UNIVERSITY OF SOUTHAMPTON

Faculty of Natural and Environmental Sciences

Exploring conformation of human fatty acid synthase inhibitors using Replica Exchange Molecular Dynamics

by

Nawel Mele

PhD Thesis, September 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

Faculty of Natural and Environmental Sciences

Chemistry

Thesis for the degree of Doctor of Philosophy

Exploring conformation of human fatty acid synthase inhibitors using Replica Exchange Molecular Dynamics

Nawel Mele

Human fatty acid synthase (hFAS) is a homodimer multienzyme complex involved in the lipogenesis and catalysis of long-chain fatty acids. hFAS is overexpressed in cancer cells and enhances tumor growth. A recent study reported a new potent and selective inhibitor of the β-ketoacyl reductase (KR) domain of hFAS, GSK2194693. An x-ray crystal structure of this inhibitor bound to the KR domain provides binding mode information regarding the druggable pocket. In this thesis simulations and analysis of the solution-phase conformational ensembles of four inhibitors of the human fatty acid synthase are required. The ensembles are generated using replica exchange enhanced sampling molecular dynamics approaches for two force fields, and analysed using a combination of dihedral and Cartesian space clustering, and principal components analysis. These ensembles are compared to experimental data derived using nuclear magnetic resonance from C4X Discovery to evaluate the convergence of our data and to analyze the influence of the force field on the quality of the sampling. We find that while the simulations are able to identify all the conformations found by NMR, their relative populations are in less satisfactory agreement. The ligand-receptor complex binding modes were also investigated by first identifying conformations of the four compounds with shape and chemical group similarity using clustering and superimposition methodologies. Then, in a ligand preorganization approach to identify if the solution phase conformations obtained from NMR and REMD bind favourably to the receptor binding pocket, the interactions made with hFAS were evaluated keeping the conformations and the receptor rigid. Potential binding modes for the compounds were generated with consistent interactions. Contacts found in the x-ray structure GSK2194069 were highly conserved in the compounds and additional hydrogen bonds were identified. Thus, this study offers valuable information for future drug development and optimization.

Table of Contents

Table of Figures

Table of Tables

Academic Thesis: Declaration Of Authorship

I, Nawel Mele
declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Exploring conformations of human fatty acid synthase inhibitors using Replica Exchange Molecular Dynamics

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either None of this work has been published before submission or Parts of this work have been published as: (Not applicable)

Signed: .....
Date: .....

Acknowledgements

Several people have contributed to make these four years an amazing experience. A page is not enough to thank them all but I would like to take this opportunity to first thank my supervisor Pr Jonathan Essex for offering me this chance of doing a PhD and all his support and guidance throughout this research. Thank you to my industrial supervisor Dr Richard Ward for his involvement throughout this project, his encouragement, support and guidance. I have learned so much and developed abilities that will be key in all part of my professional future. I would also like to thank all my collaborators in this project from Portsmouth University, Astra Zeneca and C4X for their help and inspiration.

There is one particular person I would like to thank in a different way, we started our PhD together as “team PhD” and we finished this PhD together as friend for life. To my friend, my soul sister Marija Miljak, I will never thank you enough for your friendship, your support and encouragements during all these years. You were by my side during the fun times and during the hardest times of this PhD, a thousand thanks “my beauty” (You know this one).

I would like to thank from the bottom of my heart my family for always supporting me in my decision to go abroad to study; my mum for her precious advice; my dad for always being there for me despite the difficult times he has been through; and of course my two brothers and my lovely nephew who always keep an eye on me. My thank you extends here to all my friends in France for staying along my side until the end despite the distance separating us and to the friends I made here, especially Liza Zelgay and her entire family who open their hearts and door for me during this journey, thank you so much.

Thank you for the whole Essex group - for the fun and all the help that have helped me get through these years.

Abbreviations

HTS: High Throughput Screening

ADME: Absorption, Distribution, Metabolism, Excretion

ADMET: Absorption, Distribution, Metabolism, Excretion, and Toxicity

PK: PharmacoKinetic

SOC: Standard Of Care

FDA: Food and Drug Administration

SBDD: Structure-Based Drug Design

QSAR: Quantitative Structure Activity Relationships

ITC: Isothermal Titration Calorimetry

NADPH: Nicotinamide Adenine Dinucleotide Phosphate

hFAS: Human fatty acid synthase

MAT: malonyl-acetyl CoA transferase

KS: β-keoacyl synthase

KR: β-ketoacyl reductase

DH: β-hydroxyacyl dehydratase

ER: enoyl reductase

TE: Thioesterase

ACP: Acyl Carrier Protein

ACC: acetyl-CoA carboxylase

NMR: Nuclear Magnetic Resonance

DMSO: Dimethyl sulfoxide

3D: Three Dimensional

cMD: Conventional Molecular Dynamics

MD: Molecular Dynamics

REMD: Replica Exchange Molecular Dynamics

CSD: Cambridge Structural Database

NOE: Nuclear Overhauser Effect

TMS: Tetramethylsilane

AMBER: Assisted Model Building with Energy Refinement

CHARMM: Chemistry at HARvard using Molecular Mechanics

OPLS-AA: Optimized Potential For Liquid Simulations - All Atoms

GAFF: General AMBER force field

CGenFF: CHARMM General force field

PBC: Periodic Boundary Conditions

VdW: Van der Waals

PME: particle mesh Ewald

MOE: Molecular Operating Environment

DASH: Dynamics Analysis by Salt and Hudson

RMSD: Root Mean Squared Deviation

PCA: Principal Component Analysis

DNA: DeoxyriboNucleic Acid

ROCS: Rapid Overlays of Chemicals Structures

PC: Principal Components

VMD: Virtual Molecular Dynamics

REST: Replica Exchange with Solute Tempering

Chapter 1 Introduction

1.1 Drug discovery:

The process leading to the design of new drugs is complex, time consuming and very expensive. Drug discovery can often take 10 to 15 years and £400-800 million to identify and translate a promising molecule into a commercialised drug [1-4]. Many suggest that it can be significantly higher [5]. Targeted drug discovery starts with the identification of the biological target involved in a disease, although phenotypic approaches are also used to identify compounds, which elicit a specific biological response. Nowadays, genomic, bioinformatics and proteomics enable the identification of genes or proteins involved in diseases as potential therapeutic targets. Once the link between the biological target and the disease is revealed, scientific researches can aim to inhibit or increase target function, to fulfil their deficiency or to perform the function they are not able to execute. *Figure 1.1* shows the different phases and some of the approaches employed during the drug discovery process [6].

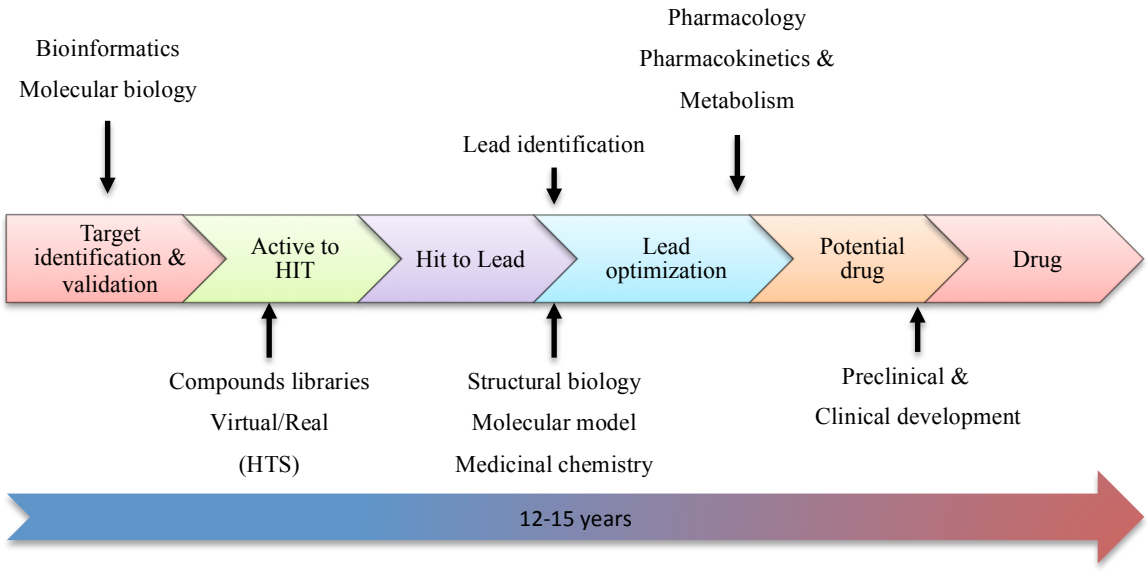


Figure 1.1: The different phases of the drug discovery process

Once a target has been validated pre-clinically to an acceptable level of understanding, small molecules modulators are often studied (if known) as starting points for a potential drug, which can modulate its action. High Throughput Screening (HTS) is a popular approach in drug discovery aiming to

identify bioactive molecules in large chemicals libraries against a chosen set of predefined targets. HTS may identify molecules, “hits”, that demonstrate activity against the target. Hit molecules will be then tested in a variety of efficacy and non-efficacy assays (solubility, stability, lipophilicity) to identify their potential as lead molecules. Lead molecules require further optimisation commonly to improve their potency toward the target and improve physicochemical properties. At this point of the research understanding the absorption, metabolism, excretion (ADME) and toxicity of the compounds is crucial. Unsuitable ADME properties or toxicity is often a major reason for failure of potential clinical drug candidates [7-10].

A key challenge in drug discovery occurs in the preclinical and clinical phases. In the preclinical phase, the optimised compounds are tested in vivo, on animals, to confirm their therapeutics activity and provide information on dosing and toxicity levels to decide whether the drug should be tested in human clinical studies. Human clinical trials occur in a number of phases. Phase I corresponds to the first clinical study, on small numbers of patients, commonly for 6 to 18 months with the aim of understanding pharmacokinetic (PK), safety and early efficacy signals. Phase II is carried out across a broader group of patients from 2 to 3 years to further evaluate the safety and efficacy of the drug at a dose(s) identified from Phase I. During this phase, the optimal dosage is determined providing the best therapeutic results, which is acceptable in terms side effects. A phase III study is carried out at optimal dose(s) and compared to either a placebo arm or standard of care (SOC). The patients are divided “blindly” in two groups; the first group of patients will be treated with the drug while a placebo (or SOC) will be used for the other group. If the clinical data supports it and the drug is approved by the relevant agency such as the U.S. FDA, a phase IV study may often monitor the safety of the drug in the longer-term.

1.2 Computational chemistry in drug discovery:

The ultimate aim of drug discovery is to identify and optimise bioactive small molecules with favourable properties, which bind to specific target(s) of interest. However, small molecule drugs can adopt various conformations in solution but often only one will be compatible with binding to the target. It is not possible to synthesise or screen all of the possible “chemical space” to identify molecules and therefore, for economical, speed and scientific reasons, in silico drug design has emerged as a valuable approach [11-15]. In silico drug design involves multiple

techniques using computers to help in the development of new therapeutic agents such as structure-based drug design (SBDD), chemoinformatics, bioinformatics, Quantitative Structure Activity Relationships (QSAR) and also the large number of biological and chemical databases to enable knowledge sharing and computational assessment. Those tools can be used in various ways, from the generation and the analysis of hit molecules to lead optimization to improve binding properties and ADMET. The power of computers nowadays enables scientists to run a broad range of simulations to positively impact on the drug discovery process. Combining computer and experimental studies appear as a good compromise to assist in the generation of future drug molecules [16-19].

1.3 Receptor-ligand recognition:

1.3.1 Receptor-ligand complementarity:

The design of drug molecules can be assisted by exploiting detailed knowledge of the 3D structure of the target, or other compounds which are know to bind to that target. Understanding the mechanism of the binding between the active site of the receptor and a ligand enables the rational design of new molecules. In an ideal case, the shape and the electronic features of a small molecule need to be complementary to the target to enable the most effective binding to the active site. In 1894, Emil Fischer used the "lock and key" model to explain the mechanism of small molecules binding to a target whereby the ligand must possess the functional shape to fit into the receptor [20] *(Figure 1.2)*.

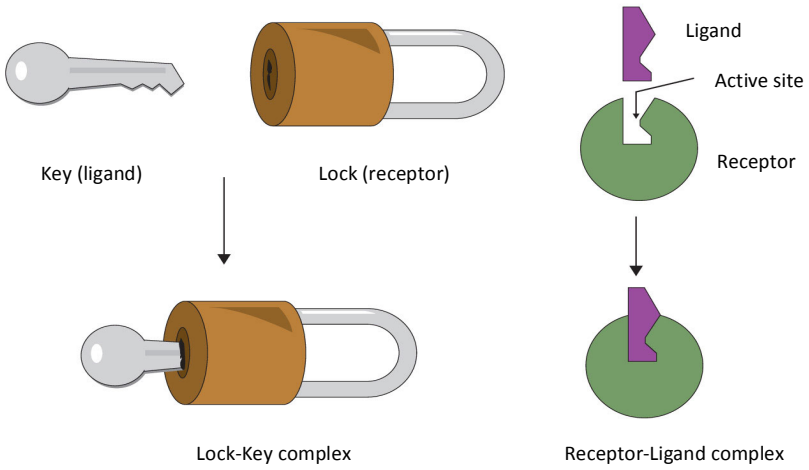


Figure 1.2: The Lock and Key model of the ligand-receptor mechanism (Source: saylordotorg.github.io)

The “lock and key” model portrays the ligand and the receptor as rigid and inflexible entities. However, we know they are both flexible and can change their conformation to facilitate the binding; it does however give a simple and practical view of the challenges and aims of structure-based drug design. Therefore, the development of the lock and key model led to the “induced fit model” by Daniel E.Koshland in 1959 [21]. In induced fit theory, the substrate plays a role in determining the final shape of the target and that the target is partially flexible. Therefore, while in the lock and key model the substrate shape was affected by the rigid active site of the target, in the induced fit model the active site can changes its shape to enfold a substrate molecule. Induced fit model has the advantage to explain how receptors may exhibit broad specificity.

The interaction between the drug and the receptor active site will induce pharmacological effects that can be measured in term of the affinity, efficacy and potency of the small molecule to its target [22-25]. Affinity corresponds to the strength of the interactions between the ligand and the receptor. Efficacy refers to the maximum biological effect that a drug can produce when bound to its target. Potency refers to the amount of drug required to achieve a defined biological effect. Affinity is measured by the inhibition constant K_i which can also be called the equilibrium dissociation constant K_D. This constant corresponds to the ligand concentration for which half of the active site is occupied. The smaller the K_i value is, the greater is the binding affinity of the ligand for its target. For a ligand L binding to receptor R forming the complex RL, the inhibition constant in a simple scheme can be defined as follow:

$$L + R \underset{k_{-1}}{\overset{k_{+1}}{\leftrightarrows}} LR$$

$$K_D = K_i = \frac{[R][L]}{[RL]} = \frac{k_{-1}}{k_{+1}} = \frac{1}{K_A}$$

[R], [L] and [RL] correspond to the concentration of the receptor, ligand and receptor-ligand complex at equilibrium respectively. k_1 and k_{-1} are the rate constants of association and dissociation respectively and K_A is the equilibrium association constant.

The potency can be defined by IC_{50} corresponding to the molar concentration of a ligand (drug) at which 50 % of the target is inhibited but does not inform if the desired biological effect was achieved. The lower is IC_{50}, the more potent is the molecule. IC_{50} can often be represented in logarithm as $pIC_{50} = -\log(IC_{50})$, thus the larger the pIC_{50} the more potent is the molecule.

However, when considering the interactions between ligand and target it is crucial not to ignore the impact of solvent (i.e. water). Indeed, in an initial state before binding has occurred, both ligand and target are surrounded by the solvent during the association step, both ligand and target, or a part of it, need to be desolvated to enable the binding. Shape complementarity between these two species is really important but for the ligand to remain in the binding pocket, specific protein-ligand interactions are necessary. For successful binding a combination of van der Waals, hydrogen bonds, electrostatics and hydrophobic interactions have to compensate the desolvation energy cost.

The global energy of the ligand-target complex is defined by the standard Gibbs free energy of binding, $\Delta G^{\Theta}_{bind,}$ by the following equation: $\Delta G^{\Theta}_{bind} = RTlnK_i = -RTlnK_A$ where R is the ideal gas constant (8.314 $J.K^{-1}.mol^{-1}$) and T is the temperature in Kelvin. A molecular system evolves to minimize this energy. So, when K_A is high ΔG^{Θ}_{bind} will be negative which will reflect a favourable association between the ligand and the target.

Additionally, ΔG^{Θ}_{bind} is composed of two independent thermodynamic terms, the change in standard enthalpy ΔH^{Θ} and the change in standard entropy ΔS^{Θ} with the Gibbs expression $\Delta G^{\Theta}_{bind} = \Delta H^{\Theta} - T\Delta S^{\Theta}$.

The change in enthalpy ΔH^{Θ} refers to the internal energy of the solute and the solvent and reflects the standard enthalpy change of the system during the binding process. Breaking and forming nonbonded interactions during association induces the change in enthalpy. ΔH^{Θ} is primarily composed of van der Waals and electrostatics interactions and hydrogen bonds. For increased affinity, the binding enthalpy of the complex receptor-ligand tends to be maximised. To maximise the enthalpy gain, it is necessary to maximise the ligand-target interactions.

The change in entropy reflects the change in terms of disorder of the system (receptor, ligand, solvent), which is related to the degree of freedom of the system and also derived from the solvation/desolvation of both receptor and ligand. A flexible molecule in solution will lose a greater number of degrees of freedom while interacting to the target, which is unfavourable for the binding in term of entropy. However, the molecule has to adjust its conformation to a higher energy state inducing the molecule to be strained and so it is enthalpically disfavoured. Therefore there is two effects on ligand binding, the ligand conformation are restricted causing an entropic disfavoured but also it could be forced to adopt a new geometry when binding to the target inducing an enthalpy

disfavoured. The rigidification (or pre-organisation) of ligands is a common approach in structure-based design to overcome this effect (see later).

So, there is a contribution to the change in binding free energy between non-covalent interactions from the change in standard enthalpy ($\Delta H^{\ominus}$) and the desolvation effects and solvent reorganisation related to change in standard entropy ($\Delta S^{\ominus}$) and it is known as enthalpy-entropy compensation. However, it is too simple to think of one effect is just enthalpic and one entropic, both effects are coupled. Changes in non-covalent interactions predominantly affect enthalpy. Solvent and ligand and protein reorganisation affect enthalpy and entropy. Desolvation/resolvation affect enthalpy and entropy. Therefore, modifying a molecule to increase its enthalpy of binding will generate a decrease in entropy of binding and so the energy of binding will stay the same.

1.3.2 Enthalpy-entropy compensation:

The phenomenon of entropy-enthalpy compensation refers to the change in enthalpy associated by a change in entropy resulting from a ligand modification during a binding process. Therefore if a ligand modification induces an increase in favourable interactions with the substrate, inducing a more negative enthalpy change $\Delta H^{\ominus}$, it will likely lead to increase ligand/substrate rigidity and so a decrease in the entropy $\Delta S^{\ominus}$ of the complex. Numerous studies have shown that the relation between these two terms is linear [26-30]. The concept of enthalpy-entropy compensation was thought to be a “phantom phenomenon” for a long time as explain by Athel Cornish-Bowden in 2002 [31]. Isothermal Titration Calorimetry (ITC) enabled us to measure the relative binding entropy and enthalpy separately to elucidate the thermodynamic contributions to the Gibbs binding free energy and entropy-enthalpy compensation is no longer consider as an artefact. The use of ITC to obtain thermodynamic data on the binding of a ligand to its target was introduced by Ladbury et al [32]. Many studies have shown the importance of enthalpy/entropy optimization in the enhancement of binding affinity in molecular design [33-39]. The introduction of ITC method brought the thermodynamic evidence to support potency enhancement of ligands toward their targets.

A recent study in 2011 measured thermodynamic data for 100 ligand-protein complexes from different target classes to evaluate the effect of enthalpy and entropy to the binding [40]. As seen in many different studies, protein-ligand complexes with higher negative enthalpy of binding will show a more positive

value of the $-T\Delta S$ and vice versa as we can see in *Figure 1.3*. *Figure 1.3* shows the correlation between the enthalpy ($\Delta H^{\ominus}$) and the entropy ($-T\Delta S^{\ominus}$) from the 100 complexes. The colours describe the different target classes. From this plot an evident correlation between these two terms can be observed.

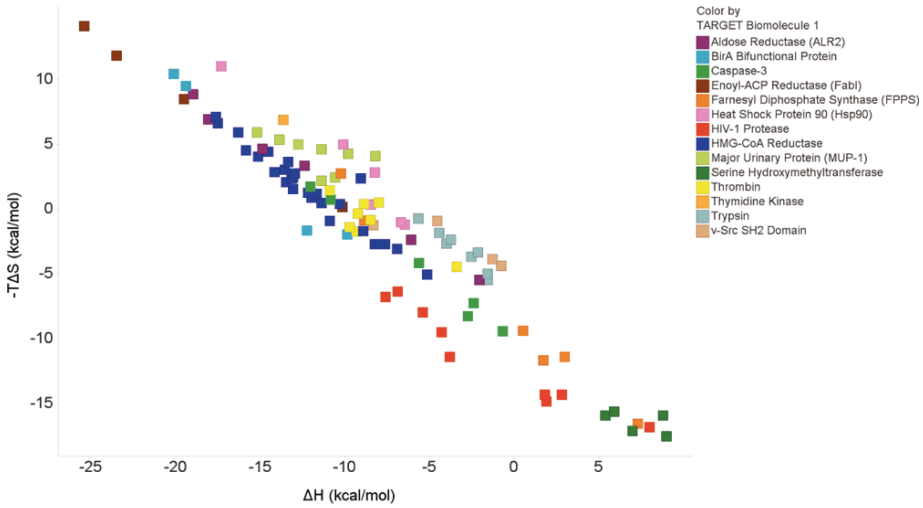


Figure 1.3: Correlation between enthalpy (ΔH) and the entropy ($-T\Delta S$) [40].

In addition, the study emphasized the fact that the size of the ligand will influence the enthalpy and entropy efficiencies. To illustrate this observation, they plotted the enthalpy and entropy efficiencies against the number of heavy atoms in the ligands *(Figure 1.4)*. From *Figure 1.4*, on average the enthalpy reflects the binding affinity for small molecules *(Figure 1.4a)* compared to the entropic term that does not change much with the molecular size *(Figure 1.4b).*

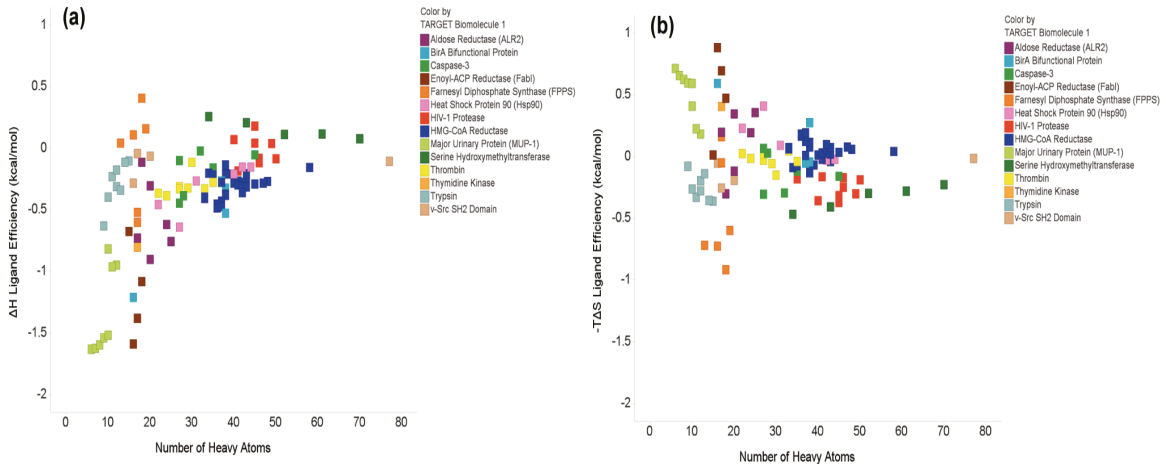


Figure 1.4: Enthalpy (a) and entropy (b) efficiency against ligand molecular size [40].

One common design method to enhance the change in entropy on binding and consequently improve the ligand binding affinity is known as preorganization.

1.3.3 Preorganization:

The complementarity of molecular interactions between a ligand and its receptor plays a major part in their mutual recognition. To reduce the binding free energy, Cram introduced the notion of preorganization defined as follow: “the more highly hosts and guests are organized for the binding and low solvation prior to their complexation, the more stable will be their complexes” [41]. The concept of preorganization means that the ligand can be designed for an optimal complementarity to bind to the receptor active site. Making the ligand more rigid decreases the entropy loss while binding to the receptor. According to the preorganization concept, a rigid molecule in a particular conformation facilitates the complex formation by increasing the binding affinity thus resulting in smaller entropic penalty.

In *Chapter 2*, the influence of preorganization in term of ligand efficiency is discussed in more detail. In this thesis we explore the concept of conformational preorganisation with particular reference to an enzyme involved in fatty acid biosynthesis.

1.4 Introduction to the system:

1.4.1 Fatty acids biosynthesis:

1.4.1.1 Introduction:

A fatty acid is a term which describes a range of carboxylic acids with an aliphatic tail that are either saturated or unsaturated. Human fatty acid synthesis takes place in the cytosol of human cells whereas their degradation by β-oxidation occurs in the mitochondria [42]. In humans, the primary site of the fatty acid synthesis is the liver and in the adipose tissue. The biosynthesis of fatty acids satisfies two main requirements of the cell. First, fatty acid biosynthesis is an essential metabolic pathway of the lipogenesis in eukaryotic and bacterial cells. Lipogenesis describes the process of formation of lipids as triglyceride using free fatty acids and glycerol as substrates. Fatty acids are significant energy resource for the human body. Indeed, in response to energetic excesses, triglycerides appear as efficient energy storage for the organism [43]. Thus, fatty acids are involved in numerous cellular processes, from major components of cellular membranes, energy storage as triglycerides to precursors for the generation of signalling molecules [42,44-45].

The synthesis as well as the oxidation of fatty acid involves the intermediate, Acetyl-CoA. It is the source of all carbon atoms of fatty acids and is formed in the mitochondria. Fatty acid biosynthesis requires the oxidation of the NADPH co-factor and human fatty acid synthase (hFAS) enzyme control the synthesis pathway.

1.4.1.2 The human fatty acid synthase complex:

Human fatty acid synthase (hFAS) is a large homodimeric multienzyme complex of 552 kDa that regulates the biosynthesis of long-chain fatty acids [46]. The grouping of the different catalytic activities into one protein improves the efficiency of all the synthesis steps. Each monomer is composed of a polypeptide chain endowed with seven distinct enzyme domains (malonyl-acetyl CoA transferase MAT, β-keoacyl synthase KS, β-ketoacyl reductase KR, β-hydroxyacyl dehydratase DH, enoyl reductase ER and thioesterase TE) bound to an acyl carrier protein (ACP) [47-49]. All of the intermediates of the synthesis pathway remain covalently bound to the ACP thiols -SH group until the fatty acid chain is long

enough to be released. hFAS catalyses the reaction of acetyl CoA plus 7 malonyl CoA to yield to the palmitate fatty acid. *Figure 1.5* shows the structure of human fatty acid synthase.

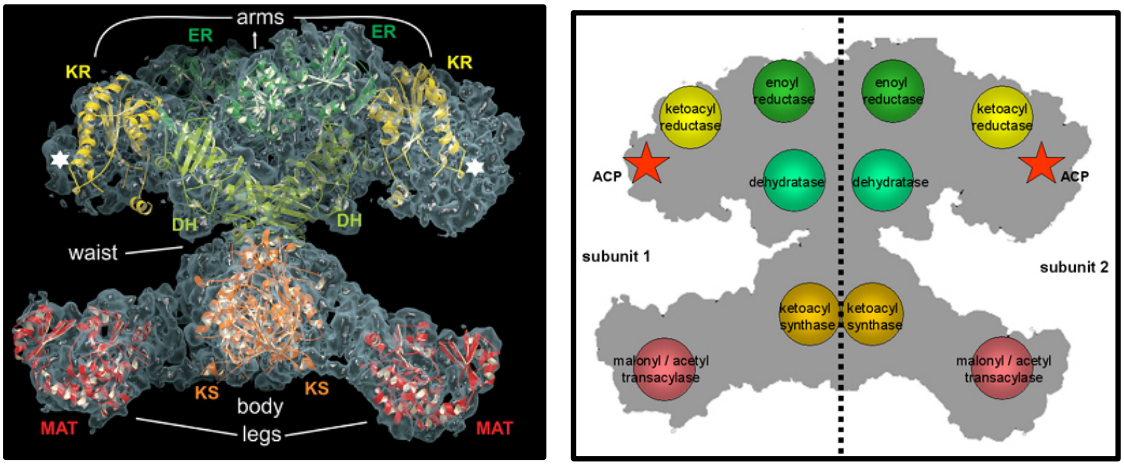


Figure 1.5: Structure of hFAS and its different catalytic domains

(pdb code: 1pqw) [48].

1.4.1.3 Steps of the fatty acid synthesis:

The biosynthesis of the fatty acids takes place in three steps, activation, elongation and termination [42]:

The **activation** step corresponds to the formation of malonyl-CoA by carboxylation of acetyl-CoA. It is a reversible reaction catalysed by the acetyl-CoA carboxylase (ACC) using ATP as co-factor. The acetyl-CoA carboxylase is an allosteric enzyme activated by citrate. To be transported from the mitochondria to the cytosol where the biosynthesis takes place, acetyl-CoA reacts with oxaloacetate to give citrate. A tricarboxylate translocase transports citrate from the mitochondria to the cytosol where the citrate is cleaved back to oxaloacetate and acetyl-CoA. This reaction is called the Citrate Shuttle. The reaction scheme of malonyl-CoA formation is as follow:

$$\mathrm{Acetyl-CoA + ATP + HCO_3^- \xrightarrow{ACC} Malonyl-CoA + ADP + P_i + H^+}$$

The next step, **elongation**, acts in a cycle with four processes (condensation, reduction, dehydration and reduction), which are repeated until a fatty acid is synthesized of the length required. Each cycle add two carbons units to the growing fatty acid resulting in the synthesis of the fatty acid palmitate (16

carbons). The reactions take place within the FAS complex. The intermediate's acyl groups are linked to the carrier portion of hFAS, the ACP, allowing them to enter the fatty acid synthesis cycle. The elongation phase starts with the transfer of acetyl-CoA and malonyl-CoA to ACP catalysed by the acetyl/malonyltransferase. Then the four reactions can start, condensation, reduction, dehydration and reduction, leading ultimately to the palmitate C_{16} fatty acid. The elongation phase is extremely energy consuming with 8 acetyl-CoA, 7 ATP and 14 NADPH consumed for one palmitate molecule synthetized.

The **termination** step consists of releasing of the newly formed fatty acid from the enzyme. The palmitate molecule is now the precursor for additional elongation or unsaturation to yields to other fatty acids molecules.

1.4.2 Human fatty acid synthase and drug development:

In human normal cells and tissues, hFAS expression is low and mainly compensated with the food we eat. However, many studies have shown that hFAS in cancer cells is overexpressed [50-51]. Excessive activation of hFAS induces an increased rate of de novo lipids biosynthesis, important components of cell growth and proliferation. The lipids newly synthesised are then used as constituents of biological membranes conferring a growth and survival environment for tumour cell proliferation. It has been suggested that multiple cancer cell lines depend on hFAS for proliferation and survival [52-54]. Thus, hFAS overexpression has been observed in various types of human tumours such as breast, prostate, thyroid, bladder, oesophagi, lung, colon, ovary, kidney, stomach, tongue cancer [55-57]. hFAS, in this manner, appears as a potential therapeutic target in cancer treatment. A number of pharmacological inhibitors of lipid biosynthesis have been identified [58-62] some of which are in clinical testing. However, early hFAS inhibitors showed pharmacological limitations that restrict their use as their modes of action are not fully understood. As example of these inhibitors is cerulenin and C75 in breast cancer [63-68], orlistat shown antiproliferative activity against prostate cancer cells [69] and C93 was developed to overcome the lack of potency and side effects of C75 [70-71]. Among these issues was irreversible activity (C75, orlistat), high chemical reactivity responsible for low specificity, low cell permeability and solubility (orlistat), low chemical stability (cerulenin), poor selectivity and oral bioavailability (orlistat) [67,66]. In an attempt to improve the limitations of these compounds, several studies through high-throughput screening or medicinal chemistry programmes identified new

potent inhibitors of hFAS. Examples of these studies came from the Merck and AstraZeneca groups where they developed a series of 3-aryl-4-hydroxyquinolin-2(1H)-one derivatives and bisamide derivatives hFAS inhibitors respectively [67-72]. The GlaxoSmithKline group also reported a potent, reversible inhibitor of the KR domain of hFAS, one example known as GSK837149A [60]. The catalytic KR domain of hFAS induces the reduction of acetoacetyl-ACP to D-3-Hydroxbutyryl-ACP in presence of the reduction agent NADPH and β-ketoacyl reductase enzyme.

A more recent study has reported the discovery of a new highly potent hFAS inhibitor, GSK2194069, showing reversible and specific activities toward the KR domain of the enzyme [62].

1.4.2.1 Discovery of GSK2194069:

In 2008 a study by GlaxoSmithKline performed a high-throughput screen on a collection of hFAS inhibitors monitoring NADPH consumption [60]. The study reported the first selective inhibitor of human FAS enzyme, GSK837149A with a pIC_{50} of 6.0 *(Figure 1.6)*. It acts reversibly on the KR domain and is competitive with respect to NADPH and uncompetitive with respect to acetoacetyl-CoA (substrates of hFAS). It was suggested that the inhibitors and the substrates followed a compulsory ordered kinetic mechanism with the acetoacetyl-CoA first to bind to hFAS, followed by GSK837149A forming acetoacetyl-CoA – hFAS – GSK837149 complex. However, GSK837149A has shown poor cell permeability.

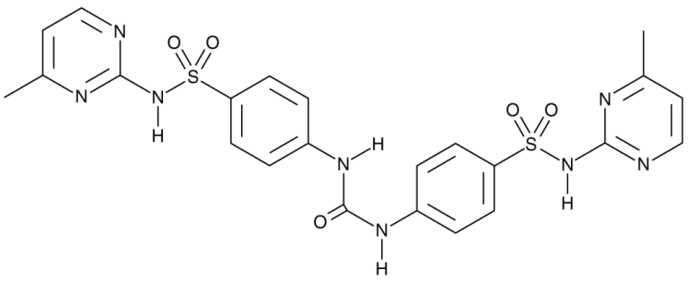


Figure 1.6: Chemical structure of GSK837149A [60]

Further optimization of the GSK837149A compound lead to the introduction of the benzimidazole, cyclopropylamide carbonyl and biphenyl moieties [73].

Additional studies led to the replacement of the benzimidazole to monocyclic core and the exploration of triazolone moieties due to the improvement of ligand efficiency [74-76]. These enhancements lead to the discovery of the first x-ray crystal structure (pdb code: 4PIV) of a highly potent, specific and reversible

inhibitor of the KR domain of hFAS, GSK2194069 with a pIC_{50} of 7.5 *(Figure 1.7)* [62]. GSK2194069 is competitive with respect to the substrate inducing the inhibition of the acetoacetyl-CoA reaction and in comparison with GSK837149A, the mechanism of inhibition of the triazolone GSK2194069 is uncompetitive with respect to the NADPH leading to a potency improvement. Both the protein and NADPH form part of the binding pocket for the inhibitor. The reported x-ray crystal structure was resolved in the presence of NADPH and GSK2194069 in the KR domain of hFAS and solved at a resolution of 2.3 Å enabling elucidation of the binding mode.

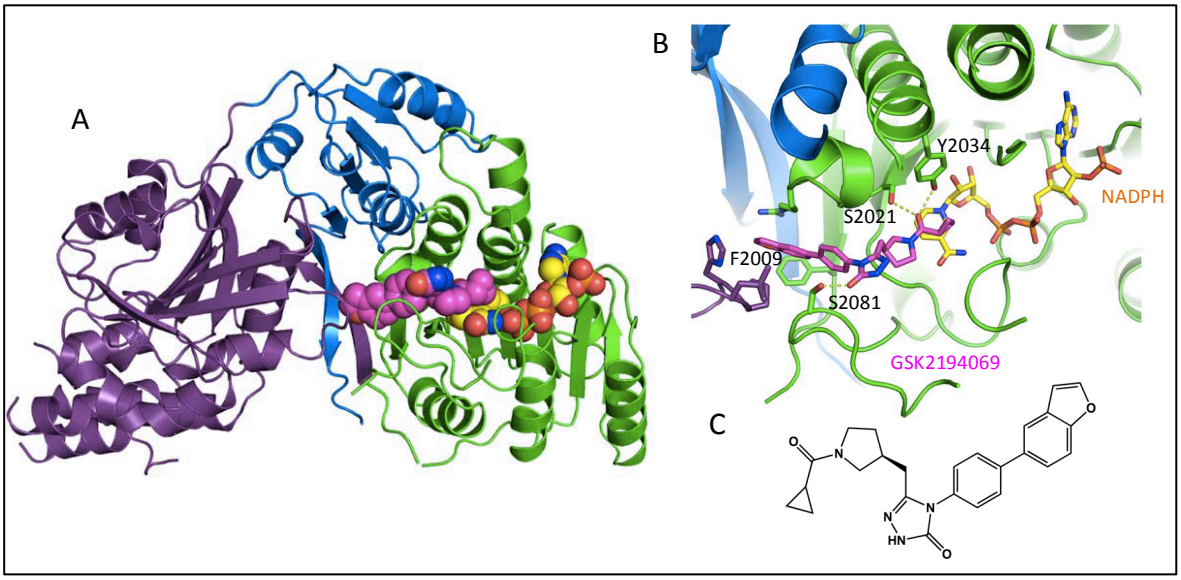


Figure 1.7: (A) GSK2194069 compound (magenta) and NADPH (orange) in the KR domain (orange) of hFAS (B) Zoom view of GSK2194069 and NADPH illustrating the active site interactions(C) Chemical structure of GSK2194069 [62].

Thus, GSK2194069 is a highly potent, reversible and selective inhibitor of the hFAS KR domain. The evidence of its inhibitory activity toward tumour cells was shown after a decrease of new lipid synthesis [62].

1.4.2.2 Binding mode information from GSK2194069:

The published x-ray crystal structure of GSK2194069 bound to the KR domain of hFAS has provided information about the binding mode and mechanism. Various interactions were reported and are shown in *Figure 1.8*. Two hydrogen bonds are

formed with the cyclopropyl carbonyl oxygen and the residues serine and tyrosine (Ser2021 and Tyr2034 green arrows). The oxygen of the triazolone core forms a hydrogen bond with the residue Serine (Ser2081). Additionally, the benzofuran moiety makes an arene-π interaction with the side-chain of phenylalanine (Phe2109).

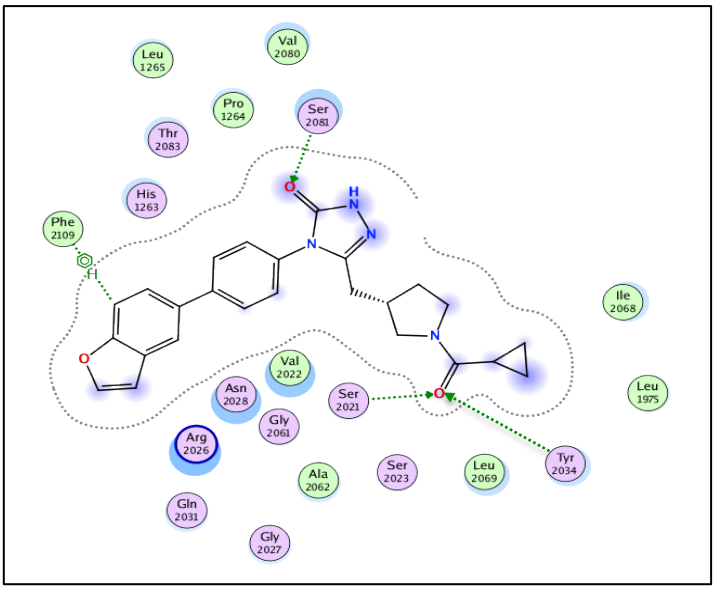


Figure 1.8: Interactions between GSK2194069 and hFAS receptor binding-site. The greens arrows represent the residues side chains acceptor hydrogen bonds [62].

hFAS multienzyme is implicated in the progression of numerous tumours. Improving and identifying novel, potent inhibitors of hFAS will potentially deliver a more effective cancer treatment.

1.4.3 Novel fatty acid synthase inhibitors:

Within this thesis, the chemical structures and activity of four inhibitors of hFAS, including GSK2194069, have been provided to us by AstraZeneca. Collaboration with the drug discovery and development company, C4X Discovery, provided us with NMR experimental data in solution of the conformations adopted by these compounds.

1.4.3.1 The C4X Discovery Company:

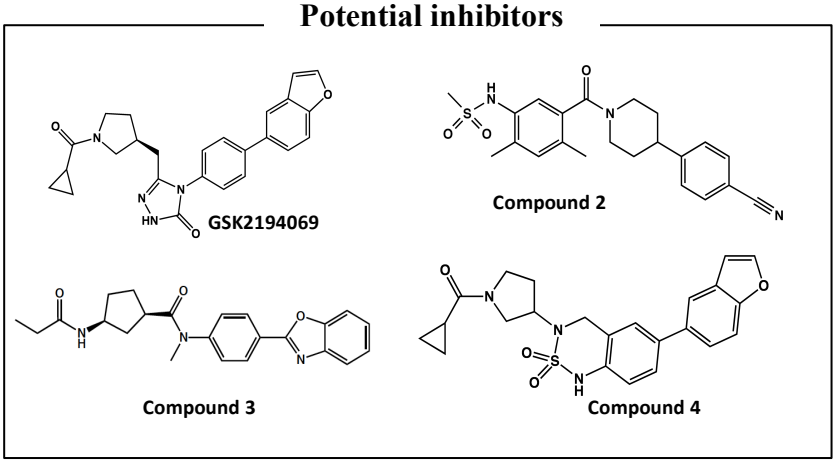
C4X Discovery is a drug discovery and development company spin out from the University of Manchester in 2008 [77]. It aims to speed up and improve the drug discovery process to identify novel therapeutic small molecules. Two unique software platforms, Taxonomy3® and Conformetrix, have been developed to help understand and solve the 3D structures of a broad range of biomolecules including peptides, cofactors, oligonucleotides and carbohydrates in solution. The Conformetrix technology platform enables the dynamic 3D shapes of free drug molecules in solution to be precisely measured from NMR experimental data providing valuable information for drugs candidate design and optimisations, showing the important shapes active molecules prefer to adopt in solution [78].

C4X Discovery works in partnerships with multiple global and international companies including AstraZeneca in the development of new medicines [78-79] as well as biotech firms and academics. In July 2016, C4X has discovered multiple novel drug targets for development in the areas of inflammation and neurodegeneration.

1.4.3.2 NMR data for potential inhibitors:

The structure and the potency of the four potential hFAS inhibitors are shown in *Table 1.1*. The structure of each compound was determined in a mixture of Dimethyl sulfoxide (DMSO) and H_2O ensuring no structural differences between optimum conditions and physiological conditions.

Potential inhibitors



Compound	Potency (pIC_{50})	Solution conditions
GSK2194069 (compound 1)	7.5	- 90% DMSO ; 10% H_2O at 278°C - pH: 6.5
2	6.8	- 90% DMSO ; 10% H_2O at 278°C - pH: 6.5
3	7.9	- 25-100% DMSO ; 75-0% H_2O at 278°C - pH: 6.5
4	4.5	- 80% DMSO ; 20% D_2O at 278°C - pH: 7.5

Table 1.1: hFAS inhibitors structures, potency and NMR experiments conditions.

1.5 Purpose of the thesis:

Conventional molecular dynamics simulations (MD) can be used to obtain the detailed description of a molecular system. However, the complex energy landscape of biological molecules often induces trapping in local minima while using molecular dynamics. A common approach to overcome the trapping issue is to use an enhanced sampling method. In the purpose of this thesis, Replica Exchange Molecular Dynamics Simulation (REMD) has been performed on four inhibitors of human fatty acid synthase. This project is in collaboration with the pharmaceutical company AstraZeneca. Nuclear Magnetic Resonance (NMR) data have been provided from C4X Discovery Company, giving us information on the conformational equilibrium of these compounds in solution. The aim of the project is to understand the conformational ensemble of these inhibitors in solution and to determine the extent of their preorganisation. An x-ray crystal structure of an inhibitor, GSK2194069, bound to the KR domain of human fatty acid synthase receptor is available and provided us with information on the binding mode (pdb code: 4PIV).

Two chapters will focus on the comparison between the computational and experimental methods and with the Cambridge Structural Database to evaluate if modelling is able to generate reliable solution phase conformational distribution. The following chapter will focus on the conformations of the compounds in solution by identifying relevant conformations for the binding. Finally, the last chapter, by using the knowledge of the x-ray binding mode, will predict the binding mode for each compound.

Chapter 2 Ligand preorganization

2.1 Introduction:

A major challenge in drug discovery is to improve the binding affinity of an inhibitor to its target. As introduced in the previous chapter, one popular strategy used is through the rational design of ligands to be preorganised for their binding pocket as presented by Cram in 1986 [41].

Cram first brought the principle of "preorganization" to his work on the design of small-organic host-guest complexes. The free energy of complexation was measured to assess the importance of molecular preorganization and complementarity. His work emphasized the idea that a preorganized host or ligand molecule will reduce the entropy and enthalpy cost necessary for binding. He demonstrated that a rigid host will induce a higher affinity with an appropriate guest and he also highlighted the importance of macrocyclization to achieve this. This work was focused on the preorganization of the host. In the case of drug discovery, only ligand preorganization can be considered as a design strategy. The aim of ligand preorganization is to reduce the flexibility of a ligand by pre-stabilizing its bound conformation in order to decrease the entropic penalty through complex formation leading to a binding affinity improvement.

Numerous studies have demonstrated that rigidifying a ligand will induce a binding affinity enhancement [80-84]. However, if this rigidification is not done in a way complementary to the binding mode significant losses in potency will be observed this may not be necessary due to affinity.

A number of preorganization strategies can be observed including macro-cyclisation to decrease flexibility and the fixing of rotatable bonds through the addition of polar and/or non-polar groups. In this chapter, different studies that show the effect of ligand preorganization on complex binding, and how some commonly assumed paradigms appear to not always be true will be presented in the next sections.

2.2 The effect of macrocyclization:

When a ligand binds to its receptor it invariably becomes less flexible. In this manner, a common approach to improve the binding affinity is to constrain the ligand by inclusion into a larger ring that will stabilize the overall conformation.

Macrocyclics are found in natural products and have broad applications in drug discovery and development thanks to the favourable proprieties including improvement in the binding affinity and metabolic stability, good solubility and oral availability, enhancement of the membrane permeation with desirable pharmacokinetic (PK) properties [82-83,85-86]. However, the synthesis of macrocyclic drugs appears to be challenging and several methodologies are used [87-89].

In the case of peptides, the success of the macrocyclization was observed in several studies. The common approach is to insert a ring into the peptide chain by the formation of a bond between a side chain and a backbone atom, or between two side chains or between the N and C terminal of the peptide [83].

To illustrate the effect of this strategy, studies of the effect of the macrocyclization in term of ligand affinity and potency, selectivity profile and PK properties are presented in the following section

2.2.1 Binding affinity and potency improvement:

Initial modelling studies of a peptide inhibitor of hepatitis C viral NS3 protease had emphasized the issue of conformational flexibility of peptide inhibitors that prevent the design of compounds that could penetrate the cell membranes and inhibit the protease [90]. To overcome this problem, several modifications have been performed using NMR spectroscopy, crystallography and molecular modelling, leading to the design and synthesis of highly potent and specific macrocyclic inhibitors of the HCV NS3. The study has shown that macrocyclic inhibitors have greater potencies (IC_{50} = 400 µM) than the linear one (IC_{50} > 1000 µM) which not necessary means a greater binding affinity and better preorganization. No binding affinity comparison between cyclic inhibitors and acyclic analogues has been shown in the study.

Another example comes with inhibitors of Penicillopepsin [91]. A macrocyclic peptidyl phosphatane inhibitor was derived and synthesis from an acyclic Penicillopepsin inhibitor using NMR and molecular modelling (Monte Carlo conformational analysis). The macrocyclic form of the inhibitors improved the binding affinity with constant inhibition K_i of 0.10 nM for the cyclic form against K_i of 1300 and 42 nM for its acyclic analogues.

A recent study in 2016 reported the discovery of non-natural peptides Mcl-1 inhibitors using macrocyclization [92]. This study demonstrated that linking the two homophenylalanine end groups of the molecule into a macrocyclic enhances

the binding potency (acyclic IC_{50} = 2 μM; cyclic IC_{50} = < 3 nM). Crystallographic studies also provided the insight of the binding mode enabling the ligand optimisation to preorganize the structure toward the bound conformation.

2.2.2 Selectivity and PK profiles:

The study of kinase inhibitors in 2011 and 2013 revealed the discovery of an inhibitor of pan-CDK/FLT3/JAK2 with balanced potency and ADME properties leading to a candidate for phase I clinical development [93-94]. Docking of lead compounds obtained from HTS revealed a semi-circular shape of the binding compound. Thus, the structure-based design (SBDD) of novel small molecules using macrocyclization was discussed and led to the design and synthesis of novel small-molecules macrocycles. Investigations of the interactions made with the kinases revealed the importance of hydrogen bonds and salt bridges in the selectivity. Studies of PK profile and dose-dependent efficacy also revealed good oral bioavailability.

Another recent study in 2016 presented the synthesis of three chemical series macrocyclic MTH1 inhibitors. These inhibitors demonstrated potent and selective activities by making hydrogen bonds [95]. *Figure 2.1* shows an example of the compounds macrocyclization synthesis with the inhibitor 15. The different reaction conditions that led to the macrocyclic inhibitor can be found in the paper [95].

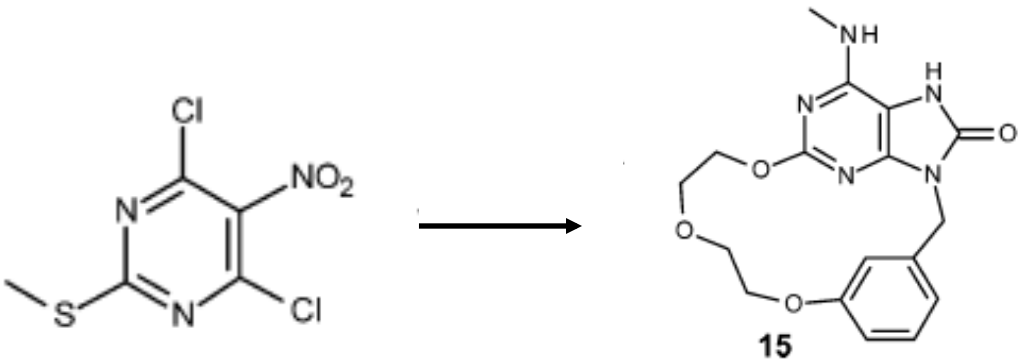


Figure 2.1: Macrocyclization of the inhibitor 15 [95].

These studies have shown that modifying a peptide or small organic molecule to a macrocyclic, if done appropriately, can improve the binding affinity as well as PK properties. However, only binding affinity was disclosed, no thermodynamic data were taken into account. Thus, although the ligand potency

was improved the reason for this was not fully evaluated. To assess the thermodynamics parameters upon complexation, it is necessary that the preorganized and the control ligands must interact with the solvent and the target in similar manner. Furthermore, structural information on the flexible and constrained ligands bound to the receptor is not always available which makes the prediction of the contributing binding enthalpy and entropy harder to obtain. Isothermal titration calorimetry (ITC) provides information on the thermodynamic binding parameters, enthalpy, entropy and free energy from a single experiment (see section 2.3) [32].

2.3 The energetic consequence of macrocyclic preorganization:

Studies of Src SH2 domain binding ligands have shown the impact of a constrained ligand comparing to a flexible ligand on binding affinity [82,97-99]. *Figure 2.2* shows the preorganized and control ligand structures; the constraint appears in the form of a carbon-carbon bond.

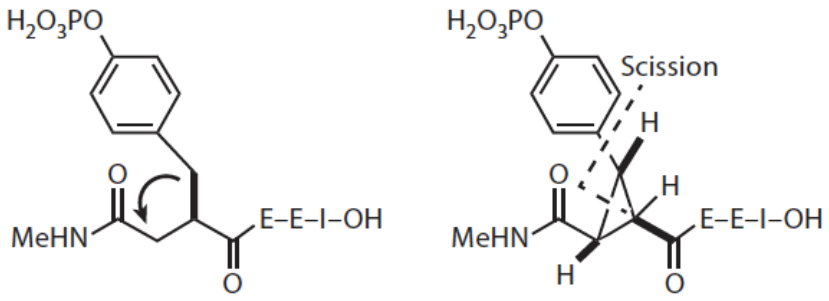


Figure 2.2: Src SH2 domains flexible and preorganized ligands [82].

The preorganized ligand binds to the domain with a more favourable entropy change than the open chain ligand. However, due to the enthalpy-entropy compensation, the improvement in the entropy change is associated with a less favourable enthalpy, which makes both ligands equipotent in terms of the overall binding affinity. To understand the origin of the enthalpy-entropy compensation, a combination of NMR, molecular dynamics simulations (MD) and crystallography on the Src SH2 domain were conducted in 2010 [100]. The studies revealed that the constraint applied to the ligand induced a modification of the geometry of the critical interactions in the binding pocket leading to the loss in enthalpy. They

suggested inserting the constraint in another part of the ligand to overcome the enthalpy penalty.

Another study in 2010, showed the structural and energetic effect of introducing a macrocyclization to stabilise the biologically active conformation of Grb2 SH2 binding peptides using Isothermal Titration Calorimetry (ITC) and x-ray crystallography [101]. However, during this study, the authors discovered that the peptide preorganization could have either favourable or unfavourable entropic effect. To illustrate this issue, the results of two macrocyclic peptides and their corresponding acyclic control on the inhibition of the Grb2 SH2 domain are presented *(Figure 2.3).*

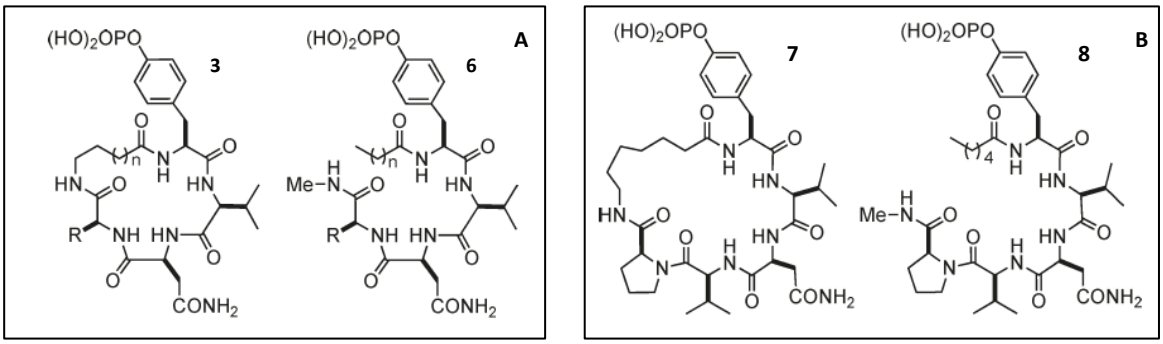


Figure 2.3: Macrocycle peptide 3 and its acyclic control 6 (A) and macrocyclic peptide 7 and its acyclic control compound (B) [101].

Figure 2.3 shows the structure of the macrocyclic peptides 3 and 7 and their corresponding acyclic controls 8 and 6 *(orange in table 2.1)* respectively. In the paper it has been shown that both macrocyclic peptides exhibit higher affinity for the Grb2 SH2 domain comparing to the linear peptides *(Table 2.1)*. However, by comparing their binding energy an interesting observation can be seen. Indeed, the macrocyclic compound 3 shows a more favourable enthalpic term than its acyclic control compensated by a less favourable entropic term *(Table 2.1)*. For compounds 7 and 8, the opposite effect is observed with a more favourable entropy for the macrocyclic compound 7 than its acyclic control 8 compensated by less favourable enthalpy than for 8 *(Table 2.1)*. This study emphasised the fact that ligand preorganization does not always induce a favourable entropy gain but both entropy and enthalpy need to be taken into account.

Compounds	K_a (M^{-1})	$\Delta H°$(kcal.mol^{-1})	$\Delta S°$ (cal.$mol^{-1}.K^{-1}$)
3	8.5 ± 0.03 * 10^5	-6.3 ± 0.68	5.9 ± 0.37
6	5.7 ± 0.65 * 10^5	-4.8 ± 0.44	10.2 ± 1
7	1.0 ± 0.20 * 10^7	-4.3 ± 0.57	17.7 ± 1.20
8	6.5 ± 0.13 * 10^6	-6.3 ± 0.38	9.9 ± 0.17

Table 2.1: Thermodynamic data and binding affinity for macrocyclic compounds and their acyclic analogues obtained by ITC at 278 K.

Two studies in 2006 have shown the effect of the permeability modulation with the formation of internal hydrogen bonds network in cyclic peptides [102-103]. The cyclization of the peptides eliminated the charges of the N and C termini of peptides making them inaccessible for solvent binding and promotes internal hydrogen bond formation and facilitates membrane permeation by preventing the backbone N-H groups to forms interactions with the solvent. Thus, the hydrogen bonds promote passive membrane permeation by reducing the energetic cost of amide N-H desolvation while entering the membrane.

These studies highlight the challenge of correlating the energetic change with the structure of the ligand-target interactions.

2.4 Knowledge of the unbound ligand:

The two previous sections emphasized ligand preorganization involving the introduction of a conformational constraint via macrocyclisation in the effort of increasing the ligand affinity. This approach limits the conformational variability and thus reduces the entropy penalty when the preorganized ligand binds to the target. However, understanding the different conformational changes that a ligand can adopt in solution could enable the design of ligand with higher affinity [104-105]. Additionally, when x-ray crystallography data of a molecule bound to its target is not available, the knowledge of the unbound ligand conformations in solution could facilitate the design of preorganized ligands toward the bioactive conformation. Therefore, two studies using NMR and Molecular Dynamics are presented in the following paragraphs.

C4X Discovery *(Chapter 1 section 4.3.1)* introduced a new approach to maximise the conformational knowledge of flexible ligands using NMR spectroscopy and

mathematical models to describe the conformations of unbound ligands in solution [77,106]. The molecules are analysed using NMR and the results are interpreted according to the method described in the paper [78]. This new NMR method provides an accurate description of the 3D conformations a ligand can adopt in solution as well as the population of each conformation. Successful applications of this method have been reported [79] such as the study of the aminoglycoside streptomycin [78]. This study revealed 12 conformational macrostates that are grouped into two main conformations, and the comparison with crystallography data confirmed the similarity of one of the main groups to the bioactive conformation.

A study in 2016 [107] investigated the conformational ensemble and dynamics of pharmaceutical compounds using molecular dynamics simulations. Understanding the conformations an unbound ligand can adopt in solution will help to evaluate the conformational energy penalty upon binding. They used 26 diverse molecules with known bound bioactive x-ray structures to compare the bound and unbound conformations. The MD simulations were performed in explicit solvent and the average internal energy of the bound and free compound were calculated from the simulations and their difference gave an estimation of the enthalpic energy contribution due to ligand strain when bound (ΔH). They found that the majority of the studied compounds have a low ΔH (< 6 kcal/mol).

The use of NMR and MD methods to obtain an accurate representation of the unbound conformational ensembles have shown promising results to help ligand preorganization to its bioactive shape. However, further work in this area is required.

2.5 Conclusion:

Throughout this short review, different studies of how ligand preorganization plays an important role in the development and design of molecules that bind with higher affinity together with how this principle has been used were presented. Different points have been emphasised here.

Primarily, the challenge is to correlate the structural information of the ligand and target to the thermodynamic effects during the complexation. Two tactics to preorganize a ligand in order to increase the binding affinity are seen in

constraining a molecule by the formation of a ring and molecule optimisation from the knowledge of unbound ligand conformations. Macrocyclic ligands have shown improvement in potency over their acyclic counterparts. Pharmacokinetic and ADME properties can be modulated through cyclization promoting selectivity and membrane permeation with the formation of hydrogen bonds.

Additionally, the common belief that a preorganized ligand will induce a lesser entropic penalty on binding is not always true even if the constrained ligand presents a higher binding affinity than its linear control. In ligand preorganization it is necessary to remember that both enthalpy and entropy are important in the binding. It is difficult to show how preorganization affects the enthalpy and entropy contribution of the ligand-protein complex because structural information on the two protagonists are often lacking as well as a structure of a flexible control and constrained ligand with similar components (heavy atoms, functional groups, hydrogen bond donors and acceptors).

Furthermore, NMR and MD studies have shown the importance of understanding the conformational ensemble a free ligand can adopt in solution in addressing ligand preorganization. The combination of both methodologies can provide richer knowledge on the unbound ligand and will be discussed in this thesis.

Chapter 3 Methodology

3.1 Introduction:

Elucidating the structure of small molecules to better study the relationship between structure and impact on the biological function of a target is very important. To determine the 3D structures of small molecules a number of experimental techniques such as Nuclear Magnetic Resonance (NMR) and x-ray crystallography are available. NMR experiments enable the dynamics of small molecules to be explored providing information on internal motions from the picosecond to millisecond timescales in solution states [108-112].

Molecular dynamics simulations (MD) are a powerful computational tool, which can enable the understanding of the structure and function of both small and large biologically relevant molecules. Simulations of hundreds of nanosecond and multi-microseconds are common and can directly be compared with appropriate experimental data such as x-ray crystallography. MD also provides a complete set of atomic level insights of a solvated system, which are difficult to obtain from experimental data, offering a more detailed analysis of the structure and dynamics of a system [18,113-114].

Therefore, combining experimentally derived data and molecular dynamics simulations should increasingly enable greater insights and description of the dynamic motions of biological systems. A range of computational tools facilitate the analysis of the generated ensembles such as clustering, principal component analysis, root mean square deviation, hydrogen bond, torsion angle and conformational sampling analyses.

To complement the observations from MD and NMR, molecular docking has also been performed in this study. Only GSK2194069 has a reported x-ray crystal structure [62], no crystallography data are available for compounds 2,3 and 4. Therefore, molecular docking predictions can help to better understand the action and drug interactions mechanism where there is a lack of experimental data and enable us to understand the binding geometries of those compounds.

3.2 Nuclear Magnetic Resonance:

3.2.1 Theory:

Nuclear magnetic resonance is an experimental technique, which is commonly used in drug discovery to determine the structure and purity of synthesised small molecules. Indeed, NMR uses the magnetics properties of this nucleus within an intense magnetic field. Nuclear particles (protons and neutrons) behave as small magnets or nuclear spin and can be described using quantum numbers I for the spin and m for the quantum magnetic spin (m=2I+1). Atoms with uneven particles in their nuclei such as hydrogen, carbon, nitrogen, have a non-zero spin that enable them to absorb/emit electromagnetic radiation and so to be studied by NMR [116-117]. The charge of the rotated nuclei produces a magnetic field tie to a magnetic moment μ define as:

$$\mu = \gamma I \text{ ; with } \gamma \text{ the gyromagnetic ratio of the nucleus}$$

In the absence of magnetic field, the magnetic spins are randomly oriented. In the presence of a magnetic field B_0, the magnetic spins align within the direction of the magnetic field (Zeeman effect). In a magnetic field B_0, the spin rotates around B_0 axis. However, the rotation around B_0 axis cannot be exactly parallel or antiparallel to the B_0 direction but precess around the magnetic field with an angular frequency defines as $\omega_0 = \gamma B_0$ *(Figure 3.1).* This phenomenon is called Larmor precession.

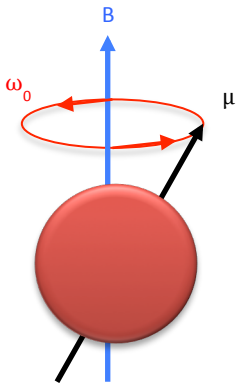


Figure 3.1: Larmor precession

According to quantum mechanical theory, spins of magnetic nucleus can take only defined orientations with the applied field. For atoms ^{1}H, ^{13}C, ^{15}N, ^{31}P corresponding to biological relevant nuclei with a spin of ½, two distinct magnetic states m= ½, and m= -½ can occur. Therefore, when B_0 is applied to these atoms, m can occupy two energy levels, a low level α (+½) and high level β (-½) with energy of spin $E = -\mu B_0$ *(Figure 3.2)*. The energy difference between the two states is defined as follow: $\Delta E = E(\beta) - E(\alpha) = \gamma B_0$ and depends on the magnetic field intensity applied. The more intense B_0 is, the greater the energy difference.

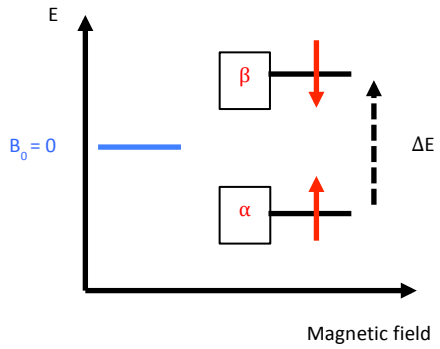


Figure 3.2: Energy levels for a spin= ½ nucleus in an external magnetic field B_0.

Therefore, when a biological system is placed in a magnetic field all protons begin to precess at the Larmor frequency ω_x. The magnetic moments m will then be aligned with and against B_0. Electromagnetic radiation in the radio frequency range is applied to excite the nucleus of the studied system and cause the orientation of the nuclear magnetic spins to flip from the low to the higher energy. The protons will thus absorb radiations at certain frequencies. The frequencies at which the radiation is absorbed as well as the intensity of each absorbance will then be recorded. Knowing that each proton resonance frequency is different according to their neighbouring atoms, the analysis of the different resonance frequencies will enable the structure of the studied system to be determined.

Therefore, NMR can provide information about molecules at the atomic level. NMR spectral parameters will enable the determination of the structure and dynamics of biomolecular systems based on chemical shifts, scalar coupling constant (J-coupling) and nuclear Overhauser effect (NOE) signals.

3.2.2 Chemical shifts:

An NMR spectrum contains signals or peaks corresponding to the protons resonance in the system. The signals are plotted on a horizontal axis and the location, the shape and the intensity of the peaks are analysed. The frequency of the signal is known as the chemical shift δ and reflects the position of the NMR signals in the spectrum. The chemical shifts show where the signals is shifted due to their chemical environment in comparison with the signal of some reference compound. Chemical shifts, as well as allowing structure determination of organic compounds, can also be used to determine the secondary structure of proteins.

The magnetic field B_0 induces circulations in the electron cloud around the nuclei. These circulations will produce an induced magnetic field B^{ind}, with the electrons acting as shield to protect the nuclei against the force of B_0 with a shielding factor σ: $B^{ind} = \sigma B_0$. The nuclei i will thus experience a local magnetic field B^{loc} as $B^{loc} = B_0 - B^{ind} = (1-\sigma)B_0$. The resonance frequency of each nucleus will, thus, be dependent on the electronic environment and it is called the chemical shift δ. To avoid having to deal with high values (MHz), the chemical shift δ is normalised with respect to a reference. The chemical shift is thus independent of B_0 and is measured in parts per million (ppm) as follow:

$$\delta_i = \frac{\nu_i - \nu_{ref}}{\nu_{ref}} * 10^6 \tag{3.1}$$

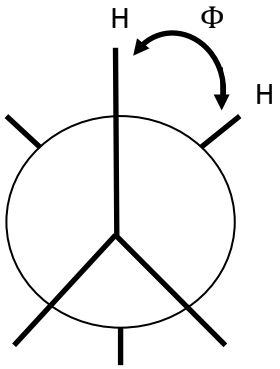
with ν_i the NMR frequency of the nuclear species and ν_{ref} the NMR frequency of a reference compound.
The most common reference compounds used are the tetramethylsilane (TMS) for referencing the ^1H and ^{13}C shifts and the liquid ammonia or nitromethane to reference the ^{15}N shifts.

3.2.3 Spin-spin coupling: J-Coupling

J-coupling can evaluate the intensity and the distance between the signals. Interactions between two nuclei can be observed in NMR spectra and is known as spin-spin coupling or J-coupling. The coupling between the two nuclei causes a split in the NMR signals into a small number of peaks with relative intensities and

characteristics spacing. The signal multiplicity depends on the neighbouring numbers of atoms involved in the coupling. Therefore, J-coupling reflects an interaction between pair of nuclei A and X through the intermediary of the electrons of N bonds. It corresponds to an indirect dipole-dipole interaction. The measurement of this interaction between two spins is done by the J-coupling constant J in Hz. The value of J can be either positive or negative and is dependent on the number, the bonding type between the nuclei and their chemical environment but independent of the magnetic field B_0. Additionally, the magnitude of the coupling constant will provide information on dihedral angles, hydrogen bonds and the type of coupled nuclei.

Indeed, the value of J depends on the molecule geometry making them very important in the structural study of chemical compounds. A particular case, $^3J_{HH}$ with 3 the number of bonds connecting the coupling nuclei H and H, the coupling constant is related to the dihedral angle ϕ by the semi empirical relation expressed by the Karplus equation *(Figure 3.3)*. When knowing the coupling constant $^3J_{HH}$, the dihedral angle H-x-x-H can be determined with x any atom.



Karplus equation

$$^3J_{HH} = A + B\cos\Phi + C\cos^2\Phi$$

With A, B and C constants

Figure 3.3: The Karplus equation

3.2.4 The nuclear Overhauser effect:

The NOE spectrum reveals couplings and coupling constant through space rather than through bonds as observed in J-coupling. A coupling will thus be observed if two nuclei are close to each other in space. NOE can be used to determine inter and intra molecular distances as well as the molecular geometry of a compound.

Two nuclear spins very close in space will induce magnetic dipole moment. It consists in magnetization transfer from one spin to another as a result of their dipolar coupling via cross relaxation. The rate, by which this effect is transmitted, is inversely proportional to the sixth power of the distance between the two nuclear spins, with a maximum distance of 5 Å [118-120].

3.2.5 The Conformetrix technology:

As mentioned in the introduction *section 1.5.3*, NMR data leading to the conformation ensemble in solution were provided from C4X Discovery Company. The Conformetrix technology platform from the C4X Discovery is used to determine the 3D conformations of biologically relevant biomolecules directly from NMR experimental data [77-79]. The 3D shape of the NMR ensemble used in the study was generated using the same methodology presented in the paper [78]. The technique uses multiple datasets and a dynamic model to match the NMR data. Conventional NMR methodology is used to calculated chemical shifts, J-coupling and NOE in various solvent and temperatures to obtain structural restraints. The NMR conditions, which are physically relevant for the 3D structure determination of our systems, can be found in *Chapter 1 Table 1.1*. Additionally, 3D structure ensembles are built by keeping bond lengths and angles fixed and rotating bonds according to a specific algorithm [78]. The newly ensemble is then used to make theoretical predictions of experimental data (Karplus equation, J-coupling, NOE) and compared against the experimentally measured data using a chi-square least-squared measure (X^2).

3.3 Molecular Dynamics:

3.3.1 Theory:

Molecular dynamics (MD) simulations are a theoretical method applied to simulate the behavior of molecules and molecular systems to obtain insight at atomistic detail and acquire descriptions of atomic and molecular interactions. MD simulations model the time evolution of a system of interacting particles from a set of initial positions and velocities. This method uses Newton's second law of motion for all particles in the system:

$$\overrightarrow{F_i} = m_i \overrightarrow{a_i} \tag{3.2}$$

where $\overrightarrow{F_i}$ is the force exerted on a particle i with a mass m_i and $\overrightarrow{a_i}$ is its acceleration.

The force can also be related to the potential energy V by the following equation:

$$\overrightarrow{F_i} = -\frac{dV_i}{d\overrightarrow{x_i}} \tag{3.3}$$

Where x is the position and V the potential energy of the particle i.

Moreover, the acceleration is related to the velocity $\vec{v}$ and as a consequence to the position x of a particle i:

$$\overrightarrow{a_i} = \frac{d\,\overrightarrow{v_i}}{dt} = \frac{d^2\overrightarrow{x_i}}{dt^2} \tag{3.4}$$

Combining these equations with the Newton law's yields:

$$\overrightarrow{F_i} = m_i\,\overrightarrow{a_i} = -\frac{dV_i}{d\overrightarrow{x_i}} = m_i\frac{d^2\overrightarrow{x_i}}{dt^2} \tag{3.5}$$

Thus, the positions, velocities and forces of each atom are described as a function of time. However, to obtain a trajectory of all atomic motions it is necessary to integrate this function along the simulation time. The MD software used in this study is implement with the Leapfrog algorithm [121] for integrating the equations of motion. In the leapfrog algorithm the positions $\overrightarrow{x_i}$ are defined at times $t_i, t_{i+1,}, t_{i+2}, \ldots$, spaced at constant intervals Δt and the velocities $\overrightarrow{v_i}$ are defined at times halfway in between noted $t_{i-1/2}, t_{i+1/2}, t_{i-3/2}, \ldots$,. Thus, the position $\overrightarrow{x_i}(t + \Delta t)$ is determined by the velocity at time $t + \frac{1}{2}\Delta t$ and the position $\overrightarrow{x_i}(t)$:

- Velocities at time $\boldsymbol{t + \frac{1}{2}\Delta t}$ are evaluated from the values at $\boldsymbol{t - \frac{1}{2}\Delta t}$:

$$\overrightarrow{v_i}\left(t + \frac{1}{2}\Delta t\right) = \overrightarrow{v_i}\left(t - \frac{1}{2}\Delta t\right) + \overrightarrow{a_i}(t)\Delta t \tag{3.6}$$

- We deduce the positions from the velocities:

$$\overrightarrow{x_i}(t + \Delta t) = \overrightarrow{x_i}(t) + \overrightarrow{v_i}\left(t + \frac{1}{2}\Delta t\right).\Delta t \tag{3.7}$$

- The forces are recalculated from the new structure and its energy

The process is then repeated and the system can evolve with time. Thus, from this algorithm the velocities are first evaluated at time $t + \frac{1}{2}\Delta t$ and they are used to calculated the positions at time $t + \Delta t$.
The potential energy and all the forces taken in account in equation (3.5) are described by the force fields terms.

3.3.2 Force Field:

A force field describes interactions between atoms in a molecular system. It is a set of equations and associated constants that are used to generate the potential energy of a system to model the molecular interactions providing a quantitative description of thermodynamic, structural and dynamic properties of molecules.

There are large numbers of force fields more or less specialized in different areas of application, depending on the systems studied. In biomolecular systems where proteins are simulated, among the most commonly used are found: AMBER [122] (Assisted Model Building with Energy Refinement), CHARMM [123] (Chemistry at HARvard using Molecular Mechanics) and OPLS-AA [124] (Optimized Potential For Liquid Simulations - All Atoms).

Generalized versions of these force fields providing parameters for small organic molecules are also available, especially the General AMBER force field (Gaff) [125] and the CHARMM General force field (CGenff) [126] which have been used in this work. CGenff and Gaff force fields aimed to simulate drug-like molecules and are designed to be as compatible as possible with the traditional CHARMM and AMBER force fields with similar functional form. AMBER and CHARMM force fields have similar functional forms of the energy terms providing parameters for bonded and non-bonded interactions. The form of the potential energy function for both force fields is shown with in following equation:

Bonded terms

$$\sum_{bonds} K_b(b-b_0)^2 \quad (3.8) + \sum_{angles} K_\theta(\theta-\theta_0)^2 \quad (3.9) + \sum_{dihedrals} \frac{v_n}{2}(1+\cos(n\phi-\delta)) \quad (3.10)$$

Non-bonded terms

$$\sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}}-\frac{B_{ij}}{R_{ij}^{6}}+\frac{q_i q_j}{\varepsilon R_{ij}}\right] \quad (3.11)$$

With two additional bonded terms with the CGenFF Charmm force field:

$$\sum_{\substack{improper\\dihedrals}} K_\varphi(\varphi-\varphi_0)^2 \quad (3.12) + \sum_{Urey-Bradley} K_{UB}(r_{1,3}-r_{1,3;0})^2 \quad (3.13)$$

- Bonded terms:

- Bond interactions (3.8) estimate the energy associated with oscillations between atomic pairs where atoms are separated by one covalent bond with $b - b_0$ is the displacement from the ideal bond.

- Angle interactions (3.9) correspond to oscillations of 3 atoms with $\theta - \theta_0$ the angle from equilibrium between three bonded atoms.

Both of these interactions are expressed by a Harmonic equation using force constant K_b and K_θ and equilibrium values b_0 and θ_0. The force constants can be obtained from experimental or theoretical vibrational analysis of the molecule.

- Torsion interactions (3.10) describe a dihedral rotation of 4 atoms about a central bond. This term contributes to the energy barrier during the rotation of a bond as a cosine function. K_Φ is the force constant with n and δ corresponding to the multiplicity and phase for the torsion angle Φ parameter.

In addition to these bonded terms, the Charmm force field has two extra bonded terms, the Urey-Bradley and improper dihedral terms. These two terms are used to optimize the fit to vibrational spectra and out of plane motions. The improper term is also used to prevent transition to a conformation of opposite chirality.

- Improper dihedral (3.12) refers to a bending movement out of the plan. It is used to keep aromatic cycles planar and atom chirality. It is a harmonic potential. $K\varphi$ corresponds to the force constant and φ_0 the equilibrium improper angle for the out of plan angle φ.

- Urey-Bradley interaction (3.13) is a based on a distance between atoms separated by two bonds accounting for angle bending with K_{UB} the force constant, $r_{1,3}$ is the distance between atoms 1 and 3 and $r_{1,3;0}$ the equilibrium value in the harmonic potential.

➢ Non-bonded terms:

The non-bonded terms (3.11) correspond to any interactions between atoms separated by at least three bonds or atoms from two different molecules. These interactions are described by two terms:

- van der Waals Interactions are modeled with the 6-12 Lennard-Jones potential accounting for both repulsive and attractive interactions of two atoms separated by a distance r.

The A_{ij} and B_{ij} are the van der Waals radii, R_{ij} is the distance between atoms i and j and the Lennard-jones well depth ε as illustrated in *Figure 3.4*.

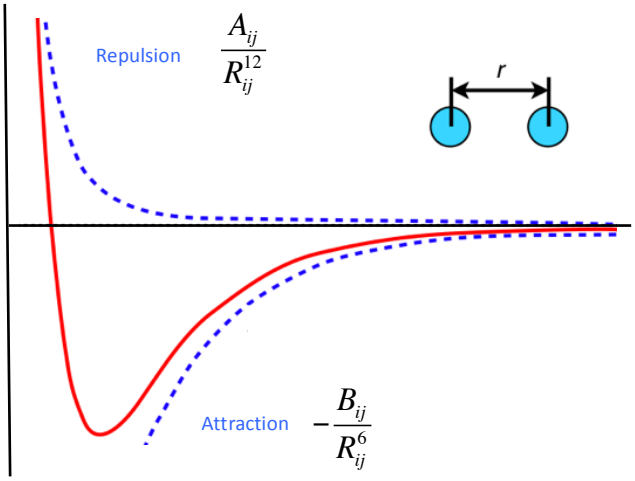


Figure 3.4: van der Waals parameters model for a molecule. Interaction energy is a function of the interatomic distance.

The $()^{6}$ term represents the attractive London dispersion between two atoms and $()^{12}$ the repulsion due to Pauli exclusion.

- The Coulomb law model electrostatic interaction. It describes the interactions of partial atomic charges q_i and q_j of atoms i and j at a set of inter-atomic distance R_{ij}.

3.3.3 Periodic boundary conditions:

To minimize surface effects and to simulate more closely the properties of a bulk system and consequently avoid problems with boundary effects caused by finite size, periodic boundary conditions (PBC) are commonly used in MD simulations.

By using PBC:

- The simulation box is replicated throughout space to form an infinite lattice
- The particles leaving the central box are returned on the opposite side
- Interactions are calculated with the particles of neighboring cells, it is called the minimum image convention. Thus, for each particles in the system, only interaction with the closest molecule is considered

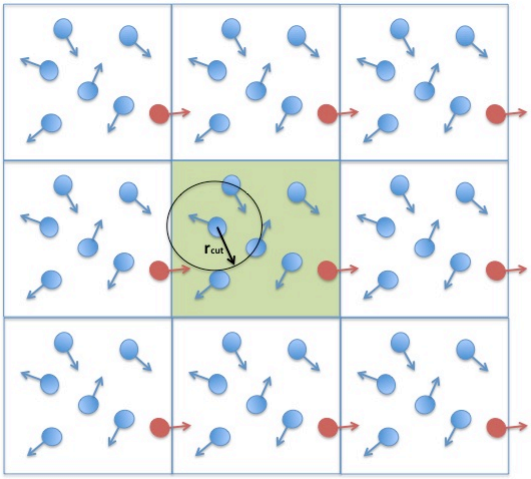


Figure 3.5: Periodic boundary condition concept for a cubic cell.

The green box represents the simulated system, which is surrounded by exact copies of this box. The arrows represent the velocities of each particle. In the *Figure 3.5*, the distance r_{cut} controls the non-bonded interactions (VdW and electrostatics). Interaction distances smaller than r_{cut} are calculated every step of the simulation; all the rest of non-bonded potentials do not contribute to the energy. For the molecule in the system, to avoid seeing its image across the periodic boundary and to stop the molecule seeing another molecule twice, the cutoff r_{cut} should not be more than half the length of the cell.
A result of this method is that whenever an atom leaves the simulation boxes (red particles), it is replaced by another with exactly the same velocity, entering from the opposite boxes face. Therefore, the number of atoms in the boxes is conserved.

3.3.4 Long-range interactions: Particle Mesh Ewald (PME)

A discontinuity of the energy due to the distance of the cutoff (r_{cut}) induces

artefacts in the motions and conformations or in the calculated energies. To correct this artefact it is possible to estimate electrostatic interactions beyond the cutoff by using the Ewald summation. The Ewald summation is a method for calculating the full long-range electrostatic energies of periodic systems [127]. By taking care of the periodicity, the electrostatic term is defined as:

$$E_{electrostatic} = \sum_{n} \sum_{ij} \frac{q_i q_j}{4 * \pi * \varepsilon_0 * R_{ij}, n} \tag{3.14}$$

Where n corresponds to the cell coordinate vector $n = n_x L_x + n_y L_y + n_z L_z$ (L_x, L_y and L_z are the cell dimensions).

The electrostatic equation converges very slowly and summing over n is not efficient. So, the idea of the Ewald sum is to separate electrostatic interactions into a short-range part, calculated in real space and a long-range part, calculated by Fourier transform in reciprocal space. The final expression of the electrostatic term is given by the following equation:

$$E_{electrostatic} = \sum_{n} \sum_{j=1}^{N} q_i \frac{erfc(\alpha |r_{ij} + n|)}{|r_{ij} + n|} + \frac{4\pi}{L^3} \sum_{k \neq 0} \sum_{j} q_j \exp\left(\frac{-|k|^2}{4\alpha^2}\right) \exp\{ik.(r_j - r_i)\} - \frac{2\alpha}{\pi^{1/2}} q_i \tag{3.15}$$

Where:

1. $k = 2\pi n / L^2$
2. $erfc(x) = 1 - \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t} d_t$: Error function
3. α Ewald convergence parameter

The α parameter is chosen in a way to optimize the convergence to 0 of the error function *erfc(x)*. In this project the particle mesh Ewald (PME) algorithm was used to recast the electrostatic energy term [128]. The PME method calculates the long-range interactions of the standard Ewald summation from particle mesh methods. The long-range interactions are calculated on a grid in reciprocal space.

3.4 Thermodynamic conditions:

During a simulation different parameters are defined to maintain the system in a particular thermodynamic ensemble. A thermodynamic ensemble corresponds to the coordinates and momenta sets of the system and the constrained thermodynamic parameters. The temperature, volume, pressure and the number of particles usually define these parameters. The three main ensembles often used in MD simulations can be defined as follow:

- Microcanonic ensemble, NVE, characterised by a fixed number of atoms, N, a fixed volume, V, and a fixed energy E.
- Canonical ensemble (NVT) characterised by a fixed number of atoms, N, a fixed volume, V, and a fixed temperature T.
- Isobaric-Isothermal Ensemble (NPT) characterised by a fixed number of atoms, N, a fixed pressure, P, and a fixed temperature T.

To simulate in the canonical ensemble, as used in this thesis, there are different approaches to control the temperature, which add and remove energy from the simulation in a realistic manner, including the Anderson thermostat [129], Berendsen thermostat [130], Nosé-Hoover thermostat [131] and Langevin thermostat [132]. The Langevin thermostat, used in this thesis, maintains the temperature by modifying Newton's equations of motion (3.5). Whereby at each time step all particles receive a random force and have their velocities lowered using a friction term.

3.5 Replica exchange molecular dynamics:

3.5.1 Introduction:

In most conventional Molecular Dynamics (cMD) simulations it is difficult to obtain accurate conformational sampling at commonly studied temperatures. At low temperatures, the system tends to get trapped in one of the numerous local minimum-energy states causing the conformational space to be underexplored. Furthermore, whereas the cMD simulation time scale is limited to nanoseconds or microseconds, many biological processes of interest occur on time scales up to seconds such as protein folding (μs-s) [133-135]. One way to overcome these limitations is to use specialized methods for capturing the long-scale, long-time conformational changes within an MD simulation. Numerous methods have been developed and the efficiency of each depends on the system and the resources

available. In this study, a version of REMD is used called temperature REMD where the temperature is exchanged. Temperature REMD is a method widely used to provide assistance in energy barrier crossing [136-138].

3.5.2 Theory:

REMD method was developed by Sugita and Okamoto [140]. REMD is one of the methods commonly used to increase the sampling of conformational space of complex biological systems. It performs a random walk in energy space and allows the simulation to cross any energy barriers and to sample more than conventional methods at low temperature.

The principle of REMD is to simulate simultaneously and independently, n replicas of the original system in parallel using conventional Molecular Dynamics method. Each replica is simulated at different temperature in the canonical ensemble (NVT: number of atoms (N), volume (V) and temperature (T) are fixed during the simulation). At regular intervals, pairs of replicas are exchanged with their nearest neighbour according to a Metropolis criterion, which gives a probability of exchange between two replicas i and j with respective energy and temperature E_i, E_j and T_i, T_j. If the exchange between two replicas is accepted by the Metropolis criterion, the two replicas swap their temperatures. In a successful exchange, a scaling factor involving the previous and the new target temperatures rescales the associated velocities of all the atoms [139-142].

Metropolis criterion:

$$P(i,j) = \min\begin{cases} 1 \\ e^{-\Delta} \end{cases} \tag{3.16}$$

With: $\Delta = \left({}^{1}/_{kT_i} - {}^{1}/_{kT_j}\right) * (E_j - E_i)$ (3.17)

k the Boltzmann constant, E and T the potential energy and the temperature of replicas i and j.

Rescaling assignment for the replica i:

$$v_{inew} = \sqrt{\frac{T_{inew}}{T_{iold}}} * v_{iold}$$

(3.18)

With T_{iold} and T_{inew} the temperatures before and after the exchange and ν_{iold} ν_{inew} the velocities before and after the exchange for the replica i.

The protein energy landscape is highly complex with numerous minima. At low temperatures, a simulation is trapped in local minima as seen with a black asterisk in *Figure 3.6*. Increasing the temperature enhances barrier crossing. Thus, when a replica is exchange to a higher temperature, the system is expected to have sufficient energy to cross energetic barriers and as a consequence to sample larger volumes of phase space whereas, at lower temperature the system explore detail energy in a local region of phase space. *Figure 3.7* shows the general workflow of REMD with 4 replicas. Each replica will perform an MD run for a pre-defined number of steps and then will stop to attempt a swapping with a nearest neighbour as indicated by the question mark. By repeating the exchange process, the replicas will be able to explore the temperature space.

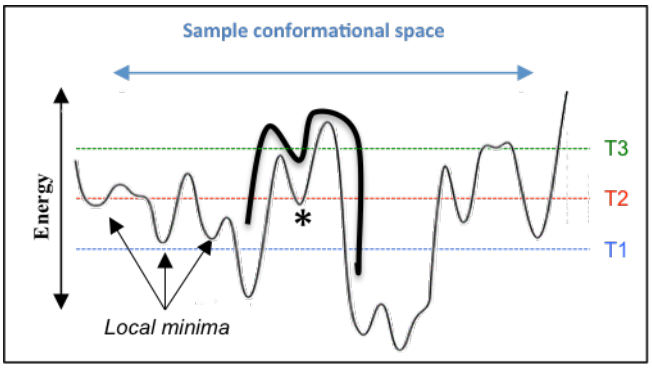


Figure 3.6: Example of a potential energy surface of a bio-molecular system. The black asterisk highlights one of the local minima of the system. At low temperature, the system can be trapped in this minimum. The dashed lines show the accessible areas at three different temperatures. Conformation swaps between lower and higher temperature, allows the system to escape from the region of the phase space where it was trapped and to sample more phase space.

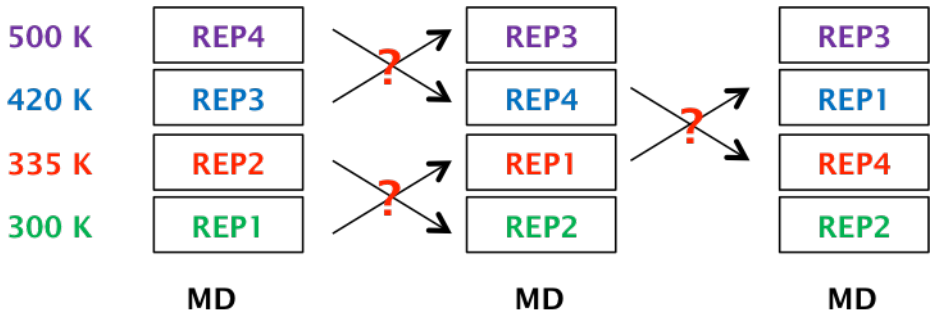


Figure 3.7: An example of REMD swaps between adjacent replicas at four different temperatures. The question mark indicates the exchange attempts. In this figure all attempted moves are successful.

Before starting a REMD simulation, the number of replicas and the temperature for each of them has to be determined. Different criteria need to be considered to evaluate the quality of a REMD simulation [143].

a) Check if the temperatures are optimally distributed. Adequate choices of temperatures are required to induce a uniform distribution of the acceptance probability between neighbouring pairs of temperatures, resulting in a free random walk in the conformational space and so each replica spends the same time at each temperature during the simulation.
b) Evaluate if the number of replicas/temperatures is sufficient. If temperatures are well distributed, their numbers will determine the energetic overlap between neighbouring replicas and in this way allow a reasonable exchange rate. An exchange success rate corresponding to at least 20% is satisfactory [140-141,143-144].
c) The highest temperature must be sufficiently high to prevent the system being trapped in a local energy minimum. The trajectory at the highest simulation temperature needs to induce important conformational change(s). Analysis of structural features of this trajectory will provide evidence for this last point.

A number of studies have proven the efficiency of REMD methods to enhance the sampling of a simulation compared to conventional MD tools [145-146]. However, REMD requires the number of replicas to increase with the number of degrees of freedom of the studied system [39]. Consequently, running REMD requires a large amount of computational resources for most complex biological systems such as proteins due to the large system size.

3.5.3 Temperature distribution and acceptance ratio:

As mentioned above, the choice of the temperature range is crucial to ensure that the overlap of the potential energy distributions is sufficient for the exchange and to cross barrier. The highest temperature needs to be high enough to cross energy barrier(s) while the lowest temperature explores local minima. However, the temperature range cannot be too wide.

If the potential energy distributions are assumed to be Gaussian, the distribution is centred on the average energy of the replica with the standard deviation depending on the temperature *(Figure 3.8)*.

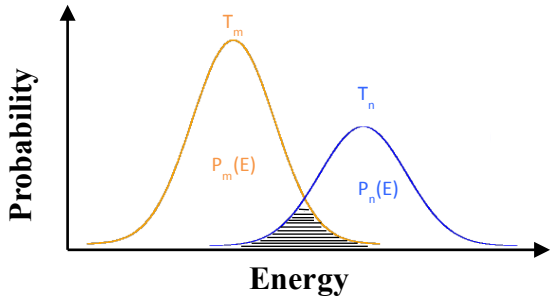


Figure 3.8: Illustration of an overlap of two potential energy distribution functions $P_m(E)$ and $P_n(E)$ of two replicas at temperatures T_m and T_n.

If the Gaussian overlap is too small, an insufficient number of exchanges will be accepted making the use of REMD inefficient. Conversely, if the overlap is too high so the temperatures are too closely spaced, the exchanges will be accepted too frequently preventing an efficient sampling. Therefore, the choice of the spacing between temperatures appears to be essential. An optimal set of temperatures, as mentioned previously, will be given by the distribution acceptance ratio over the entire temperature range. Several studies have demonstrated different approaches to produce an optimal temperature distribution.

Okamoto [147] has shown that an exponential or geometric distribution of the temperatures is preferable to an arithmetic distribution where the temperatures would be spread with a constant spacing between consecutives temperatures. Indeed, higher temperatures will induce higher energies and broader energy distributions leading to an increase in the probability of exchange. Therefore, the temperatures need to be closer together when they are lower in temperature. However, a simple geometric distribution does not yield a

constant value for the exchange probabilities [143].

In the following studies, Sanbonmatsu and Gracia [148] and Rathore *et al* [149] assume the energy distributions to be Gaussian, and produced a similar scheme with a polynomial fit of average energies as a function of temperature. The target temperatures are then determined by solving the equation P = exp ($\Delta\beta\Delta E$), with $\Delta\beta\Delta E$ corresponding to equation (3.17), the energy differences between two replicas. This approach has shown better performance than the geometric temperature distribution. However, initial energy approximations or short runs at few temperatures are required to generate the energy distributions.

Alexandra Patrikson and David van der Spoel have developed an algorithm to generate a range of temperatures given the number of atoms and water molecules in the system, the lower and higher temperatures limit, and the desired probability of exchange [150]. No detail knowledge about the energies or temperatures is required. The principle of this method is that, from a set of different structures used to parameterize the algorithm, to compute a heat capacity per structure atoms and per water molecule. From that the energy of the studied system is predicted and according to the desired exchange probability, the temperature distribution is generated.

3.6 Molecular docking:

Protein ligand docking is a technique used to predict the likely binding modes of a ligand when bound to its receptor in order to obtain the most stable ligand-receptor complex. In this study, molecular docking aimed to complement the observations from MD and NMR methodologies by providing knowledge of potential binding modes. Knowing the binding site location is essential for reliable docking along with using a reliable or appropriate conformation of the active site. For this reason self-docking of ligands back into a protein is more successful than the cross-docking of different compounds [151-153] often due to changes in the binding site. In this study, an x-ray crystal structure of an inhibitor bound to the molecule of interest is available, indicating the presence of an important cofactor in the binding mode.

The steps of a docking simulation are first to sample conformations in the active site of the target molecule and then ranking these conformations via a scoring function. Various docking program are available such as Glide [154], MOE [155], Gold [156] and AutoDock [157] among others. In this study, the AutoDock program was used.

AutoDock tool was developed to predict the possible interactions between small molecules and proteins. The aims is to identify a binding mode between a small molecule and its target by sampling the conformational space of the system by varying all the degree of freedom of the ligand while keeping the protein rigid. For each conformation, the resulting binding energy is evaluated.

Prior to the docking, the interaction energy between the ligand atoms and the protein are calculated for the entire binding pocket to speed up the evaluation of the system energy. The protein is embedded in a three-dimensional grid and a probe atom is placed at each grid point then the energy is interpolated to the nearest grid point. The energy of interaction at each grid point is then calculated. An affinity grid for each atom type in the ligand is calculated as well as electrostatics and desolvation potentials grids. Thus, when sampling different ligand conformations on the grid, the resulting binding energy can be evaluated.

AutoDock uses a Lamarckian genetic algorithm to sample the conformation of the ligand. Several simulations are performed to identify the best conformations. The ranking of the various conformations is performed via a scoring function. Autodock uses a semi-empirical free energy force field. It was parameterized using a large number of protein-inhibitors complexes with known inhibition constant. The functional form of the force field is as follow:

$$\Delta G = \left(V_{bound}^{L-L} - V_{unbound}^{L-L}\right) + \left(V_{bound}^{P-P} - V_{unbound}^{P-P}\right) + \left(V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf}\right)$$

$$V = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right) + W_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right) + W_{vdw} \sum_{i,j} \left(\frac{q_i q_j}{e(r_{ij}) r_{ij}}\right) + W_{sol} \sum_{i,j} \left(S_i V_j + S_j V_i\right) e^{-r_{ij}^2/2\sigma^2} \quad (3.19)$$

with L and P referring to the ligand and the protein respectively and W are weighting constants. The first term correspond to the Lennard-Jones interactions based on the Amber force field *(3.2.2)*. The second term is a directional hydrogen bond term based on a 10/12 potential with C and D parameters for the well depth. The third term is the Coulomb potential for electrostatic interactions and the final term is a desolvation potential with V the volume of atoms that surround a given atom, S a solvation parameter and σ a distance-weighting factor equal to 3.5 Å.

As part of the analysis, clustering is performed across the conformational

ensemble to identify and remove similar structures. The best predicted energy conformation for each cluster can be extracted and analyzed.

3.7 Analysis tools:

In this study, a number of analysis tools have been used to compare the different method ensembles. In the following section the tools including cluster analysis, root mean square deviation, principal component analysis, hydrogen bond analysis, the Cambridge structural database programs, conformational generator and structure shape comparison tool are reviewed.

3.7.1 Cluster analysis method:

Molecular Dynamics simulations and enhanced sampling methods are powerful methods for sampling the conformational space of a biomolecular system. These simulations produce a sequence of snapshots of the system (each defining a "configuration" of the studied system), specifying their 3D atomic positions, as a function of time in the case of conventional MD, providing a representation of the sampling. Data-mining techniques, such as clustering, are suitable methods for analysing this large amount of data.

Clustering is widely used to organize data to understand the important molecular motions of molecules. The term "cluster analysis" refers to numerous different methods and algorithms for grouping similar objects into meaningful structures [158-160]. It divides data into groups such that objects in one group are more similar to each other than to the objects in any other group. Thereby, clustering appears as a technique to understand, simplify and interpret large amounts of multidimensional data.

In this work, DASH, an algorithm developed to analyse MD as well as REMD trajectories based on torsion angles of rotatable bonds was used [161].

DASH analyses MD and REMD simulations to extract the major features for each torsion angle and then identify the most frequent conformations adopted by the system. The torsion angles of interest are extracted and are given as input file for the DASH program. The algorithm clusters the torsion space as a time series of DASH states in the case of conventional MD. A DASH state describes geometrically similar conformations and is characterized by mean and standard deviation torsion angles. In the case of REMD, trajectories are not continuous with respect to simulation time due to the exchanges. Therefore, a special option is used within the DASH program to remove all the elements that rely on the frames

coming from a time-series. The DASH states are defined only by their relative occupancies, not by their persistence for continuous periods of time.

Each MD and REMD snapshots have been assigned to a state and a representative frame for each state will be available corresponding to a conformation with the highest similarity to the mean torsions. As result, DASH enables us to understand the significant molecular motions and provides insight into the relative flexibility of the system.

3.7.2 Root mean squared deviation:

Root mean squared deviation (RMSD) is a measure to evaluate similarity between two molecules. It measures the average distance between selected atoms and a reference structure and is defined as follow:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{3.20}$$

With δ the distance between N atoms.

3.7.3 Principal component analysis:

Principal component analysis (PCA) is a dimensionality reduction method to retrieve the most dominant patterns of your input variables based on variance. It generates a linear combination of the input variables, which explain the majority of the variance [162-164]. In the case of MD/REMD simulations it extracts the bigger motions of the system [165-167]. The principle is to map the studied system from a multidimensional space to a reduced space spanned by few principal components (PCs) to investigate the most relevant collective fluctuations. Prior to the PCA, the centre of mass rotation and translation is removed for each frame by a least square fitting of each conformation of the ensemble to a reference or average structure. The correlation between the atomic motions will then be determined through a covariance matrix C_{ij} of the atoms coordinates i and j:

$$C_{ij} = \langle M_{ii}^{\frac{1}{2}}(x_i - \langle x_i \rangle) M_{jj}^{\frac{1}{2}}(x_j - \langle x_j \rangle) \rangle \tag{3.21}$$

Where <> denotes an average over a complete ensemble. M is a diagonal matrix

containing the masses of the atoms (mass-weighted analysis) or the unit matrix (non-masse weighted analysis). x_i and x_j are the atoms positions. C is a symmetric 3N * 3N matrix with N referring to the atoms number and can be diagonalized with an orthonormal transformation matrix R to obtain PCs and corresponding eigenvalues.

$$R^T C R = diag(\lambda_1, \lambda_2, \dots, \lambda_{3N}) \tag{3.22}$$

Where λ_i are the eigenvalues i of the PCs i.

The columns of R are the eigenvectors also called PCs or essentials modes. R defines a transformation to a new coordinate system. The eigenvectors describe direction of atomic displacement in high dimensional space and corresponding eigenvalues show mean square fluctuation of the total displacement. These eigenvalues are sorted in descending order and therefore, first few eigenvectors describe the largest variance in the atomic coordinates corresponding to the largest conformational change obtained from the MD simulation.

The first few PCs contain the largest mean square fluctuations. Typically the first few PCS often contribute significantly to the bigger motions of the system. Therefore, from PCA the largest variance motions corresponding to the larger conformational transitions of an ensemble are extracted describing the functionally relevant conformational transitions. To visualize the sampled conformations in the subspace spanned by the eigenvectors, the trajectory can be projected in a so-called “2D projection”. The trajectory is projected onto the two first eigenvectors. Each point in the projection plot will represent a frame from the simulation and the distribution will show how each frame occupies a certain region of the conformational space. The modules g_covar and g_anaeig of GROMACS software were used to perform the PCA [168].

3.7.4 Hydrogen bond:

Hydrogen bonds are interaction that can occur when a hydrogen atom from a molecule or a molecular fragment D-H (donor), where D is more electronegative than H causing a significant positive charge on the H, is close to another electronegative atom with a lone pair of electrons (acceptor). Hydrogen bonds often stabilize structures of molecules such as proteins, DNA and membranes as

well as playing an important role in the ligand-receptor recognition and small molecule conformations. Several studies shown that the distance between the donor and the acceptor is often within a range greater than 1.6 Å but less than 3.5 Å (distance between the heavy atoms A --- D) and an angle θ = A --- H-D between 90° and 180° [169-170] *(Figure 3.9)*. Hydrogen bonds with donor-acceptor distance of 2.2 to 2.5 Å are considered as "strong, mostly covalent", 2.5 to 3.2 Å as "moderate, mostly electrostatic" and 3.2 to 4.0 Å as "weak, electrostatic" [171].

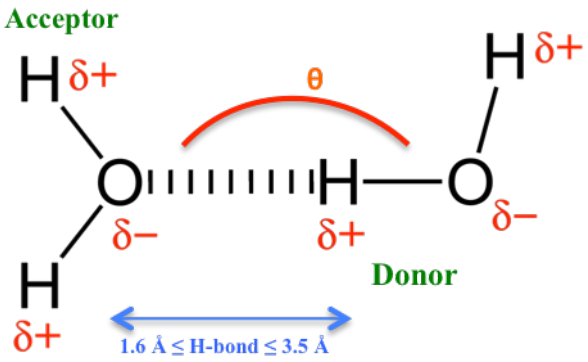


Figure 3.9: Hydrogen bond in water.

3.7.5 The Cambridge Structural database:

The Cambridge Structural Database (CSD) was established in 1965 and refers to a world repository of small molecule organic and metal organic crystal structures [172]. It contains over 900 000 entries of crystal structures from X-ray and neutron diffraction analyses. The CSD reflects the published literature, thus each entry in the CSD relates to an individual publication of a specific crystal structure and is updated on a continuous basis. Detailed information of each entry is available such as chemical name, formula, diagram, bioactivity, conformer, isomer and racemate that were provided from the published literature. Each crystal structure undergoes extensive validation and cross-checking by expert chemists and crystallographers. The CSD also provide a system (CSD system) with tools to analyse CSD entries such as Mogul, ConQuest, Mercury and many others, to rapidly access to information on the preferred values of bond lengths, valence, angles.... and geometry of isolated ring systems as well as 3D searching of structures and visualization tools. For the purpose of this study, the ConQuest program was used.

ConQuest is a program providing advanced 3D searching of the structures in the CSD [173]. Conquest permits the search of all CSD information and the

location of chemical substructures and intermolecular interactions. The idea of ConQuest is to build queries of fragments of interest and to search in the CSD structures if the query has been identified. Queries are entered graphically and can be embellished by 3D geometrical constraints to locate specific conformations. Full ranges of extensive search options are available to define something to be searched such as torsion angles, bonds and angles as well as the ability to define chemical constraints. The information about the entries found during the search, name, formula, elemental composition, literature reference and experimental details is also available.

3.7.6 Conformation generator and similarity analysis tools:

Drug discovery often involves studying potential drug candidates that share similar 3D shapes and electrostatics resulting in similar biological effect when bound to the target molecule. Therefore, the need to find efficient and fast tools to analyze the conformational ensemble obtained from experiment and computational methods is important. For the purpose of this work, two applications were used name OMEGA a conformer generation package and ROCS a similarity analysis tool, which are industry leading tools for this purpose.

- OMEGA:

OMEGA is a conformational generation tool providing conformational sampling of small molecules [174]. The process used by OMEGA can be summarized by five steps [174].

The first step named ***fragment database preparation*** involves the fragmentation of a large collection of commercially available compounds into ring conformations and small open-chain linkers. At this step, for each fragments at least one conformation is generated. Enumerating every possible combination of ring conformations and inconvertible nitrogen atoms will allow the generation of additional conformations; the lowest energy conformers are retained. The next step in the model generation, ***torsion sampling dictionary***, is a torsion search determining bonds that may freely rotate. Torsions are populated with different torsion angle values from experimental or calculated 3D structures (Cambridge Structural Database and PDB database).

The third step, ***3D structure generation***, the fragments from step one that match the query molecule are assembled using geometric and chemical rules.

The fourth step, ***torsion driving***, the rotatable bonds in the conformer generated

in step 3 will be compared with the torsion angles values library from step 2 and the appropriate values with no internal clashes or duplicates are noted.
The final step, ***sampling***, all conformers generated are ordered using a scoring function, based on force field calculation, eliminating conformations with internal clashes. Using the lowest scoring conformation, all higher scoring conformations that are less than a threshold RMSD value are eliminated (default RMSD = 0.5 Å). This process is continued until a fixed number of conformations is reached.

- Rapid Overlays of Chemicals Structures:

ROCS is a shape-based superposition method using only heavy atoms of a ligand (hydrogen are ignored) and can compute a global rate of 100-1000 similarities per second [175]. ROCS aligns each molecule of a database of pre-computed molecular conformer (MD, REMD, NMR, Docking or OMEGA ensemble) against a reference molecule(s). It uses a volume overlap procedure between a pair of molecules A and B by measuring the degree of volume overlap between the two. In addition to shape alignment, a chemistry overlay, known as "color", is available facilitating the identification of compounds that are similar in both shape and chemistry. The measure of the similarity between the reference and the molecule is based on Tanimoto coefficient from fingerprint-based similarity calculation [176-177], defined as follow:

$$Tanimoto = Sim_{Tanimoto}(A,B) = \frac{bothAB}{onlyA + onlyB + bothAB} \tag{3.23}$$

With *only*A the number of bits set in fingerprint A but in B and vice versa for *only*B. *both*AB number of bits set in both A and B. Where a bit refers to the presence of the absence of a molecular feature.

The Tanimoto coefficient is commonly used to compare fingerprints together but also can be derived in 3D space. Considering the shape and color alignments scores as $\text{Tanimoto}_{\text{Shape}}$ and $\text{Tanimoto}_{\text{Color}}$ [178], the score will be define as overlap between the molecules shape or overlap between the molecules functional groups as follow:

$$\text{Tanimoto}_{\text{Shape}} = \frac{\text{overlap}(q,t)}{\text{overlap}(q,q) + \text{overlap}(t,t) - \text{overlap}(q,t)} \tag{3.24}$$

With overlap(q,t) the volume of the intersection q and t and [overlap(q,q)+ overlap(t,t) - overlap(q,t)] the volume of the union of q and t.

The scores for their best alignment will rank the conformations in the database. A simple shape or chemistry alignment run will be scored by Tanimoto coefficient while combinations of shape and chemistry alignment will use TanimotoCombo coefficient. TanimotoCombo coefficient is simply to separate measures added together as TanimotoCombo = $Tanimoto_{Shape}$ + $Tanimoto_{Color}$. The Tanimoto ranges from 0 to 1 while Tanimoto ranges from 0 to 2 with 0 no overlay, 1 full shape overlap or 2 full shape and chemistry overlap.

3.8 Summary:

In this chapter the methods used in this thesis have been outlined. REMD and NMR ensembles will be compared to understand the conformational space of GSK2194069 and compounds 2, 3 and 4 in solution. A docking method was used to complete the study to understand the potential binding mode of these compounds.

The CSD database will be used as an indicator of the quality of the REMD simulations in term of force field parameterisation efficiency. Principal component analysis will be applied to the REMD and NMR ensembles to establish how well the REMD is able to reproduce experimental data along the principal modes. Omega will be used to evaluate how a conformer generator is able to reproduce the correct conformational distribution. Analysis tools such as RMSD, Dash, ROCS and hydrogen bond analysis will enable us to visualise how different are the conformational ensembles of the different compounds in comparison to the reported GSK2194069 x-ray structure [62] in term of shape similarities, to determine potential binding modes for these compounds.

Chapter 4 Replica exchange molecular dynamics

4.1 Aim of this Chapter:

A co-crystal structure of GSK2194069 with NADPH, confirms that the molecule binds to the KR domain of hFAS [62]. Furthermore, NMR data of GSK2194069 and other hFAS inhibitors has enabled us to understand the conformational equilibrium of these compounds in solution. The aim of these following chapters is to identify if advanced molecular simulations can be used to generate reliable solution phase conformational distributions and to identify if these compounds are preorganized in accordance with the likely binding mode.

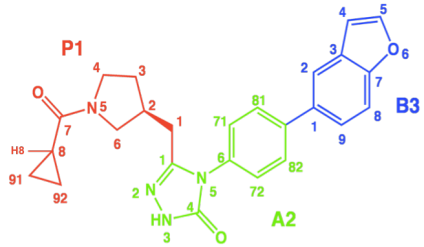
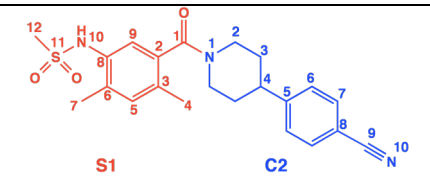
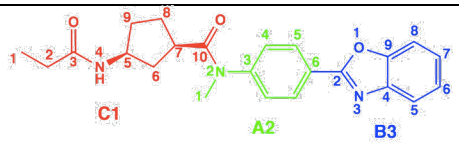
To overcome the local energy trapping seen in conventional molecular dynamics simulation and to enable the exploration of the entire possible conformation in phase space, REMD simulations have been performed on the four compounds. From the REMD trajectories many conformations of the ligands in solution are identified but only a proportion of these will be consistent with the binding mode. If the population of the bioactive conformations can be increased, the ligand can be preorganized to improve its affinity to the required target. The biochemical and cellular potencies (in pIC_{50}) of those 4 compounds as well are their chemical structure can be found in the *Table 4.1*. Published hFAS inhibitor GSK2194069 and compounds 2 and 3 appear to be fairly potent while the novel inhibitor compound 4 is significantly less active despite being similar to GSK2194069.

In this chapter the performance of the REMD simulations were examined by checking three different aspects introduced in *Section 3.5.2*. The accuracy and efficiency of the REMD trajectories was then studied by comparing the distribution of the compounds torsion angles with similar fragments from the Cambridge Structural Database via the ConQuest software package [172-173]. A significant deviation of the torsion angle distribution from those observed in the CSD is most likely to indicate an error in the force field parameterisation.

4.2 Methodology:

4.2.1 Starting conformation preparation:

To investigate the impact of initial conditions on the sampling, Replica Exchange molecular dynamics simulation (REMD) were performed on the four potential inhibitors starting from two different conformations named Confo1 and Confo2 using the Amber 12 and Gromacs 5.1.4 packages with the Amber Gaff and CGenff Charmm force fields for small molecules respectively. The molecular modelling package MOE was used to build each compound in 3D [155]. To differentiate conformers, key values of torsion angles were defined within each molecule. The choice of the torsion angle value of important chemical moieties such as E/Z conformers of amide moieties, sulphonamides and boat/chair conformations of the cyclohexane ring can influence the conformational sampling and were thus set up according to the different conformations they can adopt. More details of the initial torsion angle values as well as a detailed description of the different chemical moieties can be found in the following parts of this chapter. *Table 4.1* shows the NMR torsion angle definition for each compound. The definition of torsion angle τ1 in GSK2194069 and compound 4 is based on the hydrogen atom bound to the carbon atom number C8 and C7 respectively. The biochemical potency in pIC50 for each compound is also reported in *Table 4.1*.

Compounds	Torsion angles	Biochemical potencies pIC_{50}
1 	**τ1:** P1 N5 - P1 C7 - P1 C8 - P1 H8 **τ2:** P1 C4 - P1 N5 - P1 C7 - P1 C8 **τ3:** A2 C1 - P1 C1 - P1 C2 - P1 C6 **τ4:** P1 C2 - P1 C1 - A2 C1 - A2 N2 **τ5:** A2 C1 - A2 N5 - A2 C6 - A2 C72 **τ6:** A2 C81 - A2 C9 - B3 C1 - B3 C9	7.5
2 	**τ1:** S1 C8 - S1 N10 - S1 11 - S1 C12 **τ2:** S1 C6 - S1 C8 - S1 N10 - S1 S11 **τ3:** C2 N1 - S1 C1 - S1 C2 - S1 C3 **τ4:** S1 C2 - S1 C1 - C2 N1 - C2 C21 **τ5:** C2 C32 - C2 C4 - C2 C5 - C2 C6	6.7
3 	**τ1:** C1 C1 - C2 - C3 - N4 **τ2:** C1 C3 - N4 - C5 - C6 **τ3:** C1 C6 - C7 - C10 - A2 N2 **τ4:** C1 C7 - C10 - A2 N2 - C3 **τ5:** C1 C10 - A2 N2 - C3 - C41 **τ6:** A2 C51 - C6 - B3 C2 - O1	7.5

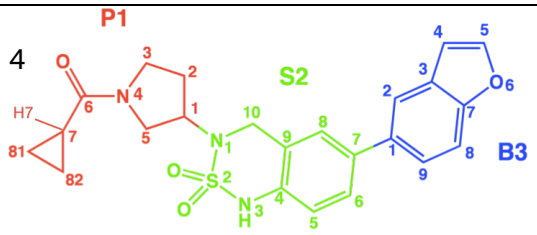
	τ1: P1 H7 - P1 C7 - P1 C6 - P1 N4 τ2: P1 C7 - P1 C6 - P1 N4 - P1 C3 τ3: P1 C2 - P1 C1 - S2 N1 - S2 S2 τ4: S2 C6 - S2 C7 - B3 C1 - B3 C9	5.1

Table 4.1: Torsion angle definition of the four potential inhibitors of the hFAS.

GSK2194069 and the three compounds contain symmetrical functional groups, which need consideration in terms of the definition of torsion angles values and for further analysis such as the superimposition of molecules. From NMR data (*Table 4.1*) a specific definition for each torsion angle based on atom name is given. However, in the NMR data files, explicit atom numbering is not used, we need this to compare with REMD simulation. In this manner, it appears difficult to define torsion angles in an exact same way in each method. *Figure 4.1* illustrates and example of the symmetry observed in GSK2194069 torsion angle τ6.

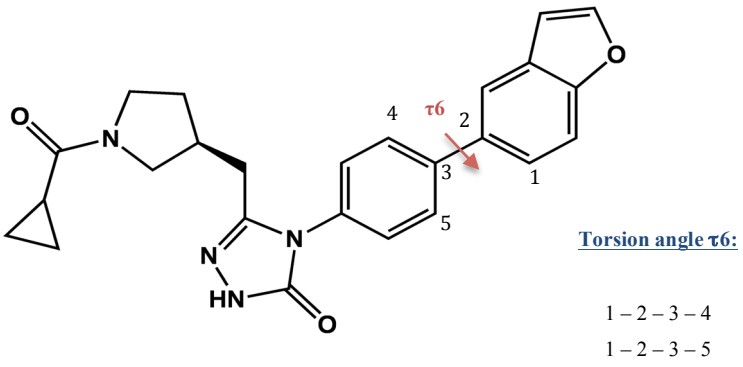


Figure 4.1: Illustration of torsion angle τ6 symmetry in GSK2194069.

Without knowing which carbon is involved in the torsion angle, two possible choices can be made corresponding to atoms 1 – 2 – 3 – 4 or 1 – 2 – 3 – 5.
From the NMR data analysis, the data from symmetrical atom were combined. To be consistent with the experimental data, the same methodology was followed for the REMD simulations.

4.2.2 REMD parameterization:

The temperature range was generated using an online generator *(Section 3.4.3)* with an overall desired exchange probability several replicas of 30 % and provided with an overall desired exchange probability among replica of 30 % and provided us with temperature for 48 replicas from 278 K to 481.02 K. Each starting conformation was solvated using explicit TIP3P water molecules in a cubic box with periodic boundary conditions (PBC). Particle Mesh Ewald was used for the long-range interactions using a 10 Å cutoff. Bonds involving hydrogen were constrained using the SHAKE algorithm with a tolerance of 0.00001 Å. REMD simulations were performed in the NVT ensemble using a Langevin thermostat for the temperature coupling with a collision frequency of 1 ps^{-1}. 200 ps of NVT simulation were used to equilibrate the initial state to the desired temperature for each replica, following a rescaling of the velocities. Using these equilibrated replicas, 500 ns of REMD simulation were performed on each replica. All exchanges between neighbouring replicas were allowed every 2 ps in the NVT ensemble with a 0.002 ps timestep. The compounds were parameterised using the General Amber force field (GAFF) and the Charmm General Force Field (CGenFF) for small molecules.

4.2.3 Post-processing:

REMD is used to obtain thermodynamic ensembles at a temperature of interest. However, the implementation of REMD in the Amber software required post-processing the trajectory data. With Amber, REMD will produce a trajectory for each replica representing the frames sampled in the continuous MD run, which changed in thermostat temperature as the replica gets exchanged. In this way, each replica trajectory appears to be uncontinuous with respect to temperature. The Amber software provides post-processing scripts that reconstruct the data for individual temperature. Therefore, it is possible to follow the trajectory through the temperature space. Compared to Amber, Gromacs REMD trajectories are written at a consistent temperature so no post processing is required.

Experimentally, NMR were performed at 278 K, so all REMD trajectories where studied at this temperature. Once the trajectories are rebuilt with respect to the temperature, the efficiency of the REMD can be examined.

4.3 REMD efficiency:

4.3.1 Acceptance probability target:

If the temperatures are appropriately distributed, the acceptance ratios will be the same through temperature range resulting in a free random walk. All simulations from each compound led to a uniform probability of exchange between 20 to 35 % in accordance with the desired probability specified in the temperature generator of 30 %, and are not too small (>10%) as required by the studies of Sugita and Okamoto [140]. More details about the acceptance ratio can be found in the Methodology part of the thesis *(Section 3.4.3)*.

4.3.2 Temperature distribution:

The efficiency of a REMD simulation depends on the capability of the replica to exchange between lower and higher temperatures. As mentioned in Chapter 3, an optimal distribution of temperature induces a free random walk in the temperature space. 48 replicas were used to perform REMD; plotting the temperature trajectories of each of them will be difficult to visualise. Therefore, the temperatures visited by four replicas, choose randomly, are shown in *Figure 4.2A* from conformation 1 *(Section 4.1.2.1)* of GSK2194069 simulated with the Amber Gaff force field and to better distinguish them 500 ns of the trajectory is shown. In *Figure 4.2*, each colour corresponds to a replica. As no major differences were observed within the compounds, only GSK2194069 is presented here. A free random walk is observed in the temperature space by visiting both lowest and highest temperatures during the simulation, which is indeed realised for all four replicas.

However, cases where the replicas at higher temperatures demonstrated difficulty to cross the energetic barrier are observed. *Figure 4.2B* displays the temperature evolution of replicas starting at 302.49 K and 481.02 K in black and cyan. From 50ns to 150ns and from 200 ns to the end of the simulation, the conformations at higher temperatures are trapped in higher energy minima and no exchanges are accepted with these temperatures.

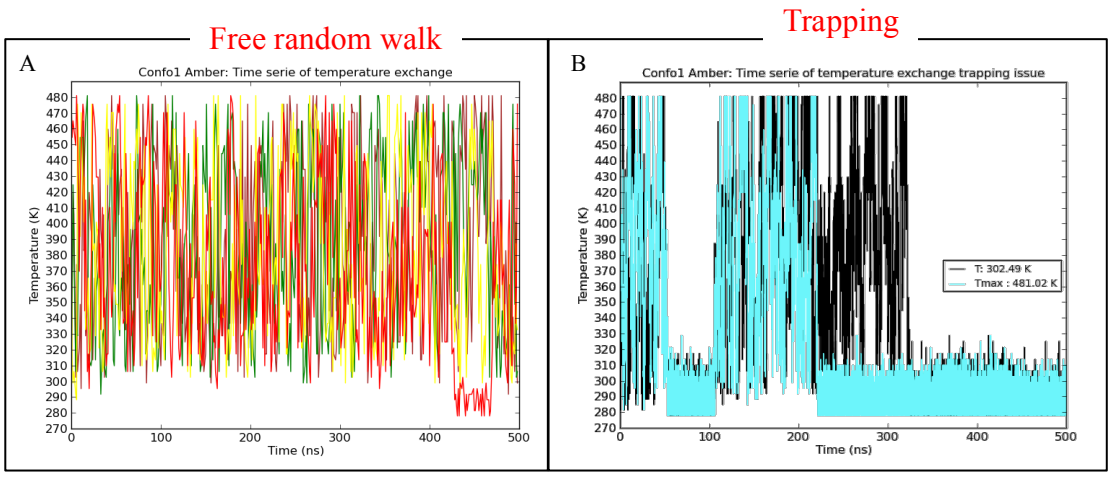


Figure 4.2: Time series of temperatures exchange for GSK2194069 in a case of a free random walk (A) and trapping in higher energy barrier (B).

Each colour represents a replica at a given temperature (brown: 278 K; green: 309.81 K, yellow: 405.51 K, red: 475.61 K, black: 302.49 K, cyan: 481.02 K).

Despite some local trapping in the temperature space, exchanges are observed throughout the trajectories. Similar analyses are for the three other compounds, showing consistency with these observations. Free random walks with a constant acceptance ratio through the trajectories and with both force fields have been observed.

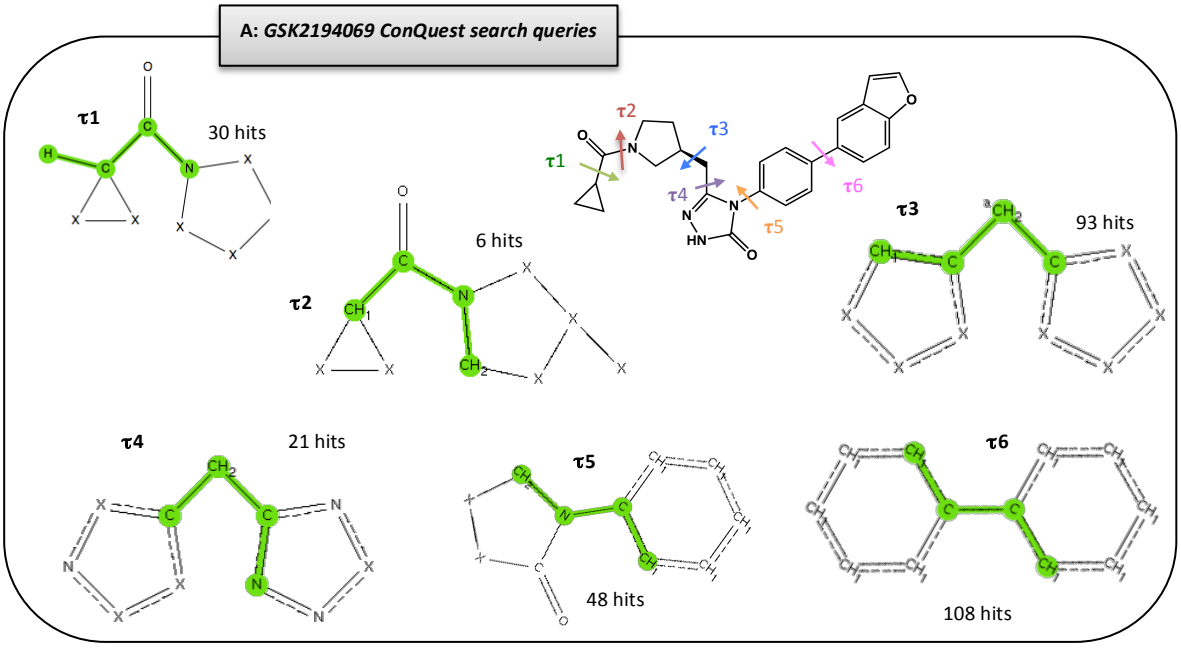
Further analyses looking at the torsion angle distributions in comparison with the Cambridge Structural Database (CSD) of experimental crystal structures at the lowest and the highest temperatures will enable us to examine if the trajectories were trapped in a local minimum-energy state as well as to evaluate if the conformers are sensible. What we consider as trapping could be due to the force field parameterization inducing inaccurate conformational sampling.

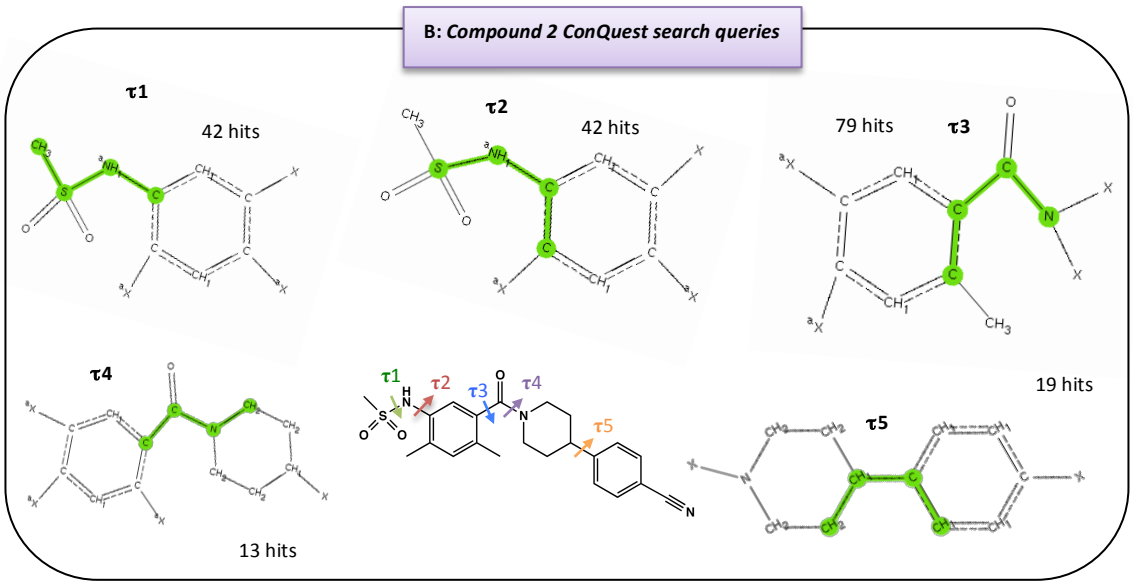
4.4 Comparison with the Cambridge Structural Database:

In this section the distributions of the compounds torsion angle will be compared with experimentally determined values from the CSD structures from a commonly used assumption that the conformations adopted by the CSD crystal structures are likely to be representative of the conformations adopted in solution.

4.4.1 ConQuest search queries:

As described in *Section 3.6*, parts of the compounds are entered as fragments in the conquest search program. The torsion angle values from the REMD simulations are then compared with molecules from the CSD with those substructures. The sub-structural queries searched within the ConQuest are shown for each compound in the following figures together with the number of hits found. The hits numbers, next to each query, refer to the number of crystal structures found with similar substructure pattern.





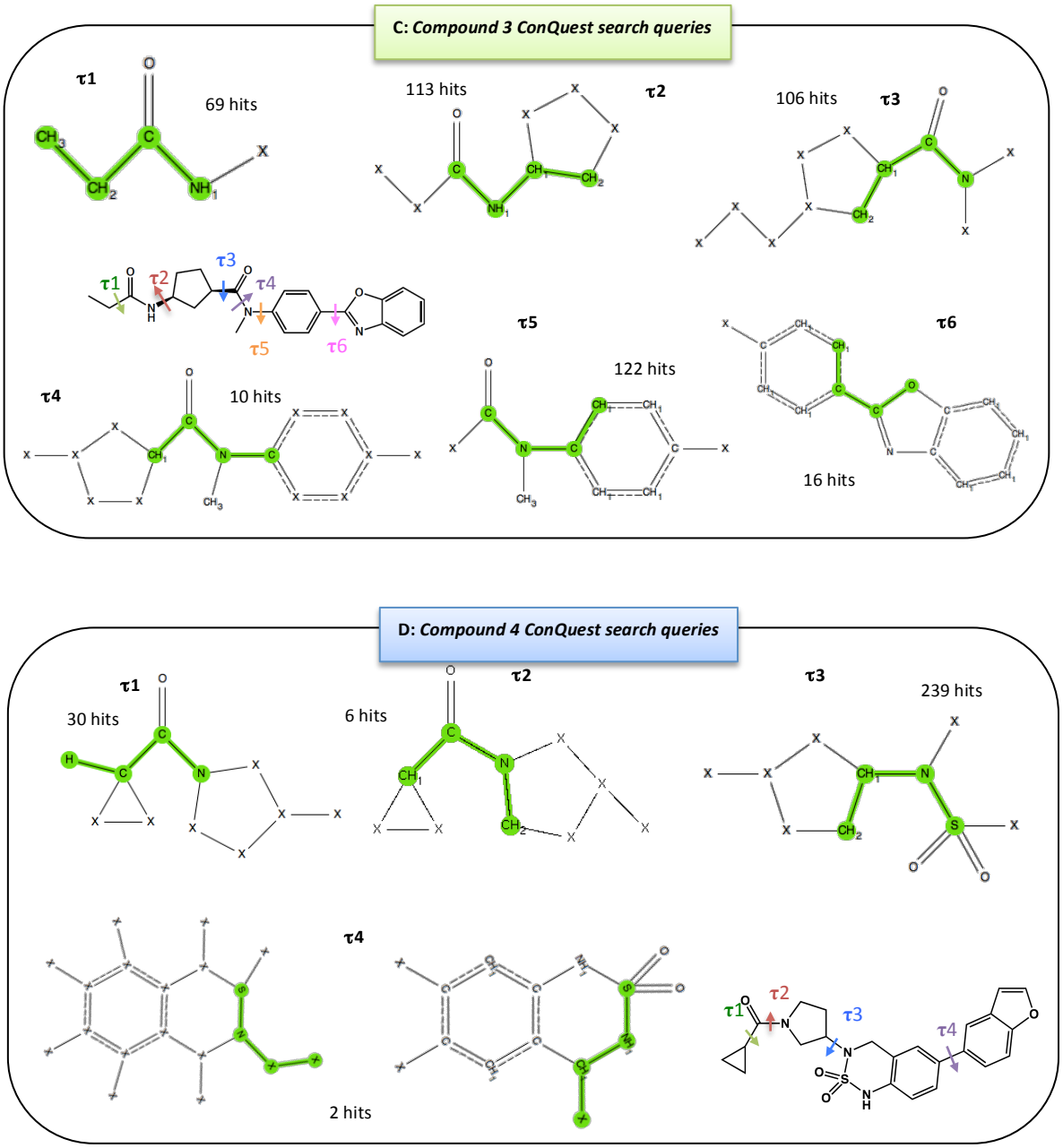


Figure 4.3: Chemical substructures queries and their corresponding torsion angle (green) for the four compounds.

Figure 4.3 describes the queries for each compound, defining a chemical pattern to be searched in the CSD. Once the desired query has been defined a search in the database is performed and the structures containing the similar fragment are listed as a hit. The 3D structures of the hit ensemble can be viewed as well as all the bibliographic and chemical information about the entry. When building a query, if the search is too specific, few (or no) compounds are often found and if it is too general the structures can be irrelevant to the molecule of interest. The

symbols ^{a}X and X in *Figure 4.3* specify the acyclic character and the choice of being any atom respectively, emphasizing the choice to be more or less specific throughout the search. Acyclic means an atom that is not part of a ring moiety. Finding a balance between the two is important and is made by checking the 3D structures of the hit molecules. Additionally, molecular geometry measurements are also available from the hits throughout the search such as distances, angles and torsion angles. In this study, torsion angles corresponding to the NMR torsion definition have been evaluated for each CSD hit found during the search, as seen in green in *Figure 4.3*. The torsion angle values are extracted from the associated CSD file and compared to the REMD trajectories in the following section.

4.4.2 Comparison with REMD simulations:

Torsion angles of each compound were extracted from all REMD simulations to be compared with the corresponding CSD data. The following figures display histograms of the observed torsion angles in the REMD trajectories and in the CSD search. The torsion angle values of the GSK2194069 x-ray crystal structure are shown in magenta lines and the torsion angles values of each starting conformation in red lines.

4.4.2.1 GSK2194069:

4.4.2.1.1 Torsion angle τ1:

The distribution of torsion angle τ1 in the CSD displays main sampling between -50° to 50° with few hits at +/- 90° and +/- 150° *(Figure 4.4A)*. However, REMD trajectories, *Figure 4B*, suggested a different torsional profile predominantly around +/-90° and minimal sampling around 0°.

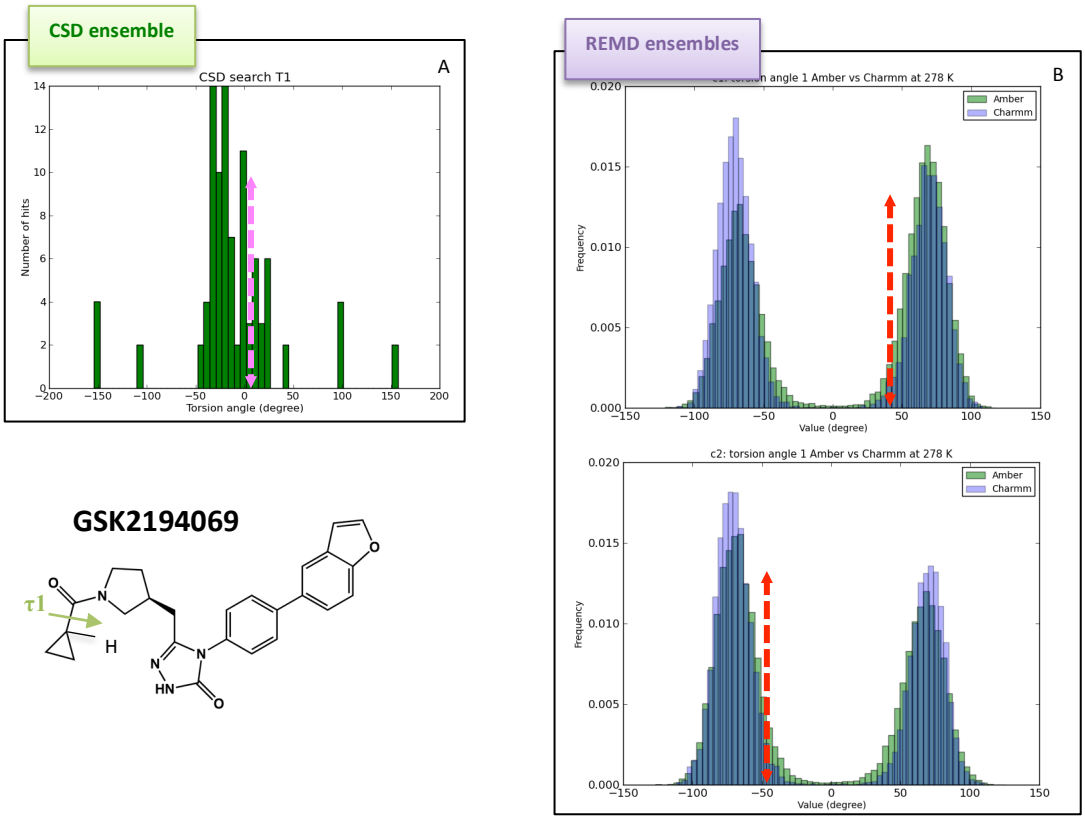


Figure 4.4: Torsion angle τ1 distribution in the CSD (A) and in the REMD trajectories of conformations 1 and 2 of GSK2194069 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K *(see Table 4.1 for definition)*. The magenta and red arrows describe the values of the torsion angles in the x-ray crystal structure and the starting conformations respectively.

After analysing the REMD trajectories at the highest temperature (481 K), no differences were observed *(Appendix A Figure A.1)*. This difference between the CSD and REMD is more likely to be due to the force fields paramerisation. Looking at the value of τ1 in the x-ray structure in *Figure 4.4* magenta arrow, it appears that the CSD result is in accordance with the x-ray. Further analysis on the torsion angle τ1 in the compound 4 and the NMR ensemble will provide us more information on the distribution of τ1.

4.4.2.1.2 Torsion angle τ2: amide isomerism

The torsion angle τ2 corresponds to a tertiary amide bond. An amide bond can adopt two main geometries, the E and the Z isomer with respective torsion angle of ≈ 0˚ and ≈ 180˚. The distinction between the two can be made according to the position of the cyclopropyl moiety illustrates in *Figure 4.5*. The red asterisk indicates the atom involves in the torsion amide enabling the differentiation between the two isomers.

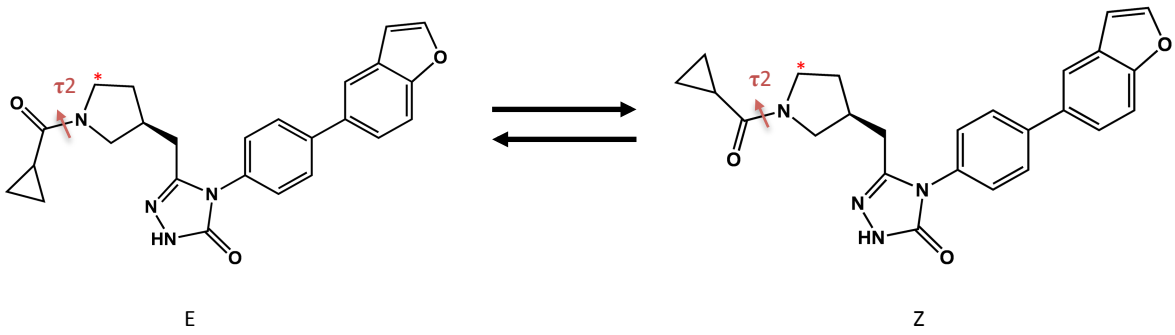


Figure 4.5: Z (180˚) and E (0˚) isomers definition in the tertiary amide bond of GSK2194069.

To minimize steric clashes between substituents, the Z isomer will appear to be more stable than the E isomer.

Figure 4.6 displays histograms for the torsion angle τ2 in the CSD and REMD ensembles. The REMD trajectories of the two starting conformations with the Amber Gaff and Charmm CGenff force fields in green and violet respectively are shown in *Figure 4.6B*. The CSD search highlights that the amide bond can adopt two main geometries, Z and E, with respective torsion angle values of ≈ 180˚/-180˚ and ≈ 0˚ (*Figure 4.6A)*. Starting with conformation 1, the REMD trajectories are able to reproduce the results expected from the CSD with both E and Z isomers observed with a higher population of the Z isomer. In contrast, conformation 2 simulations show trapping at 0˚; this might be caused by the torsion angle τ2 trapped in a local energy minimum despite the use of REMD and long simulations of 500 ns. Additionally, from *Figure 4.6*, starting with two different force fields similar results are observed so in the case of τ2 the force field does not influence the sampling.

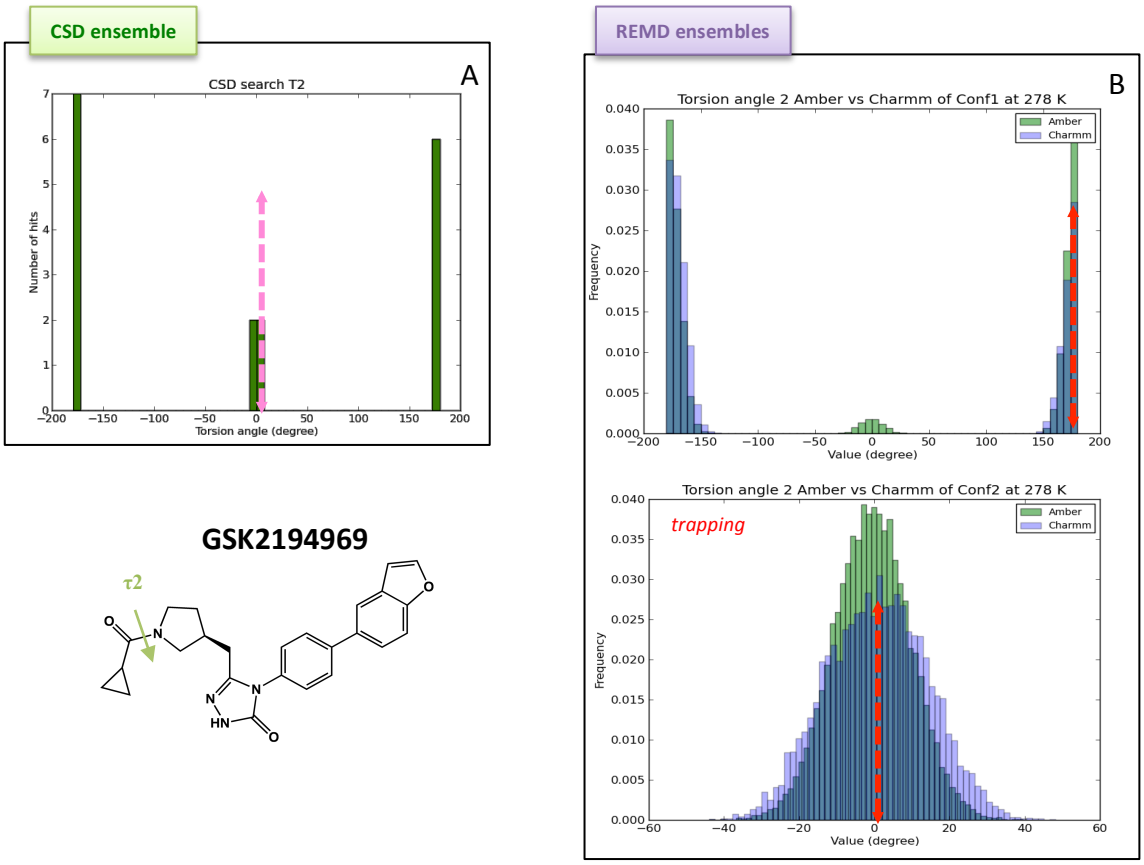


Figure 4.6: Torsion angle τ2 distribution in the CSD (A) and in the REMD trajectories of conformations 1 and 2 of GSK2194069 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The magenta and red arrows describe the values of the torsion angles in the x-ray crystal structure and the starting conformations respectively.

The distribution of torsion angle τ2 at the highest temperature has also been extracted to determine if trapping is observed. *Figure 4.7* shows the distribution of τ2 at 481 K, compared to the result form 278 K, starting from 0˚ and 180˚ sampling of both E and Z isomer, in a low population, are observed. Therefore, these results emphasised the highest temperature used in this case is sufficient to cross the energetic barrier by getting some sampling but the REMD trajectories are still not converged.

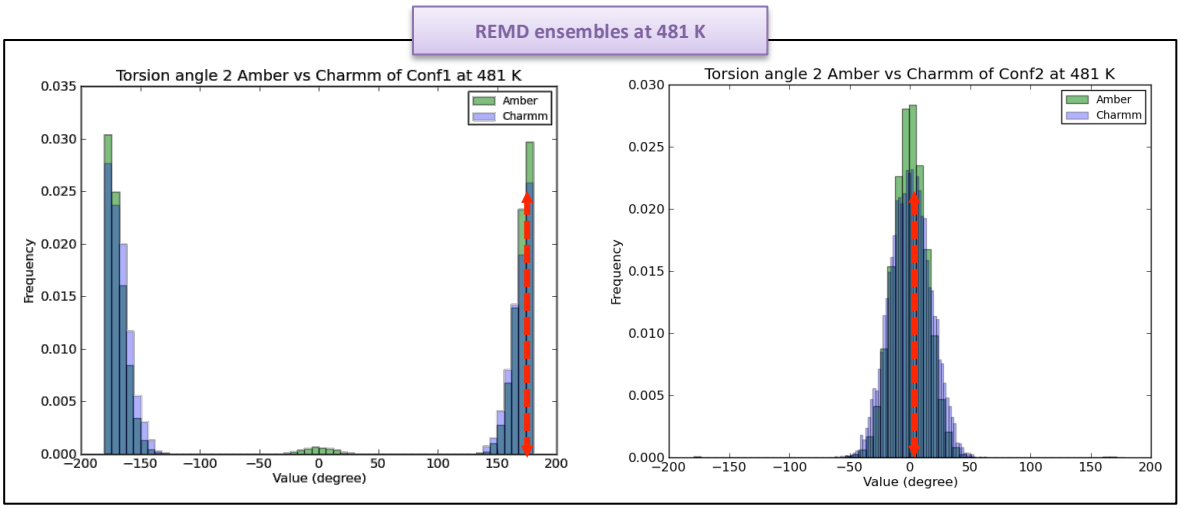


Figure 4.7: Torsion angle τ2 distribution in the REMD trajectories of conformations 1 and 2 of GSK2194069 in the Amber Gaff and Charmm CGenff force fields at 481 K. The red arrow describes the values of the torsion angles in the starting conformations.

4.4.2.1.3 Torsion angle τ3 and τ4:

Figure 4.8 displays the distribution of torsion angle τ3 and τ4 in the CSD search and in the REMD trajectories. No differences were observed between the two starting conformations in the REMD simulations, therefore only the results for conformation 1 in the Gaff and CGenff force fields are shown here. *Figure 3A* displayed the queries used to perform the CSD search. τ3 and τ4 queries required the use of a cyclopentane with double bonds *(Figure 4.3A green circle)* to obtain a sufficient number of hits making the results less specific. Indeed, zero hit were found when using cyclopentane with single bond.

REMD results for torsion angle τ3 appear broadly consistent with the CSD search and within the two force fields and both methods are able to capture the x-ray structure τ3 value. However, REMD trajectory result for τ4 display two main distributions around +/- 90˚ and are consistent within the two force fields, while CSD identifies four major torsional populations at +/- 150˚ and +/- 50˚. Further comparison with the NMR ensemble will help us understanding the distribution of τ4.

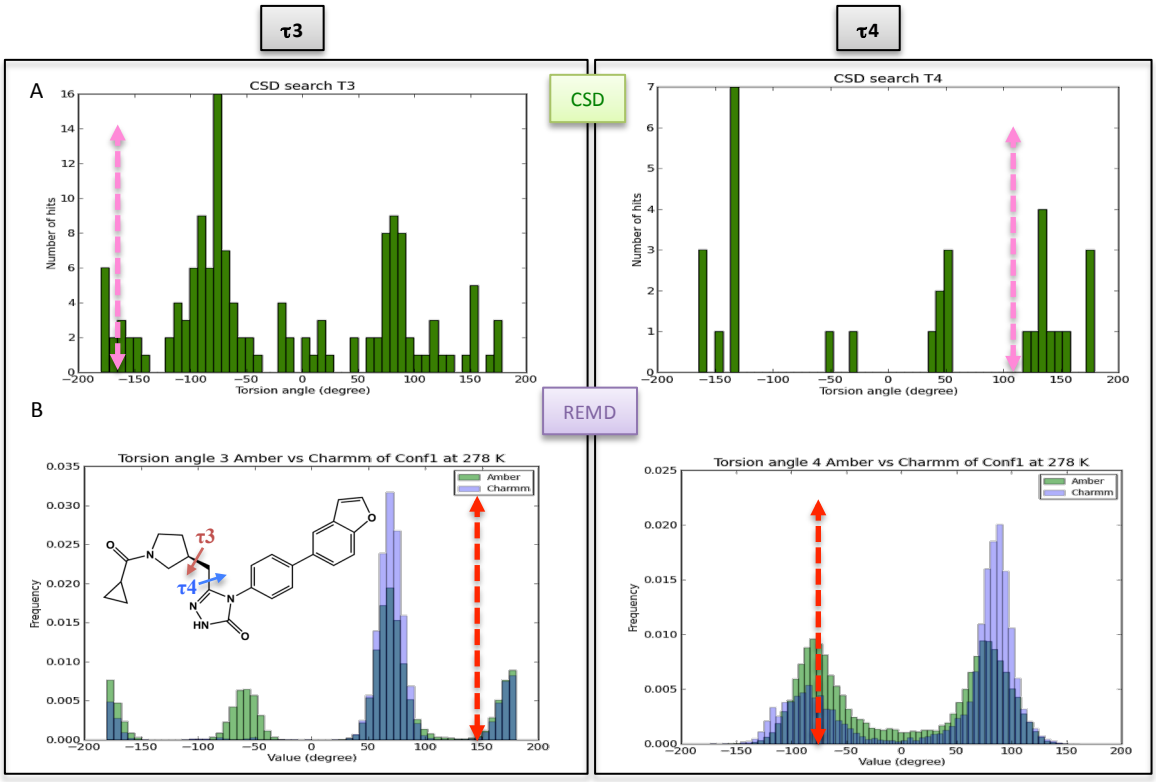


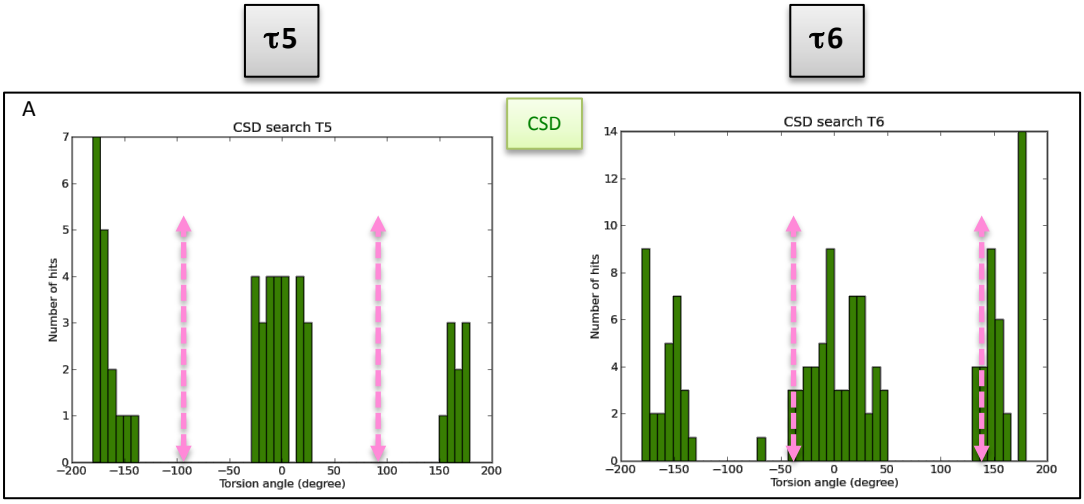
Figure 4.8: Torsion angles τ3 and τ4 distribution in the CSD (A) and in the REMD trajectories of conformation 1 of GSK2194069 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The magenta and red arrows describe the values of the torsion angles in the x-ray crystal structure and the starting conformations respectively.

Therefore, the cause of the difference in the torsion angle τ4 distributions is due to the query definition in the ConQuest program. The query may not be a good measure in this case.

4.4.2.1.4 Torsion angle τ5 and τ6: symmetry case

Figure 4.9 displays histograms of torsion angles τ5 and τ6 in the CSD and REMD ensembles. The two possibilities of the symmetry for the values of the x-ray crystal structure and the starting conformations torsion angle are shown in magenta and red arrows. The carbon atoms involved in the symmetry are shown in *Figure 4.9C* red asterisk. As for torsion angles τ3 and τ4, no differences between the two starting conformations results were observed so only conformations 2 are shown here.

Both torsion angles distributions from the REMD trajectories are not consistent with the CSD analysis. As can be seen in *Figure 4.9A*, the CSD distributions of τ5 and τ6 displays values around 0˚ while the REMD trajectories do not sample this region *(Figure 4.9B)*. However, comparing to the CSD, REMD simulations capture the x-ray torsion angle value for τ5. Additionally, the two force fields display some differences in the two torsion angles. In the distribution of τ5, two main torsions are populated with the Amber Gaff force field (green) while the torsions appear to be split when using the GCenff Charmm force field (violet). Further analysis with the NMR ensemble will enable us to evaluate if one of the force fields is performing better in the case of τ5 and τ6 dihedrals as well as understanding more the reason of the inconsistency with the CSD search.



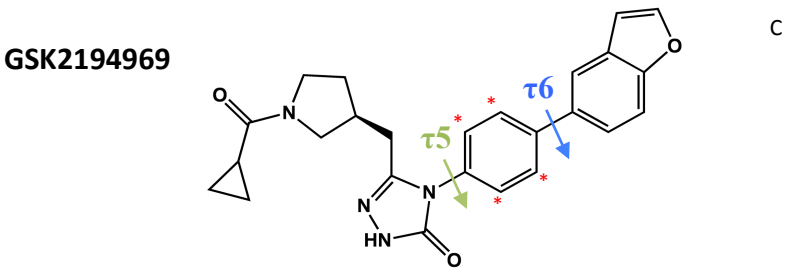


Figure 4.9: Torsion angles τ5 and τ6 distribution in the CSD (A) and in the REMD trajectories of conformation 2 of GSK2194069 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The magenta and red arrows describe the values of the torsion angles in the x-ray crystal structure and the starting conformations respectively including the two symmetry related possibilities.

Therefore, the difference observed between the CSD data and the REMD trajectories for torsions τ5 and τ6 is most likely to be caused by the force field parameterisation. Further comparison with the NMR ensemble will enable us to understand the sampling of these torsions.

4.4.2.2 Compound 2:

4.4.2.2.1 Torsion angle τ1 and τ2:

In REMD simulations, τ1 and τ2 adopt two main conformations around +/- 90˚, *Figure 4.10B (see Table 4.1 for torsions definition)*. However, variability is observed between these force fields. The Amber Gaff force field in green is shifted in the τ1 distribution and the area around 0˚ is sampled more than with the Charmm CGenFF force field in violet.

Looking at the CSD results *(Figure 4.10A)*, the Charmm CGenFF force field appears to be more consistent for the τ1 distribution with zero sampling observed around 0˚ in the CSD structures. However, sampling is observed between -40˚ to 0˚ for the τ2 distribution in the CSD, both force fields show populations close to this range. Similar results were observed within the starting conformations so only conformation 1 is presented here. The results for the conformation 2 can be found in *Appendix A Figure A.2.*

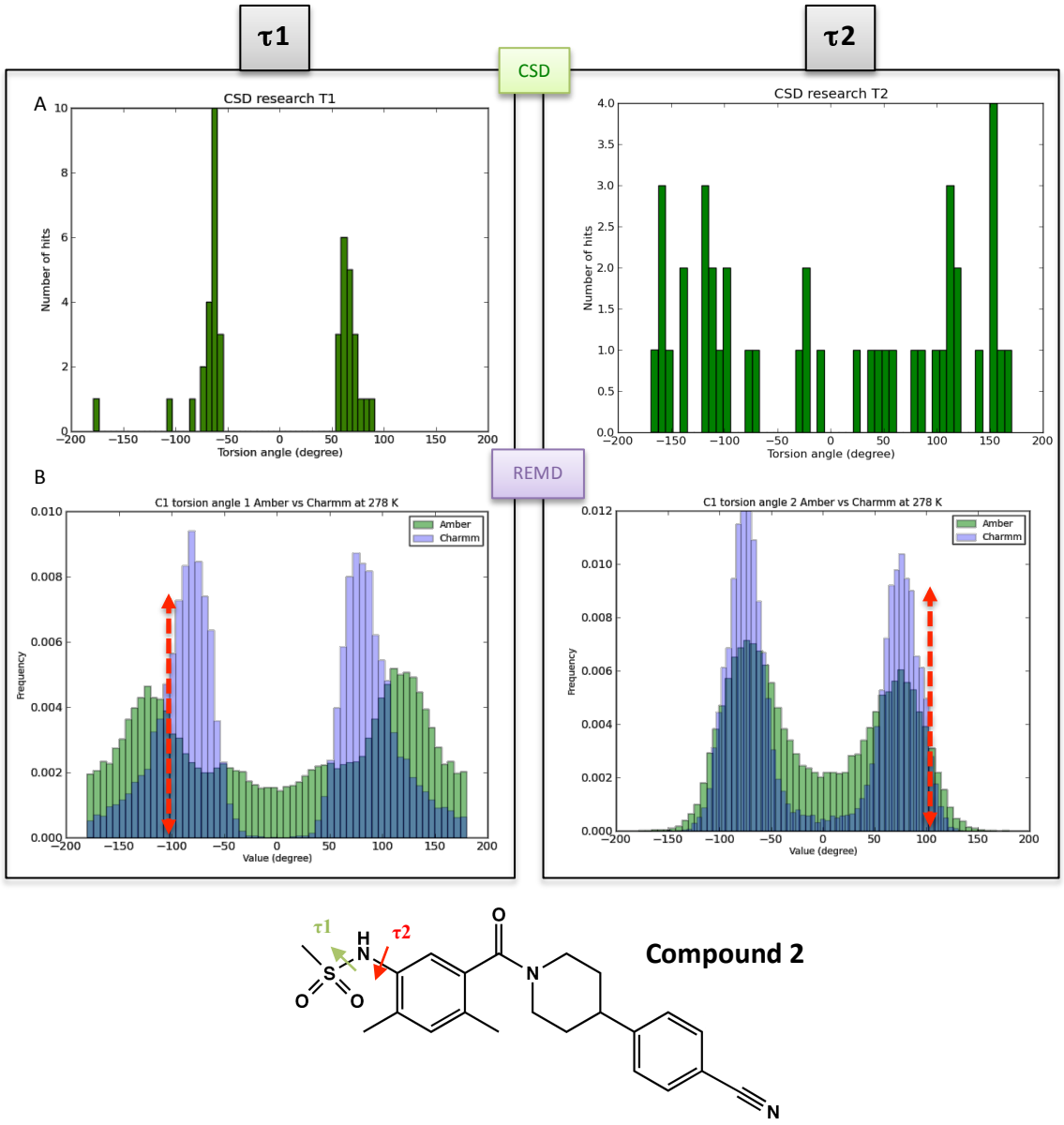


Figure 4.10: Torsion angles τ1 and τ2 distribution in the CSD (A) and in the REMD trajectories of conformation 1 of compound 2 in the Amber Gaff and Charmm CGenff force fields (B). The red arrows describe the values of the torsion angles in the starting conformation.

Inconsistency have been observed between the CSD and REMD data and consistency within the force fields for torsion angle τ2 while the opposite was seen for torsion angle τ1 with the results with the CGenFF Charmm force field being more consistent with CSD than Amber Gaff.

4.4.2.2.2 Torsion angle τ3: case of trapping

As observed with the GSK2194069 torsion angle τ2, the torsion angle τ3 suffers trapping while starting at 108° *(Figure 4.11B)* although when starting at -100° some sampling is observed at -100° which is consistent with the results observed with the CSD. Furthermore, both force fields are displaying similar results emphasizing that the trapping issue is not caused by force field parameterization itself.

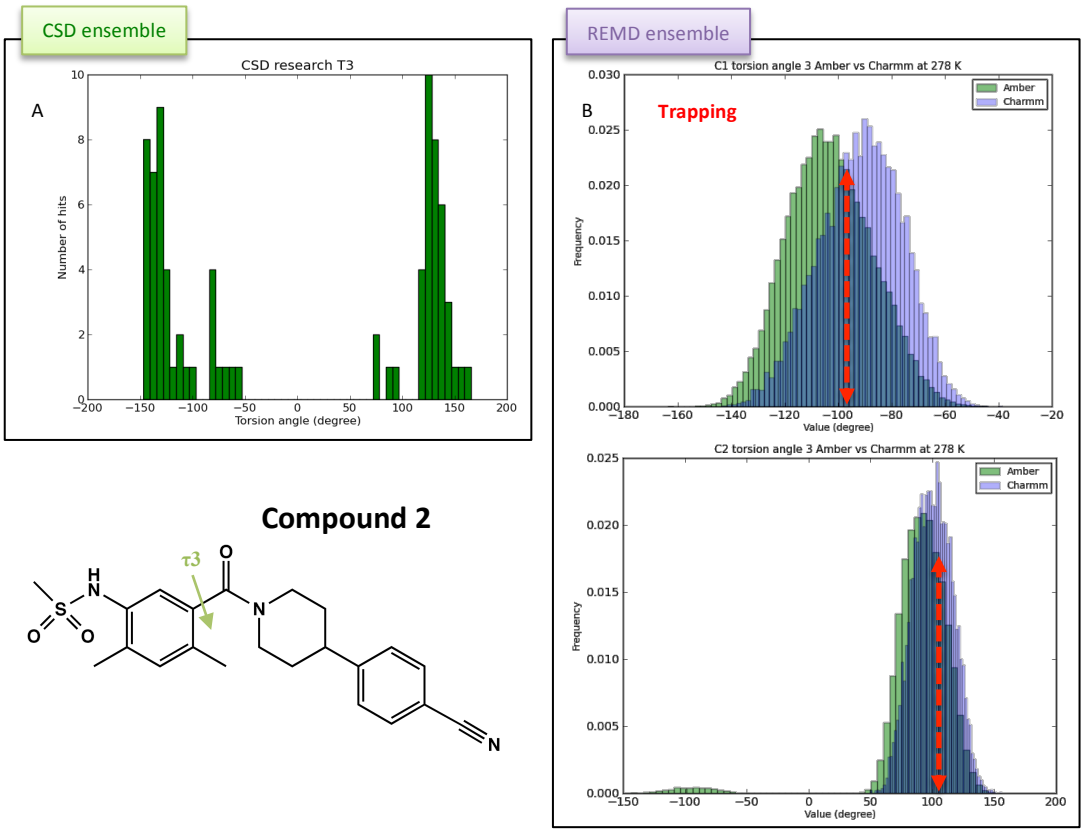


Figure 4.11: Torsion angle τ3 distribution in the CSD (A) and in the REMD trajectories of conformations 1 and 2 of compound 2 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The red arrows describe the values of the torsion angle in the starting conformations.

The distribution of τ3 at 481 K displays similar results as observed at 278 K *(Figure 4.12)*. In the case of compound 2 the choice of the 481.02 K does not seem to be high enough to overcome this trapping issue reflecting that the simulations are still not converged.

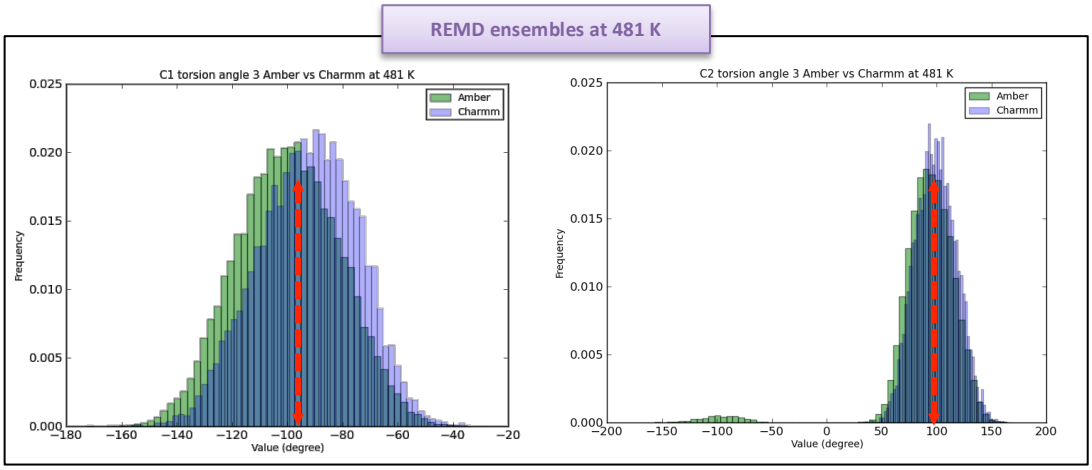


Figure 4.12: Torsion angle τ3 distribution in the REMD trajectories of conformations 1 and 2 of compound 2 in the Amber Gaff and Charmm CGenff force fields at 481 K. The red arrow describes the values of the torsion angles in the starting conformations.

Therefore, the choice of the starting conformation influence the sampling in this case causing the simulation to get trapped despite consistency with the CSD data.

4.4.2.2.3 Torsion angles τ4 and τ5: symmetry case

The values of the torsion angles, for each symmetry, are displayed in red arrows in *Figure 4.13* and the atoms involved are highlighted in red stars in *Figure 13C*. From *Figure 4.13*, both torsions appeared to broadly be in accordance with the CSD. The tertiary amide torsion angle τ4 is sampling both E and Z isomers as observed in the CSD search. The distribution of τ5 shows sampling around +/- 50˚ to +/- 150˚ with four main torsions at +/- 60˚ and +/-140˚ which are consistent with the results observed with the CSD.

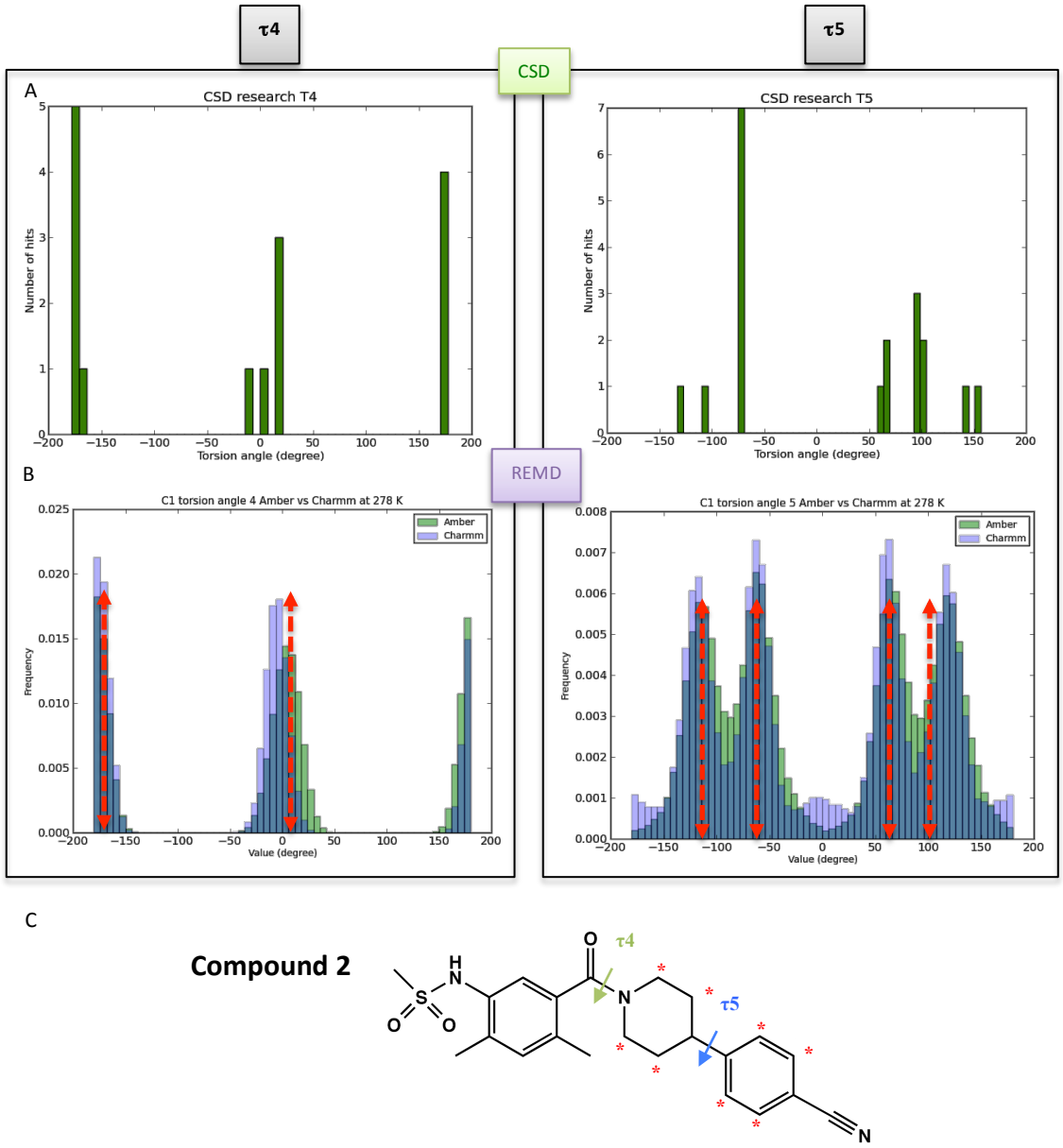


Figure 4.13: Torsion angles τ4 and τ5 distribution in the CSD (A) and in the REMD trajectories of conformations 1 and 2 of compound 2 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The red arrows describe the values of the torsion angles in the starting conformations. C: Structure of compound 2 with the red asterisks emphasizing the atoms involved in all symmetries.

Therefore, consistency is observed between the CSD and the REMD data and within the force fields with good sampling starting from both conformations 1 and 2. The symmetry of the torsion τ4 enables the sampling to sample both E and Z isomer.

4.4.2.3 Compound 3:

4.4.2.3.1 Torsion angle τ1, τ2 and τ3:

CSD distribution of torsions τ1, τ2 and τ3 is broadly symmetrical around +/- 60° to +/- 180° while the REMD simulations is showing bias sampling around the same regions but is not symmetrical with less sampling in the negative regions for torsions τ2 and τ3 *(Figure 4.14)*. The symmetrical character of the CSD data can easily be explain by the fact that comparing to us the torsions τ2 and τ3 appear symmetrical for the CSD search. The red asterisks in *Figure 4.14* indicate the atoms involve in the measure of torsions τ2 and τ3 induce an unsymmetrical distribution. Looking at the distributions of τ2 and τ3 at the highest temperature similar sampling as for 278 K was observed conforming the trapping issue of torsions τ2 and τ3 *(Appendix A Figure A.3)*.

Additionally, from *Figure 4.14* no major differences were observed within the force fields and the starting conformations so only the data for conformation 1 is shown here.

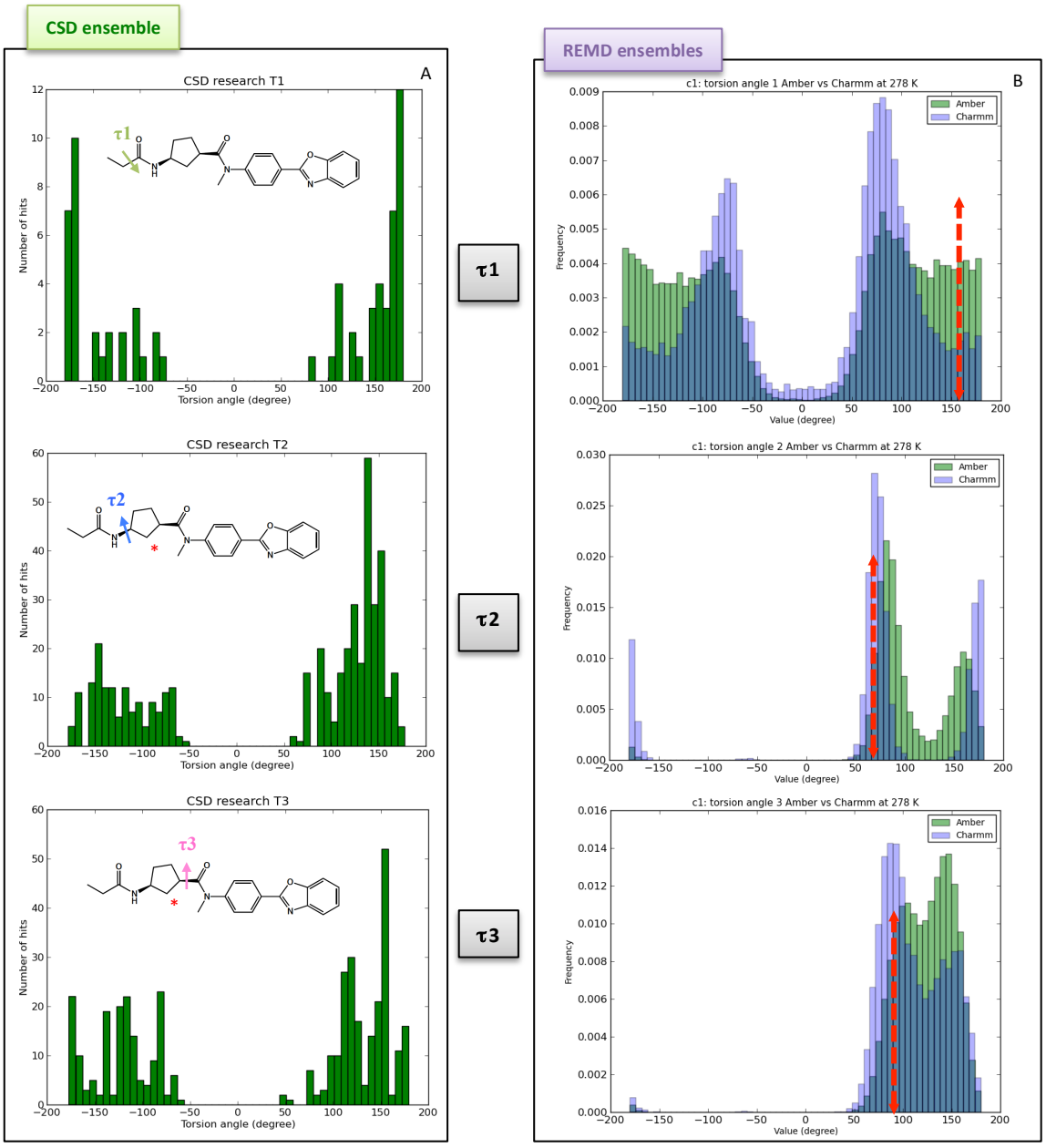
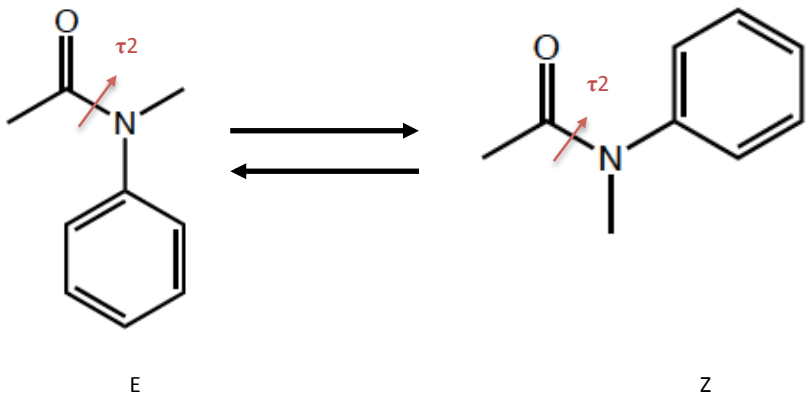


Figure 4.14: Torsion angles τ1, τ2 and τ3 distribution in the CSD (A) and in the REMD trajectories of conformation 1 of compound 3 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The red arrows describe the values of the torsion angles in the starting conformations.

Furthermore, despite consistency between the CSD data and the REMD simulations and within the force fields, the fact that poor sampling is observed in the negative values for torsions τ2 and τ3 is again due to trapping. The simulations are not converged and are not yet able to cross the energy barriers.

4.4.2.3.2 Torsion angle τ4: tertiary amide

As for the torsion angle 2 of GSK2194069, the compound 3 τ4 can adopt two isomers E and Z. The distinction between the E and Z isomers is based on the orientation of the phenyl ring as illustrated in *Figure 4.15*.



E Z

Figure 4.15: Z (180˚) and E (0˚) isomers definition in the tertiary amide bond of the compound 3.

The query used to perform the CSD search on torsion angle τ4, *Figure 4.3C*, led in structures only with the E isomer *(Figure 5.16A)*. Only 10 hits were found with multiple values for τ4, in this case, the query was build to be as similar as possible to the part of the compound studied, so the molecules found would provide the more accurate values for the torsion τ4. REMD simulations, comparing to CSD was able to sample both E and Z isomers in both starting conformations *(Figure 4.16B)*. However, according to the force field the sampling is more or less populated.

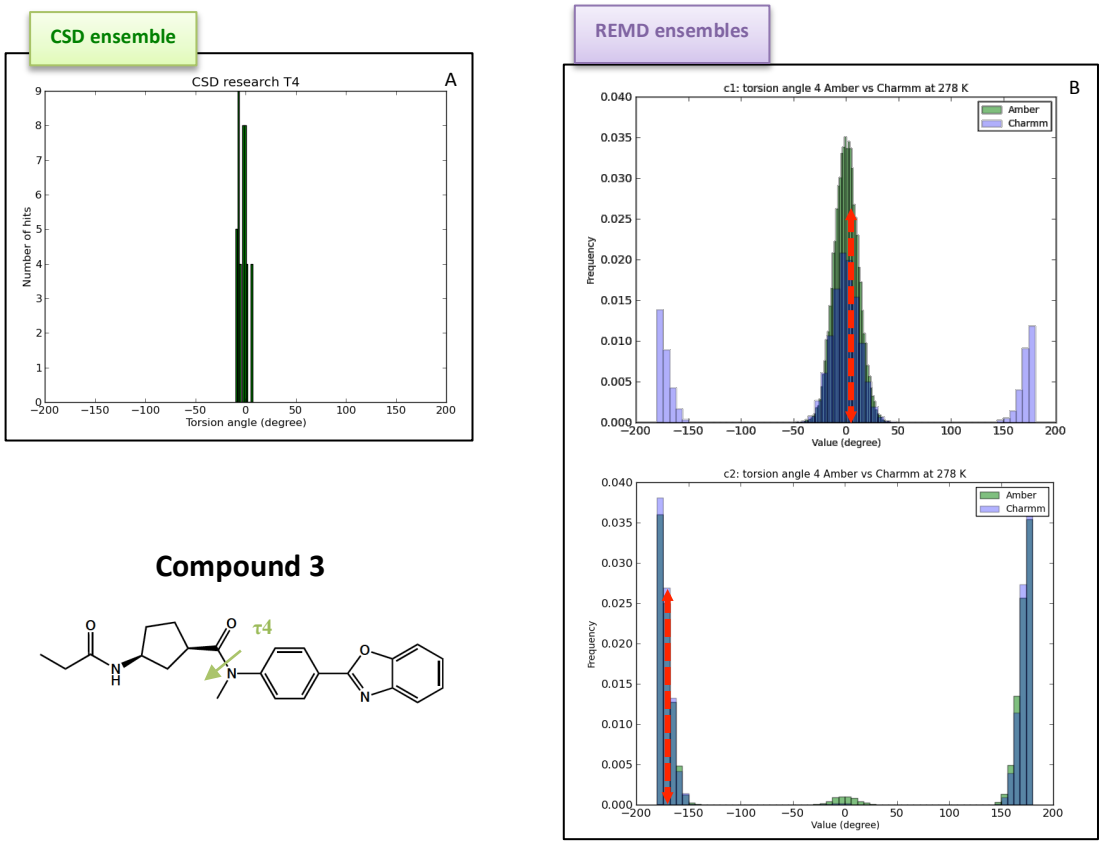


Figure 4.16: Torsion angle τ4 distribution in the CSD (A) and in the REMD trajectories of conformations 1 and 2 of compound 3 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The red arrows describe the values of the torsion angle in the starting conformations.

Despite a beginning of sampling of E isomers when starting in Z isomer and vice versa, the simulations are still not converged. Furthermore, the CGenFF force field appears to overcome the trapping faster than the CGenFF Charmm force field when starting in E isomer (conformation 1) and vice versa when stating in Z isomer (conformation 2). Therefore, it is not possible to determine if one force field perform better than the other.

4.4.2.3.3 Torsion angle τ5 and τ6: symmetry

The distributions of both torsion angles in the REMD simulations are in accordance with the CSD analysis *(Figure 4.17)*. Torsion angle τ5 adopts a symmetric ensemble at +/- 90˚ while τ6 displays two coplanar conformations at 0˚

and -180˚. No signs of trapping and differences within the two force fields and between the two starting conformations were observed.

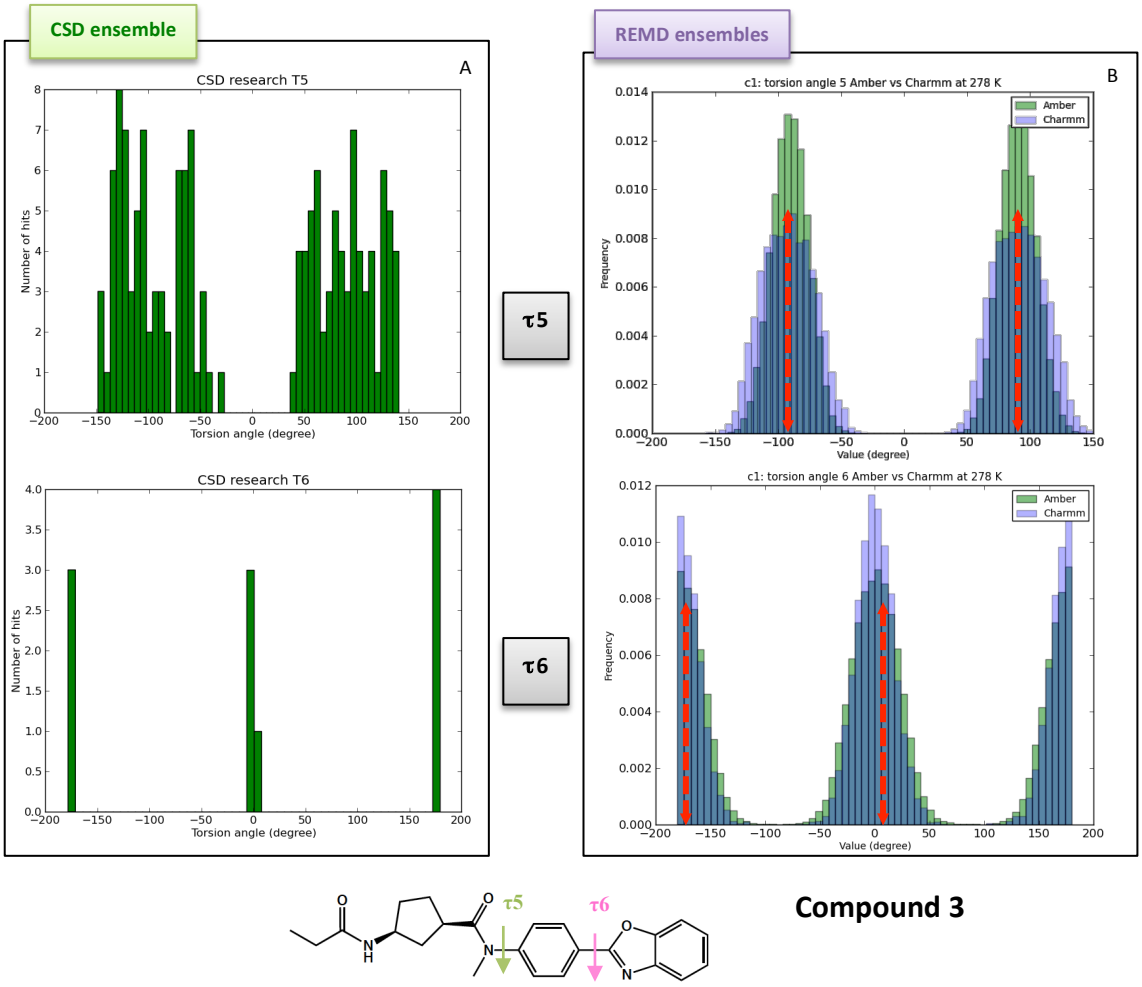


Figure 4.17: Torsion angles τ5 and τ6 distribution in the CSD (A) and in the REMD trajectories of the conformation 1 of compound 3 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The red arrows describe the values of the torsion angles in the starting conformations.

Therefore, the distributions of torsions τ5 and τ6 are consistent between the CSD data and the REMD trajectories and within the force fields and both starting conformations have shown good sampling.

4.4.2.4 Compound 4:

From Table 1, the structure of compound 4 shows similarities to GSK2194069; it was indeed designed as "scaffold-hop" from this molecule. However, compound 4 displays much lower potency, therefore exploring the differences in 3D conformations is highly valuable to see if it can explain or rationalise this. Given that the compounds share some key torsions, same query was used for the CSD search for the torsion angles τ1 and τ2 *(Figure 4.3A and 4.3D)*.

4.4.2.4.1 Torsion angle τ1:

As observed with GSK2194069 *(Figure 4.4)*, the distribution of the torsion angle τ1 is inconsistent with the CSD search *(Figure 4.18)*. The REMD simulations are consistent within the two starting conformations and within force fields, which enable us to hypothesize that the force fields parameterization influences the sampling and is not well parameterized for this torsion angle.

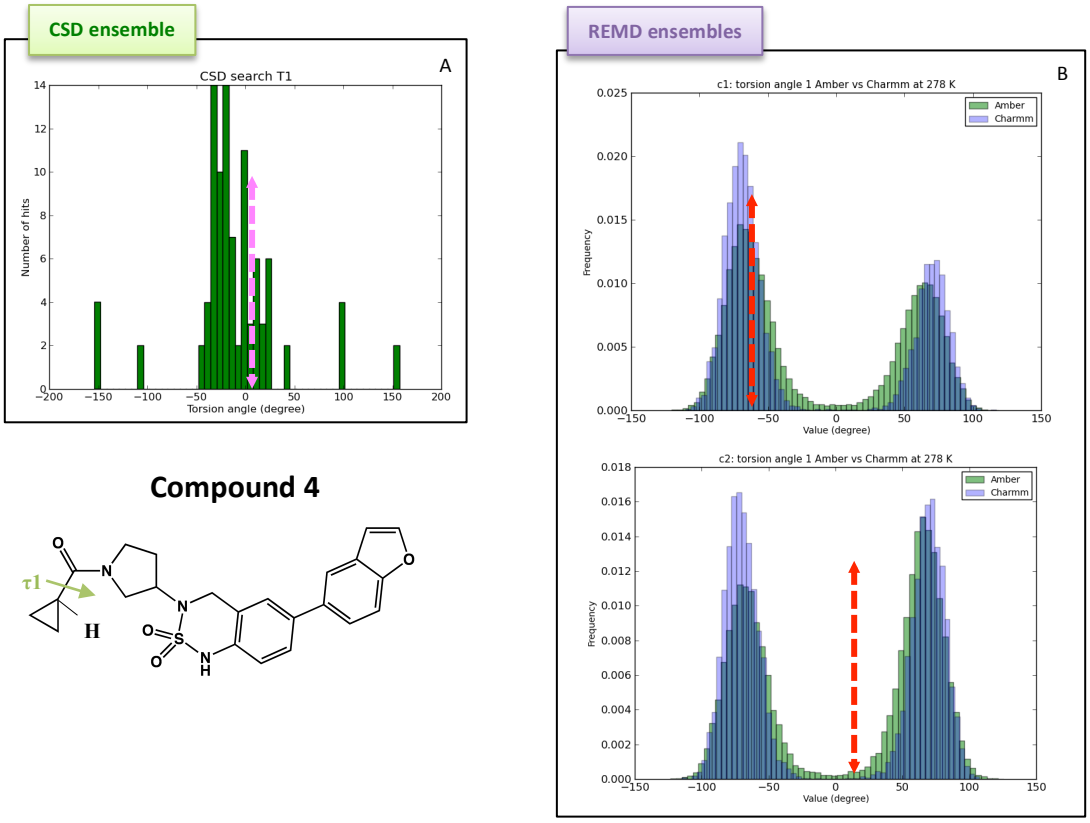


Figure 4.18: Torsion angle τ1 distribution in the CSD (A) and in the REMD trajectories of conformations 1 and 2 of compound 4 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The magenta

and red arrows describe the values of the torsion angles in the x-ray crystal structure and the starting conformations respectively.

4.4.2.4.2 Torsion angle τ2: tertiary amide

From the GSK2194069 results of torsion angle τ2 *(Figures 4.5B and 4.6B)*, the E isomer has shown trapping at 278 K and started to sample the Z isomer at 481 K. However, compound 4 shows different behaviour. Starting with E isomer *(Figure 4.19B conformation 2)*, a beginning of sampling of Z isomer is observed while no E isomer is sampled when starting with Z isomer.

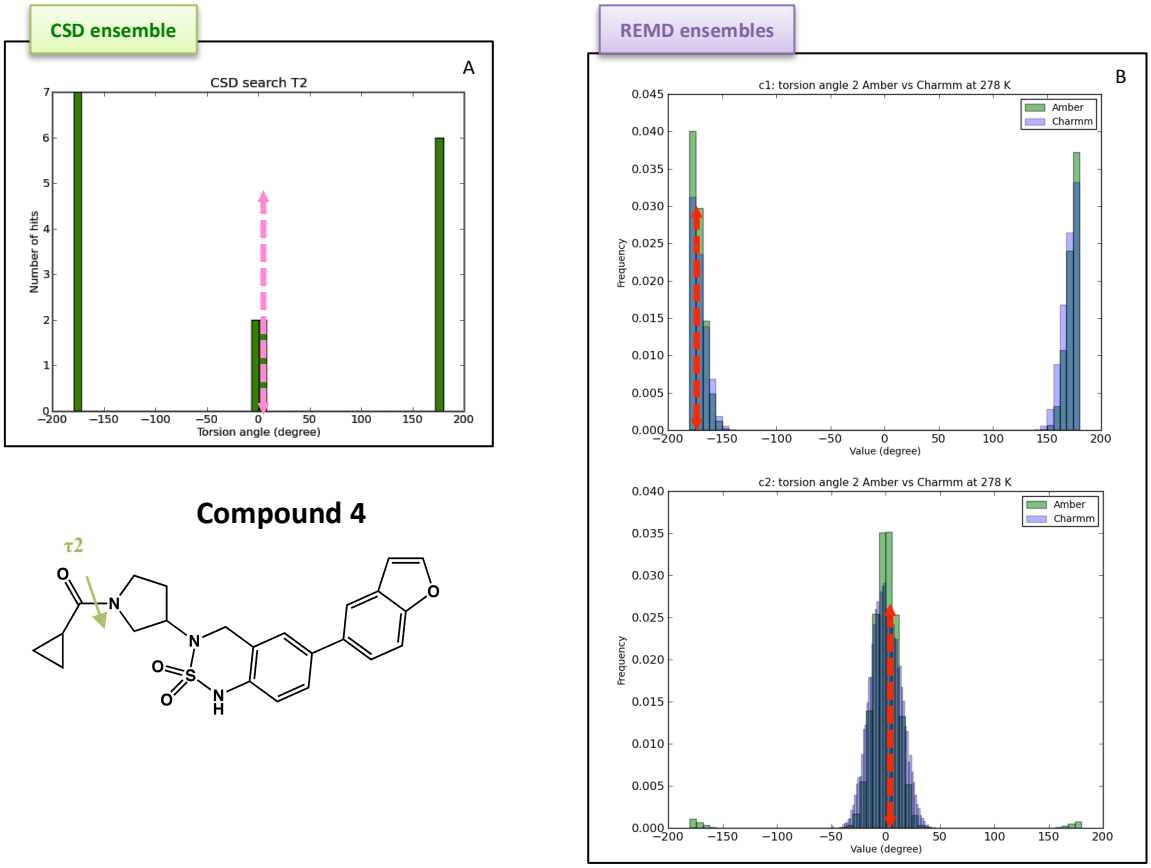


Figure 4.19: Torsion angle τ2 distribution in the CSD (A) and in the REMD trajectories of conformations 1 and 2 of compound 4 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The magenta and red arrows describe the values of the torsion angles in the x-ray crystal structure and the starting conformations respectively.

Looking at the distribution of τ2 at the highest temperature, the lack of sampling of E isomer is still observed when starting with Z isomer *(Figure 4.20)*. The results suggest trapping as it was observed for the amide torsions of GSK2194069 and the compound 3. Additionally, from the highest and the lowest simulations the results are consistent within the force fields and between the starting conformations.

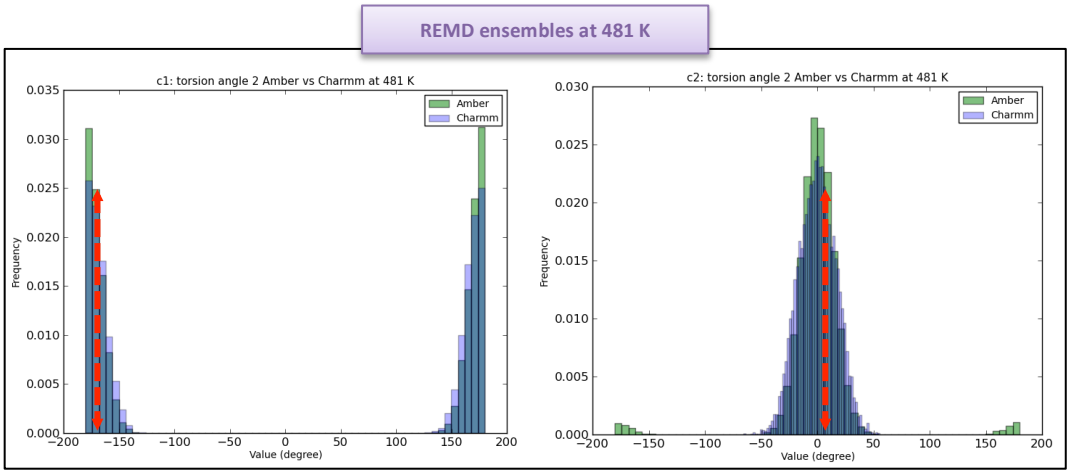


Figure 4.20: Torsion angle τ2 distribution in the REMD trajectories of the conformations 1 and 2 of compound 4 in the Amber Gaff and Charmm CGenff force fields at 481 K. The red arrow describes the values of the torsion angles in the starting conformations.

4.4.2.4.3 Torsion angle τ3:

Figure 4.21 shows the distributions of torsion angle τ3 in the CSD (A) and in the REMD trajectories of conformations 1 and 2 in the Amber Gaff and Charmm CGenFF force fields (B). The two methods are consistent within each other by both sampling around +/- 150 ˚ and between +/- 50˚ to +/- 100˚.

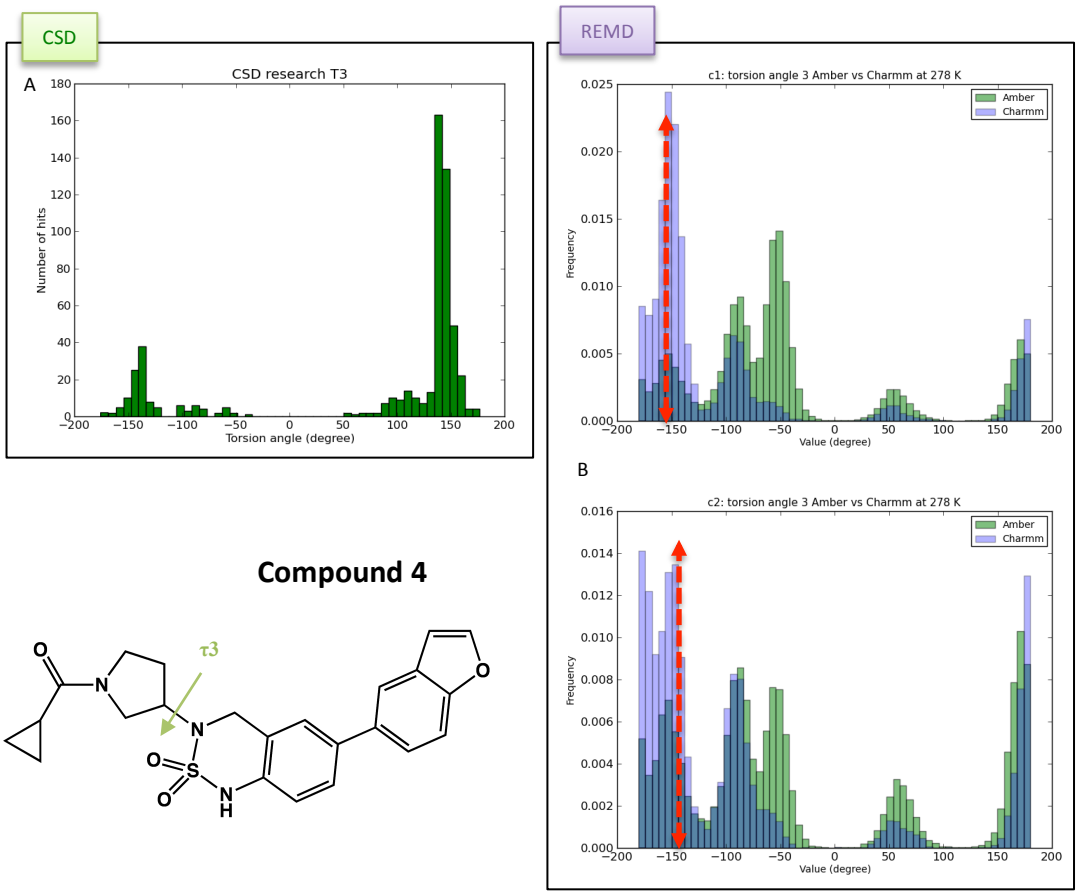


Figure 4.21: Torsion angle τ3 distribution in the CSD (A) and in the REMD trajectories of conformations 1 and 2 of compound 4 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The red arrows describe the values of the torsion angle in the starting conformations.

4.4.2.4.4 Torsion angle τ4:

Building a query for the torsion angle τ4 in the CSD was challenging due to the lack of suitable available data. Only two queries identify two structures in the CSD to compare with the REMD trajectories as shown in *Figure 4.3D* (with only 2 hits found with one hit in each query). The hit found for the first query offered a broad range of torsion angle values while the hit for the second query only had one torsion angle value as seen with the red asterisks in *Figure 4.22*. Despite only having two structures with multiple values for the torsion angle τ4 to be compared with the REMD simulations, the results appear to be consistent within each other with four main picks observed around +/- 150 ˚ and +/- 60˚ in both ensembles.

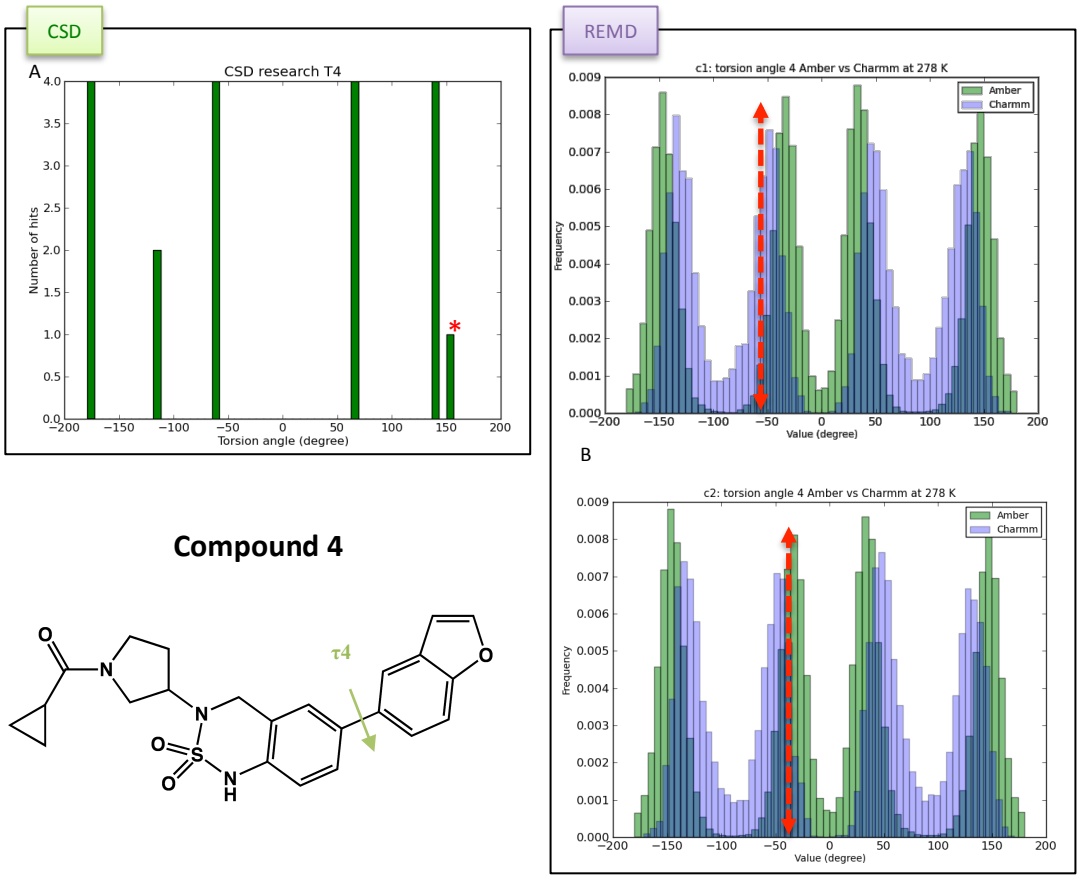


Figure 4.22: Torsion angle τ4 distribution in the CSD (A) and in the REMD trajectories of conformations 1 and 2 of compound 4 in the Amber Gaff and Charmm CGenff force fields (B) at 278 K. The red arrows describe the values of the torsion angle in the starting conformations.

Despite only having two queries from CSD to compare our REMD simulations, the distributions are consistent between the two. Additionally, consistency is observed within the two force fields and the two starting conformations have shown good sampling.

- Comparison between GSK2194069 and compound 4:

Comparing GSK2194069 and compound 4 analyses shown similar sampling in the REMD simulations for torsion τ1 and the torsions τ6 and τ4 with the τ6 and τ4 referring to the same torsion angle in GSK2194069 and compound 4 respectively. However, inconsistency was observed with the CSD data for both compounds for

these two torsion angles. The analysis of the NMR data in the next chapter will enable us to determine if the CSD data or the force fields parameterizations provide realistic sampling.

In the case of the amide bond (torsion $\tau 2$), compound 4 has shown better sampling with the ability of sampling both E and Z isomers at 278 K when starting with the E isomer which could be influenced by the sulphonamide moiety. However, both compounds shown trapping issue when starting from the Z isomer with the beginning of the sampling of E isomer with GSK2194069. Additionally, the analysis of the torsion $\tau 2$ at the highest temperature did not solve the trapping issue, which told us that the two simulations have not converged even after 500 ns.

The torsion angle that differs between GSK2194069 and compound 4, in term of functional groups definition, corresponds to torsion $\tau 3$ with the difference of having a triazolone moiety in GSK2194069 and a sulphonamide group for compound 4 in their central core. Thus, the torsion angle $\tau 3$ is not identical in the two compounds. However, the distribution of torsion $\tau 3$ is similar in GSK2194069 and compound 4 with similar torsion angle values regions sampled. Additionally, the two compounds have demonstrated consistency with the CSD data.

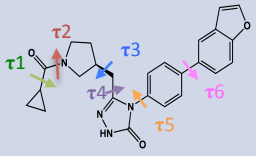
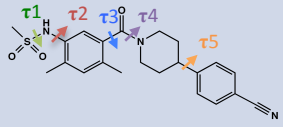
Therefore, the comparison between GSK2194069 and compound 4 did not show major differences from the torsion angles distributions that are shared within the two. The only difference that could explain the lower potency of compound 4 is the triazolone and sulphonamide moieties. Further analysis on the interactions such as possible clashes or a decrease in contacts made with the receptor will enable a better understanding of the cause of the low potency.

4.5 Discussion:

The results from this chapter demonstrate that the choice of temperatures is crucial for optimal performances of the REMD simulation. The use of 48 temperatures with the highest temperature of 481 K emphasized good exchanges between the replicas *(Figure 4.2)* and a good agreement between the acceptance rate desired and observed. However, local minima trapping was often observed suggesting that in some cases the highest temperature was not high enough to cross the energetic barrier and that the trajectories were not yet converged even after long runs.

Additionally, torsion angles distribution comparison with experimentally determined values obtained using the Cambridge Structural database emphasized

different issues according to the force field and the chemical group studied as summarized in *Table 4.2*.

Compounds	CSD	REMD Gaff	REMD CGenFF
Compound 1 Conformation 1	**τ1**: 0 ° **τ2**: 0 ° , +/-180 ° (E/Z) **τ3**: - 90 ° ↑ **τ4**: +/- 50 ° , +/- 150 ° **τ5**: 0 ° , +/- 180 ° **τ6**: 0 ° , +/- 180 °	**τ1**: +/- 90 ° **τ2**: 0 ° , +/-180 ° (E/Z) **τ3**: +90 ° ↑ **τ4**: +/- 60 ° **τ5**: +/- 90 ° **τ6**: +/- 40° , +/- 140 ° (Shifted)	**τ1**: +/- 90 ° **τ2**: +/- 180 ° (Trapped) **τ3**: + 90° **τ4**: +/- 60 ° **τ5**: +/- 100 ° (split) **τ6**: +/- 40° , +/- 140 ° (Shifted)
Compound 1 Conformation 2		**τ1**: +/- 90 ° **τ2**: 0 ° (trapped) **τ3**: +90 ° ↑ **τ4**: +/- 60 ° **τ5**: +/- 90 ° **τ6**: +/- 40° , +/- 140 ° (Shifted)	**τ1**: +/- 90 ° **τ2**: 0° (trapped) **τ3**: +90 ° ↑ **τ4**: +/- 60 ° **τ5**: +/- 90 ° (split) **τ6**: +/- 40° , +/- 140 ° (Shifted)
Compound 2 Conformation 1	**τ1**: +/- 90 ° (no 0 °) **τ2**: broad sampling **τ3**: +/- 90 ° **τ4**: 0 ° ; +/-180 ° (E/Z) **τ5**: +/-60 ° ; +/- 140 °	**τ1**: 0° , +/- 90 ° **τ2**: +/- 90 ° (sampling around 0°) **τ3**: - 90 ° (trapped) **τ4**: 0 ° , +/-180 ° (E/Z) **τ5**: +/-60 ° , +/- 140 °	**τ1**: +/- 90 ° (no 0 °) **τ2**: +/- 90 ° (sampling around 0°) **τ3**: - 90 ° (trapped) **τ4**: 0 ° , +/-180 ° (E/Z) **τ5**: +/-60 ° , +/- 140 °
Compound 2 Conformation 2		**τ1**: +0 °, /- 90 ° **τ2**: +/- 90 °(sampling around 0°) **τ3**: + 90 ° ↑ , - 90 ° (begin sampling) **τ4**: 0 ° , +/-180 ° (E/Z) **τ5**: +/-60 ° , +/- 140 °	**τ1**: +/- 90 ° (no 0 °) **τ2**: +/- 90 °(sampling around 0°) **τ3**: + 90 ° ↑ , - 90 ° (begin sampling) **τ4**: 0 ° , +/-180 ° (E/Z) **τ5**: +/-60 ° , +/- 140 °

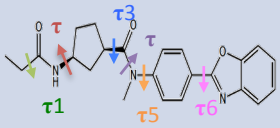
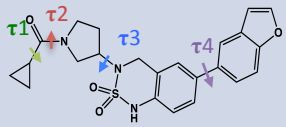
Compounds	CSD	REMD Gaff	REMD CGenFF
Compound 3 Conformation 1	**τ1**: +/- 90 ° ; +/ - 180 ° **τ2**: +/- 100 °; +/-150 ° **τ3**: +/- 100 °; +/-150 ° **τ4**: 0 ° (E) **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 °	**τ1**: +/- 90 ° ↑ **τ2**: [100 °; 150 °]↑; [-100 °; -150 °]↓ **τ3**: [100 °; 150 °]↑ **τ4**: 0 ° (E) Trapping **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 °	**τ1**: +/- 90 ° ↑ **τ2**: [100 °; 150 °]↑; [-100 °; -150 °]↓ **T3:** [100 °; 150 °]↑ **τ4**: 0 ° ; +/-180 ° ↑ (E/Z) **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 °
Compound 3 Conformation 2		**τ1**: +/- 90 ° ↑ **τ2**: [100 °; 150 °]↑; [-100 °; -150 °] ↓ **τ3**: [100 °; 150 °]↑ **τ4**: 0 ° ; +/-180 ° ↑ (E/Z) **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 °	**τ1**: +/- 90 ° ↑ **τ2**: [100 °; 150 °]↑; [-100 °; -150 °]↓ **τ3**: [100 °; 150 °]↑ **τ4**: +/-180 ° (Z) Trapping **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 °
Compound 4 Conformation 1	**τ1**: 0 ° **τ2**: 0 ; +/-180 ° (E/Z) **τ3**: +/- 100 ° ; +/- 150 ° (↑ 150°) **τ4**: +/- 60 °; +/- 150 °	**τ1**: +/- 90 ° **τ2**: +/-180 ° (Z) Trapping **τ3**: +/- 100 °; +/- 150 ° (↑ -150°) **τ4**: +/- 60 °; +/- 150 °	**τ1**: +/- 90 ° **τ2**: +/-180 ° (Z) Trapping **τ3**: +/- 100 °; +/- 150 ° (↑ -150°) **τ4**: +/- 60 °; +/- 150 ° (shifted)
Compound 4 Conformation 2		**τ1**: +/- 90 ° **τ2**: 0° ↑ E ; +/-180 ° Z trapped **τ3**: + /- 100 °; +/- 150 ° (↑ -150°) **τ4**: +/- 60 °; +/- 150 °	**τ1**: +/- 90 ° **τ2**: 0 (E) Trapping **τ3**: +/- 100 °; +/- 150 ° (↑ -150°) **τ4**: +/- 60 °; +/- 150 ° (shifted)

Table 4.2: Torsion angles comparison between the CSD structures, the REMD simulations with the Amber Gaff and Charm CGenFF force fields and the NMR ensemble for GSK2194069 and compounds 2, 3 and 4. The arrows ↑ indicate a most populated distribution while the arrows ↓ indicate a less populated distribution.

All compounds have an amide group in their 3D structures. However only compound 2 did not show trapping issue during the sampling of the amide. The cyclopropyl moiety in GSK2194069 and compound 4 has also shown difficulty to reproduce the CSD data and could influence the trapping observed with amide. Similarly, the distributions of torsions $\tau 2$ and $\tau 3$ of the compound 3 have shown poor sampling in the negative torsion angles emphasizing again trapping and the data are not converged.

In term of force field, it is difficult to determine if one perform better than the other. They both sampled similar torsion angle areas and only the population differ. The major differences between the two were observed first with GSK2194069 and compound 4 when dealing with the benzofuran moiety. The torsion angle distribution is shifted between the two force fields, which can be explained by the force fields parameterization. Then, the distribution of torsion angle $\tau 1$ in compound 2 also shown difference within the two force fields with a better sampling with the CGenff force field in term of consistency with the CSD data. The sulphonamide parameterization in the CGenff force field seems to be more adequate in this study. Analysis of the NMR data will help us determine if it is due to the force fields parameterizations or is it the chemical group itself that appears to be challenging in terms of achieving adequate sampling.

Chapter 5 Comparing computational and NMR data of inhibitors

5.1 Introduction:

Small molecule inhibitors often have a number of rotatable bonds and will adopt numerous conformations in solution. NMR methods can be used to determine the solution conformations of compounds and generate an ensemble of structures in solution that explore the conformational space of our system of interest. C4X's NMR technology [77] aims to characterise the conformations a ligand will adopt in solution, which can be then compared to a binding mode if known. A report is provided for each compound with information for each degree of freedom referring to the preferred conformations and torsion angle distributions. Combining experimental and computational data from NMR and REMD simulations will improve our understanding of these systems and we can apply that learning to others. Understanding the range of motions as well as the favourable conformational states may provide us with valuable information to improve the ligand affinity.

This chapter aims to compare the solution phase data of GSK2194069 and compounds 2,3 and 4 from NMR experiments with the described REMD ensembles. As performed in Chapter 4, the torsion angle distributions were analysed and compared between the methods. A principal component analysis (PCA) was conducted over the entire REMD simulations and the NMR ensemble to evaluate how well the computational method was able to reproduce the experimental data. Comparing the population of the NMR ensemble with the REMD trajectories followed the PCA analysis. The analysis was completed with the use of a conformational generator using OMEGA from OpenEye in order to see if a faster method was able to reproduce the correct conformational distribution.

5.2 Sampling comparison between NMR and REMD ensembles:

5.2.1 GSK2194069:

5.2.1.1 Torsion angle τ1:

From *section 4.4.2.1.1*, the comparison of the distribution of τ1 between the CSD data and the REMD ensembles emphasized the different conformation sampled between the two. However, the distribution of τ1 in the NMR ensemble *(Figure 5.1B)* is consistent with the CSD and with the x-ray suggesting than the REMD ensembles do not describe the conformations phase of τ1 to a satisfactory level.

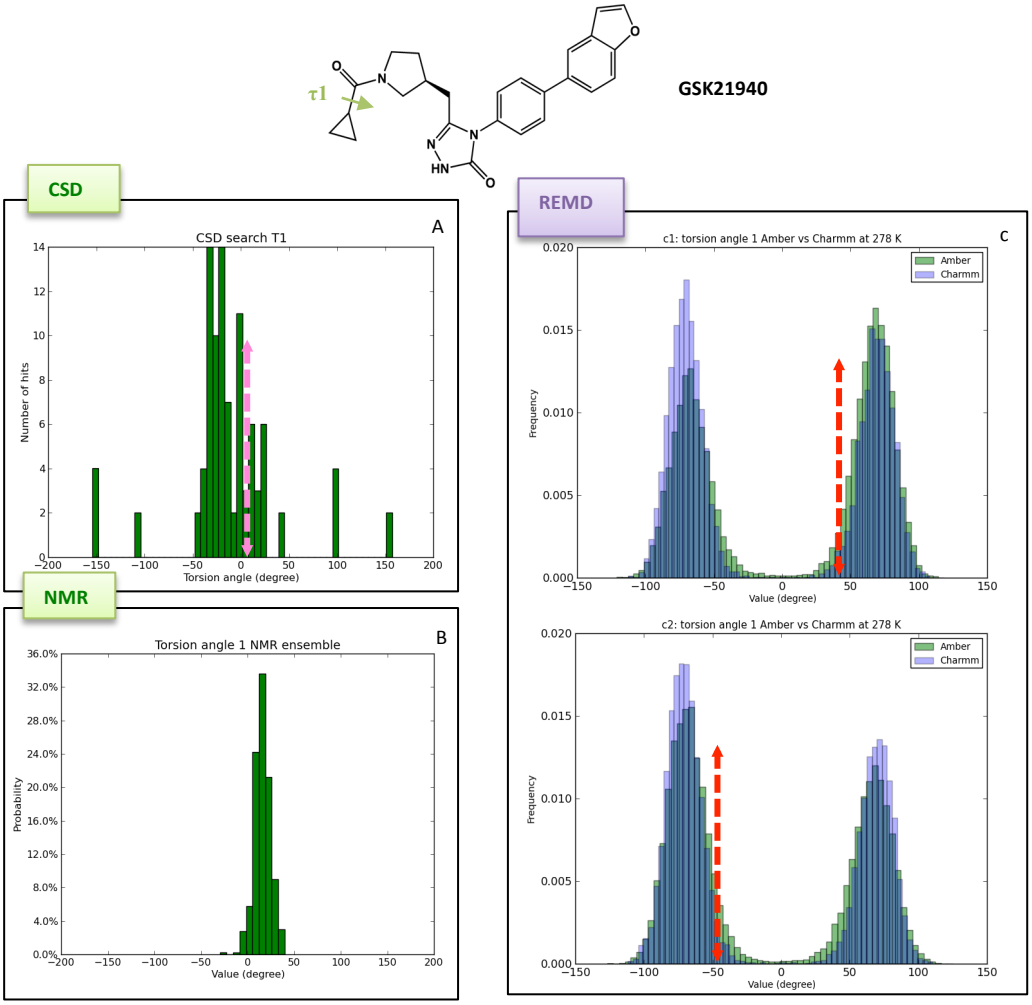


Figure 5.1: Torsion angle τ1 distribution in the NMR ensemble (A), the REMD trajectories of conformations 1 and 2 of GSK2194969 in the Amber Gaff and Charmm CGenff force fields in green and violet respectively (B) and in the CSD database (C). The magenta and the

red arrows show the values of τ1 in the x-ray and the initial conformations respectively.

Therefore, the analysis of the τ1 emphasized that the parameterization of the Gaff and CGenFF force fields are not good to sample accurate distribution of the τ1.

5.2.1.2 Torsion angle τ2:

The torsion angle τ2 corresponds to the tertiary amide bond. *Figure 5.2* shows the distributions of τ2 in the NMR (A), the REMD (B) ensembles and the CSD database (C) with the magenta and red arrows corresponding to the values of τ2 in the x-ray structure and the REMD starting conformations 1 and 2 respectively. In the REMD simulations, starting from the E isomer shows trapping issue while starting from Z a beginning of sampling of the E isomer is observed with the Amber Gaff force field. The highest temperature trajectory studied in *Chapter 4 section 4.2.1.2 Figure 4.7* highlighted that the conformations started to overcome the amide trapping but the simulations were not converged. In the NMR ensemble the amide is in slow exchange resulting in two equally populated amide isomers E and Z (E:Z – 50:50) which is consistent with CSD search results.

The combination of CSD and NMR data shows that GSK2194069 can adopt both isomers in equal population. Starting the REMD simulations with the E isomer appears to be a major problem. Additionally, the REMD results with the Amber Gaff force field shows beginning of sampling while starting with Z isomer.

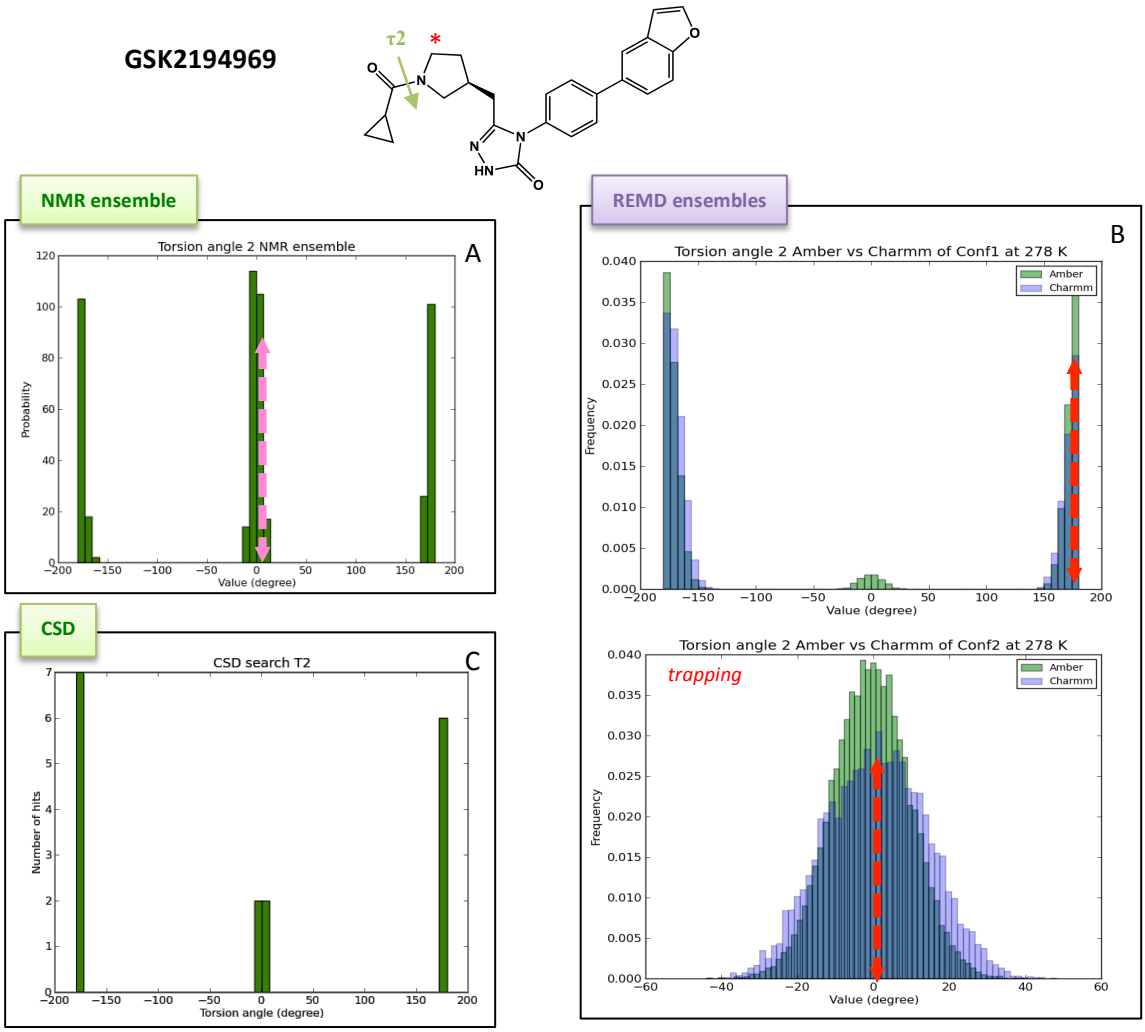


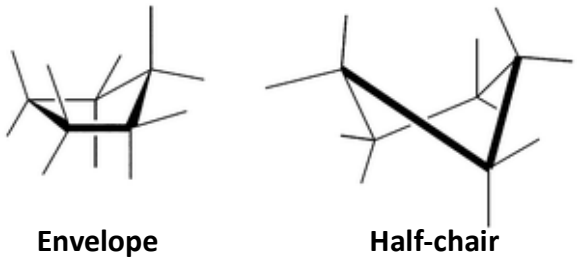
Figure 5.2: Torsion angle τ2 distribution in the NMR ensemble (A), the REMD trajectories of conformations 1 and 2 of GSK2194969 in the Amber Gaff and Charmm CGenff force fields in green and violet respectively (B) and in the CSD database (C). The magenta and the red arrows show the values of τ2 in the x-ray and the initial conformations respectively. The red asterisk shows the pyrrolidine atom involved in the defining torsion angle.

Therefore, consistency have been observed between the CSD, REMD and NMR data, however, despite a beginning of E isomer while starting in conformation 1, the REMD simulations are still not converged.

5.2.1.3 The pyrrolidine ring conformation:

Ring systems refer to a connected set of atoms forming a closed chain, which limits degrees of freedom and so has the possibility to be distorted. In the case

of the pyrrolidine ring, to minimise the steric interactions due to the eclipsed position of all substituents, a slight distortion is possible for the ring. Therefore to reduce the energy of the system, the planar form of the ring can adopt two geometries. The envelope conformation with four atoms coplanar and the fifth atom outside the plan and the "half-chair" conformation where three atoms are coplanar and the remaining two are offset with one above and one below the plan. *(Scheme 5.1)*



Scheme 5.1: Cyclopentane envelope and half-chair conformations.

The representation of the pyrrolidine ring is similar to the cyclopentane conformations in Scheme 5.1. The histogram of the pyrrolidine ring dihedral angle C1-C2-C3-N1 shows that the NMR ensemble adopts two half-chair configurations in a 30:70 ratio with values around +/- 35 ˚ *(Figure 5.3A)*. Compare to the NMR data, the pyrrolidine ring torsion angle in the REMD trajectories displays a different behaviour *(Figure 5.3B)*. Indeed, instead of having two main conformations around +/-35 ˚, the dihedral angle is spread between -45 ˚ to 45 ˚ with two major populations at +/-30 ˚. Therefore, by sampling a broader range of torsion angle values the REMD trajectories capture the pyrrolidine ring conformation observed in GSK2194069 x-ray structure *(Figure 5.3A magenta arrow)* compared to the NMR data. REMD simulations enabled the sampling of more conformational space than NMR offering a better chance to capture the bioactive conformation.

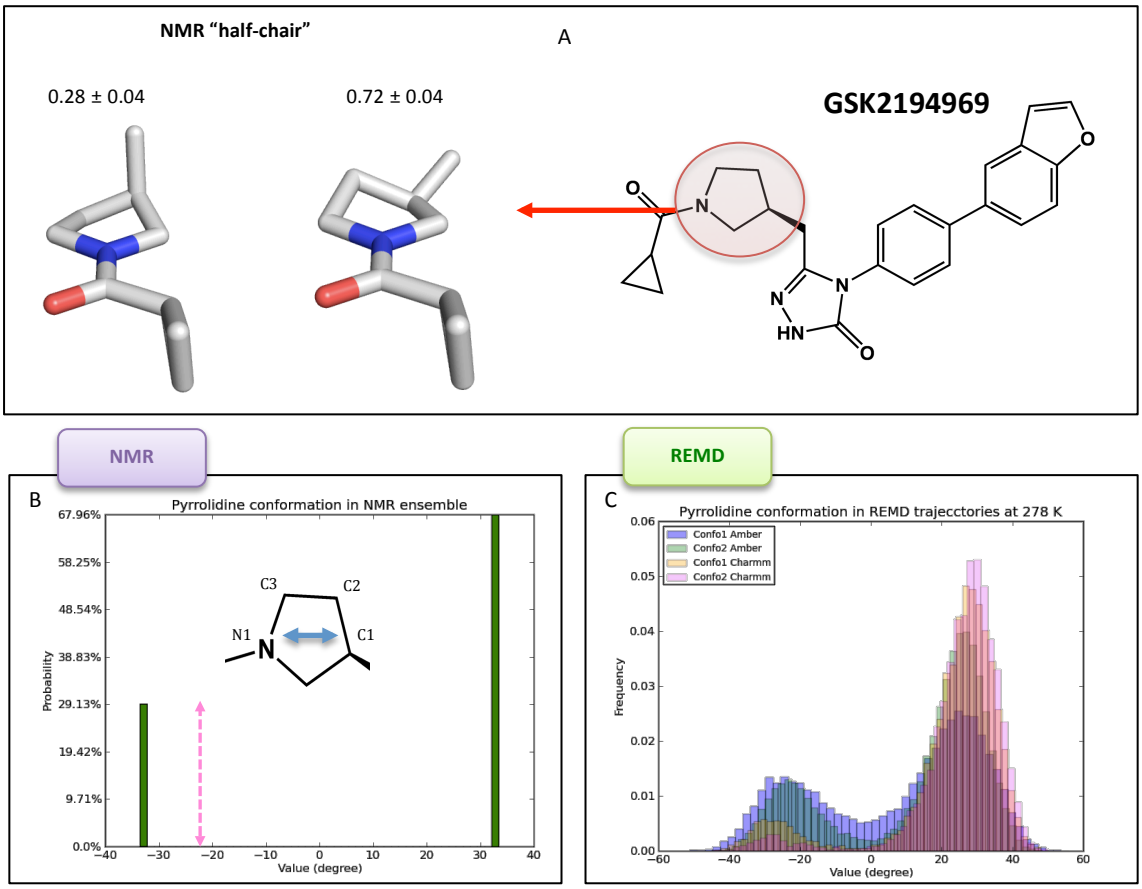


Figure 5.3: Distribution of the pyrrolidine ring dihedral angle in the NMR ensemble (B) and in the REMD trajectories of the conformations 1 and 2 of GSK2194969 in the Amber Gaff and Charmm CGenff force fields (C). The magenta arrow refers to the x-ray value of the pyrrolidine ring. The populations and the conformations of the NMR pyrrolidine ring are illustrated in (A).

5.2.1.4 Torsion angles τ3 and τ4:

According to the NMR data, the torsion angles τ3 and τ4 are mutually dependent. *Figure 5.4* shows the distributions of the torsion angles τ3 and τ4 in the entire NMR ensemble with the colours reflecting the link between them. The red values of τ3 will correspond to red values of τ4 and vice versa. From the NMR data when the torsion angle τ3 ranges around 0˚ to 130˚, the torsion angle τ4 ranges around 0˚ to 80˚ and around -100˚ to 0˚ whereas when τ3 ranges around +/-90 to +/-180˚, τ4 ranges around -100˚ to 0˚ and around 0 ˚ to 80 ˚ *(Figure 5.4)*.

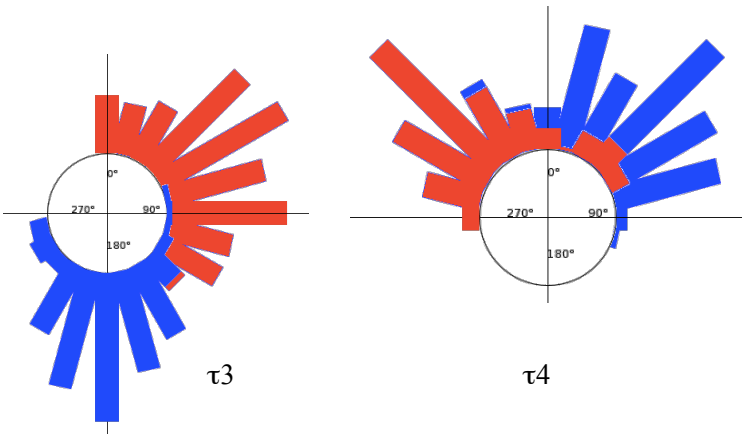


Figure 5.4: NMR ensemble torsion angles τ3 and τ4 distributions. The colours denoted the mutual dependency of these two torsion angles. Figure taken from the C4X report [77].

Equivalent results were found for the two REMD starting conformations so here only the data using conformation 1 as the starting point is presented. The distribution of the torsion angles τ3 and τ4 in NMR and REMD ensembles (*Figure 5.5)* shows consistency between the two by sampling similar torsion angles area. However, the distribution of torsion angle τ3 shows consistency between CSD, NMR and REMD, the different populations are in the same values range. Concerning the torsion angle τ4, the REMD and NMR ensembles show similar results with two main distributions ranging from 0˚ to 100˚ and -100˚ to 0˚ while the CSD displays sampling at +/- 150˚

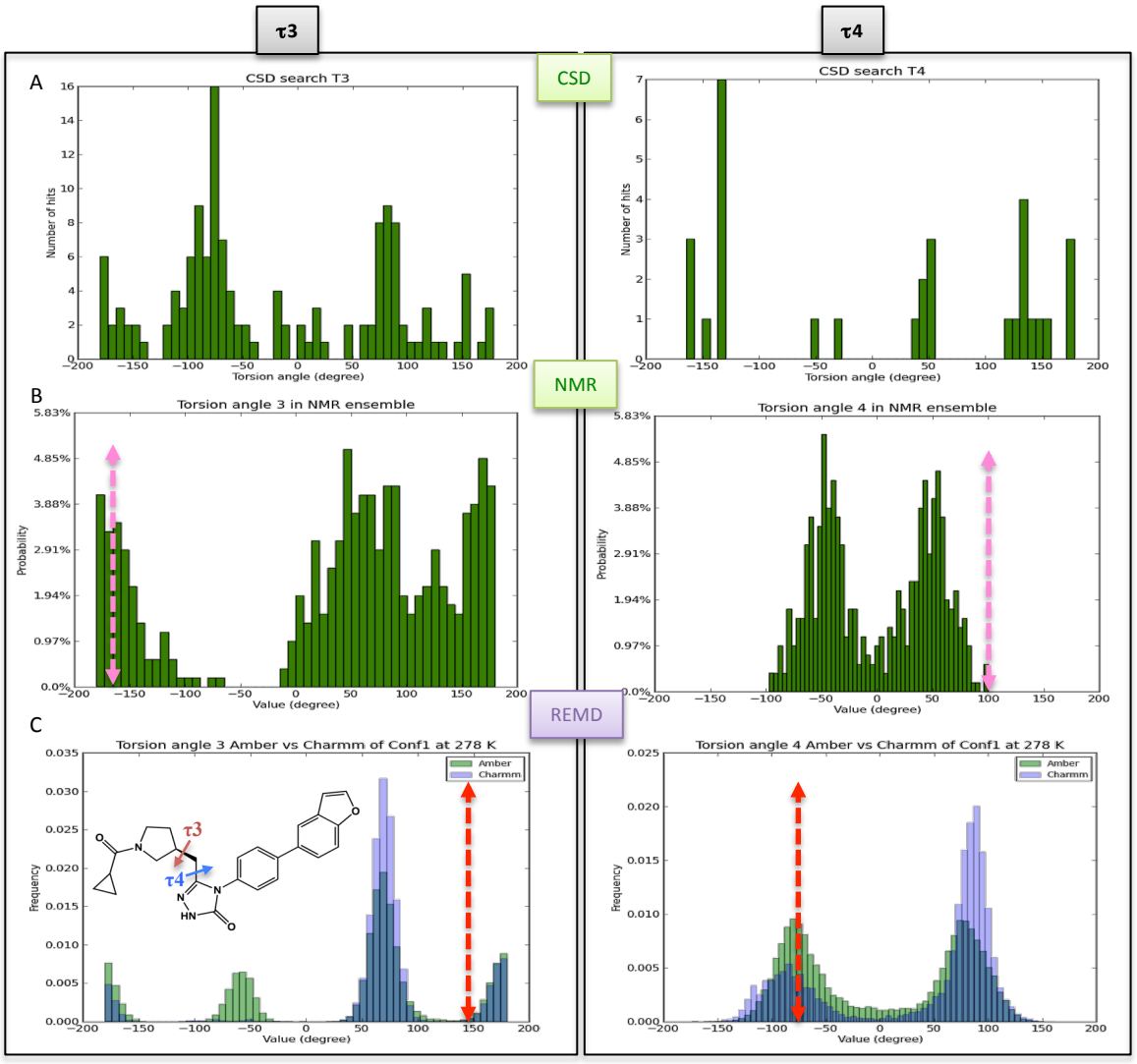


Figure 5.5: Distribution of the torsion angles τ3 and τ4 in the CSD database (A), the NMR ensemble (B) and in the REMD trajectories of the conformation 1 of GSK2194069 in the Amber Gaff and Charmm CGenff force fields in green and violet (C). The magenta and red arrows show the values of τ3 and τ4 in the x-ray and the initial conformations respectively.

In term of correlation between τ3 and τ4, a torsion angle distributions in *Figure 5.6* shows the mutual dependency of these two angles in the REMD simulations. As seen with the NMR ensemble in *Figure 5.4*, when τ3 varies between 50˚ to 100˚, τ4 varies between +/-50˚ to +/-130˚ and when τ3 varies between -50˚ to -100˚, τ4 varies between 50˚ to 90˚. Additionally, combining both results from *Figures 5.5 and 5.6*, a preferred conformations with τ3 values around 50˚ to 100˚ with an almost equal distribution of τ4 between +/-50˚ to +/-100˚ is observed in REMD simulations.

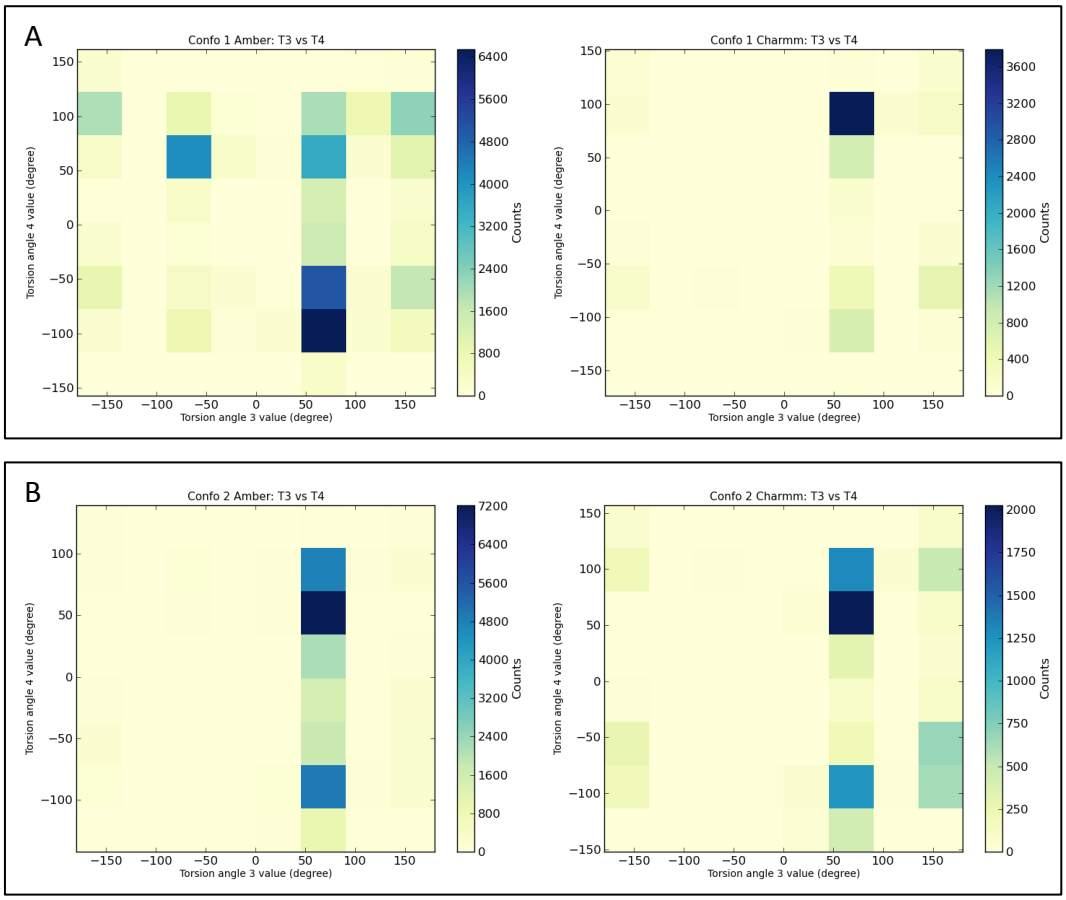


Figure 5.6: Torsion angle distribitions showing the mutual dependency of the torsion angles τ3 and τ4 in the REMD trajectories of GSK2194069. A: Conformation 1 in the Amber Gaff and Charmm CGenff force fields. B: Conformation 2 in the Amber Gaff and Charmm CGenff force fields.

Therefore, the distributions of the torsion angles τ3 and τ4 are consistent between the CSD data and the NMR and REMD ensembles. However, the REMD simulations have shown sampling differences between the force fields. The Amber Gaff force field was able to sample more the negative values in both torsion angles. Additionally, the correlation between the two torsion angles observed in the NMR data was also confirmed in the REMD simulations emphasizing consistency between the methods.

5.2.1.5 Torsion angles τ5 and τ6:

The torsion angles τ5 and τ6 show symmetrical behaviour according to the carbon atoms picked in the phenyl ring. *Figure 5.7* illustrates the two possibilities

due to the symmetry for each torsion angles. For torsion τ5, the carbon atoms 4 and 7 can be chosen creating two possible definitions 1-2-3-4 or 1-2-3-7. Similarly for torsion τ5, the carbon atoms 5 and 6 induce definitions 5-8-9-10 or 6-8-9-10.

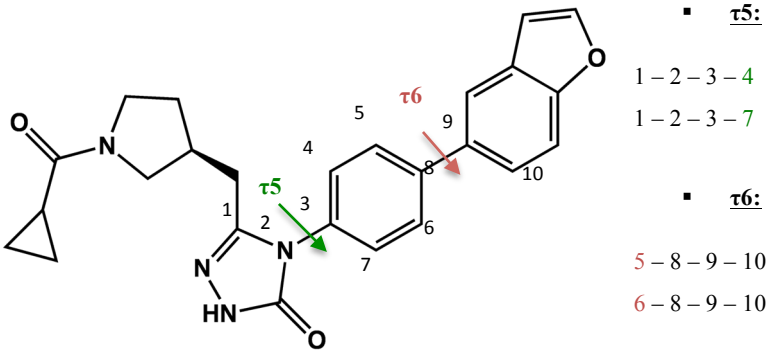


Figure 5.7: Torsion angle τ5 and τ6 symmetry in GSK2194069.

In the case of the experimental approach (NMR), the two symmetrical possibilities are not distinguishable and torsion angles τ5 and τ6 show similar behaviour. There are two equally populated orientations of the two torsion angles at +90° and -90° for the two symmetrical cases of τ5 and τ6 *(Figure 5.8B)*. The REMD trajectories starting from two different conformations shown similar results, therefore, only the conformation 2 data are presented here. Similarly to the NMR ensemble, the REMD trajectories show two main populations at +/-90° for torsion τ5. However, a difference is observed between the force fields. Indeed, while the Amber Gaff force field in green displays two main major torsions at +/-90°, the Charmm CGenFF force field in violet is split into two torsional angles around at +/- 90°.

In contrast to τ5, torsion angle τ6 shows different results between the two methods. Indeed, while the NMR data for τ6 ranges around +/-90°, the REMD trajectories sample those areas to a much lower extent, starting with two major regions between +/- 50° and +/- 150° which are also observed in the CSD analysis, thus NMR and CSD results differ in the distribution of τ6 *(Figure 5.8A)*.

Additionally, compared to the CSD *(Figure 5.8A)*, both REMD and NMR ensembles do not sample the area around 0° for the two torsion angles τ5 and τ6, results that are in consistent with the x-ray torsion angles values.

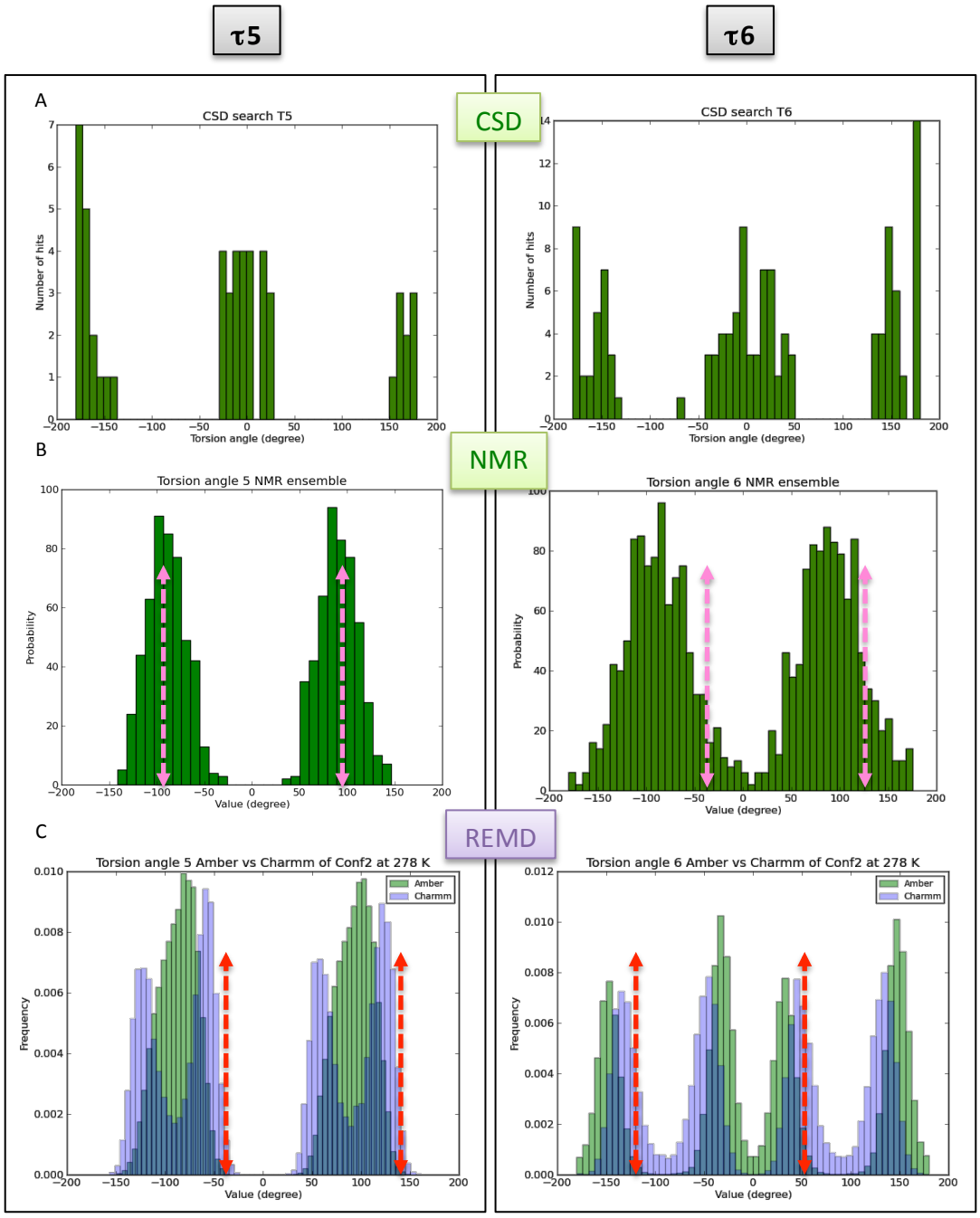


Figure 5.8: Distribution of the torsion angles τ5 and τ6 in the CSD database (A), the NMR ensemble (B) and the REMD trajectories of conformation 2 (C) of GSK2194069. The magenta and red arrows show the values of τ5 and τ6 in the x-ray structure and the initial conformation respectively for the two possible symmetries.

Therefore, the distributions of torsion angles τ5 and τ6 have shown better consistency between NMR data and REMD trajectories than with the CSD data. Indeed, the CSD analysis have shown sampling in the regions around 0° that was not observed in both NMR and REMD ensembles. Furthermore, the NMR and

REMD data were able to capture the GSK2194069 x-ray structure torsion angle values, which are not seen with the CSD. Additionally, different observations were seen within the force fields for torsion τ5 with a preference for the result for the Amber Gaff force field that shows better consistency with the NMR data.

Overall, the comparison between the NMR and REMD ensembles for GSK2194069 has shown good agreement between the methods and both of them were able to reproduce the x-ray structure torsion angles values. Differences in population were observed with the distribution of torsion τ6 between NMR and REMD. Additionally, differences were observed within the REMD simulation force field analysis. The Amber Gaff force field appeared to perform better than CGenFF Charmm force field by showing better consistency with the NMR data.

5.2.2 Compounds 2, 3 and 4:

Figure 5.9 shows the compounds 2, 3 and 4 torsion angles definition based on the NMR data.

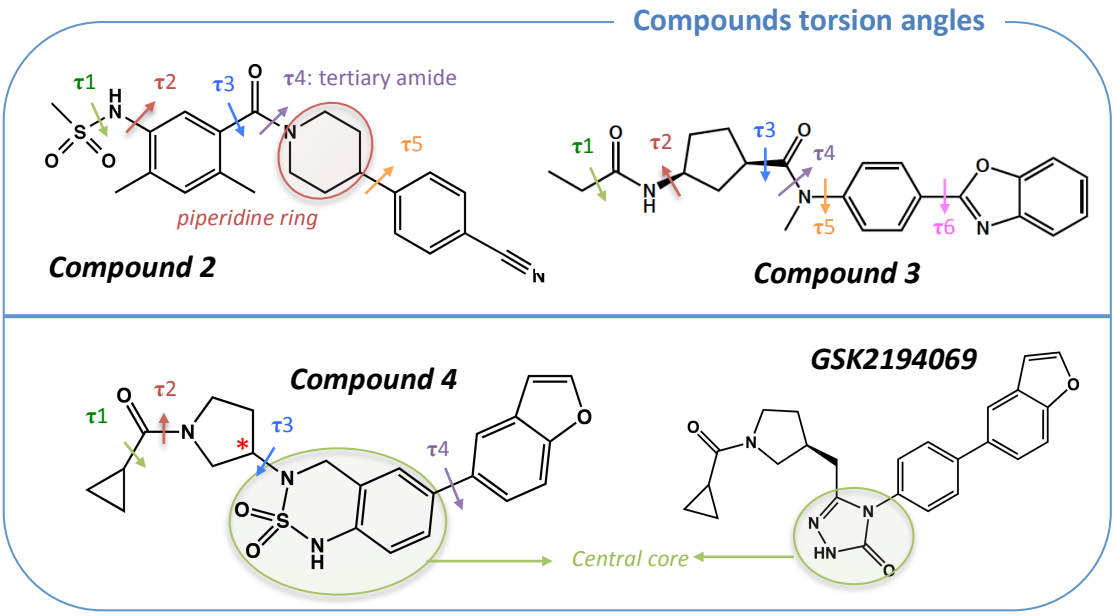


Figure 5.9: Torsion angles definitions of compounds 2, 3 and 4.

5.2.2.1 Compound 2:

5.2.2.1.1 Torsion angles τ1 and τ2:

Compound 2 can be described using five torsions including the sulphonamide τ1 and tertiary amide τ4 as well as a piperidine ring that can adopts a chair or boat conformation. The comparison between the NMR and REMD ensembles shows that the results of three torsion angles differ between the two methods corresponding to torsion angles τ1, τ2 and τ3.

In the NMR ensemble, τ1 displays three major conformations around +/- 90° and 0° *(Figure 5.10B)*. The REMD trajectories with the Charmm GCenFF force field in violet does not sample the region around 0° whilst using the Amber Gaff force field in green the sampling is improved. Compared to the CSD analysis in *Figure 5.10A*, the result with Charmm CGenFF force field are reasonably consistent. From the three different ensembles it is possible to say that Amber Gaff force field performs better than the Charmm CGenFF force field by sampling the 0° area similar as the NMR results.

The torsion angle τ2 displays consistent results as torsion τ1. Indeed, NMR and CSD torsion angle distributions are broadly consistent with each other while the REMD simulations starting with Amber Gaff force field (green) show higher sampling around 0° which was not seen by NMR and CSD. Thus in the case of torsion τ2 the CGenFF Charmm force field perform better.
Similar results were observed with the starting conformation 2 so only conformation 1 is presented here.

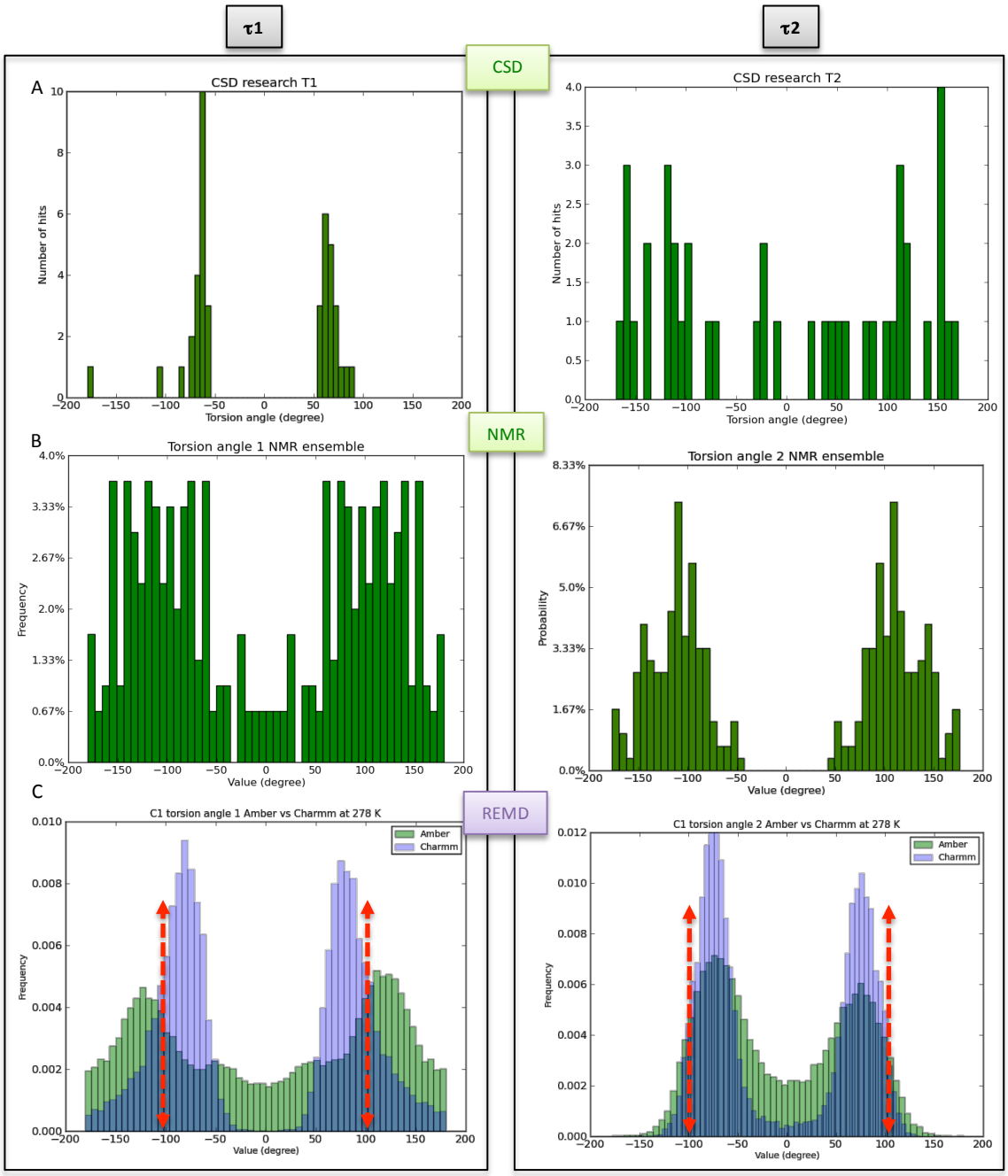


Figure 5.10: Distribution of the torsion angle τ1 and τ2 in the CSD (A), the NMR ensemble (B) and the REMD (C) trajectories for conformation 1 respectively with the Amber Gaff (green) and CGenFF Charmm force field (violet). The red arrows correspond to the values of torsion angles τ1 and τ2 in the starting conformations.

Therefore, *Figure 5.10* highlights that there is no one force field performing better than the other and it is most likely dependent on the specific molecule being simulated.

Additionally, the NMR data reported two symmetrically related conformations between the torsion angles τ1 and τ2. The conformations of the two bonds are mutually dependent as denoted by the colours in *Figure 5.11.*

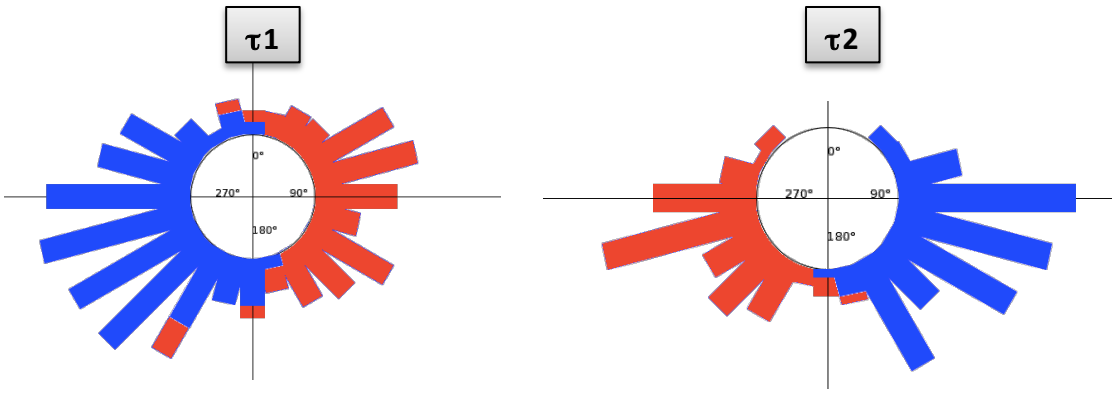


Figure 5.11: Torsion angles τ1 and τ2 distributions in the NMR ensemble. The colours denoted the mutual dependency of the two torsion angles. Figure taken from the C4X report [77].

From the NMR ensemble, τ1 is ranging from 0˚ to 180˚ and τ2 has a range of values from 0˚ to -180˚ and vice versa. To compare the mutual dependency of the two torsion angles between the REMD and the NMR ensembles, a correlation matrix was performed for each trajectory. Figure 5.12A and 5.12B illustrates the data for the conformations 1 and 2 in the Amber Gaff and CGenFF Charmm force fields respectively with torsion τ1 on the x axis and τ2 on the y axis. The colour bar determined the number of frames within the torsion angle values.

From the correlation matrices the symmetrical dependence observed in the NMR ensemble is not seen for the simulations with the CGenFF Charmm force field with τ1 ranges from 0˚ to +/- 180˚ when τ2 is ranging from 0˚ to +/-180˚. However, when the REMD trajectories are simulated with the Amber Gaff force field starting with conformation 1, some mutual dependency is observed with high population of τ1 values around 30˚ to 150˚ correlated with τ2 values ranges around -150˚ to -30˚ and vice versa, which is be broadly more consistent with the NMR data in term of torsion angle distribution and for the correlation observed between τ1 and τ2.

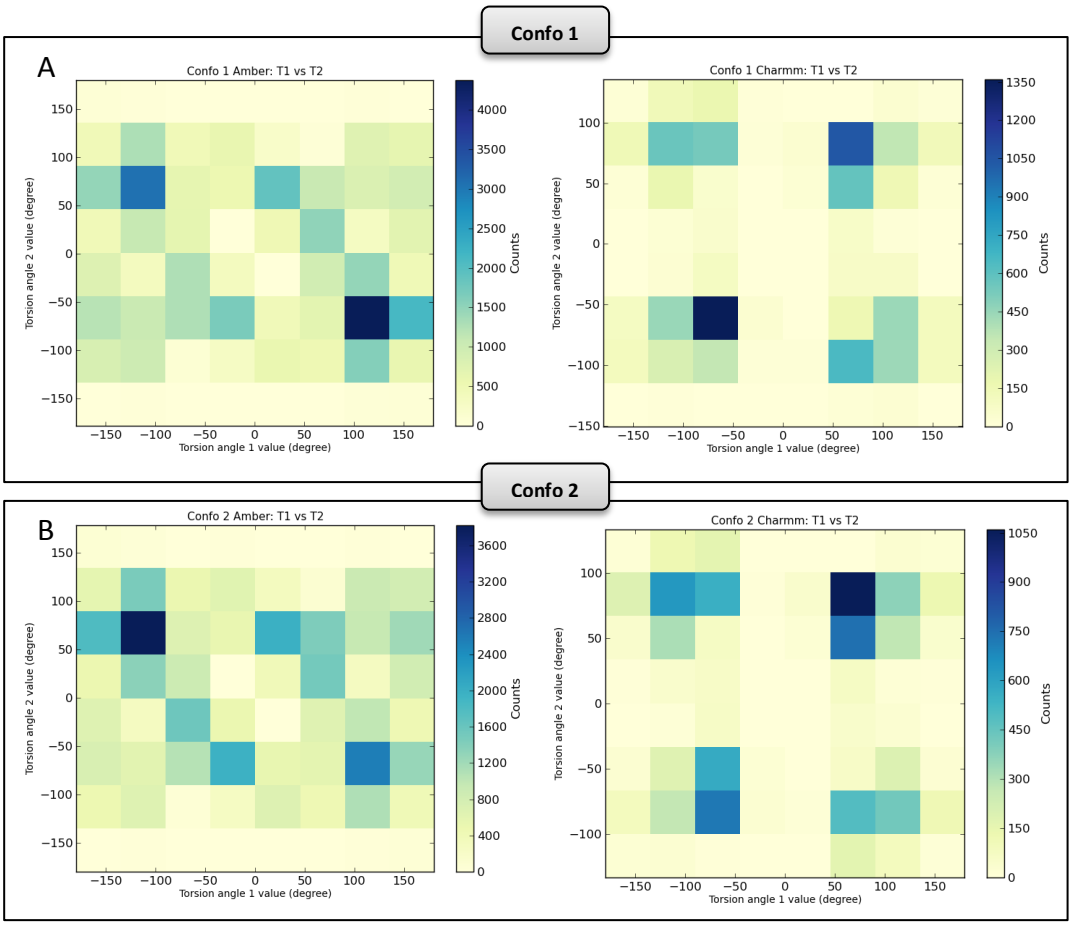


Figure 5.12: Correlation matrix showing the mutual dependency of the torsion angles τ1 and τ2 in the REMD trajectories of the compound 2. A: Conformation 1 in the Amber Gaff and Charmm CGenff force fields. B: Conformation 2 in the Amber Gaff and Charmm CGenff force fields.

Therefore, the distribution of the torsion angles τ1 and τ2 is broadly consistent between the NMR and the REMD ensembles. However, differences were observed within the force fields for the two torsions. The distribution of τ1 with the Amber Gaff force field appeared more consistent with the NMR data whilst the CGenFF Charmm force field distribution for τ1 is more consistent with the CSD data. Whereas, for the distribution of torsion τ2 the result with CGenFF Charmm force field is more consistent with both NMR and CSD by almost not sampling the area around 0˚. As for the distributions, the correlation between torsions τ1 and τ2 observed in the NMR data is observed only with the Amber Gaff force field simulation starting with conformation 1. Therefore, Amber Gaff appeared to perform better than CGenFF force field for this torsion angle.

5.2.2.1.2 Torsion angle τ3: atropisomers

The NMR data are in agreement with the CSD analysis with two populated torsions at +/- 100˚ *(Figure 5.13A and B)*. Additionally, from the NMR data, these two populations of torsion angles correspond to two slowly interconverting atropisomers in solution in a ratio 58:42. Atropisomers are stereoisomers resulting from hindered rotation about one or more single bonds, where the energy barrier to rotation is high enough to allow for the isolation of the conformers.

The REMD ensembles despite sampling similar regions as the NMR and CSD data show trapping while starting with the conformation 1, which was not overcome even at the highest temperature *(section 4.4.2.2)*. Starting with a value of 107.8˚, the distribution varies from 60˚ to 115˚ while starting at – 99.9˚ a beginning of sampling between 50˚ to 120˚ is observed in accordance with the NMR and CSD data *(Figure 5.13C)*.

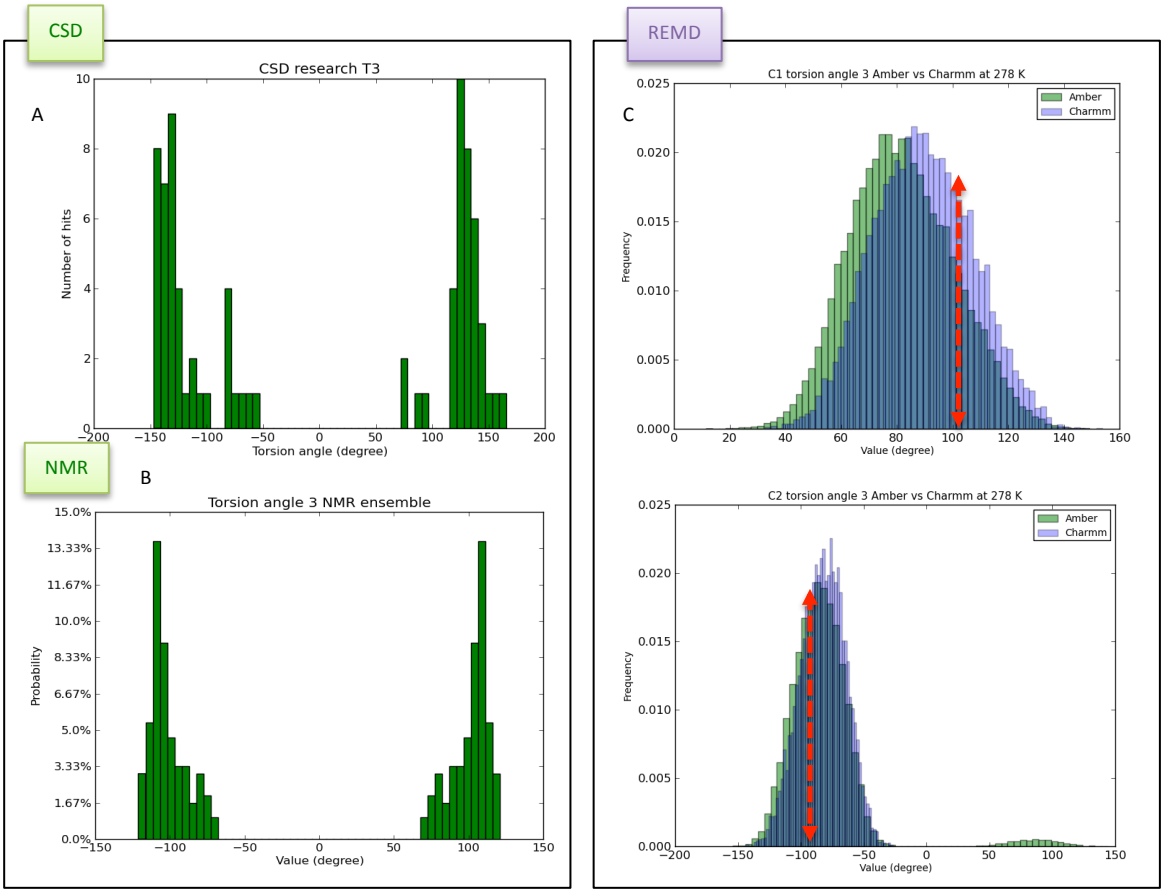


Figure 5.13: Distribution of the torsion angle τ3 in the CSD (A), the NMR ensemble (B) and REMD trajectories (C) for starting conformations 1 and 2 with the Amber Gaff (green) and CGenFF Charmm force field (violet). The red arrows show the starting conformation torsion angle values.

Therefore, the two conformations observed in the NMR data have show the challenge encountered with atropisomers. Atropisomers are observed in the REMD simulations with the conformations 1 and 2 resulting in a very slow sampling. A study in 2009 addressed the challenge of dealing with atropisomers in drug discovery [179]. They suggested several options to overcome in building analogues molecules with different features.

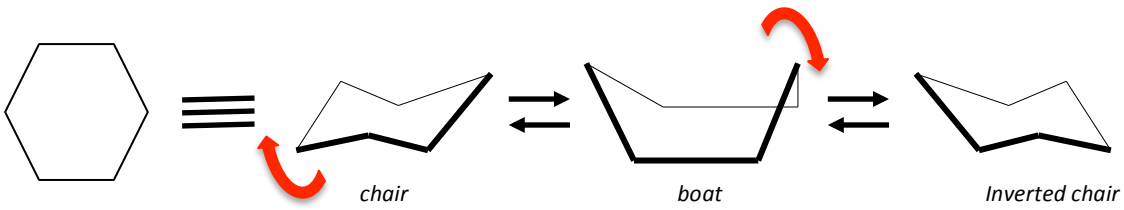
5.2.2.1.3 Torsion angle τ4 and τ5:

Torsion angles τ4 and τ5 show similar results between starting conformations, force fields and with the NMR ensemble. Indeed, the tertiary amide shows equal distributions of the E/Z configurations in the NMR ensemble, which is consistent with the CSD data. Owing to the two possible symmetries of the torsion τ4, the amide trapping observed for GSK2190469 is not seen with compound 2 by sampling both E and Z.

In the case of τ5, the two ensembles (NMR and REMD) demonstrated similar behaviour with equal distribution between two main populations around +/- 90° which are consistent with the CSD data. Additionally, the distribution of torsion τ5 of the REMD simulations with the CGenFF force field shows better consistency with the NMR ensemble by sampling the values around 0° *(Appendix B Figure B.1 and Figure 4.13).*

5.2.2.1.4 Piperidine ring:

The piperidine ring can adopt similar conformation as the cyclohexane ring. Two distinct conformations are observed for the six-membered heterocyclic piperidine ring named "chair" and "boat". Going from one conformer to another is induced by the swing of one of the ring branches as represented by red arrows in *Scheme 5.2.*



Scheme 5.2: Cyclohexane ring conformations.

Going from chair to inverted chair conformations can easily be performed by moving trough an intermediate boat conformation with the chair conformation more stable than the boat.

The piperidine ring in the NMR ensemble adopts a single chair conformation. For the REMD trajectories starting from boat and chair conformation both simulations converge to a single chair conformation, which is consistent with NMR data

5.2.2.2 Compound 3:

The distributions of the torsion angles for compound 3 demonstrate consistent results starting from two different conformations but some differences are observed between force fields and with the NMR ensemble.

5.2.2.2.1 Torsion angle τ1:

Two symmetrical conformations were reported in the NMR ensemble around +/- 90°. The REMD ensembles from the Charmm CGenFF and Amber Gaff force fields display sampling at +/- 90° regions which are consistent with the NMR data (*Appendix B Figure B.2 and Figure 4.14A).*

5.2.2.2.2 Torsion angle τ2:

Torsion angle τ2 is dependent on the cyclopentane ring pucker according to the NMR report. Indeed the conformation adopted is influenced by the steric interaction with the cyclopentane ring hydrogen atoms. The ring adopts two envelope shapes in a 60:40 ratio inducing the torsion angle τ2 to adopt two different angles at -90° and -150°. Similarly to what was performed for GSK2190469, the torsion angle of the ring was evaluated in both NMR and REMD ensembles *(Figure 5.14). Figure 5.14B* illustrated the REMD trajectories starting from conformations 1 and 2 with the two force fields. The red asterisk in the compound 3 structure shows the carbon atom involved in the definition of torsion τ2.

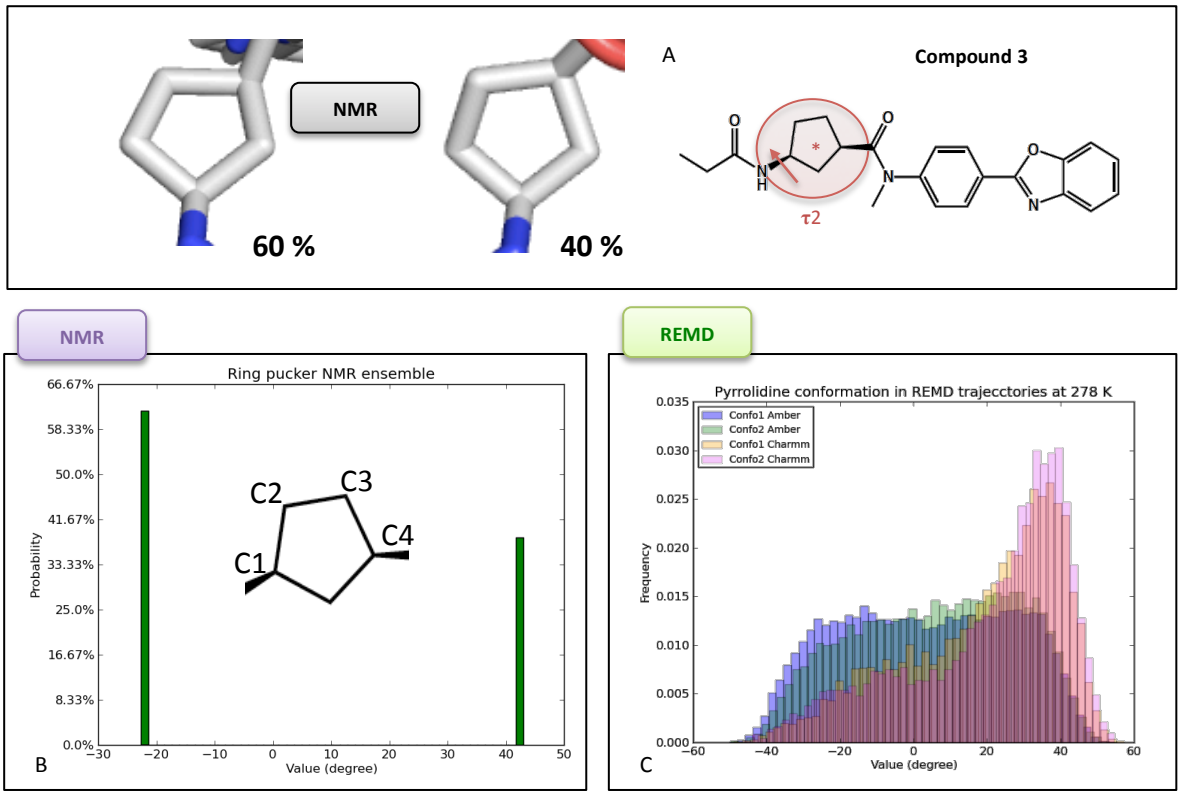


Figure 5.14: Distribution of the cyclopentane ring dihedral angle in the NMR (B) ensemble and in the REMD trajectories (C) of the conformations 1 and 2 of the compound 3 in the Amber Gaff and Charmm CGenff force fields. The populations and the conformations of the NMR cyclopentane ring are illustrated in (A).

From *Figure 5.14* the two conformations of the cyclopentane ring observed in the NMR data are at ~ -20˚ and ~ 40˚. However, the REMD simulations show a broad distribution of the ring pucker dihedral angle with no distinct separation between the two ring shapes as was observed with GSK2194069 for the pyrrolidine ring *(Figure 5.3).*

In term of torsion angle distributions, NMR and REMD ensembles display a related symmetry distributions with two mains conformations at -90˚ and -150˚ for the NMR and 90˚ and 150˚ for the REMD. A small degree of sampling emerges for REMD ensemble at - 50 ˚ REMD suggesting a trapping of the REMD simulations and a lack of sampling compared to the CSD data *(Figure 5.15)*. The CSD query used induces the symmetry for the torsion τ2. Indeed the red asterisk in *Figure 5.14A* indicates the torsion τ2 definition in the NMR and REMD ensembles, which can not be apply for the CSD query. Therefore, the CSD query search may be too

general and explain the lack of sampling in the negative values for the REMD ensemble and the positive values in the NMR data *(see Figure 4.3D)*. The CSD query is picking up conformations that are inaccessible to our compound. However, by showing opposite sampling between the REMD and NMR ensembles and the beginning of sampling at -50˚ observed for the REMD simulations shows that REMD simulations are trapped.

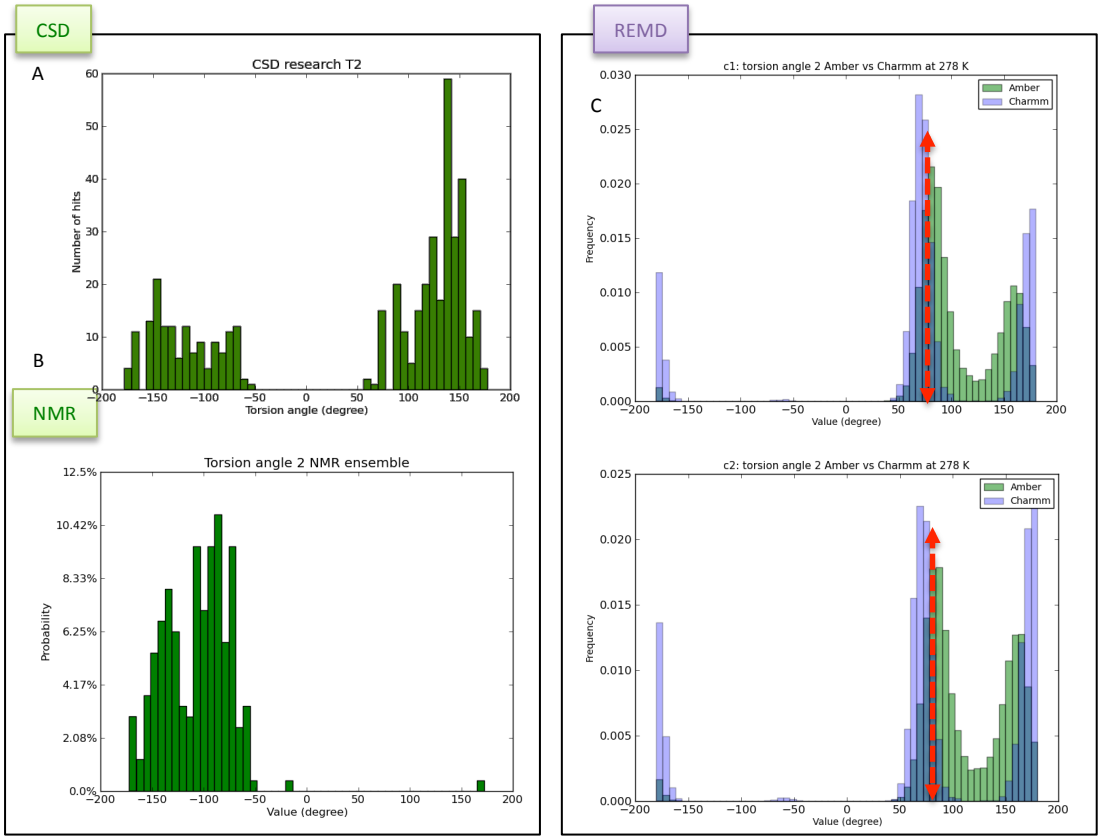


Figure 5.15: Distribution of the torsion angle τ2 in the CSD (A), the NMR ensemble (B) and REMD trajectories (C) for starting conformations 1 and 2 with the Amber Gaff (green) and CGenFF Charmm force field (violet). The red arrows show the starting conformation torsion angles values.

Therefore, from the NMR data it has been reported that the distribution of the torsion angle τ2 is dependent upon the cyclopentane ring with two main population at -80˚ and -140˚ of torsion τ2 *(see Figure 5.15B)* coupled with each ring states at ~ -20˚ and ~ 40˚ *(see Figure 5.14B).* The NMR report does not precise how the torsion τ2 populations are coupled with the ring pucker.

In the REMD ensemble two main populations of τ2 are also observed but at +60° and +180° *(see Figure 5.15C)* and the cyclopentane ring states are sampled as well *(see Figure 5.14C)*. Thus we are not getting agreement for torsion τ2 between NMR and REMD. The values of the torsion angle τ2 in the starting conformations could explain this difference by being trapped. However by sampling a broader torsion angle distribution of the cyclopentane ring in the REMD simulations, it means that the different distributions of τ2 between NMR and REMD is not due to the ring pucker.

5.2.2.2.3 Torsion angle τ3:

The NMR ensemble adopts one main conformation at 137° with some sampling at -180° that is consistent with the REMD simulations with however a split of the torsion angle distribution with two mains populations is observed at 90° and 150° *(Figure 16C)*.

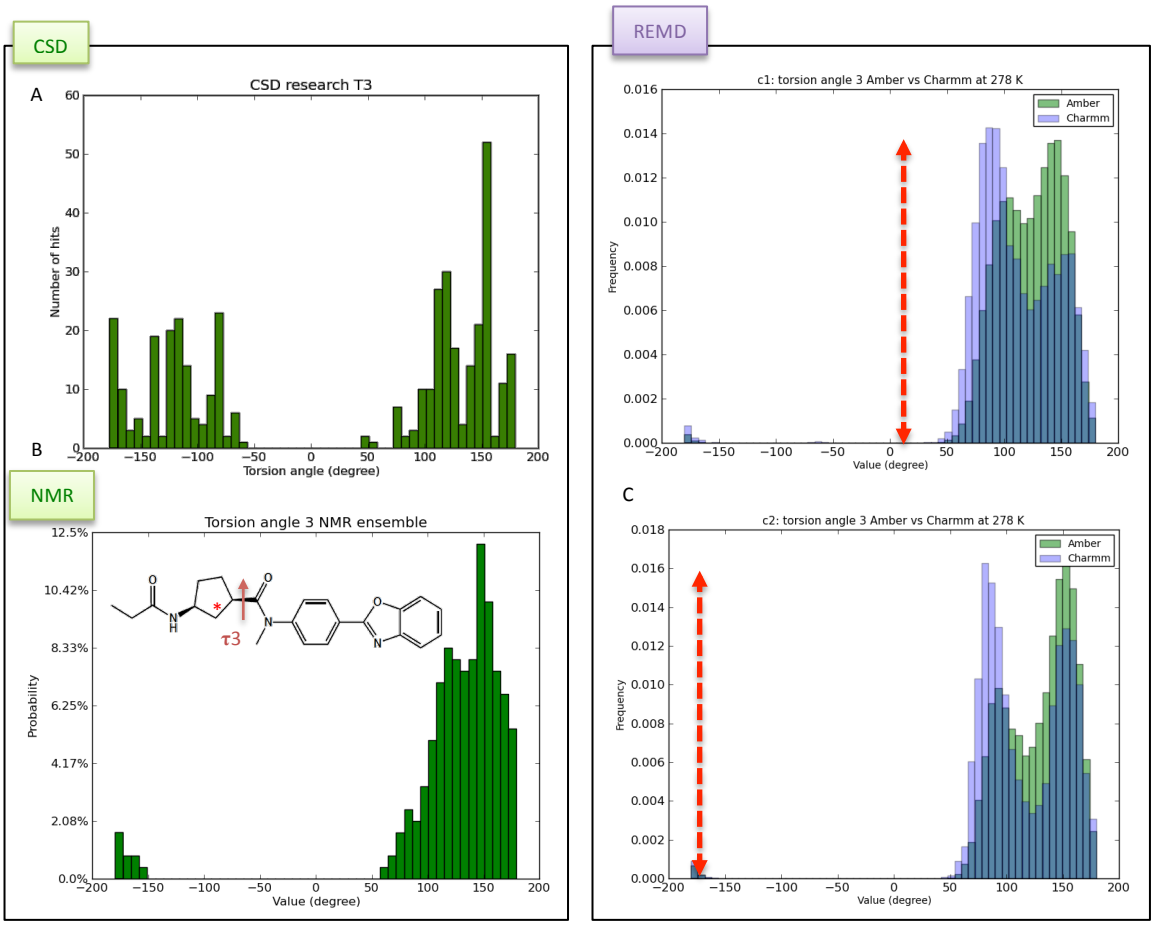


Figure 5.16: Distribution of the torsion angle τ3 in the CSD (A), the NMR ensemble (B) and REMD trajectories (C) for starting conformations 1 and 2 with the Amber Gaff (green) and CGenFF Charmm force

fields (violet). The red arrows show the starting conformation torsion angles values.

As for torsion τ2, the definition of torsion angle τ3 in the NMR and REMD ensembles is based on the atom in *Figure 5.16* red asterisk, which is not the case in the CSD search *(see Figure 4.3C)*. Therefore, the generality of CSD query could explain the lack of sampling in the negative values for the NMR and REMD ensembles. Additionally, the torsion angle τ3 values for the starting conformations highlight that the REMD simulations are not trapped but that the torsion τ3 favourable conformation is between +50˚ to +180˚ which is consistent with the NMR data.

5.2.2.2.4 Torsion angle τ4:

Torsion angle refers to an amide bond. GSK2194069 demonstrated trapping issues which such a group *(GSK2194069 torsion angle τ2 Figure 5.2).*

In the case of the distribution of compound 3 amide torsion τ4, similar behaviour as the GSK21949069 the amide torsion τ2 was observed (*Figure 5.2)*. The choice of the starting conformation influences the sampling in the REMD simulations. The NMR ensemble distribution of the amide bond of the compound 3, comparing to the two previous compounds, adopts two conformations Z/E in a ratio 82:18 *(Figure 5.17B)* which could explain the poor sampling of the E isomer of the REMD simulations when starting in Z isomer *(conformation 2 Figure 5.17C)*.

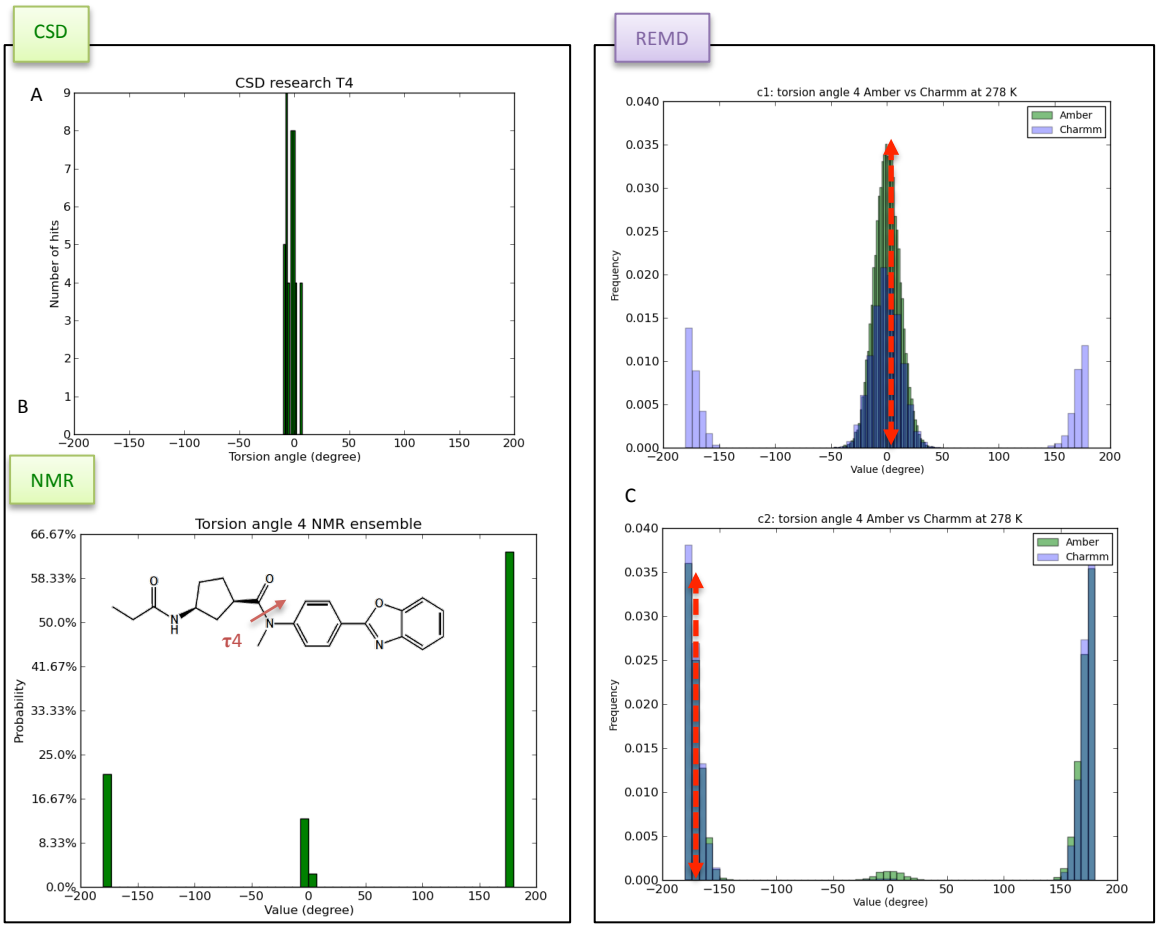


Figure 5.17: Distribution of the torsion angle τ4 in the CSD (A), the NMR ensemble (B) and REMD trajectories (C) for starting conformations 1 and 2 with the Amber Gaff (green) and CGenFF Charmm force field (violet). The red arrows show the starting conformation torsion angles values.

Therefore, there is consistency between the NMR and REMD ensembles, however the simulations of the starting conformation 1 suggest trapping by poorly sampling the Z isomer. Additionally, differences were observed between the two force fields. While, the CGenFF Charmm force field was able to sample the Z isomer when starting with the E isomer *(conformation 1 Figure 5.17C),* the opposite is observed for conformation 2 where a beginning of sampling of the E isomer is observed when starting with the Z isomer with the Amber Gaff force field *(conformation 2 Figure 5.17C)*. Thus, none of the force fields performed better than the other.

5.2.2.2.5 Torsion angle τ5 and τ6:

Torsion angles τ5 and τ6 are part of a rigid region of compound 3. In the NMR ensemble the phenyl group sits orthogonal to the amide and τ5 adopts two main angles at +/-90° which is consistent with the REMD trajectories as well as the CSD database *(Appendix B Figure B.3)*. Concerning τ6, two co-planar conformations are adopted with respect to the rest of the molecule leading to two main torsion angles at +/-180° and 0° in both REMD and NMR ensembles which are consistent with the CSD *(Appendix B Figure B.4)*.

5.2.2.3 Compound 4:

As described, compound 4 is a "scaffold-hop" based on GSK2194069, and therefore some torsions are shared between them *(see Figure 5.9* green circle). The carbon atom with the red star in *Figure 5.9* indicates the racemic character of this atom. To be consistent within the methods the S enantiomer is studied in both REMD and NMR ensembles.

5.2.2.3.1 Torsion angle τ1:

As for GSK2194069, the distribution of torsion angle τ1 is inconsistent between REMD and NMR ensembles *(Appendix B Figure B.5)*. The NMR ensemble is consistent with the CSD data results, which suggest that the force fields parameterization is not accurate for the case of torsion angle τ1.

5.2.2.3.2 Torsion angle τ2:

Torsion angle τ2 refers to a tertiary amide bond similar as in GSK2194069 *(Figure 5.2)*. Two species in slow interconversion are observed in the NMR ensemble with a ratio E:Z of 57:43. The ratio is not 50:50 due to the pyrrolidine envelope ring pucker inducing different local steric between the two isomers, which is consistent whit the CSD database result *(Figure 5.18A/B)*. However, the REMD simulations show signs of trapping issue, the Z isomer is not able to cross the energetic barrier, while starting from the conformer E the beginning of sampling of the Z conformer is seen *(Figure 5.18C)*.

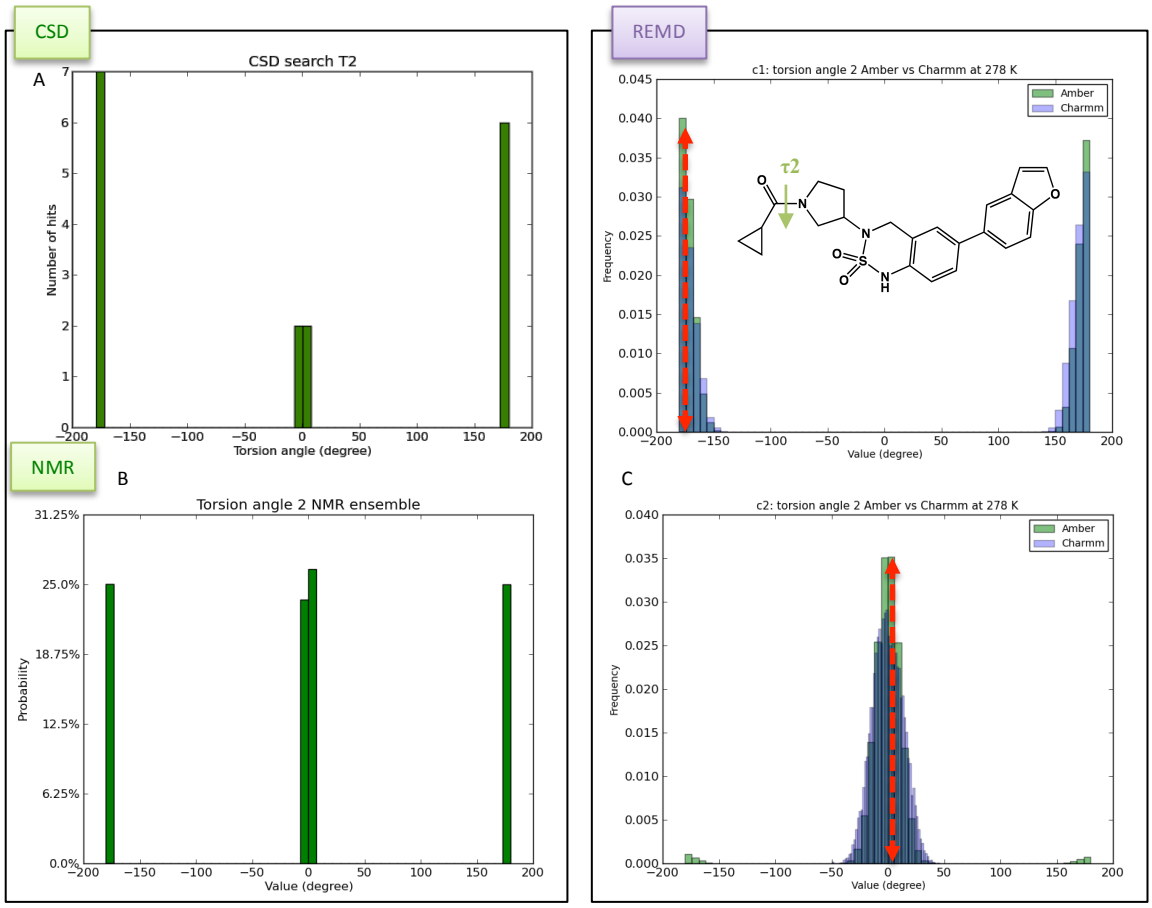


Figure 5.18: Distribution of the torsion angle τ2 in the CSD (A), the NMR ensemble (B) and REMD trajectories (C) for starting conformations 1 and 2 with the Amber Gaff (green) and CGenFF Charmm force field (violet). The red arrows show the starting conformation torsion angle values.

Therefore, compare to GSK2194069 when starting with the E isomer, compound 4 begins to overcome the trapping with the Amber Gaff force field. However, when starting with the Z isomer, none of the force fields were able to sample the E isomer while GSK2194069 has shown beginning of sampling with the Amber Gaff force field. Therefore, similar as what it was observed with GSK2194069, the Amber Gaff seems to perform better for the distribution of τ2 but neither is fully converged.

5.2.2.3.3 Torsion angle τ3 and ring pucker:

The system including torsion τ3, the sulphonamide nitrogen atom and the ring have linked behaviour according to the NMR report. From the NMR data, torsion

τ3 is trimodal for each nitrogen atom inversion with a single dominant conformation around -60˚. Thus, the nitrogen atom will invert its orientation with the change in torsion τ3 population (+/- 60˚ and 180˚). Furthermore, as the ring flexes and the nitrogen atom inverts, τ3 adopts a symmetric dihedral distribution with a main conformation at - 60˚.

From the CSD search the queries identified indicate the geometry of torsion τ3 similar to the REMD trajectories as well as in the NMR ensemble in term of the torsion angle ranges sampled *(Figure 5.19)*. However, the main conformation in the CSD data result appears at +150˚, this could be explain by the CSD query *(see Figure 4.3D)* where the nitrogen and sulphur atoms are not part of a ring system as in the compound 3 structure.

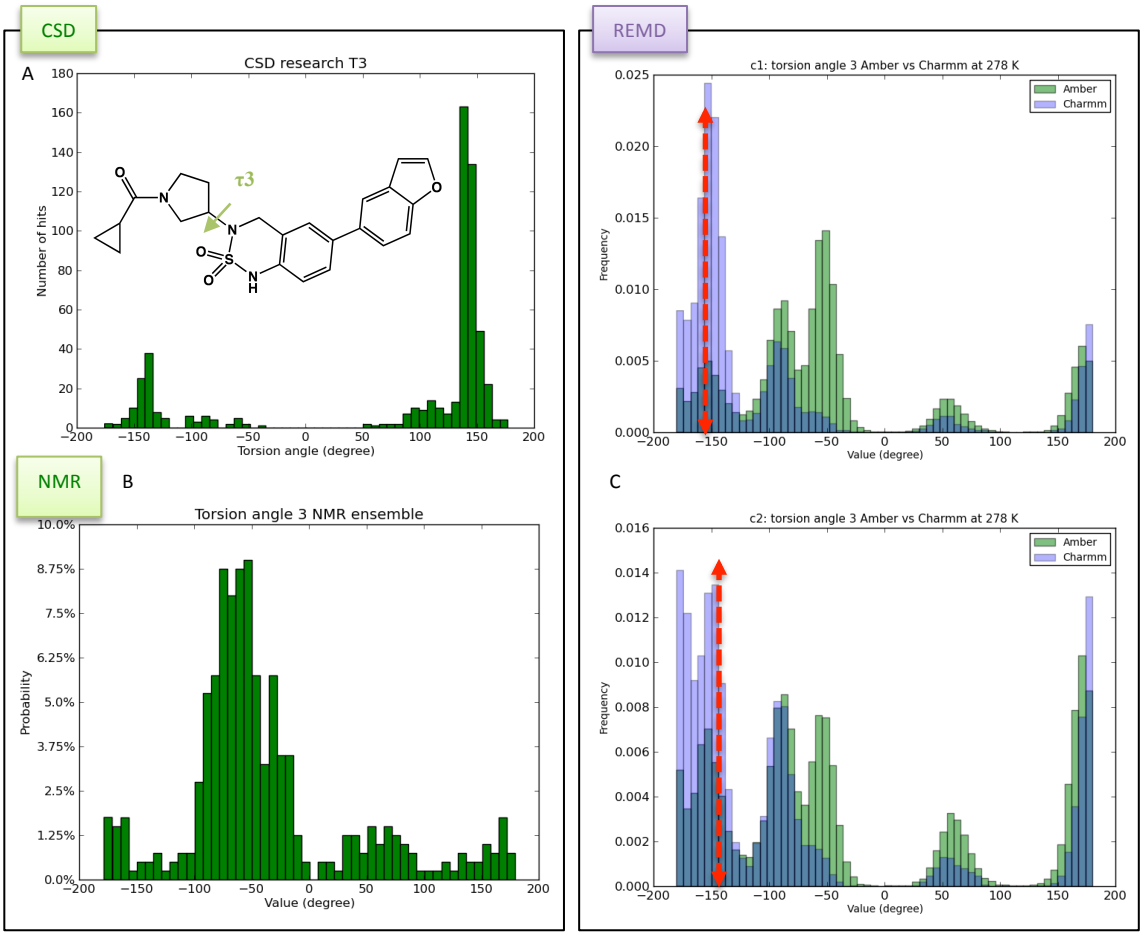


Figure 5.19: Distribution of the torsion angle τ3 in the CSD (A), the NMR (B) and REMD (C) ensembles starting form conformations 1 and 2 with the Amber Gaff and CGenff Charmm force fields. The red arrows show the values of τ3 in the starting conformations.

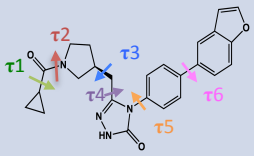
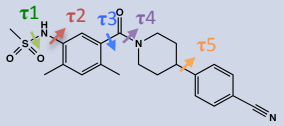
Therefore, the NMR reported three distributions of τ3 at +/- 60° and -180° torsion coupled with the inversion of the nitrogen atom with a dominant conformation at -60°. The population of the REMD simulations are consistent with the NMR data by sampling three main areas at +/-60° and 180° with same dominant conformation at -60°. The results of the CSD are less satisfactory in term of populations compared to NMR and REMD ensembles that could be explain by the generality of the CSD query.

5.2.2.3.4 Torsion angle τ4:

Both starting conformations in the REMD simulations show similar results in term of torsion angle distribution. However, the torsion angle distribution appears to be shifted between the two force fields similar to that observed with GSK2194069 and in the CSD data *(Figure 5.8C and Figure 4.9)*. Indeed, the distribution of torsion τ4 with the Amber Gaff force field is two more extremes values than with the CGenFF Charmm force field for the population -50° and 150° (right shift) and two less extremes values for the populations -150° and 50° (left shift). However, compare to the REMD trajectories, the NMR data shows two main conformations at +/- 180° and at 0° *(Appendix B Figure B.6)* which appear to be consistent with the CSD analysis of the equivalent torsion angle τ6 for GSK2194069 *(Figure 4.9)*.

5.2.2.4 Compounds torsion angle distributions summary:

The following *Table 5.1* summarises the differences observed between the CSD, the REMD simulations and the NMR ensemble in term of torsion angle distributions for the four compounds. The two starting conformations as well as the two force fields used (Amber Gaff and Charmm CGenFF) have shown differences all along the torsion angle distributions analyses and thus are presented individually in the table.

Compounds	CSD	REMD Gaff	REMD CGenFF	NMR
Compound 1 Conformation 1	**τ1**: 0 ° **τ2**: 0 °, +/-180 ° (E/Z) **τ3**: - 90 ° ↑ **τ4**: +/- 50 °, +/- 150 ° **τ5**: 0 °, +/- 180 ° **τ6**: 0 °, +/- 180 ° 	**. τ1**: +/- 90 ° **. τ2**: 0 °, +/-180 ° (E/Z) **. Pyrrolidine ring**: +/- 30 ° (broad) **τ3 and τ4 coupled** **. τ3**: +90 ° ↑ **. τ4**: +/- 60 ° **. τ5**: +/- 90 ° **. τ6**: +/- 40 ° , +/- 140 ° (Shifted)	**. τ1**: +/- 90 ° **. τ2**: +/- 180 ° (Trapped) **. Pyrrolidine ring**: +/- 30 ° (broad) **τ3 and τ4 coupled** **. τ3**: + 90 ° ↑ **. τ4**: +/- 60 ° **. τ5**: +/- 100 ° (split) **. τ6**: +/- 40 ° , +/- 140 ° (Shifted)	**. τ1**: 0 ° **. τ2**: 0 °, +/-180 ° (E/Z) **. Pyrrolidine ring**: +/- 30 **τ3 and τ4 coupled** **. τ3**: + 90 ° ↑ **. τ4**: +/- 50 ° **. τ5**: +/- 90 ° **. τ6**: +/- 90 °
Compound 1 Conformation 2		**. τ1**: +/- 90 ° **. τ2**: 0 ° (trapped) **. Pyrrolidine ring**: +/- 30 ° (broad) **τ3 and τ4 coupled** **. τ3**: +90 ° ↑ **. τ4**: +/- 60 ° **. τ5**: +/- 90 ° **. τ6**: +/- 40 ° , +/- 140 ° (Shifted)	**. τ1**: +/- 90 ° **. τ2**: 0° (trapped) **. Pyrrolidine ring**: +/- 30 ° (broad) **τ3 and τ4 coupled** **. τ3**: +90 ° ↑ **. τ4**: +/- 60 ° **. τ5**: +/- 90 ° (split) **. τ6**: +/- 40 ° , +/- 140 ° (Shifted)	
Compound 2 Conformation 1	**. τ1**: +/- 90 ° (no 0 °) **. τ2**: broad sampling **. τ3**: +/- 90 ° **. τ4**: 0 °; +/-180 ° (E/Z) **. τ5**: +/- 60 ° ; +/- 140 ° 	**τ1 and τ2 coupled** **. τ1**: 0° , +/- 90 ° **. τ2**: +/- 90 ° (sampling around 0°) **. τ3**: - 90 ° (trapped) **. τ4**: 0 °, +/-180 ° (E/Z) **. τ5**: +/-60 ° , +/- 140 ° (split)	**τ1 and τ2 not coupled** **. τ1**: +/- 90 ° (no 0 °) **. τ2**: +/- 90 ° (sampling around 0°) **. τ3**: - 90 ° (trapped) **. τ4**: 0 °, +/-180 ° (E/Z) **. τ5**: +/-60 ° , +/- 140 ° (split)	**τ1 and τ2 coupled** **. τ1**: 0° , +/- 90 ° **. τ2**: +/- 90 ° (no 0°) **. τ3**: +/- 90 ° **. τ4**: 0 °; +/-180 ° (E/Z) **. τ5**: +/- 90 °
Compound 2 Conformation 2		**τ1 and τ2 not coupled** **. τ1**: +0 °, /- 90 ° **. τ2**: +/- 90 ° (sampling around 0°) **. τ3**: + 90 ° ↑, - 90 ° (begin sampling) **. τ4**: 0 °, +/-180 ° (E/Z) **. τ5**: +/-60 ° , +/- 140 °	**τ1 and τ2 not coupled** **. τ1**: +/- 90 ° (no 0 °) **. τ2**: +/- 90 ° (sampling around 0°) **. τ3**: + 90 ° ↑, - 90 ° (begin sampling) **. τ4**: 0 °, +/-180 ° (E/Z) **. τ5**: +/-60 ° , +/- 140 °	

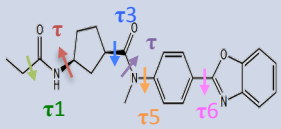
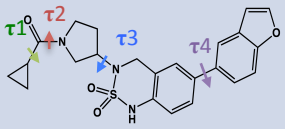
Compounds	CSD	REMD Gaff	REMD CGenFF	NMR
Compound 3 Conformation 1	**τ1**: +/- 90 ° ; +/ - 180 ° **τ2**: +/- 100 °; +/-150 ° **τ3**: +/- 100 °; +/-150 ° **τ4**: 0 ° (E) **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 ° 	**τ1**: +/- 90 ° ↑ **τ2**:[100 °; 150 °] ↑; [-100 °; -150 °] ↓ **τ3**: [100 °; 150 °] ↑ **τ4**: 0 ° (E) Trapping **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 °	**τ1**: +/- 90 ° ↑ **τ2**: [100 °; 150 °] ↑; [-100 °; -150 °] ↓ **τ3**: [100 °; 150 °] ↑; **τ4**: 0 ° ; +/-180 ° ↑ (E/Z) **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 °	**τ1**: +/- 90 ° ↑ **τ2**: - 90 °; - 150 ° ↑ **τ3**: 137 ° ↑ ; - 180 ° ↓ **τ4**: 0° E ; +/-180 ° Z ↑ **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 °
Compound 3 Conformation 2		**τ1**: +/- 90 ° ↑ **τ2**: [100 °; 150 °] ↑; [-100 °; -150 °] ↓ **τ3**: [100 °; 150 °] ↑ **τ4**: 0 ° ; +/-180 ° ↑ (E/Z) **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 °	**τ1**: +/- 90 ° ↑ **τ2**: [100 °; 150 °] ↑; [-100 °; -150 °] ↓ **τ3**: [100 °; 150 °] ↑ **τ4**: +/-180 ° (Z) Trapping **τ5**: +/- 90 ° **τ6**: 0 ° ; +/- 180 °	
Compound 4 Conformation 1	**τ1**: 0 ° **τ2**: 0 ; +/-180 ° (E/Z) **τ3**: +/- 100 ° ; +/- 150 ° (↑ 150°) **τ4**: +/- 60 °; +/- 150 ° 	**τ1**: +/- 90 ° **τ2**: +/-180 ° (Z) Trapping **τ3**: +/- 100 °; +/- 180 ° (↑ -150°) **τ4**: +/- 60 °; +/- 150 °	**τ1**: +/- 90 ° **τ2**: +/-180 ° (Z) Trapping **τ3**: +/- 100 °; +/- 180 ° (↑ -150°) **τ4**: +/- 60 °; +/- 150 ° (shifted)	**τ1**: 0 ° **τ2**: 0 ; +/-180 ° (E/Z) **τ3**: +/- 60 ° ° (↑ -60°); +/- 180 ° **τ4**: +/- 180 °; +/- 0 °
Compound 4 Conformation 2		**τ1**: +/- 90 ° **τ2**: 0° ↑ E ; +/-180 ° Z trapped **τ3**:+ /- 100 °; +/- 180 ° (↑ -150°) **τ4**: +/- 60 °; +/- 150 °	**τ1**: +/- 90 ° **τ2**: 0 (E) Trapping **τ3**: +/- 100 °; +/- 180 ° (↑ -150°) **τ4**: +/- 60 °; +/- 150 ° (shifted)	

Table 5.1: Torsion angles comparison between the CSD structures, the REMD simulations with the Amber Gaff and Charm CGenFF force fields and the NMR ensemble for GSK2194069 and compounds 2, 3 and 4. The arrows ↑ indicate a most populated distribution while the arrows ↓ indicate a less populated distribution.

5.2.2.5 Comparison between GSK2194069 and compound 4:

The amide torsion angle τ2 in GSK2194069 and compound 4 are equivalent in term of definition. From Chapter 4, the comparison of GSK2194069 and compound 4 has shown better sampling for compound 4 torsion τ2 *(see Figure 4.6 and Figure 4.19)*. The NMR data of both compounds appeared to be similar with almost equal population of the E/Z isomers confirming the trapping of torsion τ2 in the REMD simulations for both GSK2194069 and compound 4 *(see Figure 5.2 and Figure 5.18)*.

The torsion angles τ6 of GSK2194069 and the torsion τ4 of compound 4 appear challenging to sample in the REMD simulations in both structures. The distribution of the torsion has shown inconsistency between the methodologies *(Figure 5.9 and Appendix B Figure B.6)* but consistency was made within the

structure with different methods (compound 4 NMR (*Appendix B Figure B.6B*) is consistent with GSK2194069 CSD data *(Figure 5.9A)*; GSK2194069 REMD and NMR *(Figure 5.9B and C)* is consistent with compound 4 REMD and CSD (*Appendix B Figure B.5B and C*)).

GSK2194069 torsion angles τ3, τ4 and τ5 and compound 4 torsion angle τ4 correspond to the structure differences between the two compounds *(Figure 5.9 green circle)*. The NMR data of GSK2194069 has shown dependency within torsion angles τ3 and τ4 that was consistent with the REMD simulations *(Figure 5.6).* The compound 4 torsion τ3 was seen having a linked behaviour with the sulphonamide moiety in the NMR data with a dominant conformation at -60° *(Figure 5.19)* that was consistent with the REMD simulations with both presenting a dominant conformations at -60°.

Therefore, from the analysis of torsion angles distributions in the NMR, CSD and REMD data of GSK2194069 and compound 4 it is difficult to determine why the potency of compound 4 is low. Compound 4 has shown generally good agreement with the experimental conformational data and better sampling in the case of amide trapping *(Figure 5.19C)* by a beginning of sampling of the Z isomer when starting with the isomer which was not seen with GSK2194069. The analysis of the contacts made with the receptor will help us evaluate if the low potency can be explain by steric hindrance.

5.3 Principal component analysis:

5.3.1 Introduction:

Principal component analysis (PCA) is a statistical procedure that helps identify the principal fluctuations of a system in an MD trajectory. It summarizes the information among a set of observed variables into a linear combination of a smaller number of artificial variables called “principal components”. PCA can be a powerful tool to compare multiple MD trajectories with experimental structural ensemble [180-181]. In these studies Cartesian coordinates PCA was performed, using Gromacs tool *(see section 3.7.3)*, describing the positions of atoms. PCA analysis consists in the diagonalization of the covariance matrix of the atoms of the studied system to extract eigenvectors and eigenvalues. The eigenvalues will depict the motions of the particles in our systems of interest. The eigenvectors with the largest eigenvalues correspond to the principal components (PC) that account for the highest motions or variance in the dataset. Often, the relevant

motions of the system can be described by fewer than 10 principal components. Different methods can be employed to evaluate the adequate number of factors that will explain the most variance in our data. Here the method named "scree-plot" introduced by Cattell in 1966 was used [182]. The scree plot will help to determine how many components to retain that describe the most important global information. The scree plot consists of plotting the eigenvalues in descending order against the components numbers (or factors). The ideal shape of a scree plot shows a steep slope of eigenvalues followed by a bend ("elbow") in the curve and then a horizontal line. The idea is to detect the bend in the curve. The first components before the break are assumed to be the one that describe the main motions of the system and are retained for the PCA. Multiple breaks can be displayed in the plot. In this case we look at the last break before the eigenvalues start to decrease.

Prior to performing the PCA, the trajectories as well as the NMR ensemble must be superimposed onto a reference structure. The x-ray structure as well as the initial conformations was used as reference for GSK2194069 and the three other compounds respectively. The aim of this PCA analysis is to compare the ensembles obtained from the REMD trajectories with the NMR ensemble. Thus, for each compounds all the REMD trajectories (two starting conformations simulated with CGenFF and Gaff force fields) were concatenated with the NMR ensemble into a "super trajectory". Each REMD simulations as well as the NMR ensemble were projected onto the eigenvectors extracted from the "super trajectory".

5.4 Principal Component Analysis results:

5.4.1.1 Scree plot:

Figure 5.20 shows the scree plot of the PCA of GSK2194069 as well as compounds 2, 3 and 4. It shows the eigenvalues associated with the first 30 components. A large break between components 1 and 3 for all the compounds is observed. The "elbow" in red arrows displays the threshold chosen for the retention of the number of factors that maximised the accounted variance. Looking at the PCA maximized variance for the first two factors, a percentage of 73.25 %, 67.59 %, 74.96 % and 70.91 % of the overall multivariate variability for each compound respectively is obtained. Thus, by describing more than 70 % of the total variance, these results suggest that the first two components appeared

to be meaningful and represent the major dynamics for these compounds. They contain the essential information of the dynamics of our systems and will be retained for the PCA analysis. The third factor could have been included in the accounted variance but it only explains 10% of the variance in average so to make the visualisation clearer and to simplify the analysis only one and two have been retained.

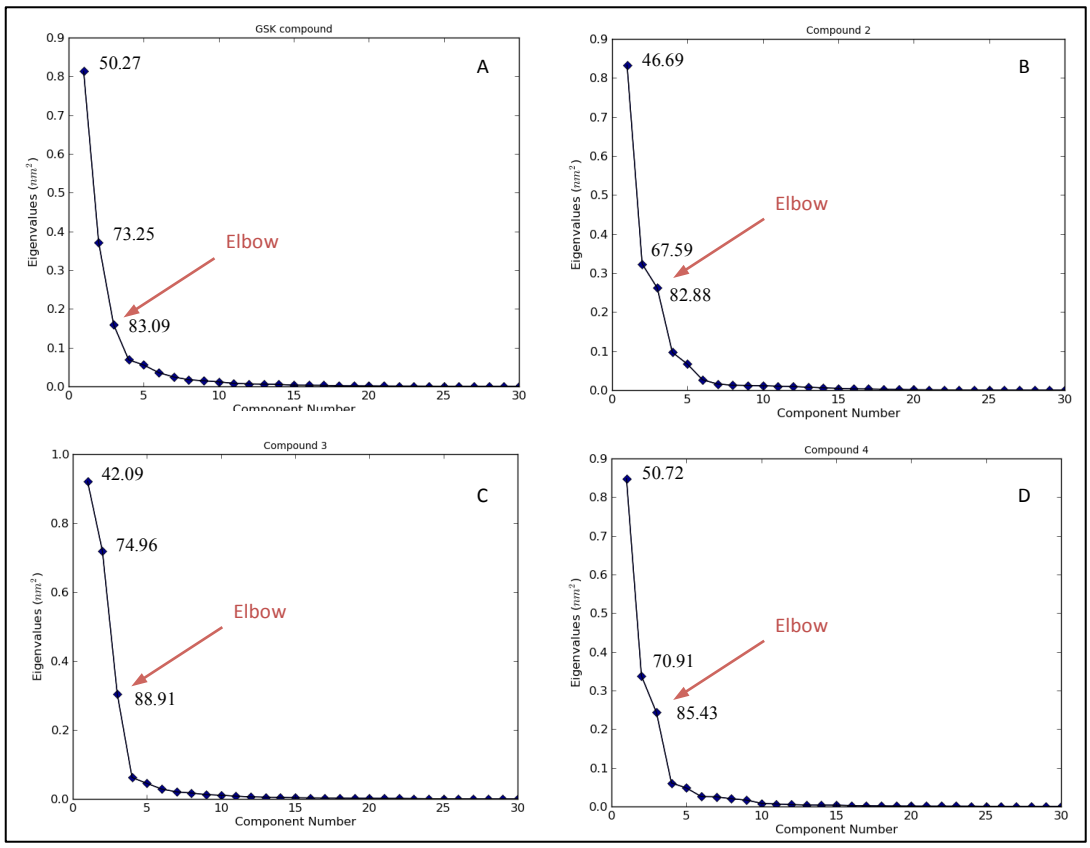


Figure 5.20: Scree plot of the principal components (X-axis) and their contribution to variance (Y-axis) of GSK2194069 (A), compounds 2 (B), 3 (C) and 4 (D) for the first 30 components. The red arrows represent the break point.

5.4.1.2 PCA simulations:

Principal component analysis was used to identify if the REMD simulations capture the NMR ensemble motions. All the REMD trajectories for each compound were analysed simultaneously. *Figure 5.21* shows the projection of the first two components of the REMD and NMR ensembles of the four compounds onto the "super trajectory". The colours correspond to the starting conformation and the force fields employed in the REMD simulations (Amber Gaff force field is in light blue and magenta and the Charmm CGenff force field in green and orange for the

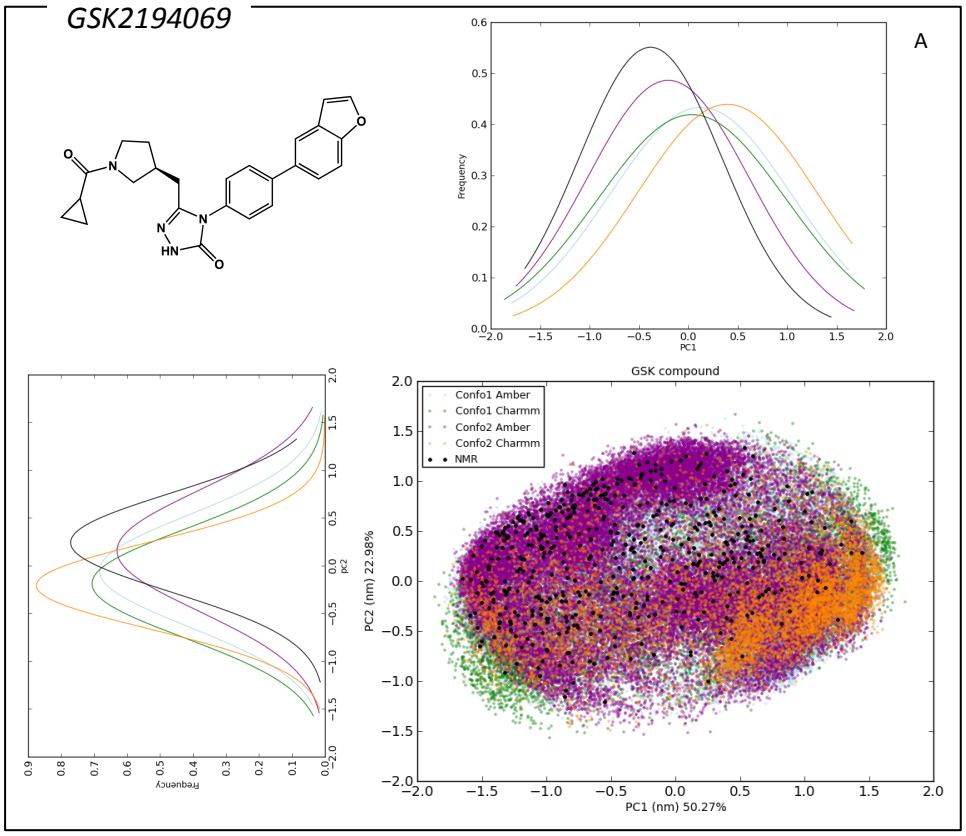
conformations 1 and 2 respectively). The NMR data are shown in black dots. To better investigate the overlap between the REMD and NMR subspaces, histograms of each ensemble projections of the PC1 and PC2 have been plotted for each compound.

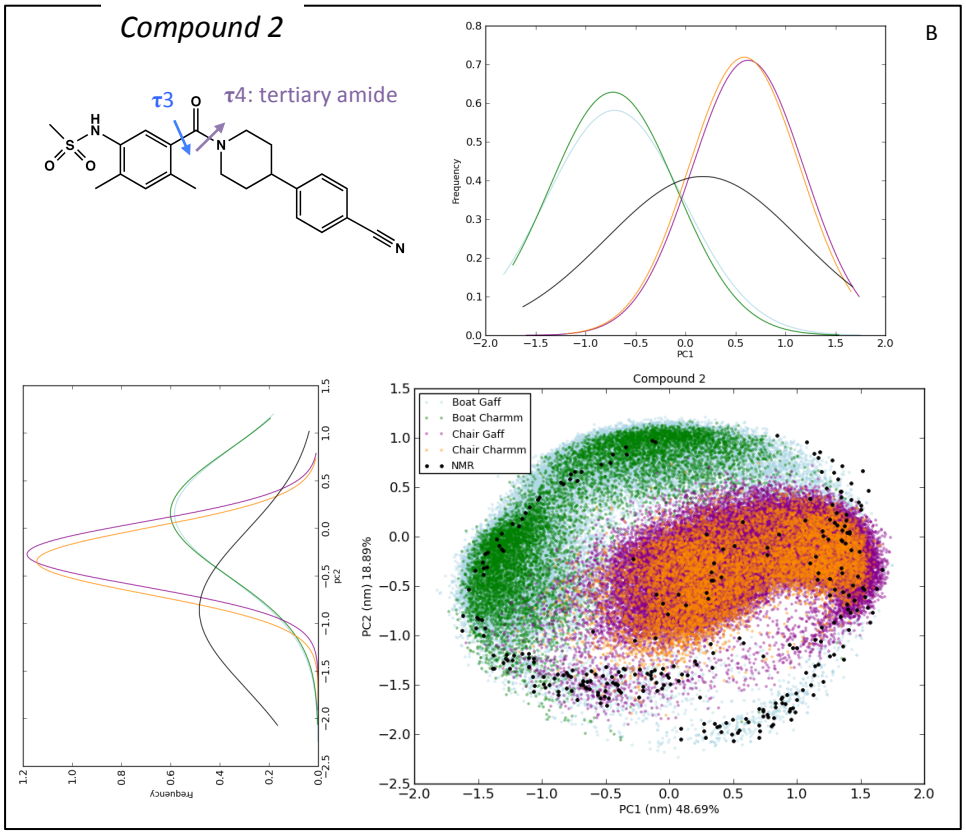
In term of comparison between the experimental and computational data, these analyses reveal that the REMD trajectories cover nearly the entire regions sampled by the NMR ensemble of the four compounds. Additionally, comparing the two methodologies subspaces, the REMD ensemble appears to sample larger regions of conformational space than the NMR data, especially for the compounds 2 and 3 where the NMR data are forming a halo on the top of the REMD projections and show regions poorly populated by the NMR. Thus, the REMD data find all the NMR conformations and sample regions not found by the NMR data.

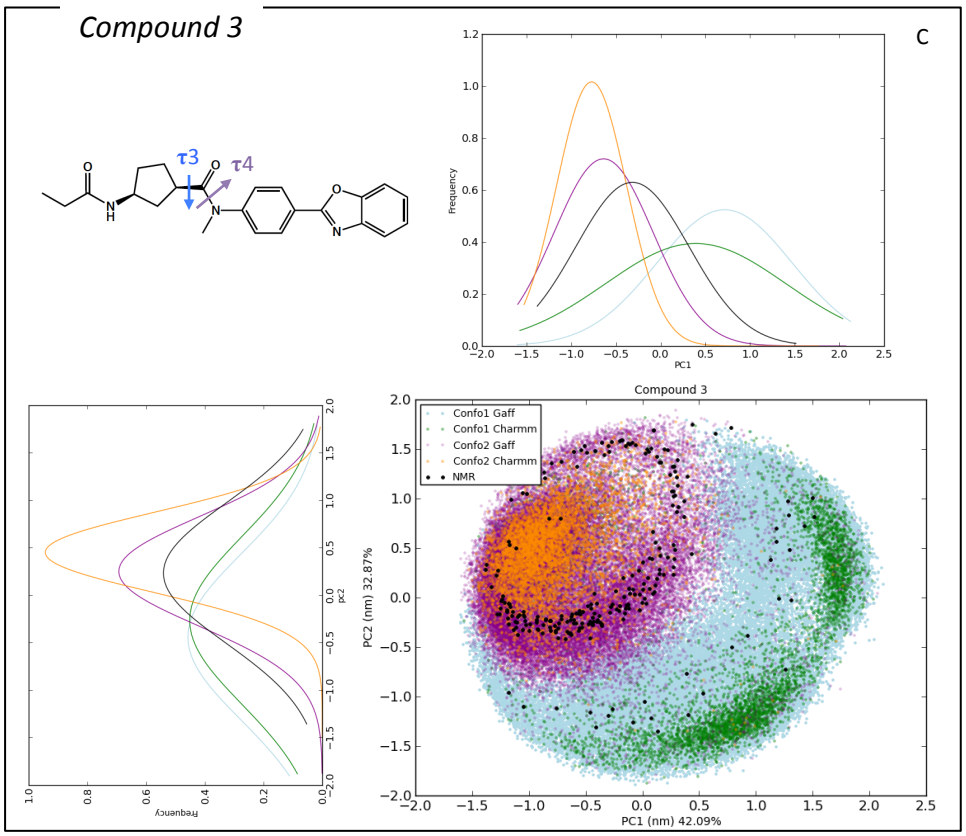
Looking at the REMD trajectories individually, two key observations can be made. GSK2194069 and compound 4, starting from two different conformations and two different force fields, the different simulations cover similar regions of the phase space. The histograms of the PC1 and PC2 emphasize the good overlap between the trajectories with each colours corresponding to the same colour in the PCA projections.

In contrast, compounds 2 and 3 show different behaviour according to the starting conformation. Indeed, from examination of the PCA analyses, two conformational subsets in light blue, green, magenta and orange corresponding to the starting structures 1 and 2 with the Amber Gaff and CGenFF Charmm force fields respectively are observed. To have a better idea of how well or poorly the different trajectories overlay together, the PC1 and PC2 projection histograms have been plotted. For both, compounds 2 and 3, the overlap between the trajectories is reasonable for PC2 and shows two distinct populations for PC1.

Thus, from compounds 2 and 3 PCA results, starting with different conformations, different regions of the subspace are sampled while the trajectories of GSK2194069 and compounds 4 populate similar basins.


GSK2194069
A
Frequency
PC1
GSK compound
Confo1 Amber
Confo1 Charmm
Confo2 Amber
Confo2 Charmm
NMR
PC2 (nm) 22.98%
PC1 (nm) 50.27%
pc2
Frequency


Compound 2
B
τ3
τ4: tertiary amide
Frequency
PC1
Compound 2
Boat Gaff
Boat Charmm
Chair Gaff
Chair Charmm
NMR
PC2 (nm) 18.89%
PC1 (nm) 48.69%
pc2
Frequency



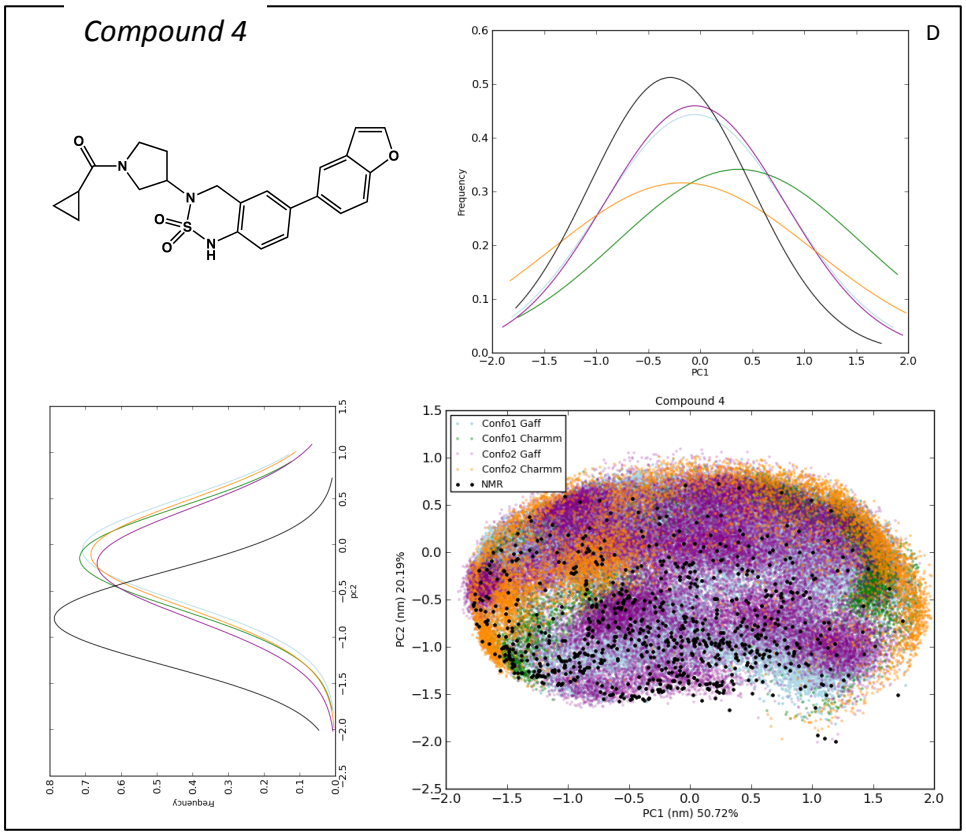


Figure 5.21: Distribution of the REMD trajectories of GSK2194069 (A) and compounds 2, 3 and 4 (B, C and D) and the NMR ensemble along the two first components identified by PCA. The colours refer to

the different conformations and force fields used. The NMR ensemble is represented by black dots. REMD using the Amber Gaff force field is in light blue and magenta and the Charmm CGenFF force field in green and orange for the conformations 1 and 2 respectively. The histogram of the ensembles projections on the corresponding PCs for each compound is displayed next to the PCs projection.

To investigate the reason for the separation between the two starting conformations observed in compounds 2 and 3 PCAs, the correlation between the torsion angles values and the PCs has been plotted. The previous analysis in section 5.2.2.1.2 and 5.2.2.2.4 has shown the trapping of the REMD trajectories for torsion angle τ3 of compound 2 and torsion τ4 of compound 3 *(Figures 5.13 and 5.17)*.

Compound 2 torsion angle τ3 sampled the region around 90° in the conformation 1 simulations while starting with conformation 2, τ3 was mainly sampled around -90° with the beginning of sampling at 90°. The projection of compound 2 torsion angle 3 onto PC1 in the two starting conformations emphasized that the trapping of τ3 in both trajectories explain the split in the PCA plot with conformation 1 τ3 predominantly projected around PC1 values of 0 to 0.5 *(Figure 5.22* red circle) while τ3 in the conformation 2 is projected between -1.5 and 0 *(Figure 5.22* green circle). The hot spot at the top of the PCA projection in black circle refers to the beginning of sampling of τ3 at 90° from conformation 2.

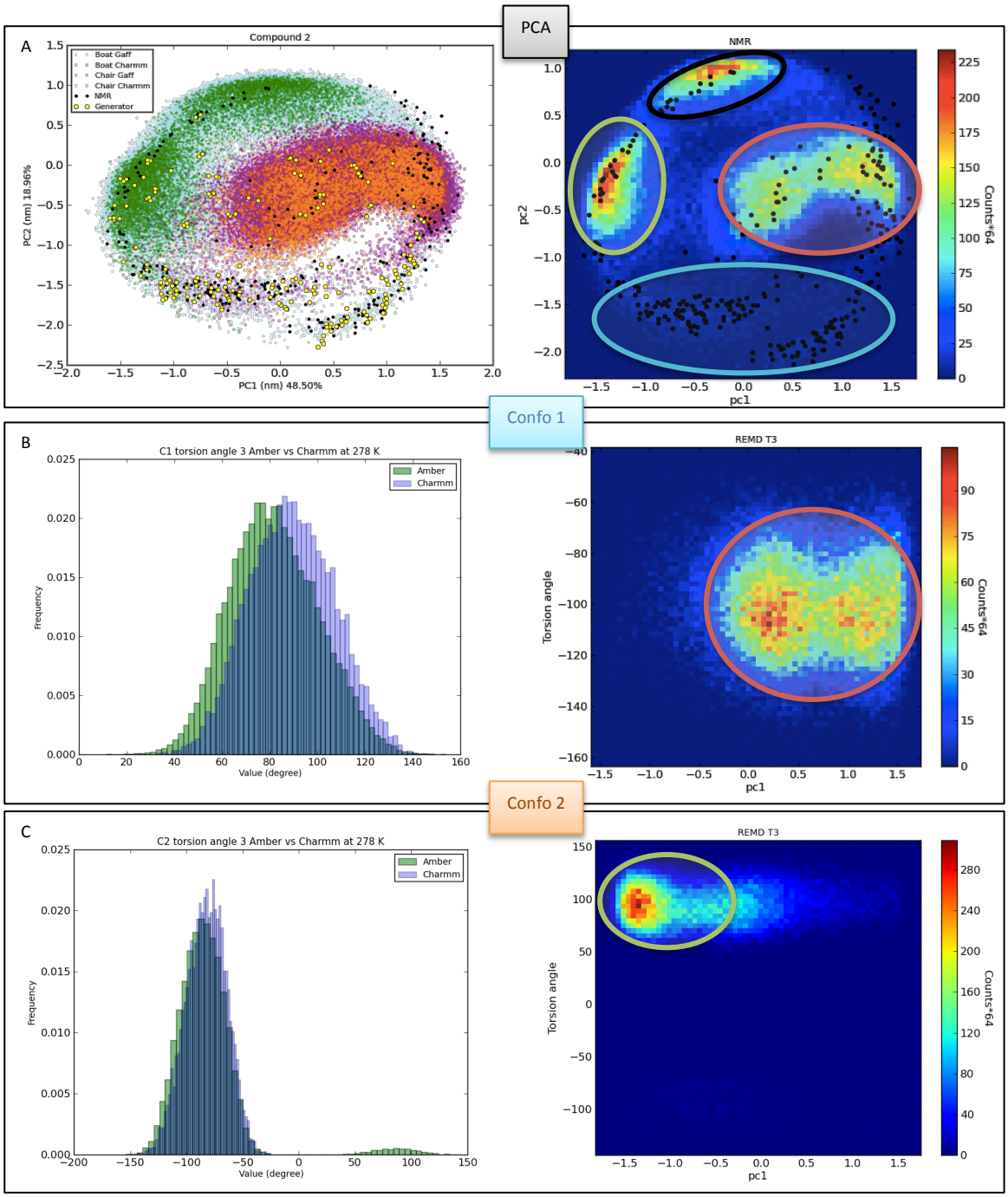


Figure 5.22: Projection of compound 2 torsion angle τ3 onto PC1 in the two starting conformations. A: PCA projection of the REMD and NMR ensembles along the two first components. The corresponding heatmap of the PCA is represented next to the conventional PCA projection representation. B: correlation plot of the torsion angle τ3 with the first PC from the REMD simulations of conformation 1. B: correlation plot of the torsion angle τ3 with the first PC from the REMD simulations of conformation 2.

Similar results were observed for compound 3. The trapping of compound 3 torsion angle τ4 *(Figure 5.17)* in the REMD simulations correlates with the split observed in the PCA projection.

However, the NMR ensemble was able to capture both torsion angle distribution but despite the fact that the NMR overlaps with the REMD simulations, it also tends to be localised in certain areas of the PCA *(Figure 5.22 cyan circle)* which tells us that the trapping observed with the torsion angle τ3 of compound 2 and τ4 of compound 3 causes the NMR data to be projected in different PCs values. PCA is a dimensionality reduction method so the fact that the NMR is seen in places that are less populated by the REMD simulations *(Figure 5.22 cyan circle)* could be explain by the poor sampling of these torsion angles.

5.4.1.3 GSK2194069 versus compound 4:

Figure 5.21A and 5.21D shows similar PCA projections from both compounds. The REMD trajectories, comparing to compounds 2 and 3, have shown good overlap between each other. However, a closer look at the plots, the NMR ensemble in black dots and in black line in the histograms for GSK2194069 shows good overlap for both PCs with the REMD simulations especially with conformation 2 with the Amber Gaff force field. Compound 4, on the contrary, shows poor overlap with PC2 *(Figure 5.21D histogram PC2)*. The difficulty of capturing similar dominant conformations between NMR and REMD could be related with the difficulty of sampling similar torsion angle distributions for torsion τ4.

Overall, the REMD simulations for all the compounds provide a reasonably accurate description of the NMR ensemble and sample larger regions of the conformational space. However, the hot spots observed in the REMD highlighted the incomplete sampling observed in the REMD trajectories and point to the influence of the starting conformation and the force field used.

5.5 Percentage of NMR structures in the REMD ensemble:

5.5.1 Introduction:

Figure 5.21 demonstrated that the REMD simulations were able to mirror a broad range of the NMR conformational motions. To determine the population of the NMR ensemble in the REMD trajectories, Dash, a torsion angle based cluster algorithm was used. Dash is a tool to analyse torsion angles from MD trajectories as a time series [161]. It aims to identify the major characteristic of each torsion angle to identify the most populated conformations of the molecule. Further details about Dash can be found in *section 3.7.1.*

Each REMD trajectory was processed with the Dash algorithm using the torsion angle definition presented in *table 4.1*. Given the torsion angle values as input, a representative torsion angle ensemble for similar conformations defined as a sequence of individual torsion angle states was obtained. Each frame from a given trajectory will therefore be associated with a Dash state with a specific torsion angle distribution. The Dash states will then be sorted according to their population in the trajectory. Thus, by associating each NMR structure to a Dash state, the population of the NMR ensemble in the REMD trajectories can be determined.

5.5.2 Results:

Figure 5.23 shows the population of the NMR structures associated with a particular Dash state as a black bar onto the REMD trajectories of GSK2194069 *(A)* and compounds 2, 3 and 4 *(B, C, D)* as red bar for conformation 1 in the Amber Gaff force field. Similar results are observed for conformation 2 in both force field and for conformation 1 in the CGenFF Charmm force field.

The NMR structures can be found in the REMD simulations of GSK2194069 and compounds 2 and 4 with a population of 100 % while less than 80 % of the compound 3 NMR data were found in the REMD trajectories. The trapping of torsion angle $\tau 4$ in compound 3 in the REMD simulations can explain the lower percentage from these simulations *(see Figure 5.17)*. Furthermore, the population of NMR structures assigned to Dash states does not reflect the entire REMD ensembles with an average of 80% of the REMD trajectories assigned to an NMR structure. The remaining simulations correspond to new states, which have not

been observed in the NMR ensemble. REMD simulations thus enhance the sampling.

Additionally, the clusters are in decreasing population in the REMD simulations and do not coincide with similar population for the corresponding NMR structures. The two ensembles are not correlated. *Figure 5.23* highlights that, what is seen in the NMR ensemble is found by the REMD trajectories but the ranking is not valuable in terms of compound preorganization, different populations are observed between the two ensembles.

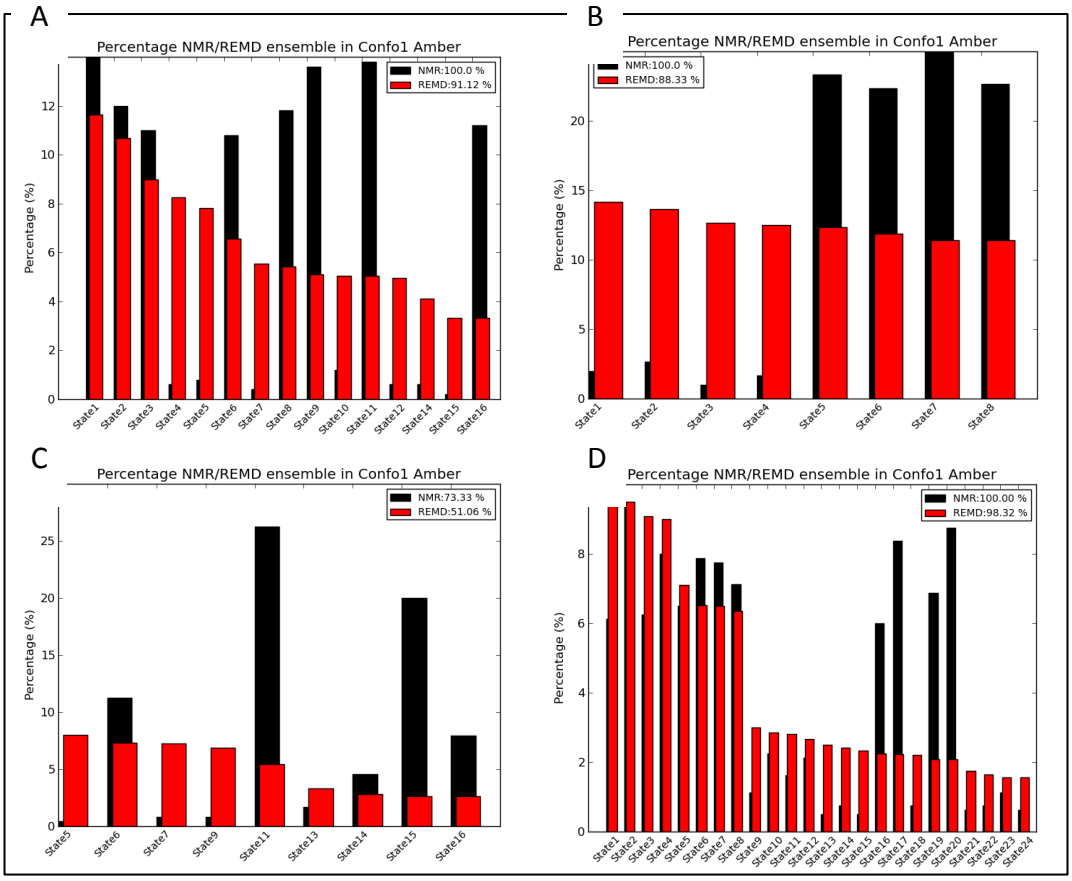


Figure 5.23: Histogram of the NMR ensemble (black bar) population into the REMD trajectories (red bar) Dash states in GSK2194069 (A) and the compounds 2 (B), 3 (C) and 4 (D) for the conformation 1 with the Amber Gaff force field. The total number of Dash states for each compound is 16 for GSK2194069, 8 for compound 2, 16 for compound 3 and 24 for compound 4.

Therefore, analysing the population of the NMR conformations observed in the REMD ensembles highlighted the increase in term of conformational sampling

while using REMD. Several new conformations sampled in the REMD trajectories were not observed by the NMR data.

5.5.3 Conformational generator:

The aim in this project is to describe the conformational ensemble of our small molecules in solution using experimental and computational methods. The REMD has shown itself to be able to reproduce the NMR data as a whole and sample more conformational phase space. However, the analyses also revealed the challenge encountered to produce reliable prediction populations. As an additional test, the REMD ensembles are compared with one obtained using OMEGA from OpenEye [174]. OMEGA can produce conformational ensembles for a given molecule at high speed but will not be able to indicate population of the conformations. More details of the OMEGA algorithm can be found in *section 3.7.6*.

To evaluate the efficiency of OMEGA generator compared to REMD, a PCA study was performed in the same way as introduced in *section 5.3.1*. The overall motions of the ensemble extracted from the generator were compared to REMD and NMR data. As for the PCA performed in the *section 5.3.1*, the REMD, NMR and generator ensembles were concatenated into a “super trajectory” and each ensemble is projected onto the combined trajectory.

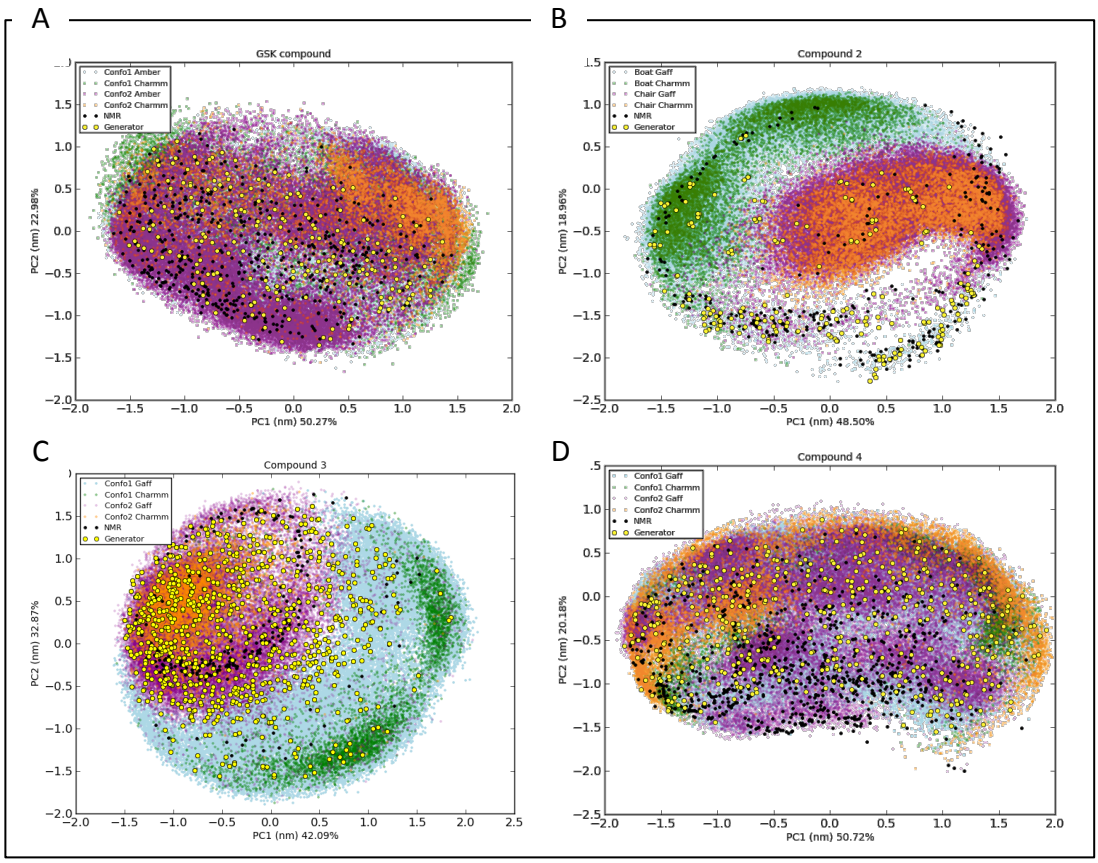
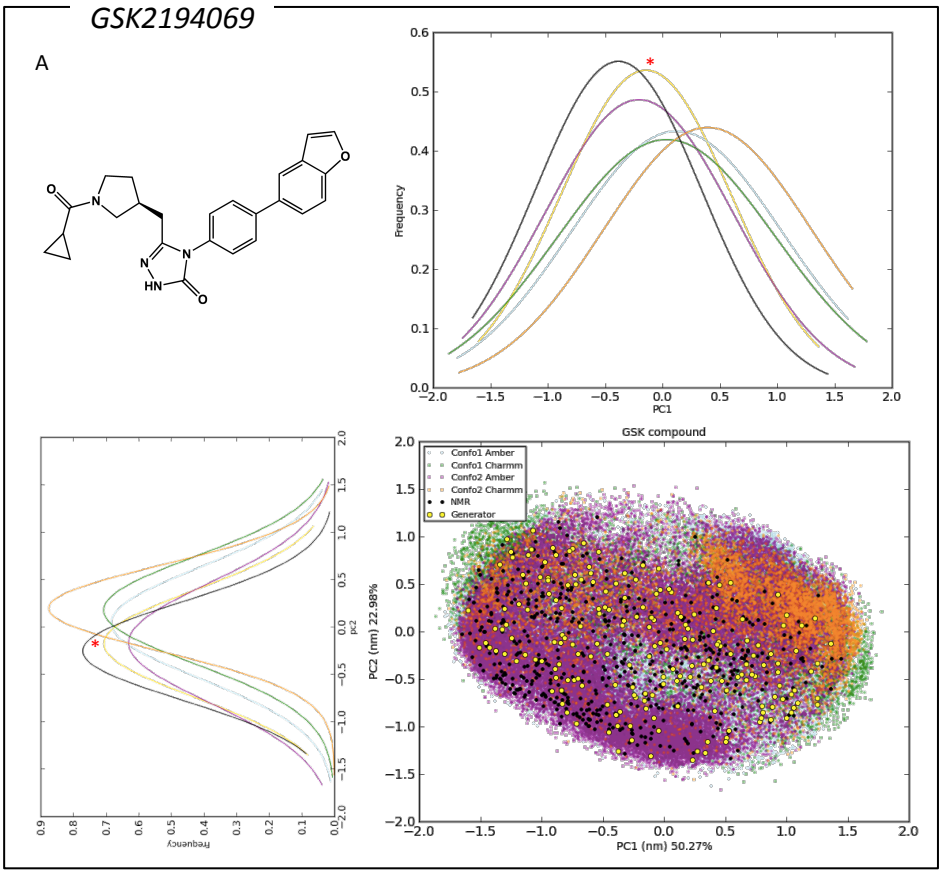


Figure 5.24: Principal component analysis of the first two components of the REMD, NMR and the OMEGA generator ensembles for the GSK2196049 (A) and compounds 2 (B), 3 (C) and 4 (D). The colours refer to the different conformations and force fields used. The NMR ensemble is shown in black dot. The REMD using the Amber Gaff force field is in light blue and magenta and the Charmm CGenff force field in green and orange for the conformations 1 and 2 respectively. The OMEGA generator ensemble is in yellow dot.

Figure 5.24 shows the conformational phase space sampled by the REMD, NMR (black) and generator (yellow) ensembles projected onto the first two principal components. The REMD ensembles capture the entire NMR and generator samples in all the compounds. Compared to OMEGA, the REMD simulations generate a diverse ensemble of structures and are able to reproduce the entire NMR ensemble.

However, the Omega dots in yellow appear to be uniform compared to REMD as it can be seen with the histogram of the omega conformations on the corresponding PCs of GSK2194069 (*Figure 5.25A red asterisk*) and compound 2 (*Figure 5.25B red asterisk*). The results for compounds 3 and 4 can be found in

Appendix B Figure B.6. Furthermore, the position of the NMR dots in the compound *(Figure 5.24B black dots and Figure 5.25B red asterisk)* appears to be consistent with the Omega conformations *(Figure 5.24B black dots and Figure 5.25B red asterisk).*



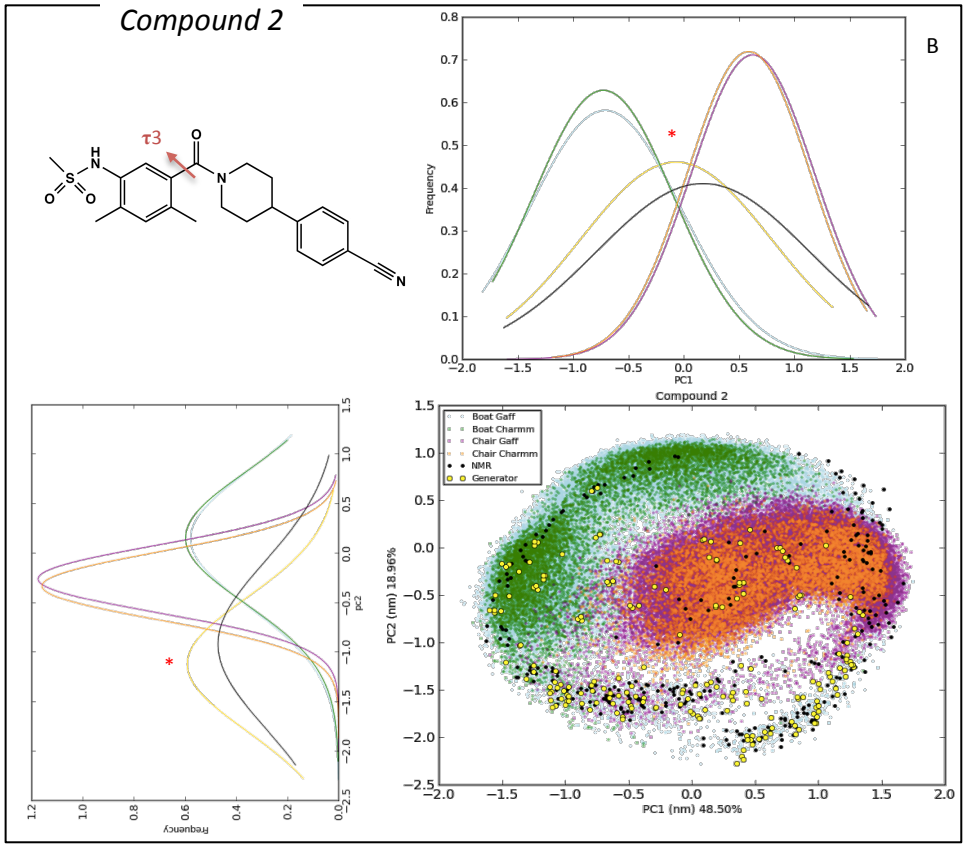


Figure 5.25: Distribution of the REMD trajectories of GSK2194069 (A) and compounds 2 (B), the NMR ensemble and the Omega generator along the two first components identified by PCA. The colours refer to the different conformations and force fields used. REMD using the Amber Gaff force field is in light blue and magenta and the Charmm CGenFF force field in green and orange for conformations 1 and 2 respectively. The NMR ensemble is represented by black dots and Omega by yellow dots. The histogram of the ensemble's projections on the corresponding PCs for each compound is displayed next to the PCs projection and the red asterisks emphasised the omega histograms.

Therefore, the results obtained from omega emphasised that the PCA is driven by the torsion angles trapping in this particular study. The omega data and NMR data are consistent in term of populations as seen with their histogram overlapping in *Figure 5.25B.*

5.6 Conclusion:

In summary, this chapter has shown that the REMD ensembles were able to reproduce the NMR data and provide new conformations of the compounds by sampling more phase space. However, it appears that some results are inconsistent between the REMD and the NMR as for torsion angle τ1 of GSK2194069 and compound 4 *(Figure 5.1 and Appendix B B.5)*. The REMD simulations of both compounds have shown inconsistency with the NMR ensemble. Furthermore, the comparison with the CSD data has also emphasized the importance of the query used to perform the search. A too general query led to a misleading or less satisfactory sampling as seen for compound 3 torsion angles τ2 and τ3 *(Figure 5.15 and Figure 5.16)* and for the torsion angle τ3 of compound 4 *(Figure 5.19).*

Additionally, differences were also observed between the starting conformations and between the Amber Gaff and CGenFF Charmm force fields. According to the starting conformation, torsion angle trapping was observed as seen with the amide torsion angles of GSK2194069 and compounds 3 and 4 *(Figures 5.2, 5.17 and 5.18)* emphasizing that the REMD simulations were not yet converged. In term of the force field, in some cases one will demonstrate better consistency with the NMR and/or the CSD data or vice versa as seen for the torsion angle τ3 of GSK2194069 with Gaff performing better than CGenFF *(Figure 5.5)* or for the torsion angles τ1 and τ2 of compound 2 with CGenFF performing better than Gaff for τ2 *(Figure 5.10)*. Determining whether the difference is due to the force field parameterization or to the conditions under which the NMR was performed appears as a difficult task. Some chemical group such as sulphonamide are difficult to parameterise in REMD and can contribute to the differences observed as well as the atropisomerism encountered for torsion τ3 in compound 2 *(Figure 5.13)* highlighted the challenge with conformational trapping in drug development.

The comparison of the torsion angle distributions and the PCA analysis of GSK2194960 with compound 4, scaffold hopping of GSK2194069, have shown good consistency between the REMD and NMR and the CSD data for the majority of the torsion angles of compound 4. Consistency between the equivalent torsion angles between GSK2194069 and compound 4 has been observed *(see sections 5.2.2.3.4 and 5.4.1.3).* Therefore, there is no evidence to suggest that the torsion angle distributions have affected the potency of compound 4 *(compound 4 pIC_{50} = 4.5, GSK2194069 pIC_{50} = 7.5)*, however the low potency could be due to different

factors. By changing the central core, the conformations of pendant groups may have changed. Additionally, the central core could induce weaker interactions when it binds to the receptor or a solvation/desolvation of the compound can also cause the lower potency of compound 4. Thus, further analysis of the contacts made with the receptor in the following chapter will help in understanding the cause of the lower potency.

Therefore, the NMR data provided us with the basic information for understanding the range of motion of our compounds. The combination of both experimental and computational methods appears to be a good start to investigate ligand preorganization in solution for conformational control to achieve improved binding affinity.

Chapter 6 Use of NMR and REMD analysis to understand small molecule inhibitor conformational preorganization

6.1 Introduction

The analyses of the NMR ensembles with the REMD trajectories in *sections 5.3 and 5.4* have shown the capability of the REMD simulations to capture the NMR ensembles for GSK2194069 and compounds 2 and 4 with all NMR conformations identified within the REMD ensemble. Less than 80 % of compound 3's NMR ensemble was found by REMD. REMD also sampled conformations not seen by the NMR. To describe the conformational ensemble sampled by these two methods in solution, additional analyses have been performed. The aim was to assess the ability of the methods not only to generate diverse conformational ensembles but also to evaluate the relative populations of the bioactive conformation within these ensembles.

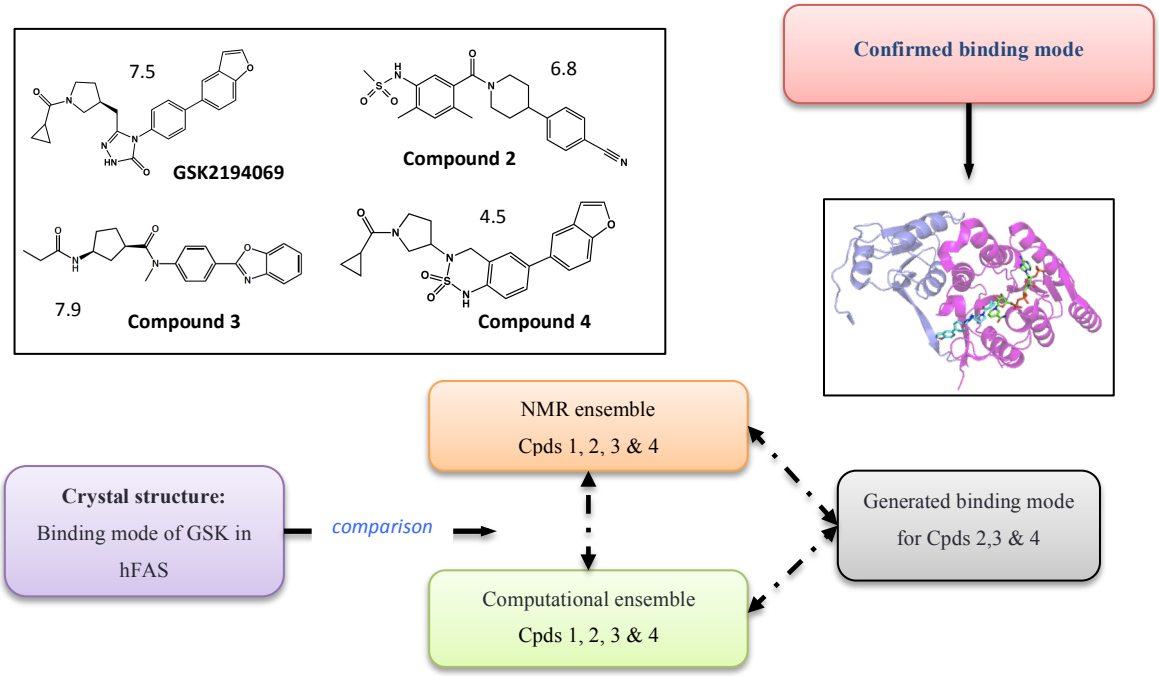
In this chapter, computational and experimental methods have been used to study the conformational dynamics of a selection of hFAS inhibitors. As part of this analysis, structural knowledge from the x-ray crystal structure of GSK2194069 in the KR domain of hFAS was used to elucidate potential binding modes for hFAS inhibitors 2, 3 and 4 [62]. Our hope is that learning from GSK2194069 could then be applied to compounds 2, 3 and 4 where the binding mode has not been elucidated, but the solution conformations are well understood. The chemical structures and activity data for the hFAS inhibitors studied here can be found in the following *Scheme 6.1*. Furthermore, to complete our analysis, docking technique was used to gain additional insights into the binding mode.

Initially, a torsion based clustering was performed using Dash [161]. The torsion angle definition for each compound was selected with respect to the NMR data *(Table 4.1)*. The GSK2194069 x-ray crystal structure was used as a reference to compare the cluster conformations using the root mean square deviation (RMSD) to measure similarity. Using the combination of Dash and RMSD raised questions about the efficiency of the overlap approach. Indeed, by using a cluster analysis the REMD ensembles were reduced to representative conformations. Thus, only few structures were used to compare with the x-ray leading to loss of

information from the REMD trajectories. Structures with possibly higher affinity could not be taken into account when performing Dash clustering. Subtle changes in conformations could have a large effect on the compound affinity. An alternative tool was then used for such overlap *(ROCS from Openeye [175], see section 3.7.6 for a description of ROCS)* to perform a similarity search on the whole REMD and NMR ensembles using the crystal ligand structure as reference [62]. In this chapter we are going to review two methodologies, a Dash/RMSD cluster on representative Dash states and ROCS including all the REMD trajectories frames. Overall, these analyses will provide a potential basis for further chemical design for more effective and highly potent analogues of these inhibitors.

6.2 Methodology:

Dash was used to reduce the number of structures within the ensembles without losing information whilst being able to compare the ensembles. The *scheme 6.1* illustrates the protocol used in this analysis.



Scheme 6.1: Illustration of the study workflow. The binding mode picture corresponds to GSK2194069 and the NADPH binding to the active pocket of KR domain of hFAS. The compounds potencies in pIC_{50} appear next to the compound's structure.

For each compound, computational and experimental ensembles were calculated from the REMD simulations and NMR data from the C4X Discovery Conformetrix platform. The x-ray crystal structure of GSK2194069 in the KR domain of hFAS enables us to use this compound as a reference. Indeed, the results from both, the NMR and REMD ensembles were combined, using the information of GSK2194069 to generate potential binding modes for the potential inhibitors of hFAS without a known crystal structure. *Figure 6.1* summarizes the detailed workflow of the different steps and methodologies used to compare the ensembles.

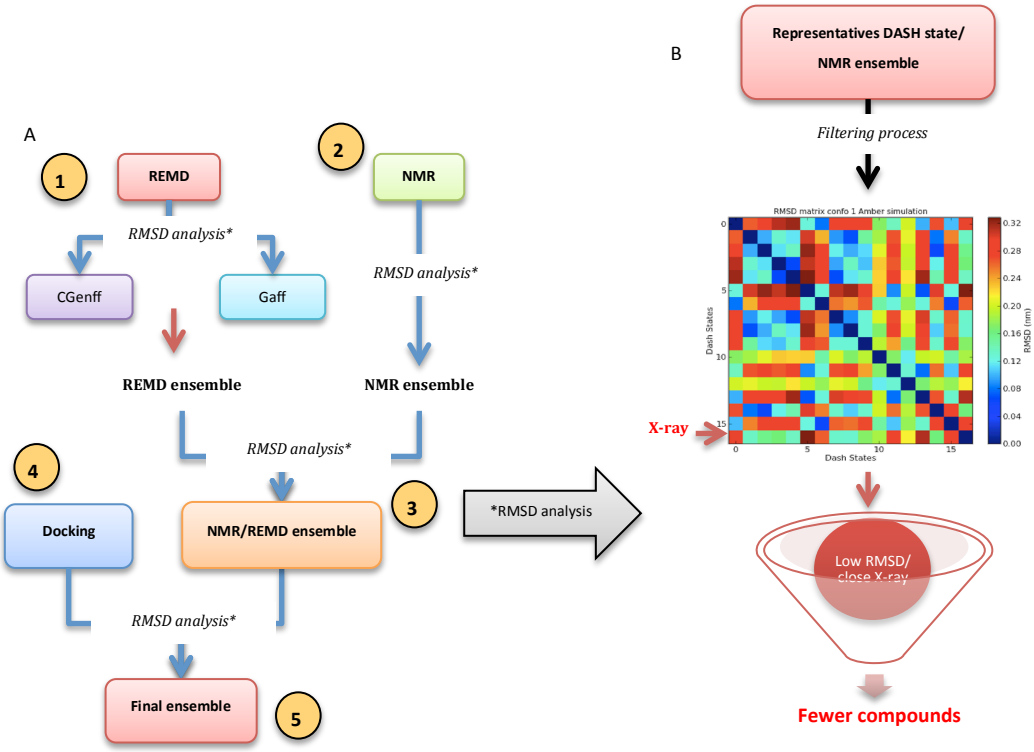


Figure 6.1: Protocol to filter and compare the NMR, REMD and docking ensembles. A: The different steps of the ensembles comparison between the different methodologies (REMD, NMR and docking). B: The tool used to cluster and reduce the ensembles to fewer representative compounds.

Four REMD simulations using two different starting conformations and the Amber Gaff and CGenff Charmm force fields were performed for each compound for 500 ns using 48 replicas. The total simulation length for each REMD trajectory

was of 24 μs; further details of the simulations are provided in *section 4.2*. To reduce the size of the ensembles, whilst minimizing the loss of information, Dash, was used. Given a list of torsion angles describing the trajectory, a series of states that correspond to the number of clusters identified in the trajectory were obtained. Each state was described by one structure, which is referred to as the representative conformation. The trajectory frames that best represent each Dash state are then processed through an RMSD search using the confirmed binding mode as reference. By not having a known x-ray crystal structure for compounds 2, 3 and 4, the x-ray crystal structure of GSK2194069 was used as a reference for those compounds as well as for GSK2194069 to find structures with a close binding mode observed in the x-ray. The MDTraj python library was used here to perform the alignment and the RMSD calculation [183]. MDTraj is an open library in python for the analysis of MD trajectories from broad range of MD software. The alignment tool allows the user to superpose an ensemble of conformations onto a reference structure using the atom indices. The RMSD protocol followed two steps:

1. Alignment of each frame of the trajectory to the reference structure using atoms indices.

For the simulations of GSK2194069, given that the x-ray structure was used as reference, the superposition of the atom indices between the two compounds is straightforward. However, when additional compounds are analysed in this way we need to understand how to perform a suitable alignment. *Figure 6.2* illustrates the superimposition of each compound onto the x-ray structure. By using the known binding mode of GSK2194069 the similarities between compounds are evaluated in order to predict the binding modes of additional compounds.

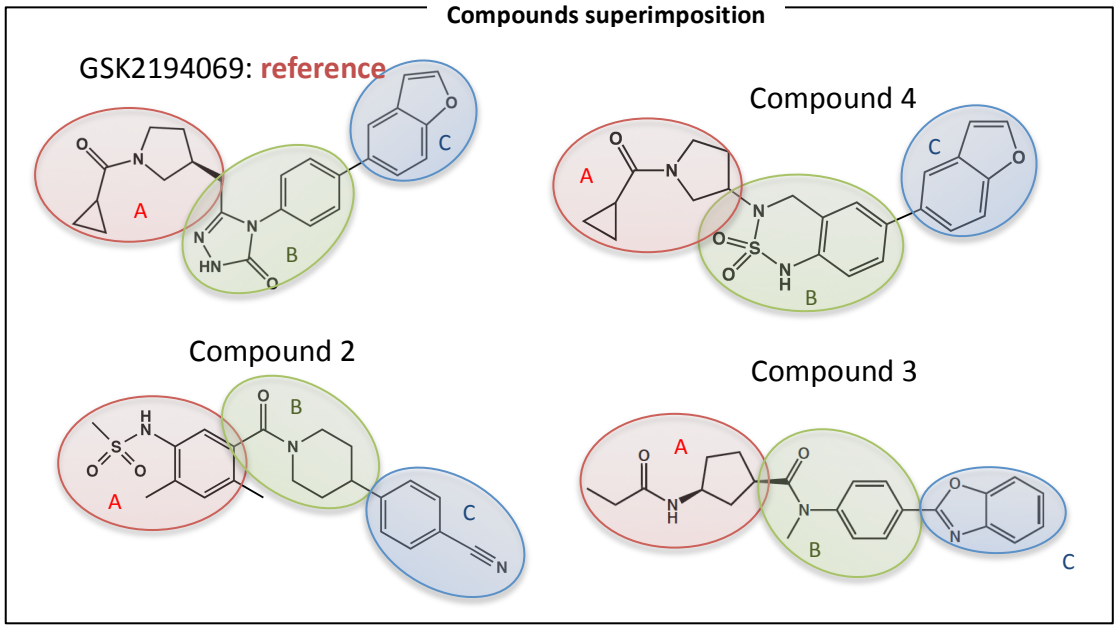


Figure 6.2: Chemical group similarities between the four hFAS potential inhibitors.

To help the visualisation of how the alignment was perform, each compound can be broadly considered in these three parts, referred to as A, B and C. Furthermore, from the crystal structure of GSK2194069 the different contacts made with the receptor are known. Thus, by matching the three groups A, B and C to the corresponding part of the x-ray structure, whether these alignments are potentially compatible with the binding mode will be understood. However, the superimposition is an atom based RMSD alignment. Thus, more than matching the three groups A, B and C from the compounds 2, 3 and 4 with the equivalent A, B and C in the x-ray structure, the atoms in the different groups from each compound need to match the equivalent atoms in the reference. The protocol used here was to evaluate the chemical similarity, in term of reproducing important native contacts, of each atom between the compound and the reference as well as hypothetically predict the possible interaction each atom could make with the receptor based on the GSK2194069 x-ray binding mode. The atom pairs used to compute the RMSD between the compounds 2, 3 and 4 and the GSK2194069 x-ray structure can be found in *Appendix C Figure C.1 and Table C.1*.

2. Computing RMSD of the different conformations to a reference structure:

The results obtained from the RMSD analysis were filtered by RMSD value (≤ 0.20 nm) and visualised within Pymol [184]. The aim was to identify conformations similar to the binding mode from the x-ray. This protocol was then applied to each REMD trajectory *(Figure 6.1B).* The following steps were taken to compare the multiple REMD trajectories with the NMR and docking ensembles.

Step 1: Comparing the results from the REMD trajectories in term of force field differences. To focus on structures most similar to the x-ray binding mode, when similar structures are found within ensembles, the structure with the lowest RMSD value is kept. Once all the trajectories have been analysed a "REMD ensemble" is obtained which describes conformations across the REMD simulations most similar to the x-ray structure.

Step 2: Similarly to the REMD ensembles, the NMR data are processed through the RMSD protocol using the GSK2194069 x-ray crystal structure as reference. After filtering, structures with low RMSD and shape similar to the x-ray, the "NMR ensemble" can be obtained.

Step 3: Comparison of the NMR and REMD ensembles with regard to similarities and differences to obtain an ensemble of structures from both methodologies. At this step the ensemble obtained is named "NMR/REMD ensemble".

Step 4: To help predict the binding modes of the compounds with hFAS a molecular docking program was used. As all compounds have been reported to bind with the same mechanism of action, the four compounds were docked into the receptor pocket in the presence of NADPH, using AutoDock [157]. Further details on the docking protocol are described in *section 3.6*. The docking method was to evaluate if a simpler approach could also reliably reproduce the x-ray (or predicted) binding mode. In this manner, two docking calculations with 500 runs were performed for each compound starting from the two conformations used in the REMD trajectories, with a rigid receptor. Clustered conformations were analysed using RMSD filtering with a visual comparison of the docked poses with the x-ray to keep conformations most similar to the x-ray binding mode. The results were then compared with the structures from "NMR/REMD ensemble" obtained in step 3. The ensembles comparison will evaluate if more than one potential binding mode can be identified.

From the RMSD protocol a final ensemble of structures similar to the x-ray binding mode from experimental and computational methods was obtained. To assess the quality of the alignment the use of the knowledge of the interactions made between GSK2194069 and the receptor was used.

6.3 Results:

6.3.1 Compound selection and filtering:

6.3.1.1 GSK2194069 RMSD search:

Cross correlation matrix of the GSK2194069 final ensemble structures (step 4) indicates the similarity of the structures within each other and with the x-ray crystal structure *(Figure 6.3)*. The rows and columns refer to the different structures kept at the end of step 4 from the various methods. The label A and B corresponds to the REMD and docking starting conformations and the force field used are identified as Charmm and Gaff. A structure from REMD simulation of the conformation A using the Charmm CGenFF force field will be identified as CharmmA in the plot. The RMSD results of the superimposition of the x-ray structure with the final ensemble appear at the last row highlighted with a red star. The range of RMSD values (nm) is indicated by various colours in the panel from o to 0.2 nm with blue to yellow colours indicating structures similar (RMSD < 0.015 nm) while red colour reflects poor correlation between the structures.

From *Figure 6.3,* looking at the similarity between the different GSK2194069 conformations from the different methodologies (REMD, NMR and docking), without comparison with the x-ray structure, shows that there is no apparent structure in term of relative similarity within the various techniques. Some are more similar to other like NMR and docking, however, no pattern is observed. Thus the rest of the matrix *(not the last row, Figure 6.3 red asterisk)* shows no particular significant correlation between the structures.

In term of comparison with the x-ray structure (*Figure 6.3 red asterisk)*, docking conformation A identifies a potential binding mode closest to the x-ray crystal structure, with an RMSD value of 0.078 nm. This structure corresponds to the second best conformation (in term of docking score) extracted from the docking analysis with energy of binding of -11.38 kcal/mol. The RMSD values of the remaining structures range from 0.12 nm to 0.14 nm. To evaluate the

accuracy of the RMSD search, the alignment of the different structures with the X-ray were visualised using Pymol [184].

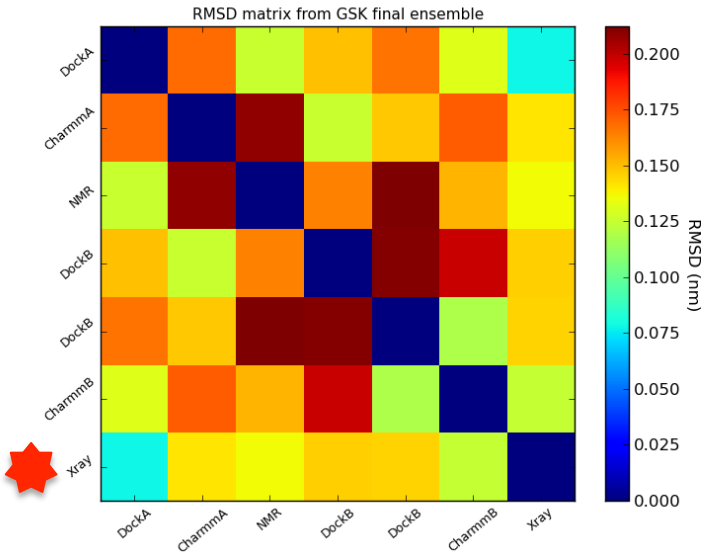


Figure 6.3: Correlation matrix between the final ensemble structures of GSK2194069. The rows and columns of the matrix represent each conformation from computational and experimental data with the last row corresponding to the x-ray structure (red asterisk). The colour of the cells corresponds to the RMSD value between the pair of configurations studied in nm. The nomenclature used here for the two starting structures is A and B, the two force fields are referred as Charmm and Gaff and the docking simulations as DockA and DockB.

The two structures with the highest docking score, named best-docked poses, from the docking simulation of conformation A were first aligned to the x-ray structure in blue and are shown in *Figure 6.4*. The benzofuran and central triazolone moiety of the two poses displays good overlap with the x-ray structure with a slightly better alignment for the second best pose (green). However, the oxygen atom of the cyclopropyl moiety, which appears to be involved in a hydrogen bond with the receptor (residues tyrosine2034 and serine2021), has an opposite orientation for the second best pose than the equivalent oxygen atom of the x-ray structure *(Figure 6.4 red circle)*. The RMSD calculation is based on atom indices superimposition and distances between the reference compound and the structure. Therefore, a lowest RMSD value for the second best pose even with the oxygen atom in other different orientation than the x-ray can be explained by a better overlap in term of other atom distances.

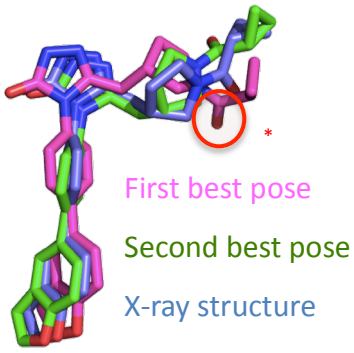


Figure 6.4: The two structures with the highest docking score aligned to the x-ray in blue.

The REMD and NMR structures from the final ensemble aligned to the x-ray binding mode are shown in *Figure 6.5* with their corresponding RMSD values. Compared to the docking result of conformation A *(Figure 6.5 cyan asterisk)*, the orientation of the oxygen atom of the cyclopropyl group is similar to that observed in the x-ray for the REMD, NMR structures and for the docking structure of the conformation B. Regarding the oxygen atom of the triazolone moiety *(red circles in Figure 6.5),* which is involved in a hydrogen bond with the receptor residue serine2081, for each of the structures the orientation is consistent with the x-ray. However, the benzofuran moiety *(green circles in Figure 6.5)* involves in a π-arene interaction with the receptor show an orientation flip that could decrease the compound affinity for the conformation A with the CGenFF force field and the docking conformation B as indicating in coloured asterisks in *Figure 6.5*. The colours of the star are the colours of the structures showing differences with the x-ray.

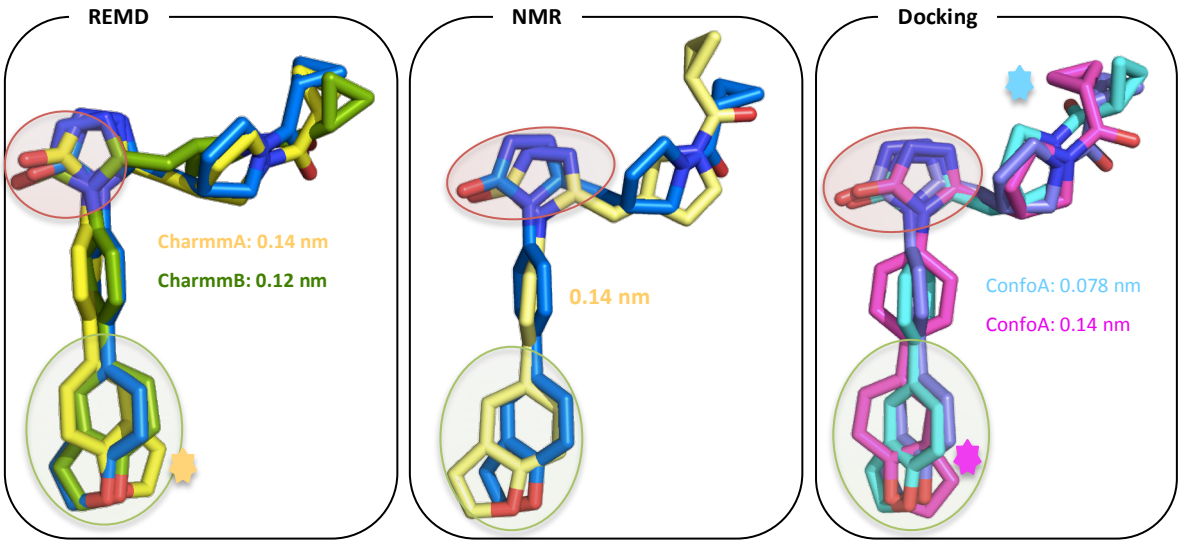


Figure 6.5: Alignment of GSK2194069 structures from step 4 with the x-ray structure in blue with their corresponding RMSD values. The structures from the REMD simulations with the CGenFF Charmm force field for conformations A and B are in yellow and green respectively. The NMR structure is in light yellow and the docking structures from the conformations A and B are in cyan and magenta respectively. The stars indicate the structures that show differences with the x-ray structure.

Therefore, the analysis of the final ensemble structures of GSK2194069 has shown structures with good overlay with the x-ray corresponding to conformation B of the REMD simulation with CGenFF force field and the NMR structure. However, the analysis of the docking structures revealed that structures with better alignment could here been discarded due to the use of the RMSD filtering criterion. Indeed, the first best docked structure was discarded from our RMSD filtering while it appears to be a good structure in term of similarity with the x-ray binding mode.

6.3.1.2 Compounds 2, 3 and 4:

Figure 6.6 shows the alignment of the final ensemble structures with the x-ray structure from the Dash/RMSD analysis for compounds 2, 3 and 4. The nomenclature used here for the two starting structures is 1 and 2 and the two force fields are referred as Charmm and Gaff. Thus, if a structure corresponds to the REMD simulation of the conformation 1 with the Gaff force field it appears as

Gaff1. The red asterisks indicate the RMSD values of the alignment between the reference (x-ray) and the final ensemble structures.

Compare to GSK2194069, the correlation matrix of the compound *2 (Figure 6.6A red asterisks)* emphasised that the structure from the REMD simulation of the conformation 2 with the Amber Gaff and CGenFF Charmm force fields (Gaff2 and Charmm2) display poor overlay with all the different structures with high RMSD values. Additionally, compared to compounds 3 and 4 *(Figures 6.6B and C)*, the correlation matrix of compound 2 shows higher RMSD values making the conformations from the Dash/RMSD filtering less similar to the x-ray binding mode.

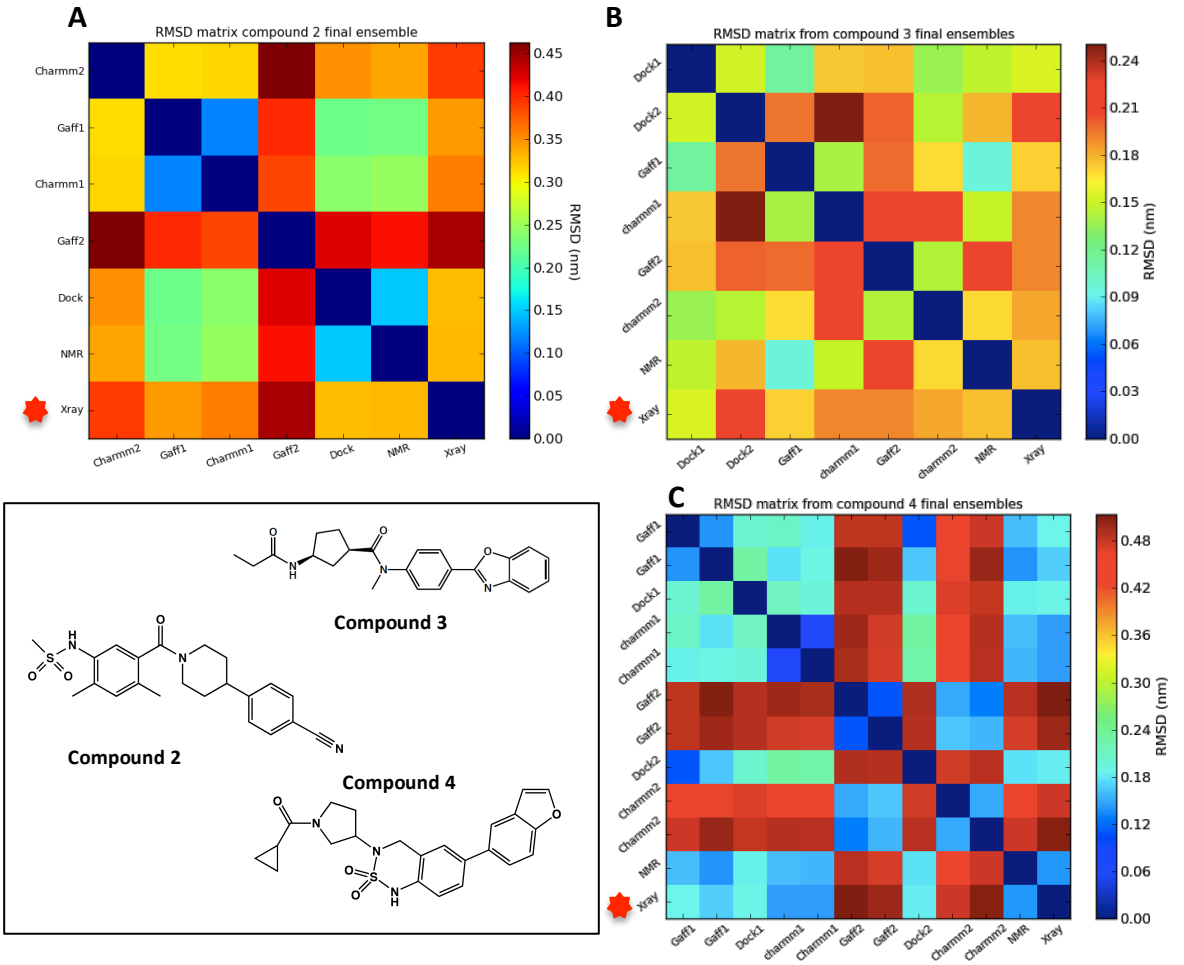


Figure 6.6: Correlation matrix between the final ensemble structures of the compounds 2 (A), 3 (B) and 4 (C). The rows and columns of the matrix represent each conformation from computational and experimental data with the last corresponding to the x-ray structure. The color of the cells corresponds to the RMSD value between the pairs of conformation studied in nm. The nomenclature used here for the two starting structures is 1 and 2,

the two force fields are referred as Charmm and Gaff and the docking simulations as Dock1 and Dock2.

Similarly to the GSK2194069 analysis, the aim of the correlation plots is to highlight the structures from the final ensemble that show similar binding mode as the x-ray. The final row with a red star refers to the RMSD values of structures aligned with the x-ray. As in *Figure 6.3* the colour panel corresponds to the RMSD value in nm with increasing values from blue to red. Compared to GSK2194069, the RMSD values are bigger and vary from 0 to 0.5 for compounds 2 and 4 and to 0.30 for compound 3. Greater values are explained by the structures differences between the reference structure and the compounds 2, 3 and 4. However, the result of compounds 4 shows lower RMSD values for most of the structures *(Figure 6.6C blue squares)* as expected by being designed as scaffold hop from GSK2194069 structure.

Compound 2: *Figure 6.6A* shows that all methods display similar results in term of alignment with the x-ray structure of GSK2194069 with RMSD values ranging from 0.33 to 0.45 nm with the REMD simulations of conformation 1 with the CGenFF Charmm and the Amber Gaff force fields showing higher RMSD values. The docking and NMR structures as well as conformations 1 with the CGenFF Charmm and with the Amber Gaff force fields appear to perform slightly better with lower RMSD values. *Figure 6.7* shows the alignment of each structure with the x-ray and their corresponding RMSD values. The colour circles emphasize regions in the molecules that show differences with the x-ray.

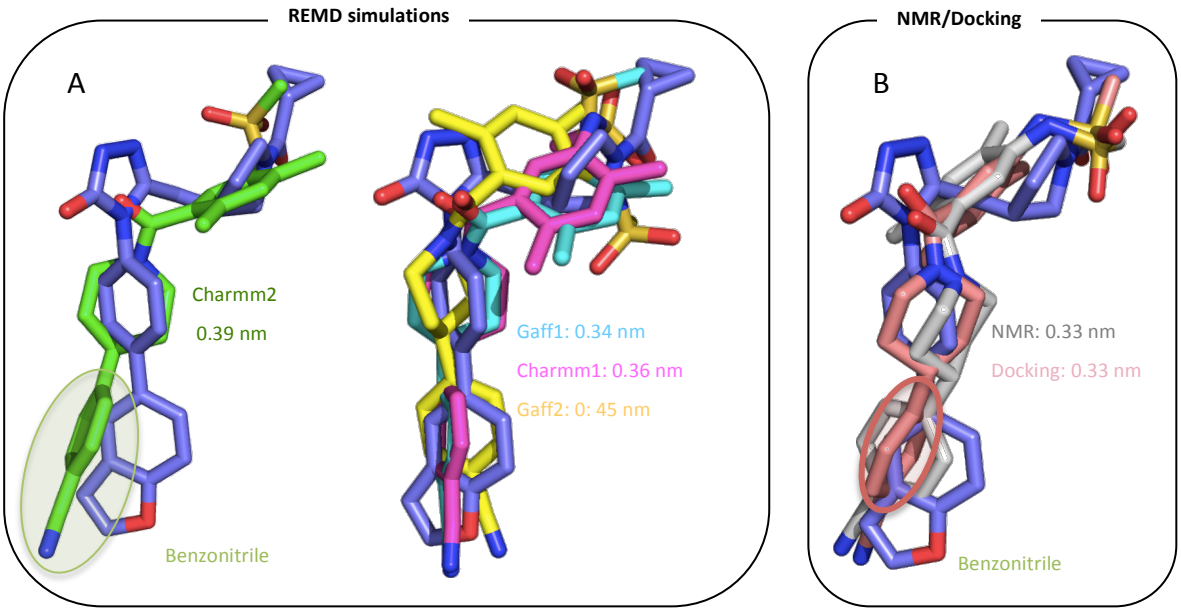


Figure 6.7: Compound 2 structures from the Dash/RMSD analysis final ensemble aligned with x-ray in blue and their corresponding RMSD values in nm. A: REMD simulation final structures, B: NMR and docking final structures.

Figure 6.7A highlights good overlay between the REMD simulations Gaff1, Gaff2 and Charmm1. From the *section 5.2*, *Figure 6.2* demonstrated how the different parts of compounds 2, 3 and 4 show complementarity with the x-ray indicated by A, B and C. Therefore, the sulphonamide moiety (A), the carbonyl group of the amide (B) and the benzonitrile (C) show similar orientation as cyclopropyl (A), triazolone (B) and benzofuran (C) moieties of x-ray structure, moieties involved in important contacts with the receptor. The REMD structure identified from the simulation of the conformation 2 with the CGenFF Charmm force field despite good overlay of the sulphonamide moiety and carbonyl group with the x-ray structure, show an orientation of the benzonitrile group away from the benzofuran moiety of the x-ray which may induce clashes if binding to the receptor in that region *(Figure 6.7A green circle).*

Regarding the NMR and docking structures, they display good agreement with the x-ray in term of shape alignments. However, despite the fact that the docking ensemble has the lowest RMSD value, there is one aspect of the binding mode that is not in full agreement. π-arene interaction was observed between the residue phenylalanine 2109 and the benzofuran group in the x-ray structure that

help stabilize GSK2194069 into the binding pocket. However, the orientation of the benzonitrile group in the docking structure *(Figure 6.7B pink circle)* could lead to loss of this important contact making the structure less favourable for binding.

Compound 3: As for compound 2, *Figure 6.6B* shows the various methods have identified a small range of RMSD values, from 0.18 nm to 0.19 nm. The reason of higher RMSD values for compound 2 could be explain by the torsion angle τ3 distribution analysis from *section 5.2.2.1.2 (Figure 5.13)*. Torsion angle τ3 refers to a case of atropisomerism inducing an isolation of the conformations leading to trapping. Therefore, the trapping observed with compound 2 torsion τ3 distribution could explain the difference of RMSD values between the compounds 2 and 3.

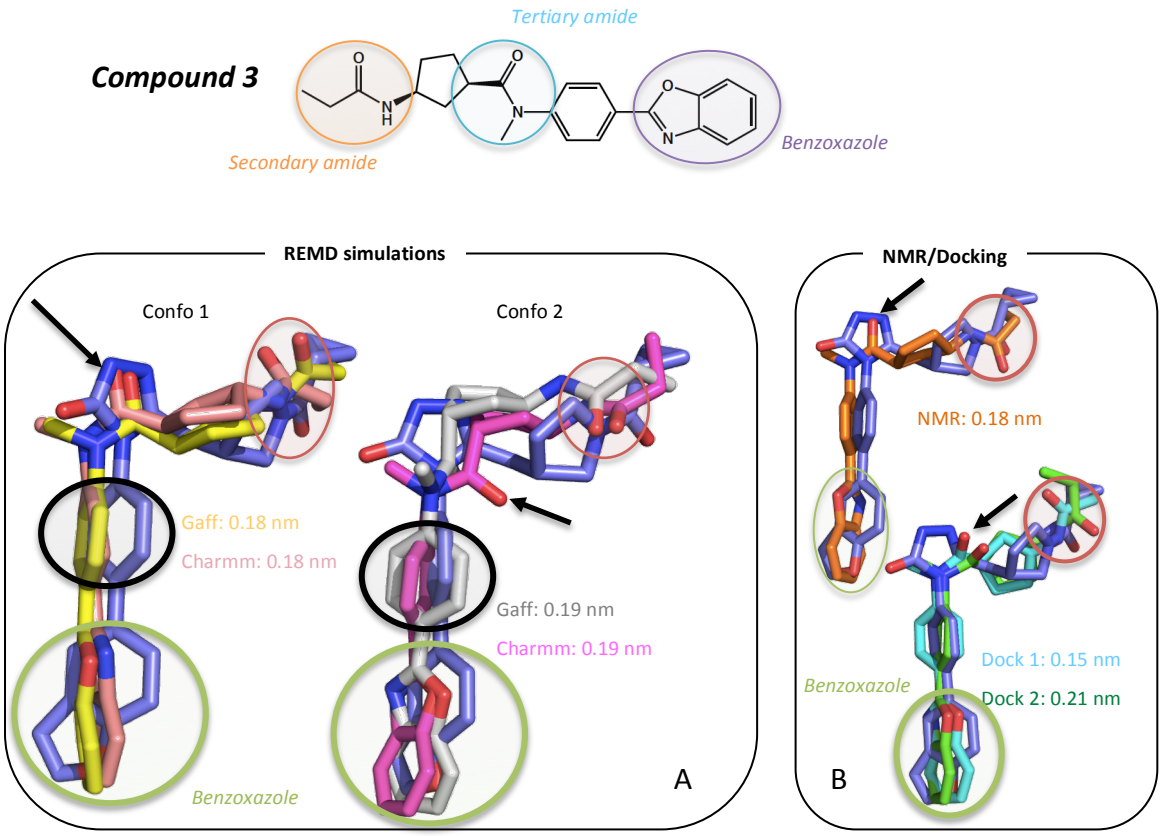


Figure 6.8: Compound 3 final ensemble structures aligned to the x-ray (blue) with their corresponding RMSD values in nm. A: REMD simulations final structures, B: Docking and NMR final structures.

The top part of *Figure 6.8* emphasizes the chemical groups considered as important in the compound 3 from the knowledge of the GSK2194069 binding mode. To assess the quality of the overlap between the structures of compound 3 from the final ensemble of Dash/RMSD analysis and the x-ray, the orientation of the three groups in colours circles in *Figure 6.8* will be compared with the x-ray structure of GSK2194069. The bottom part of *Figure 6.8* shows the alignment of structures from the different methodologies.

The alignment of the benzofuran group of the x-ray with the benzoxazole group of compound 3 shows two possible orientations adopted by the benzoxazole. The structure from the REMD simulations of conformation 1 in both Gaff and CGenFF force fields and the NMR structure show the benzoxazole orientated perpendicularly to the benzofuran moiety in the x-ray structure while the phenyl rings in both compounds are similarly oriented. However, the structures of the docking simulations of conformations 1 and 2 and of the REMD simulations of the conformation 2 in both Gaff and CGenFF force fields show similar orientation as the x-ray for the benzoxazole moiety while the phenyl rings have different orientation. *Section 5.2.2.2.5* have shown the distribution of the torsion angle $\tau 5$ of compound 3. Consistency between NMR and REMD ensembles were observed *(Appendix B Figure B.3).* Additionally, the NMR data reported that the distribution of torsion angle $\tau 5$ depends upon both steric and electronic effects leading to the phenyl ring being orthogonal to amide. However, the distribution of the compound 3 amide torsion angle $\tau 3$ have emphasized trapping *(section 5.2.2.2.4 Figure 5.17).* Therefore, the greater differences observed between compound 3 benzoxazole moiety and the x-ray structure benzofuran group could be explain by the amide trapping that influence the orientation of the phenyl ring and benzofuran moieties *(Figure 6.8 green and black circles)*.

Furthermore, the carbonyl of the secondary amide in *Figure 6.8* red circles in the REMD and the docking simulations is positioned in the opposite direction from the carbonyl oxygen atom of the cyclopropyl moiety of GSK2194069. It is therefore possible that besides causing a steric clash with the protein, an important contact may be lost.

The tertiary amide oxygen atom shows different orientations depending on the starting conformation and methodology *(Figure 6.8 black arrows)*. In the NMR ensemble and the REMD simulations of conformation 1, the oxygen atom is oriented toward the nitrogen atoms of the x-ray triazolone ring while the docking

structures and the structures from the REMD simulations of conformation 2 are oriented away from the x-ray structure.

Therefore, knowing that compound 3 is potent toward the receptor *(Figure 6.1)*, further analysis on the contacts of these structures made with the receptor will enable us to understand if the different orientation of the different moieties *(benzoxazole, phenyl ring and the carbonyl of the two amides)* will allow a different binding mode than GSK2194069.

Compound 4: In contrast to the other compounds, compound 4 was synthesis by "scaffold-hop", by changing the central core of GSK2194069 as highlighted in green circle in the *Figure 6.9D*. However, compound 4 appears to be less potent with a pIC_{50} of 4.5 *(GSK2195069* pIC_{50} = 7.5*)*. Additionally, compared to compounds 2 and 3, two conformations of each REMD simulations were kept from the Dash/RMSD filtering to assess the possibility of multiple binding modes.

As expected, the RMSD values of the final ensemble aligned to the x-ray are lower than observed with compounds 2 and 3 (≤ 0.18 nm). However, the REMD simulations starting with conformation 2 with the Gaff and Charmm force fields show higher RMSD values ≥ 0.49 nm *(Figure 6.6C)*. The alignment of the REMD, NMR and docking structures with the x-ray structure of GSK2194069 is shown in *Figure 6.9*.

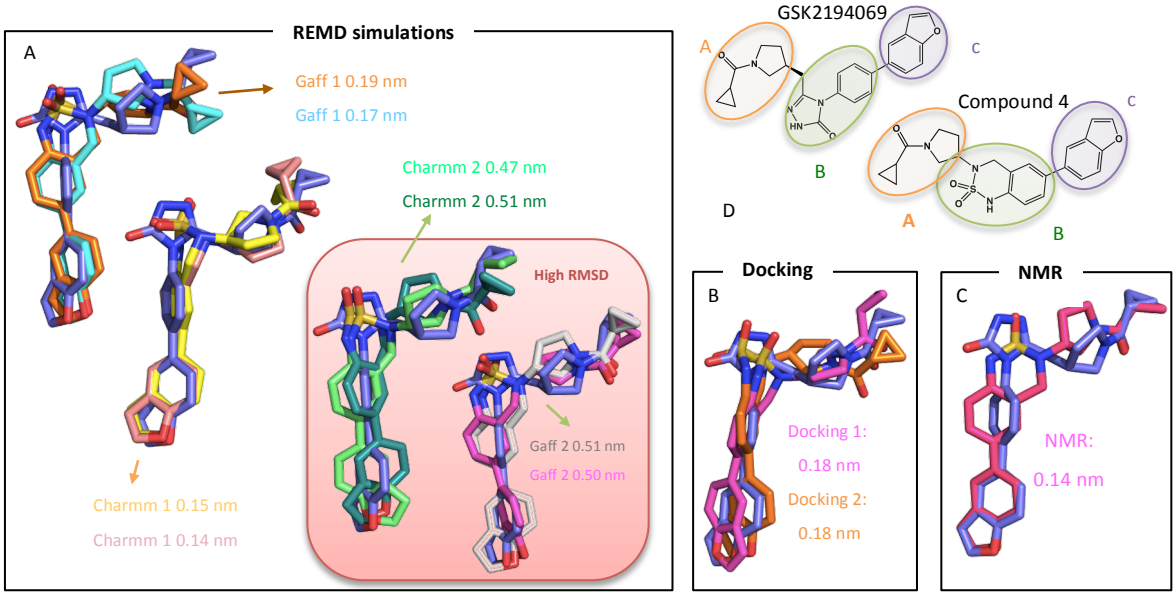


Figure 6.9: Compound 4 final ensemble structures aligned to the x-ray (blue) with their corresponding RMSD values. A: REMD simulations final structures, B: Docking and C: NMR final structures. The part D of the figure illustrates the similarities and differences between GSK2194069 and the compound 4 functional groups.

As was introduced in *Figure 6.2* and is shown in *Figure 6.9D*, the central core of compound 4 corresponds to the part B and will be compared with the corresponding part B of GSK2194069. Knowing from the x-ray binding mode that the triazolone moiety forms a hydrogen bond with the residue serine 2081 of the receptor, the orientation of the oxygen atoms of the central core of compound 4 will be compared to this atom. The part A and C are similar in both compounds and the orientation of the atoms involved in important contacts will be compared.

The REMD simulations starting with conformation 1 with the CGenFF Charmm force field *(Figure 6.9A top left)* shows good alignment with the x-ray structure of GSK2194069 with the lowest RMSD values of 0.15 nm and 0.14 nm. Indeed, the benzofuran moiety from the two compounds shows good overlap within each other. Then, the carbonyl of the cyclopropyl group of the two compounds is oriented in the same direction. The phenyl ring of the central core of the two compounds is perfectly aligned and the oxygen atoms of the sulphonamide are oriented similarly as the oxygen and nitrogen atoms of the triazolone moiety. Therefore, the alignment of the REMD simulations of conformations 1 with the CGenFf force field is promising in term of conservation of the x-ray binding mode.

As for the REMD simulations of conformation 1 with the CGenFF force field, the NMR structure demonstrated a low RMSD of 0.14 nm. The benzofuran and cyclopropyl moieties show good overlap. However, the orientation of the phenyl ring of the central core of the two compounds appears to be different. The rotation of the central core of compound 4 in comparison to GSK2194069 could induce the loss of key contacts and may cause internal clashes with the receptor.

The REMD structures of conformation 2 with the Gaff and CGenFF Charmm force fields show high RMSD values that can be explained by the relatively poor shape alignment. Indeed, despite the similar orientation of the carbonyl of the cyclopropyl group between the two compounds, which is known to be a key interaction, the central core and the benzofuran moiety shows greater differences and thus prevents a good overlap. The distribution of the torsion angle $\tau 4$ in *section 5.2.2.3.4* has shown inconsistency between the NMR and REMD ensembles as well as between the two force fields *(Appendix B Figure B.6)*. Therefore, the inaccuracy of the torsion angle $\tau 4$ distribution could explain the high RMSD observed for the REMD simulations of conformations 2. Additionally, further analysis with the receptor may inform if this observation will induce steric clashes.

The alignment of the docking structures shows greater differences in the cyclopropyl moiety between the structures and the x-ray causing higher RMSD values.

Therefore, from the analysis of the alignment of compound 4 to the x-ray structure of GSK2194069 have shown good overlay for the REMD simulations of conformation 1 with the CGenFF Charmm force field. However, large differences of the central core and the cyclopropyl moiety of the compound 4 for the other structures were observed that might cause a loss of key interactions, or possible steric clashes with the receptor that could be responsible of the low potency of the compound 4.

6.3.1.3 Conclusion of the Dash/RMSD search:

The combination of Dash clustering, RMSD search followed by a visualization of the final ensemble alignment with the x-ray binding mode of GSK2194069 has enabled us to identify conformations with similar binding mode(s). As would be expected, the analysis using GSK2194069 itself emphasized good overlap with the x-ray structure. For compounds 2, 3 and 4, similar patterns in term of the position of the functional groups were observed with some important differences highlighted.

The intent of the study is to generate binding mode(s) for novel hFAS inhibitors. Through the RMSD search, the contacts identified as important for the binding from the x-ray crystal structure were found in most of the final ensemble structures. However, the REMD simulations highlighted different behavior according to the starting conformation and the force field used but there is no clear evidence that one force field performed better than the other. The fluctuations observed in the different functional groups for compounds 2 and 3 in comparison to the x-ray structure of GSK2194069 could lead to different binding or to the lost of important contacts with the receptor. Further analysis on the contacts made with the receptor will enable us to better understand the binding mode of these potent compounds.

The analysis of compound 4 also demonstrated that the central core corresponding of the part of the compound that has been changed by “scaffold hopping” has shown large differences that could be responsible of a lower potency comparing to compounds 2 and 3.

Despite the good overlap observed between the x-ray structure and the four compounds, the choice of filtering the ensembles by Dash clustering followed by RMSD search could induce a loss of information. Indeed, by clustering the REMD trajectories using Dash, the structures analyzed were reduced to the representatives Dash states. Additionally, the RMSD calculation on the representatives Dash states, NMR and docking ensembles was made based on the atoms indices of each compound. Therefore, the structural superimposition of a compound different to that of the reference structure, before the actual RMSD calculation, may induce some variability and good overlays might be discarded.

To address these problems, the RMSD search was completed using the ROCS (Rapid Overlay of Chemical Structures) overlay method from OpenEye software to get a more satisfactory overlay. ROCS will compare all the conformation from each ensemble with the x-ray structure and measures the shape similarity combined with the chemical group resemblance between the two. The ensemble is then ranked based on the TanimotoCombo score ranging from 0 to 2.

6.3.2 Efficiency of the Dash clustering in term of relative populations:

To address the issue with the use of Dash clustering and the RMSD metric, ROCS was used to align the REMD, NMR and docking ensembles against the x-ray structure. By using ROCS it should be possible to use all structures from the REMD trajectories, therefore avoiding loss of information and also enabling us to try to exploit population information that we were not able to assess from Dash analysis. This was discussed in *section 5.2*. ROCS is a molecular shape-based superimposition method refined with a chemical complementarity measure. Once the structures are aligned, based on their molecular shape and their functional group *(see section 3.7.6 for a description of ROCS),* to the reference, ROCS ranks the ensembles based on a TanimotoCombo score ranging from 0 (poor alignment) to 2 (best alignment). To evaluate this novel approach, ROCS was performed on the REMD trajectories and a TanimitoCombo score is provided for each structure in comparison with the GSK2194069 x-ray structure. From Dash analysis each structure from a REMD trajectory is also associated with a dash state. The structure belonging to a particular Dash states against their corresponding TanimotoCombo score is plotted in the correlation matrix in *Figure 6.10* with the colours corresponding to the number of structures

associated to a certain dash state and TanimitoCombo value (dark colours indicating higher number of structures). As no significant differences were observed between the starting conformations and between the two force fields only the result for the simulations of conformation 1 with the Amber Gaff force field are shown for each compound.

From *Figure 6.10*, for all the compounds no one Dash state is showing a higher TanimotoCombo score than any other. Additionally, the Dash states are ranked based on the population of structures belonging to the Dash state. Thus, from the distribution of the Dash states through the ROCS score, the best binding conformation will not necessary belong to the highest populated Dash states.

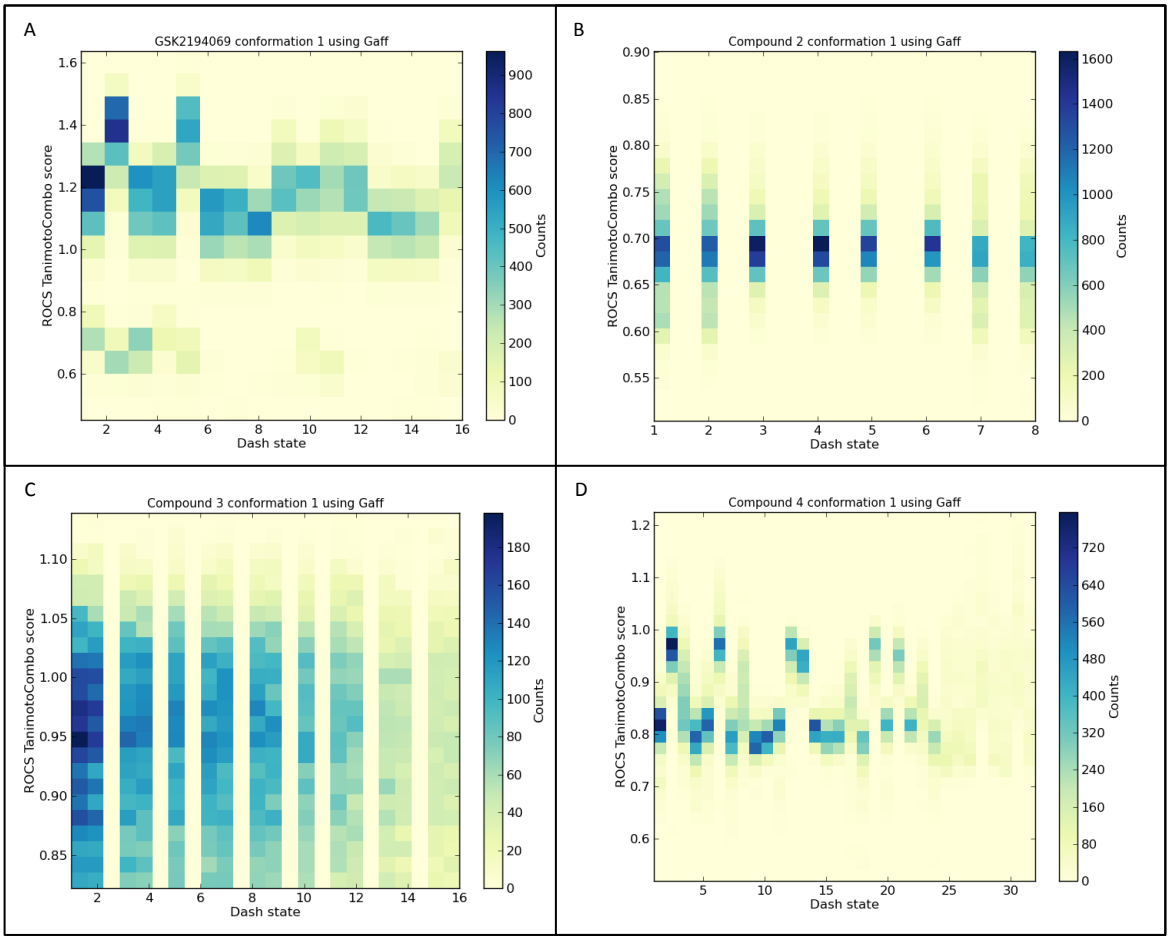


Figure 6.10: Correlation matrix of the Dash states allocated to ROCS TanimotoCombo score value for the REMD simulations of conformation 1 in the Amber Gaff force field for GSK2194069 (A), the compound 2 (B), the compound 3 (C) and the compound 4 (D).

Therefore, if the Dash states were ranked based on the best-superimposed structure with the x-ray, the structures with high TanimotoCombo score are expected to belong to the first Dash states. *Figure 6.10* clearly demonstrates that this is not the case. The Dash states are almost evenly spread through the ROCS TanimotoCombo scores so no correlation is observed and the relative population from the Dash state analysis cannot be used to look at how similar are the structures from the ensembles in term of overlap. Also, we cannot rely on using cluster representative structure from each Dash states to be representative of the cluster.

6.3.3 ROCS analysis:

6.3.3.1 GSK2194069:

Similar to the Dash/RMSD protocol, the overlay of the structures with high TanimotoCombo scores from the ROCS methodology with the x-ray structure for each ensemble were visualised *(Figure 6.11)*. Using ROCS enabled the analysis of the entire REMD trajectories starting from conformations A and B with the Amber Gaff and CGenFF Charmm force field; thus pre-filtering by clustering was not required.

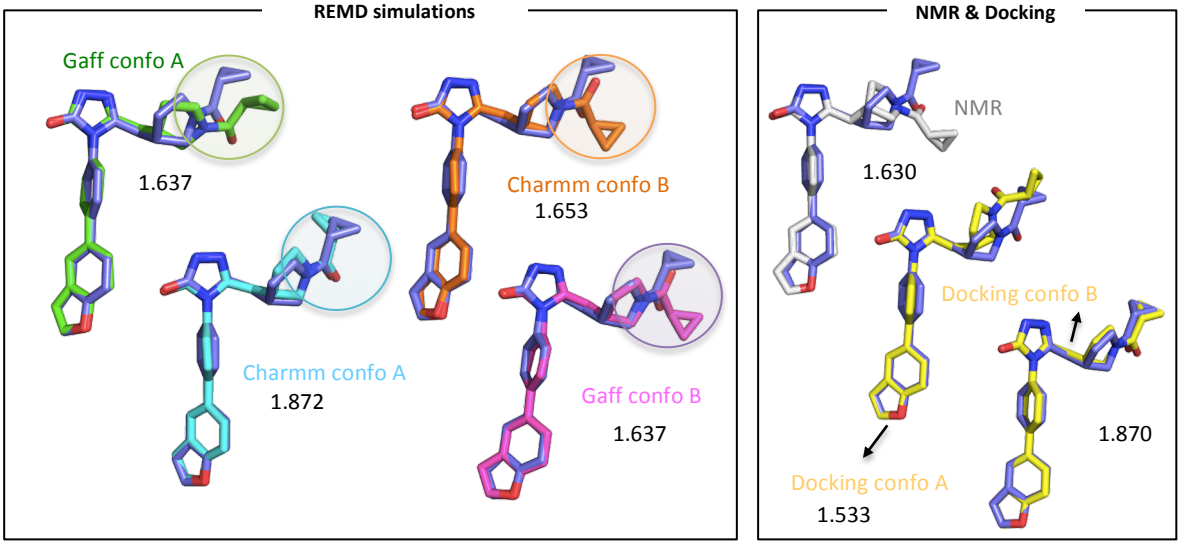


Figure 6.11: ROCS best-fitted structures superimposed with the x-ray structure of GSK2194069 (blue) for each method (REMD, NMR and Docking) with their corresponding TanimotoCombo score.

Good overlap is observed between the x-ray structure in blue and the best fitted structure from ROCS for each method (*Figure 6.11).* The shape as well as the key chemical groups such as the benzofuran, phenyl ring and triazolone moieties overlap well with the x-ray. However, the cyclopropyl group shows greater flexibility as highlighted with coloured circles and lead to lower TanimotoCombo score *(Figure 6.11 NMR, docking, conformation B with the two force fields and conformation A with Gaff)*

From each method, REMD, NMR and Docking, the overlap of structures from the final ensemble from Dash/RMSD protocol and the structures from *Figure 6.11* with the x-ray are compared (*Figure 6.12).* The ROCS score of the structures from Dash/RMSD filtering can also be found in *Figure 6.12.* The aim is to evaluate if discarding structures during the Dash/RMSD protocol induced a loss of good candidates for the binding modes.

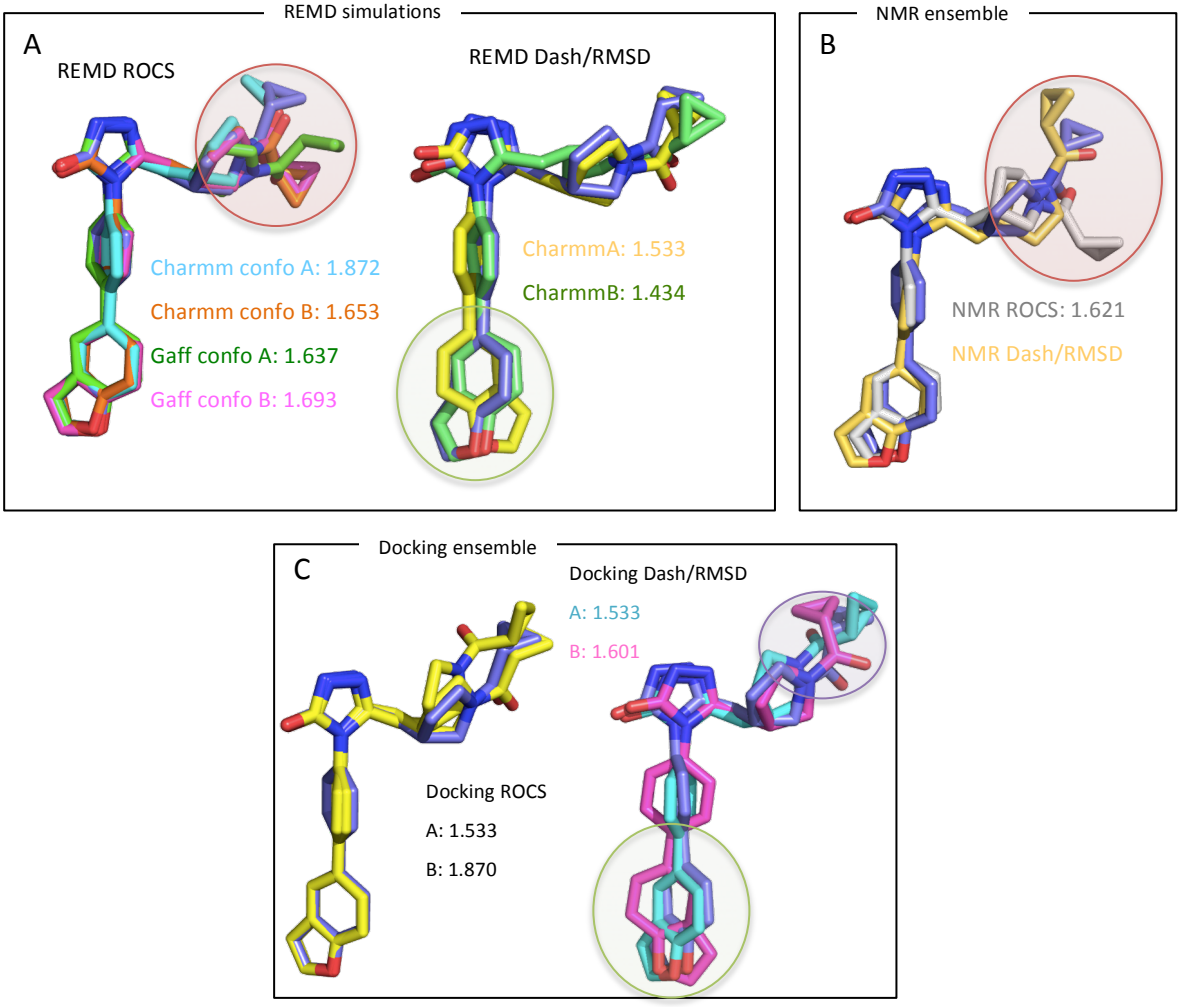


Figure 6.12: ROCS and Dash/RMSD protocol alignment for the various methods ensembles with their corresponding ROCS scores (REMD: A, NMR: B, and Docking: C). The x-ray structure is illustrated in blue

From the REMD simulations, an improvement of the superimposition is observed when using ROCS as seen with the lower ROCS score for the Dash/RMSD structures. Indeed, the green circle emphasised the large difference observed in the benzofuran and triazolone moieties when using Dash/RMSD that are not seen with ROCS data. However, the cyclopropyl group demonstrated greater variability with ROCS *(Figure 6.12A red circle).*

In both methodologies, Dash/RMSD and ROCS, the NMR ensemble was analysed completely. The ROCS alignment in grey shows more fluctuations of the cyclopropyl group *(Figure 6.12B red circle)* making the ROCS structure less favourable in terms of identifying the confirmed binding mode due to possible steric clashes with the receptor that will be discussed in the next chapter.

Figure 6.11C shows a comparison of Dash/RMSD and ROCS results for the docking ensemble. The benzofuran and cyclopropyl moieties of the structures from Dash/RMSD have shown poor overlap for the magenta structure with the x-ray structure, compared to ROCS data which find the x-ray binding mode from the same docking ensemble, as emphasized by the coloured circles. Additionally, both protocol (Dash/RMSD and ROCS) find identified the same structure from the docking (cyan) according to the ROCS score of 1.533.

Therefore, using ROCS here demonstrated better overlap, on average and enable the x-ray binding mode to be reproduced as observed for the REMD simulation of the conformation A with the Charmm force field and the docking simulation. However, the Dash/RMSD clustering has also demonstrated good overlap compared to ROCS *(Figure 6.12B)* and has also found the x-ray binding mode with larger differences than ROCS, so ROCS can appear better or worst in some cases but we cannot assume that it is a better methodology.

6.3.3.2 Compound 2:

Figure 6.13 shows the overlap of the highest scored structures from ROCS and the final ensemble structures from the Dash/RMSD clustering protocol to the x-ray in blue for each method (REMD: *Figure 6.13A*, NMR & Docking: *Figure 6.13B*) for compound 2. To evaluate the performance of ROCS versus Dash/RMSD, the regions of the compound showing differences in term of superposition have been emphasized in coloured circles.

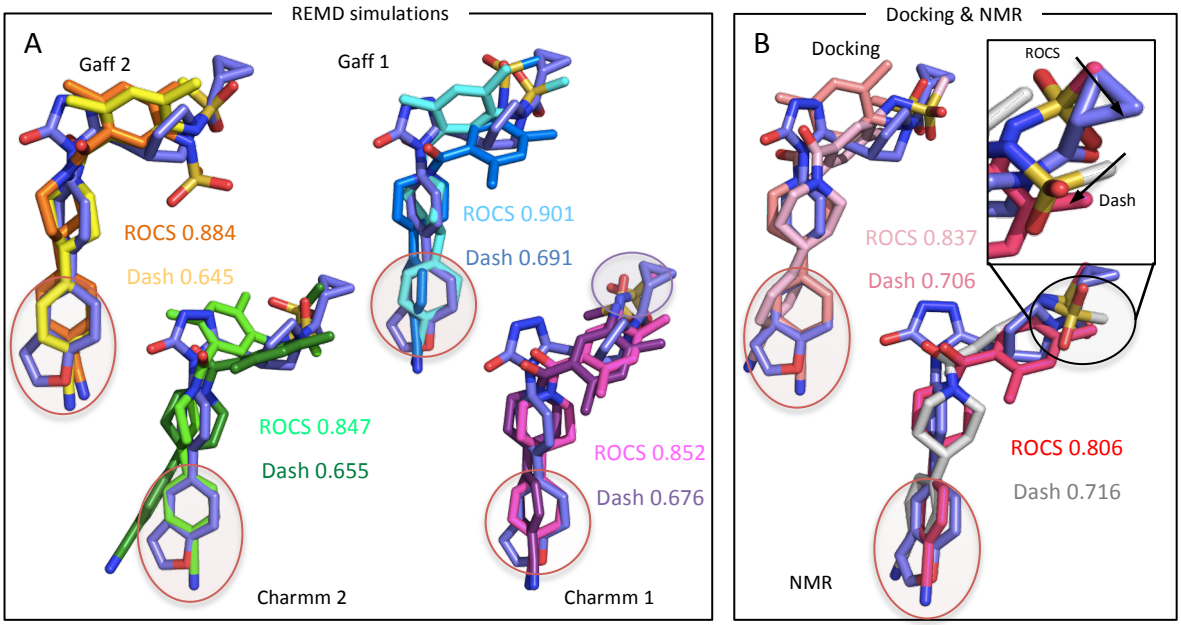


Figure 6.13: Compound 2 ROCS and Dash/RMSD comparison for the various ensembles and their corresponding ROCS scores: REMD (A), NMR and Docking (B). The x-ray structure is illustrated in blue.

In all three methods (REMD, NMR and docking), the overlay of the benzonitrile moiety with the benzofuran moiety of the x-ray structure demonstrates an improvement with the ROCS method as highlighted in red circles and with lower score values. The phenyl ring in the REMD simulations Charmm 2, Gaff 1 and Charmm 1 as well as the NMR ensemble is aligned perpendicular to the X-ray benzofuran moiety which may impact of the relative binding mode between the two molecules.

As for benzonitrile moiety, the overlap of the sulphonamide moiety onto the cylcopropyl moiety of the x-ray structure shows an improvement with the ROCS method. Indeed, the oxygen atom of the sulphonamide group is oriented in similar direction of the carbonyl of the cyclopropyl group, which is involved in a key interaction. *Figure 6.13B* shows an enlargement on the sulphonamide group overlay with the cyclopropyl group of the x-ray for the ROCS and Dash/RMSD alignment of the NMR ensemble. The black arrows show the direction of the oxygen atom of the sulphonamide moiety in both ROCS and Dash/RMSD with the x-ray structure in blue. The orientation of the oxygen atom with Dash/RMSD is away from the carbonyl of the cyclopropyl group of the x-ray that could induce the loss of the hydrogen bond made with the receptor.

However, knowing that compound 2 is a potent molecule, the regions of the compound that overlay differently to GSK2194069 could also be explained by a different binding mode of the compound 2.

6.3.3.3 Compound 3:

Figure 6.14A shows the result for the REMD simulations for the conformations 1 and 2 with the Amber Gaff and Charmm CGenFF force fields while the *Figure 6.14B* shows the NMR and docking results.

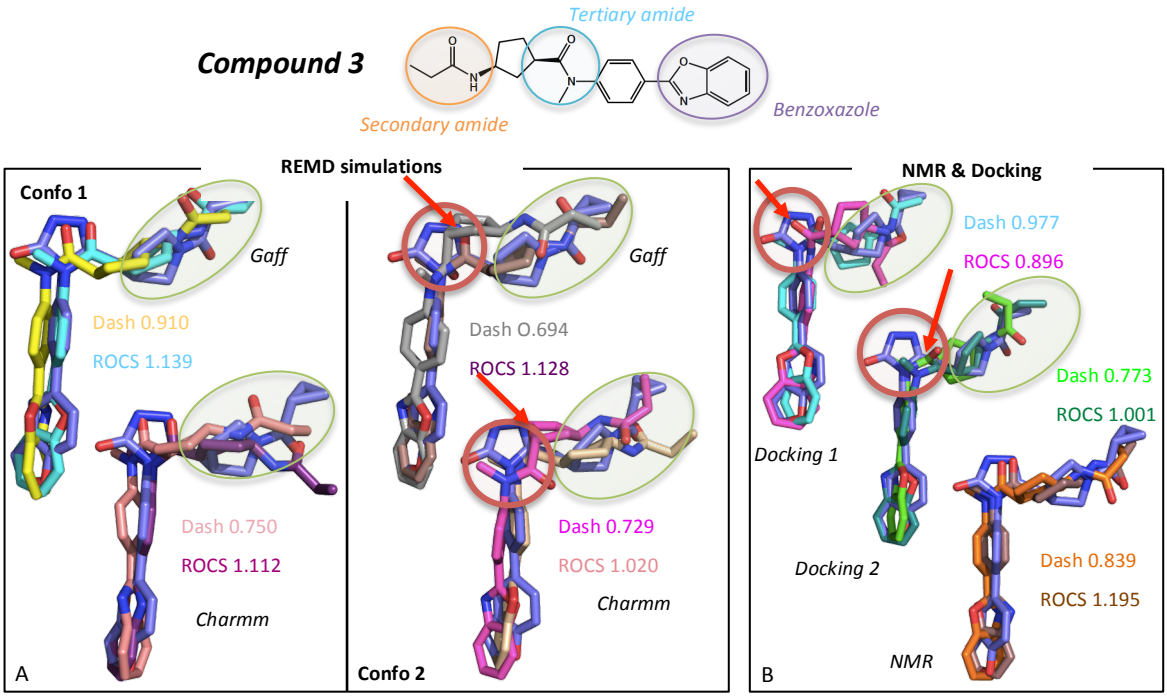


Figure 6.14: Compound 3 ROCS and Dash/RMSD comparison for the different REMD simulations (A) and for the NMR and Docking ensemble (B) with the x-ray structure in blue.

For all three methods, ROCS shows an improved alignment with the x-ray binding mode. In particular, the phenyl and the benzoxazole rings overlay well with the x-ray binding mode. In addition, the oxygen atom of the tertiary amide also shows an improved alignment with the triazolone moiety of the x-ray for the REMD simulations of conformation 2 and the docking runs (red arrows and circles). However, the results of the docking conformation 1 shows higher TanimotoCombo score value for the Dash/RMSD clustering. Thus, as for

GSK2194069 the ROCS results cannot assumed to be always better than the Dash/RMSD clustering.

In both protocol (ROCS and Dash/RMSD), the secondary amide part of compound 3 shows greater differences as highlighted in green circles. As seen with compound 2, the different orientation observed between compound 3 and the x-ray could be explained by a different biding mode. The contacts analysis in the following chapter will enable us to better understand these differences observed during the overlay.

6.3.3.4 Compound 4:

The results from the REMD simulations show improved alignment in term of molecular-shape using the ROCS method emphasized by higher ROCS scores. First, the benzofuran group of compound 4 shows fewer variations with the ROCS data, as well as the central core component, which has improved shape overlay with the x-ray phenyl ring *(Figure 6.15A)*. However, despite the good overlap alignment with ROCS, in the REMD simulations of conformations 1 with Gaff and conformation 2 with Charmm *(in red and green boxes)* the orientation of the cyclopropyl group is different to that in the x-ray structure, which may cause steric hindrance as well as loss of key contact with the receptor. On the contrary, the cyclopropyl group of the trajectories starting with conformation 1 with the Charmm force field and conformation 2 with the Amber Gaff force field demonstrated an improved overlap of this region of the molecule compared to the dash/RMSD clustering *(green and blue circles)*.

The NMR ensemble results are similar for both methodologies with good overlap of all the functional groups and high TanimotoCombo scores *(Figure 6.15B)*.

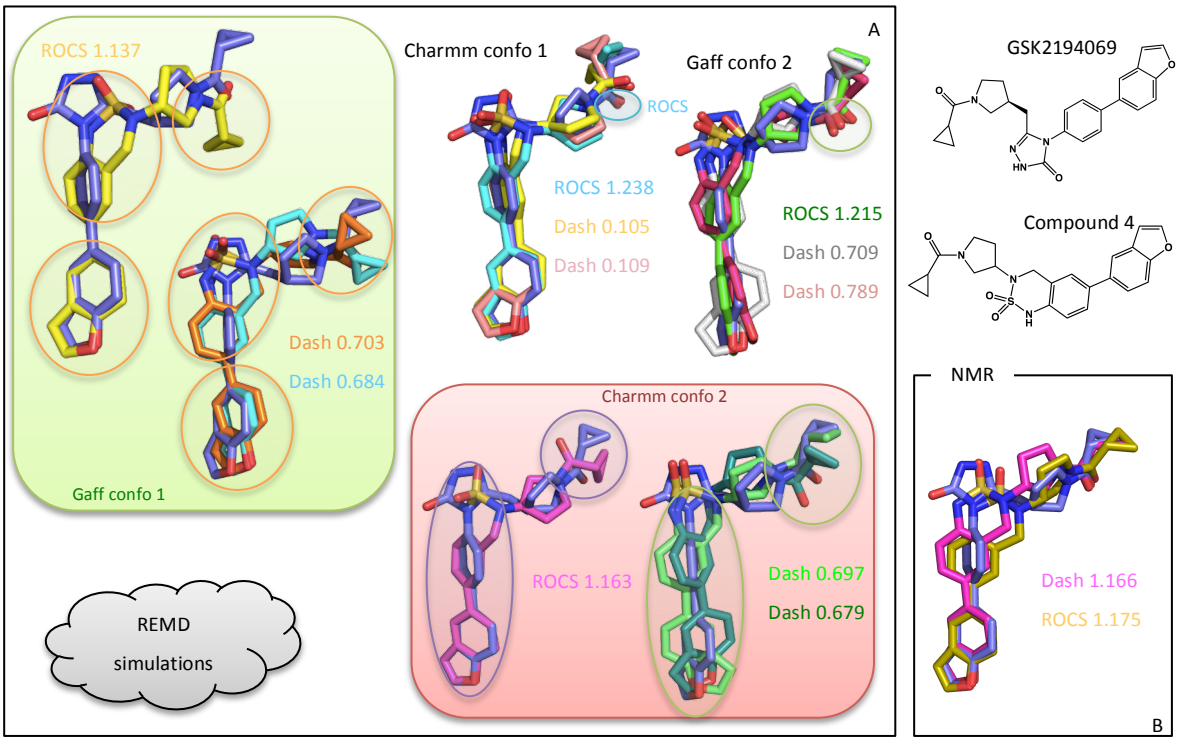


Figure 6.15: Compound 4 ROCS and Dash/RMSD results with their corresponding ROCS scores, for the different REMD simulations (A) and the NMR ensemble (B). The x-ray structure is illustrated in blue.

6.4 Conclusion:

Large trajectories require a fast, efficient and simple way to fully compare the conformational landscape of two molecules. The Dash/RMSD protocol, which involves Dash clustering by taking the cluster representative structures and superimposing each structure to the GSK2194069 x-ray molecule using an atom based RMSD tool, was used to enable us to identify the most populated conformations from solution to focus the overlays. From the analysis of GSK2194069 with the Dash/RMSD clustering, the x-ray binding mode was found but with larger differences than the x-ray *(Figure 6.5).* The analysis of compounds 2, 3 and 4 has revealed higher RMSD values for compound 2 compared to compounds 3 and 4, which could be related to the trapping observed with the distribution of torsion $\tau 3$ due to atropisomerism *(Figures 6.6 and 5.13).*

Furthermore, the sampling inconsistencies observed with the distributions of compound 3 torsion $\tau 5$ and compound 4 torsion $\tau 4$ between the NMR and REMD ensembles *(Appendix B Figure B.3 and B.6)* induced greater differences in the overlap between the compounds and the x-ray structure giving higher RMSD values.

Additionally, the Dash/RMSD clustering analyses have emphasised that the filtering process resulted in losing information by restricting the analysis to the cluster representative states. Also by choosing to use the x-ray structure as reference owing to the lack of crystal structures for compounds 2, 3 and 4, led to less favourable overlap, as has been shown in *section 3.1.1 Figure 6.4.* Thus, the use of a clustering method through Dash/RMSD filtering to identify a small number of most prevalent conformations did not appear successful *(Figure 6.10)*. Therefore, ROCS, a shape-based similarity analysis tool, was used with the advantage of being able to analyse the entire REMD ensembles without losing such information. The comparison between the two protocols *(Dash/RMSD and ROCS)* has shown improved superimposition when using ROCS for all the compounds *(Figures 6.12, 6.13, 6.14 and 6.15).* However, some cases where Dash/RMSD filtering performed better than ROCS were observed *(Figure 6.11 with the NMR structure and Figure 6.14 with the docking structure)* suggesting that the ROCS algorithm also led to approximations in terms of overlap. The combination of both protocols could enable us to better determine potential binding modes for compounds 2, 3 and 4.

To complete the analysis, the contacts made by the final ensemble structures from both Dash/RMSD and ROCS with hFAS will be evaluated and compared with the known binding mode from the x-ray structure of GSK2194069. Identifying potential steric clashes of the conformations in these simulations may enable likely binding modes to be identified for the novel inhibitors where the binding mode is not known.

Chapter 7 The exploitation of REMD simulations to identify potential binding modes of hFAS inhibitors

7.1 Introduction:

The protocol in chapter 6 was used to identify sets of conformations of the four hFAS inhibitors using two methodologies, Dash/RMSD filtering and ROCS, which were then compared against experimental data in the form of NMR and an x-ray structure. The publication of an x-ray structure of GSK2194069, (a potent and selective inhibitors of the KR domain of hFAS), enabled us to validate our approach using GSK2194069 and to potentially gain additional insights into the binding mode of compounds 2, 3 and 4 [62].

To evaluate the ability of these methods introduced in chapter 6 to confirm the binding mode of GSK2194069 and predict potential binding modes for compounds 2, 3 and 4, the interactions made with the receptor were studied. Hydrogen bonds, π and hydrophobic interactions will be thus identified and compared with those reported on analysis of the crystal structure. Good candidates for binding will thus be identified as well as a binding mode for each compound.

7.2 Receptor-ligand binding:

7.2.1 Methodology:

The analysis performed with the Dash/RMSD and ROCS protocols for each compound identified conformations, which demonstrated good overlap with the x-ray structure of GSK2194069, based on molecular shape and chemical groups alignment. However, some conformations demonstrated differences while superimposed to the x-ray structure, emphasizing the possibility of a different binding mode. To assess the potential of these conformations when binding to the x-ray structure, Molecular Operating Environment (MOE) software has been used [155]. MOE enables a user to visualise and understand details of the receptor active site and the interactions between the receptor and its ligand. Compared to other visualization tool such as Pymol and VMD, a ligand interaction diagram tool is available in MOE allowing a user to visualise the different interactions with the residues in close contact with the ligand. Hydrogen bonds,

salt bridges, hydrophobic interactions, π interactions, sulphur-LP, halogen bond and solvent exposure can easily be identified.

The overlaying procedure from *Chapter 6* allowed the compounds to be automatically centred to the position where the x-ray structure of GSK2194069 was in the binding pocket. Thus, when loading the receptor and the aligned structures from *Chapter 6* into MOE, the different conformations were automatically placed into the binding pocket. In this project, the ligand preorganization of the compounds is studied in term of how it affects the binding, thus, no energy minimization was performed after placing the structures into the pocket to maintain the information from the structures obtained in *Chapter 6*, at which point potential contacts were evaluated. Indeed, to evaluate if the solution phase conformations obtained from NMR and REMD ensembles bind favourably to the receptor binding pocket, the compound conformations as well as the receptor have not been relaxed.

Therefore, this protocol will enable us to identify conserved and new interactions made with hFAS and to evaluate the quality of each interaction as a good candidate for binding, to assess the possible binding mode for compounds 2, 3 and 4. However, we recognize that by not performing an exhaustive search and in keeping the ligand and receptor rigid, it is possible that better conformations could be identified if relaxation were allowed. This would, however, lose the conformational information associated with the solution phase ensemble.

7.2.2 Results:

The published paper of the x-ray structure of GSK2194069 reported three hydrogen bonds while bound to hFAS (*Section 1.4.2.2 Figure 1.8*) [62]. The carbonyl of the cyclopropyl moiety forms two hydrogen bonds with the residues serine 2021 and tyrosine 2034 (Ser2021, Tyr2034) and the oxygen atom of the triazolone core forms a hydrogen bond with serine 2081 (Ser2081). The benzofuran moiety also forms a π interaction with phenylalanine 2109 (Phe2109). Hydrophobic interactions were also observed around the triazolone and benzofuran moieties stabilising the whole system *(Chapter 1 section 1.4.2.1 Figures 1.7 and 1.8)*.

7.2.2.1 GSK2194069:

The interactions of the structures from the final ensemble of the Dash/ROCS protocol as well as the structures with the higher TanimotoCombo score from ROCS analysis were identified and compared with that found in the x-ray structure and can be found in *Table 7.1*. The percentage of the conserved interactions observed in the conformation and the ROCS scores for each structure are also indicated in *Table 7.1*. The percentage is based on the x-ray native contacts listed in the first row of the table. The cross-boxes correspond to the simulations where no structures were retained after Dash/RMSD analyses.

GSK2194069 structures	ROCS contacts	Dash/RMSD contacts	Native contact (%)
x-ray	Tyr2034H --- O Ser2021H --- O Ser2081H --- O Π: Phe2109 ring --- H		
Confo A Gaff	ROCS score : 1.637 Tyr2034H --- O Asn2028H --- O Ser2081H --- O Π: Phe2109 ring --- H		ROCS: 75 %
ConfoA Charmm	ROCS score : 1.872 Tyr2034H --- O Asn2028H --- O Ser2081H --- O Π: Phe2109 ring --- H	ROCS score : 1.533 Tyr2034H --- O Asn2028H --- O Ser2081H --- O	ROCS: 75 % Dash/RMSD: 50 %
ConfoB Gaff	ROCS score : 1.693 Asn2028H --- O Ser2081H --- O Π: Phe2109 ring --- H		ROCS: 50 %
ConfoB Charmm	ROCS score : 1.653 Asn2028H --- O Ser2081H --- O Π: Phe2109 ring --- H	ROCS score : 1.434 Tyr2034H --- O Gln2031O --- H pyrrolidine Π: Val2080H – triazolone ring Ser2081H --- O Π: Phe2109 ring --- H	ROCS: 50 % Dash/RMSD: 75 %
Docking confoA	ROCS score : 1.533 Gln2031H --- O Asn2028H --- O Ser2081H --- O Π: Phe2109 ring --- H	ROCS score : 1.533 Tyr2034H --- O Asn2028H --- O Ser2081H --- O Π: Phe2109 ring --- H Π: Arg2026H --- Benzofuran ring	ROCS: 50 % Dash/RMSD: 75 %
Docking confoB	ROCS score : 1.870 Tyr2034H --- O Asn2028H --- O Ser2081H --- O Π: Phe2109 ring --- H	ROCS score : 1.601 Tyr2034H --- O Asn2028H --- N triazolone ring Ser2081H --- O	ROCS: 75 % Dash/RMSD: 50%
NMR ensemble	ROCS score : 1.621 Tyr2034H --- O Ser2081H --- O Π: Val2080H --- triazolone ring	Ser2081H --- O	ROCS: 50 % Dash/RMSD: 25 %

Table 7.1: Contacts made between GSK2194069 final ensemble structures from ROCS and Dash/RMSD protocol and the hFAS receptor as well as the percentage of native contacts observed. The ROCS score for each conformation is highlighted in yellow at the top of the contacts list.

Table 7.1 shows first that none of the structures in either protocol (Dash/RMSD and ROCS) were able to find the hydrogen bond involving the residue serine 2012, whereas the hydrogen bond involving the residue serine 2081 is found in all the structures. Additionally, the hydrogen bond between the

triazolone and asparagine 2028 that was not found in the x-ray is present in all the structures except in the NMR ensemble, which could be an important bond to maintain the stability of the complex. Furthermore, the hydrogen bond and the π interaction involving residues tyrosine 2034 and phenylalanine 2109 are well represented in both ROCS and Dash/RMSD filtering methodologies.

From the overlay analysis in *Chapter 6*, two conformations were particularly able to show a good overlap with the x-ray structure corresponding to the structures using ROCS from the REMD simulation of conformation A with the CGenFF Charmm force field and the docking simulation of conformation B *(in bold in Table 7.1)*. These conformations have conserved 75 % of the native contacts.

In term of hydrophobic interactions, the structures using the ROCS protocol from the REMD simulation of conformation A with the CGenFF Charmm force field and the docking simulation of conformation B were studied, corresponding to structures that were able to reproduce the binding mode. *Figure 7.1* shows the ligand diagrams from MOE analysis. From these diagrams, the solvent exposure regions observed with the x-ray structure are conserved *(Figure 7.1 red circles and Figure 1.8).* Additionally, the cyclopropyl moiety also shows solvent exposed surface.

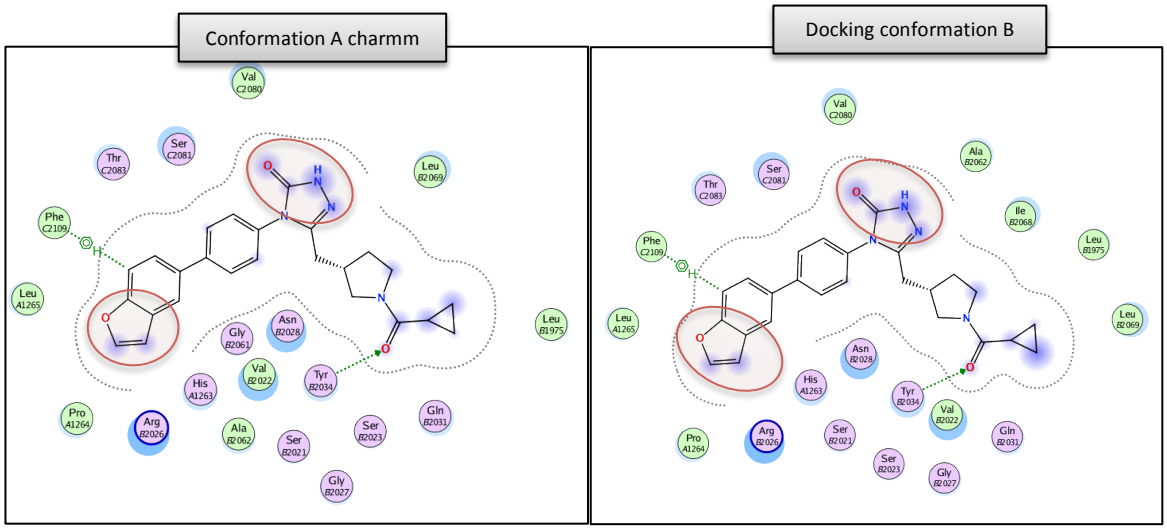


Figure 7.1: Ligand interactions diagram of conformation A in the CGenff Charmm force field and the structure from the docking simulation of conformation B using the ROCS protocol. The green arrows represent contacts made between the ligand and the receptor.

Therefore, the contacts analysis has shown the ability of the methods to largely reproduce the GSK2194069 binding mode and to be able to conserve a

high percentage of native contacts. However, the *Figure 7.2* the distribution of the TanimotoCombo ROCS score of the REMD simulation of conformation A with the Charmm CGenFF force field which was similar to the x-ray structure in term of binding mode *(Figure 7.1)* in comparison to the NMR ensemble. The conformations with high TanimotoCombo scores that are similar to the x-ray structure are low populated *(Figure 7.2 red asterisks)*.

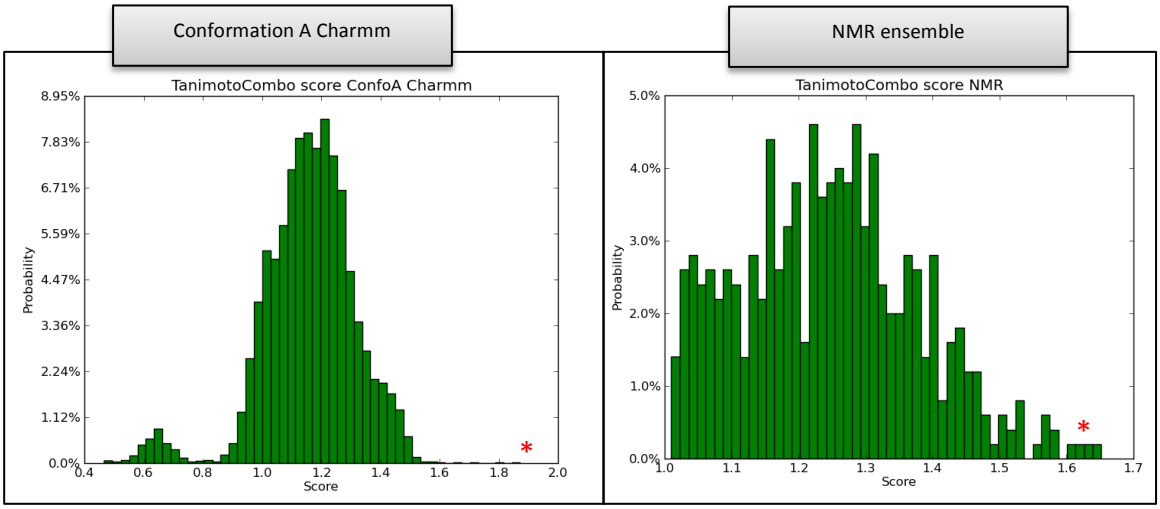


Figure 7.2: Histogram distribution of the TanimotoCombo score in the REMD simulation of the conformation A with the Charmm CGenFF force field and the NMR ensemble. The red asterisks indicate the position of the high TanimotoCombo score conformations that are similar to the x-ray structure.

7.2.2.2 Compound 2:

The contacts made between the structures from the final ensemble of Dash/RMSD filtering and the structure with the highest TanimotoCombo score from ROCS and the percentage of similar contact with the x-ray binding mode is shown in *Table 7.2*. The ROCS score for each structure is highlighted in yellow in *Table 7.2*.

Compound 2 structures	ROCS contacts	Dash/RMSD contacts	Native contact (%)
x-ray	Tyr2034H --- O Ser2021H --- O Ser2081H --- O Π: Phe2109 ring --- H		
Confo1 Gaff	ROCS score: 0.901 Tyr2034H --- O Gln2031H --- O Arg2026H --- H	ROCS score: 0.691 Gln2031H --- O Gln2031O --- HN Ser2023O --- HN	ROCS: 25 % Dash/RMSD: 0 %
Confo1 Charmm	ROCS score: 0.852 Tyr2034H --- O Gln2031H --- O Π: Phe2109 ring --- H	ROCS score: 0.676 Tyr2034H --- O Gln2031H --- O His1263H --- N	ROCS: 50 % Dash/RMSD: 25 %
Confo2 Gaff	ROCS score: 0.884 Tyr2034H --- O Gln2031H --- O Ser2081H --- O	ROCS score: 0.645 Ser2023O --- H	ROCS: 50 % Dash/RMSD: 0 %
Confo2 Charmm	ROCS score: 0.847 Tyr2034H --- O Π: Phe2109 ring --- H	ROCS score: 0.655 Gln2031H --- O Leu1265H --- N	ROCS: 50 % Dash/RMSD: 0 %
Docking confo1	ROCS score: 0.837 Tyr2034H --- O Π: Phe2109 ring --- H	None	ROCS: 50 % Dash/RMSD: 0 %
NMR ensemble	ROCS score: 0.806 Tyr2034H --- O	None	ROCS: 25 % Dash/RMSD: 0 %

Table 7.2: Contacts made between compound 2 final ensemble structures from ROCS and Dash/RMSD protocol and the hFAS receptor as well as the percentage of native contacts observed. The ROCS score for each conformation is highlighted in yellow at the top of the contacts list.

The structures obtained from the ROCS analysis show a high percentage of similar interactions with the x-ray; all of the structures from ROCS have identified native contacts in comparison to the structures from Dash/RMSD where only one structure was able to reproduce a native contact. Additionally, the analyses of the structure overlap from chapter 6 have shown greater variability with the structures obtained from Dash/RMSD in terms of molecular shape alignment. The fluctuations observed seems to influence the stability of the structures with the receptor by forming an extra hydrogen bond with the nitrile moiety and the hydrogen atom of residue serine 2023 for conformation 1 in the Amber Gaff force field, residue histidine 1263 for conformation 1 in the CGenFF Charmm force field and residue leucine 1265 for conformation 2 in the CGenFF Charmm force field. Furthermore, the REMD simulation of conformation 2 with the Amber Gaff force field from the ROCS protocol, forms extra hydrogen bond with the oxygen atom of the amide moiety and residue serine 2081, corresponding to a

native contact, emphasizing the possibility of increasing the tightness of the binding in future drug development. *Figure 7.3* illustrates which parts of the compound 2 structure are involved in the different contacts from *Table 7.2.*

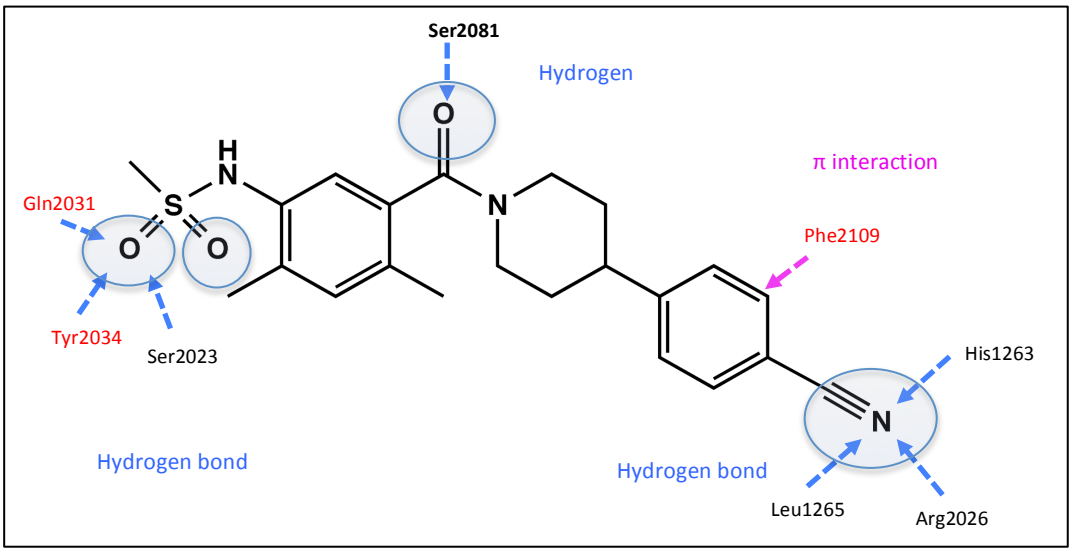


Figure 7.3: The different contacts formed by the structures of the Dash/RMSD and ROCS protocols of compound 2 with the residues hFAS receptor. The residues highlighted in red emphasize the most prevalent contacts.

Looking at the differences within the structures from ROCS and Dash/RMSD while bound to the receptor in terms of steric hindrance, the concern that the fluctuations observed for the structures from the Dash/RMSD protocol may suffer steric clashes is confirmed. Two examples are shown in *Figure 7.4* with the structures from the REMD simulation of conformation 2 using the CGenFF Charmm force field and from the docking simulation. The hFAS receptor is shown by dashed line around the compounds and the steric clashes are highlighted in red circles. The corresponding alignment of the structures with the x-ray structure of GSK2194069 from ROCS and Dash/RMSD protocols is illustrated at the bottom of the ligand diagrams.

Therefore, the orientation of the benzonitrile moiety appears to play an important role in the compound 2 binding mode as it could cause major clashes with the receptor.

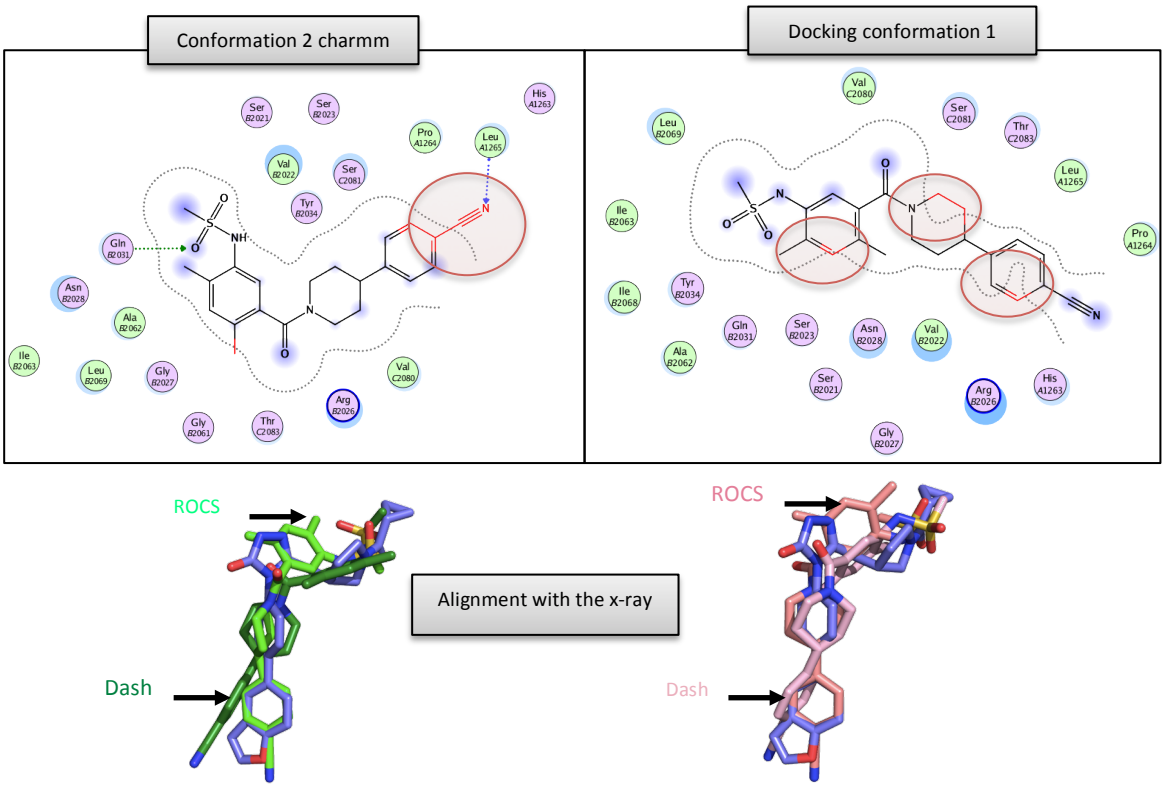


Figure 7.4: Ligand diagrams of conformation 2 using the CGenFF Charmm force field and the docking simulation of conformation 1 from the Dash/RMSD clustering. The alignment of the corresponding structure with the GSK2104069 x-ray is shown under each diagram.

Furthermore, the analysis of the solvent exposure has identified solvent exposure in the methyl groups in the ROCS structures and in the sulphonamide moiety in the Dash/RMSD structures. However, the methyl and sulphonamide moieties are hydrophobic so having solvent exposure in these groups will not improve the binding affinity of the structures. The benzene and the methyl groups may favour the conformational preorganization required for binding in which case their solvent exposure would be compensated by having a molecule better preorganized and thus with higher affinity. The hydrophobic properties of the sulphonamide have been studied in several papers [185-187].

Therefore, from the analysis of the different ensemble overlaps with the x-ray structure of GSK2194069 using Dash/RMSD and ROCS protocols followed by the identification of the interactions and the possible clashes with the receptor, a potential binding mode for compound 2 can be determined. Three main contacts appear to be important in the binding involving the sulphonamide and benzonitrile moieties and the hFAS residues tyrosine 2034, glutamine 2031 and

the phenylalanine 2109. The orientation of the structure is more likely to be similar to the one observed from the ROCS overlay of the REMD simulation of conformation 1 with the CGenFF Charmm force field with the x-ray. A predicted binding mode for compound 2 is shown in *Figure 7.5.*

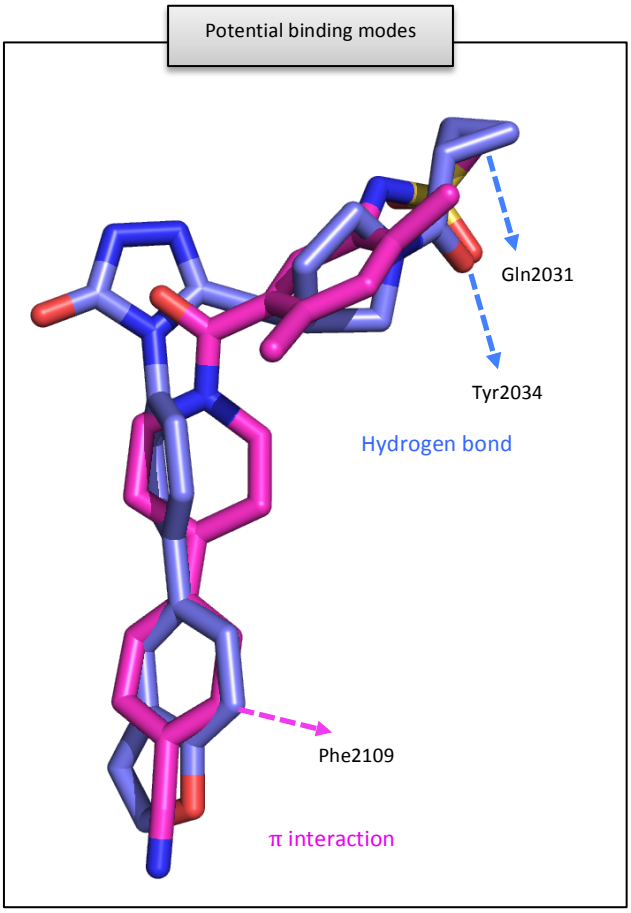


Figure 7.5: Predicted binding mode of compound 2 with the x-ray in blue

However, an exhaustive search of the binding pocket has not been performed here. Therefore, other binding modes may be available after relaxation as the *Figure 7.6* shows with the distribution of the TanimotoCombo ROCS score of the REMD simulation of conformation 1 with the Charmm CGenFF force field considered as potentially good candidates for binding *(Figure 7.5)* in comparison to the NMR ensemble. The conformations with high scores similar as the one in *Table 7.2* in the ensembles are low populated *(Figure 7.6 red and blue asterisks)*.

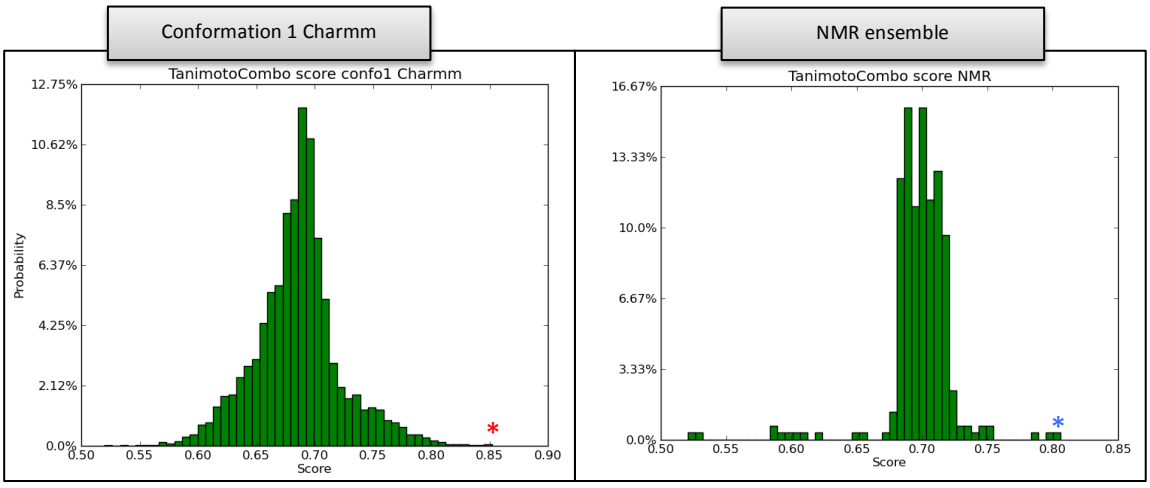


Figure 7.6: Histogram distribution of the TanimotoCombo score in the REMD simulation of the conformation 1 with the Charmm CGenFF force field and the NMR ensemble. The red asterisk indicates the position of the potential candidate for the binding from *Figure 7.5* while the blue asterisk indicates high TanimotoCombo score in the NMR ensemble.

7.2.2.3 Compound 3:

The different contacts made between the structures from the final analysis of Dash/RMSD and ROCS protocols for compound 3 can be found in *Table 7.3*.

Compound 3 structures	ROCS contacts	Dash/RMSD contacts	Native contacts (%)
x-ray	Tyr2034H --- O Ser2021H --- O Ser2081H --- O Π: Phe2109 ring --- H		
Confo1 Gaff	ROCS score:1.139 Asn2028H --- O Π: Arg2026H --- benzoxazole ring Π: Phe2109 ring --- H	ROCS score: 0.910 Ser2081O --- HN	ROCS: 25 % Dash/RMSD: 25 %
Confo1 Charmm	ROCS score: 1.112 Π: Ser2081H --- phenyl ring Asn2028H --- O	None	ROCS: 50 % Dash/RMSD: 0 %
Confo2 Gaff	ROCS score: 1.128 Asn2028O --- HN Ser2081O --- HN	ROCS score: 0.694 Π: Arg2026H --- benzoxazole ring	ROCS: 25 % Dash/RMSD: 0 %
Confo2 Charmm	ROCS score: 1.020 Π: Ser2081H --- phenyl ring	ROCS score: 0.729 Π: Arg2026H --- benzoxazole ring	ROCS: 25 % Dash/RMSD: 0 %
Docking confo1	ROCS score: 0.896 Tyr2034H --- O	None	ROCS: 25 % Dash/RMSD: 0 %
Docking confo2	ROCS score: 1.001	None	ROCS: 0 %

	Π: Arg2026H --- benzoxazole ring		Dash/RMSD: 0 %
NMR ensemble	ROCS score: 1.195 Tyr2034H --- O	ROCS score: 0.839 Ser2081O --- HN	ROCS: 25 % Dash/RMSD: 25 %

Table 7.3: Contacts made between compound 3 final ensemble structures from ROCS and Dash/RMSD protocol and the hFAS receptor as well as the percentage of native contacts observed. The ROCS score for each conformation is highlighted in yellow at the top of the contacts list.

From the structure overlap analysis, the structures from the ROCS protocol show better overlap in terms of molecular shape alignment with the x-ray. However, greater variations than the x-ray structure of the secondary amide region were observed for conformations 1 and 2 with the CGenFF force field and the docking simulation of conformation 1 from ROCS protocol *(Figure 6.14). Table 7.3* demonstrates that the structures from ROCS show a higher number of interactions with the receptor that could increase the stability of the compound inside the binding pocket. Comparing to compound 2, the relative percentages of native contacts reproduced by the structures of compound 3 are lower. However, as for compound 2, the contacts observed in the x-ray involving the residues tyrosine 2034 and phenylalanine 2109 are seen here. To evaluate how these different interactions from *Table 7.3* help the stability of the structure in the binding pocket, *Figure 7.7* illustrates the interactions of the residues involved in contact with the functional groups of compound 3.

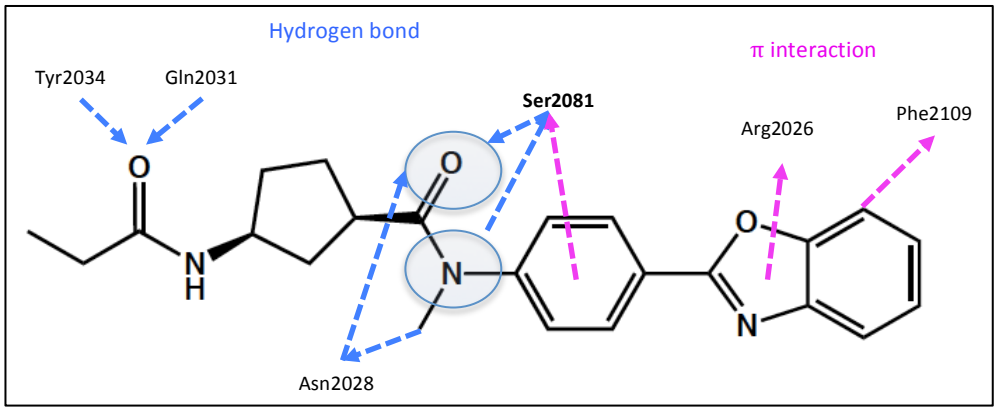


Figure 7.7: The different contacts formed by the structures of the Dash/RMSD and ROCS protocols of compound 3 with the residues hFAS receptor

Therefore, *Figure 7.7* emphasizes that an ideal conformation would have at least three main interactions with the receptor allowing the stabilisation of the structure into the binding pocket involving the two amide and the benzoxazole moieties. The REMD simulation of conformation 1 with the Charmm force field demonstrated the identification of these three main interactions, however, the higher variation of the secondary amide observed during the overlay analysis induces a steric clash when the conformation as inserted into the binding pocket *(Figure 7.8 red circle)*.

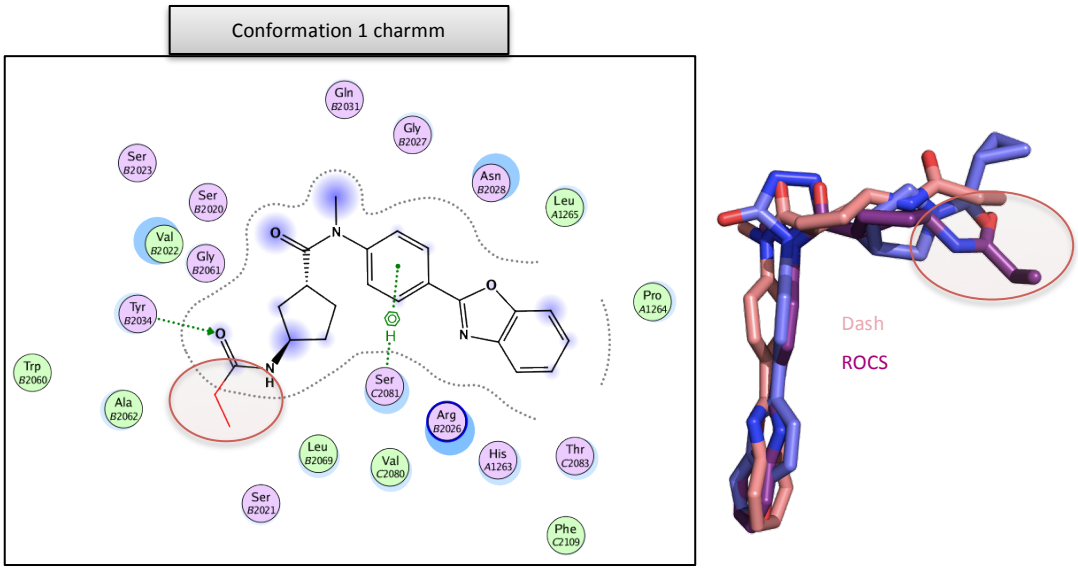


Figure 7.8: Ligand diagram of the conformation 1 using the CGenFF Charmm force field. The alignment of the corresponding structure with the x-ray is shown next to the diagram. The red circle emphasized the area where the compound induces a steric clash with the receptor.

Additionally, a look at the solvent exposure of the different structures from the ROCS protocol shows high solvent exposure on the hydrogen atom of the tertiary amide moiety regions that could improve the binding if the carbonyl oxygen of the tertiary amide moiety is not able to form a hydrogen bond.

Therefore, from a combination of the structure alignments and contact analyses from the REMD simulations of conformation 1 using the Amber Gaff force field, the docking simulation of conformation 2 and the NMR with the ROCS protocol, two potential binding modes for compound 3 can be identified. The analysis of the possible major steric clashes with the receptor has revealed that a preferred orientation of the secondary amide moiety similar to the cyclopropyl

group of the x-ray will prevent clashes as observed for the three structures mentioned above. Furthermore the orientation of the carbonyl oxygen of the secondary amide in the NMR and the docking structures also offers extra stability with the hydrogen bond involving residues tyrosine 2034 and glutamine 2031. Thus, ideal binding modes will appear with conformations having the orientation of the benzoxazole group of the REMD simulation of conformation 1 with the Amber Gaff force field, enabling two π interactions, and the orientation of the secondary amide moiety of the NMR or the docking conformation. *Figure 7.8* illustrates the two potential binding modes of compound 3.

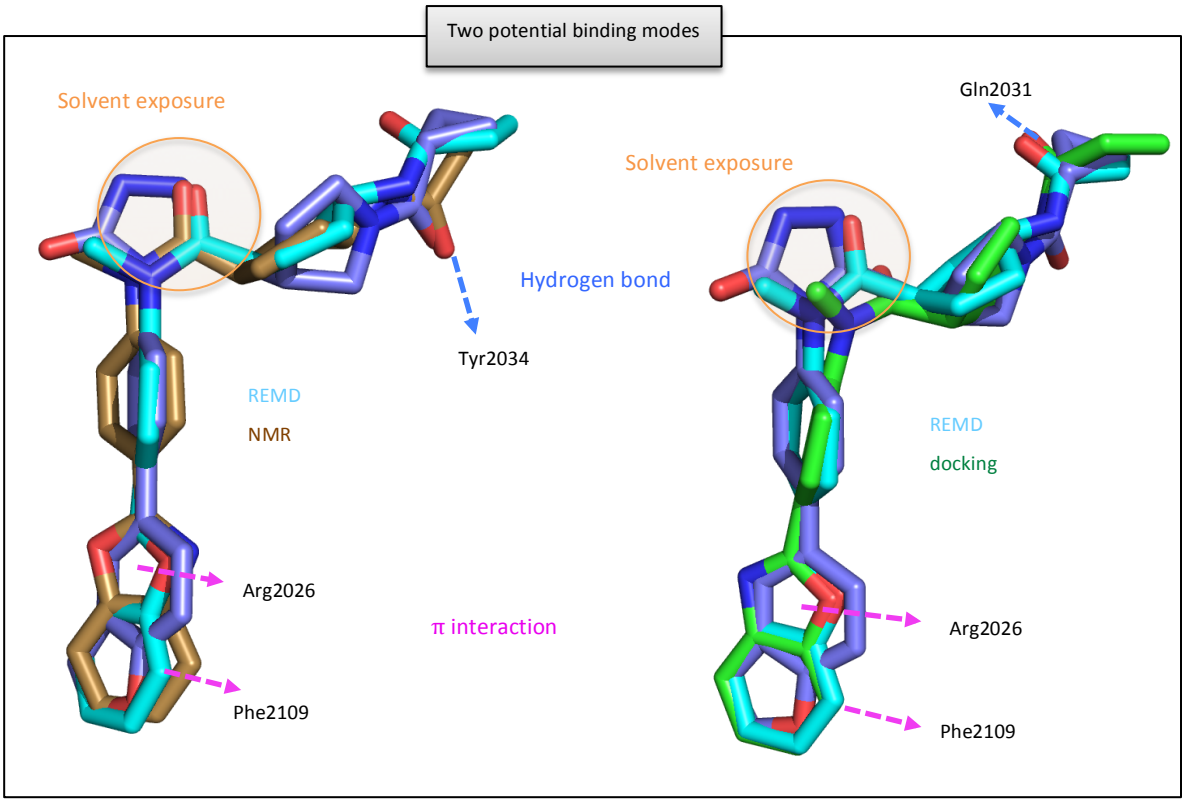


Figure 7.8: Predicted binding modes for compound 3 with the x-ray in blue.

Therefore, as for compound 2, by not allowing the compound and the receptor to relax, conformations that could possess the combined features of the predicted binding modes illustrated in *Figure 7.8* could be identified. As for compound 2, the *Figure 7.9* that shows the distribution of the TanimotoCombo ROCS score of the REMD simulation of conformation 1 with the Amber Gaff force field and the NMR ensemble, considered as potentially good candidates for binding *(Figure 7.8)* emphasised that the conformations with high scores such as these two structures are low populated *(Figure 7.9 red asterisks)*.

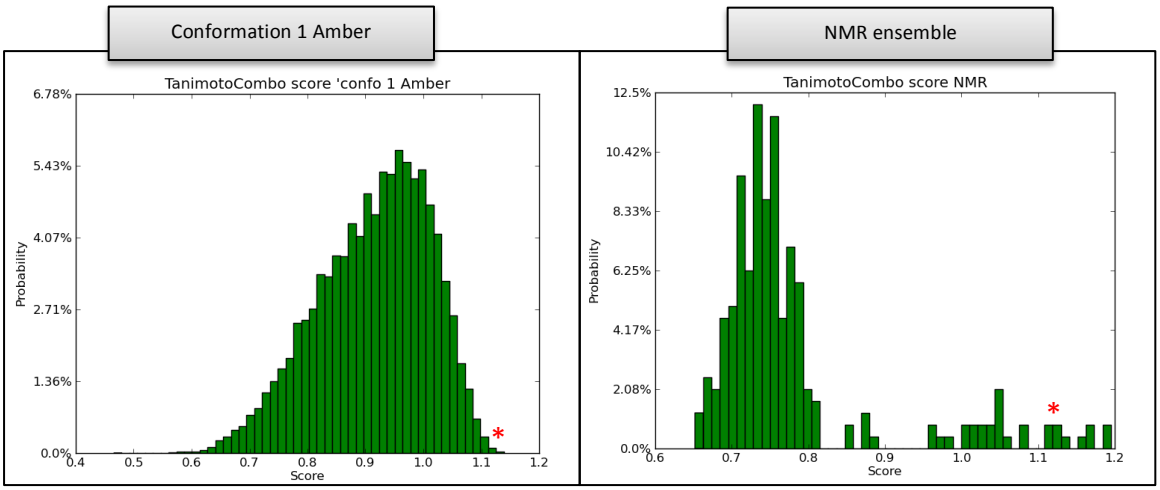


Figure 7.9: Histogram distribution of the TanimotoCombo score in the REMD simulation of the conformation 1 with the Amber Gaff force field and the NMR ensemble. The red asterisks indicate the position of the two potential candidates for the binding from *Figure 7.8*.

7.2.2.4 Compound 4:

The interactions of the final structures from the Dash/RMSD and ROCS protocols for the study of compound 4 are shown in *Table 7.4*. During the Dash/RMSD analysis of the REMD simulations, more than one structure was kept for the final ensemble and are indicated in the table based on the colours used during the overlay.

Compound 4 structures	ROCS contacts	Dash/RMSD contacts	Native contacts (%)
x-ray	Tyr2034H --- O Ser2021H --- O Ser2081H --- O Π: Phe2109 ring --- H		
Confo1 Gaff	ROCS score: 1.137 Tyr2034H --- O Asn2028O --- HN Π: Phe2109 ring --- H	**Cyan:** 0.684 Tyr2034H --- O GLN2031O --- H Asn2028H --- OS **Orange:** 0.703 Asn2028O --- HN Asn2028H --- OS Π: Phe2109 ring --- H	ROCS: 50 % Dash/RMSD: 25 %
Confo1 Charmm	ROCS score: 1.238 Tyr2034H --- O Asn2028H --- OS	**Pink:** 1.109 Tyr2034H --- O Asn2028H --- OS **Yellow:** 1.105 Asn2028H --- OS Π: His1263 ring --- H Π: Phe2109 ring --- H	ROCS: 25 % Dash/RMSD: 25 %
	ROCS score: 1.215 Tyr2034H --- O	**Grey:** 0.709 Asn2028H --- OS	ROCS: 50 %

Confo2 Gaff	Asn2028O --- HN Ser2023H --- OS Π: Phe2109 ring --- H	Ser2081O --- H **Dark pink:** 0.789 Asn2028H --- OS	Dash/RMSD: 25 % Dash/RMSD: 0 %
Confo2 Charmm	ROCS score:1.163 Asn2028H --- OS Π: Phe2109 ring --- H	**Light green:** 0.697 Tyr2034H --- O **Dark green:** 0.679 Ser2023O --- H	ROCS: 25 % Dash/RMSD: 25 % Dash/RMSD: 0 %
NMR ensemble	ROCS score:1 .175 Tyr2034H --- O Ser2023H --- OS Π: Phe2109 ring --- H	ROCS score:1 .166 Tyr2034H --- O Π: Phe2109 ring --- H	ROCS: 50 % Dash/RMSD: 50 %

Table 7.4: Contacts made between compound 4 final ensemble structures from ROCS and Dash/RMSD protocols and the hFAS receptor as well as the population of the conserved contacts.

The structures overlay analyses from chapter 6 have emphasized larger variations of the benzofuran moiety for the structures from the Dash/RMSD analysis for the REMD ensembles with the NMR structures ensemble from both Dash/RMSD filtering and ROCS protocols have shown equivalent results *(see Figure 6.15).* Additionally, the REMD simulation of conformation 1 with the Amber Gaff force field has also emphasized greater differences of the cyclopropyl moiety in both protocols *(Dash/RMSD and ROCS)*.

As expected from the similarity between compound 4 and GSK2194069, the native contacts involving the residues tyrosine 2034 and phenylalanine 2109 are highly conserved especially with the structures from the ROCS protocol due to the stability of the overlay of the benzofuran moiety *(see Figure 6.15).* Furthermore, from both ROCS and Dash/RMSD protocols, the sulphonamide moiety of the central core of compound 4 forms hydrogen bonds with the residues asparagine 2028, serine 2023 enabling stabilisation of the molecule.

In term of the concern of steric clashes due to the greater variability of the benzofuran moiety for the structures with the Dash/RMSD protocol and the fluctuation of the cyclopropyl moiety of the structure from REMD simulation of conformation 1 with the Amber Gaff force field with the ROCS protocol, the study of the insertion into the binding pocket confirms major clashes with the receptor for these conformations as seen in *Figure 7.10*. Allowing structural relaxation of the ligand will likely remove thus clashes, but this will be associated with an enthalpic and entropic cost which ligand preorganisation is trying to avoid.

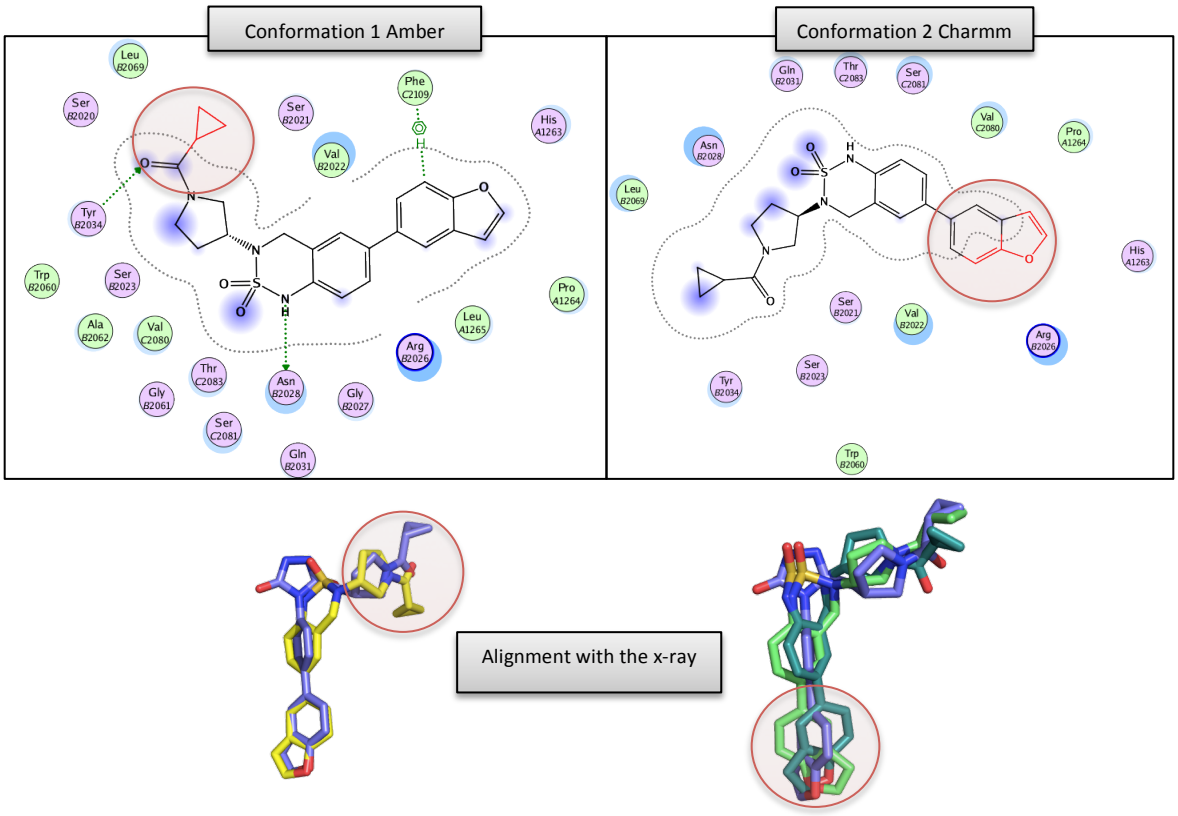


Figure 7.10: Ligand diagrams of conformation 1 and using the Amber Gaff and CGenFF Charm force fields with ROCS and Dash/RMSD protocols respectively. The alignment of the corresponding structure with the x-ray is shown under each diagram. The red circles indicate the steric clash between the conformations and the receptor.

The solvent exposed regions observed in the x-ray structure of GSK2194069 are also seen in the cyclopropyl moiety but also in the oxygen atoms of the central core group of compounds 4 for both protocols (ROCS and Dash/RMSD). The central core refers to the region of the GSK2194069 structure, which has been modified by scaffold hop to generate compound 4 *(Appendix C Figure C.2)*. The exposed solvent observed in this region may induce a cost in the binding affinity by the character hydrophobic of the sulphonamide moiety.

Therefore, the study of compound 4, a "scaffold-hop" of GSK2194069, shows that despite a lower potency the compound demonstrates good binding with the hFAS receptor in term of contacts made with receptor. Potential binding modes can be generated from the results of the structures of the ROCS data. The simulation of conformation 2 with the Gaff force field as well as the NMR structure appear as good candidates for binding and will enable us to generate potential binding modes *(Figure 7.11)*.

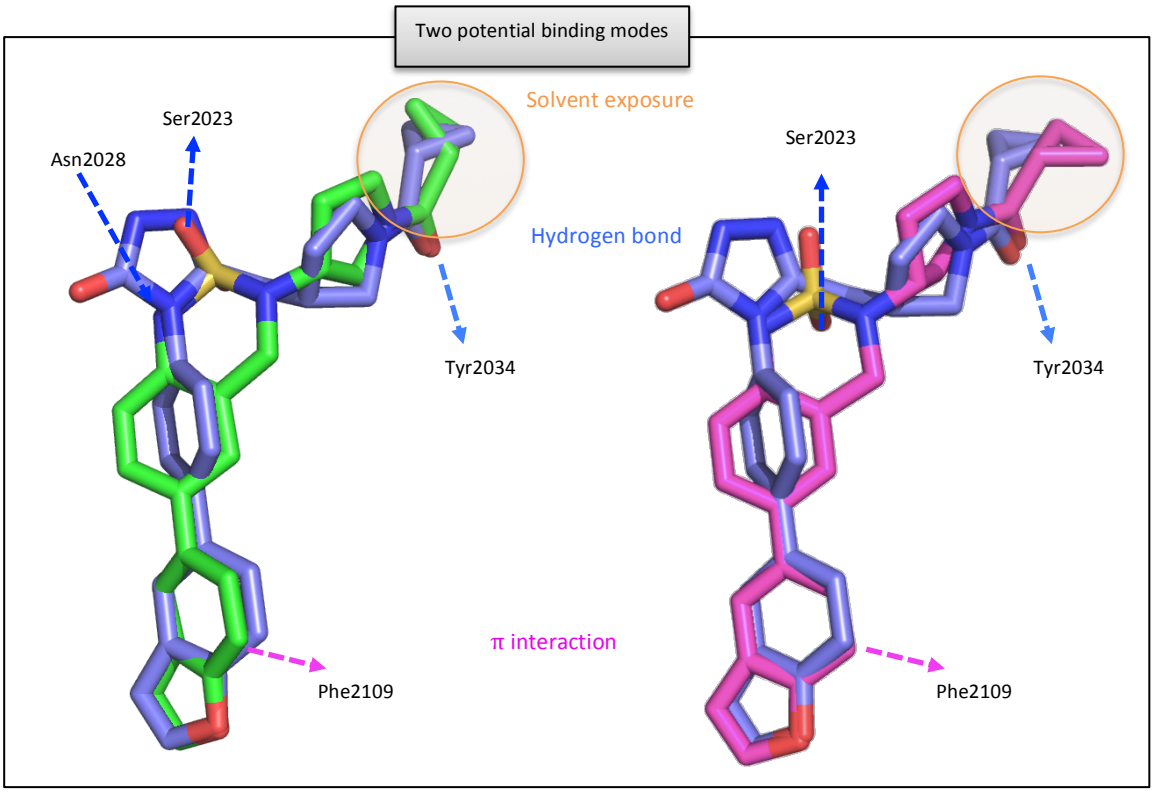


Figure 7.11: Predicted binding modes for compound 4 with the x-ray in blue.

The causes of a lower potency for compound 4 can be explained by different reason. First, despite good interactions observed with the receptor, the exposure of the sulphonamide oxygen atoms to the solvent could influence the ligand potency. To continue in this way, in this analysis the receptor-ligand solvation has not been taken into account, while solvent actively participates in binding of course. Second, owing the flexibility of the compounds, the Dash clustering tool was not able to provide relative populations and thus the conformations that are considered as good potential candidates for binding *(Figure 7.11)* could be comparatively rare compared to the whole ensemble. Furthermore, *Figure 7.12* shows the distribution of the TanimotoCombo ROCS score of the REMD simulation of conformation 2 with the Amber Gaff force field and the NMR ensemble, considered as potentially good candidates for binding *(Figure 7.11)*; the conformations with high scores such as these two structures, in the ensembles are low populated *(Figure 7.12 red asterisks)*.

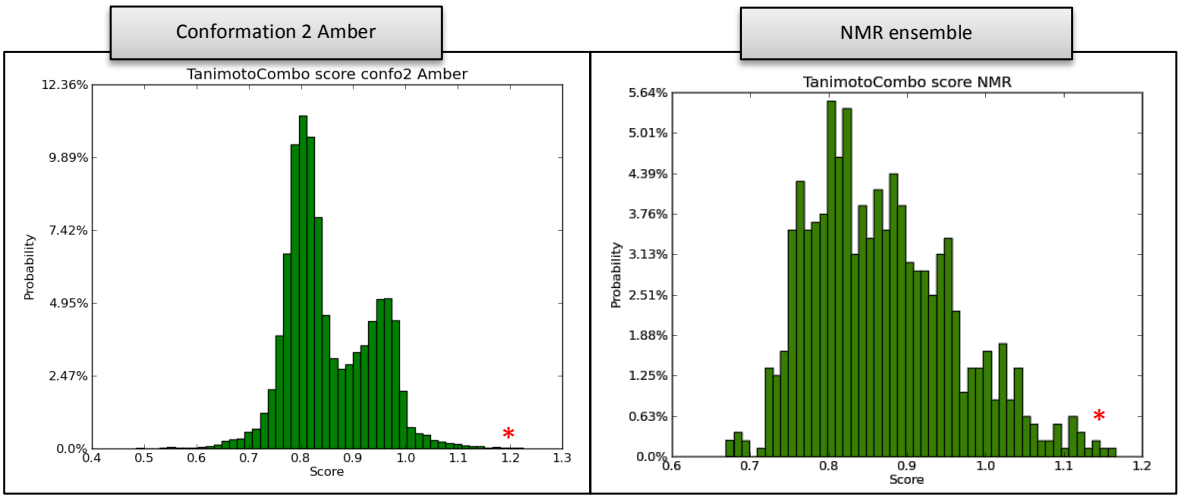


Figure 7.12: Histogram distribution of the TanimotoCombo score in the REMD simulation of the conformation 2 with the Amber Gaff force field and the NMR ensemble. The red asterisks indicate the position of the two potential candidates for the binding from *Figure 7.11*.

7.3 Conclusion:

Using the knowledge of the GSK2194069 x-ray crystal structure bound to the KR domain of hFAS, the study of the contacts made between the final ensemble structures obtained from *Chapter 6* using ROCS and Dash/RMSD protocols enabled us to try and predict the binding mode of novel hFAS inhibitors with consistent interactions. In this chapter, the extent to which the most similar conformer to the GSK2194069 x-ray structure for each compound interacts favourably with hFAS receptor with both the compound and the receptor maintained rigid.

The analysis of GSK2194069 REMD, NMR and docking ensembles has shown that compounds 2, 3 and 4 when hypothetically bound to the receptor adopt similar contacts as the GSK2194069 x-ray binding mode involving the three main parts of each compound structure as introduced in *section 6.2 Figure 6.2*. Three native contacts from the x-ray appear conserved in compounds 3 and 4 involving residues tyrosine 2034, serine 2081 and phenylalanine 2109 *(Tables 7.3 and 7.4)*. Additionally, new hydrogen bonds are formed in all the compounds involving the residue glutamine 2023 for compounds 2 and 3 *(Figures 7.5 and 7.8)*, residue arginine 2026 for compound 3 *(Figure 7.8)* and residues asparagine 2028 and serine 2023 for compound 4 *(Figure 7.11)* appearing to mimic the x-ray hydrogen bond with serine 2021 to stabilize the central core of the compound 4.

Furthermore, the hope that docking will provide us ideas and potential binding modes was not as successful as expected. Fewer contacts and poor overlays were observed making some docking conformations less favourable for binding *(Figures 7.4)*.

Additionally, the ligand-receptor complex solvation-desolvation effect has not been taken into account as well as the extent to which the most similar structure is also the most populated in the conformational ensemble which could affect the potency of the compound as observed with compound 4. The distribution of the TanimotoCombo score from compound 4 suggested that the conformations that emphasized good potential binding mode are of low population in the ensemble *(Figure 7.12).* However, the study has illustrated that the ensembles from the REMD and NMR methodologies provided good quality conformations from solution phase and enabled us to find conformations which fit well in the hFAS active site as potential candidates to a better binding mode.

Chapter 8 Conclusion and future work

In this thesis the hypothesis that computer simulations can be used to generate NMR ensemble in solution has been investigated. The rationale behind this approach is the better a ligand conformation reflects the bound geometry the better is the affinity toward the target for entropic and enthalpic reasons. The conformational space of four hFAS inhibitors including GSK2194069 has been examined using experimental and computational methods. Replica Exchange Molecular Dynamics simulations using two force fields, the Amber Gaff and CGenff Charmm, starting with two different conformations have been performed. NMR data in solution on these four compounds have been provided to us by C4X. Extensive comparisons have been performed between structures generated within the Cambridge Structural Database (CSD), NMR population and the REMD ensembles. Several points have been emphasized through the analyses. In Chapter 4 and Chapter 5, the ability of REMD to capture the experimental conformational equilibrium in solution was tested using torsion angle distributions and principal component analysis (PCA). In Chapter 6 and Chapter 7 the conformational dynamics of the experimental and computational ensembles have been studied. An x-ray crystal structure of GSK2194069, a known inhibitor of hFAS, was used to elucidate potential binding modes of the four compounds. To investigate the ligand preorganization of these conformations, clustering analysis followed by an atom based RMSD superimposition completed by a molecular shape and chemical group alignment analysis using ROCS was performed to identify conformations similar to the GSK2194069 x-ray structure. The interactions made between the most highly scoring conformations from the cluster and ROCS analyses and the receptor were investigated using a rigid receptor and rigid ligand, thus, neither the conformations nor the receptor were relaxed, to maintain the conformation of the structures.

8.1 Conformational ensemble equilibrium:

The results generated using torsion angle distributions and PCA, highlighted that despite being able to reproduce the NMR ensembles, conventional REMD does not ensure sufficient sampling of the system. Conformational trapping for certain dihedral angles was observed for each of the compounds according to the starting conformation. Furthermore, when the sampling appeared to be consistent compared to NMR and CSD data, for some cases, differences were

observed between force fields. It was not clear whether the Amber Gaff or the CGenFF force field was performing better. Neither of the two force fields was demonstrated to be superior to the other. Additionally, when using the CSD database to evaluate if the conformations sampled by REMD were correct from a commonly used assumption that the conformations adopted by the CSD crystal structures are likely to be representative of the conformations adopted in solution, it has been observed that the generality of certain queries used for the search gave misleading results.

Therefore, the standard temperature REMD is restricted by a being computationally expensive equilibrium method with the need for large number of replicas, covering a wide range of temperature, to achieve high energy barrier crossing to escape local minima. Thus, enhancing the sampling through temperature REMD appears to not always be efficient. An alternative to temperature REMD is the possibility to use a variant of REMD known as Replica Exchange with Solute Tempering (REST) [188]. REST only simulates the ligand at different temperature during the process while the water molecules remained at room temperature allowing fewer replicas to be used. Additionally, in the REST methodology, replicas differ with respect to scaling factors applied to the solute-solvent and solvent-solvent potential energies. REST uses a Hamiltonian based approach to soften the potential interactions, which could result in more efficient sampling than standard REMD. A recent study, in 2016, compared REMD with REST on the sampling efficiency of Aβ peptides showing better convergence when using REST [189]. In term of force field, to evaluate the quality of the two force fields to predict accurate conformational ensembles, ab-initio calculations can be performed [190].

8.2 Ligand preorganization:

Understanding the structural behaviour of a ligand in solution and the binding mechanism toward its target is fundamental in drug development. The superimposition of the four compounds on the GSK2194069 x-ray structure enabled us to determine a set of conformations from the computational and experimental ensembles with similar x-ray binding modes. However, the intent to investigate the relative population of the most prevalent conformations was not successful due to the high flexibility of the compounds.

The study of the contacts made between the set of conformations with hFAS receptor, using the knowledge of the GSK2194069 x-ray structure binding mode,

highlighted the importance of native contacts that appeared highly conserved in all compound and also the discovery of additional interactions that could favour the increase of binding affinity. Additionally, it has been observed that the structures from the ROCS protocol was able to generated more native contacts than the structures using Dash/RMSD protocol. Potential binding modes for each compound have been generated leading to valuable information for future rational drug design.

Therefore, the intent of the study was to evaluate if the conformational ensembles in solution from NMR and REMD bind favourably to the receptor binding pocket to investigate ligand preorganisation. Future work to identify if whether or not the ligand binds to the receptor will require, the population information to be used in the analysis. Additionally, other components, such as solvation-desolvation effect, conformational reorganization energies upon binding, which are important regulators in the stability of the complex ligand-receptor will need to be considered.

Appendix A

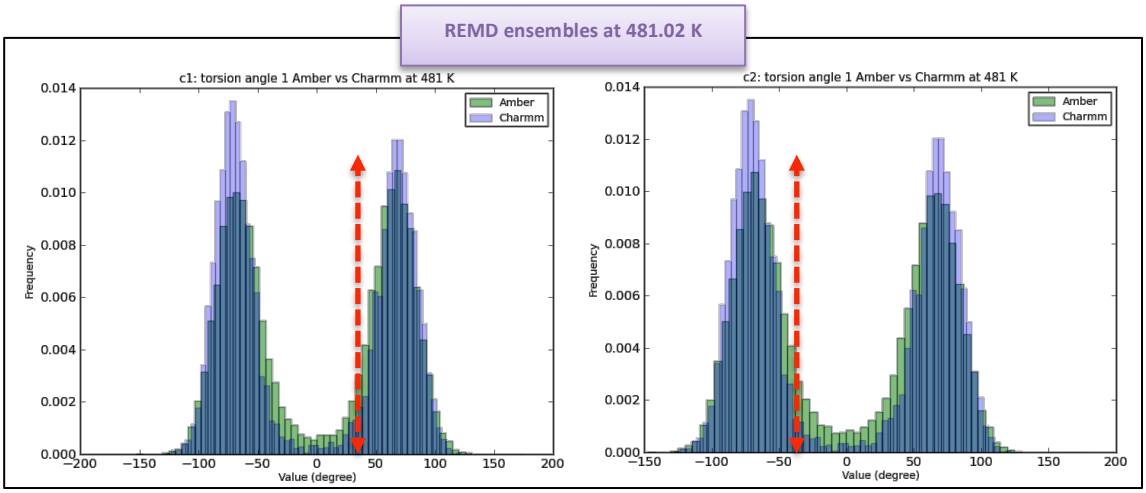


Figure A.1: Torsion angle τ1 distribution in the REMD trajectories of conformations 1 and 2 of GSK2194069 in the Amber Gaff and CGenFF Charmm force fields at 481 K. The red arrows describe the values of the torsion angle in the starting conformations.

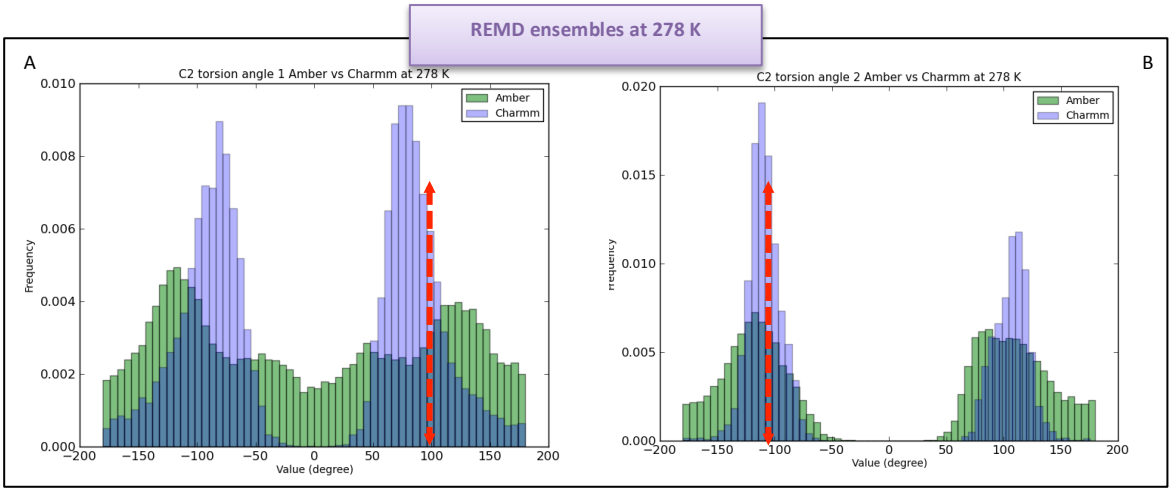


Figure A.2: Torsion angle τ1 distribution in the REMD trajectories of conformation 2 of compound 2 in the Amber Gaff and CGenFF Charmm force fields at 278 K. The red arrows describe the values of the torsion angle in the starting conformation.

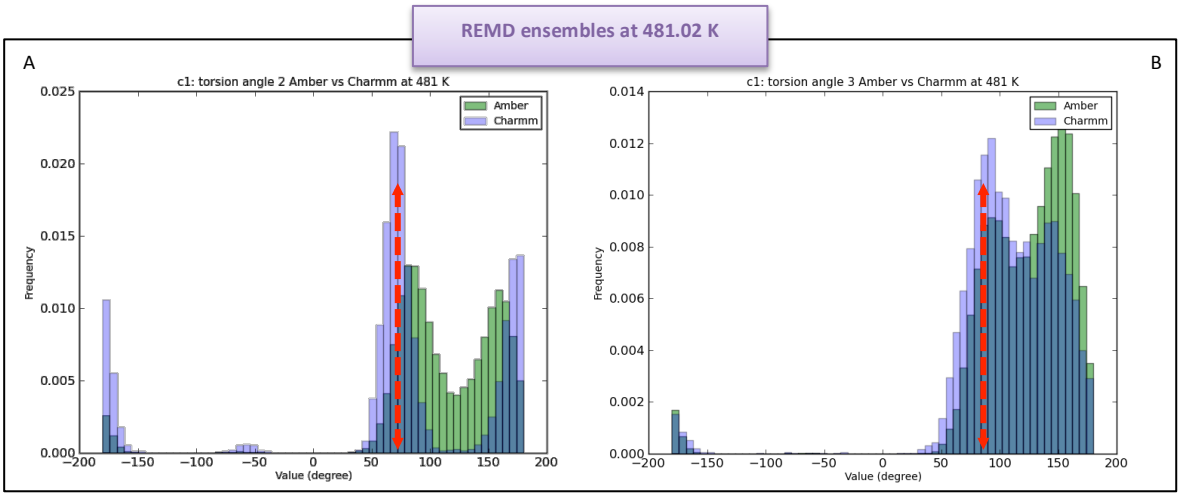


Figure A.3: Torsion angles τ2 (A) and τ3 (B) distribution in the REMD trajectories of conformation 1 of compound 3 in the Amber Gaff and CGenFF Charmm force fields at 481 K. The red arrows describe the values of the torsion angle in the starting conformations.

Appendix B

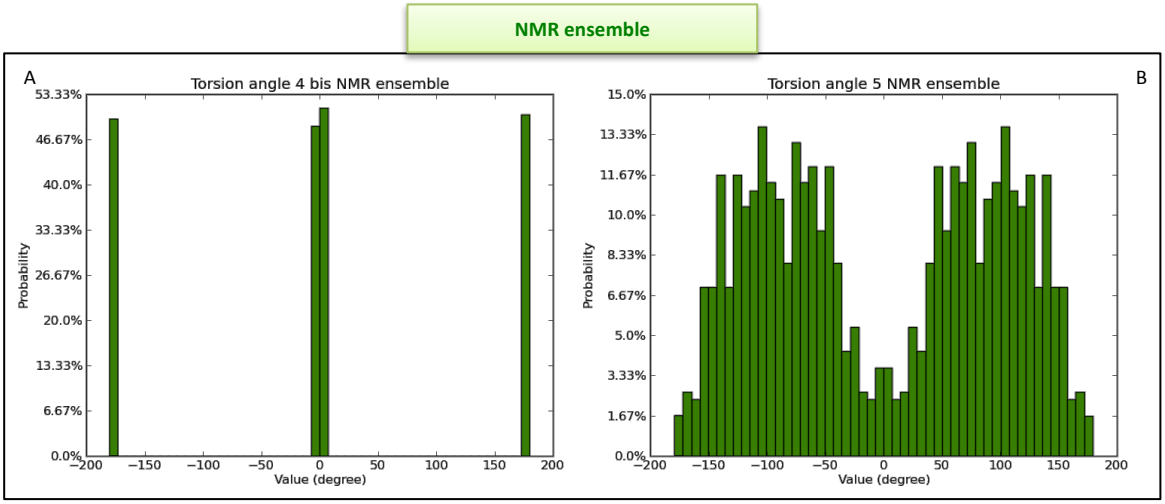


Figure B.1: Torsion angles τ4 (A) and τ4 (B) distribution in the NMR ensemble of compound 2.

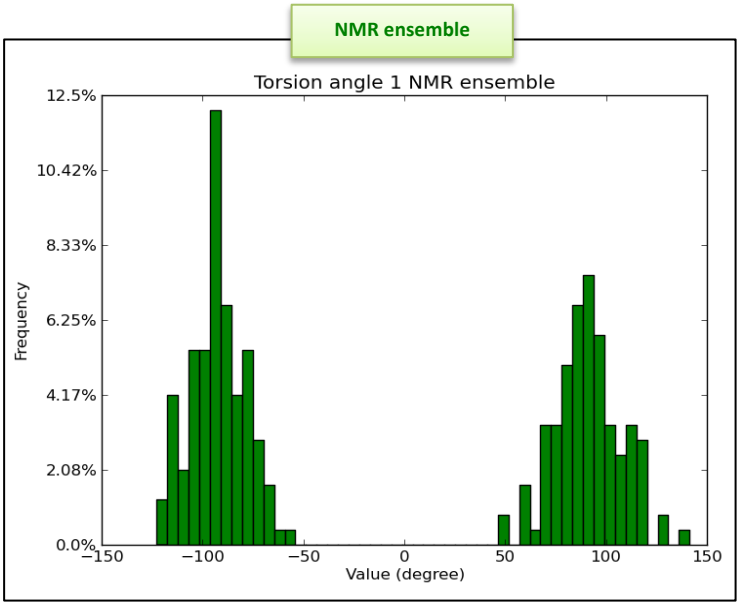
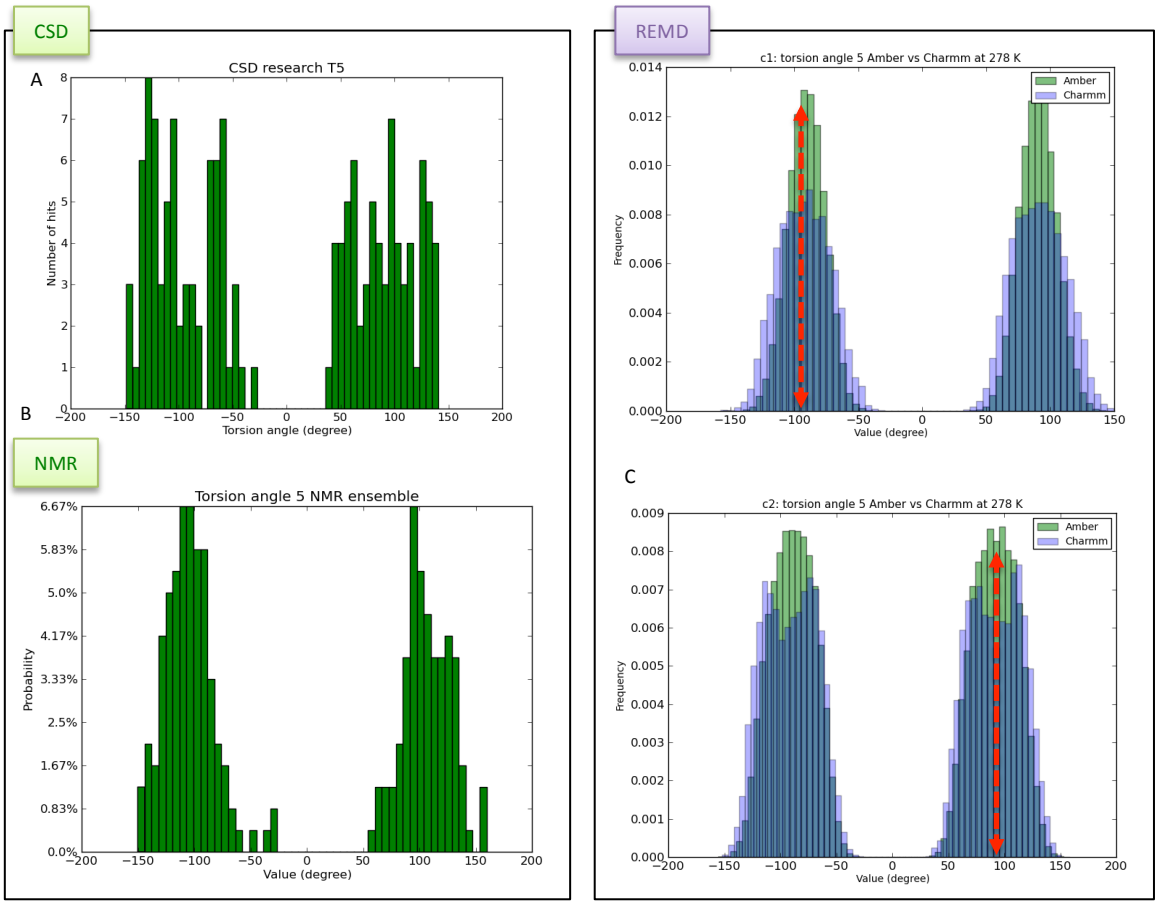


Figure B.2: Torsion angle τ1 distribution in the NMR ensemble of compound 3.



Compound 3

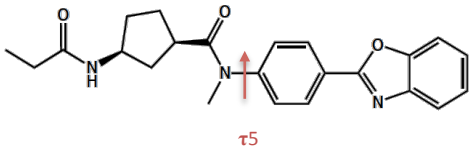
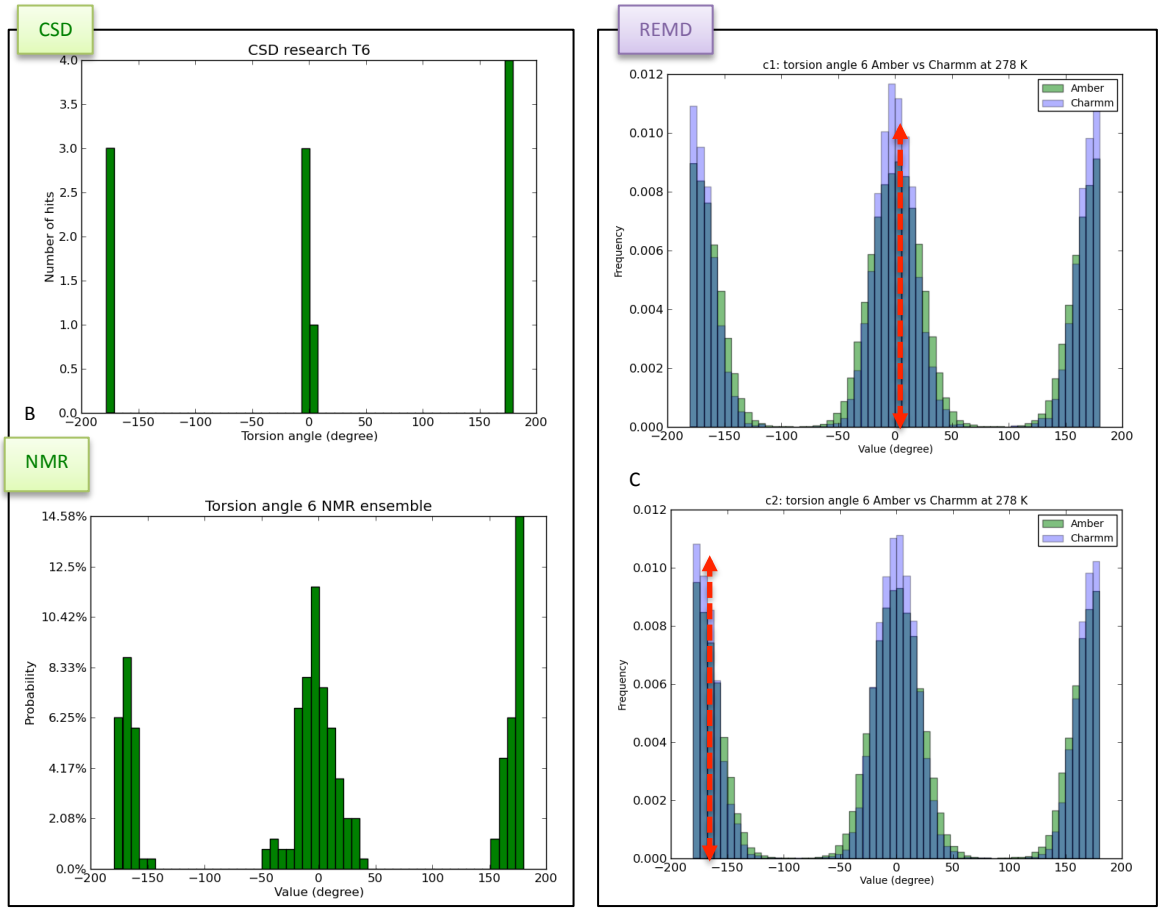
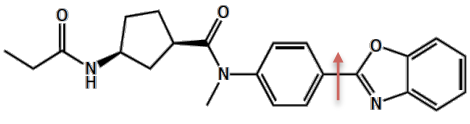


Figure B.3: Torsion angle τ5 distribution in the CSD (A), the NMR ensemble (B) and the REMD trajectories (B) for the starting conformations 1 and 2 with the Amber Gaff (green) and CGenFF Charmm force fields (violet) for compound 3. The red arrows show the starting conformation torsion values.



Compound 3



τ6

Figure B.4: Torsion angle τ6 distribution in the CSD (A), the NMR ensemble (B) and the REMD trajectories (B) for the starting conformations 1 and 2 with the Amber Gaff (green) and CGenFF Charmm force fields (violet) for compound 3. The red arrows show the starting conformation torsion values.

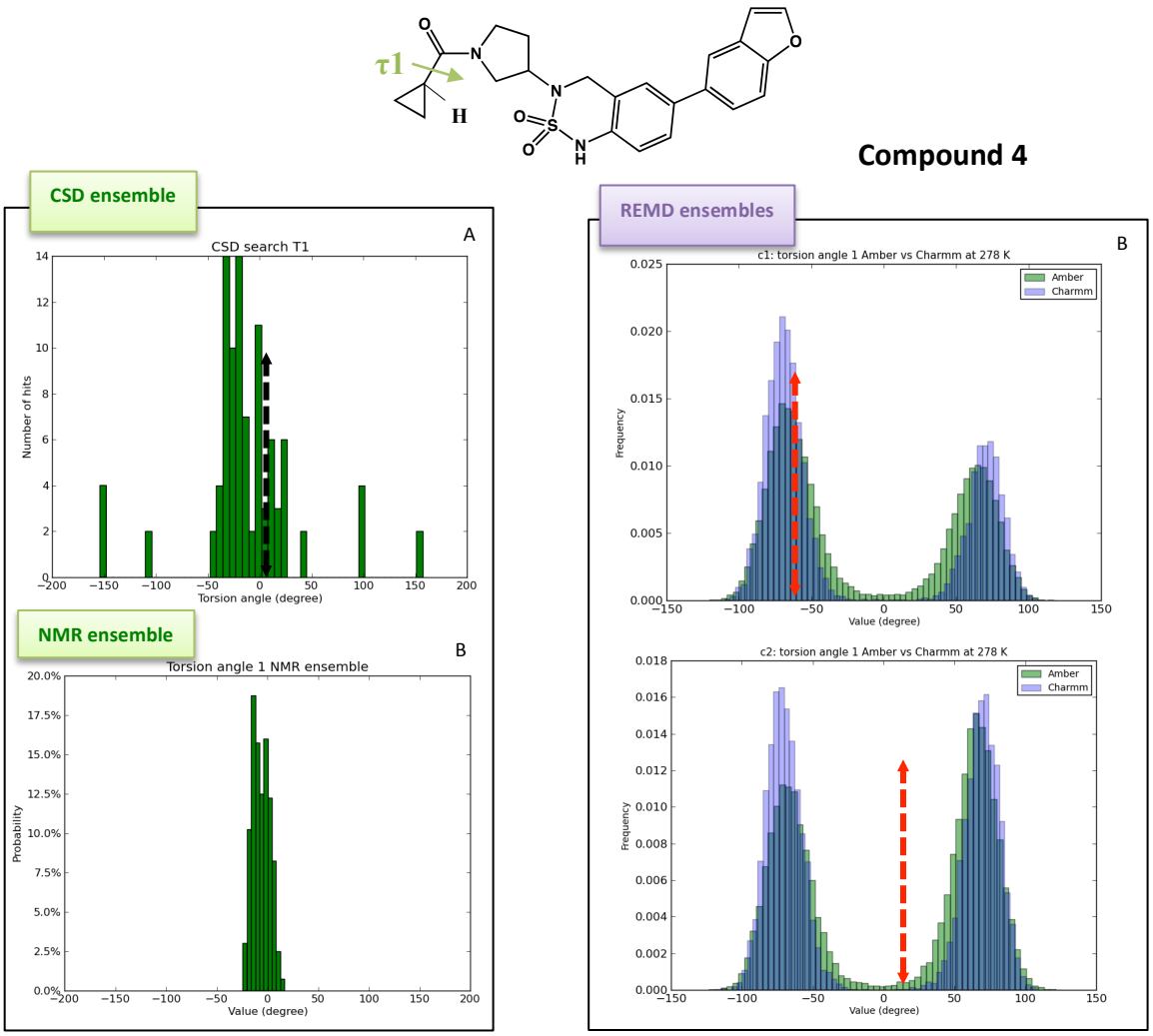


Figure B.5: Torsion angle τ1 distribution in the CSD (A), the NMR ensemble (B) and the REMD trajectories (B) for the starting conformations 1 and 2 with the Amber Gaff (green) and CGenFF Charmm force fields (violet) for compound 4. The red arrows show the starting conformation torsion values.

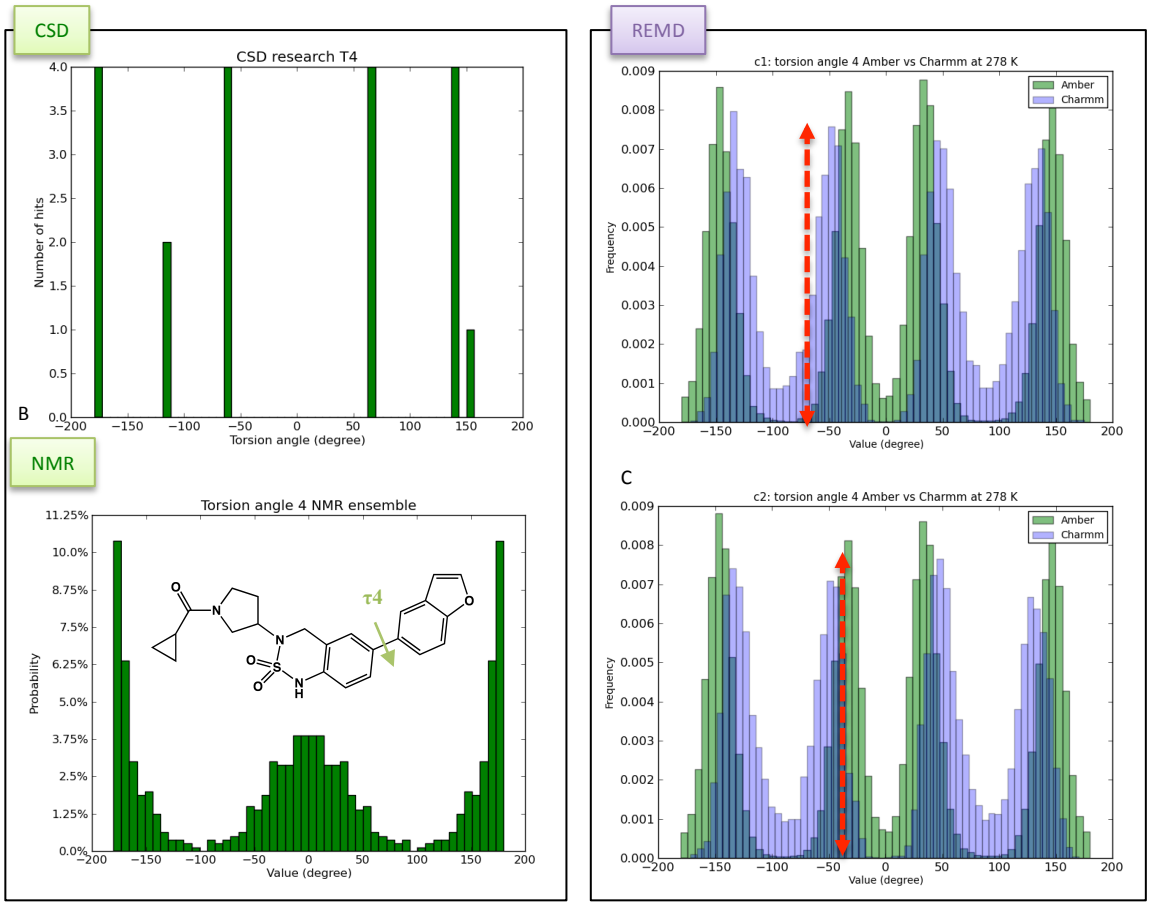
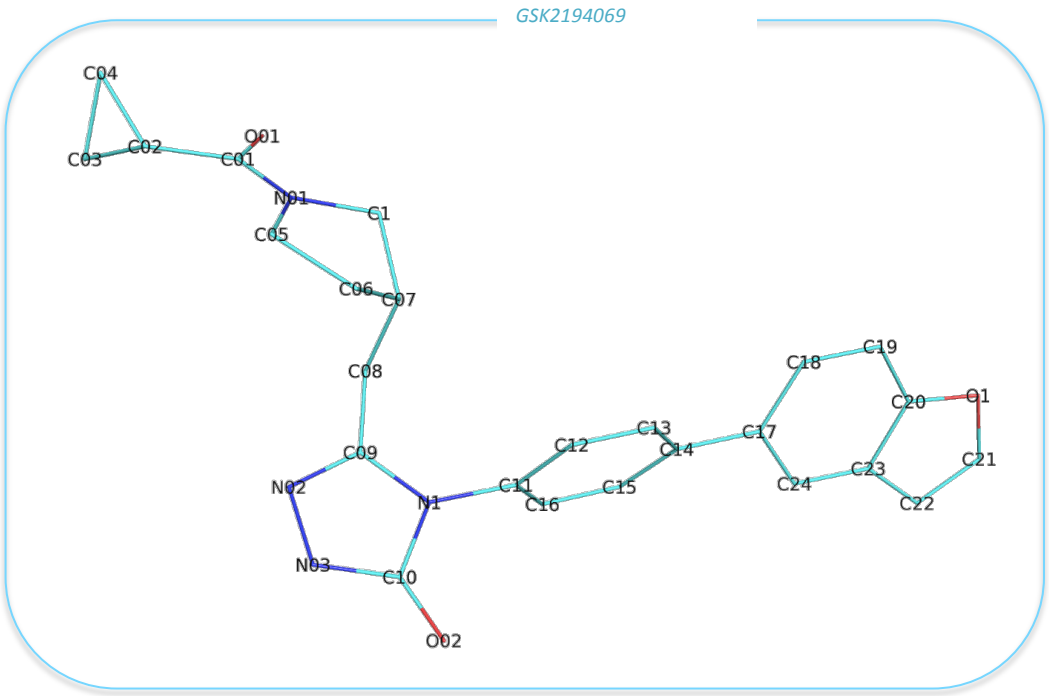
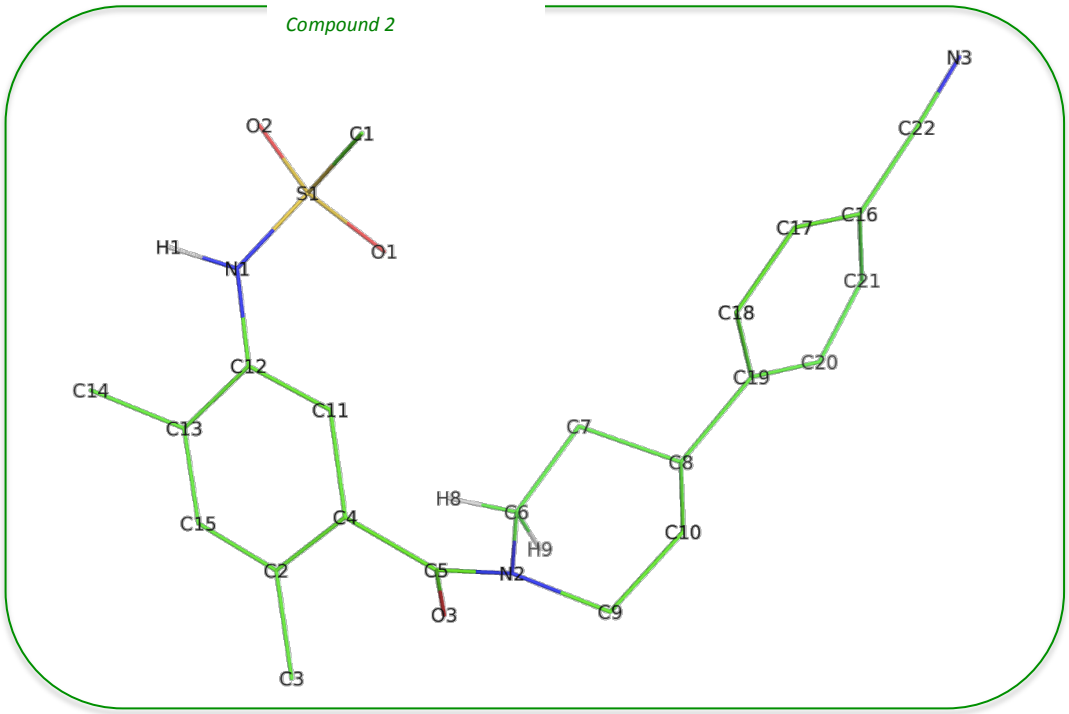


Figure B.6: Torsion angle τ4 distribution in the CSD (A), the NMR ensemble (B) and the REMD trajectories (B) for the starting conformations 1 and 2 with the Amber Gaff (green) and CGenFF Charmm force fields (violet) for compound 4. The red arrows show the starting conformation torsion values.

Appendix C


GSK2194069
C04
C03
C02
O01
C01
N01
C1
C05
C06
C07
C08
C09
N02
N1
N03
C10
O02
C11
C16
C12
C13
C14
C15
C17
C18
C19
C20
O1
C21
C22
C23
C24


Compound 2
O2
C1
S1
O1
H1
N1
C12
C14
C13
C11
C15
C4
C2
C3
C5
O3
N2
H8
C6
H9
C7
C8
C9
C10
C19
C18
C17
C16
C22
N3
C21
C20

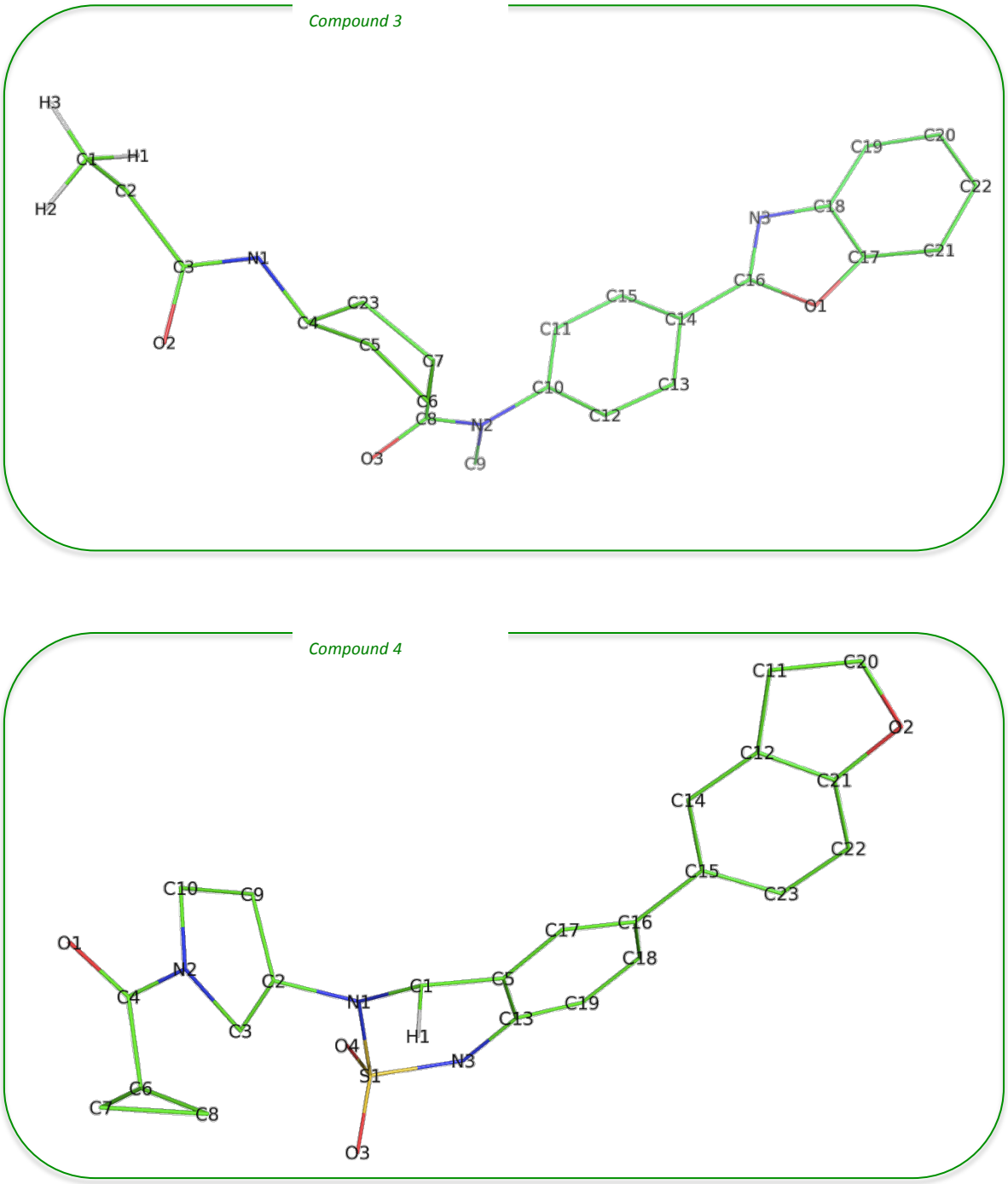


Figure C.1: Compounds 2, 3 and 4 atom names that will be used to perform RMSD alignment with GSK2194069 atom names as reference structure.

GSK2194069	Compound 2	Compound 3	Compound 4
C02	H1	C1	C1
C01	S1	C2	N1
C03	N1	C3	C2
C04	C1	N1	C3
O01	O1	C4	N2
C06	O2	C5	C4
C05	C2	C6	C5
N01	C3	C7	C6
C1	C4	C8	C7
C07	C5	N2	C8
C08	N2	C9	C9
C10	C6	C10	C10
N1	C7	C11	N3
C09	C8	C12	O1
N02	C9	C13	C11
N03	C10	C14	C12
O02	C11	C15	C13
C14	C12	C16	C14
C13	C13	N3	C15
C12	C14	O1	C16
C11	C15	C17	C17
C16	O3	C18	C18
C15	C16	C19	C19
C20	C17	C20	C20
C23	C18	C21	O2
C22	C19	C22	C21
C21	C20	O2	C22
O1	C21	O3	C23

C24	C22	C23	S1
C17	N3	H1	O3
C18	H8	H2	O4
C19	H9	H3	H1

Table C.1: Atom pairs to perform the RMSD overlay between the GSK2194069 x-ray structure (reference) and the structures from the different ensembles (REMD, NMR, Docking) of compounds 2, 3 and 4.

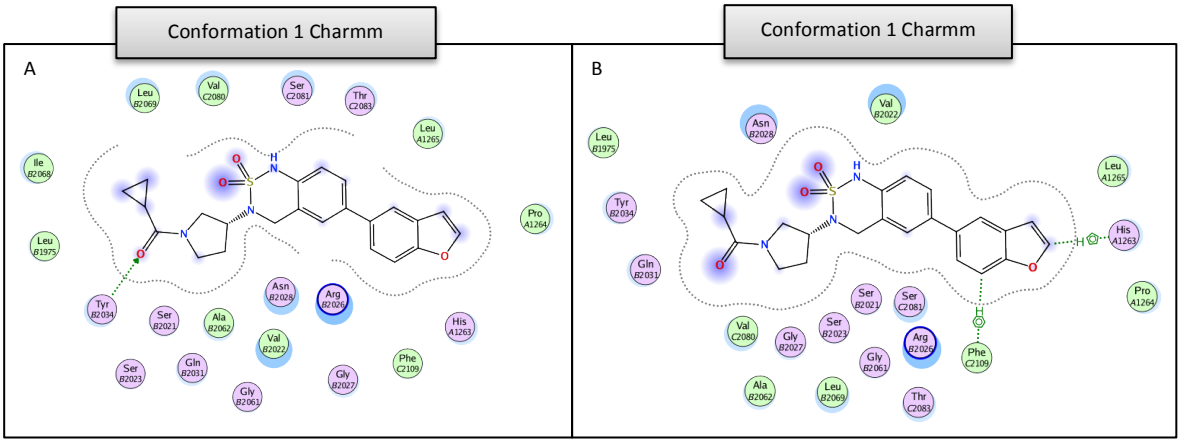


Figure C.2: Ligand interactions diagram of conformation 1 in the CGenff Charmm force field using the ROCS (A) and Dash/RMSD (B) protocol of compound 4. The blue circles illustrate the solvent exposure.

Bibliography

1. Mullin, Rick. "Cost to develop new pharmaceutical drug now exceeds $2.5 B." *Scientific American* 24 (2014).

2. DiMasi, Joseph A., Henry G. Grabowski, and Ronald W. Hansen. "The cost of drug development." *New England Journal of Medicine* 372.20 (2015): 1972-1972.

3. Billones, Junie B., et al. "in silico discovery and in vitro activity of inhibitors against Mycobacterium tuberculosis 7, 8-diaminopelargonic acid synthase (Mtb Bioa)." *Drug design, development and therapy*11 (2017): 563.

4. Light, Donald W., and Joel Lexchin. "Pharmaceutical research and development: what do we get for all that money?." (2012).

5. Prasad, Vinay, and Sham Mailankody. "Research and development spending to bring a single cancer drug to market and revenues after approval." *JAMA Internal Medicine* (2017).

6. Hughes, James P., et al. "Principles of early drug discovery." *British journal of pharmacology* 162.6 (2011): 1239-1249.

7. Wang, Jianling, and Laszlo Urban. "The impact of early ADME profiling on drug discovery and development strategy." *DDW Drug Discovery World*5.4 (2004): 73-86.

8. Waring, Michael J., et al. "An analysis of the attrition of drug candidates from four major pharmaceutical companies." *Nature reviews. Drug discovery* 14.7 (2015): 475.

9. Prueksaritanont, Thomayant, and Cuyue Tang. "ADME of Biologics—What Have We Learned from Small Molecules?." *The AAPS journal* 14.3 (2012): 410-419.

10. Cox, Christopher D., et al. "Discovery of the Dual Orexin Receptor Antagonist [(7 R)-4-(5-Chloro-1, 3-benzoxazol-2-yl)-7-methyl-1, 4-diazepan-1-yl][5-methyl-2-(2 H-1, 2, 3-triazol-2-yl) phenyl] methanone (MK-4305) for the Treatment of Insomnia." *Journal of medicinal chemistry* 53.14 (2010):

5320-5332.

11. Krumm, Brian, et al. "Identification of small molecule inhibitors of Interleukin-18." *Scientific Reports* 7 (2017).

12. Bernard, Denzil, et al. "Computer-aided drug design: Structure-activity relationships of delta opioid ligands." *Drug Design Reviews-Online* 2.4 (2005): 277-291.

13. Wadood, A., et al. "In-silico drug design: An approach which revolutionarised the drug discovery process." *OA Drug Design & Delivery* 1.1 (2013): 3.

14. Sliwoski, Gregory, et al. "Computational methods in drug discovery." *Pharmacological reviews* 66.1 (2014): 334-395

15. Pártay, Lívia B., Albert P. Bartók, and Gábor Csányi. "Efficient sampling of atomic configurational spaces." *The Journal of Physical Chemistry B* 114.32 (2010): 10502-10512

16. Alonso, Hernán, Andrey A. Bliznyuk, and Jill E. Gready. "Combining docking and molecular dynamic simulations in drug design." *Medicinal research reviews* 26.5 (2006): 531-568.

17. Stark, Jaime L., and Robert Powers. "Application of NMR and molecular docking in structure-based drug discovery." *NMR of proteins and small biomolecules*. Springer Berlin Heidelberg, 2011. 1-34

18. Zhang, Liqun, Sabine Bouguet-Bonnet, and Matthias Buck. "Combining NMR and Molecular Dynamics Studies for Insights into the Allostery of Small GTPase–Protein Interactions." *Allostery: Methods and Protocols* (2012): 235-259

19. Lindert, Steffen, et al. "Drug screening strategy for human membrane proteins: from NMR protein backbone structure to in silica-and NMR-screened hits." *Biochemical and biophysical research communications* 445.4 (2014): 724-733

20. Fischer, Emil. "Einfluss der Configuration auf die Wirkung der Enzyme." *European Journal of Inorganic Chemistry* 27.3 (1894): 2985-

2993.

21. Koshland, Daniel E. "The key–lock theory and the induced fit theory." *Angewandte Chemie International Edition* 33.23-24 (1995): 2375-2378.

22. Neubig, Richard R., et al. "International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on terms and symbols in quantitative pharmacology." *Pharmacological Reviews* 55.4 (2003): 597-606.

23. HUANG, SM. "Guidance for industry: drug interaction studies-study design, data analysis, and implications for dosing and labeling, Draft Guidance." *http://www. fda. gov/downloads/Drugs/Guidance-ComplianceRegulatoryInformation/Guidances/UCM072101. pdf* (2006).

24. Grimshaw, Jeremy M., and Ian T. Russell. "Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations." *The Lancet* 342.8883 (1993): 1317-1322.

25. Lilly, Eli. "Company and NIH Chemical Genomics Center." *Assay guidance manual version* 5 (2008).

26. Petersen, Raymond C. "The linear relationship between enthalpy and entropy of activation." *The Journal of Organic Chemistry* 29.11 (1964): 3133-3135.

27. Ford, David M. "Enthalpy– entropy compensation is not a general feature of weak association." *Journal of the American Chemical Society* 127.46 (2005): 16167-16170.

28. Lumry, Rufus, and Shyamala Rajender. "Enthalpy–entropy compensation phenomena in water solutions of proteins and small molecules: a ubiquitous properly of water." *Biopolymers* 9.10 (1970): 1125-1227.

29. Gallicchio, Emilio, Masahito Mogami Kubo, and Ronald M. Levy. "Entropy–enthalpy compensation in solvation and ligand binding revisited." *Journal of the American Chemical Society* 120.18 (1998): 4526-4527.

30. Krug, R. R., W. G. Hunter, and R. A. Grieger. "Statistical interpretation of enthalpy-entropy compensation." *Nature* 261.5561 (1976): 566-567.

31. Cornish-Bowden, Athel. "Enthalpy—entropy compensation: a phantom phenomenon." *Journal of biosciences* 27.2 (2002): 121-126.

32. Ladbury, John E., Gerhard Klebe, and Ernesto Freire. "Adding calorimetric data to decision making in lead discovery: a hot tip." *Nature reviews. Drug discovery*9.1 (2010): 23.

33. Olsson, Tjelvar SG, et al. "The thermodynamics of protein–ligand interaction and solvation: insights for ligand design." *Journal of molecular biology* 384.4 (2008): 1002-1017.

34. Chodera, John D., and David L. Mobley. "Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design." (2013).

35. Ferenczy, György G., and György M. Keserű. "Thermodynamics guided lead discovery and optimization." *Drug discovery today* 15.21 (2010): 919-932.

36. Freire, Ernesto. "A thermodynamic approach to the affinity optimization of drug candidates." *Chemical biology & drug design* 74.5 (2009): 468-472.

37. Scott, Andrew D., et al. "Thermodynamic optimisation in drug discovery: a case study using carbonic anhydrase inhibitors." *ChemMedChem* 4.12 (2009): 1985-1989.

38. Sarver, Ronald W., et al. "Thermodynamic and structure guided design of statin based inhibitors of 3-hydroxy-3-methylglutaryl coenzyme A reductase." *Journal of medicinal chemistry* 51.13 (2008): 3804-3813.

39. Sharp, Kim. "Entropy—enthalpy compensation: Fact or artifact?." *Protein Science* 10.3 (2001): 661-667.

40. Reynolds, Charles H., and M. Katharine Holloway. "Thermodynamics of ligand binding and efficiency." *ACS medicinal chemistry letters* 2.6 (2011): 433-437.

41. Cram, Donald J. "The design of molecular hosts, guests, and their

complexes." *Journal of Inclusion Phenomena and Macrocyclic Chemistry* 6.4 (1988): 397-413.

42. Berg, J. M., J. L. Tymoczko, and L. Stryer. "Section 22.4 fatty acids are synthesized and degraded by different pathways. Biochemistry. NCBI Bookshelf. A service of the National Library of Medicine, National Institutes of Health." *New York: WH Freeman. Web access on May* 25 (2002): 2013.

43. Berg, J. M., J. L. Tymoczko, and L. Stryer. "Triacylglycerols are highly concentrated energy stores." *Biochemistry* 5 (2002).

44. Neitzel, J. J. "Fatty acid molecules: fundamentals and role in signaling." *Nat Educ* 3 (2010): 57.

45. Menendez, Javier A., and Ruth Lupu. "Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis." *Nature reviews. Cancer* 7.10 (2007): 763.

46. Pappenberger, Günter, et al. "Structure of the human fatty acid synthase KS–MAT didomain as a framework for inhibitor design." *Journal of molecular biology* 397.2 (2010): 508-519.

47. Chakravarty, Bornali, et al. "Human fatty acid synthase: structure and substrate selectivity of the thioesterase domain." *Proceedings of the National Academy of Sciences of the United States of America* 101.44 (2004): 15567-15572.

48. Maier, Timm, Simon Jenni, and Nenad Ban. "Architecture of mammalian fatty acid synthase at 4.5 Å resolution." *Science* 311.5765 (2006): 1258-1262.

49. Jenni, Simon, et al. "Structure of fungal fatty acid synthase and implications for iterative substrate shuttling." *Science* 316.5822 (2007): 254-261.

50. Flavin, Richard, et al. "Fatty acid synthase as a potential therapeutic target in cancer." *Future oncology* 6.4 (2010): 551-562.

51. Röhrig, Florian, and Almut Schulze. "The multifaceted roles of fatty acid synthesis in cancer." *Nature Reviews Cancer* 16.11 (2016): 732-749.

52. Lu, Jeff Z., et al. "Fatty acid synthesis as a target for antimalarial drug discovery." *Combinatorial chemistry & high throughput screening* 8.1 (2005): 15-26.

53. Lupu, Ruth, and Javier A. Menendez. "Pharmacological inhibitors of fatty acid synthase (FASN)-catalyzed endogenous fatty acid biogenesis: a new family of anti-cancer agents?." *Current pharmaceutical biotechnology* 7.6 (2006): 483-494

54. Baenke, Franziska, et al. "Hooked on fat: the role of lipid synthesis in cancer metabolism and tumour development." *Disease models & mechanisms* 6.6 (2013): 1353-1363.

55. WEISS, Ludwig, et al. "Fatty-acid biosynthesis in man, a pathway of minor importance. Purification, optimal assay conditions, and organ distribution of fatty-acid synthase." *Biological chemistry Hoppe-Seyler* 367.2 (1986): 905-912.

56. Kuhajda, Francis P., et al. "Fatty acid synthesis: a potential selective target for antineoplastic therapy." *Proceedings of the National Academy of Sciences*91.14 (1994): 6379-6383.

57. McGarry, J. Denis, and Nicholas F. Brown. "The mitochondrial carnitine palmitoyltransferase system—from concept to molecular analysis." *The FEBS Journal* 244.1 (1997): 1-14.

58. Liu, Hailan, et al. "Biochemistry, molecular biology, and pharmacology of fatty acid synthase, an emerging therapeutic target and diagnosis/prognosis marker." *International journal of biochemistry and molecular biology* 1.1 (2010): 69.

59. Turrado, Carlos, et al. "New synthetic inhibitors of fatty acid synthase with anticancer activity." *Journal of medicinal chemistry* 55.11 (2012): 5013-5023.

60. Vázquez, María Jesús, et al. "Discovery of GSK837149A, an inhibitor of human fatty acid synthase targeting the β-ketoacyl reductase reaction." *The FEBS journal* 275.7 (2008): 1556-1567.

61. Kuhajda, Francis P. "Fatty acid synthase and cancer: new application of an old pathway." *Cancer research*66.12 (2006): 5977-5980.

62. Hardwicke, Mary Ann, et al. "A human fatty acid synthase inhibitor binds β-ketoacyl reductase in the keto-substrate site." *Nature chemical biology* 10.9 (2014): 774-779.

63. Thupari, Jagan N., et al. "C75 increases peripheral energy utilization and fatty acid oxidation in diet-induced obesity." *Proceedings of the National Academy of Sciences* 99.14 (2002): 9498-9502.

64. Kuhajda, Francis P., et al. "Synthesis and antitumor activity of an inhibitor of fatty acid synthase." *Proceedings of the National Academy of Sciences*97.7 (2000): 3450-3454.

65. Yang, Nengyu, et al. "C75 [4-methylene-2-octyl-5-oxo-tetrahydro-furan-3-carboxylic acid] activates carnitine palmitoyltransferase-1 in isolated mitochondria and intact cells without displacement of bound malonyl CoA." *Journal of Pharmacology and Experimental Therapeutics* 312.1 (2005): 127-133.

66. Rendina, Alan R., and Dong Cheng. "Characterization of the inactivation of rat fatty acid synthase by C75: inhibition of partial reactions and protection by substrates." *Biochemical Journal* 388.3 (2005): 895-903.

67. Lupu, Ruth, and Javier A. Menendez. "Pharmacological inhibitors of fatty acid synthase (FASN)-catalyzed endogenous fatty acid biogenesis: a new family of anti-cancer agents?." *Current pharmaceutical biotechnology* 7.6 (2006): 483-494.

68. Abramson, Hanley N. "The lipogenesis pathway as a cancer target." *J. Med. Chem* 54.16 (2011): 5615-5638.

69. Kridel, Steven J., et al. "Orlistat is a novel inhibitor of fatty acid synthase with antitumor activity." *Cancer research* 64.6 (2004): 2070-2075.

70. McFadden, Jill M., et al. "Application of a Flexible Synthesis of (5 R)-Thiolactomycin To Develop New Inhibitors of Type I Fatty Acid

Synthase." *Journal of medicinal chemistry* 48.4 (2005): 946-961.

71. Zhou, Weibo, et al. "Fatty acid synthase inhibition activates AMP-activated protein kinase in SKOV3 human ovarian cancer cells." *Cancer research* 67.7 (2007): 2964-2971.

72. Turrado, Carlos, et al. "New synthetic inhibitors of fatty acid synthase with anticancer activity." *Journal of medicinal chemistry* 55.11 (2012): 5013-5023.

73. Hallman, J., C. Laudeman, and R. Liu. "Benzimidazoles as fatty acid synthase inhibitors." *World Intellectual Property Organization WO2011056635* (2011).

74. Chaudhari, Amita M., et al. "Azabenzimidazoles as fatty acid synthase inhibitors." U.S. Patent Application No. 13/511,767.

75. Adams, Nicholas David, et al. "Triazolones as fatty acid synthase inhibitors." U.S. Patent No. 9,340,529. 17 May 2016.

76. Adams, Nicholas D., et al. "Fatty acid synthase inhibitors." U.S. Patent Application No. 13/824,000.

77. https://www.c4xdiscovery.com/

78. Blundell, Charles D., Martin J. Packer, and Andrew Almond. "Quantification of free ligand conformational preferences by NMR and their relationship to the bioactive conformation." *Bioorganic & medicinal chemistry* 21.17 (2013): 4976-4987.

79. Georgsson, Jennie, et al. "GPR103 antagonists demonstrating anorexigenic activity in vivo: design and development of pyrrolo [2, 3-c] pyridines that mimic the C-terminal Arg-Phe motif of QRFP26." *Journal of medicinal chemistry* 57.14 (2014): 5935-5948.

80. Driggers, Edward M., et al. "The exploration of macrocycles for drug discovery--an underexploited structural class." *Nature reviews. Drug discovery* 7.7 (2008): 608.

81. K Oyelere, Adegboyega. "Editorial [Hot topic: Macrocycles in Medicinal

Chemistry and Drug Discovery (Guest Editors: Adegboyega K. Oyelere)]." *Current Topics in Medicinal Chemistry* 10.14 (2010): 1359-1360.

82. Marsault, Eric, and Mark L. Peterson. "Macrocycles are great cycles: applications, opportunities, and challenges of synthetic macrocycles in drug discovery." *Journal of medicinal chemistry* 54.7 (2011): 1961-2004.

83. Mallinson, Jamie, and Ian Collins. "Macrocycles in new drug discovery." *Future medicinal chemistry* 4.11 (2012): 1409-1438.

84. Wittenberg, James B., and Lyle Isaacs. "Complementarity and Preorganization." *Supramolecular Chemistry: From Molecules to Nanomaterials* (2012).

85. Wessjohann, Ludger A., et al. "What can a chemist learn from nature's macrocycles?–A brief, conceptual view." *Molecular diversity* 9.1 (2005): 171-186.

86. Ruan, B-F., and H-L. Zhu. "The chemistry and biology of the bryostatins: potential PKC inhibitors in clinical development." *Current medicinal chemistry*19.16 (2012): 2652-2664.

87. Yu, Xufen, and Dianqing Sun. "Macrocyclic drugs and synthetic methodologies toward macrocycles." *Molecules* 18.6 (2013): 6230-6268.

88. Wang, Dongyu, et al. "Total synthesis of the marine cyclic depsipeptide viequeamide A." *Journal of natural products* 76.5 (2013): 974-978.

89. Parenty, Arnaud, et al. "Update 1 of: Macrolactonizations in the total synthesis of natural products." *Chemical reviews* 113.1 (2013): PR1-PR40.

90. Tsantrizos, Youla S., et al. "Macrocyclic inhibitors of the NS3 protease as potential therapeutic agents of hepatitis C virus infection." *Angewandte Chemie International Edition* 42.12 (2003): 1356-1360.

91. Smith, Whitney W., and Paul A. Bartlett. "Macrocyclic inhibitors of penicillopepsin. 3. Design, synthesis, and evaluation of an inhibitor bridged between P2 and P1 '." *Journal of the American Chemical Society* 120.19 (1998): 4622-4628.

92. Johannes, Jeffrey W., et al. "Structure Based Design of Non-Natural Peptidic Macrocyclic Mcl-1 Inhibitors." *ACS Medicinal Chemistry Letters* 8.2 (2017): 239-244.

93. William, Anthony D., et al. "Discovery of kinase spectrum selective macrocycle (16 e)-14-methyl-20-oxa-5, 7, 14, 26-tetraazatetracyclo [19.3. 1.1 (2, 6). 1 (8, 12)] heptacosa-1 (25), 2 (26), 3, 5, 8 (27), 9, 11, 16, 21, 23-decaene (SB1317/TG02), a potent inhibitor of cyclin dependent kinases (cdks), janus kinase 2 (jak2), and fms-like tyrosine kinase-3 (flt3) for the treatment of cancer." *Journal of medicinal chemistry* 55.1 (2011): 169-196.

94. Poulsen, Anders, et al. "Structure-based design of nitrogen-linked macrocyclic kinase inhibitors leading to the clinical candidate SB1317/TG02, a potent inhibitor of cyclin dependant kinases (CDKs), Janus kinase 2 (JAK2), and Fms-like tyrosine kinase-3 (FLT3)." *Journal of molecular modeling* 19.1 (2013): 119-130.

95. Kettle, Jason G., et al. "Potent and selective inhibitors of MTH1 probe its role in cancer cell survival." *Journal of medicinal chemistry* 59.6 (2016): 2346-2361.

96. Martin, Stephen F., and John H. Clements. "Correlating structure and energetics in protein-ligand interactions: paradigms and paradoxes." *Annual review of biochemistry* 82 (2013): 267-293.

97. García-Echeverría, Carlos, et al. "Mapping the X+ 1 binding site of the Grb2-SH2 domain with α, α-disubstituted cyclic α-amino acids." *Bioorganic & medicinal chemistry letters* 9.20 (1999): 2915-2920.

98. Davidson, James P., et al. "Calorimetric and structural studies of 1, 2, 3-trisubstituted cyclopropanes as conformationally constrained peptide inhibitors of Src SH2 domain binding." *Journal of the American Chemical Society* 124.2 (2002): 205-215.

99. DeLorbe, John E., et al. "Thermodynamic and Structural Effects of Conformational Constraints in Protein− Ligand Interactions. Entropic Paradoxy Associated with Ligand Preorganization." *Journal of the American Chemical Society* 131.46 (2009): 16758-16770.

100. Ward, Joshua M., et al. "Constraining binding hot spots: NMR and molecular dynamics simulations provide a structural explanation for enthalpy− entropy compensation in SH2− ligand binding." *Journal of the American Chemical Society* 132.32 (2010): 11058-11070.

101. DeLorbe, John E., et al. "Thermodynamic and structural effects of macrocyclic constraints in protein− ligand interactions." *ACS medicinal chemistry letters* 1.8 (2010): 448-452.

102. Rezai, Taha, et al. "Conformational flexibility, internal hydrogen bonding, and passive membrane permeability: successful in silico prediction of the relative permeabilities of cyclic peptides." *Journal of the American Chemical Society* 128.43 (2006): 14073-14080.

103. Rezai, Taha, et al. "Testing the conformational hypothesis of passive membrane permeability using synthetic cyclic peptide diastereomers." *Journal of the American Chemical Society* 128.8 (2006): 2510-2511.

104. Finch, Harry. "The conformational musings of a medicinal chemist." *Drug discovery today* 19.3 (2014): 320-325.

105. Foloppe, Nicolas, and I-Jen Chen. "Conformational sampling and energetics of drug-like molecules." *Current medicinal chemistry* 16.26 (2009): 3381-3413.

106. Blundell, Charles D., Thorsten Nowak, and Martin J. Watson. "Chapter Two-Measurement, Interpretation and Use of Free Ligand Solution Conformations in Drug Discovery." *Progress in medicinal chemistry* 55 (2016): 45-147.

107. Foloppe, Nicolas, and I-Jen Chen. "Towards understanding the unbound state of drug compounds: implications for the intramolecular reorganization energy upon binding." *Bioorganic & medicinal chemistry* 24.10 (2016): 2159-2189.

108. Kai, L., D. Torchia, and A. Bax. "Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to

staphyloccal nuclase." *Biochemistry* 28 (1989): 8972-8979.

109. Palmer 3rd, A. G., Christopher D. Kroenke, and J. Patrick Loria. "Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules." *Methods in enzymology*339 (2001): 204.

110. Mittermaier, Anthony, and Lewis E. Kay. "New tools provide new insights in NMR studies of protein dynamics." *Science* 312.5771 (2006): 224-228.

111. Lindorff-Larsen, Kresten, et al. "Picosecond to millisecond structural dynamics in human ubiquitin." *J. Phys. Chem. B* 120.33 (2016): 8313-8320.

112. Ban, David, et al. "Recent advances in measuring the kinetics of biomolecules by NMR relaxation dispersion spectroscopy." *Archives of Biochemistry and Biophysics* (2017).

113. Klepeis, John L., et al. "Long-timescale molecular dynamics simulations of protein structure and function." *Current opinion in structural biology* 19.2 (2009): 120-127.

114. Vettoretti, Gerolamo, et al. "Molecular dynamics simulations reveal the mechanisms of allosteric activation of Hsp90 by designed ligands." *Scientific reports* 6 (2016): 23830.

115. James, Thomas L. "Fundamentals of NMR." Online Textbook: Department of Pharmaceutical Chemistry, University of California, San Francisco (1998): 1-31.

116. Willoughby, Patrick H., Matthew J. Jansma, and Thomas R. Hoye. "A guide to small-molecule structure assignment through computation of (1H and 13C) NMR chemical shifts." *Nature protocols* 9.3 (2014): 643.

117. Dias, David M., and Alessio Ciulli. "NMR approaches in structure-based lead discovery: recent developments and new frontiers for targeting multi-protein complexes." *Progress in biophysics and molecular biology* 116.2 (2014): 101-112.

118. Lindon, John C., George E. Tranter, and David Koppenaal. *Encyclopedia of spectroscopy and spectrometry*. Academic Press, 2016.

119. Neuhaus, David, and Michael P. Williamson. "The nuclear Overhauser effect in structural and conformational analysis." (1989): 31.

120. Marion, Dominique. "An introduction to biological NMR spectroscopy." *Molecular & Cellular Proteomics*12.11 (2013): 3006-3025.

121. Hockney, R. W., S. P. Goel, and J. W. Eastwood. "Quiet high-resolution computer models of a plasma." *Journal of Computational Physics* 14.2 (1974): 148-158.

122. Case, D. A., et al. "AMBER 12; University of California: San Francisco, 2012." *There is no corresponding record for this reference*: 1-826.

123. Brooks, Bernard R., et al. "CHARMM: the biomolecular simulation program." *Journal of computational chemistry* 30.10 (2009): 1545-1614.

124. Jorgensen, William L., David S. Maxwell, and Julian Tirado-Rives. "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids." *J. Am. Chem. Soc*118.45 (1996): 11225-11236.

125. Wang, Junmei, et al. "Development and testing of a general amber force field." *Journal of computational chemistry* 25.9 (2004): 1157-1174.

126. Vanommeslaeghe, Kenno, et al. "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields." *Journal of computational chemistry* 31.4 (2010): 671-690.

127. Darden, Tom, Darrin York, and Lee Pedersen. "Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems." *The Journal of chemical physics* 98.12 (1993): 10089-10092.

128. Porto, Markus. "Ewald summation of electrostatic interactions of systems with finite extent in two of three dimensions." *Journal of Physics A: Mathematical and General* 33.35 (2000): 6211.

129. Andersen, Hans C. "Molecular dynamics simulations at constant pressure and/or temperature." *The Journal of chemical physics* 72.4 (1980): 2384-

2393

130. Berendsen, Herman JC, et al. "Molecular dynamics with coupling to an external bath." *The Journal of chemical physics* 81.8 (1984): 3684-3690.

131. Hoover, William G. "Canonical dynamics: equilibrium phase-space distributions." *Physical review A* 31.3 (1985): 1695.

132. Grest, Gary S., and Kurt Kremer. "Molecular dynamics simulation for polymers in the presence of a heat bath." *Physical Review A* 33.5 (1986): 3628.

133. Freddolino, Peter L., et al. "Ten-microsecond molecular dynamics simulation of a fast-folding WW domain." *Biophysical journal* 94.10 (2008): L75-L77.

134. Freddolino, Peter L., et al. "Microsecond Explicit Solvent Molecular Dynamics Simulations of Protein Folding." *Biophysical Journal* 96.3 (2009): 590a.

135. Pande, Vijay. "Protein folding studied with molecular dynamics simulation." *Encyclopedia of Biophysics*. Springer Berlin Heidelberg, 2013. 2016-2020.

136. Ngo, Son Tung, et al. "Replica exchange molecular dynamics study of the truncated amyloid beta (11–40) trimer in solution." *Physical Chemistry Chemical Physics* 19.3 (2017): 1909-1919.

137. Mostofian, Barmak, Xiaolin Cheng, and Jeremy C. Smith. "Replica-exchange molecular dynamics simulations of cellulose solvated in water and in the ionic liquid 1-butyl-3-methylimidazolium chloride." *The Journal of Physical Chemistry B* 118.38 (2014): 11037-11049.

138. Lu, Jianfeng, and Eric Vanden-Eijnden. "Infinite swapping replica exchange molecular dynamics leads to a simple simulation patch using mixture potentials." *The Journal of chemical physics* 138.8 (2013): 084105.

139. Kubitzki, Marcus B., and Bert L. De Groot. "Molecular dynamics simulations using temperature-enhanced essential dynamics replica exchange." *Biophysical journal* 92.12 (2007): 4262-4270.

140. Sugita, Yuji, and Yuko Okamoto. "Replica-exchange molecular dynamics method for protein folding." *Chemical physics letters* 314.1 (1999): 141-151.

141. Earl, David J., and Michael W. Deem. "Parallel tempering: Theory, applications, and new perspectives." *Physical Chemistry Chemical Physics*7.23 (2005): 3910-3916.

142. Nymeyer, Hugh. "How efficient is replica exchange molecular dynamics? An analytic approach." *Journal of chemical theory and computation* 4.4 (2008): 626-636.

143. Kofke, David A. "On the acceptance probability of replica-exchange Monte Carlo trials." *The Journal of chemical physics* 117.15 (2002): 6911-6914.

144. Nadler, Walter, and Ulrich HE Hansmann. "Dynamics and optimal number of replicas in parallel tempering simulations." *Physical Review E* 76.6 (2007): 065701.

145. Periole, Xavier, and Alan E. Mark. "Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent." *The Journal of chemical physics* 126.1 (2007): 01B601.

146. Bergonzo, Christina, et al. "Multidimensional replica exchange molecular dynamics yields a converged ensemble of an RNA tetranucleotide." *Journal of chemical theory and computation* 10.1 (2013): 492-499.

147. Okamoto, Yuko, et al. "α-Helix folding by Monte Carlo simulated annealing in isolated C-peptide of ribonuclease A." *Protein Engineering, Design and Selection* 4.6 (1991): 639-647.

148. Sanbonmatsu, K. Y., and A. E. Garcia. "Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics." *Proteins: Structure, Function, and Bioinformatics* 46.2 (2002): 225-234.

149. Rathore, Nitin, Manan Chopra, and Juan J. de Pablo. "Optimal allocation of replicas in parallel tempering simulations." *The Journal of chemical physics* 122.2 (2005): 024111.

150. Patriksson, Alexandra, and David van der Spoel. "A temperature predictor for parallel tempering simulations." *Physical Chemistry Chemical Physics*10.15 (2008): 2073-2077.

151. Sutherland, Jeffrey J., et al. "Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy." *Journal of chemical information and modeling* 47.6 (2007): 2293-2302.

152. Nithya, V. "Bioactive Compound Analysis of Coriandrum Sativum L against Microbial Keratitis." *Examining the Development, Regulation, and Consumption of Functional Foods* (2016): 109.

153. Liu, Kai, and Hironori Kokubo. "Exploring the Stability of Ligand Binding Modes to Proteins by Molecular Dynamics Simulations: A Cross-docking Study." *Journal of Chemical Information and Modeling* (2017).

154. Friesner, Richard A., et al. "Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein− ligand complexes." *Journal of medicinal chemistry* 49.21 (2006): 6177-6196.

155. ChemicalComputingGroup, M. O. E. "Molecular Operating Environment." (2008).

156. Jones, Gareth, et al. "Development and validation of a genetic algorithm for flexible docking." *Journal of molecular biology* 267.3 (1997): 727-748.

157. Morris, Garrett M., et al. "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility." *Journal of computational chemistry* 30.16 (2009): 2785-2791.

158. Phillips, Joshua L., Michael E. Colvin, and Shawn Newsam. "Validating clustering of molecular dynamics simulations using polymer models." *BMC bioinformatics* 12.1 (2011): 445.

159. Torda, Andrew E., and Wilfred F. van Gunsteren. "Algorithms for clustering molecular dynamics configurations." *Journal of computational chemistry*15.12 (1994): 1331-1340.

160. Shao, Jianyin, et al. "Clustering molecular dynamics trajectories: 1.

Characterizing the performance of different clustering algorithms." *Journal of Chemical Theory and Computation* 3.6 (2007): 2312-2334.

161. Salt, David W., et al. "DASH: a novel analysis method for molecular dynamics simulation data. Analysis of ligands of PPAR-γ." *Journal of medicinal chemistry* 48.9 (2005): 3214-3220.

162. Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010): 433-459.

163. Bro, Rasmus, and Age K. Smilde. "Principal component analysis." *Analytical Methods* 6.9 (2014): 2812-2831.

164. Constantin, Cristinel. "Principal Component Analysis-A Powerful Tool in Computing Marketing Information." *Bulletin of the Transilvania University of Brasov. Economic Sciences. Series V* 7.2 (2014): 25.

165. David, Charles C., and Donald J. Jacobs. "Characterizing protein motions from structure." *Journal of Molecular Graphics and Modelling* 31 (2011): 41-56.

166. Wolf, Antje, and Karl N. Kirschner. "Principal component and clustering analysis on molecular dynamics data of the ribosomal L11· 23S subdomain." *Journal of molecular modeling* 19.2 (2013): 539-549.

167. David, Charles C., and Donald J. Jacobs. "Principal component analysis: a method for determining the essential dynamics of proteins." *Protein Dynamics: Methods and Protocols* (2014): 193-226.

168. Abraham, M. J., et al. "The GROMACS development team GROMACS user manual version 5.0. 4." (2014): 50.

169. Hubbard, Roderick E., and Muhammad Kamran Haider. "Hydrogen bonds in proteins: role and strength." *eLS* (2010).

170. Baker, E. N., and R. E. Hubbard. "Hydrogen bonding in globular proteins." *Progress in biophysics and molecular biology* 44.2 (1984): 97-179.

171. Jeffrey, George A., and George A. Jeffrey. *An introduction to hydrogen bonding*. Vol. 32. New York: Oxford university press, 1997.

172. Groom, Colin R., et al. "The Cambridge structural database." Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials72.2 (2016): 171-179.

173. Bruno, Ian J., et al. "New software for searching the Cambridge Structural Database and visualizing crystal structures." *Acta Crystallographica Section B: Structural Science* 58.3 (2002): 389-397

174. Hawkins, Paul CD, et al. "Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database." *Journal of chemical information and modeling* 50.4 (2010): 572-584.

175. Hawkins, Paul CD, A. Geoffrey Skillman, and Anthony Nicholls. "Comparison of shape-matching and docking as virtual screening tools." *Journal of medicinal chemistry* 50.1 (2007): 74-82.

176. Haranczyk, Maciej, and John Holliday. "Comparison of similarity coefficients for clustering and compound selection." *Journal of chemical information and modeling* 48.3 (2008): 498-508.

177. Willett, Peter. "Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.3 (2011): 241-251

178. OpenEye Toolkits 2017.Jun.1 OpenEye Scientific Software, Santa Fe,NM. **https://www.eyesopen.com/**

179. Clayden, Jonathan, et al. "The challenge of atropisomerism in drug discovery." *Angewandte Chemie International Edition* 48.35 (2009): 6398-6401.

180. Wolf, Antje, and Karl N. Kirschner. "Principal component and clustering analysis on molecular dynamics data of the ribosomal L11· 23S subdomain." *Journal of molecular modeling* 19.2 (2013): 539-549.

181. Stein, Sarah A. Mueller, et al. "Principal components analysis: a review of its application on molecular dynamics data." *Annual Reports in Computational Chemistry* 2 (2006): 233-261.

182. Cattell, Raymond B. "The scree test for the number of factors." *Multivariate behavioral research* 1.2 (1966): 245-276.

183. McGibbon, Robert T., et al. "MDTraj: A modern open library for the analysis of molecular dynamics trajectories." *Biophysical journal* 109.8 (2015): 1528-1532.

184. Schrödinger, L. "PyMOL The PyMOL Molecular Graphics System." (2010).

185. Fujita, Toshio, and Corwin Hansch. "Analysis of the structure-activity relationship of the sulfonamide drugs using substituent constants." *Journal of medicinal chemistry* 10.6 (1967): 991-1000.

186. Grieco, Ciro, et al. "Quantitative structure-activity relationships of chymotrypsin-ligand interactions: An analysis of interactions in ϱ3 space." *Archives of biochemistry and biophysics* 194.2 (1979): 542-551.

187. Mengelers, M. J. B., et al. "Structure-activity relationships between antibacterial activities and physicochemical properties of sulfonamides." *Journal of veterinary pharmacology and therapeutics* 20.4 (1997): 276-283.

188. Liu, Pu, et al. "Replica exchange with solute tempering: A method for sampling biological systems in explicit water." *Proceedings of the National Academy of Sciences of the United States of America* 102.39 (2005): 13749-13754.

189. Smith, Amy K., Christopher Lockhart, and Dmitri K. Klimov. "Does Replica Exchange with Solute Tempering efficiently sample Aβ peptide conformational ensembles?." *Journal of chemical theory and computation* 12.10 (2016): 5201-5214.

190. Daintith, John, ed. *A dictionary of chemistry*. OUP Oxford, 2008.