl-bullvalene, and finally 48412 for di-hydroxymethyl bullvalene.

As a simple check of the ATB bullvalene parameters, the bond lengths of the crystal structure of unsubstituted bullvalene (13) and the optimized unsubstituted bullvalene geometry from ATB, from which the GROMOS 54A7-style force field parameters were produced, were compared and found to be highly similar.

To further test the parameters, energy minimisation, heating and equilibration (NVT), followed by 100 ns MD simulation in SPC water, was run for unsubstituted bullvalene, and the same procedure carried out in SPC water and in methanol (solvent) for the four distinct isomers of methyl-bullvalene (results are not shown). Accordingly, this GROMOS 54A7-style force field parameter set was concluded to be the best choice for simulating bullvalene.

A.1.1.3.2. Bullvalene partial charge optimisation

The partial charges assigned to the bullvalenes by the ATB were high compared to the partial charges assigned to other hydrocarbon molecules in the GROMOS 54A7 force field. For instance, in the GROMOS force field, carbon atoms in hydrocarbons or regions of molecules that have only carbons and hydrogens tend to have all zero charges, whereas the ATB suggests partial charges as high as −0.129.

To investigate the effect of these high charges on the behaviour of bullvalene, three additional partial charge schemes were constructed by scaling the ATB charges by factors of 0.00 to 1.00 in intervals of 0.25. In each case, thermodynamic integration (TI) was used to calculate the solvation free energy (ΔG_{solv}) values for the most stable isomer of methyl-bullvalene in water and methanol solvents. The trend in the results from the free energy calculations were as expected, with methyl-bullvalene preferring to be solvated in methanol more than in water. As the charges were scaled down, methyl-bullvalene became less soluble in water (Table A. 1).

Table A. 1. Solvation free energies (ΔG_{solv}) in methanol (MeOH) and in water (ΔG_{hyd}) and methanol-water partition coefficients of the lowest energy methyl-bullvalene isomer with ATB-derived partial charges scaled by factors from 0 to 1.

ATB Charge Scaling Factor	ΔG_{solv} (MeOH) (kJ·mol^{-1})	ΔG_{hyd} (kJ·mol^{-1})	log *P*
1.0	-36.80	-19.26	3.07
0.75	-26.34	-7.41	3.32
0.50	-26.15	+4.33	5.34
0.25	-26.50	+4.37	5.41
0.00	-25.66	+5.23	5.41

To better estimate the solubility of bullvalene in each solvent, the partition coefficient (log *P*), which is commonly used to describe the solvent preference of a solute molecule, was calculated according to Equation A. 1:

$$\log P = \frac{-\Delta G_{transfer}}{RT\ln(10)} = \frac{\Delta G_{transfer} - \Delta G_{solvation}}{RT\ln(10)} \quad ,$$

Equation A. 1

where $\Delta G_{transfer}$ is the transfer free energy of the solute moving from water to the organic solvent, $\Delta G_{solvation}$ and $\Delta G_{hydration}$ are the solvation free energy of the solute in the organic solvent and in water, respectively, R is the Gas Constant (R = 8.314 J·mol·K) and T is the temperature, which was 298 K.

Since log P is directly proportional to $\Delta G_{transfer}$ (268), more positive log P values show that the solute molecule prefers the non-aqueous solvent more. The results in Table A. 1 demonstrate that methyl-bullvalene favours dissolution in methanol rather than water at 298 K with all charge sets tested, in keeping with the solvation free energy values and with the known improved solubility of bullvalenes in methanol (Dr T. Fallon, *Pers. Commun*.). Reducing the partial charges significantly reduces its solubility in water, however. However, it is known from experiments that methyl-bullvalene is soluble in water (Dr Thomas Fallon, *Pers. Commun.*). As a result, the full ATB charges are most appropriate for simulation of bullvalenes.

A.1.1.3.3. 2016H66 force field parameters for bullvalene

It would be preferable to use the same force field to describe bullvalene and cyclodextrin, so while the 2016H66 force field was being considered for cyclodextrin, the key force field parameters relevant to modelling bullvalenes and methanol as a solvent were compared between the 2016H66 and GROMOS 54A7 (as used in the ATB) force fields. They were found to be either identical or highly similar, and thus it would be straightforward to generate 2016H66 parameters for the bullvalenes by analogy with the ATB parameter sets. However, given that the ATB assigned non-standard partial charges to the bullvalene atoms, and that the solvation free energy and partition coefficient calculations (Table A. 1) revealed that these charges were appropriate, these were retained for the 2016H66 bullvalene parameter set.

A test 100 ns MD simulation of unsubstituted bullvalene was performed using this 2016H66 parameter set in SPC water. The average RMSD relative to the first frame over the simulation time was 0.053 nm (data not shown), and no chemically unrealistic behaviour was observed, suggesting that the 2016H66 force field with ATB charges also provides an appropriate description of the behaviour of bullvalene in solution.

A.1.1.3.4. GROMOS 53A6_GLYC parameters for bullvalene

Ultimately, since the GROMOS 53A6_GLYC parameter set was chosen for cyclodextrins, and is compatible with the GROMOS 53A6 and 54A7 parameter sets, it was decided to use the ATB-derived GROMOS 54A7-style force field for bullvalenes.

A.1.1.4. Cyclodextrin coordinates

A.1.1.4.1. β-cyclodextrin coordinates

Initially, while testing the ATB parameter sets for βCD (Chapter 2ii, Section 2ii.2.1.2), the coordinates corresponding to the seven parameter sets (MolIDs 2698, 23854, 22007, 22072, 22069, 22120 and 19880) were used to initiate simulations of βCD in water and in methanol using their corresponding parameter set. After ATB parameter set MolID 19880 was selected, the corresponding βCD coordinates were used to undergo the process in Figure 2i. 1 up to the unbiased MD simulation after the HGBPEP method. However, visual analysis suggested that the βCD was unrealistically rigid, and this was confirmed by analysis of structural properties such as the distance between the seven sugar rings, and the angle across the oxygen atoms that link each pair of sugar units. Therefore, a crystal structure for βCD was taken from Cambridge Crystallographic Data Centre (166) (www.ccdc.cam.ac.uk , identifier: ARUXIU, and deposition number: 187189) and used for all subsequent simulations.

A.1.1.4.2. γ-cyclodextrin coordinates

The structure and symmetry of γCD, which contains eight sugar units, are different to those of βCD, which contains only seven. With a guest molecule bound, γCD is more likely to show a symmetrical conformation; however, in the absence of guest molecule inside the cavity, it assumes a less symmetrical conformation (269). Similar to βCD, initial coordinates of γCD were taken from ATB (molID: 26959) and from a crystal structure taken from ChemSpider (CSID:10469499, http://www.chemspider.com/Chemical-Structure.10469499.html). Both were tested by running an unbiased MD simulation; however, neither structure was perfectly symmetrised initially or during the simulation. An open, symmetric structure is important for HGBPEP. Hence, the crystal structure was symmetrised using VMD. Subsequent MD simulations showed this C_8 symmetrised conformation to be stable, therefore this structure was used for all further MD simulations.