Computational studies in human African trypanosomiasis

A Thesis Submitted In Fulfilment Of The Requirements For The Degree Of

DOCTOR OF PHILOSOPHY

Department of Biochemistry and Microbiology

RHODES UNIVERSITY

by

Tendai Muronzi

ORCID ID: https://orcid.org/0000-0003-2616-7178

April, 2023

Abstract

Human African trypanosomiasis (HAT) is a neglected tropical disease (NTD) caused by two subspecies of the parasite, namely *Trypanosoma brucei (Tb) gambiense* (g-HAT) and *rhodesiense* (r-HAT). HAT is endemic in sub-Saharan countries, where the parasite transmission vectors, tsetse flies, breed. An estimated 70 million people remain at risk of contracting the disease, where the infection is classified as acute or chronic for g-HAT and r-HAT, respectively, with both forms ending in fatal meningoencephalitis when left untreated.

Both g-HAT and r-HAT are responsible for widespread fatal epidemics throughout sub-Saharan African history, resulting from the complex molecular interplay between trypanosomes and humans through unique, innate immunity evasion mechanisms. Of interest, the Tbr subspecies expresses a serum resistance-associated protein (SRA), which binds to human serum lytic factor, apolipoprotein L1 (ApoL1), nullifying any trypanocidal activity. In response, ApoL1 (G1 and G2) variants found in humans of sub-Saharan African lineage have been cited for conferring resistance to the r-HAT infection in an interaction that is not fully elucidated

In the event of successful infection, current HAT chemotherapeutics are plagued with complexity of administration, poor efficacy, toxicity, and potential drug resistance, highlighting a need for improved approaches. The parasite folate pathway provides a strategic target for alternative anti-trypanosomal drug development as trypanosomatids are folate auxotrophs, requiring host folate for growth and survival. Validated drug targets pteridine reductase (*Tb*PTR1) and dihydrofolate reductase (*Tb*DHFR) are essential for salvaging cofactors folate and folate biopterin crucial to parasite survival, making them viable targets for anti-folate investigation.

The overall aims of this thesis were to a) provide insights into the molecular and dynamic basis of the SRA and ApoL1 interplay in HAT infection and b) identify safer and more efficient anti-folate anti-trypanosomal drug alternatives through in silico approaches.

To achieve our first aim, in silico structure prediction was applied to generate 3D models of ApoL1 C-terminal variants G0, G1, G1G/M, G2 and G1G2, and four SRA variants retrieved from the NCBI database. The SRA and ApoL1 structures were inspected dynamically to identify the effect of the variants through molecular dynamics (MD) simulations. Dynamic

residue network (DRN) analysis of MD trajectories was fundamental in identifying residues playing a vital role in the intramolecular communication of both proteins in the presence of mutations. Protein-protein docking was then applied to calculate plausible SRA-ApoL1 C-terminal wild-type complex structures to further elucidate the nature of SRA-mediated infection. Through MD simulations, twelve SRA-ApoL1 dimeric structures were narrowed down from five to two energetically sound complexes. The two feasible SRA-ApoL1 complexes (1 and 2) exhibited favourable communication observed through DRN analysis, including the retaining key communication residues identified in prior monomer DRN calculations. ApoL1 C-terminal variants were additionally incorporated into SRA-ApoL1 complexes 1 and 2 for further complex dynamics analysis

This investigation into the nature of SRA-ApoL1 binding resulted in five primary outcomes: 1) highlighting the intramolecular effects ApoL1 variants have on the stability of the protein, 2) the identification of crucial SRA and ApoL1 communication residues in both monomeric or dimeric form, 3) the isolation of feasible SRA-ApoL1 complexes determined through global and local structural analyses 4) identification of residues crucial to the complex formation and maintenance of SRA-ApoL1, overlapping with those identified in (1), and 5) the minimal dissociative role of the G1 mutations in the complex, but compounding effect of the G2 deletion mutation.

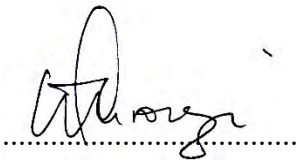
Computational modelling and drug repurposing were employed to achieve the thesis's second aim as they drastically cut down the costs involved in drug discovery and provide a more time-efficient screening method through numerous drug candidates. Using high throughput virtual screening, a subset of 2089 approved DrugBank compounds were screened against *Tb*PTR1. The outputs were filtered to 24 viable compounds in 54 binding poses using binding energy and molecular interactions. Through subsequent MD simulations of 200ns, thirteen potential hit compounds were identified. The resultant hit compounds were subjected to further blind docking against *Tb*DHFR and molecular dynamics to identify compounds with the potential for dual inhibition. The filtered subset was also tested in *in vitro* single concentration and dose-response bioassays to assess inhibitory properties against *Trypanosoma brucei*, complementing in silico findings. Post-molecular dynamics, four compounds exhibited high stabilities and molecular interactions with both *Tb*PTR1 and *Tb*DHFR, with two presenting favourable results in the *in vitro* assays. Three compounds additionally shared common structural moieties.

In all, the *in silico* repurposing highlighted drugs characterised by favourable interactions and stabilities in *Tb*PTR1, thus providing (1) a framework for further studies investigating anti-folate HAT compounds and (2) modulatory scaffolds based on identified moieties that can be used for the design of safe anti-folate trypanosomal drugs.

Declaration

I, Tendai Muronzi, declare that this is my own unaided work, except where duly acknowledged. It is being submitted for the degree of Doctor of Philosophy in Bioinformatics for the Faculty of Science at Rhodes University. It has not been submitted before for any degree for examination in any other university.

TENDAI MURONZI



.....

Date:

.....25/04/2023.....

Acknowledgements

Supervisor:

Professor Özlem Tastan Bishop for her tireless work and long nights to provide academic and financial support, and for her role an academic mentor. Thank you for all your faith and patience, without you this PhD would not have been possible.

Research:

Computational resources to perform this research were provided by the Centre for High Performance Computing (CHPC), Cape Town, South Africa.

Funding:

This work was supported through H3Africa [H3A-18-004]. The Human Heredity and Health in Africa (H3Africa) is a programme of the Alliance for Accelerating Excellence in Science in Africa (AESA) platform. AESA is a funding, agenda-setting, and programme management initiative of the African Academy of Sciences (AAS), the African Union Development Agency (AUDA-NEPAD), founding and funding global partners, and through a resolution of the summit of African Union Heads of Governments. H3Africa is supported by [the National Institute of Health (NIH), Wellcome, and African Society of Human Genetics (AfSHG)]. The views expressed herein are those of the author(s) and not necessarily those of AESA and her partners.

Personal:

It is impossible to list everyone who has been part of this journey, but specific mentions go to: To my parents and brother who been there every step of the way. My dear friends (and chosen family) Colleen Manyumwa, Chiamaka Okeke, Takudzwa Muvoti, Amanda Marumahoko, and Weston Kapungu for always having a ready ear for me. And Dr Fiona Robertson and Dr Magambo Phillip Kimuda.

Support:

I would also like to thank my extended family, the Muvotis, the Nyamatangas, the overall

Muronzi-Nyamatanga clan for providing support from all corners of the globe. And my colleagues in RUBi , for their readily available advice and assistance.

It truly takes a village.

For my grandmothers, Gogo Nyamatanga,

and the late Mbuya Muronzi.

Contents

Abstract ii

Declaration v

Acknowledgements vi

Contents ix

List of Equations xv

List of Tables xvi

List of Figures i

List of Supplemental Tables vii

List of Supplemental Figures ix

List of Abbreviations xi

Listing of Web Servers xiii

Listing of Amino Acids xiv

Thesis Overview xv

Chapter 1 1

Introduction 1

1.1 Literature Review 1
1.1.1 Disease background 1
1.1.2 Parasite Life-cycle 2
1.1.3 Disease management and treatment 3
1.1.4 Knowledge gap and motivation 5
1.2 Computational approaches 6
1.2.1 Protein structure prediction 6
1.2.2 Molecular docking 9
1.2.3 Molecular dynamics 13
1.2.4 Analyses of computational methods 17
1.3 Aims and motivations 24
1.4 Research Objectives 25

Chapter 2 27

Variations in ApoL1 27

2.1 Introduction 27
2.2 Methods 29
2.2.1 Homology modelling 29
2.2.2 Molecular dynamics simulations 30
2.2.3 Classical trajectory analysis 31
2.2.4 Principal Component Analysis 31
2.2.5 Dynamic Residue Network analysis 31
2.2.6 Contact map analysis 32
2.3 Results and Discussion 32
2.3.1 Homology modelling 32
2.3.2 Principal Component, Rg and RMSD analysis 34

2.3.3 Dynamic Residue Networks....36
2.3.4 Contact maps analysis....39
2.3.5 Hydrogen bond occupancy analysis....40
2.3 Conclusion....41

Chapter 3....43

Variations in SRA....43

3.1 Introduction....43
3.2 Methods....44
3.2.1 Data retrieval and Sequence analysis....44
3.2.2 Homology modelling....45
3.2.3 Molecular Dynamics....45
3.2.4 Intramolecular bond analysis....45
3.2.5 Comparative essential dynamics....45
3.2.6 Dynamic Residue Networks....46
3.3 Results and Discussion....46
3.3.1 Principal Component and RMSD analysis....46
3.3.2 RMSF and intermolecular hydrogen bonding....48
3.3.3 DRN analysis identified potential communication cold spots amid SRA mutations....49
3.3.4 Contact Maps....53
3.4 Conclusion....55

Chapter 4....57

Protein-protein interaction in SRA-ApoL1....57

4.1 Introduction....57

4.2 Methods 58
4.2.1 Protein-protein docking 59
4.2.2 Molecular dynamics 60
4.2.3 Trajectory analysis 60
4.2.4 Intramolecular bond analysis 60
4.2.5 Dynamic Residue Network analysis 61
4.2.6 Contact maps 61
4.2.7 Conformational Analysis 61
4.2.8 Protein-protein complex and interface validation 61
4.2.9 Interface alanine scanning 61
4.3 Results and Discussion 62
4.3.1 Identification of feasible SRA-ApoL1 complexes 62
4.3.2 Effect of ApoL1 variants on the complex 70
4.4 Conclusion 75

Chapter 5 77

Anti-folate Drug Discovery 77

5.1 Introduction 77
5.2 Methodology 79
5.2.1 Molecular Docking 79
5.2.2 Molecular Dynamics 81
5.2.3 Trypanosoma brucei experimental assays 83
5.3 Results and Discussions 84
5.3.1 Pharmacology of the hit compounds 84
5.3.2 Analysis of ligand dynamics 88
5.3.3 Analysis of receptor dynamics 90

5.3.4 Energetics and protein-ligand interactions 94
5.3.5 Four compounds exhibited potential anti-trypanosomal activity 103
5.4 Conclusion 105

Chapter 6 106

Overall Conclusions 106

6.1 Findings and their relevance 106
6.1.1 Pathogenicity 106
6.1.2 Drug discovery 107
6.2 Limitations 107
6.2.1 Molecular dynamics 107
6.2.2 Full-length structure of ApoL1 108
6.2.3 Lack of cytotoxicity assays 108
6.3 Future Work 108

References 111

Appendices 155

APPENDIX A : Chapter 2 155
APPENDIX B : Chapter 3 160
APPENDIX C : Chapter 4 162
Choice in protein-protein docking software 162
Additional calculations 162
APPENDIX D : Chapter 5 185

List of Equations

Equation 1: Simplified Force Field Calculation 16
Equation 2: Generalised RMSD calculation 17
Equation 3: Generalised Rg calculation 18
Equation 4: Generalised RMSF calculation 20
Equation 5: Averaged betweenness centrality calculations 21
Equation 6 Averaged closeness centrality calculations: 21
Equation 7: Averaged degree centrality calculations 21
Equation 8: Averaged eigencentrality calculations (i) 22
Equation 9: Averaged eigencentrality calculations (ii) 22
Equation 10: Averaged Katz centrality calculations 23
Equation 11: Calculation of the HADDOCK score 59
Equation 12: Calculation of MM-PBSA (i) 82
Equation 13: Calculation of MM-PBSA (ii) 83

List of Tables

Table 1: Homology model evaluation and validation scoring from varying programs. Model evaluation was conducted on the top Z-DOPE scoring structures. The Z-DOPE scores were calculated i) post model calculation and ii) post minimisation and a 20ns MD production run 32
Table 2: The docking input used for complex generation on the HADDOCK webserver. The AIRs were selected based on surface residues in the literature highlighted binding sites. Use of the bioinformatics settings increased the sampling size of the calculations 60
Table 3: Probable stability of SRA-ApoL1 complexes in solution as predicted by the PDBePISA server. ΔGint indicates the solvation free energy gained upon formation of the complex, while ΔGdiss indicates the free energy of complex dissociation. Complexes with ΔGdiss>0 are likely to be thermodynamically stable in solution 66
Table 4: Destabilising protein-protein interface residues identified through in silico alanine scanning for each dimeric complex and their G1, G1-M, G1-G and G2 variants. BC hubs are in bold, and CC hubs are italicised. 73
Table 5: The chemical structures and classification of the top DrugBank compounds. The post-docking binding energies of the compounds in each protein are shown. 85
Table 6: A decomposition of the binding energy components obtained from MM-PBSA calculations in TbPTR1. 95
Table 7: A decomposition of the binding energy components obtained from MM-PBSA calculations in TbDHFR. 97

List of Figures

Figure 1: A schematic illustration of the life cycle of T. brucei in the human host and tsetse fly vector. Developmental forms of the parasite are shown in both hosts. The long slender parasites proliferate in the human bloodstream, infecting the blood as well as the tissue (such as the brain). The parasites then differentiate into the non-proliferative short stumpy form, which is taken up by the tsetse fly, morphing into the procyclic form. Procyclic parasites rapidly divide in the mid-gut of the fly and transition into non-dividing metacyclic forms in the fly salivary glands for transmission into the human bloodstream via a bite..... 3

Figure 2: A schematic illustration of the methodological objectives of the overall project for research aims 1 and 2.....25

Figure 3: **A)** The C-terminal of ApoL1, showing the SRA-binding site and **B)** the location of the G1 (S342 and I384) and G2 (N388/Y389) mutations in firebrick and salmon, respectively. (Structure source : Homology model).....28

Figure 4: Model validation of the ApoL1 C-terminal structures from **A)** ProSA, **B)** PROCHECK and **C)** ERRAT servers. The RMSF of the ApoL1 variants post MD simulations in **D** reveal the flexible loop regions of each variant that are highlighted on the structure in **E**. The significant number of loops in the structures contributes to the validation scoring. (Structure source : Homology model) 34

Figure 5: Essential Dynamics of ApoL1 C-terminal structures for **A)** G0, G1,G1M and G1G and **B)** G2 and G2G1 variants of the ApoL1 C-terminal structures. The structures in A are 98 residues long, while those in B are 96 residues long due to the G2 deletion. The **C)** RMSD and **D)** Rg values of the structures over the 400ns simulations are represented in kernel density violin plots in 35

Figure 6: Heatmaps of DRN calculation, showing the top 5% values of the centrality metrics **A)** BC, **B)** CC, **C)** DC, and **D)** Katz Centrality of ApoL1 variants. Residues detected as hubs are annotated with the centrality value in the cell. The increment of centrality

values for each metric corresponds with the intensity of the colour bar on the right of each heatmap.....38
Figure 7: Contact heatmap analysis was conducted on residues that were sites of mutation in the ApoL1 C-terminal, namely residues **A)** 342, **B)** 384 and **C)** 388/390. Due to the deletion of 388 and 389 in the G2 variants, 388 in C is compared with 390. Notable changes in the contact network of high-profile residues due to the substitution and deletions are encircled in red. 39
Figure 8: Hydrogen bonding between αH1 and αH2 residues was tracked over 300ns in hydrogen bond occupancy. Occupancy is ranked from 0-1 (in the colour bar), with an occupancy of 1 signifying the continuous maintenance of the bond throughout the simulation 41
Figure 9: **A)** The structure of SRA (PDB ID:6elu), consisting of eight helices, represented in **B**, including the location of Cysteines 101 and 180 of the disulphide bridge in maroon. Helices H1, H2, H3, H4, H5, H6, H7, and H8 are highlighted in grey, black, yellow, teal, green, orange, and navy blue , respectively. The multiple sequence alignments in **C)** present the location of the SNVs in the four SRA variants identified in sub-Saharan Africa with the asterisks. The locations of the SNVs identified are mapped onto the structure in A as spheres, corresponding MSA colouring. 44
Figure 10: **A)** Essential dynamics of the SRA variants, determined through PCA calculated over times (300ns). The colour progression of the colour bar characterises the progression of the time in MD simulation. PC1 captures 55.54%, and PC2 14.44%. The RMSD of the protein backbones, in **B**, plotted in line and kernel density violin plots, is representative of the global motion exhibited during the simulation and is mirrored in the PCA. 47
Figure 11: **A)** Backbone RMSF and the **B)** hydrogen bonding occupancy of the SRA variants over the course of the 300ns simulations. Notable regions of high flexibility in SRA in A and low hydrogen bond occupancy in B coincident with the loop region of SRA highlighted navy blue in the structure in **C.**.....49
Figure 12: Persistent hubs identified in centrality metrics **A)** BC, **B)** CC, **C)** DC and **D)** KC are mapped onto the SRA structure in the coloured spheres, highlighting the position of the hubs. Each metric emphasises a different measure of importance within the network, resulting in a unique distribution of persistent hubs. (Structure source : PDB ID 6elu).....50

Figure 13: The top 5% centrality hubs of metrics **A)** BC, **B)** CC, **C)** DC, and **D)** KC per SRA variant. The annotated heatmap cells represent identified hubs meeting the 5% cut-off. Persistent hubs were those observed in all complexes per metric. Low to high centrality values for each metric correspond to the intensity of the colour bar on the right of the each heatmap. 52
Figure 14: Contact heatmaps for SRA variants highlighting sites of mutation or high communication residues identified through DRN analysis. The frequency of contacts between residues in a network was monitored over the course of the MD simulations and is represented on a scale from 0.0-1.0 in the colour on the bar on the right. 54
Figure 15: A schematic of the relationship between SRA and ApoL1, and its effect on the pathogenicity of HAT presented in a simplified correlation matrix. 58
Figure 16: **A)** The principal component analysis (PCA) of the SRA-ApoL1 complexes over the length of the MD simulations. The CoM, RMSD and Rg calculations for the complexes are shown in line **(C-E)** and violin plots **(Cii & Dii)**, with a colour key presented at the bottom of the plots. Correlation is observed between the backbone flexibility calculations and the conformational space sampled in the PCA. 63
Figure 17: The PCA calculations identify the low energy conformations presented in the figure above , which show the unique binding orientations of the monomers per complex differentiated unique ApoL1 colouring. 65
Figure 18: High **A)** BC and **B)** CC residues mapped onto energetically stable minima structures for five complexes. Paths of communication (in white) can be traced along the high centrality scoring hub residues in grey for ApoL1 and in blue for SRA, showing top 10% and top 5% residues, respectively. The persistent hubs in SRA are denoted as "significant hubs". (Structure source : HADDOCK protein docking output) 68
Figure 19: Top 5% and 10 % **A)** BC and **B)** CC hubs detected for SRA and ApoL1 are denoted by the annotation in their heatmap cells compared to their residue counterparts. Persisents hubs were those observed in all complexes per metric. Low to high centrality values for each metric correspond to the intensity of the colour bar on the right of the each heatmap. 69
Figure 20: **A)** BC and **B)** CC hubs observed in the SRA-ApoL1 complexes over a stabilised section of the MD simulations. Detected hubs within the cut-off % were annotated with their centrality values for comparison against counterpart residues in other variants. Low to high centrality values for each metric correspond to the intensity

of the colour bar on the right. Key residues were noted to maintain hub status in all the variant complexes for both metrics 71
Figure 21: Alanine scanning highlighted destabilising residues that were common across all the SRA-ApoL1 variant complexes **(i)**. The residues were found to be in the binding interface, as illustrated in **(ii)**. Destabilising alanine substitutions had free energy values (ddG) ≥1 kcal mol-1, while neutral values ranged from -0.8 to 0.99 kcal mol^{-1}, which is represented on the colour bar. (Structure source : HADDOCK protein docking output) 72
Figure 22: Contact heatmaps for ApoL1 mutation residues **A)** Ser/Gly342, **B)** Ile/Met384, **C)** Asn388, and **D)** Tyr389 in the SRA-ApoL1 wildtype complexes. The inter-residue contact frequency was tracked over 500ns MD simulations and is denoted by the colour bar at the bottom of the figure. Complexes 1 and 2 G2 contacts are absent in C and D due to the deletion of Asn388 and Tyr389 in both. 75
Figure 23: The structures of **a)** TbDHFR (PDB ID:3qfx) and **b)** TbPTR1 (PDB ID:2x9n) . Residues active in the substrate binding and reduction reactions are shown in maroon, as well as the substrate-binding loop of PTR1. The NADPH co-factor (shown in stick figure representation), crucial to the catalytic activity, is shown in green, blue, and orange. The native substrates of both proteins are shown as 2D stick figures, highlighting their structural similarity, pathway, and the end product of the reductions. 78
Figure 24: The azabicyclic and indole moieties are shared across the compounds **ai)** DB00424, **aii)** DB00572, **b)** DB00757, **c)** DB02959, and **d)** DB11699. (Structure source : DRUGBANK) 87
Figure 25: RMSD of the ligands in **a)** TbPTR1 and **b)** TbDHFR presented in line plots. The median of the RMSD values is presented in the **c)** kernel density violin plot as a crossbar. 89
Figure 26: Protein backbone RMSD of ligand-bound TbPTR1 presented as a factor of the free energy of the protein. The colour intensity in the colour bar in the right presents the free energy (ΔG) in $kcalmol^{-1}$. The inserts in each plot present the kernel plots density of each system's RMSD, with corresponding peaks showing the lowest energy conformations. 91

Figure 27: **A**) RMSD and **B**) Rg of the protein backbone in ligand-bound TbPTR1 and TbDHFR presented in violin plots. The width of the plots outlines kernel probability density. Box plots represent the median ± the standard deviation. 92
Figure 28: RMSF of the **A**) TbPTR1 backbone, with the corresponding flexible regions shown in the **B**) protein structure. The MM-PBSA, per residue energy contribution, is shown in **C**, with the energy range represented in the colour bar. Positions of residues contributing the most energy are highlighted in **D**. The catalytically essential residues are coloured red in both structures **B** and **D**. (Structure source : PDB 2x9n).....94
Figure 29: MM-PBSA decomposition of residues in TbDHFR (in **A**) represented in a heatmap for each protein-ligand complex. The energy contribution of each residue is shown in **B**, with the residues attributing most to the energy highlighted in the structure **A** in salmon. (Structure source : PDB 3qfx) 96
Figure 30: The intermolecular interactions of ligands **a)** DB00424, **b)** DB00472, **c)** DB00572, **d)** DB00757, **e)** DB00809, **f)** DB00891, **g)** DB00949, **h)** DB01151, **i)** DB02959, **j)** DB11254, **k)** DB11282, **l)** DB11285 and **m)** DB11699 in the TbPTR1 active site are shown in 3D representation.The inset present a schematic of each binding pose, highlighting each interaction according to the provided key. 99
Figure 31: Intermolecular interactions between the ligands **a)** DB00424, **b)** DB00472, **c)** DB00572, **d)** DB00757, **e)** DB00809, **f)** DB00891, **g)** DB00949, **h)** DB01151, **i)** DB02959, **j)** DB11254, **k)** DB11282, **l)** DB11285 and **m)** DB11699 and residues in the TbDHFR the active site are presented in 3D structure representation. A schematic outline of the binding pose is shown in the insets, with an accompanying key for the type of interactions. 101
Figure 33: Hydrogen bonding occupancy of the hit ligands with the binding pocket residues in **A)** TbPTR1 and **B)** TbDHFR. Occupancy is presented as a percentage of the total number of frames in the MD simulations..... 103
Figure 34: Anti-trypanosomal activity bioassays: **A)** Single concentration assays. Percentage parasite viability (% viability ± SD) of all 13 compounds after single concentration assays at 20 πM. **B)** Pentamidine was used as the assay control and run in duplicate, as well as co-crystallised compounds (cyromazine (AX3) and pyrimethamine (CP6) shown in **C**). Viability of DB00472, DB00572, DB01151, DB11254, AX3 and CP6 (highlighted in red) were determined to be 0.40 ± 0.46, 0.14 ± 0.42, 0.32 ± 0.33 and 0.10

± 0.09 μM, respectively. Pentamidine % viability was 0.025-0.026 μM. **C)** IC_{50} values of compounds represented in dose-response plots and % viability ± SD..... 104

List of Supplemental Tables

Table S 1: Statistical analysis for HADDOCK generated SRA-Apo L1 complexes from docking run 1. The HADDOCK score for each cluster was calculated using a summation equation of thermodynamic energies. From the top-scoring cluster, 1 (in blue) complexes 1 and 2 were obtained and used in further investigations. 163
Table S 2: Statistical analysis for HADDOCK generated SRA-ApoL1 complexes from docking run 2. The HADDOCK score for each cluster was calculated using a summation equation of thermodynamic energies. From the top-scoring cluster 4 (in blue), complexes 3 and 4 used in further investigations were obtained. 164
Table S 3: Statistical analysis for HADDOCK generated SRA-ApoL1 docked complexes from run 3. The HADDOCK score for each cluster was calculated using a summation equation of thermodynamic energies. From the top-scoring cluster 9 (in blue), complexes 5 and 6 used in further investigations were obtained 165
Table S 4: The code names of the complexes from each of the docking runs..... 166
Table S 5: Destabilising protein-protein interface residues identified for each dimeric complex through in silico alanine scanning..... 167
Table S 6: SRA interface residues identified by computational alanine scanning. The residues in each complex were singly mutated to alanine using the ROBETTA webserver. The change in binding free energy indicates either destabilisation or stabilisation of the complex. The free energy scores in blue have a neutral effect on the complex..... 168
Table S 7: ApoL1 interface residues identified by computational alanine scanning. The residues in each complex were singly mutated to alanine using the ROBETTA webserver. The change in binding free energy indicates either destabilisation or stabilisation of the complex. The free energy scores in blue have a neutral effect on the complex..... 169
Table S 8: Probable stability of SRA-ApoL1 complexes in solution as predicted by the PDBePISA server. ΔGint indicates the solvation free energy gained upon formation of

the complex, while ΔGdiss indicates the free energy of complex dissociation. Complexes with ΔGdiss>0 are likely to be thermodynamically stable in solution 170
Table S 9: Q-Vina W molecular docking parameters 185
Table S 10: The median values of ligand and backbone RMSD in TbDHFR 185
Table S 11: The median values of ligand and backbone RMSD in TbPTR1 187
Table S 12: The p-values of protein backbone RMSD distribution in TbPTR1. The values were calculated using the Shapiro Wilks test for normality. 188
Table S 13: The p-values of ligand RMSD distribution in TbPTR1. The values were calculated using the Shapiro Wilks test for normality. 189
Table S 14: The p-values of radius of gyration in TbPTR1. The values were calculated using the Shapiro Wilks test for normality. 190
Table S 15: The p-values of protein backbone RMSD distribution in TbDHFR. The values were calculated using the Shapiro Wilks test for normality 191
Table S 16: The p-values of ligand RMSD distribution in TbDHFR. The values were calculated using the Shapiro Wilks test for normality. 192
Table S 17: The p-values of radius of gyration distribution in TbDHFR. The values were calculated using the Shapiro Wilks test for normality. 193
Table S 18: The RMSD values at the peaks of the distribution in TbPTR1 and TbDHFR. 194
Table S 19: The BBB properties of the hit compounds identified. 195

List of Supplemental Figures

Figure S 1: ApoL1 C-terminal structure prediction. Template suggestions were included from A) PsiPred, B) Robetta C) and D) Quick2D (HHPred) 155
Figure S 2: The RMSD and Rg recorded for the ApoL1 variants over 400ns simulations, represented as line plots. 156
Figure S 3: Overlayed structures of ApoL1 C-terminal at 100ns.. The mutating residues are highlighted in the stick representation 157
Figure S 4: Overall averaged DRN values of BC, CC, DC and KC of the ApoL1 c-terminal variants over 300ns trajectories. 158
Figure S 5: Significant DRN hubs Phe343, Val346, 350 and 353 are mapped onto the ApoL1 C-terminal structure in orange. G1 and G2 mutation residues are highlighted in red and salmon, respectively. 159
Figure S 6: Overall averaged DRN values of BC, CC, DC and KC of the SRA variants over 300ns trajectories. 160
Figure S 7: The length of the Cys101-Cys180 disulphide bridge was monitored for each SRA variant over the course of the simulations. The bond length is represented in both line and violin plots. 161
Figure S 8: RMSD of server generated SRA-ApoL1 complexes. RMSD was used to investigate stability of the complexes over the course of the simulations. ZDOCK, HDOCK and ClusPro output complexes dissociated after 20ns 171
Figure S 9: Structures of SRA (in teal) and the ApoL1 C-terminal region (in pale blue) highlighting the AIR residues applied for HADDOCK protein-protein docking. SRA AIRs for Run 1 and 2 are coloured in raspberry, while Run 3 is distinguished by the orange colouring. AIRs applied for ApoL1 coiled-coil are coloured in hot pink 172
Figure S 10: HADDOCK run 1. The clusters represent are ranked based on FCC, and molecular energies. 173
Figure S 11: HADDOCK run 2. The clusters represent are ranked based on FCC, and molecular energies 174

Figure S 12: HADDOCK run 3. The clusters represent are ranked based on FCC, and molecular energies. 175
Figure S 13: Stacked bar graph of the raw BC values of complexed wild-type SRA and ApoL1 over 500ns 176
Figure S 14: Clustered bar graph of the raw CC values of complexed wild-type SRA and ApoL1 over 500ns 177
Figure S 15: Contact map analysis of key residues in wild-type SRA-ApoL1 complexes over 500ns 178
Figure S 16: Alanine scanning of wild-type SRA-ApoL1 complexes from an energy minima conformation. 179
Figure S 17: Normal Mode analysis of the wild-type SRA-ApoL1complexes.The arrows are coloured based on each chain, denoting both the extent of motion and direction with respect to each of the residues. 180
Figure S 18: RMSF analysis of wild-type SRA-ApoL1 complexes. 181
Figure S 19: All averaged centrality values of the SRA-ApoL1 variants complexes. A) SRA and B) ApoL1 values for i) BC and ii) CC over 500ns are measured 182
Figure S 20: Mapped BC an CC values onto the complex structures form the energy minima 183
Figure S 21: BC and CC hubs observed in the SRA-ApoL1 complexes over a stabilised section of the MD simulations of a duplicate 500ns run. Detected hubs within the cut-off % were annotated with their centrality values for comparison against counterpart residues in other variants. For each metric, low to high centrality values correspond to the intensity of the colour bar on the right. 184
Figure S 22: RMSD and free energy landscapes of the TbDHFR-ligand complexes. The inserts show probability density distribution of RMSD, with the peaks corresponding with the lowest energy conformation. 196
Figure S 23: RMSF of the TbDHFR-ligand complex protein backbone. The flexible regions are show in the dotted lines in both the a) line graph and b) the 3D structure. 197

List of Abbreviations

Abbreviation	Definition
3D	Three dimensional
AAT	Animal African Trypanosomiasis
ASA	Accessible Surface Area
ApoL1	Apolipoprotein 1
BBB	Blood-Brain Barrier
BC	*Betweenness Centrality*
BSA	Buried Surface Area
CC	*Closeness Centrality*
CKD	Chronic Kidney Disease
CHPC	Centre for **High-Performance Computing**
CNS	Central Nervous System
CoM	Centre of Mass
CryO-EM	Cryogenic Electron Microscopy
CSS	Complex Significance Score
DC	*Degree Centrality*
DS	BIOVIA Discovery Studio
DHFR	Dihydrofolate Reductase
DNA	Deoxyribonucleic acid
DOPE	Discrete optimised protein energy
DRN	Dynamic Residue Networks
EC	*EigenCentrality*
FMAX	Force maximum
GWAS	Genome-wide association studies
HAT	Human African Trypanosomiasis
HDL	High-Density Lipoprotein
HDX-MS	Hydrogen Deuterium Exchange Mass Spectrometry
HTVS	High-Throughput Virtual Screening
KC	*Katz Centrality*
LINCS	Linear Constraints Solver
LZD	Leucine Zipper domain
MD	Molecular dynamics
MM-PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
MSA	Multiple sequence alignment
NAD/P	Nicotinamide Adenine Dinucleotide/Phosphate
NTD	Neglected Tropical Disease

NMA	Normal Mode Analysis
PAINS	Pan Assay Interference assays
PBC	Periodic Boundary Conditions
PCA	Principal component analysis
PCR	Polymerase Chain Reaction
PME	Particle Mesh Ewald
PPI	Protein-protein interaction
PTR1	Pteridine Reductase 1
Rg	Radius of gyration
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SDR	Short-chain Dehydrogenase
SNV/P	Single nucleotide variation/polymorphism
SRA	Serum Resistance-Associated protein
SVM	Support Vector Machine
Tb	*Trypanosoma brucei*
TCA	Tricyclic Anti-depressant
TFL	Trypanosome Lytic Factor
vdW	Van der Waals
WHO	World Health Organisation
WT	Wild-type

Listing of Web Servers

Web server	URL
ALASCAN	*https://robetta.bakerlab.org/alascansubmit.jsp*
Ensembl	*https://www.ensembl.org/index.html*
HHPred	*https://toolkit.tuebingen.mpg.de/tools/hhpred*
HADDOCK	*https://wenmr.science.uu.nl/haddock2.4*
NCBI	*https://www.ncbi.nlm.nih.gov/*
PDBePISA	*https://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver*
PlayMolecule	*https://playmolecule.com*
PRODIGY	*https://wenmr.science.uu.nl/prodigy*
RCSB PDB	*https://www.rcsb.org*
ROBETTA	*https://robetta.bakerlab.org*
SAVES	*https://saves.mbi.ucla.edu*

Listing of Amino Acids

Single letter	Three letter	Amino acid
A	ALA	Alanine
C	CYS	Cysteine
D	ASP	Aspartic acid
E	GLU	Glutamic acid
F	PHE	Phenylalanine
G	GLY	Glycine
H	HIS	Histidine
I	ILE	Isoleucine
K	LYS	Lysine
L	LEU	Leucine
M	MET	Methionine
N	ASN	Asparagine
P	PRO	Proline
Q	GLN	Glutamine
R	ARG	Arginine
S	SER	Serine
T	THR	Threonine
V	VAL	Valine
W	TRP	Tryptophan
Y	TYR	Tyrosine

Thesis Overview

The main objective of the research was to investigate human African trypanosomiasis (HAT) as: 1) a result of a molecular arms race between the causative parasites and humans, and 2) a target for anti-folate drug discovery.

CHAPTER 1

This chapter introduces African trypanosomes and the disease human African trypanosomiasis (HAT) caused by the trypanosome subspecies. Additionally, computational methods typically used to study protein structural dynamics are highlighted.

CHAPTER 2

Apolipoprotein L1 (ApoL1) is introduced in chapter 2. The ApoL1 C-terminal variants G0, G1, And G2 are modelled, and the structural consequences of the mutations are highlighted.

CHAPTER 3

Chapter 3 describes variants in *Trypanosoma brucei rhodesiense (Tbr*) Serum Resistance-associated protein (SRA). Variants of the protein retrieved from the NCBI database were structurally compared, leading to the identification of mutational cold spots.

CHAPTER 4

This chapter expands on Chapters 2 and 3 by incorporating SRA and ApoL1 into protein-protein docking. Plausible SRA-ApoL1 complexes are calculated, and the effect of ApoL1 variants in the complexation investigation.

CHAPTER 5

Chapter 5 describes the use of high throughput virtual screening (HTVS) for drug discovery. Potential anti-folate compounds are identified though docking a DrugBank dataset against *Tb*PTR1 and *Tb*DHFR.

CHAPTER 6

Finally, the findings from chapters 2-5 are summarised. The significance of the research outcomes is highlighted in addition to detailing future work.

Chapter 1

Introduction

1.1 Literature Review

1.1.1 Disease background

African trypanosomiasis is a disease endemic in sub-Saharan African countries [1,2] where the transmission vectors, tsetse flies *(Glossina* spp.) [3], breed. Compared to less conventional transmission means, such as cross-placental [4–6] and sexual contact [7,8], a tsetse fly bite is the most prevalent. Protozoan parasites, African trypanosomes *Trypanosoma brucei*, cause the disease to varying infections and hosts, depending on the infecting parasite subspecies [9]. *Trypanosoma brucei brucei (Tbb)*, for example, causes animal African trypanosomiasis (AAT), also known as Nagana, in livestock but cannot infect humans due to lysis by human serum. Nagana is a wasting disease that contributes significantly to the socio-economic burden of African trypanosomes [10,11] as infected animals seldom survive.

Trypanosoma brucei rhodesiense (Tbg) and *Trypanosoma brucei gambiense (Tbg)* are responsible for human African trypanosomiasis (HAT), with each causing geographically well-delineated infections that present unique characteristics in the human host. *Trypanosoma brucei rhodesiense* HAT (r-HAT) is an acute and fast-progressing [12,13] zoonotic infection, primarily isolated to eastern and southern Africa, while *Trypanosoma brucei gambiense* (g-HAT) causes an anthroponotic [14] chronic disease localised to western and central Africa. Though both parasite subspecies are non-pathogenic in domestic and wild animals, the animals serve as epidemiologically significant reservoir hosts [9,15].

HAT pathology is divided into the early haemo-lymphatic and the meningo-encephalitic stage [16,17]. In the early haemo-lymphatic stage, following an infected tsetse fly bite, the parasites proliferate in the blood and lymphatic system. This first stage of the disease is initially asymptomatic, save a chancre at the site of inoculation [18]. As the infection progresses, it is

characterised by several symptoms, including general malaise, headaches, fever, lymphadenopathy, arthralgias, and myalgias.

Progression to the second stage has been estimated at over one year and three weeks for *Tbg,* [19,20] and *Tbr* HAT [21], respectively. In the final stages (the meningo-encephalitic stage), the parasites cross the blood-brain barrier (BBB) [22] into the central nervous system (CNS) [23] causing severe mental health complications such as sensory disturbances or insomnia and sleep cycle disruptions [16,17,24], hence the colloquial name. Additional complications and symptoms include motor problems such as tremors, ataxia, endocrinological disturbances, seizures and ultimately, fatal comas [25].

The parasites undergo antigenic variation to effectively evade the host's immune system [26–28], making preventive vaccines challenging to develop.

1.1.2 Parasite Life-cycle

Once a host is infected, *T. brucei* undergoes several developmental morphological and enzymatic changes in its life cycle in ***Figure 1*** to facilitate host-host transfer of the parasite and adapt to the different environments within the tsetse vector such as the midgut or the mammalian host lymphatic system and the central nervous system. When an infected tsetse fly bites a human host, the metacyclic forms of the parasite from its saliva are introduced into the host's bloodstream. The parasites multiply at the site of infection, passing between the lymph and blood [29,30]. The parasites differentiate into the long slender form in the bloodstream infecting various tissue and organs. These cells express the bloodstream-stage-specific VSG coat to evade the mammalian immune response [31]. During the progression of infection, the parasites morph into the short stumpy form that stays dormant in the blood, ready for vector feeding (***Figure 1***). The stumpy form exists in division-arrested forms to limit the increase in parasite numbers, prolonging host survival for further transmission [31]. The tsetse fly segment of the cycle takes approximately three weeks, beginning when the fly feeds on an infected blood meal [29]. The stumpy parasite forms differentiate into procyclic forms and resume proliferation in the insect mid-gut. The division is arrested once more as the procyclic parasite travels to the tsetse fly salivary glands, transforming into the intermediate epimastigotes, then the human-infective metacyclic trypomastigotes [29,30].

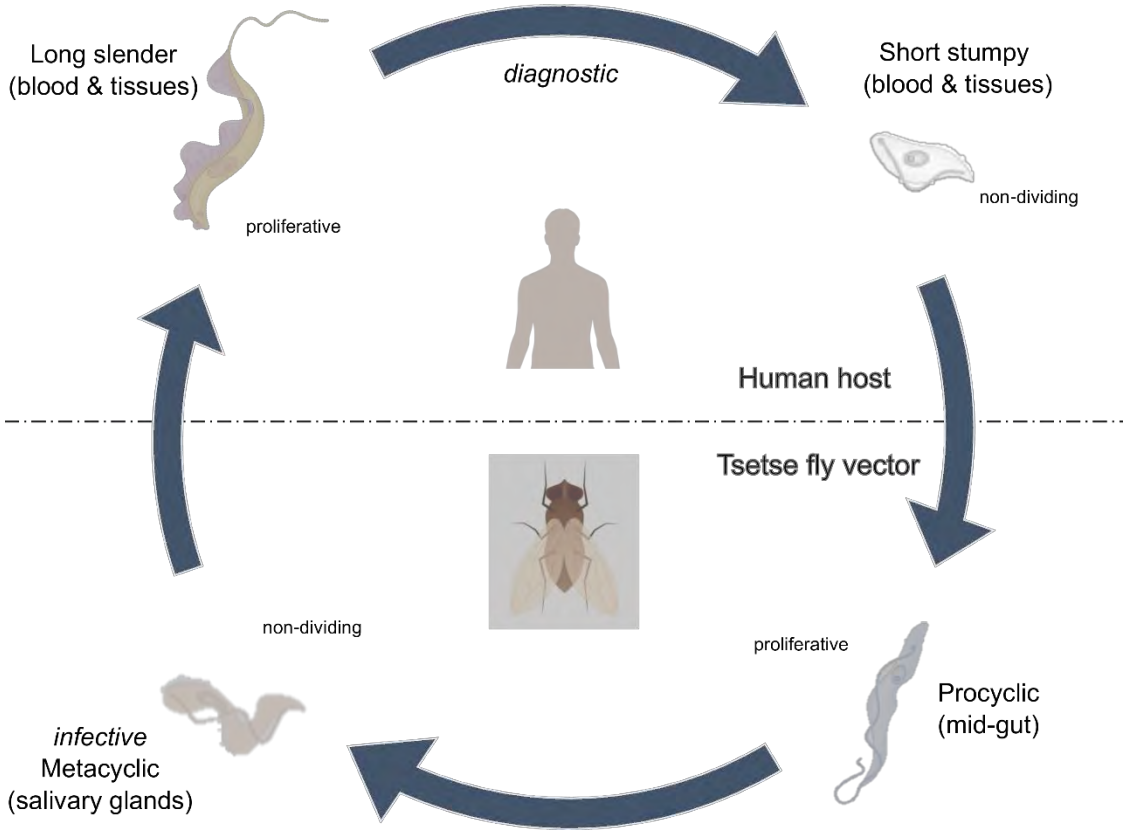


Figure 1: A schematic illustration of the life cycle of T. brucei *in the human host and tsetse fly vector. Developmental forms of the parasite are shown in both hosts. The long slender parasites proliferate in the human bloodstream, infecting the blood as well as the tissue (such as the brain). The parasites then differentiate into the non-proliferative short stumpy form, which is taken up by the tsetse fly, morphing into the procyclic form. Procyclic parasites rapidly divide in the mid-gut of the fly and transition into non-dividing metacyclic forms in the fly salivary glands for transmission into the human bloodstream via a bite.*

1.1.3 Disease management and treatment

Sleep sickness management uses several strategies, spanning prevention, diagnosis, and treatment.

- Prevention: As no vaccines are available for HAT, prevention aims to reduce disease incidence by disrupting the parasite life cycle. Controlling the tsetse fly population is vital to prevention as it curtails transmission [32,33]. Vector control measures such as using insecticides and clearing bushes near residential areas effectively reduce the incidence of the disease. These techniques, as well as rudimentary ones such as wearing protective clothing and use of repellent, work well when combined with educating the population in affected areas on risks and prevention endeavours.

Infection control in livestock is also vital to disease prevention, as they serve as a parasite reservoir [14,34,35]. This can be done through uses the same methods to protect livestock from tsetse flies through the use of insecticide-impregnated ear tags or mass anti-parasite drug administration

- Early-stage diagnosis: Screening of the population in endemic areas and any travellers to the area play a crucial role in preventing the spread of the parasite as both infected human and animal hosts act as a reservoir for the parasite [14,35]. Diagnosis is also vital for medical intervention to prevent the death of infected individuals.

  Diagnosing sleeping sickness can be challenging due to vague or absent symptoms in the early stages and the need to test cerebrospinal fluid for conclusive results, hence the need for frequent screening. Diagnosis methods such as physical examination, blood and serological tests, lymph node biopsy, and molecular (PCR) tests are often used [36].

- Treatment: The administration of drugs to kill the parasites and cure the disease requires knowledge of the stage of progression to and the parasite species causing the infection. Treatment typically involves chemotherapy that targets the parasites, such as pentamidine or suramin for the early stage of the disease [37], and melarsoprol [38] or eflornithine [39,40] for the late stage. Eflornithine is also used in a recommended combination therapy called NECT [41,42] , with nifurtimox [43], a drug used in Chagas disease treatment [44]. Fexinidazole [45] is an orally administered drug, recently available from 2020, that can be used against both stages of g-HAT, while acoziborole [46] is in the final stages of promising clinical trials [47]. Patients treated for HAT ideally receive follow-up surveillance and care to ensure parasite clearance.

Overall, the management of sleeping sickness requires a comprehensive approach that includes prevention, early diagnosis, and effective treatment. Efforts are ongoing to improve surveillance, control and treatment of the disease in affected regions [48]. To date, through the application of the HAT management techniques, Benin, Uganda, and Rwanda have received validation from the World Health Organisation (WHO) for elimination of at least one form of HAT as a public health problem [48,49]. Uganda, where both g-HAT and r-HAT are endemic, has achieved elimination of the former and continues efforts in targeting r-HAT.

1.1.4 Knowledge gap and motivation

As HAT mostly affects rural, resource-poor communities in sub-Saharan Africa, HAT is considered a neglected tropical disease (NTD). During the 20th century, it was responsible for a number of devastating epidemics [50,51], with the latest upsurge in infections in the late 1990s [52] peaking at a reported over 25,000 cases. This spurred on coordinated interventions, with WHO, in 2001, launching an initiative to reinforce HAT surveillance and control in all endemic countries.

From the year 2000, there has been a steady decrease in the prevalence and incidence of both *Tbg* and *Tbr* HAT [53–56] due to the combination of government and internationally assisted disease control measures. Despite these efforts, HAT is unfortunately far from being eradicated and remains a public health risk causing considerable morbidity and mortality [30,54,57] in regions where the parasite and its vector are present and in remote and rural areas where access to healthcare and control measures may be limited [58]. There have alarmingly been reports of new cases in ordinarily HAT-free areas [24,59,60]. While this may be due to previous underreporting [56,60] or vector distribution being affected by climate change [59], it is cause for concern.

Additionally, common HAT chemotherapies are plagued with adverse side effects, poor efficacy, and emerging drug resistance, in combination with complex treatment and administration protocols impractical in impoverished areas [61]. While effective against both stages of g-HAT and r-HAT [62], melarsoprol is restricted to second-stage r-HAT [63] due to toxicity and reported occurrences of resistance [64]. The drug contains arsenic that can accumulate in the body, causing a reactive encephalopathy fatal in 50% of cases [63] and damaging the liver and kidneys. Suramin, used for the first stage, is less toxic but requires complicated prolonged administration, while pentamidine is primarily used to treat stage one g-HAT as it is a less effective alternative for r-HAT. Eflornithine, effective in both stages of g-HAT [62,65], is also difficult to administer, coupled with a short drug half-life. The drug's shortcomings are improved by administering it in combination with nifurtimox, which is cytotoxic to both forms of HAT [63]. Fexinidazole is currently only administered for g-HAT, but clinical trials are underway to assess its alternative use for both stages of fast-progressing r-HAT (**Clinicaltrials.gov: NCT03974178**). Fexinidazole and acoziborole are considered improvements over previous treatments as both can be administered orally [61] and have less severe side effects. However, more research is needed to determine their long-term effectiveness and safety. A potential chemotherapy avenue to explore can be noted in targeting

the anti-folate drugs, which at present are in use for malaria treatment[66], and are of interest in other anti-parasite therapy such as , leishmaniasis[67,68] and Chagas disease (American trypanosomiasis)[69,70] .

Due to the continued status of HAT as an NTD and the lack of comprehensive and safe treatment for both forms of the disease, coordinated effort is still required for the elimination of *gambiense* and *rhodesiense* HAT as public health problems in central, and eastern southern Africa [71]. In the project, we aim to support elimination efforts by exploring the mechanism of r-HAT infection and potential anti-folate drug discovery as an alternative to existing chemotherapies.

1.2 Computational approaches

Various computational methods are used for the analysis of HAT to supplement and complement conventional wet-lab experimentation, as with many bioinformatics approaches.

1.2.1 Protein structure prediction

In silico protein structure prediction fills in the gap left by unsolved protein structures, calculating models comparable to what would be achieved experimentally and simultaneously providing an alternate solution where experimental techniques such as NMR analysis, cryo-EM or X-ray crystallography fall short. Calculated models can be used in numerous studies, including structure-based drug design [72,73], protein function analysis, variant analysis, and rational design of proteins for increased functionality.

Structure prediction can be achieved using either *ab initio* or template-based approaches. The template-based methods make use of alignment with experimentally solved protein structures, while the *ab initio* approach builds models from scratch using physics- or knowledge-based energy functions [74,75]. The template-based approach can further be divided into threading [76–78] and homology modelling, with the latter as the favoured approach [79], though the two can be used in tandem.

1.2.1.1 Homology modelling

Homology modelling is based on the established notion [80] of evolutionary-related proteins sharing highly conserved structural features, regardless of differences in primary amino acid sequence [81,82].

The process of homology modelling consists of four main steps:

1. Template selection
2. Target and template alignment
3. Model construction
4. Model assessment.

Template selection is the first, and thus the most crucial step in the process, as errors trickle downstream to the final step. Templates are typically identified based on sequence identity or similarity, where a match above 50% is ideal. Search tools, such as BLAST [83] and HHPred [84], use pairwise alignment to match homologous sequences and can be employed for template identification.

However, numerous studies [85–88] have revealed that protein structures can be generated with lower sequence identity. In such cases with weak sequence identity, secondary structure prediction plays a significant role [89,90]. Apart from sequence similarity, additional factors to be considered for template eligibility include phylogenetic similarity between template and target sequences, presence of ligands, as well as experimental structure resolution and R-factor quality [91].

Once a template is identified, it is aligned with the target sequences with consideration of gaps and insertions. Pair-wise alignment between the target and template or preferably multiple sequence alignment [92] including related sequences, are applied to generate the basis for model construction. Accuracy in target-template alignment is crucial to model quality[93], as a single residue misalignment can result in an error of approximately 4Å [94].

Model construction entails backbone generation, loop and side-chain modelling, and model optimisation [91]. Backbone generation provides the skeleton that the rest of the protein will be modelled onto, filling gaps and insertions during loop modelling. Loops are generally the most variable regions of a protein, often determining the functional specificity of a protein structure [95] . Protein side chains occur in energy conformations called rotamers, which are selected and fixed onto the backbone coordinates using defined energy functions and search strategies [96,97].

Proprietary standalone or webserver-based programs such as MODELLER [98], SwissModel[99], RaptorX[100] , and ROBETTA[101] are typically applied to for model generation, including a scoring functionality. Also worth mentioning is AlphaFold [102,103], an emerging tool combining

neural networks [104,105] and homology modelling to generate models. Specialised tools [106] such as DaReUS-Loop[107] , Loopy[108] and ModLoop[109] for loop modelling, or SCWRL and MolIDE[108] for side-chain prediction and modelling, exist as well.

The resultant structures from the model generation need to be optimised to resolve any conformational inaccuracies in bond lengths and angles through an energy minimisation procedure [110], as well as molecular dynamics (MD) to tackle steric clashes with minimal structural distortion [91]. Additionally, webservers can be employed for further structural validation. Examples of external servers include ProSA [111,112], ERRAT [113], PROCHECK [114,115], PROVE [116], WHATCHECK [117] and Verify3D [118].

1.2.1.2 Threading

Threading [119] , also known as fold recognition [120], is a protein structure prediction method that entails aligning the target amino acid sequence to a library of known protein structures [121,122] to identify the best match. The basic assumption of the threading algorithm is the conservation of the tertiary structure over secondary structure and sequence as in homology modelling [80]. It is particularly useful for identifying structurally similar folds between proteins with low sequence identity through profile alignment methods that consider both sequence and structural similarity, providing insights into protein function and evolution. Therefore, structural modelling can be better detected through 2D or 3D structural homology instead of sequence homology.

Threading involves steps similar to homology modelling except for the fold identification and fold fitting step in the template identification, where the protein sequence is divided into fragments threaded through 3D structural databases [123], such as the RCSB Protein Data Bank (PDB) (*https://www.rcsb.org/*). Some factors considered in determining structural similarity include predicted secondary structure and predicted solvent accessibility[120].

A scoring function[124] , typically based on a 2D matrix of inter-residue distance and solvation terms, is assigned to assess the fit of a sequence to a given fold, with the lowest energy fold being taken as the most suitable match. After the library of candidate structural fragments (templates) are identified, the target sequence is fitted onto the template scaffold using optimization algorithms, allowing for relative insertions and deletions in loop regions, adjusting side chain conformations, and refining the protein backbone.

As in homology modelling, the structures are refined and evaluated after model generation to improve accuracy and quality. Threading can typically be combined with other structure prediction methods to generate more accurate models[123].

1.2.1.3 Incorporation of SNPs/SNVs

Single point gene mutations can alter pre-mRNA splicing, amino acid sequence, and, ultimately, protein structure [125]. Mutations in nucleotide sequences are called single nucleotide variations (SNVs) and single nucleotide polymorphisms (SNPs) when observed in a population with a frequency of at least 1% [126–128]. At the protein level, non-synonymous SNPs (nsSNPs) alter the amino acid sequence of the resultant protein and can be further divided into missense and nonsense mutations which result in amino acid substitution and polypeptide sequence truncation through premature stop codons, respectively. As a result, missense mutations play a significant clinical role as they can inadvertently alter protein structure and function through the disruption of folding and stability, physiochemical properties, and modification structure flexibility.

Additionally, SNPs can play a role in the determination of the susceptibility or resistance of individuals to certain diseases [129]. Genome-wide association studies (GWAS) can be applied to analyse phenotype and mutation associations [130]. Due to this importance, these can be studied computationally to elucidate any protein function-altering properties by inclusion in protein model structures. Inclusion of mutations can easily be achieved by altering the target sequence as outlined above. Alternatively, tools such as PyMOL [131], Discovery Studio (DS) [132], and UCSF Chimera [133] offer functionality to introduce mutations to 3D protein structures, which can then be optimised, and further studies for the identification of their functional and structural effects

In the current work, protein structure prediction was applied using MODELLER [98] and Discovery Studio [132] to generate structurally sound models for the analysis of structural variants of ApoL1, SRA, and their complexes.

1.2.2 Molecular docking

Molecular interactions, such as protein-protein, enzyme-substrate, protein-nucleic acid, or drug-protein, play vital roles in essential biological processes. Naturally, there is a need to study these processes for varying fields of biological study. However, obtaining complex structures through experimental initiatives such as X-ray crystallography or NMR is often

difficult and expensive. As such, an alternative approach to understanding is required. Molecular docking aims to computationally model complex biological structures, showing interactions between the components at an atomic level and facilitating the characterisation of their behaviour and elucidation of fundamental biochemical processes.

A typical docking procedure entails:

1. definition and representation of the system
2. conformational search (for a complex)
3. scoring and ranking of possible complexes

These are achieved in varying ways depending on the participating molecules of the docking.

1.2.1.1 High-throughput virtual screening

High-throughput virtual screening (HTVS) is an increasingly valuable tool for drug discovery [73,134], functioning as a low-cost and effective [135] alternative to traditional experimental screening. HTVS aims to calculate a feasible protein (receptor) and compound (ligand) complex structure by performing molecular docking with multiple ligands [136].

1.2.1.1.1 Receptor and ligand preparation

System definition in protein-ligand docking constitutes pre-processing of both components. Receptor pre-processing involves adjustment of residue protonation states and tautomeric form dependent on functional pH, as well as general structure clean-up such as removal or inclusion of water molecules, cofactors and metals, and consideration of missing residues or atoms, according to the parameters available. Most important is the assignment of a potential binding site, usually obtained from literature on similar proteins or co-crystallized protein-ligand structures. Assigning a singularly known binding site results in targeted docking, while working without any assumption of the binding site is called blind docking. Blind docking is generally considered unbiased and closer to mimicking experimental screening [137] .

Preparation of the ligand entails assigning the correct atom types based on the ligand's appropriate ionisation states, chirality, and tautomeric states. As with the receptor, protonation states and tautomeric forms of a ligand influence ionic and hydrogen bonding abilities necessary for binding. Ligand filtering can also be included, assessing chemical property filters such as Lipinski rule of 5, pan assay interference (PAINS) assays [138–140] , and blood-brain

barrier (BBB) [141–145] to assist in discerning amongst non drug-like candidates. Ultimately, the amount of pre-processing required is chiefly dependent on the database source of ligands.

1.2.1.1.2 Docking calculation

The docking experiment, whether targeted or blind, predicts the ligand conformation, position and orientation within the binding site using assigned parameters and a predefined search algorithm [146]. Based on the "induced fit" theory [147], the receptor and ligand would ideally be treated as flexible during the calculation to accommodate conformational changes that may occur [148]. However, in the interest of computational and temporal efficiency and practicality [149], several docking programs define a flexible ligand and a rigid receptor. Docking programs such as AutoDock [150], FlexX [151] and MDock [152–154] adopt this asymmetric flexible-ligand (or semi-flexible) docking approach.

1.2.1.1.3 Scoring

The binding poses are subsequently assessed using a scoring function, applied in accordance with the force field with which atoms of the protein and the ligand were parameterised. In this study, a Lamarckian genetic algorithm (LGA) and an approximation of the AMBER force field specific to AutoDock were used [155]. A typical scoring function is developed through free-energy estimation techniques, as enthalpic and entropic effects play essential roles in ligand binding events [156]. Some estimation is employed as scoring functions aim to identify ligand binding mode and predict binding affinity within reasonable computation time. Co-crystallised protein-ligand structures [157], when available, can additionally be used to validate docking parameters through re-docking to recreate the co-crystallised pose. In addition to the aforementioned binding energy scoring, docking results can be further assessed via visual analysis of the ligand-protein interactions using software such as LigPlot+ [158,159], BIOIVIA DS [132], and Arpeggio [160].

1.2.1.2 Compound databases

Most compounds or fragments used in molecular docking are housed in compound libraries as 2D structures, molecular formulas, or SMILES strings that are parametrised for use. While generic repositories such as CheMBL [161] and ZINC [162] exist, most are compiled and categorised thematically around research areas, diseases (PfalDB [163]), geographical regions (SANCDB [164,165], ConMedNP [166], TCMSP [167]), or even compound properties and sources

(DRUGBANK [168]). The final selection of databases is based on the aims of the study and the quality of entries [169,170].

This work applies protein-ligand docking for HTVS of hit compounds against HAT drug targets *Tb*PTR1 and *Tb*DHFR (see ***Chapter 5***

Anti-folate Drug Discovery).

1.2.1.3 Protein-protein docking

As protein-ligand interactions are to drug discovery, knowledge of protein-protein interactions (PPI) is vital to understanding the molecular mechanisms in biochemical processes and cellular pathways. Multi-protein assemblies often execute complex biological processes [171] where individual proteins are assigned to specific functions and perform correlatedly [172]. As such, interactions between proteins are intrinsic to their function, and dysfunction in their inter-communication can cause or contribute to the pathogenicity of diseases. Understanding PPI is, therefore, a crucial step in biological system investigation and a possible starting point in rational drug design. To facilitate PPI study, protein-protein docking can be applied to generate protein complex models to enhance and complement classical structural biology techniques. The general procedure for protein-protein docking carries similarities to that for protein-ligand molecular docking. Some notable tools used for protein-protein docking include HADDOCK[173], ZDOCK[174], ClusPro[175], and HDOCK[176], each applying a unique algorithm to docking procedure.

1.2.1.3.1 Setting the system up

As in protein-ligand docking, one of the molecules is designated the receptor for ease of reference. A binding interface is defined for both proteins. Residues selected should be of central importance to the interaction, usually obtained from experimental procedures such as amino acid knockout, chemical shift perturbation (CSP) [177], or hydrogen−deuterium exchange mass spectrometry (HDX-MS) [178]. Throughout the simulation, these residues are restrained to being part of the interface.

1.2.1.3.2 Conformational search

As in protein-ligand docking, an appropriate degree of approximation [179] must be defined to address molecule flexibility. Both molecules are considered flexible, though to a limited or simplified extent, unlike the semi-flexible docking model where the receptor is regarded as

rigid. A two-stage docking process tends to be applied, first generating a large number of plausible candidates through systematically scanning the entire search space, then subjecting candidates to minimisation or an MD simulation using classical force fields [180,181] , such as OPLS used in HADDOCK[182], focusing on a smaller portion of the search space.

1.2.1.3.3 Complex scoring

Properties like accessible surface area (ASA) and buried surface area (BSA), interface residue conservation, hydrogen bonds, and electrostatic and hydrophobic interactions play significant roles in determining the nature of the protein interfaces. Knowledge of interface characteristics can be further used to score protein complexes.

1.2.3 Molecular dynamics

In order to obtain a molecular-level understanding of biochemical processes and systems, one needs to know the predominant conformations of the participating elements, their thermodynamic weights, and transition rates between these states. This information can be obtained from ensembles produced by molecular dynamics simulations [183].

MD simulation computationally imitates the natural motions of atoms and molecules in a molecular system in a time-dependent manner [184], which can be used to supplement and complement conventional wet-lab experimentation, as with many bioinformatics approaches. Due to the ability to manipulate numerous aspects of a simulation, it is possible to capture and analyse data in a way that would be impossible to do in wet-lab experimentation. Additionally, computer-simulated analysis is faster and less expensive than the alternative. Some biological applications of MD simulations include molecular docking and drug design [185], refinement of structure prediction [186], protein motion and functional conformation analysis [187,188] , and protein folding analysis [189].

Simulation typically entails the determination of molecular and atomic trajectories through integration to solve Newton's classical laws of motion for a system of interacting particles. Inter-particle forces and potential energy for these calculations are defined by molecular mechanics force fields [190], all run through simulation engines such as AMBER [191], CHARMM [192], GROMACS [193], and NAMD [194].

1.2.1.4 Running an MD simulation

The basic properties needed to run an MD simulation system are the topological describing connectivity of atoms, the structural providing atomic position and conformation, the energetic defining forces acting on the system, and the thermodynamic to assign experimental conditions. The application of these properties occurs in four general steps:

1. Setting up the system. The biological system is defined from either experimental structures or homology modelling data, typically as an atomistic or coarse-grained reproduction of experimental systems, from which the topology is derived. The system is contained in a simulation box defining boundaries and size. In order to combat surface interactions with the boundary and approximate an infinite system, periodic boundary conditions (PBC) are typically preferred.

   The standard procedure entails parametrisation of the biological system based on a force field of choice, definition of size of the entire system, resolving structural errors, ionisation of titratable amino acids, addition of structural water molecules, counter-ions and solvents, and energy minimisation. As many biological processes occur in an aqueous solution, solvation of the system is crucial to determining molecular conformation, electrostatic properties and binding energies [195]. Solvent representation is therefore essential to this end, with the most effective being the explicit molecule representation [196,197], whose advantage lies in the ability of the solvent molecules to maintain most of the entropic solvation effects of natural solvents.

   The system must then be neutralised to avoid polarisation or set to a desired pH at a defined salt concentration. An energy minimisation run then serves to combat steric clashes by computing an equilibrium configuration of all molecules now in the system. As atoms are reoriented mathematically from a non-equilibrium state, the lowest energy configuration is achieved, and ideally, the global minimum.

2. Equilibration. To prevent a collapse of the system post-minimisation, equilibration of the solvent and ions around the protein is conducted at desired temperatures, pressure, and density [198]. Integration of Newtonian equations of motion allows for a constant energy surface of the system that excludes temperature and pressure.

   Thermodynamic ensembles are therefore used for system equilibration, where the correct ensemble distribution for specified temperature and pressure allows

interpretation of the trajectory conventionally. The common ensembles in this regard are the constant-energy constant-volume, microcanonical ensemble (NVE); the constant-temperature constant-volume, canonical ensemble (NVT); and the constant-temperature constant-pressure, isothermal-isobaric ensemble (NPT). The NVE solves Newtonian equations without temperature and pressure control, thus conserving energy in the system [199]. The NVT ensemble, on the other hand, is obtained by controlling the temperature with the maintenance of constant volume throughout the run. The amount of substance (N), volume (V) and temperature (T) are conserved, while the energy of any endothermic and exothermic processes is regulated via a thermostat. A temperature plateau is reached at the desired value signifying that the temperature has stabilised. Finally, the NPT ensemble allows control over both the temperature and pressure, where the pressure adjustment corresponds to the volume. The NPT approach tends to be used to achieve the equilibrium density corresponding to the desired pressure and temperature. The amount of substance (N), pressure (P) and temperature (T) are conserved, and a thermostat [200,201] and a barostat [202] are needed. While choice in the thermostat and barostat is not critical, an equilibrated state is crucial.

3. Production. In order to begin production, position restraints can be released once equilibrium is achieved at the desired pressure and temperature. Ideally, the simulation time should be sufficient to allow the macromolecule to explore all possible configurations.

4. Assessment. The simulated trajectory must be analysed to collect data and extract the desired properties. Assessment of protein motions can be conducted at a global and local level based on atomic positions, velocities and even forces as a function of time.

This study used GROMACS [193,203] on the Lengau cluster of the Centre for High-Performance Computing (CHPC) in Cape Town, South Africa. The force fields to parametrise the system in each investigation were elected on a case-by-case basis. Additionally, the steepest descent algorithm was employed for the minimisation procedure, while the Particle Mesh Ewald (PME) method [204,205] and Linear Constraint Solver (LINCS) [206] were applied for handling electrostatic interactions and bond length constraints, respectively, during production runs. The short-range non-bonded interactions were defined based on the Verlet algorithm [207], using a cut-off of 1.4nm.

1.2.1.5 Force fields

Force fields are equations used to deduce the potential energy of a molecular structure, applied in the topology building segment of an MD simulation and maintained throughout the run. Typically, force fields are split into bonded and non-bonded interactions, as shown in ***Equation 1*** [208] .

$$U\vec{r} = \sum U_{bonded}(\vec{r}) + \sum U_{non-bonded}(\vec{r})$$

where:

$$U_{bonded}(\vec{r}) = U_{bond}(\vec{r}) + U_{angle}(\vec{r}) + U_{dihedral}(\vec{r})$$

$$U_{non-bonded}(\vec{r}) = U_{vdW}(\vec{r}) + U_{elec}(\vec{r})$$

Equation 1: Simplified Force Field Calculation

The representation of molecular features in force fields comprises springs for bond length and angles; periodic functions for bond rotations and Lennard-Jones potentials [209]; and Coulomb's law for vdW and electrostatic interactions, all of which assure that energy and force calculations are relatively fast. Force fields can be defined as either united-atom or all-atom, with the former joining several atoms, such as methyl groups, into a single interaction site and the latter utilising explicit representation of aliphatic hydrogen atoms. A similar molecular representation philosophy [210] is combined with bespoke empirical and quantum mechanical calculations to generate parameters for the varying force fields currently used in MD simulations. Commonly used families of force fields include AMBER [211], CHARMM [212], GROMOS [213], and OLPS [214], where the resulting simulations remain comparable [215,216] despite the consequently unique parameters and atom/molecule types. Choice in force field is determined on a case-by-case basis [217,218].

However, due to the vast number of possible atom types, numerous compounds and co-factors are not catered for in existing force fields. Non-protein topologies and parameters tend to be derived externally based on the force field of choice, using quantum mechanical (for OPLS, AMBER and CHARMM) and semi-empirical (for GROMOS) calculations. These parameters can either be fitted [219–221] or generated through automated tools such as ACPYPE [222], PRODRG [223], MKTOP [224] and Automated Topology Builder (ATB) [225].

1.2.4 Analyses of computational methods

Most computational analysis is conducted post-MD simulations as this gives the most informative run down. Analyses are conducted on either static or dynamic structures. In this study, static calculations were conducted on the low energy conformations of structures, while dynamics calculations were conducted on converged sections of MD trajectories. The extent of post-MD analysis naturally depends on the aim of the investigation and the biological system of interest, with the most robust or informative being elected per study.

1.2.1.6 Global analyses

Global protein analysis typically refers to the large-scale computational analysis of protein structures. This involves predicting the motions and conformational changes of proteins and intermolecular interaction over time or examining the overall structure properties such as folding and stability. As this analysis approach aims to garner structural information about the protein as a whole, global analyses can help to identify critical regions involved in specific functionalities such as ligand binding or allosteric regulation.

1.2.1.6.1 Root mean square deviation and Radius of gyration

Analysis of protein MD trajectories typically relies on classical measures, such as the root mean square deviation (RMSD) and radius of gyration (Rg), conceived and developed for the holistic analysis of single macromolecules.

RMSD measures the averaged distance deviation of the atom coordinates in a protein structure from those of a reference structure. The comparative analysis can be applied to investigate the similarity of protein structures during homology modelling [196], and conformational changes [226] or molecule stability in molecular docking. In its application to protein MD trajectories, RMSD is typically calculated by fitting Cα atoms or protein backbones over the simulation frames, with the MD starting point structure as the reference as depicted in ***Equation 2***:

$$RMSD = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i \left(r_i(t) - r_i^{ref}\right)^2}$$

Equation 2: Generalised RMSD calculation

where M is total mass, m_i and $r_i(t)$ is the position of atom i at time t after fitting the overall structure with respect to the reference r^{ref}.

Rg in ***Equation 3*** depicts protein compactness [227] and measures change from initial to final structure as in RMSD. Rg is calculated with respect to the centre of mass (CoM) of the molecule:

$$Rg = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i\,(r_i - R)^2}$$

Equation 3: Generalised Rg calculation

where m_i and r_i are the mass and position of atom i, respectively, and M is the total mass of the molecule consisting of N atoms with the centre of mass R.

The molecular spatial packing of amino acid residues is an essential aspect of protein stability [228,229]; as such, changes in Rg can suggest conformational deterioration.

RMSD and Rg values are typically presented in line or violin plots where conformational change, convergence or stability can be deduced.

1.2.1.6.2 Essential dynamics analysis

Proteins exist as an ensemble of conformational substates aiming to achieve dynamic equilibrium [230] across an energy landscape. In conceptualising protein dynamics [231] as a conformational energy landscape [232] capturing both discrete minima states separated by energy barriers to traverse [233], it is crucial to note the equilibrium fluctuations and non-equilibrium effects in transition phases between the initial and final states. Protein motions observed in their stable equilibrium conformations seem to govern function in biological processes [232,234,235].

MD simulations aim to explore the multidimensional energy landscape covering the amplitude or directionality of functional protein motions on a breadth of timescales [232]. Applying conformational substates to analyse MD trajectories allows for a physically reasonable interpretation of the runs and extraction of meaningful information from these large datasets.

As methods such as RMSD are inadequate in identifying molecular equilibrium or convergence in simulations [236], principal component analysis (PCA) can be applied for investigation. PCA essentially converts correlated atom movements to a set of linearly independent coordinates (principal components (PC)). Mathematically, PCA analysis transforms movement data to the coordinates, where the variance abates with each sequential coordinate. Ultimately, in the distillation of the dataset, the PCs preserve most of the differences between the data.

In protein analysis, PCA isolates the statistically meaningful conformations sampled during the trajectory. As such, it can be used to reveal the essential protein dynamics filtering observed motions from the greatest to minor spatial scales by applying the decomposition process.

In this work, PCA was applied through the essential dynamics (ED) tool of MDM-TASK-web [237] to track the conformational changes of the proteins over the MD simulation run.

In the ED tool, variance in simulation data analysed over the trajectory length is presented as a plot, where each point on the plot represents the conformation of the protein at specific time frames during the simulation.

For comparison, the trajectories were internally aligned before a covariance matrix based on the protein Cα and Cβ atoms and diagonalised [237]. The eigenvectors and eigenvalues solved from the matrix were projected onto protein coordinates and represented the directions of functional motions, respectively. As a result, the magnitude of the motions was captured in the corresponding eigenvalues. Additionally, potential equilibrium low energy states could be identified using *k*-means to sample centroid conformations [237].

1.2.1.6.3 Visual analyses

Snapshots of MD simulation at specific time frames can be visualised using software such as PyMOL [131], LigPlot+[159], DS [132], and Schrodinger Maestro [238]. These snapshots capture intermolecular interactions within the system. Additionally, all simulation frames can be visualised as video in software such as Visual Molecular Dynamics (VMD) [239] without the PBC box.

1.2.1.7 Local analyses

In contrast to global analyses, local protein analysis focuses on the dynamics and properties of specific regions or features of a protein, such as individual amino acid residues, protein

domains, or binding sites. At a residue level, the coalition of individual properties tends to influence the overall functional properties of the protein.

Both global and local protein analyses are crucial to protein study and function complementarily. Global analysis is typically useful for understanding the overall protein architecture and intermolecular interaction, while local analysis is useful for the detailed studying of intramolecular mechanisms underlying protein function.

1.2.1.7.1 Root mean square fluctuation

Root mean square fluctuation (RMSF) measures the local average deviation in the flexibility of residues over the course of a simulation. RMSF is typically calculated for the Cα or backbone atoms based on residue contribution. Where the RMSD measures structural divergence from a reference (over time), the RMSF can reveal areas of the system contributing to mobility, highlighting regions of flexibility which may be overlooked in the global analysis.

RMSF is averaged over time as depicted in ***Equation 4***, giving a value for each particle *i*, while the averaged RMSD is taken over the particles, giving time stamps.

$$RMSF = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(r_i(t) - r_i^{ref}\right)^2}$$

Equation 4: Generalised RMSF calculation

where T is the averaged time over the simulation, and r_i^{ref} is the reference position of particle i.

1.2.1.7.2 Dynamic Residue Networks

Dynamic residue network analysis (DRN) uses the concept of graph theory to explain the dynamics of residue contacts during molecular dynamics simulations [240]. In the case of proteins, nodes refer to individual residues and their connections. Centrality depicts and measures the importance of an element (node) in a network. The relationships between residues are quantified based on how specific centrality metrics are defined [241,242], the outcome of which can be used to define communication within the protein and identify key residues. In this work, we applied network analysis as defined by [237,240]:

1. *Betweenness centrality* (*BC*) is measured by the sum of the shortest paths between residues [188] that pass through the residue of interest. As a result, residues with large

betweenness values function as bridges and gatekeepers for communication between points in the network.

Averaged *BC* is calculated in *Equation 5* [243,244] :

$$\overline{BC}(v) = \frac{1}{m}\Sigma_{i=1}^{m}\Sigma_{s,t\epsilon v}\frac{\delta(s_i, t_i|v_i)}{\delta(s_i, t_i)}$$

Equation 5: Averaged betweenness centrality calculations

where *v* is the given node between *s* and *t* (amino acid residues), and *(s,t)* symbolises the number of averaged shortest paths. The number of said shortest paths going through node *v* other than *s, t* is indicated by *(s,t|v)*, at the time *i* over averaged total *m* frames.

2. *Closeness centrality* (*CC*) identifies the residues closest to the others in the network based on the length of the path. As closeness scores the residues on their distance and ability to reach others, residues with a high closeness centrality have the shortest paths to all other residues and are best positioned to influence communication in the rest of the network. The metric is calculated using *Equation 6* :

$$\overline{CC}(v) = \frac{n-1}{m}\Sigma_{i=1}^{m}\Sigma_{u=1}^{n-1}\frac{1}{d_i(v,u)}$$

Equation 6 Averaged closeness centrality calculations:

Averaged *CC* calculates the reciprocal distance of the shortest path of residue *v* to all other residues *u* as denoted in *Equation 6* [244,245]. It is normalised by the overall sum of minimum possible distance *n*-1 [237], where d_i *(v, u)* is the shortest path distance between residues *v* and *u*, and *n* is the number of residues in the network.

3. *Degree centrality (DC)* measures the number of paths a residue possesses to others, thus identifying residues with the highest connections.

$$\overline{DC}(i) = \frac{1}{m(n-1)}\Sigma_{k=1}^{m}\Sigma_{j=1,j\neq i}^{n}A_{ijk}$$

Equation 7: Averaged degree centrality calculations

DC defines immediate local connectivity of *k*, normalised by the highest possible degree in the network (*n*-1). An adjacency matrix is generated where *Aijk* in *Equation 7* [244] is a 3D tensor consisting of adjacency matrices in which *j* and *k* denote the residue indices. A central residue, therefore, presents high *DC* as it has a high local connectivity [237].

4. *Eigenvector centrality* (*EC)* measures the influence of residues by incorporating the sum of all connections to the residue of interest and those of the corresponding residues. The centrality of neighbouring residues adds to the initial residue, compounding the influence and thus ranking the importance of the neighbours [246].

$$EC_{ik} = \frac{1}{\lambda_k} \Sigma_{t=1}^{n} A_{ijk} \cdot EC_{jk}$$

Equation 8: Averaged eigencentrality calculations (i)

The eigencentrality calculation is an extension of degree centrality, assigning residue importance by solving the eigenvector *EC* with the adjacency matrix *A* to convergence. *Equation 8*, where *Aijk* is an adjacency of the frame, *k* is the eigenvalue, and *n* is the number of nodes, can be rewritten to *Equation 9* [244,247,248]. Averaged *EC* for *i*-th residue is then calculated using the time averaged from *m* frames in *Equation 9ii* [237] .

$$A \cdot \overrightarrow{EC} = \lambda \cdot \overrightarrow{EC}$$

$$\overline{EC}(i) = \frac{1}{m} \Sigma_{k=1}^{m} EC_{ik}$$

Equation 9: Averaged eigencentrality calculations (ii)

5. *Katz centrality* (*KC*) is similar to eigencentrality [246] in that it calculates the relative influence of a residue within the network by measuring the number of direct connections, and all other residues in the network connect to the residue of interest through these direct neighbours. However, the paths are weighted in *KC*, and connections made with distant nodes are penalised.

$$KC(i) = \alpha \Sigma_{j=1}^{n} A_{ij} KC_j + \beta$$

$$\overline{KC}(i) = \frac{i}{m} \Sigma_{k=1}^{m} KC_{ik}$$

Equation 10: Averaged Katz centrality calculations

KC is a permutation of *EC* where α and β in *Equation 10* [244,247] are constants denoting a damping coefficient and a basal adjacency assigned to the connected neighbours of residue *i*. Resulting residue centrality can be attenuated to desired extents, resulting in *KC* and *EC* converging [237].

A collection of MDM-TASK-web [237] and MD-TASK suite [240] scripts (GitHub: *https://github.com/RUBi-ZA/MODE-TASK* & *https://github.com/RUBi-ZA/MD-TASK/tree/mdm-task-web*) were implemented in conducting DRN analysis on MD trajectories in this research. The networks are constructed using Cβ atoms (Cα for Gly) regarded as nodes and an edge is created when two nodes are within a certain distance from each other. The cut-off distance between two nodes was specified as 6.7Å [188,249]. To identify nodes important to communication in the residue network, *cal_network.py* (*GitHub - RUBi-ZA/MD-TASK at mdm-task-web*) was used to calculate the centrality metrics *BC*, *CC*, *DC*, *EC*, and *KC*.

1.2.1.7.3 Bond analyses

Monitoring intra- and inter- molecular bonds over the course of MD trajectories provides further insight into the folding or stability of biological systems. Disulphide bridges and hydrogen bonds are typically monitored based on the bond distance. Disulphide bonds play a crucial role in proteins, with the distance and angle between the participating cysteine residues modulating protein stability and conformational dynamics [250–253]. In addition to the maintenance of protein secondary structure elements [254–256], hydrogen bonding is also vital to protein-ligand binding, [257–259] contributing to ligand stability, binding affinity [260], and site recognition [261].

As bond strength is dependent on the bond length and angle [262–264], those are the attributes monitored in protein structural analysis in tools such as VMD [239], AmberTools21 [265] and GROMACS [193,203]. Visual analysis of trajectory snapshots complements bond analysis.

1.3 Aims and motivations

HAT remains a lethal NTD responsible for enormous suffering in endemic areas in sub-Saharan Africa. While a decline in cases has been noted, coordinated effort is still required for the elimination of *gambiense* and *rhodesiense* HAT as public health problems in central, eastern, and southern Africa [71]. In the project, we aimed to support elimination efforts through:

1. Investigating the pathogenicity of HAT by studying the proteins that participate in the parasite's resistance to human serum.

2. Improving existing chemotherapies by identifying alternative drug hits that can be developed to replace current antiquated drugs.

To achieve aim 1, the following steps were followed:

- Modelling variant structures of ApoL1 C-terminal and *Tbr* SRA
- Establishing a dynamically sound SRA-ApoL1 dimeric complex
- Elucidation of communication between the two proteins
- Investigation of the effect of the structural variations on the protein-protein interaction.

To achieve aim 2, the following steps were followed:

- Identifying anti-folate hit compounds viable in both PTR1 and DHFR
- Elucidation of communication between the two proteins
- Investigation of the effect of the structural variations on the protein-protein interaction

1.4 Research Objectives

The research objectives of the project were addressed in two workflows, highlighted in ***Figure 2*:**

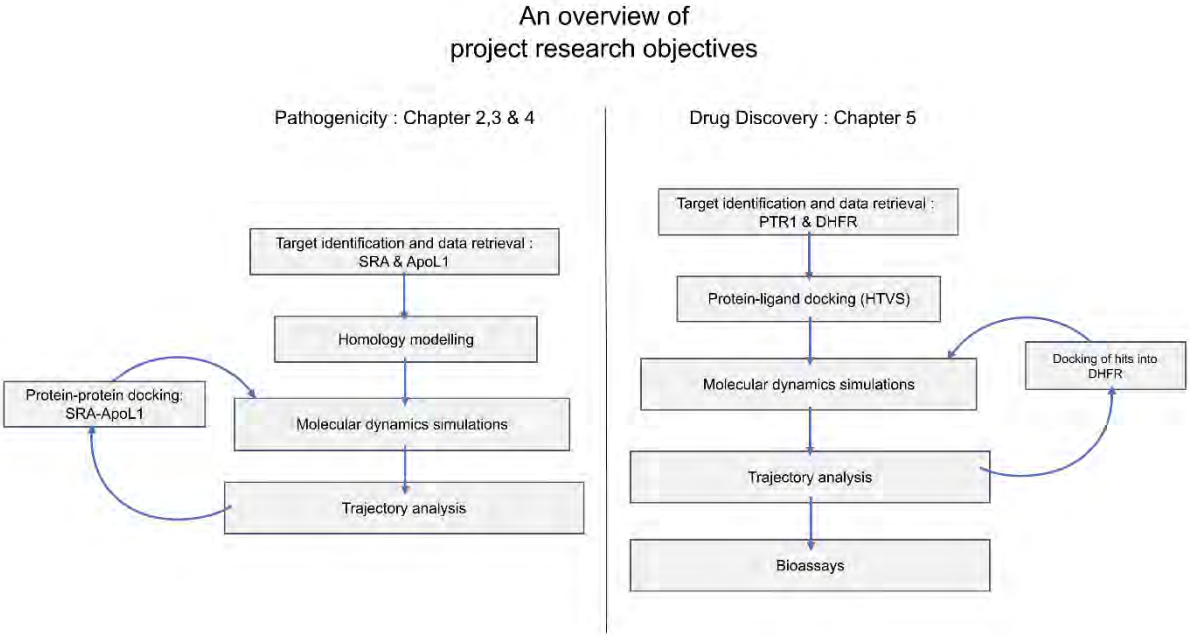


Figure 2: A schematic illustration of the methodological objectives of the overall project for research aims 1 and 2

Aim 1:

1. Homology modelling and calculation of ApoL1 C-terminal and SRA variants
2. Protein-protein docking of ApoL1 and SRA models
3. All-atom molecular dynamics simulations of the best-ranked SRA-ApoL1 C-terminal dimeric complexes
4. Trajectory analysis:
   - Stability analysis of the molecular dynamics trajectories
   - Dynamic residue network analysis
   - Conformational analysis: Essential dynamics and Normal mode analysis
   - Intermolecular interaction analysis and alanine scanning
   - Energy calculations

Aim 2:

1. Homology refinement of *Tb* PTR1 and DHFR
2. Preparation of ligand dataset
3. HTVS using protein-ligand molecular docking
4. All-atom molecular dynamics simulations of the top-scoring PTR1/DHFR-ligand complexes
5. Trajectory analysis:
    - Stability analysis of the molecular dynamics trajectories
    - Binding energy calculations
    - Intermolecular interaction analysis
6. Bioassays

Chapter 2

Variations in ApoL1

2.1 Introduction

Human African trypanosomiasis infections are only caused by *Trypanosoma brucei gambiense (Tbg)* and *Trypanosoma brucei rhodesiense (Tbr)* [266] due to their resistance to the cytotoxic action of human serum [267,268]. The innate immunity [266] resulting from parasite lysis is facilitated by trypanosome lytic factors (TFL) [269–271], where TFL1 consists of apolipoprotein L1 (ApoL1) and haptoglobin-related protein (Hpr) [272], with ApoL1 as the main lytic component.

ApoL1 is an HDL-bound lipoprotein, existing in both a membrane-bound and localised form that imparts innate immunity to the *Tbb* parasite. The lipoprotein has a five-helix pore-forming domain and a pH-sensitive membrane-addressing domain consisting of two helices [273,274], which depolarise the parasite lysosomal membrane, leading to osmotic swelling and rupture of the lysosome, and lysis of the trypanosome [273,275–278].

At full length, the human lytic factor ApoL1 is 398 residues long [279] with three main domains [273,280] the pore-forming, pH-sensitive membrane-addressing, and SRA interacting (SRA-ID) domains. The C-terminal of ApoL1 (R305-L398) consists of an α-helical hairpin with a leucine zipper domain (LZD) housed in the SRA-ID [280–283]. The LZD, a series of leucine heptad repeats from L371-L392, is suggested by [284] to play a significant role in possible SRA-ApoL1 coiled-coil formation.

Two non-synonymous coding variants, G1 and G2 [285], with high frequency in individuals of African ancestry [286–290] occur in the ApoL1 C-terminal [291,292]. The G1 (G1GM) mutation consists of missense substitutions at positions S342G and I384M occurring in nearly absolute linkage disequilibrium [287,291], while G2 (del388N389Y) is an in-frame deletion of residues Asn388 and Tyr389, in the same functional domain as G1, resulting in a truncation of the protein in the LZD. Due to the variants' location in the ApoL1 C-terminal helix cited for the sole responsibility of SRA binding [280,291], their presence disrupts the SRA-ApoL1 interaction,

restoring protection against *Tbr* [281,283,291,293,294]. The truncation of the C-terminal has particularly been shown to affect SRA-ApoL1 binding [295] thus restoring ApoL1 activity. Unfortunately, the presence of both alleles results in a "gain of function" [296] scenario by strongly increasing the risk of chronic kidney disease (CKD) in carriers, garnering increased interest in the protein.

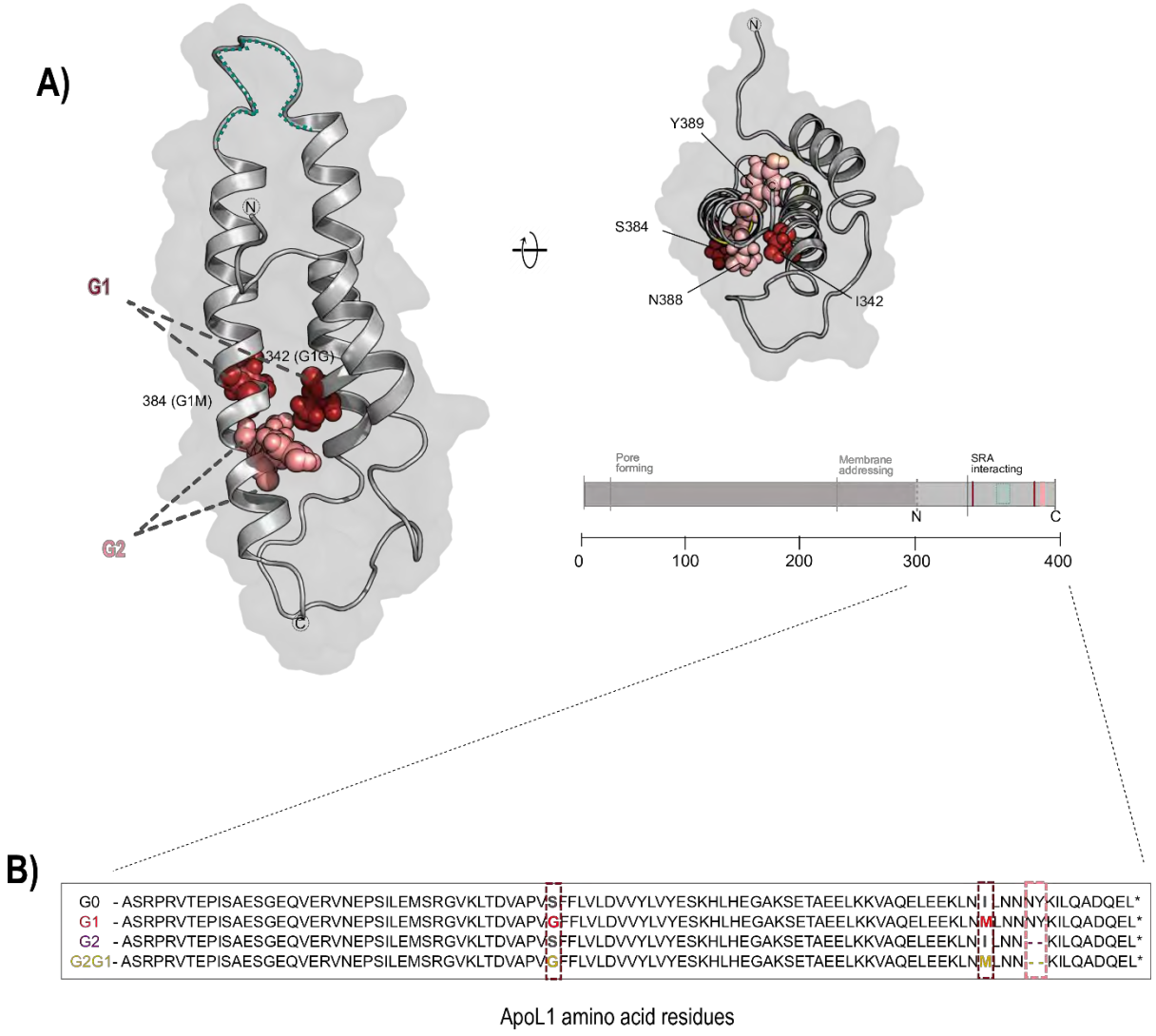


Figure 3: ***A)*** *The C-terminal of ApoL1, showing the SRA-binding site and* ***B)*** *the location of the G1 (S342 and I384) and G2 (N388/Y389) mutations in firebrick and salmon, respectively. (Structure source : Homology model)*

A third ApoL1 variant characterised by a homozygous missense substitution (N264K) in the membrane-addressing domain has been cited to affect the lytic activity of ApoL1 against *Tb. rhodesiense*. The variation increases the risk of contracting HAT when accompanied by the C-terminal SRA-interacting domain substitutions [297]. This observation highlights the possible

participation of ApoL1 regions beyond the C-terminal domain [298] in the human-trypanosome infection interplay.

Despite great interest, the full-length ApoL1 experimental structure has yet to be reported as of the compilation of this thesis. The protein and its multiple coiled-coil domains have proven unstable in concentrated solution for crystallography[284] . At the same time, NMR spectroscopy has a protein size limit, [299,300] and cryogenic electron microscopy (cryo-EM) lacks information on intramolecular interactions or protein dynamics due to high output resolutions [301,302].

Previous studies have used homology modelling [283] supported by NMR spectra [284], threading [303], and *ab initio* [295,304] to determine the ApoL1 C-terminal structure. While these have resulted in differing structures, all have supported the α-helical characterisation of the C-terminal. The full length of the protein has further been modelled by Kumar *et al.*, [305] using threading and separately [304] in AlphaFold [102]. The ApoL1 N-terminal has recently become the only portion to be the crystallographically solved [306].

In this chapter, we aim to investigate the protein dynamics of ApoL1 C-terminal through modelling of the wild-type (WT), G0, and structural variants G1, G1G/M and G2 of the protein. The effects of the mutations on the protein will provide further insight into their role in the SRA-ApoL1 interaction.

2.2 Methods

2.2.1 Homology modelling

In the absence of an experimentally solved full-length structure of ApoL1, other research groups have calculated *in silico* predicted models of the whole protein and fragments to varying degrees of success. The latest model worth noting at the time of thesis compilation was an AlphaFold predicted structure of the full-length protein [304] . Due to the limitations of AlphaFold in transmembrane protein prediction [307–309], the resultant structure comprises disordered loops (*https://alphafold.ebi.ac.uk/entry/A0A1B2FJP1*), contrary to secondary structure predictions of the protein using threading software such as ROBETTA (see ***Figure S 1***). While AlphaFold2[310] has been rolled out to improve membrane-associated protein prediction, work still needs to be done to compensate the lack of transmembrane training data [311] required for membrane predictions[307,310] .

Sharma *et al.* [312] , conducted a study focusing on a shorter 2-helix fragment (339–398) of the protein as predicted in studies prior to it[281,313] , which was not suitable to our study aiming to study a larger segment of the C-terminal. Additionally, contrasting results were observed between this and a similarly timed study by Madhavan *et al.* [312] using a longer 3-helix construct (305–398) cited to be more stable. This second study determined the C-terminal structure using threading and *ab initio* approaches, later enhancing the work with NMR analysis. However,[295] a considerable divergence was noted between the NMR and threading-generated models. This is cause for caution regarding the existing models.

Considering these observations, we aimed to generate our models of the ApoL1 C-terminal to corroborate or provide a resolution to the current knowledge on the computationally determined protein structure.

Modelling of ApoL1 G0 C-terminal (301-398) was achieved through a combination of both homology modelling and threading. Secondary structure prediction from PsiPred [314,315], HHPred's Quick2D [84,316] and ROBETTA [317] enabled suitable templates to be identified (**see *Figure S 1***). The crystal structures for Apolipoprotein E3 (APOE3) (PDB ID:1NFN [318] and 6V7M [319]) and the structural motif of *Tb* VSG (PDB ID: 2VSG) were used as templates, with 2VSG [283] functioning as a secondary template for the leucine zipper domain (LZD). MODELLER v9.20 [320] was then employed for model construction using the slow refinement option of the *automodel* class to build 100 models that were ranked by z-DOPE score. The lowest z-DOPE models were consequently structurally validated using ProSA and the SAVESv6.0 web server housing ERRAT [113], PROCHECK [114,115], PROVE [116], WHATCHECK [117] and Verify3D [118].

The models were additionally subjected to energy minimisation and 20ns MD production runs (as described in 1.2.1 and the methodology below) to curtail steric clashes. The five ApoL1 variants were modelled by incorporating the respective variations G1 (S342G/I384M), G1G (S342G), G1M (I384M), G2 (del388N/389Y) and G2G1 (S342G/I384M/del388N/del389Y) into the target sequences for each modelling run.

2.2.2 Molecular dynamics simulations

The CHARMM36 [321] force field was applied to parameterise the ApoL1 C-terminal wild-type and variant proteins in the all-atom molecular dynamic simulation runs. The protein systems were solvated using the TIP3P water model in a dodecahedron box of 1nm cut-off and

neutralised to a concentration of 0.15Mm using Na^+ and Cl^- ions. The subsequent minimisation to achieve a conformational equilibrium for the molecules was set at a F_{max} <1000.00kJmol^{-1}nm^{-1} with a maximum of 50000 steps. Equilibration applied the NVT and NPT ensembles to bring the systems to 300K temperature, using modified Berendsen temperature coupling and 1 bar of pressure using the Parrinello-Rahman barostat, respectively. The PME method [204,205] was used for long-range electrostatic interactions, and the LINCS [206] for bond length constraints. Short-range non-bonded interactions were defined based on the Verlet algorithm [207], using a cut-off of 1.4nm. Finally, the production runs of 400ns were conducted using 240CPU cores, using a step size of 2fs for data collection.

2.2.3 Classical trajectory analysis

The GROMACS *gmx trjconv* module was used to correct for the PBC of trajectories and fit the Cα atoms of the starting structures to eliminate any rotational and translation artefacts from the entire system before analyses. RMSD, Rg, and RMSF were calculated by *gmx* modules *rms, gyrate* and *rmsf* to investigate protein stability and flexibility. RStudio *ggplot* and *R-base* were used to plot these properties.

2.2.4 Principal Component Analysis

Comparative ED was calculated on the 400ns trajectories using the *compare_essential.py* script (*https://github.com/RUBi-ZA/MODE-TASK*) of the MODE-TASK functionality of MDM-TASK at a step size of 1 frame. The comparative calculations were conducted in two sets due to the two residue deletion in the G2 and G2G1 variants. To maintain corresponding Cα and Cβ, the PCA calculations for the 96 residue variants (G2 and G2G1) were run separately from the 98 residue long proteins (G0, G1, G1G, and G1M).

2.2.5 Dynamic Residue Network analysis

DRN calculations were conducted on the last 350ns of each MD trajectory. The *calc_network.py* (*https://github.com/RUBi-ZA/MD-TASK/tree/mdm-task-web*) script from the MDM-TASK suite calculated *BC*, CC, *DC* and *KC* using a step size of 50 frames and a node cut-off of 6.7Å. The top 5% of the centrality nodes (residues) per protein was then extracted from each calculated metric to shortlist residues potentially contributing highly to communication in each ApoL1 variant.

2.2.6 Contact map analysis

Additionally, weighted residue contact maps and subsequent heatmaps were calculated through the *contact_map.py* and the *contact_heatmap.py* scripts (found at *https://github.com/RUBi-ZA/MD-TASK/tree/mdm-task-web*) [237] using the same parameters as the DRN calculations. Contact map analysis was conducted on the sites of variation to assist in the elucidation of their effects on a residue-residue level.

2.3 Results and Discussion

The global dynamics of the modelled ApoL1 C-terminal structures were investigated using RMSD, Rg and ED calculations, while DRN and contact map analysis were applied for residue level analysis.

2.3.1 Homology modelling

We modelled the WT and five variants of the ApoL1 proteins, as explained in the methodology section (see 0***2.2.1*** Homology modelling). These models were first ranked according to their

z-DOPE score (see ***Table 1***).

Table 1: Homology model evaluation and validation scoring from varying programs. Model evaluation was conducted on the top Z-DOPE scoring structures. The Z-DOPE scores were calculated i) post model calculation and ii) post minimisation and a 20ns MD production run

ApoL1 model	***z-DOPE[ai] score***	***z-DOPE[aii] score***	***ERRAT score[b] (%)***	***ProSA Z-score[c]***	***VERIFY3D[d] (%)***
G0	-0.657	-1.182	98.89	-4.14	53.06
G1	-0.661	-1.092	98.87	-3.86	45.92
G1G	-0.606	-1.392	98.85	-3.34	44.69
G1M	-0.563	-1.230	98.68	-3.72	43.88
G2	-0.667	-0.960	97.88	-3.78	17.71
G2G1	-0.546	-1.171	96.69	-4.07	36.58

a. MODELLER Z-DOPE score calculated (i) before and (ii) after a 20ns MD production run.

b. The ERRAT score measures the overall structural quality factor [113] .

c. PROSA Z-score, accompanied by the plot in Figures 4A

d. The VERIFY 3D score - at least 80% of residues should score >= 0.2 in the 3D/1D profile to PASS [118]

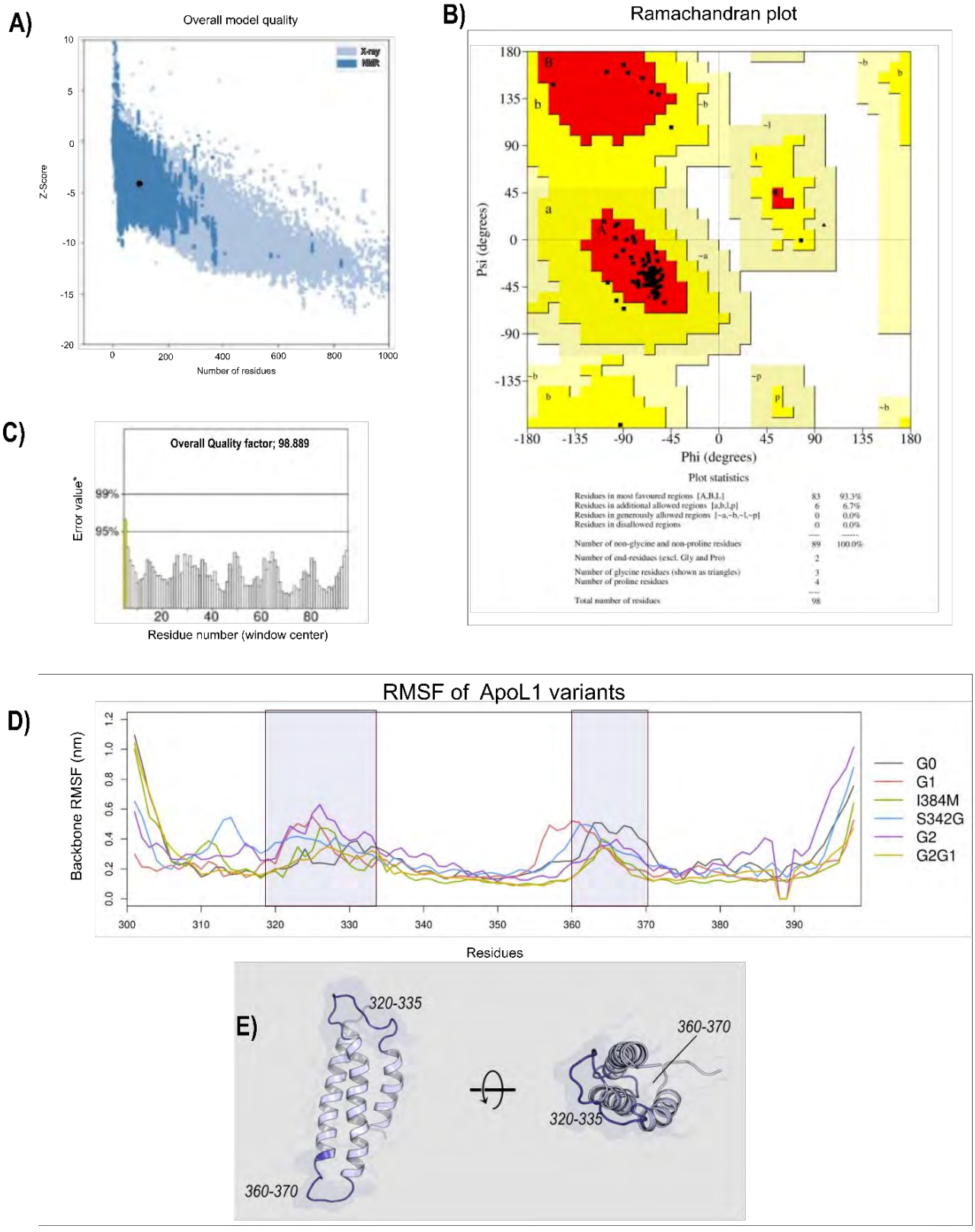


*Figure 4: Model validation of the ApoL1 C-terminal structures from **A)** ProSA, **B)** PROCHECK and **C)** ERRAT servers. The RMSF of the ApoL1 variants post MD simulations in **D** reveal the flexible loop regions of each variant that are highlighted on the structure in **E**. The significant number of loops in the structures contributes to the validation scoring. (Structure source : Homology model)*

The highest-ranked modelled structure of the C-terminal domains after minimisation and a 20ns MD simulation of G0 showed good stereochemical quality, with 96% of residues occupying favoured regions in the Ramachandran map (***Figure 4B***) and conforming to adequate backbone conformation and non-bonded interactions by scoring 98% in the ERRAT overall quality score (***Figure 4C***). The ProSA calculated Z-scores (***Table 1***) further validated the reliability of the structures, with scores within the range typical of native structures, as seen in ***Figure 4A***. It is worth noting that, despite the modelled structures showing good stereochemical quality, model assessment tools such as ProSA and Verify3D have a shortfall in accurately assessing non-globular protein folds [322] and loops [323] , and may thus present low scores.

The structural fold of the G0 model comprised an N-terminal helix from 301-335, and two α-helices, H1 (338-360) and H2 (368-397) joined into a hairpin, as observed by [284]. The ApoL1

G1 and G2 models exhibited similar secondary structure elements to G0, especially with helix αH1 (338-360). In the G2 models, the H2 helix was truncated due to the N388/Y389 deletion, resulting in the loss of secondary structure and a helix-to-loop transition from residue Asn386. RMSF of the subsequently modelled variants corroborated the secondary structure predicted. This transition was observable in the motion of the structural elements of each variant.

In the RMSF calculated throughout the MD trajectories, notable fluctuation was observed in the loop regions between defined helices N-terminal and αH1 and αH2 helices. Higher flexibility was additionally noted for the truncated ends of the ApoL1 G2 C-terminus.

2.3.2 Principal Component, Rg and RMSD analysis

The RMSD and Rg of the protein trajectories were coupled with the ED calculations to obtain a comparative overview of the global motions of the proteins.

Essential dynamics calculations of the 96 residue structures, G2 and G2G1, and 98 residue structures, G0 and G1 variants, represented a variance of over 60% (***Figure 5A***). Each of the variants was removed from the wild type to varying degrees. Due to the ordering of observed motions decreasingly, a large part of the conformational fluctuations was described by the first few principal components [324], meaning these PCs are most relevant.

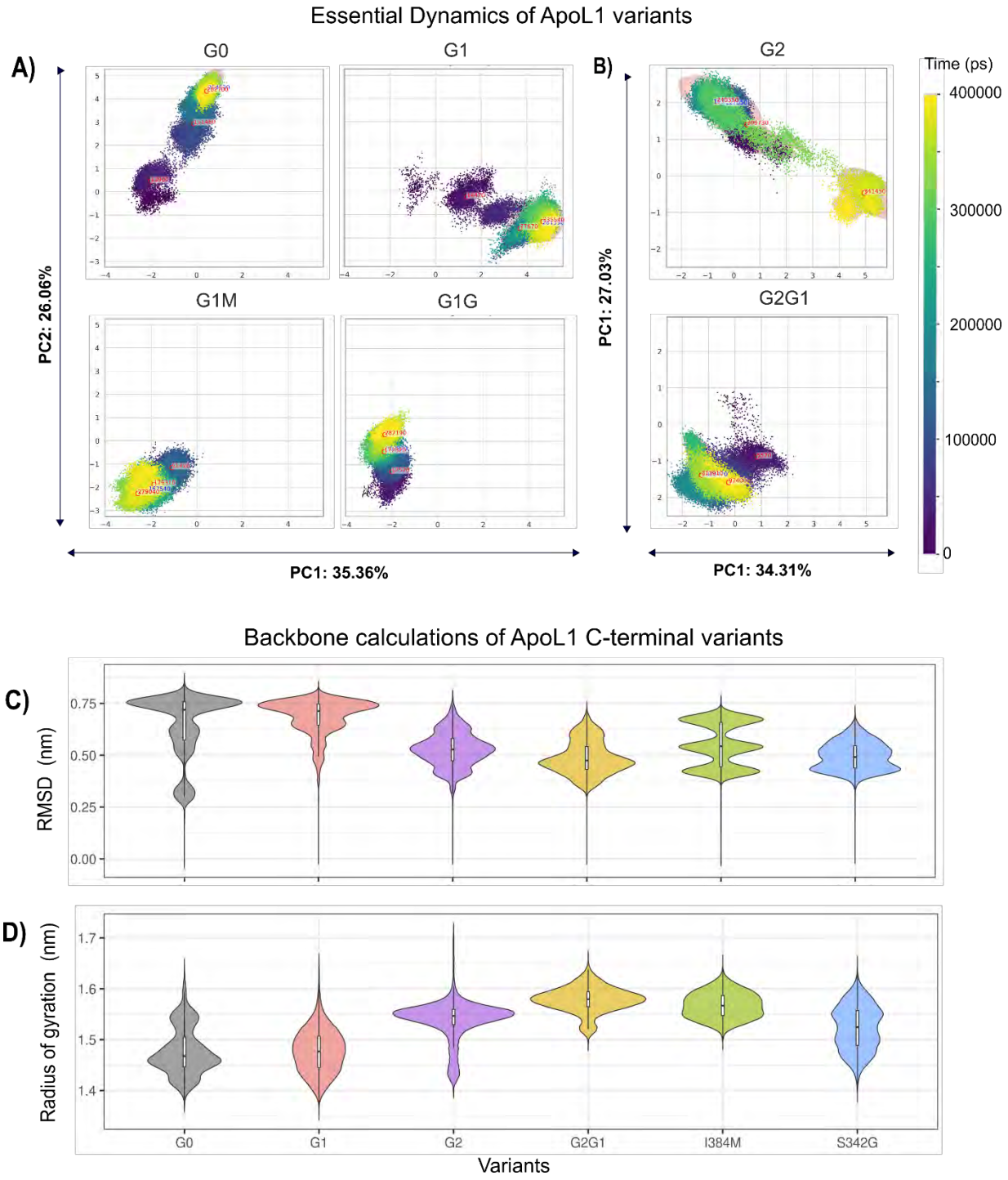


*Figure 5: Essential Dynamics of ApoL1 C-terminal structures for **A)** G0, G1,G1M and G1G and **B)** G2 and G2G1 variants of the ApoL1 C-terminal structures. The structures in A are 98 residues long, while those in B are 96 residues long due to the G2 deletion. The **C)** RMSD and **D)** Rg values of the structures over the 400ns simulations are represented in kernel density violin plots in .*

G2G1 sampled more conformational space along both PC1 and PC2 than G2. G1 traversed the most conformational space compared to G0, G1M and G1G along PC1. The G0 and G1 variants exhibited multimodal higher RMSD and inversely lower Rg than their companions, mirrored in the PCA, as they converged after the 200ns, as seen in the line graphs. Conversely, the G1G and G1M variants, which differ by a single residue, sampled similar conformational space, especially along PC1, sharing comparable medians as seen in the RMSD violin plots in ***Figure 5***. When visually inspected, the G1M N-terminal helix displayed a higher rate of fluctuation, resulting in three equally sized peaks recorded, which appeared adjacent in the PCA. Of the 96 residue variants (G2 and G2G1), G2G1 sampled less conformational subspace along both PCs, showing a higher but more stable Rg [227] than G2.

2.3.3 Dynamic Residue Networks

DRN was calculated using a top 5% cut-off for *BC, CC, DC,* and *KC*, in ***Figure 6***, to deduce the role C-terminal variants of ApoL1 had on the overall communication network in the protein. Shifts in residue importance have implications on the functionality of the protein. Overlap was observed in the centrality metrics due to the size of the ApoL1 structure. Most high centrality residues, hubs [325,326], were shared by the wild-type and at least one variant form of ApoL1. For this discussion, we defined *prominent hubs*, which are shared across at least three of the six structures being studied.

The *BC* and *DC* analyses captured the most extensive wild-type communication changes against variants. *DC* and *KC* concentrated the neighbouring residues in the surrounding residue networks, while *CC* was concentrated in the core of the α-hairpin identifying residues crucial to structural integrity.

BC highlighted residues strategically located to provide the shortest path between points in the protein segment. The variance in *BC* hubs suggested a shift in bridging residues due to the presence of the mutations. The G2 mutations resulted in the removal of two nodes in the *BC* residue network. This was manifested in the absence of hubs beyond Lys381 and the gain of Phe343 as a *prominent hub*.

Val346, centrally located in the core of the structure, was the most *prominent* in the four centrality metrics calculated, with its hub status being shared by at least four of six ApoL1 structures. For *BC,* the location of Val346 in αH1 provided access to the terminal ends of the

structure, both along the length of the structure and through interhelical contacts. In the *DC* and *KC* calculations, *KC* revealed a more extensive reach in hubs as the metric includes the added global influence of nodes over *DC,* highlighting residues with an increase each in neighbours. The result of this difference in the two metrics was the inclusion of Phe343 as *KC prominent* hub.

Most notable was the high *CC* and *prominent* status recorded for both residues Phe343 and Val346 in all variants but the G1M. While the substitutions of isoleucine and methionine are considered safe mutations [327] due to both being non-polar and the structural similarities, they allow the methionine side chain conformation to mimic that of isoleucine [328,329].

DRN analysis of both G1 and G1M mutations showed predominantly unique communication paths across all the centrality metrics compared to the other variants and the wild-type.

This observation was complemented by the unique *BC* hubs (Val349 and Asn388) shared between G1 and G1M or common losses in *DC* and *KC* (Val346 and Tyr350). The difference may be due to the change in hydrophobicity [327] between the two residues rather than steric differences. Additional hubs gained in the G1M variant include Lys345 in *BC* and *CC*, Val321, Val349 in *DC*, and Leu352 and Lys381 in *KC*.

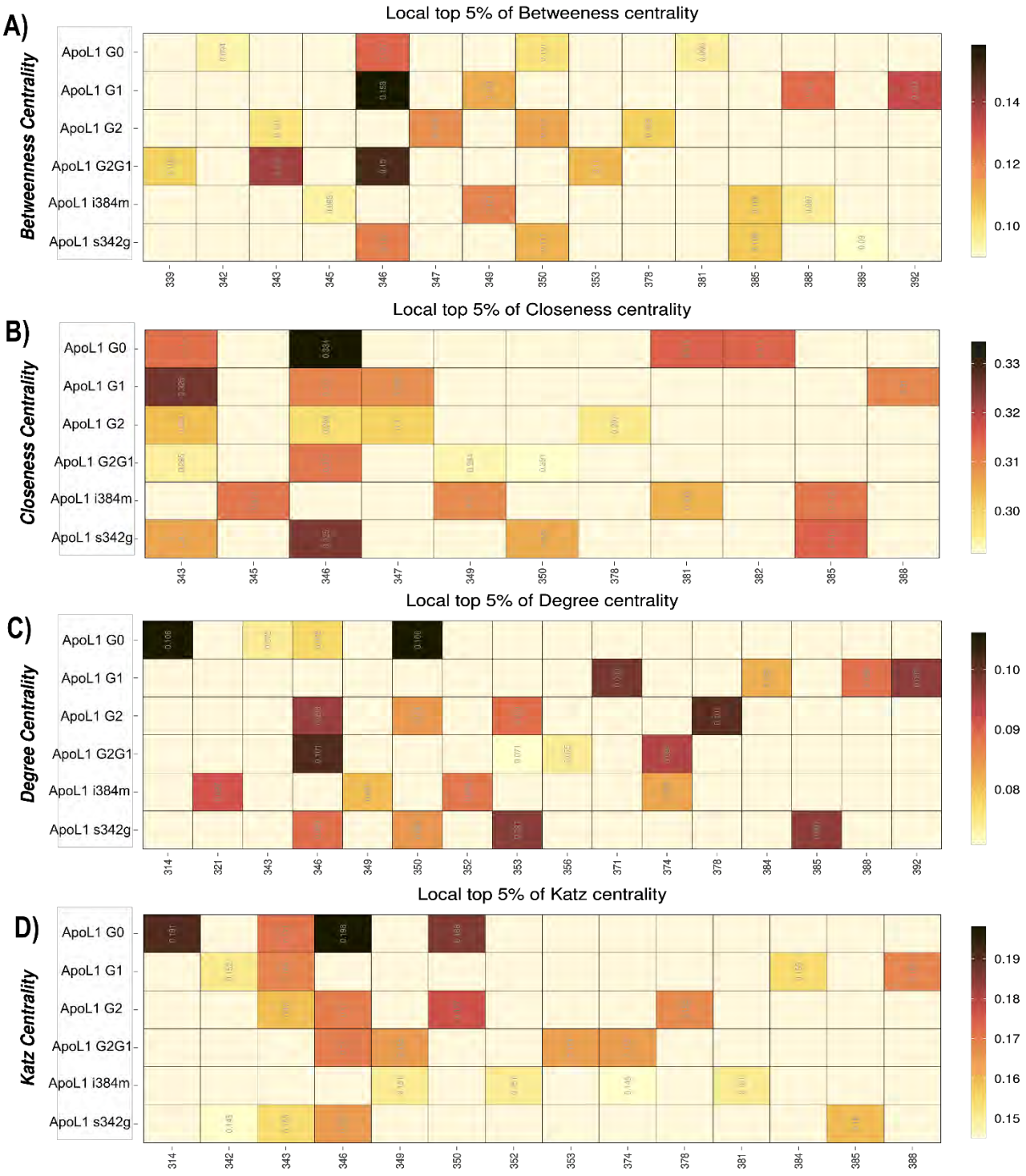


Figure 6: Heatmaps of DRN calculation, showing the top 5% values of the centrality metrics ***A)*** *BC,* ***B)*** *CC,* ***C)*** *DC, and* ***D)*** *Katz Centrality of ApoL1 variants. Residues detected as hubs are annotated with the centrality value in the cell. The increment of centrality values for each metric corresponds with the intensity of the colour bar on the right of each heatmap.*

2.3.4 Contact maps analysis

Contact map analysis revealed the contact, and intra-molecular interaction, between the adjacent α-hairpin helices (αH1 and αH2), disrupted to varying degrees by the presence of the ApoL1 variants. While ApoL1 G0 Asn388 reveals contact with residues 326-333, that is observably lost in the variants.

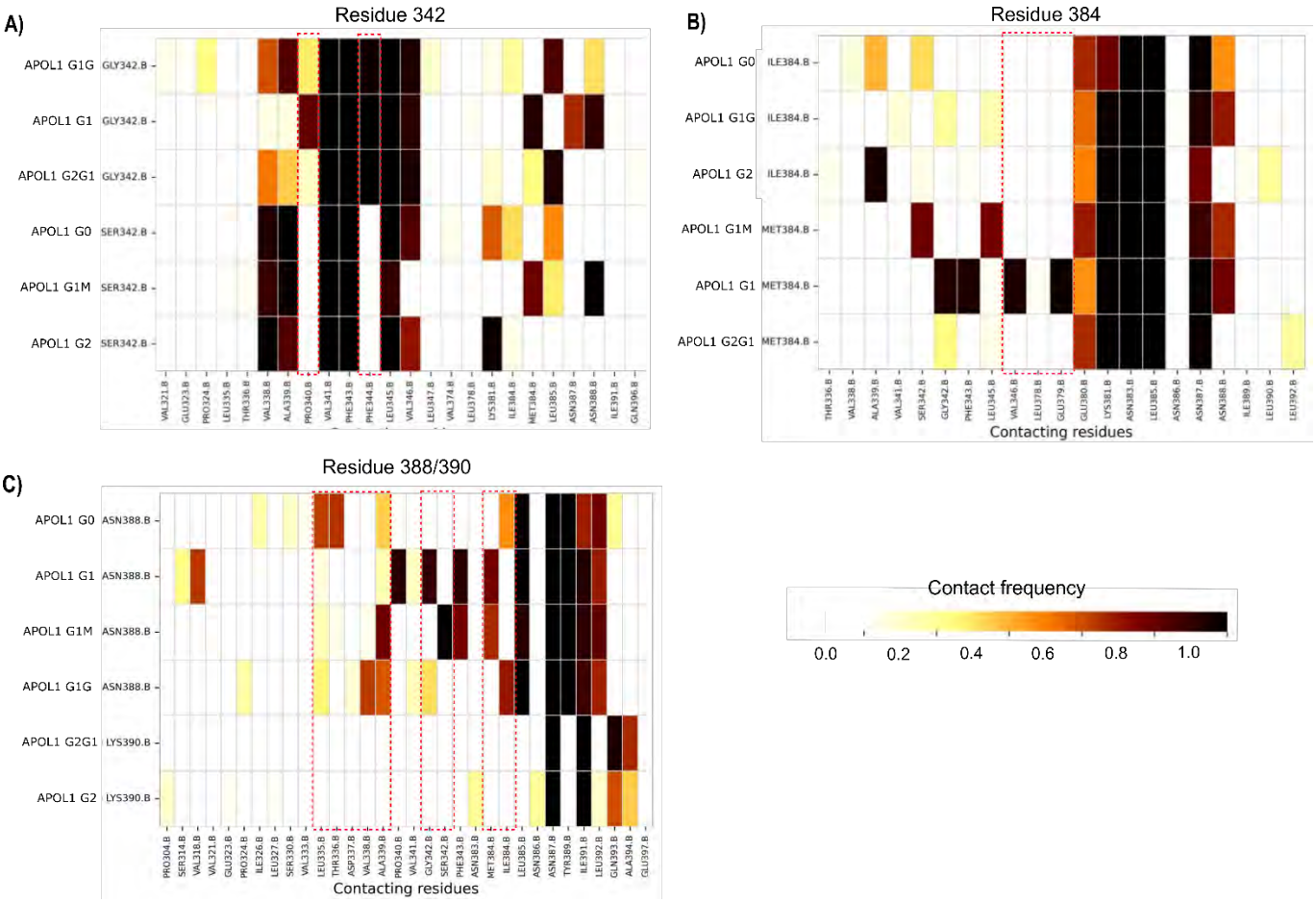


Figure 7: Contact heatmap analysis was conducted on residues that were sites of mutation in the ApoL1 C-terminal, namely residues ***A)*** *342,* ***B)*** *384 and* ***C)*** *388/390. Due to the deletion of 388 and 389 in the G2 variants, 388 in C is compared with 390. Notable changes in the contact network of high-profile residues due to the substitution and deletions are encircled in red.*

Asn388 in, ***Figure 7***, exhibits contact in the G1 variants with Leu335, Ala339, Res384 and Leu385, which is lost with its absence in the G2 bearing variants of the ApoL1 C-terminal. Residues 335 and 339 both lie in αH1, likely contributing to the stability of both helices in the α-hairpin. Interestingly, the asparagine residue also facilitates contact with both variations of residue 384.

The I384M substitution introduced minimal changes in adjacent contacts in G1, G1M, and G2G1, likely due to the identical hydrophobic side chain packing for Ile and Met [327].Notably, the presence of Met384 appears to encourage contact between Asn388 and Phe343. Lys390, which moves up in G2 deletions, lacks these contacts, additionally opting for increased contact with Asn383 and 386 versus the G1 residues observed with Asn388. We also observed that the presence of Gly342 initiates contact with Pro344 and Phe340 not present on the serine-bearing variants. All variants of residue 342, however, maintain interaction with their iteration of residue 384. Met384, especially in the G1, exhibits the most pronounced effect on the network after the deletion mutations, echoing observations noted in the DRN analysis.

2.3.5 Hydrogen bond occupancy analysis

Hydrogen bond occupancy analysis was conducted to validate and expand the observations noted in the local analysis. We tracked the occupancy between αH1 and αH2 residues in ***Figure 8***, to assess the effect of the variants on intramolecular hydrogen bonding and the stabilising effect resulting from hydrogen bonding.

We noted the hydrogen bond participation of Asn388 with αH1 residues of the G1 variants that was inadvertently absent in all G2 variants. This observation is likely the cause of the stabilising effect noted for G1G and G1M in the ED, RMSD, and Rg analyses. The contact maps showed the stabilising effect with the Asn388 contact frequency with residue 342 and higher 384 frequency compared to the wild type.

Aside from the hydrogen bonding lost due to the absence of Asn388, a drastic reduction in inter-helix was recorded in the G2G1 variant. This largely suggested a compounding effect of both G2 and G2 co-occurring, as noted in the Rg analysis.

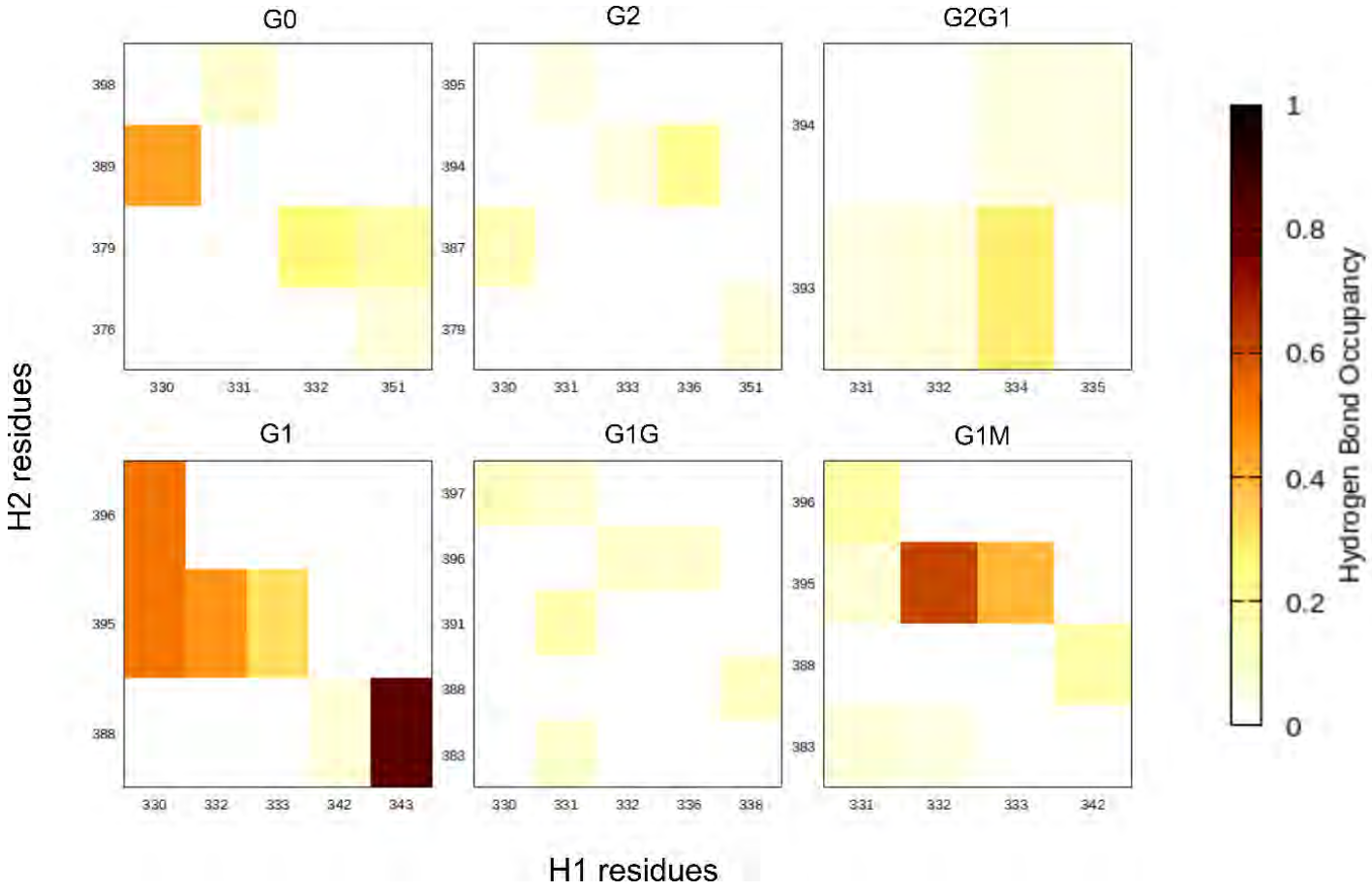


Figure 8: Hydrogen bonding between αH1 and αH2 residues was tracked over 300ns in hydrogen bond occupancy. Occupancy is ranked from 0-1 (in the colour bar), with an occupancy of 1 signifying the continuous maintenance of the bond throughout the simulation

2.3 Conclusion

Here, we analysed the WT and five mutants structures of the ApoL1 C-terminal region to elucidate the role mutations play in the intramolecular dynamics of the protein. Understanding the effect of ApoL1 variance is crucial to the pathogenicity of r-HAT.

Data analysis was run in two steps: Global analysis, which included PCA, RMSD, and Rg, and local analysis, consisting of RMSF, DRN, contact maps, and hydrogen occupancy calculations.

Our global analysis showed how the mutations affected the entire C-terminal of the protein, including conformational stability and flexibility. The differences in sequence resulted in notable changes in the global and local behaviour of the protein as exhibited in the RMSD, an

observation likely amplified by the smaller size of the protein segments allowing greater flexibility. The differences in Rg indicate these changes in the spatial conformation of the ApoL1 coiled-coil, where variants recording higher radii exhibit pronounced internal disruption [227].

Local analyses revealed that the presence of the variant residues in ApoL1 had a noted effect on the intra-molecular interactions of the hairpin. Interestingly, we noted that G1M imposed a more pronounced effect on the residue network than the combination G1 due to the disruptive nature of the methionine substitution. The effect of the variations appeared amplified in the combination of G1 and G2 as contact map analysis revealed active interaction between the residues involved 342, 384, 388 and 389. The glycine and methionine substitutions maintained contact with Asn 388, which was not re-established with Asn390 in the presence of G2. The observations were highlighted in hydrogen occupancy analysis and in agreement with studies by Mayanja *et al.* [304], which suggest a more significant effect of the G2 deletion on the structural stability of ApoL1 [304].

Site-directed mutagenesis[282] in combination with binding assays to ApoL1 partners such as the SNARE protein, vesicle-associated membrane protein 8 (VAMP8), [312] could be employed to further investigate the role residues 342, 384, 388 and 389 play in the stability of the ApoL1 C-terminal coiled-coil, as well as exploit the possible role of residues identified through DRN analysis.

Chapter 3

Variations in SRA

3.1 Introduction

The *Trypanosoma brucei rhodesiense (Tbr)* and *Trypanosoma brucei gambiense (Tbg)* subspecies of trypanosomes can cause acute and chronic HAT, respectively, due to having developed mechanisms to resist human serum ApoL1-dependent apoptosis [266–268].

The mechanisms by which the *Tbg* parasite evades the host immune system are complex but are known to involve the protein *T.b.gambiense*-specific glycoprotein (*Tg*sGP). Parasite lysis in human serum is restored upon deletion of TgsGP [330,331]. *Tbr*, on the other hand, bypasses innate immunity through the expression of the serum-resistance-associated (SRA) protein, a truncated trypanosome variant surface glycoprotein (VSG) [283,332].

The SRA gene coding for the protein was discovered and characterised from a Ugandan strain of *T. b. rhodesiense* [333–336] and has been identified in most *Tbr* variants tested [333,337,338]. The resulting protein is 396 residues long, consisting of eight helices (H1-H8) held together by ordered and disordered loops and a single disulphide bridge (Cys101-Cys180) [339] ***(Figure 9)***.

The resistance of *Tbr* to ApoL1-mediated apoptosis results from the direct binding of the SRA protein to the SRA-interacting domain of the ApoL1 C-terminus [283,340]. While it is related evolutionarily to the VSGs lining the outside of the trypanosome plasma membrane [341], SRA is localised in endocytic vessels [342,343]. As a result, the SRA-ApoL1 interaction is theorised to occur in the lysosome, leading to blockage of ApoL1 pore formation [280,332] and, ultimately, loss of lytic function [276,283].

Undoubtedly, the role of SRA in *Tbr* infectiveness is vital. As such, a detailed structural investigation of SRA is crucial to understanding immune escape by the parasite. In this study, we aim to investigate structural variants of serum resistance-associated protein using structural bioinformatics approaches and literature-identified ApoL1 binding sites [339] to assess the effect of the variations. We additionally aim to further understand SRA protein behaviour by

identifying residues critical to communication within the protein, as this may provide further insight into the nature of the interaction.

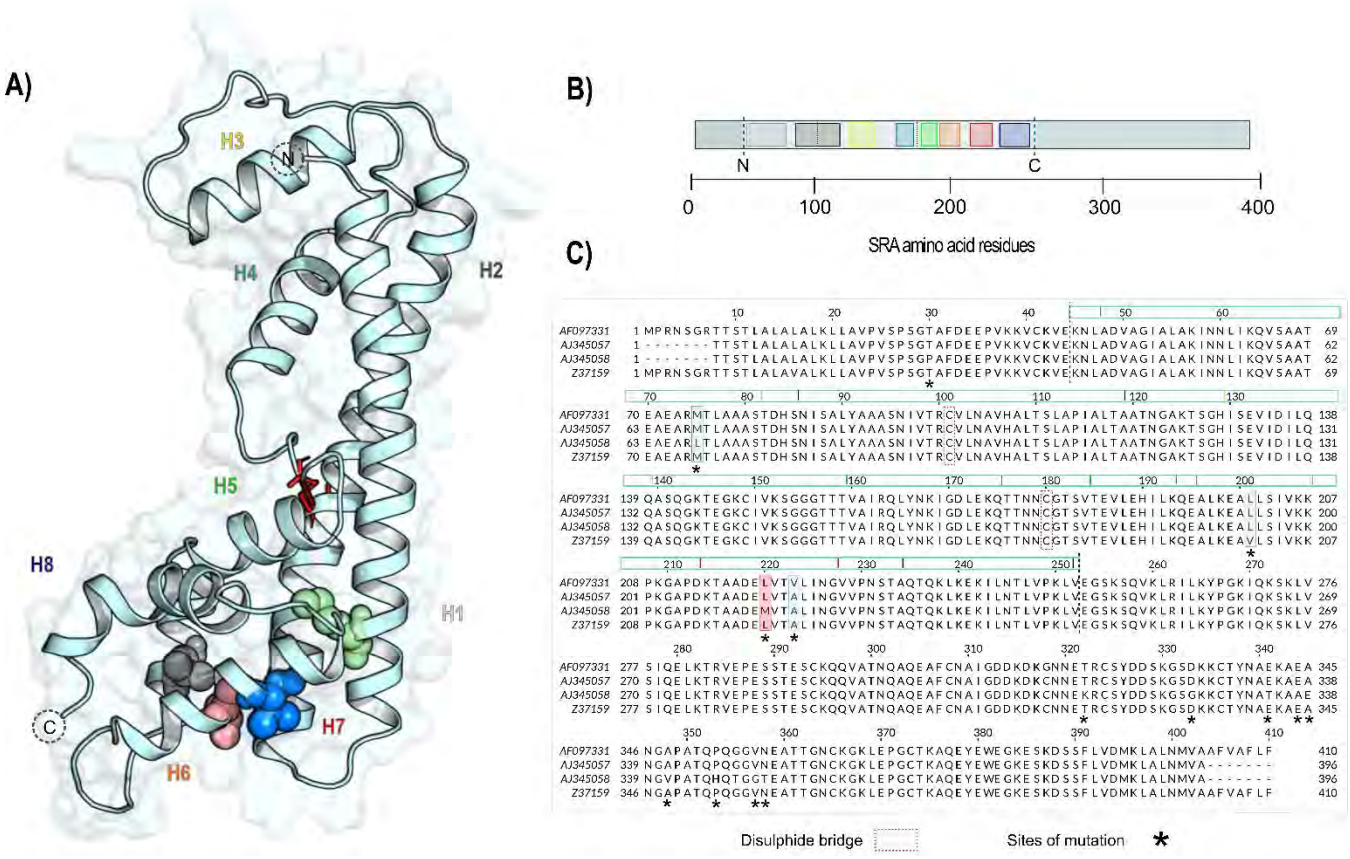


Figure 9: ***A)*** *The structure of SRA (PDB ID:6elu), consisting of eight helices, represented in* ***B****, including the location of Cysteines 101 and 180 of the disulphide bridge in maroon. Helices H1, H2, H3, H4, H5, H6, H7, and H8 are highlighted in grey, black, yellow, teal, green, orange, and navy blue , respectively. The multiple sequence alignments in* ***C)*** *present the location of the SNVs in the four SRA variants identified in sub-Saharan Africa with the asterisks. The locations of the SNVs identified are mapped onto the structure in A as spheres, corresponding MSA colouring.*

3.2 Methods

3.2.1 Data retrieval and Sequence analysis

The NCBI [344] and Uniprot [345] databases were the primary sources of the SRA sequences. In total, the sequences were filtered down to four (accession numbers: AJ345057, AJ345058, AF097331 and Z37159) which, for this study, were assigned names SRA1-SRA4 and aligned using the ClustalW [346] tool of Jalview [347] with default settings.

3.2.2 Homology modelling

The crystal structure of SRA was retrieved from RCSB PDB ((PDB ID:6ELU) [339], which had missing residues 142-158. MODELLER v9.20 [320] was used to rebuild the mentioned section using the *Trypanosoma brucei* Variant Surface Glycoprotein motif structure (PDB ID:2VSG). The rebuilt SRA structure was subsequently used as a template to generate 100 models of each SRA variant, using the modelling parameters and validation procedure applied in *Chapter 2*.

3.2.3 Molecular Dynamics

The CHARMM36 [321] force field was applied for the parametrisation of the proteins in the all-atom molecular dynamic simulations run in GROMACS 2018v6 [203,348] to achieve 300ns production runs. All the MD simulations were performed under periodic boundary conditions, maintaining the systemic parameters utilised in ***Chapter 2***.

The MD trajectories were corrected and fitted using the *gmx trjconv* module and subjected to classical conformational analyses of RMSD, RMSF and Rg using the respective *gmx* modules. The resultant values were plotted for interpretation in RStudio *ggplot* and *R-base*.

3.2.4 Intramolecular bond analysis

Hydrogen bond analysis was conducted using a maximum bond distance of 3.5 Å and a bond angle >30 in AmberTools21 [265] facilitated in CPPTRAJ [349]; where hydrogen bond occupancy measured the number of frames in which hydrogen bonds were maintained as a percentage of the whole simulation. Additionally, Visual Molecular Dynamics (VMD) [239] was used to monitor disulphide bridges and their bond length throughout the MD production runs.

3.2.5 Comparative essential dynamics

A step size of 1 frame, with a *k*-mean clustering setting of 3, were set for the comparative PCA calculations of the SRA variants. The *k*-mean clustering functioned to estimate single conformations from *k* possible probability density maxima [237] . The procedure and analyses applied were parallel to those used in ***Chapter 2***, apart from running a single calculation, as all structures were the same length.

3.2.6 Dynamic Residue Networks

MDM-TASK-web [237] was applied for the calculation of the DRN, using a distance of 6.7Å between nodes and a step size of 50. To identify nodes important to communication in the residue networks, 250ns runs from equilibrated sections of the MD trajectories were used to calculate the *BC*, CC, *DC*, and *KC*.

Additionally, weighted contact maps and subsequent heat maps were calculated for residues that were sites of variation and had high centrality scores, using a step size of 1 frame for data collection.

3.3 Results and Discussion

We applied RMSD and ED calculations to quantify the global motions of the SRA variants and RMSF for local residue motions. The intramolecular behaviour of the variants was investigated using hydrogen bond occupancy, DRN, and contact map analysis.

3.3.1 Principal Component and RMSD analysis

PCA is widely used to predict essential protein motions in low-dimensional subspaces [237]. The ED was calculated to investigate the functional dynamics and the number of low-energy conformations sampled by each SRA variant in order to deduce the probable effect of the mutations on the overall protein. The majority of the variance was captured in PC1 and PC2, at over 69.98% of the total variance over 300ns. The two PCs accounted for 55.54% and 14.44% of the variance, respectively, as shown in ***Figure 10A***. Each structure experienced overall conformational displacement that pointed to the ease with which each SRA variant settled into an energy minimum conformations.

For most variants, the largest displacement was recorded along PC1, which incidentally represented a significant portion of the variance for most variants except SRA3. Noteworthy was SRA4, which sampled the most conformational space along both PCs compared to its counterparts, in contrast to SRA1, which sampled the least. The RMSD analysis was complementary to the PCA findings of SRA4, which presented the highest RMSD median (***Figure 10B***) and a trimodal distribution in the violin plots. SRA2, which, like SRA4, sampled relatively significant space along PC2, also exhibited a trimodal RMSD density distribution, suggesting increased conformational changes, notable in the line plot flexibilities.

SRA1 and SRA3 had bimodal RMSD density distributions, with SRA1 settling at a lower RMSD peak observed in the ***Figure 10B*** violin plots and SRA2 at a higher peak. The conformational subspace sampled by these two variants along both PCs in ***Figure 10A*** mirrors the PMSD observations, suggesting increased conformational stability for SRA1 and SRA3 compared to their counterparts.

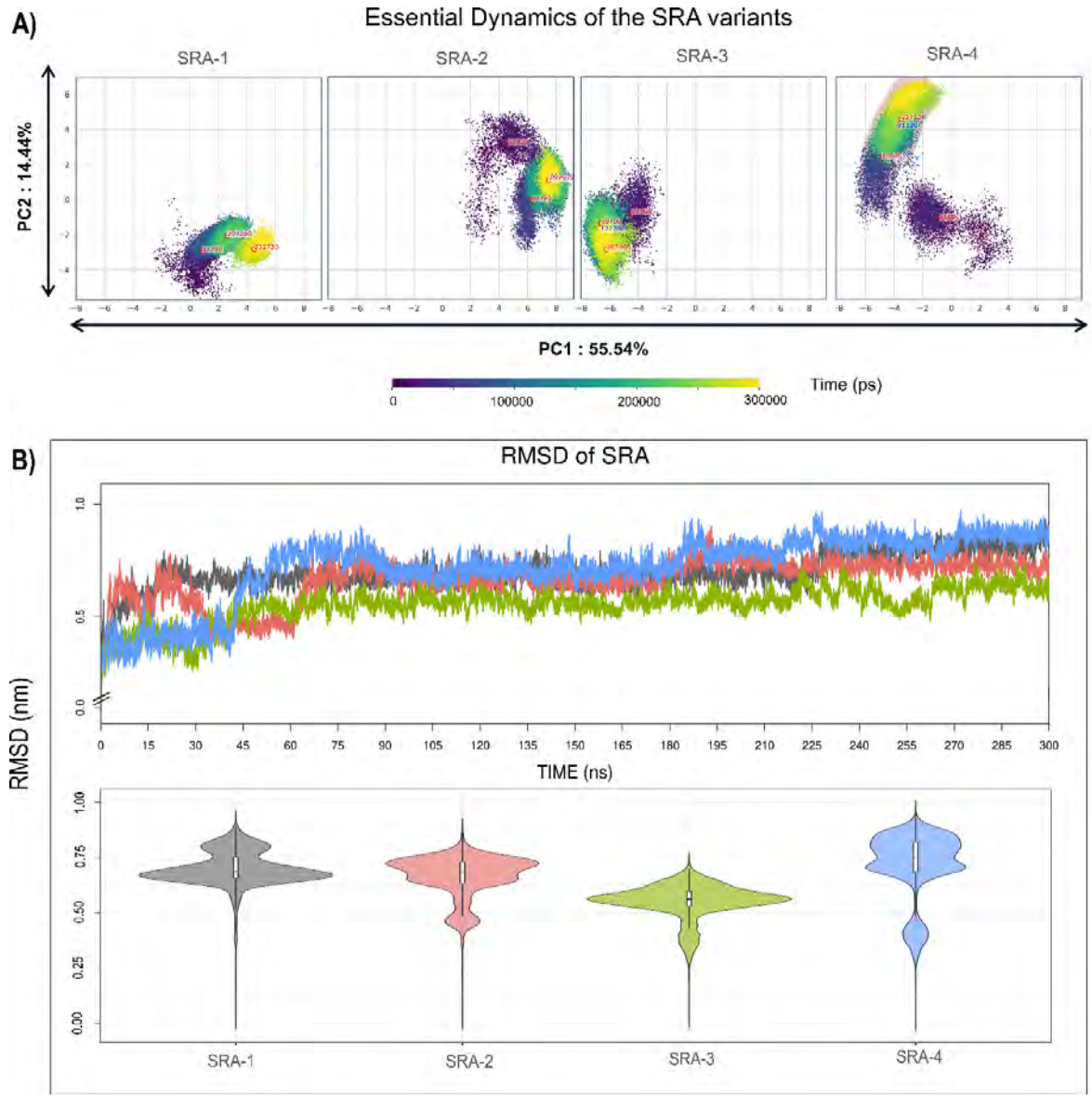


Figure 10: ***A)*** *Essential dynamics of the SRA variants, determined through PCA calculated over times (300ns). The colour progression of the colour bar characterises the progression of the time in MD simulation. PC1 captures 55.54%, and PC2 14.44%. The RMSD of the protein backbones, in* ***B****, plotted in line and kernel density violin plots, is representative of the global motion exhibited during the simulation and is mirrored in the PCA.*

3.3.2 RMSF and intermolecular hydrogen bonding

RMSF analysis was conducted to deduce residues and regions of the proteins likely contributing to RMSD fluctuations. The RMSF observations were corroborated the observations recorded in the hydrogen bonding occupancy, as hydrogen bonds play a key role in protein structure stability[256].

The regions exhibiting low hydrogen bond occupancy corresponded to regions with high RMSF, especially in SRA1 and 2. Naturally, most fluctuation can be attributed to the loops joining each helix. Notably, the high RMSF observed from residues 116-130 can be attributed to the orientation of H3. The helix packs anti-parallel to H1 and 2, resulting in decreased intramolecular hydrogen bond participation with the core of the protein and, thus, increased flexibility. Cys101 in H2 is reported to form a disulphide bridge with Cys180 [339]. The disulphide bridge likely confers stability and compactness to the predominantly helical SRA by tethering the C-terminal loop of the protein to a helix in the N-terminal. The length of this bond was monitored and shown unaffected over the course of the simulation across all variants (***Figure S 7***).

SRA4 presented increased backbone fluctuation in most regions, including some outside of the expected loops in***Figure 11A***, tying in with the observed conformational sampling noted in the RMSD and PCA calculations in section ***3.3.1***.

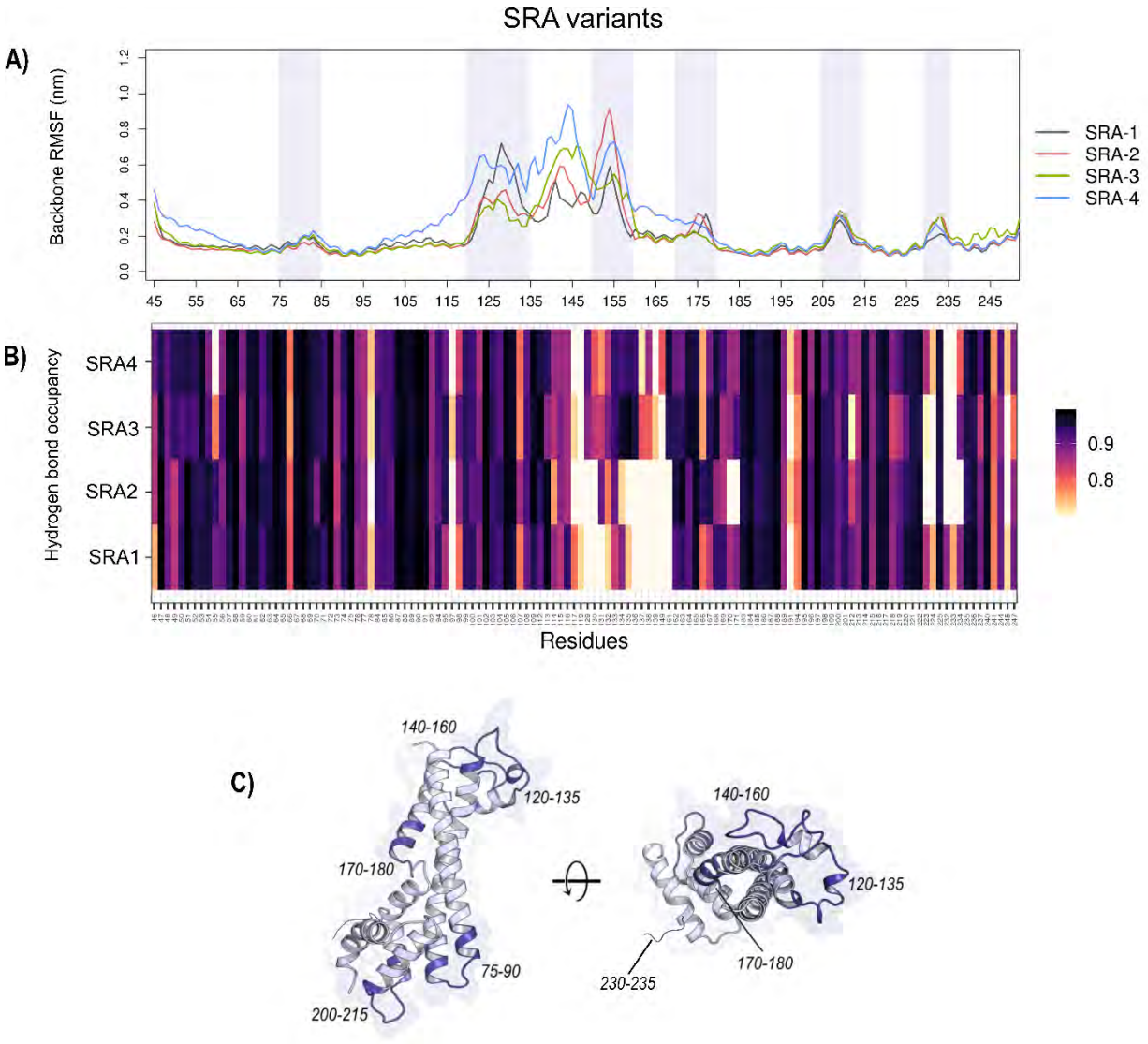


Figure 11: ***A)*** *Backbone RMSF and the* ***B)*** *hydrogen bonding occupancy of the SRA variants over the course of the 300ns simulations. Notable regions of high flexibility in SRA in A and low hydrogen bond occupancy in B coincident with the loop region of SRA highlighted navy blue in the structure in* ***C.***

(Structure source : PDB ID 6elu)

3.3.3 DRN analysis identified potential communication cold spots amid SRA mutations

A commonality in communication hubs was observed between SRA1 and SRA2 versus SRA3 and SRA4, likely due to the mutations shared between each of the duos. Despite the split in observations, residues Val98 and Ala105 were *persistent hubs*, which, as defined by Boateng *et al.* [325] , are retained across all systems compared in the analysis. Despite each metric

isolating different measures of importance within a network, the residues maintained *persistent* status in all investigated metrics highlighting the cruciality of communication (***Figure 13A-D***).

Persistent hubs aside, the *BC* captured the most divergence in the top 5% hubs compared to the subsequent metrics due to the nature of the metric.

The *BC* analysis captured the most divergence in the top 5% hubs across the SRA variants compared to the subsequent metrics. Due to the accumulative nature of the shortest paths calculation of *BC,* the changes in the residue networks brought on by the mutations, regardless of location, were discernible in the heatmap of ***Figure 12A*.** This is demonstrated in SRA3 and SRA4, which exhibited unique *BC* hub candidates in trading the inclusion of H1 residues as high scoring (top 5%) hubs for excluding a number of H2 residues.

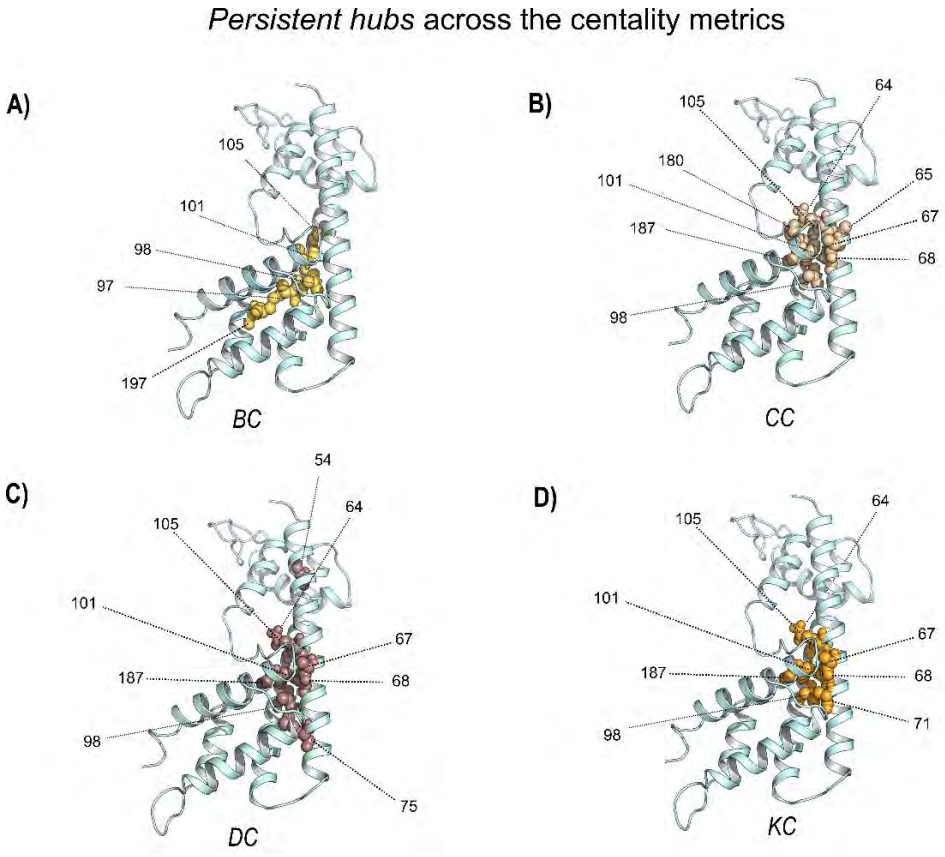


*Figure 12: Persistent hubs identified in centrality metrics **A)** BC, **B)** CC, **C)** DC and **D)** KC are mapped onto the SRA structure in the coloured spheres, highlighting the position of the hubs. Each metric emphasises a different measure of importance within the network, resulting in a unique distribution of persistent hubs. (Structure source : PDB ID 6elu)*

However, regardless of distinct *BC* profiles, *persistent hubs* were recorded for the variants. Notable *persistent hubs* in *BC* included Ile97, Ala98, Cys101, Ala105 and Leu197 (***Figure 13A***), which mostly resided in H2, except residue Leu197 in H6, playing a bridging role between the N and C terminals of the protein. These residues are well-positioned to provide the shortest path across the two ends of the protein, as shown in the structure in ***Figure 12A***.

The *CC* analysis revealed high agreement across the variants, with the top 5% hubs ***(***Figure 13***B)*** populating the core of the protein traversing H1, H2 and H5 when mapped onto structures in ***Figure 12B***. Due to their location in the protein core, residues with high *CC* values are typically important for maintaining structural integrity. A maintained consistency noted in the *CC* hubs points to the presence and location of the mutations having minimal effect on the function of the core residues. In SRA, this was encapsulated in the inclusion of both participants of the structure stabilising disulphide bridge[250], Cys101 and Cys180, in the shortlist of *persistent hubs* (see ***Figure 13 B)***.

The high connectivity nature of *DC and KC* analysis highlighted a similar residue profile to that of *CC* analysis, with the exception of mutating residue 75 in *DC* and Ala71 in *KC* analysis, denoted in ***Figure 13***. *KC* is an extension of *DC* (see ***1.2.1.7.2)***, while *CC* accrues residue

significance unrelatedly. *DC* assigns residue importance based on local contacts, while *KC* incurs penalties with increased neighbouring resides. A result of the derivative nature of the two metrics was the considerable number of shared *persistent hubs* in Gln64, Ala67, Ala68, Ala98, Ala105, and Val187 (***Figure 13C & D***). On the other hand, *CC* identified residues with the easiest and shortest access to the rest of the network, a property that, due to the helical and smaller nature of SRA, overlapped with local connectivity sought out by *DC* and *KC*. The *persistent CC hubs* Ala98, and Ala105 were both centrally located and a short path from neighbouring residues. Interestingly, we observed the most pronounced difference in the *KC* hubs of SRA1 compared to counterpart variants, including alanine residues 54, 116 and 119, but excluding Glu72 and Met75. This was likely an aggregated effect of each

mutation residue as the variant shared each of its residue states/substitutions with at least two other variants, thus lacking a unique substitution to attribute the distinct *KC* profile.

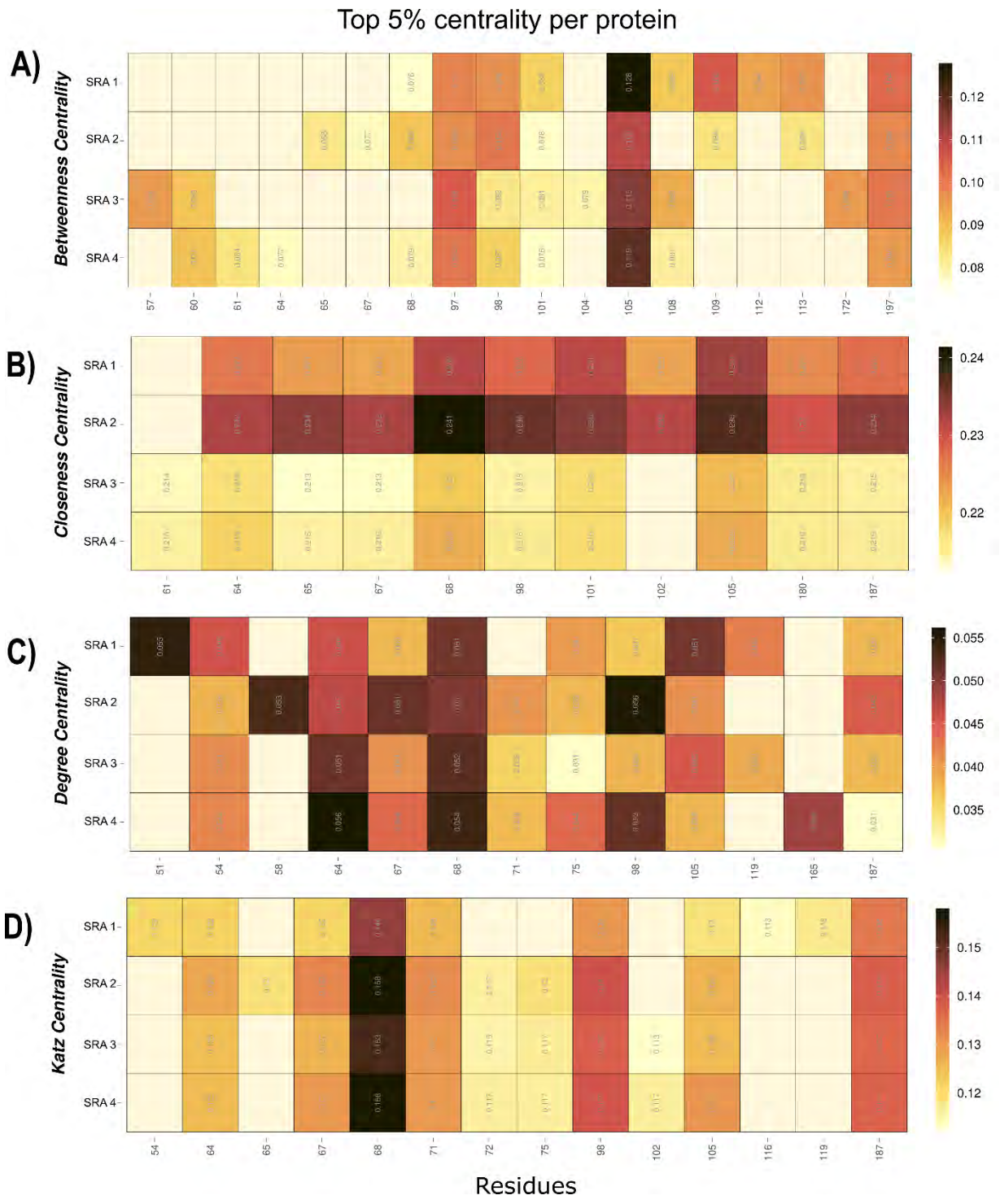


*Figure 13: The top 5% centrality hubs of metrics **A)** BC, **B)** CC, **C)** DC, and **D)** KC per SRA variant. The annotated heatmap cells represent identified hubs meeting the 5% cut-off. Persistent hubs were those observed in all complexes per metric. Low to high centrality values for each metric correspond to the intensity of the colour bar on the right of the each heatmap.*

The regions of the SRA structures housing low *CC, DC,* and *KC* residues 79-84 and 207-210 (***Figure S 6***) corresponded to loop regions, which were understandably poorly connected to the core of the protein.

Overall, despite the unique pathways induced by residue substitutions in position 75 in H1, 201 of H6, 220 and 223 of H7, *persistent hubs* were identified in all metrics. This observation infers the crucial participation of residues *persistent* in all metrics, such as Val98 and Ala105, to the functionality of SRA. Additionally, our observations revealed the minimal role the SRA mutations have on the internal functionality of SRA.

3.3.4 Contact Maps

The residue networks around prominent centrality residues Gln64, Ala65, Ala67, Ala68, Ala71, Ile97, Val98, Cys101, Ala105, Cys180, Val187 and Leu197, identified from DRN were further analysed to investigate shifts in residue contacts due to presence of the mutations. SRA2 demonstrated distinct changes, while most residues inspected maintained their network across the variants, as shown in ***Figure 14***.

Contact map analysis of *persistent hubs* and mutation residues in SRA2 revealed the most deviation in residue contacts compared to the other variants, likely due to the unique M75L and L220M substitutions housed by the variant. The Met to Leu substitution in position 75 triggered contact with Gly227 and increased frequency with Val98, while the reverse substitution at residue 220 had a nominal effect on residue contacts. Contacts introduced in SRA2 maps included S127-I130 for Ala65, Asn178 with Cys101, and Thr178-179 with Ala105, while Cys180 and Ala223 had varied frequencies with Thr178, His190, and Gln227, respectively.

Contact networks around the investigated residues in SRA1, SRA2 and SRA4 were comparable. Worth noting were Ala67, Ala68, Ala71, Ile97, Val98, Ala105, Val187 and Leu197, which at most showed minimal changes in both the contact and magnitude of contact with surrounding residues, despite the location of Ala71 and Leu197, four residues upstream of a mutation residue.

The retainment of residue contacts reinforced DRN observations, further highlighting the importance, likelihood of conservation, and the overall inconsequential local effect of the mutations. We additionally noted that the presence of mutations and change residue contacts,

as noted in SRA2 contact maps, was not an automatic analogue for a change in inter-residue communication and importance as defined by DRN analysis.

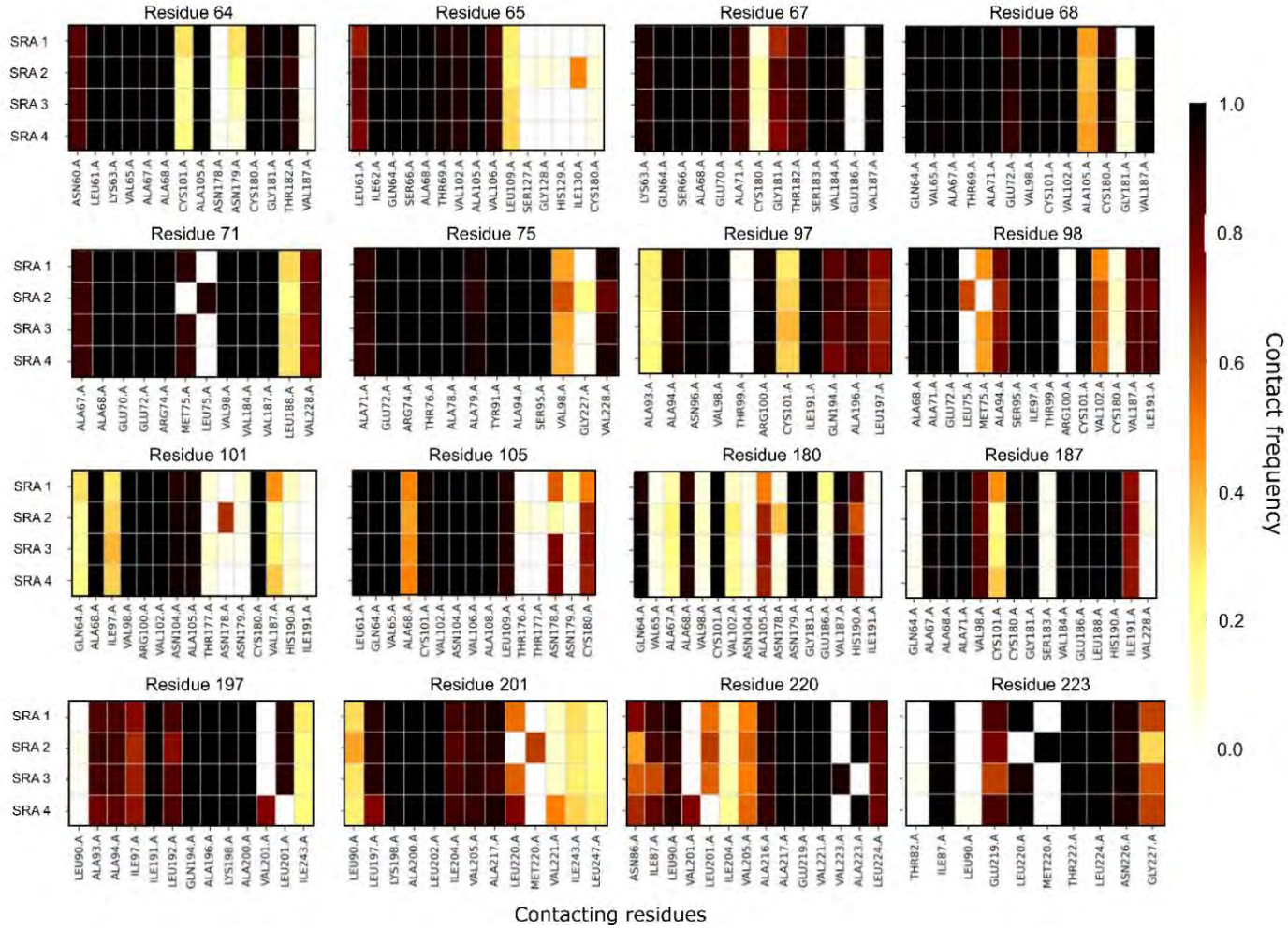


Figure 14: Contact heatmaps for SRA variants highlighting sites of mutation or high communication residues identified through DRN analysis. The frequency of contacts between residues in a network was monitored over the course of the MD simulations and is represented on a scale from 0.0-1.0 in the colour on the bar on the right.

3.4 Conclusion

In this chapter, we analysed the protein dynamics of four variants of SRA to elucidate the effect of the mutations on protein functionality and, consequently, the nature of SRA motion.

We investigated SRA variants using global and local analyses to achieve this role, as seen in ***Chapter 2***. Global analyses included RMSD and ED, which were coupled with DRN, contact map and hydrogen bond occupancy for local analyses.

Global analysis of the SRA variants revealed SRA4 as the most structurally flexible. The variant sampled a larger conformational subspace than its counterparts, exhibiting higher RMSF and, consequently, RMSD values. SRA4 carries a unique Val201 substitution compared to leucine in the other variants. While both residues were non-polar and aliphatic, leucine has a larger side chain affording it more hydrophobic interactions than valine. We, therefore, suggest SRA4 global dynamics were due to a decrease in structure stabilising hydrophobic interactions [350,351].

In contrast, substitutions in SRA1, SRA2 and SRA3 revealed minimal global differences were observed in protein behaviour across the structural SRA variants, putting to question the effect of the mutations on the SRA-ApoL1 interaction and, subsequently, the infectiveness of *Tbr*.

Our DRN analyses identified Ala98 and 105 as important across all centrality metrics studied, indicating that the residues play vital roles in bridging, influencing, and directing the flow of communication [326] within SRA. Regions expressing high centrality coincided with the experimentally determined ApoL1 binding sites [339]. High *persistence* in the *BC, CC, DC* and *KC* calculations and the general maintenance of comparable residue contacts suggested a marginal effect of the investigated mutations at residues 75, 201, 220 and 223 on protein integrity and functionality. It is likely that the parasite still manages to evade the host immune system regardless of which of the four SRA variants is expressed.

Overall, we concluded that the SRA mutations were inconsequential to the protein's function but were able to identify residues that are plausibly consequential. The identified residues through DRN analysis are likely mutational cold spots [352] key to the protein. In the case of SRA, these residues could be vital to the SRA-ApoL1 binding and represent promising targets for site-directed mutagenesis studies to disrupt SRA functionality. In addition, focusing on neighbouring residues, as informed by *KC,* could alter the protein activity and be relevant for

the disruption of a larger region of the protein, such as the dimerisation site cited to participate in ApoL1-binding [312].

Chapter 4

Protein-protein interaction in SRA-ApoL1

4.1 Introduction

The spread of HAT epidemics throughout history in sub-Saharan Africa [2] has been spurred on by a molecular arms race (***Figure 15***) between trypanosomes and humans [281,293] through unique, innate immunity evasion mechanisms by the parasite. Specifically, the expression of ApoL1-binding serum resistance-associated protein by *Trypanosoma brucei rhodesiense* results in the nullification of any trypanocidal activity [353,354] by host serum.

In turn, genetic polymorphisms in the human host have been shown to contribute to different responses to trypanosome infection [355]. Two genetic variants of ApoL1, G1 and G2, occurring at high frequency in humans of African descent [356,357] encode proteins that confer protection against the pathogenic subspecies of trypanosomes through circumvention of parasite evasion, restoring lytic activity [281,293]. As such, the inter-play between SRA and ApoL1 hinges the progression of r-HAT and is crucial for understanding and curtailing of the infection.

While it is related evolutionarily to the VSGs lining the outside of the trypanosome plasma membrane [341], SRA is localized in endocytic vessels [342,343]. As a result, the SRA-ApoL1 interaction is theorised to occur in the lysosome, leading to blockage of ApoL1 pore formation [332,340] and, ultimately, loss of lytic function [276,283]. Due to the helical nature of both proteins, SRA-ApoL1 binding has been proposed to occur via coiled-coil packing [284], or dimerization [312] between the ApoL1 C-terminal helix and the SRA N-terminal helix [295] as observed in VSGs homodimerisation [281]. However, the arrangement of helices in the SRA structure and location of the membrane-binding domain in ApoL1 bring these suggestions to question. The third N-terminal helix H3 of SRA packs against the coiled-coil of helices H1 and H2, potentially blocking the proposed dimerisation interface of the protein [339]. A separate study proposes a transmembrane pore-lining region in the ApoL1 (L335–S356) [358], which would occlude participation in a coiled-coil structure [306].

Though the study by Zoll *et al.*[339] demonstrates stable binding SRA and ApoL1, two lower affinity binding sites are identified in addition to the SRA N-terminal helix, suggesting a more complex binding. SRA has been documented to participate endosomally, localising in the cytoplasm [343,359] to confer *Tbr*-infectivity, yet it remains unknown if ApoL1 participates in trypanolytic activity in membrane-associated or cytoplasm localised form [276]. Regardless of the membrane relation of ApoL1, the dynamic nature of protein orientation in membranes [360,361], allows for the possibility of a membrane-bound ApoL1 to exposing the C-terminus to the cytoplasm to facilitate protein-protein interactions [358,362].

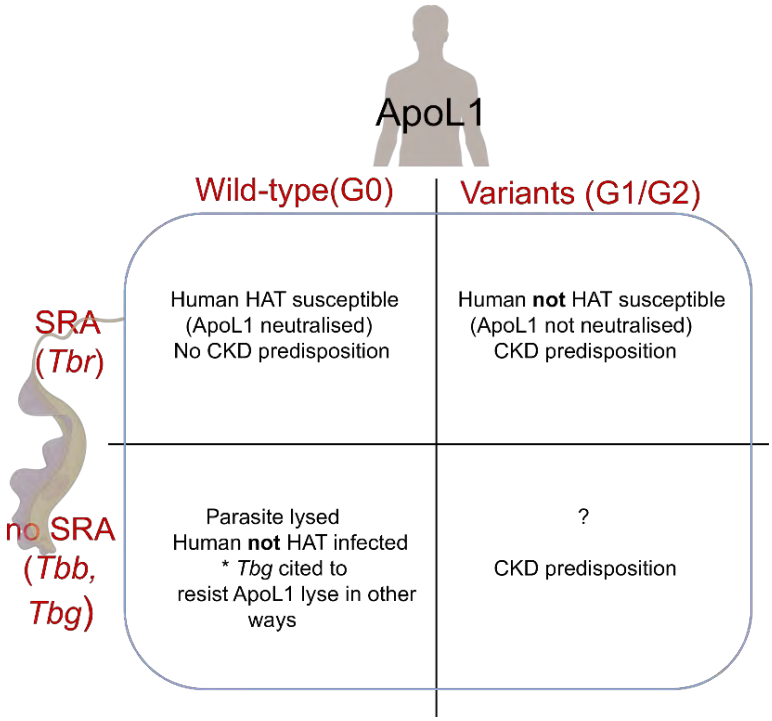


Figure 15: A schematic of the relationship between SRA and ApoL1, and its effect on the pathogenicity of HAT presented in a simplified correlation matrix.

In light of the questions surrounding the nature of the SRA-ApoL1 interaction, we designed this chapter to address the nature of the complexation. With the assumption of a localised form of ApoL1, we performed computational modelling of SRA to the ApoL1 C-terminal binding to understand the protein-protein specificity of the complex. Additionally, the role of ApoL1 variants, G1 and G2, in the protein-protein complex stability is investigated to understand the protein-protein specificity of the complex formation.

4.2 Methods

4.2.1 Protein-protein docking

Wild-type models used for ApoL1 and SRA were attained from *Chapters 2* and *3*, respectively.

Protein-protein docking was run on the HADDOCK2.4 server [173,182,363,364] using the default and bioinformatics intensity settings in ***Table 2***. Using the ApoL1 [339] and SRA [277,283,294] binding sites identified in the literature, ambiguous interaction restraints (AIRs) defining active and passive interface residues were generated. Run 3 SRA AIRs were specifically based on residues identified by *Zoll et al.*[339] for site-directed mutagenesis. The AIRs complexed the two proteins through a first iteration *(it0)* of rigid-body docking and minimisation to generate SRA-ApoL1 dimer models. The best-ranked dimers were then subjected to a second iteration *(it1)* for semi-flexible annealing using 5Å cut-off [365], and finally refined *(water)* by simulation in an explicit solvent (TIP3P) at 300K. The models were clustered on structural similarity using fraction of common contacts (FCC) and ranked in-cluster to give the top four models used for cluster scoring. Final cluster selection was based on the resultant HADDOCK score, cluster size, electrostatic energy, and the Z-score.

The HADDOCK Score is calculated based on the weighted sum of a number of energies, E_{vdw}, E_{elec},E_{desolv},E_{AIR}, and the buried surface area (BSA), broken down in ***Equation 11***:

$$_{it0}HADDOCKScore = 0.01E_{vdw} + 1.0E_{elec} + 1.0E_{desolv} + 0.01E_{AIR} - 0.01BSA$$

$$_{it1}HADDOCKScore = 1.0E_{vdw} + 1.0E_{elec} + 1.0E_{desolv} + 0.1E_{AIR} - 0.01BSA$$

$$_{water}HADDOCKScore = 1.0E_{vdw} + 0.2E_{elec} + 1.0E_{desolv} + 0.1E_{AIR}$$

Equation 11: Calculation of the HADDOCK score

The top clusters and models were visually inspected, and the top four models of the best scoring cluster were selected for further analysis [366]. All protein structures generated in this study were visually examined in the PyMOL viewer [131].

Table 2: The docking input used for complex generation on the HADDOCK webserver. The AIRs were selected based on surface residues in the literature highlighted binding sites. Use of the bioinformatics settings increased the sampling size of the calculations

Docking run	***SRA active AIR residues***	***ApoL1 active AIR residues***	***Parameters***
1	58,59,60,62,63,65,66,67,69,70,72, 73,74,76,77,78	341,342,344,345,346,347,348,349,351,352,354, 355,356,357,358,359,360,361,362,363,364,365, 366,367,368,369,370,371,372,373,374,375,376, 377,379,380,381,383,384,386,387,389,390,391, 392,393,394,395,397,398	Bioinformatics
2	58,59,60,62,63,65,66,67,69,70,72, 73,74,76,77,78	341,342,344,345,346,347,348,349,351,352,354, 355,356,357,358,359,360,361,362,363,364,365, 366,367,368,369,370,371,372,373,374,375,376, 377,379,380,381,383,384,386,387,389,390,391, 392,393,394,395,397,398	Default
3	59,63,66,70	341,342,344,345,346,347,348,349,351,352,354, 355,356,357,358,359,360,361,362,363,364,365, 366,367,368,369,370,371,372,373,374,375,376, 377,379,380,381,383,384,386,387,389,390,391, 392,393,394,395,397,398	Bioinformatics

4.2.2 Molecular dynamics

To determine the dynamic stability of the HADDOCK generated dimeric complexes, MD calculations were performed. The system was set up as described in Chapter 1 (**section *1.2.3***) using the parameters applied in both Chapters 2 (**section *2.2.2***) and 3 (**section *3.2.3***) in GROMACS 2018v6 [348]. The production runs of 500ns were conducted on twelve HADDOCK generated complexed systems to filter out any dissociating complexes.

4.2.3 Trajectory analysis

Using the *gmx trjconv* module, the trajectories were corrected and fitted in preparation for analysis as outlined in the post-MD processing established in Chapter 2 (**see *2.2.3***) . To investigate protein stability and flexibility, RMSD, Rg, and RMSF of the backbone atoms based on residue contribution were calculated. Additional analyses included CoM distance calculations between the two proteins, implementing the *gmx distance* module. RStudio [367] was used for any statistical analyses of these calculations.

4.2.2 Intramolecular bond analysis

Hydrogen bond analysis was conducted in CPPTRAJ [265,349] to monitor interactions between the two proteins, while VMD [239] tracked disulphide bridge bond length throughout the MD production runs.

4.2.3 Dynamic Residue Network analysis

The first 50ns of the MD trajectories were skipped in the calculation of DRN in favour of post-convergence last 450ns of the runs. Centrality values was calculated per monomer in each complex to isolate any shared communication hubs across the systems. The top 5% of SRA values were shortlisted, while a cut-off of 10% was used for ApoL1 instead as this was more informative in analysing the shorter length protein.

4.2.4 Contact maps

Additionally, weighted residue contact maps and subsequent heat maps were calculated through the *contact_map.py* and the *contact_heatmap.py* scripts (available at *https://github.com/RUBi-ZA/MD-TASK/tree/mdm-task-web*) [237], using the same parameters as the centrality calculations.

4.2.5 Conformational Analysis

Principal component analysis (PCA) and normal mode analysis (NMA) were additionally used to characterise and investigate the conformational changes of the proteins over the length of the MD trajectory [237,240]. For PCA comparison, the essential dynamics script of MDM-TASK-web(*https://github.com/RUBi-ZA/MD-TASK/tree/mdm-task-web*)
compare_essential_dynamics.py was used.

4.2.6 Protein-protein complex and interface validation

Additional analysis of the protein-protein interfaces was performed using the PDBePISA (Protein Interfaces, Surfaces, and Assemblies service at the European Bioinformatics Institute) webserver [368], which assessed complex stability through chemical thermodynamic calculations and provided further insight into the participating interactions.

The PISA algorithm calculated the change in Gibbs free energy on dissociation from the estimated free energy of binding and the entropy of complex dissociation to determine which complexes were probable and expected to remain stable in solution.

4.2.7 Interface alanine scanning

The computational alanine scanning was run on the ROBETTA web server [101,369]. Through defining interface residues as being within a 4Å radius of one or more atoms from the partner

monomer, residues in the respective SRA-ApoL1 interfaces were systematically mutated to alanine. The procedure then used a free energy function to calculate the changes in the binding Gibbs free energy (ddG) between the monomers upon each alanine substitution.

4.3 Results and Discussion

Protein-protein docking was implemented to generate SRA-ApoL1 complexes which were filtered using interface analysis, energy calculations and DRN analysis to isolate energetically feasible complexes. We then introduced ApoL1 mutations into the identified complexes and investigated using DRN and alanine scanning to rationalise complex formation the role key residues play.

4.3.1 Identification of feasible SRA-ApoL1 complexes

4.3.1.1 HADDOCK output

HADDOCK generated clustered datasets of protein structures from each run, to filter for identification of thermos-feasible structures.

Of the 200 water-refined structures generated by HADDOCK from each run (**see *Table 2***), 116 were clustered in 16 clusters during run 1, 131 were in 12 clusters for run 2, and 141 were placed in 13 clusters for run 3, representing 58%, 65%, and 70.5% of the models, respectively. The top ranked dimeric structures were inspected based on the HADDOCK score, in-cluster RMSD, electrostatic energy, Z-Score, and the FCC. The FCC measures the shared intermolecular contacts across the clustered complexes [370], favouring a higher score. Based on the highlighted metrics, a representative cluster was identified per run, resulting in clusters 1, 4 and 9 of runs 1,2 and 3, respectively. Of the final clusters, cluster 1 scored the lowest HADDOCK score at -78.68 ± 5.42 (***Tables S2-4***), while the largest FCC of 0.8 was noted in cluster 9 (***Figures S4-6***), increasing the possibility of either of the runs producing adequate dimeric models. Four structures from each final cluster were selected for further analysis, which entailed MD simulation runs to ascertain the structural stability in solution. Of the twelve structures filtered via MD, five potential structures remained in-complex after 500ns. For simplicity, the remaining five SRA-ApoL1 structures were renamed complexes 1-5 (***Table S 4***).

4.3.1.2 Principal Component Analysis

PCA was used to predict the protein motions in low-dimensional subspaces [237]. The ED was calculated comparatively to investigate the functional dynamics and identify low energy conformations sampled by each distinct SRA-ApoL1 complex highlighted in ***Figure 16*** and ***Figure 17***.

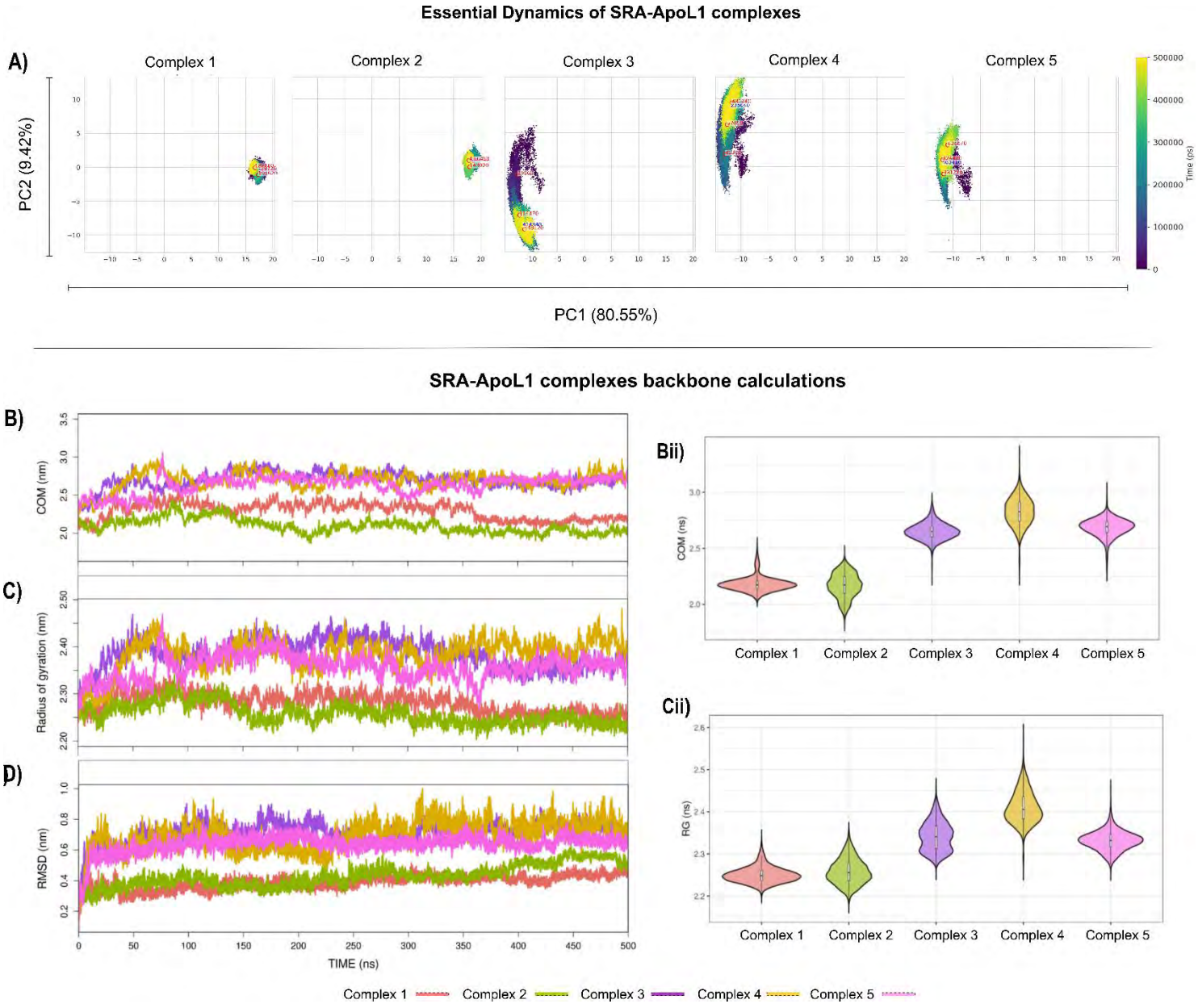


Figure 16: ***A)*** *The principal component analysis (PCA) of the SRA-ApoL1 complexes over the length of the MD simulations. The CoM, RMSD and Rg calculations for the complexes are shown in line* ***(C-E)*** *and violin plots* ***(Cii & Dii)****, with a colour key presented at the bottom of the plots. Correlation is observed between the backbone flexibility calculations and the conformational space sampled in the PCA.*

The majority of variance was captured in the first two principal components, at 89.97% of the total variance over a time frame of 500ns, as shown in ***Figure 16A***. The most significant conformational displacement of the observed PCs was recorded along PC2, which held 9.42% of the variance. Minimal fluctuation was noted in the conformational distribution of complexes 1 and 2 along both PC1 and PC2, suggesting the identification of a stable conformational state. By comparison, larger conformational space was notably sampled along PC2 in complexes 3, 4, and 5 for the 500ns, with significant variance noted along PC2 in complex 3, an observation corroborated in the RMSD, Rg and CoM calculations in ***Figure 16C-D***. Three local minima were identified in each of the five cases and used to determine structures with the lowest energy and highest conformational stability for further static calculations, illustrated in ***Figure 17***. Visual inspection of the minima structures revealed two distinct binding orientations of ApoL1 across the complexes rotated nearly 180° from each other. The contrasting ApoL1 binding conformations were incidentally split between complexes 1 and 2, and complexes 3-5, coinciding with the PCA distributions.

The ApoL1 hairpin of complexes 3-5 faces away from the N-terminal end of SRA ***(Figure 17)***, resulting in conformational mobility of these regions, as seen in ***Figure 16***. This trend was in agreement with normal mode analysis ***(Figure S 17),*** where the largest motions coincided with the loops and shorter helices of both structures across all complexes. We also observed significant displacement in the H3 of SRA and its surrounding loops and in the α-hairpin of ApoL1, consistent with findings of the RMSF analysis and favoured complexes 1 and 2.

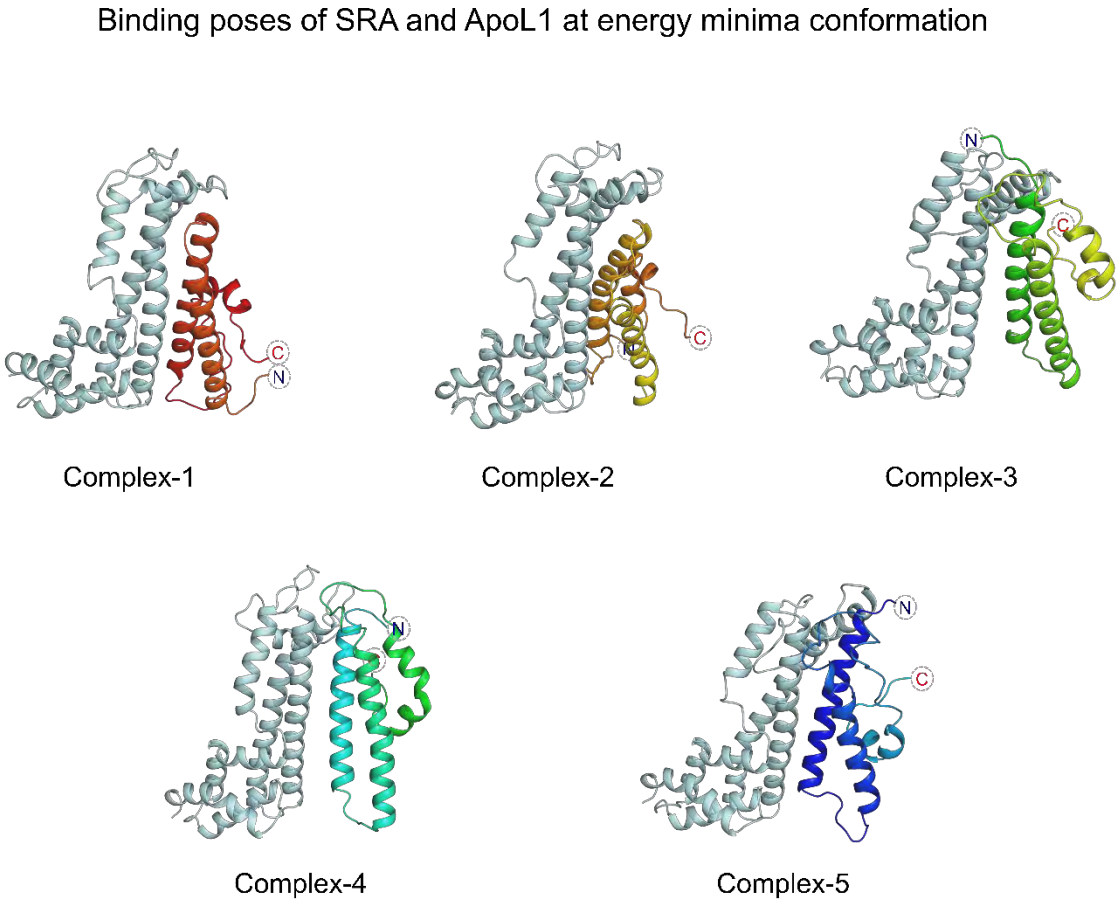


Figure 17: The PCA calculations identify the low energy conformations presented in the figure above , which show the unique binding orientations of the monomers per complex differentiated unique ApoL1 colouring.

(Structure source : HADDOCK protein docking output)

4.3.1.3 Energy analyses

The five distinct minima structures obtained through ED analysis (***Figure 16B***) were evaluated and ranked based on the solvation/dissociation energy, interaction energy, and buried surface area (BSA) along the dimerisation interface (***Table 3***). As a rule of thumb, dissociation energies greater than zero and low interaction energies complement complex formation [368,371]. Interface participation in complex formation was further scored based on the complexation significance score (CSS). Complex 1 and complex 2 notably scored more favourably than their counterpart complexes 3, 4 and 5, maintaining a trend observed in the analysis of the ED (see ***Figure 16***). The positive dissociation energies of 7.7 kcal mol^{-1} and 10.2 kcal mol^{-1}, in combination with low interaction energies of -17.4 kcal mol^{-1} and -19.9 kcal mol^{-1}, observed in

complexes 1 and 2, respectively **(*Table 3*)**, were indicative of the stability of the complexes in solution, and by extension physiological relevance.

Table 3: Probable stability of SRA-ApoL1 complexes in solution as predicted by the PDBePISA server. ΔGint indicates the solvation free energy gained upon formation of the complex, while ΔGdiss indicates the free energy of complex dissociation. Complexes with ΔGdiss>0 are likely to be thermodynamically stable in solution

Complex	*Interaction Energy ΔG^{int} (kcalmol^{-1})*	*Dissociation Energy ΔGdiss (kcalmol^{-1})*	*BSA (Å^2)*	*Interface (Å^2)*	*Complex Significance Score*	*Stability In Solution*
Complex-1	-17.4	7.7	2260.0	1129.6	1	yes
Complex-2	-19.8	10.2	2280.0	1138.3	1	yes
Complex-3	-6.5	-2.7	1467.0	733.5	0	low
Complex-4	-2.3	-5.6	1129.5	564.7	0	low
Complex-5	-6.1	-2.4	1574.7	787.3	0	low

Each dimer-forming interface buried over 1000Å² (1129.6Å² and 1138.3Å², respectively, of the 2200Å² buried surface area (***Table S12***), stabilised by hydrogen bonding and salt bridges (***Table S14***). Consistent with these observations, the CSS (scored from 0.00 to 1.00) was determined to be 1.00 for both interfaces. By contrast, CSS values of complexes 3, 4 and 5 were scored at 0.00, which was supported by the smaller dimer-interface surface areas of 735.5Å², 564.7Å², and 787.3Å² noted for the complexes, respectively. The interface areas matched the below 2000Å² BSAs determined in the three complexes. The negative dissociation energies also contributed to the unlikelihood of thermostability for the three complexes. At the lowest, complex 4 scored an interaction energy of -2.3 kcal mol^{-1} and low dissociation energy of -5.6 kcal mol^{-1}, consistent with the conclusion that the dimers would be unstable in solution.

4.3.1.4 Dynamic Residue Networks analysis

The use of DRN analysis allowed for mapping communication pathways throughout the protein complexes distinct to each metric and protein system.

A common communication pathway from H6 (Leu197) flowing into H2 of SRA was highlighted by the concentration of high *BC* values along helices H1 and H2, with H5 and H6 in ***Figure 18***. Depending on how the protein was oriented in reference to SRA, the communication channel of ApoL1 was restricted to one helix on either side of the α-hairpin. Complex 4 offered an exception to this rule, as the path diverted through the involvement of 346, positioned in the next helix **(*Figure 19***).

Persistent hubs [325] , as established in Chapter 3 (see ***0*)**, were observed in SRA across the five complex conformations in ***Figure 19***, specifically Ala105, identified as a crucial hub in **Chapter 3**, and Leu197, which correspond to the *BC* consistency of SRA. Complexes 1 and 2 were found to have a noticeable divergence in hub distribution from their last three counterparts. Except for Val346 from complex 4, most complex 3-5 *BC* hubs were missing from complexes 1-2 in ApoL1, where this trend was particularly pronounced. This absence of shared hubs in ApoL1 was probably influenced by different protein orientations in the individual complexes.

When complex 1 and 2 hubs were mapped onto the energy minima conformations, they were distinguished by probable communication across the interface (***Figure 18***). Overall, the *BC* analysis identified residues that bridged the protein binding interfaces, indicating residues crucial to the stability of the SRA-ApoL1 interaction.

The *CC* analysis revealed a non-linear communication path that traversed through the cores of both SRA and ApoL1. Accordingly, SRA with high *CC* residues were in H1 and H2, with Ile62, Gln64, and Val65 serving as the only *persistent hubs*. Shared communication hubs were absent in ApoL1, as seen in the *BC* analysis **(*Figure 19***). The high *CC* path seen in ApoL1 was isolated to a single helix in complexes 3-5 but traced both adjacent hairpin helices in complexes 1 and 2.

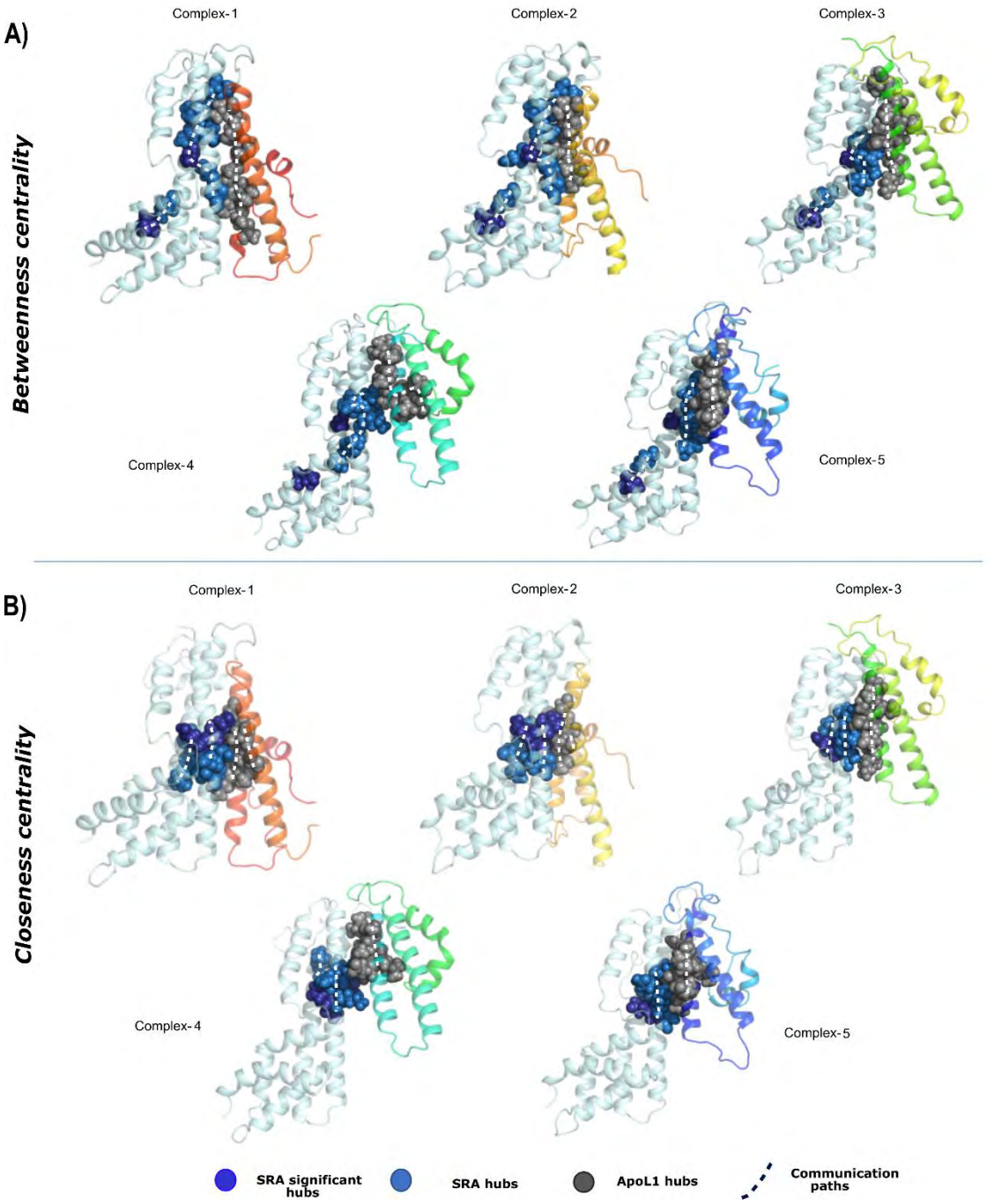


*Figure 18: High **A)** BC and **B)** CC residues mapped onto energetically stable minima structures for five complexes. Paths of communication (in white) can be traced along the high centrality scoring hub residues in grey for ApoL1 and in blue for SRA, showing top 10% and top 5% residues, respectively. The persistent hubs in SRA are denoted as "significant hubs". (Structure source : HADDOCK protein docking output)*

CC hubs in complexes 1 and 2 notably scored higher, while visibly populating the interface ***(Figure 19)*** key to protein complex stability. The positioning of the high *CC* residues such as

Ile62 and Val65, or Lys349 and Leu352, was indicative of cross-interface communication, as validated by contact map analysis in ***Figure S 15***.

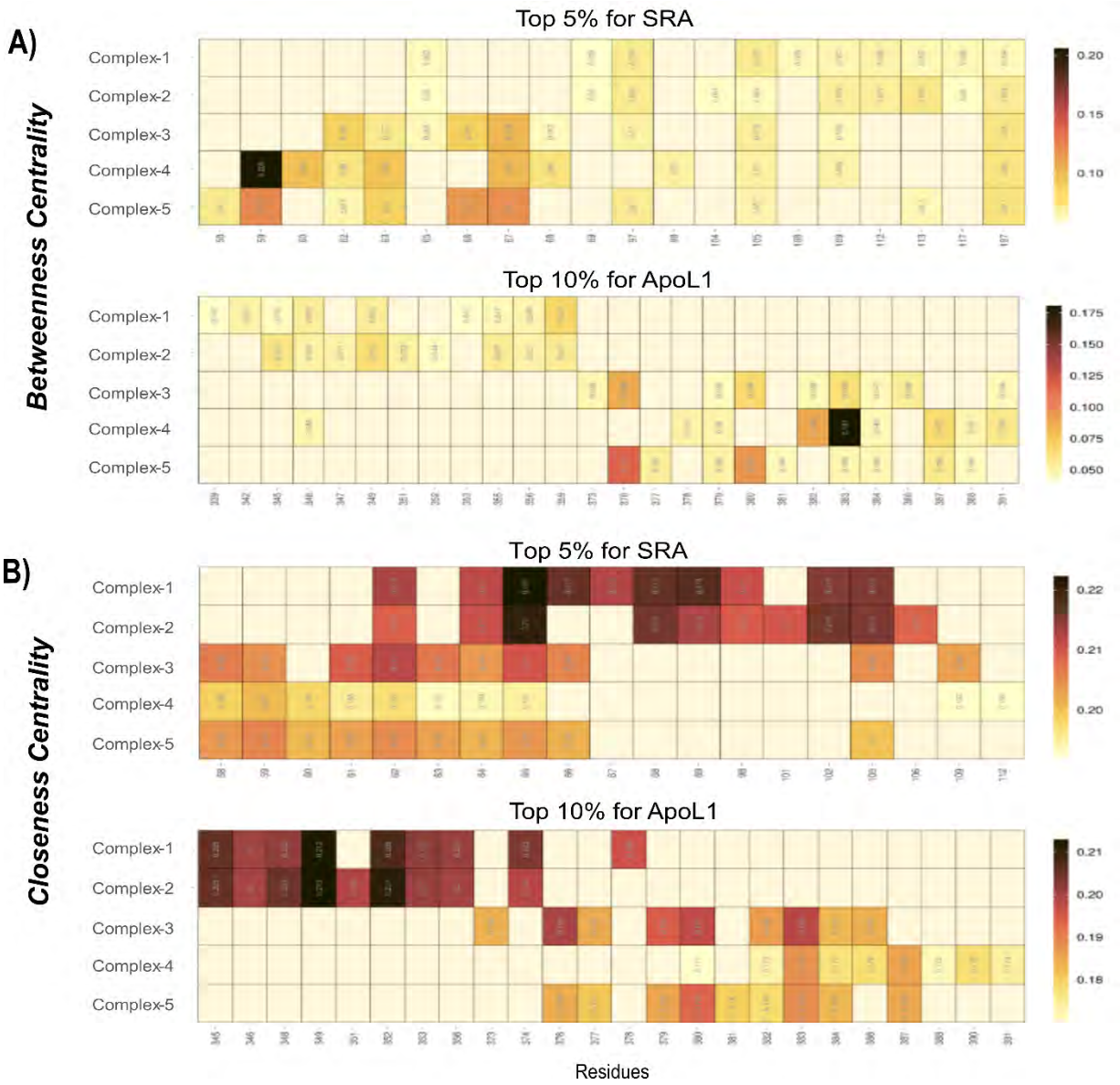


*Figure 19: Top 5% and 10 % **A)** BC and **B)** CC hubs detected for SRA and ApoL1 are denoted by the annotation in their heatmap cells compared to their residue counterparts. Persisents hubs were those observed in all complexes per metric. Low to high centrality values for each metric correspond to the intensity of the colour bar on the right of the each heatmap.*

Additionally, key monomer *CC* hubs such as Val98 and Ala105 in SRA (***Chapter 3***) and Leu346 in ApoL1 (***Chapter 2***) were maintained in both complexes, lending further confidence to the viability of the aforementioned complexes. In contrast, peripheral residues Cys180 and

Val187 were relegated from *CC persistent hub* status, as their access to neighbouring residues decreases in the complex compared to the monomer in ***Chapter 2***

4.3.2 Effect of ApoL1 variants on the complex

The effect of the ApoL1 variants on the complex integrity of the SRA-ApoL1 complexes was investigated through DRN, alanine scanning and contact map analysis.

4.3.2.1 Dynamic Residue Networks

Further DRN analysis of the mutants allowed for the identification of significant complexation residues in ApoL1.

The residues found included Val346, Val349 (*persistent hubs*) and Ser356 using *BC*, and Leu345, Asp348, Val349, Leu352, Leu353, Ser356 and Val374 through *CC*, all validated in duplicated runs (***Figure S 21***). While introducing the ApoL1 variants resulted in increased conformational mobility of the complexes, DRN analysis revealed a retainment of the communication pattern between SRA and each iteration of ApoL1 (see ***Figure S 20)***. Though some change in network contacts was expected due to the steric differences of Gly342 and Met384 substitutions, the established *BC* and *CC* communication paths were not adversely affected, reinforcing the significance of these communication hubs to maintaining the SRA-ApoL1 complex.

Noteworthy residues in this regard were Ile97, Ala105 and Leu197 in SRA, as observed in the wild-type *BC* analyses, which retained the short path within the monomer, even in both an ApoL1 bound and unbound state (see ***0***). An additional hub noted in the SRA-ApoL1 complex is Ala113 which is promoted to *persistent hub* state in the complex located in the binding interface.

In *CC* calculations, *persistent hubs* included Val65, Ala68, Val102, Ala105, Asp348, Val349, Leu352, Leu353, and Val374 as noted in ***Figure 20B***. The residues were located in the interface of the complex, thus attributing significantly to the stability of the complex.

G1M appeared to have the most pronounced effect on ApoL1 communication. In *BC*, a lone loss of significance for residues Ser356 and Leu359, and a gain of Glu377, was noted (***Figure 20A***). This resulted in a more disrupted communication path compared to the other complexes. In *CC* analysis, more disparity was observed with G1M as well as with the reduction in *CC* for H1 residues Ile62 and Thr69. Additional hubs triggered by the presence of mutated Ser342 and

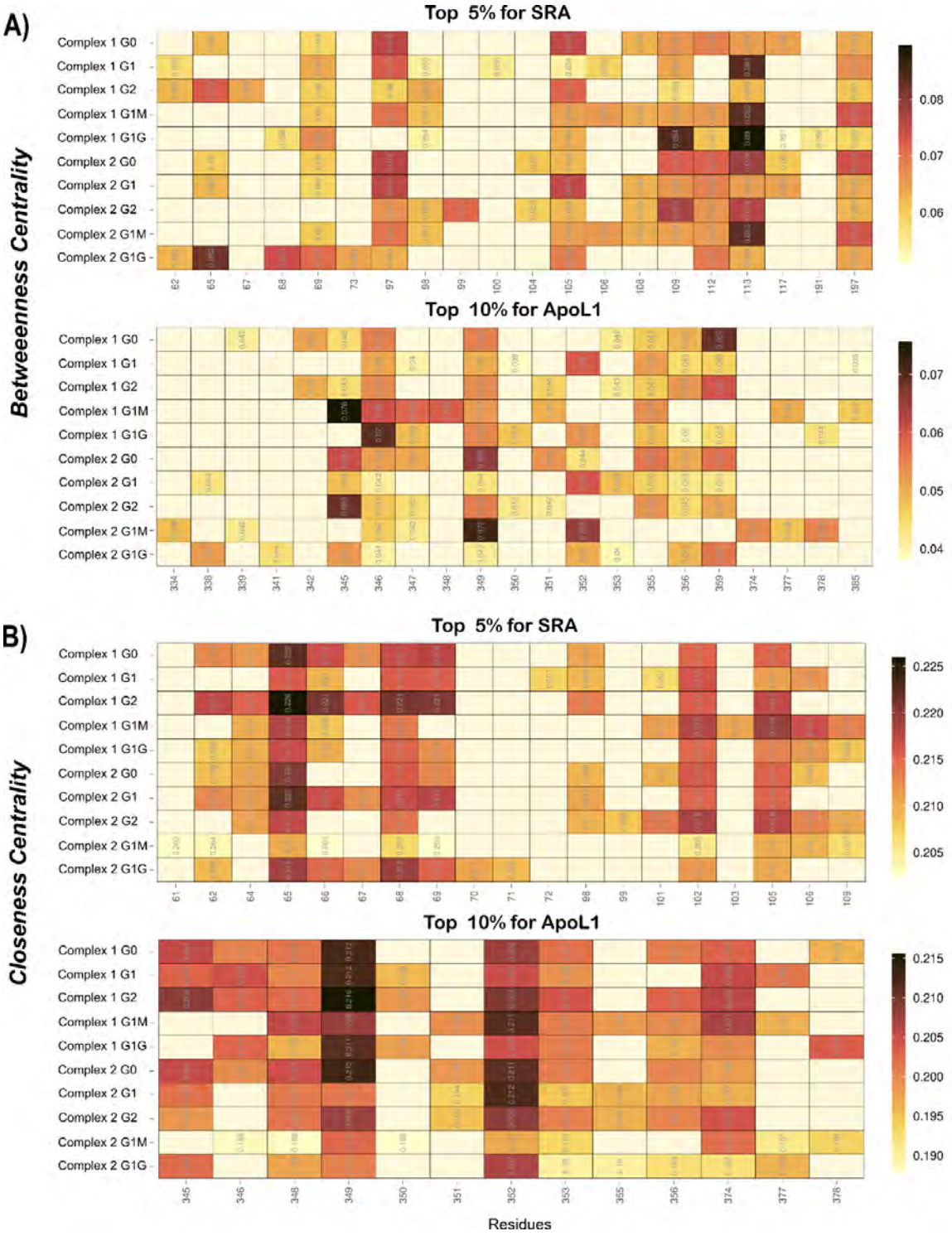


Figure 20: ***A)*** *BC and* ***B)*** *CC hubs observed in the SRA-ApoL1 complexes over a stabilised section of the MD simulations. Detected hubs within the cut-off % were annotated with their centrality values for comparison against counterpart residues in other variants. Low to high centrality values for each metric correspond to the intensity of the colour bar on the right. Key residues were noted to maintain hub status in all the variant complexes for both metrics*

Ile384 included Thr69 for *BC*. *BC* hubs gained solely in the variant complexes included Ala68 (in G1G) and Val98, and ApoL1's Val350, Tyr351, Glu355 and Glu377, absent in the ApoL1G0 *BC* pathways. In contrast, Val98 is gained as a *CC* hub only in the G1 and G0 complexes. Due to the steric differences of Gly342 and Met384 substitutions, a reduction in some network contacts was expected. However, this is outweighed by the pathways maintained regardless of the variation, as these appear significant to the retainment of the complexing across the interface.

4.3.2.2 Alanine scanning

Alanine scanning determined residues with a destabilising effect on the SRA-ApoL1 complexes. Specific residues were common in both the wild-type and the G1 mutation complexes.

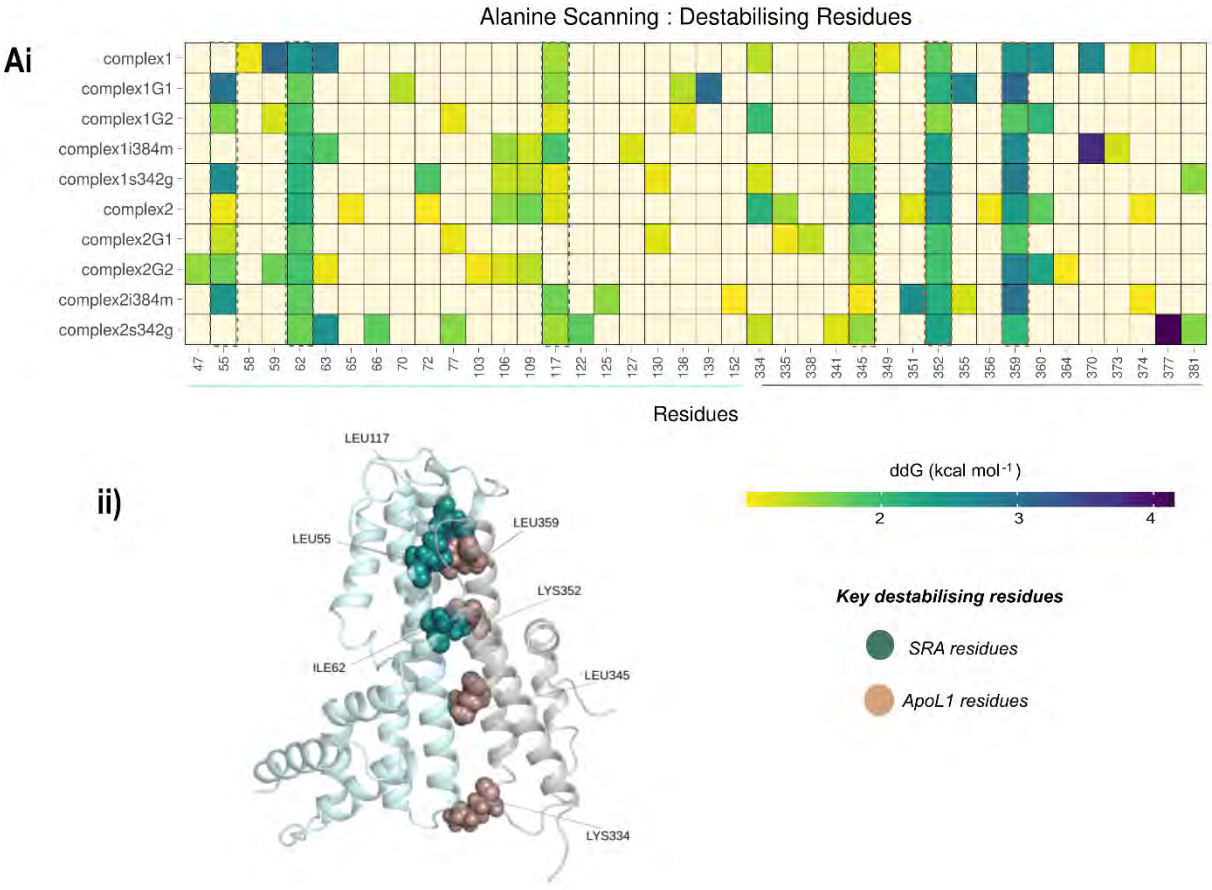


Figure 21: Alanine scanning highlighted destabilising residues that were common across all the SRA-ApoL1 variant complexes ***(i)****. The residues were found to be in the binding interface, as illustrated in* ***(ii)****. Destabilising alanine substitutions had free energy values (ddG) ≥1 kcal mol-1,*

while neutral values ranged from -0.8 to 0.99 kcal mol^{-1}, which is represented on the colour bar. (Structure source : HADDOCK protein docking output)

The destabilising residues were especially noted in the interaction interface between the two proteins, with the commonality in the identified residues suggesting a similar binding pattern of the ApoL1 G1 variants to SRA. Agreement was observed between the destabilising and high centrality residues, especially in ApoL1. Residue Ile62, located in the N-terminal H1 of SRA, is destabilising across all complexes, with a notable effect of Leu117. Leucine residues 345, 352 and 359 are chief destabilising agents in ApoL1, with Leu345 and Leu352 recording significant centrality status in both *BC* and *CC* (***Figure 21*** **&** ***Table 4***).

Table 4: Destabilising protein-protein interface residues identified through in silico alanine scanning for each dimeric complex and their G1, G1-M, G1-G and G2 variants. BC hubs are in bold, and CC hubs are italicised.

Dimeric structure	*Destabilising residues*	
	SRA	ApoL1
Complex-1	58, 59, *62*, 63, **117**	334, ***345***, ***349***, *352*, **359**, 360, 370, *374*
Complex-1 G1	55, **62**, 70, 117, 136, 139	*345*, ***352***, **355**, **359**
Complex-1 G1^{G}	55, 62, 72, ***106***, ***109***, 117, 130	334, **345**, ***352***, 359, *381*
Complex-1 G1^{M}	*62*, 63, *106*, ***109***, **117**, 127	345, ***352***, **359**, 370, 373
Complex-1 G2	55, 59, 62, 77, 117, 136	334, 345, 351, 352, 359, 360
Complex-2	55, *62*, ***65***, ***72***, *106*, **109**, **117**	334, 335, ***345***, ***351***, ***352***, ***356***, **359**, 360, *374*
Complex-2 G1	55, *62*, 77, 130	335, **338**, ***345***, ***352***, **359**
Complex-2 G1^{G}	*62*, 63, *66*, 77, 117, 122	**334**, **341**, **345**, ***352***, 359, ***377***, 381
Complex-2 G1^{M}	55, ***62***, 117, 125, 152	***345***, 351, ***352***, 355, **359**, *374*
Complex-2 G2	47, 55, 59, 62, 63, 103, 106, 109	345

Due to the helical nature of both SRA and ApoL1, leucine residues play a significant role in the structural stability of helices [372], which in turn affects the nature of the inter-protein helical interactions. In ApoL1, the leucine residues align along the outside of the helix, most exposed to binding with SRA, supporting a concept similar to that defining the LZD heptad (residues Leu371, Leu378, Leu385 and Leu392) cited to facilitate the SRA-ApoL1 coiled-coil binding [280].

Other centrality hubs with destabilising potential included Leu109, Lys334, Val374, and Val106. Overall, communication hubs identified through network analysis were corroborated key residues identified in alanine scanning, highlighting their importance to the integrity of the SRA-ApoL1 complexes.

4.2.2.1 Contact map analysis

Contact map analysis further enlightened interactions between the two proteins, as an expected shift in the residue networks was noted due to the presence of ApoL1 variants. Inter-protein interaction was chiefly observed between 72-99 (H1 and H2) of SRA in the G1G mutant and SRA. All variants with the mutated Gly342 (G1G and G1) resulted in a general reduction in contact with the linkage disequilibrium partner Ile/Met384 as opposed to the wild-type Ser342 as noted in ***Figure 22.***

This effect on the residue network is highlighted in the *BC* and *CC* analyses. The two residues are located in adjacent helices of the ApoL1 C-terminal, likely affording the section some stability. Decreased contact between residues 384 and 342 probably affects helix stability, which is enhanced in the presence of the G2 mutation resulting in the dissociation of the complex. G1-G2 contact was observed as Asn388 (part of the deletion mutation G2) interacted with both Ile/Met384. The residue also interacted with H1 residues of SRA, which is not observed with counterpart Tyr389.

Further contact was noted between Ala73 (and Thr76) with primarily the wild-type complexes and some of the Ser and Gly variants. Interaction with Pro344 and Phe344 is gained in the presence of Gly342 in both G1 and G1G complexes (as opposed to Ser342). It is also worth acknowledging that G1G infrequently occurs in the absence of the G1M mutation [285,287,373], while SRA-ApoL1 dissociation is primarily attributed to the presence of the G2 deletion

mutations. In merit to these findings, we noted minimal difference in the communication networks between the G1G and G1M ApoL1 variations.

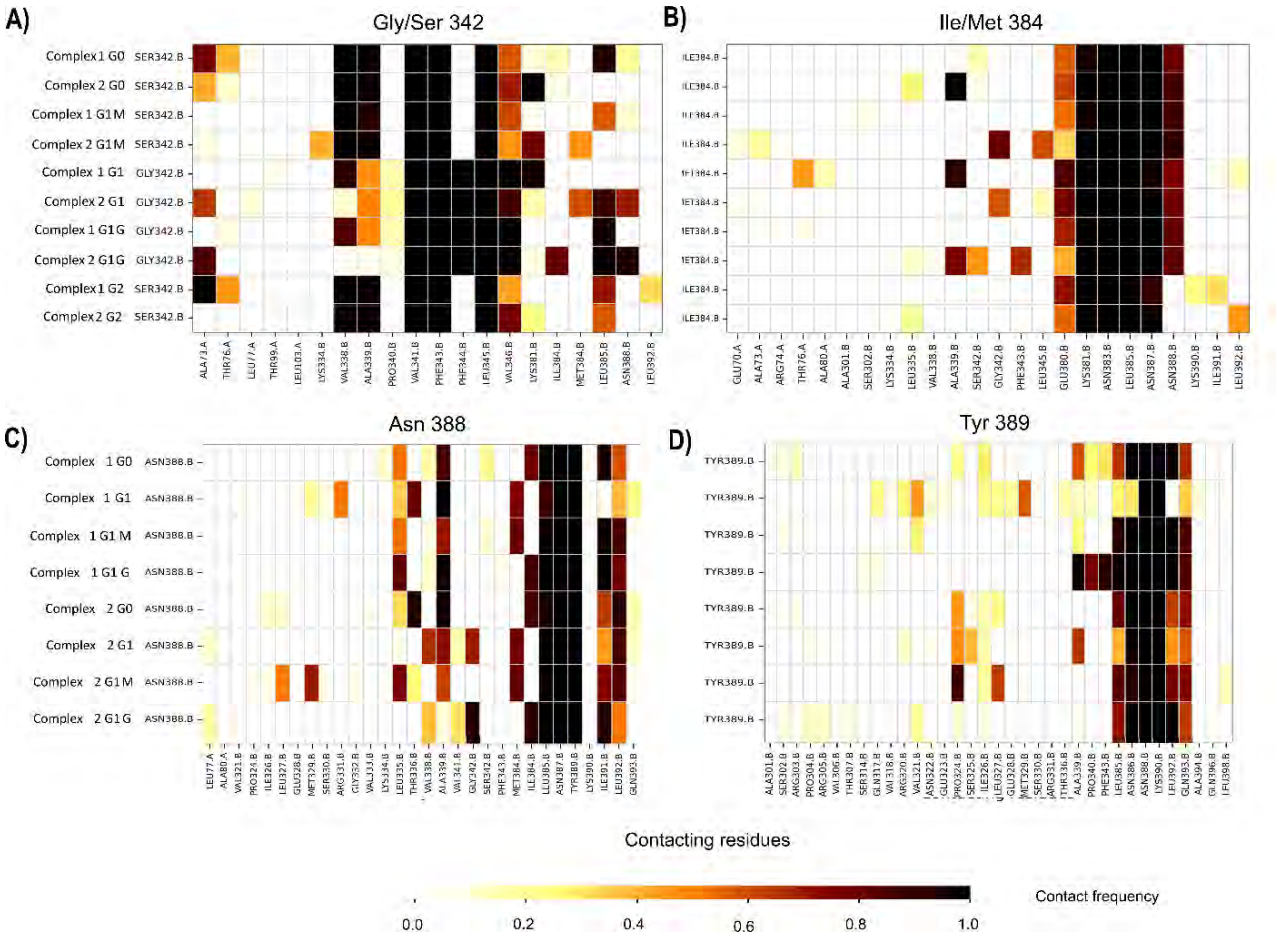


*Figure 22: Contact heatmaps for ApoL1 mutation residues **A)** Ser/Gly342, **B)** Ile/Met384, **C)** Asn388, and **D)** Tyr389 in the SRA-ApoL1 wildtype complexes. The inter-residue contact frequency was tracked over 500ns MD simulations and is denoted by the colour bar at the bottom of the figure. Complexes 1 and 2 G2 contacts are absent in C and D due to the deletion of Asn388 and Tyr389 in both.*

4.4 Conclusion

We see that binding affects the path of communication with the binding sites. As observed in ***Chapter 2***, the ApoL1 variants have an intramolecular effect on the stability of the protein.

This study aimed to further identify and elucidate the nature of SRA-ApoL1 binding, highlighting three main observations. Firstly, feasible SRA-ApoL1 complexes were determined based on global and local structural analyses. Our findings corresponded with the

external thermostability, and complexation analysis used. Additionally, the orientation of both proteins in the isolated complexes was practicable, dismissing the SRA's H3 obstructing binding [339]. Secondly, we identified residues that are credibly key to the complexation of SRA-ApoL1 and the maintenance of dimer integrity. Consistency was observed between the sites of complex destabilisation and the primary communication residues as isolated by DRN analysis.

Furthermore, identified key residues supported the theory of LZD facilitating the ApoL1 binding as opposed to a dimerisation along the entire interface [312]. Leucine residues appeared to play a significant role in the maintenance of both the individual structures as well as the SRA-ApoL1 complexes. Lastly, we noted the minimal dissociative role of the G1 mutations in the complex. This observation was consistent with wet-lab studies that highlighted the necessity of G2 for dissociation experiments [291].

Additionally, computational studies have shown the more significant effect of the G2 deletion on the structural stability of ApoL1 [304]. It is important to note that the study was conducted using a segment of ApoL1. Use of the complete protein would provide further insight into SRA-ApoL1 complexation and ApoL1 variants.

Chapter 5

Anti-folate Drug Discovery

5.1 Introduction

Apart from the parasites' efforts to evade host immunity, most drugs in chemotherapy for HAT have numerous drawbacks [374]. Drugs for the treatment of first-stage HAT have adverse effects after prolonged use [375,376], while those used for the treatment of stage two *Tbr* infection – melarsoprol – can induce encephalopathy in patients, with a death rate of 5% [377]. The treatment for the *Tbg* meningo-encephalic infection is expensive and challenging to administer [42]. Furthermore, the available drugs are plagued with resistance [64,378,379] and inefficacy [380,381]. Undoubtedly, new and better drugs to treat HAT are vital [382] to a continued decrease in disease prevalence.

An attractive target in anti-parasite and anti-microbial chemotherapy is the folate metabolic pathway [383,384]. Trypanosomes are pteridine auxotrophs, requiring folates exclusively from host organisms [385,386] for growth and survival. Reduced folates are essential for thymine and, consequently, DNA synthesis, such that a lack of the co-factors results in cell death [387]. As a result, the parasite folate pathway is an objective target for anti-trypanosome chemotherapies[388]. Anti-folate therapy has already been used to target plasmodia successfully in malaria treatment [383,389]. Chemically validated drug targets involved in folate and pterin salvage include pteridine reductase 1 (PTR1) [390] and dihydrofolate reductase (DHFR) [388,391,392]. The proteins act by reducing the co-factors folate and biopterin with NAD or NADP in eukaryotes to their tetrahydro-forms in two successive reactions. Through a Lys-Asp-Tyr triad, PTR1 and DHFR participate in a hydride and proton transfer required for substrate reduction [393]. In its reduced form, biopterin plays a vital role in folate biosynthesis and defending against oxidative stress [394]. DHFR is currently targeted using the drug methotrexate (MTX) [395]. DHFR-targeted therapy alone, unfortunately, does not work due to the presence of PTR1, shown in ***Figure 23B***, a tetrameric short-chain dehydrogenase (SDR) responsible for 10% of the tetrahydrofolate production [388,396]. While this substrate turnover is low, inhibition of DHFR triggers PTR1 up-regulation, resulting in parasite survival [396]. As such, both targets need to be

inhibited using synergistic therapy [397,398] for effective treatment. As DHFR and PTR1 have similar substrates (***Figure 23***), it is plausible that inhibitors designed for one protein may also modulate the other.

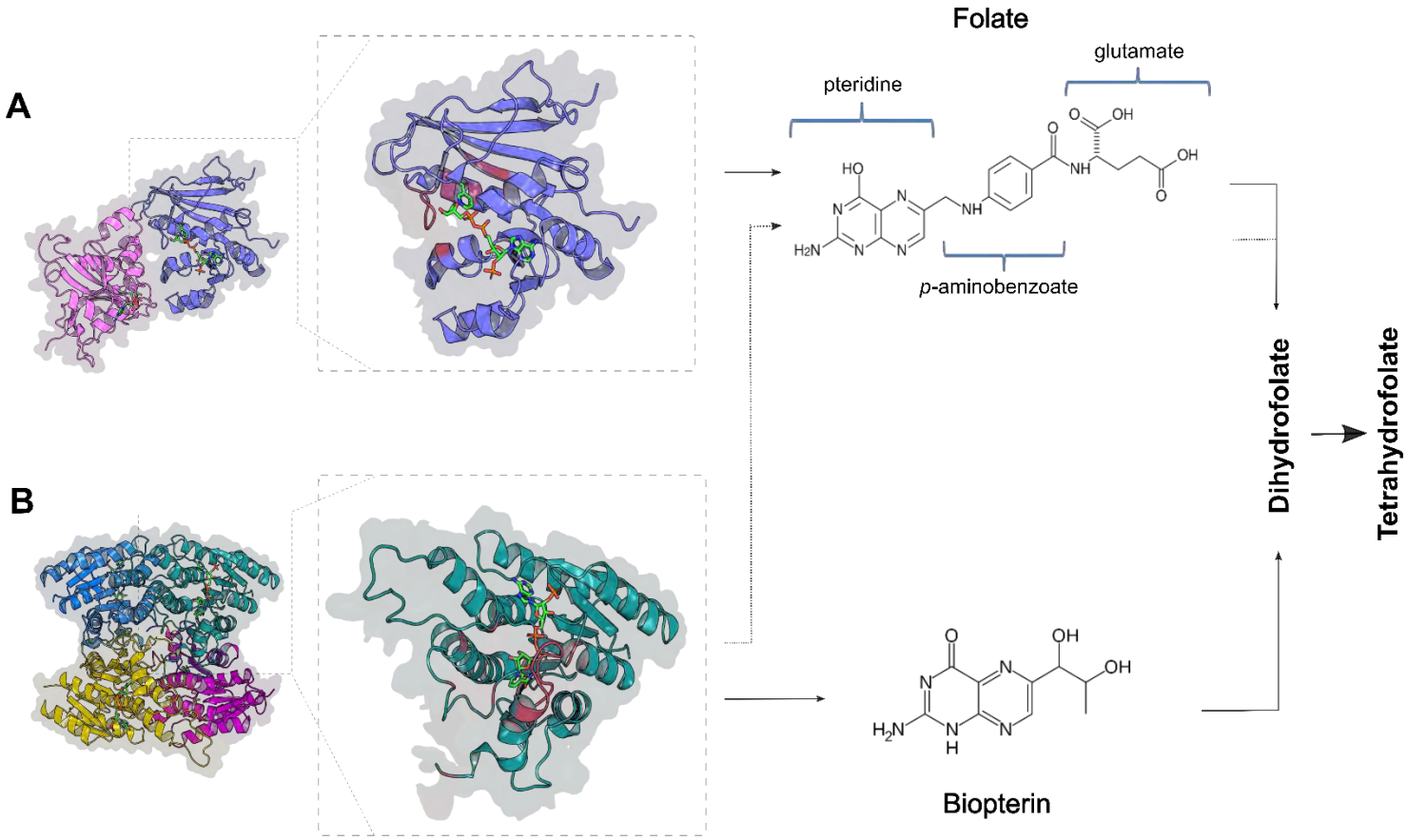


Figure 23: The structures of ***a)*** *TbDHFR (PDB ID:3qfx) and* ***b)*** *TbPTR1 (PDB ID:2x9n) . Residues active in the substrate binding and reduction reactions are shown in maroon, as well as the substrate-binding loop of PTR1. The NADPH co-factor (shown in stick figure representation), crucial to the catalytic activity, is shown in green, blue, and orange. The native substrates of both proteins are shown as 2D stick figures, highlighting their structural similarity, pathway, and the end product of the reductions.*

A key aim in drug discovery is the rapid identification of pharmacologically active compounds with minimal side effects. Drug repurposing can partially address this aim which entails reusing approved drugs for alternative medical purposes [399]. The re-profiling has the benefit of bypassing initial research on drug pharmacology, formulation, safety, and adverse effects, as well as FDA approval [400], to skip right to clinical and efficacy trials. This decreases development time [401,402] and results in higher success rates than original drugs owing to their

prior documentation [400]. In cases where identified hit drugs are not adapted for therapy, there is an opportunity to attain valuable data for scaffolding and lead optimisation. A long-standing example of successful drug repurposing is thalidomide [403], an anti-emetic drug with detrimental teratogenic side effects [404,405] which is now safely used in treating myeloma and leprosy. More recently, the drug niclosamide has shown activity to combat Zika virus [406–408], among many other conditions [409]. More relevantly, the repurposed anti-tumour agent eflornithine [410] is currently used for treating second-stage *Tbg* HAT. Unfortunately, its use is accompanied by adverse reactions characteristic of cytotoxic drugs used in cancer treatment [62,410,411]. Nonetheless, the use of eflornithine shows both the potential of drug repurposing as a strategy in HAT drug development and the need to develop less toxic alternatives for chemotherapy. While earlier examples of drug repurposing were either from accidental discoveries or based on prior knowledge of drug pharmacology being applied to a new pathology [412], computational methods can now be used to expand this scope using a larger chemical space [413,414]. In essence, computational drug discovery has drastically reduced the cost [415] and time [416] required for drug discovery. High throughput virtual screening has been applied to complement *in vitro* experimental assays [417] by narrowing down the number of drug candidates to be evaluated [418,419]. Aside from lead identification, *in silico* methods have also been used to predict and elucidate drug mechanisms and interactions with reasonable accuracy by applying biomolecular simulations [420,421]. Several compounds that have entered late-stage clinical trials and even therapeutic approval [422] are a consequence of integrating computational approaches into the drug discovery pipeline, including HIV-1 protease inhibitors [423]. In this chapter, we employed HTVS using a subset of approved compounds from DrugBank to be repurposed as potential anti-folate-based HAT chemotherapeutics. Further evaluations of the effectiveness of the top thirteen hit compounds targeting both DHFR and PTR1 were carried out by describing the molecular dynamics and energetics of the ligand-bound complexes.

5.2 Methodology

5.2.1 Molecular Docking

5.2.1.1 Ligand Preparation

Approved drugs were retrieved from the DrugBank database [168]. After examination of the data set, entries with compound mixtures were separated. Additionally, in order to perform virtual

screening using the AutoDock Vina [424] potential from QuickVina-W [425] any halogen or metal-bound complexes were removed from the docking set. The RDKit library [426] was used, by Rita Afriyie Boateng and Olivier Sheik Amamuddy, to build fully protonated 3D structures for the resulting 2089 compounds, using their corresponding SMILES strings. Gasteiger charges were assigned, and non-polar hydrogen atoms merged using MGLTools Python script *prepare_ligand4.py* from AutoDockTools (v1.5.6) [427,428].

5.2.1.2 Receptor Preparation

The crystal structure of *Tb*PTR1 (PDB ID:2X9N) [429] with missing residues modelled in was obtained from previous work done by K. Magambo [430], while that of *Tb*DHFR (PDB ID:3QFX) [431] was retrieved from the RCSB PDB. The proteins were protonated at pH 6 using the PlayMolecule ProteinPrepare tool [432], which allowed for the presence of the co-factor in the calculation of pKa values. MGLtools *prepare_receptor4.py* scripts were then employed to prepare both proteins, including the NADPH co-factors, for molecular docking. The scripts applied AutoDock-compatible atom types through the addition of Gasteiger charges to the protein.

5.2.1.3 Docking Experiment

Blind docking was conducted on all systems using docking grid dimensions and grid centre Cartesian coordinates (***Table S 9***) calculated using BIOVIA DS [132]. These parameters were validated by re-docking the co-crystallised inhibitor cyromazine (AX3) into *Tb*PTR1. The procedure was parallelised using GNU parallel [433]. Each docking run generated nine protein-ligand complexes, ranked according to their Vina binding scores. The docking procedure was repeated in *Tb*DHFR after potential hit compounds were identified. *Tb*DHFR-specific parameters (***Table S 9***) were used and validated by re-docking pyrimethamine (CP6) into 3QFX.

5.2.1.4 Binding Pose Assessment

An in-house Python script was developed, by Olivier Sheik Amamuddy, to facilitate the visualisation and analysis of the ligand binding locations in the multimeric protein after blind docking. In order to do so, each ligand from a highly exhaustive search was assigned to the protein chain closest to it before aligning the complexes to a common reference chain. This enabled the calculation of ligand pose clustering based on different distance metrics, such as

the ligand centroid, the RMSD, and the CoM distance. After calculating the distances between each pair of ligand poses, hierarchical clustering was applied, and a cut-off distance was selected. In addition to consideration of hierarchical clustering, Euclidean distance in relation to all four active sites of the tetramer and Vina scored binding energy were used to screen ligand binding poses. An energy range of less than -8kcal mol^{-1} and a clustering percentage greater than 80% were considered for filtering. Additionally, a visual assessment of the poses and intermolecular interactions was conducted using LigPlot+ [158,159], BIOVIA DS [132] and Schrödinger Suite Maestro [238] to attain a consensus. Favourable interactions such as hydrogen bonds, hydrophobic (π-stacking), and electrostatic interactions were especially noted.

5.2.2 Molecular Dynamics

5.2.2.1 Simulations

GROMACS v2018.6 [203,348] was employed for the all-atom molecular dynamics simulations on the eight favourable protein-ligand complexes. The AMBER03 force field [211] was used to parametrise the co-factors and ligands, and proteins through ACPYPE [222] and GROMACS, respectively. Thirteen complexes were solvated in a cubic box using the SPC water model, neutralised in 0.15 mM NaCl and subsequently energy-minimised terminating at an F_{max} <1000 $kJmol^{-1}nm^{-1}$ and maximum steps of 50000 using the steepest descent algorithm. The converged protein, ligand and co-factors were coupled, and the systems were equilibrated to a temperature of 300K and pressure of 1 bar using the NVT using Berendsen temperature coupling and NPT ensembles, respectively. Production runs of 200ns were conducted, at a time step of 2fs, applying the PME method [204] for long-range electrostatic interactions and LINCS [206] for bond length constraint. Short-range non-bonded interactions were defined based on the Verlet algorithm [207], using a cut-off of 1.4nm.

5.2.2.2 Trajectory Analysis

Using the *gmx trjconv* module, the trajectories were first corrected for the PBC. The trajectories were additionally fitted by the Cα atoms of the starting structures to eliminate the rotation and translation of whole systems before analyses. The *gmx make_ndx* was applied to isolate each of the components in the system to investigate their conformational properties. To assess protein and ligand stabilities, the RMSD of protein backbone atoms and heavy atoms of the ligands in relation to the protein was calculated. Ligand RMSD was mainly used to filter out unsatisfactory ligand binding conformations. The effect of ligand presence on the system

residue flexibility was evaluated using RMSF of the backbone atoms based on residue contribution. Lastly, a mass-weighted Rg calculation was also employed to investigate the compactness of the protein. RStudio [367] was used for value plotting and any statistical calculations analyses.

5.2.2.3 Interactions Assessment

The ligand-protein interactions profile from an energy minimum state of each simulation run was analysed as before the MD simulations, with the additional use of Arpeggio [160] for a 3-dimensional assessment. Furthermore, VMD [239] was used to visualise the movements of both ligand and protein throughout the trajectory. The frequency of hydrogen bonds between the protein and the ligands was analysed using a setting of 3.5Å and < 30º in CPPTRAJ [349] . The analysis aimed to identify hydrogen donors and acceptors and track and quantify the number of hydrogen bonds per frame throughout the trajectory. The tool was additionally used to calculate hydrogen bond occupancy, which measures the number of frames in which hydrogen bonds are maintained as a percentage of the entire simulation.

5.2.2.4 Blood-Brain barrier Filtering

In order to target the meningo-encephalitic phase of the disease, drugs must be able to cross the BBB. For this reason, hit ligands from all thirteen systems were filtered using the BBB Prediction Server by CBligand [413]. The tool applies the support vector machine (SVM) and LiCABEDS algorithms to predict permeability.

5.2.2.5 Binding Free Energy

The Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) approach was used to calculate the binding free energy of the complexes, using the *g_mmpbsa* [434]. During the equilibrium phase of the simulations, the calculations were run on a 10ns segment by taking samples at 50ps intervals. Here the binding free energy of the ligand-bound complex is estimated using the equations:

$$\Delta G_{bind} = G_{complex} - \left(G_{protein} + G_{ligand}\right)$$

$$\Delta G_x = E_{mm} - (TS - G_{solv})$$

Equation 12: Calculation of MM-PBSA (i)

The free energy of the complexes in solvent *(G_{bind})* results from the isolated free energy values of the complex $G_{complex}$, the protein $G_{protein}$ and the ligand G_{ligand}. The Gibbs free energy *(G_x)* of each of these elements is calculated by combining the molecular mechanics potential in a vacuum *(E_{mm})*, the desolvation energy *(G_{solv})* and the entropic contribution *(TS)* (where *T* is the temperature and *S* is entropy). The E_{mm} encompasses bonded energies as well as van der Waals and electrostatic interactions, while G_{solv} can be further broken down into G_{polar} and G_{apolar}.

$$E_{mm} = E_{bonded} - (E_{elec} + E_{vdW})$$

$$G_{solv} = G_{polar} + G_{apolar}$$

Equation 13: Calculation of MM-PBSA (ii)

Furthermore, a decomposition of the energy contribution of each residue, notably those binding to the ligand, provided further insight into the protein-ligand interactions.

5.2.3 *Trypanosoma brucei* experimental assays

These experiments were performed by our collaborators Prof Heinrich Hoppe and the team of the Centre for Chemico- and Biomedicinal Research [435].

5.2.3.1 Single concentration screen

To assess anti-trypanocidal activity, the identified compounds were added to *in vitro* cultures of *T.b.brucei* in 96-well plates at a fixed concentration of 20µM for pure compounds. The number of parasites surviving drug exposure was determined by adding a resazurin-based reagent after 48 hours. Living cells reduce resazurin to resorufin, a fluorophore (Exc_{560}/Em_{590}) that can be quantified in a multi-well fluorescence plate reader.

Using pentamidine, an existing drug treatment for trypanosomiasis, as a positive control drug standard, the compounds were tested in duplicate wells, including a standard deviation (SD). The results were expressed as percentage parasite viability, which is the resorufin fluorescence in compound-treated wells relative to untreated controls. Note was taken of compounds that reduced parasite viability to < 10-20% for further testing.

5.2.3.2 Dose response assay

Compounds noted from the single dose assays were tested in triplicate wells with resorufin fluorescence as in the single dose assays, using 3-fold serial dilutions with 100µM as the highest concentration (100 µM - 0.0457 µM). For dose response, each compound percentage viability was plotted against Log(compound concentration) and the IC_{50} (50% inhibitory concentration) obtained from the resulting dose-response curve by non-linear regression. Pentamidine was once more used as the control drug standard yielding IC_{50} values in the range of 0.01-0.05 µM.

5.3 Results and Discussions

Hit compounds will be discussed using their DrugBank IDs.

5.3.1 Pharmacology of the hit compounds

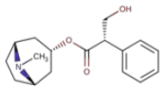
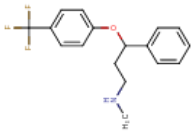
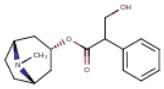
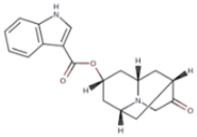
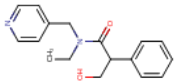
Thirteen drugs that presented favourable behaviour from the molecular docking experiments and dynamics simulations in *Tb*PTR1 and *Tb*DHFR were identified. Of these, three had anti-depressant properties, three were acetylcholine antagonists, two were anti-emetic agents, and the remaining five were an anti-convulsant, an anti-bacterial, an insect repellent, an antiseptic, and an anti-oxidant serving various purposes. All these drugs can cross the BBB [436,437], making them particularly useful in targeting the parasite in the meningo-encephalitic phase. The properties and binding energies of these drugs are summarised in ***Table 5***. For the remainder of this article, the hit compounds are referred to by their DrugBank IDs.

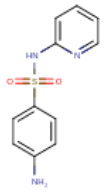
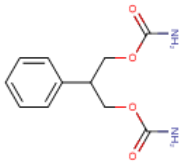
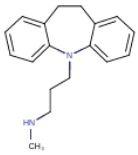
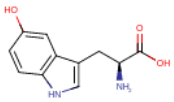
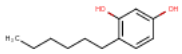
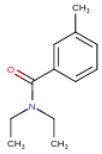
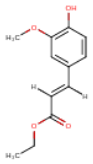
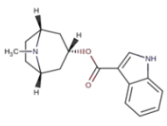
The anti-emetic agents, DB00757 and DB11699, are structurally similar, due to DB11699 being a precursor to DB00757. The former was initially designed for the treatment of migraines [438] before repurposing, thus conferring BBB permeability to the pair. DB00472, DB01511 and DB02959 are anti-depressant agents functioning through the inhibition of serotonin reuptake. DB01151 is a tricyclic anti-depressant (TCA), usually administered after resistance to other chemotherapies. This reluctance is due to the more significant anti-cholinergic side effects of TCAs [439].

DB00472 is more readily prescribed as it has less adverse effects [440]. DB02959 is a naturally occurring amino acid [441], 5-Hydroxytryptophan (5-HTP), that also functions as an appetite suppressant and sleep aid [442]. As anti-depressants are required to cross the BBB to affect neuronal function and work, the three drugs can cross the barrier. The extent of permeability

varies depending on the anti-depressant and its interaction with the P-glycoprotein protein pump at the BBB membrane [443,444]. DB00424, DB00572 and DB00809 are muscarinic acetylcholine antagonists, with the first two, hyoscyamine and atropine, being optical isomers [445]. The laevorotatory isomers are alkaloids sourced from various plants [446] to treat different conditions. Atropine is a racemic equimixture of D- and L- hyoscyamine. DB00424 is used for symptomatic relief from various gastrointestinal disorders. DB00809 is another anti-cholinergic agent, usually

Table 5: The chemical structures and classification of the top DrugBank compounds. The post-docking binding energies of the compounds in each protein are shown.

Drug Properties				***Binding Energy (kcal mol^{-1})***	
DrugBank ID	***Chemical Structure***	***Drug Name***	***Classification***	***TbPTR1***	***TbDHFR***
DB00424		Hyoscyamine	acetylcholine antagonist	-9.1	-8
DB00472		Fluoxetine	antidepressant	-9.2	-8.3
DB00572		Atropine	acetylcholine antagonist	-9	-7.8
DB00757		Dolasetron	antiemetic agent	-10.7	-9.5
DB00809		Tropicamide	acetylcholine antagonist	-8.2	-7.2

DB00891		Sulphapyridine	antibacterial agent	-8.3	-7.3
DB00949		Felbamate	anticonvulsant agent	-8.2	-7.7
DB01151		Desipramine	tricyclic antidepressant	-9.4	-8.5
DB02959		Oxitriptan	anti-depressant, appetite suppressant, and sleep aid	-8.8	-7.3
DB11254		Hexylresorcinol	antiseptic, anthelmintic, and local anaesthetic	-8	-7.7
DB11282		Diethyltoluamide	insect repellent	-8.5	-6.6
DB11285		Ethyl ferulate	experimental	-8.3	-7.2
DB11699		Tropisetron	antiemetic activity	-9.7	-8.6

administered as an anti-parasympathetic agent [447] for pupil dilation. Like anti-depressants, anti-cholinergic agents permeate the BBB [448], especially those that are neutrally charged [449]. This property made the drugs attractive candidates for the repurposing study. DB00891 is a sulphonamide anti-bacterial [450,451] that inhibits folate metabolism in bacteria through competitive inhibition of para-aminobenzoic acid (PABA). However, sulphapyridine therapy has been flagged for toxicity, affecting the frequency of prescription [452]. DB00949 is an antiepileptic drug [453] studied for administration in both children and adults [454]. The drug suppresses N-methyl-D-aspartate (NMDA) through its affinity for the glycine recognition site of the NMDA receptor [453]. Due to its target, the drug, too, is required to cross the BBB to take effect. DB11282 is an active insect repellent ingredient [455], while DB11254 has antiseptic, anthelmintic [456,457], as well as anaesthetic properties. The combined DB11285 properties have led to its use in anti-bacterial lozenges [458]. Lastly, DB11285, ethyl ferulate (FAEE), is a ferulic acid derivative with anti-oxidative properties [459,460]. FAEE has also been cited for its neuroprotective [461,462], anti-inflammatory [463] and antiproliferative [464] activity.

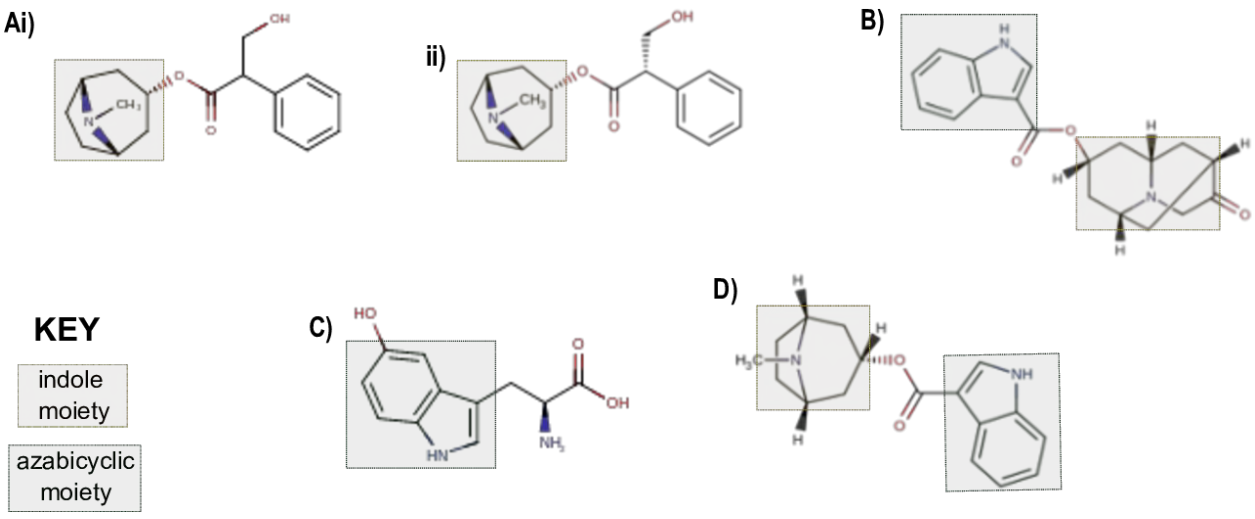


Figure 24: The azabicyclic and indole moieties are shared across the compounds ***ai)*** *DB00424,* ***aii)*** *DB00572,* ***b)*** *DB00757,* ***c)*** *DB02959, and* ***d)*** *DB11699. (Structure source : DRUGBANK)*

The indole moiety shared by DB11699 and DB00757, as well as DB02959 in ***Figure 24***, is a popular nucleus for several pharmacological purposes [465,466]. The architecture and

electrophilicity of the indole nucleus are conducive to the manipulation and synthesis of numerous derivatives [467]. To date, indole derivatives have been investigated and synthesised for the treatment of cancer [468,469], inflammation [470,471], emesis, and infections from microbes [472–474] and parasites [400,475]. In addition to the indole moiety, DB00757 and DB11699 have the azabicyclic architecture (***Figure 24***) in common with DB00424 and DB00572. Azabicyclic compounds have been investigated for anti-protozoal activity [476], and even more specifically, *Trypanosoma* [477], pointing to the modulatory potential of these drugs.

5.3.2 Analysis of ligand dynamics

5.3.2.1 Ligand RMSD

The RMSD of the ligands was used to quantify the internal fluctuation of the ligand atoms with respect to the protein Cα backbone. The RMSD values obtained stipulated the stability of the ligand binding in the protein complexes. The native ligand, folate, showed the highest fluctuation in *Tb*PTR1, with the RMSD distribution ranging from 0.1-0.4nm. DB00572, DB00949 and DB02959 present the most stability in both PTR1 and DHFR, corroborating the unimodal density seen in the violin plots in ***Figure 25***. Here, DB00572 has a median RMSD of 0.19nm in both proteins, and DB00949 has values of 0.14nm and 0.13nm in PTR1 and DHFR, respectively. Additionally, DB00891, DB01151, DB11254, DB11282 and DB11699 show the most stability in PTR1, with the RMSD values showing minimal fluctuation. As a result, the RMSD density distribution for these ligands is unimodal (***Figure 25A***). The *Tb*DHFR counterparts of these ligands exhibited multimodal RMSD distributions. DB01151 shows the most defined conformations with one prominent one. In *Tb*DHFR (***Figure 25B***), DB11699 shifted between two conformations due to twisting between the azabicyclic and the isoquinoline rings, finally settling at around 0.15nm.

DB11282, in great contrast to its mode in PTR1, sampled three conformations, which was the most across all ligands in both systems. DB00424, DB00472, DB00757, DB00809 and DB11285 took on multiple conformations in *Tb*PTR1. DB00424 presents more stability in *Tb*DHFR than *Tb*PTR1, with a unimodal RMSD frequency in the former. Upon video analysis of DB00424, rotation of the amide bond of the ligand was observed, explaining the two states. DB00757 changes conformation after 75ns, with a decrease in the RMSD. DB11285, like DB00424, exhibits a bimodal distribution in PTR1 contrasted with the more stable unimodal

one observed in DHFR. DB00472 exhibited bimodal distribution in *Tb*DHFR as well as PTR1, though it was more pronounced in PTR1.

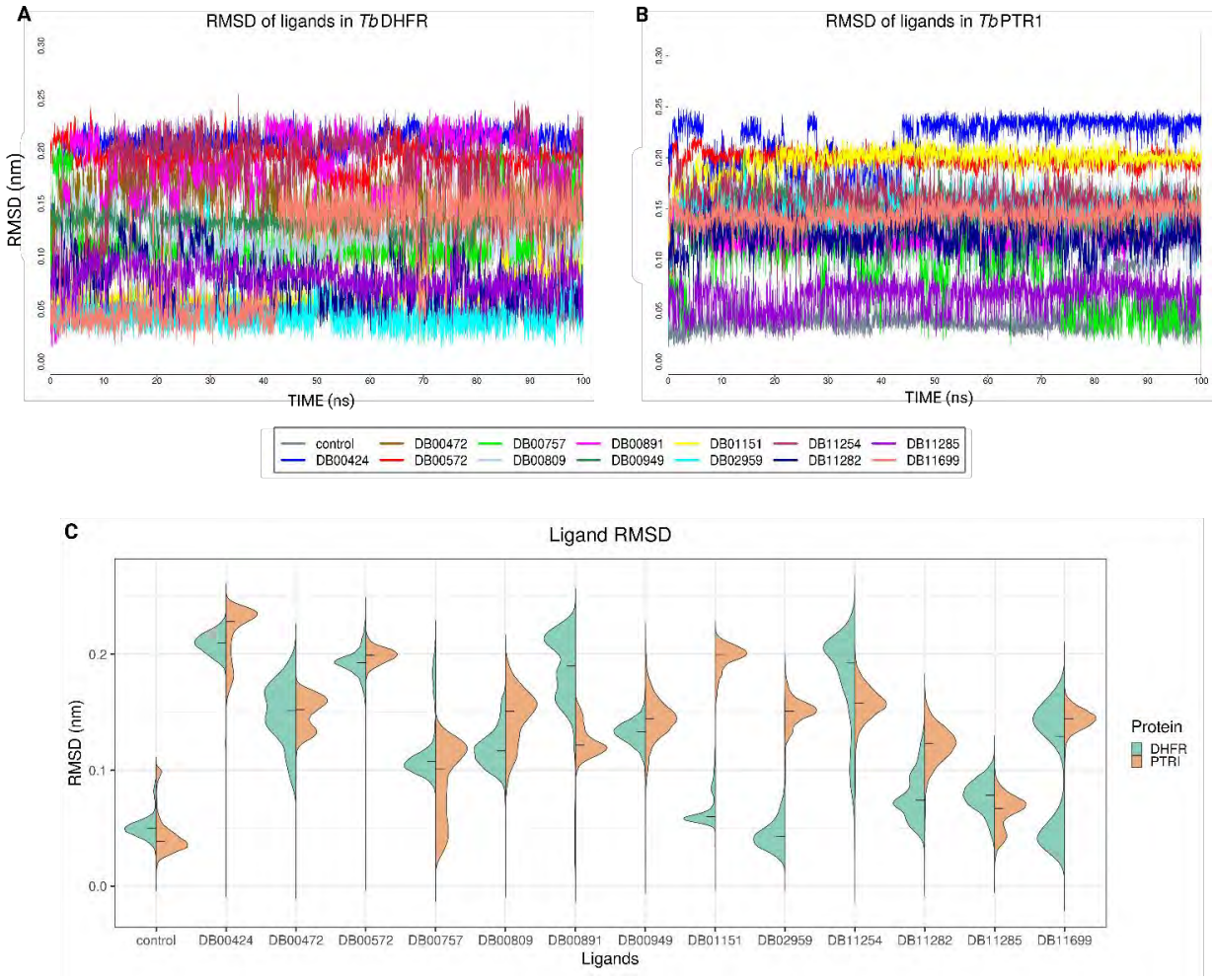


Figure 25: RMSD of the ligands in ***a)*** *TbPTR1 and* ***b)*** *TbDHFR presented in line plots. The median of the RMSD values is presented in the* ***c)*** *kernel density violin plot as a crossbar.*

5.3.3 Analysis of receptor dynamics

5.3.3.1 RMSD

The RMSD of the protein backbone was monitored to gain insight into any structural conformational changes in the protein during the simulation. The *Tb*PTR1-ligand complexes converged to equilibrium within the first 20ns of the simulation, apart from the *Tb*PTR1-DB11699 complex, which did so after the 60ns mark. The changes in backbone RMSDs were within the order of 0.1-0.3 nm, which accounts for intrinsic protein flexibility, but is indicative of minimal global conformational changes [196]. A general, albeit nominal, decrease in RMSD was observed in the ligand-bound systems compared to apo-*Tb*PTR1, apart from *Tb*PTR1-DB00424. The majority of the protein-ligand complexes sampled at most two structural conformations (**Figure 26**), with the exception of the complexes with DB00424, DB00472, DB00757, DB00809, DB00891 and DB00949. The conformations sampled in these unimodal RMSD distributions additionally had the lowest free energy.

*Tb*PTR1-DB00572, *Tb*PTR1-DB01151, *Tb*PTR1-DB02959, *Tb*PTR1-DB11254, *Tb*PTR1-DB11282, *Tb*PTR1-DB11285 and *Tb*PTR1-DB11699 presented two peaks, with one eventually becoming the dominant conformation (***Figure 26*** **&** ***Figure*** 27***A***). The *Tb*DHFR-ligand complexes, unlike those of PTR1, show slight increases in RMSD, with the majority sampling one conformation (***Figure 27*** **&** ***Figure S 23***). *Tb*DHFR-DB02959, like its PTR1 counterpart, exhibits a bimodal RMSD distribution, though with more pronounced peaks.

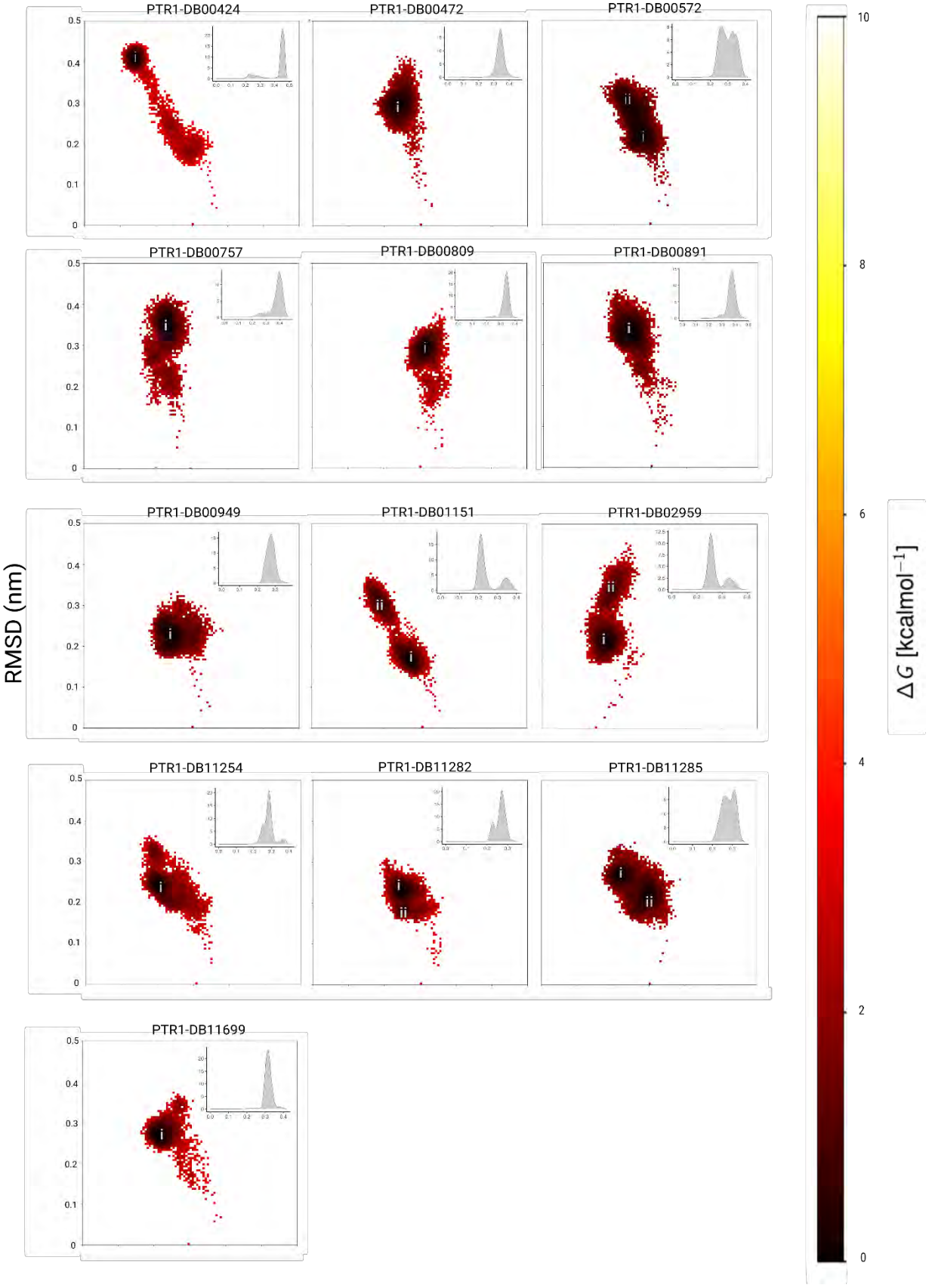


Figure 26: Protein backbone RMSD of ligand-bound TbPTR1 presented as a factor of the free energy of the protein. The colour intensity in the colour bar in the right presents the free energy (ΔG) in kcalmol^{-1} . The inserts in each plot present the kernel plots density of each system's RMSD, with corresponding peaks showing the lowest energy conformations.

5.3.3.2 Radius of gyration

The radius of gyration analysis of the complexes showed minimal change in the spatial compactness of the ligand-bound protein complexes for both proteins. The values remained within a range of 0.2nm and displayed both uni- and multi-modal distributions for the *Tb*PTR1 complexes (***Figure 27B***). A decrease in the Rg, comparable to that observed in the RMSD, was observed in the ligand-bound PTR1 complexes compared to the apo system.

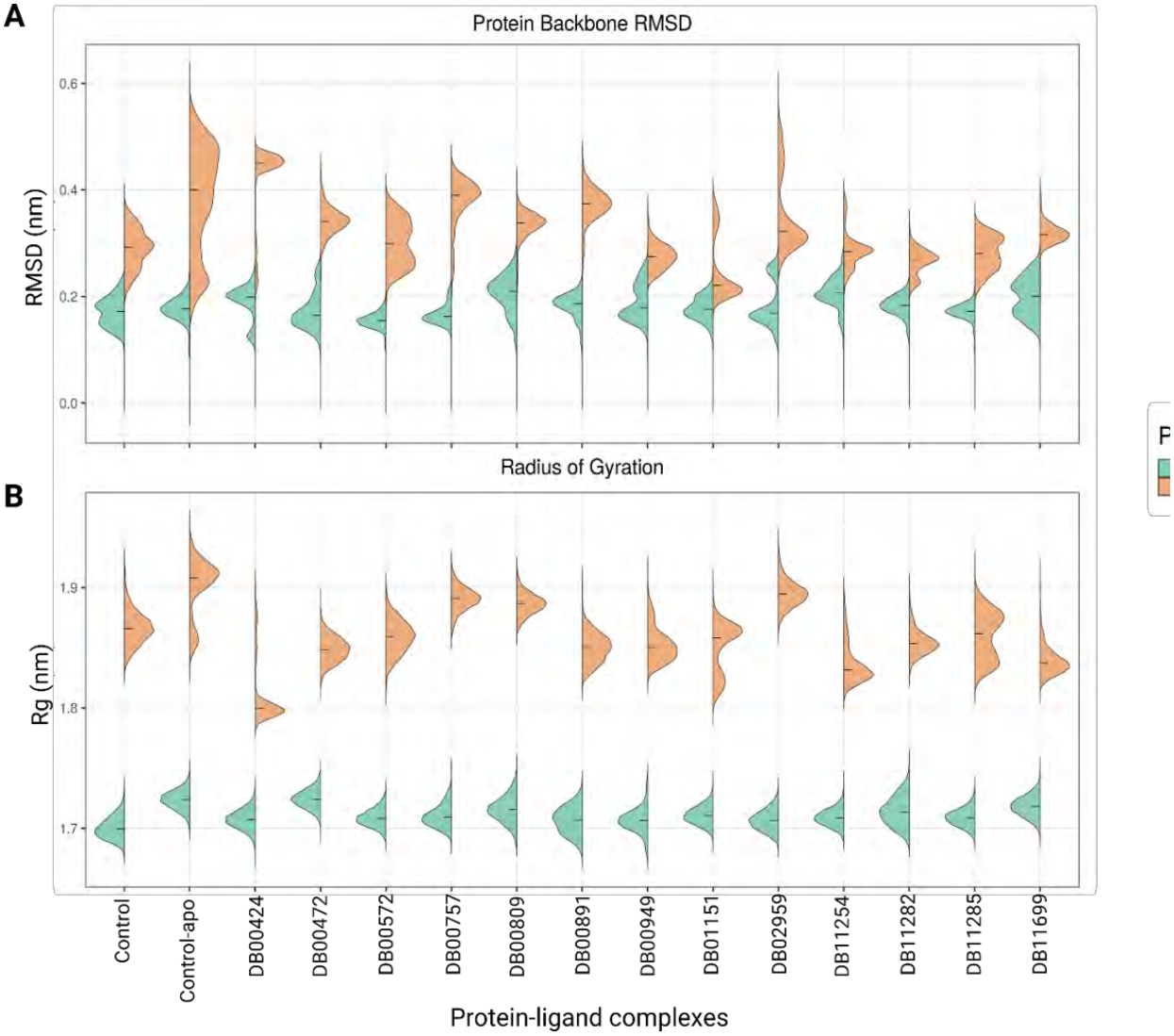


*Figure 27: **A)** RMSD and **B)** Rg of the protein backbone in ligand-bound TbPTR1 and TbDHFR presented in violin plots. The width of the plots outlines kernel probability density. Box plots represent the median ± the standard deviation.*

DHFR generally had lower Rg (***Figure 27B***), likely due to the smaller size of the monomer as compared to PTR1. The protein additionally maintained a unimodal distribution, suggesting

ligand binding did not markedly alter the compactness of the complexes. A slight decrease in the median was noted, with all values other than the apo-protein falling below 1.73nm.

5.3.3.3 RMSF

The RMSF captured provided insight into the flexibility of regions of the protein. Of significant interest are the residues 100-120 and 140-160 in both the ligand-bound and ligand-free complexes. These regions experienced high fluctuations as they are solvent-exposed loop regions. The ligand-free system showed high RMSF between residues 205-220 covering the substrate-binding loop. This increase in fluctuation was not mirrored by the ligand-bound *Tb*PTR1 complexes, suggesting the presence of the ligands and AX3 decreased the flexibility of the loop. Of note is *Tb*PTR1-DB00424, which showed both the least flexibility in the substrate-binding loop (***Figure 28A***) and the lowest Rg ***(Figure 28B***). Contrarily, *Tb*PTR1-DB02959, which had the largest Rg after the apo-PTR1, exhibited high fluctuations in the loop, thus suggesting that binding of the loop can affect the global nature of the complex.

DHFR exhibited flexibility in the loopy regions between residues 42-48 [478] shared across numerous protozoan DHFR enzymes, 64-77, 187-210, 222-230. An increase and decrease in RMSF in these regions may have a bearing on the changes in Rg seen in ***Figure 27B***. The RMSF of the DHFR-DB00472 remained similar to that of the ligand-free system, especially along all the loops. The Rg corroborated this with a median value of 1.72nm. Contrarily, the high fluctuation seen from residues 64-77 in the complexes of DB00949 and DB002959 did not appear to influence the Rg.

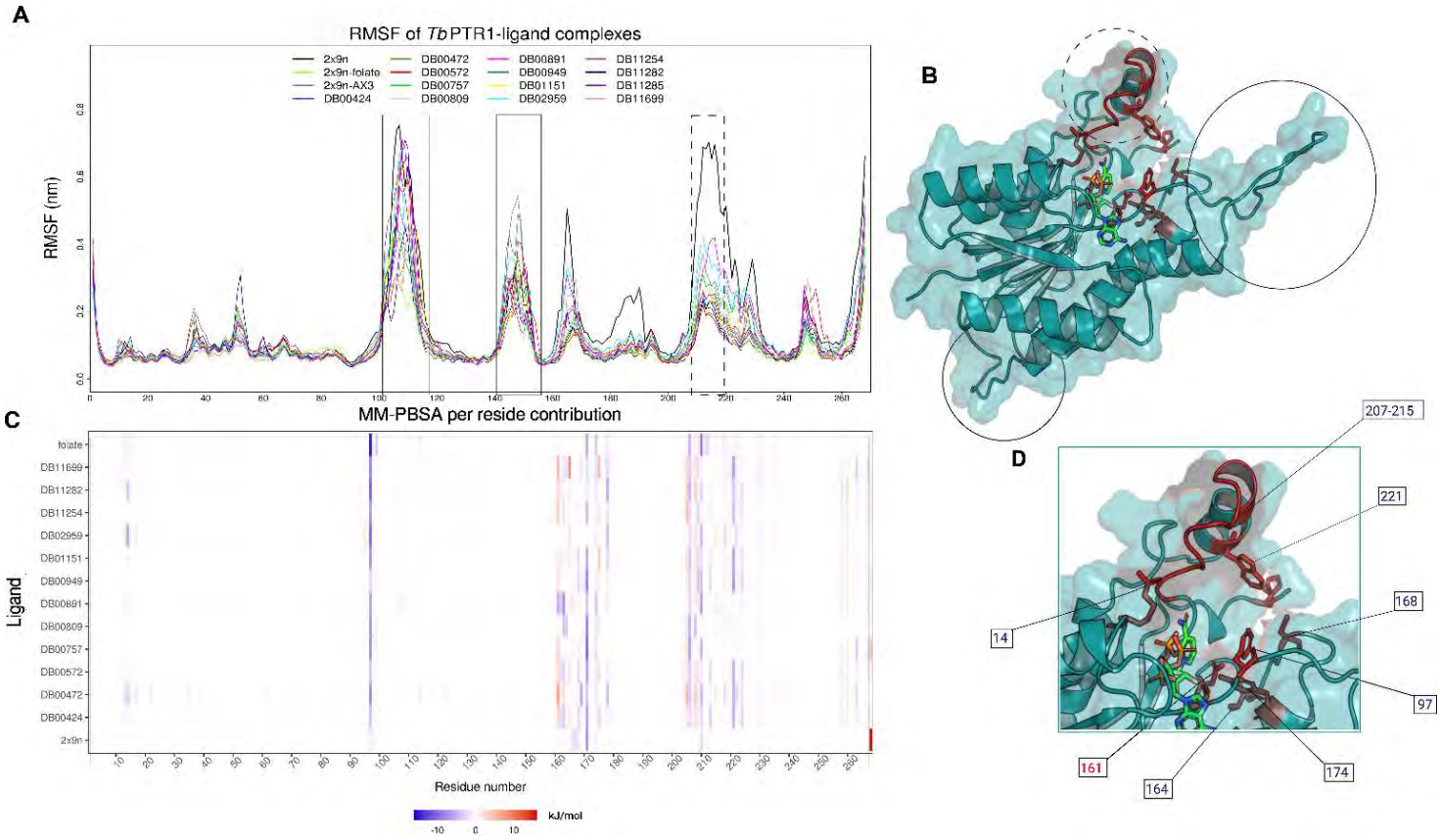


*Figure 28: RMSF of the **A**) TbPTR1 backbone, with the corresponding flexible regions shown in the **B**) protein structure. The MM-PBSA, per residue energy contribution, is shown in **C**, with the energy range represented in the colour bar. Positions of residues contributing the most energy are highlighted in **D**. The catalytically essential residues are coloured red in both structures **B** and **D**. (Structure source : PDB 2x9n)*

5.3.4 Energetics and protein-ligand interactions

5.3.4.1 MMPBSA

All eight ligands bound favourably in *Tb*PTR1 throughout the 100ns simulations, with free binding energies well below those of the native ligand AX3 at 43 ± 8 kJmol^{-1}. The binding and other energies are summarised in ***Table 6***. The ligands DB00472, DB00572 and DB01151 exhibited the lowest binding energies at -94±11 kJmol^{-1}, -84 ± 11 kJmol^{-1}, and -85 ± 8 kJmol^{-1}, respectively. DB01151 also had low energy in *Tb*DHFR, along with DB00757 and DB00809, as observed in ***Table 7***. The lack of hydrogen bonding observed in *Figure 32* may have contributed to the low binding energy of DB01151. The highest binding energies in PTR1

were recorded with DB00949, DB02959 and DB11254. This behaviour was mirrored in DHFR with DB00949 and DB002959. DB11254 exhibited the highest binding energy in *Tb*DHFR.

Table 6: A decomposition of the binding energy components obtained from MM-PBSA calculations in TbPTR1.

Ligand	***Energy Component kJmol^{-1}***				
	VDW	*Electrostatic*	*Polar solvation*	*SASA*	*Binding*
AX3	-68 ± 11	-18 ± 12	73 ± 20	-9 ± 1	-23 ± 12
FOL	-123 ± 11	-15 ± 9	80 ± 11	-14 ± 1	-73 ± 11
DB00424	-137 ± 8	-21 ± 5	91 ±9	-15 ± 1	-82 ± 10
DB00472	-158 ± 9	-28 ± 4	109 ± 7	-17 ± 1	-94 ± 11
DB00572	-135 ± 12	-21 ± 6	87 ± 9	-15 ± 1	-84 ± 11
DB00757	-115 ± 12	-1 ± 6	59 ± 12	-14 ± 1	-71 ± 12
DB00809	-92 ± 9	-33 ± 5	61 ± 5	-12 ± 1	-77 ± 8
DB00891	-113 ± 12	-61 ± 12	112 ± 8	-12 ± 1	-75 ± 11
DB00949	-104 ± 9	-12 ± 7	69 ± 10	-13 ± 1	-61 ± 10
DB01151	-121 ± 9	-5 ± 3	56 ± 7	-15 ± 1	-85 ± 8
DB02959	-82 ± 7	-9 ± 15	54 ± 16	-10 ± 1	-48 ± 11
DB11254	-106 ± 7	-2 ± 5	58 ± 7	-13 ± 1	-63 ± 10
DB11282	-114 ± 8	-2 ± 4	59 ± 9	-13 ± 1	-69 ± 11
DB11285	-105 ± 11	6 ± 4	40 ± 12	-14 ± 1	-72 ± 9
DB11699	-134 ± 9	-23 ± 5	104 ± 14	-15 ± 1	-67 ± 12

Interestingly, DB00949 participated in the largest number of hydrogen bonding **(*Figure 32*),** contrasting DB01151. The residue energy contribution ***(Figure 28***) highlighted critical residues interacting with the ligands and the nature of the interaction. Substrate binding residues within the active site region of Ser207-Trp221 and Phe97 were observed to contribute favourably to the binding energy, as did binding pocket residues between Asp161-Tyr174. Asp161, which with Tyr174 participates in proton shuttling for pterin reduction [479], formed undesirable interactions with all the ligands.

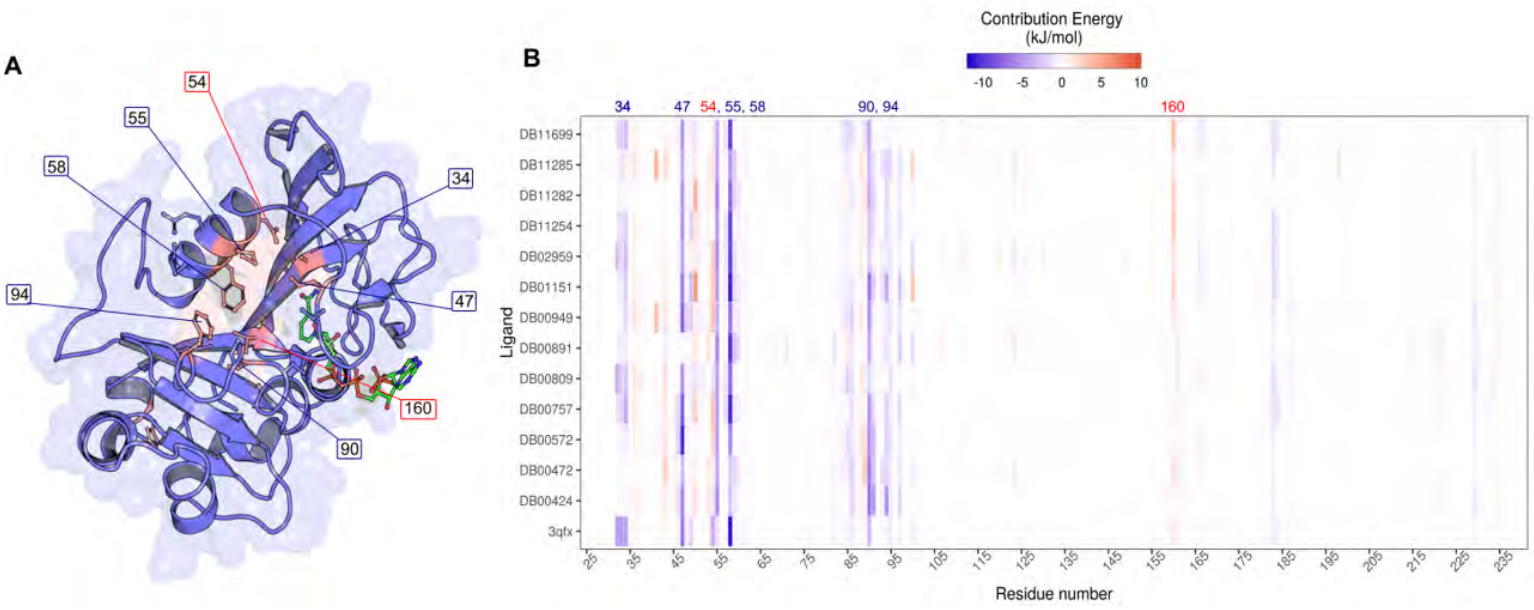


*Figure 29: MM-PBSA decomposition of residues in TbDHFR (in **A**) represented in a heatmap for each protein-ligand complex. The energy contribution of each residue is shown in **B**, with the residues attributing most to the energy highlighted in the structure **A** in salmon. (Structure source : PDB 3qfx)*

Overall, hydrophobic amino acids such as Ala, Met, Phe and Trp, with the exception of basic hydrophilic residues Arg and Lys, contributed favourably to the binding energy. The residues contributing unfavourably, Asn, Asp, Gln and Glu, were hydrophilic, except aliphatic Leu and Val.

Residues known to be involved in inhibitor binding, such as Gly45, Thr46, Met55, Arg59 and Phe94 [431,479], contributed constructively to the energy of the *Tb*DHFR ligand-bound systems. The energy contribution of the drugs was similar to that of the known DHFR inhibitor pyrimethamine (CP6), as shown in Error! Reference source not found.. All complexes revealed t he energetically unfavourable participation of Asp54.

Table 7: A decomposition of the binding energy components obtained from MM-PBSA calculations in TbDHFR.

Ligand	***Energy Component kJmol-1***				
	VDW	*Electrostatic*	*Polar solvation*	*SASA*	*Binding*
CP6	-148 ± 9	-37 ± 8	97 ± 6	-13 ± 1	-102 ± 9
DB00424	-116 ± 8	10 ± 3	37 ± 6	-15 ± 1	-82 ± 10
DB00472	-94 ± 10	-13 ± 5	47 ± 8	-13 ± 1	-73 ± 10
DB00572	-127 ± 8	-1 ± 3	57 ± 5	-15 ± 1	-85 ± 9
DB00757	-157 ± 11	-14 ± 8	85 ± 10	-16 ± 1	-102 ± 12
DB00809	-148 ± 10	-37 ± 8	101 ± 8	-16 ± 1	-101 ± 9
DB00891	-94 ± 9	-37 ± 7	79 ± 9	-10 ± 1	-63 ± 8
DB00949	-116 ± 10	-13 ± 8	89 ± 9	-14 ± 1	-59 ± 9
DB01151	-166 ± 11	-16 ± 4	86 ± 15	-16 ± 1	-114 ± 13
DB02959	-109 ± 8	-19 ± 9	82 ± 9	-12 ± 1	-59 ± 11
DB11254	-115 ± 12	-13 ± 7	64 ± 10	-14 ± 1	-78 ± 12
DB11282	-114 ± 9	-22 ± 4	72 ± 11	-14 ± 1	-78± 11
DB11285	-96 ± 13	-4 ± 6	63 ± 13	-14 ± 1	-51 ± 10
DB11699	-133 ± 8	-17 ± 7	77 ± 6	-15 ± 1	-88 ± 8

5.3.4.2 Intermolecular interactions

5.3.4.2.1 *Tb*PTR1

Based on residue numbering in *Tb*PTR1, Asp161 and Tyr174 are involved in the proton shuttling required for substrate reduction, while Arg14 plays a substrate stabilising role. The substrate-binding loop, which is conserved among SDR superfamily enzymes [430,480], spans Ser207-Gly215 and further holds the substrate in place. Other vital residues in the framing and shaping of the active site pocket include Phe97, Cys168, Tyr174, Pro210, Trp221 and His267, with the nicotinamide ring of NADPH and Phe97 creating the catalytic centre. The resultant pocket is hydrophobic, hinting that hydrogen bonding is unlikely. As such, the protein-ligand interactions were primarily hydrophobic and van der Waals (vdW). The acetylcholine antagonists all interacted with NADPH and Phe171 via π-π-stacking of the aromatic groups, with DB00424 additionally stacking with Phe97 and forming hydrogen bonds with Ala162,

Met163 and Gly205. The hydrogen bonding was mirrored in the isomer DB00572, which further participated in hydrophobic interactions with Cys168 and Trp221. DB00809 furthermore exhibited hydrogen bonding with Tyr174 and an alkyl interaction with Asp161. In agreement with the substrate interaction, Phe97 formed π-π-stacking with the aromatic groups DB00757 and DB11699, with the added π-alkyl bonding with NADPH. Due to the structural similarities between DB00757 and DB11699, the compounds exhibited analogous interactions. DB00757 had favourable binding in all four active sites of *Tb*PTR1, exhibiting numerous energetically viable binding poses [481]. In chain A, DB00757 interacted with Val206, Phe171 and Gln166, while DB11699 formed hydrogen bonds with Trp221 and Met163, in addition to π-alkyl stacking with Leu209 and Phe171.

Phe97 binds with all the anti-depressants mirrored in DHFR with Phe58. DB00472 and DB02959 interact with the co-factor, whilst DB00472 and DB01151 form hydrophobic contacts with Arg14, Met163, Phe171 and Pro210. DB00949 interacts with NADPH, forming the most hydrogen bonds among the ligands, as seen in ***Figure 30***. The anaesthetic agent, DB11254, participated in hydrogen bonding with Asp161, an interaction replicated in DHFR. DB11282, like the anti-depressant agents, formed π interactions with the active site phenylalanine residues, as well as with Met163, Tyr174, Met213 and Trp221. Lastly, DB11285 formed a hydrogen bond with NADPH, a π-π stack with Phe97, and additional alkyl and π-alkyl interactions with Leu209 and Trp221.

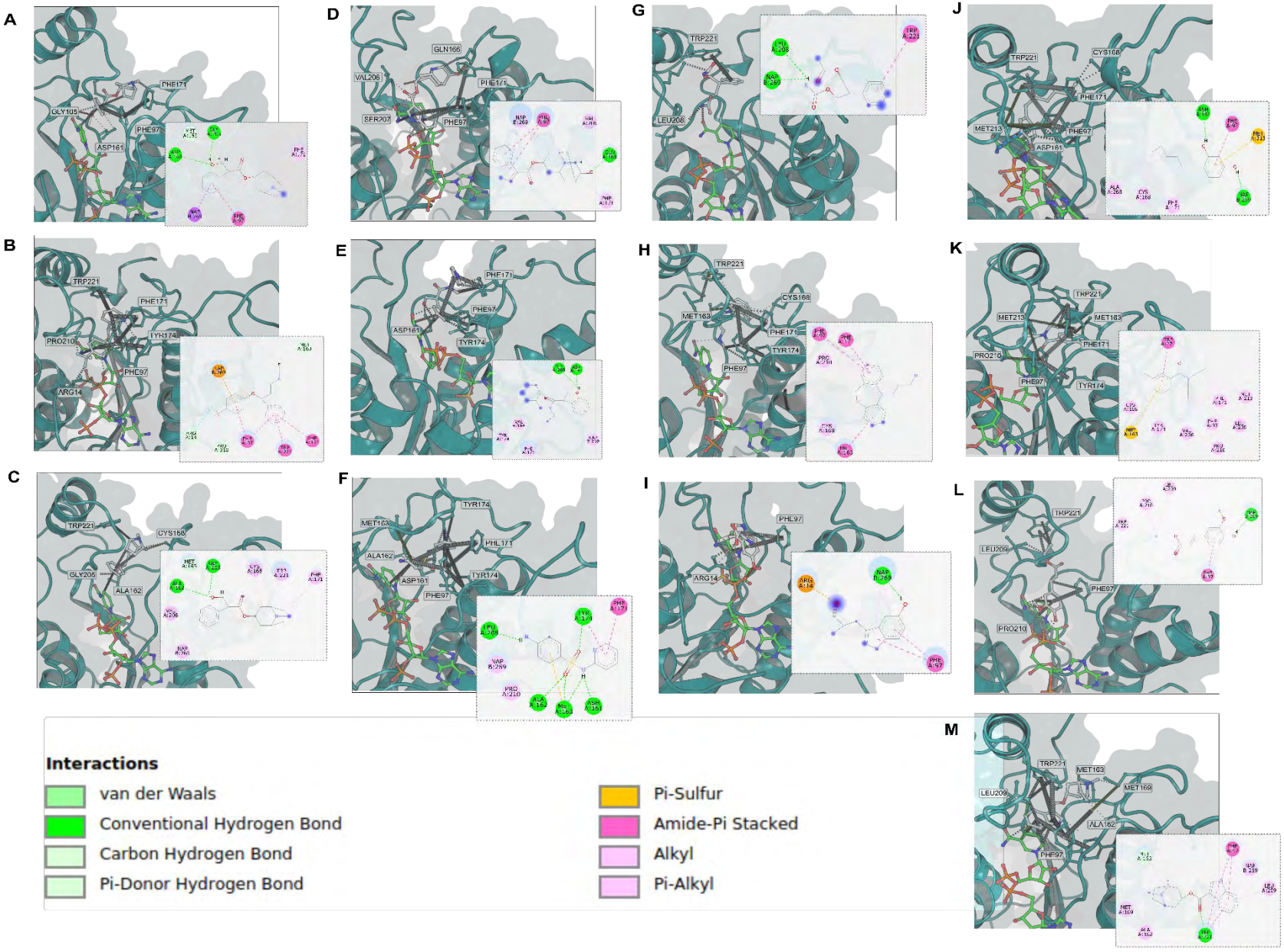


Figure 30: The intermolecular interactions of ligands ***a)*** *DB00424,* ***b)*** *DB00472,* ***c)*** *DB00572,* ***d)*** *DB00757,* ***e)*** *DB00809,* ***f)*** *DB00891,* ***g)*** *DB00949,* ***h)*** *DB01151,* ***i)*** *DB02959,* ***j)*** *DB11254,* ***k)*** *DB11282,* ***l)*** *DB11285 and* ***m)*** *DB11699 in the TbPTR1 active site are shown in 3D representation.The inset present a schematic of each binding pose, highlighting each interaction according to the provided key.*

5.3.4.2.2 *Tb*DHFR

The intermolecular interactions observed in *Tb*DHFR were highlighted in ***Figure 31***.

Residues paramount to folate reduction in *Tb*DHFR include Ala34, Val32, Met55, Phe58, Ser89, Phe94 and Tyr166. Much like in PTR1, the active site of DHFR is a neutral hydrophobic cavity [478]. Due to the hydropathy of the active site, ligands were retained in the active site via an array of hydrophobic and vdW interactions and a few hydrogen bonds. Asp54 and NADPH were responsible for most of the hydrogen bonding observed, with the former binding to DB00757, DB00809, DB00949 and DB11254. Ile47 formed alkyl interactions and π-stacking with DB11699, DB02959, DB01151, DB00757, DB00572 and DB00424, while other isoleucine residues of interest, including Ile50 and Ile160, interacted with DB00809 and DB00891.

Of the phenylalanine residues of the active site, Phe58 facilitated π-π-stacking with the aromatic ring-containing ligands, as well as the azabicyclic moieties of DB00757, expect in the case of optical isomers DB00424 and DB00572, and DB11699. The residue Met55 also attracted the ligands into π-alkyl, π-sulphur, and hydrogen bonds, with the exception of DB00757 and DB02959. Leu90 facilitated alkyl contacts with the ligands DB00472, DB00572, DB00809, DB01151 and DB11282. Ser89 bound exclusively to the anti-emetic agents DB00757 and DB11699, with the reinforcement of hydrophobic Ala34. The alanine residue also formed π-alkyl interaction with the aromatic moieties of DB00572, DB00809, DB11254 and DB11285. Interestingly, DB01151 was the only ligand to form hydrophobic interactions in both PTR1 and DHFR solely.

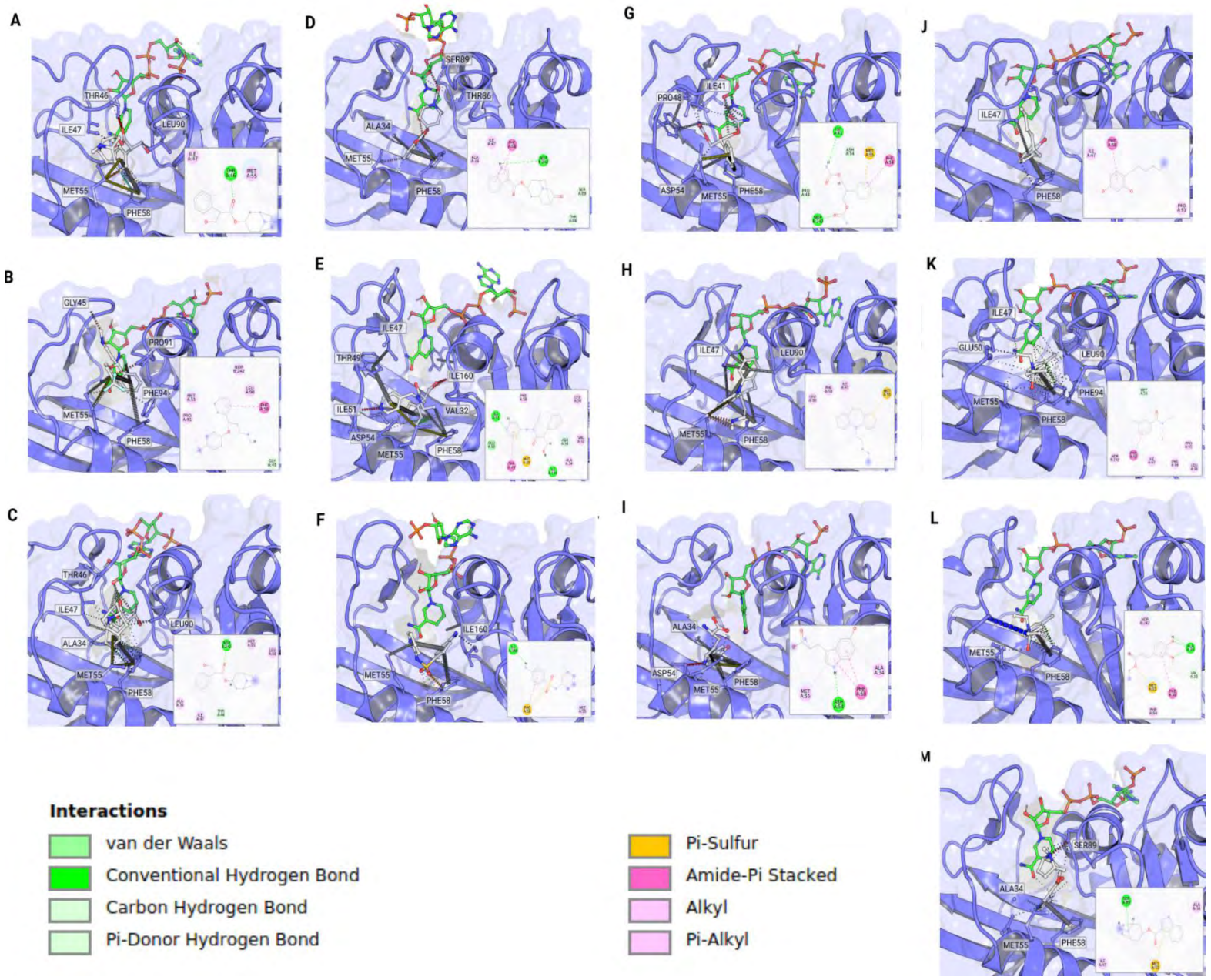


Figure 31: Intermolecular interactions between the ligands ***a)*** *DB00424,* ***b)*** *DB00472,* ***c)*** *DB00572,* ***d)*** *DB00757,* ***e)*** *DB00809,* ***f)*** *DB00891,* ***g)*** *DB00949,* ***h)*** *DB01151,* ***i)*** *DB02959,* ***j)*** *DB11254,* ***k)*** *DB11282,* ***l)*** *DB11285 and* ***m)*** *DB11699 and residues in the TbDHFR the active site are presented in 3D structure representation. A schematic outline of the binding pose is shown in the insets, with an accompanying key for the type of interactions.*

5.3.4.3 [478]Hydrogen bonding occupancy analysis

Analysis of hydrogen bonding between ligands and proteins revealed that the NADPH co-factor attracted hydrogen bonding with numerous ligands in both *Tb*PTR1 and *Tb*DHFR. Other residues of note were Asp161 and Tyr174 in PTR1. In DHFR, Asp54 and Ile160 formed the most hydrogen bonds, with other isoleucine residues, Ile41, Ile47 and Ile51, making an appearance. DB00424 formed at least four hydrogen bonds in both *Tb*PTR1 and *Tb*DHFR, with an occupancy of < 0.56 in both proteins ***(Figure 32***). DB00572 formed fewer hydrogen bonds than its isomer in *Tb*PTR1, with Ala162 and Asp161 in common. In *Tb*DHFR, the isomers had the same number of bonds, sharing bonds with NADPH, Ser89 and Thr46. DB00757 exhibited six hydrogen bonds in PTR1, with the highest occupancy in Met163 at 0.48. In contrast, the ligand formed only two bonds in DHFR, including one with NADPH, unlike its PTR1 counterpart. DB00809 formed two hydrogen bonds in both cases, with bond occupancies higher than 0.68. Isoleucine was responsible for the hydrogen bonding in DHFR, but aspartate and glutamate in PTR1. Antibiotic DB00891 displayed high hydrogen bond occupancies of 0.93, 0.92, 0.83, 0.4 and 0.35 with residues Asp161, Ala162, Met163, Tyr174 and Leu208, respectively. The highest number of hydrogen bonds was observed with DB00949, where nine bonds were recorded in *Tb*DHFR. DB00949 also exhibited the most considerable frequency in binding with the co-factor. DB02959 and DB11254 participated in a minimum of three bonds in both systems, including those with Asp and NADPH. *Tb*DHFR-DB11285 hydrogen bonding recorded five contacts, where Ala34 had the highest frequency at 0.15. The binding of the ligand in *Tb*PTR1 was low. DB11699 formed three hydrogen bonds in PTR1 and two in DHFR. In PTR1, the highest occupancy noted was 0.34 with Tyr174, and one of 0.42 was recorded with Asp54 in *Tb*DHFR.

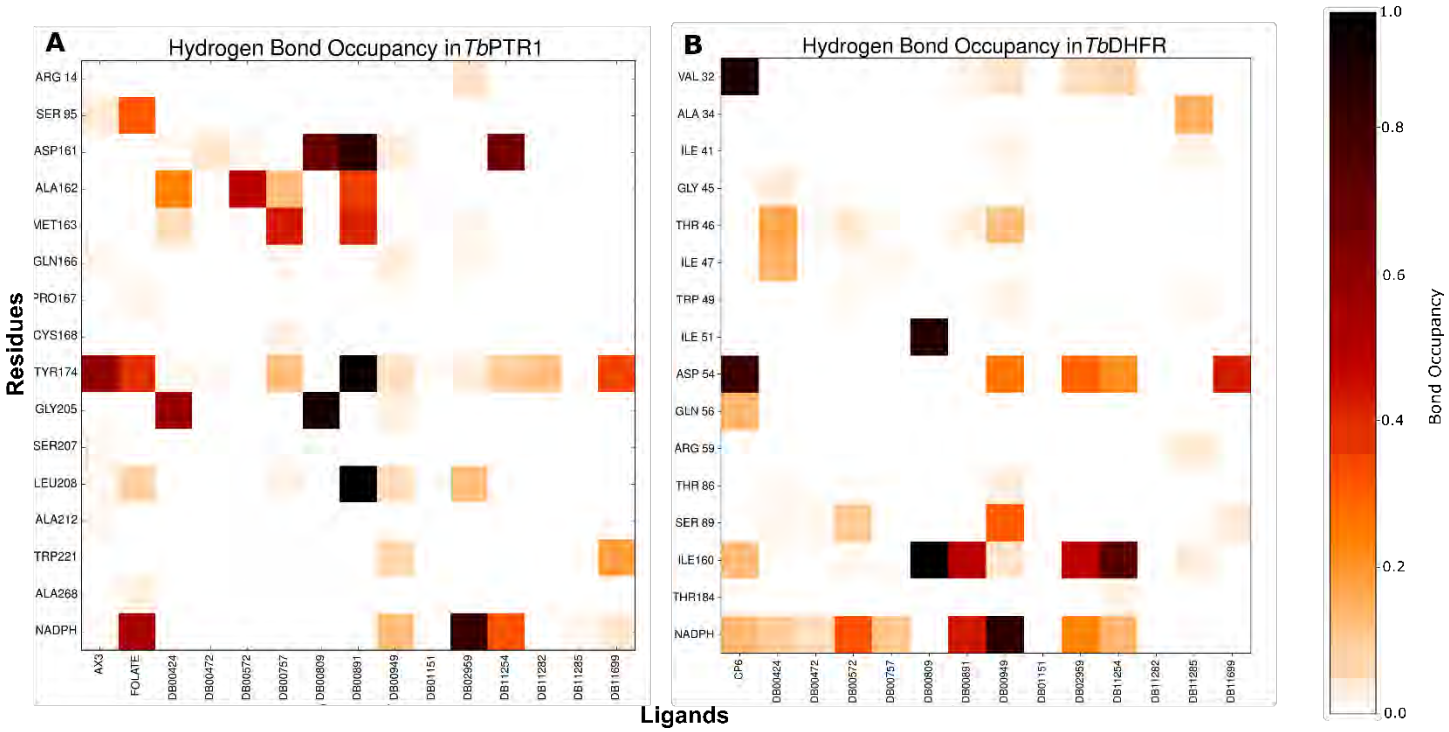


Figure 32: Hydrogen bonding occupancy of the hit ligands with the binding pocket residues in ***A)*** *TbPTR1 and* ***B)*** *TbDHFR. Occupancy is presented as a percentage of the total number of frames in the MD simulations.*

DB01151 and DB11282 exhibited poor hydrogen bonding in both proteins, with a maximum of one hydrogen bond per frame over the observed 100ns. The DB01151 hydrogen bonding profile corroborated the observations made in the interaction analysis, where the drug primarily only formed hydrophobic interactions. The lack of hydrogen bonding is likely due to the hydrophobic nature of the drug, which has a log P of 4.9.

The Rg analysis of the complexes showed minimal change in the spatial compactness of the ligand-bound protein complexes for both proteins. The values remained within a range of 0.2nm with unimodal distributions, with the exception of the co-crystallised 2X9N system.

5.3.5 Four compounds exhibited potential anti-trypanosomal activity

Of the 13 drugs isolated *in silico*, compounds DB00472, DB00572, DB01151 and DB11254, as well as AX3 and CP6, were active at 20µM, exhibiting % cell viability <10% (***Figure 33***). As a result, the samples were put forward for IC_{50} assays.

Pentamidine was used as the assay control against the co-crystallised (cyromazine (AX3) and pyrimethamine (CP6)) and potential hit compounds. The IC_{50} of DB00472, DB00572, DB01151, DB11254, AX3 and CP6 were determined to be 1.61 ± 0.08, 9.76 ± 0.99, 1.87 ±

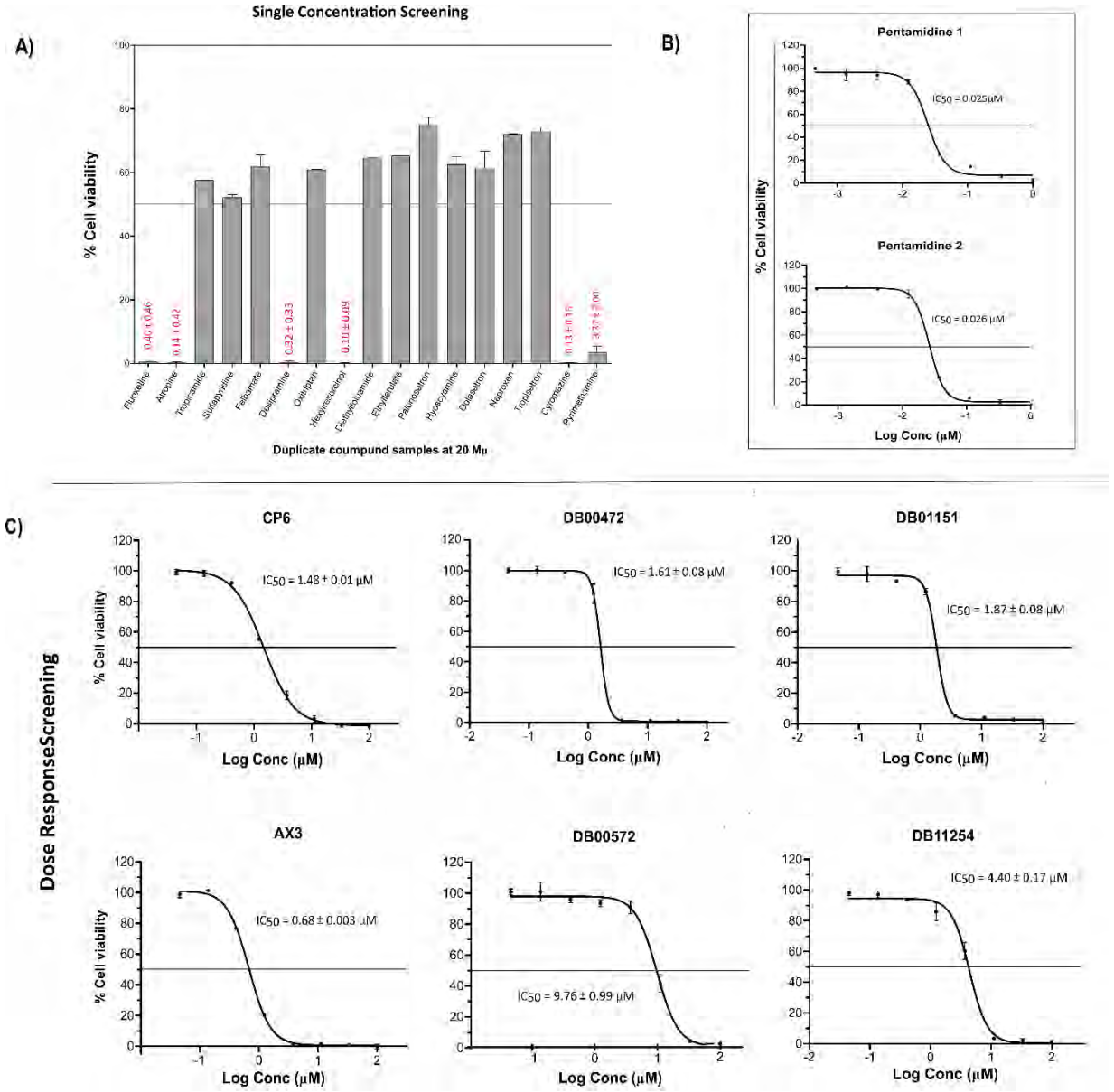


Figure 33: Anti-trypanosomal activity bioassays: ***A)*** *Single concentration assays. Percentage parasite viability (% viability ± SD) of all 13 compounds after single concentration assays at 20 πM.* ***B)*** *Pentamidine was used as the assay control and run in duplicate, as well as co-crystallised compounds (cyromazine (AX3) and pyrimethamine (CP6) shown in* ***C****). Viability of DB00472, DB00572, DB01151, DB11254, AX3 and CP6 (highlighted in red) were determined to be 0.40 ± 0.46, 0.14 ± 0.42, 0.32 ± 0.33 and 0.10 ± 0.09 μM, respectively. Pentamidine % viability was 0.025-0.026 μM.* ***C)*** *IC_{50} values of compounds represented in dose-response plots and % viability ± SD.*

. 0.08, 4.4 ± 0.17, 0.68 ± 0.003 and 1.48 ± 0.01 μM, respectively. The IC_{50} of pentamidine was 0.025-0.026 μM. CP6 had sub-micromolar activity of the compounds, as was expected of an existing anti-folate drug. While the four DrugBank compounds exhibited super-micromolar activity, further experimental investigation is required.

The potential of active compounds can be better assessed in conjunction with cytotoxicity assays. However, as the database used (DrugBank), [168] consists of well-known and -documented compounds, a considerable amount of data is available to consult regarding their toxicity in lieu of bioassay

5.4 Conclusion

In the chapter, high throughput virtual screening was employed to repurpose FDA-approved drugs for use in neglected tropical disease HAT. Drug repurposing has become an increasingly promising venture for drug discovery in ordinarily impoverished conditions. Repurposed drugs present a financial and temporal advantage to further develop and optimise these drugs for use in chemotherapy. The drugs identified through *in silico* screening process in this study showed favourable binding to the folate pathway proteins, making them viable options for modulation. Of these drugs, DB00572, DB00949, DB01151 and DB11699 were the most promising in both *Tb*PTR1 and *Tb*DHFR, opening another avenue in dual anti-folate anti-trypanosome chemotherapy. DB00572 and DB01151 also exhibited some potential *in vitro* assays screened against the parasite.

Additionally, the indole and the azabicyclic moieties were identified across at least three drugs, which could be instrumental in analogue screening and scaffold development of a safe anti-folate drug for anti-trypanosome therapy.

Chapter 6

Overall Conclusions

6.1 Findings and their relevance

In this project, we used structural bioinformatics approaches to investigate two aspects of HAT, which remains a significant health concern in sub-Saharan Africa. Our aims were divided in two, 1) investigating the proteins involved in the pathogenicity of r-HAT, *Tbr*SRA and human ApoL1 to assist the understanding of infection initiation, and 2) targeting folate pathway protein *Tb*PTR1 and *Tb*DHFR as drug targets for potential alternative drugs

6.1.1 Pathogenicity

Chapters 2-4 of this study aimed to investigate the mechanism of *Tbr*SRA mediated HAT infection through investigations of ApoL1, SRA, and the potential SRA-ApoL1 complex.

Chapter 2 entailed the study of ApoL1 variants G0, G1, G1M, G1G, G2, and G2G1, providing insight into the intramolecular effect the mutations play in protein functionality. The presence of ApoL1 C-terminal variants had a destabilising effect on SRA-ID. This plausibly attributes to the ability of the variants to evade SRA.

The study of four SRA variants in ***Chapter 3*** highlighted the minimal effect of the investigated mutation on the functional protein as predicted by DRN analysis. In lieu of dynamic consequences of the SRA variants, we identified potential mutational cold spot residues noted through DRN metrics *BC, CC, DC,* and *KC*. Residues Ala98 and 105 specifically were recorded as *persistent hubs* across the centrality metrics suggesting their conservation and importance to both the structural and functional integrity of SRA and making them viable targets for disruption of SRA anti-ApoL1 activity.

Lastly, we isolated the potential nature of the SRA-ApoL1 complex in ***Chapter 4***. Through protein-protein docking, we calculated two feasible dimeric structures of the SRA-ApoL1.

The identified complexes exhibited communication across the interface through the calculation of *BC* and *CC*, which highlighted *prominent hubs*, including residues identified in both ***Chapters 2 and 3***. Complexes carrying these prominent hubs also proved to be energetically feasible through PCA and external PDBePISA calculations.

The introduction of ApoL1 mutations into the complexes complemented observations noted for the ApoL1 monomer from ***Chapter 2***. We also recorded the significance of leucine residues to the complex formation in a coiled-coil suggestion of interaction.

These address the knowledge gap concerning the roles of both proteins in HAT susceptibility. Additionally, our observations identified key residues to the maintenance of the complex, providing more opportunities to prevent complexation or disruption of existing complexes.

6.1.2 Drug discovery

Chapter 5 of this thesis was dedicated to the search for potential anti-folate drug compounds as an alternative to existing HAT chemotherapy by implementing *in silico* drug repurposing.

Four compounds cross-binding in validated anti-folate therapy targets *Tb* PTR1, and DHFR were identified, encouraging the possibility of wholly effective anti-folate therapy against African trypanosomes. The compounds were promising in single-concentration bioassays but exhibited super-micromolar anti-trypanosome activity in IC50 bioassays. However, shared indole and the azabicyclic architecture were identified across the drugs, which could be used as scaffolding to design compounds with improved molar activity.

Overall, the aim of the chapter was achieved in the identification of DrugBank compounds DB00572, DB00949, DB01151 and DB11699 for additional investigation towards anti-folate HAT chemotherapy. We additionally highlighted the convenience of drug repurposing in combination with HTVS, which reduces the development time, encouraging investment in HAT chemotherapy endeavours.

6.2 Limitations and challenges

6.2.1 Molecular dynamics

Despite its obvious convenience, the MD simulation used in the study has several limitations. These include the inability to model quantum effects, force field quality, and time and size

limitations. High-performance computing (HPC) has alleviated some of the computational intensity of MD simulations, but more work needs to be done to improve these aspects [482–485].

Additionally, singular runs we conducted for the calculation of ***Chapter 2, 3*** and *5* in the interest of managing computational resources. The differing lengths in MD simulations noted in each chapter were dependent on the ease of convergence of each system. Larger biomolecular[486] systems tend to require more simulation time to converge and stabilise compared to smaller counterparts.

6.2.2 Full-length structure of ApoL1

Our study concentrated on the protein region where the variants primarily associated with HAT are located and did not take the whole protein available in the AlphaFold database[102]. In this setting, future studies will be critical to understanding the structure and effect of the HAT-associated variants in the context of the full-length protein.

Despite these limitations, our study provides clues to the structure and conformational dynamics of the G0 ApoL1 C-terminal and its variants, as well as their interactions with SRA. The ApoL1 C-terminal region structure suffices for the initial aim of our study, advancing our understanding of ApoL1 structure and function.

6.2.3 Lack of cytotoxicity assays

The potential of active compounds can be better assessed in conjunction with cytotoxicity assays. Since these are well-known compounds, there is considerable data regarding their toxicity [168].

However, cytotoxicity assays play a dual role as the compounds additionally need to be tested for selectivity in *in vivo* or *in vitro* models. The successfully screened compounds may bind to human homologs [487,488] of DHFR in addition to parasite's, resulting in undesirable side effects . Some antifolate drugs, such as sulfa drugs, are examples of selective antifolate therapy that were once used for many bacterial infections

6.3 Relevance and Future Work

The two overall aims of this thesis were not exhaustively solved, allowing for enhancements to our current results.

In our pathogenicity study, further structural studies of ApoL1 and SRA in membranes could be implemented to provide significant insight into the molecular basis of the critical interaction that facilitates *T. b. rhodesiense* and ApoL1 variant–mediated trypanosomiasis. As highlighted in ***section 6.2.2***, while using an ApoL1 fragment was sufficient for the current investigation, the inclusion of a full-length structure can only improve the understanding of SRA-ApoL1 interaction. In the study by Zoll, [339] , two lower affinity binding sites were identified in addition to the SRA N-terminal helix, suggesting a more complex binding which might be complemented by the presence of a membrane to facilitate the membrane addressing domain of ApoL1[282,362] .

Aside from their trypanocidal role, ApoL1 variants play a significant role in the development of CKD[287,291,356] and have been targeted in nephropathy related studies [304] . As the SRA interaction with ApoL1 prevents it from inserting into endocytic membranes it is likely a balance between the two plays a role in the development of ApoL1 nephropathy. The SRA-ApoL1 interaction is therefore a potential target for therapeutic intervention. As such the identification of the binding modes of ApoL1 and its risk variants are crucial in modulation of the protein for both HAT and CKD to varying outcomes.

Additionally, expanding understanding of SRA-ApoL1 interaction may aid in the development of druggable targets, after the example of previous studies that have inversely identified targets that prevent *Tbb* lysis [489].

In the drug discovery study, further docking of the identified hit compounds against human orthologues would enhance findings, while cytotoxicity investigations would greatly complement *in silico* and the data retrieval work conducted . While the use of these drugs via repurposing likely points to the drugs being non-toxic at the concentrations used, repurposing opens up the possibility of modulating unwanted drug targets [487,488]. As such standalone assays could provide further insight into drug selectivity, ensuring only parasite proteins are targeted. Additionally, the compounds identified can be further used as scaffolding in the development of more effective analogues that may perform better in experimental assays against the parasite. To date, drug repositioning has been used to identify plausible antifolate drug for Chagas disease in a similar fashion to this study[490] .

In the advent of AI-driven structure prediction software such as AlphaFold2[102], and AlphaFold Multimer[491], we believe improvements can be made to existing homology modelling and protein docking approaches. It is important to note that while AlphaFold has made significant

advancements in protein structure prediction, it still has limitations and may not be applicable in all scenarios[492,493]. As of CAPRI (Critical Assessment of PRedicted Interactions) (*https://www.ebi.ac.uk/pdbe/complex-pred/capri/*; *http://www.capri-docking.org/*) Round 50 [494–496] , where 63% of generated models were of acceptable quality[497], further research and development are needed to address these limitations and improve the accuracy and applicability of the predictions [493].

References

1. Brun, R., Blum, J., Chappuis, F. & Burri, C. Human African trypanosomiasis. *The Lancet* vol. 375 148–159 Preprint at https://doi.org/10.1016/S0140-6736(09)60829-1 (2010).
2. Steverding, D. The history of African trypanosomiasis. *Parasit Vectors* **1**, 3 (2008).
3. Krafsur, E. S. Tsetse flies: Genetics, evolution, and role as vectors. *Infection, Genetics and Evolution* vol. 9 124–141 Preprint at https://doi.org/10.1016/j.meegid.2008.09.010 (2009).
4. Lindner, A. K. & Priotto, G. The unknown risk of vertical transmission in sleeping sickness-a literature review. *PLoS Neglected Tropical Diseases* vol. 4 1–5 Preprint at https://doi.org/10.1371/journal.pntd.0000783 (2010).
5. Mudji, J., Künzli, E., Molyneux, D. & Blum, J. Long-term sequelae of congenital gambiense human African trypanosomiasis. *Trans R Soc Trop Med Hyg* **115**, 932–936 (2021).
6. Lestrade-Carluer de Kyvon, M. A. *et al.* Congenital trypanosomiasis in child born in France to African mother. *Emerging Infectious Diseases* vol. 22 935–937 Preprint at https://doi.org/10.3201/eid2205.160133 (2016).
7. Rocha, G., Martins, A., Gama, G., Brandão, F. & Atouguia, J. Possible cases of sexual and congenital transmission of sleeping sickness [5]. *Lancet* vol. 363 247 Preprint at https://doi.org/10.1016/S0140-6736(03)15345-7 (2004).
8. Maia, A. *et al.* GP54 Sleeping sickness: congenital case associated with a possible sexual transmission. in *Archives of Disease in Childhood* vol. 104 A51.2-A51 (BMJ Publishing Group Ltd, 2019).
9. Büscher, P., Cecchi, G., Jamonneau, V. & Priotto, G. Human african trypanosomiasis. *The Lancet* **390**, 2397–2409 (2017).
10. Bukachi, S. A., Wandibba, S. & Nyamongo, I. K. The socio-economic burden of human African trypanosomiasis and the coping strategies of households in the South Western Kenya foci. *PLoS Negl Trop Dis* **11**, (2017).

11. Solari. *Programme Against African T rypanosomiasis FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS*. (2009).

12. Bouteille, B. & Buguet, A. Human African Trypanosomiasis. in *Encyclopedia of the Neurological Sciences* 618–624 (Elsevier Inc., 2014). doi:10.1016/B978-0-12-385157-4.00391-2.

13. Kennedy, P. G. E. The continuing problem of human African trypanosomiasis (sleeping sickness). *Ann Neurol* **64**, 116–126 (2008).

14. Büscher, P. *et al.* Do Cryptic Reservoirs Threaten Gambiense-Sleeping Sickness Elimination? *Trends in Parasitology* vol. 34 197–207 Preprint at https://doi.org/10.1016/j.pt.2017.11.008 (2018).

15. Hamill, L. *et al.* Evaluating the impact of targeting livestock for the prevention of human and animal trypanosomiasis, at village level, in districts newly affected with T. b. rhodesiense in Uganda. *Infect Dis Poverty* **6**, 1–12 (2017).

16. Kennedy, P. G. E. Human African trypanosomiasis-neurological aspects. *Journal of Neurology* vol. 253 411–416 Preprint at https://doi.org/10.1007/s00415-006-0093-3 (2006).

17. Kennedy, P. G. E. Human African trypanosomiasis of the CNS: current issues and challenges. *Journal of Clinical Investigation* **113**, 496–504 (2004).

18. Greenwood, B. M. & Whittle, H. C. The pathogenesis of sleeping sickness. *Trans R Soc Trop Med Hyg* **74**, 716–725 (1980).

19. Brun, R., Don, R., Jacobs, R. T., Wang, M. Z. & Barrett, M. P. Development of novel drugs for human African trypanosomiasis. *Future Microbiology* vol. 6 677–691 Preprint at https://doi.org/10.2217/fmb.11.44 (2011).

20. Checchi, F., Filipe, J. A. N., Haydon, D. T., Chandramohan, D. & Chappuis, F. Estimates of the duration of the early and late stage of gambiense sleeping sickness. *BMC Infect Dis* **8**, 16 (2008).

21. Odiit, M., Kansiime, F. & Enyaru, J. C. K. Duration of symptoms and case fatality of sleeping sickness caused by Trypanosoma brucei rhodesiense in Tororo, Uganda. *East Afr Med J* **74**, 792–795 (1997).

22. Daneman, R. & Prat, A. The Blood–Brain Barrier. *Cold Spring Harb Perspect Biol* **7**, (2015).

23. Nikolskaia, O. V. *et al.* Blood-brain barrier traversal by African trypanosomes requires calcium signaling induced by parasite cysteine protease. *Journal of Clinical Investigation* **116**, 2739–2747 (2006).

24. Lumbala, C. *et al.* Human African trypanosomiasis in the Democratic Republic of the Congo: Disease distribution and risk. *Int J Health Geogr* **14**, (2015).

25. Jamonneau, V. *et al.* Untreated human infections by trypanosoma brucei gambiense are not 100% fatal. *PLoS Negl Trop Dis* **6**, (2012).

26. Bangs, J. D. Evolution of Antigenic Variation in African Trypanosomes: Variant Surface Glycoprotein Expression, Structure, and Function. *Bioessays* **40**, e1800181 (2018).

27. Jackson, A. P. *et al.* Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. *Proc Natl Acad Sci U S A* **109**, 3416–3421 (2012).

28. Horn, D. Antigenic variation in African trypanosomes. *Mol Biochem Parasitol* **195**, 123 (2014).

29. Dyer, N. A., Rose, C., Ejeh, N. O. & Acosta-Serrano, A. Flying tryps: Survival and maturation of trypanosomes in tsetse flies. *Trends in Parasitology* vol. 29 188–196 Preprint at https://doi.org/10.1016/j.pt.2013.02.003 (2013).

30. Franco, J. R., Simarro, P. P., Diarra, A. & Jannin, J. G. Epidemiology of human African trypanosomiasis. *Clinical Epidemiology* vol. 6 257–275 Preprint at https://doi.org/10.2147/CLEP.S39728 (2014).

31. Matthews, K. R. The developmental cell biology of Trypanosoma brucei. *J Cell Sci* **118**, 283–290 (2005).

32. Hocking, K. S., Lamerton, J. F. & Lewis, E. A. Tsetse-fly control and eradication. *Bulletin of the World Health Organisation* **28**, 811–823 (1963).

33. Kaba, D. *et al.* Use of vector control to protect people from sleeping sickness in the focus of bonon (Côte d'ivoire). *PLoS Negl Trop Dis* **15**, e0009404 (2021).

34. Muhanguzi, D. *et al.* The burden and spatial distribution of bovine African trypanosomes in small holder croplivestock production systems in Tororo District, south-eastern Uganda. *Parasit Vectors* **7**, 1–10 (2014).

35. Funk, S., Nishiura, H., Heesterbeek, H., Edmunds, W. J. & Checchi, F. Identifying Transmission Cycles at the Human-Animal Interface: The Role of Animal Reservoirs in Maintaining Gambiense Human African Trypanosomiasis. *PLoS Comput Biol* **9**, (2013).

36. Bonnet, J., Boudot, C. & Courtioux, B. Overview of the Diagnostic Methods Used in the Field for Human African Trypanosomiasis: What Could Change in the Next Years? *BioMed Research International* vol. 2015 Preprint at https://doi.org/10.1155/2015/583262 (2015).

37. Nok, A. J. Arsenicals (melarsoprol), pentamidine and suramin in the treatment of human African trypanosomiasis. *Parasitol Res* **90**, 71–79 (2003).

38. Kennedy, P. G. E. An alternative form of melarsoprol in sleeping sickness. *Trends in Parasitology* vol. 28 307–310 Preprint at https://doi.org/10.1016/j.pt.2012.05.003 (2012).

39. Doua, F., Miezan, T. W., Sanon Singaro, J. R., Boa Yapo, F. & Baltz, T. The efficacy of pentamidine in the treatment of early-late stage Trypanosoma brucei gambiense trypanosomiasis. *American Journal of Tropical Medicine and Hygiene* **55**, 586–588 (1996).

40. Doua, F. *et al.* Treatment of human late stage gambiense trypanosomiasis with α-difluoromethylornithine (eflornithine): Efficacy and tolerance in 14 cases in Cote d'Ivoire. *American Journal of Tropical Medicine and Hygiene* **37**, 525–533 (1987).

41. Priotto, G. *et al.* Nifurtimox-eflornithine combination therapy for second-stage Trypanosoma brucei gambiense sleeping sickness: A randomized clinical trial in Congo. *Clinical Infectious Diseases* **45**, 1435–1442 (2007).

42. Priotto, G. *et al.* Nifurtimox-eflornithine combination therapy for second-stage African Trypanosoma brucei gambiense trypanosomiasis: a multicentre, randomised, phase III, non-inferiority trial. *The Lancet* **374**, 56–64 (2009).

43. Pépin, J. *et al.* High-dose nifurtimox for arseno-resistant trypanosoma brucei gambiense sleeping sickness: An open trial in central zaire. *Trans R Soc Trop Med Hyg* **86**, 254–256 (1992).

44. Docampo, R. *et al.* Mechanism of nifurtimox toxicity in different forms of Trypanosoma cruzi. *Biochem Pharmacol* **30**, 1947–1951 (1981).

45. Lindner, A. K. *et al.* New WHO guidelines for treatment of gambiense human African trypanosomiasis including fexinidazole: substantial changes for clinical practice. *The Lancet Infectious Diseases* vol. 20 e38–e46 Preprint at https://doi.org/10.1016/S1473-3099(19)30612-7 (2020).

46. Nare, B. *et al.* Discovery of novel orally bioavailable oxaborole 6-carboxamides that demonstrate cure in a murine model of late-stage central nervous system African trypanosomiasis. *Antimicrob Agents Chemother* **54**, 4379–4388 (2010).

47. Betu Kumeso, V. K. *et al.* Efficacy and safety of acoziborole in patients with human African trypanosomiasis caused by Trypanosoma brucei gambiense: a multicentre, open-label, single-arm, phase 2/3 trial. *Lancet Infect Dis* **23**, 463–470 (2023).

48. Franco, J. R. *et al.* The elimination of human African trypanosomiasis: Achievements in relation to WHO road map targets for 2020. *PLoS Negl Trop Dis* **16**, e0010047 (2022).

49. World Health Organisation. Elimination of Human African Trypanosomiasis as public health problem. *Wkly. epidemiol. rec* **96**, 196–196 (2021).

50. Hide, G. History of sleeping sickness in East Africa. *Clinical Microbiology Reviews* vol. 12 112–125 Preprint at https://doi.org/10.1128/cmr.12.1.112 (1999).

51. Headrick, D. R. Sleeping Sickness Epidemics and Colonial Responses in East and Central Africa, 1900-1940. *PLoS Negl Trop Dis* **8**, (2014).

52. Smith, D. H., Pepin, J. & Stich, A. H. R. Human African trypanosomiasis: An emerging public health crisis. *British Medical Bulletin* vol. 54 341–355 Preprint at https://doi.org/10.1093/oxfordjournals.bmb.a011692 (1998).

53. Barrett, M. P. The elimination of human African trypanosomiasis is in sight: Report from the third WHO stakeholders meeting on elimination of gambiense human African trypanosomiasis. *PLoS Negl Trop Dis* **12**, e0006925–e0006925 (2018).

54. Franco, J. R. *et al.* Monitoring the elimination of human African trypanosomiasis: Update to 2014. *PLoS Negl Trop Dis* **11**, e0005585–e0005585 (2017).

55. Franco, J. R. *et al.* Monitoring the elimination of human African trypanosomiasis at continental and country level: Update to 2018. *PLoS Negl Trop Dis* **14**, 1–18 (2020).

56. World Health Organisation. *Report of the third WHO stakeholders meeting on rhodesiense human African trypanosomiasis, Geneva, Switzerland, 10–11 April 2019.* (2020).

57. Achcar, F., Kerkhoven, E. J. & Barrett, M. P. Trypanosoma brucei: Meet the system. *Current Opinion in Microbiology* vol. 20 162–169 Preprint at https://doi.org/10.1016/j.mib.2014.06.007 (2014).

58. Elenga, V. A. *et al.* Risk factors and prevalence of human African trypanosomiasis in individuals living in remote areas of the republic of Congo. *BMC Public Health* **22**, 1–9 (2022).

59. Dama, E. *et al.* Description of the first sleeping sickness case diagnosed in Burkina Faso since two decades. *PLoS Negl Trop Dis* **12**, (2018).

60. World Health Organisation. Trypanosomiasis, human African. *Special Programme for Research and Training in Tropical Diseases (TDR)* TDR Diseases and topics Trypanosomiasis, human Afr https://www.who.int/news-room/fact-sheets/detail/trypanosomiasis-human-african-(sleeping-sickness) (2019).

61. Kaiser, M. *et al.* Antitrypanosomal activity of fexinidazole, a new oral nitroimidazole drug candidate for treatment of sleeping sickness. *Antimicrob Agents Chemother* **55**, 5602–5608 (2011).

62. Babokhov, P., Sanyaolu, A. O., Oyibo, W. A., Fagbenro-Beyioku, A. F. & Iriemenam, N. C. A current analysis of chemotherapy strategies for the treatment of human African trypanosomiasis. *Pathogens and Global Health* vol. 107 242–252 Preprint at https://doi.org/10.1179/2047773213Y.0000000105 (2013).

63. Eperon, G. *et al.* Treatment options for second-stage gambiense human African trypanosomiasis. *Expert Review of Anti-Infective Therapy* vol. 12 1407–1417 Preprint at https://doi.org/10.1586/14787210.2014.959496 (2014).

64. Fairlamb, A. H. & Horn, D. Melarsoprol Resistance in African Trypanosomiasis. *Trends Parasitol* **34**, 481–492 (2018).

65. Bednarczyk, E. M. Desk Reference of Clinical Pharmacology, 2nd Edition. *Annals of Pharmacotherapy* **43**, 553–554 (2009).

66. Nzila, A. The past, present and future of antifolates in the treatment of Plasmodium falciparum infection. *Journal of Antimicrobial Chemotherapy* vol. 57 1043–1054 Preprint at https://doi.org/10.1093/jac/dkl104 (2006).

67. Bhargava, P. & Singh, R. Developments in diagnosis and antileishmanial drugs. *Interdiscip Perspect Infect Dis* **2012**, (2012).

68. Herrera-Acevedo, C. *et al.* Kaurane-Type Diterpenoids as Potential Inhibitors of Dihydrofolate Reductase-Thymidylate Synthase in New World Leishmania Species. *Antibiotics* **12**, 663 (2023).

69. Senkovich, O., Schormann, N. & Chattopadhyay, D. Structures of dihydrofolate reductase-thymidylate synthase of Trypanosoma cruzi in the folate-free state and in complex with two antifolate drugs, trimetrexate and methotrexate. *Acta Crystallogr D Biol Crystallogr* **65**, 704–716 (2009).

70. Teixeira, B. V. F. *et al.* Dual and selective inhibitors of pteridine reductase 1 (PTR1) and dihydrofolate reductase-thymidylate synthase (DHFR-TS) from Leishmania chagasi. *J Enzyme Inhib Med Chem* **34**, 1439 (2019).

71. Franco, J. R. *et al.* The elimination of human African trypanosomiasis: Achievements in relation to WHO road map targets for 2020. *PLoS Negl Trop Dis* **16**, e0010047 (2022).

72. Bishop, A., De Beer, T. A. P. & Joubert, F. Protein homology modelling and its use in South Africa. *S Afr J Sci* **104**, 2–6 (2008).

73. Hillisch, A., Pineda, L. F. & Hilgenfeld, R. Utility of homology models in the drug discovery process. *Drug Discovery Today* vol. 9 659–669 Preprint at https://doi.org/10.1016/S1359-6446(04)03196-4 (2004).

74. Floudas, C. A., Fung, H. K., McAllister, S. R., Mönnigmann, M. & Rajgaria, R. Advances in protein structure prediction and de novo protein design: A review. *Chem Eng Sci* **61**, 966–988 (2006).

75. Lee, J., Freddolino, P. L. & Zhang, Y. Ab initio protein structure prediction. in *From Protein Structure to Function with Bioinformatics: Second Edition* 3–35 (2017). doi:10.1007/978-94-024-1069-3_1.

76. Bowie, J. U., Lüthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional stucture. *Science (1979)* **253**, 164–170 (1991).

77. Jones, D. T., Taylort, W. R. & Thornton, J. M. A new approach to protein fold recognition. *Nature* **358**, 86–89 (1992).

78. Zhou, H. & Zhou, Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins: Structure, Function and Genetics* **58**, 321–328 (2005).

79. Cavasotto, C. N. & Phatak, S. S. Homology modeling in drug discovery: current trends and applications. *Drug Discov Today* **14**, 676–683 (2009).

80. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence - A study of structural response in protein cores. *Proteins: Structure, Function and Bioinformatics* **77**, 499–508 (2009).

81. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**, 823–826 (1986).

82. Sander, C. & Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics* **9**, 56–68 (1991).

83. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

84. Zimmermann, L. *et al.* A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core. *J Mol Biol* **430**, 2237–2243 (2018).

85. Tramontano, A. Homology modeling with low sequence identity. *Methods: A Companion to Methods in Enzymology* **14**, 293–300 (1998).

86. Bender, B. J., Marlow, B. & Meiler, J. Improving homology modeling from low-sequence identity templates in Rosetta: A case study in GPCRs. *PLoS Comput Biol* **16**, e1007597 (2020).

87. Khor, B. Y., Tye, G. J., Lim, T. S. & Choong, Y. S. General overview on structure prediction of twilight-zone proteins. *Theor Biol Med Model* **12**, 1–11 (2015).

88. Nikolaev, D. M. *et al.* A Comparative Study of Modern Homology Modeling Algorithms for Rhodopsin Structure Prediction. *ACS Omega* **3**, 7555–7566 (2018).

89. Boscott, P. E., Barton, G. J. & Richards, W. G. Secondary structure prediction for modelling by homology. *Protein Engineering, Design and Selection* **6**, 261–266 (1993).

90. Mautsa, N., Prinsloo, E., Bishop, Ö. T. & Blatch, G. L. The PINIT domain of PIAS3: Structure-function analysis of its interaction with STAT3. *Journal of Molecular Recognition* **24**, 795–803 (2011).

91. Krieger, E., Nabuurs, S. B. & Vriend, G. Homology modeling. *Methods Biochem Anal* **44**, 509–524 (2003).

92. Dunbrack, R. L. Sequence comparison and protein structure prediction. *Current Opinion in Structural Biology* vol. 16 374–384 Preprint at https://doi.org/10.1016/j.sbi.2006.05.006 (2006).

93. Jones, D. T. Progress in protein structure prediction. *Curr Opin Struct Biol* **7**, 377–387 (1997).

94. Fiser, A. Template-based protein structure modeling. *Methods in molecular biology (Clifton, N.J.)* vol. 673 73–94 Preprint at https://doi.org/10.1007/978-1-60761-842-3_6 (2010).

95. Papaleo, E. *et al.* The role of protein loops and linkers in conformational dynamics and allostery. *Chem Rev* **116**, 6391–6423 (2016).

96. Liang, S. & Grishin, N. V. Side-chain modeling with an optimized scoring function. *Protein Science* **11**, 322–331 (2002).

97. Mendes, J., Baptista, A. M., Carrondo, M. A. & Soares, C. M. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins: Structure, Function, and Bioinformatics* **37**, 530–543 (1999).

98. Eswar, N. *et al.* Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* **15**, 5–6 (2006).

99. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* (2018).

100. Wang, S., Li, W., Liu, S. & Xu, J. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res* **44**, W430–W435 (2016).

101. Kim, D. E., Chivian, D. & Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* **32**, W526 (2004).

102. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

103. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature 2021 596:7873* **596**, 590–596 (2021).

104. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

105. Bertoline, L. M. F., Lima, A. N., Krieger, J. E. & Teixeira, S. K. Before and after AlphaFold2: An overview of protein structure prediction. *Frontiers in Bioinformatics* **3**, 1120370 (2023).

106. Xiang, Z. Advances in homology protein structure modeling. *Curr Protein Pept Sci* **7**, 217–227 (2006).

107. Karami, Y. *et al.* DaReUS-Loop: a web server to model multiple loops in homology models. *Nucleic Acids Res* **47**, W423–W428 (2019).

108. Xiang, Z., Soto, C. S. & Honig, B. Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci U S A* **99**, 7432–7437 (2002).

109. Fiser, A. & Sali, A. ModLoop: Automated modeling of loops in protein structures. *Bioinformatics* **19**, 2500–2501 (2003).

110. Park, H., Ovchinnikov, S., Kim, D. E., DiMaio, F. & Baker, D. Protein homology model refinement by large-scale energy optimization. *Proceedings of the National Academy of Sciences* 201719115 (2018).

111. Sippl, M. J. Knowledge-based potentials for proteins. *Curr Opin Struct Biol* **5**, 229–235 (1995).

112. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* **35**, W407--W410 (2007).

113. Colovos, C. & Yeates, T. O. Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science* **2**, 1511–1519 (1993).

114. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* **26**, 283–291 (1993).

115. Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. PROCHECK : validation of protein-structure coordinates. in 684–687 (2012). doi:10.1107/97809553602060000882.

116. Pontius, J., Richelle, J. & Wodak, S. J. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J Mol Biol* **264**, 121–136 (1996).

117. Hooft, R. W. W. W., Vriend, G., Sander, C. & Abola, E. E. Errors in protein structures. *Nature* **381**, 272 (1996).

118. Eisenberg, D., Lüthy, R. & Bowie, J. U. VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods Enzymol* **277**, 396–404 (1997).

119. Bowie, J. U., Lüthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170 (1991).

120. Rost, B., Schneider, R. & Sander, C. Protein fold recognition by prediction-based threading. *J Mol Biol* **270**, 471–480 (1997).

121. Lu, L., Lu, H. & Skolnick, J. Multiprospector: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins: Structure, Function and Genetics* **49**, 350–364 (2002).

122. Skolnick, J., Kihara, D. & Zhang, Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins: Structure, Function and Genetics* **56**, 502–518 (2004).

123. Bryant, S. H. & Lawrence, C. E. An empirical energy function for threading protein sequence through the folding motif. *Proteins: Structure, Function, and Bioinformatics* **16**, 92–112 (1993).

124. Ayers, D. J., Huber, T. & Torda, A. E. Protein fold recognition score functions: unusual construction strategies. *Proteins* **36**, 454–461 (1999).

125. Mendell, J. T. & Dietz, H. C. When the message goes awry: disease-producing mutations that influence mRNA content and performance. *Cell* **107**, 411–414 (2001).

126. Riva, A. & Kohane, I. S. SNPper: retrieval and analysis of human SNPs. *Bioinformatics* **18**, 1681–1685 (2002).

127. Eichler, E. E. *et al.* Completing the map of human genetic variation. *Nature* **447**, 161–165 (2007).

128. Zou, H., Wu, L. X., Tan, L., Shang, F. F. & Zhou, H. H. Significance of Single-Nucleotide Variants in Long Intergenic Non-protein Coding RNAs. *Frontiers in Cell and Developmental Biology* vol. 8 347 Preprint at https://doi.org/10.3389/fcell.2020.00347 (2020).

129. Mitra, A. K. *et al.* Protective association exhibited by the single nucleotide polymorphism (SNP) rs1052133 in the gene human 8- oxoguanine DNA glycosylase (hOGG1) with the risk of squamous cell carcinomas of the head & neck (SCCHN) among north Indians. *Indian J Med Res* **133**, 605 (2011).

130. Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101**, 5–22 (2017).

131. Schrödinger, L. and W. D. PyMOL. Preprint at https://pymol.org/2/ (2020).

132. BIOVIA, D. S. Discovery Studio. *Discovery Studio Modeling Environment: San Diego, CA, USA* Preprint at (2019).

133. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605–1612 (2004).

134. Hillisch, A. & Hilgenfeld, R. The role of protein 3D-structures in the drug discovery process. *EXS* 157–181 Preprint at https://doi.org/10.1007/978-3-0348-7997-2_8 (2003).

135. Moitessier, N. *et al.* Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br J Pharmacol* **153**, S7--S26 (2008).

136. Zhang, B., Li, H., Yu, K. & Jin, Z. Molecular docking-based computational platform for high-throughput virtual screening. *CCF Transactions on High Performance Computing* **4**, 63–74 (2022).

137. Hetényi, C. & van der Spoel, D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. *FEBS Lett* **580**, 1447–1450 (2006).

138. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* **53**, 2719–2740 (2010).

139. Baell, J. & Walters, M. A. Chemistry: Chemical con artists foil drug discovery. *Nature* vol. 513 481–483 Preprint at https://doi.org/10.1038/513481a (2014).

140. Baell, J. B. & Nissink, J. W. M. Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017 - Utility and Limitations. *ACS Chemical Biology* vol. 13 36–44 Preprint at https://doi.org/10.1021/acschembio.7b00903 (2018).

141. Pardridge, W. M. The blood-brain barrier: Bottleneck in brain drug development. *NeuroRx* **2**, 3–14 (2005).

142. Ballabh, P., Braun, A. & Nedergaard, M. The blood-brain barrier: An overview: Structure, regulation, and clinical implications. *Neurobiology of Disease* vol. 16 1–13 Preprint at https://doi.org/10.1016/j.nbd.2003.12.016 (2004).

143. Di, L. & Kerns, E. H. *Blood-Brain Barrier in Drug Discovery: Optimizing Brain Exposure of CNS Drugs and Minimizing Brain Side Effects for Peripheral Drugs*. *Blood-Brain Barrier in Drug Discovery: Optimizing Brain Exposure of CNS Drugs and Minimizing Brain Side Effects for Peripheral Drugs* (wiley, 2015). doi:10.1002/9781118788523.

144. Banks, W. A. Characteristics of compounds that cross the blood-brain barrier. in *BMC Neurology* vol. 9 1–5 (BioMed Central, 2009).

145. Pardridge, W. M. Drug transport across the blood-brain barrier. *Journal of Cerebral Blood Flow and Metabolism* vol. 32 1959–1972 Preprint at https://doi.org/10.1038/jcbfm.2012.126 (2012).

146. Ferreira, L. G., Dos Santos, R. N., Oliva, G. & Andricopulo, A. D. Molecular docking and structure-based drug design strategies. *Molecules* vol. 20 13384–13421 Preprint at https://doi.org/10.3390/molecules200713384 (2015).

147. Koshland Jr, D. E. The key--lock theory and the induced fit theory. *Angewandte Chemie International Edition in English* **33**, 2375–2378 (1995).

148. Hammes, G. G. Multiple conformational changes in enzyme catalysis. *Biochemistry* **41**, 8221–8228 (2002).

149. Berry, M., Fielding, B. & Gamieldien, J. Practical considerations in virtual screening and molecular docking. *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology; Tran, QN, Hamid, AR, Eds* 487–502 (2015).

150. Morris, G. M. *et al.* Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**, 1639–1662 (1998).

151. Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **261**, 470–489 (1996).

152. Huang, S. Y. & Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J Comput Chem* **27**, 1876–1882 (2006).

153. Huang, S. Y. & Zou, X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J Comput Chem* **27**, 1866–1875 (2006).

154. Ma, Z. & Zou, X. MDock: A Suite for Molecular Inverse Docking and Target Prediction. in *Methods in Molecular Biology* vol. 2266 313–322 (Methods Mol Biol, 2021).

155. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J Comput Chem* **30**, 2785 (2009).

156. Böhm, H. J. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des* **6**, 593–606 (1992).

157. Najmanovich, R., Kuttner, J., Sobolev, V. & Edelman, M. Side-chain flexibility in proteins upon ligand binding. *Proteins: Structure, Function and Genetics* **39**, 261–268 (2000).

158. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. Ligplot: A program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering, Design and Selection* **8**, 127–134 (1995).

159. Laskowski, R. A. & Swindells, M. B. LigPlot+: Multiple ligand-protein interaction diagrams for drug discovery. *J Chem Inf Model* **51**, 2778–2786 (2011).

160. Jubb, H. C. *et al.* Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J Mol Biol* **429**, 365–371 (2017).

161. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res* **45**, D945--D954 (2016).

162. Irwin, J. J. & Shoichet, B. K. ZINC- A free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**, 177–182 (2005).

163. Kumar, A. *et al.* PfalDB: An Integrated Drug Target and Chemical Database for Plasmodium flaciparum. *Curr Drug Targets* **15**, 1089–1093 (2014).

164. Diallo, B. N., Glenister, M., Musyoka, T. M., Lobb, K. & Tastan Bishop, Ö. SANCDB: an update on South African natural compounds and their readily available analogs. *J Cheminform* **13**, 1–14 (2021).

165. Hatherley, R. *et al.* SANCDB: A South African natural compound database. *J Cheminform* **7**, (2015).

166. Ntie-Kang, F. *et al.* ConMedNP: a natural product library from Central African medicinal plants for drug discovery. *RSC Adv* **4**, 409–419 (2014).

167. Chen, C. Y.-C. TCM Database@ Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One* **6**, e15939 (2011).

168. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**, D1074–D1082 (2018).

169. Griffen, E. J., Dossetter, A. G., Leach, A. G. & Montague, S. Can we accelerate medicinal chemistry by augmenting the chemist with Big Data and artificial intelligence? *Drug Discovery Today* vol. 23 1373–1384 Preprint at https://doi.org/10.1016/j.drudis.2018.03.011 (2018).

170. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics* vol. 29 476–488 Preprint at https://doi.org/10.1002/minf.201000061 (2010).

171. Changeux, J. P. & Edelstein, S. Conformational selection or induced fit? 50 Years of debate resolved. *F1000 Biology Reports* vol. 3 Preprint at https://doi.org/10.3410/B3-19 (2011).

172. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).

173. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**, 1731–1737 (2003).

174. Pierce, B. G. *et al.* ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771 (2014).

175. Kozakov, D. *et al.* The ClusPro web server for protein–protein docking. *Nature Protocols 2017 12:2* **12**, 255–278 (2017).

176. Yan, Y., Tao, H., He, J. & Huang, S. Y. The HDOCK server for integrated protein–protein docking. *Nature Protocols 2020 15:5* **15**, 1829–1852 (2020).

177. Williamson, M. P. Using chemical shift perturbation to characterise ligand binding. *Progress in Nuclear Magnetic Resonance Spectroscopy* vol. 73 1–16 Preprint at https://doi.org/10.1016/j.pnmrs.2013.02.001 (2013).

178. Konermann, L., Pan, J. & Liu, Y. H. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem Soc Rev* **40**, 1224–1234 (2011).

179. Robert, F. J. M. & Editors, J. M. P. *Computer Simulations of Protein Structures and Interactions*. (Springer-Verlag, 1995).

180. Hoffmann, D. *et al.* Two-stage method for protein-ligand docking. *J Med Chem* **42**, 4422–4433 (1999).

181. Trosset, J. Y. & Scheraga, H. A. Reaching the global minimum in docking simulations: A Monte Carlo energy minimization approach using Bezier splines. *Proc Natl Acad Sci U S A* **95**, 8011–8015 (1998).

182. De Vries, S. J., Van Dijk, M. & Bonvin, A. M. J. J. The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc* **5**, 883–897 (2010).

183. Shukla, D., Hernández, C. X., Weber, J. K. & Pande, V. S. Markov state models provide insights into dynamic modulation of protein function. *Accounts of Chemical Research* vol. 48 414–422 Preprint at https://doi.org/10.1021/ar5002999 (2015).

184. Pais, F. S.-M., de Cássia Ruy, P., Oliveira, G. & Coimbra, R. S. Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology* **9**, 4 (2014).

185. Musyoka, T. M., Kanzi, A. M., Lobb, K. A. & Bishop, Ö. T. Structure based docking and molecular dynamic studies of plasmodial cysteine proteases against a South African natural compound and its analogs. *Sci Rep* **6**, 23690 (2016).

186. Raval, A., Piana, S., Eastwood, M. P., Dror, R. O. & Shaw, D. E. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics* **80**, 2071–2079 (2012).

187. Brown, D. K. & Bishop, Ö. T. Role of structural bioinformatics in drug discovery by computational SNP analysis: analyzing variation at the protein level. *Glob Heart* **12**, 151–161 (2017).

188. Penkler, D. L., Atilgan, C. & Tastan Bishop, O. Allosteric Modulation of Human Hsp90α Conformational Dynamics. *J Chem Inf Model* **58**, 383–404 (2018).

189. Scheraga, H. A., Khalili, M. & Liwo, A. Protein-folding dynamics: overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.* **58**, 57–83 (2007).

190. Hansson, T., Oostenbrink, C. & van Gunsteren, W. Molecular dynamics simulations. *Curr Opin Struct Biol* **12**, 190–196 (2002).

191. Case, D. A. *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668–1688 (2005).

192. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J Comput Chem* **30**, 1545–1614 (2009).

193. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).

194. Nelson, M. T. *et al.* NAMD: a parallel, object-oriented molecular dynamics program. *The International Journal of Supercomputer Applications and High Performance Computing* **10**, 251–268 (1996).

195. Bashford, D. & Case, D. A. Generalized born models of macromolecular solvation effects. *Annu Rev Phys Chem* **51**, 129–152 (2000).

196. Kufareva, I. & Abagyan, R. Methods of protein structure comparison. in *Homology Modeling* vol. 857 231–257 (Springer, 2011).

197. Orozco, M. & Luque, F. J. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem Rev* **100**, 4187–4226 (2000).

198. Berendsen, H. J. C. Dynamic simulation as an essential tool in molecular modeling. *J Comput Aided Mol Des* **2**, 217–221 (1988).

199. Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J Chem Phys* **72**, 2384–2393 (1980).

200. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J Chem Phys* **81**, 3684–3690 (1984).

201. Hünenberger, P. H. Thermostat algorithms for molecular dynamics simulations. *Advances in Polymer Science* vol. 173 105–147 Preprint at https://doi.org/10.1007/b99427 (2005).

202. Martyna, G. J., Tobias, D. J. & Klein, M. L. Constant pressure molecular dynamics algorithms. *J Chem Phys* **101**, 4177–4189 (1994).

203. Van Der Spoel, D. *et al.* GROMACS: fast, flexible, and free. *J Comput Chem* **26**, 1701–1718 (2005).

204. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J Chem Phys* **103**, 8577–8593 (1995).

205. Abraham, M. J. & Gready, J. E. Optimization of parameters for molecular dynamics simulation using smooth particle-mesh Ewald in GROMACS 4.5. *J Comput Chem* **32**, 2031–2040 (2011).

206. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for molecular simulations. *J Comput Chem* **18**, 1463–1472 (1997).

207. GrubmüLler, H., Heller, H., Windemuth, A. & Schulten, K. Generalized verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Mol Simul* **6**, 121–142 (1991).

208. Truhlar, D. G. & McKoy, V. Computational chemistry. *Comput Sci Eng* **2**, 19–21 (2000).

209. Lennard-Jones, J. E. On the determination of molecular fields. —II. From the equation of state of a gas. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **106**, 463–477 (1924).

210. González, M. A. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique* **12**, 169–200 (2011).

211. Ponder, J. W. & Case, D. A. Force fields for protein simulations. *Adv Protein Chem* **66**, 27–85 (2003).

212. Vanommeslaeghe, K. *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* **31**, 671–690 (2010).

213. Christen, M. *et al.* The GROMOS software for biomolecular simulation: GROMOS05. *J Comput Chem* **26**, 1719–1751 (2005).

214. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* **118**, 11225–11236 (1996).

215. Jorgensen, W. L. & Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences of the United States of America* vol. 102 6665–6670 Preprint at https://doi.org/10.1073/pnas.0408037102 (2005).

216. 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* **7**, 146–157 (2011).

217. González, M. A. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique* **12**, 169–200 (2011).

218. Martín-García, F., Papaleo, E., Gomez-Puertas, P., Boomsma, W. & Lindorff-Larsen, K. Comparing molecular dynamics force fields in the essential subspace. *PLoS One* **10**, e0121114 (2015).

219. Khairallah, A., Tastan Bishop, Ö. & Moses, V. AMBER force field parameters for the Zn (II) ions of the tunneling-fold enzymes GTP cyclohydrolase I and 6-pyruvoyl tetrahydropterin synthase. *J Biomol Struct Dyn* **39**, 5843–5860 (2021).

220. Musyoka, T., Tastan Bishop, Ö., Lobb, K. & Moses, V. The determination of CHARMM force field parameters for the Mg2+ containing HIV-1 integrase. *Chem Phys Lett* **711**, 1–7 (2018).

221. Tendwa, M. B., Chebon-bore, L., Lobb, K., Musyoka, T. M. & Tastan Bishop, Ö. Force field parameters for fe2+ 4s2−4 clusters of dihydropyrimidine dehydrogenase, the 5-fluorouracil cancer drug deactivation protein: A step towards in silico pharmacogenomics studies. *Molecules* **26**, 2929 (2021).

222. da Silva, A. W. S. & Vranken, W. F. ACPYPE-Antechamber python parser interface. *BMC Res Notes* **5**, 367 (2012).

223. Schüttelkopf, A. W. & Van Aalten, D. M. F. PRODRG: a tool for high-throughput crystallography of protein--ligand complexes. *Acta Crystallogr D Biol Crystallogr* **60**, 1355–1363 (2004).

224. Ribeiro, A. A. S. T., Horta, B. A. C. & Alencastro, R. B. de. MKTOP: a program for automatic construction of molecular topologies. *J Braz Chem Soc* **19**, 1433–1435 (2008).

225. Malde, A. K. *et al.* An automated force field topology builder (ATB) and repository: version 1.0. *J Chem Theory Comput* **7**, 4026–4037 (2011).

226. Sargsyan, K., Grauffel, C. & Lim, C. How molecular size impacts RMSD applications in molecular dynamics simulations. *J Chem Theory Comput* **13**, 1518–1524 (2017).

227. Lobanov, M. I., Bogatyreva, N. S. & Galzitskaia, O. V. Radius of gyration is indicator of compactness of protein structure. *Molekuliarnaia biologiia* **42**, 701–706 (2008).

228. Halle, B. Flexibility and packing in proteins. *Proc Natl Acad Sci U S A* **99**, 1274–1279 (2002).

229. Richards, F. M. Protein stability: Still an unsolved problem. *Cellular and Molecular Life Sciences* vol. 53 790–802 Preprint at https://doi.org/10.1007/s000180050100 (1997).

230. Wei, G., Xi, W., Nussinov, R. & Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? the Diverse Functional Roles of Conformational Ensembles in the Cell. *Chemical Reviews* vol. 116 6516–6551 Preprint at https://doi.org/10.1021/acs.chemrev.5b00562 (2016).

231. Miller, M. D. & Phillips, G. N. Moving beyond static snapshots: Protein dynamics and the Protein Data Bank. *Journal of Biological Chemistry* vol. 296 Preprint at https://doi.org/10.1016/j.jbc.2021.100749 (2021).

232. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* vol. 450 964–972 Preprint at https://doi.org/10.1038/nature06522 (2007).

233. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. *Science (1979)* **254**, 1598–1603 (1991).

234. Kurakin, A. Scale-free flow of life: On the biology, economics, and physics of the cell. *Theor Biol Med Model* **6**, (2009).

235. Latorraca, N. R., Venkatakrishnan, A. J. & Dror, R. O. GPCR dynamics: Structures in motion. *Chemical Reviews* vol. 117 139–155 Preprint at https://doi.org/10.1021/acs.chemrev.6b00177 (2017).

236. Knapp, B., Frantal, S., Cibena, M., Schreiner, W. & Bauer, P. Is an intuitive convergence definition of molecular dynamics simulations solely based on the root mean square deviation possible? *Journal of Computational Biology* **18**, 997–1005 (2011).

237. Sheik Amamuddy, O., Glenister, M., Tshabalala, T. & Tastan Bishop, Ö. MDM-TASK-web: MD-TASK and MODE-TASK web server for analyzing protein dynamics. *Comput Struct Biotechnol J* **19**, 5059–5071 (2021).

238. Release, S. Maestro , Schrödinger. *LLC, New York, NY,* Preprint at https://www.schrodinger.com/maestro (2020).

239. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J Mol Graph* **14**, 33–38 (1996).

240. Brown, D. K. *et al.* MD-TASK: a software suite for analyzing molecular dynamics trajectories. *Bioinformatics* **33**, 2768–2771 (2017).

241. Golbeck, J. *Analyzing the Social Web*. *Analyzing the Social Web* (Elsevier, 2013). doi:10.1016/C2012-0-00171-8.

242. Negre, C. F. A. *et al.* Eigenvector centrality for characterization of protein allosteric pathways. *Proc Natl Acad Sci U S A* **115**, E12201–E12208 (2018).

243. Brandes, U. On variants of shortest-path betweenness centrality and their generic computation. *Soc Networks* **30**, 136–145 (2008).

244. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. in *7th Python in Science Conference (SciPy 2008)* 11–15 (2008).

245. Freeman, L. C. Centrality in social networks conceptual clarification. *Soc Networks* **1**, 215–239 (1978).

246. Sharkey, K. J. A control analysis perspective on Katz centrality. *Sci Rep* **7**, 1–8 (2017).

247. Newman, M. *Networks: A Introduction*. *Oxford* vol. 148 (2018).

248. Bonacich, P. Power and Centrality: A Family of Measures. *American Journal of Sociology* **92**, 1170–1182 (1987).

249. Atilgan, A. R., Akan, P. & Baysal, C. Small-World Communication of Residues and Significance for Protein Dynamics. *Biophys J* **86**, 85–91 (2004).

250. Betz, S. F. Disulfide bonds and the stability of globular proteins. *Protein Science* vol. 2 1551–1558 Preprint at https://doi.org/10.1002/pro.5560021002 (1993).

251. Wedemeyer, W. J., Welker, E., Narayan, M. & Scheraga*, H. A. Disulfide bonds and protein folding. *Biochemistry* vol. 39 7032 Preprint at https://doi.org/10.1021/bi005111p (2000).

252. Manteca, A. *et al.* The influence of disulfide bonds on the mechanical stability of proteins is context dependent. *Journal of Biological Chemistry* **292**, 13374–13380 (2017).

253. Thornton, J. M. Disulphide bridges in globular proteins. *J Mol Biol* **151**, 261–287 (1981).

254. Hubbard, R. E. & Kamran Haider, M. Hydrogen Bonds in Proteins: Role and Strength. in *eLS* (John Wiley & Sons, Ltd, 2010). doi:10.1002/9780470015902.a0003011.pub2.

255. Mallamace, D., Fazio, E., Mallamace, F. & Corsaro, C. The role of hydrogen bonding in the folding/unfolding process of hydrated lysozyme: A review of recent NMR and FTIR results. *International Journal of Molecular Sciences* vol. 19 Preprint at https://doi.org/10.3390/ijms19123825 (2018).

256. Pace, C. N. *et al.* Contribution of hydrogen bonds to protein stability. *Protein Science* **23**, 652–661 (2014).

257. Chen, D. *et al.* Regulation of protein-ligand binding affinity by hydrogen bond pairing. *Sci Adv* **2**, (2016).

258. Ferreira De Freitas, R. & Schapira, M. A systematic analysis of atomic protein-ligand interactions in the PDB. *Medchemcomm* **8**, 1970–1981 (2017).

259. Huang, Y. M. M., Kang, M. & Chang, C. E. A. Switches of hydrogen bonds during ligand-protein association processes determine binding kinetics. *Journal of Molecular Recognition* **27**, 537–548 (2014).

260. Bitencourt-Ferreira, G., Veit-Acosta, M. & de Azevedo, W. F. Hydrogen bonds in protein-ligand complexes. in *Methods in Molecular Biology* vol. 2053 93–107 (Methods Mol Biol, 2019).

261. Bulusu, G. & Desiraju, G. R. Strong and Weak Hydrogen Bonds in Protein–Ligand Recognition. *Journal of the Indian Institute of Science* vol. 100 31–41 Preprint at https://doi.org/10.1007/s41745-019-00141-9 (2020).

262. Sigala, P. A. *et al.* Determination of hydrogen bond structure in water versus aprotic environments to test the relationship between length and stability. *J Am Chem Soc* **137**, 5730–5740 (2015).

263. Sun, M. an *et al.* Prediction of reversible disulfide based on features from local structural signatures. *BMC Genomics* **18**, (2017).

264. Herschlag, D. & Pinney, M. M. Hydrogen Bonds: Simple after All? *Biochemistry* vol. 57 3338–3352 Preprint at https://doi.org/10.1021/acs.biochem.8b00217 (2018).

265. Case, D. A. *et al.* Amber 2021. Preprint at http://ambermd.org/contributors.html (2021).

266. Pays, E. & Vanhollebeke, B. Human innate immunity against African trypanosomes. *Current Opinion in Immunology* vol. 21 493–498 Preprint at https://doi.org/10.1016/j.coi.2009.05.024 (2009).

267. Hawking, F. Trypanocidal substances in human serum. *Trans R Soc Trop Med Hyg* **67**, 262 (1973).

268. Hawking, F., Ramsden, D. B. & Whytock, S. The trypanocidal action of human serum and of baboon plasma. *Trans R Soc Trop Med Hyg* **67**, 501–516 (1973).

269. Bishop, J. R., Shimamura, M. & Hajduk, S. L. Insight into the mechanism of trypanosome lytic factor-1 killing of Trypanosoma brucei brucei. *Mol Biochem Parasitol* **118**, 33–40 (2001).

270. Hajduk, S. L. *et al.* Lysis of Trypanosoma brucei by a toxic subspecies of human high density lipoprotein. *Journal of Biological Chemistry* **264**, 5210–5217 (1989).

271. Smith, A. B., Esko, J. D. & Hajduk, S. L. Killing of trypanosomes by the human haptoglobin-related protein. *Science (1979)* **268**, 284–286 (1995).

272. Greene, A. S. & Hajduk, S. L. Trypanosome lytic factor-1 initiates oxidation-stimulated osmotic lysis of trypanosoma brucei brucei. *Journal of Biological Chemistry* **291**, 3063–3075 (2016).

273. Pérez-Morga, D. *et al.* Microbiology: Apolipoprotein L-I promotes trypanosome lysis by forming pores in lysosomal membranes. *Science (1979)* **309**, 469–472 (2005).

274. Uzureau, S. *et al.* APOL1 C-Terminal Variants May Trigger Kidney Disease through Interference with APOL3 Control of Actomyosin. *Cell Rep* **30**, 3821-3836.e13 (2020).

275. Molina-Portela, M. D. P., Lugli, E. B., Recio-Pinto, E. & Raper, J. Trypanosome lytic factor, a subclass of high-density lipoprotein, forms cation-selective pores in membranes. *Mol Biochem Parasitol* **144**, 218–226 (2005).

276. Thomson, R. & Finkelstein, A. Human trypanolytic factor APOL1 forms pH-gated cation-selective channels in planar lipid bilayers: Relevance to trypanosome lysis. *Proc Natl Acad Sci U S A* **112**, 2894–2899 (2015).

277. Vanhollebeke, B. & Pays, E. The function of apolipoproteins L. *Cellular and Molecular Life Sciences* **63**, 1937–1944 (2006).

278. Vanwalleghem, G. *et al.* Coupling of lysosomal and mitochondrial membrane permeabilization in trypanolysis by APOL1. *Nat Commun* **6**, 1–10 (2015).

279. Page, N. M., Butlin, D. J., Lomthaisong, K. & Lowry, P. J. The human apolipoprotein L gene cluster: Identification, classification, and sites of distribution. *Genomics* **74**, 71–78 (2001).

280. Lecordier, L. *et al.* C-terminal mutants of apolipoprotein L-I efficiently kill both Trypanosoma brucei brucei and Trypanosoma brucei rhodesiense. *PLoS Pathog* **5**, (2009).

281. Pays, E., Vanhollebeke, B., Uzureau, P., Lecordier, L. & Pérez-Morga, D. The molecular arms race between African trypanosomes and humans. *Nature Reviews Microbiology* vol. 12 575–584 Preprint at https://doi.org/10.1038/nrmicro3298 (2014).

282. Schaub, C. *et al.* Coiled-coil binding of the leucine zipper domains of APOL1 is necessary for the open cation channel conformation. *Journal of Biological Chemistry* **297**, (2021).

283. Vanhamme, L. *et al.* Apolipoprotein L-I is the trypanosome lytic factor of human serum. *Nature* **422**, 83–87 (2003).

284. Sharma, A. K., Friedman, D. J., Pollak, M. R. & Alper, S. L. Structural characterization of the C-terminal coiled-coil domains of wild-type and kidney disease-associated mutants of apolipoprotein L1. *FEBS Journal* **283**, 1846–1862 (2016).

285. Kopp, J. B. *et al.* APOL1 genetic variants in focal segmental glomerulosclerosis and HIV-associated nephropathy. *Journal of the American Society of Nephrology* **22**, 2129–2137 (2011).

286. Duran, C. E. *et al.* Prevalence of APOL1 Risk Variants in Afro-Descendant Patients with Chronic Kidney Disease in a Latin American Country. *Int J Nephrol* **2019**, (2019).

287. Limou, S., Nelson, G. W., Kopp, J. B. & Winkler, C. A. APOL1 Kidney Risk Alleles: Population Genetics and Disease Associations. *Advances in Chronic Kidney Disease* vol. 21 426–433 Preprint at https://doi.org/10.1053/j.ackd.2014.06.005 (2014).

288. Nadkarni, G. N. *et al.* Worldwide Frequencies of APOL1 Renal Risk Variants. *New England Journal of Medicine* **379**, 2571–2572 (2018).

289. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science (1979)* **324**, 1035–1044 (2009).

290. Ulasi, I. I. *et al.* High Population Frequencies of APOL1 Risk Variants Are Associated with Increased Prevalence of Non-Diabetic Chronic Kidney Disease in the Igbo People from South-Eastern Nigeria. *Nephron Clin Pract* **123**, 123–128 (2013).

291. Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science (1979)* **329**, 841–845 (2010).

292. Tzur, S. *et al.* Missense mutations in the APOL1 gene are highly associated with end stage kidney disease risk previously attributed to the MYH9 gene. *Hum Genet* **128**, 345–350 (2010).

293. Capewell, P., Cooper, A., Clucas, C., Weir, W. & MacLeod, A. A co-evolutionary arms race: Trypanosomes shaping the human genome, humans shaping the trypanosome genome. *Parasitology* **142**, S108–S119 (2015).

294. Thomson, R. *et al.* Evolution of the primate trypanolytic factor APOL1. *Proc Natl Acad Sci U S A* **111**, E2130–E2139 (2014).

295. Madhavan, S. M., Hansen, A. L., Cao, S., Sedor, J. R. & Buck, M. Towards the NMR solution Structure and the Dynamics of the C-terminal Region of APOL1 and its G1, G2 Variants with a Membrane Mimetic. *bioRxiv* 2021.03.16.435683 (2021) doi:10.1101/2021.03.16.435683.

296. Daneshpajouhnejad, P., Kopp, J. B., Winkler, C. A. & Rosenberg, A. Z. The evolving story of apolipoprotein L1 nephropathy: the end of the beginning. *Nat Rev Nephrol* **18**, 307–320 (2022).

297. Cuypers, B. *et al.* Apolipoprotein L1 variant associated with increased susceptibility to trypanosome infection. *mBio* **7**, (2016).

298. Lannon, H. *et al.* Apolipoprotein L1 (APOL1) risk variant toxicity depends on the haplotype background. *Kidney Int* **96**, 1303–1307 (2019).

299. Frueh, D. P., Goodrich, A. C., Mishra, S. H. & Nichols, S. R. NMR methods for structural studies of large monomeric and multimeric proteins. *Current Opinion in Structural Biology* vol. 23 734–739 Preprint at https://doi.org/10.1016/j.sbi.2013.06.016 (2013).

300. Yu, H. Extending the size limit of protein nuclear magnetic resonance. *Proceedings of the National Academy of Sciences of the United States of America* vol. 96 332–334 Preprint at https://doi.org/10.1073/pnas.96.2.332 (1999).

301. Benjin, X. & Ling, L. Developments, applications, and prospects of cryo-electron microscopy. *Protein Science* vol. 29 872–882 Preprint at https://doi.org/10.1002/pro.3805 (2020).

302. Lyumkis, D. Challenges and opportunities in cryo-EM single-particle analysis. *Journal of Biological Chemistry* **294**, 5181–5197 (2019).

303. Madhavan, S. M. & Buck, M. The Relationship between APOL1 Structure and Function: Clinical Implications. *Kidney360* **2**, 134–140 (2021).

304. Mayanja, R. *et al.* Molecular Dynamic Simulation Reveals Structure Differences in APOL1 Variants and Implication in Pathogenesis of Chronic Kidney Disease. *Genes (Basel)* **13**, (2022).

305. Kumar, V. *et al.* Disruption of APOL1-miR193a Axis Induces Disorganization of Podocyte Actin Cytoskeleton. *Sci Rep* **9**, (2019).

306. Ultsch, M. *et al.* Structures of the ApoL1 and ApoL2 N-terminal domains reveal a non-classical four-helix bundle motif. *Commun Biol* **4**, (2021).

307. Dobson, L. *et al.* TmAlphaFold database: membrane localization and evaluation of AlphaFold2 predicted alpha-helical transmembrane protein structures. *Nucleic Acids Res* **51**, D517–D522 (2023).

308. Perrakis, A. & Sixma, T. K. AI revolutions in biology. *EMBO Rep* **22**, e54046 (2021).

309. Azzaz, F., Yahi, N., Chahinian, H. & Fantini, J. The Epigenetic Dimension of Protein Structure Is an Intrinsic Weakness of the AlphaFold Program. *Biomolecules* **12**, (2022).

310. Hegedűs, T., Geisler, M., Lukács, G. L. & Farkas, B. Ins and outs of AlphaFold2 transmembrane protein structure predictions. *Cellular and Molecular Life Sciences* **79**, 1–12 (2022).

311. Bertoline, L. M. F., Lima, A. N., Krieger, J. E. & Teixeira, S. K. Before and after AlphaFold2: An overview of protein structure prediction. *Frontiers in Bioinformatics* **3**, (2023).

312. Madhavan, S. M. *et al.* APOL1 variants change C-terminal conformational dynamics and binding to SNARE protein VAMP8. *JCI Insight* **2**, (2017).

313. Pays, E. *et al.* The trypanolytic factor of human serum. *Nat Rev Microbiol* **4**, 477–486 (2006).

314. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res* **47**, W402–W407 (2019).

315. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405 (2000).

316. Gabler, F. *et al.* Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr Protoc Bioinformatics* **72**, (2020).

317. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science (1979)* **373**, 871–876 (2021).

318. Dong, L. M. *et al.* Novel mechanism for defective receptor binding of apolipoprotein E2 in type III hyperlipoproteinemia. *Nature Structural Biology 1996 3:8* **3**, 718–722 (1996).

319. McPherson, A. Crystal structure of a proteolytically cleaved, amino terminal domain of apolipoprotein E3. *Biochem Biophys Res Commun* **525**, 57–60 (2020).

320. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815 (1993).

321. Best, R. B. *et al.* Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain χ1 and χ2 Dihedral Angles. *J Chem Theory Comput* **8**, 3257–3273 (2012).

322. Rosato, A., Tejero, R. & Montelione, G. T. Quality assessment of protein NMR structures. *Current Opinion in Structural Biology* vol. 23 715–724 Preprint at https://doi.org/10.1016/j.sbi.2013.08.005 (2013).

323. Mitusińska, K., Skalski, T. & Góra, A. Simple selection procedure to distinguish between static and flexible loops. *Int J Mol Sci* **21**, 2293 (2020).

324. Sittel, F., Jain, A. & Stock, G. Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *J Chem Phys* **141**, 07B605_1 (2014).

325. Sheik Amamuddy, O., Boateng, R. A., Barozi, V., Wavinya Nyamai, D. & Tastan Bishop, Ö. Novel dynamic residue network analysis approaches to study allosteric modulation: SARS-CoV-2 Mpro and its evolutionary mutations as a case study. *Comput Struct Biotechnol J* **19**, 6431–6455 (2021).

326. Böde, C. *et al.* Network analysis of protein dynamics. *FEBS Letters* vol. 581 2776–2782 Preprint at https://doi.org/10.1016/j.febslet.2007.05.021 (2007).

327. Ohmura, T., Ueda, T., Hashimoto, Y. & Imoto, T. Tolerance of point substitution of methionine for isoleucine in hen egg white lysozyme. *Protein Eng* **14**, 421–425 (2001).

328. Leahy, D. J., Erickson, H. P., Aukhil, I., Joshi, P. & Hendrickson, W. A. Crystallization of a fragment of human fibronectin: Introduction of methionine by site-directed mutagenesis to allow phasing via selenomethionine. *Proteins: Structure, Function, and Bioinformatics* **19**, 48–54 (1994).

329. Skinner, M. M. *et al.* Structure of the gene V protein of bacteriophage f1 determined by multiwavelength x-ray diffraction on the selenomethionyl protein. *Proc Natl Acad Sci U S A* **91**, 2071–2075 (1994).

330. Capewell, P. *et al.* The TgsGP Gene Is Essential for Resistance to Human Serum in Trypanosoma brucei gambiense. *PLoS Pathog* **9**, e1003686 (2013).

331. Uzureau, P. *et al.* Mechanism of Trypanosoma brucei gambiense resistance to human serum. *Nature* **501**, 430–434 (2013).

332. Van Xong, H. *et al.* A VSG expression site-associated gene confers resistance to human serum in Trypanosoma rhodesiense. *Cell* **95**, 839–846 (1998).

333. De Greef, C., Chimfwembe, E., Kihang'a Wabacha, J., Bajyana Songa, E. & Hamers, R. Only the serum-resistant bloodstream forms of Trypanosoma brucei rhodesiense express the serum resistance associated (SRA) protein. *Ann Soc Belg Med Trop* **72 Suppl 1**, 13–21 (1992).

334. De Greef, C. & Hamers, R. The serum resistance-associated (SRA) gene of Trypanosoma brucei rhodesiense encodes a variant surface glycoprotein-like protein. *Mol Biochem Parasitol* **68**, 277–284 (1994).

335. De Greef, C., Imberechts, H., Matthyssens, G., Van Meirvenne, N. & Hamers, R. A gene expressed only in serum-resistant variants of Trypanosoma brucei rhodesiense. *Mol Biochem Parasitol* **36**, 169–176 (1989).

336. Milner, J. D. & Hajduk, S. L. Expression and localization of serum resistance associated protein in Trypanosoma brucei rhodesiense. *Mol Biochem Parasitol* **104**, 271–283 (1999).

337. Welburn, S. C. *et al.* Identification of human-infective trypanosomes in animal reservoir of sleeping sickness in Uganda by means of serum-resistance-associated (SRA) gene. *Lancet* **358**, 2017–2019 (2001).

338. Gibson, W., Backhouse, T. & Griffiths, A. The human serum resistance associated gene is ubiquitous and conserved in Trypanosoma brucei rhodesiense throughout East Africa. *Infection, Genetics and Evolution* **1**, 207–214 (2002).

339. Zoll, S. *et al.* The structure of serum resistance-associated protein and its implications for human African trypanosomiasis. *Nat Microbiol* **3**, 295–301 (2018).

340. Lecordier, L. *et al.* Adaptation of Trypanosoma rhodesiense to hypohaptoglobinaemic serum requires transcription of the APOL1 resistance gene in a RNA polymerase I locus. *Mol Microbiol* **97**, 397–407 (2015).

341. Campillo, N. & Carrington, M. The origin of the serum resistance associated (SRA) gene and a model of the structure of the SRA polypeptide from Trypanosoma brucei rhodesiense. *Mol Biochem Parasitol* **127**, 79–84 (2003).

342. Bart, J.-M. *et al.* Localization of serum resistance-associated protein in Trypanosoma brucei rhodesiense and transgenic Trypanosoma brucei brucei. *Cell Microbiol* **17**, 1523–35 (2015).

343. Oli, M. W., Cotlin, L. F., Shiflett, A. M. & Hajduk, S. L. Serum resistance-associated protein blocks lysosomal targeting of trypanosome lytic factor in Trypanosoma brucei. *Eukaryot Cell* **5**, 132–139 (2006).

344. Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **44**, D7 (2016).

345. Consortium, T. U. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480–D489 (2021).

346. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680 (1994).

347. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

348. Abraham, M., Van Der Spoel, D., Lindahl, E., Hess, B. & Team, G. D. GROMACS User Manual version 2018.1. *The GROMACS development teams at the Royal Institute of Technology and Uppsala University, Sweden. More* vol. 2018.1 Preprint at www.gromacs.org (2018).

349. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

350. Privalov, P. L. & Gill, S. J. Stability of Protein Structure and Hydrophobic Interaction. *Adv Protein Chem* **39**, 191–234 (1988).

351. Pace, C. N. *et al.* Contribution of hydrophobic interactions to protein stability. *J Mol Biol* **408**, 514–528 (2011).

352. Krishnamoorthy, N. & Fakhro, K. Identification of mutation resistance coldspots for targeting the SARS-CoV2 main protease. *IUBMB Life* **73**, 670–675 (2021).

353. Gibson, W. C. The SRA gene: The key to understanding the nature of Trypanosoma brucei rhodesiense. *Parasitology* vol. 131 143–150 Preprint at https://doi.org/10.1017/S0031182005007560 (2005).

354. Lai, D., Wang, Q., Li, Z., Julius, L. & Lun, Z. Evolution of the serum resistance-associated SRA gene in African trypanosomes. *Chinese Science Bulletin* **54**, 1275–1278 (2009).

355. Jamonneau, V. *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* **98**, 329–337 (2004).

356. Limou, S., Dummer, P. D., Nelson, G. W., Kopp, J. B. & Winkler, C. A. APOL1 toxin, innate immunity, and kidney injury. *Kidney International* vol. 88 28–34 Preprint at https://doi.org/10.1038/ki.2015.109 (2015).

357. Cooper, A. *et al.* APOL1 renal risk variants have contrasting resistance and susceptibility associations with African trypanosomiasis. *Elife* **6**, (2017).

358. Schaub, C. *et al.* Cation channel conductance and pH gating of the innate immunity factor APOL1 are governed by pore-lining residues within the C-terminal domain. *Journal of Biological Chemistry* vol. 295 13138–13149 Preprint at https://doi.org/10.1074/jbc.RA120.014201 (2020).

359. Stephens, N. A. & Hajduk, S. L. Endosomal localization of the serum resistance-associated protein in African trypanosomes confers human infectivity. *Eukaryot Cell* **10**, 1023–1033 (2011).

360. Bogdanov, M., Xie, J., Heacock, P. & Dowhan, W. To flip or not to flip: Lipid-protein charge interactions are a determinant of final membrane protein topology. *Journal of Cell Biology* **182**, 925–935 (2008).

361. Lu, Y. *et al.* Reorientation of aquaporin-1 topology during maturation in the endoplasmic reticulum. *Mol Biol Cell* **11**, 2973–2985 (2000).

362. Giovinazzo, J. A. *et al.* Apolipoprotein L-1 renal risk variants form active channels at the plasma membrane driving cytotoxicity. *Elife* **9**, (2020).

363. Honorato, R. V. *et al.* Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. *Front Mol Biosci* **8**, 708 (2021).

364. Van Zundert, G. C. P. *et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **428**, 720–725 (2016).

365. Jorgensen, W. L. & Tirado-Rives, J. The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J Am Chem Soc* **110**, 1657–1666 (1988).

366. Koukos, P. I. *et al.* An overview of data-driven HADDOCK strategies in CAPRI rounds 38-45. *Proteins: Structure, Function and Bioinformatics* **88**, 1029–1036 (2020).

367. Team, Rs. RStudio: Integrated Development Environment for R. Preprint at http://www.rstudio.org/ (2020).

368. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J Mol Biol* **372**, 774–797 (2007).

369. Kortemme, T., Kim, D. E. & Baker, D. Computational alanine scanning of protein-protein interfaces. *Sci STKE* **2004**, (2004).

370. Rodrigues, J. P. G. L. M. *et al.* Clustering biomolecular complexes by residue contacts similarity. *Proteins: Structure, Function and Bioinformatics* **80**, 1810–1817 (2012).

371. Krissinel, E. Macromolecular complexes in crystals and solutions. *Acta Crystallogr D Biol Crystallogr* **67**, 376–385 (2011).

372. Chou, P. Y. & Fasman, G. D. Structural and functional role of leucine residues in proteins. *J Mol Biol* **74**, 263–281 (1973).

373. Reidy, K. J. *et al.* Fetal—Not Maternal—APOL1 Genotype Associated with Risk for Preeclampsia in Those with African Ancestry. *Am J Hum Genet* **103**, 367–376 (2018).

374. Barrett, M. P. Problems for the chemotherapy of human African trypanosomiasis. *Current Opinion in Infectious Diseases* vol. 13 647–651 Preprint at https://doi.org/10.1097/00001432-200012000-00012 (2000).

375. Wiedemar, N., Hauser, D. A. & Mäser, P. 100 years of suramin. *Antimicrob Agents Chemother* **64**, (2020).

376. Bronner, U. *et al.* Pharmacokinetics and adverse reactions after a single dose of pentamidine in patients with Trypanosoma gambiense sleeping sickness. *Br J Clin Pharmacol* **39**, 289–295 (1995).

377. Blum, J., Nkunku, S. & Burri, C. Clinical description of encephalopathic syndromes and risk factors for their occurrence and outcome during melarsoprol treatment of human African trypanosomiasis. *Tropical Medicine and International Health* **6**, 390–400 (2001).

378. Vincent, I. M. *et al.* A molecular mechanism for eflornithine resistance in African trypanosomes. *PLoS Pathog* **6**, (2010).

379. Sokolova, A. Y. *et al.* Cross-resistance to nitro drugs and implications for treatment of human african trypanosomiasis. *Antimicrob Agents Chemother* **54**, 2893–2900 (2010).

380. Robays, J. *et al.* High failure rates of melarsoprol for sleeping sickness, Democratic Republic of Congo. *Emerg Infect Dis* **14**, 966–967 (2008).

381. Legros, D., Evans, S., Maiso, F., Enyaru, J. C. K. & Mbulamberi, D. Risk factors for treatment failure after melarsoprol for Trypanosoma brucei gambiense trypanosomiasis in Uganda. *Trans R Soc Trop Med Hyg* **93**, 439–442 (1999).

382. Brun, R., Don, R., Jacobs, R. T., Wang, M. Z. & Barrett, M. P. Development of novel drugs for human African trypanosomiasis. *Future Microbiology* vol. 6 677–691 Preprint at https://doi.org/10.2217/fmb.11.44 (2011).

383. Gangjee, A., Jain, H. & Kurup, S. Recent Advances in Classical and Non-Classical Antifolates as Antitumor and Antiopportunistic Infection Agents: Part II. *Anticancer Agents Med Chem* **8**, 205–231 (2008).

384. Gangjee, A., Kurup, S. & Namjoshi, O. Dihydrofolate Reductase as a Target for Chemotherapy in Parasites. *Curr Pharm Des* **13**, 609–639 (2007).

385. Beck, J. T. & Ullman, B. Nutritional requirements of wild-type and folate transport-deficient Leishmania donovani for pterins and folates. *Mol Biochem Parasitol* **43**, 221–230 (1990).

386. Scott, D. A., Coombs, G. H. & Sanderson, B. E. Folate utilisation by Leishmania species and the identification of intracellular derivatives and folate-metabolising enzymes. *Mol Biochem Parasitol* **23**, 139–149 (1987).

387. Blaney, J. M., Hansch, C., Silipo, C. & Vittoria, A. Structure-Activity Relationships of Dihydrofolate Reductase Inhibitors. *Chem Rev* **84**, 333–407 (1984).

388. Nare, B., Hardy, L. W. & Beverley, S. M. The roles of pteridine reductase 1 and dihydrofolate reductase- thymidylate synthase in pteridine metabolism in the protozoan parasite Leishmania major. *Journal of Biological Chemistry* **272**, 13883–13891 (1997).

389. Metz, J. Folic acid metabolism and malaria. *Food Nutr Bull* **28**, (2007).

390. Sienkiewicz, N., Ong, H. B. & Fairlamb, A. H. Trypanosoma brucei pteridine reductase 1 is essential for survival in vitro and for virulence in mice. *Mol Microbiol* **77**, 658–671 (2010).

391. Sienkiewicz, N., Jarosławski, S., Wyllie, S. & Fairlamb, A. H. Chemical and genetic validation of dihydrofolate reductase-thymidylate synthase as a drug target in African trypanosomes. *Mol Microbiol* **69**, 520–533 (2008).

392. Nare, B., Luba, J., Hardy, L. W. & Beverley, S. New approaches to Leishmania chemotherapy: Pteridine Reductase 1 (PTR1) as a target and modulator of antifolate sensitivity. *Parasitology* **114**, 101–110 (1997).

393. Dawson, A. *et al.* Structure and reactivity of Trypanosoma brucei pteridine reductase: Inhibition by the archetypal antifolate methotrexate. *Mol Microbiol* **61**, 1457–1468 (2006).

394. Ong, H. B., Sienkiewicz, N., Wyllie, S. & Fairlamb, A. H. Dissecting the metabolic roles of pteridine reductase 1 in Trypanosoma brucei and Leishmania major. *Journal of Biological Chemistry* **286**, 10429–10438 (2011).

395. Gourley, D. G. *et al.* Pteridine reductase mechanism correlates pterin metabolism with drug resistance in trypanosomatid parasites. *Nat Struct Biol* **8**, 521–525 (2001).

396. Bello, A. R., Nare, B., Freedman, D., Hardy, L. & Beverley, S. M. PTR1: A reductase mediating salvage of oxidized pteridines and methotrexate resistance in the

protozoan parasite Leishmania major. *Proc Natl Acad Sci U S A* **91**, 11442–11446 (1994).

397. Vickers, T. J. & Beverley, S. M. Folate metabolic pathways in Leishmania. *Essays Biochem* **51**, 63–80 (2011).

398. Teixeira, B. V. F. *et al.* Dual and selective inhibitors of pteridine reductase 1 (PTR1) and dihydrofolate reductase-thymidylate synthase (DHFR-TS) from Leishmania chagasi. *J Enzyme Inhib Med Chem* **34**, 1439–1450 (2019).

399. Oprea, T. I. & Overington, J. P. Computational and practical aspects of drug repositioning. *Assay Drug Dev Technol* **13**, 299–306 (2015).

400. Agrawal, P. Advantages and Challenges in Drug Re-Profiling. *J Pharmacovigil* **s2**, (2015).

401. Kepplinger, E. E. FDA's expedited approval mechanisms for new drug products. *Biotechnol Law Rep* **34**, 15–37 (2015).

402. Ashburn, T. T. & Thor, K. B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* **3**, 673–683 (2004).

403. Vargesson, N. Thalidomide-induced teratogenesis: History and mechanisms. *Birth Defects Research Part C - Embryo Today: Reviews* vol. 105 140–156 Preprint at https://doi.org/10.1002/bdrc.21096 (2015).

404. Greenstone, G. The revival of thalidomide: From tragedy to therapy. *British Columbia Medical Journal* vol. 53 230–233 Preprint at (2011).

405. Novac, N. Challenges and opportunities of drug repositioning. *Trends in Pharmacological Sciences* vol. 34 267–272 Preprint at https://doi.org/10.1016/j.tips.2013.03.004 (2013).

406. Cairns, D. M., Boorgu, D. S. S. K., Levin, M. & Kaplan, D. L. Niclosamide rescues microcephaly in a humanized in vivo model of Zika infection using human induced neural stem cells. *Biol Open* **7**, bio031807–bio031807 (2018).

407. Devillers, J. Repurposing drugs for use against Zika virus infection$. *SAR QSAR Environ Res* **29**, 103–115 (2018).

408. Micewicz, E. D., Khachatoorian, R., French, S. W. & Ruchala, P. Identification of novel small-molecule inhibitors of Zika virus infection. *Bioorg Med Chem Lett* **28**, 452–458 (2018).

409. Kadri, H., Lambourne, O. A. & Mehellou, Y. Niclosamide, a Drug with Many (Re)purposes. *ChemMedChem* **13**, 1088–1091 (2018).

410. Burri, C. & Brun, R. Eflornithine for the treatment of human African trypanosomiasis. *Parasitol Res* **90**, S49–S52–S49–S52 (2003).

411. Bouteille, B., Oukem, O., Bisser, S. & Dumas, M. Treatment perspectives for human African trypanosomiasis. *Fundam Clin Pharmacol* **17**, 171–181 (2003).

412. Dehpour, A. R. *Drug Repositioning: A Review*. *Journal of Iranian Medical Council* vol. 1 http://www.jimc.ir/article_66203.html (2018).

413. Liu, X. *et al.* Predicting Targeted Polypharmacology for Drug Repositioning and Multi- Target Drug Discovery. *Curr Med Chem* **20**, 1646–1661 (2013).

414. Muegge, I. Synergies of Virtual Screening Approaches. *Mini-Reviews in Medicinal Chemistry* **8**, 927–933 (2008).

415. Tan, J. J. *et al.* Therapeutic strategies underpinning the development of novel techniques for the treatment of HIV infection. *Drug Discovery Today* vol. 15 186–197 Preprint at https://doi.org/10.1016/j.drudis.2010.01.004 (2010).

416. Singh, S., Malik, B. K. & Sharma, D. K. Molecular drug targets and structure based drug design: A holistic approach. *Bioinformation* **1**, 314–320 (2006).

417. Schmidt, F., Matter, H., Hessler, G. & Czich, A. Predictive in silico off-target profiling in drug discovery. *Future Medicinal Chemistry* vol. 6 295–317 Preprint at https://doi.org/10.4155/fmc.13.202 (2014).

418. Issa, N., Byers, S. & Dakshanamurthy, S. Drug Repurposing: Translational Pharmacology, Chemistry, Computers and the Clinic. *Curr Top Med Chem* **13**, 2328–2336 (2013).

419. Al-Sha'Er, M. A. & Taha, M. O. Discovery of novel CDK1 inhibitors by combining pharmacophore modeling, QSAR analysis and in silico screening followed by in vitro bioassay. *Eur J Med Chem* **45**, 4316–4330 (2010).

420. Ayton, G. S., Noid, W. G. & Voth, G. A. Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr Opin Struct Biol* **17**, 192–198 (2007).

421. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol* **9**, 71 (2011).

422. Talele, T., Khedkar, S. & Rigby, A. Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic. *Curr Top Med Chem* **10**, 127–141 (2010).

423. Wlodawer, A. & Vondrasek, J. Inhibitors of HIV-1 protease: A major success of structure-assisted drug design. *Annual Review of Biophysics and Biomolecular Structure* vol. 27 249–284 Preprint at https://doi.org/10.1146/annurev.biophys.27.1.249 (1998).

424. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **31**, 455–461 (2010).

425. Hassan, N. M., Alhossary, A. A., Mu, Y. & Kwoh, C.-K. Protein-ligand blind docking using QuickVina-W with inter-process spatio-temporal integration. *Sci Rep* **7**, 1–13 (2017).

426. Landrum, G. RDKit: Open-Source Cheminformatics Software. *Http://Www.Rdkit.Org/* Preprint at (2021).

427. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* **30**, 2785–2791 (2009).

428. Sanner, M. F. Python: A programming language for software integration and development. *J Mol Graph Model* **17**, 57–61 (1999).

429. Dawson, A., Tulloch, L. B., Barrack, K. L. & Hunter, W. N. 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 D Biol Crystallogr* **66**, 1334–1340 (2010).

430. Kimuda, M. P., Laming, D., Hoppe, H. C. & Bishop, O. T. Identification of Novel Potential Inhibitors of Pteridine Reductase 1 in Trypanosoma brucei via

Computational Structure-Based Approaches and in Vitro Inhibition Assays. *Molecules* **24**, 142 (2019).

431. Vanichtanankul, J. *et al.* Trypanosomal dihydrofolate reductase reveals natural antifolate resistance. *ACS Chem Biol* **6**, 905–911 (2011).

432. Martínez-Rosell, G., Giorgino, T. & De Fabritiis, G. PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations. *J Chem Inf Model* **57**, 1511–1516 (2017).

433. Tange, O. GNU Parallel 2018. Preprint at https://doi.org/10.5281/zenodo.1146014 (2018).

434. Kumari, R., Kumar, R. & Lynn, A. G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations. *J Chem Inf Model* **54**, 1951–1962 (2014).

435. Veale, C. G. L. & Hoppe, H. C. Screening of the Pathogen Box reveals new starting points for anti-trypanosomal drug discovery. *Medchemcomm* **9**, 2037–2044 (2018).

436. Wang, L. *et al.* Design, synthesis, and neuroprotective effects of dual-brain targeting naproxen prodrug. *Arch Pharm (Weinheim)* **351**, 1700382 (2018).

437. Thompson, A. J. & Lummis, S. C. R. The 5-HT3 receptor as a therapeutic target. *Expert Opin Ther Targets* **11**, 527–540 (2007).

438. Sanger, G. J. & Andrews, P. L. R. A history of drug discovery for treatment of nausea and vomiting and the implications for future research. *Frontiers in Pharmacology* vol. 9 913 Preprint at https://doi.org/10.3389/fphar.2018.00913 (2018).

439. Abdollahi, M. & Mostafalou, S. Tricyclic Antidepressants. in *Encyclopedia of Toxicology: Third Edition* 838–845 (Elsevier, 2014). doi:10.1016/B978-0-12-386454-3.00796-X.

440. Benfield, P., Heel, R. C. & Lewis, S. P. Fluoxetine: A Review of Its Pharmacodynamic and Pharmacokinetic Properties, and Therapeutic Efficacy in Depressive Illness. *Drugs* **32**, 481–508 (1986).

441. Birdsall, T. C. 5-Hydroxytryptophan: A Clinically-Effective Serotonin Precursor. *Alternative Medicine Review* **3**, 271–280 (1998).

442. Amer, A., Breu, J., McDermott, J., Wurtman, R. J. & Maher, T. J. 5-Hydroxy-L-tryptophan suppresses food intake in food-deprived and stressed rats. *Pharmacol Biochem Behav* **77**, 137–143 (2004).

443. O'Brien, F. E., Dinan, T. G., Griffin, B. T. & Cryan, J. F. Interactions between antidepressants and P-glycoprotein at the blood-brain barrier: Clinical significance of in vitro and in vivo findings. *British Journal of Pharmacology* vol. 165 289–312 Preprint at https://doi.org/10.1111/j.1476-5381.2011.01557.x (2012).

444. Zheng, Y., Chen, X. & Benet, L. Z. Reliability of In Vitro and In Vivo Methods for Predicting the Effect of P-Glycoprotein on the Delivery of Antidepressants to the Brain. *Clinical Pharmacokinetics* vol. 55 143–167 Preprint at https://doi.org/10.1007/s40262-015-0310-2 (2016).

445. Shutt, L. E. & Bowes, J. B. Atropine and hyoscine. *Anaesthesia* **34**, 476–490 (1979).

446. Henry, T. A. *The plant alkaloids*. (J. & A. Churchill, 1924).

447. Sherman, A. E., Shaw, M. M., Ralay-Ranaivo, H. & Rahmani, B. Tropicamide has limited clinical effect on cycloplegia and mydriasis when combined with cyclopentolate and phenylephrine. *Journal of AAPOS* **23**, 30.e1-30.e5 (2019).

448. Chancellor, M. B. *et al.* BloodBrain barrier permeation and efflux exclusion of anticholinergics used in the treatment of overactive bladder. *Drugs Aging* **29**, 259–273 (2012).

449. Çetinel, B. & Onal, B. Rationale for the use of anticholinergic agents in overactive bladder with regard to central nervous system and cardiovascular system side effects. *Korean J Urol* **54**, 806–815 (2013).

450. Haller, J. S. & Wainwright, M. The First Miracle Drugs: How the Sulfa Drugs Transformed Medicine (review). *J Hist Med Allied Sci* **63**, 639–642 (2007).

451. Powell, H. M. & Chen, K. K. THE ACTION OF SULFAPYRIDINE (2-SULFANILYL AMINOPYRIDINE). *Journal of Pharmacology and Experimental Therapeutics* **67**, (1939).

452. Flippin, H. F., Schwartz, L. & Rose, S. B. THE COMPARATIVE EFFECTIVENESS AND TOXICITY OF SULFATHIAZOLE AND SULFAPYRIDINE IN PNEUMOCOCCIC PNEUMONIA. *Ann Intern Med* **13**, 2038 (1940).

453. Palmer, K. J. & McTavish, D. Felbamate: A Review of its Pharmacodynamic and Pharmacokinetic Properties, and Therapeutic Efficacy in Epilepsy. *Drugs* **45**, 1041–1065 (1993).

454. Ritter, F. J. S. G. in L.-G. Efficacy of Felbamate in Childhood Epileptic Encephalopathy (Lennox-Gastaut Syndrome). *New England Journal of Medicine* **328**, 29–33 (1993).

455. Katz, T. M., Miller, J. H. & Hebert, A. A. Insect repellents: Historical perspectives and new developments. *Journal of the American Academy of Dermatology* vol. 58 865–871 Preprint at https://doi.org/10.1016/j.jaad.2007.10.005 (2008).

456. Maplestone, P. A. & Mukerji, A. K. Hexylresorcinol as an Anthelmintic. *Ind Med Gaz* **67**, 610- (1932).

457. Rogers, W. P. Studies on the anthelmintic activity of hexylresorcinol and tetrachlorethylene. *Parasitology* **36**, 98–109 (1944).

458. Matthews, D., Adegoke, O. & Shephard, A. Bactericidal activity of hexylresorcinol lozenges against oropharyngeal organisms associated with acute sore throat. *BMC Res Notes* **13**, 99 (2020).

459. Piazzon, A. *et al.* Antioxidant activity of phenolic acids and their metabolites: Synthesis and antioxidant properties of the sulfate derivatives of ferulic and caffeic acids and of the acyl glucuronide of ferulic acid. *J Agric Food Chem* **60**, 12312–12323 (2012).

460. Picone, P., Nuzzo, D. & Di Carlo, M. Ferulic acid: A natural antioxidant against oxidative stress induced by oligomeric A-beta on sea urchin embryo. *Biological Bulletin* **224**, 18–28 (2013).

461. Choi, R. *et al.* Effects of ferulic acid on diabetic nephropathy in a rat model of type 2 diabetes. *Exp Mol Med* **43**, 676–683 (2011).

462. Fetoni, A. R. *et al.* In vivo protective effect of ferulic acid against noise-induced hearing loss in the guinea-pig. *Neuroscience* **169**, 1575–1588 (2010).

463. Nazaré, A. C. *et al.* Ethyl ferulate, a component with anti-inflammatory properties for emulsion-based creams. *Molecules* **19**, 8124–8139 (2014).

464. Manoj Prabhakar, M. *et al.* Anti-cell proliferative efficacy of ferulic acid against 7, 12-dimethylbenz(a) anthracene induced hamster buccal pouch carcinogenesis. *Asian Pacific Journal of Cancer Prevention* **13**, 5207–5211 (2012).

465. Chadha, N. & Silakari, O. Indoles as therapeutics of interest in medicinal chemistry: Bird's eye view. *European Journal of Medicinal Chemistry* vol. 134 159–184 Preprint at https://doi.org/10.1016/j.ejmech.2017.04.003 (2017).

466. Kaushik, N. K. *et al.* Biomedical importance of indoles. *Molecules* vol. 18 6620–6662 Preprint at https://doi.org/10.3390/molecules18066620 (2013).

467. Bandini, M. Electrophilicity: The 'dark-side' of indole chemistry. *Org Biomol Chem* **11**, 5206–5212 (2013).

468. Badria, F. A. *et al.* Anticancer indole-based chalcones: A structural and theoretical analysis. *Molecules* **24**, 3728 (2019).

469. Dadashpour, S. & Emami, S. Indole in the target-based design of anticancer agents: A versatile scaffold with diverse mechanisms. *Eur J Med Chem* **150**, 9–29 (2018).

470. Verma, M., Tripathi, M., Saxena, A. & Shanker, K. Antiinflammatory activity of novel indole derivatives. *Eur J Med Chem* **29**, 941–946 (1994).

471. Verma, N. *et al.* Computational investigation for modelling the protein- protein interaction of TasA(28-261) – TapA(33- 253): A decisive process in biofilm formation by Bacillus subtilis. *bioRxiv* (2020) doi:10.1101/2020.04.26.062109.

472. Giampieri, M. *et al.* Antiviral activity of indole derivatives. *Antiviral Res* **83**, 179–185 (2009).

473. Hishmat, O. H., Abdel Rahman, A. H., Nasef, A. M. M., Ismail, M. M. F. & Abdel-Hamid, A. H. A. Synthesis and antimicrobial activity of some indole derivatives. *Arch Pharm Res* **11**, 266–269 (1988).

474. Zhang, M. Z., Chen, Q. & Yang, G. F. A review on recent developments of indole-containing antiviral agents. *Eur J Med Chem* **89**, 421–441 (2015).

475. Agarwal, A., Srivastava, K., Puri, S. K. & Chauhan, P. M. S. Synthesis of substituted indole derivatives as a new class of antimalarial agents. *Bioorg Med Chem Lett* **15**, 3133–3136 (2005).

476. Seebacher, W. *et al.* Synthesis of 3-azabicyclo[3.2.2]nonanes and their antiprotozoal activities. *Bioorg Med Chem Lett* **25**, 1390–1393 (2015).

477. Berger, H. *et al.* Novel Azabicyclo[3.2.2]nonane derivatives and their activities against Plasmodium falciparum K1 and Trypanosoma brucei rhodesiense. *Bioorg Med Chem* **16**, 6371–6378 (2008).

478. Zuccotto, F., Martin, A. C. R., Laskowski, R. A., Thornton, J. M. & Gilbert, I. H. Dihydrofolate reductase: A potential drug target in trypanosomes and leishmania. *J Comput Aided Mol Des* **12**, 241–257 (1998).

479. Gilbert, I. H. Inhibitors of dihydrofolate reductase in leishmania and trypanosomes. *Biochim Biophys Acta Mol Basis Dis* **1587**, 249–257 (2002).

480. Tulloch, L. B. *et al.* Structure-based design of pteridine reductase inhibitors targeting African sleeping sickness and the leishmaniases. *J Med Chem* **53**, 221–229 (2010).

481. Mpamhanga, C. P. *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* **52**, 4454–4465 (2009).

482. Beauchamp, K. A., Lin, Y.-S., Das, R. & Pande, V. S. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J Chem Theory Comput* **8**, 1409–1414 (2012).

483. Hospital, A., Goñi, J. R., Orozco, M. & Gelp\'\i, J. L. Molecular dynamics simulations: advances and applications. *Adv Appl Bioinform Chem* **8**, 37 (2015).

484. Larsson, P., Hess, B. & Lindahl, E. Algorithm improvements for molecular dynamics simulations. *Wiley Interdiscip Rev Comput Mol Sci* **1**, 93–108 (2011).

485. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences* 201800690 (2018).

486. Hollingsworth, S. A. & Dror, R. O. Molecular dynamics simulation for all. *Neuron* **99**, 1129 (2018).

487. Kisliuk, R. L. Folate Biochemistry in Relation to Antifolate Selectivity. *Antifolate Drugs in Cancer Therapy* 13–36 (1999) doi:10.1007/978-1-59259-725-3_2.

488. Wang, T., Pulkkinen, O. I. & Aittokallio, T. Target-specific compound selectivity for multi-target drug discovery and repurposing. *Front Pharmacol* **13**, (2022).

489. Alsford, S., Currier, R. B., Guerra-Assunção, J. A., Clark, T. G. & Horn, D. Cathepsin-L Can Resist Lysis by Human Serum in Trypanosoma brucei brucei. *PLoS Pathog* **10**, e1004130 (2014).

490. Adasme, M. F. *et al.* Repositioned Drugs for Chagas Disease Unveiled via Structure-Based Drug Repositioning. *International Journal of Molecular Sciences 2020, Vol. 21, Page 8809* **21**, 8809 (2020).

491. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034 (2022) doi:10.1101/2021.10.04.463034.

492. Tsuchiya, Y., Yamamori, Y. & Tomii, K. Protein–protein interaction prediction methods: from docking-based to AI-based approaches. *Biophys Rev* **14**, 1341–1348 (2022).

493. Johansson-Åkhe, I. & Wallner, B. Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. *Frontiers in bioinformatics* **2**, (2022).

494. Lensink, M. F. *et al.* Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins* **87**, 1200–1221 (2019).

495. Lensink, M. F. *et al.* Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins* **89**, 1800 (2021).

496. Lensink, M. F., Nadzirin, N., Velankar, S. & Wodak, S. J. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins: Structure, Function, and Bioinformatics* **88**, 916–938 (2020).

497. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications 2022 13:1* **13**, 1–11 (2022).

498. Ross, C. *et al.* MODE-TASK: Large-scale protein motion tools. *Bioinformatics* **34**, 3759–3763 (2018).

499. Fernández-Recio, J., Totrov, M. & Abagyan, R. Identification of protein-protein interaction sites from docking energy landscapes. *J Mol Biol* **335**, 843–865 (2004).

APPENDIX A : Chapter 2

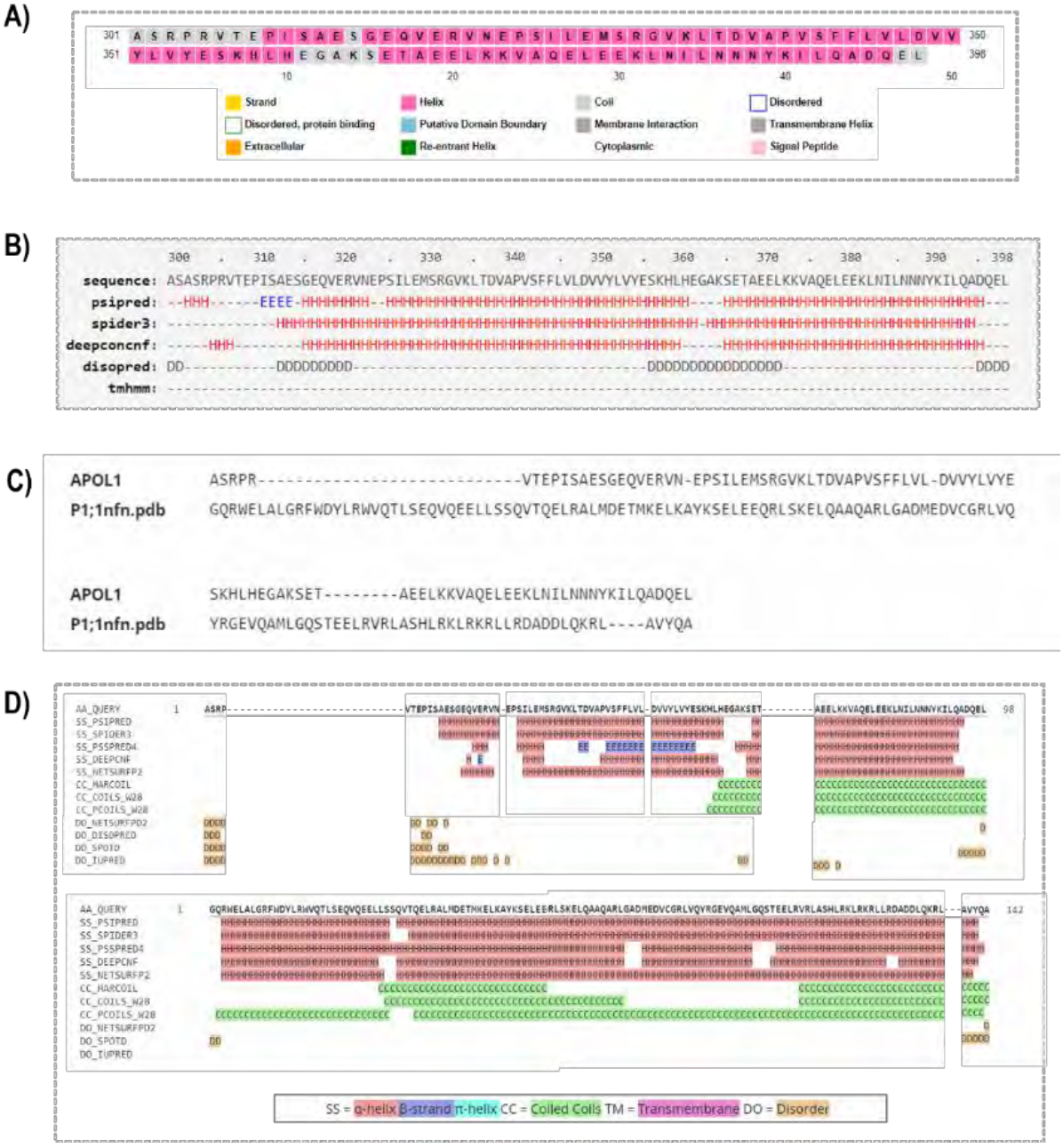


Figure S 1: ApoL1 C-terminal structure prediction. Template suggestions were included from A) PsiPred, B) Robetta C) and D) Quick2D (HHPred)

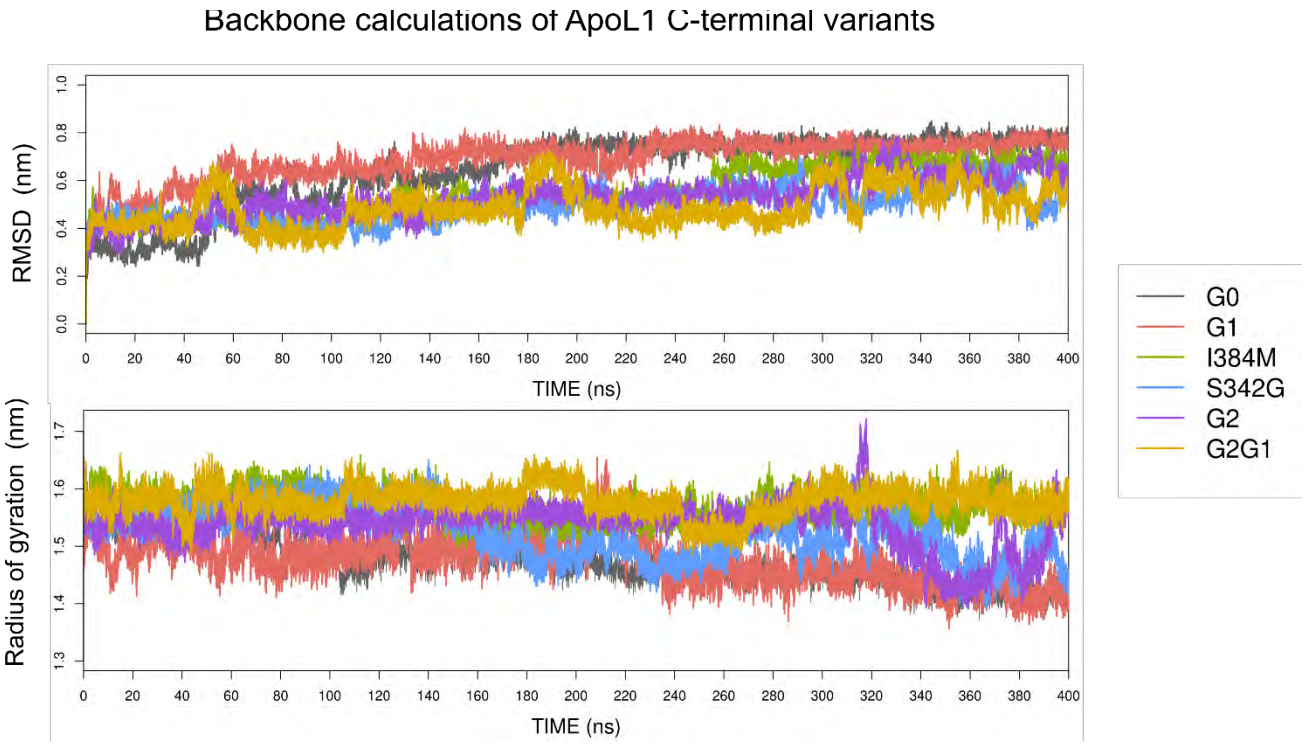


Figure S 2: The RMSD and Rg recorded for the ApoL1 variants over 400ns simulations, represented as line plots.

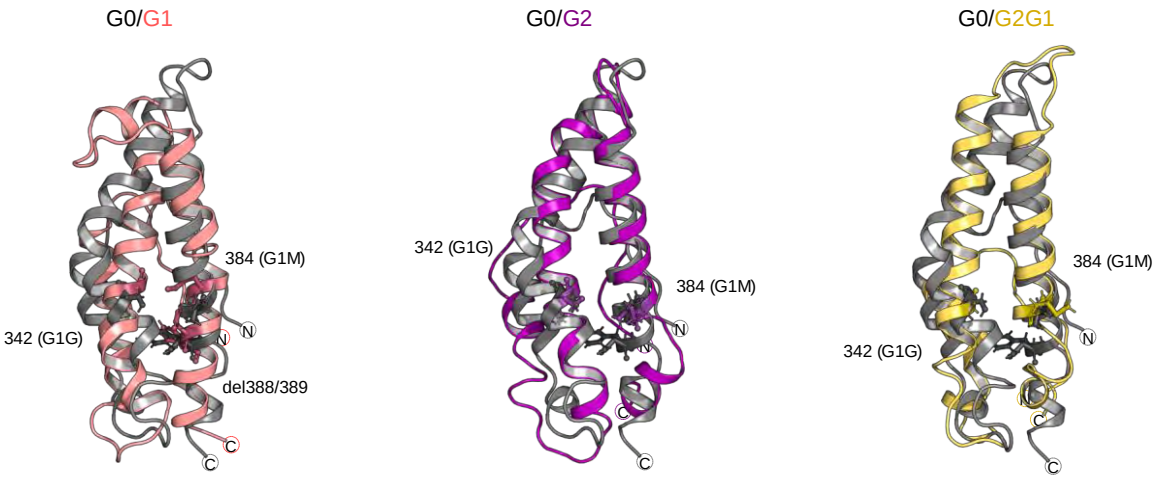


Figure S 3: Overlayed structures of ApoL1 C-terminal at 100ns.. The mutating residues are highlighted in the stick representation

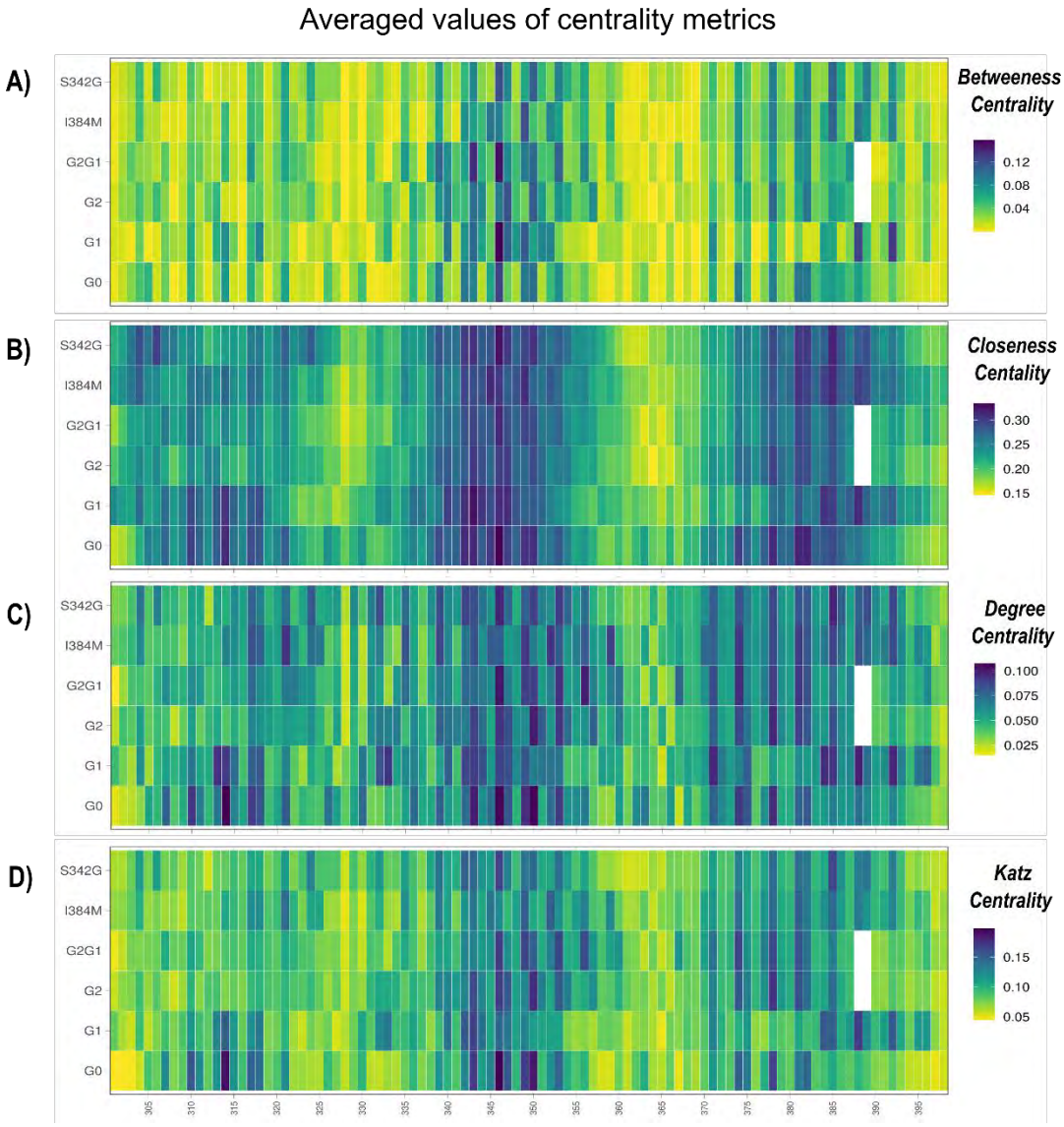


Figure S 4: Overall averaged DRN values of BC, CC, DC and KC of the ApoL1 c-terminal variants over 300ns trajectories.

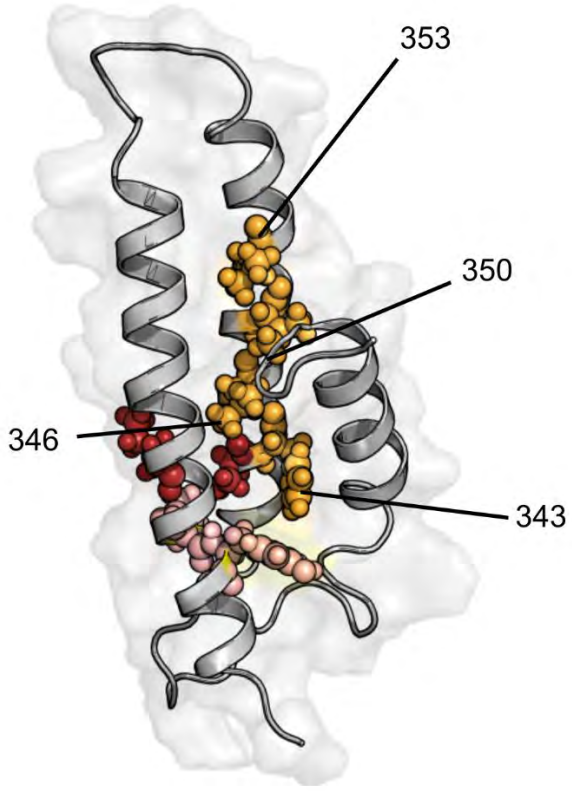


Figure S 5: Significant DRN hubs Phe343, Val346, 350 and 353 are mapped onto the ApoL1 C-terminal structure in orange. G1 and G2 mutation residues are highlighted in red and salmon, respectively.

APPENDIX B : Chapter 3

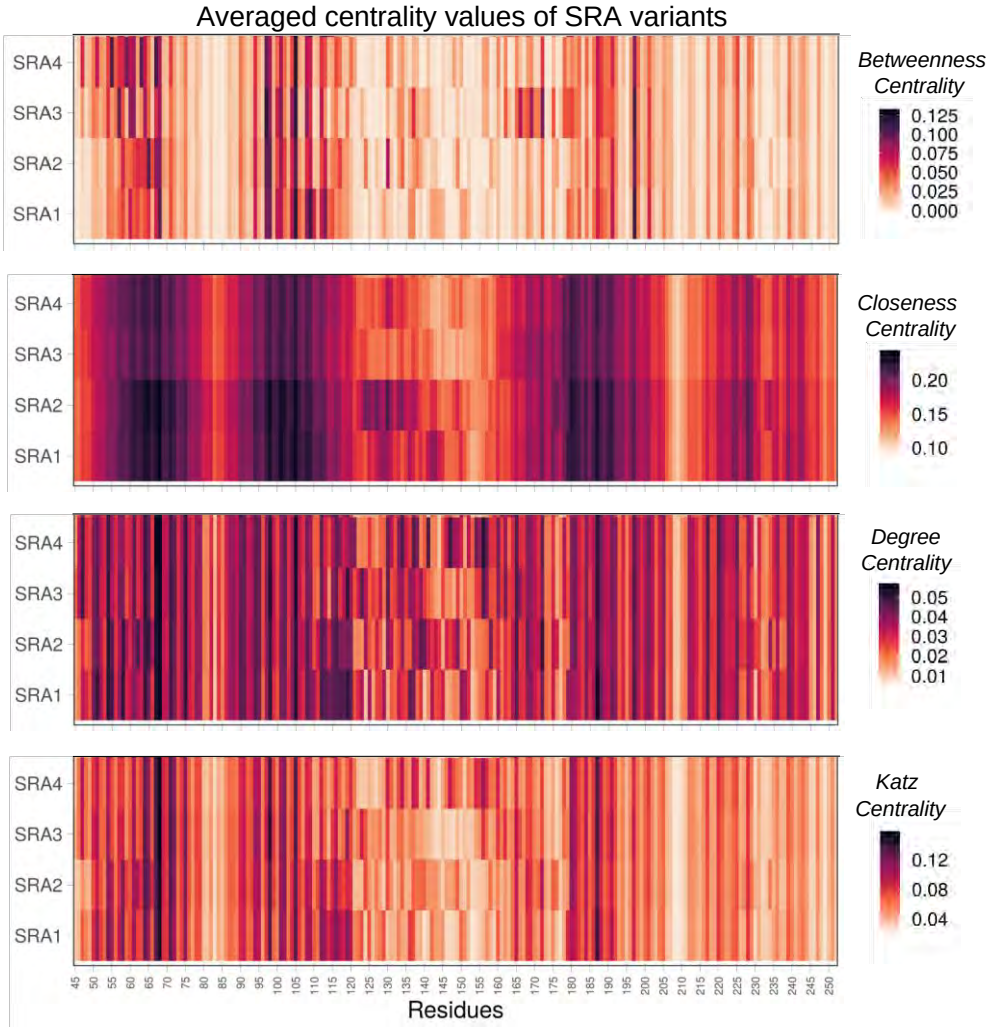


Figure S 6: Overall averaged DRN values of BC, CC, DC and KC of the SRA variants over 300ns trajectories.

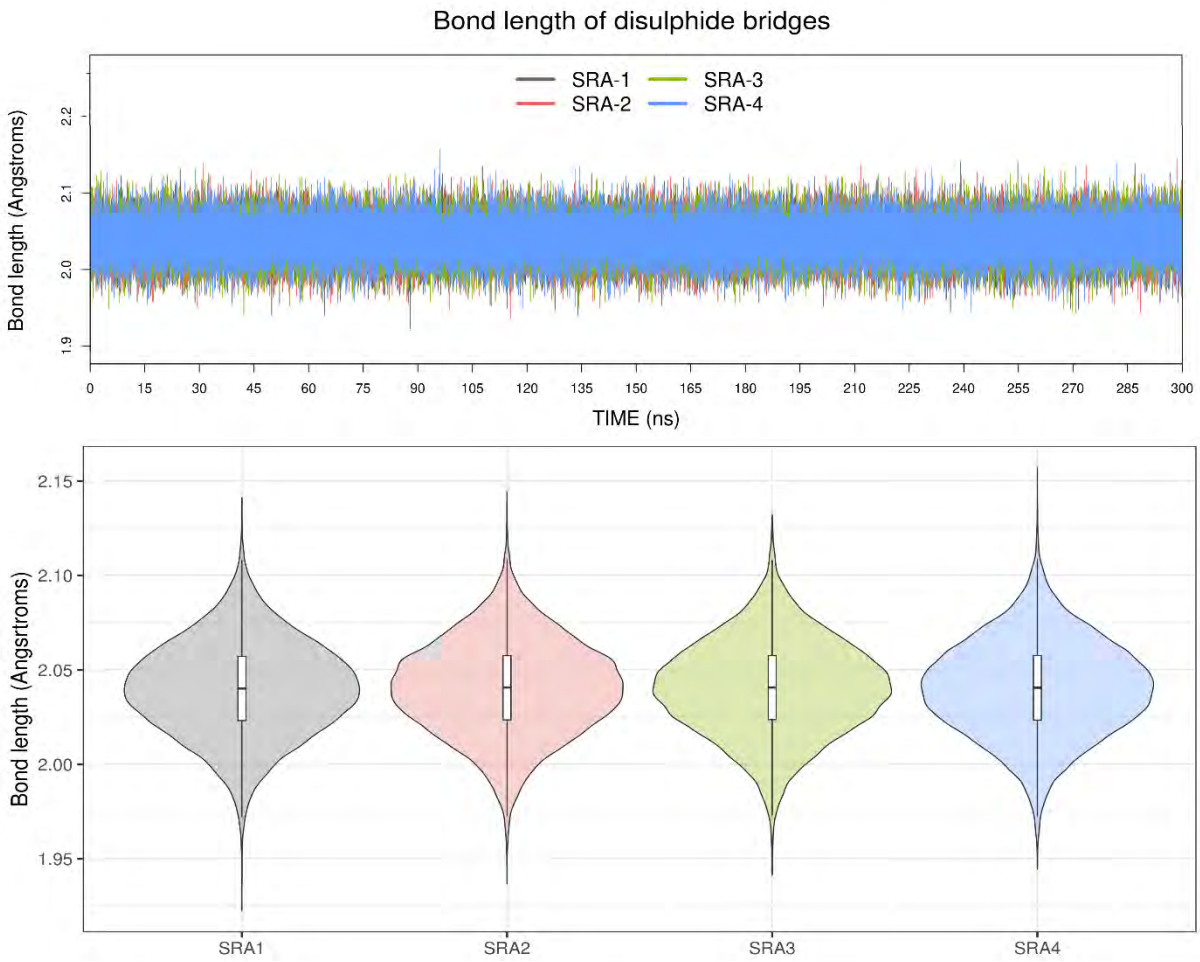


Figure S 7: The length of the Cys101-Cys180 disulphide bridge was monitored for each SRA variant over the course of the simulations. The bond length is represented in both line and violin plots.

APPENDIX C : Chapter 4

Choice in protein-protein docking software

Choice in protein-protein docking software was made using CASP-CAPRI rounds 38-46[494–496] rankings. Of the top ranking reputable servers , HADDOCK, ZDOCK, CLUSPRO and HDOCK were tested for complex formation. The resultant complexes were subjected to MD simulations using parameters outlined in ***section 5.2.2***

The output of the MD simulations is highlighted in **Figure S 8**

Additional calculations

For the NMA calculations, *NMA from MD* tool of the server MDM-TASK-web [240,498] was used. The first dominant mode was selected to analyse the larger, rotational motions of the system. In this calculation, principal components were produced through diagonalising a covariance matrix obtained through centring a transposed Cartesian coordinate tensor and dotting it with its transpose [237]. The eigen decomposition resulted in PCs of descending eigenvalue, where the percentage variance of the first 50 modes was explained. The NMA of each complex was calculated with the exclusion of 3 residues at each of the termini.

The results of the NMA calculation are depicted in **Figure S 17**

Table S 1: Statistical analysis for HADDOCK generated SRA-Apo L1 complexes from docking run 1. The HADDOCK score for each cluster was calculated using a summation equation of thermodynamic energies. From the top-scoring cluster, 1 (in blue) complexes 1 and 2 were obtained and used in further investigations.

Cluster rank	HADDOCK score [a] (a.u)	RMSD from the lowest energy structure (Å)	Cluster size	Van der Waals energy [b] (E_{VdW}) (kcal mol^{-1})	Electrostatic energy [b] (E_{elec}) (kcal mol^{-1})	Restraints violation energy (E_{air}) (kcal mol^{-1})	Desolvation energy [c] (E_{desol}) (kcal mol^{-1})	Buried Surface Area (Å^2)	Z-Score
Cluster 1	**-76,68 ± 5,42**	**8,79 ± 0,29**	**16**	**-38,36 ± 1,75**	**-212,5 ± 58,17**	**104,07 ± 52,33**	**-6,22 ± 2,69**	**2034 ± 85,35**	**-1.5**
Cluster 2	-74,68 ± 8,71	15,77 ± 0,12	12	-27,63 ± 5,96	-313,39 ± 48,13	140,68 ± 31,65	1,56 ± 5,79	1721,45 ± 119,83	-1.3
Cluster 16	-70,64 ± 5,99	6,37 ± 0,27	4	-28,46 ± 5,4	-248,11 ± 57,4	96,8 ± 12,73	-2,23 ± 2,42	1081,04 ± 79,03	-0.8
Cluster 8	-70,63 ± 16,73	15,15 ± 0,62	5	-34,62 ± 4,41	-192,83 ± 59,57	103,67 ± 63,1	-7,81 ± 3,61	1313,40 ± 215,60	-0.8
Cluster 7	-65,60 ± 4,29	9,67 ± 0,74	6	-39,02 ± 4,39	-164,04 ± 34,09	161,24 ± 51,45	-9,9 ± 3,59	1894,07 ± 183,54	-0.2
Cluster 3	-60,08 ± 9,1	17,0 ± 0,108	8	-45,68 ± 7,25	-126,7 ± 29,17	139,76 ± 47,73	-3,04 ± 2,18	1894,63 ± 91,28	0.4
Cluster 15	-59,32 ± 9,6	14,67 ± 0,36	4	-26,03 ± 3,65	-154,11 ± 45,43	151,11 ± 78,58	-17,58 ± 2,22	1390,59 ± 138,79	0.5
Cluster 10	-55,621 ± 8,71	8,69 ± 0,38	5	-41,63 ± 5,07	-133,64 ± 48,53	210,62 ± 17,86	-8,32 ± 2,641	1531,03 ± 60,26	0.9
Cluster 13	-51,91 ± 9,62	8,45 ± 0,23	4	-31,51 ± 1,42	-147,77 ± 46,38	102,68 ± 48,34	-1,11 ± 3,94	1763,68 ± 66,71	1.3
Cluster 5	-51,81 ± 6,39	15,79 ± 0,47	6	-21,58 ± 4,64	-227,95 ± 34,89	187,74 ± 32,8	-3,42 ± 3,19	1514,75 ± 347,34	1.4
Cluster 9	-49,53 ± 2,23	7,41 ± 0,5	5	-32,67 ± 6,47	-127,93 ± 13,81	136,99 ± 58,34	-4,97 ± 1,55	1351,98 ± 95,59	-
Cluster 12	-46,86 ± 10,24	6,36 ± 0,33	4	-32,11 ± 6,72	-120,58 ± 30,69	146,73 ± 47,19	-5,3 ± 4,38	1358,95 ± 169,95	-
Cluster 6	-46,68 ± 9,73	15,1 ± 0,51	6	-26,45 ± 5,82	-94,92 ± 36,35	155,94 ± 33,8	-16,84 ± 3,31	1436,41 ± 179,85	-
Cluster 4	-44,45 ± 2,67	15,01 ± 0,1	7	-37,43 ± 5,88	-84,38 ± 26,52	169,6 ± 50,44	-7,1 ± 3,44	1726,66 ± 68,45	-
Cluster 14	-43,1 ± 6,90	15,53 ± 0,16	4	-25,04 ± 8,88	-119,92 ± 47,05	105,15 ± 56,24	-4,59 ± 5,14	1232,98 ± 152,35	-
Cluster 11	-36,5 ± 3,97	9,21 ± 0,36	4	-27,93 ± 2,67	-74,58 ± 6,79	93,84 ± 23,62	-3,04 ± 1,58	1031,46 ± 93,54	-

a. *HADDOCK score = 1.0 E_{vdw} + 0.2 E_{elec} + 1.0 E_{desol} + 0.1 E_{air}*

b. Non-bonded interactions were calculated with the OPLS force field [215]

c. Empirical desolvation parameters were based on Fernandez et al [499](Fernández-Recio et al., 2004)(Fernández-Recio et al., 2004)(Fernández-Recio et al., 2004)

Table S 2: Statistical analysis for HADDOCK generated SRA-ApoL1 complexes from docking run 2. The HADDOCK score for each cluster was calculated using a summation equation of thermodynamic energies. From the top-scoring cluster 4 (in blue), complexes 3 and 4 used in further investigations were obtained.

Cluster rank	HADDOCK score [a] (a.u)	RMSD from lowest energy structure (Å)	Cluster size	Van der Waals energy[b] (E_{VdW}) (kcal mol^{-1})	Electrostatic energy[b] (E_{elec}) (kcal mol^{-1})	Restraints violation energy (E_{air}) (kcal mol^{-1})	Desolvation energy[c] (E_{desol}) (kcal mol^{-1})	Buried Surface Area (Å^2)	Z-Score
Cluster 4	**-59,61 ± 6,56**	**3,16 ± 0,61**	**11**	**-47,53 ± 11,52**	**-355,1 ± 48,80**	**532,16 ± 14,81**	**5,72 ± 4,68**	**2381,42 ±174,86**	**-1.9**
Cluster 3	-53,71 ± 9,25	15,46 ± 0,03	12	-52,72 ± 5,17	-262,14 ± 27,18	533,41 ± 58,21	-1,90 ± 5,68	2241,19 ±274,23	-1.4
Cluster 6	-41,82 ± 4,77	15,80 ± 0,138	9	-43,27 ± 13,66	-208,88 ± 54,2	506,98 ±115,83	-7,47 ± 3,74	2290,60 ±129,12	-0.5
Cluster 5	-40,18 ± 10,10	3,10 ± 0,46	9	-49,83 ± 7,37	-274,07 ± 36,71	503,99 ± 65,13	14,06 ± 8,76	2046,91 ±32,78	-0.4
Cluster 1	-35 ± 1,32	7,58 ± 0,61	33	-46,94 ± 14,34	-219,71 ± 54,37	543,24 ±81,44	1,56 ± 4,14	2080,98 ±235,63	0.0
Cluster 11	-28,50 ± 12,64	15,69 ± 0,15	5	-25,28 ±10,11	-320,36 ± 28,67	511,88 ± 75,01	9,67 ± 2,072	1653,05 ±146,62	0.5
Cluster 8	-28,32 ± 8,14	15,97 ± 0,245	7	-44,07 ± 3,59	-170,4 ±43,34	499,74 ± 55,28	-0,15 ± 3,45	1716,6 ± 95,49	0.5
Cluster 7	-26,83 ± 8,66	17,27 ± 0,06	7	-43,12 ± 7,48	-254,42 20,01	585,76 ± 37,60	8,60 ± 2,10	1813,97 ± 119,42	0.6
cluster 2	-20,02 ± 5,17	10,156 ± 0,28	12	-34,4 ± 9	-226,92 ± 71,9	548,45 ± 6,85	4,92 ± 3,64	1804,27 ±92,70	1.2
Cluster 10	-18,31 ± 8,23	17,05 ± 0,18	5	-43,31 ± 3,66	-202,05 ± 5,07	594,47 ± 58,61	5,97 ±2,51	1769,67 ± 83,50	1.3
Cluster 9	-15,95 ± 10,37	15,96 ± 0,08	5	-35,48 ± 6,69	-192,36 ± 26,73	471,72 ± 58,12	10,83 ± 1,03	1933,02 ± 41,323	-
Cluster 12	-15,7 ± 3,32	15,04 ± 0,13	4	-39,8 ± 11,01	-170,53 ± 26,93	573,92 ± 57,83	0,82 ± 1,03	1753,66 ± 123,06	-

(Jorgensen & Tirado-Rives, 1988)(Jorgensen & Tirado-Rives, 1988)(Jorgensen & Tirado-Rives, 1988)a. *HADDOCK score = 1.0 E_{vdw} + 0.2 E_{elec} + 1.0 E_{desol} + 0.1 E_{air}*

b. Non-bonded interactions were calculated with the OPLS force field [215]

c. Empirical desolvation parameters were based on Fernandez et al [499]

Table S 3: Statistical analysis for HADDOCK generated SRA-ApoL1 docked complexes from run 3. The HADDOCK score for each cluster was calculated using a summation equation of thermodynamic energies. From the top-scoring cluster 9 (in blue), complexes 5 and 6 used in further investigations were obtained

Cluster rank	HADDOCK score [a] (a.u)	RMSD from the lowest energy structure (Å)	Cluster size	Van der Waals energy [b] (E_{VdW}) (kcal mol^{-1})	Electrostatic energy [b] (E_{elec}) (kcal mol^{-1})	Restraints violation energy (E_{air}) (kcal mol^{-1})	Desolvation energy [c] (E_{desol}) (kcal mol^{-1})	Buried Surface Area (Å^2)	Z-Score
Cluster 9	**-46,54 ± 11,64**	**0,81 ± 0,48**	**7**	**-46,63 ± 9,96**	**-338,38 ± 64,89**	**569,26 ± 37,42**	**10,84 ± 2,34**	**2293,38 ± 161,22**	**-1.7**
Cluster 8	-45,4 ± 6,05	7,85 ± 0,29	8	-31,71 ± 6,79	-373,68 ± 39,2	455 ± 35,01	15,55 ± 1,366	1930,11 ± 59,86	-1.5
Cluster 1	-40,28 ± 10,26	15,93 ± 0,17	16	-58,19 ± 5,36	-186,52 ± 77,75	572,25 ± 33,52	-2,01 ± 3,23	2000,37 ± 109,76	-0.7
Cluster 10	-38,08 ± 2	4,31 ± 0,82	7	-38,3 ± 3,68	-345,39 ± 48,08	550,48 ± 84,89	14,254 ± 2,91	1866,57 ± 151,73	-0.4
Cluster 12	-34,3 ± 7,68	8,64 ± 0,12	5	-38,19 ± 6,04	-247,01 ± 31,27	477,49 ± 33	5,55 ± 0,95	1828,79 ± 176,01	0.1
Cluster 2	-33,9 ± 2,42	16,39 ± 0,1	15	-28,31 ± 1,55	-313,84 ± 21,35	504,42 ± 34,4	6,74 ± 1,45	1687,81 ± 5,73	0.2
Cluster 3	-29,9 ± 0,64	16,11 ± 0,15	14	-36,41 ± 1,1	-222,49 ± 28,17	495,28 ± 25,6	1,39 ± 3,15	1731,54 ± 96,68	0.8
Cluster 5	-28,88 ± 2,59	14,7 ± 0,11	13	-33,63 ± 3,91	-299,85 ± 41,71	534,06 ± 20,9	11,32 ± 4,7	1755,73 ± 94,41	0.9
Cluster 4	-28,34 ± 2,89	8,243 ± 0,172	13	-38,3 ± 9,38	-234,34 ± 50,03	503,07 ± 26,2	6,53 ± 2,42	1827,21 ± 58,65	1.0
Cluster 7	-26,75 ± 5,12	15,16 ± 0,186	10	-40,94 ± 3,75	-211,85 ± 36,34	529,81 ± 43,6	3,59 ± 2,71	1790,46 ± 75,29	1.3
Cluster 13	-24,55 ± 5,48	16,462 ± 0,584	4	-39,7 ± 4,06	-165,15 ± 20,6	415,09 ± 19,73	6,67 ± 2,05	1518,13 ± 40,29	-
Cluster 6	-12,18 ± 3,8	16 ± 0,052	10	-38,03 ± 3,24	-142,79 ± 41,34	473,39 ± 92,37	7,07 ± 2,86	1622,39 ± 128,47	-
Cluster 11	-11,07 ± 7,44	16,084 ± 0,205	6	-38,52 ± 9,37	-209,41 ± 50,87	640,95 ± 38,61	5,23 ± 5,90	1842,91 ± 94,96	-

a. *HADDOCK score = 1.0 E_{vdw} + 0.2 E_{elec} + 1.0 E_{desol} + 0.1 E_{air}*

b. Non-bonded interactions were calculated with the OPLS force field [215]

c. Empirical desolvation parameters were based on Fernandez et al [499]

Table S 4: The code names of the complexes from each of the docking runs

	Cluster - model	Codename
Run 1	Cluster 1-1	Complex-1
	Cluster 1-3	Complex-2
Run 2	Cluster 4-4	Complex-3
Run 3	Cluster 9-3	Complex-4
	Cluster 9-4	Complex-5

Table S 5: Destabilising protein-protein interface residues identified for each dimeric complex through in silico alanine scanning.

Dimeric structure	Destabilising residues a	
	SRA	ApoL1
Complex 1	55, 58, 59, 62, 63, 109, 117	338, 345, 352, 359, 360, 366, 367, 370, 374
Complex 2	62, 117	334, 345, 349, 352, 359, 374
Complex 3	55, 62, 117	320, 387, 391, 393, 397, 398
Complex 4	55, 59, 125	387, 391, 395
Complex 5	63, 129	380, 391

a. Destabilising alanine substitutions had free energy values ≥1 kcal mol^{-1}, while neutral values ranged from -0.8 to 0.99 kcal mol^{-1}.

Table S 6: SRA interface residues identified by computational alanine scanning. The residues in each complex were singly mutated to alanine using the ROBETTA webserver. The change in binding free energy indicates either destabilisation or stabilisation of the complex. The free energy scores in blue have a neutral effect on the complex

Interface residues	Change in complex binding free energy (ddG) kcal mol^{-1}				
	Complex 1	Complex 2	Complex 3	Complex 4	Complex 5
55	2.57	0.83	1.58	1.10	0.57
58	1.34	0.31	0.61	-	-
59	2.49	0.31	0.03	1.23	0.23
62	2.28	1.92	1.18	0.03	0.01
63	1.95	0.73	0.13	0.68	1.51
109	1.69	0.62	0.28	-	-
117	1.16	1.33	1.77	0.33	-
125	0.55	0.42	0.95	1.21	0.83
129	0.71	-	0.75	-	1.93
132	-	-	0.70	-	-

Table S 7: ApoL1 interface residues identified by computational alanine scanning. The residues in each complex were singly mutated to alanine using the ROBETTA webserver. The change in binding free energy indicates either destabilisation or stabilisation of the complex. The free energy scores in blue have a neutral effect on the complex

Interface residues	Change in complex binding free energy (ddG) kcal mol^{-1}				
	Complex 1	Complex 2	Complex 3	Complex 4	Complex 5
320	0.53	- 1	1.66	-	
334	1.19	.39	0.34	-	0.04
338	1.84	0.79	-	-	0.24
345	0.91	1.33	-	-	-
349	2.71	1.20	-	-	-
352	2.79	2.21	-	-	-
359	2.77	1.81	-	-	-
360	1.35	0.05	-	-	-
366	1.19	0.02	-	-	-
367	1.61	- -	-	-	-
370	1.09	0.05	-	-	-
374	-	1.27	-	- -	-
380	-	-	0.00	0.04	2.08
386	-	-	1.25	0.05	-
387	-	-	0.00	1.76	0.23
389	-	-	1.32	- 2	1.16
391	-	-	1.02	.43	0.03
393	-	-	-0.12	0.03	-0.17
395	-	-	1.39	2.64	0.08
397	-	-	2.45	-0.17	0.25
398		-		-	

Table S 8: Probable stability of SRA-ApoL1 complexes in solution as predicted by the PDBePISA server. ΔGint indicates the solvation free energy gained upon formation of the complex, while ΔGdiss indicates the free energy of complex dissociation. Complexes with ΔGdiss>0 are likely to be thermodynamically stable in solution

Complex	Interaction energy ΔG^{int} (kcal/mol)	Dissociation energy ΔG^{diss} (kcal/mol)	BSA (Å²)	Interface (Å²)	Complex significance score	Stab**ility** in solution
Complex1	-18.5	9.4	2410	1204.3	**1**	**yes**
Complex1 G1	-22	13.5	2680	1339.5	**1**	**yes**
Complex1 $G1^G$	-19.3	9.5	2270	1135	**1**	**yes**
Complex1 $G1^M$	-15.8	6.6	2070	1035.9	**1**	**yes**
Complex1 G2	-19.9	10.3	2240	1120.4	**1**	**yes**
Complex2	-19.9	12.7	2420	1209.7	**1**	**yes**
Complex2 G1	-13.9	5	2010	1006.5	**1**	**yes**
Complex2 $G1^G$	-12	5.1	2130	1065.9	**1**	**yes**
Complex2 $G1^M$	-16.4	8.1	2060	1027.7	**1**	**yes**
Complex2 G2	-17.8	8.1	2400	1202.4	**1**	**yes**

Interface area is calculated as the difference in total accessible surface areas of the isolated and interfacing structures divided by two. The Complexation Significance Score (CSS), which indicates how significant for assembly formation the interface is.

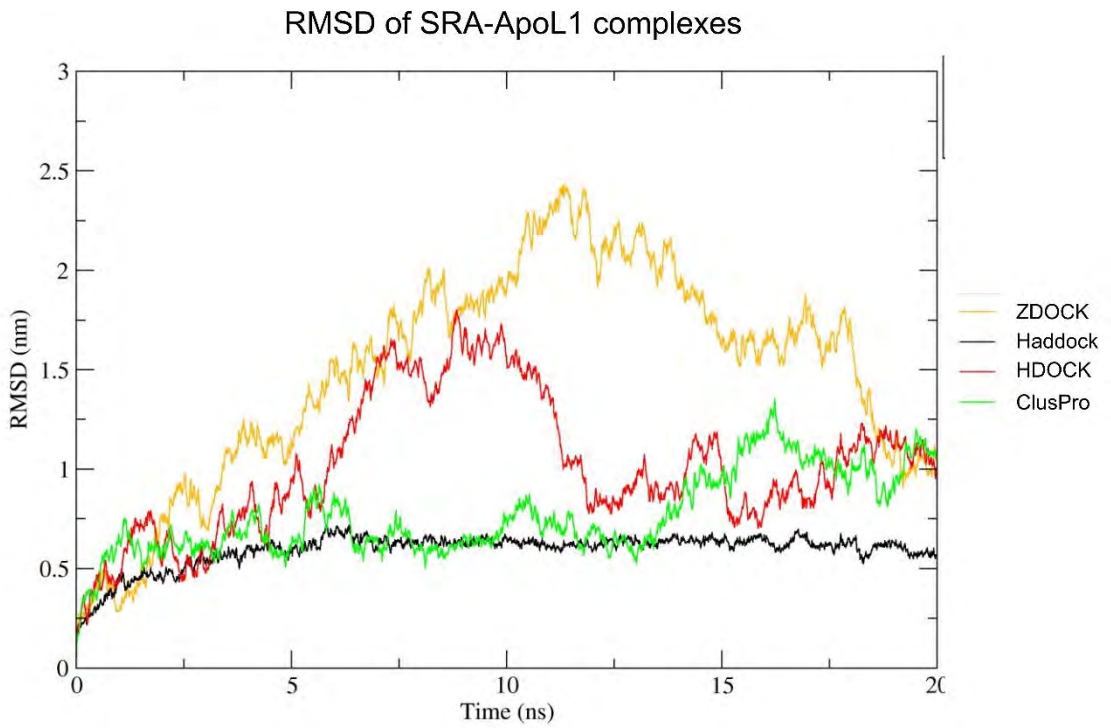


Figure S 8: RMSD of server generated SRA-ApoL1 complexes. RMSD was used to investigate stability of the complexes over the course of the simulations. ZDOCK, HDOCK and ClusPro output complexes dissociated after 20ns

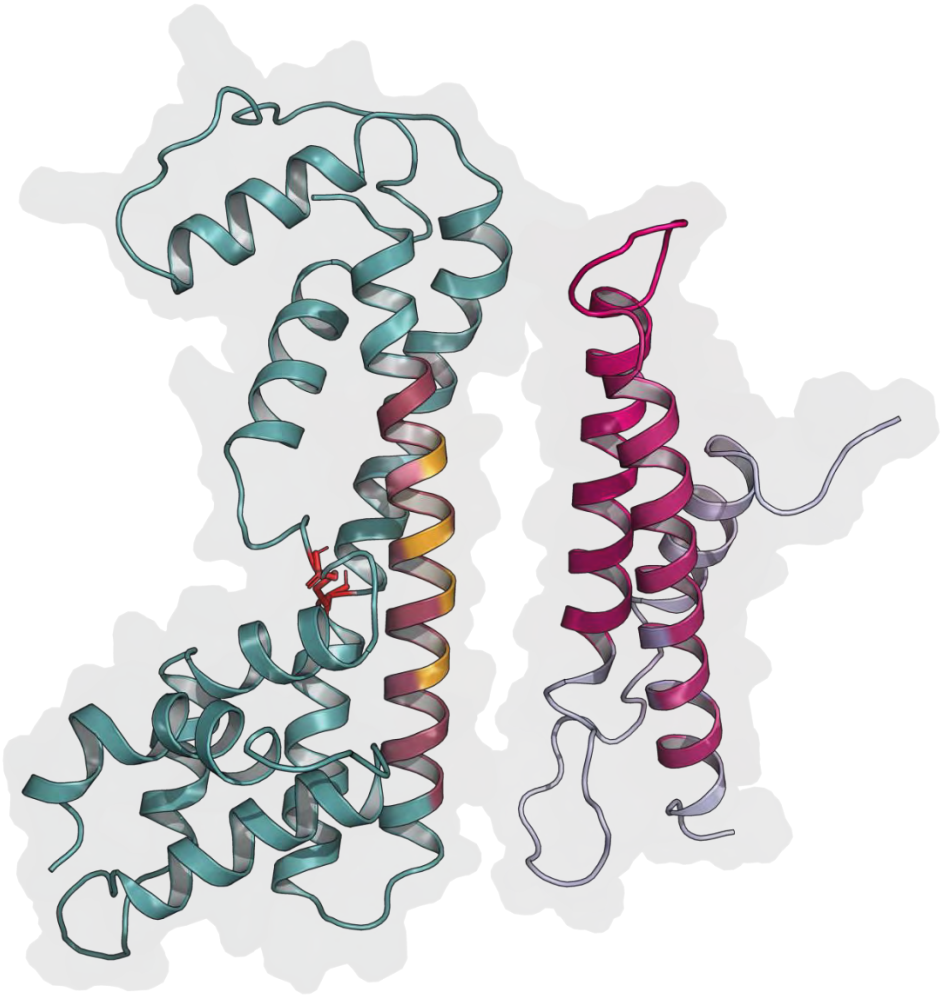


Figure S 9: Structures of SRA (in teal) and the ApoL1 C-terminal region (in pale blue) highlighting the AIR residues applied for HADDOCK protein-protein docking. SRA AIRs for Run 1 and 2 are coloured in raspberry, while Run 3 is distinguished by the orange colouring. AIRs applied for ApoL1 coiled-coil are coloured in hot pink.

(Structures source - PDB ID: 6elu (SRA) + homology model (ApoL1))

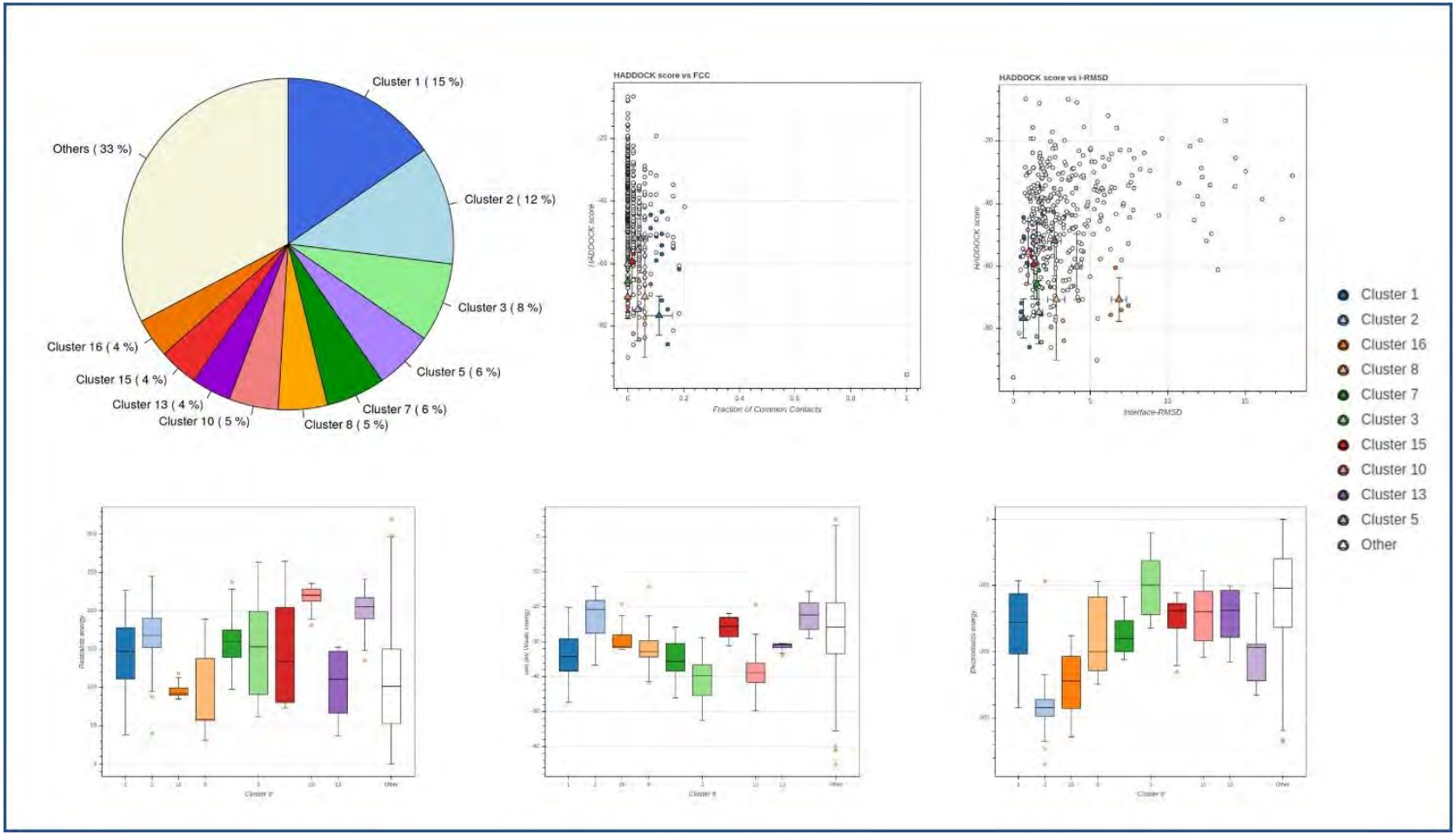


Figure S 10: HADDOCK run 1. The clusters represent are ranked based on FCC, and molecular energies.

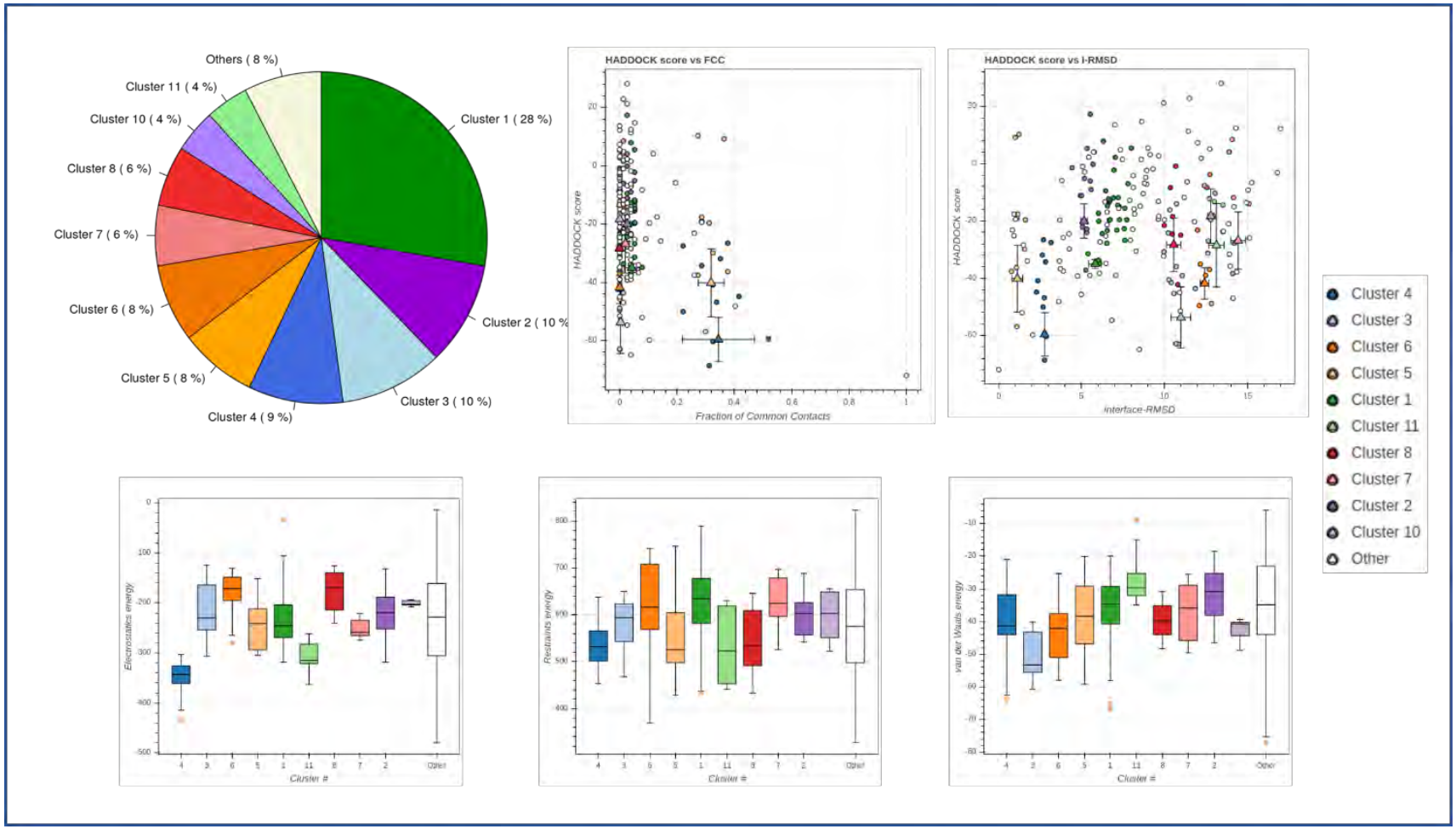


Figure S 11: HADDOCK run 2. The clusters represent are ranked based on FCC, and molecular energies

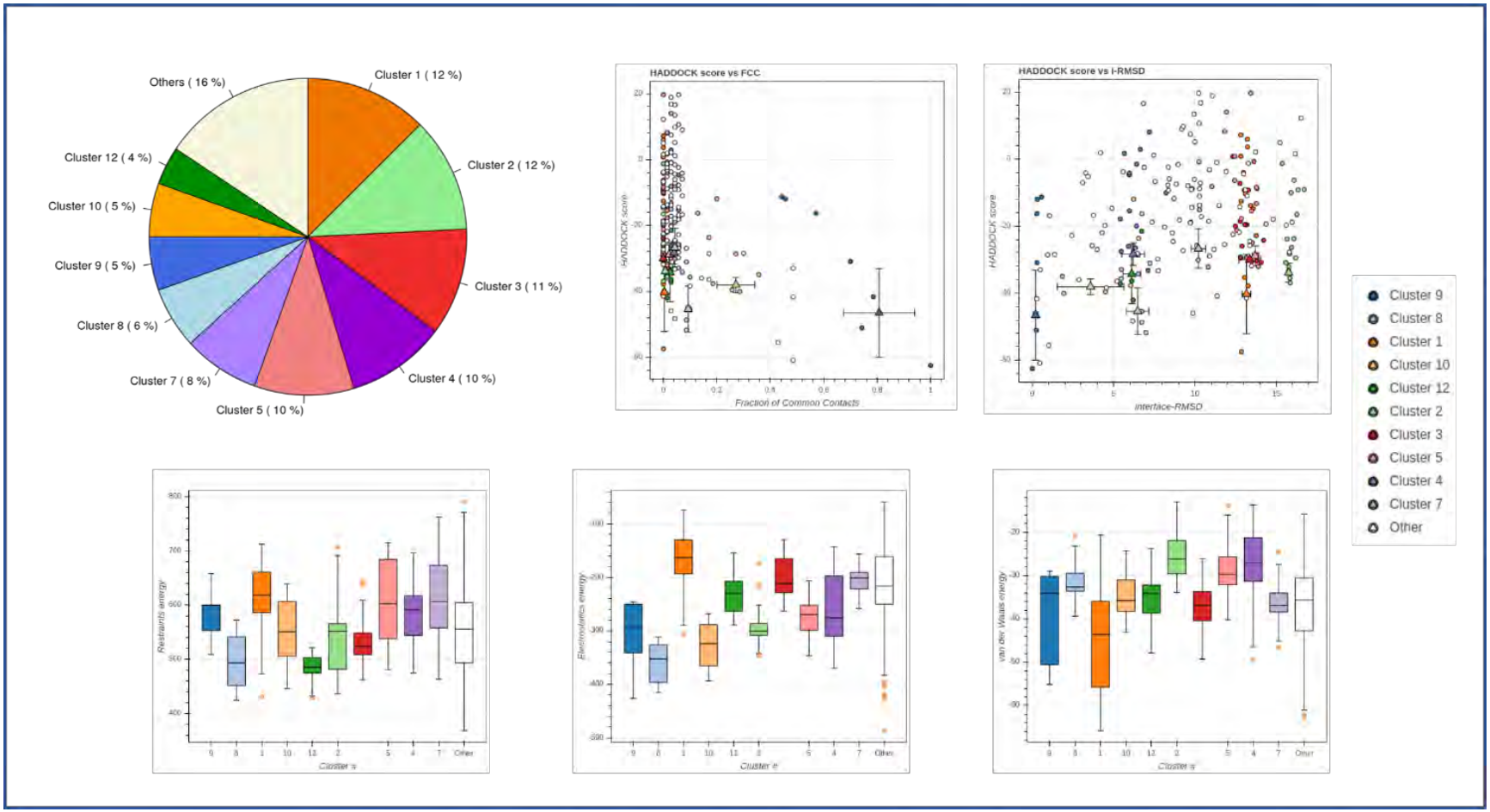


Figure S 12: HADDOCK run 3. The clusters represent are ranked based on FCC, and molecular energies.

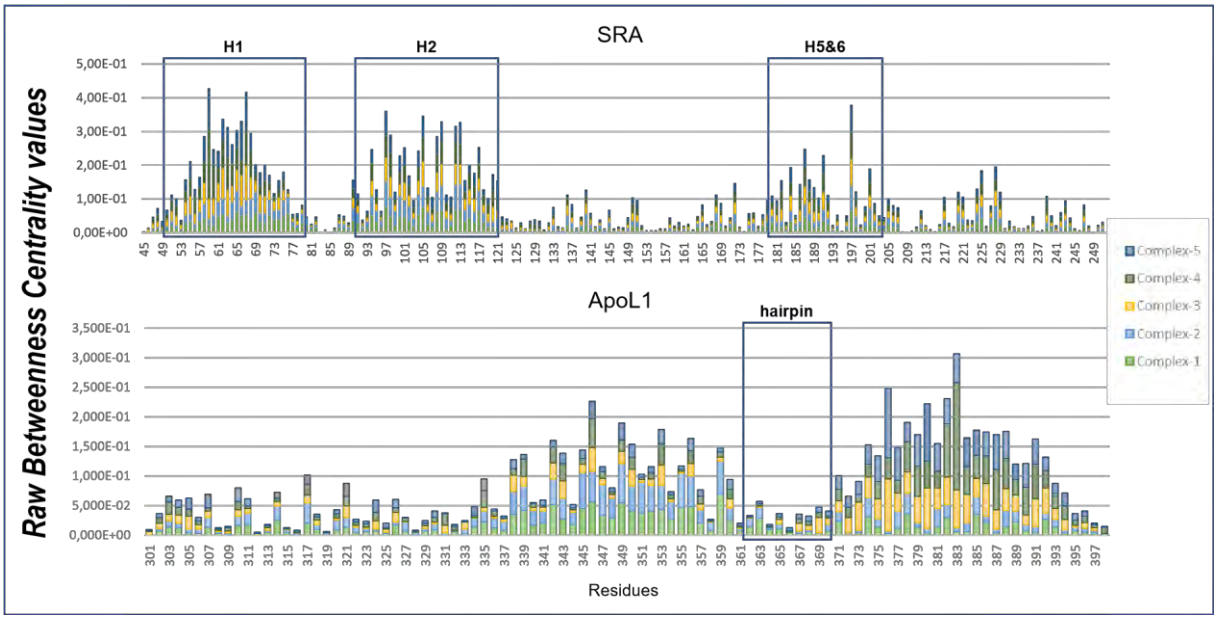


Figure S 13: Stacked bar graph of the raw BC values of complexed wild-type SRA and ApoL1 over 500ns

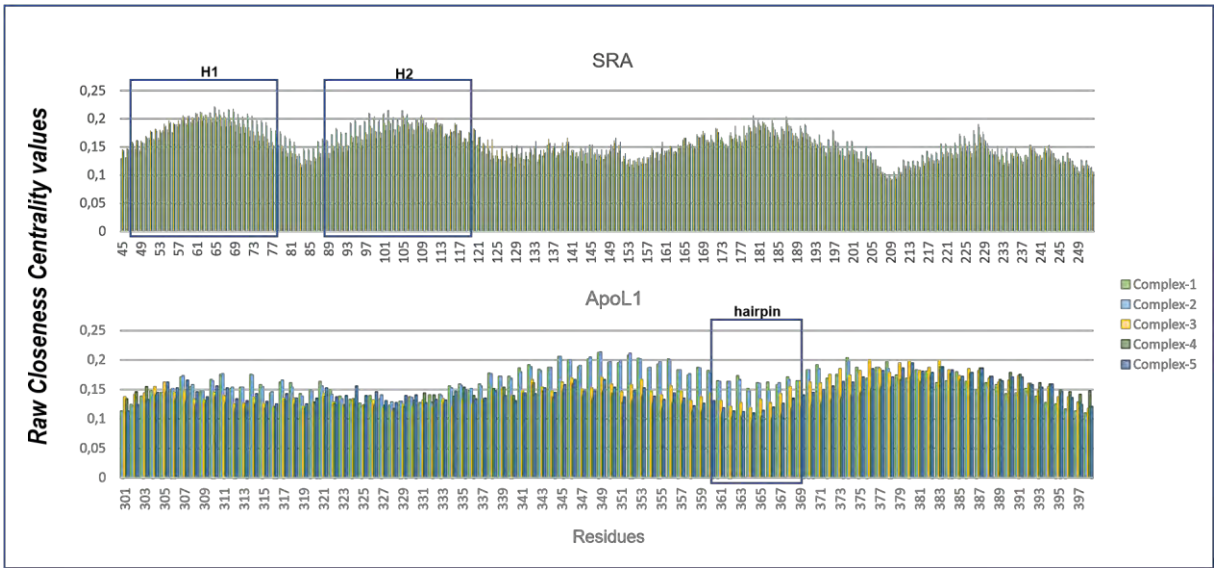


Figure S 14: Clustered bar graph of the raw CC values of complexed wild-type SRA and ApoL1 over 500ns

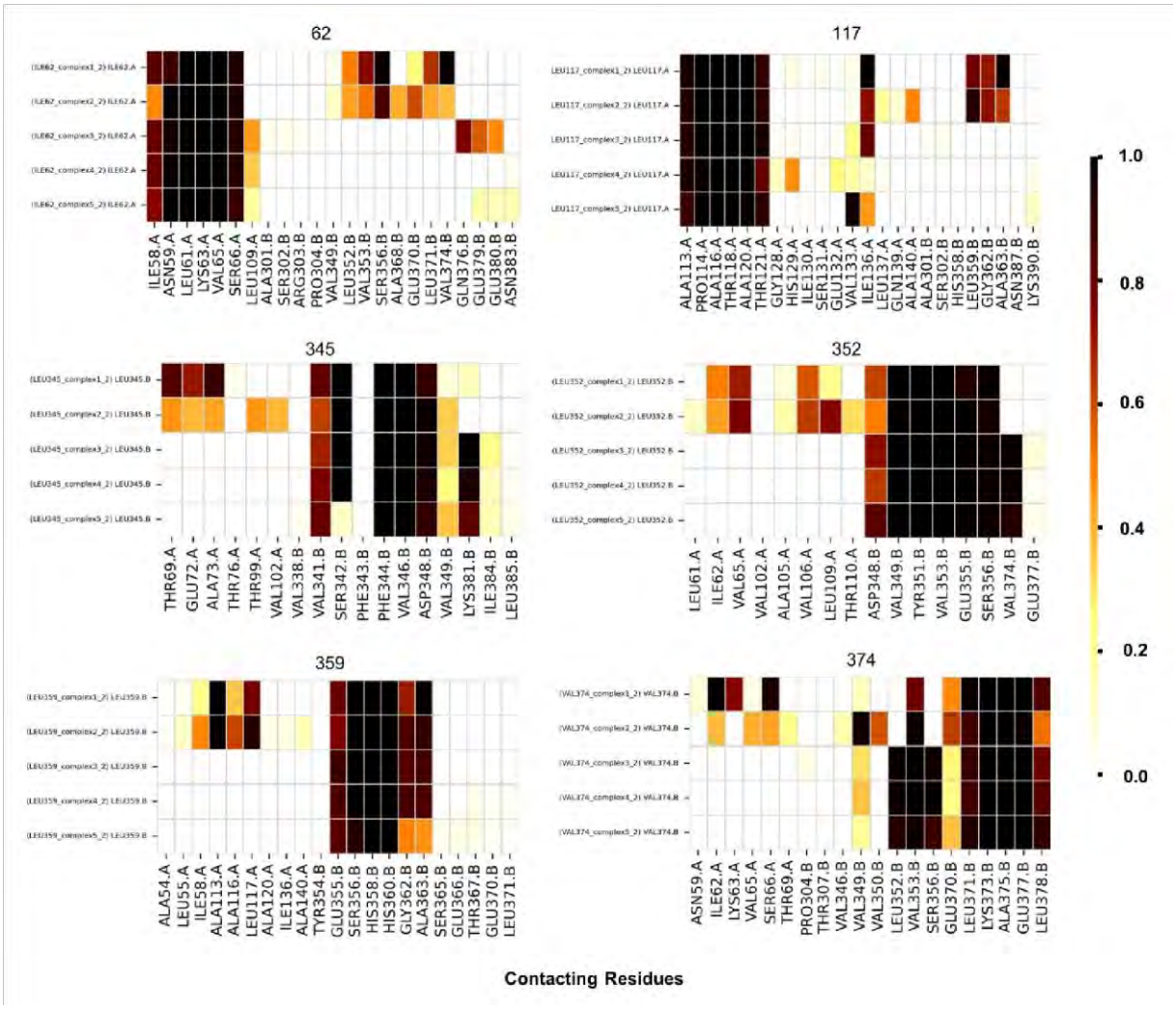


Figure S 15: Contact map analysis of key residues in wild-type SRA-ApoL1 complexes over 500ns

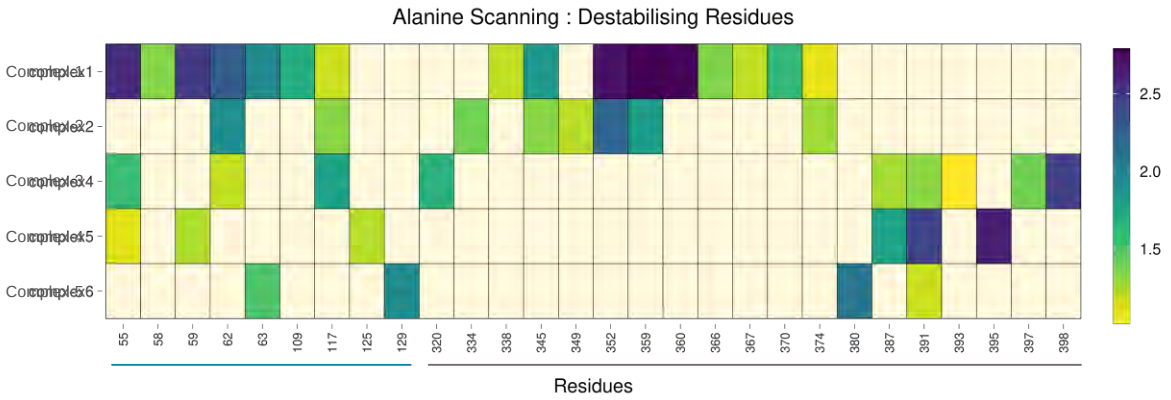


Figure S 16: Alanine scanning of wild-type SRA-ApoL1 complexes from an energy minima conformation.

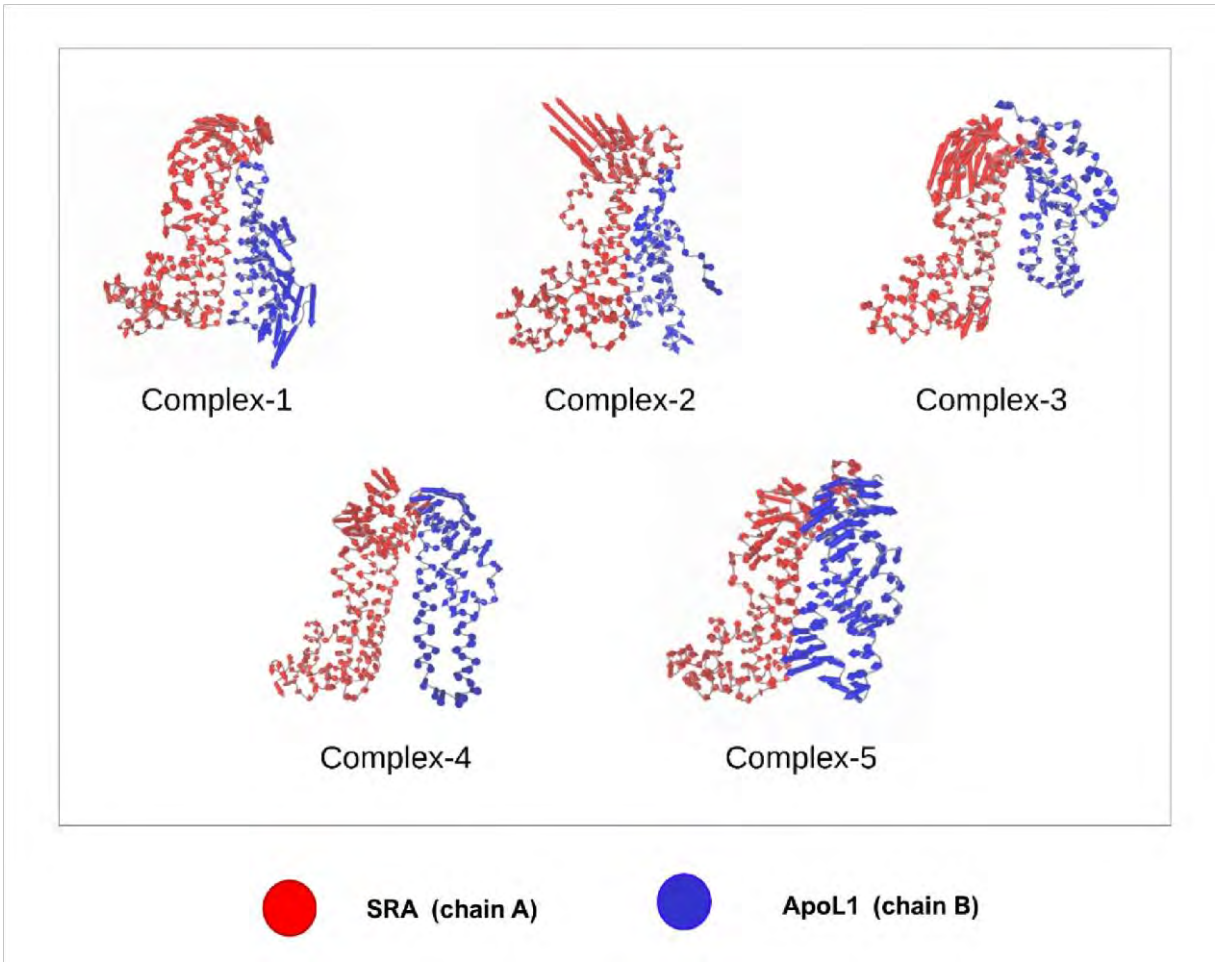


Figure S 17: Normal Mode analysis of the wild-type SRA-ApoL1complexes.The arrows are coloured based on each chain, denoting both the extent of motion and direction with respect to each of the residues.

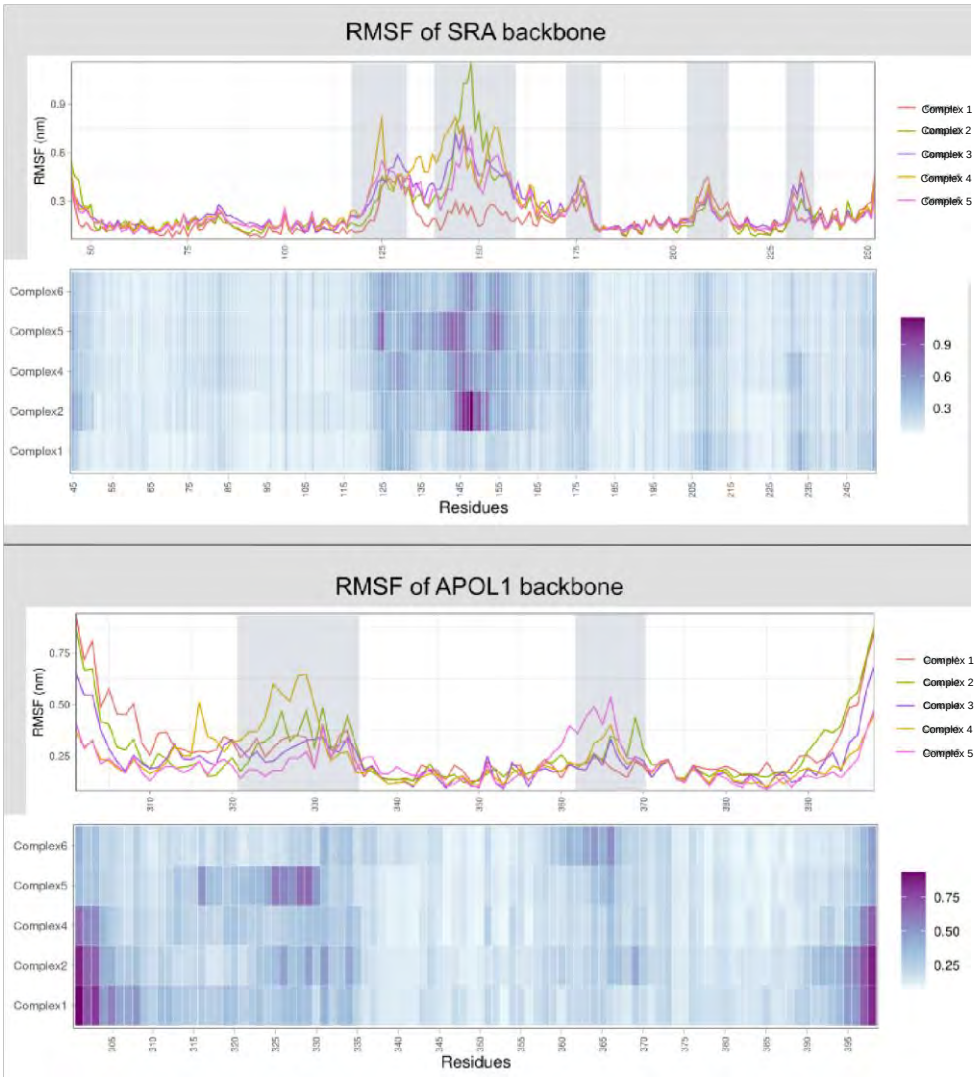


Figure S 18: RMSF analysis of wild-type SRA-ApoL1 complexes.

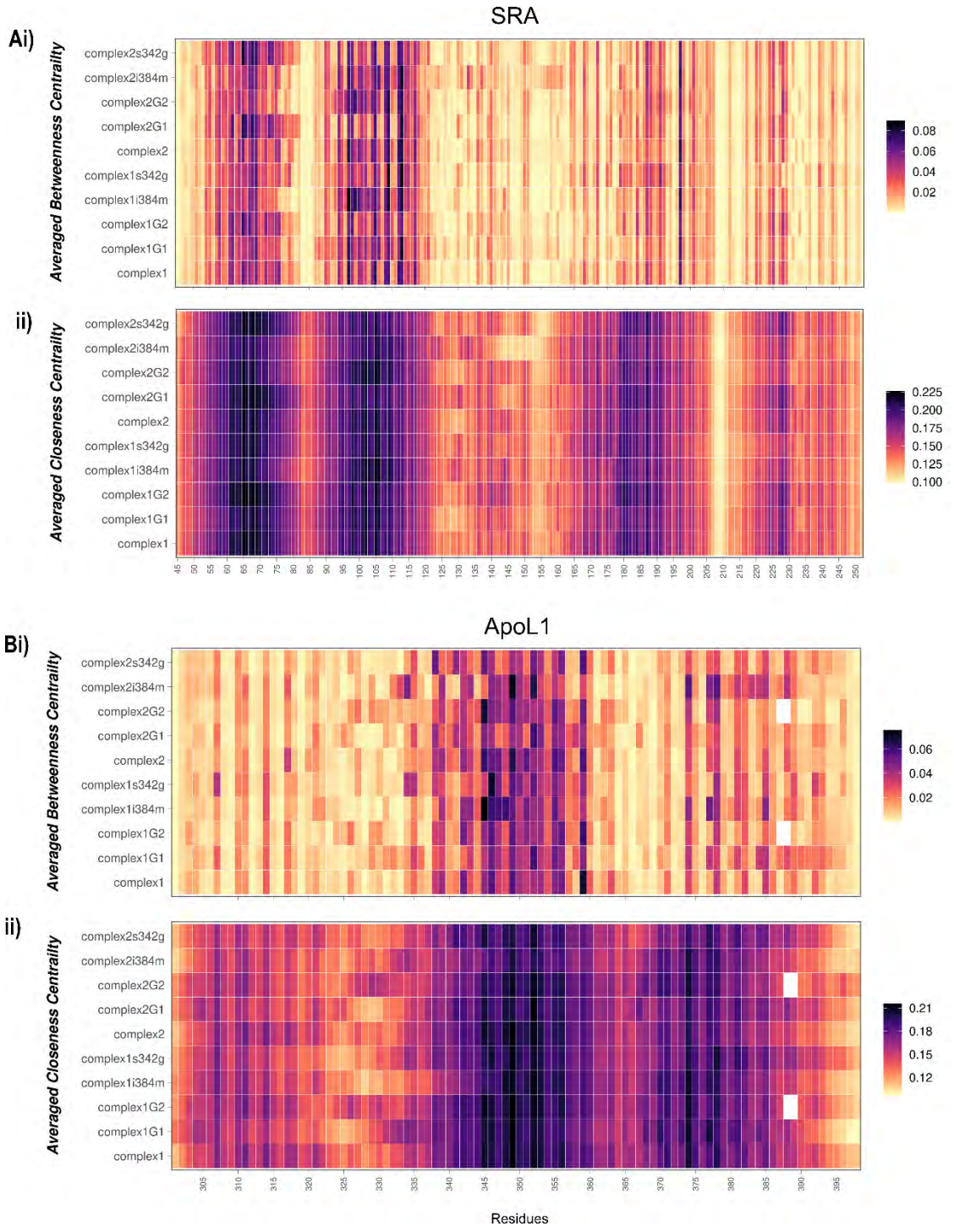


Figure S 19: All averaged centrality values of the SRA-ApoL1 variants complexes. A) SRA and B) ApoL1 values for i) BC and ii) CC over 500ns are measured

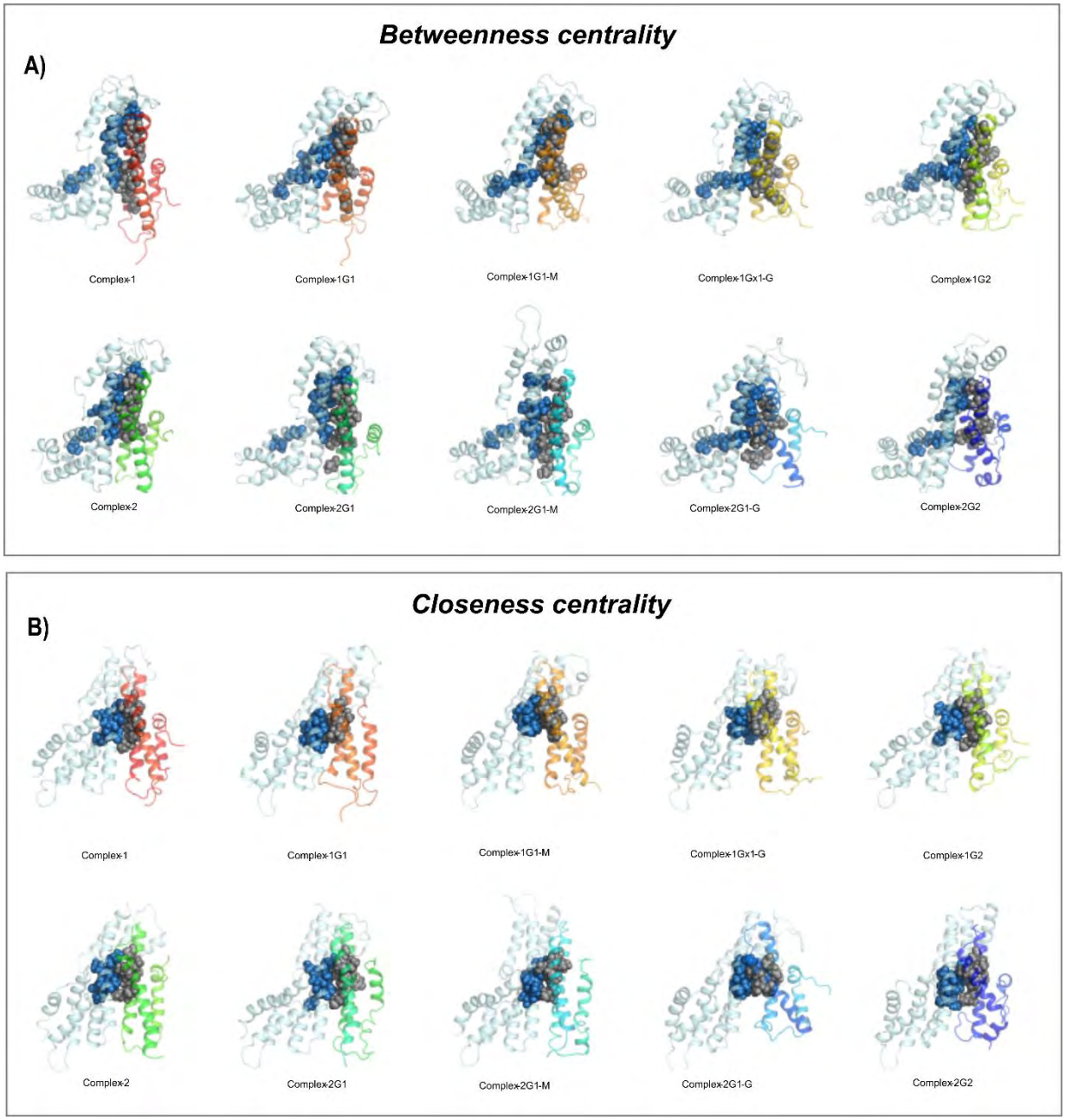


Figure S 20: Mapped BC an CC values onto the complex structures form the energy minima

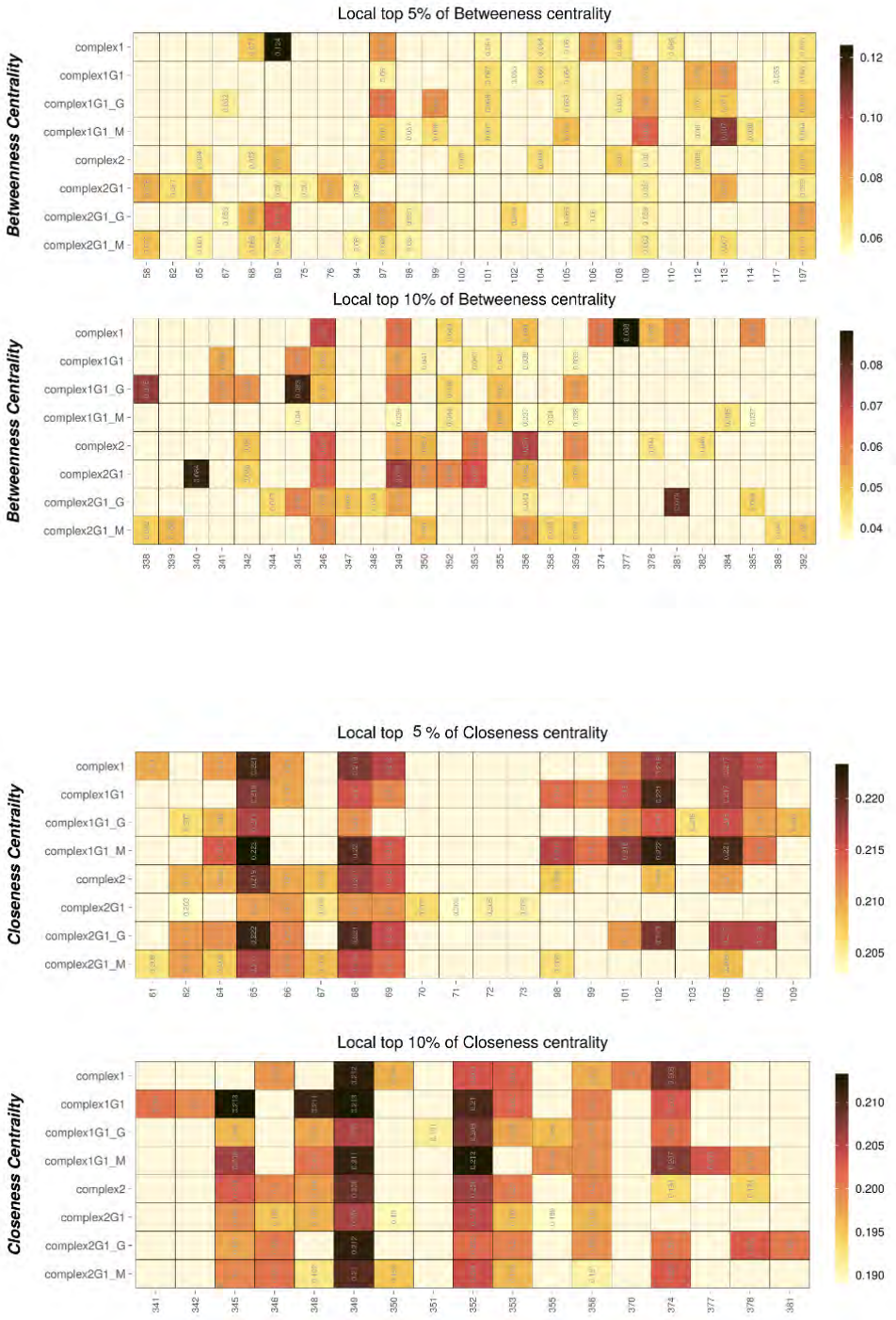


Figure S 21: BC and CC hubs observed in the SRA-ApoL1 complexes over a stabilised section of the MD simulations of a duplicate 500ns run. Detected hubs within the cut-off % were annotated with their centrality values for comparison against counterpart residues in other variants. For each metric, low to high centrality values correspond to the intensity of the colour bar on the right.

APPENDIX D : Chapter 5

Table S 9: Q-Vina W molecular docking parameters

Parameter	***Tb* PTR1**	***Tb* DHFR**
Box center X	31.006	-5.131
Box center Y	-0.043	12.638
Box center Z	92.900	26.350
Box size	101^3	101^3
Energy range	4	4
Exhaustiveness	120	120
CPU	4	4

Table S 10: The median values of ligand and backbone RMSD in TbDHFR

	median (nm)	
Ligand	*Tb*PTR1 ligand RMSD	*Tb*PTR1 protein RMSD
Control	0.03	0.293
DB00424	0.228	0.45
DB00472	0.152	0.341
DB00572	0.199	0.299
DB00757	0.101	0.390
DB00809	0.151	0.338
DB00891	0.122	0.375
DB00949	0.144	0.274
DB01151	0.200	0.221
DB02959	0.153	0.323
DB11254	0.134	0.283
DB11282	0.079	0.269
DB11285	0.137	0.282
DB11699	0.147	0.316

Table S 11: The median values of ligand and backbone RMSD in TbPTR1

	median (nm)	
Ligand	***Tb*DHFR ligand RMSD**	***Tb*DHFR protein RMSD**
Control	**0.05**	**0.172**
DB00424	**0. 209**	**0.199**
DB00472	**0.151**	**0.164**
DB00572	**0.192**	**0.154**
DB00757	**0.107**	**0.162**
DB00809	**0.117**	**0.210**
DB00891	**0.189**	**0.187**
DB00949	**0.133**	**0.179**
DB01151	**0.060**	**0.176**
DB02959	**0.043**	**0.169**
DB11254	**0.192**	**0.206**
DB11282	**0.074**	**0.184**
DB11285	**0.078**	**0.174**
DB11699	**0.129**	**0.200**

Table S 12: The p-values of protein backbone RMSD distribution in TbPTR1. The values were calculated using the Shapiro Wilks test for normality.

*Tb*PTR1 protein RMSD distribution *p-values*														
	Control	DB00424	DB00472	DB00572	DB00757	DB00809	DB00891	DB00949	DB01151	DB02959	DB11254	DB11282	DB11285	DB11699
Control	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00424	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00472	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00572	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00757	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00809	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00891	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00949	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB01151	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB02959	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11254	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11282	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11285	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$
DB11699	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1

Table S 13: The p-values of ligand RMSD distribution in TbPTR1. The values were calculated using the Shapiro Wilks test for normality.

*Tb*PTR1 ligand RMSD distribution *p-values*														
	Control	DB00424	DB00472	DB00572	DB00757	DB00809	DB00891	DB00949	DB01151	DB02959	DB11254	DB11282	DB11285	DB11699
Control	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00424	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00472	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00572	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00757	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00809	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00891	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00949	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB01151	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB02959	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11254	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11282	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11285	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$
DB11699	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1

Table S 14: The p-values of radius of gyration in TbPTR1. The values were calculated using the Shapiro Wilks test for normality.

*Tb*PTR1 radius of gyration distribution *p-values*														
	Control	DB00424	DB00472	DB00572	DB00757	DB00809	DB00891	DB00949	DB01151	DB02959	DB11254	DB11282	DB11285	DB11699
Control	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00424	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00472	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00572	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00757	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00809	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00891	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00949	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB01151	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB02959	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11254	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11282	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11285	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$
DB11699	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1

Table S 15: The p-values of protein backbone RMSD distribution in TbDHFR. The values were calculated using the Shapiro Wilks test for normality

*Tb*DHFR protein RMSD distribution *p-values*														
	Control	DB00424	DB00472	DB00572	DB00757	DB00809	DB00891	DB00949	DB01151	DB02959	DB11254	DB11282	DB11285	DB11699
Control	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00424	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00472	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00572	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00757	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00809	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00891	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00949	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB01151	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB02959	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11254	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11282	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11285	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$
DB11699	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1

Table S 16: The p-values of ligand RMSD distribution in TbDHFR. The values were calculated using the Shapiro Wilks test for normality.

*Tb*DHFR ligand RMSD distribution *p-values*														
	Control	DB00424	DB00472	DB00572	DB00757	DB00809	DB00891	DB00949	DB01151	DB02959	DB11254	DB11282	DB11285	DB11699
Control	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00424	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00472	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00572	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00757	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00809	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00891	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00949	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB01151	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB02959	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11254	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11282	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11285	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$
DB11699	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1

Table S 17: The p-values of radius of gyration distribution in TbDHFR. The values were calculated using the Shapiro Wilks test for normality.

*Tb*DHFR radius of gyration distribution *p-values*														
	Control	DB00424	DB00472	DB00572	DB00757	DB00809	DB00891	DB00949	DB01151	DB02959	DB11254	DB11282	DB11285	DB11699
Control	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00424	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00472	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00572	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00757	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00809	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00891	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB00949	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB01151	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB02959	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11254	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11282	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$
DB11285	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1	$< 2.2 \times 10^{-16}$
DB11699	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	1

Table S 18: The RMSD values at the peaks of the distribution in TbPTR1 and TbDHFR.

System	*Tb*PTR1 RMSD distribution peak (nm)		*Tb*DHFR RMSD distribution peak (nm)	
	i	ii	i	ii
Apo	0,449	-	0,159	-
Control	0,318	-	0,149	0,199
DB00424	0,449	-	0,149	-
DB00472	0,380	-	0,169	-
DB00572	0,299	0,399	0,149	-
DB00757	0,410	-	0,172	-
DB00809	0,374	-	0,249	-
DB00891	0,399	-	0,199	-
DB00949	0,299	-	0,199	0,249
DB01151	0,369	0,249	0,179	-
DB02959	0,3497	0,499	0,199	0,258
DB11254	0,299	-	0,279	-
DB11282	0,299	-	0,198	-
DB11285	0,299	0,359	0,199	-
DB11699	0,349	-	0,169	0,249

Table S 19: The BBB properties of the hit compounds identified.

DrugBank ID	DB00424	DB00472	DB00572	DB00757	DB00809	DB00891	DB00949	DB01151	DB02959	DB11254	DB11282	DB11285	DB11699
Target	Muscarinic acetylcholine receptor M1	5-hydroxy tryptamine receptor	Muscarinic acetylcholine receptor M1	5-hydroxy tryptamine receptor 3A	Muscarinic acetylcholine receptor M1	*Dihydropteroate synthase type-1	Glutamate receptor ionotropic, NMDA 2B	5-hydroxy tryptamine receptor	5-hydroxy tryptamine receptor	DNA topoisomerase 1	-	-	5-hydroxy tryptamine receptor 3A
BBB*	0,093	0,076	0,093	0,031	0,094	0,3	0,032	0,081	0,032	0,033	0,079	0,053	0,06

Compound	Methotrexate	Cyromazine	Pyrimethamine	Pentamidine
BBB*	-0.032	0.024	0.044	0.079

*values lower than 0.02 are classified as BBB- and are incapable of crossing the barrier.

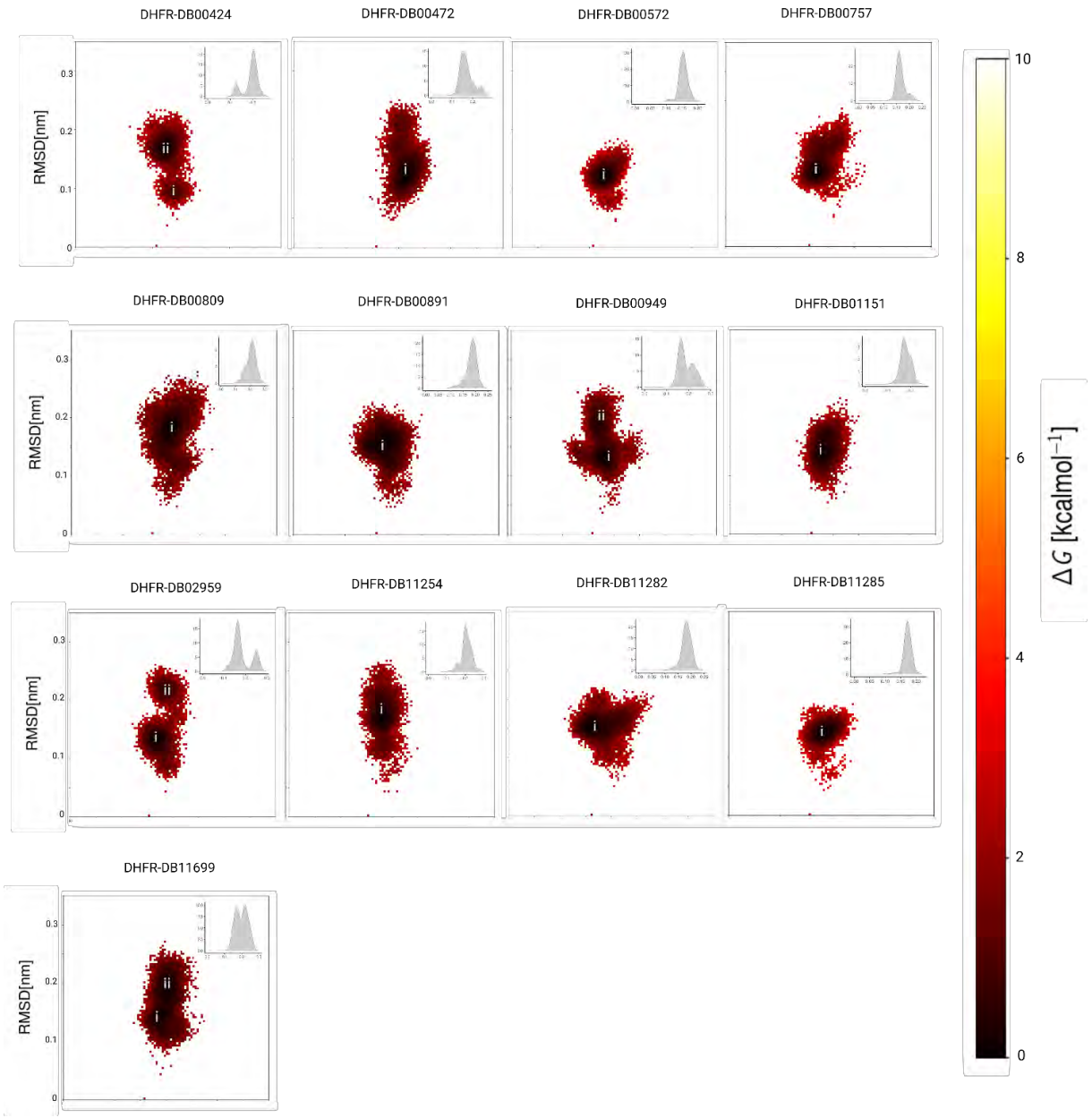


Figure S 22: RMSD and free energy landscapes of the TbDHFR-ligand complexes. The inserts show probability density distribution of RMSD, with the peaks corresponding with the lowest energy conformation.

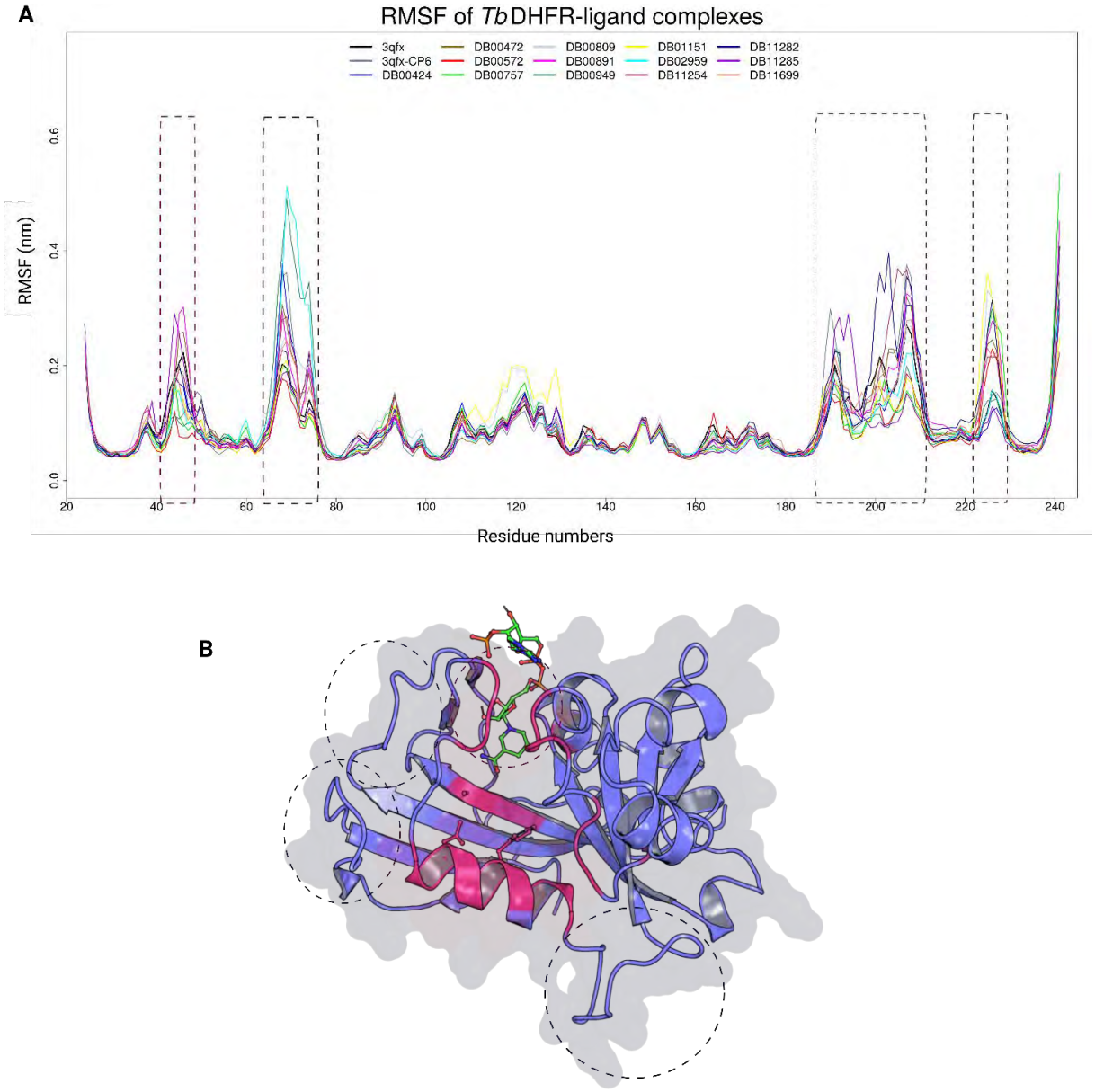


Figure S 23: RMSF of the TbDHFR-ligand complex protein backbone. The flexible regions are show in the dotted lines in both the a) line graph and b) the 3D structure.