



HAL
open science

Structure-based design of glycomimetic ligands for the N-terminal domain of BC2L-C lectin

Kanhaya Lal

► **To cite this version:**

Kanhaya Lal. Structure-based design of glycomimetic ligands for the N-terminal domain of BC2L-C lectin. Biochemistry, Molecular Biology. Université Grenoble Alpes [2020-..]; Università degli studi (Milan, Italie), 2021. English. NNT: 2021GRALV072 . tel-03675242

HAL Id: tel-03675242

<https://theses.hal.science/tel-03675242>

Submitted on 23 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



UNIVERSITÀ DEGLI STUDI DI MILANO



UNIVERSITÀ DEGLI STUDI DI MILANO
CORSO DI DOTTORATO IN CHIMICA
XXXIII CICLO

UNIVERSITÉ GRENOBLE ALPES
DOCTORATE in CHEMISTRY-BIOLOGY

Dipartimento di Chimica

Doctoral School of Chemistry
and Life Sciences

TESI DI DOTTORATO DI RICERCA

STRUCTURE-BASED DESIGN OF GLYCOMIMETIC LIGANDS FOR THE N-
TERMINAL DOMAIN OF BC2L-C LECTIN

Settore scientifico disciplinare CHIM/06

Dottorando: Kanhaya LAL
Matr. 12139

Tutor: Prof. Anna BERNARDI
Prof. Laura BELVISI

Tutor: Dr. Anne IMBERTY
Dr. Annabelle VARROT

COORDINATORE: Prof. Emanuela LICANDRO



PhD4GlycoDrug
Marie Skłodowska-Curie Innovative Training Network
H2020-MSCA-ITN-2017-EJD-765581

A.A.
2019/2020

THESIS

To obtain the degree of

DOCTOR OF GRENOBLE ALPES UNIVERSITY

**Prepared as part of a joint supervision between the
Grenoble Alpes University and the *University of
Milan***

Specialty: **CHEMISTRY - BIOLOGY** and **CHEMISTRY**

Ministerial decree: January 6, 2005 - May 25, 2016

Presented by

Kanhaya LAL

Thesis supervised by **Prof. Anna BERNARDI**, **Prof. Laura BELVISI**, **Dr. Anne IMBERTY** and **Dr. Annabelle VARROT**

prepared in the **Center for Research on Plant Macromolecules** and the **Chemistry Department of University of Milan**

in the **Doctoral School of Chemistry and Life Sciences** and the **Doctoral School of Chemistry of University of Milan**

Structure-based design of glycomimetic ligands for the N-terminal domain of BC2L-C lectin

Thesis publicly defended on **15th December 2021**, before the jury composed by

Dr. Sandrine PY

Research Director CNRS - UGA, Jury President

Prof. Paola CONTI

University of Milan, Examiner

Prof. Marko ANDERLUH

University of Ljubljana, Examiner

Prof. Elisa FADDA

Maynooth University, Referee

Prof. Sonsoles MARTIN-SANTAMARIA

Center for Biological Research, Spanish National Research Council (CSIC), Referee



ACKNOWLEDGEMENTS

I gladly take this opportunity to acknowledge those great people who have inspired, supported and motivated me during the course of my research work. I find myself short of words while expressing my sense of gratitude and indebtedness towards my supervisors, Prof. Anna Bernardi, Prof. Laura Belvisi, Dr. Anne Imberty and Dr. Annabelle Varrot. It is my honour and rare privilege to have worked under the guidance and supervision of revered mentors. Their wise and timely advice, constant encouragement and painstaking supervision are greatly and generously recognized. I would like to express my sense of gratitude towards Dr. Martin Lepsik and Dr. Monica Civera for providing their suggestions, necessary help and support all the way through my project. I am highly obliged to Dr. Serge Perez, Prof. Francesca Vasile, Dr. Jonathan Cramer and Prof. Beat Ernst for their great collaboration, valuable suggestions and ever available help. I would like to express my appreciation to Dr. Marko Anderluh for being a great coordinator in the PhD₄GlycoDrug network and for actively working for all the research fellows in the network.

I'm thankful to all the research fellows in the network for being such a great co-worker. Their enthusiasm, energy and sense of humour have made my experience more memorable. I appreciate all the help, encouragement and moral support provided by them in all stages of this uphill task. I extend my sincere appreciation to Rafael Bermeo for amazing collaboration to achieve the planned objectives of the project. Special thanks to Valérie Chazalet and Emilie Gillon for their full hearted co-operation and support during my research work at the CERMAV. Last but not the least I am deeply respectful to my beloved parents and family members for their support and continuous remembrance in their prayers for my success.

Kanhaya Lal

LIST OF ABBREVIATIONS

AAT	Anti-adhesion therapy	GlcNAc	N-Acetylglucosamine
ADP	Adenosine diphosphate	HTS	High-throughput screening
AMR	Anti-microbial resistance	HMO	Human milk oligosaccharide
APC	Antigen-presenting cells	Ig	Immunoglobulin
BCC	<i>Burkholderia cepacia</i> complex	IPTG	Isopropyl β -D-thiogalactoside
CF	Cystic fibrosis	ITC	Isothermal titration calorimetry
CRD	Carbohydrate recognition domain	IMAC	Immobilized metal affinity chromatography
CLR	C-type lectin receptors	K _d	Dissociation constant
DC-SIGN	Dendritic cell-specific intercellular adhesion molecule-3-grabbing non-integrin	LB	Luria broth
DMSO	Dimethyl sulfoxide	LPS	Lipolysaccharide
DSF	Differential scanning fluorimetry	MD	Molecular dynamics
DSC	Differential scanning calorimetry	MDR	Multidrug-resistant
EM	Electron microscopy	Me- α -L-Fuc	Methyl alpha-L-fucopyranoside
FITC	Fluorescein 5-isothiocyanate	MST	Microscale thermophoresis
Fuc	Fucose	MVMp	Virus of mice prototype strain
GAGs	Glycosaminoglycans	NMR	Nuclear magnetic resonance
GalNAc	N-Acetylgalactosamine	NOE	Nuclear Overhauser Effect
Gal	Galactose	OD	Optical density
GIST	Grid inhomogeneous solvation theory	PDB	Protein data bank
		RESP	Restrained electrostatic potential
		PMEMD	Particle mesh Ewald molecular dynamics

SARS-CoV	Severe acute respiratory syndrome coronavirus
SAXS	Small angle X-ray scattering
SBDD	Structure-based drug design
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SSL	Superantigen-like protein
STD-NMR	Saturation transfer difference nuclear magnetic resonance
TB	Tuberculosis
TCEP	Tris(2-carboxyethyl)phosphine
T _m	Melting temperature
TNF	Tumor necrosis factor
TRIC	Temperature related intensity change
TSA	Thermal shift assay
UPEC	Uropathogenic <i>Escherichia coli</i>
UTI	Urinary tract infection
WHO	World health organisation

Table of contents

1. Introduction.....	7
1.1 Antimicrobial resistance (AMR): A serious global public health threat.....	7
1.2 Anti-adhesion therapy: A promising alternative.....	11
1.3 Carbohydrates and lectins in bacterial adhesion.....	14
1.3.1 Pathogen strategies to invade cells	15
1.3.2 Glycocalyx at the surface of human cells.....	20
1.3.3 Bacterial adhesins and soluble lectins involved in adhesion	23
1.4 Glycans and glycomimetics as anti-adhesion compounds.....	27
1.4.1 FimH and LecA antagonists as anti-adhesion compounds.....	31
1.5 <i>Burkholderia cenocepacia</i> : An opportunistic pathogen that targets glycans.....	34
1.5.1 The Burkholderia cepacia complex.....	34
1.5.2 Lectins from Burkholderia cenocepacia.....	35
1.6 Thesis objective.....	42
1.7 The PhD4GlycoDrug consortium.....	44
1.8 References	45
2. Research methodology	60
2.1 Introduction: The domain of structure-based drug design	60
2.2 Prediction of druggable sites using SiteMap	61
2.3 Virtual screening: docking studies	64
2.3.1 Sampling.....	65
2.3.2 Scoring.....	67
2.3.3 Glide software.....	71
2.4 Molecular dynamics (MD) simulations	73
2.4.1 Statistical mechanics.....	73

2.4.2	Thermodynamic state, microscopic state and ensembles.....	74
2.4.3	Calculation of averages from molecular dynamics simulations.....	74
2.4.4	Classical mechanics.....	76
2.4.5	Integration algorithms.....	77
2.4.6	Constant temperature dynamics.....	78
2.4.7	Constant pressure dynamics.....	80
2.4.8	Molecular mechanics.....	82
2.4.9	Periodic boundary conditions.....	83
2.5	Analysis of water thermodynamics using Grid Inhomogeneous Solvation Theory (GIST) ...	84
2.6	Recombinant protein (BC2L-C-nt) expression and purification.....	86
2.7	Thermal shift assay (TSA).....	88
2.8	Microscale thermophoresis (MST).....	90
2.9	Saturation transfer difference (STD)-NMR.....	92
2.10	X-ray crystallography.....	93
2.10.1	Crystallogensis.....	94
2.10.2	Data processing, structure determination, and refinement.....	96
2.11	Isothermal titration calorimetry (ITC).....	99
2.12	References.....	102
3.	Analysis of crystal structures and fragment screening in BC2L-C-nt.....	109
3.1	Analysis of the binding site in crystal structures.....	109
3.2	Resolution of first crystal structure of the apo form of BC2L-C-nt.....	111
3.3	Prediction of druggable sites.....	115
3.4	Virtual screening.....	117
3.4.1	Ligand preparation.....	117
3.4.2	Models for docking study.....	118
3.4.3	Docking analysis.....	119
3.4.4	Selection of best fragments from the docking results.....	121

3.5	Hit expansion	122
3.6	References	125
4.	Experimental validation of fragment binding	127
4.1	Protein production and purification	127
4.2	Thermal shift assay (TSA)	129
4.3	Microscale Thermophoresis (MST)	131
4.4	STD-NMR analysis of fragment binding	134
4.5	Crystal structure of the complex KL3-BC2L-C-nt.....	139
4.6	Affinity analysis using ITC	144
4.7	Summary	145
4.8	Experimental section	146
4.8.1	Protein expression and purification.....	146
4.8.2	Thermal shift assay (TSA)	147
4.8.3	Microscale Thermophoresis (MST)	148
4.8.4	STD-NMR interaction studies.....	148
4.8.5	X-ray crystallography, data collection, and structure determination.....	148
4.8.6	ITC measurements	149
4.9	References	150
5.	Design of bifunctional glycomimetic ligands	152
5.1	Strategies to connect fragments to the sugar core	152
5.2	A new model for the docking of fucoside-linker conjugates	152
5.3	Selection and building of bifunctional glycomimetic ligands.....	154
5.4	Docking studies of bifunctional glycomimetic ligands.....	157
5.5	MD simulations of BC2L-C-nt complexes with glycomimetics.....	164
5.5.1	Force field parameterization of glycomimetic ligands.....	164
5.5.2	MD simulation setup.....	165
5.5.3	Molecular dynamics simulations of methyl α -L-fucoside (Me- α -L-Fuc) in complex with BC2L-C-nt	167

5.5.4	MD simulations of glycomimetic ligands Lfuc-Aky-KL07 and Lfuc Aky-KL08 in complex with BC2L-C-nt	169
5.5.5	MD simulation of glycomimetic ligand Lfuc-Amd-KL13 in complex with BC2L-C-nt ..	170
5.6	Experimental validation of glycomimetic binding using ITC and X-ray crystallography	172
5.7	Strategies to improve binding affinity of the glycomimetics.....	174
5.8	Analysis of water thermodynamics at binding site using Grid Inhomogeneous Solvation Theory (GIST).....	175
5.9	Docking studies of additional ligands with guanidine moiety	177
5.10	References	179
Annexure: Chapter 5.....		182
6.	Conclusions and perspectives	193
7.	Scientific communication: Short secondment at Glycopedia	196
8.	Identification of druggable allosteric pockets in β-propeller lectins	218
8.1	Methods.....	219
8.1.1	Binding site prediction	219
8.1.2	Preparation of protein model	219
8.1.3	Preparation of ligand models.....	220
8.1.4	Models for docking study.....	220
8.2	Results.....	220
8.2.1	Binding site analysis	220
8.2.2	Docking analysis	223
8.3	Conclusions	224
8.4	References	225
Appendix 1		226

1. Introduction

1.1 Antimicrobial resistance (AMR): A serious global public health threat

The emergence and spread of infectious diseases with pandemic potential have been reported in the history and some of them are still present in the current times. Several infectious diseases such as cholera, plague, flu, Middle East respiratory syndrome coronavirus (MERS-CoV) and severe acute respiratory syndrome coronavirus (SARS-CoV) turned into epidemics and pandemics that have afflicted humanity. In the 20th century, the rapid development of several vaccines provided a prophylactic means to combat some of these infectious diseases. More importantly, the discovery of antibiotics paved the way for the direct fight against various bacterial infections. Thus, introduction of antibiotics has been a bedrock of greatest medical advances of the 20th century. Because of these discoveries, deadly illnesses such as pneumonia and tuberculosis (TB) could be treated effectively, and a routine surgery was no longer potentially fatal. However, within a short period, it was also discovered that bacteria and other pathogens have evolved to resist the drugs used to combat them. Antimicrobial resistance (AMR) enables the pathogens to resist to the effects of an antibiotic or drug that would usually kill them or limit their growth.² Hence, all microbes that have the potential to mutate can render the drugs ineffective. Consequently, over several decades, pathogens causing common or severe infections have developed resistance (to various extent) to the new antibiotic coming to market. Penicillin, discovered by Alexander Fleming in 1928, was the first antibiotic to treat bacterial infections in soldiers during World War II. In the early 1940s, the use of penicillin to treat staphylococcal infection showed dramatic success. However, after a short period of time, in early 1942, penicillin-resistant staphylococci were reported.³ Surprisingly, by the late 1960s, more than 80% of

staphylococcal isolates were resistant to penicillin. Similar pattern of resistance, has been well-established with several antibiotics discovered later (**Table 1.1**).

Table 1.1 A brief history of antibiotics and resistance. Adapted from the U.S. Centers for Disease Control and Prevention.⁴

Antibiotic	Year released	Resistant strain	Year identified
Penicillin	1941	Penicillin-resistant <i>Staphylococcus aureus</i>	1942
		Penicillin-resistant <i>Streptococcus pneumoniae</i>	1967
		Penicillinase-producing <i>Neisseria gonorrhoeae</i>	1976
Vancomycin	1958	Plasmid-mediated vancomycin-resistant <i>Enterococcus faecium</i>	1988
		Vancomycin-resistant <i>Staphylococcus aureus</i>	2002
Amphotericin B	1959	Amphotericin B-resistant <i>Candida auris</i>	2016
Methicillin	1960	Methicillin-resistant <i>Staphylococcus aureus</i>	1960
Extended-spectrum cephalosporins	1980	Extended-spectrum beta-lactamase-producing <i>Escherichia coli</i>	1983
Azithromycin	1980	Azithromycin-resistant <i>Neisseria gonorrhoeae</i>	2011
Imipenem	1985	<i>Klebsiella pneumoniae</i> carbapenemase (KPC)-producing <i>Klebsiella pneumoniae</i>	1996
Ciprofloxacin	1987	Ciprofloxacin-resistant <i>Neisseria gonorrhoeae</i>	2007
Fluconazole	1990 (FDA approved)	Fluconazole-resistant <i>Candida</i>	1988
Caspofungin	2001	Caspofungin-resistant <i>Candida</i>	2004
Daptomycin	2003	Daptomycin-resistant methicillin-resistant <i>Staphylococcus aureus</i>	2004
Ceftazidime-avibactam	2015	Ceftazidime-avibactam-resistant KPC-producing <i>Klebsiella pneumoniae</i>	2015

The period between 1960 and 1980 is known as the golden age of antibiotic discovery, as one-half of the drugs commonly used today were discovered in these two decades. After the 1980s, the rate of discovery of new antibiotic classes had dramatically decreased. The recent report from the World Health Organization (WHO) shows that almost all the new

antibiotics that have obtained marketing authorisation in recent decades are the derivatives of the existing antibiotic classes.⁵ However, two recently approved agents, vaborbactam (a β -lactamase-inhibitor based on a cyclic boronate pharmacophore) and lefamulin (a pleuromutilin) belong to new drug classes.⁵⁻⁹ In addition, the same report highlights that the antibacterial drugs already in the early stages of clinical development are not effective against extensively drug-resistant bacteria.⁵ This analysis indicates a weak pipeline for antibiotic agents. Furthermore, due to the extensive use of antibiotics in 20th century, several of the bacterial pathogens have evolved into multidrug-resistant (MDR) forms. Such microbes with enhanced morbidity and mortality due to multiple mutations, providing high levels of resistance to the antibiotics are called “superbugs”. For instance, fluoroquinolones are antibiotics that target bacterial enzymes known as DNA topoisomerases II (DNA gyrase) and IV. These enzymes are essential for the supercoiling of bacterial DNA. Both enzymes consist of subunits which are encoded by *gyrA* and *gyrB* (for DNA gyrase) or *parC* and *parE* (for topoisomerase IV).¹⁰ The resistance to fluoroquinolones is developed by accumulation of amino-acid substitutions in these subunits.¹¹ Likewise, linezolid is another example of antibiotic which belongs to oxazolidinone class.¹² This antibiotic prevents protein synthesis in bacteria by inhibiting formation of the 70S ribosomal initiation complex. Bacterial strains such as *enterococci*, *Staphylococcus aureus* and *streptococci* display resistance to linezolid. The resistance is mediated mainly by mutations in the genes that encode 23S rRNA. Particularly, G2576T mutation is common among the resistant clinical isolates.¹³ These highly resistant strains sometimes show increased virulence and transmissibility. Infections to patients by such pathogens are usually hospital-linked as they are associated with several risk factors and comorbidities such as cancer, diabetes and immunosuppression etc. These patients with existing medical conditions are usually prone to the ‘opportunistic’ infections.¹⁴⁻¹⁶ Some of

the examples of these “superbugs” are *Burkholderia cepacia*, *Acinetobacter baumannii*, *Campylobacter jejuni*, *Citrobacter freundii*, *Klebsiella pneumoniae*, *Proteus mirabilis*, *Clostridium difficile*, *Enterobacter* spp., *Enterococcus faecium*, *Enterococcus faecalis*, *Escherichia coli*, *Haemophilus influenzae*, *Pseudomonas aeruginosa*, *Salmonella* spp., *Serratia* spp., *Staphylococcus aureus*, *Staphylococcus epidermidis*, *Stenotrophomonas maltophilia*, and *Streptococcus pneumoniae*.¹⁷

The emergence of AMR has challenged the effective prevention and treatment of an ever-increasing range of infections by these highly drug-resistant pathogens. It has emerged as one of the principal public health threat of the 21st century. As per the estimates (**Figure 1.1**), at least 700,000 people die annually from drug-resistance infections with a future prediction of 10 million deaths per year by 2050.¹ This ten-fold increase can overtake cancer by 2050 to become one of the biggest threats to public health. In terms of economy, the loss due to AMR is predicted in trillions of US dollars which is estimated about 7 percent of the gross domestic product (GDP) of the world.¹

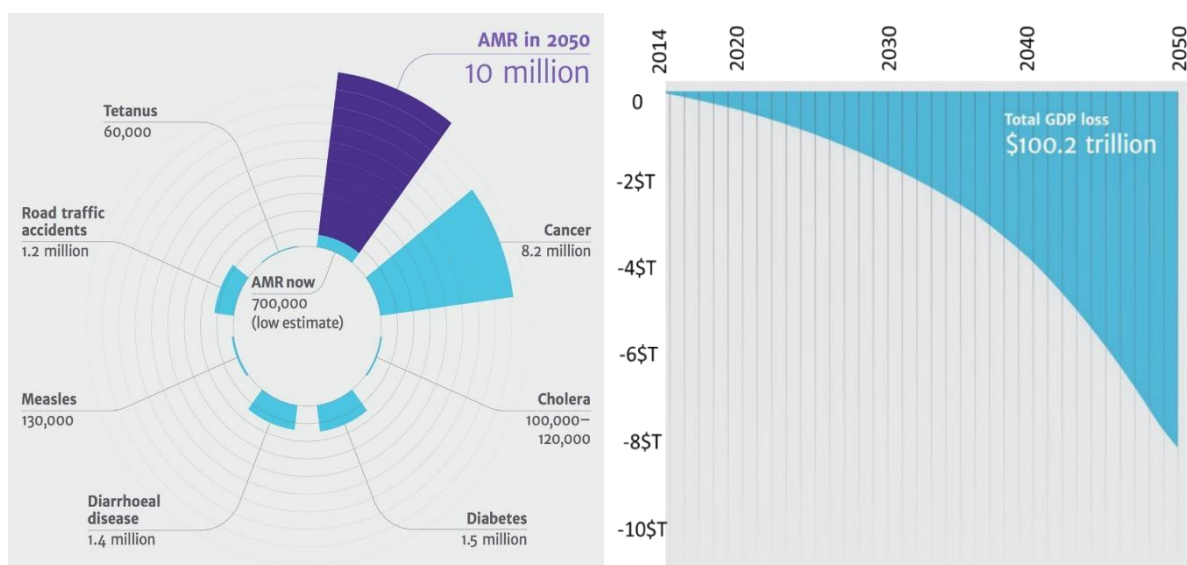


Figure 1.1 (A) As per the estimates, deaths due to antimicrobial resistance could grow to 10 million by 2050. (B) The economic loss due to the impact of AMR on the world's economy is predicted in trillions of US dollars between 2014 and 2050. Adapted from O’Neill (2014).¹

As a consequence of increasing antimicrobial resistance and a narrow antimicrobial pipeline of new drugs, rise in bacterial infections due to multidrug-resistant or extensively drug-resistant pathogens has already rang the alarm bell. To counteract AMR, it is highly important to gear up the efforts toward the discovery and development of novel antimicrobials. Identification of new drug targets and use of innovative strategies to decrease the infections by drug-resistant pathogens are therefore a necessary public health concern. An attractive strategy is the use of anti-adhesive agents that interfere with the ability of bacteria to adhere to host tissues, since bacterial adhesion is one of the initial stages of infections.

1.2 Anti-adhesion therapy: A promising alternative

Adhesion of pathogen to host cell is a universal prerequisite to efficiently deploy repertoire of virulence factors and exert effects on host cells. Upon encountering the host cell surface, the initial attachment of bacteria is mediated by weak non-specific interactions which depend on the physicochemical and electrostatic interaction between the bacteria and the substrate. Thus, the primary attachment during the planktonic phase is weak and reversible that relies on the environmental factors like pH, ionic forces and temperature.^{18,19} Finally, irreversible attachment takes place with the help of fimbria.²⁰ The cells that attach irreversibly to surfaces (i.e. not removed by gentle rinsing) undergo cell division and form microcolonies. This leads to the formation of extracellular polymers that define a biofilm.²⁰ These strong interactions are mediated by molecules on the host-bacterial surface which are mostly sugars, proteins or lipids. Subsequently, the mature microbial cells are detached and dispersed. This process continues as the released microbial cells from a mature biofilms attach to new surfaces. **Figure 1.2** depicts the stages of biofilm formation.

After adhesion, bacterium's ability to colonize its host highly depends on the mechanisms to withstand the host's mechanical and immunological clearance mechanisms.

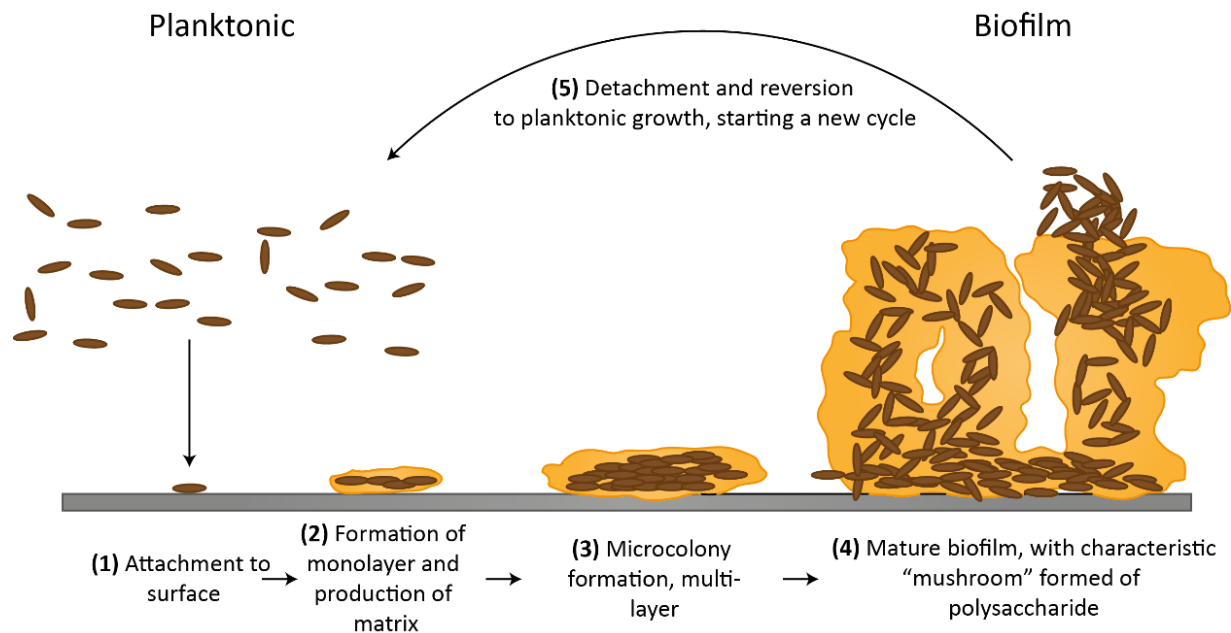


Figure 1.2 Representation of different stages in a biofilm formation. (1) As a first step, planktonic cells attach reversibly and then adhere to the surface. (2) Subsequently, bacteria form a monolayer with irreversible adhesion and also produce an extracellular matrix. (3) Next, multilayers appear to form a microcolony. (4) In later stages, the mature biofilm forms “mushroom like” structure containing polysaccharides. (5) Finally, the cells are dispersed from the biofilm to start a new cycle. Adapted from Vasudevan (2014).²⁰

Different regions of the human body are equipped with a natural clearance mechanism which protects the body from infections. Airflow in the respiratory tract, mucus removal from the airway, urine flow in the urinary tract, lining of tracts and tissue with antibodies are some of the natural cleansing mechanisms of the host. All these mechanisms prevent the host cells from the bacterial adhesion. In addition, there are host-derived anti-adhesion molecules which can specifically bind to the entrapped pathogens and prevent their attachment to the underlying epithelial cells.²¹⁻²² For example, sulfated gastric mucin can prevent the bacterial binding to host cells.²³ However, all these natural mechanisms of the host cell are not sufficient to prevent the bacterial infection because bacteria with their adhesives resist this

mechanism and causes infections. To facilitate the infection process, bacteria have to be able to quickly and effectively attach to host cells. Subsequently, through biofilm formation, the bacterial cells acquire essential nutrients by entrapping minerals and host components such as fibrin, red blood cells (RBCs), and platelets which further enhance their ability to survive and infect the host.²⁴ The direct contact between bacteria and host cells also facilitates translocation of effector proteins from the bacterial cytoplasm into the host cell's cytoplasm.²⁵ The direct interaction should be long enough to allow the transfer of proteins over the time. The studies have shown that the transfer of proteins between host and bacterial cell follow a sequence and do not occur simultaneously.²⁶⁻²⁷ Therefore, if bacteria are removed from host cells prematurely, infection can be prevented.²⁸ Secondly, microbes which will be disabled to bind to the host cell will not be subjected to sustained selective pressure, that may occur with antibiotic therapy. Persistent use of antibiotics usually kills the non-resistant bacteria while a small population of bacteria with the mutation in the binding site residues can multiply further. Adhesion of these bacteria to host cells can resist the natural cleaning mechanisms of the body, therefore allowing the bacteria to reach a population whereby an infection can start. Anti-adhesion strategies can prevent these host-pathogen interactions and facilitate the removal by the host. Several strategies can be employed to counter with bacterial adhesion to the host cells¹⁸ (**Figure 1.3**) which involved interfering with the receptor biosynthesis,²⁹ coating the target substrate,³⁰ use of anti-adhesion antibodies³¹ or adhesin analogs.³² Thus, anti-adhesion therapy could be used as a novel approach to prevent or treat bacterial infections by targeting bacterial virulence properties (e.g. adhesion, colonization, invasion) as alternative strategy to antibiotic therapy.³³

The mechanism of bacterial adhesion has evolved with the better understanding of the host pathogen interactions. The role of different carbohydrate-binding molecules (lectins, toxins, adhesins) as virulence factors specifically targeting their corresponding epitopes in the host-pathogen interface has been already identified.³⁴ The studies show that microbial

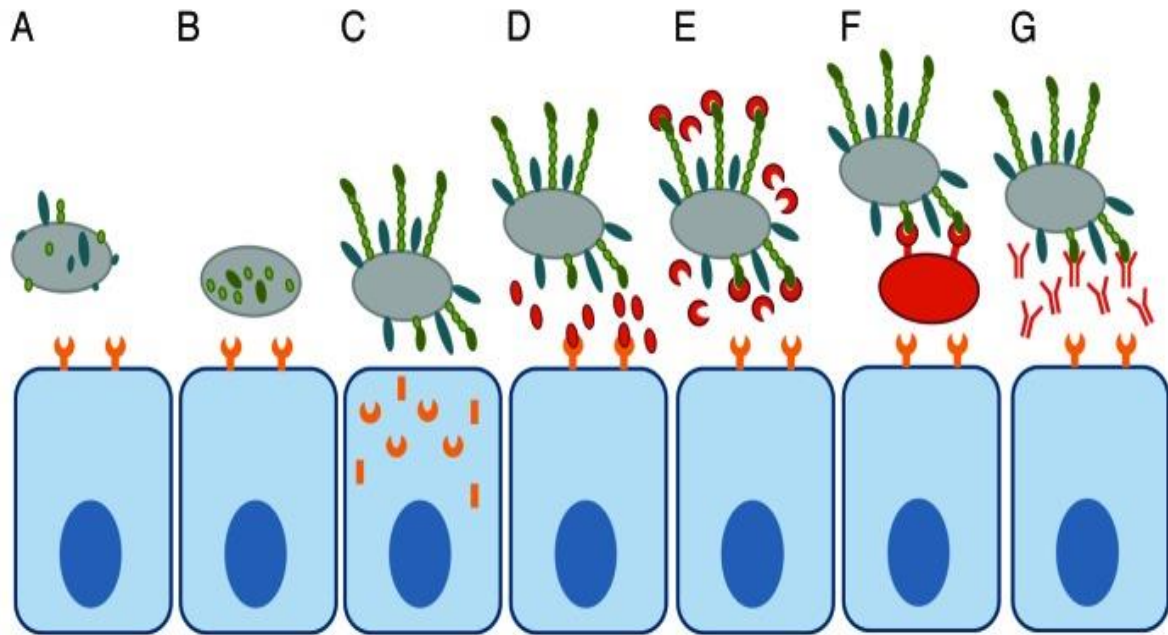


Figure 1.3 Different strategies for anti-adhesion therapy to counteract bacterial infection. Bacterial adhesion can be hampered by interfering with the biosynthesis of adhesin (A), assembly of adhesin (B), or the assembly of host-receptor (C). Binding inhibition can be achieved by competitive replacement of the adhesin from the receptor (D) or by competitive replacement of the host from the adhesin (E) using either soluble molecules or designer microbes (F). In order to block the surface epitopes (which are required for binding), antibodies against bacterial adhesins can be used (G). Adapted from Krachler and co-workers (2013).¹⁸

colonization is largely associated with the glycoconjugate decoration of the host cells, named the 'glycocalyx'.³⁵⁻³⁶

1.3 Carbohydrates and lectins in bacterial adhesion

Several microorganisms use glycans as target on the host cell surface to establish interactions and initiate infection. Proteins such as adhesins or lectins at the surface also mediate the binding to host cells. In addition, toxins secreted by pathogens also exploit

surface glycan for internalization using various mechanisms. Thus, a large number of pathogenic species of microorganisms depend on these interactions for infection.

1.3.1 Pathogen strategies to invade cells

Viruses, bacteria, fungi and parasites, employ carbohydrate-binding molecules (lectins, toxins, adhesins) as virulence factors to infect the host cells. Complex carbohydrates, also known as glycans decorate the surfaces of host cells and viruses mediating interactions and thus, promoting viral pathogenesis.³⁷⁻³⁹ The structural complexity and diversity of the glycans present on surface glycoproteins or at the host cell surface primarily arises from the complex biosynthetic mechanism involving several enzymes of the host cells that display different expression patterns.³⁹⁻⁴¹ As a result of post-translational modifications, the complex glycans are attached by N- or O-glycosidic bonds to proteins or they are attached to lipids. The envelop proteins or spike glycoproteins on the surface of many viruses^{38-39,42} recognised the glycan motifs on the host cells surface and establish specific interactions that contribute to virus entry into host cells and also play role in host tissue tropism. In addition, the host cell surface glycans act as general attachment factors or primary receptors for different viruses that ultimately mediate viral infection and entry. Generally, acidic glycans such as linear heparan sulfate glycosaminoglycan polysaccharides or branched glycans with terminal sialic acid on host cell surfaces may serve as initial attachment factors to the host cell. Influenza virus hemagglutinin is the well-studied example of a viral glycan-binding protein which binds to sialic acid-containing glycans present on the host cell surface.⁴³ The binding event involves fusion of the viral envelope with the endosomal membrane followed by internalization of the virus by endocytosis and subsequent release of the viral RNA into the cytosol. The host glycan-hemagglutinin interaction shows different specificities depending on the subtypes of influenza virus. These differences are caused due to structural differences in the

hemagglutinin. For example, human strains of influenza-A and -B viruses specifically bind to cells with receptor expressing *N*-acetylneuraminic acid (Neu5Ac α)2–6Gal while avian influenza viruses specifically bind to receptors with Neu5Ac α 2–3Gal-. Likewise, porcine strains show binding to the receptors with both types of linkages.

Similar to viruses, adhesion to host cell surface is an essential step of bacterial infection and pathogenesis which helps in developing resistance to natural defence mechanism and mechanical stress. Bacteria usually express several types of adherence factors (adhesins) that bind to the carbohydrate motifs on glycoproteins or glycosphingolipids located at the cell-surface. On bacterial surface, the thread like protein appendages known as fimbriae, act as adhesins and mediate carbohydrate-specific binding to the cell surfaces.⁴⁴ For instance, *Escherichia coli* uses FimH (adhesin) in fimbriae to bind specifically to mannosylated residues on human epithelial cells, thus facilitating urinary tract infection (UTI).⁴⁴⁻⁴⁵ In addition, these adhesins play role in 'catch bonds': bonds that are strengthened by tensile mechanical force.³⁴ For example, a catch bond has been reported in *E. coli* where FimH binds to mannose on epithelial cells.⁴⁶ The underlying mechanism involves force-induced structural alterations in the receptor protein from a low- to a high-affinity conformation. These catch bonds are known to mediate urinary tract infections by causing strong adhesion due to sheer stress induced during urine flow.⁴⁶ Other than adhesins, toxins and lectins are soluble molecules known to facilitate the infection process. Toxins are proteins which consist of different subunits. They usually have glycan-binding subunits that allow the toxin to combine with membrane glycoconjugates followed by internalization to deliver the functionally active toxic subunit across the membrane which ultimately leads to cell death. AB5 toxin family is a classic example of such toxins. AB5 toxins in organisms such as *E. coli* and *Bordella pertussis* consist of cytotoxic ADP-ribosyltransferase (A) domain linked to five (B5) lectin subunits which

recognizes endothelial surfaces.⁴⁷⁻⁴⁸ Few examples of adhesins are illustrated in section 1.3.3 of this chapter.

The interface of the host and pathogen also consists of specialized carbohydrate-specific proteins which play the complementary role of 'reader'. They were first reported at the end of 19th century,⁴⁹ for their ability to agglutinate erythrocytes. Later, in 1954, the term 'lectin' was proposed for these substances with blood group-specific agglutination properties.⁵⁰ In the 1990s, this term was generalised for all proteins of non-immune origin that can agglutinate erythrocytes or any other cell types. With the advancement in the elucidation of structural details of lectins, they were later classified into different classes. The initial classification was based on their specificities for carbohydrates, structural features and similarity in carbohydrate recognition domains (CRDs), or simply called as carbohydrate binding sites.⁵¹⁻⁵³ The most recent classification built on three different levels involves 35 lectin domain folds, 109 classes with 20% sequence similarity and 350 families with at least 70% sequence similarity.⁵⁴ According to their structural characters, animal lectins are divided into five main groups: C-type lectins, galectins, I-type lectins, pentraxins and P-type lectins. C-type lectins depend on the presence of Ca^{2+} ions and have conserved carbohydrate recognition domains with different specificities while the galectins bind to β -galactosides.⁵⁵ I-type lectins have immunoglobulins (Ig) like carbohydrate recognition domain (CRD) and pentraxins are composed of five monomers that produce pentameric lectins. P-type are the lectins have multiple domains including terminal one that recognize mannose 6-phosphate. However, with the increase in the number of described sequences, this classification includes several new groups.⁵⁶ UniLectin3D is a curated database with classification of lectins on the basis of origin, fold and their specificity towards carbohydrates.⁵⁶⁻⁵⁷ Thus, lectins present different types of folds such as β -sandwich, C-type, I-type (Ig fold) etc. The basic folds can serve as a

platform to form several types of carbohydrate recognition domains (CRDs) with their own characteristic properties (**Figure 1.4**). Usually, lectins have millimolar affinity for the monosaccharide ligands which is often compensated by establishing multivalent interactions,

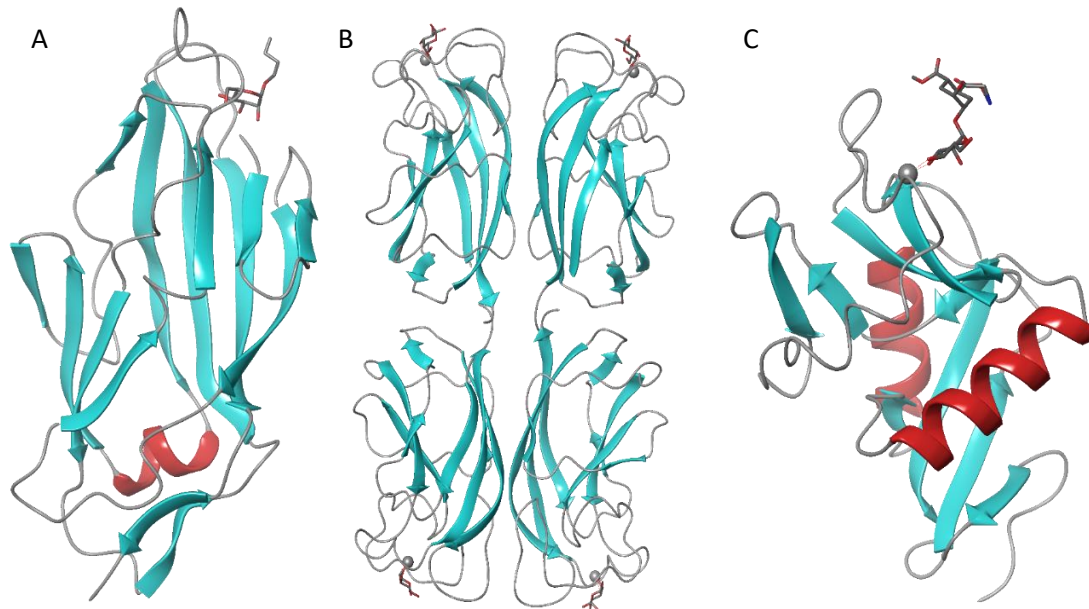


Figure 1.4 Examples of common lectin folds complexed with the ligands. Complex of (A) FimH from *E. coli* with mannoside (PDB 1TR7) and (B) LecA tetramer from *P. aeruginosa* with D-galactose showing β -sandwich fold (PDB 1OKO). (C) The complex of DC-SIGN (PDB 2XR5) from human with a mannoside showing a C-type lectin carbohydrate recognition domain with a mixed fold. The Ca^{2+} ion is shown in dark grey sphere.

mediated by the presence of several binding sites. Likewise, the topology of the lectin binding sites in space is also of importance for generating high specificity for the biological function. Recent studies have demonstrated the role of specific arrangements of lectin binding sites that favour glycolipid clusters and ultimately affect the structure and dynamics of the cell membranes.⁵⁸ Moreover, the specificity and topology in lectins have marked them as interesting tools for generating engineered 'nelectins' with applications in diagnostics, therapy and material science etc.⁵⁹

Lectins in bacteria often show specificity to sugar epitopes presented by the glycocalyx. In addition, they play other complex roles in biofilm formation, quorum sensing

etc.⁶⁰⁻⁶¹ Several pathogens use these proteins on the surface to initiate the infection process by adhesion, infection and toxicity. Some of the examples are *E.coli*, *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Vibrio cholera*, *Clostridium tetani*, etc.⁶² Few examples of lectins from bacteria are discussed in later section (1.3.3) of this chapter.

Similar to viruses and bacteria, fungi with a parasitic life cycle can act as pathogens for plants or animals. Many fungal lectins are also reported to play role in the early stages of infection. Lectins from microfungi have been shown to be directly involved in human pathogenesis.⁶³ For example, a GlcNAc binding lectin from *Paracoccidioides brasiliensis*, known as paracoccin, binds to laminin and stimulates the release of TNF- α and nitric oxide by macrophages that leads to paracoccidioidomycosis.⁶⁴ Similarly, lectins from *A. fumigatus* and *C. glabrata* which are responsible for major hospital-acquired diseases are reported to display specific binding to human oligosaccharides.⁶⁵ FleA (AFL) located on the surface of conidia in *A. fumigatus* shows binding to fucosylated human blood group oligosaccharides. The lectin demonstrated a strong pro-inflammatory effect on human bronchial cells that could be antagonised by addition of exogenous fucose.⁶⁵ **Figure 1.5** illustrates some infections mediated by adhesins, lectins, toxins etc. at host-pathogen interface. C-type Lectin Receptors (CLRs), a class of proteins expressed on the membrane of antigen-presenting cells (APCs) such as macrophages and dendritic cells (DCs) are often used as entry receptors by viral pathogens.⁶⁶⁻⁶⁷ CLRs recognised conserved carbohydrate epitopes on diverse pathogens and initiate specific adjustments of the immune response. DC-SIGN (Dendritic Cell-Specific Intercellular adhesion molecule-3-Grabbing Non-integrin) is a major player in the recognition of pathogenic viruses thus also involved in the pathology of HIV infections.⁶⁸⁻⁷⁰ DC-SIGN mediate adhesion and internalization of virus particles that allow trafficking to non-lysosomal compartments and virus persistence in a protected intracellular environment. DC-SIGN is

attached to the cell membrane by a hydrophobic neck domain and that consist of a carbohydrate recognition domain with the primary binding site containing calcium ion as a co-factor. DC-SIGN binds to *N*-linked high mannose oligosaccharides such as Man₉GlcNAc₂.⁷¹⁻⁷³ These are abundantly present on viral envelop glycoproteins. In addition, DC-SIGN recognizes fucosylated glycans such as Lewis-type and ABO antigens.⁶⁸ Besides this, DC-SIGN is also known as a receptor for many other viruses, such as zika, dengue, Ebola and coronaviruses.⁶⁸ Therefore, carbohydrate molecules and glycomimetic drugs which have potential to interfere with the viral adhesion are interesting for the treatment of viral

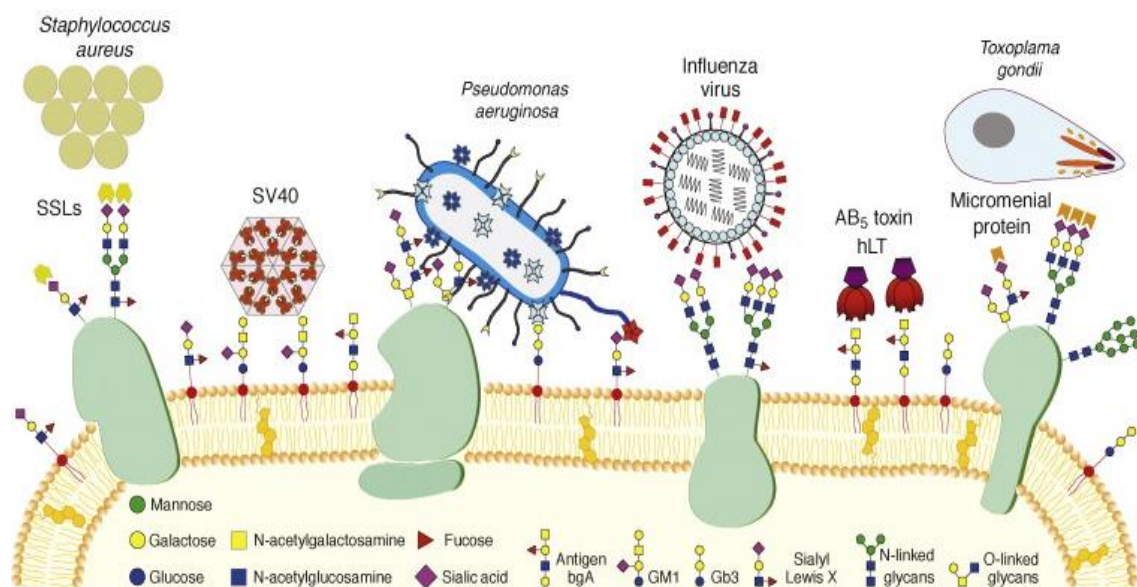


Figure 1.5 Strategies used by pathogens for host recognition and adhesion. Adapted from Imberty and co-workers (2008).⁶²

infection.

1.3.2 Glycocalyx at the surface of human cells

The glycocalyx is an extracellular compartment that covers the cell membranes in some bacteria, epithelia and other cells.³⁶ It comprises of a huge variety of different glycoconjugates and acts as anchoring platform for pathogens (**Figure 1.6**). The most important adhesion components, expressed at the host-pathogen interface by numerous

bacteria are the surface lectins (carbohydrate-binding proteins), which can bind to glycoconjugates present in the glycocalyx and act as virulence factors.⁷⁴⁻⁷⁶ The glycocalyx consists of glycoproteins, proteoglycans such as glypicans and syndecans and glycosaminoglycans (GAGs) including heparan sulfate and hyaluronic acid.⁷⁴⁻⁷⁶ The sialic acid residues are often located at the termini of glycan chains in surface glycoproteins or glycolipids and are targeted by pathogens for host cell-recognition, attachment and host specificity. For example, the human pathogen *Staphylococcus aureus* secreted superantigen-like proteins (SSLs) known as SSL5 and SSL11 shows binding to granulocytes and monocytes in a sialic acid dependent manner and promote infection by subverting the host immune system.⁷⁷ Their structures in complex with sialyl Lewis X confirmed the presence of carbohydrate-binding site located in a V-shaped surface depression on the C-terminal β -grasp domain.⁷⁸⁻⁷⁹ Likewise, sialic acid acts as attachment factor of parvoviruses. The crystal structure complex of the minute virus of mice prototype strain (MVMp) capsid in complex with sialic acid showed that the parvovirus bind to sialic acid using a binding site at the twofold axis called the dimple.⁸⁰

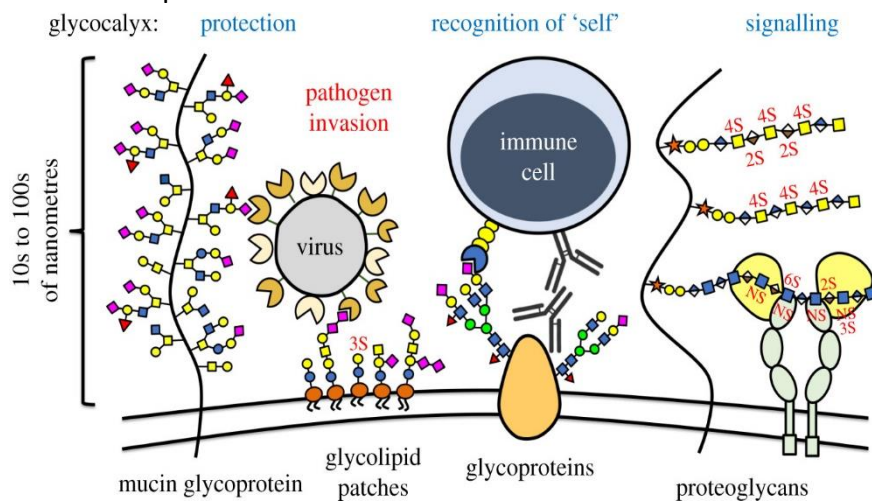


Figure 1.6 The cellular glycocalyx acts as biological interface that mediates the exchange of information between cells and their surroundings. The glycans in glycocalyx are molecular target for opportunistic pathogens. Adapted from Purcell and co-workers (2019).³⁵

In addition, A, B, and H antigens are also complex fucosylated oligosaccharides located at endothelial cells and erythrocytes of all individuals (**Figure 1.7**).⁸¹ These antigens are also expressed in saliva, tears and mucus secretions in the digestive tract of individuals.⁸² The studies in past indicates the susceptibility to diseases in certain phenotypes such as higher susceptibility of O phenotype for plague and cholera.⁸³ Since ABO structures present on

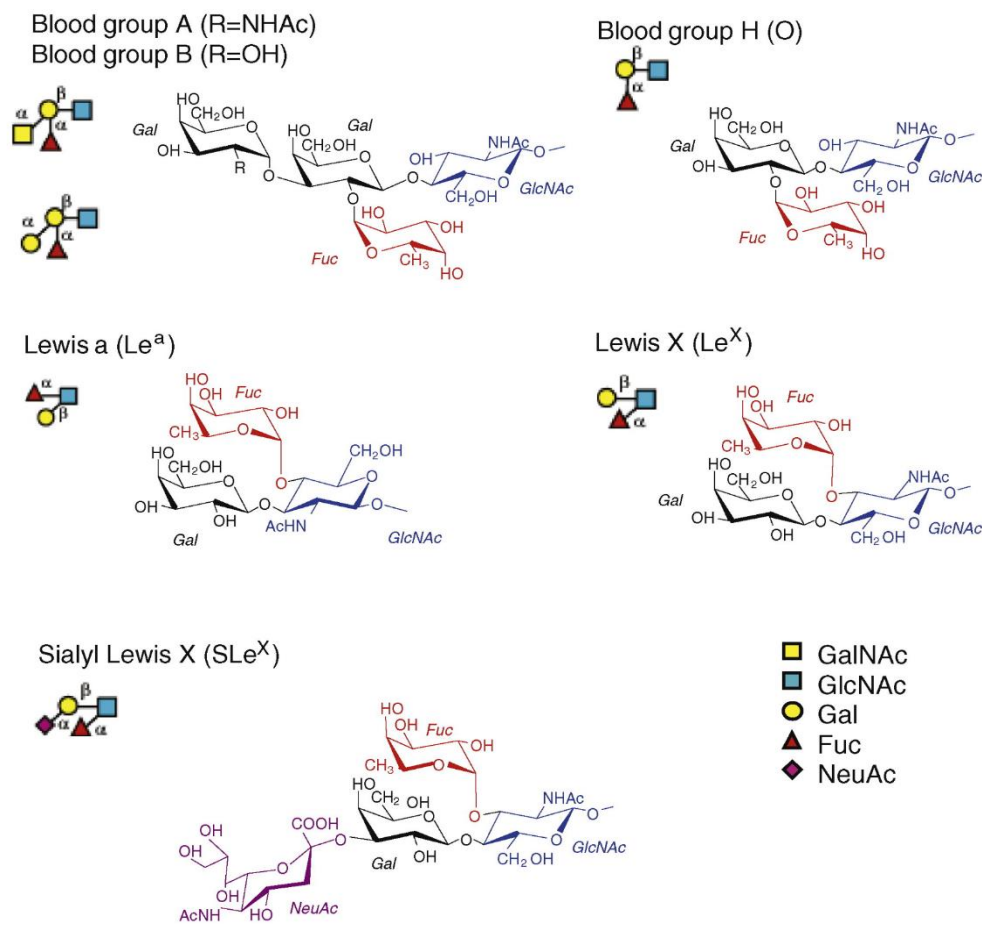


Figure 1.7 Structures of oligosaccharide epitopes of histo-blood groups with schematic representation. Fucose, sialic acid and GlcNAc are shown according to the SNFG nomenclature.⁹⁰ Adapted from Imberty and co-workers (2008).⁶²

the intestinal mucus, they play role as receptors for a wide range of pathogens.⁸⁴⁻⁸⁵ The studied on Norwalk virus (responsible for acute gastroenteritis) demonstrated the role of fucosylated epitopes (such as H-type oligosaccharides) for infection.⁸⁶ These blood-group-related epitopes are also located in lung mucus. The presence of oligosaccharides in lungs

also depends on certain conditions such as genetic disorders, chronic bronchitis, cystic fibrosis (CF) etc. In certain diseases, fucosylation is increased in epithelial glycoconjugates and also in mucus in airways that can increase susceptibility towards infection by opportunistic bacteria.⁸⁷ For example, *P. aeruginosa* produced a soluble fucose-binding lectin which acts a virulence factor and displays high affinity for fucose and fucose-containing oligosaccharides such as Lewis a trisaccharide.⁸⁸⁻⁸⁹

Apart from microbial colonization, the latest research has demonstrated the role of glycocalyx in various cellular processes (**Figure 1.6**).³⁵⁻³⁶ It acts as the additional molecular filter for endothelium and plays an important role in sensing and communicating with their environment. For instance, endothelial tissue assembly is ensured by glycocalyx-mediated communication.³⁶ Likewise, glycoconjugates assist the immune system to recognise own and foreign cells, and act accordingly. Moreover, alterations of the glycosylation are a hallmark of diseases such as cancer.⁹¹⁻⁹² Similarly, bacteria are surrounded by glycocalyx-enclosed microcolony⁹³ to protect them from harmful phagocytes by creating capsules and also help them to attach to host cells or surfaces via biofilm formation. Therefore, glycocalyx can be directly targeted in therapeutic contexts. The oligo- and polysaccharide structures that are expressed on cell surfaces are recognised as a 'glycocode'.⁹⁴ Thus, every glycan of glycocalyx can be presented as a 'message' to its environment. Hence, carbohydrates are recognised as 3rd alphabet of life in parallel to nucleobases and amino acids.⁹⁵ However, the vast information held by the glycocode is yet to be investigated.

1.3.3 Bacterial adhesins and soluble lectins involved in adhesion

On bacterial surface, adhesins and lectins mediate carbohydrate-specific binding to the cell surfaces. *E. coli* is a known pathogen to cause a range of diseases such as urinary tract infections (UTIs), enteritis and meningitis. 90 percent diagnosed cases of UTIs are reported to

be caused by Uropathogenic *E. coli* (UPEC).⁹⁶⁻⁹⁷ The infection process involves multi step cascade that has been demonstrated in mouse cystitis model and human UTIs.⁹⁸ UPEC adhesion is mediated by type 1 fimbriae (pili) that binds to mannosylated glycoprotein receptors such as uroplakin-Ia (UPIa) located on the surface of urinary bladder mucosa.⁹⁹ This event prevents clearance of UPEC by shear stress of urine flow and promotes bacterial invasion. The pilus rod forms right-handed helical structure which consists of numerous immunological-like (Ig) FimA subunits with terminal tip composed of FimF, FimG and lectin FimH (Figure 1.8A).¹⁰⁰

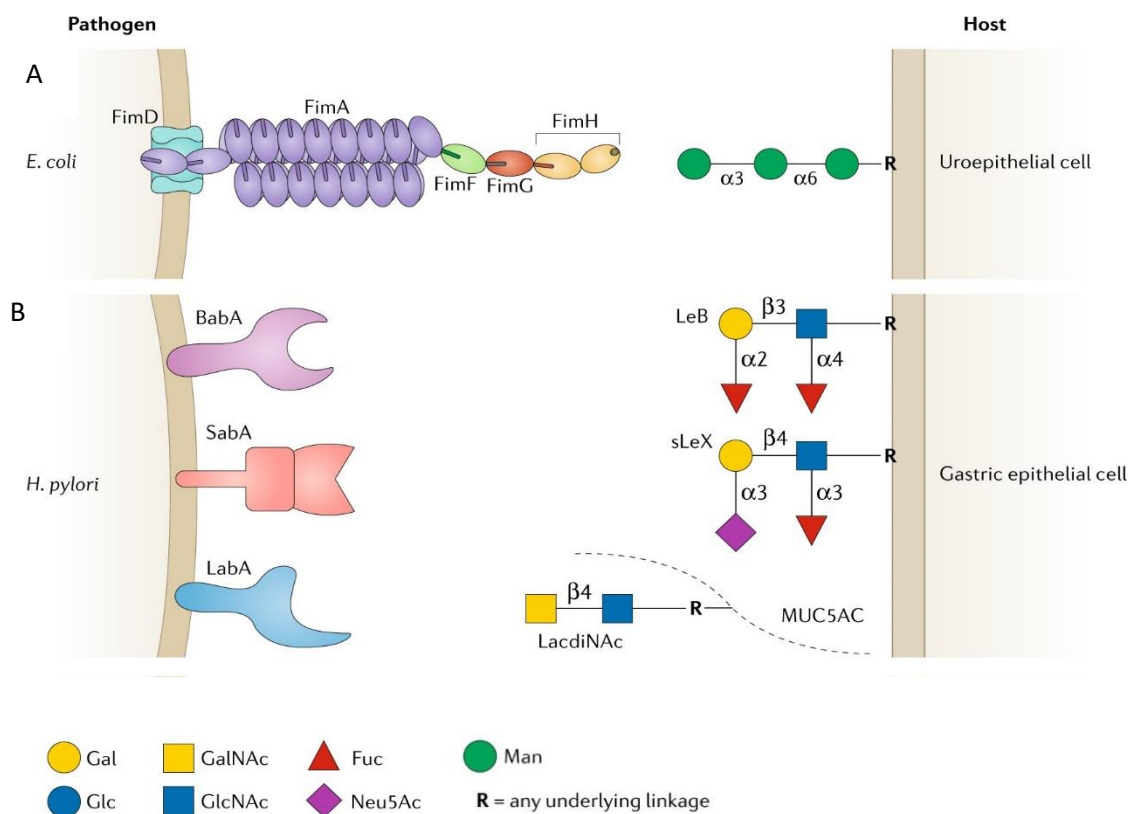


Figure 1.8 (A) Type 1 fimbrial adhesin FimH of *E. coli* binds to oligomannose present on the host uroepithelial cell during urinary tract infections. (B) Adhesins (BabA, SabA and LabA) of *Helicobacter pylori*. Adapted from Poole and co-workers (2018).¹⁰⁰

The binding preference of FimH towards oligomannose has been used as the basis for rational drug design of α -mannoside-based FimH antagonists that has been tested in mouse models for their potential as UTI therapeutics.¹⁰¹⁻¹⁰³ In addition, pathogenic *E. coli* strains use

several different types of pili to adhere to host cell surface such as fimbrial adhesin PapG¹⁰⁴, and UclD.¹⁰⁵ The binding specificities of these adhesions determine host and tissue tropism. PapG adhesin is located at the tips of the fimbriae that mediates UPEC adhesion to host cell by recognizing the Gal α 1-4Gal epitope in the globo-series of glycolipids.¹⁰⁶⁻¹⁰⁸ Similarly, a pilus adhesion known as UclD also has lectin function in *E.coli*. that interacts with O-linked glycans on host cells.¹⁰⁵

Helicobacter pylori is another Gram-negative bacterium that colonizes the human gastric mucosa and causes symptomatic infection.¹⁰⁹ In some individuals, the infection can develop into gastric cancer, thus *H. pylori* is considered a carcinogen by the WHO.¹¹⁰⁻¹¹¹ It uses lectins to bind the glycosylated human gastric mucosa. The two lectins known as -blood group antigen-binding adhesin BabA and sialic acid (neuraminic acid)-binding adhesin SabA. BabA shows specificity towards fucosylated glycoconjugates such as Lewis B¹¹² (**Figure 1.8B**), while SabA binds to Lewis blood group antigens sialyl Lewis X and sialyl Lewis A.¹¹³ The stomach lining displays increased expression of sialic acid after *H. pylori* has established a chronic infection.¹¹¹ LabA (also known as LacdiNAc-binding adhesin) is another adhesin of *H. pylori* that binds to LacdiNAc structures (GalNAc β 1-4GlcNAc) in the gastric mucosal layer.¹¹⁴ This adhesion helps in retaining *H. pylori* in the mucosal layer and thus have role in persistence.¹¹⁴

Streptococcus pneumoniae (pneumococcus) is an opportunistic pathogen responsible for infections like community acquired pneumonia, otitis media and bacteremia. The Pilus-1 proteins, RrgA, RrgB and RrgC of *S. pneumoniae* are known to play role in adhesion and infection.¹¹⁵⁻¹¹⁶ The adhesin RrgA was shown to bind to α/β linked galactose, maltose/cellobiose and blood group A and H antigens as RrgC whilst RrgB shows binding to mannose.¹¹⁷ In addition, the pathogen contains many complex surface proteins such as beta-galactosidase BgaA, which specifically binds to terminal galactose residues with beta-1-4

linkage to glucose or N-acetylglucosamine in host cell and known to play an important role in bacterial adhesion.¹¹⁸⁻¹²⁰

Pseudomonas aeruginosa is also a Gram-negative opportunistic bacterium that is involved in acute and chronic lung infections.¹²¹ This pathogen is highly resistant to many antibiotics, thus limiting the therapeutic options for the infections. *P. aeruginosa* utilizes lectins and adhesins for host-pathogen adhesion which is a critical step in initiating *P. aeruginosa* pathogenesis. The soluble lectins LecA and LecB specifically bind to galactosides and fucosides, respectively.¹²²⁻¹²⁴ Both lectins act as virulence factors and show cytotoxic effect on respiratory epithelial cells.¹²⁵ LecA interacts with α -galactosylated glycosphingolipids present in the lung epithelial cell membranes, while LecB binds to the fucosylated and mannosylated epitopes but preferentially to the Lewis^a oligosaccharides.¹²⁶ LecA and LecB are tetrameric lectins that require Ca²⁺ ion for carbohydrate binding. LecA binds to α -D-galactosides with a submicromolar binding affinity thanks to one Ca²⁺ ion while LecB shows strong micromolar affinity for L-fucose thanks to two Ca²⁺ ions. Another well-known example of bacterial soluble lectin is BambL from *Burkholderia ambifaria*.¹²⁷ BambL is a fucose-binding lectin with binding affinity in low micromolar range for monosaccharide. Like *P. aeruginosa*, this pathogen is also responsible for opportunistic infections that affect immune-compromised or cystitis fibrosis patients. Given an increasing rate of antibiotic resistance, these lectins are interesting targets to block bacterial adhesion using anti-adhesive molecules and treat drug resistant infections.¹²⁸⁻¹³⁰

The presence of surface adhesins, lectins and toxins in pathogens formed the basis of anti-adhesion approach to counteract the drug-resistant infections at the initial stage. One of the basic examples of anti-adhesive therapy that exists in Nature is related to high concentrations of human milk oligosaccharides (HMOs) in milk from breast-feeding women.

The HMOs among other function, compete with glycoconjugates in infant guts for binding to receptors on pathogens and protect the breast-fed infants.¹³¹ Very recent studies also demonstrated that SARS-CoV-2 uses its spike glycoprotein to target human lectin DC-SIGN and others.¹³²⁻¹³³ In another study, it was also shown that the recognition of spike glycoprotein by DC-SIGN can be inhibited by a glycomimetic antagonist.¹³⁴

The anti-adhesive strategies aimed at blocking the interaction between host and pathogen offer an attractive means of preventing infection at an early stage. Therefore, high-affinity anti-adhesive therapeutic agents can be designed to mimic and compete against epitopes targeted by virulence factors. In case of lectins, glycomimetic ligands can act efficiently by competitive inhibition to impede the adhesion of pathogens.

1.4 Glycans and glycomimetics as anti-adhesion compounds

It was more than four decades ago when mannose receptor in host cell were reported to mediate adhesion of *E. coli* to human mucosal cells.¹³⁵ Since then, the carbohydrate specificities of many pathogens have been determined, which provided the bases for the rational design of anti-adhesive molecules.¹³¹ While carbohydrates are exciting targets for drug design/development, native carbohydrates also suffer from a number of drawbacks when considered as therapeutic agents¹³⁶⁻¹³⁸ such as poor pharmacokinetic properties^{137,139} and poor or limited oral bioavailability. The passive permeation through intestinal enterocyte layer requires drug-like molecules with acceptable molecular mass, limited polar surface area and low numbers of H-bond donors and acceptors as per the Lipinski and Veber rules.¹⁴⁰⁻¹⁴¹ Due to high polarity of carbohydrates, they are unable to cross passively through epithelial layer in the small intestine that results into poor oral bioavailability. In addition, bulk solvent can easily displace the native carbohydrate ligands that often bind to shallow binding sites of

the lectins. The weak binding results into higher k_{off} rates of lectin interactions that leads to a short residence time. In addition, once available in bloodstream by parenteral administration, carbohydrates undergo fast renal excretion, which contributes to poor pharmacokinetic profile.

Various strategies to design glycomimetic agents have been used to address and overcome these properties of carbohydrates. Glycomimetics are compounds which mimic the structure and function of native carbohydrates, developed for their application as therapeutic candidates for carbohydrate-binding proteins such as lectins.^{137,139} A rational approach to design glycomimetic agents can enhance the binding affinities, bioavailability and also improve plasma half-lives of the molecules. To improve the binding-affinity and selectivity against a target, glycomimetics usually establish additional interactions which are not present in the native counterpart.¹³⁸ Additional strategies such as reducing ligand polarity, ligand pre-organization, optimization of entropic and enthalpic binding components can be employed to overcome the poor drug-like properties of the carbohydrates.¹³⁸ A rational approach for glycomimetic design needs sufficient information about the targets to understand the protein-ligand interactions.¹³⁷

The most sophisticated methods to study ligand-protein binding involves X-ray crystallography, nuclear magnetic resonance (NMR) experiments and molecular modelling methods.¹⁴²⁻¹⁴⁴ The structural details convey important information about the ligand binding mode and important moieties/functional groups that are available for further modifications. In addition, protein crystal structures also provide information about the amino acid residues in the vicinity of the binding site which can be explored to establish additional interactions with the ligands. In the absence of crystal structure, computational approaches such as development of homology models can be useful whenever reasonable because of high

sequence identity with proteins with known 3D structure. In addition, information related to ligand binding epitope and conformation can be obtained from saturation transfer difference (STD) NMR and transfer Nuclear Overhauser Effect (NOE) methods.¹⁴⁵ Other than direct interactions between the ligand and protein surface, particularly in the targets like lectins, it is important to consider the role of structural waters on ligand binding. Usually, the presence of structural water molecules, which are highly ordered and present in the binding site can afford favourable enthalpic gains, thus they are important for the ligand design.¹⁴⁶⁻¹⁴⁷ While the weakly bound water molecules in the binding site can have significant entropic penalties by restricting their movement in the bound state. Therefore, highly conserved water molecules are usually considered important in glycomimetic design against several carbohydrate-binding proteins, for example L-arabinose binding protein and FimH.^{146,148} By combining information from multiple approaches, appropriate strategies like bioisostere replacement, derivatization, or deoxygenation etc. can be used for the development of glycomimetics.¹³⁸ In addition, ligand design strategies related to the development of covalent inhibitors can be used for the enhancement of the binding affinity of glycomimetic ligands.¹⁴⁹ Likewise, multivalency can also enhance the binding affinity of the ligands by mimicking the multivalent presentation of native ligands.¹⁵⁰⁻¹⁵⁵ The approach has been implemented in designing multivalent inhibitors, against the targets in pathogens such as *H. pylori*,¹⁵⁶ *E.coli* and *P. aeruginosa*.¹⁵⁷⁻¹⁵⁸ Multivalency usually enhances the binding affinity and selectivity of the ligands using different mechanisms which involved chelation, statistical rebinding effects or clustering of soluble binding partners.¹⁵⁹⁻¹⁶⁰ In *chelation*, the molecule can engage two or more binding sites of a target simultaneously. In some cases, multivalency improves binding by increasing local concentration of the ligand which is called *statistical rebinding*. This mechanism reduces the off-rate of the ligands and thus increases the binding

affinity. In addition, *receptor clustering* also plays role in multivalency. In this case, the multivalent ligands lead to recruitment and aggregation or precipitation of masses of protein and also known to elicit signalling cascades.¹⁶¹⁻¹⁶² The multivalent scaffolds of the ligands should be carefully designed in order to have flexibility and spacing that allows correct fit into the binding site and also minimize the entropic costs of ligand binding. A recent study has demonstrated the role of scaffold hydrophobicity in the affinity of multivalent constructs.¹⁶³ In another excellent study on the design of multivalent glycomimetic antagonists of DC-SIGN using molecular rods, antagonists with nanomolar binding affinity have been reported.^{160,164} In this approach, a rigidified core based on phenylene-ethynylene units were designed to an ideal length (approx. 4 nm) for chelation to bridge the carbohydrate recognition domains (CRD) on the neighbouring subunits of DC-SIGN. The excellently designed construct was also able to individually probe the effects of ligand, rigid rod, and proximity etc.

The development of glycomimetics as therapeutic agents have been successful for several drug targets. These drug molecules already reached the market after successful clinical trials. For example, Oseltamivir (Tamiflu), Zanamivir (Relenza) used in the treatment of influenza infection (**Figure 1.9**) are glycomimetic inhibitors of influenza neuraminidase.¹⁶⁵⁻¹⁶⁶ Another drug named Miglustat is a glycomimetics inhibitor of glucosylceramide synthase used in the treatment of type I Gaucher disease to prevent the harmful accumulation of glucosylceramide.¹⁶⁷ The drugs used as α -glucosidase inhibitors (miglitol, voglibose, acarbose) for the treatment of diabetes and lysosomal storage disorders¹⁶⁸⁻¹⁷⁰ are also some examples of glycomimetics. Various studies on design of multivalent constructs are focused towards fucose-binding soluble receptors in pathogens, in order to have improved therapeutic effect on patients with cystic fibrosis (CF). Alternatively, these studies can also afford design and synthesis of simplified mimetics of sialyl Lewis X that can mimic its native structure.¹⁵⁴

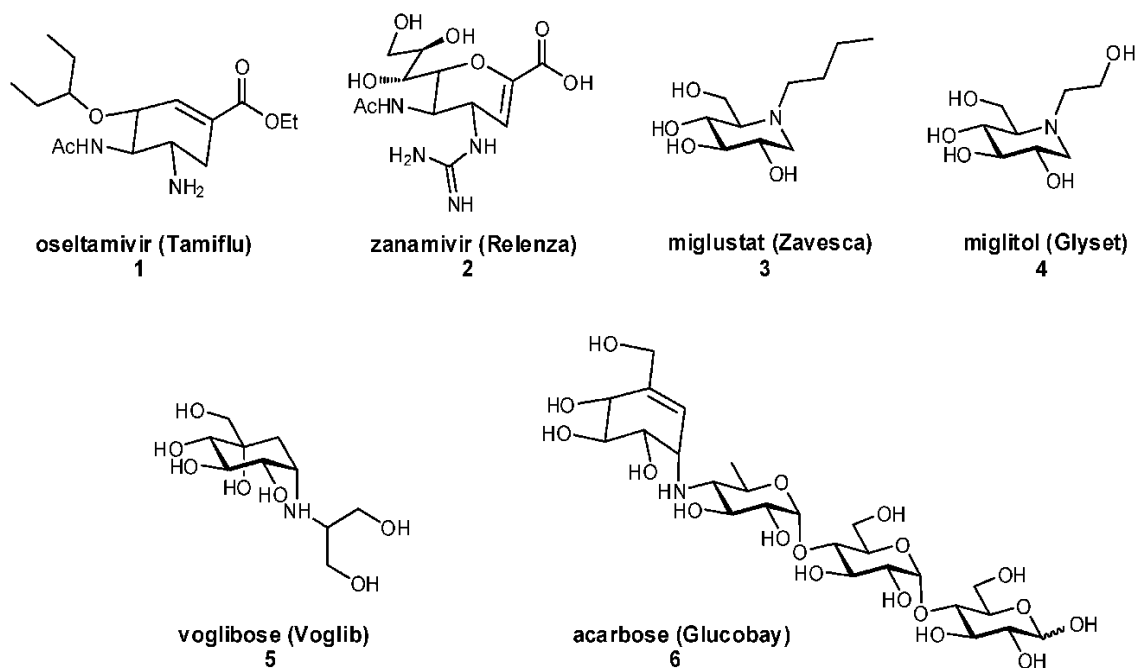


Figure 1.9. Examples of glycomimetic inhibitors that have successfully reached the market. Adapted from Hevey (2019).¹³⁸

1.4.1 FimH and LecA antagonists as anti-adhesion compounds

In 1977, the anti-adhesive activity of mannose was described in detail by Ofek, *et al.* for *E. coli*.¹³⁵ Later in vivo activity of methyl α -D-mannoside was reported in a UTI mouse model.¹⁷¹ Since FimH was identified as an important target for the development of anti-adhesive therapeutic strategies, various studies were performed to investigate the effects of FimH antagonists.¹⁷²⁻¹⁷⁴ In the following years, with an aim to improve binding affinity, various structural modifications resulted mono- and polyvalent inhibitors involving chemically modified derivatives of α -D-mannosides. For example, Bouckert *et al.* reported a series of alkyl α -D-mannosides as potent FimH antagonists (\rightarrow 1 **Figure 1.10**).¹⁷⁵ The best representative of alkyl mannosides series, *n*-heptyl α -D-mannoside, was used later as reference compound. The studies involving aromatic aglycones provided the basis for design of molecules to reach the hydrophobic tyrosine gate formed by Tyr48, Tyr137 and Ile52 (\rightarrow 2, **Figure 1.10**).¹⁷⁶ These

finding were further rationalized to design squaric acid derivatives (\rightarrow 3, **Figure 1.10**).¹⁷⁷ Later on the basis of crystallized complexes with biphenyl mannosides (showing π - π stacking with Tyr48, Tyr137), potent inhibitors were obtained (\rightarrow 4, **Figure 1.10**).¹⁷⁸ Further, substitution and extended aromatic moieties showed improvement in the binding affinity. The final ligands showed affinities in low nanomolar range however indicated poor oral bioavailability (\rightarrow 5, 6, **Figure 1.10**).¹⁷⁹⁻¹⁸⁰ The attempts to optimize oral bioavailability using prodrug

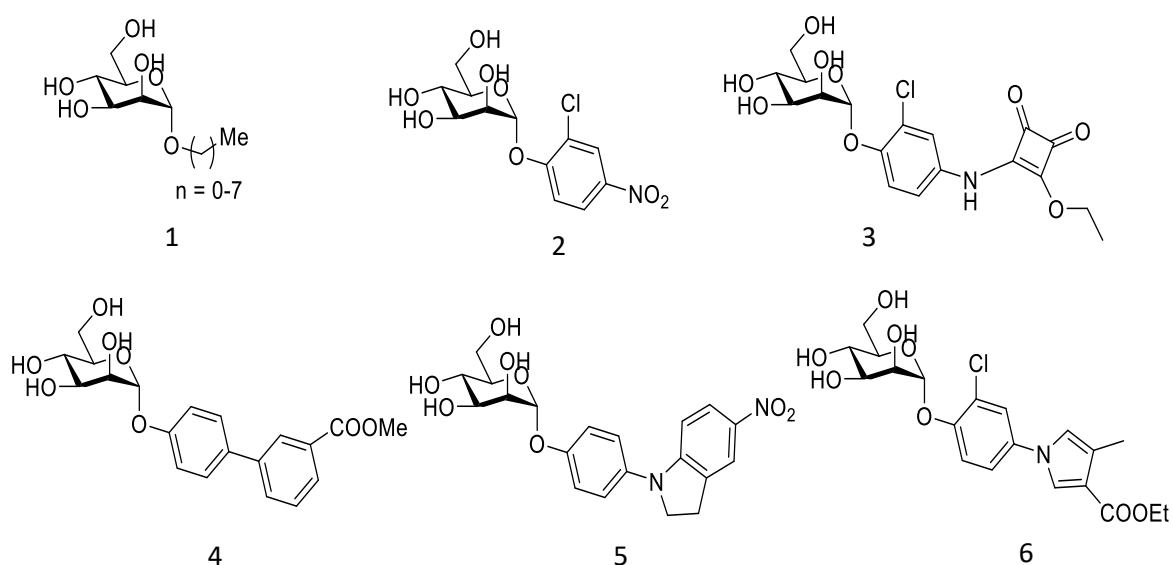


Figure 1.10 Structures of potent FimH antagonists designed using different optimizations strategies.

approach and structural modifications such as biosisosteric replacement improved the oral bioavailability in vivo. However, the final molecules with better therapeutic potential showed poor solubility. Further structural modifications improved the solubility and the final antagonist GSK3882347 (structure not disclosed) has already entered Phase 1 clinical trial.

LecA is another interesting target for the design of anti-adhesive molecules against *P. aeruginosa*. In Nature, LecA displays binding affinity towards α -linked galactosides, whereas studies using synthetic glycomimetics showed that it can bind to β -aryl galactosides (\rightarrow 1, **Figure 1.11**).¹⁸⁰⁻¹⁸² Based on rational drug-design approaches, various mono- and multivalent glycomimetics have been designed for LecA. The ligands are composed of a galactose moiety

linked to an aglycone region which establishes additional interactions in LecA CRD and enhances the binding affinity. Using the multivalency approach, presentation of aglycone part in tetravalent form resulted in a ligand with nanomolar binding affinity (\rightarrow 2, **Figure 1.11**).¹⁸³

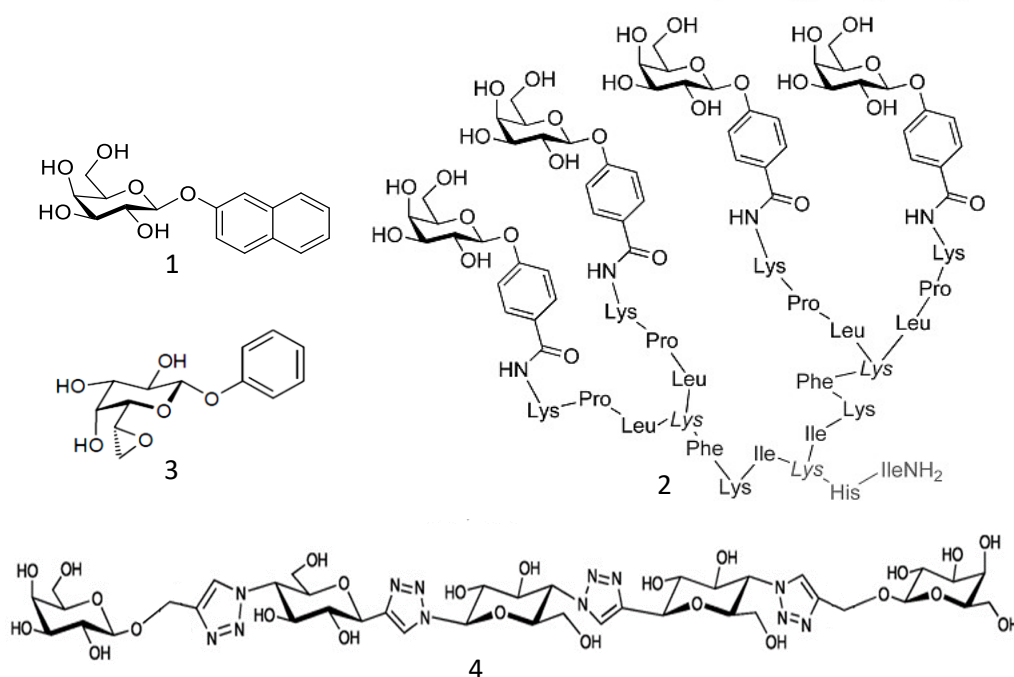


Figure 1.11 Representative structures of antagonists of LecA.

Introduction of diverse tetravalent glycopeptides also showed enhanced binding affinity. Similarly, a rationally designed divalent molecule has been demonstrated as the ligand with highest binding affinity ($K_d = 176$ nM) against LecA to date (\rightarrow 4, **Figure 1.11**). In another study, a molecule with nanomolar binding affinity, designed using glycoclusters functionalized with galactosides demonstrated almost complete protection against *P. aeruginosa* in a lung infection model in mice, thus providing a promising drug candidate.¹⁸⁴ Compared to LecB, LecA has been proven a more challenging target for the development of high-affinity ligands and was reported to have large k_{off} for LecA-ligand interactions. Therefore, a covalent inhibitor (\rightarrow 3, **Figure 1.11**) targeting a nearby cysteine (Cys62) was designed to circumvent the inherently weak affinity caused by a short lifetime of lectin-ligand complexes.¹⁴⁹

1.5 *Burkholderia cenocepacia*: An opportunistic pathogen that targets glycans

1.5.1 The *Burkholderia cepacia* complex

Opportunistic infections usually occur due to a declined in innate or adaptive immune responses. This situation allows organisms with usually low virulence, to initiate infection which can be life threatening.¹⁸⁵⁻¹⁸⁶ The most common type of infection caused by these opportunistic pathogens involves lung infection that is a major cause of morbidity and mortality for immunocompromised patients having other medical conditions like HIV, haematological malignancy, aplastic anaemia, chemotherapy treatment, cystic fibrosis (CF), recipients of solid-organ or stem cell transplants etc. The infection caused by such opportunistic pathogens depends on the type and level of immune defect in the host.

B. cenocepacia is a Gram negative bacterium which is a member of the *Burkholderia cepacia* complex (Bcc),¹⁸⁷⁻¹⁹⁰ a group of 22 bacterial species. Bcc species can be found in natural sources including water, soil and vegetation, and thus, are widely spread in the Nature. These bacteria can have both beneficial and detrimental effects on plants¹⁹¹ but they are also identified as opportunistic human pathogens. In particular, *B. cenocepacia* is responsible for deadly infections in patients with immunocompromised conditions like chronic granulomatous disease¹⁹² and cystic fibrosis (CF) resulting in severe decline in lung functions.¹⁹³ Further progression involves systemic infection known as cepacia syndrome which is characterized by an uncontrolled deterioration of lungs with septicaemia and necrotizing pneumonia that usually leads to early death.¹⁹¹ *B. cenocepacia* are highly resistant to different classes of the existing antibiotics.¹⁹¹ The pathogenicity of these bacteria is caused by several virulence determinants.¹⁹⁴⁻¹⁹⁵ In addition, they can adapt to different environmental changes inside the infected lungs resulting in more challenging treatments of infections. Inside the lungs, these pathogenic bacteria get exposed to different stressful

conditions as consequence of host immune responses, antimicrobials, fluctuation in nutrient oxygen level, low pH and the presence of co-infecting microbes.¹⁹⁶⁻¹⁹⁸ Therefore, due to long-term infection, the bacterial population undergoes transcriptional reprogramming that causes genetic and phenotypic alterations leading to heterogeneous bacterial community which is very difficult to eradicate with the existing drugs.¹⁹⁹⁻²⁰⁴ Therefore, it is highly important to identify new drug targets and apply a rational approach to design new drug molecules to effectively treat the infections caused by these drug resistant pathogens. The information from the available mutant library and genomic studies can help in the identification of essential genes responsible for virulence, adaptation and antimicrobial resistance.²⁰⁵⁻²⁰⁶ It can be very useful to identify new drug targets for the purpose of antimicrobial drug discovery against these pathogenic bacteria.

1.5.2 Lectins from *Burkholderia cenocepacia*

A similarity search based on the *lecB* sequence (from *P. aeruginosa*) identified the genes coding for related proteins in the genomes of *Burkholderia* species belonging to the Bcc and also reported in CF isolates.²⁰⁷ The analysis of *B. cenocepacia* strain J2315 identified three genes coding for LecB-like proteins on chromosome 2 with sizes of 384, 732 and 816 base pairs (bp).²⁰⁷ A fourth one (877 bp) has been identified on chromosome 3 of *B. cepacia* R18194-1TCC17660. These four putative lectins are referred to as BC2L-A (128 aa), BC2L-B (244 aa), BC2L-C (272 aa), and BC2L-D (289 aa).²⁰⁷⁻²⁰⁸ All four polypeptides display a LecB-like domain at the C-terminal domain (**Figure 1.12**). An additional domain (120–160 aa) at the N-terminus was also identified in three polypeptides (except BC2L-A). These three N-terminal domains were different from each other and do not show similarity to any other protein as

well. The N-terminal domain of BC2L-C has been characterized as a novel fucose binding domain with a TNF- α -like fold.²⁰⁹

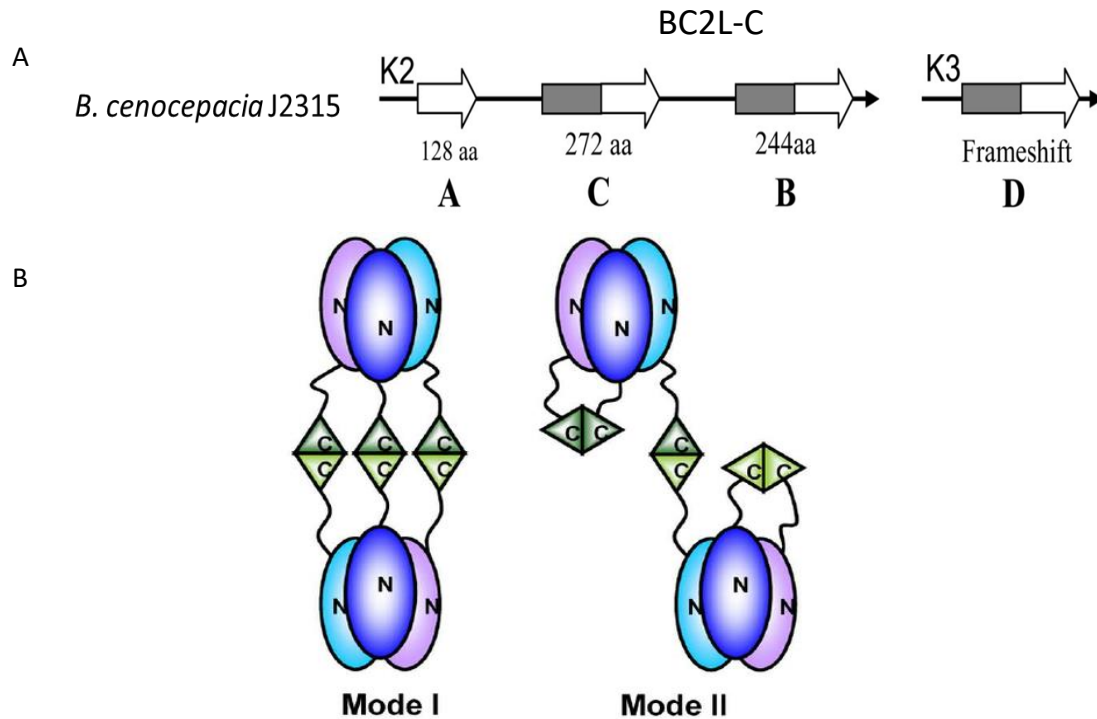


Figure 1.12 (A) Schematic representation of LecB-like proteins with number of amino acid (aa) in *B. cenocepacia*. The arrows with shaded part indicate the predicted lectin domains. Adapted from Lameignere and co-workers (2008).²⁰⁷ (B) Schematic representation of the BC2L-C hexamer with N-terminal domains (blue) and C-terminal domains (green) in two models. Both the domains are connected through a peptide linker. Adapted from Šulák and co-workers (2011).²⁰⁸

The initial studies of these *lecB*-like lectin family was done for BC2L-A.²⁰⁷ The *lecB*-like gene, *bclA* was identified on the six *Burkholderia* strains, showed 32% similarity with *lecB*.²⁰⁷ BC2L-A have successfully purified in native form from *B. cenocepacia* strain J2315, and later cloned in *E. coli* to produce recombinant form. The experimental studies also revealed that BC2L-A specifically binds to mannose.²⁰⁷ Similar to LecB lectin from *P. aeruginosa*.¹²⁴ it consists of one carbohydrate recognition domain with two Ca²⁺ ions directly involved in ligand

binding. The difference in their specificities has been reported due to the presence of different residues in a loop featuring two serine residues (22 & 23) in LecB, and two alanine residues (29 & 30) in BC2L-A.²¹⁰ Moreover, structural comparison of both the lectins shows different features (**Figure 1.13**). The lectin LecB from *P. aeruginosa* forms homotetramers while the *Burkholderia* lectins BC2L-A and the C-terminal domain of BC2L-C are known to form homodimer^{207,209} (**Figure 1.13, A, B, C**).

In order to investigate the location of these lectins, studies have been already performed which indicate that they are present in the extracellular medium, although the secretion system is not identified.²⁰⁹ In particular, these studies demonstrated that BC2L-B and -C lectins are released from the bacterial cells upon mannose treatment. These results provide the important clue that they are located at the external envelope of the bacteria. However, BC2L-A was not detected on the surface probably due to its much lower expression

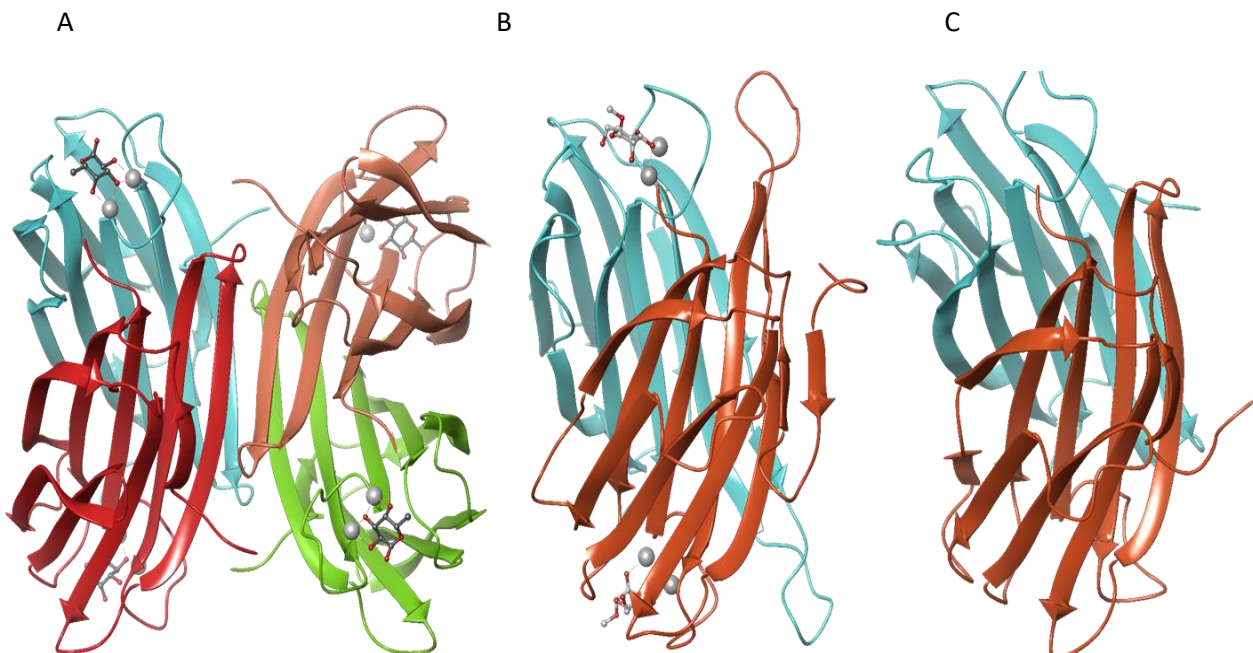


Figure 1.13 Structural similarities between LecB, BC2L-A and C-terminal domain of BC2L-C. (A) LecB from *P. aeruginosa* (PDB code 1GZT); (B) BC2L-A (PDB code 2VNV) lectin from *B. cenocepacia* and (C) the apo form of the C-terminal domain of BC2L-C (PDB code 2XR4) from *B. cenocepacia*. LecB forms a tetramer while BC2L-A and the C terminal domain of BC2L-C are arranged as dimers.

in *B. cenocepacia*.²⁰⁷ Nonetheless, *B. cenocepacia* cells incubated with fluorescein 5-isothiocyanate (FITC) labelled BC2L-A, displayed its accumulation at the surface of *B. cenocepacia* indicating its presence in the biofilm and possible role in cell adhesion. In addition, it has been reported that BC2L-A and BC2L-C bind to the epitopes²⁰⁹ which are abundantly found as a component of the *B. cenocepacia* lipopolysaccharide (LPS)²¹¹ or other Gram negative bacteria.²¹² The structural and functional details of BC2L-A was further used to probe the mannosides and glycomimetics for the successful inhibition.²¹³⁻²¹⁵ Moreover, BC2L-A has been used as a model to perform optimization of multivalent glycomimetic design.²¹⁶

These studies have been extended to understand the role of BC2L-B and -C in the virulence of *B. cenocepacia*. The transcriptomic analysis along the chronic infection suggests that of genes for BC2L-B and -C were up regulated while the corresponding gene for BC2L-A was down regulated. These results indicate the possibility of secondary roles of the additional N-termini in BC2L-B and -C.²¹⁷⁻²¹⁸ The experimental studies demonstrated that operon *bclACB* coding for the three lectins (BC2L-A, -B, -C) is regulated by quorum sensing which ultimately regulates the biofilm formation and plays an important role in maintaining the structure of biofilm.²¹⁷⁻²¹⁸ Further, gene knock-out strategies confirmed that lack of any one of the lectins led to defective biofilm formation.²¹⁹ The results from these studies indicate that these lectins could be interesting targets for drug design.

To further decipher the strategies used by *B. cenocepacia* to recognize glycans, the investigation of the N-terminal domain of BC2L-C was also important. These studies revealed a *superlectin* with dual specificity and proinflammatory activity.²⁰⁸ BC2L-C consists of lectin domains with 272 amino acids long peptide sequence. The C-terminal domain consists of 116 residues that shows 43% sequence identity with LecB and uses two-calcium ions for sugar

binding, hence considered as “Lec-B like” lectin. In addition, the presence of Ala-Ala-Asn sequence in the “specificity loop” suggests its specificity for mannosides.²¹⁰ The N-terminal domain of BC2L-C (BC2L-C-nt) consists of 130 amino acids which are separated from the LecB-like C-terminal domain by a 26 amino acid linker rich in glycine and serine residues. This domain displays sequence identity up to 92 percent with other species of *Burkholderia* and only one hypothetical protein from another bacterium *Photorhabdus luminescens*, displayed 59 % sequence identity.²²⁰ In order to study its structure and function, Šulák and co-workers expressed the recombinant form of BC2L-C-nt in *E.coli*. The initial design consisted of the sequence coding for the 156 amino acids of BC2L-C-nt (including linker) and a C-terminal histidine (HisTag) to facilitate purification, which resulted a 187 amino acid long polypeptide (19.26 kDa). The size exclusion chromatography eluted a 58 kDa protein which indicated a homotrimeric assembly. Further characterization of the protein revealed its specific millimolar affinity towards L-fucose by surface plasmon resonance (SPR). The construct was further probed against a glycan array, indicating the preference for fucosylated histo-blood group epitopes. In addition, isothermal titration calorimetry (ITC) showed human oligosaccharides binding affinities in micromolar range (**Table 1.2**).

Table 1.2 Binding affinity of monosaccharides and oligosaccharides for the two domains of BC2L-C, measured by ITC. Adapted from Šulák and co-workers (2010 and 2011).²⁰⁸⁻²⁰⁹

Ligand	Terminal epitope	Affinity (K _d) in μM	Reference
BC2L-C-ct			Šulák and co-workers (2011)
D-Man	Man	37.4	
αMeMan	Man	27.6	
Trimannose	Manα1-3(Manα1-6)Man	28.8	
αMeHept	L, D-manHep	236	
Diheptose	L,D-manHepα1-3L,D-manHep.	88.1	
BC2L-C-nt			
αMe-L-Fuc	Fuc	2700	

H-type 2	Fuca1-2Galβ1-4GlcNAc	1236	Šulák and co-workers (2010)
Lewis b	Fuca1-2Galβ1-3[αFuc1-4]GlcNAc	213	
Lewis x	Galβ1-4[αFuc1-3]GlcNAc	196	
Lewis a	Galβ1-3[αFuc1-4]GlcNAc	132.1	
H-type 1	Fuca1-2Galβ1-3GlcNAc	77.2	
Lewis y	Fuca1-2Galβ1-4[αFuc1-3]GlcNAc	53.9	
BC2L-C			Šulák and co-workers (2011)
D-Mannose	Man	28.8	
αMeMan	Man	18.3	
Lewis y	Fuca1-2Galβ1-4[αFuc1-3]GlcNAc	47.5	

BC2L-C is therefore a *superlectin* that binds independently to mannose/heptose glycoconjugates and fucosylated human histo-blood group epitopes. In the same study, the first crystal structure of BC2L-C-nt (PDB ID: 2WQ4) complexed with methylseleno- α -L-fucopyranoside (MeSe- α -L-Fuc) has been reported at 1.42 Å which confirmed the trimeric arrangement of the lectin (**Figure 1.14**). The structure of BC2L-C-nt domain revealed a compact jellyroll architecture composed of 11 β strands and a short helix. The β strands show Greek-key topology usually reported for human tumor necrosis factor (TNF)-like proteins,²²¹ despite no sequence identity. In the trimeric structure, three identical binding sites could be identified at each interface between two adjacent monomers (**Figure 1.14**). The key residues Tyr48, Ser82, Thr83, Arg85 from one chain and Tyr58, Thr74, Tyr75, Arg111 from the neighbouring protomer play an important role in the ligand binding. In addition, water mediated interactions to bridge the sugar and the protein have been reported. Other than MeSe- α -L-Fuc, crystallized complexes with larger ligands were not available.

In another study from Šulák and co-workers attempted to uncover the whole architecture of superlectin, recombinant C-terminal domain (BC2L-C-ct) and full protein were expressed and probed for the specificities towards sugars. The protein showed structural similarities with BC2L-A, and also displayed affinity for different mannosides and manno-

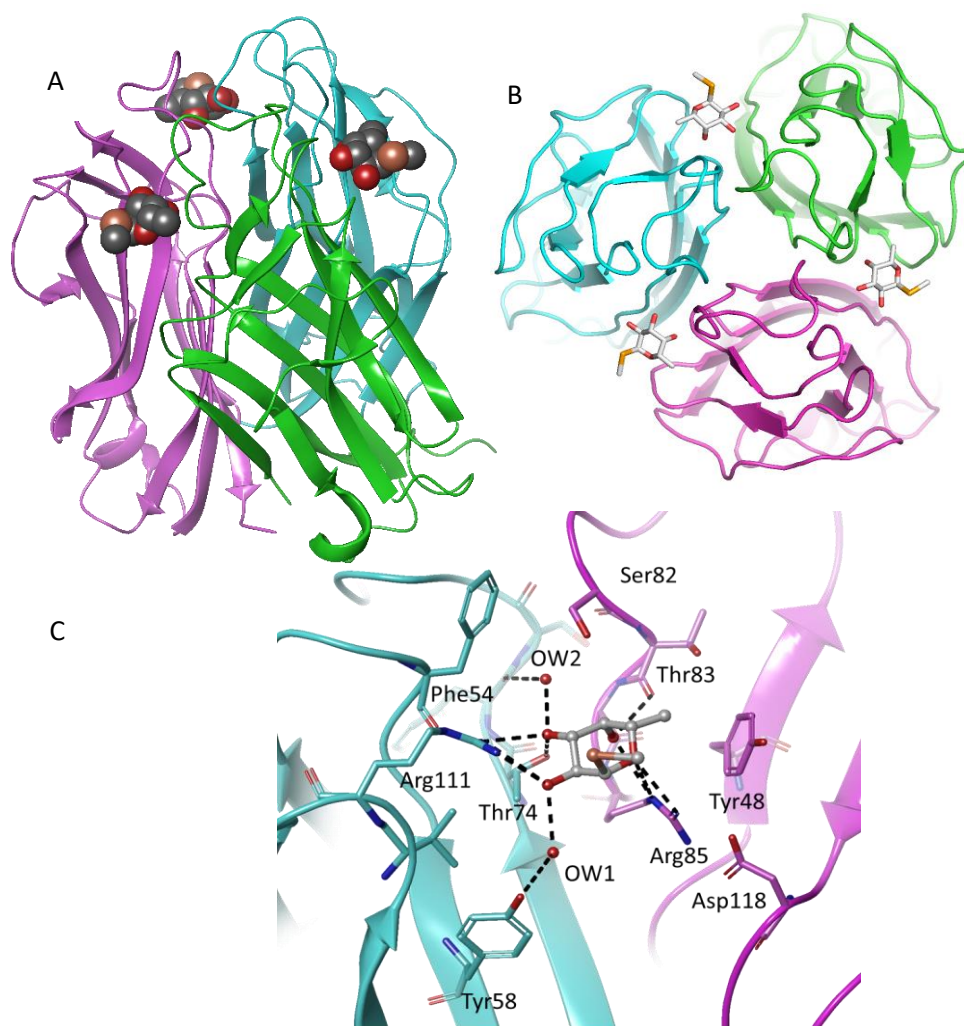


Figure 1.14 Trimeric arrangement and fucoside binding site of the N-terminal domain of BC2L-C (PDB code 2WQ4). (A) Side and (B) top view of the crystal structure (homotrimer) of the N-terminal domain of BC2L-C. (C) The H-bond interactions with the ligand (methylseleno- α -L-fucopyranoside) and water molecules (depicted as red spheres) are shown as black lines. Adapted from Šulák and co-workers (2010).²⁰⁹

configured heptose ligands (**Table 1.2**).²⁰⁸ Thus, experimental studies showed that the whole lectin also binds to both types of oligosaccharides. In addition, a crystal structure of the recombinant BC2L-C-ct dimeric lectin domain in the apo form was obtained (PDB code 2XR4, **Figure 1.13 C**). In order to understand the binding of mannosides, the crystal structure was further used for computational docking to generate the complex.²⁰⁸ Structural elucidation by small angle X-ray scattering (SAXS) and electron microscopy (EM) revealed a flexible hexameric arrangement of the lectin, which accounted for the (three) dimeric C-terminal and

(two) trimeric N-terminal domains (**Figure 1.15**). This unique hexameric architecture appears suitable for cross-linking between bacteria and epithelial cells. The flexible linker connecting two domains may also assist in adapting a suitable conformation under shear force and allow tight binding as already reported for some adhesins.²²² As already discussed, BC2L-C is reported to be released into extracellular matrix and play role in bacterial adhesion.^{208,217,223} Further studies based on gene knock-out strategies confirmed that the lack of any one of the lectins (BC2L-A, -B, -C) led to defective biofilm formation.²¹⁷ Thus, BC2L-C could be an interesting target for anti-adhesive therapy.

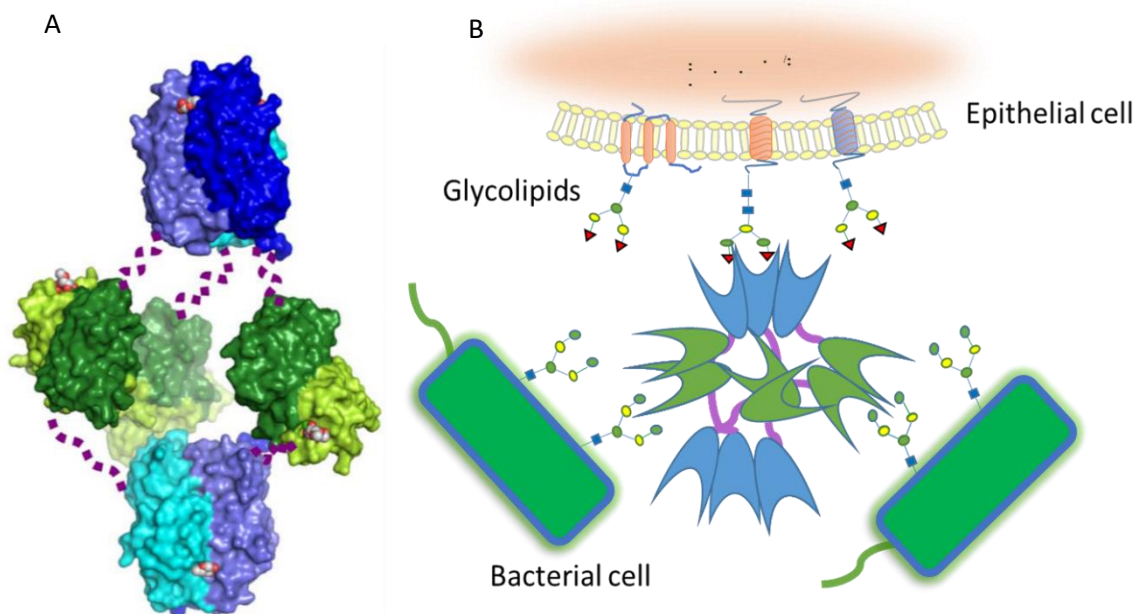


Figure 1.15 (A) Model of the BC2L-C hexamer obtained by best manual fit of the different domains in the *ab-initio* SAXS envelop.¹⁹⁹ The N- and C-terminal domains are shown in blue and green, respectively. (B) Schematic representation of BC2L-C hexamer showing its role in crosslinking bacteria surface and host epithelial cells.

1.6 Thesis objective

B. cenocepacia is known to cause life-threatening systemic infection which is extremely difficult to treat due to antibiotic resistance. The experimental studies have confirmed the presence of lectins (e.g. BC2L-C) as virulence factors responsible for adhesion

of pathogen to host cell. These studies opened the route for the design of anti-adhesion therapy which aims to disrupt the virulence mechanisms of the bacterium. In particular, the superlectin BC2L-C is an interesting target, which shows similarities with the known virulence factors through C-terminus and also composed of an additional N-terminal domain with unique structural features. Since the N-terminal domain presents a novel lectin domain and already reported with a high resolution crystal structure, it provides an opportunity to explore this domain for structure based drug-design of small-molecule antagonists to block lectin-mediated adhesion.

There are following objectives of the project:

- Virtual screening of fragment library at identified sites.
- Experimental validation of fragment binding.
- Identification of strategies to connect selected fragments to the sugar core to design glycomimetic ligands.
- Molecular dynamics simulation studies of the designed ligands.
- Biophysical evaluation of affinity and identification of binding mode for fragments and designed ligands (in collaboration with Rafael Bermeo at CERMAV-CNRS, Grenoble Alps University and the University of Milan).
- Structure-based elaboration of selected compounds into high affinity ligands.

The initial studies in the project involve computational analysis of the target to identify druggable regions that could host drug like molecules. In addition, the project involves validation of the interaction/binding of small fragments at the identified site using biophysical assays. The identification of these new sites and small fragments will enable the rational design of glycomimetic ligands. The second part of the project is based on the computational design and modelling of the glycomimetic ligands which involves studies of the interactions of the ligands with the target to prioritize the best ligands for synthesis in collaboration with

Rafael Bermeo. The additional studies are focused towards the structure-based strategies to improve the binding affinity of glycomimetic ligands.

1.7 The PhD4GlycoDrug consortium

Objectives of this thesis have been completed under the framework of the PhD4GlycoDrug- Marie Skłodowska-Curie Innovative Training Network (MSCA ITN); a European project, funded by the European Union's Horizon 2020 research and innovation programme.²²⁴ The PhD4GlycoDrug consortium (**Figure 1.16**) involves academic and non-academic partners including a research institute to ensure quality training in glyco-drug discovery and development. The project aims to pursue the research on the development of carbohydrate-based therapeutic molecules (glycodrugs) for the different targets. Hence, using the methods under structural biology, molecular modelling, organic synthesis and

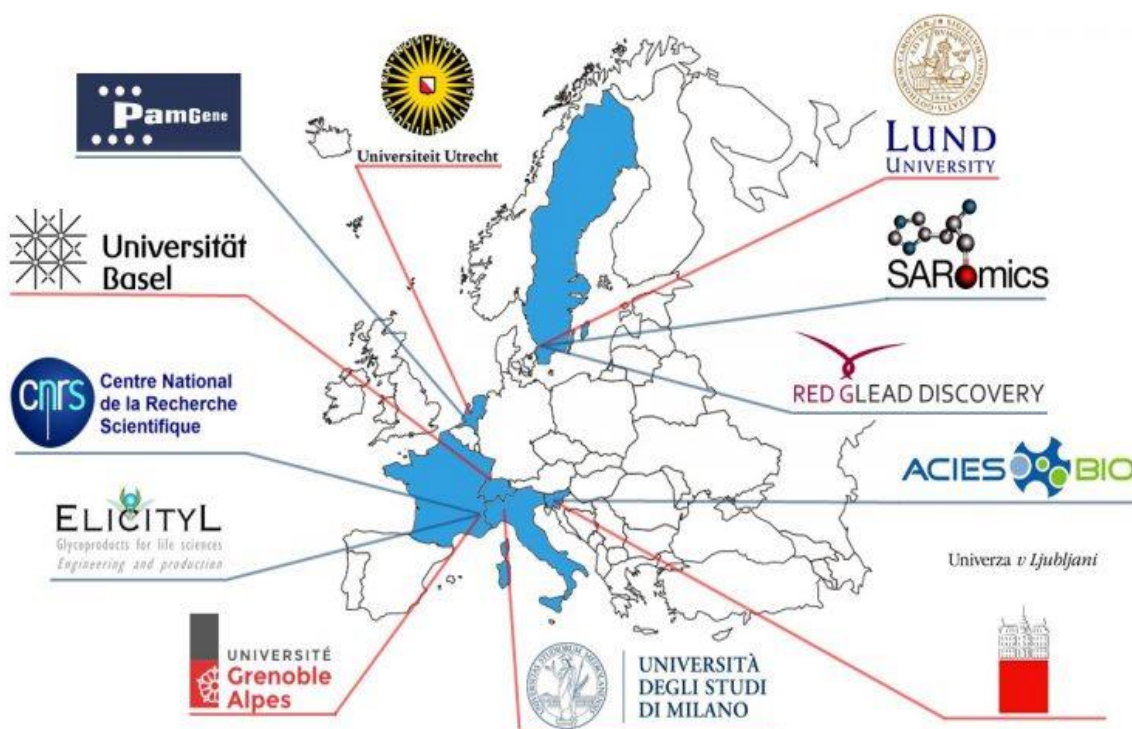


Figure 1.16 The PhD4GlycoDrug consortium.

medicinal chemistry, pathogenic lectins have been characterized and targeted for antagonist design. In addition, the research work under the consortium was performed in collaboration

between the universities that allowed the PhD students to be enrolled in two universities and combine their unique domains of expertise.

Particularly, the objectives of this project targeting the BC2L-C lectin were accomplished at the University of Milan and the Grenoble Alps University. The initial studies based on the design of antagonists using molecular modelling involve expertise in the group of Prof. Bernardi and Prof. Belvisi at the University of Milan. The later studies focused on the computational and biophysical evaluation of the designed ligands were performed at the Structural and Molecular Glycobiology (GMBS) group at CERMAV-CNRS (Grenoble Alps University) under the supervision of Dr. Imberty and Dr. Varrot. Finally, the designed molecules have been synthesized in collaboration with another PhD student (Rafael Bermeo) in the PhD4GlycoDrug network. The contribution from the expertise in different fields helped to achieve the initial objectives of the study. In addition, a short secondment for the dissemination of scientific information (related to glycoscience) to the scientific community and the general public was completed in collaboration with the Glycopedia²²⁵ platform under the supervision of Dr. Serge Perez. This involves production of a review article attached in the **Chapter 6: Scientific communication: Short secondment at Glycopedia.**²²⁶

1.8 References

[1] *The Review on Antimicrobial Resistance, chaired by Jim O'Neill. Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations, 2014.*

[2] *Antimicrobial Resistance: Global Report on Surveillance, Geneva: WHO 2014, 1-252*

[3] Rammelkamp, C. H. Maxon, T., Resistance of *Staphylococcus aureus* to the Action of Penicillin, *Proc. Soc. Exp. Biol. Med.* **1942**, *51*, 386-389.

[4] *CDC. Antibiotic Resistance Threats in the United States, U.S. Department of Health and Human Services, CDC, Atlanta, GA, 2019.*

[5] *2020 Antibacterial agents in clinical and preclinical development: an overview and analysis, WHO, Geneva, 2021.*

- [6] Lomovskaya, O. *et al.*, Vaborbactam: Spectrum of Beta-Lactamase Inhibition and Impact of Resistance Mechanisms on Activity in Enterobacteriaceae, *Antimicrob. Agents Chemother.* **2017**, *61*, 01443-01417.
- [7] Veve, M. P.Wagner, J. L., Lefamulin: Review of a Promising Novel Pleuromutilin Antibiotic, *Pharmacotherapy* **2018**, *38*, 935-946.
- [8] *FDA approves new antibiotic to treat community-acquired bacterial pneumonia*, Food and Drug Administration (FDA) (Press release), **19 August 2019**.
- [9] *FDA approves new antibacterial drug*, Food and Drug Administration (FDA) (Press release), **29 August 2017**.
- [10] Hawkey, P. M., Mechanisms of quinolone action and microbial response, *J. Antimicrob. Chemother.* **2003**, *1*, 29-35.
- [11] Ruiz, J., Mechanisms of resistance to quinolones: target alterations, decreased accumulation and DNA gyrase protection, *J. Antimicrob. Chemother.* **2003**, *51*, 1109-1117.
- [12] Ford, C. W. *et al.*, Oxazolidinones: New antibacterial agents, *Trends Microbiol.* **1997**, *5*, 196-200.
- [13] Wolter, N. *et al.*, Novel Mechanism of Resistance to Oxazolidinones, Macrolides, and Chloramphenicol in Ribosomal Protein L4 of the Pneumococcus, *Antimicrob. Agents Chemother.* **2005**, *49*, 3554-3557.
- [14] Appelgren, P. *et al.*, Risk factors for nosocomial intensive care infection: a long-term prospective analysis, *Acta Anaesthesiol. Scand.* **2001**, *45*, 710-719.
- [15] Ozer, B. *et al.*, Nosocomial infections and risk factors in intensive care unit of a university hospital in Turkey, *Cent. Eur. J. Med.* **2010**, *5*, 203-208.
- [16] van Duin, D.Paterson, D. L., Multidrug-Resistant Bacteria in the Community: Trends and Lessons Learned, *Infect. Dis. Clin. North Am.* **2016**, *30*, 377-390.
- [17] Davies, J.Davies, D., Origins and evolution of antibiotic resistance, *Microbiology and molecular biology reviews : MMBR* **2010**, *74*, 417-433.
- [18] Krachler, A. M.Orth, K., Targeting the bacteria-host interface: strategies in anti-adhesion therapy, *Virulence* **2013**, *4*, 284-294.
- [19] Anderson, B. N. *et al.*, Weak Rolling Adhesion Enhances Bacterial Surface Colonization, *J. Bacteriol.* **2007**, *189*, 1794-1802.
- [20] Vasudevan, R., Biofilms: Microbial Cities of Scientific Significance, *J. Microbiol. Exp.* **2014**, *1*.
- [21] Sherman, P. M.Boedeker, E. C., Pilus-mediated interactions of the Escherichia coli strain RDEC-1 with mucosal glycoproteins in the small intestine of rabbits, *Gastroenterology* **1987**, *93*, 734-743.
- [22] Pak, J. *et al.*, Tamm-Horsfall protein binds to type 1 fimbriated Escherichia coli and prevents E. coli from binding to uroplakin Ia and Ib receptors, *J. Biol. Chem.* **2001**, *276*, 9924-9930.

- [23] Piotrowski, J. *et al.*, Inhibition of *Helicobacter pylori* colonization by sulfated gastric mucin, *Biochem. Int.* **1991**, *24*, 749-756.
- [24] Evans, L. V., *Biofilms: Recent Advances in their Study and Control*, CRC Press, **2000**.
- [25] Hayes, C. S. *et al.*, Bacterial contact-dependent delivery systems, *Annu. Rev. Genet.* **2010**, *44*, 71-90.
- [26] Winnen, B. *et al.*, Hierarchical effector protein transport by the *Salmonella Typhimurium* SPI-1 type III secretion system, *PLoS One* **2008**, *3*, 0002178.
- [27] Schlumberger, M. C. *et al.*, Real-time imaging of type III secretion: *Salmonella* SipA injection into host cells, *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 12548-12553.
- [28] Krachler, A. M. *et al.*, Outer membrane adhesion factor multivalent adhesion molecule 7 initiates host cell binding during infection by gram-negative pathogens, *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 11614-11619.
- [29] Ofek, I. *et al.*, Anti-adhesion therapy of bacterial diseases: prospects and problems, *FEMS Immunol. Med. Microbiol.* **2003**, *38*, 181-191.
- [30] Bernbom, N. *et al.*, Bacterial adhesion to stainless steel is reduced by aqueous fish extract coatings, *Biofilms* **2006**, *3*, 25-36.
- [31] Okuda, K. *et al.*, Inhibition of *Streptococcus mutans* adherence and biofilm formation using analogues of the SspB peptide, *Arch. Oral Biol.* **2010**, *55*, 754-762.
- [32] Moon, H. W. Bunn, T. O., Vaccines for preventing enterotoxigenic *Escherichia coli* infections in farm animals, *Vaccine* **1993**, *11*, 213-200.
- [33] Signoretto, C. *et al.*, Functional foods and strategies contrasting bacterial adhesion, *Curr. Opin. Biotechnol.* **2012**, *23*, 160-167.
- [34] Viela, F. *et al.*, What makes bacterial pathogens so sticky?, *Mol. Microbiol.* **2020**, *113*, 683-690.
- [35] Purcell, S. C. Godula, K., Synthetic glycoscapes: addressing the structural and functional complexity of the glycocalyx, *Interface Focus* **2019**, *9*, 20180080.
- [36] Reitsma, S. *et al.*, The endothelial glycocalyx: composition, functions, and visualization, *Pflugers Archiv : Eur. J. Appl. Physiol.* **2007**, *454*, 345-359.
- [37] Van Breedam, W. *et al.*, Bitter-sweet symphony: glycan–lectin interactions in virus biology, *FEMS Microbiol. Rev.* **2014**, *38*, 598-632.
- [38] Ströh, L. J. Stehle, T., Glycan Engagement by Viruses: Receptor Switches and Specificity, *Annu. Rev. Virol.* **2014**, *1*, 285-306.
- [39] Smith, D. F. Cummings, R. D., Investigating virus-glycan interactions using glycan microarrays, *Curr. Opin. Virol.* **2014**, *7*, 79-87.

- [40] Stambach, N. S. Taylor, M. E., Characterization of carbohydrate recognition by langerin, a C-type lectin of Langerhans cells, *Glycobiology* **2003**, *13*, 401-410.
- [41] Lowe, J. B. Marth, J. D., A Genetic Approach to Mammalian Glycan Function, *Annu. Rev. Biochem.* **2003**, *72*, 643-691.
- [42] Raman, R. *et al.*, Glycan receptor specificity as a useful tool for characterization and surveillance of influenza A virus, *Trends Microbiol.* **2014**, *22*, 632-641.
- [43] Sriwilaijaroen, N. Suzuki, Y., Molecular basis of the structure and function of H1 hemagglutinin of influenza virus, *Proc. Jpn. Acad., Ser. B, Phys. Biol. Sci.* **2012**, *88*, 226-249.
- [44] Chahales, P. Thanassi, D. G., Structure, Function, and Assembly of Adhesive Organelles by Uropathogenic Bacteria, *Microbiol. Spectr.* **2015**, *3*, 3.5.11.
- [45] Sarshar, M. *et al.*, FimH and Anti-Adhesive Therapeutics: A Disarming Strategy Against Uropathogens, *Antibiotics (Basel, Switzerland)* **2020**, *9*, 397.
- [46] Sauer, M. M. *et al.*, Catch-bond mechanism of the bacterial adhesin FimH, *Nat. Commun.* **2016**, *7*, 10738.
- [47] Imberty, A. *et al.*, Structural basis of high-affinity glycan recognition by bacterial and fungal lectins, *Curr. Opin. Struct. Biol.* **2005**, *15*, 525-534.
- [48] Merritt, E. A. Hol, W. G., AB5 toxins, *Curr. Opin. Struct. Biol.* **1995**, *5*, 165-171.
- [49] Sharon, N. Lis, H., History of lectins: from hemagglutinins to biological recognition molecules, *Glycobiology* **2004**, *14*, 30.
- [50] Boyd, W. C. Shapleigh, E., Specific Precipitating Activity of Plant Agglutinins (Lectins), *Science* **1954**, *119*, 419.
- [51] Ambrosi, M. *et al.*, Lectins: tools for the molecular understanding of the glycode, *Org. Biomol. Chem.* **2005**, *3*, 1593-1608.
- [52] Drickamer, K., Ca²⁺-dependent carbohydrate-recognition domains in animal proteins, *Curr. Opin. Struct. Biol.* **1993**, *3*, 393-400.
- [53] Harvey, D. J. *et al.*, Structural and quantitative analysis of N-linked glycans by matrix-assisted laser desorption ionization and negative ion nanospray mass spectrometry, *Anal. Biochem.* **2008**, *376*, 44-60.
- [54] Bonnardel, F. *et al.*, LectomeXplore, an update of UniLectin for the discovery of carbohydrate-binding proteins based on a new lectin classification, *Nucleic Acids Res.* **2021**, *49*, D1548-D1554.
- [55] André, S. *et al.*, Lectins: getting familiar with translators of the sugar code, *Molecules* **2015**, *20*, 1788-1823.
- [56] Bonnardel, F. *et al.*, LectomeXplore, an update of UniLectin for the discovery of carbohydrate-binding proteins based on a new lectin classification, *Nucleic Acids Res.* **2020**, *49*, D1548-D1554.

- [57] Bonnardel, F. *et al.*, UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands, *Nucleic Acids Res.* **2018**, *47*, D1236-D1244.
- [58] Arnaud, J. *et al.*, Reduction of lectin valency drastically changes glycolipid dynamics in membranes but not surface avidity, *ACS Chem. Biol.* **2013**, *8*, 1918-1924.
- [59] Notova, S. *et al.*, Structure and engineering of tandem repeat lectins, *Curr. Opin. Struct. Biol.* **2020**, *62*, 39-47.
- [60] Turkina, M. V. Vikström, E., Bacteria-Host Crosstalk: Sensing of the Quorum in the Context of *Pseudomonas aeruginosa* Infections, *J. Innate Immun.* **2019**, *11*, 263-279.
- [61] Winzer, K. *et al.*, The *Pseudomonas aeruginosa* lectins PA-IL and PA-IIL are controlled by quorum sensing and by RpoS, *J. Bacteriol.* **2000**, *182*, 6401-6411.
- [62] Imberty, A. Varrot, A., Microbial recognition of human cell surface glycoconjugates, *Curr. Opin. Struct. Biol.* **2008**, *18*, 567-576.
- [63] Varrot, A. *et al.*, Fungal lectins: structure, function and potential applications, *Curr. Opin. Struct. Biol.* **2013**, *23*, 678-685.
- [64] Coltri, K. C. *et al.*, Paracoccin, a GlcNAc-binding lectin from *Paracoccidioides brasiliensis*, binds to laminin and induces TNF- α production by macrophages, *Microb. Infect.* **2006**, *8*, 704-713.
- [65] Houser, J. *et al.*, Protein oligomerization in Aleuria aurantia lectin family - importance and difficulties, *Mater. Struct. Chem. Biol. Phys. Technol.* **2012**, *12* 20-21.
- [66] Geijtenbeek, T. B. Gringhuis, S. I., Signalling through C-type lectin receptors: shaping immune responses, *Nat. Rev. Immunol.* **2009**, *9*, 465-479.
- [67] Takeuchi, O. Akira, S., Pattern Recognition Receptors and Inflammation, *Cell* **2010**, *140*, 805-820.
- [68] Monteiro, J. T. Lepenies, B., Myeloid C-Type Lectin Receptors in Viral Recognition and Antiviral Immunity, *Viruses* **2017**, *9*, 59.
- [69] Bermejo-Jambrina, M. *et al.*, C-Type Lectin Receptors in Antiviral Immunity and Viral Escape, *Front. Immunol.* **2018**, *9*.
- [70] Garcia-Vallejo, J. J. van Kooyk, Y., DC-SIGN: The Strange Case of Dr. Jekyll and Mr. Hyde, *Immunity* **2015**, *42*, 983-985.
- [71] Feinberg, H. *et al.*, Structural basis for selective recognition of oligosaccharides by DC-SIGN and DC-SIGNR, *Science* **2001**, *294*, 2163-2166.
- [72] Guo, Y. *et al.*, Structural basis for distinct ligand-binding and targeting properties of the receptors DC-SIGN and DC-SIGNR, *Nat. Struct. Mol. Biol.* **2004**, *11*, 591-598.
- [73] Mitchell, D. A. *et al.*, A novel mechanism of carbohydrate recognition by the C-type lectins DC-SIGN and DC-SIGNR. Subunit organization and binding to multivalent ligands, *J. Biol. Chem.* **2001**, *276*, 28939-28945.

- [74] Nizet, V. *et al.* *Microbial Lectins : Hemagglutinins, Adhesins , and Toxins*, Cold Spring Harbor Laboratory Press, **2017**.
- [75] Sharon, N., Bacterial lectins, cell-cell recognition and infectious disease, *FEBS Lett.* **1987**, *217*, 145-157.
- [76] Rostand, K. S.Esko, J. D., Microbial adherence to and invasion through proteoglycans, *Infect. Immun.* **1997**, *65*, 1-8.
- [77] Fraser, J. D.Proft, T. *The Streptococcal Superantigens*, **2007**, 1-20.
- [78] Baker, H. M. *et al.*, Crystal Structures of the Staphylococcal Toxin SSL5 in Complex with Sialyl Lewis X Reveal a Conserved Binding Site that Shares Common Features with Viral and Bacterial Sialic Acid Binding Proteins, *J. Mol. Biol.* **2007**, *374*, 1298-1308.
- [79] Chung, M. C. *et al.*, The crystal structure of staphylococcal superantigen-like protein 11 in complex with sialyl Lewis X reveals the mechanism for cell binding and immune inhibition, *Mol. Microbiol.* **2007**, *66*, 1342-1355.
- [80] López-Bueno, A. *et al.*, Host-selected amino acid changes at the sialic acid binding pocket of the parvovirus capsid modulate cell binding affinity and determine virulence, *J. Virol.* **2006**, *80*, 1563-1573.
- [81] Watkins, W. M.Morgan, W. T. J., Neutralization of the Anti-H Agglutinin in Eel Serum by Simple Sugars, *Nature* **1952**, *169*, 825-826.
- [82] Henry, S. *et al.*, Lewis Histo-Blood Group System and Associated Secretory Phenotypes, *Vox Sang.* **1995**, *69*, 166-182.
- [83] Berger, S. A. *et al.*, Relationship between infectious diseases and human blood type, *Eur. J. Clin. Microbiol. Infect. Dis.* **1989**, *8*, 681-689.
- [84] Bishop, J. R.Gagneux, P., Evolution of carbohydrate antigens—microbial forces shaping host glycomes?, *Glycobiology* **2007**, *17*, 23R-34R.
- [85] Henry, S. M., Molecular diversity in the biosynthesis of GI tract glycoconjugates. A blood group related chart of microorganism receptors, *Transfus. Clin. Biol.* **2001**, *8*, 226-230.
- [86] Lindesmith, L. *et al.*, Human susceptibility and resistance to Norwalk virus infection, *Nat. Med.* **2003**, *9*, 548-553.
- [87] Rhim, A. D. *et al.*, Altered terminal glycosylation and the pathophysiology of CF lung disease, *J. Cyst. Fibros.* **2004**, *3*, 95-96.
- [88] Perret, S. *et al.*, Structural basis for the interaction between human milk oligosaccharides and the bacterial lectin PA-III of *Pseudomonas aeruginosa*, *Biochem. J.* **2005**, *389*, 325-332.
- [89] Mitchell, E. *et al.*, Structural basis for oligosaccharide-mediated adhesion of *Pseudomonas aeruginosa* in the lungs of cystic fibrosis patients, *Nat. Struct. Biol.* **2002**, *9*, 918-921.

- [90] Varki, A. *et al.*, Symbol Nomenclature for Graphical Representations of Glycans, *Glycobiology* **2015**, *25*, 1323-1324.
- [91] Mereiter, S. *et al.*, Glycosylation in the Era of Cancer-Targeted Therapy: Where Are We Heading?, *Cancer Cell* **2019**, *36*, 6-16.
- [92] Pinho, S. S.Reis, C. A., Glycosylation in cancer: mechanisms and clinical implications, *Nat. Rev. Cancer* **2015**, *15*, 540-555.
- [93] Costerton, J. W. *et al.*, The bacterial glycocalyx in nature and disease, *Annu. Rev. Microbiol.* **1981**, *35*, 299-324.
- [94] Gabius, H. J. *et al.*, From lectin structure to functional glycomics: principles of the sugar code, *Trends Biochem. Sci.* **2011**, *36*, 298-313.
- [95] Feizi, T.Chai, W., Oligosaccharide microarrays to decipher the glyco code, *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 582-588.
- [96] Increasing Antimicrobial Resistance and the Management of Uncomplicated Community-Acquired Urinary Tract Infections, *Ann. Intern. Med.* **2001**, *135*, 41-50.
- [97] Roberts, J. A. *et al.*, The Gal(alpha 1-4)Gal-specific tip adhesin of Escherichia coli P-fimbriae is needed for pyelonephritis to occur in the normal urinary tract, *Proc. Natl. Acad. Sci.* **1994**, *91*, 11889.
- [98] Cegelski, L. *et al.*, The biology and future prospects of antivirulence therapies, *Nat. Rev. Microbiol.* **2008**, *6*, 17-27.
- [99] Capitani, G. *et al.*, Structural and functional insights into the assembly of type 1 pili from Escherichia coli, *Microbes Infect.* **2006**, *8*, 2284-2290.
- [100] Poole, J. *et al.*, Glycointeractions in bacterial pathogenesis, *Nat. Rev. Microbiol.* **2018**, *16*, 440-452.
- [101] Cusumano, C. K. *et al.*, Treatment and prevention of urinary tract infection with orally active FimH inhibitors, *Sci. Transl. Med.* **2011**, *3*, 3003021.
- [102] Guiton, P. S. *et al.*, Combinatorial small-molecule therapy prevents uropathogenic Escherichia coli catheter-associated urinary tract infections in mice, *Antimicrob. Agents Chemother.* **2012**, *56*, 4738-4745.
- [103] Totsika, M. *et al.*, A FimH inhibitor prevents acute bladder infection and treats chronic cystitis caused by multidrug-resistant uropathogenic Escherichia coli ST131, *J. Infect. Dis.* **2013**, *208*, 921-928.
- [104] Roberts, J. A. *et al.*, The Gal(alpha 1-4)Gal-specific tip adhesin of Escherichia coli P-fimbriae is needed for pyelonephritis to occur in the normal urinary tract, *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 11889-11893.
- [105] Spaulding, C. N. *et al.*, Selective depletion of uropathogenic E. coli from the gut by a FimH antagonist, *Nature* **2017**, *546*, 528-532.

- [106] Bock, K. *et al.*, Specificity of binding of a strain of uropathogenic *Escherichia coli* to Gal alpha 1--4Gal-containing glycosphingolipids, *J. Biol. Chem.* **1985**, *260*, 8545-8551.
- [107] Leffler, H.Svanborg-Edén, C., Glycolipid receptors for uropathogenic *Escherichia coli* on human erythrocytes and uroepithelial cells, *Infect. Immun.* **1981**, *34*, 920-929.
- [108] Lund, B. *et al.*, The PapG protein is the alpha-D-galactopyranosyl-(1----4)-beta-D-galactopyranose-binding adhesin of uropathogenic *Escherichia coli*, *Proc. Natl. Acad. Sci. U. S. A.* **1987**, *84*, 5898-5902.
- [109] Hatakeyama, M., A Sour Relationship between BabA and Lewis b, *Cell Host Microbe* **2017**, *21*, 318-320.
- [110] Dunne, C. *et al.*, Factors that mediate colonization of the human stomach by *Helicobacter pylori*, *World J. Gastroenterol.* **2014**, *20*, 5610-5624.
- [111] Magalhães, A. *et al.*, *Helicobacter pylori* chronic infection and mucosal inflammation switches the human gastric glycosylation pathways, *Biochim. Biophys. Acta, Mol. Basis Dis.* **2015**, *1852*, 1928-1939.
- [112] Ilver, D. *et al.*, *Helicobacter pylori* adhesin binding fucosylated histo-blood group antigens revealed by retagging, *Science* **1998**, *279*, 373-377.
- [113] Mahdavi, J. *et al.*, *Helicobacter pylori* SabA adhesin in persistent infection and chronic inflammation, *Science* **2002**, *297*, 573-578.
- [114] Rossez, Y. *et al.*, The lacdiNac-specific adhesin LabA mediates adhesion of *Helicobacter pylori* to human gastric mucosa, *J. Infect. Dis.* **2014**, *210*, 1286-1295.
- [115] Hilleringmann, M. *et al.*, Molecular architecture of *Streptococcus pneumoniae* TIGR4 pili, *The EMBO Journal* **2009**, *28*, 3921-3930.
- [116] Nelson, A. L. *et al.*, RrgA is a pilus-associated adhesin in *Streptococcus pneumoniae*, *Mol. Microbiol.* **2007**, *66*, 329-340.
- [117] Day, C. J. *et al.*, Lectin activity of the pneumococcal pilin proteins, *Sci. Rep.* **2017**, *7*, 17784.
- [118] King, S. J. *et al.*, Deglycosylation of human glycoconjugates by the sequential activities of exoglycosidases expressed by *Streptococcus pneumoniae*, *Mol. Microbiol.* **2006**, *59*, 961-974.
- [119] Limoli, D. H. *et al.*, BgaA acts as an adhesin to mediate attachment of some pneumococcal strains to human epithelial cells, *Microbiology* **2011**, *157*, 2369-2381.
- [120] Singh, A. K. *et al.*, Unravelling the Multiple Functions of the Architecturally Intricate *Streptococcus pneumoniae* β -galactosidase, BgaA, *PLoS Path.* **2014**, *10*, e1004364.
- [121] de Bentzmann, S.Plésiat, P., The *Pseudomonas aeruginosa* opportunistic pathogen and human infections, *Environ. Microbiol.* **2011**, *13*, 1655-1665.

- [122] Imberty, A. *et al.*, Structures of the lectins from *Pseudomonas aeruginosa*: insight into the molecular basis for host glycan recognition, *Microbes Infect.* **2004**, *6*, 221-228.
- [123] Cioci, G. *et al.*, Structural basis of calcium and galactose recognition by the lectin PA-IL of *Pseudomonas aeruginosa*, *FEBS Lett.* **2003**, *555*, 297-301.
- [124] Mitchell, E. *et al.*, Structural basis for oligosaccharide-mediated adhesion of *Pseudomonas aeruginosa* in the lungs of cystic fibrosis patients, *Nat. Struct. Biol.* **2002**, *9*, 918-921.
- [125] Bajolet-Laudinat, O. *et al.*, Cytotoxicity of *Pseudomonas aeruginosa* internal lectin PA-I to respiratory epithelial cells in primary culture, *Infect. Immun.* **1994**, *62*, 4481-4487.
- [126] Perret, S. *et al.*, Structural basis for the interaction between human milk oligosaccharides and the bacterial lectin PA-III of *Pseudomonas aeruginosa*, *Biochem. J.* **2005**, *389*, 325-332.
- [127] Audfray, A. *et al.*, Fucose-binding lectin from opportunistic pathogen *Burkholderia ambifaria* binds to both plant and human oligosaccharidic epitopes, *J. Biol. Chem.* **2012**, *287*, 4335-4347.
- [128] Ligeour, C. *et al.*, Mannose-centered aromatic galactoclusters inhibit the biofilm formation of *Pseudomonas aeruginosa*, *Org. Biomol. Chem.* **2015**, *13*, 8433-8444.
- [129] Wagner, S. *et al.*, Novel Strategies for the Treatment of *Pseudomonas aeruginosa* Infections, *J. Med. Chem.* **2016**, *59*, 5929-5969.
- [130] Meiers, J. *et al.*, Directing Drugs to Bugs: Antibiotic-Carbohydrate Conjugates Targeting Biofilm-Associated Lectins of *Pseudomonas aeruginosa*, *J. Med. Chem.* **2020**, *63*, 11707-11724.
- [131] Sharon, N., Carbohydrates as future anti-adhesion drugs for infectious diseases, *Biochim. Biophys. Acta* **2006**, *4*, 527-537.
- [132] Gao, C. *et al.*, SARS-CoV-2 Spike Protein Interacts with Multiple Innate Immune Receptors, *bioRxiv : the preprint server for biology* **2020**, 2020.2007.2029.227462.
- [133] Lempp, F. A. *et al.*, Lectins enhance SARS-CoV-2 infection and influence neutralizing antibodies, *Nature* **2021**.
- [134] Thépaut, M. *et al.*, DC/L-SIGN recognition of spike glycoprotein promotes SARS-CoV-2 trans-infection and can be inhibited by a glycomimetic antagonist, *PLoS Pathog.* **2021**, *17*.
- [135] Ofek, I. *et al.*, Adherence of *Escherichia coli* to human mucosal cells mediated by mannose receptors, *Nature* **1977**, *265*, 623-625.
- [136] Zhang, Y. Wang, F., Carbohydrate drugs: current status and development prospect, *Drug Discov. Ther.* **2015**, *9*, 79-87.
- [137] Ernst, B. Magnani, J. L., From carbohydrate leads to glycomimetic drugs, *Nat. Rev. Drug Discov.* **2009**, *8*, 661-677.
- [138] Hevey, R., Strategies for the Development of Glycomimetic Drug Candidates, *Pharmaceuticals (Basel)* **2019**, *12*, 55.

- [139] Magnani, J. L. Ernst, B., Glycomimetic drugs--a new source of therapeutic opportunities, *Discov. Med.* **2009**, *8*, 247-252.
- [140] Lipinski, C. A. *et al.*, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug. Deliv. Rev.* **2001**, *46*, 3-26.
- [141] Veber, D. F. *et al.*, Molecular properties that influence the oral bioavailability of drug candidates, *J. Med. Chem.* **2002**, *45*, 2615-2623.
- [142] de Bentzmann, S. *et al.*, Monitoring lectin interactions with carbohydrates, *Methods Mol. Biol.* **2014**, 0473-0470_0432.
- [143] Pérez, S. Tvaroška, I., Carbohydrate-protein interactions: molecular modeling insights, *Adv. Carbohydr. Chem. Biochem.* **2014**, *71*, 9-136.
- [144] del Carmen Fernández-Alonso, M. *et al.*, Protein-carbohydrate interactions studied by NMR: from molecular recognition to drug design, *Curr. Protein Peptide Sci.* **2012**, *13*, 816-830.
- [145] Hemmi, H. *NMR Analysis of Carbohydrate-Binding Interactions in Solution: An Approach Using Analysis of Saturation Transfer Difference NMR Spectroscopy*, (Ed. J. Hirabayashi), Springer New York, NY, **2014**, 501-509.
- [146] Modenutti, C. *et al.*, Using crystallographic water properties for the analysis and prediction of lectin-carbohydrate complex structures, *Glycobiology* **2015**, *25*, 181-196.
- [147] Sager, C. P. *et al.*, What contributes to an effective mannose recognition domain?, *Beilstein J. Org. Chem.* **2017**, *13*, 2584-2595.
- [148] Quiocho, F. A. *et al.*, Substrate specificity and affinity of a protein modulated by bound water molecules, *Nature* **1989**, *340*, 404-407.
- [149] Wagner, S. *et al.*, Covalent Lectin Inhibition and Application in Bacterial Biofilm Imaging, *Angew. Chem. Int. Ed.* **2017**, *56*, 16559-16564.
- [150] Imberty, A. *et al.*, Glycomimetics and Glycodendrimers as High Affinity Microbial Anti-adhesins, *Chem. Eur. J.* **2008**, *14*, 7490-7499.
- [151] Bertolotti, B. *et al.*, Polyvalent C-glycomimetics based on l-fucose or d-mannose as potent DC-SIGN antagonists, *Org. Biomol. Chem.* **2017**, *15*, 3995-4004.
- [152] Bücher, K. S. *et al.*, Heteromultivalent Glycooligomers as Mimetics of Blood Group Antigens, *Chem. Eur. J.* **2019**, *25*, 3301-3309.
- [153] Cecioni, S. *et al.*, Glycomimetics versus multivalent glycoconjugates for the design of high affinity lectin ligands, *Chem. Rev.* **2015**, *115*, 525-561.
- [154] Moog, K. E. *et al.*, Polymeric Selectin Ligands Mimicking Complex Carbohydrates: From Selectin Binders to Modifiers of Macrophage Migration, *Angew. Chem. Int. Ed.* **2017**, *56*, 1416-1421.

- [155] Zhang, X. *et al.*, Synthesis of Fucosylated Chondroitin Sulfate Glycoclusters: A Robust Route to New Anticoagulant Agents, *Chem. Eur. J.* **2018**, *24*, 1694-1700.
- [156] Simon, P. M. *et al.*, Inhibition of *Helicobacter pylori* binding to gastrointestinal epithelial cells by sialic acid-containing oligosaccharides, *Infect. Immun.* **1997**, *65*, 750-757.
- [157] Horst, A. *et al.*, Binding inhibition of type 1 fimbriae to human granulocytes: a flow cytometric inhibition assay using trivalent cluster mannosides, *Med. Microbiol. Immunol.* **2001**, *190*, 145-149.
- [158] Palmioli, A. *et al.*, Targeting Bacterial Biofilm: A New LecA Multivalent Ligand with Inhibitory Activity, *ChemBioChem* **2019**, *20*, 2911-2915.
- [159] García-Moreno, M. I. *et al.*, The Impact of Heteromultivalency in Lectin Recognition and Glycosidase Inhibition: An Integrated Mechanistic Study, *Chem. Eur. J.* **2017**, *23*, 6295-6304.
- [160] Ordanini, S. *et al.*, Designing nanomolar antagonists of DC-SIGN-mediated HIV infection: ligand presentation using molecular rods, *Chem. Commun.* **2015**, *51*, 3816-3819.
- [161] Lundquist, J. J. Toone, E. J., The cluster glycoside effect, *Chem. Rev.* **2002**, *102*, 555-578.
- [162] Kiessling, L. L. *et al.*, Synthetic multivalent ligands as probes of signal transduction, *Angew. Chem. Int. Ed. Engl.* **2006**, *45*, 2348-2368.
- [163] Boden, S. *et al.*, Sequence-Defined Introduction of Hydrophobic Motifs and Effects in Lectin Binding of Precision Glycomacromolecules, *Macromol. Biosci.* **2019**, *19*, 1800425.
- [164] Berzi, A. *et al.*, Pseudo-Mannosylated DC-SIGN Ligands as Immunomodulants, *Sci. Rep.* **2016**, *6*, 35373.
- [165] Kim, C. U. *et al.*, Influenza neuraminidase inhibitors possessing a novel hydrophobic interaction in the enzyme active site: design, synthesis, and structural analysis of carbocyclic sialic acid analogues with potent anti-influenza activity, *J. Am. Chem. Soc.* **1997**, *119*, 681-690.
- [166] McClellan, K. Perry, C. M., Oseltamivir, *Drugs* **2001**, *61*, 263-283.
- [167] Cox, T. *et al.*, Novel oral treatment of Gaucher's disease with N-butyldeoxynojirimycin (OGT 918) to decrease substrate biosynthesis, *The Lancet* **2000**, *355*, 1481-1485.
- [168] Campbell, L. K. *et al.*, Miglitol: assessment of its role in the treatment of patients with diabetes mellitus, *Ann. Pharmacother.* **2000**, *34*, 1291-1301.
- [169] Chen, X. *et al.*, Voglibose (Basen[®], AO-128), One of the Most Important α -Glucosidase Inhibitors, *Curr. Med. Chem.* **2006**, *13*, 109-116.
- [170] Truscheit, E. *et al.*, Chemistry and Biochemistry of Microbial α -Glucosidase Inhibitors, *Angew. Chem., Int. Ed. Engl.* **1981**, *20*, 744-761.
- [171] Aronson, M. *et al.*, Prevention of colonization of the urinary tract of mice with *Escherichia coli* by blocking of bacterial adherence with methyl α -D-mannopyranoside, *J. Infect. Dis.* **1979**, *139*, 329-332.

- [172] Waksman, G.Hultgren, S. J., Structural biology of the chaperone-usher pathway of pilus biogenesis, *Nat. Rev. Microbiol.* **2009**, *7*, 765-774.
- [173] Firon, N. *et al.*, Carbohydrate specificity of the surface lectins of Escherichia coli, Klebsiella pneumoniae, and Salmonella typhimurium, *Carbohydr. Res.* **1983**, *120*, 235-249.
- [174] Neeser, J. R. *et al.*, Oligomannoside-type glycopeptides inhibiting adhesion of Escherichia coli strains mediated by type 1 pili: preparation of potent inhibitors from plant glycoproteins, *Infect. Immun.* **1986**, *52*, 428-436.
- [175] Bouckaert, J. *et al.*, Receptor binding studies disclose a novel class of high-affinity inhibitors of the Escherichia coli FimH adhesin, *Mol. Microbiol.* **2005**, *55*, 441-455.
- [176] Firon, N. *et al.*, Aromatic alpha-glycosides of mannose are powerful inhibitors of the adherence of type 1 fimbriated Escherichia coli to yeast and intestinal epithelial cells, *Infect. Immun.* **1987**, *55*, 472-476.
- [177] Sperling, O. *et al.*, Evaluation of the carbohydrate recognition domain of the bacterial adhesin FimH: Design, synthesis and binding properties of mannoside ligands, *Org. Biomol. Chem.* **2006**, *4*, 3913-3922.
- [178] Han, Z. *et al.*, Structure-based drug design and optimization of mannoside bacterial FimH antagonists, *J. Med. Chem.* **2010**, *53*, 4779-4792.
- [179] Jiang, X. *et al.*, Antiadhesion Therapy for Urinary Tract Infections—A Balanced PK/PD Profile Proved To Be Key for Success, *J. Med. Chem.* **2012**, *55*, 4700-4713.
- [180] Pang, L. *et al.* FimH antagonists – solubility vs. permeability, *R. Soc. Chem.* **2017**, 248-273.
- [181] Rodrigue, J. *et al.*, Aromatic thioglycoside inhibitors against the virulence factor LecA from Pseudomonas aeruginosa, *Org. Biomol. Chem.* **2013**, *11*, 6906-6918.
- [182] Kadam, R. U. *et al.*, Structure-Based Optimization of the Terminal Tripeptide in Glycopeptide Dendrimer Inhibitors of Pseudomonas aeruginosa Biofilms Targeting LecA, *Chem. Eur. J.* **2013**, *19*, 17054-17063.
- [183] Kadam, R. U. *et al.*, A Glycopeptide Dendrimer Inhibitor of the Galactose-Specific Lectin LecA and of Pseudomonas aeruginosa Biofilms, *Angew. Chem. Int. Ed.* **2011**, *50*, 10631-10635.
- [184] Boukerb, A. M. *et al.*, Antiadhesive properties of glycoclusters against Pseudomonas aeruginosa lung infection, *J. Med. Chem.* **2014**, *57*, 10275-10289.
- [185] José, R. J. *et al.*, Opportunistic bacterial, viral and fungal infections of the lung, *Medicine (Abingdon, England : UK ed.)* **2020**, *48*, 366-372.
- [186] Zanoni, B. C.Gandhi, R. T., Update on opportunistic infections in the era of effective antiretroviral therapy, *Infect. Dis. Clin. North Am.* **2014**, *28*, 501-518.

- [187] De Smet, B. *et al.*, *Burkholderia stagnalis* sp. nov. and *Burkholderia territorii* sp. nov., two novel *Burkholderia cepacia* complex species from environmental and human sources, *Int. J. Syst. Evol. Microbiol.* **2015**, *65*, 2265-2271.
- [188] Ong, K. S. *et al.*, *Burkholderia paludis* sp. nov., an Antibiotic-Siderophore Producing Novel *Burkholderia cepacia* Complex Species, Isolated from Malaysian Tropical Peat Swamp Soil, *Front. Microbiol.* **2016**, *7*.
- [189] Vanlaere, E. *et al.*, *Burkholderia latens* sp. nov., *Burkholderia diffusa* sp. nov., *Burkholderia arboris* sp. nov., *Burkholderia seminalis* sp. nov. and *Burkholderia metallica* sp. nov., novel species within the *Burkholderia cepacia* complex, *Int. J. Syst. Evol. Microbiol.* **2008**, *58*, 1580-1590.
- [190] Weber, C. F. King, G. M., Volcanic Soils as Sources of Novel CO-Oxidizing Paraburkholderia and Burkholderia: Paraburkholderia hiiakae sp. nov., Paraburkholderia metrosideri sp. nov., Paraburkholderia paradisi sp. nov., Paraburkholderia peleae sp. nov., and Burkholderia alpina sp. nov. a Member of the Burkholderia cepacia Complex, *Front. Microbiol.* **2017**, *8*.
- [191] Mahenthiralingam, E. *et al.*, The multifarious, multireplicon Burkholderia cepacia complex, *Nat. Rev. Microbiol.* **2005**, *3*, 144-156.
- [192] Winkelstein, J. A. *et al.*, Chronic granulomatous disease. Report on a national registry of 368 patients, *Medicine* **2000**, *79*, 155-169.
- [193] Butler, S. L. *et al.*, Burkholderia cepacia and cystic fibrosis: do natural environments present a potential hazard?, *J. Clin. Microbiol.* **1995**, *33*, 1001-1004.
- [194] Loutet, S. A. Valvano, M. A., A decade of Burkholderia cenocepacia virulence determinant research, *Infect. Immun.* **2010**, *78*, 4088-4100.
- [195] Sousa, S. A. *et al.*, Burkholderia cepacia Complex: Emerging Multihost Pathogens Equipped with a Wide Range of Virulence Factors and Determinants, *Int. J. Microbiol.* **2011**, *607575*, 3.
- [196] Hogardt, M. Heesemann, J., Adaptation of Pseudomonas aeruginosa during persistence in the cystic fibrosis lung, *Int. J. Med. Microbiol.* **2010**, *300*, 557-562.
- [197] Döring, G. *et al.*, Differential adaptation of microbial pathogens to airways of patients with cystic fibrosis and chronic obstructive pulmonary disease, *FEMS Microbiol. Rev.* **2011**, *35*, 124-146.
- [198] Cullen, L. McClean, S., Bacterial Adaptation during Chronic Respiratory Infections, *Pathogens* **2015**, *4*, 66-89.
- [199] Lieberman, T. D. *et al.*, Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes, *Nat. Genet.* **2011**, *43*, 1275-1280.
- [200] Madeira, A. *et al.*, Quantitative proteomics (2-D DIGE) reveals molecular strategies employed by Burkholderia cenocepacia to adapt to the airways of cystic fibrosis patients under antimicrobial therapy, *Proteomics* **2011**, *11*, 1313-1328.

- [201] Moreira, A. S. *et al.*, Burkholderia dolosa phenotypic variation during the decline in lung function of a cystic fibrosis patient during 5.5 years of chronic colonization, *J. Med. Microbiol.* **2014**, *63*, 594-601.
- [202] Nunvar, J. *et al.*, What matters in chronic Burkholderia cenocepacia infection in cystic fibrosis: Insights from comparative genomics, *PLoS Pathog.* **2017**, *13*.
- [203] Silva, I. N. *et al.*, Mucoïd morphotype variation of Burkholderia multivorans during chronic cystic fibrosis lung infection is correlated with changes in metabolism, motility, biofilm formation and virulence, *Microbiology* **2011**, *157*, 3124-3137.
- [204] Silva, I. N. *et al.*, Long-Term Evolution of Burkholderia multivorans during a Chronic Cystic Fibrosis Infection Reveals Shifting Forces of Selection, *mSystems* **2016**, *1*, e00029-00016.
- [205] Gislason, A. S. *et al.*, Competitive Growth Enhances Conditional Growth Mutant Sensitivity to Antibiotics and Exposes a Two-Component System as an Emerging Antibacterial Target in Burkholderia cenocepacia, *Antimicrob. Agents Chemother.* **2016**, *61*, 00790-00716.
- [206] Wong, Y. C. *et al.*, Candidate Essential Genes in Burkholderia cenocepacia J2315 Identified by Genome-Wide TraDIS, *Front. Microbiol.* **2016**, *7*.
- [207] Lameignere, E. *et al.*, Structural basis for mannose recognition by a lectin from opportunistic bacteria Burkholderia cenocepacia, *Biochem. J.* **2008**, *411*, 307-318.
- [208] Sulák, O. *et al.*, Burkholderia cenocepacia BC2L-C is a super lectin with dual specificity and proinflammatory activity, *PLoS Path.* **2011**, *7*, e1002238-e1002238.
- [209] Sulák, O. *et al.*, A TNF-like trimeric lectin domain from Burkholderia cenocepacia with specificity for fucosylated human histo-blood group antigens, *Structure* **2010**, *18*, 59-72.
- [210] Adam, J. *et al.*, Engineering of PA-III lectin from Pseudomonas aeruginosa – Unravelling the role of the specificity loop for sugar preference, *BMC Struct. Biol.* **2007**, *7*, 36.
- [211] De Soya, A. *et al.*, Chemical and biological features of Burkholderia cepacia complex lipopolysaccharides, *Innate Immun.* **2008**, *14*, 127-144.
- [212] Silipo, A. Molinaro, A., The diversity of the core oligosaccharide in lipopolysaccharides, *Subcell. Biochem.* **2010**, *53*, 69-99.
- [213] Beshr, G. *et al.*, Development of a competitive binding assay for the Burkholderia cenocepacia lectin BC2L-A and structure activity relationship of natural and synthetic inhibitors, *MedChemComm* **2016**, *7*, 519-530.
- [214] Csávás, M. *et al.*, Tri- and tetravalent mannoclusters cross-link and aggregate BC2L-A lectin from Burkholderia cenocepacia, *Carbohydr. Res.* **2017**, *437*, 1-8.
- [215] Lameignere, E. *et al.*, Structural basis of the affinity for oligomannosides and analogs displayed by BC2L-A, a Burkholderia cenocepacia soluble lectin, *Glycobiology* **2010**, *20*, 87-98.

- [216] Reynolds, M. *et al.*, Influence of ligand presentation density on the molecular recognition of mannose-functionalised glyconanoparticles by bacterial lectin BC2L-A, *Glycoconj. J.* **2013**, *30*, 747-757.
- [217] Inhülsen, S. *et al.*, Identification of functions linking quorum sensing with biofilm formation in *Burkholderia cenocepacia* H111, *Microbiologyopen* **2012**, *1*, 225-242.
- [218] Schmid, N. *et al.*, The AHL- and BDSF-dependent quorum sensing systems control specific and overlapping sets of genes in *Burkholderia cenocepacia* H111, *PLoS One* **2012**, *7*, 20.
- [219] Diggle, S. P. *et al.*, The galactophilic lectin, LecA, contributes to biofilm development in *Pseudomonas aeruginosa*, *Environ. Microbiol.* **2006**, *8*, 1095-1104.
- [220] Duchaud, E. *et al.*, The genome sequence of the entomopathogenic bacterium *Photorhabdus luminescens*, *Nat. Biotechnol.* **2003**, *21*, 1307-1313.
- [221] Réty, S. *et al.*, The Crystal Structure of the *Bacillus anthracis* Spore Surface Protein BclA Shows Remarkable Similarity to Mammalian Proteins *J. Biol. Chem.* **2005**, *280*, 43073-43078.
- [222] Thomas, W. E. *et al.*, Bacterial adhesion to target cells enhanced by shear force, *Cell* **2002**, *109*, 913-923.
- [223] Marchetti, R. *et al.*, *Burkholderia cenocepacia* lectin A binding to heptoses from the bacterial lipopolysaccharide, *Glycobiology* **2012**, *22*, 1387-1398.
- [224] Phd4GlycoDrug ITN webpage. <https://www.phd4glycodrug.eu/> (Accessed August 2021).
- [225] Glycopedia Glycoscience Portal. <https://www.glycopedia.eu/> (Accessed August 2021).
- [226] Lal, K. *et al.*, Computational tools for drawing, building and displaying carbohydrates: a visual guide, *Beilstein J. Org. Chem.* **2020**, *16*, 2448-2468.

2. Research methodology

2.1 Introduction: The domain of structure-based drug design

The past few decades have experienced a fundamental change in preclinical drug discovery with structure-based drug design (SBDD) growing in popularity while the traditional methods of high-throughput screening (HTS) have continued to witness unsatisfactory results.¹ One of the disadvantages of using HTS is that it does not provide mechanistic information about ligand-receptor interactions. Likewise, HTS screening can have influence of physical artefact such as precipitation or any component of the assay. Moreover, the results of HTS screening in form of large amount of data require significant efforts to be analyzed.

SBDD approach plays an important role in investigating the interactions between the ligands and receptors and contributes towards in-depth understanding of molecular recognition/interactions. This also helps to optimize lead molecule and bridge the gap between an identified hit to a preclinical drug candidate.¹ The SBDD is based on the hypothesis that a molecule's ability to exert a desired biologic effect depends on its ability to favourably interact with a particular binding site on a protein. Molecules that share those favourable interactions usually exert similar biological effects. The initial projects based on SBDD were under progress in the mid-80s, and by the early 1990s few success stories were published.²⁻³ Structural information about the target is a prerequisite for SBDD project. Recent advancements in the field of genomics and proteomics have led to the discovery of a large number of drug targets.⁴⁻⁵ Efficient use of advanced biophysical techniques such as X-ray crystallography and NMR spectroscopy has also increased the motivation for elucidation of the structures of a large number of protein from pathogenic microorganisms.⁶ Consequently, the field of structure-based drug design became an integral part of many

industrial drug discovery projects and a major subject of research in academic laboratories.⁷

Different methods involving computational and experimental techniques to underpin the structure-based drug design are discussed in this chapter. These methods are particularly useful to design the ligands and to further investigate the ligand-protein interactions. They are divided into two categories given below.

Computational methods to identify druggable sites in proteins and to investigate ligand-protein interactions:

- Prediction of druggable sites using SiteMap
- Virtual screening: Glide docking
- MD simulations
- Grid Inhomogeneous Solvation Theory (GIST) to study water thermodynamics

Experimental (biophysical) techniques to validate the fragment/ligand-protein interactions:

- Production of BC2L-C-nt
- Thermal shift assay (TSA)
- Microscale thermophoresis (MST)
- Saturation Transfer Difference (STD) -NMR
- X-ray crystallography
- Isothermal titration calorimetry (ITC)

The specific protocols and materials used for experiments will be discussed under a relevant section in each chapter.

2.2 Prediction of druggable sites using SiteMap

The ligand binding site on a receptor such as a protein is usually known from the structure of the receptor in complex with a ligand. In order to enhance the binding affinity of ligands, understanding of ligands and receptor complementarity is highly important. This can

further guide in extending the ligands into adjacent regions near binding site to promote efficient binding. Therefore, initial investigation of the receptor is required to study and identify nearby sites that might be useful for fragment/ligand binding. In some cases, the apo form of the receptor is available which lacks the information about binding sites for protein-ligand interactions. In this situation, computational studies can help to suggest likely binding sites, and even to predict the druggability/ligandability of the site. In literature, such approaches have been already reported.⁸⁻⁹

SiteMap¹⁰ is a tool from Schrödinger¹¹ that generates important information on the key properties of binding sites in proteins. This tool uses a search and analysis approach and generates important information which can be visualised in Maestro.¹¹ The calculations involve initial search stage to determine different druggable regions on or near the protein surface which are known as *sites*. These regions that can be suitable for ligand binding are represented by grid of points, called *site points*. The site-finding algorithm places a 1-Å grid of possible site points around the entire protein or a ligand. Identification of sites involves different steps. In the first step, grid points are categorized as being “inside” or “outside” the protein. The distance of grid point from the protein atoms is compared to the van der Waals radius of each protein atom. In case the ratio of the squares of the distances is greater than a threshold distance, the grid point is considered outside the protein. Next, the “outside” points are analysed to understand good van der Waals contact with the receptor as well as their enclosure by the receptor to serve as site points. To define the enclosure, all directions are sampled from the grid point to determine the fraction of directions that strike the surface within a defined distance. If the fraction is higher than a defined threshold, the point is sufficiently enclosed and considered as a site point. At a site point, if the magnitude of van der Waals interaction energy is too small, the point is rejected. Only site points that meet the

defined criteria are added to the list. In the third step, site points are combined into different site-point groups. A group must have a minimum number of candidate site points within a given distance otherwise they are discarded. Finally, site-point groups (with small distances between them) located in a solvent-exposed region are merged. In the next stage, 3D grid is generated by placing a probe (that simulates water) at each of the grid point to calculate van der Waals and electrostatic interactions. Thereafter, contour maps (*site maps*) displaying hydrophobic and hydrophilic maps are generated. Further, the hydrophilic maps can be shown as donor and acceptor, and metal-binding regions etc. The results are evaluated based on the calculations which involve assessment of each site for various properties. This usually includes the different features like, van der Waals interaction energy, depth, size, hydrophilicity and hydrophobicity that contribute to the druggability of a protein region. Based on these characteristics, a single scoring function called SiteScore is assigned to potential druggable regions. SiteMap can assist in the design of high-affinity ligands, by predicting druggable regions that can accommodate the desired groups, for example, larger hydrophobic groups of ligands in the hydrophobic regions. In addition, it can also be useful to select the appropriate site for ligand docking using docking tools like Glide¹² followed by evaluation of docking hits for their complementarity to the receptor. Generally, the regions that are neither very hydrophobic nor very hydrophilic are interesting for drug design purpose because these sites allow further modifications in the physical properties of the ligand. For instance, changing the solubility might have minimal effect on the binding affinity of the ligands. The results of the SiteMap depend on the site as a whole and explicitly show the shape and extent of hydrophilic and hydrophobic regions, which is not possible with a surface-based display.

2.3 Virtual screening: docking studies

Most of the biological processes involve the interaction between two or more biological systems. The molecular characterization of these recognition processes is important to understand various mechanisms including disease research and the development of drugs.¹³⁻¹⁴ These important tasks can be accomplished computationally with the help of molecular docking approaches. Molecular docking is the computer-aided prediction of the bound geometry by utilizing the coordinates of the unbound components (which can be proteins, carbohydrates, peptides or small molecules).

The initial coordinates for the receptor (proteins) are usually available from crystallographic and NMR experiments or from homology models. Likewise, initial conformation of ligands or small molecules can be created using computational tools.¹⁵ The docking procedure usually generates several possible complexes known as docking poses. These multiple poses represent the different local minima. The docking procedures aim to achieve free energy minimum corresponding to the native binding mode of the ligands.

The docking methods usually include two steps: the first step involves conformational sampling of the ligand in the active site of the protein while in the second step, conformations are ranked via a scoring function. The sampling step employs a search algorithm that generates several possible binding modes of the ligands by varying their conformations (flexible docking) and orientations in the binding site of the receptor. The scoring procedure is based on scoring functions which approximate the binding affinity between two molecules and allow ranking the binding poses. These two steps (sampling and scoring) are iterated to converge to a solution of minimum energy. In addition to prediction of ligand-binding poses, docking can be used for different studies based on drug-receptor interactions such as lead optimization, virtual screening and library design. In the field of computer-aided drug design,

docking software are used for virtual screening campaigns as well as in the lead optimization of the compounds.¹⁶ The description of molecular interactions is useful to identify key residues involved in the ligand binding.¹⁷⁻¹⁸ Likewise, analysis of the mutations in the receptor is helpful to investigate the bases of drug-resistance.¹⁵ Docking techniques also assist in the understanding of the molecular mechanisms of selectivity, by exploring the interactions of the same molecule with different receptor targets.¹⁹ Various software packages are devoted to perform molecular docking with different scoring functions and sampling methods.^{12,20-25}

As discussed above, docking is based on two steps involving a sampling procedure to generate a wide variety of possible binding modes and a scoring step that aims to rank the poses.^{16,26}

2.3.1 Sampling

In general, the binding process involves changes in the respective orientation of the ligand and receptor as well as the conformational degrees of freedom of ligand and protein can generate several conformations. Consequently, a huge number of binding modes are possible between protein and ligand. Unfortunately, it would be highly expensive from the computational point of view to exhaustively sample all the possible binding modes. To tackle this issue, three strategies with a different degree of exhaustiveness are usually adopted in the docking algorithms. The first approach is based on rigid ligand and rigid receptor, in which search space is very limited by exploring only the six rotational and translational degrees of freedom while both the receptor and the ligand are treated as rigid bodies. In this approach, pre-computed set of ligand conformations can be used for docking calculations. Second method is based on rigid receptor and flexible ligand approach in which the conformational degrees of freedom of the ligand are sampled. The third approach is based on fully or partially flexible protein and flexible ligand in which all the degrees of freedom of the ligand are

investigated while all or only few relevant conformational degrees of freedom of the protein are sampled.²⁷

In order to balance the speed and the accuracy of the docking calculations, currently most popular docking methods treat the ligand as flexible and the receptor as a rigid-body. Various sampling algorithms have been developed and being used in different molecular docking software.²⁶ Systematic and stochastic sampling are two methods that are generally used to perform ligand sampling. In the systematic sampling, incremental variations in each structural parameter explores the free-energy landscape to achieve convergence to minimum energy conformation.²⁸ However, this method has higher computational cost while handling highly flexible ligands due to an enormous increase in the combination of structural parameters. In opposite, the stochastic methodologies involve a random change in the structural parameters at each step that generates a wide variety of possible solutions. These methods either accept or reject the proposed solutions according as per the requirement, thus limit the computational cost. Monte Carlo and genetic algorithms are examples of stochastic methods. However, this methodology may not guarantee the convergence to the global minimum, thus multiple independent runs are required to achieve the optimal solution.²⁷ In addition to the ligand conformations, receptor flexibility also constitutes one of the major challenges in the docking methodologies. It also plays an essential role in the process of ligand binding as the receptor undergoes structural rearrangements due to induced effect. These changes in the receptor can have small effects limited to binding site region or larger effects that may result in conformational rearrangements affecting the whole receptor.²⁸ The flexibility of receptor is also crucial when the free receptor could exist in multiple conformational states.²⁹ Therefore, various methods can be adopted to deal with the receptor flexibility. These methods differ in the degree of exhaustiveness and their

accuracy.^{27,30} In Soft-docking approach, the repulsive contribution of the van der Waals potential is softened to allow small atoms overlaps, thus ligands get accommodated more easily in the binding site. This method is applicable when small local receptor motions occur. Another method involves side-chain flexibility by sampling the side chain conformations that can be employed only for local motions of the receptor. Likewise, methods based on post-processing involve molecular relaxation using Monte Carlo or MD simulations. Additional methods involve ensemble docking, in which multiple conformations of the receptor are generated to dock the ligand. These putative conformations of the receptor can come from experimental methods like NMR and X-ray crystallography or from computational models (e.g. molecular modelling and MD). However, this method has limitations when receptor undergoes large structural rearrangements. Likewise, approaches based on collective degrees of freedom consider the full flexibility of the receptor by reducing its high-dimensional conformational landscape that capture only the dominant motion modes. This can be achieved by using methods such as normal mode analysis or principal component analysis.

2.3.2 Scoring

The purpose of the scoring function is to rank the binding poses that are generated during the sampling of molecules. The scoring functions estimate the binding affinity between the protein and ligand by adopting various assumptions and simplifications and perform the calculations in a reasonable time to save on computation cost.²⁷ They can guide the search methods towards relevant ligand conformations and distinguish the experimentally observed binding modes from all the other predicted poses. The scoring functions are mathematical equations that consist of different terms representing physical properties of the two interacting molecules. For instance, simple scoring functions may consider different interaction terms such as hydrogen bonds, hydrophobic contacts and salt bridges, etc.¹⁶

Conversely, more complicated and time-consuming functions involve additional terms, for instance, contribution of entropic effects and desolvation.^{28,31} Scoring functions can be classified into three types: force-field-based, empirical and knowledge-based.³²

Force field-based functions are derived from a classical force field where the binding energy is computed as the sum of bonded (bond stretching, angle bending and torsional energy) and non-bonded terms (electrostatic and van der Waals interactions). For example, D-score involves van der Waals and electrostatic energy terms to describe interactions between ligand and receptor (equation 2.1 and 2.2).³³ The van der Waals energy term is given by a Lennard–Jones potential function.

$$D - Score = E_{vdw} + E_{electrostatic} \quad (2.1)$$

$$\sum_{prot} \sum_{lig} \left[\left(\frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^6} \right) + 332.0 \frac{q_i q_j}{\epsilon(d_{ij}) d_{ij}} \right] \quad (2.2)$$

where $E_{electrostatic}$, represents Coulomb energy term and E_{vdw} is van der Waals energy term, i and j are two atoms, A_{ij} and B_{ij} are van der Waals parameters for given atom types, d_{ij} is the interatomic distance, q_i and q_j are atomic partial charges, and $\epsilon(d_{ij})$ is a distance-dependent dielectric function. One of the limitations of these scoring functions is that they often overestimate the interactions between charged atoms.²⁷ Further extension of these scoring functions includes desolvation and entropic effects which are described by desolvation energy and conformational entropy terms, respectively.³³ Moreover, the recent developments involve quantum mechanics (QM) to address the challenges of covalent interactions, charge transfer and polarization in docking. However, the QM-based scoring functions have greater computational cost.^{27,33}

A second class of scoring-functions involves empirical methods.³⁴⁻³⁸ In this approach, the scoring-function is based on sum of the terms which reflect the general features of the

complex such as hydrogen bond, ionic interaction, hydrophobic effect and binding entropy etc.²⁷ Each component of the scoring function is simplified and weighted with proper coefficients, and then summed up to give a final score to the complexes. The coefficients for these scoring functions are usually obtained from regression analysis or machine learning approaches.¹⁵ Machine-learning-based scoring functions employ a variety of algorithms, such as neural network, deep-learning, support vector machine that usually rely on the training dataset. Thus, accuracy of these scoring functions is usually related to the quality of the training set.¹⁵

Knowledge-based scoring functions rely on statistical analysis of experimentally determined receptor-ligand structures by obtaining interatomic contact frequencies and/or distances between the ligand and protein. These scoring functions are based on the assumption that the frequent contacts appearing in several complexes, correspond to favourable interactions. These frequency distributions are further used to construct the pairwise atom-type potentials. The calculation of final score is based on the favourable contacts and unfavourable repulsive interactions between each atom in the ligand and protein. These scoring-functions are fast and their interaction potential also includes the features that cannot be modelled explicitly.¹⁶ However, the accuracy of such methods rely on the number and diversity of the training set employed to create the potential.

All the scoring functions have their limitations, therefore in order to improve the performance of these scoring-functions, consensus scoring approach is usually recommended. This strategy involves the use of multiple docking software and scoring functions. Finally, the scores obtained are combined through a consensus scheme. This approach usually provides notable improvement in terms of accuracy.³⁹ At present, most scoring-functions are able to predict native or near-native binding modes, however accurate

estimation of the free energy of binding and ranking of different compounds toward a target receptor is still a challenging task. Consequently, the available scoring functions show a weak correlation with experimental binding affinities and target dependent performance.³¹ These challenges in reproducing experimental results are because of several reasons, for example, poor quality of the input ligand and receptor structures, improper treatment of long-range interactions, an inappropriate handling of solvent and entropic effects.⁴⁰ To avoid the computational cost, contribution of conformational entropy in the determination of binding free energy is usually neglected or oversimplified in most of the docking software.²⁷ However, methodologies to deal with solvent effects have been developed. In this regards, structural waters are considered for the solvent effect, which play important role in mediating receptor-ligand interactions.¹⁵ The consideration of structural waters in docking calculations significantly improves the accuracy of result.^{15,28} Some methods that explicitly account for water molecules during the docking process have been proposed.⁴¹⁻⁴⁴ However, such methods are computationally expensive due to the large number of degrees of freedom associated to the solvent molecules. To reduce the computational time, some methods treat the water molecules implicitly by representing them as a continuum dielectric medium. Molecular Mechanics generalized-Born/surface area (MM-GB/SA) and Molecular Mechanics Poisson-Boltzmann surface area (MM-PB/SA) are the most widely used methods based on this strategy.⁴⁵⁻⁴⁷ In order to improve the accuracy of ranking and binding energy predictions, these approaches are generally employed as post-processing step to rescore the docking poses. MM-GB/SA and MM-PB/SA have been successfully employed to estimate the free energy of binding in different studies on protein–ligand interactions.⁴⁸⁻⁴⁹

2.3.3 Glide software

Glide (Grid-based Ligand Docking with Energetics)¹² is a ligand docking program for predicting protein-ligand binding poses and ranking them after virtual screening. This program uses systematic conformational search and employs a scoring-function that relies on the empirical and force-field-based terms. Glide uses three different docking protocols with different accuracies. The Standard-Precision (SP) approach employs a *soft* scoring function which can identify a wide pool of possible binders and minimize false negatives. Similar to SP, HTVS (high-throughput virtual screening) approach uses the same scoring function but it decreases the number of intermediate conformations by reducing the thoroughness of sampling and final torsional refinement. However, the algorithm itself is essentially the same. Another docking approach known as Extra-Precision (XP) uses more extensive sampling than SP and is characterized as *harder* scoring-function with greater requirements for ligand-receptor shape complementarity. This attempts to minimize the false positives by including additional penalties.^{12,21,50} This weeds out false positives that are passed through SP approach.

Glide uses a series of filters to generate the best binding pose of the ligand in the binding-site region of the receptor. At a first stage, several conformations of the ligands are generated by exploring its torsional angle space. The conformations with lower torsional energy are selected using a model energy function that uses force-field (OPLS) based terms.¹² In the second step, these conformers are screened on the binding site of the receptor to identify their possible positions and orientations. The placement of the conformations is then evaluated by assigning energetic-like properties of the protein on a grid which assigns a precomputed score (derived from a discretized version of the empirical ChemScore function)⁵¹ to boxes of 1 Å³ dimensions. The penalties are also assigned to the scores in case

of steric clashes. This score is able to recognize favourable hydrophobic interactions, hydrogen-bonds and metal-ligation interactions.

In the next stage, the best poses of the ligands are minimized in the field of the receptor, using a molecular mechanics scoring function and a multi-grid strategy. In this approach, the side-dimension of the boxes, which store the Coulomb/van der Waals fields of the receptor, are gradually decreased in the area where the two molecules are in contact that improves the accuracy of the calculations. The best three to six lowest-energy poses are minimized with a Monte Carlo approach. This last phase plays an important role in the accurate prediction of docking poses. Finally, *GlideScore* is used to re-score the minimized poses. *GlideScore* is based on ChemScore⁵¹ and includes terms such as steric-clash term, rewards, penalties for electrostatic mismatches, amide twist penalties, hydrophobic enclosure terms etc. The components of *GlideScore* (GScore) can be described as:

$$GScore = 0.05 * vdW + 0.15 * Coul + Lipo + Hbond + Metal + Rewards + RotB + Site \quad (2.3)$$

where vdW is Van der Waals energy, Coul is Coulomb energy, Lipo is lipophilic term, HBond is hydrogen-bonding term, RotB is the penalty awarded for freezing rotatable bonds, Metal is metal-binding term, Site is the term for polar interactions in the active site, Rewards represents the rewards and penalties for hydrophobic enclosure, buried polar groups, amide twists. Hydrophobic enclosure is a reward for the displacement of water molecules by a ligand from the binding site region with many proximal lipophilic protein atoms. Formation of protein-ligand hydrogen bonds within regions of hydrophobic enclosure are beneficial to binding. Likewise, electrostatic mismatches are penalties for buried polar terms and amide twists are energy penalties for the non-planar conformation of amides. This penalty is applied

to twisted and cis conformations. Therefore, it is possible to obtain poses with twisted amides if there are compensating interactions.

2.4 Molecular dynamics (MD) simulations

The molecular dynamics method was introduced in the late 1950's.⁵² The next major advancement took place in 1964 when first simulation for liquid argon was performed using a realistic potential.⁵³ Subsequently, the first realistic system was simulated for liquid water in 1974.⁵⁴ Later on 1977, protein simulation of the bovine pancreatic trypsin inhibitor was reported.⁵⁵ The advancements in the MD simulation methods and computation have made it possible to expand these methods for biological complexes addressing the thermodynamics of ligand binding and the folding of small proteins. Increasing computational power has also advanced the molecular simulations studies allowing simulations up to millisecond timescales.⁵⁶ Similarly, emerging new protocols for MD simulations studies have also increased scientific interest in correlation of molecular simulations results with experimental data.⁵⁷ These new molecular simulation protocols can be used for the better understanding of conformational dynamics of ligand-receptor complexes.

2.4.1 Statistical mechanics

The aim of a molecular dynamics simulation is to infer the macroscopic properties of a system from studies of its microscopic behaviour. For example, calculation of changes in the binding free energy of a particular drug candidate or the study of the energetics and mechanisms of conformational change.⁵⁸ Statistical mechanics make a connection between microscopic and macroscopic properties by providing the mathematical expressions that relate macroscopic properties to the distribution and motion of the atoms and molecules of the N body system.⁵⁸ Therefore statistical mechanics is the branch of physical sciences which deals with the study of macroscopic systems from molecular point of view.

2.4.2 Thermodynamic state, microscopic state and ensembles

Thermodynamic state of a system can be defined by a set of parameters, the temperature, T , the pressure, P , and the number of particles, N .⁵⁸ The microscopic state for each atom in a system of N particles is can be defined by the atomic positions, \mathbf{r}^N , and momenta \mathbf{p}^N ; these can also be considered as coordinates in a multidimensional space also known as phase space. For a system of N particles, this space has $6N$ dimensions.⁵⁸ A single point in phase space, denoted by Γ , describes the state of the system $\Gamma = (\mathbf{p}^N, \mathbf{r}^N)$.

An ensemble is a collection of all possible systems which have different microscopic states but an identical macroscopic or thermodynamic state.⁵⁸ Therefore, it is a collection of points in phase space satisfying the conditions of a particular thermodynamic state.⁵⁸ Microcanonical ensemble (NVE) is characterized by a fixed number of atoms, N , a constant volume, V , and a constant energy, E . This ensemble represents an isolated system.⁵⁸ Another important ensemble known as canonical ensemble (NVT) is a collection of all microstates whose thermodynamic state is characterized by a constant number of atoms, N , a constant volume, V , and a constant temperature, T . In this ensemble temperature is regulated by thermostats.⁵⁸ Isobaric-Isothermal Ensemble (NPT) ensemble consists of a fixed number of atoms, N , a constant, P , and a constant temperature, T where temperature and pressure are regulated by thermostats and barostats respectively.⁵⁹⁻⁶⁰

2.4.3 Calculation of averages from molecular dynamics simulations

In statistical mechanics, macroscopic quantities are defined as averages over ensembles of microstates.⁵⁸ The ensemble average of a property A is given by equation 2.4 :

$$\langle A \rangle_{ensemble} = \iint d\mathbf{p}^N d\mathbf{r}^N A(\mathbf{p}^N, \mathbf{r}^N) \rho(\mathbf{p}^N, \mathbf{r}^N) \quad (2.4)$$

where $A(\mathbf{p}^N, \mathbf{r}^N)$ is the observable of interest and it is expressed as a function of the momenta, \mathbf{p} , and the positions, \mathbf{r} , of the system, ρ is the probability density.⁵⁸

The probability density of finding the ensemble with momenta, \mathbf{p} , and the positions, \mathbf{r} is calculated as:

$$\rho(\mathbf{p}^N, \mathbf{r}^N) = \frac{1}{Q} \exp[-H(\mathbf{p}^N, \mathbf{r}^N)/k_B T] \quad (2.5)$$

where H is the Hamiltonian, T is the temperature, k_B is Boltzmann's constant and Q is the partition function given by equation 2.6.⁵⁸

$$Q = \iint d\mathbf{p}^N d\mathbf{r}^N \exp[-H(\mathbf{p}^N, \mathbf{r}^N)/k_B T] \quad (2.6)$$

In a molecular dynamics simulation, the points in the ensemble are calculated sequentially in time; therefore, to calculate an ensemble average, the molecular dynamics simulations must pass through all possible states corresponding to the particular thermodynamic constraints. Alternatively, it is possible to determine a time average of A , which is expressed in equation 2.7:

$$\langle A \rangle_{time} = \lim_{n \rightarrow \infty} \frac{1}{n} \int_{t=0}^{\tau} A(\mathbf{p}^N(t), \mathbf{r}^N(t)) dt \approx \frac{1}{M} \sum_{t=1}^M A(\mathbf{p}^N(t), \mathbf{r}^N(t)) \quad (2.7)$$

where t is the simulation time, M is the number of time steps in the simulation and $A(\mathbf{p}^N(t), \mathbf{r}^N(t))$ is the instantaneous value of A . n also represents the number of observations (time steps) that goes to infinity.⁵⁸ The experimental observables are assumed to be ensemble averages. This leads a fundamental principle of statistical mechanics, called ergodic hypothesis, which states that an ensemble average of an observable is equivalent to the time average of an observable.⁵⁸

$$\langle A \rangle_{ensemble} = \langle A \rangle_{time}$$

The general idea is that if system is allowed to evolve in time indefinitely, it will eventually pass through all possible states. Therefore, it is advisable to generate enough representative conformations during MD simulations.

2.4.4 Classical mechanics

The molecular dynamics simulation method is based on Newton's second law which is also known as the equation of motion (equation 2.8).^{58,61}

$$\mathbf{F}_i = m_i \mathbf{a}_i \quad (2.8)$$

Where \mathbf{F} is the force acting on a particle i , m_i is its mass and \mathbf{a}_i is acceleration of the particle. It is possible to determine the acceleration of each atom in the system from the knowledge of the force on each atom. The equations of motion are integrated to yield a trajectory that describes the positions, velocities and accelerations of the particles.^{58,61} The trajectory generated is averaged to determine the properties. Once the positions and velocity of atoms are known, the state of the system can be computed at any time (future or past).^{58,61} The force can be expressed as negative of the gradient of the potential energy (equation 2.7).⁵⁸

$$\mathbf{F}_i = -\nabla_i V \quad (2.9)$$

The above equations (2.6 and 2.7) can be combined to rewrite as follows:

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \quad (2.10)$$

where V is the potential energy of the system. Newton's equation of motion can relate the derivative of the potential energy to the changes in position as a function of time. In MD

simulations the initial distribution of velocities is usually computed from a random Maxwell-Boltzmann distribution with the magnitudes conforming to the required temperature and corrected so that overall momentum (\mathbf{P}) is zero.⁵⁸

$$\mathbf{P} = \sum_{i=1}^N m_i \mathbf{v}_i = 0 \quad (2.11)$$

The velocities, \mathbf{v}_i are often chosen randomly from a Maxwell-Boltzmann distribution at a given temperature (equation 2.12).⁵⁸

$$p(\mathbf{v}_{ix}) = \sqrt{\frac{m_i}{2\pi k_B T}} \exp\left[-\frac{1}{2} \frac{m_i v_{ix}^2}{k_B T}\right] \quad (2.12)$$

Where p is the probability of atom i which has mass m_i , a velocity \mathbf{v}_i in the x direction at a temperature T .

2.4.5 Integration algorithms

The MD trajectories describe the time evolution of the system in phase space which is defined by both position and velocity vectors. Therefore, integrators are used for the propagation of the positions and velocities with a finite time interval. Various numerical algorithms have been developed for integrating the equations of motion.⁵⁸ There is no analytical solution to the equations of motion because of the complexity of the function. Therefore, numerical methods have been developed for integrating the equations of motion. The integration algorithms work on the assumption that positions, velocities and acceleration can be approximated by a Taylor series.⁵⁸ For instance the Verlet algorithm uses position and acceleration at time t and positions from time $t-dt$ to calculate new positions at time $t+dt$ (equation 2.13, 2.14).⁵⁸

$$\mathbf{r}(t + dt) = \mathbf{r}(t) + \mathbf{v}(t)dt + \frac{1}{2} \mathbf{a}(t)dt^2 \quad (2.13)$$

$$\mathbf{r}(t - dt) = \mathbf{r}(t) - \mathbf{v}(t)dt + \frac{1}{2}\mathbf{a}(t)dt^2 \quad (2.14)$$

combining equations 2.13 and 2.14, the following equation can be obtained:

$$\mathbf{r}(t + dt) = 2\mathbf{r}(t) - \mathbf{r}(t - dt) + \mathbf{a}(t)dt^2 \quad (2.15)$$

where \mathbf{r} is the position, \mathbf{v} is the velocity, \mathbf{a} is the acceleration (second derivative with respect to time). The disadvantages of this method are that the algorithm is not self-starting because estimate is required for the initial position and the results are also of moderate precision. Therefore, an alternative algorithm called velocity Verlet algorithm that uses the velocity to yield positions, velocities, and acceleration is more frequently used (equation 2.16 and 2.17).⁵⁸

$$\mathbf{r}(t + dt) = \mathbf{r}(t) + \mathbf{v}(t)dt + \frac{1}{2}\mathbf{a}(t)dt^2 \quad (2.16)$$

$$\mathbf{v}(t + dt) = \mathbf{v}(t) + \frac{1}{2}[\mathbf{a}(t) + \mathbf{a}(t + dt)]dt \quad (2.17)$$

Similarly, more sophisticated algorithms include additional terms. Few examples of integration algorithm include leap-frog algorithm, Beeman's algorithm.⁵⁸

2.4.6 Constant temperature dynamics

Molecular dynamics is naturally performed in the constant microcanonical (NVE) ensemble because Newtonian mechanics conserve the energy (E). To perform molecular dynamics (MD) in the canonical ensemble (NVT) or Isobaric-Isothermal Ensemble (NPT), a thermostat is introduced to modulate the temperature of a system in some fashion. Various methods are available to add and remove energy from the boundaries of an MD system to approximate the ensemble. The goal of introducing thermostats is to ensure the desired average temperature of a system.

The temperature of the system is related to the time average of kinetic energy (K) which can be represented by equation 2.18:

$$\langle K \rangle_{NVT} = \frac{3}{2} N k_B T \quad (2.18)$$

the most intuitive strategy to keep the temperature constant would be to multiply at every step the velocity of the particles by a scaling factor (λ) given in equation 2.19.⁶⁰

$$\lambda = \sqrt{\frac{T_{required}}{T_{current}}} \quad (2.19)$$

Where $T_{current}$ is the current temperature and $T_{required}$ is the desired temperature. This however suppresses fluctuations in instantaneous temperature and does not lead to simulation of a correct canonical ensemble. An alternative way to maintain the temperature is to couple the system to an external heat bath that is fixed at desired temperature.⁶² The bath supplies or removes the heat from the system as required. The velocities are scaled at each step so that the rate of change of temperature is proportional to the difference in temperature between the bath and the system. Scaling factor for the velocities is given by equation 2.20:

$$\lambda^2 = 1 + \frac{\delta t}{\tau} \left(\frac{T_{bath}}{T_{current}} - 1 \right) \quad (2.20)$$

where τ is coupling parameter which determines how tightly is the system and bath coupled together, T_{bath} is bath temperature and δt is the time step.⁶² This simple method, known as Berendsen thermostat does not generate a rigorous canonical ensemble.⁶² Therefore alternative stochastic collision and extended system methods are usually implemented for this purpose. In the stochastic collision method, a particle is randomly chosen at intervals and its velocity is reassigned by random selection from Maxwell-Boltzmann distribution.⁶³ The method represents the system in contact with heat bath emitting thermal particles which

collide with atoms in the system. The system is simulated at constant energy while each collision takes place which overall represents a number of microcanonical simulations at slightly different energies. In this approach, momentum transfer is destroyed because of the random velocities therefore it is unadvisable to use this method to compute diffusion coefficients. Similarly, a trajectory generated with this method is not smooth which might be another disadvantage.

Other methods for performing constant temperature molecular dynamics are called extended system methods.^{59,64} In this method a thermal reservoir is introduced in the system to provide additional degree of freedom (s). The reservoir has potential energy (P) given by equation 2.21:

$$P = (f + 1)k_B T \ln s \quad (2.21)$$

where f is the number of degree of freedom in physical system and T is the desired temperature.⁶⁴ The kinetic energy (K) of reservoir can be given as equation 2.22:

$$K = \left(\frac{Q}{2}\right) \left(\frac{ds}{dt}\right)^2 \quad (2.22)$$

where Q is the fictitious mass of the extra degree of freedom. The magnitude of Q also determines the coupling between reservoir and real system.⁶⁴

2.4.7 Constant pressure dynamics

A macroscopic system can maintain constant pressure by changing its volume. In isothermal-isobaric ensemble (NPT), the constant pressure is maintained by changing the volume of simulation cell. The amount of volume fluctuations depends on the isothermal compressibility. The isothermal compressibility (k) is related to mean square volume (V) displacement given by equation 2.23:

$$k = \frac{1}{k_B T} \frac{\langle V^2 \rangle - \langle V \rangle^2}{\langle V^2 \rangle} \quad (2.23)$$

most of the methods used for the control of pressure are analogous to those used for the constant temperature. The rate of change of pressure is given by equation 2.24:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_p} (P_{bath} - P(t)) \quad (2.24)$$

where τ_p is the coupling constant, P_{bath} is the pressure of the bath and $P(t)$ is the actual pressure at time t . The volume of simulation box is scaled by factor (λ) (equation 2.25) which scales the atomic coordinates by a factor $\lambda^{1/3}$.⁵⁸

$$\lambda = 1 - k \frac{\delta t}{\tau_p} (P - P_{bath}) \quad (2.25)$$

where k is the isothermal compressibility and P is the required pressure. Thus the new position (\mathbf{r}) would be given by equation 2.26:

$$\mathbf{r} = \lambda^{\frac{1}{3}} \mathbf{r}_i \quad (2.26)$$

where \mathbf{r}_i is initial position. In extended pressure-coupling system methods, an extra degree of freedom is added which corresponds to the volume of box.^{59,64} The kinetic energy associated with this degree of freedom behaves like a piston acting on the system. The kinetic energy (k_p) of piston can be given by equation 2.27:

$$k_p = \frac{1}{2} Q \left(\frac{dV}{dt} \right)^2 \quad (2.27)$$

where Q is the mass of the piston and V is the volume of the system.^{59,64} The potential energy of the piston is given as PV , where P is the desired pressure and V is volume of the system. Piston with small mass gives rapid oscillations, while a large mass shows opposite effect.⁵⁸

2.4.8 Molecular mechanics

Energy calculation is based on either ab initio, semi-empirical quantum chemistry calculations or empirical methods. Although the ab initio description using quantum mechanics is more accurate, the limited computational power restricts its use only to very small systems containing up to a few hundred atoms. Molecular mechanics provides a faster method to determine the energy of the system which depends solely on the derived set of parameters (force-field) for the potential energy estimation.⁵⁸ The potential energy calculated using molecular mechanics usually consists of bonding terms (bond lengths, angles and dihedral angles) and non-bonding terms (van der Waals and electrostatic interactions) given by equation 2.28:

$$U = \sum_{bonds} \frac{1}{2} K_r (r - r_{eq})^2 + \sum_{angles} \frac{1}{2} K_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2} [1 + \cos(n\theta - \gamma_{eq})] + \sum_{i < j} 4\epsilon_{ij} \left[\left(\frac{A_{ij}}{r_{ij}} \right)^{12} - \left(\frac{B_{ij}}{r_{ij}} \right)^6 \right] + \sum_{i < j} \frac{1}{4\pi\epsilon} \frac{q_i q_j}{r_{ij}} \quad (2.28)$$

where U is the total potential energy and K_r , K_θ , and V_n are the force constants for bond-stretching, angle bending and dihedral angle deformations respectively.⁵⁸ Similarly, r , θ , and φ represents the bond-length, valence and dihedral angle values respectively; r_{eq} , θ_{eq} , and γ_{eq} are the equilibrium values of bond-length, angles and phase angle respectively; ϵ_{ij} is the depth of the potential well, A_{ij} and B_{ij} are the finite distances at which the inter-particle potential is zero; ϵ is dielectric constant; q_i and q_j are charges of atoms i and j , and r_{ij} is the distance between them. Molecular mechanics force fields can be used to calculate the potential energy of large biological systems with millions of atoms. The total energy of the system is calculated as the sum of overall interaction terms.⁵⁸ In larger systems the calculation of energy is a time consuming process, therefore, interactions between atoms separated by more than a specific cut-off distance are ignored.⁶² The cut-off of distance for non-bonded terms can result

discontinuities in the potential energies. Therefore, to resolve this problem, the non-bonded terms may be multiplied by a switching function.^{58,61} There are other methods like particle mesh Ewald (PME) to deal with long range electrostatics where cut-off potential is not set to zero. This method is particularly important for the MD simulation of nucleic acids, which have polyanionic backbones.⁶¹ Most of the MD simulations are currently performed in explicit solvent conditions that involve the use of different water models. The examples of these models include TIP3P, TIP4P, TIP5P and SPC.⁶⁵⁻⁶⁷ All the models make use of a rigid geometry however there is a difference in the number of interaction points used to represent the water molecules. For instance, TIP3P and SPC use three interaction points that represent the three atoms of the molecule. Likewise, TIP4P and TIP5P have four and five interaction points respectively, by involving one or two dummy atoms with negative charge on them. The dummy atoms represent the lone pairs of oxygen. The additional interaction sites usually improve the electrostatic distribution around the molecule but at the same time also enhance the computational cost. Therefore, three-site models are the most widely used in the MD simulation studies. In addition to these models, implicit solvation model can also be used to reduce the computational cost. In this model, the average action of water molecules is represented by a potential that gives equivalent properties. Hence, explicit solvent coordinates are not used in the model. Although this model is as not accurate as the explicit solvent model but it can be useful for the MD simulation studies of huge systems.

2.4.9 Periodic boundary conditions

Periodic boundary conditions are used for approximating a large system. About half of the total atoms arranged in a box are on the outer faces in the simulation box have less neighbouring atoms than atoms inside. This will have a large effect on the measured properties. Therefore, surrounding the simulation box with replicas of each cell may help to

overcome this problem.⁶² When the potential range is not too long, the principle of minimum image convention can be adopted which states that each atom interacts with the nearest atom or image in the periodic array.⁶¹ When an atom leaves the basic simulation box during simulation, incoming image can be considered for calculations. To calculate the particle interactions within the cut off range real and image neighbours are included.

2.5 Analysis of water thermodynamics using Grid Inhomogeneous Solvation

Theory (GIST)

Ordered water molecules on protein surfaces are frequently observed in the X-ray crystal structures. When a macromolecule binds another macromolecule or a drug-like small molecule, the displacement of surface water to bulk can have significant contribution to the overall free energy of binding.⁶⁸⁻⁷³ Therefore, the study of water at molecular surfaces is crucial in molecular recognition and ligand binding. The water molecules can act as H-bond acceptors or donors and also establish favourable van der Waals contacts with polar and apolar parts of the surface. These strong interactions make the displacement of these water molecules more challenging by the drug molecules. Conversely, the hydrophobic surfaces have unfavourable effects, due to combination of entropic and enthalpy costs, and thus allow the displacement of water. The water-protein interactions in a binding pocket display a pattern which is strongly replicated by the binding of small molecules.⁷⁴ Hence, these hydration sites in the binding pocket can provide valuable information about the key features that a molecule could mimic to favour its binding to the target. Grid Inhomogeneous Solvation Theory (GIST) is a computational method to calculate thermodynamic values of the water molecules occupying the binding pocket.⁷⁵ This approach discretizes the integrals of Inhomogeneous Solvation Theory (IST) onto a three-dimensional grid that includes the

binding site region involving high- and low-occupancy sites (**Figure 2.1**). Thermodynamic values of solvent are calculated within a defined region that involves boxes, or voxels, of a three-dimensional grid. The energy of each voxel represents the full interaction between the water molecules located in that voxels, with the solute (protein) and other water molecules

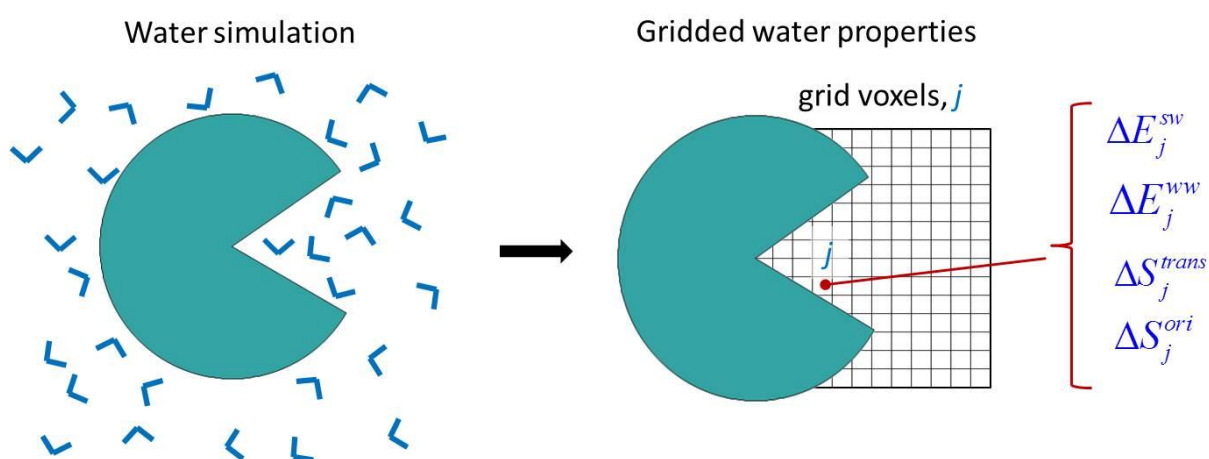


Figure 2.1 GIST calculates different thermodynamic properties of water using a grid-based approach (right panel) near the binding site region of protein or solute (green). The calculations are based on the molecular dynamics simulation results (left panel). Adapted from Ramsey and co-workers (2016).⁷⁵

in the region. GIST produces quantitative thermodynamic data for each grid box, or voxel. In the calculation of energy of voxels, the interaction between a water molecule and solute is assigned in its totality to the grid box while the water-water energies are divided by two, to avoid double counting. This approach provides a detailed analysis of water thermodynamic and structural quantities in a defined region of interest such as water occupancy within each voxel, water first order entropic penalty for each voxel etc. This information can be useful to understand the water thermodynamics and to know whether the water at a given site is thermodynamically favourable or not when compared to the bulk distribution. The tool helps to estimate a local, density-weighted free energy of solvation, $\Delta G(\mathbf{r}_j)$ for voxel (j),

$$\Delta G(\mathbf{r}_j) = \Delta E_{total}(\mathbf{r}_j) - T\Delta S_{sw}^{total}(\mathbf{r}_j) \quad (2.29)$$

$$\Delta E_{total}(\mathbf{r}_j) = \Delta E_{sw}(\mathbf{r}_j) + \Delta E_{ww}(\mathbf{r}_j) \quad (2.30)$$

$$\Delta S_{sw}^{total}(\mathbf{r}_j) = \Delta S_{sw}^{trans}(\mathbf{r}_j) + \Delta S_{sw}^{orient}(\mathbf{r}_j) \quad (2.31)$$

where r may be defined as the location of a water oxygen relative to the solute, T is absolute temperature. The total solute-water entropy (ΔS_{sw}^{total}) is derived using translational (ΔS_{sw}^{trans}) and orientational (ΔS_{sw}^{orient}) solute-water entropy terms. Likewise, for solvation energy (ΔE_{total}), ΔE_{sw} and ΔE_{ww} are the solute-water and water-water terms, respectively. Finally, the normalized (per water) free energy ($\Delta G^{R,norm}$) of region \mathbf{R} can be derived as

$$\Delta G^{R,norm} = \Delta E_{sw}^{R,norm} + \Delta E_{ww}^{R,norm} - T\Delta S_{sw}^{R,trans,norm} - T\Delta S_{sw}^{R,orient,norm} \quad (2.32)$$

Thus, by assembling the corresponding energy and entropy terms, change in solvation free energy after displacing the water from a site into bulk, and the normalized (per water) solvation free energy can be computed.

2.6 Recombinant protein (BC2L-C-nt) expression and purification

The biophysical evaluation of fragments or ligands requires sufficient production of the protein, here lectin of interest. To achieve this, the protein was expressed in recombinant form in heterologous system such as the bacteria *Escherichia coli*. The plasmidic construct was performed by Rafael Bermeo.⁷⁶ Primers on both 5' and 3' ends were designed using short single-strand DNA sequences that complement and delineate the genetic material to be amplified. Then amplification was performed using polymerase chain reaction (PCR) which can produce billions of copies of the gene. Subsequently, ligation of the genetic material is achieved by mixing with ligase that allows insertion of genetic material into an expression vector i.e pCold TF-TEV⁷⁷ in this study (**Figure 2.2**). The recombinant plasmid is then transformed into *E. coli* competent cells by applying heat shock at 42 °C and the colonies

bearing these plasmids are cultured for protein expression. Protein expression was performed in baffled culture flasks and induced with isopropyl β -D-1-thiogalactopyranoside (IPTG). The overnight incubation was required for the expression of the target protein with higher yield. The cspA (cold-shock protein A) promoter, upregulates the expression of the target protein on induction at low temperature while the expression of other proteins is hampered.

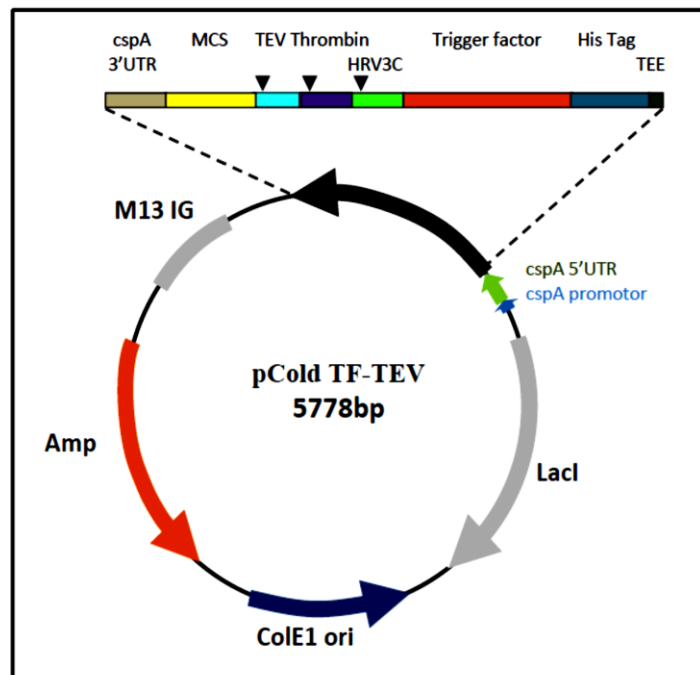


Figure 2.2 pCold TF-TEV Plasmid which is used for cloning and expression of BC2L-C-nt.

Likewise, trigger factor (TF) chaperone is expressed to enhance the solubility and improve the yield. The antibiotic ampicillin is added to ensure selective expression of the cells with acquired resistance due to the plasmid (Amp^R gene) and thus get the desired protein.

The protein expressed by the bacterial cells are released from the cytoplasm after cell disruption using high pressure. After centrifugation of the disrupted cells, the target protein released in the supernatant which is filtered using 0.45 μ M filter to discard the cell particles before the purification. To facilitate purification, the construct contains tobacco etch virus (TEV) cleavage site and histidine tag (HisTag) at the N-terminal. The His-tag allows separation of the fusion protein from the other bacterial proteins using immobilized metal affinity

chromatography (IMAC). The desired protein can be eluted with the help of a gradient of imidazole which disrupts the interactions of the proteins with the metal immobilized (nickel) on the column matrix. After elution, TEV protease is used to act on the TEV cleavage site to separate the HisTag from the target protein.⁷⁸ Further purification is performed by repeating IMAC to separate cleaved tag and the desired protein. Final step involves purification using size exclusion chromatography (SEC). This step separates the desired protein from the contaminants by size. In this method, smaller particles are trapped by the porous matrix in the column of SEC apparatus while the larger particles cannot enter the pores and passes quickly. Thus, smaller particles have longer retention time. This procedure is the first quality control of the recombinant lectin as is eluted as trimer in the SEC. In addition, electrophoresis performed using sodium dodecyl sulfate-polyacrylamide gel electrophoresis helps to evaluate the purity of the lectin.

2.7 Thermal shift assay (TSA)

The biophysical characterization of protein-ligand interactions plays an important role in structural biology. Several methods used to screen compound libraries are available ranging from highly specialized techniques such as NMR⁷⁹ and mass spectrometry⁸⁰ to simpler methods like thermal shift assays (TSA).⁸¹ TSA measures a shift in thermal denaturation temperature and hence stability of a protein under different conditions such as change in drug concentration, sequence mutation, buffer pH etc. This change in melting temperature is usually measured by fluorescence methods. These fluorescence-based techniques are also known as thermofluor assays or differential scanning fluorimetry (DSF).⁸¹⁻⁸³ TSA methods are easy to perform, hence most laboratories can routinely use them in high-throughput screening to identify new ligands.⁸⁴ The identified hits can be further characterized for bimolecular interactions, using more sophisticated biophysical techniques. In addition, these

methods can be used to determine the storing solution where the protein is the most stable and monodisperse to favour crystallization.^{83,85-88} In thermofluor assay, a compound such as SYPRO Orange dye, presents a low fluorescence signal in aqueous solution but emits high fluorescence⁸² when binding non-specifically to hydrophobic surfaces exposed in denatured proteins. When the protein unfolds upon heating, the hydrophobic core is exposed to the dye which results in stronger fluorescence signals. As the temperature increases, the protein becomes completely denatured. The stability curve and the temperature (T_m) at which half of the protein is unfolded is determined (**Figure 2.3**). The curves can be measured for protein in the absence or presence of ligand to calculate the difference in melting temperature (ΔT_m). The significant difference in melting temperature suggests interaction (binding) of ligand and protein. Usually, the assays work with most protein samples. However sometimes a clear signal is not observed due to several factor such

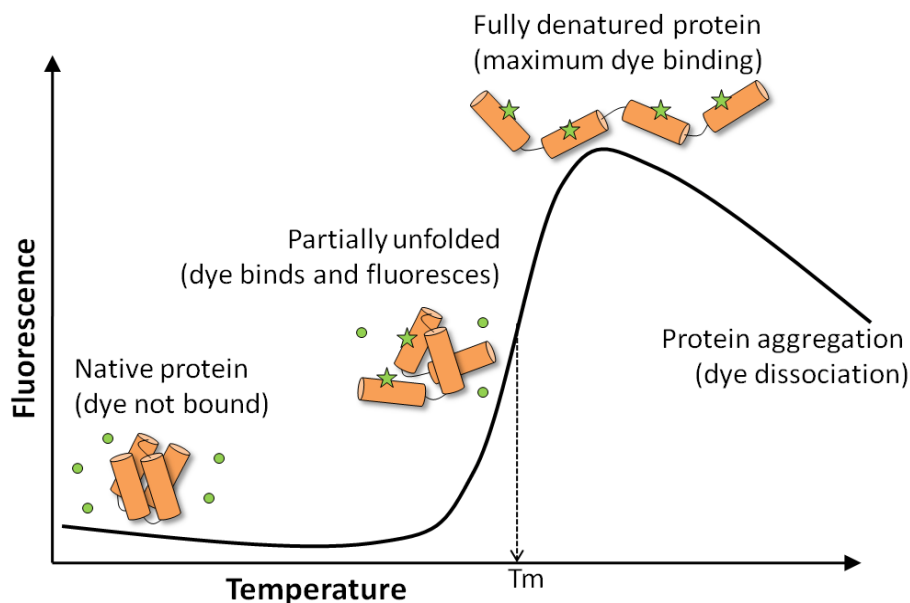


Figure 2.3 Principle of thermal shift assays (thermofluor). Fluorescent dye (SYPRO Orange) can be used to monitor the thermal denaturation of protein. The melting temperature (T_m) provides important information about the protein's thermal stability. Retrieved from <https://www.biotrend.com/en/other-products-186/glomelt-thermal-shift-protein-stability-514058784.html>

as high background noise because of fluorophore binding to native state of protein, insufficient hydrophobic core etc. In addition to measuring ligand binding, TSA is commonly used for determining buffer conditions, or additive nature, that would stabilize a protein and enhance the probability of crystallization.⁸⁹

2.8 Microscale thermophoresis (MST)

Thermophoresis can be and described as the directed motion of molecules through a temperature gradient.⁹⁰ Microscale thermophoresis (MST) is a biophysical method for the detection of directed movement of fluorescent molecules through microscopic temperature gradients in a very small volume (μl) of solution, that allows accurate analysis of binding events.⁹¹⁻⁹² The basic principle involves that the spatial temperature difference dT results into a change in molecule concentration in the region of elevated temperature which can be quantified by the Soret coefficient S_T . This can be represented as

$$c_{Ti} = c_i \exp(-S_{Ti}dT) \quad (2.33)$$

where c_i is the initial concentration, c_{Ti} is the concentration at the position where the temperature is increased by dT and i indicates the index for different states and types of molecules. A temperature difference dT causes the depletion of solvated biomolecules in the region of elevated temperature. This phenomenon of depletion depends on the interface between solvent and molecule. In general, thermophoresis depends on the charge, size, and solvation entropy of the molecules. The thermophoresis of a protein is usually different from the thermophoresis of a protein-ligand complex. This difference is observed due the ligand binding which results a change in size, charge and solvation energy.⁹³⁻⁹⁴ Even if one of these parameters is changed by ligand binding, a wide range of molecular interactions can be analysed.

MST is performed in thin capillaries (**Figure 2.4**) containing ligand and protein (with a fluorescent probe) as a free solution without immobilization. At the time of MST experiment, infrared laser is used to induce a microscopic temperature gradient. This results into temperature related intensity change (TRIC) of the fluorescent probe and also generates

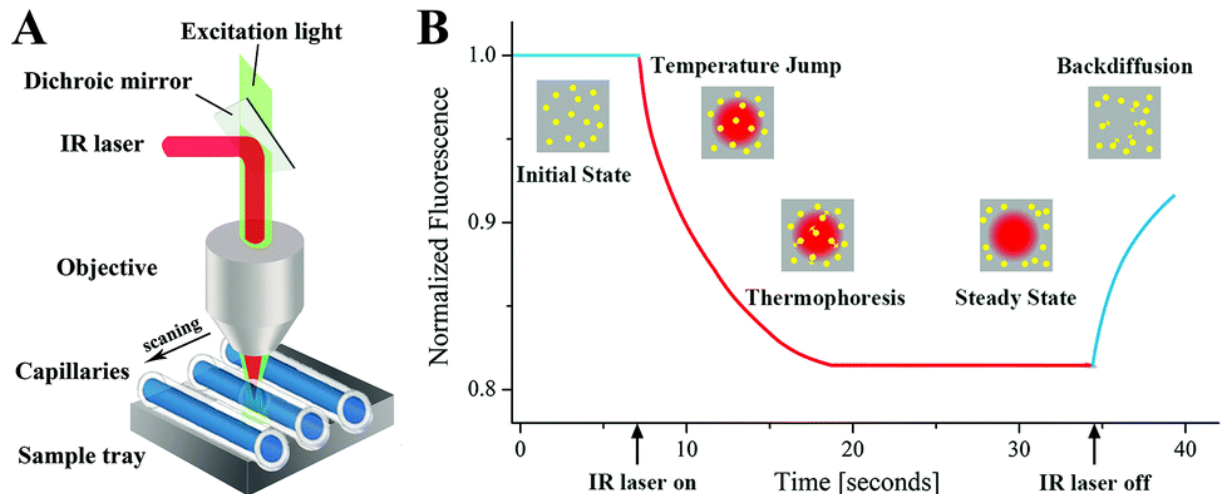


Figure 2.4 Principle of the MST and Experimental setup. (A) MST is performed using a capillary with a small volume (4 μL) of sample. An IR-Laser is used to heat the sample volume which is detected as fluorescence. (B) MST signal for a given capillary Initially shows the homogeneous distribution of molecules which can be measured as constant "initial fluorescence". Once IR-Laser is turned on, a temperature-jump is observed, followed by thermophoresis of molecules. Thereafter, a decrease in fluorescence is detected for about 30 seconds. Next, when the IR-Laser is turned off, a reversed temperature-jump is observed. Finally, "backdiffusion" of molecules is detected, which is driven by mass diffusion and gives information on the molecule size. Adapted from Liu and co-workers (2015).⁹⁵

thermophoresis. The binding events influence the fluorophore's microenvironment which ultimately have effect on the TRIC. Similarly, the movement of the molecule in the temperature gradient, depends on the parameters (such as charge, size, and solvation entropy of the molecules) that typically change upon interaction. Finally, a dose-response curve is obtained by plotting overall MST signal against the ligand concentration that can be further used to deduce binding affinity. Other than proteins-ligand binding, MST can also detect binding substrates to enzymes and ligands to liposomes.⁹⁴ Unlike some other methods

(such as surface plasmon resonance, SPR),⁹⁵ MST experiment is performed without surface immobilization and requires only small volume of sample. Further advancements in MST involves label-free approach that uses intrinsic fluorescence of tryptophan-containing proteins to follow their thermophoresis.⁹⁶

2.9 Saturation transfer difference (STD)-NMR

Among different screening techniques, nuclear magnetic resonance (NMR) is a powerful tool known to provide structural knowledge. In particular, saturation transfer difference NMR (STD NMR)⁹⁷ has emerged as a popular method due to its robustness and ease of implementation.⁹⁸ This method is generally used to investigate interactions of weak binding molecules with dissociation constants (K_d) in the millimolar to micromolar range. Thus, making it ideal for the identification of early lead molecules. In addition, STD NMR has been successfully used to investigate protein-carbohydrate interactions.⁹⁹⁻¹⁰⁰ The STD NMR experiment is based upon the principle of the nuclear Overhauser effect (NOE). This phenomenon involves the perturbation of signal intensity of a specific proton (for example) because of the cross-relaxation with another perturbed proton located nearby in space.¹⁰¹ In case of STD NMR, protons in the proteins can be selectively saturated by irradiating with radiofrequency. Subsequently, the saturation is rapidly transferred across the whole protein, in a process known as spin diffusion (**Figure 2.5**). When a ligand is present in the binding site of the protein, the saturation is transferred to the protons of the ligand. The ligand protons in closest contact, receive most of the saturation. Due to fast exchange, the ligand returns to solution and its signal is acquired. The signal produced by these protons helps to map the binding epitope of the ligand.

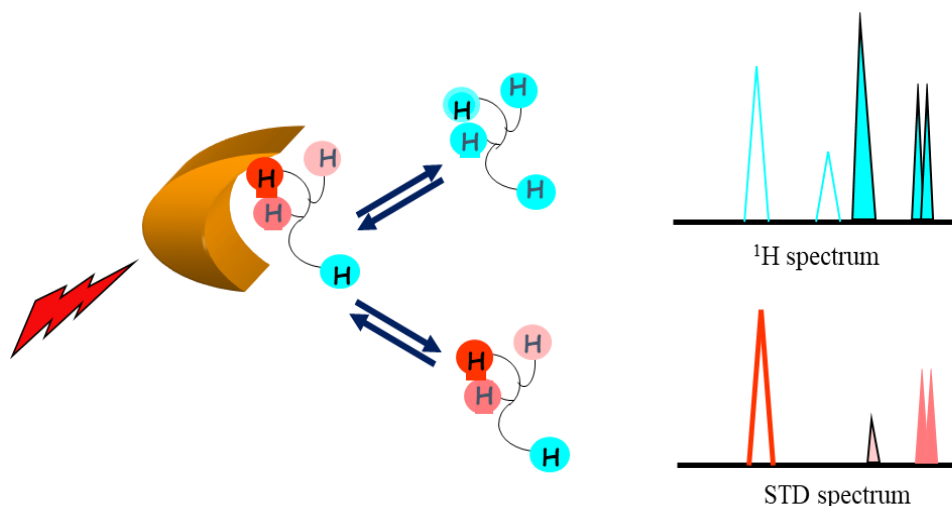


Figure 2.5 STD-NMR experiment. When the protein is selectively irradiated, the bound ligand receives the magnetization from protein: only the protons in intimate contact with receptor site show an STD signal in the NMR experiment. The differences between STD and reference (proton) spectra allow to map the interaction.

This method is easy to implement for ligand screening as the sample preparation involves a cocktail of ligand which can be used to identify hits simply by recording the STD NMR signals.¹⁰² Further, the different signals for the same compound can be used to construct a ligand binding epitope map, which may provide useful information to understand the relative position of the ligand in the binding site.¹⁰³ This method can be used to validate the results of docking studies by comparing STD NMR data with the predicted binding poses of the ligands.¹⁰⁴

2.10 X-ray crystallography

X-ray crystallography has become a powerful screening method in structure-based drug discovery that provides structural information on complexes to fuel the drug discovery process. From practical aspect, crystallization involves different necessary steps. In order to crystallize the protein, different crystallization conditions are screened which involve the identification of physical, chemical and biochemical conditions to initiate crystallization and

then systematic alteration and optimization of these initial conditions is required to further improve the crystallization process. In the next step, X-ray diffraction is used to collect the quality data that are further processed with the help of appropriate mathematic and informatics tools to solve the crystal structure. Finally, the structure is validated and deposited in the Protein Data Bank.¹⁰⁵

2.10.1 Crystallogenesi

Crystallization of macromolecules is largely empirical since the number of variables is very high, and requires numerous trials to obtain crystals. Crystallization of any molecule including proteins, involves two important steps: nucleation and growth.¹⁰⁶ Nucleation represents a first-order phase transition by which molecules pass from a disordered state to an ordered and thermodynamically stable state. This process ultimately yields small, ordered assemblies which are known as critical nuclei. They provide suitable surface for crystal growth which depends on the diffusion of particles to the surface of the nuclei and their ordered assembling onto the growing crystals.¹⁰⁷⁻¹⁰⁸ Both nucleation and growth, depend on supersaturation of the mother liquor giving rise to the crystals. When the concentration of a protein solution is increased beyond its solubility limit, the solution becomes supersaturated. The overall phenomenon of crystal growth can be described with the help of phase diagram (**Figure 2.6**) which illustrates three zones. The first zone with very high supersaturation is called “precipitation zone”, that leads to the formation of amorphous aggregates.¹⁰⁹ The next zone with intermediate supersaturation is called “labile zone” where growth and nucleation occur while the zone with lower supersaturation supports crystal growth only.¹⁰⁹⁻¹¹⁰ Thus creating a supersaturated protein solution is the immediate objective in growing protein crystals. This can be achieved by altering different parameters such as temperature, buffers, pH, precipitant nature: salt; polymers or organic compounds, addition of ions etc.¹⁰⁹ The

crystals of the proteins are usually grown in a physical apparatus which allows the alteration

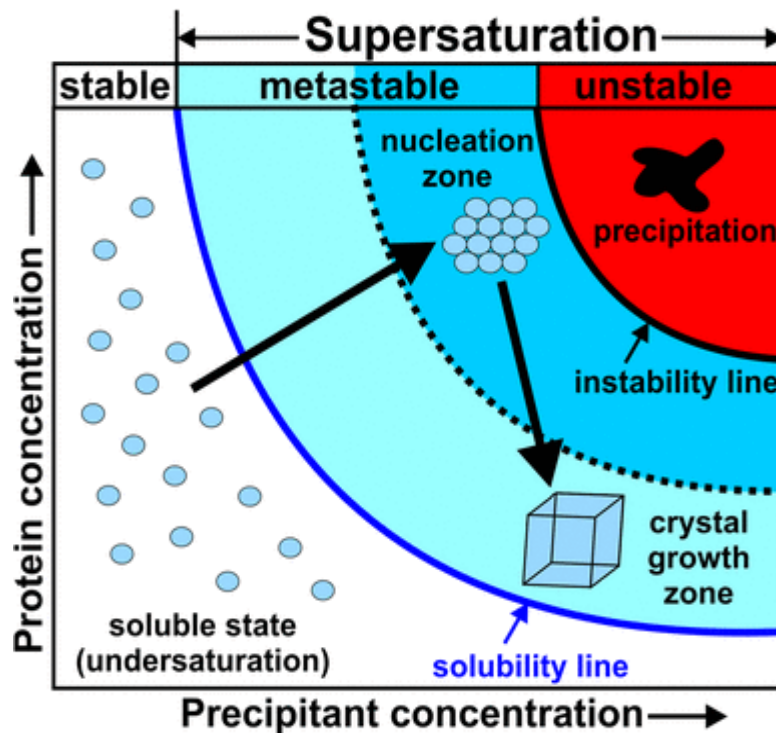


Figure 2.6 Protein crystallization phase diagram. Supersaturated solution is required for the formation of crystals. Nucleation zone supports the nucleation while the crystal growth occurs in the metastable zone. When a protein is undersaturated, none of the crystallization steps occur. Adapted from Bijelic and co-workers (2018).¹⁰⁶

of the properties of the mother liquor and the solubility of the protein by different approaches. One of the such approach is based on the vapor diffusion (**Figure 2.7**).¹⁰⁶

In this method, droplets containing purified protein under crystallization conditions are allowed to equilibrate in a large reservoir solution which generally contains similar precipitants and buffers in slightly higher concentrations. As the drop and reservoir equilibrate, the concentrations of precipitant and protein are increased in the drop that may result in crystal growth in the drop.¹¹¹ The drop can either hang from a cover slip (*hanging drop*) over the reservoir or allowed to sit at top on a small platform (*sitting drop*). The chamber is sealed and kept at a constant temperature to allow equilibration between the reservoir and the drop. After crystallization, single crystals are mounted in cryoloops after transfer in

cryoprotectant solution if necessary and flash-frozen in liquid nitrogen. This allows diffraction experiments to be done at 100K in order to limit the impact of radiation damages.

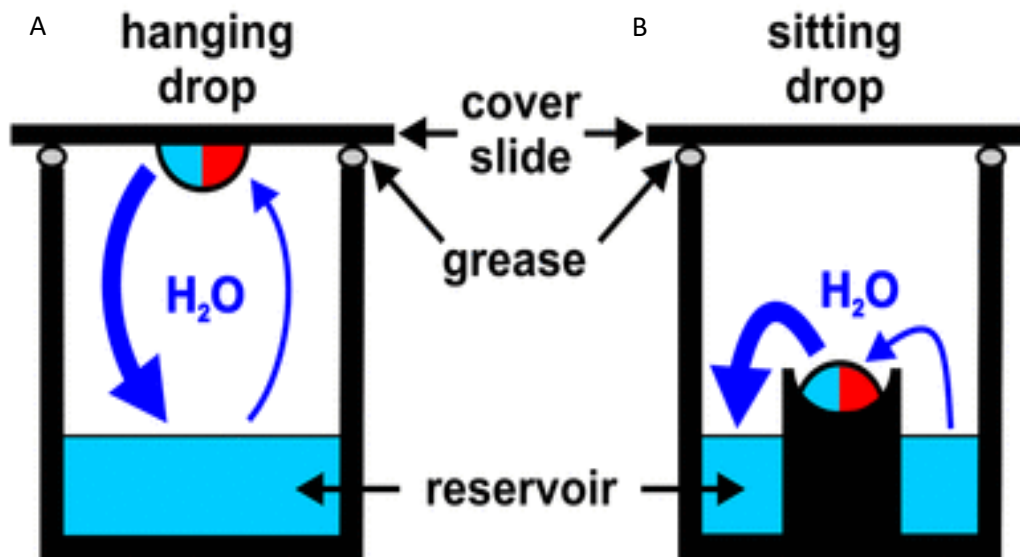


Figure 2.7 Vapor diffusion based crystallization techniques. Setup for hanging drop (A) and sitting drop (B) methods. The drop is placed over the slide which consists of the precipitant (light blue) and the protein solution (red). Finally, the wells are sealed to allow equilibration via vapor diffusion (dark blue arrows). Adapted from Bijelic and co-workers (2018).¹⁰⁶

2.10.2 Data processing, structure determination, and refinement

A protein crystal is composed of a highly ordered array of protein molecules that forms repeating units called unit cells (**Figure 2.8**). These unit cells are described by three vectors which have defined angles (α , β , γ) and lengths (a , b , c). Unit cell can build up the whole crystal by applying translational operations. Unit cells can be reduced to bare components called asymmetric units. These minimal arrangements, after applying symmetry operations can generate a whole crystal (**Figure 2.8**). X-rays are electromagnetic waves with wavelengths in the range between 0.1-100 Å. Since the interatomic distances also fall within this range (for example, C-C bond \sim 1.5 Å), X-rays can reveal the atomic detail of the proteins. In X-ray diffraction experiment, an incident X-ray beam is scattered by the electrons of the protein

crystal. When X-rays pass through a crystal, their interaction with the electrons induces oscillation in the crystal, causing electrons to emit partial waves. These partial waves then

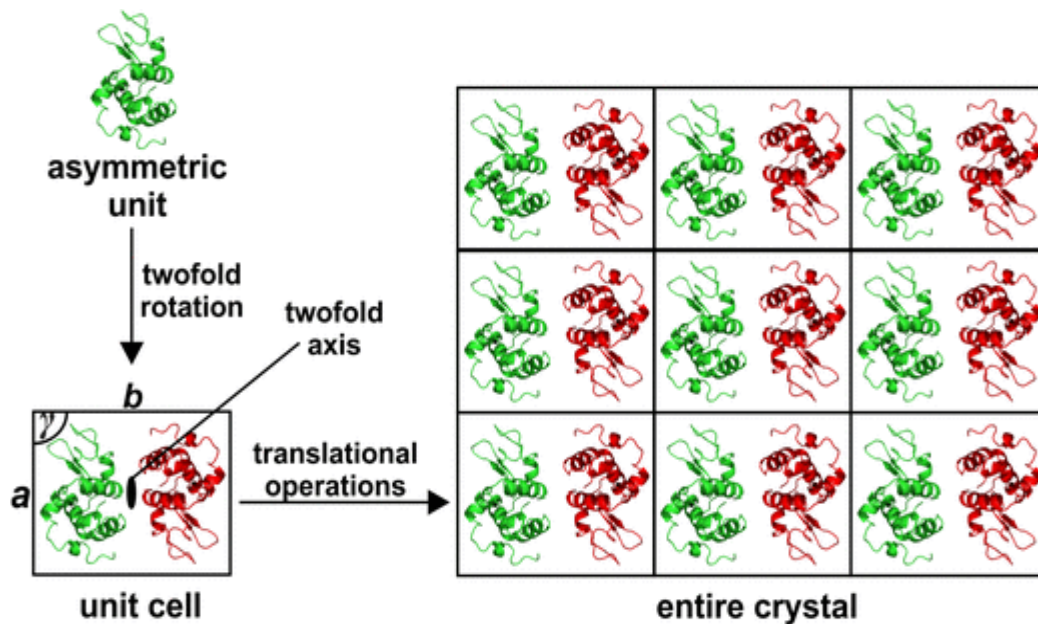
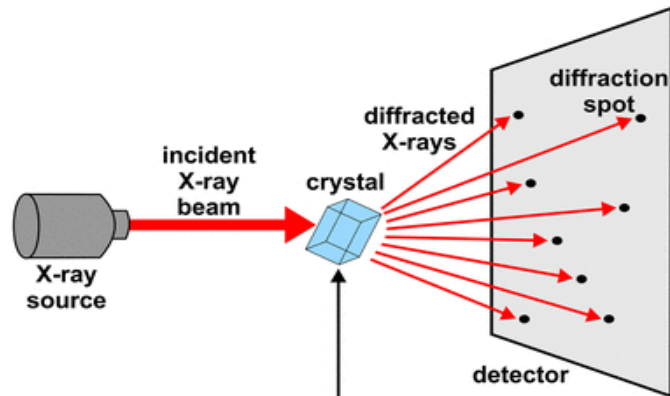


Figure 2.8 Assembly of a crystal structure (shown in 2D space). The asymmetric unit consists of one protein (green) molecule used to generate the unit cell by producing a symmetry mate (red molecule). A whole crystal can be built by translationally stacking up unit cells (shown in 2D space for clarity but exist as 3D in the real crystal). Adapted from Bijalic and co-workers (2018).¹⁰⁶

superimpose constructively (in phase) in certain directions and generate “reflections”, which are recorded on the detector (**Figure 2.9**). The probability of observing diffraction in a particular direction depends on the amplitude of the resulting wave (structure factor F). The X-ray diffraction event can be explained with the help of Bragg’s law. According to this, crystals consist of Bragg’s planes that are separated by the interplanar distance d . When incident X-ray wave of the wavelength λ hits a lattice plane at an angle θ followed by reflection at the same angle, the constructive interference is possible, only if the path difference ($d \sin\theta$) between the waves is an integer multiple of the wavelength λ . This can be written as:

$$n\lambda = 2d\sin\theta \quad (2.34)$$

A



B

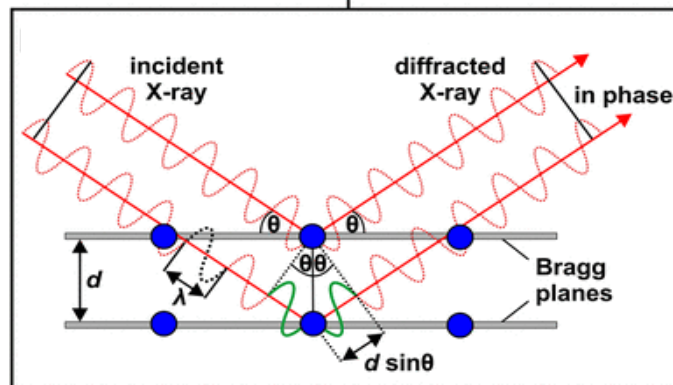


Figure 2.9 (A) Basic setup of an X-ray diffraction experiment. An incident X-ray beam after passing the crystal produces a diffraction pattern in the forms of dots and recorded on a detector. (B) A constructive interference is only observed when the path difference ($d \sin \theta$) between the waves (red dotted waves) is an integer multiple of the wavelength λ (green solid wave). Thus, the amplitudes of the diffracted waves add up to generate a signal on the detector. The two Bragg planes are shown as grey bars with an interplanar distance d and lattice points (blue dots), which represent atoms of the protein. Adapted from Bijalic and co-workers (2018).¹⁰⁶

For all the incident X-rays which do not exhibit angles that fulfil Bragg's Law, the resulting scattered waves will show destructive interference and thus no reflection will be observed.¹⁰⁶

Since protein contains numerous atoms, the crystal need to be rotated during data collection to be able to measure complete data (all necessary reflections) and the number of degrees to be collected depend on crystal orientation and symmetry.

Following the data collection, the processing of the diffraction data is done using well established algorithms available in many software packages and program suites. These tools

process data and calculate an electron density map. During data processing, the first step is to accurately determine crystal system and unit cell. In addition, orientation of the crystal in the beam is also determined and then indexing can be carried out.¹¹² In this process, an index is assigned to each reflection on the image, represented as three integers: h , k , and l . Then, reflection intensities are measured using various software tools leading to computer file containing the measured intensity and index of each reflection.¹¹³ The intensity is determined from the amplitude of the diffracted waves incident on the detector and by the phase difference which is expressed as an angle, between them. Amplitudes can be calculated directly from the intensities but information on the phases are lost and needs to be calculated from indirect ways that involves methods like molecular replacement and isomorphous replacement.^{106,114} Once the amplitudes and phases are obtained, structure factors can be calculated using the fast Fourier transform (FFT) method.¹¹⁵ This will generate an electron density map and in the form of the three dimensional contours which can be used to build the protein structure. The quality of the electron density map may be further improved by refinement and model building. The final coordinates for the structure are validated against geometry rules and quality fit and then deposited to the Protein Data Bank (PDB) repository.¹¹⁶

2.11 Isothermal titration calorimetry (ITC)

Isothermal titration calorimetry (ITC) is a technique to determine the thermodynamic parameters of binding between two molecules in solution.¹¹⁷⁻¹¹⁸ ITC determines the heat released or absorbed as the result of interaction between molecules, depending if of exothermic or endothermic character. The equation of heat exchange can be used to determine binding affinity. This can be defined as:

$$Q = V_o \Delta H_b [M]_t \left\{ \frac{K_a [L]}{1 + K_a [L]} \right\} \quad (2.35)$$

where Q is the heat evolved or absorbed, V_o is volume of sample cell, $[M]_t$ is the total concentration of protein in sample cell, ΔH_b represents the enthalpy of binding per mole of ligand, $[L]$ corresponds to the concentration of free ligand and K_a is the binding constant. Furthermore, the heat exchange corresponds to the enthalpy of binding (ΔH), and with K_a being related to the free energy of binding ΔG , it is possible to determine also the entropy ΔS , giving access to the thermodynamic parameters of the interaction.

$$-RT \ln K_a = \Delta G = \Delta H - T\Delta S \quad (2.36)$$

where R is the gas constant and T is the absolute temperature. The dissociation constant (K_d) can be directly obtained as the inverse of the binding constant (K_a).

The instrument consists of two cells which are enclosed in an adiabatic jacket. One of the cell is called sample cell where the compounds to be studied are placed while the other cell is used as reference cell, which is meant only for the buffer used to dissolve the sample (**Figure 2.10**). In ITC experiment, ligand is usually titrated into the sample cell containing macromolecule (protein). To maintain the isotherm, the apparatus provides higher (or lower) electric power to the system depending on the interactions. The heat change during the titration is calculated by integration of the power over the time (seconds). Therefore, heat released or absorbed during the complete titration corresponds to the fraction of bound ligand. As the titration proceeds, the protein gets saturated with the increased concentration of ligand and less heat is released or consumed. In order to obtain the reliable results using ITC, it is important to use the appropriate concentration of protein and ligand which relies on the parameter “ c ” value, which can be defined as:

$$c = N \times K_a \times [P_t] \quad (2.37)$$

where K_a is binding affinity constant (M^{-1}), P_t is sample cell concentration (M) and N is stoichiometry. The c value determines the shape of the binding curve. A higher c value (~ 1000)

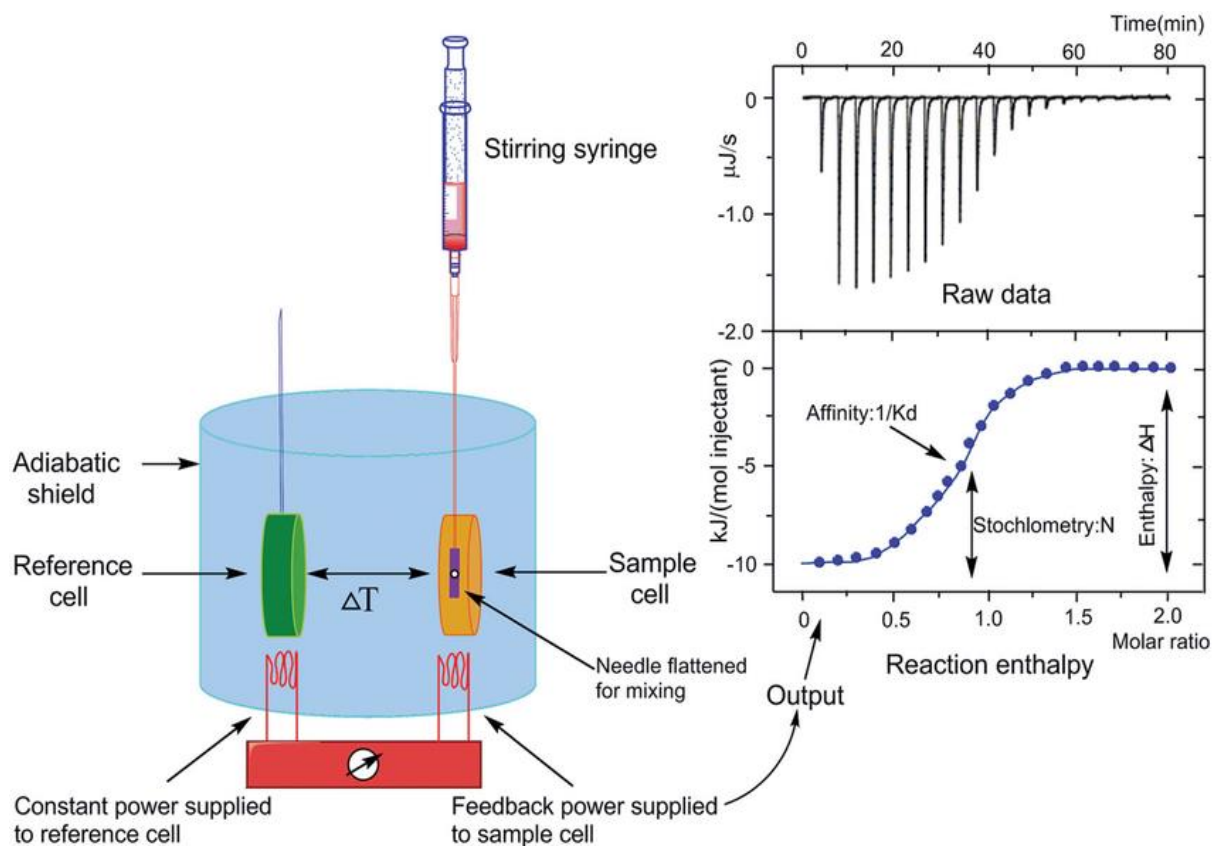


Figure 2.10 Schematic representation of the isothermal titration calorimeter (left) and annotated thermogram (upper right) with its evaluation (lower right). Adapted from Song and co-workers (2015).¹¹⁸

will result a very steep curve, hence make it difficult to estimate the K_d value. Similarly, if the c value is <5 , then shape of the binding curve will be too shallow. This will not allow accurate estimation of thermodynamic parameters. Therefore, in order to get a good sigmoidal shape of the binding curve, that will allow estimation of K_d and thermodynamic parameters, it is recommended to keep the c value between 20 and 100.

2.12 References

- [1] Ballester, P. J. *et al.*, Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification, *J. R. Soc. Interface* **2012**, *9*, 3196-3207.
- [2] Roberts, N. A. *et al.*, Rational design of peptide-based HIV proteinase inhibitors, *Science* **1990**, *248*, 358-361.
- [3] Erickson, J. *et al.*, Design, activity, and 2.8 Å crystal structure of a C₂ symmetric inhibitor complexed to HIV-1 protease, *Science* **1990**, *249*, 527-533.
- [4] Bambini, S. Rappuoli, R., The use of genomics in microbial vaccine development, *Drug Discov. Today* **2009**, *14*, 252-260.
- [5] Lundstrom, K., *Genomics and drug discovery*, *Future Med Chem.* **2011**, *15*, 1855-1858.
- [6] Wang, R. *et al.*, The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures, *J. Med. Chem.* **2004**, *47*, 2977-2980.
- [7] Mountain, V., Astex, Structural Genomix, and Syrrx. I can see clearly now: structural biology and drug discovery, *Chem. Biol.* **2003**, *10*, 95-98.
- [8] Nayal, M. Honig, B., On the nature of cavities on protein surfaces: application to the identification of drug-binding sites, *Proteins* **2006**, *63*, 892-906.
- [9] Xie, Z. R. Hwang, M. J., Methods for predicting protein-ligand binding sites, *Methods Mol. Biol.* **2015**, 383-398.
- [10] Halgren, T. A., Identifying and characterizing binding sites and assessing druggability, *J. Chem. Inf. Model.* **2009**, *49*, 377-389.
- [11] Schrödinger Release 2018-1: Maestro, Schrodinger, LLC, New York, NY, **2018**.
- [12] Friesner, R. A. *et al.*, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, *J. Med. Chem.* **2004**, *47*, 1739-1749.
- [13] Silakari, O. Singh, P. K. *Chapter 6 - Molecular docking analysis: Basic technique to predict drug-receptor interactions*, Eds.: O. Silakari and P. K. Singh), *Academic Press*, **2021**, 131-155.
- [14] Xu, M. Lill, M. A., Induced fit docking, and the use of QM/MM methods in docking, *Drug discovery today. Technologies* **2013**, *10*, e411-e418.
- [15] Grinter, S. Zou, X., Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design, *Molecules* **2014**, *19*, 10150-10176.
- [16] Audie, J. Swanson, J., Recent work in the development and application of protein-peptide docking, *Future Med. Chem.* **2012**, *4*, 1619-1644.
- [17] Grosdidier, A. *et al.*, SwissDock, a protein-small molecule docking web service based on EADock DSS, *Nucleic Acids Res* **2011**, *39*, 29.

- [18] Röthlisberger, D. *et al.*, Kemp elimination catalysts by computational enzyme design, *Nature* **2008**, 453, 190-195.
- [19] Cai, J. *et al.*, Peptide deformylase is a potential target for anti-Helicobacter pylori drugs: reverse docking, enzymatic assay, and X-ray crystallography validation, *Protein Sci.* **2006**, 15, 2071-2081.
- [20] Dominguez, C. *et al.*, HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information, *J. Am. Chem. Soc.* **2003**, 125, 1731-1737.
- [21] Halgren, T. A. *et al.*, Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening, *J. Med. Chem.* **2004**, 47, 1750-1759.
- [22] Pagadala, N. S. *et al.*, Software for molecular docking: a review, *Biophys. Rev.* **2017**, 9, 91-102.
- [23] Chen, Y.-C., Beware of docking!, *Trends Pharmacol. Sci.* **2015**, 36, 78-95.
- [24] Pantsar, T. Poso, A., Binding Affinity via Docking: Fact and Fiction, *Molecules* **2018**, 23, 1899.
- [25] Pinzi, L. Rastelli, G., Molecular Docking: Shifting Paradigms in Drug Discovery, *Int. J. Mol. Sci.* **2019**, 20, 4331.
- [26] Vajda, S. *et al.*, Sampling and scoring: a marriage made in heaven, *Proteins* **2013**, 81, 1874-1884.
- [27] Guedes, I. A. *et al.*, Receptor-ligand molecular docking, *Biophys. Rev.* **2014**, 6, 75-87.
- [28] Ferreira, L. G. *et al.*, Molecular docking and structure-based drug design strategies, *Molecules* **2015**, 20, 13384-13421.
- [29] Changeux, J.-P. Edelstein, S., Conformational selection or induced fit? 50 years of debate resolved, *F1000 Biol. Rep.* **2011**, 3, 19-19.
- [30] Teodoro, M. L. Kavraki, L. E., Conformational flexibility models for the receptor in structure based drug design, *Curr. Pharm. Des.* **2003**, 9, 1635-1648.
- [31] Jain, A. N., Scoring functions for protein-ligand docking, *Curr Protein Pept Sci* **2006**, 7, 407-420.
- [32] Kitchen, D. B. *et al.*, Docking and scoring in virtual screening for drug discovery: methods and applications, *Nat. Rev. Drug Discov.* **2004**, 3, 935-949.
- [33] Meng, X. Y. *et al.*, Molecular docking: a powerful approach for structure-based drug discovery, *Curr. Comput. Aided Drug Des.* **2011**, 7, 146-157.
- [34] Böhm, H. J., Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs, *J. Comput. Aided Mol. Des.* **1998**, 12, 309-323.
- [35] Gehlhaar, D. K. *et al.*, Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming, *Chem. Biol.* **1995**, 2, 317-324.
- [36] Head, R. D. *et al.*, VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands, *J. Am. Chem. Soc.* **1996**, 118, 3959-3969.

- [37] Jain, A. N., Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities, *J. Comput. Aided Mol. Des.* **1996**, *10*, 427-440.
- [38] Verkhivker, G. M. *et al.*, Deciphering common failures in molecular docking of ligand-protein complexes, *J. Comput. Aided Mol. Des.* **2000**, *14*, 731-751.
- [39] Feher, M., Consensus scoring for protein-ligand interactions, *Drug Discov. Today* **2006**, *11*, 421-428.
- [40] Kastiris, P. L. Bonvin, A. M. J. J., Molecular origins of binding affinity: seeking the Archimedean point, *Curr. Opin. Struct. Biol.* **2013**, *23*, 868-877.
- [41] Forli, S. Olson, A. J., A Force Field with Discrete Displaceable Waters and Desolvation Entropy for Hydrated Ligand Docking, *J. Med. Chem.* **2012**, *55*, 623-638.
- [42] Rarey, M. *et al.*, The particle concept: placing discrete water molecules during protein-ligand docking predictions, *Proteins* **1999**, *34*, 17-28.
- [43] van Dijk, A. D. J. Bonvin, A. M. J. J., Solvated docking: introducing water into the modelling of biomolecular complexes, *Bioinformatics* **2006**, *22*, 2340-2347.
- [44] Verdonk, M. L. *et al.*, Modeling water molecules in protein-ligand docking using GOLD, *J. Med. Chem.* **2005**, *48*, 6504-6515.
- [45] Huang, S. Y. Zou, X., Advances and challenges in protein-ligand docking, *Int. J. Mol. Sci.* **2010**, *11*, 3016-3034.
- [46] Liu, H.-Y. Zou, X., Electrostatics of Ligand Binding: Parametrization of the Generalized Born Model and Comparison with the Poisson-Boltzmann Approach, *The Journal of Physical Chemistry B* **2006**, *110*, 9304-9313.
- [47] Thompson, D. C. *et al.*, Investigation of MM-PBSA rescoring of docking poses, *J. Chem. Inf. Model.* **2008**, *48*, 1081-1091.
- [48] Bradshaw, R. T. *et al.*, Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction, *Protein Eng. Des. Sel.* **2011**, *24*, 197-207.
- [49] Huo, S. *et al.*, Computational alanine scanning of the 1:1 human growth hormone-receptor complex, *J. Comput. Chem.* **2002**, *23*, 15-27.
- [50] Friesner, R. A. *et al.*, Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes, *J. Med. Chem.* **2006**, *49*, 6177-6196.
- [51] Eldridge, M. D. *et al.*, Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes, *J. Comput. Aided Mol. Des.* **1997**, *11*, 425-445.
- [52] Alder, B. J. Wainwright, T. E., Phase Transition for a Hard Sphere System, *J. Chem. Phys.* **1957**, *27*, 1208.

- [53] Rahman, A., Correlations in the Motion of Atoms in Liquid Argon, *Phys. Rev.* **1964**, *136*, A405-A411.
- [54] Stillinger, F. H. Rahman, A., Improved simulation of liquid water by molecular dynamics, *J. Chem. Phys.* **1974**, *60*, 1545-1557.
- [55] McCammon, J. A. *et al.*, Dynamics of folded proteins, *Nature* **1977**, *267*, 585-590.
- [56] Dror, R. O. *et al.*, Biomolecular simulation: a computational microscope for molecular biology, *Annu. Rev. Biophys.* **2012**, *41*, 429-452.
- [57] Michel, J., Current and emerging opportunities for molecular simulations in structure-based drug design, *Phys. Chem. Chem. Phys.* **2014**, *16*, 4465-4477.
- [58] Leach, A. R., *Molecular Modelling, Principles and Applications*, Longman: Harlow, UK, **1996**.
- [59] Hoover, W. G., Canonical dynamics: Equilibrium phase-space distributions, *Phys. Rev. A* **1985**, *31*, 1695-1697.
- [60] Woodcock, L. V., Isothermal molecular dynamics calculations for liquid salts, *Chem. Phys. Lett.* **1971**, *10*, 257-261.
- [61] Allen, M. P. Tildesley, D. J., *Computer Simulation of Liquids: Second Edition*, Oxford University Press, Oxford, **2017**, 640.
- [62] Berendsen, H. J. C. *et al.*, Molecular dynamics with coupling to an external bath, *J. Chem. Phys.* **1984**, *81*, 3684-3690.
- [63] Andersen, H. C., Molecular dynamics simulations at constant pressure and/or temperature, *J. Chem. Phys.* **1980**, *72*, 2384-2393.
- [64] Nosé, S., A unified formulation of the constant temperature molecular dynamics methods, *J. Chem. Phys.* **1984**, *81*, 511-519.
- [65] Berendsen, H. J. C. *et al.* *Interaction Models for Water in Relation to Protein Hydration*, (Ed. B. Pullman), Springer Netherlands, Dordrecht, **1981**, 331-342.
- [66] Jorgensen, W. L. *et al.*, Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys.* **1983**, *79*, 926-935.
- [67] Mahoney, M. W. Jorgensen, W. L., A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions, *J. Chem. Phys.* **2000**, *112*, 8910-8922.
- [68] Abel, R. *et al.*, A displaced-solvent functional analysis of model hydrophobic enclosures, *J. Chem. Theory Comput.* **2010**, *6*, 2924-2934.
- [69] Bissantz, C. *et al.*, A Medicinal Chemist's Guide to Molecular Interactions, *J. Med. Chem.* **2010**, *53*, 5061-5084.

- [70] Poornima, C. S. Dean, P. M., Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions, *J. Comput. Aided Mol. Des.* **1995**, *9*, 500-512.
- [71] Riniker, S. *et al.*, Free enthalpies of replacing water molecules in protein binding pockets, *J. Comput. Aided Mol. Des.* **2012**, *26*, 1293-1309.
- [72] Wong, S. E. Lightstone, F. C., Accounting for water molecules in drug design, *Expert Opin. Drug Discov.* **2011**, *6*, 65-74.
- [73] Baron, R. *et al.*, Water in cavity-ligand recognition, *J. Am. Chem. Soc.* **2010**, *132*, 12091-12097.
- [74] Abel, R. *et al.*, Role of the active-site solvent in the thermodynamics of factor Xa ligand binding, *J. Am. Chem. Soc.* **2008**, *130*, 2817-2831.
- [75] Ramsey, S. *et al.*, Solvation thermodynamic mapping of molecular surfaces in AmberTools: GIST, *J. Comput. Chem.* **2016**, *37*, 2029-2037.
- [76] Bermeo, R. *et al.*, BC2L-C N-Terminal Lectin Domain Complexed with Histo Blood Group Oligosaccharides Provides New Structural Information, *Molecules* **2020**, *25*, 248.
- [77] Qing, G. *et al.*, Cold-shock induced high-yield protein production in Escherichia coli, *Nat. Biotechnol.* **2004**, *22*, 877-882.
- [78] Kapust, R. B. *et al.*, Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency, *Protein Eng.* **2001**, *14*, 993-1000.
- [79] Sillerud, L. O. Larson, R. S., Advances in nuclear magnetic resonance for drug discovery, *Methods Mol. Biol.* **2012**, *910*, 195-266.
- [80] Hofstadler, S. A. Sannes-Lowery, K. A., Applications of ESI-MS in drug discovery: interrogation of noncovalent complexes, *Nat. Rev. Drug Discov.* **2006**, *5*, 585-595.
- [81] Semisotnov, G. V. *et al.*, Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe, *Biopolymers* **1991**, *31*, 119-128.
- [82] Pantoliano, M. W. *et al.*, High-density miniaturized thermal shift assays as a general strategy for drug discovery, *J. Biomol. Screen.* **2001**, *6*, 429-440.
- [83] Vedadi, M. *et al.*, Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15835-15840.
- [84] Winter, A. *et al.*, Biophysical and computational fragment-based approaches to targeting protein-protein interactions: applications in structure-guided drug discovery, *Q. Rev. Biophys.* **2012**, *45*, 383-426.
- [85] Ericsson, U. B. *et al.*, Thermofluor-based high-throughput stability optimization of proteins for structural studies, *Anal. Biochem.* **2006**, *357*, 289-298.

- [86] Geders, T. W. *et al.*, Use of differential scanning fluorimetry to optimize the purification and crystallization of PLP-dependent enzymes, *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2012**, *68*, 596-600.
- [87] Nettleship, J. E. *et al.*, Methods for protein characterization by mass spectrometry, thermal shift (ThermoFluor) assay, and multiangle or static light scattering, *Methods Mol. Biol.* **2008**, *426*, 299-318.
- [88] Reinhard, L. *et al.*, Optimization of protein buffer cocktails using ThermoFluor, *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **2013**, *69*, 209-214.
- [89] Krintel, C. *et al.*, L-Asp is a useful tool in the purification of the ionotropic glutamate receptor A2 ligand-binding domain, *FEBS J.* **2014**, *281*, 2422-2430.
- [90] Jerabek-Willemsen, M. *et al.*, MicroScale Thermophoresis: Interaction analysis and beyond, *J. Mol. Struct.* **2014**, *1077*, 101-113.
- [91] Molecular Interaction Studies Using Microscale Thermophoresis, *Assay Drug Dev. Technol.* **2011**, *9*, 342-353.
- [92] Zillner, K. *et al.* *Microscale Thermophoresis as a Sensitive Method to Quantify Protein: Nucleic Acid Interactions in Solution*, (Eds.: M. Kaufmann and C. Klinger), Springer New York, NY, **2012**, 241-252.
- [93] Seidel, S. A. I. *et al.*, Microscale thermophoresis quantifies biomolecular interactions under previously challenging conditions, *Methods* **2013**, *59*, 301-315.
- [94] Liu, Y. *et al.*, Highly specific detection of thrombin using an aptamer-based suspension array and the interaction analysis via microscale thermophoresis, *Analyst* **2015**, *140*, 2762-2770.
- [95] Schuck, P., Use of surface plasmon resonance to probe the equilibrium and dynamic aspects of interactions between biological macromolecules, *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 541-566.
- [96] Seidel, S. A. I. *et al.*, Label-Free Microscale Thermophoresis Discriminates Sites and Affinity of Protein–Ligand Binding, *Angew. Chem. Int. Ed.* **2012**, *51*, 10656-10659.
- [97] Mayer, M. Meyer, B., Characterization of Ligand Binding by Saturation Transfer Difference NMR Spectroscopy, *Angew. Chem. Int. Ed.* **1999**, *38*, 1784-1788.
- [98] Viegas, A. *et al.*, Saturation-Transfer Difference (STD) NMR: A Simple and Fast Method for Ligand Screening and Characterization of Protein Binding, *J. Chem. Educ.* **2011**, *88*, 990-994.
- [99] Enríquez-Navas, P. M. *et al.*, A Solution NMR Study of the Interactions of Oligomannosides and the Anti-HIV-1 2G12 Antibody Reveals Distinct Binding Modes for Branched Ligands, *Chem. Eur. J.* **2011**, *17*, 1547-1560.
- [100] Muñoz-García, J. C. *et al.*, Langerin–Heparin Interaction: Two Binding Sites for Small and Large Ligands As Revealed by a Combination of NMR Spectroscopy and Cross-Linking Mapping Experiments, *J. Am. Chem. Soc.* **2015**, *137*, 4100-4110.

- [101] Hore, P. J., *Nuclear magnetic resonance*, Oxford University Press Inc, New York, **1995**.
- [102] Gao, J. *et al.*, Automated NMR Fragment Based Screening Identified a Novel Interface Blocker to the LARG/RhoA Complex, *PLoS One* **2014**, *9*, e88098.
- [103] Angulo, J. *et al.*, Saturation Transfer Difference (STD) NMR Spectroscopy Characterization of Dual Binding Mode of a Mannose Disaccharide to DC-SIGN, *ChemBioChem* **2008**, *9*, 2225-2227.
- [104] Krishna, N. R. Jayalakshmi, V. *Quantitative analysis of STD-NMR spectra of reversibly forming ligand-receptor complexes*, **2008**, *273*, 15-54.
- [105] Protein Data Bank: the single global archive for 3D macromolecular structure data, *Nucleic Acids Res.* **2019**, *47*, D520-D528.
- [106] Bijelic, A. Rompel, A., Polyoxometalates: more than a phasing tool in protein crystallography, *ChemTexts* **2018**, *4*, 10.
- [107] Malkin, A. J. *et al.*, Mechanisms of growth for protein and virus crystals, *Nat. Struct. Biol.* **1995**, *2*, 956-959.
- [108] McPherson, A. *et al.*, Atomic force microscopy in the study of macromolecular crystal growth, *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 361-410.
- [109] McPherson, A. Gavira, J. A., Introduction to protein crystallization, *Acta Crystallogr F Struct Biol Commun.* **2014**, *70*, 2-20.
- [110] *GloMelt Thermal Shift Protein Stability kit*, **2021**.
- [111] Rhodes, G., *Crystallography Made Crystal Clear*, Academic Press, **2006**.
- [112] Kabsch, W., Automatic indexing of rotation diffraction patterns, *J. Appl. Crystallogr.* **1988**, *21*, 67-72.
- [113] Otwinowski, Z. Minor, W., Processing of X-ray diffraction data collected in oscillation mode, *Methods Enzymol.* **1997**, *276*, 307-326.
- [114] The CCP4 suite: programs for protein crystallography, *Acta Crystallogr. D Biol. Crystallogr.* **1994**, *50*, 760-763.
- [115] Ten, L., Crystallographic fast Fourier transforms, *Acta Cryst.* **1973**, *29*, 183-191.
- [116] Berman, H. *et al.*, Announcing the worldwide Protein Data Bank, *Nat. Struct. Mol. Biol.* **2003**, *10*, 980-980.
- [117] Pierce, M. M. *et al.*, Isothermal Titration Calorimetry of Protein-Protein Interactions, *Methods* **1999**, *19*, 213-221.
- [118] Song, C. *et al.*, Choosing a suitable method for the identification of replication origins in microbial genomes, *Front. Microbiol.* **2015**, *6*, 1049.

3. Analysis of crystal structures and fragment screening in BC2L-C-nt

Understanding the structure of target protein and its binding sites is one of the most important steps in drug design. Therefore, accurate prediction of druggable binding sites is equally important at the initial stage of ligand design. The identified pockets can be used to design the ligands against the target protein. Moreover, these predictions can also provide insight into ligand-receptor interactions for lead optimization by suggesting effective strategies to improve receptor complementarity.

As discussed under the objectives of the project, this thesis work aims to design high-affinity glycomimetic antagonists against BC2L-C-nt using molecular modelling and biophysical methods. In this chapter, I will discuss the approaches used to identify druggable sites in the vicinity of fucose-binding site in the BC2L-C-nt. These sites were further employed to virtually screen a small fragment library. Later, the interaction of the identified fragments with a predicted site was confirmed using a set of biophysical techniques (discussed in the next chapter).

3.1 Analysis of the binding site in crystal structures

The X-ray crystal structure of the N-terminal domain of BC2L-C (BC2L-C-nt) in complex with methylseleno- α -L-fuco-pyranoside (MeSe- α -L-Fuc, PDB code 2WQ4)¹ is available at high resolution ($R = 1.42 \text{ \AA}$) in PDB.² The structure analysis shows that asymmetric unit contains three peptide chains and three carbohydrate ligands (MeSe- α -L-Fuc), organised around a 3-fold pseudo axis of symmetry (**Figure 3.1**). The three sugar binding sites are located at the interface between neighbouring protomers (A, B, C), and separated by a distance of $\sim 20 \text{ \AA}$ (**Figure 3.1**). Further analysis shows that the binding mode for the fucose ring is identical in

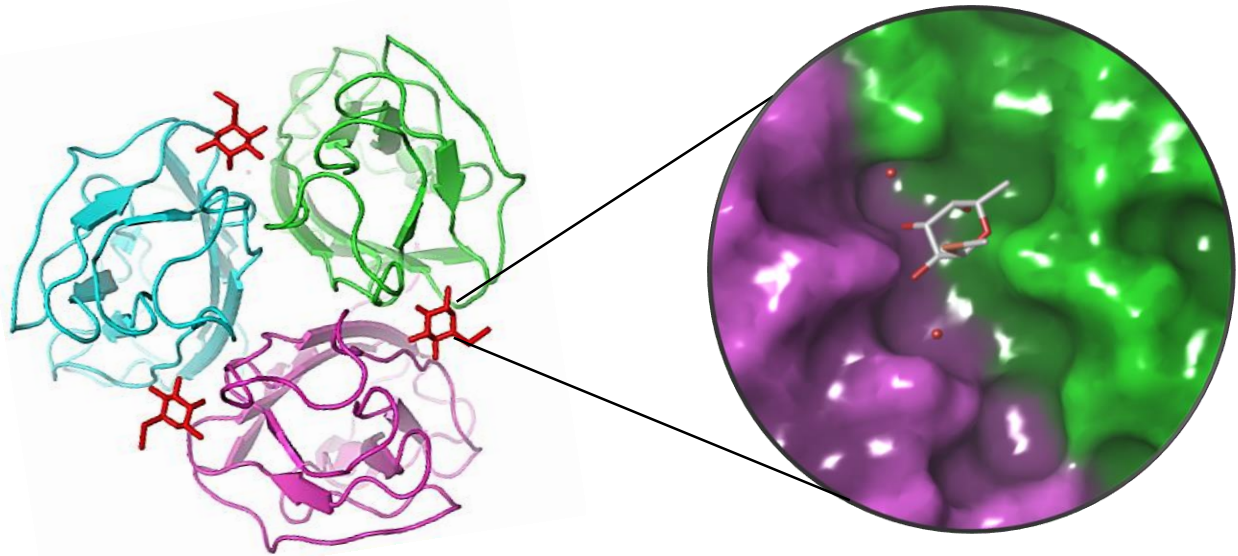


Figure 3.1 Crystal structure of BC2L-C N-terminal domain (PDB 2WQ4) showing three identical fucoside binding sites (top view) at the interface of monomers.

the three binding sites: the key residues Tyr48, Ser82, Thr83, Arg85 from one chain (e.g. chain A) and Tyr58, Thr74, Tyr75, Arg111 from the neighbouring chain (e.g. chain C) play an important role in ligand binding (**Figure 3.2**). In addition, two water molecules, W1 and W2, bridge the sugar and the protein. Both water molecules are conserved in the available X-ray structures of BC2L-C N-terminal domain in complex with fucoside and fucosylated

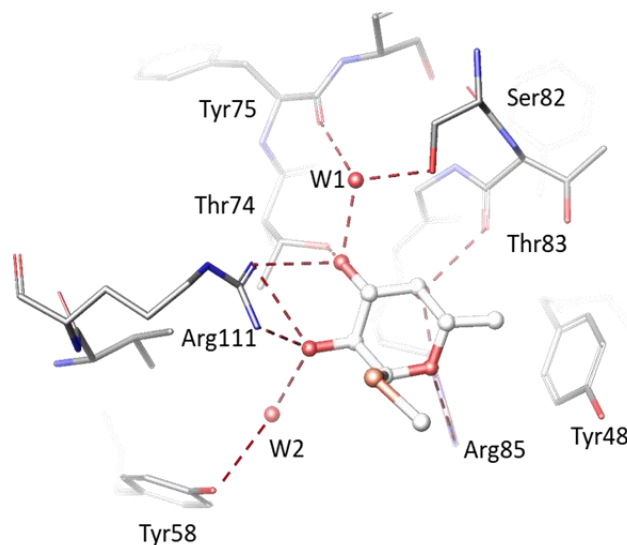


Figure 3.2 Binding site in BC2L-C-nt showing key residues and water molecules (W1 and W2) involved in fucose binding. Hydrogen bonds are represented as dashed lines.

oligosaccharides.^{1,3} In particular, the crystal structures of trimeric BC2L-C-nt complexed with H-type 1 and Globo-H oligosaccharides are available, as recently solved by Rafael Bermeo in the framework of his PhD thesis in the PhD4GlycoDrug network.³ The water molecule W1 is deeply buried in the binding site and sandwiched between the protein and the ligand, forming an H-bonding interaction with HO-3 atom of fucose (**Figure 3.2**). W2 is more exposed to the solvent and mediates an H-bonding interaction between HO-2 atom of fucose and the side chain of Tyr58. The region surrounding the fucose binding site appears interesting as it shows pockets which could be analysed for their druggability and relevance to host small molecules (fragments).

3.2 Resolution of first crystal structure of the apo form of BC2L-C-nt

All available crystal structures of the BC2L-C-nt so far were in the holo form , i.e. complexed with fucosides or fucosylated oligosaccharides. The crystal structures of the complexes displayed some promising pockets in the vicinity of the fucoside binding site. The investigation of such pockets in the apo-protein was needed to verify their occurrence and stability in both the forms. This may also confirm that these additional pockets were not merely formed as crevice due to the ligand (fucose/oligosaccharides) binding at the interface of protomers. Therefore, attempts were made to successfully crystallize the apo-protein

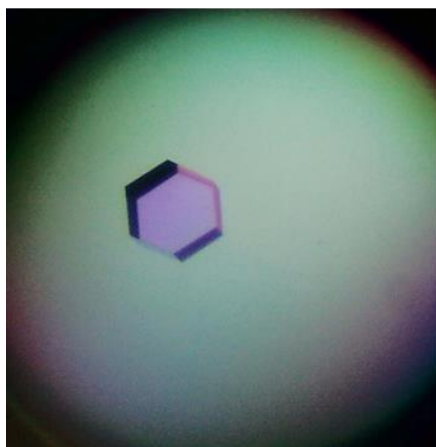


Figure 3.3 Crystals of apo form of BC2L-C-nt were obtained as stacked cluster.

(Figure 3.3). Crystals were obtained using 2 μ L hanging drops containing 50:50 (v/v) mix of protein and reservoir solution (1.2-1.4 M tri sodium citrate pH 7.0). X-ray data were collected at the beamline Proxima 1, synchrotron SOLEIL, Saint Aubin, France. The data for best crystal was collected at a high resolution (1.5 \AA). The structure was solved by molecular replacement using one monomer of PDB 2WQ4 as search model. The asymmetric unit contains one monomer in the $P6_3$ space group. Hence, crystal symmetry was applied to construct the trimer which was then compared with the other crystal structures available in holo form (**Figure 3.4.**

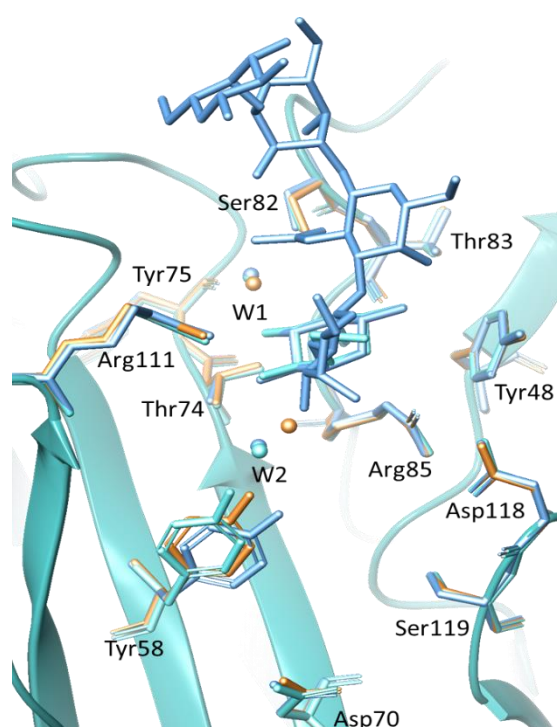


Figure 3.4 Superimposition of crystal structures of the apo (orange, PDB 7BFY) and holo forms (cyan PDB 2WQ4, azure PDB 6TIG) of BC2L-C N-terminal domain. The binding site residues do not display any significant difference.

The comparison with trimer complexes with MeSe- α -L-Fuc and Globo-H shows root-mean-square values of 0.21 \AA and 0.24 \AA , respectively for atoms (Ca of backbone). Similarly, the binding sites of both the forms did not display any significant differences in the orientation of residues in the fucose binding pockets and the surrounding regions. However, a minor difference at the surface loop (Asn52-Phe54) was observed (**Figure 3.5**). This loop shows

interactions with the methyl group of N-acetylgalactosamine (GalNAc) in the complex with Globo-H.³ Likewise, small differences in the conformation of residues at the N- and C-terminal were noticed due to the H-bond interactions between them (**Figure 3.5 B**). The minor

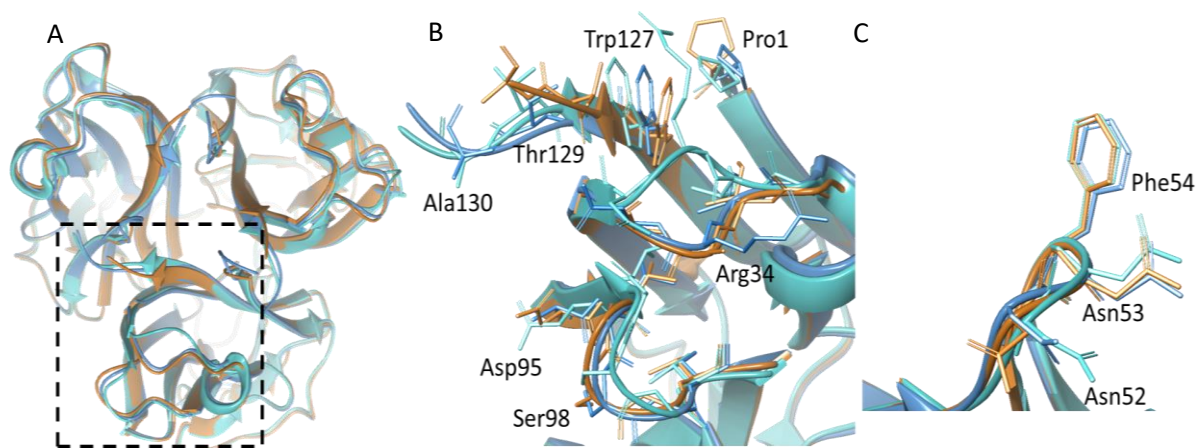


Figure 3.5 (A) Overview of the surface loops in the C-terminus of BC2L-C-nt. (B) Superimposition of different structures in holo (cyan, 2WQ4 and azure, 6TIG) and apo form (orange, 7BFY) shows different orientation of the residues in the C-terminus. (C) Only minor differences were observed in a small loop (Asn52-Phe54) located near the fucose binding site.

differences in the conformation of the termini also caused a small displacement (0.6 to 1.0 Å) of surface loops (Val28-Asp35, Asp95-Val100). Analysis of water molecules bridging the fucose to protein showed that W1 is conserved and deeply buried in all the crystal structures while W2 located in the more exposed region was displaced by 1.9 Å in the apo structure (**Figure 3.4**). The data collection and refinement details for the crystal structure are given in

Table 3.1.

These studies illustrated that the apo form of the trimer does not show any significant difference when aligned to the previously solved structures in complex with fucoside or oligosaccharides.^{1,3} This also indicates that the residues in the fucose binding site and surrounding pockets do not display strong flexibility and maintained the conformations of loops and side chains necessary for key interactions with the fucoside, even when bound to

the larger ligands like fucosylated oligosaccharides. The new crystal structure of BC2L-C-nt

Table 3.1 X-ray data collection and refinement for the apo form of BC2L-C-nt.

Data set	BC2L-C-nt apo form
Data Collection	
PDB code	7BFY
Beamline	PROXIMA1 (SOLEIL)
Wavelength (Å)	0.9801
Space Group	P6 ₃
a, b, c (Å)	42.99, 42.99, 94.68
α, β, γ (°)	90.0, 90.0, 120.0
Resolution (Å) ^a	19.97-1.50 (1.53-1.50)
Total observations	261006
Unique reflections	15915
Multiplicity ^a	16.4 (15.5)
Mean I/σ(I) ^a	21.7 (5.9)
Completeness (%) ^a	99.9 (100)
$R_{\text{merge}}^{\text{a,b}}$	0.081 (0.445)
R_{pim}	0.030 (0.274)
$CC_{\frac{1}{2}}^{\text{a,c}}$	0.999 (0.935)
Refinement	
Reflections: working/free ^d	15120 / 762
$R_{\text{work}} / R_{\text{free}}^{\text{e}}$	0.149 / 0.178
Ramachandran plot: allowed/favoured/outliers (%)	100 / 97 / 0
R.m.s. bond deviations (Å)	0.014
R.m.s. angle deviations (°)	1.855
R.m.s. chiral deviations	0.085
No. atoms / Mean B-factors (Å ²)	Chain A
Protein	952 / 15.6
carbohydrate ligand	-
ligand ^f	-
water	150 / 28.2

^a Values for the outer resolution shell are given in parentheses.

^b $R_{\text{merge}} = \sum_{\text{hkl}} \sum_i |I_i(\text{hkl}) - \langle I(\text{hkl}) \rangle| / \sum_{\text{hkl}} \sum_i I_i(\text{hkl})$.

^c $CC_{\frac{1}{2}}$ is the correlation coefficient between symmetry-related intensities taken from random halves of the dataset.

^d The data set was split into "working" and "free" sets consisting of 95 and 5% of the data, respectively. The free set was not used for refinement.

^e The R-factors R_{work} and R_{free} are calculated as follows: $R = \sum (|F_{\text{obs}} - F_{\text{calc}}|) / \sum |F_{\text{obs}}|$, where F_{obs} and F_{calc} are the observed and calculated structure factor amplitudes, respectively

^f refers to ligands bound in the active site and potential surface binding sites.

(apo form) confirms that the region in the vicinity of the fucose binding site could be interesting for drug (glycomimetic) design. The additional region does not display any significant difference in the binding site surfaces (**Figure 3.6**). Hence, the identified sites were explored for their druggability.

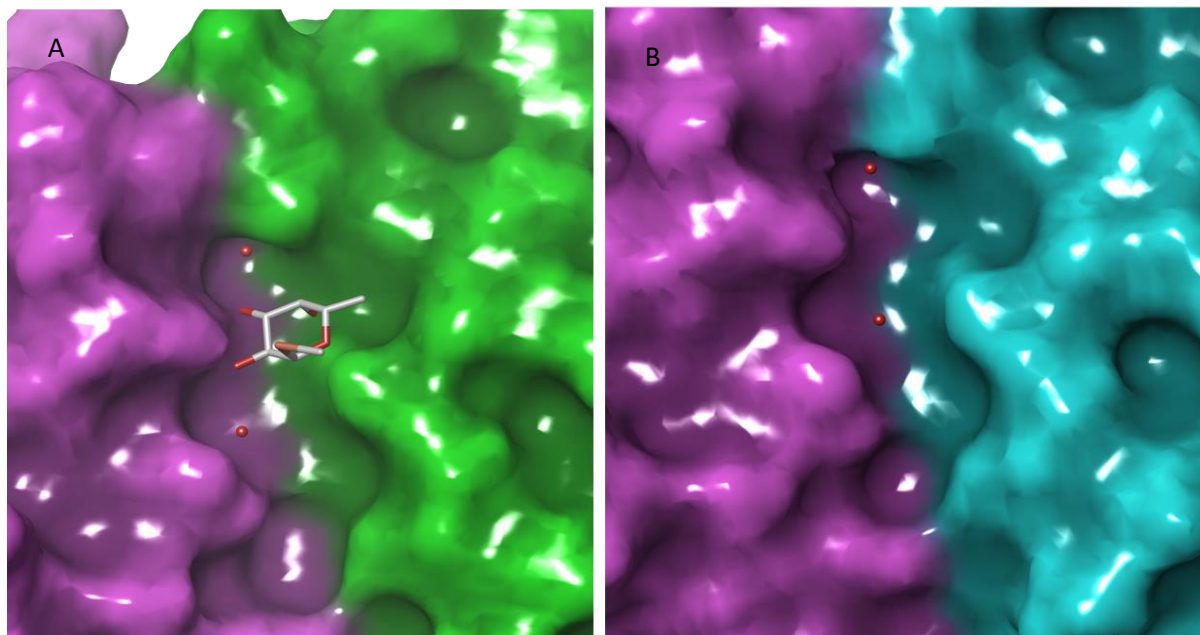


Figure 3.6 Comparison of binding site surfaces in holo and apo form of BC2L-C-nt. The additional pockets exist in both forms and do not exhibit any significant difference.

3.3 Prediction of druggable sites

After structural analysis of BC2LC-nt, the additional region was explored for the drugability using SiteMap⁴ tool. SiteMap creates a grid of points on the surface of protein and based on the depth, size, van der Waals interaction energy, hydrophilicity and hydrophobicity it determines the druggability of the region. Based on these characteristics, a single scoring function called SiteScore is assigned to the potential druggable regions that helps to prioritize the best one for further studies. To run the SiteMap calculations, the PDB structure (PDB 2WQ4) was prepared using Maestro.⁵ Crystallographic water molecules were removed from the structure to avoid their effect in binding site prediction. The protein preparation procedure involved adding hydrogen and missing atoms followed by pKa prediction for the

protein residues using the PROPKA⁶⁻⁸ method at pH 7.4. The protonation state was also assigned to the N(ϵ) atom of histidine (His116) residue. Finally, the protein was subjected to restrained minimization. The SiteMap calculations performed using the PDB structure identified three regions as potentially druggable sites in the vicinity of fucoside binding site which we labelled as X, Y and Z (**Figure 3.7**). Region Y consisting of residues Ser82, Thr83 and Phe54 in each monomer, corresponds to the area where larger, fucosylated oligosaccharides were observed to bind, including the recently described Globo H hexasaccharide and H-type 1 tetrasaccharide.³ In addition, two other regions (X and Z) appear promising to host the small fragments. The site X forms a deep cleft extending along the dimerization interface while the site Z is formed by the region between residues Val110 and Arg111. All the three sites are interesting to explore for druggability/ligandability. Therefore, docking protocol was setup involving all the (three) sites in order to identify suitable hits in the identified sites. Thereafter, best fragments could be connected to the fucose core to design high-affinity glycomimetic ligands.

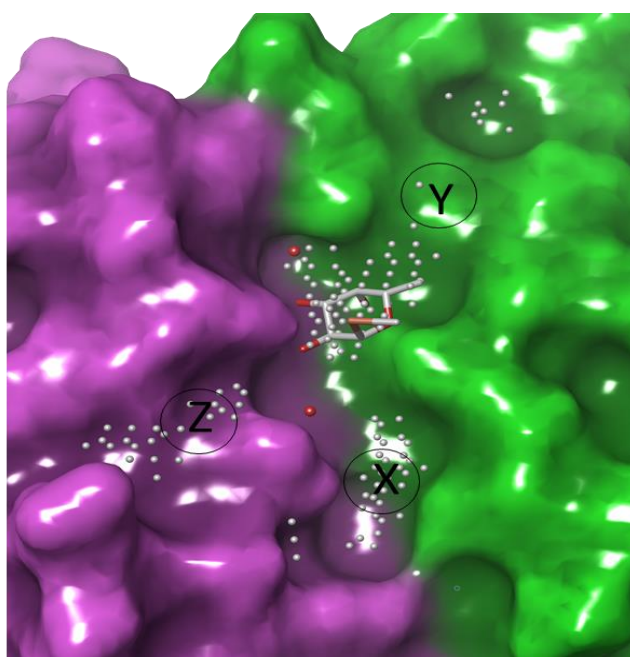


Figure 3.7 Identification of additional regions (site points) near fucoside binding site suitable for small fragment binding.

3.4 Virtual screening

After analysis of the druggable pockets, virtual screening was performed to identify the fragments (small molecules) binding at these sites. Finally, in order to design glycomimetic ligands, the best fragments could be connected to the fucose core by using a suitable chemical linker. All the calculations for virtual screening were performed using the Schrödinger Suite through Maestro (version 2018-1) graphical interface.⁵ The *in silico* protocol based on docking calculations is shown in **Figure 3.8**.

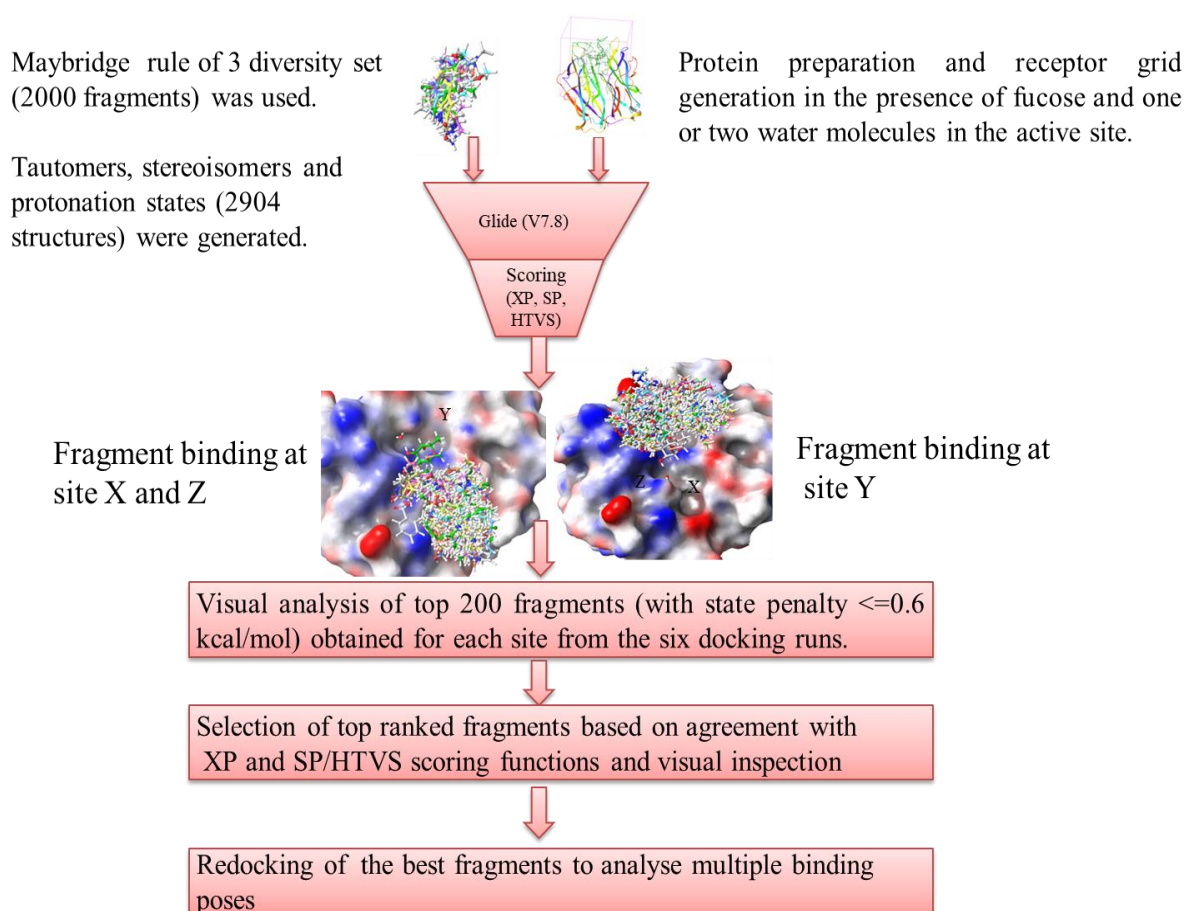


Figure 3.8 Schematic representation of the virtual screening protocol.

3.4.1 Ligand preparation

The Maybridge library of small fragments (rule of 3 diversity set with molecular weight < 300 , hydrogen bond donors/acceptors ≤ 3 , cLogP ≤ 3 , and rotatable bonds ≤ 3) containing 2000

fragments (available at <https://www.maybridge.com/>) was used for in silico screening. The small molecules (fragments) in the library allow efficient sampling of the chemical space compared to the molecules with greater complexity. This approach also provides room for ligand optimization which can result the final ligand with relatively low molecular weight.

The ligands were prepared for docking using LigPrep tool which generated tautomers, stereoisomers and protonation states at pH 7±2. The calculations resulted in 2904 structures useful for virtual screening.

3.4.2 Models for docking study

The coordinates of the complex (PDB 2WQ4)¹ were obtained from PDB database² and prepared using Maestro.⁵ Since the three binding sites are almost identical, only one binding site located between chains A and C was used for docking calculations. The two structural water molecules (W1 and W2) bridging fucose and protein were retained (**Figure 3.2**). The hydrogen atoms were added and pKa was predicted for protein residues using the PROPKA⁶⁻⁸ method at pH 7.4 and assigned protonation state to the N(ε) atom of histidine (His116) residue. Finally, protein-ligand complex was subjected to restrained minimization with convergence of heavy atoms to an RMSD of 0.3 Å using the OPLS3 force field.⁹

For docking studies, the residues from chain A (Tyr48, Ser82, Thr83, Arg85) and chain C (Tyr58, Thr74, Tyr75, Arg111) were selected as the center of a cubic grid box of size 32×32×32 to define the binding site region. The two important water molecules (W1 and W2) and ligand (MeSe-α-L-Fuc) were also retained as a part of protein during the grid generation. Another docking grid, was also generated using the same residues but retaining only one water (W1) molecule and the fucose. This docking model could identify some fragments, suitable to replace the water molecule and occupy the site. The docking studies were performed by employing both the grids (models) and XP (extra precision), SP (standard

precision) and HTVS (high throughput virtual screening) scoring functions. All the calculations were accomplished using the OPLS3 force field⁹ and the flexible docking approach in the Glide (Grid-based ligand docking with energetics) version 7.8.¹⁰

3.4.3 Docking analysis

Docking calculation were performed using both the docking models: using either one or two water molecules in the docking grid while retaining the fucose as part of protein. The results showed that the fragments were docked mainly in the region X and Y in both the models. The region Y is a shallow and exposed binding site which mostly hosted the hydrophobic fragments on the surface. Some of the ligands binding in this region show interactions with the key residues involved in the binding of fucosylated oligosaccharides.³ Likewise, region Z accommodated a few hydrophilic fragments in the narrow channel formed by the surrounding residues. The region X comparatively forms a deeper binding site where fragments appear to be nestling and generating some specific pattern of interactions. Therefore, our initial efforts were mainly focused on this region. Analysis of top 200 fragments was performed for the docking results obtained from 6 docking runs involving two types of docking models and three (XP, SP and HTVS) scoring functions.¹⁰ HTVS and SP use the same scoring function, however HTVS protocol is based on the criteria that reduce the sampling, number of intermediate conformations, and torsional refinement. On the other side, XP protocol is based on greater requirements for ligand-receptor shape complementarity and thus uses a more complex scoring function. This leaves out false positives that SP or HTVS scoring functions may let through. The docking results with the two waters model indicated that SP and HTVS methods identify almost the same number of hits at site X while the number was reduced to half using XP approach. In the docking calculations using one water model, the number of hits at site X were increased by a factor of two, due to the occupancy of the

hydration site (W2) by fragments. From each docking model, the molecules ranked within top 200 (obtained by XP, SP and HTVS) were analyzed to identify the key residues involved in ligand binding and subsequently to prioritize the best fragments with consensus scoring. The interaction pattern of the top ranked fragments (obtained using three scoring functions) at site X showed that the fragments with benzylamine moiety have better binding affinity. This also identified Tyr58, Asp70 (from one chain) and Asp118 (from neighboring chain) as key residues (**Figure 3.9**) involved in fragment binding. In most of the top ranked fragments, Tyr58 forms π stacking while Asp70 shows salt bridge interactions with benzylamino group of the fragments. Based on consensus scoring criteria involving XP and SP/HTVS or all the three scoring functions, a total of 94 and 89 fragments were identified as top ranked fragments (at site X) using one and two water models, respectively.

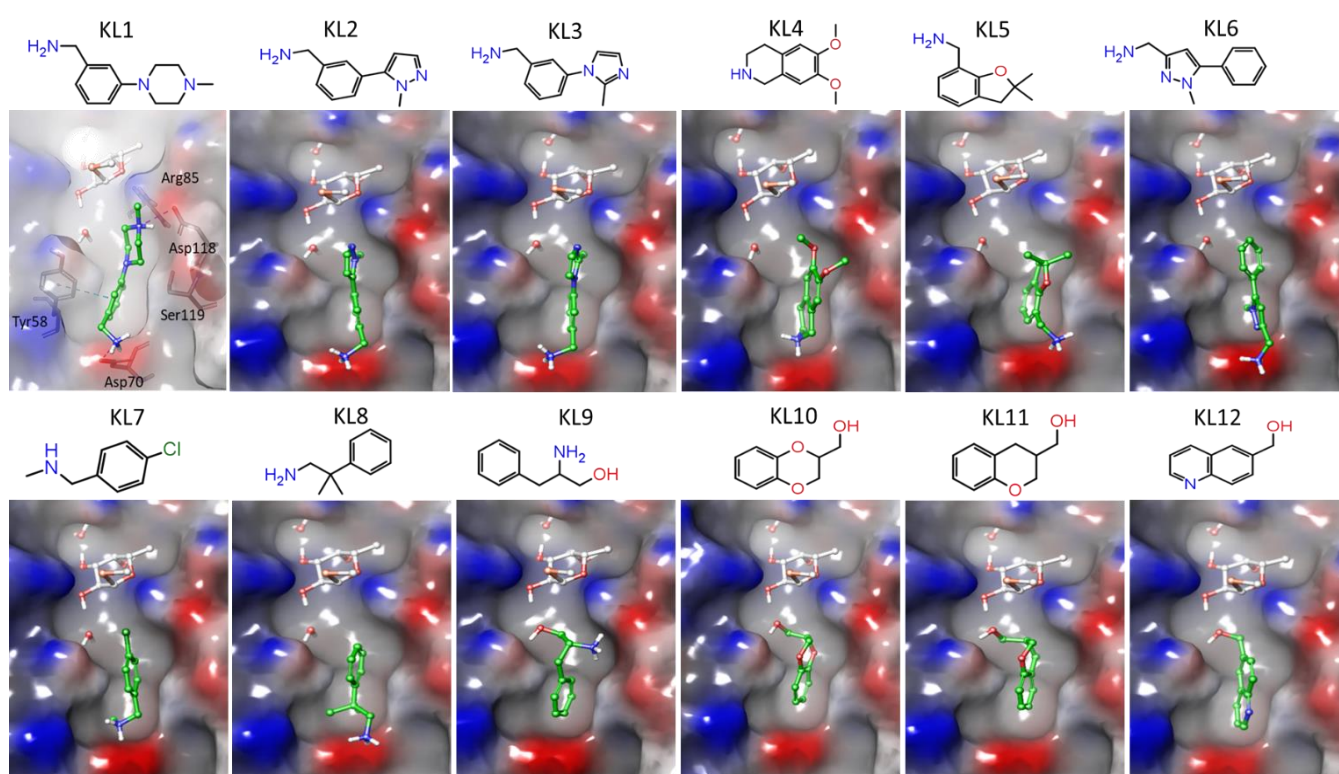
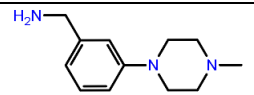
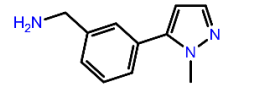
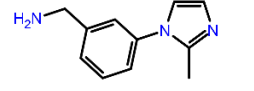
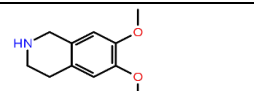
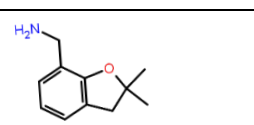
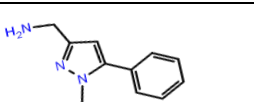
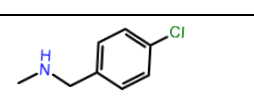
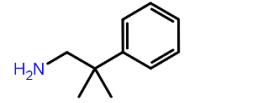


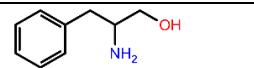
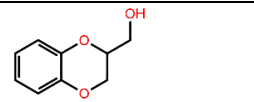
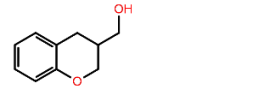
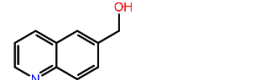
Figure 3.9 Predicted binding pose (at the site X) for the top ranked fragments (KL1-KL12) using docking studies. The key residues involved in the fragment binding are shown in the docking pose of KL1.

3.4.4 Selection of best fragments from the docking results

The identified top ranked fragments were visually analyzed based on different parameters such as structural diversity, possibility to connect them to the fucose core by considering the binding pose, size of the fragments and distance from the fucose core. The fragments which docked at regions significantly far ($>6 \text{ \AA}$) from the fucose core and bind on the shallow surface outside the site X were discarded. The remaining 32 fragments were redocked at the site X in order to save the multiple (10) binding poses and know the stability of ligand interactions. Consideration of additional factors like synthetic feasibility, commercial availability and purchasing cost allowed to finalize the best 12 fragments (**Figure 3.9, Table 3.2**) for experimental studies.¹¹

Table 3.2 Top ranked fragments identified for site X.

Fragment name	Structure	Molecular weight	Aqueous Solubility(mM)
KL1		205.3	20
KL2		187.2	50 ^[b]
KL3		187.2	100
KL4		193.2	100
KL5		177.2	1 ^[a]
KL6		187.2	1 ^[a]
KL7		155.6	20
KL8		149.2	50

KL9		151.2	100
KL10		166.2	1 ^[a]
KL11		164.2	20
KL12		159.2	25 ^[b]

[a] Solubility at higher concentration is not known

[b] dissolved in 10 percent DMSO

Among the final molecules, fragments KL1-KL8 are the top scorer in the two waters model, while the fragments KL9-KL12 are the top scoring fragments in the one water model.

3.5 Hit expansion

The identified fragments from docking studies were further employed for hit expansion by using 2D fingerprint based similarity (95%) search based on Tanimoto coefficient¹² in PubChem database.¹³⁻¹⁴ Total 28 molecules from both the models (12 from two water and 16 from one water model) were used as input for similarity search which identified 6385 fragments (3339 molecules from two waters model and 3046 molecules from one water model). The fragments were prepared for docking by generating tautomers, stereoisomers and protonation states using LigPrep¹⁵ tool. The calculations generated 15374 structures: 6937 for one water model and 8437 for two waters model. The structures were docked using one or two waters docking models. The schematic of the protocol is given in **Figure 3.10**. Only some of the identified molecules (from similarity search) showed docking score higher than the query molecules. These molecules were redocked to analyse their multiple binding poses so that the molecules binding outside the binding pocket (site X) or significantly far (>6 Å) from the fucose ring can be discarded. Finally, 19 fragments (7 fragments from two waters model and 12 fragments from one water model) were obtained.

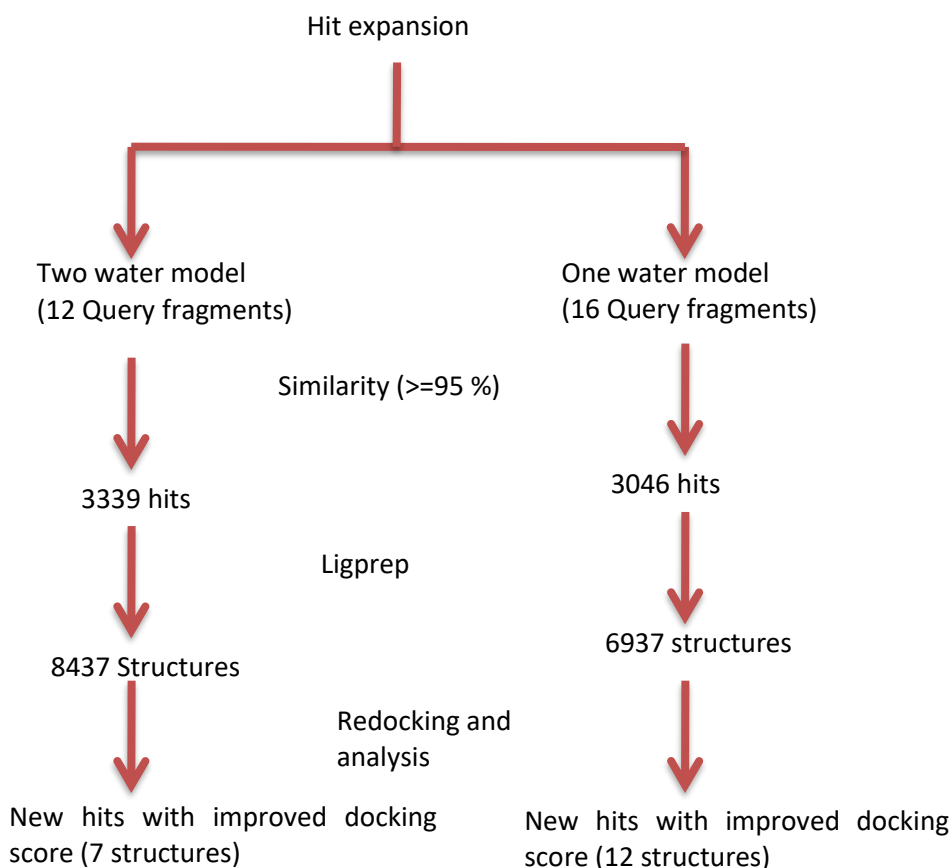
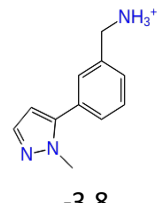
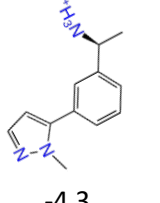
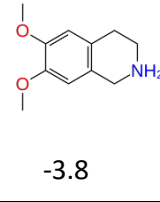
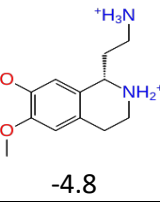
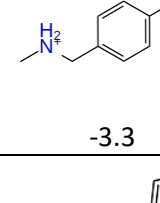
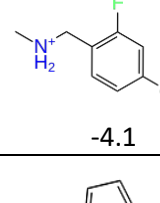
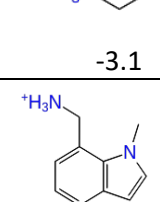
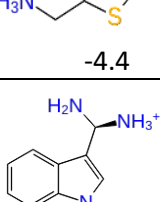
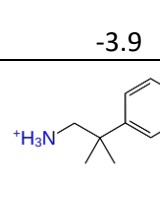
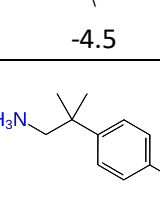
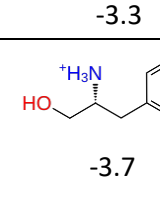
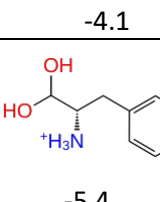
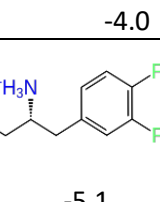

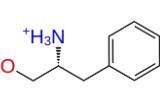
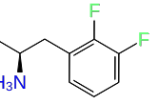
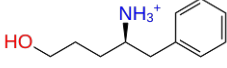
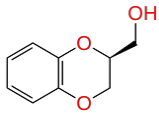
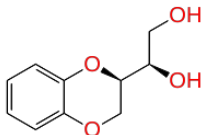
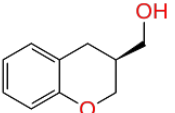
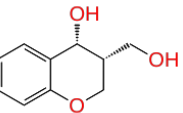
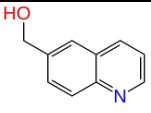
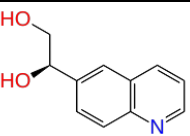
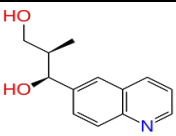
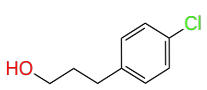
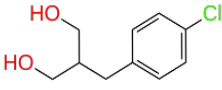
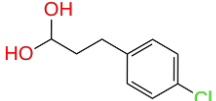
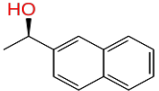
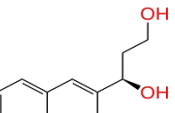


Figure 3.10 Schematic of the protocol used for hit expansion.

These new molecules from hit expansion consist of additional flexible moieties in the structure (**Table 3.3**). The introduction of additional groups in the molecules increased the stereocenters and the structural complexity that ultimately increased the purchasing cost. Moreover, most of the molecules were not available with the commercial vendors. Therefore, we decided to proceed further with the best molecules selected from the query molecules without using the fragments from the hit expansion. However, the fragments obtained from the hit expansion can be synthesized in future or they can also be used further for glycomimetic ligand design in later studies.

Table 3.3 Molecules with improved binding affinity (docking score) obtained from similarity search. Docking scores (from XP docking, kcal/mol) are mentioned below the molecule. Fragments KL1-KL8 were docked using two waters model while the remaining fragments were docked using one water model. Only those query fragments which identified new fragments with improved binding affinity are shown in the table.

Molecule	Query structure	Similar molecules with improved GlideScore (kcal/mol)	
KL2	 -3.8	 -4.3	
KL4	 -3.8	 -4.8	
KL7	 -3.3	 -4.1	
KL*	 -3.1	 -4.4	
KL*	 -3.9	 -4.5	
KL8	 -3.3	 -4.1	 -4.0
KL9	 -3.7	 -5.4	 -5.1

		-4.9	-4.5
			
		-4.5	
KL10			
	-4.7	-6.0	
KL11			
	-4.6	-5.8	
KL12			
	-3.7	-5.2	-4.8
KL*			
	-3.8	-5.2	-5.0
KL*			
	-4.0	-5.3	

*These fragments were not included in the best 12 fragments (Table 3.2) but used for similarity search.

3.6 References

- [1] Šulák, O. *et al.*, A TNF-like Trimeric Lectin Domain from Burkholderia cenocepacia with Specificity for Fucosylated Human Histo-Blood Group Antigens, *Structure* **2010**, *18*, 59-72.
- [2] Berman, H. M. *et al.*, The Protein Data Bank, *Nucleic Acids Res.* **2000**, *28*, 235-242.
- [3] Bermeo, R. *et al.*, BC2L-C N-Terminal Lectin Domain Complexed with Histo Blood Group Oligosaccharides Provides New Structural Information, *Molecules* **2020**, *25*, 248.
- [4] Halgren, T. A., Identifying and characterizing binding sites and assessing druggability, *J. Chem. Inf. Model.* **2009**, *49*, 377-389.
- [5] Schrödinger Release 2018-1: Maestro, Schrodinger, LLC, New York, NY, **2018**.

- [6] Olsson, M. H. *et al.*, PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions, *J. Chem. Theory Comput.* **2011**, *7*, 525-537.
- [7] Li, H. *et al.*, Very fast empirical prediction and rationalization of protein pKa values, *Proteins* **2005**, *61*, 704-721.
- [8] Bas, D. C. *et al.*, Very fast prediction and rationalization of pKa values for protein-ligand complexes, *Proteins* **2008**, *73*, 765-783.
- [9] Harder, E. *et al.*, OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins, *J. Chem. Theory Comput.* **2016**, *12*, 281-296.
- [10] Friesner, R. A. *et al.*, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, *J. Med. Chem.* **2004**, *47*, 1739-1749.
- [11] Lal, K. *et al.*, Prediction and Validation of a Druggable Site on Virulence Factor of Drug Resistant Burkholderia cenocepacia, *Chem. Eur. J.* **2021**, *27*, 10341-10348.
- [12] Willett, P., Similarity-based virtual screening using 2D fingerprints, *Drug Discov. Today* **2006**, *11*, 1046-1053.
- [13] Kim, S. *et al.*, PubChem in 2021: new data content and improved web interfaces, *Nucleic Acids Res.* **2020**, *49*, D1388-D1395.
- [14] Chen, X.Reynolds, C. H., Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407-1414.
- [15] Sudakevitz, D. *et al.*, A new Ralstonia solanacearum high-affinity mannose-binding lectin RS-III structurally resembling the Pseudomonas aeruginosa fucose-specific lectin PA-III, *Mol. Microbiol.* **2004**, *52*, 691-700.

4. Experimental validation of fragment binding

All the selected fragments were tested using a series of biophysical methods to investigate their interactions with the lectin.¹ The results obtained from each method are discussed in this chapter.

4.1 Protein production and purification

As already discussed under the methods section in chapter 2, the primer design, gene amplification and ligation were performed by Rafael Bermeo.² For ligation, the genetic sequence coding for the 132 first amino acids of BC2L-C-nt amplified via PCR was inserted into pCold-TEV (**Figure 4.1**). By employing the same vector, production and purification of the BC2L-C-nt was performed using the protocol described by Rafael Bermeo.²

The vector (pCold-TEV) includes a promoter (cspA), a translation enhancement element (TEE), trigger factor, and 6-Histidine tag at the N-terminus resulting in fusion protein of about 52.4 kDa. This fusion could be cleaved using the tobacco etch virus (TEV) protease using TEV-cleavage site while the 6-Histidine tag could ease the purification. pCold-TEV-BC2L-

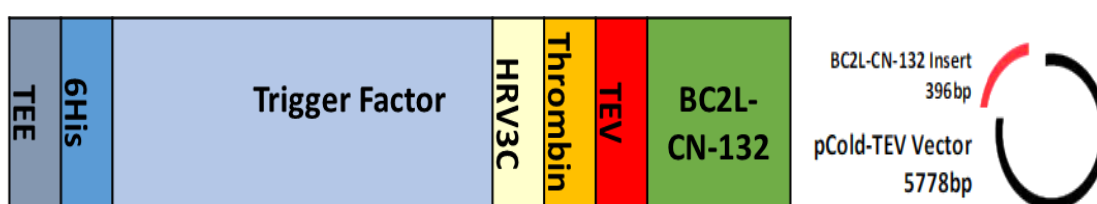


Figure. 4.1 Schematic of the expression construct of BC2L-C-nt in pCold-TEV.

CN-132 was transformed by heat shock in the BL21 (DE3) Star strain of *E. coli* for recombinant production. Then a successful expression was obtained after overnight protein expression at 16 °C induced by addition of 0.01 mM isopropyl β -D-thiogalactoside (IPTG) in Luria Broth (LB) medium. The fusion was cut using TEV protease overnight and the desired protein (BC2L-C-nt) was purified using immobilized nickel affinity chromatography thanks to imidazole

gradient. The presence of His-tag facilitated the purification before and after TEV cleavage (Figure 4.2 and 4.3). The results from SDS-PAGE (Sodium dodecyl-sulfate polyacrylamide gel electrophoresis) indicated that the fusion was partially cut probably due to poorly accessible TEV cleavage sites upon oligomerisation (Figure 4.5 A). Finally, an average yield of 4.5 mg.L⁻¹

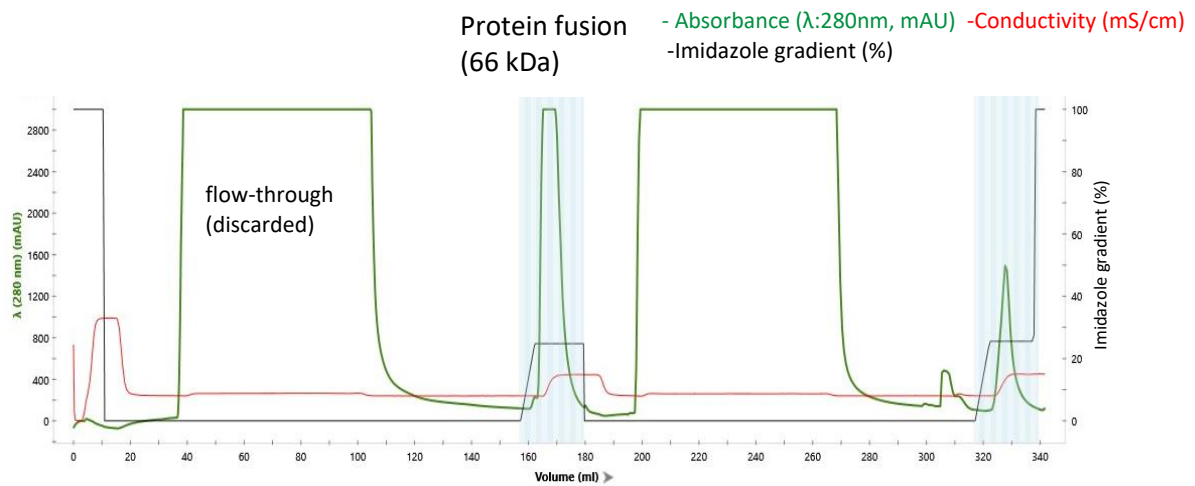


Figure 4.2 Purification of BC2L-C-nt. First, all proteins with no affinity to IMAC column were discarded. Then, imidazole gradient allowed elution of his-tag protein by disrupting the nickel/protein interactions.

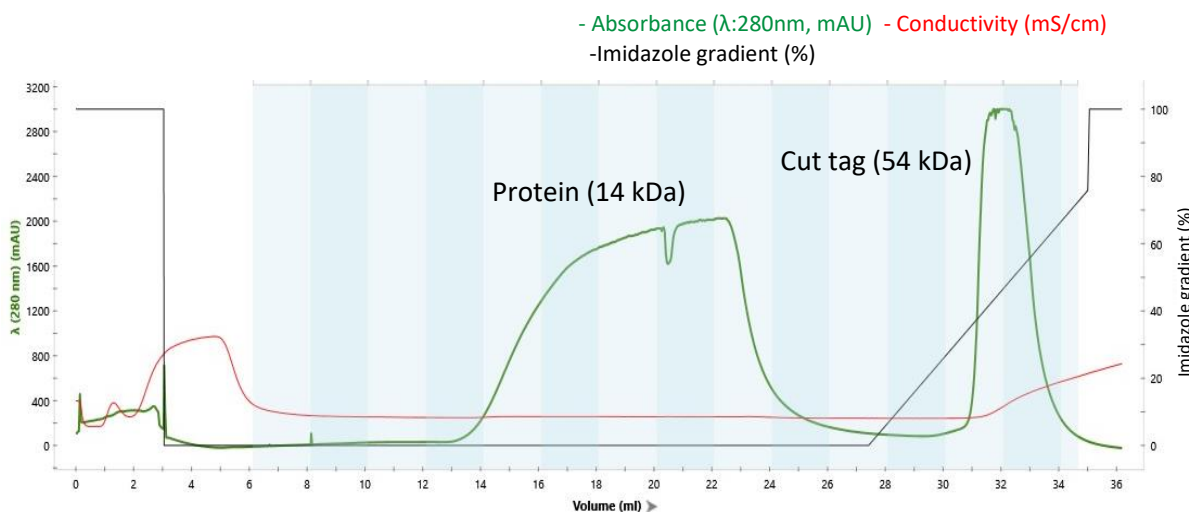


Figure 4.3 Purification of BC2L-C-nt after TEV cleavage on IMAC column.

protein was obtained as pure fraction (Figure 4.4 and 4.5 B) after size exclusion chromatography (SEC). The protein was stored at 4 °C.

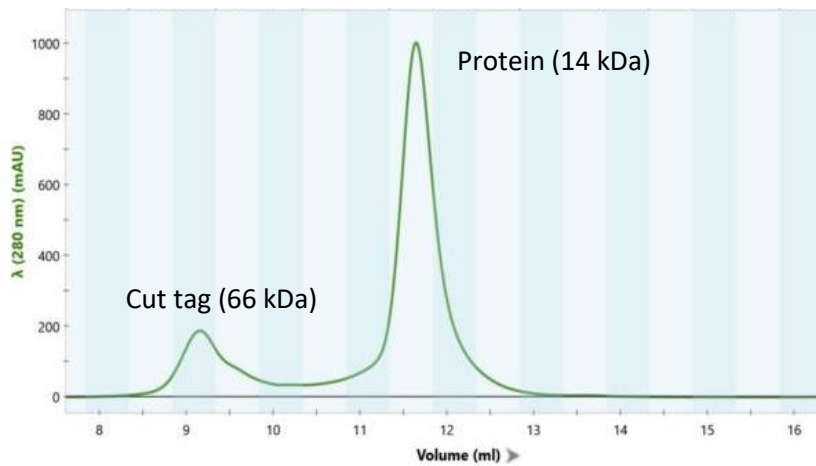


Figure 4.4 Purification of BC2L-C-nt by size exclusion chromatography (SEC) using Enrich 70 column.

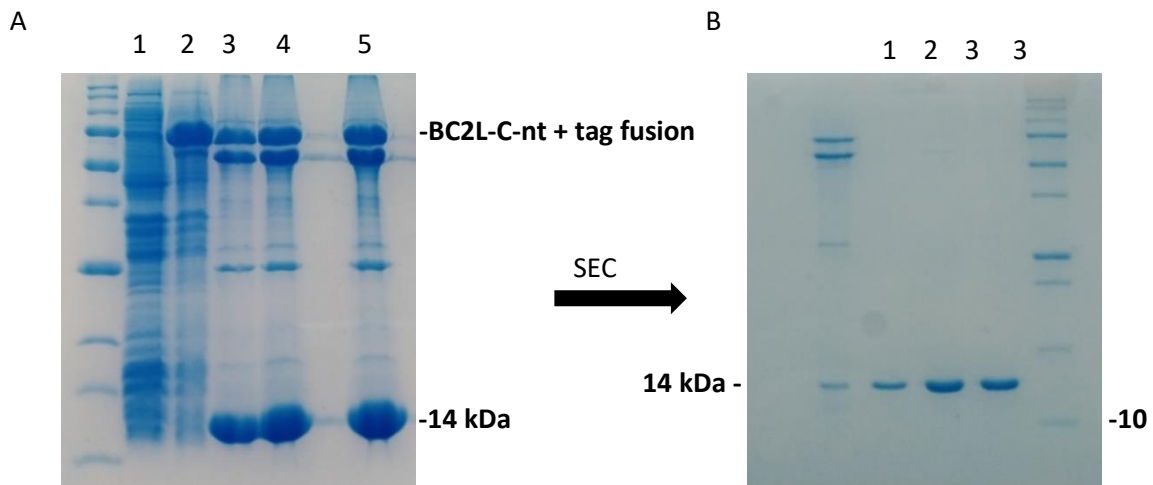


Figure 4.5 (A) 10 % SDS-PAGE showing partial cutting of histidine tag in wells 3, 4 and 5. Other wells (1 and 2) correspond to uncut protein fraction. (B) A pure fraction of BC2L-C-nt was collected (wells 1, 2 and 3) after SEC.

4.2 Thermal shift assay (TSA)

A solution of each of the 12 fragments (refer **Table 3.2, Chapter 3**) of interest (2.5 mM) was used to test the interaction with BC2L-C-nt using thermal shift assay (TSA, ThermoFluor).³ To validate the protocol, methyl α -L-fucoside (Me- α -L-Fuc) was used as a reference for positive control. The results displayed an expected positive shift (~ 2 °C) in the melting temperature (T_m) upon Me- α -L-Fuc binding (**Figure 4.6**), thus validating the protocol.

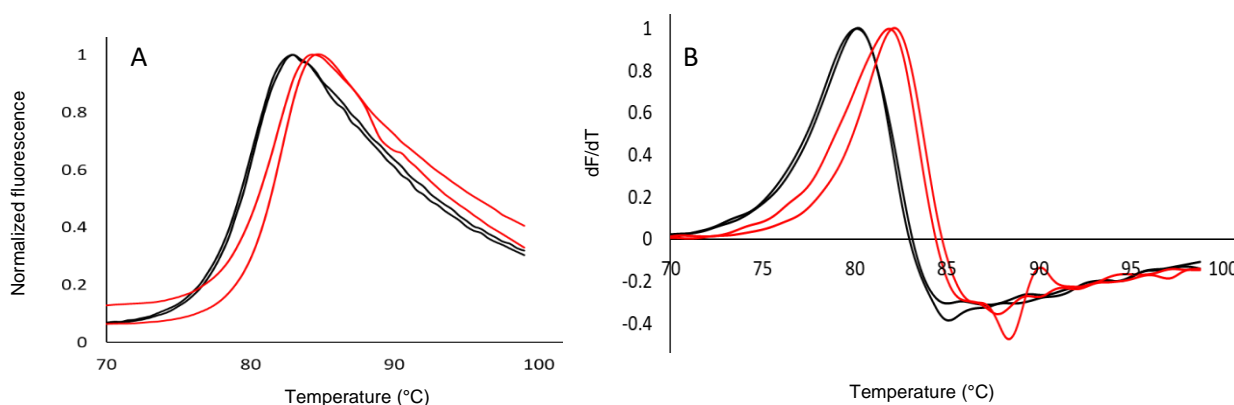


Figure 4.6 (A) The melting curves were obtained for BC2L-C-nt at 5 μM in the presence (red) and absence (black) of 20 mM Me- α -L-Fuc. (B) First derivatives of fluorescence (melting) curves.

Subsequently, the experiment was performed using the 12 fragments in the presence of 20 mM Me- α -L-Fuc. All fragments affect the T_m , indicating a probable interaction with the protein. For all of them, a negative shift of the melting temperature (T_m) was observed, with amplitude between 0.15 to 1.65 $^{\circ}\text{C}$ (**Figure 4.7**). The fragments KL1-KL9 gave the strongest effects while the remaining fragments (KL10-KL2) displayed a weaker negative shift. The possible reason is that the fragments KL10-KL12 do not contain ammonium moiety

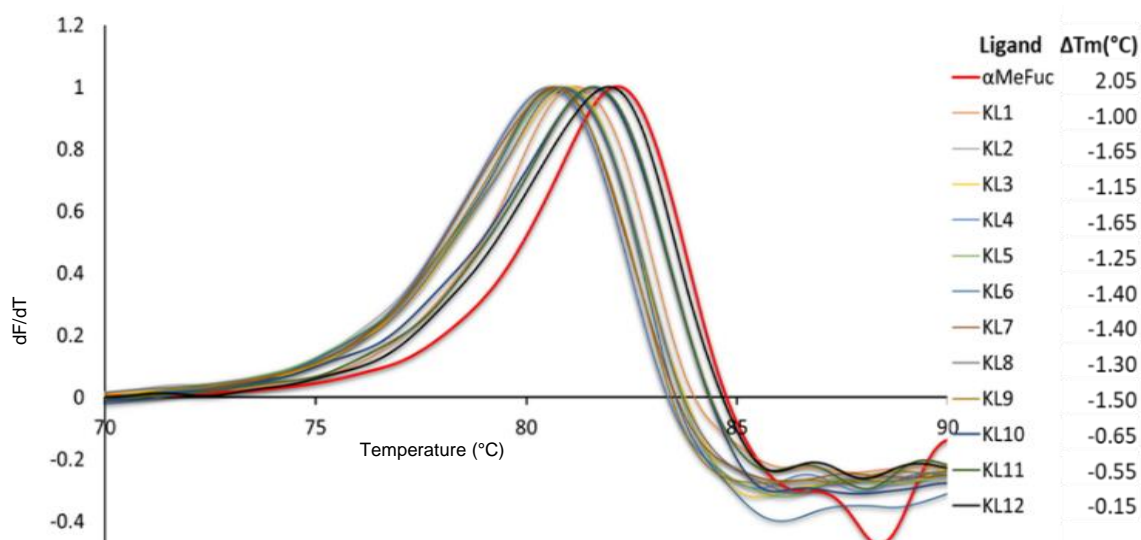


Figure 4.7 First derivatives of fluorescence curves of BC2L-C-nt (5 μM) in the presence of the fragments (KL1-12, 2.5 mM). The fragments induce a negative shift in the melting temperature (T_m) of BC2L-C-nt which indicates ligand interaction.

to establish H-bonding interactions with the residues (for example Asp70) in the binding site which probably makes them weak binders. Since negative shift corresponds to destabilization of the protein, this possibly suggests that the fragments bind at the interface of the monomers or to a non-native/partially unfolded state of the protein.⁴ The experiment was repeated in the absence of Me- α -L-Fuc to ensure that fucoside does not have any influence in the fragment binding. The results showed similar pattern of smaller negative shift in the melting temperature (**Figure 4.8**) upon fragment binding/interactions.

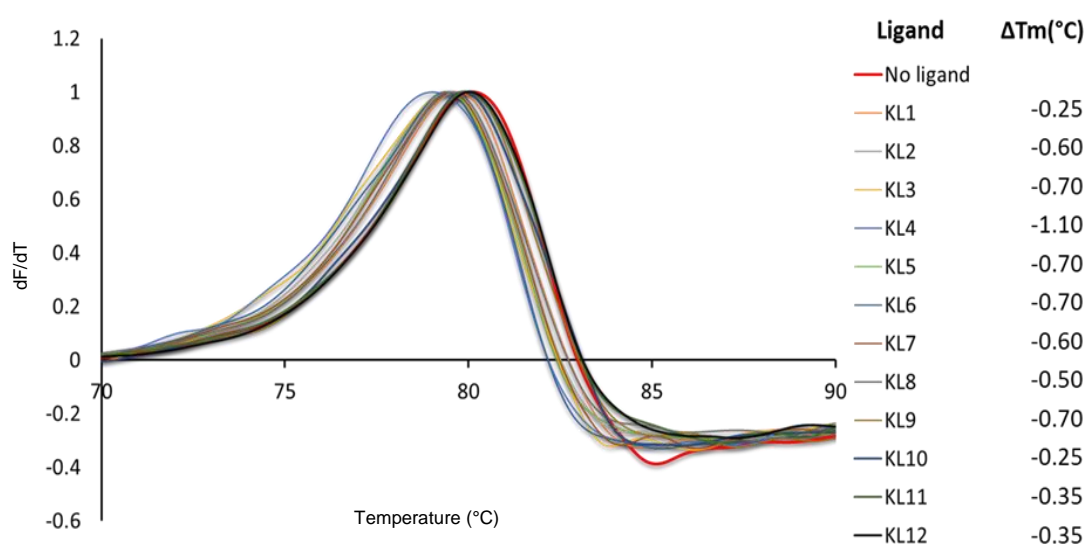


Figure 4.8 First derivatives of fluorescence curves of BC2L-C-nt (5 μ M) in the presence of the fragments (KL1-12, 2.5 mM) and the absence of Me- α -L-Fuc. The fragments induce a negative shift in the melting temperature (T_m) of the protein which indicates ligand interaction.

The results from both the experiments indicate that fragments interact with the target protein. However other biophysical methods are necessary to obtain more insights on the interactions.

4.3 Microscale Thermophoresis (MST)

The screening was repeated using the microscale thermophoresis (MST)⁵⁻⁶ method. All the fragments were tested using microscale thermophoresis to confirm their binding/

interactions with BC2L-C-nt. The lysine residues of the protein were labelled with the fluorescein isothiocyanate (FITC) dye and a ratio of 1.6 :1 (protein to dye ratio) was obtained as determined by UV spectroscopy. The protein at 326 nM was used to test the fragments at 2.5 mM. Me- α -L-Fuc was first assayed as a control experiment, but no significant response was obtained, using different dilutions (50 mM to .0015 mM) (**Figure 4.9**). It is likely that the modification of surface lysines by the fluorescein isothiocyanate (FITC) dye interferes with the fucose binding as one of the lysine residues (Lys78) was in the proximity (~ 11.20 Å) to the fucose binding site. The MST experiment at 2.5 mM concentration was repeated for all the fragments in the presence (**Figure 4.10 A, B**) and the absence (**Figure 4.10 C, D**) of Me- α -L-Fuc to identify the fragments with binding affinity for the BC2L-C-nt.

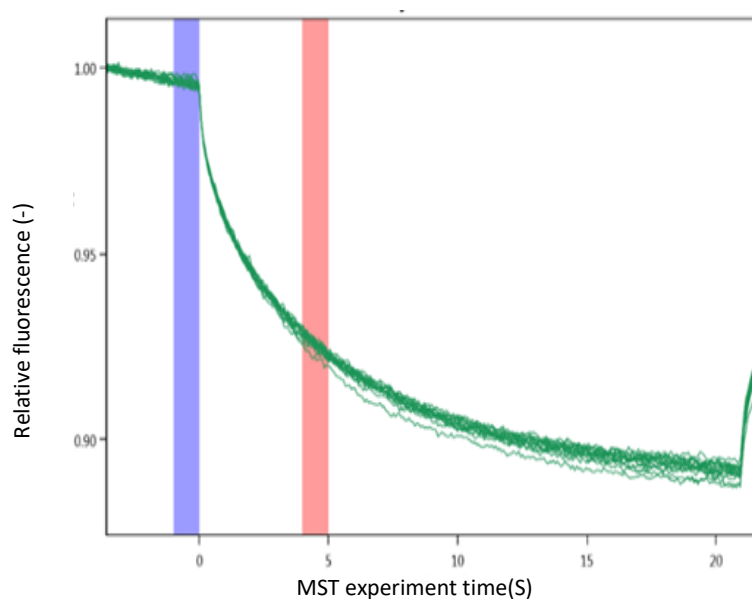


Figure 4.9. Fluorescence signal of BC2L-C-nt-FITC for different dilutions (50 mM to 0.0015 mM) of Me- α -L-Fuc.

The resulting graphs (**Figure 4.10**) from the MST experiments show a change in fluorescence pattern when different fragments were tested for their binding to BC2L-C-nt. Thus, both the screening experiments indicate interaction of the fragments to the protein domain. However, it is also likely that the fluorescein isothiocyanate (FITC) dye might have an influence on the

binding interactions of the fragments as the lysine residue (Lys108 from chain C) is located near ($\sim 5 \text{ \AA}$) to the expected fragment binding site (site X).

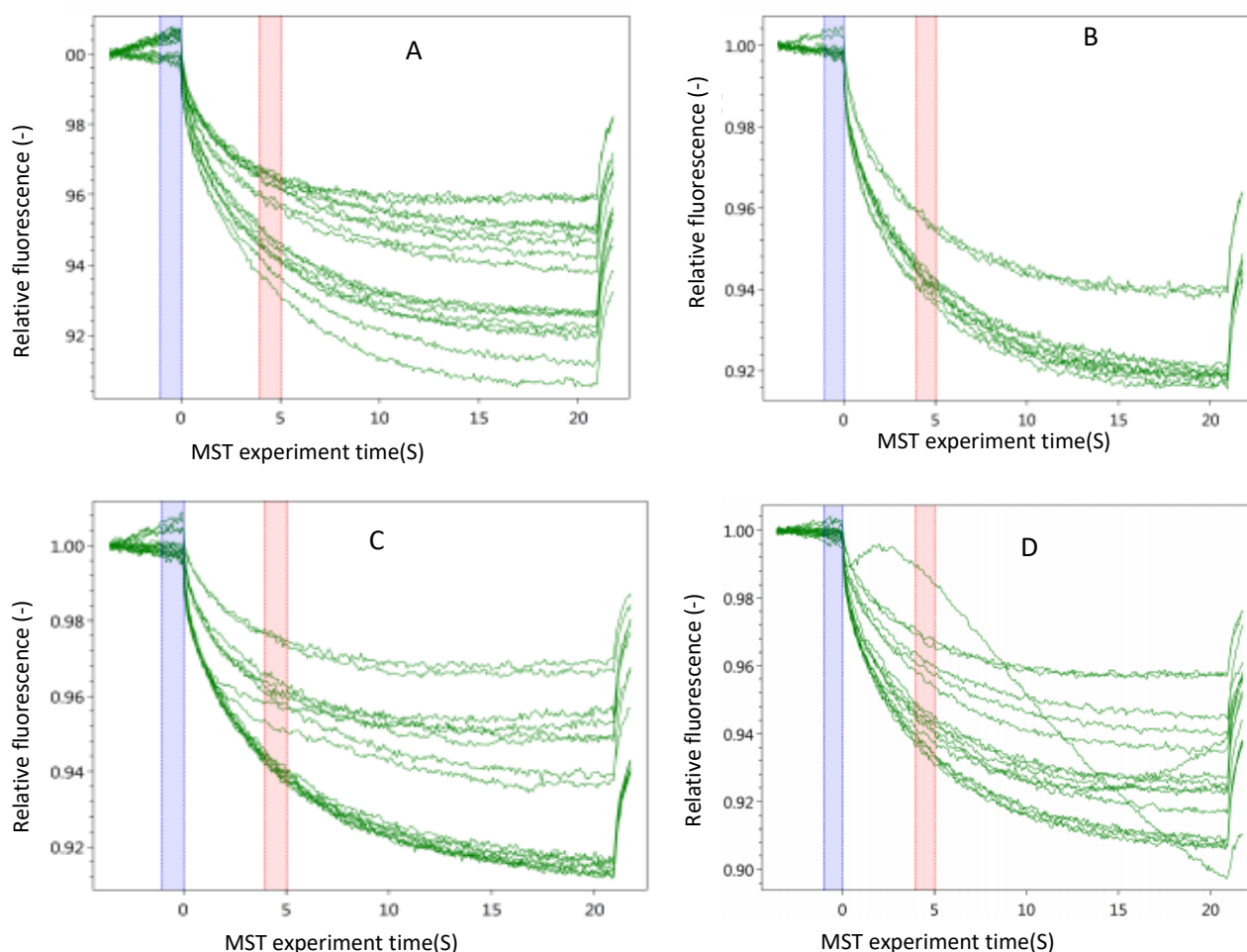


Figure 4.10 Fluorescence (binding) signals of BC2L-C-nt (A) for the fragments KL1-7 (2.5 mM) and (B) the fragments KL8-12 (2.5 mM) in the presence of Me- α -L-Fuc (326 nM). The experiment was repeated for (A) the fragments KL1-7 and (D) KL8-12 in the absence of Me- α -L-Fuc.

The experimental results of TSA and MST indicate that the fragments are binding to the lectin, but they do not afford structural information concerning the interaction, thus cannot be conclusive about the location of binding site of the fragments. Therefore, another screening was performed using STD-NMR⁷ and X-ray crystallography⁸ for the protein and the fragments.

4.4 STD-NMR analysis of fragment binding

Saturation transfer difference (STD) NMR has become a popular technique to characterize weak fragment-macromolecule interactions in solution.⁹⁻¹³ This technique is sensitive to weak binding events (dissociation constant in μM to mM range)¹¹⁻¹³, thus widely used to characterize protein-carbohydrate interactions.⁹⁻¹³ In the STD-NMR experiments, macromolecule (protein) is selectively irradiated at the resonance frequency of side chains of specific amino acids which causes transfer of the magnetization from macromolecule to the ligand. Spin diffusion is responsible for the spread of the magnetization from the irradiated residues to the rest of the protein. STD can confirm the binding event and provide information on the regions of the fragment that are in contact with the protein. The protons of the ligand that receive a higher magnetization produce more intense signals on a 1D ^1H -STD spectrum. In general, methyl group of amino acids such as valine, leucine, or isoleucine, that are often present in the binding site of proteins, are irradiated (between 1 and -1 ppm).¹¹ The binding fragments usually have preferred interaction with aliphatic or aromatic amino acids of the protein, therefore experiment are performed at different irradiation frequency.¹⁴ The experiments discussed here were performed using irradiation at -0.05 ppm and 10 ppm which correspond to the aliphatic and aromatic protons, respectively.

A first STD-NMR experiment was performed using Me- α -L-Fuc alone with irradiation at -0.05 ppm. This resulted in transfer to resonance for the peaks at 1.1 ppm that corresponds to the methyl group carries by carbon C5 of the fucose ring. This is in agreement with the strong involvement of the methyl group (**Figure 4.11**) in its binding to BC2L-C-nt, as seen in the crystal structure (PDB 2WQ4).

To investigate the interactions of fragments with BC2L-C-nt, STD-NMR experiments (irradiating at -0.05 ppm) were performed using fragments KL3, KL8 (top scoring fragments in two waters model) and KL9 (top scorer from one water model) in the presence of the protein and of 2mM Me- α -L-Fuc. Two protocols for sample preparation were followed, either by adding the fragment to a pre-incubated solution of protein and Me- α -L-Fuc, or by adding the fucoside to a pre-incubated solution of protein and fragment. The resulting STD spectra were very similar independent of the set up. The samples were prepared at 1:1 ratio between the

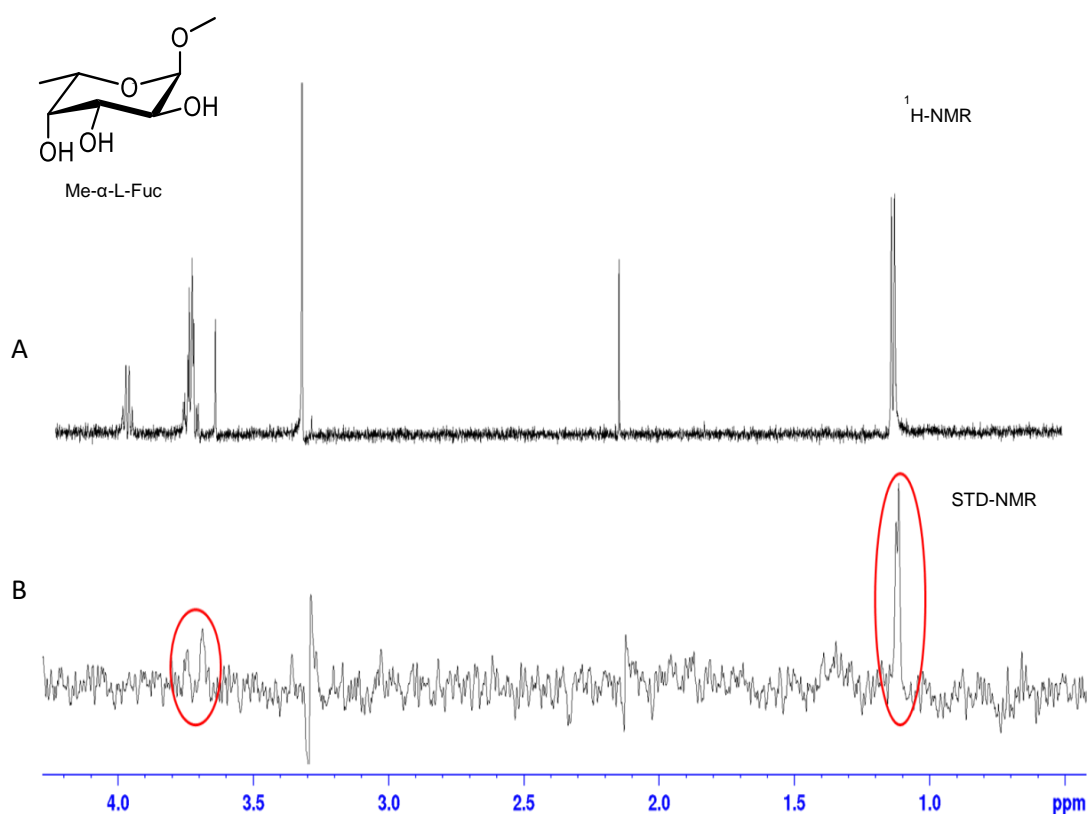


Figure 4.11 (A) $^1\text{H-NMR}$ and (B) STD spectrum of Me- α -L-Fuc in the presence of BC2L-C-nt (1000:1). The red circles represent the signals of the fucose ring (at 3.7 ppm) and of the methyl group carries by carbon C5 (at 1.1 ppm). The spectrum was recorded with a Bruker Avance 600 MHz spectrometer at 298K with irradiating frequency -0.05 ppm.

sugar and the fragment. The resulting STD spectra for KL3, KL8 and KL9 are shown in **Figures 4.12, 4.13** and **4.14**, respectively.

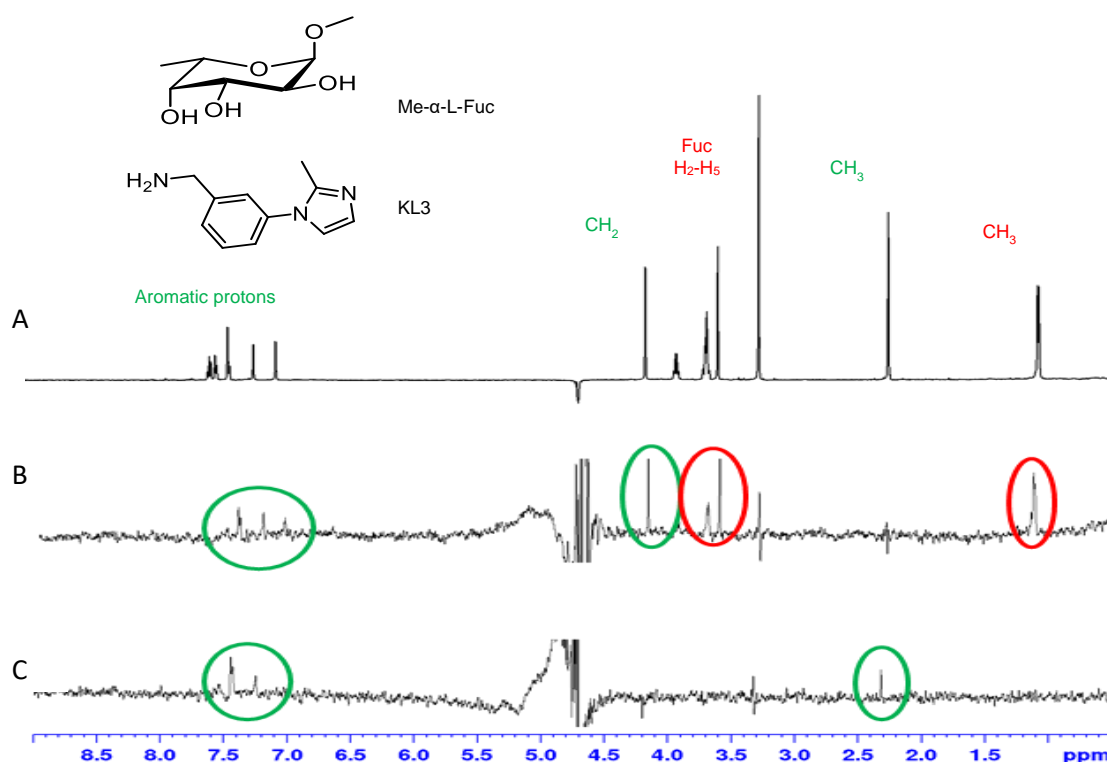


Figure 4.12 A) $^1\text{H-NMR}$ spectrum, B) STD spectrum at irradiation frequency -0.05 ppm and C) 10 ppm of fragment KL3 and Me- α -L-Fuc in the presence of BC2L-C-nt (1000:1). The signals produced by the fucose ring and its methyl group are highlighted with red circles at 3.7 ppm and 1.1 ppm, respectively. The green circles (at 4.2 ppm for $-\text{CH}_2-$ and in the range 7.05-7.4 ppm for aromatic protons) highlight the signals of the fragment. The irradiation at 10 ppm shows the signal (C) for aromatic protons (range 7.1-7.4 ppm) and the methyl group (at 2.3 ppm) of the fragment. The spectrum was recorded using a Bruker Avance 600 MHz spectrometer at 298K.

The spectra for fragment KL3 at -0.05 ppm shows the enhanced signal for aromatic protons and $-\text{CH}_2-$ group of the fragment. Likewise, fragment KL8 displayed the signals for $-\text{CH}_2\text{-NH}_2$ and aromatic protons while the fragment KL9 exhibited the STD signals for $-\text{CH}_2\text{-Ph}$ and aromatic protons at -0.05 ppm irradiation. For all the fragments, interaction with BC2L-C-nt was observed in the presence of Me- α -L-Fuc which confirmed the binding of the top scoring fragments from both the models. This further confirmed that interaction of the fragments

and Me- α -L-Fuc with BC2L-C-nt takes place simultaneously. In the STD spectra, signals induced by Me- α -L-Fuc and the fragments show comparable intensities which suggests a

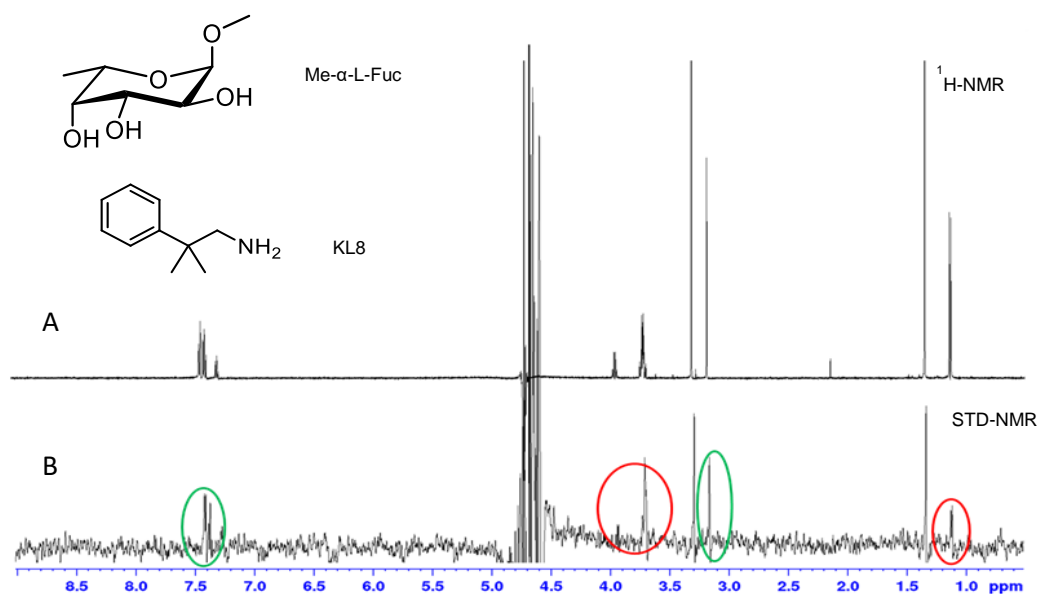


Figure 4.13 A) $^1\text{H-NMR}$ and B) STD spectrum of fragment KL8 and Me- α -L-Fuc in the presence of BC2L-C-nt (1000:1) at irradiating frequency -0.05 ppm. The red circles highlight the signals of the fucose ring (3.7 ppm) and of the methyl group (1.1 ppm). The signals of fragment (KL8) are highlighted with green circles; $-\text{CH}_2-\text{NH}_2$ at 3.2 ppm and aromatic protons at 7.4 ppm. The spectrum was recorded at 298 K with a Bruker Avance 600 MHz spectrometer.

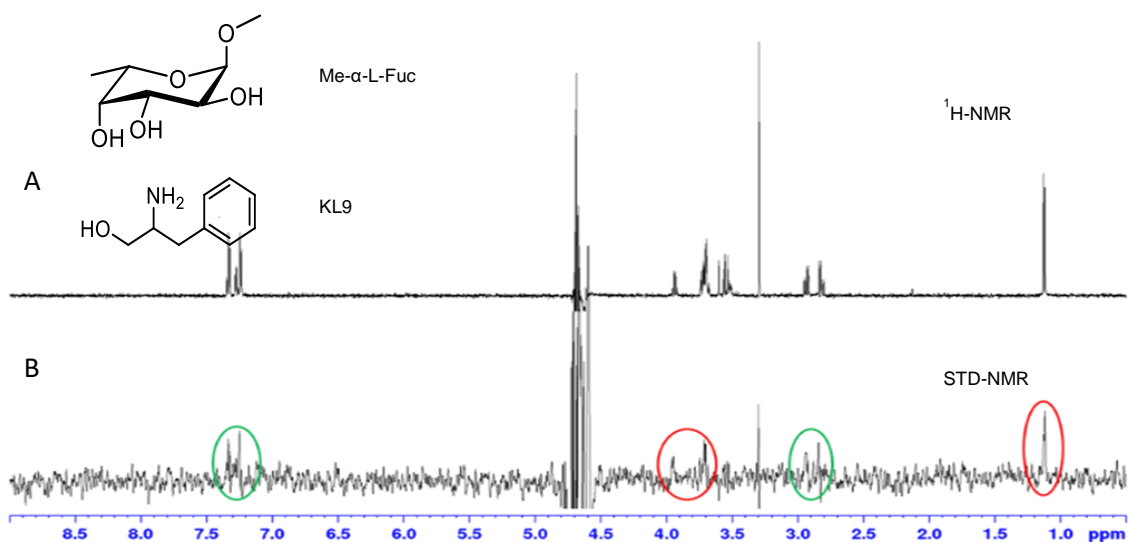


Figure 4.14 A) $^1\text{H-NMR}$ and B) STD spectrum of fragment (KL9) and Me- α -L-Fuc in the presence of BC2L-C-nt (1000:1) at -0.05 ppm irradiation frequency. The signals produced by the fragment are highlighted with green circles for $-\text{CH}_2-\text{Ph}$ (at 2.9 ppm) and aromatic protons (at 7.3 ppm). The red circles highlight the signals of fucose ring and its methyl group at 3.7 ppm and 1.1 ppm, respectively. The spectrum was recorded at 298 K with a Bruker Avance 600 MHz spectrometer.

similar binding affinity of the fucoside and the fragments.

In addition, STD spectra were also acquired for the fragment KL3, KL8 and KL9 at 10 ppm irradiation frequency. In this case, the aromatic protons of the fragments involved in interactions were observed while no signal was detected for Me- α -L-Fuc (**Figures 4.12 C and Figure 4.15 A, B**). These observations suggest that the fragments bind in the proximity of aromatic residues of the target protein which further support the predicted binding pose in docking studies.

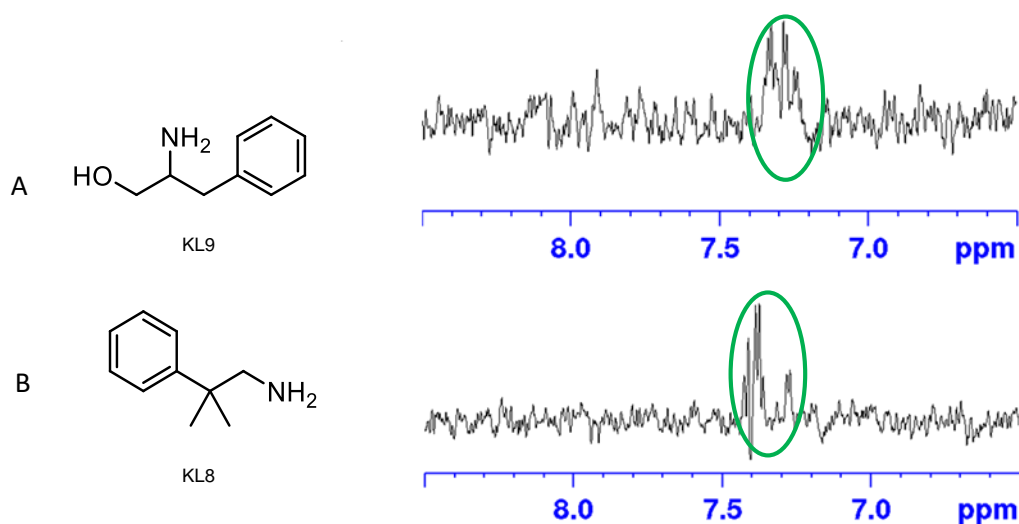


Figure 4.15 The STD spectra of fragment KL8 (A) and KL9 (B) were recorded in the presence of BC2L-C-nt and Me- α -L-Fuc (at 10 ppm irradiating frequency). The signals generated only for the aromatic protons of the fragments are highlighted in green circles. The spectra were recorded at 298 K with a Bruker Avance 600 MHz spectrometer.

The docking pose suggests that fragment KL3 binds near an aromatic residue (Tyr58) located in a protein binding pocket **Figure 4.16**. Notably, the STD spectrum of KL3 (at 10 ppm irradiation frequency) shows a clear signal (**Figure 4.12 C**) for the methyl group of the fragment (at 2.3 ppm), which was not observed when irradiated at -0.05 ppm (**Figure 4.12 B**). This indicates that methyl group is located near an aromatic side chain of the

protein. Conversely, the signal corresponding to the methyleneamino benzylic protons of the fragment (at 4.2 ppm), which was observed clearly at -0.05 ppm irradiation (Figure 4.12 B), disappeared from the spectrum. Thus, this moiety is likely to be surrounded by aliphatic protons of the binding site. Likewise, signals for the protons of fucose were also disappeared on irradiating at 10 ppm. The results from STD-NMR experiments indicate that the fragments are interacting with the target protein and probably bind at the predicted site (X).

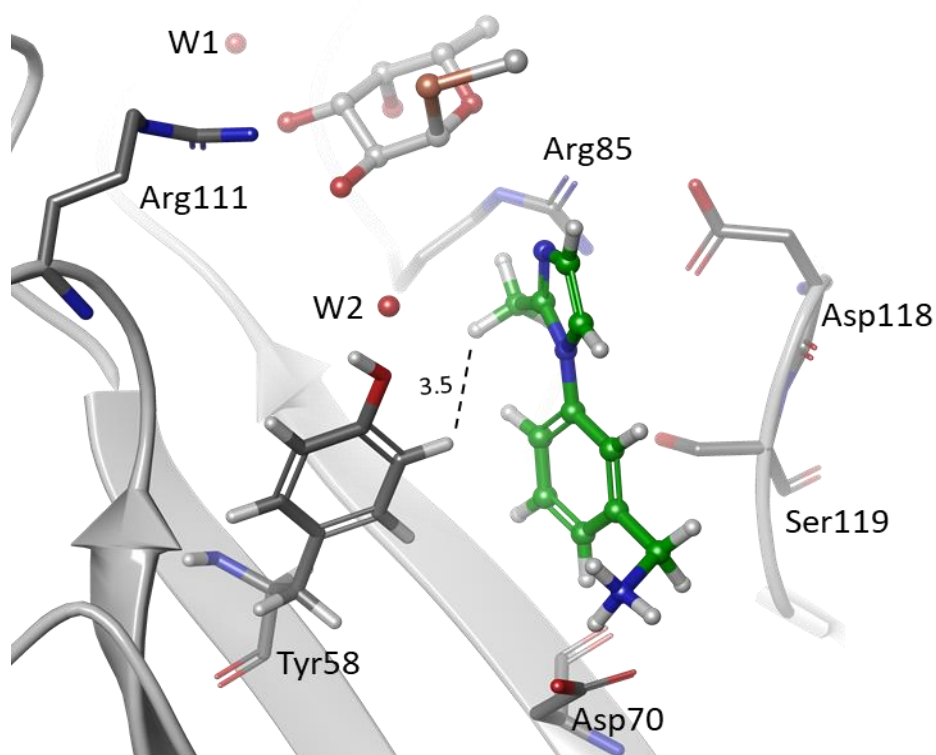


Figure 4.16 Docking pose of KL3 shows that methyl group binds near an aromatic residue (Tyr58) located in a protein binding pocket. The STD spectrum of KL3 (at 10 ppm irradiation frequency) also shows a clear signal for the methyl group of the fragment.

4.5 Crystal structure of the complex KL3-BC2L-C-nt

To further investigate the binding of fragments, attempts were made to obtain the crystal structure of fragment-protein complexes. All fragments (KL1-KL12) were

soluble in water at high concentration allowing their use for soaking experiment. First, the crystals of BC2L-C-nt complexed with Globo H hexasaccharide were obtained by following the method described by Rafael Bermeo.² Crystals were obtained within 48 hours in the form of clusters of plates (**Figure 4.17**) which were used for the soaking in fragment solution. Thereafter, these crystals were tested for the X-ray diffraction.

The crystals containing fragments KL10, KL11 and KL12 did not diffract at sufficient resolution

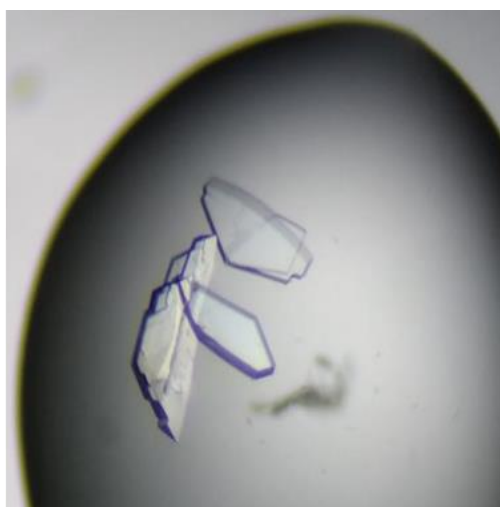


Figure 4.17 Crystals of BC2L-C-nt were obtained as clusters of plates.

to allow data collection. However, crystals soaked with the other fragments (KL1-KL9) diffracted at a resolution near 2 Å, and thus data were collected. All nine structures were solved using molecular replacement using 6TIG as search model and the electron density was analysed in the binding sites. In all cases, Globo H was present, but no electron density could be located for the fragments at the predicted site (X), except for KL3. This indicates that the fragments are not able to bind to the target protein in the experimental conditions used. Nonetheless, crystal soaked with KL3 (3-(2-Methyl-1H-imidazol-1-yl) benzylamine) diffracted at 1.9 Å resolution and electron density was visible for the fragment at the interface between monomers which clearly indicated the binding at the expected site. Data collection and refinement details of the complex are given in **Table 4.1**

Table 4.1 X-ray data collection and refinement for the BC2L-C-nt complex with Globo H and fragment KL3.

Data set	BC2L-C-nt complex with KL3 and Globo H		
PDB code	6ZZW		
Data collection			
Beamline	PROXIMA1 (SOLEIL)		
Wavelength (Å)	0.9786		
Space Group	C2		
a, b, c (Å)	74.46, 42.91, 103.34		
α, β, γ (°)	90.0, 96.10, 90.0		
Resolution (Å) ^a	37.13-1.90 (1.94-1.90)		
Total observations	175918		
Unique reflections	25522		
Multiplicity ^a	6.9 (7.1)		
Mean I/σ(I) ^a	10.2(4.0)		
Completeness (%) ^a	98.8 (98.1)		
$R_{\text{merge}}^{\text{a,b}}$	0.13 (0.53)		
R_{pim}			
$CC_{1/2}^{\text{a,c}}$	0.99 (0.89)		
Refinement			
Reflections: working/free ^d	24288 / 1224		
$R_{\text{work}} / R_{\text{free}}^{\text{e}}$	0.181 / 0.238		
Ramachandran plot: allowed/favoured/outliers (%)	100 / 97 / 0		
R.m.s. bond deviations (Å)	0.014		
R.m.s. angle deviations (°)	1.840		
R.m.s. chiral deviations	0.093		
No. atoms / Mean B-factors (Å ²)	Chain A	Chain B	Chain C
Protein	962 / 22.1	971 / 22.2	951 / 21.3
carbohydrate ligand	58 / 36.1	58 / 33.3	47 / 30.6
ligand ^f	14 / 32.0	14 / 38.7	14 / 36.0
water	69 / 26.9	65 / 26.6	72 / 24.9

^a Values for the outer resolution shell are given in parentheses.

^b $R_{\text{merge}} = \sum_{\text{hkl}} \sum_i |I_i(\text{hkl}) - \langle I(\text{hkl}) \rangle| / \sum_{\text{hkl}} \sum_i I_i(\text{hkl})$.

^c $CC_{1/2}$ is the correlation coefficient between symmetry-related intensities taken from random halves of the dataset.

^d The data set was split into "working" and "free" sets consisting of 95 and 5% of the data, respectively. The free set was not used for refinement.

^e The R-factors R_{work} and R_{free} are calculated as follows: $R = \sum (|F_{\text{obs}} - F_{\text{calc}}|) / \sum |F_{\text{obs}}|$, where F_{obs} and F_{calc} are the observed and calculated structure

factor amplitudes, respectively

^f refers to ligands bound in the active site and potential surface binding sites

The analysis shows that the residue Tyr58 forms π - π stacking (T-shaped) interactions with the benzene ring of the ligand while Asp70 establishes salt bridge interactions with the amino group in the ligand. In addition, the free nitrogen of the imidazole ring of the fragment forms water mediated H-bond with the side chain of Arg85 and the OH-4 of the GlcNAc moiety of Globo H (**Figure 4.18**). The orientation of the fragment and the key interactions with the protein correspond very well with the predicted binding pose using docking studies (**Figure 4.18 and Figure 4.19**)

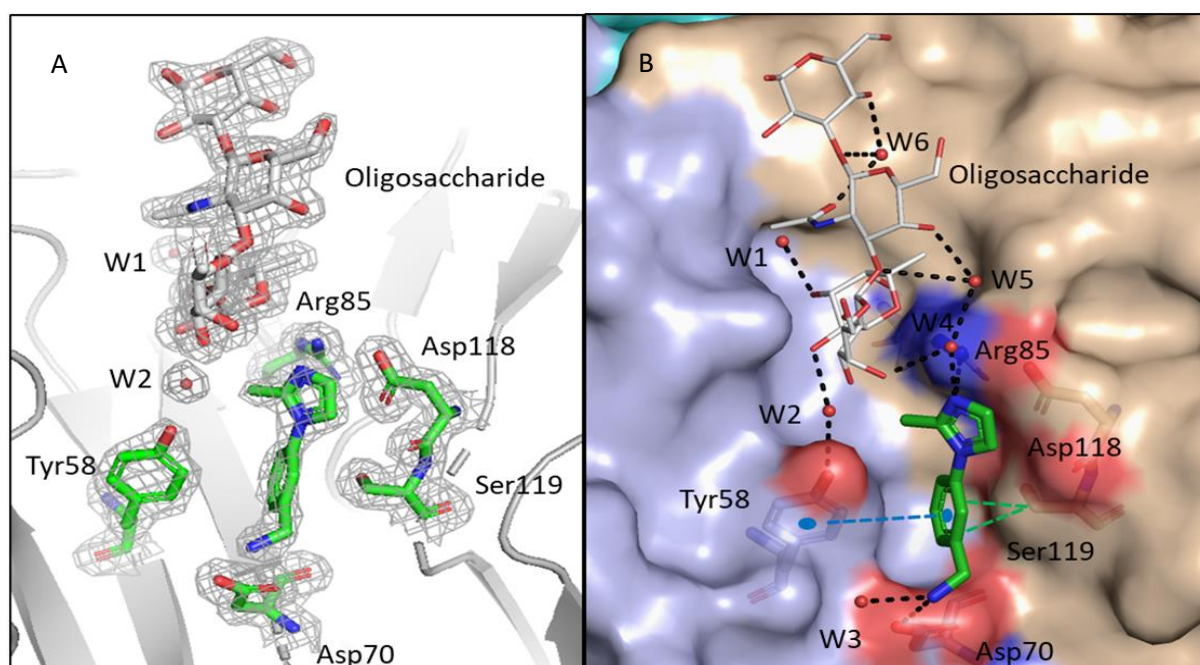


Figure 4.18 Crystal structure of BC2L-C-nt with Globo H and the fragment (KL3). (A) Enlarged view of the binding site of BC2L-C-nt with 2Fo-DFc electron density represented at 1 σ (B) Network of key interactions observed in the binding site (site X). Analysis shows that the key interactions and residues predicted from docking studies were involved in the fragment (KL3) binding. The complex shows π - π stacking interactions with Tyr58 and salt bridge interactions between Asp70 (side chain) and benzylamino group of the fragment. In addition, water molecules (other than W1 and W2) form bridging H-bond interactions with the fragment and the protein. Hydrophobic and H-bonding interactions are displayed in green and black dashed lines respectively. π - π stacking interactions are highlighted with blue dashed lines.

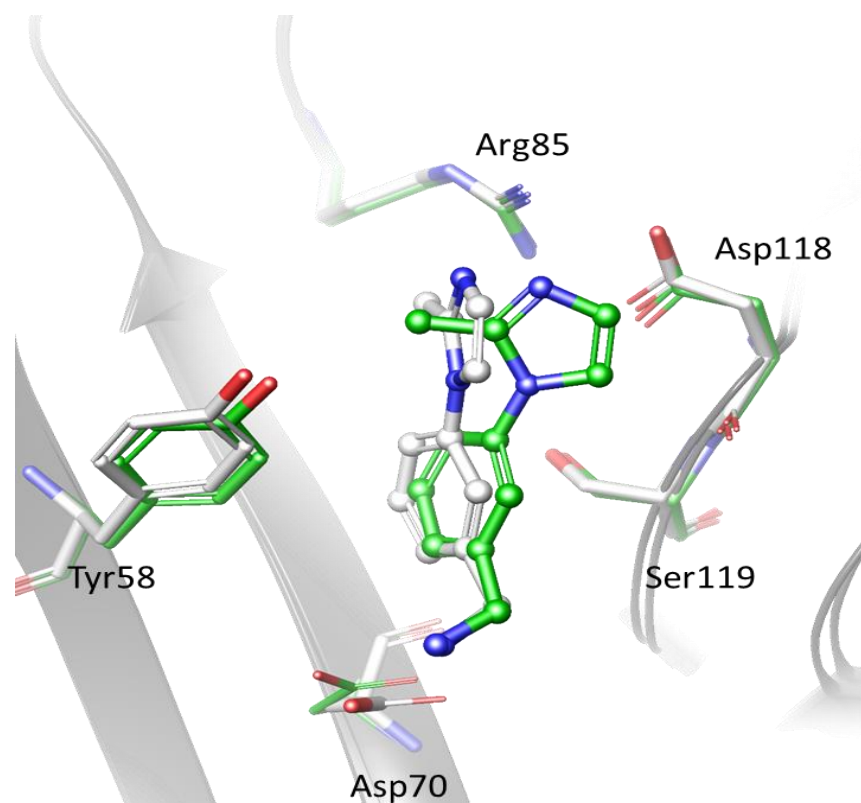


Figure 4.19 Comparison of binding pose of docked (grey) and crystallized complex (green) with KL3 (RMSD 0.4 Å).

The analysis of three binding sites located at the interface of trimer showed that the fragment reproduces the same binding interactions and thus binds with the identical pose in the three binding sites (**Table 4.2**).

Table 4.2 Analysis of interactions of BC2L-C-nt with KL3 in three binding sites in the trimer.

Ligand atom	Protein or water atom	Distance (Å)
N3	Asp70 (OD2)	3.20
	W3 (HOH161) ^[a]	2.75 ± 0.15
N1	Arg85 (NH2)	3.30 ± 0.07
	W4 (HOH108)	2.46 ± 0.05
C ^[b]	Ser119 (CB)	3.60 ± 0.04
	Tyr58 (CE1)	3.50 ± 0.07

[a] Only present in two binding sites

[b] The distances were calculated from the nearest atoms in the ligand and the protein for hydrophobic contacts and π - π interactions. Mean distance and standard deviation were calculated from the distance of ligand and protein atoms in each binding site.

The crystal structure complex of BC2L-C-nt and the fragment KL3 showing binding at the site X is fully consistent with the predictions of binding pose obtained from STD-NMR experiment (**Figure 4.12**). The results indicated proximity of the benzylic methyleneamino group of the fragment (KL3) to aliphatic residues (Asp70 and Ser119 in the X-ray structure) of BC2L-C-nt. Similarly, the crystal complex showed that the methyl group of KL3 is located in the vicinity of Tyr58 side chain, hence responds to irradiation at 10 ppm in the STD experiment. In addition, the water mediated H-bond interactions with Globo H were identical to the previously studied complex.² The results of X-ray crystallographic screening finally validated the results of virtual screening (docking) and the ability of the site X (ligandability) to host the fragments.

4.6 Affinity analysis using ITC

The biophysical methods including X-ray crystallography confirmed the binding of KL3 at the expected site. Therefore, isothermal titration calorimetry (ITC) method was used to measure the binding affinity of KL3 for BC2L-C-nt.¹⁵ Before running the titration using protein and ligand, control measurements were recorded for the titration of fragment with the buffer (**Figure 4.20 A**). Subsequently, titration of the lectin by the fragment (KL3) was performed (**Figure 4.20 B**). The measurements showed small exothermic peaks after the correction for buffer mismatch (**Figure 4.20 C**). The integrated curve was fitted using one-site model with stoichiometry of one. The final fit determined the binding affinity (K_d) of 877 μM . However, the thermodynamic contributions could not be estimated due the low c-value of the experiment. The fragment titration was repeated using the same experimental setup that gave the similar results indicating that the fragment (KL3) binds at the expected site with a sub-millimolar binding affinity.

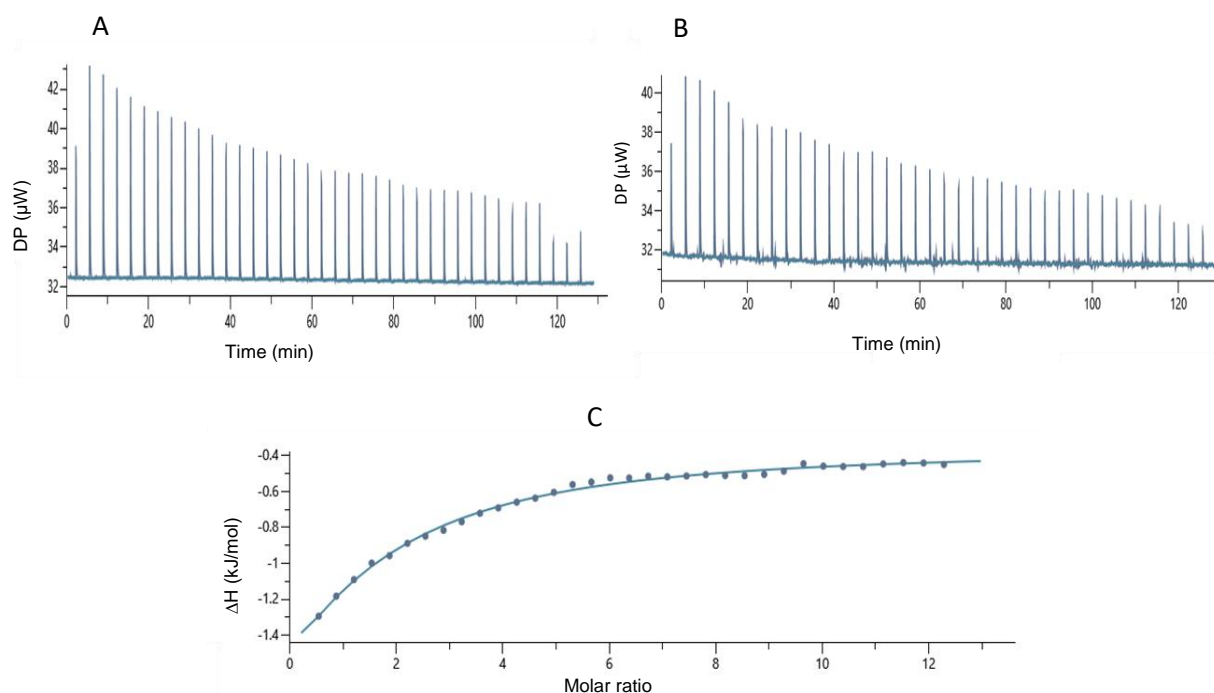


Figure 4.20 Isothermal microcalorimetry for the affinity analysis of KL3. (A) Titration of buffer by fragment (KL3, 15 mM) (B) Titration of BC2L-C-nt (225 μ M) by fragment (KL3, 15 mM) at 25 $^{\circ}$ C. (C) Final curve obtained after integration of peaks and point-by-point differences between fragment-in-buffer and fragment-in-protein for KL3. The fitting of curve was done using the “one binding site” model.

4.7 Summary

The biophysical methods used to investigate the binding of the fragments showed interaction/binding with the BC2L-C-nt. Remarkably, the structural complex of BC2L-C-nt with the fragment KL3 confirmed the druggability (ligandability) of the predicted site (site X) in the vicinity of the fucoside binding site. Thus, the experimental studies validated the results obtained from computational screening. The information from the screening results can be used further for structure-based design of high-affinity glycomimetic ligands by connecting the best fragments to the fucose core. This can be achieved using suitable linkers. Importantly, a robust synthetic route to glycomimetics will help in designing the high-affinity ligands against the target.

4.8 Experimental section

4.8.1 Protein expression and purification

The construct of BC2L-C-nt was designed as 132 amino acids-long with a cleavable histidine tag. The sequence coding for the 132 first amino acids of BC2L-C-nt amplified via PCR was inserted into pCold-TEV by Rafael Bermeo. The protein expression and purification was performed using the aforementioned vector. The vector was transformed by heat shock into *Escherichia coli* BL21 Star (DE3) cells. The bacterial cells with the plasmid were then cultured in Luria Broth (LB) medium containing 100 µg/mL ampicillin under constant shaking (170 rpm) at 37 °C. The temperature was decreased to 16 °C when the value of optical density of the culture (OD_{600nm}) reached 0.4. When OD_{600nm} reached 0.7, 0.1 mM isopropyl β -D-thiogalactopyranoside (IPTG) was added to induce the overnight protein expression. The cells were then centrifuged at room temperature for 5 minutes at 5000 x g and the pellets were obtained. These pellets can be used further or stored at -20 °C.

In the next step, wet cell pellets were resuspended in 5 mL of Buffer 1 (Tris-HCl 50 mM, NaCl 100 mM, pH 8.5) followed by treatment with DENARASE® endonuclease (c-LEcta GMBH, Leipzig, Germany) for a short duration (10 minutes) at room temperature while placed on a rotating wheel. The suspended cells were lysed by applying a pressure at 1.9 MPa in a one-shot table-top cell disruptor (Constant Systems Ltd., UK). The resulted lysate was centrifuged for 30 minutes at 24,000x g and 4 °C temperature. The supernatant was filtered through a 0.45 µm polyethersulfone (PES) syringe filter. The HisTrap™ fast flow (FF) 5ml column (GE Healthcare Life Sciences, Marlborough, MA, USA) was equilibrated with buffer 1 for affinity chromatography using NGC system (Bio-Rad, Marnes-la-Coquette, France) and then the filtered supernatant was loaded into the column. The unbound proteins are washed with buffer 1 and BC2L-C-nt was eluted using a 20 column volumes (CV) gradient of 0–500

mM imidazole. The eluted fractions were examined for the protein on 15% SDS-PAGE gel and the imidazole was removed from the pooled fractions using a PD10 desalting column (GE Healthcare Life Sciences, Marlborough, MA, USA). The protein was concentrated to at least 0.7 mg/mL by centrifugation (Vivaspin 3kDa, Sartorius, Goettingen, Germany) and treated overnight at 19 °C with TEV protease (1:50 w/w, enzyme:protein ratio), 1 mM ethylenediaminetetraacetic acid (EDTA) and 0.5 mM tris (2-carboxyethyl) phosphine (TCEP) for tag cleavage. The sample was again loaded into the affinity chromatography column and purification was repeated using the same conditions. This allows the separation of the desired protein (14 kDa) and the cleaved fusion (52 kDa) which could be assessed by SDS-PAGE 15 %. The protein was again concentrated by centrifugation and the concentration was determined by UV absorbance at 280 nm using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Illkirch-Graffenstaden, France). Finally, SEC was performed by employing an ENrich™ SEC 70 10 × 300 column (Bio-Rad, Marnes-la-Coquette, France) in a NGC™ systems (Bio-Rad Ltd.). The analytical column was pre-equilibrated with a buffer (20 mM Tris-HCl pH 7.0 and 100 mM NaCl) which was optimized for protein stability using TSA. The sample was injected using 240 µL volume with a flow rate of 1.0 mL/min. A column calibration curve based on gel-filtration standards (GE Healthcare, Life Sciences) was performed which helped in the calculation of protein molecular weight.

4.8.2 Thermal shift assay (TSA)

The fragments KL1, KL2, KL3, KL5, KL6, KL7, KL9, KL10, KL11 (**Table 3.2, Chapter 3**) were purchased from the Maybridge Company (Fisher Scientific International) and the other fragments KL4, KL8 and KL12 were purchased from abcr GmbH. The purity of the fragments was tested using liquid chromatography-mass spectrometry (LC-MS). For the dye-based TSA, BC2L-C-nt (5 µM) in assay buffer (20 mM Tris HCl, 100 mM NaCl, pH 8.0) was incubated with

50x SYPRO orange and 2.5 mM KL1-12 in the presence or absence of 20 mM Me- α -L-Fuc. A Qiagen Rotor-Gene Q instrument was used to apply a heat ramp of 1 °C/min from 25-95 °C and SYPRO orange fluorescence was monitored at 620 nm using the appropriate optical channel.

4.8.3 Microscale Thermophoresis (MST)

All the fragments (KL1-KL12) were tested using microscale thermophoresis. Prior to labelling, the protein was transferred to 0.1 M carbonate-bicarbonate, pH 9.5 buffer using buffer exchange column. The protein (lysine residue) labelling by the fluorescein isothiocyanate (FITC) dye was obtained at a ratio of 1.6 :1 (protein to dye ratio) as determined by UV spectroscopy. Tween-20 (0.2%) was also added to avoid protein sticking to the surface of tubes at low concentration. The protein at 326 nM concentration was used to test the fragments (2.5 mM).

4.8.4 STD-NMR interaction studies

STD-NMR studies were performed by Prof. Francesca Vasile at the University of Milan, Italy. More details related to the experiment can be found in the publication (**Appendix 1**).¹

4.8.5 X-ray crystallography, data collection, and structure determination

Crystals of BC2L-C-nt in complex with Globo H oligosaccharide were obtained following the procedure described previously.² Globo H hexasaccharide at 10 mM concentration in water was added to BC2L-C-nt (5 mg.mL⁻¹) for a ligand concentration of 1 mM. The mixture of protein and ligand (Globo H) was incubated for 30 minutes at room temperature (22 °C) and finally 2- μ L hanging drops containing 50:50 (v/v) mix of protein and reservoir solution (1.2–1.4 M tri sodium citrate pH 7.0) was used for crystallization using vapor diffusion method. Crystals were obtained in a few days. Cubic shape crystals of apo form were

obtained from the solution while complexes led to clusters of plates which were broken to single plates.

The fragments were tested for their aqueous solubility at high concentration and stock solutions were prepared. The globo-H complexes and apo form crystals were soaked overnight in solution containing 0.5 μ l volume of fragment solution (from stock) and 4.5 μ l of 2.5 M sodium malonate used for cryoprotection. This resulted in final concentration of 2 mM, for the fragments KL1, KL7 and KL11, 2.5 mM for KL12, 5 mM for KL2, KL5, KL6, KL8 and KL10 and 10 mM for KL3, KL4, KL9. For KL2 and KL12. 10 percent DMSO was added to achieve the above concentration. The crystals were flash-cooled in liquid nitrogen prior to data collection. The data was collected on the beamline Proxima 1, synchrotron SOLEIL, Saint Aubin, France, using an Eiger 16 m detector (Dectris, Baden, Switzerland). The data was processed using XDS and XDSME.¹⁶⁻¹⁷ The CCP4 suite was used for all further processing.¹⁸ The coordinates of the monomer A of PDB code 2WQ4 were used as search model to solve the structures of the apo form and the complexes with BC2L-C-nt by molecular replacement using PHASER.¹⁹ Refinement was performed using restrained maximum likelihood refinement and REFMAC 5.8²⁰ interspaced with using manual rebuilding in Coot.²¹ for cross validation, 5% of the data were set aside. Hydrogen atoms were added in their riding positions during refinement. Library for the fragment was made using ligand builder in Coot. All carbohydrates were validated using Privateer in CCP4i2 prior validation using the PDB validation server and deposition to the Protein Data Bank under code 7BFY for the apo form and 6ZZW for the complex.

4.8.6 ITC measurements

The ITC experiments were performed at 25 °C with an ITC200 isothermal titration calorimeter (Microcal-Malvern Panalytical, Orsay, France). The protein (BC2L-C-nt) and ligand

(KL3) were dissolved in the same buffer composed of 100 mM Tris HCl pH 7.0 and 100 mM NaCl. A total of 38 injections of 1 μ L of ligand solution (15 mM) were added at intervals of 200 s while stirring at 850 rpm was maintained to ensure proper mixing in the 200 μ L sample cell containing the protein, at 225 μ M. A control experiment was performed by injecting same concentration of KL3 in buffer. The subtraction of control for integrated peaks was performed using the Microcal PEAQ-ITC analysis software. The binding thermodynamics was further processed with a “one set of sites” fitting model. The experiment determined experiment affinity (K_d), binding enthalpy (ΔH) while the stoichiometry was fixed to 1. Free energy change (ΔG) and entropy contributions ($T\Delta S$) were derived from the equation $\Delta G = \Delta H - T\Delta S$. The experiments were performed in duplicates and the standard deviation was in 20% range for K_d .

4.9 References

- [1] Lal, K. *et al.*, Prediction and Validation of a Druggable Site on Virulence Factor of Drug Resistant Burkholderia cenocepacia, *Chem. Eur. J.* **2021**, *27*, 10341-10348.
- [2] Bermeo, R. *et al.*, BC2L-C N-Terminal Lectin Domain Complexed with Histo Blood Group Oligosaccharides Provides New Structural Information, *Molecules* **2020**, *25*.
- [3] Pantoliano, M. W. *et al.*, High-density miniaturized thermal shift assays as a general strategy for drug discovery, *J. Biomol. Screen.* **2001**, *6*, 429-440.
- [4] Cimmperman, P. *et al.*, A quantitative model of thermal stabilization and destabilization of proteins by ligands, *Biophys. J.* **2008**, *95*, 3222-3231.
- [5] Asmari, M. *et al.*, Thermophoresis for characterizing biomolecular interaction, *Methods* **2018**, *146*, 107-119.
- [6] Seidel, S. A. *et al.*, Label-free microscale thermophoresis discriminates sites and affinity of protein-ligand binding, *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 10656-10659.
- [7] Haselhorst, T. *et al.*, Saturation transfer difference NMR spectroscopy as a technique to investigate protein-carbohydrate interactions in solution, *Methods Mol. Biol.* **2009**, *534*, 375-386.
- [8] Davies, D. R. Screening Ligands by X-ray Crystallography, (Ed. W. F. Anderson), *Springer* New York, NY, **2014**, 315-323.

- [9] Haselhorst, T. *et al.* Saturation Transfer Difference NMR Spectroscopy as a Technique to Investigate Protein-Carbohydrate Interactions in Solution, Eds.: N. H. Packer and N. G. Karlsson), *Humana Press*, Totowa, NJ, **2009**, 375-396.
- [10] Hemmi, H., NMR analysis of carbohydrate-binding interactions in solution: an approach using analysis of saturation transfer difference NMR spectroscopy, *Methods Mol. Biol.* **2014**, 501-509.
- [11] Meyer, B.Peters, T., NMR spectroscopy techniques for screening and identifying ligand binding to protein receptors, *Angew. Chem. Int. Ed.* **2003**, *42*, 864-890.
- [12] Vasile, F. *et al.*, NMR interaction studies of Neu5Ac- α -(2,6)-Gal- β -(1-4)-GlcNAc with influenza-virus hemagglutinin expressed in transfected human cells, *Glycobiology* **2018**, *28*, 42-49.
- [13] Vasile, F. *et al.*, Diffusion-Ordered Spectroscopy and Saturation Transfer Difference NMR Spectroscopy Studies of Selective Interactions between ELAV Protein Fragments and an mRNA Target, *Eur. J. Org. Chem.* **2014**, *2*, 5.
- [14] Monaco, S. *et al.*, Differential Epitope Mapping by STD NMR Spectroscopy To Reveal the Nature of Protein-Ligand Contacts, *Angew. Chem. Int. Ed.* **2017**, *56*, 15289-15293.
- [15] Duff, M. R., Jr. *et al.*, Isothermal titration calorimetry for measuring macromolecule-ligand affinity, *J. Vis. Exp.* **2011**.
- [16] Kabsch, W., Xds, *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 125-132.
- [17] Legrand, P., XDSME: XDS Made Easier, *GitHub Repos.* **2017**, 2017.
- [18] Winn, M. D. *et al.*, Overview of the CCP4 suite and current developments, *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 235-242.
- [19] McCoy, A. J., Solving structures of protein complexes by molecular replacement with Phaser, *Acta Crystallogr. D Biol. Crystallogr.* **2007**, *63*, 32-41.
- [20] Murshudov, G. N. *et al.*, REFMAC5 for the refinement of macromolecular crystal structures, *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 355-367.
- [21] Emsley, P. *et al.*, Features and development of Coot, *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 486-501.

5. Design of bifunctional glycomimetic ligands

5.1 Strategies to connect fragments to the sugar core

The studies on fragment screening was performed with fucose present in the binding site with the aim of designing glycomimetic ligands by connecting the best fragments to the fucose core. The general idea for ligand design was to obtain the bifunctional molecules which can occupy the fucoside binding site as well as the site X simultaneously. To achieve this, selection of suitable linkers was made by considering the synthetic feasibility of the glycomimetic ligands and the possibility to maintain the binding pose at the site X. Since the best fragments were docked at 3-6 Å distance from the anomeric carbon of the fucose, the carbohydrate functionalization through the anomeric position was preferred to design glycomimetic ligands. In the bifunctional ligands, the fragments can make interactions with the hydrophobic residues in the site X that might contribute to counterbalance the inherent hydrophilicity of the sugar. Such molecules can further improve the selectivity and the binding-affinity. In order to maintain the expected binding pose of the glycomimetic ligands, different factors like rigidity, orientation and length of the linkers were considered to link the fragments to the fucose core. Finally, considering the synthetic feasibility, chemical linkers such as alkyne, amide, triazole and alkene moieties were considered suitable for the present studies (**Figure 5.1**).

5.2 A new model for the docking of fucoside-linker conjugates

In order to investigate the orientation of the chemical linkages, docking was performed for the several fucoside-linker conjugates in the sugar binding region. The crystal structure of BC2L-C-nt (PDB 2WQ4) was prepared using Maestro for docking as described previously under the fragment docking section (**Chapter 3, section 3.4.2**). Docking grid was

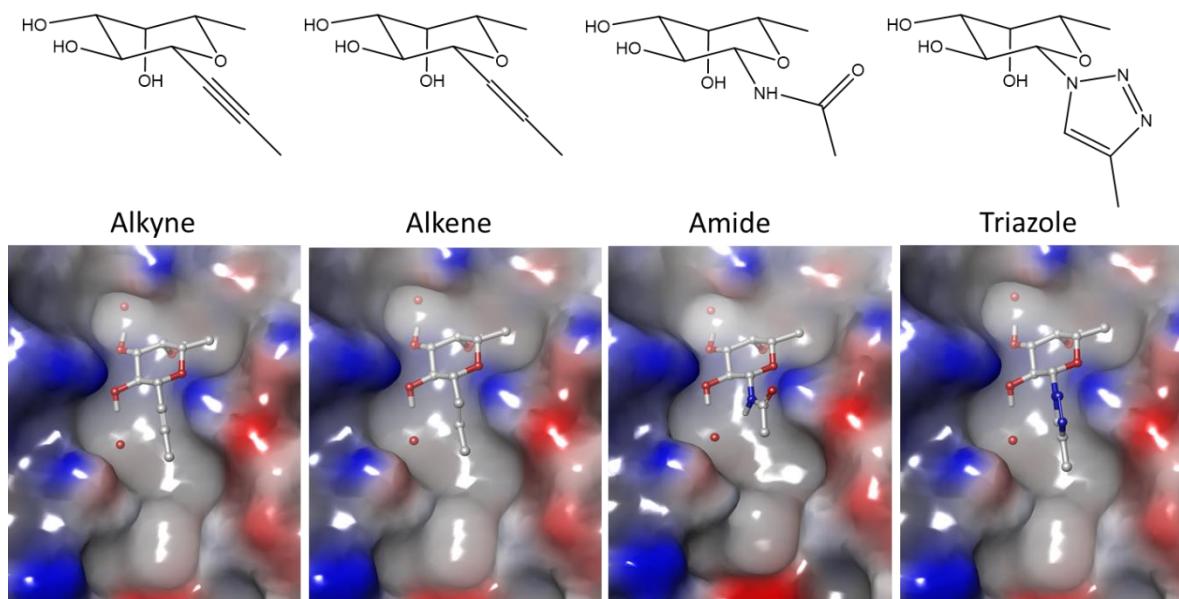


Figure 5.1 Strategies to link the fucose core to the selected fragments.

prepared without fucose but including two water molecules (W1 and W2). The centroid of fucoside was located in the active site between chain A and chain C in order to define a cubic grid box with dimensions 32×32×32 Å. Selenium atom in the fucoside of the crystal structure (MeSe- α -L-Fuc) was replaced by oxygen and then redocked at the sugar binding. The protocol reproduced the co-crystallized pose (RMSD 0.1 Å), hence validating the docking protocol using Glide (version 7.8).¹ The selected chemical linkers were then attached to the C-1 of fucose in β -configuration (equatorial) to be directed towards the hydrophobic pocket. The docking procedure allowed to determine their predicted preferred orientation and conformation.

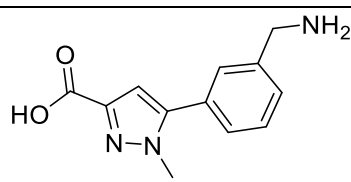
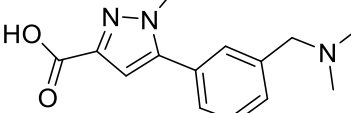
The docking results show that the alkyne function has the desired orientation in β -fucosylacetylene with an acceptable length (4.2 Å) required to connect the best fragments from the virtual screening in the site X. Similarly, amide linker (**Figure 5.1**) also appears interesting offering polar interactions with the structurally conserved water molecule W2. In addition, triazole function and alkene bond could be additional options for ligand design.

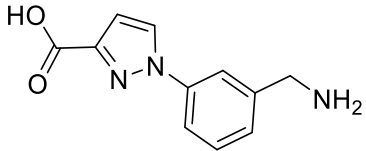
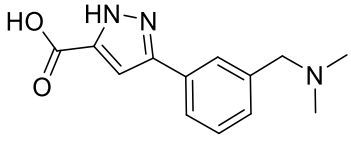
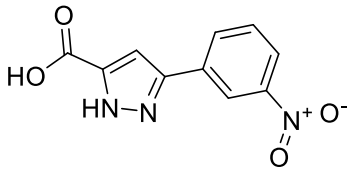
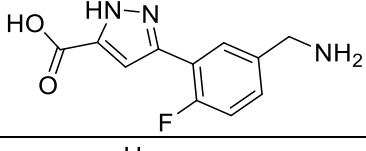
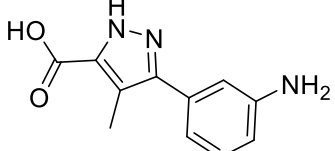
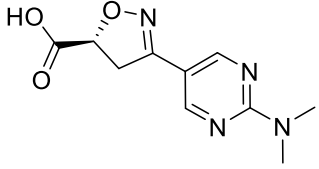
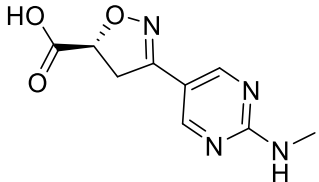
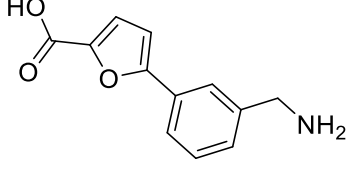
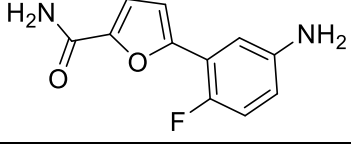
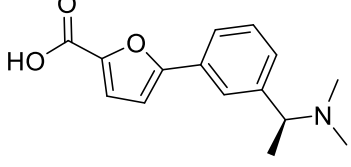
After selection of the linkers, strategies to chemically connect the fragments to the fucose core were designed. The grid defined above was further used for the docking studies of glycomimetic ligands by employing extra precision (XP) and standard precision (SP) scoring functions in Glide (version 7.8).¹

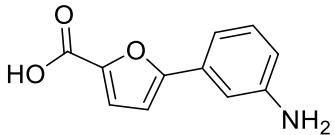
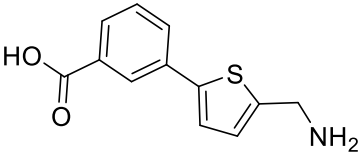
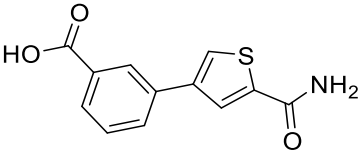
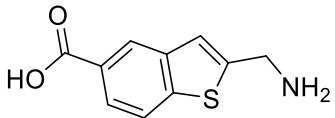
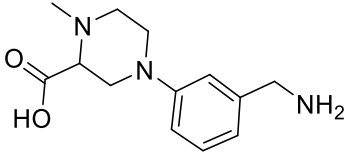
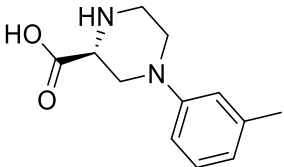
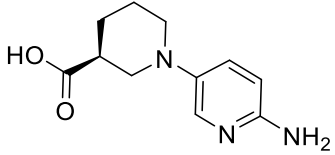
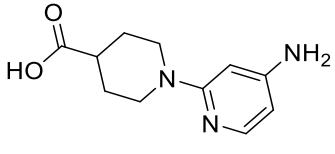
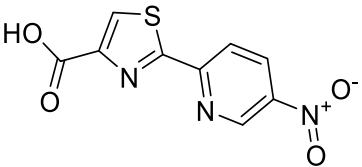
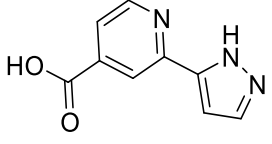
5.3 Selection and building of bifunctional glycomimetic ligands

The fragments of interest determined earlier (see **Table 3.2, Chapter 3**) were functionalized with additional groups (such as -COOH) and used as query structure for Tanimoto coefficient based similarity (95%) search to obtain the functionalized fragments from the PubChem database.² Some of the functionalized fragments were commercially available that would help to facilitate the synthesis of glycomimetic ligands in collaboration with Rafael Bermeo. For example, the fragments KL1 and KL2 provide the possibility to be linked to the fucose anomeric position with the amide linkage. Hence, the fragments similar to KL1 and KL2 with carboxyl and amide moieties have been identified (**Table 5.1**). These fragments were later used to design the glycomimetic ligands.

Table 5.1 List of the fragments with carboxyl and amide moieties identified from PubChem similarity (95 percent) search using COOH functionalized fragments KL1 and KL2 (see **Table 3.2, Chapter 3**) as query molecules. Fragments 1-15 are similar to the fragment KL2 and the fragments 16-20 are similar to the fragment KL1.

S.No.	PubChem ID	Structure
	Query molecule	
1	117402114	

2	91662832	
3	83395866	
4	2771732	
5	83410863	
6	91667363	
7	91670979	
8	91670957	
9	68633489	
10	91679354	
11	56709929	

12	1509177-63-2	
13	84756412	
14	56722394	
15	82610995	
	Query molecule	
16	84129598	
17	82019337	
18	17864765	
19	83309537	
20	57525825	

In **Table 5.1**, the fragment number 15 is likely to have low solubility due to fused ring system, therefore similar fragments with carboxyl functionalization have been identified using similarity search based on Tanimoto coefficient in the SciFinder (<https://scifinder-n.cas.org/>). The resulted fragments (**Annexe, Table 1**) have heterocyclic rings which can increase the polarity of the fragments and improve the solubility. The substitution at different positions in the fused ring system may also influence the binding affinity of the designed (glycomimetic) molecules. After selection of the linkers and functionalized fragments, the first generation of antagonists was designed as β -C- and β -N-fucosides using Maestro³ to target the fucoside binding site and the site X simultaneously in BC2L-C-nt.

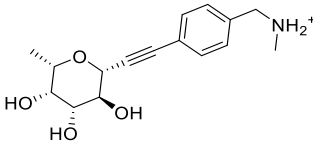
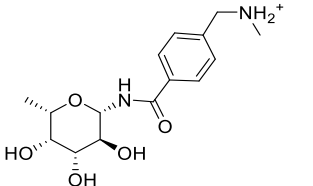
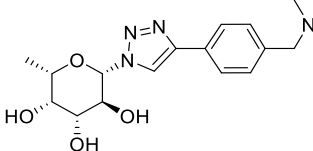
5.4 Docking studies of bifunctional glycomimetic ligands

All the designed bifunctional ligands were then studied for binding using docking studies. The ligands were prepared for docking using the LigPrep⁴ tool. The protonation states were generated at pH 7 ± 2 . The new docking model (discussed in previous section) was employed as the receptor grid for the target (BC2L-C-nt). The glycomimetic ligands (**Table 5.2 and 5.3**) designed using the best fragments from virtual screening were studied using XP and SP approaches in Glide.¹ Likewise, the ligands (**Table 5.4**) derived from the functionalized fragments resulting from similarity search (using KL1 and KL2 from **Table 3.2**) were also studied using two waters docking model and the XP and SP scoring functions. In addition, the glycomimetic molecules with heterocyclic rings (**Annexe, Table 2**) designed using the functionalized fragments were also docked using the two water model and the same (XP and SP) docking approaches.

The docking results for the molecules with alkyne and amide linkers indicate that the sugar and the non-sugar part of glycomimetic ligands establish interactions with the key residues already identified by the docking studies of the fragments. The main interactions

between the ligands and the protein involve π - π stacking with Tyr58 and the salt bridge or H-bond interactions between ammonium group (NH_3^+) and Asp70 (O) in the site X. Moreover, the amide linker forms additional H-bonding interactions with the structurally conserved water molecule W2. Thus, the expected binding pose with all the key interactions were maintained in the docking studies which was again confirmed by the analysis of multiple (10) binding poses of the ligands. The ligands with fused ring system involving different heterocyclic rings, do not display any significant difference in their docking scores (**Annexe, Table 2**) due to minor structural differences. These ligands have been classified into different subclasses based on the heterocyclic system of the non-sugar part (**Annexe, Table 2**).

Table 5.2 Docking studies of glycomimetic ligands designed using best fragments from two water model. These ligands were prioritized for synthesis in collaboration with Rafael Bermeo. The last column shows the results from experimental studies performed by Rafael Bermeo to calculate the binding affinity.

S.NO.	Molecule name	Structure	GlideScore (kcal/mol)		Binding affinity (K_d)
			XP	SP	
1	Lfuc-Aky-KL07		-9.6	-7.3	7.85 mM (SPR) 1.24 mM (ITC)
2	Lfuc-Amd-KL07		-9.7	-7.1	1.57 mM (SPR) 3.66 mM (ITC)
3	Lfuc-Trz-KL07		-10.0	-6.9	2.45mM (SPR) 6.25mM (ITC)

4	Lfuc-Ake-KL07		-9.4	-6.9	1.02 mM (SPR) 3.37 mM (ITC)
5	Lfuc-Aky-KL08		-9.7	-6.9	1.33 mM (SPR) 281 μM (ITC)
6	Lfuc-Amd-KL08		-9.6	-7.0	0.94 mM (SPR) 2.55 mM (ITC)
7	Lfuc-Trz-KL08		-9.8	-6.9	1.19 mM (SPR) 2.49 mM (ITC)
8	Lfuc-Ake-KL08		-9.8	-6.8	tbs
9	Lfuc-Amd-I1		-9.4	-7.1	3.42 mM (SPR) 3.49 mM (ITC)
10	Lfuc-Amd-I2		-9.5	-7.2	1.42mM (SPR)
11	Lfuc-Aky-KL03		-10.4	-7.0	tbs

12	Lfuc-Amd-KL13*		-9.6	-7.6	2.36 mM (SPR)
----	----------------	--	------	------	------------------

tbs = to be synthesized * Designed using the fragment from similarity search (see Table 5.1)

Table 5.3 Docking studies of glycomimetic ligands using one water model (KL9-KL12). These molecules can be synthesized in future.

S.No.	Molecule name	Structure	GlideScore (kcal/mol)	
			XP	SP
1	Lfuc-Aky-KL09-RR		-8.9	-6.4
2	Lfuc-Aky-KL09-SR		-9.5	-7.1
3	Lfuc-Aky-KL10-RR		-9.2	-5.9
4	Lfuc-Aky-KL10-SR		-9.8	-7.3
5	Lfuc-Aky-KL11-RR		-9.1	-6.0
6	Lfuc-Aky-KL11-SR		-9.0	-7.7

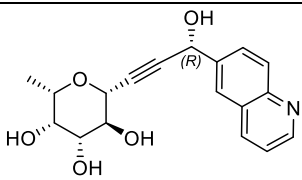
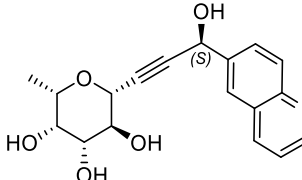
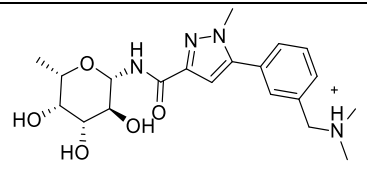
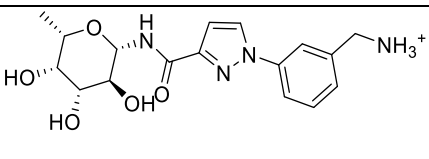
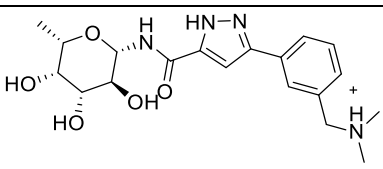
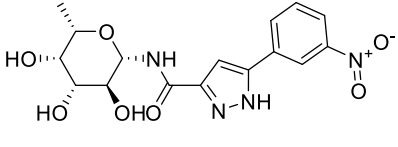
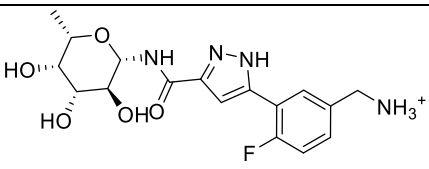
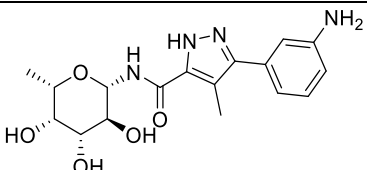
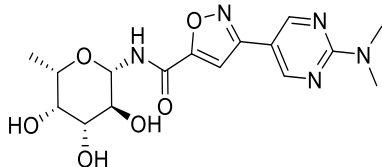
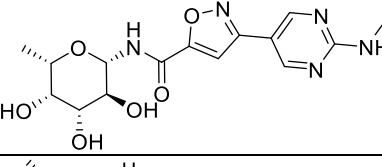
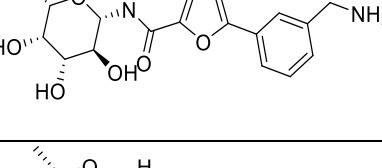
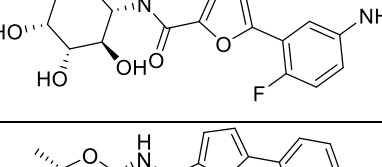
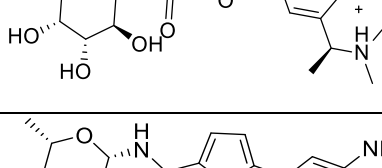
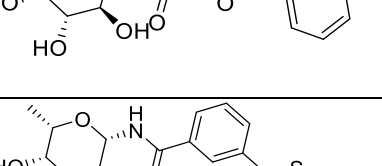
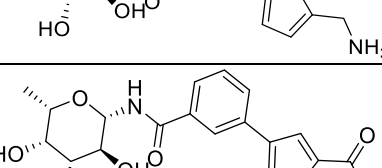
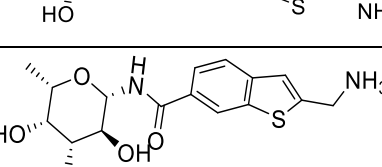
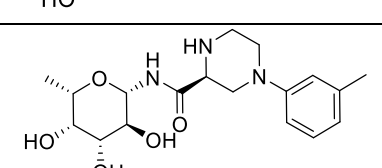
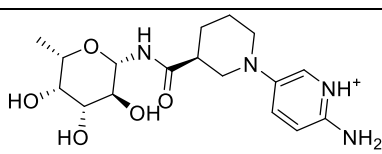

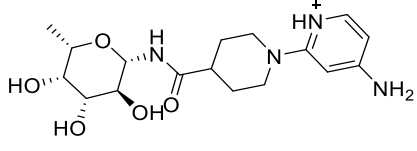
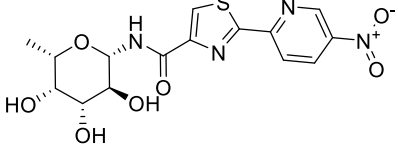
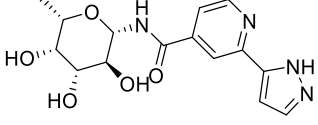
7	Lfuc-Aky-KL12-R		-8.7	-6.2
8	Lfuc-Aky-KL12-S		-8.7	-6.1

Table 5.4 Docking studies of the ligands derived from -COOH functionalized fragments (see **Table 5.1**). The fragments were obtained from similarity search (using KL1 and KL2) in PubChem database and connected to the fucoside using an amide linker.

S.No.	PubChem ID of the fragment	Structure	GlideScore(kcal/mol)	
			XP	SP
1	117402114		-10.18	-7.10
2	91662832		-10.31	-7.54
3	83395866		-10.60	-7.15
4	2771732		-9.86	-7.15
5	83410863		-10.96	-7.51
6	91667363		-9.94	-7.15

7	91670979		-9.02	-6.46
8	91670957		-9.49	-6.51
9	68633489		-9.98	-7.25
10	91679354		-10.07	-7.76
11	56709929		-9.94	-6.89
12	1509177-63-2		-9.60	-7.63
13	84756412		-9.35	-7.02
14	56722394		-9.17	-7.38
15	82610995		-10.28	-7.28
16	84129598		-9.73	-6.75
17	82019337		-9.94	-6.81

18	17864765		-9.03	-6.76
19	83309537		-8.90	-6.77
20	57525825		-9.56	-7.12

After docking studies of the designed glycomimetic ligands, fragment 12 (-COOH functionalized fragment) in **Table 5.1** was purchased from a commercial vendor. Other fragments from this extended set as well as fragments with heterocyclic rings (**Annexe, Table 2**) were not considered for purchasing due to their high cost or non-availability with the vendors. However, they can be considered for design and in-house synthesis of glycomimetic molecules in future.

The ligands with alkyne and amide chemical linkers appear more interesting as they are more rigid to direct the ligands towards site X and, in particular, the amide linker establishes additional interactions with water molecule (W2). Hence, some of the prioritized ligands with alkyne and amide linkers were further studied using MD simulations (discussed later) to evaluate the conformational behaviour of the ligand-protein complexes, that could be compared with the results of molecular docking studies and with experimental results. The glycomimetic molecules (**Table 5.2**) designed based on the best fragments from the initial fragment screening were selected for synthesis in collaboration with Rafael Bermeo. After synthesis, the glycomimetics were further tested for their binding using a series of biophysical methods such as isothermal titration calorimetry (ITC), surface plasmon resonance (SPR) and

X-ray crystallography (discussed in next chapter). The values of binding affinity (K_d) for some of the synthesized ligands are given in **Table 5.2**.

Most of the affinities were measured equivalent to the binding affinity of the monosaccharide (Me- α -L-Fuc) in the millimolar range. Nevertheless, the two alkyne-bound ligands Lfuc-Aky-KL07 and Lfuc-Aky-KL08 showed better binding affinity than the monosaccharide with nearly 2-fold and 9-fold improvement, respectively. The results indicate that the use of alkyne linker could be an efficient approach to design the bifunctional ligands with a non-sugar part binding at the site X in BC2L-C-nt.

5.5 MD simulations of BC2L-C-nt complexes with glycomimetics

After docking studies, molecular dynamics (MD) simulation can be employed to further investigate the key residues or interactions involved in ligand binding. The static view of ligand-receptor interactions provided by docking studies can be verified by performing MD simulation studies of ligand-receptor complexes in solution with ions. Thus, the studies can be performed using a setup much closer to real physiological condition. The results can provide insights into docking studies and help rationalizing experimental binding data. The combination of two *in silico* techniques (docking and MD) could improve the reliability of the results.

5.5.1 Force field parameterization of glycomimetic ligands

Based on the docking studies and the feasibility of synthesis using suitable chemical linkers (alkyne and amide), three bifunctional glycomimetic ligands, Lfuc-Aky-KL07, Lfuc-Aky-KL08 and Lfuc-Amd-KL13, were prioritized for MD simulations in complex with BC2L-C-nt. The results can be compared with the docking studies by analyzing the major contacts during the MD. This can further rationalize the protein-ligand interactions.

To accomplish the calculations using MD simulations, partial atomic charges and force field parameters for the non-sugar part of the ligands were required. Therefore, PyRED server⁵ was used to derive the partial atomic charges for the aglycone moiety of the glycomimetic ligands. PyRED derives charges by restrained electrostatic potential (RESP) fit at the HF/6-31G(d) level using Gaussian16.⁶ The program optimizes the input molecular structures using the Gaussian program and carries out single point energy calculations. The minimized structure is used to compute the corresponding molecular electrostatic potential (MEP). Subsequently, the RESP program is executed to fit the atom-centred charges to the MEP. Finally, mol2 files with derived partial charges are generated (**Annexe, Table 3 A, B, C**). The program can generate charges compatible with the AMBER force fields.⁷ **Figure 5.2** illustrates the schematic representation of PyRED workflow.

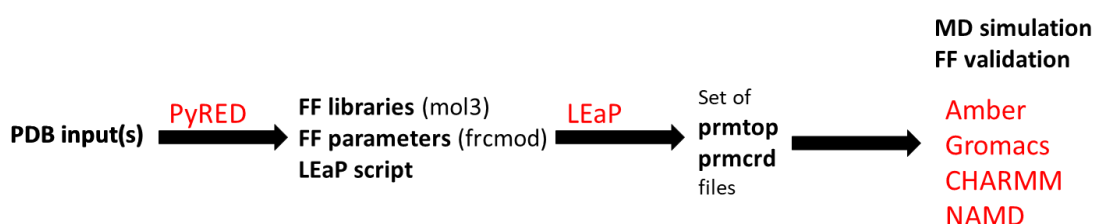


Figure 5.2 Schematic representation of PyRED workflow.

The parameters for the bond, angles and torsion of the aglycone linkage were obtained from the AMBER force field (GAFF2). Finally, the charges derived by PyRED server were employed for generating the input files for MD simulations using AMBER (version 18) biomolecular simulation package.⁸

5.5.2 MD simulation setup

The X-ray crystal structure of the N-terminal domain of the lectin in complex with methylseleno- α -L-fucopyranoside (PDB code 2WQ4) was prepared for MD simulations. Initially, the complex with the fucoside in all three binding sites of the trimer was prepared for MD simulation. In another setup, the system involved BC2L-C-nt trimer in complex with

the glycomimetic molecule (generated using docking) in the binding site between chain A and C while maintaining the co-crystallized fucosides (methylseleno- α -L-fucopyranoside) in the two other binding sites. Selenium in the fucosides was changed to oxygen linked to anomeric carbon to create Me- α -L-Fuc. The AMBER force field ff14SB⁹ and GLYCAM06-j¹⁰ parameters were used for protein and fucoside respectively. The charges derived using PyRED server and the force field parameters from AMBER force field (GAFF2) were used for the non-sugar part of the molecules to generate input files for the MD simulations. Missing side chains and hydrogen atoms were added using tleap program of AMBER18.⁸ The first conformation of the residues was retained for the alternate side chain conformations. The ligand-protein complex was solvated by adding TIP3P¹¹ water molecules to fill a truncated octahedral box extending 12 Å from the solute and with a van der Waals closeness parameter set to 0.7. Sodium and chloride ions were added to neutralize the system at the physiological concentration of 150 mM. Hydrogen mass repartitioning was performed on the topology file using parmed.¹² This allowed 4 fs time step for MD simulation run. The system was minimized for 3000 cycles using the steepest descent (1000 cycles) and conjugate gradient (2000 cycles) algorithms and then subjected to 100 ps MD in NVT (canonical) ensembles to gradually raise the temperature to 300 K. The system was equilibrated for 500 ps stepwise in isothermal-isobaric (NPT) ensemble with decreasing constraints on the backbone of 50, 25, 10, 5, 2.5, and 1 kcal mol⁻¹Å⁻². MD was run in periodic boundary conditions using the PMEMD module of AMBER18. Nonbonded interactions were treated with cutoff radius of 9.0 Å. 600 ns and 1 μ s production ensued for the complexes with Me- α -L-Fuc and glycomimetics, respectively at 300 K and 1 atm. The temperature and pressure was controlled using the Langevin thermostat¹³ and Berendsen barostat¹⁴, respectively. Frames were saved for both the complexes: every 200 ps for Me- α -

L-Fuc complex and every 40 ps for glycomimetic complex. Finally, a part of the total saved frames was analysed.

5.5.3 Molecular dynamics simulations of methyl α -L-fucoside (Me- α -L-Fuc) in complex with BC2L-C-nt

The 600 ns second simulation demonstrated that the protein backbone of BC2L-C-nt trimer was stabilized (RMSD 1 Å) during the MD simulation (**Annexe, Figure 1**). The ligand formed key interactions with the residues in the binding site and similar to the co-crystallized conformation, the anomeric position of the ligand was oriented towards the site X. The analysis of ligand also showed that hydrogen bond interactions with key residues in the binding sites were maintained in the sampled structures. The frames in the trajectory were used to calculate the occupancy (average occupancy) of H-bonds between the individual residue pairs of the ligand and the receptor. The hydrogen bonds involving the key residues displayed an overall occupancy of 85 to 100 percent in all the three sites (**Figure 5.3 and 5.4**). Hydrogen bonding interactions between fucoside and side chains of Arg85 (chain A) and Arg111 (chain C) were stabilizing the fucoside at the interface between monomers. Similarly, H-bond interactions between backbone atom (O) of Thr83 (chain A) and side chain of Thr74

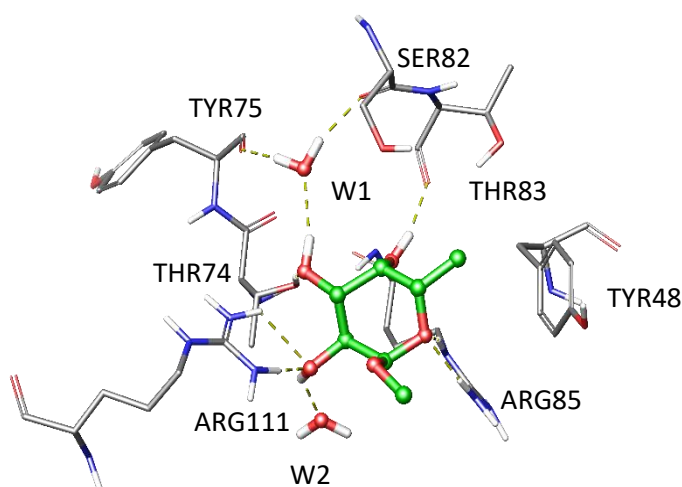


Figure 5.3 A representative snapshot from MD simulation trajectory of BC2L-C-nt and fucoside complex showing key H-bond interactions. Hydrogen bonds are represented as dashed lines.

(chain C) also appear to play important role in ligand binding. The multiple H-bonding interactions ultimately resulted a very low RMSD (0.1 Å) for the fucoside in all the binding sites in the MD simulation performed for 600 ns. Moreover, the key water molecules (W1 and W2) also showed bridging H-bonds between the ligand and the protein. Thus, the water molecules at these two sites can be important for the ligand binding.

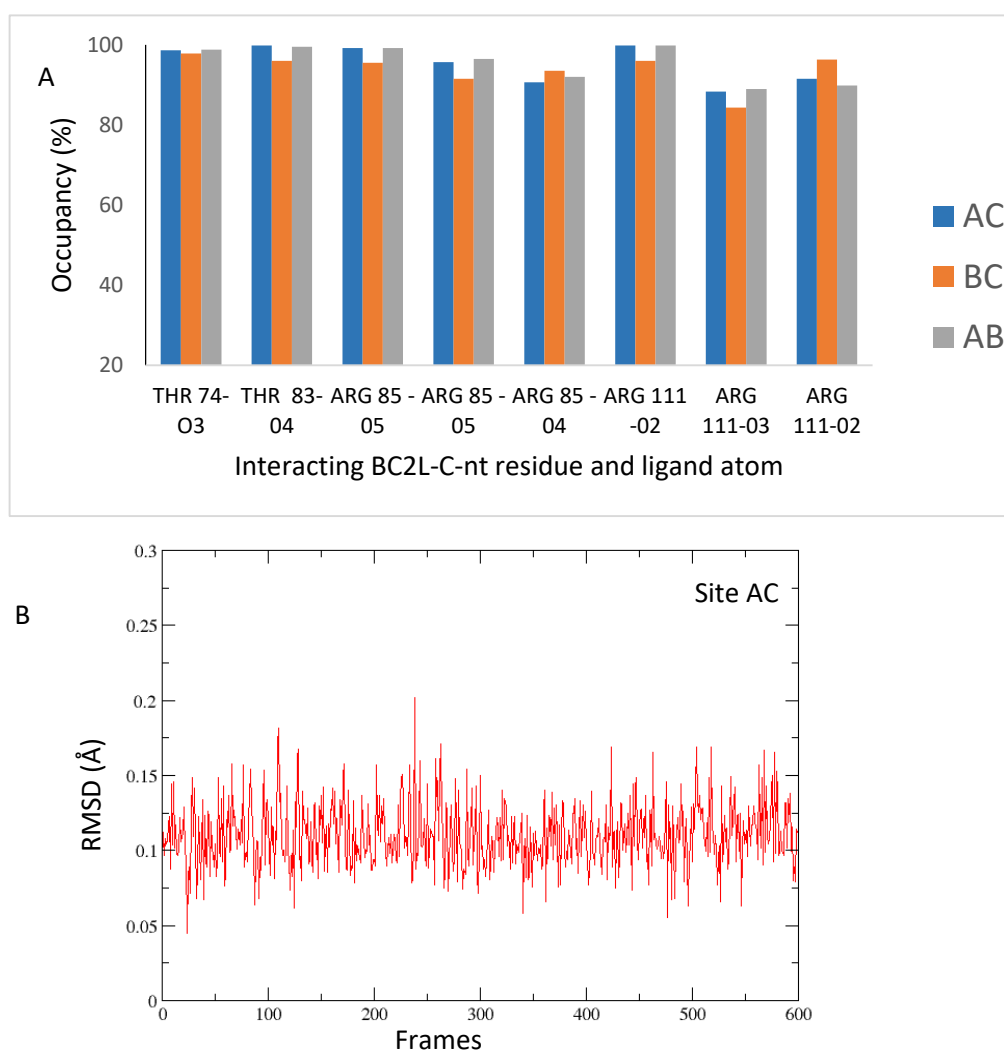


Figure 5.4 (A) Hydrogen bond interaction pattern (showing percentage occupancy) obtained from the MD simulation of BC2L-C-nt and fucoside complex. AB, BC and AB are the binding site interfaces of the monomers in BC2L-C-nt trimer. The interactions with arginine shown (twice) in the plot involve atoms of -NH1 and -NH2 groups of the side chain of the residue. (B) The RMSD calculation for the ligand (fucoside) binding at the interface of chain A and C. The results show a small value of RMSD which demonstrates that the ligand maintains key interactions in the binding site.

5.5.4 MD simulations of glycomimetic ligands Lfuc-Aky-KL07 and Lfuc Aky-KL08 in complex with BC2L-C-nt

The analysis of the trajectory (1 μ s) for the BC2LC-nt complex with Lfuc-Aky-KL07 shows that the designed ligand is maintained in both sites: fucose binding and the site X. Values of RMSD of 0.5 Å for sugar and 1.5 Å for the linked non-sugar (KL7) part (**Annexe, Figure 2**) are obtained. The key interactions of fucose ring are maintained. For the non-sugar part of the ligand, H-bond with Asp70 and π - π stacking with the Tyr58 in the binding sites were maintained during the MD simulation. In addition, the ammonium group (-NH₂⁺) orients towards Ser119 and forms H bond interactions with both Asp70 and Ser119 (**Figure 5.5**).

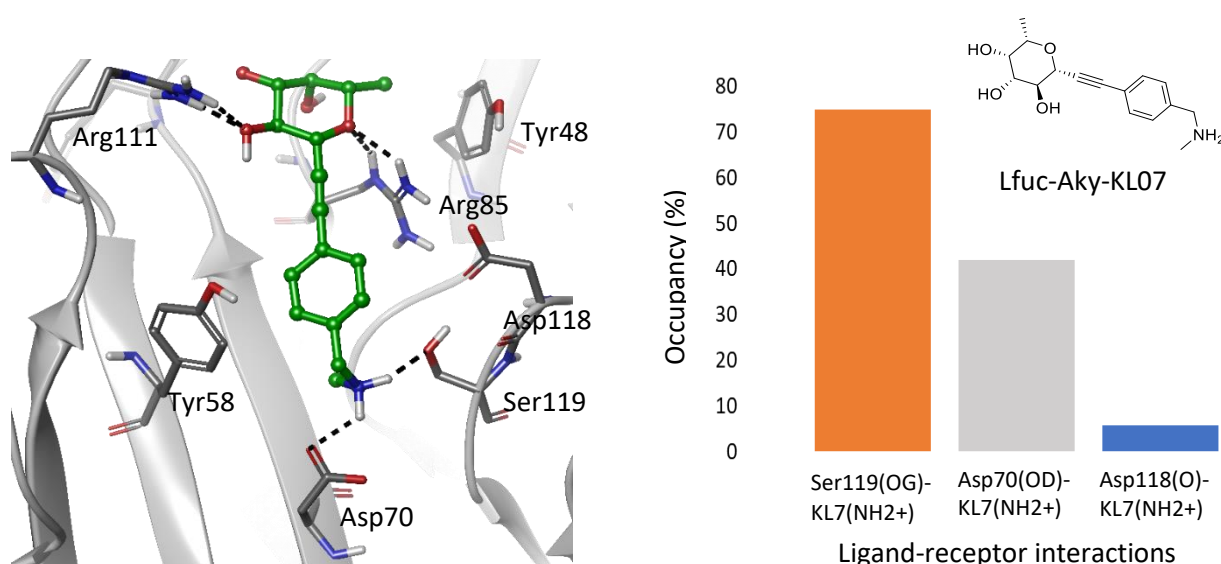


Figure 5.5 (A) A representative snapshot from the MD simulations of Lfuc-Aky-KL07 in complex with BC2L-C-nt showing key residues involved in ligand binding. The H bond interactions are shown in dashed lines. (B) Hydrogen bond interaction pattern (showing percentage occupancy) obtained from the MD simulation of Lfuc-Aky-KL07 in complex with BC2L-C-nt.

This interaction with Ser119 was not observed in the docking studies. Thus, the results from MD simulation of the complex with the ligand provides important information for the design of new ligands that can establish additional interactions with Ser119. The analysis of Lfuc-Aky-KL08 also provides similar results confirming that the ligand maintains key interactions in

agreement with the docking studies. Like Lfuc-Aky-KL07, the ammonium group (NH₃⁺) in the ligand establishes additional interactions with Ser119 (**Figure 5.6**). However, due to the

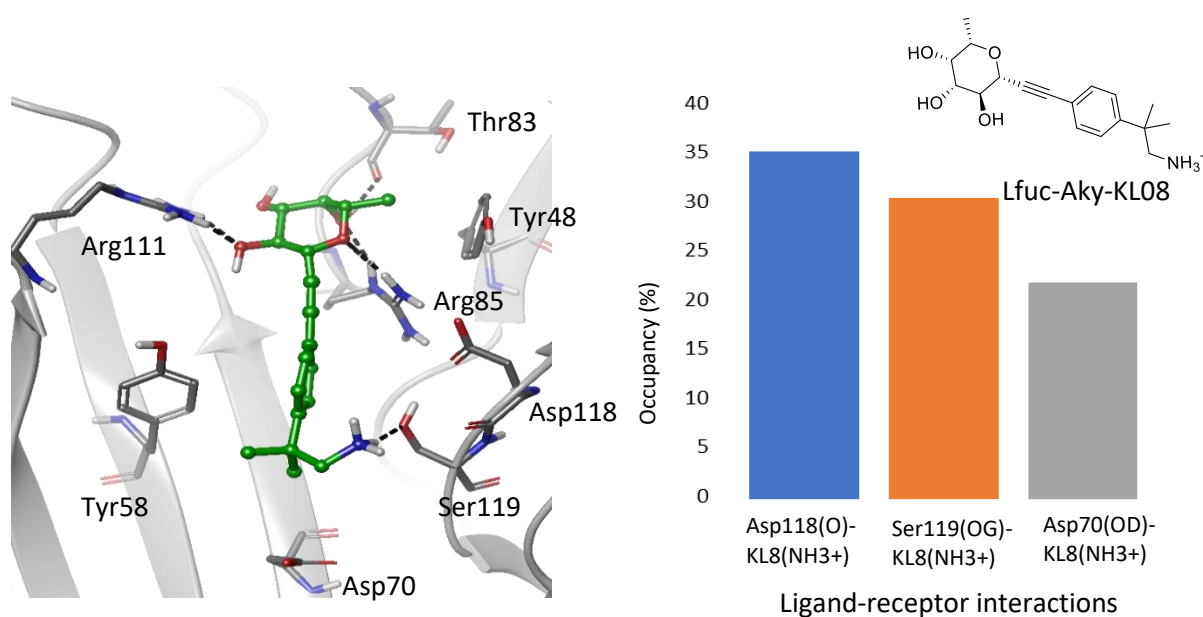


Figure 5.6 A representative snapshot from the MD simulations of Lfuc-Aky-KL08 in complex with BC2L-C-nt showing key residues involved in ligand binding. The H bond interactions are shown in dashed lines. (B) Hydrogen bond interaction pattern (showing percentage occupancy) obtained from the MD simulation of Lfuc-Aky-KL08 in complex with BC2L-C-nt.

presence of two methyl groups occupying the site X, the non-sugar part is pushed towards Asp118 and forms H-bonding interactions. Consequently, a higher RMSD (2.0 Å) was observed for the non-sugar part of the ligand (**Annexe, Figure 3**). Similar to Lfuc-Aky-KL07, the sugar part of Lfuc-Aky-KL08 displayed a lower RMSD (0.5 Å). The results from computational studies thus indicated that the ligands designed using alkyne linker might be interesting for the synthesis and ligand binding studies.

5.5.5 MD simulation of glycomimetic ligand Lfuc-Amd-KL13 in complex with BC2L-C-nt

In addition to the MD simulation studies of the ligands with alkyne linkers, MD simulation were performed for the docking complex of BC2L-C-nt with Lfuc-Amd-KL13. The ligand consists of fragment (KL13) connected through an amide linker. The sugar part of the

ligand maintained predicted binding pose with an RMSD of 0.15 Å while the non-sugar part displayed a lower RMSD (1.0 Å) than the other two ligands (**Annexe, Figure 4**). The non-sugar part containing aniline moiety maintained the key interactions in the site X (**Figure 5.7**) similar to other ligands with benzylamine moiety. However, aniline moiety lacks salt bridge interactions with Asp70. Non-sugar part of the ligand also shows alternate binding modes (in less than 10 percent sampled structures) with rotation of aniline ring. These additional

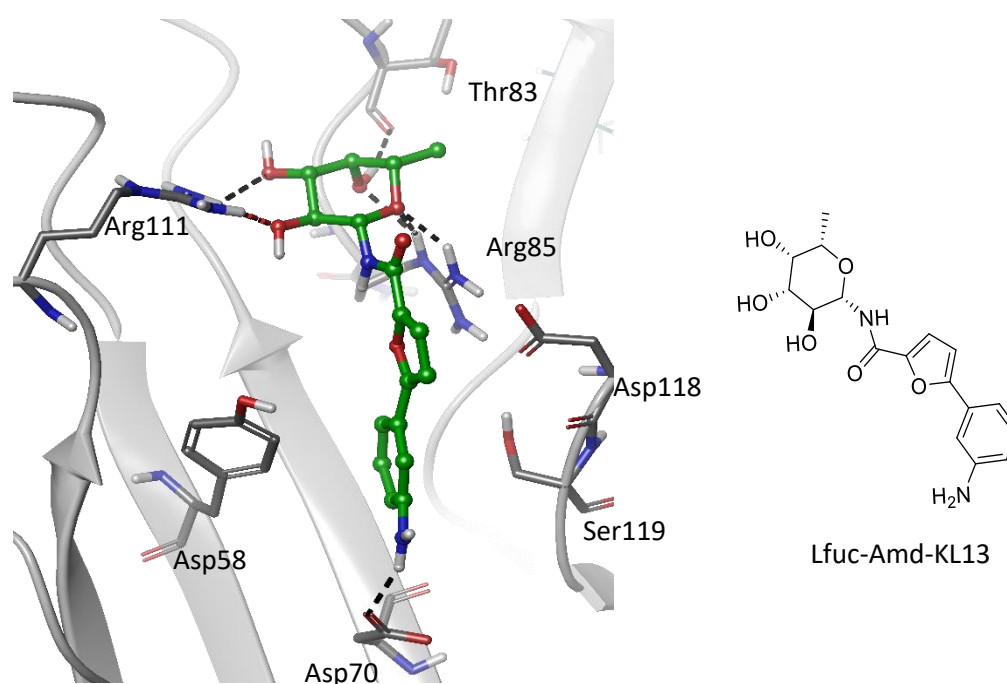


Figure 5.7 A representative snapshot from the MD simulations of Lfuc-Amd-KL13 in complex with BC2L-C-nt showing key residues involved in ligand binding. Most of the binding poses in trajectory maintain Asp70(OD)-KL13(NH₂) interactions. Other H-bond interactions were observed in less than 10 percent of sampled structures.

binding poses indicates H-bond interactions with Ser119. The results from computational studies for the ligand with amide conformed the possibility of ligand binding at the site X. Therefore, in addition to the ligands designed using alkyne linkers, it was also prioritized for synthesis and ligand binding studies with the help of collaborators.

5.6 Experimental validation of glycomimetic binding using ITC and X-ray crystallography

Synthesis and evaluation of the designed molecules (glycomimetics) against BC2L-C-nt was performed by Rafael Bermeo. The ligands were first tested for their binding using STD NMR and then later further characterized for the interactions and binding affinity using surface plasmon resonance (SPR) and isothermal titration calorimetry (ITC). The results indicated that the ligands bind with different binding affinities (K_d) ranging from millimolar to micromolar (**Table 5.5**).

Table 5.5 Comparison of the binding affinities determined using ITC for monosaccharide (Me- α -L-Fuc) and the designed glycomimetics (Lfuc-Aky-KL07 and Lfuc-Aky-KL08). The binding affinity of Lfuc-Amd-KL13 was estimated using SPR, therefore required further evaluation using ITC.

Compound/Glycomimetics	Binding affinity (K_d)
Monosaccharide (Me- α -L-Fuc)	2.4 mM
Lfuc-Aky-KL07	1.24 mM
Lfuc-Aky-KL08	281 μ M
Lfuc-Amd-KL13	2.36 mM

The alkyne-bound molecules appear strongest binders, with molecule Lfuc-Aky-KL08 showing almost 9-fold increase in affinity when compared to the monosaccharide (e.g Me- α -L-Fuc). Moreover, crystallographic complexes (PDB not released) with two glycomimetic ligands: Lfuc-Aky-KL08 and Lfuc-Amd-KL13 were also solved by Rafael Bermeo. Lfuc-Aky-KL08 consists of fragment KL08 connected via alkyne linker while Lfuc-Amd-KL13 is composed of fragment KL13 and an amide linker. The complexes of both the ligands (**Figure 5.8 and 5.9**) show that the co-crystallized binding poses of the ligands with all the key interactions were matching the binding poses predicted by docking and MD simulations studies. The top ranked binding

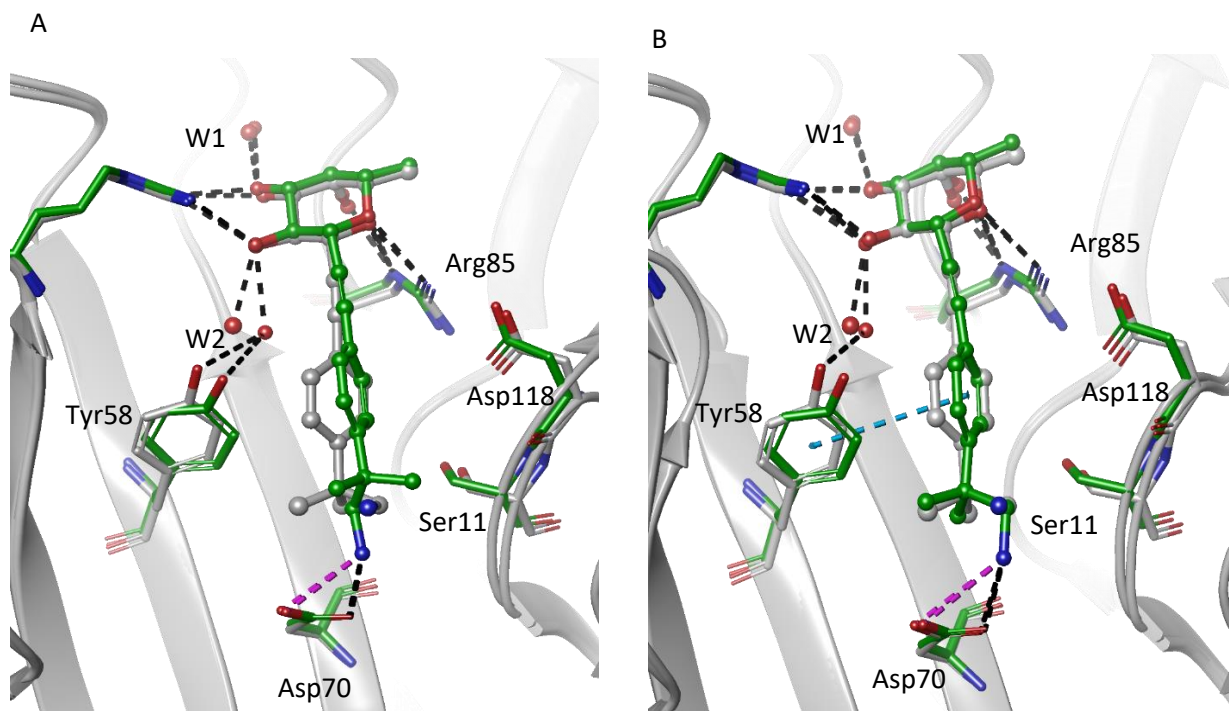


Figure 5.8 Superimposition of docked (green) and co-crystallized (grey) complex of Lfuc-Aky-KL08.

(A) The best pose from docking studies shows difference in the orientation of the methyl groups while (B) the pose ranked third in the docking studies shows orientation almost identical to the crystallized conformation of the ligand.

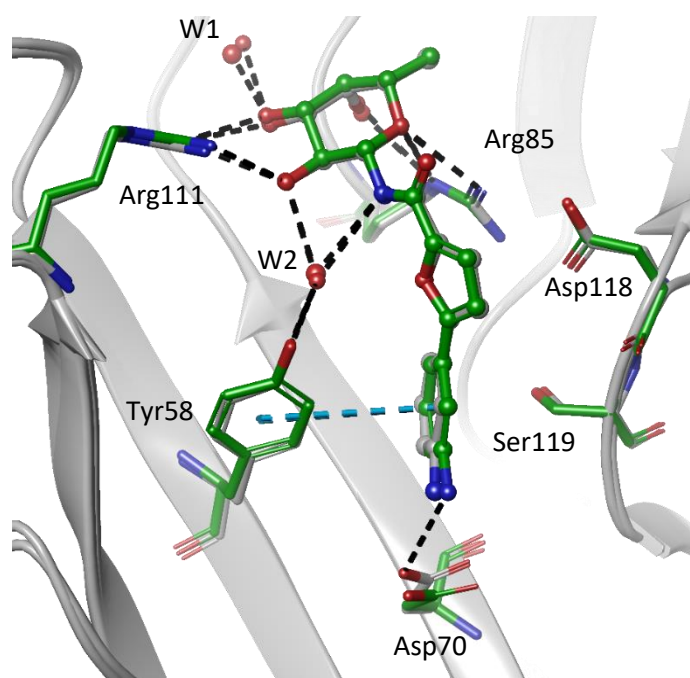


Figure 5.9 Superimposition of docked (green, best pose) and co-crystallized (grey) complex of Lfuc-Amd-KL13.

pose from docking predictions shows RMSD 0.5 Å and 0.3 Å with the co-crystallized complex

of the Lfuc-Aky-KL08 and Lfuc-Aky-KL13, respectively. Thus, these experimental studies further confirmed the druggability of the binding site and also validated the *in silico* ligand design for the BC2L-C-nt. The two ligands in the co-crystallized complexes and compound Lfuc-Aky-KL07 from ITC studies can be used as 'hit' compounds for further optimization studies towards a 'lead' structure aiming functional inhibition of BC2L-C-nt. Thus, these molecules were further used to design new molecules with additional substitution in the existing structures that might improve their binding affinity.

5.7 Strategies to improve binding affinity of the glycomimetics

The binding of a macromolecule to another macromolecule or a drug-like small molecule can cause displacement of surface water to bulk that can contribute to the overall free energy of binding.¹⁵⁻²⁰ The carbohydrate binding site in lectins usually consist of water molecules which established bridging H-bonding interactions with protein and the ligand. Therefore, the study of the water at molecular surfaces can be crucial in molecular recognition and ligand binding. The water molecules at the hydrophobic surfaces have unfavourable effects, due to combination of entropic and enthalpy costs, and that allows their displacement to bulk. The water-protein interactions in a binding pocket display a pattern which is strongly replicated by the binding of small molecules.²¹ Hence, these hydration sites in the binding pocket can provide valuable information about the key features that a molecule could mimic to favour its binding to the target.

In addition to the analysis of water molecules, the analysis of the region in the vicinity of the binding site can guide towards the structure optimization of the existing ligands to improve their binding affinity. The substitution in the ligand by a suitable chemical group could help to establish interactions with the additional residues in the binding site, thus can improve the ligand efficiency and binding affinity.

5.8 Analysis of water thermodynamics at binding site using Grid Inhomogeneous Solvation Theory (GIST)

Water molecules play an important role in the free energy of molecular recognition.²²⁻
²⁴ High-resolution crystal structures of ligand-protein complexes have demonstrated that water molecules frequently mediate protein-ligand interactions.²⁵ The studies have shown that the hydrogen bonds formed between the ordered water molecules and ligands can have significant impact on the ligand specificity and binding affinity.²⁶⁻²⁸ The displacement of high-energy water molecules from the binding site surface to the bulk solvent can improve the binding of the ligands by several folds.²⁹⁻³⁰ Therefore, introduction of substituents in the ligands to displace these water molecules can show remarkable improvements in the binding affinity.³¹⁻³² Hence, analysis of solvation thermodynamics in protein-ligand binding studies could be an efficient approach in structure-based drug design that can contribute towards ligand optimization to enhance the binding affinity.²⁹⁻³⁰

Grid inhomogeneous solvation theory (GIST)³³ method calculates thermodynamic values of solvent located within a defined region using the grid discretized inhomogeneous solvation theory formulation and produces quantitative thermodynamic data for each grid box or voxel. This calculation provides a detailed map of water thermodynamic and structural quantities in a defined region of interest including: interaction energy of water found in a voxel with all other water molecules or solute molecules and water occupancy within each voxel. The information can be used to decide whether the water at a given location is favourable or not compared to the bulk distribution.

For the calculations of thermodynamic properties of the solvent near glycomimetic binding site, the docked complex of BC2L-C-nt with glycomimetic molecule (Lfuc-Aky-KL07)

was solvated (using TIP3P water model) in a cubic box and topology and coordinate files were generated (using tleap) as input for MD simulation. Protein was restrained ($200 \text{ kcal}/\text{\AA}^2$) to avoid translation or rotation and only water was minimized for 20000 cycles using steepest descent (1500 cycles) and conjugate gradient algorithms. Next, water and the protein hydrogen atoms were minimized to another 20000 cycles. System was heated to 50 K while restraining the protein heavy atoms in NVT (canonical) ensembles, followed by increment from 50 K to 300 K for 200 ps. The system was relaxed restraining the protein heavy atoms in NPT ensemble for 10 ns followed by restraining all atoms in the protein at 300 K for 5 ns in NVT ensemble. MD simulation was run using PMEMD module of AMBER18 in NVT ensemble for 30 ns sampling every 1 ps frame.

Analysis of trajectory using GIST post processing (GISTPP)³³ tool identified three regions (A, B, C) with higher occupancy (**Figure 5.10**) of water molecules within 3.5 \AA from the ligand. The grid points in these regions were grouped and thermodynamic properties of the sites were calculated (**Table 5.6**).

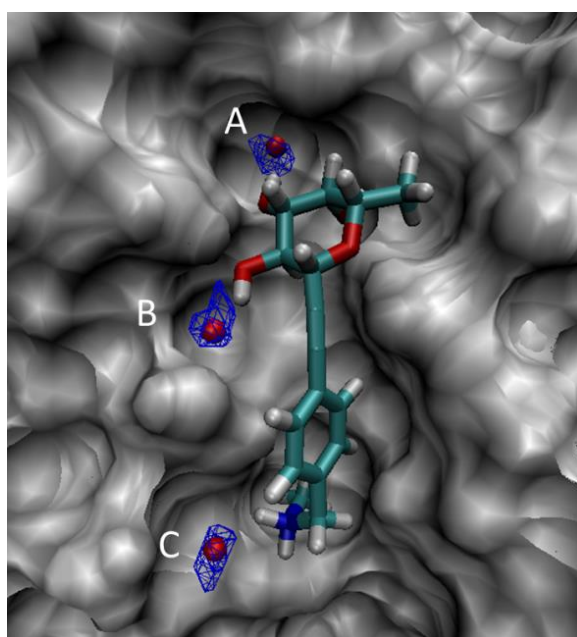


Figure 5.10 Three hydration sites A, B, C have been identified with higher solvent occupancy.

The site A was predicted with lowest free energy while the other two sites (B and C) indicated the region of slightly higher energy but both the water molecules are directly connected to the fucose ring and appears important for ligand binding in all the crystallographic complexes of BC2L-C-nt with glycomimetics and monosaccharide. Therefore, site C can be explored further for the possibility of displacement of water by suitable moiety to further improve the ligand affinity. Ligands with additional polar group such as ammonium can be designed to mimic the interactions of water molecule at site C.

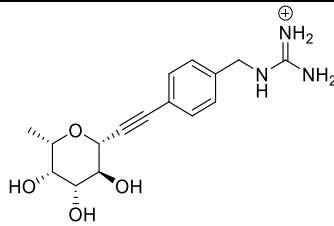
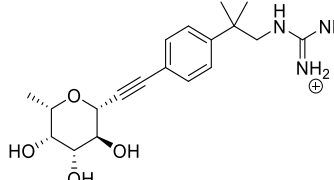
Table 5.6. Thermodynamic properties (in kcal/mol) of the three hydration sites near ligand binding site.

Site	ΔG_{solv}	ΔH_{solv}	$T\Delta S_{\text{solv}}$	Occupancy (%)
A	-4.24	-10.86	-6.62	99.22
B	-1.42	-3.07	-1.65	56.10
C	-1.41	-4.81	-3.40	87.70

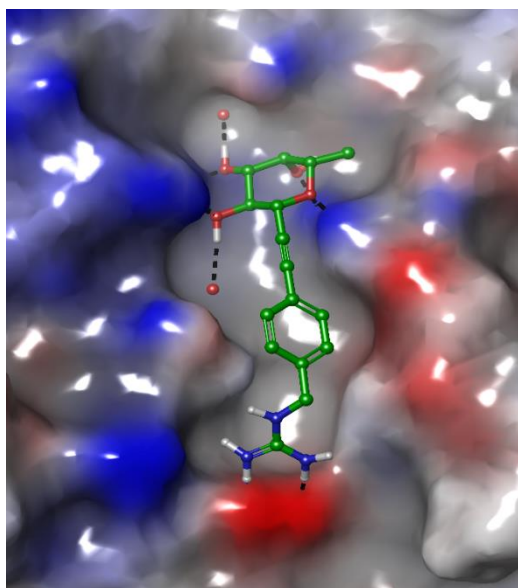
5.9 Docking studies of additional ligands with guanidine moiety

In order to improve the binding affinity of the glycomimetic ligands (Lfuc-Aky-KL07, and Lfuc-Aky-KL08), the benzylamine moiety have been further modified with a substitution of guanidine moiety (**Table 5.7**). This can probably improve ligand efficiency by establishing additional H-bonding interactions with the binding site residues (e.g. Asp70). The newly designed ligands were docked using the existing model (with two water molecules) for the docking studies of glycomimetics. The results show that the key interactions were maintained by the newly designed ligands. In addition, bidentate H-bonding interactions with two oxygen atoms of the side chain of Asp70 were established. Likewise, $-\text{CH}_2-$ group in the ligand shows hydrophobic contacts with $-\text{CH}_2-$ group in the side chain of Ser119 (**Figure 5.11 and 5.12**).

Table 5.7 Docking studies of the glycomimetic ligands with guanidine moiety.

Molecule name	Structure	GlideScore (kcal/mol)	
		XP	SP
Lfuc-Aky-KL07g		-11.1	-6.8
Lfuc-Aky-KL08g		-11.6	-6.5

A



B

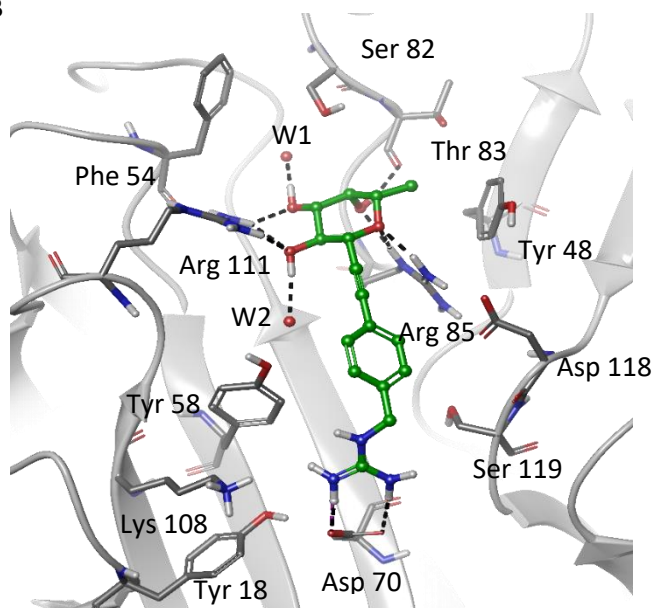


Figure 5.11 (A) Docking pose of Glycomimetic ligand (Lfuc-Aky-KL07g) with guanidine moiety. As compared to existing ligand with benzylamine moiety, this ligand shows (B) bidentate H-bonding interactions with two oxygen atoms of the side chain of Asp70.

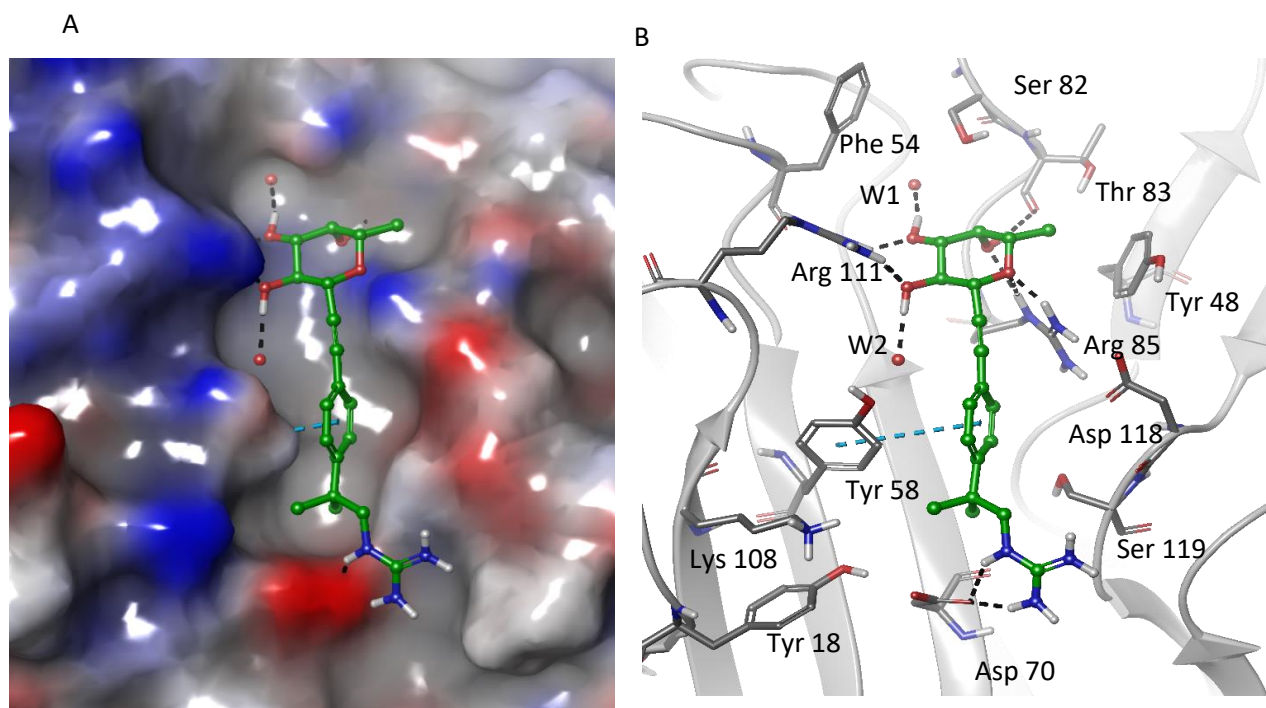


Figure 5.12 (A) Docking pose of glycomimetic ligand (Lfuc-Aky-KL08g) with guanidine moiety. (B) The ligand establishes bidentate H-bonding interactions through guanidine moiety.

These additional interactions ultimately enhanced the binding affinity (glideScore) of the newly designed ligands by 1 kcal/mol (approx.). These two ligands have been already synthesized by another collaborator at the University of Milan and further studies related to their binding to the target are still to be performed at the Université Grenoble Alpes (UGA), France.

5.10 References

- [1] Friesner, R. A. *et al.*, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, *J. Med. Chem.* 2004, *47*, 1739-1749.
- [2] Kim, S. *et al.*, PubChem in 2021: new data content and improved web interfaces, *Nucleic Acids Res.* 2020, *49*, D1388-D1395.
- [3] Schrödinger Release 2018-1: Maestro, Schrodinger., LLC, New York, NY, 2018.
- [4] Schrödinger Release 2018-1: LigPrep, Schrodinger, LLC, New York, NY, 2018.

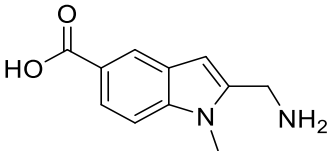
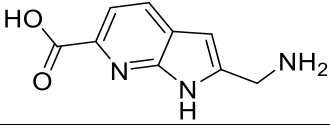
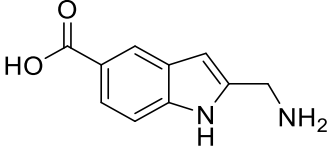
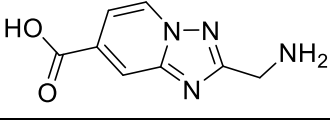
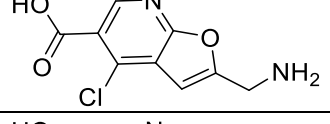
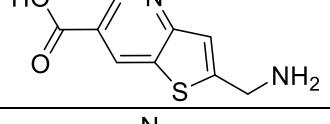
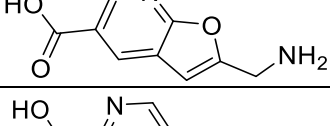
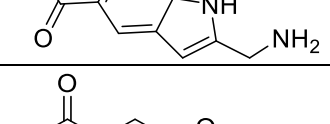
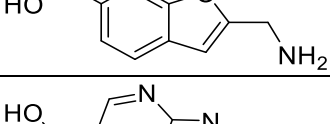
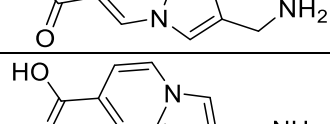
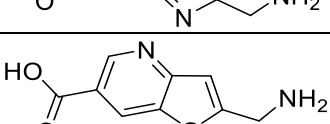
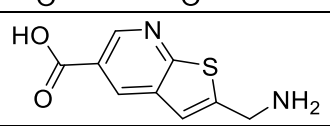
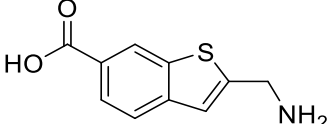

- [5] Vanquelef, E. *et al.*, R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments, *Nucleic Acids Res.* 2011, *39*, W511-517.
- [6] Frisch, M. J., *et al.* Gaussian, Inc, Wallingford CT, 2016.
- [7] Ponder, J. W. Case, D. A., Force fields for protein simulations, *Adv. Protein Chem.* 2003, *66*, 27-85.
- [8] D.A. Case, I. Y. B.-S., S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, *et al.*, AMBER 2018, *University of California, San Francisco* 2018.
- [9] Maier, J. A. *et al.*, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB, *J. Chem. Theory Comput.* 2015, *11*, 3696-3713.
- [10] Kirschner, K. N. *et al.*, GLYCAM06: a generalizable biomolecular force field. *Carbohydrates, J. Comput. Chem.* 2008, *29*, 622-655.
- [11] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys.* 1983, *79*, 926-935.
- [12] Hopkins, C. W. *et al.*, Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning, *J. Chem. Theory Comput.* 2015, *11*, 1864-1874.
- [13] Allen, M. P., Tildesley D. J., *Computer simulation of liquids, Oxford University Press, USA, 1989.*
- [14] Berendsen, H. J. C. P., J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath, *J. Chem. Phys.* 1984, *81*, 3684-3690.
- [15] Abel, R. *et al.*, A displaced-solvent functional analysis of model hydrophobic enclosures, *J. Chem. Theory Comput.* 2010, *6*, 2924-2934.
- [16] Bissantz, C. *et al.*, A Medicinal Chemist's Guide to Molecular Interactions, *J. Med. Chem.* 2010, *53*, 5061-5084.
- [17] Poornima, C. S. Dean, P. M., Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions, *J. Comput. Aided Mol. Des.* 1995, *9*, 500-512.
- [18] Riniker, S. *et al.*, Free enthalpies of replacing water molecules in protein binding pockets, *J. Comput. Aided Mol. Des.* 2012, *26*, 1293-1309.
- [19] Wong, S. E. Lightstone, F. C., Accounting for water molecules in drug design, *Expert Opin. Drug Discov.* 2011, *6*, 65-74.
- [20] Baron, R. *et al.*, Water in cavity-ligand recognition, *J. Am. Chem. Soc.* 2010, *132*, 12091-12097.
- [21] Abel, R. *et al.*, Role of the active-site solvent in the thermodynamics of factor Xa ligand binding, *J. Am. Chem. Soc.* 2008, *130*, 2817-2831.
- [22] Darby, J. F. *et al.*, Water Networks Can Determine the Affinity of Ligand Binding to Proteins, *J. Am. Chem. Soc.* 2019, *141*, 15818-15826.

- [23] Maurer, M.Oostenbrink, C., Water in protein hydration and ligand recognition, *J. Mol. Recognit.* 2019, *32*, e2810.
- [24] Zsidó, B. Z.Hetényi, C., The role of water in ligand binding, *Curr Opin Struct Biol* 2021, *67*, 1-8.
- [25] Lu, Y. *et al.*, Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes, *J. Chem. Inf. Model.* 2007, *47*, 668-675.
- [26] Lu, Y. *et al.*, Binding free energy contributions of interfacial waters in HIV-1 protease/inhibitor complexes, *J. Am. Chem. Soc.* 2006, *128*, 11830-11839.
- [27] Levinson, N. M.Boxer, S. G., A conserved water-mediated hydrogen bond network defines bosutinib's kinase selectivity, *Nat. Chem. Biol.* 2014, *10*, 127-132.
- [28] Li, Z.Lazaridis, T., Thermodynamic contributions of the ordered water molecule in HIV-1 protease, *J. Am. Chem. Soc.* 2003, *125*, 6636-6637.
- [29] de Beer, S. B. *et al.*, The role of water molecules in computational drug design, *Curr. Top. Med. Chem.* 2010, *10*, 55-66.
- [30] Spyrakis, F. *et al.*, The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery, *J. Med. Chem.* 2017, *60*, 6781-6827.
- [31] Lam, P. Y. *et al.*, Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors, *Science* 1994, *263*, 380-384.
- [32] Michel, J. *et al.*, Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization, *J. Am. Chem. Soc.* 2009, *131*, 15403-15411.
- [33] Ramsey, S. *et al.*, Solvation thermodynamic mapping of molecular surfaces in AmberTools: GIST, *J. Comput. Chem.* 2016, *37*, 2029-2037.

Annexure: Chapter 5

Table 2 Fragments resulted from similarity search in the SciFinder using a query molecule (fragment no. 15 from **Table 5.1, Chapter 5**) with carboxyl group (-COOH). Similar fragments with different heterocyclic rings and -COOH moiety were identified.

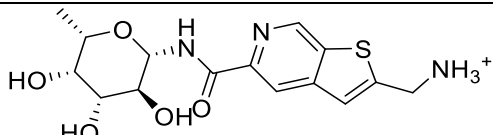
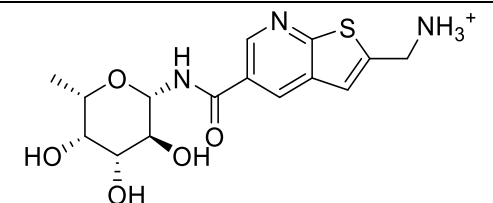
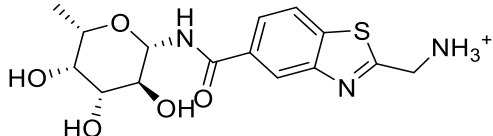
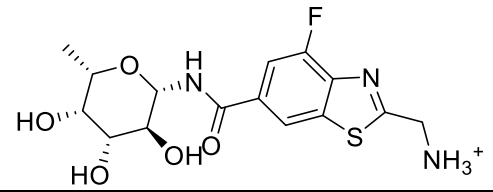
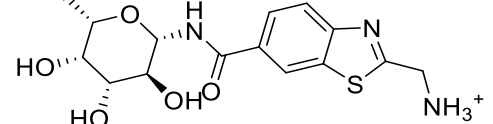
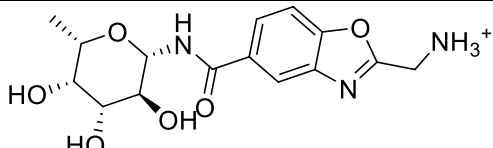
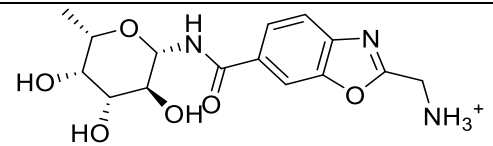
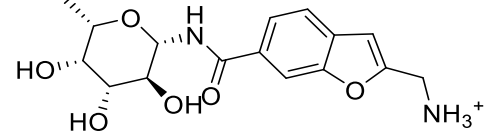
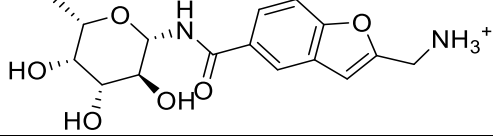
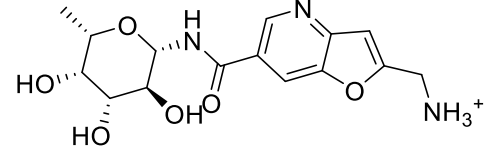
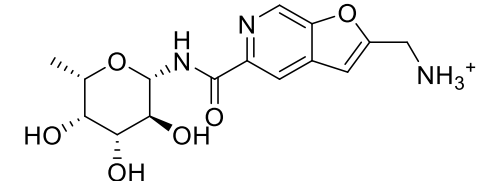
S.No	Structure
Query molecule	
1	
2	
3	
4	
5	
6	
7	
8	
9	

10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	

24	
25	
26	
27	
28	
29	
30	

Table 2 Docking studies and classification of the glycomimetic ligands designed using the fragments with heterocyclic rings (see **Table 1**) and amide linker.

Subclass	Fragment No.	Structure	GlideScore	
			XP	SP
Fragment 25 and analog (Benzothio- phene)	25 (query molecule from table 1)		-10.22	-7.20
	23		-10.44	-7.26
	15		-10.08	-7.29

Analog of fragment 25 with pyridine (thieno-pyridine)	24		-10.26	-7.23
	22		-10.59	-7.21
Benzothiazole	6		-10.08	-6.64
	2		-10.16	-7.23
	1		-10.19	-7.32
Benzoxazole	7		-10.24	-6.77
	8		-10.02	-7.11
Benzofuran and analog with pyridine	18		-9.90	-7.05
	30		-10.31	-7.30
	21		-10.05	-7.05
	26		-10.43	-7.28

	16		-10.51	-7.25
	14		-10.76	-7.40
Indole and pyrrolo-pyridine	9		-10.14	-7.14
	12		-10.58	-7.30
	10		-10.58	-7.24
	11		-9.67	-7.03
	17		-10.41	-7.34
	Heterocycle with 4N, 3N and 2N	5		-9.80
19			-10.32	-7.35

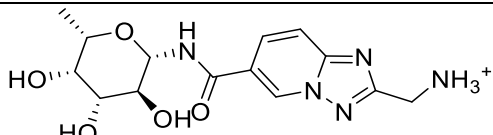
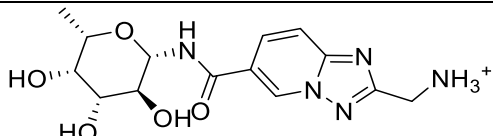
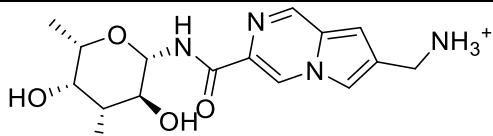
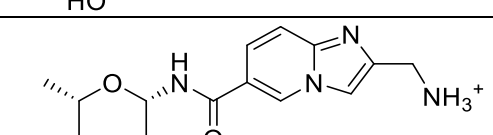
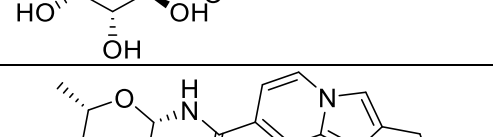
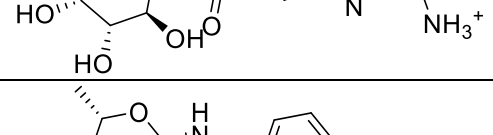
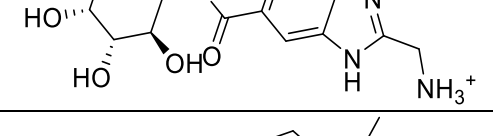
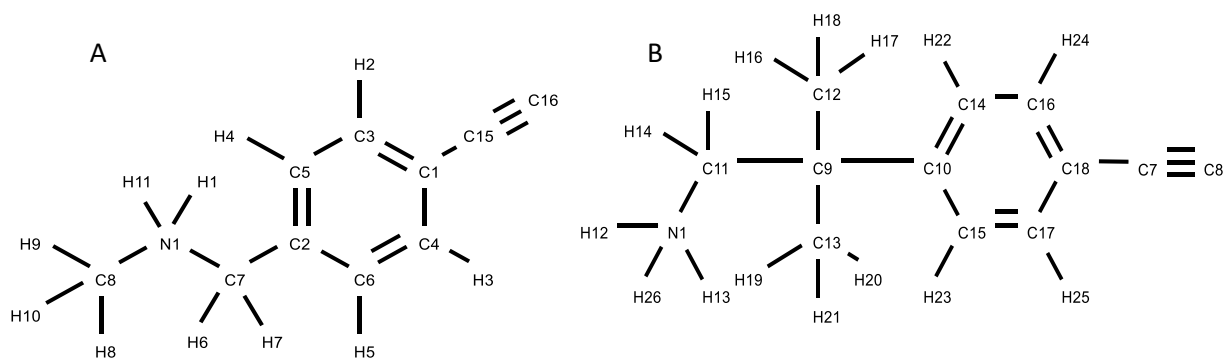
	28		-9.71	-7.35
	13		-9.71	-7.32
	3		-10.33	-7.42
	27		-10.58	-7.34
	20		-9.93	-6.89
Benzimidazole and analogs	4		-10.18	-7.20
	29		-9.96	-6.85

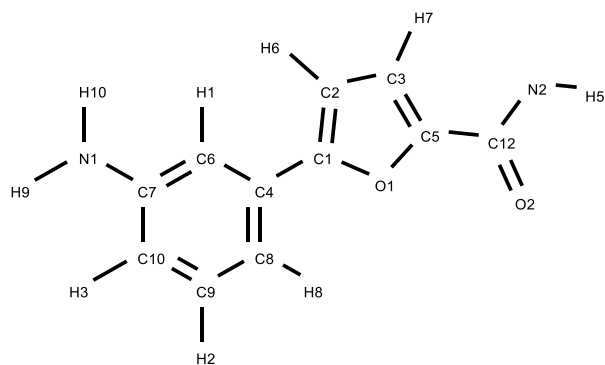
Table 3 PyRED derived charges for atoms in the non-sugar part of the glycomimetic ligands (A) Lfuc-Aky-KL07 (B) Lfuc-Aky-KL07 and (C) Lfuc-Amd-KL13.



Atom name	Atom type	Charge
C1	CA	0.2502
N1	N3	-0.1460
H1	H	0.2855
H11	H	0.2855
C8	CT	-0.0917
H8	HP	0.1140
H9	HP	0.1140
H10	HP	0.1140
C2	CA	0.0373
C3	CA	-0.1440
H2	HA	0.1710
C4	CA	-0.1440
H3	HA	0.1710
C5	CA	-0.1944
H4	HA	0.1514
C6	CA	-0.1944
H5	HA	0.1514
C7	CT	-0.0941
H6	HP	0.1357
H7	HP	0.1357
C15	CZ	-0.2735
C16	CZ	-0.0286

Atom name	Atom type	Charge
C7	CZ	-0.5193
C8	CZ	0.3019
C9	CT	0.1716
C12	CT	0.0278
H16	HC	0.0087
H17	HC	0.0087
H18	HC	0.0087
C13	CT	0.0278
H19	HC	0.0087
H20	HC	0.0087
H21	HC	0.0087
C10	CA	0.0177
C11	CT	-0.2662
H14	HP	0.1574
H15	HP	0.1574
N1	N3	-0.1660
H12	H	0.2748
H13	H	0.2748
H26	H	0.2748
C14	CA	-0.2531
H22	HA	0.1572
C15	CA	-0.2531
H23	HA	0.1572
C16	CA	-0.1452
H24	HA	0.1685
C17	CA	-0.1452
H25	HA	0.1685
C18	CA	0.3585

C



Atom name	Atom type	Charge
C1	CW	0.2140
C2	CD	-0.3058
H6	HA	0.1751
C3	CD	-0.0385
H7	HA	0.1525
C4	CA	-0.0233
O1	OS	-0.1749
C5	CW	-0.1052
C6	CA	-0.2371
H1	HA	0.1666
C7	CA	0.3150
N1	N2	-0.8711
H9	H	0.3680
H10	H	0.3680
C8	CA	-0.0859
H8	HA	0.1023
C9	CA	-0.2157
H2	HA	0.1702
C10	CA	-0.2044
H3	HA	0.1576
N2	N	-0.1570
H5	H	0.2367
C12	C	0.5436
O2	O	-0.5507

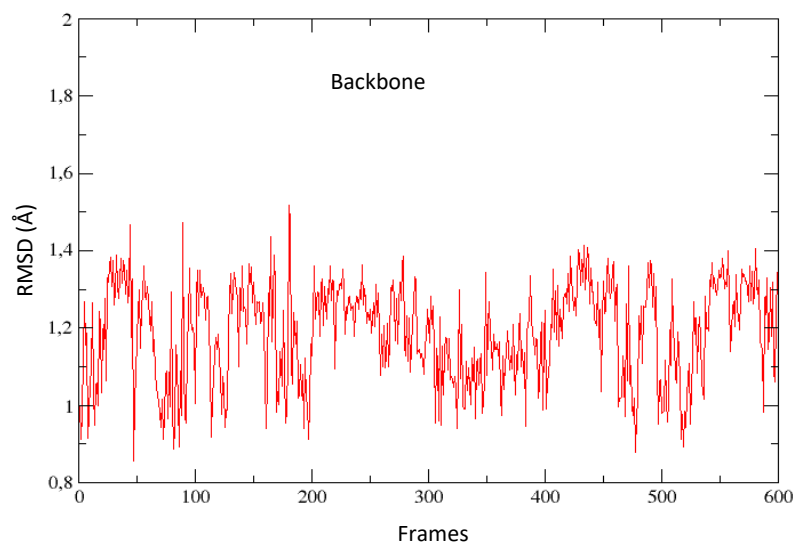


Figure 1 RMSD analysis of the backbone atoms of BC2L-C-nt complex with Me- α -L-Fuc. The results demonstrate that the complex with the trimer was stabilized.

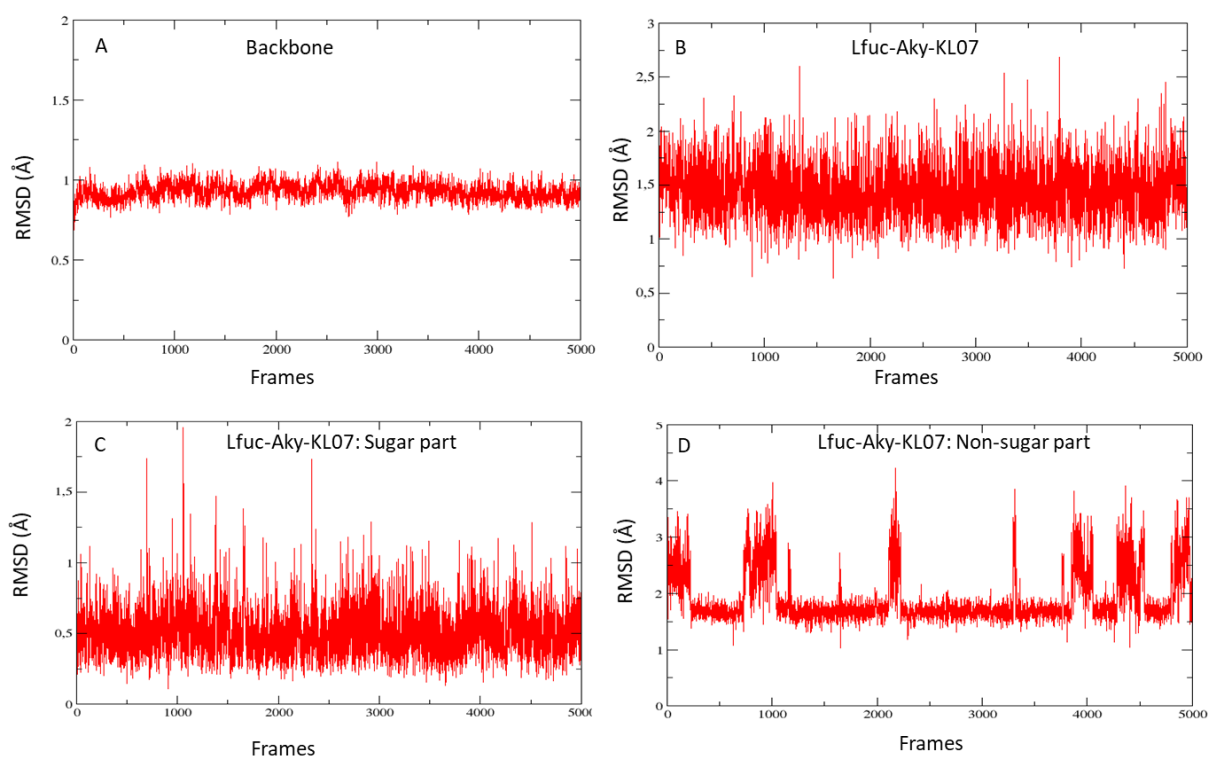


Figure 2 RMSD analysis of (A) backbone atoms of BC2L-C-nt trimer in complex with Lfuc-Aky-KL07. The analysis was also performed for (B) the glycomimetic ligand and its (C) sugar and (D) non-sugar part. The results indicate a low value of RMSD (0.5 Å) for the sugar part of the ligand while the non-sugar part displayed a higher value of RMSD due to additional H-bonding interactions with Ser119 located near the binding site.

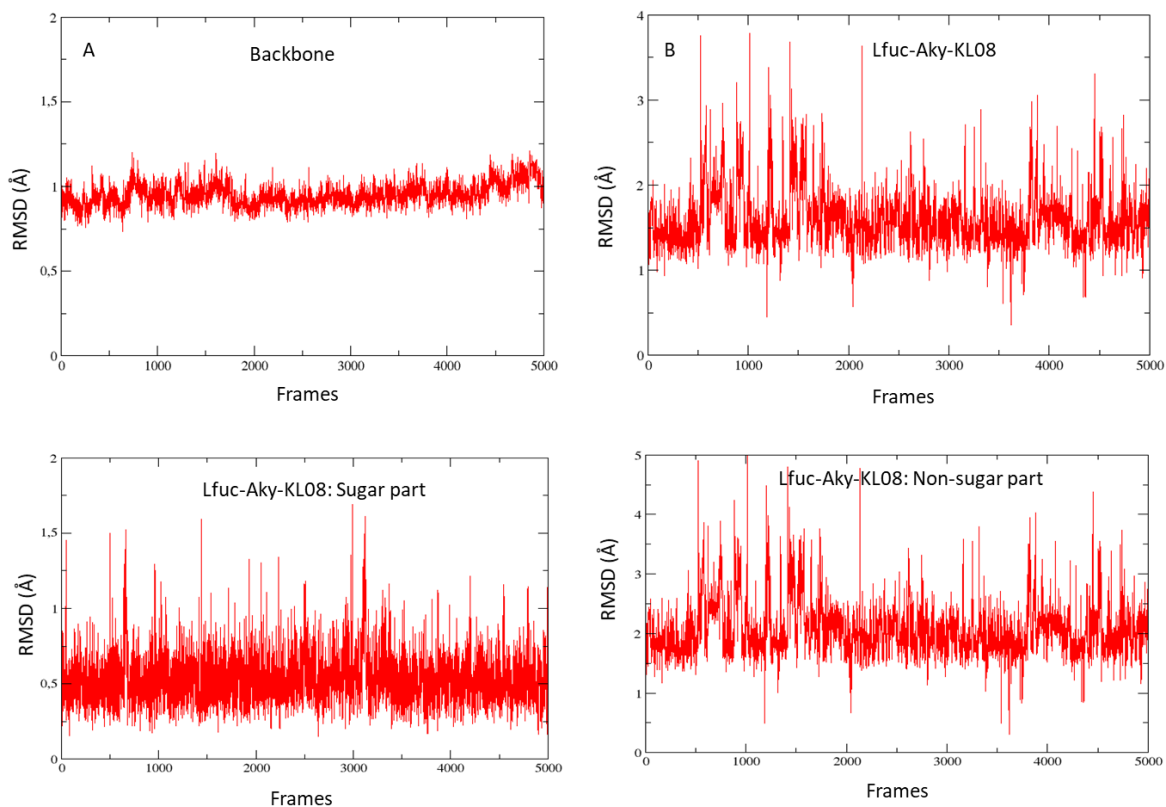


Figure 3 RMSD analysis of (A) backbone atoms of BC2L-C-nt trimer in complex with Lfuc-Aky-KL08. RMSD analysis was performed for (B) the glycomimetic ligand and its (C) sugar and (D) non-sugar part to investigate the stability of the ligand in the binding site. The results indicate a low value of RMSD (0.5 Å) for the sugar part of the ligand while the non-sugar part displayed a higher value of RMSD probably due to the presence of additional methyl groups in the ligand. The additional groups enforce the non-sugar part toward Asp118 and thus this part establishes H-bonding interactions between -NH₃⁺ of the ligand and the backbone oxygen atom in the residue (Asp118).

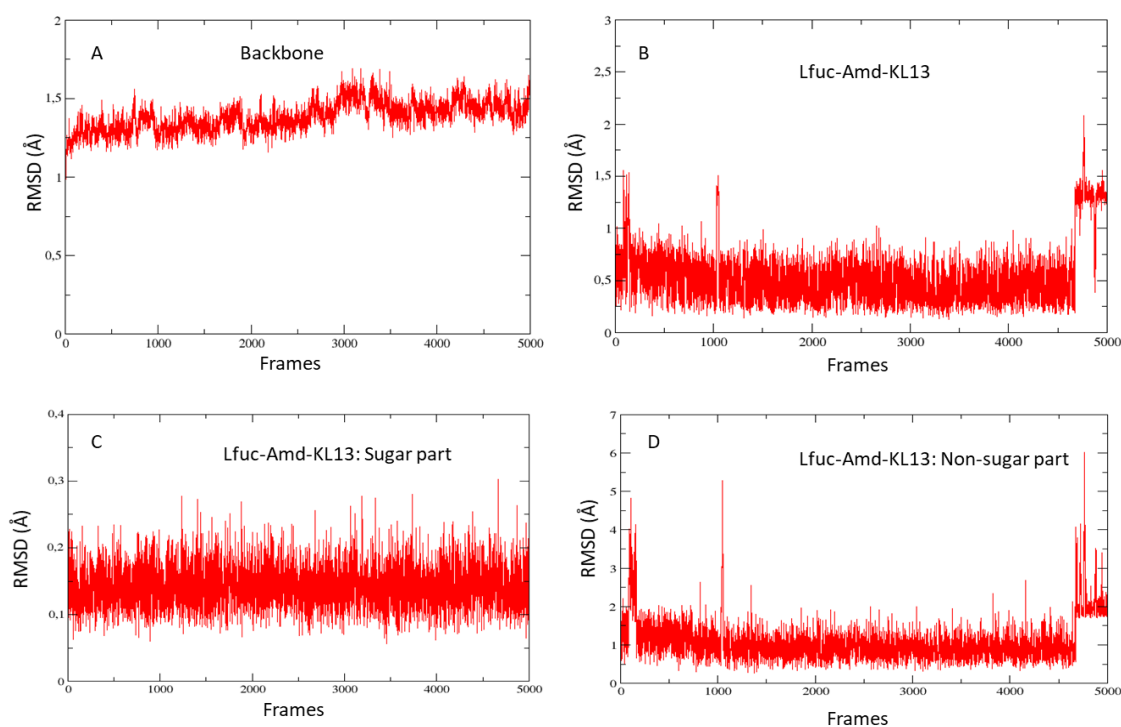


Figure 4 RMSD analysis of (A) backbone atoms of BC2L-C-nt trimer in complex with Lfuc-Amd-KL13. The analysis was also performed for (B) the glycomimetic ligand and its (C) sugar and (D) non-sugar part. As compared to the other two ligands with alkyne linker (Lfuc-Aky-KL07 and Lfuc-Aky-KL08) the results for this ligand indicate a lower value of RMSD for the sugar (0.1 Å) and non-sugar (0.5 Å) part of the ligand. In most of the sampled structures, H-bonding interactions between the ligand and Asp70 were maintained but it also displayed additional interactions with Ser119 in few (< 10%) sampled structures. The possible reason is that the non-sugar part of the ligand is shorter (lacks methyl groups or benzylamine moiety) compared to other two ligands thus accommodated well in the site X.

6. Conclusions and perspectives

The prevalence of drug resistant infections has challenged the existing treatment regimen using antibiotics. There is a need to discover and employ alternative and complementary therapies to counteract these life threatening infections. In the fast few decades, the use of anti-adhesion molecules targeting virulence factors such as lectins has been proven an attractive approach to counteract the infections by disarming the pathogens. This has been achieved by the development of glycomimetic inhibitors of various lectin targets.

Within the scope of the PhD4GlycoDrug consortium, this thesis work aimed to design glycomimetic antagonists of the N-terminal domain of the BC2L-C lectin (BC2L-C-nt) from the drug resistant pathogen known as *B. cenocepacia*. To achieve the objectives, this project employed a fragment-based approach to design glycomimetic antagonists of the target protein (BC2L-C-nt). This approach under the structure-based drug design involved the application of the computational and experimental methods to achieve the objectives of the research.

The initial studies were focused towards the binding site prediction and target evaluation by computational tools which identified additional druggable regions near the fucoside binding site in the lectin (BC2L-C-nt). These additional regions have been explored further to evaluate the druggability by employing virtual screening of a small fragment library in the vicinity of the fucose-binding site. This identified an interesting region (region 'X') that could host the drug like fragment by establishing some key interactions in the site. The interactions of the fragments with the lectin have been confirmed using a group of biophysical techniques, including X-ray crystallography. Remarkably, the binding mode of one of the

fragment (KL3) has been validated by X-ray crystallography at high resolution confirming the ability of site X to host drug-like fragments.

The computational studies also identified suitable linkers that could be used to chemically connect the fragments to the fucose core to obtain high-affinity bifunctional glycomimetic ligands. In addition, the MD simulation studies of some of the glycomimetics indicated that they establish stable interactions in the region X, thus reproducing the results in agreement with the docking studies. Further studies involving glycomimetic synthesis and evaluation of their binding to the target were performed in collaboration with another colleague (Rafael Bermeo) in the Phd4GlycoDrug network. Interestingly, the crystal complexes of BC2L-C-nt with three bifunctional glycomimetic ligands (designed connecting fragments and fucose core) again confirmed the druggability of the identified site, thus also validated the computational predictions. Hence, the first generation of glycomimetic ligands with binding affinities in micromolar range have been successfully designed.

The planned objectives of the project under the framework of the Phd4GlycoDrug-Marie Skłodowska-Curie Innovative Training Network (MSCA ITN) have successfully completed. The first generation glycomimetic ligands provide further opportunities to design high-affinity glycomimetic ligands.

Future studies to design high affinity ligands can be benefitted from the structural information obtained from the solved crystal structures of BC2L-C-nt complexes with a fragment, glycomimetic ligands and oligosaccharides. This can further help in designing new ligands with additional moieties that can mimic the interactions of the oligosaccharides. In addition, the druggable region near the fucose binding site that was identified as site Y and site Z can be targeted to further enhance the binding affinity of the ligands. The best fragments binding at these two additional regions can be prioritized and strategies to

chemically connect them to the fucose ring can be briefly explored. The ligands designed based on the targeted sites can be used to generate libraries of glycomimetics with potential as lectin antagonists. *In silico* studies on the ligands can be further extended to optimize the binding affinity using MD simulations and binding free energy calculations. Likewise, the multimeric structure of BC2L-C-nt can provide opportunity for multivalent approach to improve the binding affinity of the ligands that can antagonize the target. Hence, future studies based on structure-based approaches and robust synthetic routes to synthesize glycomimetics can lead towards the high-affinity ligands as anti-adhesive agents against *B. cenocepacia*. Further, it will be interesting to test the synthesized molecules in functional assays such as biofilm formation, cell-adhesion and hemagglutination.

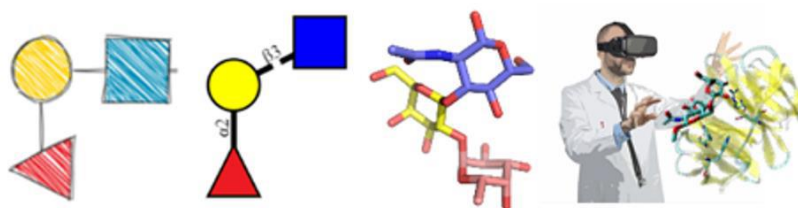
7. Scientific communication: Short secondment at Glycopedia

Under the framework of PhD4GlycoDrug, a short secondment for 15 days was organised in coordination with Glycopedia for the purpose of scientific communication. During this time, a chapter was drafted in collaboration with Rafael Bermeo and Dr. Serge Perez. The chapter will be added to the glycopedia platform although it was published as an open-access review article in the Beilstein Journal of Organic Chemistry. The review article entitled 'Computational tools for drawing, building and displaying carbohydrates: a visual guide' is attached in the next section.



Computational tools for drawing, building and displaying carbohydrates: a visual guide

Kanhaya Lal, Rafael Bermeo and Serge Perez





Computational tools for drawing, building and displaying carbohydrates: a visual guide

Kanhaya Lal^{†1,2}, Rafael Bermeo^{†1,2} and Serge Perez^{*1}

Review

Open Access

Address:

¹Univ. Grenoble Alpes, CNRS, CERMAV, 38000 Grenoble, France and ²Dipartimento di Chimica, Università Degli Studi di Milano, via Golgi 19, I-20133, Italy

Email:

Serge Perez* - spsergeperez@gmail.com

* Corresponding author ‡ Equal contributors

Keywords:

bioinformatics; carbohydrate; glycan; glycobiology; nomenclature; oligosaccharide; polysaccharide; representation; structure

Beilstein J. Org. Chem. **2020**, *16*, 2448–2468.
<https://doi.org/10.3762/bjoc.16.199>

Received: 18 June 2020

Accepted: 17 September 2020

Published: 02 October 2020

This article is part of the thematic issue "GlycoBioinformatics".

Guest Editor: F. Lisacek

© 2020 Lal et al.; licensee Beilstein-Institut.
License and terms: see end of document.

Abstract

Drawing and visualisation of molecular structures are some of the most common tasks carried out in structural glycobiology, typically using various software. In this perspective article, we outline developments in the computational tools for the sketching, visualisation and modelling of glycans. The article also provides details on the standard representation of glycans, and glycoconjugates, which helps the communication of structure details within the scientific community. We highlight the comparative analysis of the available tools which could help researchers to perform various tasks related to structure representation and model building of glycans. These tools can be useful for glycobiologists or any researcher looking for a ready to use, simple program for the sketching or building of glycans.

Introduction

Glycoscience is a rapidly surfacing and evolving scientific discipline. One of its current challenges is to keep up and adapt to the increasing levels of data available in the present scientific environment. Indeed, the rise of accessible experiment data has changed the landscape of how research is performed. The accessibility of this information, coupled with the emergence of new platforms and technologies, has benefitted glycoscience to the point of enabling the detection and high-resolution determination and representation of complex glycans [1]. Increasing

numbers of carbohydrate sequences have accumulated throughout extensive work in areas of chemical and biochemical fragmentations followed by analysis using mass spectroscopy, nuclear magnetic resonance, crystallography and computational modelling. There have been some initiatives by independent research groups worldwide, that pushed the development of visual tools to improve some aspects of glycan identification, quantification and visualisation, some of which will be further developed throughout this article.

Biological molecules express their function throughout their three-dimensional structures. For this reason, structural biology places great emphasis on the three-dimensional structure as a central element in the characterisation of biological function. An adequate understanding of biomolecular mechanisms inherently requires our ability to model and visualise them. Visualisation of molecular structures is thus one of the most common tasks performed by structural biologists. As an essential part of the research process, data visualisation allows not only to communicate experimental results but also is a crucial step in the integration of multiple data derived resources, such as thermodynamics and kinetic analysis, glycan arrays, mutagenesis, etc. Data visualisation remains a challenge in glycoscience for both the developers and the end-users even for the simple task of describing molecular structures. Progress in this area allows to translate a static visualisation of single molecules into dynamic views of complex interacting large macromolecular assemblies, which increases our understanding of biological processes.

Representing the structures of carbohydrates has historically been considered to be a complicated task. Starting from the linear form of the Fischer projection, which is certainly not a realistic representation of a carbohydrate structure, there has been a continuous development and evolution of the description of monosaccharides [2]. Glycans are puzzles to many chemists, and biologists as well as bioinformaticians. This complexity occurs at different levels (which makes it incremental). Amongst the most recognisable “sugars”, glucose is merely one of 60+ monosaccharides, all of which are, in truth, pairs of mirror-image enantiomers (L and D).

Moreover, monosaccharides occur as two forms: 5-atom ring (furanose) and 6-atom ring (pyranose). With the occurrence of a statistically rarer “open form,” we obtain at least 6 “correct” representations of glucose. And yet, monosaccharides are only the chemical units and the individual building blocks of much more complex molecules; the carbohydrates, also referred to as glycans. The glycan family can be grouped in the following categories: (i) oligosaccharides (comprising two to ten monosaccharides linked together either linearly or branched); (ii) polysaccharides (for glycan chains composed of more than ten monosaccharides); (iii) glycoconjugates (where the glycan chains are covalently linked to proteins (glycoproteins), lipids (glycolipids). The complexity of glycans is a consequence of their branched structure and the range of building blocks available. Other levels of complexity include the nature of the glycosidic linkage (anomeric configuration, position and angles), the number of repeating units (polysaccharides) as well as the substitutions of the monosaccharides. Regardless of the different nomenclatures available to describe each monosaccharide,

representing and encoding a glycan structure into a file is required for communication among scientists as well as for data processing.

As a consequence, glycobiochemists have proposed different graphical representations, with symbols or chemical structures replacing monosaccharides. The description of carbohydrate structures using standard symbolic nomenclature enables easy understanding and communication within the scientific community. Research groups working on carbohydrates have developed schematic depictions with symbols [3] and expansions with greyscale colouring as the so-called Oxford nomenclature (UOXF) [4,5], and even fully coloured schemes later on. Among these, some of the proposed representation forms have been accepted and implemented by several groups and initiatives, namely the Consortium for Functional Glycomics (CFG) [6]. Whereas the initial versions of such representation were limited to mammalian glycans, an extension of the graphical representation of glycans, called SNFG Symbol Nomenclature for Glycans (SNFG) [7,8] resulted from a joint international agreement. The newly proposed nomenclature covers 67 monosaccharides aptly represented in eleven shapes and ten colours. There is the hope that it will cope better with the rapidly growing information on the structure and functions of glycans and polysaccharides from microbes, plants and algae. The rendering of glycan drawing and symbol representations motivated the development of several computer applications using a standardised notation. The earliest glycan editors allowed manual drawing similar to ChemDraw or used input files with glycan sequence KCF (KEGG Chemical Function) [9] in text format for similarity search against other structures deposited in the databases. Later developments supported the construction and representation of glycan structures in symbolic form by computational tools like GlycanBuilder [10]. Since then, several advancements have been made to allow the user to both draw glycans manually or by importing and exporting the structure files in different text formats [11].

Along the same line, the development of various other applications allowed the users to sketch 2D-glycan structures by dragging and dropping monosaccharides to canvas to generate 3D structures for further usages. These depictions comply with protein data bank (PDB) [12] format, or in the form of images [13,14]. Besides, these tools for representing glycans in 2D and 3D shape [15] allowed the integration of glycans into protein structures or complexes. The tools developed in the last few years have automated the sketching of glycans and glycopeptides, allowing rapid display of structures using IUPAC format [16] as input. This article explores and illustrates the concepts of “sketching”, “building” and “viewing” glycans (Figure 1). It provides a descriptive analysis of the tools available for such

desired format. The tools have been attributed to categories such as “sketcher”, “builder” and “viewer”, with eventual overlaps. A brief analysis of each application ordered by category is given in the next section.

Sketching with the free hand

As a preview of the following parts of this study, we performed an initial test of the tools available for the representation of a simple disaccharide: lactose (β -D-Galp-(1 \rightarrow 4)-D-Glcp).

Figure 2 shows how different web-available platforms rendered it. On the one hand, thanks to the unified nomenclature, there is no ambiguity regarding the nature of the carbohydrate represented. On the other hand, small differences between sketches appear. Such variations will multiply with the increasing complexity of the carbohydrates. It is, therefore, essential to choose which tools to use before starting an hour-long “drawing-spreer”. The variations of the colour code used to represent the monosaccharides show striking differences across platforms even though the appropriate colours to be used are strictly defined (<https://www.ncbi.nlm.nih.gov/glycans/snfg.html#tab2>). The colour discrepancy observed here means that some of the tools do not conform to SNFG standards. For some purposes, this conformity might not be a strict necessity. Another pronounced disparity concerns the representations of the glycosidic linkage. Across sketches the length/width of the linkage varies, which will result in either compact or extended images, to be taken into account when considering the size available for the intended figures. Finally, the sketches provide further information about the linkage type: anomericity and position. These details can be either useful or superfluous depending on what is the intended use for the finished design. The main characteristic of a helpful sketching tool should be its adaptability. By allowing to modify colours, sizes, lengths/widths and turn some features on/off, a “sketcher” would allow maximum flexibility to depict carbohydrates in any desired or necessary form, size, orientation. However, this adaptability should become available without hampering the sketching effort. The perfect sketching tool would, therefore, combine flexibility and high usability.

Building with scientific accuracy

The necessity for precision is what, at some point, turns carbohydrate sketching into building. What defines this turning point (besides a certain level of accuracy) is the intended purpose for the produced figures/images. Scientific communication, comparison between similar yet different structures, or merely showcasing the complexity of carbohydrates: all three cases cannot rely on a sketching tool to convey their message. Consequently, a new set of considerations appears. The requirement for accurate depiction comes from the complexity mentioned

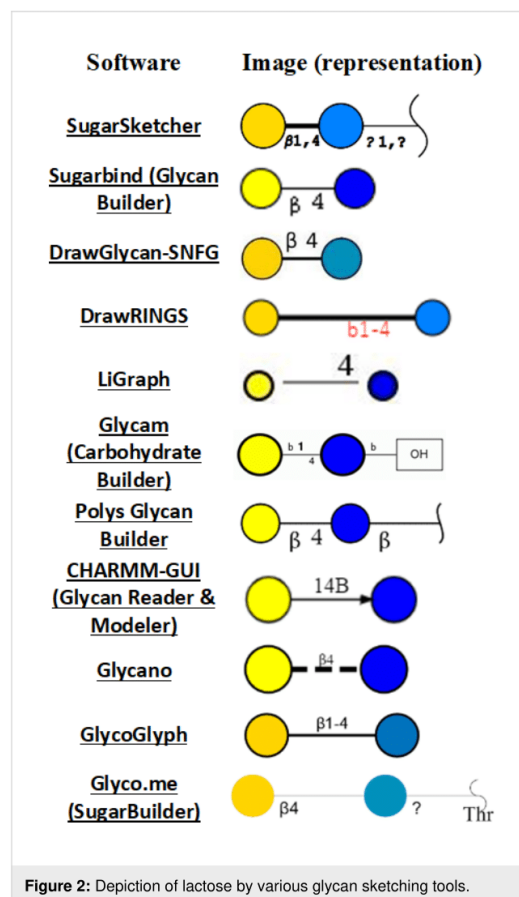


Figure 2: Depiction of lactose by various glycan sketching tools.

above of carbohydrates: anomeric configuration, substitution, glycosidic bond position, and repeating units (as well as tethering to larger macromolecules, and more). For the sake of accuracy, only the right combination of characteristics should be depicted, leaving no ambiguity: every relevant piece of data should be detailed. The glycosidic linkage is a perfect example to illustrate the necessity for accuracy in building, as opposed to sketching. While a simple line is enough to link two monosaccharides, it is necessary to define the linkage as alpha or beta (or unknown) and to state the positions of the glycosyl acceptor and even donor. Cellulose and amylose are two glucose-based polysaccharides that differ only in the nature of their glycosidic bond, and yet they have entirely different shapes and so, biological roles. For the sake of completion, the full description of a monosaccharide should obey the following rules: *<anomeric prefix><prefix for absolute configuration><the monosaccharide code><prefix for ring configuration>[<O-ester and O-ether substitutions and positions>]*. It is thus necessary to include such information when depicting carbohydrates, but

such features are simply absent in most of the existing glycan sketching tools.

Another feature that may become essential when the carbohydrate at hand is a polysaccharide is the possibility of building repeating units. Without this option, it would be simply impossible to build the required depiction. It emerges that an efficient carbohydrate builder must offer a wide array of options to characterise and personalise each monosaccharide. This would, in turn, entail a multitude of buttons, switches, etc.; which would result in a very complex interface. Consequently, unless the interface is rather straightforward and the building dynamic is well-designed, the software would be too difficult to use effectively. The ideal carbohydrate builder pick would also allow liberty for the user in terms of levels of precision since it has to fit every level of complexity above sketching. Lastly, once the building process is complete, a good builder must not only render all the provided data in the form of a precise figure but also allow the transfer of the data to other platforms (for example, by exporting the generated code).

Force fields for carbohydrates, 3D model building and beyond

Carbohydrates present various challenges to the development of force fields [23]. The tertiary structures of monosaccharides usually have a high number of chiral centres which increases the structural diversity and complexity. The structural diversity changes the electrostatic landscape of molecules; thus, it provides challenges in the development of force fields for accurate modelling of such variations in charge distributions. The monosaccharides can further form a large number of oligosaccharides which can enormously increase the conformational space, due to a high number of rotatable bonds. Nonetheless, recent developments in carbohydrate force fields enable to model and reproduce the energies associated with minute geometrical changes. The currently available force fields which are parameterised for carbohydrates are also capable of carrying out simulations of the oligosaccharides containing additional groups like sulfates, phosphates etc. [24] Generally used force fields for the Molecular Mechanics (MD) simulation of carbohydrates are CHARMM [25], GLYCAM [26], and GROMOS [27]. The structural complexity increases the computational cost, which makes simulations of large systems more challenging. Therefore, coarse-grained models [28] for carbohydrates are generally used for molecular modelling of large systems.

In terms of 3D model building, the complex topologies of glycans require dedicated molecular building procedures to convert sequence information into reliable 3D models. These tools generally use 3D molecular templates of monosaccharides

to reconstruct a 3D model. Energy minimisation methods can further refine the models. These models are essential for structure-based studies and complex calculations like Molecular Dynamics simulations. Therefore, the accurate model building requires the use of reliable databases to generate atomic coordinates and topology to provide an acceptable model. Some of the computational tools usually contain atom coordinates of generally used monosaccharides (as templates) and also use libraries of bond and angle parameters from various force fields dedicated for carbohydrates. The accurately predicted oligosaccharide conformations are good starting points for further investigations. Of particular interest are the evaluations of the dynamics of glycans and their interactions with proteins which is a most significant concern in glycoscience. The joint need to better perceive and manipulate the three-dimensional objects that make up molecular structures is leading to a rapid appropriation of techniques of Virtual Reality (VR) by the molecular biology community. Generic definitions describe VR as being immersion in an interactive virtual reactive world. The computer-generated graphics provide a realistic rendering of an immersive and dynamic environment that responds to the user's requests. One finds in these definitions the three pillars that define VR: Immersion, Interaction, Information. Although it is difficult to extract a single, simple definition of VR, the main idea is to put the user at the centre of a dynamic and reactive VR environment, artificially created and which will supplant the real world for the time of the experiment.

Input and output for sketching, building and displaying applications

The variety and complexity of carbohydrate structures hamper the use of a unique nomenclature. The choice of notation depends on whether the study is focused on chemistry or has a more biological approach. The IUPAC-IUBMB (International Union for Pure and Applied Chemistry and International Union for Biochemistry and Molecular Biology) terminologies, in their extended and condensed forms [16], govern the naming of the primary structure or sequence.

Further down the line, the complexity of the existing nomenclatures for carbohydrate-containing molecules remains a significant hurdle to their practical use and exchanges within and outside the glycoscience cenacle. The linearisation of the description of the structure is a way to cope with the description of the structural complexity. The proposed formats provide rules to extract the structure of the branches and create a unique sequence for the carbohydrate. The most commonly used formats are IUPAC [16], GlycoCT [29], KCF [9], and WURCS [30].

The sketching of carbohydrates using computational tools generally requires the textual input and output in at least one of


Input Output formats																									
IUPAC condensed	Man(a1-3)[Man(a1-6)]Man(b1-4)GlcNAc(b1-4)b-GlcNAc																								
LINUXS	[][b-D-GlcpNAc]{((4+1))}[b-D-GlcpNAc]{((4+1))}[b-D-Manp]{((3+1))}[a-D-Manp]{((6+1))}[a-D-Manp]{((3+1))}																								
GlycoCT	<table border="1"> <thead> <tr> <th>RES</th> <th>LIN</th> </tr> </thead> <tbody> <tr> <td>1b:b-dglc-HEX-1:5</td> <td>1:1d(2+1)2n</td> </tr> <tr> <td>2s:n-acetyl</td> <td>2:1o(4+1)3d</td> </tr> <tr> <td>3b:b-dglc-HEX-1:5</td> <td>3:3d(2+1)4n</td> </tr> <tr> <td>4s:n-acetyl</td> <td>4:3o(4+1)5d</td> </tr> <tr> <td>5b:b-dman-HEX-1:5</td> <td>5:5o(3+1)6d</td> </tr> <tr> <td>6b:a-dman-HEX-1:5</td> <td>6:5o(6+1)7d</td> </tr> <tr> <td>7b:a-dman-HEX-1:5</td> <td></td> </tr> </tbody> </table>	RES	LIN	1b:b-dglc-HEX-1:5	1:1d(2+1)2n	2s:n-acetyl	2:1o(4+1)3d	3b:b-dglc-HEX-1:5	3:3d(2+1)4n	4s:n-acetyl	4:3o(4+1)5d	5b:b-dman-HEX-1:5	5:5o(3+1)6d	6b:a-dman-HEX-1:5	6:5o(6+1)7d	7b:a-dman-HEX-1:5									
RES	LIN																								
1b:b-dglc-HEX-1:5	1:1d(2+1)2n																								
2s:n-acetyl	2:1o(4+1)3d																								
3b:b-dglc-HEX-1:5	3:3d(2+1)4n																								
4s:n-acetyl	4:3o(4+1)5d																								
5b:b-dman-HEX-1:5	5:5o(3+1)6d																								
6b:a-dman-HEX-1:5	6:5o(6+1)7d																								
7b:a-dman-HEX-1:5																									
KCF	<table border="1"> <thead> <tr> <th>ENTRY</th> <th>G12157</th> <th>Glycan</th> </tr> </thead> <tbody> <tr> <td>NODE</td> <td>6</td> <td>EDGE 5</td> </tr> <tr> <td></td> <td>1 Asn 18 0</td> <td>1 2:b1 1:4</td> </tr> <tr> <td></td> <td>2 GlcNAc 9 0</td> <td>2 3:b1 2:4</td> </tr> <tr> <td></td> <td>3 GlcNAc -1 0</td> <td>3 4:b1 3:4</td> </tr> <tr> <td></td> <td>4 Man -11 0</td> <td>4 5:a1 4:6</td> </tr> <tr> <td></td> <td>5 Man -19 3</td> <td>5 6:a1 4:3</td> </tr> <tr> <td></td> <td>6 Man -19 -3</td> <td>///</td> </tr> </tbody> </table>	ENTRY	G12157	Glycan	NODE	6	EDGE 5		1 Asn 18 0	1 2:b1 1:4		2 GlcNAc 9 0	2 3:b1 2:4		3 GlcNAc -1 0	3 4:b1 3:4		4 Man -11 0	4 5:a1 4:6		5 Man -19 3	5 6:a1 4:3		6 Man -19 -3	///
ENTRY	G12157	Glycan																							
NODE	6	EDGE 5																							
	1 Asn 18 0	1 2:b1 1:4																							
	2 GlcNAc 9 0	2 3:b1 2:4																							
	3 GlcNAc -1 0	3 4:b1 3:4																							
	4 Man -11 0	4 5:a1 4:6																							
	5 Man -19 3	5 6:a1 4:3																							
	6 Man -19 -3	///																							
WURCS	WURCS=2.0/3,5,4/[a2122h-1b_1-5_2*NCC/3=O][a1122h-1b_1-5][a1122h-1a_1-5]/1-1-2-3-3/a4-b1_b4-c1_c3-d1_c6-e1																								

Figure 3: Examples of different glycan structure text formats for the same glycan. Data in these formats are generally used as input/output in glycan drawing and 3D structure building tools.

these formats (Figure 3). An alternate input method involves manual sketching of 2D glycan structures by dragging and dropping monosaccharide symbols on canvas (with or without grids) to connect them further. This method makes the sketching tools more friendly and interactive as it does not require large text code as input. Both input methods are compliant to the Symbol Nomenclature for Glycans (SNFG). Another symbolic representation that could clearly distinguish monosaccharides in monochrome colours is the Oxford notation [5]. In this method, dashed and solid lines represent the alpha and beta glycosidic linkages, respectively. There are few tools which have implemented this method while other tools use text to highlight this information in the structures. In addition to sketching tools, some applications, specific to the field of carbohydrates, provide the possibility to visualise and display 3D structures. These visualisation tools accept strings or files in text formats (GlycoCT, IUPAC-condensed, KCF) to display the structure via a graphical user interface. For instance, the DrawGlycan-SNFG [31] tool uses IUPAC-condensed nomenclature for input string and converts it into a 2D image represented in SNFG symbols. At the same time, the 3D-SNFG [15] can generate glycan structures by incorporating SNFG symbols

in 3D space for further visualisation using the computational tools like visual molecular dynamics (VMD) [21] LiteMol [22] and Sweet Unity Mol [32].

Glycan sketchers

SugarSketcher. SugarSketcher [14] is a JavaScript interface module currently included in the tool collection of Glycomics@ExpASY (available at <https://glycoproteome.expasy.org/sugarsketcher/>) for online drawing of glycan structures. The interactive graphical interface (Figure 4, top) allows glycan drawing by glycobioinformaticists and non-expert users. In particular, a “Quick Mode” helps users with limited knowledge of glycans to build up a structure quickly as compared to the normal mode, which offers options related to the structural features of complex carbohydrates (for example additional monosaccharides, isomers, ring types, etc.). The building of glycan structures uses mouse and proceeds via a selection of monosaccharides, substituents and linkages from the list of symbols. However, some wrong combinations of choices can block the interface, resulting in the need to re-start the process (SugarSketcher is on version beta 1.3). Alternatively, SugarSketcher also uses GlycoCT or a native template library as an input. A list of pre-built core N- and O-linked carbohydrate moieties, which are usually present in glycoproteins structures, can be used as a template for further modification. A shortlist of glycan epitopes is also included providing templates for drawing more complex molecules. The software uses the Symbol Nomenclature for Glycans (SNFG) notation for structure representation and exports the obtained sketch to text format (GlycoCT) or image (.svg) files. The software SugarSketcher is featured in the web portal GlyCosmos (<https://glycosmos.org/glytoucans/graphic>) [33]. Under the name “SugarDrawer”, it provides an interface for generating carbohydrate structures to query the database included in GlyCosmos: GlyTouCan [34].

GlyCosmos is a web portal that integrates resources linking glycosciences with life sciences. Besides elements such as “SugarDrawer” and GlyTouCan (carbohydrate database), the platform GlyCosmos assembles data resources ranging from glycoscience standard ontologies to pathologies associated with glycans. GlyCosmos is recognized as the official portal of the Japanese Society for Carbohydrate Research and provides information about genes, proteins, lipids, pathways and diseases.

GlyTouCan (Figure 5) is a repository for glycans which is freely available for the registry of glycan structures. The repository can register structures ranging from monosaccharide compositions to fully defined structures of glycans. It assigns a unique accession number to any glycan to identify its structure and even allows to know its ID number in other databases. Al-

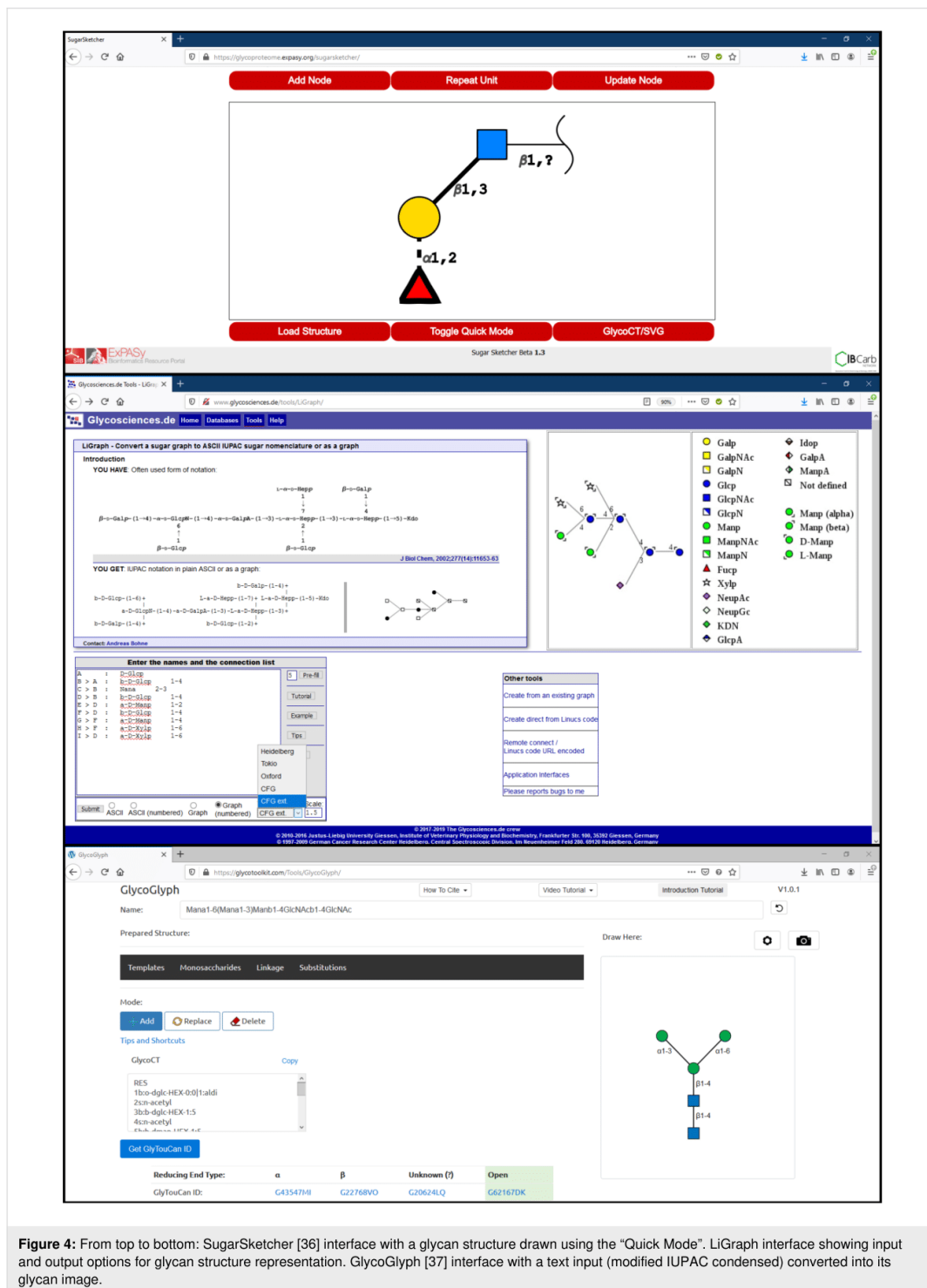
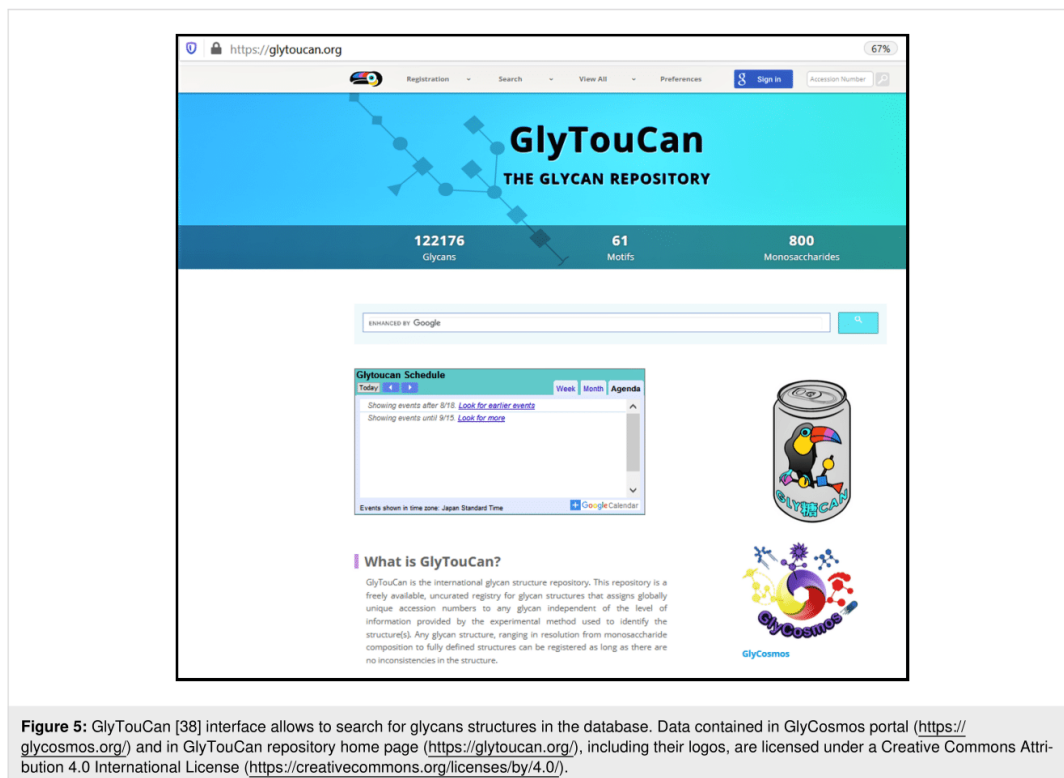


Figure 4: From top to bottom: SugarSketcher [36] interface with a glycan structure drawn using the "Quick Mode". LiGraph interface showing input and output options for glycan structure representation. GlycoGlyph [37] interface with a text input (modified IUPAC condensed) converted into its glycan image.



ternatively, users can search and retrieve information about the glycan structures and motifs that have been already registered into the repository. The structures can be searched simply by browsing through the list of already registered glycans or by specifying a particular sub-structure to retrieve structurally similar glycans (<https://glytoucan.org/Structures/graphical>). The software tool featured in the GlyTouCan website is called GlycanBuilder and is presented in a later section of our analysis.

Recapitulating, SugarSketcher can be an efficient tool for non-glycobiologists or glycobiologists to sketch glycans. However, it does not accept different input or output formats like IUPAC, WURCS (Web3 Unique Representation of Carbohydrate Structures), which would make the tool more versatile.

LiGraph. LiGraph [35] (<http://www.glycosciences.de/tools/LiGraph/>) is an online tool based on the concept of schematic drawings of oligosaccharides to display glycan structures. This tool also renders images of glycans in different notation using a text input. The input for the carbohydrate structure consists of a list of names and connections. The glycan structure is output in the specified notation: either ASCII IUPAC sugar nomencla-

ture or a graph which can be rendered in different themes which include Heidelberg, Oxford, Tokyo, CFG and extended CFG (Figure 4, middle). The output images for the glycan structure and the legends can be saved and downloaded in .svg format. This tool is useful for glycan sketching using text templates, but its shortcomings include a limited number of monosaccharide symbols and restricted compatibility with other input file formats.

GlycoGlyph. GlycoGlyph [39] is a web-based application (available at <https://glycotoolkit.com/Tools/GlycoGlyph/>) built using JavaScript which allows users to draw structures using a graphical user interface or via text string in the CFG linear (also known as modified IUPAC condensed) nomenclature dynamically. The interface (Figure 4, bottom) is equipped with templates for N- and O-linked glycans and terminals. Also, it provides 80+ monosaccharide (SNFG) symbols and a selection for substituents. The selected template or text string (in CFG linear nomenclature) input directly gets converted into an image in canvas and also appears as text in GlycoCT format. The output can be saved as a .svg file or as GlycoCT text. The interface also provides additional options to add, replace or delete each monosaccharide, modify the sizes of symbols and text

fonts, and turn off the linkage annotations or change their orientation; all of which increases the usability of the software. The input structure can be further used to search the GlyTouCan [34] database to explore the literature details related to the input structure.

GlycoGlyph is an efficient tool for sketching or building glycans with a highly usable interface that can significantly help researchers to improve the uniformity in glycan formats in literature/manuscripts. It can also be a tool of choice for text mining for the query structure.

GlycanBuilder2. GlycanBuilder2 [40] is a Java-based glycan drawing tool which runs locally as an application on different platforms including Windows, macOS and Linux. It is freely available for downloading at <http://www.rings.t.soka.ac.jp/downloads.html>. GlycanBuilder2 is a newer version of GlycanBuilder [20] with additional features. This version is capable of supporting various ambiguous glycans consisting of monosaccharides from plants and bacteria. The tool uses the SNFG notation to display glycan structures. Moreover, this updated version can convert a drawn structure into WURCS sequences for further use as a query for glycan search or registration in databases like GlyTouCan. GlycanBuilder2 provides an excellent interface (Figure 6, top) for glycan drawing. Glycan structures can be drawn manually using the mouse or by importing text input files. The interface provides a list of templates: N- O-glycans, glycosphingolipids, glycosaminoglycans(GAGs). Rows of CFG notations for monosaccharides assist with glycan structure drawing on canvas. The application also supports the glycan symbol notations for the University of Oxford (UOXF) format. The input complies with various linear sequence and text formats. They include GlycoCT, GLYcan structural Data Exchange using Connection Tables (GLYDE-II), Bacterial Carbohydrate Structures DataBase (BCSDB) [41], carbohydrate sequence markup language (CabosML) [42], CarbBank [43], LinearCode [44], LINUCS, IUPAC-condensed and GlycoSuiteDB [45]. The output yields structures in the following formats: GlycoCT, LinearCode, GLYDE-II and LINUCS. Thus, GlycanBuilder2 is a versatile tool which can be used for glycan sketching or building and also as a glycan sequence converter from one format to another.

Original GlycanBuilder. GlycanBuilder [10,20] was originally part of the GlycoWorkbench platform [49]. This interface is integrated in most tools of the Glycomics@ExPASy collection that require a drawing interface to query data. GlycanBuilder is written in Java Programming language and can be used as standalone or as an applet for embedding in web pages for glycan search. For example, GlycanBuilder is integrated in SugarbindDB [50] to draw glycan structures and search the

database (<https://sugarbind.expasy.org/builder>), and in GlycoDigest or GlyS3 [10,20] to define the input of these tools.

Technically, the tool provides an interactive interface which allows an automated glycan rendering using a library of individual monosaccharides or pre-built template structures (Figure 6, middle). GlycanBuilder provides access to 41 templates. They include N- and O-linked glycans, GAGs (glycosaminoglycans), glycosphingolipids and milk oligosaccharides. It also contains 68 entries from MonosaccharideDB (<http://www.monosaccharidedb.org/>) including monosaccharides, modifications (e.g. deoxy) and substituents. The tool provides options to modify a monosaccharide by adding substituents and alterations. Free movement of the monosaccharides is allowed through movement and orientation buttons. GlycanBuilder offers multiple options for glycan notation which include CFG, CFG colour, UOXF, UOXF colour and text only. GlycanBuilder can also calculate the masses of glycan structures according to the options selected by the user. GlycanBuilder is a versatile tool for building carbohydrates, with multiple options for exporting the generated structures in the form of text format (GlycoCT, LINUCS, Glycominds, Glyde II) or image (.svg, .png, .jpg, .bmp, .pdf, etc.) files.

DrawRINGS. DrawRINGS [17] is a Java-based applet for rendering glycan structures on canvas (<http://www.rings.t.soka.ac.jp/drawRINGS-js/>). The different drawing features in an interactive interface (Figure 6, bottom) can be selected with the mouse by surfing the buttons and scroll-down menus. Alternatively, KCF files or KCF text format can be used as input. The free movement of the monosaccharides allows drawings with flexible geometry, for example, for schematic studies of carbohydrates. The drawn glycan structure can be exported in the KCF or IUPAC text format or saved in .png format. The drawn structure can further be used as a query for the search in glycan databases; using match percentage (Similarity) or by the number of components matched (Matched) criteria. Four predefined score matrices are available, named: N-glycans, O-glycans, Sphingolipids and Link_similarity. The “Link_similarity” matrix is based on glycosidic linkages and monosaccharides that may be more highly substituted with other glycosidic linkages and monosaccharides, respectively. There is a query to search the generated structure in the RINGS or GlycomeDB databases (or both). The former compiles data from the KEGG GLYCAN and GLYCOSCIENCES.de databases. DrawRINGS is an efficient tool for sketching glycan figures as well as translating to (and from) the KCF and IUPAC text formats.

DrawGlycan-SNFG. DrawGlycan-SNFG [31] is an open-source program available with a web interface (Figure 7, top) at

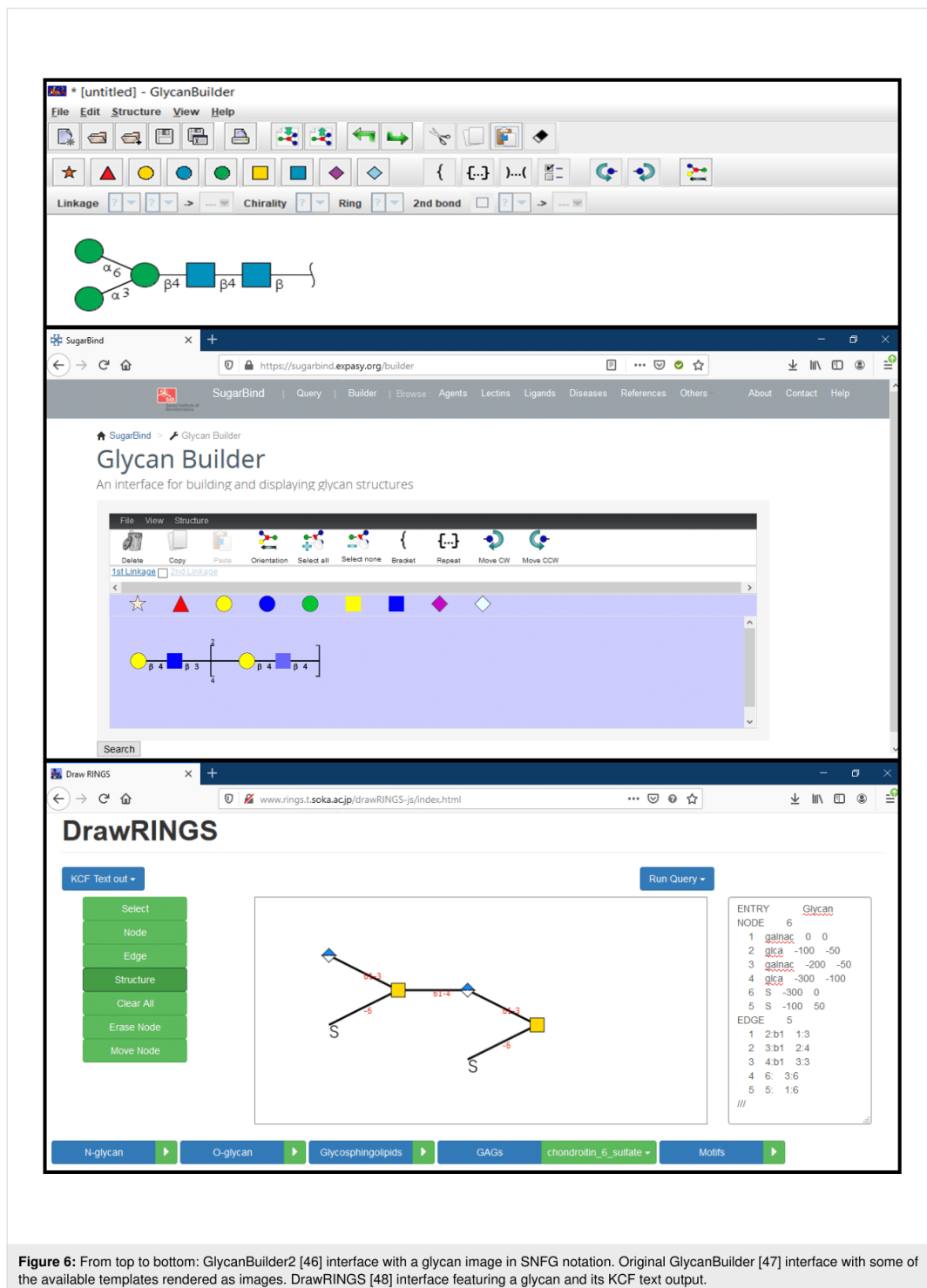


Figure 6: From top to bottom: GlycanBuilder2 [46] interface with a glycan image in SNFG notation. Original GlycanBuilder [47] interface with some of the available templates rendered as images. DrawRINGS [48] interface featuring a glycan and its KCF text output.

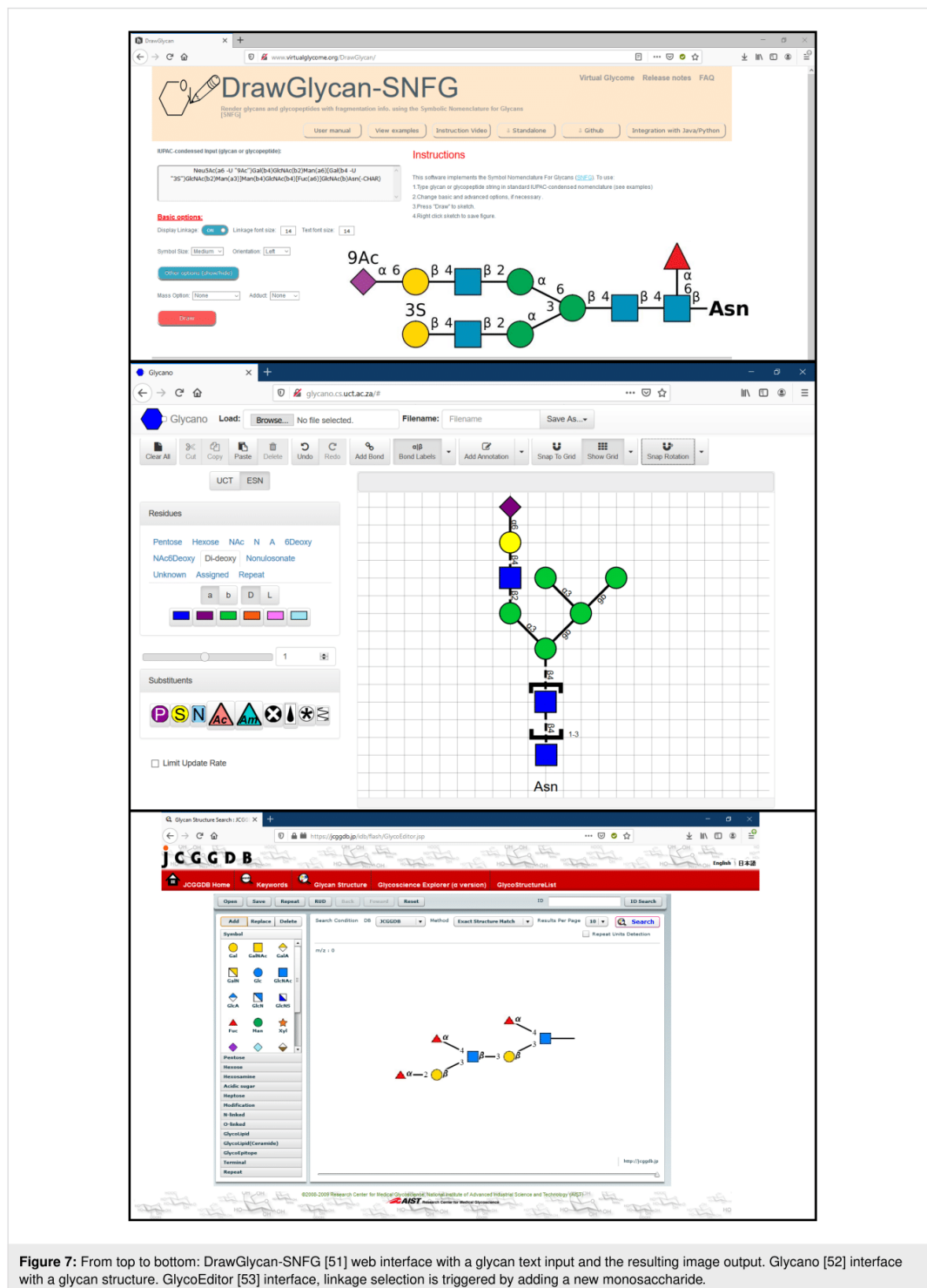


Figure 7: From top to bottom: DrawGlycan-SNFG [51] web interface with a glycan text input and the resulting image output. Glycano [52] interface with a glycan structure. GlycoEditor [53] interface, linkage selection is triggered by adding a new monosaccharide.

<http://www.virtualglycome.org/DrawGlycan>. The same web page gives access to a downloadable, standalone Graphical User Interface (GUI) version of this tool with additional functionality. It can be launched from different platforms including Windows, Mac or Linux. The program can be used to render glycans and glycopeptides using SNFG and uses IUPAC-condensed text inputs. The DrawGlycan-SNFG version with command-line operations makes it more versatile as it allows integration of multiple features of the program using custom scripts. The tool uses automatic operations for the majority of the drawing, which could meet the needs of researchers, but additional intervention may sometimes be required to get the desired output. For example, manual input in IUPAC-condensed language allows to generate, among others: repeating units, adducts, tethering to other structures (represented by text), and complex branching (the examples section showcases these options). The drawn glycan structure can be saved as .jpg image and modified through parameters such as symbol and text size, the thickness of lines, orientation of drawing and spacing. This software provides all the guidance and tools needed to generate high-quality pictures. DrawGlycan-SNFG is a reliable choice for building glycans.

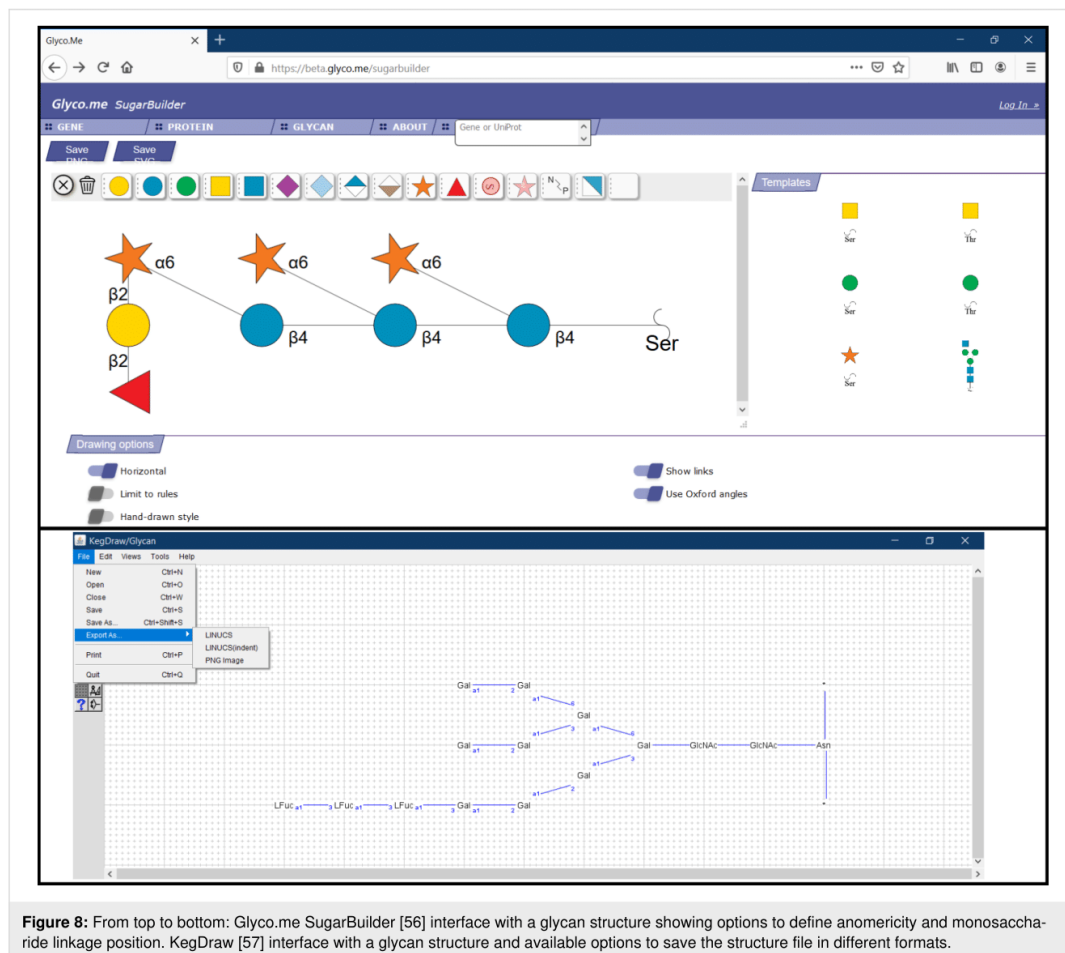
In addition to glycan structure drawing, DrawGlycan-SNFG (version 2) [54] is equipped with a wide range of options to enhance the usability of the original code [32]. The new version is capable to accommodate the latest updates to the SNFG [7]. This tool has been particularly upgraded for MS spectrum annotation by adding an intuitive interface with additional features. The upgraded version can depict bond fragmentation, repeating structural unit anomeric groups, adduct ions, different types of glycosidic linkages etc. These advanced features make this tool ideal for integrated use with various glycoinformatics software and also for applications in glycoproteomics, glycomics and mass spectrometry (MS). One of the illustrations involves combined use with the gpAnnotate application, dedicated to score and annotate MS/MS glycopeptide spectrums in different fragmentation modes [54].

Glycano. Glycano (available at <http://glycano.cs.uct.ac.za>) is a software tool for drawing glycans. This tool is based on JavaScript, which can be used without the requirement of any server or browser dependency. The interactive interface allows sketching via the drag-and-drop method on canvas (with or without grid). The software is provided with “UCT” and “ESN”, interchangeable interfaces (Figure 7, middle) with different symbols for monosaccharides. These names (UCT and ESN) correspond to the University of Cape Town, South Africa, where Glycano was developed, and to the “Essentials of Glycobiology Symbol Nomenclature”, precursor of the SNFG symbol set [55]. The interface provides a wide choice of monosaccha-

rides and substituents represented in SNFG symbols but lacks the standard colour scheme. The user can easily modify the structure with by click and drag, which allows to either cut/copy, delete or move a portion of the structure. The drawn structure can be saved in text format, in .gly format or as an image (PNG and SVG formats). A drawback to note is that linking the monosaccharides at specific positions is only possible in the UCT mode, which means that back-and-forth between the two symbol systems is necessary to define the linkages correctly. Despite some drawbacks, this is an excellent tool due to its ease-of-use, tenable degree of freedom, and functionalities/options for sketching and building glycan structures.

GlycoEditor. GlycoEditor [19] (available at <https://jcgdb.jp/idb/flash/GlycoEditor.jsp>) is an online software for drawing glycans. Through a straightforward interface, three ways of input are possible: by JCGGDB ID, through a library of common oligosaccharides and by direct input. A list of most common monosaccharides is presented, and the rest can be found categorised by family. The click and drag addition of new monosaccharides trigger the selection of linkage-type and configuration (Figure 7, bottom). The tool provides an option to create repeating units. Additionally, several functionalisation options are also available. Once the structure is ready, the user can save it as an .xml file. GlycoEditor allows searching a given structure across many databases in four ways: exact structure match (with or without anomer and linkage specifics) and the same for substructure match. The central database featured is the JCGGDB, to which can be added, among others: Glaxy, GlycomeDB, GlycoEpitope, GMDB, KEGG, etc. Searching by ID is also possible. GlycoEditor is a now dated tool that allows efficiently building glycans and performing databases searches.

GLYCO.ME (SugarBuilder). Glyco.me-SugarBuilder (available at <https://beta.glyco.me/sugarbuilder>) is online software for drawing glycans. The interface leads to rapid carbohydrate construction. A panel of monosaccharide templates complements the drawing interface (one pre-built oligosaccharide is available (Figure 8, top). The user can start a chain from amino acid residues: Asn, Ser or Thr, then structure building is limited by to a set of “rules” (limiting building options to known carbohydrates). These rules may be deactivated with a switch button to draw freely. A list of 13 monosaccharides is deployed, and sequential clicking allows their addition to the existing structure and definition of the associated glycosidic bond (the relative sizes of the options available related to their real statistical value for that particular linkage). Upon building some specific motifs, if they are recognised, an option for repeating units appears. Other switch buttons allow the user to change the orientation of the drawing, show/hide linkage information etc. The Oxford notation can be enabled for glycosidic bonds only. The



structure obtained can be rendered as .png or .svg images. Glyco.me-SugarBuilder is still under development: more monosaccharides/substitutions/templates will complete an already very functional platform. The quick and easy options put forward offer natural building and liberty for tailoring the rendered image.

KegDraw. KegDraw (<https://www.kegg.jp/kegg/download/kegtools.html>) is a freely available Java application for rendering glycan structures. It can be downloaded and installed locally as a platform-independent tool. This tool can be used in two different modes: “Compound mode” which can be used for drawing small molecules (similarly to any chemical structure drawing software), and “Glycan mode” which is dedicated for rendering glycan structures using different monosaccharide units. The simplest method for drawing involves a selection of monosaccharides and glycosidic linkages from an available list

to generate a glycan structure. Alternatively, a text box option provides a way to draw uncommon types of monosaccharides. The tool also contains templates from KEGG GLYCAN and their importation using their accession number. Besides, input files in KCF can be used while the output can be saved in LINUCS, KCF or an image in PNG format (Figure 8, bottom). The glycan structure in text format can be further used as a query for search in KEGG GLYCAN and CarbBank databases. Hence, KegDraw can be an option for the freely available tool for drawing and querying chemical structures. However, there are similar tools already available for glycan drawing with more advanced and acceptable notations.

Glycan builders

Sweet II. Sweet [58] is a web-based program for constructing 3D models of glycans from a sequence using standard nomenclature accessible at <http://www.glycosciences.de/modeling/>

[sweet2/doc/index.php](#) (Figure 9, top). This tool is available as a part of the glycosciences.de website, which also provides other options for analysing glycans in three-dimensional space. This program uses a glycan sequence in a standard format and generates a 3D model in the form of a .pdb file. The glycan input can come from a library of relevant oligosaccharides, available through one of the sub-menus. Alternatively, manual input is possible in three platforms adapted for increasing complexity. The model can be further minimised using MM2 [59] and MM3 [60] methods. The 3D models can be viewed using molecular viewers like JMol, WebMol-applet, Chemis3D-applet, etc. Besides, the program also generates additional files which can be used for molecular mechanics and molecular dynamics using molecular modelling tool like Tinker [61]. This tool is as a versatile tool for generating a 3D model for glycans.

GLYCAM-web (Carbohydrate Builder). Carbohydrate builder [65] is an online tool (at <http://glycam.org/>) for carbohydrate structure drawing and subsequent 3D structure building. With a flexible interface, it uses three methods for glycan building. The first method is manual building (“Carbohydrate Builder” button). It allows selection of monosaccharide, as well as defining linkages, branching and substitution (Figure 9, middle). The second method involves the use of a template library (using “Oligosaccharide libraries” button) containing commonly relevant structures (<http://glycam.org/Pre-builtLibraries.jsp>). The third option (direct input from a text sequence) becomes relevant when the glycan structure does not exist in the library or challenging to build due to structural complexity. In this case, a text for the oligosaccharide in GLYCAM-Web’s condensed notation can be entered as an input to create the glycan structure. Once the glycan is generated, the options include the solvation of the structure and the manual input of the glycosidic linkages. The tool allows structure minimisation and generates rotamers which can be visualised using JSmol viewer. Information about the force field that is used to build the structure is also provided. The multiple structures can be downloaded compressed as .tar, .gz or .zip files containing .pdb files. Similarly, the 2D image can be saved in GIF format. GLYCAM-web- Carbohydrate Builder can be used to prepare the system for MD simulation as it solvates the glycans and also generates the topology and coordinate files. In addition to its carbohydrate builder, Glycam-web consists of additional tools like glycoprotein builder and glycosaminoglycans (GAG) builder.

CHARMM-GUI (Glycan Reader and Modeler). The CHARMM-GUI (<http://www.charmm-gui.org>) is a web-based graphical user interface which provides various functional modules to prepare complex biomolecular systems and input files for molecular simulations. Glycan Reader and Modeler

[65–67] is a part of CHARMM-GUI (Figure 9, bottom) and available as a freely accessible online tool at <http://charmm-gui.org/input/glycan>. It can read input files in PDB, PDBx/mmCIF and CHARMM formats containing glycans and automatically detects the carbohydrate molecules and glycosidic linkage information. Alternatively, it can also read a glycan sequence (GRS format) to generate a 3D model and input files for MD simulation of the carbohydrate-only system. GRS carbohydrate sequences can be made through a straightforward interface: monosaccharides (20+ options) and their linkages are added incrementally from drop-down menus. A useful feature of this tool is the real-time rendering of the carbohydrate image: each added monosaccharide and modified linkage is directly reported to the image as well as to a text (GRS) format. Option for numerous chemical modifications is also available.

On the other hand, the Glycan Modeler allows in silico N-/O-glycosylation for glycan-protein complexes and generates a “most relevant” glycan structure through Glycan Fragment Database (GFDB) [68] search which gives proper orientations relative to the target protein. In the absence of target glycan sequence in GFDB, the structures are generated by using the valid internal coordinate information (averaged phi, psi, and omega glycosidic torsion angles) in the CHARMM force field. Input files for CHARMM can be generated for the purpose of MD simulation. Amongst other possible outputs, 3D representations of the glycans are available as .pdb files. This tool can be helpful for researchers to generate 2D depictions of a glycan and then obtain the corresponding 3D representation, which can be useful for modelling studies of glycans and glycoconjugates.

doGlycans. doGlycans [69] is a compilation of tools dedicated for preparing carbohydrate structures for atomistic simulations of glycoproteins, carbohydrate polymers and glycolipids using GROMACS [70,71] In the form of Python scripts; the tools are used to prepare the system, which generally includes the processing of a.pdb file using the *pdb2gmx* tool. Subsequently, a glycosylation model can be prepared for carbohydrate polymer simulation using the *prepreader.py* script. Similarly, the *doglycans.py* script can be used to develop models for glycoproteins and glycolipids. Together, these tools are called doGlycans toolset. Although doGlycans is highly flexible, it only uses the sugar units that are defined in GLYCAM. The topologies generated for glycosylated proteins and glycolipids are compatible with the OPLS [72] and AMBER [73] force fields. The topology for carbohydrate polymers is based on the GLYCAM force field. The user needs to provide the ceramide topology as input to generate the topologies for glycolipids. The tools contained in doGlycans create 3D models and simulation files as a starting point for more complex molecular simulation studies.

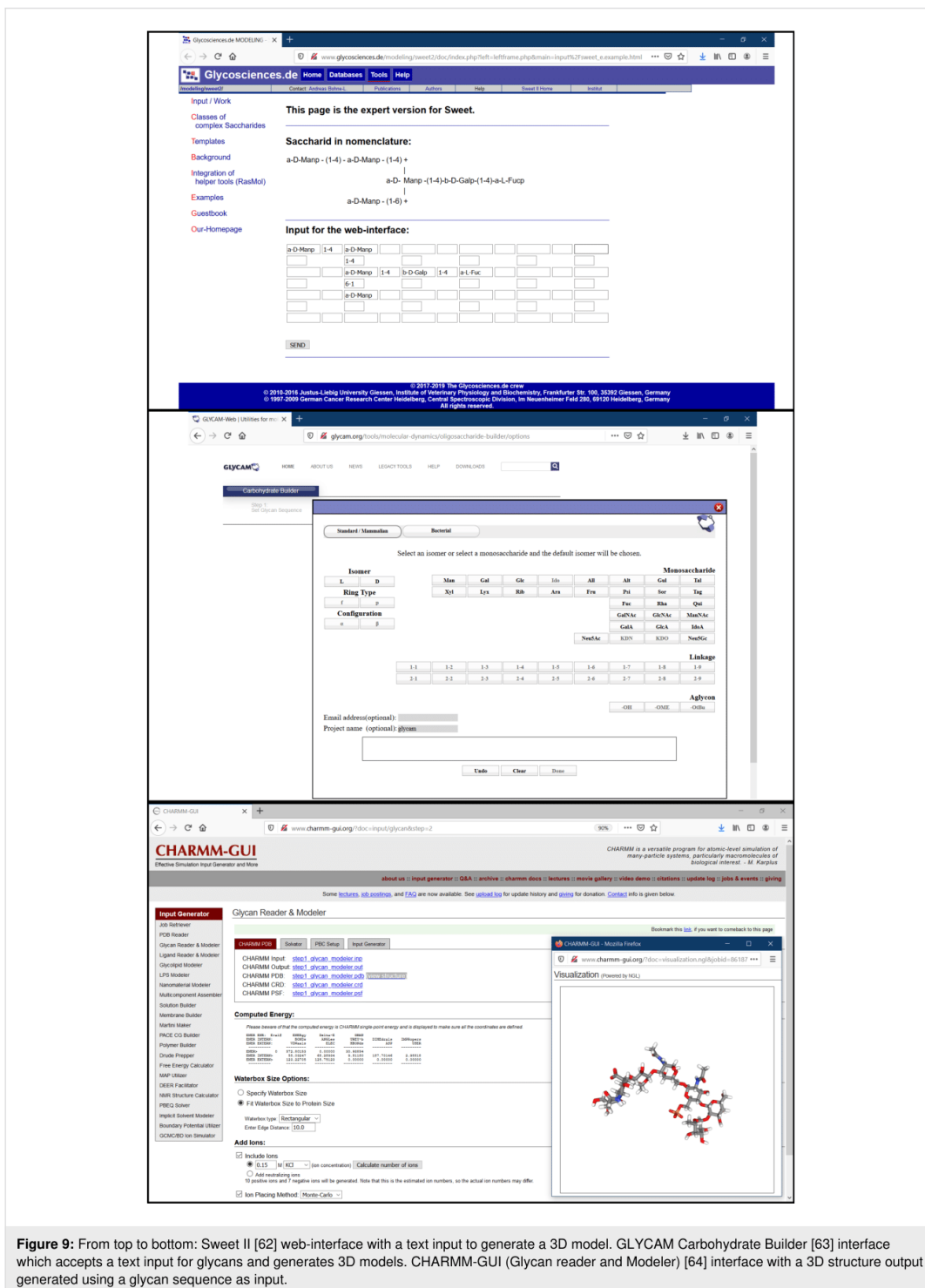


Figure 9: From top to bottom: Sweet II [62] web-interface with a text input to generate a 3D model. GLYCAM Carbohydrate Builder [63] interface which accepts a text input for glycans and generates 3D models. CHARMM-GUI (Glycan reader and Modeler) [64] interface with a 3D structure output generated using a glycan sequence as input.

RosettaCarbohydrate. Rosetta is a software suite for macromolecular modelling as an extensive collection of computer code mostly written in C++ and Python languages. Rosetta is available to academic and commercial researchers through a license available at <https://www.rosettacommons.org/software/license-and-download>. The licence is free for academic users. The tool runs best on Linux or macOS platforms only. It can be installed on a multiprocessor computing cluster to increase efficiency. RosettaCarbohydrate [74,75] tool provides the methods for general modelling and docking applications for glycans and glycoconjugates. The application accepts the standard PDB, GLYCAM, and GlycoWorkbench (.gws) file formats and the available utilities (codes) helps with the general problems in sampling, scoring, and nomenclature related to glycan modelling. It samples glycosidic bonds, ring forms, side-chain conformations, and utilises a glycan-specific term within its scoring function. The tool also consists of utilities for virtual glycosylation, protein–glyco–ligand docking, and glycan “loop” modelling. This tool is best for the researcher with basic knowledge and skills to work with a command-line interface (Linux).

PolysGlycanBuilder. PolysGlycanBuilder [76] is a web-based tool (<http://glycan-builder.cermav.cnrs.fr/>) with an interactive and more usable interface (Figure 10). The software translates a glycan sequence or polysaccharide repeat unit into the coordi-

nate set of the corresponding tertiary structure, in one or several of its low energy conformations. The construction follows an intuitive scheme which is as close as possible to the way glycoscientists draw the sequence of their structures. The simplest method for model building involves dragging and dropping monosaccharide units to the canvas or workspace grid. The software displays rows of monosaccharides in the form of standard SNFG symbols with 3D information (furanose/pyranose shape, configuration, anomericity, and ring conformation). Glycosidic linkages can be easily defined, as the values of the dihedral angles (Φ , Ψ , Ω). They can be manually set or extracted from a database of low energy conformations of 600 disaccharide segments. The monosaccharides have been subjected to geometry optimisation using molecular mechanics approach. For a given input sequence, the corresponding 3D coordinates are generated at the PDB format. Within the process of construction, the structure is displayed via the LiteMol and eventually optimised to remove any steric clashes. The image for the glycan can be downloaded and saved in SVG format. Keeping the glycan/polysaccharide structure in text format (condensed IUAPC, GlycoCT, SNFG and INP) offers several ways to connect to other applications. Other than drag and drop method, PolysGlycan-Builder also accepts input of files in INP, IUPAC and GlycoCT formats. An interactive interface accompanies the application, which makes it more versatile for glycan drawing and 3D model building.

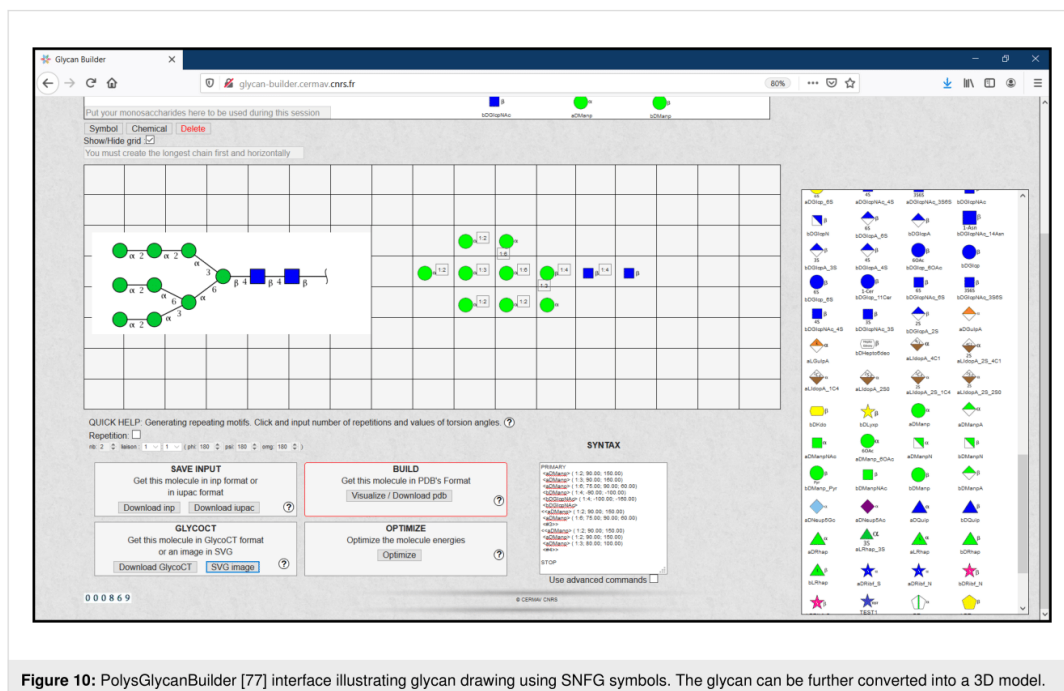


Figure 10: PolysGlycanBuilder [77] interface illustrating glycan drawing using SNFG symbols. The glycan can be further converted into a 3D model.

Displaying 3D structures of glycans

3D-SNFG VMD interface and visualisation algorithms. The recently introduced 3D-Symbol Nomenclature for Glycans (3D-SNFG) [15] allows the representation of carbohydrates in an unusual way: the SNFG symbols are added to a three-dimensional structure. The 3D-SNFG script must be integrated into the visual molecular dynamics (VMD) [21,78] viewer software to enable the representation of glycans as large SNFG-matching 3D shapes that can either replace the molecular monosaccharides or stay lodged at the geometric centre of the cycle (Figure 11, top left). Upon the input of a glycan-containing structure (in PDB format), the integrated script in VMD automatically recognises the common monosaccharide names and generates the 3D shapes. The embedded script also enables shortcuts keys from keyboard to quickly change between large and small 3D-SNFG shapes and also label the reducing terminus. The 3D structure displayed in VMD can be saved as a .bmp image file. Thanks to 3D-SNFG, the standardised representation of glycan structures can finally take a step into the 3D space. The obtained images can become very useful for quick assessment of 3D glycan models.

In addition to the 3D-SNFG script, *PaperChain* and *Twister* [83] are two visualisation algorithms available with the Visual Molecular Dynamics (VMD) package. These algorithms are useful to visualize complex cyclic molecules and multi-branched polysaccharides. {Cross, 2009 #69} *PaperChain* displays rings in a molecular structure with a polygon and colours them according to the ring pucker. The other algorithm (*Twister*) traces glycosidic bonds in a ribbon representation that twists and changes its orientation according to the relative position of following sugar residues, hence provides an important conformational detail in polysaccharides. Combination of these algorithms with other visualisation features available in VMD can enhance the flexibility of displaying structural details of glycoconjugate, glycoprotein and cyclic structures.

LiteMol. The LiteMol [22] viewer is a freely available web application (Figure 11, top right) for 3D visualisation of macromolecules and other related data. LiteMol enables standard visualisation of macromolecules in different representation modes like surface, cartoons, ball-and-stick, etc. The software can be accessed at v.litemol.org and also available for integration in a webpage from the github (<https://github.com/dsehnal/LiteMol>). LiteMol is compatible with all modern browsers without the support of additional plugins. The viewer automatically depicts any carbohydrate residues and displays 3D structures of carbohydrates with 3D-SNFG symbols, which allows the viewer to identify the monosaccharides readily. The presented structure can be saved as a .png image file. Any monosaccharide with a residue name in PDB can be visualised

using 3D-SNFG in LiteMol. However, a significant portion of the carbohydrates may contain some form of error in annotation, which would result in either no symbol or an incorrect symbol. Although LiteMol is an efficient and rapid 3D viewer for glycans, 3D representation does not provide any information about the glycosidic linkage type (e.g. α 1-3 or β 1-4). Also, it does not display any information about connection and configuration. If this information is required, returning to the classic molecular representation is possible.

PyMOL- Azahar plugin. Azahar [84] is a plugin in PyMOL [85] which enables building, visualization and analysis of glycans and glycoconjugates. This tool is based on Python and provides additional computing environment within the PyMOL package. The tool is provided with a template list of saccharide structures to facilitate structure building and visualisation. The interface provides three option menus to assist glycan structure building. The two first options help to specify residues to be connected from a list of available templates, and the third one allows selection of the chemical bond between the residues. The visualisation using PyMOL includes three cartoon-like representations. These display modes provided in the tool simplify the representation of glycan structures in cartoon, wire and bead representations. In cartoon and wire representations, the rings in sugars are shown as non-flat polygons connected by rods while in the bead representation mode, these cycles are represented as a sphere. In addition of visualization of static structures, the tool also allows analysis of trajectories of MD simulations. The tool can be used for conformational search using a Monte Carlo approach [86]. The conformational search is done by perturbing a torsional angle, followed by an energy minimization using the MMFF94 force field. Azahar is freely accessible from <http://www.pymolwiki.org/index.php/Azahar>.

UnityMol/SweetUnityMol. Sweet UnityMol [32] is a molecular structure viewer (Figure 11, middle) developed from the game engine Unity3D. The software is available for free download (https://sourceforge.net/projects/unitymol/files/UnityMol_1.0.37/) from the SourceForge project website. It can be installed in Mac, Windows and Linux platforms. The program reads files in PDB, mmCIF, Mol2, GRO, XYZ, and SDF formats, OpenDX potential maps and XTC trajectory files. It efficiently displays specific structural features for the simplest to the most complex carbohydrate-containing biomolecules. Sweet UnityMol displays 3D carbohydrate structures with different modes of representation, such as: liquorice, ball-and-stick, hyperBalls, RingBlending, hydrophilic/hydrophobic character of sugar face etc. The most recent version is fully compatible with the SNFG colour coding, which also uses acceptable pictorial representation, generally used in carbohydrate chemistry, biochemistry and glycobiology.

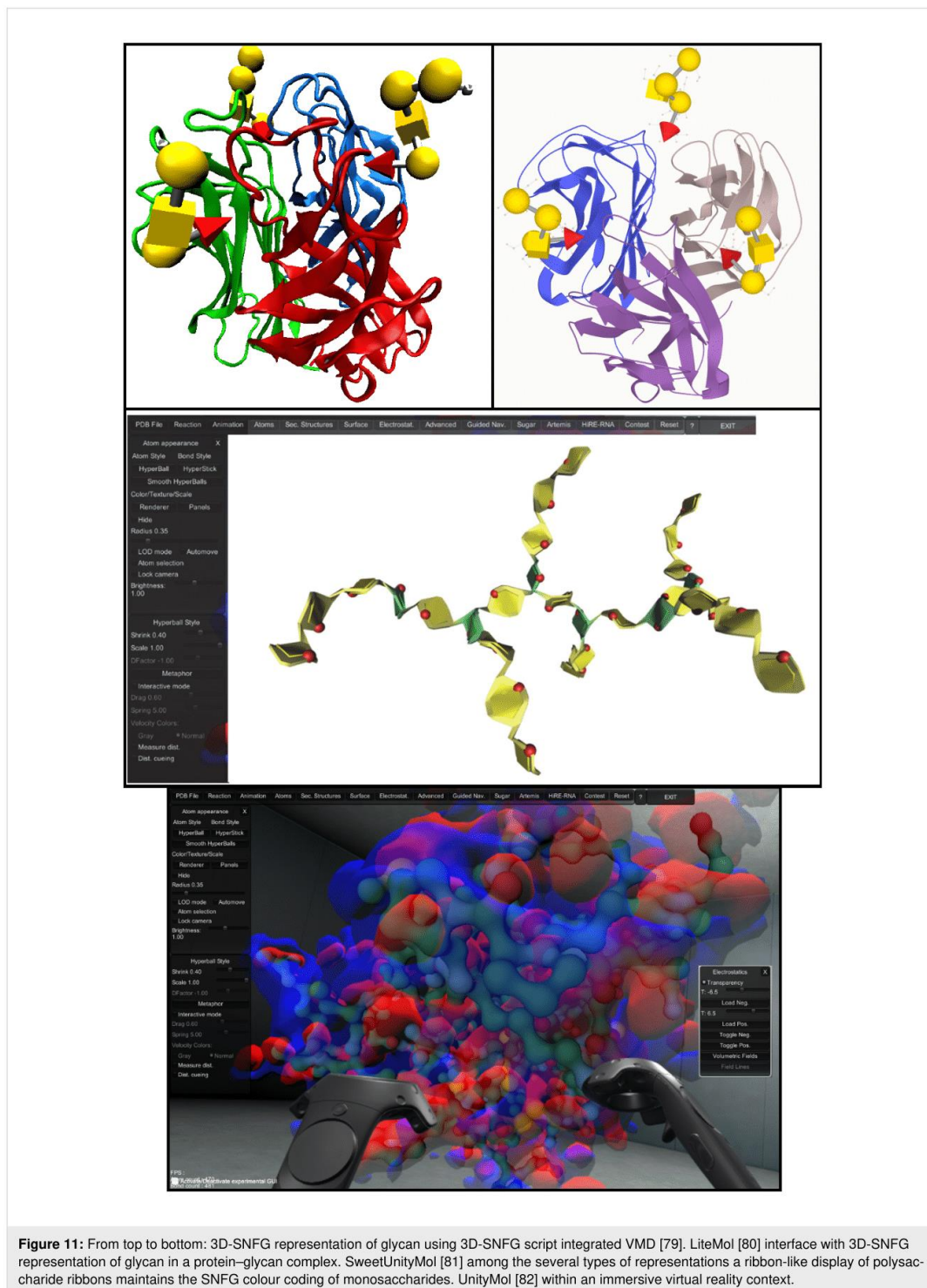


Figure 11: From top to bottom: 3D-SNFG representation of glycan using 3D-SNFG script integrated VMD [79]. LiteMol [80] interface with 3D-SNFG representation of glycan in a protein–glycan complex. SweetUnityMol [81] among the several types of representations a ribbon-like display of polysaccharide ribbons maintains the SNFG colour coding of monosaccharides. UnityMol [82] within an immersive virtual reality context.

SweetUnityMol provides a continuum from the conventional ways to depict the primary structures of complex carbohydrates all the way to visualising their 3D structures. Several options are offered to the user to select the most relevant type of depictions, including new features, such as “Coarse-Grain” representation while keeping the option to display the details of the atomic representations. Powerful rendering methods produce high-quality images of molecular structures, bio-macromolecular surfaces and molecular interactions.

A recently developed version of UnityMol has been implemented with the immersive Virtual Reality context using head-mounted displays [87]. It offers high-quality visual representations, ease of interactions with multiple molecular objects, powerful tools for visual manipulations, accompanied by the evaluation of intermolecular interactions. Consequently, simultaneous investigations of multiple objects such as macromolecular interactions gain in efficiency and accuracy. (Figure 11, bottom).

Conclusion

The set of computational tools presented above illustrates the rich contributions of a community devoted to enabling the accurate representation of complex carbohydrates via the development and implementation of a versatile informatics toolbox. These legitimate efforts aim at facilitating communication within the scientific community. To establish a comparative analysis of the several available applications, we evaluated 17 selected items that characterise best their availability, implementation, maintenance and field of use. The comparative analysis of tools could be useful for glycobiologists or any researcher looking for a ready to use, simple application for the sketching, building and display of glycans.

This article provides an overview of the computational tools and resources available for glycan sketching, building and representing. It also provides a descriptive analysis of the recently developed software tools dedicated explicitly to glycans and glycoconjugates. The newly developed tools are more advanced and use the standard nomenclature and symbols for glycan representation. These tools can further help to standardise the description of glycans in research, communication and databases.

Supporting Information

Supporting Information File 1

Features of glycan sketchers, builders and viewers.
[<https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-16-199-S1.pdf>]

Acknowledgements

Appreciation is extended to Drs. A. Imberty, A. Varrot, L. Belvisi and A. Bernardi for their support.

Funding

This research was performed within the framework of the PhD4GlycoDrug Innovative Training Network and was funded from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765581. The work was supported by the Cross-Disciplinary Program Glyco@Alps, within the framework “Investissement d’Avenir” program [ANR-15IDEX-02].

ORCID® iDs

Kanhaya Lal - <https://orcid.org/0000-0001-8555-7948>

Rafael Bermeo - <https://orcid.org/0000-0002-4451-878X>

Serge Perez - <https://orcid.org/0000-0003-3464-5352>

References

1. Alocci, D.; Lisacek, F.; Perez, S. *A Traveler's Guide to Complex Carbohydrates in the Cyber Space*. http://www.glycopedia.eu/IMG/pdf/traveler_s_guide_to_cyber_space.pdf
2. Perez, S.; Aoki-Kinoshita, K. F. *Development of Carbohydrate Nomenclature and Representation*; Springer, 2017; pp 7–25. doi:10.1007/978-4-431-56454-6_2
3. Kornfeld, S.; Li, E.; Tabas, I. *J. Biol. Chem.* **1978**, *253*, 7771–7778.
4. Royle, L.; Dwek, R. A.; Rudd, P. M. *Curr. Protoc. Protein Sci.* **2006**, *43*, 12.6.1–12.6.45. doi:10.1002/0471140864.ps1206s43
5. Harvey, D. J.; Merry, A. H.; Royle, L.; Campbell, M. P.; Dwek, R. A.; Rudd, P. M. *Proteomics* **2009**, *9*, 3796–3801. doi:10.1002/pmic.200900096
6. Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Marth, J. D.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E. *Proteomics* **2009**, *9*, 5398–5399. doi:10.1002/pmic.200900708
7. Neelamegham, S.; Aoki-Kinoshita, K.; Bolton, E.; Frank, M.; Lisacek, F.; Lütke, T.; O’Boyle, N.; Packer, N. H.; Stanley, P.; Toukach, P.; Varki, A.; Woods, R. J.; Darvill, A.; Dell, A.; Henrissat, B.; Bertozzi, C.; Hart, G.; Narimatsu, H.; Freeze, H.; Yamada, I.; Paulson, J.; Prestegard, J.; Marth, J.; Vliegthart, J. F. G.; Etzler, M.; Aebi, M.; Kanehisa, M.; Taniguchi, N.; Edwards, N.; Rudd, P.; Seeberger, P.; Mazumder, R.; Ranzinger, R.; Cummings, R.; Schnaar, R.; Perez, S.; Kornfeld, S.; Kinoshita, T.; York, W.; Knirel, Y. *Glycobiology* **2019**, *29*, 620–624. doi:10.1093/glycob/cwz045
8. Varki, A.; Cummings, R. D.; Aebi, M.; Packer, N. H.; Seeberger, P. H.; Esko, J. D.; Stanley, P.; Hart, G.; Darvill, A.; Kinoshita, T.; Prestegard, J. J.; Schnaar, R. L.; Freeze, H. H.; Marth, J. D.; Bertozzi, C. R.; Etzler, M. E.; Frank, M.; Vliegthart, J. F. G.; Lütke, T.; Perez, S.; Bolton, E.; Rudd, P.; Paulson, J.; Kanehisa, M.; Toukach, P.; Aoki-Kinoshita, K. F.; Dell, A.; Narimatsu, H.; York, W.; Taniguchi, N.; Kornfeld, S. *Glycobiology* **2015**, *25*, 1323–1324. doi:10.1093/glycob/cwv091
9. Aoki, K. F.; Yamaguchi, A.; Ueda, N.; Akutsu, T.; Mamitsuka, H.; Goto, S.; Kanehisa, M. *Nucleic Acids Res.* **2004**, *32*, W267–W272. doi:10.1093/nar/gkh473

10. Ceroni, A.; Dell, A.; Haslam, S. M. *Source Code Biol. Med.* **2007**, *2*, No. 3. doi:10.1186/1751-0473-2-3
11. Damerell, D.; Ceroni, A.; Maass, K.; Ranzinger, R.; Dell, A.; Haslam, S. M. Annotation of Glycomics MS and MS/MS Spectra Using the GlycoWorkbench Software Tool. In *Glycoinformatics. Methods in Molecular Biology*; Lütteke, T.; Frank, M., Eds.; Humana Press: New York, NY, 2015; Vol. 1273, pp 3–15. doi:10.1007/978-1-4939-2343-4_1
12. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242. doi:10.1093/nar/28.1.235
13. Engelsen, S. B.; Hansen, P. I.; Pérez, S. *Biopolymers* **2014**, *101*, 733–743. doi:10.1002/bip.22449
14. Alloci, D.; Suchánková, P.; Costa, R.; Hory, N.; Mariethoz, J.; Vařeková, R.; Toukach, P.; Lisacek, F. *Molecules* **2018**, *23*, 3206. doi:10.3390/molecules23123206
15. Thieker, D. F.; Hadden, J. A.; Schulten, K.; Woods, R. J. *Glycobiology* **2016**, *26*, 786–787. doi:10.1093/glycob/cww076
16. McNaught, A. D. *Adv. Carbohydr. Chem. Biochem.* **1997**, *52*, 44–177. doi:10.1016/s0065-2318(08)60090-6
17. Akune, Y.; Hosoda, M.; Kaiya, S.; Shinmachi, D.; Aoki-Kinoshita, K. F. *OMICS* **2010**, *14*, 475–486. doi:10.1089/omi.2009.0129
18. Hashimoto, K.; Goto, S.; Kawano, S.; Aoki-Kinoshita, K. F.; Ueda, N.; Hamajima, M.; Kawasaki, T.; Kanehisa, M. *Glycobiology* **2006**, *16*, 63R–70R. doi:10.1093/glycob/cwj010
19. Maeda, M.; Fujita, N.; Suzuki, Y.; Sawaki, H.; Shikanai, T.; Narimatsu, H. JCGGDB: Japan Consortium for Glycobiology and Glycotechnology Database. In *Glycoinformatics. Methods in Molecular Biology*; Lütteke, T.; Frank, M., Eds.; Humana Press: New York, NY, 2015; Vol. 1273, pp 161–179. doi:10.1007/978-1-4939-2343-4_12
20. Damerell, D.; Ceroni, A.; Maass, K.; Ranzinger, R.; Dell, A.; Haslam, S. M. *Biol. Chem.* **2012**, *393*, 1357–1362. doi:10.1515/hsz-2012-0135
21. Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38. doi:10.1016/0263-7855(96)00018-5
22. Sehmal, D.; Grant, O. C. *J. Proteome Res.* **2019**, *18*, 770–774. doi:10.1021/acs.jproteome.8b00473
23. Foley, B. L.; Tessier, M. B.; Woods, R. J. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 652–697. doi:10.1002/wcms.89
24. Mallajosyula, S. S.; Guvench, O.; Hatcher, E.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2012**, *8*, 759–776. doi:10.1021/ct200792v
25. Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2009**, *5*, 2353–2370. doi:10.1021/ct900242e
26. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008**, *29*, 622–655. doi:10.1002/jcc.20820
27. Lins, R. D.; Hünenberger, P. H. *J. Comput. Chem.* **2005**, *26*, 1400–1412. doi:10.1002/jcc.20275
28. Molinero, V.; Goddard, W. A. *J. Phys. Chem. B* **2004**, *108*, 1414–1427. doi:10.1021/jp0354752
29. Hergert, S.; Ranzinger, R.; Maass, K.; Lieth, C.-W. v. d. *Carbohydr. Res.* **2008**, *343*, 2162–2171. doi:10.1016/j.carres.2008.03.011
30. Tanaka, K.; Aoki-Kinoshita, K. F.; Kotera, M.; Sawaki, H.; Tsuchiya, S.; Fujita, N.; Shikanai, T.; Kato, M.; Kawano, S.; Yamada, I.; Narimatsu, H. *J. Chem. Inf. Model.* **2014**, *54*, 1558–1566. doi:10.1021/ci400571e
31. Cheng, K.; Zhou, Y.; Neelamegham, S. *Glycobiology* **2017**, *27*, 200–205. doi:10.1093/glycob/cww115
32. Perez, S.; Tubiana, T.; Imbert, A.; Baaden, M. *Glycobiology* **2015**, *25*, 483–491. doi:10.1093/glycob/cwu133
33. Yamada, I.; Shiota, M.; Shinmachi, D.; Ono, T.; Tsuchiya, S.; Hosoda, M.; Fujita, A.; Aoki, N. P.; Watanabe, Y.; Fujita, N.; Angata, K.; Kaji, H.; Narimatsu, H.; Okuda, S.; Aoki-Kinoshita, K. F. *Nat. Methods* **2020**, *17*, 649–650. doi:10.1038/s41592-020-0879-8
34. Tiemeyer, M.; Aoki, K.; Paulson, J.; Cummings, R. D.; York, W. S.; Karlsson, N. G.; Lisacek, F.; Packer, N. H.; Campbell, M. P.; Aoki, N. P.; Fujita, A.; Matsubara, M.; Shinmachi, D.; Tsuchiya, S.; Yamada, I.; Pierce, M.; Ranzinger, R.; Narimatsu, H.; Aoki-Kinoshita, K. F. *Glycobiology* **2017**, *27*, 915–919. doi:10.1093/glycob/cwx066
35. Lütteke, T.; Bohne-Lang, A.; Loss, A.; Goetz, T.; Frank, M.; von der Lieth, C.-W. *Glycobiology* **2006**, *16*, 71R–81R. doi:10.1093/glycob/cwj049
36. *Sugar Sketcher*. <https://glycoproteome.expasy.org/sugarsketcher/> (accessed April 2020).
37. *GlycoGlyph*. <https://glycotoolkit.com/Tools/GlycoGlyph/> (accessed April 2020).
38. *GlyTouCan*. <https://glytoucan.org/> (accessed April 2020).
39. Mehta, A. Y.; Cummings, R. D. *Bioinformatics* **2020**, *36*, 3613–3614. doi:10.1093/bioinformatics/btaa190
40. Tsuchiya, S.; Aoki, N. P.; Shinmachi, D.; Matsubara, M.; Yamada, I.; Aoki-Kinoshita, K. F.; Narimatsu, H. *Carbohydr. Res.* **2017**, *445*, 104–116. doi:10.1016/j.carres.2017.04.015
41. Toukach, P. V.; Egorova, K. S. *Nucleic Acids Res.* **2016**, *44*, D1229–D1236. doi:10.1093/nar/gkv840
42. Kikuchi, N.; Kameyama, A.; Nakaya, S.; Ito, H.; Sato, T.; Shikanai, T.; Takahashi, Y.; Narimatsu, H. *Bioinformatics* **2005**, *21*, 1717–1718. doi:10.1093/bioinformatics/bti152
43. Doubet, S.; Bock, K.; Smith, D.; Darvill, A.; Albersheim, P. *Trends Biochem. Sci.* **1989**, *14*, 475–477. doi:10.1016/0968-0004(89)90175-8
44. Banin, E.; Neuberger, Y.; Altschuler, Y.; Halevi, A.; Inbar, O.; Nir, D.; Dukler, A. *Trends Glycosci. Glycotechnol.* **2002**, *14*, 127–137. doi:10.4052/tigg.14.127
45. Cooper, C. A.; Harrison, M. J.; Wilkins, M. R.; Packer, N. H. *Nucleic Acids Res.* **2001**, *29*, 332–335. doi:10.1093/nar/29.1.332
46. *GlycanBuilder2*. Downloaded from <http://www.rings.t.soka.ac.jp/downloads.html> (accessed April 2020).
47. *SugarBind GlycanBuilder*. <https://sugarbind.expasy.org/builder> (accessed April 2020).
48. *DrawRINGS*. <http://www.rings.t.soka.ac.jp/drawRINGS-js/> (accessed April 2020).
49. Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. *J. Proteome Res.* **2008**, *7*, 1650–1659. doi:10.1021/pr7008252
50. Mariethoz, J.; Khatib, K.; Alloci, D.; Campbell, M. P.; Karlsson, N. G.; Packer, N. H.; Mullen, E. H.; Lisacek, F. *Nucleic Acids Res.* **2016**, *44*, D1243–D1250. doi:10.1093/nar/gkv1247
51. *DrawGlycan-SNFG*. <http://www.virtualglycome.org/DrawGlycan/> (accessed April 2020).
52. *Glycano*. <http://glycano.cs.uct.ac.za/> (accessed April 2020).
53. *GlycoEditor*. <https://jcgdb.jp/idb/flash/GlycoEditor.jsp> (accessed April 2020).
54. Cheng, K.; Pawlowski, G.; Yu, X.; Zhou, Y.; Neelamegham, S. *Bioinformatics* **2019**. doi:10.1093/bioinformatics/btz819

55. Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Darvill, A. G.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H., *Essentials of Glycobiology [Internet]*, 3 ed.; Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press: 2015–2017.
56. *Glyco.me SugarBuilder*. <https://beta.glyco.me/sugarbuilder> (accessed April 2020).
57. *KegDraw*. Downloaded from <https://www.kegg.jp/kegg/download/kegtools.html> (accessed April 2020).
58. Bohne, A.; Lang, E.; von der Lieth, C. W. *Bioinformatics* **1999**, *15*, 767–768. doi:10.1093/bioinformatics/15.9.767
59. Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127–8134. doi:10.1021/ja00467a001
60. Lii, J. H.; Allinger, N. L. *J. Am. Chem. Soc.* **1989**, *111*, 8566–8575. doi:10.1021/ja00205a002
61. Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardère, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. *J. Chem. Theory Comput.* **2018**, *14*, 5273–5289. doi:10.1021/acs.jctc.8b00529
62. *Sweet*. <http://www.glycosciences.de/modeling/sweet2/doc/index.php> (accessed April 2020).
63. *GLYCAM Web*. (2005–2020) Complex Carbohydrate Research Center, University of Georgia, Athens, GA. (<http://glycam.org>).
64. *CHARMM-GUI Glycan Reader & Modeler*. <http://www.charmm-gui.org/?doc=input/glycan> (accessed April 2020).
65. Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *J. Comput. Chem.* **2008**, *29*, 1859–1865. doi:10.1002/jcc.20945
66. Jo, S.; Song, K. C.; Desaire, H.; MacKerell, A. D., Jr.; Im, W. *J. Comput. Chem.* **2011**, *32*, 3135–3141. doi:10.1002/jcc.21886
67. Park, S.-J.; Lee, J.; Qi, Y.; Kern, N. R.; Lee, H. S.; Jo, S.; Joung, I.; Joo, K.; Lee, J.; Im, W. *Glycobiology* **2019**, *29*, 320–331. doi:10.1093/glycob/cwz003
68. Jo, S.; Im, W. *Nucleic Acids Res.* **2013**, *41*, D470–D474. doi:10.1093/nar/gks987
69. Danne, R.; Poojari, C.; Martinez-Seara, H.; Rissanen, S.; Lolicato, F.; Rög, T.; Vattulainen, I. *J. Chem. Inf. Model.* **2017**, *57*, 2401–2406. doi:10.1021/acs.jcim.7b00237
70. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447. doi:10.1021/ct700301q
71. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718. doi:10.1002/jcc.20291
72. Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. *J. Chem. Theory Comput.* **2016**, *12*, 281–296. doi:10.1021/acs.jctc.5b00864
73. *AMBER 2018*; University of California: San Francisco, 2018.
74. Labonte, J. W.; Adolf-Bryfogle, J.; Schief, W. R.; Gray, J. J. *J. Comput. Chem.* **2017**, *38*, 276–287. doi:10.1002/jcc.24679
75. Frenz, B.; Rämisch, S.; Borst, A. J.; Walls, A. C.; Adolf-Bryfogle, J.; Schief, W. R.; Veessler, D.; DiMaio, F. *Structure* **2019**, *27*, 134–139.e3. doi:10.1016/j.str.2018.09.006
76. Perez, S.; Rivet, A., *Methods in Molecular Biology, Glycoinformatics, Methods and Protocols*. 2nd ed.; 2020, (in press).
77. *PolysGlycanBuilder*. <http://glycan-builder.cermav.cnrs.fr/> (accessed April 2020).
78. Kuttel, M.; Gain, J.; Burger, A.; Eborn, I. *J. Mol. Graphics Modell.* **2006**, *25*, 380–388. doi:10.1016/j.jmgs.2006.02.007
79. *3D-SNFG*. Downloaded from <http://glycam.org/3d-snfg> (accessed April 2020).
80. *LiteMol*. <https://v.litemol.org/> (accessed April 2020).
81. *SweetUnityMol*. <https://sourceforge.net/projects/unitymol/files/OtherVersions/UnityMol-676-SweetUnityMol/> (accessed April 2020).
82. *UnityMol*. <https://sourceforge.net/projects/unitymol/files/> (accessed April 2020).
83. Cross, S.; Kuttel, M. M.; Stone, J. E.; Gain, J. E. *J. Mol. Graphics Modell.* **2009**, *28*, 131–139. doi:10.1016/j.jmgs.2009.04.010
84. Arroyuelo, A.; Vila, J. A.; Martin, O. A. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 619–624. doi:10.1007/s10822-016-9944-x
85. *PyMOL: An open-source molecular graphics tool*; DeLano Scientific, 2002. <http://www.pymol.org>.
86. Li, Z.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U. S. A.* **1987**, *84*, 6611–6615. doi:10.1073/pnas.84.19.6611
87. Martinez, X.; Chavent, M.; Baaden, M. *Biochem. Soc. Trans.* **2020**, *48*, 499–506. doi:10.1042/bst20190621

License and Terms

This is an Open Access article under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>). Please note that the reuse, redistribution and reproduction in particular requires that the authors and source are credited.

The license is subject to the *Beilstein Journal of Organic Chemistry* terms and conditions: (<https://www.beilstein-journals.org/bjoc>)

The definitive version of this article is the electronic one which can be found at: <https://doi.org/10.3762/bjoc.16.199>

Table S1. Schematic summary of important features of glycan sketchers, builders and viewers.

S.No.	Software Tool	2D sketchers and builders														3D Builders					3D (representation) viewers				
		Sugar Sketcher	UGraph	Glyco Glyph	Glycan Builder2 SNFG	Sugarbird Glycan Builder	Draw Ring5	Draw Glycan SNFG	Glycano	Glyco Editor	Glycone Sugar Builder	KeyDraw	Polys Glycan Builder	Sweet	CHARMM Gui	Glycan Carbo Builder	do Glycans	Rosetta Carbohydrate	Carb Builder	3D-SNFG	Pymol	Sweet UnityMol	LiteMol		
1	Available online	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					✓		✓		
2	Download for local installation	✓		✓	✓		✓	✓			✓						✓	✓	✓	✓	✓	✓	✓		
3	Instructions		✓	✓	✓	✓	✓				✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓		
4	Self-explanatory interface	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓								✓		
5	Image output (file type)	.SVG	.SVG	.SVG	.PNG, .SVG, .BMPPNG (save as)	.JPG (save as)	.PNG, .SVG	(screen shot)	.PNG, .SVG	.PNG	.SVG			(screen shot)	.GIF (save as)				.BMP	.PNG	.PNG	.PNG		
6	SNFG colours	CMYK	CMYK	RGB	RGB	CMYK	RGB		RGB	RGB		CMYK			CMYK				CMYK	RGB	RGB		RGB		
7	Oxford linkage geometry	✓			✓	✓	✓	✓	✓	✓		✓													
8	Functionalization	✓		✓	✓		✓	✓	✓		✓			✓		✓	✓	✓			✓				
9	Repeating units []	✓			✓	✓	✓	✓	✓		✓										✓				
10	Linking as glyco-conjugates				✓		✓				✓					✓	✓	✓	✓	✓	✓	✓	✓		
11	Text input	✓	✓	✓	✓	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓	✓				✓		
12	Option to modify by coding	✓	✓	✓			✓					✓	✓	✓	✓	✓				✓	✓				
13	3D-model visualization											✓	✓	✓	✓	✓				✓	✓	✓	✓		
14	3D model output											✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
15	Text/String output (any format)	✓	✓	✓	✓	✓	✓	✓	✓		✓				✓	✓			✓	✓					
16	Glycan library	✓		✓	✓	✓	✓		✓												✓				
17	No. of templates (monosaccharides)	70+	15+	80+	70+	50+	70+	60+	40+	10+	10+	100+	40+	20+	30+				40+				40+		

8. Identification of druggable allosteric pockets in β -propeller lectins

In addition to the research work related to my PhD project, I have been working on another side project with the collaborators of Dr. Imberty at the Max Planck Institute of Colloids and Interfaces, Department of Biomolecular Systems, Germany. The work has been submitted for publication as a research article in a scientific journal.

In this research, we identified a druggable allosteric binding site in a lectin from the opportunistic human pathogen *Burkholderia ambifaria* (BamBL)¹ using the SiteMap² analysis and docking predictions. The computational analysis combined with experimental studies revealed a promising compound (dissociation constant of 0.3 ± 0.1 mM) binding at the allosteric site. Interestingly, the fragment binding at the allosteric site affected the carbohydrate-binding site as determined by protein-observed fluorine NMR (PrOF).³ These findings were further supported by the studies involving site-directed mutagenesis in the orthosteric and secondary pocket which showed the effect on fragment binding, indicating additional insights into the communication path between the sites. In addition, computational and experimental studies performed on the structurally similar β -propeller lectins from *Ralstonia solanacearum* (RSL)⁴ and *Aspergillus fumigatus* (AFL)⁵⁻⁶ also suggest the presence of druggable secondary pockets. These observations can be useful for drug-discovery campaigns to develop allosteric inhibitors for bacterial and fungal lectins as a new therapeutic approach against antibiotic-resistant pathogens.

The computational approaches used to perform the studies are briefly discussed in this chapter.

8.1 Methods

8.1.1 Binding site prediction

The crystal structures of BamBL (PDB code 3ZZV and 3ZW0) were used for prediction of the possible secondary binding sites using the SiteMap tool. SiteMap creates a grid of points on the protein surface based on depth, size, van der Waals interaction energy, hydrophilicity and hydrophobicity and assign a single scoring function (SiteScore) to the potential druggable regions. The score helps to assess a site's propensity for ligand binding and prioritize the pharmaceutically relevant regions in the target protein. For BamBL, the calculations identified three regions at the interface in the trimer as potential druggable sites. These sites were explored further, thus the docking model was set up involving the key residues in the identified region. The same approach was applied to predict the druggable sites in the apo (PDB code 3ZI8) and holo (5AJB) forms of another bacterial lectin (RSL) from *Ralstonia solanacearum*. In addition, the calculations were performed on a fungal lectin known as AFL (PDB code 4AGI) from *Aspergillus fumigatus*.

8.1.2 Preparation of protein model

All the calculations were performed using the Schrödinger Suite through Maestro (version 2018-1) graphical interface.⁷ Atomic coordinates from the high resolution crystal structures of BamBL (PDB code 3ZZV and 3ZW0) were taken from the Protein Data Bank.⁸ The asymmetric unit contains three peptide chains and carbohydrate ligands, around a 3-fold pseudo axis of symmetry. The water molecules were removed and hydrogen atoms were added. pKa was predicted for protein residues using the PROPKA⁹⁻¹¹ method at pH 7.4. Protonation state (δ -nitrogen protonated) was assigned to the histidine (His58) residue. Finally, the complex was subjected to restrained minimization with convergence of heavy atoms to an RMSD of 0.3 Å using the OPLS3 force field.¹²

8.1.3 Preparation of ligand models

All (three) ligands were prepared for docking using the LigPrep¹³ tool and generated tautomers, stereoisomers and protonation states at pH 7.4. The calculations yield 13 structures.

8.1.4 Models for docking study

For docking grid generation (using PDB 3ZW0), the centroids of residues from chain B (Gly67, Thr69, Gly86, Leu87) and chain C (Thr18, Asn20, Lys23, Thr25) were selected to define a cubic grid box with dimensions 32×32×32 Å. The grid was used for docking studies using extra precision (XP) and standard precision (SP) scoring functions. All the calculations were accomplished by Glide (version 7.8)¹⁴ using the flexible docking approach.

8.2 Results

8.2.1 Binding site analysis

The crystal structures of BambL show that it forms a trimer involving three identical chains. The complex with oligosaccharides revealed 6 binding sites in both the PDB structures (PDB code 3ZZV and 3ZW0). In addition, the SiteMap tool identified three secondary sites (**Figure 9.1**) which can potentially host the ligands. The predicted sites located at the interface of the monomers near C-terminal form a narrow channel near the termini involving residues Thr18, Asn20, Lys23, Thr25, Gly67, Thr69, Gly86 and Leu87. The deep cavity can maximize the occupancy of suitable ligands. The binding site surrounded by hydrophilic residues makes it suitable to accommodate ligands with polar groups. Comparison of the key residues in both PDB structures shows differences in orientation of side chains, probably due to their location at the surface and flexible terminal region of the protein (**Figure 9.2**).

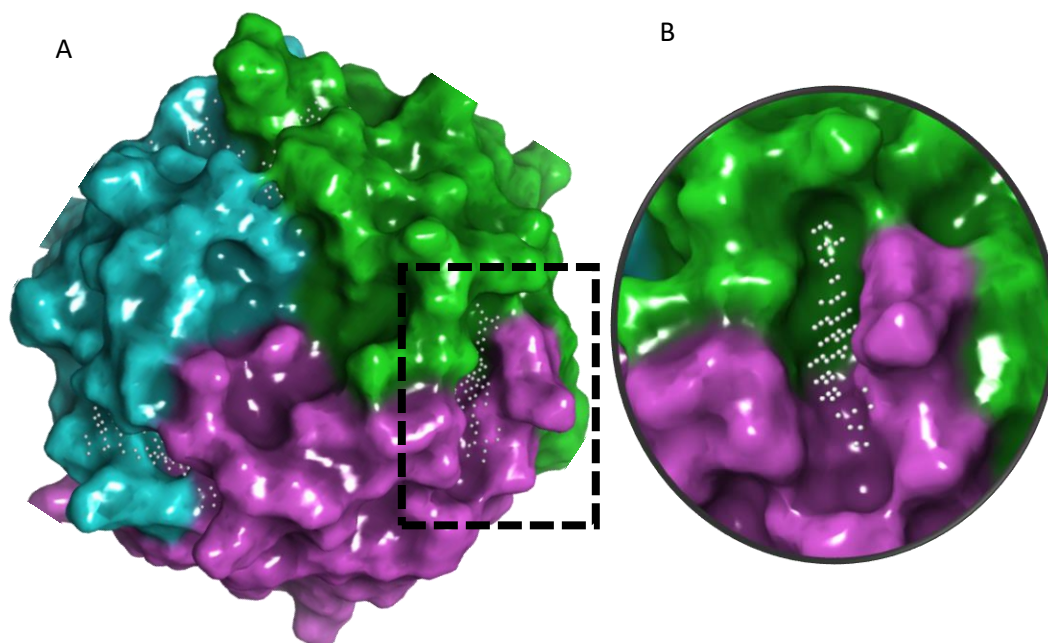


Figure 9.1 (A) Crystal structure BamBL (PDB 3ZW0) with druggable regions (site points) predicted using SiteMap. (B) Enlarged view of one of the binding sites.

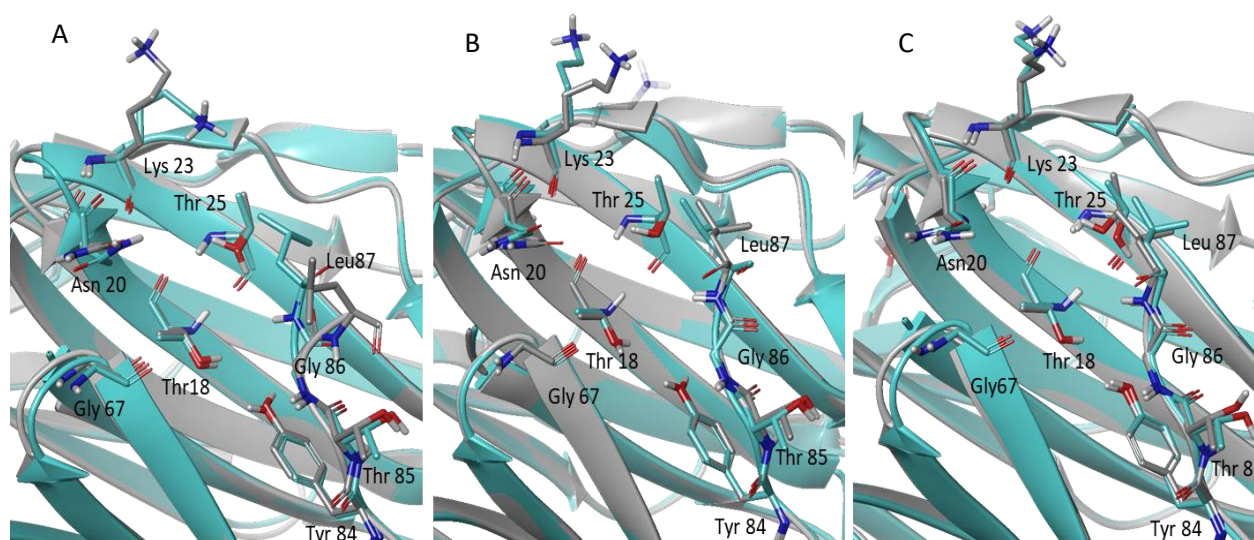


Figure 9.2 superimposition of the predicted three binding sites at the interface of chains A, B (A) A, C (B) and B, C (C) in the crystal structures of BamBL (PDB 3ZW0, grey and 3ZVV, cyan). Lys23 (in all the sites) and Leu87 (in one site) show significant difference in the side chain orientation which changes the shape and size of the predicted site.

Lys23 (in three sites) and Leu87 (in one site) illustrate significant difference in side chain orientation which slightly changes the shape and the size of the predicted site. Nonetheless, these sites were top ranked by the SiteMap tool for their propensities to bind a ligand.

Similar as before, computational pocket prediction algorithms using SiteMap tool were applied on the apo and holo forms of two lectins: RSL (from *Ralstonia solanacearum*) and AFL (from *Aspergillus fumigatus*). Interestingly, SiteMap calculations identified three secondary pockets in RSL trimer equivalent to the newly identified pockets in BamBL (**Figure 9.3**).

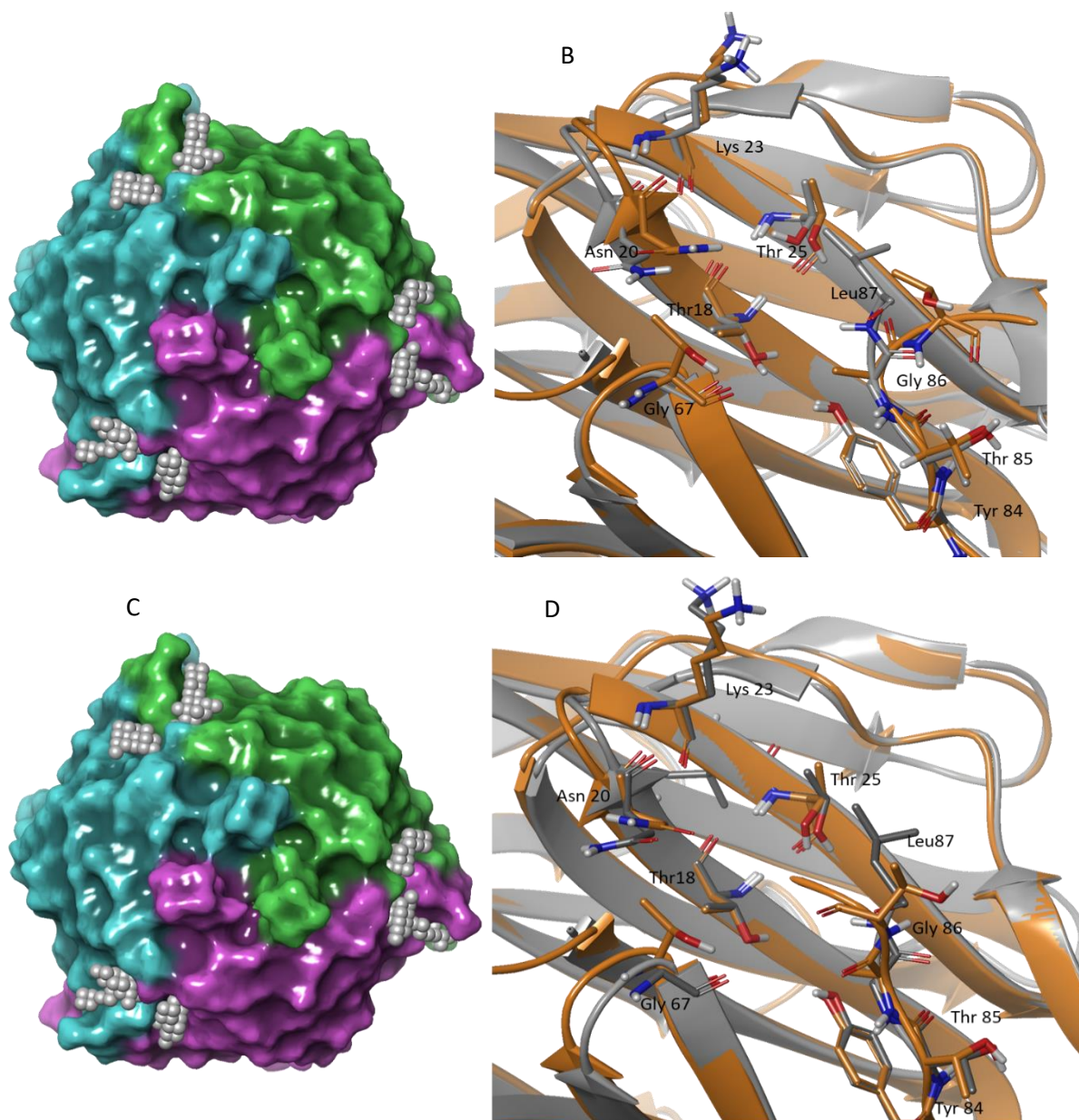


Figure 9.3 Computational analysis of RSL. Computational analysis of potential druggable binding sites in (A, B) apo and (C, D) holo forms of RSL. Three binding pockets comparable to BamBL were identified in PDB structures 3Z8I (apo) and 5AJB (holo) using SiteMap tool. Superimposition of one of the predicted binding sites at the interface of chains B, C of apo (B) and holo (D) forms of BamBL (grey) and RSL (orange) shows similarity in some of the binding site residues. However, terminal residues in the flexible loop region indicate significant differences.

However, shape and size of the predicted sites were slightly different due to differences in residues in the binding sites. In fungal lectin AFL, SiteMap identified a completely different druggable region which is structurally more distant from both bacterial lectins (**Figure 9.4**).

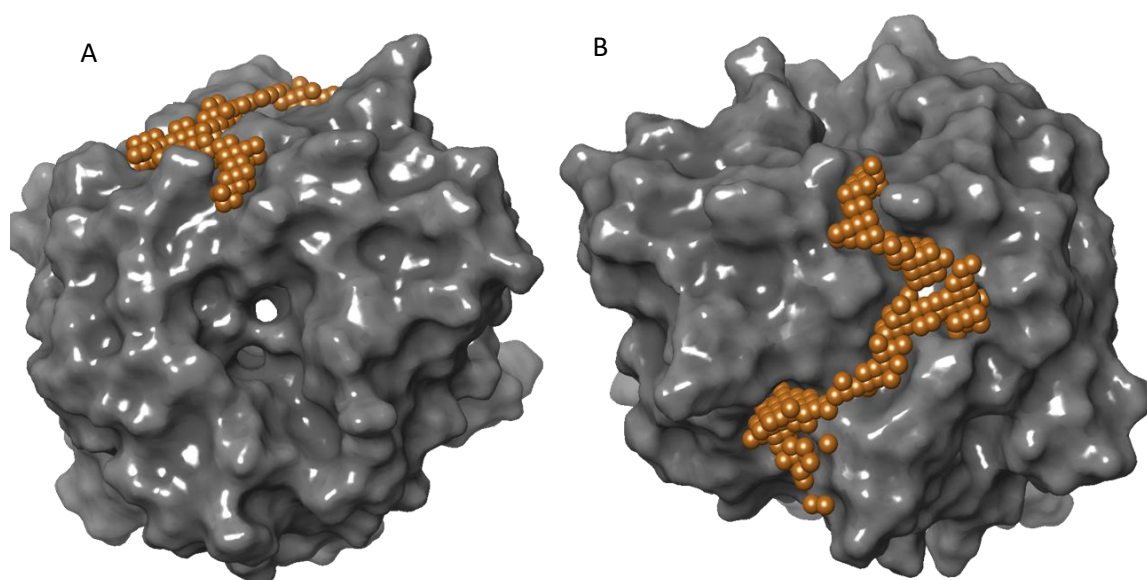


Figure 9.4 Computational analysis of potential druggable binding sites (dots) in AFL reveals one druggable site as shown in the (A) top and (B) side views.

8.2.2 Docking analysis

Docking calculations were done using XP and SP approaches to examine any difference in the ligand binding pose. The results from both the docking approaches show that the ligands bind at the predicted site and the residues Thr18, Lys23, Thr25, Gly67, Tyr84 and Leu87 play key role in the binding (**Figure 9.5 and 9.6**). In both the methods, Ligand 16A02 binds with an almost identical pose with a small difference due to a rotation of morpholine ring. Likewise, 15B05 and 14H04 generated similar binding poses with a slight difference in the orientation of fluorobenzene and benzoic acid, respectively. Results from docking studies indicate that the ligands possibly bind to the identified binding sites.

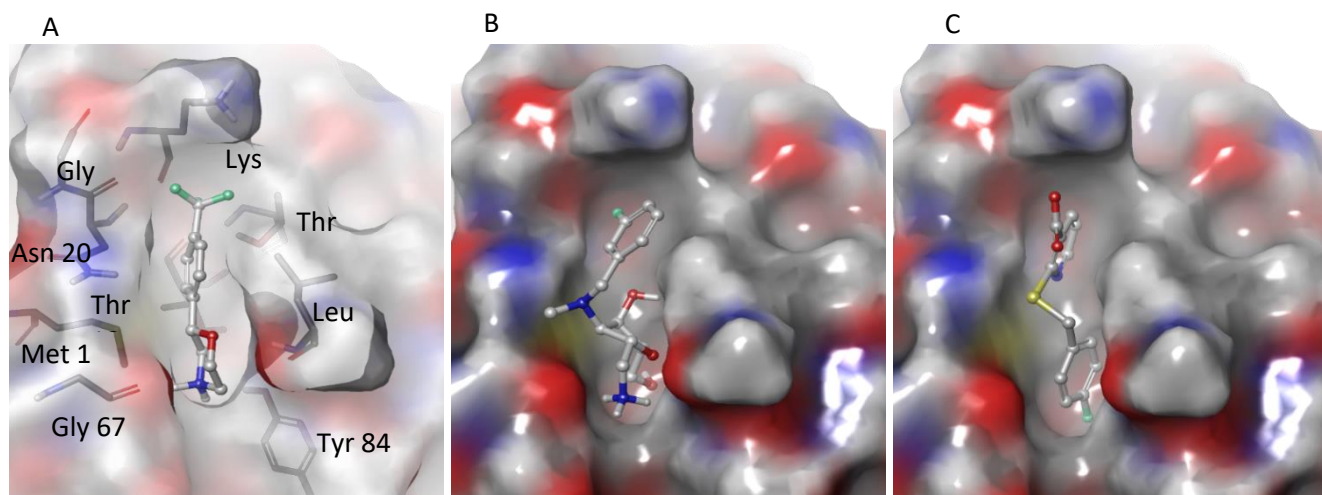


Figure 9.5 Binding pose of the ligands (A) 16A02, (B) 15B05 and (C) 14H04 predicted by docking (XP) studies. The key residues identified in the binding site are shown in the binding pose of 16A02.

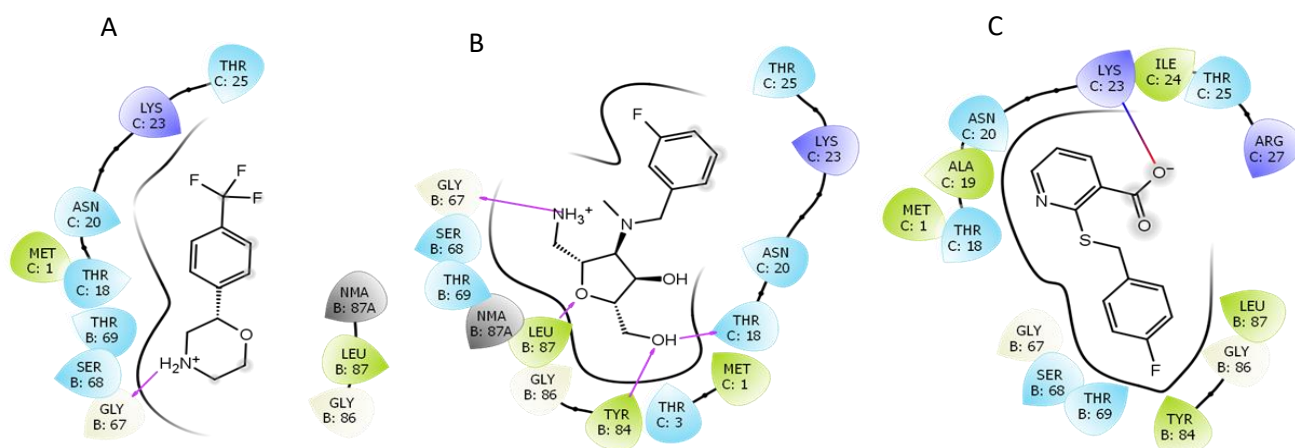


Figure 9.6 Key residues involved in the interaction with the molecules (A) 16A02, (B) 15B05 and (C) 14H04 predicted by XP docking.

8.3 Conclusions

The results indicate the presence of druggable sites in a bacterial lectin BamBL, which could be used to design allosteric inhibitors. The computational and experimental methods demonstrated the binding of drug-like molecules in the predicted binding site in BamBL. Further studies using PrOF NMR demonstrated that fragment binding to the secondary site induced conformational changes in the carbohydrate-binding site in BamBL. This indicates a communication between two spatially distant binding sites in the lectin. Site-directed

mutagenesis within the predicted site and the carbohydrate binding pocket also demonstrated conformational changes in distal regions from the mutation sites. The results suggest the presence of allosteric site in BambL. Additional computational analysis and experimental studies on RSL and AFL lectins showed structural similarities and demonstrated hit rates comparable to BambL. The results suggest the presence of allosteric sites in other β -propeller lectins. These observations can support future studies that aim to develop drug-like allosteric inhibitors against bacterial and fungal lectins.

8.4 References

- [1] Audfray, A. *et al.*, Fucose-binding lectin from opportunistic pathogen *Burkholderia ambifaria* binds to both plant and human oligosaccharidic epitopes, *J. Biol. Chem.* **2012**, *287*, 4335-4347.
- [2] Halgren, T. A., Identifying and Characterizing Binding Sites and Assessing Druggability, *J. Chem. Inf. Model.* **2009**, *49*, 377-389.
- [3] Arntson, K. E. Pomerantz, W. C. K., Protein-Observed Fluorine NMR: A Bioorthogonal Approach for Small Molecule Discovery, *J. Med. Chem.* **2016**, *59*, 5158-5171.
- [4] Kostlánová, N. *et al.*, The Fucose-binding Lectin from *Ralstonia solanacearum*: a new type of propeller architecture formed by oligomerization and interacting with fucoside, fucosyllactose, and plant xyloglucan, *J. Biol. Chem.* **2005**, *280*, 27839-27849.
- [5] Wimmerova, M. *et al.*, Crystal structure of fungal lectin: six-bladed beta-propeller fold and novel fucose recognition mode for *Aleuria aurantia* lectin, *J. Biol. Chem.* **2003**, *278*, 27059-27067.
- [6] Fujihashi, M. *et al.*, Crystal structure of fucose-specific lectin from *Aleuria aurantia* binding ligands at three of its five sugar recognition sites, *Biochemistry* **2003**, *42*, 11093-11099.
- [7] Schrödinger Release 2018-1: Maestro, S., LLC, New York, NY, **2018**.
- [8] Berman, H. M. *et al.*, The Protein Data Bank, *Nucleic Acids Res* **2000**, *28*, 235-242.
- [9] Olsson, M. H. *et al.*, PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions, *J Chem Theory Comput* **2011**, *7*, 525-537.
- [10] Li, H. *et al.*, Very fast empirical prediction and rationalization of protein pKa values, *Proteins* **2005**, *61*, 704-721.
- [11] Bas, D. C. *et al.*, Very fast prediction and rationalization of pKa values for protein-ligand complexes, *Proteins* **2008**, *73*, 765-783.
- [12] Harder, E. *et al.*, OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins, *J Chem Theory Comput* **2016**, *12*, 281-296.

[13] Schrödinger Release 2018-1: LigPrep, S., LLC, New York, NY *Vol.* **2018**.

[14] Friesner, R. A. *et al.*, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, *J. Med. Chem.* **2004**, *47*, 1739-1749.

Prediction and Validation of a Druggable Site on Virulence Factor of Drug Resistant *Burkholderia cenocepacia***

Kanhaya Lal,^[a, b] Rafael Bermeo,^[a, b] Jonathan Cramer,^[c] Francesca Vasile,^[a] Beat Ernst,^[c] Anne Imberty,^{*,[b]} Anna Bernardi,^{*,[a]} Annabelle Varrot,^{*,[b]} and Laura Belvisi^{*,[a]}

Abstract: *Burkholderia cenocepacia* is an opportunistic Gram-negative bacterium that causes infections in patients suffering from chronic granulomatous diseases and cystic fibrosis. It displays significant morbidity and mortality due to extreme resistance to almost all clinically useful antibiotics. The bacterial lectin BC2L-C expressed in *B. cenocepacia* is an interesting drug target involved in bacterial adhesion and subsequent deadly infection to the host. We solved the first high resolution crystal structure of the apo form of the lectin N-terminal domain (BC2L-C-nt) and compared it with the ones complexed with carbohydrate ligands. Virtual screening of a small fragment library identified potential hits predicted

to bind in the vicinity of the fucose binding site. A series of biophysical techniques and X-ray crystallographic screening were employed to validate the interaction of the hits with the protein domain. The X-ray structure of BC2L-C-nt complexed with one of the identified active fragments confirmed the ability of the site computationally identified to host drug-like fragments. The fragment affinity could be determined by titration microcalorimetry. These structure-based strategies further provide an opportunity to elaborate the fragments into high affinity anti-adhesive glycomimetics, as therapeutic agents against *B. cenocepacia*.

Introduction

Antimicrobial resistance enables pathogens to resist to the effects of an antibiotic or drug that would usually kill them or limit their growth.^[1] The emergence and spread of multidrug-resistant bacteria have challenged the existing treatment regimen which has enormous implications for worldwide healthcare delivery and community health.^[1–2] *Burkholderia cenocepacia* is a Gram-negative bacterium belonging to a group of more than 20 species called *Burkholderia cepacia* complex (BCC).^[3] BCC species survive in natural sources including water, soil and vegetation. In Nature, BCC bacteria can have both beneficial and detrimental effects on plants^[4] but they are also

identified as opportunistic human pathogens. In particular, *B. cenocepacia* is responsible for deadly infections in patients with immunocompromised conditions like chronic granulomatous diseases^[5] and cystic fibrosis.^[6] The treatment of the infection is really challenging, as *B. cenocepacia* strains show extreme resistance to almost all clinically useful antibiotics^[7] and cause significant morbidity and mortality. *B. cenocepacia* produces a large number of virulence factors that play an important role in host cell infection.^[8] Among them, four soluble lectins (BC2L-A, -B, -C and -D) have been identified, displaying very high sequence similarity with the virulence factor LecB (PA-IIL) from *Pseudomonas aeruginosa*.^[9] LecB forms a tetramer with high affinity for fucose,^[10] while BC2L-A is a dimer with significant affinity for mannose and oligomannose-type N-glycans.^[9,11] Except BC2L-A, the other three *B. cenocepacia* lectins present additional N-terminal domains.^[11–12] For BC2L-C, the C-terminal domain (LecB like) specifically binds to mannose, while the N-terminal domain (BC2L-C-nt) has been structurally characterized as a novel fucose-binding domain with a trimeric TNF- α -like architecture.^[13] Thus, BC2L-C represents a novel type of superlectin with dual specificity for fucose and mannose in the N- and C-terminal domains, respectively.^[14] BC2L-C as a virulence factor binds to carbohydrates present on the epithelial cells of the host. BC2L-C-nt has higher affinity for fucosylated oligosaccharides and its complexes with H-type 1 and Globo H (H-type 3) oligosaccharides have been recently solved.^[15] The superlectin is proposed to be involved in adhesion and inflammation processes.^[14] Bacterial adhesion represents the first step of infection, it also enables bacteria to have access to nutrients and to better resist to immune factors, bacteriolytic enzymes and antibiotics.^[16] Therefore, preventing glycoconjugate-lectin interactions by anti-adhesive therapy can counteract the

[a] K. Lal, R. Bermeo, Prof. Dr. F. Vasile, Prof. Dr. A. Bernardi, Prof. Dr. L. Belvisi
 Università degli Studi di Milano, Dipartimento di Chimica
 via Golgi 19, I-20133, Milano (Italy)
 E-mail: anna.bernardi@unimi.it
 laura.belvisi@unimi.it

[b] K. Lal, R. Bermeo, Dr. A. Imberty, Dr. A. Varrot
 Université Grenoble Alpes, CNRS, CERMAV, 38000 Grenoble (France)
 E-mail: Anne.imberty@cermav.cnrs.fr
 annabelle.varrot@cermav.cnrs.fr

[c] Dr. J. Cramer, Prof. Dr. B. Ernst
 University of Basel, Department of Pharmaceutical Sciences
 Klingelbergstrasse 50, 4056 Basel (Switzerland)

** A previous version of this manuscript has been deposited on a preprint server (<https://doi.org/10.26434/chemrxiv.13006382.v1>).

Supporting information for this article is available on the WWW under <https://doi.org/10.1002/chem.202100252>

© 2021 The Authors. Chemistry - A European Journal published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

infection process at its initial stage.^[16–17] This inhibition can be achieved by means of carbohydrate-based synthetic molecules which can compete for the lectin working as antagonists and, thus reduce the level of infection.

Here we describe the development of a structure-based approach to the design of such antagonists. First, we solved the crystal structure of apo BC2L-C-nt and compared the protein surface and bound water molecules with the fucose-bound structure. The X-ray crystal structure in complex with methylseleno- α -L-fuco-pyranoside (MeSe- α -L-Fuc, PDB code 2WQ4) was then used for virtual screening of a small fragment library in the vicinity of the fucose-binding site. This procedure identified a region (region 'X') that was most likely to host potential hits. The results were analysed with the main objective of identifying suitable fragments that docked in region X and could be chemically connected to the fucose core to obtain high-affinity ligands. The interaction of the fragments with the protein domain was confirmed using a group of biophysical techniques including STD-NMR,^[18] ITC and X-ray crystallography performed on one fragment confirmed binding at the expected location and therefore the ability of site X to host drug-like fragments. This study provides the rational design tools to elaborate the selected fragments into high-affinity ligands.

Results and Discussion

Analysis of the binding site in crystal structures

Crystal structure of trimeric BC2L-C-nt complexed with H-type 1 and Globo-H oligosaccharides are available^[14–15] revealing three sugar binding sites located at the interface between neighbouring chains (A, B, C), and separated by a distance of ~ 20 Å (Figure 1A). In each fucose binding site (Figure 1B), the key residues Tyr48, Ser82, Thr83, Arg85 from one chain (e.g. chain A) and Tyr58, Thr74, Tyr75, Arg111 from the neighbouring chain (e.g. chain C) play an important role in ligand binding. In addition, two water molecules bridge the sugar and the protein. Both water molecules are conserved in the available X-ray structures of BC2L-C-nt in complex with fucoside and fucosylated oligosaccharides.^[14–15] One is deeply buried in the binding site and sandwiched between the protein and the ligand, forming an H-bonding interaction with the HO-3 of fucose (Figure 1B). The second water molecule is more exposed to the solvent and mediates an H-bonding interaction between HO-2 of fucose and the side chain of Tyr58.

The crystal structures of complexes evidenced some promising pockets on the protein surface near the fucose binding site. The occurrence of such pockets in the apo-protein needed to be verified and therefore, we solved the crystal structure of the apo form of BC2L-C-nt at high resolution (1.5 Å). The asymmetric unit in the $P6_3$ space group contains one monomer and crystal symmetry was applied to build the trimer for comparison with other structures. Root-mean-squares values of 0.21 Å and 0.24 Å were obtained when comparing with the trimer complexed with MeSe- α -L-Fuc and Globo-H, respectively.

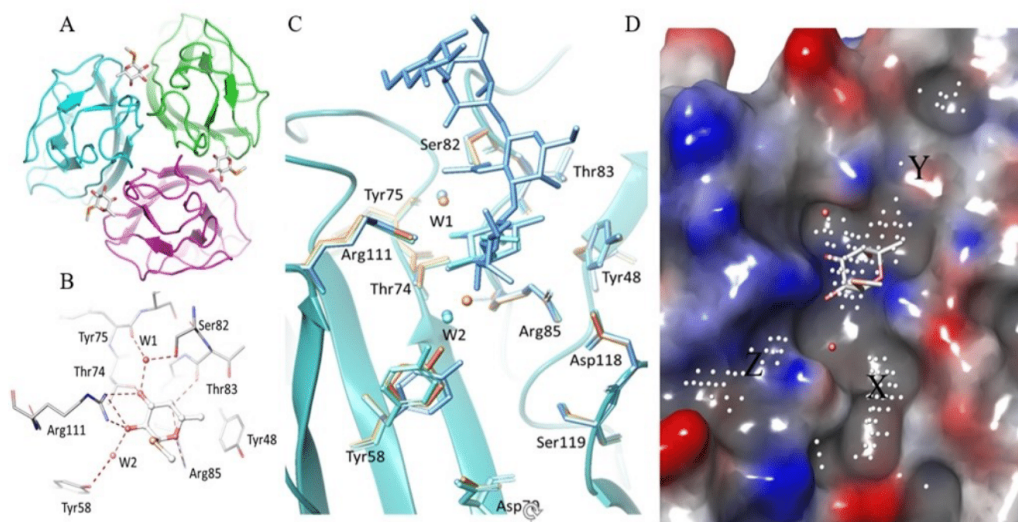


Figure 1. A) Crystal structure of BC2L-C N-terminal domain (PDB 2WQ4) showing three identical fucoside binding sites at the interface of monomers B) Fucoside binding site with MeSe- α -L-Fuc. Hydrogen bonds are represented as dashed lines. C) Superimposition of binding sites in the apo (orange, PDB 7BFY) and the holo forms (Cyan PDB 2WQ4, azure PDB 6TIG) of BC2L-C N-terminal domain. D) Identification of additional regions (site points) near fucose binding site suitable for fragment binding.

Comparison of binding sites (Figure 1C) did not display significant differences in the amino acid of fucose binding pockets and environment. Only a minor difference is observed at the surface loop (Asn52-Phe54), that is involved in the interaction with methyl group of N-acetylgalactosamine (GalNAc) in the complex with Globo-H (Figure S1). Likewise, small differences in the conformation of the N- and C-terminal residues were noticed due to H-bond interaction between them. The changes in the conformation of the termini further caused a small displacement (0.6 to 1.0 Å) of surface loops (Val28-Asp35, Asp95-Val100). Analysis of water molecules involved in bridging fucose to protein indicated that the more buried one (W1) is conserved in all structures, while the more exposed one (W2) moves by 1.9 Å in the apo structure.

The new crystal structure therefore confirms that the region surrounding the fucose binding site is of interest for drug design. This surface was analysed for druggability (ligandability)^[19] using the SiteMap^[20] tool. SiteMap creates a grid of points based on the depth, size, van der Waals interaction energy, hydrophilicity and hydrophobicity to determine the druggability of a protein region. Based on these characteristics, a single scoring function called SiteScore is assigned to potential druggable regions. For BC2L-C-nt the calculations identified three regions, which we labelled X and Y and Z (Figure 1D) in the vicinity of the fucose. Region Y, consisting of residues Ser82, Thr83 and Phe54 in each monomer, corresponds to the area where larger, fucosylated oligosaccharides were observed to bind, including the recently described Globo H hexasaccharide and H-type 1 tetrasaccharide.^[15] Of the two other regions (X and Z), site X is a deep crevice extending along the binding interface. The site Z consists of the region between Val110 and Arg111. All the sites are worth exploring further, thus the docking protocol was built to include them in the analysis.

Identification of top-ranked fragments

Docking analysis

2000 molecular fragments were retrieved from the Maybridge library of small fragments (rule of 3 diversity set available at <https://www.maybridge.com/>). In the first docking model, all the fragments were docked in the presence of the two conserved water molecules and the MeSe- α -L-Fuc. In the second docking model, only the buried water molecule and the ligand were retained since the second water molecule is close to region X and rather exposed to the solvent. This second model allowed us to examine the fragments that might be able to replace it. Fragments were found to dock mainly in regions X and Y. The region Y forms a very shallow and exposed binding site, which mostly hosted lipophilic fragments on the surface. Similarly, region Z also hosted a few hydrophilic fragments on the shallow surface. Region X is comparatively deeper and fragments appear to be nestling in it, generating some specific interactions. Therefore, we focused our further efforts on this region.

Binding analysis of top 200 fragments was done for 6 docking runs with XP, SP and HTVS protocols^[21] and involving either one water or two water molecules. HTVS and SP use the same scoring function but the HTVS protocol reduces the number of intermediate conformations, torsional refinement and sampling. The XP protocol employs a different, more complex scoring function with greater requirements for ligand-receptor shape complementarity. This screens out false positives that SP or HTVS may let through. From each model, the best fragments with consensus scoring (ranked within top 200 fragments) obtained by XP, SP and HTVS were selected for analysis of key residues involved in ligand binding. The docking results with the two waters model showed that the number of hits obtained at site X using SP and HTVS methods were almost same, while the hits obtained using XP were reduced to half. In the one water model, the number of hits at site X increased almost by a factor of two, due to the omitted water molecule near site X. The interaction pattern identified using three scoring functions at site X indicated that the fragments including a benzylamine moiety have good binding affinity. The key residues involved in binding are Tyr58, and Asp70 whilst Asp118 from neighbouring protomer can also be recognized by some of the top scoring fragments that form a salt bridge interaction with it (Figure 2).

The main interactions observed for the majority of the top ranked fragments are a salt bridge between Asp70 side chain and the benzylamino group of the fragments and π - π stacking interactions with Tyr58. A total of 94 and 89 fragments for site X were identified for one and two waters models, respectively, as top ranked fragments according to XP and SP/HTVS or all the three scoring functions.

Selection of best fragments

The fragments were carefully analysed based on different parameters such as structural diversity, possibility to connect them to the fucose core, size and distance from the fucose core.

Small fragments which were found significantly far (> 6 Å) from the fucose core and docked on the shallow surface surrounding site X were discarded. The remaining 32 fragments for site X were redocked to analyse the stability of the ligand interactions in multiple binding poses (10 poses). Other factors like commercial availability, synthetic feasibility and purchasing cost allowed to select 12 fragments (Figure 2 and Table S1) for experimental validation. Within this group, fragments KL1-8 were among the top scorer in the two waters model, while fragments KL9-12 were predicted to bind in the one water model.

Experimental validation of fragment binding

For each fragment, a 2.5 mM solution was used to test the interaction with BC2L-C-nt using thermal shift assay (TSA, ThermoFluor).^[22] Methyl α -L-fucoside (Me- α -L-Fuc) was used as a reference in the experiment to observe fucose binding and

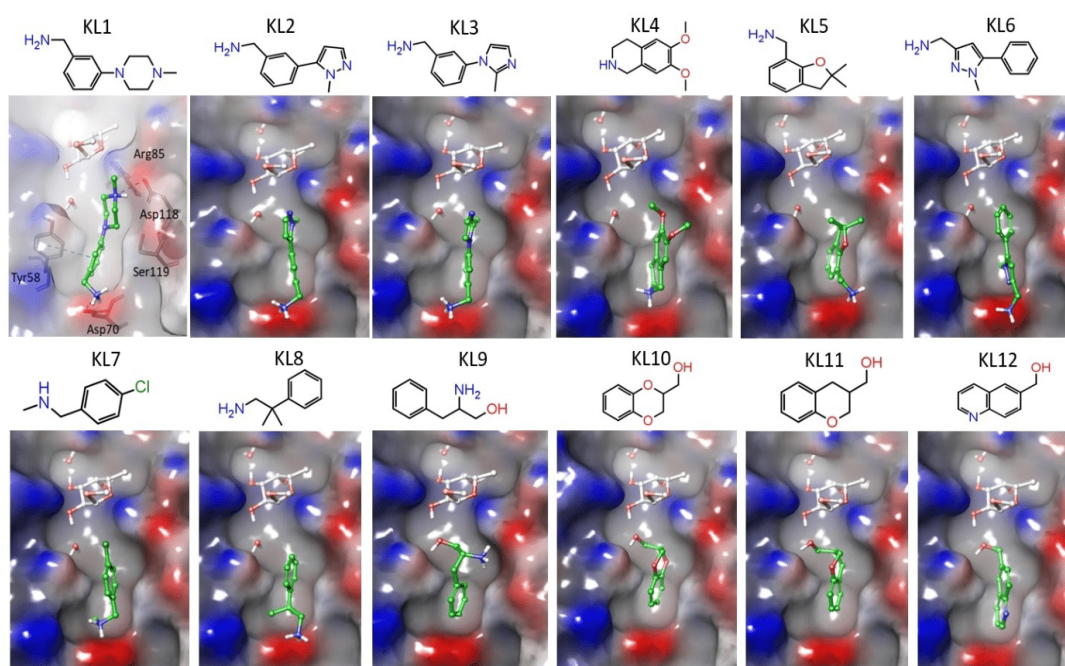


Figure 2. Binding pose for the top ranked fragments (KL1-KL12) predicted by docking studies at site X. The key residues identified in the binding site are shown in the docking pose of KL1.

hence validate the protocol. Then, the fragments were tested in the presence of Me- α -L-Fuc (20 mM). The results show the expected positive shift ($\sim 2^\circ\text{C}$) upon Me- α -L-Fuc binding (Figure S2) while all of the complexes with fragments exhibit a small negative shift between 0.15 to 1.65 $^\circ\text{C}$ (Figure S3) in the melting temperature (T_m), which possibly suggests that the fragments destabilize the binding interface and bind to a native or partially unfolded state of the protein.^[23]

We repeated the experiment for all the fragments in the absence of Me- α -L-Fuc and the results show similar behaviour with a smaller negative shift in the melting temperature (Figure S4). The experimental results of TSA do not afford any structural information concerning the interaction. Therefore, we performed another screening using STD-NMR and X-ray crystallography.

STD-NMR analysis of fragment binding

Saturation transfer difference (STD) NMR has become a leading technique to characterize fragment–macromolecule interaction in solution, because it is sensitive to weak binding events (dissociation constant in a low μM to mM range).^[18,24] In general, STD experiments are performed by irradiating the methyl group of valine, leucine, or isoleucine residues (between 1 and -1 ppm), that are often present in the binding site of proteins.^[18]

The irradiation frequency of STD can also be varied in order to investigate whether the fragment has a preferred interaction with aliphatic or aromatic amino acids of the protein.^[25]

STD-NMR was used to analyse the interaction of BC2L-C-nt with fragments KL3, KL8 and KL9 in the presence of Me- α -L-Fuc, irradiating at -0.05 ppm. Me- α -L-Fuc was initially tested alone in the experiment, verifying that it binds BC2L-C-nt, with a strong involvement of the methyl group (Figure S5). Then, fragment KL3, KL8 (among the top scorers in the two waters docking model) and fragment KL9 (predicted to bind by the one water model) were analysed in the presence of the protein and of 2 mM Me- α -L-Fuc. The sample was prepared at 1:1 ratio between sugar ligand and fragment. The resulting spectra for fragment KL9 and KL3 are shown in Figure 3 and Figure 4B, respectively. The spectra of fragment KL8 are reported in the supplementary information (Figure S6). In all cases, simultaneous interaction of the fragment and Me- α -L-Fuc with BC2L-C-nt was observed, confirming the binding event for the three fragments in the BC2L-C-nt/fucose complex. In the STD spectra, the signals of Me- α -L-Fuc and of the fragment appear with comparable intensities, indicating a similar affinity for sugar and fragment.

STD spectra were also acquired using 10 ppm as irradiation frequency. In this case, the aromatic protons of the fragments are observable, while no signals of Me- α -L-Fuc can be detected (Figures 4C and Figure S7). This finding suggests that the fragments bind in the proximity of aromatic residues of the protein and thus supports the docking prediction that they are

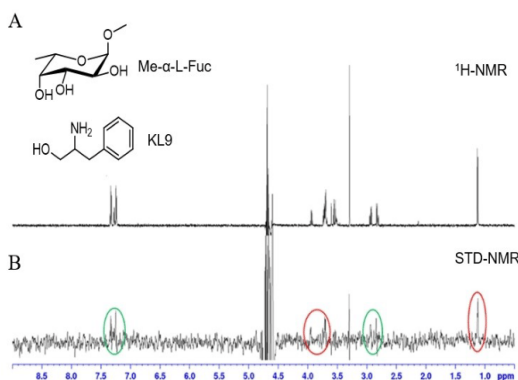


Figure 3. A) $^1\text{H-NMR}$ and B) STD spectrum of fragment KL9 and Me- α -L-Fuc in the presence of BC2L-C-nt (1000:1) recorded with a Bruker Avance 600 MHz spectrometer. The spectrum is recorded at 298 K with irradiation frequency at -0.05 ppm. In the STD spectrum, the signals at 3.7 ppm and 1.1 ppm, produced respectively by the fucose ring and by its methyl group, are highlighted with red circles. The signals of the fragment are highlighted with a green circle (at 2.9 ppm for $-\text{CH}_2\text{-Ph}$ and 7.3 ppm for the aromatic protons).

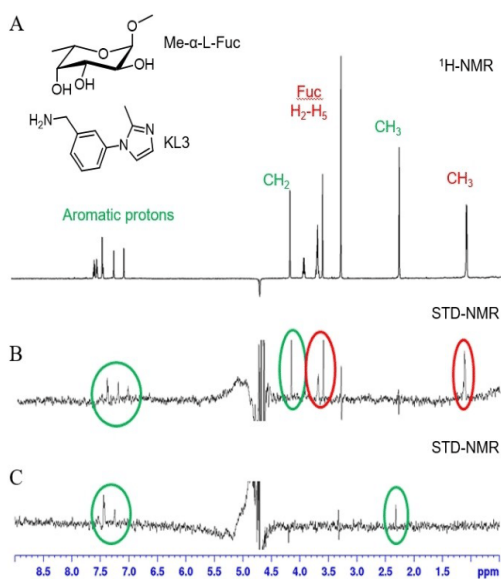


Figure 4. A) $^1\text{H-NMR}$ spectrum, B) STD spectrum (irradiation frequency -0.05 ppm) and C) STD spectrum (irradiation frequency 10 ppm) of fragment KL3 and Me- α -L-Fuc in the presence of BC2L-C-nt (1000:1) recorded with a Bruker Avance 600 MHz spectrometer at 298 K. In the STD spectrum at -0.05 ppm (B), the signals at 3.7 ppm and 1.1 ppm, produced respectively by the fucose ring and by its methyl group, are highlighted with red circles. The signals of the fragment are highlighted with a green circle (at 4.2 ppm for $-\text{CH}_2-$ and in the range 7.05–7.4 ppm for aromatic protons). The STD spectrum at 10 ppm (C) shows the aromatic protons (in the range 7.1–7.4 ppm) and the methyl group (at 2.3 ppm) of the fragment

located in a protein binding pocket that includes an aromatic residue (Tyr58). It is interesting to note that the STD spectrum

of KL3 obtained irradiating at 10 ppm (Figure 4C) also shows a clear signal for the methyl group of the fragment (at 2.3 ppm), which is not visible when irradiating at -0.05 ppm (Figure 4B). This suggests that also this moiety is proximal to an aromatic side chain of the protein. On the contrary, the singlet at 4.2 ppm, corresponding to the methyleneamino benzylic protons of the fragment, which is clearly visible when irradiating at -0.05 ppm (Figure 4B), disappears from the spectrum, like the fucose protons, when irradiating at 10 ppm. Thus, this moiety is expected to be surrounded by aliphatic protons of the protein.

KL3-BC2L-C-nt crystal structure analysis

All fragments (KL1–KL12) were soluble enough to be used for soaking experiment with crystals of BC2L-C-nt complexed with Globo H hexasaccharide obtained as described previously.^[15]

After soaking, crystals containing KL10, KL11 and KL12 did not diffract at sufficient resolution for data collection. Crystal soaked with the remaining fragments (KL1–KL9) diffracted at a resolution close to 2 Å or better, but examination of the electron density after molecular replacement only revealed electron density for the sugar and not for the fragment indicating that they did not bind to the protein in the experimental conditions used. Only in the complex with KL3 (3-(2-Methyl-1H-imidazol-1-yl) benzylamine) at 1.9 Å resolution, electron density corresponding to the expected fragment could be seen in site X located at the interface between two monomers. The orientation of the fragment, and the observed interactions correspond very well with those predicted by the docking studies (Figure 5 and Figure S8). Residue Tyr58 forms T-shaped π - π stacking interactions with the benzene ring and Asp70 forms a salt bridge with the amino group in the fragment. The free nitrogen of the imidazole ring makes water mediated interaction with the side chain of Arg85 and the OH-4 of the GlcNAc moiety of Globo H (Figure 5B). The fragment binds with identical pose and reproduce the same binding interactions in the three binding sites of the trimer (Table 1).

The position of the fragment in site X is fully consistent with the STD-NMR data (Figure 4), which indicate proximity of the benzylic methyleneamino group of KL3 to aliphatic residues of the protein (Asp70 and Ser119 in the X-ray structure). The

Table 1. Summary of the interactions of BC2L-C-nt with KL3 in three binding sites.

Ligand atom	Protein or water atom	Distance [Å]
N3	Asp70 (OD2)	3.20
	W3 (HOH161) ^[a]	2.75 ± 0.15
N1	Arg85 (NH2)	3.30 ± 0.07
	W4 (HOH108)	2.46 ± 0.05
C ^[b]	Ser119 (CB)	3.60 ± 0.04
	Tyr58 (CE1)	3.50 ± 0.07

[a] only present in two binding sites; [b] For hydrophobic contacts and π - π interactions, the distance is calculated from the nearest atoms in the ligand and the protein. Mean distance and standard deviation were calculated from the distance of ligand and protein atoms in each binding site.

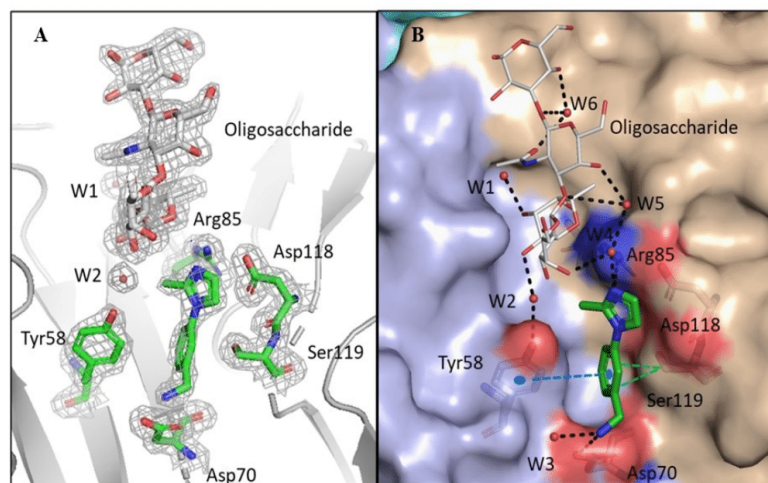


Figure 5. Crystal structure of BC2L-C-nt with Globo H and KL3. A) Zoom in the binding site with 2Fo-DFc electron density represented at 1 σ . B) Network of interaction in the binding site. Analysis of the complex shows that the key interactions and residues predicted from docking studies were involved in the ligand (KL3) binding. The salt bridge between Asp70 side chain and benzylamino group and π - π stacking interactions with Tyr58 are maintained in the crystallized complex. In addition to the water molecules from two waters model, a new network of water molecules involved in key interactions between the ligand and the protein is also highlighted. H-bonding interactions and hydrophobic interactions are displayed in black and green dashed lines respectively. π - π stacking interactions are shown in blue dashed lines.

methyl group of KL3, which in the STD spectra responds to irradiation at 10 ppm, is in fact close to the Tyr58 side chain in the X-ray structure. The water mediated interactions with Globo H were identical to the previous complex.^[15] The results of X-ray crystallographic screening validated the docking results and the ligandability of site X.

Affinity analysis and activity validation

The affinity of BC2L-C-nt for KL3 was determined by isothermal titration calorimetry (ITC) measurements.^[26] Titration of the lectin by KL3 resulted in small exothermic peaks after correction for buffer mismatch (Figure S9). The integrated curve could be fitted with one-site model with stoichiometry of one, resulting in the determination of a K_d of 877 μ M. Because of the low c -value of the experiment, the thermodynamic contributions cannot be safely estimated.

Conclusions

Crystal structure analysis of the apo and the holo form of BC2L-C-nt demonstrated the presence of a druggable (ligandable) region (site X) in the vicinity of the fucoside binding site. The computational and experimental screenings identified fragments interacting with BC2L-C-nt. The study indicates that the fragments bind in a newly identified binding region in BC2L-C-nt when the fucoside binding site is occupied. Different biophysical techniques including TSA and STD-NMR spectroscopy, confirmed fragment-protein interaction. Remarkably, the

binding mode of one fragment (KL3) could be validated by X-ray crystallography at high resolution, further confirming the ability of site X to host drug-like fragments. The affinity measured by ITC is sub-millimolar, which is very promising for such small fragment. The complementary structural and thermodynamic data give clear view of the relative importance of apolar and polar interactions for fragment KL3. This could be used in the future for structure-based optimization of this first hit.

Most interestingly, this study provides an opportunity to connect the best fragments to the fucose core to obtain high affinity glycomimetic ligands. The selection of suitable linkers can be done based on the distance (measured 4.8 Å) between the nearest atoms of fragment and the fucose core. Other factors like synthetic feasibility and possibility to maintain the binding pose at the site X can be considered to identify suitable linkers. A robust synthetic route to glycomimetics comprising fucose linked fragments will help in designing high affinity ligands as anti-adhesive agents against *B. cenocepacia*.

Experimental Section

Protein expression and purification: Protein production and purification of the BC2L-C-nt was performed as described previously.^[15] An average yield 5.2 mgL⁻¹ of culture medium was obtained and stored at 4 °C.

Preparation of protein model: All the calculations were performed using the Schrödinger Suite through Maestro (version 2018-1) graphical interface.^[27] Atomic coordinates from the crystal structure of BC2L-C-nt complexed with MeSe- α -L-Fuc (PDB code 2WQ4) were taken from the Protein Data Bank.^[28] The asymmetric unit contains

three peptide chains and three carbohydrate ligands (MeSe- α -L-Fuc), around a 3-fold pseudo axis of symmetry. The mode of binding for the sugar is identical in the three binding sites, therefore only one binding site located between chains A and C was used for the calculations. The two structural water molecules HOH2195 (W1) and HOH2194 (W2) bridging fucose and protein were also retained. The hydrogen atoms were added and pK_a was predicted for protein residues using the PROPKA^[29] method at pH 7.4 and assigned HIE protonation state to the histidine (His116) residue. Finally, protein-ligand complex was subjected to restrained minimization with convergence of heavy atoms to an RMSD of 0.3 Å using the OPLS3 force field.^[30]

Preparation of ligand models: The Maybridge library of small fragments (rule of 3 diversity set) containing 2000 fragments was used for *in silico* screening. The LigPrep^[31] tool was used to generate tautomers, stereoisomers and protonation states at pH 7 \pm 2. The calculation generated 2904 structures.

Models for docking study: For docking grid generation, the centroids of residues from chain A (Tyr48, Ser82, Thr83, Arg85) and chain C (Tyr58, Thr74, Tyr75, Arg111) were selected to define a cubic grid box of 32 \times 32 \times 32 Å. The ligand (MeSe- α -L-Fuc) and the water molecules (HOH2194 and HOH2195) were retained. The same residues were used to generate the second grid with one water (HOH 2195) molecule. Both the grids (models) were used for docking studies using XP, SP and HTVS scoring functions. All the calculations were accomplished by Glide (version 7.8)^[21] using the flexible docking approach.

Thermal shift assay (TSA): The fragments KL1, KL2, KL3, KL5, KL6, KL7, KL9, KL10, KL11 (Table S1) were purchased from the Maybridge (Fisher Scientific International) and the other fragments; KL4, KL8 and KL12 were purchased from the abcr GmbH. The fragments were tested for the purity using liquid chromatography-mass spectrometry (LC–MS).

For the dye-based TSA, BC2L-C-nt (5 μ M) in assay buffer (20 mM Tris HCl, 100 mM NaCl, pH 8.0) was incubated with 50x SYPRO orange and 2.5 mM KL1-12 in the presence or absence of 20 mM Me- α -L-Fuc. A Qiagen Rotor-Gene Q instrument was used to apply a heat ramp of 1 °C/min from 25–95 °C and SYPRO orange fluorescence at 620 nm was monitored using the appropriate optical channel.

STD-NMR interaction studies: The interaction between ligands and isolated protein was investigated using STD-NMR experiments. The spectra were acquired with a Bruker Avance 600 MHz instrument at 298 K, in a 3 mm NMR tube and in the phosphate buffer previously described (200 μ L). All protein–ligand samples were prepared in a 100:1 and 1000:1 ligand/protein ratio in concentration. In STD experiments water suppression was achieved by using the WATERGATE 3-9-19 pulse sequence. The on-resonance irradiation of the protein was kept at –0.05 ppm and 10 ppm. Off-resonance irradiation was applied at 200 ppm, where no protein signals were visible. Selective presaturation of the protein was achieved by a train of Gauss shaped pulses of 49 ms length each. The total length of the saturation train depends on the L7 parameter (the loop counter). STD experiments were acquired with L7=60 leading 2.94 s of total saturation. Two protocols for sample preparation were followed in all cases: either by adding the fragment to a pre-incubated solution of protein and Me- α -L-Fuc, or by adding the fucoside to a pre-incubated solution of protein and fragment. The resulting STD spectra were very similar independent of the set up. So, the results reported here correspond to the experiments obtained by adding the fragments to a solution of protein and Me- α -L-Fuc.

X-ray crystallography, data collection, and structure determination: The apo form of BC2L-C-nt was crystallized using the vapour diffusion method and 2 μ L hanging drops containing a 50:50 (v/v) mix of protein (5.5 mg/ml) and reservoir (sodium citrate 1.2 M at pH 7.0). Cubic crystals were obtained from the solution after 3 weeks. For the soaking experiments, crystals of BC2L-C-nt in complex with Globo H oligosaccharide were obtained as described previously.^[15] The fragments were tested for the aqueous solubility at higher concentration and a stock solution was prepared. The crystals were soaked overnight in the 0.5 μ L volume of fragments (from stock) in 4.5 μ L of 2.5 M sodium malonate used for cryoprotection that makes a final concentration of 2 mM, for the fragments KL1, KL7 and KL11, 2.5 mM for KL12, 5 mM for KL2, KL5, KL6, KL8 and KL10 and 10 mM for KL3, KL4, KL9. For KL2 and KL12, 10 percent DMSO was added to achieve the above concentration. The crystals were flash-cooled in liquid nitrogen prior to data collection. The data was collected on the beamline Proxima 1, synchrotron SOLEIL, Saint Aubin, France, using an Eiger 16 m detector (Dectris, Baden, Switzerland). The data was processed using XDS and XDSME.^[32] The CCP4 suite was used for all further processing.^[33] The coordinates of the monomer A of PDB code 2WQ4 were used as search model to solve the structures of the apo form and the complexes with BC2L-C-nt by molecular replacement using PHASER.^[34] Refinement was performed using restrained maximum likelihood refinement and REFMAC 5.8^[35] interspaced with using manual rebuilding in Coot.^[36] for cross validation, 5% of the data were set aside. Riding atoms were added during refinement (Table S2). Library for the fragment was made using ligand builder in Coot. All carbohydrates were validated using Privateer in CCP4i2 prior validation using the PDB validation server and deposition to the Protein Data Bank under code 7BFY for the apo form and 6ZZW for the complex.

ITC measurements: The ITC experiments were performed at 25 °C with an ITC200 isothermal titration calorimeter (Microcal-Malvern Panalytical, Orsay, France). The protein (BC2L-C-nt) and ligand (KL3) were dissolved in the same buffer composed of 100 mM Tris HCl pH 7.0 and 100 mM NaCl. A total of 38 injections of 1 μ L of ligand solution (15 mM) were added at intervals of 200 s while stirring at 850 rpm was maintained to ensure proper mixing in the 200 μ L sample cell containing the protein, at 225 μ M. A control experiment was performed by injecting same concentration of KL3 in buffer. The differences of integrated peaks were performed using the Microcal PEAQ-ITC analysis software. The binding thermodynamics was further processed with a “one set of sites” fitting model. The experiment determined experiment affinity (K_d), binding enthalpy (ΔH) while the stoichiometry was fixed to 1. Free energy change (ΔG) and entropy contributions ($T\Delta S$) were derived from the equation $\Delta G = \Delta H - T\Delta S$. The experiments were performed in duplicates and the standard deviation was in 20% range for K_d .

Acknowledgments

This research was funded from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 765581. The authors acknowledge support by the ANR PIA Glyco@Alps (ANR-15-IDEX-02) and Labex Arcane-CBH-EUR-GS (ANR-17-EURE-0003) and by Università degli Studi di Milano (open access agreement CARE-CRUI). The authors are grateful to SOLEIL Synchrotron, Saint Aubin, France for provision of synchrotron radiation facilities and access to the beamline Proxima 1. The STD-NMR

experiments were performed using the Unitech COSPECT platform at the University of Milan.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: Antimicrobial resistance · BC2L-C · Glycomimetics · Ligand design · Virtual screening

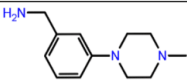
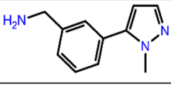
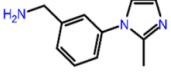
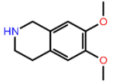
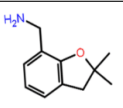
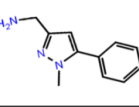
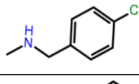
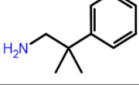
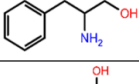
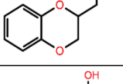
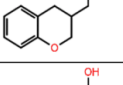
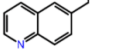
- [1] *Antimicrobial Resistance: Global Report on Surveillance*. Geneva: WHO 2014, pp. 1–252.
- [2] R. Smith, J. Coast, *BMJ* 2013, 346, f1493.
- [3] E. Mahenthiralingam, A. Baldwin, C. G. Dowson, *J. Appl. Microbiol.* 2008, 104, 1539–1551.
- [4] E. Mahenthiralingam, T. A. Urban, J. B. Goldberg, *Nat. Rev. Microbiol.* 2005, 3, 144–156.
- [5] J. A. Winkelstein, M. C. Marino, R. B. Johnston, Jr., J. Boyle, J. Curnutte, J. I. Gallin, H. L. Malech, S. M. Holland, H. Ochs, P. Quie, R. H. Buckley, C. B. Foster, S. J. Chanock, H. Dickler, *Medicine* 2000, 79, 155–169.
- [6] S. L. Butler, C. J. Doherty, J. E. Hughes, J. W. Nelson, J. R. Govan, *J. Clin. Microbiol.* 1995, 33, 1001–1004.
- [7] S. Nzula, P. Vandamme, J. R. Govan, *J. Antimicrob. Chemother.* 2002, 50, 265–269.
- [8] a) S. A. Loutet, M. A. Valvano, *Infect. Immun.* 2010, 78, 4088–4100; b) M. S. Saldias, M. A. Valvano, *Microbiology* 2009, 155, 2809–2817.
- [9] A. Imberty, M. Wimmerova, E. P. Mitchell, N. Gilboa-Garber, *Microbes Infect.* 2004, 6, 221–228.
- [10] E. P. Mitchell, C. Sabin, L. Snajdrova, M. Pokorna, S. Perret, C. Gautier, C. Hofr, N. Gilboa-Garber, J. Koca, M. Wimmerova, A. Imberty, *Proteins* 2005, 58, 735–746.
- [11] E. Lameignere, L. Malinowska, M. Slavikova, E. Duchaud, E. P. Mitchell, A. Varrot, O. Sedo, A. Imberty, M. Wimmerova, *Biochem. J.* 2008, 411, 307–318.
- [12] E. Lameignere, T. C. Shiao, R. Roy, M. Wimmerova, F. Dubreuil, A. Varrot, A. Imberty, *Glycobiology* 2010, 20, 87–98.
- [13] O. Sulak, G. Cioci, M. Delia, M. Lahmann, A. Varrot, A. Imberty, M. Wimmerova, *Structure* 2010, 18, 59–72.
- [14] O. Sulak, G. Cioci, E. Lameignere, V. Balloy, A. Round, I. Gutsche, L. Malinowska, M. Chignard, P. Kosma, D. F. Aubert, C. L. Marolda, M. A. Valvano, M. Wimmerova, A. Imberty, *PLoS Pathog.* 2011, 7, e1002238.
- [15] R. Bermeo, A. Bernardi, A. Varrot, *Molecules* 2020, 25, 248.
- [16] I. Ofek, D. L. Hasty, N. Sharon, *FEMS Immunol. Med. Microbiol.* 2003, 38, 181–191.
- [17] a) B. Ernst, J. L. Magnani, *Nat. Rev. Drug Discovery* 2009, 8, 661–677; b) S. Sattin, A. Bernardi, *Trends Biotechnol.* 2016, 34, 483–495.
- [18] B. Meyer, T. Peters, *Angew. Chem. Int. Ed.* 2003, 42, 864; *Angew. Chem.* 2003, 115, 890–890.
- [19] S. Vukovic, D. J. Huggins, *Drug Discovery Today* 2018, 23, 1258–1266.
- [20] T. A. Halgren, *J. Chem. Inf. Model.* 2009, 49, 377–389.
- [21] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, P. S. Shenkin, *J. Med. Chem.* 2004, 47, 1739–1749.
- [22] M. W. Pantoliano, E. C. Petrella, J. D. Kwasnoski, V. S. Lobanov, J. Myslik, E. Graf, T. Carver, E. Asel, B. A. Springer, P. Lane, F. R. Salemme, *J. Biomol. Screening* 2001, 6, 429–440.
- [23] P. Cimpmperman, L. Baranauskienė, S. Jachimovičienė, J. Jachno, J. Torresan, V. Michailoviene, J. Matuliene, J. Sereikaite, V. Bumelis, D. Matulis, *Biophys. J.* 2008, 95, 3222–3231.
- [24] a) F. Vasile, F. Gubinelli, M. Panigada, E. Soprana, A. Siccardi, D. Potenza, *Glycobiology* 2018, 28, 42–49; b) F. Vasile, D. Rossi, S. Collina, D. Potenza, *Eur. J. Org. Chem.* 2014, 2, 5.
- [25] S. Monaco, L. E. Tailford, N. Juge, J. Angulo, *Angew. Chem. Int. Ed.* 2017, 56, 15289; *Angew. Chem.* 2017, 129, 15491–15293.
- [26] M. R. Duff, Jr., J. Grubbs, E. E. Howell, *J. Visualization* 2011.
- [27] Schrödinger Release 2018–1: Maestro, LLC, New York, NY, 2018.
- [28] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* 2000, 28, 235–242.
- [29] a) M. H. Olsson, C. R. Sondergaard, M. Rostkowski, J. H. Jensen, *J. Chem. Theory Comput.* 2011, 7, 525–537; b) H. Li, A. D. Robertson, J. H. Jensen, *Proteins* 2005, 61, 704–721; c) D. C. Bas, D. M. Rogers, J. H. Jensen, *Proteins* 2008, 73, 765–783.
- [30] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, J. W. Kaus, D. S. Cerutti, G. Krilov, W. L. Jorgensen, R. Abel, R. A. Friesner, *J. Chem. Theory Comput.* 2016, 12, 281–296.
- [31] Schrödinger Release 2018–1: LigPrep, LLC, New York, NY in Vol. 2018.
- [32] a) W. Kabsch, *Acta Crystallogr. D Biol. Crystallogr.* 2010, 66, 125–132; b) P. Legrand, *GitHub Repos* 2017, 2017.
- [33] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Pottertton, H. R. Powell, R. J. Read, A. Vagin, K. S. Wilson, *Acta Crystallogr. Sect. D* 2011, 67, 235–242.
- [34] A. J. McCoy, *Acta Crystallogr. Sect. D* 2007, 63, 32–41.
- [35] G. N. Murshudov, P. Skubak, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, A. A. Vagin, *Acta Crystallogr. Sect. C* 2011, 67, 355–367.
- [36] P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, *Acta Crystallogr. Sect. C* 2010, 66, 486–501.

Manuscript received: January 21, 2021

Accepted manuscript online: March 26, 2021

Version of record online: May 1, 2021

Table S1 Top ranked fragments identified for site X.

Fragment name	Structure	Molecular weight	Aqueous Solubility(mM)
KL1		205.3	20
KL2		187.2	50 ^[b]
KL3		187.2	100
KL4		193.2	100
KL5		177.2	1 ^[a]
KL6		187.2	1 ^[a]
KL7		155.6	20
KL8		149.2	50
KL9		151.2	100
KL10		166.2	1 ^[a]
KL11		164.2	20
KL12		159.2	25 ^[b]

[a] Solubility at higher concentration is not known

[b] dissolved in 10 percent DMSO

Table S2 X-ray data collection and processing of apo form of BC2L-C-nt and its complex with Globo H and fragment KL3.

Data set	BC2L-C-nt complex with KL3 and Globo H			BC2L-C-nt apo form
PDB code	6ZZW			7BFY
Data Collection				
Beamline	PROXIMA1 (SOLEIL)			PROXIMA1 (SOLEIL)
Wavelength (Å)	0.9786			0.9801
Space Group	C2			P6 ₃
a, b, c (Å)	74.46, 42.91, 103.34			42.99, 42.99, 94.68
α, β, γ (°)	90.0, 96.10, 90.0			90.0, 90.0, 120.0
Resolution (Å) ^a	37.13-1.90 (1.94-1.90)			19.97-1.50 (1.53-1.50)
Total observations	175918			261006
Unique reflections	25522			15915
Multiplicity ^a	6.9 (7.1)			16.4 (15.5)
Mean $I/\sigma(I)$ ^a	10.2(4.0)			21.7 (5.9)
Completeness (%) ^a	98.8 (98.1)			99.9 (100)
$R_{\text{merge}}^{\text{a,b}}$	0.13 (0.53)			0.081 (0.445)
R_{pim}				0.030 (0.274)
$CC_{1/2}^{\text{a,c}}$	0.99 (0.89)			0.999 (0.935)
Refinement				
Reflections: working/free ^d	24288 / 1224			15120 / 762
$R_{\text{work}} / R_{\text{free}}^{\text{e}}$	0.181 / 0.238			0.149 / 0.178
Ramachandran plot: allowed/favoured/outliers (%)	100 / 97 / 0			100 / 97 / 0
R.m.s. bond deviations (Å)	0.014			0.014
R.m.s. angle deviations (°)	1.840			1.855
R.m.s. chiral deviations	0.093			0.085
No. atoms / Mean B-factors (Å ²)	Chain A	Chain B	Chain C	Chain A
Protein	962 / 22.1	971 / 22.2	951 / 21.3	952 / 15.6
carbohydrate ligand	58 / 36.1	58 / 33.3	47 / 30.6	-
ligand ^f	14 / 32.0	14 / 38.7	14 / 36.0	-
water	69 / 26.9	65 / 26.6	72 / 24.9	150 / 28.2

^a Values for the outer resolution shell are given in parentheses.

^b $R_{\text{merge}} = \sum_{\text{hkl}} \sum_i |I_i(\text{hkl}) - \langle I(\text{hkl}) \rangle| / \sum_{\text{hkl}} \sum_i I_i(\text{hkl})$.

^c $CC_{1/2}$ is the correlation coefficient between symmetry-related intensities taken from random halves of the dataset.

^d The data set was split into "working" and "free" sets consisting of 95 and 5% of the data, respectively. The free set was not used for refinement.

^e The R-factors R_{work} and R_{free} are calculated as follows: $R = \sum (|F_{\text{obs}} - F_{\text{calc}}|) / \sum |F_{\text{obs}}|$, where F_{obs} and F_{calc} are the observed and calculated structure

factor amplitudes, respectively

^f refers to ligands bound in the active site and potential surface binding sites

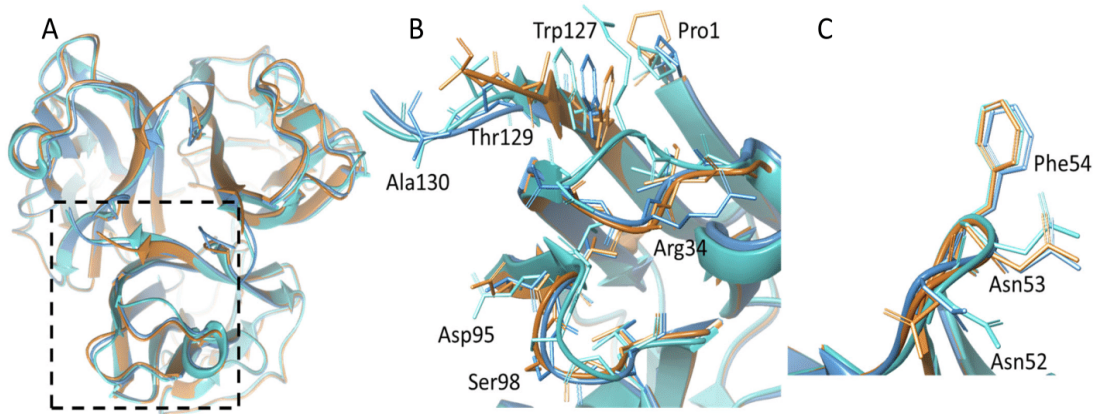


Figure S1 A) Overview of the surface loops in the lower face of BC2L-C-nt holo (cyan (2WQ4) and azure (6TIG)) and apo forms (orange (7BFY)). B) Close-up view of the loop conformation changes observed in both forms. C) Minor differences observed in the loop (Asn52-Phe54) near oligosaccharide binding site.

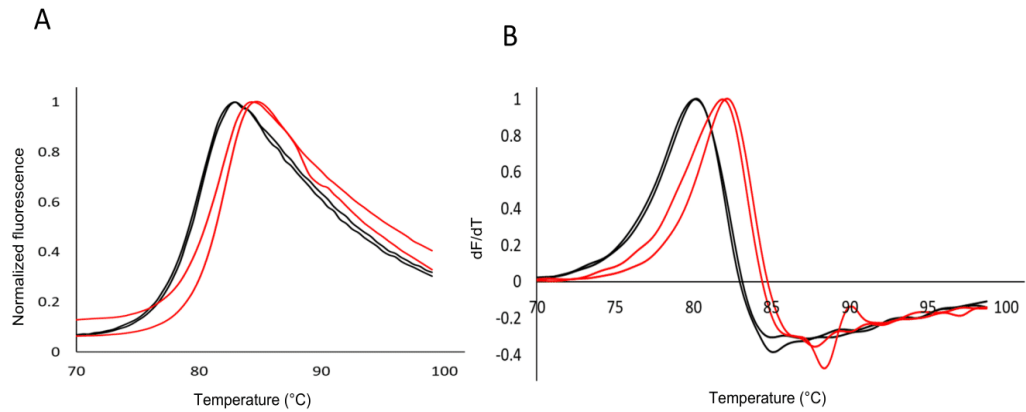


Figure S2 A) Methyl α -L-fucoside (20 mM) binding studies showed a positive shift in the melting temperature (T_m) of BC2L-C-nt in the presence (red) and absence (black) of methyl α -L-fucoside. B) First derivatives of fluorescence curves.

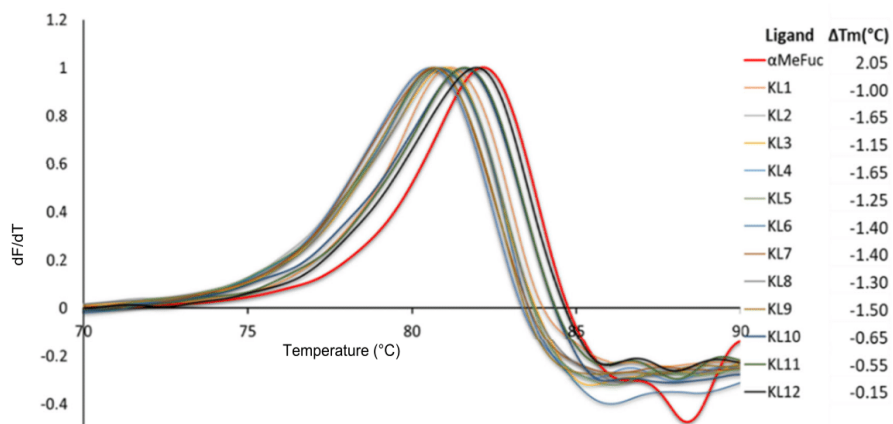


Figure S3 First derivatives of fluorescence curves of the fragments (KL1-12) in the presence of methyl α -L-fucoside (20 mM). Fragments KL1-12 causes a negative shift in the melting temperature (T_m) of BC2L-C-nt which indicates ligand interaction.

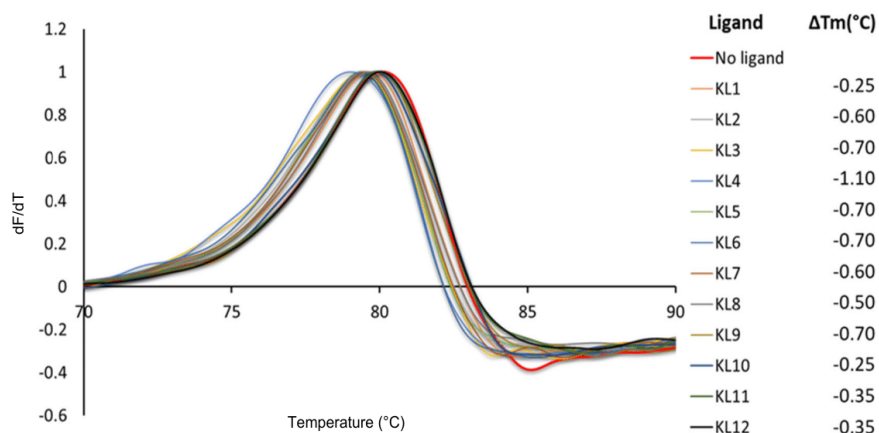


Figure S4 First derivatives of fluorescence curves of the fragments (KL1-12) in the absence of methyl α -L-fucoside. Fragments KL1-12 show a negative shift in the melting temperature (T_m) of the protein which indicates ligand interaction.

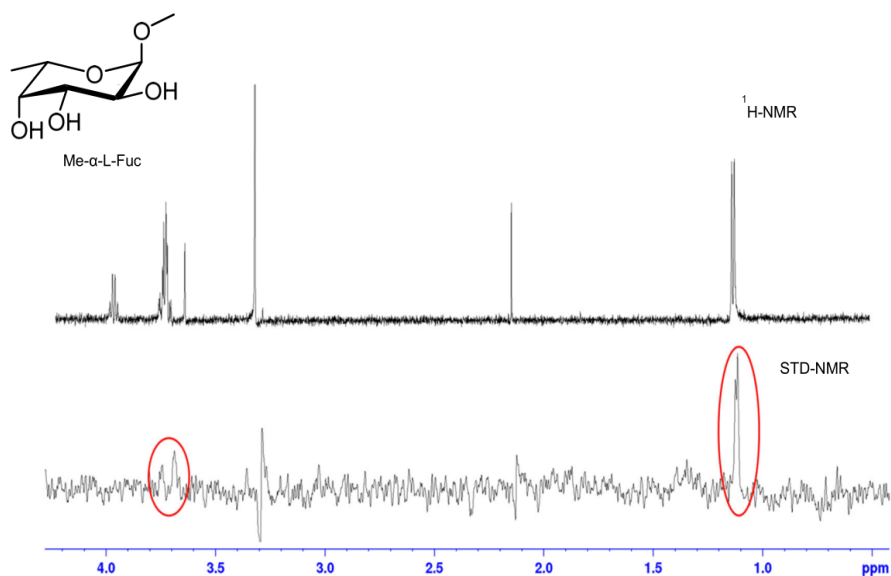


Figure S5 $^1\text{H-NMR}$ (upper) and STD spectrum (lower) methyl α -L-fucoside in the presence of BC2L-C-nt (1000:1) recorded with a Bruker Avance 600 MHz spectrometer. The spectrum is recorded at 298K with irradiating frequency -0.05 ppm. In the STD spectrum, the signals of the fucose ring at 3.7 ppm and of the methyl group at 1.1 ppm are highlighted with red circles.

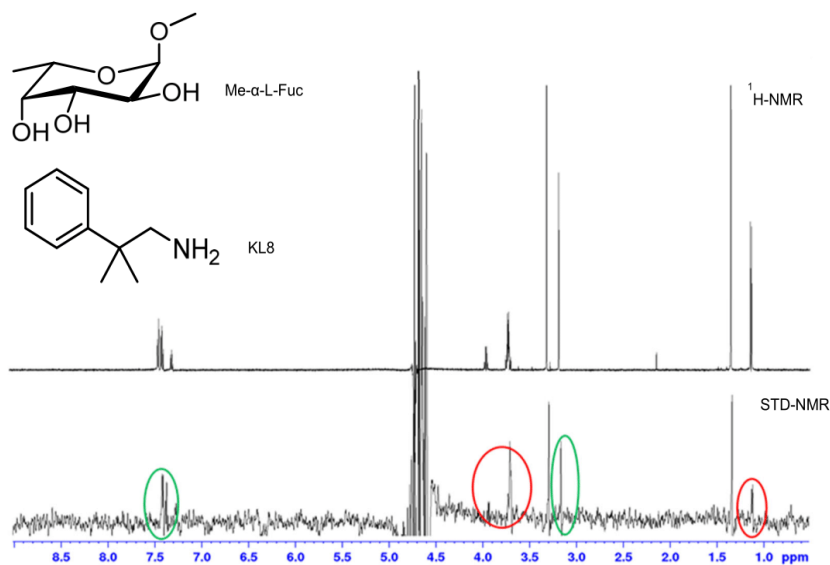


Figure S6 $^1\text{H-NMR}$ (upper) and STD spectrum (lower) of fragment KL8 and methyl α -L-fucoside in the presence of BC2L-C-nt (1000:1) recorded with a Bruker Avance 600 MHz spectrometer. The spectrum is recorded at 298K with the irradiating frequency at -0.05 ppm. In STD spectrum, the signals of the fucose ring around 3.7 ppm and of the methyl group at 1.1 ppm are highlighted with red circles. The signals of fragment KL8 are highlighted with a green circle (at 3.2 ppm for $-\text{CH}_2\text{-NH}_2$ and 7.4 ppm for aromatic protons).

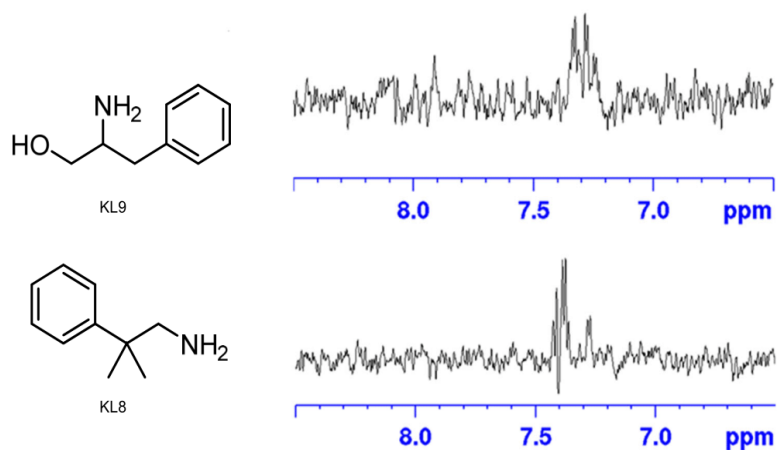


Figure S7 The STD spectra of fragment KL8 (lower) and KL9 (upper) in the presence of the BC2L-C-nt (1000:1) recorded in the presence of Methyl α -L-Fucoside with a Bruker Avance 600 MHz spectrometer. The spectra are recorded at 298K with the irradiating frequency at 10 ppm. In these spectra only the aromatic protons of the fragments are observable.

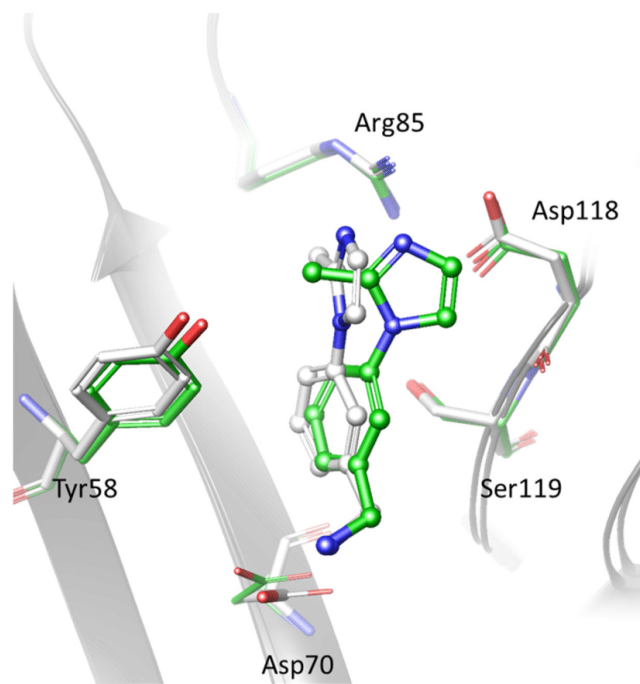


Figure S8 Comparison of binding pose of docked (grey) and crystallized complex (green) with KL3 (RMSD 0.4 Å).

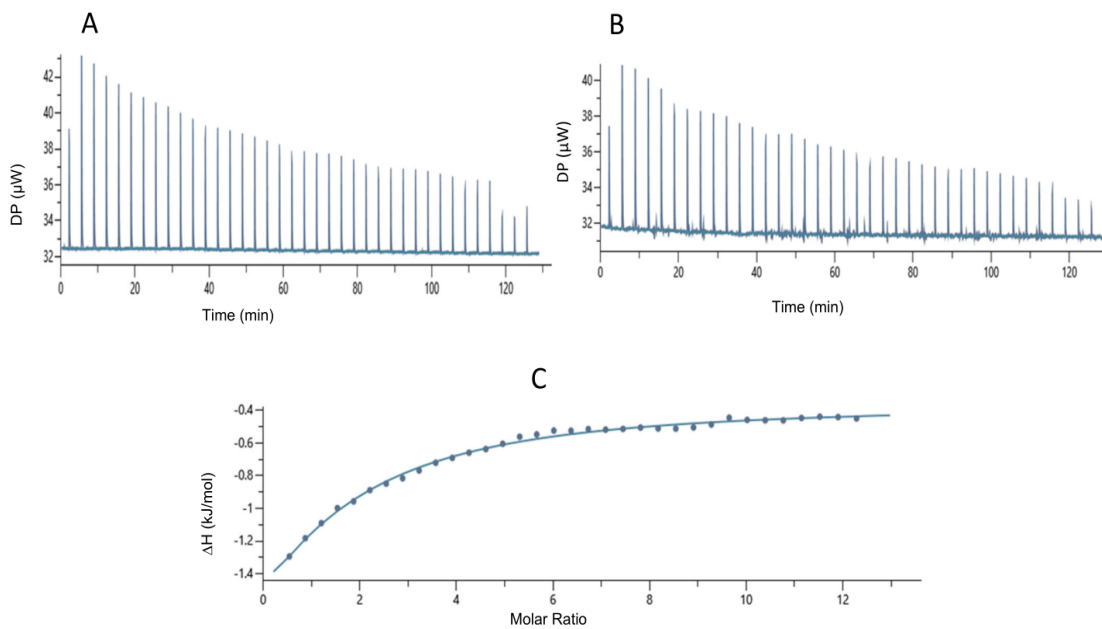


Figure S9 Isothermal microcalorimetry data. Titration of buffer by ligand (KL3, 15 mM) (A) and titration of BC2L-C-nt (225 µM) by KL3 (15 mM) at 25 °C (B). Point-by-point differences between ligand-in-protein and ligand-in-buffer for KL3 (C). The curve was fitted using the "one binding site" model.