

In silico study of *Plasmodium* 1-deoxy-d-
xylulose 5-phosphate reductoisomerase
(DXR) for identification of novel inhibitors
from SANCDB

A thesis submitted in partial fulfilment of the requirements for the degree

of

Master of Science in Bioinformatics and Computational Molecular Biology

(Coursework and Thesis)

of

RHODES UNIVERSITY, SOUTH AFRICA

Research Unit in Bioinformatics (RUBi)

DEPARTMENT OF BIOCHEMISTRY and MICROBIOLOGY

Faculty of Science

by

Bakary N'tji Diallo

January 2018

ABSTRACT

Malaria remains a major health concern with a complex parasite constantly developing resistance to the different drugs introduced to treat it, threatening the efficacy of the current ACT treatment recommended by WHO (World Health Organization). Different antimalarial compounds with different mechanisms of action are ideal as this decreases chances of resistance occurring. Inhibiting DXR and consequently the MEP pathway is a good strategy to find a new antimalarial with a novel mode of action..

From literature, all the enzymes of the MEP pathway have also been shown to be indispensable for the synthesis of isoprenoids. They have been validated as drug targets and the X-ray structure of each of the enzymes has been solved. DXR is a protein which catalyses the second step of the MEP pathway. There are currently 255 DXR inhibitors in the Binding Database (accessed November 2017) generally based on the fosmidomycin structural scaffold and thus often showing poor drug likeness properties.

This study aims to research new DXR inhibitors using *in silico* techniques. We analysed the protein sequence and built 3D models in close and open conformations for the different *Plasmodium* sequences. Then SANCDB compounds were screened to identify new potential DXR inhibitors with new chemical scaffolds.

Finally, the identified hits were submitted to molecular dynamics studies, preceded by a parameterization of the manganese atom in the protein active site.

DECLARATION

I, Bakary N'tji Diallo, declare that this thesis submitted to Rhodes University is wholly my own work and has not previously been submitted for a degree at this or any other institution.

The research described in this thesis was carried out as part of the one-year MSc coursework and research thesis programme in Bioinformatics and Computational Molecular Biology, from July 15th 2017 to January 23rd 2018 under the supervisions of Dr Kevin Lobb and Prof Ozlem Tastan Bishop.

Signature :

Date

ACKNOWLEDGEMENTS

Firstly, I would like to thank the Developing Excellence in Leadership and Genetics Training for Malaria Elimination in sub-Saharan Africa (DELGEME) for the scholarship to study bioinformatics and all the other financial support.

I acknowledge my supervisors, Pr. Özlem Tastan Bishop and Dr. Kevin Lobb. My sincere gratitude to Prof. Ozlem Tastan Bishop for assisting us to be at Rhode University and to be part of her group. And to Dr. Lobb for all the help and support and for making me appreciate computational chemistry throughout course work and project work. It was a great pleasure learning at your side.

Finally, special thanks to all RUBi colleagues, I thank all my colleagues at RUBi (Rhodes Bioinformatics Research Group) for their help and assistance.

This work was supported through the DELTAS Africa Initiative [grant 107740/Z/15/Z]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant #] and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

Table of content

CHAPTER 1: LITERATURE REVIEW.....	1
1.1 Introduction.....	1
1.2 Biology of Plasmodium.....	2
1.3 Isoprenoid biosynthesis in Plasmodium.....	4
1.1.1 Biochemistry of the non-mevalonate pathway.....	5
1.1.2 The non-mevalonate pathway as drug target.....	5
1.4 Catalytic Mechanism of 1-Deoxy-D-xylulose-5-phosphate reductoisomerase.....	6
1.1.3 1-Deoxy-D-xylulose-5-phosphate reductoisomerase (DXR).....	6
1.1.4 DXR Inhibition.....	9
1.5 Research problem statement and justification.....	18
1.6 Aims.....	19
1.7 Objectives.....	20
CHAPTER 2: SEQUENCE ANALYSIS.....	21
2.1 Introduction.....	21
2.1.1 Biological databases and information search.....	22
2.1.2 Sequence alignment.....	22
2.1.3 Phylogenetic trees.....	24
2.2 Methodology.....	24
2.2.1 Data retrieval.....	24
2.2.2 Multiple Sequence Alignment.....	25
2.2.3 Phylogenetic analysis.....	25
2.2.4 Motifs analysis.....	26
2.3 Results and Discussion:.....	26
2.3.1 Sequence data.....	26
2.3.2 Multiple Sequence Alignment.....	28
2.3.3 Phylogenetic tree.....	30
2.3.4 Motifs analysis.....	31
2.4 Conclusion.....	33
CHAPTER 3: HOMOLOGY MODELLING.....	34
3.1 Introduction.....	34
3.2 Steps in homology modeling.....	35

3.3	Methodology	38
3.3.1	Template identification	38
3.3.2	Template-target alignment	39
3.3.3	Homology modeling	39
3.3.4	Model evaluation	40
3.4	Results and Discussion.....	40
3.4.1	Template identification	40
3.4.2	Template-Target alignment.....	46
3.4.3	Modeling and Model Evaluation	47
3.5	Conclusion	52
CHAPTER 4:	MOLECULAR DOCKING	53
4.1	Introduction	53
4.2	Docking strategies: AUTODOCK4 and AUTODOCK VINA.....	54
4.3	Methodology	55
4.3.1	Ligand preparation.	55
4.3.2	Receptor preparation	56
4.3.3	Molecular Docking.....	56
4.3.4	Docking validation	57
4.3.5	Analysis and hit identification	59
4.4	Results and Discussion.....	61
4.4.1	Ligand preparation.	61
4.4.2	Receptor preparation.	61
4.4.3	Docking validation	63
4.4.4	Analysis.....	64
4.5	Conclusion	90
CHAPTER 5:	MOLECULAR DYNAMICS	91
5.1	Introduction	91
5.1.1	System preparation	95
5.1.2	Defining simulation box and solvation	96
5.1.3	Energy minimization.....	97
5.1.4	Equilibration	97
5.1.5	MD simulation	98
5.1.6	Analysis.....	98

5.2	Methodology	98
5.2.1	Metal parameterization.	98
5.2.2	Implementation in GROMACS.....	101
5.2.3	Steps of Molecular dynamics	102
5.3	Results and Discussion.....	104
5.3.1	Metal parameterization	104
5.3.2	Implementation in GROMACS.....	111
5.3.3	Molecular Dynamics.....	112
5.4	Conclusion	122
	Future work.....	124
	REFERENCES	125
	APPENDIX	148

List of figures

Figure 1-1: Time to first detection resistances to antimalarials (Kennedy and Read 2017).....	2
Figure 1-2: Plasmodium life cycle. Adaptated from (Winzeler 2008)	4
Figure 1-3: PfDXR domains and residue numbers.....	7
Figure 1-4: Subunit structure of fosmidomycin-bound quaternary complex of PfDXR.....	7
Figure 1-5: The overall three-dimensional structure of PfDXR (Umeda et al. 2011).....	8
Figure 1-6: The alpha-ketol rearrangement (A) and the retroaldol (B) mechanism.....	9
Figure 1-7: Fosmidomycin chemical structure (Wiesner et al. 2016)	9
Figure 1-8: Metal chelation	11
Figure 1-9: Left: Fosmidomycin complex with PfDXR. The carbon atoms of fosmidomycin (in yellow), the four buried water molecules (cyan), and the bound Mg ²⁺ ion (green).	12
Figure 1-10: Fosmidomycin and FR-900098 structures (Wiesner et al. 2016).	14
Figure 1-11: Important structural features of fosmidomycin.	15
Figure 1-12: Some DXR inhibitors	18
Figure 2-1: MSA of Plasmodium DXR and it homologs..	28
Figure 2-2: Phylogenetic Tree and sequence identity heatmap.	30
Figure 2-3: Motif analysis results. The coloring shows the degree of conservation of the motif (red: highly conserved to bleu: less conserved). The related motif logos are in appendix A.....	32
Figure 2-4: Rossman fold (GXXGXXG) motif found in motif 4 of the MEME (Bailey et al. 2006) analysis.	33
Figure 3-1: Molecular overlay of 5JAZ (chain B, closed loop with inhibitor in active site) and 1K5H (chain A – open loop conformation). Arrows show the orientation of the loop.....	41
Figure 3-2: 5JAZ PDB metrics for structure quality	43
Figure 3-3: 5JAZ, Chain B Assessment.....	44
Figure 3-4: 1K5H, Chain A assessment.	45
Figure 3-5: 1K5H PDB Percentile Ranks.....	46
Figure 3-6: Graphical representation of the PIR file (target-template: PvDXR-5JAZ) alignment viewed in Jalview (Waterhouse et al. 2009)..	46
Figure 3-7: Graphical representation of the PIR file (target-template: PfDXR-1K5H) alignment viewed in Jalview (Waterhouse et al. 2009).	47
Figure 3-8: Effect of SCWRL(Krivov, Shapovalov, and Dunbrack 2009) on MODELLER (Šali et al. 2017) Dope-Z score.	51
Figure 3-9: Final models superimposed in Discovery Studio 2016 (Biovia, San Diego, CA).....	52
Figure 4-1: Setting of the grid box on both cofactor binding site (residues in bleu) and active site (showed with ligand in stick).....	57
Figure 4-2: Hit selection process.	61
Figure 4-3: Bisubstrate inhibition approach.	63
Figure 4-4: Molecular overlay Original LC5 (color by element) in the crystal structure and redocked LC5 (in light bleu). RMSD= 0.58 Å.....	63
Figure 4-5: Predicted Binding Energies by X-score and Autodock Vina.....	64
Figure 4-6: Histogram of binding energies.....	65

Figure 4-7: Heatmap of the binding energies for DXR in closed conformation (targetted docking of the protein active site).	67
Figure 4-8: Best binding poses for fosmidomycin FR98 in the protein active site, metal ion in black.	68
Figure 4-9: Binding energies in rigid and flexible receptors.	69
Figure 4-10: 2D plots of SANC00191 binding poses in flexible (left) and rigid (right) receptors.	70
Figure 4-11: 2D plot SANC00303 docked in the rigid crystal structure 5JAZ.	71
Figure 4-12: Left to Right: SANC00302 (3,6-Dibromoindole), SANC00303 (6-Bromo-3-chloroindole), SANC00304 (6-Bromo-2-oxindole).	71
Figure 4-13: Left: Open conformation with three openings of the large active/cofactor binding site. Right: Clustering of compounds in blind docking of DXR open conformation.	72
Figure 4-14: Cofactor in PfDXR active site.	73
Figure 4-15: Clustering of compounds in the closed conformation.	73
Figure 4-16: Compounds binding on DXR active site opposite face.	74
Figure 4-17: Reference ligands in blind docking on open conformation: 2D poses and binding energies.	75
Figure 4-18: Binding energies in DXR closed and open conformation in blind docking.	76
Figure 4-19: Top 10 compounds binding better in open conformation.	77
Figure 4-20: Compounds identified as potential bisubstrate inhibitor.	79
Figure 4-21: Phosphonate binding pocket (red), its adjacent pocket (green) and the third pocket (blue).	81
Figure 4-22: 2D plot and fit in the active site for identified hits.	85
Figure 5-1: Simplified diagram of MD simulation. Adaptated from (Badrinarayan, Choudhury, and Sastry 2015).	93
Figure 5-2 Header of the .com input file for Gaussian	99
Figure 5-3: Footer of the .com input file for Gaussian	100
Figure 5-4: Left optimized subset. Right geometry in the crystal structure (5JAZ chain B). The backbones of the residues were removed for clarity.	104
Figure 5-5: Water molecule and Mn coordination in the crystal structure.	105
Figure 5-6: Least square error fitting for PES data for bond 44-61.	107
Figure 5-7: Least square error fitting for PES data from Angle 9-61-52.	107
Figure 5-8: Implication of the step zise in PES scan.	109
Figure 5-9: Example of energy variation during PES scan on the ONIOM system.	110
Figure 5-10: Force field parameters files modified and their modifications.	112
Figure 5-11: Temperature variation during NVT equilibration. The legends display the SANCDB ID of the different ligands in the different systems.	113
Figure 5-12: Pressure variation during NPT equilibration.	114
Figure 5-13: Bond distances and protein RMSD for force fied parameters validation.	115
Figure 5-14: Complex 5JAZ-SANC00152 during simulation	117
Figure 5-15: Complex 5JAZ-SANC00236 during simulation	118
Figure 5-16: Complex 5JAZ-SANC00339 during simulation	119
Figure 5-17: Complex 5JAZ-SANC00438 during simulation	120
Figure 5-18: Complex 5JAZ-SANC00570 during simulation	121

List of tables

Table 1-1: Some PfDXR residues and their identified/suggested roles in inhibition.	13
Table 2-1: Models and site coverage cut-offs for phylogenetic tree construction.....	25
Table 2-2: Plasmodium DXR sequences retrieved from PlasmoDB. New ID is the ID used in this thesis.	27
Table 2-3: Plasmodium DXR homologs sequences retrieved from NCBI.....	27
Table 3-1: Metrics for modeling from 1K5H obtained using NCBI BLASTp search (Word size: , Expect value: 10, Hitlist size: 100, Gapcosts: 11.1, Matrix: BLOSUM62).	42
Table 3-2: Selected hits from HHpred result.....	42
Table 3-3: 5 Best Models for each sequence in Open conformation according to Dope-Z score. The best Q-mean score is highlighted in green.....	48
Table 3-4: 5 Best Models for each sequence in Closed conformation according to Dope-Z score. The best Q-mean score is highlighted in green. (PfDXR_closed is absent from this table as it already has a crystal structure 5JAZ).....	49
Table 3-5: Complete table of all final models' evaluation.	50
Table 4-1: Summary of the different docking experiments.....	58
Table 4-2: X-score and Vina predicted binding energies for LC5 for docking validation.....	64
Table 4-3: Top compounds showing highest difference in binding energy and their interacting residues. In bold, residues set as flexible.....	70
Table 4-4: Interacting residues with inhibitors in closed and open conformation. Red: residues interaction with fosmidomycin phosphonate. Green: residues interaction with fosmidomycin hydroxamate.....	75
Table 4-5: Identified bisubstrates molecular interactions.....	80
Table 4-6: Identified hits.....	80
Table 4-7: Hits molecular interactions. Red: residues interaction with fosmidomycin phosphonate. Green: residues interaction with fosmidomycin hydroxamate.....	86
Table 4-8: Preselected compounds.....	88
Table 4-9: Drug-like properties of top SANCDB compound according to QED score calculated using FAF Drugs4.....	89
Table 4-10: Hits drug likeness.....	89
Table 5-1: GROMACS units used and their conversion.....	102
Table 5-2: Bond lengths in initial crystal and optimized structure.....	105
Table 5-3: Angles in initial crystal and optimized structure.....	106
Table 5-4: Force force parameters derived from least-square fitting of PES data.....	107
Table 5-5: Parameters converted to GROMACS compliant units.....	111
Table 5-6: Potential energies and maximum forces attained during minimization.....	113

List of web servers and applications

MEME: (Multiple EM for Motif Elicitation) (Bailey et al. 2009)

<http://www.meme.nbcr.net/meme/cgi-bin/meme.cgi>

PROMALS3D: <http://prodata.swmed.edu/promals3d/promals3d.php>,

MUSCLE: <http://www.ebi.ac.uk/Tools/msa/muscle/>

SANCDDB: <https://sancdb.rubi.ru.ac.za/>.

The HHpred Interactive Server for Protein Homology Detection and Structure Prediction:

<https://toolkit.tuebingen.mpg.de>

QMEAN: <https://swissmodel.expasy.org/qmean/>

SWISS-MODEL assessment Workspace : <https://swissmodel.expasy.org/workspace>

The Binding Database: <https://www.bindingdb.org/bind/index.jsp>

ProQ3/ProQ3D: <http://proq3.bioinfo.se/>

NCBI BLAST: <http://blast.ncbi.nlm.nih.gov/>

Protein Data Bank RCSB: <https://www.rcsb.org/>

List of acronyms

3D:	Three-dimensional
ACT:	Artemisinin Combination Therapy
ADME:	Absorption, distribution, metabolism, and excretion
CDP-ME:	Methylerythritol cytidyl diphosphate
CTP:	Cytidine 5'-triphosphate
D-GLP:	D-glyceraldehyde-3-phosphate
DMAPP:	Dimethylallyl diphosphate, Dimethylallyl diphosphate
DOPE:	Discrete Optimized Protein Energy
DRL:	1-deoxy-d-xylulose 5-phosphate reductoisomerase-like
DXP:	1-deoxy-D-xylulose-5-phosphate, 1-deoxy-d-xylulose 5-phosphate
DXR:	1-deoxy-d-xylulose 5-phosphate reductoisomerase
Fos:	Fosmidomycin
GROMACS:	GRONingen MACHine for Chemical Simulations
Hbond	Hydrogen bonds
HMBPP:	4-hydroxy-3-methylbutenyl 1-diphosphate
IPP:	Isopentenyl diphosphate
LR:	Linker Region
MD:	Molecular Dynamics
MEME:	Multiple EM for Motif Elicitation
MEP:	2-C-methyl-d-erythritol 4-phosphate
MUSCLEL	MULTiple Sequence Comparison by Log-Expectation
MVA:	Mevalonate pathway
NADH:	Nicotinamide adenine dinucleotide, 20
NADPH:	Nicotinamide adenine dinucleotide phosphate
NCBI:	National Center for Biotechnology Information
PES:	Potential Energy Surface.
PDB:	Protein Data Bank

PROCHECK: PROgram to CHECK the stereochemical quality of protein structures

PROMALS3D: Profile Multiple Sequence Alignment with Local Structure and 3D constraints

QSAR: Quantitative structure–activity relationship

RMSD Root Mean Square Deviation

RMSF Root Mean Square Fluctuation

Rg: Radius of gyration

SANCDDB: South Africa Natural Compounds Database

SAR: Structure Activity Relationship

SCWRL: Side Chains With Rotamer Library

WHO: World Health Organization

CHAPTER 1: LITERATURE REVIEW

1.1 Introduction

Malaria is a disease caused by a protozoan parasite of the genus *Plasmodium*. There are four main species of *Plasmodium* that cause the disease in human: *P. falciparum*, *P. vivax*, *P. ovale* and *P. malariae* (Control et al. 1991). Rare cases of the simian parasite *P. knowlesi* causing human malaria in Southeast Asia have been reported (Barber et al. 2017). These parasites are transmitted to human by the bite of the female anopheles mosquitoes (Winzeler 2008). The clinical manifestations of malaria include fever, headache, nausea and vomiting, diarrhea and abdominal pain (Crutcher and Hoffman 1996).

Considerable advancements have been made in the fight against malaria. According to the WHO (World Health Organization) report of the year 2016 on malaria, the disease incidence rate has fallen by 21% in the world during the period 2010-2015. The global mortality rates fell by an estimated 29% globally and by 31% in the African region (WHO | World Malaria Report 2016 n.d.).

Despite these advances, many challenges remain to be addressed as malaria remains a major health concern. The same report indicates a global tally of malaria of 212 million new cases and 429 000 deaths in 2015 (WHO | World Malaria Report 2016 n.d.). The main burden of malaria is carried by the Sub-Saharan region of Africa which accounts for 90% of malaria and 92% of malaria deaths. The most vulnerable segment of the population is children under the age of five years with an estimated 70% of all malaria deaths (WHO | Malaria Control Improves for Vulnerable in Africa, but Global Progress off-Track n.d.).

The parasite's ability to develop resistance is another key concern. There has been a continual development of resistance to each class of drugs introduced to fight malaria (see Figure 1-1): quinine, chloroquine, proguanil, sulfadoxine-pyrimethamine, mefloquine, atovaquone (Cui et al. 2015). Artemisinin-based combination therapies (ACTs) are currently the WHO recommended drugs for malaria treatment (WHO | Overview of Malaria Treatment n.d.). The use of combining drugs with different modes of action greatly decrease the chances of resistance occurring (White 1999), requiring thus various antimalarial compounds with orthogonal mechanisms (Lunev et al. 2016).

Cases of resistance to artemisinin have been recorded in five countries in Southeast Asia: Cambodia, Laos, Myanmar, Thailand and Viet Nam. The spread of this resistance to other regions could severely impact all the previous effort done to fight the disease. Separate to antimalarial treatment is the plasticity of the mosquitoes, which may lead to the development of resistance to insecticides remains also another major concern (Hemingway et al. 2016; WHO | 10 Facts on Malaria n.d.).

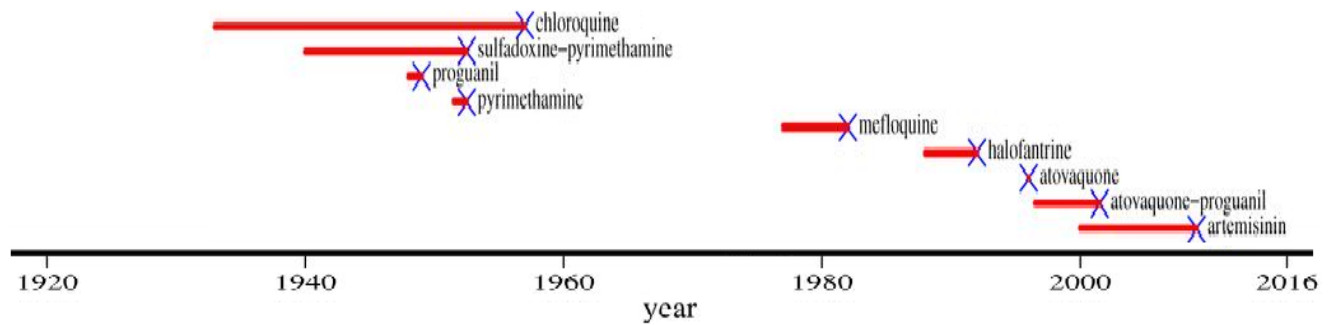


Figure 1-1: Time to first detection resistances to antimalarials (Kennedy and Read 2017). (Reproduced with permission).

In vaccine research, intense efforts have been made by diverse groups during the last two decades. The most advanced one, RTS,S/AS01, an anti-sporozoite has progressed through phase III clinical trials (Cowman et al. 2016). In spite of all these efforts, there is still no licensed vaccine for malaria (WHO | Tables of Malaria Vaccine Projects Globally n.d.). These diverse reasons show the urgent need of new antimalarial drugs to always keep a step ahead of the resistance curve (Hemingway et al. 2016).

1-Deoxy-d-xylulose 5-phosphate reductoisomerase of *P. falciparum* (PfIspC or PfDXR) is currently one of the clinically validated malaria drug targets (Konzuch et al. 2014). The enzyme is involved in the nonmevalonate pathway. It converts 1-deoxy-d-xylulose 5-phosphate (DXP) to 2-C-methyl-d-erythritol 4-phosphate (MEP), thus playing a key role in the production of isoprenoid precursors (Cobb et al. 2015). DXR is also essential to different phases of the parasite's life cycle (Saggu et al. 2016).

Thus, due to its uniqueness and its key role for the parasite, it has become an attractive drug target in the fight against malaria. The natural product fosmidomycin, a promising antimalarial drug is an inhibitor of PfDXR. However, fosmidomycin shows poor pharmacokinetic properties hampering its usage. During the past years, several research groups focused on the development of fosmidomycin analogues offering better drug-like properties (Saggu et al. 2016).

1.2 Biology of *Plasmodium*

Plasmodium spp are unicellular eukaryotes and belong to the large phylum of protozoan parasites, the apicomplexan. The *Apicomplexa* are characterized by the apicoplast, an essential organelle responsible for the synthesis of key molecules like isoprenoids and fatty acids required for the growth of the parasite. They are obligate intracellular parasites interacting with diverse hosts (Wirth 2002; Morrissette and Sibley 2002).

Plasmodium spp causing human malaria have a complex life cycle (see Figure 1-2) that are characterized by distinct phases and host-parasite interactions (Antinori et al. 2012). Through this cycle, the parasite moves between the mosquito vector and the human host, in three phases: the liver phase, the blood phase, and the mosquito phase (Control et al. 1991).

The parasite is first inoculated to the human through the bite of the mosquito vector. From the salivary glands of the mosquito, thousands of sporozoites are released into the blood stream. These sporozoites then migrate to the host liver (Control et al. 1991). They invade the host's hepatocytes and commence an endogenous asexual multiplication also known as schizogony, which takes 5 to 15 days depending on the specie of *Plasmodium* (Antinori et al. 2012). The sporozoites mature into schizonts containing about 10,000 merozoites in *P. vivax/P. ovale* and up to 30,000 merozoites in *P. falciparum*. In the case of *P. vivax*, and *P. ovale*, some sporozoites differentiate to the hypnozoite form, a latent form, that can proliferates days to years later leading to a relapse (Soulard et al. 2015; Campo et al. 2015). The mature schizont ruptures (together with the infected hepatocytes) releasing merozoites into the blood stream (Antinori et al. 2012).

During the blood phase, the merozoites invade the erythrocytes where they multiply asexually (within 48 to 72 hours). In the erythrocytes, the merozoites take different forms: rings, trophozoites, schizonts (Biamonte, Wanner, and Le Roch 2013). Each schizont contains about 6 to 36 merozoites. The rupture of erythrocytes releases into the blood these merozoites which would infect other erythrocytes. The clinical symptoms of the disease are visible during this stage (Biamonte, Wanner, and Le Roch 2013; Antinori et al. 2012). During this erythrocytic cycle, some parasites further differentiate into male and female gametocytes (Biamonte, Wanner, and Le Roch 2013).

These gametes can then be ingested by a female anopheline mosquito (Bennink, Kiesow, and Pradel 2016; Antinori et al. 2012) corresponding to the transmission of the parasite to the mosquito. The parasites can now develop into their sexual forms, the female macrogametes and male microgametes, and then commence sexual reproduction (Bennink, Kiesow, and Pradel 2016). During this mosquito phase, the fusion of the macrogamete with a single microgamete results in fertilization and the formation of the ookinete (Control et al. 1991). The ookinete converts to oocysts, in which sporogonic replication takes place. This takes roughly 2 weeks and results in the formation of infective sporozoites. These sporozoites will migrate to the salivary glands and will be released into the human dermis during the next blood meal of the mosquito. The life-cycle of the *Plasmodium* will be thus completed (Bennink, Kiesow, and Pradel 2016).

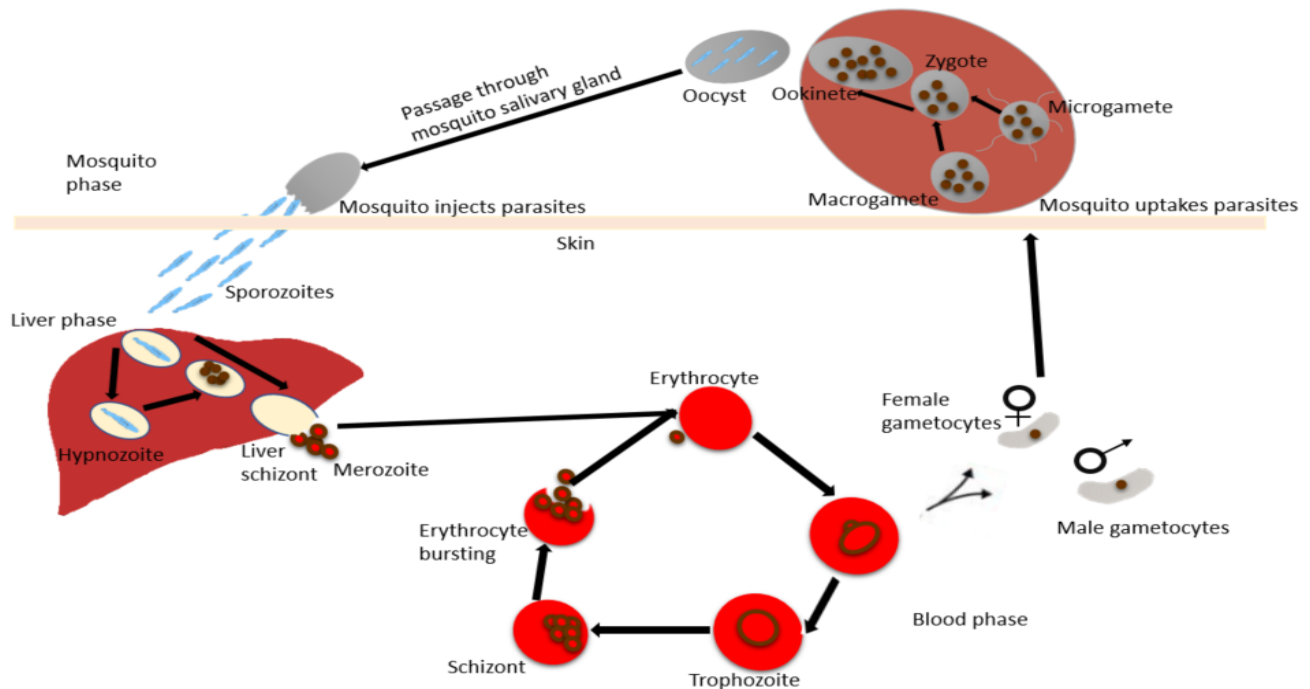


Figure 1-2: *Plasmodium* life cycle. Adaptated from (Winzeler 2008)

1.3 Isoprenoid biosynthesis in Plasmodium

Also known as terpenoids, isoprenoids constitute one of the largest classes of biological compounds and they are ubiquitous in all domains of life: bacteria, archaea, and eukaryotes (Chang et al. 2013). They include the carotenoids as photosynthetic biopigments, the sterols as cell membrane components, the steroid hormones for the regulation of growth and development, the quinones involved in the electron transport chain, dolichol in glycoprotein and bacterial cell wall biosynthesis, linear prenyl diphosphates as protein prenylation units for intra-cellular protein targeting, and PfPRL prenylation for erythrocyte invasion (Sacchetti and Poulter 1997; Holstein and Hohl 2004; van der Meer and Hirsch 2012).

All this structural and functional diversity is derived from two precursors: isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP) (Biamonte, Wanner, and Le Roch 2013). To produce IPP and DMAPP, two main pathways exist: MVA pathway (mevalonate pathway or the isoprenoid pathway or HMG-CoA reductase pathway) and the MEP pathway (methylerythritol phosphate or non-mevalonate pathway). Two other pathways discovered from late 1990: a modified MVA pathway and the 5-Methylthioadenosine shunt pathway also exist (Chang et al. 2013).

The mevalonate (MVA) pathway was first discovered in the 1950s. Over the decades, it was accepted as the main source of IPP and DMAPP (Chang et al. 2013). In the 1990s, eventually, the groups of Arigony and Rohmer independently showed the existence of another route for isoprenoid biosynthesis, the non-mevalonate pathway (Rohmer et al. 1993; van der Meer and Hirsch 2012).

1.1.1 Biochemistry of the non-mevalonate pathway

The MEP pathway is comprised of seven enzymatic steps.

The first step in the non-mevalonate pathway involves the condensation of pyruvate (Pyr) and D-glyceraldehyde-3-phosphate (D-GLP) to form 1-deoxy-D-xylulose-5-phosphate (DXP) and CO₂. This first reaction of the pathway is catalysed by 1-deoxy-D-xylulose-5-phosphate synthase (DXS).

The following reaction is the NADPH-dependent reductive rearrangement of DXP to 2-C-methyl-D-erythritol 4-phosphate (MEP). It is catalysed by D-xylulose-5-phosphate reductoisomerase (DXR also known as IspC) (Chang et al. 2013). In the next reactions, MEP is coupled with cytidine 5'-triphosphate (CTP) to produce methylerythritol cytidyl diphosphate (CDP-ME, 14). The step is catalysed by CDP-ME synthetase (IspD) (Zhao et al. 2013).

This product is then phosphorylated by IspE to produce 4-diphosphocytidyl-2-C-methyl-d-erythritol-2-phosphate (CDP-MEP) which will be cyclized by IspF to 2-C-methyl-D-erythritol-2,4-cyclodiphosphate (MEcPP) (Chang et al. 2013).

Then, we have the ring-opening and reductive dehydration of MEcPP to produce 4-hydroxy-3-methylbutenyl 1-diphosphate (HMBPP) catalysed by IspG. And finally, IspH catalyses HMBPP by reductive dehydration to produce both IPP and DMAPP (Imlay and Odom 2014).

1.1.2 The non-mevalonate pathway as drug target

The building blocks isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) for isoprenoids are synthesized in the apicoplast of *Plasmodium spp* (Wiley et al. 2015). The Apicomplexans do not have the MVA pathway and rely entirely on the MEP pathway to generate isoprenoids (Odom 2011). The first evidence for MEP pathway in *Plasmodium* and its presence in the apicoplast by Jomaa et al. in 1999, and its validity as a drug target was also proven. DOXP reductoisomerase was also shown to be a key enzyme of this pathway as its inhibition by fosmidomycin and its derivative FR-900098 showed antimalarial activity *in vivo* and *in vitro* (Jomaa et al. 1999). The existence of the natural product fosmidomycin attracted much attention toward the DXR enzyme, and made it the most widely cited therapeutic target in the MEP pathway (Hale et al. 2012). Still, its poor pharmacokinetic properties are hampering its usage. Fosmidomycin has been shown to be only effective for short-term treatment. Researches on finding a potential partner drug to improve the efficiency of fosmidomycin are needed (Bhagavathula, Elnour, and Shehab 2016).

Up to now, several DXR inhibitors have been identified. The natural acetyl derivative of fosmidomycin, FR900098, has shown better activity (IC₅₀ = 0.018 μM) than fosmidomycin. It also shows a good toxicity profile, supporting its future development as an antimalarial drug (Wiesner et al. 2016).

More studies could help in the development of fosmidomycin derivatives and similar compounds, to be used as anti-malarial (Saggu et al. 2016). Also, the combination with a potential drug partner

is another area of research. Fosmidomycin-clindamycin and fosmidomycin-piperaquine are in Phase II clinical trials (Mishra et al. 2017).

The other enzymes of the pathway have also been shown to be indispensable for the synthesis of isoprenoids. As the MEP pathway is linear, each enzyme is essential for isoprenoid biosynthesis (Odom 2011). Several research groups studied these enzymes identify/design probable inhibitors. They have been validated as drug targets and the X-ray structure of each of the enzymes has been solved. But most inhibitors have been shown to be weak inhibitors, or do not possess drug like characteristics (Hale et al. 2012). In view of the essentiality and the uniqueness of the pathway for the survival of the parasite, more extensive studies on the remaining enzymes of the pathway are needed (Saggu et al. 2016).

In addition, the pathway is indispensable at different stages of the parasite life cycle. During the asexual phase, it has been reported to be essential for both liver and blood stages. In recent studies, products of the pathway are also required during for the early phases of parasite gamete development, thus showing the pathway as a valid drug target for the development of malaria transmission-blocking inhibitors (Saggu et al. 2016).

The MEP pathway is indispensable for most eubacteria, and none of its enzymes has a homolog in human or other mammalian cells. Many human pathogens, *Escherichia coli*, *Mycobacterium tuberculosis*, *Mycobacterium leprae*, *Helicobacter pylori*, *Vibrio cholera*, *Bacillus anthracis* rely exclusively on the MEP pathway for the biosynthesis of their isoprenoid compounds (Murkin, Manning, and Kholodar 2014; van der Meer and Hirsch 2012). With the problem of drug resistance in Gram-negative bacteria, *M. tuberculosis*, and *P. falciparum*, inhibition of the MEP pathway hold great potential for broad-spectrum agents, and their use in combinational treatments (Odom 2011).

Nevertheless, a distinct DXR-like (DRL) protein can also catalyse the same reaction to produce MEP in some MEP dependant organisms. This includes the human and animal pathogens *Bartonella* and *Brucella* (Pérez-Gil et al. 2012; Sangari et al. 2010).

Elsewhere, the MEP pathway is exclusively responsible for IPP, DMAPP, and the isoprenoids in a number of problematic weeds. So the enzymes of the pathway can also be used as targets for the development of novel herbicide (van der Meer and Hirsch 2012).

1.4 Catalytic Mechanism of 1-Deoxy-D-xylulose-5-phosphate reductoisomerase

1.1.3 1-Deoxy-D-xylulose-5-phosphate reductoisomerase (DXR)

1.1.3.1 Structure

Encoded in the *ispC* gene, DXR is the most studied drug target in the MEP pathway. There are more than thirty published crystal structures of DXR from different organisms, including *P. falciparum*, *M. tuberculosis* and *E. coli*. The general structure of PfDXR is comparable to the structure of the enzyme in other species. In its active form (Lys75 to Ser488), PfDXR is a homodimer in a V shape with a molecular mass of approximately 47 kDa. Each monomer contains an NADPH molecule and a divalent metal ion (Mg^{2+} , Co^{2+} or Mn^{2+}) required for the catalytic activity of the enzyme (Umeda et al. 2011).

Each monomer has two large domains, a linker region, and a small C-terminal domain. The two large domains are separated by a cleft containing a deep pocket. One of them will bind NADPH, and the other domain provides the groups necessary for catalysis (metal and substrate binding). The NADPH domain contains 154 residues (residues 77 to 230) (see Figure 1-3), while the catalytic domain, in the centre of the V shape, covers 139 residues (residue 231 to 369). The residues' numbers refer to the *P. falciparum* sequence. The linker region spans from residue 370 to 395 and the C-terminal domain from residue 396 to 486. The first 74 residues are similar to an endoplasmic reticulum signal (the first 30 residues) and plastidial targeting sequences (for the next 44 amino acids) (Umeda et al. 2011). A flexible loop region (residues 291 to 299) is inserted in the catalytic domain. Apart from for Pro294, buried residues (His293, Trp296, and Met298) in this region are completely conserved (Umeda et al. 2011).

Comparative studies between different DXR structures revealed three conformations: the open form with the loop opened (no substrate/inhibitor), the open form with the flexible loop closed (with substrate/inhibitor, prepared by soaking), and the configuration with the flexible loop covering the active site (with substrate/inhibitor, prepared by co-crystallization) (Takenoya et al. 2010). The flexible loop undergoes a movement of about 16 Å to close the active site, ordering and acting as a lid isolating the active site from bulk solvent after inhibitor binding (Kholodar and Murkin 2013).

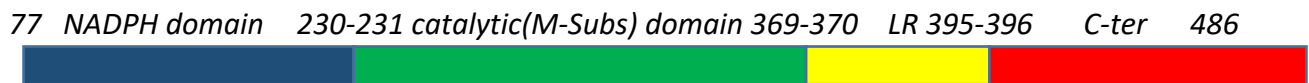


Figure 1-3: PfDXR domains and residue numbers. NADPH binding domain (residues 77-230), catalytic (Metal and Substrate binding, residues 231-369), Linker region (residues 370-395), C-terminal domain (396-486).

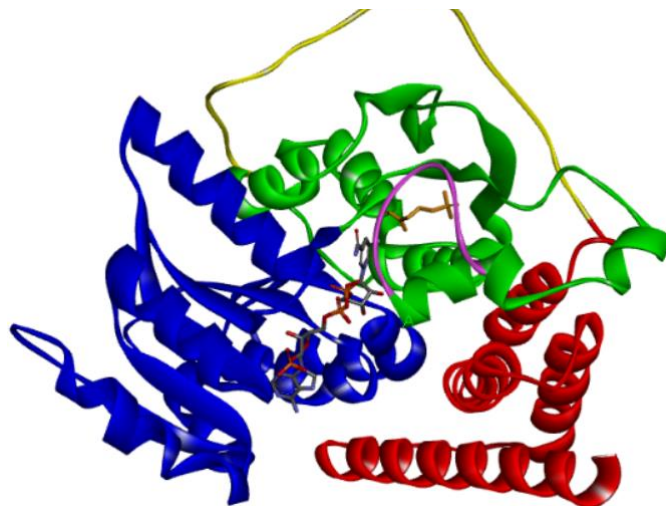


Figure 1-4: Subunit structure of fosmidomycin-bound quaternary complex of PfDXR.

Subunit structure of fosmidomycin-bound quaternary complex of PfDXR. The subunit structure of the fosmidomycin-bound quaternary complex of PfDXR. The NADPH-binding (blue), catalytic (green), linker

(yellow), active site flap (pink) and C-terminal domains (red). The bound fosmidomycin (gold) and NADPH (grey and red) molecules are shown as ball-and-stick models. Adapted from (Umeda et al. 2011).

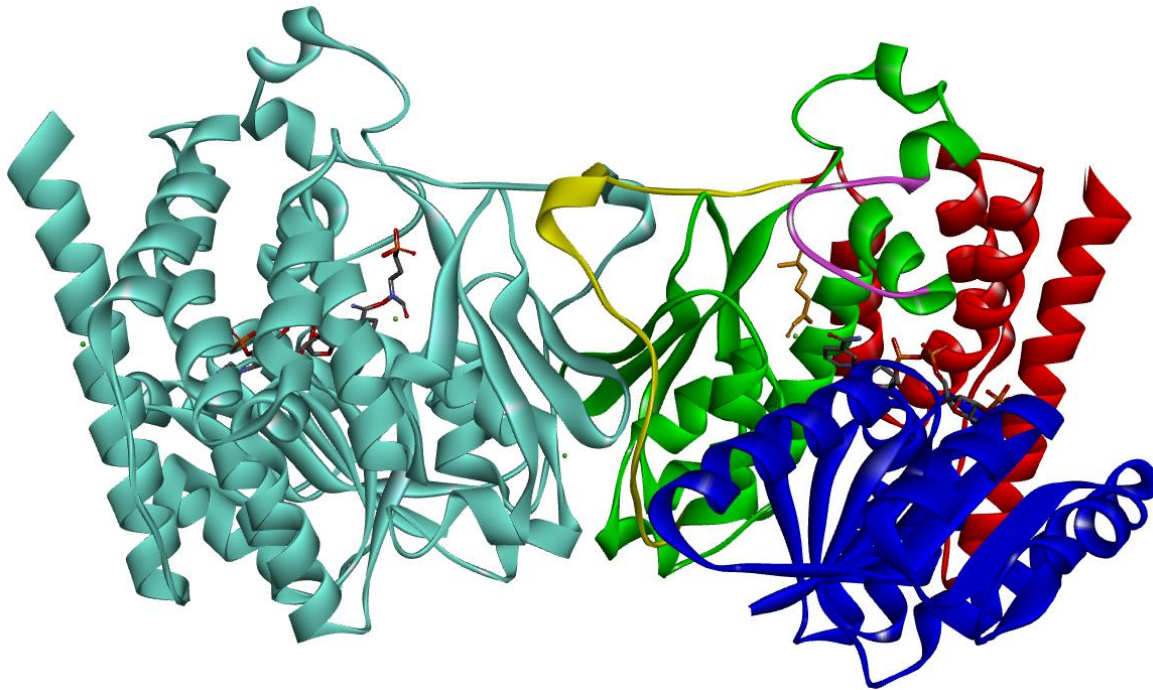


Figure 1-5: The overall three-dimensional structure of PfDXR (Umeda et al. 2011). The overall structure of PfDXR. One monomer is colored in cyan the other as in Figure 1-4. Adapted from PDB structure 3AU9 (Umeda et al. 2011). Each monomer contains fosmidomycin, the cofactor NADPH and the metal ion.

The NADPH binding domain (in blue Figure 1-4 and Figure 1-5) at the N-terminal region is a member of dinucleotide binding fold known as Rossmann fold, composed of two $\beta\alpha\beta\alpha\beta$ units. In the PfDXR NADPH-binding domain, an additional $\alpha\beta$ motif is inserted after β_3 . This domain consists of a seven-stranded β -sheet in the centre of the domain sandwiched by two arrays of three α -helices (Umeda et al. 2011; Saggu et al. 2016).

Then we have the catalytic domain (in green Figure 1-4 and Figure 1-5). It comprises the binding sites for the inhibitor fosmidomycin and for the bivalent cation. The catalytic flap covers residues 291 to 299. Its structure is of an α/β -type made up of five α -helices and four β -strands (Umeda et al. 2011).

The catalytic domain is connected to the C-terminal domain (red) through the linker region (residues 370 to 395, in yellow in Figure 1-5). The linker region spans the open face of the catalytic domain. The C-terminal domain (residues 396 to 486) is comprised of a four-helix bundle structure (Umeda et al. 2011).

1.1.3.2 Mechanism of action

DXR catalyses the second step of the MEP pathway. The enzyme is classified as a class B dehydrogenase as it uses the pro-S hydride of NADPH. The reaction converts DOXP to 2-C-methyl-D-erythritol-4-phosphate (MEP) by isomerization and followed by NADPH reduction. It is a rate limiting step of the pathway. Its mechanism of action of the enzyme has been widely studied (Murkin, Manning, and Kholodar 2014).

Two main chemical mechanisms were first proposed for the chemical reaction: the alpha-ketol rearrangement and the retro aldol sequence (see Figure 1-6). In the first one, the C-3 hydroxyl group is first deprotonated, followed by a 1,2-migration to yield methylerythrose phosphate. This is followed by a reduction to MEP by NADPH. In the second mechanism, a three-carbon and a two-carbon phosphate intermediate are generated through the cleavage by DXR of DXP C3-C4 bond in a retro aldol way. The two compounds are then combined in an aldol reaction to form a new C-C bond, giving an aldehyde intermediate which would be reduced to MEP by NADPH. Finally, the retro aldol is adopted as the most plausible mechanism supported by multiple studies (Li et al. 2013; Munos et al. 2009; Murkin, Manning, and Kholodar 2014).

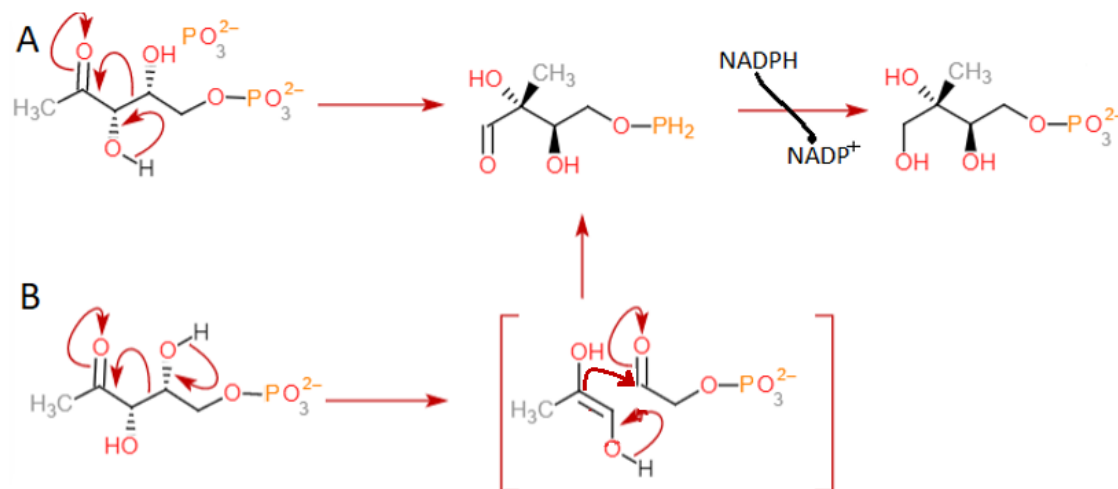


Figure 1-6: The alpha-ketol rearrangement (A) and the retroaldol (B) mechanism adapted from (Munos et al. 2009).

1.1.4 DXR Inhibition

1.1.4.1 Fosmidomycin - PfDXR

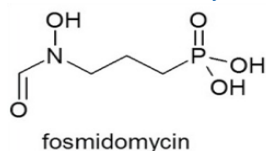


Figure 1-7: Fosmidomycin chemical structure (Wiesner et al. 2016)

Currently, there are 255 DXR inhibitors in the Binding Database (accessed November 2017) (Liu et al. 2007). These inhibitors are mainly based on the fosmidomycin chemical scaffold.

Fosmidomycin (FR31564 or 3-(N-formyl-N-hydroxyamino) propylphosphonic acid) (chemical structure in Figure 1-7) is a natural antibiotic isolated from *Streptomyces lavendulae* and first discovered by Fujisawa Company in the 1970s (Iguchi et al. 1980). With its structural similarity to DXOP, it inhibits the DXR enzyme by imitating the binding mode of the substrate. The binding region of the molecule consists of three different parts: the phosphonate moiety binding pocket, a hydrophobic patch for the carbon backbone and the hydroxamate group binding pocket (Umeda et al. 2011).

The crystal structures of PfDXR with fosmidomycin and other analogues inhibitors show high conservation of the residues involved in the binding region. The catalytic domain presents the binding sites for divalent cations (Mn^{2+} , Mg^{2+} or Co^{2+}) and the substrate. The domain has been shown to be highly conserved across multiple organisms through diverse multiple sequence alignments (Murkin, Manning, and Kholodar 2014). The metal ion is binding to residues Asp231, Glu233, and Glu315 in the bottom of a cleft in the catalytic domain in a distorted trigonal bipyramidal geometry (Umeda et al. 2011; Xue et al. 2012; Kunfermann et al. 2013; Konzuch et al. 2014). The affinity of fosmidomycin to DXR does not depend on the type of the metal cation Mg^{2+} or Mn^{2+} (Murkin, Manning, and Kholodar 2014). In general, DXR showed higher specificity for Mn^{2+} than Co^{2+} and more than Mg^{2+} while in plants and the parasite *Toxoplasma gondii*, the enzyme has comparable degrees of activation by both Mn^{2+} and Mg^{2+} and less or no activation by Co^{2+} . In general, Mn^{2+} remains the most effective metal cation (Argyrou and Blanchard 2004). In the case of PfDXR, studies showed that Mn^{2+} and Mg^{2+} were the metal cations used by the parasite enzyme, and it could not use Co^{2+} (Jessica L. Goble n.d.). Bodill et al. underlined the importance of the charge and parameters of the metal cation in DXR docking studies (Bodill et al. 2011).

The hydroxamate group chelates that metal cation in a binding pocket with carboxylate groups of Asp231, Glu233 and Glu315 residues (see Figure 1-8) (Umeda et al. 2011). Many other inhibitors have also been reported to bind the same residues (Kunfermann et al. 2013; Xue et al. 2012; Konzuch et al. 2014) . The oxygen atoms of the inhibitor hydroxamate group should be in a *cis* conformation. This is required for the tight binding of the inhibitor to the active site metal (Umeda et al. 2011) but this chelation is not an indispensable interaction (Murkin, Manning, and Kholodar 2014).

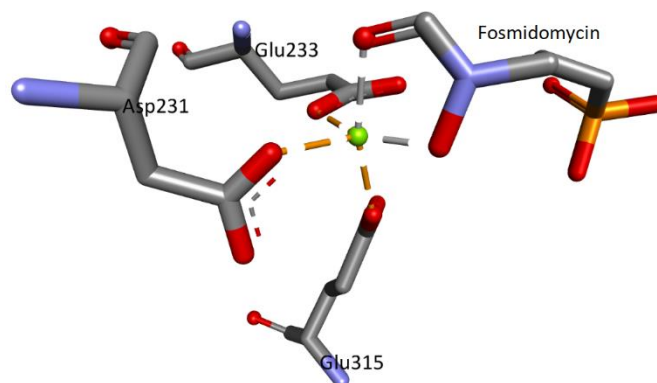


Figure 1-8: Metal chelation with hydroxamate group of fosmidomycin and carboxylate groups of Asp231, Glu233 and Glu315. Image Discovery Studio created from PDB ID 3AU9 (Umeda et al. 2011).

The negatively charged phosphonate moiety of the substrate binds in a polar environment, forming hydrogen bonds with, Ser270, Asn311, His293 and two water molecules (Kunfermann et al. 2013; Xue et al. 2012; Umeda et al. 2011), with the histidine residue having a key interaction with one of the phosphonate oxygens. This interaction is implied in the pre-orientation of the ligand in the active site. The phosphonate moiety is supposed to trigger the closure of a flexible loop over the active site of DXR thus acting as an allosteric effector (Kholodar and Murkin 2013; Murkin, Manning, and Kholodar 2014).

The three-carbon spacer in fosmidomycin backbone lies parallel to the indole ring of the tryptophan in the flexible loop region (Trp296) (see Figure 1-9). There is interaction with Met298 (Umeda et al. 2011). Mutagenesis studies on this Trp296 have shown its role in substrate discrimination by DXR and the condition of an aromatic residue for binding and catalysis. In the same way, mutating Met298 made DXP binding and turnover less efficient (Murkin, Manning, and Kholodar 2014; Fernandes and Proteau 2006).

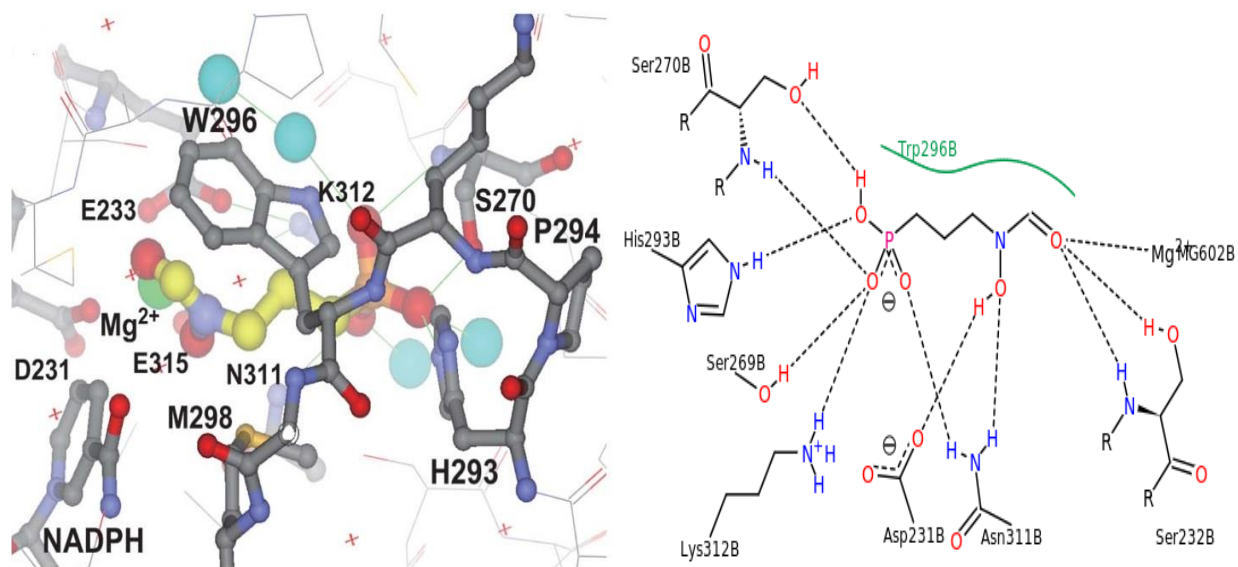


Figure 1-9: Left: Fosmidomycin complex with PfDXR. The carbon atoms of fosmidomycin (in yellow), the four buried water molecules (cyan), and the bound Mg²⁺ ion (green). Right: 2D diagram of fosmidomycin binding mode (Umeda et al. 2011; Konzuch et al. 2014). (Reproduced with permission).

The cofactor binds to the cavity composed of conserved residues D231, E233, S269, S270, W296, M298, S306, N311, K312, and E315 (Saggu et al. 2016). These residues are conserved across all human malaria species (Kunfermann et al. 2013). The enzyme in *Escherichia coli* showed a preference for NADPH contrary to NADH. Combining the enzyme with NADH in *EcDXR*, Takahashi et al. observed a reduction to 1% of the reaction rate (Takahashi et al. 1998).

After inhibitor binding, an induced fit movement of the enzyme accommodates the bound inhibitor in the active site (Umeda et al. 2011). The catalytic site is then covered by a movement of the flexible loop critical to the enzymatic reaction (Reuter et al. 2002; Yajima et al. 2002; Henriksson et al. 2007). In fact, the loop has been shown to be determinant in inhibition through interaction with the ligand directly or via active site water molecule. The kinetics of inhibition first requires the binding of the NADPH to DXR. The complex DXR-NADPH is essential for fosmidomycin competitive inhibition against DXP. Fosmidomycin showed a two-step mechanism for a slow tight binding mechanism. A first slow-onset phase is observed, then the ternary complex (DXR-NADPH-Fosmidomycin) undergoes a conformational change to form a stronger binding (Murkin, Manning, and Kholodar 2014; Kholodar et al. 2014).

Table 1-1: Some PfDXR residues and their identified/suggested roles in inhibition.

PfDXR Residues	Role	References
Ser269, Ser270, Ser306, Asn311, Lys312, His293	Binding phosphonate moiety of fosmidomycin	(Kunfermann et al. 2013; Xue et al. 2012; Umeda et al. 2011)
Thr86, Gly87, Ser88, Ile89, Asn115, Lys116, Ser117, Glu206, Gly299	NADPH binding	(Umeda et al. 2011)
Asp231, Glu233, and Glu315	Binding hydroxamate group of fosmidomycin. Divalent metal cation coordination (pentacoordinate trigonal bipyramidal geometry)	(Umeda et al. 2011; Xue et al. 2012; Kunfermann et al. 2013; Wadood et al. 2017; Konzuch et al. 2014) .
His293	Suggested to be important in ligand pre-orientation in the active site and loop closure. Hydrogen bonding and/or salt bridging with the phosphodianion of the ligand.	(Murkin, Manning, and Kholodar 2014)
Pro294	Important for maintaining the structure of the flexible loop.	(Umeda et al. 2011)
Gly299	May contribute to the flexibility of the flexible loop as an active site flap.	(Umeda et al. 2011)
Met298	Hydrophobic interactions with the backbone of the inhibitor or the nicotinamide moiety of NADPH. Mutation of this residue to alanine or valine was reported to significantly impair DXP binding and turnover.	(Murkin, Manning, and Kholodar 2014)
Trp296	Important role in discriminating DXR inhibitors, Induced-fit conformational change upon fosmidomycin binding, closing over and interacting with the bound inhibitor. Interact better with electron-deficient, hydrophobic group.	(Deng et al. 2010) (Sooriyaarachchi et al. 2016) (Umeda et al. 2011)
Linker region	Structural support for the catalytic domain	(Saggu et al. 2016).

1.1.4.2 FR900098

Also, isolated from *Streptomyces*, FR900098 was first identified by Fujisawa Pharmaceutical in the 1970s as a new antibiotic. The molecule differs from fosmidomycin by a methyl group (Iguchi et al. 1980). In 1999, Jomaa et al. established the antimalarial activity of FR900098 and fosmidomycin (Jomaa et al. 1999:199). The molecule is twice as active as fosmidomycin and has good toxicity profile in animal models (Wiesner et al. 2016). It shows similar binding mode to

fosmidomycin in its quaternary complex (PfDXR-NADPH-Mg²⁺ - FR900098). FR900098 methyl group is accommodated in the active site through a suggested induced-fit movement. The van der Waals interaction between Trp296 indole ring in PfDXR and FR900098 methyl group could explain its increased activity compared to fosmidomycin (Umeda et al. 2011). Also, by the presence of the acetyl group in place of the formyl group of fosmidomycin, FR900098 closely mimics the natural substrate DXP (Murkin, Manning, and Kholodar 2014).

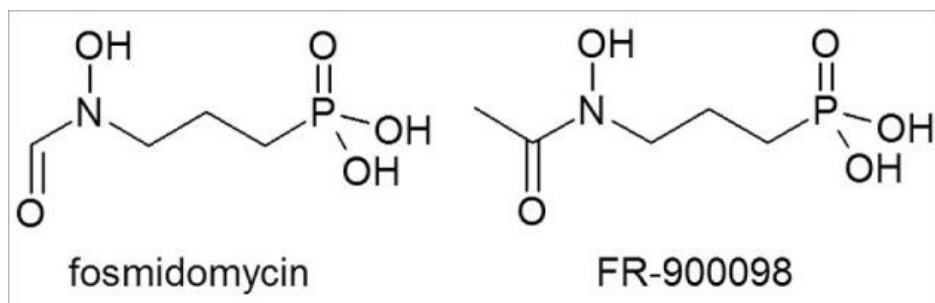


Figure 1-10: Fosmidomycin and FR-900098 structures (Wiesner et al. 2016).

Fosmidomycin and FR900098 (see chemical structures in Figure 1-10) are the two most studied inhibitors of DXR. Unfortunately, fosmidomycin has poor pharmacological properties: short half-life in plasma as well as poor oral availability and low lipophilicity, and this emphasizes the need for new and more efficient inhibitors.

1.1.4.3 Other DXR Inhibitors

1.1.4.4 Chemical synthesis approaches

Since its discovery, diverse research groups have extensively worked on the development of fosmidomycin/FR900098 analogues for DXR inhibition. Maintaining the main frame of these molecules, efforts for new inhibitors focused on improving the lipophilicity for better druglike properties. Different strategies were used for that purpose: modification of the phosphonate group, modification of the hydroxamate group, modification on the three-carbon spacer. Hybrid methods combining previous ones and prodrug approaches were also used. Several compounds have been synthesized and tested *in vitro* or/and *in vivo* for their antiplasmodial activity. Many of these compounds showed good inhibitory activities against PfDXR compared to the reference molecules, fosmidomycin/FR900098, supporting their further studies as potential drug candidate (Aneja et al. 2016; Chofo et al. 2014).

These studies have revealed some key features on the structure-activity relationships of DXR inhibition. The two main chemical groups, the phosphonate group and the hydroxamate group have been shown to be essential for inhibitory activity (see Figure 1-11). A reverse orientation of the hydroxamate is also as effective as the normal one. The group chelates the metal ion, essential for inhibitory activity. N-methyl substituted hydroxamic are preferred. The three-carbon propyl provides the optimal length between these two previous chemical groups to maintain them in good binding position (Chofo et al. 2015; Murkin, Manning, and Kholodar 2014; Saggu et al. 2016; Aneja et al. 2016).

In review studies, Saggu et al. 2016 and Aneja et al. 2016 reported numerous synthesized DXR inhibitors. Many of these compounds showed high potent inhibitory activities against PfDXR with IC50 and K_i values in the low nanomolar range. An equal or improved inhibitory activity with respect to fosmidomycin has been recorded for many of them. Notably, among lipophilic inhibitors, a pyridine-containing fosmidomycin derivative showed 11-fold improved inhibitory activity against PfDXR compared to fosmidomycin, the reference molecule. The compound also showed better antiplasmodial activity. Some of these inhibitors, thus, show high potential as a drug candidate for further pharmacological studies (Saggu et al. 2016; Aneja et al. 2016).

Despite these advancements, Deng et al. synthesized compounds structurally different from fosmidomycin missing either the hydroxamate or the phosphonate group, which showed unexpectedly interesting biological activities illustrating the challenge of rational chemical design (Deng et al. 2010; Chofor et al. 2015). According to Mercklé et al. the rational design of new inhibitors of DXR difficult at best (Mercklé et al. 2005). Chofor et al. underlined the daunting challenge of finding different and efficient bidentate ligands, mirroring the fosmidomycin hydroxamate for metal chelation (Chofor et al. 2014).

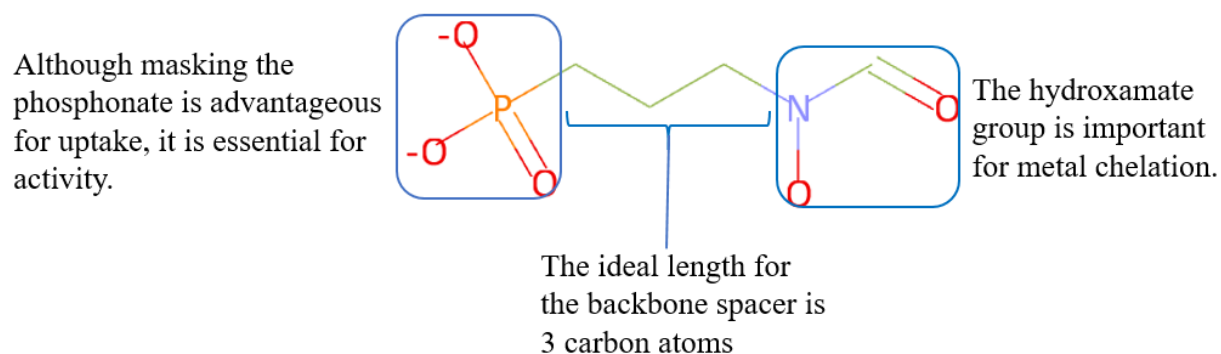


Figure 1-11: Important structural features of fosmidomycin.

1.1.4.5 *In silico* drug discovery approaches in DXR inhibition

The first solved DXR crystal structure was of *E.coli* (EcDXR, PDB IDRE 1K5H) (Reuter et al. 2002). With no crystal structure available, homology modeling was used to generate 3D structures for the *Plasmodium* enzyme (Goble 2011; Singh et al. 2007). Goble et al. and Singh et al. developed and validated a model for PfDXR to analysis functional and structural features of the enzyme inhibition. Docking studies were performed on the model, revealing structural information on the importance of the flexible loop, and ligand binding sites. Good correlations were found between inhibitors' activity and their binding energies and thus the potential of docking to identify good inhibitors.

In parallel to these efforts, data from several DXR X-ray structures from multiple organisms helped to improve the precision of prediction of structural affinity for related ligands without losing the capability to estimate the affinities of structurally distinct inhibitor (Silber et al. 2005).

Elsewhere the different potential binding pockets in DXR have been analysed and the crystal structure (PDB ID: 1Q0Q) presented mainly two binding pockets: the substrate binding and the cofactor binding pockets. Both were found to be druggable assessed by DoGSiteScorer (Volkamer et al. 2012) and presenting a Dscore of 0.8 for the fosmidomycin binding pocket and 0.76 for the NADPH binding pocket. The Dscore evaluates the druggability of binding pockets in protein and varies from 0 (poor druggability score) to 1 (high druggability score). Even though this latter pocket presents a low lipophilic character, it is still worth exploring especially for bisubstrate analogues (Masini, Kroezen, and Hirsch 2013; Deng et al. 2010).

Using fosmidomycin fragments, Mercklé et al. tested different compounds. None of them showed time-dependent inhibition of DXR. It was then suggested that these compounds were not able to induce the required reorganisation of the active site and the full-loop closure for inhibition. More, computational modeling and docking studies showed that a close structural analogue of fosmidomycin ((S)-N-hydroxy-4-(phosphoryloxy) methyloxazolidin-2-one) could bind to DXR. Unfortunately, this compound also could not show cooperative nor time dependent inhibition against the enzyme revealing challenges with the computational approaches: the problem of force field parameters for metal ions and the challenge to predict or simulate DXR flexible loop movement upon inhibitor binding (de Ruyck, Wouters, and Poulter 2011). Also, Deng et al. observed inhibitors binding in reverse mode with the phosphonate binding to Mg^{2+} and the hydroxamate located in the phosphonate binding site in docking studies on Ec-DXR (Deng et al. 2010). A similar observation was made in a docking study on PfDXR (Bodill et al. 2013).

Umeda et al. solved the first PfDXR crystal structure, revealing the intrinsic flexibility of the molecule and especially of its active site (Umeda et al. 2011). Subsequently, other crystal structures, complexed with different inhibitors were solved. This provided much insight on structural and kinetics information of DXR and inhibitors interactions (Xue et al. 2012; Kunfermann et al. 2013; Konzuch et al. 2014; Chofer et al. 2015; Sooriyaarachchi et al. 2016). Currently (July 2017), there are PfDXR crystal structures complexed with diverse ligands available in the PDB database (Berman et al. 2000).

Based on previous quantitative structure–activity relationship (QSAR) and crystallographic studies, Xue et al. synthesized pyridine-containing fosmidomycin derivatives (see Figure 1-12) which showed to be potent inhibitors with K_i values of 1.9–13 nM. Crystallographic complexes containing these pyridine fosmidomycin derivatives showed the movement of the flexible loop upon ligand binding away from the active site centre to hold the pyridine group and allow the receptor to have favourable hydrophobic interactions with the inhibitor (Xue et al. 2012).

Through crystallographic and kinetics studies of chiral inhibitors, Kunfermann et al. showed that PfDXR has a high level of enantioselectivity for an α -substituted fosmidomycin derivative (see Figure 1-12). More potent inhibitory activity was observed for S-(+)-enantiomer derivatives. (+)-enantiomers showed IC_{50} values of 94 nM (+) while (–)-enantiomers showed $>10 \mu M$ (–) against Pf IspC. A similar observation was made in EclspC MtlspC. Bulkier substituents like the phenyl group could adapt in the active site thanks to the mobility of the loop region (Kunfermann et al. 2013).

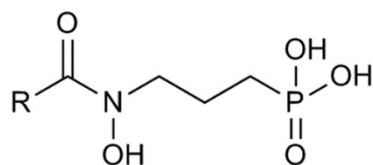
Chofor et al. synthesized β substituted fosmidomycin analogues (see Figure 1-12) with high inhibitory activity on *P. falciparum* growth *in vitro*. Crystal structures of complexes with the most active ligands revealed a different flap structure with the aromatic group of the ligand lying between the tryptophan of the flexible loop and the hydroxamate's methyl group. This reorganization of the flap results in favourable interactions between the phenyl ring of the inhibitors and the tryptophan of the flexible loop for better inhibitory activity (Chofor et al. 2015).

Based on a similar observation that the N-methyl group of FR900098 has favourable van der Waals contact with the indole ring of Trp-296 resulting in higher inhibition, Cobb et al. synthesized a n-propionyl FR-900098 derivative (FR-900098P) (see Figure 1-12). This could extend the inhibitor into the hydrophobic pocket flanked by Met-298 and Met-360 for more favourable van der Waals interactions and better binding affinity. FR-900098P showed inhibition constants K_i value of 0.92 ± 0.19 nM much better than the parent compounds (Cobb et al. 2015).

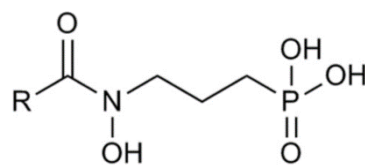
In recent years, new ligands with potential inhibitory effect on *P. falciparum* were identified through successfully multiple structure-guided designs and virtual screenings (Tangyuenyongwatana and Gritsanapan 2017; Wadood et al. 2017; Chaudhary and Prasad 2014; Cobb et al. 2015; de Ruyck et al. 2016).

Drug like molecules from the ZINC (Irwin et al. 2012) database were screened with the help of FRED module of Open Eye software against followed by docking study of selected hits. These hits showed better binding energies than fosmidomycin. A final compound with good toxicological profile was assessed through OSIRIS Property explorer (Chaudhary and Prasad 2014). Using a pharmacophore model of PfDXR active site and molecular docking, Wadood et al. identified new potent inhibitors from the ChemBridge (<http://www.chembridge.com/>) database. Their pharmacokinetic properties were also assessed through *in silico* ADME studies. Computational promiscuity binding data revealed that the identified hits could as well bind others *P. falciparum* drug targets (Wadood et al. 2017). Through a shape-based search approach using ArgusLab, the ZINC12 database was docked against PfDXR to find hits with similar functional group to fosmidomycin interacting with residues in the active site (Tangyuenyongwatana and Gritsanapan 2017).

By means of chemical and computational approaches, many advancements have been made in the research of pharmacological effective DXR inhibitors and in the understanding of its inhibition mechanism. Despite all these efforts, an exhaustive structural description and detailed reaction mechanism of DXR inhibition remain incomplete and there is still room for exploring clinically effective DXR inhibitors.

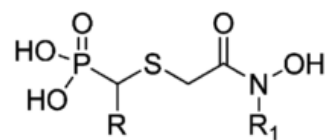


R = H, fosmidomycin
R = CH₃, FR900098



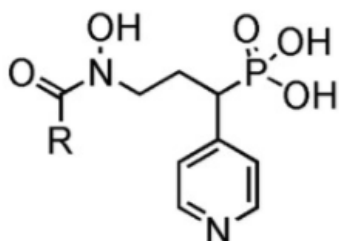
R = CH₂CH₃, FR900098P

N-propionyl FR-900098
derivative
(Cobb et al. 2015)

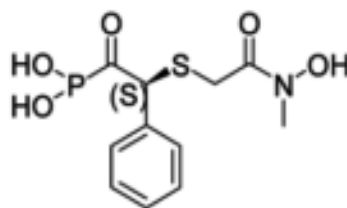


R: Aryl, R₁ = H, CH₃

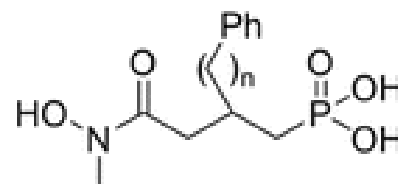
Reverse orientation of the hydroxamate
with N-methyl substituted hydroxamine
(Konzuch et al. 2014).



R = CH₃
Pyridine-containing
fosmidomycin derivative
(Xue et al. 2012)



Chiral enantiomers S-(+)- DXR
inhibitors (Kunfermann et al.
2013).



β-Substituted Fosmidomycin Analogues.
(Chofor et al. 2015)

Figure 1-12: Some DXR inhibitors

1.5 Research problem statement and justification

Malaria remains a major health concern with a complex parasite constantly developing resistance to the different drug introduced to treat malaria, threatening the current ACT treatment recommended by WHO. Different antimalarial compounds with different mechanisms of action are ideal as this decreases chances of resistance occurring (Lunev et al. 2016). With no currently available effective vaccine, these factors underscore the necessity for new antimalarial to constantly feed the antimalarial pipeline and to stay ahead of resistance (Hemingway et al. 2016:201). Inhibiting DXR and consequently the MEP pathway is a good strategy to find a new antimalarial with a novel mode of action.

Furthermore, while developing new antimalarials, an ideal drug should combine the following profiles: matching or improving the ACT 3 days treatment, curing the blood-stage, blocking the parasite transmission and hypnozoitocidal properties (Burrows et al. 2014). The MEP pathway has been shown to be present in all intra-erythrocytic stages of the parasite and is required for optimal development of the hepatic stage (Guggisberg, Amthor, and Odom 2014). Even though fosmidomycin does not inhibit gametocytes, MEP intermediates are present in gametocyte and isoprenoid are essentials to gametocytogenesis, showing thus potentiality for transmission blocking compounds (Wiley et al. 2015). In addition, the MEP pathway is indispensable for many human and animal pathogens, and also problematic weeds. None of its enzymes has homologs in

human. It thus presents an interesting potential as broad-spectrum antimicrobial agents with a novel mode of action and novel herbicides (Odom 2011).

Currently, numerous DXR crystal structures from diverse organisms *M. tuberculosis* and *E. coli* and *P. falciparum* are available. Also, numerous *in silico* structure-based studies have been done on these structures complexed with various ligands. These studies revealed much information on SAR of DXR inhibition but also revealed its mechanistic and structural complexity (Murkin, Manning, and Kholodar 2014). Currently, except for *P. falciparum* there is no crystal structure available for any of the other *Plasmodium* species. Comparative protein modeling can be used here to obtain protein structural information in the absence of an experimentally determined structure (Webb and Sali 2014).

Although previous studies shed much light on the SAR of DXR inhibition, these findings are not exhaustive as the protein active site remains very flexible (Murkin, Manning, and Kholodar 2014). Other compounds with significantly different scaffolds from fosmidomycin also showed interesting inhibitory potency, and reverse binding modes in docking studies were also observed (Deng et al. 2010; Bodill et al. 2013). Thus, exploring new compounds can contribute to extend the understanding of DXR inhibitory mechanism.

Natural products have been a major source of pharmaceutical drugs. Still only 6% of existing plant species have been systematically studied for pharmacological activity (Atanasov et al. 2015). The South Africa Natural Compounds Database (SANCDDB) (Hatherley et al. 2015) has not been explored yet for potential activity against DXR. SANCDDB is a database of diverse natural compounds from aquatic and land-based origin. These natural products offer compounds with diverse chemical and biological properties. Evaluated through the Lipinski's rules of five, 60% of the compounds met the conditions, and up to 80% violate at most only one rule (Hatherley et al. 2015). They, thus, present good potential for exploring compounds with good pharmacological properties. Currently (July 2017) the database contains around 700 compounds from 166 different organisms. These compounds often present better properties for drug development compared to synthetic ones. They play an important role in drug discovery and development, and much of their potentiality remains to be discovered (Newman and Cragg 2016). SANCDDB database has not previously been screened against DXR. It, thus, constitutes an excellent source for potential DXR inhibitors with good drug likeness properties. For that purpose, computational docking provides a fast and efficient method for virtual screening of large libraries of compounds (Lionta et al. 2014).

1.6 Aims

The aim of the project is to investigate the SANCDDB database for new DXR inhibitors. First homology models will be built for the different *Plasmodium* species (*P. malariae*, *P. vivax*, *P. ovale*, *P. knowlesi*, *P. berghei*, *P. yoelii yoelii* and *P. chabadi*.). This will be followed by the exploration of hits from SANCDDB using molecular docking. Finally, the identified hits will be further evaluated through molecular dynamics.

1.7 Objectives

1. Retrieve PfDXR and ortholog protein sequences *P. falciparum*, *P. malariae*, *P. vivax*, *P. ovale*, *P. knowlesi*, *P. berghei*, *P. yoelii yoelii* and *P. chaubadi* from PlasmoDB (Aurrecochea et al. 2009) database and their homologs (*Escherichia coli*, *Synechocystis sp.*, *Bacillus anthracis*, *Mycobacterium tuberculosis*, *Arabidopsis thaliana*, *Chlamydia trachomatis*, *Clostridium tetani*, *Vibrio cholerae*, *Zymomonas mobilis*) from NCBI (NCBI Resource Coordinators 2017) to perform sequence analysis.
2. Perform comparative multiple sequence analysis of PfDXR orthologs to deduce residue variations and to investigate their effect on substrate binding.
3. Build and validate 3D structures in open and closed conformations through homology modeling for *Plasmodium* sequences: *P. falciparum*, *P. malariae*, *P. vivax*, *P. ovale*, *P. knowlesi*, *P. berghei*, *P. yoelii yoelii* and *P. chaubadi*.
4. Perform high throughput screening of SANCDB (Hatherley et al. 2015) compounds against PfDXR crystal structure and the 3D models of other *Plasmodium* species to identify potential inhibitors and assess their drug likeness properties.
5. Develop and validate force field parameters for the metal ion in DXR active site.
6. Molecular dynamic studies using GROMACS protein-ligand complexes of hits to evaluate the stability of each complex.

CHAPTER 2: SEQUENCE ANALYSIS

2.1 Introduction

Protein functions are related to their 3D structures. Sequence variations can impact protein 3D structure resulting in differences in functions and in inhibition. Sequence variations may be related to key residues or regions in the protein sequences, consequently impacting the different functions the protein can be implied in. For example, in enzyme inhibition, the same ligand can show different type of binding depending on the sequence and structural variation of the targets. This is important especially for key residues. Two or more biological sequences can share similar regions, domains and/or residues. Aligning these regions/residues may suggest that these sequences share their related functionality. Conservation of residues in an enzyme catalytic site, for example, suggests the common activity shared across the different sequences. The validity of such hypothesis also depends on the quality of the alignment conducted (Baxevanis and Ouellette 2001).

Animal models are often used during drug development. Rodent models are important translational models for research on protozoan parasites. In malaria, rodent models (*P. yoelii yoelii*, *P. chabaudi* and *P. Berghei*) are often used *in vivo* experiments for drug and vaccine development. Elucidating sequence and structural variations through comparative studies can help understand and interpret experimental results variation across the different disease models (Ehret et al. 2017).

This chapter aims to analyse homolog protein sequences of different *Plasmodium* species. Analysis of sequence features relating to substrate binding and interactions compared with similar types of DXR found in *Plasmodium*, Apicomplexa and eukaryotes is to be performed with the main aim to identify regions that are conserved. Important sequence features related to the protein function and its inhibition will be analysed across the different species (*P. falciparum*, *P. malariae*, *P. vivax*, *P. ovale*, *P. knowlesi*, *P. berghei*, *P. yoelii yoelii* and *P. chaubadi*). Residues of interest will be those involved in substrate/inhibitor binding and enzyme inhibition. These variations will be considered during docking analysis (Chapter 4) to find out if they have significant effects on inhibitor binding modes.

DXR is well studied enzyme. Multiple crystal structures of the protein with different ligands from different organisms have been solved. Critical residues implied in binding to the protein and protein inhibition has been identified not only in *P. falciparum* but also in other organisms. Also, residues involved in binding of the cofactor and in the active site flap region have been identified.

In this chapter, a comparative analysis including some non-apicomplexan species (*Escherichia coli* (*Ec*), *Mycobacterium tuberculosis* (*Mt*), *Helicobacter pylori* (*Hp*), *Vibrio cholera* (*Vc*), *Bacillus anthracis* (*Bc*), *Arabidopsis thaliana* (*At*), *Synechocystis sp* (*Sy*)) will also be done. This will allow us to identify region and residue differences within the apicomplexans which may impact on inhibitors selectivity across the different *Plasmodium* DXR. Our results will be compared to the literature.

2.1.1 Biological databases and information search

Biological information comprises different types of data including: literature, genomic and protein sequences, sequence annotation information, motifs and domains and protein 3D structures. These data are saved in different file formats and often organized in biological databases. This allows for efficient storage of data but also facilitates the search of this data using different criteria, access to this data, and allows for downloading and the management of information. These different features help in knowledge discovery but also this is very important as the development of new technologies and their decreasing costs have caused an increase in the amount of available data. As a consequence, a large community of scientists can now submit information into biological databases which raises the need for curation of data (Marx 2013).

Depending on their level of curation, biological databases can be classified into primary databases (example GenBank), and secondary databases (example SWISS-PROT). The first category contains sequences or structural data submitted by the scientific community, while secondary databases contain data automatically treated using computers or manually curated information (Baxevanis and Ouellette 2001). There are also specialized databases which tend to focus on information related to a specific area of research or an organism such as PlasmoDB (Aurrecochea et al. 2009). To retrieve information from these databases, a classic word search based method, searching with key/identifiers, usage of Boolean operators and ability to specify some additional criteria can be done. Some advanced features also as the use of API (Application Programming Interface) are implemented on some of these databases. More specific to the biological field is the search for similarity between biological sequences through sequence alignment (Xiong 2006).

2.1.2 Sequence alignment

Sequence alignment is a fundamental method used in bioinformatics to compare and find the similarity and identity between DNA, RNA and proteins sequences by arranging them in a certain way. This can be done between two sequences (pairwise alignment) or multiple sequences (multiple sequence alignment; MSA). Similarities and differences are thus revealed with their implication for structural, functional, and evolutionary relationships between the sequences (Baxevanis and Ouellette 2001). A key purpose of sequence alignment is to infer sequence homology based on their identity and length, especially important for homology modeling studies in the next chapter. When sequence identity and length falls in the safe zone of the homology detection graph established by Rost, the proteins are homologous. The graph has a twilight zone in which inferring homology is risky and a midnight zone, where in which inferring homology is not reliable. A cutoff of 20% of identity is enough for sequences longer than 250 amino acids. Shorter sequences require higher cutoffs for inferring homologous relationships than longer sequences (Rost 1999).

Most of the alignment programs use two main algorithms: Smith-Waterman and Needleman-Wunsch. The first method uses a local alignment strategy trying to spot regions of similarity within sequences that may not be evident when maximizing the alignment over the entire sequences. Whereas, Needleman-Wunsch algorithm makes use of a global alignment approach trying to find the maximum similarity between sequences across their full span. A hybrid method, semi-global alignment compromises between the two previous approaches, and is ideal when

regions of similarity are found in the ending part of the sequences. The results produced by these procedures can be manually adjusted to reflect biological meanings related to research questions (Baxevanis and Ouellette 2001).

To characterize the biological meaning of amino acid similarity in protein sequences, substitution matrices are used. They characterize the likelihood of sequence character to replace each other. PAM (Point Accepted Mutation) and BLOSUM (BLOck SUBstitution Matrix) are the two most commonly used series of matrices. For the PAM1 matrix approach, probabilities of substitution are calculated when 1% of the amino acids had changed; other PAM matrices are then derived using mathematical matrix operations. PAM was obtained by observing global alignment of closely related sequences while the BLOSUM series was derived from local alignment of evolutionary divergent sequences. Different identity thresholds were fixed for the set of sequences, giving thus the different BLOSUM matrices. The matrix probabilities were calculated from observation of the conserved regions (blocks) (Henikoff and Henikoff 1992; Pevsner 2009).

Ideally, computational implementations of these algorithms use dynamic programming as it produces the best alignment(s) possible for the sequences. Unfortunately, the method is computationally expensive when it comes to data as large as the biological data used. Heuristic approaches are then used. BLAST (Basic Local Alignment Tool) is a fast and heuristic approach for sequence similarity search in large biological databases. As algorithm, a word-based search method is used to find short matches and extend upon them by local alignment if the alignment score is above a set threshold. The significance of an alignment is estimated by the E-value (Expected value) (Altschul et al. 1997).

MSA is used when aligning more than two sequences. As for pairwise alignment, dynamic programming provides the most efficient algorithm to find the exact solution. Unfortunately, as the number of sequences increase, the approach becomes impractical. Heuristic approaches which do not guarantee the most accurate solution are then used (Altschul et al. 1997). They are the progressive and the iterative methods. In the first approach, two sequences from the set are first aligned. These two sequences are the closest ones selected from a guide tree. A third sequence following the order in the guide tree is alignment with the resulting alignment. This process is repeated to progressively align all sequences in the set (Wallace, Orla, and Higgins 2005). Some examples of tools using this method are Clustal and T-Coffee. The second approach starts with a low-quality alignment of the sequences. Then improvements are done by realigned in an iterative way until no more improvement can be done. A program like PRRN uses an iterative alignment method (Xiong 2006).

This helps to identify conserved regions and residues across these sequences but also insertions and deletions. These features can be difficult to identify in a pairwise alignment. To quote Arthur M. Lesk, "Two homologous sequences whisper, a multiple alignment shouts loudly". Adding more sequences to the alignment improves its accuracy. Residues or regions implied in key structural or mechanistic functions in a protein will show high degrees of conservation across this protein. For example, the conservation of an enzyme's catalytic residues across sequences can indicate the conservation of the same catalytic activity across these organisms. But, it is notable that there are some exceptions. One interesting example is the recognition region in antibodies and MHC molecules. They remain key functional regions but hypervariable (Reche and Reinherz 2003).

Aligning multiple sequences is also used in protein classification into families, in motif search and phylogenetic tree construction (Pevsner 2009).

2.1.3 Phylogenetic trees

Phylogenetic trees are diagram-like representations depicting evolutionary relationships between organisms. They have a broad sphere of application, such as in helping species classification, the study of disease origin, resistance evolution, protein families coevolution and gene transfers and speciation events (Brown 2002). In homology modeling, they can help reconstruct an ancestral sequence to be modelled (Studer et al. 2014). They are based on organism molecular (genes and proteins) or physical information and give the information on sequence's origin, identify paralog and ortholog sequences and gene transfers. Trees can be built using different UPGMA (Unweighted Pair Group Method with Arithmetic Mean), neighbor-joining, maximum parsimony, maximum likelihood and Bayesian methods. The first two methods are fast, based on distance-matrices (based on sequence similarity) and do not imply a biological model of evolution. The other methods are character based, trying to estimate dynamic of mutations traced back to ancestral sequences. Maximum parsimony uses a minimum evolution model by minimizing the number of mutations. Finally, maximum likelihood and Bayesian methods use an explicit evolution model that fit best the data. They are more robust methods, but also are more computationally demanding (Xiong 2006; Pevsner 2009; Baxevanis and Ouellette 2001).

2.2 Methodology

2.2.1 Data retrieval

PlasmoDB (Aurrecochea et al. 2009) and the NCBI (National Center for Biotechnology Information) databases were used to retrieve the sequence data. Protein sequences in Fasta format were retrieved from PlasmoDB (Aurrecochea et al. 2009) for the *Plasmodium* species.

On PlasmoDB, the sequence of the most documented crystal structure 3AU9 was used to search for *Plasmodium* orthologs through BLAST. The BLAST search was conducted with the following parameters: Matrix: BLOSUM62, Gap Penalties: Existence: 11, Extension: 1, Neighboring words threshold: 11, Window for multiple hits: 40. The search was limited to the databases of the different *Plasmodium* species to select only ones from the organisms of interest.

The protein sequences for the other organisms were obtained from the NCBI. A BLAST search was conducted with 3AU9 sequence as query and the following default parameters were used: Gapcosts 11.1, Matrix BLOSUM62, window size 40, word size 6. Endoplasmic reticulum signal and plastidial targeting sequences were trimmed from the N-terminal for the search as these regions are absent in the mature protein.

2.2.2 Multiple Sequence Alignment

PROMALS3D (available at <http://prodata.swmed.edu/promals3d/promals3d.php>), and MUSCLE (<http://www.ebi.ac.uk/Tools/msa/muscle/>) were used to perform the MSA. PROMALS3D has the advantage to introduce structural constraints in the alignment. PSI-BLAST and secondary structure prediction from PSIPRED are used to build hidden Markov model (HMM) forming sequence constraints. PROMALS3D default parameters were used. When using PROMALS3D, a 3D structure from the PDB (Protein Data Bank) (Berman et al. 2000) database if available for each sequence was added in the alignment.

MUSCLE stands for MUltiple Sequence Comparison by Log- Expectation. As previously described, it uses a progressive alignment method (Pei, Kim, and Grishin 2008; Edgar 2004). The default parameters were used: BLOSUM62 matrix, gap opening of -12.0 and gap extension of -1.0 with CLUSTALW as weighting scheme and UPGMB for clustering.

The outputs of the alignments were visually inspected for misalignment(s) and edited with Jalview (Waterhouse et al. 2009) where necessary, and the best alignment based on regions and residues conservation was selected. The best results were obtained with MUSCLE.

2.2.3 Phylogenetic analysis

From the alignment obtained from MUSCLE a phylogenetic analysis was conducted. MEGA (Molecular Evolutionary Genetic Analysis) version 7.0.26 was used. The tool provides sophisticated approaches for phylogenetic analysis.

Any of the tree construction methods is not guaranteed to produce the most accurate tree. Combining different construction methods and validating the consistency of the produced tree provides a strong support for the tree accuracy (Xiong 2006). MEGA allows to select between different evolutionary models depending on the datatype (amino-acids, DNA, RNA) to estimate the pairwise evolutionary distance. Then using the bootstrap method and analytical formulas MEGA computes the standard errors for the estimates. The number of bootstrap replication for test of phylogeny was set to 1000. The statistical method of Maximum Likelihood was used for the phylogeny analysis. The substitution type was amino acid. The lowest BIC score (Bayesian Information Criterion) model describes the best substitution pattern. The three lowest BIC scores models (LG+G32, LG+G+I33, WAG+G+I33) were selected. Different site coverage cut-offs to remove gaps in the alignment were tried: 100%, 90%, 85%, 80% (see Table 2-1). The tree inference method was the Nearest Neighbor Interchange (NNI) with a strong branch swap filter and a bootstrap value of 1000. Depending on the model selected and the site coverage cut-offs, different trees were generated.

Table 2-1: Models and site coverage cut-offs for phylogenetic tree construction

Models	Site coverage cut-offs
LG+G32	100%, 90%, 85%, 80%.
LG+G+I33,	100%, 90%, 85%, 80%.
WAG+G+I33	100%, 90%, 85%, 80%.

The trees were assessed using the bootstrapping method through their consistency with the bootstrap tree and the bootstrap support values. The tree with the highest log likelihood is selected.

2.2.4 Motifs analysis

MEME (Multiple EM for Motif Elicitation) (Bailey et al. 2009) webserver (<http://www.meme.nbcn.net/meme/cgi-bin/meme.cgi>) was used for motifs finding and analysis. The program uses expectation maximization (EM) algorithm and can take multiple sequences to identify motifs across these sequences. The search was conducted on the set of 17 sequences. The minimum motif width was 3 and the maximum width was 20. A gradual search was performed to identify all motifs with no repetition. The search was first performed for a high number of motif which may return repetitive motifs. Then this number was reduced to finally only unique motifs. 21 unique motifs were identified. The 'E-value' of the motif, the probability of finding it in random sequences (Bailey et al. 2006), was used to assess the significance of a motif. A Python script: Motif analyser written by Ngonidzashe Faya and Pr. Ozlem Tastan Bishop was used to produce a heatmap for the identified motifs. The motif search was performed to analyse similarity and variations of motifs across the different *Plasmodium* species, especially motif located in the binding sites.

2.3 Results and Discussion:

2.3.1 Sequence data

PfDXR ortholog sequences from PlasmoDB showed an expected high percentage of similarity (see Table 2-2). Sequences showing lowest sequence identity to 3AU9 were PBANKA_1330600 and PCHAS_1335200 from *P. berghei* and *P. chabaudi chabaudi* respectively. As the crystal structure, the PF3D7_1467300 sequence from *P. falciparum* 3D7 shows 100% identity to 3AU9. These sequences were further confirmed through PlasmoDB orthology classification. All the sequences belonged to the same ortholog group [OG5_130462](#).

Table 2-2: *Plasmodium* DXR sequences retrieved from PlasmoDB. New ID is the ID used in this thesis.

Accession numbers	New ID	Organism	Score	E-Value	%identity to 3AU9	Protein length
PBANKA_1330600	PbDXR	<i>P. berghei</i> ANKA	696	0,00E+00	72%	495
PCHAS_1335200	PcDXR	<i>P. chabaudi chabaudi</i>	687	0,00E+00	72%	495
PF3D7_1467300	PfDXR	<i>P. falciparum</i> 3D7	991	0,00E+00	100%	488
PKNH_1214000	PkDXR	<i>P. knowlesi</i> strain H	653	0,00E+00	76%	519
PVP01_1239500	PvDXR	<i>P. vivax</i> P01	648	0,00E+00	73%	528
PY05578	PyDXR	<i>P. yoelii yoelii</i> 17XNL	694	0,00E+00	74%	493
PmUG01_12049600	PmDXR	<i>P. malariae</i> UG01	717	0,00E+00	82%	498
PocGH01_12047500	PoDXR	<i>P. ovale curtisi</i> GH01	700	0,00E+00	75%	478

The homolog sequences from the other organisms showed a lower sequence identity percentage compared to the *Plasmodium* ones as expected (see Table 2-3). Nonetheless the high sequence identities for the different sequences retrieved (with a minimum of 34% sequence identity) and the good coverage (at least 91%) indicate their homology relationship. These sequences fall in the safe zone of the graph for homology deduction (Rost 1999).

Table 2-3: *Plasmodium* DXR homologs sequences retrieved from NCBI

Accession numbers	New ID	Organism	Score	E-Value	%identity to 3AU9	Query coverage	Protein Length
WP_000811927.1	EcDXR	<i>Escherichia coli</i>	284	9e-91	37%	93%	398
WP_041426024.1	SsDXR	<i>Synechocystis</i> sp. PCC 6803	305	8e-102	42%	96%	393
WP_072191693.1	BaDXR	<i>Bacillus anthracis</i>	295	6e-97	39%	97%	385
WP_031675280.1	MtDXR	<i>Mycobacterium tuberculosis</i>	233	2e-71	34%	97%	413
OAO90921.1	AtDXR	<i>Arabidopsis thaliana</i>	315	2e-103	41%	97%	479
CRH69109.1	CtrDXR	<i>Chlamydia trachomatis</i>	293	8e-96	41%	92%	387
WP_011099497.1	CteDXR	<i>Clostridium tetani</i>	328	1e-110	49%	91%	384
WP_001229020.1	VcDXR	<i>Vibrio cholerae</i>	272	2e-87	36%	99%	402
WP_014848273.1	ZmDXR	<i>Zymomonas mobilis</i>	269	2e-87	37%	92%	388

2.3.2 Multiple Sequence Alignment

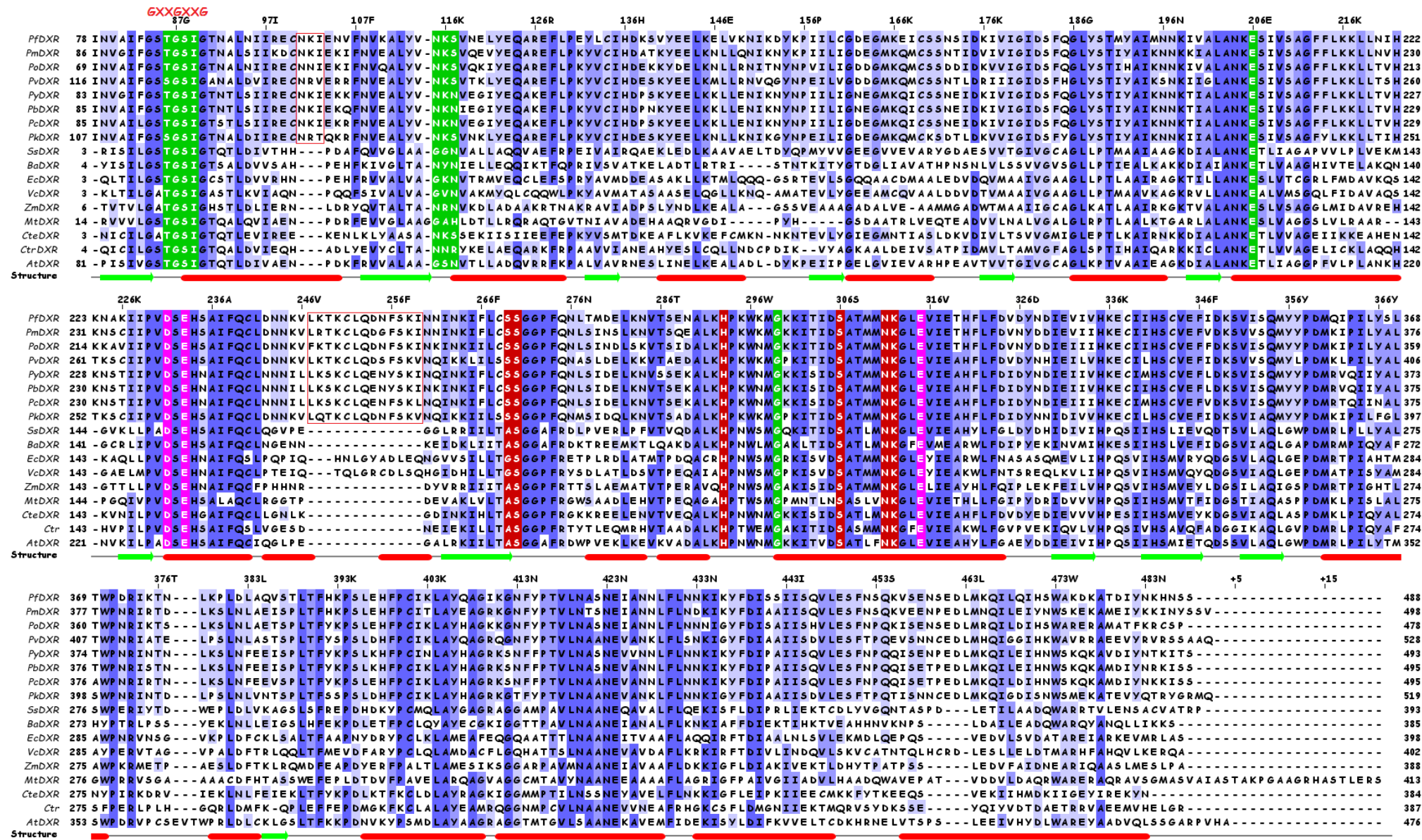


Figure 2-1: MSA of *Plasmodium* DXR and its homologs. The alignment is coloured by percentage identity.

Only residues falling into the highest percentage identity (consensus sequence) are coloured. The numbering of residues (scale above) is based on the sequence of reference PfDXR (PDB ID: 3AU9). Columns are coloured per percentage identity. For each column, only residues that agree with the consensus sequence are coloured. Hyphens represent the absence of amino acid residues. Secondary structures annotations were extracted from the structure of 3AU9 and displayed below the alignment (Red is helices and green is sheets). Sequences are truncated to have portions of the N-terminus. In *Plasmodium*, the first 77 residues are similar to be an endoplasmic reticulum signal and plastidial targeting sequences (Umeda et al. 2011).

> 80 %
> 60 %
> 40%
< 40%

The alignment shows the expected higher degree of conservation among *Plasmodium* sequences. Compared to other organisms, *Plasmodium* species present three main inserts (red frames): a three residue insert N101 -K102- I103, and 13 residues insert (L247 to I259) and a less evident insert toward the end of the alignment (see Figure 2-1).

The NADPH binding domain is characterised by the Rossmann fold at the N-terminal. This fold is composed of two $\beta\alpha\beta\alpha\beta$ units as illustrated by the secondary structures in the alignment. Residues interacting with NADPH are highlighted in green in the alignment. The turn region between the first strand and the next helix is characterised by a consensus binding pattern GXXGXXG in which the first two (2) glycines participate in NAD(P)-binding, and the third facilitates close packing of the helix to the beta-strand. The conserved structural motif GXXGXXG as shown in the alignment is known for dinucleotide-binding proteins and the second glycine of the motif play an important role in the recognition of both cofactor and substrate (Jang et al. 2007). *P. vivax* and *P. knowlesi* present GSSGSI motif while the remaining *Plasmodium* species have a GSTGSI motif instead. The threonine (THR86) residue implied in the cofactor binding is substituted into a serine residue in *P. vivax* and *P. knowlesi*. SER117 is changed to an asparagine in *P. yoelii*, *P. berghei* and *P. chabaudi*. They all are rodent parasites.

In the flexible loop region (residues 291 to 299), except for LYS297 in PfDXR, all residues are completely conserved across all *Plasmodium* species but presented little variations compared to those observed in other species. The loop presents a highly conserved motif HPXWXMG (Kholodar et al. 2014) which is illustrated in the alignment. Residues (His293, Pro294, Trp296, Met298 and Gly299), reported to be important for enzyme ligand/substrate interaction, are completely conserved in all organisms. The conservation of the residues His293, Trp296, and Met298 in the flexible loop region were also reported in the literature (Umeda et al. 2011; Murkin, Manning, and Kholodar 2014; Deng et al. 2010; Sooriyaarachchi et al. 2016). These buried residues have been associated with loop closure (Umeda et al. 2011). LYS297 remains the most variable residue in that position. The residue is substituted by an Asparagine in *P. ovale*, *P. yoelii*, *P. berghei* and *P. chabaudi*.

Residues Asp231, Glu233, and Glu315 (highlighted in pink) are highly conserved in all DXR family members as shown in the alignment. They coordinate the metal ion in the active site (Singh et al. 2007; Wadood et al. 2017; Kunfermann et al. 2013). Except for SER269 which is conserved in *Plasmodium* species but showed variations in the other species (see Figure 2-1), residues implied in the phosphonate moiety binding, Ser270, Ser306, Asn311, Lys312, His293 (highlighted in red), are also completely conserved.

2.3.3 Phylogenetic tree

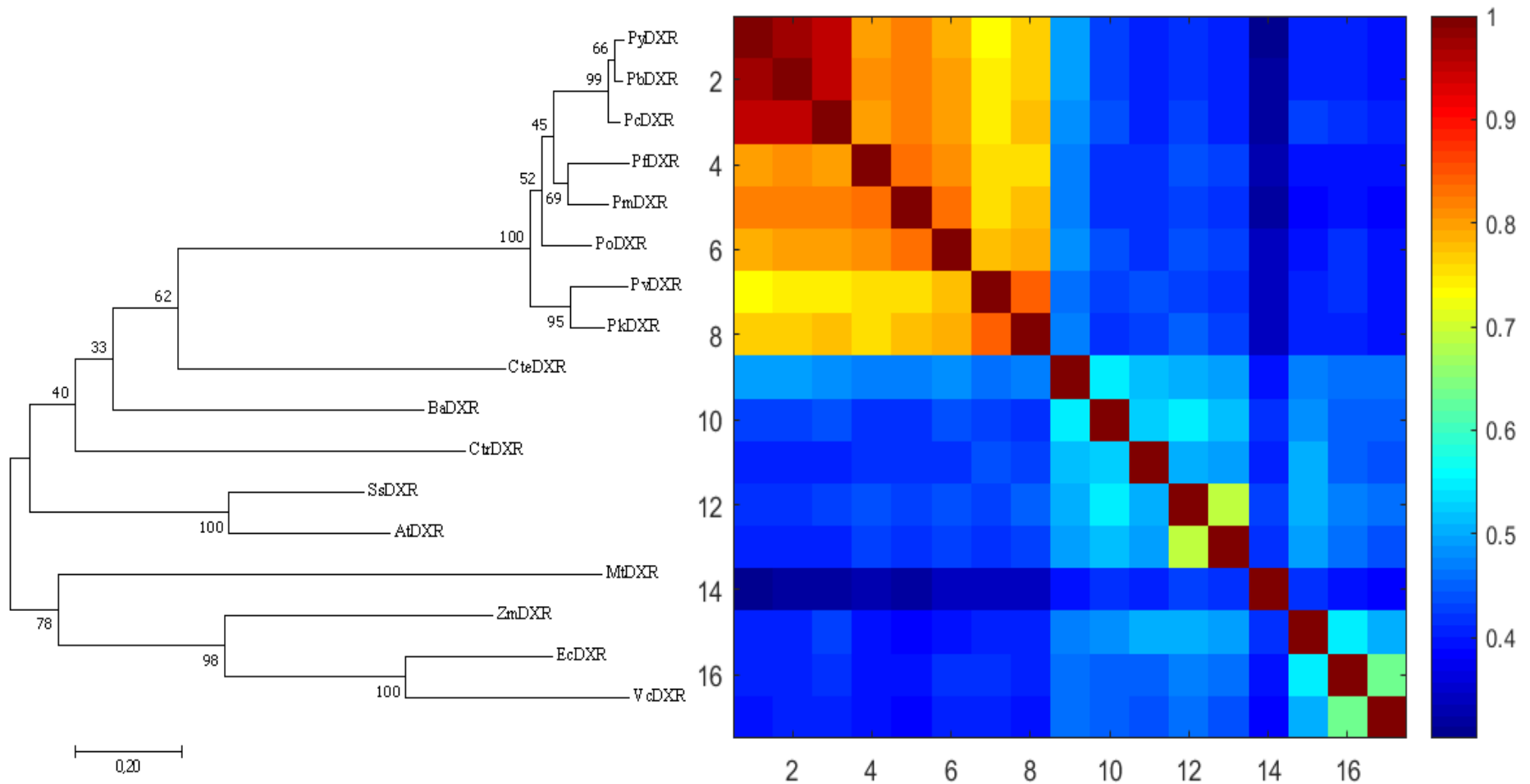


Figure 2-2: Left: Phylogenetic Tree for 17 DXR protein sequences constructed using the software MEGA7 and the Maximum Likelihood method. A threshold of 100% was applied for site coverage. Left bottom: Scale for evolutionary change, substitution per site. Right: Pairwise sequence identity matrix. Colours represent degree of similarity (Red (1) 100% identical to Bleu (0) 0% percent identical.) The pairwise sequence identity is calculated based on the same alignment used to construct the phylogenetic tree. The order of sequence in the tree is the same as in the identity matrix (1: PyDXR to 17 VcDXR).

LG (Le and Gascuel 2008)+G (discrete Gamma distribution to model evolutionary rate differences), LG+G+I, and WAG (Whelan and Goldman 2001)+G+I models showed the lowest BIC scores. The LG+G model was selected as it generated the tree with the highest likelihood (-8048,56).

Phylogenetic analysis was performed to analyse evolutionary relationship between the different DXR protein sequences. As expected all sequences share the same root, as they are all from the same DXR (1-deoxy-D-xylulose 5-phosphate reductoisomerase) protein family. The *Plasmodium* species (PyDXR, PbDXR, PcDXR, PfDXR, PmDXR, PoDXR, PvDXR, PkDXR) clustered under the same clade. In the same way, gram positive (CteDXR, BaDXR, CtrDXR), gram negative (MtDXR, ZmDXR, EcDXR, VcDXR) and plant (AtDXR) clustered together (see Figure 2-2). Furthermore, the rodent (*P. yoelii yoelii*, *P. chabaudi* and *P. berghei*) DXRs formed a sub-cluster independent from the other *Plasmodium* species. Interestingly SsDXR, from *Synechocystis*, a freshwater cyanobacterium, had the closest relationship with DXR from *Arabidopsis thaliana* as reflected by their clustering in the tree and their clear green colour in the identity matrix. This coloring corresponds to sequence identity between 60% and 70%. Indeed, the two sequences share 65% sequence identity. This is consistent with previous findings. In fact, plant DXRs are from an endosymbiotic origin and obtained by gene transfer from *Synechocystis* (Lange et al. 2000).

Some of the bootstrap support values were low (<70%), a bootstrap proportion $\geq 70\%$ corresponding to a probability of $\geq 95\%$ that the respective clade is real (Hillis and Bull 1993). Nonetheless, these values were consistent across the different generated trees using different the three best models with different site coverage thresholds.

2.3.4 Motifs analysis

MEME (Bailey et al. 2006) identified a total of 21 unique motifs across the set of sequences (see Figure 2-3). All motifs were statistically significant, the lowest in term of significance being motif 21 with an E-value of $1.6e-019$.

The first and most highly conserved motif 1 starting at position 293 in PfDXR corresponded to mainly residues in the loop region of the protein (residues HPKWKMGKKITIDSATMMNK in PfDXR). HIS293, PRO293, MET298, GLY299, TRP296 in this motif have been reported in different literatures to play major role in the protein inhibition (see Table 1-1 Table 1-1: Some PfDXR residues and their identified/suggested roles in inhibition.). Motif 3 starting at position 224 in PfDXR covers the protein active site. It has two significant residues: APS231 and GLU233 implied in binding to the inhibitor/substrate hydroxamate moiety (Umeda et al. 2011). Motif 5 follows motif 1 in the protein sequence and cover GLU315 of the active site, and is implied in substrate binding and the metal coordination (Kholodar et al. 2014).

As general observation, conserved motifs might be related to important functional regions of the protein and often reported in literature. Nonetheless, some conserved motifs that we identified, for example, motif 2 and motif 6, cover residues of the protein with no related known function reported in literature. It would be interesting to further investigate the functionality related to these motifs.

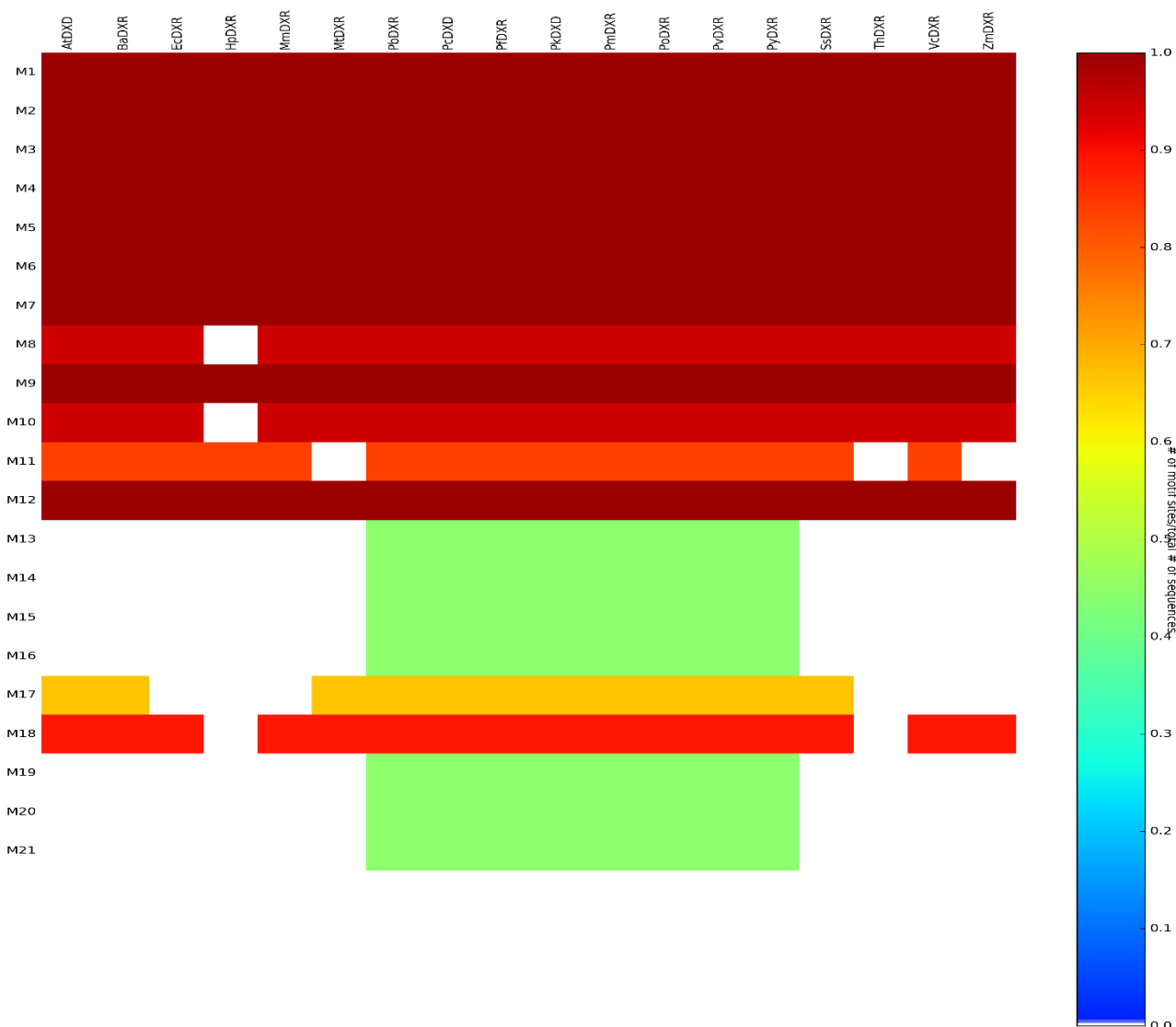


Figure 2-3: Motif analysis results. The coloring shows the degree of conservation of the motif (red: highly conserved to blue: less conserved). The related motif logos are in appendix A.

During the iterative search varying the number of motifs to be searched and the minimum and maximum length of motif. Long length motifs corresponded to the protein domains, for example, the NADPH binding domain while shorter ones were around key conserved smaller regions of the protein covering known motifs or key conserved residues. An interesting example was the case of motif 4 which is 20 amino acid long. It contains a common DXR motif, the Rossman fold motif (GXXGXXG) which is 7 amino acid long and present in all DXR sequences (see Figure 2-4). As illustrated in the motif logo, we can see GXXGXXG of the NADPH binding domain standing out with 3 to 4 bits.



Figure 2-4: Rossman fold (GXXGXXG) motif found in motif 4 of the MEME (Bailey et al. 2006) analysis.

As expected, some motifs were unique to *Plasmodium* sequences due to the higher degree of conservation in these sequences. They are motif 13, 14, 15, 16, 19, 20, and 21. Motif 15 corresponds precisely to the large insert in the red box seen in the sequence alignment. In the same way, motif 14 corresponds to the insert in the ending region of the alignment starting at position 447 in PfDXR. A similar observation is made on motif 16 (see Figure 2-1). The remaining motifs (13, 19, 20, 21) show high conservation of residues. These motifs are clear inserts in *Plasmodium* sequences as shown in the MSA (see Figure 2-1). Interestingly the first insert in the alignment did not result in any unique motif to *Plasmodium* sequences. The closest motif to this region is motif 19 which starts with the last (Isoleucine) residue of the insert in the alignment.

2.4 Conclusion

17 DXR (1-deoxy-D-xylulose 5-phosphate reductoisomerase) protein sequences were retrieved from PlasmoDB and NCBI. Sequences showed high identity and coverage with respect to the query sequence (*P. falciparum*) underlining their homology relationship. Sequence analysis through MSA revealed the key conservation of residues implied in the enzyme active site and implied in the cofactor binding. The Rossmann fold motif, the GXXGXXG motif at the N-terminal, residues in the active site and binding cofactor, and the flexible loop show high degree of conservation. The main noticeable different among the *Plasmodium* sequence is certainly the substitution of SER117 to an asparagine in *P. yoelii*, *P. berghei* and *P. chabaudi*, all rodent parasites.

The phylogenetic analysis showed the clustering of *Plasmodium* sequences. The tree also highlights the origin of plants DXR which was transferred from *Synechocystis* through an endosymbiotic process. The motif analysis was generally in agreement with the sequence alignment.

In the next chapter, homology models of the different *Plasmodium* species will be built. As shown in the sequence analysis, a conservation of residues implied in the cofactor and enzyme active site is expected.

CHAPTER 3: HOMOLOGY MODELLING

3.1 Introduction

High-throughput sequencing technologies by offering cost-efficient and fast techniques have revolutionized biology. These new sequencing techniques have made it easy to generate large amounts of genomic data (Marx 2013). As of August 2015, 187 million biological sequences were deposited in Genbank databases with 50 million protein sequences were deposited in the UniprotKB database (Kc 2016). Unfortunately, protein sequences do not provide much insight on their functionality. Protein functions are mainly related to their three dimensional structures. Enzymes, for example, can provide the spatial arrangement for their substrates to allow for a catalytic reaction. A 3D structure thus provides much better understanding of protein's mechanism of action. This provides greater insight to understand a protein and elucidate its interactions with other proteins and ligands, thus giving the opportunity to modify it. Knowledge of structural information is thus essential for structure-guided drug discovery (Meng et al. 2011).

There are around 100 000 (August 2015) experimentally solved proteins structures in the Protein Data Bank (PDB) (Berman et al. 2000). So only around 0.2% of the known protein sequences have an associated solved structure. Hence there exist a huge sequence-structure gap (Kc 2016). NMR (Nuclear magnetic resonance spectroscopy), X-ray crystallography, electron microscopy are the main experimental techniques used to determine protein structure. These techniques are expensive, time consuming and require much expertise. Comparing them to the efficiency of sequencing techniques, the challenges associated with the experimental methods for solving protein structures associated with experimental technique do not clearly support an optimism for a reduction of the sequence-structure gap. That existing situation has paved the way for computational techniques of solving protein structures. Comparative modeling (CM or homology modeling), threading and free modeling (FM or *ab initio*) approaches offer a different approach of solving protein structure (Kc 2016; Zheng et al. 2015).

Comparative modeling (also known as homology modeling) is a computational technique to predict a protein 3D structure from its amino acid sequence using an experimentally determined homologous protein structure. The technique is supported by two main assumptions.

Historically, firstly Christian Anfinsen established the basis for the sequence–structure–function paradigm. All the information required for a protein to fold into its functional structure is encoded in its sequence. Secondly, since 1995, 80% to 90% protein structures deposited in the PDB database do not introduce a significant number of different folds and since 2012 the number of folds in the database has remained stagnant (RCSB PDB - Content Growth Report n.d.). Protein structures tend to be 10 times more conserved than their sequences. During evolution, changes in the protein sequence generally (amino acid substitutions) result in little or no effect on the overall protein structure. This is particularly true for homologous sequences. However, some exceptions exist to this general observation. For example, Alexander et al. designed two proteins with 88% sequence identity but having dissimilar structure and function (Alexander et al. 2007). Except in rare cases, proteins presenting similar sequences, even distantly related, fold into

similar structures. Based on protein sequence similarity and length Rost (1999) determined a safe zone to assume similar structures (Rost 1999).

Based on these observations, it is possible to predict a protein structure from its amino acid sequence, and homology modeling is a successful *in silico* approach. During the April-June 2003 SARS (Severe acute respiratory syndrome) epidemic, a homology model for SARS coronavirus (SARS-CoV) Mpro, was used to dock the compound AG7088 into the substrate binding site. AG7088 indeed was shown to have anti-SARS activity *in vitro* (Anand et al. 2003). Using homology modeling, Cui et al. investigated the agonist binding mode to the Dopamine (D₃) receptor for the design of new inhibitors (Cui et al. 2010). In many other cases in drug design, homology modeling has been successfully applied (França 2015).

This chapter aims to build accurate 3D models for *P. malariae*, *P. vivax*, *P. ovale*, *P. knowlesi*, *P. berghei*, *P. yoelii yoelii* and *P. chaubadi* preceded by a detailed description of the methodology used and the steps in homology modeling. These structures will be modelled in two different conformations: from template 5JAZ (closed conformation) with active site ligand and metal maintained and from 1K5H which presents an open loop conformation of the protein.

3.2 Steps in homology modeling

Comparative modeling determines a protein structure based on an already known structure. The structure to be determined is often referred as the target and the known structure used as the template. These two sequences need to be homologous hence the name homology modeling.

Thus, the first step in the modeling process is the identification of a suitable template. The Protein Data Bank (Berman et al. 2000) where protein structures are deposited is used to search for a template. A simple method is a BLAST (Altschul et al. 1997) search against the PDB (Berman et al. 2000) database. Beside the classic BLAST search, other homology detection tools exist. PSI-Blast (Position-Specific Iterative Basic Local Alignment Search Tool) (Altschul et al. 1997) is more sensitive for remote homologs detection. HHPred incorporates Hidden Markov Models (HMMs) which includes insertion and deletion information. Using HMM-HMM comparison has proven to improve search sensitivity and selectivity (Söding, Biegert, and Lupas 2005).

The search result can present different templates and choosing an appropriate one is a critical step. These initial steps are important and they can significantly impact the remaining processes and the quality of the produced homology model. Different criteria need to be considered when selecting a template. The most important is to find a homologous protein for the target protein. One can conclude a homology relationship between two proteins based on their sequence identity and length. Rost established a graph to determine protein homology. From it two major observations could be made in the context of homology modeling: short sequences require much higher similarity to be inferred as homologous and the graph presents a twilight zone in which inferring homology is risky (Rost 1999). Other than the sequence identity, other factors need to be considered. The next important condition is the query coverage, the fraction of the target sequence covered by the template sequence. One needs to also consider the predicted secondary structures of the target and compare its consistency with the one of the template. Finally, an important feature is the quality of the template structure. The Protein Data Bank has quality criteria for the deposited structures. They include the resolution, the R-value, Rfree-value,

clashscore, Ramachandran and sidechain outliers and the presence of ligand(s). One should also investigate the presence of missing residues; this is especially important if long sections of residues are missing, or the missing residues are present in important regions of the protein (Sliwoski et al. 2014; Pevsner 2009).

An approach at this step is to assess the template using the model quality assessment tools.

If no suitable unique template is found, an alternative approach is multi-template modeling. Different suitable templates can be chosen to fit different portions of the target sequence. Regions without template coverage remain the most difficult to model and the most error prone in homology modeling (Fiser and Šali 2003; Webb and Sali 2016).

After selecting a template, the next step is sequence-target alignment. Even though, the template search itself is based on alignments, such alignment may not be optimal as they use heuristic approaches favorizing speed. Different CASP experiments have underscored the accuracy of this alignment as a critical step for the final model quality. A 3.8 Å distortion can be introduced in the final model as the result of only one residue shift in the alignment (di Luccio and Koehl 2011) and alignment errors are practically unrecoverable. MODELLER can use an iterative alignment to improve alignment quality and thus improve the quality of the resulting models. Using a genetic algorithm, alignments are improved and when tested; in a test using this approach the resulting model's accuracy increased from 43% to 54% (accuracy: percentage of the model C α atoms within 5 Å of the matching C α atoms in the native structure) (Webb and Sali 2016). A MSA, including structural data and evolutionary information can be useful at this step, and this is especially important for distantly related proteins. However, it is noteworthy that including too many sequences in the MSA may decrease the model accuracy (Kryshtafovych, Fidelis, and Moult 2014). One must avoid gaps in the alignment, especially large gaps, as they result in no template to model from. Also, visually inspecting the template structure and analysing the effect of alignment gaps at the structural level and correcting them can be useful (Krieger, Nabuurs, and Vriend 2005). MSA tools such as PROMALS3D (Pei, Kim, and Grishin 2008) and MUSCLE (Edgar 2004) can be used. The alignment guides the modeling process.

The modeling process first maps the template backbone coordinates to the target based on the alignment. For identical residues, side chains can also be mapped (Krieger, Nabuurs, and Vriend 2005). Different approaches can be used at this step: satisfaction of spatial restraint used by MODELLER (Fiser and Šali 2003), or segment matching and assembly of rigid bodies. MODELLER uses satisfaction of spatial restraint as its modeling approach. The method works in two steps. First, based on the target-template alignment, homology-based restraints are generated. A probability density function is used to generate the protein 3D structure. Considering the type of residue, the dihedral angles and the backbone C α atoms distances, the most probable 3D structure that optimizes the density function is generated (Šali and Blundell 1993). Secondly, a force field CHARMM-22 is used to derive the stereochemical restraints for bond lengths and bond angles. The two types of restraints are finally combined into an objective function. The model is then obtained by optimization of the objective function. When the target sequence has regions which are not aligned with the template, these regions are modelled through loop modeling, a difficult problem in this type of modeling. As unstructured regions, they are very flexible and can adopt very diverse conformations. However, loops can play important functional roles in proteins,

in the formation of active sites for example. Two approaches can be used to address this issue: a database-based method by searching in a specific database of loops facilitated by the availability of more and more structures, and a conformational search approach. This latter is much more diverse and based on *ab initio* predictions. MODELLER implements a loop-modeling module using an optimization scoring function (Webb and Sali 2014; Krieger, Nabuurs, and Vriend 2005).

It is notable that at this step, a model can be refined using an energy function. Energy minimization is employed to remove unfavourable geometric imperfections such as bond lengths and angles to finally produce a refined structure. Another approach is through molecular dynamics simulations. However, these techniques should be carefully used as they may result in incorrect structures with atoms removed from their correct places (Xiong 2006). About side chain positioning, this remains a very difficult and challenging task as these can adopt various conformations and errors can often happen in side packing. These chains remain important as they can be involved in interactions with other proteins and ligands. MODELLER also uses spatial restraints to model side chains. Another approach is through exhaustive conformational searching, but this is computationally expensive. As a compromise, a search is conducted in a rotamer library of preferred side chains to select a rotamer with lowest energy. The software program SCWRL4 has shown good precision in side chain positioning (Krivov, Shapovalov, and Dunbrack 2009:4).

Finally, after model building, its quality is assessed. In the absence of the real structure of the protein, it remains difficult to know how far is the model from the real structure. Still, an approach is to evaluate the model self-consistency. The model compliance with the physicochemical rules such as bond lengths and distance, dihedral angles is assessed. PROCHECK (Laskowski et al. 1993) and WHATCHECK (Hooft et al. 1996) are examples of programs which assess the stereochemistry of a protein structure. The second approach is knowledge-based and derives statistical potential, 3D-profiles from experimental structures. MODELLER uses the DOPE-Z score (Discrete Optimized Protein Energy and Z for normalized and the score considers the full length of the protein), a statistical potential derived from structures in the Protein Data Bank using probability density functions. ANOLEA (Melo and Feytmans 1998) and VERIFY3D (Lüthy, Bowie, and Eisenberg 1992) are examples of tools using statistical method for model assessment.

Other measures for a model assessment consider the similarity between a model and the template it was modelled from. Some of these scores include the RMSD (Root Mean Square Deviation), the GDT-HA (Global Distance Test-High Accuracy), and the LDDT (Local Distance Difference Test). The RMSD computes the average distance between of atoms in the superimposed structure (model and the template). Generally, only the backbone atoms are considered (Kufareva and Abagyan 2012). The TM-score like the RMSD evaluates two structure similarity and was introduced to solve two limitations with the RMSD. It does depend on protein's length and evaluates global fold similarity being thus less sensitive to local variations (Gadzała et al. 2017). As for the GDT is a global score for the model reflecting the accuracy $C\alpha$ of positions using specific cut-off distances. GDT-HA is an improved version using more stringent cut-off distances (Huang et al. 2014). Finally, the LDDT is a measure for a model local quality. It is based on the local distance difference of all atoms in a model as well as validation of stereo chemical quality of the model (Mariani et al. 2013).

Various errors can happen in the modeling process. They include errors in the side chain packing, those resulting from misalignment, and selection of an inappropriate template. It is noteworthy that even in correctly aligned regions, distortions and shifts can happen due to sequence divergence. Although some errors can happen in homology modeling, it remains the most accurate *in silico* approach to determine protein structure. The technique can get very close to experimental techniques such as NMR and X-ray crystallography producing models with RMSD as low as 1.5 for high sequence identity (90%). The technique has many applications, particularly the study of catalytic mechanism, defining antibodies epitopes, refining NMR structures and molecular replacement in x-ray crystallography and proteins ligands interactions in docking and virtual screening (Webb and Sali 2016).

3.3 Methodology

3.3.1 Template identification

As the most important criterion for template selection in homology modeling is sequence identity (Webb and Sali 2016), *P. falciparum* DXR structures available in the PDB database were first assessed. The other sequences of the *Plasmodium* will present highest sequence identity to *falciparum* than other organisms. A Python script was written to count missing residues in the structures and their positions.

As the application for the models was molecular docking, the second important criterion was the protein configuration and also the presence of ligand. As shown in the literature, DXR active site is highly flexible. The protein undergoes conformational changes after ligand binding with movement of the flexible loop over the active site. That open conformation presents the advantage to accommodate larger ligands (Mac Sweeney et al. 2005). So an inhibitor free structure would be interesting to study in view of the large structural motifs present in SANCDB (Hatherley et al. 2015).

Also, as 1-deoxy-D-xylulose 5-phosphate reductoisomerase is a homodimer, only one chain is necessary for docking (the active site is far-removed from the interface). So, the chain presenting the highest quality in the dimer from the selected structure was used as template.

HHPred was then used for our template identification. HHpred is a web server for protein homology detection. It automatically can search in multiple databases including PDB (Berman et al. 2000), CATH (Class, Architecture, Topology and Homology) (Sillitoe et al. 2015) , SCOP (Structural Classification of Proteins) (Murzin et al. 1995) and makes use of PSI-BLAST and HMM-HMM (Hidden Markov Model) profile comparison for remote homologs detection. Using this latter greatly improves the sensitivity and selectivity. The server also scores the matches between target predicted secondary structure using PSIPRED and the template one (Söding, Biegert, and Lupas 2005). The server was queried on PDB_mmCIF70 database for modeling using its default parameters with each of the *Plasmodium* protein sequences. The sequences were trimmed to not include the endoplasmic reticulum signal and plastidial targeting sequences.

For template selection, the following criteria were considered: the sequence % identity, the query coverage, the E-value, and the quality of the match between secondary structures. Threshold values were set for the coverage of the target (at least 20%), the E-value ($1e-3$) and the probability

(>10%). As sequence identity is an important criterion, available *Plasmodium* structures were first assessed.

Two templates were finally selected: 5JAZ (chain B) for modeling *Plasmodium* DXRs in closed conformation and 1K5H (chain A) for their modeling in open conformation.

For template evaluation, the following characteristics of the crystal structure were considered: the resolution of the 3D structure, the R factor and R free values, the presence of ligands in the structure and the completeness of the structure. HHpred server results give the convenience to directly compare the secondary structure of the crystal and the predicted one for the target sequence by scoring secondary structure similarity.

3.3.2 Template-target alignment

Our template and target sequences were approximately of same length and had high sequence identity (above 70%), any alignment method with reasonable parameters would result in the same alignment. Hence, MODELLER's ALIGN2D command was used. The method uses dynamic programming and considers also structural information of the template. The method also tends to eliminate gaps from secondary structure regions and locate them in between close alpha carbon in space, curved and solvent exposed regions. This reduces error in the model building. The command (ALIGN2D) aligns only residues present in the atom section of the PDB file and also produces the alignment in PIR format to be used with modeler (Webb and Sali 2016). A simple Python script (fasta_to_ali.py) was written to quickly convert every *Plasmodium* Fasta file into the Ali format as input for ALIGN2D.

For the template 1K5H, the sequence identity was relatively low across the different species compared to 5JAZ (see Table 3-1 and Table 3-2). So, MSA was used for a more accurate alignment in contrast to a simple pairwise alignment used in the case of 5JAZ. The sequence from 1K5H was added to the previous MSA generated in the sequence analysis chapter. The set of sequences was then realigned using MUSCLE (MULTiple Sequence Comparison by Log- Expectation) (Edgar 2004).

Finally, Jalview (version 2.10.1) was used for alignment visualization and manual corrections. The sequences were trimmed at the N- and C-terminal. An important adaptation of the alignment file was the specification of ligands and metal ion in the active site.

3.3.3 Homology modeling

Modeller (Webb and Sali 2016) version 9.19 32-bits was used for the modeling. The modeling script "get-model.py" was obtained from the tool documentation and adapted to each model to be produced. The alignment PIR files and sequence codes were passed to the script. 100 models were generated for each protein using very slow refinement. The models were produced while maintaining the template ligands at their positions. The ligands were thus treated as rigid body and transferred to the models. This maintained the binding site geometry and environment reasonable similar to the template. The active site thus maintained an environment similar to ligand bound conditions. Its geometry is thus preserved for docking application (Šali et al. 2017).

All modelings were carried out on a laptop: Processor: Intel(R) Core(TM) i7-3740QM CPU @ 2.70GHz (8 CPUs), ~2.7GHz, Memory: 8192MB RAM, Operating System: Windows 10 Home 64-

bit. Discovery Studio 2016 (Biovia, San Diego, CA) was used for visualization of the produced models.

SCWRL4 (Krivov, Shapovalov, and Dunbrack 2009:4) was used in an effort to improve side chains positioning. The original models from MODELLER were compared with the outputs from SCWRL4 to judge significant improvement of the models using their Dope-Z scores.

After modeling of the open conformation for the different *P. falciparum* sequence, a Python script was used to reassign the correct residue numbers as in the closed conformation.

3.3.4 Model evaluation

There are many different software and web-servers for to evaluate model quality. These tools can employ different approaches for model assessment. So, using different tools with different algorithms is good practice to have a comprehensive analysis of the model quality. Some assessment programs also provide means to verify local quality of models, i.e. at the residue level. Here, model quality assessment programs combining consistency with physicochemical rules, knowledge-based methods and assessment on model global and local quality were used. Modeller provides DOPE, GA341 and SOAP (statistically optimized atomic potentials) (Dong et al. 2013) scores for model assessment. DOPE and SOAP showed to be better at distinguishing good models from bad models compared to GA341 (Modeller Tutorial available at <https://salilab.org/modeller/tutorial/basic.html>). MODELLER DOPE Z-score was used and provided a mean to assess using a statistical potential.

The tools used for model assessment include Modeller (its DOPE Z-score), QMEAN, PROCHEK, ProQ3D and DFIRE. The models were first filtered by the DOPE Z-score. The best five models for each protein per DOPE Z-score were selected. These models were then assessed using a Python script on the available QMEAN API (Application Programming Interface). QMEAN provides an evaluation of 'degree of nativeness' of a model. The tool can also be used for local quality assessment (Benkert, Biasini, and Schwede 2011). For each species, models having the best QMEAN Z-score were finally selected.

PROCHECK was used to assess models' self-consistency and their stereochemistry. A more recent method to model assessment is through machine learning, ProQ3D uses deep learning and combining ProQ2 and Rosetta energies (Leaver-Fay et al. 2011). The method achieved state of arts performances in CASP12 in the MQA category (Uziela et al. 2017). DFIRE (distance-scaled, finite ideal-gas reference), a knowledge-based all-atom potential based on a distance-scaled finite ideal-gas reference state was used. The tool evaluates non-bonded atomic interactions in the protein model (Zhou and Zhou 2002). Assessment was mainly conducted in the SWISS-MODEL assessment Workspace (Arnold et al. 2006). Models were assessed by using their template as reference.

3.4 Results and Discussion

3.4.1 Template identification

The different crystal structures of *Plasmodium* DXR were first assessed (Complete table in appendix B. All potential templates had missing residues at the terminal regions. It is a common

observation to have missing residues in protein structure at the terminal regions. These regions pose difficulty in crystallography as they are often flexible. Templates with too many missing residues were eliminated in the selection process. This was especially important for potential template 3AU8 (chain A) as it has all residues in the flexible loop region covering the active site (residues 291-299) missing. Although alternative in these cases can be to use multi-template modeling in MODELLER, but there were enough structure available to avoid that backup solution.

About the different crystal structure conformations, Reuter et al. solved a DXR crystal structure revealing its highly flexibility, especially for the flexible loop region. In the three independent molecules (A, B, C) of the asymmetric unit, the loop shows very different conformations. 1K5H (chain C) is in open conformation (flexible loop covering the active open) with unbound inhibitor (Reuter et al. 2002). It would have been interesting to use an open conformation for in the docking experiment (see Chapter 4). The choice of this template is thus mainly motivated by its unique conformation (Figure 3-1). Another considerable difference between the templates used for modeling is the absence of metal ion in 1K5H. The difference of conformation could also influence significant differences in molecular docking to these structures.

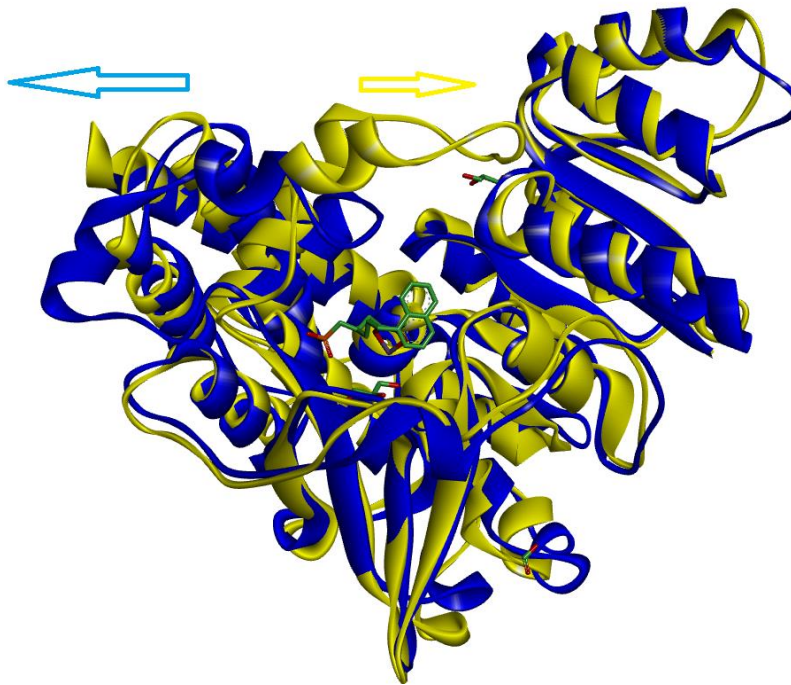


Figure 3-1: Molecular overlay of 5JAZ (chain B, closed loop with inhibitor in active site) and 1K5H (chain A – open loop conformation). Arrows show the orientation of the loop.

Table 3-1: Metrics for modeling from 1K5H obtained using NCBI BLASTp search (Word size: , Expect value: 10, Hitlist size: 100, Gapcosts: 11.1, Matrix: BLOSUM62).

Query	BLAST Hit	Score	E-value	Identity	Query Length	Query coverage
PfDXR	1K5H	268	1e-87	37%	411	91%
PbDXR	1K5H	273	2e-89	39%	411	90%
PcDXR	1K5H	278	3e-91	40%	411	90%
PkDXR	1K5H	273	4e-89	39%	413	92%
PvDXR	1K5H	275	6e-90	38%	413	98%
PmDXR	1K5H	264	9e-86	38%	413	89%
PyDXR	1K5H	272	5e-89	39%	411	90%
PoDXR	1K5H	275	4e-90	38%	410	99%

Table 3-2: Selected hits from HHpred result.

Query	1 st HHpred Hit	Score	E-value	Probability	Identity	Score Secondary structure	Cols	Target Length
PbDXR	5JAZ_B	703.81	2.2E-92	100	78%	47.3	407	411
PcDXR	5JAZ_B	709.44	2.5E-9	100	79%	48.1	407	411
PkDXR	5JAZ_B	701.98	3.1E-92	100	74%	46.5	404	413
PvDXR	5JAZ_B	725.82	1.8E-95	100	73%	43.5	406	413
PmDXR	5JAZ_B	770.87	2.8E-101	100	82%	50	407	413
PyDXR	5JAZ_B	699.54	2.6E-92	100	77%	47.5	406	411
PoDXR	5JAZ_B	698.42	2.1E-92	100	80%	47.9	406	410

The column 'Cols' (see Table 3-2) indicates the number of matches in the target-template alignment. The probability (in percentage) that the hit is a true positive, a homolog to the query sequence, at least in some core part (Söding, Biegert, and Lupas 2005).

The HHpred search consistently returned the same template for the different *Plasmodium* species. All results indicated 100% probability confirming the homology relationship. The lowest E-value was 2.8E-101 for PmDXR (see Table 3-2). For the sequence identity, the lowest value was 73%. These high sequence identity values support the suitability of the template 5JAZ for the homology modeling of other *Plasmodium* sequences.

Interestingly, HHpred did not return the other *Plasmodium* structures present in the PDB database which were found in our assessment of available *Plasmodium* structure. HHpred search is limited to "PDB_mmCIF70 for modeling" which seems to be a filtered PDB for modeling.

The template is a recent crystal structure from *Plasmodium falciparum*. It has a resolution of 1.4 Å, a R-Value Free of 0.207 and a R-Value Work of 0.185 (see Figure 3-2). The crystal structure presents highest values of zero (0) for the Clashscore and the Ramachandran outliers. It has only 0.9% sidechain outliers and though present a low RSRZ outlier value of 8%. It contains a glycerol

molecule and a formic acid molecule. At the active site, there is a manganese ion Mn^{2+} and an arylpropyl substituent on the reference inhibitor fosmidomycin. The bulky substituent displaces the key tryptophan in the active-site flap to accommodate the ligand (Sooriyaarachchi et al. 2016). This structure, thus provides a wider binding pocket for docking purposes. As the application was docking and molecular dynamics, the presence of ligand(s) in the different crystal structure was also considered. The template was also selected as it a good inhibitor bound with the tryptophan residues (TRP296) of the flexible moved to accommodate the ligand.

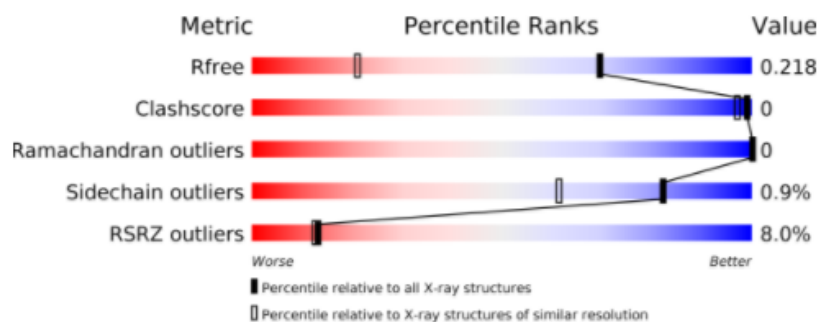


Figure 3-2: 5JAZ PDB metrics for structure quality

About the structure completeness, the first 10 and the last 2 residues of this template were missing. Missing residues were found in all *Plasmodium* DXR crystal structures as indicated in the table above. The beginning region (residues 67 to 76) were missing in all structures. Coordinates for missing residues are not available and MODELLER does not automatically handle missing residues (Webb and Sali 2016). A solution is to use MODELLER to “fill in” the missing residues. Using either MODELLER automodel or loopmodel class, a new model with missing residues filled in is built using the original crystal structure. But both classes can move the non-missing residues from their position. To avoid that, the select_atoms method can be overridden to select only the missing residues. In our case, missing residues were the first 10 residues of the disordered region and two residues at the end of the sequence.

Interesting HHpred results consistently indicated chain B for modeling. The chain B for 5JAZ was selected as template. The reason being that residues in Chain A had poor fit the electron density. In fact, the structure validation report indicates that 11% of residues in chain A have a RSRZ normalised real-space R-value (Kleywegt et al. 2004) above 2 compared to 4% in chain B (Sooriyaarachchi et al. 2016). Thus, very few RSRZ outliers were on chain B making it of better quality than chain A and more suitable for modeling.

Significant errors can be found in experimentally determined protein structure (Wlodawer et al. 2008). Hence the need for us to thoroughly evaluate our templates before proceeding to modeling.

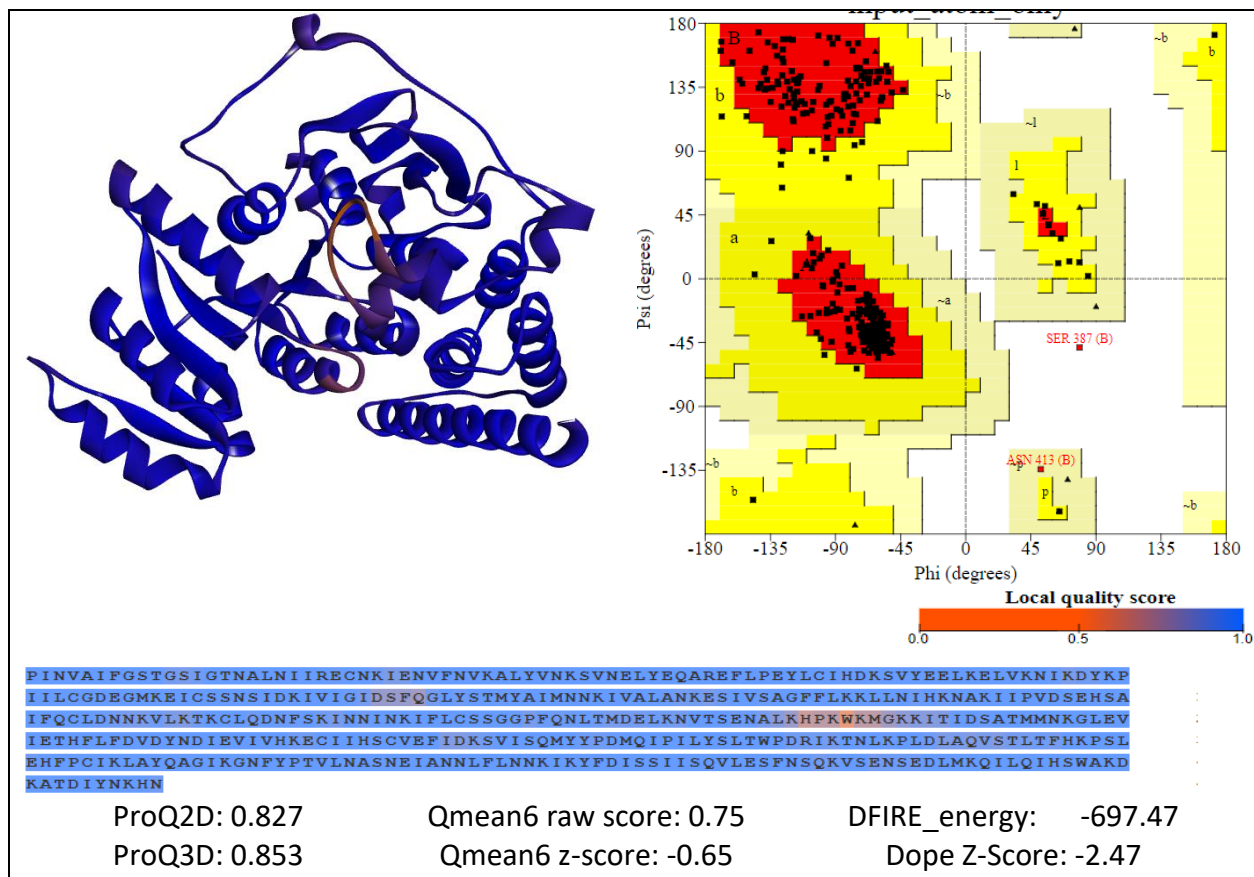


Figure 3-3: 5JAZ, Chain B Assessment.

On the Ramachandran plot, red regions indicate most sterically favoured regions, dark-yellow: the additional allowed regions, light yellow shows the generously allowed regions. The disallowed regions are in white. The black dots indicate the residues in good regions and the red those in bad regions based on the Psi and Phi angles. The structure was visualized in Discovery Studio and coloured from Orange (low quality regions) to blue (Good quality regions).

The template 5JAZ showed global good quality scores (see Figure 3-3). Only residues in the flexible loop of the active site shows an average to low quality local quality. The lowest quality being attributed to the TRP296. This can be expected as this structure contain an inhibitor with phenyl rings removing the TRP296 indole ring from its 'usual' position (Sooriyaarachchi et al. 2016). The remaining majority of the chain residues show good quality. From the Ramachandran plot, any residue is found in the outlier region and only 2% are in allowed region, the remaining 98% being in favourite regions. It is noteworthy that SER387 and ASN413 are not in favourable regions. Nonetheless these residues are not in the active site regions and are in the C-terminal region.

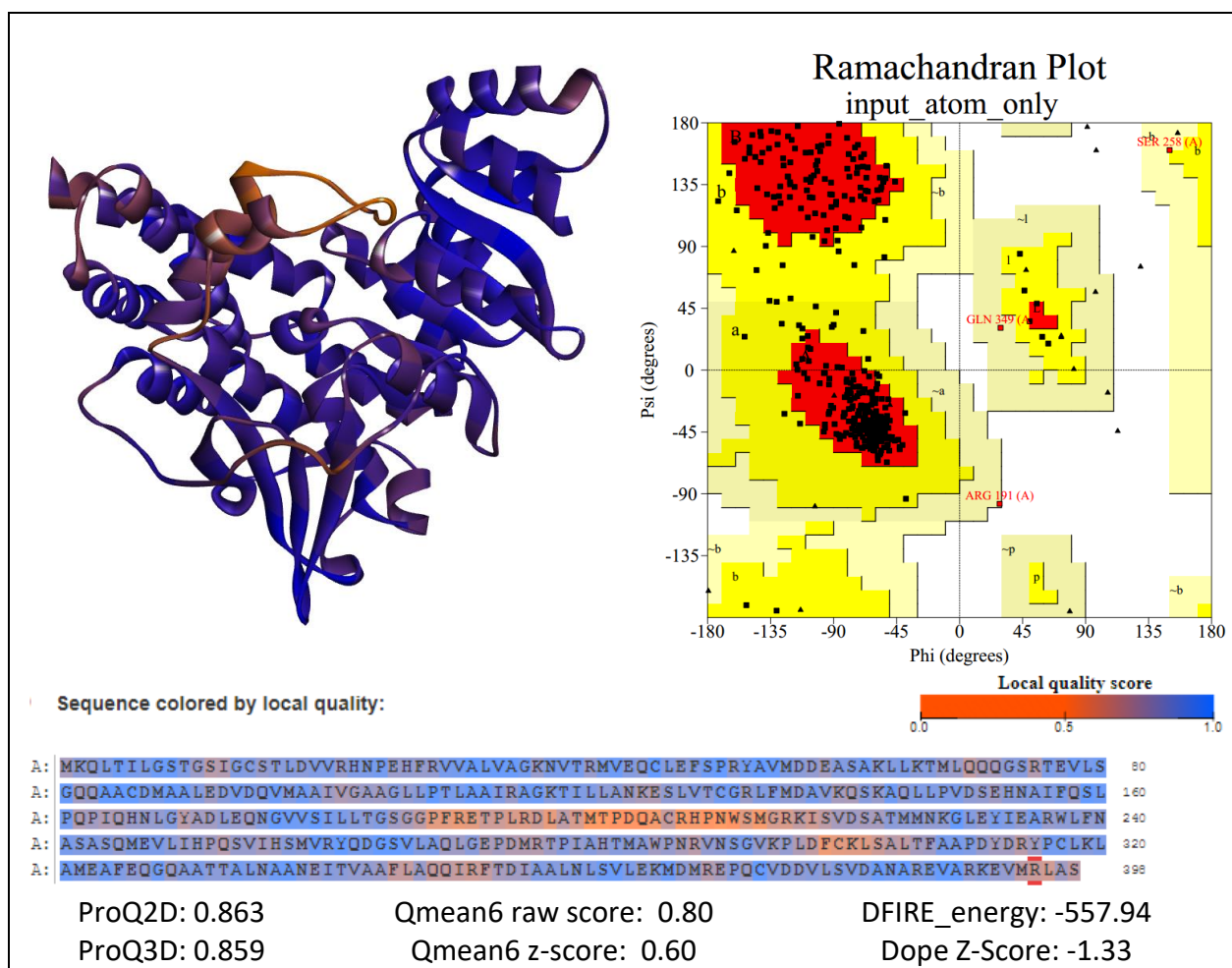


Figure 3-4: 1K5H, Chain A assessment.

On the Ramachandran plot, red regions indicate most sterically favoured regions, dark-yellow: the additional allowed regions, light yellow shows the generously allowed regions. The disallowed regions are in white. The black dots indicate the residues in good regions and the red those in bad regions based on the Psi and Phi angles (see Figure 3-4). The structure was visualized in Discovery Studio and coloured from Orange (low quality regions) to blue (Good quality regions).

1K5H is a crystal structure from *Escherichia coli* with a resolution of 2.5 Å and a R-Value Free of 0.284 (see Figure 3-5). Its resolution remains low for molecular docking and the structure is solved without any ligand. The PDB metric percentiles show lower value compared to the structures of same resolution. Chain A in the structure presents any missing residue ideal for homology modeling. The template assessment shows few residues (SER258, GLN34, ARG191) in the generously allowed regions of the Ramachandran plot and none in the disallowed regions. The Dope-Z score is -1.33, considered thus to be native. Qmean local scores illustrated in the sequence

shows only the flexible loop region (in orange color) with bad local quality. This is also common to template 5JAZ.

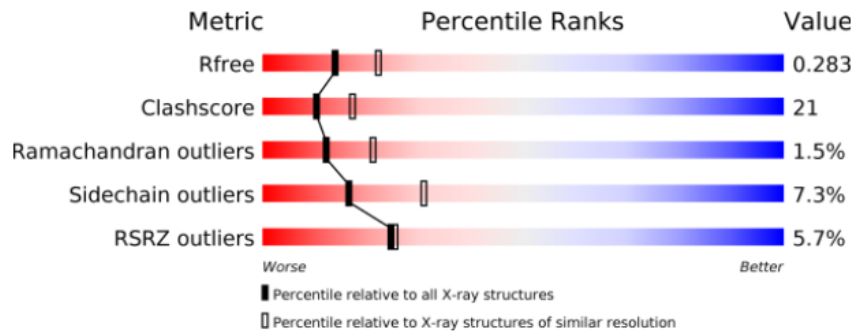


Figure 3-5: 1K5H PDB Percentile Ranks

Comparing the two templates, they both showed low quality scores for the residues in the flexible loop region. This can be expected as the mobility of these regions can present difficulty in crystallization. Overall 5JAZ showed to be of better quality. The template present higher resolution (1.4 against 2.5). The different metrics used in their assessment (Dfire, Qmean6, ProQ3D, ProQ2D) show better values for 5JAZ. However, both structures remain suitable for homology modeling purposes.

3.4.2 Template-Target alignment

```

5jazB 78  I N V A I F G S T G S I G T N A L N I I R E C N K I E N V F N V K A L Y V N K S V N E L Y E Q A R E F L P E Y L C I H D K S V Y E E L K E 146
PvDXR 1  I N V A I F G S S G S I G A N A L D V I R E C N R V E R R F N V E A L Y V N K S V T K L Y E Q A R E F L P K Y V C I H D E S K Y E E L K M 69

5jazB 147 L V K N I K D Y K P I I L C G D E G M K E I C S S N S I D K I V I G I D S F Q G L Y S T M Y A I M N N K I V A L A N K E S I V S A G F F L 215
PvDXR 70  L L R N V Q G Y N P E I L V G D D G M K Q M C S S N T L D R I I I G I D S F H G L Y S T I Y A I K S N K I I G L A N K E S I V S A G F F L 138

5jazB 216  K K L L N I H K N A K I I P V D S E H S A I F Q C L D N N K V L K T K C L Q D N F S K I N N I N K I F L C S S G G P F Q N L T M D E L K N 284
PvDXR 139 K K L L T S H T K S C I I P V D S E H S A I F Q C L D N N K V L K T K C L Q D S F S K V N Q I K K L I L S S S G G P F Q N A S L D E L K K 207

5jazB 285  V T S E N A L K H P K W K M G K K I T I D S A T M M N K G L E V I E T H F L F D V D Y N D I E V I V H K E C I I H S C V E F I D K S V I S 353
PvDXR 208 V T A E D A L K H P K W K M G P K I T I D S A T M M N K G L E V I E A H F L F D V D Y N H I E I L V H K E C I L H S C V E F I D K S V V S 276

5jazB 354  Q M Y Y P D M Q I P I L Y S L T W P D R I K T N L K P L D L A Q V S T L T F H K P S L E H F P C I K L A Y Q A G I K G N F Y P T V L N A S 422
PvDXR 277 Q M Y L P D M K L P I L Y A L T W P N R I A T E L P S L N L A S T S P L T F Y S P S L D H F P C I K L A Y Q A G R Q G N F Y P T V L N A A 345

5jazB 423  N E I A N N L F L N N K I K Y F D I S S I I S Q V L E S F N S Q K V S E N S E D L M K Q I L Q I H S W A K D K A T D I Y N K H N - - w w 488
PvDXR 346 N E V A N K L F L S N K I G Y F D I A A I I S D V L E S F T P Q E V S N N C E D L M H Q I G G I H K W A V R R A E E V Y R V R S - - w w 411

```

Figure 3-6: Graphical representation of the PIR file (target-template: PvDXR-5JAZ) alignment viewed in Jalview (Waterhouse et al. 2009). The two dots at the end indicate the positions for the ligand and the metal ion. The “w” indicate the two water molecules in the active site.

The dots at the end of the alignment (see Figure 3-6) instruct MODELLER to read from the HETATM section to include the ligands (Šali et al. 2017). *Plasmodium vivax* had the lowest sequence identity (73%) with 5JAZ compared to other *Plasmodium* species from the results in HHpred search. Nonetheless the identity was high enough for modeling for purposes and the alignment shows only a few mismatches with no gap and 100% coverage of the target sequence. The remaining *Plasmodium* had a higher sequence identity in the alignment and thus make of 5JAZ an ideal template for modeling.

```

1k5h 1 MKQLTILGSTGSIGCS TLDVVRHN - - - PEHFRVVALVAGKNVT RMV EQCLEFS PRYAVMDD EASAKLLK 66
PfDXR 1 - INVAIFGSTGSIGTNALNI IRECNKIENV FNVKALYVNKSVNELYEQAREFLPEYLCIHDKSVYEELK 68

1k5h 67 TMLQQQ- GS RTEVLSGQQAA CDMAALEDVDQVMAA IVGAAGLLPTLAA I RAGKT ILLANKES LVT CGRL 134
PfDXR 69 ELVKNIKDYKPI ILCGDEGMKEICSSNS IDKIVIG IDS FQGLYS TMYA IMNKN IVA LANKES IVSAGFF 137

1k5h 135 FMDAVKQSK - AQLLPVDS EHNA IFQS LPQPIQ - - - HNLGYADLEQNGVVS ILLTGSGGPFRETPLRDLA 199
PfDXR 138 LKKLLNIHKNAKI I PVDS EHS AIFQCLDNNKVLKTKCLQDNFSKINNINKIFLCSGGPFQNLTMDELK 206

1k5h 200 TMT PDQA CRHPNWS MGRKISVDSATMMNKGLEY IEARWLFNASAS QMEVLIHPQSV IHS MVRYQDGSVL 268
PfDXR 207 NVTSENALKHPKWKMGKKIT IDSATMMNKGLEV IETHFLFDVDYNDIEVIVHKECI IHS CVEFIDKSVI 275

1k5h 269 AQLGEPDMRTP IAHTMAWPNRVNSGVKPLDFCKLSALTFAAPDYDRYPCLKLAMEAFEQGQAATTA LNA 337
PfDXR 276 SQMYYPDMQIP ILYSLTWPDR IKTNLKPLDLAQVSTLTFHKPSLEHFPCKI KLAYQAGIKGNFYPTV LNA 344

1k5h 338 ANEITVA AFLAQQIRFTDIAALNLSVLEKMDMREPQC - - - - VDDVLSVDANA REVARKEVMRLAS - 398
PfDXR 345 SNEIANNLFLNNKIKYFDISS IISQVLESFNSQKVS ENSEDLMKQILQIHSWAKDKATDIYNKHNS S 411

```

Figure 3-7: Graphical representation of the PIR file (target-template: PfDXR-1K5H) alignment viewed in Jalview (Waterhouse et al. 2009).

The sequence of PfDXR showed the lowest sequence identity to the template 1K5H with 37%. However, the template presents good coverage for the sequence 91% (see Table 3-1). The alignment presents 3 gap regions (see Figure 3-7). They are not present in key regions of the protein: GXXGXXG motif, NADPH or fosmidomycin binding residues or in the flexible loop region. More they are short enough (the longest gap being of 5 residues) to be handled by MODELLER.

5JAZ presented better characteristics compared to 1K5H with respect to every *Plasmodium* sequence, but this latter remains satisfactory for homology modeling.

3.4.3 Modeling and Model Evaluation

The script “`assess_complete_models.py`” (appendix C) assessed the models’ Dope scores and ranked them according to their Dope-Z scores. The following table report for each model, the 5 best Dope and their Qmean scores (Best Qmean score highlighted in green).

Table 3-3: 5 Best Models for each sequence in Open conformation according to Dope-Z score. The best Q-mean score is highlighted in green.

PbDXR_open				PcDXR_open				PfDXR_open				PkDXR_open			
Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean	
		Raw score	Z-score			Raw score	Z-score			Raw score	Z-score			Raw score	Z-score
98	-0.96	0.62	-3.59	98	-0.92	0.64	-3.20	91	-0.99	0.64	-3.31	52	-0.90	0.64	-3.07
37	-0.95	0.63	-3.44	22	-0.92	0.64	-3.25	83	-0.92	0.62	-3.66	97	-0.90	0.63	-3.36
18	-0.95	0.63	-3.41	57	-0.90	0.64	-3.11	78	-0.90	0.64	-3.31	24	-0.88	0.63	-3.46
46	-0.94	0.63	-3.52	83	-0.89	0.65	-3.04	40	-0.90	0.63	-3.34	83	-0.88	0.65	-3.02
49	-0.94	0.63	-3.42	15	-0.87	0.65	-3.02	98	-0.90	0.63	-3.44	10	-0.87	0.63	-3.49
PmDXR_open				PoDXR_open				PvDXR_open				PyDXR_open			
Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean	
		Raw score	Z-score			Raw score	Z-score			Raw score	Z-score			Raw score	Z-score
25	-0.95	0.64	-3.15	16	-0.84	0.63	-3.53	93	-0.77	0.61	-3.96	19	-0.96	0.63	-3.36
87	-0.91	0.64	-3.11	84	-0.81	0.62	-3.59	16	-0.76	0.64	-3.27	70	-0.95	0.63	-3.51
90	-0.88	0.65	-2.94	50	-0.80	0.64	-3.11	34	-0.74	0.60	-4.06	100	-0.95	0.62	-3.78
27	-0.88	0.66	-2.73	12	-0.78	0.63	-3.52	66	-0.74	0.63	-3.30	71	-0.93	0.62	-3.57
60	-0.87	0.65	-3.03	38	-0.77	0.62	-3.62	63	-0.73	0.63	-3.34	49	-0.92	0.63	-3.52

Table 3-4: 5 Best Models for each sequence in Closed conformation according to Dope-Z score. The best Q-mean score is highlighted in green. (PfDXR_closed is absent from this table as it already has a crystal structure 5JAZ).

PbDXR_closed				PcDXR_closed				PkDXR_closed							
Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean					
		Raw score	Z-score			Raw score	Z-score			Raw score	Z-score				
07	-1.80	0.72	-1.47	04	-1.72	0.72	-1.43	-1.82	-1.82	0.71	-1.65				
41	-1.82	0.71	-1.55	31	-1.72	0.71	-1.53	-1.84	-1.84	0.72	-1.47				
58	-1.79	0.71	-1.54	33	-1.72	0.72	-1.34	-1.82	-1.82	0.71	-1.62				
59	-1.81	0.72	-1.40	42	-1.74	0.73	-1.12	-1.82	-1.82	0.71	-1.51				
88	-1.79	0.71	-1.50	77	-1.73	0.72	-1.28	-1.82	-1.82	0.72	-1.41				
PmDXR_closed				PoDXR_closed				PvDXR_closed				PyDXR_closed			
Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean		Model	Dope-Z scores	Qmean	
		Raw score	Z-score			Raw score	Z-score			Raw score	Z-score			Raw score	Z-score
13	-1.86	0.72	-1.35	06	-1.75	0.71	-1.63	78	-1.72	0.72	-1.44	06	-1.79	0.70	-1.83
64	-1.84	0.72	-1.41	46	-1.74	0.69	-2.04	84	-1.74	0.71	-1.52	30	-1.77	0.70	-1.97
89	-1.84	0.72	-1.46	55	-1.75	0.70	-1.86	92	-1.76	0.72	-1.45	38	-1.77	0.70	-1.88
93	-1.85	0.73	-1.18	87	-1.71	0.70	-1.96	93	-1.72	0.72	-1.45	60	-1.77	0.71	-1.59
97	-1.86	0.72	-1.35	89	-1.72	0.70	-1.87	97	-1.73	0.72	-1.34	87	-1.76	0.70	-1.82

Table 3-5: Complete table of all final models' evaluation.

Template/ Models	PROCHECK				Dope Z- Score	Qmean6		DFire	PROQ3D		Evaluation
	Most favored	Allowed	Generously allowed	Disallowed		Raw score:	Z-score		ProQ2D	ProQ3D	
5JAZ_B	92.1%	7.3%	0.3%	0.3%	-2.47	0.75	-0.65	- 697.47	0.827	0.853	Approved
<i>PbDXR_closed</i>	93.4%	5.6%	1.1%	0.0%	-1.81	0.72	-1.40	- 628.18	0.812	0.840	Approved
<i>PcDXR_closed</i>	93.4%	5.8%	0.8%	0.0%	-1.74	0.73	-1.11	- 621.84	0.807	0.849	Approved
<i>PkDXR_closed</i>	93.4%	5.6%	1.1%	0.0%	-1.82	0.72	-1.41	- 628.18	0.790	0.828	Approved
<i>PmDXR_closed</i>	92.3%	7.1%	0.3%	0.3%	-1.85	0.72	-1.18	- 630.61	0.814	0.838	Approved
<i>PoDXR_closed</i>	93.1%	6.6%	0.3%	0.0%	-1.74	0.71	-1.63	- 628.03	0.809	0.838	Approved
<i>PvDXR_closed</i>	93.3%	5.6%	1.1%	0.0%	-1.73	0.72	-1.34	- 621.20	0.795	0.830	Approved
<i>PyDXR_closed</i>	92.6%	6.6%	0.8%	0.0%	-1.77	0.71	-1.59	- 624.03	0.811	0.840	Approved
1K5H_A	89.4%	9.8%	0.8%	0.0%	-1.33	0.77	-0.09	- 557.94	0.863	0.859	Approved
<i>PbDXR_open</i>	89.2%	7.4%	1.6%	1.8%	-0.95	0.63	-3.41	- 580.16	0.710	0.751	Approved
<i>PcDXR_open</i>	88.2%	9.5%	1.1%	1.3%	-0.87	0.65	-3.02	- 573.93	0.703	0.753	Approved
<i>PkDXR_open</i>	89.4%	9.3%	1.1%	0.3%	-0.88	0.65	-3.02	- 574.44	0.695	0.731	Approved
<i>PmDXR_open</i>	88.7%	8.4%	1.8%	1.0%	-0.88	0.66	-2.73	- 579.01	0.739	0.709	Approved
<i>PoDXR_open</i>	88.9%	9.2%	1.1%	0.8%	-0.80	0.64	-3.11	- 574.44	0.751	0.715	Approved
<i>PvDXR_open</i>	88.9%	9.5%	1.1%	0.5%	-0.76	0.64	-3.27	- 570.45	0.698	0.720	Approved
<i>PyDXR_open</i>	89.2%	7.7%	2.1%	1.1%	-0.96	0.63	-3.36	- 575.63	0.708	0.756	Approved
<i>PfDXR_open</i>	90.1%	7.6%	1.6%	0.8%	-0.99	0.64	-3.31	- 574.44	0.731	0.760	Approved

Combining the Dope-Z score with the Q-mean score approach allowed to merge the results coming from different assessment tools in the early stages of the assessment. An interesting tool was, MetaMQAPII is a Meta Model Quality Assessment Program, which integrate scores from eight other quality assessment programs VERIFY3D, PROSA2003, PROVE, ANOLEA, BALASNAPP, TUNE, REFINER, and PROQRES, using multivariate regression model (Pawlowski et al. 2008). Nonetheless, during the model assessment, some of the eight (8) servers were not responding which may impact the final quality of the assessment.

All produced models have Dope Z-Score lower than -0.5 (see Table 3-5) thus considered to be acceptable. Only open DXR models consistently showed a Dope Z-Score greater than -1.0, threshold for native state structure. Nonetheless, the models showed values closer to -1.0 (<-

0.90). In all cases, the closed conformation models presented better assessment scores, presenting a Dope Z-Score lower than -1.0. This was expected as models' quality decreases with sequence identity (Webb and Sali 2016). All the models presented a low local quality score (Qmean) in the flexible loop region. This observation was previously made in the two templates used for the modeling.

An alternative to improve those models' quality especially for models from the open conformation as all their Dope-Z scores were superior to -1 was to use SCWRL (Krivov, Shapovalov, and Dunbrack 2009). The side chain positioning problem is an NP-hard problem in modeling. MODELLER uses a heuristic algorithm for that problem which may find the optimal solution. SCWRL guarantees to find a global optimal solution for the problem and remains very fast for many proteins (Miyano et al. 2005). Using the command line (`Scwrl4 -i model_in.pdb -o model_out.pdb`), 100 models were submitted to SCWRL4.0. Comparing models showed that the best models remain the ones before submission to SCWRL4.0. A Welch Two Sample t-test between 100 models before treatment with SCWRL and after gave $t = -0.3199$, $df = 167.731$, $p\text{-value} = 0.7495$. The p-value was inferior to 0.05, so at 5% level of significance, the data provided enough evidence that the mean is equal in the two populations (see Figure 3-8). It is noteworthy that the models were compared with only the Dope-Z score.

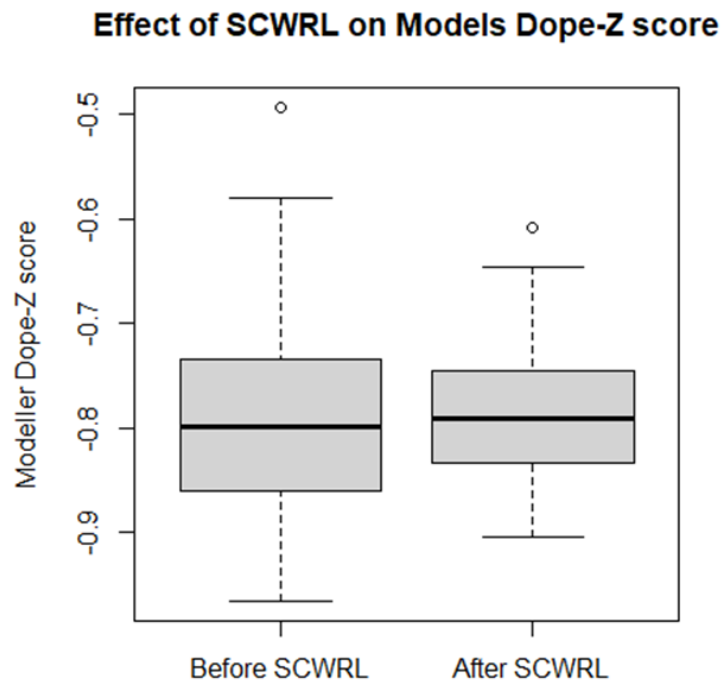


Figure 3-8: Effect of SCWRL(Krivov, Shapovalov, and Dunbrack 2009) on MODELLER (Šali et al. 2017) Dope-Z score.

All built models for the closed conformation presented at least 90% of their residues in the most favourable region of the Ramachandran plot. For open conformation six (6) out of the eight (8) models have 89% of residues in the most favourable regions, the two (2) remaining models having 90.3%. Nonetheless the template 1K5H_A itself has 89,6 % of its residues in the most favourable

regions. Interestingly all models showed better Dfire energy score than the template 1K5H_A. On contrast, it showed the best Q-mean and PROQ3D scores, even better than 5JAZ_B.

Assessment methods measuring the similarity between a model and the template are generally based on the comparison of the coordinates the c-alpha alpha atoms of a crystal structure and a model produced from. As MODELLER methodology is to copy coordinates of matching residues in the alignment from the structure to the model, thus, as result, we can expect good to very good scores from these measurements especially when the identity (template-sequence) is high. These scores should thus be used cautiously when judging a model quality. When two structures match perfectly, the resulting TM-score is 1. 0 is the lowest TM-score. Unrelated proteins show a score below 0.17 while greater than 0.5 assume generally the same fold in SCOP/CATH (Gadzała et al. 2017:201). Thus between 0.17 and 0.5 seems to be a twilight zone.

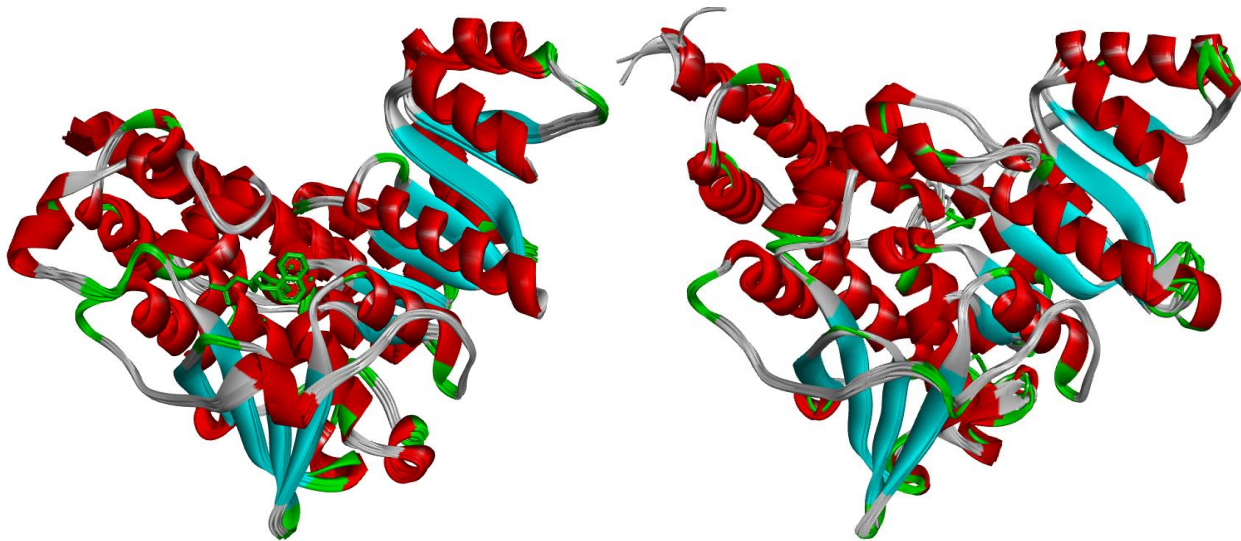


Figure 3-9: Final models superimposed in Discovery Studio 2016 (Biovia, San Diego, CA). Left: closed conformation (PbDXR, PcDXR, PoDXR, PmDXR, PkDXR, PyDXR, PvDXR + 5JAZ). Ligands in green in the active site. Right: open conformation (PbDXR, PcDXR, PoDXR, PmDXR, PkDXR, PyDXR, PvDXR and PfDXr).

3.5 Conclusion

A 3D structure of 1-deoxy-d-xylulose 5-phosphate reductoisomerase was modelled for the different *Plasmodium* sequences. These structures were modelled in two different conformations: from template 5JAZ (closed conformation) with active site ligand and metal maintained and from 1K5H which present an open loop conformation of the protein (see Figure 3-9). Structure validation showed the suitability of both templates for modeling with a notable low local quality around the flexible loop region in both templates. As expected, models derived from 5JAZ showed to be of better quality than the ones from 1K5H. This later had low resolution and low sequence identity with *Plasmodium* sequences.

CHAPTER 4: MOLECULAR DOCKING

4.1 Introduction

During the past decades, computer aided drug design has been successfully applied for the research of new drug molecules. Computational methods showed to be a fast and cost-effective and have contributed much to recent drug research. Luminespib also known as NVP-AUY922 is a drug candidate for the treatment of cancer discovered using CADD techniques. The molecule inhibits Hsp90, a protein implied in multiple oncogenic processes regulation. Starting from a library of 0.7 million compounds, virtual screening and lead optimization allowed to identify NVP-AUY922 (Sliwoski et al. 2014). Another success story was the discovery of agonists molecule for the M₁ acetylcholine receptor. These agonists have potential for treating dementia, including Alzheimer's disease. Using a homology model of the receptor, compounds were tested using computational methods. This helped in finding lead compounds which optimization lead to effective M₁ mAChR agonists with excellent pharmacokinetic properties (Budzik et al. 2010). In the case of malaria, our topic of interest, computational methods have been recently used to study atovaquone drug resistance in *P. falciparum*. The drug acts by binding to the parasite Cytochrome b protein. Using modeling, docking and molecular dynamics simulations, Akhoun et al. showed that a single point mutation in the active site of Cytochrome b protein results in atovaquone losing its binding affinity on that site and thus leading to resistance (Akhoun et al. 2014). These studies often use different approaches.

Methods in computer aided drug design can be divided into two main groups: ligand-based and structure-based methods. The ligand-based methods predict the activity of tested compounds by comparing them to active and inactive known ligands using structure-activity information. In the structure-based approach, structural information from both target and ligand are used. For the target-based approach, a solved 3D structure or a generated model where relevant is used to calculate interaction energies for the tested ligands (Sliwoski et al. 2014).

Molecular docking falls into this second category. It is used to study the interactions between a receptor/target and a ligand/small molecule. Two methods can be used: a shape based complementarity between the receptor and the ligand surface, and the calculation of interaction energies between the receptor and the ligand. Molecular docking tries to determine where the molecule can bind to the receptor and if so how strong is the intermolecular interaction. The number of hydrogen bonds and hydrophobic contacts are major contributors to the strength of the intermolecular interaction. Many potential ligands can be therefore tested on a target using a docking program, a method also known as virtual screening. The technique is a structure-based drug design technique in which large database of compounds can be tested against a receptor. These compounds do not require the labour of compounds chemical synthesis and/or laboratory testing. It thus offers a very time and cost saving method to test large database of compounds on a target (Sliwoski et al. 2014; Meng et al. 2011).

The main aim of this chapter is to perform *in silico* ligand docking studies of the SANCDB compounds database on the PfDXR using Autodock Vina (Trott and Olson 2010) to identify hits, ideally with new scaffolds. Secondly, docked compounds will be analysed for potential

bisubstrate hits, showing affinity for both active and cofactor binding sites. They will also be investigated for potential bidentate ligand chelating the metal ion. Multiple DXR conformations are used: closed, closed with flexible residues and open. This can help to gain insight into the protein dynamic and conformational space. Finally, compounds' drug likeness will be evaluated and will constitute an integral part of the hit selection process.

4.2 Docking strategies: AUTODOCK4 and AUTODOCK VINA

Molecular docking strategies have two main components: the search algorithm and the scoring function. As its name indicates, the search algorithm searches through the different possible poses of the ligand with respect to the protein. A pose is a binding mode, an orientation of the ligand within the target. The binding affinities of the different poses are evaluated by the scoring function. Based on the different scores, the poses are ranked and the most favourable binding mode is found (Huang 2014).

Ideally, all possible combinations of protein and ligand orientations should be considered by the search algorithm. In most protein-ligand biological interaction both remain flexible. A “hand-in-glove” analogy describe much better these systems than a “lock-and-key” one. As it is the case of DXR an induced-fit movement of the protein and the conformational changes of the ligand allows to find the best-fit (Mukhopadhyay 2014). Although this is the most effective solution, yet, this full exploration is impractical with current computational power and search algorithms. To find a compromise between time and efficiency, search strategies include genetic algorithms, systematic searches and molecular dynamics simulations. The search algorithm should also consider the flexibility of the entities in the system, especially for the receptor. Extensive sampling of the protein conformational space and all possible degrees of flexibility remain challenging. Three levels of search can be considered: rigid-body and flexible-ligand methods, and the flexible ligand–flexible protein methods. The scoring function is then applied on the different generated poses (Du et al. 2016).

Scoring functions predict the binding affinities based mainly on the strength of the non-covalent interactions. These include hydrogen bonds, ionic bonds, van der Waals interactions, hydrophobic bonds, salt bridge, metal and lipophilic interactions with the hydrogen bonds being the most contributing ones (Bissantz, Kuhn, and Stahl 2010; Lodish et al. 2000). Ideally, other interactions as the solvent effect and entropic effect also should be considered. Unfortunately, this is impractical as it increases the system complexity and thus will require large computational time (Du et al. 2016). Three main methods are used to calculate binding free energies: the empirical, the knowledge-based (or statistical potential) and the force-field-based methods.

In the force field approach, the binding affinity is estimated by the sum of the strength of the different interactions. The strength of these interactions is calculated using parameters from experiments and quantum mechanical calculations. Receptor and ligand intramolecular energies are also often considered. Finally, explicitly or implicit solvent models such as GBSA (generalized-Born surface area) and PBSA (Poisson–Boltzmann surface area) methods are used to account for the desolvation energy (Genheden and Ryde 2015). DOCK, GOLD and AutoDock are some docking tools using a force field scoring function (Meng et al. 2011).

The empirical scoring functions use ligand binding affinities from experimental structures to develop statistical regression models. These models are then used to estimate the binding affinities (Pason and Sotriffer 2016). LUDI, PLP and ChemScore are some docking programs making use of an empirical scoring function (Meng et al. 2011).

As the previous approach, knowledge-based scoring functions make use of ligand-protein complex crystal structures information. From these structures, interatomic contact frequencies and/or distances between the ligand and protein are obtained. The general assumption of the approach is that frequent close inter-atomic interactions are more likely to make favourable contributions to the binding affinity. Frequent contacts will thus have better scores. Some knowledge-based functions include DrugScore, SMOG, PMF, and Bleep (Du et al. 2016; Meng et al. 2011).

To sum up, a molecular docking process should specify three aspects: the search algorithm, the scoring function and the ligand/protein flexibility.

As search algorithm, AutoDock4 uses primarily a Lamarckian genetic algorithm. And as a scoring function, a combination of semi empirical free energy force field is used to estimate binding free energies. The two approaches are combined in a grid-based method to speed up evaluation of binding energies. The energy cost of placing a probe atom at each point in the grid is calculated. These values are stored and can then be used as a lookup table to avoid unnecessary recalculation and thus speed up the simulation. In its evaluation of free energy binding, ligand and protein intramolecular energies of both bound and unbound states are considered. AutoDock4 allows to have portion of the protein, for example, sidechains of the receptor, to be flexible. Other search methods as simulated annealing and traditional genetic algorithms are also available in the program (Morris et al. 2009).

As for Autodock Vina, it uses the iterated local search global optimizer based on stochastic global and local optimization procedures. Its scoring function is a hybrid empirical and knowledge based function inspired from X-Score (Wang, Lai, and Wang 2002). The tool has improved speed compared to Autodock4 by using multithreading. This allowed the tool to parallelly use computer's multiple processors. The tool also automates grid calculation. Finally, comparing accuracy, Autodock Vina also showed significantly better binding mode prediction (Trott and Olson 2010).

4.3 Methodology

4.3.1 Ligand preparation.

The compounds were retrieved in PDB format from the SANCDB website <https://sancdb.rubi.ru.ac.za/>. They were already minimized at RAM1 semi-empirical molecular orbital model using GAMESS (Schmidt et al. 1993). Some reference ligands were added to the set of SANCDB compound: DXP (the natural substrate for DXR), NADPH (the cofactor), fosmidomycin (from PDB structure 3AU9), FR98, LC5 a beta-substituted fosmidomycin analogue (from crystal structure 5JAZ). They will serve as reference for comparison and filtering of the docking results based on their binding poses and binding energies. The ligands were prepared using the Python script `prepare_ligand4.py`:

prepare_ligand4.py -l filename

The script assigns the correct AutoDock 4 atom types, the Gasteiger charges if necessary. The resulting file is saved in the .pdbqt file format. The script will also merge non-polar hydrogens as AutoDock and its AutoDock Vina use the 'United-Atom' model, and set up the 'torsion tree'.

4.3.2 Receptor preparation

The set of receptors was composed of PfDXR open (modelled from 1K5H), 5JAZ and its derived models. They were mainly prepared for docking using the Python script for docking “*prepare_receptor4.py*”. The different parameters of the script allow to automatically remove the water molecules (waters), in the case of the crystal structure, to assign atom types and Gasteiger charges. It also merges non-polar hydrogens (using thus the 'United-Atom' model as with the ligand). The option “deleteAltB” allowed to remove alternate coordinate present in 5JAZ chain B.

prepare_receptor4.py -A checkhydrogens -U nphs_lps_waters_deleteAltB -r + protein_pdb

Autodock Vina does not provide parameters for the Mn ion present in the protein active site. To calculate a more realistic charge on the metal ion, a QM (Quantum Mechanics) calculation using Gaussian09 (Frisch et al. 2009) at the B3LYP/6-31G(d) level of theory was done on the metal ion and its coordinating residues. The resulting charge was then assigned to the Mn atom in the resulting pqbqt file. The following residues of the active-site (ASP231, GLU233, GLU315, SER117, ILE89, SER88) were assigned as flexible for the crystal structure 5JAZ. Ideally, receptor should be treated as flexible. This describes protein ligand interactions such as induced-fit in the case of DXR more accurately allowing for a better fit of ligands in the protein active site (Umeda et al. 2011; Meng et al. 2011).

prepare_flexreceptor4.py -r 5JAZ_apoB.pdbqt -s ASP231_GLU233_GLU315_SER_117_ILE89_SER_88

The flexible residues were only added to the crystal structure (5JAZ). A rigid 5JAZ was also used. The setting of flexible residues poses an additional challenge, to merge the flexible and rigid portion of the protein after docking. A bash script flexrigidpdbqt2pdb (Moman 2011) was used for merging.

4.3.3 Molecular Docking

PyMOL Autodock/Vina Plugin was used to set the docking search area. Three docking experiences were set up: blind docking on 5JAZ, blind docking on PfDXR in open conformation (PfDXR_open) and targetted docking on 5JAZ and its derived models. The box was set to simultaneously cover both substrate and cofactor binding site in the targetted docking. Both DXOP and NADPH binding site will be indeed targetted. Targeting the cofactor binding site will be used in the investigation of potential bisubstrate inhibitors. Many SANCDB compounds are large with phenol and fused rings (Hatherley et al. 2015) and thus less likely to mimic fosmidomycin and fit into its small binding pocket. Masini et al. showed the druggability of both sites despite and the low lipophilicity of the NADPH binding site. Targeting both binding sites is interesting for the development of bisubstrate analogues (Masini, Kroezen, and Hirsch 2013). It noteworthy that targeting only the cofactor binding site would not be ideal. Cofactor analogues can cause cross-reactivity because of the widespread of nucleotide-binding pockets on numerous different proteins (Srinivasan et al. 2017). Our approach is thus to cover the two adjacent binding sites for searching for bisubstrate ligand (see Figure 4-1). The cofactor binding site is near the active site which open

and extends towards it. Such approach has been previously used in docking studies on MtDXR leading to interesting binding modes and a compound with an IC50 of 17.8 μ M (San Jose et al. 2013).

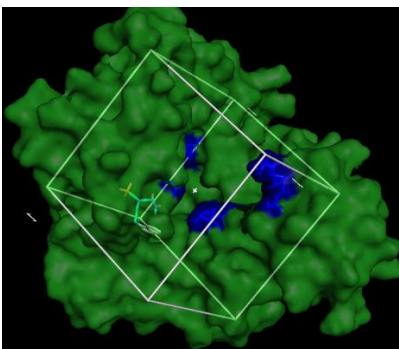


Figure 4-1: Setting of the grid box on both cofactor binding site (residues in bleu) and active site (showed with ligand in stick).

Although DXR is a well-studied protein, with the different binding pockets assessed (Masini, Kroezen, and Hirsch 2013; Deng et al. 2010) two blind docking experiments were set up: one on the open conformation and one on the closed conformation. These experiments will help confirm the binding sites of the protein but more importantly help filter the compounds. Compounds not binding in the binding sites in these experiments can thus be filtered out.

In the blind docking experiments, the grid was centred on the protein centre using PyMOL Autodock/Vina Plugin and then enlarge to cover the entire protein. With the resulting larger grid size, the exhaustiveness was proportionally increased. The high exhaustiveness allows to assure consistency of binding poses for selection of the best conformation of binding.

A Python script (see appendix D) was used to automatically generate the vina files required for high throughput screening. A job file was then generated and contained the command “*vina – config file.vina*” for all pairs of protein-ligand. The job file was then submitted on PBS (portable batch system) cluster on CHPC using a walltime=48:00:00 and the normal queue provided. The individual dockings were performed in parallel. Ten (10) poses were generated for each compound docking on a receptor and ranked per binding affinities.

Linux commands, Python and Perl scripts, PyMOL and Discovery Studio were used for analysis and visualization of the docking results. In analysing the docking results, two main criteria were considered: the binding pose and the binding energy. These two criteria were compared to the fosmidomycin binding pose and energy.

4.3.4 Docking validation

The quality of reproduction of a known binding pose is often used to validate a docking procedure. The docking process will be validated by redocking the ligand in the PDB (Berman et al. 2000) structure, the original crystallographic binding of LC5 and comparing the resulting RMSD on all atoms. The RMSD value between the two binding poses was computed in Discovery Studio. The other reference ligands DXP (the natural substrate for DXR), NADPH (the cofactor), fosmidomycin

(from PDB structure 3AU9), FR98 will secondarily be used. These ligands have well known binding poses from the literature.

The independent docking validation tool, X-score (Wang, Lai, and Wang 2002) was used to assess the predicted binding energies. Xscore is a command line tool that scores the binding using several independent methods. The tool uses an empirical scoring function to predict protein ligand binding affinity between the docked conformation of a ligand and the receptor. It can predict the binding free energies with a standard deviation of 2.2 kcal/mol (Wang, Lai, and Wang 2002). The best poses of the ligands from the output (pdbqt files) on the receptor 5JAZ_B were used for validation. The spearman correlation coefficient of binding energies between X-score predicted binding energies and the ones from Autodock Vina (Trott and Olson 2010) was calculated and the test of correlation Pearson's product-moment was conducted using the statistical tool RStudio Version 1.0.44 (Team 2014).

Each docking experiment was validated following the same procedure as above except for the plotting of the correlation graph which was conducted only for the targeted docking on DXR in closed conformation. All the docking experiments followed the same methodology. For purpose of validation, the predicted binding energies by X-score and Vina for LC5 in each experiment were compared (see Table 4-1).

Table 4-1: Summary of the different docking experiments

Parameters	Blind docking on DXR_closed	Targeted docking on DXR_closed	Blind docking on DXR_open
Proteins	5JAZ_B	5JAZ_B + 7 Models (closed)	1K5H_A + PfDXR_open
Grid	Full protein size_x=63.75 size_y=78.75 size_z=63.75	size_x=30, size_y=30, size_z=30	Full protein size_x = 63.75 size_y = 63.75 size_z = 63.75
Grid centre coordinates	center_x=-3.94 center_y=24.93 center_z=-18.38	Fosmidomycin+ NADPH binding sites center_x=-10, center_y=30, center_z=-19	center_x = 65.00 center_y = 81.42 center_z = 79.39
Ligands	SANCDDB + FOM + FR98 LC5 + NADPH	SANCDDB + FOM + FR98 LC5 + NADPH	SANCDDB + FOM + FR98 LC5+ NADPH
CPU	12	12	12
Exhaustiveness	384	192	408
Validation	LC51 from 5JAZ X-score	LC51 from 5JAZ X-score	Reference ligands X-score
Flexible Residues	None	ASP231, GLU233, GLU315, SER117, ILE89, SER88	None

4.3.5 Analysis and hit identification

The main ranking criterion for hit selection was binding energy on the rigid crystal structures (5JAZ rigid). A threshold of binding energy of -8 Kcal/mol across the 7 *Plasmodium* docked proteins was used. Ligand efficiency (LE) and ligand lipophilic efficiency (LLE) were also used to rank and filter the compounds.

The ligand efficiency was calculated by dividing the binding energy by the number of non-hydrogen atoms in the compound (As was shown in Equation below) (Hopkins, Groom, and Alex 2004). A Python script (see appendix E) was used to calculate the number of non-hydrogen atom in the compound from their mol2 format. A threshold of -0.25 Kcal/mol/Non Hydrogen-atom, considered to be the acceptable lower limit for ligand efficiency in screening (Hopkins et al. 2014) was used.

LLE is linked to compounds' permeability and showed good correlation between experimental and calculated values of binding free energy (García-Sosa, Hetényi, and Maran 2010).

$$LE = \frac{\Delta G}{\text{number of heavy atoms}}$$
$$LLE = \log\left(-\frac{\Delta G}{P}\right)$$

ΔG : free energy of binding

P: octanol-water partition coefficient (García-Sosa, Hetényi, and Maran 2010).

The use of these metrics has been encouraged in early stages of hits identification. It has been observed that compound molecular weight and lipophilicity increases during lead optimization. Thus, LE and LLE metrics help maintaining compound with reasonable molecular weight and lipophilicity to facilitate further optimization (Doak et al. 2014).

Across the different solved crystal structures of PfDXR, the binding pattern of the different inhibitors is well known. Literature showed the importance of some residues of the protein active site but also in the flap covering it in the protein inhibition (see Table 1-1). Discovery Studio provides scripting capabilities were used to extract all interactions (protein residue/atom, ligand atom and the type of interaction). The scripting tool was adapted (see appendix F) to parse all docked ligands and derive all interactions between the ligands and the protein. The resulting file could then be parsed using Python (see appendix G) to derive residues mostly implied residues in interaction with the ligands, the number of hydrogen bond for each ligand and its interacting residues. Finally, the interactions were compared with the main residues involved in DXR inhibition from literature (see Table 1-1). The tool was also helpful in the investigation of possible bisubstrate inhibitors. As residues involved in interaction with NADPH are well known (Table 1.1), ligand presenting interaction with both substrate binding site and NADPH binding site were also analysed. These ligands were visualized in Discovery Studio, analysing the pose of the compound, its fit in the protein active site especially in the pockets and its drug likeness properties. Bisubstrate hits constitute a separate cluster of hits.

The fitting of the ligands in the protein active site was evaluated by the distance (see equation below) to the active site. The measure was estimated between the centre (averaging the x, z, y coordinate below) of the ligands and the coordinates of the C2 carbon of LC5. This allowed a quick filtering of the ligand before visual inspection.

$$\frac{1}{n} \sum_{i=1}^n x_i, \frac{1}{n} \sum_{i=1}^n y_i, \frac{1}{n} \sum_{i=1}^n z_i$$

$$distance = \sqrt{(x_l - x_{lc5})^2 + (y_l - y_{lc5})^2 + (z_l - z_{lc5})^2}$$

x_l, y_l, z_l : coordinates of the ligands

$x_{lc5}, y_{lc5}, z_{lc5}$: coordinates of LC5

The selected compounds were then cross validated against the compounds in blind docking. Any compound found not in the active site in the blind docking was removed from the set.

Pharmacological properties have been a major drawback in DXR inhibitors development reason for us to undertake assessment of the SANCDB compounds in the early stages of virtual screening. The FAF-Drugs4 (Free ADME-Tox Filtering Tool) was used to evaluation the ADME-Tox properties (Adsorption, Distribution, Metabolism, Excretion and Toxicity) properties. In general, the use of ADMET properties for hit selection is also recommended (Zhu et al. 2013). The set of SANCDB compounds, FR98, fosmidomycin and LC5 were assessed. They were converted in mol2 format using Babel (O'Boyle et al. 2011) and assembled in sdf format for submission to the server. No FAF-Drugs4 pre-defined filters was used. The XLOGP3 method was used for the octanol/water partition coefficient (logP) computation. A detailed exploration of the pharmacological properties of the compounds goes beyond the scope of this project. An approach could for example, prioritize compounds with chemical properties ideal in the context of antimalarial. The simple QED (Quantitative Estimate of Druglikeness) score was used to filter compounds based on their drug likeness. The score integrates eight (8) relevant characteristics of chemical compounds related to their drug likeness: number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), molecular weight (MW), octanol-water partition coefficient (ALOGP), number of rotatable bonds (ROTB), the number of aromatic rings (AROM) molecular polar surface area (PSA), and number of structural alerts (ALERTS). The resulting scoring from 0-1 provides a more flexible way of filtering of scoring different from a binary black and white assessment of the compounds and allowing to tolerate more compounds which can later be optimized. A survey on the opinion of chemists' views of chemical attractiveness on 17117 compounds associated "unattractive compounds" considered "too complex" with a QED score of 0.34 (and a standard deviation of 0.24) (Bickerton et al. 2012). Therefore, a threshold of 0.4 was used for filtering.

As different conformations of the protein (rigid, open and rigid flexible residues) were used, a comparative analysis of the results especially in term of compounds' ranking was conducted. The nonparametric statistics test weighted Kendall's tau (Vigna 2014) was used to compare the ranking of compounds between the different configurations of the receptor. The tau ranges from 1 to -1, corresponding to observations having similar rank -1 when their rank is anti-correlated. 0 correspond to no correlation (Vigna 2014). The measure was computed using the Python package SciPy (Jones, Oliphant, and Peterson 2001). Other statistical test in analysing the docking results were done using Microsoft Excel 2016 and RStudio Version 1.0.44 (Team 2014).

All processes in the methodology except in the visual analysis of the compounds' poses have been automated and can easily be implemented in further studies. To conclude this section process followed the diagram below to finally select 5 hits (see Figure 4-2).

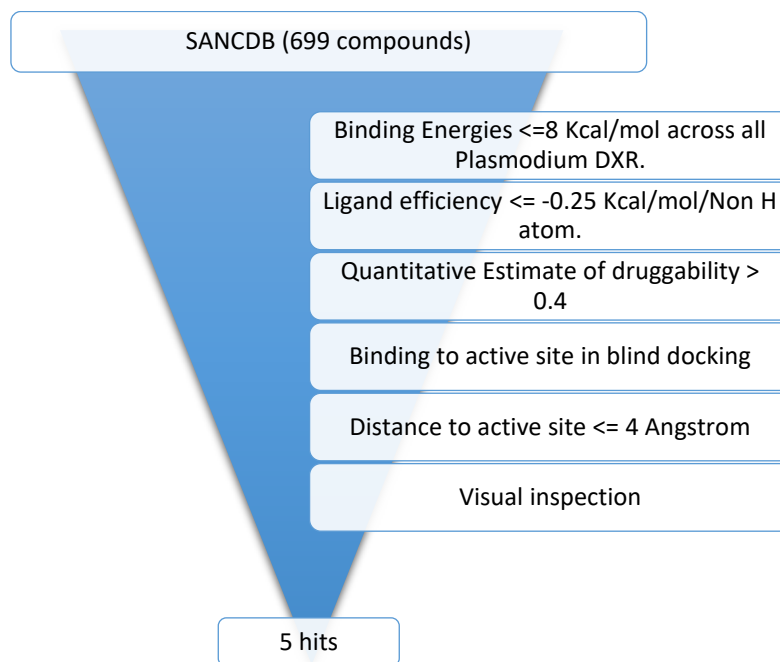


Figure 4-2: Hit selection process.

4.4 Results and Discussion

Analysing docking results has been recognized as one of the most difficult and subjective steps in virtual screening. Inaccuracies of the scoring functions may result in errors in ranking (Cosconati et al. 2010). In this study, different ranking methods were used: binding energies, ligand efficiency and ligand lipophilic efficiency.

4.4.1 Ligand preparation.

A total of 699 compounds was retrieved from the SANCDB database. Using minimized ligands will allow Autodock Vina to start with a reasonable conformation of the ligands (bonds angles, lengths and torsion angles). The compounds could be further optimized using a higher level of theory, B3LYP density functional with the 6-31G* basis set using Gaussian (Frisch et al. 2009). Although its noteworthy that the search algorithm of Vina utilizes flexible ligand docking. Thus, it generates and searches through different conformations of the ligands. Further optimization is hence not indispensable. The 699 SANCDB compounds in .pdbqt format and the five (5) ligands of reference were successfully generated. The total number of compounds was then of 704.

4.4.2 Receptor preparation.

The script *prepare_receptor4.py* produced the protein in .pdbqt file format for the docking. Water molecules were removed from the crystal structure. This latter does not contain the cofactor

NADPH. Keeping water molecules increases the number of interacting species and thus increases the complexity of calculation and the computational time.

Force-field parameters for metal binding simulation in docking remain a major problem (de Ruyck, Wouters, and Poulter 2011). When no Gasteiger parameter is available, Autodock tools assigns a charge of 0.00. Nonetheless, as previously reported, this charge doesn't describe realistically the metal and its surrounding negatively charged residues, nor will a charge of +2 (Bodill et al. 2011). The charge calculation set a charge of 1.4729 on the manganese metal, preferable to either extremes of charge given by the formal charge state +2 of the cation or the charge of 0 automatically given by Autodock Vina when no parameters available.

The docking results clearly shows the importance of the flexible residues. Ideally all residues in the active site should be set to be flexible. This will result in better reproducibility of the experiments and it models perfectly the flexible biological system and the concept of hand in glove (Mukesh and Rakesh 2011), especially important in the case of a DXR. In fact, the protein follows an induced fit movement in its active upon ligand binding. However, such set up would require excessive computational cost. This is especially important in the case of high throughput screening of ligands. A trade-off is thus to select some of the residues to be flexible. In our case three residues (ASP231, GLU233, GLU315) implied in metal binding near the NADPH binding site were selected with three other residues (SER117, ILE89, SER88) implied in NADPH near the substrate binding site were set to be flexible (see Figure 4-3).

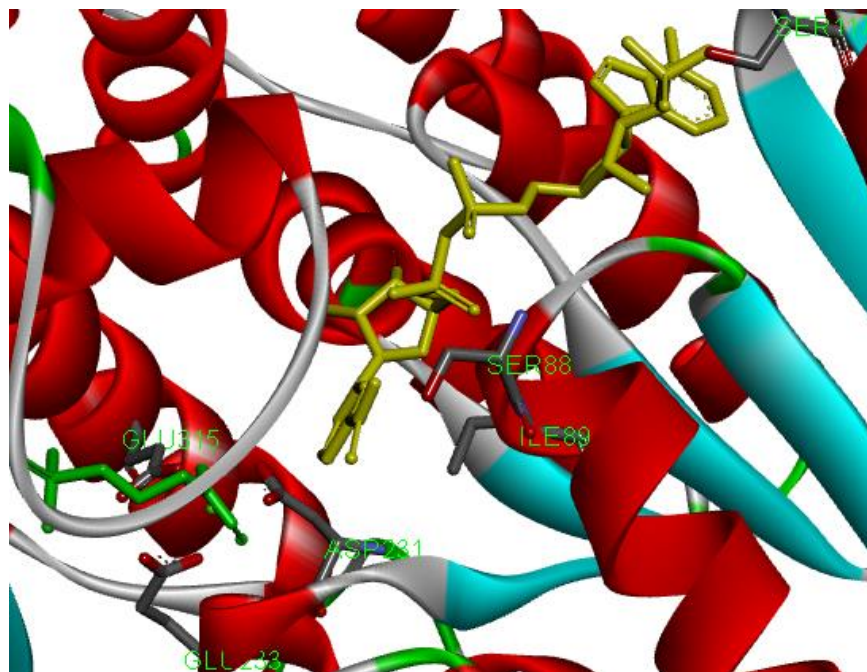


Figure 4-3: Bisubstrate inhibition approach. In yellow the cofactor, in green, fosmidomycin, residues are colored in atom types.

4.4.3 Docking validation

The RMSD value between the original crystallographic ligand (LC5) and the redocked ligand was 0.58 Å (see Figure 4-4). That value is less than 1 Å, indicating good reproduction of the correct pose. For the cofactor NADPH, it is important to note that it mostly bound in the substrate binding site in the 10 poses generated by Vina. Thus, the correct pose for the cofactor was not reproduced as illustrated in Figure 4-14.

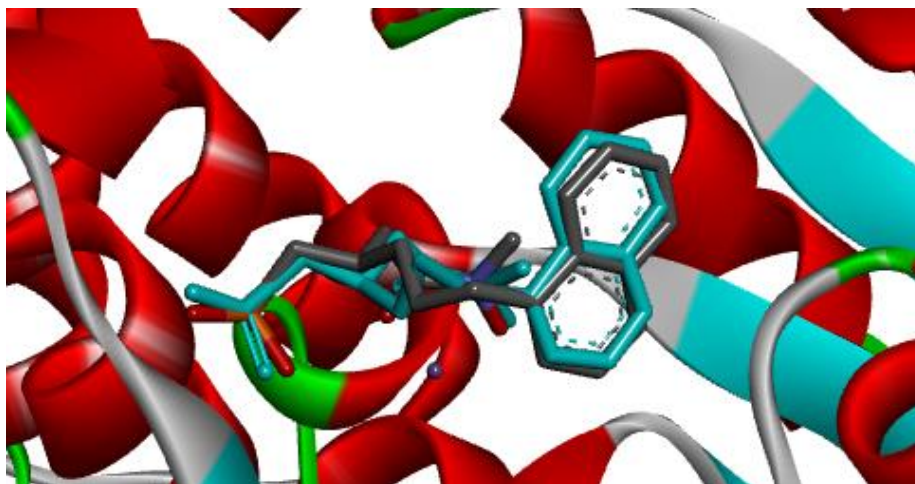


Figure 4-4: Molecular overlay Original LC5 (color by element) in the crystal structure and redocked LC5 (in light blue). RMSD= 0.58 Å.

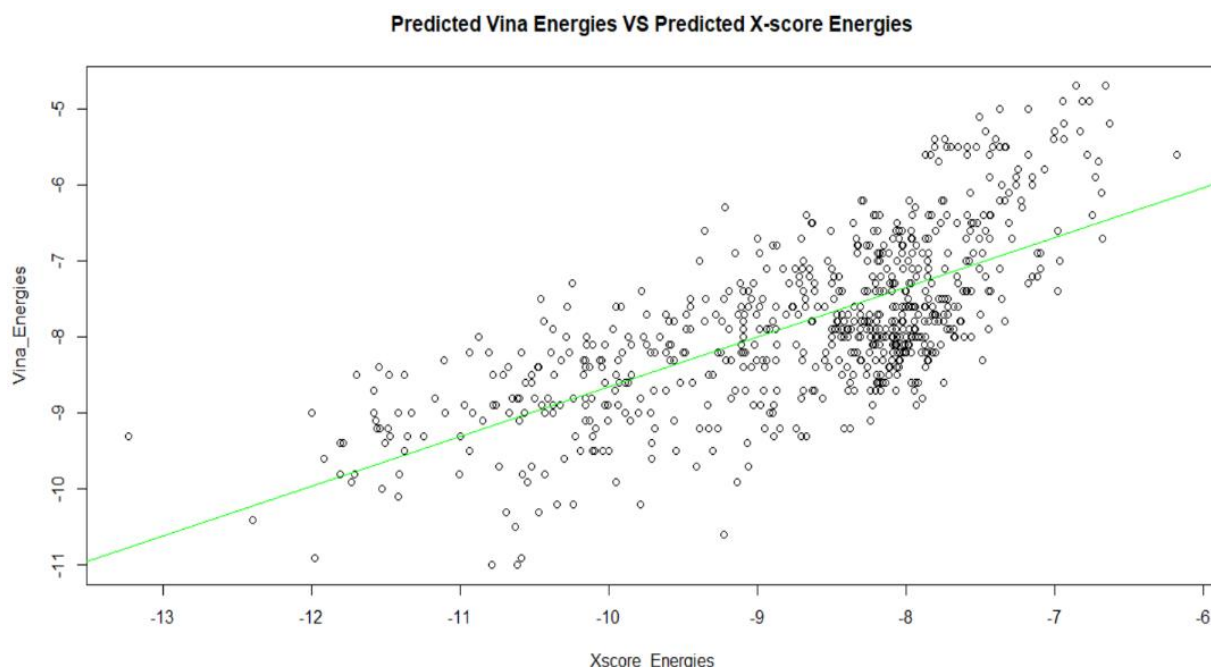


Figure 4-5: Predicted Binding Energies by X-score and Autodock Vina.

The spearman correlation of binding energies between the predicted binding energies by Autodock Vina (Trott and Olson 2010) and the ones by X-score (Wang, Lai, and Wang 2002:200) was of 0.70 (see Figure 4-5). The Pearson's product-moment correlation test was conducted at the 5% level of significance. The results indicated a p-value $< 2.2e-16$ with $t = 26.0419$, $df = 702$. So at the 5% level of significance, the data do provide sufficient evidence that the two predicted binding energies are correlated.

Table 4-2: X-score and Vina predicted binding energies for LC5 for docking validation.

Binding energy prediction	Blind docking on DXR_closed	Blind docking on DXR_open
Vina	-8.7 kcal/mol	-7.8 kcal/mol
X-score	-9.00 kcal/mol	-7.82 kcal/mol

For the two blind docking experiments, the predicted binding energies by Autodock Vina and X-score are very similar (see

Table 4-2).

4.4.4 Analysis

LC5, FR98, and fosmidomycin in decreasing order of potency in inhibiting DXR (Sooriyaarachchi et al. 2016; Wiesner et al. 2016:98). That order is also reflected in the predicted binding energies by Autodock Vina (Trott and Olson 2010). Indeed, these ligands bounded with the following

energies: FR98 6.44 Kcal/mol, fosmidomycin -5.89 Kcal/mol, DXP -6.86 Kcal/mol, LC5 -8.7 Kcal/mol.

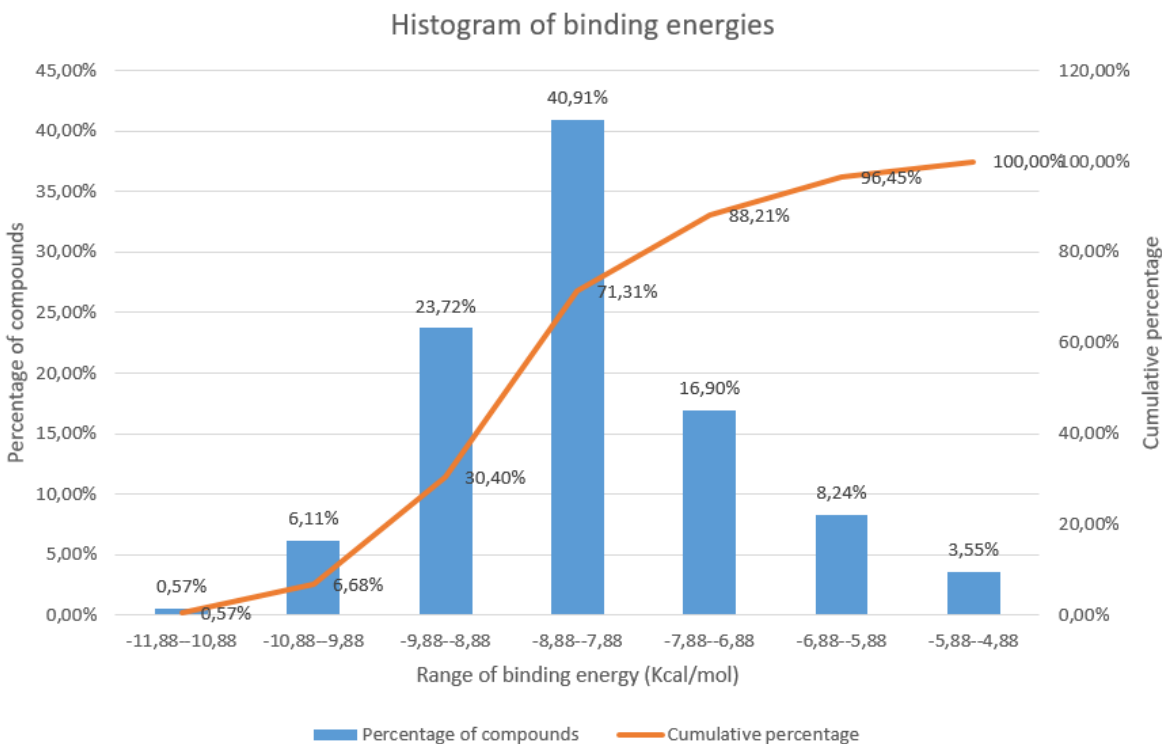
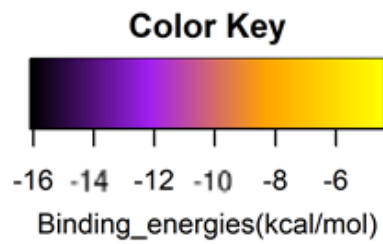


Figure 4-6: Histogram of binding energies

Interestingly, more than 90% of the compounds showed binding energy better than FR98 and fosmidomycin, the reference inhibitor with DXR (see Figure 4-6). This makes the filtering process challenging as these compounds' binding energies were planned to be used as a threshold for filtering. Thirty percent of the compounds had better binding energy than LC5 (the most potent among the reference ligands used and the original ligand in the crystal structure). An explanation for that observation is the presence of many heavy atoms in the SANCDB compounds. These compounds are large, and present many chemical reactive groups. As reported in a previous study, virtual screening has a strong bias toward large compounds, the larger the compound the higher the binding energy, a common problem in lead likeness of hits attributed to large compounds' molecular weight and logP values (Keseru and Makara 2006). A similar observation was made in this study. Plotting the binding energy against the number of non-hydrogen atoms gave a moderate correlation coefficient of 0.58. This observation can lead to false positives in compound ranking. Thus, for compound filtering setting a threshold based on the binding energy would have been biased. Using potency alone to rank compounds from a high throughput screening can lead to false positives. A study with Autodock and Autodock Vina indicated the same size related bias affecting compounds with <20 heavy atoms (Shityakov and Förster 2014). Moreover, Autodock and Autodock Vina both achieve a comparable standard error of ± 2 kcal/mol for Autodock and 2.85 kcal/mol for Vina in the prediction of free energies of binding. It was then suggested that for compound selection, the binding energy must not be the only measure (Trott and Olson 2010; Cosconati et al. 2010). Ligand efficiency provides a mean to normalize binding

energy among compounds of different molecular weight, countering thus the strong bias of virtual screening towards large molecules. The metric is recommended for hit identification (Zhu et al. 2013). Values of ligand efficiency less than -0.3 kcal/mol per heavy atom are considered good (Hopkins, Groom, and Alex 2004; Cosconati et al. 2010). Nonetheless, according to more recent study, the threshold -0.3 kcal/mol/atom should be increased. A value of -0.3 kcal/mol/atom for a compound of roughly 500 Daltons or 35 heavy atoms corresponds to a roughly 10 nM activity, impractical for initial hit identification. Different ligand efficiency values have been recommended depending on the compound size (Zhu et al. 2013). For simplicity, a ligand efficiency threshold of -0.25 kcal/mol/atom was used.

More, large molecules are more likely to present toxicophores or structural alerts, chemical structures known for having noxious properties. These compounds also are not ideal for drug development. Also, about the number of aromatic rings, current studies support that more than 3 aromatic rings are undesirable in drug design and that heteroaromatics perform better than carboaromatic (Ward and Beswick 2014). And the SANCDB database is known for presenting numerous compounds with many aromatic rings (Hatherley et al. 2015).



Protein-Ligand binding energies

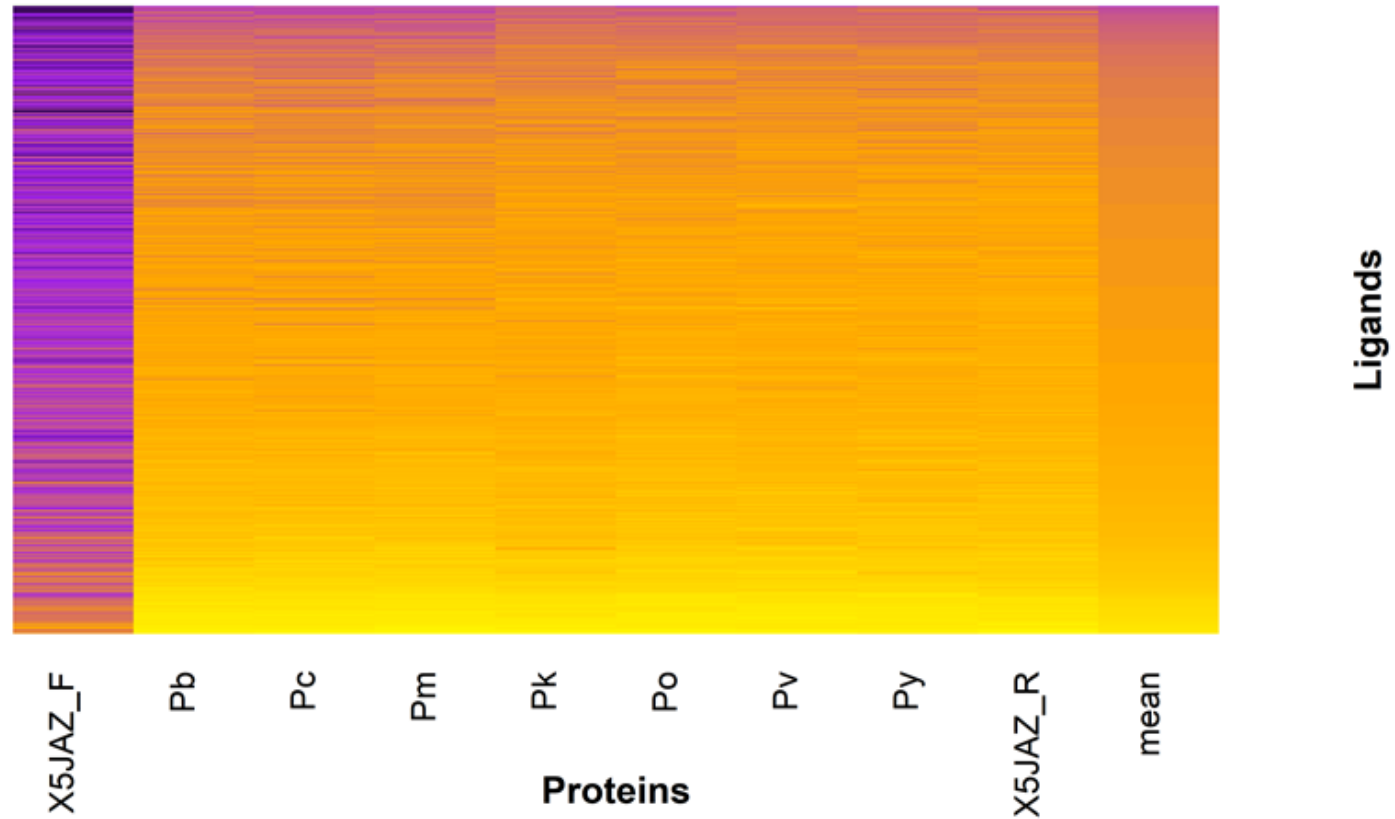


Figure 4-7: Heatmap of the binding energies for DXR in closed conformation (targetted docking of the protein active site).

The heatmap shows overall the binding energies remain did not differ across the different proteins (see Figure 4-7). This can be attributed to the highly conserved sequences of DXR in the genus *Plasmodium*. Also, as previously underlined in the sequence analysis chapter through the MSA, residues in the protein active site are highly conserved.

4.4.4.1 Docking with flexible residues

Generally, for docking with flexible residues, we observe better poses, and better fitting of the ligands in the protein active site. For example, flexible residues allowed fosmidomycin and FR98 to coordinate the metal ion contrary to rigid docking. DXP, fosmidomycin, and FR98 showed the correct orientation in the active site with the hydroxamate moiety oriented on the metal ion and the phosphonate moiety at the opposite site. This general observation of a better pose is also reflected in the binding energies. A t-test of mean (Welch Two Sample) between the binding energies in the two (2) configurations showed a significance difference. The mean was of -7.8 kcal/mole for the rigid receptor and of -11.54 kcal/mole with a p-value of 2.2e-16 ($t = 58.6939$, $df = 1370.898$).

It is also notable that a reverse binding mode was observed for fosmidomycin and FR98 as often reported in the literature from docking studies (Deng et al. 2010; Bodill et al. 2013). This was observed for fosmidomycin, FR98 and DXOP. Also, these compounds often showed a deviation of the hydroxamate moiety from its correct position in the active site.

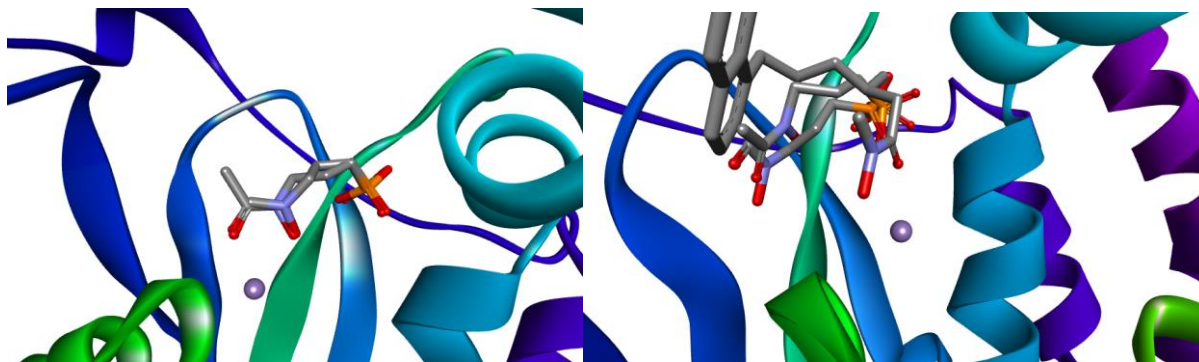


Figure 4-8: Best binding poses for fosmidomycin FR98 in the protein active site, metal ion in black.

In the rigid receptor, the two molecules turn away from the metal (see Figure 4-8). This orientation is also observed in all *Plasmodium* species. While the phosphonate moiety interacting residues is consistent with literature (see Table 1-1), the hydroxamate moiety interacts with HIS341 forming a hydrogen bond with the carbonyl group. LC5 can find the right conformation in both settings of flexible and rigid receptors. In all receptors (open, closed, closed with flexible residues), NADPH showed an unlikely binding pose, attempting to fit in the protein active site.

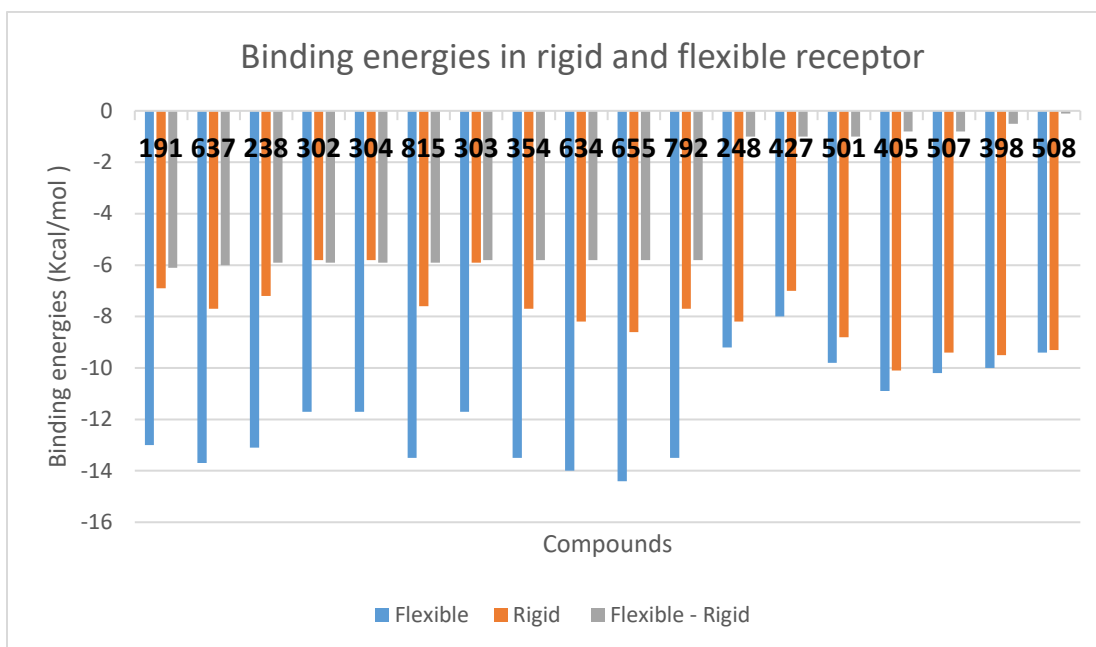


Figure 4-9: Binding energies in rigid and flexible receptors. The graph shows the highest (10) differences in energy and the lowest (7) differences in energy indicate. The compound identification number (SANC00XXX) at the base of the histogram.

Compounds showed very significant difference of their binding energies when comparing the rigid and flexible receptors (see Figure 4-9). The average difference was of -3.73 Kcal/mol. A change of 1.36 kcal/mol-1 of the binding energy results in 10-fold change in the equilibrium constant (Berg, Tymoczko, and Stryer 2002). As expected, none of the compound showed better binding energy with the rigid receptor than the flexible one. The ligands showed the highest binding energies when docked on the flexible receptor, ranging from -16.1 Kcal/mol (SANC00585) to -7.3 Kcal/mol (SANC00763). On the other hand, the rigid receptor 5JAZ, which had an energy range between -11 Kcal/mol (SANC00585) to -4.7 Kcal/mol (SANC00631).

Comparing the compound's ranking in the two configurations of receptor, the test result gave a weighted Kendall's tau rank-correlation (τ) of 0.61. The ranking in the two settings (flexible and rigid receptor) was correlated. The weighted Kendall's tau coefficient varies $-1 \leq \tau \leq 1$, with ideally $\tau = 1$ with two ranking are strongly correlated and -1 when they are anticorrelated.

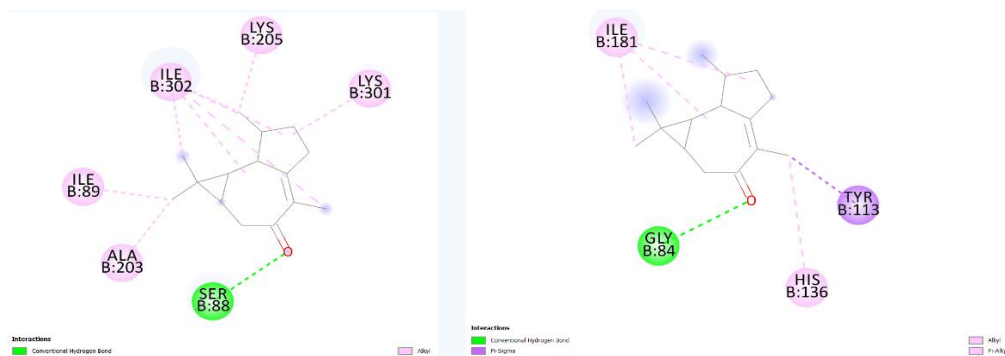


Figure 4-10: 2D plots of SANC00191 binding poses in flexible (left) and rigid (right) receptors.

SANC00191 showed a variation in binding energy of -6.1 Kcal/mol. The compound binds to different residues of the two different conformations of the protein (see Figure 4-10). Flexible residues ILE89, SER88 play a role in its binding on the flexible receptor. The compound is forming a hydrogen bond with SER88 and an alkyl type of interaction with ILE89. It was expected to see ligand binding with better energies to the flexible receptor, and greater interaction with the flexible residues. This was not the case overall. The higher binding energies in the flexible receptor are indirectly linked to the flexible residues (see Table 4-3). Compounds showing the highest differences in their binding energies do not necessarily interact with the flexible residues.

Table 4-3: Top compounds showing highest difference in binding energy and their interacting residues. In bold, residues set as flexible.

Compounds	Flexible Receptor	Rigid Receptor	Delta Binding Energy (Flexible-Rigid) Kcal/mol
SANC00191	LYS205, LYS301, ILE302, SER88 , ALA203, ILE89	ILE181, TYR113, HIS136, GLY84	-6.1
SANC00637	HIS341, CYS338, GLU233 , TRP296, PRO358, SER306	TRP296, CYS338, GLU233 , HIS341, SER306, PRO358	-6
SANC00238	SER270, GLY272, PRO273, LYS295	ASN115, GLY84, TYR113, ILE181	-5.9
SANC00302	LYS295, GLY272, PRO273	LYS295, GLY272, PRO273	-5.9
SANC00304	MET298, PRO273, LYS295, GLY272, PRO294, SER270	PRO273, LYS295, GLY272	-5.9
SANC00815	SER232, GLU233 , CYS268, HIS341, LYS312, TRP296, MET298	CYS338, GLU233 , MN502, SER232, HIS341	-5.9
SANC00303	TRP296, GLY272, THR303, ALA290, PRO294, MET298, PRO273, SER306, SER270, LYS295, LYS336	LYS295, GLY272, PRO273, TRP296, SER270	-5.8
SANC00354	ILE181, GLY84, ASN115, TYR113	ILE181, GLY84, ASN115, TYR113	-5.8
SANC00634	GLY272, PRO273, LYS295	GLY272, PRO273, LYS295	-5.8
SANC00655	GLY272, PRO273, LYS295, TRP296	PRO273, GLY272, LYS295, TRP296	-5.8
SANC00792	MET360, ASP182, ILE302, GLU206, LYS205, VAL230, ALA203, ILE89	HIS341, TRP296, CYS338, LYS297, SER270, GLU233	-5.8

SANC00302, SANC00303, SANC00304 are some of the compounds showing high difference in binding energy: -5.9 Kcal/mol, -5.8 Kcal/mol, -5.9 Kcal/mol respectively. The 2D plots showed very similar to identical interacting residues for these compounds. Interestingly, their binding did not differ in both flexible and rigid receptors, having the same interacting residues. The difference in binding, thus is not related to nor the binding pose, neither the binding residues. An explanation could be the possible more energetically favourable rearrangement of the flexible residues in the protein.

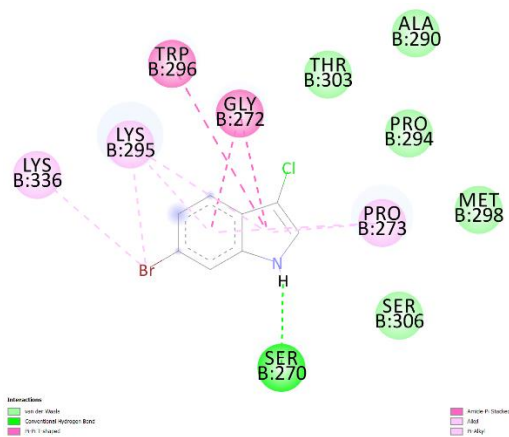


Figure 4-11: 2D plot SANC00303 docked in the rigid crystal structure 5JAZ.

These ligands have very similar chemical structures (see Figure 4-12) and are all extracted from *Distaplia skoogi* (Other name: Sea squirt). They are alkaloids known for the anticancer activity (Bromley et al. 2013).

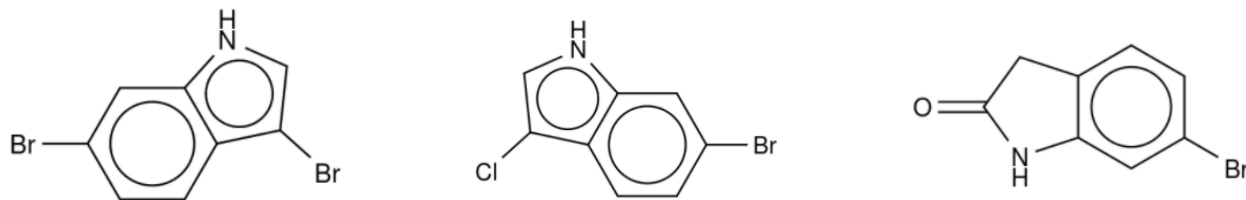


Figure 4-12: Left to Right: SANC00302 (3,6-Dibromoindole), SANC00303 (6-Bromo-3-chloroindole), SANC00304 (6-Bromo-2-oxindole).

Compounds with no significant change in the binding energy in both configurations, flexible and rigid receptor, were found to be large compounds with high molecular weights in general: SANC00248 (408.57 Da), SANC00427 (606.83 Da), SANC00501 (1095.23 Da), SANC00405 (953.12 Da), SANC00507 (1389.48 Da), SANC00398 (736.89 Da), SANC00508 (1225.32 Da).

For the selection of flexible residues, an interesting approach would have been to set residues in the flexible loop as flexible. This can allow to try to simulate the movement of the loop, to account for the open and closed conformations of the protein. However, the loop region is composed of around 10 residues. Setting such number of residues to be flexible would increase the computation time drastically. Another approach was to generate a set of representative conformers through molecular dynamic simulations and finally use these representatives for

docking simulations. Nonetheless, using the open conformation in this study allowed to gain insight into the possible conformational ensemble.

4.4.4.2 Docking on open conformation

It was of interest to study the screening on the open conformation. The analysis here focusses in the comparison of the compound ranking according to their binding energies, but also and more their poses on the open conformation. Various proteins are known for adopting a closed conformation upon inhibitor binding (Sandak, Wolfson, and Nussinov 1998). More than the movement of the flap covering the active site to form the closed conformation, DXR active site undergoes an induced fit movement to accommodate the substrate/inhibitor in the active site. More the open conformation may accommodate larger ligands (Mac Sweeney et al. 2005) in the protein active site.

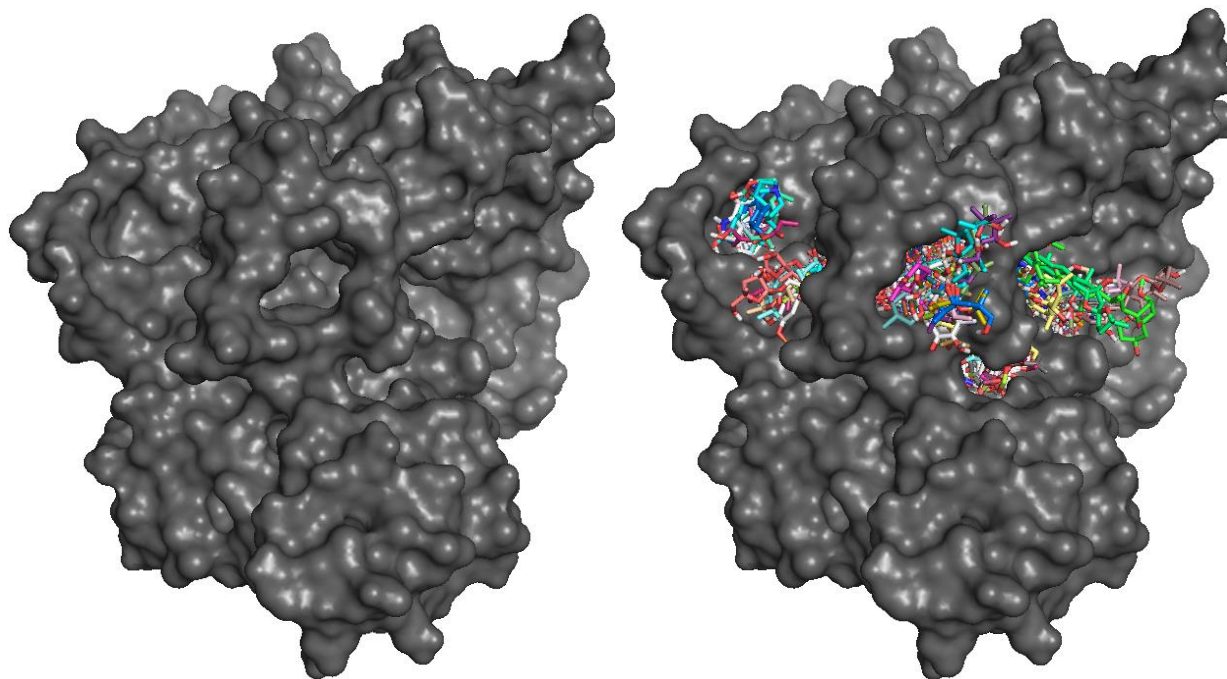


Figure 4-13: Left: Open conformation with three openings of the large active/cofactor binding site. Right: Clustering of compounds in blind docking of DXR open conformation.

In blind docking on the DXR open conformation, the majority of the compounds bind to the large cavity formed by the active site, the hole formed by the loop region in the middle and the cofactor binding site (see Figure 4-13). This cavity forms a canal crossing the NADPH binding site and opening on the protein active site on the opposite side. Few compounds bind to the opposite site. Interestingly, the cofactor bound in the substrate binding site in most of the 10 poses generated by Vina. The cofactor molecule spans across the bottom of the active site to the flexible region at the top, adopting an unlikely conformation (see Figure 4-14). Many ligands also bind to the semi-circle formed by the flexible loop. These ligands thus occupy a superficial region around the loop and do not reach the bottom of the large open cavity.

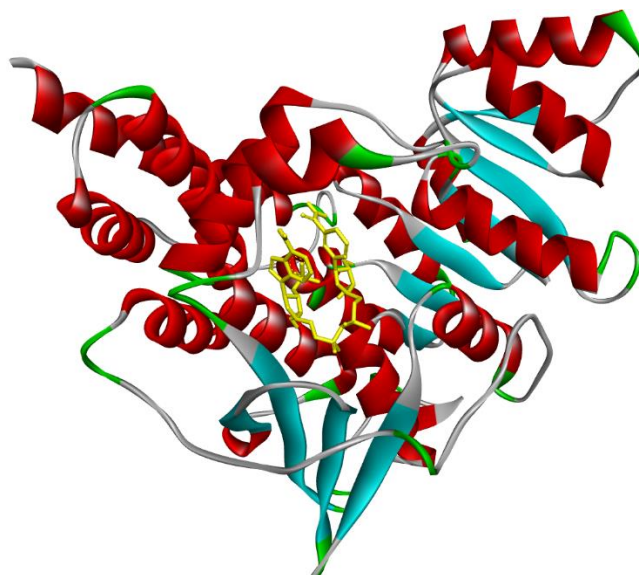


Figure 4-14: Cofactor in PfDXR active site

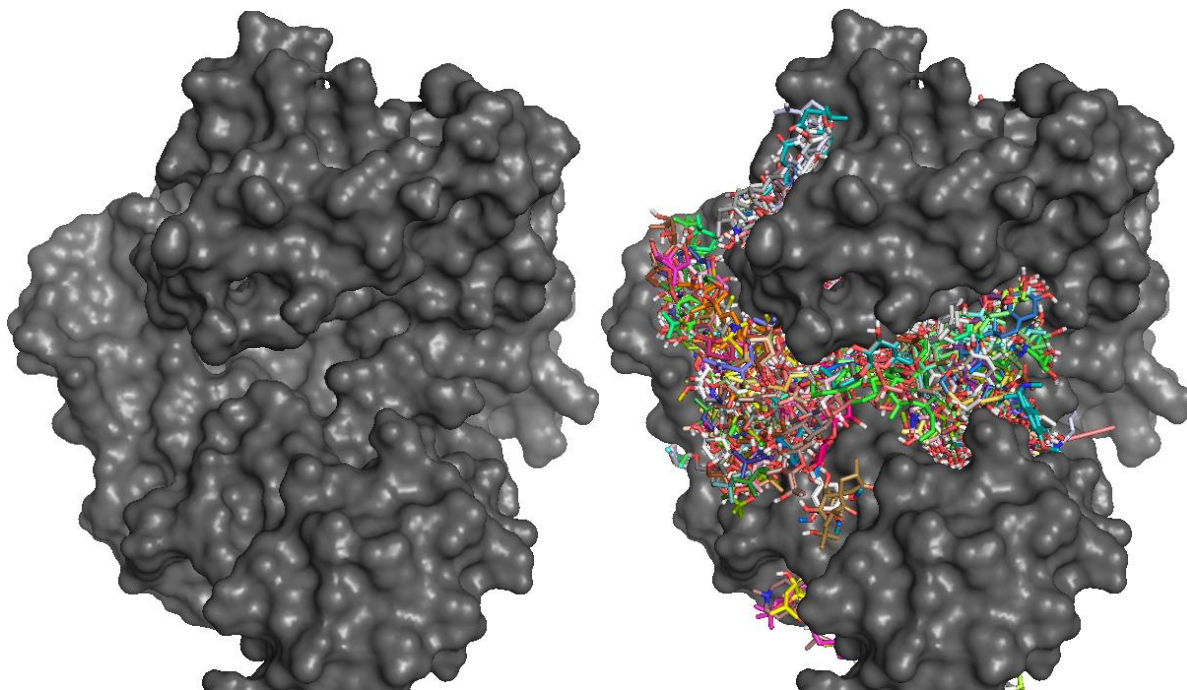


Figure 4-15: Clustering of compounds in the closed conformation. Left: before docking. Right: after docking.

As in the closed conformation, compounds clustered mainly in the protein active site (see Figure 4-15), cofactor binding site and the surrounding region. Though more compounds were found on the opposite face of the protein active site (see Figure 4-16). Compared to the open conformation, more ligands are scattered on the protein surface, some clustering on the opposite

face to the active site. Some large ligands may be unable to bind in the smaller pocket of the protein in closed conformation.

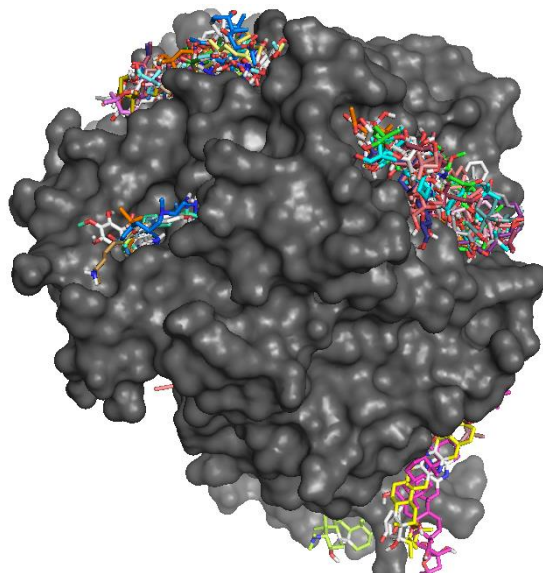
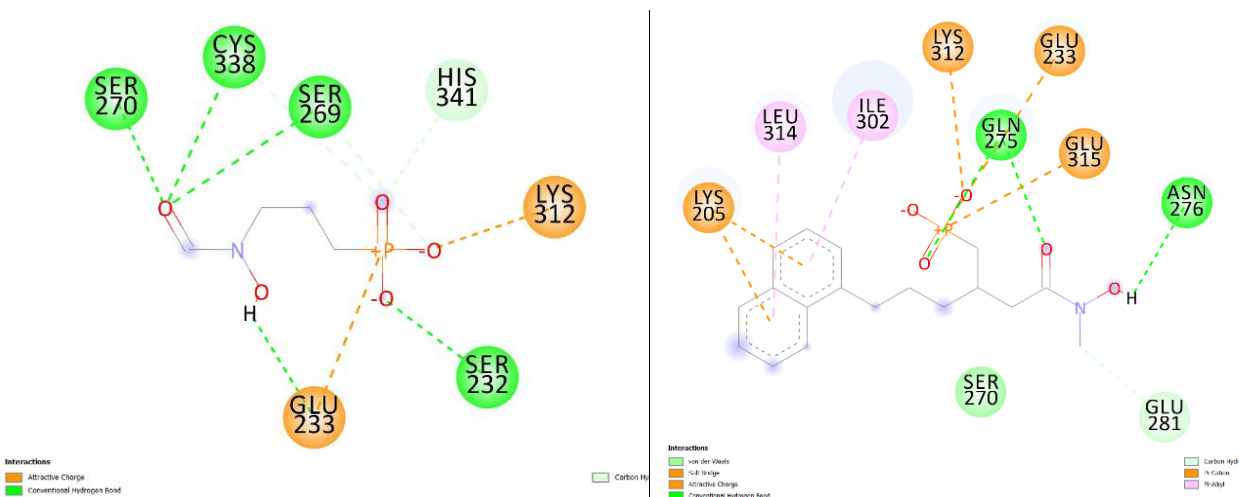
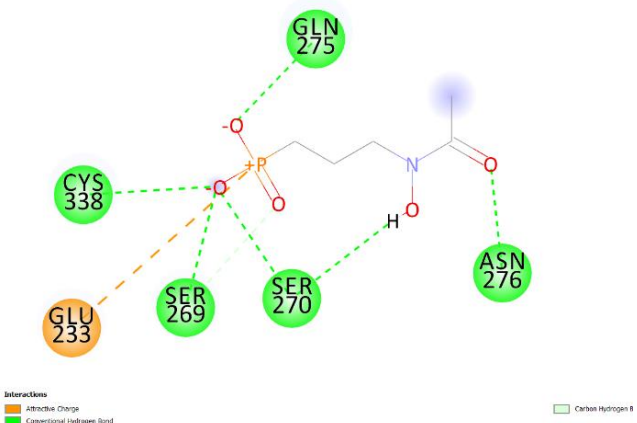


Figure 4-16: Compounds binding on DXR active site opposite face.

All reference ligands and most of the compounds bind in the protein active site. Residues interacting with reference ligands often imply known residues implied in DXR inhibition: SER270, SER269, ASN311, GLU315, LYS312 (see Figure 4-17 and Table 1-1). CYS338 is also commonly found to interact with the reference ligands. DXP forms a strong network of hydrogen bond around the phosphate group which is also often involved in charged interaction in which negatively charged residues such as GLU233 and LYS312 are involved. Oxygens on the phosphate are frequently involved in the hydrogen bonding interactions. LC5 does not show any interaction with any residues in the flexible loop region. Its aromatic rings are associated with LYS205, ILE302 and LEU314. We can also note the absence of interaction with any residue in the flexible loop region (residues 290 to 299) for the reference ligands.

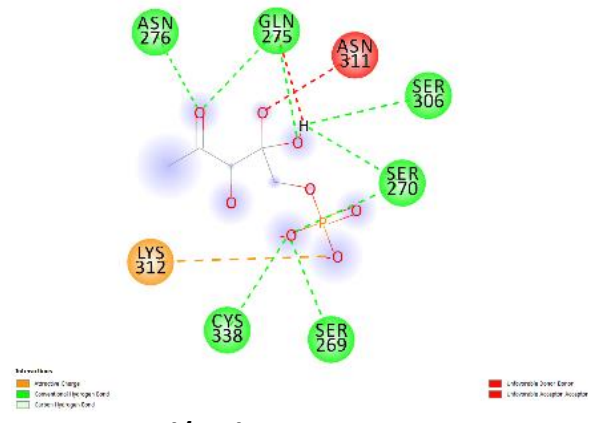


Fosmidomycin: -5 Kcal/mol



FR98: -5.6 Kcal/mol

LC5: -7.8 Kcal/mol



DXP: -6 Kcal/mol

Figure 4-17: Reference ligands in blind docking on open conformation: 2D poses and binding energies

Table 4-4: Interacting residues with inhibitors in closed and open conformation. Red: residues interaction with fosmidomycin phosphonate. Green: residues interaction with fosmidomycin hydroxamate

ID	Molecular Interactions (Residue and Type of interaction)		ID	Molecular Interactions (Residue and Type of interaction)	
FR98 OPEN	GLU233 Charge attraction	SER270 Hydrogen bond	FR98 CLOSED	LYS312 Charge attraction	SER270 Hydrogen bond
	SER269 Hydrogen bond	GLN275 Hydrogen bond		LYS312 Charge attraction	SER232 Unfavorable acceptor acceptor
FOS OPEN	SER270 Hydrogen bond	ASN276 Hydrogen bond	FOS CLOSED	GLU233 Charge attraction	HIS341 Hydrogen bond
	SER270 Hydrogen bond	CYS388 Hydrogen bond		SER270 Hydrogen bond	HIS341 Carbon Hbond
LC5 OPEN	GLU233 Charge attraction	SER270 Hydrogen bond	LC5 CLOSED	LYS312 Charge attraction	SER269 Hydrogen bond
	SER269 Hydrogen bond	CYS388 Hydrogen bond		LYS312 Charge attraction	GLY271 Hydrogen bond
FR98 OPEN	SER269 Carbon Hbond	ASN276 Hydrogen bond	FOS CLOSED	GLU233 Charge attraction	ASN311 Hydrogen bond
	SER270 Carbon Hbond	CYS388 Hydrogen bond		SER270 Hydrogen bond	ASN311 Hydrogen bond
FOS OPEN	HIS341 Carbon Hbond	CYS388 Carbon Hbond	FOS CLOSED	SER270 Hydrogen bond	MN502 Metal acceptor
				SER270 Hydrogen bond	CYS388 Pi-sulfur
LC5 OPEN	LYS312 Salt bridge	ILE302 Pi-alkyl	LC5 CLOSED	SER270 Hydrogen bond	PRO358 Pi-alkyl
	GLU315 Charge attraction	LYS205 Pi-alkyl		SER232 Hydrogen bond	PRO358 Pi-alkyl
LC5 OPEN	GLU233 Charge attraction	LEU315 Pi-alkyl			
	GLN275 Hydrogen bond	ASN276 Hydrogen bond			
LC5 OPEN	GLN275 Hydrogen bond	GLU281 Carbon Hbond			
	GLN275 Hydrogen bond				

ASN311 in the closed conformation receptor interacts with all inhibitors, forming hydrogen bonds, while this interaction is absent with all ligands in the open conformation. GLU233 through its negative charge is a key player in all inhibitors' interactions in both open and closed conformations. In contrast, ASP231 is absent from all interactions while often reported as binding to the inhibitors hydroxamate group. GLN275 and ASN276 are forming hydrogen bonds with the inhibitors

in the open conformation. As previously mentioned the deviation from the metal in the closed rigid conformation, FR98 and fosmidomycin interact with HIS341 through hydrogen binding.

In terms of binding energy (see Figure 4-18), a paired t-test was conducted, comparing each ligand binding energy in both configurations of the protein. The test results were p-value = 0.01762, t = 2.3791, df = 702. The p-value was inferior to 0.05. So, at 5% level of significance, the data provided enough evidence that the compound binding energies are different in the two configurations. The difference of mean between the closed and open conformation (binding energy in closed conformation – binding energy in open conformation) was of 0.0596 Kcal/mol. The binding energy is significantly better in the open conformation than in the closed one. Some compound showed significantly increased binding energies (see Figure 4-19).

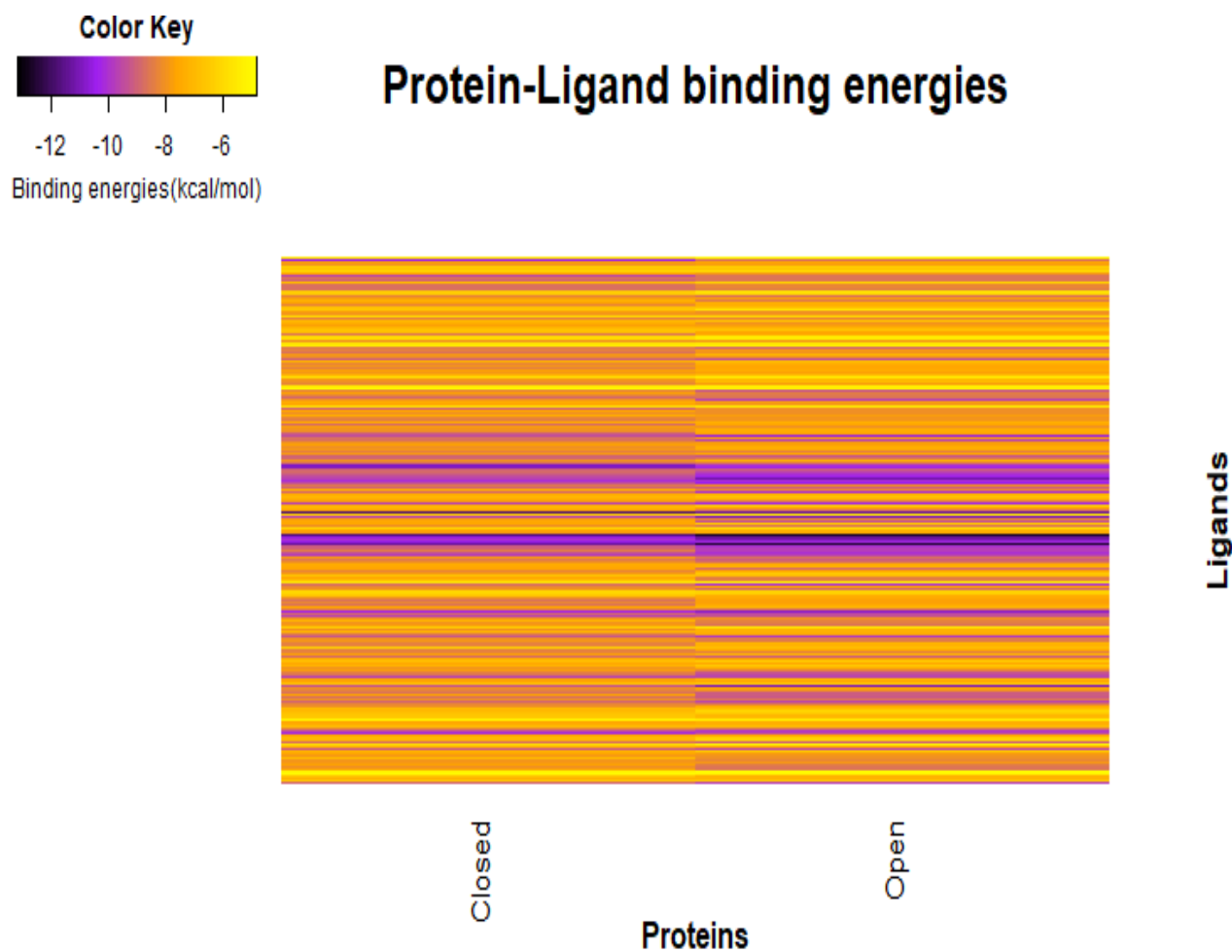


Figure 4-18: Binding energies in DXR closed and open conformation in blind docking.

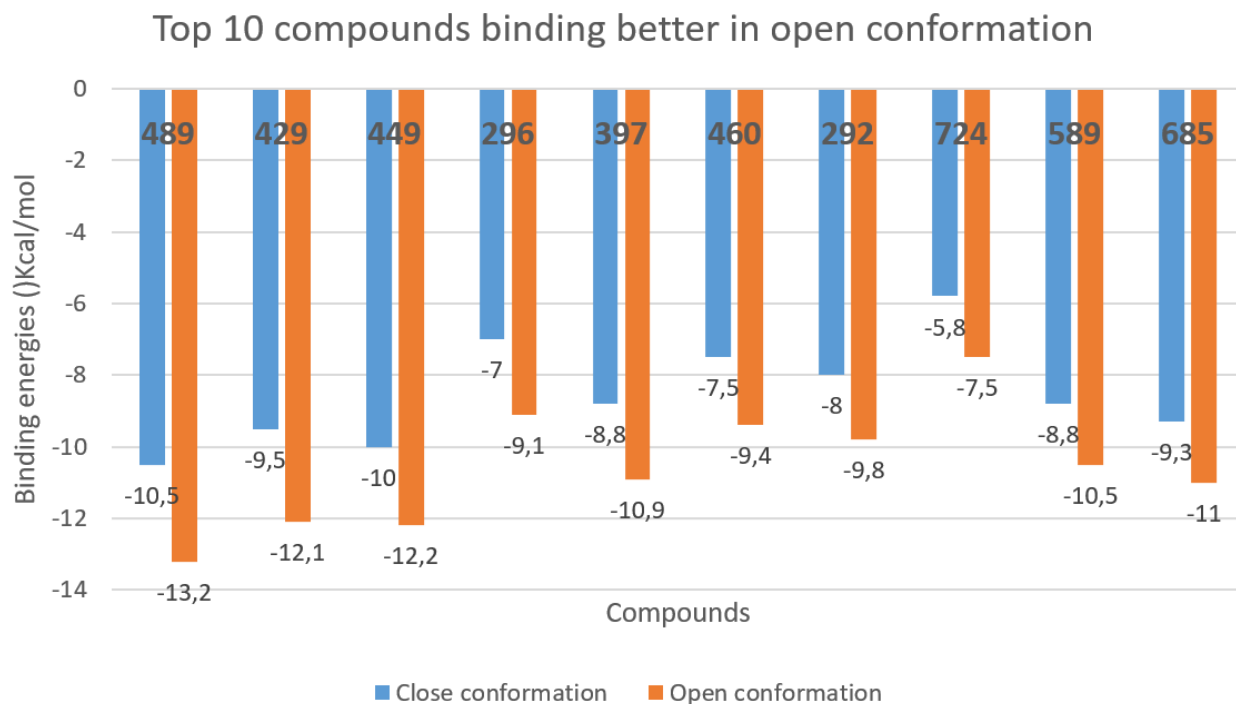


Figure 4-19: Top 10 compounds binding better in open conformation.

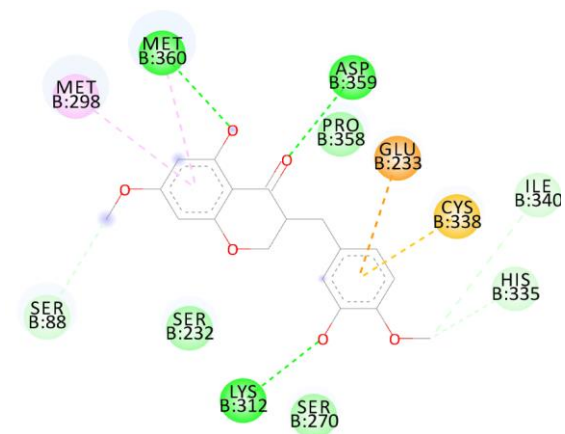
Comparing compound rankings in the two configurations, the weighted Kendall's tau rank-correlation (Kendall 1938) was used. The test result gave a (τ) of 0.85. Hence, the ranking in the two settings (open and closed conformation) was strongly correlated. This correlation coefficient was higher than the one with flexible residues (0.61).

To identify hits in the open conformation, compounds were first ranked by binding energy. The first 32 compound had a QED score under 0.376. These were large compounds with the number of carbon atom ranging from 38 to 81 and molecular weight between 1033.11 Da and 510.58 Da. The best ligand efficiency score for these compounds was -0.289 kcal/mol/heavy-atom. The smallest of these compounds doesn't fit the molecular weight component of the Lipinski rules. Investigating the correlation between the binding energies and the number of non-hydrogen atoms showed a high correlation coefficient of 0.773 and a R^2 value of 0.5989. These values illustrate bias toward large compounds observed in a similar study with Autdock Vina (Shityakov and Förster 2014). More these compounds were less likely to fit in the protein active site due to their size.

4.4.4.3 Bisubstrate inhibitors

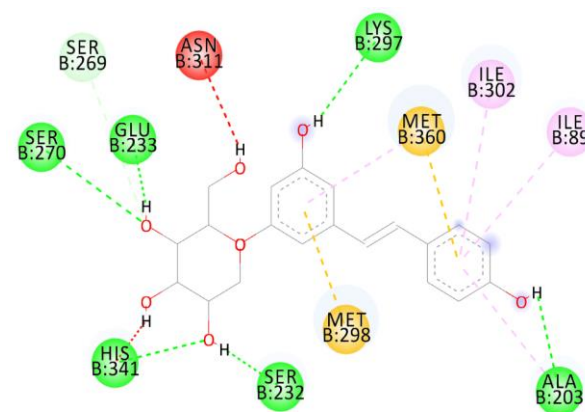
The setting of flexible residues (ASP231, GLU233, GLU315, SER117, ILE89, SER88) was planned to identify bisubstrate inhibitors. None of the docked compounds bind simultaneously to residues interacting with phosphonate moiety and the hydroxamate moiety of fosmidomycin. The identified hits had good binding energies ranging from -7 Kcal/mol to -9.7 Kcal/mol on the rigid receptor (5JAZ). Visually inspecting their binding poses revealed large compounds. This could be

expected as compound fitting such criteria would extend across the cofactor binding site and the protein active site. Among these compounds, SANC00615 showed a good druggability score: 0.896. Moreover, the compound displayed an interesting binding pose ranging across the cofactor binding site and the protein binding site, interacting with GLU233, SER270, SER88, and LYS312 (see Figure 4-20).



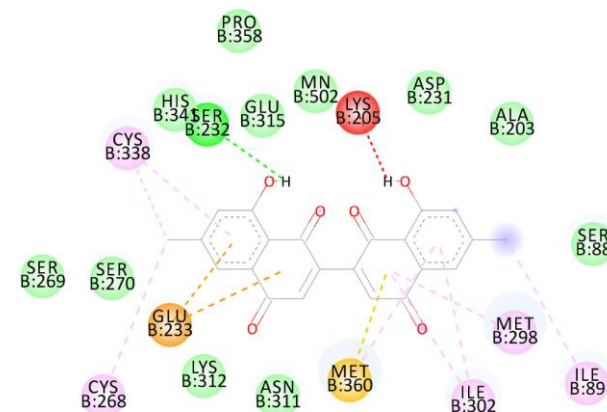
Interaktionen
 von der Wirtszelle
 Conventional Hydrogen Bond
 Carbon Hydrogen Bond
 Pi-Anion
 Pi-Sulfur
 Pi-Allyl

SANC00615
 Binding energy: -8.2 Kcal/mol (Rigid receptor). -11.4 Kcal/mol (Flexible receptor). QED: 0.896.



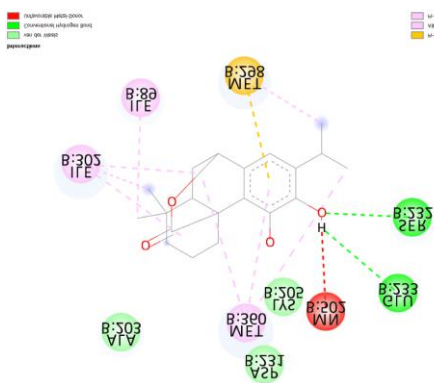
Interaktionen
 Conventional Hydrogen Bond
 Carbon Hydrogen Bond
 Unfavorable Donor-Donor
 Pi-Sulfur
 Pi-Allyl

SANC00562
 Binding energy: -8.4 Kcal/mol (Rigid receptor). -11.2 Kcal/mol (Flexible receptor). QED: 0.416.

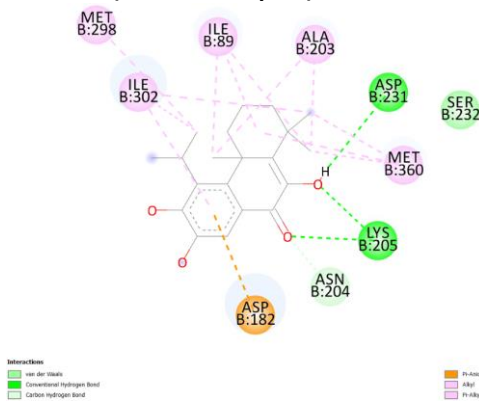


Interaktionen
 von der Wirtszelle
 Conventional Hydrogen Bond
 Unfavorable Donor-Donor
 Pi-Anion
 Pi-Sulfur
 Pi-Allyl

SANC00436
 Binding energy: -9.7 Kcal/mol (Rigid receptor). -13.8 Kcal/mol (Flexible receptor). QED: 0.787.



SANC00443
 Binding Energy: -8.2 Kcal/mol (Rigid receptor). -11.3 Kcal/mol (Flexible receptor). QED: 0.592.



Interaktionen
 von der Wirtszelle
 Conventional Hydrogen Bond
 Carbon Hydrogen Bond
 Pi-Anion
 Pi-Allyl

SANC00556
 Binding energy: -7.6 Kcal/mol (Rigid receptor). -11.5 Kcal/mol (Flexible receptor). QED: 0.639.

Figure 4-20: Compounds identified as potential bisubstrate inhibitor.

Table 4-5: Identified bisubstrates molecular interactions.

ID	Molecular Interactions (Residue and type of interaction)		ID	Molecular Interactions (Residue and type of interaction)	
SANC00562	MET298 pi-Sulfur	ASN311 Unfavorable donor-donor	SANC00615	MET360 pi-Alkyl	ILE340 carbon H Bond
	ILE302 pi-Alkyl	MET360 pi-Sulfur		SER88 carbon H Bond	MET298 pi-Alkyl
	HIS341 Unfavorable donor-donor	GLU233 Hydrogen bond		MET360 Hydrogen bond	CYS338 pi-Sulfur
ALA203 pi-Alkyl	HIS341 Hydrogen bond	LYS297 Hydrogen bond	HIS335 carbon H Bond	GLU233 pi-Anion	
SER232 Hydrogen bond	LYS297 Hydrogen bond	MET360 pi-Alkyl	ASP359 Hydrogen bond	LYS312 Hydrogen bond	
SER270 Hydrogen bond	MET360 pi-Alkyl	ALA203 Hydrogen bond			
SER269 carbon H Bond	ILE89 pi-Alkyl				
SANC00436	MET360 pi-Alkyl	MET298 pi-Alkyl	SANC00443	MET298 pi-Sulfur	MET360 alkyl
	CYS268 alkyl	CYS338 pi-Alkyl		ILE302 alkyl	MN502 Unfavorable metal donor
	GLU233 pi-Anion	LYS205 Unfavorable donor-donor		ILE302 alkyl	MET360 pi-Alkyl
	SER232 Hydrogen bond	ILE89 alkyl		ILE302 alkyl	MET298 alkyl
	MET360 pi-Sulfur	CYS338 alkyl		SER232 Hydrogen bond	ILE89 alkyl
ILE302 pi-Alkyl				GLU233 Hydrogen bond	
SANC00556	ILE89 alkyl	ILE302 pi-Alkyl			
	MET360 alkyl	ASP231 Hydrogen bond			
	MET360 alkyl	LYS205 Hydrogen bond			
	ALA203 alkyl	ILE302 alkyl			
	ASN204 carbon H Bond	MET298 alkyl			
		ILE89 alkyl			
	ASP182 pi-Anion				

All potential substrates presented at least an interaction with reported residues binding in the NADPH binding domain (in red) and in the fosmidomycin binding domain (in green) (see Table 4-5). ILE89 is the main contributor to these interactions.

4.4.4.4 Identified Hits

Table 4-6: Identified hits

Compound SANCDB ID	CAS ID	Name	Formula	Source Organisms	Known Activity	Binding energy	LE	LLE	QED
SANC00152	168075-14-7	Tsitsixenicin D	C24H32O6	Capnella thyrsoidea		-8.4	-0.28	-2.73	0.466
SANC00236	91236-90-7	Aplysulphurin-1	C22H28O5	Aplysilla sulphurea	Anticancer activity	-10.2	-0.38	-4.15	0.686
SANC00438	33916-25-5	Neodiospyrin	C22H14O6	Euclea natalensis	Antibacterial	-9.9	-0.35	-2.73	0.787
SANC00339	1162-10-3	Buphanidrine	C18H21NO4	Boophane disticha	Anti-serotonin transporter	-9.2	-0.40	-1.46	0.838
SANC00570	1000888-69-6	3,5,7-Trihydroxy-3-(3'-hydroxy-4'methoxybenzyl)-4-chromanone	C17H16O7	Pseudoprospero firmifolium		-8.1	-0.34	-1.15	0.68

Binding energy on the rigid crystal structure: kcal/mol / LE: Ligand efficiency kcal/mol/heavy atom
LLE: Ligand lipophilic efficiency (kcal/mol)

Comparing compounds ranking across the different receptors gave high Kendall's tau rank-correlation (τ) coefficients (0.61 between flexible and rigid receptors and 0.85 between closed and open conformations). Compounds thus conserved comparable ranking across the different receptors. Hits were selected from the rigid crystal structure. From the initial total of 699

compounds, five (5) final fits were chosen based on their binding energies, poses in the active site and interactions with the protein (see Table 4-6). Other metrics such as the QED score, LE, LLE were also used for filtering as described in the methodology section (see Figure 4-2). Hence, these ligands result from a careful selection process implying multiple metrics, presenting thus good qualities of hits compounds. Hits' binding energies ranged from -8.1 kcal/mol (SANC00570) to -10.2 kcal/mol (SANC00236). It is important to note that SANC00152 presented binding energies above the planned threshold of -8 kcal/mol. Indeed, the compound had -7.9 kcal/mol on PcDXR, -7.9 kcal/mol on PmDXR and -7.8 kcal/mol on PoDXR. However, the compound presented in its binding a strong network of hydrogen bonds with contact with two of the pockets in the active site.

About their binding poses, all ligands were in the protein active site, overlapping well with LC5. The most distant one (SANC00570) was 3.41 Å away. All hits present at least a moiety fitting well in one of the pockets. Deeply fitting in both pockets would cause an energetically unfavourable conformation of the molecule as observed with some compounds. A general observation is that all ligands fitting in the active site presented a preference of binding for the pocket adjacent to the phosphonate moiety of fosmidomycin binding pocket. This could be explained by the higher hydrophobicity of the pocket. Visually inspecting ligand binding poses revealed a third pocket. Any of the visualized ligands was binding in that pocket (see Figure 4-21).

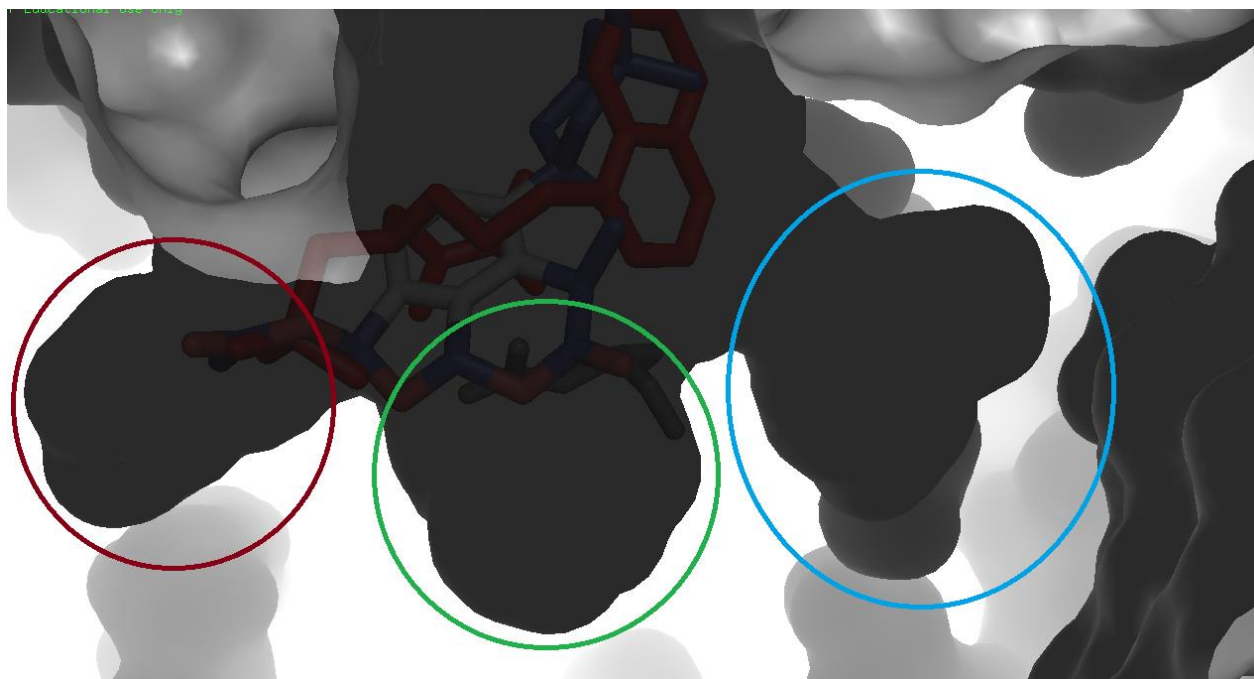


Figure 4-21: Phosphonate binding pocket (red), its adjacent pocket (green) and the third pocket (blue).

Figure 4-22 and Table 4-7 describe the molecular interactions and the fit of the hits in the protein active site. In bold, residues known from literature (see Table 1-1) as important for DXR inhibition. Among the selected hits, SANC00236 has a favourable interaction with the manganese metal in the active site. SANC00152 forms a strong network of hydrogen bonds and

interestingly has a moiety fitting in each of the two adjacent pockets, possible through an adopted conformation of the macrocycle present in its structure.

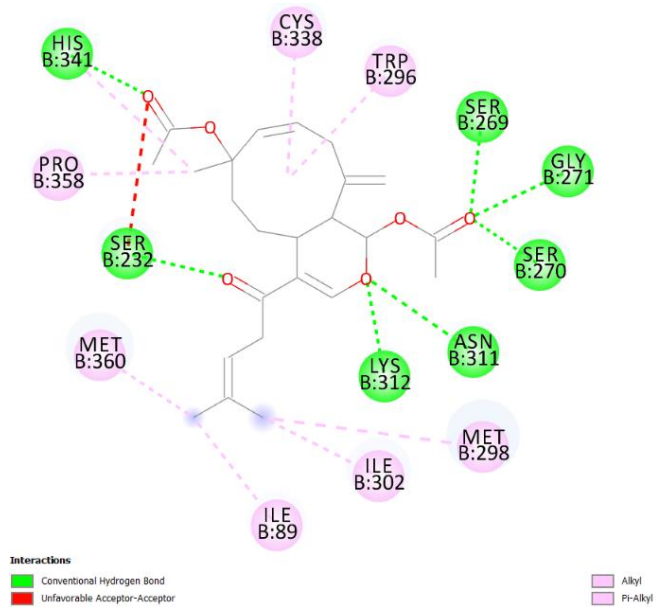
The hits showed interaction with all reported residues binding to the phosphonate moiety except HIS293. More this residue showed few interactions with the different ligands. Indeed, it showed only 4 interactions with the entire set of all SANCDB compounds. In contrast, ligands showed increased number of interaction with the aromatic ring of TRP296 (pi-Alkyl type of interaction), and MET298. This latter showed interesting pi-sulfur type of interaction with SANC00438, SANC00570, and SANC00339. These residues are in the loop region of the protein and appear among the most interacting residues. Aspartic acid 231 and the glutamic acid 315 are absent from the interactions while often reported in fosmidomycin-like binding particularly for the metal coordination.

Except for SANC00152, all ligands interact with GLU233, also known for metal coordination (Murkin, Manning, and Kholodar 2014). The residue generally displayed a pi-anion through its carboxylate group and a benzene ring found on the compound. As for SANC00152, the compound forms a hydrogen bond with SER232 in the same region. It is also the only compound showing an extension toward the cofactor binding site through an alkyl interaction with the aliphatic chain of ILE89. Among the hits, only SANC00438 interacted with the metal ion in the active site and through a hydrogen bond with SER232 and GLU233 in the same region. At the same time, the compound kept some interactions with SER270 and LYS312 implied in fosmidomycin binding.

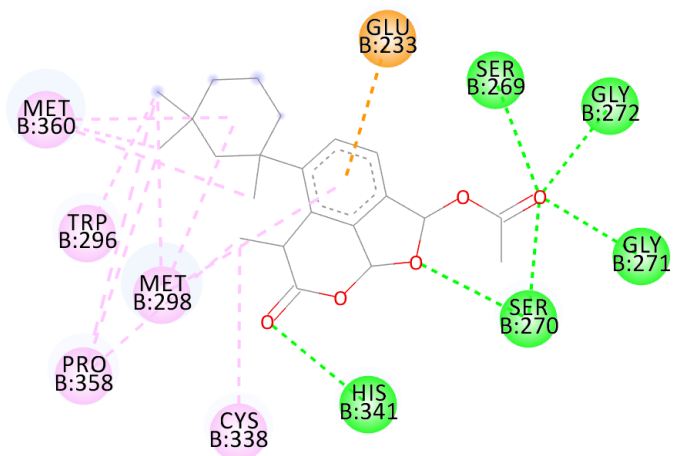
MET360 is present in all ligand interactions having a pi-Sulfur or an alkyl type of interaction. This residue is not reported as interacting with fosmidomycin-like inhibitors. Indeed, the residue is located outside the binding pocket, toward the NADPH binding region, closed to ILE89. It is the residue with the highest number of interactions with the ligands. As the residue is not located in a deep pocket, it could thus interact with ligand not fitting the pocket and even with those extending outside the pockets. In the same way, CYS338 interacts with all the identified hits through an alkyl type of interaction with its carbon atom bound to its sulfur. The residue is located near the entry of the second pocket. PRO358 is also located near the same region but closer to the third pocket and displayed an alky type of interaction through its pyrrole ring with the ligands. In general, CYS338, MET360 and MET298 were implied in numerous sulfuric interactions with the ligands.

All identified hits lack the phosphate moiety and hydroxamic acid moiety of fosmidomycin. The identified compounds can thus present new scaffolds with better pharmacological properties for DXR inhibition. These moieties were associated with fosmidomycin's poor drug likeness properties (Chofor et al. 2014). The interactions from closed conformation docking were analysed within Discovery Studio in a search for alternative bidentate ligands forming two or more bonds with the metal cation. Finding alternatives to the fosmidomycin hydroxamate group to chelate the metal is challenging (Chofor et al. 2014). None of the compounds were found to chelate the metal ion.

Molecular Interactions 2D Plot

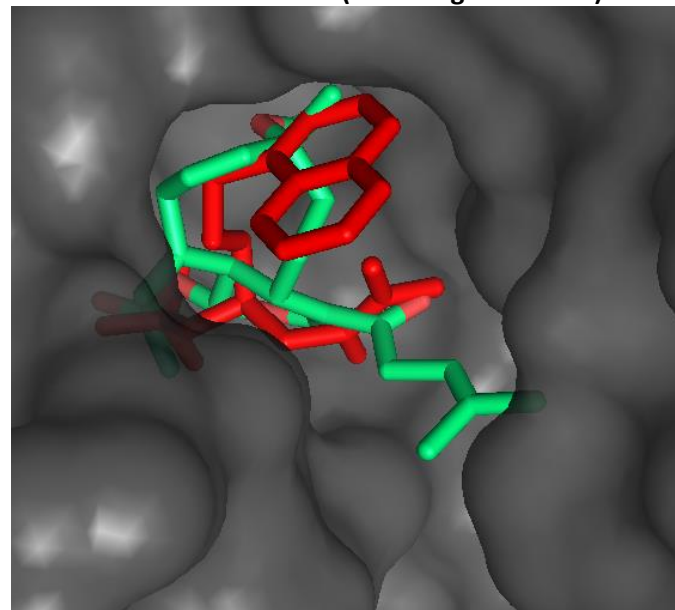


SANC00152

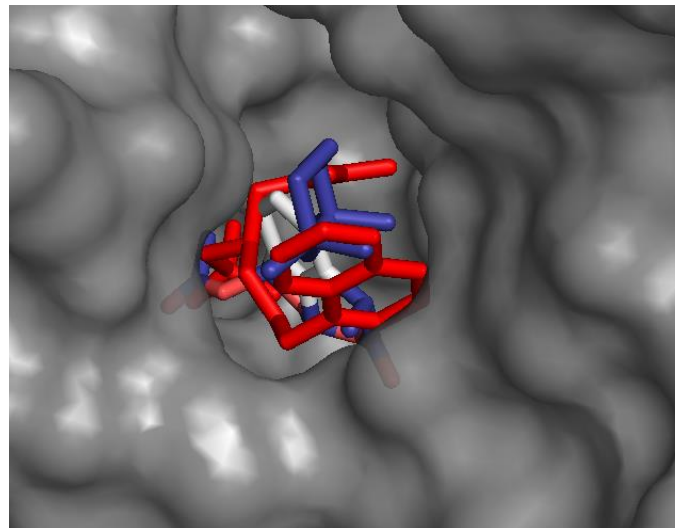


SANC00236

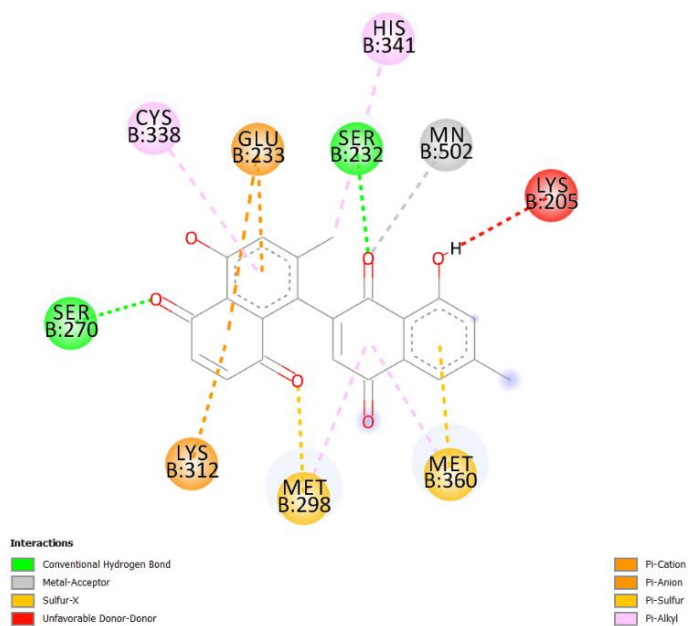
Fit in the active site (Native ligand in Red)



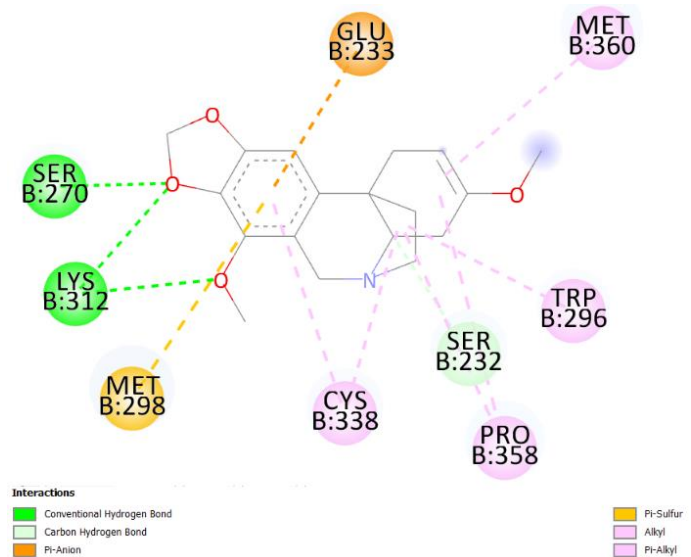
SANC00152



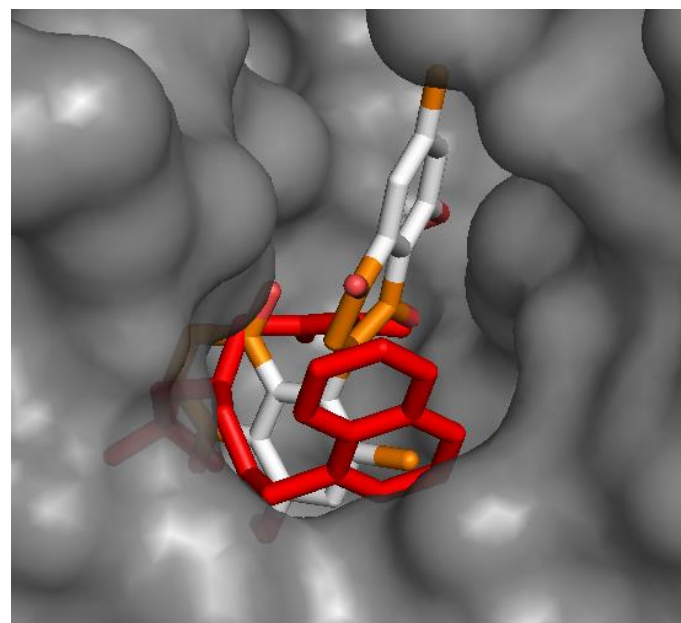
SANC00236



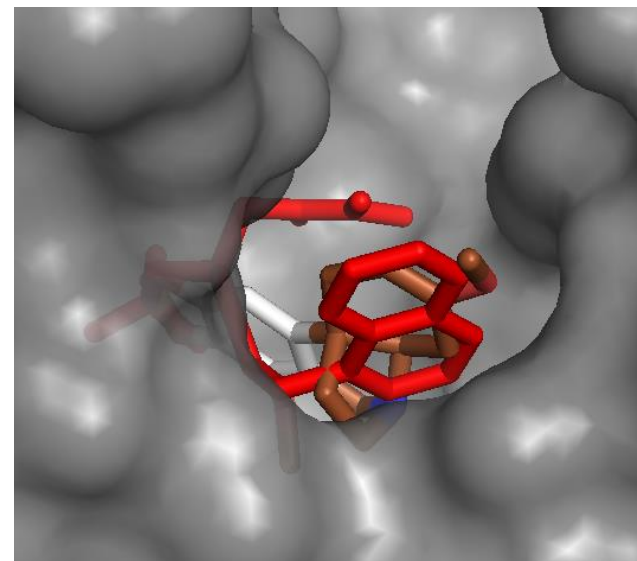
SANC00438



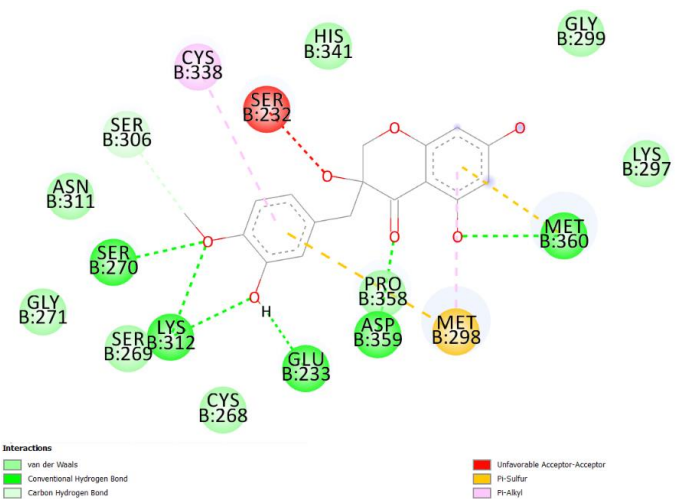
SANC00339



SANC00438

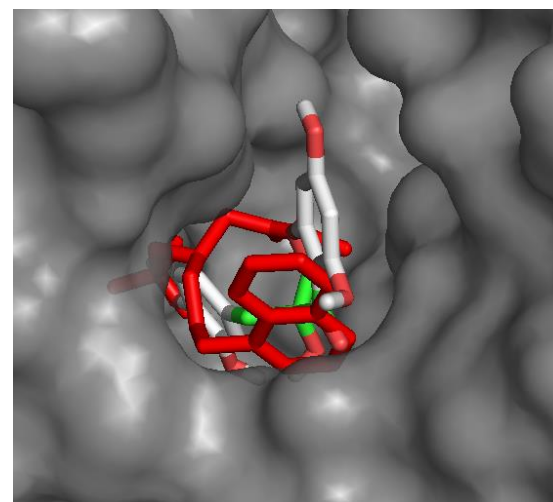


SANC00339



SANC00570

Figure 4-22: 2D plot and fit in the active site for identified hits.



SANC00570

Table 4-7: Hits molecular interactions. Red: residues interaction with fosmidomycin phosphonate. Green: residues interaction with fosmidomycin hydroxamate.

ID	Molecular Interactions (Residue and Type of interaction)		ID	Molecular Interactions (Residue and Type of interaction)	
SANC00152	TRP296 pi-Alkyl MET360 alkyl MET298 alkyl PRO358 alkyl ASN311 Hydrogen bond LYS312 Hydrogen bond ILE89 alkyl SER270 Hydrogen bond	SER269 Hydrogen bond CYS338 alkyl HIS341 Hydrogen bond GLY271 Hydrogen bond HIS341 pi-Alkyl ILE302 alkyl SER232 Unfavorable acceptor-acceptor	SANC00236	TRP296 pi-Alkyl GLU233 pi-Anion MET360 alkyl MET298 alkyl PRO358 alkyl SER270 Hydrogen bond SER269 Hydrogen bond	MET298 alkyl GLY272 Hydrogen bond MET360 alkyl HIS341 Hydrogen bond GLY271 Hydrogen bond CYS338 alkyl MET298 pi-Alkyl
SANC00438	MN502 Metal acceptor MET298 sulfur LYS312 pi-Cation CYS338 pi-Alkyl GLU233 pi-Anion SER270 Hydrogen bond	MET360 pi-Sulfur HIS341 pi-Alkyl SER232 Hydrogen bond MET298 pi-Alkyl MET360 pi-Alkyl LYS205 Unfavorable donor-donor	SANC00339	MET298 pi-Sulfur SER270 Hydrogen bond GLU233 pi-Anion SER232 Carbon HBond MET360 alkyl	TRP296 pi-Alkyl CYS338 pi-Alkyl CYS338 alkyl PRO358 alkyl LYS312 Hydrogen bond
SANC00570	SER270 Hydrogen bond SER306 Carbon HBond ASP359 Hydrogen bond MET298 pi-Sulfur MET360 pi-Sulfur	SER232 Unfavorable acceptor-acceptor MET298 pi-Alkyl MET360 Hydrogen bond LYS312 Hydrogen bond CYS338 pi-Alkyl GLU233 Hydrogen bond			

SANC00236, SANC00438, SANC00339 have recorded anticancer activity, antibacterial and anti-serotonin transporter respectively but none has known anti-malarial activity according to the information on the SANCDB database. SANC00438 (Neodiospyrin) belongs to the class of naphthoquinones, well known for their antimalarial activity (example atovaquone). Nonetheless, naphthoquinones act by inhibition of mitochondrial electron transport (Schuck et al. 2013) while the current protein is present in the non-mevalonate pathway. SANC00339 (Buphanidrine) belongs to the class of alkaloid and amaryllidaceae. Quinidine, cinchonine and cinchonidine are all alkaloid and effective against malaria (Achan et al. 2011). Lastly, SANC00570 is a flavonoid and flavonoids from *Artemisia annua L.* as have been to have potential synergism effect when combined with artemisinin. As for SANC00236 (Aplysulphurin-1), it is a terpenoid, class of compound synthesized through the non-mevalonate pathway (Guggisberg, Amthor, and Odom 2014).

Ranked according to ligand efficiency, top compounds showed relatively low binding energies. The top 20 compounds had an affinity ranging from -7.6 kcal/mol to 4.7 kcal/mol. However, this was expected considering the few number of carbon atoms (less than 15) among these compounds. Fosmidomycin and FR98 were ranked 32nd and 35th. They thus have better ranking than when using the sole binding energy criterion.

Compounds were also ranked according the number of hydrogen bond formed according to the hydrogen bond from Discovery Studio. Even though the hydrogen bonding term is integrated in the scoring function to calculate the binding energies (Trott and Olson 2010), hydrogen bonds have been shown as major contributors to protein ligand interactions (Du et al. 2016). This was also supported by the observation that fosmidomycin, FR98 and LC5 showed higher number of hydrogen bonds 11, 9, 14 respectively compared to the rest of the SANCDB compounds. Among the hits, SANC00236 showed to be a promising compound with five (5) hydrogen bonds.

It is noteworthy that the five (5) hits result from a set of preselected compounds satisfying the conditions described in the methodology (see Figure 4-2). These preselected compounds still present suitable features to be considered for further investigation. The list of compounds is in Table 4-8 below.

Table 4-8: Preselected compounds

compound ID	Binding energy (kcal/mol)	LE (kcal/mol/heavy atom)	LLE (kcal/mol)	# Hydrogen bonds	Distance to active site (Angstrom)	logP	QEDw	Lipinski Violation
SANC00346	-10.3	-0.32	-3.13	6	3.58	4.14	0.566	0
SANC00355	-10.2	-0.32	-3.46	6	3.40	4.47	0.522	0
SANC00434	-9.7	-0.35	-2.74	5	3.12	3.73	0.787	0
SANC00436	-9.7	-0.35	-2.75	3	3.51	3.74	0.787	0
SANC00344	-9.5	-0.31	-4.54	1	3.94	5.52	0.566	1
SANC00435	-9.3	-0.33	-2.75	4	3.31	3.72	0.727	0
SANC00374	-9.2	-0.38	-1.96	2	1.72	2.92	0.874	0
SANC00183	-8.9	-0.31	-2.73	8	3.92	3.68	0.625	0
SANC00326	-8.7	-0.33	-1.42	2	2.00	2.36	0.826	0
SANC00345	-8.7	-0.36	-0.40	5	3.54	1.34	0.816	0
SANC00661	-8.6	-0.39	-0.50	13	2.15	1.43	0.34	0
SANC00575	-8.6	-0.33	-1.64	8	3.85	2.57	0.846	0
SANC00225	-8.5	-0.37	-1.89	9	3.77	2.82	0.754	0
SANC00562	-8.4	-0.30	-0.11	13	3.40	1.03	0.416	1
SANC00229	-8.3	-0.42	-2.29	10	3.82	3.21	0.473	0
SANC00335	-8.3	-0.36	-2.03	8	3.02	2.95	0.672	0
SANC00626	-8.3	-0.35	-2.23	8	2.17	3.15	0.67	0
SANC00529	-8.2	-0.34	-1.53	10	2.63	2.44	0.493	0
SANC00337	-8.2	-0.36	-1.13	8	2.92	2.04	0.619	0
SANC00566	-8.1	-0.29	-1.77	10	3.87	2.68	0.776	0
SANC00260	-8.1	-0.35	-1.91	9	3.73	2.82	0.806	0

A general observation on these compounds was that majority of them preferred binding in the pocket next to the phosphonate binding pocket, with an extension toward the NADPH binding region for large compounds. None of them were observed to bind in the third pocket. With SANC00355, SANC00346, SANC00562, SANC00434 and SANC00436 a moiety of the compound fits well in the adjacent pocket to the phosphonate binding pocket. SANC00374, SANC00326 present an interesting pose by fitting in the two pockets. A hydroxyl on a benzene ring fits in phosphonate moiety binding pocket while a pyrrole ring fits in the adjacent pocket. The compounds present a pyrrole ring fitting in the pocket adjacent to the phosphonate binding pocket. SANC00661 presents an interesting pose by fitting well in the two pockets and forming a strong network of hydrogen bonds. But as for SANC00152, the compound presents a macrocycle allowing such pose, which may latter be identified as a problematic moiety for drug likeness. These macrocycles appear to be the reason for their fitting in the two adjacent pockets.

4.4.4.5 Pharmacological properties

The properties for the set of SANCDB compounds were evaluated (see Table 4-9). Ten compounds were identified, presenting the best weighted QED scores (≥ 0.9). All the identified compounds follow the Lipinski of 5 ($MWT \leq 500$, $\text{LogP} \leq 5$, H-bond donors ≤ 5 , and H-bond acceptor ≤ 10) (Lipinski 2004). Except for SANC00387, their binding energies are below -7.8 Kcal/mol on the rigid receptor in its close conformation (5JAZ). Their bioavailability was also assessed according to the

VEBER rule (good oral bioavailability for range (rotatable bonds ≤ 10) and ($tPSA \leq 140 \text{ \AA}$ or H-Bonds Donors+H-Bonds Acceptors ≤ 12)) and EGAN rules (good orally available ($0 \geq tPSA \leq 132$) and ($-1 \geq \log P \leq 6$)). All compounds presented good bioavailability according to these metrics. In their chemical structure, they present at most 2 aromatic rings with no structural alert involved in toxicity problem.

Table 4-9: Drug-like properties of top SANCDB compound according to QED score calculated using FAF Drugs4.

SANCDB ID	MW	LogP	HBA	HBD	TPSA	Lipinski Violation	QEDw	RB	AR	SA	OBV	OBE	Binding Energy
SANC00565	314.33	2.95	5	1	64.99	0	0.94	4	2	0	Good	Good	-8.3
SANC00379	327.37	1.95	5	1	60.2	0	0.9	2	1	0	Good	Good	-8.3
SANC00358	300.31	3.18	5	2	75.99	0	0.91	3	2	0	Good	Good	-8.1
SANC00336	300.31	2.63	5	2	75.99	0	0.911	3	2	0	Good	Good	-8
SANC00688	300.31	3.18	5	2	75.99	0	0.91	3	2	0	Good	Good	-8
SANC00689	298.29	3.29	5	2	79.57	0	0.909	3	2	0	Good	Good	-8
SANC00385	329.39	3.01	5	2	63.36	0	0.903	4	2	0	Good	Good	-8
SANC00376	341.4	3.05	5	1	51.16	0	0.93	3	2	0	Good	Good	-7.9
SANC00362	314.33	3.51	5	1	64.99	0	0.935	4	2	0	Good	Good	-7.8
SANC00613	314.33	3.51	5	1	64.99	0	0.935	4	2	0	Good	Good	-7.8

BE: Binding energy 5JAZ closed conformation (Kcal/mol). RB: Rotatable Bond. AR: Aromatic Ring. SA: Structural Alert. OBV: Oral Bioavailability (VEBER). OBE: Oral Bioavailability (EGAN).

These compounds can be interesting in high throughput virtual screening using multiple known drug targets. As they present excellent drug likeness properties, additionally testing their binding affinities and identifying any hit could provide a very good starting point for lead compounds.

Table 4-10: Hits drug likeness

SANCDB ID	MW	LogP	HBA	HBD	TPSA	Lipinski Violation	QEDw	RB	AR	SA	OBV	OBE
SANC00152	416.51	3.65	6	0	78.9	0	0.466	7	0	2	Good	Good
SANC00236	372.45	5.16	5	0	61.83	1	0.686	3	1	1	Good	Good
SANC00438	374.34	3.73	6	2	108.4	0	0.787	1	2	0	Good	Good
SANC00339	315.36	2.42	5	0	41.36	0	0.838	2	1	0	Good	Good
SANC00570	332.3	2.06	7	4	116.5	0	0.68	3	2	0	Good	Good

RB: Rotatable Bond. AR: Aromatic Ring. SA: Structural Alert. OBV: Oral Bioavailability (VEBER). OBE: Oral Bioavailability (EGAN).

The lowest QED score among the hits was 0.466 (see Table 4-10). Only SANC00236 violates one of the Lipinski rule of logP with a value of 5.16. Nonetheless, one can note that a variant of the rule has a cut-off of 5.6 (Schneider 2013). SANC00152 presents two structural alerts linked to

the macrocycle in its structure. Except for these two observations, the identified hits present good drug likeness scores.

4.5 Conclusion

The binding energies for the reference ligands did not provide a significant threshold for filtering the compounds. Hence the necessity to introduce other metrics for hit selection. Though the binding energy remained the main selection criteria. LE, LLE and drug likeness properties were introduced as recommended in virtual screening studies (Zhu et al. 2013). Combining these different metrics and a visual examination of the compound allowed to finally identify 5 hits: SANC00152 (Tsitsixenicin D), SANC00236 (Aplysulphurin-1), SANC00438 (Neodiospyrin), SANC00339 (Buphanidrine) and SANC00570 (3,5,7-Trihydroxy-3-(3'-hydroxy-4'methoxybenzyl)-4-chromanone) as new potential DXR inhibitors. More, the absence of the phosphate moiety or hydroxamic acid moiety of fosmidomycin and their good QED scores allow to expect better drug likeness properties for these hits. Considering its scores on the different used metrics (binding energy -10.2 kcal/mol, LE:-0.38 kcal/mol/heavy atom, LLE:-4.15 and QED score of 0.686), SANC00236 appears to be the most promising compound.

Docking on multiple conformations of the receptor (closed, open, closed with flexible residues) revealed a correlated ranking of the compounds across the different settings of the receptor. Flexible residues allowed to have an insight into the induced fit binding mechanism of fosmidomycin in the protein active site. For example, in this setting, we observed a correct orientation of fosmidomycin around the metal atom in the active site. A significant increase in the binding energy was observed for the receptor with flexible residues. These configurations can represent different frames of the protein dynamic. Nonetheless, molecular dynamic simulations would certainly provide better insight in the landscape of protein conformations and help apprehend other factors such as the importance of solvation.

Exploring possibility for bisubstrate inhibitors revealed an interesting compound SANC00615 (5-Hydroxy-7-methoxy-3-(3-hydroxy-4-methoxybenzyl) chroman-4-one) with a good QED score of 0.896, showing promiscuity for both NADPH and fosmidomycin binding site. A molecular dynamic simulation will certainly provide more insight into the ligand stability and its interactions with the two sites. Finally, none of the docked compounds were found to chelate the metal ion.

CHAPTER 5: MOLECULAR DYNAMICS

5.1 Introduction

Molecular dynamics (MD) is the movement of molecules and atoms through time. A MD simulation uses computational methods to follow the evolution of these particles. The ensemble of particles constitutes the system. Algorithms are used to follow the positions and speeds of the elements in the system. Time has a discrete evolution characterized by the simulation number of steps. The algorithms compute the position and potential energy of each element at each step (see Figure 5-1). Classical and *ab initio* molecular dynamics are the two approaches used to compute the positions and potential energies of atoms. In the *ab initio* technique, quantum mechanics principles through solving the Schrödinger equation are used to calculate forces and hence particles' position evolution while in classical MD, force-field approaches determine the forces. In either case, the trajectory of particles proceeds using Newton's Laws of Motion. Hybrid models combining quantum and molecular mechanics (QM/MM) can also be used (Allen and others 2004; Petrenko and Meller 2001).

In any case the nuclear positions are critical. From the QM perspective, two main approximations are used: the Born–Oppenheimer approximation and neglecting electrons masses compared to the nuclei one which is consequently treated as a point particle following classical Newtonian dynamics (Datta, Datta, and Davim 2016). In the MM approach, the system can be described by a method analogous to treatment as network of balls and springs. Atoms are represented by balls and the interactions between atoms with springs (for the bonding interaction; other force-field interactions are included). The simulation thus results in simply solving the Newton's equations of motion (see Equation 1) for the particles in the system during the defined number of steps. Different configurations of the systems are thus produced at each step resulting in a trajectory which describes the variations of positions and speeds of the elements in the system over time (Petrenko and Meller 2001).

Equation 1 .

$$F_i = m_i a_i = m_i \frac{d^2 r_i(t)}{dt^2}$$

F_i : Force acting on particle i

m_i : mass of particle i

a_i : acceleration of particle i

r_i : position vector

t : time

Solution to equation 1 helps find positions at time $t + \Delta t$ based on previously known positions and velocities of the particles in the system. Solving the equation for the particles in the system requires knowledge of the forces, initial positions, velocities, and algorithms used to solve this

equation involves the use of numerical integrators. Initial positions can be derived from crystallographic data and initial velocities from a Maxwellian distribution at the temperature of interest. The distribution describes the probability that an atom i of mass m_i has velocity v_i in the direction x at a temperature T (González 2011). Examples of commonly used algorithms include Verlet, leap-frog and Velocity Verlet schemes (Petrenko and Meller 2001). For these, the last elements for use in the algorithm are the forces derived from the force fields.

Classical MD uses empirical potentials or force fields. Forces and potentials are important in MD simulation as they describe the system potential energy. They describe, in the form of mathematical expressions, the conditions governing the interactions between the elements in the system. The considered interactions are the non-bonded interactions (the short-range and the long distance) and the bonded or intra-molecular interactions. This latter includes the stretching (between two atoms), the bending (three atoms) and the torsional (4 atoms) terms. The non-bonded forces are related to van der Waals forces and electrostatic charge. The total energy of the system can thus be expressed in the following form (Vanommeslaeghe, Guvench, and MacKerell 2014; MacKerell et al. 1998).

Equation 2

$$E = E_{bonded} + E_{nonbonded}$$

In which:

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral}; E_{nonbonded} = E_{electrostatic} + E_{vand\ der\ Waals}$$

Some of the most popular force fields developed for simulation of macromolecules include AMBER (Assisted Model building with Energy Refinement) (Cornell et al. 1995), CHARMM (Chemistry at HARvard Molecular Mechanics) (Brooks et al. 1983), and GROMOS (GRONingen MOlecular Simulation) (van Gunsteren et al. 1996).

Functional form of the AMBER03 force field:

Equation 3

$$E_{total} = \sum_{bonds} K_b (b - b_{eq})^2 + \sum_{angles} k_\theta (\theta - \theta_{eq})^2 + \sum_{torsions} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)] + \sum_{i < j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

The first three terms in the above equation describe energies of the bonded interactions. Covalent bond stretching and angle bending are estimated by a harmonic potential with K_b and k_θ being the force constants for bonds and angles and b and θ being the bond length and bond angle and θ_{eq} , b_{eq} being the equilibrium bond lengths and angles. V_n , ϕ , γ are respectively the force constant, dihedral angle and the phase angle for dihedrals. The last term describes the Van der Waals interactions (A_{ij}), the London dispersion terms (B_{ij}) and the Coulombic interactions with q_i and q_j being the partial atomic charges and ϵ the dielectric constant (Duan et al. 2003). These parameters need to be known for all atoms in the system.

Many research efforts allowed to have force field parameters for common biological molecules such as amino acids, nucleic acids, carbohydrates and small druglike molecules. Nonetheless, developing parameters for metal centers remain challenging. One of the reasons is the complex geometries of metal centers. Metals can bind different ligands and multiple atoms around them. More these bonds have strengths that range between covalent and hydrogen-bonding strength resulting in flexible geometries. Furthermore, for transition metals, quantum mechanical ligand-field, spin-state, trans, and Jahn–Teller effects are more significant (Neves et al. 2013; Hu and Ryde 2011).

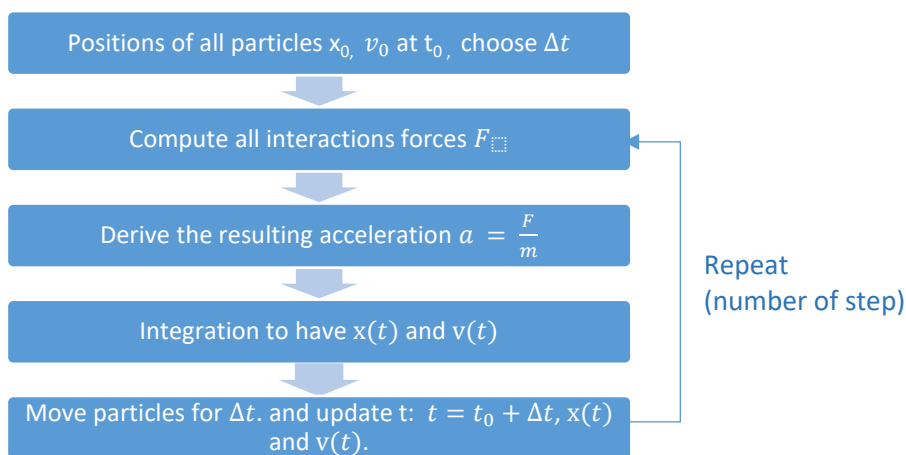


Figure 5-1: Simplified diagram of MD simulation. Adapted from (Badrinarayan, Choudhury, and Sastry 2015).

Other important aspects of MD simulations include the statistical mechanical ensembles, the solvent models and the periodic boundary conditions.

Statistical mechanical ensembles refer to all possible microstates (conformations) of the system. These states depend on the number of particles in the system, the volume, the pressure, the temperature and the energy. The volume and pressure are dependent parameters, the same applies for the energy and the temperature. Combining these parameters allows to have three main ensembles. The microcanonical or NVE in which the number of particles (N), the volume (V) and energy (E) are constant. The canonical ensemble or NVT in which the number of particles (N), the volume (V) and the temperature are constant. And finally, the isothermal-isobaric ensemble which maintains constant the number of particles (N), the temperature (T) and the pressure (P). Newton’s Laws of motion implies the conservation of energy, thus MD trajectories become a set of configuration sampled from the NVE ensemble (Petrenko and Meller 2001). In reality, systems exchange energy, matter and volume with the environment. But such implementing such systems can result in much more complex algorithms (Rigden 2017).

Water is an important parameter and can have important effects on solutes in the system. For example, water molecules can form hydrogen bond networks between protein and ligand and therefore play important role in binding affinities (Nguyen, Viet, and Li 2014; Ben-Naim 2002). To account for the water effect, implicit and explicit solvent models have been developed. Explicit models are more accurate, but computationally expensive, thus slower. While implicit ones are

less accurate but computationally faster. The main difference is that this latter does not integrate the density fluctuation behaviour around solutes (Roux and Simonson 1999). This is some of the MD limitations. Some of the most popular developed water solvation models include TIP3P, TIP4P, TIP5P, SPC and SPC/E (González 2011).

Another limitation is the space occupied by the system. Ideally, molecular systems behave in an “infinite” space. However, due to the computational cost associated with such systems, unit cells combined with periodic boundary conditions are used in MD simulation. Periodic boundary conditions are used at the frontiers of the simulation box to mimic an infinite space by stacking simulation boxes. When any element in the box crosses the box on one side, it re-enters it by the opposite side (González 2011).

Computer’s computational capacities are limited. In a system with N particles, there are $N(N-1)/2$ possible interactions. The number of interactions, thus grows in the order of $O(n^2)$. As a result, the number of particles in a system is also limited, limiting the size of the simulation box to prevent computationally demanding simulations. Interactions are limited to closest ones, using thresholds. However, using cut-offs can introduce error for significant long range interactions. Thus, methods such as Ewald summation or the particle–particle–particle–mesh (P3M) are used to calculate long-range interactions (Kolafa and Perram 1992; Toukmaji and Board 1996). As with space, MD has also a time limitation severely hampering the exploration of many biological processes. Due to the fast motions of molecules, time steps between integrations should be very small, in femtosecond (10^{-15} s) order. As a result, highly increased number of steps are required for long simulations (Petrenko and Meller 2001).

More, force fields make use of a set of predefined bonding parameters for atoms in the system. Harmonical functions for bonds and angles in force field equations imply no bond breaking or forming. They are thus not able to model chemical reactivity with bond forming and breaking explicitly (González 2011). Furthermore, most of the force fields do not take into account the electronic polarization effect (Ganesan, Coote, and Barakat 2017).

In terms of applications, MD has been used to study protein folding and unfolding, protein-ligand, protein-protein interactions, and to decipher cryptic pockets in proteins. Molecular dynamics is commonly used in drug discovery. Docking provides an initial estimation of protein-ligand affinity while MD provides a more accurate description of protein-ligand complexes in terms of thermodynamics and kinetics during recognition and binding. This is achieved through explicitly treating structural flexibility and entropic effects and it provides more insight into the induced-fit and conformational selection paradigms beyond the lock-and-key one (De Vivo et al. 2016). MD studies have contributed to the success stories of many commercialized drugs (Mortier et al. 2015). For example, allosteric inhibitors of the M2 muscarinic acetylcholine receptor have been found through MD simulations technique (Dror et al. 2013).

Many software tools exist to conduct MD simulation. GROMACS, NAMD, LAMMPS are available in the public domain while GROMOS, CHARMM and AMBER are commercial ones (González 2011).

In this chapter, we intend to further investigate the hits identified in the docking chapter for their stability in the protein-ligand complexes in GROMACS (Abraham et al. 2015). DXR has a metal ion

in the active site. The crystal structure 5JAZ used in this study has a manganese ion (Sooriyaarachchi et al. 2016). Currently, there is no parameter for the manganese atom in the force fields implemented in GROMACS 5.1.4 (Abraham et al. 2015). We will then start by developing force field parameters for that metal using PES (Potential Energy Surfaces) scans followed by their implementation in GROMACS. These parameters will finally be validated with MD simulation.

Written in C and C++, GROMACS (Abraham et al. 2015) is a free molecular dynamics software that achieves state of the art performance in MD simulation by using multi-level parallelism across computers' cores and optimized algorithms for modern MD. A simulation is decomposed into domains using MPI and load balancing, while SIMD registers parallelize bonded interactions on cores and nonbonded interactions are handled by GPUs and other accelerators. GROMACS proposes different force fields for running simulations including GROMOS96, OPLS-AA, AMBER and CHARMM (Abraham et al. 2015).

The general steps for the molecular dynamic simulation in GROMACS are the following: system preparation, solvation, neutralization, energy minimization, equilibration, the MD run and finally the analysis of the generated data (Abraham et al. 2015).

GROMACS offers a convenient command line interface. Each command run output a log file of the process and results which is useful for monitoring and debugging. Most inputs and output files from the tool are in plain-text format, making it easy to track all simulation parameters. For example, an simulation .log file informs on how a simulation run and its performances. The tool automatically backup old output files, renaming them by using the prefix “#”.

The following experimental method section is derived from the GROMACS online documentation (GROMACS Documentation n.d.), its user manual version 5.0.7 (Mark Abraham, Berk Hess, David van der Spoel, and Erik, Ren, and Vanden-Eijnden 2002), and from the GROMACS tutorials Protein-Ligand Complex (Protein-Ligand Complex n.d.) available at <http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/complex/index.html> and Lysozyme in Water (Lysozyme in Water n.d.) available at <http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/lysozyme/index.html>.

5.1.1 System preparation

In this step, the system is defined. This includes protein structures retrieved from the PDB (Berman et al. 2000), from modeling or protein-ligand complexes from docking experiments. Some aspects to consider about the structures are their quality: resolution, missing residues, and the presence of ligand. The GROMACS command `pdb2gmx` is used to create topologies for the structures. It reads in a structure (.pdb or .gro) file, adds hydrogens and output a structure file (.gro), a topology (.top), and a position restraint file (.itp) in GROMACS format. The .gro contains atoms coordinates and their velocities information. The topology describes the system: force field, components, atoms, bonds, angles, position restraints... The topology file is divided in multiple sections. It uses the “#include” mechanism to combine the different components of the system listed in its “[molecules]” section. The sequence of molecules in the .gro must match the one in that section. The position restraint file is used to restrain heavy atoms positions.

An important step here is the choice of the force field. GROMACS proposes different force fields (CHARMM) (Brooks et al. 1983), AMBER (Cornell et al. 1995), GROMOS (van Gunsteren et al. 1996), OPLS (GROMACS Documentation n.d.).

For the ligands, different external tools can be used to generate their topologies as GROMACS does not automatically recognize some of the species. For example, AcPype (Antechamber PYthon Parser interface) (Sousa da Silva and Vranken 2012; Batista et al. 2006) based on the Antechamber (Wang et al. 2006; Wang et al. 2004) module can be used. Other external tools include PRODRG web server (Schüttelkopf and van Aalten 2004), CGenFF (Vanommeslaeghe et al. 2010) and ATB (Automated Topology Builder) (Malde et al. 2011) also can be used.

This information will then be incorporated in the final coordinate to build the complex (protein-ligand) and topology file using the “;include” statement and updating the molecules section.

5.1.2 Defining simulation box and solvation

The editconf module is used to set up the simulation box. Some important options are: -bt to specify the box type, -c to center the protein in the box, -d to specify the distance solute-box (in other terms the box size). The box should be large enough so that an re-entering element cannot bump to itself. Though the box should not be too large to prevent computationally demanding simulation. GROMACS proposes different box shapes: triclinic, cubic, dodecahedron, octahedron having different implications on the box size. For example, using the same periodic boundary conditions (-d) the dodecahedron box is ~71% the size of the cubic one, saving thus some space for faster computation.

Next, the solvate command is used to add water to the box. Options -cp specify the box and -cs the solvent. As previously described, different water models can be used. Non water solvent also can be used. The -p option specifies the topology file which is a result automatically updated with the number of solvent molecules added.

Finally, to maintain the chemical neutrality of the system, ions are added to the system. The previous solvated system is charged based on the protein amino acid composition. Prior to adding ions, a .trp (portable binary run input) file is first generated using gmx grompp. The command gmx grompp check and process the generated topology file from a molecular description to an atomic description of the system. For that process, it requires an mdp (molecular dynamics parameters) file, here ions.mdp.

The command gmx genion is then used to neutralize the system. Ions are added to the system to neutralize using the following options -pname, -nname (example: -pname NA -nname CL to add NaCl), -nn and -np to specify the number of ions, either positively or negatively charged to neutralize the system. More conveniently, -conc and -neutral options allow to neutralize the system with a certain concentration. The user is prompted to specify entities to be replaced in the system, resulting in a neutral solvated system.

5.1.3 Energy minimization

This step allows to remove from the system its imperfections, steric clashes, improper solvent and ions orientations by finding a local energy minimum. The structures from the PDB database or the ligand structure may have bond lengths, angles, interatomic distances) that are not taken into account perfectly by the force field. A .tpr file is first created using the grompp command, then the structure energy is minimized using the created .tpr file. The energy module helps to monitor the minimization process which is performed using the mdrun program. The main energy terms of the process are in the resulting .edr file. Accessorily, a .trr (binary full-precision trajectory) and .gro (energy minimized structure) are also obtained.

The parameters for minimization are defined in the .mdp file and passed to the command. Some important ones include the algorithm used specified by value of *integrator* (example: steepest-descent), the maximum number of steps for minimization (nstep) and emtol in $\text{kJ}\cdot\text{mol}^{-1}\cdot\text{nm}^{-1}$ minimum force as target value to reach in order to achieve minimization. Meeting one of the two criteria (number of steps or the maximum force) results in minimization ending. Depending on its size, the system should ideally have a negative E_{pot} in the order of 10^{-5} and 10^{-6} , with the maximum force lower than the set target. Though it is possible to reach a reasonable E_{pot} with F_{max} still greater than emtol depending on the defined value.

5.1.4 Equilibration

At this step, the system will be equilibrated by bringing it at proper temperature and pressure. Water molecules and ions are relaxed around the restrained solutes, optimizing thus the system. Both steps require .tpr files generated using the grompp command. During equilibration, it is important to restrain the solutes in the system to avoid its collapse. The restraining is achieved by applying a force on the heavy atoms of the protein and ligand and using the parameter “define = -DPOSRES”. The pdb2gmx generated a position restraint file (.itp) for the protein and a similar restraint file should be generated for a ligand using the genrestr command. These information are included in the topology file in their correct location.

The first phase is run under a canonical (NVT) ensemble (constant Number of particles, Volume, and Temperature). During it, the temperature should reach a plateau at the specified value and stabilizes around it. The next step, NPT equilibration, conducted under an "isothermal-isobaric" ensemble, stabilizes the system pressure. The parameters for the two steps are specified in their respective .mdp files. The modified Berendsen thermostat is used for temperature coupling and the Parrinello-Rahman is used for pressure coupling.

An important aspect at this stage is the temperature coupling. Coupling every single entity in the system to its own thermostat group can be problematic for the coupling algorithm. The ideal approach is to merge some system components using the gmx make_ndx command. For example, one can set “tc_grps = Protein Non-Protein” in the parameters (.mdp) file after dividing the system into two parts, the protein and non-protein group using gmx make_ndx.

One can monitor the equilibration by plotting the generated .edr files from the equilibrations. The temperature and pressure should reach the respective specified values and stabilize around

them. One can also compare the experimental values of density for the used water model to the ones obtained during equilibration. Incorrect temperature or pressure behaviour may indicate an incorrect system behaviour.

5.1.5 MD simulation

After equilibration, the system is ready for the simulation. Restraints on the protein and ligands are first removed. Again here the run is preceded by a grompp command with the simulation parameter file. Important parameters include the length defined by the number of steps, the integrator for Newton's equations of motion and the time step for integration which are defined in the .mdp file.

At the end of the simulation, the produced files will be analysed.

5.1.6 Analysis

GROMACS offers diverse analysis modules to study MD results. Analysed properties depend on the type of study. Generally, the simulation quality and stability should first be validated and then subsequent analysis related to the research question can be conducted. Different ways of analysis include protein conformations analysis, trajectory visualization, distance, angles and measurements, principal component and interactions (such as hydrogen bonds) analysis.

Some of the important modules for analysis include gmx trjconv, gmx rms, gmx rmsf, gmx gyrate. GROMACS also proposes the group modules which can be helpful in grouping atoms of the system for specific analysis for example when it comes to measuring distances.

The stability and the dynamics of interactions of the complex protein-ligand were studied using the following measurements: the RMSD (Root Mean Square Deviation), RMSF (Root Mean Square Fluctuation) and the radius of gyration (Rg). Hydrogen bonds are essential for the stability of protein-ligand complexes. The gmx hbond module analyses the hydrogen bonds between different components of the system. The program gmx energy allows to analyse system properties such as temperature, potential energy, pressure, density...

At the end of the simulation, the output files structural (.gro), compressed trajectory (.xtc), energy (.edr) and (.log) are often used for analysis. Visualization tools such as VMD (Humphrey, Dalke, and Schulten 1996) are used for trajectories and XMGRACE (Turner 2005) to plot graphs resulting from the different properties analysed.

5.2 Methodology

The methodology used for the different MD simulations and the metal parameterization are described in this section.

5.2.1 Metal parameterization.

The manganese metal ion in DXR active site is involved in substrate and inhibitor binding. Fosmidomycin-like inhibitors showed metal chelation as key characteristics (Umeda et al. 2011). The different force fields implemented in GROMACS 5.1.4 (Abraham et al. 2015) do not provide parameters for the Mn atom. The metal center was of interest in the investigation of DXR inhibition, reason to undertake its parameterization.

Various methods exist to develop force fields parameters. They include fast methods such as the nonbonded model with restraints, the bonded parameterization based on the method of Seminario or more accurate technique as the Norrby and Liljefors method (Hu and Ryde 2011). As often in computational biology, a compromise needs to be found between speed and accuracy taking into account the intended application. In this study, we used QM calculations to obtain parameters to be included with the AMBER03 (Duan et al. 2003) force field in GROMACS. Potential energy surface scans were performed on the QM optimized subset of the metal center. The following parameters were calculated: equilibrium bond lengths, angles, and their corresponding values of force constants for each potential.

5.2.1.1 Subset Setup and geometry optimization

Gaussian09 was used for geometry optimization. The optimization finds a local energy minima on the PES. This section contextualizes information required from the Gaussian online documentation available at <http://gaussian.com/> and its use in this study. By default, Gaussian optimizes the geometry in redundant internal coordinates. During the process, the geometry is adjusted until a stationary point is found on the potential energy surface using the Berny algorithm using GEDIIS (Li and Frisch 2006). The energy is calculated in Hartree (1 Hartree = 627.15 kcal mol⁻¹; or 1 Hartree = 2625.5 kJ/mol). Two convergence criteria are used, the changes in the energy gradient and the structure on two consecutive calculations. Optimization completes when criteria fall below the fixed threshold values for these two parameters.

The subset of the protein active site including the metal and its coordinating residues (ASP231, GLU233, GLU315) and the ligand were selected in Discovery Studio maintaining the geometry coordination. Only the hydroxamate moiety of the ligand (LC5) was maintained. A charge of -1, (+2 for the Mn²⁺ and -1 for each of the three negatively charged residues (ASP231, LU233, GLU315) and multiplicity 1 were used when preparing the subset for optimization in Gaussian. A split basis set employing an ECP (effective core potential) basis LANL2DZ (Los Alamos National Laboratory 2 double ζ) (Chiodo, Russo, and Sicilia 2006) commonly used for the optimization of transition metals (Neves et al. 2013) and B3LYP/6-31G(d) for the rest of the atoms (C H O N) and using the keyword “pseudo=read” for taking into account the pseudo potentials. Figure 5-2 and 5-3 show the header and footer of the .com input file for Gaussian, used to achieve these ends.

Header of the .com file:

```
%nprocshared=24  
%mem=50GB  
%chk=s_5JAZ_B_Mn_LC51_modified.chk  
# opt b3lyp/3-21g geom=connectivity pseudo= read
```

Figure 5-3 : Header for gaussian input.

Footer used:

```
CHONO
6-31G(d)
****
MnO
LANL2DZ
****

MnO
LANL2DZ
```

Figure 5-4: Footer of the .com input file for Gaussian

The created .com file was submitted for calculation by Gaussian09 on the CHPC using 1 node and 24 cores for the computation. The optimized geometry was finally submitted to Metalizer (Bietz and Rarey 2016; Meyder et al. 2017) for prediction of the coordination geometry.

5.2.1.2 Potential Energy Surface Scan (PES)

A PES was performed for the bonds, angles and dihedrals around the metal. All the bonds implying the manganese were scanned. Due to the high number of possible angles and dihedrals combinations, some of them were selected for scan. This later was performed using Internal Redundant coordinate using the Gaussian 09 keyword *Opt=ModRedundant* was used. Also known as Relaxed PES Scan, in that scheme, the subset geometry is optimized at each step while keeping the scanned parameters constant. The parameters are the bond stretching, bending and torsional movements values. This allows to have energy profile depending only on these maintained constant. Force field parameters were then obtained by fitting the energy profiles to the bonded terms in Equation 3 using the least squares method. Microsoft Excel “solver” module was used for that purpose.

Gaussian header used for PES scan:

```
# opt=modredundant b3lyp/gen geom=connectivity pseudo=read scf=xqc
```

The scan parameters are specified in the footer section Gaussian .com file.

The following parameters were used during the scan.

1. Bonds: 15 steps, step size of 0.01 Angstrom
2. Angles: 15 steps, step size of 1 Degree
3. Dihedrals: 15 steps, step size 1 Degree

Due to time constraint, the derived parameters from the PES for the dihedrals were not included in the implementation in AMBER03 (Duan et al. 2003) force field in GROMACS (Abraham et al. 2015).

5.2.1.3 ONIOM Setup

In the early stages of this project, a hybrid system combining QM/MM (ONIOM) was used in order to derive force field parameters. The high level system included atoms for which force-field parameters were to be refined, while most of the protein was defined with well-known force-field parameters. It was hoped that residue geometries and hence the geometry of the central metal would be more accurately defined using this approach for surface scans. Here we describe the methodology used to set up the ONIOM calculations. The PES scan was performed using the same methodology as in the subset in terms of redundant coordinates.

The chain B of the protein, the metal ion and the only the hydroxamate moiety of fosmidomycin were extracted from the PDB structure (5JAZ). The ONIOM system was set up in Gaussian. The H++ web server was used to protonate the structure. The server returned a charge of +5 for the full protein (excluding the metal ion) at pH 6.5. The metal ion and its three coordinating residues were selected at the high layer and the rest of the system at the low layer. A charge of -1 was set for the high layer (adding charges of its elements Mn +2, Glutamic acid -1, Glutamic acid -1, Aspartic acid -1) and +7 for the full protein (adding the +2 charges for Mn atom to the charge returned by H++). A multiplicity of one (1) was used. The low layer was treated at molecular mechanical level using UFF (Universal Force Field) while the high layer was treated using quantum mechanic using the b3lyp level of theory. The following steps of optimization and PES scan were similarly to the one used for the subset.

5.2.2 Implementation in GROMACS

The derived force field parameters were then implemented in the AMBER03 (Duan et al. 2003) in GROMACS following the following guide of the GROMACS online documentation (http://www.gromacs.org/Documentation/How-tos/Adding_a_Residue_to_a_Force_Field).

Adding a new atom in a force field in GROMACS requires modifications on the parameter files of the chosen force field. The following paragraph summarizes the different required modifications to undertake.

Before any modification to the force field, the directory containing it and the GROMACS file residuetypes.dat (/usr/local/gromacs/share/gromacs/top/amber03.ff/), were first copied to the working directory. Then when running MD and prompted for force field selection, GROMACS proposes both force field options. The new atom Mn was added to the .rtp (aminoacids.rtp) of the force field. As the Mn atom does not require hydrogen, no modification was done to the .hdb file. Similarly, no modification was done to the specbond.dat file to add special connectivity to other residues. The new atom type was added in the atomtypes.atp and the non-bonded parameters in the ffnonbonded.itp file and the bonded parameters in the ffbonded.itp one. Finally, the Mn atom was added to the residuetypes.dat file. The required modifications were done following the parameters files formatting system (see result section 5.3.2 for details).

The values for the Lennard-Jones potentials were obtained from the literature (Babu and Lim 2006). Chapter 2 of the GROMACS manual was used for unit conversion. The following table summarizes the different units used and their conversion.

Table 5-1: GROMACS units used and their conversion

Parameters	Unit from PSES fitting	Conversion factor	GROMACS Unit
Energies	Kcal mol ⁻¹	1 kcal mol ⁻¹ = 4.184 kJ mol ⁻¹	kJ mol ⁻¹
Force	kcal mol ⁻¹ Å ⁻²	1 nm ⁻² = 100 Å ⁻² 1 kcal mol ⁻¹ = 4.184 kJ mol ⁻¹	kJ mol ⁻¹ nm ⁻¹
Distance	Å	1 nm = 10 Å	nm
Angles	Degree		Degree

5.2.3 Steps of Molecular dynamics

The different systems were prepared using the same procedure.

The following software tools were used:

1. GROMACS 5.1.4 (GRONingen MACHine for Chemical Simulations) (Abraham et al. 2015) on a local machine for preliminary system preparation and GROMACS version 2016.1 on CHPC to run the simulation and/or for energy minimization and NVT and NPT equilibrations.
2. Babel: the tool offers a solution for interconversion of a large variety of chemical file formats (O'Boyle et al. 2011).
3. Discovery Studio 2016 (Biovia, San Diego, CA).
4. VMD Visual Molecular Dynamics 1.3.9: a visualization and analysis program for molecular structures, especially biomolecules (Humphrey, Dalke, and Schulten 1996).
5. Acyppe (AnteChamber PYthon Parser interfacE) (Sousa da Silva and Vranken 2012; Batista et al. 2006) was used to generate topology for the ligands. The resulting pdbqt format from docking were first converted to .pdb format using Babel (O'Boyle et al. 2011) to be used with Acyppe.

A MD simulation was conducted to validate the developed parameters for manganese and to investigate the docking complexes which also included the cofactor (NADPH). Since docking of the cofactor in the previous chapter did not result in the correct pose, and since the crystal structure used in this study (5JAZ) did not include the cofactor, this necessitated the copying of the cofactor pose from the PDB structure 3AU9 after molecular overlay of the two structures using Discovery Studio (Biovia, San Diego, CA). Both structures are DXR in a closed conformation. Only a monomer (Chain B) of the protein was used and all water molecules were removed. Acyppe (Sousa da Silva and Vranken 2012; Batista et al. 2006) was used to generate topologies for the cofactor and the ligands. The generated topologies were adapted to GROMACS ones by renaming the atom types to compliant GROMACS ones.

The adapted AMBER03 (Duan et al. 2003) force field, with the parameters for the manganese atom was used for all the simulations. The different systems simulated are the following:

1. Force field validation: Protein (5JAZ) + Mn (the ligand was not included in the simulation).
2. Hit studies: Protein + NADPH + Mn + Ligand (Hits from docking: SANC00152, SANC00236, SANC00339, SANC00438, SANC00570).

After preparing the topologies, the simulation box and the periodic boundary conditions were defined using the command *editconf*. A cubic box was chosen to run the simulation in with distance between the solute and the box set to 1.0 nm. The Simple Point Charge (spc216) water was used as solvent model. A concentration of 0.15 M (Na⁺ (sodium) and Cl⁻ (chloride) ions) was used and the system was neutralized using the *-neutral* option.

The system energy was minimized using a steepest descent method with a maximum force set at <1000.0 kJ/mol/nm and a maximum number of steps of 50000.

All the preparations were done on a local machine (using GROMACS 5.1.4) up to minimization step. The next steps were alternatively run on a remote machine at CHPC (Center for High Performance Computing) depending on the availability of resources. GROMACS version 2016.1 was used. The simulation was submitted using a PBS (portable batch system) file (an example of job file in appendix H) on CHPC, using 10 nodes with 24 cores each (total cores of 240) using a *walltime=48:00:00* and a normal queue.

The temperature was set to 300K and the pressure at 1 atmosphere. A simulation of 100 nanoseconds was run by setting 50.000.000 steps in the *.mdp* file, and *dt* of 0.002 (ps) time step for integration with a leap-frog algorithm leap-frog integrator.

5.2.3.1 Analysis

For the force field parameters, the protocol for validation consisted in first tracking the stability of the protein through its RMSD and then the stability of the coordination sphere. The GROMACS distance module was used to monitor the distance between the Mn atom and the bound oxygen atoms.

For the hit-protein complexes (Protein-Ligand-Cofactor), the analysis focused on their stability. Compounds with poor binding can engender unstable trajectories and vice-versa. They may also leave the protein binding site. An unstable RMSD over time is a good indicator. On the other hand stable and specific interactions, for example, maintaining stable hydrogen bonds between the ligand and the protein is a good indicator of good binding.

Different properties of the systems temperature, pressure, and the potential energy before MD run will be analysed to assure systems' qualities.

After simulation, the *trjconv* was used to correct for periodicity, centering the protein in the simulation box and avoiding jumps across box sides. The root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF) of the backbone atoms and the radius of gyration (Rg) were used to study the protein stability. The interactions between the ligands and the protein were studied through the number of hydrogen bonds.

5.3 Results and Discussion

5.3.1 Metal parameterization

5.3.1.1 Initial structure preparation and geometry optimization

The reported geometry is described as distorted bipyramidal trigonal. It has been argued that the increased flexibility of the metal–ligand bonds showed by Mn^{2+} more easily accommodates distortions in coordination geometry at the transition state (Murkin, Manning, and Kholodar 2014; Chofo 2016).

Using Gaussian09, the geometry of the subset was optimized. The optimization process can be tracked in the .log file through the line containing the keyword “SCF”. The energy value is given in atomic unit (Hartree) for example:

```
“SCF Done: E(RB3LYP) = -1815.14444090 A.U. after 17 cycles”.
```

The final optimized subset had an energy of -1815.14 Hartree after 53 steps of optimization. Visually, the optimized structure tended toward an octahedral geometry (see Figure 5-5). Metalizer predicted a square pyramidal geometry would have the best score. Although, the octahedral geometry also has the same lowest angle RMSD of 8.12 as the square pyramidal one (See complete table of all possible predicted geometries in appendix I). Metalizer uses a weighted score that combines the RMSD, the number of free sites and the overlap of presumed free sites with any heavy atom in the first coordination sphere or with any non-water heavy atom for geometry prediction (Bietz and Rarey 2016).

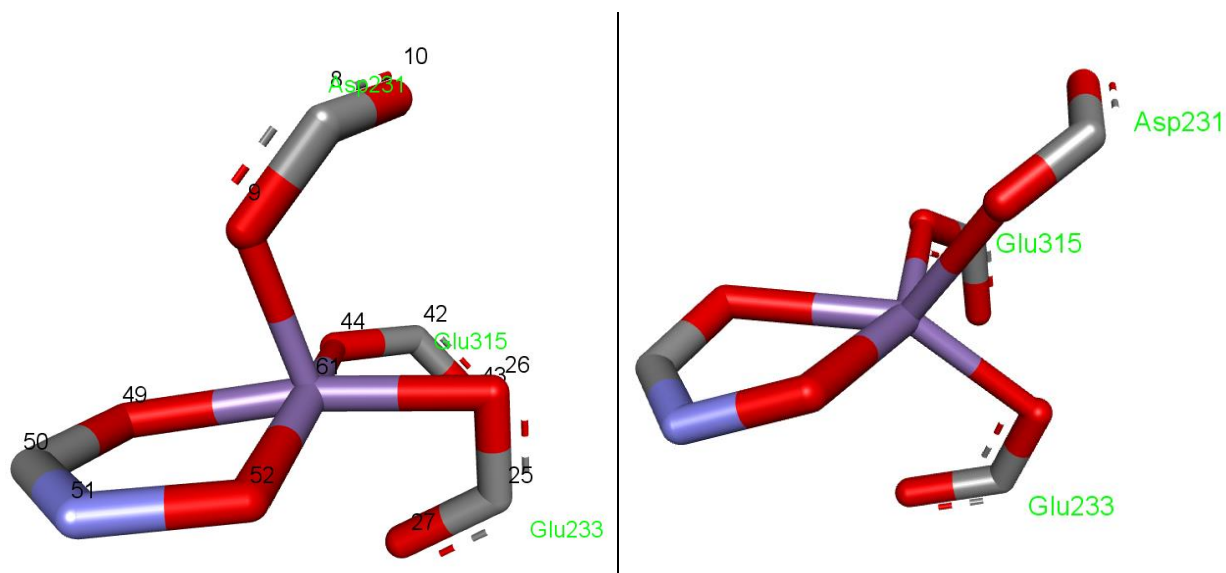


Figure 5-5: Left optimized subset. Right geometry in the crystal structure (5JAZ chain B). The backbones of the residues were removed for clarity.

The crystal structure has a distorted bipyramidal trigonal (see Figure 5-5). This is consistent with the different geometries observed in the same series of crystal structures (5JBI, 5JC1, 5JMP, 5JMW, 5JO0) also having a Mn atom in their active sites with the ligand involved in the coordination. Coordinations implying the Mn atom and a bidentate ligand in DXR active site have often reported to be octahedral especially in *E.coli* (Mac Sweeney et al. 2005; Yajima et al. 2002; Steinbacher et al. 2003). In these cases, a water molecule fills the 6th coordination position. In 5JAZ, no water molecule was found in a radius of 3 angstroms. A water molecule was within the radius of 4 angstroms (3.68 angstroms, see Figure 5-6) but still too far to occupy the 6th coordination position. This position in the close-to octahedral geometry of the optimized subset is occupied by oxygen 27 (GLU233) (see Figure 5-5).

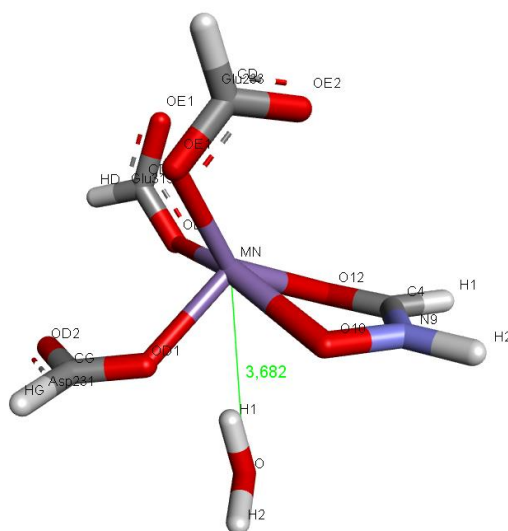


Figure 5-6: Water molecule and Mn coordination in the crystal structure.

Table 5-2: Bond lengths in initial crystal and optimized structure

Bonds					Length (Å)		
Atom Number In subset	Residue	Manganese	Atom Number In subset	Simplified notation	Initial	Optimized	Initial – optimized
9	ASP231	MN	61	9 - 61	2.13	1.91	0.22
26	GLU233	MN	61	26 - 61	2.1	2.01	0.09
27	GLU233	MN	61	27 - 61	2.74	2.08	0.66
44	GLU315	MN	61	44 - 61	2.13	1.85	0.28
49	LC5501	MN	61	49 - 61	2.17	2.07	0.1
52	LC5501	MN	61	52 - 61	2.17	1.89	0.28

Table 5-3: Angles in initial crystal and optimized structure

Angles						Angles (degree)			
Residues	Atom Number In subset	Manganese (angle vertex)		Residues	Atom Number In subset	Simplified Notation	Crystal structure	Optimized	Cryst-Opt
GLU233	26	MN	61	ASP231	9	26 - 61 - 9	95.2	101.3	-6.1
GLU315	44	MN	61	ASP231	9	44 - 61 - 9	96.8	99.09	-2.29
LC5501	49	MN	61	ASP231	9	49 - 61 - 9	122.4	84.62	37.78
LC5501	52	MN	61	ASP231	9	52 - 61 - 9	82.5	90.5	-8
GLU315	44	MN	61	GLU233	26	44 - 61 - 26	96.4	95.7	0.7
LC5501	52	MN	61	GLU233	26	52 - 61 - 26	107.6	93.31	14.29
LC5501	49	MN	61	GLU315	44	49 - 61 - 44	85.6	89.16	-3.56
LC5501	49	MN	61	LC5501	52	49 - 61 - 52	74.9	80.58	-5.68

Comparing bond lengths in the initial crystal structure with the optimized one (see Table 5-2), values are not significantly different. However, one oxygen on GLU233 (atom number 27 see Figure 5-5) has a significant decrease of 0.66 Å in its bond length with the manganese. The atom is not part of the trigonal bipyramidal geometry of the initial structure but gets closer in the close-to octahedral geometry as it is part of the coordination. Comparing the angles, only 52-61-26 and 49-61-9 showed significant differences getting compressed by 14.29° and 37.78° respectively (see

Table 5-3 highlighted in brown). Again, these changes are explained by the change in geometry.

The larger free degree of movement of the coordinating residues during optimization can explain the observed closed-to octahedral geometry. The sixth coordination position is occupied by the second oxygen atom of GLU233 (atom number 27).

5.3.1.2 PES scans

After geometry optimization, the force field parameters were derived through PES scan. Here we present the graphs obtained after least square error fitting for the PES data for bond 44-61.

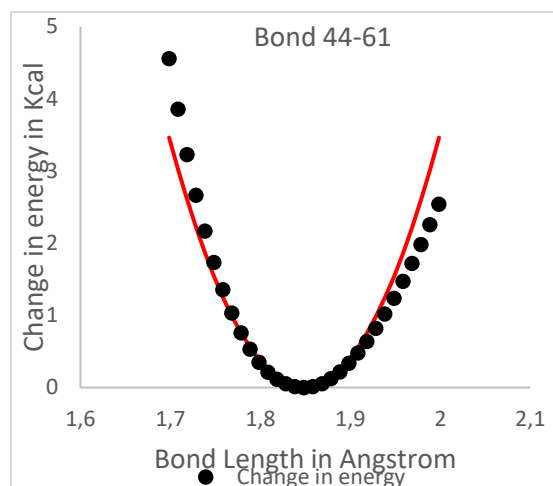


Figure 5-7: Least square error fitting for PES data for bond 44-61.

The change in energy represents the difference between each energy point with the minimum energy point on the PES. In this case, the bond length corresponding to the minimum energy was 1.85 Å. The continuous curve represents the harmonic potential model used for fitting. In this case the bonded term for bond 44-61. The fitting gave a force constant of $153.89 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$. The sum of errors was 5.03, the 2nd highest value among all scanned bonds.

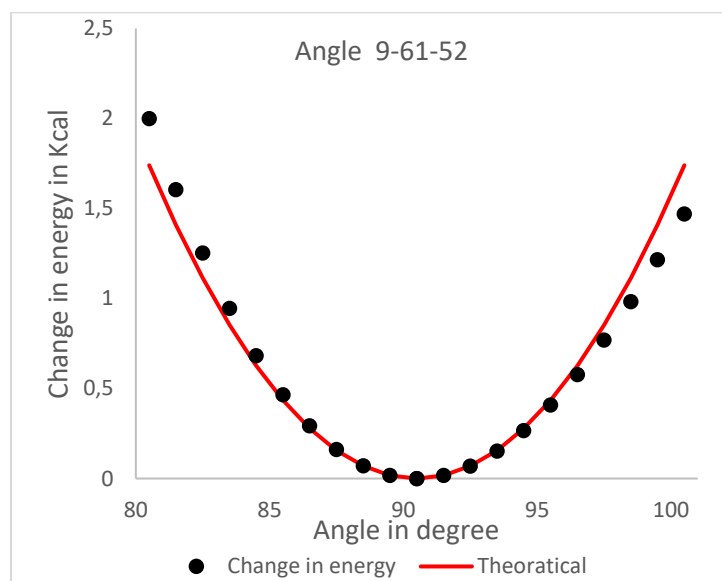


Figure 5-8: Least square error fitting for PES data from Angle 9-61-52.

PES scan data points for angle 9-61-52 achieved one of the best fitting to the harmonic potential (see Figure 5-8), achieving a sum error of 0.2755 with all the data points included in the error calculation. In total 7 angles were scanned. The graphs for fitting the data point for all angles are in appendix J.

Table 5-4: Force parameters derived from least-square fitting of PES data.

Bonds	Force constant / equilibrium bond length	
	Kb	b0
	Kcal mol ⁻¹ Å ⁻²	Å
9-61	110.51	1.91
26-61	84.26	2.01
27-61	47.78	2.08
44-61	153.89	1.85
49-61	62.47	2.07
52-61	112.59	1.89

Angles	Force constant / equilibrium bond length	
	cth	th0
	Kcal mol ⁻¹ degree ⁻²	°
9-61-52	0.0173	90.50
9-61-49	0.0144	84.64
9-61-44	0.0144	84.64
9-61-26	0.0144	84.64
26-61-44	0.0144	84.64
26-61-52	0.0229	84.64
49-61-52	0.0148	84.77

Kb: bond force constant. b0: equilibrium bond distance.
Cth: angle force constant. th0: equilibrium angle.

The bonds' force constants range between 153.89 kcal mol⁻¹ Å⁻² (bond 44-61) and 47.78 kcal mol⁻¹ Å⁻² (bond 27-61), having respectively the lowest and highest b0. However, the b0 values were not significantly different, ranging from 1.85 Å to 2.08 Å (see Table 5-4).

The different equilibrium angles were around 84.64 degrees except for 9-61-52 (see Table 5-4) which present a slightly different value of 90.50 degrees. This similarity can be explained by the symmetry in the octahedral geometry. This same similarity is observed for the force constants which is about 0.0144 Kcal mol⁻¹ degree⁻². Again, here 9-61-52 has a slightly different value for its force constant 0.0173 Kcal mol⁻¹ degree⁻². Angle 26-61-52 presents a significantly different force constant of 0.0229 Kcal mol⁻¹ degree², two (2) times higher than the other angles. This could be explained that both oxygens implied in this angle are from molecules having their second oxygens implied in the coordination. Indeed, atom 52 belongs to the ligand and 26 to GLU233. Both of these molecules have two oxygens implied in the coordination. As a result, they can restrain the movement of atoms 52 and 26 resulting in a higher force constant.

Compared to literature, Neves et al. determined parameters for 12 models of manganese metal centres for the AMBER force field using a similar procedure. A model from *Mandelate racemase* (PDB ID: 2MNR) is similar to present subset, presenting a Mn coordination with a distorted trigonal geometry implying two glutamic acid, one aspartic acid and a bidentate ligand. However, the optimized model kept the same similar geometry while the subset here tends toward an octahedral one. They found equilibrium bond lengths in the interval [2.00; 2.45] Å for the 12 models, while a range of [1.85 ; 2.08] Å was observed in this study. Shorter b0 were thus observed. Their bond force constants were in the range 60–80 kcal mol⁻¹ Å⁻² while this subset presents a range of 153.89 kcal mol⁻¹ Å⁻² to 47.78 kcal mol⁻¹ Å⁻². However, similarly low force constant were observed for the *Mandelate racemase* with a range 33.3 kcal mol⁻¹ Å⁻² to 49.5 kcal mol⁻¹ Å⁻². It is noteworthy that a semi-flexible approach in which the non-scanned ligands and the backbone of the scanned ones were frozen during the optimization and PES (Neves et al. 2013).

Geometry optimization and PES are tedious processes. Indeed, many optimizations and PES scan fail due to multiple types of error in Gaussian09. Syntax errors often occur when setting up

the scan in the .com files. A second common error is the “Convergence failure” occurring when SCF (self-consistent field) procedure fails to converge. Different convergence procedures using the “SCF” keyword can then be used. Mode QC can be used for a quadratically convergent SCF convergence and XQC adds an extra “SCF=QC” step if first-order SCF has not converged. Gaussian also has a “tight” optimization mode. In that scheme, the RMS force criterion is set to lower value (1.10^{-5}) from its default value (3.10^{-4}). The other convergence conditions are also scaled accordingly. Another difficulty in optimization or PES scan is the “Linear angle in Tors” error. During optimization, atoms of the dihedral angles may be aligned, blocking the optimization process. In that case, using optimization mode in cartesian using the keyword “opt = cartesian” can be helpful. In any case, all of these problems were solved for successful results.

Also, one needs to carefully adjust the value for the number of steps and the step size. This is done in order to only scan the bottom of the Morse potential. The reason being that only near the equilibrium bond length, the harmonic potential approximates well the Morse potential (Lewars 2016). Moving away from the equilibrium bond distance, for example, for bond displacements greater than 10%, the harmonic potential becomes a poor approximation (González 2011). This is well illustrated in Figure 5-9 below. A step size of 0.01 yield better fitting with a sum of errors of 0.88 while using a step size of 0.05 resulted in a sum of error of 1162.26. A similar result could be attained by also reducing the number of steps. During the series of scans, multiple values for these parameters were then tried in order to scan only around the equilibrium bonds and angles parameters for better fitting.

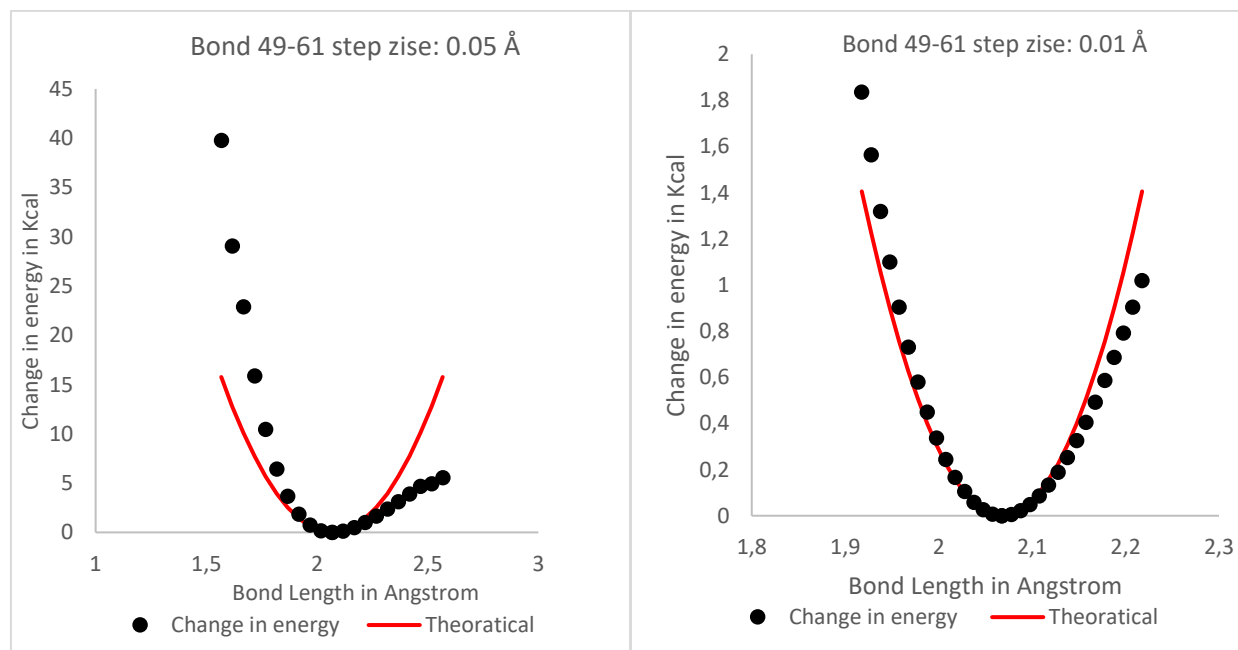


Figure 5-9: Implication of the step size in PES scan.

Considering this difference of the two models, some data points obtained from the PES scan may need to be adjusted. The most important points are obtained around the equilibrium bond lengths and angles and should always, ideally be kept for the fitting. Points on the extremities (far from the equilibrium bond, angles) often increase the sum of error during the least square

fit. This may result in incorrect parameter values. Thus, manual adjustments by removing these points in the extremities for a better fitting may be required.

5.3.1.3 ONIOM

As mentioned earlier in the methodology (section 5.2.1.3), an ONIOM system was first used to parameterize the metal ion. The system was optimized using molecular mechanics for the low layer and quantum mechanics for the high layer. It is interesting to note here that the geometry of the coordination in the ONIOM system was still bipyramidal trigonal as in the crystal structure (see appendix KK). This supports the hypothesis that the change to octahedral geometry is due to the “free-flying” residues in the subset. After optimization, PES scans were conducted on the metal center, scanning bonds, angles and dihedrals implying the metal ion as with the subset. Scans’ data were challenging for least square fitting to the harmonic potentials. Indeed, unexpected energy variations were observed in the resulting graphs, see Figure 5-10 below. Similar observations were made on the different angles, bonds and dihedral scanned.

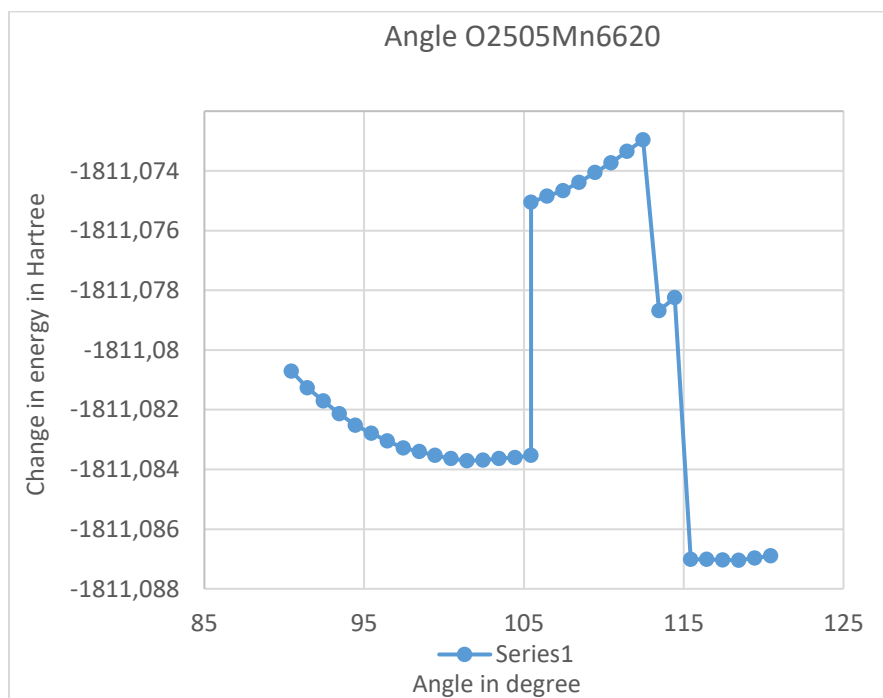


Figure 5-10: Example of energy variation during PES scan on the ONIOM system.

Due to these variations, these graphs were difficult to fit to the harmonic potential model. They may be explained by twists in the structure. Gaussian optimizes the system energy at each step of the scan by changing the system geometry. All successful optimizations locate a stationary point, although not always the one that was intended. The nearest stationary point from the initial geometry is the point often reach in geometry minimization (Ramachandran, Deepa, and Namboori 2008). As the ONIOM system is large, even the small step size used in PES scan may engender a significant structural change in the system. Consequently, the system energy can be greatly affected as seen in the energy graphs. The structure may thus fall into other local energy minima or have an increase in energy.

5.3.2 Implementation in GROMACS

After fitting, the obtained parameters were first converted to GROMACS compliant units.

Table 5-5: Parameters converted to GROMACS compliant units.

Bonds	PES		GROMACS		Angles	PES		GROMACS	
	Kb	b0	Kb	b0		cth	th0	cth	th0
	kcal mol ⁻¹ Å ⁻²	Å	kJ mol ⁻¹ nm ⁻¹	nm		kcal mol ⁻¹ degree ⁻²	degree	kJ mol ⁻¹ rad ⁻²	degree
9-61	110.51	1.91	92471.06	0.19	9-61-52	0.0173	90.50	477.8973	90.50
26-61	84.26	2.01	70504.61	0.20	9-61-49	0.0144	84.64	395.8788	84.64
27-61	47.78	2.08	39985.89	0.21	9-61-44	0.0144	84.64	395.8788	84.64
44-61	153.89	1.85	128777.11	0.18	9-61-26	0.0144	84.64	395.8788	84.64
49-61	62.47	2.07	52276.69	0.21	26-61-44	0.0144	84.64	395.8788	84.64
52-61	112.59	1.89	94215.36	0.19	26-61-52	0.0229	84.64	629.5327	84.64
					49-61-52	0.0148	84.77	407.8382	84.77

ffbonded.itp

[bondtypes]

; i j func b0 kb

Mn O2 1 0.20753 39985.8 ; Adding parameter for Mn

[angletypes]

; i j k func th0 cth

O2 Mn O2 1 84.6487 395.878 ; Adding angle parameters for Mn

ffnonbonded.itp

[atomtypes]

; name at.num mass charge ptype sigma epsilon

Mn 25 54.94 0.0000 A 1.32940e-01 0.16736e+00 ;

aminoacids.rtp

[MN]

[atoms]

MN Mn 2.00000 1

atomtypes.atp

Mn 54.94 ; manganese

residuetypes.dat

Mn Ion

MN Ion

Figure 5-11: Force field parameters files modified and their modifications.

Figure 5 10 illustrates the different parameters files modified and the added values. When implementing the parameters, there is only one oxygen type for the carboxyl group. Indeed, in the AMBER03 force field, the following oxygen types exist in the atomtypes.atp with 16.0 being the mass.

atomtypes.atp		
O	16.00000	; carbonyl group oxygen
OW	16.00000	; oxygen in TIP3P water
OH	16.00000	; oxygen in hydroxyl group
OS	16.00000	; ether and ester oxygen
O2	16.00000	; carboxyl and phosphate group oxygen

Both oxygens of the carboxylate group (bold line) have the same atom type (O2). In other words, they will be treated with the same parameters. As a result, the different oxygen atoms coordinating the manganese will be treated with the same set of unique parameters (bond and angles force constants...). However, the parameters derived from the bonds PES scans after fitting showed different force constants and equilibrium lengths for the oxygens (see Table 5-4: Force parameters derived from least-square fitting of PES data.). But only one parameter could be used. The values for oxygen 26 from GLU233 were chosen to be implemented.

Other approaches may have been to calculate an average value of the parameters. Elsewhere, one could also introduce new atom type to describe the two oxygens of the carboxylate group. This is, for example, the case in PDB files format in which OE1 and OE2 are used to describe the two oxygens for the glutamic acid and OD1 and OD2 to describe the ones on the aspartic acid.

The parameters were included using the same oxygen (O2) of the force field. After their implementation, the parameters were validated through molecular dynamics.

5.3.3 Molecular Dynamics

5.3.3.1 Monitoring system stability before MD run

After setting up the different MD systems, their solvation and neutralization, the systems' energies minimized. The following table summarizes the values of the potential energies and maximum force on atom, obtained after energy minimization.

Table 5-6: Potential energies and maximum forces attained during minimization.

Systems	Properties at the end Minimization	
	E_{pot} (kJ/mol)	Maximum force (kJ/mol)
5JAZ + Mn	-1.2999762e+06	9.0311633e+02
SANC00152	-1.3123954e+06	8.7838123e+02
SANC00236	-1.3095252e+06	8.5790820e+02
SANC00339	-1.3033710e+06	9.9254956e+02
SANC00438	-1.3105418e+06	8.3803802e+02
SANC00570	-1.3115348e+06	9.3925336e+02

During minimization, the energies of the systems dropped at some local minimum. All systems showed a negative E_{pot} in the order of 10^{-5} and 10^{-6} and all the maximum forces were inferior to 10.0 kJ/mol (see Table 5-6). The systems were hence minimized ensuring an optimization of the atoms with the force field parameters. These values were thus satisfactory to proceed to dynamics with NVT and NPT equilibration.

Temperature during NVT equilibration

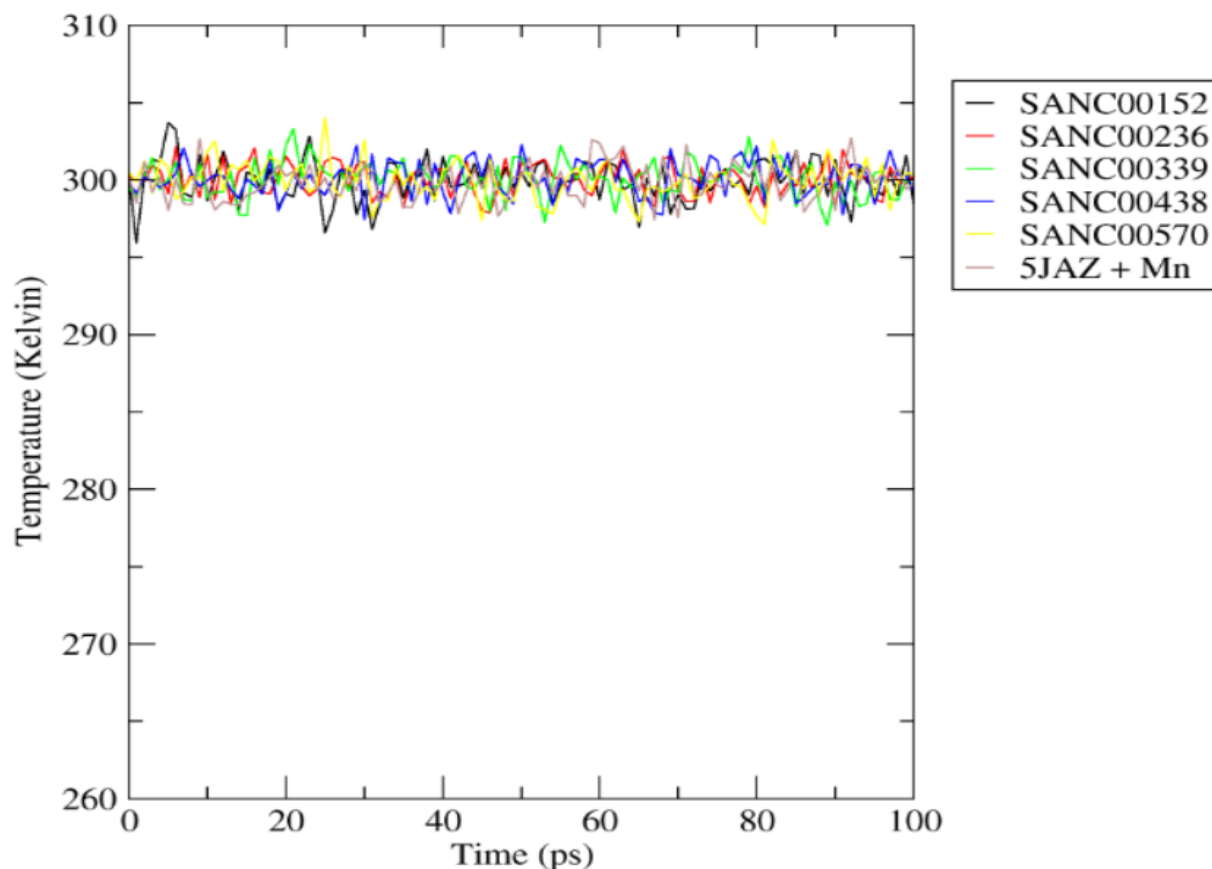


Figure 5-12: Temperature variation during NVT equilibration. The legends display the SANCDB ID of the different ligands in the different systems.

During NVT equilibration, the temperatures of the different systems stabilize around the target value of 300 k (see Figure 5-12). The systems quickly attain that value and then show little fluctuations around it during the remaining picoseconds of the equilibration. The temperatures oscillate between approximately 296 Kelvin (22.85 Celsius) and 304 Kelvin (30.85 Celsius).

Pressure during NPT equilibration

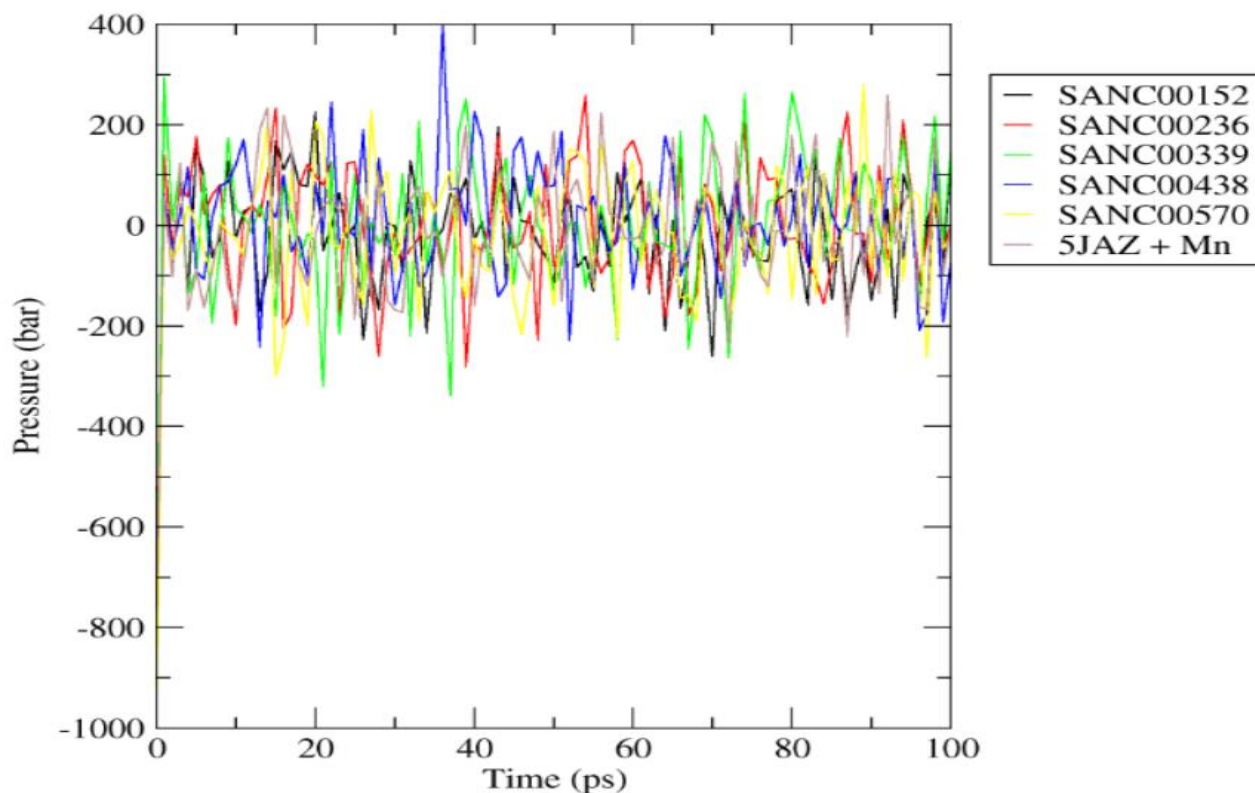


Figure 5-13: Pressure variation during NPT equilibration. The legends display the SANCDB ID of the different ligands in the different systems.

The pressures of the different systems quickly reach the set value of 1.0 bar and then fluctuate around it during NPT equilibration (see Figure 5-13). The pressure is hence equilibrated around that value. After temperature and pressure equilibration, the system is now ready for dynamics.

5.3.3.2 Force field parameters validation

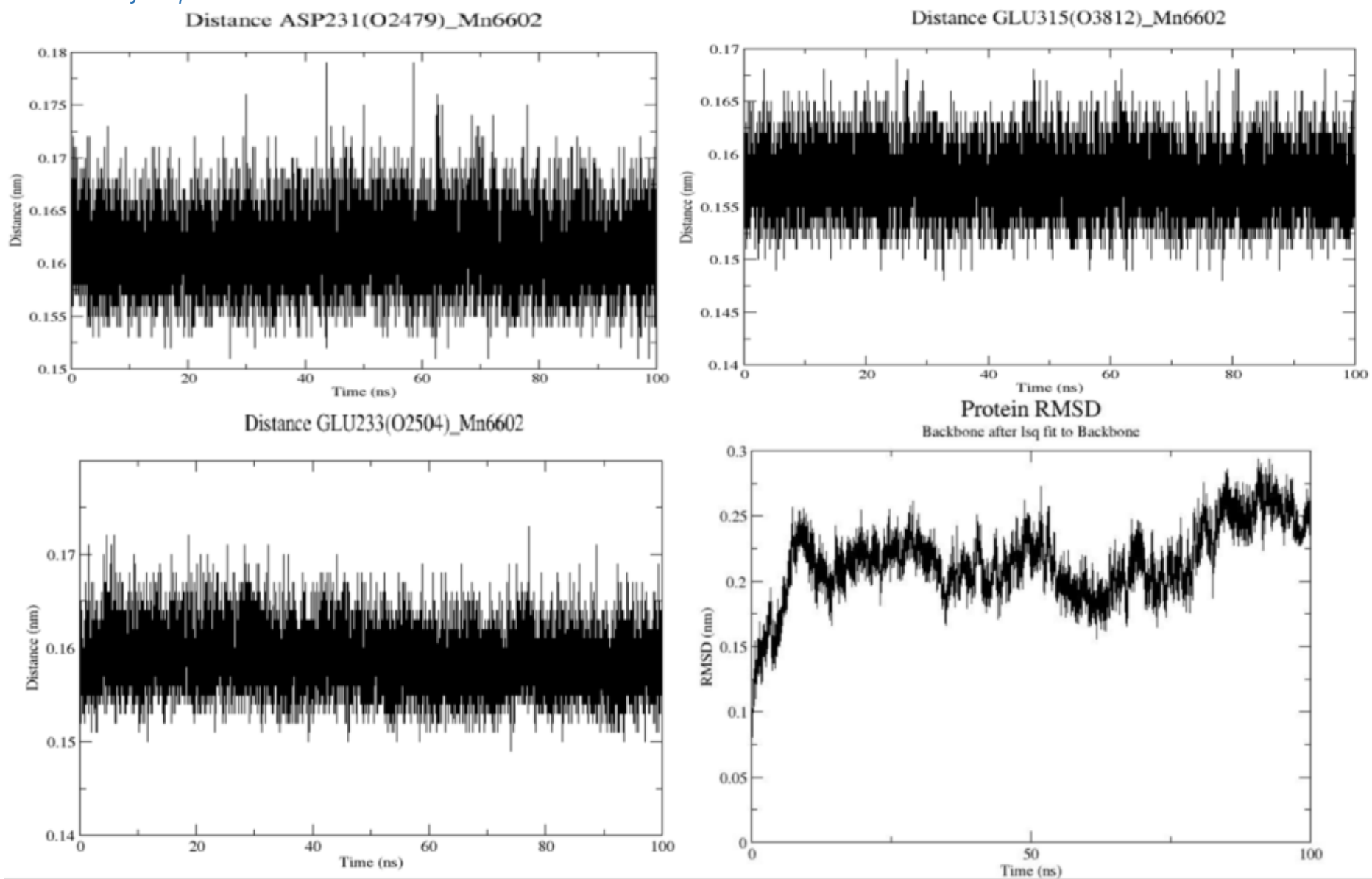


Figure 5-14: Bond distances and protein RMSD for force field parameters validation.

The RMSD of the protein backbone converged to around 0.225 nm at around 10 ns of simulation. The protein stabilizes with an RMSD around that value during the majority of the simulation. However, a little shift to 0.275 nm is visible at about 75 ns of simulation (see Figure 5-14). This value is maintained during the last nanoseconds. The protein, thus, is fairly stable during with an RMSD inferior to 0.3 nm during the 100 ns.

The bond distances between manganese and the oxygen atoms were also measured to track the stability of the metal center (see Figure 5-14). A first observation is that the manganese was coordinated by three oxygens, one from each coordinating residues (GLU233, ASP231 and GLU315). This was confirmed by visualizing the subset in the last frame of simulation. As mentioned in the methodology (section 5.2.3), the ligand was not included in the simulation. The three oxygens are in the same range of bond length, 1.5 Å to 1.7 Å. As indicated earlier (see section 5.3.2), the different oxygen atoms coordinating the manganese will be treated with the same set of unique parameters. This result of similar behaviour was thus expected. Also the bond distances remain within acceptable bounds. All the bond distances are in the range 1.5 Å to 1.7 Å during the entire simulation showing very small fluctuations.

The observed distances during simulation for parameter validation for the metal ion are significantly different from the ones in the crystal structure, but also from the ones observed during PES. Indeed, bond distances during simulation (see Figure 5-14) are in the range 1.5 Å to 1.7 Å which is different from the range from the obtained parameters 1.85 Å to 2.08 Å.

A plausible explanation can be the absence of ligand in this system. Indeed, both subset and crystal structure have a ligand. Nonetheless, the geometry of the optimized subset was also different from the one of the metal center in the crystal structure. This latter was more in accord with the optimized metal center in the ONIOM system. The ONIOM is thus definitely more suitable to derive force field parameters.

Another explanation can be the low force constant of the bond. Indeed the lowest force constant ($39985.89 \text{ kJ mol}^{-1} \text{ nm}^{-1}$) for bond Mn-O (see Table 5-4) of the different force constants obtained was introduced in the force field. This value showed to be 10 times lower than other bond force constants in the force field which are in the order of 10^6 while the force constants obtained were in the order of 10^5 . This significant difference in the force constants may explain the compression of the oxygen atoms on the manganese resulting in the shorter bond distances observed in simulation.

5.3.3.3 Complex 5JAZ-SANC00152 during simulation

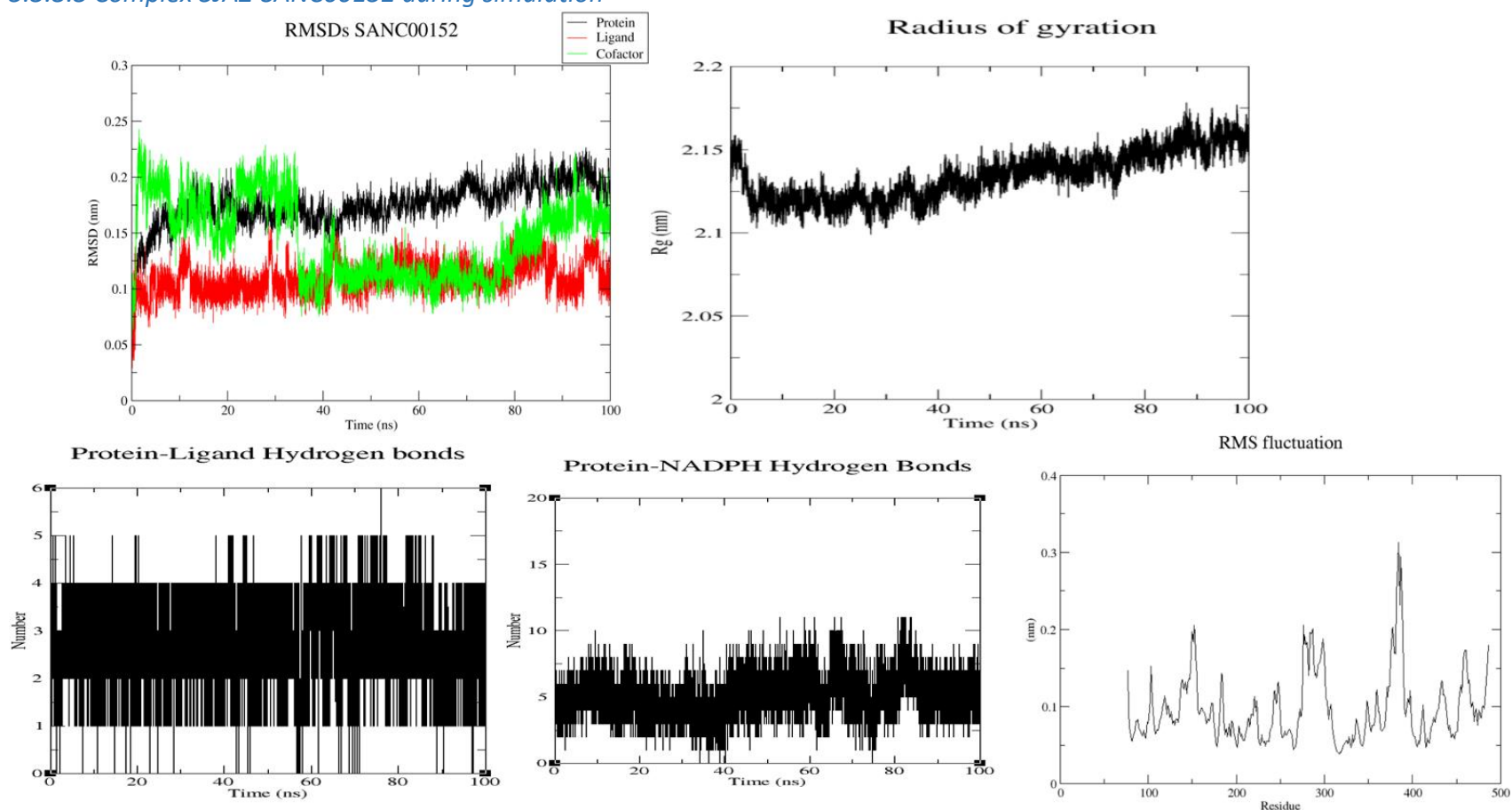


Figure 5-15: Complex 5JAZ-SANC00152 during simulation

The protein RMSD converged to around 0.17 nm (see Figure 5-15). The structure showed very little fluctuations around that value, indicating its stability. The system compactness (described in the radius of gyration) seems to decrease in the first steps of the simulation (first 5 ns) but then starts increasing in the remaining part of the simulation. The system is in increasing expansion from around 2.11 nm to 2.17 nm. This contrasts a little with the protein RMSD curve. A longer simulation would certainly inform more in that expansion. About, the hydrogen bonds, the compound showed 3 to 4 bonds during the majority of the simulation. This is in accord with the ligand RMSD which remains also stable at around 0.1 nm. The cofactor showed some fluctuations. Indeed, we can see a shift in its RMSD at around 35 ns but also in the first 2 ns of simulation. This may indicate a rearrangement of the molecule in the protein. The change can also be linked to the decrease in the number of hydrogen bonds at about 37 ns.

5.3.3.4 Complex 5JAZ-SANC00236 during simulation

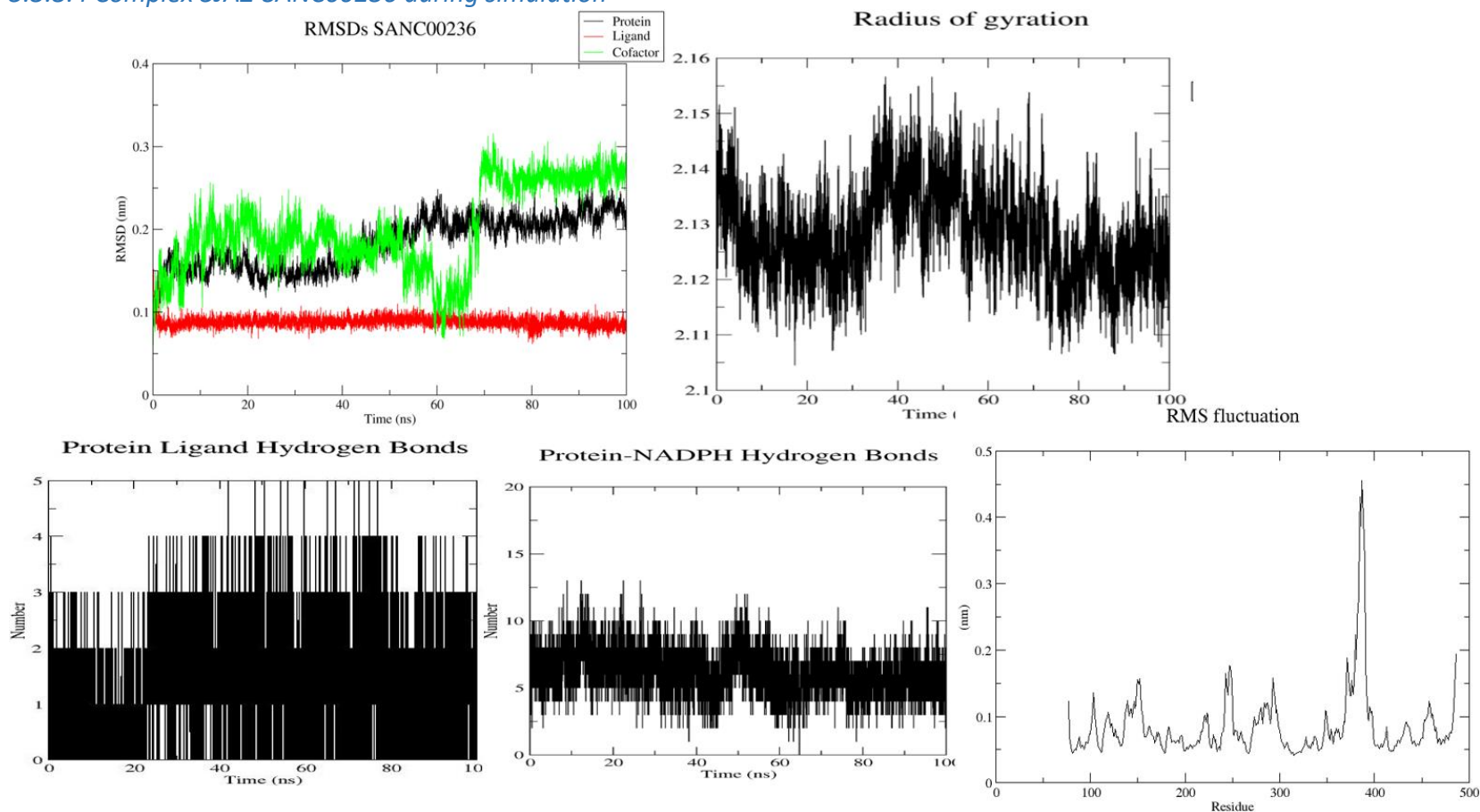


Figure 5-16: Complex 5JAZ-SANC00236 during simulation

The RMSD showed a little shift at about 50 ns. The system then levelled off to around 2.1 nm, during the 2nd part (last 50 ns) of simulation (see Figure 5-16). Concerning its compactness, the system radius of gyration varies between the minimum 2.11 nm and maximum 2.15 nm indicating a stable system. Though we can note a little augmentation in the Rg at 35 ns. This change does not coincide with the same timeframe with the one of the RMSD plot. The ligand showed a very stable RMSD. Indeed, its value showed very little fluctuations near 0.1 nm. This can also be related to the number of hydrogen bonds with the protein. The ligand is consistently keeping at least two (2) hbonds with this later. Finally, the cofactor showed a major change in RMSD from 0.1 nm to 0.3 nm at 70 ns. The molecule then stabilizes at 0.3 nm. The protein-NADPH hbonds plot consistently showed a relatively high number of bonds during the entire simulation making it difficult to explain the change in its RMSD.

5.3.3.5 Complex 5JAZ-SANC00339 during simulation

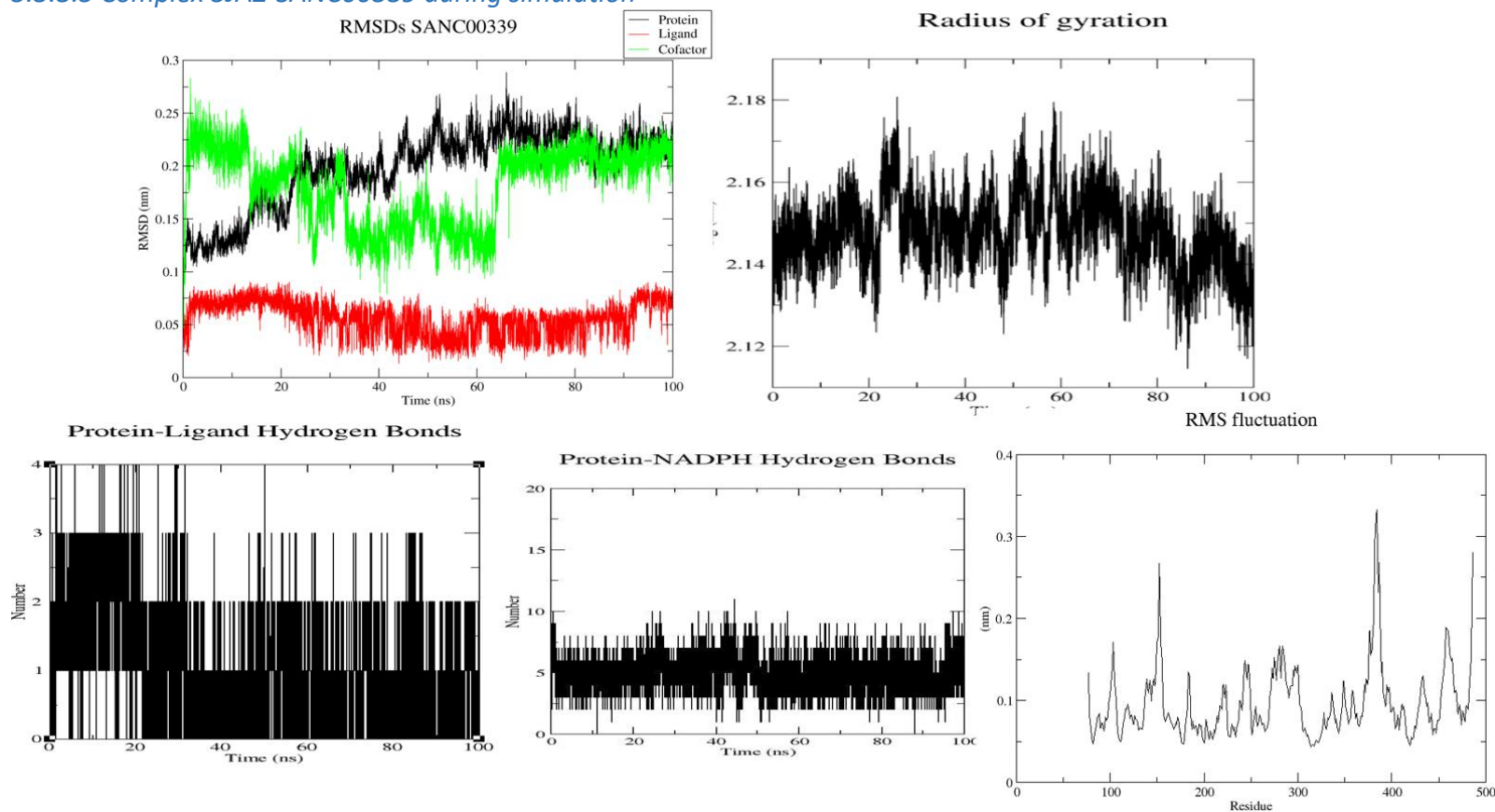


Figure 5-17: Complex 5JAZ-SANC00339 during simulation

The structure RMSD converged at around 0.22 nm during the 2nd half of MD (see Figure 5-17). The protein showed a relatively stable radius of gyration throughout the entire simulation. The Rg fluctuates around 2.15 nm attaining the lowest Rg at 2.12 nm and the highest one at 2.18 nm. The system is getting slightly more compact in the last 10 ns of the simulation where it attains the lowest Rg. This could be explained by a rearrangement of the ligand and/or cofactor in the in protein. Indeed, we can observe a shift in the cofactor RMSD at around 65 ns of simulation which can be linked to the change in the protein Rg. Regarding the number of hbond for the ligand, from the plot, we observed a decrease. Indeed, the molecule showed peaks attaining up to four (4) hydrogens in the early 20 ns of MD. But after, it showed a diminishing trend to 1-2 hbonds towards the end. At the same time, the RMSD also decreases from 0.08 nm to 0.04 nm. Nonetheless, the compound stabilizes with at least one hydrogen bond with the protein during the major part of the simulation (last 60 ns).

5.3.3.6 Complex 5JAZ-SANC00438 during simulation

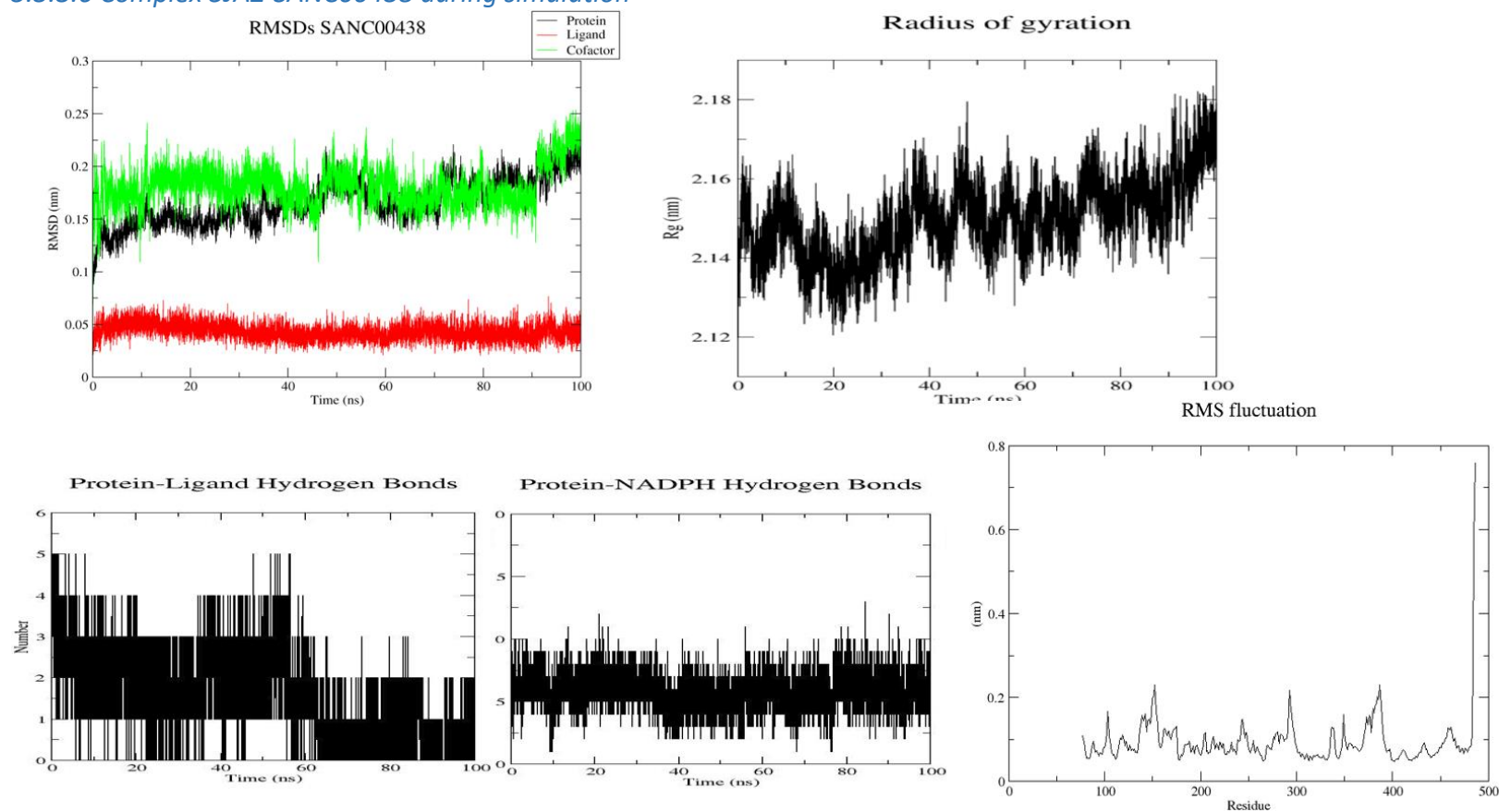


Figure 5-18: Complex 5JAZ-SANC00438 during simulation

The complex with SANC00438 showed an increasing radius of gyration especially after the first 20 ns of simulation. This value changes from 2.13 nm to 2.19 nm. It is noteworthy that the system displays a little compression during the first 20 ns of simulation. Concerning the ligand, it remains very stable during the simulation, with an RMSD fluctuating just between 0.04 nm and 0.06 nm. The compound showed the lowest RMSD value. In terms of its interactions with the protein, the number of hydrogen bond decreases. Showing peaks up to 5 hbonds, this number falls to 1-2 during the last nanoseconds of simulation. This contrasts with the stable ligand RMSD. Other interactions may be involved. However, it is notable that this decrease in hydrogen bonding can be linked to the increase in the protein radius of gyration. About the cofactor, its RMSD remains stable, and the hydrogen bonds plot shows about five (5) hbonds during the simulation. The RMSD shows very low fluctuations around 0.17 nm except for a small shift during the last nanoseconds.

5.3.3.7 Complex 5JAZ-SANC00570 during simulation

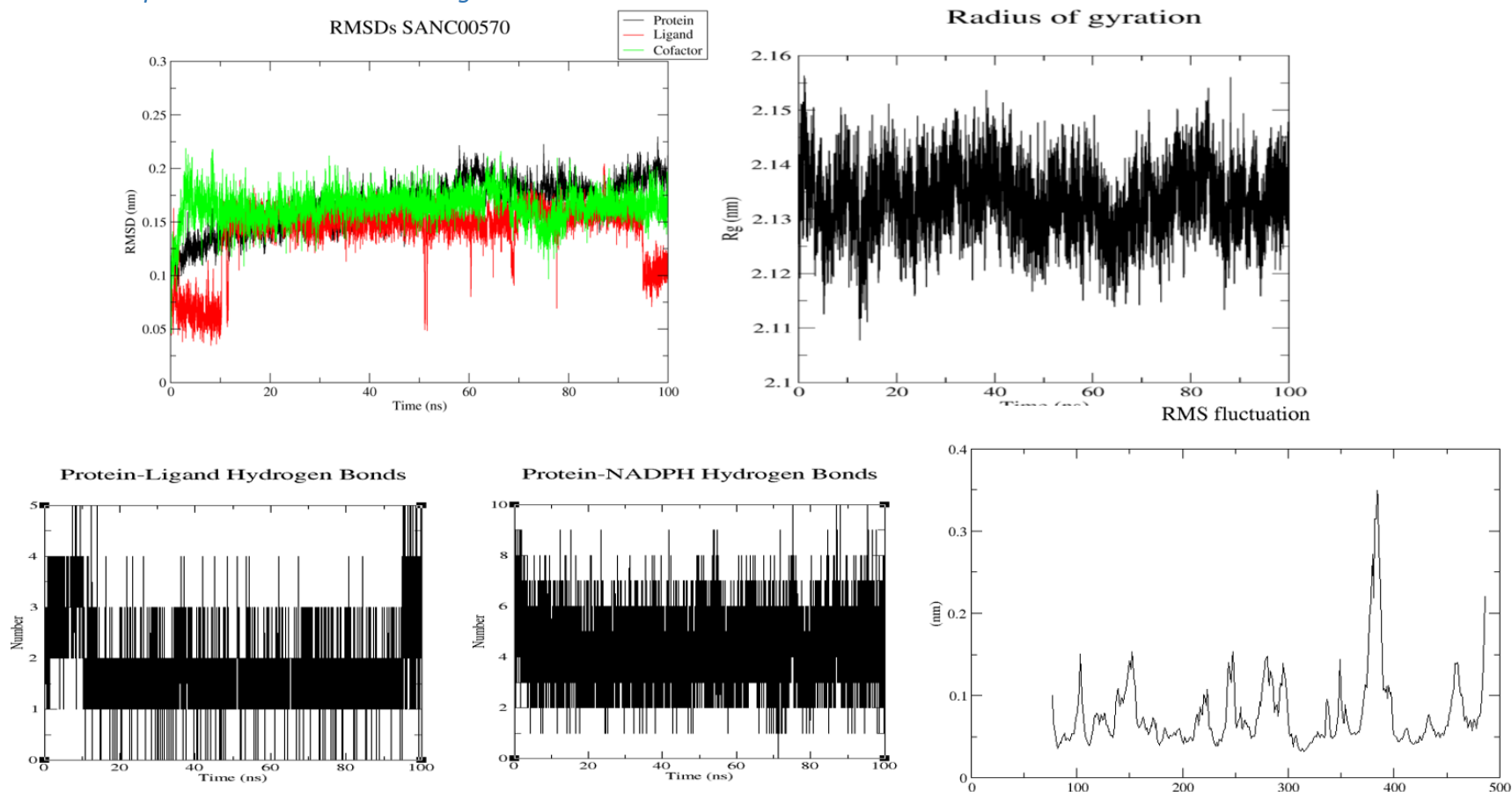


Figure 5-19: Complex 5JAZ-SANC00570 during simulation

The protein with SANC00570 is very stable during the 100 ns (see Figure 5-19). The radius of gyration indeed varies little, fluctuating around 2.13 nm. For the hydrogen bonds, the ligand kept one hydrogen bond with the protein during the majority of the simulation even though it has about four (4) hbonds during the first 10 ns. It is also interesting that we observe more interactions during the last five (5) ns. In fact, we note a number of hydrogen bonds attaining 4 to 5. A longer simulation could have helped to explore more. The pattern of the hbond graph seems to be in correlation with the one of the RMSD. Actually, the changes in the number of hydrogen bonds happen during the same time frames of the RMSD ones. We can note a lower RMSD (0.05 nm to 0.1 nm) at the starting and end of the simulation consistent with the higher number of hbonds during the same period.

5.3.3.8 Conclusions on Hits Simulation

Regarding their backbone RMSD, all proteins' RMSDs were lower than 2.75 nm (the maximum value observed with SANC00339). We can thus conclude to the relative stability of the complexes.

Comparatively, about their radius of gyration, SANC00570, SANC00339 exhibited the most stable binding. On the other hand, SANC00438 and SANC00152 showed an increase in their Rg attaining values of 2.17 nm and 2.15 nm respectively. Longer simulations could help investigate more of the behaviour of the ligands in this binding.

Among the different cofactors, one can note a sharp change in the RMSD in the beginning of the simulation (first 3 ns) for compounds SANC00570, SANC00339, SANC00152. This can be linked to a better rearrangement of the cofactor within the complex. This may be due to the transfer of the cofactor from another structure (3AU9) as previously described in the methodology (section 5.2.3). These changes in RMSD can be linked to the adaptation of the molecule to the new protein.

Comparing the NADPHs to the ligands, the ligands exhibited greater stability. Indeed the different RMSD graphs for NADPH were characterized by much larger fluctuations except in complexes with SANC00438 and SANC00570. In these systems, the cofactor is more stable with an RMSD levelling off around 0.15 nm. In the remaining complexes, the molecule displayed some change in the RMSD. Interestingly, in both complexes with SANC00339 and SANC00236, the cofactor shows a significant change in RMSD at around 60 ns of simulation. Nonetheless, in all complexes, in regard to the hydrogen bonding, the cofactor shows good interactions with the protein, showing an average of five (5) hydrogen bonds to the macromolecule during the entire simulation.

About the residues' fluctuations, a common pattern in the different RMSFs, is the higher degree of fluctuation observed for residues in the ranges 280-300 and 380-400. The region 290-300 corresponds to the flexible loop covering the protein active site. These residues being in the loop explains their higher fluctuations on the RMSF plots.

The different systems' proteins showed enough stable RMSD (lower than 2.75 nm) during the simulation. More, the different ligands' RMSDs showed very low values, the maximum being 0.15 nm. SANC00438 for example, had a very low RMSD value of 0.05 nm which remained stable during the simulation. The cofactors showed a high number of hydrogen bond, consistent throughout the simulations. This related to its initial binding mode with many interactions among which many hydrogen bonds. We also note variations in its RMSD probably related to its adaptation to the molecule.

5.4 Conclusion

In this chapter force field parameters have been developed for the Mn atom in DXR active site, implemented in GROMACS and validated using MD simulation. However, this work may be further refined by the inclusion of further parameters such as for dihedrals. Moreover, a validation

protocol including the bidentate (LC5) ligand in the protein active site can be more accurate to observe the close-to octahedral geometry (obtained after optimization) during simulation or the trigonal bipyramidal observed in the crystal structure.

The stability of the hits complexes has been studied. The five identified hits from the docking chapter (SANC00152, SANC00236, SANC00339, SANC00438 and SANC00570) exhibited stable binding in the protein active site confirming their high binding affinities and with the good poses observed in docking. More, combining their stability, with their good predicted pharmacological properties motivates for further laboratory investigation for these compounds.

Future work

In future, high throughput molecular dynamics for all preselected compounds in docking may be completed, bisubstrate hits may be explored, and all of these may also be applied to the open conformation of the protein. Other databases of chemical compounds can also be explored and/or extend the work to the other proteins of the non-mevalonate pathway. A more detailed study of the different residues implied in the interactions with the ligands may be followed using an automated pipeline. This could help shed more light in the different ligand binding modes. QM/MM can also be applied to explore reactivity in the protein active site, especially the reaction mechanisms and the involvement of the metal center. Other approaches to evaluate docking hits through MD such as MM-PB/GBSA for free-energy calculations and GROMACS sasa module for computing solvent accessible surface areas could be used to explore more the complexes. Extension of the simulations to longer one will provide a more comprehensive *in silico* assessment. Also using free energy calculations such as the Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA) can help explore the energy landscape of the hit complexes. Finally, the identified hits can be further tested in laboratory assays to investigate their potential antimalarial activity.

REFERENCES

- Abraham, Mark James, Teemu Murtola, Roland Schulz, et al.
2015 GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* 1–2: 19–25.
- Achan, Jane, Ambrose O Talisuna, Annette Erhart, et al.
2011 Quinine, an Old Anti-Malarial Drug in a Modern World: Role in the Treatment of Malaria. *Malaria Journal* 10: 144.
- Akhood, Bashir A., Krishna P. Singh, Megha Varshney, et al.
2014 Understanding the Mechanism of Atovaquone Drug Resistance in Plasmodium Falciparum Cytochrome b Mutation Y268S Using Computational Methods. *PLoS One* 9(10): e110041.
- Alexander, Patrick A., Yanan He, Yihong Chen, John Orban, and Philip N. Bryan
2007 The Design and Characterization of Two Proteins with 88% Sequence Identity but Different Structure and Function. *Proceedings of the National Academy of Sciences of the United States of America* 104(29): 11963–11968.
- Allen, Michael P, and others
2004 Introduction to Molecular Dynamics Simulation. *Computational Soft Matter: From Synthetic Polymers to Proteins* 23: 1–28.
- Altschul, S F, T L Madden, A A Schäffer, et al.
1997 Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Research* 25(17): 3389–3402.
- Anand, Kanchan, John Ziebuhr, Parvesh Wadhvani, Jeroen R. Mesters, and Rolf Hilgenfeld
2003 Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs. *Science* 300(5626): 1763–1767.
- Aneja, Babita, Bhumika Kumar, Mohamad Aman Jairajpuri, and Mohammad Abid
2016 A Structure Guided Drug-Discovery Approach towards Identification of Plasmodium Inhibitors. *RSC Adv.* 6(22): 18364–18406.
- Antinori, Spinello, Laura Galimberti, Laura Milazzo, and Mario Corbellino
2012 Biology of Human Malaria Plasmodia Including Plasmodium Knowlesi. *Mediterranean Journal of Hematology and Infectious Diseases* 4(1).
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3340990/>, accessed April 13, 2017.

Argyrou, Argyrides, and John S. Blanchard

2004 Kinetic and Chemical Mechanism of Mycobacterium Tuberculosis 1-Deoxy-D-Xylulose-5-Phosphate Isomeroeductase. *Biochemistry* 43(14): 4375–4384.

Arnold, Konstantin, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede

2006 The SWISS-MODEL Workspace: A Web-Based Environment for Protein Structure Homology Modelling. *Bioinformatics* 22(2): 195–201.

Atanasov, Atanas G., Birgit Waltenberger, Eva-Maria Pferschy-Wenzig, et al.

2015 Discovery and Resupply of Pharmacologically Active Plant-Derived Natural Products: A Review. *Biotechnology Advances* 33(8): 1582–1614.

Aurrecochea, Cristina, John Brestelli, Brian P. Brunk, et al.

2009 PlasmoDB: A Functional Genomic Database for Malaria Parasites. *Nucleic Acids Research* 37(Database issue): D539–D543.

Babu, C. Satheesan, and Carmay Lim

2006 Empirical Force Fields for Biologically Active Divalent Metal Cations in Water. *The Journal of Physical Chemistry. A* 110(2): 691–699.

Badrinarayan, Preethi, Chinmayee Choudhury, and G Narahari Sastry

2015 Molecular Modeling.

Bailey, Timothy L., Nadya Williams, Chris Misleh, and Wilfred W. Li

2006 MEME: Discovering and Analysing DNA and Protein Sequence Motifs. *Nucleic Acids Research* 34(Web Server issue): W369–W373.

Barber, Bridget E., Giri S. Rajahram, Matthew J. Grigg, Timothy William, and Nicholas M. Anstey

2017 World Malaria Report: Time to Acknowledge Plasmodium Knowlesi Malaria. *Malaria Journal* 16. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5374563/>, accessed April 11, 2017.

Batista, Paulo R., Alan Wilter, Elza H. A. B. Durham, and Pedro G. Pascutti

2006 Molecular Dynamics Simulations Applied to the Study of Subtypes of HIV-1 Protease Common to Brazil, Africa, and Asia. *Cell Biochemistry and Biophysics* 44(3): 395–404.

Baxevanis, Andreas D., and B. F. Francis Ouellette, eds.

2001 *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. 2nd ed. *Methods of Biochemical Analysis*, v. 43. New York: Wiley-Interscience.

Benkert, Pascal, Marco Biasini, and Torsten Schwede

2011 Toward the Estimation of the Absolute Quality of Individual Protein Structure Models. *Bioinformatics (Oxford, England)* 27(3): 343–350.

- Ben-Naim, A.
2002 Molecular Recognition—viewed through the Eyes of the Solvent. *Biophysical Chemistry* 101–102(Supplement C). Special Issue in Honour of John A Schellman: 309–319.
- Bennink, Sandra, Meike J. Kiesow, and Gabriele Pradel
2016 The Development of Malaria Parasites in the Mosquito Midgut. *Cellular Microbiology* 18(7): 905–918.
- Berg, Jeremy M., John L. Tymoczko, and Lubert Stryer
2002 Free Energy Is a Useful Thermodynamic Function for Understanding Enzymes. <https://www.ncbi.nlm.nih.gov/books/NBK22584/>, accessed November 1, 2017.
- Berman, Helen M., John Westbrook, Zukang Feng, et al.
2000 The Protein Data Bank. *Nucleic Acids Research* 28(1): 235–242.
- Bhagavathula, Akshaya Srikanth, Asim Ahmed Elnour, and Abdulla Shehab
2016 Alternatives to Currently Used Antimalarial Drugs: In Search of a Magic Bullet. *Infectious Diseases of Poverty* 5(1): 103.
- Biamonte, Marco A., Jutta Wanner, and Karine G. Le Roch
2013 Recent Advances in Malaria Drug Discovery. *Bioorganic & Medicinal Chemistry Letters* 23(10): 2829–2843.
- Bickerton, G. Richard, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins
2012 Quantifying the Chemical Beauty of Drugs. *Nature Chemistry* 4(2): 90–98.
- Bietz, Stefan, and Matthias Rarey
2016 SIENA: Efficient Compilation of Selective Protein Binding Site Ensembles. *Journal of Chemical Information and Modeling* 56(1): 248–259.
- Bissantz, Caterina, Bernd Kuhn, and Martin Stahl
2010 A Medicinal Chemist's Guide to Molecular Interactions. *Journal of Medicinal Chemistry* 53(14): 5061–5084.
- Bodill, Taryn, Anne C. Conibear, Gregory L. Blatch, Kevin A. Lobb, and Perry T. Kaye
2011 Synthesis and Evaluation of Phosphonated N-Heteroarylcarboxamides as DOXP-Reductoisomerase (DXR) Inhibitors. *Bioorganic & Medicinal Chemistry* 19(3). *Imaging Probes*: 1321–1327.
- Bodill, Taryn, Anne C. Conibear, Marius K. M. Mutorwa, et al.
2013 Exploring DOXP-Reductoisomerase Binding Limits Using Phosphonated N-Aryl and N-Heteroarylcarboxamides as DXR Inhibitors. *Bioorganic & Medicinal Chemistry* 21(14): 4332–4341.

- Bromley, Candice L., Shirley Parker-Nance, Jo-Anne de la Mare, et al.
2013 Halogenated Oxindole and Indoles from the South African Marine Ascidian *Distaplia Skoogi*. *South African Journal of Chemistry* 66: 00–00.
- Brooks, Bernard R., Robert E. Bruccoleri, Barry D. Olafson, et al.
1983 CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry* 4(2): 187–217.
- Brown, Terence A.
2002 *Molecular Phylogenetics*. Wiley-Liss. <https://www.ncbi.nlm.nih.gov/books/NBK21122/>, accessed August 13, 2017.
- Budzik, Brian, Vincenzo Garzya, Dongchuan Shi, et al.
2010 Novel N-Substituted Benzimidazolones as Potent, Selective, CNS-Penetrant, and Orally Active M1 MACHR Agonists. *ACS Medicinal Chemistry Letters* 1(6): 244–248.
- Burrows, JEREMY N, EMILIE BURLOT, BRICE CAMPO, et al.
2014 Antimalarial Drug Discovery – the Path towards Eradication. *Parasitology* 141(1): 128–139.
- Campo, Brice, Omar Vandal, David L. Wesche, and Jeremy N. Burrows
2015 Killing the Hypnozoite – Drug Discovery Approaches to Prevent Relapse in *Plasmodium Vivax*. *Pathogens and Global Health* 109(3): 107–122.
- Chang, Wei-chen, Heng Song, Hung-wen Liu, and Pinghua Liu
2013 Current Development in Isoprenoid Precursor Biosynthesis and Regulation. *Current Opinion in Chemical Biology* 17(4): 571–579.
- Chaudhary, Kamal Kumar, and C.V.S. Siva Prasad
2014 Virtual Screening of Compounds to 1-Deoxy-Dxylulose 5-Phosphate Reductoisomerase (DXR) from *Plasmodium Falciparum*. *Bioinformatics* 10(6): 358–364.
- Chiodo, S., N. Russo, and E. Sicilia
2006 LANL2DZ Basis Sets Recontracted in the Framework of Density Functional Theory. *The Journal of Chemical Physics* 125(10): 104107.
- Chofor, René
2016 Synthesis and Evaluation of 1-Deoxy-D-Xylulose 5-Phosphate Reductoisomerase Inhibitors as Antimalarial and Antituberculosis Agents. Ghent University. <https://biblio.ugent.be/publication/8043366/file/8043367>, accessed May 4, 2017.
- Chofor, René, Martijn D. P. Risseeuw, Jenny Pouyez, et al.
2014 Synthetic Fosmidomycin Analogues with Altered Chelating Moieties Do Not Inhibit 1-

Deoxy-D-Xylulose 5-Phosphate Reductoisomerase or Plasmodium Falciparum Growth In Vitro. *Molecules* 19(2): 2571–2587.

Chofor, René, Sanjeewani Sooriyaarachchi, Martijn D. P. Risseeuw, et al.
2015 Synthesis and Bioactivity of β -Substituted Fosmidomycin Analogues Targeting 1-Deoxy-d-Xylulose-5-Phosphate Reductoisomerase. *Journal of Medicinal Chemistry* 58(7): 2988–3001.

Cobb, Ryan E., Brian Bae, Zhi Li, et al.
2015 Structure-Guided Design and Biosynthesis of a Novel FR-900098 Analogue as a Potent Plasmodium Falciparum 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase (Dxr) Inhibitor. *Chemical Communications (Cambridge, England)* 51(13): 2526–2528.

Control, Institute of Medicine (US) Committee for the Study on Malaria Prevention and, Jr Stanley C. Oaks, Violaine S. Mitchell, Greg W. Pearson, and Charles C. J. Carpenter
1991 Parasite Biology. National Academies Press (US).
<https://www.ncbi.nlm.nih.gov/books/NBK234327/>, accessed April 14, 2017.

Cornell, Wendy D, Piotr Cieplak, Christopher I Bayly, et al.
1995 A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* 117(19): 5179–5197.

Cosconati, Sandro, Stefano Forli, Alex L. Perryman, et al.
2010 Virtual Screening with AutoDock: Theory and Practice. *Expert Opinion on Drug Discovery* 5(6): 597–607.

Cowman, Alan F., Julie Healer, Danushka Marapana, and Kevin Marsh
2016 Malaria: Biology and Disease. *Cell* 167(3): 610–624.

Crutcher, James M., and Stephen L. Hoffman
1996 Malaria. *In* Medical Microbiology. 4th edition. Samuel Baron, ed. Galveston (TX): University of Texas Medical Branch at Galveston.
<http://www.ncbi.nlm.nih.gov/books/NBK8584/>, accessed April 10, 2017.

Cui, Liwang, Sungano Mharakurwa, Daouda Ndiaye, Pradipsinh K. Rathod, and Philip J. Rosenthal
2015 Antimalarial Drug Resistance: Literature Review and Activities and Findings of the ICEMR Network. *The American Journal of Tropical Medicine and Hygiene* 93(3 Suppl): 57–68.

Cui, Wei, Zhuo Wei, Quan Chen, et al.
2010 Structure-Based Design of Peptides against G3BP with Cytotoxicity on Tumor Cells. *Journal of Chemical Information and Modeling* 50(3): 380–387.

Datta, Shubhabrata, Shubhabrata Datta, and J. Paulo Davim

2016 Computational Approaches to Materials Design: Theoretical and Practical Aspects. 1st edition. Hershey, PA, USA: IGI Global.

De Vivo, Marco, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli

2016 Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry* 59(9): 4035–4061.

Deng, Lisheng, Kiwamu Endo, Masahiro Kato, et al.

2010 Structures of 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase/Lipophilic Phosphonate Complexes. *ACS Medicinal Chemistry Letters* 2(2): 165–170.

Doak, Bradley Croy, Björn Over, Fabrizio Giordanetto, and Jan Kihlberg

2014 Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chemistry & Biology* 21(9): 1115–1142.

Dong, Guang Qiang, Hao Fan, Dina Schneidman-Duhovny, Ben Webb, and Andrej Sali

2013 Optimized Atomic Statistical Potentials: Assessment of Protein Interfaces and Loops. *Bioinformatics* 29(24): 3158–3166.

Dror, Ron O., Hillary F. Green, Celine Valant, et al.

2013 Structural Basis for Modulation of a G-Protein-Coupled Receptor by Allosteric Drugs. *Nature* 503(7475): 295–299.

Du, Xing, Yi Li, Yuan-Ling Xia, et al.

2016 Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. *International Journal of Molecular Sciences* 17(2).
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4783878/>.

Duan, Yong, Chun Wu, Shibasish Chowdhury, et al.

2003 A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *Journal of Computational Chemistry* 24(16): 1999–2012.

Edgar, Robert C.

2004 MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research* 32(5): 1792–1797.

Ehret, Totta, Francesca Torelli, Christian Klotz, Amy B. Pedersen, and Frank Seeber

2017 Translational Rodent Models for Research on Parasitic Protozoa—A Review of Confounders and Possibilities. *Frontiers in Cellular and Infection Microbiology* 7.
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5461347/>.

Eisenreich, Wolfgang, Matthias Schwarz, Alain Cartayrade, et al.

1998 The Deoxyxylulose Phosphate Pathway of Terpenoid Biosynthesis in Plants and Microorganisms. *Chemistry & Biology* 5(9): R221–R233.

Fernandes, Roberta P. M., and Philip J. Proteau

2006 Kinetic Characterization of *Synechocystis* Sp. PCC6803 1-Deoxy-d-Xylulose 5-Phosphate Reductoisomerase Mutants. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1764(2): 223–229.

Fiser, András, and Andrej Šali

2003 Modeller: Generation and Refinement of Homology-Based Protein Structure Models. *In* . *BT - Methods in Enzymology*, ed. Pp. 461–491. *Macromolecular Crystallography, Part D*. Academic Press. <http://www.sciencedirect.com/science/article/pii/S0076687903740208>, accessed April 4, 2017.

França, Tanos Celmar Costa

2015 Homology Modeling: An Important Tool for the Drug Discovery. *Journal of Biomolecular Structure & Dynamics* 33(8): 1780–1793.

Frisch, MJ, GW Trucks, HB Schlegel, et al.

2009 Gaussian 09, Revision D. 01.

Gadzała, M., B. Kalinowska, M. Banach, L. Konieczny, and I. Roterman

2017 Determining Protein Similarity by Comparing Hydrophobic Core Structure. *Heliyon* 3(2): e00235.

Ganesan, Aravindhan, Michelle L. Coote, and Khaled Barakat

2017 Molecular Dynamics-Driven Drug Discovery: Leaping Forward with Confidence. *Drug Discovery Today* 22(2): 249–269.

García-Sosa, Alfonso T., Csaba Hetényi, and Uko Maran

2010 Drug Efficiency Indices for Improvement of Molecular Docking Scoring Functions. *Journal of Computational Chemistry* 31(1): 174–184.

Genheden, Samuel, and Ulf Ryde

2015 The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opinion on Drug Discovery* 10(5): 449–461.

Goble, Jessica Leigh

2011 The Druggable Antimalarial Target 1–deoxy–D–Xylulose–5–phosphate Reductoisomerase: Purification, Kinetic Characterization and Inhibition Studie. Rhodes University.

González, M. A.

2011 Force Fields and Molecular Dynamics Simulations. *École Thématique de La Société Française de La Neutronique* 12: 169–200.

Gromacs Documentation

N.d. <http://manual.gromacs.org/documentation/5.1.4/>, accessed June 11, 2017.

Guggisberg, Ann M., Rachel E. Amthor, and Audrey R. Odom

2014 Isoprenoid Biosynthesis in *Plasmodium Falciparum*. *Eukaryotic Cell* 13(11): 1348–1359.

van Gunsteren, Wilfred, SR Billeter, AA Eising, et al.

1996 Biomolecular Simulation: The {GROMOS96} Manual and User Guide.

Hale, Ian, Paul M. O'Neill, Neil G. Berry, Audrey Odom, and Raman Sharma

2012 The MEP Pathway and the Development of Inhibitors as Potential Anti-Infective Agents. *MedChemComm* 3(4): 418–433.

Hatherley, Rowan, David K Brown, Thommas M Musyoka, et al.

2015 SANCDDB: A South African Natural Compound Database. *Journal of Cheminformatics* 7. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4471313/>.

Hemingway, Janet, Rima Shretta, Timothy N. C. Wells, et al.

2016 Tools and Strategies for Malaria Control and Elimination: What Do We Need to Achieve a Grand Convergence in Malaria? *PLoS Biology* 14(3). <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4774904/>, accessed April 11, 2017.

Henikoff, S., and J. G. Henikoff

1992 Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences* 89(22): 10915–10919.

Henriksson, Lena M., Torsten Unge, Jens Carlsson, et al.

2007 Structures of Mycobacterium Tuberculosis 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase Provide New Insights into Catalysis. *Journal of Biological Chemistry* 282(27): 19905–19916.

Hillis, David M., and James J. Bull

1993 An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology* 42(2): 182–192.

Holstein, Sarah A., and Raymond J. Hohl

2004 Isoprenoids: Remarkable Diversity of Form and Function. *Lipids* 39(4): 293–309.

Hooft, R. W., G. Vriend, C. Sander, and E. E. Abola

1996 Errors in Protein Structures. *Nature* 381(6580): 272.

- Hopkins, Andrew L., Colin R. Groom, and Alexander Alex
2004 Ligand Efficiency: A Useful Metric for Lead Selection. *Drug Discovery Today* 9(10): 430–431.
- Hopkins, Andrew L., György M. Keserü, Paul D. Leeson, David C. Rees, and Charles H. Reynolds
2014 The Role of Ligand Efficiency Metrics in Drug Discovery. *Nature Reviews. Drug Discovery* 13(2): 105–121.
- Hu, LiHong, and Ulf Ryde
2011 Comparison of Methods to Obtain Force-Field Parameters for Metal Sites. *Journal of Chemical Theory and Computation* 7(8): 2452–2463.
- Huang, Sheng-You
2014 Search Strategies and Evaluation in Protein–protein Docking: Principles, Advances and Challenges. *Drug Discovery Today* 19(8): 1081–1096.
- Huang, Yuanpeng J., Binchen Mao, James M. Aramini, and Gaetano T Montelione
2014 Assessment of Template Based Protein Structure Predictions in CASP10. *Proteins* 82(0 2): 43–56.
- Humphrey, W., A. Dalke, and K. Schulten
1996 VMD: Visual Molecular Dynamics. *Journal of Molecular Graphics* 14(1): 33–38, 27–28.
- Iguchi, E., M. Okuhara, M. Kohsaka, H. Aoki, and H. Imanaka
1980 Studies on New Phosphonic Acid Antibiotics. II. Taxonomic Studies on Producing Organisms of the Phosphonic Acid and Related Compounds. *The Journal of Antibiotics* 33(1): 19–23.
- Imlay, Leah, and Audrey R. Odom
2014 Isoprenoid Metabolism in Apicomplexan Parasites. *Current Clinical Microbiology Reports* 1(3–4): 37–50.
- Irwin, John J., Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman
2012 ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling* 52(7): 1757–1768.
- Jessica L. Goble, Hailey Johnson
N.d. The Druggable Antimalarial Target PfDXR: Overproduction Strategies and Kinetic Characterization. [Http://Www.Eurekaselect.Com](http://www.Eurekaselect.Com).
<http://www.eurekaselect.com/105826/article>, accessed March 16, 2017.
- Jomaa, Hassan, Jochen Wiesner, Silke Sanderbrand, et al.

1999 Inhibitors of the Nonmevalonate Pathway of Isoprenoid Biosynthesis as Antimalarial Drugs. *Science* 285(5433): 1573–1576.

Jones, Eric, Travis Oliphant, and Pearu Peterson

2001 {SciPy}: Open Source Scientific Tools for {Python}. <http://www.scipy.org>.

Kc, Dukka B.

2016 Recent Advances in Sequence-Based Protein Structure Prediction. *Briefings in Bioinformatics*.

Kendall, M. G.

1938 A NEW MEASURE OF RANK CORRELATION. *Biometrika* 30(1–2): 81–93.

Kennedy, David A., and Andrew F. Read

2017 Why Does Drug Resistance Readily Evolve but Vaccine Resistance Does Not? *Proc. R. Soc. B* 284(1851): 20162562.

Keseru, György M., and Gergely M. Makara

2006 Hit Discovery and Hit-to-Lead Approaches. *Drug Discovery Today* 11(15–16): 741–748.

Kholodar, Svetlana A., and Andrew S. Murkin

2013 DXP Reductoisomerase: Reaction of the Substrate in Pieces Reveals a Catalytic Role for the Nonreacting Phosphodianion Group. *Biochemistry* 52(13): 2302–2308.

Kholodar, Svetlana A., Gregory Tomblin, Juan Liu, et al.

2014 Alteration of the Flexible Loop in 1-Deoxy-d-Xylulose-5-Phosphate Reductoisomerase Boosts Enthalpy-Driven Inhibition by Fosmidomycin. *Biochemistry* 53(21): 3423–3431.

Kleywegt, G. J., M. R. Harris, J. Zou, et al.

2004 The Uppsala Electron-Density Server. *Acta Crystallographica Section D: Biological Crystallography* 60(12): 2240–2249.

Kolafa, Jiri, and John W. Perram

1992 Cutoff Errors in the Ewald Summation Formulae for Point Charge Systems. *Molecular Simulation* 9(5): 351–368.

Konzuch, Sarah, Tomonobu Umeda, Jana Held, et al.

2014 Binding Modes of Reverse Fosmidomycin Analogs toward the Antimalarial Target IspC. *Journal of Medicinal Chemistry* 57(21): 8827–8838.

Krieger, Elmar, Sander B. Nabuurs, and Gert Vriend

2005 Homology Modeling. *In* *Methods of Biochemical Analysis*. Philip E. Bourne and Helge Weissig, eds. Pp. 509–523. Hoboken, NJ, USA: John Wiley & Sons, Inc. <http://doi.wiley.com/10.1002/0471721204.ch25>, accessed April 3, 2017.

- Krivov, Georgii G., Maxim V. Shapovalov, and Roland L. Dunbrack
2009 Improved Prediction of Protein Side-Chain Conformations with SCWRL4. *Proteins* 77(4): 778–795.
- Kryshtafovych, Andriy, Krzysztof Fidelis, and John Moult
2014 CASP10 Results Compared to Those of Previous CASP Experiments. *Proteins* 82(0 2): 164–174.
- Kufareva, Irina, and Ruben Abagyan
2012 Methods of Protein Structure Comparison. *Methods in Molecular Biology* (Clifton, N.J.) 857: 231–257.
- Kunfermann, Andrea, Claudia Lienau, Boris Illarionov, et al.
2013 IspC as Target for Antiinfective Drug Discovery: Synthesis, Enantiomeric Separation, and Structural Biology of Fosmidomycin Thia Isosters. *Journal of Medicinal Chemistry* 56(20): 8151–8162.
- Lange, B. Markus, Tamas Rujan, William Martin, and Rodney Croteau
2000 Isoprenoid Biosynthesis: The Evolution of Two Ancient and Distinct Pathways across Genomes. *Proceedings of the National Academy of Sciences* 97(24): 13172–13177.
- Laskowski, R. A., M. W. MacArthur, D. S. Moss, and J. M. Thornton
1993 PROCHECK: A Program to Check the Stereochemical Quality of Protein Structures. *Journal of Applied Crystallography* 26(2): 283–291.
- Le, Si Quang, and Olivier Gascuel
2008 An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution* 25(7): 1307–1320.
- Leaver-Fay, Andrew, Michael Tyka, Steven M. Lewis, et al.
2011 Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods in Enzymology* 487: 545–574.
- Lewars, Errol G.
2016 *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*. Springer.
- Li, Heng, Jie Tian, Wei Sun, Wei Qin, and Wen-Yun Gao
2013 Mechanistic Insights into 1-Deoxy-d-Xylulose 5-Phosphate Reductoisomerase, a Key Enzyme of the MEP Terpenoid Biosynthetic Pathway. *FEBS Journal* 280(22): 5896–5905.
- Li, Xiaosong, and Michael J. Frisch

2006 Energy-Represented Direct Inversion in the Iterative Subspace within a Hybrid Geometry Optimization Method. *Journal of Chemical Theory and Computation* 2(3): 835–839.

Lionta, Evanthia, George Spyrou, Demetrios K. Vassilatis, and Zoe Cournia

2014 Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Current Topics in Medicinal Chemistry* 14(16): 1923–1938.

Lipinski, Christopher A.

2004 Lead- and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discovery Today: Technologies* 1(4): 337–341.

Liu, T., Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson

2007 BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Research* 35(Database): D198–D201.

Lodish, Harvey, Arnold Berk, S. Lawrence Zipursky, et al.

2000 Noncovalent Bonds. <https://www.ncbi.nlm.nih.gov/books/NBK21726/>, accessed July 13, 2017.

di Luccio, Eric, and Patrice Koehl

2011 A Quality Metric for Homology Modeling: The H-Factor. *BMC Bioinformatics* 12: 48.

Lunev, Sergey, Fernando A. Batista, Soraya S. Bosch, Carsten Wrenger, and Matthew R. Groves

2016 Identification and Validation of Novel Drug Targets for the Treatment of Plasmodium Falciparum Malaria: New Insights. <http://www.intechopen.com/books/current-topics-in-malaria/identification-and-validation-of-novel-drug-targets-for-the-treatment-of-plasmodium-falciparum-malar>, accessed April 15, 2017.

Lüthy, Roland, James U. Bowie, and David Eisenberg

1992 Assessment of Protein Models with Three-Dimensional Profiles. *Nature* 356(6364): 83–85.

Lysozyme in Water

N.d. <http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/lysozyme/index.html>, accessed December 7, 2017.

Mac Sweeney, Aengus, Roland Lange, Roberta P.M. Fernandes, et al.

2005 The Crystal Structure of E.Coli 1-Deoxy-d-Xylulose-5-Phosphate Reductoisomerase in a Ternary Complex with the Antimalarial Compound Fosmidomycin and NADPH Reveals a Tight-Binding Closed Enzyme Conformation. *Journal of Molecular Biology* 345(1): 115–127.

MacKerell, A. D., D. Bashford, M. Bellott, et al.

1998 All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry. B* 102(18): 3586–3616.

- Malde, Alpeshkumar K., Le Zuo, Matthew Breeze, et al.
2011 An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *Journal of Chemical Theory and Computation* 7(12): 4026–4037.
- Mariani, Valerio, Marco Biasini, Alessandro Barbato, and Torsten Schwede
2013 LDDT: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests. *Bioinformatics* 29(21): 2722–2728.
- Mark Abraham, Berk Hess, David van der Spoel, and Erik, E., Weiqing Ren, and Eric Vanden-Eijnden
2002 Gromacs User Manual Version 5.0.7.
- Marx, Vivien
2013 Biology: The Big Challenges of Big Data. *Nature* 498(7453): 255–260.
- Masini, Tiziana, Blijke S. Kroezen, and Anna K. H. Hirsch
2013 Druggability of the Enzymes of the Non-Mevalonate-Pathway. *Drug Discovery Today* 18(23–24): 1256–1262.
- van der Meer, Jan-Ytzen, and Anna K. H. Hirsch
2012 The Isoprenoid-Precursor Dependence of Plasmodium Spp. *Natural Product Reports* 29(7): 721–728.
- Melo, F., and E. Feytmans
1998 Assessing Protein Structures with a Non-Local Atomic Interaction Energy. *Journal of Molecular Biology* 277(5): 1141–1152.
- Meng, Xuan-Yu, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui
2011 Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer-Aided Drug Design* 7(2): 146–157.
- Mercklé, Ludovic, Ana de Andrés-Gómez, Bethany Dick, Russell J. Cox, and Christopher R. A. Godfrey
2005 Fragment-Based Approach to Understanding Inhibition of 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase. *Chembiochem*. <http://agris.fao.org/agris-search/search.do?recordID=US201301942494>, accessed November 20, 2017.
- Meyder, Agnes, Eva Nittinger, Gudrun Lange, Robert Klein, and Matthias Rarey
2017 Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-Ray Structures. *Journal of Chemical Information and Modeling* 57(10): 2437–2447.
- Mishra, Mitali, Vikash K. Mishra, Varsha Kashaw, Arun K. Iyer, and Sushil Kumar Kashaw

2017 Comprehensive Review on Various Strategies for Antimalarial Drug Discovery. *European Journal of Medicinal Chemistry* 125: 1300–1320.

Miyano, Satoru, Jill Mesirov, Simon Kasif, et al.

2005 Research in Computational Molecular Biology: 9th Annual International Conference, RECOMB 2005, Cambridge, MA, USA, May 14-18, 2005, Proceedings. Springer Science & Business Media.

Modeller Tutorial

N.d. <https://salilab.org/modeller/tutorial/basic.html>, accessed August 15, 2017.

Moman, Edelmiro

2011 A Useful Script for AutoDock4 and Vina That Merges Flexible and Rigid PDBQT Output « Prosciens. <http://prosciens.com/prosciens/a-useful-script-for-autodock4-and-vina-that-merges-flexible-and-rigid-pdbqt-output/>, accessed November 1, 2017.

Morris, Garrett M., Ruth Huey, William Lindstrom, et al.

2009 AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of Computational Chemistry* 30(16): 2785–2791.

Morrisette, Naomi S., and L. David Sibley

2002 Cytoskeleton of Apicomplexan Parasites. *Microbiology and Molecular Biology Reviews* 66(1): 21–38.

Mortier, Jérémie, Christin Rakers, Marcel Bermudez, et al.

2015 The Impact of Molecular Dynamics on Drug Design: Applications for the Characterization of Ligand–macromolecule Complexes. *Drug Discovery Today* 20(6): 686–702.

Mukesh, Backwani, and Kumar Rakesh

2011 Molecular Docking: A Review. *Int J Res Ayurveda Pharm* 2: 746–1751.

Mukhopadhyay, Mayukh

2014 A Brief Survey on Bio Inspired Optimization Algorithms for Molecular Docking. *International Journal of Advances in Engineering & Technology* 7: 868–878.

Munos, Jeffrey W., Xiaotao Pu, Steven O Mansoorabadi, Hak Joong Kim, and Hung-wen Liu

2009 A Secondary Kinetic Isotope Effect Study of the 1-Deoxy-d-Xylulose-5-Phosphate Reductoisomerase-Catalyzed Reaction: Evidence for a Retroaldol-Aldol Rearrangement. *Journal of the American Chemical Society* 131(6): 2048–2049.

Murkin, Andrew S., Kathryn A. Manning, and Svetlana A. Kholodar

2014 Mechanism and Inhibition of 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase. *Bioorganic Chemistry* 57: 171–185.

NCBI Resource Coordinators

2017 Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 45(D1): D12–D17.

Neves, Rui P. P., Sérgio F. Sousa, Pedro A. Fernandes, and Maria J. Ramos

2013 Parameters for Molecular Dynamics Simulations of Manganese-Containing Metalloproteins. *Journal of Chemical Theory and Computation* 9(6): 2718–2732.

Newman, David J., and Gordon M. Cragg

2016 Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products* 79(3): 629–661.

Nguyen, Trang Truc, Man Hoang Viet, and Mai Suan Li

2014 Effects of Water Models on Binding Affinity: Evidence from All-Atom Simulation of Binding of Tamiflu to A/H5N1 Neuraminidase. Research article. *The Scientific World Journal*. <https://www.hindawi.com/journals/tswj/2014/536084/>, accessed December 4, 2017.

O'Boyle, Noel M., Michael Banck, Craig A. James, et al.

2011 Open Babel: An Open Chemical Toolbox. *Journal of Cheminformatics* 3(1): 33.

Odom, Audrey R.

2011 Five Questions about Non-Mevalonate Isoprenoid Biosynthesis. *PLOS Pathogens* 7(12): e1002323.

Pason, Lukas P., and Christoph A. Sotriffer

2016 Empirical Scoring Functions for Affinity Prediction of Protein-Ligand Complexes. *Molecular Informatics* 35(11–12): 541–548.

Pawlowski, Marcin, Michal J Gajda, Ryszard Matlak, and Janusz M Bujnicki

2008 MetaMQAP: A Meta-Server for the Quality Assessment of Protein Models. *BMC Bioinformatics* 9: 403.

Pei, Jimin, Bong-Hyun Kim, and Nick V. Grishin

2008 PROMALS3D: A Tool for Multiple Protein Sequence and Structure Alignments. *Nucleic Acids Research* 36(7): 2295–2300.

Pérez-Gil, Jordi, Bárbara M. Calisto, Christoph Behrendt, et al.

2012 Crystal Structure of *Brucella Abortus* Deoxyxylulose-5-Phosphate Reductoisomerase-like (DRL) Enzyme Involved in Isoprenoid Biosynthesis♦. *The Journal of Biological Chemistry* 287(19): 15803–15809.

Petrenko, Roman, and Jarosław Meller

2001 Molecular Dynamics. *In* ELS. John Wiley & Sons, Ltd.
<http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0003048.pub2/abstract>.

Pevsner, Jonathan

2009 Bioinformatics and Functional Genomics. 2nd ed. Hoboken, N.J: Wiley-Blackwell.

Protein-Ligand Complex

N.d. <http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/complex/index.html>, accessed December 7, 2017.

Ramachandran, K. I., Gopakumar Deepa, and Krishnan Namboori

2008 Computational Chemistry and Molecular Modeling: Principles and Applications. Springer Science & Business Media.

RCSB PDB - Content Growth Report

N.d. <http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=fold-scop>, accessed July 24, 2017.

Reche, Pedro A., and Ellis L. Reinherz

2003 Sequence Variability Analysis of Human Class I and Class II MHC Molecules: Functional and Structural Correlates of Amino Acid Polymorphisms. *Journal of Molecular Biology* 331(3): 623–641.

Reuter, Klaus, Silke Sanderbrand, Hassan Jomaa, et al.

2002 Crystal Structure of 1-Deoxy-d-Xylulose-5-Phosphate Reductoisomerase, a Crucial Enzyme in the Non-Mevalonate Pathway of Isoprenoid Biosynthesis. *Journal of Biological Chemistry* 277(7): 5378–5384.

Rigden, Daniel J.

2017 From Protein Structure to Function with Bioinformatics. Springer.

Rohmer, M, M Knani, P Simonin, B Sutter, and H Sahm

1993 Isoprenoid Biosynthesis in Bacteria: A Novel Pathway for the Early Steps Leading to Isopentenyl Diphosphate. *Biochemical Journal* 295(Pt 2): 517–524.

Rost, B.

1999a Twilight Zone of Protein Sequence Alignments. *Protein Engineering* 12(2): 85–94.

1999b Twilight Zone of Protein Sequence Alignments. *Protein Engineering* 12(2): 85–94.

Roux, Benoît, and Thomas Simonson

1999 Implicit Solvent Models. *Biophysical Chemistry* 78(1): 1–20.

de Ruyck, Jerome, Guillaume Brysbaert, Ralf Blossey, and Marc F Lensink

2016 Molecular Docking as a Popular Tool in Drug Design, an in Silico Travel. *Advances and Applications in Bioinformatics and Chemistry* : AABC 9: 1–11.

de Ruyck, Jérôme, Johan Wouters, and C. Dale Poulter

2011 Inhibition Studies on Enzymes Involved in Isoprenoid Biosynthesis: Focus on Two Potential Drug Targets: DXR and IDI-2 Enzymes. *Current Enzyme Inhibition* 7(2).
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3856697/>, accessed May 14, 2017.

Sacchettini, J. C., and C. D. Poulter

1997 Creating Isoprenoid Diversity. *Science (New York, N.Y.)* 277(5333): 1788–1789.

Saggu, Gagandeep S., Zarna R. Pala, Shilpi Garg, and Vishal Saxena

2016 New Insight into Isoprenoids Biosynthesis Process and Future Prospects for Drug Designing in Plasmodium. *Frontiers in Microbiology* 7: 1421.

Šali, Andrej, and Tom L. Blundell

1993 Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology* 234(3): 779–815.

Šali, Andrej, Ben Webb, M. S. Madhusudhan, et al.

2017 MODELLER A Program for Protein Structure Modeling Release 9.19.
<https://www.salilab.org/modeller/9.18/manual.pdf>, accessed August 15, 2017.

San Jose, Géraldine, Emily R. Jackson, Eugene Uh, et al.

2013 Design of Potential Bisubstrate Inhibitors against Mycobacterium Tuberculosis (Mtb) 1-Deoxy-D-Xylulose 5-Phosphate Reductoisomerase (Dxr)—Evidence of a Novel Binding Mode. *MedChemComm* 4(7): 1099–1104.

Sandak, Bilha, Haim J. Wolfson, and Ruth Nussinov

1998 Flexible Docking Allowing Induced Fit in Proteins: Insights from an Open to Closed Conformational Isomers. *Proteins: Structure, Function, and Bioinformatics* 32(2): 159–174.

Sangari, Félix J., Jordi Pérez-Gil, Lorenzo Carretero-Paulet, Juan M. García-Lobo, and Manuel Rodríguez-Concepción

2010 A New Family of Enzymes Catalyzing the First Committed Step of the Methylerythritol 4-Phosphate (MEP) Pathway for Isoprenoid Biosynthesis in Bacteria. *Proceedings of the National Academy of Sciences of the United States of America* 107(32): 14081–14086.

Schmidt, Michael W., Kim K. Baldrige, Jerry A. Boatz, et al.

1993 General Atomic and Molecular Electronic Structure System. *Journal of Computational Chemistry* 14(11): 1347–1363.

Schneider, Gisbert

2013 Prediction of Drug-Like Properties. Landes Bioscience.
<https://www.ncbi.nlm.nih.gov/books/NBK6404/>, accessed November 13, 2017.

Schuck, Desiree C., Sabrina B. Ferreira, Laura N. Cruz, et al.
2013 Biological Evaluation of Hydroxynaphthoquinones as Anti-Malarials. *Malaria Journal* 12: 234.

Schüttelkopf, Alexander W., and Daan M. F. van Aalten
2004 PRODRG: A Tool for High-Throughput Crystallography of Protein-Ligand Complexes. *Acta Crystallographica. Section D, Biological Crystallography* 60(Pt 8): 1355–1363.

Shityakov, Sergey, and Carola Förster
2014 In Silico Predictive Model to Determine Vector-Mediated Transport Properties for the Blood–brain Barrier Choline Transporter. *Advances and Applications in Bioinformatics and Chemistry : AABC* 7: 23–36.

Silber, Katrin, Philipp Heidler, Thomas Kurz, and Gerhard Klebe
2005 AFMoC Enhances Predictivity of 3D QSAR: A Case Study with DOXP-Reductoisomerase. *Journal of Medicinal Chemistry* 48(10): 3547–3563.

Sillitoe, Ian, Tony E. Lewis, Alison Cuff, et al.
2015 CATH: Comprehensive Structural and Functional Annotations for Genome Sequences. *Nucleic Acids Research* 43(D1): D376–D381.

Singh, Nidhi, Gweneal Chevé, Mitchell A. Avery, and Christopher R. McCurdy
2007 Targeting the Methyl Erythritol Phosphate (MEP) Pathway for Novel Antimalarial, Antibacterial and Herbicidal Drug Discovery: Inhibition of 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase (DXR) Enzyme. *Current Pharmaceutical Design* 13(11): 1161–1177.

Sliwoski, Gregory, Sandeepkumar Kothiwale, Jens Meiler, and Edward W. Lowe
2014 Computational Methods in Drug Discovery. *Pharmacological Reviews* 66(1): 334–395.

Söding, Johannes, Andreas Biegert, and Andrei N. Lupas
2005 The HHpred Interactive Server for Protein Homology Detection and Structure Prediction. *Nucleic Acids Research* 33(Web Server issue): W244–W248.

Sooriyaarachchi, Sanjeewani, René Chofor, Martijn D. P. Risseeuw, et al.
2016 Targeting an Aromatic Hotspot in Plasmodium Falciparum 1-Deoxy-d-Xylulose-5-Phosphate Reductoisomerase with β -Arylpropyl Analogues of Fosmidomycin. *ChemMedChem* 11(18): 2024–2036.

Soulard, Valérie, Henriette Bosson-Vanga, Audrey Lorthiois, et al.
2015 Plasmodium Falciparum Full Life Cycle and Plasmodium Ovale Liver Stages in Humanized Mice. *Nature Communications* 6: 7690.

Sousa da Silva, Alan W., and Wim F. Vranken

2012 ACPYPE - AnteChamber PYthon Parser InterfacE. *BMC Research Notes* 5: 367.

Srinivasan, Bharath, João V. Rodrigues, Sam Tondast-Navaei, Eugene Shakhnovich, and Jeffrey Skolnick

2017 Rational Design of Novel Allosteric Dihydrofolate Reductase Inhibitors Showing Antibacterial Effects on Drug-Resistant *Escherichia Coli* Escape Variants. *ACS Chemical Biology* 12(7): 1848–1857.

Steinbacher, Stefan, Johannes Kaiser, Wolfgang Eisenreich, et al.

2003 Structural Basis of Fosmidomycin Action Revealed by the Complex with 2-C-Methyl-D-Erythritol 4-Phosphate Synthase (IspC). Implications for the Catalytic Mechanism and Anti-Malaria Drug Development. *The Journal of Biological Chemistry* 278(20): 18401–18407.

Studer, Romain A., Pascal-Antoine Christin, Mark A. Williams, and Christine A. Orengo

2014 Stability-Activity Tradeoffs Constrain the Adaptive Evolution of RubisCO. *Proceedings of the National Academy of Sciences* 111(6): 2223–2228.

Takahashi, Shunji, Tomohisa Kuzuyama, Hiroyuki Watanabe, and Haruo Seto

1998 A 1-Deoxy-d-Xylulose 5-Phosphate Reductoisomerase Catalyzing the Formation of 2-C-Methyl-d-Erythritol 4-Phosphate in an Alternative Nonmevalonate Pathway for Terpenoid Biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 95(17): 9879–9884.

Takenoya, Mihoko, Akashi Ohtaki, Keiichi Noguchi, et al.

2010 Crystal Structure of 1-Deoxy-d-Xylulose 5-Phosphate Reductoisomerase from the Hyperthermophile *Thermotoga Maritima* for Insights into the Coordination of Conformational Changes and an Inhibitor Binding. *Journal of Structural Biology* 170(3): 532–539.

Tangyuenyongwatana, Prasan, and Wandee Gritsanapan

2017 Virtual Screening for Novel 1-Deoxy-d-Xylulose-5-Phosphate Reductoisomerase Inhibitors: A Shape-Based Search Approach. *Thai Journal of Pharmaceutical Sciences (TJPS)* 41(1). <http://www.tjps.pharm.chula.ac.th/ojs/index.php/tjps/article/view/323>, accessed April 10, 2017.

Team, R Core

2014 R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.

Toukmaji, Abdulnour Y., and John A. Board

1996 Ewald Summation Techniques in Perspective: A Survey. *Computer Physics Communications* 95(2): 73–92.

Trott, Oleg, and Arthur J. Olson

2010 AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading. *Journal of Computational Chemistry* 31(2): 455–461.

Turner, PJ

2005 XMGRACE, Version 5.1. 19. Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology, Beaverton, OR.

Umeda, Tomonobu, Nobutada Tanaka, Yoshio Kusakabe, et al.

2011 Molecular Basis of Fosmidomycin's Action on the Human Malaria Parasite *Plasmodium Falciparum*. *Scientific Reports* 1: 9.

Uziela, Karolis, David Menéndez Hurtado, Nanjiang Shu, Björn Wallner, and Arne Elofsson

2017 ProQ3D: Improved Model Quality Assessments Using Deep Learning. *Bioinformatics (Oxford, England)* 33(10): 1578–1580.

Vanommeslaeghe, K., E. Hatcher, C. Acharya, et al.

2010 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): 671–690.

Vanommeslaeghe, Kenno, Olgun Guvench, and Alexander D. MacKerell

2014 Molecular Mechanics. *Current Pharmaceutical Design* 20(20): 3281–3292.

Vigna, Sebastiano

2014 A Weighted Correlation Index for Rankings with Ties. ArXiv:1404.3325 [Cs]. <http://arxiv.org/abs/1404.3325>.

Volkamer, Andrea, Daniel Kuhn, Friedrich Rippmann, and Matthias Rarey

2012 DoGSiteScorer: A Web Server for Automatic Binding Site Prediction, Analysis and Druggability Assessment. *Bioinformatics* 28(15): 2074–2075.

Wadood, Abdul, Mehreen Ghufuran, Syed Fahad Hassan, et al.

2017 In Silico Identification of Promiscuous Scaffolds as Potential Inhibitors of 1-Deoxy-d-Xylulose 5-Phosphate Reductoisomerase for Treatment of *Falciparum* Malaria. *Pharmaceutical Biology* 55(1): 19–32.

Wallace, Iain M., O'Sullivan Orla, and Desmond G. Higgins

2005 Evaluation of Iterative Alignment Algorithms for Multiple Alignment. *Bioinformatics* 21(8): 1408–1414.

Wang, Junmei, Wei Wang, Peter A. Kollman, and David A. Case

2006 Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *Journal of Molecular Graphics and Modelling* 25(2): 247–260.

Wang, Junmei, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case
2004 Development and Testing of a General Amber Force Field. *Journal of Computational Chemistry* 25(9): 1157–1174.

Wang, Renxiao, Luhua Lai, and Shaomeng Wang
2002 Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *Journal of Computer-Aided Molecular Design* 16(1): 11–26.

Ward, Simon E., and Paul Beswick
2014 What Does the Aromatic Ring Number Mean for Drug Design? *Expert Opinion on Drug Discovery* 9(9): 995–1003.

Waterhouse, Andrew M., James B. Procter, David M. A. Martin, Michèle Clamp, and Geoffrey J. Barton
2009 Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics* 25(9): 1189–1191.

Webb, Benjamin, and Andrej Sali
2014 Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics* 47: 5.6.1-32.

2016 Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]* 54: 5.6.1-5.6.37.

Whelan, S., and N. Goldman
2001 A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution* 18(5): 691–699.

White, N
1999 Antimalarial Drug Resistance and Combination Chemotherapy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 354(1384): 739–749.

WHO | 10 Facts on Malaria
N.d. WHO. <http://www.who.int/features/factfiles/malaria/en/>, accessed April 10, 2017.

WHO | Malaria Control Improves for Vulnerable in Africa, but Global Progress off-Track
N.d. WHO. <http://www.who.int/mediacentre/news/releases/2016/malaria-control-africa/en/>, accessed April 10, 2017.

WHO | Overview of Malaria Treatment
N.d. WHO. <http://www.who.int/malaria/areas/treatment/overview/en/>, accessed April 12, 2017.

WHO | Tables of Malaria Vaccine Projects Globally

N.d. WHO. http://www.who.int/immunization/research/development/Rainbow_tables/en/, accessed April 11, 2017.

WHO | World Malaria Report 2016

N.d. WHO. <http://www.who.int/malaria/publications/world-malaria-report-2016/report/en/>, accessed April 10, 2017.

Wiesner, Jochen, Christina Ziemann, Martin Hintz, et al.

2016 FR-900098, an Antimalarial Development Candidate That Inhibits the Non-Mevalonate Isoprenoid Biosynthesis Pathway, Shows No Evidence of Acute Toxicity and Genotoxicity. *Virulence* 7(6): 718–728.

Wiley, Jessica D., Emilio F. Merino, Priscilla M. Krai, et al.

2015 Isoprenoid Precursor Biosynthesis Is the Essential Metabolic Role of the Apicoplast during Gametocytogenesis in *Plasmodium Falciparum*. *Eukaryotic Cell* 14(2): 128–139.

Winzeler, Elizabeth Ann

2008 Malaria Research in the Post-Genomic Era. *Nature* 455(7214): 751–756.

Wirth, Dyann F.

2002 The Parasite Genome: Biological Revelations. *Nature* 419(6906): 495–496.

Wlodawer, Alexander, Wladek Minor, Zbigniew Dauter, and Mariusz Jaskolski

2008 Protein Crystallography for Non-Crystallographers, or How to Get the Best (but Not More) from Published Macromolecular Structures. *The FEBS Journal* 275(1): 1–21.

Xiong, Jin

2006 *Essential Bioinformatics*. New York: Cambridge University Press.

Xue, Jian, Jiasheng Diao, Guobin Cai, et al.

2012 Antimalarial and Structural Studies of Pyridine-Containing Inhibitors of 1-Deoxyxylulose-5-Phosphate Reductoisomerase. *ACS Medicinal Chemistry Letters* 4(2): 278–282.

Yajima, Shunsuke, Takamasa Nonaka, Tomohisa Kuzuyama, Haruo Seto, and Kanju Ohsawa

2002 Crystal Structure of 1-Deoxy-D-Xylulose 5-Phosphate Reductoisomerase Complexed with Cofactors: Implications of a Flexible Loop Movement upon Substrate Binding. *The Journal of Biochemistry* 131(3): 313–317.

Zhao, Lishan, Wei-chen Chang, Youli Xiao, Hung-wen Liu, and Pinghua Liu

2013 Methylerythritol Phosphate Pathway of Isoprenoid Biosynthesis. *Annual Review of Biochemistry* 82: 497–530.

Zheng, Heping, Katarzyna B Handing, Matthew D Zimmerman, et al.

2015 X-Ray Crystallography over the Past Decade for Novel Drug Discovery – Where Are We Heading Next? *Expert Opinion on Drug Discovery* 10(9): 975–989.

Zhou, Hongyi, and Yaoqi Zhou

2002 Distance-Scaled, Finite Ideal-Gas Reference State Improves Structure-Derived Potentials of Mean Force for Structure Selection and Stability Prediction. *Protein Science : A Publication of the Protein Society* 11(11): 2714–2726.

Zhu, Tian, Shuyi Cao, Pin-Chih Su, et al.

2013 Hit Identification and Optimization in Virtual Screening: Practical Recommendations Based Upon a Critical Literature Analysis. *Journal of Medicinal Chemistry* 56(17): 6560–6572.

APPENDIX

A. Motif logos

	Logo	E-value	Sites	Width		Logo	E-value	Sites	Width
1.		1.5e-254	18	20	12.		2.1e-099	18	20
2.		2.2e-195	18	20	13.		1.3e-077	8	20
3.		4.5e-179	18	20	14.		3.2e-074	8	20
4.		1.4e-161	18	20	15.		2.9e-061	8	20
5.		2.1e-144	18	20	16.		1.4e-043	8	16
6.		8.2e-142	18	20	17.		2.5e-037	12	16
7.		4.7e-137	18	20	18.		7.9e-034	16	7
8.		8.5e-137	17	20	19.		1.1e-029	8	12
9.		4.9e-126	18	20	20.		8.2e-026	8	9
10.		1.4e-110	17	20	21.		1.6e-019	8	9
11.		7.6e-104	15	20					

B. *Plasmodium* Crystal structures present in the PDB database (August 2017)

PDB ID	Resolution	Mis_Residues	Chain_A	Chain_B	Ligand in active site	Cofactor
3au8	1.86	172	1-76-291-299-485-486-487-488	1-76-297-298-299-485-486-487-488	No	Yes
3au9	1.90	156	1-76-487-488	1-76-487-486	Yes	Yes
3aua	2.15	156	1-76-487-488	1-76-487-487	Yes	Yes
3wq q	2.25	156	1-76-487-488	1-76-487-488	Yes	Yes
3wqr	1.97	156	1-76-487-488	1-76-487-488	Yes	Yes
3wqs	2.35	78	0	1-76, 487-488	Yes	Yes
4gae	2.30	26	63-70-487-488	63-76-487-488	Yes	Yes
4kp7	2.00	43	62-76-486-487-488	62-76 292-297-485-486-487-488	Yes	Yes
4y67	1.60	24	67-76 487-488	67-76-487-488	Yes	No
4y6p	1.90	24	67-76 487-488	67-76-487-489	Yes	No
4y6r	1.90	24	67-76 487-488	67-76-487-490	Yes	No
4y6s	2.10	24	67-76 487-488	67-76-487-491	Yes	No
5JAZ	1.40	24	67-76 487-488	67-76-487-492	Yes	No
5jbi	1.70	24	67-76 487-488	67-76-487-493	Yes	No
5jc1	1.65	24	67-76 487-488	67-76-487-494	Yes	No
5jmp	1.70	24	67-76 487-488	67-76-487-495	Yes	No
5jm w	1.55	24	67-76 487-488	67-76-487-496	Yes	No
5jnl	1.60	24	67-76 487-488	67-76-487-497	Yes	No
5jo0	1.80	24	67-76 487-488	67-76-487-498	Yes	No

C. Python script to calculate Dope-Z score using modeller and rank protein accordingly

```

from modeller import *
from modeller.scripts import complete_pdb
env = environ()
env.libs.topology.read(file='${LIB}/top_heav.lib')
env.libs.parameters.read(file='${LIB}/par.lib')
# Read a model previously generated by Modeller's automodel class

import os
score_dict = {}
for file_name in os.listdir('.'):
    if not file_name.endswith('.pdb'):
        continue

mdl = complete_pdb(env, file_name)

```

```
score = mdl.assess_normalized_dope()
score_dict[file_name] = score
```

```
with open('model_assess_zDOPE.txt', 'w') as report_file:
    for i, v in sorted(score_dict.iteritems(), key=lambda x: x[1]):
        report_file.write("%s \t %s\n" % (i.ljust(20), v))
```

D. Python script to create vina file for high throughput virtual screening.

```
import os
Ligand_files = os.listdir('../Ligand')
print "Ligands in", len(Ligand_files)

PDB_files = os.listdir('../Target')
print "Receptors in", PDB_files, len(PDB_files), "Receptor(s)"

#go_vina = raw_input("Create vina files (y or n:)?")
#if go_vina == "y":
for ligand in Ligand_files:
    if ".pdb" in ligand:
        ligand_name = ligand[:-6]
        for PDB in PDB_files:
            vina_name = PDB+"_"+ligand_name+".vina"
            with
open("/mnt/lustre/users/bdiallo/Dockings/5JAZB_SANCDDB_both_sites/Vina/"+vina_name, "w") as vw:

        vw.writelines(["receptor=/mnt/lustre/users/bdiallo/Dockings/5JAZB_SANCDDB_both_sites/Target/"+PDB+"
\n"])

        vw.writelines(["ligand=/mnt/lustre/users/bdiallo/Dockings/5JAZB_SANCDDB_both_sites/Ligand/"+ligand_n
ame+".pdbqt", "\n"])

        vw.writelines(["out=/mnt/lustre/users/bdiallo/Dockings/5JAZB_SANCDDB_both_sites/Out/"+vina_name+"a
ll.pdbqt", "\n"])

        vw.writelines(["log=/mnt/lustre/users/bdiallo/Dockings/5JAZB_SANCDDB_both_sites/Log/"+vina_name+"a
l.log", "\n"])

        #Fix your grid center
        vw.writelines(["center_x=-10", "\n"])
        vw.writelines(["center_y=30", "\n"])
        vw.writelines(["center_z=-19", "\n"])
        #Spacing
        vw.writelines(["size_x=30", "\n", "size_y=30"])
        vw.writelines(["\n", "size_z=30", "\n"])
        vw.writelines(["cpu=12", "\n", "exhaustiveness=192"])
```

E. Python script for counting non-hydrogen atoms in mol2 file format

```
import os
def count_non_H(molecule_mol2):
    """
    Count the number of non H atom in a mol2 file
    """
    file = open(molecule_mol2, "r")

    file = file.readlines()
    mol = 0          #tracking when line is still in @<TRIPOS>ATOM section
    for line in file:
        #print line
        if "@<TRIPOS>ATOM" in line:
            non_H = 0
            mol = 1
        if "@<TRIPOS>BOND" in line:
            mol = 0
        if mol == 1:
            #print line[8:11]
            if line[8:11] != " H ":
                non_H +=1

    return non_H - 1      # -1 because the first atom doesnt include any h atom and is not part of the molecule

for file in os.listdir("."):
    if file.endswith(".mol2"):
        print file, count_non_H(file)
```

F. Perl script for analysing extraction all ligand interaction from Discovery Studio

```
#!/usr/bin/perl -w
#
# File: CountHydrogenBonds.pl
#
# Function: Counts the number of hydrogen bonds in a trajectory file.
#
# Syntax: <perl> CountHydrogenBonds.pl
#
# Product: Scripting, MdmDiscoveryScript
#
# Copyright (C) 2013 by Dassault Systèmes Biovia Corp., All rights reserved.
# _____

# Modification. Dr. Kevin Lobb
# Bakary N'tji Diallo
# Original adapted to extract all protein ligand interactions from docking results.

use strict;
use MdmDiscoveryScript;
```

```

#Opening protein
opendir(PROT,"C:\\Users\\Diallo-Pc\\Google Drive\\M_T\\docking\\5JAZB_SANCDDB_both_sites\\Target");
my @proteins = readdir PROT;
closedir PROT;

#Opening ligands
opendir(LIGANDS,"C:\\Users\\Diallo-Pc\\Google
Drive\\M_T\\docking\\5JAZB_SANCDDB_both_sites\\Out\\5JAZB\\Best_5JAZ_B\\");
my @ligands = readdir LIGANDS;
closedir LIGANDS;

foreach my $protein (@proteins)
{
    if($protein =~ m/.pdbqt/)
    {
        my $proteinname = $protein;
        $proteinname =~ s/_apo.pdbqt//;
        print "$proteinname\n";
        open(my $tee, '>', 'C:\\Users\\Diallo-Pc\\Google
Drive\\M_T\\docking\\5JAZB_SANCDDB_both_sites\\Target\\'.$proteinname.'_ligands_interactions.txt'); #opening
an output file

        foreach my $ligand (@ligands)
        {
            if($ligand =~ m/.pdbqt/ and $ligand =~ m/$proteinname/)
            {
                printf $tee "-----\n-----\n$ligand\n";

                #Inserting the protein
                my $document = Mdm::Document::Create();
                $document->Insert("C:\\Users\\Diallo-Pc\\Google
Drive\\M_T\\docking\\5JAZB_SANCDDB_both_sites\\Target\\$protein");

                #Inserting the ligand and putting it into focus
                $document->Insert("C:\\Users\\Diallo-Pc\\Google
Drive\\M_T\\docking\\5JAZB_SANCDDB_both_sites\\Out\\5JAZB\\Best_5JAZ_B\\$ligand");

                #Using the CreateLigandNonbondMonitor function
                my $monitor = $document->CreateLigandNonbondMonitor( "True", "False", "True");
                $document->UpdateViews();

                # Analyse the identified (favorable/unfavorable) interactions.
                my $nonbonds = $monitor->Nonbonds;
                my $count = $nonbonds->Count;
                my $favorableNonbonds = $monitor->FavorableNonbonds;
                my $favorableCount = $favorableNonbonds->Count;
                my $unfavorableCount = $monitor->UnfavorableNonbonds->Count;

                printf $tee "\nFound %d non-bond interactions (total):", $count;
                printf $tee "\n %d of these are favorable interactions (such as H-bonds)", $favorableCount;
                printf $tee "\n %d of these are unfavorable interactions (such as bumps).", $unfavorableCount;
                printf $tee "\n";
            }
        }
    }
}

```

```

#Analyse all type of interaction
printf $tee "\nAnalyse all non-bond interaction:";
printf $tee "\nThe NonbondTypes property can be used to identify all interaction types of a non-bond.\n";
foreach my $nonbond (@$nonbonds)
{
  printf $tee "- %s (%s) and %s (%s):",
    $nonbond->FromSite->Name,
    $nonbond->FromSite->ChemistryName,
    $nonbond->ToSite->Name,
    $nonbond->ToSite->ChemistryName;

  my $nonbondTypes = $nonbond->NonbondTypes;
  foreach my $type (@$nonbondTypes) {
    printf $tee "$type";
  }

  printf $tee "\n"; }

#Use sleep because code can go faster than display
sleep(4);
$document->Close();

}
}
}
}

close my $tee;

```

G. Python script for analysing protein-ligands' interactions

```

# Author: Bakary N'tji Diallo
# Date: October 2017

```

```

import os
import operator

```

```

data = open("5JAZB_ligands_1_interactions.txt", "r")      #interactions file from Discovery Studio
data = data.readlines()
#List of residues and their residues in PfDXR
amino_acids = ["MN502", 'PRO77', 'ILE78', 'ASN79', 'VAL80', 'ALA81', 'ILE82', 'PHE83', 'GLY84', 'SER85', 'THR86',
'GLY87', 'SER88', 'ILE89', 'GLY90', 'THR91', 'ASN92', 'ALA93', 'LEU94', 'ASN95', 'ILE96', 'ILE97', 'ARG98', 'GLU99',
'CYS100', 'ASN101', 'LYS102', 'ILE103', 'GLU104', 'ASN105', 'VAL106', 'PHE107', 'ASN108', 'VAL109', 'LYS110',
'ALA111', 'LEU112', 'TYR113', 'VAL114', 'ASN115', 'LYS116', 'SER117', 'VAL118', 'ASN119', 'GLU120', 'LEU121',
'TYR122', 'GLU123', 'GLN124', 'ALA125', 'ARG126', 'GLU127', 'PHE128', 'LEU129', 'PRO130', 'GLU131', 'TYR132',
'LEU133', 'CYS134', 'ILE135', 'HIS136', 'ASP137', 'LYS138', 'SER139', 'VAL140', 'TYR141', 'GLU142', 'GLU143',
'LEU144', 'LYS145', 'GLU146', 'LEU147', 'VAL148', 'LYS149', 'ASN150', 'ILE151', 'LYS152', 'ASP153', 'TYR154', 'LYS155',
'PRO156', 'ILE157', 'ILE158', 'LEU159', 'CYS160', 'GLY161', 'ASP162', 'GLU163', 'GLY164', 'MET165', 'LYS166',
'GLU167', 'ILE168', 'CYS169', 'SER170', 'SER171', 'ASN172', 'SER173', 'ILE174', 'ASP175', 'LYS176', 'ILE177', 'VAL178',
'ILE179', 'GLY180', 'ILE181', 'ASP182', 'SER183', 'PHE184', 'GLN185', 'GLY186', 'LEU187', 'TYR188', 'SER189',
'THR190', 'MET191', 'TYR192', 'ALA193', 'ILE194', 'MET195', 'ASN196', 'ASN197', 'LYS198', 'ILE199', 'VAL200',
'ALA201', 'LEU202', 'ALA203', 'ASN204', 'LYS205', 'GLU206', 'SER207', 'ILE208', 'VAL209', 'SER210', 'ALA211',
'GLY212', 'PHE213', 'PHE214', 'LEU215', 'LYS216', 'LYS217', 'LEU218', 'LEU219', 'ASN220', 'ILE221', 'HIS222', 'LYS223',

```

```
'ASN224', 'ALA225', 'LYS226', 'ILE227', 'ILE228', 'PRO229', 'VAL230', 'ASP231', 'SER232', 'GLU233', 'HIS234', 'SER235',
'ALA236', 'ILE237', 'PHE238', 'GLN239', 'CYS240', 'LEU241', 'ASP242', 'ASN243', 'ASN244', 'LYS245', 'VAL246',
'LEU247', 'LYS248', 'THR249', 'LYS250', 'CYS251', 'LEU252', 'GLN253', 'ASP254', 'ASN255', 'PHE256', 'SER257',
'LYS258', 'ILE259', 'ASN260', 'ASN261', 'ILE262', 'ASN263', 'LYS264', 'ILE265', 'PHE266', 'LEU267', 'CYS268', 'SER269',
'SER270', 'GLY271', 'GLY272', 'PRO273', 'PHE274', 'GLN275', 'ASN276', 'LEU277', 'THR278', 'MET279', 'ASP280',
'GLU281', 'LEU282', 'LYS283', 'ASN284', 'VAL285', 'THR286', 'SER287', 'GLU288', 'ASN289', 'ALA290', 'LEU291',
'LYS292', 'HIS293', 'PRO294', 'LYS295', 'TRP296', 'LYS297', 'MET298', 'GLY299', 'LYS300', 'LYS301', 'ILE302', 'THR303',
'ILE304', 'ASP305', 'SER306', 'ALA307', 'THR308', 'MET309', 'MET310', 'ASN311', 'LYS312', 'GLY313', 'LEU314',
'GLU315', 'VAL316', 'ILE317', 'GLU318', 'THR319', 'HIS320', 'PHE321', 'LEU322', 'PHE323', 'ASP324', 'VAL325',
'ASP326', 'TYR327', 'ASN328', 'ASP329', 'ILE330', 'GLU331', 'VAL332', 'ILE333', 'VAL334', 'HIS335', 'LYS336', 'GLU337',
'CYS338', 'ILE339', 'ILE340', 'HIS341', 'SER342', 'CYS343', 'VAL344', 'GLU345', 'PHE346', 'ILE347', 'ASP348', 'LYS349',
'SER350', 'VAL351', 'ILE352', 'SER353', 'GLN354', 'MET355', 'TYR356', 'TYR357', 'PRO358', 'ASP359', 'MET360',
'GLN361', 'ILE362', 'PRO363', 'ILE364', 'LEU365', 'TYR366', 'SER367', 'LEU368', 'THR369', 'TRP370', 'PRO371',
'ASP372', 'ARG373', 'ILE374', 'LYS375', 'THR376', 'ASN377', 'LEU378', 'LYS379', 'PRO380', 'LEU381', 'ASP382',
'LEU383', 'ALA384', 'GLN385', 'VAL386', 'SER387', 'THR388', 'LEU389', 'THR390', 'PHE391', 'HIS392', 'LYS393',
'PRO394', 'SER395', 'LEU396', 'GLU397', 'HIS398', 'PHE399', 'PRO400', 'CYS401', 'ILE402', 'LYS403', 'LEU404',
'ALA405', 'TYR406', 'GLN407', 'ALA408', 'GLY409', 'ILE410', 'LYS411', 'GLY412', 'ASN413', 'PHE414', 'TYR415',
'PRO416', 'THR417', 'VAL418', 'LEU419', 'ASN420', 'ALA421', 'SER422', 'ASN423', 'GLU424', 'ILE425', 'ALA426',
'ASN427', 'ASN428', 'LEU429', 'PHE430', 'LEU431', 'ASN432', 'ASN433', 'LYS434', 'ILE435', 'LYS436', 'TYR437',
'PHE438', 'ASP439', 'ILE440', 'SER441', 'SER442', 'ILE443', 'ILE444', 'SER445', 'GLN446', 'VAL447', 'LEU448', 'GLU449',
'SER450', 'PHE451', 'ASN452', 'SER453', 'GLN454', 'LYS455', 'VAL456', 'SER457', 'GLU458', 'ASN459', 'SER460',
'GLU461', 'ASP462', 'LEU463', 'MET464', 'LYS465', 'GLN466', 'ILE467', 'LEU468', 'GLN469', 'ILE470', 'HIS471',
'SER472', 'TRP473', 'ALA474', 'LYS475', 'ASP476', 'LYS477', 'ALA478', 'THR479', 'ASP480', 'ILE481', 'TYR482',
'ASN483', 'LYS484', 'HIS485', 'ASN486']
```

```
interactions = {} #Dictionary containing the ligand and its interactions:
# interactions[ligand][first_res, second_res] = interaction_type
```

```
#Counting residues interacting the most with ligands, hotspot of binding
```

```
residues_count = {} #dictionary containing the residue and the number of interaction
```

```
for line in data:
```

```
if ".pdbqt" in line: #Every new ligand has .pdbqt in the name
```

```
ligand = line[16:24] #extracting the ligand name
```

```
interactions[ligand] = {} #creation a dictionary containing the different type of interactions
```

```
if line.startswith("- "):
```

```
#Identifying the two interacting residues (Ligand-Protein)
```

```
first_res = line.split("(")[0][line.index("("):].strip() #First Residue:atom in the interaction
```

```
second_res = line.split(" ")[1]
```

```
second_res = second_res[second_res.index(":")+1:second_res.index(")"].strip()
```

```
interaction_type = line.split(" ")[-1] #The type of interaction
```

```
#counting the number of time each residues is implied in the interactions
```

```
for res in amino_acids:
```

```
if res in first_res:
```

```
if ":" in first_res[1:7]:
```

```
first_res = first_res[1:6]
```

```
else:
```

```
first_res = first_res[1:7]
```

```
residues_count[first_res] = residues_count.get(first_res, 0) + 1 #updating residues_count
```

```
if res in second_res:
```

```
second_res = second_res.split(":")[0]
```

```
residues_count[second_res] = residues_count.get(second_res, 0) + 1 #updating residues_count
```

```
interactions[ligand][first_res, second_res] = interaction_type
```



```

#Printing the residues and their count
print "Residues in protein and how much they are implied in interaction with the ligands, binding hotspots"
for residue in residues_count:
    for aa in amino_acids:
        if aa in residue:
            print residue, residues_count[residue]

#Sorting the number of interaction from the residue that interacts the most to the lowest
print "\nHighest to lowest interaction per residue"
x = residues_count
sorted_x = sorted(x.items(), key=operator.itemgetter(1))
for e in sorted_x[::-1]:
    print e[0], e[1]

#Ligand that match fosmidomycin-like inhibitors binding pattern

hydroxamate = ["ASP231", "GLU233", "GLU315"] #: Binding hydroxamate group of fosmidomycin Divalent metal
cation coordination (pentacoordinate trigonal bipyramidal geometry).

phosphonate = ["SER269", "SER270", "SER306", "ASN311", "LYS312", "HIS293"] # binding phosphonate moiety

NADPH = ["THR86", "GLY87", "SER88", "ILE89", "ASN115", "LYS116", "SER117", "GLU206", "GLY299"] #: NADPH
binding residues

print "\nLigand matching known inhibitor binding pattern"
for ligand in interactions:
    bisubstrate_score1 = [0, 0, 0] #binding hydroxamate, phosphonate, NADPH
    fosmidomycin_motif = [0, 0] #binding hydroxamate and phosphonate
    for interaction in interactions[ligand]:
        interaction_type = interactions[ligand][interaction[0], interaction[1]]
        res1, res2 = interaction[0], interaction[1]
        tot_res = res1 + res2 #all residues in the interaction

    if not bisubstrate_score1[0]: #
        bisubstrate_score1[0] = any(aa in tot_res for aa in hydroxamate)
    if any([aa in tot_res for aa in hydroxamate]):
        fosmidomycin_motif[0] += 1
    if not bisubstrate_score1[1]:
        bisubstrate_score1[1] = any(aa in tot_res for aa in phosphonate)
    if any([aa in tot_res for aa in phosphonate]):
        fosmidomycin_motif[1] += 1

    if not bisubstrate_score1[2]:
        bisubstrate_score1[2] = any(aa in tot_res for aa in NADPH)

    # if bisubstrate_score1 == [True, False, True]: #compound binding to at least one residue in each
group
    # print "Potential bisubstrate: hydroxamate residues + NADPH", ligand, interactions[ligand]
    #if bisubstrate_score1 == [True, True, True]: #compound binding to at least one residue in each
group
    # print "Potential bisubstrate: ", ligand, interactions[ligand]

```

```

    if fosmidomycin_motif[1] > 2 and fosmidomycin_motif[0] > 2:           #compound binding to at least
one residue in each group
    print "Potential match to fosmidomycin: ", ligand, interactions[ligand]

#Number of hydrogen bond per ligand

print "\nHydrogen bonding"
interactions[ligand][first_res, second_res] = interaction_type
ligand_Hbond = {}
for ligand in interactions:
    total_H_bond = 0
    for interaction in interactions[ligand]:
        interaction_type = interactions[ligand][interaction[0], interaction[1]]
        if "conventionalHBond" in interaction_type:
            total_H_bond += 1
    ligand_Hbond[ligand] = total_H_bond
    print ligand, total_H_bond, "Hydrogen bond(s)"

##### Sorting the number of hbonding from the ligands the most to the lowest
#####
print "Highest to lowest count of hbond per ligand:"
x = ligand_Hbond
sorted_x = sorted(x.items(), key=operator.itemgetter(1))
print sorted_x[::-1]

# ##### Printing interactions for specific ligand
#####
hits = "SANC00152 SANC00236 SANC00438 SANC00339 SANC00570"
bisubstrate_hits = "SANC00615 SANC00436 SANC00556 SANC00443 SANC00562"
for ligand in interactions:
    if ligand.strip() in bisubstrate_hits:
        print ligand,
        for interaction in interactions[ligand]:
            if interaction[0] in amino_acids or interaction[1] in amino_acids:
                print interaction[0], interaction[1], interactions[ligand][interaction[0], interaction[1]]

# ##### Search for bidentate ligands chelating the metal
#####
print "#####SEARCH FOR BIDENTATE LIGANDS#####\n"
for ligand in interactions:
    for interaction in interactions[ligand]:
        interaction_type = interactions[ligand][interaction[0], interaction[1]]
        if "etalAcceptor" in interaction_type:
            print ligand, interaction_type      #if ligand doubly printed --> bidentate ligand

```

H. MD simulation job file for submission on CHPH

```

#!/bin/bash
#PBS -q normal

```

```
#PBS -l select=10:ncpus=24:mpiprocs=24
#PBS -l walltime=48:00:00
#PBS -V
#PBS -P CBB10867
#PBS -N Cl_152_cof
#PBS -e /mnt/lustre/users/bdiallo/MDs/Systems/Cl_152_cof/out.err
#PBS -o /mnt/lustre/users/bdiallo/MDs/Systems/Cl_152_cof/out.txt
#PBS -m bea
#PBS -M diallobakary4@gmail.com
```

```
module load chpc/GROMACS /v2016.1dev-noomp-openmpi-2.0.0-gcc-6.2.0
```

```
#Setting open mpi parameters
```

```
OMP_NUM_THREADS=1
```

```
NP='cat ${PBS_NODEFILE} | wc -l'
```

```
#Usage qsub -P CBB10867 MD_protein.sh
```

```
#Ensure that only a single (prepared) pdb file is in the directory
```

```
cd $PBS_O_WORKDIR
```

```
mkdir pbserr pbsout
```

```
#Performs MD on protein-ligand complex
```

```
#Energy minimization
```

```
gmx_mpi mdrun -v -deffnm em
```

```
#Check minimization
```

```
gmx_mpi energy -f em.edr -o potential_energy.svg > potential_energy.txt
```

```
#NVT equilibration
```

```
gmx_mpi grompp -f nvt.mdp -c em.gro -p topol.top -n index.ndx -o nvt.tpr
```

```
mpirun -np $NP -machinefile ${PBS_NODEFILE} gmx_mpi mdrun -cpi -maxh 48 -deffnm nvt
```

```
#Check NVT equilibration
```

```
echo -e "16\n0\n" | gmx energy -f nvt.edr -o temperature.svg > temperature.txt
```

```
#NPT equilibration
```

```
gmx_mpi grompp -f npt.mdp -c nvt.gro -t nvt.cpt -p topol.top -n index.ndx -o npt.tpr
```

```
mpirun -np $NP -machinefile ${PBS_NODEFILE} gmx_mpi mdrun -cpi -maxh 48 -deffnm npt
```

```
#Check NPT equilibration
```















```
echo -e "17\n0\n" | gmx energy -f npt.edr -o pressure.svg > pressure.txt
```

```
#MD run
```

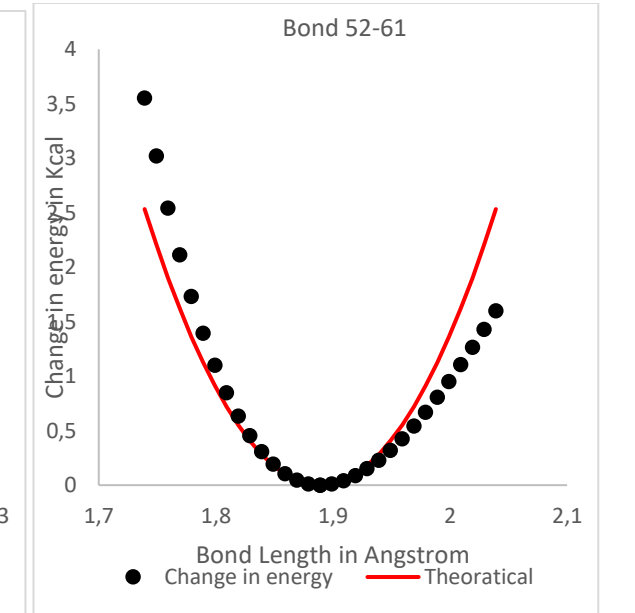
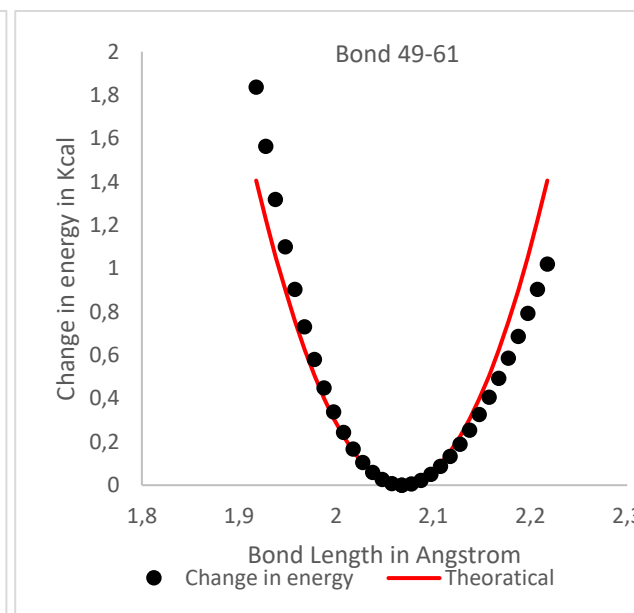
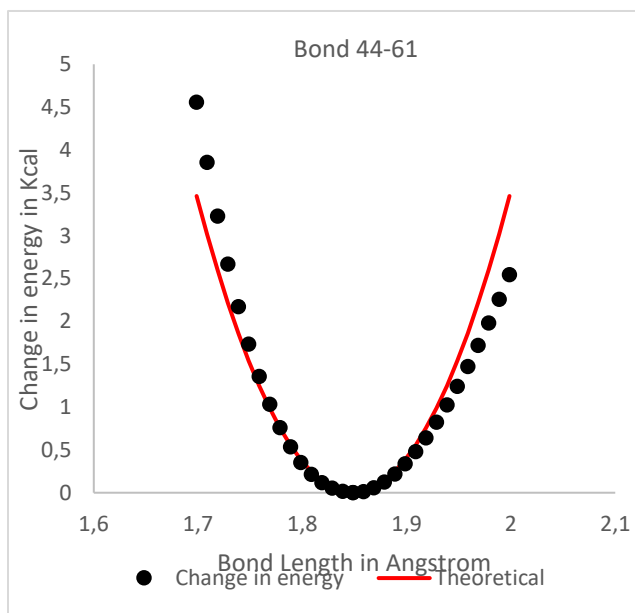
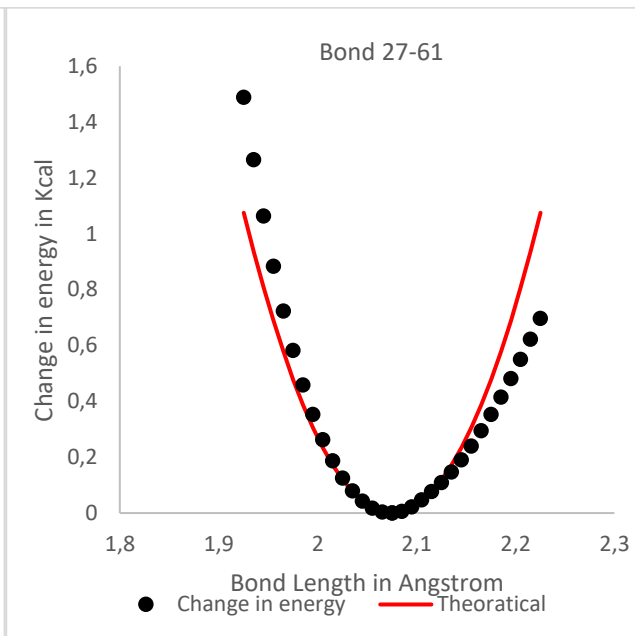
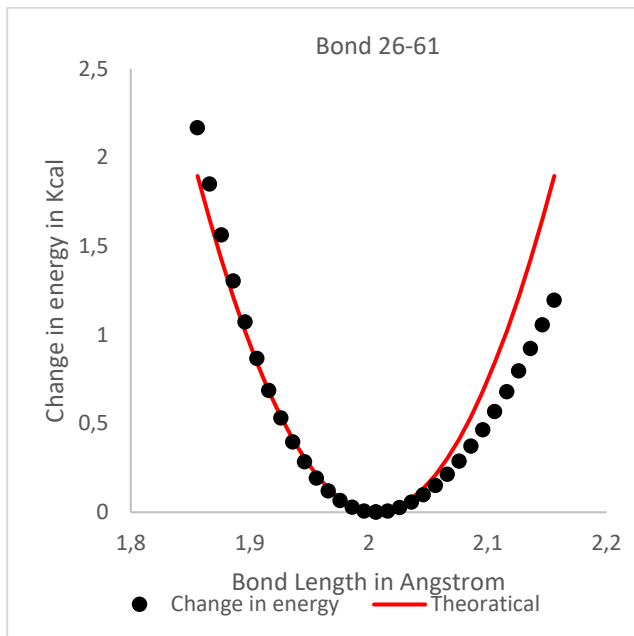
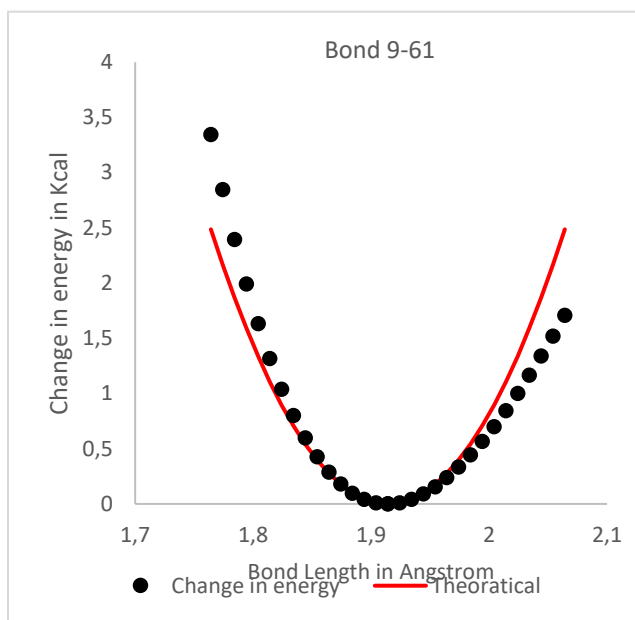
```
gmx_mpi grompp -f md.mdp -c npt.gro -t npt.cpt -p topol.top -n index.ndx -o md_0_1.tpr
```

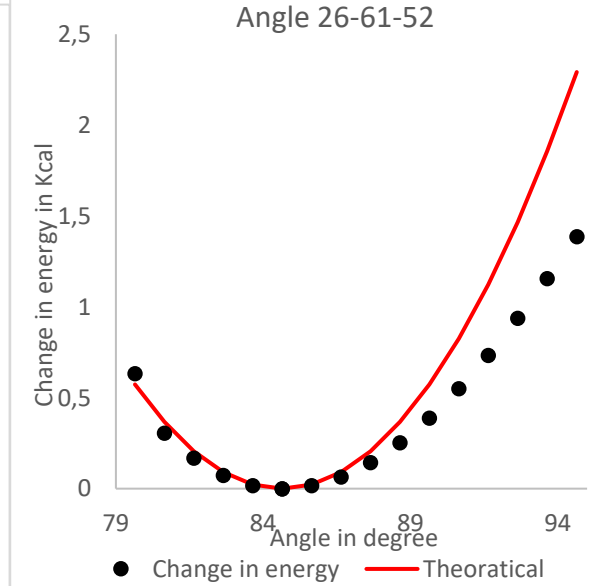
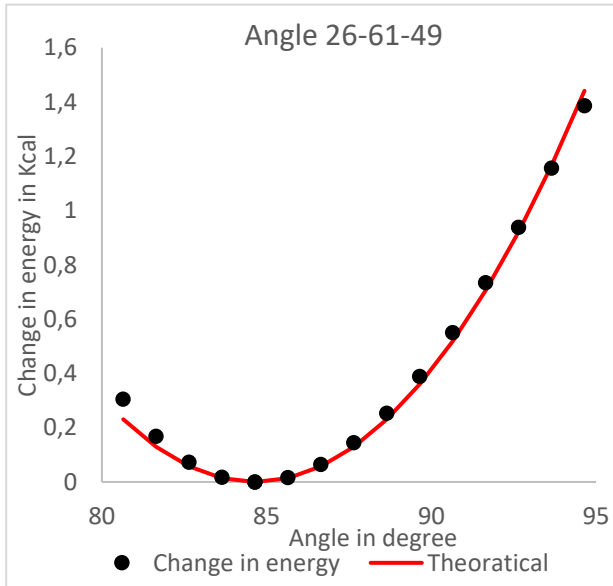
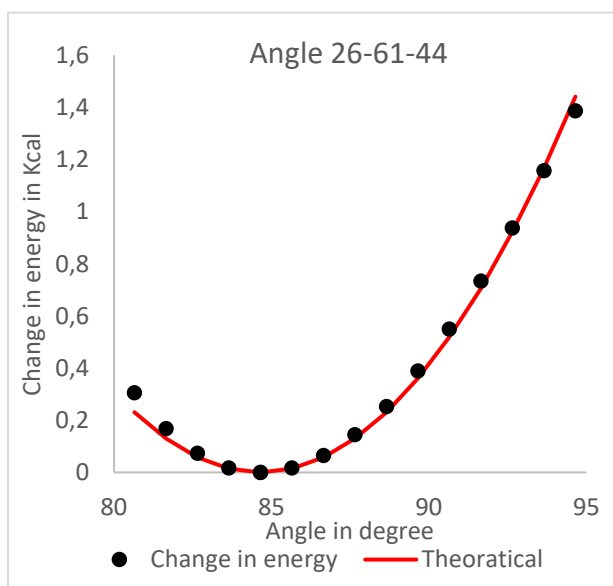
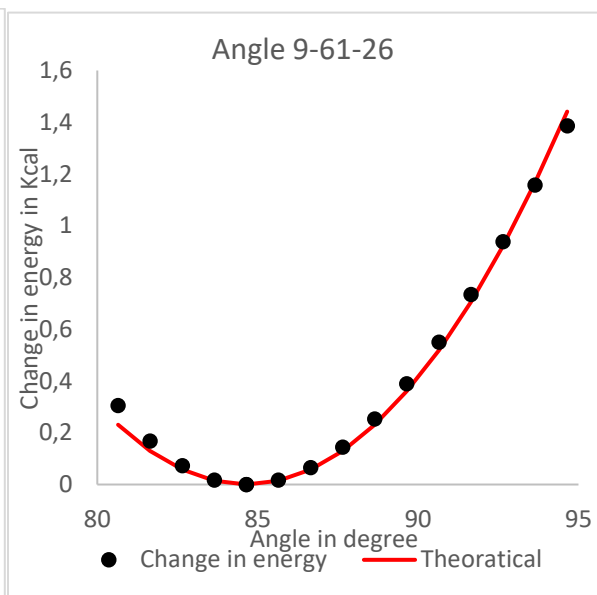
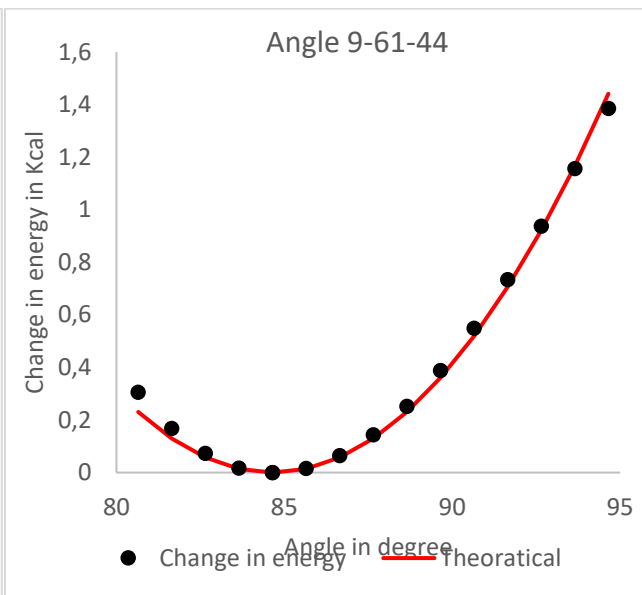
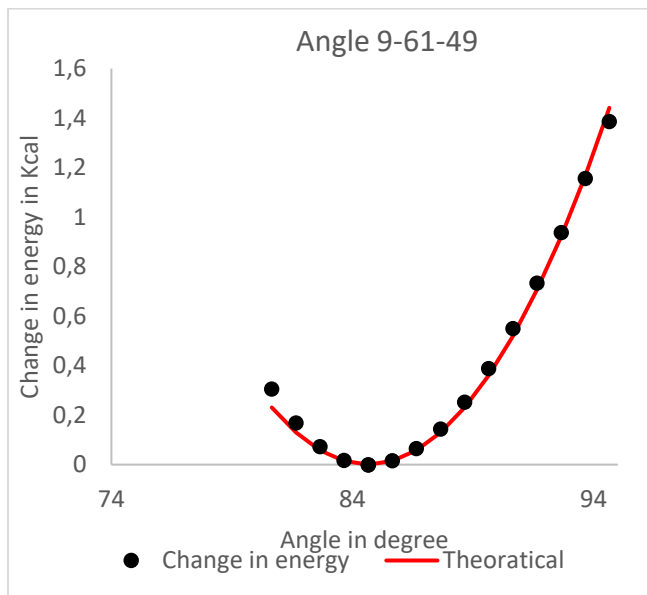
```
mpirun -np $NP -machinefile ${PBS_NODEFILE} gmx_mpi mdrun -cpi -maxh 48 -deffnm md_0_1
```

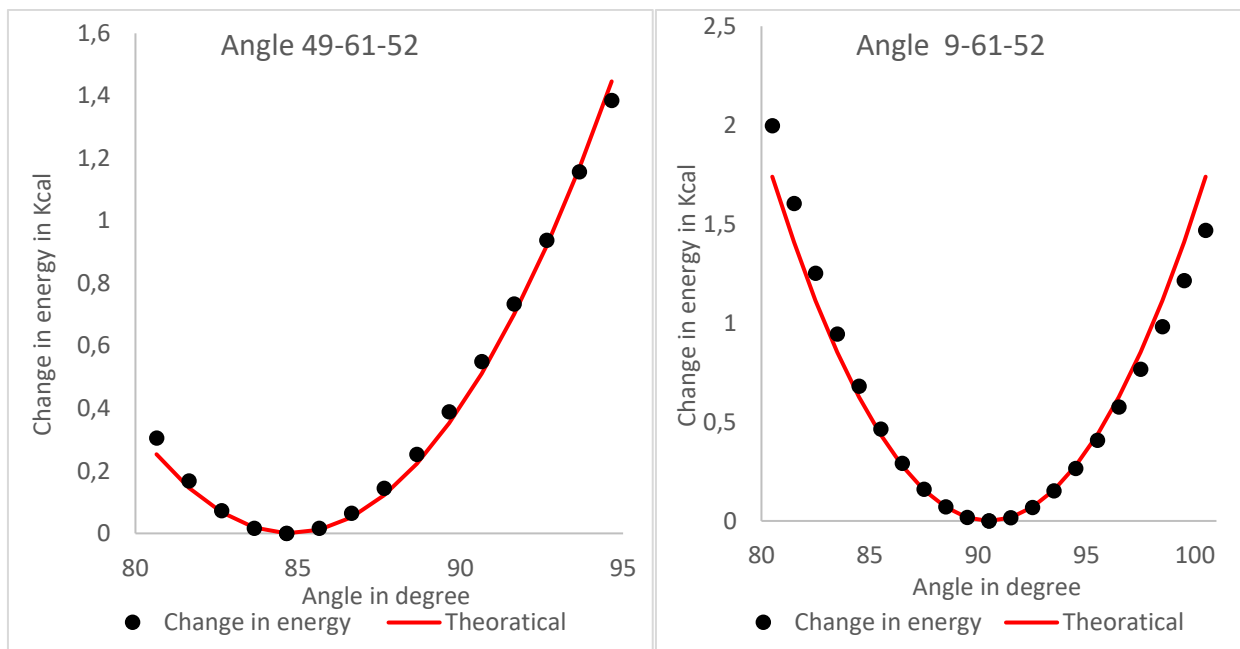
I. Table of the different predicted geometries by the Metalizer webserver for the optimized subset

Geometry  	Coordination #  	Symm. bidentates  	Angle RMSD  	Free sites  	Overlap penalty  	Score  
square pyramid	5	0	8.12	0	0.000	8.12
trigonal bipyramid	5	1	11.78	0	0.000	11.78
trigonal prismatic	5	1	15.99	1	0.093	22.89
pentagonal bipyramid	5	1	7.40	2	0.819	30.69
octahedral	5	0	8.12	1	1.625	38.00

J. Graphs used for fitting PES data to models used for the force field.







K. Molecular overlay of the optimized geometry from the ONIOM (bleu) system and the geometry in the crystal structure. The backbones of the residues were removed for clarity.

