Computer aided approaches against Human African Trypanosomiasis

A thesis submitted in fulfilment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

IN BIOINFORMATICS

of

RHODES UNIVERSITY, SOUTH AFRICA

In the Department of Biochemistry and Microbiology

Faculty of Science

by

Magambo Phillip Kimuda

February 2019

ABSTRACT

The thesis presented here is divided into two parts under a common theme that is the use of computer based tools, genomics, and *in vitro* experiments to develop innovative ways of tackling Human African Trypanosomiasis (HAT). Part I of this thesis focused on the human host genetic determinants while Part II focused on the discovery of novel chemotherapeutics against the parasite.

Part I is further sub-divided into two parts: The first involves a Candidate Gene Association Study (CGAS) on an African population to identify genetic determinants associated with disease and/or susceptibility to HAT. The second involves studying the effects of missense Single Nucleotide Variants (SNVs) on protein structure, dynamics, and function using Macrophage Migration Inhibitory Factor (MIF) as a case study.

Part II is also sub-divided into two parts: The first involves a computer based rational drug discovery of potential inhibitors against the *Trypanosoma* the folate pathway; particularly by targeting *Trypanosoma brucei* Pteridine Reductase (*Tb*PTR1) which is an enzyme used by trypanosomes to overcome *T. brucei* Dihydrofolate Reductase (*Tb*DHFR) inhibition. Lastly the derivation of CHARMM force-field parameters that can be used to accurately model the geometry and dynamics of the *T. brucei* Phosphodiesterase B1 enzyme (TbrPDEB1) bi-metallic active site center. The derived parameters were then used in MD simulations to characterise protein-ligand residue interactions that are important in TbrPDEB1 inhibition with the goal of targeting the cyclic Adenosine Monophosphate (cAMP) signalling pathway.

In the CGAS we were unable to detect any genetic associations in the Ugandan cohort analysed that passed correction for multiple testing in spite of the study being sufficiently powered. Additionally, our study found no association of the Apo lipoprotein 1 (*APOL1*) G2 allele association with protection against acute HAT that has been previously reported. Future investigations for example, Genome Wide Association Studies using larger samples sizes (>3000 cases and controls) are required.

Macrophage migration inhibitory factor (MIF) is a cytokine that is important in both innate and adaptive immunity that has been shown to play a role in *T. brucei* pathogenicity using murine models. A total of 27 missense SNVs were modelled using homology modelling to create MIF protein mutants that were investigated using *in silico* effect prediction tools, molecular dynamics (MD), Principal Component Analysis (PCA), and Dynamic Residue Network (DRN) analysis. Our results demonstrate that mutations P2Q, I5M, P16Q, L23F, T24S, T31I, Y37H, H41P, M48V, P44L, G52C, S54R, I65M, I68T, S75F, N106S, and T113S caused significant conformational changes. Further, DRN analysis showed that residues P2, T31, Y37, G52, I65, I68, S75, N106, and T113S are part of a similar local residue interaction network with functional significance. These results show how polymorphisms such as missense SNVs can affect protein conformation, dynamics, and function.

Trypanosomes are auxotrophic for folates and pterins but require them for survival. They scavenge them from their hosts. PTR1 is a multifunctional enzyme that is unique to trypanosomatids that reduces both pterins and folates. In the presence of DHFR inhibitors, PTR1 is over-expressed thus providing an escape from the effects of DHFR inhibition. Both *Tb*PTR1 and *Tb*DHFR are pharmacologically and genetically validated drug targets. In this study 5742 compounds were screened using molecular docking, and 13 promising binding modes were further analysed using MD simulations. The trajectories were analysed using RMSD, Rg, RMSF, PCA, Essential Dynamics Analysis (EDA), Molecular Mechanics Poisson–Boltzmann surface area (MM-PBSA) binding free energy calculations, and DRN analysis. The computational screening approach allowed us to identify five of the compounds, named RUBi004, RUBi007, RUBi014, RUBi016 and RUBi018 that exhibited anti-trypanosomal growth activities against trypanosomes in culture with IC_{50} values of 12.5 ± 4.8

μM, 32.4 ± 4.2 μM, 5.9 ± 1.4 μM, 28.2 ± 3.3 μM, and 9.7 ± 2.1 μM, respectively. Further when used in combination with WR99210 a known *Tb*DHFR inhibitor RUBi004, RUBi007, RUBi014 and RUBi018 showed antagonism while RUBi016 showed an additive effect. These results indicate that the four compounds might be competing with *Tb*DHFR while RUBi016 might be more specific for *Tb*PTR1. These compounds provide scaffolds that can be further optimised to improve their potency and specificity.

Lastly, using a systematic approach we derived CHARMM force-field parameters to accurately describe the TbrPDEB1 bi-metal catalytic center. For dynamics, we employed mixed bonded and non-bonded approach. We optimised the structure using a two-layer QM/MM ONIOM (B3LYP/6-31(g): UFF). The TbrPDEB1 bi-metallic center bonds, angles, and dihedrals were parameterized by fitting the energy profiles from Potential Energy Surface (PES) scans to the CHARMM potential energy function. The parameters were validated by means of MD simulations and analysed using RMSD, Rg, RMSF, hydrogen bonding, bond/angle/dihedral evaluations, EDA, PCA, and DRN analysis. The force-field parameters were able to accurately reproduce the geometry and dynamics of the TbrPDEB1 bi-metal catalytic center during MD simulations. Molecular docking was used to identify 6 potential hits, that inhibited trypanosome growth *in vitro*. The derived force-field parameters were used to simulate the 6 protein-ligand complexes with the aim of elucidating crucial protein-ligand residue interactions. Using the most potent ligand RUBi022 that had an IC_{50} of 14.96 μM we were able to identify key residue interactions that can be of use in *in silico* prediction of potential TbrPDEB1 inhibitors.

Overall we demonstrate how bioinformatics tools can complement current disease eradication strategies. Future work will focus on identifying variants identified in Genome Wide Association Studies and partnering with wet labs to carry out further enzyme-ligand activity relationship studies, structure determination or characterisation of appropriate protein-ligand complexes by crystallography, and site specific mutation studies.

DECLARATION

I declare that this thesis is my own, unaided work, unless otherwise stated. It is being submitted for the degree of Doctor of Philosophy at Rhodes University. It has not been submitted before for any degree or examination in any other university.

Signature: …………………………………

Date: ………………………………………

ACKNOWLEDGEMENTS

I owe a debt of gratitude to my PhD supervisor Prof Özlem Taştan Bishop, her guidance has been instrumental in the successful completion of this work.

I also would like to thank Prof Enock Matovu for mentoring me and giving me the opportunity to venture into the field of bioinformatics.

I would also like to take a moment to appreciate the guidance given to me by Prof Kevin Lobb particularly in teaching and guiding me through computational chemistry.

I thank Prof Heinrich Hoppe and Dusting Laming for their help in carrying out the *in vitro* inhibition screening assays and the Promega cAMP-Glow assay.

I would like to thank Dr David Brown, Dr Rowan Hatherley, Dr Vuyani Moses, Margaret Nabatanzi, Tendai Muronzi, Michael Glenister, and Caroline Ross for their support and guidance. Thank you to Olivier Sheik Amamudy for kindly providing scripts that I used to run the MIF Molecular Dynamics simulations.

Lastly, a big thank you to the H3Africa, TrypanoGEN, RUBi members, and MOBILA members.

Personal acknowledgements

I would like to appreciate the support given to me by my friends and family during my PhD journey.

I especially thank all the wonderful life friends I made at Oakdene, Vanessa Ashipala, Nnamdi Nwahara, and Dalitso Chindipha (you are too many to mention).

Thank you Mr Daniel Waiswa Kimuda and Mrs Miriam Catherine Magambo Kimuda my dear parents.

My siblings, Sarah, Simon, Saul, Joyce, Sharon, and Sam thanks for encouraging me. My other family; Simon Gwapa Kimuda, Eseza Muwanga, Andrew Byaruhanga, and Belinda Mwerinde thank you for keeping me grounded.

Acknowledgement of funding

This work was supported by the National Research Foundation (NRF) South Africa (Grant Numbers 105267 and 116F229) and the Wellcome Trust through an award to the TrypanoGEN consortium which is a member of the H3Africa initiative (Grant number 099310/Z/12/Z). The bioassays were supported by funding from the S.A. Medical Research Council. The Molecular Dockings, QM/MM optimization calculations, PES scans, and CHARMM/GROMACS Molecular Dynamics simulations were carried out on the Centre for High Performance Computing (CHPC), Cape Town, South Africa.

DEDICATION

I dedicate this work to all our mothers and to all our sisters in honour and gratitude for their sacrifices. I especially dedicate this work to the late Nyanya Halima Kidawa Magambo and Jajja Kirowa Kantono Kimuda who are my grandmothers, and my mother Miriam Catherine Magambo Kimuda, thank you for raising a generation.

TABLE OF CONTENTS

LIST OF EQUATIONS

LIST OF FIGURES

LIST OF TABLES

LIST OF WEBSERVERS

H++

http://biophysics.cs.vt.edu/

False Positive Remover

www.cbligand.org/PAINS/

METALizer

https://proteins.plus/

Multivariate Analysis of Protein Polymorphism (MAPP)

http://www.ngrl.org.uk/Manchester/page/mapp-multivariate-analysis-protein-polymorphism

Online BBB Predictor

www.cbligand.org/BBB

PhD-SNP

http://snps.biofold.org/phd-snp/phd-snp.html

Phenotyping Polymorphism (PolyPHEN)

http://genetics.bwh.harvard.edu/pph2/

Protein Analysis Through Evolutionary Relationships

(PANTHER)http://www.pantherdb.org/tools/csnpScoreForm.jsp

Protein Structure Analysis (ProSA)

https://prosa.services.came.sbg.ac.at/prosa.php

Protein Variation Effect Analyser (PROVEAN)

http://provean.jcvi.org/index.php

SNAP

https://rostlab.org/services/snap/

SNPs&GO

http://snps-and-go.biocomp.unibo.it/snps-and-go/

Sorting Intolerant From Tolerant (SIFT)

http://blocks.fhcrc.org/sift/SIFT.html

SUPPLEMENTARY INFORMATION

ACRONYMS

1000G	One thousand genomes
APOL1	Apolipoprotein 1
BBB	Blood Brain Barrier
BT	Biopterin Transporter
CA	Carbon Alpha
cAMP	cyclic Adenosine Monophosphate
CATT	Card Agglutination Test
CB	Carbon Beta
CFH	Complement Factor H
CGAS	Candidate Gene Association Study
CHARMM	Chemistry at HARvard Macromolecular Mechanics
COVAB	College of Veterinary Medicine, Animal resources, and Biosecurity
CXCR2	Chemokine Receptor 2
CXCR4	Chemokine Receptor 4
CV	Coefficient of Variation
dbSNP	Single Nucleotide Polymorphism Database
DCC	Dynamic Cross Correlation
DNA	Deoxyribonucleic Acid
DRN	Dynamic Residue Network
EDTA	Ethylenediaminetetraacetic acid
FDR	False Discovery Rate
FT	Folate Transporter
GWAS	Genome Wide Association Study
HAT	Human African Trypanosomiasis

HBA	H-bond Acceptor
HLA-G	Human Leucocyte Antigen G
HLAA	Human Leucocyte Antigen A
HP	Haptoglobin
HPR	Haptoglobin-related protein
HWE	Hardy-Weinberg Equilibrium
IFNG	Interferon Gamma
IL10	Interleukin 10
IL12B	Interleukin 12 Beta
IL12R	Interleukin 12R
IL1B	Interleukin 1 Beta
IL4	Interleukin 4
IL4R	Interleukin 4R
IL6	Interleukin 6
IL8	Interleukin 8
IMDM	Iscove’s Modified Dulbecco’s Medium
IRB	Internal Review Board
LJ	Lennard-Jones
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MAPP	Multivariate Analysis of Protein Polymorphism
MD	Molecular Dynamics
MIF	Macrophage Migration Inhibitory Factor
MT	Mutant
MWT	Molecular Weight

NADPH	Nicotinamide Adenine Dinucleotide Phosphate
NAPQI	N-acetyl-p-benzoquinone imine
NC	Net Charge
NR	Number of rotatable bonds
nsSNP	Non-synonymous Single Nucleotide Polymorphism
ONIOM	Our Own N-layer Integrated Orbital Molecular Mechanics
OR	Odds Ratio
PAINS	Pan-assay Interference compounds
PCA	Principal Component Analysis
PDE	Phosphodiesterase
PES	Potential Energy Surface
PolyPhen	Polymorphism Phenotyping
PROVEAN	Protein Variation Effect AnalyzerQM/MM Quantum Mechanics/Molecular Mechanics
RESP	Restrained Electrostatic Potential
Rg	Radius of Gyration
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RRIF	Residue-Residue Interaction Frequency
SAR	Structure Activity Relationship
SDR	Short Dehydrogenase Reductase
SIFT	Sorting Intolerant From Tolerant
SLE	Systemic Lupus Erythematosus
SNAP	Screening for Non Acceptable Polymorphisms
SNP	Single Nucleotide Polymorphism

SNV	Single Nucleotide Variant
SVM	Support Vector Machine
*Tb*DHFR	*Trypanosoma brucei* Dihydrofolate Reductase
*Tb*PTR1	*Trypanosoma brucei* Pteridine Reductase 1
TbrPDEB1	*Trypanosoma brucei* Phosphodiesterase B1
TNFA	Tumor Necrosis Factor Alpha
TNX4LB	Thioredoxin 4LB
tPSA	Topological Surface Area
UNCST	Uganda National Council of Science and Technology
UTR	Untranslated Region
vdW	van der Waals
VMD	Visual Molecular Dynamics
WHO	World Health Organisation
WT	Wildtype

LIST OF AMINO ACIDS

Table full name, three letter code, one letter code

AMINO ACID	THREE LETTER CODE	ONE LETTER CODE
ALANINE	ALA	A
ARGININE	ARG	R
ASPARAGINE	ASN	N
ASPARTIC ACID	ASP	D
ASPARAGINE OR ASPARTIC ACID	ASX	B
CYSTEINE	CYS	C
GLUTAMIC ACID	GLU	E
GLUTAMINE	GLN	Q
GLUTAMINE OR GLUTAMIC ACID	GLX	Z
GLYCINE	GLY	G
HISTIDINE	HIS	H
ISOLEUCINE	ILE	I
LEUCINE	LEU	L
LYSINE	LYS	K
METHIONINE	MET	M
PHENYLALANINE	PHE	F
PROLINE	PRO	P
SERINE	SER	S
THREONINE	THR	T
TRYPTOPHAN	TRP	W
TYROSINE	TYR	Y
VALINE	VAL	V

LIST OF RESEARCH OUTPUTS

Conference Poster Presentations

1. Magambo Phillip Kimuda, Enock Matovu, Nicky Mulder, and Özlem Taştan Bishop. "Giving protein structural context to genetic association studies: Macrophage Inhibitory Factor (MIF) and Human African Trypanosomiasis (HAT)". 25th South African Society of Biochemistry and Molecular Biology (SASBMB) Congress, 2016. East London, South Africa, 2016. (Winner of best Bioinformatics poster)

2. Magambo Phillip Kimuda, Harry Noyes, Julius Mulindwa, Özlem Tastan Bishop, and Enock Matovu. "Association between IL-6, MIF, HLA-G, TNF-A, APOL1, IFN-G, IL-8, IL-4, and IL-10 gene polymorphisms with acute and chronic Human African Trypanosomiasis in a Ugandan population". 9th H3Africa consortium meeting in Mauritius, 2016. Mauritius, 2016.

Conference Oral Presentations

1. Magambo Phillip Kimuda, Harry Noyes, Julius Mulindwa, Özlem Tastan Bishop, and Enock Matovu. "Association between IL-6, MIF, HLA-G, TNF-A, APOL1, IFN-G, IL-8, IL-4, and IL-10 gene polymorphisms with acute and chronic Human African Trypanosomiasis in a Ugandan population". 9th H3Africa consortium meeting in Mauritius, 2016. Mauritius, 2016. (Winner of best presentation)

Publications

1. Sente C, Erume J, Naigaga I, Mulindwa J, Ochwo S, Magambo PK, Namara BG, Kato CD, Sebyatika G, Muwonge K, Ocaido M. Prevalence of pathogenic free-living amoeba and other protozoa in natural and communal piped tap water from Queen Elizabeth protected area, Uganda. Infectious diseases of poverty. 2016 Dec;5(1):68.

Contribution: For this paper I set up the xenic cultivation protocol for the culture of field isolated Acanthamoeba, sample preparation, DNA extraction, and PCR amplifications.

2. Sente C, Erume J, Naigaga I, Magambo PK, Ochwo S, Mulindwa J, Namara BG, Kato CD, Sebyatika G, Muwonge K, Ocaido M. Occurrence and genetic characterisation of Acanthamoeba spp. from environmental and domestic water sources in Queen Elizabeth Protected Area, Uganda. Parasites & vectors. 2016 Dec;9(1):127.

Contribution: For this paper I set up the xenic cultivation protocol for the culture of acanthomoeba, DNA extraction, and PCR amplifications.

3. Sente C, Erume J, Naigaga I, Gabriela Namara B, Mulindwa J, Ochwo S, Magambo PK, Kato CD, Tamale A, Ocaido M. Xenic cultivation and genotyping of pathogenic free-living amoeba from public water supply sources in Uganda. New Journal of Science. 2016;2016.

Contribution: For this paper I set up the xenic cultivation protocol for the culture of Acanthamoeba, DNA extraction, and PCR amplifications.

4. Jongeneel CV, Achinike-Oduaran O, Adebiyi E, Adebiyi M, Adeyemi S, Akanle B, Aron S, Ashano E, Bendou H, Botha G, Chimusa E, Choudhury A, Donthu R, Drnevich J, Oluwadamila F, Christopher J, Liesl SH, Isewon I, Radhika S, Judit KK, Kimuda MP, Magosi L, Liudmila SM, Maslamoney S, Mbiyavanga M, Meintjes A, Mugutso N, Mpangase P, Munthali P, Nembaware V, Ndhlovu A, Odia T, Okafor A, Oladipo O, Panji S, Pillay V, Rendon G, Sengupt D, Mulder N. Assessing computational genomics skills: Our experience in the H3ABioNet African bioinformatics network. PLoS computational biology. 2017 Jun 1;13(6):e1005419.

Contribution: For this paper I was part of the team at the CBIO node where I helped in setting up a variant calling pipeline.

5. Ilboudo H, Noyes H, Mulindwa J, Kimuda MP, Koffi M, Kaboré JW, Ahouty B, Ngoyi DM, Fataki O, Simo G, Ofon E. Introducing the TrypanoGEN biobank: A valuable resource for the elimination of human African trypanosomiasis. PLoS neglected tropical diseases. 2017 Jun 1;11(6):e0005438.

Contribution. For this paper, I was involved in sample collection, DNA extraction and sample curation.

6. Kimuda MP, Noyes H, Mulindwa J, Enyaru J, Alibu VP, Sidibe I, Ngoyi DM, Hertz-Fowler C, MacLeod A, Bishop ÖT, Matovu E. No evidence for association between APOL1 kidney disease risk alleles and Human African Trypanosomiasis in two Ugandan populations. PLoS neglected tropical diseases. 2018 Feb 22;12(2):e0006300.

Contribution: For this paper, I was involved in sample collection, DNA extraction, sample curation, experiments, data analysis, and contributed to writing up the manuscript.

7. Kimuda MP, Laming D, Hoppe H, Tastan Bishop Ö. Identification of Novel Potential Inhibitors of Pteridine Reductase 1 in *Trypanosoma brucei* via Computational Structure-Based Approaches and *in Vitro* Inhibition Assays. Molecules. 2019 Jan;24(1):142.

Contribution: For this paper I performed all the computational experiments, data analysis and contributed to the writing up of the manuscript.

Part of the work presented in this thesis has been published in the following articles:

1. Kimuda MP, Noyes H, Mulindwa J, Enyaru J, Alibu VP, Sidibe I, Ngoyi DM, Hertz-Fowler C, MacLeod A, Bishop ÖT, Matovu E. No evidence for association between APOL1 kidney disease risk alleles and Human African Trypanosomiasis in two Ugandan populations. PLoS neglected tropical diseases. 2018 Feb 22;12(2):e0006300.

2. Kimuda MP, Laming D, Hoppe H, Tastan Bishop Ö. Identification of Novel Potential Inhibitors of Pteridine Reductase 1 in *Trypanosoma brucei* via Computational Structure-Based Approaches and *in Vitro* Inhibition Assays. Molecules. 2019 Jan;24(1):142.

THESIS MOTIVATION AND OVERVIEW

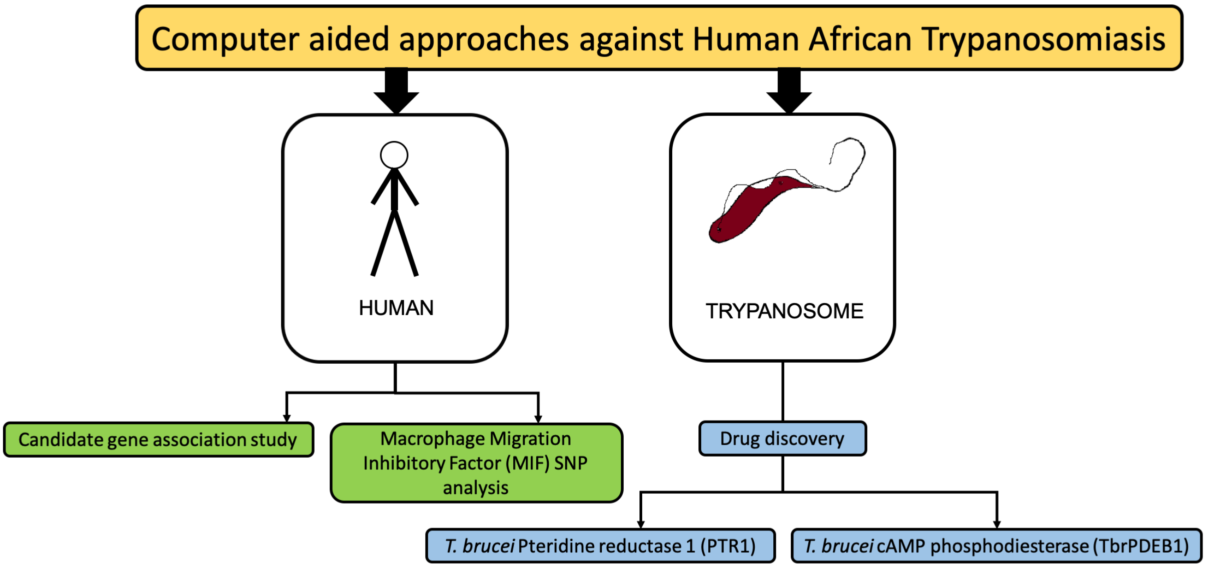
The motivation of this thesis was to use computer aided bioinformatics tools in a multi-pronged approach to develop new treatment and management options for Human African Trypanosomiasis (HAT). To achieve this goal, we used computational tools and *in vitro* experimentation explore host genetic determinants and previously under explored potential parasite drug targets.

This thesis is composed of six Chapters but is broadly divided into two thematics. Chapter 1 introduces HAT and gives a general overview of current disease and management strategies in HAT while Chapter 6 gives of a conclusion of the research work presented herein.

Part I of the thesis is composed of Chapter 2 and Chapter 3 which focus on how human host genetic factors are associated with disease susceptibility and/or resistance. Chapter 2 covers a candidate gene association study to the discover host genetic determinants associated with HAT susceptibility and/or resistance in a Ugandan population. Chapter 3 follows, covering the use of Macrophage Migration Inhibitory Factor (MIF) as a case study for how non-synonymous missense SNVs affect protein structure and function.

Part II of the thesis is composed of Chapter 4 and Chapter 5 which focus on potential parasite drug targets with the aim of developing safe and effective HAT chemotherapeutics. Chapter 4 covers the targeting of Pteridine Reductase 1 (PTR1) to make the trypanosome folate pathway a viable drug target. Chapter 5 covers the development of accurate CHARMM force-field parameters for the simulation of the bi-metal catalytic centre of *T. brucei* phosphodiesterases B1 (TbrPDEB1) that offered a much needed contribution to TbrPDEB1 computational drug discovery. Further, how we used the derived force-field parameters in combined MD simulations with the goal of discovering novel TbrPDEB1 inhibitors.

Lastly, Chapter 6 gives an overall conclusion of the thesis and outlines future work. The graphical abstract below summarizes the thesis layout.



Graphical representation of thesis layout

CHAPTER 1

Literature Review

Chapter Overview

Included in this Chapter are an introduction to Human African Trypanosomiasis (HAT) and a review of literature concerning current HAT control, treatment, and management.

1.1 Human African Trypanosomiasis (HAT)

HAT is a vector-borne zoonotic disease caused by flagellated hemoparasites known as Trypanosoma. It is transmitted to mammalian hosts by the bite of an infected tsetse fly of the *Glossina* genus and cases of mechanical transmission by Tabanids have also been reported [1]. Other less common means of transmission include mother-to-child (trypanosomes can cross the placental barrier) and sexual contact [2–4]. In animals the disease is known as Nagana while in Humans it is known as Human African Trypanosomiasis or sleeping sickness [5]. HAT causes considerable morbidity and mortality [6–8]. This is especially more pronounced in resource limited settings where the disease is most prevalent resulting in declined productivity and poor quality of life [9]. Due in part to mass screening of at risk populations, treatment of diagnosed cases, intensive vector control strategies, and treatment of animal reservoirs, reports of new HAT cases have declined in the recent decades [10]. Altogether there are over 65 million people at risk HAT and the number of new cases reported per year has fallen to well below 5,000 in recent years [11]. There is however some concern that the number of new cases might be underestimated because many cases go unreported [11]. The number of new cases reported in 2017 was 1447 [11]. It is also important to note that while the annual incidence of HAT has been on the decline over the decade there are reports of new cases in previously HAT-free areas [9,11,12]. The number of HAT cases has reached control levels that were never seen in the past. However it is worth mentioning that HAT normally occurs in sporadic epidemics and the risk of re-mergence is further compounded by climate change affecting vector distribution [12]. The low number of cases is also a major constraint for the development of clinical trials to test new drugs.

HAT is caused by members of the subgenus *Trypanozoon* that is composed of three microscopically indistinguishable sub-species, namely *Trypanosoma brucei brucei* that is non-human-infective but causes Nagana in animals, *Trypanosoma brucei gambiense* causes chronic HAT, and *Trypanosoma brucei rhodesiense* causes acute HAT. Chronic HAT also known as Tbg HAT is prevalent in West and Central Africa where infection is characterised by patients taking years to become patent [9]. Acute HAT also known as Tbr HAT is prevalent in East and Southern Africa and infection is characterised by disease developing within a few weeks or months [9]. Uganda is endemic for both Tbr and Tbg HAT within two geographically isolated regions [13].

HAT pathology is divided into the early hemo-lymphatic and the meningo-enciphalitic stage. In the early hemo-lymphatic stage, following an infected tsetse fly bite the parasites proliferate in the blood and lymphatic system. This is followed by the meningo-enciphalitic stage where the parasites cross the blood brain barrier (BBB) into the central nervous system (CNS) causing mental health disturbances such as sleep cycle alterations that are characteristic of the disease [9].

1.2 HAT Lifecycle

When an infected tsetse fly bites an infected human host the following events unfold. The infected fly introduces metacyclic forms of the parasite that are in its saliva during feeding. The parasites multiply at the site of infection causing a swelling known as a chancre and then pass from the lymph into the blood [3,14]. The parasites are then carried into the bloodstream where they differentiate into the long slender form that is also known as the bloodstream form of the parasite [3,14]. During the course of infection, the parasites differentiate into the short stumpy form that are present in the blood. The insect part of the cycle that takes about three weeks begins when a hungry tsetse fly feeds on an infected host [14]. The stumpy forms of the parasites present in the blood meal differentiate into procyclic forms in the insect mid-gut. The procyclic forms then make their way to the tsetse fly salivary glands where they eventually transform into epimastigotes and finally metacyclic trypomastigotes that are the human-infective form [3,14].

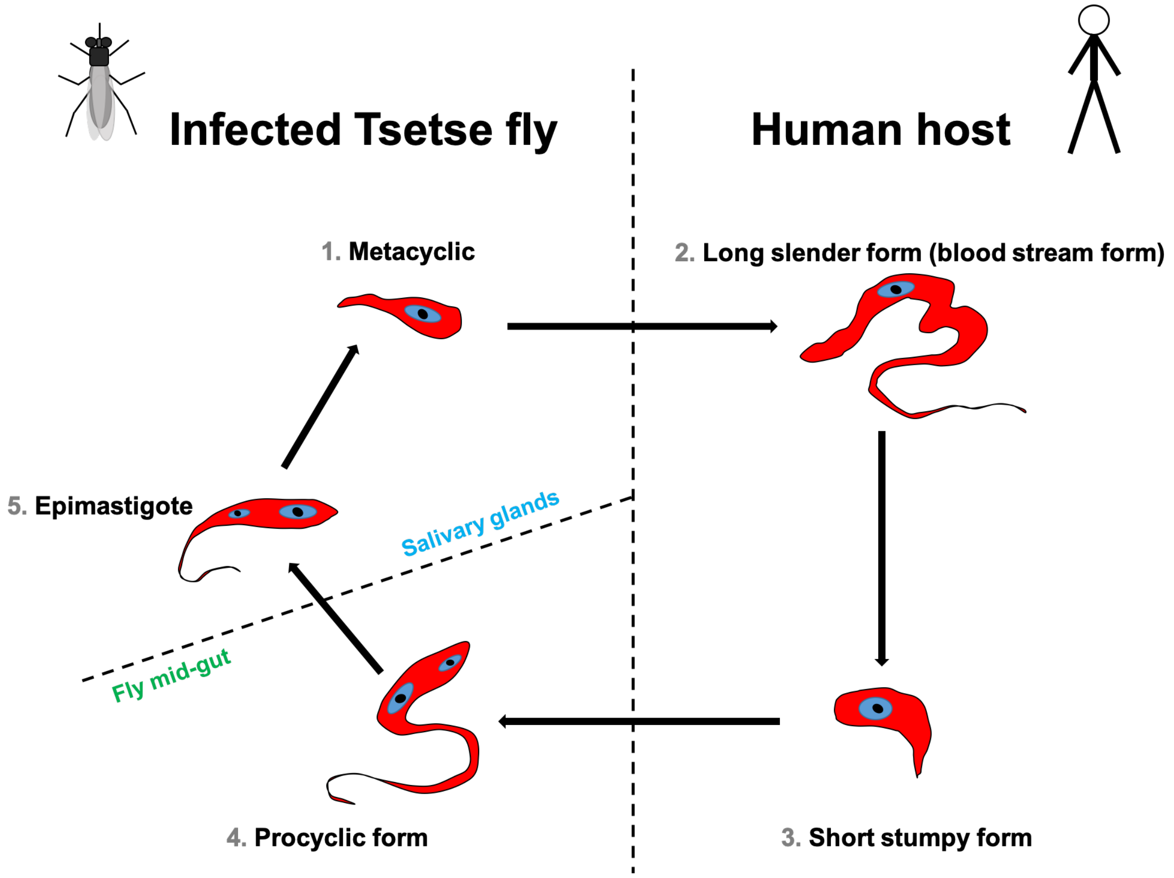


Figure 1.1: A schematic illustration of the life cycle of *T. brucei* in the human host and tsetse fly vector.

1.3 Current Disease Treatment and Management Modalities

The primary means of HAT control and management are chemotherapeutics and vector control [9]. Chemotherapy remains the main means of disease management because there are no vaccines available for HAT. HAT chemotherapeutics are further hampered by the use of few drugs that are old, toxic, difficult to administer, and increasingly ineffective [15,16]. The drugs currently used to treat HAT are stage specific. HAT clinical pathology can be broken down into two stages. This is important because treatment strategies for the stages is different [17]. The first stage is the hemo-lymphatic stage following the bite of an infected tsetse fly. The parasites in this stage of infection are mainly in the blood and lymphatic systems [9]. This is followed by the second stage where the parasites cross the blood brain barrier (BBB) and invade the central nervous system (CNS) [9]. Staging during treatment is important because the drugs that are used to treat the stage 2 have to be able to cross the BBB and are invariably more toxic [17]. The drugs used to treat stage 1 HAT include: Pentamidine [18,19] and Suramin [19,20] that were discovered in 1940 and 1920 respectively. While those used to treat stage 2 include: Melarsoprol [19–21], Eflornithine [19–21], and Nifurtimox [19,22] that were discovered in 1949, 1990, and 2009 respectively. Though effective, the current combination of Nifurtimox and Eflornithine requires 14 doses administered intravenously over a week [19]. There is a dire need for new, safe, easy to administer, and efficacious drugs. The most recent developments in HAT chemotherapeutics include oral administered drug Fexinidazole that has shown positive results of a Phase III clinical trial for the treatment of chronic HAT [19,23,24]. Additionally, Fexinidazole has also recently received a positive response from the European Medicines Agency and will start being implemented in the field in the coming months (2019) [25]. Also of importance is the orally-active anti-HAT drug SCYX-7158 (Benzoxaborole) that has also entered Phase III clinical trials [26].

1.4 Knowledge Gap

Current reports indicate that the number of new HAT cases reported per year is on the decline. The goal as indicated by World Health Organisation (WHO) is the elimination of the disease as a public health problem (less than 1 case per 10,000 inhabitants) by 2020 and the ultimate goal defined as the interruption of transmission by 2030 [17]. However due to factors such as climate change, conflict, and migration the risk of sporadic outbreaks still remains [9,12]. HAT was generally considered to be fatal if left untreated however there have been reports of asymptomatic carriers (latent infections) as well as spontaneous recovery from infection i.e. self-cure/resistance in some HAT cases in West Africa [27–30]. This could point to human trypanotolerance; a phenomenon well described in the animal disease Nagana in some cattle species such as the N'dama [31]. Further, the existence of *T. b. gambiense* asymptomatic carriers and potential animal reservoirs of the parasite are major challenges to reaching the 2030 goal. There is a dire need to develop new, effective, easy to administer drugs and furthermore, explore alternative treatment options such host genetic determinants.

In Part I (Chapter 2 and Chapter 3) of this work we sought to determine host genetic factors that might be associated with disease susceptibility and/or resistance using a Candidate Gene Association Study (CGAS). The host genetic factors analysed were Single Nucleotide Polymorphisms (SNPs) in selected genes identified in the literature to be of importance in HAT. We further analysed the effects of missense SNVs on protein function using Macrophage migration inhibitory factor (MIF) as a case study.

In Part II (Chapter 4 and Chapter 5) of this work we sought to identify novel inhibitors that targeted: 1) the trypanosome folate pathway by inhibiting *T. brucei* Pteridine reductase 1 (*Tb*PTR1) and 2) the Cyclic Adenosine Monophosphate signalling by inhibiting *T. brucei* Phosphodiesterase B1 (TbrPDEB1). Due to a lack of force field parameters describing the TbrPDEB1 bi-metal catalytic centre, we calculated force-field parameters that accurately described this in order to carry out accurate molecular dynamics simulations.

PART I

HOST GENETIC DETERMINANTS

CHAPTER 2

A Candidate Gene Association Study on a *Tbr* and *Tbg* HAT Endemic Cohort from Uganda

Chapter Overview

The host genetic contributions to trypanosomiasis infection outcomes in humans is gaining recognition, as more loci have been suggested to play a role in disease susceptibility. Included in this Chapter is an introduction to the concepts concerning genetic association testing studies and how they relate to HAT. It includes a review of literature on variation, genetics, association testing, and SNP analysis. The Chapter also outlines the motivation and research objective of Part I of the thesis. Lastly, the Chapter includes a case-control CGAS where we looked for susceptibility/resistance variants associated with *Tbr* and *Tbg* HAT in two Ugandan populations.

In my Master's mini-thesis we studied a smaller cohort from NW Uganda that is endemic for only *Tbg* HAT and was composed of 50 samples (20 cases and 30 controls) [32]. Only a single gene was studied, a MIF gene vcf-subset was generated from the whole genome sequence data that was used for a CGAS [32]. The work presented herein differs in that we used 440 samples from the *Tbr* and *Tbg* endemic areas of Uganda that were genotyped on a chip based on selected genes. The *IL10, IL8, IL4, HLAG, TNFA, TNX4LB, IL6, IFNG, MIF, APOL1, HLAA, IL1B, IL4R, IL12B, IL12R, HP, HPR*, and *CFH* genes were selected for this study. The CGAS approach was used to analyse the SNP frequencies between cases and controls to identify host genetic determinants associated with disease susceptibility and/or resistance.

The work presented in this Chapter has been presented in the following publication: Kimuda MP, Noyes H, Mulindwa J, Enyaru J, Alibu VP, Sidibe I, Ngoyi DM, Hertz-Fowler C, MacLeod A, Bishop ÖT, Matovu E. No evidence for association between APOL1 kidney disease risk alleles and Human African Trypanosomiasis in two Ugandan populations. PLoS neglected tropical diseases. 2018 Feb 22;12(2):e0006300.

2.1 Introduction

The annual incidence of HAT on the African continent has been on the decline over the recent decades however there are still reports of new cases in previously HAT free foci [9,12,17]. Adequate surveillance and response is still needed especially in areas affected by civil unrest and conflicts that lead to uncontrolled migrations that interrupt established epidemiological surveys and control strategies [9,33,34].

The general consensus used to be that HAT was fatal if left untreated, however reports of asymptomatic carriers (latent infections), self-cure, and spontaneous recovery in Tbg HAT cases in West Africa has this in question [27–30]. Trypanotolerance is not rare and has been previously observed in some cattle breeds such as the N'dama that are resistant to Nagana [31]. Disease outcome can be affected by a broad range of factors including, environmental factors, genetic factors, diet, and parasite genetic polymorphisms. In the era of precision medicine genetic factors influencing disease outcomes are becoming more and more important [35].

Host genetic factors have also been explored in murine and human models. Stijlemans *et al.* showed that macrophage migration inhibitory factor (MIF) deficient mice, as well as normal ones treated with anti-MIF antibody, exhibited lowered inflammatory responses, liver damage and anaemia when experimentally challenged with *T. b. brucei* [36]. There are several studies in the Democratic Republic of Congo (DRC), Guinea, Cameroon, Cote D'Ivoire, Guinea and Uganda that have found evidence for polymorphisms in *HP*, *IL6* and *APOL1* associated with outcome of infection [13,37–42]. In this study, we carried out a CGAS using case and control samples from Tbg and Tbr endemic areas of Uganda in order to identify risk variants for HAT.

The candidate gene studies included *IL10*, *IL8*, *IL4*, *HLAG*, *TNFA*, *TNX4LB*, *IL6*, *IFNG*, *MIF*, *APOL1*, *HLAA*, *IL1B*, *IL4R*, *IL12B*, *IL12R*, *HP*, *HPR*, and *CFH* genes.

2.1.1 HAT Self-cure

HAT is widely considered to be fatal if left untreated. However, there are reports of asymptomatic carriers (latent infections) as well as spontaneous recovery from infection i.e. self-cure/resistance in some HAT cases in West Africa [27–30]. This points to human trypanotolerance, a phenomenon well described in the animal disease Nagana [31]. There are several factors leading to this spectrum of HAT susceptibility and infection outcomes, including host and parasite genetic polymorphisms that have been associated with HAT presentation [43–45]. It is on this basis that we chose to identify host genetic determinants that may affect HAT outcome i.e. disease susceptibility and/or resistance.

2.1.2 Genetic Variation and Disease

2.1.2.1 Genetic Variation

Genetic variation refers to the differences in the DNA sequences in the human genome between individuals [46,47]. Genetic variation is important because it affects how individuals respond to drugs and also disease susceptibility/resistance [48,49]. Mutations are one of the key forces driving evolution because they allow genetic variation in populations [50]. The most common type of genetic variants are Single Nucleotide Polymorphisms (SNPs) which are defined as changes in a single nucleotide base (A, C, G or T) that occur in >1% of a population [46,51]. Single Nucleotide Variants (SNV) are any nucleotide changes without any limitations of frequency. Other genetic variants include insertion and deletion (INDEL) which refer to insertions or deletion of bases in the DNA sequence [52]. SNPs can occur in a variety of locations within the DNA, for example in non-coding regions such as introns or expression regulation sites [53]. SNPs that occur in coding regions of genes can be subdivided into two groups; synonymous that do not result in a change in the encoded amino acid and non-synonymous that result in a change in the encoded amino acid [53]. Further, non-synonymous SNPs are divided into those that result in a change in the encoded amino known as missense SNPs and those that result in a premature stop codon being encoded known as nonsense [53]. As sequencing gets cheaper there is more and more data on human genetic variation available; the largest publically available database is dbSNP [54].

2.1.2.2 Genetic Variation and Disease

In Mendelian disease, a single SNP in a single gene can be responsible for a disease state [55]. A common example is sickle cell anaemia where a single nsSNP results in a GAG codon changing to GTG [56,57]. This results in an E to V amino acid substitution that yields a defective haemoglobin protein and hence disease [56,57]. However, in complex diseases, SNPs causing disease might work in coordination with several other SNPs, environmental factors, and/or numerous genes in order to result in a disease state [55]. HAT is an example of a complex disease and for this study we focused on SNPs as a genetic determinant for disease susceptibility and/or resistance.

2.1.3 SNP Based Association Studies and Linkage Disequilibrium (LD)

Population genetic association studies have been used to identify genetic variants linked to complex diseases [58]. This is carried out by comparing the frequencies of SNPs in affected (cases) and unaffected (controls) groups within an at risk population [58]. Association studies fall under two categories depending on the initial hypothesis. Genome Wide Association Studies (GWAS) which are designed to study variants across the entire genome and require no prior hypothesis concerning possibly involved genes or molecular markers [59]. GWAS is a powerful tool for detecting genetic associations in complex traits. However, it does have a few drawbacks, in order to boot statistical power GWAS requires a large sample size, and many variant associations fail to reach the stringent genome wide significance in the independent testing [60,61]. Candidate gene based studies (CGAS) require a prior hypothesis concerning which genes might be involved in the disease [60]. This requires suitable knowledge of the candidate genes and their significance in the different biological pathways related to the disease [60,61]. In spite of a higher statistical power, a drawback of a prior hypothesis is that CGAS are only able to discover associations in the selected genes unlike GWAS that can discover new associations in any number of genes [61].

SNPs that are inherited together are said to be in complete Linkage disequilibrium (LD) [62]. Even though they are located at different loci the presence of one automatically implies the other. As a result, including both SNPs does not contribute any additional information to the association testing, if anything it reduces on the statistical power. One copy of SNPs in LD is therefore used in association testing to increase power [63]. Additionally, another implication of LD is that majority of the SNPs displaying association signals are not the functional polymorphisms but rather they are in LD with the functional polymorphism.

2.1.4 Statistical Analyses in Association Testing

2.1.4.1 Hardy-Weinberg Equilibrium (HWE)

The Hardy-Weinberg Equilibrium (HWE) principle states that the genotype and allele frequencies within a randomly mating population are constant in different generations [64]. Evolutionary influences that can affect HWE include; inbreeding, founder effect, assertive mating, and population bottlenecks for example. In population genetic association testing SNP deviations in HWE can be lead to spurious association results. Additionally, HWE deviations can also be indicative of genotyping errors. SNPs showing any such deviations in HWE are excluded from the study [63].

2.1.4.2 Admixture

Population stratification of population structure refers to differences in allele frequency that are result of sub-populations within a population [65]. This is usually due to differences in ancestry for example genetic drift or population divergence [65]. In population genetic association testing it is normally important to control for population stratification as it can result in spurious associations. In this study, admixture, which refers to when two or more populations with distinct lineages interbreed, was used to detect population structure [65]. The cases and controls in the population were expected to have similar admixture.

2.1.4.3 Fisher's Exact Test

Fisher's exact test is the standard test for CGAS given a limited sample size [66]. The Fisher's exact test was used to compare the SNP frequencies between the cases and controls to determine which were associated with susceptibility and/or resistance. For a single test a p-value below 0.05 is considered significant.

2.1.4.4 Correction for Multiple Testing

In genetic association testing the major limitations of multiple testing is that of significance and teasing out false positives. In this study controlling for multiple testing was implemented using a Bonferroni correction [67] and the Benjamini-Hochberg false discovery rate (FDR) [68]. In the Bonferroni correction; $\alpha^* = \alpha/n$, where α is the level of significance, n is the number of independent SNP association tests and α^* is the adjusted threshold of significance [67]. The Bonferroni correction is very stringent in that it assumes that each statistical test is independent which is not necessarily true due to possible LD between SNPs. This results in increased false negatives. Hence a less stringent correction was included, the Benjamini-Hochberg FDR, which is used to give an estimate of the proportion of false positives out of the total number of significant associations ($p < 0.05$) [68].

2.1.5 Candidate Gene selection

Genes included in this study was based on prior knowledge of and their association with parasitic diseases like HAT.

2.1.5.1 Apolipoprotein 1 (*APOL1*)

APOL1 is a human serum protein that lyses trypanosomes and has been associated with resistance to African trypanosomes. There are also reports of strong association of *APOL1* G1 and *APOL1* G2 alleles with risk for chronic and end stage kidney disease [69]. A study by Cooper *et al.* in the Eastern Uganda foci of Soroti and Kaberamaido found the *APOL1* G2 variant to be protective against Tbr HAT with an odds ratio of 0.2 [40]. In contrast, there was an association of *APOL1* G2 variant with a higher risk while *APOL1* G1 was associated with a lower risk of developing HAT in a Guinean population [42].

2.1.5.2 Complement Factor H (*CFH*)

CFH plays an important role in the alternative complement pathway and is crucial in preventing damage to host cells by the complement system. CFH is known to contribute to pathogenicity of several pathogenic bacteria [70], *Candida albicans* [71], and *Plasmodium falciparum* [72]. CFH was previously explored unsuccessfully as a diagnostic marker for staging HAT [73].

2.1.5.3 Human Leucocyte Antigen G (*HLA-G*) and Human Leucocyte Antigen A (HLAA)

HLA-G is involved in the modulation of host immunity and its immunosuppressive properties have been shown to contribute to infectious disease susceptibility, viral infections, and inflammatory conditions [52]. *HLA-G* variants have been associated with HAT susceptibility [37]. The *HLA-G* gene, unlike other classical HLA class 1 genes, characteristically displays low polymorphism, showing a strong selection pressure for invariance [55, 56]. HLAA which is also involved in immune modulation was shown to be associated with plasmodium falciparum infection severity in a cohort of Malian children [77].

2.1.5.4 Haptoglobin (*HP*) and Haptoglobin- Related Protein (*HPR*)

HP is a plasma protein involved in free heme scavenging while HPR is part of the Trypanolytic Factor 1 (TLF-1) together with APOL1 [78]. *HP* and *HPR* copy number variation have been previously associated with HAT protection [78].

2.1.5.5 Interleukin 4 (*IL4*), Interleukin 6 (*IL6*), Interleukin 8 (IL8), Interleukin 10 (*IL10*), Interleukin 1 Beta (*IL1B*), Interleukin 4R (*IL4R*), Interleukin 12 Beta (*IL12B*), and Interleukin 12R (*IL12R*)

IL4 is a cytokine that is important in regulating immune responses. It is produced by T-cells and natural killer (NK) cells. It is important in determining the phenotype of naïve CD4$^+$ T cells. In murine models IL4 was important in determining the outcome of *Tbr* infection where it was involved in the control of parasitaemia levels [70]. In addition, polymorphisms in the *IL4* gene were shown to be associated with susceptibility for *T. cruzi* infection in a Bolivian cohort where it was important in maintaining inflammation and parasite persistence [71]. IL6 is a cytokine with several different functions and is known to possess both inflammatory and counter inflammatory effects [72, 73]. Elevated serum levels of IL6 have been reported in chronic HAT patients [74] and previous genetic association studies have shown *IL6* gene polymorphisms to be associated with disease susceptibility in a Congolese population [17]. The *IL8* gene has been associated with control of infection and *IL10* with HAT development in a sample cohort from Guinea [84,85]. *IL1B*, *IL4R*, and *IL12RB* were associated with *Plasmodium vivax* malaria susceptibility in a cohort from Brazil [86].

2.1.5.6 Interferon Gamma (*IFNG*)

IFNG is a pro-inflammatory cytokine that has roles in both innate and adaptive immune responses. It is secreted by T-cells and NK cells. Courtin *et al.* analysed *IFNG* polymorphisms for association with chronic HAT in DRC and no associations were observed [38]. However murine models have shown that IFNG is important in parasite control [75]. Recent human studies on cytokine profiles suggest that IFNG might be associated with neurological symptoms in *Tbr* HAT [87]

2.1.5.7 Tumor Necrosis Factor Alpha (TNFA)

Produced by macrophages, TNFA is a pro-inflammatory cytokine that is important in innate immune response to intracellular pathogens. Previous studies showed that *TNFA* polymorphisms play a role in *Tbg* HAT susceptibility in Ivorian and Congolese populations [38,85]. Furthermore, studies in murine models also demonstrated that TNFA is important in trypanosomiasis associated anemia [57]. *TNFA* polymorphisms have also been shown to be associated with occurrence and severity of malaria in a Tanzanian cohort, underscoring its role in inflammation responses due to parasite infection [58].

2.1.5.8 Thioredoxin 4LB (*TNX4LB*)

TNX4LB is part of the glutathione/glutathione reductase system [90]. Trypanosomes were thought not to possess a thioredoxin system [91]. They instead possess a unique trypanothione/trypanothione reductase instead [91,92]. To detoxify hydroperoxides, trypanosomes rely on a cascade made up of trypanothione reductase, trypanothione, tryparedoxin, and tryparedoxin peroxidase [91,93]. TNX4LB is important in regulating cellular redox balance, inhibition of apoptosis, promoting the growth of cells, and modulating inflammation [94].

2.1.5.9 Macrophage Migration Inhibitory Factor (MIF)

MIF is a cytokine that is important in both adaptive and innate immunity [59–62]. It is an important player in the induction of systematic inflammation and as such is implicated in many inflammatory diseases [63]. Functional polymorphisms in the *MIF* gene, for example in the promoter, have been linked with diseases such as sepsis [64], rheumatic arthritis [65], scleroderma [66], tuberculosis [29], juvenile inflammatory arthritis [67], and systemic lupus erythematosus [68]. Its major function is the recruitment of myeloid cells to inflammation sites

[59, 69]. While there are no human studies directly linking MIF to HAT, murine studies show that MIF plays a role as a mediator of the inflammation cascade which is a key feature in trypanosomiasis-associated pathology [15,59]. MIF deficient mice featured limited anaemia, increased iron bio-availability, improved erythropoiesis and a marked reduction in red blood cell (RBC) clearance during chronic stages of infection when the mice were experimentally challenged with *T. b. brucei* [15].

2.2 Research Aims and Objectives

In this study we sought to identify host genetic determinants associated with HAT. The specific objective for this section of work was to identify SNPs in the selected candidate genes that are associated with HAT.

2.3 Materials and Methods

2.3.1 Ethics Statements

This study received ethical clearance from the Uganda National Council of Science (UNCST; assigned code HS 1344) and was reviewed by the Internal Review Board (IRB) of the Uganda Ministry of Health. Recruitment of participants was through community engagements and active field surveys. A consent form was administered to each of the participants in their local language by a trained local health worker that informed them of the project aims and objectives. All participants had to be over 18 years of age and were required to provide written consent in order to join the study.

2.3.2 Study Population

The Tbr HAT samples were collected from Iganga district that is a traditional HAT foci found in the South East Uganda [10]. The Tbr HAT sample comprised individuals from Basoga, Baganda, Banyole, Balamogi, Basiginyi, Itesot and Japhadola ethnicities. The Tbg HAT samples were collected from the traditional North West Uganda foci [10] of Adjumani, Arua, Koboko, Maracha, and Moyo districts. The Tbg HAT sample comprised of individuals from the Kakwa, Lubgbara and Madi ethnicities. A HAT case was defined as an individual in whom trypanosome parasites had been found in either blood, lymph node aspirate, cerebral spinal fluids (CSF) or body tissues such skin snips. A HAT control was defined as an individual matched to a case in terms of age, sex, ethnicity, living environment but had never had HAT infection. This was demonstrated them having no serological reaction to the CATT or Trypanolysis test. It was important to ensure the both groups had similar exposure to HAT. Blood samples were collected from individuals who had had HAT, active cases, and controls in the period from 2014 to 2018. Venous blood was collected using venepuncture into EDTA or heparin vacutainer tubes (BD). This was followed by the preparation of buffy coats in the field laboratories that were then stored in liquid nitrogen and transported to the Molecular Biology Laboratory, COVAB, Makerere University. At the lab the DNA was extracted using a QIAGEN extraction kit and quantified using a Qubit™ (Life Technologies).

2.3.3 Study Design

This study was designed to have 80% power to detect odds ratios (OR) > 2 for loci with disease allele frequencies of 0.15 – 0.65 and 100 cases and 100 controls with the 96 SNPs genotyped. The statistical power was calculated using the genetics analysis package gap in r [107]. The study had a total of 462 samples of which 239 samples were from the *Tbr* HAT region (120 cases and 119 controls) and 223 samples from the *Tbg* HAT region (110 cases and 232 controls).

2.3.4 Gene Selection

In this study, candidates for genotyping were selected based on previous reports in the literature. *HPR* [108,109] and *APOL1*, which has alleles that are lytic to *Tbr in vitro* and whose upregulation was previously associated with *Tbg* HAT infection, although an *in vitro* assay did not demonstrate toxicity to *Tbg* [110–112]. *HLA-G* is a polymorphic gene that plays an important role in determining the strength of immune responses [75]. *HLA-G* polymorphisms have previously been shown to be associated with HAT [37]. CFH was included in this study because it plays an important role in the regulation of complement cascade which is important in parasite lysis [113,114]. *IL6* polymorphisms have previously been associated with HAT in a Congolese population [38], while MIF depletion resulted in lowered *T. brucei* associated immunopathology in mice [15]. The latter has also been shown to play a role in several other inflammatory diseases such as Tuberculosis [29], Leishmaniasis [30] and Malaria [31]. Additionally, *TNFA* [32, 33] , *IFNG* [34] , *IL8* [33] , *IL4* [33] , and *IL10* [32, 33, 35] polymorphisms have been previously associated with HAT. While *IL1B*, *IL4R*, *IL12B*, and *IL12RB1* gene polymorphisms have been associated with *Plasmodium vivax* Malaria [36].

2.3.5 SNP Selection

A total of 96 SNPs were selected for this study. The SNPs panel was generated using two approaches:

1) SNPs previously reported to be associated HAT in literature
2) SNPs that were in LD across a given gene ($r^2 > 0.5$) based on a merged SNP dataset obtained from 10X coverage whole genome sequence data generated from 230 samples from HAT endemic regions (DRC, Guinea Conakry, Ivory Coast, and Uganda; TrypanoGEN consortium, sequences at European Nucleotide Archive Study; EGAS00001002602) and 1000 Genomes Project data from African populations

The PLINK v1.9 package (https://www.cog-genomics.org/plink/1.9/) was used to calculate Linkage (r^2) [120,121]. SNPs that failed genotyping were excluded.

2.3.6 Genotyping

Genotyping was done by INRA (Plateforme Genome Transcriptome de Bordeaux, France) using approximately 1µg of genomic DNA per sample. A multiplex analysis composed of two sets of 80 SNPs each was designed using Assay Design Suite v2.0 (Agena Biosciences). SNP genotyping was carried out using the iPLEX Gold genotyping kit (Agena Biosciences) for the MassArray iPLEX genotyping assay, following the manufacturer's instructions. Detection of products was done using a MassArray mass spectrophotometer and the data acquired in real time with MassArray RT software (Agena Biosciences). SNPs that failed genotyping at INRA and additional SNPs were genotyped at LGC Genomics, Hoddesden, UK where the SNPs were genotyped using the PCR based KASP assay [37]. A summary of the candidate genes and SNPs is shown in Supplementary Table 1.

2.3.7 Statistical Analysis

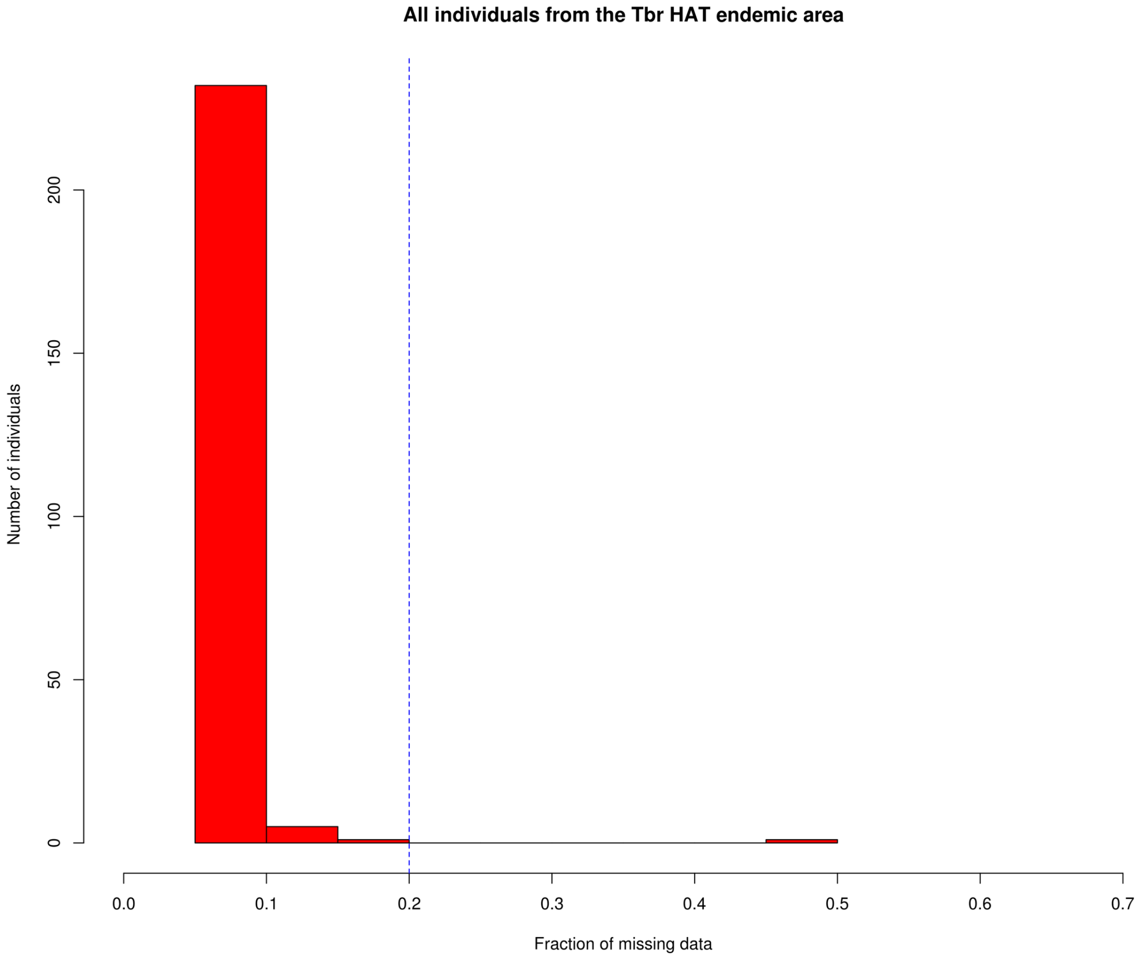
Quality control procedures were implemented using PLINK v1.9 package (http://pngu.mgh.harvard.edu/purcell/plink/) [120]. This involved converting the raw genotype data to PLINK format, determining the level of individual and genotype missingness, HWE, LD, and estimating allele frequencies. Admixture was used to test for population stratification and was implemented using Admixture 1.3 [123] and visualized using StructurePlot2 [124]. The HAT association testing for the SNPs was carried out using the Fisher's exact test [40] and was implemented using PLINK at 95% confidence level. Correction for multiple testing was implemented using the Bonferroni correction [67] and the Benjamini-Hochberg false discovery rate (FDR) [68].

2.4 Results

The Tbr HAT endemic population was composed of 239 individuals while the Tbg HAT endemic population was composed of 223 individuals. The Tbr HAT sample was made up of 120 cases and 119 cases with a mean age of 43 ± 5 years and 1:2 male to female ratio. The Tbg HAT sample was composed of 110 cases and 113 controls with a mean age of 37 ± 5 years, and a male to female ratio of 1:1.

2.4.1 Genotyping and Data Quality Control

Ninety-six (96) SNPs in 15 genes were genotyped from each of the Tbr and Tbg HAT endemic area samples as shown in Supplementary Table 1. Before association testing, individuals with missing data, SNPs that were not in HWE, SNPs with missing data or those that were poorly genotyped were removed using PLINK [38, 44]. High levels of individual missingness can indicate poor DNA quality and/or technical problems while high levels of genotype missingness can lead to potential bias in the subsequent association testing [127,128]. Individuals with more than 20% or 15% missing data were excluded from the *Tbr* and the *Tbg* HAT endemic datasets respectively. The final dataset had 237 (99.2%) individuals from the Tbr HAT sample and 202 (90.6%) individuals from the Tbg HAT sample as shown in Figure 2.1 and Figure 2.2.



All individuals from the Tbr HAT endemic area

Figure 2.1: Histogram of missing data rate in all individuals from the Tbr HAT endemic area https://doi.org/10.1371/journal.pntd.0006300.s008

SNPs that were missing more than 30% or 40% data were excluded from the *Tbr* and the *Tbg* HAT samples respectively as shown in Figures 2.3 and 2.4. Further filtering of the SNPs was done using a HWE p-value cut-off of 1 x 10^{-8} and genotype scatter plots. To ensure a high LD between the markers and causal SNPs, any loci that were in a 5 SNP window after a single step with a variance inflation factor (VIF; VIF = 1/(1-R2)]) beyond 0.2 were excluded. A high LD between markers and SNPs increases redundancy and hence reduces statistical power. After quality pruning procedures, 79 loci from Tbr and 85 loci from the Tbg HAT endemic samples were included in the total sample.

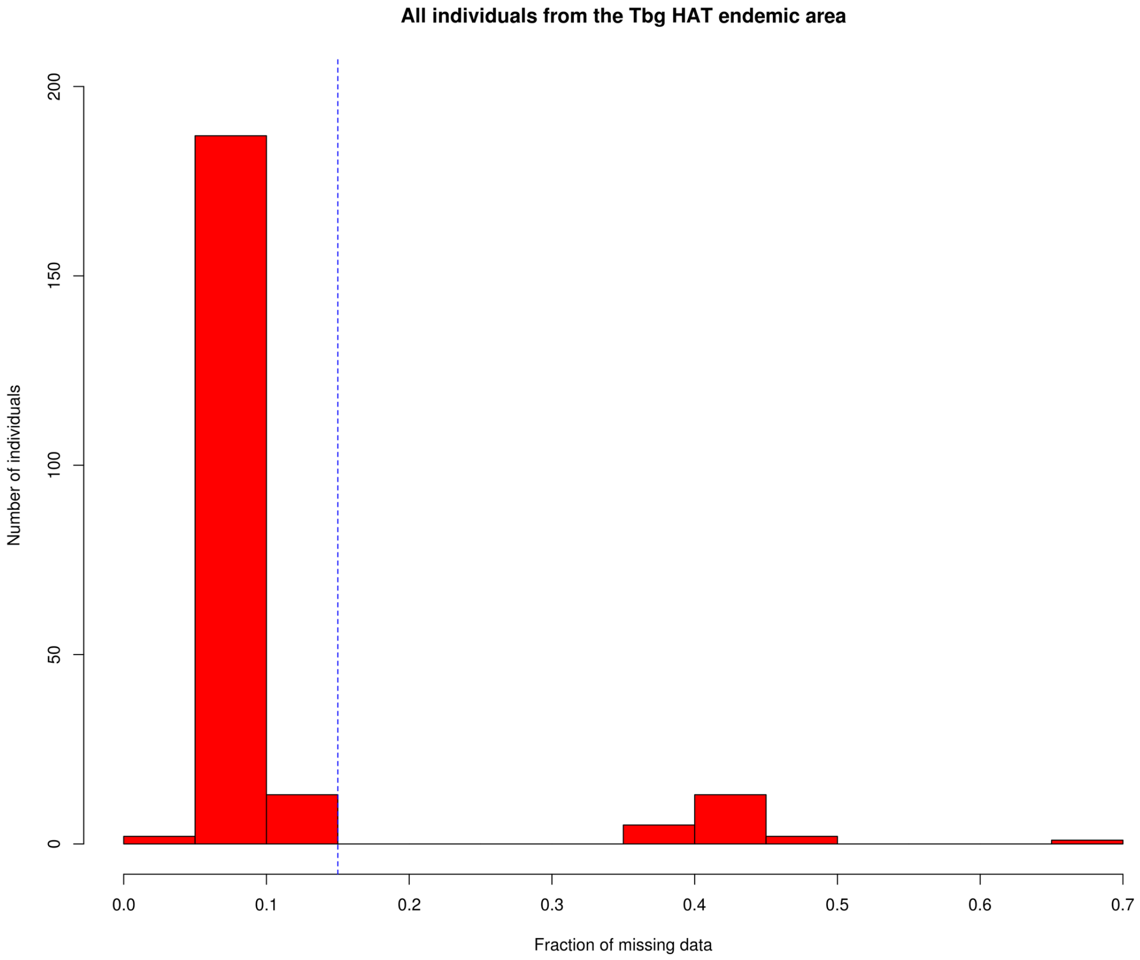


Figure 2.2: Histogram of missing data rate in all individuals from the Tbg HAT endemic area https://doi.org/10.1371/journal.pntd.0006300.s009

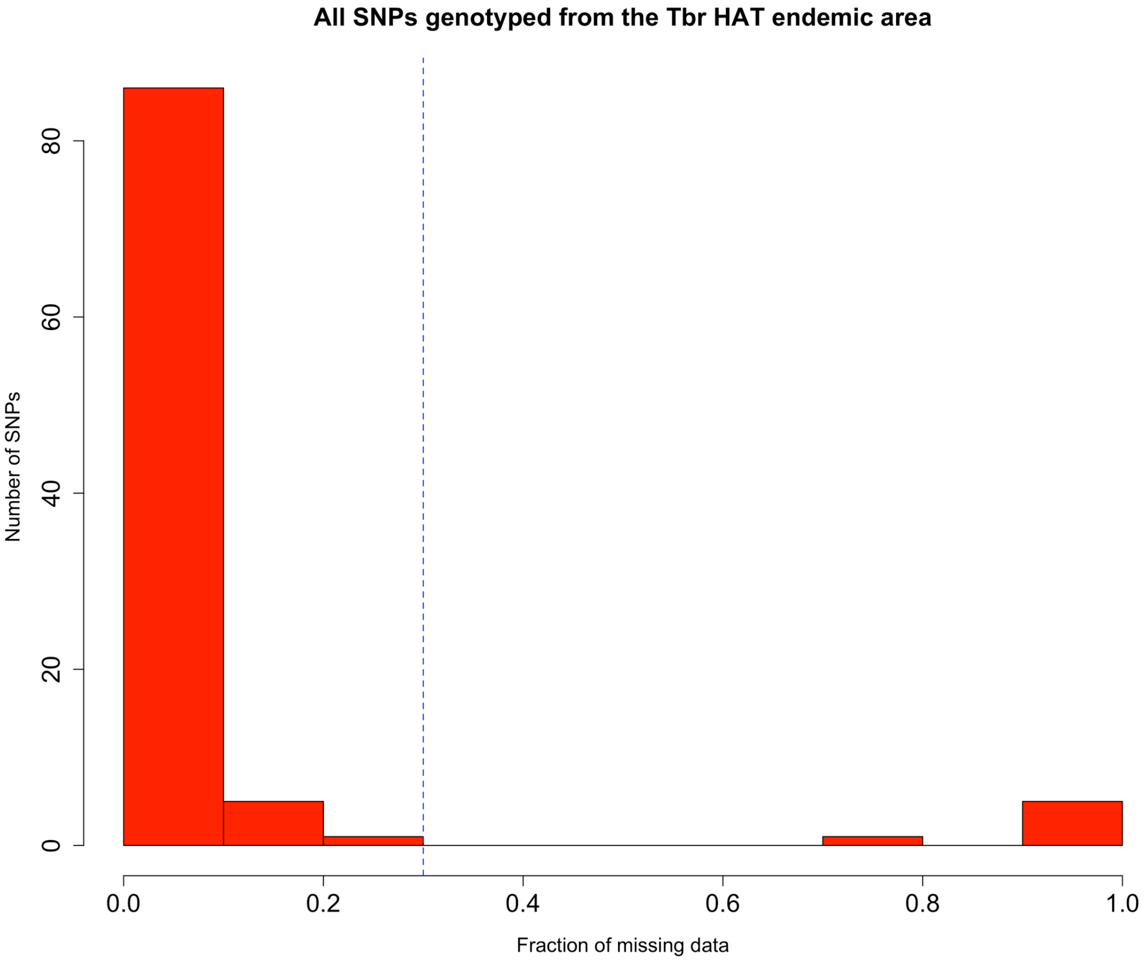


Figure 2.3: Histogram of missing data rate in all SNPs from the Tbr HAT endemic area passing https://doi.org/10.1371/journal.pntd.0006300.s010

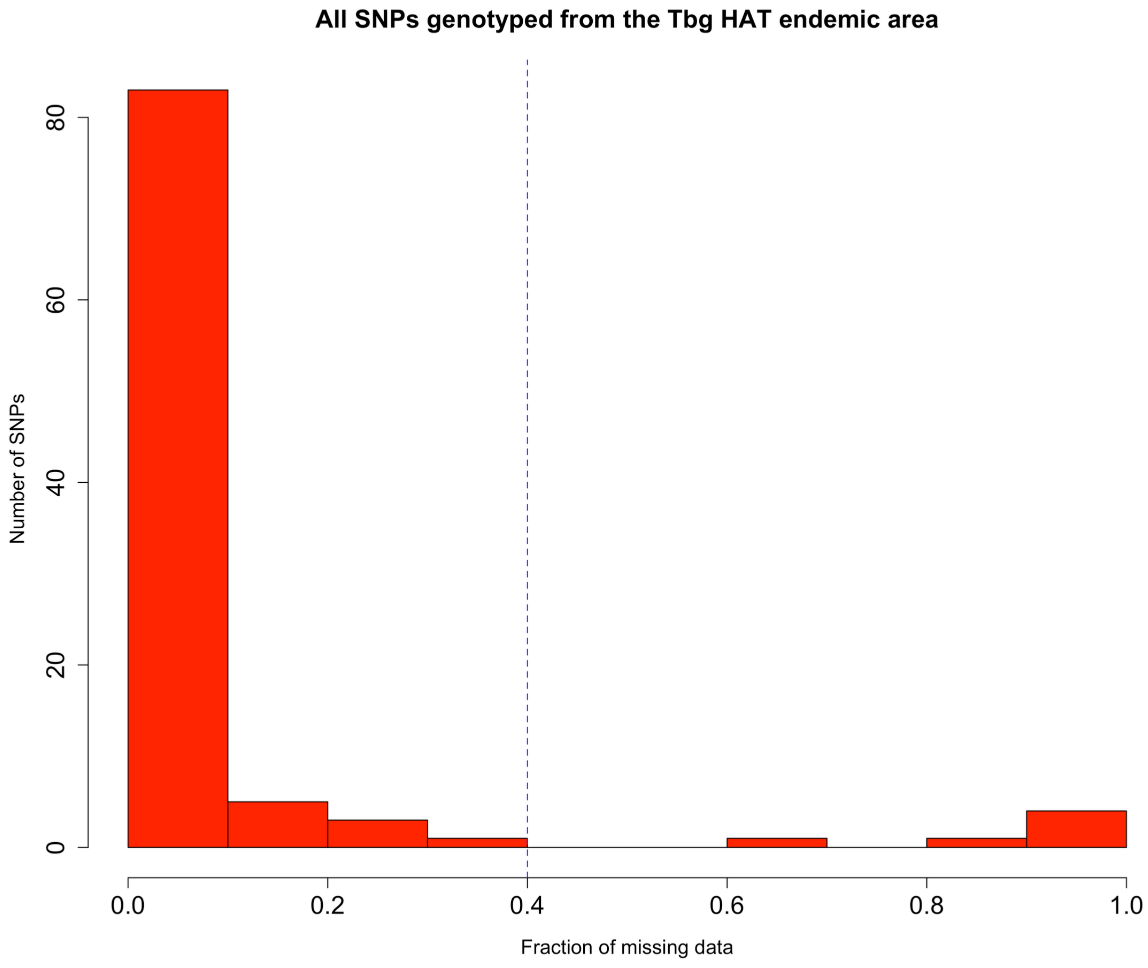


Figure 2.4: Histogram of missing data rate in all SNPs from the Tbg HAT endemic area passing https://doi.org/10.1371/journal.pntd.0006300.s011

2.4.2 Admixture Analysis

Testing for population structure was carried out using Admixture analysis. Of the 8 ancestral populations K=4 had the lowest coefficient of variation (CV) error. The CV errors were very similar across K1-8 (0.42 – 0.46). The Admixture plot (Figure 2.5) showed no evidence of gross population structure and as a result no correction for population structure was required.

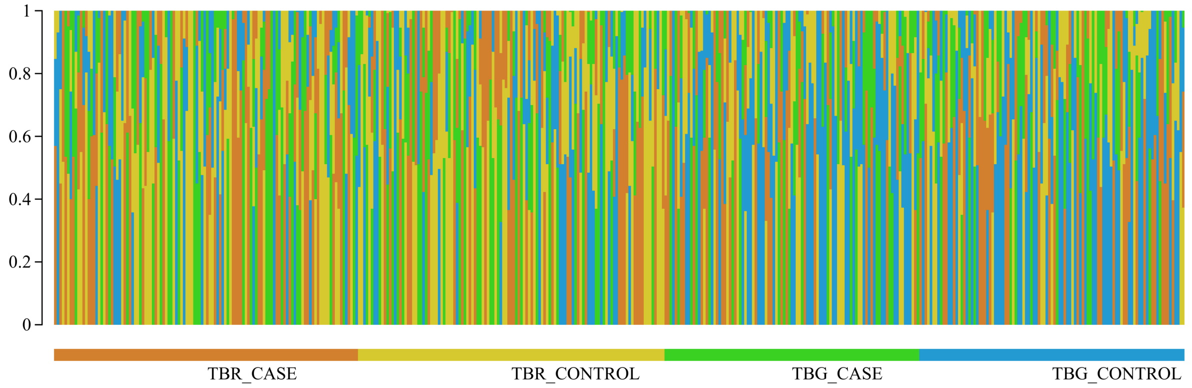


Figure 2.5: Bar plot showing the admixture analysis performed for K=4. Individuals are shown as vertical bars https://doi.org/10.1371/journal.pntd.0006300.s012

2.4.3 Association Testing Yielded Suggestive SNPs in *CFH*, *IL4*, *IL6*, *HLA-G*, *TNFA*, *IFNG*, and *MIF* Genes

In the study, we tested for association between Tbg HAT and Tbr HAT separately because HAT occurs as two distinct forms of the disease i.e. chronic or acute HAT. The following SNPs in the *Tbr* HAT endemic area achieved statistical significance at a threshold of $P < 0.05$ indicative of possible association with HAT: rs9380142 (*HLA-G*) on chromosome 6 (odds ratio (OR) = 1.8, $P = 0.0028$, 95%, confidence interval (CI): 1.2-2.7), rs1800630 (*TNFA*) on chromosome 6 (OR = 1.8, $P = 0.038$, 95% odds ratio confidence interval: 1.0-3.2), rs9282783 (*MIF*) on chromosome 22 (OR = 2.2, $P = 0.033$, 95% odds ratio confidence interval: 1.0-4.8), rs2243283 (*IL4*) on chromosome 5 (OR = 1.6 , $P = 0.040$, 95% odds ratio confidence interval: 1.0-2.6), and rs34383331 (*MIF*) on chromosome 22 (OR = 1.6, $P = 0.030$, 95% odds ratio confidence interval: 1.0-2.6) (Table 2.1). The following SNPs in the *Tbg* HAT endemic area populations achieved statistical significance at a threshold of $P < 0.05$, indicative of possible association with HAT: rs1061170 (*CFH*) on chromosome 1 (OR = 0.62, $P = 0.019$, 95% odds ratio confidence interval: 0.42-0.93), rs2069843 (*IL6*) on chromosome 7 (OR = 2.0, $P = 0.033$, 95% odds ratio confidence interval: 1.1-3.9), rs78554979 (*IFNG*) on chromosome 12 (OR =

3.6, $P = 0.035$, 95% odds ratio confidence interval: 0.97-13.4), and rs1233330 (*HLA-G*) on chromosome 6 (OR = 0.52, $P = 0.044$, 95% odds ratio confidence interval: 0.27-1.0) (Table 2.1). The association test results for all the SNPs in the Tbr HAT and Tbg HAT endemic samples are shown in Supplementary Table 2 and Supplementary Table 3 respectively.

A total of 6 SNPs from the Tbr HAT sample and 4 from the Tbg HAT sample had $p < 0.005$, but were not significant after Bonferroni correction (Table 2.1). Of note was the absence of any association with any SNPs in the *APOL1* gene.

Table 2.1: SNP *p* values after association testing with Tbr and Tbg HAT

Tbr HAT endemic sample (N= 238; 119 cases, 119 controls)

CHR	SNP	GENE	BP	Allele 1	Cases	Controls	Allele 2	P	OR	95% lower CI	95 % upper CI	BONF	FDR_BH
6	rs9380142	HLA-G	29798794	G	0.369	0.242	A	0.003	1.834	**1.231**	**2.731**	0.1805	0.181
6	rs1233330	HLA-G	29799103	A	0.076	0.136	G	0.03	0.522	**0.284**	**0.958**	1	0.541
5	rs2243283	IL4	132016593	G	0.275	0.188	C	0.04	1.644	**1.031**	**2.621**	1	0.434
22	rs34383331	MIF	24238079	A	0.24	0.16	T	0.03	1.657	**1.049**	**2.621**	1	0.434
22	rs9282783	MIF	24236359	G	0.089	0.042	C	0.033	2.227	**1.025**	**2.621**	1	0.434
6	rs1800630	TNFA	31542476	A	0.156	0.092	C	0.038	1.807	**1.031**	**3.169**	1	0.434
22	rs73885319	APOL1	36661906	G	0.076	0.101	A	0.339	0.730	**0.385**	**1.383**	1	0.875
22	rs71785313	APOL1	36662046	T	0.081	0.086	A	0.803	0.928	**0.482**	**1.788**	1	0.9281

Tbg HAT endemic sample (N=202; 99 cases, 103 controls)

CHR	SNP	GENE	BP	Allele 1	Cases	Controls	Allele 2	P	OR	95% lower CI	95 % upper CI	BONF	FDR_BH
1	rs1061170	CFH	196659237	C	0.409	0.525	T	0.019	0.627	**0.4221**	**0.9313**	1	0.611
6	rs1233330	HLA-G	29799103	A	0.076	0.136	G	0.045	0.521	**0.5211**	**0.2693**	1	0.611
12	rs78554979	IFNG	68554636	C	0.051	0.015	T	0.035	3.638	**0.9861**	**13.42**	1	0.611
7	rs2069843	IL6	22769994	A	0.147	0.078	G	0.033	2.038	**1.07**	**3.882**	1	0.611
22	rs73885319	APOL1	36661906	G	0.015	0.015	A	0.843	1.041	0.208	5.220	1	0.898
22	rs71785313	APOL1	36662046	T	0.056	0.070	A	0.612	0.782	0.346	1.766	1	0.8653

*Abbreviations: CHR = Chromosome, SNP = SNP ID, BP = Physical position (base-pair) **(**Human genome build GRCh37), Allele 1 = Minor allele (based on whole sample), Cases = Frequency of allele 1 in cases, Controls = Frequency of allele 1 in controls, Allele 2 = Major allele, P = p-value for this test, OR = Estimated odds ratio (for Allele 1, i.e. Allele 2 is reference), BONF = Bonferroni single-step adjusted p-values, FDR_BH = Benjamini & Hochberg (1995) step-up FDR control https://doi.org/10.1371/journal.pntd.0006300.t001

2.5 Discussion

The host genetic contribution to trypanosomiasis infection outcomes in humans is gaining recognition, as certain loci have been suggested to play a role in disease susceptibility. In this case-control CGAS, we looked for susceptibility variants associated with *Tbr* and *Tbg* HAT in two Ugandan populations. We found no evidence for variants associated with Tbr or Tbg HAT after correction for multiple testing.

The two populations were from broadly different ethnolinguistic groups that are geographically isolated [10]. Further *T. b. rhodesiense* and *T. b. gambiense* the causative agents of Tbr HAT and Tbg HAT respectively have different mechanisms by which they mediate serum resistance [129]. This has implications on selection pressure on the host genes in the different populations. From the Admixture analysis we found no evidence of population structure within the studied SNPs which might have reduced the power of the study (Figure 5). The power calculations indicated that we had power to detect odds ratios > 2, but 7 of the 10 SNPs with $P < 0.05$ had odds ratios < 2.0, which the study was not powered to detect.

The most important finding was the absence of any association of HAT with *APOL1* variants G1 or G2. The *APOL G2* allele (rs71785313) is known to be lytic to *T. b. rhodesiense in vitro* [112]. Cooper *et al.* reported an association of *APOL1* G2 with protection from Tbr HAT with an odds ratio of 0.2 [40]. This study is well powered to discover such a strong effect but the frequencies in the Tbr HAT sample between cases and controls were very similar (8.1% and 8.6%, 95% odds ratio confidence interval: 0.37 - 2.34). Further, in Guinea the *APOL1* G1 and *APOL1* G2 allele frequencies only showed associations with HAT risk/susceptibility after asymptomatics were included [42]. The absence of asymptomatics in our Uganda study cohort might thus in part explain the absence of association with *T. b. gambiense* infections. In Côte d'Ivoire, a suggestive association of *APOL1* G1 with protection against HAT was reported, it however did not pass correction for multiple testing [41]. The role of *APOL1 G2* in response to Tbr HAT remains to be clarified.

2.6 Conclusion

None of the suggestively significant associations at nine SNP loci passed correction for multiple testing. The Benjamini-Hochberg FDR showed the rate of type 1 errors or false positives, for example rs9380142 in HLA-G indicated that there was a 18% chance that it was a false positive and conversely a 82% chance that it is a true positive. There was a greater than 38% probability that these nine SNPs were true positives [68,130]. This gives an indication that larger sample size might probably increase the likelihood that these are indeed genuine associations with the disease [131]. Our findings indicate that *HLA-G* variants might be important in both Tbr HAT and Tbg HAT but further studies are required.

CHAPTER 3

In silico Analysis of the Effects of Single Nucleotide Variants on the Structure and Function of Macrophage Migration Inhibitory Factor (MIF)

Chapter Overview

In the previous Chapter we investigated a set of genes using a case-control CGAS in order to identify SNPs that are associated with HAT. In the current Chapter we present a case-study using Macrophage Migration inhibitory factor (MIF) protein to show how computational tools can be used to explore the molecular mechanisms by which missense Single Nucleotide Variants (SNVs) might affect protein structure, dynamics and function. This was done using a set of SNVs retrieved from the Short Genetic Variations Database (dbSNP). It also includes a review of literature on SNV analysis using computational tools. This work follows up on work presented in my Master's mini thesis [32] that also carried out SNV analysis of the MIF protein, but differs substantially in the following aspects;

1) The current study focuses on a different set of missense SNVs all of which are validated either by 1000 Genomes or cluster [54]. Cluster validation indicates that the at least one of the variants in the cluster was confirmed using non-computational methods or has associated frequency information.
2) In the previous work we carried out 1 ns MD simulations that were too short to yield any substantial inferences. In the current study we carried out extensive all atom 100 ns MD simulations with more in depth analyses.

3) The current study also includes recently developed Dynamic Residue Network (DRN) analysis using MD-TASK [132].

3.1 Introduction

MIF (EC:5.3.2.1) is a cytokine that affects the activity of multiple cells and is a key player in both innate and adaptive immunity [97]. MIF, as the name suggests is involved in the recruitment of leucocytes to sites of inflammation [106]. As such it is important in several inflammatory pathologies and is classified as an inflammatory cytokine [98,133]. Further, this pluripotent cytokine is involved in several biological activities including phagocytosis, macrophage activation, tumoricidal activity, and induction of nitrogen oxide production [99,133,134]. MIF is expressed by several cells for example; macrophages, monocytes, cardiomyocytes, and vascular smooth muscle cells [97,135]. It is also secreted by tumor cells and parenchymal cells when they are under stress [97]. MIF recruitment of leukocytes to sites of inflammation is dependent on interaction with MIF-CXCR2 (C-X-C Motif Chemokine Receptor 2) and MIF-CXCR4 (C-X-C Motif Chemokine Receptor 4) receptors [106].

MIF has been implicated in several diseases and cancers [99,136]. Polymorphisms in the *MIF* gene have been associated with diseases such as the following: juvenile inflammatory arthritis [104], systemic lupus erythematosus (SLE) [105], rheumatic arthritis [101], scleroderma [102], and tuberculosis [115]. MIF depletion has been also shown to alleviate pathology when mice are experimentally infected with trypanosomes [36].

The majority of the studies done on MIF are association studies [101,102,104,105,115] and very little has been done to explore how polymorphisms such as missense SNVs affect MIF structure and function. In the current study we sought to use computational tools to explore the molecular mechanisms by which missense SNVs impact MIF protein structure and function. The computational tools and approaches used included: *in silico* effect prediction, homology modelling, Molecular Dynamics (MD), Principal Component Analysis (PCA), and Dynamic

Residue Network (DRN) analysis. The general theory of each of these approaches is explained below.

3.1.1. *In silico* Analysis of SNVs

SNPs that are found to be significantly associated with disease and/or susceptibility can be further analysed using bioinformatics tools. In this study we carried out bioinformatics analyses on missense SNVs in MIF to elucidate molecular mechanisms by which they might impact protein structure and function. We used *in silico* effect prediction and structural based approaches. The *in silico* effect prediction tools used were mainly based on machine learning algorithms and relied on a mixture of sequence and structural data. These included: MutPred [137], PredictSNP [138], SIFT [139], PROVEAN [140], and PANTHER [141]. The structural based analyses included: structure prediction using homology modelling and Molecular Dynamics simulations.

3.1.1.1 Web Based *in silico* SNV Effect Prediction Tools

Sorting Intolerant From Tolerant (SIFT) (http://blocks.fhcrc.org/sift/SIFT.html) [139] predictions are based on a comparison of similar sequences in the SWISS-PROT and TREMBLE. MutPred [137] predictions are based on SIFT [139] and a gain or loss of 14 different functional and structural factors. PredictSNP [138] is a consensus classifier that uses several other prediction tools and these included:

a) Multivariate Analysis of Protein Polymorphism MAPP [142] (http://www.ngrl.org.uk/Manchester/page/mapp-multivariate-analysis-protein-polymorphism) that uses physical chemical properties and an alignment score.

b) PhD-SNP [143] (http://snps.biofold.org/phd-snp/phd-snp.html) and SNPs&GO (http://snps-and-go.biocomp.unibo.it/snps-and-go/) [144] that use methods based on Support Vector Machines (SVMs) that predicted disease related mutations from a protein sequence and structure respectively.

c) Phenotyping Polymorphism (PolyPHEN) [145] (http://genetics.bwh.harvard.edu/pph2/) based its prediction on a naïve Bayesian classification based on sequence and specific physical comparative empirical rules.

d) SNAP [146] (https://rostlab.org/services/snap/) predicted the consequence of an amino acid substitution on protein functionality using a neural network method and required both the protein sequence and mutation residue as inputs.

Protein Variation Effect Analyser (PROVEAN) [140] (http://provean.jcvi.org/index.php) uses pre-computed PROVEAN predictions based on pairwise sequence alignment map scores. Protein Analysis Through Evolutionary Relationships (PANTHER) (http://www.pantherdb.org/tools/csnpScoreForm.jsp) [141], uses a probability estimation of the impact of the amino acid substitution on the protein based on evolutionary conservation.

3.1.2 Homology Modelling

Homology modelling also known as comparative modelling, refers to the generation of an accurate 3D model of an amino acid sequence using the atomic coordinates of a template structure [147–149]. The goal of homology modelling is to generate a quality structure similar to that achieved experimentally through X-ray crystallography. It consists of four main steps which are;

i) Template identification

ii) Multiple sequence alignment (MSA)

iii) Model building and refinement

iv) Model validation.

In the current study proteins that had no crystal structures were modelled for subsequent structural analyses. The models were created using the MODELLER ‘automodel’ class that automates many of the modelling steps [149,150]. The reliability of the calculated model can be assessed based on the lowest normalized Discrete Optimized Protein Energy (z-DOPE) score [149] and online webservers such as ProSA [151], VERIFY-3D [152] and PROCHECK [153].

3.1.3 Molecular Dynamics

MD is a computer simulation that allows the simulation of the natural motion of biological molecules for a fixed amount of time where they are governed by standard laws of physics and energetics [154–157]. Proteins in nature are not simply static but actually in motion to varying degrees [158]. MD allows mimicking of protein motions in physiological environments that is crucial to elucidating protein dynamics and function [159]. MD simulations have been successfully used to analyse protein flexibility, folding stability, structure, and internal motions [160,161]. In the current study, comparisons between wild-type and mutant structures during MD simulations formed the basis of elucidating nsSNP effects on the protein structure, dynamics and function.

3.1.4 Principal Component Analysis (PCA)

During MD simulation the protein undergoes changes in molecular conformation as a function of time in a multidimensional space [162,163]. PCA is a multivariate statistical analysis that can be used to reduce dimensionality of data [162]. PCA has been used to show the most important motions of a protein during MD simulation [162]. In MD analysis, PCA can be used to generate a 2D representation of the variance i.e. the protein conformational sampling in phase space. The modes of fluctuation are represented by the eigenvectors and the top eigenvalues represent the large overall motions [162,164]. PCA is solved by calculating eigenvalues of a covariance matrix which is a $3N \times 3N$ matrix where the element (i, j) describes the covariance between coordinates i and j. The principal components of the covariance matrix are the eigenvectors. The eigenvalue for each eigenvector describes the amount of variance explained [162,164]. When the Cα atoms are used to represent the system in a PCA it is referred to as Essential Dynamics Analysis (EDA) which allows describing of the motions (using eigenvectors) relevant to protein function while reducing the computational cost [165,166].

3.1.5 Dynamic Residue Network (DRN) Analysis

MD-TASK [132] was used to carry out DRN analysis which is a method to analyse MD trajectories using graph theory where the protein residues are treated as nodes connected by edges to form a network. The networks are generated using a 6.7 Å cut-off between nodes (Cα atoms for glycine residues and Cβ atoms for other amino acids). The specific networks analyses were;

3.1.5.1 Residue Contact Map

A weighted adjacency matrix was generated using edge lists based on a method described by Doshi *et al.* [132,167] and the networks graphed using R [168].

3.1.5.2 Average Shortest Path (Average L)

The shortest path (L_{ij}) in a network is defined as the least number of edges traversed to reach *j* from *i* nodes [132]. The average shortest path (*Average L*) is defined as the average of all the shortest paths (*L*) for a residue during MD simulation [132]. The *Average L* is defined by the equation shown below:

$$\alpha = \sum_{i,j \in V} \frac{L(i,j)}{n(n-1)}$$

Equation 3.1: ***Average L*** **Equation**

Where ***V*** refers to a set of nodes in the network, and *L*(*i,j*) is the shortest path from node *i* to node *j* and n is the total number of nodes [169]. The average shortest path was used to determine spatial displacement of the residues during simulation [170].

3.1.5.3 Average Betweenness Centrality (Average BC)

Betweenness centrality (BC_i) is the measurement of the number of the shortest paths in a network that run through node i [132]. The *Average BC* of a single node v is defined by the equation below:

$$BC(v) = \sum_{i,j \in V} \frac{\sigma(i,j|v)}{\sigma(i,j)}$$

Equation 3.2: ***Average BC*** **Equation**

Where V is the total number of nodes, $\sigma(i, j)$ refers to the number of shortest paths (L) and $\sigma(i, j| v)$ refers to the number of paths passing through node v other than i, j [169].

Residues with high BC values indicate that the node is important for communication within in the network [132]. High *Average BC* values have previously been shown to be located at positions that are critical to inter- and intra-protein domain communication [170]. Changes in *average BC* due to mutations therefore can indicate changes in communication flow that has an impact on protein functionality, flexibility, and folding.

3.1.5.4 Dynamic Cross Correlation

In the MD-TASK tool DCC is used to investigate residue movement relative to other residues [132]. The tool produces a correlation network N x N matrix, where N is the number of Cα atoms and each element in the matrix corresponds to the DCC, C_{ij}, between each i^{th} and j^{th} atom. The DCC between each atom was calculated using the equation shown below [171]:

$$\mathrm{C}_{ij} = \frac{\langle \Delta \mathrm{r}_i \cdot \Delta \mathrm{r}_j \rangle}{\sqrt{\langle \Delta \mathrm{r}_i \rangle} \cdot \sqrt{\langle \Delta \mathrm{r}_j \rangle}}$$

Equation 3.3: DCC Equation

where $\Delta \mathrm{r}_i$ equals the displacement from the average position of atom i and < > represents the time average over the trajectory. The values of C_{ij} range from -1 to 1, where a value of 1 indicates correlation, -1 indicates anti-correlation and 0 no correlation. The matrices are then used to generate heat maps that can be used to study major changes in physical motions of the same atom positions between the wildtype and mutants during the MD simulation.

DRN was used to identify mutations (P40L, A188V, and D104N) that destabilized the Renin-Angiotensinogen complex [172] and mutation Y99A that resulted in MIF impaired function [173].

3.2 Research Aims and Objectives

In this study we sought to understand the molecular mechanisms by which missense SNVs affect protein structure and function using the MIF protein as a case-study. The specific objective for this section of work was to analyse missense SNVs with the aim of elucidating the molecular mechanism by which they impact structure and function.

3.3 Materials and Methods

3.3.1 Data Retrieval

Missense SNVs in the human *MIF* gene (NM_002415.1) were retrieved from dbSNP148 [54]. The missense SNVs were filtered on validation status (either by 1000 Genomes project or by frequency cluster) [174]. The protein sequence for MIF was retrieved from the UniProtKB/Swiss-Prot database (UniProt: P14174) [175] and the crystal structure from the Protein Databank (PDB: 3DJH) [176,177].

3.3.2 Prediction of Functional Effects of Missense SNVs via Web Servers

A consensus of five SNP web servers was used to predict the functional effects of the missense SNVs. This was done to improve the accuracy of the prediction as the different tools utilized varying algorithms to determine the effect outcome as presented in section 2.7.1. The analysis web servers used were; MutPred [137], PredictSNP [138], SIFT [139], PROVEAN [140], and PANTHER [141]. The missense SNVs were categorized as either ‘neutral’ which indicated no negative impact on the protein or ‘deleterious’ which indicated a negative impact on the protein using a consensus agreement cut-off of ≥ 60%. MutPred gave its results in the form of a probability score where any value above 0.5 was considered deleterious, with an increase beyond that indicating a higher confidence in the score [137]. The other web servers gave their results as either deleterious/damaging or neutral/benign. Each of the web server tools were run using their default user settings.

3.3.3 Homology Modelling

The MIF mutant structures were calculated using homology modelling utilizing an in-house Python script. The mutations were introduced into each of the three chains in the homotrimer. The models were generated from a 3DJH template using MODELLER (version 9.19) and the ‘automodel’ class [178,179]. For each mutant a total of 100 models were generated and the best selected based on the lowest normalized Discrete Optimized Protein Energy (z-DOPE) score [149]. The models were validated using ProSA [151], VERIFY-3D [152] and PROCHECK [153] online servers. The MIF structures in this study were renumbered to start from M1 to A115.

3.3.4 Molecular Dynamics Simulations

MD simulations were performed using the GROMACS 5.1.4 molecular dynamics package [157]. MD simulations were carried out using the AMBER03 force field in a cubic box of 5.07 × 5.18 × 5.16 (nm) with a 2.37 × 2.86 × 3.30 (nm) center. The box volume was 2462.31 (nm^3) and the temperature 300 K. At least 1.5 nm distance between any protein atom and the wall was allowed. Each system was solvated using a simple point-charge (SPC) water model in a cubic box with periodic boundary conditions. The systems were neutralized using Na^+ and Cl^- counter ions. Energy minimization was done using a steepest descent algorithm with a 100 kJ/mol/nm tolerance value. This was done to correct any inappropriate structural distortion or incorrect geometry. Equilibration was done using the NVT (constant number of particles, volume, and temperature) and NPT (constant number of particles, pressure, and temperature) ensembles. A total of 100 ns production were performed with trajectories generated at a time step of 2 fs and frames saved every 10 ps. MD simulations were analysed in respect to;

Radius of gyration (Rg), Root Mean Square Fluctuation (RMSF), and Root Mean Square Deviation (RMSD). Graphical representations, diagrammatic visualizations, and analyses were generated using PyMOL [180], GROMACS [157], and GRACE software (http://plasma-gate.weizmann.ac.il/Grace/). MD simulations were carried out on the whole MIF homotrimer for both the wildtype and mutant proteins.

3.3.5 Principal Component Analysis

The calculation and diagonalisation of the weighted covariance matrix was carried out using a script that utilized Python libraries that included MDTraj [181], Scikit-learn [182,183], and the 2D plot generated using Matplotlib [184]. The PCA was based on the elements based on the fluctuation positions of the Cα carbon atoms during simulation. A simple linear transformation was applied to the covariance matrix in the Cartesian space followed by a diagonalisation to yield eigenvectors and eigenvalues [162,164]. The eigenvectors were then used to prepare 2D projections of the MD trajectories.

3.3.6 Dynamic Residue Network Analysis

DRN was carried out on the MD trajectories using MD-TASK to identify changes in protein dynamics. The networks were generated using a cut-off 6.7 Å between nodes. The default MD-TASK parameters for *Average L*, *Average BC*, Residue Contact Maps, and DCC were used.

3.4 Results and Discussion

The MIF wildtype structure (PDB:3DJH) and mutants were renumbered to start from M1 to A115. In this Chapter, the amino acid notation includes the amino acid single letter abbreviation followed by the amino acid position and then the chain e.g. F50.C indicates that it is a phenylalanine at amino acid position 50 in chain C.

3.4.1 MIF structure and Function

MIF is a 12.5 kDA homotrimer, each of the monomers is composed of 115 amino acid residues. Each monomeric unit consists of two anti-parallel α-helices and four-stranded β-sheet [177,185] as shown in Figure 3.1. MIF leukocyte recruitment is contingent on its interaction with its CXCR2, CXCR4 and CD74 receptors [186–189]. MIF is a pluripotent cytokine that possesses a broad range of immunological and enzymatic activities. Protein-protein interactions appear to be important for MIF function and several important surface residues have been identified in the literature [186–189].

3.4.1.1 MIF/CD74 Interaction

MIF interaction with the CD74 receptor is important for its leucocyte recruitment activity, the MIF/CD74 binding site is composed of MIF residues Y37, I65, K67, W109, and N110 [190]. Wet-lab mutation of these residues to alanine resulted in varying losses in MIF neutrophil recruiting activity. Mutation of I65 and Y109 to alanine, resulted in partial loss of neutrophil recruiting activity, while mutation of Y37, K67 and N110 to alanine resulted in a complete loss of neutrophil recruiting activity [190]. *In silico* protein-protein docking studies have identified more residues that might be critical for MIF and CD74 interaction and these include: P35, Q36, E55, P56, S75, K78, S91, P92, D93, N111, T113, F114, and A115 [191]. The last 10 carboxy terminus residues have also been shown to be critical for MIF macrophage recruiting activity with their deletion resulting in the loss of MIF immunological activity [192,193].

3.4.1.2 MIF/CXCR2 Interaction

MIF residues L47-P56 (LMAFGGSSEP) form an N-like loop that has been shown to be important in MIF and CXCR2 interaction that is crucial for macrophage recruitment [106,188,189,194].

3.4.1.3 MIF Tautomerase Activity

An N-terminal proline residue (P2) has been shown to be important for MIF tautomerase activity that appears to be conserved among prokaryotic isomerases [187]. Other residues involved in the dopachrome and phenyl keto-enol tautomerase activity include surface residues K3 and I65, and buried core residues Y96 and N98 [187,195,196]. Mutation of P2 to alanine results in loss of MIF tautomerase activity and impaired MIF immunological activity [190,197–199].

3.4.1.4 MIF Oxidoreductase Activity

Residues C57-A58-L59-C60 form a CALC motif that has been shown to be important for MIF oxidoreductase activity and consequently macrophage recruiting activities [200]. The CALC motif is essential for a cysteine dependent intramolecular disulphide bond that is important for oxidoreductase activity [200,201]. Mutation of C57 or C60 to serine resulted in a mutant MIF that lacked both macrophage recruiting activity and oxidoreductase activity [200].

3.4.1.5 MIF Exonuclease and Endonuclease Activity

MIF has the ability to cleave genomic DNA because it possesses 3’ exonuclease and endonuclease activity [202]. This has made MIF a potential therapeutic target in diseases such as arthritis, atherosclerosis, diabetic complications, and nephritis for example, that are characterized by excessive PARP-1 activation [202].

3.4.2 SNV Selection

From dbSNP148 a total of 232 SNVs were retrieved from the human *MIF* gene. The SNVs were broken down as follows: 58 missense, 55 intron, 32 coding synonymous, 31 3-UTR, and 29 5-UTR SNVs. Any missense SNVs that were not validated by either 1000 Genomes (1000G) or by clustering or frequency were excluded from the study [54,174]. Table 3.1 below lists the 27 missense SNVs that were selected for further analysis. The Minor allele frequency (MAF) refers to second most frequent allele value. For the 1000 Genomes validated variants, the Global MAF refers to the MAF of the variant in the global population (1094 worldwide individuals from the phase 1 genotype data) and are included in Table 3.1 where applicable.

Table 3.1: Missense SNVs selected for further analysis and their amino acid mutations

SNV ID (rs)	Validation	Amino acid Mutation
rs200500959	By cluster	P2Q
rs201465617	By cluster	I5M
rs201862457	By 1000G and Cluster (Global MAF: T=0.0002/1)	P16Q
rs533563568	By 1000G (Global MAF: T-0.0002/1)	L23F
rs200995600	By cluster	T24S
rs766674911	By frequency	A28V
rs765744422	By frequency	T31I
rs201631604	By 1000G (Global MAF: A=0.0002/1)	P34T
rs530185151	By 1000G (Global MAF: C=0.0002/1)	Y37H
rs200005486	By cluster	H41P
rs11548059	By cluster	P44L
rs747064040	By frequency	P44A
rs372575900	By 1000G (Global MAF: G=0.0002/1)	M48V
rs560265113	By 1000G (Global MAF: T=0.0002/1)	G52C
rs768313072	By cluster	S54R
rs761381011	By frequency	A58T
rs767129873	By frequency	A58V
rs1049829	By cluster	L59F
rs532485317	By 1000G (Global MAF: G=0.0004/2)	I65M
rs11548056	By cluster	I68T
rs200394994	By cluster	A71T
rs182012324	By 1000G (Global MAF: T=0.0004/2)	S75F
rs780425386	By frequency	L84R
rs200286358	By cluster	Y99C
rs575017934	By 1000G (Global MAF: A=0.0002/1)	A105T
rs1803976	By 1000G (Global MAF: G=0.0004/2)	N106S
rs139210892	By 1000G (Global MAF: T-0.0004/2)	T113S

3.4.3 SNV Structural Properties

The missense SNV resulting amino acid mutations were mapped on the MIF structure and are shown in Figure 3.1 and summarized in Table 3.2. Mutations P2Q (loop) and I65M (loop) are located in the MIF tautomerase active site that is considered vestigial in humans (Figure 3.1 and Table 3.2) [190]. Mutations Y37H (beta sheet), I65M (loop), I68T (loop), S75F (helix), N106S (loop) and T113S (beta sheet) are associated with the MIF/CD74 and MIF/CXCR2 binding sites which are important for MIF biological function [190,191,203] (Figure 3.1 and Table 3.2).

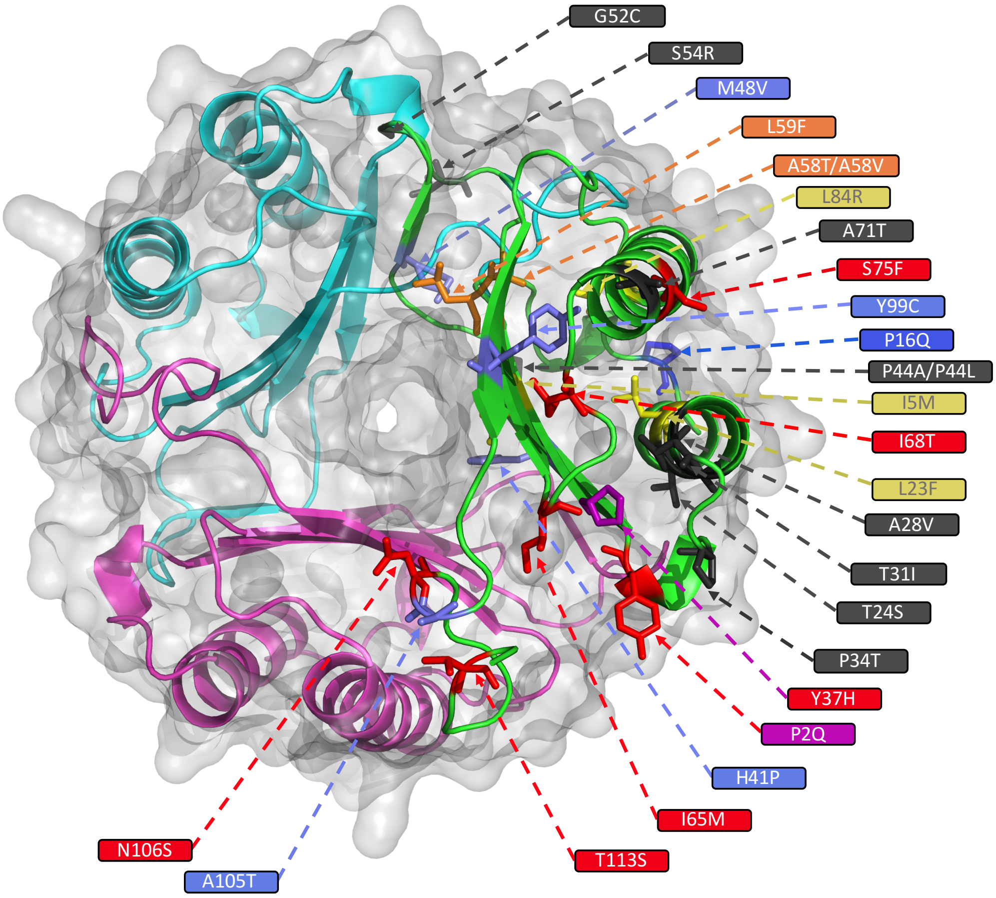


Figure 3.1: The MIF homotrimer 3D structure with chain A colored green, chain B colored cyan and chain C colored magenta. The mutations are highlighted in the chain A monomeric unit with the amino acids shown in stick format. Mutation P2Q which is located in the tautomerase active site (P2Q) was colored purple. P16Q which is located in the endonuclease and exonuclease site was colored blue. Mutations Y37H, I65M, I68T, S75F, N106S, and T113S which are located in the MIF/CD74 and MIF/CXCR2 binding sites were colored red. Mutations A58T, A58V, and L59F are located within the CALC motif were colored orange. Mutations H41P, M48V, Y99C and A105T are located on the chain-chain interface and were colored slate. Mutations I5M, L23F, and L84R are located within the hydrophobic buried core and were colored yellow. Lastly, mutations T24S, A28V, T31I, P34T, P44A, P44L, G52C, S54R, A71T which are located on the protein outer surface were colored gray.

Mutations A58T (sheet), A58V (sheet), and L59F (sheet) are located within the CALC motif that is known to be important for MIF oxidoreductase and macrophage recruiting activities [200]. Mutation P16Q (loop) is located within the nuclease-like superfamily PD-D/E(X)K motif that is common among many mammalian nucleases [202,204]. Mutations I5M (sheet), L23F (helix), H41P (loop), M48V (sheet), and L84R (helix) are located within the buried hydrophobic core while T24S (helix), A28V (helix), T31I (helix), P34T (helix), P44A (loop), P44L (loop), G52C (loop), S54R (loop), A71T (helix), Y99C (sheet), and A105T (helix) are located on the protein outer surface (Figure 3.1 and Table 3.2). Lastly, mutations H41P, M48V, Y99C and A105T are located at chain-chain interfaces (Figure 3.1 and Table 3.2).

Table 3.2: Amino acid physicochemical properties and location on the structure and functional significance

Protein residue mutation	Wildtype-mutant amino acid properties (pH 7.0)	Location on the secondary structure	Location on the 3D structure	Functionally significant sites
P2Q	Hydrophobic-Hydrophilic	Loop	Outer surface	Chain interface & Tautomerase catalytic site [190]
P16Q	Hydrophobic-Hydrophilic	Loop	Outer surface	MIF 3' endonuclease and exonuclease site [202]
T24S	Hydrophilic-Hydrophilic	Helix	Outer surface	-
A28V	Hydrophobic-Hydrophobic	Helix	Outer surface	-
T31I	Hydrophilic-Hydrophobic	Helix	Outer surface	-
P34T	Hydrophobic-Hydrophilic	Helix	Outer surface	-
Y37H	Hydrophobic-Hydrophilic	Sheet	Outer surface	Chain interface & CD74 binding site [190,191]
P44A	Hydrophobic-Hydrophobic	Loop	Outer surface	-
P44L	Hydrophobic-Hydrophobic	Loop	Outer surface	-
M48V	Hydrophobic-Hydrophobic	Sheet	Outer surface	Chain interface & N-like loop [188]
G52C	Hydrophobic-Hydrophobic	Loop	Outer surface	N-like loop [188]
S54R	Hydrophilic-Hydrophilic	Loop	Outer surface	N-like loop [188]
I65M	Hydrophobic-Hydrophobic	Loop	Outer surface	CD74 binding site & Tautomerase catalytic site [190,191]
I68T	Hydrophobic-Hydrophilic	Loop	Outer surface	CD74 binding site [191]
A71T	Hydrophobic-Hydrophilic	Helix	Outer surface	-
S75F	Hydrophilic-Hydrophobic	Helix	Outer surface	CD74 binding site [191]
Y99C	Hydrophilic-Hydrophobic	Sheet	Outer surface	Chain interface
A105T	Hydrophobic-Hydrophilic	Helix	Outer surface	Chain interface
N106S	Hydrophilic-Hydrophilic	Loop	Outer surface	Chain interface & CXCR2 binding site [203]
T113S	Hydrophilic-Hydrophilic	Sheet	Outer surface	Chain interface & CD74 binding site [191]
I5M	Hydrophobic-Hydrophobic	Sheet	Buried hydrophobic core	-
L23F	Hydrophobic-Hydrophobic	Helix	Buried hydrophobic core	-
H41P	Hydrophilic-Hydrophobic	Loop	Buried hydrophobic core	Chain interface
A58T	Hydrophobic-Hydrophilic	Sheet	Buried hydrophobic core	CALC motif [200]
A58V	Hydrophobic-Hydrophobic	Sheet	Buried hydrophobic core	CALC motif [200]
L59F	Hydrophobic-Hydrophobic	Sheet	Buried hydrophobic core	CALC motif [200]
L84R	Hydrophobic-Hydrophilic	Helix	Buried hydrophobic core	-

3.4.4 *In silico* Missense SNV Effect Prediction using Web-based Servers

A consensus prediction was generated by using five different web servers (≥ 60% consensus cut-off). The individual prediction from each of the algorithms is shown in Supplementary Table 3. The consensus results are summarized in Figure 3.2. MutPred predicted 23 out of 27 missense SNVs as deleterious. PredictSNP, SIFT and PROVEAN predicted 18 out of 27 missense SNVs as likely damaging, while Panther predicted 8 out of 27 missense SNVs as disease causing. Mutants P2Q, P34T, Y37H, H41P, G52C, I65M, L84R and Y99C were predicted as detrimental by all the programs, while I5M, S54R, and A71T were all predicted to be neutral (Figure 3.2).

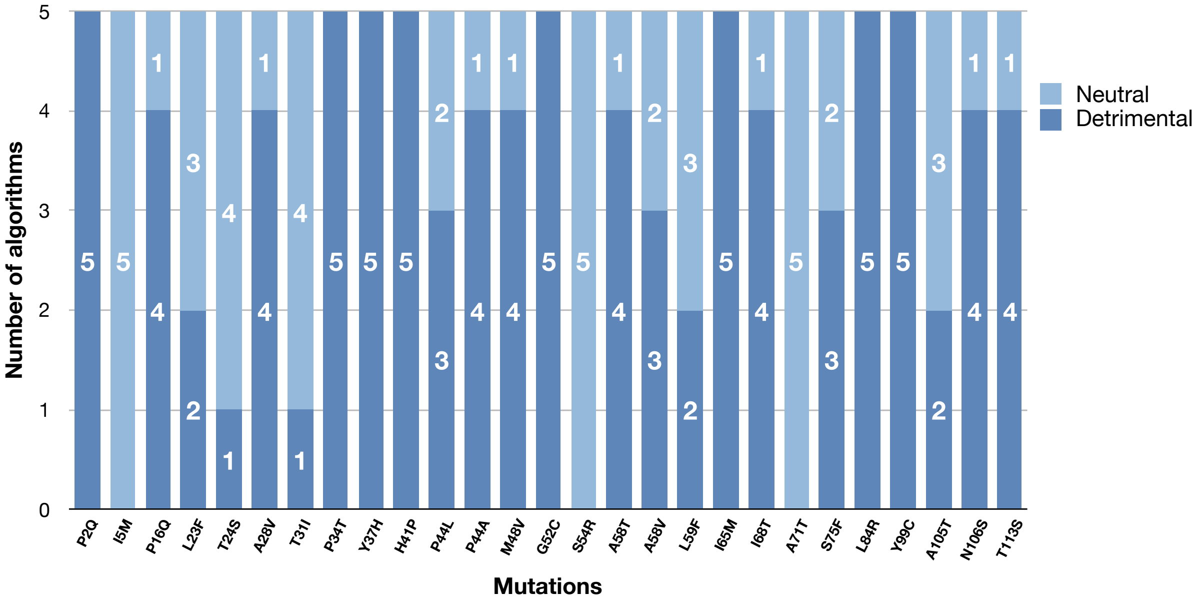


Figure 3.2: Number of 'deleterious' and 'neutral' predictions for each of the MIF non-synonymous mutations via web servers. MutPred, PredictSNP, SIFT, PROVEAN, and PANTHER were used. The darker blue bars indicate deleterious predictions while the lighter blue indicates neutral predictions.

3.4.5 Effects of Missense SNVs on Protein Dynamics

All the MD systems had equilibrated by 50ns of simulation and generally showed linear trajectories.

3.4.5.1 MIF Wildtype Analysis

The MIF wildtype homo-trimeric protein showed the most stable conformation relative to its starting structure (Figure 3.3) with a final RMS deviation of ~ 0.17 nm. For each protein residue in the different systems the RMSF was calculated from the equilibrated MD trajectories.

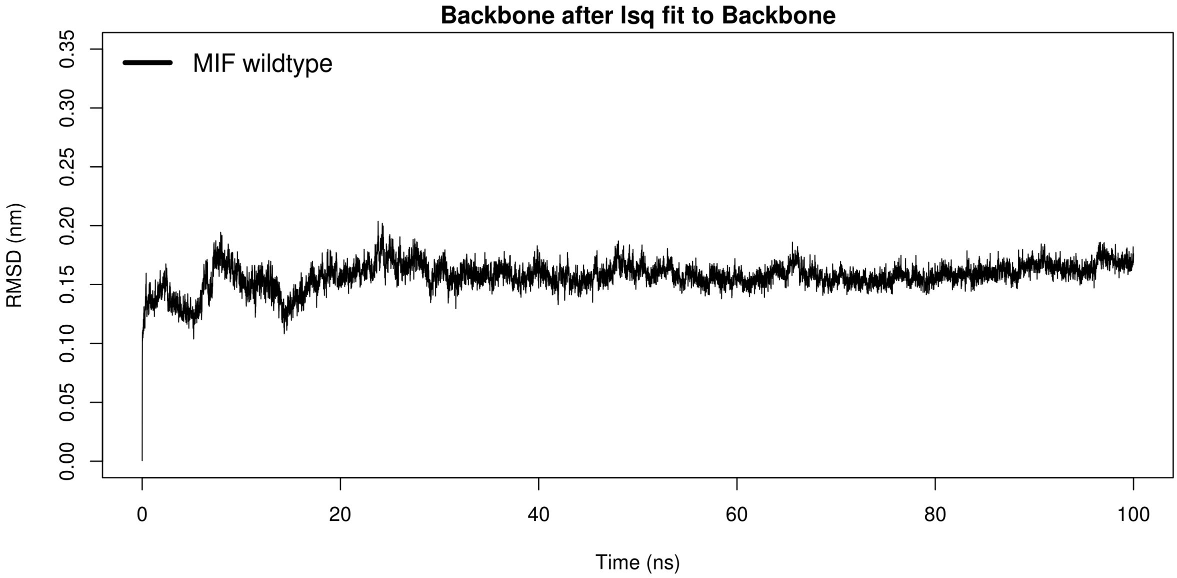


Figure 3.3: Time dependent RMSD of the backbone atoms of the MIF wildtype (black) protein. Backbone least square fitting (lsq) to the backbone of the initial structure.

The residues that showed the highest flexibility were N-like loop residues F50.C - S54.C which are part of the N-like loop and the C-terminus residue A115.A (Figure 3.4). The N-like loop from D48-C57 is known to be important in MIF-CXCR2 binding [188] while C-terminal residues N106 - A115 are known to be key modulators of MIF catalytic activity and stability [193].

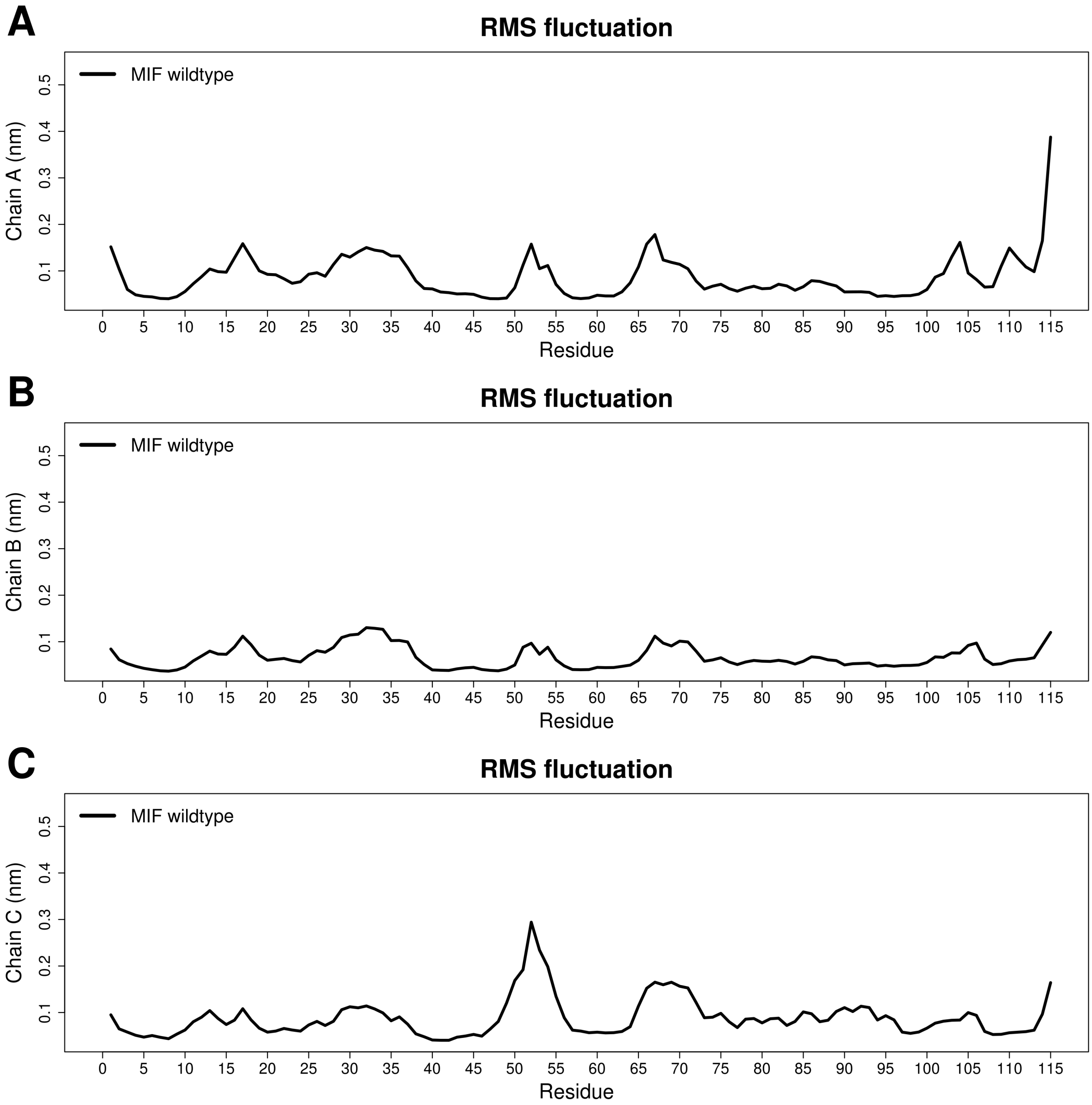


Figure 3.4: RMSF of backbone atoms MIF wildtype (black). A) chain A, B) chain B, and C) chain C.

The *average BC*, 1/Average L, and 1/RMSF generally showed a similar trend as previously reported by Penkler *et al.* [205]. The feature scaled *Average BC*, 1/*Average L*, and 1/RMSF values are shown in Figure 3.5. Across the three chains *Average BC* and 1/*Average L* had an average Pearson correlation coefficient of 0.902, while *Average BC* and 1/RMSF was 0.594, and lastly *Average L* and RMSF had a 0.517 correlation coefficient. The full Pearson correlation coefficients for the metrics for all the systems are shown in Supplementary Table

5. *Average BC* and *Average L* were strongly correlated the wildtype and mutant protein systems.

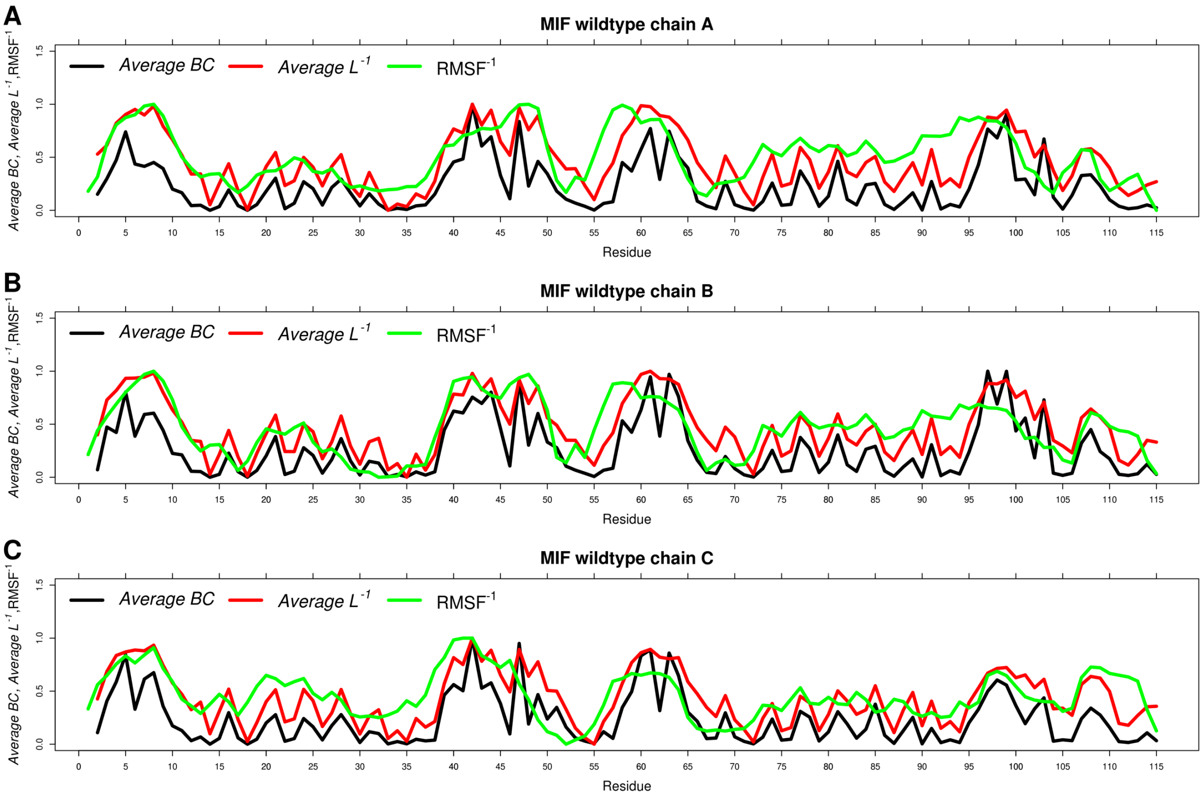


Figure 3.5: Changes in *Average BC* (black), *1/Average L* (red) and 1/RMSF (red). A) chain A, B) chain B, and C) chain C.

Investigation of the protein network properties using *Average BC* showed that the most crucial residues for communication in the MIF wildtype were residues V42.A and Y99.A (Figure 3.5). These were the most central residues in the protein network, residues showing high centrality are important in mediating signals across the protein network [206]. In the MIF wildtype, residues V42.C and S61.B had the shortest *Average L* and were both located in sheets at the chain-chain interface (Figure 3.5). Residues K33.A, G18.A, P35.A, E55.C and Q72.A had the highest *Average L* and were all located on the protein outer surface (Figure 3.1 and Figure 3.5). Residues P35 and E55 were proposed through *in silico* protein-protein docking studies to be involved in MIF/CD74 interaction [191].

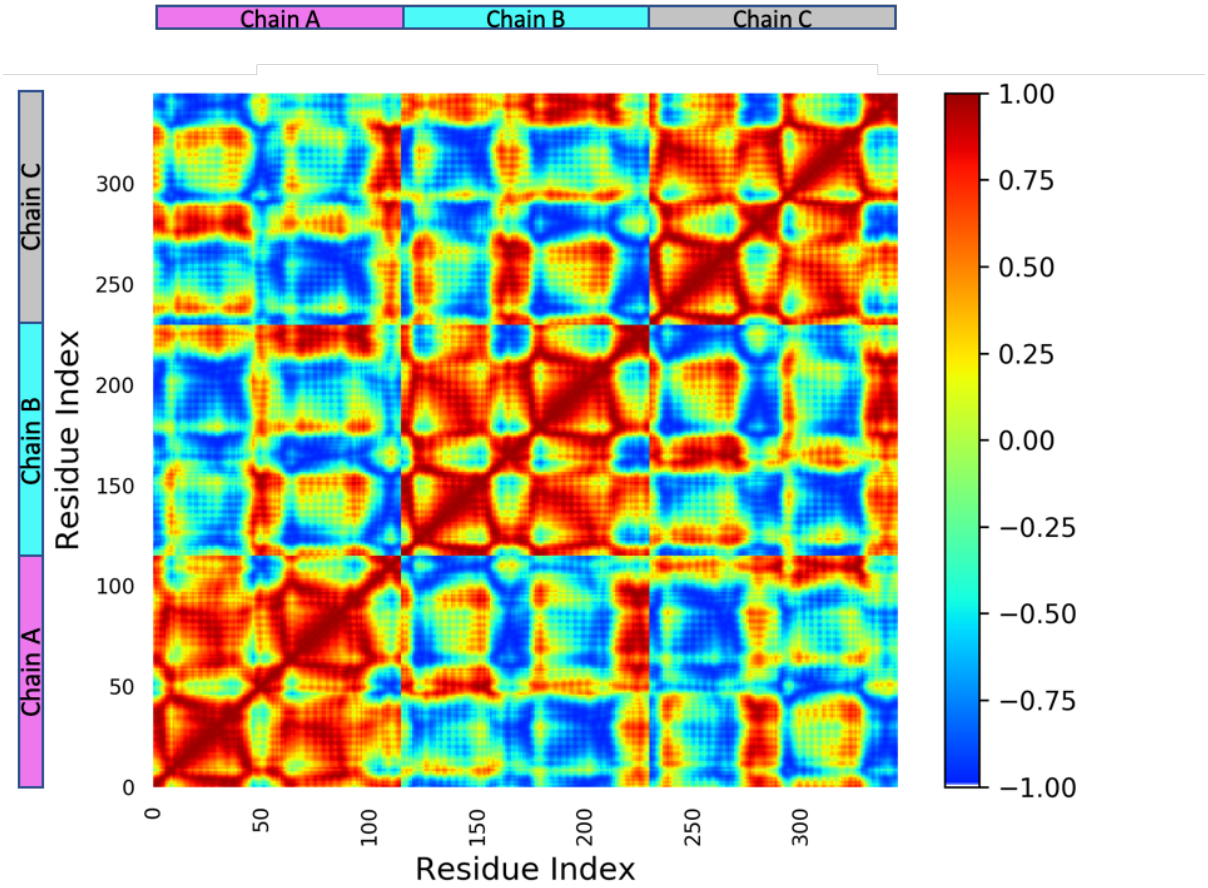


Figure 3.6: Dynamic cross correlation Cα matrices of the MIF wildtype.

Dynamic cross correlation (DCC) was used to examine how the changes in the fluctuations affected the motions of residue pairs in each system using MD-TASK. The movements of P2 and Y37, T31, I65, K67, I68, S75, T113 were highly correlated (> 0.7) and anti-correlated with the movements of G52 (< -0.4) as shown in Figure 3.6. The N-like loop residues F50 – C57 showed high anti-correlation (< -0.7) with the C-terminal residues A105 – A115. Previous studies have showed that the motions of P2 were strongly correlated with Y37, I65, N106, and T113 in the chain A monomeric unit [173] all of which are surface residues important in MIF/CD74 interaction [190].

3.4.5.2 MIF Mutant Analysis

Comparison of the MIF wildtype to the mutants A28V, P34T, P44A, A58T, A58V, L59F, A71T, L84R, Y99C, and A105T showed that the mutants relative to their starting structures did not show any significant conformational changes during MD simulation (Figure 3.7).

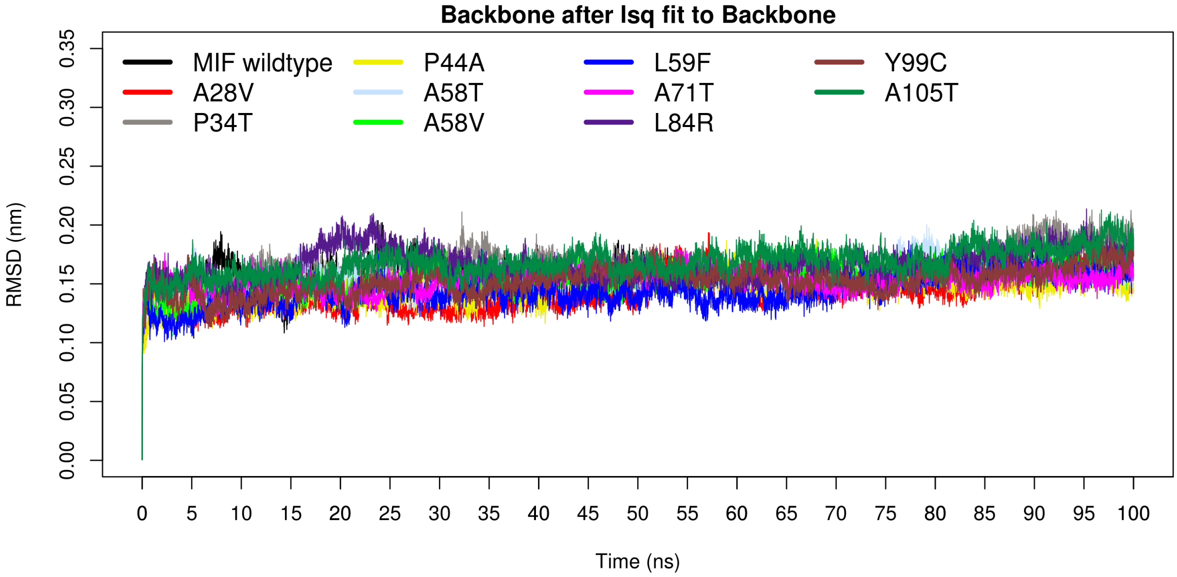


Figure 3.7: Time dependent RMSD of the backbone atoms of the MIF wildtype (black) and the mutant proteins. Shown are the mutations that did show any significant conformational changes in comparison to the MIF wildtype during MD simulation. A28V, P34T, P44A, A58T, A58V, L59F, A71T, L84R, Y99C, and A105T mutants.

Mutants P2Q, I5M, P16Q, L23F, T24S, T31I, Y37H, H41P, M48V, P44L, G52C, S54R, I65M, I68T, S75F, N106S, and T113S on the other hand showed significant conformational changes during MD simulation (Figure 3.8-A) and are the main focus of discussion in the subsequent sections. During simulation there were no significant changes in the protein compactness between the wildtype and mutant structures as measured by the radius of gyration (Rg) (Figure 3.9) which is likely due to the rigidness of the MIF homotrimer [196].

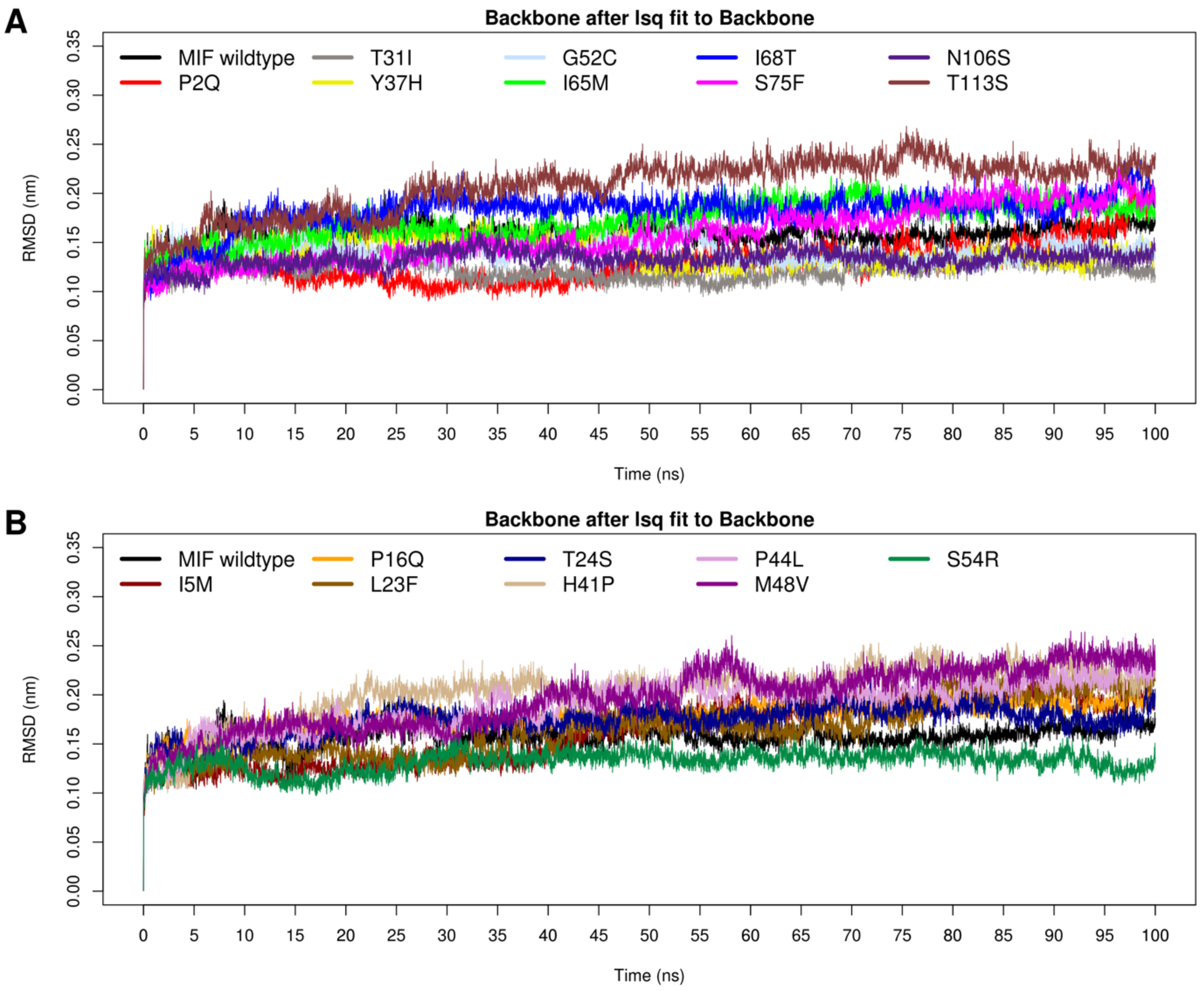


Figure 3.8: Time dependent RMSD of the backbone atoms of the MIF wildtype (black) and the mutant proteins. Shown are the mutations that resulted in conformational changes in comparison to the MIF wildtype during MD simulation. A) P2Q, T31I, Y37H, G52C, I65M, I68T, S75F, N106S and T113S mutants. B) I5M, P16Q, L23F, T24S, H41P, P44L, M48V, and S54R mutants.

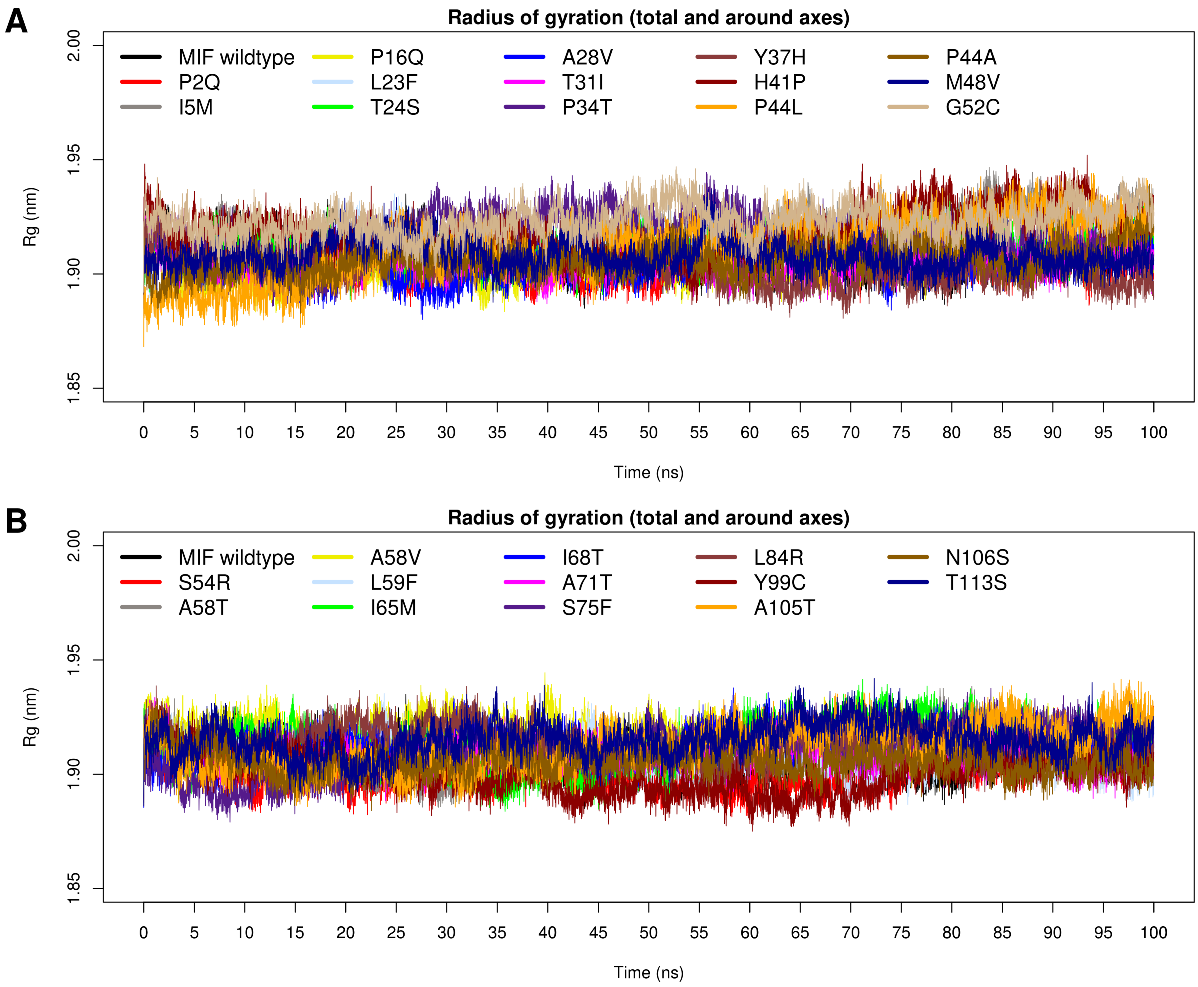


Figure 3.9: MIF 3DJH wildtype (black) and mutant proteins radius of gyration (Rg) of based on Cα atoms versus time at 300K. A) P2Q, I5M, P16Q, L23F, T24S, A28V, T31I, P34T, Y37H, H41P, P44L, P44A, M48V, and G52C mutants. B) S54R, A58T, A58V, L59F, I65M, I68T, A71T, S75F, L84R, Y99C, A105T, N106S, and T113S mutants.

3.4.5.3 Tautomerase Active Site Mutants P2Q and I65M

In comparison to the MIF wildtype the mutants P2Q, T31I, Y37H, G52C, I65M, I68T, S75F, N106S, and T113S showed an overall reduction in the flexibility of the chain C N-like loop (Figure 3.10). Mutants P2Q and I65M were predicted to be detrimental (Figure 3.2) and are located in the Tautomerase active site (Figure 3.1). Mutant I65M is also important in MIF/CD74 interaction (Figure 3.1). MIF tautomerase activity is not considered to be important for MIF leucocyte recruitment activity which is primarily driven by interaction with its CD74, CXCR2 and CXCR4 receptors [190,197,198].

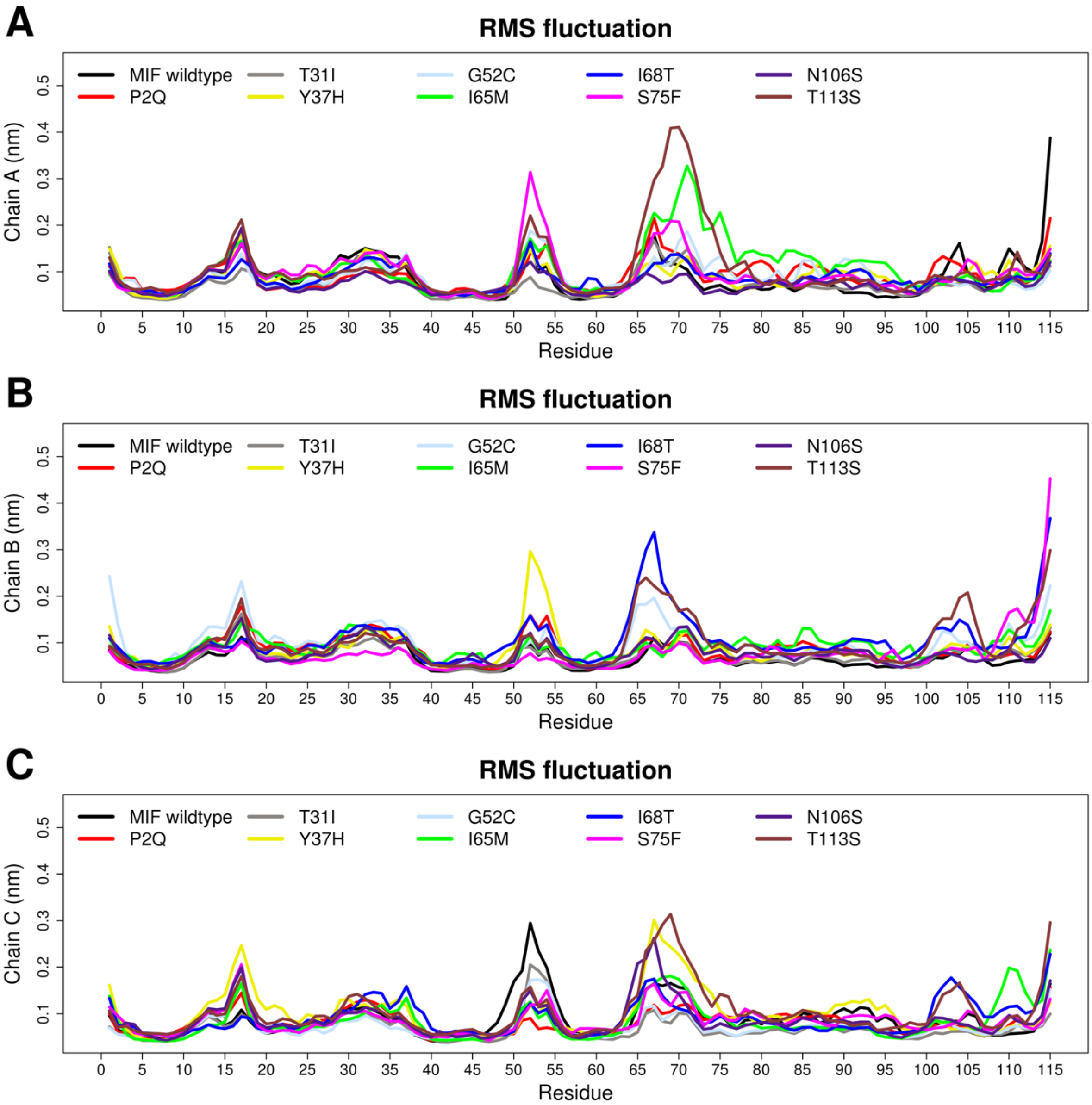


Figure 3.10: RMSF of backbone atoms MIF wildtype (black), P2Q, T31I, Y37H, G52C, I65M, I68T, S75F, N106S, and T113S mutant proteins as a function of time. A) Chain A RMS fluctuations, B) Chain B RMS fluctuations, and D) Chain C RMS fluctuations.

However, studies have shown that MIF tautomerase inhibitors such as N-acetyl-p-benzoquinone imine (NAPQI) [177,207] and (*S,R*)-3-(4-hydroxyphenyl)-4,5-dihydro-5-isoxazole acetic acid methyl ester (ISO-66) [208] result in loss of tautomerase activity and also impaired MIF physiological activity. The loss of MIF leucocyte recruiting activity is due to inhibitor NAPQI forming aromatic hydrophobic interactions with Y37 (PDB:3DJI) [177] while inhibitor ISO-66 forms a hydrogen bond with I65 (PDB:4K9G) which are both important in MIF/CD4 interaction. This explains why MIF tautomerase inhibitors such as NAPQ1 appear to impede MIF physiological activity.

Conformation, flexibility and PCA analysis

Mutant P2Q and I65M protein systems showed conformational changes when compared to the MIF wildtype using RMSD (Figure 3.8-A). Their RMSF further showed an overall reduction in the flexibility of the chain C N-like loop residues F50.C – S54.C in both mutants (Figure 3.4a). Mutant I65M also showed increases in flexibility of residues G70.A – S75.A and N110.C (Figure 3.4a). Figure PCA analysis based on the Cartesian coordinates showed that mutants P2Q and I65M showed differing conformational transitioning from the MIF wildtype as shown by the 2D projection of PC1 and PC2 (Figure 3.11). In the MIF wildtype PC1 and PC2 explained 47% and 22% of the observed variance respectively. In the P2Q mutant PCA, PC1 and PC2 explained 51% and 22% of the observed variance respectively (Figure 3.11-B). In the I65M mutant PCA, PC1 and PC2 explained 53% and 23% of the variance observed (Figure 3.11-C). P2Q had a trace of 8.24 nm^2 while I65M had one of 10.68 nm^2 while the MIF wildtype had one of 8.25 nm^2.

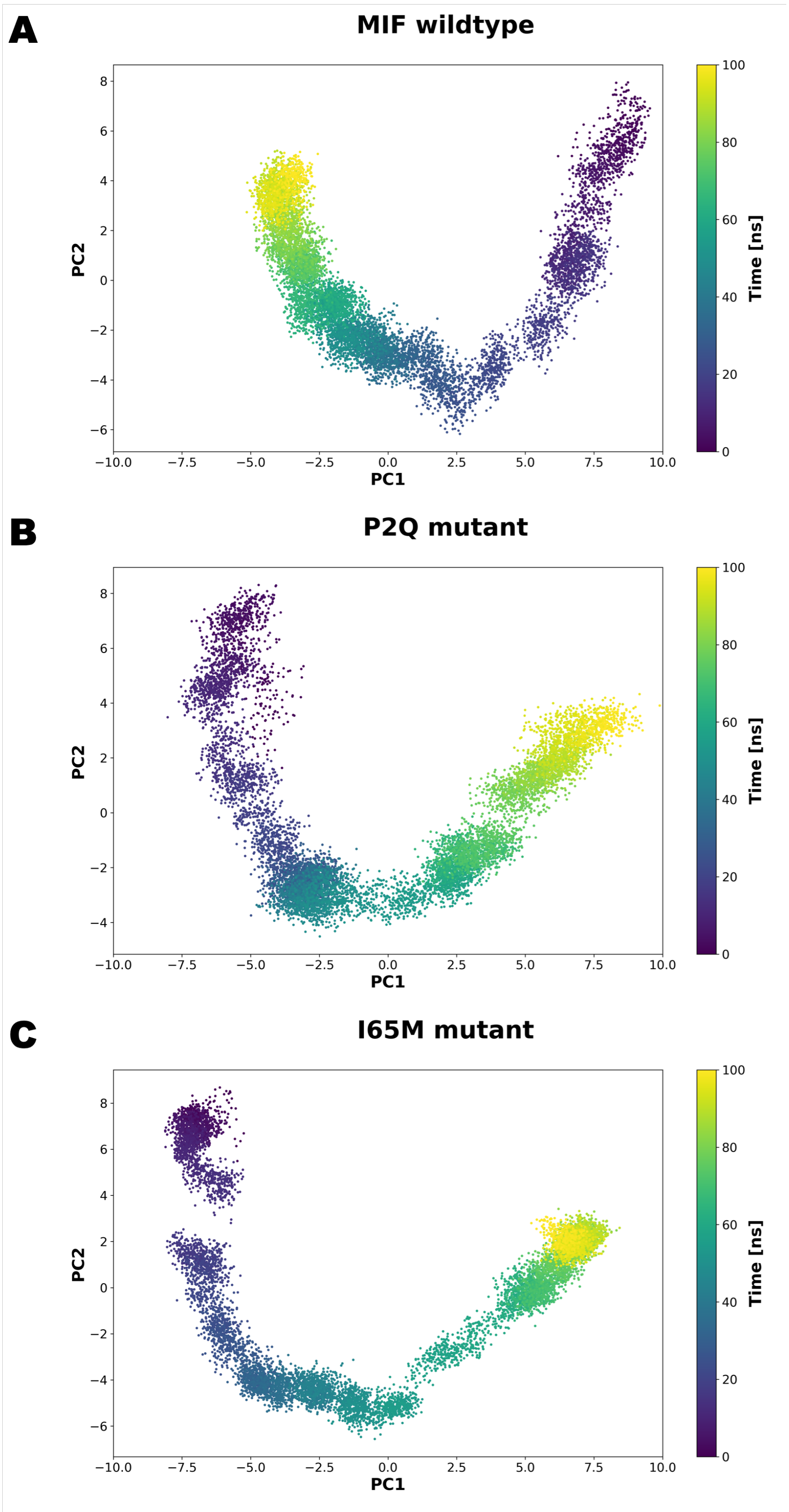


Figure 3.11: Principal component analysis of MIF wildtype and mutant protein MD systems. The motions of the proteins during 100 ns all-atom MD production are shown along the first and second principal components (PC1 and PC2). A) MIF wildtype, B) P2Q mutant, and C) I65M mutant.

Dynamic residue network (DRN) analysis

Compared to the MIF wildtype mutants P2Q and I65M showed a reduction in the *Average L* of N-loop residues S54.C and E55.C (Figure 3.12). The MIF chain A and B show a similar *Average L* but not with chain C because MIF has been reported to have a pseudo 2-fold symmetry [185]. The changes in the *Average L* of the N-like loop especially as shown by a similar trend in the RMSF (Figure 3.4a) are indicative of changes in spatial displacement due to the introduction of a mutation [209].

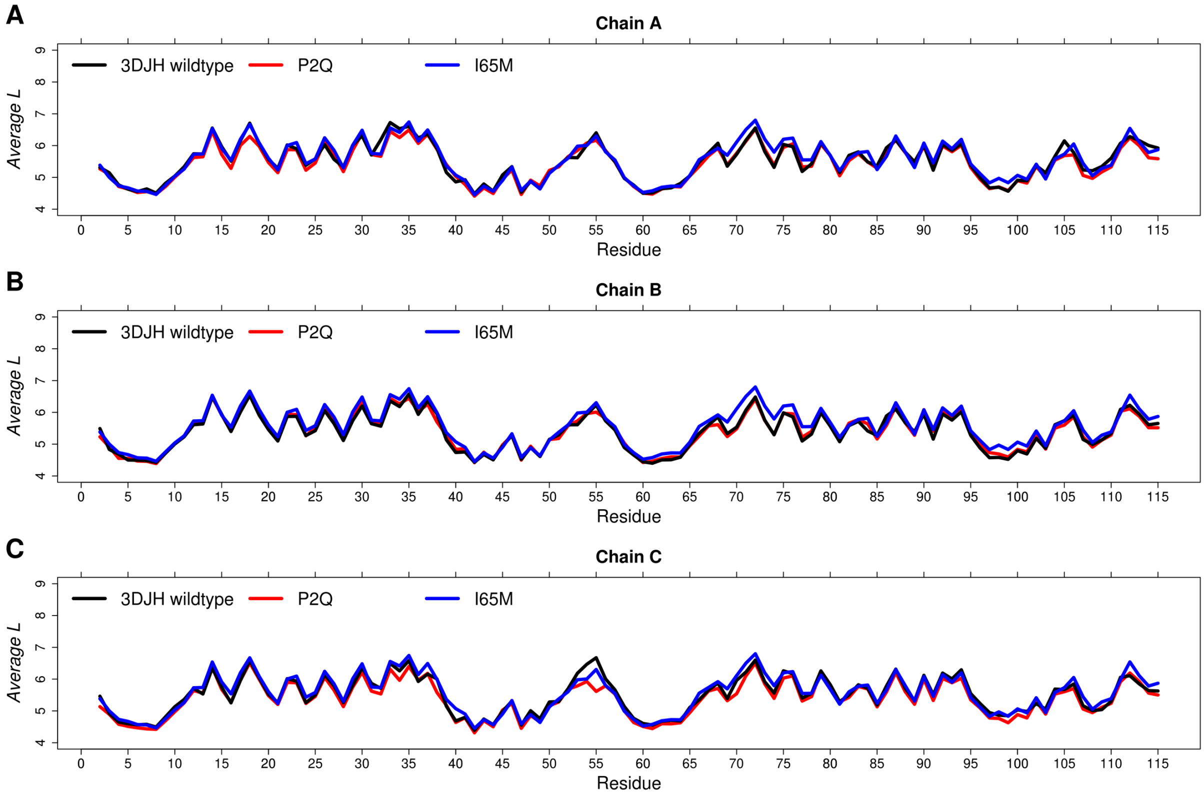


Figure 3.12: Changes in *Average L* of the MIF wildtype, P2Q and I65M mutant proteins during MD production. A) Chain A, B) Chain B, and C) Chain C.

The two mutants appear to cause changes in the dynamic communication flow within the protein network as indicated by changes in residue Betweenness centrality (BC) measure *Average BC*. Mutant P2Q had residues I97.B and Y99.B showing lowered centrality when compared to the MIF wildtype (Figure 3.13). In addition, mutant I65M showed local changes in centrality close to the mutation site with residues H63.B, S64.B, Y96.A-N98.A, and Y96.B-N98.B showing lowered centrality while Y99.C showed increased centrality (Figure 3.13). In the two mutants *Average BC*, 1/*Average L* and 1/RMSF correlations generally showed a similar trend as the MIF wildtype (Supplementary Table 5).

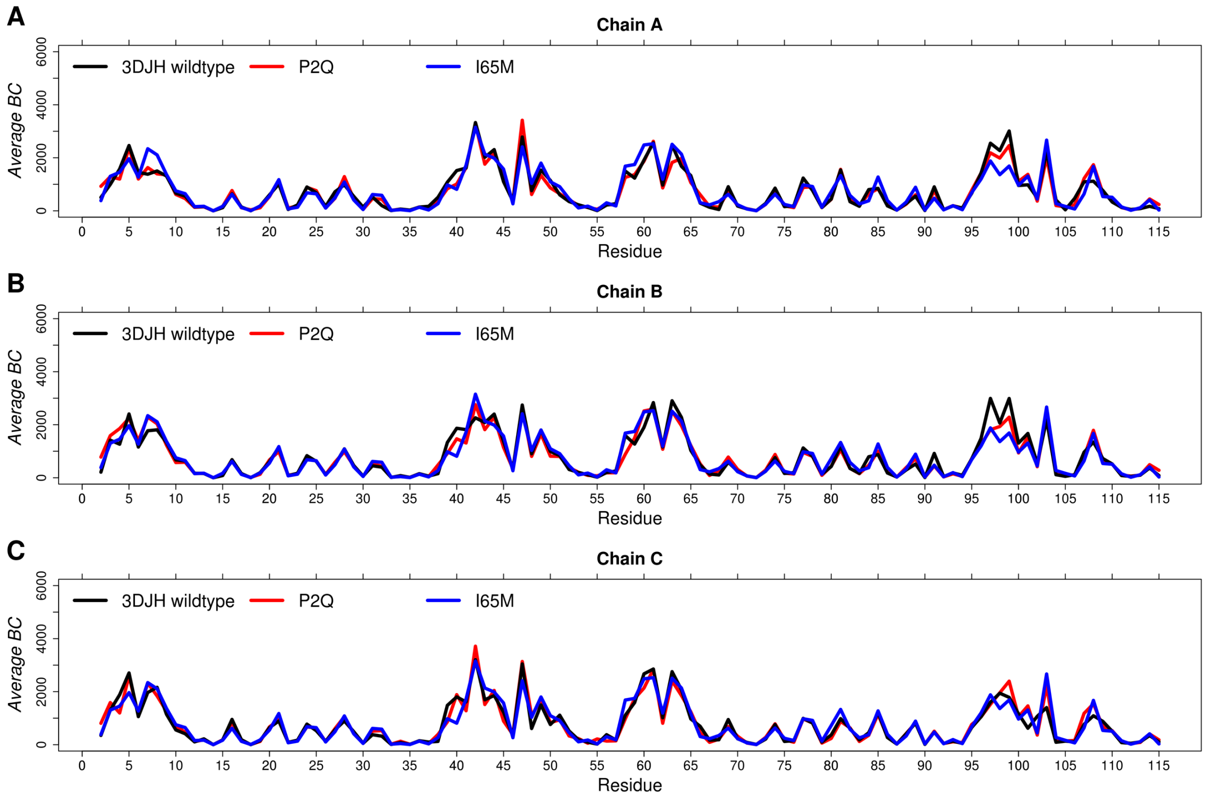


Figure 3.13: Changes in *Average BC* of the MIF wildtype, P2Q and I65M mutant proteins during MD production. A) Chain A, B) Chain B, and C) Chain C.

N-terminal P2 residue is central and arguably critical to the both MIF tautomerase activity and MIF/CD74 interaction. Contact map analysis revealed that in the MIF wildtype, P2 appears to make contacts with T31, Y37, H63, S64, I65 and I68 during MD simulation some of which are associated with MIF tautomerase activity and MIF/CD74 interaction (Figure 3.14).

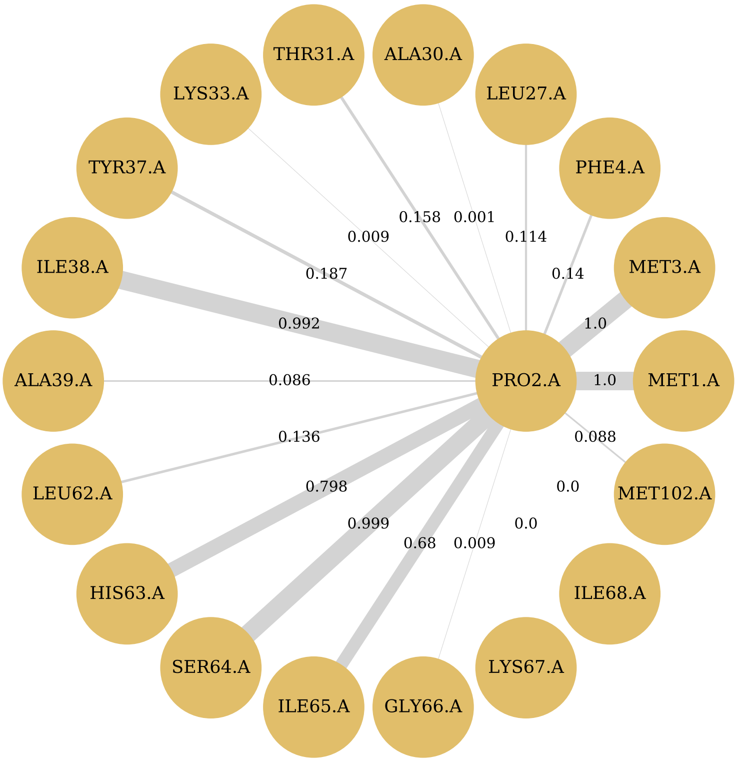


Figure 3.14: A weighted residue interaction map generated across the trajectory for the chain A residue P2 of the MIF wildtype protein.

A comparative residue contact map showed that the mutation of P2 to Q2 altered the residue interaction network around the amino acid position (Figure 3.15). In the mutant P2Q there was an increased residue-residue interaction frequency (RRIF) between Q2.A with Y37.A are hydrogen bonded [Wildtype (WT) RRIF: 0.187; Mutant (MT) RRIF: 0.355] and with K67.A (WT RRIF: 0; MT RRIF 0.013) (Figure 3.15). There was a reduced frequency of interaction with H63.A (WT RRIF 0.798; MT 0.462) and I65.A (WT RRIF: 0.680; MT RRIF: 0.341) (Figure 3.15).

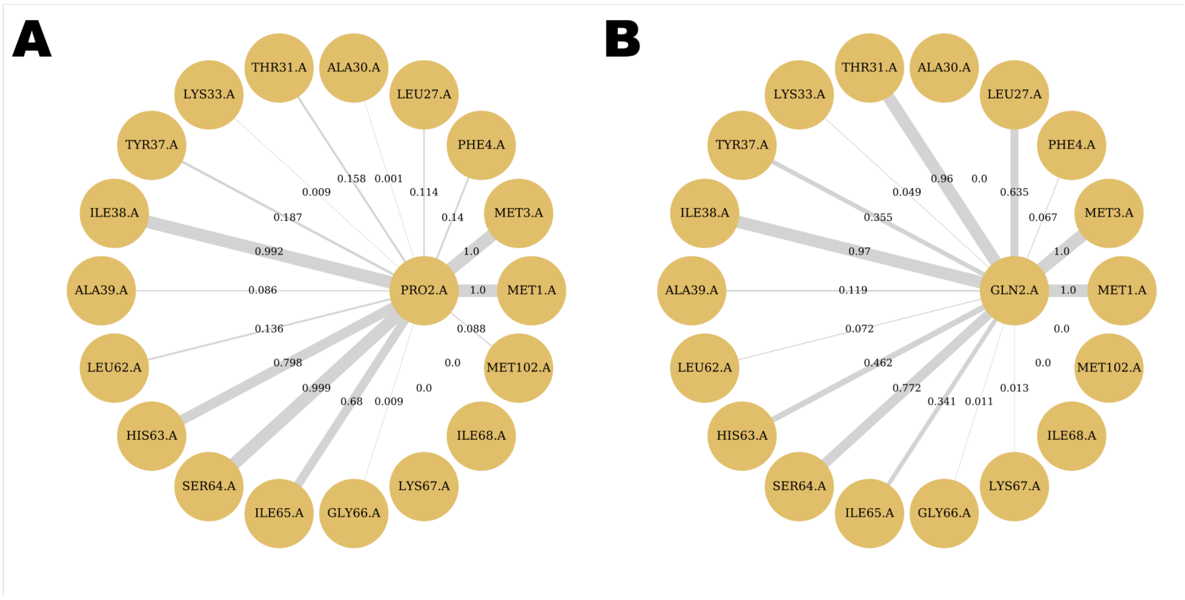


Figure 3.15: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and P2Q mutant generated using MD-TASK. A) MIF wildtype chain A wildtype and B) P2Q chain A.

The changes in the residue interaction network around the position appear to also affect the motion of protein residues as shown in the DCC where in the P2Q mutant residue pairs I65.B & K67.C and K67.C & 1M.B became anti-correlated in their motions when compared to the MIF wildtype (Figure 3.16 and Supplementary Table 6). Residues I65 and K67 are both known to be important in MIF/CD74 interaction with their mutation resulting in impaired MIF biological activity [190].

The N-terminal P2 residue has long been known to be an important determinant of MIF biological function [190,197,198]. In our study we show using MD and DRN how mutation can impair its residue interaction network resulting in changes in protein dynamics and possibly function.

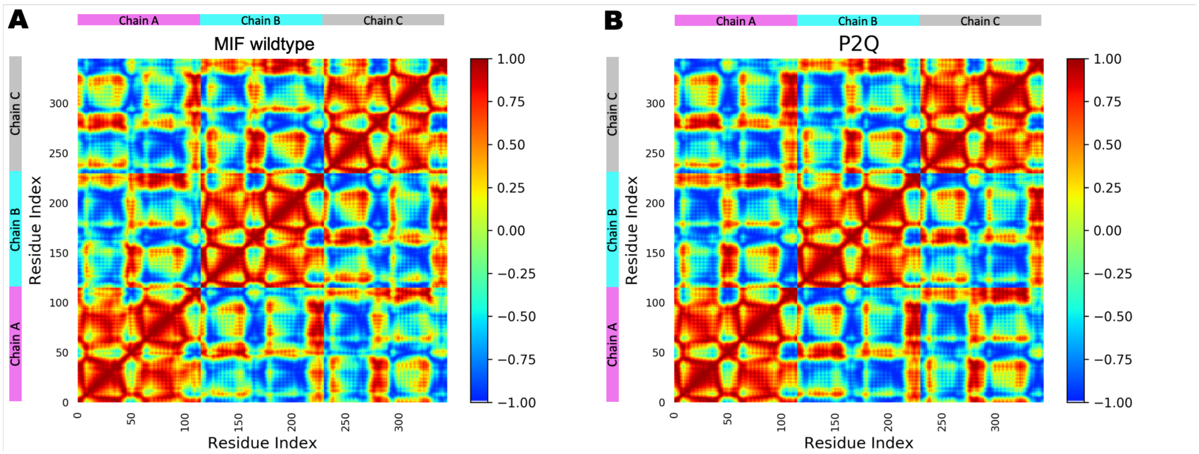


Figure 3.16: Dynamic cross correlation Cα matrices of A) MIF wildtype and B) P2Q. Blue indicates indicating anti-correlation, green no correlation, and red indicating correlation. The side bar indicates chain A (pink), chain B (teal) and chain C (gray). A) DCC for MIF wildtype and B) P2Q mutant

Lastly mutant I65M, showed that residue M65.A gained an interaction with K67.A that forms a vdW interaction with G69.A resulting in the increased flexibility observed in chain A residues G69.A - S75.A (Figure 3.17 and Figure 3.10).

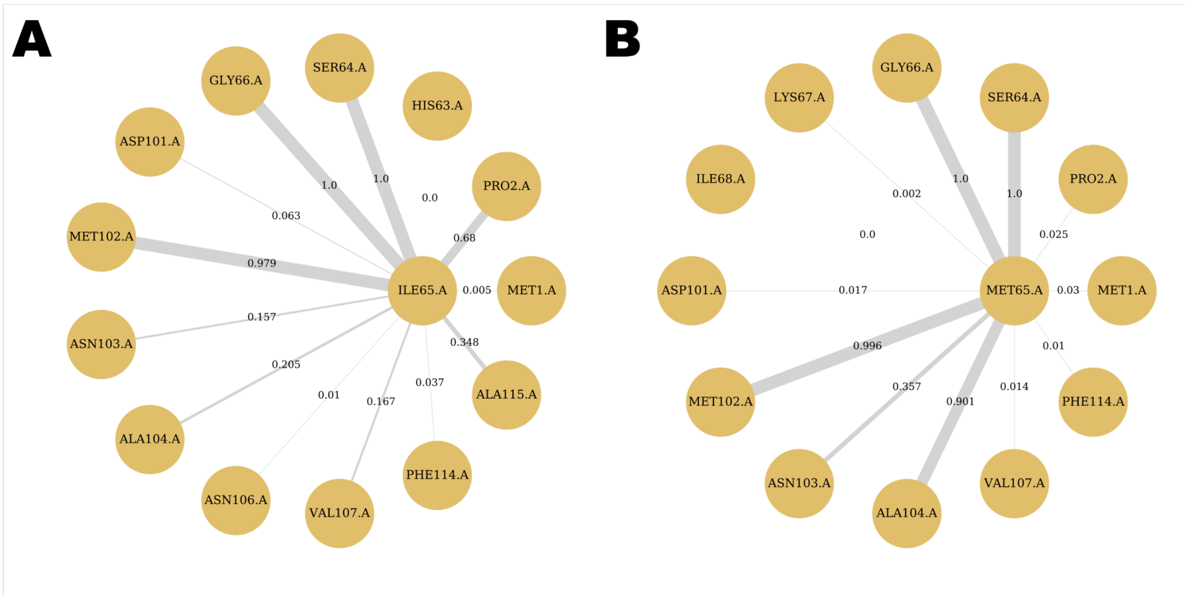


Figure 3.17: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and I65M mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) I65M chain A.

Additionally M65.A when compared to the wildtype I65.A had a markedly reduced RRIF with P2.A (WT RRIF: 0.680; MT RRIF: 0.025; Figure 3.15). In the MIF wildtype structure amino acid positions P2 and I65 are very close to each other, their Cβ atoms are ~ 7 Å apart and appear to part of a similar residue interaction network.

3.4.5.4 MIF/CD74 Binding Site Mutants T31I, Y37H, G52C, I68T, S75F, N106S, and T113S

Our DRN analyses lead us to believe that residues P2, T31, Y37, G52, I65, I68, S75, N106, and T113 appear to be part of a similar local residue interaction network. This local residue interaction network is located on the MIF/CD74 interaction surface [190]. As shown in Figure 3.18, based on measurement of their Cβ backbone carbons, P2 is 7.0 Å from I65 and 6.5 Å from Y37 (Figure 3.18-A). While I65 is 7.8 Å from I65, 6.6 Å from K67, and 9.2 Å from N106 (Figure 3.18-B). N106 is in turn 6.9 Å from T113 that is 5.4 Å from W109 (Figure 3.18-C). T31 is 6.2 Å from K67 and 9.8 Å from S75 (Figure 3.18-D). Further, as shown in the MIF wildtype DCC the motions of P2 were strongly correlated with Y37, I65, N106, and T113 in the chain A monomeric unit [173]. These outer surface residues appear to be part of a similar local residue interaction network with functional significance.

Residues Y37, I65, K67, W109, and N110 are experimentally validated MIF/CD74 interaction residues [190] (Table 3.2). While residues N106 and T113 are located in the carboxy-terminal whose flexibility is known to be a key modulator of MIF immunological activity [193] (Table 3.2). Residue G52 is located within the MIF N-like loop that is of importance in MIF-CXCR2 binding and MIF physiological function [188] (Table 3.2).

Conformation, flexibility and PCA analysis

Relative to their initial structures and comparison with the MIF wildtype structure T31I, Y37H, G52C, I68T, S75F, N106S, and T113S mutant proteins showed marked conformational changes, more so for (Figure 3.8-A).

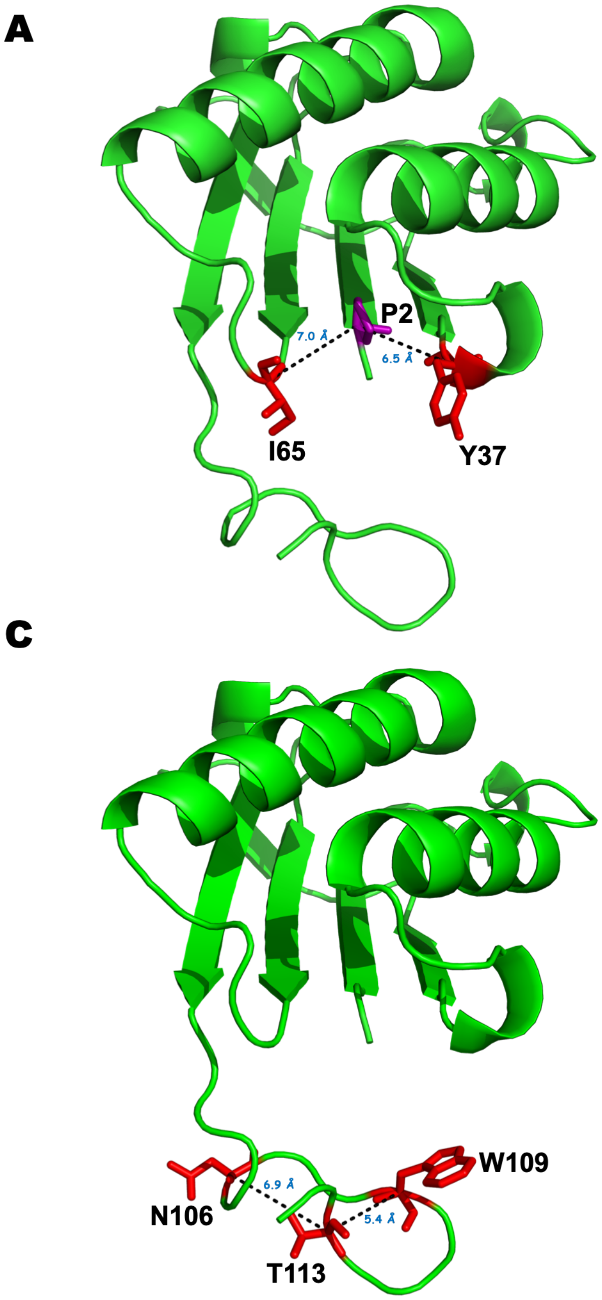


Figure 3.18: The 3D structure of MIF homology model numbered to start from M1 to A115 (PDB:3DJH). A cartoon representation of the MIF chain A that is colored green with the MIF/CD74 interacting residues shown in stick format and colored red.

Compared to the MIF wildtype protein, all the mutants showed a reduction in the flexibility of the chain C N-like loop (residues F50.C – S54.C; Figure 3.10). Mutant T31I also showed reduction in the flexibility of residue A115.A (Figure 3.10) while Y37H showed additional increased flexibility of chain B N-like loop residues G52.B – E55.B, P16.C, D17.C, K67.C and

I68.C (Figure 3.10). Mutant G52C showed increased flexibility of residues M1.B and D17.B (Figure 3.10) while I68T showed increased flexibility of G66.B – I68.C, A115.B and reduced flexibility of A115.A (Figure 3.10). Mutant S75F showed increased flexibility in the N-like loop residues G52.A – S53.A, G70.A, G69.A, A115.B, D17.C, and a decreased flexibility in A115.A (Figure 3.10). Mutant N106S showed increased flexibility in K67.C and decreased flexibility of A115.A (Figure 3.10). Lastly, mutant T113S showed increased flexibility of residues K67.A – G72.A, I65.B, G69.B, A104.B, A115.B, G69.C, G70.C and decreased flexibility in A115.A (Figure 3.10).

PCA of the MD systems revealed differences in the conformational transitioning of the mutants and wildtype protein as shown by Figure 3.19. In the T31I mutant PCA, PC1 and PC2 explained 38% and 24% of the observed variance respectively (Figure 3.19-B). In the Y37H mutant PCA, PC1 and PC2 explained 62% and 17% of the observed variance respectively (Figure 3.19-C). In the G52C mutant PCA, PC1 and PC2 explained 55% and 20% of the observed variance respectively (Figure 3.19-D). In the I68T mutant PCA, PC1 and PC2 explained 46% and 23% of the observed variance respectively (Figure 3.19-E). In the S75F mutant PCA, PC1 and PC2 explained 55% and 18% of the observed variance respectively (Figure 3.19-F). In the N106S mutant PCA, PC1 and PC2 explained 44% and 26% of the observed variance respectively (Figure 3.19-G). Lastly, in the T113 mutant PCA, PC1 and PC2 explained 58% and 16% of the observed variance respectively (Figure 3.19-H). On the PC1 axis T31I showed the least difference between the initial and final structural conformations, while T113S showed the greatest difference.

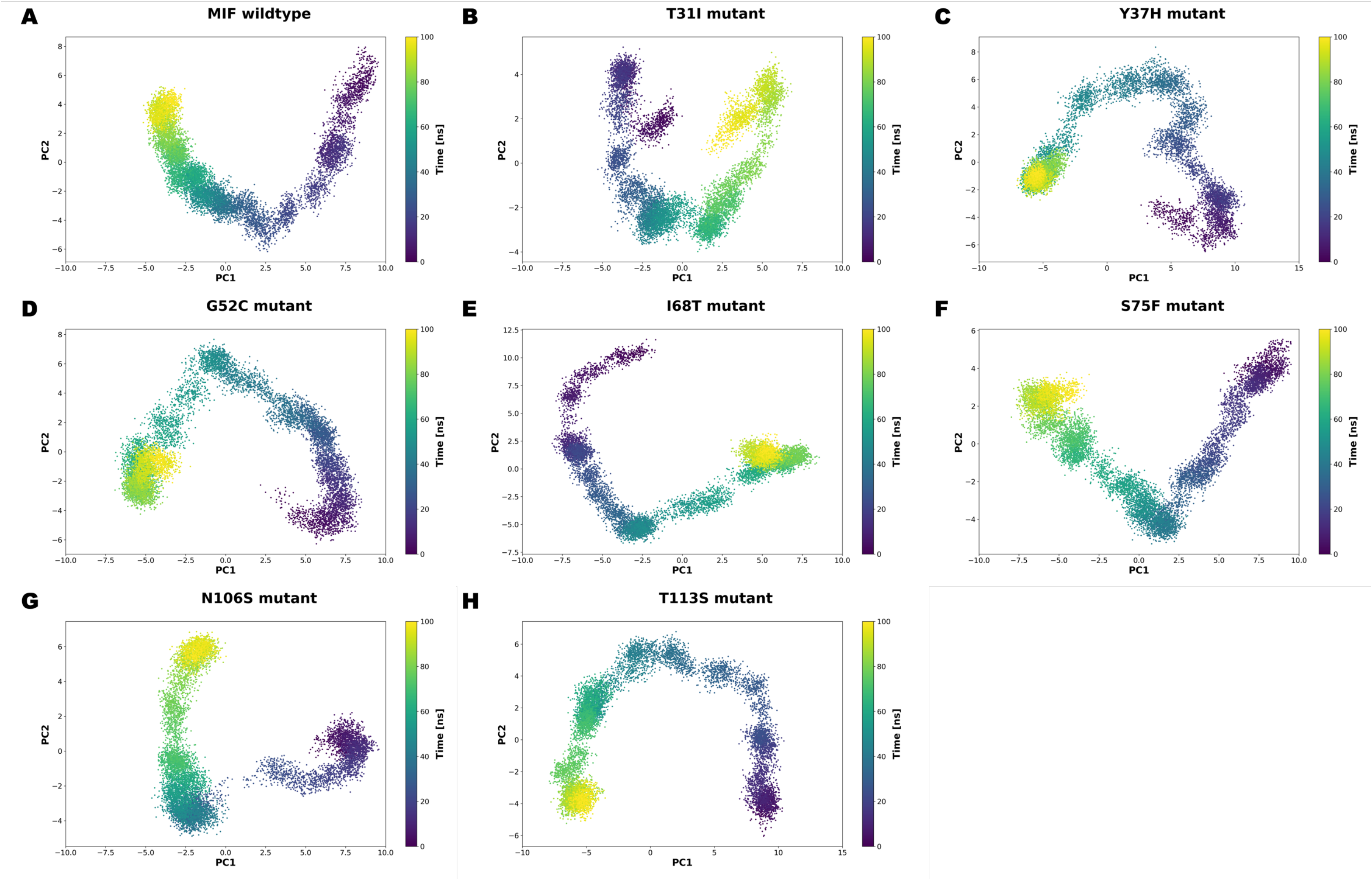


Figure 3.19: Principal component analysis of MIF wildtype and mutant protein MD systems. The motions of the proteins during 100 ns all-atom MD production are shown along the first and second principal components (PC1 and PC2). A) MIF wildtype, B) T31I mutant, C) Y37H mutant, D) G52C mutant, E) I68T mutant, F) S75F mutant, G) N106 mutant, and H) T113S mutant.

Dynamic residue network (DRN) analysis

Changes in the flexibility of the chain A and chain C N-like loop were also observed in mutants T31I, Y37H, G52C, I68T, S75F, N106S and T113S (Figure 3.10). These changes in fluctuation also coincided with changes in spatial displacement of the N-like loop as measured by *Average L* (Figure 3.20).

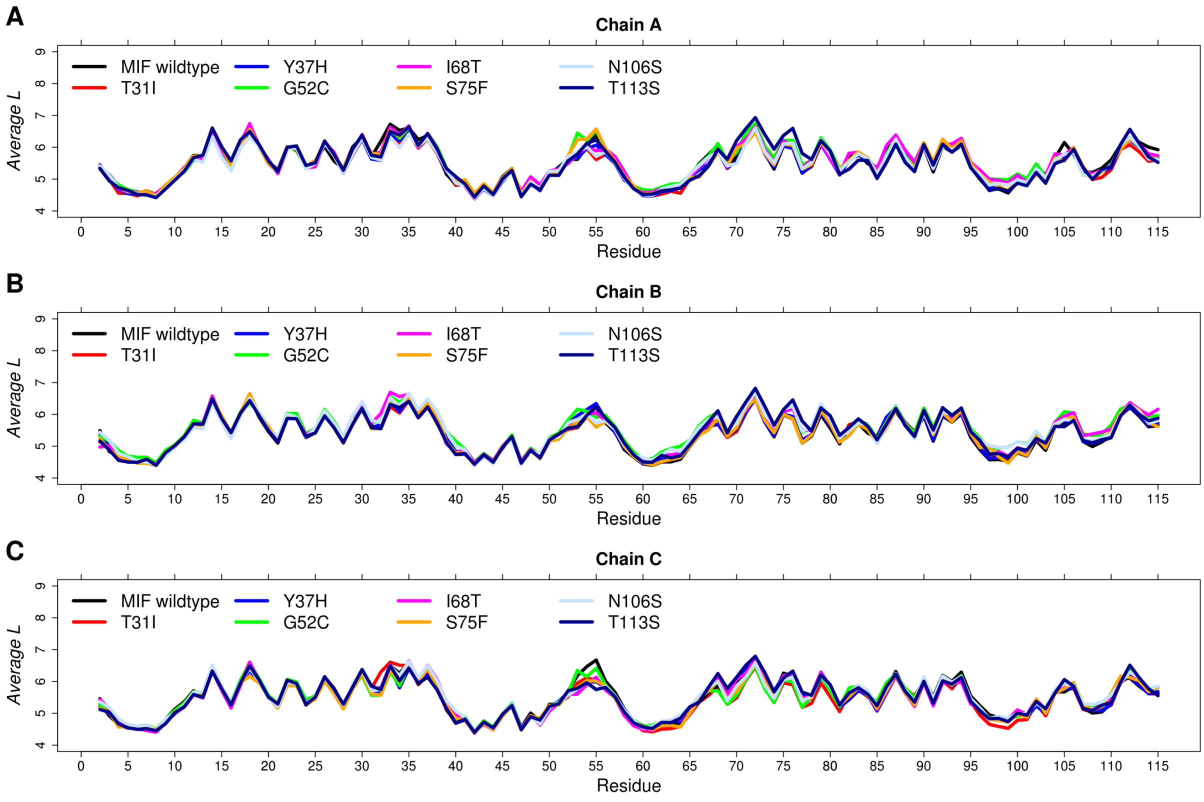


Figure 3.20: Changes in *Average L* of the MIF wildtype, T31I, Y37H, G52C, I68T, S75F, N106S, and T113S mutant proteins during MD production. A) Chain A, B) Chain B, and C) Chain C.

The changes observed changes in the N-like loop of the mutants is an important finding because the N-like loop L47-P56 (LMAFGGSSEP) has been shown to be of importance in MIF-CXCR2 binding which is critical for MIF macrophage recruitment [188]. Mutant also showed further spatial displacement of helix residues Y76.A and Y76.B (Figure 3.20). The mutants when compared to the MIF wildtype showed changes in communication flow within their respective protein networks, particularly of not was the change in the *Average BC* of residues I97 and Y99 which is known to be functionally significant [173] (Figure 3.21).

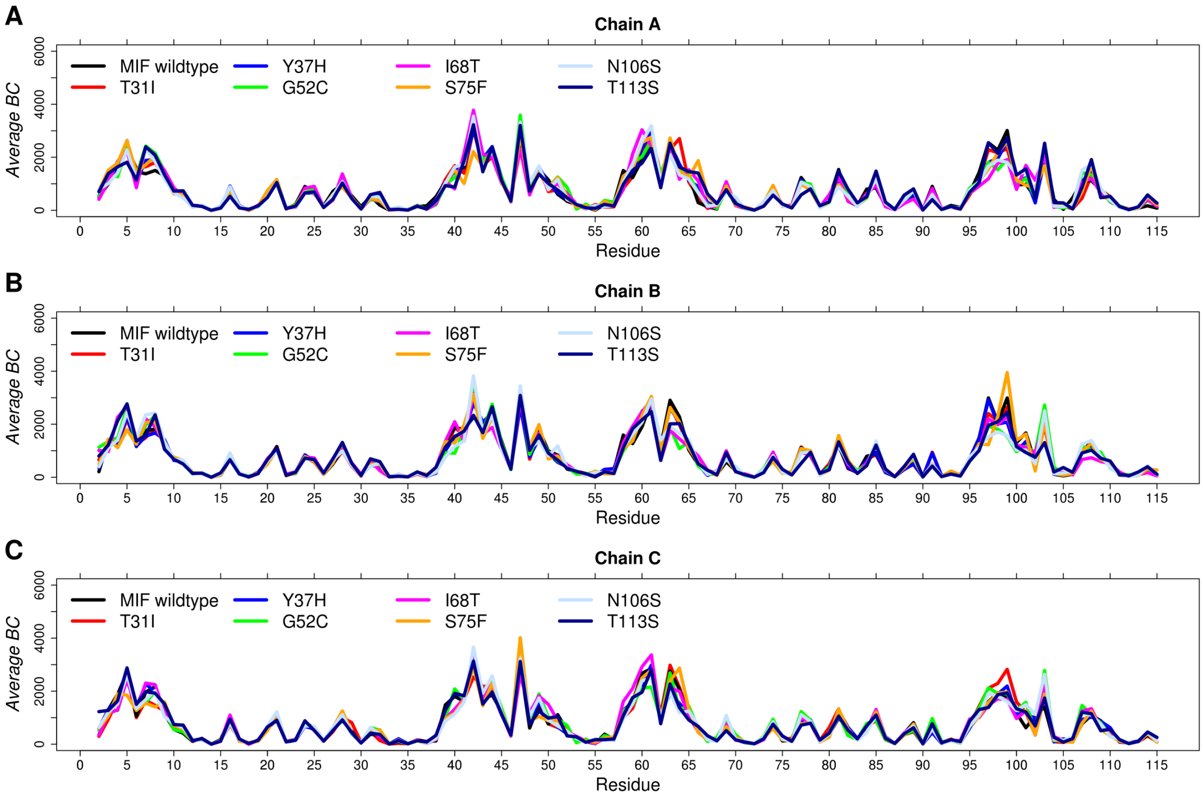


Figure 3.21: Changes in *Average BC* of the MIF wildtype, T31I, Y37H, G52C, I68T, S75F, N106S, and T113S mutant proteins during MD production. A) Chain A, B) Chain B, and C) Chain C.

Mutant T31I showed increased centrality of residue S64.A while Y37H had residue H63.C showing decreased centrality (Figure 3.21-C). Mutant G52C showed decreased centrality of residues H63.C, C60.C and increased centrality of N106.C (Figure 3.21-C). I68T showed increased centrality of N7.A and C60.A (Figure 3.21-A). S75F showed decreased centrality of H63.B and increased centrality of G66.A (Figure 3.21-A). N106 showed decreased centrality of residues H63.B and increased centrality of V42.C and N103.C (Figure 3.21-B and 4.21-C). Lastly, T113S showed decreased centrality of H63.B and increased centrality of N7.A and G66.A (Figure 3.21-A).

Of note, as seen in mutant P2Q and I65M mutation appears to alter the residue-residue interaction network arounds around mutation sites. For the T31I mutant, I31.A showed an increase in RRIF with P2.A (WT RRIF: 0.158; MT RRIF: 0.589), K67.A (WT RRIF: 0.024; MT RRIF: 0.470), and S64.A (WT RRIF: 0.002; MT RRIF: 0.354) and a reduction in RRIF with Y37.A (WT RRIF: 0.089; MT RRIF: 0.056) (Figure 3.22).

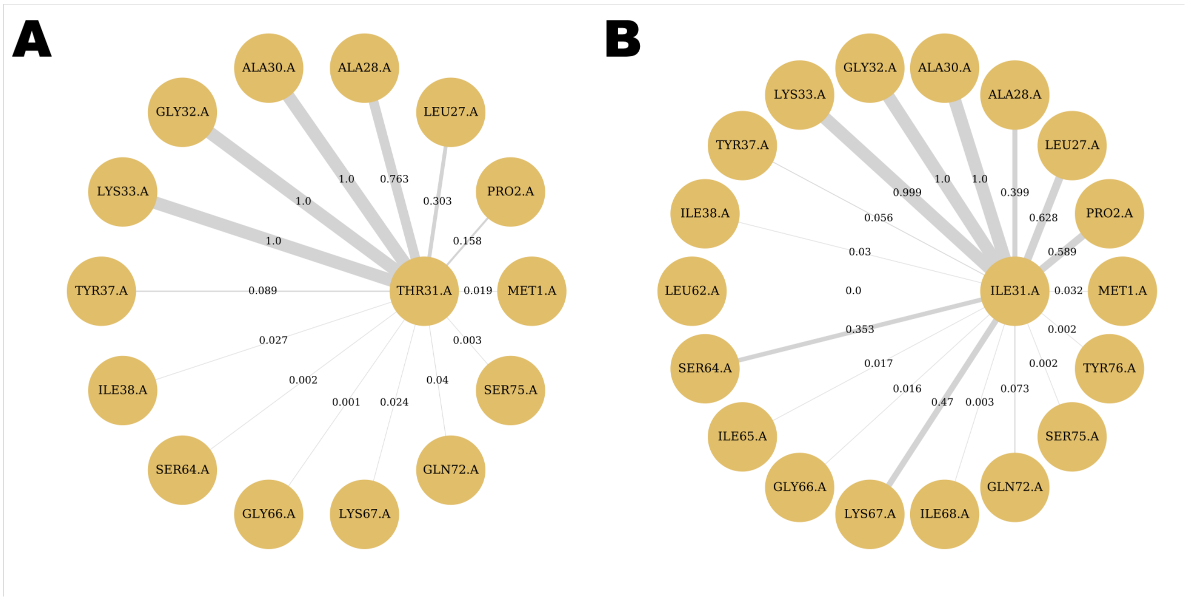


Figure 3.22: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and T31I mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) T31I chain A.

The Y37H mutant, H37.A showed an increased RRIF with catalytic residue P2.A (WT RRIF: 0.187; MT RRIF: 0.35) and residue T31.A (WT RRIF: 0.089; MT RRIF: 0.345) (Figure 3.23).

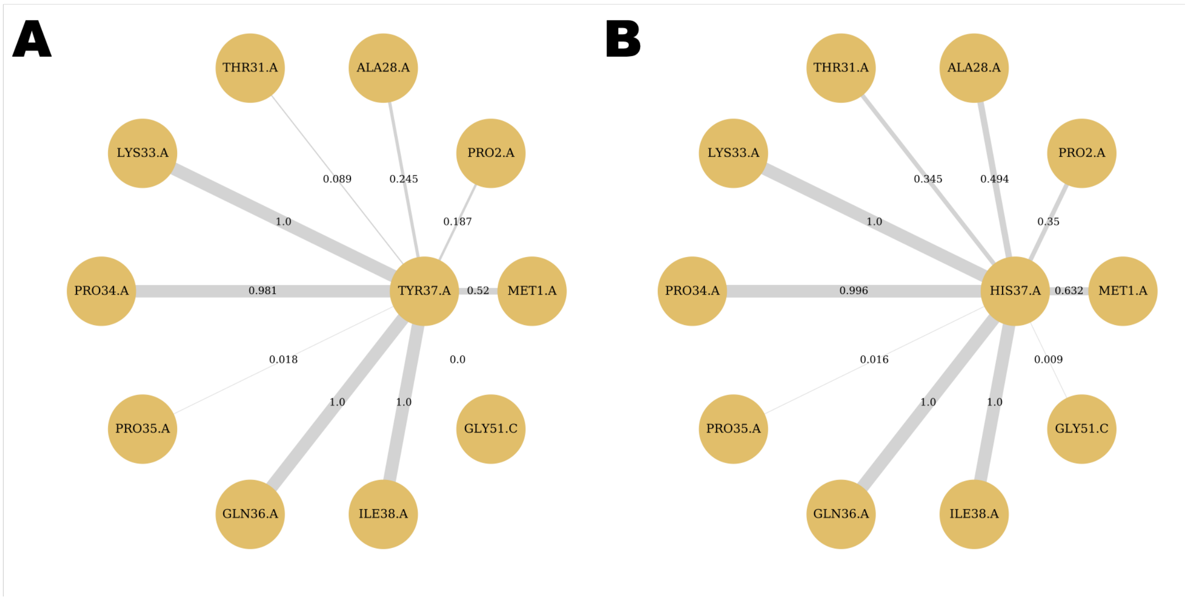


Figure 3.23: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and Y37H mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) Y37H chain A.

In mutant G52C, C52.A lost a contact with S54.A, an increased contact with P35.B (WT 0.185; MT 0.625) and Q36.B (WT 0.102; MT 0.869) (Figure 3.24).

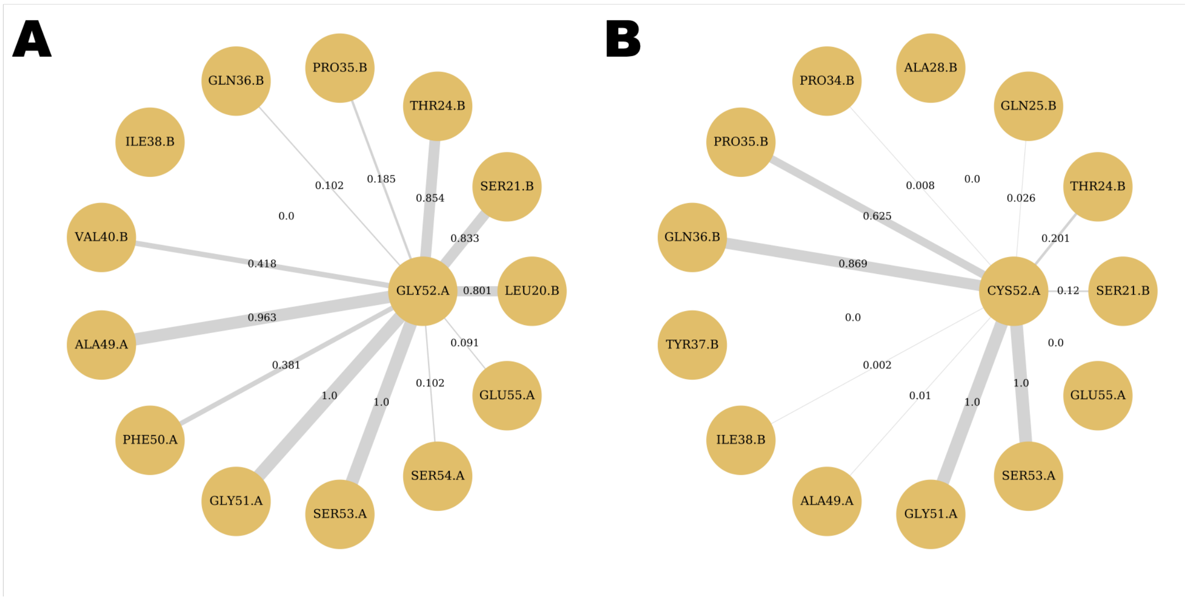


Figure 3.24: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and G52C mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) G52C chain A.

In mutant I68T we observed a reduction in the RRIF between T68.A and G66.A (WT 0.328; T: 0.176) and a reduced RRIF with N106 (WT 0.657; T: 0.077) (Figure 3.25).

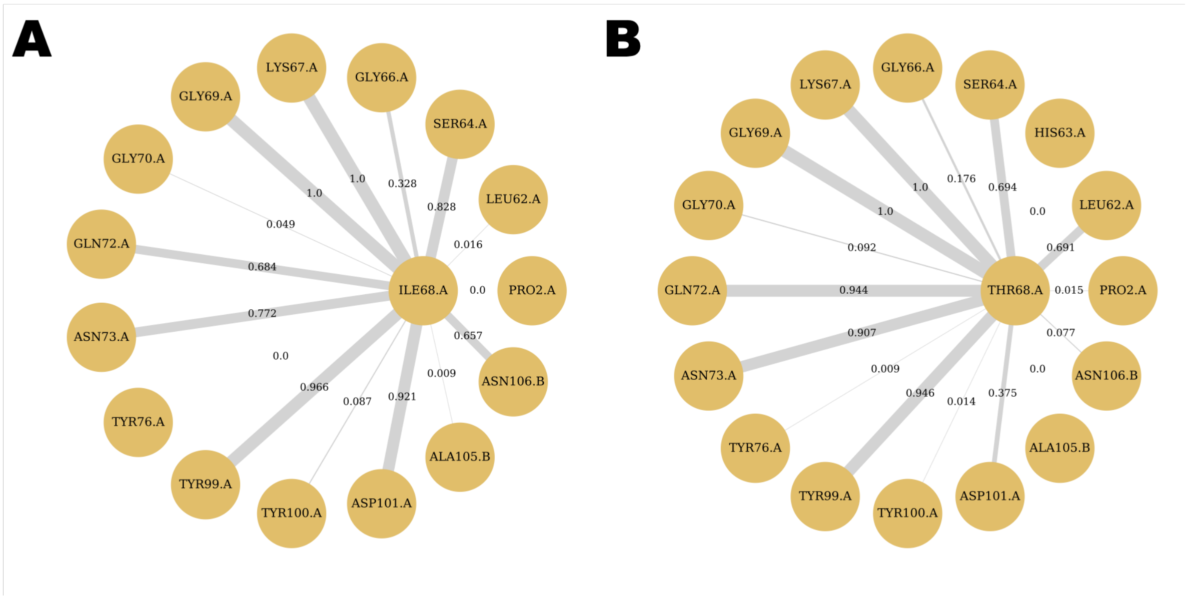


Figure 3.25: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and I68T mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) I68T chain A.

In mutant S7F we observed a reduction in the RRIF between F75.A and A71.A (WT RRIF 0.700; MT RRIF 0.493) (Figure 3.26). Residue A71.A forms vdW interactions with G69.A and the loss of this interaction resulted in increased flexibility of A71.A and G69.A in the mutant relative to the MIF wildtype protein (Figure 3.10).

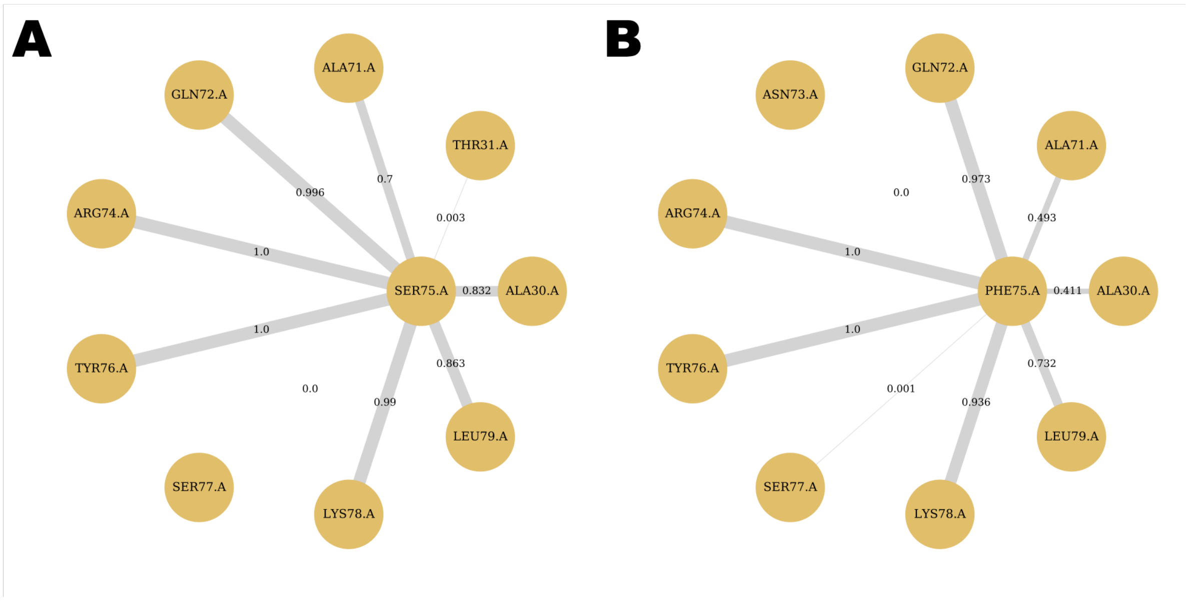


Figure 3.26: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and S75F mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) S75F chain A.

In mutant N106S we observed a loss of interaction between S106.A and K67.C and reduced RRIF with I68.C (WT RRIF 0.436; MT RRIF: 0.365). Residue K67.C shows increased flexibility in the N106S mutant relative to the MIF wildtype protein (Figure 3.10).

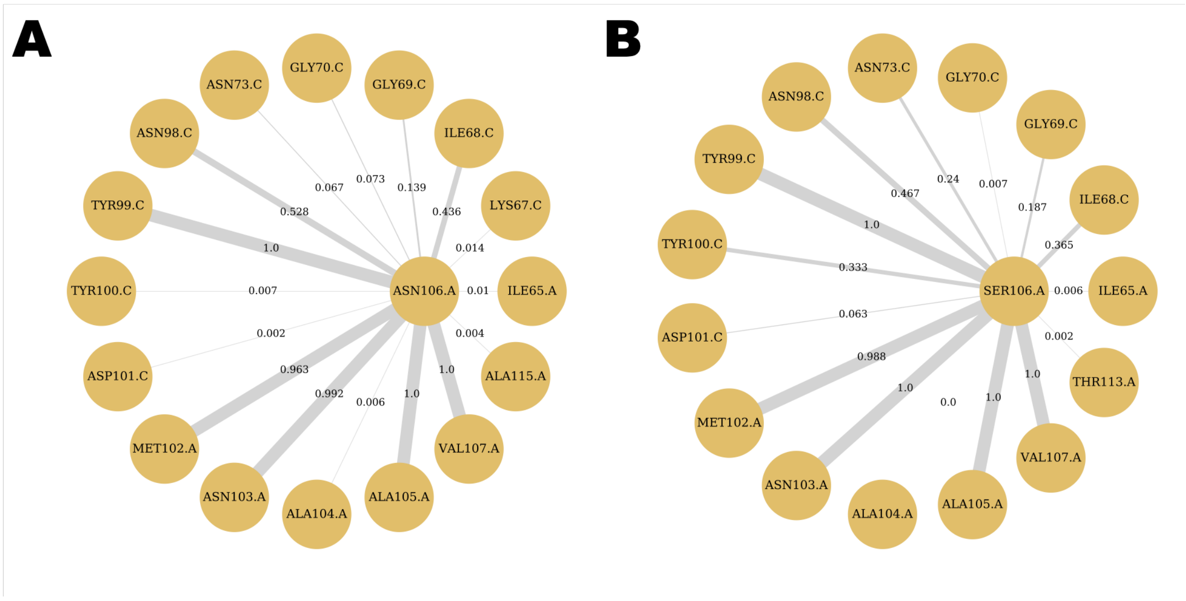


Figure 3.27: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and N106S mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) N106S chain A.

In mutant T113S we observed a reduction in the RRIF between S113.A and G70.A (WT RRIF 0.466; MT RRIF 0.001), N73.A (WT RRIF 1; MT RRIF 0.073), and R74.A (WT RRIF 1; MT RRIF 0.312) (Figure 3.28). In T113S mutant, residues K67.A - S75.A and K67.C - S75.C showed increased flexibility when compared to the MIF wildtype protein (Figure 3.10).

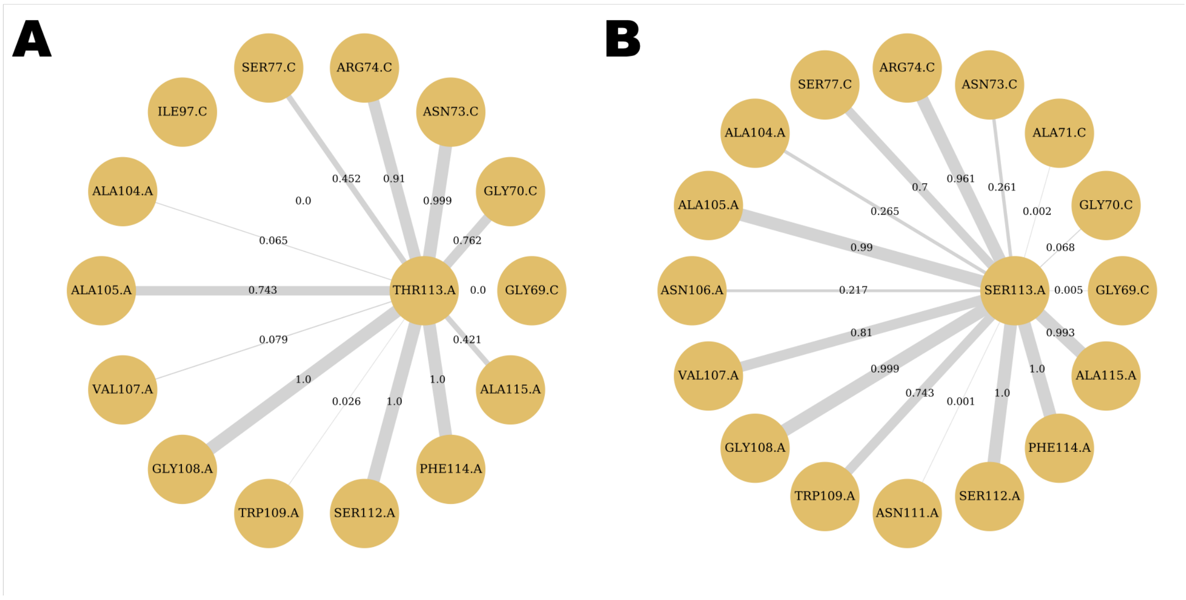


Figure 3.28: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and T113S mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) T113S chain A.

Further, the mutations resulted in alterations of protein concerted motions as shown by changes in the DCC when compared to the MIF wildtype. The T31I mutant had the following residue pairs whose motions became anti-correlated: N98.A & V107.B, M1.B & F114.B, K67C & N98.A, and K67C & F114.B (Supplementary Table 6). The Y37H mutant had the motions of the following residue pairs become anti-correlated: VI5.A & S75.C, N9.A & M48.B, N9.A & A49.B, N9.A & P50.B while the following residue pairs had their motions become correlated: N9.A & A71.A, N9.A & G72.A, N9.A & N73.A, N9.A & R74.A, N9.A & S75.A (Supplementary Table 6). The G52C mutant had the motions of the following residue pairs become correlated: A30C & K33A, T31.C & S64.A, G32.C & S64A (Supplementary Table 6). The I68T mutant had the motions of the following residue pairs become anti-correlated: D17.A

& A115.A; G18.B & V15.C while the following residue pairs had their motions become correlated: S14.B & D17.A; P16.B & D17.A (Supplementary Table 6). The S75F mutant had the motions of the following residue pairs become anti-correlated: P16.A & V6.B, P16.A &V10.B, D17.A & V10.B, G18.A & V10.B while the following residue pairs had their motions become correlated V6.A & G66.A,V6.A & I65.A (Supplementary Table 6). The N106S mutant had the motions of the following residue pairs become anti-correlated: Q46.B & K67.C, L47.B & K67.C while the following residue pairs had their motions become correlated: K67.A & G66.C, K67.A & K67.C (Supplementary Table 6). Lastly, the T113S mutant had the motions of the following residue pairs become anti-correlated G69.A & K78.C, G70.A & S75.C, G70.A & K78.C while the following residue pairs had their motions become correlated L59.A & G66.A, G69.A & L59.A (Supplementary Table 6).

From the DRN analysis it can be seen that these residues P2, T31, Y37, G52, I65, I68, S75, N106, and T113 are part of a similar local residue interaction network. From our results, mutation in any one of these positions causes alterations in residue-residue interactions adversely affecting the protein's dynamics. We further propose that T31, G52, I68, S75 and T113 are interesting candidates for mutagenic studies to validate to investigate their role in MIF/CD74 interaction.

4.3.5.5 Mutant I5M, P16Q, L23F, T24S, H41P, P44L, M48V, and S54R

I5M, P16Q, L23F, T24S, H41P, P44L, M48V, and S54R mutants appear to be important for MIF physiological function. Interestingly I5M, L23F, T24S, and S54R mutations were predicted to be neutral (Figure 3.2). P16Q, H41P, P44L, and M48V mutations were predicted to be detrimental (Figure 3.1). Residues M48 and S54 are both located within the MIF N-like loop [188] while P16 is located within the MIF 3' exonuclease and endonuclease activity site [202]. Residues I5, L23, and H41 are located within the homotrimer hydrophobic buried core while T24 and P44 are located on the protein outer surface (Table 3.2). H41 and P44 are part of a highly conserved hydrophobic pocket whose disruption has been shown to affect MIF oligomerization [210]. Furthermore, I5, H41, and M48 are also located at the chain-chain interface (Figure 3.2).

Conformation, flexibility and PCA analysis

Compared to the MIF wildtype, I5M, P16Q, L23F, T24S, H41P, P44L M48V and S54R mutant protein MD systems showed pronounced conformational changes (Figure 3.8-B). Further, when compared to the MIF wildtype I5M, P16Q, L23F, T24S, H41P, P44L, and S54R showed also a reduction in the flexibility of the N-like loop in chain C (Figure 3.29). The I5M mutant showed increased flexibility of residues A71.A – N73.A, G69.B – A71.B, and A115.C (Figure 3.29). The L23F mutant showed increased flexibility of residues K67.B and G66.B (Figure 3.29). The H41P mutant showed increased flexibility of residues K67.B – G72.B and A104.C (Figure 3.29) while P44L showed increased flexibility of residues G52.B – S53.B (Figure 3.29). The M48V mutant showed increased flexibility of residues K67.B, I90.B – P92.B, L47.C, V48.C, S64.C and I65.C (Figure 3.29). Mutation of L23F, H41P, and M48V appears to cause changes in the flexibility of MIF/CD74 interaction residues I65 and K67 (Figure 3.29) [190].

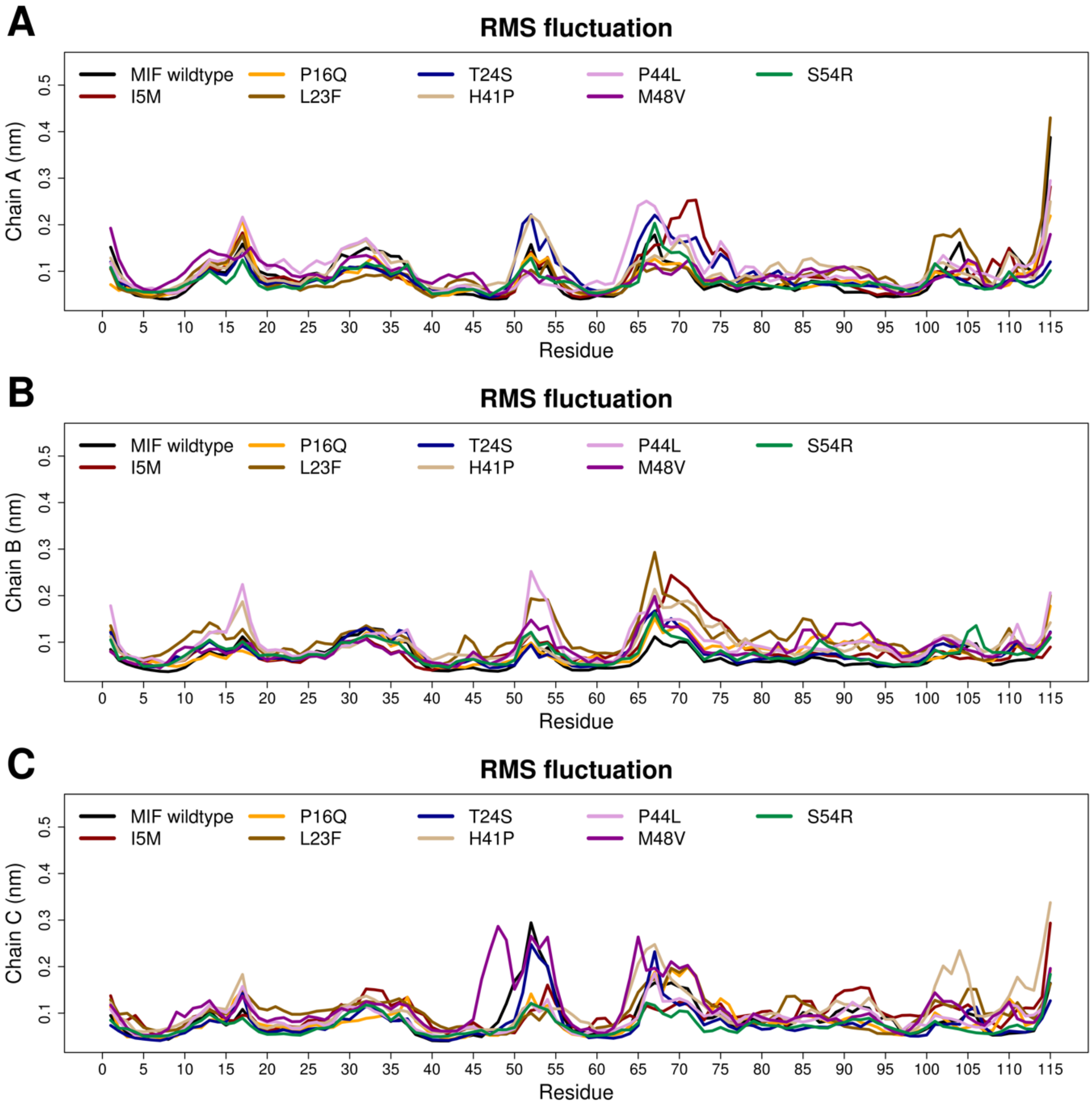


Figure 3.29: RMSF of backbone atoms MIF wildtype (black), I5M, P16Q, L23F, T24S, H41P, P44L, M48V, and S54R mutant proteins as a function of time. A) Chain A RMS fluctuations, B) Chain B RMS fluctuation, and D) Chain C RMS fluctuations.

Dynamic residue network analysis

In comparison to the MIF wildtype I5M, P16Q, L23F, T24S, H41P, P44L, M48V and S54R mutants showed a reduction in the *Average L* of the chain C N-like loop, most notably in residue S54.C and E55.C (Figure 3.30-C). H41P and P44L mutants also showed increased *Average L* of residues S53.A and S53.B respectively (Figure 3.30).

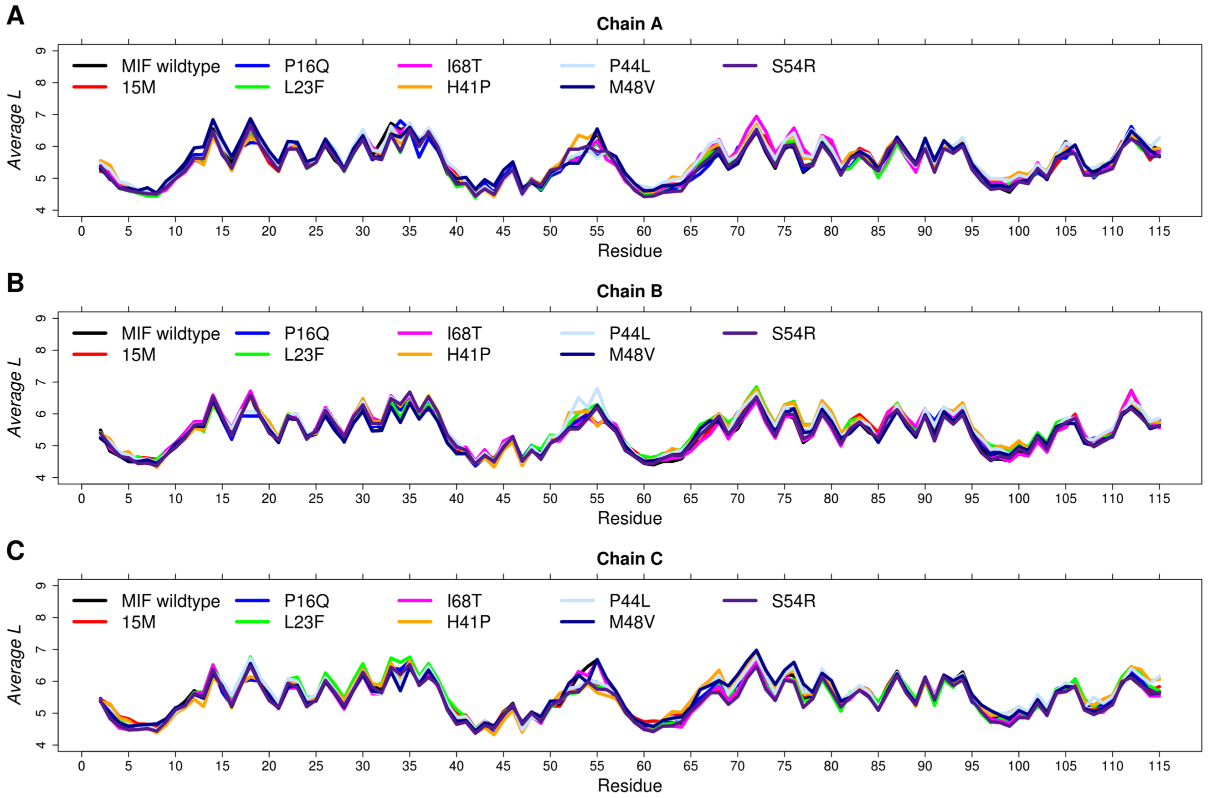


Figure 3.30: Changes in *Average L* of the MIF wildtype, I5M, P16Q, L23F, T24S, H41P, P44L, M48V and S54R mutant proteins during MD production. A) Chain A, B) Chain B, and C) Chain C.

The mutations resulted in changes in communication flow within the protein network with the centrality of residue Y99.A reduced in L23F, H41P, P44L, M48V, and S54R mutants (Figure 3.31). Residue H63.B showed a reduction in centrality in I5M, P16Q, H41P, P44L, M48V, and S54R mutants (Figure 3.31). In the 3DJH crystal structure residue H63.B is centrally located and forms vdW interactions with both P2.B and a hydrogen bond with Y100.A.

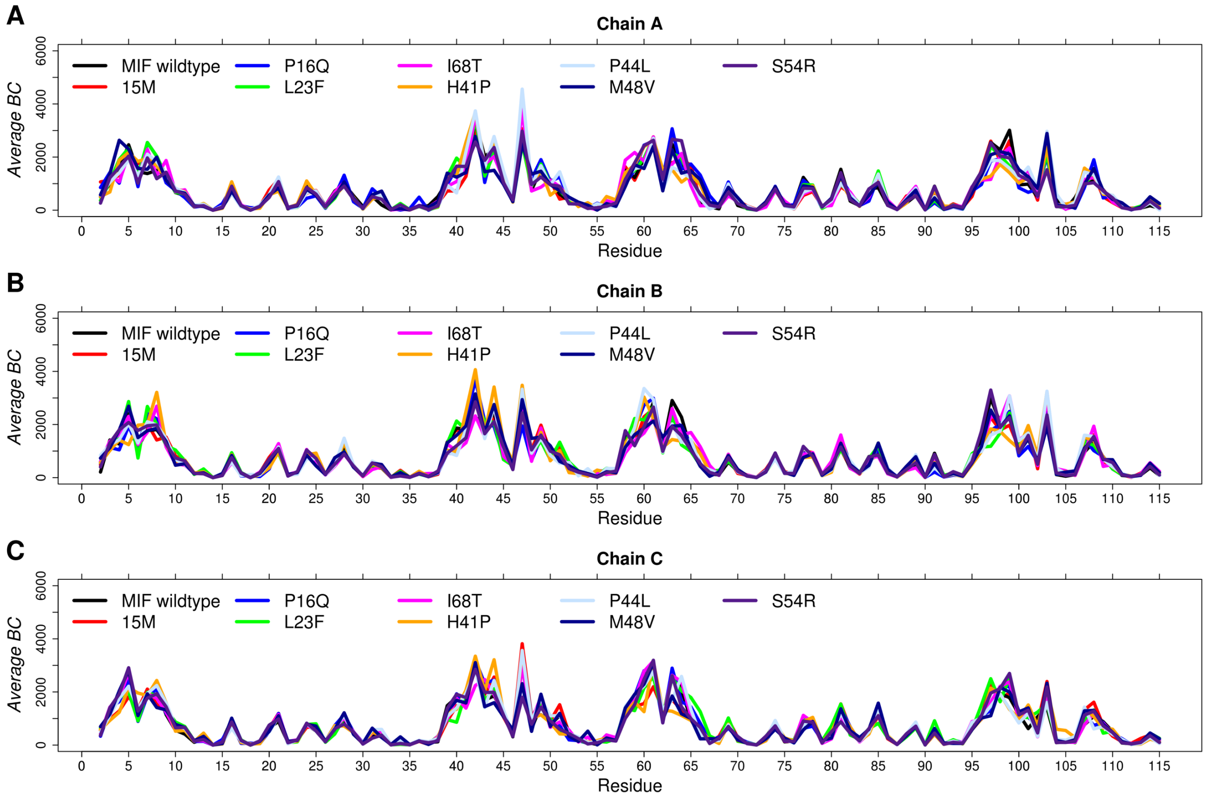


Figure 3.31: Changes in *Average BC* of the MIF wildtype, I5M, P16Q, L23F, T24S, H41P, P44L, M48V and S54R mutant proteins during MD production. A) Chain A, B) Chain B, and C) Chain C.

The I5M mutant had residues, S64.B, I97.B, C60.C and H63.C becoming less central (Figure 3.31-C). In the I5M mutant, residue M5.A has an increased RRIF with C60.A (MT 0.394; WT 0.745) and a gained interaction with H63.A (Figure 3.32). This is an indication of how mutation causes a change in the residue-residue interaction network and thus affects communication flow as measured by centrality.

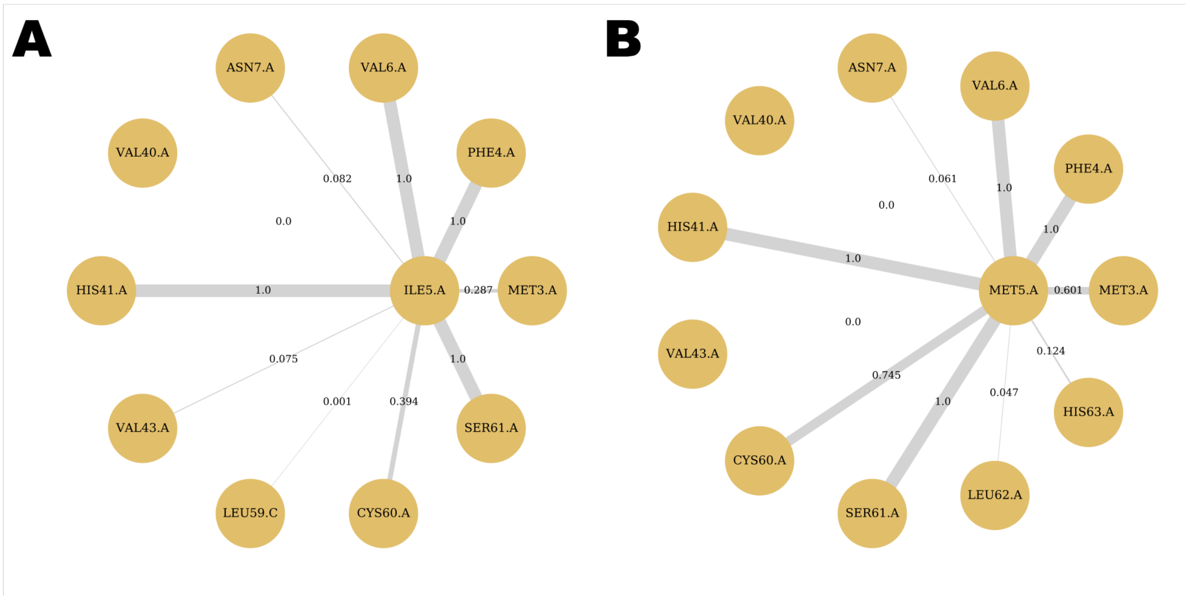


Figure 3.32: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and I5M mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) I5M chain A.

The P16Q mutant had residues L47.B, and I97.B becoming less central while V40.B showed increased centrality (Figure 3.21). Further residue Q16.A shows a reduced RRIF with R87.A (MT 0.820; WT 0.022) and gained interactions with L47.C (Figure 3.33).

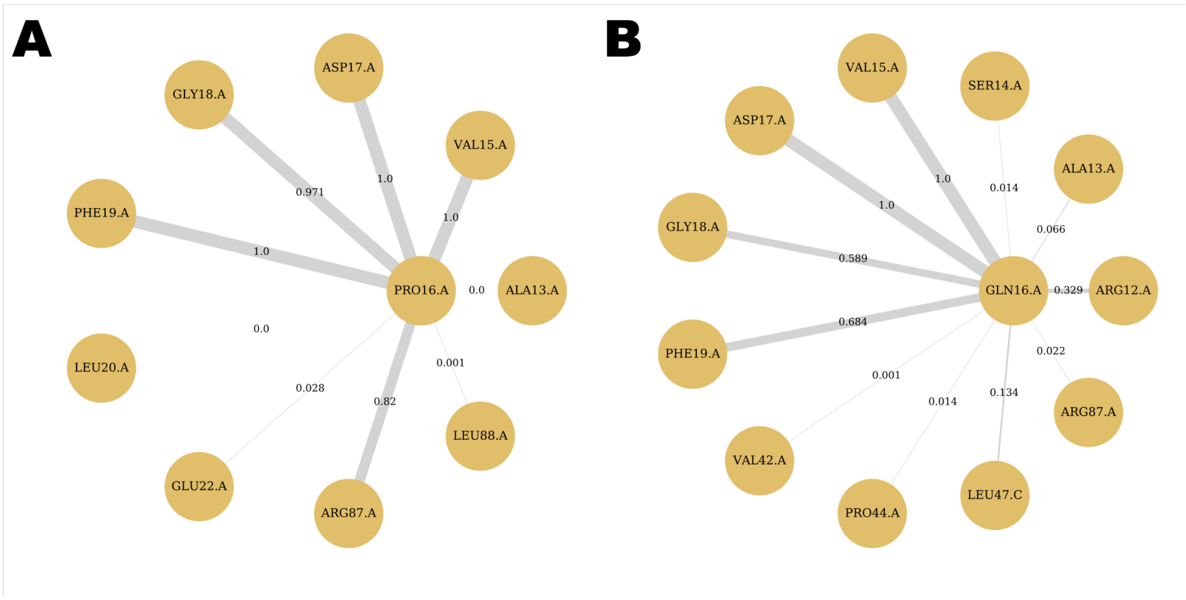


Figure 3.33: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and P16Q mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) P16Q chain A.

The L23F mutant had residues S64.B and I97.B showing reduced centrality (Figure 3.31). Residue F23.B showed a reduction in the interaction with L83.B (WT 0.743; MT 0.222) and an increased RRIF with F4.B (WT 0.112; MT 0.382) (Figure 3.34).

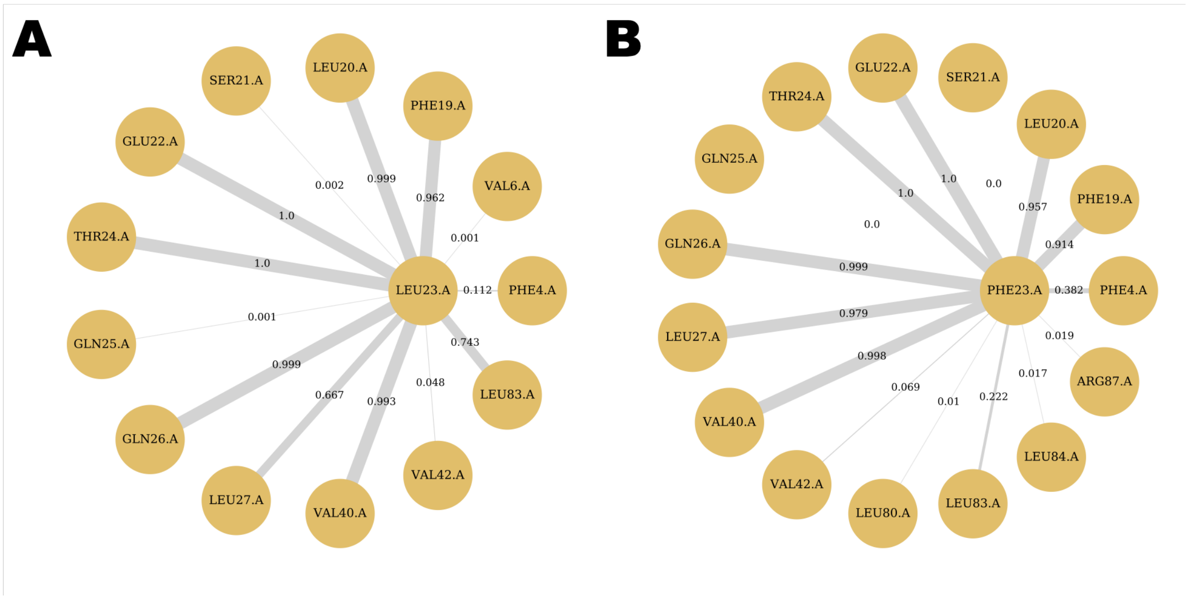


Figure 3.34: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and L23F mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) L23F chain A.

The T24S mutant had residues I97.A, V40.B, I97.B and V42.C showing reduced centrality (Figure 3.31). There was a reduced RRIF between S24.B and A49.A (WT 0.300; MT 0.079) that is located within the N-like loop (Figure 3.35). Further, the motions of I65.B&K67.C became anti-correlated (Supplementary Table 6).

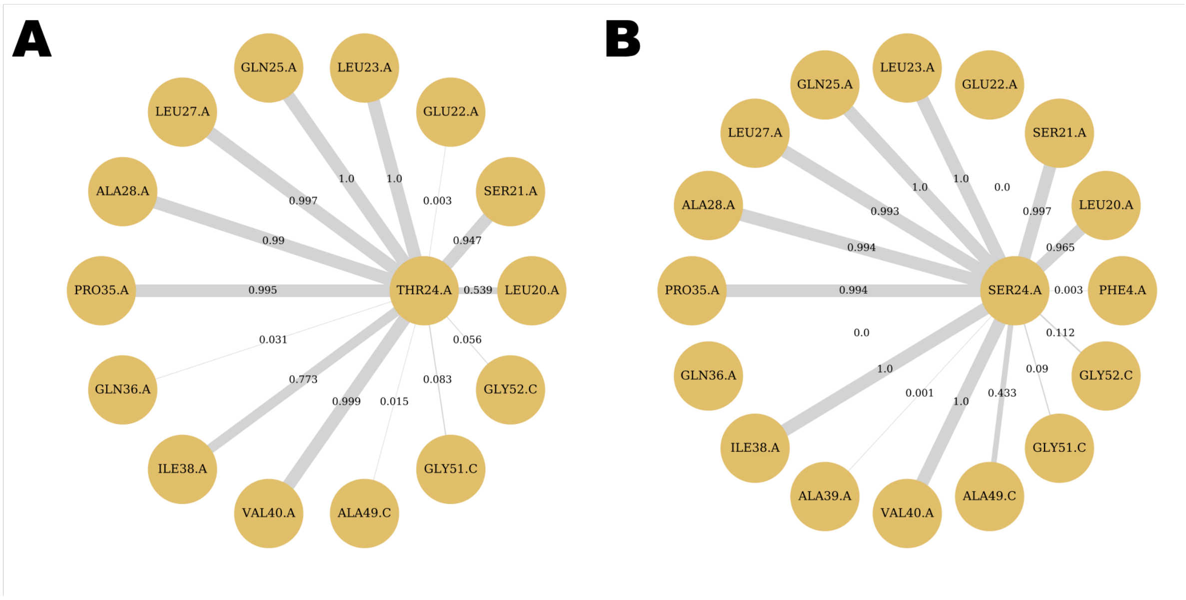


Figure 3.35: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and T24S mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) T24S chain A.

Mutations within the hydrophobic pocket positions H41P and P44L showed similar changes in centrality (Figure 3.31). H41P had residues I97.A, I97.B, Y99.B, and H63.C showing reduced centrality (Figure 3.31). In the H41P mutant, P41.A had increased RRIF with N7.B (WT 0.451; MT 0.936), L47.C (WT 0.038; MT 0.131), L59.C (WT 0.256; MT 0.809), a gained contact with C57.A, and a lost contact with F4.A (WT 0.641; MT 0.000). Both H41 and P41 can be seen to be within 6.7 Å of I5 and M48 (Figure 3.36).

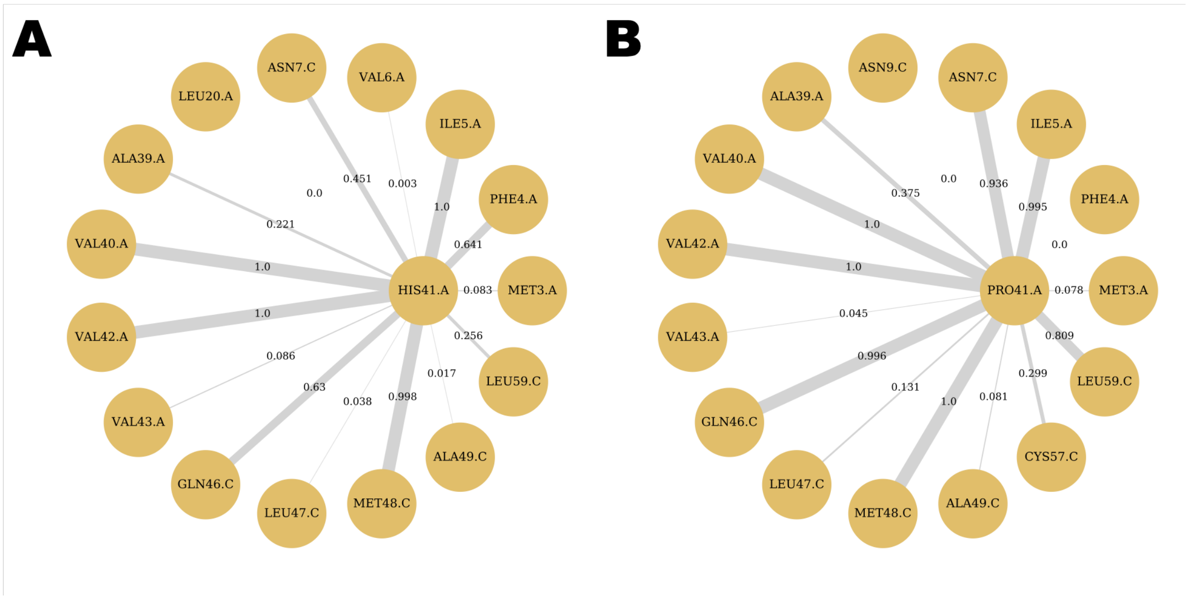


Figure 3.36: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and H41P mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) H41P chain A.

In the P44L mutant, residues I97.A, and I97.B showed increased centrality. Residue L44.A had reduced RRIF with residues V15.A (WT 0.998; MT 0.507), V42.A (WT 0.812; MT 0.679) and increased RRIF with L47.C (WT 0.062; MT 0.124) (Figure 3.37).

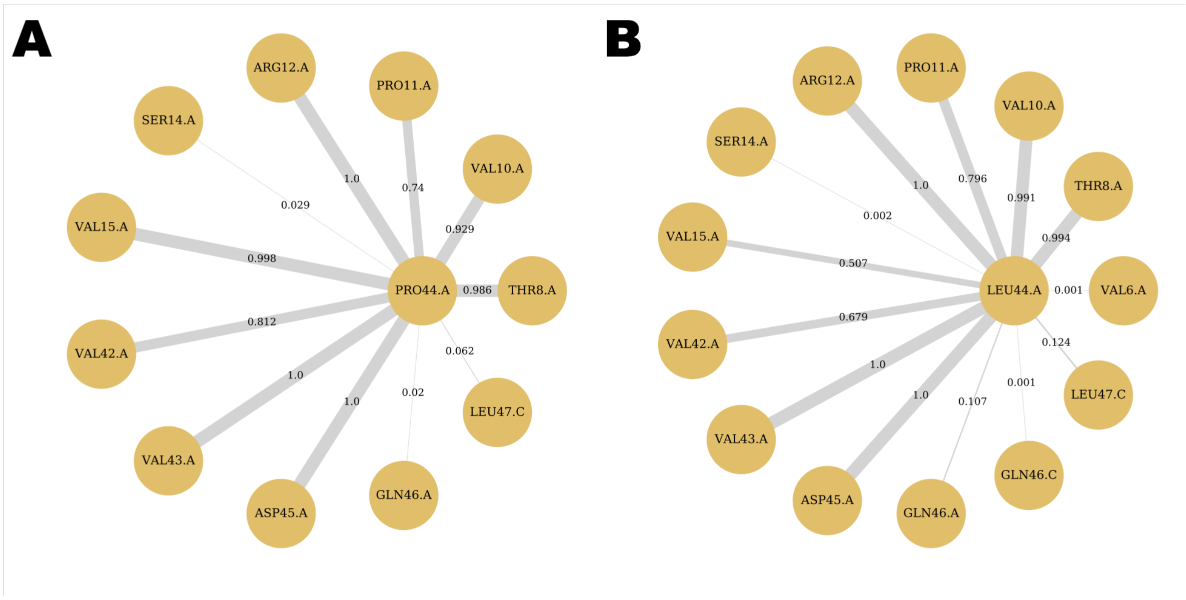


Figure 3.37: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and P44L mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) P44L chain A.

In the M48V mutant, residues H63.C and I97.B showed reduced centrality while residue N103.C showed increased centrality (Figure 3.31). Residue V48.A showed an increased RRIF with residues T8.A (WT 0.010; MT 0.831), N-like loop residues C57.A (WT 0.554; MT 0.916) and L59.A (WT 0.517;MT 0.997) and an decreased RRIF with residue V40.B (WT 0.745; MT 0.047) (Figure 3.38).

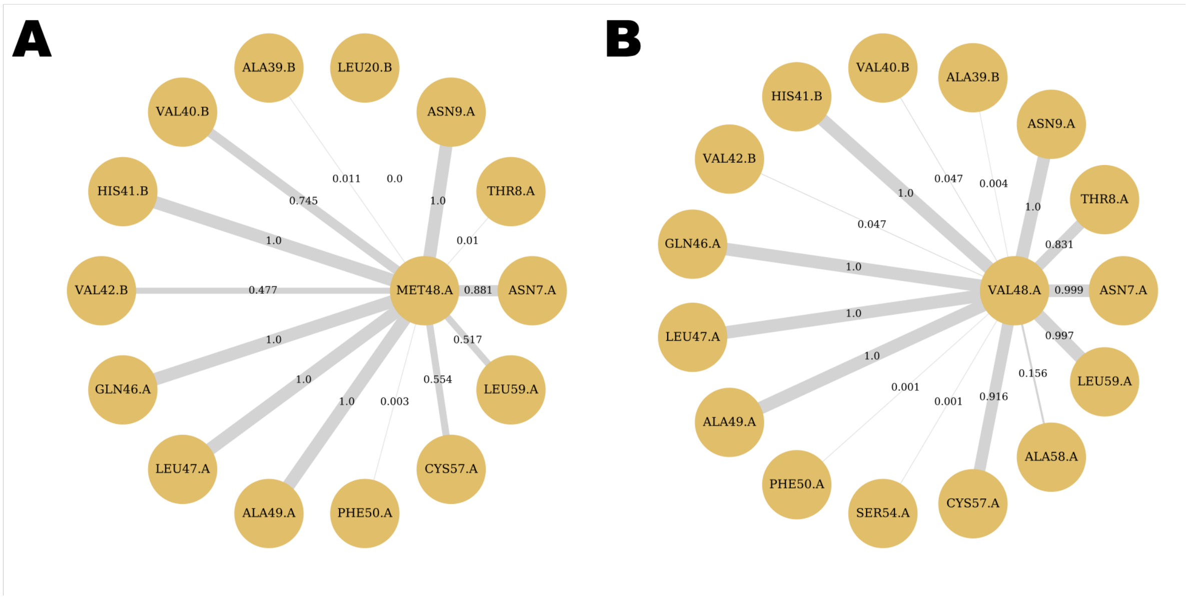


Figure 3.38: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and M48V mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) M48V chain A.

In the S54R mutant, residues Y99.C and N103.C showed increased centrality (Figure 3.3.1). Residue R54.A showed increased RRIF with residues N-like loop residue A49.A (WT 0.177; MT 0.989) (Figure 3.3.9).

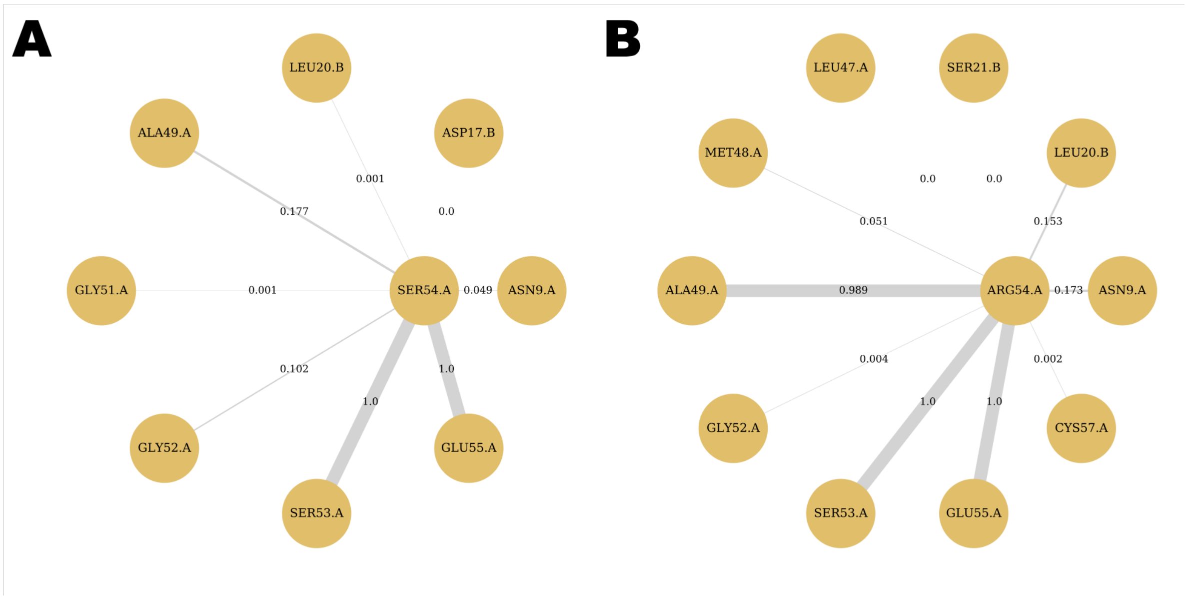


Figure 3.39: Weighted residue interaction maps generated across the trajectory for the chain A of the MIF wildtype and S54R mutant generated using MD-TASK. A) 3DJH chain A wildtype and B) S54R chain A.

When compared to the MIF wildtype the motions of T31.A &S54.A became correlated in the I5M mutant (Supplementary Table 6). In the L23F mutant, Y37.A & T31.C, and Y37.A & G32.C had motions that became anti-correlated (Supplementary Table 6). In the H41P mutant, the motions of C57.A & K67.B became correlated (Supplementary Table 6) while in the M48V mutant the motions of N110.A & I65.C became anti-correlated (Supplementary Table 6).Our findings agree with previous wet-lab analyses that showed that residues such as Y37 and I65, which are experimentally validated MIF/CD74 interaction residues, are important for protein stability [190].

Out of the the 27 mutations analyzed, mutations P2Q, P16Q, Y37H, H41P, M48V, P44L, G52C, I65M, I68T, S75F, N106S and T113S were predicted to be detrimental while mutations I5M, L23F, T24S, T31I, and S54R were predicted to be neutral (Figure 3.2). In contrast, MD simulations showed mutations P2Q, I5M, P16Q, L23F, T24S, T31I, Y37H, H41P, M48V, P44L, G52C, S54R, I65M, I68T, S75F, N106S, and T113S were destabilizing. Mutations I5M, L23F, T24S, T31I and S54R that were predicted to be neutral using the consensus prediction based on MutPred, PredictSNP, SIFT, PROVEAN, and PANTHER were destabilizing according to our MD results. There is a need to complement in silico predictions with more robust approaches that take into consideration protein dynamics and energetics [211,212].

Many of the destabilizing mutants were located in or near functionally important regions. For example, mutations P2Q and I65M are located within the Tautomerase active site (Figure 3.1). Mutations I65M and Y37M are located within the MIF and C74 interaction site (Figure .1). Residues Y37, I65, K67, W109, and N110 are experimentally validated MIF/CD74 interaction residues [190]. Further, mutations N106S and T113S are located at the MIF carboxy-terminal whose deletion results in loss of MIF immunological activity (Figure 3.1) [193]. Mutations M48V, G52C, and S54R are located within are located within the N-like loop that is important in MIF and CXCR2 binding [188]. Mutation P16Q is located within MIF's 3' exonuclease and endonuclease activity site while mutations H41P and P44L are located within the highly conserved hydrophobic pocket (Figure 3.1) [210]. Mutations I5M, H41P, and M48V are located at the chain-chain interface that is known to be important for intra-protein communication and oligomerization (Figure 3.1) [213,214].

An investigation of the effects of the mutations on MIF structure, stability and dynamics using all-atom MD simulations at 300 K revealed that mutants P2Q, I5M, P16Q, L23F, T24S, A28V, T31I, P34T, Y37H, H41P, P44A, P44L, M48V, G52C, S54R, A58T, A58V, L59F, I65M, I68T, A71T, S75F, L84R, Y99C, A105T, N106S, and T113S differed in dynamics when compared to the wildtype MIF protein. Changes in conformation were measured using RMSD, while changes in flexibility were measured using RMSF, *Average L,* and *Average BC.* Most of the mutants showed a reduction in the flexibility of the N-like loop that is of importance in MIF and CXCR2 binding [188]. Amino acid substitution as a result of mutation appears to alter residue-residue interactions around mutation the positions which is a key driver for the changes in flexibility and stability observed. The findings in this study reveal how amino acid mutation changes the interactions around the mutation site result in changes in local interactions and thus protein dynamics.

3.4 Conclusion

In this study we investigated the effects of 27 missense SNVs on MIF protein structure and function. We used *in silico* effect prediction tools, structural analyses, modeled the mutants and subjected them to molecular dynamic simulations and network analysis.

From our findings, *in silico* effect prediction using Web-based servers indicated that P2Q, P16Q, A28V, P34T, Y37H, H41P, P44L, P44A, M48V, G52C, A58T, A58V, I65M, I68T, S75F, L84R, Y99C, N106S and T113S were predicted to be deleterious while I5M, L23F, T24S, T31I, S54R, L59F, A71T, and A105T were predicted to be neutral. MD and DRN analysis showed that mutations P2Q, I5M, P16Q, L23F, T24S, T31I, Y37H, H41P, M48V, P44L, G52C, S54R, I65M, I68T, S75F, N106S, and T113S caused pronounced conformational changes. Further, DRN showed that P2, T31, Y37, G52, I65, I68, S75, N106, and T113S are part of a similar local residue interaction network.

From our findings we show that there is need to combine several computational tools/techniques especially more robust approaches that take into account protein dynamics and energetics. Sequence based *in silico* prediction tools alone though helpful require supplementation. Network analysis has been previously used by Pantouris *et al.* to successfully identify residues crucial for MIF biological function; residue Y99 that showed high centrality when mutated to alanine (Y99A) resulted in the failure of MIF to activate CD74 *in vivo* [173]. Altered residue-residue interactions around the mutation positions appears to be a key driver of the changes in protein flexibility and stability. Our results are of use in informing future MIF function research such as; mutagenic studies and accelerating phenotype prediction.

PART II: COMPUTER AIDED DRUG DISCOVERY

CHAPTER 4

Identification of Novel Pteridine Reductase 1 (PTR1) Inhibitors

Chapter Overview

In the previous part of the thesis (Part I) we looked at host genetic determinants. In Part II of the thesis we used bioinformatics tools to identify novel compounds that could provide a starting point for the development of safe and efficacious HAT chemotherapeutics. According to the Drugs for Neglected Diseases Initiative (DND*i*) new potential HAT drugs should offer an improvement to currently available chemotherapeutics. Further, in the HAT target product profile, what would be ideal is a safe, effective, and practical drug that can be used to treat both stages of the disease (further details available at: https://www.dndi.org/diseases-projects/hat/hat-target-product-profile/).

This Chapter includes a computational rational drug discovery of HAT chemotherapeutics targeting the trypanosome folate pathway, in particular PTR1 which is responsible for folate inhibitor resistance in trypanosomatids. Also included in this Chapter is a brief introduction to the concepts and a review of literature concerning computer aided rational drug discovery and the folate pathway in trypanosomes. The Chapter includes our computational screening approach and *in vitro* testing of the potential hit compounds.

The wet lab aspects of this work, which included *in vitro* Trypanosome inhibition, human cell cytotoxicity, and the Isobologram assays were carried out by our collaborators Prof Heinrich Hoppe and Dustin Laming of the Centre for Chemico- and Biomedicinal Research, Rhodes University, Grahamstown 6140, South Africa.

4.1 Introduction

Trypanosomes are folate auxotrophs which means that they are unable to synthesize folates *de novo* [215]. Reduced pterins and folates are essential for the survival of the parasite and are needed in pathways such as nucleic acid and protein synthesis [216]. To acquire required folates and pterins the parasites scavenge extracellular folates and pterins from their mammalian hosts [217,218]. The main enzymes in the pathways are a fused Dihydrofolate Reductase Thymidylate Synthase (DHFR-TS; EC 1.5.1.3 & EC 2.1.1.45) and Pteridine Reductase 1 (PTR1; EC 1.5.1.33) that is present only in trypanosomatids [219–221].

Anti-folates have been used in the treatment of infections such as malaria, shigellosis, listeria, and toxoplasmosis [222–226]. The majority of anti-folates for example: pyrimethamine, proguanil, and trimethoprim rely primarily on inhibiting DHFR [222–227]. Despite the wide use of anti-folates, the pathway remains largely under explored in trypanosomes.

The two main enzymes involved in folate metabolism in trypanosomes are the NADPH dependent enzymes *Tb*DHFR-TS (EC 1.5.1.3 & EC 2.1.1.45) and *Tb*PTR1 (EC 1.5.1.33). *Tb*DHFR-TS catalyzes the reduction of folates to dihydrofolate (H_2F), and dihydrofolate to tetrahydrofolate (H_4F) (Figure 4.1). PTR1 catalyzes the reduction of both folates and pterins (Figure 4.1). This is because of the structural similarity between pterins and folates. Folate is a pteridine that has been conjugated to *p*-aminobenzoic acid (pABA) that is glutamylated (Figure 4.1) [228].

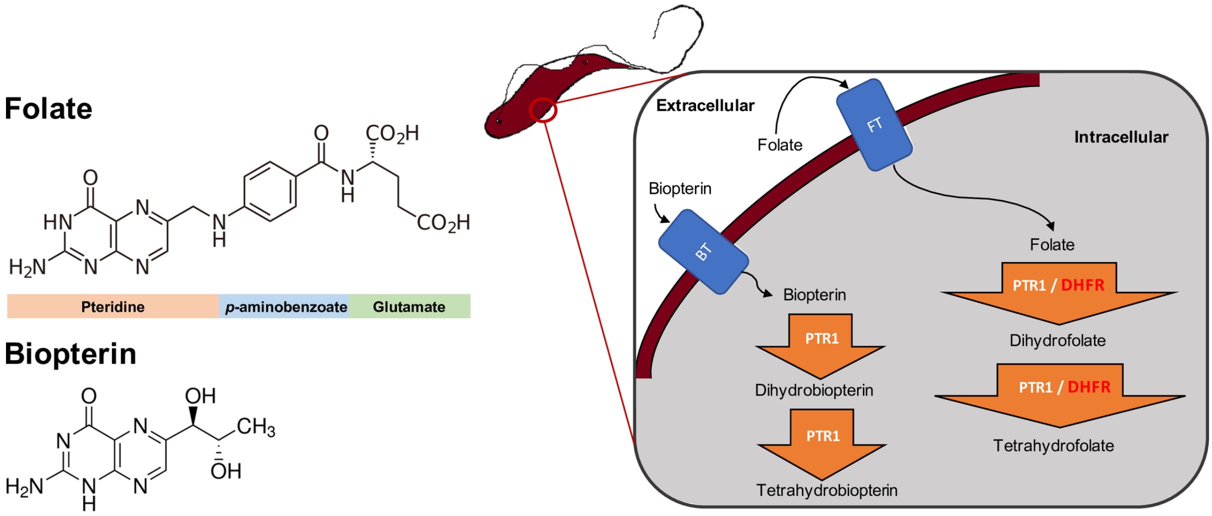


Figure 4.1: The role of *Tb*DHFR and *Tb*PTR1 in trypanosome folate and pterin metabolism. The structures of folate and biopterin are shown to the left. Folates and pterins are taken up by transporters (folate-biopterin transporter superfamily, includes biopterin transporter 1 [BT1] and folate transporter 1 [FT1]) after which they are reduced to their functional cofactors (right).

PTR1 is a short-chain dehydrogenase family member unique to trypanosomatids [217]. It reduces biopterin to dihydrobiopterin, and dihydrobiopterin to tetrahydrobiopterin (Figure 4.1) [217]. Further it also reduces folate to H_2F, and H_2F to H_4F (Figure 4.1) [217]. Under physiological conditions PTR1 only contributes about 10% of trypanosomatid folate metabolism [221]. However, under DHFR inhibition in trypanosomatids such as *L. major* and *T. cruzi*, PTR1 is over-expressed thus promoting anti-folate resistance [217,221,229,230]. Additionally, gene knock down and knock out of *Tb*PTR1 has shown that is essential for *T. brucei* parasite survival [231,232].

As demonstrated in previous studies, successful targeting of the trypanosome folate pathway required dual inhibition of *Tb*DHFR-TS and *Tb*PTR1 [232–236]. This can be achieved by combination of two drugs or possibly a single ligand inhibitor motif that can target both proteins. In this study we sought to use computational tools to discover novel *Tb*PTR1 inhibitors that can be used on their own or in combination with known *Tb*DHFR-TS inhibitors to target the trypanosome folate pathway.

4.1.1 Computer Aided Rational Drug Discovery

Drug discovery is still dependent on wet-lab work, costs a lot, is a long and arduous process and requires technical skills. Computer Aided Drug Design (CADD) is rapidly gaining popularity in pharmaceutical drug discovery because it significantly accelerates the drug discovery process with better predictions of efficacy, toxicity, and contraindications [237–239]. The innovative aspect of rational drug design is that it relies on prior knowledge about the biological target [238]. This knowledge based approach has the advantage of narrowing drug discovery efforts, allowing for specific characterisation of protein-ligand interactions throughout the research and development process.

Computer aided drug design can be divided into two general categories, namely structure based and ligand based approaches [237–239]. In ligand based drug design 3D structural information of the protein target is not required as it relies on knowledge of small molecules (both active and inactive) to generate similarity searches or to make predictions [237–239]. Examples of ligand based approaches include Quantitative Structure Relation (QSAR) modeling. In structure based drug design 3D structural information of the biological target is required because binding affinities have to be calculated, examples include ligand docking and *de novo* design [237–239]. The desired effects from ligand binding include enzyme inhibition blocking signalling cascades, protein oligomerisation disruption, and impaired enzymatic reaction rates to mention but a few.

The initial hit identification is followed by optimisation that involves structure modification/optimisation, testing the compounds for *in vivo* or *in vitro* efficacy and potency. This also includes testing the compounds for selectivity by assessing their cytotoxicity in *in vivo* or *in vitro* models. Successfully screened compounds are then referred to lead compounds which are modified to generate lead series for pre-clinical testing. Lead compounds generally have to go through a process of modification and optimisation. Even at later stages further modifications might have to be made, it is a long and technical process. Eventually compounds that pass pre-clinical testing and clinical phases can eventually be approved for animal or human use.

4.1.2 Concepts in Drug Discovery

4.1.2.1 Sensitivity

Sensitivity refers to the ability of a cell or enzyme to respond to an external factor, which in this case is the ligand inhibitor [240].

4.1.2.2 Selectivity

Selectivity refers to a measure of how specific a ligand inhibitor is when binding to its receptor target instead of other receptors [240]. In the case of targeting parasite proteins, an ideal outcome is the ligand inhibitor preferentially binding to the specific parasite protein target and not the host proteins as this is the primary cause of undesired drug side effects.

4.1.2.3 Half Maximal Inhibitory Concentration IC_{50}

IC_{50} refers to the concentration of a ligand inhibitor where the number of parasites or enzyme activity (response) is reduced by half [241]. A lower IC_{50} concentration indicates higher potency.

4.1.2.4 Synergy, Additive effect, and Antagonism

In some instances, in order to achieve a higher efficacy than from monotherapy, drugs have to be used in combination. Drug combination therapy has several advantages for example staving off the development of drug resistance problems, lowered toxicity, and improved efficacy [242]. This is because there is more than a single drug and as such more than one protein target therefore lowering the likelihood of simultaneous resistance development. In malarial chemotherapy, Artemisinin-based combination therapies (ACTs) are widely used in the treatment of uncomplicated malaria with great success [243,244]. When describing the interaction of drugs used in combination there are three possible outcomes:

a) *Synergy,* where the effect observed as a result of combination is higher than the individual IC_{50}s observed when the drugs are used on their own.

b) *Additive effect*, where the effect observed as a result of combination is equal to the sum of the of the effects of the two drugs.

c) *Antagonism*, refers to the opposite of synergy where the combination of the drugs results in a lowered efficacy of one or both drugs.

Determination of the observed effects from combination therapies in this study were determined using an Isobologram assay in which the concentrations of either drug is varied while dose responses are measured [242,245].

4.1.3 Blood Brain Barrier (BBB) Permeability

As mentioned earlier in section Chapter 1, the second stage of HAT pathology is characterised by the parasites invading the CNS. Therefore, chemotherapeutics used in the treatment of the second stage of HAT have to be able to cross the BBB. The BBB is formed by tight junctions between epithelial cells in CNS capillaries [246]. It prevents large molecules and several small molecule drugs from entering the CNS [246]. There are several chemical properties that have been identified to influence BBB permeability and these include: rotatable bonds, hydrogen donors, molecular weight, lipophilicity as calculated by XlogP, and net charge [246,247]. Rotatable bonds in this case refers to the number of rings and rotatable bonds in the compound. BBB permeable dugs have been shown to have fewer rotatable bonds than non-CNS permeable drugs [248]. Compounds with high hydrogen bonding potential for example peptides with their amide groups poorly cross the BBB [249]. A study of CNS drugs on the market showed that BBB permeable drugs have an average MW of 310 compared to a mean MW of 377 for orally active drugs [248]. Highly lipophilic drugs permeate the BBB better with better retention [250]. BBB permeable drugs tend to be more basic with a positive charge at pH 7-8 [251].

4.1.4 Pan Assay Interferences Compounds (PAINS)

PAINS refers to group of compounds that are associated with high false-positive rates during screening [252]. They are often poor choices for lead optimization and further drug development [253]. This is because compounds that contain PAINS features tend to react non-specifically as such it is difficult to determine if they are interacting with the desired target [254]. There are over 450 compound classes that have been designated as PAINS and these include: hydrazones, toxoflavin, isothiazolones, rhodanines, hydroxyphenyl, curcumin, phenol-sulfonamides, enones, catechols, and quianones [255]. PAINS cause false positives through a variety of means, some include: protein-scaffold mimicry [256], chemical aggregation [257], chelation [258], redox activity [259], membrane disruption [260], singlet oxygen production [261], and cysteine oxidation [262]. In the current study we used a webserver located at http://www.cbligand.org/PAINS/ to identify any PAINS features that might be present in the test compounds [253].

4.1.5 Multiple Sequence Alignment (MSA)

Multiple sequence alignment (MSA) was used to identify conserved amino acid residues among the aligned PTR1 orthologue sequences [263]. This was done in order to highlight the biological relationships between the orthologues and identify residues critical for PTR1 function [263]. MSA computation largely relies on heuristic algorithms as exact calculations are computationally very expensive [263]. Alignments are generally scored either using a protein scoring matrix or a consistency based algorithm. Programs that utilize the protein scoring matrix algorithm include MUSCLE [264], CLUSTALW [265], and Kalign [266]. The protein scoring matrix contains information of substitution rates of amino acids over time and can thus determine how favourable a substitution is [264–266]. Consistency based methods incorporate further information in their evaluation such as substitution matrices that are specific for particular protein types and pairwise or global alignments [267,268]. Examples of consistency based programs are T-COFFEE [267] and ProbCons [269].

4.1.6 Molecular Docking

Molecular docking is a structure based computational prediction of protein-ligand interactions [270–272]. This method has been successfully used in the discovery of several lifesaving anti-retroviral drugs such as: Amprenavir, Nelfinavir, and Raltegravir [273–275]. The molecular docking process can be broken down into three essential steps:

a) *Sampling*, which involves the generation of several ligand poses, orientations and configurations within the protein binding site(s). A key factor in this process is the number of rotatable bonds present in the ligand as they result in increased torsion flexibility and thus a larger sampling space. Protein flexibility is often reduced or excluded in many docking algorithms as this would substantially increase the sampling space and thus computational expense.

b) *Scoring*, which involves the calculation of the estimated Gibbs free energy of interaction $\Delta G°$ for the different ligand posed complexes. The calculation of the binding affinity is dependent on the scoring function used.

c) *Ranking*, which involves ordering the binding free energies in order to determine the pose with the best binding free energy.

4.1.6.1 Autodock Vina

Autodock Vina 1.1.2 is a molecular docking tool for fast and accurate virtual screening [276,277]. It is open source, utilizes an iterative local search algorithm, and is capable of parallelization by using multithreading on multi-core machines [277]. The local search in the algorithm is carried out using a Broyden-Fletcher-Goldfarb-Shanno (BFGS) method which is a quasi-Newton method [277]. The local search involves several mutations and local optimisations with each step being accepted according to the Metropolis criterion [277]. The arguments of the scoring function are: the ligand position and orientation, the values of the torsions of the ligand rotatable bonds, and flexible ligand residues [277]. The number of evaluation steps is determined depending on the apparent complexity of the problem starting from random initial arguments [277]. The general form of the Autodock Vina scoring function is shown in the equation 4.1 below:

$$c = \sum_{i<j} f_{t_i, t_j}\left(r_{ij}\right)$$

Equation 4.1: The general form of the Autodock Vina scoring function

Where the summation is for all the pairs of atoms that can move relative to each other (excluding 1-4 interations). For an atom i, t_i is its assigned atom type and the symmetric set of interaction functions are defined by $f_{ti,tj}$ and interatomic distance (r_{ij}) [277]. The interaction functions $f_{ti,tj}$ are defined relative to the surface distance that is defined by $d_{ij} = r_{ij} - R_{ti} - R_{tj}$, where R_t refers to the van derWaals radius of atom type t and is shown below:

$$f_{t_i, t_j}\left(r_{ij}\right) = h_{t_i, t_j}\left(d_{ij}\right)$$

Equation 4.2: The Autodock Vina interaction functions $f_{ti,tj}$

In the Autodock scoring function, $h_{ti,tj}$ defines the weighted sum of the energy terms shown below:

$$gauss_1(d) = e^{-(d/0.5\text{Å})^2}$$

$$gauss_2(d) = e^{-((d-3\text{Å})/2\text{Å})^2}$$

$$repulsion(d) = \begin{cases} d^2, & if\ d < 0 \\ 0, & if\ d \geq 0 \end{cases}$$

$$hydrophobic(d) = \begin{cases} 1, & if\ d < 0.5Å \\ linear\ interpolation & \\ 0, & if\ d > 0.5Å \end{cases}$$

$$H - bond(d) = \begin{cases} 1, & if\ d < 0.7Å \\ linear\ interpolation & \\ 0, & if\ d > 0Å \end{cases}$$

Equation 4.3: The Autodock Vina energy terms used in the scoring function

The first three terms are the steric interactions followed by hydrophobic interactions and hydrogen bonds where applicable [277]. All interaction functions $f_{ti,tj}$ are cut-off at $r_{ij} = 8Å$. The conformation-independent function *g* that is used for the prediction of binding free energy of intermolecular interactions C_{inter} is given by:

$$g(C_{inter}) = \frac{C_{inter}}{1 + wN_{rot}}$$

Equation 4.4: Autodock Vina binding free energy

Where N_{rot} refers to the number of rotatable bonds between heavy atoms in the ligand while *w* refers to the associated weight [277].

4.1.7 Molecular Dynamics (MD) Simulations and Molecular Mechanics/ Poisson–Boltzmann Surface Area (MM/PBSA)

Molecular docking does not factor in protein dynamics. Proteins are highly flexible and dynamic, and we postulated that the binding of the ligand might have an impact on protein dynamics. We constructed an all-atom MD simulation in explicit water that was used to create representations of all the atoms in the MD system as it evolved (conformational sampling) as governed by the laws of classical mechanics over a given time period at 300 K.

4.1.7.1 Molecular Mechanics/Poisson-Boltzman Surface Area (MM-PBSA) Free Energy Calculations

The MD simulations were also used in MM/PBSA calculations to provide approximations of the binding free energy in the protein-ligand complexes [278–280]. The general terms used in the g_mmpbsa binding free energy calculation are shown in equation 5.5 below:

$$\Delta G_{\mathrm{binding}} = G_{\mathrm{complex}} - (G_{\mathrm{protein}} + G_{\mathrm{ligand}}) \quad (1)$$

$$G_{x} = \langle E_{\mathrm{MM}} \rangle - TS + \langle G_{\mathrm{solvation}} \rangle \quad (2)$$

$$E_{\mathrm{MM}} = E_{\mathrm{bonded}} + E_{\mathrm{nonbonded}} = E_{\mathrm{bonded}} + (E_{\mathrm{electrostatic}} + E_{\mathrm{vdW}}) \quad (3)$$

$$G_{\mathrm{solvation}} = G_{\mathrm{polar}} + G_{\mathrm{nonpolar}} \quad (4)$$

Equation 4.5: The general terms for the calculation of the binding free energy

(1) The binding free energy of the protein-ligand complex in solvent ($\Delta G_{\mathrm{binding}}$) where G_{complex} refers to the total energy of the protein-ligand complex. G_{protein} refers to the isolated free energy of the protein while G_{ligand} refers to the isolated free energy of the ligand [280].

(2) The free energy of either the ligand, protein or protein ligand complex (x). The average mechanical potential in a vacuum is described as $\langle E_{\mathrm{MM}} \rangle$. While TS refers to the entropic contribution (T is temperature and S is entropy) and $G_{\mathrm{solvation}}$ refers to the free energy of solvation [280].

(3) The vacuum molecular mechanics potential energy E_{MM} where E_{bonded} describes the bonded interactions such as bonds, dihedrals, angles and improper interactions. The non-bonded interactions ($E_{\mathrm{nonbonded}}$) are modelled using the Coulomb and Lennard-jones (LJ) potential function. They include: van der Waals (E_{vdW}) interactions and electrostatic interactions ($E_{\mathrm{electrostatic}}$) [280].

(4) The energy required to transfer the protein-ligand solute from a vacuum into a solvent is described as the free energy of solvation ($G_{\mathrm{solvation}}$). G_{polar} and G_{nonpolar} describe the electrostatic and non-electrostatic energy contributions respectively [280].

Lastly by carrying out a per residue energy decomposition one can identify which residues are important in protein-ligand interactions [280].

4.1.8 Research Motivation

The broad theme of this thesis is the use of bioinformatics to investigate different aspects of HAT, to advance knowledge of the parasite-host interactions, of the parasite's biology and to inform or guide efforts to eradicate the disease. Part I focused on host genetic determinants while Part II is focused on parasite intervention in the form of discovery of novel chemotherapeutics. This is because there is still a need to discover and develop safe, easy to administer and effective HAT chemotherapeutics. The drugs currently used for HAT chemotherapy are few, old, difficult to administer, toxic, and growingly inefficient. The work presented in this chapter was directed towards the identification of novel *Tb*PTR1 inhibitors that can be used on their own or in combination with known *Tb*DHFR inhibitors to inhibit trypanosome parasite growth *in vitro* using computational based tools and approaches.

4.2 Materials and Methods

4.2.1 Multiple Sequence Alignment

MSA was performed using MUSCLE on the PTR1 orthologue sequences [264]. The PTR1 orthologue sequences used were *Tb*PTR1 (Uniprot: O76290), *T. cruzi* (Uniprot: O44029*)*, *L. major* (Uniprot: Q01782), and *H. sapiens* Dehydrogenase/reductase SDR family member 4 (DHRS4) (Uniprot: Q9BTZ2).

4.2.2 Ligand Library Preparation

The small-molecule ligands were obtained from the South African Natural Compounds database (SANCDB) [281] and the Drugs Now subset from the ZINC database (ZINC12) [282]. The Drugs now subset was already filtered for 'drug-likeness' using Lipinski's rule of five [282,283]. The initial ligand library was composed of 10 639 555 compounds from the

Drugs Now subset and 635 compounds from SANCDB. To filter for BBB permeability, the Drugs Now compounds were reduced to 5107 based on: compounds with XlogP ≤ 3, fewer than four rotatable bonds, at least 2 hydrogen donors, a net charge of zero and a molecular weight ≤ 490. The final ligand library was composed of 5742 compounds of which 635 were from SANCDB and 5107 from the ZINC subset.

4.2.3 Preparation of Protein-Ligand Complexes

The PTR1 and *Tb*DHFR crystal structures were retrieved from from RCSB Protein Data Bank [176]. Missing residues in the *Tb*PTR1 crystal structure (PDB: 2X9N) were resolved using homology modelling (residues Q104 – G113 and K143 – S151) using in-house Python scripts. Further, homology modelling was used to calculate a structure for *Tc*PTR1 using its isoform *Tc*PTR2 (Uniprot: Q8I814, PDB: 1MXH). Modelling of the protein structures that included their NADPH cofactors was done using MODELLER (version 9.19) using the 'automodel' class [178,179]. Out of the 100 models generated per modelling job, the top models were selected based on their z-DOPE score and validation using ProSA [151]. The structures used in the study are shown in Table 4.1 below:

Table 4.1: PTR1 orthologue and *Tb*DHFR structures

Enzyme	Organism	UNIPROT	PDB	Resolution (Å)	Chains	Residues	RMS (Å) (ref: 2X9N)
*Tb*PTR1	*T. b. brucei*	O76290	2X9N	1.15	A,B,C,D	288	0
*Tc*PTR1	*T. cruzi*	O44029	-	Homology model (1MXH: 2.2 Å)	A,B,C,D	276	0.5
*Lm*PTR1	*L. major*	Q01782	1E92	2.2	A,B,C,D	288	0.5
DHSR4	*H. sapiens*	Q9BTZ2	3O4R	1.7	A,B,C,D	261	1.6
*Tb*DHFR	*T. b. brucei*	Q27783	3QFX	2.2	A, B	241	13.5

4.2.3.1 Molecular Docking

The NADPH cofactor is essential for organising the PTR1 and DHFR active sites, as such it was included in the structure when carrying out blind docking [284,285]. Molecular docking was carried using Autodock Vina (version 1.1.2) on the tetrameric *Tb*PTR1, *Tc*PTR1 homology model, *Lm*PTR1, DHSR4 and dimeric *Tb*DHFR. The Autodock Vina parameters for each of the proteins is shown in Table 4.2 below:

Table 4.2: Molecular Docking Parameters for Autodock Vina

Parameter	***Tb*PTR1**	***Tc*PTR1**	***Lm*PTR1**	**DHSR4**	***Tb*DHFR**
Protein center X	30.859	15.330	11.600	29.790	-19.789
Protein center Y	-0.064	32.050	45.880	2.200	23.569
Protein center Z	92.956	-0.021	69.630	17.480	8.146
Box size X	126	126	126	126	126
Box size Y	126	126	126	126	126
Box size Z	126	126	126	126	126
Energy range	4	4	4	4	4
Exhaustiveness	120	120	120	120	120
CPU	24	24	24	24	24

Protein-ligand complexes were evaluated based on binding modes, parasite selectivity, binding free energy Vina score, and hydrogen bonding patterns.

4.2.4 Prediction of Blood-brain Barrier Permeability

We carried out a PCA based on the chemical descriptors of the top hit compounds, ZINC Food and Drug Administration (FDA) approved drugs, and FDA approved CNS permeable drugs [282]. The chemical descriptors used included: topological polar surface area (tPSA), XlogP, number of H-bond acceptors (HBA), number of H-bond donors (HBD), net charge (NC), molecular weight (MWT), number of rotatable bonds (NRB), polar and apolar desolvation. The first and second principal components were plotted against each other to give a representation of chemical space of the compounds.

4.2.5 Molecular Dynamics (MD) Simulations

MD simulations of the apo protein and protein-ligand complexes were carried out using the GROMACS (5.1.4) simulation engine [157]. The ligand and NADPH cofactors were parametrized using AMBER03 force field utilizing ACPYPE [286]. The MD systems were solvated using a Simple Point Charge (SPC) water model in a cubic box of 5.07 × 5.18 × 5.16 (nm) with a 2.37 × 2.86 × 3.30 (nm) center. A distance minimum of 1.5 nm was allowed between the system and the wall. The system was neutralized using sodium and chloride ions. Minimization was done using the steepest descent algorithm with a 100 kJ/mol/nm tolerance value. This was followed by temperature and pressure equilibration using the NPT (constant number of particles, pressure, and temperature) and NVT (constant number of particles, volume, and temperature) ensembles. Each of the systems were simulated for 200 ns at 300 K without the application of any restraints with energy evaluations carried out every 2 fs and saved every 10 ps.

4.2.5.1 MD Analysis

Analysis of the MD trajectories included conformational evaluations that included: RMSD, Rg, and RMSF that were carried out using the GROMACS toolbox. Visualization tools used in this study included PyMOL [180], Discovery Studio [287], and GRACE software (http://plasma-gate.weizmann.ac.il/Grace/).

4.2.6 MM-PBSA Free Energy Calculations

Equilibrated trajectories from the last 50 ns of the protein-ligand simulations were used in binding free energy calculations using the g_mmpbsa package (version 1.6) [280].

4.2.7 Average shortest path (Average L), and Average Betweenness Centrality (Average BC)

The same equilibrated trajectories (including the apo protein) were used to carry out DRN analysis using MD-TASK [132]. The DRN of the protein systems was generated using a 6.7 Å cut-off. The DRN analyses included: *Average L*, *Average BC*, and Residue contact maps that were all carried out using the default MD-TASK parameters.

4.2.8 *In vitro* inhibition Assays

4.2.8.1 Trypanosoma *in vitro* Inhibition Assay

Compounds with promising binding modes and dynamics during the computational screening were later commercially purchased and assayed for anti-trypanosomal activity and human cytotoxicity *in vitro*. The assays described in section 6.2.8 and 6.2.9 were carried out by our collaborators Prof Heinrich Hoppe and Dustin Laming.

The compounds were assayed for trypanocidal activity by adding 20 μM of each compound to cultures of *T. b. brucei* (strain Lister 427) in 96-well plates (37 °C and 5% CO_2 in IMDM medium). After 48 h incubation, the parasite percentage viability was determined by counting the number of parasites that survived using the resazurin method [288] on a Spectramax M3 microplate reader. Trypanocidal activity of the compounds was reported as the percentage of viable parasites in the compound treated wells when compared to untreated controls (% viability). Pentamidine an FDA approved trypanocidal drug was used as the control drug standard [18]. For compounds that showed < 20% viability their IC_{50} were subsequently determined. To assess any synergistic effects, the compounds were assayed for trypanocidal activity when used in combination with WR99210 a known *Tb*DHFR inhibitor [289].

4.2.8.2 *In vitro* Human Cytotoxicity Assay

The compounds assayed for trypanocidal activity were also tested to determine if they caused adverse effects against human cells *in vitro*. For this assay HeLa (human cervix adenocarcinoma) cells were used (37 °C and 5% CO_2 in DMEM supplemented with 10 % fetal calf serum). The compounds were assayed for cytotoxic activity by adding 20 μM of each compound to 96-well plates containing HeLa cells followed by incubation for 48 h. Cell viability was determined using the resazurin method [288]. Emetine a drug that which induces cell apoptosis was used as a control [290].

4.2.9 Pan-Assay Interference Compounds (PAINS) Assay

PAINS screening was carried out using a webserver available at http://www.cbligand.org/PAINS/.

4.3 Results and Discussion

4.3.1 Overview of PTR1 Structure and Conservation

*Tb*PTR1 has a homo-tetrameric structure with each chain containing an NADPH cofactor. Each monomeric unit is made up of a α/β- domain subunit that is constructed around an NADPH binding Rossmann-fold repeat that is composed of seven parallel β-sheets that are between three α-helices on either side as shown in Figure 4.2 [230]. In Figure 4.2, the substrate binding loop is colored red and is composed of S207-E215, while the SDR family signature which is colored brown is composed of D161-A193.

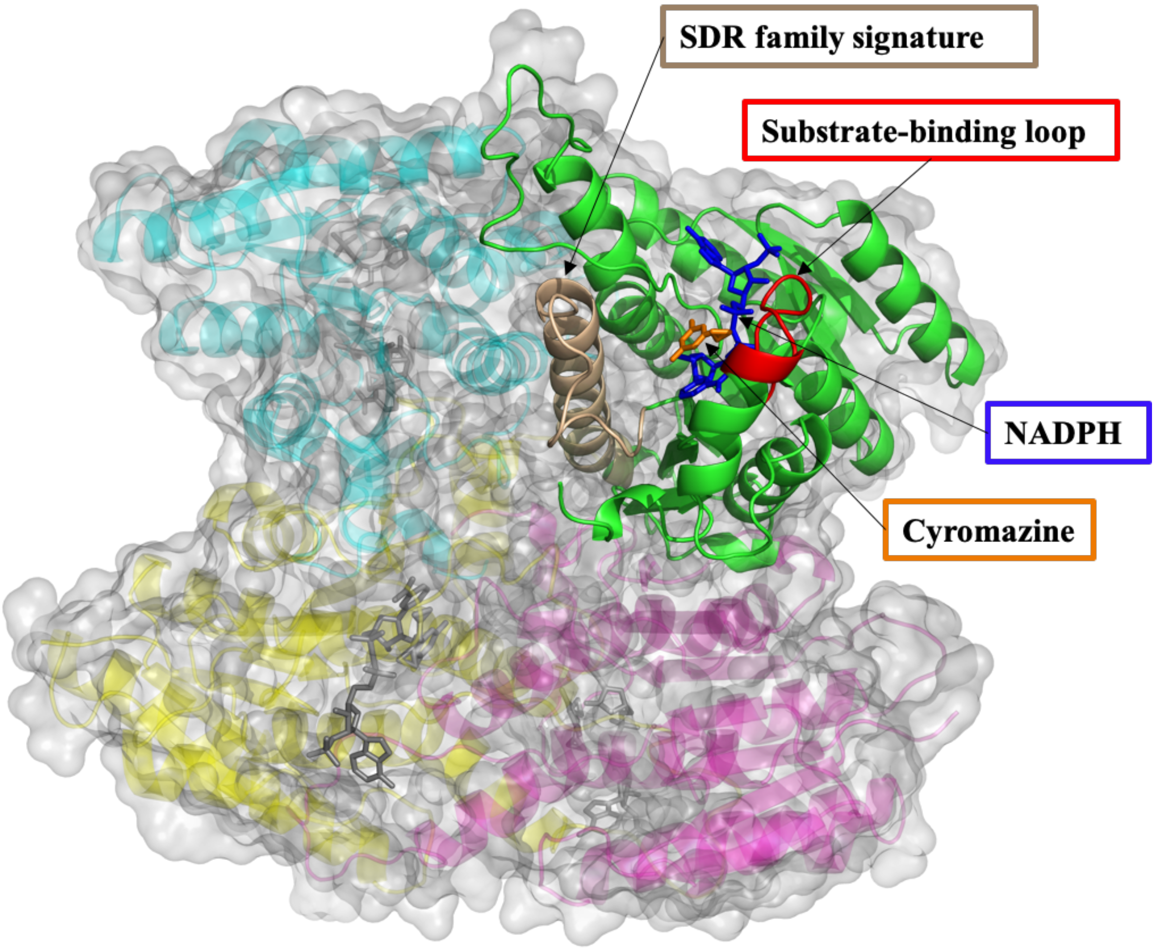
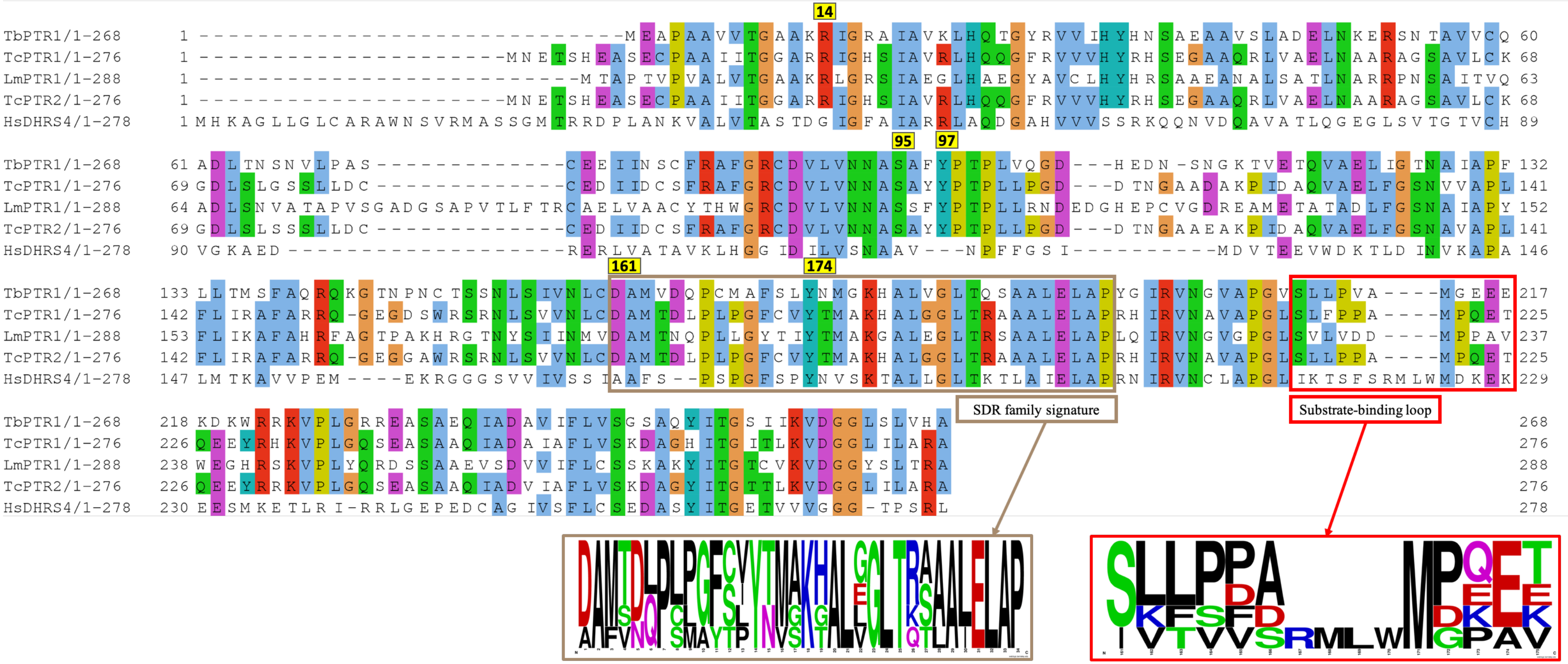


Figure 4.2: A cartoon representation of the *Tb*PTR1 protein structure (PDB: 2X9N). The protein is colored by chain, the NADPH cofactor blue, and cyromazine inhibitor colored orange.

MSA analysis of *Tb*PTR1, *Tc*PTR1, *Lm*PTR1 and *Hs*DHRS4 orthologue protein sequences revealed several highly conserved and important residues (Figure 4.3). The SDR family signature and substrate binding loop were highly conserved among the trypanosomatids. Pterins and folates are structurally similar (Figure 4.1) and as a result interact with the PTR1 active site quite similarly, as is the case with known PTR1 inhibitors. A common binding motif for substrates and inhibitors alike is a π-sandwich formed between the NADPH nicotinamide ring and residue F97 [233,285]. MSA showed that R14, S95, F97, D161, and Y174 are highly conserved among the trypanosomatids (Figure 4.3). These residues are known to be involved in PTR1 substrate and inhibitor binding [233,285]. The PTR1 orthologues are quite similar, more so the trypanosomatids (RMS 0.4 – 0.5 Å) as shown by whole structure superimposition in Table 4.1. The validation of the calculated *Tc*PTR1 and *Tb*PTR1 fixed residues homology models are shown in Supplementary Figure 1.

4.3.2 Eighteen Potential Hits out of 5742 Compounds are Identified via Virtual Screening

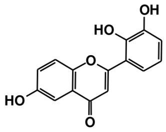
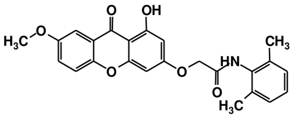
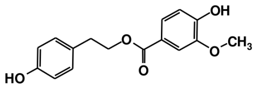
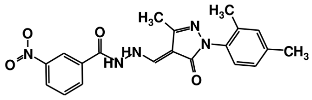
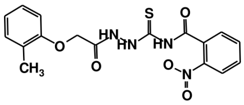
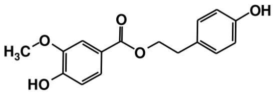
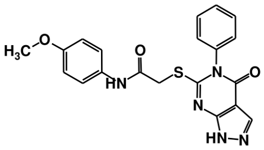
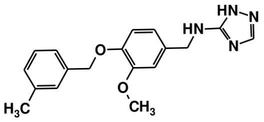
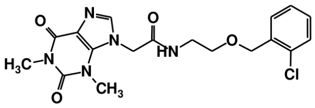
Out of the 5742 compounds that were screened using molecular docking, we identified 18 compounds as potential hits using Autodock Vina as shown in Table 4.3. Each of the top compounds had an Autodock Vina binding free energy > -33.472 kJ/mol and showed good selectivity for the parasite PTR1 with the exception of RUBi006 (Tale 6.3). RUBi006 however bound *Hs*DHSR4 (-32.217 kJ/mol) with a weaker binding energy than to *Tb*PTR1 (-42.677 kJ/mol) (Table 4.3). The protein ligand systems were further assessed with PyMOL [180] and Discovery Studio [287] to characterize their protein-ligand interactions particularly interaction with residues such as R14, S95, F97, D161 and Y174 that were shown to be of catalytic importance and highly conserved (Figure 4.3).

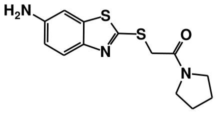
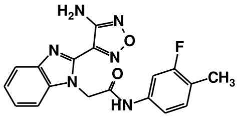
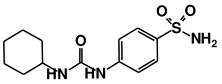
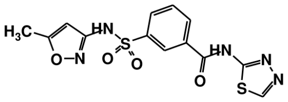
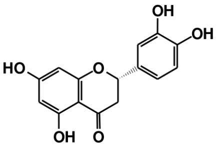
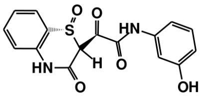
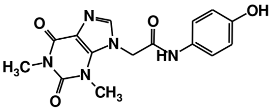
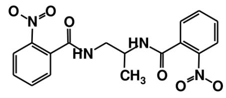
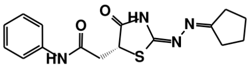


```
TbPTR1/1-268     1 -----------------------------MEAPAAVVTGAAKRIGRAIAVKLHQTGYRVVIHYHNSAEAAVSLADELNKERSNTAVVCQ 60
TcPTR1/1-276     1 ---------------------MNETSHEASECPAAIITGGARRIGHSIAVRLHQQGFRVVVHYRHSEGAAQRLVAELNAARAGSAVLCK 68
LmPTR1/1-288     1 --------------------------MTAPTVPVALVTGAAKRLGRSIAEGLHAEGYAVCLHYHRSAAEANALSATLNARRPNSAITVQ 63
TcPTR2/1-276     1 ---------------------MNETSHEASECPAAIITGGARRIGHSIAVRLHQQGFRVVVHYRHSEGAAQRLVAELNAARAGSAVLCK 68
HsDHRS4/1-278    1 MHKAGLLGLCARAWNSVRMASSGMTRRDPLANKVALVTASTDGIGFAIARRLAQDGAHVVVSSRKQQNVDQAVATLQGEGLSVTGTVCH 89

TbPTR1/1-268    61 ADLTNSNVLPAS-------------CEEIINSCFRAFGRCDVLVNNASAFYPTPLVQGD---HEDN-SNGKTVETQVAELIGTNAIAPF 132
TcPTR1/1-276    69 GDLSLGSSLLDC-------------CEDIIDCSFRAFGRCDVLVNNASAYYPTPLLPGD---DTNGAADAKPIDAQVAELFGSNVVAPL 141
LmPTR1/1-288    64 ADLSNVATAPVSGADGSAPVTLFTRCAELVAACYTHWGRCDVLVNNASSFYPTPLLRNDEDGHEPCVGDREAMETATADLFGSNAIAPY 152
TcPTR2/1-276    69 GDLSLSSSLLDC-------------CEDIIDCSFRAFGRCDVLVNNASAYYPTPLLPGD---DTNGAAEAKPIDAQVAELFGSNAVAPL 141
HsDHRS4/1-278   90 VGKAED-------------------RERLVATAVKLHGGIDILVSNAAV---NPFFGSI----------MDVTEEVWDKTLDINVKAPA 146

TbPTR1/1-268   133 LLTMSFAQRQKGTNPNCTSSNLSIVNLCDAMVDQPCMAFSLYNMGKHALVGLTQSAALELAPYGIRVNGVAPGVSLLPVA----MGEEE 217
TcPTR1/1-276   142 FLIRAFARRQ-GEGDSWRSRNLSVVNLCDAMTDLPLPGFCVYTMAKHALGGLTRAAALELAPRHIRVNAVAPGLSLFPPA----MPQET 225
LmPTR1/1-288   153 FLIKAFAHRFAGTPAKHRGTNYSIINMVDAMTNQPLLGYTIYTMAKGALEGLTRSAALELAPLQIRVNGVGPGLSVLVDD----MPPAV 237
TcPTR2/1-276   142 FLIRAFARRQ-GEGGAWRSRNLSVVNLCDAMTDLPLPGFCVYTMAKHALGGLTRAAALELAPRHIRVNAVAPGLSLLPPA----MPQET 225
HsDHRS4/1-278  147 LMTKAVVPEM----EKRGGGSVVIVSSIAAFS--PSPGFSPYNVSKTALLGLTKTLAIELAPRNIRVNCLAPGLIKTSFSRMLWMDKEK 229

TbPTR1/1-268   218 KDKWRRKVPLGRREASAEQIADAVIFLVSGSAQYITGSIIKVDGGLSLVHA 268
TcPTR1/1-276   226 QEEYRHKVPLGQSEASAAQIADAIAFLVSKDAGHITGITLKVDGGLILARA 276
LmPTR1/1-288   238 WEGHRSKVPLYQRDSSAAEVSDVVIFLCSSKAKYITGTCVKVDGGYSLTRA 288
TcPTR2/1-276   226 QEEYRRKVPLGQSEASAAQIADVIAFLVSKDAGYITGTTLKVDGGLILARA 276
HsDHRS4/1-278  230 EESMKETLRI-RRLGEPEDCAGIVSFLCSEDASYITGETVVVGGG-TPSRL 278
```

Figure 4.3: A multiple sequence alignment of *Tb*PTR1, *Tc*PTR1, *Tc*PTR2, *Lm*PTR1 and *Hs*DHSR4 and the sequence logos of the conserved SDR family signature and substrate binding loop.

Table 4.3. The chemical structures of the top *Tb*PTR1 docking compounds and Autodock Vina molecular docking results.

Compound Information			Docking Binding Energy (kJ/mol)			
Code	**Chemical structure**	**Database ID**	***Tb*PTR1**	***Tc*PTR1**	***Lm*PTR1**	***Hs*DHSR4**
RUBi001		ZINC00057846	-42.258	-40.166	-38.493	-
RUBi002		ZINC08992677	-42.677	-42.258	-41.003	-
RUBi003		SANC00368	-38.074	-35.982	-33.890	-
RUBi004		ZINC00809143	-43.095	-42.258	-38.074	-
RUBi005		ZINC02690799	-37.656	-36.819	-35.982	-
RUBi006		SANC00470	-42.677	-39.748	-35.982	-32.217[1]
RUBi007		ZINC00630525	-40.166	-36.819	-38.074	-
RUBi008		ZINC06556964	-35.564	-37.238	-35.982	-
RUBi009		ZINC02177983	-37.238	-34.727	-34.309	-

RUBi010		ZINC00359797	-28.870	-31.799	-32.635	-
RUBi011		ZINC00677623	-40.585	-41.003	-40.585	-
RUBi012		ZINC01003765	-38.074	-33.890	-33.054	-
RUBi013		ZINC02184332	-36.401	-39.330	-33.054	-
RUBi014		ZINC0058117 / SANC00320	-40.585	-38.074	-40.585	-
RUBi015		ZINC04671320	-38.074	-38.074	-34.727	-
RUBi016		ZINC00612219	-37.238	-37.238	-32.217	-
RUBi017		ZINC04523829	-36.819	-38.074	-35.146	-
RUBi018		ZINC04313814	-35.146	-35.146	-36.819	-

4.3.3 BBB Permeability Prediction and 'Drug-likeness'

The 18 hit compounds occupied the same chemical space as FDA approved CNS permeable drugs (n = 221) and FDA approved drugs (n = 3180) as shown by a PCA based on the molecular descriptors (Figure 4.4). PC1 explained 76% of the observed variance and PC2 explained 18%

of the observed variance (Figure 4.4). This showed that the hit compounds were ‘drug-like’, possibly BBB permeable, and good starting scaffolds for HAT chemotherapeutic discovery [291].

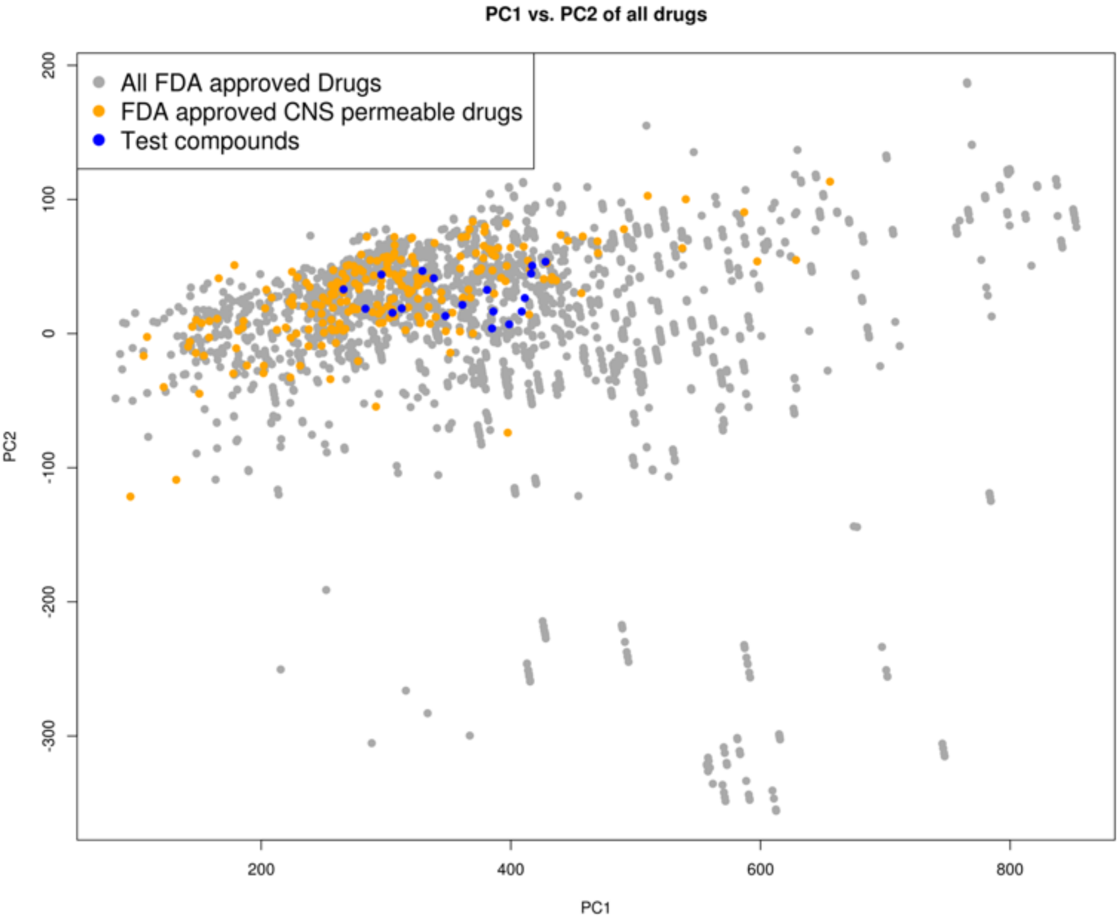


Figure 4.4: A PCA representation of the chemical space of the test compounds (n= 18; colored blue), FDA approved CNS permeable drugs (n = 221; colored yellow), and FDA approved drugs (n = 3180; colored grey) based on their molecular descriptors. The PCA scatter plot was based on PC1 and PC2 that explained 94% of the observed variance in the data.

This was followed by the 18 *Tb*PTR1-ligand complexes being subjected to 200ns production run MD simulations. The MD trajectories were analysed in respect to RMSD, Rg, RMSF, and MM-PBSA binding free energy calculations.

4.3.4 Five Hit Compounds Show Anti-trypanosomal Activity *in vitro*

From the *in silico* analyses the 18 compounds showed stable protein-ligand complexes. Thirteen compounds were commercially available and were tested for anti-trypanosomal activity *in vitro*. Compounds RUBi003, RUBi006, RUBi009, RUBi013, and RUBi017 were not commercially available for *in vitro* testing. Compounds RUBi001, RUBi002, RUBi005, RUBi008, RUBi010, RUBi011, and RUBi012 in spite of good binding modes and stable MD trajectories did not show anti-trypanosomal activity when tested on their own or in combination with WR99210 which is a known *Tb*DHFR inhibitor. Compounds RUBi004, RUBi007, RUBi014, RUBi016 and RUBi018 showed *in vitro* anti-trypanosomal activity in the micromolar range and are further discussed in the subsequent sections (Figure 4.5).

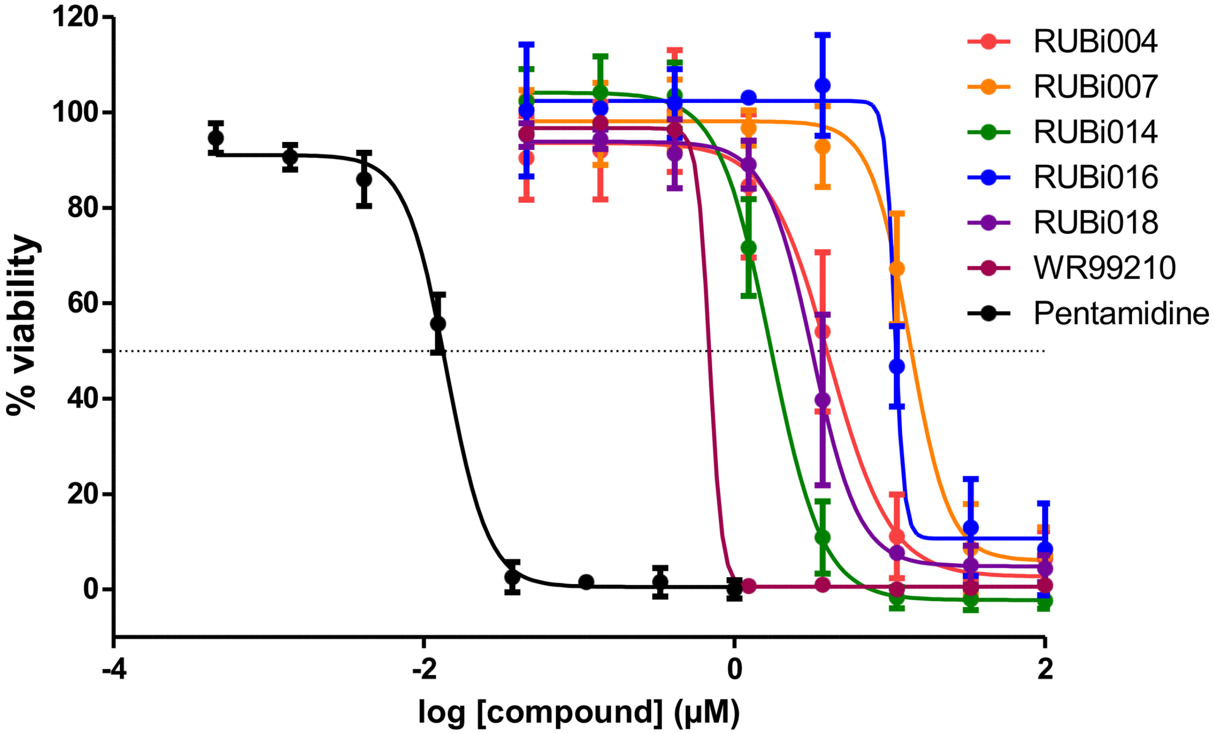


Figure 4.5: IC_{50} determination of compound RUBi004 (red), RUBi014 (orange), RUBi007 (green), RUBi016 (blue), RUBi018 (purple), WR99210 (maroon), and Pentamidine (black). Pentamidine was used as the positive control. The IC_{50} of RUBi004, RUBi007, RUBi014, RUBi016, and RUBi018 were determined to be 12.5 ± 4.8 µM, 32.4 ± 4.2 µM, 5.9 ± 1.4 µM, 28.2 ± 3.3 µM, and 9.7 ± 2.1 µM respectively. The IC50 of WR99210 and pentamidine were determined to be 1 ± 0.1 µM and 0.014 µM respectively.

4.3.5 *Tb*PTR1 Hit Compounds have either Antagonist or Additive Activity when Used in Combination with a *Tb*DHFR inhibitor

RUBi004, RUBi007, RUBi014, RUBi016 and RUBi018 showed anti-trypanosomal activities against trypanosomes in culture with IC_{50} values of 12.5 ± 4.8 µM, 32.4 ± 4.2 µM, 5.9 ± 1.4 µM, 28.2 ± 3.3 µM, and 9.7 ± 2.1 µM, respectively (Figure 4.5). Compounds RUBi007, RUBi016 and RUBi018 showed no significant human cell cytotoxicity at 100 µM while RUBi004 and RUBi014 had cytotoxicity IC_{50}s of 23.6 ± 5.8 µM and 32.9 ± 2.2 µM respectively (Figure 4.6).

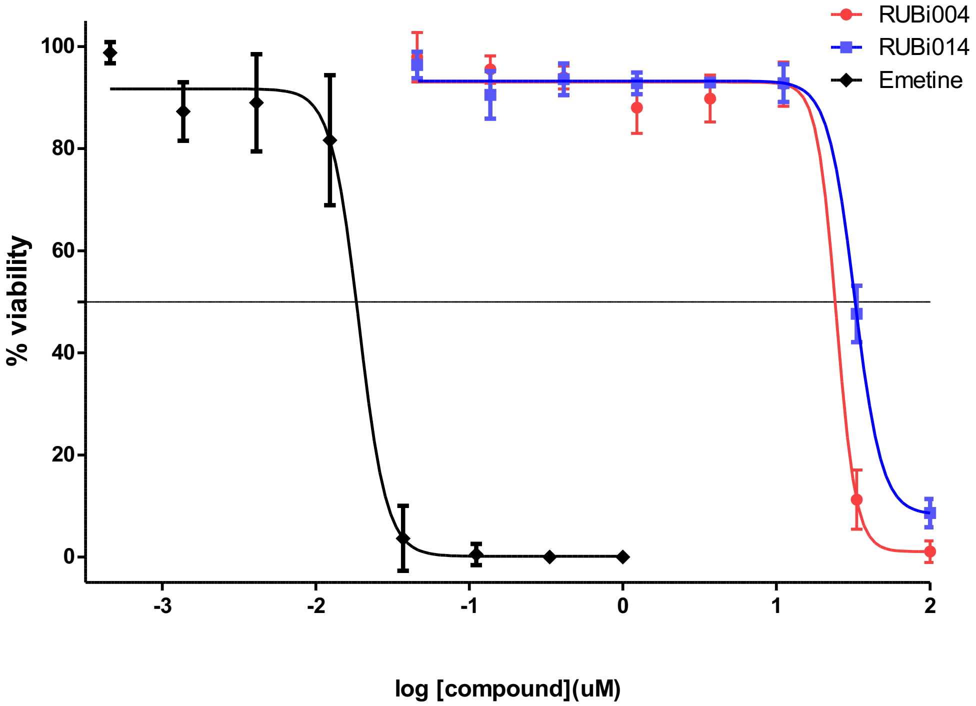


Figure 4.6: IC_{50} cytotoxicity determination of compound RUBi004 (red) and RUBi014 (blue) against HeLa cells. The positive control used was Emetine (black). RUBi004, RUBi014 had IC_{50} cytotoxicities of 23.6 ± 5.8 µM, 32.9 ± 2.2 µM, and 0.018 ± 0.005 µM respectively (error bars refer to SD of the biological replicates).

The compounds were also assayed for activity when used in combination with WR99210 which is a known *Tb*DHFR inhibitor. When the compounds were tested in combination with WR99210, RUBi004, RUBi007, RUBi014 and RUBi018 displayed antagonism while RUBi016 displayed an additive effect (Figure 4.7).

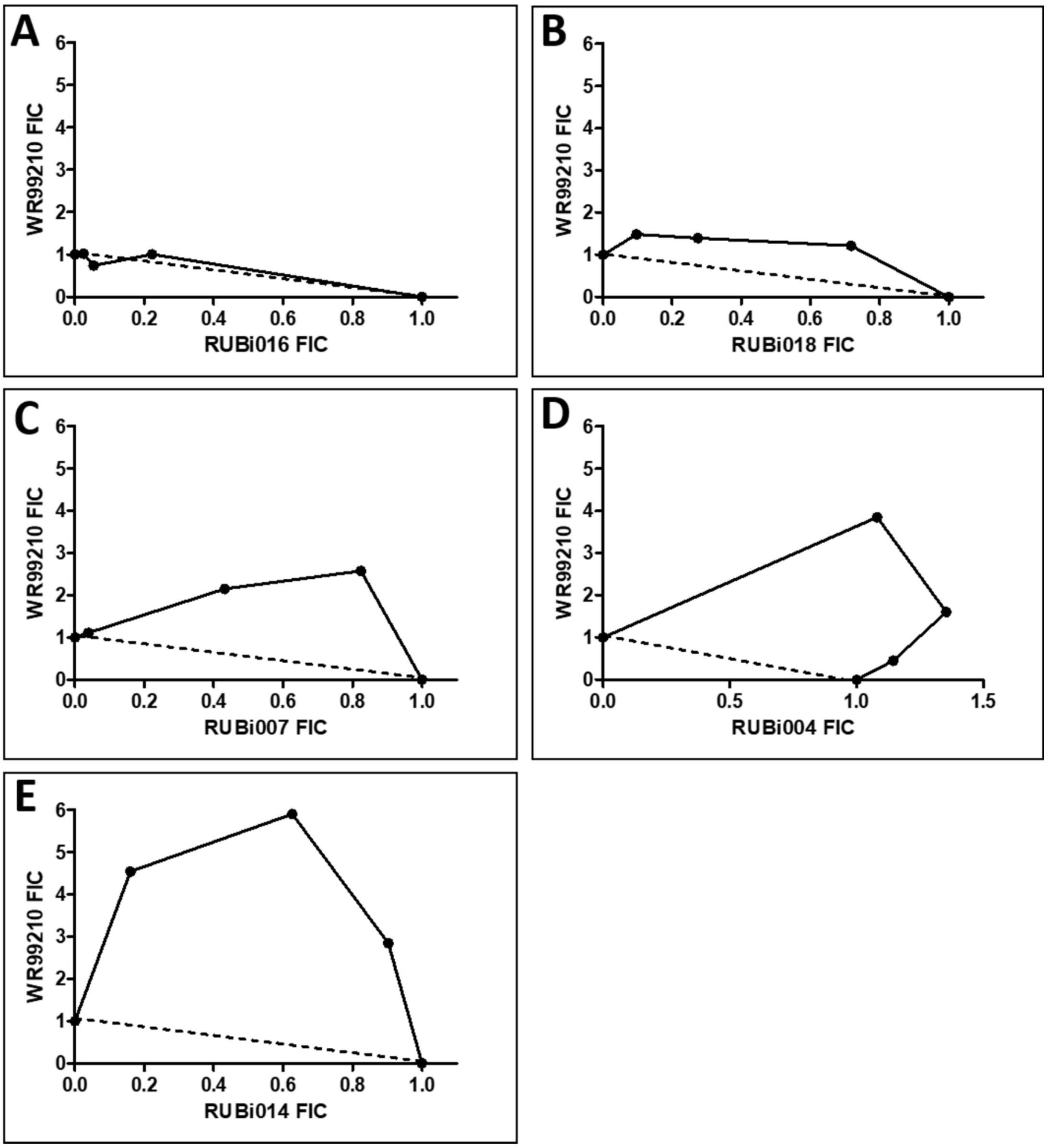


Figure 4.7. Isobologram analysis of compounds RUBi016 (A), RUBi018 (B), RUBi007 (C), RUBi004 (D) and RUBi014 (E) in combination with WR99210. RUBi compounds and WR99210 were employed alone at starting concentrations of 100 µM and 20 µM respectively, and in combination ratios of 75:25, 50:50 and 25:75. IC50 values obtained for the RUBi compounds and WR99210 alone and in combination were used to calculate and plot their fractional inhibitory concentrations (FIC). The dotted line in the graphs denotes an additive effect; FIC values above the line indicate compound antagonism, while values below the line indicate synergism.

The isobologram gave indications of competitive inhibition in some of the compounds, which required us to carry out molecular docking of the compounds against *Tb*DHFR. The five compounds bound to *Tb*DHFR with high binding affinities. RUBi004, RUBi007, RUBi014, RUBi016 and RUBi018 bound to *Tb*DHFR with Autodock Vina binding energies of -39.330 kJ/mol, -39.748 kJ/mol, -36.401 kJ/mol, -31.799 kJ/mol, and -33.054 kJ/mol respectively. This result was interesting because even though *Tb*PTR1 and *Tb*DHFR both contribute to the metabolism of folates their substrate binding sites are distinct [292]. *Tb*DHFR is also less rigid than *Tb*PTR1, and is capable of undergoing significant conformational change when complexed with Thymidine Synthase (TS) [284,293,294]. A detailed analysis of each of the five compounds is presented in the sections below.

4.3.5.1 *Tb*PTR1 Apo Protein

The *Tb*PTR1 apo protein showed a stable and linear trajectory (Figure 4.8-A) with no substantial change in compactness (Figure 4.8-B). The most flexibility was observed in the modeled missing residue loops Q104-K113 and K143-S151 followed by the substrate binding loop (Figure 4.8-C). Changes in displacement as measured by *Average L* were similar to the per residue RMSF (Figure 4.8-D). Residues T9, S95 and A238 showed the highest centrality (Figure 4.8-E).

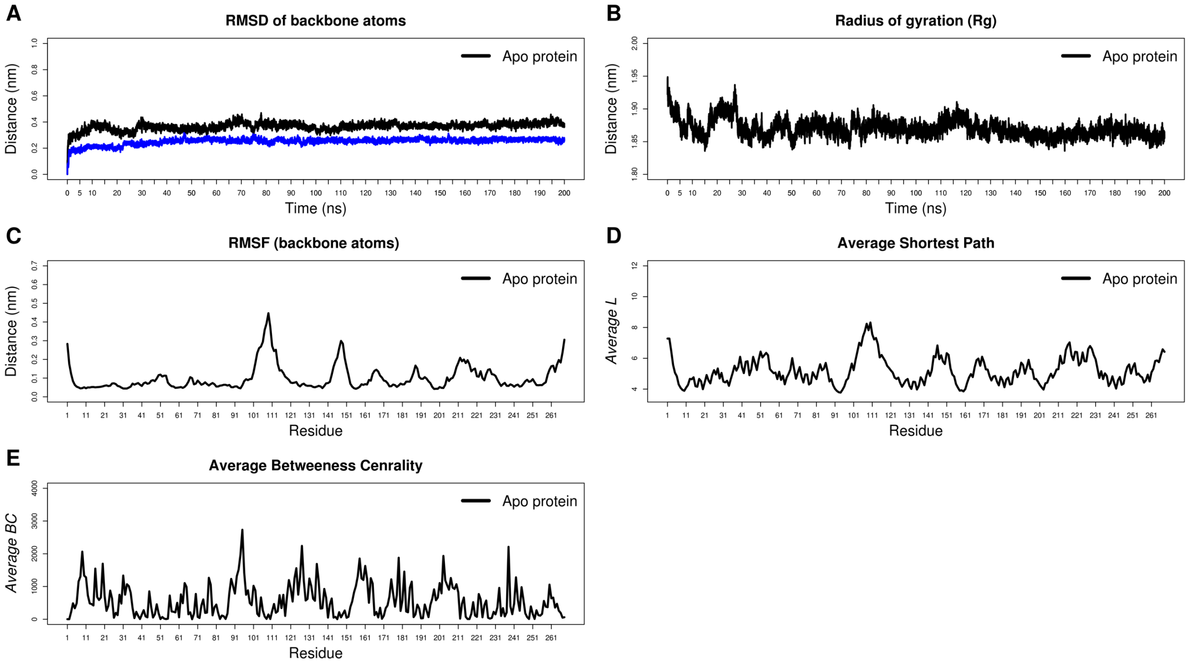


Figure 4.8: Analysis of the *Tb*PTR1 apo protein during 200 ns all-atom MD simulation at 300 K. A) RMSD of the protein back bone heavy atoms, B) Rg, C) Per residue RMSF, D) *Average L*, and E) *Average BC*.

4.3.5.2 RUBi004

Molecular docking: RUBi004 formed a π-sandwich between the NADPH nicotinamide ring and F97 (Figure 4.9-A). It also formed a T-shaped π-π interaction with Y174 and two hydrogen bonds with K13 and NADPH (Figure 4.9-A). Unfavorable interactions (colored in red) included positive-positive interaction between K13 and RUBi004, along with a donor-donor interaction between Y98 and RUBi004 (Figure 4.9-A).

Furthermore, when RUBi004 was docked to *Tb*DHFR it formed hydrogen bonds with the NADPH cofactor and *Tb*DHFR residues S89, T46, and I47 (Figure 4.10-A). RUBi004 also formed π interactions with *Tb*DHFR residues V32, A34, T46, F58, L90, P91, and I169 (Figure 4.10-A). *Tb*DHFR residues A34, V32, M55, F58, S89, F94, Y166, and the NADPH cofactor are known to be involved in the binding of *Tb*DHFR inhibitors such as pyrimethamine and WR99210 [289]. This docking analysis shows that RUBi004 binds to *Tb*DHFR in a similar manner as known *Tb*DHFR inhibitors [289].

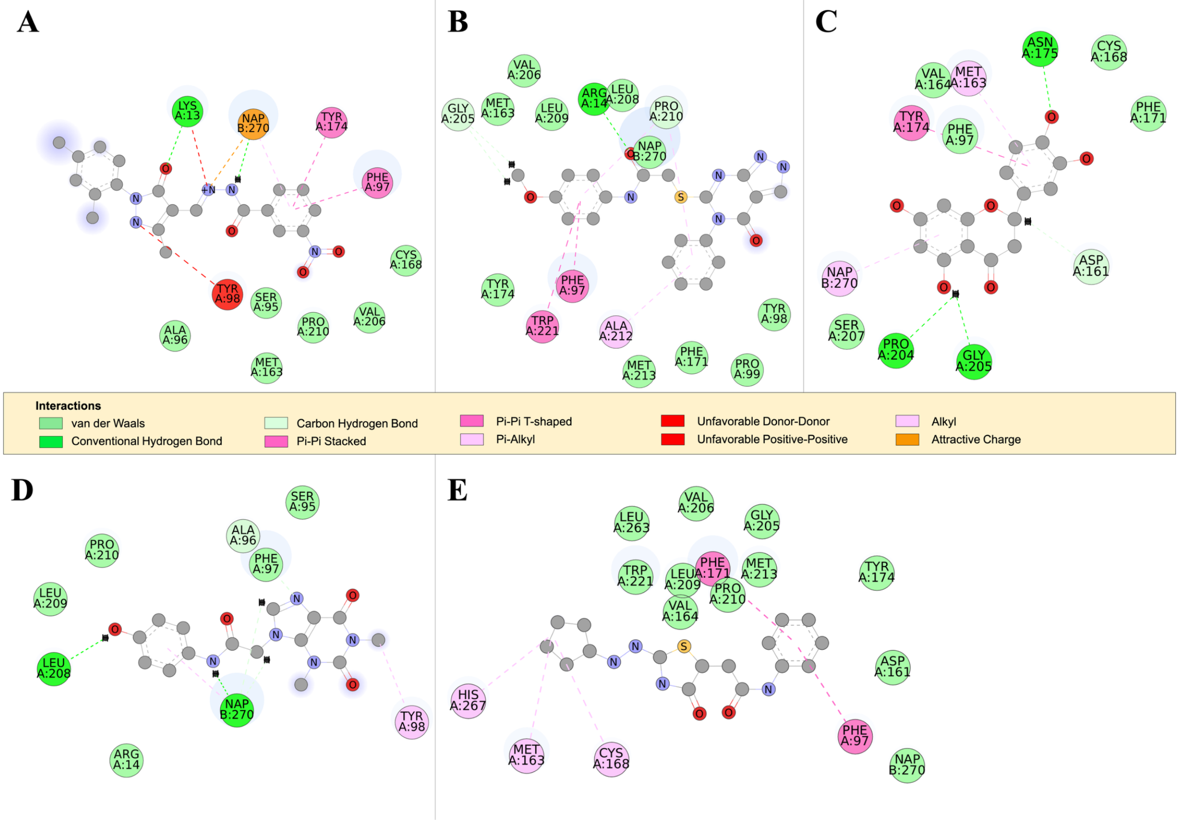


Figure 4.9: The 2D structure of *Tb*PTR1-RUBi compound binding modes. A) *Tb*PTR1-RUBi004, B) *Tb*PTR1-RUBi007, C) *Tb*PTR1-RUBi014, D) *Tb*PTR1-RUBi016, and E) *Tb*PTR1-RUBi018.

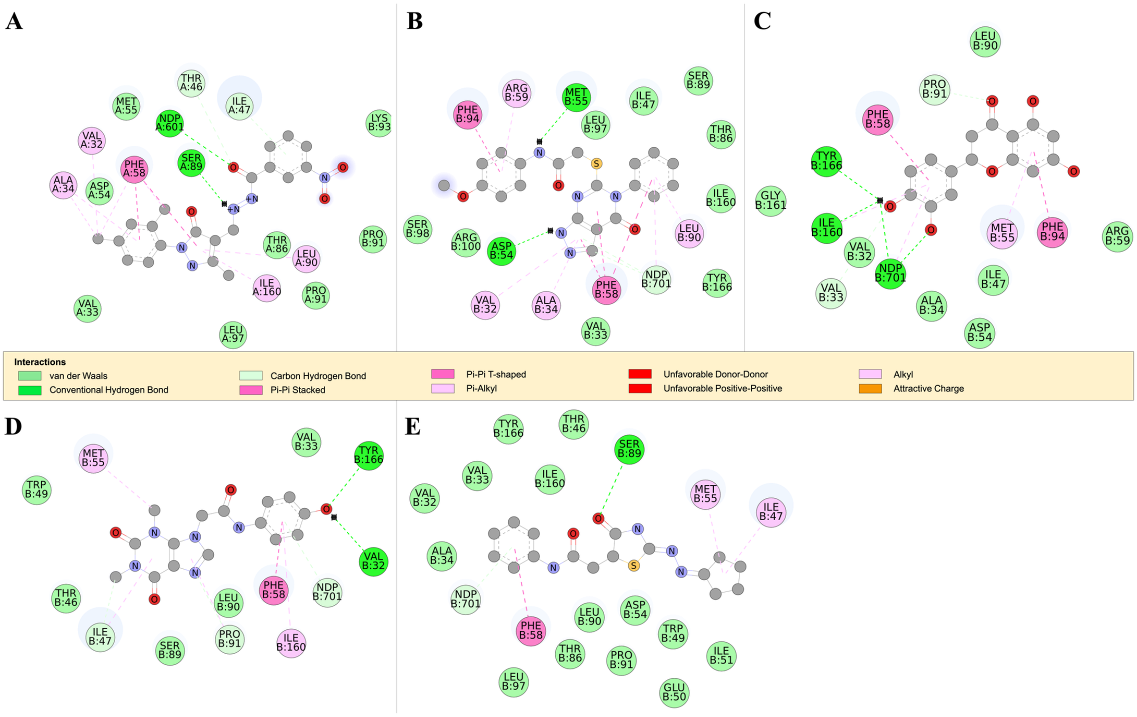


Figure 4.10. The 2D structure of *Tb*DHFR-RUBi compound binding modes. A) *Tb*DHFR-RUBi004, B) *Tb*DHFR-RUBi007, C) *Tb*DHFR-RUBi014, D) *Tb*DHFR-RUBi016, and E) *Tb*DHFR-RUBi018.

Molecular dynamics: The RMSD calculations revealed that the protein backbone and the NADPH cofactor in the *Tb*PTR1-RUBi004 complex to be stable with no conformational changes observed in the ligand (Figure 4.11-A). The binding of the ligand led the protein to become more compact as observed in the Rg analysis (Figure 4.11-B). We observed slight increases in the flexibility of residues K13, F97, Y98 and Y174 via RMSF calculations (Figure 4.11-B). Loop residues M169 and A170 showed increased flexibility while helix residues A188 and A189 showed reduced flexibility (Figure 4.11-B). The substrate binding loop S207 – E 217 was stable (Figure 4.11-B). On average the ligand formed two hydrogen bonds during simulation with *Tb*PTR1 residue K13 and the NADPH cofactor (Figure 4.11-C and Figure 4.9-A).

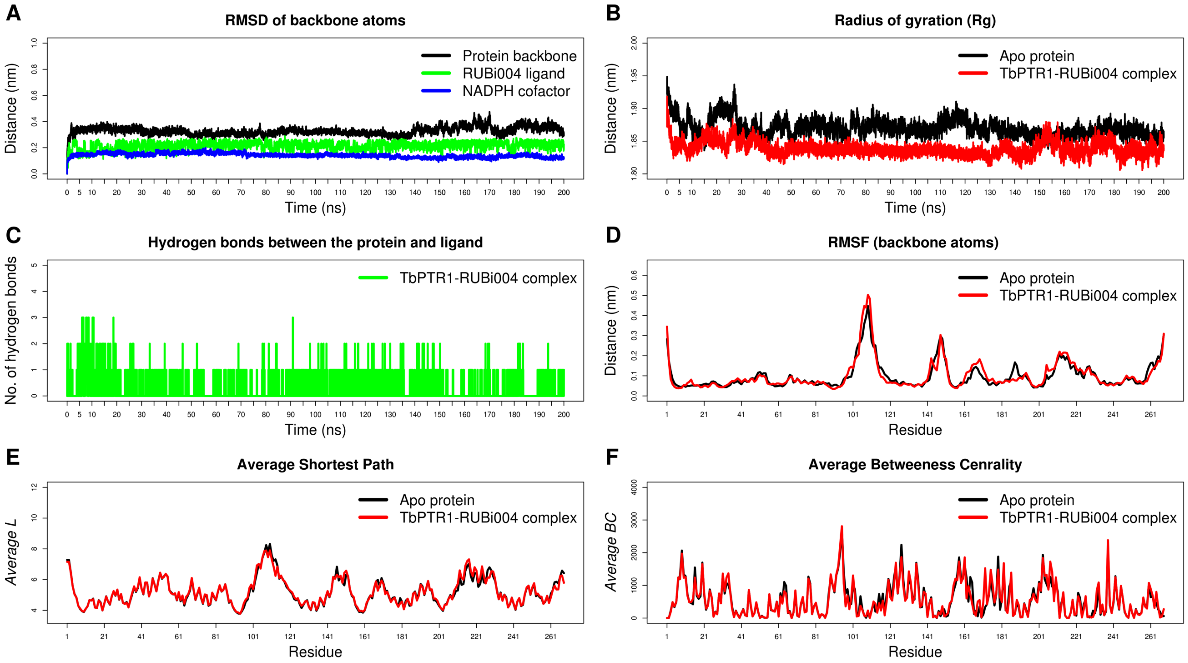


Figure 4.11: Analysis of the *Tb*PTR1-RUBi004 complex during 200 ns all-atom MD simulation at 300 K. A) RMSD of the protein back bone heavy atoms, B) Rg, C) Protein-ligand Hydrogen bond analysis, D) Per residue RMSF, E) *Average L*, and F) *Average BC*.

Binding free energy: RUBi004 bound stably to the protein throughout the MD simulation with a free binding energy of -63.127 ± 14.401 kJ/mol (Table 4.4). A per residue energy decomposition showed that R14, F97, V206, and K218 contributed -8.172 kJ/mol, -9.178 kJ/mol, -3.010 kJ/mol, and -4.170 kJ/mol respectively to binding (Figure 4.12-A and Figure 4.12-B-i). Unfavorable energy contributions included D161 and E231 that contributed 3.790 kJ/mol and 3.110 kJ/mol respectively (Figure 4.12-A and Figure 4.12-B-i).

Table 4.4: A decomposition of the binding energy components obtained from MM-PBSA

Ligand	Van der Waal energy (kJ/mol)	Electrostatic energy (kJ/mol)	Polar solvation energy (kJ/mol)	SASA energy (kJ/mol)	Binding energy (kJ/mol)
RUBi001	-147.290 ± 9.684	-72.775 ± 7.789	113.647 ± 6.897	-13.131 ± 0.561	-119.549 ± 9.302
RUBi002	-158.331 ± 10.799	-54.957 ± 17.424	167.445 ± 25.372	-20.181 ± 0.987	-66.024 ± 17.960
RUBi003	-120.331 ± 9.856	-45.157 ± 16.390	88.501 ± 14.847	-12.178 ± 0.714	-89.166 ± 10.016
RUBi004	-89.386 ± 8.660	-2.494 ± 19.739	40.462 ± 29.099	-11.708 ± 1.480	-63.127 ± 14.401
RUBi005	-176.822 ± 12.683	-21.677 ± 11.102	105.804 ± 13.441	-18.946 ± 0.896	-111.641 ± 15.287
RUBi006	-129.667 ± 9.297	-2.552 ± 7.030	72.606 ± 6.948	-14.686 ± 0.722	-74.298 ± 9.423
RUBi007	-124.684 ± 8.329	-45.487 ± 7.400	97.333 ± 9.983	-15.092 ± 0.730	-87.931 ± 9.776
RUBi008	-133.753 ± 8.370	-46.219 ± 11.226	110.761 ± 15.299	-16.040 ± 0.753	-85.250 ± 10.888
RUBi009	-143.867 ± 18.665	1.809 ± 20.494	94.287 ± 13.356	-18.040 ± 1.565	-65.812 ± 21.921
RUBi010	-144.193 ± 11.717	-138.108 ± 32.060	230.551 ± 48.093	-16.209 ± 0.689	-67.960 ± 23.305
RUBi011	-158.890 ± 9.774	-22.047 ± 12.754	106.149 ± 15.520	-16.788 ± 0.854	-91.577 ± 11.944
RUBi012	-115.925 ± 12.327	41.151 ± 9.268	103.048 ± 20.694	-13.047 ± 1.114	15.227 ± 13.133
RUBi013	-139.991 ± 10.650	-48.508 ± 27.677	142.199 ± 30.711	-15.508 ± 1.132	-61.808 ± 12.718
RUBi014	-98.534 ± 17.175	-17.122 ± 19.998	70.730 ± 28.464	-11.472 ± 1.506	-56.399 ± 11.696
RUBi015	-106.414 ± 10.967	-28.566 ± 11.142	76.073 ± 17.307	-13.479 ± 1.334	-72.386 ± 12.019
RUBi016	-87.250 ± 8.743	21.304 ± 8.487	53.635 ± 10.796	-11.042 ± 0.799	-23.353 ± 10.361
RUBi017	-129.769 ± 12.116	-32.613 ± 13.539	91.673 ± 19.346	-15.262 ± 1.235	-85.971 ± 11.865

RUBi018	-132.408 ± 11.987	-14.638 ± 9.183	74.896 ± 10.635	-15.764 ± 1.018	-87.914 ± 12.026

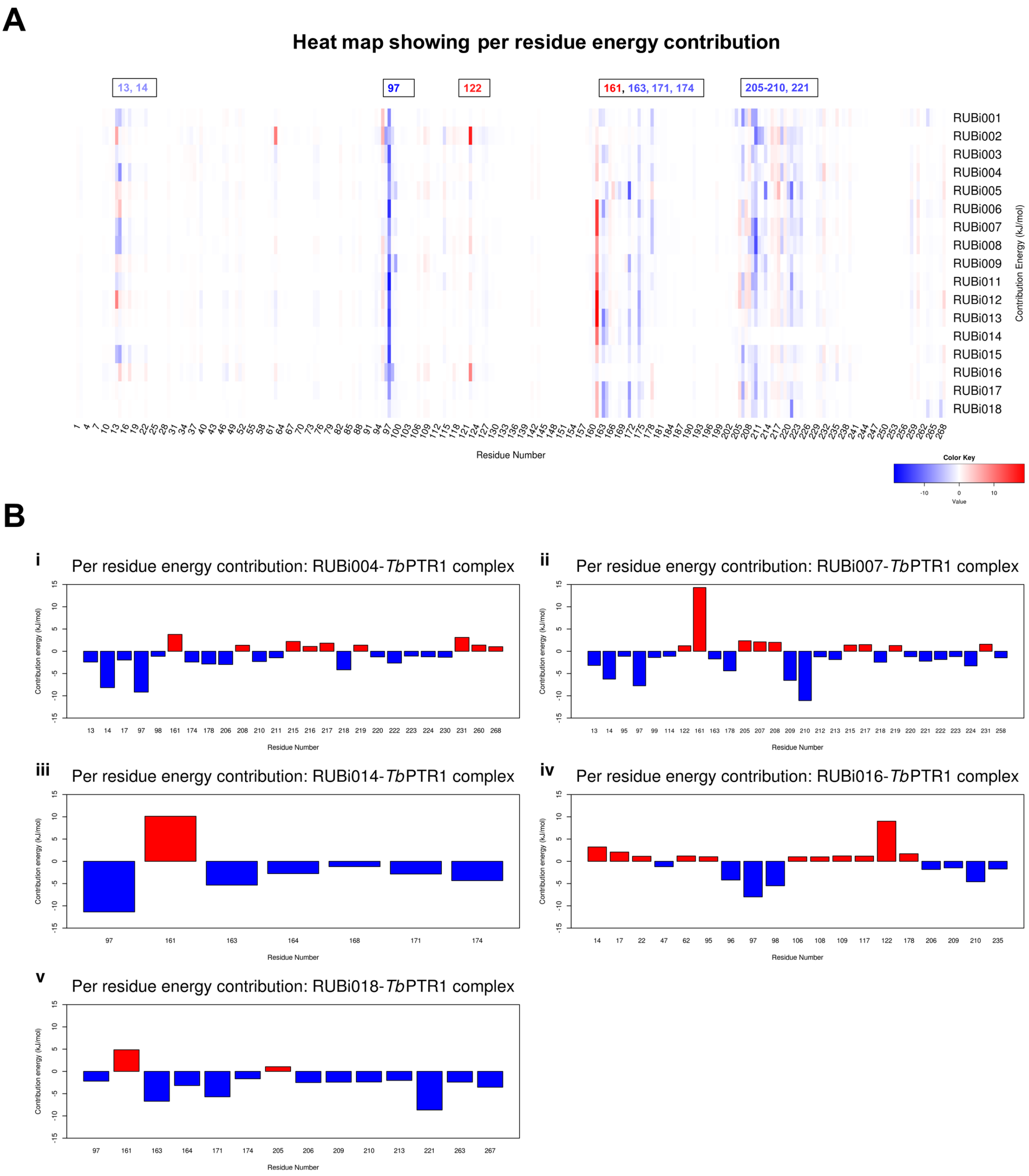


Figure 4.12: A heat map and histogram showing the per residue energy contributions to binding (energy in kJ/mol). A) A heat map showing the per residue energy contributions for all the protein-ligand systems B) A histogram showing the main residues contributing energetically to binding in i) *Tb*PTR1-RUBi004, ii) *Tb*PTR1-RUBi007, iii) *Tb*PTR1-RUBi014, iv) *Tb*PTR1-RUBi016, and v) *Tb*PTR1-RUBi018 complexes.

Dynamic residue network: As shown previously by a general correlation between *average BC*, L^{-1} and RMSF^{-1} was observed [205,209]. Pearson correlation coefficient values are presented in the Supplementary Table 5.

There was not much spatial displacement of the substrate binding loop and the small α6 helix as shown by comparison of *Average L* between apo protein and *Tb*PTR1-RUBi004 complex (Figure 4.11-E).

Average BC is an important metric to identify residues critical for communication flow within the protein network [132]. Residues T9, S95, and A238 showed the highest *average BC* in the *Tb*PTR1 protein in both apo and *Tb*PTR1-RUBi004 complex (Figure 4.11-F). Residue T9 is conserved among all the PTR1 orthologues while S95 is conserved among the trypanosomatid PTR1 orthologues only (Figure 4.3). A238 is conserved in all the PTR1 orthologues except *Lm*PTR1 where it is replaced by a serine residue (Figure 4.3). In the *Tb*PTR1-RUBi004 complex V164, S172, and S207 showed increases in *average BC* compared to apo protein (Figure 4.11-F). S207 is also conserved among all the PTR1 orthologues (Figure 4.3).

BC related residue interaction analysis: A close examination of the residue interaction networks of important residues identified via *Average BC* calculations showed that T9 formed vdW interactions with A11, N92, A94, and hydrogen bonds with V91 and N93 in the *Tb*PTR1 apo protein (Figure 4.13-A). A11 in turn formed vdW interactions with K13, R14 which contributed -8.172 kJ/mol to RUBi004 binding during simulation (Figure 4.12-B-i). While S95 formed hydrogen bonds with A96, N127, and the NADPH cofactor (Figure 4.13-B). Furthermore, S95 formed vdW interactions with Y174 and F97 both of which are involved in ligand binding (Figure 4.13-C). Lastly A238 formed an alkyl interaction with A18 that is hydrogen bonded to I15 (Figure 4.14-C). These interactions with functionally important residues indicate that *Average BC* is helpful in identifying residues crucial to communication flow within the protein-ligand dynamic network.

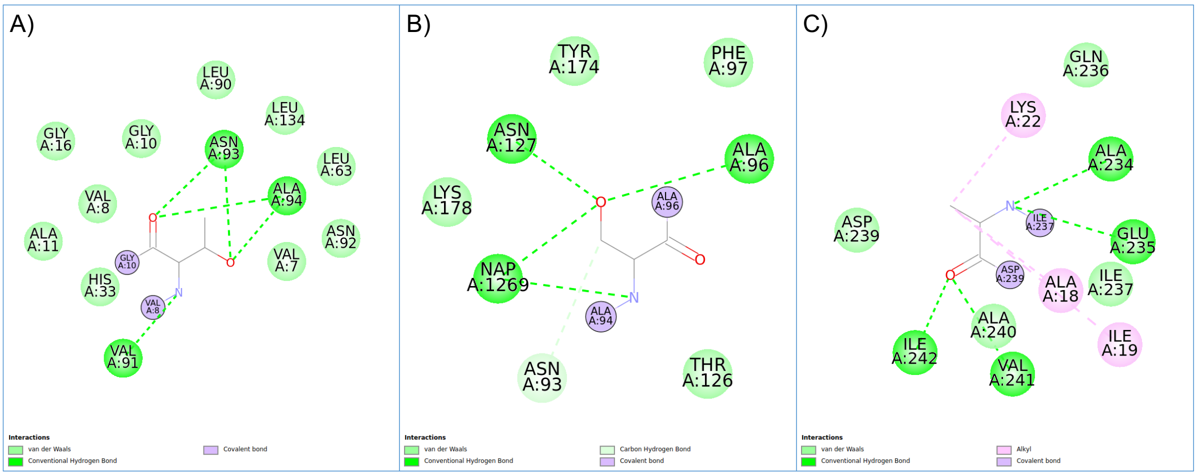


Figure 4.13: 2D interaction network around residues in *Tb*PTR1 (PDB:2X9N). A) T9, B) S95, and C) A238

The binding of the RUBi004 resulted in altered information flow across the protein network especially at V164, S172, and S207 all of which were interacting with residues involved in ligand and NADPH cofactor binding (Figure 4.14).

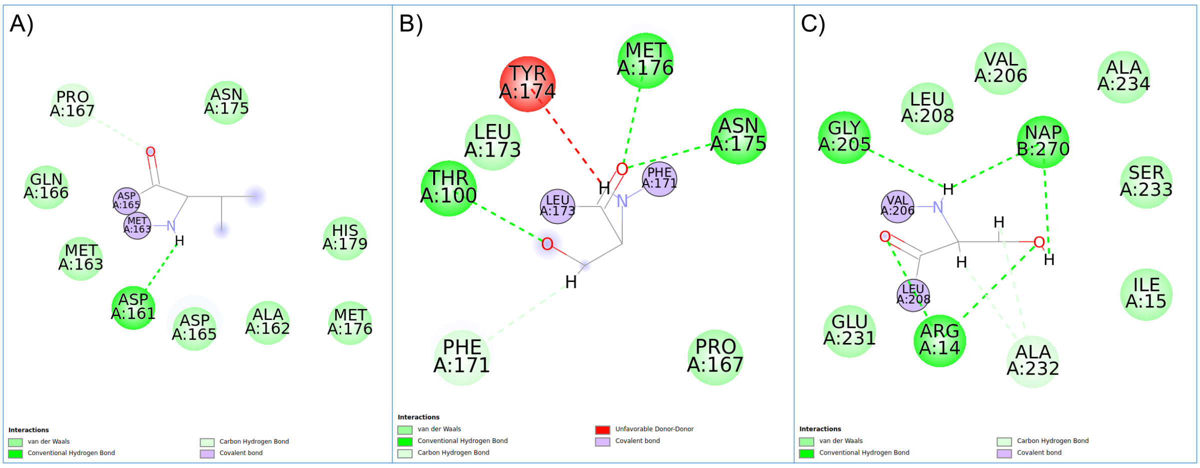


Figure 4.14: 2D interaction network around residues *Tb*PTR1- RUBi004 complex (PDB:2X9N). A) V164, B) S172, and C) S207

VAL164 formed vdW interactions with N175 and had a hydrogen bond with D161 (Figure 4.14-A). S172 had an unfavorable donor-donor interaction with Y174 (Figure 4.14-B). Lastly, residue S207 formed a hydrogen bond with the NADPH cofactor and residue R14 (Figure 4.14-C). All these interactions were of note because docking analysis showed that K13, F97, Y98 and Y174 were involved in RUBi004 binding (Figure 4.9-A).

PCA Analysis: PCA analysis revealed differences in the concerted motions between the apo protein and protein-ligand complexes (Figure 4.15 and Figure 4.16). The largest motions observed in both the *Tb*PTR1 apo protein and *Tb*PTR1-RUBi004 complex were in the substrate binding loop (S207-E215), the α6 helix (G214-V225), C160-Y174 loop region, C-terminal residues H267-A268, the modelled loop 1 (Q104-G113), and modelled loop 2 (K143-S151) (Figure 4.15 and Figure 4.16).

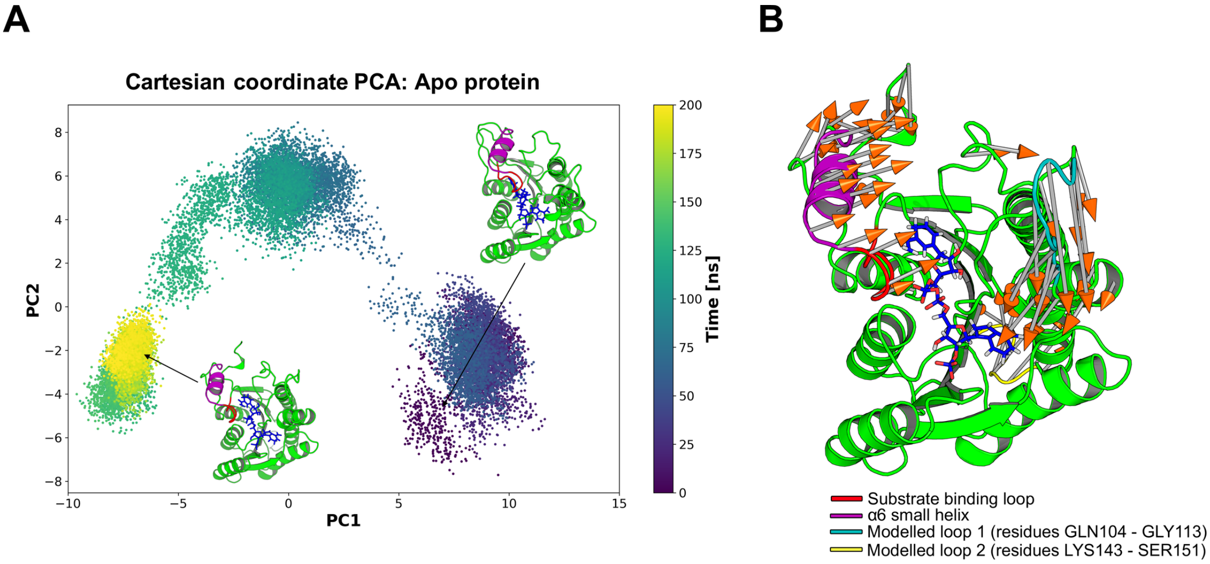


Figure 4.15: Principal component analysis of the *Tb*PTR1 apo protein. The motions of the protein during 200ns of all-atom MD simulation are shown along the first and second principal components (PC1 and PC2). The substrate binding loop (S207-E215) is colored red, α6 alpha helix (G214 - V225) is colored magenta, modelled missing residues loop 1 colored cyan, modelled missing loop residues loop2 colored yellow, and NADPH cofactor colored blue. PC1 explained 54% of the variance while PC2 explained 18%. A) Projection of the protein-ligand complex dynamics along the PC1 and PC2. B) The differential motions described by PC1 and PC2 are shown by light gray arrows with orange tips.

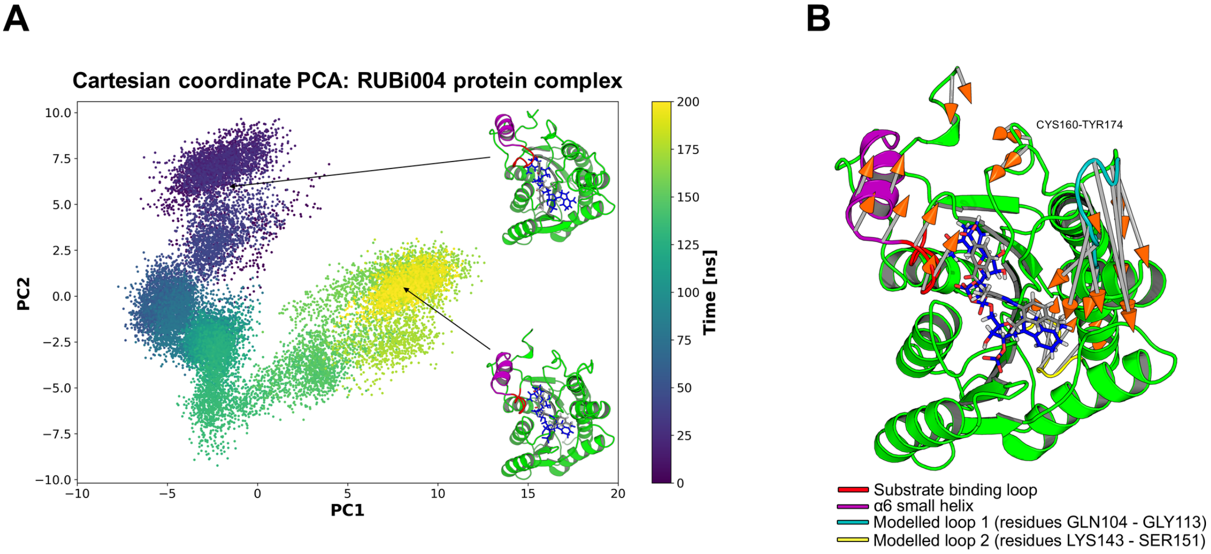


Figure 4.16: Principal component analysis of the RUBi004 - protein complex. The motions of the complex during 200ns of all-atom MD simulation are shown along the first and second principal components (PC1 and PC2). The substrate binding loop (S207-E215) is colored red, α6 alpha helix (G214 - V225) is colored magenta, modelled missing residues loop 1 colored cyan, modelled missing loop residues loop2 colored yellow, NADPH cofactor colored blue, and RUBi004 colored gray. PC1 explained 46% of the variance while PC2 explained 26%. A) Projection of the protein-ligand complex dynamics along the PC1 and PC2. B) The differential motions described by PC1 and PC2 are shown by light gray arrows with orange tips.

4.3.5.3 RUBi007

Molecular docking: RUBi007 formed T-shaped π-π interactions with F97 and W221. Along with π-alkyl interactions with P210 and A212 (Figure 4.9-B). It also formed hydrogen bonds with R14, and G205 (Figure 4.9-B). Molecular docking of RUBi007 to *Tb*DHFR showed that it formed hydrogen bonds with the NADPH cofactor, T46, I47, and S89 along with π interactions with V32, A34, F58, L90, and I160 (Figure 4.10-B).

These residues are known to be involved in the binding of known *Tb*DHFR inhibitors such as pyrimethamine and WR99210 [289].

Molecular dynamics: The *Tb*PTR1-RUBi007 complex showed a linear and stable conformation for the protein, RUBi007 and the NADPH cofactor (Figure 4.17-A). The binding of the ligand caused increase in the compactness of the protein as shown in Figure 4.17-B. In addition, we also observed a decrease in the flexibility of the substrate binding loop (G205-W221) as a result of ligand binding (Figure 4.17-D) which coincided with a change in spatial displacement as measured by *Average L* (Figure 4.17-E). Residues that showed decreased flexibility included R14, F97, G205, P210, A212, and W221 while those that showed increased flexibility included G247, S248, and A249 (Figure 4.17-D). RUBi007 formed on average 2 hydrogen bonds with R14 and G205 during the MD simulation (Figure 4.17-B and Figure 4.9-B).

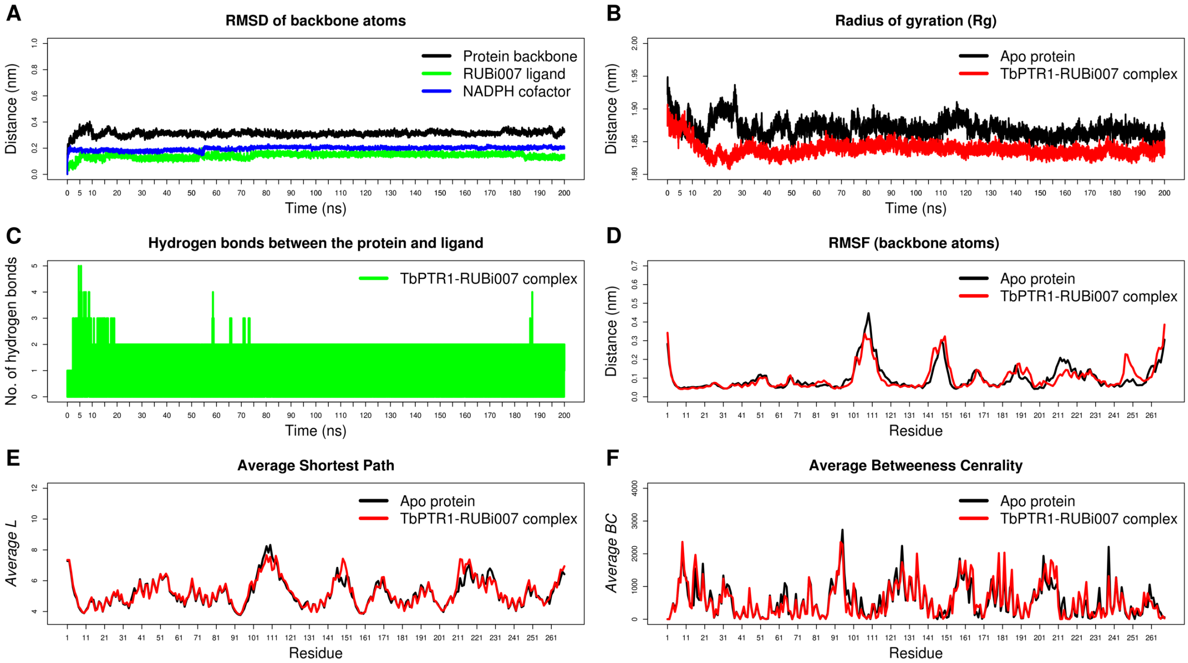


Figure 4.17: Analysis of the *Tb*PTR1-RUBi007 complex during 200 ns all-atom MD simulation at 300 K. A) RMSD of the protein back bone heavy atoms, B) Rg, C) Protein-ligand Hydrogen bond analysis, D) Per residue RMSF, E) *Average L*, and F) *Average BC*.

Binding free energy: RUBi007 bound to the protein stably throughout the MD simulation with a free binding energy of -87.931 ± 9.776 kJ/mol (Table 4.4). A per residue decomposition showed that K13, R14, F97, K178, L209, P210, and K224 contributed -3.200 kJ/mol, -6.260 kJ/mol, -7.750 kJ/mol, -4.4 kJ/mol, -6.55 kJ/mol, -11.08 kJ/mol, and -3.30 kJ/mol (Figure 4.12-A and Figure 4.12-B-ii). D161 gave an unfavorable energy contribution of 14.32 kJ/mol (Figure 4.13-B-ii).

Dynamic residue network: Compared to the apo protein the *Tb*PTR1-RUBi007 complex, *L* indicated spatial displacement of the substrate binding loop and the small α6 helix (Figure 4.17-E). In comparison to the apo protein, V211-G214 showed increases in *L* (Figure 4.17-E). The observed increases in *L* correlated well with the per residue RMSF (Figure 4.17-E) [209].

In the *Tb*PTR1-RUBi007 complex, I15, M163, S207, L208, and P210 showed increases in average BC (Figure 4.18-F). Multiple sequence analysis showed that residue I15 was conserved in all the PTR1 orthologues except *Lm*PTR1 where it is substituted for a leucine (Figure 4.3). Residue M163 was conserved only among the trypanosomatids, while L208 was conserved only in *T. brucei and T. cruzi* (Figure 4.3).

BC related residue interaction analysis: As seen in the other *Tb*PTR1-ligand complexes, these residues with higher average BC values were involved in ligand binding. Residue I15 was covalently boded to R14, while M163 formed vdW interactions with the RUBi007 ligand (Figure 4.18-A and Figure 4.18-B). Residue 207 had a hydrogen bond with the NADPH cofactor (Figure 4.18-C). Residue L208 had vdW interactions with the RUBi007 ligand and had a hydrogen bond with both R14 and the NADPH cofactor (Figure 4.18-D). Residue P210 had a π alkyl interaction with F97, vdW interactions with R14, and had a hydrogen bond with the RUBi007 ligand (Figure 4.18-E).

PCA Analysis: The largest motions observed in the *Tb*PTR1-RUBi007 complex which differed from the apo protein, were in the α6 helix, C-terminal residues H267-A268, the modelled loop 1, and modelled loop 2 (Figure 4.19-A and Figure 4.19-B).

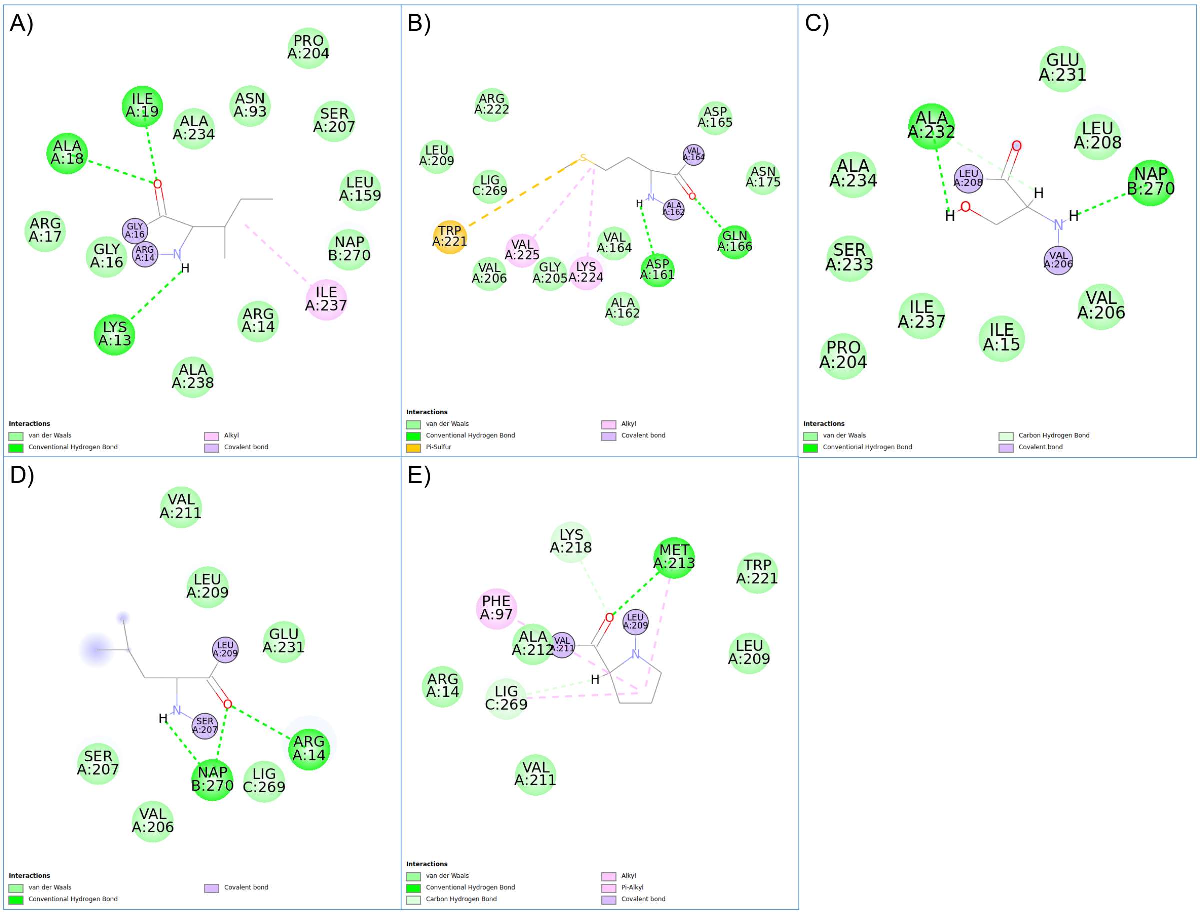


Figure 4.18: 2D interaction network around residues *Tb*PTR1-RUBi007 complex (PDB:2X9N). A) I15, B) M163, C) S207, D) L208, and E) P210

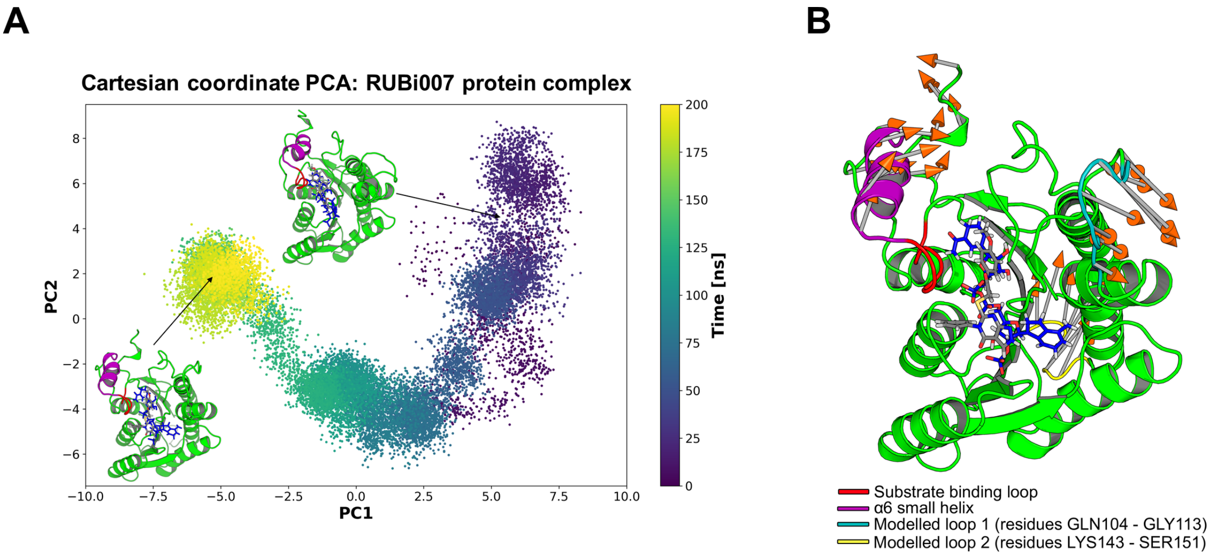


Figure 4.19: Principal component analysis of the RUBi007 - protein complex. The motions of the complex during 200ns of all-atom MD simulation shown along the first and second principal components (PC1 and PC2). The

substrate binding loop (S207-E215) is colored red, α6 alpha helix (G214 - V225) is colored magenta, modelled missing residues loop 1 colored cyan, modelled missing loop residues loop2 colored yellow, NADPH cofactor colored blue, and RUBi007 colored gray. PC1 explained 40% of the variance while PC2 explained 21%. A) Projection of the protein-ligand complex dynamics along the PC1 and PC2. B) The differential motions described by PC1 and PC2 are shown by light gray arrows with orange tips.

4.3.5.4 RUBi014

Molecular docking: RUBi014, also known as eriodictyol, has previously been reported to be selectively anti-protozoal with activity against *T. brucei* and *Leishmania* in culture [295,296]. RUBi014 formed a T-shaped π-π interaction with Y174, π-alkyl interactions with M163 and a π-alkyl interaction between with the NADPH nicotinamide ring (Figure 4.9-C). It also formed hydrogen bonds with D161, N175, P204, and G205 (Figure 4.9-C).

Docking of RUBi014 to *Tb*DHFR showed that it bound in a similar mode to known *Tb*DHFR inhibitors [289]. RUBi014 formed hydrogen bonds with the NADPH cofactor and V33, P91, I160, and Y166 (Figure 4.10-C). It also formed π interactions with F58, M55, and F94 (Figure 4.10-C).

Molecular dynamics: The *Tb*PTR1-RUBi014 showed a linear trajectory but significantly differed from its initial structure (Figure 4.20-A). The protein backbone showed a notable conformational change, while the NADPH cofactor and ligand were generally stable (Figure 4.20-A). There were no significant changes in protein compactness (Figure 4.20-B). We observed a significant increase in the flexibility of the substrate binding loop and the small α6 helix (Figure 4.20-D). We also observed increases in the flexibility of D161, M163, Y174, N175, and P204 – R229 (Supplementary Figure 5N). RUBi014 formed an average of 4 hydrogen bonds with D161, N175, P204 and G205 during the MD simulation (Figure 4.20-C and Figure 4.9-C).

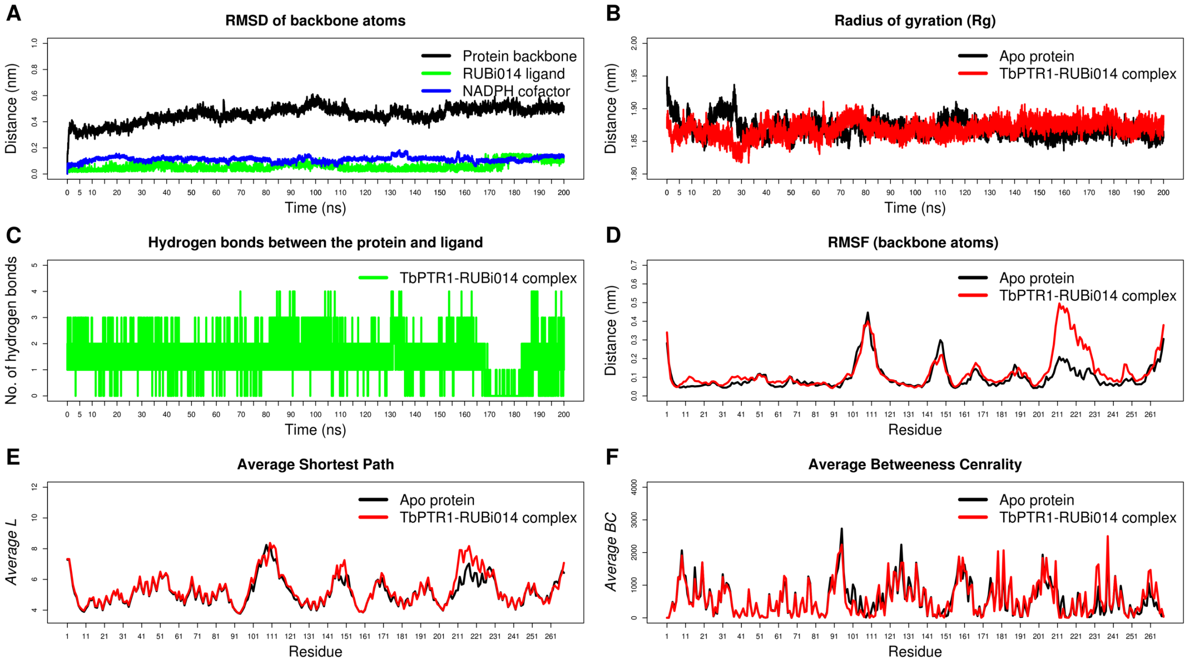


Figure 4.20: Analysis of the *Tb*PTR1-RUBi014 complex during 200 ns all-atom MD simulation at 300 K. A) RMSD of the protein back bone heavy atoms, B) Rg, C) Protein-ligand Hydrogen bond analysis, D) Per residue RMSF, E) *Average L*, and F) *Average BC*.

Binding free energy: RUBi014 bound to the protein stably throughout the MD simulation with a free binding energy of -56.399 ± 11.696 kJ/mol (Table 4.4). A per residue decomposition showed that residues F97, M163, F171, and Y174 contributed -11. 369 kJ/mol, -5.340 kJ/mol, -3.000 kJ/mol, and -4.34 kJ/mol respectively (Figure 4.12-B-iii). D161 gave an unfavorable energy contribution of 10.130 kJ/mol (Figure 4.12-B-iii). D161, M163, Y174 all showed increases in flexibility in their RMSF (Figure 4.20-D).

Dynamic residue network: Compared to the apo protein the *Tb*PTR1-RUBi014 complex showed significant differences especially in the substrate binding loop and the small α6 helix which correlated well with the per residue RMSF (Figure 4.20-E and Figure 4.20-D).

In the *Tb*PTR1-RUBi014 complex, C160, G205, P210 and S233 showed increases in average BC. Our MSA results show that C160 and P210 are conserved only among *T. brucei* and *T.*

cruzi while S233 is not conserved in the human orthologue (Figure 4.3). Residue G205 is, on the other hand, conserved among all the PTR1 orthologues (Figure 4.3).

BC related residue interaction analysis: As shown in the previous protein-ligand complexes residues that showed increased *Average BC* were involved in ligand binding. Residue C160 was covalently bound to D161 and had vdW interactions with P204 and (Figure 4.21-A). Residue P205 formed a hydrogen bond with RUBi014 (Figure 4.21-B). Residue P210 formed vdW interactions with L208 and V211 (Figure 4.21-C). Lastly, residue S233 had vdW interactions with S207 (Figure 4.21-D). *Tb*PTR1 D161, P204, and G205 were shown to be important in RUBi014 binding (Figure 4.9-C). While S207, and L208 and V211 are located in the substrate binding loop (Figure 4.3).

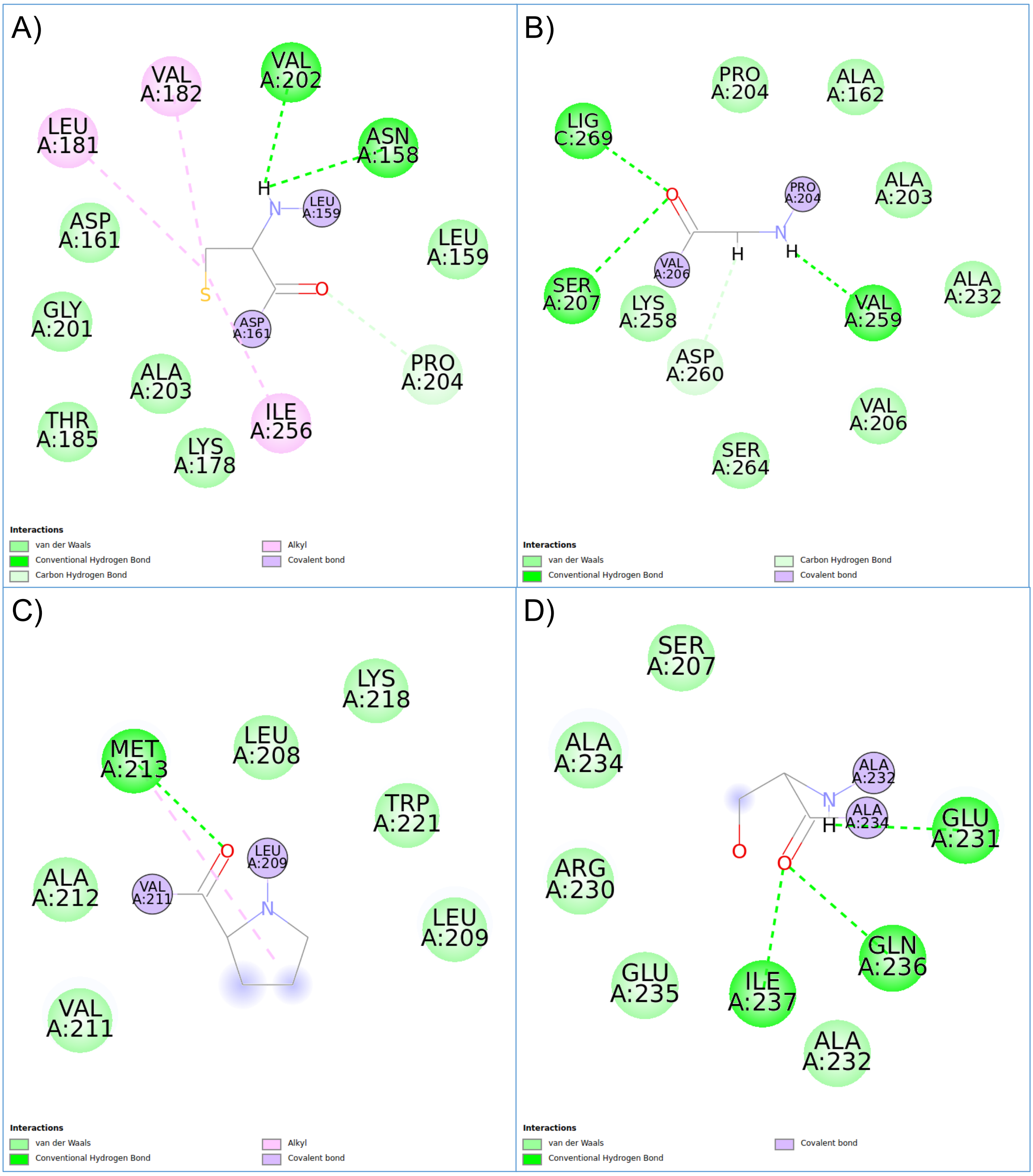


Figure 4.21: 2D interaction network around residues *Tb*PTR1- RUBi014 complex (PDB:2X9N). A) C160, B) G205, C) P210, and D) S233

PCA Analysis: *Tb*PTR1-RUBi014 complex showed the largest motions in the substrate binding loop, the α6 helix, C-terminal residues H267-A268, the modelled loop 1, and modelled loop 2

(Figure 4.22-A and Figure 4.22-B). The changes in the substrate binding loop were in keeping with the observed changes in RMSF and *Average L* (Figure 4.20-D and Figure 4.20-E).

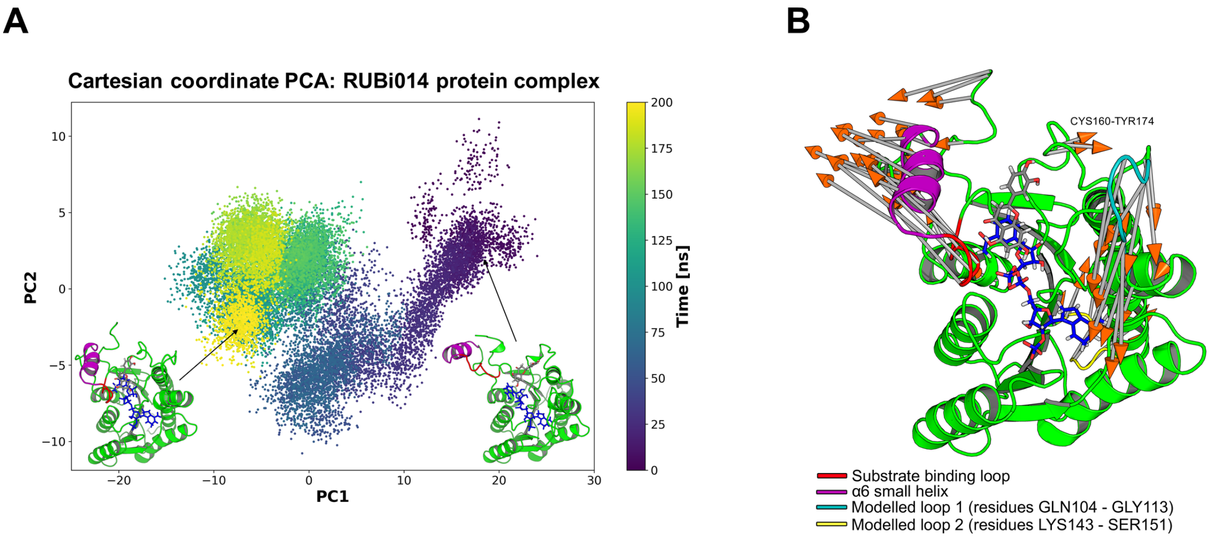


Figure 4.22: Principal component analysis of the RUBi014 - protein complex. The motions of the complex during 200ns of all-atom MD simulation shown along the first and second principal components (PC1 and PC2). The substrate binding loop (S207-E215) is colored red, α6 alpha helix (G214 - V225) is colored magenta, modelled missing residues loop 1 colored cyan, modelled missing loop residues loop2 colored yellow, NADPH cofactor colored blue, and RUBi014 colored gray. PC1 explained 66% of the variance while PC2 explained 12%. A) Projection of the protein-ligand complex dynamics along the PC1 and PC2. B) The differential motions described by PC1 and PC2 are shown by light gray arrows with orange tips.

4.3.5.5 RUBi016

Molecular docking: RUBi016 formed π-alkyl interactions with the NADPH nicotinamide ring and Y98. RUBi016 formed hydrogen bonds with A96, and L208 (Figure 4.9-D). When docked to *Tb*DHFR, RUBi016 formed hydrogen bonds with the NADPH cofactor, Y166 and V32 along with π interactions with F58 and I160 (Figure 4.10-D).

Molecular dynamics: *Tb*PTR1-RUBi016 complex was very stable with minimal conformational changes (Figure 4.23-A). There was a slight increase in the protein compactness due to ligand binding (6.23-B). There was a significant increase in the flexibility of the substrate binding loop and the small α6 helix because of the ligand binding (Figure 4.23-D). As observed in the *Tb*PTR1-RUBi014 complex it appears the binding of the RUBi016 ligand destabilizes the active site (Figure 4.20-D and Figure 4.23-D). Further, we observed decreased flexibility of S96 and Y98 while P167 and L208 – G228 showed increased flexibility (Figure 4.23-D). RUBi016 formed three hydrogen bonds with the protein during the MD simulation and these were with A96, L208 and the NADPH cofactor (Figure 4.23-C and Figure 4.9-D).

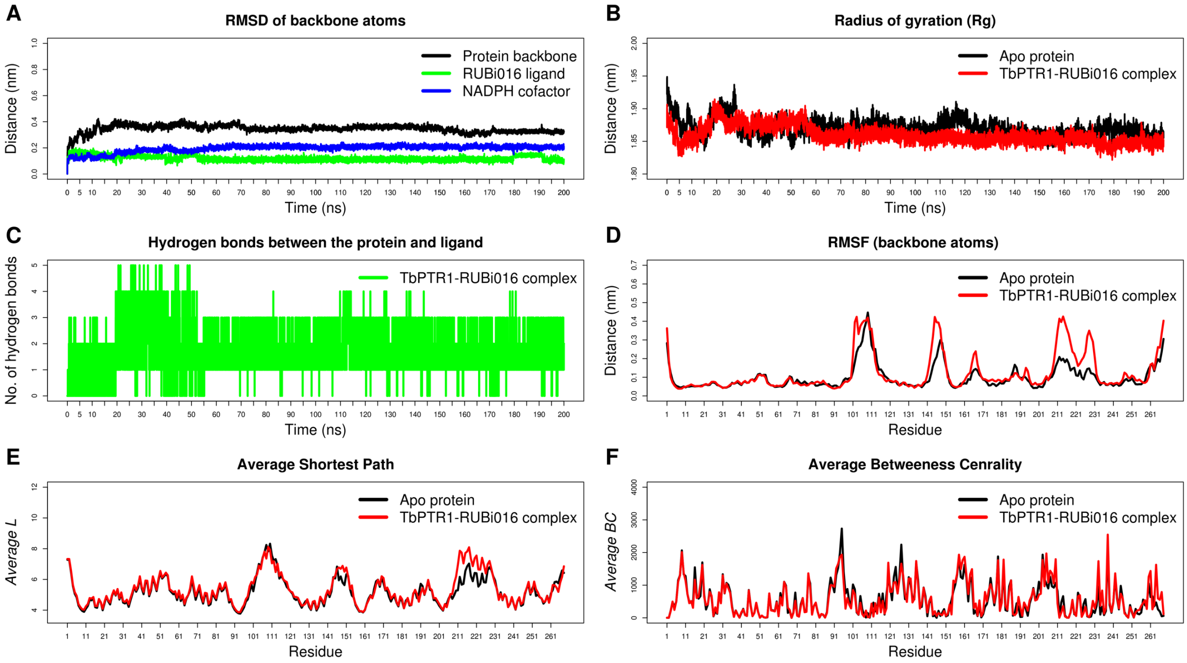


Figure 4.23: Analysis of the *Tb*PTR1-RUBi016 complex during 200 ns all-atom MD simulation at 300 K. A) RMSD of the protein back bone heavy atoms, B) Rg, C) Protein-ligand Hydrogen bond analysis, D) Per residue RMSF, E) *Average L*, and F) *Average BC*.

Binding free energy: RUBi016 bound to the protein stably throughout the MD simulation with a free binding energy of -23.353 ± 10.361 kJ/mol (Table 4.4). A per residue decomposition showed that A96, F97, Y98, and P210 contributed -4.202 kJ/mol, -7.992 kJ/mol, -5.497 kJ/mol, and -4.590 kJ/mol respectively. R14 and E122 had unfavorable energy contributions of 3.230 kJ/mol and 8.980 kJ/mol respectively (Figure 4.12-B-iv).

Dynamic residue network: Comparison of the apo protein to the *Tb*PTR1-RUBi016 complex using *L* indicated spatial displacement of the substrate binding loop and the small α6 helix (Figure 4.23-E). In comparison to the apo protein, V211-LYS218 showed increased *Average L* (Figure 4.23-E). The observed increases in *L* correlated well with residue fluctuations (Figure 4.23-E and Figure 4.23-D) [209].

Residues S95 and F97 became less central while C160, G205, P210, and S264 became more central in the *Tb*PTR1-RUBi016 complex (Figure 4.23-F). Multiple sequence analysis showed that residue S95 was conserved among the trypanosomatids only while S264 was conserved only in *Tb*PTR1 and *Lm*PTR1 (Figure 4.3).

BC related residue interaction analysis: S95 formed vdW interactions with the RUBi016 ligand, A96, and the NADPH cofactor (Figure 4.24-A). Residue C160 formed vdW interactions with P204 and alkyl interactions with A203 (Figure 4.24-B). Residue G205 formed vdW interactions with the NADPH cofactor, A203, P204, V206 and S264 (Figure 4.24-C). Residue P210 formed vdW interactions with the ligand while S264 formed vdW interactions with G205 and V206 (Figure 4.24-D).

PCA Analysis: The *Tb*PTR1-RUBi016 complex showed the largest motions in the substrate binding loop, the α6 helix, C-terminal residues H267-A268, the modelled loop 1, and modelled loop 2 (Figure 4.25-A and Figure 4.25-B).

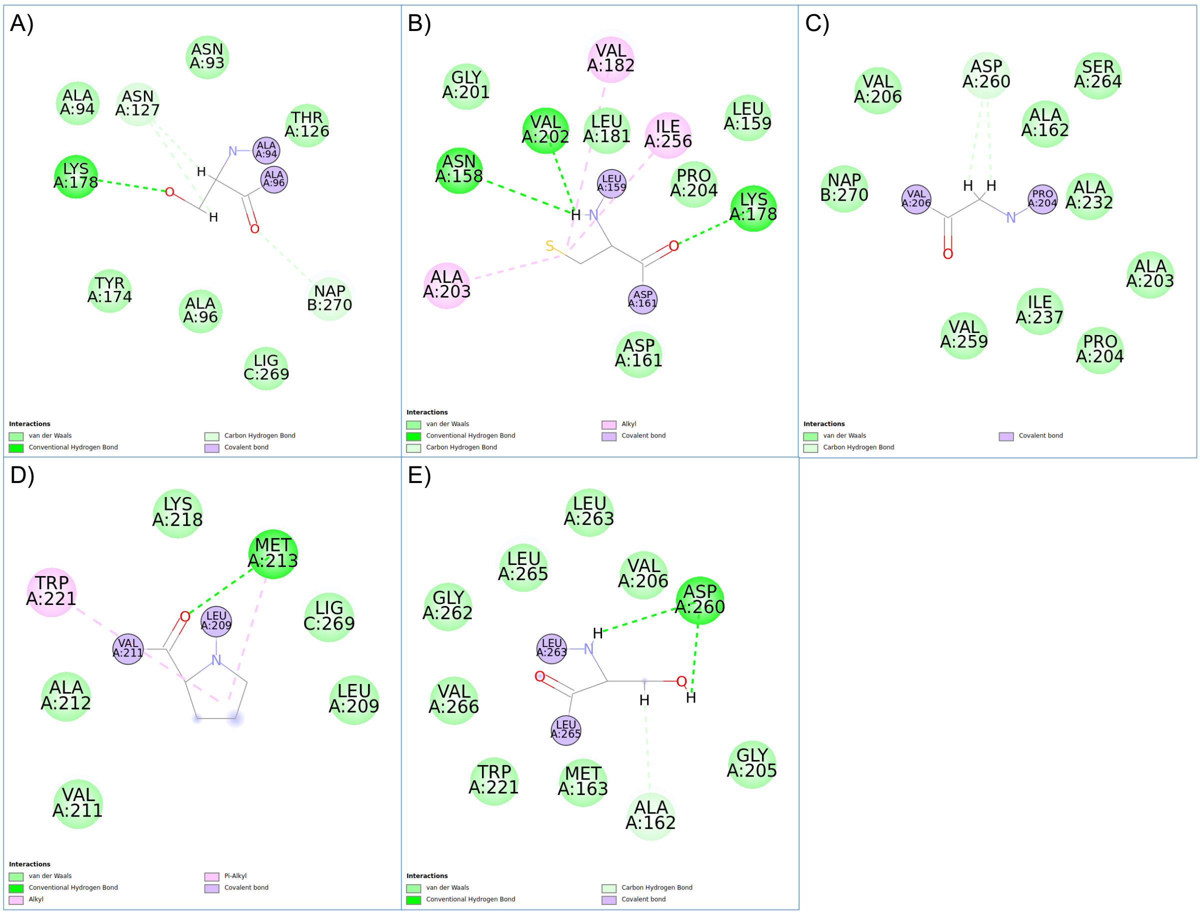


Figure 4.24: 2D interaction network around residues RUBi016-*Tb*PTR1 complex (PDB:2X9N). A) S95, B) C160, C) G205, D) P210, and E) S264

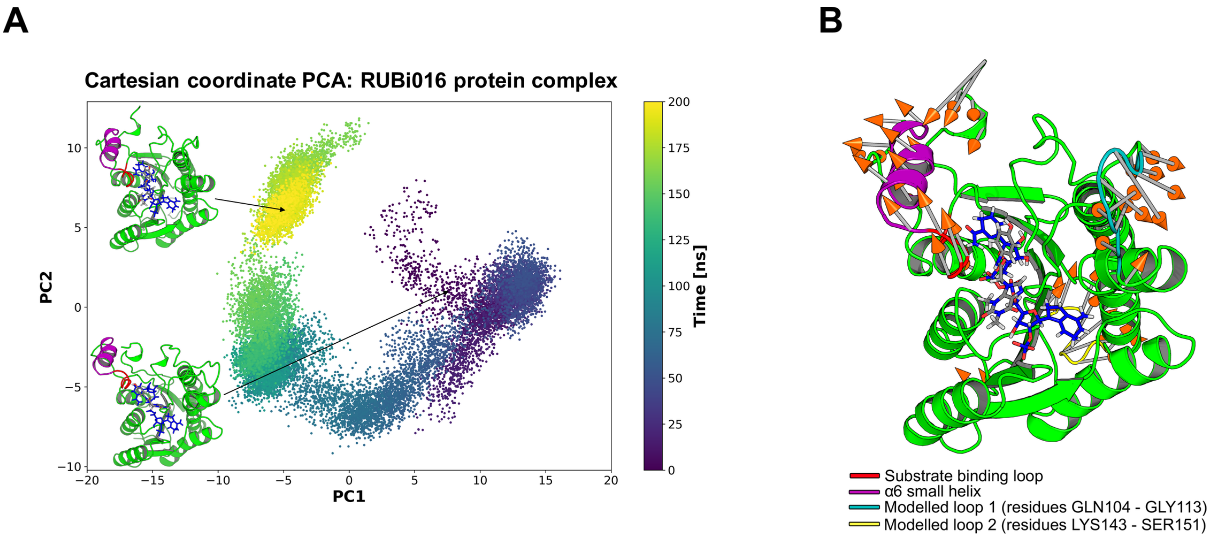


Figure 4.25: Principal component analysis of the RUBi016 - protein complex. The motions of the complex during 200ns of all-atom MD simulation shown along the first and second principal components (PC1 and PC2). The

substrate binding loop (S207-E215) is colored red, α6 alpha helix (G214 - V225) is colored magenta, modelled missing residues loop 1 colored cyan, modelled missing loop residues loop2 colored yellow, NADPH cofactor colored blue, and RUBi016 colored gray. PC1 explained 57% of the variance while PC2 explained 20%. A) Projection of the protein-ligand complex dynamics along the PC1 and PC2. B) The differential motions described by PC1 and PC2 are shown by light gray arrows with orange tips.

4.3.5.6 RUBi018

Molecular docking: Analysis of the *Tb*PTR1-RUBi016 complex showed that RUBi018 formed T-shaped π-π interactions with F97 and F171 (Figure 4.9-E). It also formed hydrophobic interactions with M163, C168, and H267 (Figure 4.9-E). In the *Tb*DHFR complex, RUBi018 formed hydrogen bonds with the NADPH cofactor and S89 along with π-π interactions with F58 in a similar binding mode to known *Tb*DHFR inhibitors [289] (Figure 4.10-E).

Molecular dynamics: The RMSD calculations showed that the *Tb*PTR1-RUBi018 complex was stable (Figure 4.26-A). There was no change in protein compactness as a result of ligand binding (Figure 4.26-B). We observed an increase in the flexibility of the substrate binding loop (P210 – M123) because of the ligand binding (Figure 4.26-D). Further, we observed decreased flexibility of F97 and M163 while P167, C168, and F171 showed increased flexibility (Figure 4.26-D).

Binding free energy: RUBi018 bound to the protein stably throughout the MD simulation with a free binding energy of -87.914 ± 12.026 kJ/mol (Table 4.4). A per residue decomposition showed that M163, V164, F171, W221, and H267 contributed -6.747 kJ/mol, -3.210 kJ/mol, -5.715 kJ/mol, -8.716 kJ/mol, and -3.560 kJ/mol respectively (Figure 4.12-B-v). D161 had an unfavorable energy contribution of 4.850 kJ/mol (Figure 4.12-B-v).

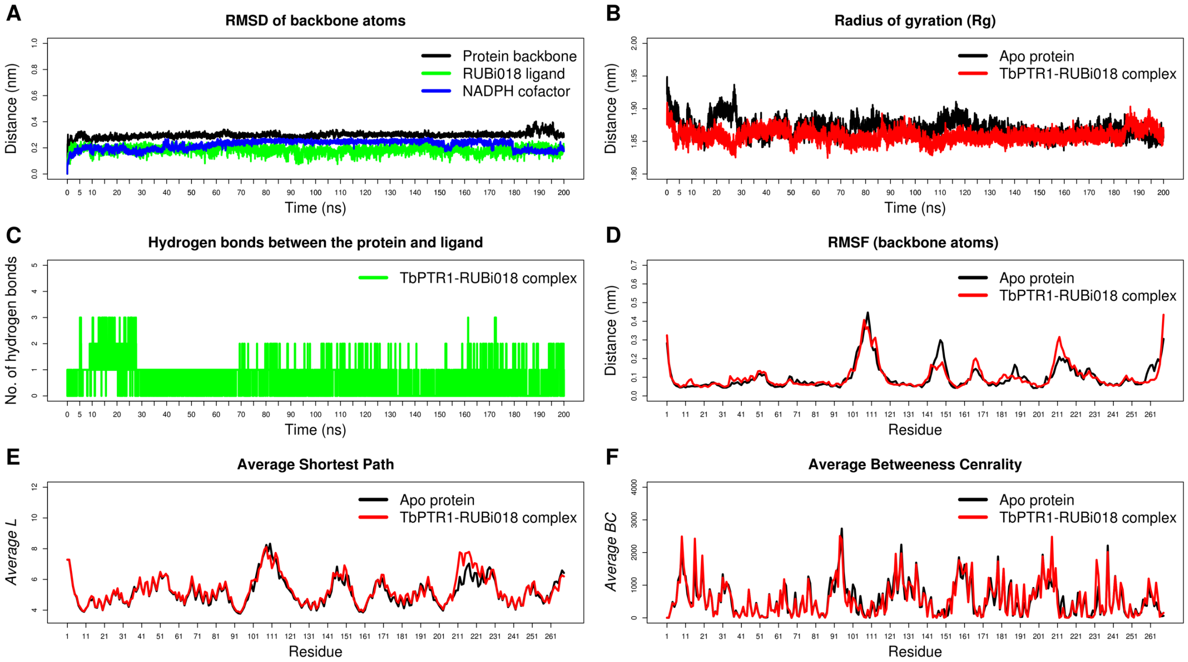


Figure 4.26: Analysis of the *Tb*PTR1-RUBi018 complex during 200 ns all-atom MD simulation at 300 K. A) RMSD of the protein back bone heavy atoms, B) Rg, C) Protein-ligand Hydrogen bond analysis, D) Per residue RMSF, E) *Average L*, and F) *Average BC*.

Dynamic residue network: Comparison of the apo protein to the *Tb*PTR1-RUBi018 complex using *L* indicated spatial displacement of the substrate binding loop and the small α6 helix (Supplementary Figure 7E). In comparison to the apo protein, V211-E215 showed increased *Average L* (Figure 4.26-E). The observed increases in *Average L* which correlated well with residue fluctuations (Figure 4.26-E and Figure 4.26-D) [209].

Residues G16, D165, V206, L208, P210, and A232 showed increases in average BC (Figure 4.26-F). Multiple sequence analysis showed that G16 was conserved among all the PTR1 orthologues. Residue V206 was only present in *Tb*PTR1 while the other PTR1 orthologues had a leucine instead. Residue D165 and A232 were conserved only among the *Trypanosoma* (Figure 4.3).

BC related residue interaction analysis: G16 formed vdW interactions with the NADPH cofactor and N 93 (Figure 4.27-A). D165 had alkyl interactions with the RUBi018 ligand and formed vdW interactions with M163 (Figure 4.27-B). Residue V206 formed vdW interactions with both the RUBi018 ligand and the NADPH cofactor (Figure 4.27-C). Residue L208 had a hydrogen bond with the NADPH cofactor and formed vdW interactions with P210 (Figure 4.27-D). Residue P210 formed vdW interactions with the RUBi018 ligand and F97 (Figure 4.27-E). Lastly, residue A232 formed a hydrogen bond with V206 and S207 (Figure 4.27-F).

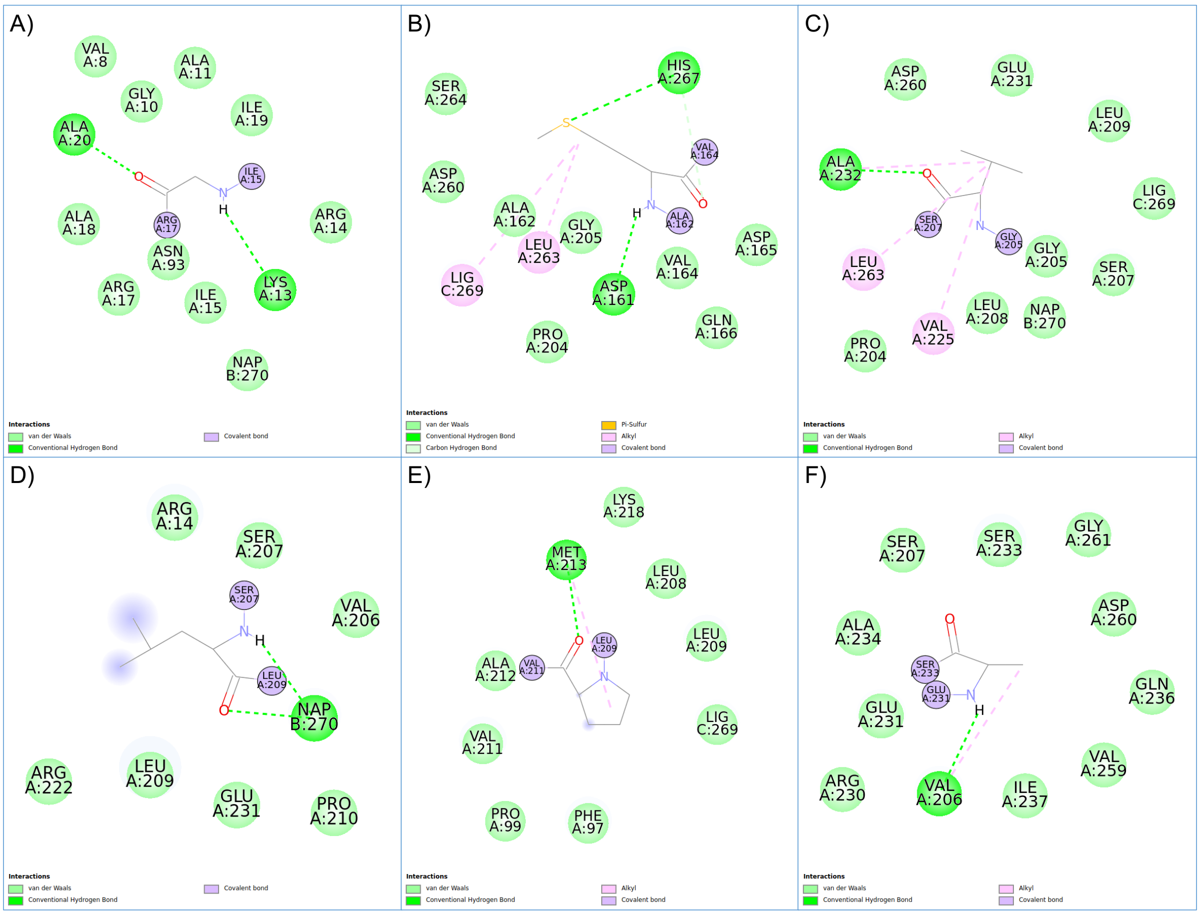


Figure 4.27: 2D interaction network around residues RUBi018-*Tb*PTR1 complex (PDB:2X9N). A) G16, B) D165, C) V206, D) L208, E) P210, and F) A232

PCA Analysis: In the TbPTR1-RUBi018 complex, the largest motions were in the substrate binding loop, the α6 helix, C160-Y174 loop region, C-terminal residues A268, the modelled loop 1, and modelled loop 2 (K143-S151) (Figure 4.28-A and Figure 4.28-B).

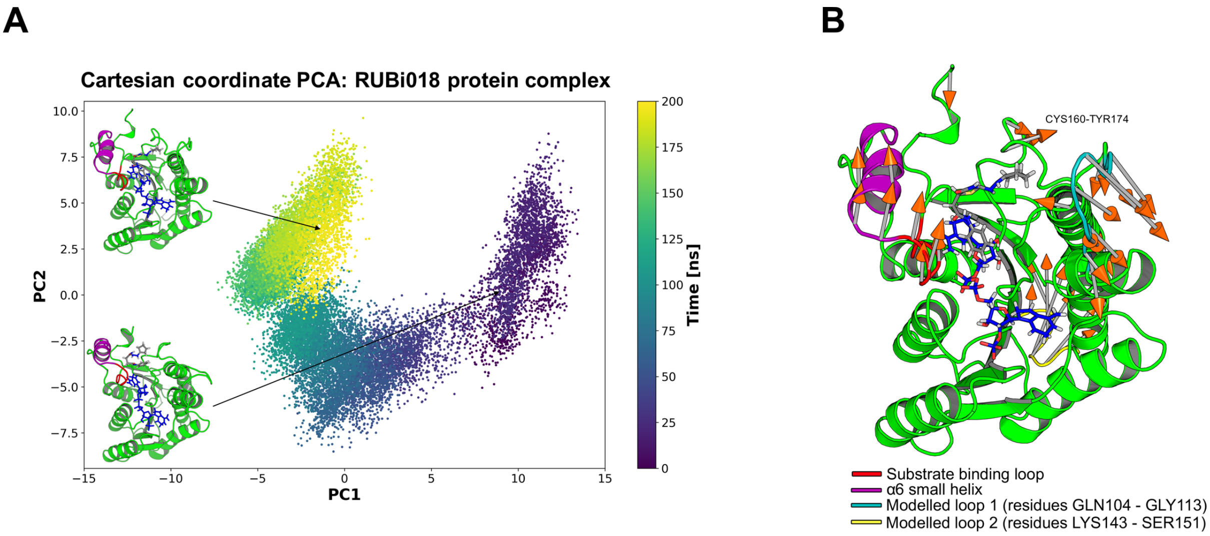


Figure 4.28: Principal component analysis of the RUBi018 - protein complex. The motions of the complex during 200ns of all-atom MD simulation shown along the first and second principal components (PC1 and PC2). The substrate binding loop (S207-E215) is colored red, α6 alpha helix (G214 - V225) is colored magenta, modelled missing residues loop 1 colored cyan, modelled missing loop residues loop2 colored yellow, NADPH cofactor colored blue, and RUBi018 colored gray. PC1 explained 44% of the variance while PC2 explained 21%. A) Projection of the protein-ligand complex dynamics along the PC1 and PC2. B) The differential motions described by PC1 and PC2 are shown by light gray arrows with orange tips.

Our *in silico* rational based drug discovery approach was able to identify five compounds that showed anti-trypanosomal *in vitro* activity. The per residue energy contribution shows the study compounds interacted with the *Tb*PTR1 protein similarly often interacting with residues known to be involved in catalysis and *Tb*PTR1 inhibitor binding [230,233,285,297] (Figure 4.12-A). We identified residues such as K13, R14, F97, M163, Y174, and substrate binding residues L209-W221 that appear to contribute favorably to binding energetics (Figure 4.12-A). With the exception of RUBi016, residue D161 gave an unfavorable energy contribution to binding in the active compounds (Figure 4.12-B). Residue D161 is involved in catalysis where it forms hydrogen bonds with M163 and Y174 and is important in proton transfer to the substrate [292]. The efficacy assessment of the compounds in this study was done using IC_{50}s based on growth inhibition of BSF *T. b. brucei* parasites *in vitro*. A more specific assay is an enzyme-inhibition assay whose results might have correlated better with the docking binding energy results.

4.4 Conclusion

In our study, structure based molecular docking was used to screen 5742 compounds. From the successful binding modes 18 were carried on for molecular docking, structure based molecular docking, molecular dynamics simulations, network analysis, and molecular mechanics Poisson–Boltzmann surface area (MMPBSA) free energy calculations. Of the 18, 13 were tested for anti-trypanosomal activity using *in vitro* inhibition assays. We report compounds RUBi004, RUBi007, RUBi014, RUBi16, and RUBi018 that exhibited anti-trypanosomal activity against trypanosomes in culture with IC_{50}s of 12.5 ± 4.8 µM, 32.4 ± 4.2 µM, 5.9 ± 1.4 µM, 28.2 ± 3.3 µM, and 9.7 ± 2.1 µM respectively. Additionally, RUBi007, RUBi016 and RUBi018 showed no significant human cell cytotoxicity at 100 µM. RUBi004 and RUBi014 had cytotoxicity IC_{50}s of 23.6 ± 5.8 µM and 32.9 ± 2.2 µM respectively (Figure 4.6). RUBi004 and RUBi014 were also shown to contain PAINS features [298]. RUBi004 contained an ene_five_het_A moiety while RUBi014 contained a catechol moiety both of which are well documented PAINS moieties [255]. When the compounds were assayed in combination with WR99210 which is a *Tb*DHFR inhibitor, compounds RUBi004, RUBi007, RUBi014 and RUBi018 displayed antagonistic effects, while RUBi016 showed an additive effect in the isobologram.

We show in this study that combined inhibition of *Tb*PTR1 and *Tb*DHFR is essential for successful targeting of the trypanosome folate pathway [217,221]. Our docking results indicate that the compounds can also bind to the *Tb*DHFR active site with good binding affinity (Figure 4.10). Our isobologram and computational experiments lead us to believe that the antagonism shown by compounds RUBi004, RUBi007, RUBi014, and RUBi018 might possibly be due to the compounds binding both *Tb*PTR1 and *Tb*DHFR but further experiments are required to validate these findings. We theorize that the compounds might be competing with the WR99210 hence the observed antagonism which would be compounded by *Tb*PTR1 over-expression that would further reduce their efficacy. Compound RUBi016 bound *Tb*DHFR with a lower binding energy than *Tb*PTR1, -31.799 kJ/mol and -37.238 kJ/mol respectively. Its additive effect when used in combination with *Tb*DHFR shows that it appears to bind selectively to *Tb*PTR1 than the other compounds. As such the addition of WR99210 further reduces folate reduction hence the observed additive effect in the isobologram. Drug combination therapy offers several advantages for one it reduces the likelihood of drug resistance. Further investigations such as ligand structure activity relationship (SAR) analysis and structure determination or characterisation of appropriate protein-ligand complexes by crystallography, are required to validate these findings.

CHAPTER 5

CHARMM Force-Field Parameterization of the Bi-metallic Catalytic Center of the *Trypanosoma brucei* Phosphodiesterase B1 (TbrPDEB1) for Computational Drug Discovery

Chapter Overview

This Chapter includes a systematic approach for the derivation of CHARMM force field parameters describing the bi-metallic catalytic center of TbrPDEB1 for computational based drug discovery.

Included in this Chapter is a brief introduction to the concepts and a review of literature concerning: the cyclic Adenosine Monophosphate (cAMP) signalling pathway in trypanosomes, Quantum Mechanics/Molecular Mechanics (QM/MM), Potential Energy Surface (PES) scans, the CHARMM Force-field, Parameterization, and CHARMM MD simulations.

As in the previous Chapter, the wet lab aspects of this work which included: *in vitro* Trypanosome growth inhibition, human cell cytotoxicity, and the Promega cAMP-Glow assays were carried out by our collaborators Prof Heinrich Hoppe and Dustin Laming of the Centre for Chemico- and Biomedicinal Research, Rhodes University, Grahamstown 6140, South Africa.

5.1 Introduction

Trypanosoma brucei phosphodiesterase B1 (TbrPDEB1) and *Trypanosoma brucei phosphodiesterase* B2 (TbrPDEB2) are isoforms that are essential for the degradation of the cyclic adenosine monophosphate (cAMP) an important intracellular second messenger [299]. Both TbrPDEB1 and TbrPDEB1 are pharmacologically and genetically validated anti-HAT drug targets [299–302]. The knock down of TbrPDEB1 and TbrPDEB2 resulted in the death of parasites *in vitro* while the RNAi parasites did not cause pathogenicity in mice [299].

In computer based drug discovery approaches MD simulations are important in understanding the protein-ligand interactions and have been widely used to identify potential lead compounds [303–305]. Accurate MD simulations are dependent on a proper description of the active site. Computational based drug discovery of TbrPDEB1 is largely hindered by a lack of force-field parameters that accurately describe the bi-metallic active center. There is need for force-field parameters that accurately describe the geometry, coordination, and dynamics of the TbrPDEB1 and TbrPDEB2 proteins.

In this study we sought to derive CHARMM force-field parameters for the simulation of TbrPDEB1 and related phosphodiesterases using a mixed bonded and non-bonded approach that can be utilized in computational based drug discovery [306–309].

5.1.1 The cAMP Signalling Pathway

Isoforms *T. brucei* phosphodiesterase B1 (TbrPDEB1; EC 3.1.4.53) and *T. brucei* phosphodiesterase B2 (TbrPDEB2; EC 3.1.4.53) are genetically and pharmacologically validated HAT chemotherapeutic targets [299–302]. The two phosphodiesterases are essential for the degradation of cAMP, which is an intracellular signaling molecule which is important in communication with the extracellular environment [299]. Simultaneous knock down of TbrPDEB1 and TbrPDEB2 resulted in parasite cell death *in vitro* and mutant parasites that were avirulent in mice [299]. Further, intracellular cAMP levels have been shown to regulate social motility behavior of *T. brucei* parasites [310]. Altogether this makes them ideal drug targets for chemotherapeutic intervention.

5.1.2 Quantum Mechanics (QM) and Molecular Mechanics (MM)

In the computation of the energy of a molecule there are two approaches that are commonly used. The first is QM which is a first principles approach that is used to describe molecules at the level of atoms and sub-atomic particles [311,312]. The nuclei are arranged in space with their corresponding electrons spread in the system using a continuous electron density computed using the Schrödinger equation [311,312]. Density Functional Theory (DFT) is the most commonly used semi-empirical computational QM modelling method in chemistry [313]. In the second approach, MM, classical mechanics are used to describe the behaviour of molecular systems. In comparison to QM, MM ignores electron properties and rather computes the potential energy surface of a system using potential functions described in classical physics [314]. These functions are known as a force-field [314,315]. In most widely used MM models there are a few assumptions made:

a) The electrons arounds around a nucleus and the nucleus are treated as perfectly spherical.

b) The bonds between molecules are treated as springs.

c) Force constants and equilibrium parameters are determined experimentally.

d) The potential energy is a sum of the bond stretching, angle bending, dihedral angle bending, and non-bonded interactions.

As a result, MM is less computationally expensive than QM and can be used to calculate the energy of large molecular systems like proteins [314].

5.1.3 Parameterisation of the TbrPDEB1 Bi-metallic Catalytic Center

5.1.3.1 CHARMM Force Field

The CHARMM energy function was used to describe the internuclear and intramolecular interactions and is defined by the equation shown below [316,317]:

$$U(\vec{R}) = \sum_{bonds} K_b\ (b - b_0)^2 + \sum_{angles} K_\theta\ (\theta - \theta_0)^2 + \sum_{dihedrals} K_\varphi\ (1 + \cos(n\varphi - \delta))$$
$$+ \sum_{impropers} K_\omega\ (\omega - \omega_0)^2 + \sum_{Urey-Bradley} K_{UB}(S - S_o)^2$$
$$+ \sum_{non-bonded} \left\{ \varepsilon_{ij} \left[\left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\varepsilon_1 r_{ij}} \right\}$$

Equation 5.1: The CHARMM energy function

The first term in this function describes bond stretching where $b - b_0$ is the distance moved from the equilibrium by an atom and K_b the bond force constant (Figure 5.1-A) [316,317]. Angle bending is described similarly where $\theta - \theta_0$ is the angle from the equilibrium between three atoms with an angle force constant of K_θ (Figure 5.1-B). Dihedral bending is described in terms of a cyclic cosine function, where $K\varphi$ is the torsional angle constant, n is the multiplicity and φ is the torsion angle and δ is the phase shift (Figure 5.1-C) [316,317]. The impropers term, deals with out of plane bending where K_ω is the force constant for a ω - ω_0 out of plane angle (Figure 5.1-D) [316,317]. For three bonded atoms a-b-c the Urey-Bradley component was described by K_{UB} which was the respective force constant for a distance S-S_0 between atoms a and c in the harmonic potential [316,317]. Non-bonded terms between atom pairs (*i,j*) are described by a combination of the van der Waals energy (calculated using standard 12-6 Lennard-Jones potential) and the electrostatic energy calculated using a Coulombic potential [316,317]. These non-bonded terms are applied only to atoms that were separated by at least three bond terms [316,317]. The CHARMM36 force-field now also includes an energy correction map (CMAP) term that offers a better description of peptide backbone dihedrals [156,318].

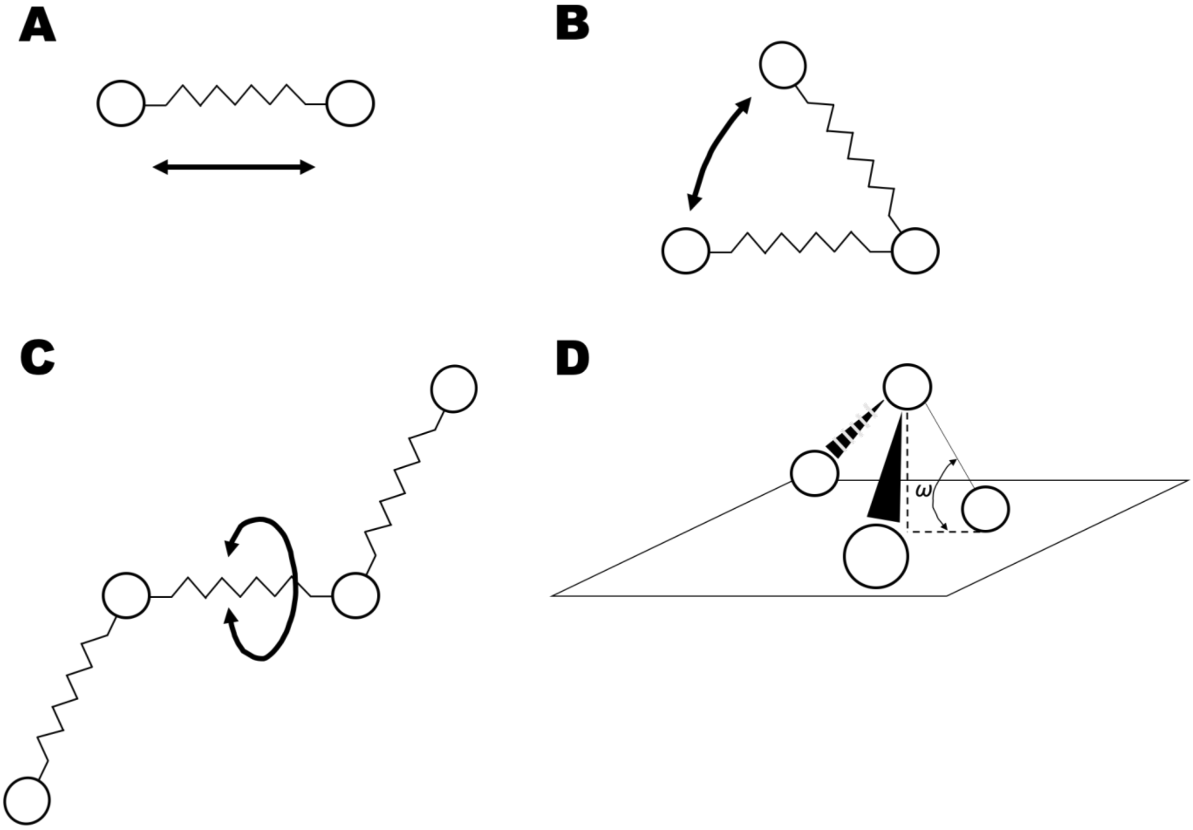


Figure 5.1: The bonded terms of the CHARMM potential energy function. A) The stretching and contraction of a bond, B) The bending of a three bonded atoms, C) The twisting around a central bond between four bonded atoms, and D) Impropers also known as out of plane bending.

5.1.3.2 Force-Field Parameterization

Force fields are generally accurate, transferable, and robust however there isn't a force-field that is generalizable for every protein system. This especially true for metal ion containing proteins which constitute about 40% of the proteins in the PDB. Metal ions play important roles protein function particularly in catalysis, structural properties, and electron transfer [319,320]. In order to accurately reproduce the geometry, conformation and dynamics of a protein active site parameters describing the atomic interactions are required. The parameters are calculated and optimised to reproduce experimental active site geometries (as close as possible to the crystal structure). The parameters typically calculated include bonded terms which include: bond stretching, angle bending, torsions and non-bonded terms such as van der Waals interactions and electrostatics. In our study we mainly focused on the bonded terms which were bond stretching, angle bending, and dihedral bending. The non-bonded interactions were described by the 12-6 Lennard-Jones (LJ) potential function which is accurate and transferable [321–324].

There are three main approaches when parameterizing metals in proteins and these are:

a) The bonded model, where the interactions between the metal ions and coordinating amino acid residues are treated as part of a bonded system. This involves describing bonds, angles, dihedrals, and non-bonded terms (electrostatics and vdW terms). The a parameters are obtained by QM optimisations and Potential Energy Surface (PES) scans [306–309,325].
b) The non-bonded model, where the interactions between the metal and amino acid residues is handled by the electrostatic and vdW terms only [306–309,322,325]. It's a much easier approach but the forces are often insufficient to maintain the metal coordination geometries.
c) The cationic dummy atom model, where charges that are placed around the metal ions are used to mimic covalent bonds [326].

In this study, we opted to use a mixed bonded and non-bonded approach to describe the bi-nuclear metal center during dynamics [306–309]. This is because fully bonded models do not allow for the simulation of bond formation, dissociation, and distortions that occur as a result of transitions between common coordination geometries (for example the Zn^{2+} ion can range from 4 to 6) [306–309,325,327].

5.1.4 Bi-metallic Active Sites

Bi-metallic motifs are common in nature and present in several important enzyme families such as phosphodiesterases, hydrolases, peptidases to mention but a few. This is because they offer a broad range of physicochemical properties, for example: Lewis acidity, positive charge, flexible and/or rigid coordination spheres, activating nucleophiles, stabilizing intermediate/transition states, and ligand affinity [328].

5.1.5 Research Motivation

This Chapter falls under the broad theme of Part II that was to discover novel anti-trypanosomal chemotherapeutics. The aim of this Chapter was to derive CHARMM force-field parameters that can be used to accurately reproduce the geometry, conformation, and the dynamics of the TbrPDEB1 bi-metallic catalytic active site center. The derived TbrPDEB1 force-field parameters were then utilized to characterize molecular protein-ligand interactions that can be used in computational drug discovery.

5.2 Materials and Methods

5.2.1 MSA

MSA was performed using MUSCLE on the PDE orthologue sequences which included TbrPDEB1 (Uniprot: Q8WQX9), *T. cruzi* (TcPDEC1; Uniprot: Q53I60), *L. major* (LmPDEB1; Uniprot: Q6S996), and *H. sapiens* (HsPDE4; Uniprot: Q07343).

5.2.2 CHARMM Potential Energy Function

The CHARMM energy function was used to describe the internuclear and intramolecular interactions and is defined by Eq.(5.1) [316,317].

5.2.3 Two layer ONIOM QM/MM Optimization of the Structure

For optimization of the structure, a two-layer QM/MM ONIOM method was used which enabled the application of two different levels of theory to the protein system; a higher level of theory was used to treat the region of the active site containing both metal cations, while a lower level molecular mechanics description could be applied to the rest of the protein [329]. This had the advantage of reproducing reliable geometries and thermodynamics of the active site with reasonable computational cost [329]. The ONIOM optimization was effected with the use of the Gaussian 09 software package using the default convergence settings [330]. The atom coordinates used as input for this optimization were from a single chain from the catalytic domain homodimer PDB:4i15 and the three ligands Zn^{2+}, Mg^{2+}, and OH^-, absent of protein substrate or inhibitor [300].

The protein system was protonated at pH 7.4 (physiological pH using the H++ webserver, http://biophysics.cs.vt.edu/). H673 and H709 that are part of the Zn^{2+} ion coordination sphere were both unprotonated at the N_ε (NE2) (HSD CHARMM residue type) (Figure 5.2). The hybrid density functional B3LYP/6-31G(d) [331] model was used to treat the high-level region (bi-metallic active site) which comprised a Zn^{2+}, Mg^{2+}, OH^-, HSD673, HSD709, D710, D822, and four water molecules identified in the crystal structure (Figure 5.2). The UFF [332] model was applied to the low-level region. The high-level region was composed of 68 atoms while the low-level region comprised 5174 atoms. The QM/MM boundary was treated using the pseudobond approach [333] and the 6-31G(d) basis set was used because it reproduced the experimentally observed metal-ligand distances with minimal experimental errors [328,334,335]. The QM/MM optimization calculations were carried out on a single computational node using 24 cores (Intel Xeon) on the CSIR Center for High Performance Computing (CHPC) cluster.

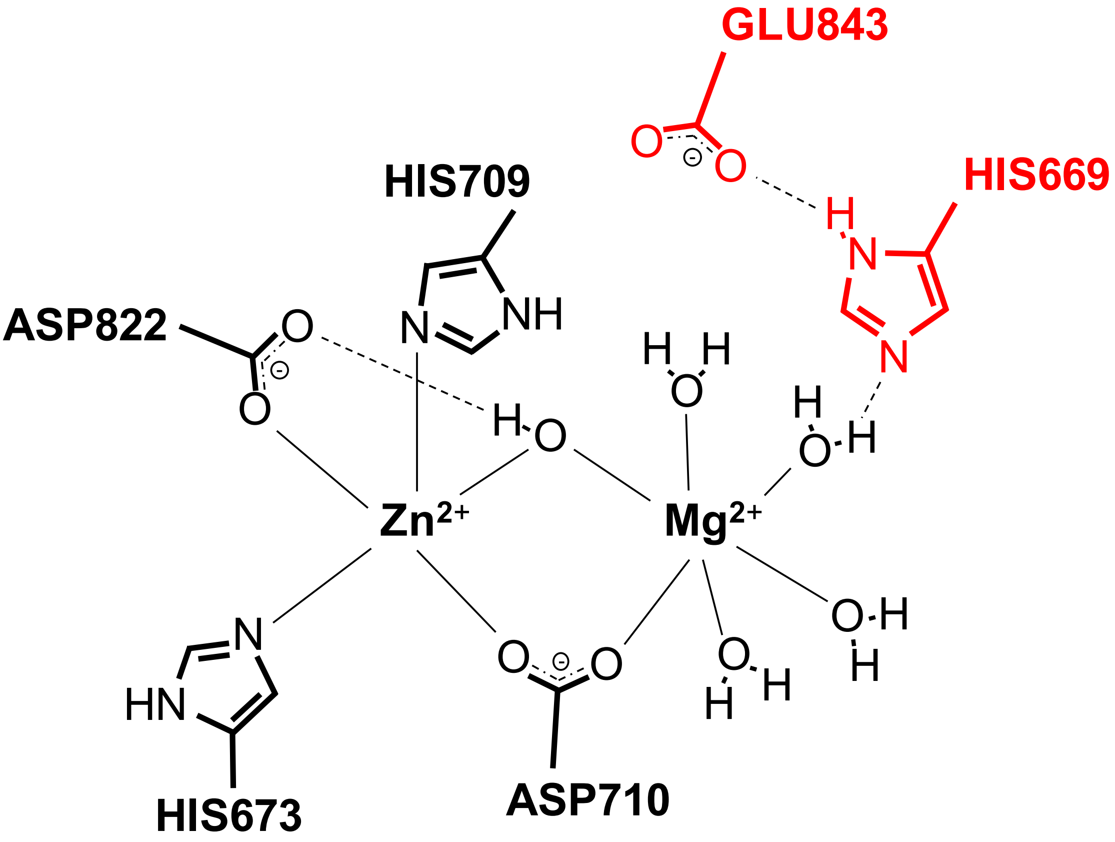


Figure 5.2: A schematic diagram of the bi-metallic active site. The metal coordinating amino acid residues H673, H709, D710, D822 are colored in black. Also shown but not included in the QM high layer are the catalytic residues H669 and E843 that are colored red.

5.2.4 Bond, Angle, and Dihedral Parameters

Bond, angle and dihedral equilibria and force constants were determined using Potential Energy Surface (PES) scans. The starting position for the PES scans was the structure from the two-layer QM/MM ONIOM optimisation. The PES scans were carried out using a step size of 0.005 Å for the bonds, 0.5 degrees for the angles, and 0.05 degrees for the dihedral angles. The PES scans were for 20 steps on either side of the minima point. The force constants were determined by best fitting of the harmonic potentials from the PES scans to the CHARMM energy potential. For dihedrals it is possible to use any value for the phase, CHARMM however recommends that values of 0 and 180° be used as the parameters as they are appropriate for different stereoisomers [336]. The PES scans were carried out using 24 cores (Intel Xeon) on the Center for High Performance Computing (CHPC) cluster.

5.2.5 Van der Waals Parameters and Atomic Single Charges

The van der Waals parameters for the $Zn^{2+,}$ Mg^{2+} and OH^- ions were used directly from the CHARMM 36 force field water and ions parameter file [337]. These values were as follows: Mg^{2+}: $\delta = 0.015$ kcal/mol and $\varepsilon = 1.185$ Å [338]; Zn^{2+}: $\delta = 0.250$ kcal/mol and $\varepsilon = 1.090$ Å [339]; and the OH^- ion oxygen: $\delta = 0.120$ kcal/mol and $\varepsilon = 1.700$ Å, with the OH^- hydrogen the same as TIP3 hydrogen [337]. The formal charges for Zn and Mg were +2 and -1 for the OH^- ion.

5.2.6 Determining the Most Optimal Configuration of the Calculated Parameters for MD Simulations

In this study we used a mixed bonded and non-bonded approach; the Zn^{2+} ion, Mg^{2+} ion, bridging OH^- ion, HSD673, HSD709, D710, and D822 were treated as part of a bonded system while the waters coordinated to the Mg^{2+} were treated as non-bonded. Further, in order to determine the most optimal configuration of the parameters to best describe the TbrPDEB1 catalytic center we set up four models:

a) **Model 1**: The Zn^{2+} ion, Mg^{2+} ion, bridging OH^- ion, HSD673, HSD709, D710, and D822 were treated as a part of a bonded system where bonds, angles and dihedral parameters were described. The water molecules in the the Mg^{2+} ion coordination sphere were handled by the CHARMM force-field LJ parameters and not bonded. The HSD673 and HSD709 were unprotonated at the NE2 Nitrogen while H669 (HSE669) was protonated at the NE2. All the other amino acid residues were protonated according to their standard protonation states at pH 7.4.

b) **Model 2**: Model 2 differed from model 1 in one aspect, no bonding parameters were included to bind residue D710 to either of the metals, so both metals were independent of D710 (with the exception of non-bonded parameters).

c) **Model 3**: Model 3 was similar to model 2 except in the way residue H669 was protonated. HSD669 was unprotonated at the NE2 atom.

d) **Model 4**: Model 4 was similar to model 1 except that residue HSD669 was unprotonated at the NE2 atom.

5.2.7 Parameter Validation using Molecular Dynamics Simulations

The derived parameters were validated using CHARMM MD simulations. The parameters were patched into the CHARMM36 protein force field for MD simulation and the four separate MD simulations and each of the four models assessed [337]. The TbrPDEB1 single chain protein (PDB: 4i15, Resolution: 1.65 Å) [300] was used as starting coordinates for the MD simulation of the apo protein and later docked protein-ligand systems. MD simulations were carried out using the CHARMM simulation program (version 35b2) [156,336,340]. Where ligands docked to the protein were included in the simulations, the ligands were parametrized using the CHARMM General Force Field (CGenFF) [336].

The MD systems were solvated with water (using the TIP3 water model), neutralized with Cl^- and Na^+ ions, run at a temperature of 310 K, and pressure of 1 atmosphere. Periodic boundary conditions were applied along with the isothermal-isobaric ensemble (NPT and NVT) along with a smooth particle mesh Ewald method to cater for long range electrostatic interactions. Further, a group-based cut-off of 12 Å was used to treat non-bonded interactions. All covalent bonds involving hydrogen are constrained with the SHAKE algorithm. A time step of 2 fs was used for numerical integration of the Newtonian equations using a Leapfrog Verlet algorithm for a total of 2500000 steps with frames saved every 500 steps. The energy bail-out was 100 kcal/mol. The MD simulations were carried out using 64 cores on the CHPC cluster.

5.2.8 Molecular Dynamics Analyses

The MD trajectories were analysed with respect to conformation and stability using RMSD, Rg, RMSF, metal coordination bond distance, angle, and dihedral angle measurements, DRN, PCA, and EDA. Protein-ligand interactions were analysed using hydrogen bonding analysis and protein-ligand interaction energy [336]. DRN analysis was carried out using MD-TASK with a 6.7 Å cut-off used when constructing the protein network [132]. Additionally, for the weighted residue contact map we included the Zn^{2+} ion, the Mg^{2+} ion, O atoms of the TIP3 water and OH^- ion, and atoms from the docked ligands [132].

Potential energy calculations, angle bending, bond distances, dihedral bending, interaction energies measurements were carried out using CHARMM and Visual Molecular Dynamics (VMD) [341]. Monitoring of the planarity of the H669 histidine-ring was done by measuring the CA-CB-CG-ND1 dihedral angle relative to its starting point. Lastly, we used the following visualization and analysis programs to generate the figures: VMD [341], Rstudio [342], Matplotlib [184], PyMoL [180], Discovery studio [287], Ligplot [343], and GaussView [344].

5.2.9 Preparation of Protein-Ligand Complexes

5.2.9.1 Molecular Docking

The molecular docking investigations were carried out using the orthologue PDEs from *T. brucei*, *Leishmania major*, *Trypanosoma cruzi* and a human host PDE4B. High quality protein structures were retrieved from PDB and are shown in Table 5.1 below:

Table 5.1: PDE orthologue structural comparison

Enzyme	**ORGANISM**	**UNIPROT**	**PDB**	**Resolution (Å)**	**Chains**	**Residues**	**RMS (ref: 4I15)**
TbrPDEB1	*T. brucei*	Q8WQX9	4I15 [300]	1.65	A,B	343	1.000
TcPDEC1	*T. cruzi*	Q53I60	3V94 [345]	2.00	A - H	345	0.161
LmjPDEB1	*L. major*	Q6S996	2R8Q [346]	1.50	A,B	359	0.407
HsPDE4B	*H. sapiens*	Q07343	1XM4 [347]	2.31	A	398	0.395

The ligand library was prepared using 635 compounds from SANCDB and 5742 compounds from the ZINC Drugs Now subset [282,283]. The pre-filtering strategy was as described in section 4.2.2 in Chapter 4. The Gasteiger charges for the metals and OH^- were assigned using the calculated Restrained Electro Static Potential (RESP) charges. Molecular docking was carried using Autodock Vina (version 7.4) and the docking parameters are summarized in Table 5.2 below:

Table 5.2: Autodock Vina blind docking parameters

Parameter	***T. brucei*** **(PDB: 4i15)**	***T. cruzi*** **(PDB: 3V93)**	***L. major*** **(PDB: 2R8Q)**	***H. sapiens*** **(PDB: 1XM4)**
Protein center X	27.59	-24.08	14.11	67.20
Protein center Y	17.69	0.03	-17.67	2.59
Protein center Z	26.65	-48.49	17.64	52.15
Box size X	126	126	126	126
Box size Y	126	126	126	126
Box size Z	126	126	126	126
Energy range	4	4	4	4
Exhaustiveness	120	120	120	120
CPU	24	24	24	24

In the blind docking, TbrPDEB1 and LmPDEB1 structures were homodimers, the TcPDEC1 was a homo-octamer, and the HsPDE4B a monomer. Protein-ligand interactions were analysed on the basis of hydrogen bonding, interaction with the metal ions, and Autodock Vina binding free energy and binding modes.

5.2.9.2 BBB Permeability Prediction

Additionally, BBB permeability prediction was done using a webserver available at www.cbligand.org/BBB [348]. The BBB prediction webserver located at www.cbligand.org/BBB relies on Support Vector Machine (SVM) machine learning and Ligand Classifier of Adaptively Boosting Ensemble Decision Stumps (LiCABEDS) that is based on four types of fingerprints of 1593 reported CNS permeable compounds [349–351].

5.2.10 *In vitro* Inhibition Assays and Promega cAMP-Glow Assay

The *in vitro* inhibition assays and cytotoxicity were carried out by Prof Heinrich Hoppe and Dustin Laming as described in section 4.2.8.1 in Chapter 4. Additionally, a Promega cAMP-Glow assay was carried out to measure the effect of the ligands on phosphodiesterase activity *in vitro*. Successful inhibition of TbrPDEB1 and TbrPDEB2 would be shown by impaired cAMP degradation in the treated trypanosomes. PDE activity was measured in 96-well plates following the manufacturer instructions (Promega, Madison, WI, USA).

5.3 Results and Discussion

5.3.1 TbrPDEB1 Overview

In this study we used a high quality unliganded TbrPDEB1 crystal structure (PDB:4i15; Resolution: 1.65 Å) that is a homodimer of the catalytic domain (V586 – R918) [300] (Figure 5.3). The catalytic domain is composed of 16 α-helices and 3_{10} helices while the active site includes two metal cations (*viz* Zn^{2+} and Mg^{2+}) and a hydroxide ion per chain [300] (Figure 5.3-B).

The TbrPDEB1 and LmjPDEB1 orthologues were the most similar and showed high sequence identity (Table 5.3). There is also a parasite-specific cavity called the 'P-pocket' that is made up of residues: A837, T841, Y845, N867, M868, E869, and L870 [300,301]. Targeting of the P-pocket has been shown to improve drug selectivity [301].

Table 5.3: Sequence comparison of the TbrPDEB1 orthologues

Enzyme	**Organism**	**UNIPROT**	**Coverage (%)**	**Percentage sequence identity (ref *T. brucei*; %)**
TbrPDEB1	*T. brucei*	Q8WQX9	100	100
TcPDEC1	*T. cruzi*	Q53I60	90	15
LmjPDEB1	*L. major*	Q6S996	100	66
HsPDE4B	*H. sapiens*	Q07343	92	24

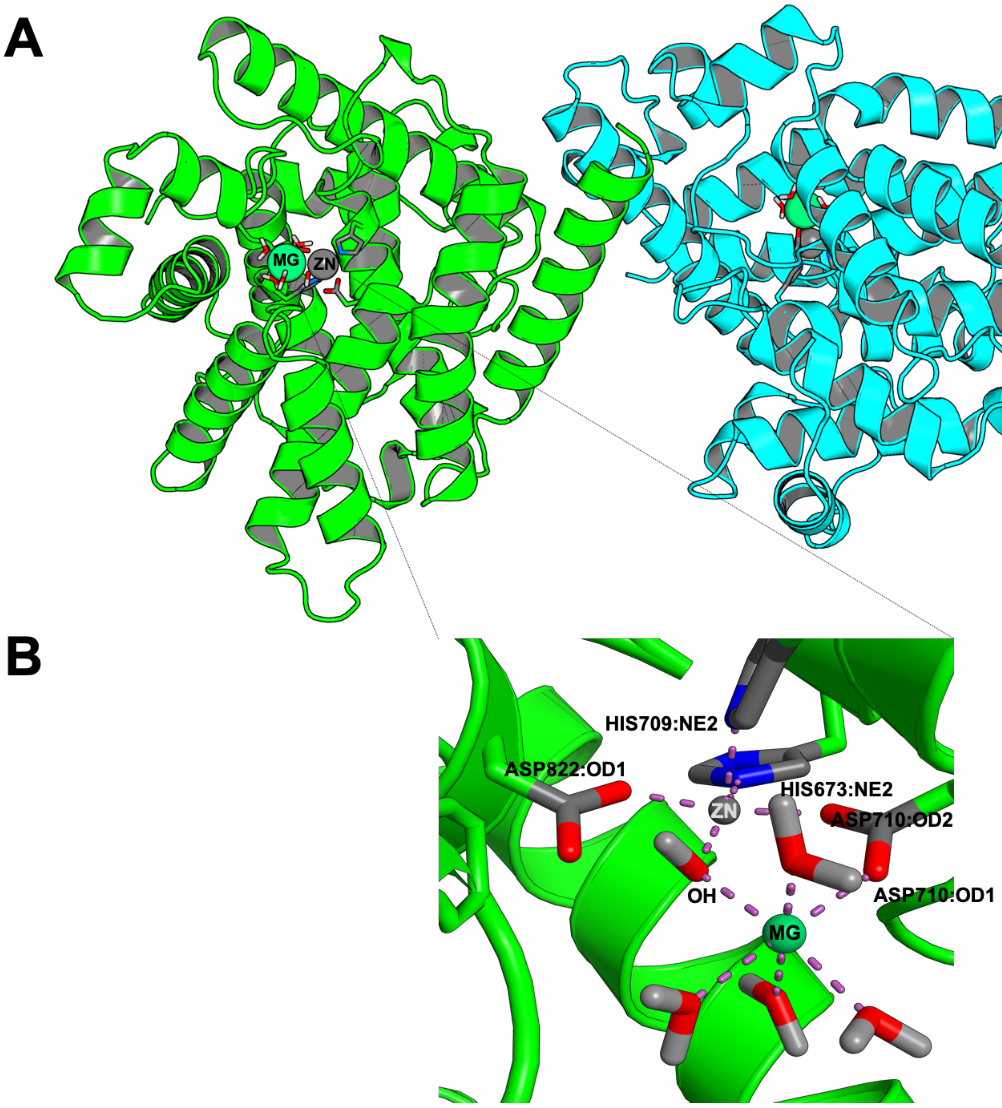


Figure 5.3: A cartoon representation of the TbrPDEB1 homodimer (PDB:4i15) and its bi-metal active site. A) The protein is colored by chain, the Zn^{2+} and Mg^{2+} cations are rendered as spheres. B) The TbrPDEB1 bi-metal active site with the coordinating residues rendered as sticks.

The PDE orthologues show a highly conserved bi-metal catalytic center within the catalytic domain (Figure 5.4). The metal-coordinating residues were highly conserved along with the PDE characteristic [HDxxHxxxxN] motif (Figure 5.4) [352–354]. PDE superfamily enzymes show similar catalytic hydrolysis of cAMP as shown by the highly conserved catalytic site [355] (Figure 5.4).

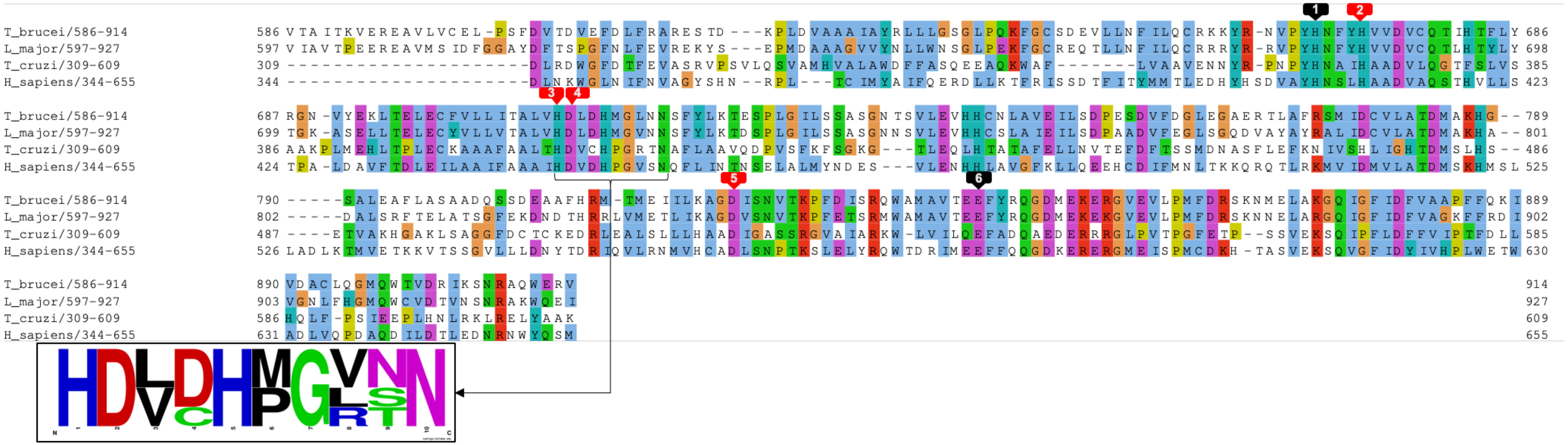


Figure 5.4: A multiple sequence alignment of *T. brucei* (TbrPDEB1), *T. cruzi* (TcPDEC1), *L. major* (LmPDEB1) and *H. sapiens* (HsPDE4B) PDE orthologues. Labels within the MSA: 1, 2, 3, 4, 5, and 6 indicate the highly conserved H669, H673, H709, D710, D822, and E843 residues respectively that are responsible for bi-metal coordination within the active site (red) and catalysis (black). A sequence logo shows the highly conserved PDE [HDxxHxxxxN] motif.

Magnesium and zinc containing active site centers are a common and functionally important biological motif [320,355–358]. The Mg^{2+}/Zn^{2+} bimetallic active center offer several physicochemical properties such as flexible or rigid coordination, positive charge, Lewis acidity, stabilizing intermediate/transition states, activating nucleophiles and ligand affinity [328].

5.3.2 PDE Hydrolysis

The PDE superfamily shows high conservation of the catalytic center and cAMP hydrolysis and because of this we can use the human PDE4 hydrolysis of cAMP to gain insights into TbPDEB1 catalysis [355]. The superscript on the PDE4 residue indicates the equivalent residue in TbrPDEB1 unless indicated otherwise (so, for example, $H234^{H669}$ indicates that discussion is about residue H234 on PDE4 or the corresponding H669 residue on TbrPDEB1).

The first step of the catalytic reaction involves the NE2 (N_ε) atom on $H234^{H669}$ extracting a proton from one of the water molecules in the Mg^{2+} ion coordination sphere yielding a OH^- ion [355,359]. The OH^- ion then serves as a bridging ligand between the two metal ions [355]. Concomitantly, residue $E413^{E843}$ extracts a proton from the ND1 (N_δ) atom of residue $H234^{H669}$ resulting in a protonated $H234^{H669}$ [355,359]. This is followed by the substrate cAMP entering the active site and binding to the two metals where the nucleophilic OH^- ion attacks the phosphorous atom and breaks the phosphodiester bond (P — O3) in cAMP causing the ring to open [302,355,359]. This then yields a bound hydrolyzed product which then dissociates from the two metals and the coordinating waters are restored [355,359]. The two metals appear to serve as electrophiles that stabilize the nucleophile and the transition state as well as serving a structural role. In the TbrPDEB1 crystal structure (PDB:4i15) residue H669 is seen to form π-π stacks with H673, Y668, and H713. H669 also forms conventional hydrogen bonds with E843, a carbon hydrogen bond with D710, and water hydrogen bonds with two water molecules that are in the Mg^{2+} coordination sphere (Figure 5.2). The bi-metal coordinating amino residues i.e. H673, H709, D710, and D822 are highly conserved among the PDE orthologues (Figure 5.3). Catalytic residues H669 and E843 are also highly conserved among the PDE orthologues (Figure 5.3).

5.3.3 Two-layer QM/MM ONIOM Optimization

METALizer, a web-based server that predicts the coordination geometry of metals in metalloproteins (https://proteins.plus/) predicted that the Zn^{2+} and Mg^{2+} ion both had octahedral geometries. Comparison of the active site geometry of the crystal structure and the ONIOM optimised structure showed that the two-layer QM/MM ONIOM optimization was able to reproduce the coordination sphere of the Mg^{2+} ion, however it gave a distorted trigonal bipyramidal geometry for the Zn^{2+} ion (Figure 5.5). The Zn^{2+} ligand interactions with its coordinating amino acid atoms and OH^- ligand were reproduced with the exception of a water molecule that was not energetically favourable. From the ONIOM optimisation the Zn^{2+} ion had a distorted trigonal bipyramidal geometry and was coordinated by an O atom from D822 (OD1), two N atoms from H673 (NE2) and H709 (NE2), and the OH^- ion. The Mg^{2+} ion had an octahedral geometry and was coordinated by an O atom from the OH^- ion, an O atom from D710 (OD2) and four O atoms from the four water molecules identified in the crystal structure. The two metals were bridged by a hydroxide ion (OH^-) and residue D710.

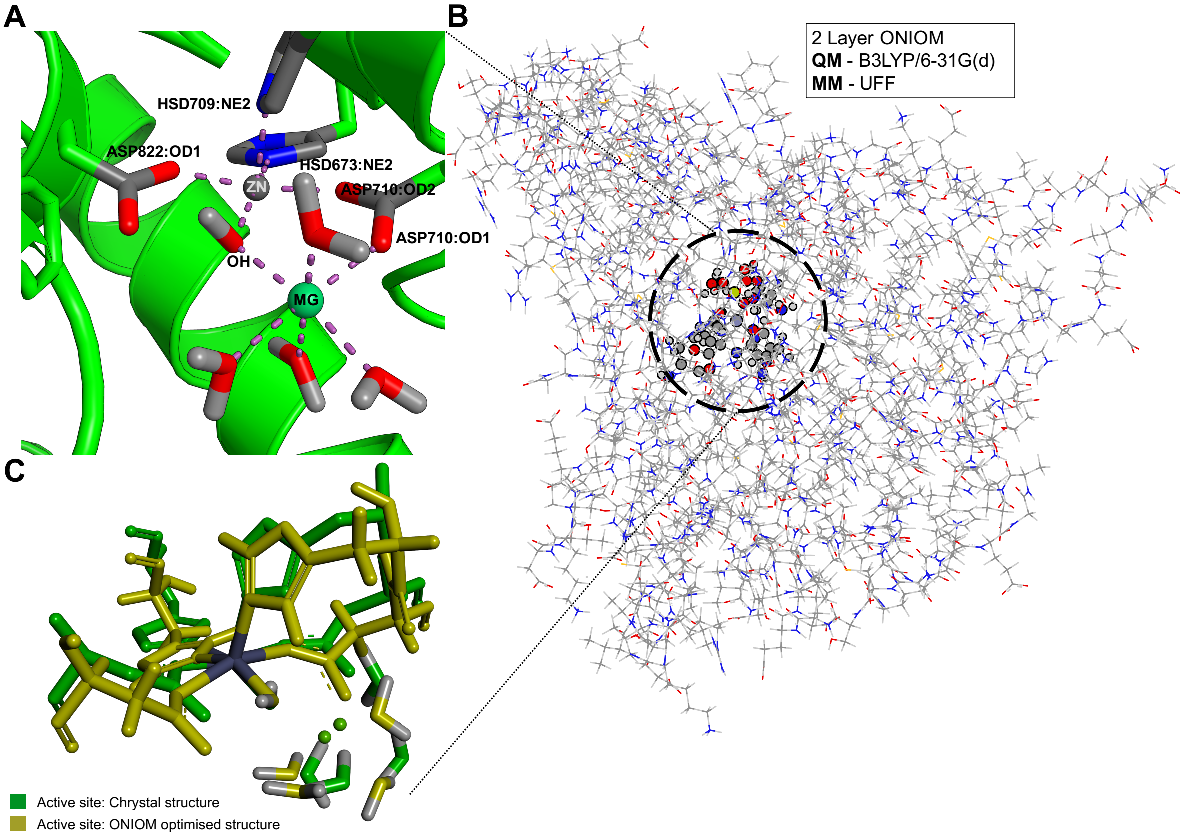


Figure 5.5: A cartoon representation of the two-layer QM/MM ONIOM optimization. The Zn^{2+} is colored gray while the Mg^{2+} is colored green and rendered as a sphere. A) The crystal structure bi-metal active center, B) The two-layer QM/MM ONIOM, and C) Overlay of the crystal (green) and two-layer QM/MM optimized (yellow) structures.

The metal internuclear distance in the ONIOM optimized structure deviated by 0.233 Å from the initial crystal structure (Figure 5.5 and Table 5.4). Previous studies report variation in the Zn^{2+} and Mg^{2+} internuclear distance among PDEs during cAMP hydrolysis [359–362]. It is proposed to be a key contributing factor in transition state stabilization during reaction [359–362]. Analysis of several TbrPDEB1 crystal structures showed variation in the metal internuclear distance between the different structures (Table 5.4) [300,301]. OX denotes the O and HX denotes the H atom in the OH- ion. The Zn^{2+}-OX and Mg-OX bond distance deviated by 0.225 Å and 0.257 Å respectively while the Zn^{2+}-OX-Mg angle showed a deviation of 12 degrees (Table 1). With these considerations, the ONIOM demonstrated good modelling of the TbrPDEB1 bi-metallic active site.

Table 5.4: : Comparison of experimental crystal structure and QM/MM ONIOM geometries

Atoms	b_0 **(Å)**	**QM/MM ONIOM optimised bond length (Å)**	**Experimental bond length (Å)**									
			4i15 (ref)	4PHL	5G2B	5G5V	5G57	5L8C	5L8Y	5L9H	Average bond length (Å)	SD (Å)
ZN - MG	3.584	3.541	3.774	3.995	3.809	3.725	3.825	3.769	3.938	3.855	3.836	± 0.091
ZN - OX	1.927	1.924	2.149	2.092	2.262	2.104	2.158	2.278	2.179	2.112	2.167	± 0.070
ZN - OC	2.176	2.247	2.068	2.138	2.115	2.122	2.105	2.054	2.172	2.188	2.120	± 0.046
ZN - NR2	2.075	2.083	2.196	2.078	2.261	2.302	2.208	2.273	2.222	2.310	2.231	± 0.075
MG - OX	1.941	1.943	2.200	2.174	2.018	2.099	2.121	1.777	1.992	2.068	2.056	± 0.133
MG - OC	2.120	2.040	2.144	2.130	1.919	2.163	2.085	2.046	2.021	2.129	2.079	± 0.081
Atoms	θ_0 **(Degrees)**	**QM/MM ONIOM optimised angle (Degrees)**	**Experimental angle (Degrees)**									
ZN-OX-MG	136.185	132.617	120.423	138.900	125.700	124.800	126.700	136.400	141.500	134.500	131.115	± 7.661

5.3.4 Bond Stretching Parameters, Angle Bending Parameters, and Dihedral Bending Parameters

We calculated six bond stretching parameters (bonds and force constants), thirteen angle bending parameters (angles and force constants), and seven angle bending parameters (angles and force constants) that were used to describe the Zn^{2+}-O(H)-Mg^{2+} ion active site. The fitting of the CHARMM potential energy function to the energy profiles from the PES scans obtained for the bond stretching, angle bending, and dihedral angle bending are shown in Figure 5.6, Figure 5.7, and Figure 5.8 respectively. A summary of the bond, angle, and dihedral parameters calculated are presented in Table 5.5.

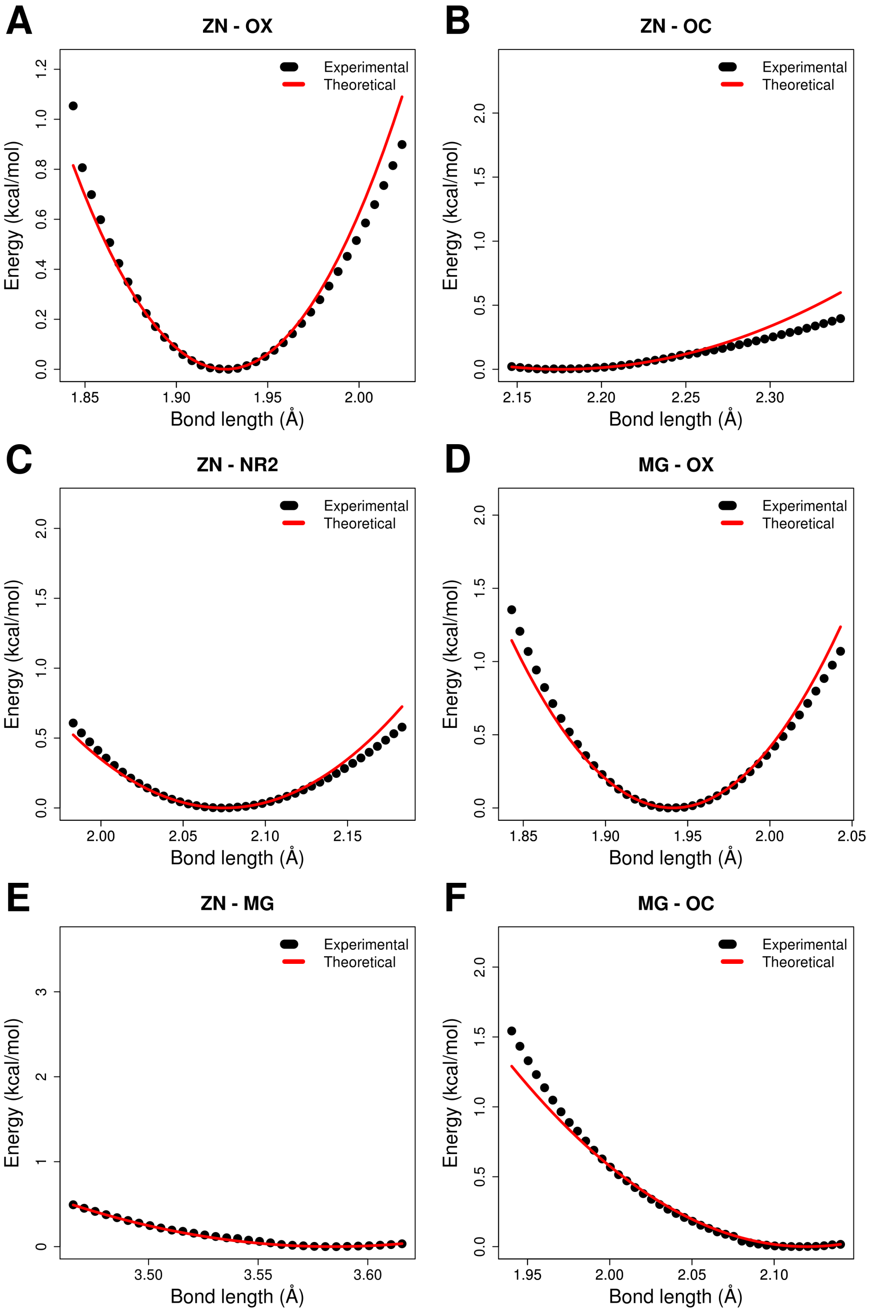


Figure 5.6: Energy profiles generated from the bond stretching PES scans and fitting to the CHARMM potential energy function. The MM fitting curve is shown in red while the black bond stretching is shown by black dots.

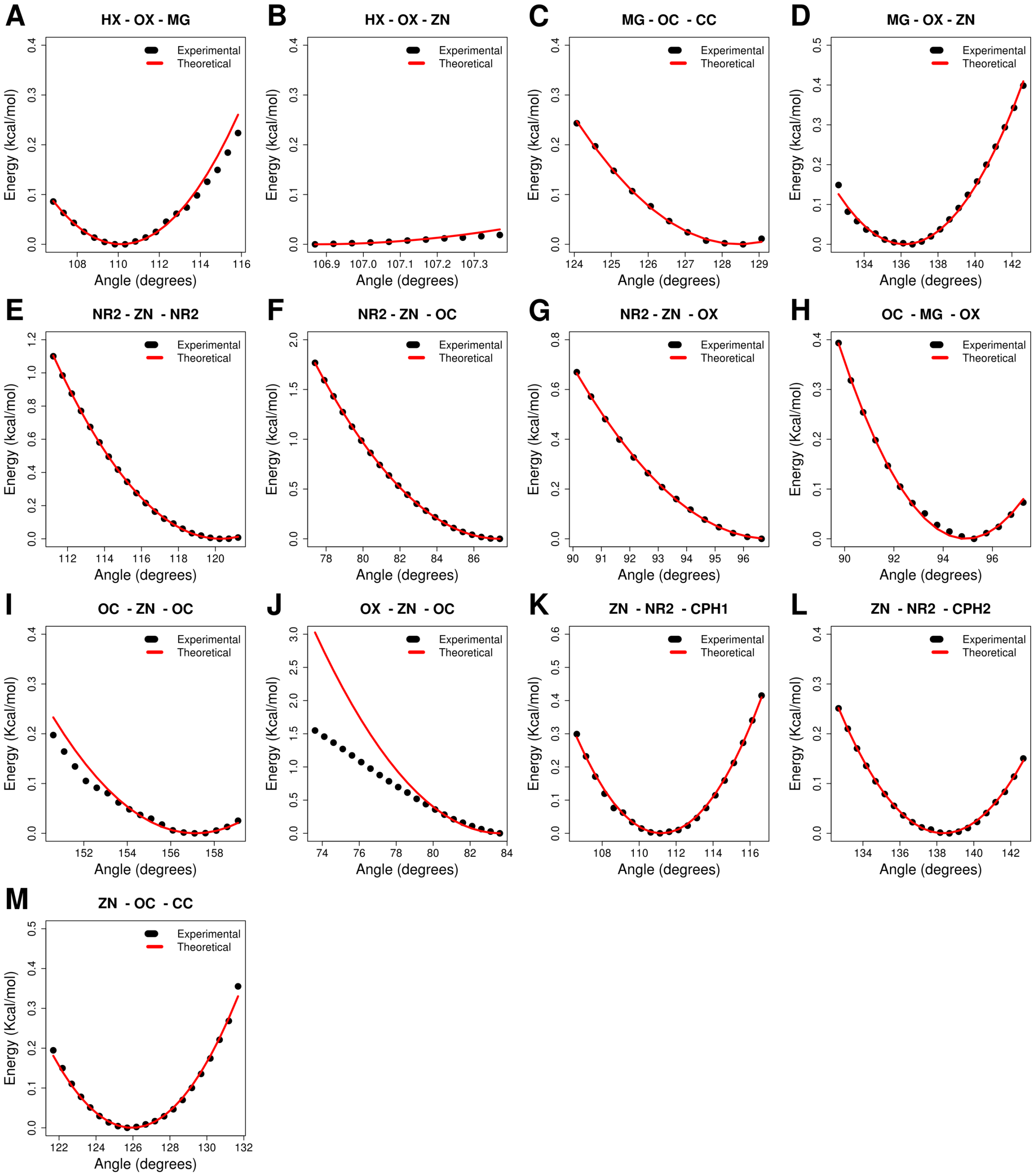


Figure 5.7: Energy profiles generated from the angle bending PES scans and fitting to the CHARMM potential energy function. The MM fitting curve is shown in red while the black angle bending is shown by black dots.

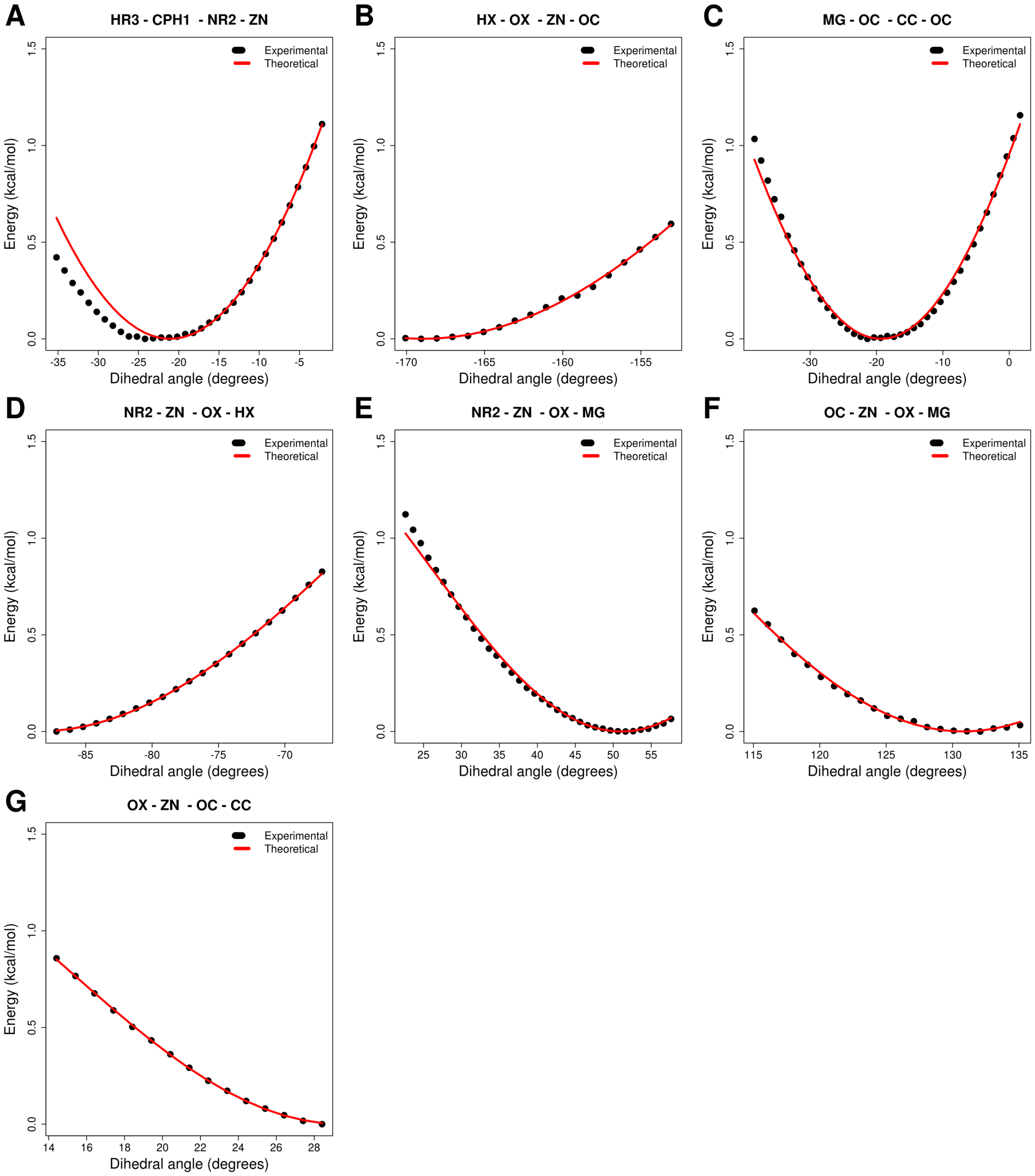


Figure 5.8: Energy profiles generated from the dihedral angle bending PES scans and fitting to the CHARMM potential energy function. The MM fitting curve is shown in red while the dihedral angle bending is shown by black dots.

Table 5.5: Bond stretching, Angle bending, and Dihedral angle bending parameters

Atom type	K_b (kcal/mol/Å^2)	b_0 (Å)
ZN - OX	117.000	1.927
ZN - OC	21.800	2.176
ZN - NR2	62.000	2.075
MG - OX	119.000	1.941
MG - OC	40.000	2.120
ZN - MG*	35.000	3.580

Atom type	K_θ (kcal/mol/rad^2)	θ_0 (Degrees)
OX - ZN - OC	98.480	83.650
OC - ZN - OC	17.880	157.130
OC - MG - OX	48.250	94.920
NR2 - ZN - OX	45.960	96.940
NR2 - ZN - NR2	42.660	120.270
NR2 - ZN - OC	59.090	87.390
MG - OX - ZN	32.480	136.180
HX - OX - ZN	39.390	106.870
HX - OX - MG	26.260	110.130
ZN - OC - CC	32.820	125.940
ZN - NR2 - CPH1	45.960	111.190
ZN - NR2 - CPH2	26.260	138.360
MG - OC - CC	42.680	128.450

Atom type	$K\varphi$(kcal/mol)	n	δ (Degrees)
NR2 - ZN - OX - MG	0.850	3.580	2.864
MG - OC - CC - OC	2.705	2.570	130.060
NR2 - ZN - OX - HX	1.095	3.500	590.140
OC - ZN - OX - MG	1.047	4.150	2.865
HR3 - CPH1 - NR2 - ZN	9.637	1.470	149.030
HX - OX - ZN - OC	1.625	3.174	3.724
OX - ZN - OC - CC	0.802	6.192	2.807

* The fitted Zn^{2+} and Mg^{2+} repulsive potential that mimics a covalent bond. OX denotes the O and HX denotes the H atom in the OH$^-$ ion.

Stabilizing the metal internuclear distance during dynamics proved quite challenging. The Zn^{2+} and Mg^{2+} ions are 1-3 neighbors that are linked via an angle with the highly charged OH$^-$ ion [316]. As a result, their non-bonded van der Waals and electrostatic interaction terms were excluded by the CHARMM program [316]. The omission of these electrostatic and van der Waals interactions resulted in a loss of nuclear-nuclear repulsion between the two weak Lewis acids since they were being treated as part of a bonded system [316,363].

The inclusion of the bridging OH^- anion in the bonded system was essential because it serves as a nucleophile to initiate the catalytic hydrolysis of cAMP [352,355,359]. We wanted to include this chemistry in the dynamics of our system. Given the complexity of the system and given that the OH^- ion resulted in the metal cations becoming bonded atoms, we opted to introduce an experimentally determined repulsive constraint between the two metals based on a fitted bond term with an equilibrium bond length and bond force constant. The fitted potential was close to the PES equilibrium metal internuclear distance calculated from the 2-layer QM/MM ONIOM (Table 5.5 and Figure 5.6). The fitting of this repulsive constraint did not significantly alter the overall chemical properties or geometry of the enzyme during CHARMM MD simulation. Our approach offers a reasonable work-around when describing this bi-nuclear metal center for dynamics simulation.

5.3.5 Validation of the TbrPDEB1 Parameters

The derived force-field parameters were validated using all-atom MD simulations in explicit water at 310 K for 30 ns. The TbrPDEB1 chain A structure (PDB:4i15) was used to provide the starting coordinates for CHARMM MD simulation.

On the 30 ns time scale the four models had equilibrated by 18 ns (Figure 5.9-A). Model 4 showed the most stable conformation as shown by its lowest RMSD relative to its initial structure and protein compactness (Figure 5.9-A and Figure 5.9-B). Models 1, 2, 3, and 4 deviated from the initial crystal structure by an average RMSD of 2.56 ± 0.12 Å, 3.20 ± 0.11 Å, 2.83 ± 0.18 Å, and 2.4 ± 0.08 Å respectively (Figure 5.9-A). Models 2 and model 3 showed the most deviations in conformation from their initial structure.

In all the four models, the residues coordinating the two metals showed low flexibility (Figure 5.9-C). The most flexible regions in all the four models were the H14 helix (E843 - R854) and M-loop (G855 - K866), followed by the H8 helix (N718 - K723), H loop (residues T724 - P727) and H9 helix (L728 - G736) (Figure 5.9-C).

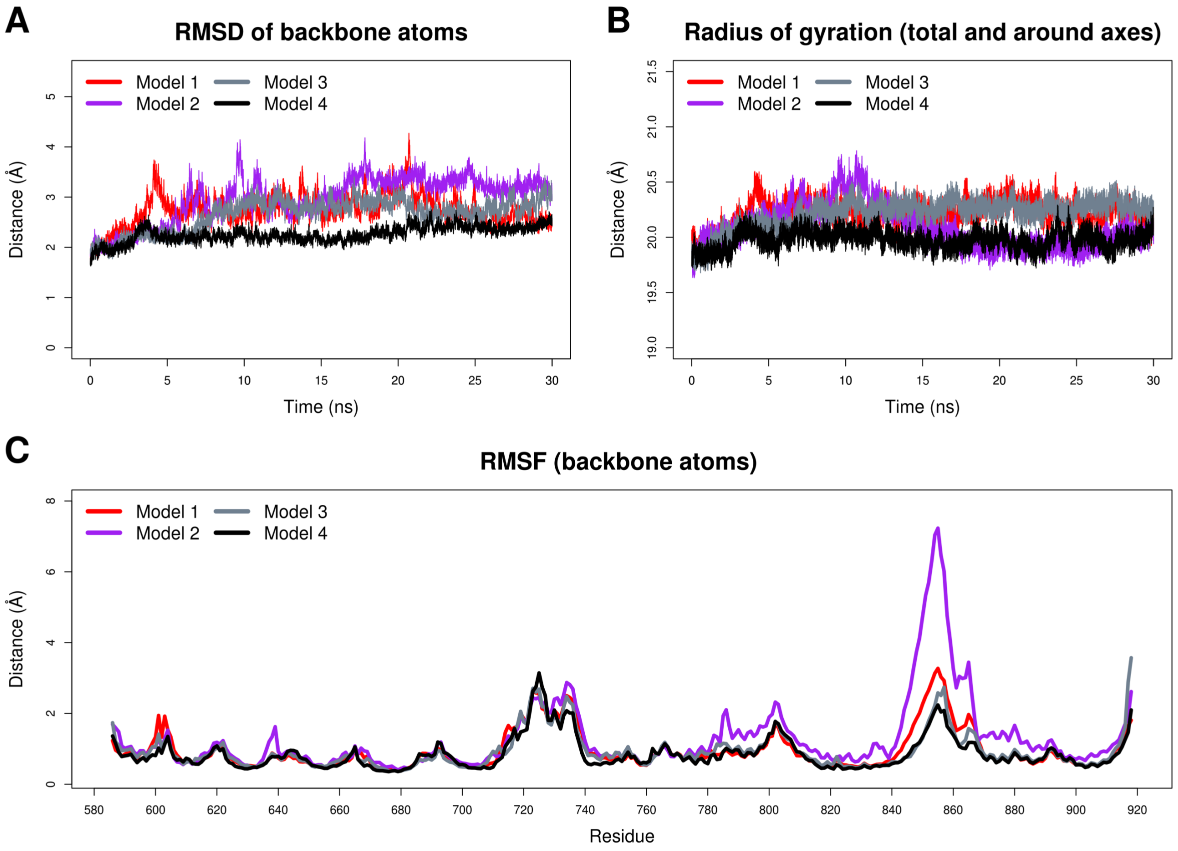


Figure 5.9: MD simulation of the four models. The protein stability was assessed using RMSD, The Rg, and per residue RMSF over the course of the MD simulation at 310k and 1 atm. A) The RMSD of the four models, B) The Rg of the four models, and C) The RMSF of the four models.

Model 2 showed pronounced fluctuation of the H14 helix residues E843 – R854 and M loop residues G855 – K866 (Figure 5.9-C). This was a consequence of residue D710 not being bonded to the two metals. All the models showed very stable potential energy profiles during simulation (Figure 5.10-D).

Analysis of the bi-metal active site dynamics showed that there was a conformational change occurring in model 1 and model 2 that was driven by a histidine ring flip of catalytic residue H669 (Figure 5.10-A – 5.10-C). The conformational change resulted in a drop in the metal internuclear distance from 3.5 Å to approximately 3.0 Å and a reduction in the Zn^{2+} - OX -

Mg^{2+} angle from 120° to ~ 90° (Figure 5.10-A and Figure 5.10-B). A detailed analysis of all four modes follows.

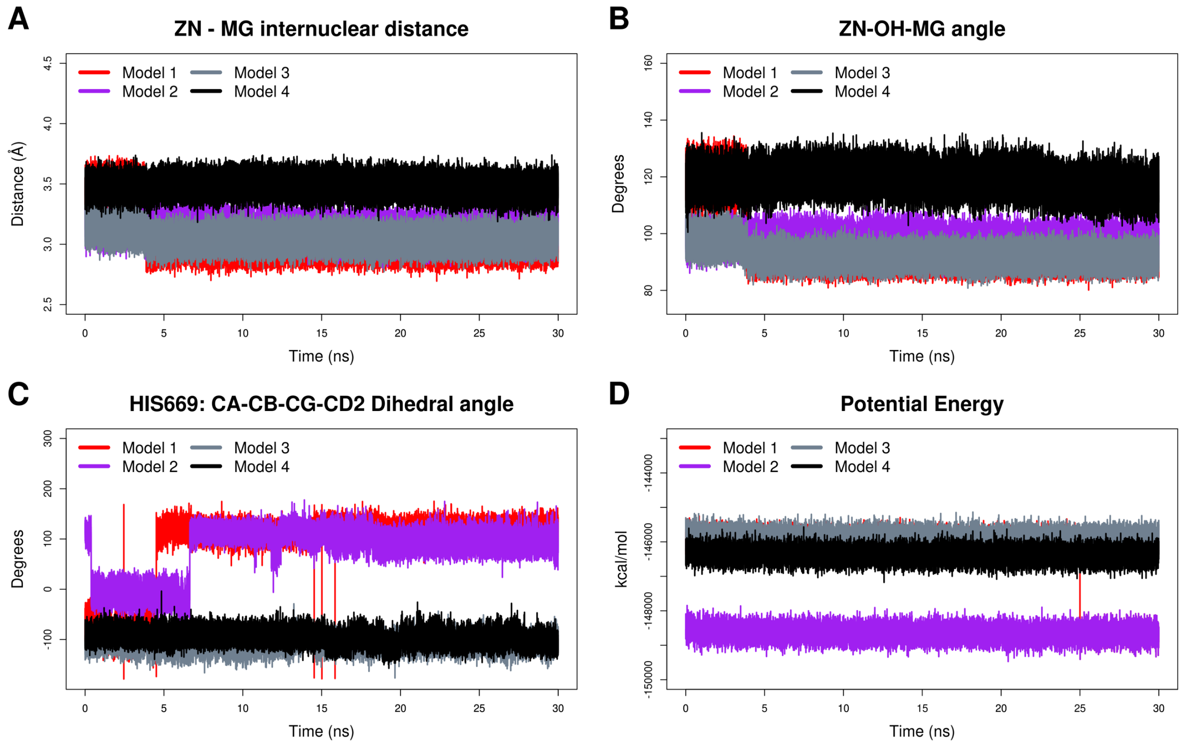


Figure 5.10: The structural properties of the TbrPDEB1 bi-metallic catalytic center during 30 ns all-atom MD simulation at 310 K. A) A plot of the zinc and magnesium internuclear distance, B) A plot of the zinc, hydroxide oxygen atom, and magnesium angle, C) A plot of the H669 ring planarity as monitored by the dihedral angle of CA-CB-CG-CD2, and D) The potential energy of the MD systems in model 1, 2, 3, and 4.

5.3.5.1 Model 1

In model 1, the drop in the internuclear metal distance and Zn^{2+} - OH^{-} - Mg^{2+} ion angle coincides with a histidine ring flip in H669 as demonstrated by a change in planarity by monitoring the H669 CA-CB-CG-ND1 dihedral angle relative to its starting point (Figure 5.10-C). This conformational change occurred at around 4ns into the MD simulation.

As result of the protonation of the NE2 the H669 histidine-ring flips so that it can re-orient itself in a manner that allows the unprotonated ND1 to attack a hydrogen atom from a water molecule in the Mg^{2+} coordination sphere. This is shown by the reduction in the H669:ND1 – TIP3924:H1 distance from ~ 4.70 Å to ~ 2.08 Å (Figure 5.11-A).

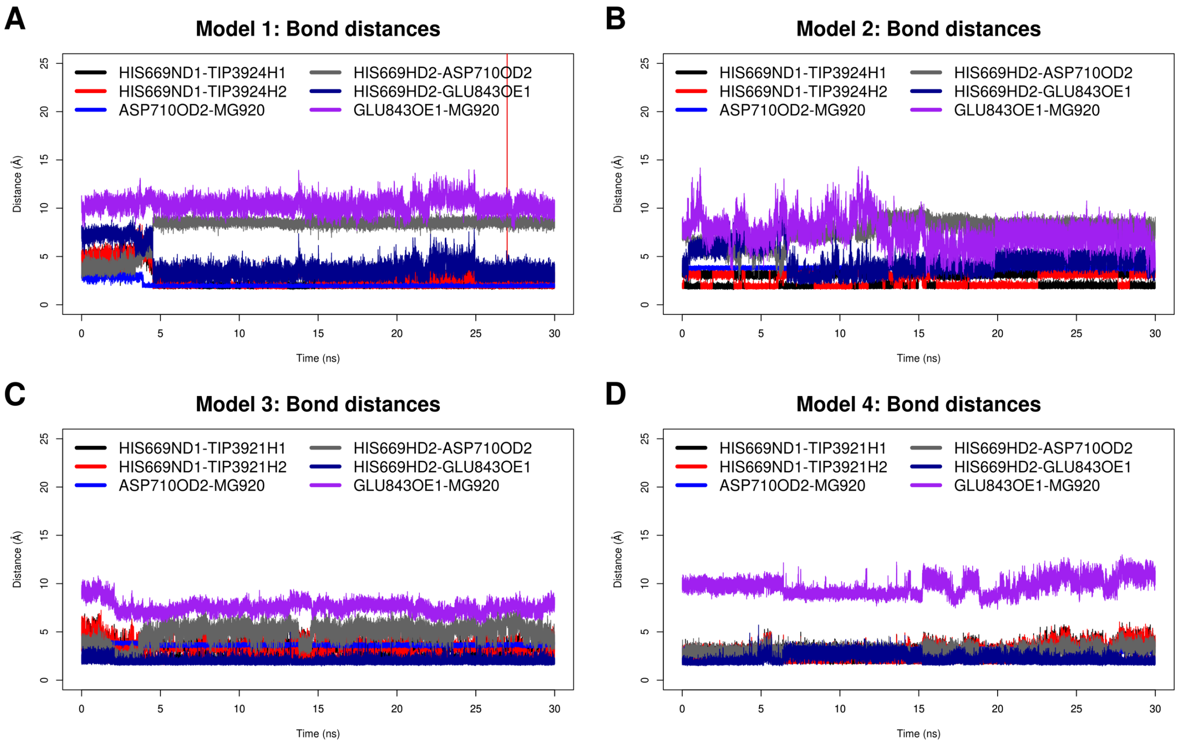


Figure 5.11: The structural properties of the TbrPDEB1 bi-metallic catalytic center. A plot of distances between H66ND1-TIP3924H1, H669ND1-TIP3924H2, D710OD2-MG920, H669HD2-ASP710OD2, H669HD2-GLU843OE1, and E843OE1-MG920.

Further, the carbon hydrogen-oxygen (CH···O) bond between H669:HD2 and D710:OD2 is lost, the distance between the two atoms increases from ~ 3.50 Å to ~ 8.50 Å (Figure 5.11-A). As a result, D710:OD2 displaces a water molecule from the Mg^{2+} coordination sphere. The exchange between the H_2O - Mg ⇌ -COO – Mg can be seen by the reduction in the Mg – D710:OD2 distance from ~ 3.00 Å to ~ 1.90 Å (Figure 5.11-A). The Mg^{2+} is then coordinated by three waters and two oxygen atoms from D710. The intermolecular metal distance is shortened to accommodate the OD2 – H_2O change. E843:OE1 formed a hydrogen bond with H669:HD2, as shown by a reduction in distance from 7 Å to 3.11 Å (Figure 5.11-A). The Zn^{2+}-

Mg^{2+} internuclear distance stabilized after at ~ 3 Å after 30ns simulation which was significantly lower than the PES scan equilibrium distance (Table 5.4 and Figure 5.10-A).

5.3.5.2 Model 2

In model 2, the conformational change occurs much earlier on during the equilibration stages and additionally the histidine ring can be seen flipping between 1ns and 7ns during simulation. Additionally, the increased flexibility of D710 significantly altered the residue interactions with the H14 helix and M-loop as seen by their increased flexibility (Figure 5.9-C). The CH···O bond between H669:HD2 and D710:OD2 was also lost in model 2; however, in this model the D710:OD2 did not replace any of the water molecules in the Mg^{2+} coordination sphere (Figure 5.11-B).

Catalytic residue E843, which is located in the H14 helix, had interactions with a water molecule within the Mg^{2+} coordination sphere. E843:OE1 formed a hydrogen bond with H669:H1 while E843:OE2 formed a hydrogen bond with TIP3924:H1/TIP3924:H2 (Figure 5.11-B). As a result, the E843 residue was brought closer to the Mg^{2+} ion as shown by a reduction in the distance between E843:OE1 and Mg^{2+} dropping from 9.5 Å to 4.5 Å (Figure 5.11-B). This explains the high flexibility observed in the H14 helix and M loop in model 2 (Figure 5.9-C). The Zn^{2+}-Mg^{2+} internuclear distance stabilized after 30ns simulation at ~ 3 Å which was also significantly lower than the PES scan equilibrium distance (Table 5.4 and Figure 5.10-A).

5.3.5.3 Model 3

In model 3 and model 4, the H669 was not protonated the NE2 which prevented the histidine ring flip observed in model 1 and model 2 (Figure 5.10-C). The CH···O bond between H669:HD2 and D710:OD2 was lost in model 3 as shown by an increase in the distance from 2.35 to 5.63 Å between the two atoms (Figure 5.11-C). HSD669:HD2 and D710:OD2 were 4.81 Å apart while the H669:HD2 – E843:OE1 hydrogen bond was maintained (Figure 5.11-

C). The Zn^{2+}-Mg^{2+} internuclear distance stabilized after 30ns simulation at ~ 3 Å which was significantly lower than the PES scan equilibrium distance (Table 5.4 and Figure 5.10-A).

5.3.5.4 Model 4

A key difference between the models is that in model 4, HSD669:HD2 formed a stable hydrogen bond with D710:OD2 as shown by a stable ~ 2.95 Å distance between the two atoms unlike the situation in model 3 where they were 4.81 Å apart (Figure 5.11-C and Figure 5.11-D). In both model 3 and model 4 H669 is seen to form interactions with the hydrogen atoms from a water molecule in the Mg^{2+} coordination sphere and the H669:HD2 – E843:OE1 hydrogen bond was maintained (Figure 5.11-C and Figure 5.11-D). The Zn^{2+}-Mg^{2+} internuclear distance stabilized after 30ns simulation at ~ 3.57 Å in model 4 (Figure 5.10-A).

Model 3 and model 4 show the important role of the D710 bridging ligand, H669, and E843 in regulating the internuclear metal distance.

Model 4 best described the TbrPDEB1 bi-metallic active site. The metal intermolecular distance was stable and fluctuated around 3.5 Å, with the Zn^{2+} having a trigonal bipyramidal geometry and the Mg^{2+} ion an octahedral geometry. The initial amino acid coordination was maintained: Zn^{2+}–NE2 distances, 2.03–2.12 Å (Figure 5.12-A); Zn^{2+}–OX distance, 1.87 Å (Figure 7-A); Zn^{2+}–OD1/OD2 distances 1.91–2.00 Å (Figure 5.12-A); Mg^{2+}–OD1 distance 2.32 Å (Figure 5.12-B); MG^{2+}–OX distance, 2.18 Å (Figure 5.12-B); Mg^{2+}–O1 distances 1.97–1.93 Å (Figure 5.12-B); NE2–Zn^{2+}–NE2 angle, 120.5° (Figure 5.12-C); OD1–Zn^{2+}–OD2 angle, 168.1° (Figure 5.12-C); OD1–Mg^{2+}–OX angle, 87.1° (Figure 5.12-C).

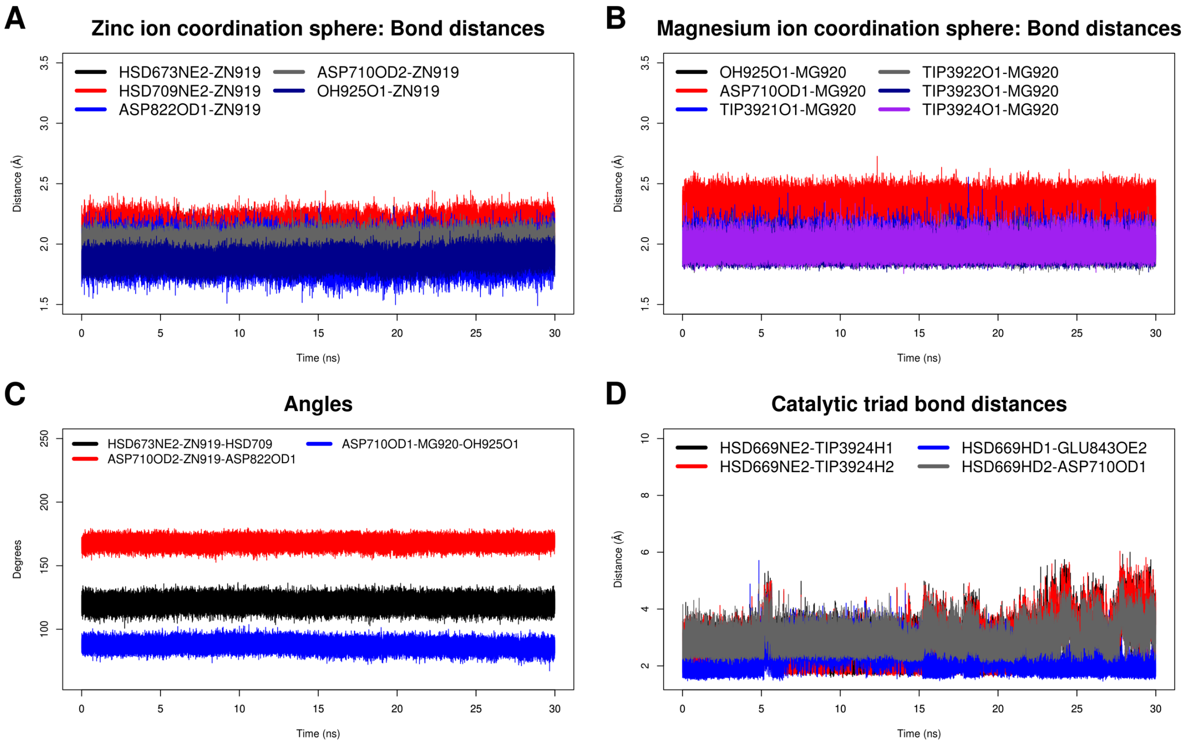


Figure 5.12: The structural properties of the TbrPDEB1 bi-metallic catalytic center during 30 ns all-atom MD simulation at 310 K. A) Bond distances in the zinc coordination sphere, B) Bond distances in the magnesium coordination sphere, C) Angle bending of NE2-ZN-NE2, OD2-ZN-OD1, and OD1-MG-OX, and D) Bond distances in of catalytic residue H669 and E843.

Model 4 gave a metal internuclear distance that was closest to the PES scan equilibrium (Table 5.5). Additionally, it reproduced the bi-metallic catalytic center including the H669, D710, and E843 hydrogen bonding that we have demonstrated is important for protein stability (Figure 5.12-D and Figure 5.13). The D710 - H669 - E843 hydrogen bonding in the “catalytic triad” shown in Figure 5.13 appears to be important in imparting a mechanical force necessary in maintaining the metal internuclear distance and active site geometry.

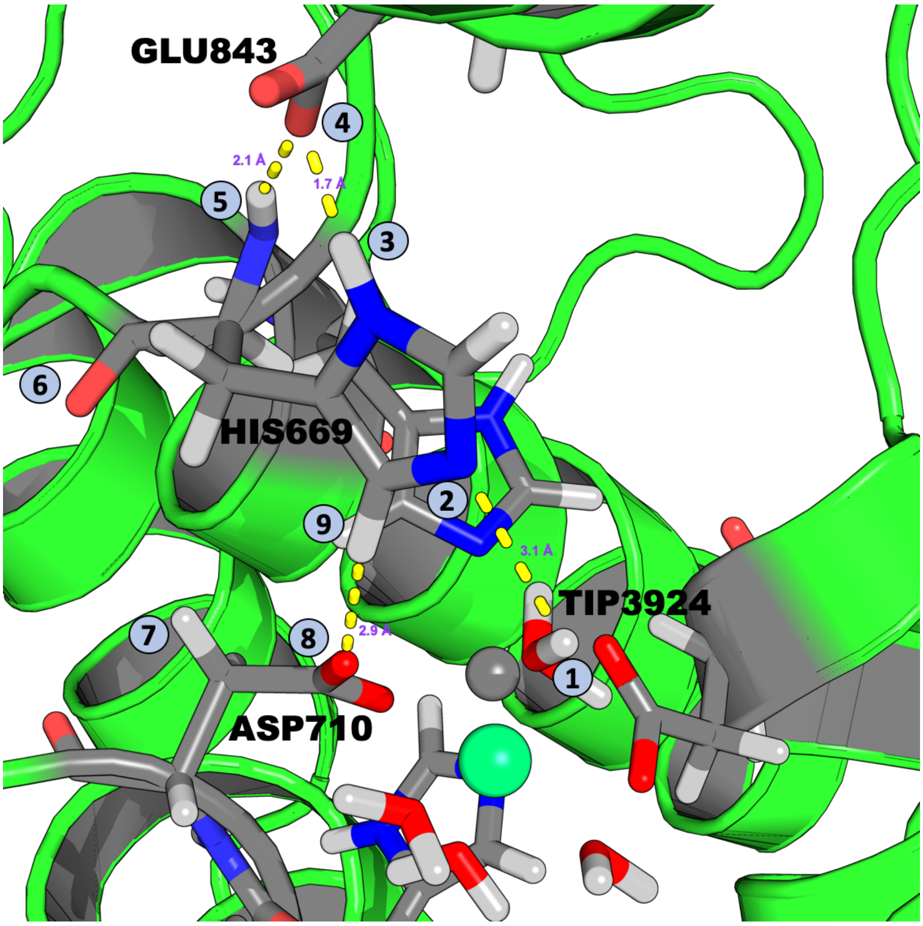


Figure 5.13: A cartoon representation of the TbrPDEB1 bi-metallic active site from Model 4 showing the H669, D710, and E843 "catalytic triad". The two ions are shown as spheres with the zinc ion colored gray and the magnesium ion green. The hydrogen bond network between the three residues D710, H669 and E843 is important for TbrPDEB1 catalysis. 1) TIP3924:H1 (water molecule), 2) H669:NE2, 3) H669:HD1, 4) E843:OE2, 5) H669:HN, and 6) H669:H, and 7) D710:OD1

5.3.6 Molecular Docking

From the 5742 compounds screened, we identified 11 compounds that bound selectively to the parasite PDEB1 orthologues with ≤ -33.472 kJ/mol Autodock Vina binding free energies as shown in Table 5.6. All the 11 compounds were predicted to be PAINS feature free and BBB permeable using CBLIGAND [348]. Two-dimensional protein-ligand interactions for RUBi019, RUBi020, RUBi021, RUBi022, RUBi028 and RUBi030 are shown in Figure 5.14, while those for RUBi023, RUBi025, RUBi026, RUBi027, and RUBi029 are shown in Supplementary Figure 2.

Table 5.6: The top TbrPDEB1 docked compounds and their corresponding Autodock Vina binding energies.

Lab Code	**ZINC database ID**	**IUPAC name**	**TbrPDEB1 (kJ/mol)**	**TcPDEC1 (kJ/mol)**	**LmjPDEB1 kJ/mol**	**HsPDE4B kJ/mol**
RUBi019	**ZINC00082024**	N-{4-[(propan-2-yl)sulfamoyl]phenyl}bicyclo[2.2.1]hept-5-ene-2-carboxamide	-34.309	-36.401	-34.727	-
RUBi020	**ZINC04347173**	N-[4-(ethylsulfamoyl)phenyl]bicyclo[2.2.1]hept-5-ene-2-carboxamide	-33.054	-34.727	-33.472	-
RUBi021	**ZINC00646711**	N-[6-(tert-butylsulfamoyl)-1,3-benzothiazol-2-yl]oxolane-2-carboxamide	-33.890	-36.819	-33.890	-
RUBi022	**ZINC00142276**	N-[4-(diethylsulfamoyl)phenyl]-2-phenylacetamide	-34.727	-37.238	-34.727	-
RUBi023	**ZINC02679474**	2-({5,6-dimethylthieno[2,3-d]pyrimidin-4-yl}sulfanyl)-N-(4-sulfamoylphenyl)acetamide	-35.982	-38.493	-33.054	-
RUBi025	**ZINC00674393**	2-(4-nitrophenoxy)-N-[4-(pyrrolidine-1-sulfonyl)phenyl]acetamide	-33.890	-40.166	-33.054	-
RUBi026	**ZINC00676648**	2-(4-methoxyphenoxy)-N-{4-[(5-methyl-1,2-oxazol-3-yl)sulfamoyl]phenyl}acetamide	-34.727	-38.493	-35.564	-
RUBi027	**ZINC00624981**	N-(5-{[2-oxo-2-(4-phenylpiperazin-1-yl)ethyl]sulfanyl}-1,3,4-thiadiazol-2-yl)butanamide	-33.890	-33.054	-34.309	-
RUBi028	**ZINC00385011**	5-amino-3-ethyl-1-(6-phenoxypyrimidin-4-yl)-1H-pyrazole-4-carboxamide	-34.727	-38.493	-34.727	-
RUBi029	**ZINC06196339**	5-({[4-(2-hydroxyethanesulfonyl)phenyl]amino}methylidene)-1-phenyl-1,3-diazinane-2,4,6-trione	-34.727	-40.166	-34.309	-
RUBi030	**ZINC00358169**	2-cyano-2-((Z)-2-{4-[(4-methyl-1-piperidinyl)sulfonyl]phenyl}hydrazono)acetamide	-33.054	-36.401	-32.635	-

5.3.7 Six Hit Compounds Show Anti-trypanosomal Activity *in vitro*

The eleven compounds were purchased from Molport and subjected to *in vitro* testing against bloodstream form (BSF) trypanosomes. RUBi019, RUBi020, RUBi021, RUBi022, RUBi028 and RUBi030 showed anti-trypanosomal activity *in vitro* with IC_{50}s of 51.49 µM, 51.09 µM, 44.64 µM, 14.96 µM, 50.26 µM, and 40.78 µM respectively. The protein-ligand interactions from molecular docking for the six compounds are shown in Figure 5.13 and their IC_{50} results in Figure 5.14.

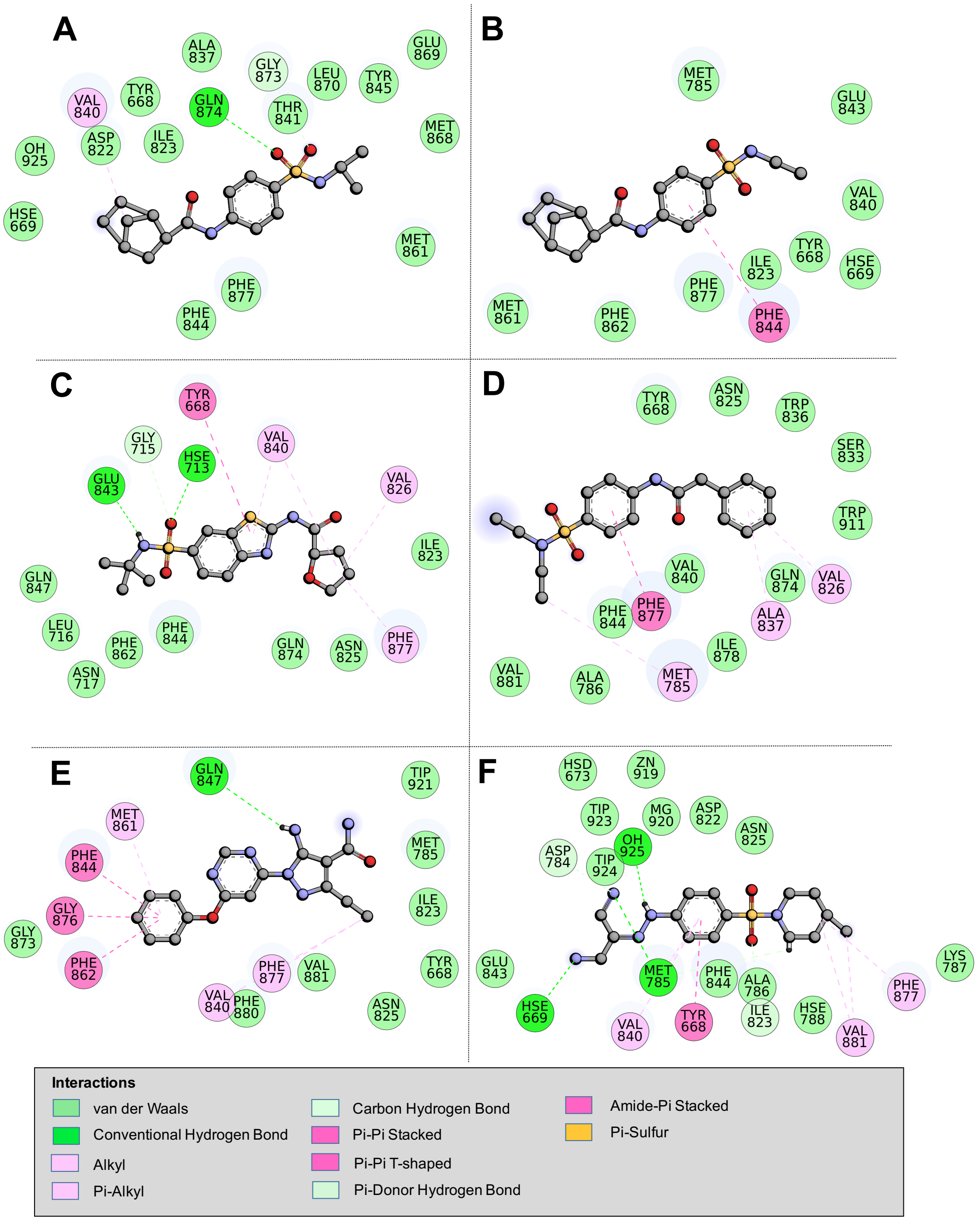


Figure 5.14: The 2D TbrPDEB1 protein-ligand interaction maps showing the binding modes of A) RUBi019, B) RUBi020, C) RUBi021, D) RUBi022, E) RUBi028, and F) RUBi030.

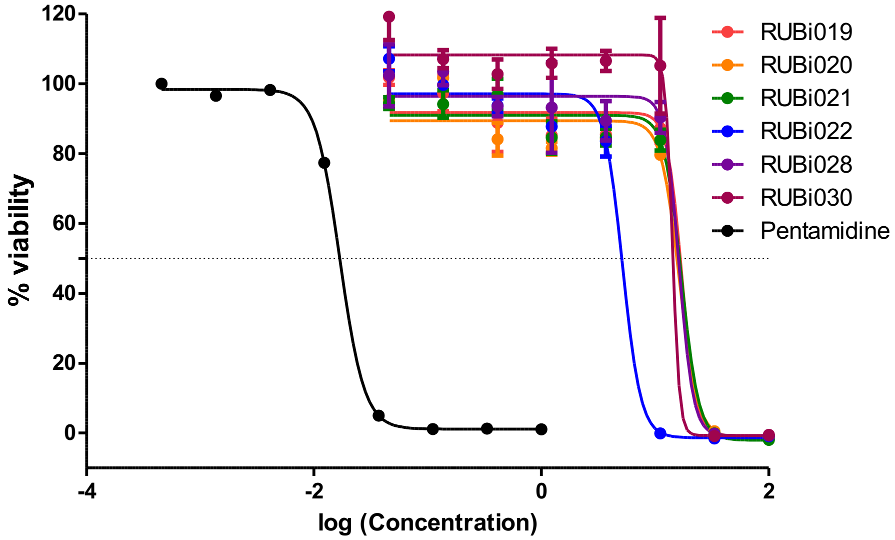


Figure 5.15: Inhibition constant determination of RUBi019 (red), RUBi020 (orange), RUBi021 (green), RUBi022 (blue), RUBi028 (purple), RUBi030 (maroon), and the Pentamidine positive control (black). The IC_{50} of RUBi019, RUBi020, RUBi021, RUBi022, RUBi028 and RUBi030 were determined to be 51.49 µM, 51.09 µM, 44.64 µM, 14.96 µM, 50.26 µM, and 40.78 µM respectively (error bars refer to SD of the biological replicates).

The six compounds also did not show any significant human cytotoxicity at 20 µM as shown in Figure 5.16. Further, when RUBi022 was assayed in triplicate at 100 µM it showed 53 ± 4% HeLa cell viability.

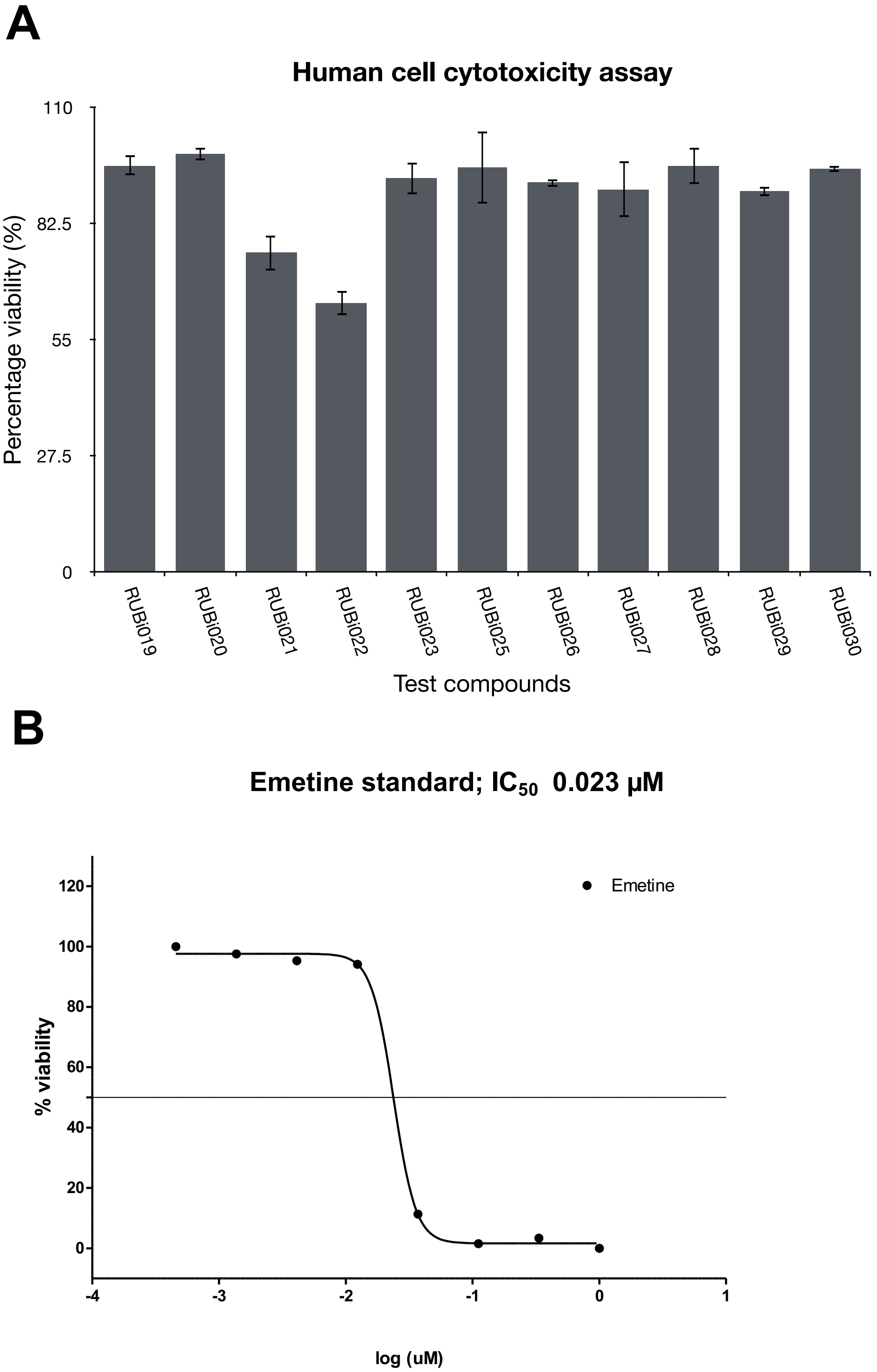


Figure 5.16: Human cytotoxicity screening assay. A) The percentage viability of HeLa cells at 20 µM of the test compounds, and B) The cytotoxicity IC50 of the positive control emetine that was determined to be 0.023 µM (error bars refer to SD of the biological replicates).

We report 6 hit compounds whose *Trypanosoma in vitro* growth inhibition might be as a result of interacting with TbrPDEB1. This was evidenced by a dose dependent increase in intracellular cAMP levels as a result of ligand inhibition. The positive control pentamidine which is an FDA approved anti-trypanosomal drug had no effect on intracellular cAMP levels. Increasing the amount of the ligands three-fold (3 × IC_{50} concentration) resulted in increased cAMP levels as measured by the relative luminescence units.

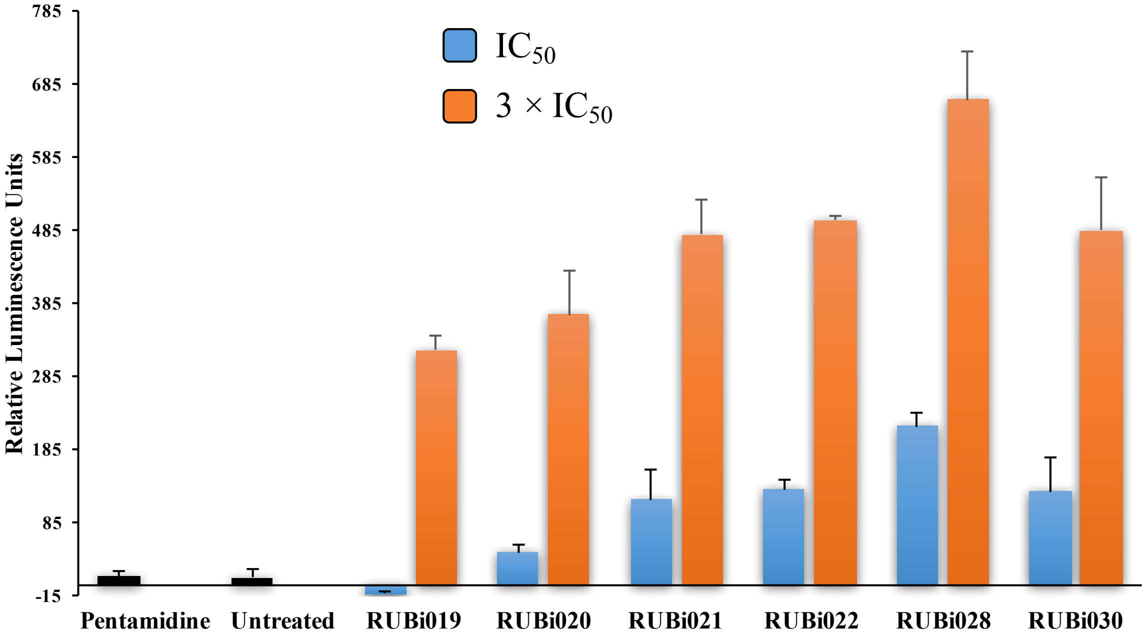


Figure 5.17: The addition of compounds RUBi019, RUBi020, RUBi021, RUBi022, RUBi028 and RUBi030 triggers an increase in intracellular cAMP levels. The trypanosome samples (24 × 10^4 cells/well) were treated with inhibitor concentrations at their respective IC_{50}s (RUBi19 – 50 µM; RUBi20 – 50 µM; RUBi21 - 45 µM; RUBi22 – 15 µM; RUBi28 – 50 µM; RUBi30 – 40 µM) and at 3 x IC_{50} for two hours before measuring intracellular cAMP concentration (Promega cAMP-Glow assay). Controls included an untreated cell sample as well as a pentamidine (1 µM) treated sample. Background luminescence was subtracted from all samples. N=3

Protein-ligand MD simulations were used to propose the molecular interactions of the compounds with the TbrPDEB1 protein.

5.3.8 TbrPDEB1 MD Applications in Drug Discovery to Elucidate Protein-inhibitor Interactions

The 6 hit compounds were used to generate protein-ligand complexes that were subjected to all-atom CHARMM MD simulations. This was done to gain insights into the protein-ligand interactions. This information will be useful in future computational based approaches to predict which inhibitors are likely to inhibit Trypanosome growth in culture by targeting TbrPDEB1. The MD systems were analysed using: RMSD, Rg, per residue RMSF, Hydrogen bond analysis, PCA, EDA and DRN analyses.

5.3.8.1 Conformational Stability of Protein-ligand Complexes

A comparison with the apo protein showed that the metal internuclear distance was stable and held between of 3.50 - 3.79 Å during the 30 ns all-atom MD simulations (Figure 5.18). All the 6 protein-ligand systems showed stable and linear trajectories, with the exception of the TbrPDEB1-RUBi022 complex as shown in Figure 5.19-A to Figure 5.24-A.

MD trajectory analysis of the last 5 ns of TbrPDEB1-RUBi022 complex (protein backbone) showed that it was stable as shown its RMSD of 2.97 ± 0.42 Å when compared to its initial structure (Figure 5.22-A). Further, the TbrPDEBB1-RUBi022 complex showed a reduction in the compactness of the protein when compared to the apo protein as measured by the Rg (Figure 5.22-B). RUBi022 was also the most potent of the 6 compounds (Figure 5.15).

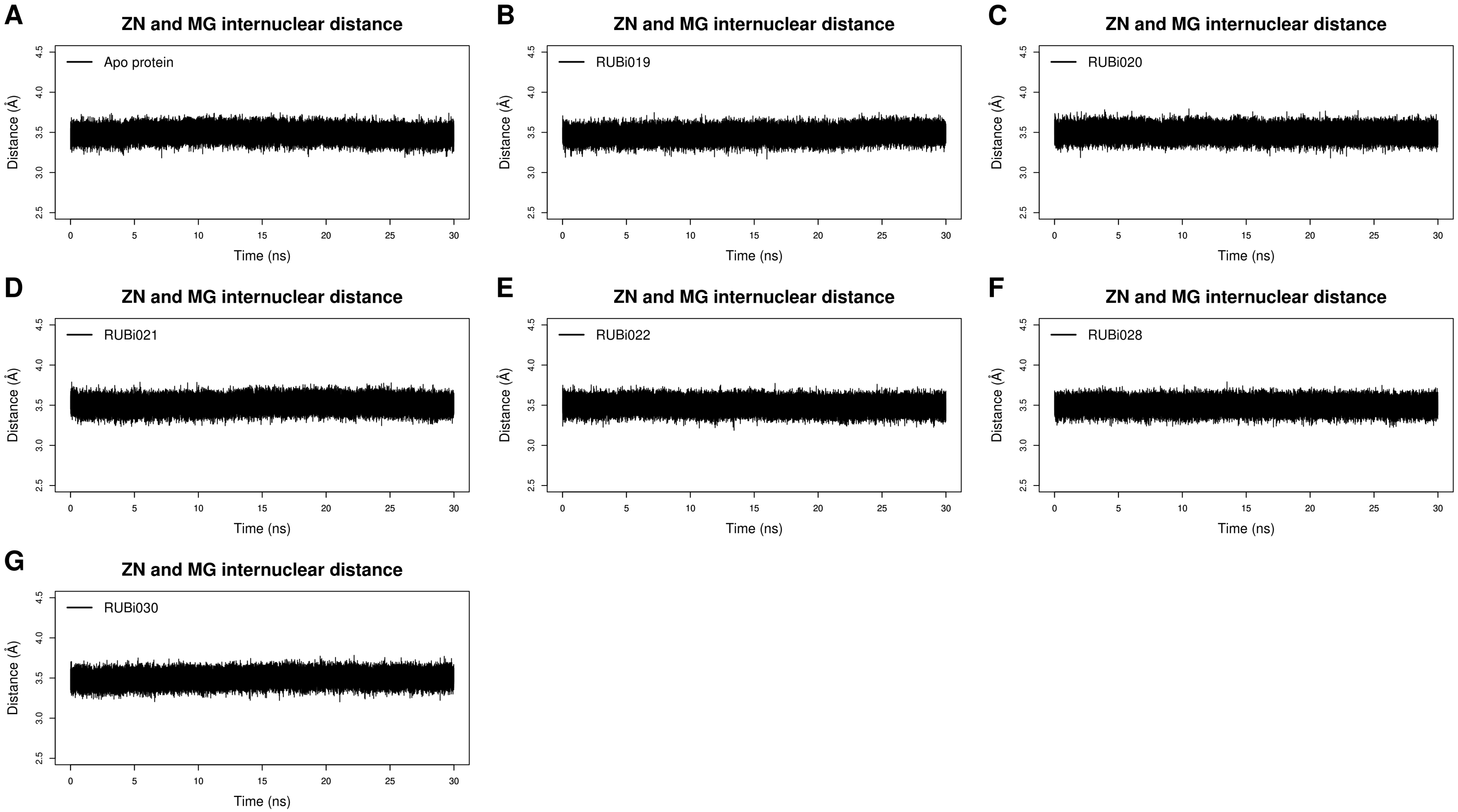


Figure 5.18: Zinc and Magnesium internuclear distance during 30 ns all-atom MD simulation at 310 K. A) Apo protein, B) TbrPDEB1-RUBi019, C) TbrPDEB1-RUBi020, D) TbrPDEB1-RUBi021, E) TbrPDEB1-RUBi022, F) TbrPDEB1-RUBi028, and G) TbrPDEB1-RUBi030.

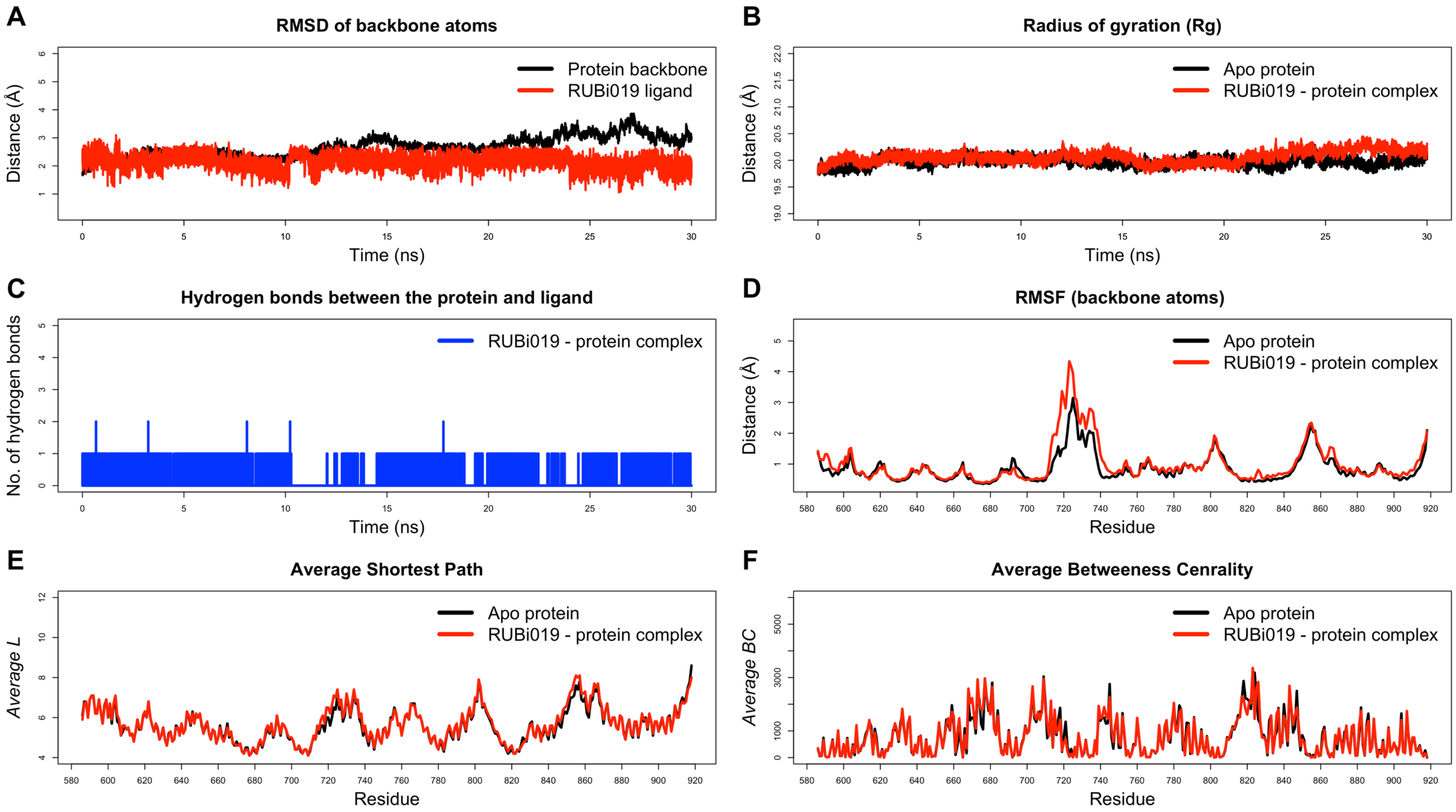


Figure 5.19: Analysis of the TbrPDEB1-RUBi019 complex during 30 ns all-atom MD simulation at 310 K. A) RMSD B) Rg, C) Protein-ligand Hydrogen bond analysis, D) RMSF, E) *Average L*, and F) *Average BC*.

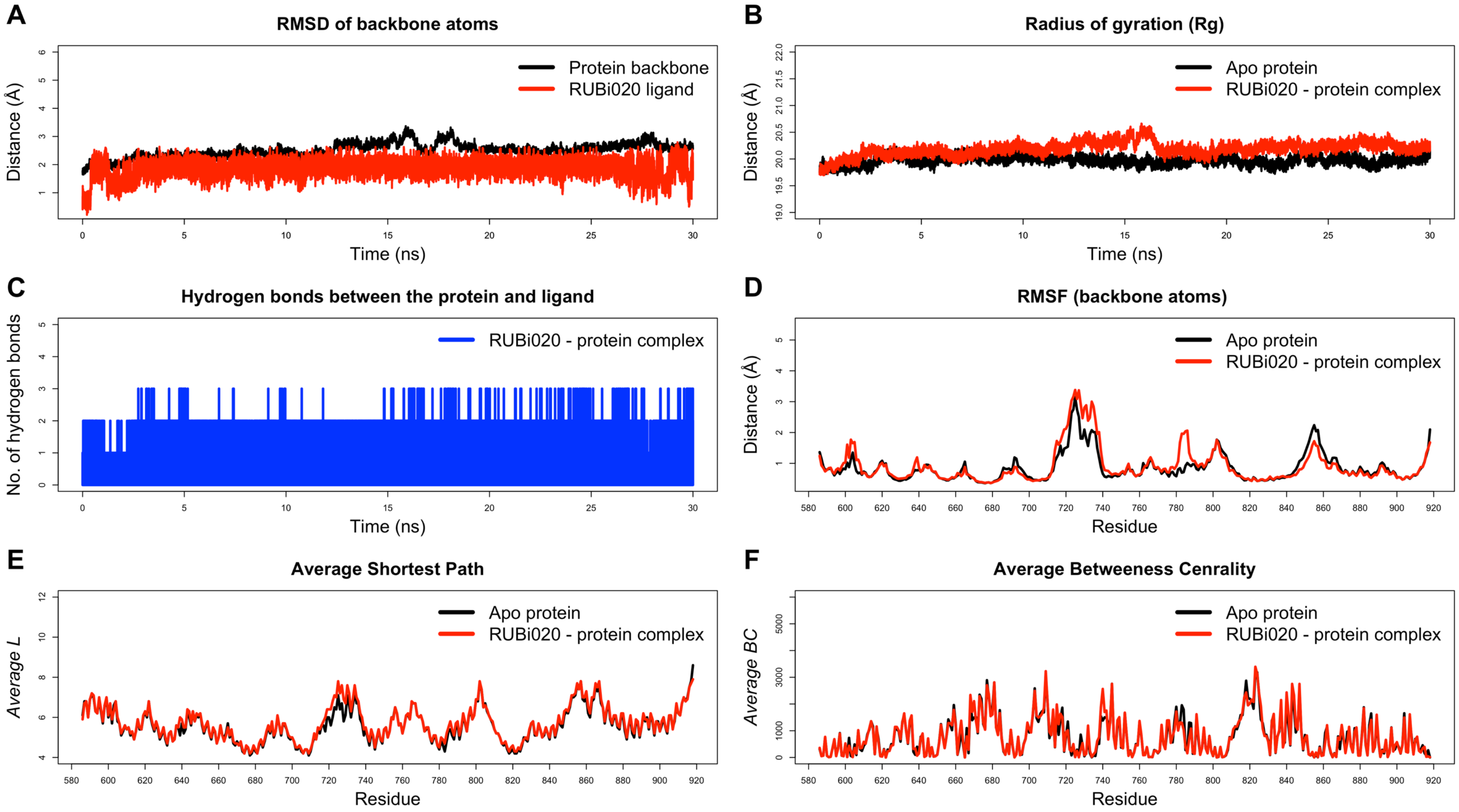


Figure 5.20: Analysis of the TbrPDEB1-RUBi020 complex during 30 ns all-atom MD simulation at 310 K. A) RMSD B) Rg, C) Protein-ligand Hydrogen bond analysis, D) RMSF, E) *Average L*, and F) *Average BC*.

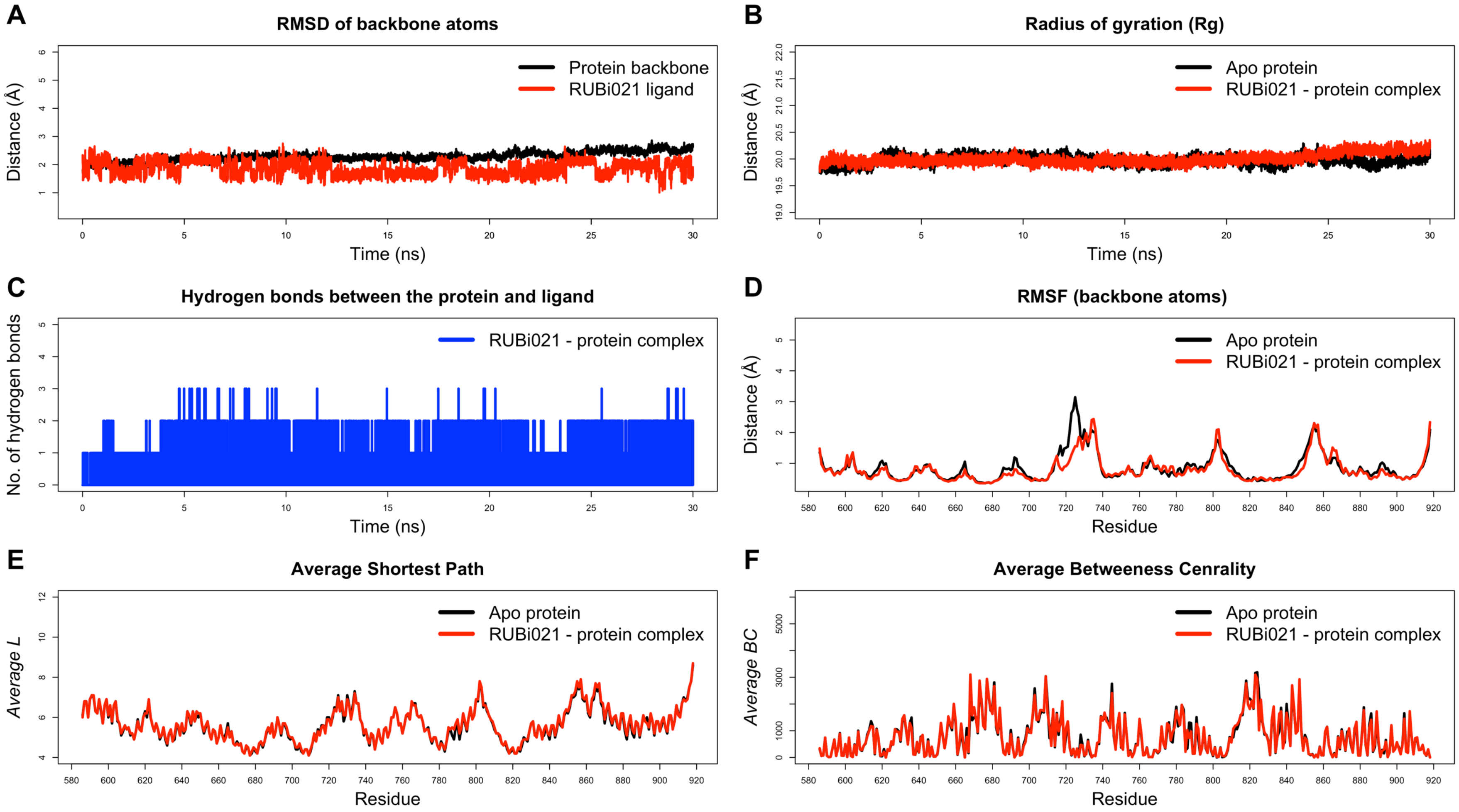


Figure 5.21: Analysis of the TbrPDEB1-RUBi021 complex during 30 ns all-atom MD simulation at 310 K. A) RMSD B) Rg, C) Protein-ligand Hydrogen bond analysis, D) RMSF, E) *Average L*, and F) *Average BC*.

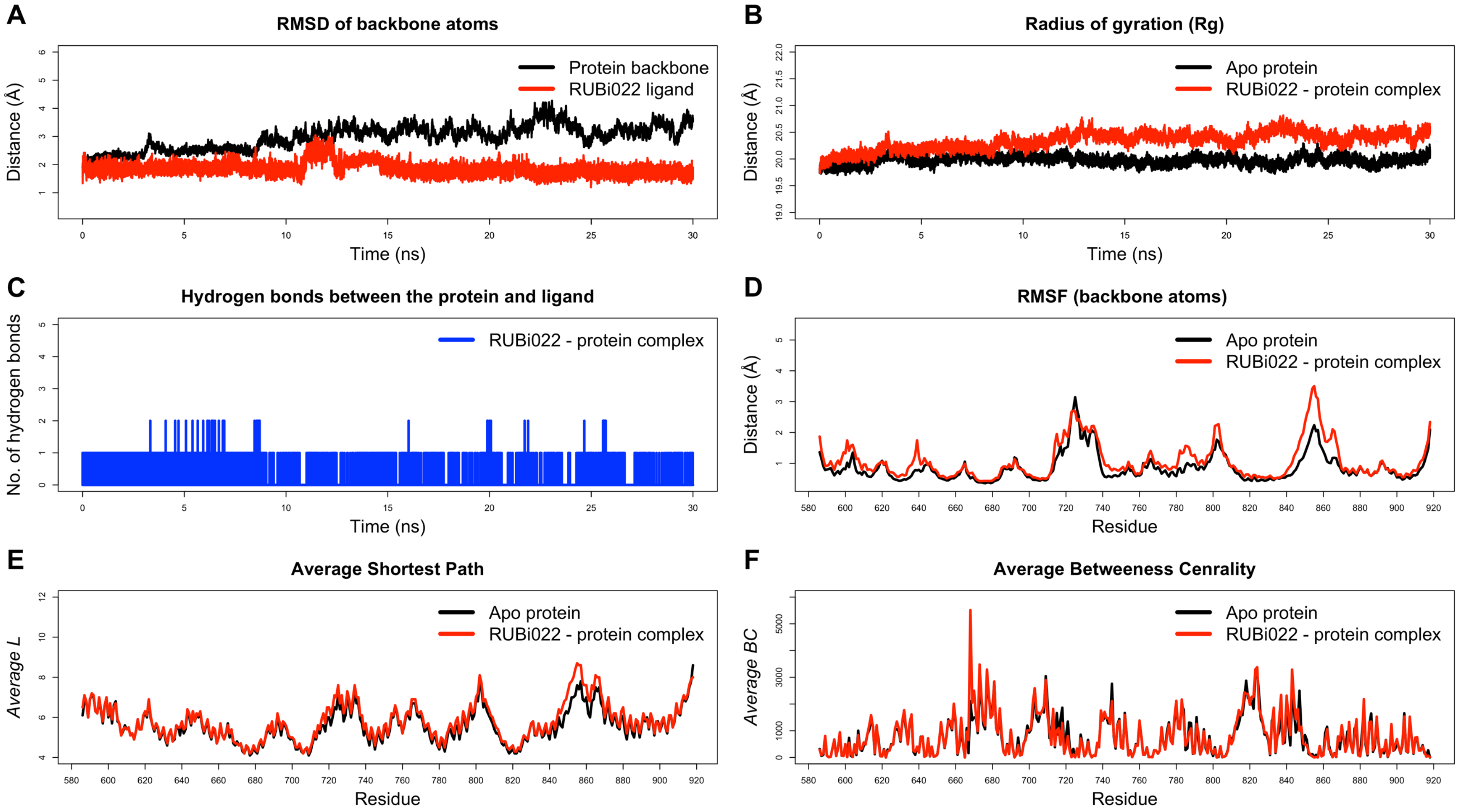


Figure 5.22: Analysis of the TbrPDEB1-RUBi022 complex during 30 ns all-atom MD simulation at 310 K. A) RMSD B) Rg, C) Protein-ligand Hydrogen bond analysis, D) RMSF, E) *Average L*, and F) *Average BC*.

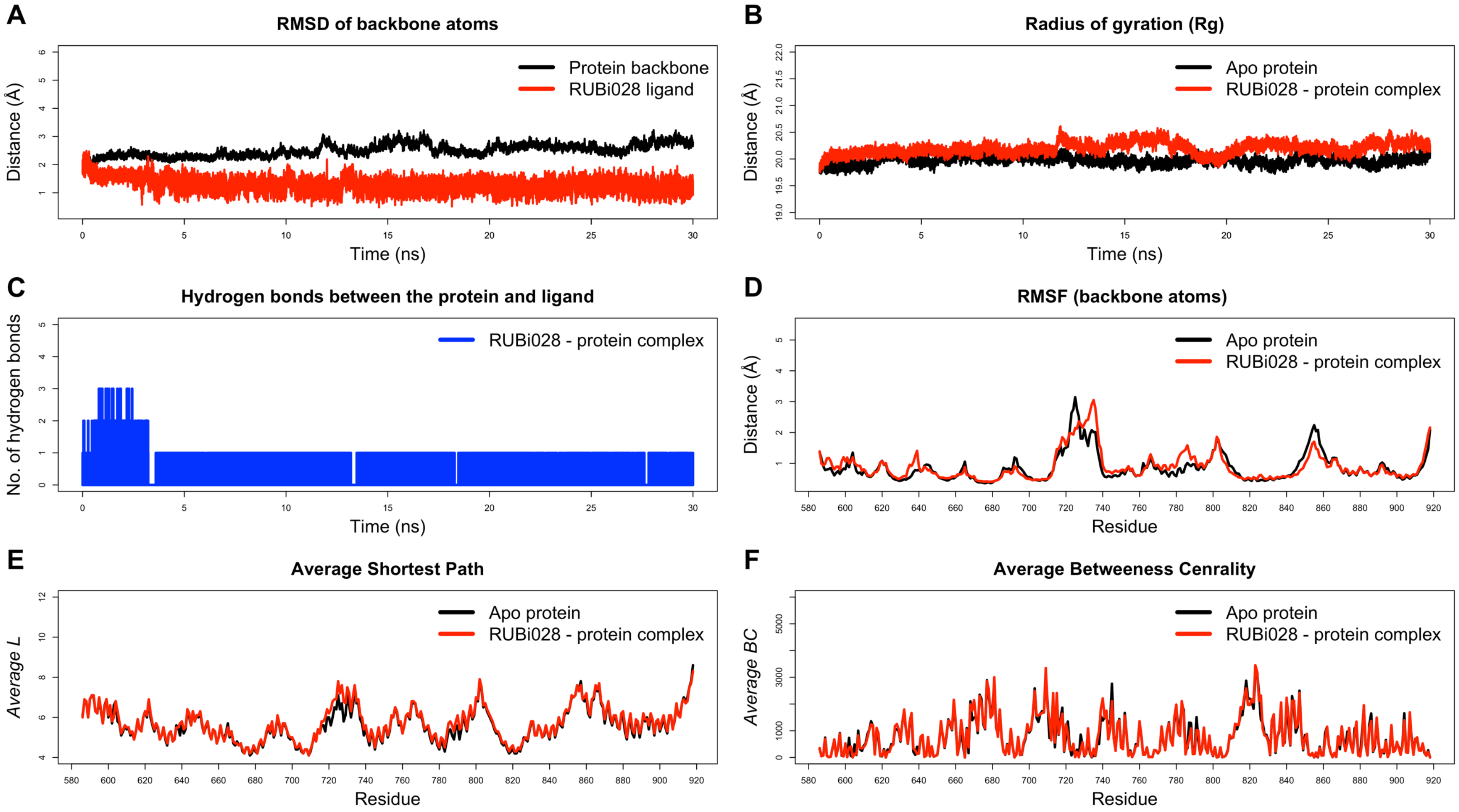


Figure 5.23: Analysis of the TbrPDEB1-RUBi028 complex during 30 ns all-atom MD simulation at 310 K. A) RMSD B) Rg, C) Protein-ligand Hydrogen bond analysis, D) RMSF, E) *Average L*, and F) *Average BC*.

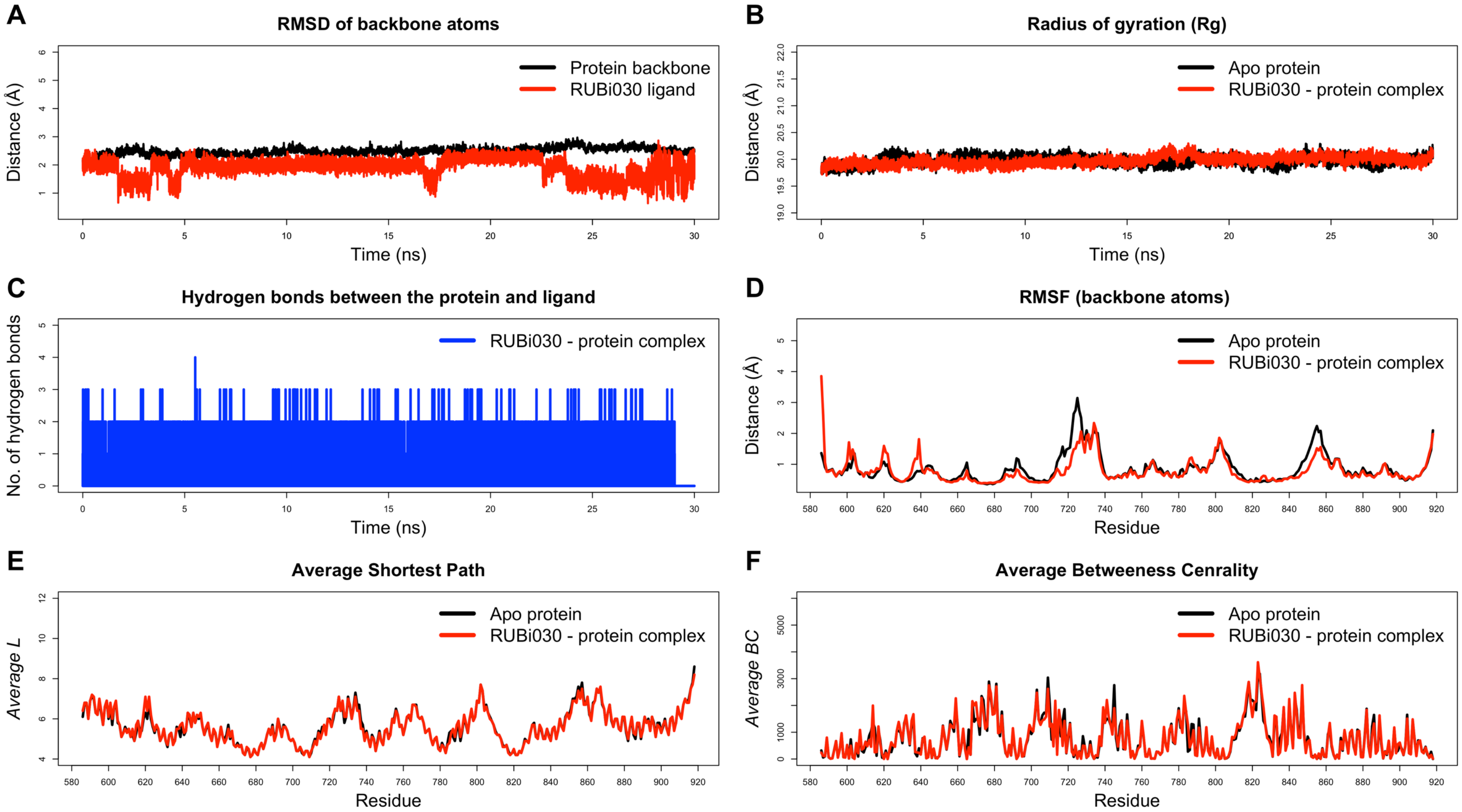


Figure 5.24: Analysis of the TbrPDEB1-RUBi030 complex during 30 ns all-atom MD simulation at 310 K. A) RMSD B) Rg, C) Protein-ligand Hydrogen bond analysis, D) RMSF, E) *Average L*, and F) *Average BC*.

5.3.8.2 Protein-ligand Hydrogen Bond Analysis

Hydrogen bond analysis of the 30 ns MD simulations showed that the compound RUBi019 formed a hydrogen bond with G873 while RUBi020 formed hydrogen bonds with N825, T841, and G873 (Figure 5.19-C and Figure 5.20-C). RUBi021 formed hydrogen bonds with Y668, N717, and Q874 (Figure 5.21-C). RUBi022 (which was the most potent) and RUBi028 formed hydrogen bonds with residue H669 and Y668 respectively (Figure 5.22-C and Figure 5.23-C). Lastly, RUBi030 formed hydrogen bonds with L716 and N718 (Figure 5.24-C). Y668 and H669 appear to be important residues for successful TbrPDEB1 inhibition.

5.3.8.3 Protein Flexibility

The calculated per residue RMSF showed that the binding of the ligands resulted in altered residue fluctuations. In comparison to the apo protein, the TbrPDEB1-RUBi019 complex backbone atoms of N718, S719, F720, Y721, L722, K723, T724, and T738 showed increased flexibility (Figure 5.19-D). In the TbrPDEB1-RUBi020 complex L728, L731, A734, T783, D784, M785, and A786, showed increased flexibility (Figure 5.20-D). In the TbrPDEB1-RUBi021 we observed a reduced flexibility in L722, K723, T724, E725, and S726 (Figure 5.21-D). In the TbrPDEB1-RUBi022 we observed an increased flexibility in M850, E851, K852, E853, R854, G855, V856, and E857 (Figure 5.22-D). In the TbrPDEB1-RUBi028 we observed a decreased in the flexibility of T724 and E725 while we observed an increased flexibility of S732, A734, and S735 (Figure 5.23-D). Lastly, in the TbrPDEB1-RUBi030 complex we observed a decreased flexibility of N717, K723, T724, E725, and S726 while we observed an increased flexibility of V586, T587 and S639 (Figure 5.24-D).

5.3.8.4 DRN Analysis

Average L was used to measure changes in spatial displacement due to ligand binding which correlated well with observed changes in RMSF. In comparison to the apo protein the TbrPDEB1-RUBi019 complex showed an increase in the *Average L* of Y721, K723, T724, R854, and V856 (Figure 5.19-E). In the TbrPDEB1-RUBi020 complex we observed an increase in the *Average L* of S639, Y721, S726, P727, L728 ((Figure 5.20-E). In the TbrPDEB1-RUBi021 complex we observed an increase in the *Average L* of H788 and K787 (Figure 5.21-E). In the TbrPDEB1-RUBi022 the increase in *Average L* was observed in Q847, M850, E851, E853, R854, G855, and V856 (Figure 5.22-E). In the TbrPDEB1-RUBi028 complex we observed an increase in the *Average L* of Y721, L722, P727, and L728 (Figure 5.23-E). Lastly, in the TbrPDEB1-RUBi030 complex we observed an increase in the *Average L* of S620 and S639 (Figure 5.24-E).

Average BC was used to measure changes in the communication flow in the protein network. In the TbrPDEB1-RUBi019 complex we observed a decreased centrality of residue S824 and increased centrality of residue V826 (Figure 5.19-F). In the TbrPDEB1-RUBi020 complex we observed an increased centrality of residue V740 (Figure 5.20-F). In the TbrPDEB1-RUBi021 complex we observed a decreased centrality of residue N825 and increased centrality of residue Y668 (Figure 5.21-F). In the TbrPDEB1-RUBi022 complex we observed a decreased centrality of residue Q847 and markedly increased centrality of H673, Y668, and N670 (Figure 5.22-F). In the TbrPDBE1-RUBi028 complex we observed an increased centrality of residue N825 (Figure 5.23-F). Lastly, in the TbrPDEB1-RUBi030 complex we observed an increased centrality of residue M785 and decreased centrality of residue N717 (Figure 5.24-F).

5.3.8.5 Protein-ligand Interaction Energy

An estimate of the protein-ligand interaction energy was calculated using CHARMM [336]. The compounds bound stably with the following interaction energies: RUBi019 -39 ± 5 kJ/mol, RUBi020 -55 ± 3 kJ/mol, RUBi021 -60 ± 4 kJ/mol, RUBi022 -45 ± 10 kJ/mol, RUBi028 -41 ± 4 kJ/mol, and RUBi030 -46 ± 7 kJ/mol. A per residue contribution of the main residues involved in the interaction energies are shown in Figure 5.25. RUBi022 and RUBi030 were the only compounds that had a favorable energy contribution from the Mg^{2+} ion.

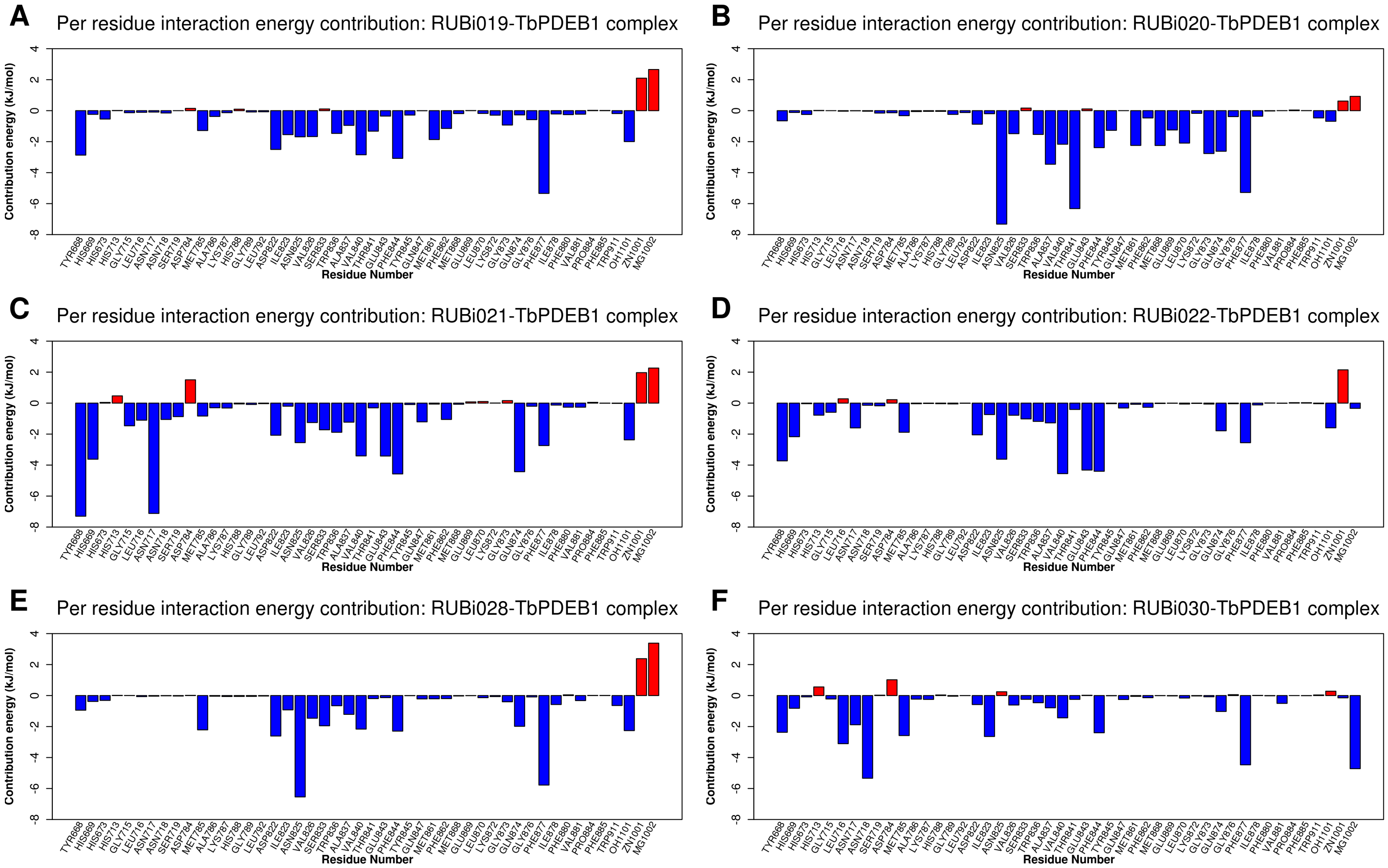


Figure 5.25: A histogram showing the main per residue interaction energy contributions in the protein-ligand complexes.

5.3.9 TbrPDEB1 Protein-ligand Complex MD Simulations Help to Identify Important Molecular Interactions

RUBi022 was the most potent ligand and showed different dynamics from the other five compounds and the apo protein (Figure 5.15 and Figure 5.22). It formed hydrogen bonds with H669 and Y668 which resulted in altered flexibility of H14-helix residues M850, E851, K852, E853, R854, G855, V856, and E857 (Figure 5.22-D). The change in flexibility resulted in altered spatial displacement of the H14 helix in the same region as measured using *Average L* (Figure 5.22E). RMSF has been previously reported to correlate well with spatial displacement measure *Average L* [205,209].

Average BC is important in identifying residues important for protein information flow and this has been shown to coincide with functionality. We observed an increased centrality of Y668, H673, and N670 along with a decrease in the centrality of Q847 (Figure 5.22-F). N670, Zn^{2+} coordinating residue H673, Y668, and Q847 are all highly conserved among all the PDE orthologues (Figure 5.4). Analysis of the TbrPDEB1 chain A crystal structure (PDB:4i15) showed that Y668 forms conventional hydrogen bonds with catalytic residue E843, N670, and N825.

The observed an increased displacement and decreased centrality of E847 in the TbrPDEB1-RUBi022 complex is noteworthy. This highly conserved residue formed hydrogen bonds with F844 and catalytic residue E843 (Figure 5.26). Q847 also formed vdW interactions with the RUBi022 ligand, H14 helix residues Y845, R846, G848, D849, and M850 and H8 helix residue L716 (Figure 5.26). This accounts for the observed an increased flexibility and spatial displacement of the H8 and H14 helices as a result of ligand binding (Figure 5.22-D and Figure 5.22-E).

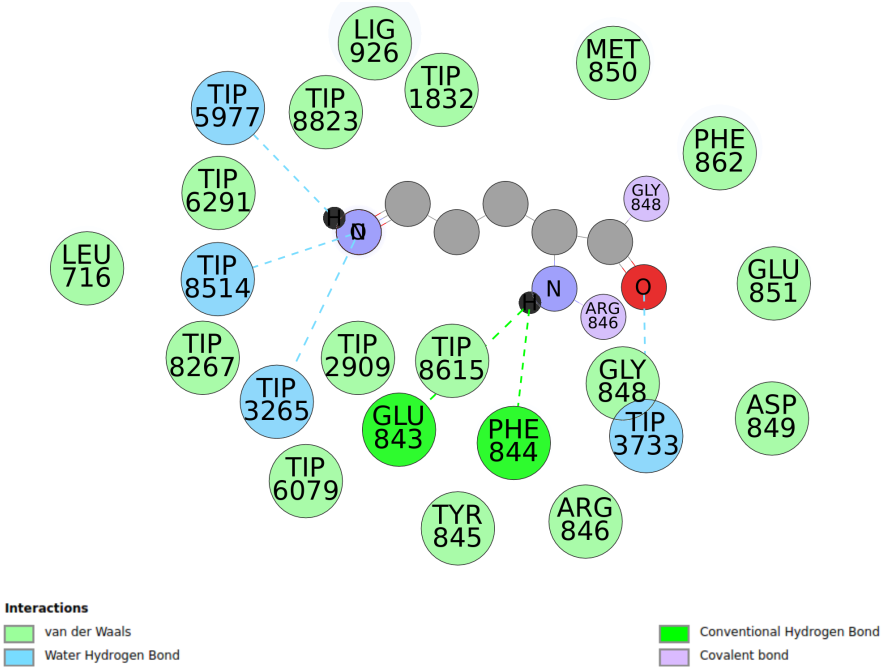


Figure 5.26: A 2D residue interaction map around Q847 in the TbrPDEB1-RUBi022 complex

Further, PCA analysis showed that RUBi022 altered the dynamics of the protein and that the largest observed motions in the complex were in the M-loop, H8, H9, and H14 helices (Figure 5.27 and Figure 5.28).

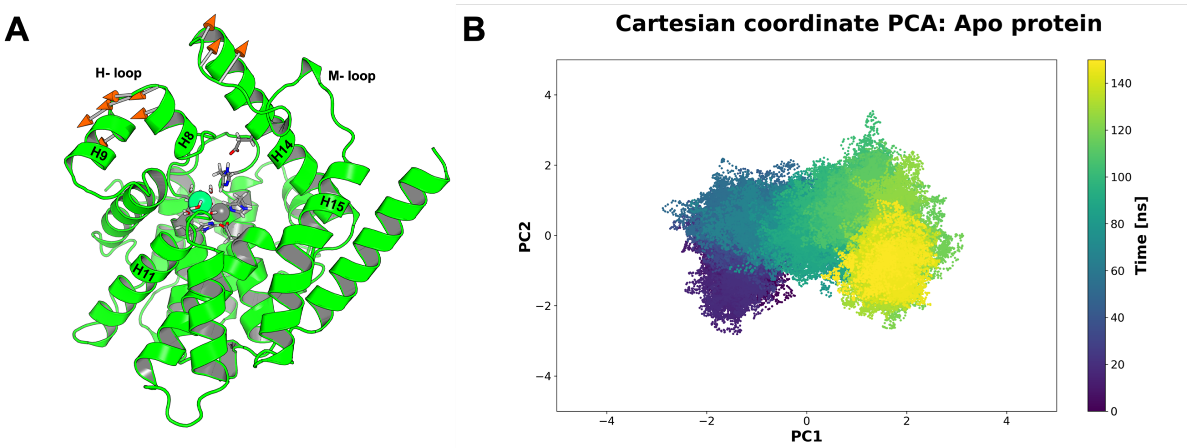


Figure 5.27: Principal component analysis of the TbrPDEB1 apo protein. A) The largest differential motions during the 30 ns all-atom MD simulation are indicated using orange tip, the magnitude of the motions are indicated by the arrow length, B) The motions of the Cβ heavy atoms are shown along the first and second components (PC1 and PC2, 49% and 23% variance explained respectively).

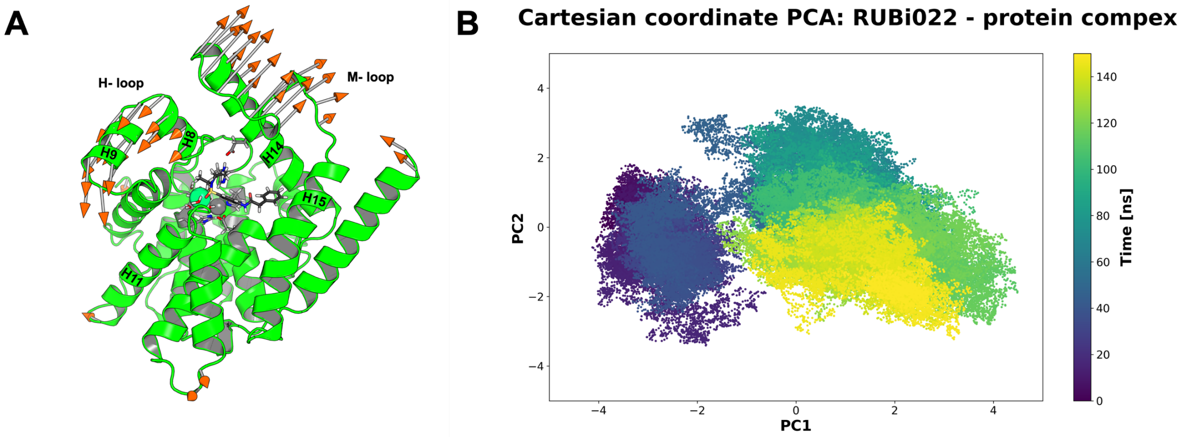


Figure 5.28: Principal component analysis of the TbrPDEB1-RUBi022 complex. A) The largest differential motions during the 30 ns all-atom MD simulation are indicated using orange tip, the magnitude of the motions are indicated by the arrow length, B) The motions of the Cβ heavy atoms are shown along the first and second components (PC1 and PC2, 58% and 19% variance explained respectively).

In addition, DRN analysis of the residue contacts of the catalytic residues H669, E843 and nucleophilic OH^- ion showed that compound RUBi022 was within 6.7 Å of H669 and E843 (Figure 5.29 and Figure 5.30-A-B). In comparison to the apo protein, the TbrPDEB1-RUBi022 complex showed reduced RRIF between H669 with the OH^- ion (Apo protein: 0.560; TbrPDEB1-complex: 0.392), HSD669 with E843 (Apo protein: 0.214; TbrPDEB1-complex: 0.005), HSD669 with G715 (Apo protein: 0.588; TbrPDEB1-complex: 0.034), E843 with Q847 (Apo protein: 0.468; TbrPDEB1-complex: 0.117), and OH^- with T783 (Apo protein: 1.0; TbrPDEB1-complex: 0.478) (Figure 5.29-A-C and Figure 5.30-A–C). Both H669 and Y668 are seen to contribute favorably to the protein-ligand interaction energy with values of -3.728 and -2.168 kJ/mol respectively (Figure 5.25-D).

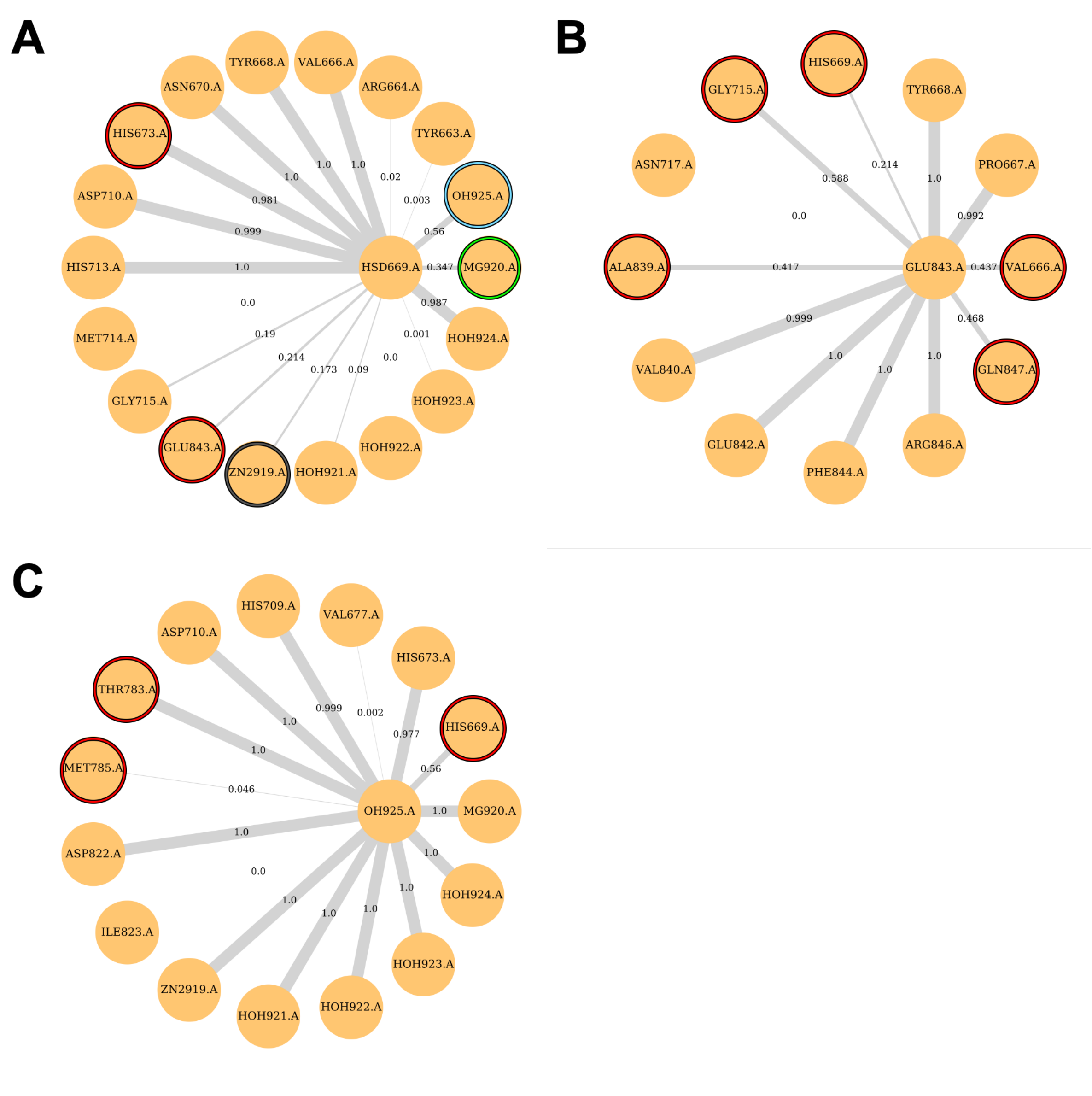


Figure 5.29: Dynamic residue contact maps around catalytic residues HSD669, E843, and the nucleophilic OH$^-$ ion of the TbrPDEB1 apo protein. The dynamic residue contact maps for A) catalytic residue HSD669, B) catalytic residue E843, and C) nucleophilic OH$^-$ ion. A red ring indicates an amino acid, a green ring the Mg^{2+} ion, a grey ring the Zn^{2+} ion, a light blue ring the OH$^-$ ion, and a dark blue ring a ligand atom.

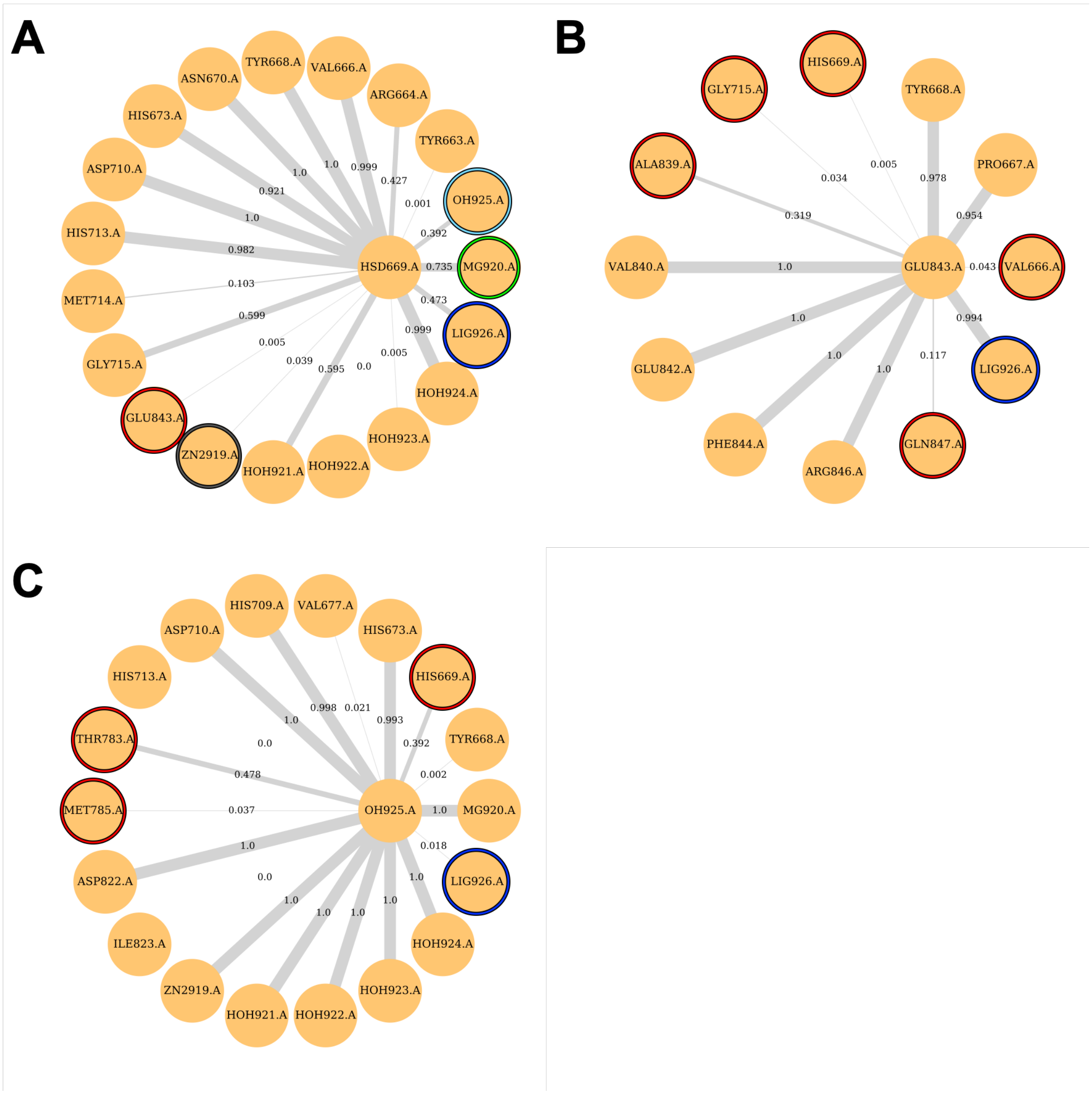


Figure 5.30: Dynamic residue contact maps around catalytic residues HSD669, E843, and the nucleophilic OH$^-$ ion of the TbrPDEB1-RUBi022 complex. The dynamic residue contact maps for A) catalytic residue HSD669, B) catalytic residue E843, and C) nucleophilic OH$^-$ ion. A red ring indicates an amino acid, a green ring the Mg^{2+} ion, a grey ring the Zn^{2+} ion, a light blue ring the OH$^-$ ion, and a dark blue ring a ligand atom.

In conclusion, from our dynamics studies, stable ligand interaction with H669 and Y668 especially in the form of hydrogen bonds appears to disrupt several functional residues. This can form a basis for computational screening of potential hit compounds before wet lab testing. These findings may be further investigated using a larger dataset of active compounds. We did not observe a strong correlation between the Autodock Vina binding free energy and the CHARMM interaction energy approximation. This is due in part to the fact that during docking the protein backbone is static, long range electrostatic effects are not taken into account and there are no water molecules in the system. Altogether this contributes to a poor description of H-bonding and poorly modelled system dynamics.

5.4 Conclusion

In this study, we derived CHARMM all-atom force-field parameters for the simulation of the zinc, magnesium and hydroxide containing bi-metallic catalytic center of TbrPDEB1. In our mixed bonded and non-bonded system, we tested different protonation states of residue H669 and bonding of D710 to the system. Our MD simulation studies show that the unprotonated H669 is essential for active site dynamics and a bonded coordinating ASP (D710) is essential for stabilizing the metal internuclear distance. Accurate description of protein catalytic centers is important for understanding the protein motions during MD simulation. For metalloproteins, failure to include accurate parameters that describe metal ion interactions can affect the modeling of the active site motions. As a result, several possibly druggable binding pocket conformations that are pharmacologically relevant might be missed during simulation. Parameters describing metal ion interactions are often very specific and depend on several factors (metal ion type, coordination spheres, charge, and geometrical arrangement). As such, they have to be tailored to each individual system and this includes: accurate calculation, testing and several optimizations.

The force-field parameters described herein in model 4 are able to accurately describe the geometry and dynamics of the TbrPDEB1 bi-metallic catalytic center as compared to the crystal structure. Further protein-ligand MD simulations using the calculated parameters were able to reveal important residue interactions between the TbrPDEB1 protein and ligands. This study lays the ground work for further use of MD simulations of TbrPDEB1 for drug discovery purposes.

CHAPTER 6

Conclusion and Future Directions

Conclusion

This thesis was presented in two parts under a broad theme of using computational based tools in HAT eradication efforts. Part I focused on host genetic determinants while Part II focused on the rational based discovery of novel chemotherapeutics.

6.1 A candidate Gene Association Study on a *Tbr* and *Tbg* HAT Endemic Cohort from Uganda

In this study we carried out a CGAS on Ugandan *Tbr* and *Tbg* HAT endemic cohorts to test the presence of genetic association within *IL10*, *IL8*, *IL4*, *HLAG*, *TNFA, TNX4LB*, *IL6, IFNG*, *MIF, APOL1*, *HLAA*, *IL1B*, *IL4R*, *IL12B*, *IL12R*, *HP*, *HPR*, and *CFH* candidate genes with HAT. In spite of some SNPs giving suggestive associations, none of the 6 SNPs passed correction for multiple testing. Interestingly the study found no association for the *APOL1* G2 allele that was previously reported to be strongly associated with protection against *Tbr* HAT. Further studies are required to determine the role of infection history and ethnicity on HAT outcome. Studies like Genome Wide Association Studies (GWAS) can be helpful in identifying undiscovered genetic associations as it does not require a prior hypothesis (candidate genes).

6.2 *In silico* analysis of the Effects of Missense SNVs on the Structure and Function of Macrophage Migration Inhibitory Factor (MIF)

The purpose of this study was to use MIF as a case study for how one could proceed to investigate the role of identified CGAS associated missense SNVs on protein structure and function. We used computational tools to investigate the effects of 27 missense SNVs on MIF protein structure and dynamics. This involved the use of *in silico* effect prediction tools, structural analyses, modeled the mutants and subjected them to molecular dynamic simulations and network analysis.

Our findings showed that mutations P2Q, I5M, P16Q, L23F, T24S, T31I, Y37H, H41P, M48V, P44L, G52C, S54R, I65M, I68T, S75F, N106S, and T113S caused pronounced protein conformational changes. We were also to demonstrate using DRN that P2, T31, Y37, G52, I65, I68, S75, N106, and T113S are part of a similar local residue interaction network with functional significance. These results will help compliment future mutation studies in the discovery of residues crucial to MIF function with the aim of drug discovery or investigating varying disease states.

6.3 Identification of Novel Pteridine Reductase 1 (PTR1) Inhibitors

In this study we discovered five potential *Tb*PTR1 inhibitors using a systematic computational based screening approach that started with 5742 compounds. The approach involved subjecting the top 18 binding modes identified using molecular docking to MD simulations. The trajectories were analysed for conformational and energetic stability. The analyses included RMSD, Rg, RMSF, PCA, EDA, DRN analysis, and Molecular mechanics Poisson–Boltzmann surface area (MMPBSA) free energy calculations. We report compounds RUBi004, RUBi007, RUBi014, RUBi16, and RUBi018 that exhibited anti-trypanosomal activity against trypanosomes in culture with IC_{50}s of 12.5 ± 4.8 μM, 32.4 ± 4.2 μM, 5.9 ± 1.4 μM, 28.2 ± 3.3 μM, and 9.7 ± 2.1 μM respectively. Compounds RUBi007, RUBi016 and RUBi018 showed no significant human cell cytotoxicity at 100 μM while RUBi004 and RUBi014 had cytotoxicity IC_{50}s of 23.6 ± 5.8 μM and 32.9 ± 2.2 μM respectively.

In order to test for dual inhibition of *Tb*PTR1 and *Tb*DHFR which is necessary for successful targeting of the Trypanosoma folate pathway the drugs were tested in combination with WR99210 which is a known DHFR inhibitor. RUBi004, RUBi007, RUBi014, and RUBi018, displayed antagonistic effects, while RUBi016 showed an additive effect in the isobologram. Of these compounds RUBi016 appears to be the most selective for *Tb*PTR1 while the others appear to compete for the *Tb*DHFR active site with WR99210. More specific enzyme-ligand activity relationship studies and structure determination or characterisation of appropriate protein-ligand complexes by crystallography are however required to validate these findings.

6.4 CHARMM Force-Field Parameterization of the Bi-metallic Catalytic Center of the *Trypanosoma brucei* Phosphodiesterase B1 (TbrPDEB1) for Computational Drug Discovery

In this study we successfully derived CHARMM force-field parameters needed to describe the TbrPDEB1 bi-metal center for MD simulations. Structure optimisations were carried out using a two-layer QM/MM ONIOM that yielded a similar active site geometry to the crystal structure. Parameters were derived from PES describing bond stretching, angle bending and dihedral bending. MD studies carried out for 30 ns at 310 K showed that our mixed bonded and non-bonded system was able to reproduce the active site geometry and dynamics well. MD studies further showed that the protonation of catalytic residue H669 is important for active site conformational stability and dynamics. Residue D710 which is among the metal bridging ligands plays an important role in regulating the metal internuclear distance.

These parameters were used to model the behaviour of potential TbrPDEB1 inhibitors in complex with the protein. Using the most potent ligand RUBi022 (IC_{50} = 14.96 μM) that was discovered using molecular docking, we illustrate that ligand interaction with H669, Y668, and Q847 is important for TbrPDEB1 inhibition. Further studies using a larger sample of potential TbrPDEB1 inhibitors may elucidate other important protein-ligand residue interactions.

6.5 Future work

The results presented in this thesis though interesting can be improved. Future work may include selecting a larger number of candidate genes, increasing the effectiveness of the cohort. A GWAS study to investigate SNPs that might be associated with HAT is underway by the TrypanoGEN project. Further, chain-chain interaction energy calculations can help shed light on the effects of the missense SNVs on the MIF protein oligomerisation. Scaffold merging of the *Tb*PTR1 inhibitors can be useful in improving the potency and specificity of RUBi016. While the TbrPDEB1 parameters were able to accurately describe the TbrPDEB1 bi-metal active site, they can be further improved by customising the CHARMM cut-off scheme so that derived LJ parameters can be applied to the Zn^{2+}-OH-Mg^{2+} system.

REFERENCES

1. Baldacchino F, Muenworn V, Desquesnes M, Desoli F, Charoenviriyaphap T, Duvallet G. Transmission of pathogens by *Stomoxys* flies (Diptera, Muscidae): a review. Parasite. 2013;
2. Bouteille B, Buguet A. Human African Trypanosomiasis. In: Encyclopedia of the Neurological Sciences. 2014.
3. Franco JR, Simarro PP, Diarra A, Jannin JG. Epidemiology of human African trypanosomiasis. Vol. 6, Clinical Epidemiology. 2014. p. 257–75.
4. Rocha G, Martins A, Gama G, Brandão F, Atouguia J. Possible cases of sexual and congenital transmission of sleeping sickness [5]. Lancet. 2004.
5. Barrett MP, Burchmore RJS, Stich A, Lazzari JO, Frasch AC, Cazzulo JJ, et al. The trypanosomiases. In: Lancet. 2003. p. 1469–80.
6. Bozza MT, Martins YC, Carneiro LAM, Paiva CN, Achcar F, Kerkhoven EJ, et al. Trypanosoma brucei: Meet the system. Curr Opin Microbiol. 2013;3(November):78.
7. Franco JR, Simarro PP, Diarra A, Jannin JG. Epidemiology of human African trypanosomiasis. Clin Epidemiol. 2014;6:257–75.
8. Funk S, Nishiura H, Heesterbeek H, Edmunds WJ, Checchi F. Identifying Transmission Cycles at the Human-Animal Interface: The Role of Animal Reservoirs in Maintaining Gambiense Human African Trypanosomiasis. PLoS Comput Biol. 2013;9(1).
9. Lumbala C, Simarro PP, Cecchi G, Paone M, Franco JR, Kande Betu Ku Mesu V, et al. Human African trypanosomiasis in the Democratic Republic of the Congo: disease distribution and risk. Int J Health Geogr. 2015;14(1):20.
10. Simarro PP, Cecchi G, Paone M, Franco JR, Diarra A, Ruiz JA, et al. The Atlas of

human African trypanosomiasis: a contribution to global mapping of neglected tropical diseases. Int J Health Geogr. 2010;9(1):57.

11. World Health Organization (WHO). Trypanosomiasis, human African (sleeping sickness). WHO fact sheets. 2018.

12. Oue E, Dama E, Drabo A, Kabore J, Compaore CF, Coulibaly B, et al. Description of the first sleeping sickness case diagnosed in Burkina Faso since two decades. 2018;1–9.

13. Kimuda MP, Noyes H, Mulindwa J, Enyaru J, Alibu VP, Sidibe I, et al. No evidence for association between APOL1 kidney disease risk alleles and Human African Trypanosomiasis in two Ugandan populations. PLoS Negl Trop Dis. 2018;

14. Dyer N a., Rose C, Ejeh NO, Acosta-Serrano A. Flying tryps: Survival and maturation of trypanosomes in tsetse flies. Trends Parasitol. 2013;29(4):188–96.

15. Matovu E, Seebeck T, Enyaru JCK, Kaminsky R. Drug resistance in Trypanosoma brucei spp., the causative agents of sleeping sickness in man and nagana in cattle. Microbes Infect. 2001;3(9):763–70.

16. Matovu E, Enyaru JCK, Legros D, Schmid C, Seebeck T, Kaminsky R. Melarsoprol refractory T. b. gambiense from Omugo, north-western Uganda. Trop Med Int Heal. 2001;6(5):407–11.

17. World Health Organization. WHO | Trypanosomiasis, human African (sleeping sickness). WHO. 2016.

18. Doua F, Miezan TW, Sanon Singaro JR, Boa Yapo F, Baltz T. The efficacy of pentamidine in the treatment of early-late stage Trypanosoma brucei gambiense trypanosomiasis. Am J Trop Med Hyg. 1996;55(6):586–8.

19. Brun R, Don R, Jacobs RT, Wang MZ, Barrett MP. Development of novel drugs for human African trypanosomiasis. Future Microbiol. 2011;6(6):677–91.

20. Steverding D. The development of drugs for treatment of sleeping sickness: A historical review. Parasites and Vectors. 2010.
21. Balasegaram M, Harris S, Checchi F, Ghorashian S, Hamel C, Karunakara U. Melarsoprol versus eflornithine for treating late-stage Gambian trypanosomiasis in the Republic of the Congo. Bull World Health Organ. 2006;
22. Boiani M, Piacenza L, Hernández P, Boiani L, Cerecetto H, González M, et al. Mode of action of Nifurtimox and N-oxide-containing heterocycles against Trypanosoma cruzi: Is oxidative stress involved? Biochem Pharmacol. 2010;
23. Torreele E, Trunz BB, Tweats D, Kaiser M, Brun R, Mazué G, et al. Fexinidazole - a new oral nitroimidazole drug candidate entering clinical development for the treatment of sleeping sickness. PLoS Negl Trop Dis. 2010;
24. Mesu VKBK, Kalonji WM, Bardonneau C, Mordt OV, Blesson S, Simon F, et al. Oral fexinidazole for late-stage African Trypanosoma brucei gambiense trypanosomiasis: A pivotal multicentre, randomised, non-inferiority trial. The Lancet. 2017;
25. Pelfrene E, Harvey Allchurch M, Ntamabyaliro N, Nambasa V, Ventura F V., Nagercoil N, et al. The European Medicines Agency's scientific opinion on oral fexinidazole for human African trypanosomiasis. PLoS Negl Trop Dis. 2019;
26. Jacobs RT, Nare B, Wring SA, Orr MD, Chen D, Sligar JM, et al. Scyx-7158, an orally-active benzoxaborole for the treatment of stage 2 human african trypanosomiasis. PLoS Negl Trop Dis. 2011;
27. Jamonneau V, Ilboudo H, Kaboré J, Kaba D, Koffi M, Solano P, et al. Untreated human infections by trypanosoma brucei gambiense are not 100% fatal. PLoS Negl Trop Dis. 2012;6(6).
28. Bucheton B, Macleod A, Jamonneau V. Human host determinants influencing the outcome of Trypanosoma brucei gambiense infections. Vol. 33, Parasite Immunology.

2011. p. 438–47.

29. Garcia A, Courtin D, Solano P, Koffi M, Jamonneau V. Human African trypanosomiasis: connecting parasite and host genetics. Trends Parasitol. 2006;22(9):405–9.

30. Sternberg JM, Maclean L. A spectrum of disease in human African trypanosomiasis: the host and parasite genetics of virulence. Parasitology. 2010;137(14):2007–15.

31. Naessens J. Bovine trypanotolerance: A natural ability to prevent severe anaemia and haemophagocytic syndrome? Vol. 36, International Journal for Parasitology. 2006. p. 521–8.

32. Kimuda MP. In silico analysis of the effects of non-synonymous single nucleotide polymorphisms on the Human Macrophage Migration Inhibitory Factor gene and their possible role in Human African Trypanosomiasis susceptibility. (Unpublished Master's thesis), Rhodes Univ South Africa. 2016;(January).

33. Simarro PP, Diarra A, Ruiz Postigo JA, Franco JR, Jannin JG, Postigo JAR, et al. The Human African trypanosomiasis control and surveillance programme of the World Health Organization 2000-2009: the way forward. PLoS Negl Trop Dis. 2011;5(2):e1007.

34. Report of the first WHO stakeholders meeting on gambiense human African trypanosomiasis elimination. 2014;(March):25–7.

35. Manrai AK, Patel CJ, Ioannidis JPA. In the era of Precisionmedicine and big data, Who is normal? JAMA - Journal of the American Medical Association. 2018.

36. Stijlemans B, Leng L, Brys L, Sparkes A, Vansintjan L, Caljon G, et al. MIF Contributes to Trypanosoma brucei Associated Immunopathogenicity Development. PLoS Pathog. 2014;10(9):e1004414.

37. Courtin D, Milet J, Sabbagh A, Massaro JD, Castelli EC, Jamonneau V, et al. HLA-G

3’ UTR-2 haplotype is associated with Human African trypanosomiasis susceptibility. Infect Genet Evol. 2013;17:1–7.

38. Courtin D, Milet J, Jamonneau V, Yeminanga CS, Kumeso VKB, Bilengue CMM, et al. Association between human African trypanosomiasis and the IL6 gene in a Congolese population. Infect Genet Evol. 2007;7(1):60–8.

39. Ofon E, Noyes H, Mulindwa J, Ilboudo H, Simuunza M, Ebo’o V, et al. A polymorphism in the haptoglobin, haptoglobin related protein locus is associated with risk of human sleeping sickness within Cameroonian populations. PLoS Negl Trop Dis. 2017;11(10).

40. Cooper A, Ilboudo H, Alibu VP, Ravel S, Enyaru J, Weir W, et al. APOL1 renal risk variants have contrasting resistance and susceptibility associations with African trypanosomiasis renal risk variants have contrasting resistance and susceptibility associations with African trypanosomiasis. Elife. 2017;6:1–18.

41. Ahouty B, Koffi M, Ilboudo H, Simo G, Matovu E, Mulindwa J, et al. Candidate genes-based investigation of susceptibility to Human African Trypanosomiasis in Côte d’Ivoire. PLoS Negl Trop Dis. 2017;

42. Kaboré JW, Ilboudo H, Noyes H, Camara O, Kaboré J, Camara M, et al. Candidate gene polymorphisms study between human African trypanosomiasis clinical phenotypes in Guinea. PLoS Negl Trop Dis. 2017;11(8).

43. Jamonneau V, Garcia A, Ravel S, Cuny G, Oury B, Solano P, et al. Genetic characterization of Trypanosoma brucei gambiense and clinical evolution of human African trypanosomiasis in Cote d’Ivoire. Trop Med Int Health. 2002;7(7):610–21.

44. Jamonneau V, Ravel S, Garcia a, Koffi M, Truc P, Laveissière C, et al. Characterization of Trypanosoma brucei s.l. infecting asymptomatic sleeping-sickness patients in Côte d’Ivoire: a new genetic group? Ann Trop Med Parasitol.

2004;98(4):329–37.

45. Goodhead I, Capewell P, Wendi Bailey J, Beament T, Chance M, Kay S, et al. Whole-genome sequencing of Trypanosoma brucei reveals introgression between subspecies that is associated with virulence. MBio. 2013;4(4).

46. Brookes AJ. The essence of SNPs. Gene. 1999.

47. Serre D, Pääbo S. Evidence for gradients of human genetic diversity within and among continents. Genome Res. 2004;

48. Orr N, Chanock S. Chapter 1 Common Genetic Variation and Human Disease. Advances in Genetics. 2008.

49. Roden DM, Wilke RA, Kroemer HK, Stein CM. Pharmacogenomics: The genetics of variable drug responses. Circulation. 2011;

50. Loewe L, Hill WG. The population genetics of mutations: Good, bad and indifferent. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010.

51. Cotton RGH, Scriver CR. Proof of “disease causing” mutation. Human Mutation. 1998.

52. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. 2006;

53. Tomoko O. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol. 1995;

54. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–11.

55. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008.

56. Pauling L, Itano HA, Singer SJ, Wells IC. Sickle cell anemia, a molecular disease. Science (80-). 1949;

57. Steinberg MH, Sebastiani P. Genetic modifiers of sickle cell disease. Vol. 87, American Journal of Hematology. 2012. p. 795–803.

58. Cordell HJ, Clayton DG. Genetic association studies. Lancet. 2005.

59. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005;6(2):95–108.

60. Patnala R, Clements J, Batra J. Candidate gene association studies: a comprehensive guide to useful in silico tools. BMC Genet. 2013;14:39.

61. Amos W, Driscoll E, Hoffman JI. Candidate genes versus genome-wide associations: Which are better for detecting genetic susceptibility to infectious disease? Proc R Soc B Biol Sci. 2011;

62. Reich DE, Cargili M, Boik S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. Nature. 2001;

63. Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case-control association studies. Nat Protoc. 2007;2(10):2492–501.

64. Mayo O. A century of Hardy-Weinberg equilibrium. Twin Research and Human Genetics. 2008.

65. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev. 2010;11(June):459–63.

66. Upton GJG. Fisher's Exact Test. J R Stat Soc. 1992;155(3):395–402.

67. Dunnett CW. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. J Am Stat Assoc. 1955;

68. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Vol. 57, Journal of the Royal Statistical Society. 1995. p. 289–300.

69. Dummer PD, Limou S, Rosenberg AZ, Heymann J, Nelson G, Winkler CA, et al.

APOL1 Kidney Disease Risk Variants: An Evolving Landscape. Seminars in Nephrology. 2015.

70. Zipfel PF, Hallstr??m T, Riesbeck K. Human complement control and complement evasion by pathogenic microbes - Tipping the balance. Vol. 56, Molecular Immunology. 2013. p. 152–60.

71. Luo S, Poltermann S, Kunert A, Rupp S, Zipfel PF. Immune evasion of the human pathogenic yeast Candida albicans: Pra1 is a Factor H, FHL-1 and plasminogen binding surface protein. Mol Immunol. 2009;47(2–3):541–50.

72. Kennedy AT, Schmidt CQ, Thompson JK, Weiss GE, Taechalertpaisarn T, Gilson PR, et al. Recruitment of Factor H as a Novel Complement Evasion Strategy for Blood-Stage Plasmodium falciparum Infection. J Immunol. 2016;196(3):1239–48.

73. Tiberti N, Hainard A, Lejon V, Robin X, Ngoyi DM, Turck N, et al. Discovery and verification of osteopontin and Beta-2-microglobulin as promising markers for staging human African trypanosomiasis. Mol Cell Proteomics. 2010;9(12):2783–95.

74. Amiot L, Vu N, Samson M. Immunomodulatory properties of HLA-G in infectious diseases. Vol. 2014, Journal of Immunology Research. 2014.

75. Donadi EA, Castelli EC, Arnaiz-Villena A, Roger M, Rey D, Moreau P. Implications of the polymorphism of HLA-G on its function, regulation, evolution and disease association. Cell Mol Life Sci. 2011;68(3):369–95.

76. Matte C, Lacaille J, Zijenah L, Ward B, Roger M, Chidawanyika H, et al. HLA-G and HLA-E polymorphisms in an indigenous african population. In: Human Immunology. 2000. p. 1150–6.

77. Lyke KE, Fernández-Viňa MA, Cao K, Hollenbach J, Coulibaly D, Kone AK, et al. Association of HLA alleles with Plasmodium falciparum severity in Malian children. Tissue Antigens. 2011;77(6):562–71.

78. Hardwick RJ, Ménard A, Sironi M, Milet J, Garcia A, Sese C, et al. Haptoglobin (HP) and Haptoglobin-related protein (HPR) copy number variation, natural selection, and trypanosomiasis. Hum Genet. 2014;133(1):69–83.

79. Bakhiet M, Jansson L, Buscher P, Holmdahl R, Kristensson K, Olsson T. Control of parasitemia and survival during Trypanosoma brucei brucei infection is related to strain-dependent ability to produce IL-4. J Immunol. 1996;157(8):3518–26.

80. Alvarado Arnez LE, Venegas EN, Ober C, Thompson EE. Sequence variation in the IL4 gene and resistance to Trypanosoma cruzi infection in Bolivians. J Allergy Clin Immunol. 2011;127(1).

81. Sternberg JM, Rodgers J, Bradley B, MacLean L, Murray M, Kennedy PGE. Meningoencephalitic African trypanosomiasis: Brain IL-10 and IL-6 are associated with protection from neuro-inflammatory pathology. J Neuroimmunol. 2005;167(1–2):81–9.

82. Ng DPK, Nurbaya S, Ye SHJ, Krolewski AS. An IL-6 haplotype on human chromosome 7p21 confers risk for impaired renal function in type 2 diabetic patients. Kidney Int. 2008;74(4):521–7.

83. Lejon V, Lardon J, Kenis G, Pinoges L, Legros D, Bisser S, et al. Interleukin (IL)-6, IL-8 and IL-10 in serum and CSF of Trypanosoma brucei gambiense sleeping sickness patients before and after treatment. Trans R Soc Trop Med Hyg. 2002;96(3):329–33.

84. Ilboudo H, Bras-Gonçalves R, Camara M, Flori L, Camara O, Sakande H, et al. Unravelling human trypanotolerance: IL8 is associated with infection control whereas IL10 and TNFα are associated with subsequent disease development. PLoS Pathog. 2014;10(11):e1004469.

85. Courtin D, Argiro L, Jamonneau V, N'dri L, N'quessan P, Abel L, et al. Interest of tumor necrosis factor-alpha - 308 G/A and interleukin-10 -592 C/A polymorphisms in

human African trypanosomiasis. Infect Genet Evol. 2006;6(2):123–9.

86. Sortica VA, Cunha MG, Ohnishi MDO, Souza JM, Ribeiro-Dos-Santos AKC, Santos NPC, et al. IL1B, IL4R, IL12RB1 and TNF gene polymorphisms are associated with Plasmodium vivax malaria in Brazil. Malar J. 2012;11:409.

87. Hertz CJ, Filutowicz H, Mansfield JM. Resistance to the African trypanosomes is IFN-gamma dependent. JImmunol. 1998;161(0022–1767 (Print)):6775–83.

88. Naessens J, Kitani H, Nakamura Y, Yagi Y, Sekikawa K, Iraqi F. TNF-α mediates the development of anaemia in a murine Trypanosoma brucei rhodesiense infection, but not the anaemia associated with a murine Trypanosoma congolense infection. Clin Exp Immunol. 2005;139(3):405–10.

89. Gichohi-Wainaina WN, Melse-Boonstra A, Feskens EJ, Demir AY, Veenemans J, Verhoef H. Tumour necrosis factor allele variants and their association with the occurrence and severity of malaria in African children: a longitudinal study. Malar J. 2015;14:249.

90. Holmgren A, Lu J. Thioredoxin and thioredoxin reductase: Current research with special reference to human disease. Biochem Biophys Res Commun. 2010;

91. Reckenfelderbaumer N, Ludemann H, Schmidt H, Steverding D, Krauth-Siegel RL. Identification and functional characterization of thioredoxin from Trypanosoma brucei brucei. J Biol Chem. 2000;275(11):7547–52.

92. Krauth-Siegel RL, Meiering SK, Schmidt H. The parasite-specific trypanothione metabolism of trypanosoma and leishmania. Biological Chemistry. 2003.

93. Fairlamb A. Metabolism and Functions of Trypanothione in the Kinetoplastida. Annu Rev Microbiol. 1992;

94. Burke-Gaffney A, Callister MEJ, Nakamura H. Thioredoxin: Friend or foe in human disease? Trends in Pharmacological Sciences. 2005.

95. Stijlemans B, Beschin A, Magez S, Van Ginderachter JA, De Baetselier P. Iron Homeostasis and Trypanosoma brucei Associated Immunopathogenicity Development: A Battle/Quest for Iron. Biomed Res Int. 2015;2015:1–15.

96. Stijlemans B, Vankrunkelsven A, Caljon G, Bockstal V, Guilliams M, Bosschaerts T, et al. The central role of macrophages in trypanosomiasis-associated anemia: rationale for therapeutical approaches. Endocr Metab Immune Disord Drug Targets. 2010;10(1):71–82.

97. Calandra T, Roger T. Macrophage migration inhibitory factor: a regulator of innate immunity. Nat Rev Immunol. 2003;3(10):791–800.

98. Larson DF, Horak K. Macrophage migration inhibitory factor: controller of systemic inflammation. Crit Care. 2006;10(2):138.

99. Rosado JDD, Rodriguez-Sosa M. Macrophage migration inhibitory factor (MIF): A Key player in protozoan infections. Int J Biol Sci. 2011;7(9):1239–56.

100. L.E. L, M. B, W. H, S.U. W, J. S, S. K, et al. A MIF haplotype is associated with the outcome of patients with severe sepsis: A case control study. J Transl Med. 2009;7:no pagination-no pagination.

101. Chitnis S, Monteiro J, Lin X, Wolfe F, Bucala R. A functional promoter polymorphism in the macrophage migration inhibitory factor (MIF) gene associated with disease severity in rheumatoid arthritis. Genes Immun. 2002;170–6.

102. Wu SP, Leng L, Feng Z, Liu N, Zhao H, McDonald C, et al. Macrophage migration inhibitory factor promoter polymorphisms and the clinical expression of scleroderma. Arthritis Rheum. 2006;54(11):3661–9.

103. Das R, Koo M-S, Kim BH, Jacob ST, Subbian S, Yao J, et al. Macrophage migration inhibitory factor (MIF) is a critical mediator of the innate immune response to Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2013;110(32):E2997-3006.

104. Donn R, Alourfi Z, Zeggini E, Lamb R, Jury F, Lunt M, et al. A Functional Promoter Haplotype of Macrophage Migration Inhibitory Factor Is Linked and Associated with Juvenile Idiopathic Arthritis. Arthritis Rheum. 2004;50(5):1604–10.

105. Sreih A, Ezzeddine R, Leng L, Lachance A, Yu G, Mizue Y, et al. Dual effect of the macrophage migration inhibitory factor gene on the development and severity of human systemic lupus erythematosus. Arthritis Rheum. 2011;63(12):3942–51.

106. Bernhagen J, Krohn R, Lue H, Gregory JL, Zernecke A, Koenen RR, et al. MIF is a noncognate ligand of CXC chemokine receptors in inflammatory and atherogenic cell recruitment. Nat Med. 2007;13(5):587–96.

107. Zhao JH. gap: Genetic Analysis Package. J Stat Softw. 2007;23(8):1–18.

108. Harrington JM, Nishanova T, Pena SR, Hess M, Scelsi CL, Widener J, et al. A retained secretory signal peptide mediates High Density Lipoprotein (HDL) assembly and function of haptoglobin-related protein. J Biol Chem. 2014;289(36):24811–20.

109. Muranjan M, Nussenzweig V, Tomlinson S. Characterization of the human serum trypanosome toxin, haptoglobin-related protein. J Biol Chem. 1998;273(7):3884–7.

110. Ilboudo H, Berthier D, Camara M, Camara O, Kabore J, Leno M, et al. APOL1 expression is induced by Trypanosoma brucei gambiense infection but is not associated with differential susceptibility to sleeping sickness. Infect Genet Evol. 2012;12(7):1519–23.

111. Cuypers B, Lecordier L, Meehan CJ, Van den Broeck F, Imamura H, Büscher P, et al. Apolipoprotein L1 Variant Associated with Increased Susceptibility to Trypanosome Infection. MBio. 2016;7(2):1–8.

112. Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, et al. Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. Science (80-). 2010;329(5993):841–5.

113. Figueroa JE, Densen P. Infectious diseases associated with complement deficiencies. Clin Microbiol Rev. 1991;4(3):359–95.

114. Simon N, Scheuermayer M, Sologub L, Zipfel PF, Skerka C, Pradel G. Malaria parasites hijack human factor H to protect from complement-mediated lysis in the mosquito midgut. Int J Med Microbiol. 2011;Conference:14.

115. Das R, Koo M-S, Kim BH, Jacob ST, Subbian S, Yao J, et al. Macrophage migration inhibitory factor (MIF) is a critical mediator of the innate immune response to Mycobacterium tuberculosis. Proc Natl Acad Sci U S A. 2013;110(32):E2997-3006.

116. de Jesus Fernandes Covas C, Cardoso CC, Gomes-Silva A, Oliveira JRS, Da-Cruz AM, Moraes MO. Candidate gene case-control and functional study shows macrophage inhibitory factor (MIF) polymorphism is associated with cutaneous leishmaniasis. Cytokine. 2013;61(1):168–72.

117. Jha AN, Sundaravadivel P, Pati SS, Patra PK, Thangaraj K. Variations in ncRNA gene LOC284889 and MIF-794CATT repeats are associated with malaria susceptibility in Indian populations. Malar J. 2013;12(1):345.

118. Maclean L, Odiit M, Macleod A, Morrison L, Sweeney L, Cooper A, et al. Spatially and genetically distinct African Trypanosome virulence variants defined by host interferon-gamma response. J Infect Dis. 2007;196(11):1620–8.

119. MacLean L, Odiit M, Sternberg JM. Intrathecal cytokine responses in Trypanosoma brucei rhodesiense sleeping sickness patients. Trans R Soc Trop Med Hyg. 2006;100(3):270–5.

120. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1).

121. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al.

PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

122. Semagn K, Babu R, Hearne S, Olsen M. Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. 2014;1–14.

123. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.

124. Ramasamy R, Ramasamy S, Bindroo B, Naik V. STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. Springerplus. 2014;3(1):431.

125. Anderson C a, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5(9):1564–73.

126. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

127. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;

128. Rodriguez S, Gaunt TR, Day INM. Hardy-Weinberg equilibrium testing of biological ascertainment for Mendelian randomization studies. Am J Epidemiol. 2009;

129. MacLean L, Chisi JE, Odiit M, Gibson WC, Ferris V, Picozzi K, et al. Severity of human African trypanosomiasis in East Africa is associated with geographic location, parasite genotype, and host inflammatory cytokine response profile. Infect Immun. 2004;72(12):7040–4.

130. Van Den Oord EJCG. Controlling false discoveries in genetic studies. Am J Med Genet B Neuropsychiatr Genet. 2008;147B(5):637–44.

131. Kato CD, Matovu E, Mugasa CM, Nanteza A, Alibu VP. The role of cytokines in the pathogenesis and staging of Trypanosoma brucei rhodesiense sleeping sickness. Allergy Asthma Clin Immunol. 2016;12:4.

132. Brown DK, Penkler DL, Sheik Amamuddy O, Ross C, Atilgan AR, Atilgan C, et al. MD-TASK: a software suite for analyzing molecular dynamics trajectories. Bioinformatics. 2017 May;

133. Simpson KD, Templeton DJ, Cross J V. Macrophage migration inhibitory factor promotes tumor growth and metastasis by inducing myeloid-derived suppressor cells in the tumor microenvironment. J Immunol. 2012;189(12):5533–40.

134. Churchill Jr. WH, Piessens WF, Sulis CA, David JR. Macrophages activated as suspension cultures with lymphocyte mediators devoid of antigen become cytotoxic for tumor cells. J Immunol. 1975;

135. Burger-Kentischer A, Goebel H, Seiler R, Fraedrich G, Schaefer HE, Dimmeler S, et al. Expression of macrophage migration inhibitory factor in different stages of human atherosclerosis. Circulation. 2002;105(1524–4539 (Electronic)):1561–6.

136. Mitchell RA, Metz CN, Peng T, Bucala R. Sustained mitogen-activated protein kinase (MAPK) and cytoplasmic phospholipase A2 activation by macrophage migration inhibitory factor (MIF): Regulatory role in cell proliferation and glucocorticoid action. J Biol Chem. 1999;

137. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. MutPred: Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744–50.

138. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, et al. PredictSNP:

Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. PLoS Comput Biol. 2014;10(1).

139. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.

140. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015;31(April):btv195.

141. Tang H, Thomas PD. PANTHER-PSEP: Predicting disease-causing genetic variants using position-specific evolutionary preservation. Bioinformatics. 2016;32(14):2230–2.

142. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005;15(7):978–86.

143. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006;22(22):2729–34.

144. Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics. 2013;14 Suppl 3(Suppl 3):S6.

145. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. In: Current Protocols in Human Genetics. 2013. p. 7.20.1-7.20.41.

146. Bromberg Y, Rost B. SNAP: Predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35(11):3823–35.

147. Krieger E, Nabuurs SB, Vriend G. Homology Modeling. In: Structural Bioinformatics. 2005.

148. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence - A study of structural response in protein cores. Proteins Struct Funct Bioinforma. 2009;

149. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M-Y, et al. Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci. 2007;Chapter 2:Unit 2.9.

150. Fiser A, Sali A. ModLoop: Automated modeling of loops in protein structures. Bioinformatics. 2003;19(18):2500–1.

151. Wiederstein M, Sippl MJ. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007;35(SUPPL.2).

152. Eisenberg D, Lüthy R, Bowie JU. VERIFY3D: Assessment of protein models with three-dimensional profiles. Methods Enzymol. 1997;277:396–406.

153. Laskowski R a., MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. 1993;26(November):283–91.

154. Alder BJ, Wainwright TE. Studies in molecular dynamics. I. General method. J Chem Phys. 1959;

155. Piana S, Klepeis JL, Shaw DE. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Curr Opin Struct Biol. 2014;

156. Brooks BR, Brooks III CL, Mackerell Jr. AD, Nilsson L, Petrella RJ, Roux B, et al. CHARMM: The Biomolecular Simulation Program. J Comput Chem. 2009;

157. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. Vol. 26, Journal of Computational Chemistry.

2005. p. 1701–18.

158. Phillips G. Describing protein conformational ensembles: beyond static snapshots. F1000 Biol Rep. 2009;

159. Adcock SA, McCammon JA. Molecular dynamics: Survey of methods for simulating the activity of proteins. Chemical Reviews. 2006.

160. Stephens DE, Khan FI, Singh P, Bisetty K, Singh S, Permaul K. Creation of thermostable and alkaline stable xylanase variants by DNA shuffling. J Biotechnol. 2014;187:139–46.

161. Khan FI, Shahbaaz M, Bisetty K, Waheed A, Sly WS, Ahmad F, et al. Large scale analysis of the mutational landscape in ??-glucuronidase: A major player of mucopolysaccharidosis type VII. Gene. 2016;576(1):36–44.

162. David CC, Jacobs DJ. Principal component analysis: A method for determining the essential dynamics of proteins. Methods Mol Biol. 2014;

163. Kitao A, Go N. Investigating protein dynamics in collective coordinate space. Current Opinion in Structural Biology. 1999.

164. Amadei A, Linssen ABM, Berendsen HJC. Essential dynamics of proteins. Proteins Struct Funct Bioinforma. 1993;17(4):412–25.

165. Bakan A, Meireles LM, Bahar I. ProDy: Protein dynamics inferred from theory and experiments. Bioinformatics. 2011;

166. Ross C, Nizami B, Glenister M, Sheik Amamuddy O, Atilgan AR, Atilgan C, et al. MODE-TASK: large-scale protein motion tools. Bioinformatics. 2018;

167. Doshi U, Holliday MJ, Eisenmesser EZ, Hamelberg D. Dynamical network of residue-residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. Proc Natl Acad Sci U S A. 2016;113(17):4735–40.

168. Csárdi G, Nepusz T. The igraph software package for complex network research. Vol.

1695, InterJournal Complex Systems. 2006. p. 1695.

169. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. Proc 7th Python Sci Conf. 2008;

170. Ozbaykal G, Rana Atilgan A, Atilgan C. In silico mutational studies of Hsp70 disclose sites with distinct functional attributes. Proteins Struct Funct Bioinforma. 2015 Nov;83(11):2077–90.

171. Di Marino D, D'Annessa I, Coletta A, Via A, Tramontano A. Characterization of the differences in the cyclopiazonic acid binding mode to mammalian and P. Falciparum Ca2+ pumps: A computational study. Proteins Struct Funct Bioinforma. 2015;83(3):564–74.

172. Brown DK, Sheik Amamuddy O, Tastan Bishop Ö. Structure-Based Analysis of Single Nucleotide Variants in the Renin-Angiotensinogen Complex. Global Heart. 2017;

173. Pantouris G, Ho J, Shah D, Syed MA, Leng L, Bhandari V, et al. Nanosecond Dynamics Regulate the MIF-Induced Activity of CD74. Angew Chemie - Int Ed. 2018;1–5.

174. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.

175. The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):D204-12.

176. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–42.

177. Crichlow G V, Lubetsky JB, Leng L, Bucala R, Lolis EJ. Structural and kinetic analyses of macrophage migration inhibitory factor active site interactions. Biochemistry. 2009;48(1):132–9.

178. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234(3):779–815.

179. Eswar N, Webb B, Marti-Renom M a, Madhusudhan MS, Eramian D, Shen M-Y, et al. Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci. 2007;Chapter 2:Unit 2.9.

180. DeLano WL. The PyMOL Molecular Graphics System. Schrödinger LLC wwwpymolorg. 2002;Version 1.:http://www.pymol.org.

181. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, et al. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys J. 2015;

182. Hauck T. Scikit-learn Cookbook. Packt Publishing. 2014.

183. Pedregosa F, Weiss R, Brucher M. Scikit-learn : Machine Learning in Python. J Mach Learn Res. 2011;

184. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;

185. Sun HW, Bernhagen J, Bucala R, Lolis E. Crystal structure at 2.6-A resolution of human macrophage migration inhibitory factor. Proc Natl Acad Sci U S A. 1996;93(11):5191–6.

186. Schwartz V, Lue H, Kraemer S, Korbiel J, Krohn R, Ohl K, et al. A functional heteromeric MIF receptor formed by CD74 and CXCR4. FEBS Lett. 2009;583(17):2749–57.

187. Fingerle-Rowson G, Kaleswarapu DR, Schlander C, Kabgani N, Brocks T, Reinart N, et al. A tautomerase-null macrophage migration-inhibitory factor (MIF) gene knock-in mouse model reveals that protein interactions and not enzymatic activity mediate MIF-dependent growth regulation. Mol Cell Biol. 2009;29(7):1922–32.

188. Kraemer S, Lue H, Zernecke A, Kapurniotu A, Andreetto E, Frank R, et al. MIF-

chemokine receptor interactions in atherogenesis are dependent on an N-loop-based 2-site binding mechanism. FASEB J. 2011;25(3):894–906.

189. Rajasekaran D, Gröning S, Schmitz C, Zierow S, Drucker N, Bakou M, et al. Macrophage migration inhibitory factor-CXCR4 receptor interactions: Evidence for partial allosteric agonism in comparison with CXCL12 chemokine. J Biol Chem. 2016;

190. Pantouris G, Syed MA, Fan C, Rajasekaran D, Cho TY, Rosenberg EM, et al. An Analysis of MIF Structural Features that Control Functional Activation of CD74. Chem Biol. 2015;22(9):1197–205.

191. Meza-Romero R, Benedek G, Leng L, Bucala R, Vandenbark AA. Predicted structure of MIF/CD74 and RTL1000/CD74 complexes. Metab Brain Dis. 2016;31(2):249–55.

192. Mischke R, Gessner A, Kapurniotu A, Jüttner S, Kleemann R, Brunner H, et al. Structure activity studies of the cytokine macrophage migration inhibitory factor (MIF) reveal a critical role for its carboxy terminus. FEBS Lett. 1997;414(2):226–32.

193. El-turk F, Cascella M, Ouertatani-sakouhi H, Lakshmi R, Leng L, Bucala R, et al. The conformational flexibility of the carboxy terminal residues 105–114 is a key modulator of the catalytic activity and stability of Macrophage Migration Inhibitory Factor (MIF). Biochemistry. 2008;47(40):10740–56.

194. Weber C, Kraemer S, Drechsler M, Lue H, Koenen RR, Kapurniotu A, et al. Structural determinants of MIF functions in CXCR2-mediated inflammatory and atherogenic leukocyte recruitment. Proc Natl Acad Sci. 2008;

195. Swope M, Sun HW, Blake PR, Lolis E. Direct link between cytokine activity and a catalytic site for macrophage migration inhibitory factor. EMBO J. 1998;17(13):3534–41.

196. Lubetsky JB, Swope M, Dealwis C, Blake P, Lolis E, February R V, et al. Pro-1 of Macrophage Migration Inhibitory Factor Functions as a Catalytic Base in the

Phenylpyruvate Tautomerase Activity. 1999;7346–54.

197. Rajasekaran D, Gröning S, Schmitz C, Zierow S, Drucker N, Bakou M, et al. Macrophage Migration Inhibitory Factor-CXCR4 Receptor Interactions. J Biol Chem. 2016;291(30):15881–95.

198. Lacy M, Kontos C, Brandhofer M, Hille K, Gröning S, Sinitski D, et al. Identification of an Arg-Leu-Arg tripeptide that contributes to the binding interface between the cytokine MIF and the chemokine receptor CXCR4. Sci Rep. 2018;8(1).

199. Stamps SL, Taylor AB, Wang SC, Hackert ML, Whitman CP. Mechanism of the phenylpyruvate tautomerase activity of macrophage migration inhibitory factor: Properties of the P1G, P1A, Y95F, and N97A mutants. Biochemistry. 2000;39(32):9671–8.

200. Kleemann R, Kapurniotu A, Mischke R, Held J, Bernhagen J. Characterization of catalytic centre mutants of macrophage migration inhibitory factor (MIF) and comparison to Cys81Ser MIF. Eur J Biochem. 1999;261(3):753–66.

201. Kleemann R, Kapurniotu A, Frank RW, Gessner A, Mischke R, Flieger O, et al. Disulfide analysis reveals a role for macrophage migration inhibitory factor (MIF) as thiol-protein oxidoreductase. J Mol Biol. 1998;280(1):85–102.

202. Wang Y, An R, Umanah GK, Park H, Nambiar K, Eacker SM, et al. A nuclease that mediates cell death induced by DNA damage and poly(ADP-ribose) polymerase-1. Science (80-). 2016;354(6308):aad6872-aad6872.

203. Xu L, Li Y, Li D, Xu P, Tian S, Sun H, et al. Exploring the binding mechanisms of MIF to CXCR2 using theoretical approaches. Phys Chem Chem Phys. 2015;

204. Laganeckas M, Margeleviius M, Venclovas E. Identification of new homologs of PD-(D/E)XK nucleases by support vector machines trained on data derived from profile-profile alignments. Nucleic Acids Res. 2011;39(4):1187–96.

205. Penkler D, Atilgan C, Tastan Bishop O. Allosteric Modulation of Human Hsp90α Conformational Dynamics. J Chem Inf Model. 2018;

206. Del Sol A, Fujihashi H, Amoros D, Nussinov R. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol Syst Biol. 2006;

207. Senter PD, Al-Abed Y, Metz CN, Benigni F, Mitchell RA, Chesney J, et al. Inhibition of macrophage migration inhibitory factor (MIF) tautomerase and biological activities by acetaminophen metabolites. Proc Natl Acad Sci U S A. 2002;99(1):144–9.

208. Ioannou K, Cheng KF, Crichlow G V, Birmpilis AI, Lolis EJ, Tsitsilonis OE, et al. ISO-66, a novel inhibitor of macrophage migration, shows efficacy in melanoma and colon cancer models. Int J Oncol. 2014;45(4):1457–68.

209. Atilgan AR, Akan P, Baysal C. Small-World Communication of Residues and Significance for Protein Dynamics. Biophys J. 2004;86(1):85–91.

210. El-Turk F, Fauvet B, Ashrafi A, Ouertatani-Sakouhi H, Cho MK, Neri M, et al. Characterization of Molecular Determinants of the Conformational Stability of Macrophage Migration Inhibitory Factor: Leucine 46 Hydrophobic Pocket. PLoS One. 2012;

211. Ponzoni L, Bahar I. Structural dynamics is a determinant of the functional significance of missense variants. Proc Natl Acad Sci. 2018;

212. Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol. 2006;356:1263–74.

213. Di Paola L, Platania CBM, Oliva G, Setola R, Pascucci F, Giuliani A. Characterization of Protein–Protein Interfaces through a Protein Contact Network Approach. Front Bioeng Biotechnol. 2015;

214. Brinda K V., Vishveshwara S. Oligomeric protein structure networks: Insights into protein-protein interactions. BMC Bioinformatics. 2005;

215. Barrett MP, Gilbert IH. Targeting of Toxic Compounds to the Trypanosome's Interior. Vol. 63, Advances in Parasitology. 2006. p. 125–83.

216. Fox JT, Stover PJ. Folate-mediated one-carbon metabolism. Vitam Horm. 2008;79(08):1–44.

217. Bello a R, Nare B, Freedman D, Hardy L, Beverley SM. PTR1: a reductase mediating salvage of oxidized pteridines and methotrexate resistance in the protozoan parasite Leishmania major. Proc Natl Acad Sci U S A. 1994;91(24):11442–6.

218. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome Trypanosoma brucei. Science (80-). 2005;309(5733):416–22.

219. Zuccotto F, Martin ACR, Laskowski RA, Thornton JM, Gilbert IH. Dihydrofolate reductase: a potential drug target in trypanosomes and leishmania. J Comput Aided Mol Des. 1998;12(3):241–57.

220. Sienkiewicz N, Jarosławski S, Wyllie S, Fairlamb AH. Chemical and genetic validation of dihydrofolate reductase-thymidylate synthase as a drug target in African trypanosomes. Mol Microbiol. 2008;69(2):520–33.

221. Nare B, Hardy LW, Beverley SM, Ptrs L. The Roles of Pteridine Reductase 1 and Dihydrofolate Reductase-Thymidylate Synthase in Pteridine Metabolism in the Protozoan Parasite Leishmania major *. 1997;272(21):13883–91.

222. Nzila A. The past, present and future of antifolates in the treatment of Plasmodium falciparum infection. Journal of Antimicrobial Chemotherapy. 2006.

223. FALCO EA, GOODWIN LG, HITCHINGS GH, ROLLO IM, RUSSELL PB. 2:4-diaminopyrimidines- a new series of antimalarials. Br J Pharmacol Chemother. 1951;

224. Ogwang S, Nguyen HT, Sherman M, Bajaksouzian S, Jacobs MR, Boom WH, et al. Bacterial conversion of folinic acid is required for antifolate resistance. J Biol Chem.

2011;

225. Allegra CJ, Kovacs JA, Drake JC, Swan JC, Chabner BA, Masur H. Potent in vitro and in vivo antitoxoplasma activity of the lipid-soluble antifolate trimetrexate. J Clin Invest. 1987;

226. Nixon GL, Moss DM, Shone AE, Lalloo DG, Fisher N, O'neill PM, et al. Antimalarial pharmacology and therapeutics of atovaquone. Journal of Antimicrobial Chemotherapy. 2013.

227. Srivastava IK, Vaidya AB. A mechanism for the synergistic antimalarial action of atovaquone and proguanil. Antimicrob Agents Chemother. 1999;

228. Vickers TJ, Beverley SM. Folate metabolic pathways in Leishmania. 2011;1–19.

229. Robello C, Navarro P, Castanys S, Gamarro F. A pteridine reductase gene ptr1 contiguous to a P-glycoprotein confers resistance to antifolates in Trypanosoma cruzi. Mol Biochem Parasitol. 1997;90(2):525–35.

230. Gourley DG, Schüttelkopf AW, Leonard GA, Luba J, Hardy LW, Beverley SM, et al. Pteridine reductase mechanism correlates pterin metabolism with drug resistance in trypanosomatid parasites. Nat Struct Biol. 2001;8(6):521–5.

231. Sienkiewicz N, Ong HB, Fairlamb AH. Trypanosoma brucei pteridine reductase 1 is essential for survival in vitro and for virulence in mice. Mol Microbiol. 2010;77(3):658–71.

232. Cavazzuti A, Paglietti G, Hunter WN, Gamarro F, Piras S, Loriga M, et al. Discovery of potent pteridine reductase inhibitors to guide antiparasite drug development. Proc Natl Acad Sci U S A. 2008;105:1448–53.

233. Tulloch LB, Martini VP, Iulek J, Huggan JK, Lee JH, Gibson CL, et al. Structure-based design of pteridine reductase inhibitors targeting African sleeping sickness and the leishmaniases. J Med Chem. 2010;53(1):221–9.

234. Mpamhanga CP, Spinks D, Tulloch LB, Shanks EJ, Robinson DA, Collie IT, et al. One scaffold, three binding modes: Novel and selective pteridine reductase 1 inhibitors derived from fragment hits discovered by virtual screening. J Med Chem. 2009;52(14):4454–65.

235. Hardy LW, Matthews W, Nare B, Beverley SM. Biochemical and genetic tests for inhibitors of Leishmania pteridine pathways. Exp Parasitol. 1997;87(3):157–69.

236. Nare B, Luba J, Hardy LW, Beverley S. New approaches to Leishmania chemotherapy: pteridine reductase 1 (PTR1) as a target and modulator of antifolate sensitivity. Parasitology. 1997;114 Suppl:S101–10.

237. Song CM, Lim SJ, Tong JC. Recent advances in computer-aided drug design. Briefings in Bioinformatics. 2009.

238. Lounnas V, Ritschel T, Kelder J, McGuire R, Bywater RP, Foloppe N. CURRENT PROGRESS IN STRUCTURE-BASED RATIONAL DRUG DESIGN MARKS A NEW MINDSET IN DRUG DISCOVERY. Comput Struct Biotechnol J. 2013;

239. Kore PP, Mutha MM, Antre R V., Oswal RJ, Kshirsagar SS. Computer-Aided Drug Design: An Innovative Tool for Modeling. Open J Med Chem. 2012;

240. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, et al. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PLoS One. 2013;

241. Soothill JS, Ward R, Girling AJ. The IC50: An exactly defined measure of antibiotic sensitivity. J Antimicrob Chemother. 1992;

242. Fivelman QL, Adagu IS, Warhurst DC. Modified fixed-ratio isobologram method for studying in vitro interactions between atovaquone and proguanil or dihydroartemisinin against drug-resistant strains of Plasmodium falciparum. Antimicrob Agents Chemother. 2004;

243. Gogtay N, Kannan S, Thatte UM, Olliaro PL, Sinclair D. Artemisinin-based combination therapy for treating uncomplicated Plasmodium vivax malaria. Cochrane Database of Systematic Reviews. 2013.

244. Visser BJ, Wieten RW, Kroon D, Nagel IM, Bélard S, Van Vugt M, et al. Efficacy and safety of artemisinin combination therapy (ACT) for non-falciparum malaria: A systematic review. Malaria Journal. 2014.

245. Tallarida RJ. An Overview of Drug Combination Analysis with Isobolograms. J Pharmacol Exp Ther. 2006;

246. Pardridge WM. The blood-brain barrier: Bottleneck in brain drug development. NeuroRx. 2005.

247. Hitchcock SA. Blood-brain barrier permeability considerations for CNS-targeted compound library design. Current Opinion in Chemical Biology. 2008.

248. Leeson PD, Davis AM. Time-related differences in the physical property profiles of oral drugs. J Med Chem. 2004;

249. Pardridge WM. CNS Drug Design Based on Principles of Blood-Brain Barrier Transport. J Neurochem. 2002;

250. Van De Waterbeemd H, Camenisch G, Folkers G, Chretien JR, Raevsky OA. Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors. J Drug Target. 1998;

251. Yan A, Liang H, Chong Y, Nie X, Yu C. In-silico prediction of blood-brain barrier permeability. SAR QSAR Environ Res. 2013;

252. Baell JB, Nissink JWM. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017 - Utility and Limitations. ACS Chemical Biology. 2018.

253. Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J

Med Chem. 2010;53(7):2719–40.

254. Baell JB. Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). Journal of Natural Products. 2016.

255. Baell J, Walters MA. Chemistry: Chemical con artists foil drug discovery. Nature. 2014.

256. Reker D, Bernardes GJL, Rodrigues T. Computational advances in combating colloidal aggregation in drug discovery. Nature Chemistry. 2019.

257. Feng BY, Simeonov A, Jadhav A, Babaoglu K, Inglese J, Shoichet BK, et al. A high-throughput screen for aggregation-based inhibition in a large compound library. J Med Chem. 2007;

258. Schorpp K, Rothenaigner I, Salmina E, Reinshagen J, Low T, Brenke JK, et al. Identification of small-molecule frequent hitters from alphascreen high-throughput screens. J Biomol Screen. 2014;

259. Soares KM, Blackmon N, Shun TY, Shinde SN, Takyi HK, Wipf P, et al. Profiling the NIH Small Molecule Repository for Compounds That Generate H_2O_2 by Redox Cycling in Reducing Environments. Assay Drug Dev Technol. 2010;

260. Ingólfsson HI, Thakur P, Herold KF, Hobart EA, Ramsey NB, Periole X, et al. Phytochemicals perturb membranes and promiscuously alter protein function. ACS Chem Biol. 2014;

261. Pai JJK, Kirkup MP, Frank EA, Pachter JA, Bryant RW. Compounds capable of generating singlet oxygen represent a source of artifactual data in scintillation proximity assays measuring phosphopeptide binding to SH2 domains. Anal Biochem. 1999;

262. Crowe A, James MJ, Virginia MYL, Smith AB, Trojanowski JQ, Ballatore C, et al. Aminothienopyridazines and methylene blue affect Tau fibrillization via cysteine

oxidation. J Biol Chem. 2013;

263. Notredame C. Recent evolutions of multiple sequence alignment algorithms. PLoS Computational Biology. 2007.

264. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

265. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;

266. Lassmann T, Sonnhammer ELL. Kalign - An accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005;

267. Notredame C, Higgins DG, Heringa J. T-coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;

268. Morgenstern B, Dress A, Werner T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc Natl Acad Sci. 1996;

269. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 2005;

270. Morris GM, Lim-Wilby M. Molecular docking. Methods Mol Biol. 2008;

271. Meng X-Y, Zhang H-X, Mezei M, Cui M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. Curr Comput Aided-Drug Des. 2011;

272. Sliwoski G, Kothiwale S, Meiler J, Lowe EW. Computational methods in drug discovery. Pharmacol Rev. 2014;

273. Shen CH, Wang YF, Kovalevsky AY, Harrison RW, Weber IT. Amprenavir complexes with HIV-1 protease and its drug-resistant mutants altering hydrophobic clusters. FEBS J. 2010;

274. Zhang KE, Wu E, Patick AK, Kerr B, Zorbas M, Lankford A, et al. Circulating

metabolites of the human immunodeficiency virus protease inhibitor nelfinavir in humans: Structural identification, levels in plasma, and antiviral activities. Antimicrob Agents Chemother. 2001;

275. Cocohoba J, Dong BJ. Raltegravir: The first HIV integrase inhibitor. Clin Ther. 2008;
276. Trott O, Olson AJ. Autodock vina. J Comput Chem. 2010;
277. Trott O, Olson A. AutoDock Vina: inproving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J Comput Chem. 2010;31(2):455–61.
278. Homeyer N, Gohlke H. Free energy calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area method. Mol Inform. 2012;31(2):114–22.
279. Hou T, Wang J, Li Y, Wang W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. J Chem Inf Model. 2011;51(1):69–82.
280. Kumari R, Kumar R, Lynn A. G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations. J Chem Inf Model. 2014;54(7):1951–62.
281. Hatherley R, Brown DK, Musyoka TM, Penkler DL, Faya N, Lobb K a, et al. SANCDB: a South African natural compound database. J Cheminform. 2015;7(1):29.
282. Irwin JJ, Shoichet BK. ZINC - A free database of commercially available compounds for virtual screening. J Chem Inf Model. 2005;45(1):177–82.
283. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. J Pharmacol Toxicol Methods. 2000;44(1):235–49.
284. Schnell JR, Dyson HJ, Wright PE. Structure, Dynamics, and Catalytic Function of Dihydrofolate Reductase. Annu Rev Biophys Biomol Struct. 2004;33(1):119–40.
285. Dawson A, Tulloch LB, Barrack KL, Hunter WN. High-resolution structures of Trypanosoma brucei pteridine reductase ligand complexes inform on the placement of

new molecular entities in the active site of a potential drug target. Acta Crystallogr Sect D Biol Crystallogr. 2010;66(12):1334–40.

286. Sousa da Silva AW, Vranken WF, Wang J, Wang W, Kollman P, Case D, et al. ACPYPE - AnteChamber PYthon Parser interfacE. BMC Res Notes. 2012;5(1):367.

287. San Diego: Accelrys Software Inc. Discovery Studio Modeling Environment, Release 3.5. Accelrys Software Inc. 2012.

288. Bowling T, Mercer L, Don R, Jacobs R, Nare B. Application of a resazurin-based high-throughput screening assay for the identification and progression of new treatments for human African trypanosomiasis. Int J Parasitol Drugs Drug Resist. 2012;2:262–70.

289. Vanichtananankul J, Taweechai S, Yuvaniyama J, Vilaivan T, Chitnumsub P, Kamchonwongpaisan S, et al. Trypanosomal dihydrofolate reductase reveals natural antifolate resistance. ACS Chem Biol. 2011;6(9):905–11.

290. S. Akinboye E. Biological Activities of Emetine. Open Nat Prod J. 2011;4(1):8–15.

291. Legros D, Ollivier G, Gastellu-Etchegorry M, Paquet C, Burri C, Jannin J, et al. Treatment of human African trypanosomiasis--present situation and needs for research and development. Lancet Infect Dis. 2002;2(7):437–40.

292. Dawson A, Gibellini F, Sienkiewicz N, Tulloch LB, Fyfe PK, McLuskey K, et al. Structure and reactivity of Trypanosoma brucei pteridine reductase: Inhibition by the archetypal antifolate methotrexate. Mol Microbiol. 2006;61(6):1457–68.

293. Sawaya MR, Kraut J. Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: Crystallographic evidence. Biochemistry. 1997;36(3):586–603.

294. Schüttelkopf AW, Hardy LW, Beverley SM, Hunter WN. Structures of Leishmania major pteridine reductase complexes reveal the active site features important for ligand

binding and to guide inhibitor design. J Mol Biol. 2005;352(1):105–16.

295. van Baren C, Martino V, Di Leo Lira P, Anao I, Houghton P, Debenedetti S, et al. Triterpenic Acids and Flavonoids from Satureja parvifolia. Evaluation of their Antiprotozoal Activity. Zeitschrift fur Naturforsch - Sect C J Biosci. 2006;61(3–4):189–92.

296. Tasdemir D, Kaiser M, Brun R, Yardley V, Schmidt TJ, Tosun F, et al. Antitrypanosomal and antileishmanial activities of flavonoids and their analogues: In vitro, in vivo, structure-activity relationship, and quantitative structure-activity relationship studies. Antimicrob Agents Chemother. 2006;50(4):1352–64.

297. Luba J, Nare B, Liang PH, Anderson KS, Beverley SM, Hardy LW. Leishmania major pteridine reductase 1 belongs to the short chain dehydrogenase family: stereochemical and kinetic evidence. Biochemistry. 1998;37(12):4093–104.

298. Kenny PW. Comment on the Ecstasy and Agony of Assay Interference Compounds. Vol. 57, Journal of Chemical Information and Modeling. 2017. p. 2640–5.

299. Oberholzer M, Marti G, Baresic M, Kunz S, Hemphill A, Seebeck T. The Trypanosoma brucei cAMP phosphodiesterases TbrPDEB1 and TbrPDEB2: flagellar enzymes that are essential for parasite virulence. FASEB J. 2007;21(3):720–31.

300. Jansen C, Wang H, Kooistra AJ, Graaf C De, Tenor H, Seebeck T, et al. Discovery of novel Trypanosoma brucei phosphodiesterase B1 inhibitors by virtual screening against the unliganded TbrPDEB1 crystal structure. J Med Chem. 2014;56(5):2087–96.

301. Blaazer AR, Singh AK, De Heuvel E, Edink E, Orrling KM, Veerman JJN, et al. Targeting a Subpocket in Trypanosoma brucei Phosphodiesterase B1 (TbrPDEB1) Enables the Structure-Based Discovery of Selective Inhibitors with Trypanocidal Activity. J Med Chem. 2018;61(9):3870–88.

302. Orrling KM, Jansen C, Vu XL, Balmer V, Bregy P, Shanmugham A, et al. Catechol pyrazolinones as trypanocidals: Fragment-based design, synthesis, and pharmacological evaluation of nanomolar inhibitors of trypanosomal phosphodiesterase B1. J Med Chem. 2012;55(20):8745–56.

303. Bollini M, Domaoal RA, Thakur V V., Gallardo-Macias R, Spasov KA, Anderson KS, et al. Computationally-guided optimization of a docking hit to yield catechol diethers as potent anti-HIV agents. J Med Chem. 2011;

304. De Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of Molecular Dynamics and Related Methods in Drug Discovery. Journal of Medicinal Chemistry. 2016.

305. Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biology. 2011.

306. Vedani A, Huhta DW. A New Force Field for Modeling Metalloproteins. J Am Chem Soc. 1990;

307. Merz KM, Murcko MA, Kollman PA. Inhibition of carbonic anhydrase. J Am Chem Soc. 1991;

308. Concha NO, Rasmussen BA, Bush K, Herzberg O. Crystal structure of the wide-spectrum binuclear zinc beta-lactamase from Bacteroides fragilis. Structure. 1996;

309. Oelschlaeger P, Schmid RD, Pleiss J. Insight into the mechanism of the IMP-1 metallo- -lactamase by molecular dynamics simulations. Protein Eng Des Sel. 2003;16(5):341–50.

310. Oberholzer M, Saada EA, Hill L. Cyclic AMP Regulates Social Behavior in African Trypanosomes. MBio. 2015;6(3):1–11.

311. Pounds AJ. Introduction to Quantum Mechanics: A Time-Dependent Perspective (David J. Tannor). J Chem Educ. 2008;

312. Stewart ET. Quantum Chemistry. Nature. 1970;

313. Kohn W, Becke AD, Parr RG. Density functional theory of electronic structure. J Phys Chem. 1996;

314. Zgarbová M, Otyepka M, Šponer J, Hobza P, Jurečka P. Large-scale compensation of errors in pairwise-additive empirical force fields: Comparison of AMBER intermolecular terms with rigorous DFT-SAPT calculations. Phys Chem Chem Phys. 2010;

315. Hagler AT, Lifson S, Dauber P. Consistent Force Field Studies of Intermolecular Forces in Hydrogen-Bonded Crystals. 2. A Benchmark for the Objective Comparison of Alternative Force Fields. J Am Chem Soc. 1979;

316. Brooks BR, III CLB, A. D. Mackerell J, Nilsson L, Petrella RJ, Roux B, et al. CHARMM: The Biomolecular Simulation Program B. J Comput Chem. 2009;

317. MacKerell AD, Banavali N, Foloppe N. Development and current status of the CHARMM force field for nucleic acids. Biopolymers. 2000;

318. Mackerell AD, Feig M, Brooks CL. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulation. J Comput Chem. 2004;

319. Holm RH, Kennepohl P, Solomon EI. Structural and functional aspects of metal sites in biology. Chem Rev. 1996;

320. Andreini C, Bertini I, Cavallaro G, Holliday GL, Thornton JM. Metal ions in biological catalysis: From enzyme databases to general principles. J Biol Inorg Chem. 2008;

321. Zhen S, Davies GJ. Calculation of the Lennard-Jones n–m potential energy parameters for metals. Phys status solidi. 1983;

322. Li P, Roberts BP, Chakravorty DK, Merz KM. Rational design of particle mesh ewald

compatible lennard-jones parameters for +2 metal cations in explicit solvent. J Chem Theory Comput. 2013;

323. Li P, Song LF, Merz KM. Systematic parameterization of monovalent ions employing the nonbonded model. J Chem Theory Comput. 2015;
324. Li P, Song LF, Merz KM. Parameterization of highly charged metal ions using the 12-6-4 LJ-type nonbonded model in explicit water. J Phys Chem B. 2015;
325. Peters MB, Yang Y, Wang B, Füsti-Molnár L, Weaver MN, Merz KM. Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). J Chem Theory Comput. 2010;
326. Pang YP. Successful molecular dynamics simulation of two zinc complexes bridged by a hydroxide in phosphotriesterase using the cationic dummy atom method. Proteins Struct Funct Genet. 2001;
327. Lu X, Gaus M, Elstner M, Cui Q. Parametrization of DFTB3/3OB for magnesium and zinc for chemical and biological applications. J Phys Chem B. 2015;
328. Yang TY, Dudev T, Lim C. Mononuclear versus binuclear metal-binding sites: Metal-binding affinity and selectivity from PDB survey and DFT/CDM calculations. J Am Chem Soc. 2008;130(12):3844–52.
329. Chung LW, Sameera WMC, Ramozzi R, Page AJ, Hatanaka M, Petrova GP, et al. The ONIOM Method and Its Applications. Chemical Reviews. 2015.
330. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. Gaussian09. Gaussian 09. 2009.
331. Becke A. B3LYP. J Chem Phys. 1993;
332. Rappé AKK, Casewit CJJ, Colwell KSS, Goddard III WA, Skiff WM. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. J Am Chem Soc. 1992;

333. Zhang Y. Improved pseudobonds for combined ab initio quantum mechanical/molecular mechanical methods. J Chem Phys. 2005;

334. Chiodo S, Russo N, Sicilia E. LANL2DZ basis sets recontracted in the framework of density functional theory. J Chem Phys. 2006;

335. Dudev T, Lin Y lin, Dudev M, Lim C. First-second shell interactions in metal binding sites in proteins: A PDB survey and DFT/CDM calculations. J Am Chem Soc. 2003;

336. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem. 2010;

337. Huang J, Mackerell AD. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. J Comput Chem. 2013;

338. Beglov D, Roux B. Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations. J Chem Phys. 1994;

339. Stote RH, Karplus M. Zinc binding in proteins and solution: a simple but accurate nonbonded representation. Proteins. 1995;

340. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem. 1983;

341. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. J Mol Graph. 1996;

342. R Development Core Team RFFSC, Millsap RE, Chase CWJ. R: A Language and Environment for Statistical Computing. J Bus Forecast Methods Syst. 1995;

343. Wallace AC, Laskowski RA, Thornton JM. Ligplot - a Program To Generate Schematic Diagrams of Protein Ligand Interactions. Protein Eng. 1995;

344. Dennington R, Keith T, Millam J. GaussView, Version 5. Semichem Inc. , Shawnee

Mission, KS. 2009.

345. Wang H, Kunz S, Chen G, Seebeck T, Wan Y, Robinson H, et al. Biological and structural characterization of Trypanosoma cruzi phosphodiesterase C and implications for design of parasite selective inhibitors. J Biol Chem. 2012;

346. Wang H, Yan Z, Geng J, Kunz S, Seebeck T, Ke H. Crystal structure of the Leishmania major phosphodiesterase LmjPDEB1 and insight into the design of the parasite-selective inhibitors. Mol Microbiol. 2007;

347. Card GL, England BP, Suzuki Y, Fong D, Powell B, Lee B, et al. Structural basis for the activity of drugs that inhibit phosphodiesterases. Structure. 2004;

348. Liu H, Wang L, Lv M, Pei R, Li P, Pei Z, et al. AlzPlatform: An Alzheimer's disease domain-specific chemogenomics knowledgebase for polypharmacology and target identification research. J Chem Inf Model. 2014;

349. Wager TT, Hou X, Verhoest PR, Villalobos A. Moving beyond rules: The development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. ACS Chem Neurosci. 2010;

350. Geldenhuys WJ, Mohammad AS, Adkins CE, Lockman PR. Molecular determinants of blood-brain barrier permeation. Therapeutic Delivery. 2015.

351. Ghose AK, Herbertz T, Hudkins RL, Dorsey BD, Mallamo JP. Knowledge-based, central nervous system (CNS) lead selection and lead optimization for CNS drug discovery. ACS Chem Neurosci. 2012;

352. Xiong Y, Lu HT, Li Y, Yang GF, Zhan CG. Characterization of a catalytic ligand bridging metal ions in phosphodiesterases 4 and 5 by molecular dynamics simulations and hybrid quantum mechanical/molecular mechanical calculations. Biophys J. 2006;91(5):1858–67.

353. Degerman E, Belfrage P, Manganiello VC. Structure, localization, and regulation of

cGMP-inhibited phosphodiesterase (PDE3). Journal of Biological Chemistry. 1997.

354. Soderling SH, Beavo JA. Regulation of cAMP and cGMP signaling: New phosphodiesterases and new functions. Current Opinion in Cell Biology. 2000.

355. Salter EA, Wierzbicki A. The mechanism of cyclic nucleotide hydrolysis in the phosphodiesterase catalytic site. J Phys Chem B. 2007;

356. Stamper C, Bennett B, Edwards T, Holz RC, Ringe D, Petsko G. Inhibition of the Aminopeptidase from Aeromonas proteolytica by L -Leucinephosphonic Acid . Spectroscopic and Crystallographic Characterization of the Transition State of Peptide Hydrolysis †. 2001;7035–46.

357. Englert M, Xia S, Okada C, Nakamura A, Tanavde V, Yao M, et al. Structural and mechanistic insights into guanylylation of RNA-splicing ligase RtcB joining RNA between. 2012;

358. Lu X, Gaus M, Elstner M, Cui Q. Parametrization of DFTB3/3OB for magnesium and zinc for chemical and biological applications. J Phys Chem B. 2015;119(3):1062–82.

359. Wong KY, Gao J. Insight into the phosphodiesterase mechanism from combined QM/MM free energy simulations. FEBS J. 2011;

360. Wong KY, Gao J. The reaction mechanism of paraoxon hydrolysis by phosphotriesterase from combined QM/MM simulations. Biochemistry. 2007;

361. Li Z, Wu Y, Feng LJ, Wu R, Luo H Bin. Ab initio QM/MM study shows a highly dissociated SN2 hydrolysis mechanism for the cGMP-specific phosphodiesterase-5. J Chem Theory Comput. 2014;10(12):5448–57.

362. Wong KY, Gao J. Insight into the phosphodiesterase mechanism from combined QM/MM free energy simulations. FEBS J. 2011;278(14):2579–95.

363. Cieplak P, Dupradeau FY, Duan Y, Wang J. Polarization effects in molecular mechanical force fields. Journal of Physics Condensed Matter. 2009.

SUPPLEMENTARY DATA

S1 Table: Candidate genes included in the study

https://doi.org/10.1371/journal.pntd.0006300.s001 [13]

Gene	SNP	Location (GRCh37 assembly)	Consequence
CFH	rs1061170	1:196659237	Missense variant
IL10	rs1800872	1:206946407	Upstream gene variant
IL1B	rs1143629	2:113593518	Intron variant
IL8	rs114259658	4:74605639	Upstream gene variant
IL8	rs2227307	4:74606669	Intron variant
IL8	rs2227545	4:74608727	3 prime UTR variant
IL8	rs13112910	4:74609755	Downstream gene variant
IL8	rs58478511	4:74610033	Downstream gene variant
IL8	rs62312369	4:74610397	Downstream gene variant
IL4	rs2243250	5:132009154	Upstream gene variant
IL4	rs2070874	5:132009710	5 prime UTR variant
IL4	rs734244	5:132010726	Intron variant
IL4	rs2243255	5:132011737	Intron variant
IL4	rs2243256	5:132011753	Intron variant
IL4	rs2243258	5:132012110	Intron variant
IL4	rs2243261	5:132012806	Intron variant
IL4	rs71889624	5:132013430	Intron variant
IL4	rs2243268	5:132013963	Intron variant
IL4	rs9282745	5:132014000	Intron variant
IL4	rs2243270	5:132014109	Intron variant
IL4	rs2243279	5:132016227	Intron variant
IL4	rs2243282	5:132016554	Intron variant
IL4	rs2243283	5:132016593	Intron variant
IL4	rs2243285	5:132016993	Intron variant
IL4	rs73269366	5:132018749	Downstream gene variant
IL12B	rs3212227	5:158742950	UTR variant 3 prime
IL12B	rs2546890	5:158759900	Intron variant
HLA-A	rs1059563	6: 29911928	Missense variant
HLA-G	rs142798055	6:29793404	Upstream gene variant
HLA-G	rs1736936	6:29794317	Upstream gene variant
HLA-G	rs17875389	6:29794484	Upstream gene variant
HLA-G	rs1130355	6:29795993	Synonymous variant
HLA-G	rs17875406	6:29797448	Synonymous variant
HLA-G	rs1130363	6:29797696	Synonymous variant
HLA-G	rs1130363	6:29797696	Synonymous codon
HLA-G	rs1632932	6:29798039	Intron variant
HLA-G	rs371194629	6:29798581	3 prime UTR variant
HLA-G	rs17179108	6:29798642	3 prime UTR variant
HLA-G	rs9380142	6:29798794	3 prime UTR variant
HLA-G	rs1610696	6:29798803	3 prime UTR variant
HLA-G	rs1233330	6:29799103	Downstream gene variant
HLA-G	rs1611139	6:29799116	Downstream gene variant
HLA-G	rs2517898	6:29799746	Downstream gene variant
HLA-G	rs141206123	6:29799849	Downstream gene variant
HLA-G	rs12661041	6:29800062	Downstream gene variant
HLA-G	rs2517897	6:29800101	Downstream gene variant
HLA-G	rs12662618	6:29800211	Downstream gene variant
Intergenic between HLA-G and HLA-A	rs34783406	6:29832415	indel
HLA-A	rs1136754	6:29911921	Synonymous variant
HLA-A	rs1136754	6:29911921	Synonymous codon

HLA-A	rs1059564	6:29911930	Synonymous variant
TNFA	rs1800630	6:31542476	Downstream gene variant
TNFA	rs1800629	6:31543031	Downstream gene variant
IL6	rs62449495	7:22764338	Upstream gene variant
IL6	rs2069830	7:22767137	Missense variant
IL6	rs2069834	7:22767828	Intron variant
IL6	rs2069837	7:22768027	Intron variant
IL6	rs1474347	7:22768124	Intron variant
IL6	rs2066992	7:22768249	Intron variant
IL6	rs2069842	7:22769310	Intron variant
IL6	rs1548216	7:22769773	Intron variant
IL6	rs2069843	7:22769994	Intron variant
IL6	rs2069845	7:22770149	Intron variant
IL6	rs2069855	7:22772624	Downstream gene variant
IL6	rs1818879	7:22772727	Downstream gene variant
IFN-G	rs2069728	12:68547784	Intron variant
IFN-G	rs2069723	12:68548594	3 prime UTR variant
IFN-G	rs2069722	12:68548953	3 prime UTR_variant
IFN-G	rs2069720	12:68549710	Intron variant
IFN-G	rs2069718	12:68550162	Intron variant
IFN-G	rs1861493	12:68551196	Intron variant
IFN-G	rs2069713	12:68552476	Intron variant
IFN-G	rs2430561	12:68552522	Intron variant
IFN-G	rs78554979	12:68554636	Upstream gene variant
IFN-G	rs2069705	12:68555011	Upstream gene variant
TNX4LB	rs152828	16: 72123886	Intron variant
IL4R	rs1801275	16:27374400	Missense variant
TXNL4B	rs1424241	16:72078907	Intron variant
HP	rs8062041	16:72088964	Intron variant
HPR	rs7185840	16:72102112	Intron variant
HPR	rs2021171	16:72110541	Missense variant
IL12RB1	rs375947	19:18180451	Missense variant
IL12RB1	rs11575934	19:18186618	Missense variant
MIF	rs12483859	22:24234807	Upstream gene variant
MIF	rs36086171	22:24235455	Upstream gene variant
MIF	rs9282783	22:24236359	5 prime UTR variant
MIF	rs36070976	22:24236864	Intron variant
MIF	rs2070766	22:24237221	Splice region variant, intron variant
MIF	rs35235644	22:24237822	Downstream gene variant
MIF	rs2000466	22:24237862	Downstream gene variant
MIF	rs34383331	22:24238079	Downstream gene variant
APOL1	rs136174	22:36661536	Synonymous codon
APOL1	rs73885316	22: 36661674	Missense variant
APOL1	rs136177	22:36661842	Synonymous codon
APOL1	rs73885319	22:36661906	Missense variant
APOL1	rs143830837	22:36662042-36662047	Coding sequence indel
APOL1	rs71785313	22: 36662046:36662051	Coding sequence indel

S2 Table: Association results of 65 SNPs with Acute HAT https://doi.org/10.1371/journal.pntd.0006300.s002 [13]

CHR	SNP	GENE	BP	A1	F_A	F_U	A2	P	OR	L95	U95	BONF	FDR_BH	FST	MAF
1	rs1061170	CFH	196659237	C	0.4915	0.4788	T	0.7833	1.052	0.7333	1.51	1	0.9257	-0.00405298	0.4852
1	rs1800872	IL-10	206946407	T	0.4068	0.4286	G	0.6095	0.9143	0.6346	1.317	1	0.9257	-0.00354243	0.4177
2	rs1143629	IL-1B	113593518	A	0.4118	0.4576	G	0.3319	0.8296	0.5767	1.193	1	0.8752	0.000171276	0.4346
4	rs114259658	IL-8	74605639	A	0.05042	0.04202	T	0.6706	1.211	0.5127	2.858	1	0.9257	-0.00363002	0.04622
4	rs13112910	IL-8	74609755	G	0.3655	0.3782	A	0.777	0.9475	0.6533	1.374	1	0.9257	-0.0044355	0.3718
4	rs2227307	IL-8	74606669	T	0.4788	0.5	G	0.6144	0.9187	0.6408	1.317	1	0.9257	-0.00293707	0.4895
4	rs2227545	IL-8	74608727	C	0.07563	0.08824	A	0.6215	0.8455	0.4383	1.631	1	0.9257	-0.00327016	0.08193
4	rs58478511	IL-8	74610033	A	0.2348	0.2261	G	0.8259	1.05	0.6804	1.621	1	0.9281	-0.004256	0.2304
5	rs2546890	IL-12B	158759900	A	0.3051	0.3193	G	0.7294	0.9358	0.6345	1.38	1	0.9257	-0.00402212	0.3122
5	rs3212227	IL-12B	158742950	G	0.4412	0.411	T	0.4872	1.131	0.7859	1.629	1	0.9257	-0.00260248	0.4262
5	rs2243250	IL-4	132009154	C	0.2353	0.2437	T	0.831	0.9549	0.6268	1.455	1	0.9281	-0.00446726	0.2395
5	rs2243255	IL-4	132011737	A	0.1096	0.1681	G	0.06764	0.6093	0.3542	1.048	1	0.5496	0.0100856	0.1388
5	rs2243256	IL-4	132011753	DEL	0.1303	0.09664	T	0.2523	1.4	0.79	2.481	1	0.8752	0.00158722	0.1134
5	rs2243258	IL-4	132012110	T	0.07563	0.06723	C	0.7265	1.135	0.5644	2.283	1	0.9257	-0.0036442	0.07143
5	rs2243261	IL-4	132012806	T	0.07983	0.05462	G	0.2799	1.502	0.7239	3.115	1	0.8752	0.00115422	0.06695
5	rs2243268	IL-4	132013963	C	0.4202	0.3739	A	0.3049	1.213	0.8399	1.752	1	0.8752	0.000652226	0.3971
5	rs2243279	IL-4	132016227	A	0.06723	0.06723	G	0.9276	1	0.488	2.049	1	0.9519	-0.0039319	0.06723
5	**rs2243283**	**IL-4**	**132016593**	**G**	**0.275**	**0.1875**	**C**	**0.04006**	**1.644**	**1.031**	**2.621**	**1**	**0.4339**	**0.0158649**	**0.2304**
5	rs73269366	IL-4	132018749	T	0.01681	0.03361	C	0.2612	0.4915	0.146	1.655	1	0.8752	0.00163907	0.02521
5	rs734244	IL-4	132010726	C	0.4703	0.5042	T	0.4362	0.8732	0.609	1.252	1	0.9257	-0.00300466	0.4873
5	rs9282745	IL-4	132014000	A	0.02119	0.04202	T	0.2477	0.4935	0.1661	1.466	1	0.8752	0.00914955	0.03165
6	**rs1233330**	**HLA-G**	**29799103**	**A**	**0.1008**	**0.1597**	**G**	**0.05823**	**0.5903**	**0.3419**	**1.019**	**1**	**0.4339**	**0.0142926**	**0.1055**
6	**rs1233330**	**HLA-G**	**29799103**	**A**	**0.07563**	**0.1356**	**G**	**0.03048**	**0.5216**	**0.284**	**0.9581**	**1**	**0.5407**	**0.00914955**	**0.1303**
6	rs12662618	HLA-G	29800211	C	0.08475	0.09664	T	0.6921	0.8655	0.4618	1.622	1	0.9257	-0.00340373	0.09072
6	rs141206123	HLA-G	29799849	C	0.07983	0.07983	DEL	0.933	1	0.5154	1.94	1	0.9519	-0.00459676	0.07983
6	rs142798055	HLA-G	29793404	DEL	0.09709	0.1179	TCT	0.4807	0.8043	0.4318	1.498	1	0.9257	-0.00293754	0.1077
6	rs1610696	HLA-G	29798803	G	0.1441	0.1538	C	0.7472	0.9257	0.557	1.539	1	0.9257	-0.00532233	0.1489
6	rs1632932	HLA-G	29798039	A	0.3551	0.2929	G	0.1896	1.329	0.878	2.013	1	0.8752	0.005317	0.3252
6	rs17875389	HLA-G	29794484	G	0.05085	0.03361	A	0.3122	1.54	0.6179	3.839	1	0.8752	-0.00126392	0.04219
6	rs2517898	HLA-G	29799746	G	0.2941	0.339	C	0.301	0.8125	0.5513	1.197	1	0.8752	0.000642655	0.3165
6	**rs9380142**	**HLA-G**	**29798794**	**G**	**0.3686**	**0.2415**	**A**	**0.002777**	**1.834**	**1.231**	**2.731**	**0.1805**	**0.1805**	**0.030427**	**0.3051**
6	rs1800629	TNF-A	31543031	A	0.1092	0.09664	G	0.6549	1.146	0.6341	2.073	1	0.9257	-0.00307857	0.1029
6	**rs1800630**	**TNF-A**	**31542476**	**A**	**0.1555**	**0.09244**	**C**	**0.03809**	**1.807**	**1.031**	**3.169**	**1**	**0.4339**	**0.0141093**	**0.1239**
7	rs1818879	IL-6	22772727	A	0.1325	0.1282	G	0.8919	1.038	0.6062	1.779	1	0.9519	-0.00423326	0.1303
7	rs2066992	IL-6	22768249	T	0.06303	0.07627	G	0.5321	0.8146	0.4004	1.657	1	0.9257	-0.00285209	0.06962
7	rs2069830	IL-6	22767137	T	0.09244	0.07143	C	0.4102	1.324	0.6843	2.562	1	0.9257	-0.00185698	0.08193
7	rs2069834	IL-6	22767828	T	0.04202	0.07983	C	0.08883	0.5055	0.2299	1.112	1	0.6415	0.00791584	0.06092
7	rs2069837	IL-6	22768027	G	0.105	0.1429	A	0.2148	0.7042	0.4059	1.222	1	0.8752	0.00263809	0.1239
7	rs2069843	IL-6	22769994	A	0.1639	0.1975	G	0.344	0.7964	0.4984	1.273	1	0.8752	-0.000431748	0.1807
7	rs2069845	IL-6	22770149	G	0.3193	0.3613	A	0.3356	0.8292	0.5672	1.212	1	0.8752	-0.000164277	0.3403
7	rs2069855	IL-6	22772624	C	0.02941	0.03782	T	0.6234	0.771	0.2824	2.105	1	0.9257	-0.0035445	0.03361
7	rs62449495	IL-6	22764338	A	0.07983	0.05462	G	0.2799	1.502	0.7239	3.115	1	0.8752	0.00115422	0.06723
12	rs1861493	IFN-ϒ	68551196	G	0.05882	0.05042	A	0.6933	1.177	0.5327	2.601	1	0.9257	-0.003994	0.05462
12	rs2069705	IFN-ϒ	68555011	G	0.4286	0.3824	A	0.3066	1.212	0.8399	1.748	1	0.8752	0.000294768	0.4055
12	rs2069713	IFN-ϒ	68552476	C	0.07563	0.06303	T	0.5948	1.216	0.598	2.474	1	0.9257	-0.00295927	0.06933
12	rs2069720	IFN-ϒ	68549710	T	0.1017	0.1261	C	0.4295	0.7849	0.444	1.388	1	0.9257	-0.00199975	0.1139
12	rs2069722	IFN-ϒ	68548953	A	0.02809	0.04717	G	0.364	0.5838	0.1958	1.741	1	0.8764	-0.000786883	0.03846
12	rs2069723	IFN-ϒ	68548594	C	0.008475	0.01261	T	0.8424	0.6695	0.1109	4.044	1	0.9281	-0.00338787	0.01055
12	rs2069728	IFN-ϒ	68547784	T	0.2269	0.2479	C	0.5924	0.8904	0.5835	1.359	1	0.9257	-0.00297077	0.2374
12	rs2430561	IFN-ϒ	68552522	A	0.1767	0.1765	T	0.9519	1.002	0.6235	1.61	1	0.9519	-0.00450227	0.1766
12	rs78554979	IFN-ϒ	68554636	C	0.0678	0.07143	T	0.9294	0.9455	0.4659	1.919	1	0.9519	-0.0041117	0.06962
16	rs1424241	TXNL4B	72078907	A	0.1176	0.1154	G	0.9432	1.022	0.5825	1.794	1	0.9519	-0.00456464	0.1165
16	rs7185840	HPR	72102112	A	0.1597	0.1765	G	0.6267	0.8867	0.5481	1.434	1	0.9257	-0.00427472	0.1681
16	rs8062041	HP	72088964	C	0.4496	0.4831	T	0.4912	0.8741	0.6092	1.254	1	0.9257	-0.00181108	0.4662

16	rs1801275	IL-4R	27374400	A	0.2034	0.2203	G	0.6545	0.9034	0.5808	1.405	1	0.9257	-0.00366904	0.2119
19	rs1736936	HLAG	29794317	A	0.4286	0.416	G	0.7817	1.053	0.7319	1.515	1	0.9257	-0.0040236	0.4223
19	rs11575934	IL-12RB1	18186618	C	0.06723	0.08974	T	0.3501	0.731	0.3715	1.439	1	0.8752	-0.000385708	0.07806
22	rs136177	APOL1	36661842	G	0.1398	0.1092	A	0.2993	1.326	0.7657	2.295	1	0.8752	0.000494383	0.1245
22	rs71785313	APOL1	36662046	T	0.08051	0.08621	A	0.8035	0.9281	0.4817	1.788	1	0.9281	-0.00394606	0.08333
22	rs73885316	APOL1	36661674	A	0.04202	0.0339	C	0.7255	1.25	0.4846	3.225	1	0.9257	-0.00318007	0.03797
22	rs73885319	APOL1	36661906	G	0.07563	0.1008	A	0.3386	0.7295	0.3849	1.383	1	0.8752	0.000139201	0.08824
22	**rs34383331**	**MIF**	**24238079**	**A**	**0.2395**	**0.1597**	**T**	**0.03006**	**1.657**	**1.049**	**2.618**	**1**	**0.4339**	**0.0150296**	**0.1996**
22	rs35235644	MIF	24237822	C	0.09244	0.08403	G	0.7501	1.11	0.5889	2.093	1	0.9257	-0.0040495	0.08824
22	rs36086171	MIF	24235455	G	0.3077	0.3277	A	0.6574	0.9117	0.6187	1.343	1	0.9257	-0.00510933	0.3178
22	**rs9282783**	**MIF**	**24236359**	**G**	**0.08898**	**0.04202**	**C**	**0.03328**	**2.227**	**1.025**	**4.837**	**1**	**0.4339**	**0.012872**	**0.0654**

*Abbreviations: CHR = Chromosome, SNP = SNP ID, BP = Physical position (base-pair), A1 = Minor allele (based on whole sample), F_A = Frequency of allele 1 in cases, F_U = Frequency of allele 1 in controls, A2 = Major allele, P = p-value for this test, OR = Estimated odds ratio (for A1, i.e. A2 is reference), BONF = Bonferroni single-step adjusted p-values, FDR_BH = Benjamini & Hochberg (1995) step-up FDR control, FST = Fixation index, and MAF = Minor allele frequency. The level of significance is 0.05.

S3 Table: Association results of 65 SNPs with Chronic HAT https://doi.org/10.1371/journal.pntd.0006300.s003 [13]

CHR	SNP	GENE	BP	A1	F_A	F_U	A2	P	OR	L95	U95	BONF	FDR_BH	FST	MAF
1	**rs1061170**	**CFH**	**196659237**	**C**	**0.4091**	**0.5248**	**T**	**0.0188**	**0.627**	**0.4221**	**0.9313**	**1**	**0.6107**	**0.022134**	**0.4675**
1	rs1800872	IL-10	206946407	T	0.4457	0.4646	G	0.7198	0.9263	0.619	1.386	1	0.8928	-0.00550129	0.4555
2	rs1143629	IL-1B	113593518	A	0.3265	0.35	G	0.6338	0.9004	0.5936	1.366	1	0.8653	-0.00378649	0.3384
4	rs114259658	IL-8	74605639	A	0.05556	0.02427	T	0.1006	2.365	0.8065	6.934	1	0.6902	0.00808187	0.0396
4	rs2227307	IL-8	74606669	T	0.3333	0.4078	G	0.1334	0.7262	0.4824	1.093	1	0.7228	0.00717949	0.3719
4	rs2227545	IL-8	74608727	C	0.1111	0.07767	A	0.2716	1.484	0.7552	2.918	1	0.8653	0.00152592	0.09406
4	rs58478511	IL-8	74610033	A	0.4105	0.3788	G	0.5009	1.142	0.7599	1.717	1	0.8653	-0.00441686	0.3943
5	rs2243250	IL-4	132009154	C	0.2323	0.1942	T	0.3639	1.256	0.7791	2.024	1	0.8653	-0.000908447	0.2129
5	rs734244	IL-4	132010726	T	0.4242	0.4417	C	0.7259	0.9312	0.6281	1.381	1	0.8928	-0.00458497	0.4332
5	rs2243256	IL-4	132011753	DEL	0.05051	0.03398	T	0.3937	1.512	0.564	4.054	1	0.8653	-0.00137587	0.04208
5	rs2243258	IL-4	132012110	T	0.07071	0.1262	C	0.05637	0.5268	0.2665	1.041	1	0.6107	0.0116914	0.09901
5	rs2243261	IL-4	132012806	T	0.2424	0.2039	G	0.3718	1.25	0.7813	1.998	1	0.8653	-0.000686061	0.2228
5	rs9282745	IL-4	132014000	A	0.03061	0.02913	T	0.8859	1.053	0.3337	3.321	1	0.914	-0.00483513	0.02985
5	rs2243270	IL-4	132014109	A	0.1939	0.1942	G	0.9498	0.9981	0.6087	1.637	1	0.9498	-0.00493724	0.194
5	rs2243279	IL-4	132016227	A	0.06566	0.1019	G	0.1819	0.619	0.301	1.273	1	0.7884	0.00338443	0.08416
5	rs2243283	IL-4	132016593	G	0.1986	0.2216	C	0.6331	0.8707	0.5073	1.494	1	0.8653	-0.00612429	0.2112
5	rs73269366	IL-4	132018749	T	0.0101	0.01942	C	0.5634	0.5153	0.09332	2.845	1	0.8653	-0.00194264	0.01485
5	rs3212227	IL-12B	158742950	G	0.3939	0.3627	T	0.505	1.142	0.7628	1.709	1	0.8653	-0.00318415	0.3781
5	rs2546890	IL-12B	158759900	A	0.4545	0.4596	G	0.9201	0.9799	0.6598	1.455	1	0.9344	-0.0051162	0.4571
6	rs142798055	HLA-G	29793404	DEL	0.2065	0.137	TCT	0.09539	1.64	0.9075	2.963	1	0.6902	0.0110264	0.1758
6	rs17875389	HLA-G	29794484	G	0.06566	0.04902	A	0.4582	1.363	0.5835	3.185	1	0.8653	-0.0026029	0.05721
6	rs17179108	HLA-G	29798642	T	0.1162	0.1408	C	0.5074	0.8022	0.4465	1.441	1	0.8653	-0.00198705	0.1287
6	rs9380142	HLA-G	29798794	G	0.2419	0.2157	A	0.5085	1.161	0.723	1.863	1	0.8653	-0.00742882	0.2282
6	rs1610696	HLA-G	29798803	G	0.2368	0.17	C	0.09037	1.515	0.9209	2.493	1	0.6902	0.00734915	0.2026
6	**rs1233330**	**HLA-G**	**29799103**	**A**	**0.07576**	**0.1359**	**G**	**0.04453**	**0.5211**	**0.2693**	**1.008**	**1**	**0.6107**	**0.0120013**	**0.1064**
6	rs1233330	HLA-G	29799103	A	0.07576	0.1068	G	0.2659	0.6855	0.3447	1.363	1	0.8653	0.000130209	0.09158
6	rs141206123	HLA-G	29799849	C	0.0202	0.03883	DEL	0.3165	0.5103	0.1512	1.722	1	0.8653	0.00120087	0.0297
6	rs2517897	HLA-G	29800101	A	0.2727	0.267	C	0.8669	1.03	0.6634	1.598	1	0.9089	-0.0075944	0.2698
6	rs12662618	HLA-G	29800211	C	0.2121	0.1814	T	0.4168	1.215	0.7423	1.989	1	0.8653	-0.00141927	0.1965
6	rs1059564	HLA-A	29911930	T	0	0.00495	C	0.7437	0	0	nan	1	0.8928	-0.000252777	0.002538
6	rs1800630	TNF-A	31542476	A	0.09596	0.08738	C	0.798	1.109	0.5637	2.18	1	0.8928	-0.00464719	0.09158
6	rs1800629	TNF-A	31543031	A	0.08586	0.05825	G	0.2934	1.518	0.7057	3.267	1	0.8653	0.00112916	0.07178
7	rs62449495	IL-6	22764338	A	0.05051	0.03883	G	0.5533	1.316	0.5087	3.407	1	0.8653	-0.00373978	0.04455
7	rs2069830	IL-6	22767137	T	0.005051	0.009709	C	0.8084	0.5178	0.04658	5.756	1	0.8928	-0.00348262	0.007426
7	rs2069834	IL-6	22767828	T	0.0202	0.01456	C	0.5845	1.395	0.3083	6.315	1	0.8653	0.0184311	0.01733
7	rs2069837	IL-6	22768027	G	0.1212	0.1311	A	0.8241	0.9144	0.5079	1.647	1	0.8928	-0.00450927	0.1262
7	rs1474347	IL-6	22768124	C	0.1111	0.1019	A	0.8108	1.101	0.585	2.073	1	0.8928	-0.00500229	0.1064
7	rs2066992	IL-6	22768249	T	0.1211	0.1716	G	0.179	0.665	0.3769	1.173	1	0.7884	0.00531762	0.1472
7	**rs2069843**	**IL-6**	**22769994**	**A**	**0.1465**	**0.07767**	**G**	**0.03313**	**2.038**	**1.07**	**3.882**	**1**	**0.6107**	**0.0184311**	**0.1114**
7	rs2069845	IL-6	22770149	G	0.303	0.2379	A	0.1315	1.393	0.896	2.166	1	0.7228	0.00576148	0.2698
7	rs2069855	IL-6	22772624	C	0.005051	0.01456	T	0.4934	0.3435	0.03543	3.33	1	0.8653	0.00576148	0.009901
7	rs1818879	IL-6	22772727	A	0.2576	0.2822	G	0.614	0.8826	0.5673	1.373	1	0.8653	-0.0033028	0.27
12	rs2069728	IFN-ϒ	68547784	T	0.3081	0.3301	C	0.6321	0.9036	0.5945	1.374	1	0.8653	-0.00426685	0.3193
12	rs2069723	IFN-ϒ	68548594	C	0.01562	0	T	0.05568	NA	NA	NA	1	0.6107	0.00787836	0.007538
12	rs2069722	IFN-ϒ	68548953	A	0.075	0.05263	G	0.4018	1.459	0.5752	3.703	1	0.8653	-0.00282761	0.06129
12	rs2069720	IFN-ϒ	68549710	T	0.1263	0.1456	C	0.6144	0.8478	0.4791	1.5	1	0.8653	-0.00302951	0.1361
12	rs1861493	IFN-ϒ	68551196	G	0.06566	0.07767	A	0.6336	0.8345	0.3905	1.783	1	0.8653	-0.00389626	0.07178
12	rs2069713	IFN-ϒ	68552476	C	0.0101	0.009709	T	0.8117	1.041	0.1452	7.461	1	0.8928	-0.00494404	0.009901
12	rs2430561	IFN-ϒ	68552522	A	0.1429	0.1068	T	0.2598	1.394	0.7679	2.53	1	0.8653	0.000765078	0.1244
12	**rs78554979**	**IFN-ϒ**	**68554636**	**C**	**0.05102**	**0.01456**	**T**	**0.03478**	**3.638**	**0.9861**	**13.42**	**1**	**0.6107**	**0.0163105**	**0.03234**
12	rs2069705	IFN-ϒ	68555011	G	0.3131	0.2913	A	0.6268	1.109	0.7253	1.697	1	0.8653	-0.00439728	0.302
16	rs1801275	IL-4R	27374400	A	0.1414	0.1553	G	0.7283	0.8956	0.517	1.551	1	0.8928	-0.00453209	0.1485
16	rs7185840	HPR	72102112	A	0.1869	0.2524	G	0.1062	0.6806	0.4229	1.095	1	0.6902	0.00749627	0.2203
16	rs2021171	HPR	72110541	A	0.2172	0.1961	G	0.5809	1.137	0.7015	1.844	1	0.8653	-0.00311486	0.2065

19	rs11575934	IL-12RB1	18186618	C	0.1818	0.2206	T	0.3532	0.7852	0.4811	1.281	1	0.8653	0.00000416	0.2015
19	rs1736936	HLA-G	29794317	A	0.5303	0.4554	G	0.1483	1.35	0.9111	2	1	0.7416	0.00623214	0.4925
22	rs36086171	MIF	24235455	G	0.2576	0.297	A	0.4043	0.8211	0.5295	1.273	1	0.8653	-0.00273115	0.2775
22	rs9282783	MIF	24236359	G	0.08081	0.06796	C	0.639	1.206	0.5721	2.541	1	0.8653	-0.00519226	0.07426
22	rs35235644	MIF	24237822	C	0.08081	0.07282	G	0.7817	1.119	0.5378	2.33	1	0.8928	-0.00518057	0.07673
22	rs34383331	MIF	24238079	A	0.2879	0.301	T	0.7857	0.9389	0.6119	1.441	1	0.8928	-0.00452386	0.2946
22	rs136174	APOL1	36661536	C	0.0404	0.02451	A	0.3349	1.676	0.5387	5.213	1	0.8653	-0.000799842	0.03234
22	rs73885316	APOL1	36661674	A	0.0303	0.02451	C	0.656	1.244	0.3734	4.143	1	0.8702	-0.00424983	0.02736
22	rs136177	APOL1	36661842	G	0.07576	0.04902	A	0.2608	1.59	0.6967	3.629	1	0.8653	0.00103504	0.06219
22	rs73885319	APOL1	36661906	G	0.01515	0.01456	A	0.8428	1.041	0.2076	5.22	1	0.898	-0.00491464	0.06281
22	rs71785313	APOL1	36662046	T	0.05556	0.07	A	0.6115	0.7815	0.3458	1.766	1	0.8653	-0.00295232	0.06281

*Abbreviations: CHR = Chromosome, SNP = SNP ID, BP = Physical position (base-pair), A1 = Minor allele (based on whole sample), F_A = Frequency of allele 1 in cases, F_U = Frequency of allele 1 in controls, A2 = Major allele, P = p-value for this test, OR = Estimated odds ratio (for A1, i.e. A2 is reference), BONF = Bonferroni single-step adjusted p-values, FDR_BH = Benjamini & Hochberg (1995) step-up FDR control, FST = Fixation index, and MAF = Minor allele frequency. The level of significance is 0.05.

S4 Table: Prediction of nsSNP effect by MutPred, PredictSNP, SIFT, PROVEAN, and PANTHER algorithms

SNP ID (rs)	Amino acid Mutation	MutPred	PredictSNP	SIFT	PROVEAN	PANTHER
rs200500959	P2Q	0.85	DELETERIOUS	DELETERIOUS	DAMAGING	DISEASE
rs201465617	I5M	0.42	NEUTRAL	NEUTRAL	TOLERATED	NEUTRAL
rs201862457	P16Q	0.71	DELETERIOUS	DELETERIOUS	DAMAGING	NEUTRAL
rs533563568	L23F	0.74	NEUTRAL	NEUTRAL	TOLERATED	NEUTRAL
rs200995600	T24S	0.67	NEUTRAL	NEUTRAL	TOLERATED	NEUTRAL
rs766674911	A28V	0.88	DELETERIOUS	DELETERIOUS	DAMAGING	NEUTRAL
rs765744422	T31I	0.53	NEUTRAL	NEUTRAL	TOLERATED	NEUTRAL
rs201631604	P34T	0.86	DELETERIOUS	DELETERIOUS	DAMAGING	DISEASE
rs530185151	Y37H	0.66	DELETERIOUS	DELETERIOUS	DAMAGING	DISEASE
rs200005486	H41P	0.61	DELETERIOUS	DELETERIOUS	DAMAGING	DISEASE
rs11548059	P44L	0.70	NEUTRAL	DELETERIOUS	DAMAGING	NEUTRAL
rs747064040	P44A	0.70	DELETERIOUS	DELETERIOUS	DAMAGING	NEUTRAL
rs372575900	M48V	0.72	DELETERIOUS	DELETERIOUS	DAMAGING	NEUTRAL
rs560265113	G52C	0.73	DELETERIOUS	DELETERIOUS	DAMAGING	DISEASE
rs768313072	S54R	0.40	NEUTRAL	NEUTRAL	TOLERATED	NEUTRAL
rs761381011	A58T	0.88	DELETERIOUS	DELETERIOUS	TOLERATED	NEUTRAL
rs767129873	A58V	0.87	DELETERIOUS	NEUTRAL	DAMAGING	NEUTRAL
rs1049829	L59F	0.63	NEUTRAL	NEUTRAL	TOLERATED	NEUTRAL
rs532485317	I65M	0.89	DELETERIOUS	DELETERIOUS	DAMAGING	DISEASE
rs11548056	I68T	0.90	DELETERIOUS	DELETERIOUS	DAMAGING	NEUTRAL
rs200394994	A71T	0.47	NEUTRAL	NEUTRAL	TOLERATED	NEUTRAL
rs182012324	S75F	0.45	DELETERIOUS	DELETERIOUS	DAMAGING	NEUTRAL
rs780425386	L84R	0.89	DELETERIOUS	DELETERIOUS	DAMAGING	DISEASE
rs200286358	Y99C	0.88	DELETERIOUS	DELETERIOUS	DAMAGING	DISEASE
rs575017934	A105T	0.56	NEUTRAL	NEUTRAL	TOLERATED	NEUTRAL
rs1803976	N106S	0.53	DELETERIOUS	DELETERIOUS	DAMAGING	NEUTRAL
rs139210892	T113S	0.77	DELETERIOUS	DELETERIOUS	DAMAGING	NEUTRAL

S5 Table: Pearson Correlation coefficients between Average BC, 1/Average L, and RMSF

System	Chain	Pearson correlation		
		Average BC vs 1/Average L	*Average BC vs 1/RMSF*	*Average L vs RMSF*
MIF wildtype	A	0.893	0.571	0.498
	B	0.914	0.632	0.628
	C	0.898	0.580	0.427
P2Q	A	0.897	0.525	0.396
	B	0.922	0.715	0.665
	C	0.911	0.654	0.639
I5M	A	0.894	0.485	0.438
	B	0.900	0.433	0.381
	C	0.896	0.553	0.494
P16Q	A	0.897	0.608	0.515
	B	0.906	0.567	0.601
	C	0.899	0.664	0.569
L23F	A	0.920	0.554	0.300
	B	0.903	0.527	0.442
	C	0.905	0.450	0.500
T24S	A	0.866	0.515	0.434
	B	0.897	0.455	0.407
	C	0.907	0.650	0.480
A28V	A	0.908	0.591	0.611
	B	0.915	0.632	0.489
	C	0.903	0.620	0.526
T31I	A	0.921	0.623	0.561
	B	0.919	0.618	0.572
	C	0.903	0.689	0.540
P34T	A	0.922	0.654	0.566
	B	0.900	0.501	0.416
	C	0.889	0.515	0.487
Y37H	A	0.912	0.690	0.668
	B	0.912	0.655	0.501
	C	0.919	0.691	0.567
H41P	A	0.903	0.629	0.529
	B	0.884	0.472	0.503

	C	0.884	0.581	0.447
P44L	A	0.867	0.506	0.507
	B	0.882	0.615	0.526
	C	0.916	0.667	0.615
P44A	A	0.920	0.559	0.416
	B	0.875	0.599	0.543
	C	0.880	0.566	0.506
M48V	A	0.902	0.605	0.616
	B	0.926	0.581	0.408
	C	0.885	0.445	0.245
G52C	A	0.896	0.626	0.634
	B	0.908	0.580	0.518
	C	0.904	0.591	0.485
S54R	A	0.918	0.593	0.439
	B	0.914	0.661	0.556
	C	0.917	0.600	0.489
A58T	A	0.879	0.563	0.490
	B	0.920	0.581	0.498
	C	0.897	0.544	0.467
A58V	A	0.907	0.624	0.525
	B	0.901	0.619	0.553
	C	0.879	0.521	0.344
L59F	A	0.910	0.571	0.400
	B	0.885	0.625	0.567
	C	0.908	0.593	0.444
I65M	A	0.911	0.632	0.582
	B	0.887	0.604	0.649
	C	0.913	0.590	0.534
I68T	A	0.902	0.558	0.591
	B	0.914	0.567	0.361
	C	0.910	0.479	0.438
A71T	A	0.914	0.663	0.586
	B	0.897	0.603	0.552
	C	0.891	0.421	0.343
S75F	A	0.911	0.664	0.519
	B	0.896	0.576	0.306
	C	0.900	0.588	0.523

L84R	A	0.899	0.631	0.577
	B	0.915	0.598	0.528
	C	0.916	0.627	0.528
Y99C	A	0.907	0.588	0.444
	B	0.885	0.644	0.511
	C	0.911	0.555	0.504
A105T	A	0.909	0.623	0.497
	B	0.895	0.618	0.486
	C	0.913	0.618	0.429
N106S	A	0.903	0.505	0.406
	B	0.896	0.597	0.573
	C	0.903	0.508	0.354
T113S	A	0.885	0.540	0.437
	B	0.919	0.584	0.411
	C	0.911	0.566	0.462

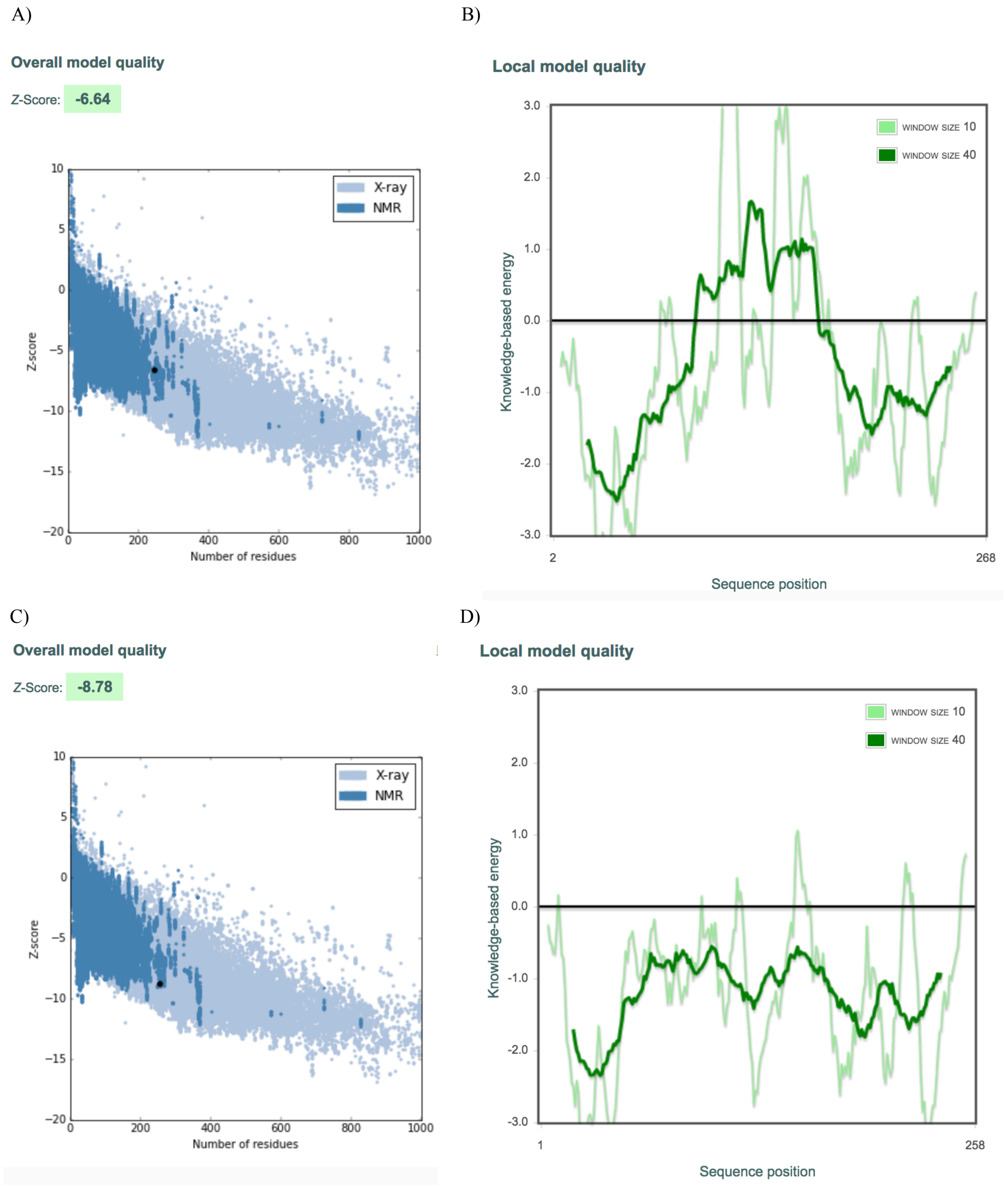
S6 Table: The change in dynamic cross correlation of residues in mutated MIF proteins

3DJH wildtype vs Mutant	Residues pairs whose motion changed from correlated to anti-correlated	Residues pairs whose motion changed from anti-correlated to correlated
P2Q	I65.B&K67.C, K67C&M1.B	
T31I	N98.A&V107.B,M1.B&F114.B, K67C&N98.A, K67C&F114.B	
Y37H	VI5.A & S75.C, N9.A & M48.B, N9.A & A49.B, N9.A & P50.B	N9.A & A71.A, N9.A & G72.A, N9.A & N73.A, N9.A & R74.A, N9.A & S75.A
G52C		A30C&K33A, T31.C&S64.A, G32.C&S64A
I65M	P16.B & D17.A, S14.B & D17.A, V15.B & D17.A	G18.A & V107.A, F19.A & V107.A
I68T	D17.A & A115.A; G18.B & V15.C	S14.B & D17.A; P16.B & D17.A
S75F	P16.A & V6.B, P16.A & V10.B, D17.A & V10.B, G18.A & V10.B	V6.A & G66.A,V6.A & I65.A
N106S	Q46.B & K67.C, L47.B & K67.C	K67.A & G66.C, K67.A & K67.C
T113S	G69.A & K78.C, G70.A & S75.C, G70.A & K78.C	L59.A & G66.A, G69.A & L59.A
I5M	G70.A & A71.C	A30.A & S54.A, T31.A & S54.A
P16Q	P35.A&P44.A, Q36.A&P44.A	
L23F	Y37.A & T31.C, Y37.A & G32.C	P44.A & G32.A, P44.A & K33.A
T24S	I65.B&K67.C	
H41P	A49.A & A13.C, S21.B & A13.C	P50.A & G69.A, P50.A & G70.A, E55.A & M102.B, C57.A & K67.B
P44L	G52.B & L62.B, G52.B & H63.B, G52.B & S64.B, Y100.B & G52.B	G52.B & D45.C, G52.B & R89.C, G52.B & I90.C
M48V	N110.A & G70.B, N110.A & M102.C , N110.A & I65.C	R12.A & N110.A, A13.A & N110.A, D17.A & N110.A, V43.A & N110.A
S54R		S64.A&L27.C, S64.A&Q29.C, S64.A&A30.C

S7 Table: The Pearson correlation coefficients for RMSF vs Average L, Average BC vs 1/(Average L), and Average BC vs 1/(RMSF).

	Pearson correlation coefficient		
System	**RMSF vs Average L**	**Average BC vs 1/Average L**	**Average BC vs 1/RMSF**
Apo protein	0.716	0.742	0.434
RUBi004	0.743	0.664	0.424
RUBi007	0.714	0.682	0.505
RUBi014	0.714	0.632	0.326
RUBi016	0.761	0.636	0.383
RUBi018	0.808	0.646	0.443

S1 Figure: Validation of *T. brucei* and *T. cruzi* PTR1 homology models using z-DOPE score and residue score using ProSA. The structural validation of *Tb*PTR1 is shown in A) and B) while that of *Tc*PTR1 in C) and D). Both models show overall reliable structural conformations.



S2 Figure: The 2D TbrPDEB1 protein-ligand interaction maps showing the binding modes of A) RUBi023, B) RUBi025, C) RUBi026, D) RUBi027, and E) RUBi029.