

Computer simulations of biological systems: from protein dynamics to drug discovery

**A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville**

**Rupesh Agarwal
December 2020**

Copyright © 2020 by Rupesh Agarwal.
All rights reserved.

DEDICATION

I dedicate this work to my grandparents.

ACKNOWLEDGEMENTS

I am grateful to my siblings, Satish and Shradha, who have sparked my interest in science since childhood. I would also like to thank my parents for their unconditional support throughout my undergraduate and graduate education. I would also like to recognize the support provided by my undergraduate research mentors Dr. Deepak Modi, Dr. T.S. Vasulu, and Dr. Subhabrata Sengupta.

A very special gratitude goes out to Payal Chirania for putting up with my stress during the degree and continuously motivating me. My friend, Rakesh, has been there to support me and has shown me the humorous side of things. My colleagues Dr. Adam Green and Dr. Utsab Shrestha for being great friends in reviewing my presentations and the long, fruitful discussions.

Also, much gratitude is owed to Dr. Jerome Baudry, who has motivated me throughout my degree and also supported tremendously in all my drug discovery projects. With a special mention to Dr. Micholas Dean Smith, and Dr. Loukas Petridis for helping me throughout my graduate studies. Moreover, I would like to acknowledge Dr. Tessa Calhoun for being on my committee. I am also grateful to Dr. Tina Iverson, Dr. Barbara Bensing, Dr. Michael Duff, and all my other collaborators for helping me with my research.

Lastly, I would like to thank my advisor, Dr. Jeremy Smith, for believing in me, for his willingness to listen to my concerns and lightening things up with his sense of humor.

ABSTRACT

Computational biophysics methods such as molecular dynamics (MD) simulations are often used in combination with experimental techniques like neutron scattering, NMR, and FTIR to explore protein conformational landscapes. With the improvements in experimental techniques, there is also a need to continually optimize the MD forcefield parameters to make precise predictions that match experimental results. To complement many of these experiments, an accurate model of deuteration is frequently required, but has been elusive. In our work, we developed a novel method to capture isotope effects in classical MD simulations by re-parameterization of the bonded terms of the CHARMM forcefield using quantum mechanical (QM) calculations.

Apart from this, MD simulations can also be applied to explore a range of protein motions over different timescales, which are otherwise experimentally challenging. This work captures three such studies on protein dynamics- 1) the role of a) global and b) local motions in facilitating ligand binding to various adhesin protein homologs, and 2) a comparative study on the effect of temperature and pressure changes on the dynamics of thermophilic and mesophilic pyrophosphatases. In the case of adhesin proteins, we identified that the local motion in the loops near their binding pockets is critical for ligand selectivity, whereas the global inter-domain orientation in the protein is important for binding to the platelets. In the case of pyrophosphatases, our studies revealed that the number of hydrogen bonds, in the respective catalytic pockets of the two homologs, vary with temperature which potentially causes the observed differences in their experimental enzymatic activities.

Finally, another emerging application of computational biophysics is in the field of therapeutic research, i.e., to identify new drugs and therapies to cure lethal diseases by incorporating information about the target protein's dynamics into structure-based drug discovery. Implementing a pipeline that includes ensemble

docking and consensus scoring, we successfully targeted two proteins: 1) Histone deacetylase (HDAC) 4 and 2) adhesin protein Hsa, which are known to cause prostate cancer and infective endocarditis, respectively. For both the targets, we identified multiple novel small molecules that also inhibit these proteins in-vitro.

TABLE OF CONTENTS

1	Introduction	1
1.1	Protein flexibility.....	2
1.2	From Static to Ensemble	2
1.3	Exploring the protein free energy landscape	6
1.4	Molecular Dynamics simulation	6
1.5	Enhanced sampling simulation	8
1.6	Protein dynamics in drug discovery	9
1.7	Early stage drug discovery pipeline	10
1.8	Proposed study.....	16
2	Parameterization of Small Deuterated Molecules	18
2.1	Abstract	19
2.2	Introduction.....	20
2.3	Methods.....	23
2.4	Results.....	29
2.5	Discussion	33
2.6	Conclusion.....	35
2.7	Appendix.....	37
3	Studying global and local motion/dynamics of Bacterial proteins.....	62
3.1	Promiscuous binding of bacterial Siglec-like adhesin Hsa to human and animal sialoglycan receptors	63
3.1.1	Abstract	63
3.1.2	Introduction.....	63
3.1.3	Methods.....	66
3.1.4	Results and Discussion	67
3.1.5	Conclusion.....	72
3.1.6	Appendix	73
3.2	Precise orientation of bacterial adhesin is critical for host adherence ..	78

3.2.1	Abstract	78
3.2.2	Introduction.....	79
3.2.3	Methods.....	81
3.2.4	Results and Discussion	84
3.2.5	Conclusion.....	87
3.2.6	Appendix	89
4	Understanding the effect of non-native conditions on global and local dynamics of a protein.....	96
4.1	Abstract	97
4.2	Introduction.....	97
4.3	Methods.....	100
4.4	Results.....	102
4.5	Discussion and Conclusions.....	105
4.6	Appendix.....	110
5	Early Stage Structure-Based Drug Discovery	119
5.1	Competitive inhibitors for Serine-Rich Repeat (Srr) Adhesin	120
5.1.1	Abstract	120
5.1.2	Introduction.....	121
5.1.3	Methods.....	123
5.1.4	Results and Discussion	128
5.1.5	Conclusions.....	133
5.1.6	Appendix	135
5.2	Disruption of histone deacetylases 4 (HDAC4) complexation by novel selective inhibitors	150
5.2.1	Abstract	150
5.2.2	Introduction.....	150
5.2.3	Methods.....	154
5.2.4	Results and Discussion	156
5.2.5	Conclusions.....	158

5.2.6	Appendix	160
6	Conclusions and Future Directions	167
6.1	Conclusions	168
6.2	Future directions	172
	References	175
	Vita.....	201

LIST OF TABLES

Table 2.1 Bonded terms of new D ₈ THF parameters	57
Table 2.2 Angle term of new D ₈ THF parameters	57
Table 2.3 Dihedral term for new D ₈ THF parameters.....	58
Table 2.4 Lennard-Jones potential term of new D ₈ THF parameters	58
Table 2.5 Bonded terms of new THFK parameters.....	59
Table 2.6 Angle term of new THFK parameters.....	59
Table 2.7 Dihedral term for new THFK parameters	60
Table 2.8 Lennard-Jones potential term of new THFK parameters.....	60
Table 2.9 Cutoffs for Coordination Number Calculations	61
Table 3.1 System and simulation details of GspB.....	95
Table 4.1 Binding pocket residues used for the calculations	117
Table 4.2 Molecular dynamics simulations details	117
Table 4.3 Hexameric crystal structure comparison	118
Table 5.1 Structure of the nine hits and interaction details in the binding pocket of the Hsa _{BR}	148
Table 5.2 Residues included during clustering of MD simulation snapshots to form an 'ensemble'.....	149
Table 5.3 Validated compound list with their respective percentage control.....	149
Table 5.4 IC ₅₀ values for compounds that inhibited HDAC3 and HDAC4 (Selective inhibitors shown in green and Known inhibitors shown in Yellow)	166

LIST OF FIGURES

Figure 1.1 Schematic representation of various protein motions at different time scales.	4
Figure 1.2 Representative image showing the free energy landscape of a protein with multiple minima. Each well corresponds to a minimum energy conformation and the arrows represent the transition states. Adapted from ref ³¹	5
Figure 1.3 Flowchart showing different steps of a molecular dynamics simulation: 1) System preparation; 2) Minimization/Relaxation; 3) Equilibration and 4) Production.....	12
Figure 1.4 Ensemble docking: A cartoon showing multiple protein conformations targeted with small molecules.....	13
Figure 1.5 Drug discovery pipeline.....	15
Figure 1.6 Dissertation division into chapters	17
Figure 2.1 Initial configuration of MD simulation of the 16.5nm-side box (Red: THF ; Blue: Water).....	37
Figure 2.2: Infrared spectra of D8THF (green), HTHF (blue), THF (red) as calculated by MD simulations (lines) and experiments (peaks)	37
Figure 2.3: Density profiles of THF for different time intervals at 90°C. The top row is the density profiles averaged over X, Y and Z directions for first 50 ns and the bottom row is the density profiles averaged over X, Y and Z directions for last 50 ns	38
Figure 2.4: Density profiles of HTHF for different time intervals at 90°C. The top row is the density profiles averaged over X, Y and Z directions for first 50 ns and the bottom row is the density profiles averaged over X, Y and Z directions for last 50 ns	39
Figure 2.5 Density profiles of D8THF for different time intervals at 90°C. The top row is the density profiles averaged over X, Y and Z directions for first 50 ns	

and the bottom row is the density profiles averaged over X, Y and Z directions for last 50 ns	40
Figure 2.6 Density profiles of THFK for different time intervals at 90°C. The top row is the density profiles averaged over X, Y and Z directions for first 50 ns and the bottom row is the density profiles averaged over X, Y and Z directions for last 50 ns	41
Figure 2.7 Individual components of the second virial coefficient: a) D ₈ THF-D ₈ THF, HTHF-HTHF, and THF- THF, b) D ₈ THF, HTHF, and THF -water c) water-water. The standard error of the mean (SEM) are shown as error bars, and those not visible are at most the size of the symbol.....	41
Figure 2.8 Average number of hydrogen bonds normalized by the number of water molecules: a) D ₈ THF, HTHF, and THF -water; b) water-water. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.	42
Figure 2.9 Mean hydrogen bond lifetimes: a) D ₈ THF, HTHF, and THF -water; b) water-water. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.	42
Figure 2.10 Visual representation of coordination sites	43
Figure 2.11 Water-water coordination numbers. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.	43
Figure 2.12 Coordination numbers between D ₈ THF- D ₈ THF, HTHF- HTHF, and THF-THF. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.	44
Figure 2.13 Coordination numbers between D ₈ THF, HTHF, and THF -water. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.	45
Figure 2.14 Infrared spectra of THF and THFK as calculated by MD simulations (lines) and experiments (peaks).....	46

Figure 2.15 Density of THF for different time interval of 50ns at 90°C.	47
Figure 2.16 Density of HTHF for different time interval of 50ns at 90°C.	48
Figure 2.17 Density of D ₈ THF for different time interval of 50ns at 90°C.	49
Figure 2.18 Density of THFK for different time interval of 50ns at 90°C.	50
Figure 2.19 Individual components of the second virial coefficient: a) D ₈ THF- D ₈ THF, HTHF-HTHF, THF- THF, and THFK-THFK b) D ₈ THF, HTHF, THFK and THF -water c) water-water. Error bars show the SEM, and those not visible are at most the size of the symbol.	51
Figure 2.20 Average number of hydrogen bonds normalized by the number of water molecules: a) D ₈ THF, HTHF, THFK, and THF -water; b) water-water. Error bars show the SEM, and those not immediately visible are at most the size of the symbol.	51
Figure 2.21 Mean hydrogen bond lifetimes: a) D ₈ THF, HTHF, THFK, and THF - water; b) water-water. Error bars show the SEM, and those not immediately visible are at most the size of the symbol.	52
Figure 2.22 Co-ordination numbers between water-water. Error bars show the SEM, and those not immediately visible are at most the size of the symbol.	52
Figure 2.23 Co-ordination numbers between D ₈ THF- D ₈ THF, HTHF- HTHF, THFK- THFK, and THF-THF. Error bars show the SEM, and those not immediately visible are at most the size of the symbol.	53
Figure 2.24 Co-ordination numbers between D ₈ THF, HTHF, THFK, and THF - water. Error bars show the SEM, and those not immediately visible are at most the size of the symbol.	54
Figure 2.25 Probability distribution of angles for THF (red) and D ₈ THF(green) in presence of water molecules at 105 °C	55
Figure 2.26 QM IR spectra (top) and corresponding QM normal modes for THF and D ₈ THF (bottom)	56

Figure 3.1 Two-dimensional structure of disaccharide: N-glycolylneuraminic acid (Neu5Gc)-Galactose (Gal), showing the important atom names and the additional C11- hydroxyl group, substitution (with hydrogen) of which makes it N-acetylneuraminic acid (Neu5Ac).....	73
Figure 3.2 a) B-factor from the crystal structures of Hsa _{BR} bound to Neu5Ac-Gal (in blue) and Neu5Gc-Gal (in red); b) the superimposition of the overall two crystal structures showing two domains (Siglec and Unique), and focused view of the binding pocket showing side-chain residues and the hydrogen bonds between backbone carbonyl-O atom in Lys 335 of the EF loop and Neu5Ac-Gal's hydroxyl-O (O4) group in white dotted line.	73
Figure 3.3 Root mean square fluctuation (RMSF) of the Siglec domain of Hsa _{BR} bound to Neu5Ac-Gal (in blue) and Neu5Gc-Gal (in red). The error bars are standard error of mean (SEM) over three independent simulations.....	74
Figure 3.4 (a) Superimposition of the crystal structures of Hsa _{BR} bound Neu5Ac-Gal (in blue), Neu5Gc-Gal (in red), and MD snapshots (in white) with distance (D) between Lys 335 backbone carbonyl and Neu5Ac-O4 atom shown with a black dotted line. Distance (D) over simulation time of three independent runs of Hsa _{BR} bound to (b) Neu5Gc-Gal and (c) Neu5Ac-Gal.	74
Figure 3.5 Probability distribution of root mean square deviation (RMSD) of disaccharide: Neu5Ac-Gal (a) and Neu5Gc-Gal (b) bound to Hsa _{BR} over three independent simulation runs.	75
Figure 3.6 Number of hydrogen bonds between Tyr 368 in SrpA _{BR} and C11 hydroxyl group of Neu5Gc.	75
Figure 3.7 Binding pocket of a) SrpA _{BR} ; b) Hsa _{BR} ; c) GspB _{BR} adhesin proteins. Top panel shows the lipophilicity surface view of the pocket bound to Neu5Gc (red stick); Bottom panel shows the key residues (white stick) in the pockets of the corresponding proteins.	76
Figure 3.8 Interaction map of Neu5Ac-Gal and Neu5Gc-Gal in Hsa.....	77

Figure 3.9 Open and closed conformation of Hsa loop bound to different substrate.	77
Figure 3.10 Pictorial representation of the hypothesis showing that the binding site orientation is critical for binding to the platelets. Note: The picture is not to scale.	89
Figure 3.11 Probability distributions of the torsion angle (Φ) between the Siglec and Unique domains of adhesin proteins GspB, SK150, Hsa, NCTC10712, SK678, and SrpA (left) as calculated from MD simulations. Crystal structures showing the Φ angle for both GspB-like and Hsa-like proteins (right). The interdomain torsion angles in the corresponding crystal structures are as follows: GspB: $\sim 100^\circ$; SK150: $\sim 100^\circ$; NCTC10712: $\sim 228^\circ$; Hsa: $\sim 230^\circ$; SrpA: $\sim 216^\circ$; SK678: $\sim 240^\circ$	90
Figure 3.12 Two dimensional histogram of interdomain torsion (Φ) vs angle (θ) between Siglec and Unique for GspB systems: SigUn _{Wild} and SigUn _{mutALL} ..	91
Figure 3.13 Two dimensional histogram of interdomain torsion (Φ) vs angle (θ) between Siglec and Unique for GspB systems: SigUn _{Wild} , SigUn _{R585D} and SigUn _{R588D}	92
Figure 3.14 Comparative binding of wild-type and mutant BRs of GspB to immobilized platelets	93
Figure 3.15 Root mean square fluctuation (RMSF) of Siglec domain for GspB systems: SigUn _{Wild} , SigUn _{R585D} and SigUn _{R588D}	93
Figure 3.16 a) Binding of biotinylated sTa at different concentrations to immobilized wild-type and mutants R585D, R588D, L442K, Y485F of GspB _{BR} . Binding is reported as the mean $\pm$ standard deviation, with n = 3.	94
Figure 4.1 Mean square displacement (MSD) of C-alpha atoms for (a) hexameric TtPPase and (b) hexameric EcPPase. Higher values of MSD for TtPPase at all conditions indicate higher structural flexibility than EcPPase. Error bars shown in this and the subsequent figures are the standard error of the mean, and those not immediately visible are at most the size of the symbol.....	110

Figure 4.2 a) Probability distribution for solvent accessible surface area (SASA) of the hexameric forms for TtPPase (triangle) and EcPPase (circle); b) The normalized number of interchain hydrogen bonds.	110
Figure 5.1 Structure based virtual screening strategy workflow showing number of compounds which were passed on to the next step.	135
Figure 5.2 a) Root mean square fluctuation of Hsa _{BR} from MD simulation showing DC, EF, FG loop regions; b) Superimposed structures (in ribbon) of Hsa _{BR} obtained from different clusters showing residues (in stick) used to during the clustering: crystal structure (in red) and ensemble structure (in shade of blue).	136
Figure 5.3 Density profile of physicochemical properties of small molecule database: molecular weight (a), number of rotatable bonds (b), polar surface area (c), and Log of the octanol/water partition coefficient: SLogP (d).	136
Figure 5.4 a) Interaction map of native ligand (sTa) from crystal structure (PDB 6EFD) ¹⁹⁸ ; b) Bar plot showing number of compounds (within the top 50 compounds) and the number of interactions (Hbond or Pi (π - π / π -H/ π -cation) made with protein residues.....	137
Figure 5.5 Alpha Screen assay. Experimentally validated effectors are marked by green boxes on the x-axes. Error bar represent the standard deviation.	138
Figure 5.6 a) Root mean square fluctuation of HsaBR, 10712BR and SK678BR (a); GspBBR and SK150BR (b) from MD simulation showing DC, EF, FG loop regions. The residue numbers of HsaBR and GspBBR were used in (a) and (b) respectively. The error bar is SEM from three independent simulations.	138
Figure 5.7 a) Inhibition curve of untagged sTa (Alpha Signal vs Log (untagged sTa)) used to calculate IC ₅₀ ; b) Table showing the percentage control at different concentration of untagged sTa.	139
Figure 5.8 Hierarchical clustering based on MACCS fingerprint of the nine validated hits and the native ligand.....	140

Figure 5.9 Crystal structure pose of native ligand (sTa) (in red) and best docked pose of 9 validated compounds (in blue) in the binding pocket of HsaBR .	141
Figure 5.10 Number of compounds within the top 50 compounds interacting (forming Hbond or Pi (π - π / π -H/ π -cation) interaction) with key residues in the HsaBR binding pocket	142
Figure 5.11 Interaction map of the nine hits in the binding pocket of the Hsa _{BR}	143
Figure 5.13 Crystal structure of open conformation ³⁵⁷ (a & b) and closed conformation ³⁵⁷ (c & d). The red sticks and red region represent residues identified as the “hot spot” residues identified previously ³⁶¹ ; black sphere represents the two zincs (catalytic & structural). a & c are ribbon model of the crystal structures, and b & d are surface model.....	160
Figure 5.14 Schematic diagram explaining the ensemble docking strategy.....	161
Figure 5.15 Closed conformation representative snapshots generated from simulations of crystal structures. Red sticks represent “hot spot” residues, and blue mesh represents pockets in & near the PPI region. a) Crystal structure and b, c, d) Snapshots from aMD and MD simulations.	162
Figure 5.16 The novel pocket discovered in the close conformation (as shown in Figure 3b). a) The pocket is shown in blue mesh with residues in the two loops, which are present in the gate of this pocket, b) probability distribution of the distance of the residues.	163
Figure 5.17 Open conformation representative snapshots generated from simulations. Red sticks represent “hot spot” residues, and blue mesh represents pockets in & near the PPI region. a) Crystal structure and b) Snapshot generated from simulations.	163
Figure 5.18 Top-ranked hits bound a) closed conformation and b) open conformation snapshot	164
Figure 5.19 Top-ranked hits bound a) closed conformation and b) open conformation snapshot	165

Figure 5.20 Effect of compounds on the activities of HDACs. The relative percent activity of HDAC3 (blue bars) and HDAC4 (red bars) in the presence of 100 μ M inhibitor compound compared to the absence of inhibitor. Error bars are the standard deviation of at least two assays. 166

1 INTRODUCTION

1.1 Protein flexibility

Proteins are flexible macromolecules, and this flexibility is often the key determinant for their biological functionality^{1, 2}. These functionalities range from enzymatic catalysis and protein-protein interactions to even signaling and transport¹. The dynamics of a protein allow it to sample different conformations and hence enable it to respond to changes in its environment, for example, pH³ and temperature fluctuations⁴, pressure changes^{5, 6}, and even the presence/absence of binding partners (other proteins or small molecules)^{7, 8}. These responses to the environment could be driven by both local and global motions. For instance, global motion like the domain movements in ABC transporters⁹, cellular locomotion/motor proteins¹⁰, and in several enzymes like glutamate dehydrogenase (GDH)¹¹ are critical for their functions. On the other hand, local motion in proteins comprises the movement of side chains and loops, and have been shown to be required for catalysis^{12, 13}, for ligand binding, and even for binding to other protein partners. Interestingly, these local and global motions of a protein occur over a wide range of timescales from a few picoseconds to hours¹ (**Fig 1.1**). The fastest events like covalent bond vibrations and side-chain rotations occur over a period of one femtosecond to a few picoseconds, whereas, motions like secondary structure formation and loop movement occur over a comparatively longer period of one nanosecond to a few microseconds. Similarly, domain motions (the relative movement of domains in a multi-domain protein) can take somewhere between a few nanoseconds to a few seconds to occur, whereas protein folding can take up to an hour (**Fig 1.1**).

1.2 From Static to Ensemble

For a long time, protein structure has been considered to control the function of a protein, which is commonly referred to as the structure-function relationship of a protein^{14, 15}. However, in the last decade, researchers have provided considerable support suggesting that the dynamics/motion of a protein is also critical for its function, such that a protein exists as an ensemble¹⁶ of inter-converting structures

rather than a single structural conformation and the dynamics/motions enabling this interconversion are also critical for protein function¹⁷⁻²⁰. In an ensemble, each conformation of the protein has an associated free energy (which determines its stability), that can be represented by a point on the resulting free energy landscape (FEL) of the entire protein ensemble²¹ (**Fig 1.2**). This idea of the FEL is a well-established concept in the field of protein folding²²⁻²⁴ and has also been used to define the protein-solvent system²⁵⁻²⁷. The FEL of a protein characterizes two properties of the system – 1) thermodynamic (*i.e.*, the relative probabilities of the conformational states) and 2) kinetic (*i.e.*, the free energy barrier between these conformational states), both of which are crucial for understanding protein dynamics^{1, 28}.

Studying the FEL of a protein has shown that the states with lower free energy have more favorable conformations, but these states can often transition to other conformations depending on the ambient conditions or as a result of their biological function^{19-21, 25, 29}. Hence, the FEL of a protein is largely a rugged space with different free energy wells resulting in different conformations (or transient microstates) separated by barriers that strongly correlate with timescales/kinetics of protein function^{19-21, 25, 29, 30} (**Fig 1.2**). Therefore, understanding how a protein explores its FEL over the course of time and how it responds to changes in the environment is very important for a complete description of its function.

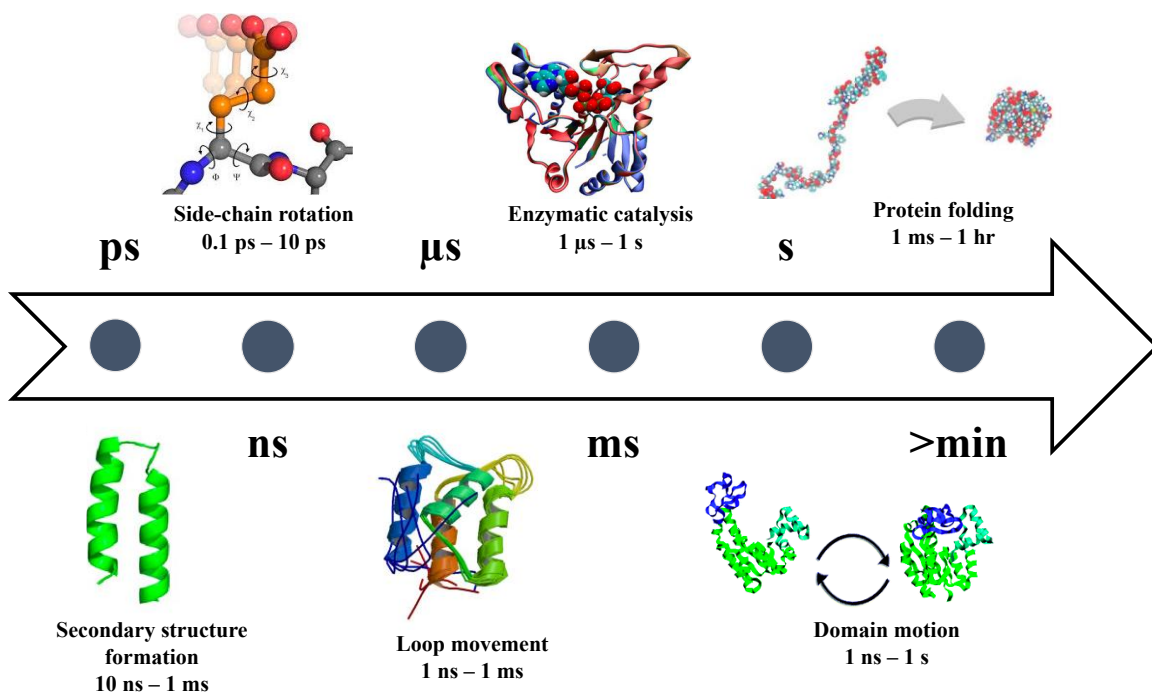


Figure 1.1 Schematic representation of various protein motions at different time scales.

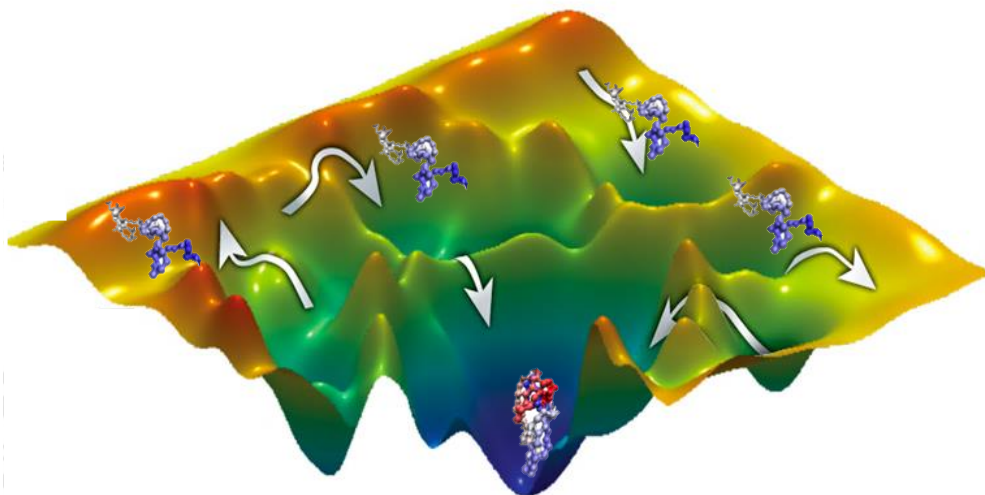


Figure 1.2 Representative image showing the free energy landscape of a protein with multiple minima. Each well corresponds to a minimum energy conformation and the arrows represent the transition states. Adapted from ref³¹.

1.3 Exploring the protein free energy landscape

Since the first protein structure was determined by X-ray crystallography in the late 1950s, the number of protein crystal structures has increased exponentially over the last two decades³². Apart from X-ray crystallography, numerous advancements have also been made in other structure determination techniques such as high-throughput time-resolved X-ray crystallography³³⁻³⁶, cryo-Electron Microscopy (Cryo-EM)³⁷⁻³⁹, Mass spectrometry, Nuclear Magnetic Resonance (NMR)⁴⁰⁻⁴⁴, and Small Angle X-ray Scattering (SAXS)⁴⁵⁻⁴⁷. These developments have enabled researchers to even resolve the structures of complex systems, which has otherwise been extremely challenging⁴⁸. Despite these improvements in experimental techniques, the exploration of the complete free energy landscape of protein conformations at an atomistic detail has still not been achieved⁴⁸. Computer simulations (like MD simulation), on the other hand, can be used to broadly sample the conformational space of a protein system and provide the required spatial and temporal information at an atomic resolution, which has been difficult to access experimentally^{49, 50}. Additionally, this information can further be combined with experimental methods, described above, to attain a complete picture of the protein energy landscape⁵¹⁻⁵³.

1.4 Molecular Dynamics simulation

There are a range of different computer simulation techniques available for the study of protein structures and dynamics. Techniques such as energy minimization and Monte Carlo (MC) simulations do not rely on the equations of motion, and hence fail to provide direct information about the dynamics of a protein system⁵⁴. On the contrary, methods such as Brownian Dynamics (BD)^{55, 56} and Molecular Dynamics (MD) calculate the positions and momenta of all particles in a system as a function of time by integrating the equations of motion in discrete time steps and therefore are applied to generate information about protein dynamics⁵⁴. Among these different simulation techniques, MD is the most commonly used method in the field of biomolecules (like protein, DNA, RNA) because it can closely replicate

experimental observations and can also capture a range of timescales and motions⁵⁰. In MD, atomic motion is simulated by solving Newton's equations of motion, and this can be applied to obtain both thermodynamic and kinetic properties of a biomolecule (i.e., ligand binding to enzymes, protein assembly, transport through a channel and so on)⁵⁰.

Mechanistically, in MD, the atoms are considered as classical particles, where the electrons are not treated explicitly. The separation between electronic and nuclear motions is justified by the Born-Oppenheimer (BO) approximation (BO states that if all the electrons are relaxed in their ground state configuration, then they can be assumed to be fixed on the nuclei)⁵⁴. Therefore, in MD, the atoms are considered as inert spheres, and the bonds/vibrations are represented as springs between two spheres. Additionally, in MD, the inter- and intra- atomic forces between the atoms are calculated using a "potential", which are collectively referred to as molecular mechanics (MM) force fields⁵⁴. A typical MM force field potential function incorporates van der Waals and coulombic electrostatic forces for the non-bonded part of the potential function and harmonic energy function for bonded interactions (i.e., bonds, angles, dihedrals)⁵⁴. Generally, the timestep between MD simulation is in femtoseconds (fs), and the interactions between atoms of the system are computed for ns - μ s timescales. The most commonly used MM forcefields to study biomolecules are AMBER^{57, 58}, CHARMM^{59, 60}, GROMOS⁶¹, OPLS⁶². However, there are some challenges to these force fields especially in the accurate modeling of intrinsically disordered proteins and in incorporating the effects of charge polarization⁶³. Therefore, there is a need to continually improve the MD forcefield parameters to not only make more accurate predictions that match experimental results but also to overcome current limitations.

One such limitation is an incomplete representation or the absence of specific force field parameters for deuterated molecules. Accurate modeling of deuterated molecules is critical for multiple studies, especially because MD is often used in conjunction with experiments such as neutron scattering^{64, 65}, NMR⁶⁶, and FT-IR⁶⁷, which usually require an accurate model of deuteration to complement these

experiments. To this end, we have addressed this issue and developed a novel method for parametrizing deuterated molecules in **Chapter 2**.

Apart from employing an accurate MM force field, simulation protocol has several basic steps, which include: 1) System preparation, 2) Minimization/Relaxation, 3) Equilibration and 4) Production (**Fig 1.3**). The respective parameters for each step mentioned above are specific to each protein, and individual details will be documented in each of the following chapters.

1.5 Enhanced sampling simulation

Although classical MD simulations are regularly employed to study most biological systems of interest, the total simulation time is usually restricted to a few hundred ns to a μ s because of the requirement for extensive computing resources. Since a protein exhibits different types of motions and that some of them occur at timescales beyond a microsecond, classical MD cannot always capture the entire dynamics of a protein. Furthermore, in many cases, the FEL of a protein has multiple potential energy wells (or minima) with high free energy barriers, and it is quite likely for a system to get trapped in one or another local minimum for long periods of simulation time thus preventing accurate exploration of FEL. Hence, to allow the exploration of multiple minima in a rugged FEL, modified simulation methods known as “enhanced” sampling techniques have been developed. These methods use certain biases (for example, modification of the potential energy surface by adding bias potential to decrease the energy barrier) to accelerate the escape from a local free energy minimum enabling greater exploration of conformational space. Within enhanced sampling, multiple methods exist, of these the most popular enhanced sampling methods are: a) Replica Exchange Molecular Dynamics (REMD)- enhances the sampling by allowing systems of similar potential energies to sample conformations at different temperatures^{68, 69} or variations in a Hamiltonian (HREMD), therefore, overcoming the energy barriers on the potential energy surface; b) Metadynamics (MetaD)- enhances the sampling by reconstructing the free-energy surface as a function of few selected degrees of

freedom or collective variables (CVs)^{70, 71}; and c) accelerated Molecular Dynamics (aMD)^{72, 73} - enhances the sampling by defining a simple and robust bias potential using boost factor which flattens the energy landscape. All these methods have their advantages and limitations, and their usage often depends on the biomolecular system and the biochemical phenomena under investigation.

1.6 Protein dynamics in drug discovery

Along with understanding protein functionality, protein dynamics are also critical for drug discovery research. A majority of approved drugs in the market target a protein in the mammalian system, hence necessitating a need for a concrete understanding of protein structures and their functional mechanisms to develop new drugs. This process, also known as structure-based drug design (SBDD), started in the late 1970s soon after the first crystallographic protein structure was resolved. Researchers realized that studying the protein structure can lead to a “rational” drug design, and that one can mimic the ligand/drug binding process in silico by docking potential ligands (generally small molecules) to protein targets in three dimensions⁷⁴. The early docking studies were very successful and have led to the discovery of antivirals for HIV and influenza^{75, 76}. For a long time, however, a single protein structure, often the crystal structure, was used for docking of small molecules, but ever since the dynamic picture of the protein has been accepted⁷⁷, a new research avenue of including protein dynamics in light of drug discovery have emerged to better target the proteins of interest⁷⁸. This process of including an ensemble of protein structures instead of one conformation for docking is commonly referred to as Ensemble docking⁷⁸ (**Fig. 1.4**). This ensemble of structures for a protein can be generated from MD simulation(s). Previous studies have shown that this process of using multiple conformations of a protein is important because the shape of the binding pocket can change during the course of the simulation and this would be missed if a single structure was used. Since receptor shape complementarity is one of the important criteria in docking protocols, this approach can lead to the identification of additional molecules

binding in the same pocket^{74, 79}. Moreover, previous research has also shown that novel and functionally relevant pockets appear during the course of the simulation^{80, 81}, which are not present in the initial static structure, and these pave a path to develop newer strategies to target the protein of interest.

1.7 Early stage drug discovery pipeline

Discovery and development of a drug is a time and resource-intensive process. There is thus a growing effort to apply alternative approaches to the combined space of chemistry and biology to modernize drug discovery, design, development, and optimization process⁸². A promising approach is to harness the ever-increasing computational power to the plaguing problem of novel drug development, as mentioned above⁸³. The stage where computational power has been monumental in speeding up the drug discovery process is the initial hit identification process⁸³. This *in-silico* hit discovery or computer-aided drug discovery (CADD) pipeline consists of multiple different steps as described below and shown in **Fig 1.5**:

- 1) **Target identification and validation**: The first step of a CADD pipeline involves the identification of potential therapeutic targets (mostly proteins) by investigating their functions and association with specific diseases. This step is the most critical step and requires extensive support of experimental studies⁸².
- 2) **Protein structure/s**: Once the protein target has been identified, knowledge of the protein structure is necessary for CADD. The two common approaches to get the initial structures of the protein are: a) crystal structure and b) homology modeling⁸⁴. In homology modeling, a known crystal structure of a homologous protein is used as a template to build the structure using an array of *in silico* tools. In the absence of a homologous crystal structure, *ab-initio* modeling is performed. In the case of “ensemble”

docking, an extensive MD is the performed on the initial structure to get an ensemble of conformations.

- 3) **Binding site prediction:** After getting the ensemble of protein conformations, the binding site or docking site is identified. The binding sites are mostly pockets or cavities, and there are several tools such as FTMAP⁸⁵ and MOE Site Finder⁸⁶ that can be used to predict these sites. In the case of a competitive inhibitor (i.e., an inhibitor which competes with the native ligand/substrate), this step is often not required since the site is already known. For example, an inhibitor for an enzyme would compete for the substrate binding site in its catalytic pocket and hence site prediction using advanced tools would not be required.
- 4) **Docking:** After getting the ensemble of protein conformations and the site of interest, the next step is to dock a database of small molecules (such as ZINC⁸⁷, Enamine⁸⁸, NCI) using docking tools such as Autodock4⁸⁹, AutodockVina⁹⁰, MOE⁸⁶, etc. to identify the molecules that bind the target protein at the site of interest. Different docking protocols use different scoring functions and search strategies to dock the small molecules.

System preparation

To build a starting structure, solvating, applying a force field, etc.

Minimization/Relaxation

To find a local energy minimum of the starting structure

Equilibration

To bring the system to the appropriate state point (e.g. target energy, temperature and pressure)

Production

To retain and analyze the collected data

Figure 1.3 Flowchart showing different steps of a molecular dynamics simulation: 1) System preparation; 2) Minimization/Relaxation; 3) Equilibration and 4) Production

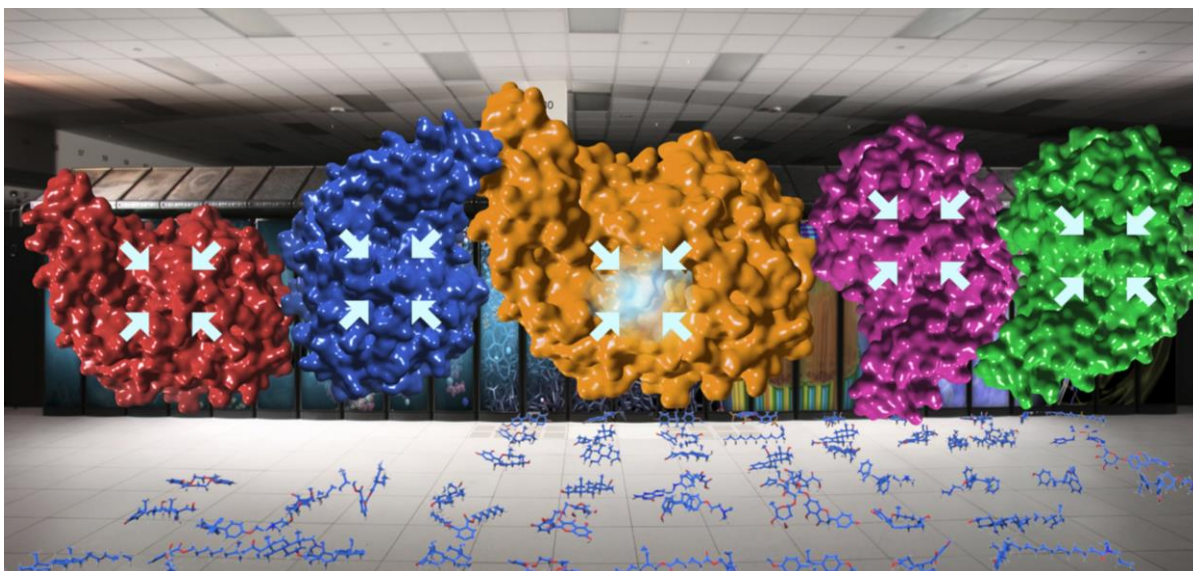


Figure 1.4 Ensemble docking: A cartoon showing multiple protein conformations targeted with small molecules.

- 5) **Scoring:** Once docking has been performed and multiple docked poses have been produced, the next step is to prioritize them and generate a ranked list of the compounds such that compounds with better scores (or better affinity) get a higher rank and are listed first. This is often done based on scores given by the different docking tools by approximating the free energy of binding. Another optional step after obtaining the prioritized list of small molecules is consensus docking. In this step, the docked poses can be re-scored using an additional scoring function to get another prioritized list, and then a consensus of both is used as the final ranked list. This approach is usually a powerful addition and has been shown to improve the hit rate and also reduce the number of false positives.
- 6) **Experimental validation:** The last step for the hit discovery pipeline is experimental validation. Here, based on the ranked list, priority compounds are selected to determine their binding to target protein, which is preferably performed *in-vitro* using techniques such as Surface plasma resonance (SPR), ITC, fluorescence assay, etc⁸².

This CADD approach not only greatly reduces the number of molecules for experimental screening, which require extensive labor and resources but also provides a more rational approach to the traditional brute-force method of experimental testing⁸³. After the initial hits are identified, the next steps in the drug discovery pipeline are hit expansion or lead optimization where the hit molecules are further modified to make them more selective and to satisfy ADME (Absorption, Distribution, Metabolism, and Excretion) criteria⁹¹. This is then followed by a three-phase clinical trial and approval. Traditionally, the whole drug discovery process can take between 10-12 years, from target identification to FDA approval⁸².

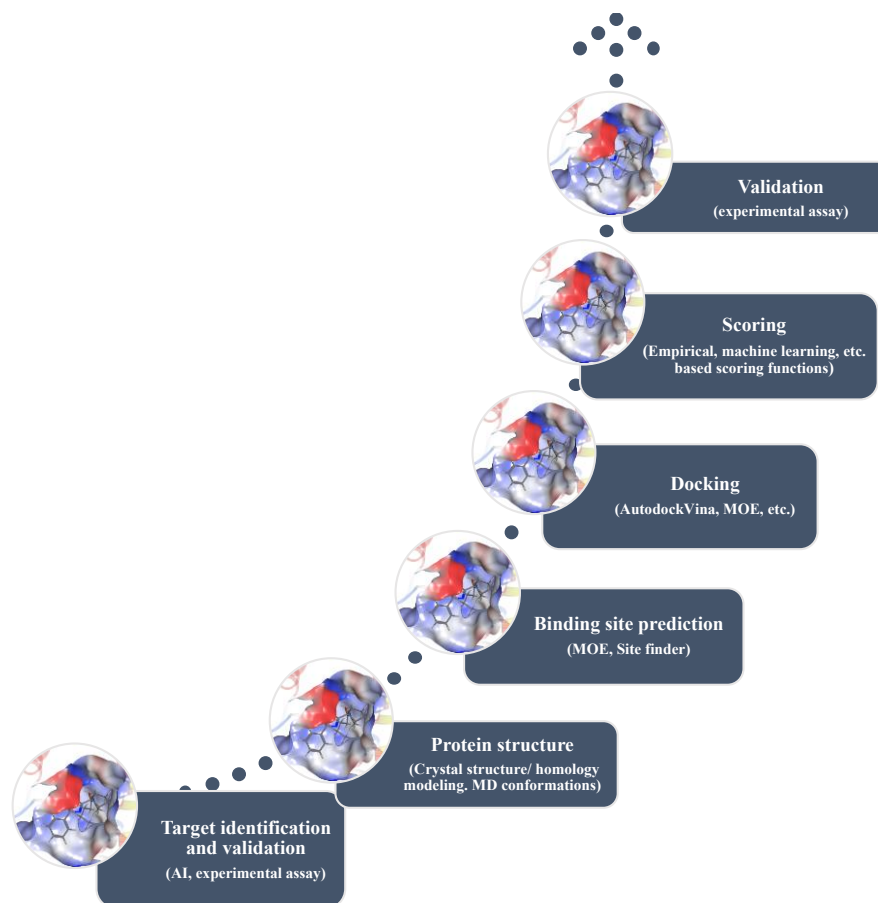


Figure 1.5 Drug discovery pipeline

1.8 Proposed study

Protein dynamics are critical to understand the function of a protein. Computer simulations have emerged as very powerful tools to uncover the mechanism of function and stability of a protein. In this dissertation, I have used an array of computational simulation methods to study different types of protein dynamics in multiple proteins of interest. This work was further expanded to apply the results in early-stage drug discovery. Specifically, in **Chapter 2**, I studied the pitfalls of existing force fields to accurately represent deuterated molecules and successfully developed a strategy to overcome this limitation. In **Chapter 3**, we studied the protein dynamics of adhesin proteins and characterized the functional relevance of global motions (like domain motion) and local motions (like loop motion) in the binding pocket. In **Chapter 4**, I studied the effect of temperature and pressure on protein dynamics and enzymatic function in inorganic pyrophosphatases from mesophilic and thermophilic organisms. Lastly, in **Chapter 5**, I applied the knowledge of protein dynamics to perform ensemble docking on two clinically relevant protein targets: adhesin protein Hsa and histone deacetylase 4 (HDAC4). Using a combination of computer simulations, ensemble docking, and experimental validations, I have successfully identified hits for both these target proteins.

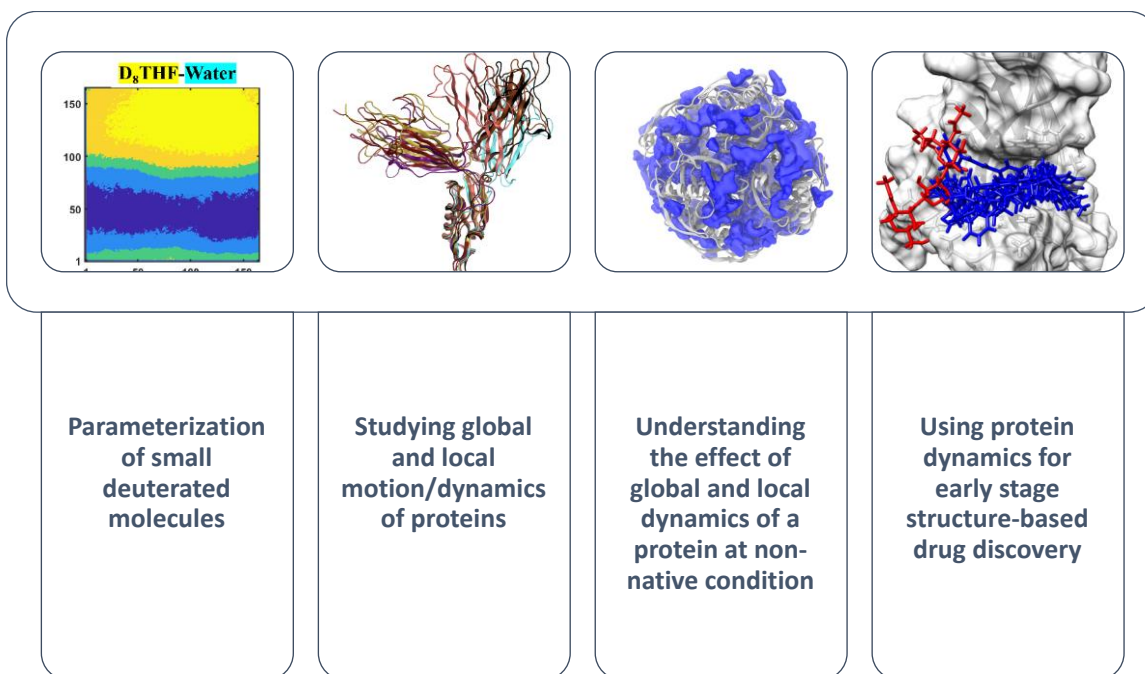


Figure 1.6 Dissertation division into chapters

2 PARAMETERIZATION OF SMALL DEUTERATED MOLECULES

A version of this chapter was originally published by me:

Agarwal, Rupesh, Micholas Dean Smith, and Jeremy C. Smith. "Capturing Deuteration Effects in a Molecular Mechanics Force Field: Deuterated THF and the THF–Water Miscibility Gap." *Journal of Chemical Theory and Computation* 16.4 (2020): 2529-2540.

In this work, I executed the research, performed analyses, and wrote the paper. Dr. Micholas Dean Smith conceived the idea and made substantive contributions to the manuscript's content and Dr. Jeremy C. Smith helped in the discussion of results and edited the manuscript.

2.1 Abstract

Deuteration is a common chemical modification used in conjunction with experiments such as neutron scattering, NMR, and FT-IR for the study of molecular systems. Under the Born-Oppenheimer (BO) approximation, while the underlying potential energy surface remains unchanged by isotopic substitutions, isotopic substitution still alters intra-molecular vibrations, which in turn may alter inter-molecular interactions. Molecular mechanics (MM) force fields used in classical molecular dynamics (MD) simulations are assumed to represent local approximations of the BO potential energy surfaces, and hence MD simulations using simple isotopic mass substitutions should capture BO-compatible isotope effects. However, standard MM force-field parameterizations, do not directly fit to the local harmonic quantum mechanical (QM) Hessian that describes the BO surface, but rather to QM normal-modes and/or mass-dependent internal-coordinate derived distortion energies. Here, using THF-Water mixtures as our model system, we show that not only does a simple mass-substitution approach fail to capture an experimentally characterized deuteration effect (the loss of the closed-loop miscibility gap associated with the complete deuteration of THF), but that it is necessary to generate new MM force-field parameters that correctly describe isotopic dependent vibrations to capture the experimental deuteration effect. We show that the origin of this failure is a result of using mass-dependent

features to fit the THF MM force-field, which unintentionally biases the bonded terms of the force-field to represent only the isotopologue used during the original force-field parameterization. In addition, we make use of our isotopologue corrected force-field for D₈THF to examine the molecular origins of the isotope-dependent loss of the THF-water miscibility gap.

2.2 Introduction

The use of hydrogen/deuterium (H/D) isotope substitution to probe the structure and dynamics of molecular systems is well established, with applications including neutron scattering⁹²⁻⁹⁴, NMR^{95, 96}, FT-IR⁹⁷⁻¹⁰¹, and others¹⁰²⁻¹⁰⁴. Deuteration is well known to change hydrogen bonding^{105, 106}, vibrational dynamics^{107, 108}, reaction kinetics^{109, 110}, and phase behavior¹¹¹⁻¹¹³.

For decades molecular dynamics (MD) simulation has been used in combination with experiment to provide atomic-resolution descriptions of isotope effects in small molecular systems¹¹⁴⁻¹¹⁶. Explicit studies of deuteration have, traditionally, required the use of Path-Integral Molecular Dynamics (PIMD)¹¹⁷⁻¹¹⁹ or Quantum Trajectory/Dynamics (QMD) approaches¹²⁰, which for large molecular systems remains computationally expensive. Although both PIMD and QMD, in principle, provide a means to completely capture the impact of deuteration on the dynamics, it is important to consider that although the quantum nature of the nuclei is changed upon deuteration, cheaper calculations that make use of the Born-Oppenheimer (BO) approximation also exhibit isotope dependencies^{121, 122}. Indeed, while the potential energy surface (PES) of a molecular system is invariant of mass, within the BO approximation, the dynamics of system is altered by isotopic substitution (via mass-dependent changes to the kinetic energy operator)^{121, 122}. Given that dynamics on the BO PES are altered, a computationally cheap means of studying the impacts of isotopic substitution phenomena, where the quantum nature of the nuclei may be hypothesized to be unimportant, would be to perform traditional molecular dynamics on the BO surface with the appropriate mass weighting.

Classical MD simulations using typical classical class I and/or biomolecular molecular mechanics (MM) force-fields are parameterized to represent a local harmonic approximation of Quantum Mechanics (QM) derived BO PES^{54, 123} by fitting to QM vibrational analyses and (when available) experimental geometries and thermodynamic quantities. In the absence of experimental geometries/thermodynamics, QM geometry optimization and vibration calculations are used to obtain starting guesses for the equilibrium geometry of the system of interest and the force-constants that describe the intra-molecular dynamics of the system. During the QM vibrational analysis, the decoupling between the translational and rotational degrees of freedom of the molecular system is required. This decoupling is achieved via the use of a mass-dependent coordinate system that is constructed to satisfy the Eckart-Sayvetz conditions¹²¹. Essentially this isotopic dependence of the QM vibrational analysis implies that normal modes and distortion energies (computed along the internal coordinates derived to satisfy the Eckart-Sayvetz conditions), which are used to fit the MM bonded terms, may carry an implicit isotopic dependence. This indicates that MM force-fields themselves may not necessarily reflect a classical approximation of the BO PES, but rather isotopic specific dynamics on the BO PES. We hypothesize that 1) this isotopic dependence does exist and as a result a simple (naïve) substitution of different isotopic masses in the MM force-fields used in MD simulations would fail to capture accurate isotopic intra-molecular dynamics; 2) if the MM force-field is re-parameterized using properly weighted QM vibrational analyses derived features using the isotopic masses of interest it would be possible to accurately capture isotope dependent intra-molecular vibrations and in turn accurately capture isotope-dependent inter-molecular interactions.

To test these two hypotheses, we study the macroscopic behavior of the miscibility gap between tetrahydrofuran (THF) and water. THF is a ubiquitous, heterocyclic organic solvent with applications ranging from organometallic chemistry¹²⁴ and polymer synthesis¹²⁵⁻¹²⁸ to NMR spectroscopy¹²⁹ and lignocellulosic biomass pretreatment (when used as a co-solvent)¹³⁰⁻¹³⁶. Along with these uses, THF, when

mixed with water, serves as a model system for the investigation of a mixture exhibiting a closed-loop miscibility gap (a miscibility gap with two critical temperatures)^{112, 137}. This miscibility gap occurs in mixtures of fully hydrogenated THF and water with mass ratios ranging from ~0.3 to 2.8 and at temperatures above ~71 °C and below ~137 °C. At these conditions, THF and water have been shown experimentally^{112, 113, 138} and computationally (at timescales on the order of tens of nanoseconds)^{139, 140} to spontaneously de-mix, whereas outside of these ranges, they remain mixed. Most interesting for the present study of isotope effects; however, is that upon complete H/D substitution of THF (i.e., D₈THF), this miscibility gap disappears^{112, 113}.

Prior experimental work on (hydrogenated) THF-water solutions has indicated that small-scale phase separation occurs even at temperatures outside of the miscibility gap, resulting in the presence of both micro-voids and THF-rich domains¹⁴¹⁻¹⁴³. Furthermore, computational studies have reported that within the miscibility gap temperature regime prior to phase separation, the diffusivities of the two species diverge¹⁴⁰. However, despite these micro-domains and variations in diffusion, THF and water mixtures contain a significant population of mutual hydrogen bonds (between the oxygen atoms of THF and the hydrogen atoms of water) while largely preserving water-water hydrogen bonding¹⁴¹. Given these competing tendencies, it has been hypothesized that THF-water hydrogen-bonds are necessary for the two solvents to remain mixed¹⁴¹⁻¹⁴³. And, it was recently suggested that the onset of the miscibility gap is due to mismatches in the rate of change (as a function of temperature) of THF-water and water-water hydrogen-bonding and THF-THF, THF-water, and water-water interactions¹⁴⁰. While these studies have advanced our understanding of the molecular origins of THF-water miscibility gap, the underpinnings of the peculiar influence of isotopic substitution on the gap remain unclear. Additionally, two points that justify the choice of the THF-water system as the model system to explore potential implicit isotopic effects are: 1) prior simulations have demonstrated that it is possible to simulate THF-water phase separation within sub-microsecond timescales^{139, 140}, suggesting the

macroscopic phase behavior can be readily simulated with current computational resources, and 2) given the isotopic substitution occurring in deuterated THF is at the sites of non-polar hydrogens (i.e., hydrogen atoms that do not form inter-molecular hydrogen-bonds), the isotope-dependent phase separation phenomena is not likely driven by non-local nuclear quantum effects, suggesting that an accurate classical approach should be able to capture the isotope-dependent behavior of the system.

The work reported here thus addresses two questions: 1) Does the standard MM force-field contain implicit isotopic biases, and, if so, can a re-parameterization approach provide a means to overcome this bias for classical MD simulations? and 2) what is the origin of the loss of the THF-water miscibility gap upon deuteration?

2.3 Methods

Several MD simulation campaigns of different THF-water systems were performed to examine substitution and re-parameterization approaches to capturing deuteration within classical simulations. All of the simulation calculations were carried out using the GROMACS (version 2016.3) simulation suite^{144, 145}, and analysis was performed using a combination of in-house VMD Tcl scripts and GROMACS analysis tools. Center-of-mass radial distribution functions (RDFs), site-specific coordination numbers, hydrogen bonds (HB), and diffusion constants were calculated. Additionally, for the re-parameterization approach, we have the ancillary aim of providing force-field parameters, which may be of use in other studies, and as such, we restricted ourselves to the development of parameters for completely deuterated THF (D_8 THF) that are compatible with the CHARMM potential¹⁴⁶.

Details of the various simulations, the deuteration parameterization method and analyses of trajectories are provided in the sections below.

Model systems

All simulations were performed using the CHARMM potential¹⁴⁶ on four systems for which the parameters that describe THF vary:

- i) with THF represented by the CHARMM32 ether force-field¹⁴⁷, denoted as THF.
- ii) with the mass of each hydrogen doubled (the traditional mass substitution method), but no changes in the force-field parameters i.e., “heavy” THF, denoted as HTHF,
- iii) with both the masses of the hydrogen atoms doubled and the bonded parameters re-parametrized (see parameterization details below) denoted as D₈THF,
- iv) with bonded parameters of hydrogenated THF calculated using the same methodology as (iii), denoted THFK.

D₈THF Parameterization

As noted above, we consider two possible approaches for capturing isotope effects in classical MD:

- i) ‘Substitution’ approach: increasing the mass of hydrogen to 2.014u while preserving all force-field parameters (here we use the CHARMM32 ether force-field¹⁴⁷ parameters) of the hydrogenated molecule, and
- ii) Re-parameterization approach: by which the new force-field parameters are generated using the mass of deuterium in place of hydrogen during the fitting of bonded (bonds, angles, and dihedrals) force-field parameters. It should be noted that a prior attempt to capture isotope effects in a classical manner (beyond the mass substitution approach) has been reported in the literature: a parameter-scaling modification to the MM2 force-field where van der Waals interactions and bond-lengths were altered to reflect H/D size differences¹⁴⁸. Although this attempt was reasonably successful failures were noted for heterocyclic compounds.

At this point it is helpful to recall that the bonded terms of molecular mechanics force-fields are frequently obtained via fitting to the mass-dependent quantities)¹⁴⁹

(see *Introduction*). By substituting the mass of deuterium for hydrogen during the QM frequency calculations and then fitting new force-field parameters using distortion energies derived from mass-weighted internal coordinates, as is done in the FFTK¹⁵⁰ parameterization protocol, we expect to obtain a parameterization that capture the energetics of the deuterated isotopologues of interest. In principle, the re-parameterization method aims to derive classical force-field parameters that better reflect how given an unchanged PES (as the BO surface is invariant to isotopic substitutions), the frequencies and vibrational modes for the nuclei, are altered as a result of the changed masses¹⁵¹. To be clear, MM force-fields generated during the parameterization process that use mass-dependent target data are designed to capture the intra-molecular motions driven by the BO surface not the BO surface itself and as such we would expect that the bonded force-field parameters derived from mass-dependent dynamics (normal modes and mass-weighted internal coordinate distortions) are themselves implicitly mass-dependent.

For the re-parameterization approach, CHARMM-compatible¹⁵⁰ force-field parameters for deuterated THF (D₈THF) were obtained using Gaussian09¹⁵² and an in-house modified Force-field Toolkit (FFTK)¹⁵⁰ which updates the standard Gaussian input files for D₈THF with the keyword iso=2 after H atoms. In the bond optimization step of FFTK, Gaussian09 includes the isotope effect by calculating the Hessian matrix by converting force constants to mass-weighted Cartesian coordinates (MWC)¹⁴⁹ and then calculating mass-dependent internal coordinates (ICs). FFTK uses these mass-dependent ICs to generate target distortion energies for use in parameter fitting. Parameterizing all bonds, angles, and dihedrals was required because local isotopic effects not only modify the H/D-C bonds but can also propagate throughout the entire molecule, resulting in changes to all angles and dihedrals that involve any carbons bound to an isotopic substitution site^{123, 153}. All the partial charges and Lennard Jones parameters were kept the same as the CHARMM32 ether force-field¹⁴⁷. To check whether the use of FFTK alone does not result in parameters for THF that would result in the loss of the THF-water

miscibility gap, we also re-parametrized the fully hydrogenated THF (denoted as THFK in the remainder of the text). The bonded parameters from FFTK for hydrogenated THF (THFK) and deuterated THF (D_8 THF) generated using FFTK are provided in **Tables 2.1-2.4** and **SI Tables 2.5-2.8**, respectively. The parameters generated from FFTK are in CHARMM/NAMD format and were converted to GROMACS format using the Topotools¹⁵⁴ (Topogromacs¹⁵⁵) plugin in VMD¹⁵⁶.

Simulation details

Two stages of simulations were performed in this study. The first of these was aimed at providing an initial comparison (and validation check) of the mass substitution method and the re-parameterization approach in modeling the behavior of bulk D_8 THF and the second stage examined the impact of the two approaches in the representation of water- D_8 THF binary systems. For the first stage, a single bulk environment was used throughout while in the second stage two separate sets of simulations were performed: 1) large-scale simulations of (D_8)THF-water mixing behavior within the THF-water miscibility gap temperature regime and 2) small-scale simulations of well-mixed (D_8)THF (at multiple temperatures) to obtain well-sampled detailed descriptions of D_8 THF-water interactions.

For the first stage of simulations (i.e., the bulk organic solvent case), small boxes of 1,000 molecules of the D_8 THF, HTHF, THF, and THFK systems were prepared using the gromacs gmx solvate command. Following box construction an energy minimization was performed using the steepest descent algorithm for a maximum of 100,000 steps and with a stopping tolerance of 1 kJ/mol/nm. Post-minimization, the systems were subjected to two rounds of NPT MD equilibration for 10 ns, each with timestep of 0.2 fs. In the first of the NPT equilibration simulations, the pressure and temperature were controlled using the Berendsen¹⁵⁷ thermostat and barostat and the second NPT equilibration simulation the V-Rescale thermostat¹⁵⁸ and the Parrinello-Rahman¹⁵⁹ barostat were used. Production simulations (of length ~0.8

ns with timestep of 0.05 fs) of the systems were performed using the Berendsen algorithm¹⁵⁷ thermostat and barostat. These simulations were used to calculate IR-spectra, the density and diffusion constants (see analysis details).

For the large-scale simulations, large cubic boxes (side length ~16.5 nm) were constructed of THF: water, HTHF:water, THFK:water and D₈THF: water with a mass ratio of 0.4 (total system size: ~473,540 atoms). In the initial configuration, the water and isotopologes were completely phase-separated, as shown in **Fig. 2.1**. Energy minimization was performed using the steepest descent algorithm for a maximum of 10,000 steps and with a stopping tolerance of 100 kJ/mol/nm. Following the minimization, a short (1 ns) equilibration simulation in the NPT ensemble was performed, followed by an NPT production simulation for 400 ns with a timestep of 1 fs at 90°C.

For the detailed small-scale D₈THF-water system, MD simulations at five different temperatures (30°C, 75°C, 105°C, 130°C and 165°C) with 1250 TIP3P (water) molecules and 125 THF molecules (mass ratio of THF to TIP3P of ~0.4 m/m or mole fraction of 0.09 THF/water) were performed. The number of atoms was chosen to coincide with prior computational work¹⁴⁰. These simulations were performed for all four systems; THF, HTHF, D₈THF, and THFK. Each temperature was simulated with three independent runs, with an integration step of 1fs, for 150 ns. Frames were saved for analysis every 10ps.

All THF-Water systems were prepared using a standard three-step process: energy minimization (to remove potential clashes during model construction), followed by short pressure relaxation simulations, and finally canonical (NVT) ensemble production simulations. Energy minimization was performed using the steepest-descent algorithm, as implemented in the GROMACS 2016 software package^{144, 145} with a convergence criterion of 100 kJ/ (mol·nm). Relaxation simulations in the NPT ensemble were performed for 2 ns at each temperature with the pressure fixed to 1 bar with the Berendsen barostat¹⁵⁷. Following the NPT relaxation, the volume was taken to be fixed, and the V-Rescale thermostat¹⁵⁸ was used to control the temperature, and the simulation protocol (as noted above) was

used to generate production runs for analyses. In both the pressure relaxation and production simulations, bond constraints were not used.

Analysis details

For the initial comparison of two approaches to modeling isotope substitution (i.e., the naïve substitution and re-parameterization approaches), the average density, self-diffusion coefficient, and IR spectra of THF and deuterated THF were calculated. IR spectra were calculated from the autocorrelation function of the total dipole moment of the simulation box, as shown in Prothro et al¹⁶⁰.

The molecular behavior of the D₈THF, HTHF, THFK, and THF–water systems was characterized using a combination of calculated radial distribution functions (RDFs), hydrogen-bond counts, and hydrogen-bond lifetimes. Additionally, the RDFs were used to calculate coordination numbers and second virial coefficients (pair interaction strengths)¹⁶¹. Each of these metrics was calculated for all temperatures and all independent trajectories, with the average and standard error of the mean reported. A comparison between THFK with THF, HTHF, and D₈THF is provided as supplementary material accompanying this manuscript.

Coordination numbers were obtained by computing the site–site RDFs between the THF oxygen and hydrogen atoms and the water atoms and then integrating the resulting profiles up to the first minima. RDFs were obtained using the GROMACS gmx rdf utility. **Fig. 2.10** illustrates our pair selections, and **table 2.9** provides integration cutoffs. The cutoffs correspond to the first minima in the site–site RDF.

Hydrogen bond counts were calculated with the built-in gmx hbond utility in GROMACS. Hydrogen bonds were taken to be within donor-acceptor distances of 0.32 nm with a cutoff angle of 20°. As a proxy for the strength¹⁶²⁻¹⁶⁴, lifetimes were computed from autocorrelation functions of each hydrogen bond, as described by Van der Spoel *et al*¹⁶².

Diffusion constants were obtained by first calculating the velocity autocorrelation function (using gmx velacc) and then integrating the function to obtain a diffusion

constant (with the integration performed using the gmx analysis tool). The functional form of this (Green-Kubo) relationship between the velocity autocorrelation function and the diffusion constant is:

$$D = \frac{1}{3} \int_0^{\infty} \langle \vec{v}(t') | \vec{v}(t'') \rangle dt$$

where $\langle \vec{v}(t') | \vec{v}(t'') \rangle$ is the velocity autocorrelation function, and D is the diffusion constant. Standard errors were calculated by splitting the trajectories into three equal time windows and calculating the diffusion constants in each window.

The pair interaction strengths (D₈THF-D₈THF, HTHF-HTHF, THFK-THFK, and THF- THF; D₈THF-water, HTHF-water, THFK-water, and THF-water; water-water) were quantified using the second virial coefficient (B₂). The virials were calculated using the standard McMillan-Mayer formalism^{161, 165} :

$$B_2 = -2\pi \int_0^{\infty} [g(r) - 1] r^2 dr$$

where r is the distance and g(r) is the radial pair distribution function.

Density profiles for all the four large systems (D₈THF, THF, HTHF, THFK) were calculated by averaging over all the three dimensions using the GROMACS gmx density utility. The GROMACS gmx energy utility was used to calculate the mean reported throughout the text, and the error estimate reported is for the standard error of the mean (SEM).

2.4 Results

Parameter validation studies

A) Density

From the MD simulation of THF and THFK, the density was calculated to be 879.82 (+/- 0.5) kg/m³ and 881.84 (+/- 1.1) kg/m³ respectively, values somewhat lower

than the experimental value¹⁶⁶ of 889 kg/m³. The experimentally determined value^{167, 168} of D₈THF is much higher, at 985 kg/m³. Most of this increase arises from the mass change, with the density of the simulated HTHF calculated as 975.47 (+/- 1.5) kg/m³. However, the average density of D₈THF was found to be 980.33 (+/- 1.2) kg/m³, and thus the change in parameters for D₈THF leads to a model more closely matching experiment.

B) Diffusion

The self-diffusion constant (D) of D₈THF, HTHF, THF, and THFK was calculated for all four systems. Although the absolute values of D are approximately 30% lower than the experiment (diffusion constants are notoriously difficult to reproduce quantitatively¹⁶⁹), the effects of deuteration are correctly reproduced. The increase in mass in HTHF (for which $D=2.18 (+/- 0.0001) \times 10^{-9} \text{ m}^2/\text{s}$) slightly decreases D relative to that of THF ($D_H=2.20 (+/- 0.0001) \times 10^{-9} \text{ m}^2/\text{s}$). The D value of THFK ($2.10 (+/- 0.00005) \times 10^{-9} \text{ m}^2/\text{s}$) is slightly lower than THF. Interestingly D₈THF shows a further decrease in the value of D ($D_D=1.93 (+/- 0.0001) \times 10^{-9} \text{ m}^2/\text{s}$) as compared to HTHF. This difference in the values of the diffusion constants (D) between THF and D₈THF calculated from MD simulations ($0.27 \text{ m}^2/\text{s}$) is consistent with the difference ($0.23 \text{ m}^2/\text{s}$) in the experimentally determined values (THF $2.84 (+/-0.03) \times 10^{-9} \text{ m}^2/\text{s}$ and D₈THF $2.61 (+/-0.03) \times 10^{-9} \text{ m}^2/\text{s}$)¹⁷⁰. Moreover, the ratio ($D_r^{-1}=D_H/D_D$) for D₈THF calculated from MD simulations ($1.13 +/- 0.00007$) is also comparable to the experimental value ($1.09 +/- 0.02$)¹⁷⁰ which corresponds to a difference of ~3.5% from the experimental values, whereas for HTHF the ratio is $1.009 +/- 0.00007$ which corresponds to a difference of ~11% from experimental values.

C) Infrared Spectra

IR spectra calculated from the MD simulations of D₈THF, HTHF, THF are compared with experimental peak positions¹⁷¹ in **Fig. 2.2**, and for THF and THFK are compared separately in Fig S1. For THF, experimental peaks assigned to C-H

stretch (2977 cm^{-1} , 2861 cm^{-1}) are blue-shifted for D_8THF (2226 cm^{-1} , 2133 cm^{-1}) and align well with theoretical peaks (THF: $\sim 2913\text{ cm}^{-1}$, $\sim 2870\text{ cm}^{-1}$; D_8THF : $\sim 2225\text{ cm}^{-1}$, $\sim 2130\text{ cm}^{-1}$; HTHF: 2159 cm^{-1} , 2131 cm^{-1} ; THFK: 2987 cm^{-1} , 2910 cm^{-1}). The experimental ring stretch assignment (THF: 1177 cm^{-1} ; D_8THF : 1162 cm^{-1}) aligns well with the corresponding theoretical peaks (THF: $\sim 1120\text{ cm}^{-1}$; D_8THF : $\sim 1163\text{ cm}^{-1}$; HTHF: $\sim 1179\text{ cm}^{-1}$; THFK: $\sim 1189\text{ cm}^{-1}$). All these peak positions¹⁷¹ show that D_8THF is closer to experimental peaks than HTHF in this region.

The fingerprint region of IR spectra (~ 500 to 1000 cm^{-1}) from MD simulations is known to be noisy and to exhibit non-uniform shifting^{172, 173}. The fingerprint region calculated from MD simulations of existing THF force-field does not properly align with the experimental peaks, and thus isotopic shifts in this region should not be interpreted in detail.

Effect of THF deuteration on the THF-Water miscibility gap

In order to capture macroscopic properties and examine the temperature-dependent closed-loop miscibility gap, 2D-density profiles for the large systems (~ 473540 atoms) were calculated and averaged over the axes (X, Y, and Z), as shown in **Fig. 2.3-2.6**. The initial configuration of all these systems was a phase-separated state (**Fig 2.1**). The 0-50 ns density profiles clearly show a separation of water molecules for all the four systems. The profiles for the THF-water, HTHF-water, and THFK-water systems show only minor changes throughout the simulations, as is seen in **Fig. 2.15, 2.16 and 2.18**, and remain phase-separated over the entire simulation (as shown by the continued phase-separation in the 350-400 ns density profile window). In the D_8THF -water system, however, the density profile at the end of the simulation (350-400 ns) differs when compared to the initial configuration, and there is no-longer a phase separation. **Fig. 2.17** demonstrates that this mixing starts relatively early, after the first 50 ns. The mixing in the D_8THF -water system and phase separation in THF, HTHF, and THFK systems shown by density profile plots further indicate that the new parameterization agrees with experimental results^{112, 113}.

THF-water interaction studies

A) Pair Interaction Strengths (Second Virials)

The second virial coefficient/pair interaction strengths (B_2), derived from center-of-mass to center-of-mass RDFs, were computed for each temperature and are shown in **Figs 2.7** and **2.19** (in this and the plots that follow THFK is plotted separately for clarity). In most of the cases, no significant differences (within the error bars) are observed between THF and HTHF. However, the plots also show that B_2 is more positive (repulsive) for HTHF-water, THF-water, and THFK-water interactions than for D₈THF-water interactions. Similarly, an increase in D₈THF-D₈THF interaction virials is observed at all temperatures relative to the THF, HTHF, and THFK systems. Interestingly, for all three systems, the D₈THF, HTHF, THF, and THFK -water interaction virials follow a near parabolic trend with temperature: D₈THF-D₈THF, HTHF-HTHF, THF- THF, and THFK-THFK and water-water virials are concave and D₈THF, HTHF, THF, and THFK -water virials convex.

B) Hydrogen Bonds

To further investigate the intermolecular effects reported in **Fig 2.7**, the number of hydrogen bonds was calculated and is presented in **Figs 2.8** and **2.20**. For all three systems, the number of hydrogen-bonds was observed to decrease with increasing temperature.

Also, again for all three systems, no significant differences were found for water-water hydrogen-bonding. However, at all five temperatures the normalized number of HBs in the D₈THF-water system is larger than in the HTHF-water, THF-water, and the THFK-water systems.

As a proxy for the strengths of the HBs¹⁶²⁻¹⁶⁴, their lifetimes were calculated and are presented in **Fig 2.9** and **2.21**. The D₈THF, HTHF, THF, and THFK -water and water-water HBs are longer-lived at lower temperatures than at higher temperatures. Interestingly, there are no significant differences in the HB lifetimes of THF, HTHF, THFK, and D₈THF-water. The lack of changes in the HB lifetimes

indicates that, although there is an observed increase in the number of HBs between D₈THF-water, the overall stability of these hydrogen bonds is not substantially altered.

C) Coordination Numbers

To further quantify the effect of deuteration on the microscopic structure of the (D₈)THF-water mixtures coordination numbers were calculated. RDFs between atoms of D₈THF, HTHF, THF, THFK, and water were calculated (with the cutoffs mentioned in **table 2.9** and atom names in **Fig. 2.10**). The coordination numbers as a function of temperature indicate that the D₈THF atoms H1H4/H2H3 and O1 coordinate more with water atoms, OW, and HW, respectively at all temperatures than do THF, HTHF and THFK, as shown in **Figs. 2.13** and **2.22**. Correspondingly, the atoms of THF, HTHF, and THFK coordinate slightly more with themselves compared to D₈THF, as shown in **Fig. 2.12** and **2.23**. At the same time, the self-coordination number of the water atoms (HW-HW, OW-HW, OW-OW) is slightly lower for D₈THF than for the HTHF, THF, and THFK systems, as seen in **Fig. 2.11** and **2.24**. No changes in coordination are observed between HTHF and THF with itself or water at any temperature.

2.5 Discussion

The aims of this work were: 1) examine if a classical MM force-field derived from distortion energies contained an implicit isotopic bias (which would limit its ability to capture isotope substitution effects upon simple mass-substitution) and, if so, whether a re-parameterization method could take advantage of this bias and be used to generate new MM parameters which could be used to capture the impact of isotopic substitution, and 2) to examine, in molecular detail, what drives the isotope-dependent loss of the THF-water miscibility gap upon the complete deuteration of THF. To achieve these aims, we performed classical MD simulation and force-field parameterization calculations for completely deuterated THF.

With regards to our first aim, we clearly demonstrated that a simple mass substitution method (the HTHF systems above) fails to quantitatively reproduce shifts in density, self-diffusion, IR-spectra. Further, the use of this naïve mass substitution approach also fails to capture the mixing of D_8 THF with water, in contrast to what is found experimentally. From the failures of the HTHF simulations, two conclusions are possible: 1) the MM force-field may have a bias that limits the applicability of mass-substitution for the study of isotopologues or 2) quantum effects neglected in the classical representation are non-negligible for the study of perdeuterated THF. By considering the success of the re-parameterization approach (i.e., the D_8 THF simulations) capturing bulk perdeuterated THF properties and the loss of the THF-water miscibility gap upon isotopic substitution, it is clear that the second possible conclusion can be ruled out, as the re-parameterization approach would only work in the case where the force-field parameterization method was to capture the “proper” implicit isotopic bias (mass corrected Eckart-Sayvetz conditions) and that the dynamics of interest contained negligible non-local nuclear quantum effects.

With our results showing that classical MM force-fields (at least those fit to features that contain mass-dependencies) contain implicit isotopic biases and that it is possible to use classical MD (with a properly re-parameterized force-field) to capture the isotope dependent loss of the THF-water miscibility gap, we return to the second aim of this work: understanding the molecular origins of macroscopic isotope-dependent behavior – in this case, the loss of the THF-water miscibility gap upon per-deuteration of THF. Our results indicate that consistent with demixing, at all temperatures, the pair interaction strength of D_8 THF- D_8 THF is more repulsive than THF-THF (THFK-THFK, and the HTHF-HTHF) and D_8 THF-water interactions are more attractive than in the THF, THFK, and HTHF systems. Additionally, comparing the number of hydrogen bonds between the three systems (HTHF-water, D_8 THF-water, THFK-water, and THF-water) shows that there is no change between HTHF and THF containing systems across all the five temperatures; while a comparison between the THF- and D_8 THF-containing

systems show an increase in the number of water-organic solvent (i.e., water- D_8 THF) hydrogen bonds upon isotope substitution. However, the lifetime of hydrogen bonds remains unaffected between all three systems. When we combine these results with a closer look at the site-site coordination behavior of THF (both deuterated and protonated), we find that the atoms O1, H1H4, and H2H3 of D_8 THF have significantly higher coordination numbers with both water hydrogen (HW1/HW2) and oxygen (OW) atoms compared to the corresponding atoms in HTHF and THF. The increase in coordination indicates that the change in the force-field parameters D_8 THF increases the solvent accessibility of the oxygen for D_8 THF (compared to fully hydrogenated THF). This increased accessibility then serves to increase the number of favorable co-solvent (D_8 THF-water) hydrogen bonds and provides additional favorable co-solvent-water interactions. We observe that in the presence of water molecules, there is a change in all the angles that make the overall structure squeeze and make the oxygen more exposed (**Fig. 2.25**). Interestingly, the QM normal modes also show that D_8 THF molecule has normal modes which result in more exposed O atom and THF has a normal mode that opposes the accessibility of oxygen (**Fig. 2.26**). These physical changes in the structure of the D_8 THF are responsible for the disappearance of the miscibility gap. Hence, it is here that we find the origin of the isotope effect, arising from minor changes to the structure of THF upon deuteration that increases the solvent exposure of the O1 atom, permitting D_8 THF to accept more hydrogen bonds.

2.6 Conclusion

In this work, we demonstrated that classical MD simulations, using the CHARMM force-field, contain implicit isotopic biases. Further, we demonstrated that careful re-parameterization of the force-field parameters to account for the “proper” isotopic bias permit the construction of new force-field parameters for use in the examination of the impact of isotopic substitutions using classical MD simulations (in the absence of nuclear quantum effects). This success suggests that a similar methodology may be implemented in MD simulation of other deuterated systems.

The methodology presented here may be of importance in the study of small deuterated molecules of growing importance as pharmaceuticals^{174, 175}, and also serve as an aid in the interpretation of experiments such as NMR¹⁷⁶ and neutron scattering¹⁷⁷ where isotope substitution can be of critical importance.

In addition to demonstrating the potential of using classical MD simulations to account for physical isotope effects, the present MD simulations provide detailed molecular insights into the origin of the loss of the THF-water miscibility gap upon the complete deuteration of THF. The results indicate that upon deuteration, the intramolecular interactions within THF are modified so as to modify the structure of the molecule and allow for an increase in the formation of THF-water hydrogen bonds (**Fig 2.8, Fig 2.13**). These new inter-molecular interactions then compensate for the overall reduction of favorable THF-water interactions upon heating and prevent D₈THF and water from phase separating, thus eliminating the miscibility gap.

2.7 Appendix

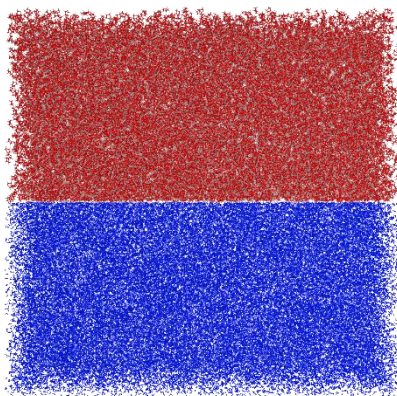


Figure 2.1 Initial configuration of MD simulation of the 16.5nm-side box (Red: THF ; Blue: Water)

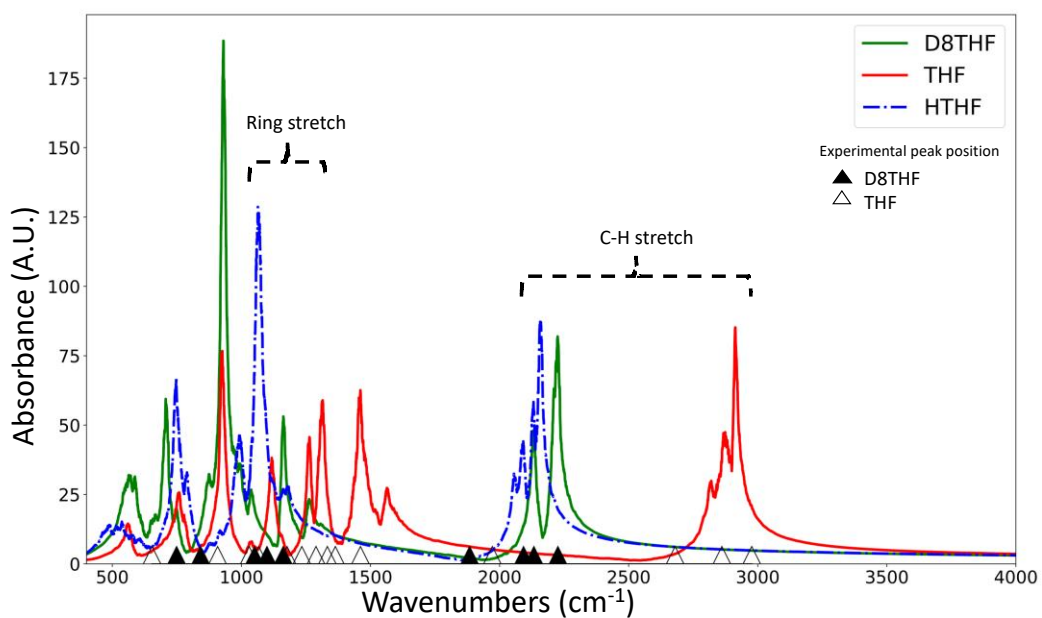


Figure 2.2: Infrared spectra of D8THF (green), HTHF (blue), THF (red) as calculated by MD simulations (lines) and experiments (peaks)

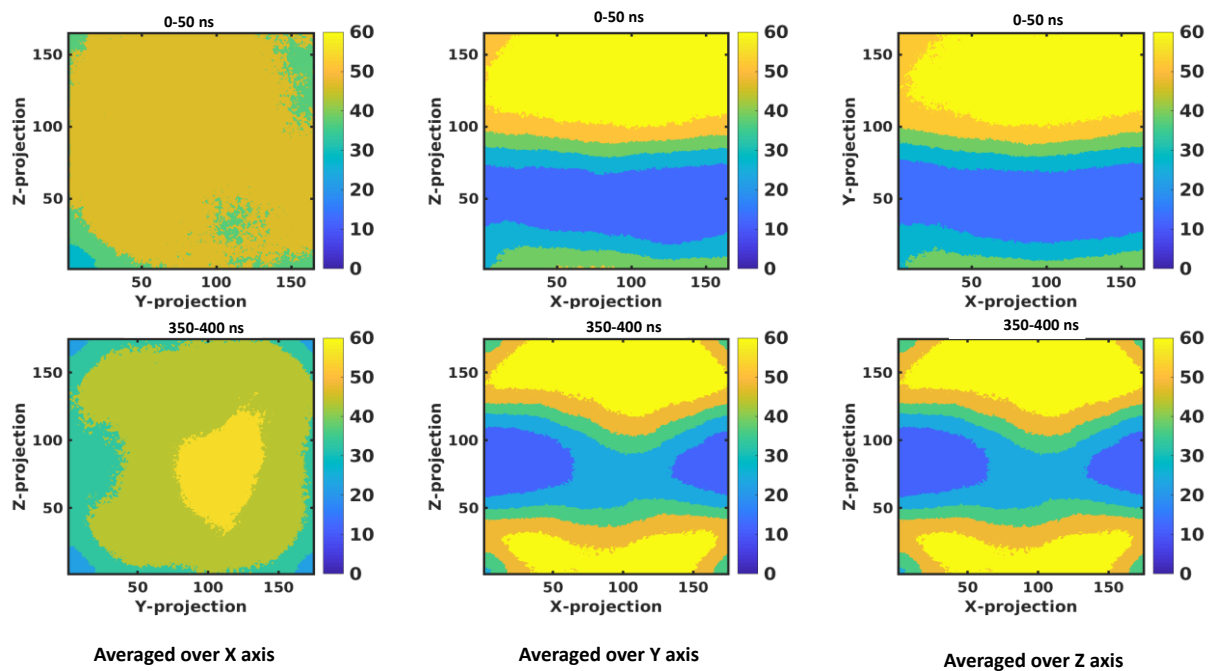


Figure 2.3: Density profiles of THF for different time intervals at 90°C. The top row is the density profiles averaged over X, Y and Z directions for first 50 ns and the bottom row is the density profiles averaged over X, Y and Z directions for last 50 ns

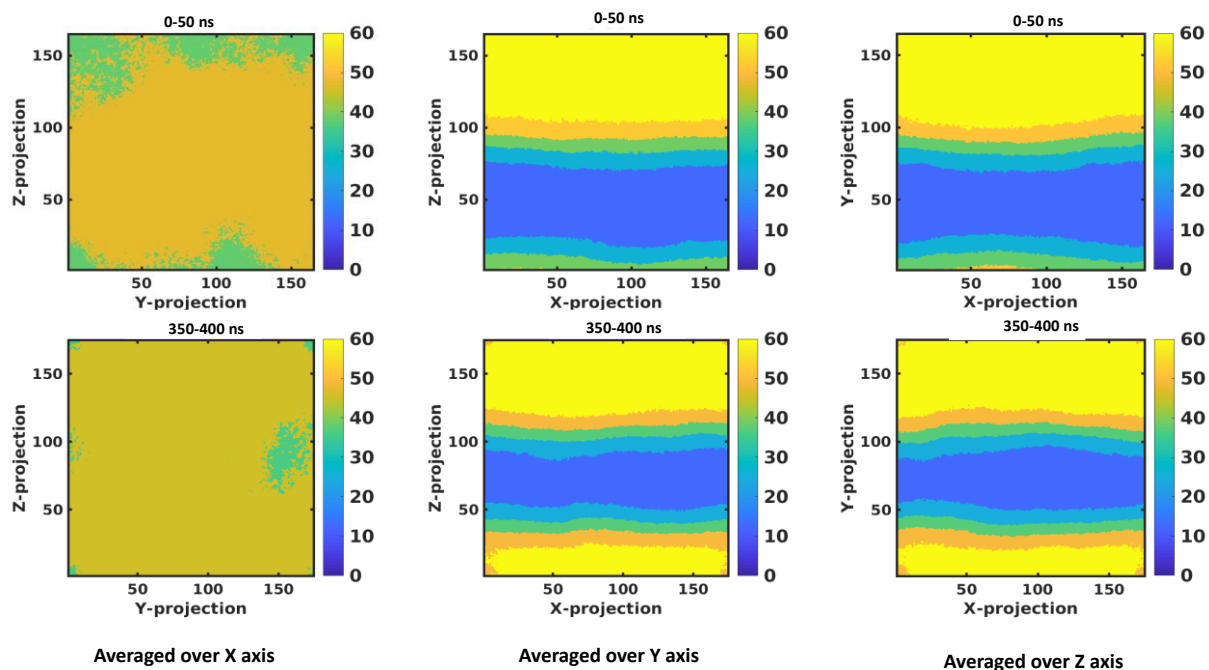


Figure 2.4: Density profiles of HTHF for different time intervals at 90°C. The top row is the density profiles averaged over X, Y and Z directions for first 50 ns and the bottom row is the density profiles averaged over X, Y and Z directions for last 50 ns

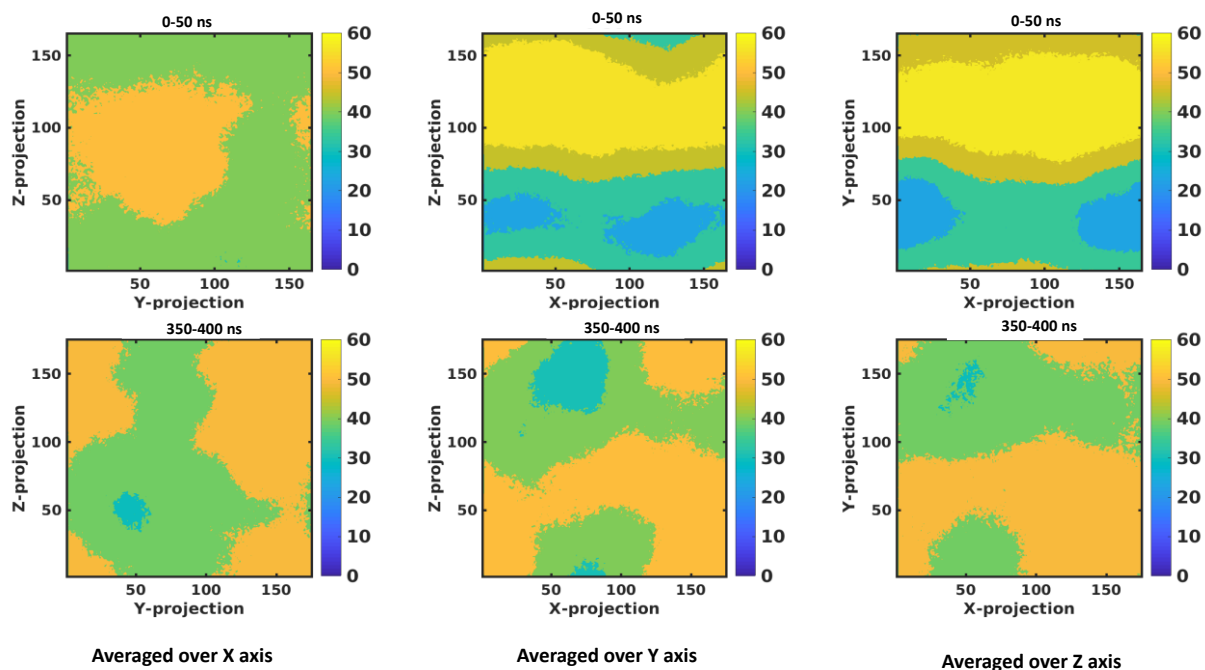


Figure 2.5 Density profiles of D8THF for different time intervals at 90°C. The top row is the density profiles averaged over X, Y and Z directions for first 50 ns and the bottom row is the density profiles averaged over X, Y and Z directions for last 50 ns

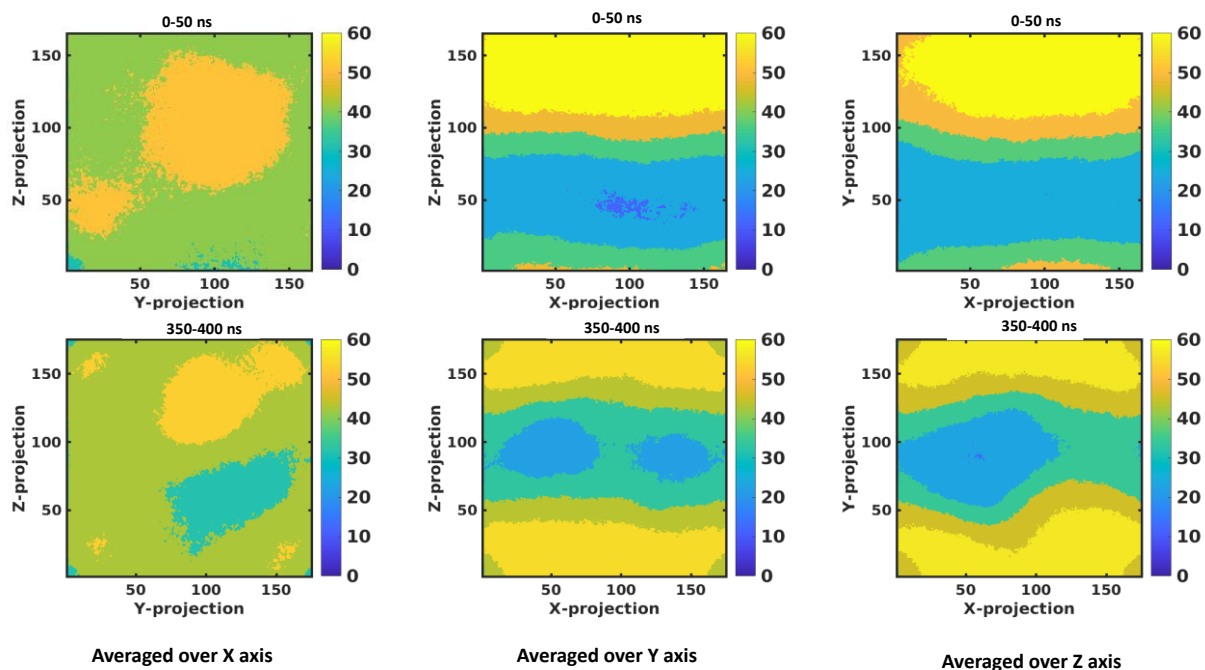


Figure 2.6 Density profiles of THFK for different time intervals at 90°C. The top row is the density profiles averaged over X, Y and Z directions for first 50 ns and the bottom row is the density profiles averaged over X, Y and Z directions for last 50 ns

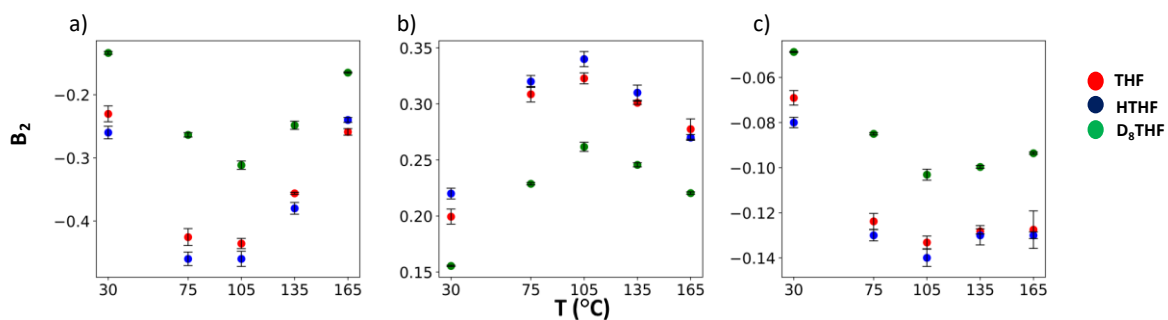


Figure 2.7 Individual components of the second virial coefficient: a) D₈THF-D₈THF, HTHF-HTHF, and THF- THF, b) D₈THF, HTHF, and THF -water c) water-water. The standard error of the mean (SEM) are shown as error bars, and those not visible are at most the size of the symbol.

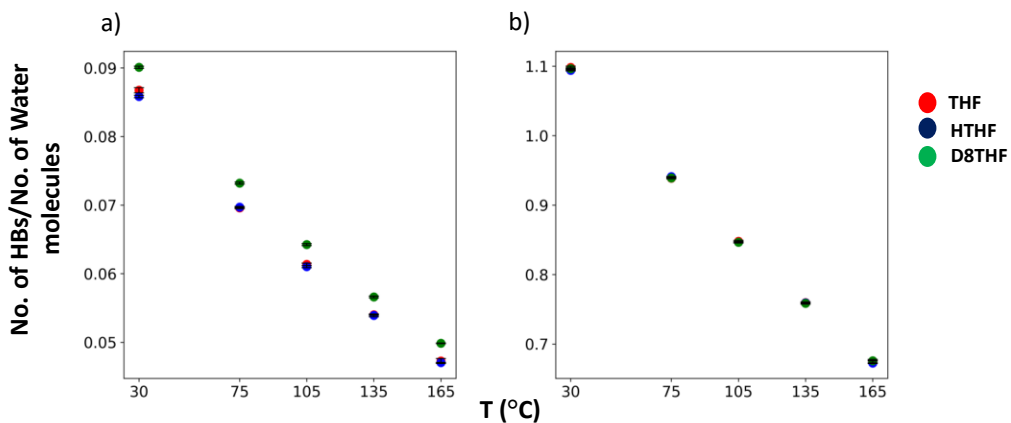


Figure 2.8 Average number of hydrogen bonds normalized by the number of water molecules: a) D₈THF, HTHF, and THF -water; b) water-water. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.

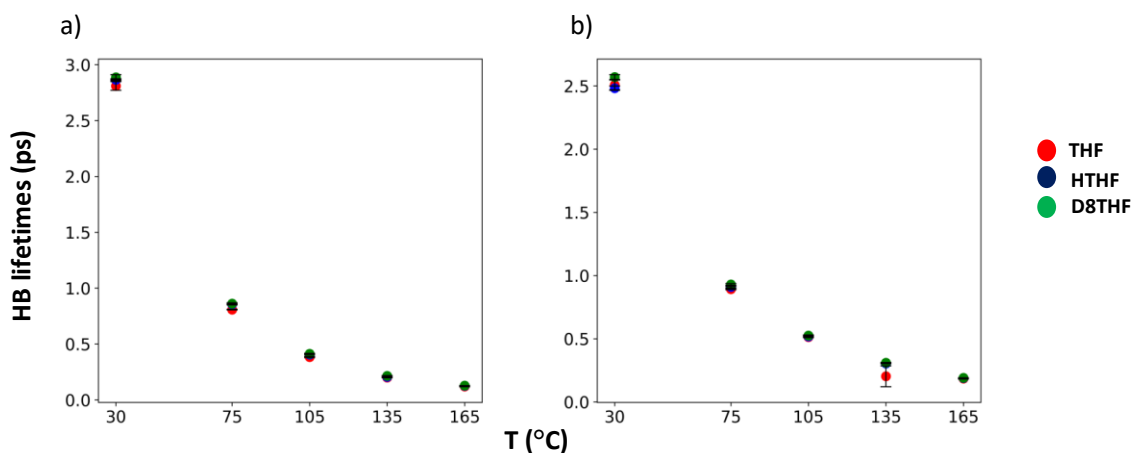


Figure 2.9 Mean hydrogen bond lifetimes: a) D₈THF, HTHF, and THF -water; b) water-water. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.

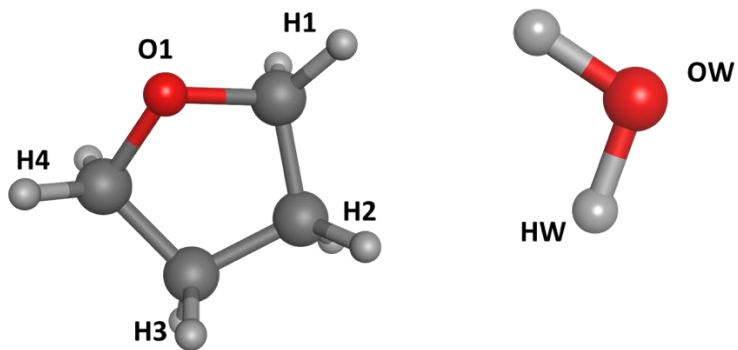


Figure 2.10 Visual representation of coordination sites

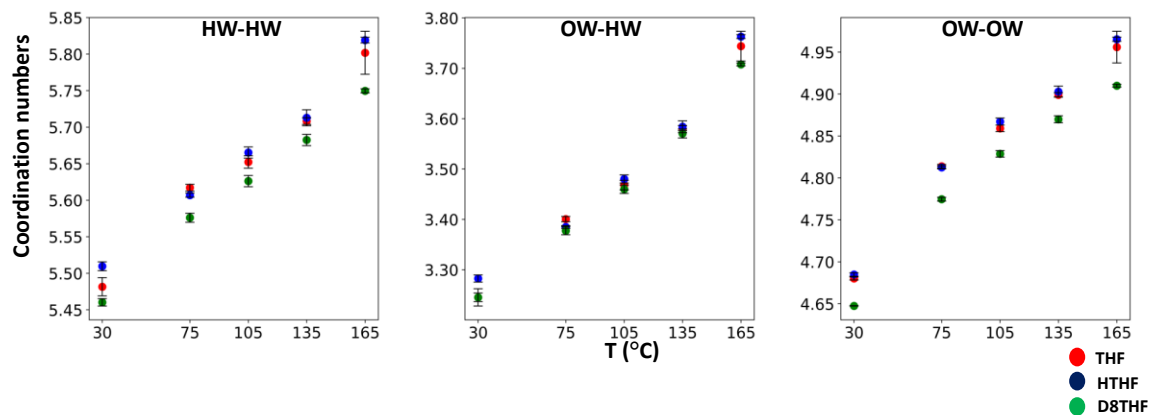


Figure 2.11 Water-water coordination numbers. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.

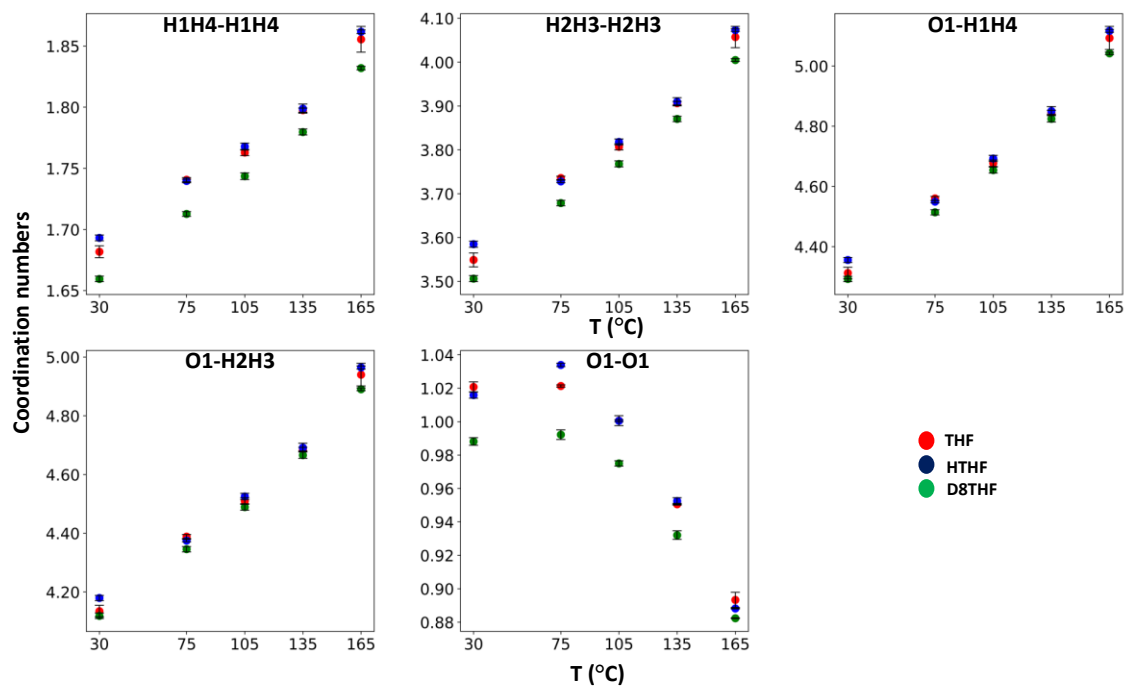


Figure 2.12 Coordination numbers between D₈THF- D₈THF, HTHF- HTHF, and THF-THF. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.

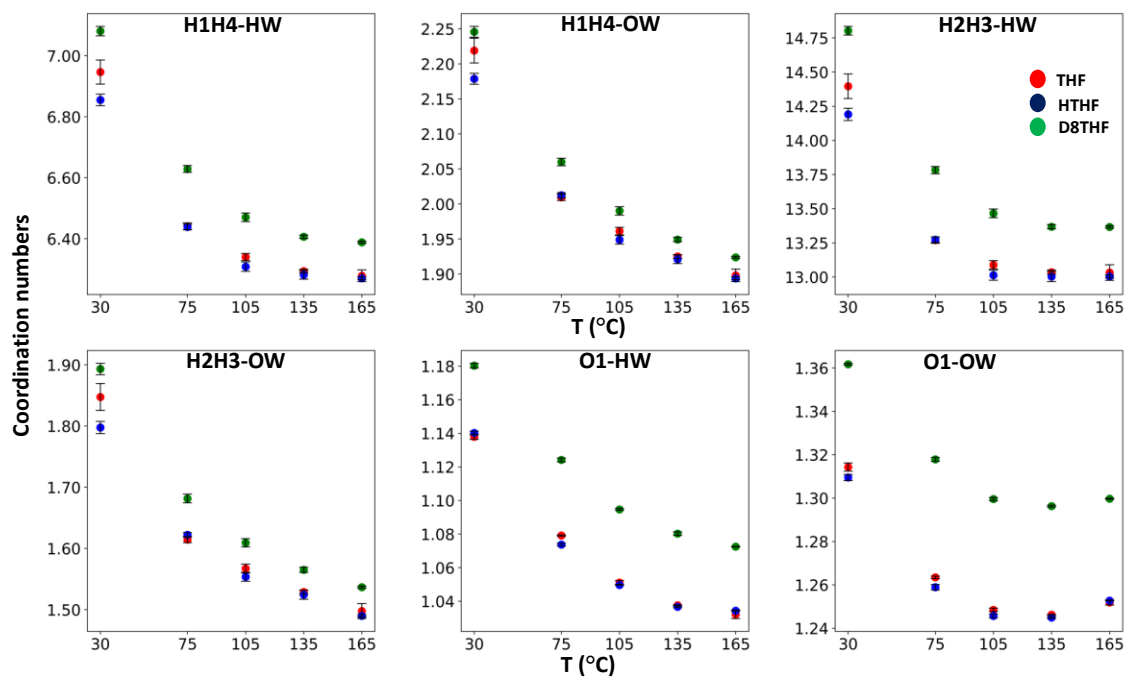


Figure 2.13 Coordination numbers between D8THF, HTHF, and THF -water. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.

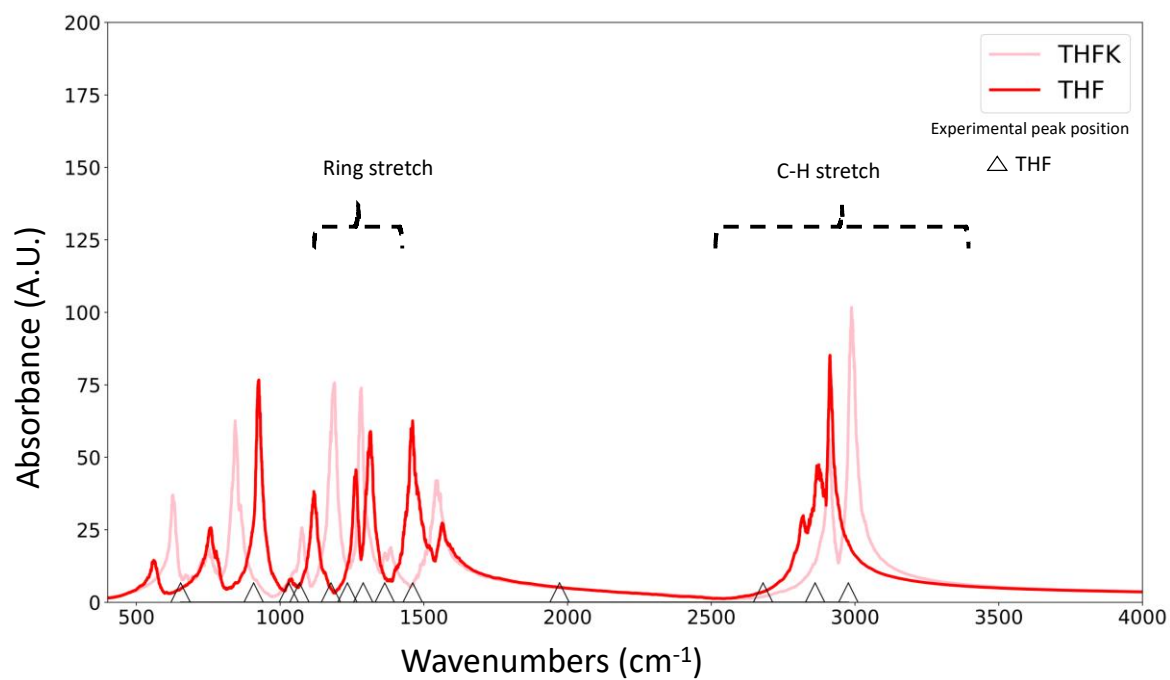


Figure 2.14 Infrared spectra of THF and THFK as calculated by MD simulations (lines) and experiments (peaks)

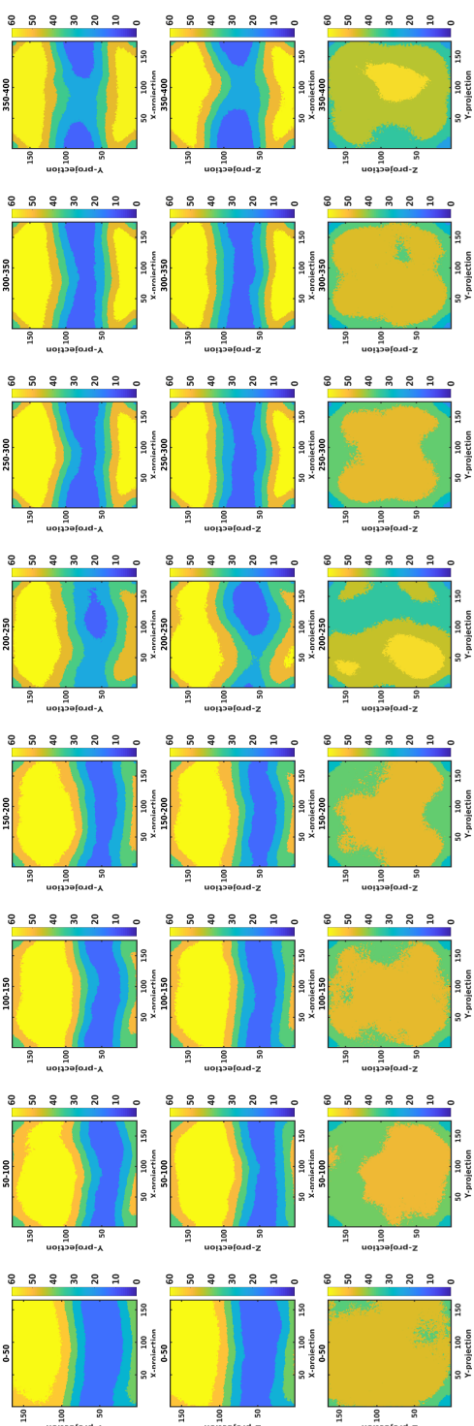


Figure 2.15 Density of THF for different time interval of 50ns at 90°C.

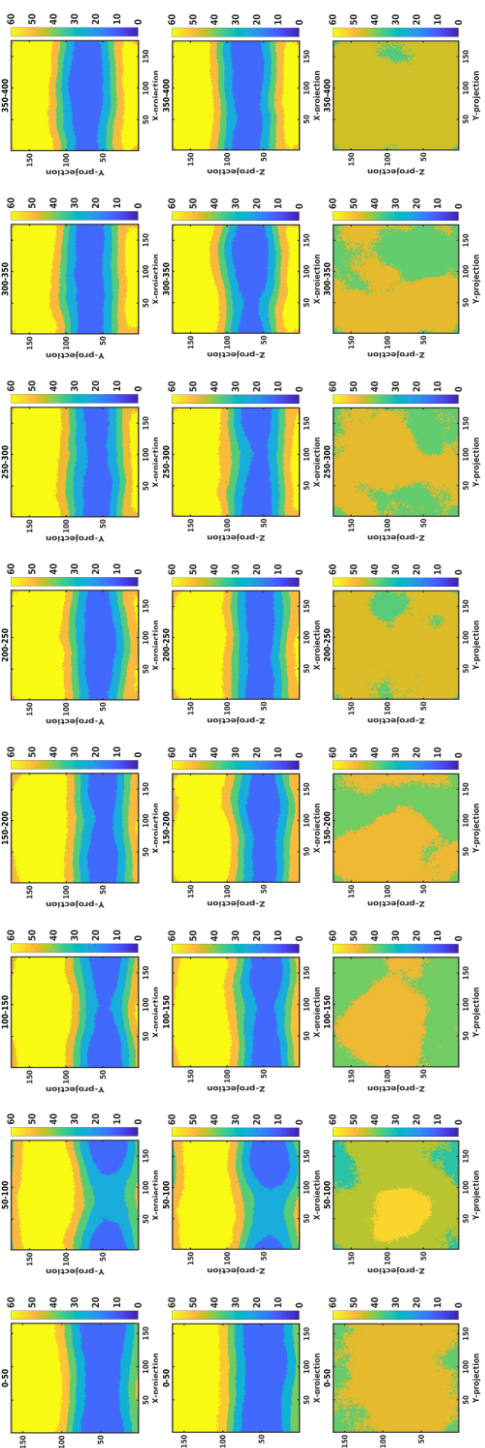


Figure 2.16 Density of HTHF for different time interval of 50ns at 90°C.

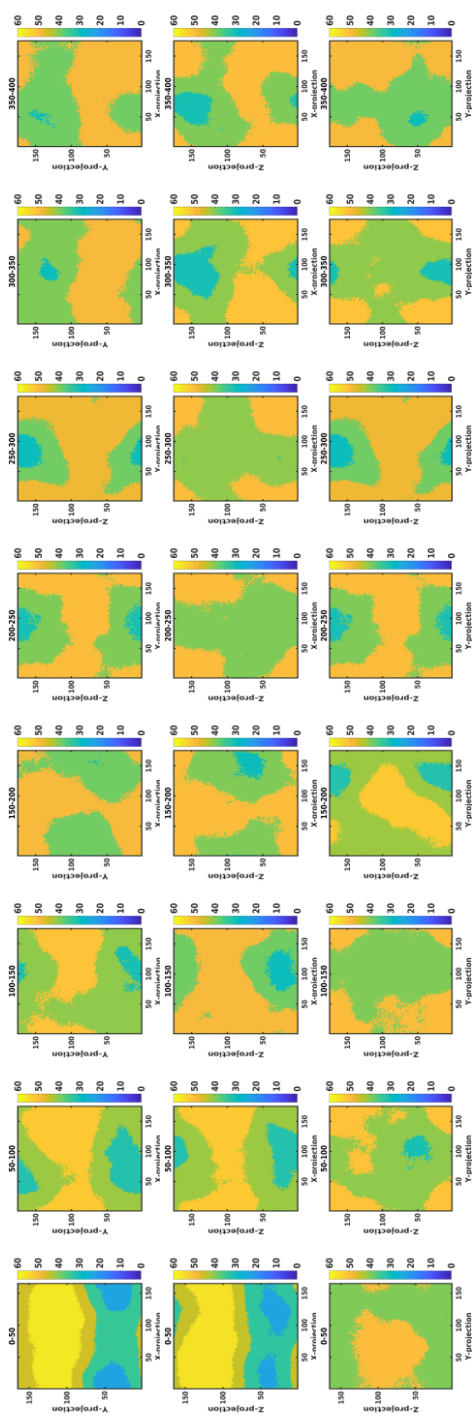


Figure 2.17 Density of D₈THF for different time interval of 50ns at 90°C.

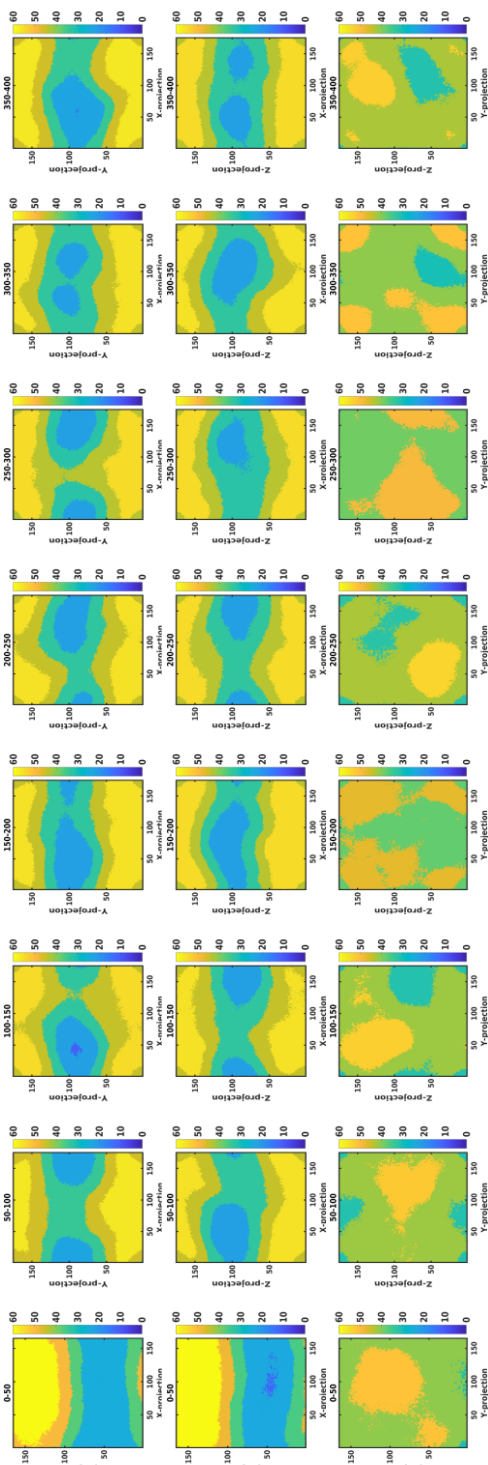


Figure 2.18 Density of THFK for different time interval of 50ns at 90°C.

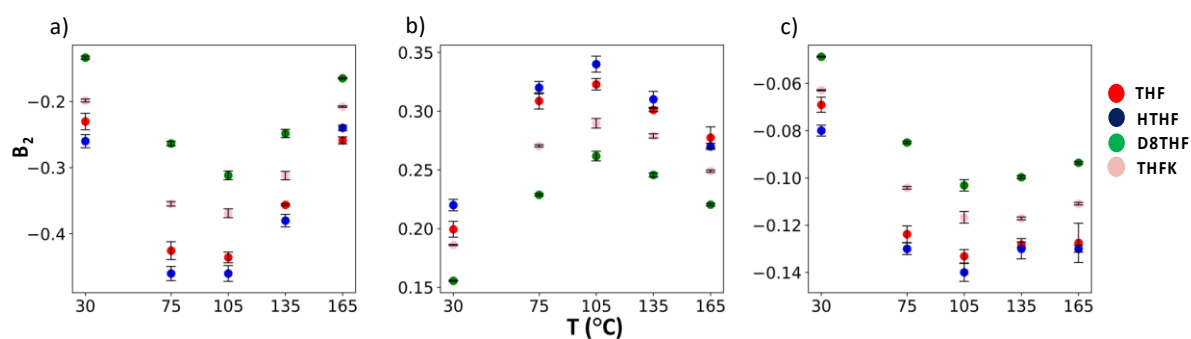


Figure 2.19 Individual components of the second virial coefficient: a) D₈THF-D₈THF, HTHF-HTHF, THF- THF, and THFK-THFK b) D₈THF, HTHF, THFK and THF -water c) water-water. Error bars show the SEM, and those not visible are at most the size of the symbol.

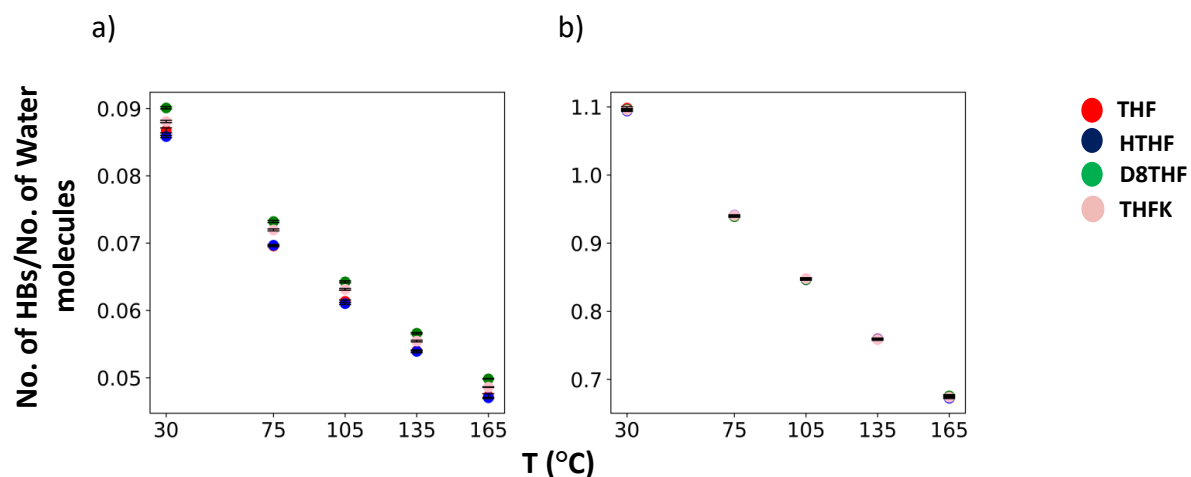


Figure 2.20 Average number of hydrogen bonds normalized by the number of water molecules: a) D₈THF, HTHF, THFK, and THF -water; b) water-water. Error bars show the SEM, and those not immediately visible are at most the size of the symbol.

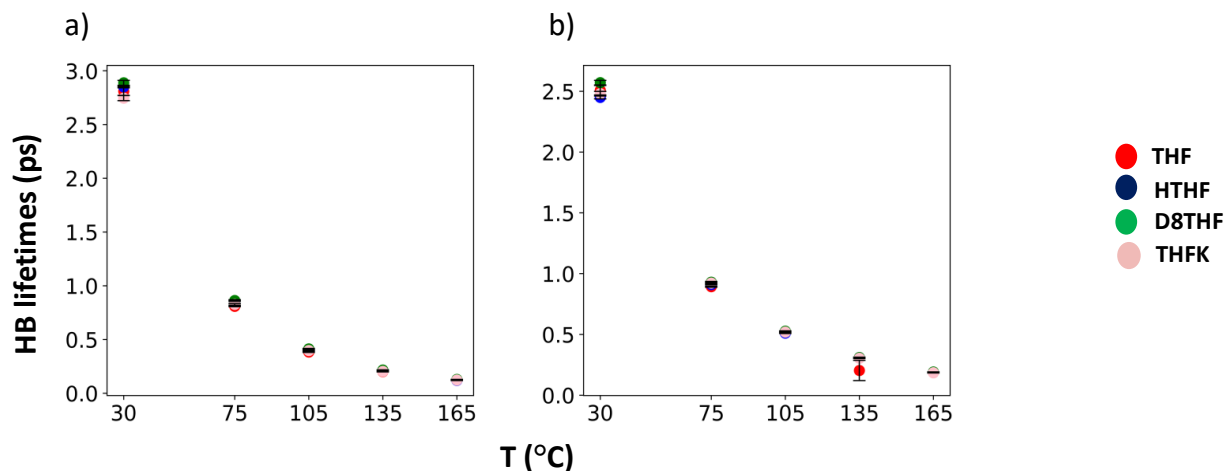


Figure 2.21 Mean hydrogen bond lifetimes: a) D₈THF, HTHF, THFK, and THF-water; b) water-water. Error bars show the SEM, and those not immediately visible are at most the size of the symbol.

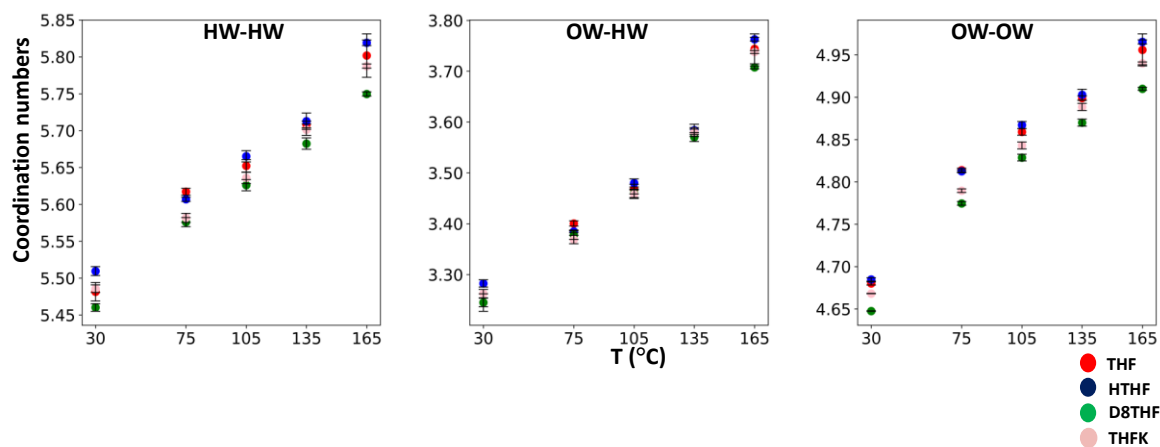


Figure 2.22 Co-ordination numbers between water-water. Error bars show the SEM, and those not immediately visible are at most the size of the symbol.

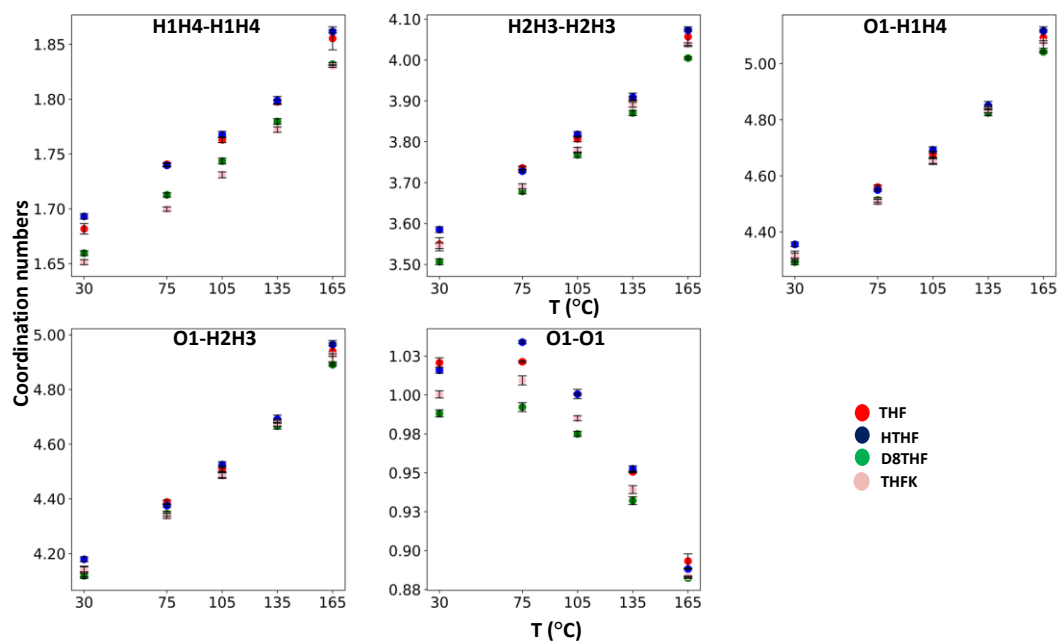


Figure 2.23 Co-ordination numbers between D₈THF- D₈THF, HTHF- HTHF, THFK-THFK, and THF-THF. Error bars show the SEM, and those not immediately visible are at most the size of the symbol.

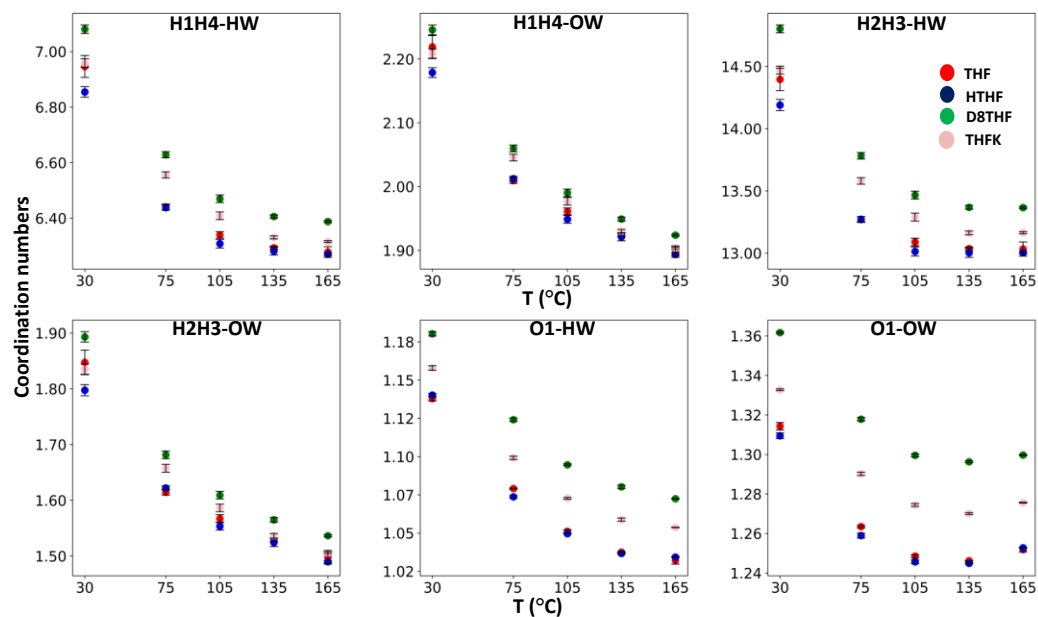


Figure 2.24 Co-ordination numbers between D₈THF, HTHF, THFK, and THF - water. Error bars show the SEM, and those not immediately visible are at most the size of the symbol.

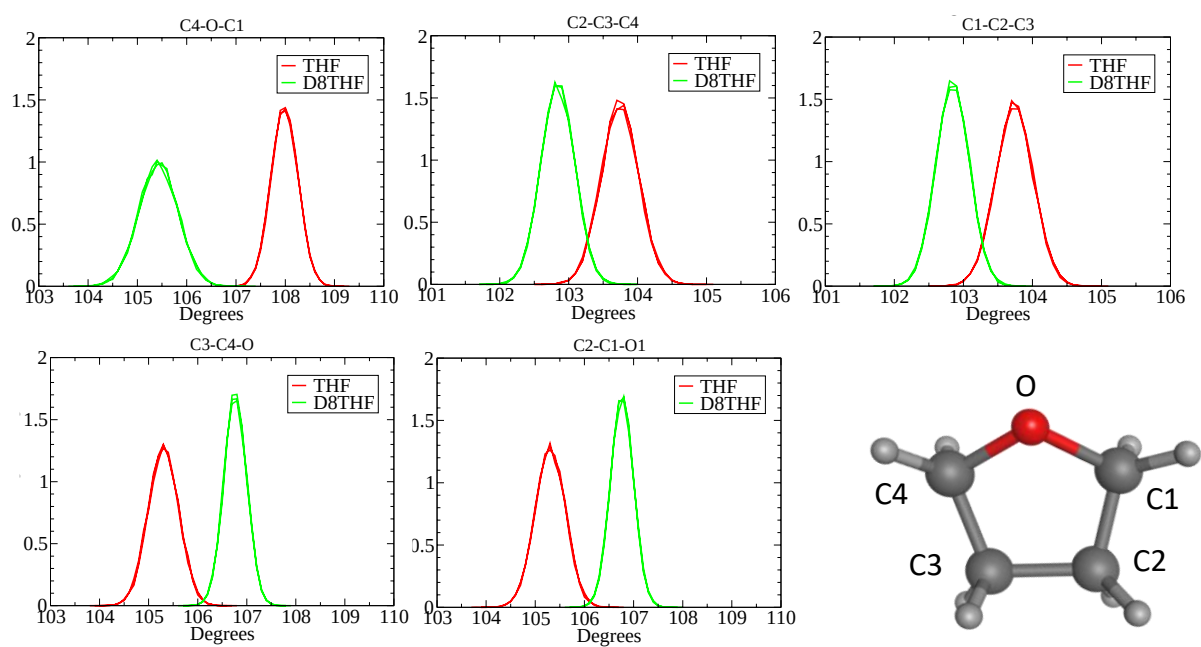


Figure 2.25 Probability distribution of angles for THF (red) and D₈THF(green) in presence of water molecules at 105 °C

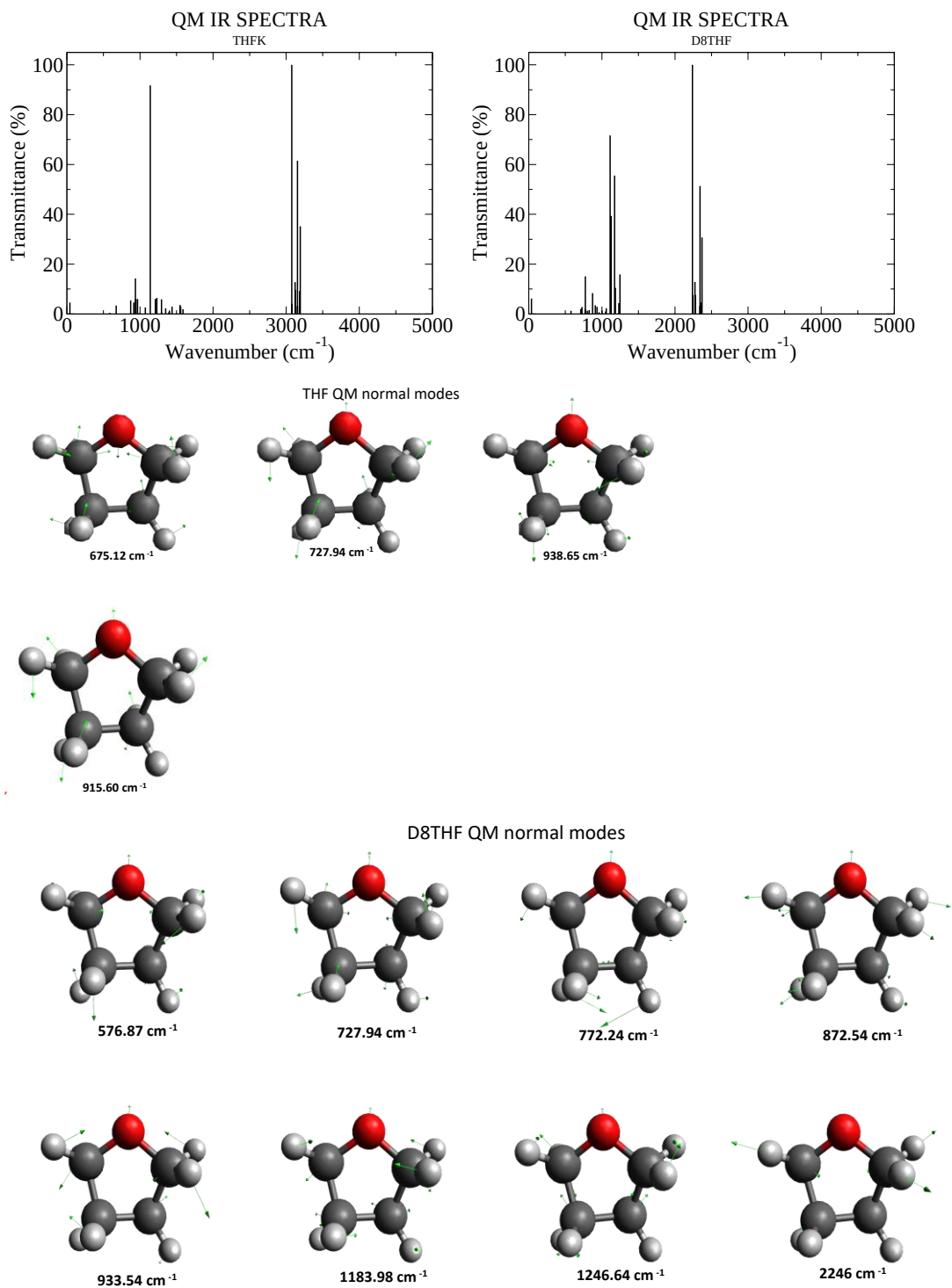


Figure 2.26 QM IR spectra (top) and corresponding QM normal modes for THF and D8THF (bottom)

Table 2.1 Bonded terms of new D₈THF parameters

Atom type	Atom type	K _b	b ₀
OC305A	CC325B	267.635	1.441
CC325B	CC325B	241.951	1.523
CC325B	HCA25A	344.015	1.094

$$V_{(\text{bond})} = K_b (b - b_0)^2$$

$$K_b: \text{kcal/mole/\AA}^2$$

$$b_0: \text{\AA}$$

Table 2.2 Angle term of new D₈THF parameters

Atom type	Atom type	Atom type	K _θ	θ ₀
OC305A	CC325B	HCA25A	62.999	108.462
OC305A	CC325B	CC325B	141.069	106.72
CC325B	CC325B	HCA25A	44.901	110.984
CC325B	CC325B	CC325B	108.129	100.701
CC325B	OC305A	CC325B	32.544	110.792
HCA25A	CC325B	HCA25A	24.863	107.157

$$V_{(\text{angle})} = K_\theta (\theta - \theta_0)^2$$

$$K_\theta: \text{kcal/mole/rad}^2$$

$$\theta_0: \text{degrees}$$

Table 2.3 Dihedral term for new D₈THF parameters

Atom type	Atom type	Atom type	Atom type	K χ	n	δ
HCA25A	CC325B	CC325B	HCA25A	2.644	1	0
CC325B	CC325B	CC325B	CC325B	2.597	1	180
CC325B	OC305A	CC325B	CC325B	1.148	1	180
CC325B	CC325B	CC325B	OC305A	2.992	1	180
CC325B	CC325B	CC325B	HCA25A	2.465	1	0
HCA25A	CC325B	CC325B	OC305A	2.807	1	180
CC325B	OC305A	CC325B	HCA25A	2.423	1	180

$$V_{(\text{dihedral})} = K\chi(1 + \cos(n(\chi) - \delta))$$

K χ : kcal/mole

n: multiplicity

delta: degrees

Table 2.4 Lennard-Jones potential term of new D₈THF parameters

Atom type	epsilon	R _{min} /2	Ignored	eps ₁₋₄	R _{min} /2 ₁₋₄
CC325B	-0.06	2.02	0	-0.01	1.9
HCA25A	-0.035	1.3			
OC305A	-0.1	1.65			

Table 2.5 Bonded terms of new THFK parameters

Atom type	Atom type	K_b	b_0
OC305A	CC325B	262.599	1.438
CC325B	CC325B	241.762	1.526
CC325B	HCA25A	345.662	1.095

$$V_{(\text{bond})} = K_b (b - b_0)^2$$

$$K_b : \text{kcal/mole/\AA}^2$$

$$b_0 : \text{\AA}$$

Table 2.6 Angle term of new THFK parameters

Atom type	Atom type	Atom type	K_θ	θ_0
OC305A	CC325B	HCA25A	58.594	108.351
OC305A	CC325B	CC325B	146.195	106.406
CC325B	CC325B	HCA25A	36.183	110.953
CC325B	CC325B	CC325B	109.996	101.417
CC325B	OC305A	CC325B	41.030	108.634
HCA25A	CC325B	HCA25A	42.288	107.785

$$V_{(\text{angle})} = K_\theta (\theta - \theta_0)^2$$

$$K_\theta : \text{kcal/mole/rad}^2$$

$$\theta_0 : \text{degrees}$$

Table 2.7 Dihedral term for new THFK parameters

Atom type	Atom type	Atom type	Atom type	K χ	n	δ
HCA25A	CC325B	CC325B	HCA25A	2.999	1	0
CC325B	CC325B	CC325B	CC325B	2.996	1	180
CC325B	OC305A	CC325B	CC325B	1.773	1	180
CC325B	CC325B	CC325B	OC305A	2.996	1	180
CC325B	CC325B	CC325B	HCA25A	0.993	1	0
HCA25A	CC325B	CC325B	OC305A	0.037	1	180
CC325B	OC305A	CC325B	HCA25A	3.000	1	180

$$V_{(\text{dihedral})} = K\chi(1 + \cos(n(\chi) - \delta))$$

K χ : kcal/mole

n: multiplicity

delta: degrees

Table 2.8 Lennard-Jones potential term of new THFK parameters

Atom type	epsilon	R _{min} /2	Ignored	eps ₁₋₄	R _{min} /2 ₁₋₄
CC325B	-0.06	2.02	0	-0.01	1.9
HCA25A	-0.035	1.3			
OC305A	-0.1	1.65			

Table 2.9 Cutoffs for Coordination Number Calculations**(Cutoff = location of first minima in corresponding site-site RDFs)**

Site 1	Site 2	Cut off (nm)
THF-O1	water-OW	0.34
THF-O1	water-HW	0.25
THF-O1	THF-O1	0.53
THF-O1	THF-H1H4	0.38
THF-O1	THF-H2H3	0.36
THF-H1H4	water-OW	0.37
THF-H1H4	water-HW	0.41
THF-H1H4	THF-H1H4	0.31
THF-H1H4	THF-H2H3	0.32
THF-H2H3	water-OW	0.37
THF-H2H3	water-HW	0.51
THF-H2H3	THF-H2H3	0.35
water-OW	water-OW	0.37
water-OW	water-HW	0.25
water-HW	water-HW	0.31

3 STUDYING GLOBAL AND LOCAL MOTION/DYNAMICS OF BACTERIAL PROTEINS

3.1 Promiscuous binding of bacterial Siglec-like adhesin Hsa to human and animal sialoglycan receptors

3.1.1 Abstract

Sialoglycan binding proteins such as adhesin proteins play a crucial role in many biological activities. Multiple adhesin proteins exist with distinct affinities towards different sialoglycans. Adhesin proteins like GspB and SrpA exhibit selective binding or preferential binding towards Neu5Ac-Gal or Neu5Gc-Gal, respectively, whereas adhesin protein Hsa exhibits promiscuity and binds both these sugars equally. However, the mechanisms behind the selectivity and binding of different adhesin proteins to the respective sialoglycans are still not very well understood. To explore this further, we first determined the co-crystal structures of Neu5Ac-Gal and Neu5Gc-Gal disaccharides, respectively, with the sialoglycan binding region of the unselective adhesin protein Hsa (Hsa_{BR}) from *S. gordonii* strain Challis. We then used molecular dynamics (MD) simulations to study the promiscuity in the binding of Hsa_{BR} to both Neu5Ac-Gal and Neu5Gc-Gal disaccharides. We further contrasted the binding mechanism in Hsa_{BR} with that observed in selective adhesins mentioned above. Here, we report a potential mechanism behind the promiscuity or decreased selectivity of Hsa_{BR} when compared to other selective adhesin proteins GspB and SrpA.

3.1.2 Introduction

Sialic acid carrying glycans (sialoglycans), present on the surface of platelets, play a crucial role in the pathogenesis of *Streptococci viridians*-associated infective endocarditis (IE), a life-threatening cardiovascular infection¹⁷⁸⁻¹⁸⁰. This group of *Streptococci* can bind to sialoglycans using adhesin proteins present on their cell surfaces to initiate adherence¹⁸¹⁻¹⁸³. Sialic acid (Sia) collectively refers to a broad range of neuraminic acid derivatives. Although over 50 distinct forms of Sia are found in nature, N-acetylneuraminic acid (Neu5Ac) and N-glycolylneuraminic acid (Neu5Gc) are most commonly observed¹⁸⁴. Unlike most animals, human cells have

lost the ability to synthesize Neu5Gc and hence only contain Neu5Ac¹⁸⁴. The only difference between the structures of Neu5Gc and Neu5Ac is the presence of an extra hydroxyl group in Neu5Gc at the C11 position as shown in **Fig. 3.1**. It has been proposed that the loss of Neu5Gc biosynthesis confers an evolutionary advantage in humans by allowing resistance to ancient forms of the malaria parasite that used Neu5Gc as a receptor¹⁸⁵, whereas on the contrary, *V. cholerae* has evolved into a human-specific pathogen partly by adapting to this evolutionary loss of Neu5Gc¹⁸⁶. Sialoglycans and their binding proteins appear to play important roles in many biological recognition mechanisms and diseases¹⁸⁷⁻¹⁸⁹. Interestingly, their binding properties suggest potential applications as specific probes for investigating the role of cell surface carbohydrates during the development and malignant transformation of cancer cells¹⁹⁰⁻¹⁹³. A recent paper has also reported that Neu5Ac has a high affinity towards the SARS-CoV-2 spike glycoprotein¹⁹⁴. Although these studies provide details about the different sialoglycan binding proteins, the respective binding mechanisms are still not well understood.

Similar to most sialoglycan-binding proteins, the streptococcal adhesin proteins are Sia-recognizing immunoglobulin superfamily lectin (Siglec) like and interact with a range of sialoglycans with different selectivities, as has been characterized by a glycan array analysis¹⁹⁵. Some Siglec-like adhesins, such as GspB from *Streptococcus gordonii* strain M99, bind selectively to Neu5Ac-containing sialoglycans¹⁹⁵. In contrast, Siglec-like adhesins such as SrpA from *S. sanguinis* strain SK36 can bind to both Neu5Ac and Neu5Gc but bind to Neu5Gc-containing sialoglycans with a higher affinity than those containing Neu5Ac¹⁹⁶. Other adhesins, such as Hsa from *S. gordonii* strain Challis, have dual specificity and bind with similar affinities to both Neu5Ac and Neu5Gc¹⁹⁵. The mechanism or reasoning behind the specificities of these adhesins is still not well understood.

Previous studies on Neu5Gc bound to the binding region (BR) of SrpA have suggested that the formation of a specific hydrogen bond with the extra hydroxyl on Neu5Gc (C11 position of Sia, **Fig. 3.1**) potentially allows SrpA_{BR} to bind preferentially to Neu5Gc¹⁹⁷, thus concluding that specific contacts to the C11

hydroxyl on Neu5Gc dictate human versus animal Sia selectivity in members of this family of bacterial adhesins. However, this mechanism fails to explain the promiscuity observed in the case of Hsa- like adhesin proteins, which bind both Neu5Gc and Neu5Ac -containing sialoglycans with equivalent affinities¹⁹⁵. Prior structural studies on the Neu5Ac/Neu5Gc promiscuous *S. gordonii* Hsa_{BR} have identified that binding to a Neu5Ac-based trisaccharide (sialyl T antigen (sTa)) involves a conformational selection mechanism¹⁹⁸ in which the EF loop in the binding site of Hsa changes its orientation to a “closed” state to form a hydrogen bond¹⁹⁸. Thus, one compelling hypothesis for the observed equivalent binding affinities in Hsa for Neu5Ac/Neu5Gc is that distinct conformational states accompany the binding of Neu5Ac- and Neu5Gc-based sialoglycans that promote different contacts to the C11 position, similar to what has been observed for sTa. The current study thus aims to examine in detail how an adhesin protein such as Hsa can exhibit dual specificity for both Neu5Ac and Neu5Gc. To this end, we first determined the co-crystal structures of Neu5Ac-Gal and Neu5Gc-Gal disaccharides, respectively, with the sialoglycan binding region of the adhesin Hsa (Hsa_{BR}) from *S. gordonii* strain Challis (**Fig. 3.2b**). We found that the overall structures of Hsa_{BR} bound to each sialoglycan were very similar, with an RMSD of ~1Å over Cα atoms. Interestingly, we observed that in the crystal structures Hsa_{BR} was bound to Neu5Ac-Gal and Neu5Gc-Gal with different loop conformations- in case of Neu5Gc-Gal the EF loop was in the “closed” state but in the Neu5Ac-Gal, it was in the “open” state, which contradicted a previous Hsa_{BR} study with sTa (Neu5Ac containing trisaccharide). However, corroborating the previous study with sTa, analyses using molecular dynamics (MD) simulations showed that this conformational difference between the two Hsa_{BR} crystal structures was a crystallization artifact and that both disaccharides bound to the closed conformation of the adhesin protein. Furthermore, using MD simulation of both these crystal structures, we uncovered a potential mechanism behind the binding promiscuity in Hsa. To provide further insights behind the comparatively higher fidelity binding in more specific adhesin proteins, we performed *in silico* modeling

and MD simulations on GspB_{BR} (with non-native ligand Neu5Gc) and SrpA_{BR} (with preferential ligand Neu5Gc) adhesin proteins that do not bind both sialic acid forms with the same affinity.

3.1.3 Methods

System preparation and molecular dynamics simulation

The experimental work in this chapter was done by Dr. Tina Iverson's group in Vanderbilt University and Dr. Barbara Bensing's group at University of California, SF. Crystal structures of the binding region of adhesin proteins: Hsa_{BR} with Neu5Ac-Gal and Neu5Gc-Gal explained above were used. Molecular dynamics (MD) simulations were performed on the proteins and glycan ligands using Amber14 ff14SB and Glycam06¹⁹⁹ parameters, respectively. The protein-glycan bound system was hydrated by water model TIP3P²⁰⁰ in an octahedral box of 10 Å around the protein in each direction. First, the protein was held fixed with a force constant of 500 kcal mol⁻¹ Å⁻² while the system was energy minimized with 500 steps of steepest descent followed by 500 steps with the conjugate gradient method. In a second minimization step, the restraints on the protein were removed, and 1000 steps of steepest descent minimization was performed, followed by 1500 steps of conjugate gradient. The system was heated to 300 K while holding the protein fixed with a force constant of 10 kcal mol⁻¹ Å⁻² for 1000 steps. Then, the restraints were removed, and 1000 MD steps were performed. The SHAKE algorithm²⁰¹ was used to constrain all bonds involving hydrogen in the simulations. MD production runs were performed at 300 K using the NPT ensemble and a 2 fs time step with nonbonded cutoff of 10 Å. Temperature was fixed with the Langevin dynamics thermostat²⁰² and the pressure was fixed with Monte Carlo barostat²⁰³. Three independent runs were performed for each simulation. All the analyses were performed using Pytraj python package²⁰⁴.

3.1.4 Results and Discussion

Crystal structure comparison of Hsa_{BR} with Neu5Ac-Gal and Neu5Gc-Gal

The X-ray crystal structures of Hsa_{BR} with Neu5Ac-Gal and Neu5Gc-Gal (residues 245–449 of the full-length protein) were determined to 1.5 Å resolution using de novo phasing calculated from single-wavelength anomalous diffraction of an osmium derivative. Each Hsa_{BR} protomer folds into two domains (Siglec and Unique) (**Fig. 3.2b**). Siglec is the sialoglycan binding domain, and the function of the Unique domain is still not well understood²⁰⁵. The Hsa_{BR} co-crystals with Neu5Gc-Gal and Neu5Ac-Gal contain 280 and 295 water molecules, respectively and bound calcium (Ca²⁺) and sodium (Na⁺) ions. The calcium (Ca²⁺) and sodium (Na⁺) ions were bound at the same place, as has been previously observed in the apo state crystal structure (PDB 6EFC)¹⁹⁸. The overall backbone structure of the holo form with Neu5Ac-Gal and Neu5Gc-Gal bound was found to be very close to the apo form (PDB 6EFC)¹⁹⁸ with the root mean square deviation (RMSD) of C α atoms being 0.34 Å and 1.12 Å, respectively. To identify the differences between the two bound structures, the crystal structures were superimposed. The overall structures were very similar, with a root mean square deviation (RMSD) over the C α atoms of only 1.06 Å. Furthermore, we plotted the B-factors from both the crystal structures to identify the flexibility of protein residues (**Fig. 3.2a**). The flexibility of the C α atoms of both the proteins was also very similar. The only significant difference between the systems was in the EF loop (330-336) region. This loop is particularly interesting because our previous molecular dynamics (MD) simulation studies of the apo form of Hsa_{BR} have shown that the EF loop can sample two states: “open” and “closed”, and that the Neu5Ac containing the sialyl T antigen (sTa) structure binds in the “closed” state (**Fig. 3.9**)¹⁹⁸. In the case of Neu5Gc-Gal, we observed that the carbonyl-O atom of Lys 335 (present in the EF loop) forms a hydrogen bond with the hydroxyl group (O4) of Neu5Gc-Gal and EF loop was crystallized in a “closed” state, similar to that seen in the sTa bound crystal structure¹⁹⁸ and in MD studies¹⁹⁸. In the case of Neu5Ac-Gal, no such interaction between Neu5Ac-Gal and Lys 335 was observed, and the protein EF

loop was crystallized in an “open” state. These findings suggested that although disaccharide (Neu5Ac-Gal) and trisaccharide (sTa) forms of Neu5Ac bind the protein in the same pose in their respective crystal structures, Neu5Ac forms no interaction with the Lys 335 in the disaccharide form (Neu5Ac-Gal) and is crystallized in an “open” state, whereas it is crystallized in a “closed” state in trisaccharide form (**Fig. 3.9**). Apart from the above, both the disaccharide molecules (Neu5Ac-Gal and Neu5Gc-Gal) aligned perfectly on top of each other and formed similar interactions in the binding pocket of the protein (**Fig. 3.2b, 3.8**). Moreover, both Neu5Ac and Neu5Gc groups formed four common hydrogen bonds with the residues in the binding pocket (with Tyr 337, Thr 339, Arg 340, Tyr 341), and Neu5Gc formed an additional hydrogen bond with Lys 335, as mentioned above. The C11-hydroxyl group (the additional hydroxyl group present only in Neu5Gc), however, did not form any hydrogen bonds with the residues in the binding pocket but is present in close proximity of residues Lys 335, Ser 336, Tyr 337, Tyr 338 and hence could have some short-range electrostatic interactions (**Fig. 3.2b**).

Molecular dynamics simulation of the two crystal structures

Since a crystal structure provides only a static structure of the protein, hence to derive detailed spatial and temporal information including a picture of the accessible conformational sub-states, we performed MD simulations on both the sialoglycan bound crystal structures. We calculated the root mean square fluctuation (RMSF) of the respective Siglec domains in both the structures to estimate the flexibility of the protein over the course of the simulations. Similar to the X-ray B-factors, the RMSFs showed that in case of the Neu5Ac-Gal bound structure, the EF loop region (330-336) was more flexible than that in the Neu5Gc-Gal bound structure (**Fig. 3.3**). Other parts of the protein, however, had very similar fluctuations.

To further characterize the EF loop flexibility we calculated the distance of the EF loop (Lys 335 backbone carbonyl) to the hydroxyl group present in Neu5Ac (labeled as O4) as shown in (**Fig. 3.1**) to measure the “open” and “closed” state distributions, as that was the only difference between the crystalized structures (**Fig. 3.4a, 3.9**). Surprisingly, in contrast to what was seen in the crystal structure, during the course of the simulation Neu5Ac-Gal bound structure also changed from “open” to “closed” state. The latter also formed a hydrogen bond with the Lys 335, as was observed for the Neu5Gc- and sTa-bound structures (**Fig. 3.4c**). It was seen that the EF loop Neu5Ac-Gal bound structure remained in the “closed” state once the hydrogen bond between Neu5Ac-Gal and Lys 335 was formed. This result suggests that the “open” state of the Neu5Ac-Gal bound structure observed earlier could be a crystal packing effect or an artifact of crystallization. On the other hand, similar to the observation in the initial Neu5Gc-Gal bound crystal structure, the EF loop in Neu5Gc-Gal structure remained in the “closed” state throughout the course of the simulation (**Fig. 3.4b**).

To further examine the flexibility of Neu5Gc-Gal and Neu5Ac-Gal, we calculated the root mean square deviations (RMSD) of the heavy (non-hydrogen) atoms of the two disaccharides. The resulting probability distributions showed that the peak of RMSD was around 1-1.5 Å for both the disaccharides, suggesting that they are very inflexible in both the cases such that they remain very close to the initial pose observed in the crystal structure (**Fig. 3.5**). Additionally, in the case of Neu5Ac-Gal, we observed smaller peaks around 2-2.5 Å, which likely happened before Neu5Ac-O4 hydroxyl formed the hydrogen bond with Lys 335.

Apart from the relative flexibility of the structures, it was observed that during the course of the simulation the C11-OH group of Neu5Gc was faced/pointed in a direction away from the protein and was completely solvent-exposed (44 ($\pm$ 3.6) % of 300 ns) (no protein atoms were found within 3 Å of C11-OH atoms). Interestingly, this observation did not agree with a previous study on another Neu5Gc binding protein, SrpA¹⁹⁶, where it was suggested that a specific hydrogen bonding interaction between the C11-OH group and a tyrosine residue of the protein (SrpA)

is critical for binding to Neu5Gc¹⁹⁶. In our study, in the case of Hsa_{BR}, the extra C11-OH group in Neu5Gc did not have a bonding interaction with the protein, making the overall sialoglycan-protein interaction similar to that with Neu5Ac. This suggests that the extra C11-OH group in Neu5Gc does not play a role in binding to Hsa_{BR}, which could potentially be a reason behind the adhesin protein's promiscuous equivalent affinity binding to both the sialoglycans.

Comparison with other adhesin protein: SrpA and GspB

Hence, to get further insights on this binding mechanism of Neu5Ac-Gal and Neu5Gc-Gal in the adhesin binding pocket, we explored the crystal structure (PDB 5EQ3)¹⁹⁶ of another adhesin protein SrpA_{BR} (from *S. sanguinis* strain SK36) that binds preferentially to Neu5Gc.

To further explore if the lack of C11-OH interaction in Neu5Gc could be driving Hsa's promiscuous binding to both the sialoglycans, we performed MD simulations on other selective adhesin proteins to contrast and identify potential differences in the sialoglycan binding mechanisms compared to Hsa. We first explored the bound structure of adhesin protein SrpA_{BR} (PDB 5EQ3¹⁹⁶, from *S. sanguinis* strain SK36) that preferentially binds to Neu5Gc over Neu5Ac. As with Hsa_{BR} bound with Neu5Gc-Gal, the C11-hydroxyl group of Neu5Gc is in close proximity to the residues in the binding pocket of SrpA_{BR}: Arg 342, Ala 341, Gln 344 and Phe 345. However, unlike Hsa_{BR}, Tyr 368 in SrpA_{BR} formed a hydrogen bond with the C11-hydroxyl group of Neu5Gc. Notably, it was observed that in Hsa_{BR}, a tyrosine residue corresponding to that seen in SrpA_{BR} was missing, thus precluding the formation of a "selective/selectivity defining" hydrogen bond with the C11-hydroxyl group of Neu5Gc in Hsa_{BR}. This further explains the observation that C11 hydroxyl group is facing outward (away from the protein) in Hsa_{BR} crystal structure and in MD simulations, and instead forms interactions with water molecules (**Fig. 3.7b**). In line with this, in the SrpA_{BR} crystal structure (PDB 5EQ3) the C11-hydroxyl group of Neu5Gc faces inward (towards the protein)¹⁹⁶ (**Fig. 3.7a**) as a more favorable interaction with the Tyr residue is available. Additionally, MD simulations of SrpA_{BR}

with Neu5Gc-Gal showed that, similar to Hsa_{BR}, the C11-hydroxyl group of Neu5Gc did exhibit intermittent interactions with the solvent but primarily maintained at least a single hydrogen bond with Tyr 368 in SrpA_{BR} (**Fig 3.6**). This further confirmed that the hydrogen bond with Tyr 368 in SrpA_{BR} is the reason behind its preferential binding to Neu5Gc, as has been reported before¹⁹⁶.

The above observations explain/confirm the mechanism behind the selective binding of certain adhesin proteins to Neu5Gc but fail to elucidate the mechanism behind the selective binding of Neu5Ac by other adhesin proteins. To explore this further, we compared the crystal structures and performed MD simulations on the adhesin protein GspB_{BR} (from *Streptococcus gordonii* strain M99) which selectively binds Neu5Ac. Hence, a good comparison would be that of GspB_{BR} structures with Neu5Ac and Neu5Gc. However, the crystal structure of GspB_{BR} bound to Neu5Gc does not exist. Therefore, we modified *in-silico* Neu5Ac to Neu5Gc in the existing structure of GspB_{BR} bound to sTa (PDB 5IUC)²⁰⁶. Following this we performed short simulations of this model. We found that the C11-hydroxyl group of Neu5Gc was surrounded by residues (Thr 478, Ile 479, Gly 480, Asp 481, Tyr 482 and Arg 585) when placed similar to the OH group in Hsa_{BR} and SrpA_{BR} crystal structures but it is also in close proximity to a more hydrophobic region (Ile 479 and Thr 478) compared to Hsa_{BR} and SrpA_{BR} (**Fig. 3.7 a,b**). These hydrophobic residues would preferentially interact with hydrophobic C-11 (methyl) of Neu5Ac more than hydrophilic C-11 hydroxyl of Neu5Gc. The overall surface representation of the Hsa_{BR} and SrpA_{BR} was observed to be slightly less hydrophobic and hence more accommodating to the hydrophilic hydroxyl group when compared to GspB_{BR} (**Fig. 3.7c**).

Unlike SrpA_{BR} and Hsa_{BR}, during the course of the simulation GspB_{BR} forms a hydrogen bond with Arg 585 in the Unique domain. Although we were not able to capture the unbinding of Neu5Gc from GspB_{BR} during the course of the simulation, we hypothesize that these extra hydrophobic residues and the hydrogen bond with Arg 585 potentially leads to the unbinding of Neu5Gc, due to the interdomain motion (between Siglec and Unique domains).

3.1.5 Conclusion

Here, we report crystal structures of Hsa_{BR} bound to Neu5Ac-Gal and Neu5Gc-Gal combined with extensive MD simulations. The results provide insights into the mechanism behind the dual specificity of Hsa_{BR} adhesin towards the commonly found sialic acids. We found that Hsa_{BR} can bind to Neu5Gc even in the absence of a specific tyrosine residue (present in SrpA_{BR}), which has previously been considered essential for binding Neu5Gc¹⁹⁶. In line with this, we also observed that unlike SrpA_{BR}, the additional C11-OH group of Neu5Gc does not form a direct interaction with the residues in Hsa_{BR} because the selectivity defining Tyr residue is not present in Hsa_{BR}, and this prevents its preferential binding to Neu5Gc. We also find that Hsa_{BR} does not have the hydrophobic residues (Ile 479 and Thr 478) as in GspB_{BR} which potentially make GspB_{BR} selective towards Neu5Ac. Taken together, these results show that Hsa lacks the features present in the selective adhesins (GspB_{BR} and SrpA_{BR}) and this forms the basis behind the observed dual specificity/binding promiscuity of Hsa towards both the sialic acids. Moreover, this study also provides insight into the selectivity of GspB_{BR} and SrpA_{BR} for Neu5Ac and Neu5Gc, respectively.

Lastly, since we were not able to observe the unbinding of Neu5Gc in the GspB_{BR} simulations and because our simulation was performed starting with a Neu5Gc bound structure, we hypothesize that the C11-OH group of Neu5Gc possibly affects the protein's interaction prior to binding (i.e. in the process of binding). Nevertheless, we acknowledge that there is a possibility that the MD force-field parameters of sugar molecules may not have sampled some interactions accurately. Thus, further work needs to be done to test these hypotheses but are orthogonal to this paper.

3.1.6 Appendix

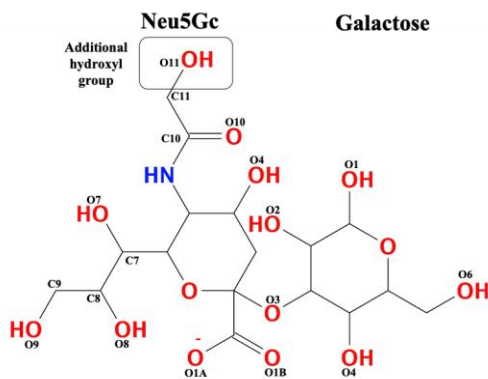


Figure 3.1 Two-dimensional structure of disaccharide: N-glycolylneuraminic acid (Neu5Gc)-Galactose (Gal), showing the important atom names and the additional C11- hydroxyl group, substitution (with hydrogen) of which makes it N-acetylneuraminic acid (Neu5Ac).

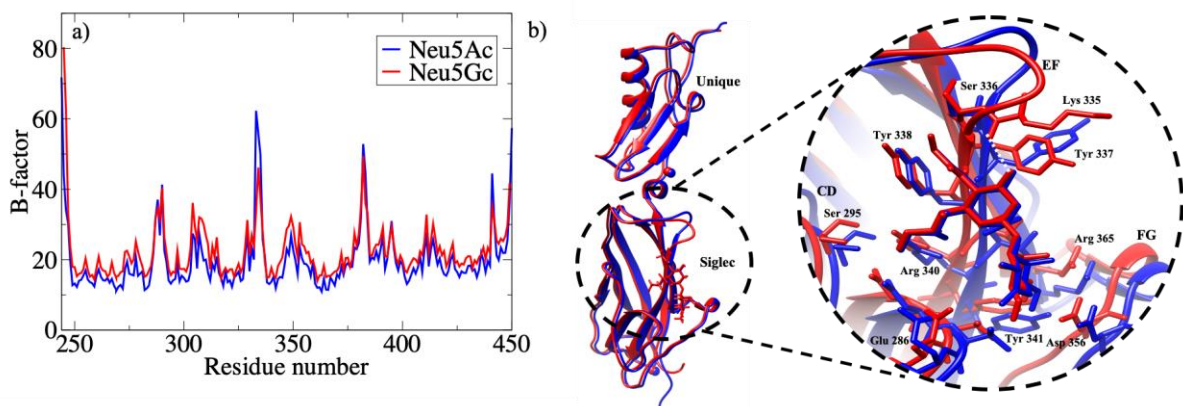


Figure 3.2 a) B-factor from the crystal structures of Hsa_{BR} bound to Neu5Ac-Gal (in blue) and Neu5Gc-Gal (in red); b) the superimposition of the overall two crystal structures showing two domains (Siglec and Unique), and focused view of the binding pocket showing side-chain residues and the hydrogen bonds between backbone carbonyl-O atom in Lys 335 of the EF loop and Neu5Ac-Gal's hydroxyl-O (O4) group in white dotted line.

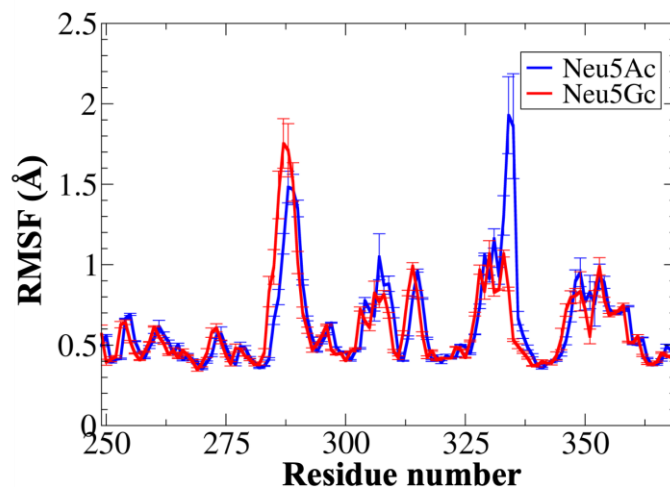


Figure 3.3 Root mean square fluctuation (RMSF) of the Siglec domain of Hsa_{BR} bound to Neu5Ac-Gal (in blue) and Neu5Gc-Gal (in red). The error bars are standard error of mean (SEM) over three independent simulations.

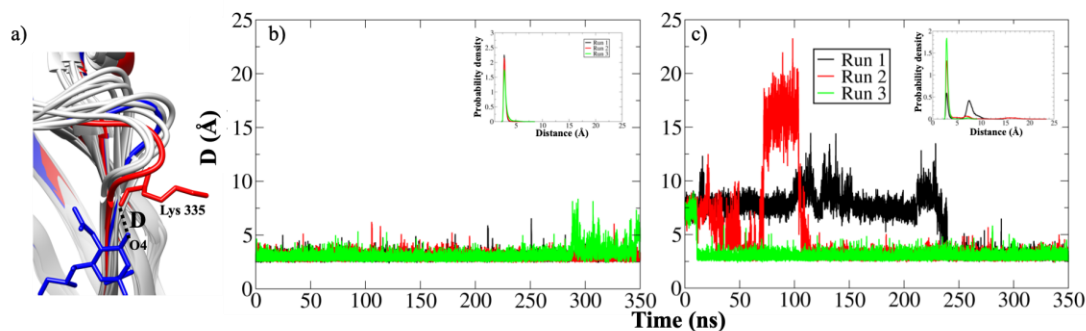


Figure 3.4 (a) Superimposition of the crystal structures of Hsa_{BR} bound Neu5Ac-Gal (in blue), Neu5Gc-Gal (in red), and MD snapshots (in white) with distance (D) between Lys 335 backbone carbonyl and Neu5Ac-O4 atom shown with a black dotted line. Distance (D) over simulation time of three independent runs of Hsa_{BR} bound to (b) Neu5Gc-Gal and (c) Neu5Ac-Gal.

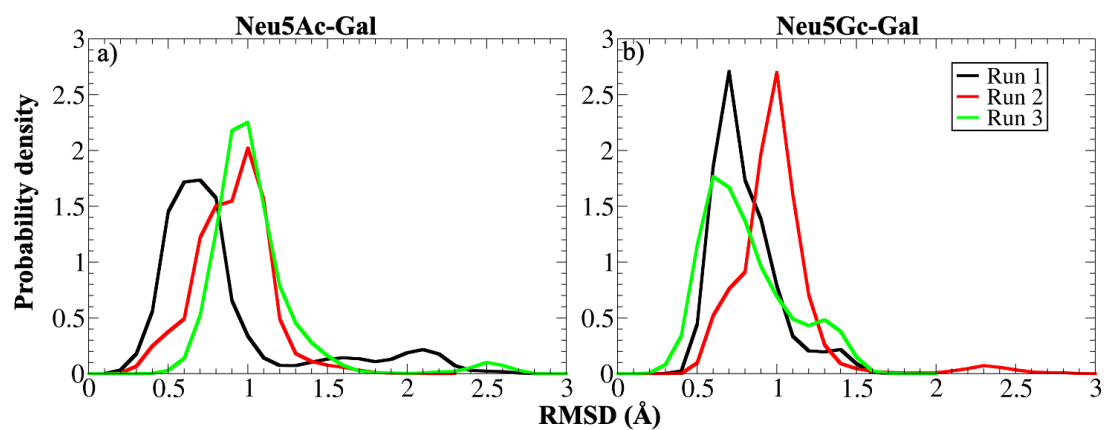


Figure 3.5 Probability distribution of root mean square deviation (RMSD) of disaccharide: Neu5Ac-Gal (a) and Neu5Gc-Gal (b) bound to Hsa_{BR} over three independent simulation runs.

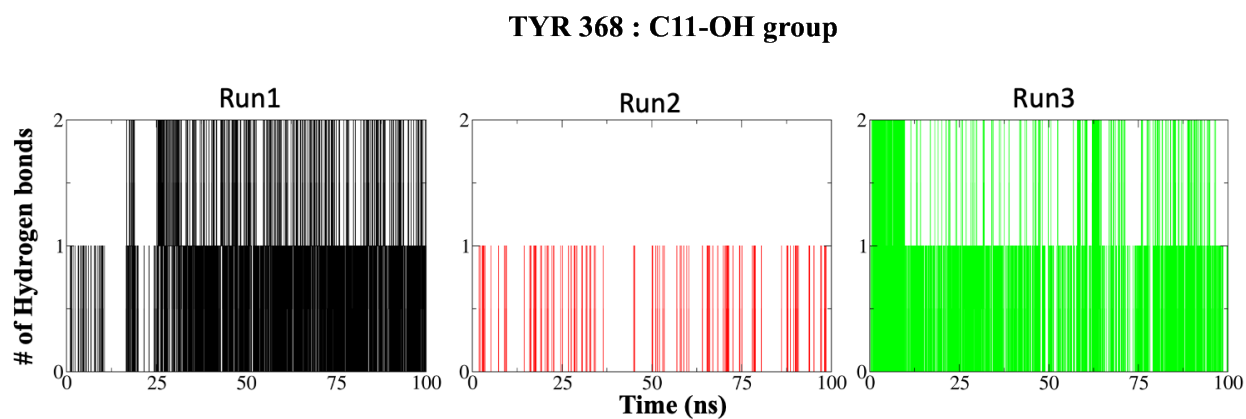


Figure 3.6 Number of hydrogen bonds between Tyr 368 in SrpA_{BR} and C11 hydroxyl group of Neu5Gc.

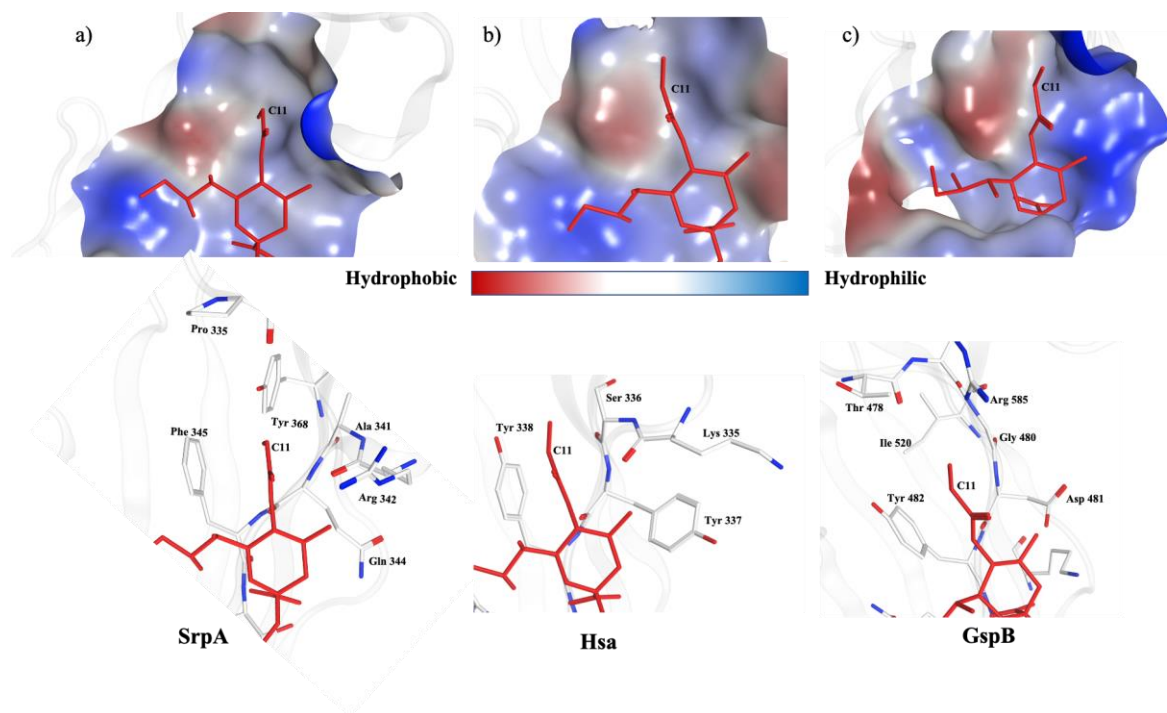


Figure 3.7 Binding pocket of a) SrpA_{BR}; b) Hsa_{BR}; c) GspB_{BR} adhesin proteins. Top panel shows the lipophilicity surface view of the pocket bound to Neu5Gc (red stick); Bottom panel shows the key residues (white stick) in the pockets of the corresponding proteins.

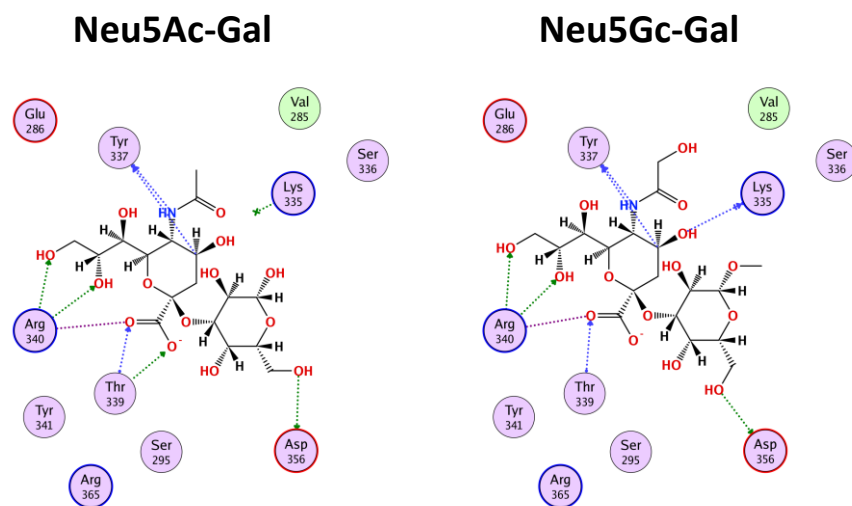


Figure 3.8 Interaction map of Neu5Ac-Gal and Neu5Gc-Gal in Hsa

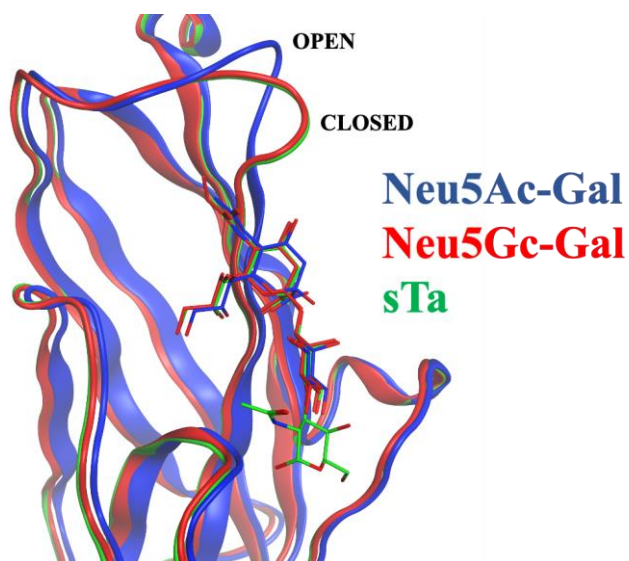


Figure 3.9 Open and closed conformation of Hsa loop bound to different substrate.

3.2 Precise orientation of bacterial adhesin is critical for host adherence

3.2.1 Abstract

Proteins present on bacterial cell surfaces play a major role in facilitating their adherence to the host cell. Previous work has reported that the cellular environment, attachment surfaces, and the selectivity of these adhesin proteins plays a role during surface adhesion. A protein's orientation during binding is also considered important for effective binding but such studies on adhesin proteins are currently lacking. There is thus a need to characterize the role of orientation and positioning of these proteins on the bacterial surfaces to understand their significance in maximizing binding probability to the host cells. Here, we focus on one such group of adhesin proteins, known as Serine Rich Repeat (SRR) adhesin proteins, which are present on the surfaces of *streptococci* and *staphylococci*. These bacteria are notorious for causing infective endocarditis, a potentially fatal cardiovascular infection, and their adherence to the host platelets via adhesin proteins is a critical step during pathogenesis. In our study, we used a combination of biochemical and computational approaches to explore if the orientation of these adhesin proteins is critical for binding to the host platelets and whether changes to the orientation have an influence on the binding. Here, we report that the orientation of adhesin proteins is crucial for binding and slight changes significantly reduces binding affinity/ability. Additionally, our results suggest that although the Unique domain in these SRR adhesin proteins does not affect its binding pocket, it is essential for retaining the binding pocket orientation by maintaining the required rigidity in the inter-domain region.

3.2.2 Introduction

Proteins presented on the surfaces of cells are critical for interactions with the environment surrounding them, including other cells^{207, 208}. This is especially important for disease causing bacteria, who use many different mechanisms and specialized cell-surface proteins to facilitate adherence to host cells²⁰⁸. These attachments are often a prerequisite for colonization and constitute the first stage of numerous microbial infections²⁰⁷. In most cases, bacterial adherence to a eukaryotic cell or the tissue surface requires a receptor and a complimentary ligand. The host receptors are usually some specific carbohydrate or peptide residues presented on their cell surfaces. The bacterial ligands are typically proteins, such as adhesin proteins, which are presented on the bacterial cell surfaces and interact with the host cell receptors. These adherences can range from being non-specific to extremely specific depending on the mode of action. Non-specific adherence involves overall physicochemical properties of the bacterial and host surfaces, such as charge and hydrophobicity; whereas specific adherence involves many high-affinity interactions (similar to a lock and key model) between the complementary receptor on the host cells and the adhesin proteins. Although countless studies have shown the importance of bacterial adhesion in numerous life-threatening diseases, the mechanism of these interactions is still not well understood.

Previous studies have identified three main factors that influence bacterial adhesion to cell surfaces - i) the liquid environment, ii) the attachment surface, and iii) the bacterium itself²⁰⁹⁻²¹². The third factor is usually under the control of the invading bacteria, and extensive work has been done on how bacteria have evolved and adapted to increase the specificity and selectivity of their adhesin proteins^{198, 205, 213}. Apart from specificity, the orientation of the interacting partners is also considered as a deciding factor for any successful binding²¹⁴. Although numerous studies have focused on the (sequences and) structures of these surface proteins, there is still limited information about the precise orientation and positioning of these proteins on the bacterial surfaces and their role in facilitating

binding. This is thus a critical question because bacterial adhesin proteins *ipso facto* evolved to maximize their binding probability to the host receptors.

Hence, to elucidate the role of orientation and positioning of surface proteins on the bacterial surfaces, here we took the case of serine-rich repeat (SRR) adhesin glycoproteins which bind to small carbohydrate molecules present on the human platelets¹⁷⁹. These adhesin proteins form fibril-like protrusions and are present on the surfaces of *streptococci* and *staphylococci*^{205, 215-217}. These bacteria are opportunistic pathogens implicated in multiple human diseases, including a fatal cardiac infection called the Infective endocarditis (IE)^{178, 218}. In IE, the attachment of the bacterial adhesin proteins to host platelets has been shown to be critical for disease onset^{178, 218}, and thus is also a target for disease intervention. Broadly, these adhesins have a conserved architecture with five linearly-arranged modules: 1) N-terminal signal sequence; 2) short serine-rich repeat (SRR1); 3) host binding region (BR) (contains two domains - Siglec and Unique domains); 4) second serine-rich repeat (SRR2); and 5) C-terminal cell-wall anchoring motif^{181, 195, 196, 198, 205, 219}. As the name suggests, the BR region is important for binding to host receptors and based on the size and domain organization of the adhesin protein, it has previously been hypothesized that the extremely long SRR2 domain serves to extend this BR through the bacterial capsular polysaccharide to mediate host cell adhesion. In the BR, the Siglec domain has been well studied as it contains the binding pocket, is responsible for interacting with the host and also known to exhibit a high sequence diversity. The binding pocket of the Siglec domain and its promiscuity in some SRR adhesins has been characterized in an earlier study. Nevertheless, the role of the Unique domain in the BR has still been elusive. However, we found that the adhesin proteins maintain a rigid interdomain (between the respective Siglec and Unique domains) motion within the BR, and that this interdomain torsion/angle is conserved for adhesins within the same clade of their phylogenetic tree. Although, the exact structural role of the Unique domain is not understood, it has been shown to not directly affect the binding pocket present in the Siglec domain but could still be important for facilitating binding.

We thus hypothesized that the interdomain rigidity allows the adhesin protein to maintain a precise orientation on the surface of the bacteria which enables it to form specific interactions with the sialoglycans present on the surfaces of host platelets and that disrupting this rigidity/orientation would abolish their adherence to the platelets (**Fig 3.10**). To this end, we performed molecular dynamics (MD) simulations and biochemical assays to understand the evolutionary pressure behind the conserved rigidity between the Siglec and Unique domains in SRR adhesin BR and further explored if this is important for maintaining a particular orientation of the binding pocket. The results not only suggest that orientation of the binding pocket is important but also elucidate the role of the Unique domain in the sugar-binding, dynamics, and structural integrity.

3.2.3 Methods

System preparation

Crystal structures of the sialoglycan binding proteins Hsa_{BR} (PDB 6EFC)¹⁹⁸, GspB_{BR} (PDB 6EFA)¹⁹⁸, 10712_{BR} (PDB 6EFF)¹⁹⁸, SK150_{BR} (PDB 6EFB)¹⁹⁸, SrpA_{BR} (PDB 5EQ2)¹⁹⁶, and SK678_{BR} (PDB 6EFI)¹⁹⁸ were used in this study. Adhesin protein GspB was chosen for further investigation and MD simulations of GspB protein was performed under four systems: i) SigUn_{wild}: all the residues of Siglec and Unique domain present in the crystal structure; ii) SigUn_{mutALL}: all residues in the Unique domain that formed hydrogen bonds with Siglec domain were mutated to alanine; iii) SigUn_{R585D}: Arginine (R) in position 585 was mutated to Aspartic acid (D); iv) SigUn_{R588D}: Arginine (R) in position 588 was mutated to Aspartic acid (D). Further details are present in **Table 3.1** below.

Simulation details

For MD simulations, each system was solvated in a 10 Å octahedral box of TIP3P²²⁰ water. The Amber16 ff14SB²²¹ force field was used for the protein. In the first step of MD simulations, the backbone and side chains of the protein were restrained using 500 kcal mol⁻¹ Å⁻² harmonic potentials while the system was energy minimized

for 500 steps of steepest descent²²². This step was followed by 500 steps according to the conjugate gradient method⁵⁸. In a second minimization step, restraints on the protein were removed and 1000 steps of steepest descent minimization were performed followed by 1500 steps of conjugate gradient. The system was then subjected to MD and heated to 300 K with the backbone and side chains of the protein restrained using 10 kcal mol⁻¹ Å⁻² harmonic potentials for 1000 steps. The restraints were then released, and 1000 MD steps were performed. The SHAKE²²³ algorithm was used to constrain all bonds involving hydrogen in the simulations. MD runs were performed at 300 K in the NPT ensemble and a 2 fs time step. Accelerated molecular dynamics (aMD) simulations⁷² were performed by using a dual energy boost (i.e. Dihedral energy and Potential energy) with an acceleration parameter (α) of 0.2. The average total potential energy and average dihedral energy parameters for aMD were calculated from classical MD simulations. The details about the duration of the simulations is mentioned in **Table 3.1**.

Analysis details

All analyses were performed on 1 μ s simulations where 3 independent runs for each system were considered. All the analyses were performed using cpptraj and pytraj²⁰⁴ python modules.

The structure and the dynamics of the systems were characterized by Root Mean Square Fluctuation (RMSF). Each of the metrics was calculated from the last 500 ns of simulation trajectory for each system.

Experimental assays

The experimental work in this chapter was done by Dr. Tina Iverson's group in Vanderbilt University and Dr. Barbara Bensing's group at University of California, SF.

Protein expression and purification

Cloning of the gspB BRs in pGEX-3X was performed as described previously in ref ¹⁸³. Additional BRs were identified and then cloned similarly. Briefly, the corresponding DNA coding regions, along with 5' BamHI and 3' EcoRI linkers, were obtained as commercially synthesized products, or were amplified from chromosomal DNA by PCR. After cloning in the pGEX-3X expression vector, the wild-type and mutant BR coding sequences were confirmed by DNA sequence analysis (Sequetech). Cultures of *E. coli* strain BL21 carrying the pGEX expression plasmids were grown in LB with 50 µg/mL carbenicillin until an OD600 of ~0.9, and the expression of GST fusion proteins was induced by the addition of IPTG to a final concentration of 1 mM. Cultures were incubated for 4 h at 24°C. Cells were harvested by centrifugation and lysed by sonication, and the GST fusion proteins were purified using glutathione-sepharose beads according to the manufacturer's instructions. The eluted proteins were exchanged into DPBS and stored at -80°C.

Binding of GST-BRs to platelet monolayers

To assess binding to platelets, fresh human platelets were washed, fixed and immobilized in 96-well plates as described in ref²²⁴. All subsequent binding steps were carried out at room temperature. To reduce non-specific adherence, the wells were treated with 50 µL of 1× Blocking Reagent in DPBS for 1 h. The blocking solution was replaced with 50 µL of purified GST-BRs, ranging from 0.16 to 2500 nM in 1× blocking solution. The plates were incubated for 1 h with vigorous rocking, wells were rinsed three times with 100 µL DPBS, and 50 µL of a rabbit polyclonal anti-GST diluted 1:500 in 1× blocking solution was added to each well. After 1 h, wells were rinsed three times with 100 µL DPBS, and 50 µL of a peroxidase conjugated anti-rabbit antibody (1:5000 dilution in DPBS) was added. After incubation for 1 h, wells were rinsed three times with 100 µL DPBS, and 200 µL of a solution of 0.4 mg mL⁻¹ OPD was added. The absorbance at 450 nm was read after ~30 min, and the values of wells containing the GST-BRs were adjusted by subtracting the average absorbance value of wells containing a GST control.

Binding of biotinylated glycans to immobilized GST-BRs

Purified GST-BRs (500 nM in DPBS) were immobilized in 96-well plates, and the binding of biotinylated glycans was assessed as described previously in ref ¹⁹⁵, with minor modifications. In brief, multivalent biotinylated glycans were added to wells at the indicated concentrations in DPBS containing 1× Blocking Reagent. After 90 min at RT, wells were rinsed three times to remove the unbound glycans, and bound glycans were detected with streptavidin-conjugated horse radish peroxidase, along with a solution of 0.4 mg OPD per mL phosphate citrate buffer. The absorbance at 450 nm was measured after ~20 min.

3.2.4 Results and Discussion

Orientation of binding pocket are critical for platelet binding

The interdomain rigid body motion between the Siglec and Unique domains was first analyzed to understand the potential importance of binding pocket orientation. The interdomain torsion angles (Φ) between the Siglec and Unique domains were calculated for the binding regions (BRs) of six different adhesin proteins (SrpA, NCTC, Hsa, SK678, SK150 and GspB) as shown in **Fig. 3.11**. The results revealed that the interdomain torsion angles (Φ) between the Siglec and Unique domains for these adhesin proteins were grouped into two clusters. Interestingly, this grouping was similar to what has been observed after phylogenetic analyses where proteins from each group here were part of the same clade. Within each group, it was seen that the Φ angles were similar and somewhat conserved- The range of angles in GspB and SK150 adhesin BRs fell between 60°-120° (called the GspB-like proteins), whereas the range of angles in Hsa, SK678, NCTC, and SrpA adhesin BRs fell between 200°-260° (called the Hsa-like proteins) (**Fig. 3.11**). These observations showed that interdomain rigidity is prevalent in all different adhesin proteins and could be important adaptation in the adhesin proteins. Additionally, the fact that the Φ angles are conserved within a group provide impetus to the view that the orientation of the binding pocket and thus its

maintenance potentially by the Unique domain could be significant for subsequent binding.

Therefore, to further understand the importance of binding pocket orientation and the role of Unique domain for sialoglycan binding in these proteins, we used GspB as the test system and modeled two new systems- SigUn_{wild} and SigUn_{mutALL} (SigUn_{mutALL} was generated by alanine mutation of all the residues (R523A, R559A, D562A, R585A, R588A) in the Unique domain that formed hydrogen bonds with the residues in the Siglec domain so as to disrupt the interdomain rigidity). MD simulations were then performed for 500ns on both these systems and the interdomain angles and torsion between the Siglec and Unique domains were calculated as shown in **Fig. 3.12**. We observed that as expected the 2-D histogram of the wild type system (SigUn_{wild}) sampled a tighter range of interdomain angles that were centered around a minimum of 100°. On the other hand, the mutant system (SigUn_{mutALL}) had a comparatively much broader distribution of interdomain angles and torsion to the point that the minimum was shifted to the left and centered around 90°. (**Fig. 3.12**)

These observations suggested that the disruption of the hydrogen bonds between the Unique and Siglec domains had made the mutant system very flexible and that the Siglec domain or the binding pocket could no longer be restricted to a particular orientation. To further focus on the key residues important for facilitating the interdomain rigidity and thus maintaining the binding pocket orientation in the wild type, we generated two more systems: SigUn_{R585D} and SigUn_{R588D}, by mutating residues R585 and R588 from the residues identified in SigUn_{mutALL} to aspartic acid (D) and performed an enhanced sampling MD for 1μs. Both these residues were chosen based on the location and the number of hydrogen bonds. We then calculated the interdomain torsion and angles for both these systems for the last 500 ns of the simulation and compared it with an equivalent simulation of the wild type (SigUn_{wild}) (**Fig. 3.13**). For both the point mutation systems (SigUn_{R585D} and SigUn_{R588D}), we observed a slight shift in the minima with a broader distribution of angles when compared to wild type (SigUn_{wild}). Although this change was less

pronounced compared to SigUn_{mutALL}, these single mutations were enough to cause a shift in the minima (i.e. a change in binding pocket orientation) when compared to wild type. Taken together, these observations confirmed that the Unique domain is responsible for providing a stiffness to the Siglec domain so as to maintain the binding pocket in a particular orientation, which is potentially critical for binding the sialoglycans present on the platelet surfaces.

Based on the above, we hypothesized that a small change/relaxation in this stiffness would modify the binding pocket orientation and prevent/inhibit bacterial adhesion to platelets. Hence to test this hypothesis, we decided to construct the two single point mutants (R585D and R588D) studied above and conduct an ELISA to examine their binding to the platelets. As a positive control, we constructed two other single mutants- L442K and Y485F, where L442 and Y485 are conserved residues in the binding pocket for GspB and are known to be critical for sialoglycan binding.

We observed that both the experimental mutants (R585D and R588D) caused a significant reduction in binding of the adhesin protein to the platelets when compared to wild type (**Fig 3.14**). Interestingly, the reduction in the experimental mutants was either similar or lower than that in the known binding mutants (Y485F and L442K). These results confirmed that the orientation of the adhesin protein binding pocket indeed plays an important role in their binding to the platelets equivalent to that played by conserved residues in the binding pocket and that a slight change in this orientation can significantly reduce the intended binding.

Mutation in the Unique domain has no global or local effect on the structure and dynamics of Siglec domain

Sometimes mutations in one region of the protein can induce allosteric changes in another region which can prevent binding, although this was unlikely because the mutations were made in the Unique domain and were not close to the binding pocket. Nevertheless, to further confirm that the mutations (R585D and R588D) in the Unique domain reduced the binding as a result of change in domain orientation

and not due to an allosteric-like effect in the Siglec domain binding pocket, we decided to explore the local effect of the mutation. To this end, we first calculated the root mean square fluctuation (RMSF) of C-alpha atoms of the Siglec domain for both the wild type and the mutant systems. Similar RMSF trends were observed for all these systems, suggesting that the fluctuations within the Siglec domain were independent or not affected by the mutations in the Unique domain (**Fig. 3.15**).

To further experimentally validate these observations obtained from RMSF, we conducted a fluorescence binding assay after immobilizing the wild type and mutant proteins to explore the effect of these mutations on the binding of known sialoglycan substrates irrespective of binding pocket orientation. Interestingly, we observed that the binding to the known substrate (sTa) remained unaffected in the case of R585D and R588D mutants after they were immobilized (their orientation fixed) and that this signal was similar to that of wild type. On the other hand, both the active site mutants (L442K and Y485F) still had significantly reduced binding to the substrate when compared to wild type. These results confirmed that the point mutations R585D and R588D in the Unique domain did not affect the structure of the binding pocket and hence, the observed reduction in platelet binding was driven by changes in the orientation of the binding pocket.

3.2.5 Conclusion

It is well established that the proteins present on the surface of the cells play a major role in signaling, transport, movement and adhesion to other cells^{207, 225}. A considerable amount of work has been done to understand the conservation of structure, function, and binding selectivity of these surface proteins, especially in bacteria that cause human infections^{207, 225}. Bacterial adhesion to human cells using the bacterial adhesion proteins is a known cause of the pathogenesis in many diseases²²⁶. Although orientation of binding sites is known to be important for any binding interaction, nothing is known about the importance of orientation of the binding sites in these bacterial surface adhesin proteins. Here, we studied a group

of proteins called the SRR adhesin proteins that enable bacterial binding to human platelets during the pathogenesis of infective endocarditis^{178, 218, 227}. We applied a combination of computational and biochemical methodologies to show that indeed the orientation of the adhesin binding pocket is critical for the binding the platelets and a small shift in this binding site orientation is enough to significantly reduce platelet binding. Overall, this study provides an improved understanding of the binding mechanism of adhesin proteins and also identifies novel therapeutic targets disease intervention.

3.2.6 Appendix

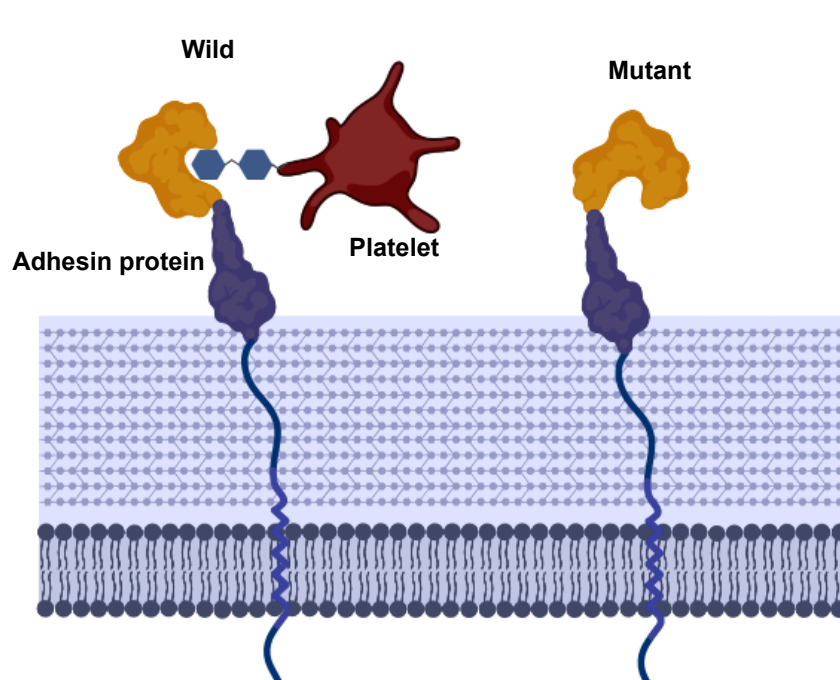


Figure 3.10 Pictorial representation of the hypothesis showing that the binding site orientation is critical for binding to the platelets. Note: The picture is not to scale.

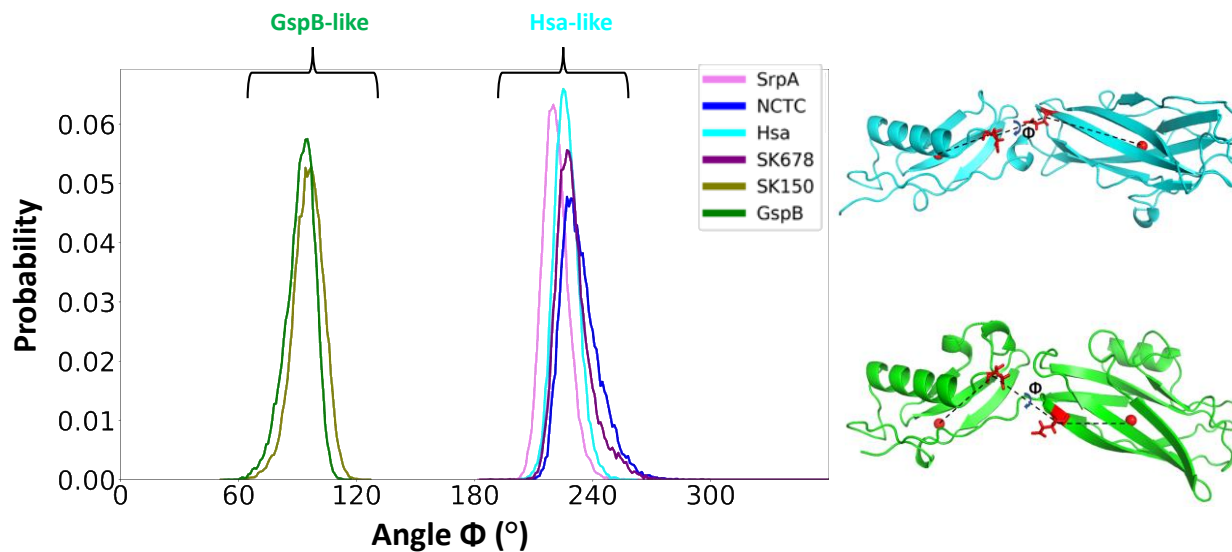


Figure 3.11 Probability distributions of the torsion angle (Φ) between the Siglec and Unique domains of adhesin proteins GspB, SK150, Hsa, NCTC10712, SK678, and SrpA (left) as calculated from MD simulations. Crystal structures showing the Φ angle for both GspB-like and Hsa-like proteins (right). The interdomain torsion angles in the corresponding crystal structures are as follows: GspB: $\sim 100^\circ$; SK150: $\sim 100^\circ$; NCTC10712: $\sim 228^\circ$; Hsa: $\sim 230^\circ$; SrpA: $\sim 216^\circ$; SK678: $\sim 240^\circ$.

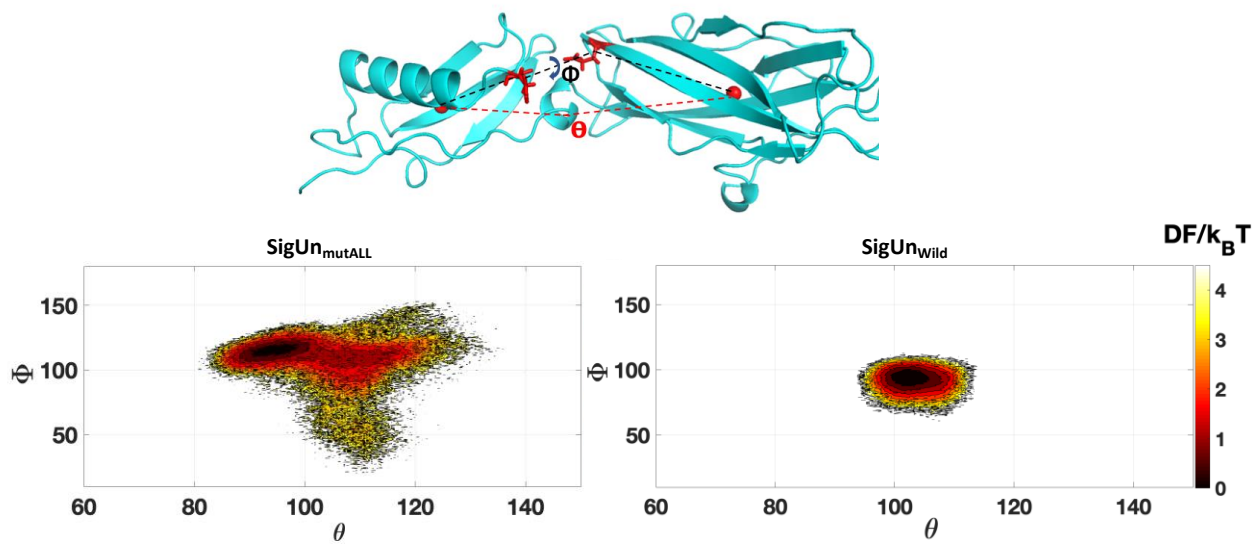


Figure 3.12 Two dimensional histogram of interdomain torsion (Φ) vs angle (θ) between Siglec and Unique for GspB systems: $\text{SigUn}_{\text{Wild}}$ and $\text{SigUn}_{\text{mutALL}}$

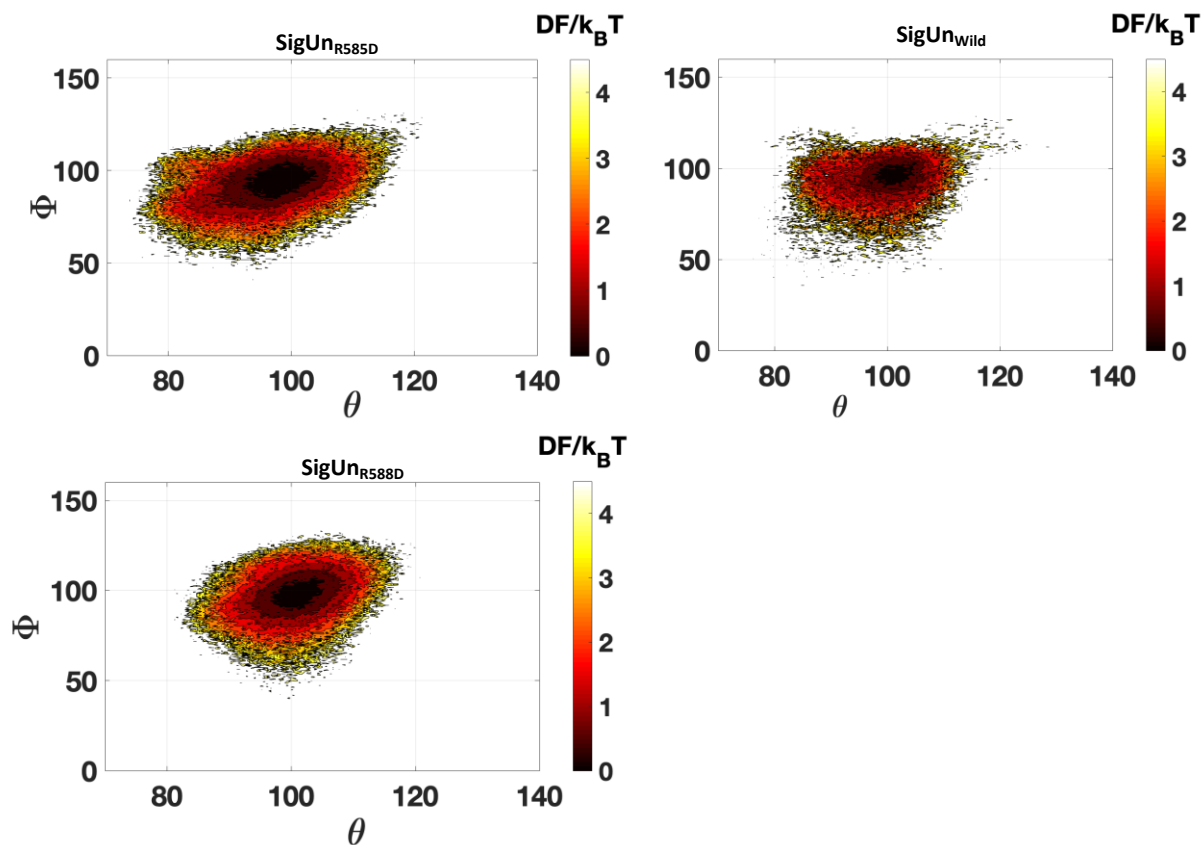


Figure 3.13 Two dimensional histogram of interdomain torsion (Φ) vs angle (θ) between Siglec and Unique for GspB systems: SigUn_{Wild} , SigUn_{R585D} and SigUn_{R588D}

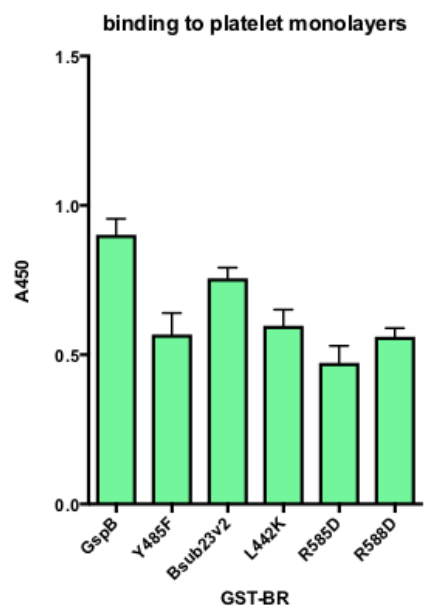


Figure 3.14 Comparative binding of wild-type and mutant BRs of GspB to immobilized platelets

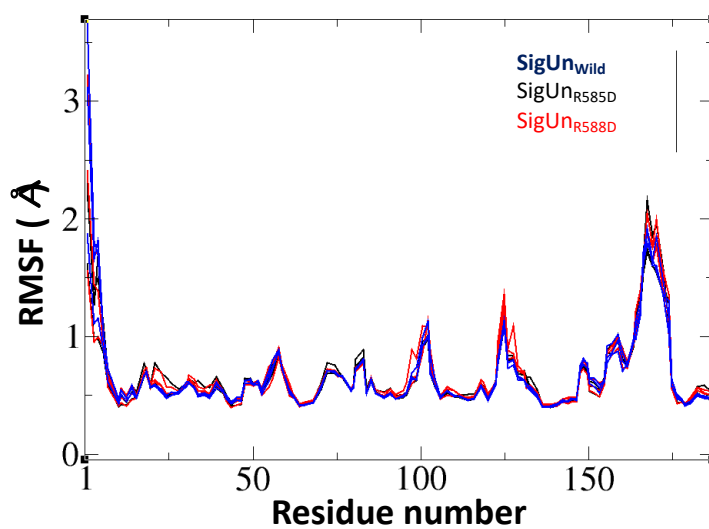


Figure 3.15 Root mean square fluctuation (RMSF) of Siglec domain for GspB systems: SigUn_{Wild}, SigUn_{R585D} and SigUn_{R588D}

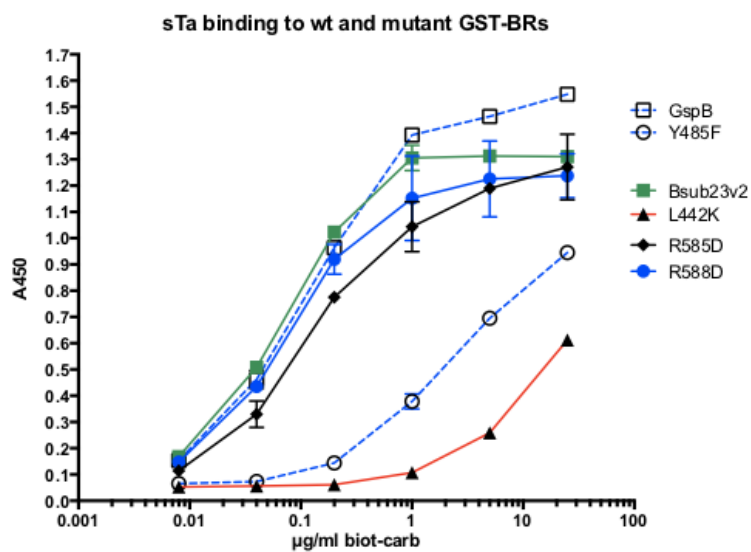


Figure 3.16 a) Binding of biotinylated sTa at different concentrations to immobilized wild-type and mutants R585D, R588D, L442K, Y485F of GspB_{BR}. Binding is reported as the mean $\pm$ standard deviation, with $n = 3$.

Table 3.1 System and simulation details of GspB

System	Residue number	Simulation time	
		Classical MD	Accelerated MD
Siglec & Unique (SigUn)	398-600	500 ns	1 μ s
Siglec & Unique mutant (SigUn _{mutALL})	398-600 (R523A, R559A, D562A, R585A, R588A)	500 ns	-
Siglec & Unique mutant (SigUn _{R585D})	398-600 (R585D)	-	1 μ s
Siglec & Unique mutant (SigUn _{R588D})	398-600 (R588D)	-	1 μ s

4 UNDERSTANDING THE EFFECT OF NON-NATIVE CONDITIONS ON GLOBAL AND LOCAL DYNAMICS OF A PROTEIN

A version of this chapter was originally published by me:

Agarwal, Rupesh, Utsab R. Shrestha, Xiang-Qiang Chu, Loukas Petridis, and Jeremy C. Smith. "Mesophilic pyrophosphatase function at high temperature: a molecular dynamics simulation study." *Biophysical Journal* (2020).

In this work, I executed research, performed analysis, and wrote the paper. Dr. Utsab Shrestha conceived the idea and made substantive contributions to the manuscript's content and Dr. Jeremy C. Smith, Dr. Loukas Petridis and Dr. Xiang-Qiang Chu provided help in results discussion and edited the manuscript.

4.1 Abstract

The mesophilic inorganic pyrophosphatase from *Escherichia coli* (*EcPPase*) retains function at 353 K, the physiological temperature of hyperthermophilic *Thermococcus thio还原ens*, whereas, the homolog protein from the hyperthermophilic organism (*TtPPase*) cannot function at room temperature. To explain this asymmetric behavior, we examined structural and dynamical properties of the two proteins using molecular dynamics simulations. The global flexibility of *TtPPase* is significantly higher than its mesophilic homolog at all tested temperature/pressure conditions. However, at 353 K, *EcPPase* reduces its solvent-exposed surface area and increases subunit compaction while maintaining flexibility in its catalytic pocket. In contrast, *TtPPase* lacks this adaptability and has increased rigidity and reduced protein:water interactions in its catalytic pocket at room temperature, providing a plausible explanation for its inactivity near room temperature.

4.2 Introduction

The enzymatic activity of proteins from hyperthermophilic microorganisms thriving in extreme conditions has been an active area of research for several decades^{228, 229}. These microbes have an optimal temperature range for their growth and survival of about 80°C - 100°C²³⁰. In addition to extreme temperatures, these microorganisms can also withstand high hydrostatic pressures ranging from 60 to

100 MPa (the atmospheric pressure at sea-level is 0.1 MPa)^{228, 231}. Since most mesophilic proteins denature under such high temperatures²³² and pressures²³³, it is intriguing to examine how proteins from hyperthermophilic organisms retain activity. Previous studies have suggested structural characteristics²³⁴⁻²³⁶ of proteins that enable these extremophilic microbes to thrive in severe conditions. However, there appears to be no universal adaptive mechanism, but rather a complex combination of different factors, which frequently differs according to the protein or protein family and is thus difficult to generalize^{235, 237}. Furthermore, the role of dynamic characteristics such as conformational stability^{238, 239} and flexibility²⁴⁰ for protein adaptability is still not well understood^{228, 241}.

Recent studies have reported that the conformational sub-states of a protein are significantly perturbed by changes in temperature and pressure^{6, 240, 242-245}. Temperature enhances the internal fluctuations of a protein¹, and an optimum temperature may provide an appropriate balance of flexibility and rigidity required for function^{240, 246, 247}. Temperatures higher than the optimum can lead to loss of function through unfolding or denaturation²⁴⁸. However, in the case of hyperthermophilic proteins, high native-state flexibility can reduce their entropy of unfolding, thus increasing their melting temperature²⁴⁹. Similarly, high pressure conditions can cause a protein to become inactive by the collapse of its intra-protein cavities, giving rise to an unfolded state²⁵⁰. Pressure drives the reduction in the volume of a protein, which results in a negative entropy change of the system, which may destabilize the native state²⁵¹. Nonetheless, there are exceptions, and the stability and/or activity of some proteins, such as a thermolysin²⁵² and a hydrogenase from *Methanococcus jannaschii*²⁵³, have been shown to increase with pressure.

Overall, a dualistic picture of protein flexibility^{249, 254} and rigidity^{255, 256} has been recognized as a possible factor behind the thermostability of thermophilic and hyperthermophilic proteins^{240, 245}. On the one hand, flexibility is required by a protein to function. On the other hand, flexible residues trigger protein unfolding due to their large thermal fluctuations at high temperature. Hence, rigidifying

flexible residues may be an effective way to improve thermostability^{255, 256}. In addition to flexibility, oligomerization has been reported to be critically important for the stability of some proteins^{257, 258} but the structure-based reasoning behind this stability is not understood²⁵⁹.

We compare here a particularly interesting pair of proteins - the inorganic pyrophosphatase (PPase) from *Thermococcus thio-reducens* (Tt), a hyperthermophilic archaea found near hydrothermal vents of the Mid-Atlantic Ridge²⁶⁰, with a homolog from the mesophilic bacterium, *Escherichia coli* (Ec)²⁶¹. PPase (EC 3.6.1.1) is a homo-hexameric enzyme (~120 kDa), which catalyzes the conversion of pyrophosphate into two phosphate ions. This conversion is important for many critical biochemical processes, such as the production of proteins, nucleic acid polymerization, and lipid metabolism²⁶⁰. Although TtPPase and EcPPase have ~60% sequence similarity and ~40% identity and share similar oligomeric crystal structures (**Fig. 4.6**), the temperatures and pressures for their optimal enzymatic activities are very dissimilar. The catalytic activity of TtPPase has been reported to be maximal at ~353 K but negligible (1.3% of the maximum activity in 10 mins) near room temperature and standard atmospheric pressure (1 bar or 0.1 MPa)²⁶². In contrast, the optimal conditions for EcPPase have been shown to be room temperature (298 K) and standard atmospheric pressure. Interestingly, unlike other mesophilic proteins, EcPPase retains up to 95 to 100% of its enzymatic activity at 353 K for about 10 minutes, after which the activity slowly decays²⁶³.

Most crystal structures of thermophilic proteins have been resolved at room temperature instead of at their native temperature or pressure conditions. Therefore, to obtain a comprehensive picture of protein structure, dynamics, and function, static structure determination (using X-ray and neutron crystallography) must be complemented with dynamic information using various spectroscopic techniques (e.g., neutron scattering^{240, 246} or NMR^{264, 265}) and computer simulations^{49, 266} under varying external conditions. Molecular dynamics (MD) simulation, which provides spatial and temporal information at an atomic

resolution²⁶⁷ has thus become one of the most powerful methods to explore the protein energy landscapes and their flexibility-function relationships^{52, 244, 245} at near-native conditions and is the technique applied here.

Here, we perform MD simulations to understand the effects of pressure and temperature on the structural and dynamic behavior of the two PPase homologs in their native and non-native environments. We find that *Tt*PPase possesses higher global flexibility at both native and non-native conditions than its mesophilic homolog (*Ec*PPase). However, this effect is not reflected locally in the catalytic pocket. Additionally, we determined that factors accompanying enzymatic activity of PPase are the number of hydrogen bonds and water molecules in its catalytic pocket. Furthermore, we provide potential factors behind the observed enzyme activity and/or adaptability of *Ec*PPase at high temperature and alternatively, the inability of *Tt*PPase to do so at low temperature.

4.3 Methods

All-atom MD simulations of monomeric and hexameric *E. coli* PPase (*Ec*PPase) and *Thermococcus thio-reducens* PPase (*Tt*PPase) were carried using GROMACS (version 2016.3) suite^{145, 268}. Analysis of the MD trajectories was performed using python package pytraj²⁰⁴ and GROMACS analysis tools. Hexameric *Ec*PPase (PDB ID 1I6T²⁶¹) and *Tt*PPase (PDB ID 3Q5V²⁶⁰) were solvated in a TIP3P cubic box of water (~150,000 atoms). The crystal structures for both the proteins were resolved at the same conditions (298 K and 0.1 MPa) and are also in the same space group (H32). The monomeric structures for both the proteins were modeled by taking only one chain from the hexameric structures. These monomeric models were then solvated in a TIP3P cubic box (~36,000 atoms).

Simulation details

All simulations were performed using the AMBER ff99SB potential²⁶⁸. This combination of AMBER ff99SB force-field and TIP3P water model has previously been shown to work well, reproducing experimental results at different temperature

and pressure conditions^{4, 269-271}. In ref⁴, the adaptation of extremophilic protein (Initiation Factor 6) at high temperatures (300 - 350 K) and pressures (0.1 – 100 MPa) agreed with experimental neutron scattering data. In ref²⁶⁹, the folding thermodynamics and structure of Trp-cage in solution revealed that the stability of peptide is more realistic in the simulation using TIP3P rather than using TIP4P-Ew when compared to the calorimetry and circular dichroism (CD) spectroscopy experiments²⁶⁹. The calculated partial molar enthalpy change and partial molar volume change values of unfolding of chignolin reasonably agree well with the results from CD spectroscopy and NMR experiments²⁷⁰. Hata et al. 2020 studied the pressure-induced dissociation of the CheY-FliM complex and found an increased density of water in the first hydration shell²⁷¹. In each of the above studies the force-field and water model used was the same as in the present work. Each protein was simulated at four different conditions: two temperatures: 298 K (ambient for *Ec*PPase) and 353 K (ambient for *Tt*PPase); and two pressures: 0.1 MPa (ambient for *Ec*PPase) and 100 MPa (ambient for *Tt*PPase). All systems were prepared in a three-step process: initial energy minimization, NVT equilibration and a NPT production run. Energy minimization was performed with the steepest-descent algorithm^{145, 268} to a tolerance of 1000 kJ/ (mol·nm). NVT equilibration was performed for 1 ns at each temperature fixed with the V-Rescale thermostat¹⁵⁸. Following this step, the NPT ensemble protocol was followed to generate production runs (500 nanoseconds for the hexamer and 1 microsecond for the monomers) with the V-Rescale thermostat¹⁵⁸ used to maintain the required temperature and Parrinello-Rahman¹⁵⁹ to maintain the pressure at 0.1 MPa or 100 MPa. Since the compressibility of water changes drastically with pressure, the compressibility of water was taken to be $4.5 \times 10^{-5} \text{ bar}^{-1}$ at 0.1 MPa and $3.5 \times 10^{-5} \text{ bar}^{-1}$ at 100 MPa, based on previous experimental work²⁷². Particle Mesh Ewald²⁷³ was used for long-range electrostatics with a short-range electrostatics and van der Waals cutoff of 1 nm. Three independent runs were performed for each simulation (**Table 4.2**). The simulations were considered converged when

the fluctuations in the root mean square deviation of C-alpha atoms ($\text{RMSD}_{\text{C}\alpha}$) reached a plateau with time.

Analysis details

Each of the metrics was calculated from the last 300 ns of each simulation trajectory. The average value of each metric from 3 independent trajectories starting with different velocities has been reported, where the standard error of the mean has been used as the error bar.

The crystal structure RMSD comparison was performed using the SuperPose webserver²⁷⁴. The number of hydrogen bonds was computed using the “hbond” utility in GROMACS using a donor–acceptor cutoff distance of 0.32 nm and a cutoff angle of 20°. The solvent accessible surface area (SASA) and volume were calculated using a 0.14 nm probe size for the whole protein. The volume of buried cavities was determined using the “trj_cavity” module²⁷⁵ in GROMACS. The volume of the completely buried cavities (CBC) was calculated with a 1.4 Å grid spacing (-spacing) and degree of buriedness of 6 Å. The number of water molecules in the cavities was estimated by dividing the total cavity volume by the volume of a buried water molecule near the protein surface (2.29 nm³) as reported in previous work²⁷⁶. Intra and interchain hydrogen bonds were calculated using the pytraj python package²⁰⁴. Inter chain hydrogen bonds were normalized by dividing it by the number of chains (n=6).

4.4 Results

Structural flexibility

The structural flexibility of proteins is associated with many biological functions, such as catalytic activity, substrate binding, and molecular recognition^{1, 277}. To understand the effect of temperature and pressure on the protein global flexibility, we calculated the average mean-square displacement (MSD) of the C α atoms. The MSD is often used to compare the stability of thermophilic and mesophilic protein homologs^{244, 245}. Overall, the MSD results show that *Tt*PPase has higher

flexibility compared to *EcPPase* at all the conditions studied here (**Fig. 4.1**). This is, therefore, an example of a thermophilic protein for which increased rigidity is not associated with thermostability.

For both the homologs, a two-fold increase in MSD is observed at 353 K relative to 298 K. Unlike temperature, the effect of pressure was less pronounced, an increase in pressure causing a small reduction in MSD for *TtPPase*, and no change for *EcPPase*. Interestingly, the MSD values for both the monomers showed an effect of pressure at 353 K, contrary to the respective hexameric forms (**Fig. 4.1 and 4.9**). These results suggest that oligomerization may confer stability to increased pressure at high temperature.

Flexibility is often correlated with activity or stability of mesophilic and thermophilic homologs^{1, 277}. However, although at 298 K *TtPPase* is more flexible than *EcPPase*, the activity of *TtPPase* is negligible near room temperature²⁶². Further, even though the flexibility of *EcPPase* at 298 K is smaller than that at 353 K, *EcPPase* is catalytically active at both temperatures. Hence, the global flexibility of the PPase does not follow the observed functional properties.

Compactness

Compactness is another common adaptive mechanism of thermophilic proteins^{259, 278}. To characterize the compactness of the hexameric structures over the course of the simulations, we calculated the solvent accessible surface area (SASA) of the proteins, which is inversely proportional to the number of native contacts²⁷⁹. The density plots of the SASA show that *EcPPase* has larger solvent accessibility than *TtPPase*, irrespective of the temperature and pressure conditions studied here (**Fig. 4.2a and 4.5b**). A similar difference is found in the hexameric crystal structures (*TtPPase*: 374 nm² and *EcPPase*: 393 nm²; **Table 4.3**). However, the solvent accessibility of *EcPPase* is reduced at 353 K compared to 298 K, whereas that of *TtPPase* remains unaffected by the change in temperature. Pressure did not significantly affect the SASA of either homolog. The SASA probability distribution for monomeric structures showed a similar change in compactness, as

observed in the hexameric form (**Fig 4.8a**). To further quantify compactness, we calculated the number of intra- and inter-chain hydrogen bonds (HBs) of the hexameric form. The normalized mean number of inter-chain HBs was nearly double for *Tt*PPase than for *Ec*PPase and remained constant at all conditions (**Fig. 4.2b**), confirming the closer packing of the subunits in *Tt*PPase than in *Ec*PPase. The mean number of intra-chain HBs did not change greatly between the two homologs (**Fig. 4.5a**) at all pressure and temperature conditions. These results suggest that the higher compactness of the *Tt*PPase hexamer is due to compactness between the subunits and also within each monomer.

Protein cavities

Intra-protein cavities have been recognized to be important for the stability and function of proteins^{280, 281}, and water molecules buried within these cavities have been reported to be influential in temperature and pressure-mediated unfolding^{250, 282}. We calculated the number of water molecules enclosed in the completely buried cavities of both proteins. This was found to be significantly higher for *Ec*PPase than *Tt*PPase (**Fig. 4.3**) and not significantly affected by temperature and pressure. These results are consistent with the inter-chain HB results discussed earlier and with calculations that show that the crystal structure of *Tt*PPase has smaller buried cavity volume (303 water molecules) than *Ec*PPase (427 water molecules) (**Table 4.3**).

Flexibility in the catalytic pocket and its interaction with water

The precise 3D preorganization of residues and their interaction with water molecules in an enzyme's catalytic pocket determines the function²⁸³⁻²⁸⁵. Previous studies have identified conserved residues in the catalytic pocket of PPase enzymes (**Table 4.1 and Figs. 4.6a and 4.6b**) that are critical for catalysis and are also required for coordinating the divalent ions^{286, 287}. Moreover, experimental studies based on several crystal structures have shown that the hydrogen bond network in the binding pocket is critical for hydrolysis and the catalysis requires precise preorganization^{288, 289}. The catalytically inactive *Tt*PPase at 298 K has

~60% more HBs (~15 HBs, i.e., 2.5 per monomer) between the catalytic residues than the catalytically-active forms of PPase (*Tt*PPase at 353 K, and *Ec*PPase at 298 and 353 K) (**Figs. 4.4 a, c**). These results, for the hexameric form were also found in the simulations of the monomeric form (**Fig. 4.8b**). We did not observe any effect of pressure on the number of HBs between the residues in the catalytic pocket (**Figs. 4.4 a, c**).

Water molecules in the catalytic pockets can be critical for enzymatic function^{290, 291, 292}. Interestingly, the catalytic pockets of both homologs have ~110 protein:water HBs at 353 K, consistent with the observed catalytic activity of *Ec*PPase at this elevated temperature²⁶³. However, at 298 K, *Ec*PPase significantly increases the number of protein:water HBs, to around ~160 HBs, whereas *Tt*PPase does this to a much lesser degree, to ~130 HBs (**Figs. 4.4 b, d**). We also quantified the flexibility and solvent exposure of the catalytic pocket by calculating the root mean square fluctuation (RMSF) of these conserved residues. The RMSF of these residues in *Ec*PPase varies weakly with temperature, whereas *Tt*PPase has significantly lower fluctuations at 298 K compared to 353 K (**Fig. 4.7**). *Ec*PPase therefore maintains its local flexibility of the catalytic pocket at both temperatures, whereas *Tt*PPase becomes rigid at the lower temperature.

4.5 Discussion and Conclusions

In this work, we compare structural and dynamic properties of hyperthermophilic and mesophilic PPases using MD simulations mimicking deep-sea and ambient conditions. The results indicate that *Tt*PPase has been designed to function at high temperatures with a more compact structure and a reduced number of intra-protein hydrogen bonds in the catalytic pocket. However, *Tt*PPase does not maintain both of these properties at room temperature, where it cannot catalyze the enzymatic reaction. Interestingly, we found *Ec*PPase adapts and retains its activity at high temperature by incorporating similar strategies used by *Tt*PPase to function at high temperature: maintaining hydrogen bonds in the catalytic pocket and increasing its compactness.

The overall structural flexibility of a protein is sometimes assumed to be relevant for its enzymatic activity¹ and the MSD is often used to quantify this. A general framework for understanding the stability and function of hyperthermophilic proteins in their native conditions has been proposed based upon the hypothesis that enhanced rigidity underlies increased thermal stability²⁴⁹. However, other experimental and computational studies have reported that hyperthermophilic proteins have larger conformational flexibility than their mesophilic homologs^{245, 293}. Here, we also observe higher overall flexibility of *Tt*PPase than *Ec*PPase at all temperature/pressure conditions. Indeed, even at room temperature, where *Tt*PPase is enzymatically inactive, it is more flexible than *Ec*PPase. Thus, differences in the overall flexibility of PPase are not directly associated with differences in its enzymatic activity.

Although the crystal structures of both homologs were not resolved at their native conditions, the RMSD between them is small (0.18 nm), and both homologs have similar radii of gyration (2.9 nm). However, the buried cavity size and solvent accessibility of the crystal structure of *Ec*PPase are greater than for *Tt*PPase. Likewise, from the simulations, the solvent-accessible surface area and number of inter-chain HBs (both of which are related to compactness) of *Tt*PPase indicate that it is more compact than *Ec*PPase at all the conditions investigated here. The main contributing factors to the compactness of *Tt*PPase are inter-subunit compaction, as shown by increased inter-chain HBs, together with the reduced solvent-exposed surface area of each monomer as demonstrated by the simulations of the monomeric forms. These results are not surprising since compactness is a known adaptive factor in common between many thermophilic proteins.

Moreover, a decrease in the number of water molecules in the buried cavities of *Tt*PPase compared to *Ec*PPase was observed at all conditions. These results are consistent with previous work²⁹⁴ which showed that proteins that are active at extremely low temperatures (psychrophilic proteins) have a comparatively larger average cavity size and the properties of water buried in the protein cavities of a

hyperthermophilic protein significantly differ from its mesophilic homolog ^{295, 296}. Similarly, a reduction in buried cavity volume has been reported as an adaptation (thermostability) mechanism at high-temperature conditions because large water-filled cavities are known to be a driving factor for protein denaturation at high temperature and pressure ²⁹⁷.

Interestingly, *EcPPase* exhibited a change in its compactness with a change in temperature, becoming more compact at a higher temperature than at its native temperature. In contrast, the compactness of *TtPPase* remains the same at room and high temperatures. Additionally, we also studied the structure and dynamics of the catalytic pocket. Here, we observed an increase in rigidity at 298 K compared to 353 K, as shown by the lower RMSF of catalytic pocket residues and a greatly increased number of HBs between the catalytic pocket residues. In contrast, we observed similar catalytic pocket residue HBs in *EcPPase* as for *TtPPase* at 353 K, and this number remained unaltered upon a change in temperature. This local flexibility (based upon the number of HBs and RMSF) of the catalytic pocket agrees with the existing “corresponding states” hypothesis, according to which a thermophilic protein has a more rigid catalytic pocket than its mesophilic homolog at room temperature²⁹⁸. However, both of these homologs should exhibit similar flexibility at their respective functional temperatures²⁹⁸. Furthermore, we see that the trend observed in global flexibility measured by the MSD is not reflected locally in the catalytic pocket (i.e., there is no direct correlation between local and global motions).

The above picture is further supported by calculations of the number of HBs of the catalytic pocket residues with water molecules. In the case of *EcPPase*, we observe a dramatic decrease in the number of HBs with water at 353 K when compared to 298 K, whereas there is a negligible change for *TtPPase*. This result is consistent with the compactness results showing that *EcPPase* has the capability to adapt to a high temperature, whereas *TtPPase* lacks the ability to adapt to a low temperature environment.

Based on these results, we suggest that at lower temperatures, both of these homologs would need comparatively more water molecules in their catalytic pockets and a larger exposed surface area to function and that the opposite is needed at higher temperatures. *EcPPase* is able to make the aforementioned changes to its structure and hence is able to adapt to high temperature/pressure, but *TtPPase* fails to do so at low temperature. Although the overall structure of *TtPPase* barely responds to change in temperature/pressure, the catalytic pocket becomes more rigid at 298 K than at 353 K due to an increase in the number of intra-protein HBs. This rigidity may explain its enzymatic inactivity at room temperature. In contrast, *EcPPase* adapts to a high temperature by reducing solvent-exposed surface and adopting a more compact oligomeric structure. In addition, *EcPPase* preserves the number of hydrogen bonds within the residues in the catalytic pocket at all the conditions and reduces its interactions with water molecules at a higher temperature. This provides a possible explanation behind its activity even at a high temperature. Nevertheless, the current work does not provide information on the rate-determining step (i.e., hydrolysis of pyrophosphate²⁸⁸) but focuses on the local environment and the preorganization of the catalytic pocket prior to hydrolysis which has been known to be important for activity^{288, 289}. We observed an effect of pressure on the MSD of monomeric form, but not for hexameric form. However, the SASA and the catalytic pocket behavior of the monomeric and hexameric forms are relatively unaffected. Consequently, our work sheds little light on the effect of pressure on oligomerization in the case of PPase. Overall, our results suggest the change in temperature significantly affects the conformational properties of PPase homologs, whereas the effect of pressure is very subtle or insignificant in vitro. However, pressure can have a significant impact on PPase in a cellular environment. Indeed, previous studies have shown that the archaeon *Thermococcus barophilus* accumulates osmolytes inside the cell to adapt and function under varying stress (pressure and temperature) conditions²⁹⁹.

300

Finally, our results provide a structural and dynamic basis for the experimentally observed activity of *EcPPase* at both high and room temperature. Such an intriguing ability of adaptation of mesophilic PPase could be possibly explained by existing evolutionary theory since hyperthermophilic archaea are thought to be a universal ancestor³⁰¹. Indeed, recent work³⁰² has suggested that thermophilic proteins (isolated from thermophilic archaea *Pyrococcus furiosus*) were “*de novo*” designed in a hot environment and then used a “structure-based mechanism” to adapt to a mesophilic environment on recolonization. In the structure-based mechanism adaptation to high temperatures is through compaction, which here is mostly inter-subunit, with the monomer structures relatively unchanged. Phylogenetic analyses³⁰³ show that *T. thio还原ens* (also hyperthermophilic archaea) shares the same clade as *P. furiosus*; hence, PPase homologs studied here may have evolved according to the structure-based adaptation theory. Thus, *EcPPase* may have been evolved from its ancestral counterpart *TtPPase* to recolonize at room temperature and may, therefore, be able to adapt when introduced to a higher temperature. Moreover, since thermophiles do not live at room temperature, there is no selective need for high activity at room temperature.

4.6 Appendix

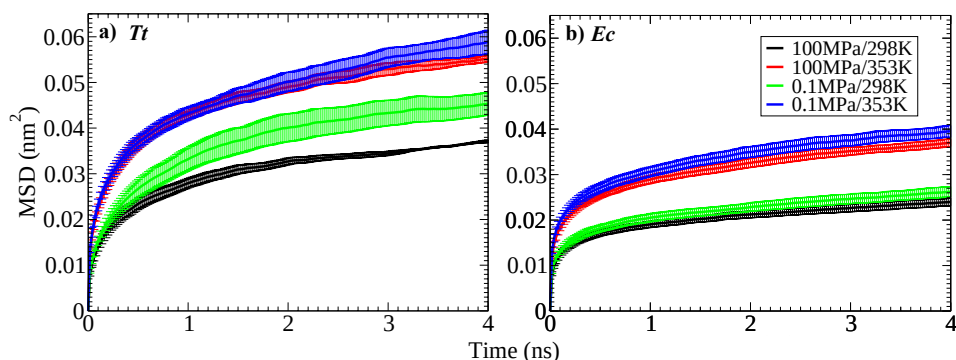


Figure 4.1 Mean square displacement (MSD) of C-alpha atoms for (a) hexameric TtPPase and (b) hexameric EcPPase. Higher values of MSD for TtPPase at all conditions indicate higher structural flexibility than EcPPase. Error bars shown in this and the subsequent figures are the standard error of the mean, and those not immediately visible are at most the size of the symbol.

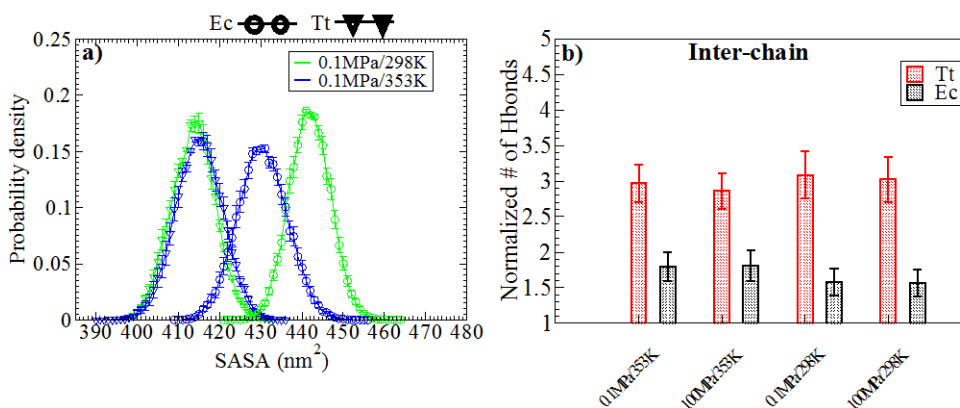


Figure 4.2 a) Probability distribution for solvent accessible surface area (SASA) of the hexameric forms for TtPPase (triangle) and EcPPase (circle); b) The normalized number of interchain hydrogen bonds.

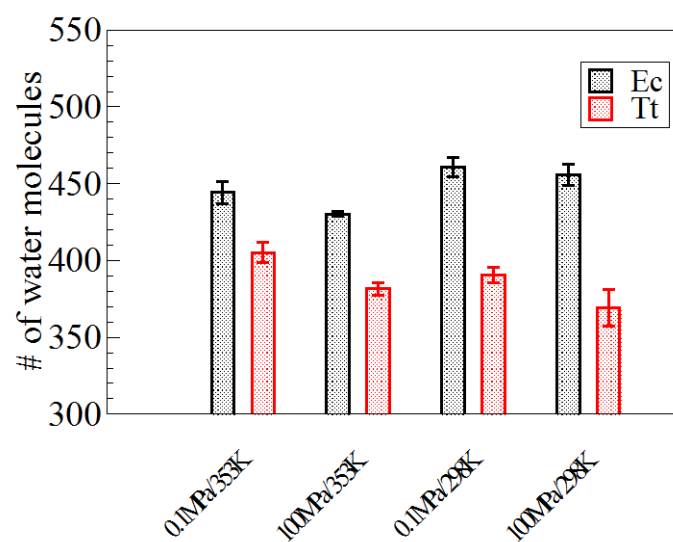


Figure 4.3 Number of water molecules in completely buried cavities of hexameric form at different temperature and pressure conditions. TtPPase (black shaded bars) vs. EcPPase (red shaded bars).

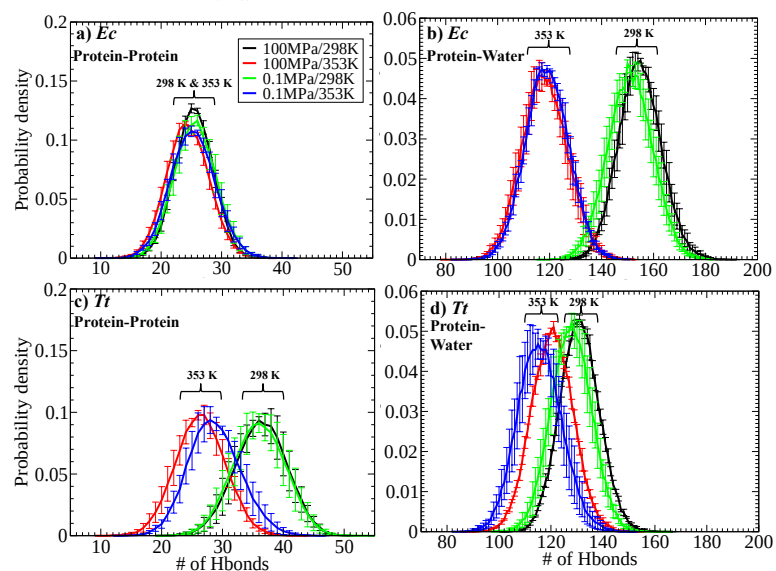


Figure 4.4 Probability distributions of the number of hydrogen bonds between the residues in the catalytic pocket: hexameric EcPPase (a) and hexameric TtPPase (c). Probability distributions of the number of hydrogen bonds between the catalytic pocket residues and water molecules in the catalytic pocket: hexameric EcPPase (b) and hexameric TtPPase (b).

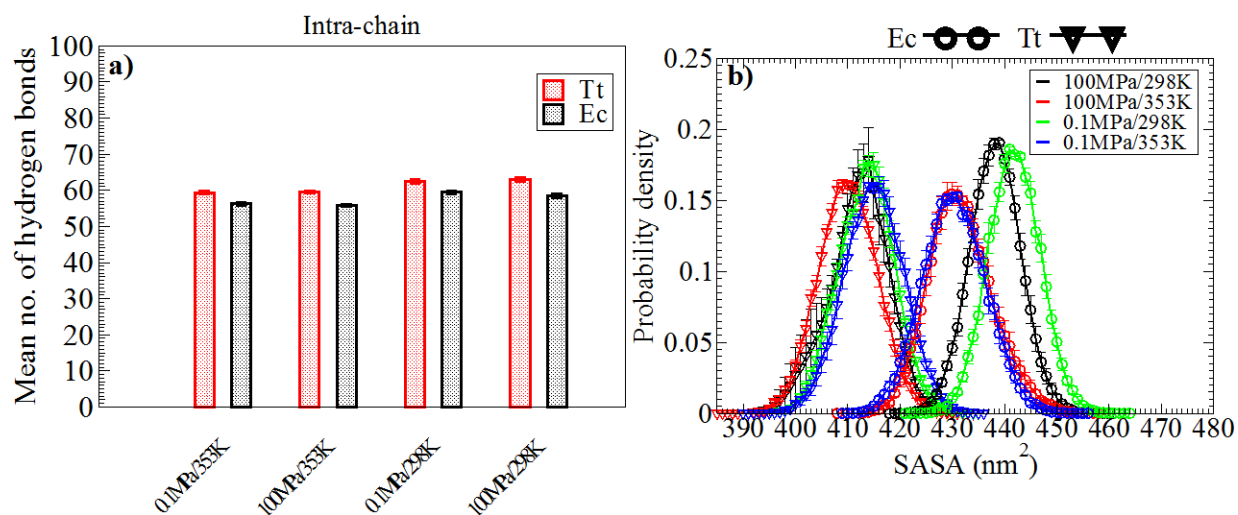


Figure 4.5 Mean number of hydrogen bonds within each chain (intrachain) for hexameric *Tt*PPase and hexameric *Ec*PPase at different conditions; b) Probability distribution for solvent accessible surface area (SASA) of the hexameric forms for *Tt*PPase (triangle) and *Ec*PPase (circle); Error bars show the standard error of the mean (SEM), and those not immediately visible are the same size of the symbol. The cutoff distance of 0.32 nm and a cutoff angle of 20° was considered for hydrogen bond analyses.

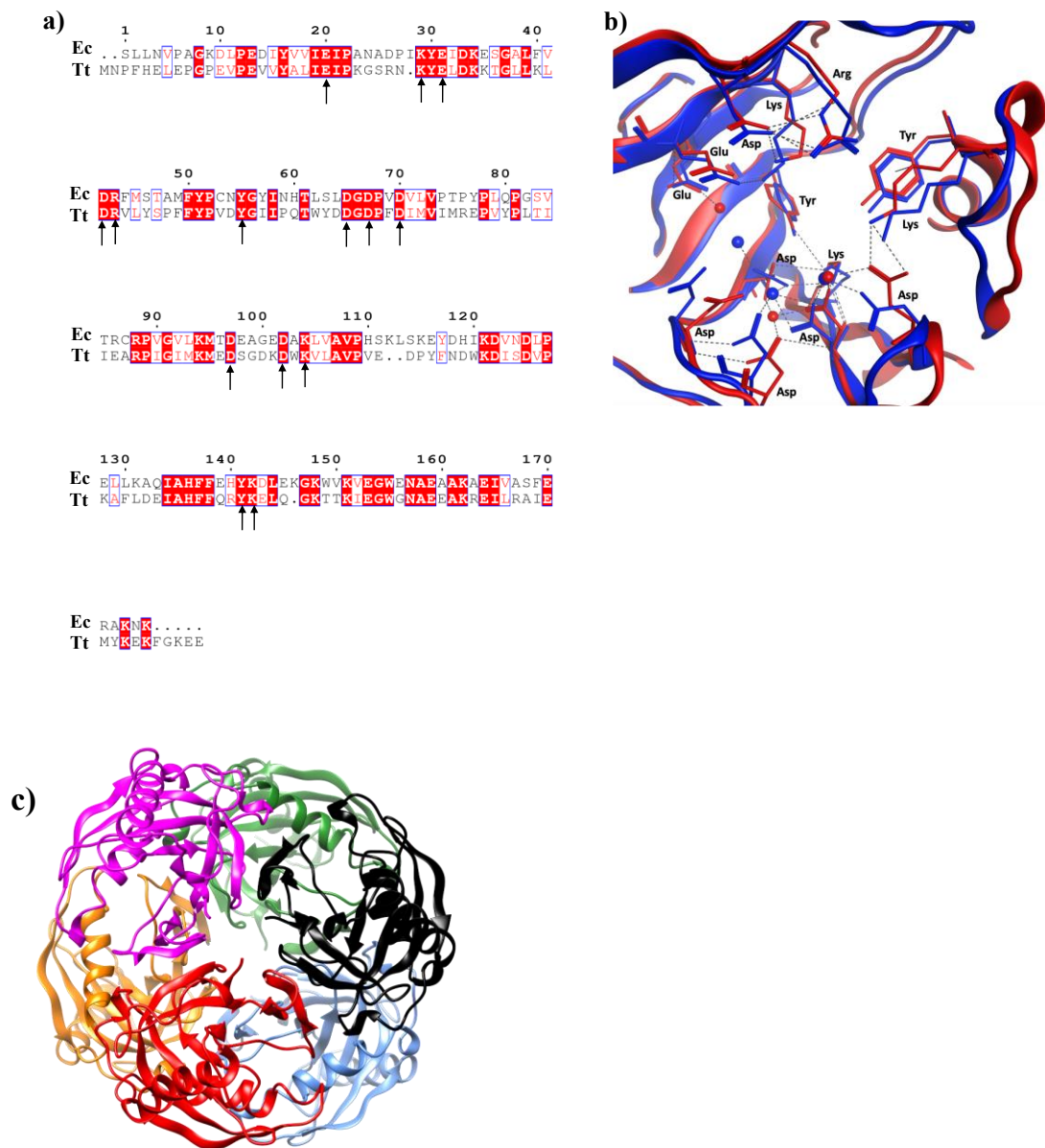


Figure 4.6 a) Sequence alignment of *Ec*PPase (*Ec*) and *Tt*PPase (*Tt*) showing identical residues (in red box), similar residues (in white box) and conserved catalytic pocket residues (black arrow); b) Catalytic pocket of *Ec*PPase (red) and *Tt*PPase (blue) showing conserved critical residues (as sticks) and hydrogen bonds (black dotted line); c) *Tt*PPase hexameric crystal structure with each subunit colored in different color.

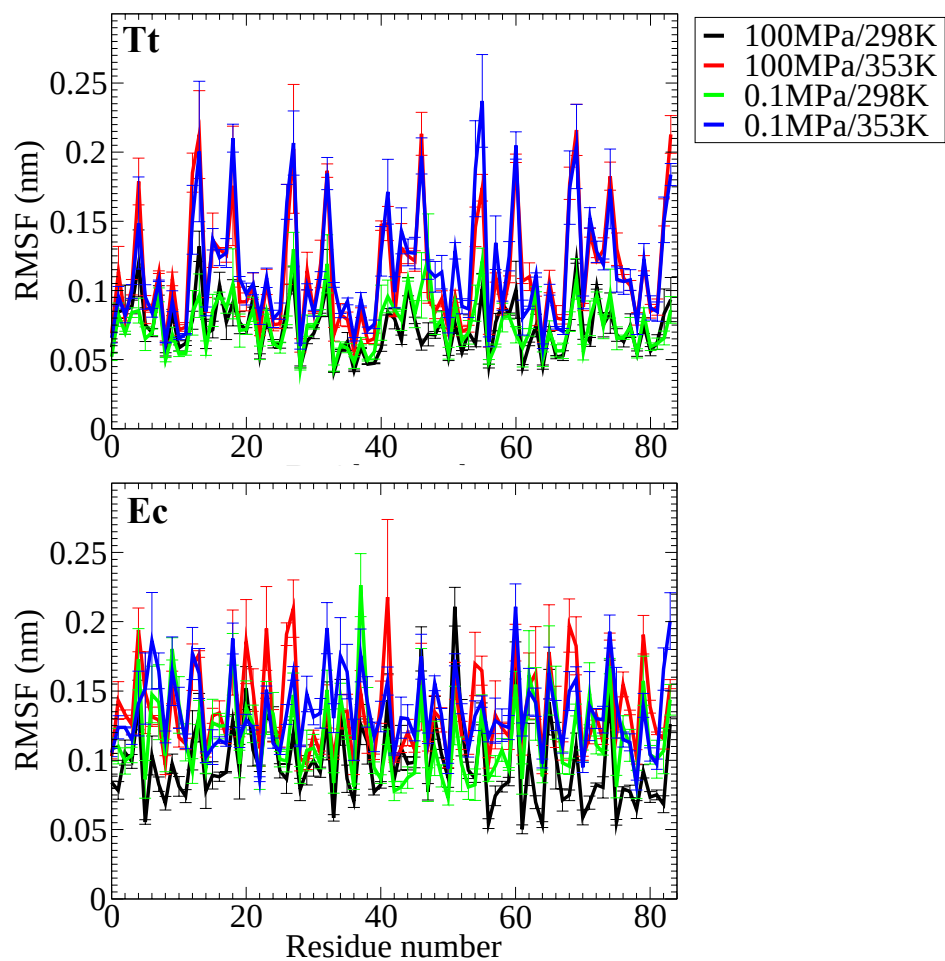


Figure 4.7 Root mean square fluctuation (RMSF) of all atoms of catalytic pocket residues (averaged over per residue) for hexameric *Tt*PPase (top) and hexameric *Ec*PPase (bottom) at different conditions.

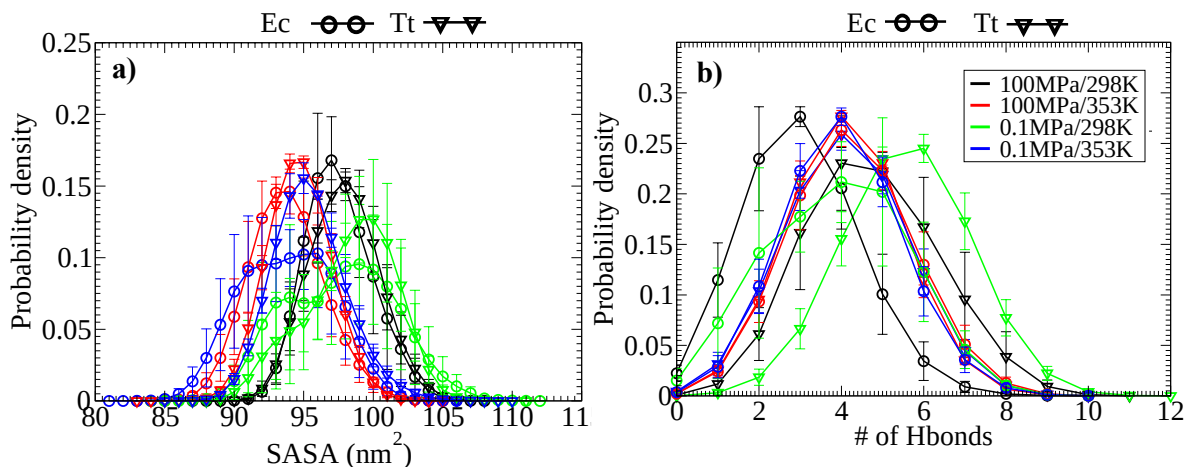


Figure 4.8 Probability distribution for solvent accessible surface area (SASA) for both *Tt*PPase monomer (triangle) and *Ec*PPase monomer (circle); b) Number of hydrogen bonds between the residues in the catalytic pocket for both *Tt*PPase monomer (triangle) and *Ec*PPase monomer (circle). The cutoff distance of 0.32 nm and cutoff angle of 20° were considered for hydrogen bond analyses. Error bars show the standard error of the mean (SEM), and those not immediately visible are at most the size of the symbol.

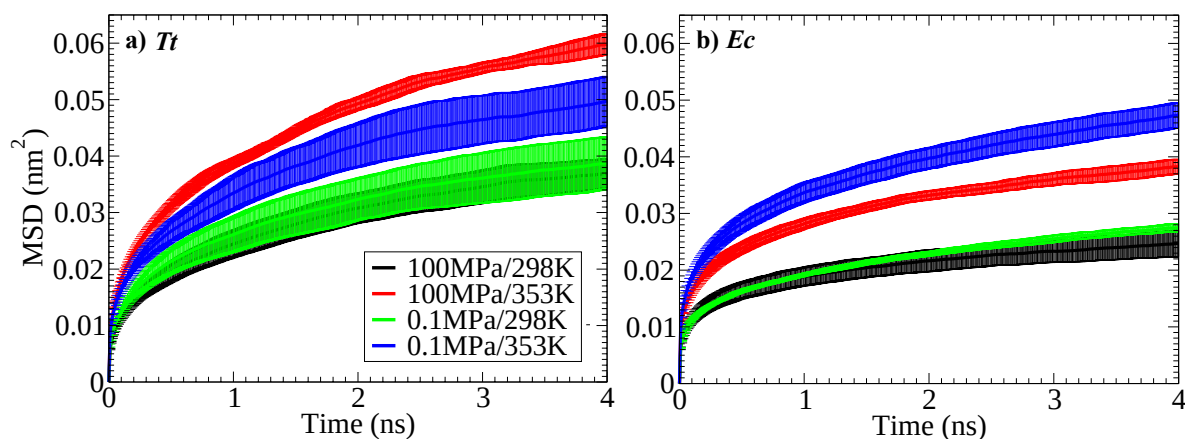


Figure 4.9 Conformational flexibility determined from mean square displacement (MSD) of C-alpha atoms for (a) *Tt*PPase monomer and (b) *Ec*PPase monomer. Error bars show the standard error of the mean (SEM).

Table 4.1 Binding pocket residues used for the calculations

Protein	Binding pocket residues
<i>Ec</i>PPase	Asp 42,65,67,70,97,102 Glu 20,31 Arg 43 Lys 29, 104,142 Tyr 55,141
<i>Tt</i>PPase	Asp 43,66,68,71,98,103 Glu 22, 32 Arg 44 Lys 30,105,141 Tyr 56, 140

Table 4.2 Molecular dynamics simulations details

System type	298K (0.1 and 100 MPa), 353K (0.1 and 100MPa)	Replicates
Hexamer	500 ns	3
Monomer	1000 ns	3

Table 4.3 Hexameric crystal structure comparison

Properties	Ec (1I6T)	Tt (3Q5V)
Residues	1049	1062
Mass (KDa)	117.3	124.2
Radius of gyration (nm)	2.9	2.9
SASA (nm²)	393	374
# of water in CBC	427	303
# of Hbonds between catalytic pocket residues	33	42

5 Early Stage Structure-Based Drug Discovery

5.1 Competitive inhibitors for Serine-Rich Repeat (Srr) Adhesin

A version of this chapter was originally published by me:

Agarwal, R., Bensing, B. A., Mi, D., Vinson, P. N., Baudry, J., Iverson, T. M., & Smith, J. C. Structure based virtual screening identifies novel competitive inhibitors for the sialoglycan binding protein Hsa. *Biochemical Journal* (2020)

In this work, I have designed the computational strategies, performed all analysis, and written the paper; BAB: made substantive contributions to the manuscript's content; DM and PNV: performed screening assay and helped in result interpretation; TMV: conceived the idea and made substantive contributions to the manuscript's content; JB and JCS: provided help in results discussion and edited the manuscript.

5.1.1 Abstract

Infective endocarditis (IE) is a cardiovascular disease often caused by bacteria of the *viridans* group of streptococci, which includes *Streptococcus gordonii* and *Streptococcus sanguinis*. Previous research has found that a serine-rich repeat (SRR) proteins on the *S. gordonii* bacterial surface play a critical role in pathogenesis by facilitating bacterial attachment to sialylated glycans displayed on human platelets. Despite its important role in disease progression, there are currently no anti-adhesive drugs available on the market. Here, we performed structure-based virtual screening using an ensemble docking approach followed by consensus scoring to identify novel small molecule effectors against the sialoglycan binding domain of the SRR adhesin protein Hsa from the *S. gordonii* strain DL1. The screening successfully predicted nine compounds which were able to displace the native ligand (sialyl-T antigen) in an *in vitro* assay and bind competitively to Hsa. Furthermore, hierarchical clustering based on the MACCS fingerprint showed that eight of these small molecules do not share a common scaffold with the native ligand. This study indicates that SRR family of adhesin proteins can be inhibited by diverse small molecules and thus prevent the

interaction of the protein with the sialoglycans. This opens new avenues for discovering potential drugs against infective endocarditis.

5.1.2 Introduction

Infective endocarditis (IE) (or bacterial endocarditis (BE)) is a life-threatening infection of cardiac valves and the interior surface of the heart (endocardium)¹⁷⁸. Oral streptococci account for ~17-45% of all cases of IE^{304, 305}. If untreated, infection destroys the valves and results in heart failure^{180, 306, 307}. Moreover, bacteria may also form clots (emboli) that enter the blood stream and produce strokes. IE affects 10,000-20,000 patients in the US every year and is associated with an in-hospital mortality rate of ~20% and a five year mortality rate of ~40-70 %^{178, 308, 309}. Treatment for endocarditis currently requires prolonged antimicrobial therapy, often combined with surgery. The rise in antibiotic resistance³¹⁰ has limited our pharmacological options^{311, 312}, and resistant organisms have increased the mortality rate³¹³. Although medical therapy alone often resolves infection, 47% or more of the patients eventually require valve replacement due to the damage incurred^{306, 307, 314}. Given the associated morbidity and rising mortality rate, there is an urgent need to develop novel therapies against IE.

Previous studies have reported that the binding of bacteria to host platelets contributes to the colonization of damaged aortic valves¹⁸⁰. A cell wall-anchored serine-rich repeat (SRR) protein mediates the adherence of *S. gordonii* and *S. sanguinis* to sialoglycans displayed on the human platelet²¹⁶ glycoprotein GPIb^{181, 182}. SRR proteins have been demonstrated to be virulence factors for endocarditis^{181, 182}, and disrupting the interaction between the SRR protein and sialoglycans on host platelets may therefore reduce virulence. *Streptococcus gordonii* is one of the well-studied species that cause IE and is a normal component of the human oral microbiota¹⁸³. Platelet binding by *S. gordonii* strains M99 and DL1 are facilitated by the homologous SRR proteins GspB and Hsa, respectively³¹⁵. Although these adhesins have high sequence identity, their ligand binding regions (BRs) differ significantly and have different sialoglycan

selectivity^{181, 195, 205}. GspB binds with narrow selectivity to sialyl-T antigen (sTa) whereas Hsa binds promiscuously to a range of glycans^{181, 195, 205}. Anti-adhesive therapies have been explored for the treatment of a wide range of other bacterial infections^{226, 316, 317}, but have not yet been pursued for IE. Anti-adhesives can, in principle, complement traditional antibiotics and improve their efficacy, potentially eliminating the need for surgical intervention. Moreover, an inhibitor might also reduce "re-seeding" (bacteremia is a hallmark of IE), or could be used as a prophylactic in some situations. Additionally, anti-adhesive agents are not bactericidal and hence the propagation and spread of resistant strains is much less likely to occur than as a result of exposure to bactericidal agents, such as antibiotics.

The crystal structures of the BRs (Hsa_{BR} (PDB 6EFC)¹⁹⁸, GspB_{BR} (PDB 6EFA)¹⁹⁸, 10712_{BR} (PDB 6EFF)¹⁹⁸, SK150_{BR} (PDB 6EFB)¹⁹⁸, SrpA_{BR} (PDB 5EQ2)¹⁹⁶, and SK678_{BR} (PDB 6EFI)¹⁹⁸) from a number of *S. gordonii* and *S. sanguinis* SRR proteins have been solved^{196, 198, 206}. These all have two domains which are associated with sialoglycan binding: the Siglec (Sialic acid-binding immunoglobulin-like lectin) domain and the Unique domain (for which function is not known completely). Furthermore, recent studies have identified that the three loops (CD, EF and FG) adjacent to the sialoglycan binding site are critical for the affinity and selectivity between ligands¹⁹⁸. Additionally, it has been reported that a partially conserved "YTRY" motif in the binding site is necessary for formation of hydrogen bond interactions with the sialic acid of the native ligand¹⁹⁸ and a crystal structure of Hsa_{BR} bound to sTa (PDB 6EFD¹⁹⁸) has been resolved showing these interactions. Importantly, there are also human sialoglycan-binding proteins³¹⁸, that contain a sialoglycan binding site but the site differs significantly in both geometry and in the location of hydrogen-bonding donors and acceptors from that found in the streptococcal Siglec-like proteins¹⁹⁷. Moreover, the mode of interaction with sialoglycans is distinct between human Siglecs and bacterial sialoglycan binding adhesin proteins¹⁹⁷.

These structural information can be leveraged by structure-based approaches to the identification of new molecular effectors to target SRR adhesin proteins. In our current pipeline, we targeted the BR of the well-characterized SRR protein Hsa (Hsa_{BR}), using *in-silico* virtual screening of ~105,000 small molecules, with the goal to predict and prioritize which small molecules could bind to Hsa_{BR} and disrupt its interactions with sialoglycan. Moreover, since Hsa_{BR} binds promiscuously to many glycans using a conformation selection mechanism¹⁹⁸ and crystal structure of both the apo and holo forms exists, it is potentially a good target for such a chemical biology approach.

Here, instead of using only the crystal structure in our computational approaches, we used molecular dynamics (MD) simulations to describe the flexibility of the binding pocket and generate an ensemble of protein conformations, which has been shown to yield large and diverse sets of molecular effectors to control protein functions^{74, 319}. Following subsequent high throughput ensemble docking, we prioritized the compounds using consensus scoring, which has previously shown to reduce the number of false positives and increase the success rate³²⁰. To further improve our predictions, we cross screened the compounds against the BRs from five Hsa homologues and identified compounds which bound to Hsa_{BR} with relatively higher docking scores compared to other BRs. From our virtual screening predictions, we were able to achieve a high success rate of ~20%, finding that 9 out of 50 compounds that were suggested for experimental validation were indeed able to displace the native ligand from the Hsa_{BR} binding pocket. Moreover, we were also able to identify scaffolds distant from the native ligand that bind to Hsa_{BR}. To our knowledge, these are the first small molecules described to inhibit binding of SRR family of adhesin protein to its native sialoglycan.

5.1.3 Methods

System preparation and molecular dynamics simulation

Crystal structures of the sialoglycan binding proteins Hsa_{BR} (PDB 6EFC)¹⁹⁸, GspB_{BR} (PDB 6EFA)¹⁹⁸, 10712_{BR} (PDB 6EFF)¹⁹⁸, SK150_{BR} (PDB 6EFB)¹⁹⁸, SrpA_{BR}

(PDB 5EQ2)¹⁹⁶, and SK678_{BR} (PDB 6EFI)¹⁹⁸ were used in this study. Molecular dynamics (MD) simulations were performed on all these proteins using the Amber14 ff14SB force-field parameters^{58, 221}. Each of these proteins was surrounded by an octahedral box of water model TIP3P²⁰⁰ of 10 Å. First, the protein structure was held fixed with a force constant of 500 kcal mol⁻¹ Å⁻² while the system was minimized with 500 steps of steepest descent followed by 500 steps with the conjugate gradient method. In the second minimization step, the restraints on the protein were removed and 1000 steps of steepest descent minimization were performed followed by 1500 steps of conjugate gradient. The system was heated to 300 K while holding the protein fixed with a force constant of 10 kcal mol⁻¹ Å⁻² for 1000 steps. Then, the restraints were removed, and 1000 MD steps were performed. The SHAKE algorithm²⁰¹ was used to constrain all bonds involving hydrogen in the simulations. 200 ns production MD was performed at 300 K using the NPT ensemble and a 2 fs time step with nonbonded cutoff of 10 Å. The temperature was fixed with the Langevin dynamics thermostat²⁰² and the pressure was fixed with a Monte Carlo barostat²⁰³. Similar MD simulation protocol was used on all the adhesins. This procedure yielded a total of 20,000 snapshots for subsequent analyses. Three independent runs were performed for each simulation.

In silico screening

Ensemble docking is an *in-silico* structure-based chemical biology method using an ‘ensemble’ of protein target conformations to discover novel protein effectors⁷⁴. The workflow used is shown in **Fig. 1**. The ensemble was constructed by clustering snapshots from molecular dynamics (MD) simulation trajectories by root mean square deviation (RMSD) of the binding pocket residues and loops (**Table S1**) surrounding the binding pocket with the hierarchical agglomerate clustering algorithm using Cpptraj module²⁰⁴.

The Vanderbilt small molecule collection (“The Discovery Collection”) containing ~105,000 compounds was docked to an ensemble of 5 conformations (4

representative structures obtained from clustering from MD and 1 crystal structure) with a cubic box with edges of ~30 Å. This small molecule library has been used in multiple high-throughput screens resulting in hits that have moved to hit-to-lead stages of early drug discovery programs³²¹⁻³²³. The docking box was centered on the C α atom of conserved residue THR 339 (Hsa numbering). VinaMPI³²⁴, a parallel version of AutodockVina⁹⁰, was used to perform the *in silico* screening. Similar docking procedure was performed for all the adhesin proteins. The docked poses were then ranked by the AutodockVina scoring function⁸⁹. The compounds were not only screened for Hsa_{BR} but also cross screened with 5 adhesin proteins (GspB_{BR}, 10712_{BR}, SK150_{BR}, SrpA_{BR}, SK678_{BR}). The cross screening was performed to remove the promiscuous compounds or “frequent hitters” (i.e., compounds which are always scored high for all the target) and thus reducing the number of false positives³²⁵. However, it must be noted that the goal of the cross screening was not to get selectivity towards Hsa_{BR}. From this ranked list of compounds, we tested compounds which were within the top 1% (~1050 compounds) for Hsa_{BR} but not within the top 1% of the other 5 BRs (GspB_{BR}, 10712_{BR}, SK150_{BR}, SrpA_{BR}, SK678_{BR}) and narrowed the list down to 250 compounds. We note that we only experimentally tested binding to Hsa_{BR} and not the selectivity of the predicted binders. Next, the resulting ~250 compounds were refined and rescored using two MOE scoring functions⁸⁶. The non-polar hydrogens (not included in Vina docking protocol) were added before performing the “induced fit” docking protocol in MOE⁸⁶. The docking poses were ranked using “GBVI-WSA dG” and “Affinity DG” scoring functions⁸⁶. Using consensus scoring, the top 50 compounds were then suggested for experimental validation. A flowchart of the screening methodology used is shown in **Fig. 5.1**.

Cheminformatics

All the physicochemical properties and fingerprints of small molecules were calculated using combination of MOE⁸⁶, ChemBioServer³²⁶ and RDkit³²⁷. MACCS fingerprints were calculated for each compound and similarity between them were

compared with the Tanimoto coefficient, followed by hierarchical clustering to cluster the similarity matrix.

Experimental assays

Protein expression and purification

GST-tagged Hsa_{BR} was expressed and purified as described in ref¹⁹⁵. GST-Hsa_{BR} was expressed under the control of the pGEX-3X vector in *E. coli* BL21 (*DE3*) in a Terrific Broth medium with 50 µg/ml kanamycin at 37 °C. When the OD₆₀₀ reached 1.0, expression was induced with 1 mM IPTG at 24 °C for 5 hrs. Cells were harvested by centrifugation at 5,000 × *g* for 15 min, optionally washed with 0.1 M Tris-HCl, pH 7.5, and stored at –20 °C before purification. The frozen cells were resuspended in homogenization buffer (20-50 mM Tris-HCl, pH 7.5, 150-200 mM NaCl, 1mM EDTA, 1 mM PMSF, 2 µg/ml Leupeptin, 2 µg/ml Pepstatin) then disrupted by sonication. The lysate was clarified by centrifugation at 38500 × *g* for 35-60 min and passed through a 0.45 µm filter. Benchtop purification was performed at 4 °C using Glutathione Sepharose 4B beads, with pure GST-Hsa were eluted with 30 mM GSH in 50 mM Tris-HCl, pH 8.0.

AlphaScreen high-throughput screening assay

We used the AlphaScreen modification of an ELISA as the primary target-based proximity assay to monitor ligand displacement. AlphaPlate (Cat # PE 6005351, Lot # 8220-16081) with 384-well was used for the screening. In the experimental setup, biotinylated sialyl T antigen (sTa) was coupled to a streptavidin donor bead and GST-tagged Hsa was coupled to an anti-GST conjugated acceptor bead in PBS (phosphate buffered saline). The reaction was excited at 680 nm to stimulate singlet oxygen-mediated energy transfer to the acceptor bead, which can be detected at 615 nm.

The dose-dependent signal reflects the number of bead-coupled adhesins bound to bead-coupled glycans. To determine the optimal ratio of Hsa-GST to biotinylated sTa for occupying binding sites on the beads and, therefore, maximal signal

production, the Hsa-GST concentration was titrated in a 10 point-3 fold dilution starting from 1000 nM and the biotinylated sTa concentration was titrated in a 9 point-3 fold dilution starting from 100 nM and the resulting Alpha signal measured. The maximal signal, representing the “hooking point” where either the donor or acceptor beads are saturated, was found to be 3 nM for Hsa-GST and 3 nM for biotinylated sTa. The final chosen concentrations used in the screen was 1 nM of Hsa-GST and 2 nM of biotinylated sTa, slightly below the hook point, to avoid potential excess Hsa that may sequester inhibitors and interfere with signal disruption. DMSO was used as the negative control and unbiotinylated sTa was used as a positive control at a concentration of 30 uM which was determined to provide maximal disruption of the Alpha signal.

We applied this assay to the evaluation of the test compounds that were predicted as binding to Hsa_{BR} using virtual screening. This initial screen was performed with all test compounds in duplicate at a final concentration of 10 µM and DMSO was used as the negative control and unbiotinylated sTa was used as a positive control.

Z' factor calculation

The Z' factor³²⁸ is an indicator of high throughput screening assay performance and was calculated as follows:

$$Z' = 1 - 3(\delta_p + \delta_n) / |(\mu_n - \mu_p)|$$

The standard deviations and means of the positive (p) and negative (n) controls are denoted by δ_p , μ_p and δ_n , μ_n respectively. DMSO and untagged sTa are the positive and negative control respectively.

Effector identification analyses

The alpha value of each test compound was measured and was filtered using 1-fold, 2-fold or 3- fold of either standard deviation (SD) from the mean of the negative control group, or absolute deviation from the median (MAD) of negative control group. We determined which tested points lay outside the mean of the vehicle control (there were 9 replicates of the negative control (DMSO), 4

replicates of the positive control (untagged sTa)). We used a threshold of both 3 SD and 3 MAD from the negative control group. This was followed by taking the union of the two. Then, it was further filtered by only keeping those molecules which hit twice in the confirmation duplicates. Finally, we calculated the Percentage Control from the control group to identify compounds that disrupted the Hsa-sTa interaction. This serves as an initial hit identifier but should be followed-up to confirm true actives and rule out false positives.

Percentage control (PC) calculated as follows:

$$\text{Percentage control (PC)} = 1 - \frac{\alpha_{NC} - \alpha_{Com}}{\alpha_{NC} - \alpha_{PC}}$$

where, α is the average alpha value for negative control (α_{NC}), positive control (α_{PC}) and compounds tested (α_{Com}). PC is a measure of the alpha signal of the 10 μ M test compound in percentage of the controls.

5.1.4 Results and Discussion

Protein dynamics and conformations

We used MD simulations to capture the internal dynamics of the proteins and find binding site conformations not seen in the crystal structure⁴⁹. We calculated the root mean square fluctuation (RMSF) to identify the flexible regions (**Fig 5.2a**). Although the overall structure of the Siglec domain is rigid, we observed that the loops (CD, EF and FG) close to the binding pocket are flexible for all the adhesin proteins (**Fig 5.2a, 5.6**).

In the case of Hsa_{BR}, we observed that the CD and EF loops constitute the most flexible region of the protein. Moreover, critical binding pocket residues other than in these loops were identified from the crystal structure of Hsa_{BR} and the native ligand (sTa) (**Table 5.2**).

To capture new conformations of the binding pocket that deviate from the initial crystal structure, the root mean square deviation (RMSD) of the loop residues and other critical residues (previously known to bind to the native ligand) (**Table 5.2**) were used to cluster the MD trajectories. The clustering resulted in four different

clusters. The structure closest to the centroid of each cluster was used for docking. The “ensemble” of structures obtained from clustering and crystal structure were superimposed to observe the deviation of structures (as shown in **Fig 5.2b**). We observe that all the structures had similar secondary and backbone structures and the RMSD_{Calpha} of the Siglec domain was calculated to be within ~1.5 Å. However, as seen in the superimposed structures (**Fig 5.2b**), the loop regions (especially CD and EF loop) have different orientations in the “ensemble” when compared to the crystal structure. Similarly, we observed that the side chains in the binding pocket residues orient differently between the structures, which can be critical for rigid body docking.

Physicochemical properties of small molecule database

The five structures obtained from MD simulations and the existing crystal structure were screened against the Vanderbilt small molecule collection (“The Discovery Collection”) containing ~105K compounds. Although, this database has been used in several early drug discovery programs³²¹⁻³²³, it has not been characterized yet. Therefore, before performing the virtual screening, and although the goal of the present work is just to identify a molecular effector, we wanted to calculate the physicochemical properties to understand its use as a potential drug precursor for future studies. Firstly, we calculated the molecular weight (MW) of the compounds (**Fig 5.3a**), which is known to be critical for safety and tolerability reasons⁸². The Vanderbilt database has compounds with MW less than 500 Da that are considered to improve druglikeness^{329, 330} and also has low MW compounds (<300 Da) that are considered better initial precursor because they serve as effective chemical starting points for lead optimization³³¹. The polar surface area and the number of rotatable bonds have been found to better discriminate between compounds that are orally active³³². It has been predicted that compounds with 10 or fewer rotatable bonds and those having a polar surface area of less than 140 Å² have a good oral bioavailability³³². In our database, we observed that most compounds had a mean polar surface area of ~150 Å² and less than 10 rotatable

bonds (**Figs 5.3b, c**). Lipophilicity (SLogP) is another factor which is known to influence drug potency, pharmacokinetics, and toxicity^{329, 333}. Compounds with SLog P values between -0.4 to +5.6 range are known to be more “druglike”^{330, 334}. Here, we found that most of the compounds fall within this range (**Fig 5.3d**). Although the above is a set of physicochemical properties that are considered to be important for different aspects of druggability, there have been numerous FDA approved drugs which violate one or more of these rules^{335, 336}.

Docking results and poses

After our virtual screening, we first ranked all the top poses for each compound based on the Autodock Vina scoring function⁸⁹. Subsequently, we selected those compounds (~250) that were in the top 1% for Hsa_{BR} but did not rank within the top 1% for any other adhesin protein (GspB_{BR}, 10712_{BR}, SK150_{BR}, SrpA_{BR}, SK678_{BR}). This was followed by implementing consensus scoring in which the poses (obtained from AutodockVina) were energy-minimized and then rescored using two MOE scoring functions⁸⁶ (as mentioned in the Methods section). In the end, compounds that ranked within the top 50 for all the three scoring functions were suggested for experimental validation. Next, we examined the number of electrostatic intermolecular interaction (hydrogen bonds (HBs) and Pi (π - π / π -H/ π -cation)) which are important for protein-ligand binding³³⁷⁻³³⁹. The importance of hydrogen bonding in drug design is well recognized and the hydrogen-bonding capabilities deeply influence the transport and ADME (Adsorption, Distribution, Metabolism and Excretion) properties of a molecule as well as its specific interaction with biological receptors³⁴⁰. Many QSAR studies have been reported in which hydrogen-bonding interactions play a key role in modeling a particular target activity^{341, 342}. Therefore, we calculated the number of compounds which form an interaction with the residues known to bind (or in close proximity) sTa to get information about which residues were targeted the most and are easily accessible to interact with a small molecule (**Fig 5.10**). Thr 339, Tyr 337 and Lys 335, which form HBs with the native ligand (sTa) in the crystal structure (PDB 6EFD)¹⁹⁸ as

shown in **Fig. 5.4a**, are also some of the residues which form interactions (HBs or Pi) with majority of the compounds. Moreover, from our strategy we found that majority of the compounds form 3 or more HBs or Pi interactions with the binding site residues, whereas 10 compounds make more than 5 HBs or Pi interactions (**Fig. 5.4b**). Furthermore, out of these 50 compounds, 25 compounds (50%) were predicted to bind to two of the “ensemble” structures generated from MD simulations with higher score than to the crystal structure. This further illustrates the usefulness of using ensemble docking.

Experimental validation

Alpha assay screening was performed for the top 50 compounds predicted to displace sTa (the highest affinity native ligand) from Hsa_{BR}. The Z' factor value of the DMSO (negative control) versus untagged sTa (positive control) was 0.32, which denotes that there is a separation between the high and low signals of the assay in that 3x the sum of the standard deviations of the high and low signals of the assay divided by the difference of the mean of these two experimental groups is 0.68 (the error is relatively small compared to the separation of the mean of the two groups).

After filtering the small molecules using the experimental data based on the percentage control (PC), nine small molecules were retained. These nine compounds showed a statistically significant decrease in the signal when the two replicates were averaged (**Fig. 5.5**). These compounds have a PC three standard deviations outside the mean of the negative control (DMSO).

The IC₅₀ of the untagged sTa (positive control) was calculated as 8.67 μ M (**Fig. 5.7a**). At 10 μ M concentration, the PC was 39% (**Fig. 5.7b**). At the same concentration, the PC of the 9 small molecules ranges from 23% to 70% and, out of these, two compounds have PC values of less than 39% and one has a PC of 41% (**Table 5.3**).

Computational and binding pose analyses of experimentally validated effectors

The nine small molecules were screened for 25 known toxic and carcinogenic fragments, such as anthracene, quinone, hydroquinone, butenone--Michael acceptor, chloroethane--Michael acceptor³²⁶. Of the 9 experimentally validated compounds (C1-C9) (**Table 5.1**), only Compound 1 (C1) was identified as potentially toxic, containing a benzo-dioxane and a catechol group. Moreover, to test the similarity between these effectors and the native ligand, fingerprint-based hierarchical clustering was performed. We found four clusters (**as shown in Fig. 5.8**), which showed that the compounds identified from the screen are diverse among themselves and are not similar to the native ligand. Additionally, we also tested the compounds for Lipinski's rule³³⁰, to evaluate druglike-ness of the compounds. C4 was the only molecule with one violation (with 11 hydrogen bond acceptors), whereas all the other compounds satisfied all the 4 rules.

Following the above cheminformatics analyses of the experimentally validated effectors, we examined the computational models of the best poses and the interactions of the nine small molecules shown to have inhibition experimentally (**Fig. 5.11 and Table 5.1**). Interestingly, the inhibitor binding site is adjacent to the sTa binding site (**Fig. 5.9**) but both the sites does partially overlap. This site can be further explored for more selective inhibitors in future studies.

Next, we calculated interaction map of each of the nine effector and looked at the hydrogen bonds between the ligand and the side chain and backbone atoms of the binding site residues. In the models C1, C2, C4, and C5 form backbone HBs with Asp 255 and compounds C2-4 form backbone HBs with Val 367 while other compounds (C1, C5-9) form side chain HBs with Val 367. Interestingly, C3 forms four backbone HBs with Gly 362, Phe 366 and Val 367, and five sidechain HBs with Tyr 337, Thr 339, Ser 253, Asn 361 as shown in the interaction map (**Fig. 5.11 and Table 5.1**). Other residues that form HBs with most of the compounds are Thr 339, Val 285 and Asn 361 (**Fig. 5.11 and Table 5.1**). The interaction maps of all the nine small molecule effectors are shown in **Fig. 5.11**. The orientation of the

best docked pose of the nine validated effectors and crystal structure pose of the native ligand in the binding pocket are very different and bind in adjacent sites (**Fig. 5.9**). However, the residues that form HBs with the effectors, also form HBs with native ligand or are in close proximity of it. Hence, it is likely that these nine compounds are able to displace the native ligand in part because they form HBs with these critical residues.

5.1.5 Conclusions

The SRR protein Hsa has been considered an attractive molecular target for drug discovery due to its role in infective endocarditis (IE). It is noteworthy that there is no vaccine or anti-adhesive drug approved against IE. Hence, it is important to use chemical biology approaches to identify and describe how small molecules could inhibit the adhering function of the protein, which may, down the road, open the door to drug discovery. Here, we performed structure-based virtual screening to identify competitive small molecule effectors for Hsa_{BR}. We combined three different SBDD strategies; ensemble docking, cross screening, and consensus scoring in one pipeline. For the ensemble docking, we generated an ensemble of receptor conformations from MD simulation, and then cross screened against five homologs (GspB_{BR}, 10712_{BR}, SK150_{BR}, SrpA_{BR}, SK678_{BR}). In the last step, three scoring functions (AutodockVina⁸⁹ and MOE⁸⁶) were used to rank and prioritize the list of compounds. The Vanderbilt database was used for the small molecules since it covers a wide distribution of different physicochemical properties.

The goal of combining these strategies was to improve the hit rate and reduce the number of false positives. Indeed, we were able to achieve a hit rate of ~20% and identified nine compounds that could displace the native ligand in the experimental assay. The binding poses of all the nine compounds identified from docking show that they are in close proximity with residues known to form HBs with the native ligand (sialyl-T antigen). These compounds may be used as a starting point for medicinal chemistry optimization. Further studies need to be conducted to characterize the binding affinities and poses of these identified

compounds, and similar analyses for other sialoglycan-binding SRR proteins are ongoing.

5.1.6 Appendix

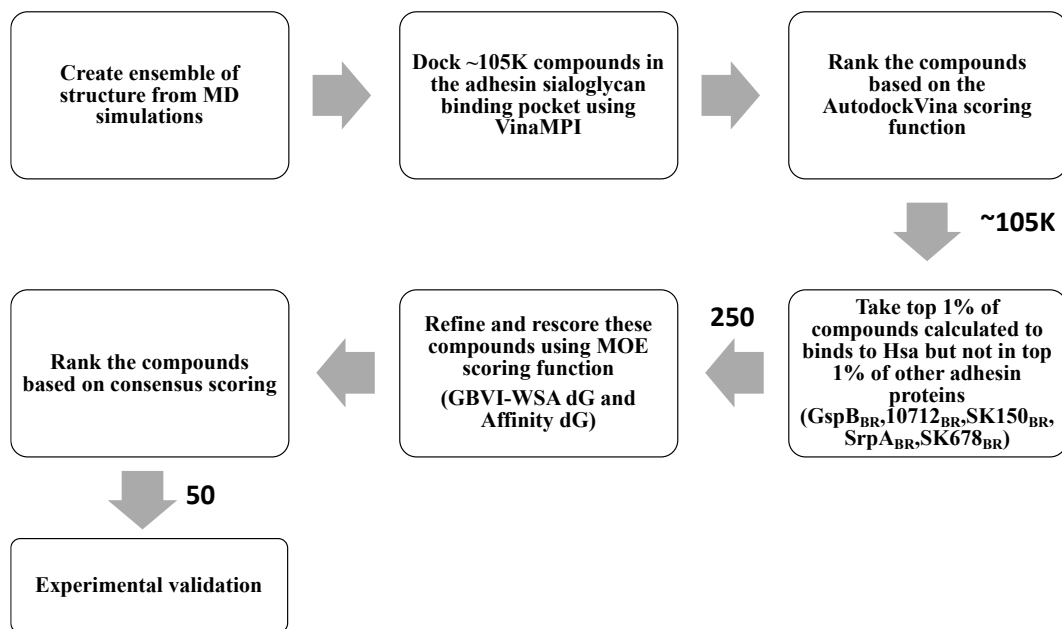


Figure 5.1 Structure based virtual screening strategy workflow showing number of compounds which were passed on to the next step.

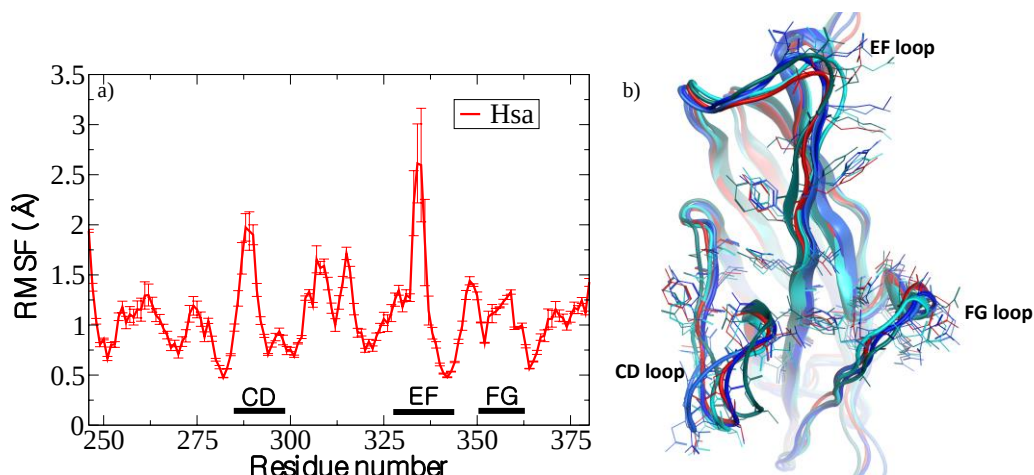


Figure 5.2 a) Root mean square fluctuation of Hsa_{BR} from MD simulation showing DC, EF, FG loop regions; **b)** Superimposed structures (in ribbon) of Hsa_{BR} obtained from different clusters showing residues (in stick) used to during the clustering: crystal structure (in red) and ensemble structure (in shade of blue).

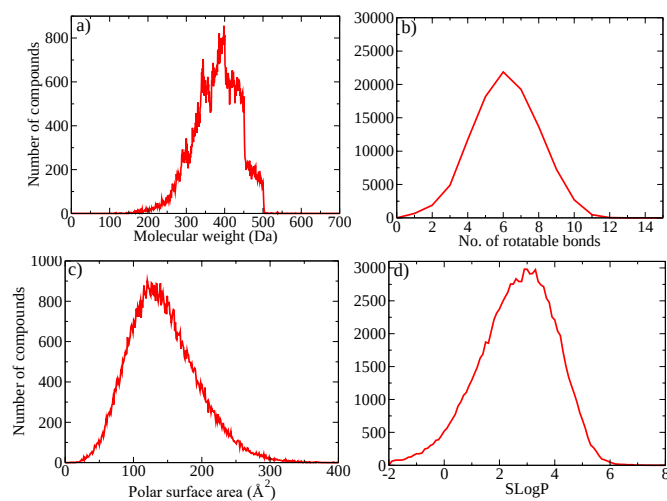


Figure 5.3 Density profile of physicochemical properties of small molecule database: molecular weight (a), number of rotatable bonds (b), polar surface area (c), and Log of the octanol/water partition coefficient: SLogP (d).

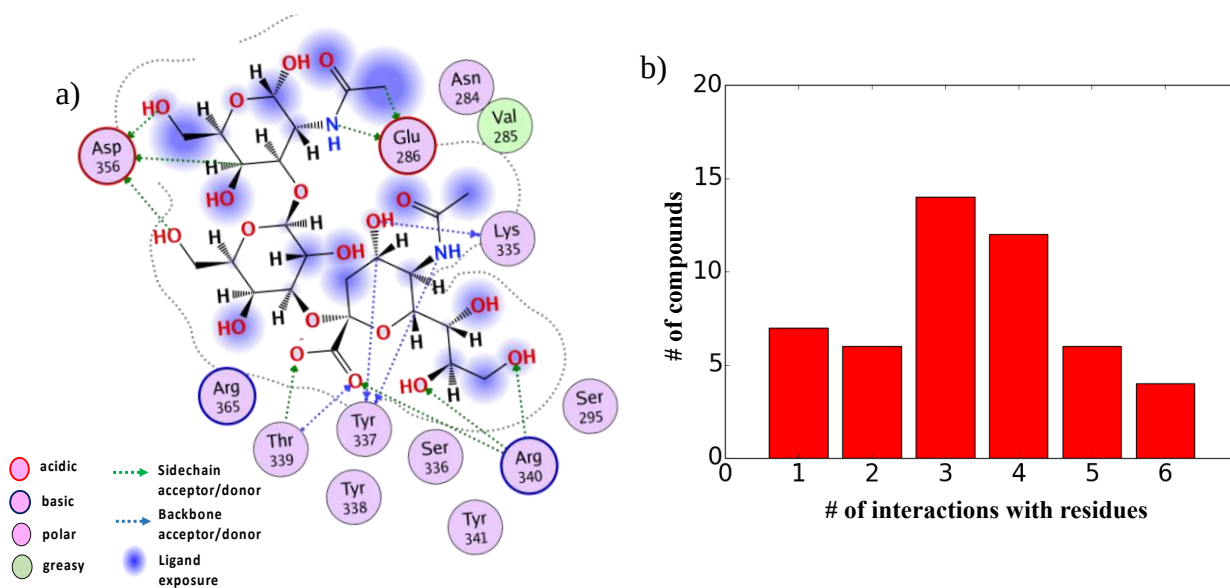


Figure 5.4 a) Interaction map of native ligand (sTa) from crystal structure (PDB 6EFD)¹⁹⁸; b) Bar plot showing number of compounds (within the top 50 compounds) and the number of interactions (Hbond or Pi (π - π / π -H/ π -cation) made with protein residues

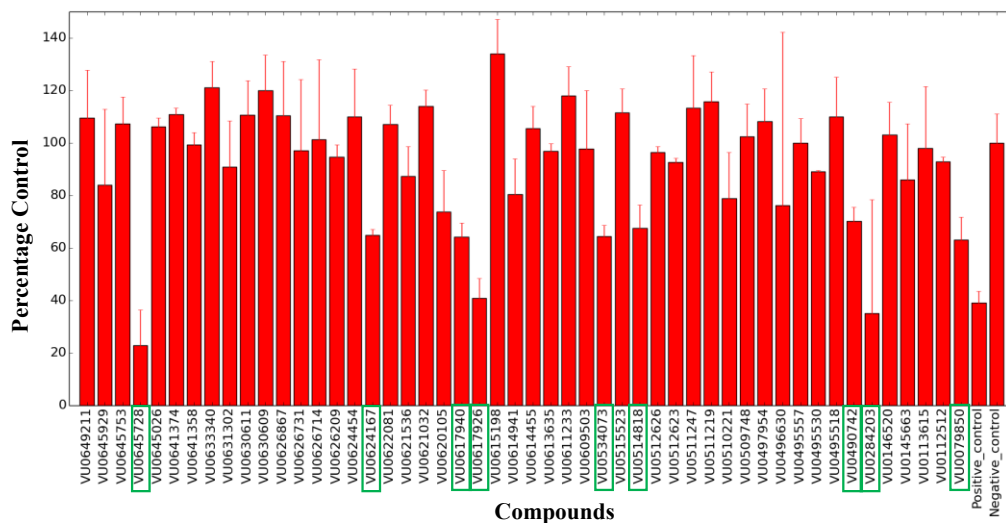


Figure 5.5 Alpha Screen assay. Experimentally validated effectors are marked by green boxes on the x-axes. Error bar represent the standard deviation.

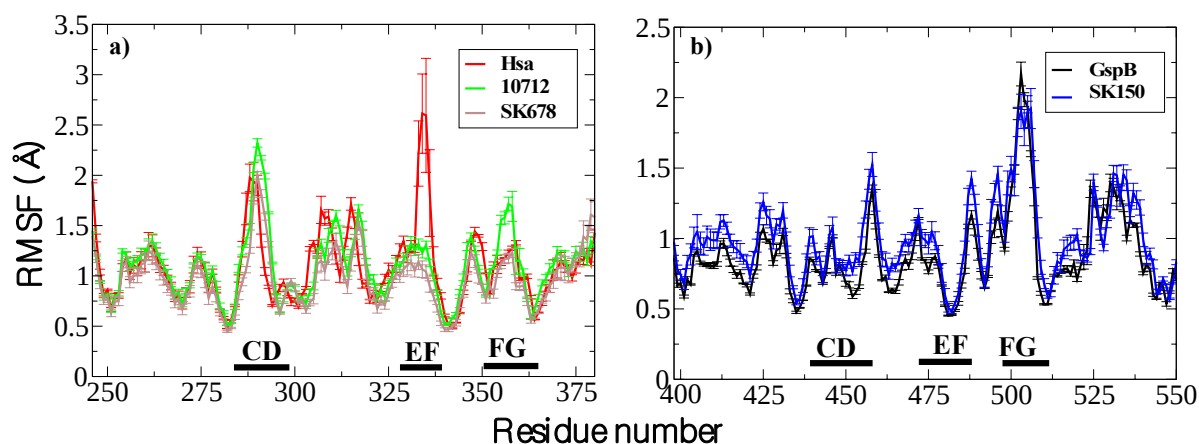
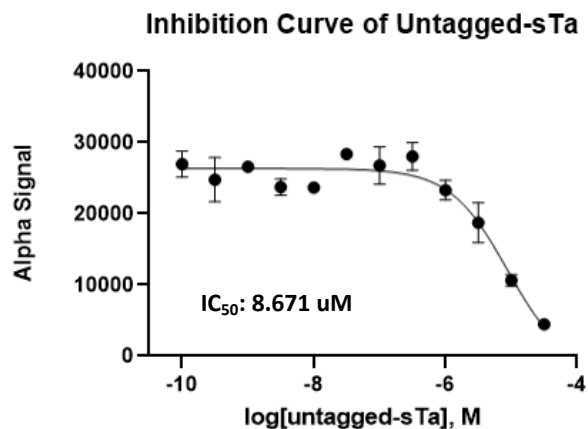


Figure 5.6 a) Root mean square fluctuation of HsaBR, 10712BR and SK678BR (a); GspBBR and SK150BR (b) from MD simulation showing DC, EF, FG loop regions. The residue numbers of HsaBR and GspBBR were used in (a) and (b) respectively. The error bar is SEM from three independent simulations.

a)



b)

Log (1x) untagged-sTa (Log M)	(1x) untagged- sTa (nM)	Percentage control
-4.5	30000	17%
-5.0	10000	39%
-5.5	3000	69%
-6.0	1000	86%
-6.5	300	104%
-7.0	100	99%
-7.5	30	105%
-8.0	10	88%
-8.5	3	88%
-9.0	1	99%
-9.5	0.3	92%
-10.0	0	

Figure 5.7 a) Inhibition curve of untagged sTa (Alpha Signal vs Log (untagged sTa)) used to calculate IC₅₀; b) Table showing the percentage control at different concentration of untagged sTa.

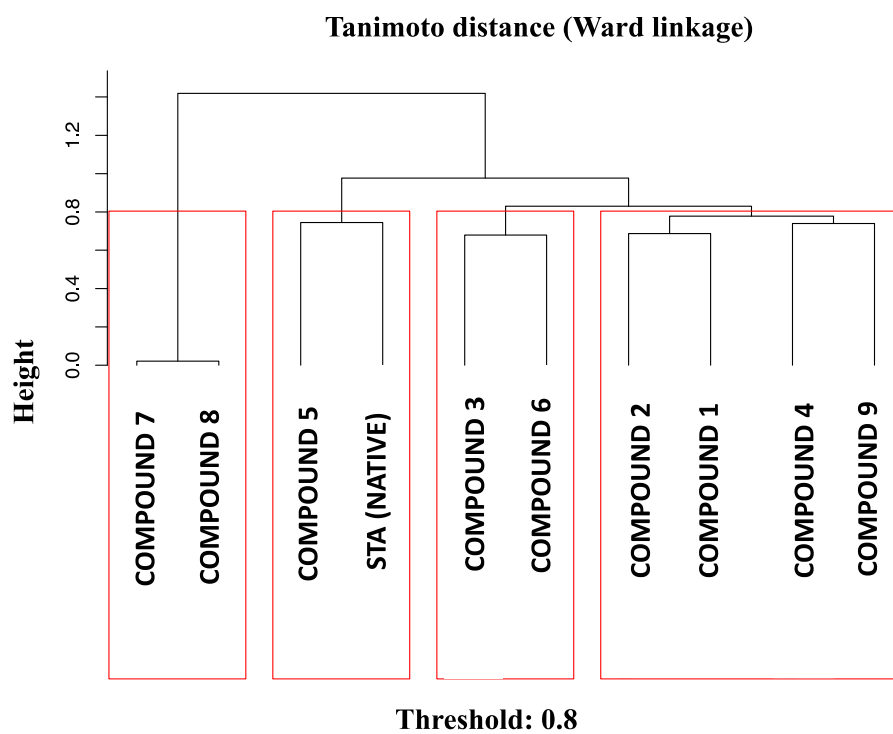


Figure 5.8 Hierarchical clustering based on MACCS fingerprint of the nine validated hits and the native ligand

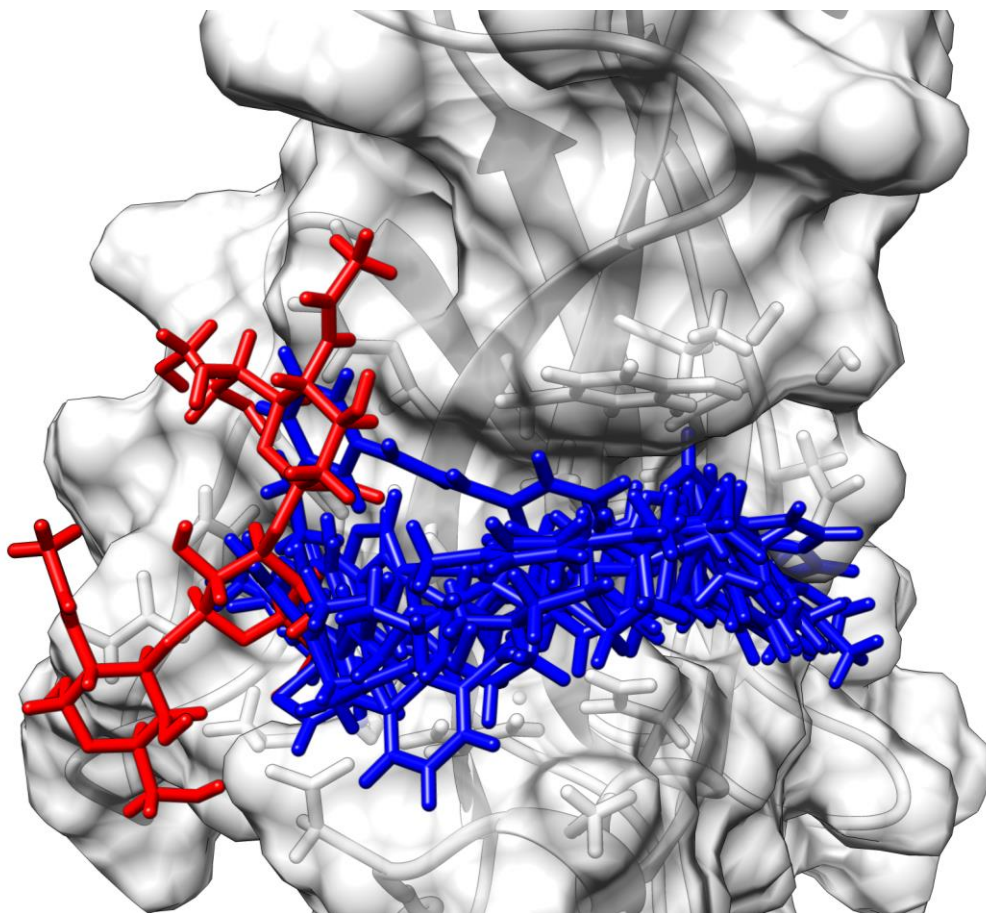


Figure 5.9 Crystal structure pose of native ligand (sTa) (in red) and best docked pose of 9 validated compounds (in blue) in the binding pocket of HsaBR

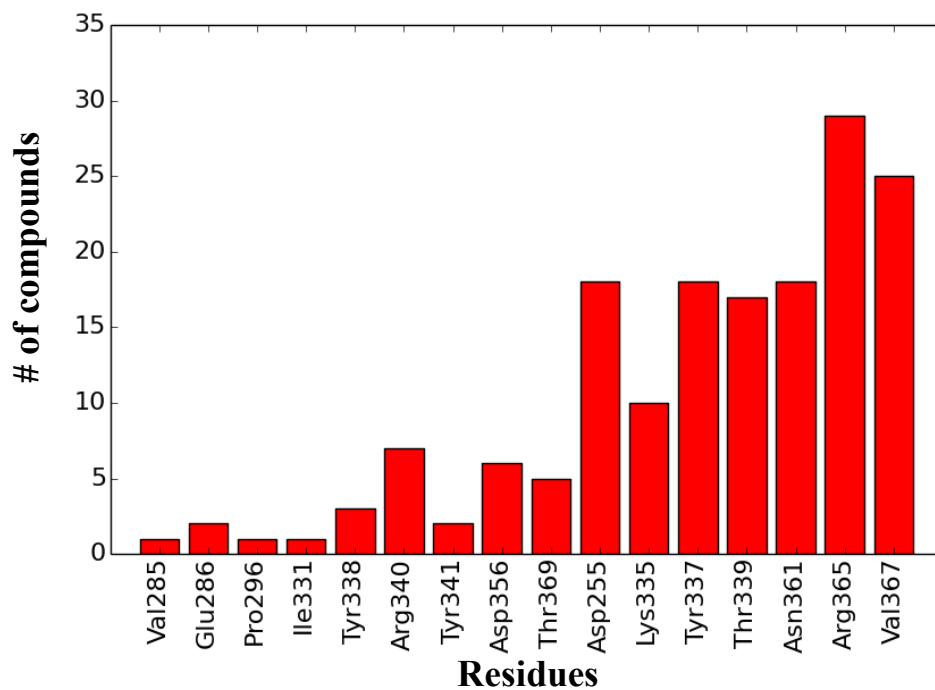


Figure 5.10 Number of compounds within the top 50 compounds interacting (forming Hbond or Pi (π - π / π -H/ π -cation) interaction) with key residues in the HsaBR binding pocket

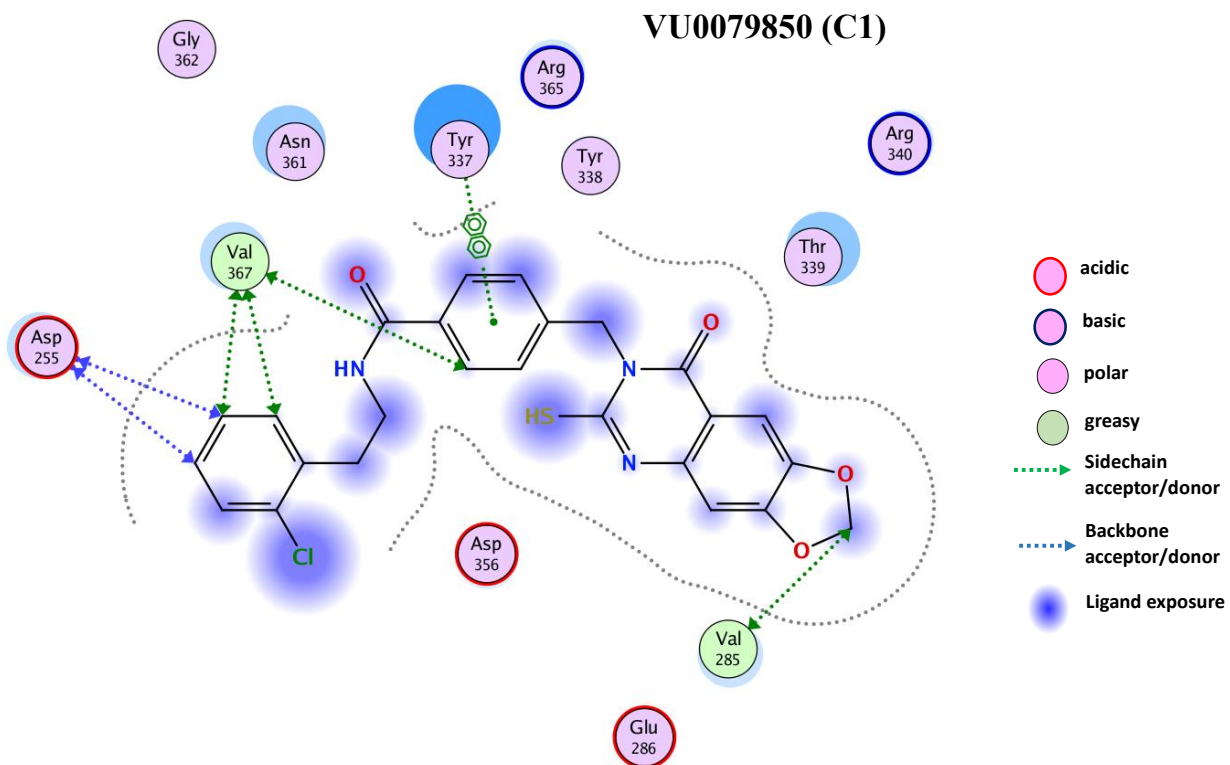


Figure 5.11 Interaction map of the nine hits in the binding pocket of the Hsa_{BR}

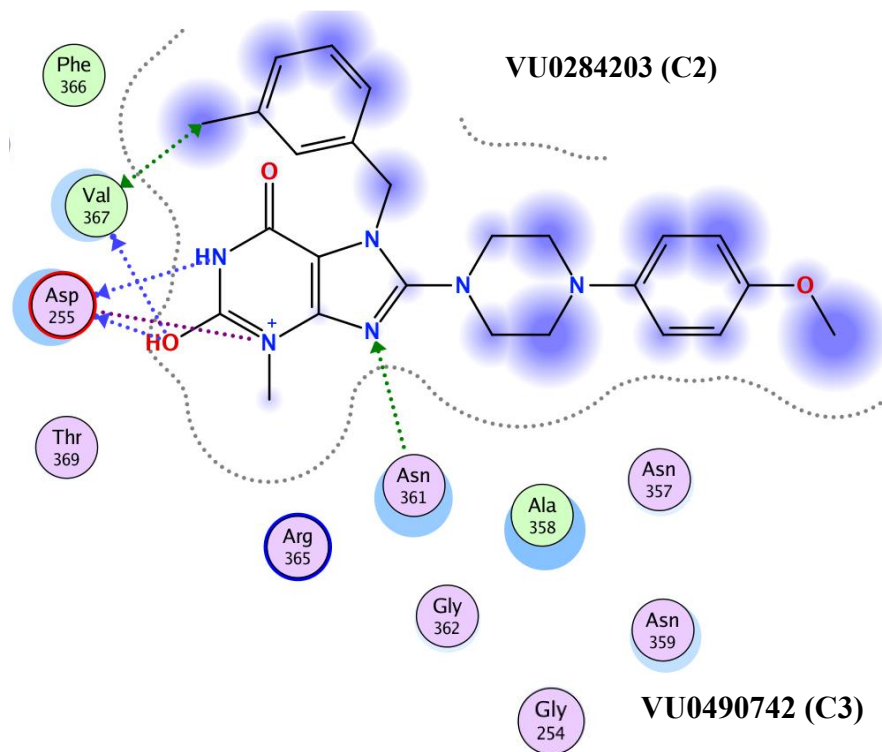


Figure 5.11 continued

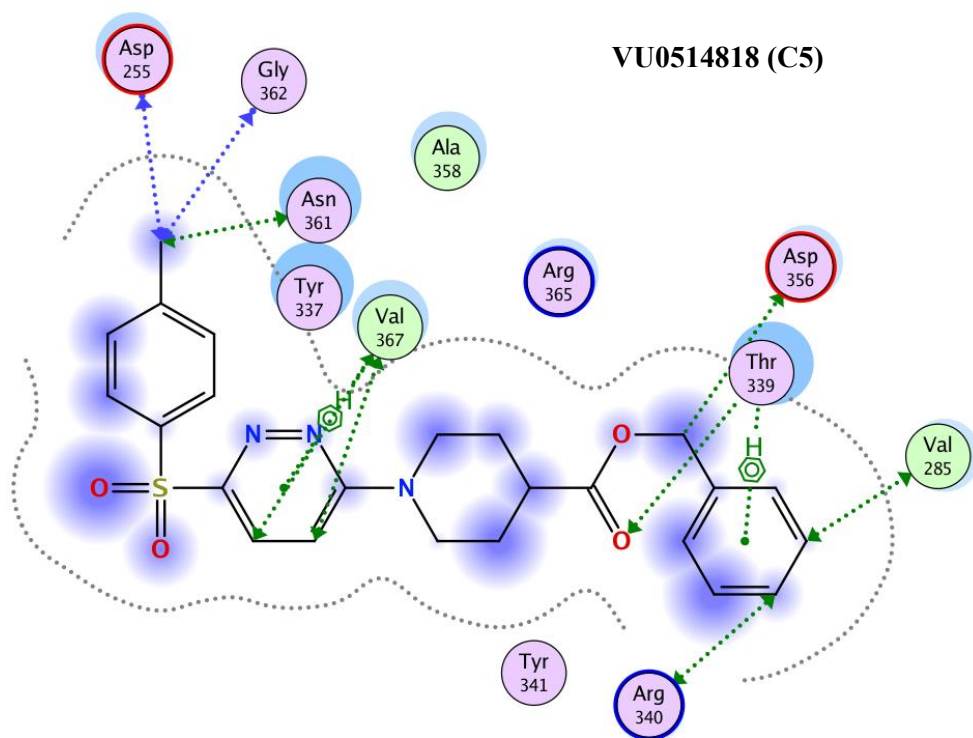
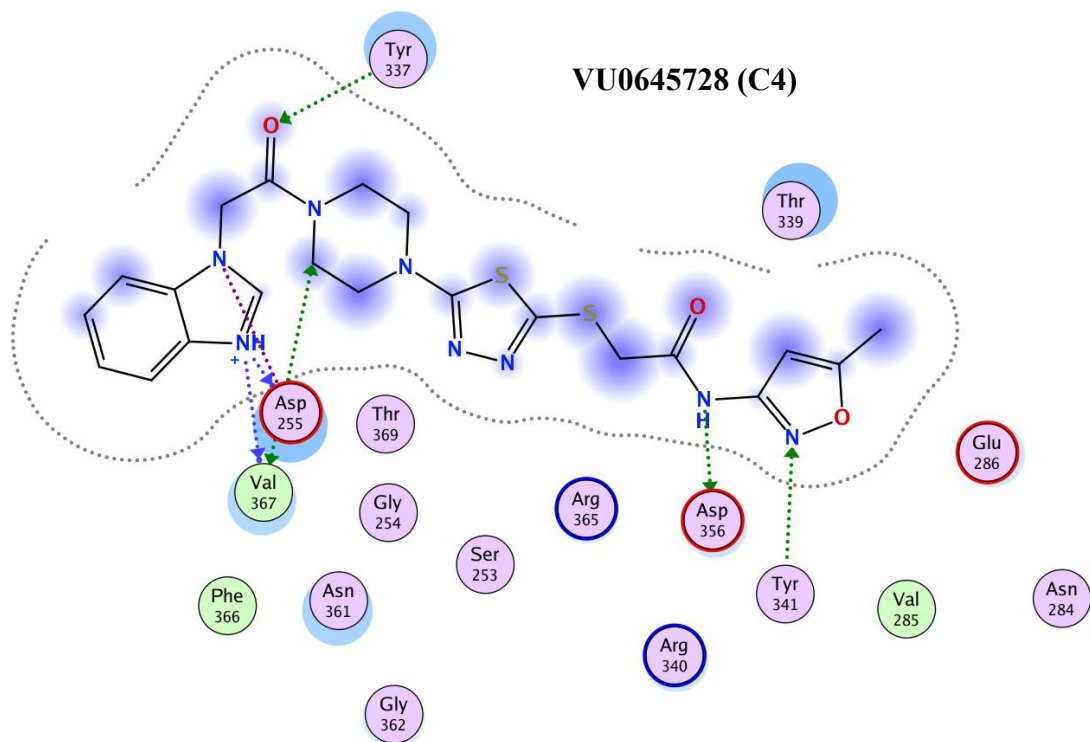


Figure 5.11 continued

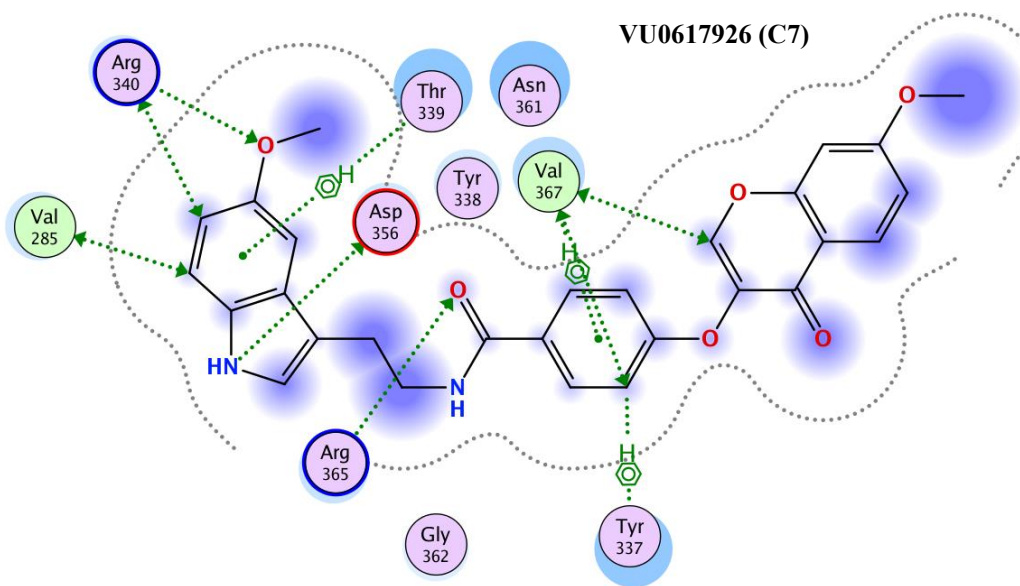
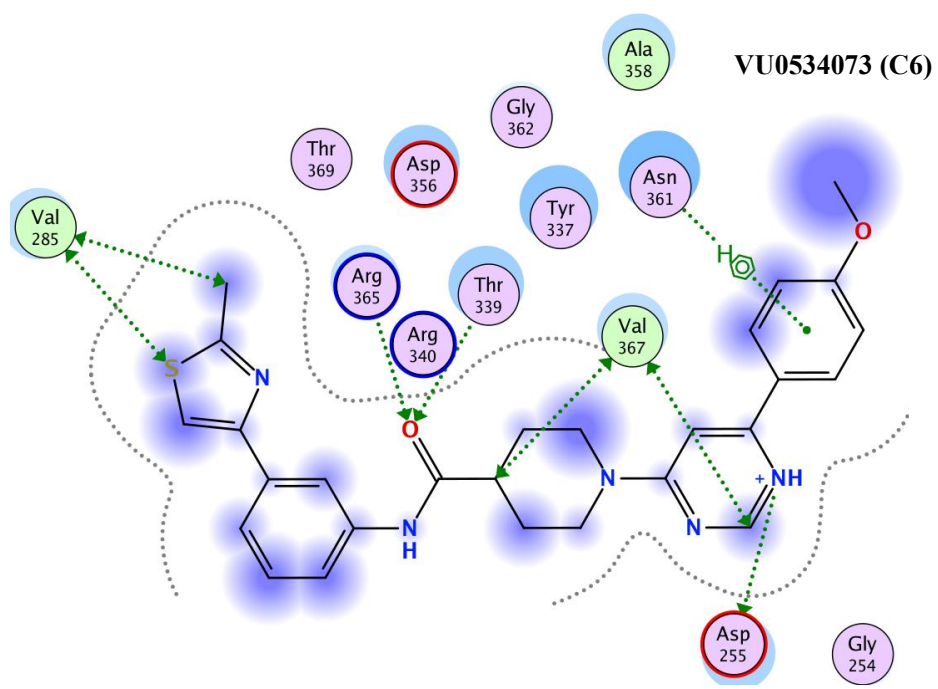


Figure 5.11 continued

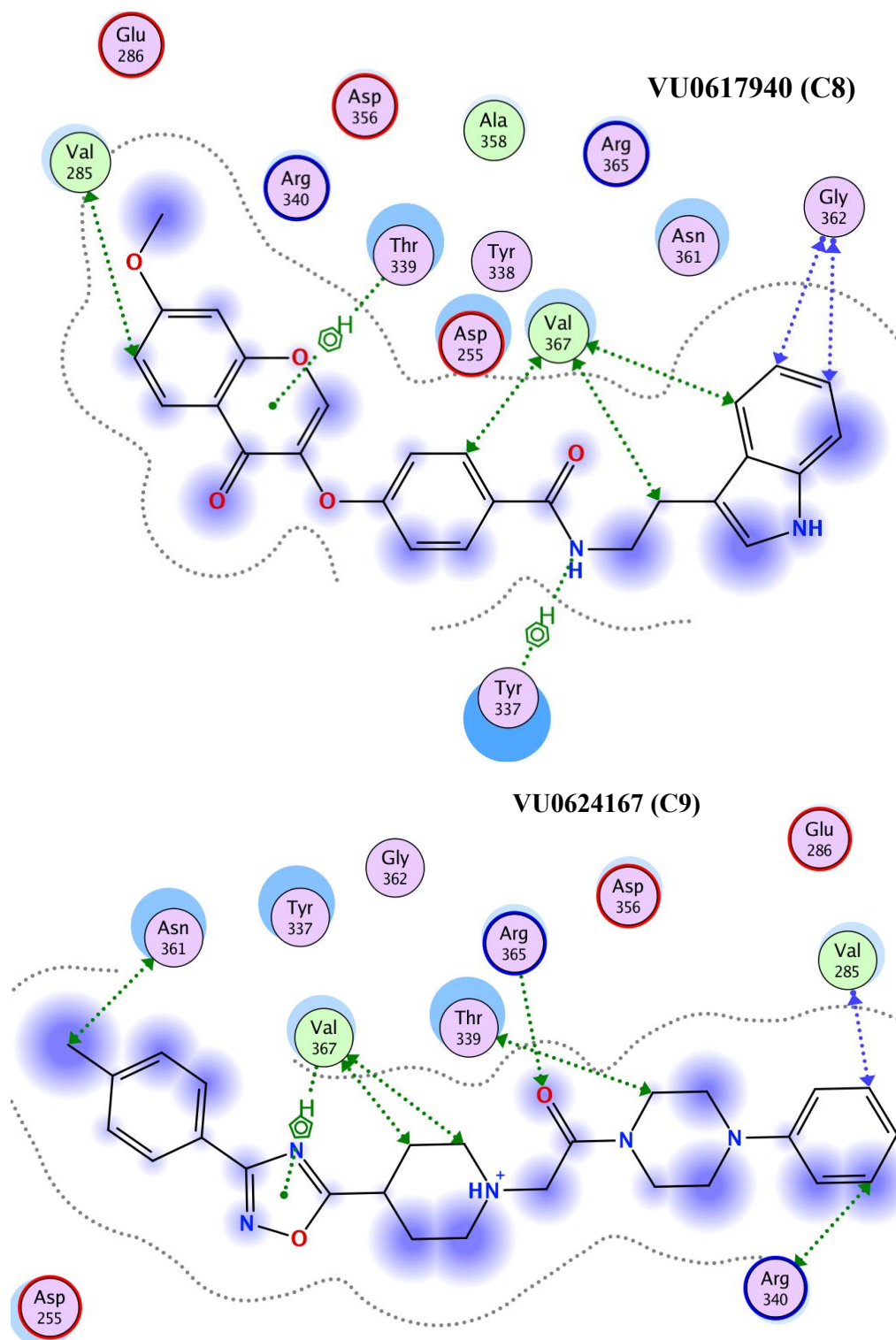


Figure 5.11 continued

Table 5.1 Structure of the nine hits and interaction details in the binding pocket of the Hsa_{BR}

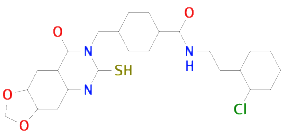
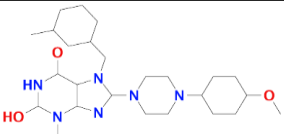
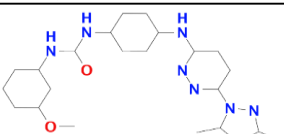
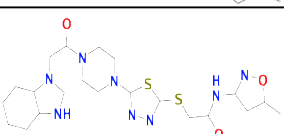
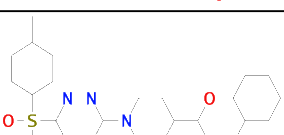
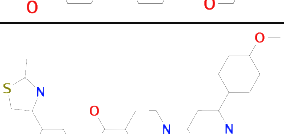
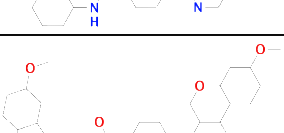
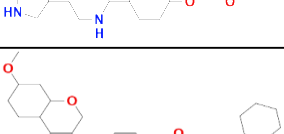
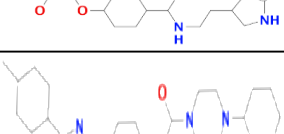
Compound number	Structure	ID	Protein residues forming Backbone HBs	Protein residues forming Sidechain HBs
C1		VU0079850	255	285,337,367
C2		VU0284203	255,367	361,367
C3		VU0490742	362,366,367	253,337, 339
C4		VU0645728	255,367	255,337,341,356
C5		VU0514818	255, 362	285,339,340,356,361,367
C6		VU0534073	-	255,285,339,361,365,367
C7		VU0617926	-	285,337,339,340,356,365,367
C8		VU0617940	362	285,337,339,367
C9		VU0624167	285	339,340,361,365,367

Table 5.2 Residues included during clustering of MD simulation snapshots to form an ‘ensemble’

Protein	Residues
Hsa_{BR}	285-295, 333-343, 356-365
SK678_{BR}	297-307, 342-352, 366-376
SrpA_{BR}	288-298, 338-348, 363-368
10712_{BR}	286-299, 335-345, 355-366
GspB_{BR}	437-447, 477-487, 503-510
SK150_{BR}	296-306, 336-346, 356-366

Table 5.3 Validated compound list with their respective percentage control

Test compound (1x) conc (nM)	Compound Name	Percentage control
10000	VU0490742	70%
10000	VU0514818	68%
10000	VU0624167	65%
10000	VU0534073	64%
10000	VU0617940	64%
10000	VU0079850	63%
10000	VU0617926	41%
10000	VU0284203	35%
10000	VU0645728	23%

5.2 Disruption of histone deacetylases 4 (HDAC4) complexation by novel selective inhibitors

5.2.1 Abstract

Histone deacetylases (HDACs) are a biologically important family of proteins which also comprise known cancer drug targets. Currently, there are several FDA approved drugs that target the catalytic pocket of HDACs, which is conserved within the family, and hence are not selective towards a particular subfamily of HDACs. This non-selectivity leads to off-site binding and toxicity. HDAC4 belongs to one such subfamily of HDACs (class IIA HDAC), and has been linked to the development of multiple cancers (Prostate, colon, ovarian, and gastric); however, there are no known inhibitors that bind specifically to HDAC4. Since the catalytic pocket in HDACs is conserved, using the currently available inhibitors invariably targets other HDACs and potentially causes toxicity. Hence to specifically target HDAC4, here we employed a new strategy to inhibit HDAC4 by disrupting its complexation with the nuclear receptor co-repressor (NCoR), which is critical for its function. Using classical and accelerated molecular dynamics (MD) simulations, we identified novel pockets in the protein-protein interface, which are not present in the available crystal structures. These pockets were then targeted using an ensemble docking approach combined with consensus scoring. Using this approach, we were able to identify 9 novel hits out of 45 suggested compounds (hit rate 20%), which successfully inhibited the catalytic activity of HDAC4 in an in-vitro assay. Out of these 9 hits, 3 compounds were found to be selective towards HDAC4 when compared to HDAC3, which belongs to a different family of HDACs (class I).

5.2.2 Introduction

Histone deacetylases (HDACs) catalyze the deacetylation of histone tail lysines, resulting in the compaction of DNA and hence the suppression of transcription³⁴³. HDAC enzymes work in opposition to Histone acetyltransferases (HAT), which

transfer an acetyl group to lysine in histone tail, and enhance transcription³⁴⁴. In a normal cell, the transcription levels regulated by HATs and HDACs are well balanced³⁴³. However, abnormality in the function of HDACs has been reported to contribute to the initiation and progression of several tumors³⁴³. Specifically, irregularity in HDAC function can cause abnormal transcription of critical genes that control vital cell functions, namely proliferation, cell cycle regulation, and apoptosis^{345, 346}. HDACs have also been thought to play a role in several other important genomic functions such as DNA repair, chromatin assembly, and recombination³⁴⁷. Hence, given their role in cellular proliferation, HDAC inhibitors have emerged to be important drugs for cancer therapy, as they can suppress abnormal histone deacetylation leading to normal levels of acetylated histones³⁴⁸. To date, four HDAC inhibitors: Vorinostat (SAHA), Romidepsin (FK-228), Belinostat (PXD-101), and Panobinostat (LBH-589) have been granted US FDA approval for cancer treatment and many other HDAC inhibitors are currently in various phases of clinical trials^{349, 350}. All these US FDA approved HDAC inhibitors target the catalytic pocket region and share common structural features that comprise a metal-binding moiety, a linker region, and a surface recognition domain^{350, 351}. The catalytic pockets in HDACs, however, is evolutionarily conserved among all zinc-dependent HDACs, thus explaining why these inhibitors are nonspecific binders³⁵⁰ and leads to off-site binding causing and hence toxicity. Hence for the structure-based design of specific inhibitors, there is a need to focus on the structural differences between HDACs and move away from the current strategies³⁵⁰.

HDACs have been divided into four major classes based on their catalytic mechanisms and sequence homology: Class I (HDAC 1, 2, 3, 8), IIa (HDAC 4, 5, 7, 9), IIb (HDAC 6, 10), III, and IV (HDAC 11)³⁵². All these HDACs, excepting class III, have a zinc (Zn^{+2}) ion in the catalytic pocket for the enzymatic activities, whereas class III HDACs or sirtuins³⁵³ require nicotinic adenine dinucleotide as a cofactor. HDACs cannot bind to DNA by themselves and exist as components in a variety of multiprotein complexes (HDAC1 and HDAC2 are found in three distinct

corepressor complexes, called SIN3A³⁵⁴, NURD/Mi2³⁵⁵ and CoREST, and HDAC3 is found in SMRT/NCoR complexes³⁵⁶). Although characterized as HDACs, class IIa enzymes lack deacetylase activity because a key tyrosine residue, which stabilizes the tetrahedral intermediate during deacetylation of the native substrate (acetylated lysine) is absent³⁵⁷. The catalytic domain contains two zincs: structural and catalytic. Structural zinc is one of the distinguishing features of Class IIa from Class I HDACs; it holds together the two loops, whereas catalytic zinc performs the deacetylation mechanism. Hence, it is believed that class IIa HDACs, especially HDAC4, function by forming multiprotein complexes with other enzymatically active class I HDACs. HDAC4 forms a complex with HDAC3-NCoR and is believed to operate as a scaffold for the recruitment of multiprotein complexes, thus increasing the range of deacetylation in specific regions of chromatin³⁵⁸.

HDAC4 has been implicated in promoting tumor growth through the suppression of p21 expression in colon cancer, glioblastoma, ovarian cancer, and gastric cancer cells³⁵⁹ and therefore is a potential drug target for anti-cancer therapy. However, there are no known inhibitors that target HDAC4 selectively. Here, we used a structure-based drug discovery approach to identify novel inhibitors that selectively bind to HDAC4, which is an appropriate approach in this case since there are two reported crystal structures of HDAC4 catalytic domain in the open and closed conformations³⁵⁷. In the open conformation, HDAC4 is bound to an inhibitor at the catalytic pocket, whereas in the closed conformation, there is a gain of function mutation of H332Y. The two loops bound to its structural zinc are closer to the active site in the closed conformation, making it different from the open conformation. Most of the enzymatic activity associated with HDAC4 expressed has been reported to be due to endogenous HDAC3. HDAC4 can shuttle between cytoplasm and nucleus and can bind to NCoR, which is also bound to enzymatically active HDAC3³⁶⁰. It has been previously reported that HDAC4 binds to the repression domain 3 (RD3) of NCoR whereas, HDAC3 binds to SANT domain of NCoR^{356, 360}. There are, however, no structures of the HDAC4-NCoR

complex; but based on mutational studies performed by Kim et al³⁶¹, 21 “hot spot” residues (C667, C669, C751, D759, T760, S767, A774, P799, P800, G801, H803, A804, F812, C813, H842, H843, G844, N845, G846, G868, and F871) present on the surface of HDAC4 have been identified as the residues that prevent its binding to the NCoR-HDAC3 complex in vivo (**Fig 5.12**). These interfacial residues are conserved across all class IIa HDACs, while some of these residues (C667, C669, C751, D759, T760, and F871) are only found in class IIa HDACs and not in class I, which form the basis for their interaction specificity with NCoR.

With this knowledge, our strategy was to focus on the HDAC4 interface with NCoR rather than targeting the catalytic pocket. Protein-protein interfaces (PPIs) have emerged as a major drug target since a large number of proteins critical in biological pathways related to various diseases, function after complex formation³⁶². Although PPIs are promising drug targets, there are major challenges in targeting interfaces since they mostly lack cavities for small molecules to bind or are intrinsically disordered³⁶³. Nevertheless, it has been shown that small molecules are capable of interacting at the position of “hot-spots” and compete with the binding of the target’s cognate partner without necessarily covering the entire PPI surface. In this study, we have used classical molecular dynamics (MD) simulation and enhanced sampling method accelerated molecular dynamics (aMD) simulation approaches to generate an ensemble of structures and to identify potential pockets in the experimentally characterized interface of HDAC4 and NCoR. aMD samples the conformational space more efficiently compared to MD, which results in receptor conformers, which would not have been sampled in a short time classical MD. Unlike other enhanced sampling methods, aMD does not require reaction coordinates or collective variables and is well suited for creating an ensemble of structures to characterize the conformational flexibility of a receptor. Using these techniques, we have identified novel pockets in the PPI which were not present in the crystal structures. These pockets have not been reported before, and we hypothesized that small molecules binding in these pockets would lead to disruption of the complex formation with NCoR. We thus

performed high throughput *in silico* screening of small molecule databases on the ensemble of structures targeting these newly identified pockets. Furthermore, we identified forty-five compounds that were predicted to bind HDAC4 in these pockets, of which nine were validated experimentally to bind HDAC4, and out of these nine, at least three compounds were selective towards HDAC4 over HDAC3.

5.2.3 Methods

System preparation

Crystal structures of HDAC4 in the open (PDB 2VQM)³⁵⁷ and in the closed states (PDB 2VQW)³⁵⁷ were used for this study. Both the protein structures and water molecules were modeled using Amber ff14SB parameters^{58, 221}. Using xleap, cysteines 667, 669, and 751 were deprotonated (-ve charge) and converted to CYM for closed conformation and cysteines 667 and 751 were converted to CYM for open conformation since they coordinate structural zinc in respective crystal structures. The open structure was crystallized in the presence of an inhibitor, which was deleted for the simulation. The closed structure was crystallized with a mutation H332Y, and so the tyrosine was mutated to wild type histidine for the simulation. The protein was surrounded by an octahedral box of water model TIP3P box of 10 Å.

Classical molecular dynamics simulation

Molecular dynamic (MD) simulations were performed on both the crystal structures. First, the HDAC4 structure was held fixed with force constant of 500 kcal mol⁻¹ Å⁻² while the system was minimized with 500 steps of steepest descent followed by 500 steps with the conjugate gradient method. In the second minimization step, the restraints on HDAC4 were removed, and 1000 steps of steepest descent minimization were performed, followed by 1500 steps of a conjugate gradient. The system was heated to 300 K while holding the protein fixed with a force constant of 10 kcal mol⁻¹ Å⁻² for 1000 steps. Then, the restraints were released, and 1000 MD steps were performed. The SHAKE algorithm was used to

constrain all bonds involving hydrogen in the simulations. A 50 ns MD was performed at 300 K using the NPT ensemble and a 2 fs time step. This procedure yielded a total of 250,000 snapshots for subsequent analyses.

Accelerated molecular dynamics simulation

The aMD simulation was performed by using a dual energy boost (i.e. Dihedral energy and Potential energy) with acceleration parameter (α) of 0.2. The average total potential energy and average dihedral energy parameters for aMD were calculated from the classical MD simulations. We performed 100 ns of accelerated MD (aMD) simulation. This procedure yielded a total of 500,000 snapshots for subsequent analyses. Snapshots from classical MD and aMD were clustered by root mean square deviation (RMSD) of hot spot residues with the hierarchical agglomerate clustering algorithm present in the CPPTRAJ module²⁰⁴.

In silico screening

We generated an ensemble of 4 representative snapshots (RS) each from classical MD and aMD for both open and closed conformations. NCI Diversity V (~1600 compounds)³⁶⁴ was docked to a total ensemble of 9 receptors (4 RS from MD, 4 RS from aMD and 1 crystal structure) with a cubic box size of ~30 Å. VinaMPI³²⁴, a parallel version of AutodockVina⁹⁰, was used to perform the *in silico* screening. The docked poses were then ranked by AutodockVina scoring function⁸⁹. The compound list was further narrowed by excluding compounds that were within 10 Å of the catalytic zinc. From this narrowed list, the top 45 ranked compounds were taken for experimental testing. The steps followed for ensemble docking are shown below in schematic (**Fig. 5.13**)

Experimental validation

The experimental validation was performed by Dr. Michael Duff at the University of Tennessee.

HDAC3 and HDAC4 activity assays were performed using fluorogenic assay kits obtained from BPS Bioscience (San Diego, CA). Compounds were prepared as 10 mM stocks in DMSO and were screened at 100 μ M in the assays.

The activity of the enzymes in the absence and presence of compounds was measured on a BioTek Cytation 5 plate reader, exciting the samples at 350 nm and measuring emission at 450 nm. IC₅₀ values were determined for the compounds that decreased HDAC activity by at least 25% relative to no inhibitor.

5.2.4 Results and Discussion

Novel pockets in the PPI

The PPI residues on HDAC4 are present in the region between the catalytic pocket and the structural zinc. In the crystal structures, this region is smooth and there are no druggable cavity present for a small molecule to bind, which is generally a major drawback of targeting PPI of complexes. In our ensemble of receptor structures generated from classical MD and aMD simulations for both open and closed conformations, we identified novel cavities in the PPI which were not present in the crystal structures (**Fig 5.14. 5.16**). These cavities were present in the “hot-spot” region and contained a cavity for small molecules to bind.

In the snapshots for closed structures generated from simulations, we observed some shallow cavities (**Fig 5.14 c, d**), and a deep pocket (~ 14 Å) near the structural zinc which were present in the PPI as shown in **Figs 5.14 and 5.16**. The opening of this novel pocket was located between the two loops with the residues (T668, H675, P676, C751, D759, T760) forming the gate of the pocket (**Fig 5.15a**). The distances between these residues were measured for the trajectory. The fluctuation in the width of the opening of this pocket has been shown by a probability distribution (**Fig 5.15b**). In the ensemble of the open structures, there were snapshots with an extended cavity with depth ~ 9 Å in the PPI region. This cavity extends ~ 16 Å from the tip of the catalytic pocket to the structural zinc (**Fig 5.15**). These pockets are important since they stretch throughout the PPI region.

These results emphasize the power of docking to an ensemble of structures as compared to traditional screens to only the crystal structure.

Virtual screening

The docking search box was made big enough to include the catalytic pocket as well as the region identified as PPI. The aim of this approach was to identify compounds that bind with a higher score in the PPI, even in the presence of a hydrophobic catalytic pocket. In our final list of compounds, we excluded all the compounds within 10 Å of the catalytic zinc and only included those compounds whose best docked poses were present in the PPI. We found small molecules that bound in the novel pockets discovered in the PPI (**Figs 5.14, 5.16**). After carefully examining the binding poses and sites, we picked 45 compounds that were able to bind in these different binding pockets. These compounds were then suggested for experimental validations.

Experimental validation

Compounds were screened for their ability to inhibit HDAC4 and their potential selectivity against HDAC3 (from a different HDAC subfamily) was also tested (**Figure 5.19**). Two compounds (67436 and 134199) decreased both HDAC3 and HDAC4 activities to 40%, or less, compared to the control condition without any inhibitor. A few of the compounds were specific towards only one of the two enzymes tested. HDAC4 was significantly inhibited by 88402 (<40% activity retained), while HDAC3 activity was only mildly perturbed (75% activity relative to no inhibitor). Compounds 34488, 44584, and 195327 decreased HDAC3 activity 50%, 60%, and 80%, respectively, while HDAC4 activity remained unchanged in the presence of those compounds (**Figure 5.19**).

The compounds that decreased the activities of the HDAC4 by at least 25% were further screened for IC₅₀ values (**Table 5.4**). Compounds 67436 and 134199 had IC₅₀ values in the low micromolar range for both HDACs, which was comparable to that of previously known inhibitors of HDAC4, trichostatin A and SAHA.

Interestingly, both these compounds were significantly weaker inhibitors of HDAC3 compared to trichostatin A and SAHA. The seemingly HDAC4-specific inhibitors 88402, 299968, and 319435 had IC₅₀ values 2-fold lower for HDAC4 compared to that for HDAC3. The latter two compounds did not inhibit HDAC3 appreciably at a concentration of at least 500 μ M, indicating that they could not inhibit HDAC3. The other compounds that weakly inhibited HDAC4 (4135, 36425, and 51936) had IC₅₀ values not much lower than the highest concentration used in the assay (500 μ M) and may not be significant inhibitors of either enzymes. From these results, we identified 9 compounds that could bind to HDAC4 and to inhibit its deacetylase activity in the assay. Out of these 9, 3 compounds were selectively inhibitory to HDAC4 when compared to HDAC3.

5.2.5 Conclusions

Histone deacetylase 4 (HDAC4) is a drug target for multiple cancers; however, all the FDA approved drugs that target HDACs are pan inhibitors and hence are not selective towards a specific sub-family. All these known inhibitors bind in the conserved catalytic pocket. Therefore, there is a need to find inhibitors selective for HDAC4. Since the catalytic pocket is conserved, selectivity is difficult to achieve by identifying small molecules that bind in the catalytic pocket by chelating the catalytic zinc. Hence, in this study, we moved away from the common strategy of targeting the catalytic pocket and focused on the protein-protein interface of HDAC4 with NCoR. This complexation is considered critical for HDAC4 function and disrupting it can lead to functional inhibition. Since the HDAC4 crystal structures had smooth PPI with no cavities/pockets that could be targeted by small molecules, we performed classical and enhanced sampling MD simulations to identify different conformations of the PPI. Indeed, we were able to identify pockets in PPI region which had not been characterized before and were not present in either of the crystal structures used. Using ensemble docking, we targeted these novel pockets in the PPI region. To further improve the hit rate and reduce the number of false positive, we used a consensus scoring scheme.

By combining computational and experimental methods, we were able to identify nine compounds (hit rate 20%) that were able to bind HDAC4 and inhibit its catalytic activity. Further experimental validations showed that out of these nine, three compounds could selectively inhibit to HDAC4 when compared to HDAC3 (from class I HDAC family). These compounds are currently being tested with Prostate and Breast cancer cell lines to measure their cell-killing activity. Moreover, these compounds can also be used as a starting point for medicinal chemistry optimization.

5.2.6 Appendix

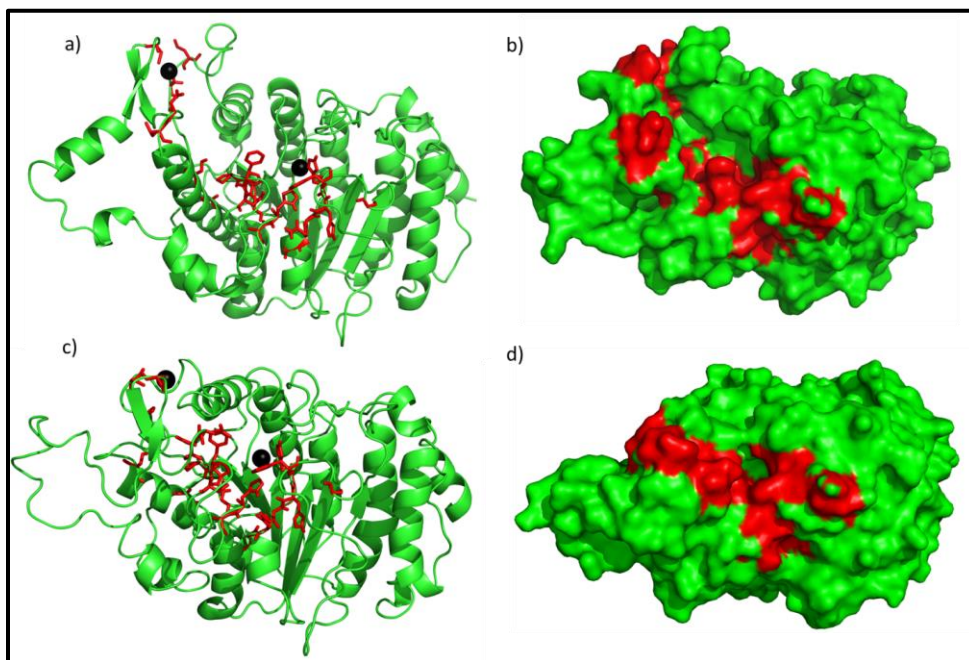


Figure 5.12 Crystal structure of open conformation³⁵⁷ (a & b) and closed conformation³⁵⁷ (c & d). The red sticks and red region represent residues identified as the “hot spot” residues identified previously³⁶¹; black sphere represents the two zincs (catalytic & structural). a & c are ribbon model of the crystal structures, and b & d are surface model.

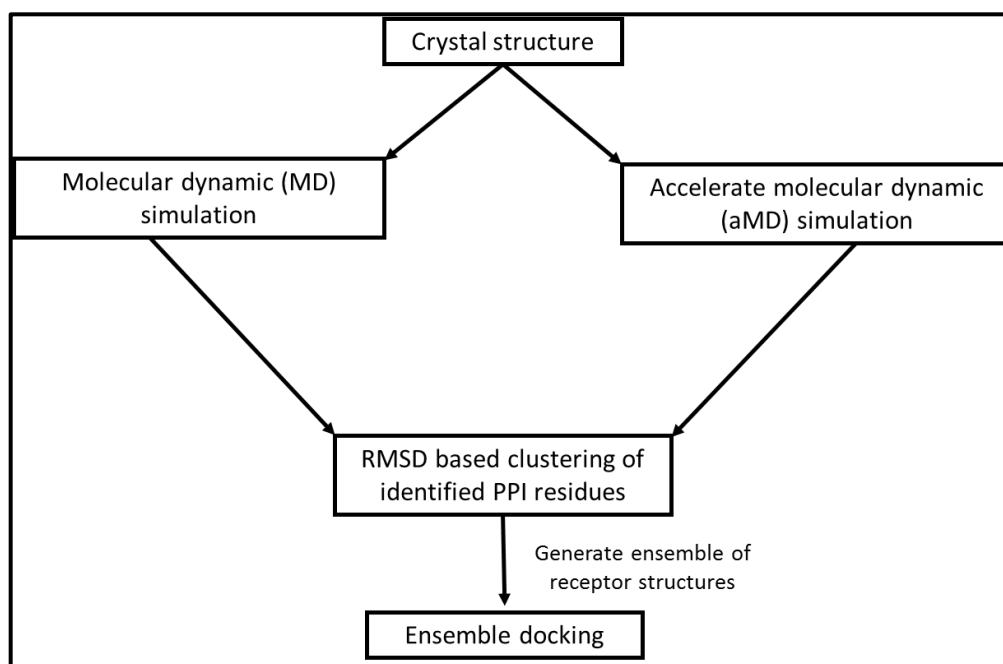


Figure 5.13 Schematic diagram explaining the ensemble docking strategy

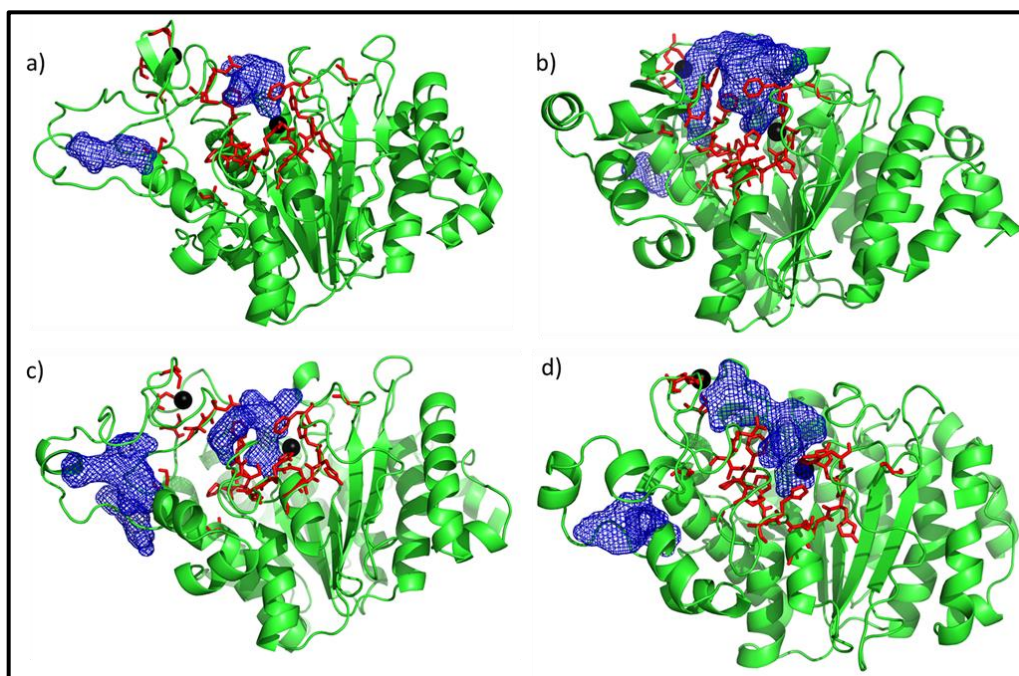


Figure 5.14 Closed conformation representative snapshots generated from simulations of crystal structures. Red sticks represent “hot spot” residues, and blue mesh represents pockets in & near the PPI region. a) Crystal structure and b, c, d) Snapshots from aMD and MD simulations.

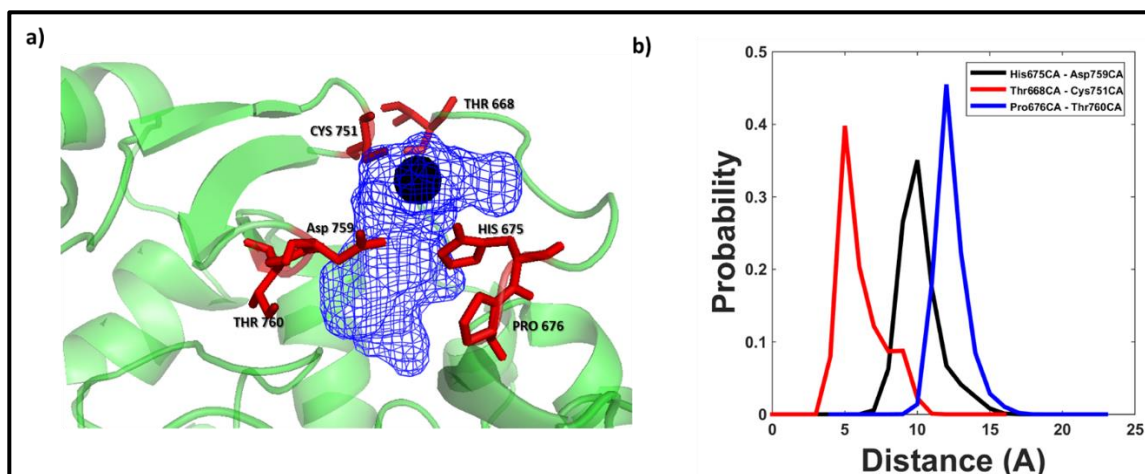


Figure 5.15 The novel pocket discovered in the close conformation (as shown in Figure 3b). a) The pocket is shown in blue mesh with residues in the two loops, which are present in the gate of this pocket, b) probability distribution of the distance of the residues.

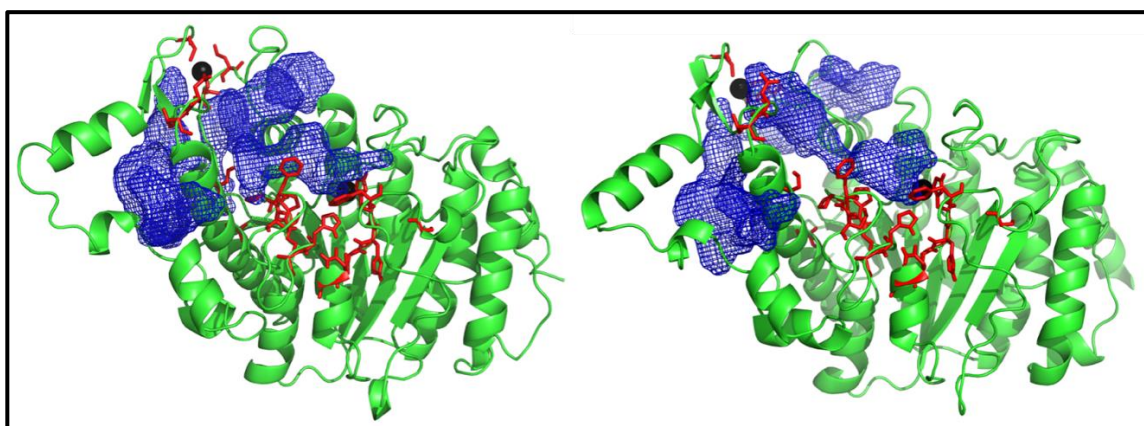


Figure 5.16 Open conformation representative snapshots generated from simulations. Red sticks represent “hot spot” residues, and blue mesh represents pockets in & near the PPI region. a) Crystal structure and b) Snapshot generated from simulations.

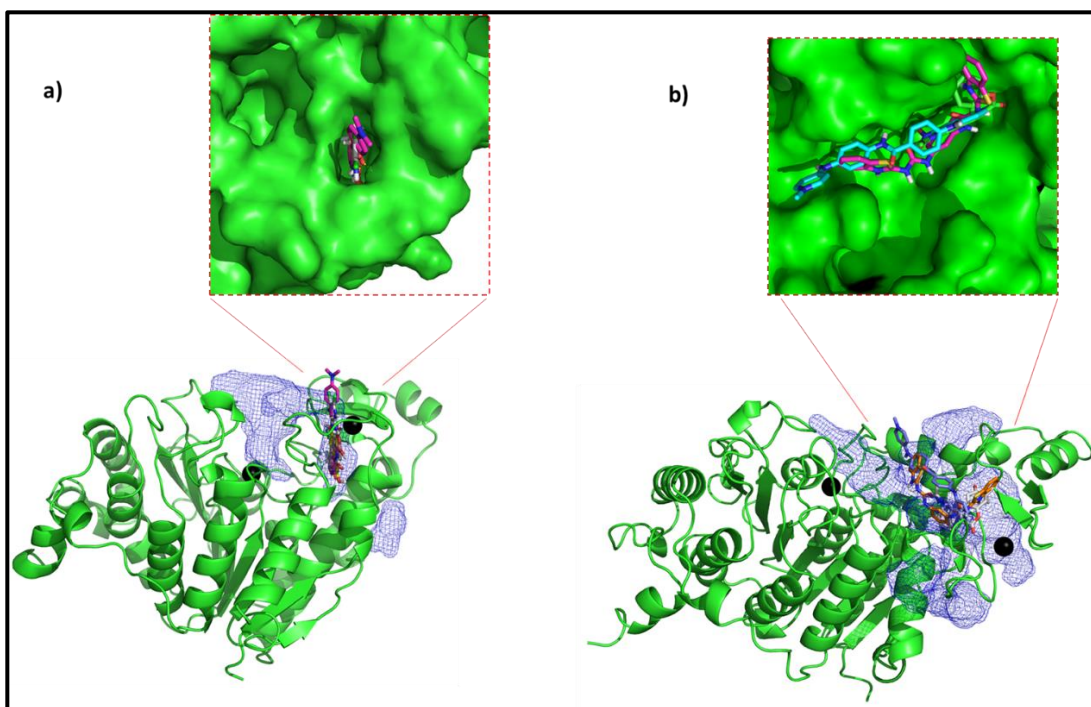


Figure 5.17 Top-ranked hits bound a) closed conformation and b) open conformation snapshot

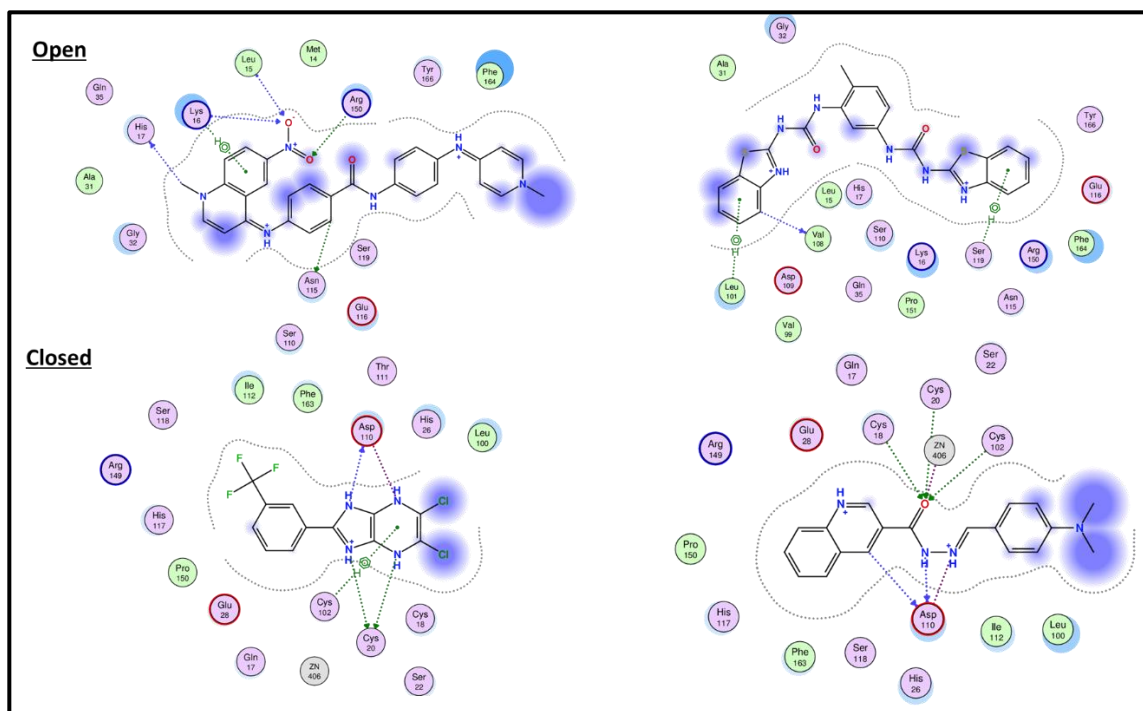


Figure 5.18 Top-ranked hits bound a) closed conformation and b) open conformation snapshot

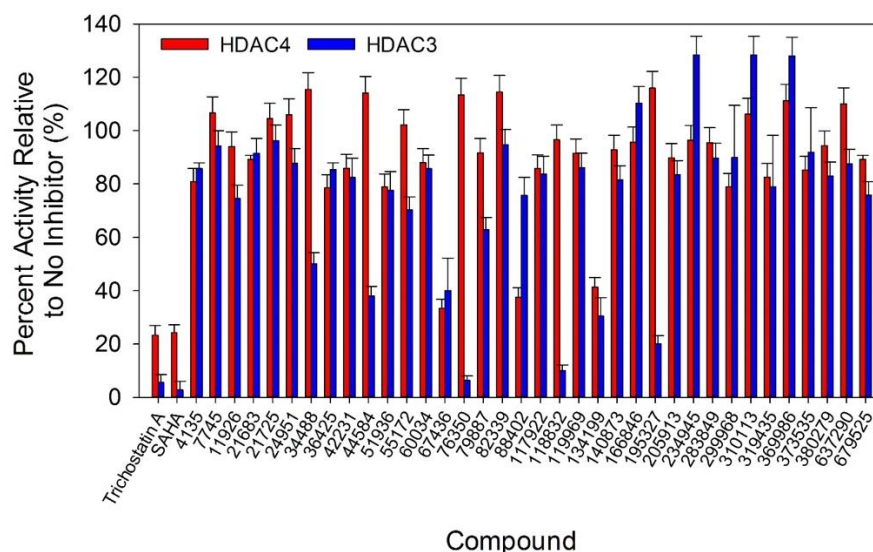


Figure 5.19 Effect of compounds on the activities of HDACs. The relative percent activity of HDAC3 (blue bars) and HDAC4 (red bars) in the presence of 100 μM inhibitor compound compared to the absence of inhibitor. Error bars are the standard deviation of at least two assays.

Table 5.4 IC₅₀ values for compounds that inhibited HDAC3 and HDAC4 (Selective inhibitors shown in green and Known inhibitors shown in Yellow)

Compound	IC ₅₀ (μM) for HDAC3	IC ₅₀ (μM) for HDAC4
Trichostatin A	0.011 $\pm$ 0.002	1.5 $\pm$ 0.8
SAHA	0.024 $\pm$ 0.006	25 $\pm$ 13
4135	>500	410 $\pm$ 240
36425	>500	480 $\pm$ 260
51936	>500	550 $\pm$ 350
67436	9.2 $\pm$ 5.2	47 $\pm$ 28
88402	210 $\pm$ 60	100 $\pm$ 40
134199	25 $\pm$ 7	33 $\pm$ 12
299968	>500	250 $\pm$ 170
319435	>500	150 $\pm$ 50

6 CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Conclusions

The motivation behind the work described in this thesis was: i) to develop a novel method to accurately parameterize small deuterated molecules for classical MD simulations; ii) to characterize the role of global and local motions in the function of adhesin proteins; iii) to examine the effect the temperature and pressure in the catalytic pocket of thermophilic inorganic pyrophosphatases and how it differs from its mesophilic homolog; iv) to identify novel inhibitors against two clinically relevant protein targets (Hsa and HDAC4). These goals were achieved using a combination of computational biophysics and experimental methods. The major contributions of the work are summarized below.

In **Chapter 2**, I demonstrated that classical MD simulations using the CHARMM forcefield contain implicit isotopic biases. I also demonstrated that careful re-parameterization of the forcefield parameters to account for the “proper” isotopic bias permitted the construction of new forcefield parameters, which can accurately represent isotopic substitutions using classical MD simulations (in the absence of nuclear quantum effects). In addition, MD simulations with these new forcefield parameters provided detailed molecular insights into the origin of the loss of the THF-water miscibility gap upon the complete deuteration of THF. These results indicated that upon deuteration, the intramolecular interactions within THF are modified, which changes the structure of the molecule and allows for an increase in the formation of THF-water hydrogen bonds. These new inter-molecular interactions then compensate for the overall reduction of favorable THF-water interactions upon heating and prevent D₈THF and water from phase separating, thus eliminating the miscibility gap.

In **Chapter 3.1**, I performed extensive MD simulations on the first crystal structures of Hsa_{BR} bound to Neu5Ac-Gal and Neu5Gc-Gal. I was able to provide insights into the mechanism behind the dual specificity of Hsa_{BR} adhesins towards the

commonly found sialic acids, Neu5Ac and Neu5Gc. Here, I demonstrated that Hsa_{BR} is able to bind Neu5Gc even in the absence of a specific tyrosine residue (present in SrpA_{BR}), which had previously been considered essential for binding Neu5Gc. Furthermore, I also demonstrated that the additional C11-OH group of Neu5Gc does not form a direct interaction with the residues in Hsa_{BR}, which seems to be the primary reason preventing its preferential binding to Neu5Gc. I also found that Hsa_{BR} did not have specific hydrophobic residues (Ile 479 and Thr 478) like GspB_{BR}, which are considered important for making GspB_{BR} selective towards Neu5Ac. Taken together, I uncovered that adhesin Hsa lacks the features present in the selective adhesins (GspB_{BR} and SrpA_{BR}). The results suggest a potential mechanism behind the observed dual specificity in Hsa towards both the sialic acids. Moreover, this study also provided insights behind the selectivity of GspB_{BR} and SrpA_{BR} adhesins towards Neu5Ac and Neu5Gc, respectively.

In **Chapter 3.2**, I have studied SRR adhesin proteins (especially GspB), which are present on the surface of *streptococci* and *staphylococci*. The adherence of the adhesin proteins (present on the bacteria) to the host platelet cells has been known to be critical for the pathogenesis of a cardiovascular disease called infective endocarditis. In our study, I have combined biochemical and computational approaches to reveal that the orientation of these adhesin proteins, specifically the binding pocket, is critical for the binding to the host platelet cells and that a slight change in the native orientation can cause significant reduction in the binding to the platelets. Additionally, I have also demonstrated that the Unique domain in these SRR adhesin proteins does not affect the binding pocket but plays an important role in maintaining the required rigidity in the inter-domain so as to retain the binding pocket orientation.

In **Chapter 4**, I have compared the structural and dynamic properties of two pyrophosphatases- a hyperthermophilic (*Tt*PPase) and a mesophilic (*Ec*PPase) pyrophosphatase using MD simulations mimicking both deep-sea and ambient

conditions. The results of this study indicated that *Tt*PPase has been “designed” to function at high temperatures and exhibits a comparatively more compact structure with reduced number of intra-protein hydrogen bonds in the catalytic pocket. However, simulations of *Tt*PPase at its non-native conditions revealed that it fails to preserve both of these properties at room temperatures, and thus becomes enzymatically inactive. Interestingly, I found that on the contrary, *Ec*PPase adapts and retains its activity at high temperatures by incorporating the strategies used by *Tt*PPase to function at high temperatures.

In **Chapter 5.1**, I performed structure-based virtual screening to identify competitive small molecule effectors against Hsa_{BR}. I have combined three different SBDD strategies: ensemble docking, cross screening, and consensus scoring in one pipeline. For ensemble docking, I generated an ensemble of receptor conformations from MD simulations, and also performed cross screening against five other homologs (GspB_{BR}, 10712_{BR}, SK150_{BR}, SrpA_{BR}, SK678_{BR}). In the last step, three scoring functions (AutodockVina and MOE) were used to rank and prioritize the list of compounds obtained after ensemble docking and cross-screening. The Vanderbilt database was used for the small molecules since it covers a wide distribution of different physicochemical properties. The goal of combining these strategies was to improve the hit rate and reduce the number of false positives. Indeed, I was able to achieve a hit rate of ~20% and identified nine compounds that could displace the native ligand in experimental assays. The binding poses of all the nine compounds identified from docking reveal that they are in close proximity of the residues known to form hydrogen bonds with the native ligand (sialyl-T antigen). These compounds may be used as a starting point for medicinal chemistry optimization.

In **Chapter 5.2**, I targeted histone deacetylase 4 (HDAC4), which is a cancer drug target. All the FDA approved drugs that target HDACs are pan inhibitors and hence are not selective towards a specific sub-family. All these known inhibitors bind in

the conserved catalytic pocket. Therefore, there is a need to find selective inhibitors selective for HDAC4. Since the catalytic pocket is conserved, selectivity is difficult to achieve by identifying small molecules that bind in the catalytic pocket. Hence, in this study, we moved away from the common strategy of targeting the catalytic pocket and focused on the protein-protein interface (PPI) of HDAC4 with NCoR. This complexation is critical for protein function and disrupting it can lead to inhibition. However, the PPI interface present in the crystal structures are smooth and do not contain any cavities/pockets that can be targeted by small molecules. Therefore, I performed classical and enhanced sampling MD simulations to identify different conformations of the PPI. Indeed, I was able to identify pockets in PPI region which have not been characterized before and are not present in either of the crystal structures. Using ensemble docking, we targeted these novel pockets present in the PPI region. To further improve the hit rate and reduce number of false positive, we used a consensus scoring scheme. By combining computational and experimental methods, I was able to identify nine compounds (hit rate of 20%) which bind to HDAC4 and inhibit the catalytic activity. Furthermore, three compounds (out of the nine) can selectively inhibit HDAC4 when compared to HDAC3 (from class I HDAC family).

6.2 Future directions

The work done in this thesis has contributed to the field of computational biophysics, drug discovery, protein evolution, and dynamics. The findings described in this thesis also generate more ideas which can be extended with additional computational and experimental studies. Below is a brief list of potential ideas and some interesting research topics, which are worth investigating further.

The success of reparameterization approach to accurately capture deuteration effect shown in **Chapter 2** suggests that a similar methodology may be implemented in MD simulations of other deuterated systems. The methodology presented here can be employed in a drug discovery pipeline of small deuterated molecules that are of growing importance as pharmaceuticals^{174, 175}, especially since the first deuterated drug Austedo (deutetrabenazine), from Teva, was FDA approved in 2017 to target Huntington's chorea³⁶⁵. Additionally, accurate modeling of the small deuterated molecules will also be critical to target a protein/RNA/DNA. Moreover, this methodology can serve as an aid in the interpretation of experiments such as NMR¹⁷⁶ and neutron scattering¹⁷⁷ where isotope substitution is of critical importance.

The local loop motion in the binding pocket of adhesin proteins provided an insight into the binding mechanism of different glycan molecules. The binding of adhesin proteins to platelet glycans is known to be critical for infective endocarditis. Therefore, knowing the binding of the native glycan ligands to these adhesins can lead to new strategies to target the protein. A series of mutagenesis experiments and simulations can be performed to further validate some of the predictions made in **Chapter 3.1**. Moreover, adhesin proteins have also been considered as probes to detect sialoglycans that are presented on the surfaces of cancer cells as potential biomarkers. Knowledge obtained from this study can further be used to engineer the binding pocket to make these adhesin proteins selective towards particular sialoglycans for use as biomarkers. Additionally, we showed that the

binding pocket orientation of adhesin protein on the surface of the bacterial cell is important in **Chapter 3.2**. This knowledge can also be used as a novel strategy to target adhesin proteins. Moreover, a similar study should be performed for other adhesin proteins and surface proteins to explore if structural domain rigidity and precise orientations are conserved.

The effect of non-native condition on mesophilic and thermophilic inorganic pyrophosphatase homologs was studied in **Chapter 4**. To take this work further, experiments like mutagenesis and measurement of enzymatic activities can be performed to validate the predictions made using MD simulations. We introduced a potential mechanism on how a mesophilic enzyme is more adaptive and can function at both low and high temperatures. This mechanism can be tested on other homologous pairs to evaluate if it is widespread or if it is unique for these particular type of proteins. Moreover, further study can be done to understand the effect of solvation on the dynamics of proteins.

Lastly, we have identified nine novel competitive hits that target the adhesin protein Hsa in **Chapter 5.1**. Further studies need to be conducted to characterize the binding affinities and poses of these identified compounds and similar analyses can be performed for other sialoglycan-binding SRR proteins. With the known hits, the plausible next step would be to crystalize them with the protein and perform more rigorous free energy calculations before moving to the next step of lead optimization. Another possible step is to experimentally test for the selectivity of these nine hits against other human and bacterial sialoglycans binding proteins. In the case of HDAC4, we have identified hits that are selective and potentially do not bind in the catalytic pocket as described in **Chapter 5.2**. The next steps would be similar to those for Hsa. In addition, these selective compounds should be screened against various cancer cell lines. Since in this study I was able to potentially find novel compounds which do not bind in the conserved catalytic

pocket, the novel scaffold can also be further explored to find more hits (i.e., hit expansion).

REFERENCES

1. Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins. *Nature* **2007**, *450* (7172), 964-72.
2. Henzler-Wildman, K. A.; Lei, M.; Thai, V.; Kerns, S. J.; Karplus, M.; Kern, D., A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* **2007**, *450* (7171), 913-6.
3. Socher, E.; Sticht, H., Mimicking titration experiments with MD simulations: A protocol for the investigation of pH-dependent effects on proteins. *Sci Rep* **2016**, *6* (1), 22523.
4. Calligari, P. A.; Calandrini, V.; Ollivier, J.; Artero, J. B.; Hartlein, M.; Johnson, M.; Kneller, G. R., Adaptation of Extremophilic Proteins with Temperature and Pressure: Evidence from Initiation Factor 6. *J Phys Chem B* **2015**, *119* (25), 7860-73.
5. Agarwal, R.; Shrestha, U. R.; Chu, X. Q.; Petridis, L.; Smith, J. C., Mesophilic Pyrophosphatase Function at High Temperature: A Molecular Dynamics Simulation Study. *Biophys J* **2020**, *119* (1), 142-150.
6. Frauenfelder, H.; Alberding, N. A.; Ansari, A.; Braunstein, D.; Cowen, B. R.; Hong, M. K.; Iben, I. E. T.; Johnson, J. B.; Luck, S.; Marden, M. C.; Mourant, J. R.; Ormos, P.; Reinisch, L.; Scholl, R.; Schulte, A.; Shyamsunder, E.; Sorensen, L. B.; Steinbach, P. J.; Xie, A. H.; Young, R. D.; Yue, K. T., Proteins and Pressure. *J Phys Chem-Us* **1990**, *94* (3), 1024-1037.
7. Tao, Y.; Rao, Z. H.; Liu, S. Q., Insight derived from molecular dynamics simulation into substrate-induced changes in protein motions of proteinase K. *J Biomol Struct Dyn* **2010**, *28* (2), 143-58.
8. Liu, S. Q.; Tao, Y.; Meng, Z. H.; Fu, Y. X.; Zhang, K. Q., The effect of calciums on molecular motions of proteinase K. *J Mol Model* **2011**, *17* (2), 289-300.
9. Wilkens, S., Structure and mechanism of ABC transporters. *F1000Prime Rep* **2015**, *7*, 14.
10. Kolomeisky, A. B., Motor proteins and molecular motors: how to operate machines at the nanoscale. *J Phys Condens Matter* **2013**, *25* (46), 463101.
11. Oide, M.; Kato, T.; Oroguchi, T.; Nakasako, M., Energy landscape of domain motion in glutamate dehydrogenase deduced from cryo-electron microscopy. *FEBS J* **2020**.
12. Aglietti, R. A.; Floor, S. N.; McClendon, C. L.; Jacobson, M. P.; Gross, J. D., Active site conformational dynamics are coupled to catalysis in the mRNA decapping enzyme Dcp2. *Structure* **2013**, *21* (9), 1571-80.
13. Pal, N.; Wu, M.; Lu, H. P., Probing conformational dynamics of an enzymatic active site by an in situ single fluorogenic probe under piconewton force manipulation. *Proc Natl Acad Sci U S A* **2016**, *113* (52), 15006-15011.
14. Nelson, D. L.; Cox, M. M., *Lehninger Principles of Biochemistry*. 6th ed.; W.H. Freeman: 2012.
15. Berg, J. M.; Tymoczko, J. L.; Stryer, L., *Protein Structure and Function*. W H Freeman: 2002.
16. 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. *Chem Rev* **2016**, *116* (11), 6516-51.
17. Austin, R. H.; Beeson, K. W.; Eisenstein, L.; Frauenfelder, H.; Gunsalus, I. C., Dynamics of ligand binding to myoglobin. *Biochemistry* **1975**, *14* (24), 5355-73.
 18. Frauenfelder, H.; Leeson, D. T., The energy landscape in non-biological and biological molecules. *Nat Struct Biol* **1998**, *5* (9), 757-9.
 19. Frauenfelder, H.; Parak, F.; Young, R. D., Conformational substates in proteins. *Annu Rev Biophys Biophys Chem* **1988**, *17*, 451-79.
 20. Frauenfelder, H.; Petsko, G. A.; Tsernoglou, D., Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature* **1979**, *280* (5723), 558-63.
 21. Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G., The energy landscapes and motions of proteins. *Science* **1991**, *254* (5038), 1598-603.
 22. Lazaridis, T.; Karplus, M., "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science* **1997**, *278* (5345), 1928-31.
 23. Leopold, P. E.; Montal, M.; Onuchic, J. N., Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc Natl Acad Sci U S A* **1992**, *89* (18), 8721-5.
 24. Wolynes, P. G., Recent successes of the energy landscape theory of protein folding and function. *Q Rev Biophys* **2005**, *38* (4), 405-10.
 25. RH, A.; KW, B.; L, E.; H, F.; IC, G., Dynamics of ligand binding to myoglobin. *Biochemistry* **1975**, *14* (24).
 26. Brooks, C. L., 3rd; Karplus, M., Solvent effects on protein motion and protein effects on solvent motion. Dynamics of the active site region of lysozyme. *J Mol Biol* **1989**, *208* (1), 159-81.
 27. Dill, K. A., Polymer principles and protein folding. *Protein Sci* **1999**, *8* (6), 1166-80.
 28. Liu, S.-Q.; Ji, X.-L.; Tao, Y.; Tan, D.-Y.; Zhang, K.-Q.; Fu, Y.-X., Protein Folding, Binding and Energy Landscape: A Synthesis. In *Protein Engineering*, InTech: 2012.
 29. Yang, L. Q.; Sang, P.; Tao, Y.; Fu, Y. X.; Zhang, K. Q.; Xie, Y. H.; Liu, S. Q., Protein dynamics and motions in relation to their functions: several case studies and the underlying mechanisms. *J Biomol Struct Dyn* **2014**, *32* (3), 372-93.
 30. Yang, L. Q.; Ji, X. L.; Liu, S. Q., The free energy landscape of protein folding and dynamics: a global view. *J Biomol Struct Dyn* **2013**, *31* (9), 982-92.
 31. Single Molecule Protein Dynamics | The Haran Group. <https://www.weizmann.ac.il/chemphys/cfharan/research-activities/single-molecule-protein-dynamics-0>.
 32. Jaskolski, M.; Dauter, Z.; Wlodawer, A., A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits. *FEBS J* **2014**, *281* (18), 3985-4009.
 33. Srajer, V.; Schmidt, M., Watching Proteins Function with Time-resolved X-ray Crystallography. *J Phys D Appl Phys* **2017**, *50* (37).

34. Levantino, M.; Yorke, B. A.; Monteiro, D. C.; Cammarata, M.; Pearson, A. R., Using synchrotrons and XFELs for time-resolved X-ray crystallography and solution scattering experiments on biomolecules. *Curr Opin Struct Biol* **2015**, *35*, 41-8.
35. Neutze, R.; Branden, G.; Schertler, G. F., Membrane protein structural biology using X-ray free electron lasers. *Curr Opin Struct Biol* **2015**, *33*, 115-25.
36. Ourmazd, A., Cryo-EM, XFELs and the structure conundrum in structural biology. *Nat Methods* **2019**, *16* (10), 941-944.
37. Murata, K.; Wolf, M., Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochim Biophys Acta Gen Subj* **2018**, *1862* (2), 324-334.
38. Nogales, E.; Scheres, S. H., Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. *Mol Cell* **2015**, *58* (4), 677-89.
39. Shoemaker, S. C.; Ando, N., X-rays in the Cryo-Electron Microscopy Era: Structural Biology's Dynamic Future. *Biochemistry* **2018**, *57* (3), 277-285.
40. Englander, S. W., Hydrogen exchange and mass spectrometry: A historical perspective. *J Am Soc Mass Spectrom* **2006**, *17* (11), 1481-1489.
41. Bai, Y., Protein folding pathways studied by pulsed- and native-state hydrogen exchange. *Chem Rev* **2006**, *106* (5), 1757-68.
42. Baker, L. A.; Baldus, M., Characterization of membrane protein function by solid-state NMR spectroscopy. *Curr Opin Struct Biol* **2014**, *27*, 48-55.
43. Jiang, Y.; Kalodimos, C. G., NMR Studies of Large Proteins. *J Mol Biol* **2017**, *429* (17), 2667-2676.
44. Opella, S. J.; Marassi, F. M., Applications of NMR to membrane proteins. *Arch Biochem Biophys* **2017**, *628*, 92-101.
45. Korasick, D. A.; Tanner, J. J., Determination of protein oligomeric structure from small-angle X-ray scattering. *Protein Sci* **2018**, *27* (4), 814-824.
46. Kikhney, A. G.; Svergun, D. I., A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett* **2015**, *589* (19 Pt A), 2570-7.
47. Vestergaard, B., Analysis of biostructural changes, dynamics, and interactions - Small-angle X-ray scattering to the rescue. *Arch Biochem Biophys* **2016**, *602*, 69-79.
48. Slabinski, L.; Jaroszewski, L.; Rodrigues, A. P.; Rychlewski, L.; Wilson, I. A.; Lesley, S. A.; Godzik, A., The challenge of protein structure determination--lessons from structural genomics. *Protein Sci* **2007**, *16* (11), 2472-82.
49. Karplus, M.; Kuriyan, J., Molecular dynamics and protein function. *Proc Natl Acad Sci U S A* **2005**, *102* (19), 6679-85.
50. Karplus, M.; McCammon, J. A., Molecular dynamics simulations of biomolecules. *Nat Struct Biol* **2002**, *9* (9), 646-52.
51. Barducci, A.; Bonomi, M.; Prakash, M. K.; Parrinello, M., Free-energy landscape of protein oligomerization from atomistic simulations. *Proc Natl Acad Sci U S A* **2013**, *110* (49), E4708-13.

52. Tavernelli, I.; Cotesta, S.; Di Iorio, E. E., Protein dynamics, thermal stability, and free-energy landscapes: a molecular dynamics investigation. *Biophys J* **2003**, 85 (4), 2641-9.
53. Maisuradze, G. G.; Liwo, A.; Scheraga, H. A., Relation between free energy landscapes of proteins and dynamics. *J Chem Theory Comput* **2010**, 6 (2), 583-595.
54. Vanommeslaeghe, K.; Guvench, O.; MacKerell, A. D., Jr., Molecular mechanics. *Curr Pharm Des* **2014**, 20 (20), 3281-92.
55. Huber, G. A.; Kim, S., Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys J* **1996**, 70 (1), 97-110.
56. Zuckerman, D. M.; Chong, L. T., Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annu Rev Biophys* **2017**, 46, 43-57.
57. Case, A. D., *AMBER*. 2010.
58. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *Journal of computational chemistry* **2005**, 26 (16), 1668-88.
59. Huang, J.; MacKerell, A. D., Jr., CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *Journal of computational chemistry* **2013**, 34 (25), 2135-45.
60. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmuller, H.; MacKerell, A. D., CHARMM36: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Biophysical Journal* **2017**, 112 (3), 175a-176a.
61. Pol-Fachin, L.; Rusu, V. H.; Verli, H.; Lins, R. D., GROMOS 53A6GLYC, an Improved GROMOS Force Field for Hexopyranose-Based Carbohydrates. *J Chem Theory Comput* **2012**, 8 (11), 4681-90.
62. Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996**, 118 (45), 11225-11236.
63. Lopes, P. E.; Guvench, O.; MacKerell, A. D., Jr., Current status of protein force fields for molecular dynamics simulations. *Methods Mol Biol* **2015**, 1215, 47-71.
64. Shu, F.; Ramakrishnan, V.; Schoenborn, B. P., Enhanced visibility of hydrogen atoms by neutron crystallography on fully deuterated myoglobin. *Proc Natl Acad Sci U S A* **2000**, 97 (8), 3872-7.
65. Nakamura, A.; Ishida, T.; Samejima, M.; Igarashi, K., The use of neutron scattering to determine the functional structure of glycoside hydrolase. *Curr Opin Struct Biol* **2016**, 40, 54-61.
66. Sattler, M.; Fesik, S. W., Use of deuterium labeling in NMR: Overcoming a sizeable problem. *Structure* **1996**, 4 (11), 1245-1249.
67. Bateman, J. E.; Eagling, R. D.; Horrocks, B. R.; Houlton, A., A deuterium labeling, FTIR, and ab initio investigation of the solution-phase thermal reactions

- of alcohols and alkenes with hydrogen-terminated silicon surfaces. *Journal of Physical Chemistry B* **2000**, *104* (23), 5557-5565.
68. Sugita, Y.; Okamoto, Y., Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999**, *314* (1-2), 141-151.
 69. Ostermeir, K.; Zacharias, M., Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins. *Biochim Biophys Acta* **2013**, *1834* (5), 847-53.
 70. Laio, A.; Gervasio, F. L., Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics* **2008**, *71* (12), 126601.
 71. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proc Natl Acad Sci U S A* **2002**, *99* (20), 12562-6.
 72. Hamelberg, D.; Mongan, J.; McCammon, J. A., Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J Chem Phys* **2004**, *120* (24), 11919-29.
 73. Pierce, L. C.; Salomon-Ferrer, R.; Augusto, F. d. O. C.; McCammon, J. A.; Walker, R. C., Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *J Chem Theory Comput* **2012**, *8* (9), 2997-3002.
 74. Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, O.; McCammon, J. A.; Miao, Y.; Smith, J. C., Ensemble Docking in Drug Discovery. *Biophys J* **2018**, *114* (10), 2271-2278.
 75. von Itzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Van Phan, T.; Smythe, M. L.; White, H. F.; Oliver, S. W.; et al., Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, *363* (6428), 418-23.
 76. Davies li, J. F., VIRACEPT (R) (NELFINAVIR MESYLATE, AG1343): A POTENT ORALLY BIOAVAILABLE INHIBITOR OF HIV-1 PROTEASE. Worldwide Protein Data Bank: 1998.
 77. Elber, R.; Karplus, M., Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science* **1987**, *235* (4786), 318-21.
 78. Carlson, H. A.; Masukawa, K. M.; McCammon, J. A., Method for including the dynamic fluctuations of a protein in computer-aided drug design. *Journal of Physical Chemistry A* **1999**, *103* (49), 10213-10219.
 79. Amaro, R. E.; Baron, R.; McCammon, J. A., An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J Comput Aided Mol Des* **2008**, *22* (9), 693-705.
 80. Pradhan, M. R.; Siau, J. W.; Kannan, S.; Nguyen, M. N.; Ouaray, Z.; Kwoh, C. K.; Lane, D. P.; Ghadessy, F.; Verma, C. S., Simulations of mutant p53 DNA binding domains reveal a novel druggable pocket. *Nucleic Acids Res* **2019**, *47* (4), 1637-1652.
 81. Di Pietro, O.; Juarez-Jimenez, J.; Munoz-Torrero, D.; Laughton, C. A.; Luque, F. J., Unveiling a novel transient druggable pocket in BACE-1 through molecular simulations: Conformational analysis and binding mode of multisite inhibitors. *PLoS One* **2017**, *12* (5), e0177683.

82. Hughes, J. P.; Rees, S.; Kalindjian, S. B.; Philpott, K. L., Principles of early drug discovery. *Br J Pharmacol* **2011**, 162 (6), 1239-49.
83. Lavecchia, A.; Di Giovanni, C., Virtual screening strategies in drug discovery: a critical review. *Curr Med Chem* **2013**, 20 (23), 2839-60.
84. Vyas, V. K.; Ukawala, R. D.; Ghatge, M.; Chintha, C., Homology Modeling a Fast Tool for Drug Discovery: Current Perspectives. In *Indian J Pharm Sci*, 2012; Vol. 74, pp 1-17.
85. Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S., The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc* **2015**, 10 (5), 733-55.
86. *Molecular Operating Environment (MOE)* 2016; Chemical Computing Group Inc: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013.08.
87. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* **2012**, 52 (7), 1757-68.
88. Enamine real database: Making chemical diversity real. **2020**.
89. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J., AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry* **2009**, 30 (16), 2785-91.
90. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **2010**, 31 (2), 455-61.
91. Cramer, J. W.; Mattioni, B. E.; Savin, K. A., Strategies for conducting ADME studies during lead generation in the drug discovery process. *IDrugs* **2010**, 13 (12), 857-61.
92. Mitra, S.; Mukhopadhyay, R.; Venu, K., Molecular motions in liquid crystal BBBA (4O.4): QENS study. *Chemical Physics* **2000**, 261 (1-2), 149-156.
93. Panesar, K. S.; Judeinstein, P.; Zanotti, J. M., Selective deuteration reveals interference caused by side-chain dynamics on measurements of self-diffusion in ionic liquid cations. *Journal of the Physical Society of Japan* **2013**, 82 (Suppl. A), SA013.
94. Jacrot, B., The study of biological structures by neutron scattering from solution. *Reports on progress in physics* **1976**, 39 (10), 911.
95. Smith, R. L.; Oldfield, E., Dynamic structure of membranes by deuterium NMR. *Science* **1984**, 225 (4659), 280-8.
96. Roder, H.; Elove, G. A.; Englander, S. W., Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR. *Nature* **1988**, 335 (6192), 700-4.
97. Hofstetter, K.; Hinterstoesser, B.; Salmen, L., Moisture uptake in native cellulose - the roles of different hydrogen bonds: a dynamic FT-IR study using Deuterium exchange. *Cellulose* **2006**, 13 (2), 131-145.

98. Gericke, A.; Moore, D. J.; Erukulla, R. K.; Bittman, R.; Mendelsohn, R., Partially deuterated phospholipids as IR structure probes of conformational order in bulk and monolayer phases. *Journal of Molecular Structure* **1996**, 379 (1-3), 227-239.
99. Thielges, M. C.; Case, D. A.; Romesberg, F. E., Carbon-deuterium bonds as probes of dihydrofolate reductase. *Journal of the American Chemical Society* **2008**, 130 (20), 6597-6603.
100. Zimmermann, J.; Gundogdu, K.; Cremeens, M. E.; Bandaria, J. N.; Hwang, G. T.; Thielges, M. C.; Cheatum, C. M.; Romesberg, F. E., Efforts toward developing probes of protein dynamics: vibrational dephasing and relaxation of carbon-deuterium stretching modes in deuterated leucine. *J Phys Chem B* **2009**, 113 (23), 7991-4.
101. Dollinger, G.; Eisenstein, L.; Lin, S. L.; Nakanishi, K.; Termini, J., Fourier transform infrared difference spectroscopy of bacteriorhodopsin and its photoproducts regenerated with deuterated tyrosine. *Biochemistry* **1986**, 25 (21), 6524-33.
102. Trapp, C.; Kulkarni, S. V., Isotope Effects in the Electron-Spin Resonance-Spectra of Deuterated Triphenylmethyl Radicals. *J Phys Chem-Us* **1984**, 88 (13), 2703-2708.
103. McCloskey, J. A., [16] Introduction of deuterium by exchange for measurement by mass spectrometry. In *Methods in enzymology*, Elsevier: 1990; Vol. 193, pp 329-338.
104. Mendelsohn, R.; Koch, C. C., Deuterated phospholipids as Raman spectroscopic probes of membrane structure: phase diagrams for the dipalmitoyl phosphatidylcholine (and its d62 derivative)-dipalmitoyl phosphatidylethanolamine system. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **1980**, 598 (2), 260-271.
105. Shi, C.; Zhang, X.; Yu, C. H.; Yao, Y. F.; Zhang, W., Geometric isotope effect of deuteration in a hydrogen-bonded host-guest crystal. *Nat Commun* **2018**, 9 (1), 481.
106. Rao, C. N. R., Effect of Deuteration on Hydrogen-Bonds. *J Chem Soc Farad T 1* **1975**, 71 (4), 980-983.
107. Habershon, S.; Fanourgakis, G. S.; Manolopoulos, D. E., Comparison of path integral molecular dynamics methods for the infrared absorption spectrum of liquid water. *J Chem Phys* **2008**, 129 (7), 074501.
108. Ohta, Y.; Ohta, K.; Kinugawa, K., Ab initio centroid path integral molecular dynamics: Application to vibrational dynamics of diatomic molecular systems. **2003**.
109. Li, Y.; Suleimanov, Y. V.; Green, W. H.; Guo, H., Quantum rate coefficients and kinetic isotope effect for the reaction $\text{Cl} + \text{CH}_4 \rightarrow \text{HCl} + \text{CH}_3$ from ring polymer molecular dynamics. *J Phys Chem A* **2014**, 118 (11), 1989-96.
110. Suleimanov, Y. V.; Aoiz, F. J.; Guo, H., Chemical Reaction Rate Coefficients from Ring Polymer Molecular Dynamics: Theory and Practical Applications. *J Phys Chem A* **2016**, 120 (43), 8488-8502.

111. Bryant, G.; Taylor, M. B.; Darwish, T. A.; Krause-Heuer, A. M.; Kent, B.; Garvey, C. J., Effect of deuteration on the phase behaviour and structure of lamellar phases of phosphatidylcholines - Deuterated lipids as proxies for the physical properties of native bilayers. *Colloids Surf B Biointerfaces* **2019**, *177*, 196-203.
112. Balevicius, V.; Weiden, N.; Weiss, A., Deuteron Nmr Relaxation, Phase-Diagrams, and Isotope Effects in Liquid-Mixtures of Tetrahydrofuran D₂O Salt. *Ber Bunsen Phys Chem* **1994**, *98* (6), 785-792.
113. Oleinikova, A.; Weingartner, H., Isotope effect on the coexistence curve and crossover behavior of water plus tetrahydrofuran. *Chemical Physics Letters* **2000**, *319* (1-2), 119-124.
114. Kucerka, N.; Holland, B. W.; Gray, C. G.; Tomberli, B.; Katsaras, J., Scattering density profile model of POPG bilayers as determined by molecular dynamics simulations and small-angle neutron and X-ray scattering experiments. *The Journal of Physical Chemistry B* **2011**, *116* (1), 232-239.
115. Pan, J. J.; Cheng, X. L.; Sharp, M.; Ho, C. S.; Khadka, N.; Katsaras, J., Structural and mechanical properties of cardiolipin lipid bilayers determined using neutron spin echo, small angle neutron and X-ray scattering, and molecular dynamics simulations. *Soft Matter* **2015**, *11* (1), 130-138.
116. Guarini, E.; Sampoli, M.; Venturi, G.; Bafile, U.; Barocchi, F., Inelastic Neutron Scattering and Molecular Dynamics Determination of the Interaction Potential in Liquid CD₄. *Physical review letters* **2007**, *99* (16), 167801.
117. Balog, E.; Hughes, A. L.; Martyna, G. J., Constant pressure path integral molecular dynamics studies of quantum effects in the liquid state properties of n-alkanes. *Journal of Chemical Physics* **2000**, *112* (2), 870-880.
118. Troitzsch, R. Z.; Martyna, G. J.; McLain, S. E.; Soper, A. K.; Crain, J., Structure of aqueous proline via parallel tempering molecular dynamics and neutron diffraction. *J Phys Chem B* **2007**, *111* (28), 8210-22.
119. Zeidler, A.; Salmon, P. S.; Fischer, H. E.; Neuefeind, J. C.; Simonson, J. M.; Markland, T. E., Isotope effects in water as investigated by neutron diffraction and path integral molecular dynamics. *Journal of Physics-Condensed Matter* **2012**, *24* (28), 284126.
120. Wang, L.; Jakowski, J.; Garashchuk, S.; Sumpter, B. G., Understanding How Isotopes Affect Charge Transfer in P3HT/PCBM: A Quantum Trajectory-Electronic Structure Study with Nonlinear Quantum Corrections. *J Chem Theory Comput* **2016**, *12* (9), 4487-500.
121. Szalay, V., Eckart-Sayvetz conditions revisited. *J Chem Phys* **2014**, *140* (23), 234107.
122. Hiratsuka, M.; Ohmura, R.; Sum, A. K.; Yasuoka, K., Vibrational spectra of deuterated methane and water molecules in structure I clathrate hydrate from ab initio MD simulation. <http://dx.doi.org/10.1080/08927022.2014.929125> **2014**.
123. Vanommeslaeghe, K.; Yang, M.; MacKerell, A. D., Jr., Robustness in the fitting of molecular mechanics parameters. *Journal of computational chemistry* **2015**, *36* (14), 1083-101.

124. Kern, R. J., Tetrahydrofuran Complexes of Transition Metal Chlorides. *J Inorg Nucl Chem* **1962**, 24 (Dec), 1105-1109.
125. Yoshida, S.; Suga, H.; Seki, S., Thermodynamic Studies of Solid Polyethers 0.3. Poly(Tetrahydrofuran), $[-(\text{CH}_2)_4\text{O}-]_n$. *Polym. J.* **1973**, 5 (1), 25.
126. Naemura, K.; Ebashi, I.; Matsuda, A.; Chikamatsu, H., Enantiomer Recognition of Crown Ethers and Open-Chain Polyethers Containing the Trans-Tetrahydrofuran-2,5-Diylbis(Methylene) Subunit as the Chiral Center. *Journal of the Chemical Society-Chemical Communications* **1986**, 9 (9), 666-668.
127. Theiler, S.; Hovetborn, T.; Keul, H.; Moller, M., Synthesis and Characterization of Amphiphilic Polyethers Based on Tetrahydrofuran and Glycidol: Antibacterial Assessment. *Macromolecular Chemistry and Physics* **2009**, 210 (8), 614-630.
128. Hummer, G.; Garde, S.; Garcia, A. E.; Pratt, L. R., New perspectives on hydrophobic effects. *Chemical Physics* **2000**, 258 (2-3), 349-370.
129. Lichtscheidl, A. G.; Janicke, M. T.; Scott, B. L.; Nelson, A. T.; Kiplinger, J. L., Syntheses, structures, and $(1)\text{H}$, $(13)\text{C}\{(1)\text{H}\}$ and $(119)\text{Sn}\{(1)\text{H}\}$ NMR chemical shifts of a family of trimethyltin alkoxide, amide, halide and cyclopentadienyl compounds. *Dalton Trans* **2015**, 44 (36), 16156-63.
130. Cai, C. M.; Nagane, N.; Kumar, R.; Wyman, C. E., Coupling metal halides with a co-solvent to produce furfural and 5-HMF at high yields directly from lignocellulosic biomass as an integrated biofuels strategy. *Green Chem* **2014**, 16 (8), 3819-3829.
131. Cai, C. M.; Zhang, T. Y.; Kumar, R.; Wyman, C. E., THF co-solvent enhances hydrocarbon fuel precursor yields from lignocellulosic biomass. *Green Chem* **2013**, 15 (11), 3140-3145.
132. Smith, M. D.; Mostofian, B.; Cheng, X. L.; Petridis, L.; Cai, C. M.; Wyman, C. E.; Smith, J. C., Cosolvent pretreatment in cellulosic biofuel production: effect of tetrahydrofuran-water on lignin structure and dynamics. *Green Chem* **2016**, 18 (5), 1268-1277.
133. Mostofian, B.; Cai, C. M.; Smith, M. D.; Petridis, L.; Cheng, X.; Wyman, C. E.; Smith, J. C., Local Phase Separation of Co-solvents Enhances Pretreatment of Biomass for Bioenergy Applications. *J Am Chem Soc* **2016**, 138 (34), 10869-78.
134. Smith, M. D.; Petridis, L.; Cheng, X.; Mostofian, B.; Smith, J. C., Enhanced sampling simulation analysis of the structure of lignin in the THF-water miscibility gap. *Phys Chem Chem Phys* **2016**, 18 (9), 6394-8.
135. Smith, M. D.; Cheng, X.; Petridis, L.; Mostofian, B.; Smith, J. C., Organosolv-Water Cosolvent Phase Separation on Cellulose and its Influence on the Physical Deconstruction of Cellulose: A Molecular Dynamics Analysis. *Sci Rep* **2017**, 7 (1), 14494.
136. Patri, A. S.; Mostofian, B.; Pu, Y.; Ciaffone, N.; Soliman, M.; Smith, M. D.; Kumar, R.; Cheng, X.; Wyman, C. E.; Tetard, L.; Ragauskas, A. J.; Smith, J. C.; Petridis, L.; Cai, C. M., A Multifunctional Cosolvent Pair Reveals Molecular Principles of Biomass Deconstruction. *J Am Chem Soc* **2019**, 141 (32), 12545-12557.

137. Vnuk, F., Phase equilibria in binary mixtures. Part 2.—Miscibility gap with two critical temperatures (closed-loop phase diagram). *J. Chem. Soc., Faraday Trans. 2* **1983**, 79 (1), 57-64.
138. Narayanan, T.; Kumar, A., Reentrant Phase-Transitions in Multicomponent Liquid-Mixtures. *Phys Rep* **1994**, 249 (3), 135-218.
139. Smith, M. D.; Cai, C. M.; Cheng, X. L.; Petridis, L.; Smith, J. C., Temperature-dependent phase behaviour of tetrahydrofuran-water alters solubilization of xylan to improve co-production of furfurals from lignocellulosic biomass. *Green Chem* **2018**, 20 (7), 1612-1620.
140. Smith, M. D.; Mostofian, B.; Petridis, L.; Cheng, X.; Smith, J. C., Molecular Driving Forces behind the Tetrahydrofuran-Water Miscibility Gap. *J Phys Chem B* **2016**, 120 (4), 740-7.
141. Takamuku, T.; Nakamizo, A.; Tabata, M.; Yoshida, K.; Yamaguchi, T.; Otomo, T., Large-angle X-ray scattering, small-angle neutron scattering, and NMR relaxation studies on mixing states of 1,4-dioxane-water, 1,3-dioxane-water, and tetrahydrofuran-water mixtures. *Journal of Molecular Liquids* **2003**, 103, 143-159.
142. Bowron, D. T.; Finney, J. L.; Soper, A. K., Structural characteristics of a 0.23 mole fraction aqueous solution of tetrahydrofuran at 20 degrees C. *J Phys Chem B* **2006**, 110 (41), 20235-45.
143. Bowron, D. T.; Finney, J. L.; Soper, A. K., The structure of liquid tetrahydrofuran. *J Am Chem Soc* **2006**, 128 (15), 5119-26.
144. Abraham, M.; van der Spoel, D.; Lindahl, E.; Hess, B., The GROMACS Development Team (2016) GROMACS User Manual version 5.0. 5. In *The GROMACS Development Team (2016) GROMACS User Manual version 5.0. 5*, 2016.
145. Abraham, M. J.; Gready, J. E., Optimization of parameters for molecular dynamics simulation using smooth particle-mesh Ewald in GROMACS 4.5. *Journal of computational chemistry* **2011**, 32 (9), 2031-40.
146. MacKerell Jr, A. D.; Brooks, B.; Brooks III, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M., CHARMM: the energy function and its parameterization. *Encyclopedia of computational chemistry* **2002**, 1.
147. Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D., Additive and Classical Drude Polarizable Force Fields for Linear and Cyclic Ethers. *J Chem Theory Comput* **2007**, 3 (3), 1120-33.
148. Allinger, N. L.; Flanagan, H. L., Isotope Effects in Molecular Mechanics (Mm2) Calculations on Deuterium Compounds. *Journal of computational chemistry* **1983**, 4 (3), 399-403.
149. Ochterski, J. W., Vibrational analysis in Gaussian. *help@ gaussian. com* **1999**.
150. Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C., Rapid Parameterization of Small Molecules Using the Force Field Toolkit. *Journal of computational chemistry* **2013**, 34 (32), 2757-2770.
151. Liboff, R. L., *Introductory quantum mechanics*. Pearson Education India: 2003.
152. Frisch, M., Gaussian09. <http://www.gaussian.com/> **2013**.

153. Bigeleisen, J.; Ishida, T., Application of Finite Orthogonal Polynomials to Thermal Functions of Harmonic Oscillators .5. Isotope Chemistry and Molecular-Structure - Simplified Theory of End Atom Isotope-Effects. *Journal of the American Chemical Society* **1973**, 95 (19), 6155-6158.
154. A. Kohlmeyer TopoTools Plugin, Version 1.7. <http://www.ks.uiuc.edu/Research/vmd/plugins/topotools/>.
155. Vermaas, J. V.; Hardy, D. J.; Stone, J. E.; Tajkhorshid, E.; Kohlmeyer, A., TopoGromacs: Automated Topology Conversion from CHARMM to GROMACS within VMD. *J Chem Inf Model* **2016**, 56 (6), 1112-6.
156. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *Journal of molecular graphics* **1996**, 14 (1), 33-8, 27-8.
157. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R., Molecular-Dynamics with Coupling to an External Bath. *Journal of Chemical Physics* **1984**, 81 (8), 3684-3690.
158. Bussi, G.; Donadio, D.; Parrinello, M., Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, 126, 014101.
159. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method. **1998**.
160. Praprotnik, M.; Janezic, D., Molecular dynamics integration and molecular vibrational theory. III. The infrared spectrum of water. *J Chem Phys* **2005**, 122 (17), 174103.
161. Mcmillan, W. G.; Mayer, J. E., The Statistical Thermodynamics of Multicomponent Systems. *Journal of Chemical Physics* **1945**, 13 (7), 276-305.
162. van der Spoel, D.; van Maaren, P. J.; Larsson, P.; Timneanu, N., Thermodynamics of hydrogen bonding in hydrophilic and hydrophobic media. *J Phys Chem B* **2006**, 110 (9), 4393-8.
163. Xu, H. F.; Berne, B. J., Hydrogen-bond kinetics in the solvation shell of a polypeptide. *Journal of Physical Chemistry B* **2001**, 105 (48), 11929-11932.
164. Xu, H.; Stern, H. A.; Berne, B. J., Can Water Polarizability Be Ignored in Hydrogen Bond Kinetics? *The Journal of Physical Chemistry B* **2002**, 106 (8), 2054-2060.
165. Hill, T. L., *Statistical Mechanics: Principles and Selected Applications*. 1987.
166. Lide, D. R.; Frederikse, H., CRC Handbook of Chemistry and Physics, seventy. CRC Press, Inc., Boca Raton, FL: 1995.
167. <http://www.chemspider.com/Chemical-Structure.72531.html>.
168. Solutions, N. C., TETRAHYDROFURAN-D8 supplier distributor- CAS 1693-74-9. **2019**.
169. Pranami, G.; Lamm, M. H., Estimating error in diffusion coefficients derived from molecular dynamics simulations. *J Chem Theory Comput* **2015**, 11 (10), 4586-92.
170. Holz, M.; Mao, X. a.; Seiferling, D.; Sacco, A., Experimental study of dynamic isotope effects in molecular liquids: Detection of translation-rotation coupling. *The Journal of Chemical Physics* **1996**, 104 (2), 669-679.
171. Palm, A.; Bissell, E. R., The Infrared Spectra of Deuterated Tetrahydrofurans. *Spectrochim Acta* **1960**, 16 (4), 459-466.

172. Sakurai, A.; Tanimura, Y., Does variant Planck's over 2π play a role in multidimensional spectroscopy? Reduced hierarchy equations of motion approach to molecular vibrations. *J Phys Chem A* **2011**, *115* (16), 4009-22.
173. Ueno, S.; Tanimura, Y.; Ten-no, S., Molecular dynamics simulation for infrared spectroscopy with intramolecular forces from electronic properties of on-the-fly quantum chemical calculations. *International Journal of Quantum Chemistry* **2013**, *113* (3), 330-335.
174. Timmins, G. S., Deuterated drugs; updates and obviousness analysis. *Expert Opin Ther Pat* **2017**, *27* (12), 1353-1361.
175. Mullard, A., FDA approves first deuterated drug. *Nat Rev Drug Discov* **2017**, *16* (5), 305.
176. Hoff, B.; Strandberg, E.; Ulrich, A. S.; Tieleman, D. P.; Posten, C., ^2H -NMR study and molecular dynamics simulation of the location, alignment, and mobility of pyrene in POPC bilayers. *Biophys J* **2005**, *88* (3), 1818-27.
177. Shrestha, U. R.; Smith, S.; Pingali, S. V.; Yang, H.; Zahran, M.; Breunig, L.; Wilson, L. A.; Kowali, M.; Kubicki, J. D.; Cosgrove, D. J.; O'Neill, H. M.; Petridis, L., Arabinose substitution effect on xylan rigidity and self-aggregation. *Cellulose* **2019**, *26* (4), 2267-2278.
178. Hoen, B.; Duval, X., Clinical practice. Infective endocarditis. *N Engl J Med* **2013**, *368* (15), 1425-33.
179. Fitzgerald, J. R.; Foster, T. J.; Cox, D., The interaction of bacterial pathogens with platelets. *Nat Rev Microbiol* **2006**, *4* (6), 445-57.
180. Sullam, P. M.; Sande, M. A., Role of platelets in endocarditis: clues from von Willebrand disease. *J Lab Clin Med* **1992**, *120* (4), 507-9.
181. Bensing, B. A.; Lopez, J. A.; Sullam, P. M., The Streptococcus gordonii surface proteins GspB and Hsa mediate binding to sialylated carbohydrate epitopes on the platelet membrane glycoprotein Ibalpha. *Infect Immun* **2004**, *72* (11), 6528-37.
182. Plummer, C.; Wu, H.; Kerrigan, S. W.; Meade, G.; Cox, D.; Ian Douglas, C. W., A serine-rich glycoprotein of Streptococcus sanguis mediates adhesion to platelets via GPIb. *Br J Haematol* **2005**, *129* (1), 101-9.
183. Takamatsu, D.; Bensing, B. A.; Cheng, H.; Jarvis, G. A.; Siboo, I. R.; Lopez, J. A.; Griffiss, J. M.; Sullam, P. M., Binding of the Streptococcus gordonii surface glycoproteins GspB and Hsa to specific carbohydrate structures on platelet membrane glycoprotein Ibalpha. *Mol Microbiol* **2005**, *58* (2), 380-92.
184. Varki, A.; Schauer, R., Sialic Acids. In *Essentials of Glycobiology*, nd; Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E., Eds. Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2009.
185. Dankwa, S.; Lim, C.; Bei, A. K.; Jiang, R. H.; Abshire, J. R.; Patel, S. D.; Goldberg, J. M.; Moreno, Y.; Kono, M.; Niles, J. C.; Duraisingh, M. T., Ancient human sialic acid variant restricts an emerging zoonotic malaria parasite. *Nat Commun* **2016**, *7*, 11187.
186. Alisson-Silva, F.; Liu, J. Z.; Diaz, S. L.; Deng, L.; Gareau, M. G.; Marchelletta, R.; Chen, X.; Nizet, V.; Varki, N.; Barrett, K. E.; Varki, A., Human

evolutionary loss of epithelial Neu5Gc expression and species-specific susceptibility to cholera. *PLoS Pathog* **2018**, *14* (6), e1007133.

187. Bornhofft, K. F.; Goldammer, T.; Rebl, A.; Galuska, S. P., Siglecs: A journey through the evolution of sialic acid-binding immunoglobulin-type lectins. *Dev Comp Immunol* **2018**, *86*, 219-231.

188. Varki, A., Sialic acids in human health and disease. *Trends Mol Med* **2008**, *14* (8), 351-60.

189. Wang, B.; Brand-Miller, J., The role and potential of sialic acid in human nutrition. *Eur J Clin Nutr* **2003**, *57* (11), 1351-69.

190. Bull, C.; Stoel, M. A.; den Brok, M. H.; Adema, G. J., Sialic acids sweeten a tumor's life. *Cancer Res* **2014**, *74* (12), 3199-204.

191. Du, J.; Hong, S.; Dong, L.; Cheng, B.; Lin, L.; Zhao, B.; Chen, Y. G.; Chen, X., Dynamic Sialylation in Transforming Growth Factor-beta (TGF-beta)-induced Epithelial to Mesenchymal Transition. *J Biol Chem* **2015**, *290* (19), 12000-13.

192. Teoh, S. T.; Ogradzinski, M. P.; Ross, C.; Hunter, K. W.; Lunt, S. Y., Sialic Acid Metabolism: A Key Player in Breast Cancer Metastasis Revealed by Metabolomics. *Front Oncol* **2018**, *8*, 174.

193. Zhou, X.; Yang, G.; Guan, F., Biological Functions and Analytical Strategies of Sialic Acids in Tumor. *Cells* **2020**, *9* (2).

194. Alexander N., B.; Sarah-Jane, R.; Collette S., G.; Thomas R., C.; Muhammad, H.; Alexander J., Z.; anne, s.; Marc, W.; Simona, C.; Giulia, P.; Simone, D.; Robert, F.; Matthew, G., *The SARS-COV-2 Spike Protein Binds Sialic Acids, and Enables Rapid Detection in a Lateral Flow Point of Care Diagnostic Device*. 2020.

195. Deng, L.; Bensing, B. A.; Thamadilok, S.; Yu, H.; Lau, K.; Chen, X.; Ruhl, S.; Sullam, P. M.; Varki, A., Oral streptococci utilize a Siglec-like domain of serine-rich repeat adhesins to preferentially target platelet sialoglycans in human blood. *PLoS Pathog* **2014**, *10* (12), e1004540.

196. Bensing, B. A.; Loukachevitch, L. V.; McCulloch, K. M.; Yu, H.; Vann, K. R.; Wawrzak, Z.; Anderson, S.; Chen, X.; Sullam, P. M.; Iverson, T. M., Structural Basis for Sialoglycan Binding by the Streptococcus sanguinis SrpA Adhesin. *J Biol Chem* **2016**, *291* (14), 7230-40.

197. Loukachevitch, L. V.; Bensing, B. A.; Yu, H.; Zeng, J.; Chen, X.; Sullam, P. M.; Iverson, T. M., Structures of the Streptococcus sanguinis SrpA Binding Region with Human Sialoglycans Suggest Features of the Physiological Ligand. *Biochemistry* **2016**, *55* (42), 5927-5937.

198. Bensing, B. A.; Loukachevitch, L. V.; Agarwal, R.; Yamakawa, I.; Luong, K.; Hadadianpour, A.; Yu, H.; Fialkowski, K. P.; Castro, M. A.; Wawrzak, Z.; Chen, X.; Baudry, J.; Smith, J. C.; Sullam, P. M.; Iverson, T. M., Selectivity and engineering of the sialoglycan-binding spectrum in Siglec-like adhesins. *bioRxiv* **2019**, 796912.

199. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J., GLYCAM06: a generalizable

- biomolecular force field. Carbohydrates. *Journal of computational chemistry* **2008**, 29 (4), 622-55.
200. Price, D. J.; Brooks, C. L., 3rd, A modified TIP3P water potential for simulation with Ewald summation. *J Chem Phys* **2004**, 121 (20), 10096-103.
201. Miyamoto, S.; Kollman, P. A., Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. *Journal of computational chemistry* **1992**, 13 (8), 952-962.
202. Zwanzig, R., Nonlinear generalized Langevin equations. *Journal of Statistical Physics* **1973**, 9 (3), 215-220.
203. Åqvist, J.; Wennerström, P.; Nervall, M.; Bjelic, S.; Brandsdal, B. O., Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm. *Chemical physics letters* **2004**, 384 (4-6), 288-294.
204. Roe, D. R.; Cheatham, T. E., 3rd, PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013**, 9 (7), 3084-95.
205. Bensing, B. A.; Khedri, Z.; Deng, L.; Yu, H.; Prakobphol, A.; Fisher, S. J.; Chen, X.; Iverson, T. M.; Varki, A.; Sullam, P. M., Novel aspects of sialoglycan recognition by the Siglec-like domains of streptococcal SRR glycoproteins. *Glycobiology* **2016**, 26 (11), 1222-1234.
206. Pyburn, T. M.; Bensing, B. A.; Xiong, Y. Q.; Melancon, B. J.; Tomasiak, T. M.; Ward, N. J.; Yankovskaya, V.; Oliver, K. M.; Cecchini, G.; Sulikowski, G. A.; Tyska, M. J.; Sullam, P. M.; Iverson, T. M., A structural model for binding of the serine-rich repeat adhesin GspB to host carbohydrate receptors. *PLoS Pathog* **2011**, 7 (7), e1002112.
207. Fischetti, V. A., Surface Proteins on Gram-Positive Bacteria. *Microbiol Spectr* **2019**, 7 (4).
208. Jahns, A. C.; Rehm, B. H., Relevant uses of surface proteins--display on self-organized biological structures. *Microb Biotechnol* **2012**, 5 (2), 188-202.
209. An, Y. H.; Friedman, R. J., Handbook of Bacterial Adhesion: Principles, Methods and Applications. Humana Press: Totowa, 2001.
210. Amorena, B.; Baselga, R.; Aguilar, B., Factors influencing the degree of in vitro bacterial adhesion to ovine mammary gland epithelial cells. *Vet Microbiol* **1990**, 24 (1), 43-53.
211. Dutta, D.; Cole, N.; Willcox, M., Factors influencing bacterial adhesion to contact lenses. *Mol Vis* **2012**, 18, 14-21.
212. Sun, X. Y.; Shi, H. C.; Quan, X. C.; Hu, L. J., [Factors influencing bacterial cell-surface hydrophobicity and bacterial adhesion ratio to activated sludge flocs]. *Huan Jing Ke Xue* **2008**, 29 (10), 2809-14.
213. Back, C. R.; Higman, V. A.; Le Vay, K.; Patel, V. V.; Parnell, A. E.; Frankel, D.; Jenkinson, H. F.; Burston, S. G.; Crump, M. P.; Nobbs, A. H.; Race, P. R., The streptococcal multidomain fibrillar adhesin CshA has an elongated polymeric architecture. *J Biol Chem* **2020**, 295 (19), 6689-6699.
214. Swamy, N.; Dutta, A.; Ray, R., Roles of the structure and orientation of ligands and ligand mimics inside the ligand-binding pocket of the vitamin D-binding protein. *Biochemistry* **1997**, 36 (24), 7432-6.

215. Foster, T. J.; Geoghegan, J. A.; Ganesh, V. K.; Hook, M., Adhesion, invasion and evasion: the many functions of the surface proteins of *Staphylococcus aureus*. *Nat Rev Microbiol* **2014**, *12* (1), 49-62.
216. Lizcano, A.; Sanchez, C. J.; Orihuela, C. J., A role for glycosylated serine-rich repeat proteins in gram-positive bacterial pathogenesis. *Mol Oral Microbiol* **2012**, *27* (4), 257-69.
217. Sanchez, C. J.; Shivshankar, P.; Stol, K.; Trakhtenbroit, S.; Sullam, P. M.; Sauer, K.; Hermans, P. W.; Orihuela, C. J., The pneumococcal serine-rich repeat protein is an intra-species bacterial adhesin that promotes bacterial aggregation in vivo and in biofilms. *PLoS Pathog* **2010**, *6* (8), e1001044.
218. Bashore, T. M.; Cabell, C.; Fowler, V., Jr., Update on infective endocarditis. *Curr Probl Cardiol* **2006**, *31* (4), 274-352.
219. Bensing, B. A.; Li, L.; Yakovenko, O.; Wong, M.; Barnard, K. N.; Iverson, T. M.; Lebrilla, C. B.; Parrish, C. R.; Thomas, W. E.; Xiong, Y.; Sullam, P. M., Recognition of specific sialoglycan structures by oral streptococci impacts the severity of endocardial infection. *PLoS Pathog* **2019**, *15* (6), e1007896.
220. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics* **1983**, *79* (2), 926-935.
221. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **2015**, *11* (8), 3696-713.
222. Arfken, G. B.; Weber, H. J.; Harris, F. E., *Mathematical methods for physicists: a comprehensive guide*. Academic press: 2011.
223. Jean-paul Ryckaert, G. C., Herman J. C. Berendsen, Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys* **1977**.
224. Bensing, B. A.; Sullam, P. M., An accessory sec locus of *Streptococcus gordonii* is required for export of the surface protein GspB and for normal levels of binding to human platelets. *Mol Microbiol* **2002**, *44* (4), 1081-94.
225. Desvaux, M.; Candela, T.; Serror, P., Surfaceome and Proteosurfaceome in Parietal Monoderm Bacteria: Focus on Protein Cell-Surface Display. *Front Microbiol* **2018**, *9*, 100.
226. Ofek, I.; Hasty, D. L.; Sharon, N., Anti-adhesion therapy of bacterial diseases: prospects and problems. *FEMS Immunol Med Microbiol* **2003**, *38* (3), 181-91.
227. Durack, D. T.; Petersdorf, R. G., Chemotherapy of experimental streptococcal endocarditis. I. Comparison of commonly recommended prophylactic regimens. *J Clin Invest* **1973**, *52* (3), 592-8.
228. Vieille, C.; Zeikus, G. J., Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* **2001**, *65* (1), 1-43.

229. Rigoldi, F.; Donini, S.; Redaelli, A.; Parisini, E.; Gautieri, A., Review: Engineering of thermostable enzymes for industrial applications. In *APL Bioeng*, 2018; Vol. 2.
230. Mohammad, B. T.; Al Daghistani, H. I.; Jaouani, A.; Abdel-Latif, S.; Kennes, C., Isolation and Characterization of Thermophilic Bacteria from Jordanian Hot Springs: *Bacillus licheniformis* and *Thermomonas hydrothermalis* Isolates as Potential Producers of Thermostable Enzymes. *Int J Microbiol* **2017**, 2017, 6943952.
231. Yayanos, A. A., Piezophiles. *e LS* **2001**.
232. Daniel, R. M.; Dines, M.; Petach, H. H., The denaturation and degradation of stable enzymes at high temperatures. *Biochem J* **1996**, 317 (Pt 1) (Pt 1), 1-11.
233. Roche, J.; Royer, C. A., Lessons from pressure denaturation of proteins. *J R Soc Interface* **2018**, 15 (147).
234. Vogt, G.; Woell, S.; Argos, P., Protein thermal stability, hydrogen bonds, and ion pairs. *J Mol Biol* **1997**, 269 (4), 631-43.
235. Pucci, F.; Rooman, M., Physical and molecular bases of protein thermal stability and cold adaptation. *Curr Opin Struct Biol* **2017**, 42, 117-128.
236. Salminen, T.; Teplyakov, A.; Kankare, J.; Cooperman, B. S.; Lahti, R.; Goldman, A., An unusual route to thermostability disclosed by the comparison of *Thermus thermophilus* and *Escherichia coli* inorganic pyrophosphatases. *Protein Sci* **1996**, 5 (6), 1014-25.
237. Folch, B.; Dehouck, Y.; Rooman, M., Thermo- and Mesostabilizing Protein Interactions Identified by Temperature-Dependent Statistical Potentials. In *Biophys J*, 2010; Vol. 98, pp 667-77.
238. Szilágyi, A.; Závodszky, P., Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure* **2000**, 8 (5), 493-504.
239. Argos, P.; Rossman, M. G.; Grau, U. M.; Zuber, H.; Frank, G.; Tratschin, J. D., Thermal stability and protein structure. *Biochemistry* **1979**, 18 (25), 5698-703.
240. Shrestha, U. R.; Bhowmik, D.; Copley, J. R.; Tyagi, M.; Leao, J. B.; Chu, X. Q., Effects of pressure on the dynamics of an oligomeric protein from deep-sea hyperthermophile. *Proc Natl Acad Sci U S A* **2015**, 112 (45), 13886-91.
241. Brock, T. D.; Freeze, H., *Thermus aquaticus* gen. n. and sp. n., a nonsporulating extreme thermophile. *J Bacteriol* **1969**, 98 (1), 289-97.
242. Librizzi, F.; Carrotta, R.; Peters, J.; Cupane, A., The effects of pressure on the energy landscape of proteins. *Sci Rep* **2018**, 8 (1), 2037.
243. Mallamace, F.; Corsaro, C.; Mallamace, D.; Vasi, S.; Vasi, C.; Baglioni, P.; Buldyrev, S. V.; Chen, S. H.; Stanley, H. E., Energy landscape in protein folding and unfolding. *Proc Natl Acad Sci U S A* **2016**, 113 (12), 3159-63.
244. Meinhold, L.; Smith, J. C.; Kitao, A.; Zewail, A. H., Picosecond fluctuating protein energy landscape mapped by pressure temperature molecular dynamics simulation. *Proc Natl Acad Sci U S A* **2007**, 104 (44), 17261-5.
245. Meinhold, L.; Clement, D.; Tehei, M.; Daniel, R.; Finney, J. L.; Smith, J. C., Protein dynamics and stability: the distribution of atomic fluctuations in

thermophilic and mesophilic dihydrofolate reductase derived using elastic incoherent neutron scattering. *Biophys J* **2008**, 94 (12), 4812-8.

246. Shrestha, U. R.; Perera, S.; Bhowmik, D.; Chawla, U.; Mamontov, E.; Brown, M. F.; Chu, X. Q., Quasi-elastic Neutron Scattering Reveals Ligand-Induced Protein Dynamics of a G-Protein-Coupled Receptor. *J Phys Chem Lett* **2016**, 7 (20), 4130-4136.

247. Ferrand, M.; Dianoux, A. J.; Petry, W.; Zaccai, G., Thermal motions and function of bacteriorhodopsin in purple membranes: effects of temperature and hydration studied by neutron scattering. *Proc Natl Acad Sci U S A* **1993**, 90 (20), 9668-72.

248. Borzova, V. A.; Markossian, K. A.; Chebotareva, N. A.; Kleymenov, S. Y.; Poliansky, N. B.; Muranov, K. O.; Stein-Margolina, V. A.; Shubin, V. V.; Markov, D. I.; Kurganov, B. I., Kinetics of Thermal Denaturation and Aggregation of Bovine Serum Albumin. *PLoS One* **2016**, 11 (4), e0153495.

249. Karshikoff, A.; Nilsson, L.; Ladenstein, R., Rigidity versus flexibility: the dilemma of understanding protein thermal stability. *FEBS J* **2015**, 282 (20), 3899-917.

250. Roche, J.; Caro, J. A.; Norberto, D. R.; Barthe, P.; Roumestand, C.; Schlessman, J. L.; Garcia, A. E.; Garcia-Moreno, B. E.; Royer, C. A., Cavities determine the pressure unfolding of proteins. *Proc Natl Acad Sci U S A* **2012**, 109 (18), 6945-50.

251. Doster, W.; Gebhardt, R., High pressure - unfolding of myoglobin studied by dynamic neutron scattering. *Chemical Physics* **2003**, 292 (2-3), 383-387.

252. Fukuda, M.; Kunugi, S., Pressure dependence of thermolysin catalysis. *Eur J Biochem* **1984**, 142 (3), 565-70.

253. Konisky, J.; Michels, P. C.; Clark, D. S., Pressure stabilization is not a general property of thermophilic enzymes: the adenylate kinases of *Methanococcus voltae*, *Methanococcus maripaludis*, *Methanococcus thermolithotrophicus*, and *Methanococcus jannaschii*. *Appl Environ Microbiol* **1995**, 61 (7), 2762-4.

254. Hernandez, G.; Jenney, F. E., Jr.; Adams, M. W.; LeMaster, D. M., Millisecond time scale conformational flexibility in a hyperthermophile protein at ambient temperature. *Proc Natl Acad Sci U S A* **2000**, 97 (7), 3166-70.

255. Yu, H.; Huang, H., Engineering proteins for thermostability through rigidifying flexible sites. *Biotechnol Adv* **2014**, 32 (2), 308-15.

256. Zhang, X. F.; Yang, G. Y.; Zhang, Y.; Xie, Y.; Withers, S. G.; Feng, Y., A general and efficient strategy for generating the stable enzymes. *Sci Rep* **2016**, 6 (1), 33797.

257. Ogasahara, K.; Khechinashvili, N. N.; Nakamura, M.; Yoshimoto, T.; Yutani, K., Thermal stability of pyrrolidone carboxyl peptidases from the hyperthermophilic Archaeon, *Pyrococcus furiosus*. *Eur J Biochem* **2001**, 268 (11), 3233-42.

258. Thoma, R.; Hennig, M.; Sterner, R.; Kirschner, K., Structure and function of mutationally generated monomers of dimeric phosphoribosylanthranilate isomerase from *Thermotoga maritima*. *Structure* **2000**, 8 (3), 265-76.

259. Kumar, S.; Tsai, C. J.; Nussinov, R., Factors enhancing protein thermostability. *Protein Eng* **2000**, *13* (3), 179-91.
260. Hughes, R. C.; Coates, L.; Blakeley, M. P.; Tomanicek, S. J.; Langan, P.; Kovalevsky, A. Y.; Garcia-Ruiz, J. M.; Ng, J. D., Inorganic pyrophosphatase crystals from *Thermococcus thioireducens* for X-ray and neutron diffraction. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **2012**, *68* (Pt 12), 1482-7.
261. Samygina, V. R.; Popov, A. N.; Rodina, E. V.; Vorobyeva, N. N.; Lamzin, V. S.; Polyakov, K. M.; Kurilova, S. A.; Nazarova, T. I.; Avaeva, S. M., The structures of *Escherichia coli* inorganic pyrophosphatase complexed with Ca(2+) or CaPP(i) at atomic resolution and their mechanistic implications. *J Mol Biol* **2001**, *314* (3), 633-45.
262. Chu, X. Q.; Gajapathy, M.; Weiss, K. L.; Mamontov, E.; Ng, J. D.; Coates, L., Dynamic behavior of oligomeric inorganic pyrophosphatase explored by quasielastic neutron scattering. *J Phys Chem B* **2012**, *116* (33), 9917-21.
263. Josse, J., Constitutive inorganic pyrophosphatase of *Escherichia coli*. 1. Purification and catalytic properties. *J Biol Chem* **1966**, *241* (9), 1938-47.
264. Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M., Protein structure determination from NMR chemical shifts. *Proc Natl Acad Sci U S A* **2007**, *104* (23), 9615-20.
265. Markwick, P. R.; Malliavin, T.; Nilges, M., Structural biology by NMR: structure, dynamics, and interactions. *PLoS Comput Biol* **2008**, *4* (9), e1000168.
266. Hollingsworth, S. A.; Dror, R. O., Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129-1143.
267. Ulmschneider, J. P.; Ulmschneider, M. B., Molecular Dynamics Simulations Are Redefining Our View of Peptides Interacting with Biological Membranes. *Acc Chem Res* **2018**, *51* (5), 1106-1116.
268. Abraham, M.; van der Spoel, D.; Lindahl, E.; Hess, B., the GROMACS Development Team (2016) GROMACS User Manual version 5.0. 5.
269. Paschek, D.; Day, R.; Garcia, A. E., Influence of water-protein hydrogen bonding on the stability of Trp-cage miniprotein. A comparison between the TIP3P and TIP4P-Ew water models. *Phys Chem Chem Phys* **2011**, *13* (44), 19840-7.
270. Okumura, H., Temperature and pressure denaturation of chignolin: folding and unfolding simulation by multibaric-multithermal molecular dynamics method. *Proteins* **2012**, *80* (10), 2397-416.
271. Hata, H.; Nishihara, Y.; Nishiyama, M.; Sowa, Y.; Kawagishi, I.; Kitao, A., High pressure inhibits signaling protein binding to the flagellar motor and bacterial chemotaxis through enhanced hydration. *Sci Rep* **2020**, *10* (1), 2351.
272. Fine, R. A.; Millero, F. J., Compressibility of Water as a Function of Temperature and Pressure. *Journal of Chemical Physics* **1973**, *59* (10), 5529-5536.
273. Darden, T.; York, D.; Pedersen, L., Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *Journal of Chemical Physics* **1993**, *98* (12), 10089-10092.

274. Maiti, R.; Van Domselaar, G. H.; Zhang, H.; Wishart, D. S., SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Res* **2004**, 32 (Web Server issue), W590-4.
275. Paramo, T.; East, A.; Garzon, D.; Ulmschneider, M. B.; Bond, P. J., Efficient Characterization of Protein Cavities within Molecular Simulation Trajectories: trj_cavity. *J Chem Theory Comput* **2014**, 10 (5), 2151-64.
276. Gerstein, M.; Chothia, C., Packing at the protein-water interface. *Proc Natl Acad Sci U S A* **1996**, 93 (19), 10167-72.
277. Henzler-Wildman, K. A.; Thai, V.; Lei, M.; Ott, M.; Wolf-Watz, M.; Fenn, T.; Pozharski, E.; Wilson, M. A.; Petsko, G. A.; Karplus, M.; Hubner, C. G.; Kern, D., Intrinsic motions along an enzymatic reaction trajectory. *Nature* **2007**, 450 (7171), 838-44.
278. Robinson-Rechavi, M.; Alibes, A.; Godzik, A., Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of *Thermotoga maritima*. *J Mol Biol* **2006**, 356 (2), 547-57.
279. Bogatyreva, N. S.; Ivankov, D. N., [The relationship between protein accessible surface area and number of native contacts in its structure]. *Mol Biol (Mosk)* **2008**, 42 (6), 1048-55.
280. Hubbard, S. J.; Argos, P., Cavities and packing at protein interfaces. *Protein Sci* **1994**, 3 (12), 2194-206.
281. Sonavane, S.; Chakrabarti, P., Cavities and Atomic Packing in Protein Structures and Interfaces. In *PLoS Comput Biol*, 2008; Vol. 4.
282. Cioni, P., Role of protein cavities on unfolding volume change and on internal dynamics under pressure. *Biophys J* **2006**, 91 (9), 3390-6.
283. Dundas, J.; Adamian, L.; Liang, J., Structural signatures of enzyme binding pockets from order-independent surface alignment: a study of metalloendopeptidase and NAD binding proteins. *J Mol Biol* **2011**, 406 (5), 713-29.
284. Kumar, P.; Agarwal, P. K.; Waddell, M. B.; Mittag, T.; Serpersu, E. H.; Cuneo, M. J., Low-Barrier and Canonical Hydrogen Bonds Modulate Activity and Specificity of a Catalytic Triad. *Angew Chem Int Ed Engl* **2019**, 58 (45), 16260-16266.
285. Cannon, W. R.; Singleton, S. F.; Benkovic, S. J., A perspective on biological catalysis. *Nat Struct Biol* **1996**, 3 (10), 821-33.
286. Yang, L.; Liao, R. Z.; Yu, J. G.; Liu, R. Z., DFT study on the mechanism of *Escherichia coli* inorganic pyrophosphatase. *J Phys Chem B* **2009**, 113 (18), 6505-10.
287. Cooperman, B. S.; Baykov, A. A.; Lahti, R., Evolutionary conservation of the active site of soluble inorganic pyrophosphatase. *Trends Biochem Sci* **1992**, 17 (7), 262-6.
288. Oksanen, E.; Ahonen, A. K.; Tuominen, H.; Tuominen, V.; Lahti, R.; Goldman, A.; Heikinheimo, P., A complete structural description of the catalytic cycle of yeast pyrophosphatase. *Biochemistry* **2007**, 46 (5), 1228-39.

289. Heikinheimo, P.; Lehtonen, J.; Baykov, A.; Lahti, R.; Cooperman, B. S.; Goldman, A., The structural basis for pyrophosphatase catalysis. *Structure* **1996**, 4 (12), 1491-508.
290. Kurkal, V.; Daniel, R. M.; Finney, J. L.; Tehei, M.; Dunn, R. V.; Smith, J. C., Enzyme activity and flexibility at very low hydration. *Biophys J* **2005**, 89 (2), 1282-7.
291. Kurkal, V.; Daniel, R. M.; Finney, J. L.; Tehei, M.; Dunn, R. V.; Smith, J. C., Low frequency enzyme dynamics as a function of temperature and hydration: A neutron scattering study. *Chemical Physics* **2005**, 317 (2-3), 267-273.
292. Carugo, O., Structure and function of water molecules buried in the protein core. *Curr Protein Pept Sci* **2015**, 16 (3), 259-65.
293. Liu, Z.; Lemmonds, S.; Huang, J.; Tyagi, M.; Hong, L.; Jain, N., Entropic contribution to enhanced thermal stability in the thermostable P450 CYP119. *Proc Natl Acad Sci U S A* **2018**, 115 (43), E10049-E10058.
294. Paredes, D. I.; Watters, K.; Pitman, D. J.; Bystroff, C.; Dordick, J. S., Comparative void-volume analysis of psychrophilic and mesophilic enzymes: Structural bioinformatics of psychrophilic enzymes reveals sources of core flexibility. *BMC Struct Biol* **2011**, 11, 42.
295. Rahaman, O.; Kalimeri, M.; Melchionna, S.; Henin, J.; Sterpone, F., Role of Internal Water on Protein Thermal Stability: The Case of Homologous G Domains. *J Phys Chem B* **2015**, 119 (29), 8939-49.
296. Rahaman, O.; Melchionna, S.; Laage, D.; Sterpone, F., The effect of protein composition on hydration dynamics. *Phys Chem Chem Phys* **2013**, 15 (10), 3570-6.
297. Xue, M.; Wakamoto, T.; Kejlberg, C.; Yoshimura, Y.; Nielsen, T. A.; Risor, M. W.; Sanggaard, K. W.; Kitahara, R.; Mulder, F. A. A., How internal cavities destabilize a protein. *Proc Natl Acad Sci U S A* **2019**, 116 (42), 21031-21036.
298. Závodszy, P.; Kardos, J.; Svingor, Á.; Petsko, G. A., Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. In *Proc Natl Acad Sci U S A*, 1998; Vol. 95, pp 7406-11.
299. Salvador-Castell, M.; Golub, M.; Martinez, N.; Ollivier, J.; Peters, J.; Oger, P., The first study on the impact of osmolytes in whole cells of high temperature-adapted microorganisms. *Soft Matter* **2019**, 15 (41), 8381-8391.
300. Cario, A.; Jebbar, M.; Thiel, A.; Kervarec, N.; Oger, P. M., Molecular chaperone accumulation as a function of stress evidences adaptation to high hydrostatic pressure in the piezophilic archaeon *Thermococcus barophilus*. *Sci Rep* **2016**, 6, 29483.
301. Di Giulio, M., The universal ancestor and the ancestor of bacteria were hyperthermophiles. *J Mol Evol* **2003**, 57 (6), 721-30.
302. Berezovsky, I. N.; Shakhnovich, E. I., Physics and evolution of thermophilic adaptation. *Proc Natl Acad Sci U S A* **2005**, 102 (36), 12742-7.
303. Catchpole, R.; Forterre, P., Positively twisted: The complex evolutionary history of Reverse Gyrase suggests a non-hyperthermophilic Last Universal Common Ancestor. **2019**.

304. Murdoch, D. R.; Corey, G. R.; Hoen, B.; Miro, J. M.; Fowler, V. G., Jr.; Bayer, A. S.; Karchmer, A. W.; Olaison, L.; Pappas, P. A.; Moreillon, P.; Chambers, S. T.; Chu, V. H.; Falco, V.; Holland, D. J.; Jones, P.; Klein, J. L.; Raymond, N. J.; Read, K. M.; Tripodi, M. F.; Utili, R.; Wang, A.; Woods, C. W.; Cabell, C. H.; International Collaboration on Endocarditis-Prospective Cohort Study, I., Clinical presentation, etiology, and outcome of infective endocarditis in the 21st century: the International Collaboration on Endocarditis-Prospective Cohort Study. *Arch Intern Med* **2009**, *169* (5), 463-73.
305. Thornhill, M. H.; Dayer, M. J.; Forde, J. M.; Corey, G. R.; Chu, V. H.; Couper, D. J.; Lockhart, P. B., Impact of the NICE guideline recommending cessation of antibiotic prophylaxis for prevention of infective endocarditis: before and after study. *BMJ* **2011**, *342*, d2392.
306. Nilson, B.; Olaison, L.; Rasmussen, M., Clinical presentation of infective endocarditis caused by different groups of non-beta haemolytic streptococci. *Eur J Clin Microbiol* **2016**, *35* (2), 215-8.
307. Walls, G.; McBride, S.; Raymond, N.; Read, K.; Coomarasamy, C.; Morris, A. J.; Chambers, S.; Holland, D.; Murdoch, D. R., Infective endocarditis in New Zealand: data from the International Collaboration on Endocarditis Prospective Cohort Study. *N Z Med J* **2014**, *127* (1391), 38-51.
308. Dadon, Z.; Cohen, A.; Szterenlicht, Y. M.; Assous, M. V.; Barzilay, Y.; Raveh-Brawer, D.; Yinnon, A. M.; Munter, G., Spondylodiskitis and endocarditis due to *Streptococcus gordonii*. *Ann Clin Microb Anti* **2017**, *16* (1), 68.
309. Khalid, N.; Shlofmitz, E.; Ahmad, S. A., Aortic Valve Endocarditis. **2020**.
310. Carratala, J.; Alcaide, F.; Fernandez-Sevilla, A.; Corbella, X.; Linares, J.; Gudiol, F., Bacteremia due to viridans streptococci that are highly resistant to penicillin: increase among neutropenic patients with cancer. *Clin Infect Dis* **1995**, *20* (5), 1169-73.
311. Fowler, V. G., Jr.; Boucher, H. W.; Corey, G. R.; Abrutyn, E.; Karchmer, A. W.; Rupp, M. E.; Levine, D. P.; Chambers, H. F.; Tally, F. P.; Vigliani, G. A.; Cabell, C. H.; Link, A. S.; DeMeyer, I.; Filler, S. G.; Zervos, M.; Cook, P.; Parsonnet, J.; Bernstein, J. M.; Price, C. S.; Forrest, G. N.; Fatkenheuer, G.; Gareca, M.; Rehm, S. J.; Brodt, H. R.; Tice, A.; Cosgrove, S. E.; Endocarditis, S. a.; Bacteremia Study, G., Daptomycin versus standard therapy for bacteremia and endocarditis caused by *Staphylococcus aureus*. *N Engl J Med* **2006**, *355* (7), 653-65.
312. Gould, I. M.; Cauda, R.; Esposito, S.; Gudiol, F.; Mazzei, T.; Garau, J., Management of serious methicillin-resistant *Staphylococcus aureus* infections: what are the limits? *Int J Antimicrob Agents* **2011**, *37* (3), 202-9.
313. Boucher, H.; Miller, L. G.; Razonable, R. R., Serious infections caused by methicillin-resistant *Staphylococcus aureus*. *Clin Infect Dis* **2010**, *51* Suppl 2 (S2), S183-97.
314. Tornos, P.; Almirante, B.; Olona, M.; Permanyer, G.; Gonzalez, T.; Carballo, J.; Pahissa, A.; Soler-Soler, J., Clinical outcome and long-term prognosis of late prosthetic valve endocarditis: a 20-year experience. *Clin Infect Dis* **1997**, *24* (3), 381-6.

315. Takamatsu, D.; Bensing, B. A.; Prakobphol, A.; Fisher, S. J.; Sullam, P. M., Binding of the streptococcal surface glycoproteins GspB and Hsa to human salivary proteins. *Infect Immun* **2006**, 74 (3), 1933-40.
316. Cozens, D.; Read, R. C., Anti-adhesion methods as novel therapeutics for bacterial infections. *Expert Rev Anti Infect Ther* **2012**, 10 (12), 1457-68.
317. Ribic, R.; Mestrovic, T.; Neuberg, M.; Kozina, G., Effective anti-adhesives of uropathogenic Escherichia coli. *Acta Pharm* **2018**, 68 (1), 1-18.
318. Duan, S.; Paulson, J. C., Siglecs as Immune Cell Checkpoints in Disease. *Annu Rev Immunol* **2020**, 38, 365-395.
319. Evangelista, W.; Weir, R. L.; Ellingson, S. R.; Harris, J. B.; Kapoor, K.; Smith, J. C.; Baudry, J., Ensemble-based docking: From hit discovery to metabolism and toxicity predictions. *Bioorg Med Chem* **2016**, 24 (20), 4928-4935.
320. Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P., Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* **1999**, 42 (25), 5100-9.
321. Kharade, S. V.; Kurata, H.; Bender, A. M.; Blobaum, A. L.; Figueroa, E. E.; Duran, A.; Kramer, M.; Days, E.; Vinson, P.; Flores, D.; Satlin, L. M.; Meiler, J.; Weaver, C. D.; Lindsley, C. W.; Hopkins, C. R.; Denton, J. S., Discovery, Characterization, and Effects on Renal Fluid and Electrolyte Excretion of the Kir4.1 Potassium Channel Pore Blocker, VU0134992. *Mol Pharmacol* **2018**, 94 (2), 926-937.
322. Bewley, B. R.; Spearing, P. K.; Weiner, R. L.; Luscombe, V. B.; Zhan, X.; Chang, S.; Cho, H. P.; Rodriguez, A. L.; Niswender, C. M.; Conn, P. J.; Bridges, T. M.; Engers, D. W.; Lindsley, C. W., Discovery of a novel, CNS penetrant M4 PAM chemotype based on a 6-fluoro-4-(piperidin-1-yl)quinoline-3-carbonitrile core. *Bioorg Med Chem Lett* **2017**, 27 (18), 4274-4279.
323. Kozek, K. A.; Du, Y.; Sharma, S.; Prael, F. J., 3rd; Spitznagel, B. D.; Kharade, S. V.; Denton, J. S.; Hopkins, C. R.; Weaver, C. D., Discovery and Characterization of VU0529331, a Synthetic Small-Molecule Activator of Homomeric G Protein-Gated, Inwardly Rectifying, Potassium (GIRK) Channels. *ACS Chem Neurosci* **2019**, 10 (1), 358-370.
324. Ellingson, S. R.; Smith, J. C.; Baudry, J., VinaMPI: facilitating multiple receptor high-throughput virtual docking on high-performance computers. *Journal of computational chemistry* **2013**, 34 (25), 2212-21.
325. McGovern, S., Promiscuous Ligands. *Comprehensive Medicinal Chemistry II* **2007**, 737-52.
326. Athanasiadis, E.; Cournia, Z.; Spyrou, G., ChemBioServer: a web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery. *Bioinformatics* **2012**, 28 (22), 3002-3.
327. rdkit.org RDKit.
328. Zhang, J. H.; Chung, T. D.; Oldenburg, K. R., A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J Biomol Screen* **1999**, 4 (2), 67-73.

329. Leeson, P. D.; Springthorpe, B., The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat Rev Drug Discov* **2007**, 6 (11), 881-90.
330. Lipinski, C. A., Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol* **2004**, 1 (4), 337-41.
331. Giordanetto, F.; Jin, C.; Willmore, L.; Feher, M.; Shaw, D. E., Fragment Hits: What do They Look Like and How do They Bind? *J Med Chem* **2019**, 62 (7), 3381-3394.
332. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D., Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* **2002**, 45 (12), 2615-23.
333. Edwards, M. P.; Price, D. A., Role of Physicochemical Properties and Ligand Lipophilicity Efficiency in Addressing Drug Safety Risks. *Annu Rep Med Chem* **2010**, 45, 381-391.
334. Oprea, T. I., Property distribution of drug-related chemical databases. *J Comput Aided Mol Des* **2000**, 14 (3), 251-64.
335. Choy, Y. B.; Prausnitz, M. R., The rule of five for non-oral routes of drug delivery: ophthalmic, inhalation and transdermal. *Pharm Res* **2011**, 28 (5), 943-8.
336. Khanna, V.; Ranganathan, S., Physiochemical property space distribution among human metabolites, drugs and toxins. *BMC Bioinformatics* **2009**, 10 Suppl 15 (Suppl 15), S10.
337. Salentin, S.; Haupt, V. J.; Daminelli, S.; Schroeder, M., Polypharmacology rescored: protein-ligand interaction profiles for remote binding site similarity assessment. *Prog Biophys Mol Bio* **2014**, 116 (2-3), 174-86.
338. Sawada, T.; Fedorov, D. G.; Kitaura, K., Role of the key mutation in the selective binding of avian and human influenza hemagglutinin to sialosides revealed by quantum-mechanical calculations. *J Am Chem Soc* **2010**, 132 (47), 16862-72.
339. Imai, Y. N.; Inoue, Y.; Nakanishi, I.; Kitaura, K., Cl-pi interactions in protein-ligand complexes. *Protein Sci* **2008**, 17 (7), 1129-37.
340. Chen, D.; Oezguen, N.; Urvil, P.; Ferguson, C.; Dann, S. M.; Savidge, T. C., Regulation of protein-ligand binding affinity by hydrogen bond pairing. *Sci Adv* **2016**, 2 (3), e1501240.
341. Gancia, E.; Montana, J. G.; Manallack, D. T., Theoretical hydrogen bonding parameters for drug design. *J Mol Graph Model* **2001**, 19 (3-4), 349-62.
342. Fujita, T.; Nishioka, T.; Nakajima, M., Hydrogen-bonding parameter and its significance in quantitative structure--activity studies. *J Med Chem* **1977**, 20 (8), 1071-81.
343. Lawlor, L.; Yang, X. B., Harnessing the HDAC-histone deacetylase enzymes, inhibitors and how these can be utilised in tissue engineering. *Int J Oral Sci* **2019**, 11 (2), 20.
344. Berger, S. L., The complex language of chromatin regulation during transcription. *Nature* **2007**, 447 (7143), 407-12.

345. Di Gennaro, E.; Bruzzese, F.; Caraglia, M.; Abruzzese, A.; Budillon, A., Acetylation of proteins as novel target for antitumor therapy: review article. *Amino acids* **2004**, 26 (4), 435-41.
346. Bolden, J. E.; Peart, M. J.; Johnstone, R. W., Anticancer activities of histone deacetylase inhibitors. *Nat Rev Drug Discov* **2006**, 5 (9), 769-84.
347. Stengel, K. R.; Hiebert, S. W., Class I HDACs Affect DNA Replication, Repair, and Chromatin Structure: Implications for Cancer Therapy. *Antioxid Redox Signal* **2015**, 23 (1), 51-65.
348. Minucci, S.; Pelicci, P. G., Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer. *Nature reviews. Cancer* **2006**, 6 (1), 38-51.
349. Suresh, P. S.; Devaraj, V. C.; Srinivas, N. R.; Mullangi, R., Review of bioanalytical assays for the quantitation of various HDAC inhibitors such as vorinostat, belinostat, panobinostat, romidepsin and chidamine. *Biomedical chromatography : BMC* **2017**, 31 (1).
350. Zhao, C.; Dong, H.; Xu, Q.; Zhang, Y., Histone deacetylase (HDAC) inhibitors in cancer: a patent review (2017-present). *Expert Opin Ther Pat* **2020**, 30 (4), 263-274.
351. Xu, W. S.; Parmigiani, R. B.; Marks, P. A., Histone deacetylase inhibitors: molecular mechanisms of action. *Oncogene* **2007**, 26 (37), 5541-52.
352. de Ruijter, A. J.; van Gennip, A. H.; Caron, H. N.; Kemp, S.; van Kuilenburg, A. B., Histone deacetylases (HDACs): characterization of the classical HDAC family. *Biochem J* **2003**, 370 (Pt 3), 737-49.
353. Finkel, T.; Deng, C. X.; Mostoslavsky, R., Recent progress in the biology and physiology of sirtuins. *Nature* **2009**, 460 (7255), 587-91.
354. Jamaladdin, S.; Kelly, R. D.; O'Regan, L.; Dovey, O. M.; Hodson, G. E.; Millard, C. J.; Portolano, N.; Fry, A. M.; Schwabe, J. W.; Cowley, S. M., Histone deacetylase (HDAC) 1 and 2 are essential for accurate cell division and the pluripotency of embryonic stem cells. *Proc Natl Acad Sci U S A* **2014**, 111 (27), 9840-5.
355. Yang, T.; Jian, W.; Luo, Y.; Fu, X.; Noguchi, C.; Bungert, J.; Huang, S.; Qiu, Y., Acetylation of histone deacetylase 1 regulates NuRD corepressor complex activity. *J Biol Chem* **2012**, 287 (48), 40279-91.
356. Fischle, W.; Dequiedt, F.; Hendzel, M. J.; Guenther, M. G.; Lazar, M. A.; Voelter, W.; Verdin, E., Enzymatic Activity Associated with Class II HDACs Is Dependent on a Multiprotein Complex Containing HDAC3 and SMRT/N-CoR. *Mol Cell* **2002**, 9 (1), 45-57.
357. Bottomley, M. J.; Lo Surdo, P.; Di Giovine, P.; Cirillo, A.; Scarpelli, R.; Ferrigno, F.; Jones, P.; Neddermann, P.; De Francesco, R.; Steinkuhler, C.; Gallinari, P.; Carfi, A., Structural and functional analysis of the human HDAC4 catalytic domain reveals a regulatory structural zinc-binding domain. *J Biol Chem* **2008**, 283 (39), 26694-704.
358. Di Giorgio, E.; Gagliostro, E.; Brancolini, C., Selective class IIa HDAC inhibitors: myth or reality. *Cellular and molecular life sciences : CMLS* **2015**, 72 (1), 73-86.

359. Guerriero, J. L.; Sotayo, A.; Ponichtera, H. E.; Castrillon, J. A.; Pourzia, A. L.; Schad, S.; Johnson, S. F.; Carrasco, R. D.; Lazo, S.; Bronson, R. T.; Davis, S. P.; Lobera, M.; Nolan, M. A.; Letai, A., Class IIa HDAC inhibition reduces breast tumours and metastases through anti-tumour macrophages. *Nature* **2017**, 543 (7645), 428-432.
360. Hudson, G. M.; Watson, P. J.; Fairall, L.; Jamieson, A. G.; Schwabe, J. W., Insights into the Recruitment of Class IIa Histone Deacetylases (HDACs) to the SMRT/NCoR Transcriptional Repression Complex. *J Biol Chem* **2015**, 290 (29), 18237-44.
361. Kim, G. S.; Jung, H. E.; Kim, J. S.; Lee, Y. C., Mutagenesis Study Reveals the Rim of Catalytic Entry Site of HDAC4 and -5 as the Major Binding Surface of SMRT Corepressor. *PLoS One* **2015**, 10 (7), e0132680.
362. Jubb, H.; Blundell, T. L.; Ascher, D. B., Flexibility and small pockets at protein-protein interfaces: New insights into druggability. *Prog Biophys Mol Bio* **2015**, 119 (1), 2-9.
363. Laraia, L.; McKenzie, G.; Spring, D. R.; Venkitaraman, A. R.; Huggins, D. J., Overcoming Chemical, Biological, and Computational Challenges in the Development of Inhibitors Targeting Protein-Protein Interactions. *Chem Biol* **2015**, 22 (6), 689-703.
364. http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html.
365. Lowe, D., The First Deuterated Drug Arrives. **2017**.

VITA

Rupesh Agarwal got his bachelor's degree in Biotechnology at West Bengal University of Technology in Kolkata, India. In 2014, he moved to Knoxville, Tennessee, USA to join the Ph.D. program in Genome Science and Technology at the University of Tennessee, Knoxville with a major in Life Sciences. During his Ph.D., he performed research at the Center for Molecular Biophysics (CMB) at the Oak Ridge National Laboratory under the supervision of Dr. Jeremy C. Smith. During his time at CMB, he worked on multiple projects and published two first author papers along with multiple co-author papers.

List of papers published (as of August 2020):

1. **Rupesh Agarwal**, Micholas D. Smith, Jeremy C. Smith. "Capturing deuteration effects in classical molecular dynamics simulations: Mixing behavior of tetrahydrofuran in aqueous solution". *Journal of Chemical Theory and Computation*, doi.org/10.1021/acs.jctc.9b01138 (2020)
2. **Rupesh Agarwal**, Utsab R. Shrestha, Loukas Petridis, Xiang-Qiang Chu, Jeremy C. Smith. "Mesophilic pyrophosphatase function at high temperature: a molecular dynamics simulation study". *Biophysical Journal*, doi.org/10.1016/j.bpj.2020.05.021 (2020)
3. **Rupesh Agarwal**, Barbara A. Bensing, Dehui Mi, Paige N. Vinson, Jerome Baudry, Tina M. Iverson, and Jeremy C. Smith. "Structure based virtual screening identifies novel competitive inhibitors for the sialoglycan binding protein Hsa." preprint: <https://doi.org/10.1101/2020.03.27.006247>. *Biochemical Journal* (2020)
4. Aranha, Michelle P., Thomas A. Penfound, Jay A. Spencer, **Rupesh Agarwal**, Jerome Baudry, James B. Dale, and Jeremy C. Smith. "Structure-based group A streptococcal vaccine design: Helical wheel homology predicts antibody cross-reactivity among streptococcal M protein-derived peptides." *Journal of Biological Chemistry* 295, no. 12: 3826-3836 (2020)
5. Gupta, Madhulika, Khanh Ha, **Rupesh Agarwal**, L. Darryl Quarles, and Jeremy C. Smith. "Molecular Dynamics Analysis of The Binding of Human Interleukin-6 with Interleukin-6 α -Receptor." *Proteins*.

6. Scott Le Grand, Aaron Scheinberg, Andreas Tillack, Mathialakan Thavappiragasam, Josh Vermaas, Ada Sedova, **Rupesh Agarwal**, Jeremy C. Smith. "GPU-Accelerated Drug Discovery with Docking on the Summit Supercomputer: Porting, Optimization, and Application to COVID-19 Research", ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB) 2020
7. Barbara A. Bensing, Lioudmila V. Loukachevitcha, **Rupesh Agarwal**, Izumi Yamakawaa, Kelvin Luong, Azadeh Hadadianpour, Hai Yu et al. "Selectivity and engineering of the sialoglycan-binding spectrum in Siglec-like adhesins." preprint: <https://doi.org/10.1101/796912> (2019). Under revision (Proceedings of the National Academy of Sciences)
8. Acharya, Atanu, **Rupesh Agarwal**, Matthew Baker, Jerome Baudry, Debsindhu Bhowmik, Swen Boehm, Kendall Byler et al. "Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19". preprint: doi.org/10.26434/chemrxiv.12725465.v1. Under review (Journal of Chemical Information and Modeling)

For updated list: [Google Scholar profile](#)