STRUCTURAL BASIS OF *O*-SIALOGLYCAN RECOGNITION AND SELECTIVITY IN α2,3 SELECTIVE STREPTOCOCCAL SERINE RICH REPEAT ADHESINS AND α2,6 LINKAGE SPECIFIC BACTERIAL SIALYLTRANSFERASES

By

Haley Elizabeth Stubbs

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

December 16, 2023

Nashville, Tennessee

Approved:

Vsevolod Gurevich, Ph.D. (Chair)

Jennifer Gaddy, Ph.D.

Ivelin Georgiev, Ph.D.

Steve Townsend, Ph.D.

T. M. Iverson, Ph.D.

DEDICATION

ACKNOWLEDGEMENTS

I would like to acknowledge both the Vision Training Grant (5T32EY007135) and the Molecular Biophysics training program (2T32GM00832) for funding. This research was also supported by the National Institutes of Health-National Institute of General Medicine and American Heart Association (AI106987) grants to T.M Iverson. I would like to thank all past and present members of the Iverson lab: Ali Kaya Ph.D., Brandon Butler, Ivette Perez Ph.D., Izumi Yamakawa, Jordan Stacey, KeAndreya Morrison, Kole Martin, Laura Frigo, Lioudmila Loukachevitch, Nicole Perry Ph.D., Pankaj Sharma Ph.D., Prashant Singh, and Sandra Berndt Ph.D. I would like to specifically highlight Ali Kaya, Izumi Yamakawa, Ivette Perez, Pankaj Sharma, Prashant Singh, and Sandra Berndt for their training and patience. I would like to thank Barbara Bensing Ph.D. for hosting me at the VA at UCSF and coauthoring two papers with me. I would like to thank Stacey Satchell for her guidance in developing my professional skills. I would like to thank the BRET Office for hosting the ASPIRE on the Road trip and professional skill modules. I would specifically like to thank Kate Stuart for always making time to help me sort out my thoughts on my future steps. I would also like to thank Tina Iverson Ph.D., for welcoming me into her lab and her time and effort put into my training.

I would like to thank my friends and family for their constant support. I would like to thank Laura Teal Ph.D., for being a steadfast friend, in good times and bad. I would like to thank Karen Buchmueller Ph.D., my undergraduate research PI, for modeling strong mentorship. I would like to thank my undergraduate labmates Daniela Mesa-Sanchez Ph.D., Kaitlyn Browning Ph.D., and Ralph White III Ph.D., for their constant encouragement and solidarity.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

i. INTRODUCTION

Glycan structure characteristics

Glycans, or polysaccharides, play an essential role in biological processes. However, our knowledge of their structure and function is limited as compared to other biological molecules, like DNA and protein. Glycan structure is highly variable due to differences in monosaccharide sequence, linkage, degree of branching, and other chemical modifications. A monosaccharide is simply a carbohydrate molecule that cannot be hydrolyzed into a smaller carbohydrate molecule. Monosaccharide structure varies in the number of carbons, the presence of a terminal aldehyde or ketone, and the number and position of hydroxyl groups on each carbon. These base monosaccharide structures can also be modified with acetyl, sulfo, amino, carboxyl, and methyl groups. To make a polysaccharide, monosaccharides are connected through several types of glycosidic bonds. These bonds can be either α or β linkages. These bonds are further diversified by the carbon positions that are linked on each monosaccharide. The type of linkage can significantly change the shape of the overall glycan. There is an established icon-based nomenclature, where monosaccharides are represented by colored circles, diamonds, triangles, or squares.[1, 2] Linkages are shown in a black line and labeled with the corresponding type of linkage and carbon residues that are linked. A glycan modifying a protein can either be described as *N*-linked or *O*-linked depending on what amino acid residue the glycan is modifying. *O*-linked glycans are linked via the hydroxyl group of serine, threonine, and in more rare cases a tyrosine residue. *N*-linked glycans are attached to the sidechain nitrogen atom of asparagine. The rest of this thesis will focus solely on *O*-linked glycosylation. Additionally, *O*-linked glycans can be described as “core” glycans. There are three types of *O*-glycan core structures. Core 1 glycans consist of an N-acetylgalactosamine connected to a serine or threonine amino acid. Core 2 glycans are *N*-acetylglucosamine β1,6 N-acetylgalactosamine connected to a serine or threonine residue. The final type of core *O*-glycan is core 3, N-acetylglucosamine β1,3 N-acetylgalactosamine connected to a serine or threonine residue.

The most common type of monosaccharides in mammals are sialic acids, which are defined by having a nine-carbon backbone. Over 50 forms of sialic acid have been identified in nature with the most abundant being

Neu5Ac and Neu5Gc. Abundance of a specific sialic acid is dependent on the organism and cell type. For example, humans evolutionarily lost the enzyme that synthesizes Neu5Gc from Neu5Ac. Thus, over 90% of sialic acid found in humans is Neu5Ac,[3] although trace amounts of Neu5Gc absorbed from food sources have been detected. Other higher order primates like chimpanzees have ~70% Neu5Gc.[4] In mammals many glycans are terminated with sialic acid.

Structures of glycans have been investigated extensively by chemists using nuclear magnetic resonance (NMR) spectroscopy and computational modeling. Glycans are highly flexible and therefore have a myriad of conformational states they can occupy. Methods of glycan structure modeling and models built thus far are further reviewed in the literature.[5]

Sialoglycan binding proteins

To use glycans for cellular signaling and recognition, organisms need proteins that can specifically bind glycans. Mammals do this with four families of binding proteins: lectins, galectins, Sialic acid binding immunoglobulin-like lectins (Siglecs; I-type lectins), and C-type lectins (including selectins which primarily recognize sLe^X motifs). Mammalian immune systems typically use surface expressed Siglecs for self-recognition. These proteins are primarily expressed on leukocytes with the specific Siglec expression profile depending on the cell type and differentiation.[6] Siglecs bind a wide range of sialoglycans and most play an inhibitory role. Homeostatic human epithelial cells express many sialoglycans on their cell surface. Siglecs on immune cells bind these sialoglycans and most inhibit an immune response. This is how the immune system identifies cells as self.

Mammalian glycosylation and extracellular glycan presentation

Mammalian cells abundantly present glycans on their cell surface through expression of glycoproteins and glycolipids. These glycolipids and glycoproteins make up a viscous layer around the cell surface known as the glycocalyx. The composition of the glycocalyx, sometimes referred to as a self-associated molecular pattern (SAMP) is cell type specific. *In vivo,* glycan structure is determined by enzymes that synthesize, transfer, and cleave monosaccharides. These enzymes are known as synthases, glycosyltransferases, and glycosidases

respectively. Most of these enzymes are housed within the Golgi apparatus, with a few exceptions located in the smooth endoplasmic reticulum. Unlike biomolecules like DNA, RNA, and proteins, glycans are not made in a template dependent manner. When synthesizing DNA, a template strand serves as instructions for the order of addition of nucleotides. For RNA synthesis, DNA serves as a template. Finally, for proteins, RNA serves as a template. When glycans are attached to proteins or lipids and elongated, there is no template to produce certain glycans on specific protein or lipid sites with high fidelity. Instead, factors like substrate availability, cell-specific expression levels of glycosylation enzymes and glycosylated proteins, sublocalization of glycosylation enzymes within the Golgi apparatus, and glycosylation enzyme substrate specificity create a complex and dynamic equilibria that control the composition of the glycocalyx. The complexity of this highly dynamic process is reflected in the fact that 1-2% of human genes are involved in glycosylation.[7]

Enzymes responsible for transferring a monosaccharide to another biomolecule are called glycosyltransferases. It is estimated that humans have approximately 200 different glycosyltransferases. They can be subdivided into lipid or polypeptide glycosyltransferases depending on whether they transfer a glycan to a lipid or polypeptide chain. These can be further subdivided depending on the glycan they transfer. For example, GalNAc-Ts transfer N-acetylatedgalactosamine. Each family has many proteoforms that vary slightly in expression level, sublocalization, and acceptor substrate specificity. This redundancy creates a more finely tunable system to regulate glycosylation patterns. GalNAc-Ts are glycosyltransferases that transfer GalNAc to a serine or threonine residue. Humans have twenty GalNAc-Ts that are categorized into nine subfamilies. GalNAc-T localization is dependent upon different domains depending on the specific GalNAc-T.[8]

Glycosyltransferases that transfer sialic acid are known as sialyltransferases. Humans have twenty sialyltransferase enzymes that can be identified by the presence of four different sequence motifs.[9, 10, 11, 12] There are four families: ST3Gal, ST6Gal, ST6GalNAc, and ST8Sia. These sialyltransferase families transfer sialic acid to a galactose with an α2,3 linkage, galactose with an α2,6 linkage, *N*-acteylgalactosamine with an α2.6 linkage, or to sialic acid with an α2,8 linkage respectively. Humans have six ST6GalNAcs.

Disease state glycosylation in humans

Misregulation of any of these glycan processing steps can lead to human disease and presentation of a unique glycocalyx or damage-associated molecular pattern (DAMP). For example, mutations in many types of cancer perturb glycoprocessing pathways. This can lead to overrepresentation of specific glycans on cancer cells. These glycans are sometimes referred to as tumor associated markers (TAMs) or to glycoproteins they modify called tumor associated glycoproteins (TAGs). The role of glycans in cancer progression remains elusive. Overall, a positive correlation between sTn positive cancer cases and poor prognosis has been found. However, cancer characteristics correlating with aberrant sTn expression like local invasion, decreased patient survival, and lymph node metastasis are dependent upon cancer type.[13] One hypothesis as to how glycans may impact cancer progression is by helping cancer cells to evade the immune system.

One particular cancer-related glycan of interest is the disaccharide sialyl Thompson-nouvelle antigen (sialosyl-Tn antigen; sTn; Neu5Acα2,6GalNAc)[16, 17], which has been detected at elevated serum levels in colon, gastric, breast, lung, esophageal, pancreatic, endometrial, bladder, and prostate cancers.[16] Efforts to develop cancer diagnostics through sTn detection have included antibody detection, electrochemical biosensors[18, 19], and multivariate -omics analyses.[20] In humans, sTn is produced by two types of glycosyltransferases: a polypeptide-GalNAc-transferase and an α2,6 linkage specific sialyltransferase. Humans have two sTn-producing α2,6 linkage specific sialyltransferases, ST6GalNAc I and II. First, a GalNAc monosaccharide is transferred to either a serine or threonine residue by one of twenty human polypeptide-GalNAc-transferases. Interestingly, GalNAc-T3 has been found to be expressed in adenocarcinomas, but this expression correlates with a greater survival rate.[21] At this point, galactose (Gal), N-acetylated galactose (GalNAc), N-acetylated glucose (GlcNAc), or sialic acid (Neu5Ac) could be transferred to the *O*-linked GalNAc. Healthy tissues commonly display more elongated glycans as compared to cancerous ones. This is because the glycosyltransferases have a high enough expression level and enzymatic activity to compete with sialyltransferases. In many epithelial cancers however, changes in expression level or mutations that affect enzymatic activity of any of the glycosyltransferases shift

the cellular equilibrium to a higher relative activity of sialyltransferases and thus increased production of shorter glycans such as sTn.

Other glycans have also been associated with human disease. For example, similar to Neu5Gc, humans do not produce the glycan Galα1,3Gal (αGal). This glycan is however found on glycoproteins and glycolipids of many other mammals. Humans can absorb αGal through consumption of red meat. αGal can elicit production of several types of antibodies and in some cases cause an allergic reaction known as αGal syndrome. This condition was recently reviewed by Román-Carrasco et al.[14]

Proposed mechanism of cancer immune evasion via overexpression of sTn

It has long been demonstrated that sTn overexpression in many epithelial cancers correlates with poor prognosis. However, the mechanism for this phenomenon remains somewhat elusive. Studies have shown that sTn overexpression with clinical characteristics associated with a poor prognosis are dependent on the cell type. For example, sTn overexpression positively correlates with lymph node metastasis in esophageal cancer, but not in cervical cancer. This indicates that the role sTn plays in cancer signaling and progression is cell type dependent.[22] One overarching hypothesis is that sTn aids cancer cells in evading the immune system. Cancerous cells aberrantly express sialoglycans, particularly sTn. Therefore, cancer cells present a plethora of ligands for Siglecs to bind and prevent the immune system from recognizing cancer cells as an antigen. A recent study with a HEK293 cell-based glycan array demonstrated that Siglec-15 is specific for binding sTn epitopes.[23]

Probe development against sialylated glycans

Given that glycosylation is a biomarker for many diseases, probes targeting glycans have been highly sought after for decades. Several challenges have prevented development of a probe reliable and specific enough for regular clinical use, whether for diagnostic or therapeutic use. First, as explained previously, glycans have an extreme amount of structural diversity. Second, antibody development against glycans is challenging due to sialoglycans having poor immunogenicity.

Development of antibodies to recognize sTn.

Antibodies have long been a good scaffold for probe development. This class of proteins is well-characterized across several scientific disciplines including structurally, biochemically, and biologically. Innovative methods have increased our ability and throughput in developing antibodies against novel targets. Antibody development against glycans has produced useful tools to date, but still faces significant challenges in producing a probe specific enough for diagnostic or therapeutic use. One of the biggest challenges in targeting glycans with antibodies is poor immunogenicity. Since mammals use glycans to recognize tissues as self, immunization with glycans often does not produce an immunologic response. Even if one is produced, higher affinity IgG antibodies are not often produced.

To leverage the immune system against cancer, vaccinations against sTn have been developed with mixed clinical success.[25, 26, 27] One explanation of variability in clinical outcomes is the poor immunogenicity of sTn. Homeostatic sTn expression is limited to small amounts which are mostly within the digestive tract. Since sTn is endogenously expressed in some adult tissues, injection of sTn only induces a T-cell independent immune response, leading to immunotolerance of sTn. To better understand the efficacy of targeting sTn for cancer and to provide improved tools for cancer detection, probes specific for sTn are needed.

The first antibodies, TKH1, TKH2, B72.3, and MLS102, developed against sTn were produced by mouse immunization with epithelial cancer cell lines. It was not discovered till later that these antibodies target the disaccharide sTn. B72.3 was found to immunoprecipitate a glycoprotein named tumor associated glycoprotein 72 (TAG72).[28, 29] This protein was then used for immunization and resulted in production of a second generation antibody named CC49.[30] A couple of other antibodies were then developed through immunization with purified sTn glycan and mucins known to be modified with sTn. Free glycan immunization only produced low affinity IgM antibodies. Later a protein known as keyhole limpet hemocyanin (KLH), known to be unreactive in humans, was used as a carrier protein to increase immunogenicity of sTn. A single KLH protein monomer can be modified with ~3000 glycans. Injection of sTn conjugated to a carrier protein induces a T cell dependent immune response leading to production of high affinity IgG antibodies. Conjugation to a carrier

protein has two main advantages over injection of the glycan alone: 1) glycan conjugated peptides can be displayed on the surface of dendritic cells whereas a free glycan cannot, 2) a carrier protein can be conjugated with hundreds of glycans, allowing for greater amounts of glycan presentation on dendritic cells. Immunization with a highly glycosylated protein rather than a free glycan presents opportunities for recognition of sTn clusters rather than single sTn glycans. sTn conjugated KLH was then developed into a vaccine that ultimately failed during phase 3 clinical trials, which has been reviewed in greater depth.[22]

Name/catalog number/company	Year created	specificity	Method of creation
TKH2[31]	1988	sTn[31]	Immunization with ovine submaxillary mucin
B72.3	1981	sTn[31], Tn[33]	Immunization with human metastatic breast cancer cell line
MLS102[34, 35]	1988		Mice immunized with human colonic cancer cell line LS180
CC49[30]	1988	Clustered sTn-serine, Tn-serine	Mouse immunized with tumor associated glycoprotein 72 (TAG-72) that immunoprecipitated with B72.3 antibody
MA54[36]	1989	Carbohydrate on mucin-type glycoprotein	Immunization with lung adenocarcinoma cell line
MA61[36]	1989	Neu5Ac α2,6 terminally linked glycans	
3P9[37]	2013		Immunization with colorectal adenocarcinoma cell line SW1116
L2A5[38]	**2018**	Neu5Ac α2,6 terminally linked short glycans	Hybridoma technology

Table i-1 anti-sTn antibodies.

The efficacy of immunization against sTn

After it was discovered that sTn abundance was associated with many types of epithelial cancer, it has been a highly sought therapeutic target. Primary efforts in targeting sTn were for diagnostic purposes, however, it was soon discovered that sTn expression positively correlates with a more invasive and aggressive cancer.[22] This could mean that targeting sTn could have a therapeutic effect. To increase the specificity and affinity of the antibodies produced, studies have also investigated infusion with an sTn conjugated protein found on the surface of cancer cells. sTn modified MUC1 peptides conjugated to KLH produce high titers of antibodies in mice.[39, 40] Small molecule and nanoparticle conjugates have also been synthesized and tested.[41]

A vaccine against sTn called Theratope was developed by a company called Biomira. Theratope is an QS21 adjuvanted sTn conjugated to keyhole limpet hemocyanin (KLH), a carrier protein known to not induce adverse side effects.[42] Theratope failed its phase three clinical trials, failing to provide evidence that vaccination extended patient life or time to remission for women with metastatic breast cancer.[25] Researchers have postulated many reasons for the failure of this trial, which have been reviewed.[22] One reason may be that patients were not screened for sTn presence before being admitted to the trial. While many epithelial cancers contain aberrant sTn expression, many factors can affect sTn abundance. For example, specific cell types display a higher sTn abundance. Second, sTn vaccination may be more effective in preventing cancer than treating it. While sTn abundance varies significantly between organ and cell type, it is generally supported that an overexpression of sTn occurs early in carcinogenesis. sTn presence is typically detected with blood samples. For sTn to be detected in blood samples, cancerous cell debris must enter the bloodstream. By the time cancer has been detected, aberrant expression has already begun. Since sTn expression can aid cancers in evading the immune system, it is likely more difficult for the immune system to recognize sTn as an antigen and develop antibodies against it. It has been suggested that sTn vaccination prior to the development of cancer could be effective, but clinical trials for this have not been designed to date.

Available sialoglycan non-antibody probes

Instead of relying on antibodies as probe scaffolds, we can use proteins with native glycan binding properties. There has been an ongoing effort to better understand the selectivity of glycan binding proteins in the hopes of engineering a probe for various glycan structures. Glycan binding proteins (GBPs) are evolutionarily ancient and are encoded for in viruses, bacteria, plants, and animals. These efforts have mostly been targeting sialylated glycans due to their prevalence in humans. Probes against sialylated glycans would therefore have a greater clinical impact. Recently, a number of mammalian, viral, and bacterial glycan binding proteins were screened against glycan arrays. Several proteins were identified and characterized for recognition of specific terminal sialic acids and sialic acid linkages.[3]

sTn, however, has proven to be a difficult glycan to target. No known bacterial or viral binding proteins have been found to bind sTn to date. Human cluster of differentiation-22 (CD22), a Siglec found on the surface of B cells known to prevent overactivation of the immune system, binds only a few α2,6 linked sialoglycans. The few glycans it does bind, it has a weak binding strength for. Notably, human CD22 does not bind sTn. In contrast, mouse CD22 bound many different α2,6 linked sialoglycans, but had very little binding strength for sTn.[3] The non-antibody probe most widely used for detecting α2,6 linked sialoglycans is a lectin from Sambucas nigra (elderberry; SNA/EBLI).[3] While in vitro studies have shown that SNA selectively binds Neu5Acα2,6GalNAc (sTn), a recent cell-based glycan array displayed SNA-I selectivity for Neu5Acα2,6Galβ1,4GlcNAc.[23] Another study using a cell-based glycan array demonstrated that Siglec-15 seems to be selective for sTn. Siglec-15 binds a range of mucin proteins, but the level of binding to each is different. This could be caused by two things. First, Siglec-15 could prefer one mucin over another due to both the mucin and the Siglec's three-dimensional structure. Second, certain mucins may display more glycosylation sites, particularly sTn.[23] Engineering other glycan binding proteins to create affinity for Sia has been shown possible. A galactose binding earthworm lectin EW29 was engineered using evolutionary strategies to have specificity for Neu5Ac α2,6 terminally linked glycans.[43]

Rationale for using a sialyltransferase for sTn probe development.

The protein described above currently being sought after for sTn probe development are less than ideal. An effective probe is stable, has low nonspecific and off target binding, and high affinity for a single target. Unfortunately, to date no known GBPs meet these criteria for α2,6 linked sialoglycans. Human CD22 demonstrated little to no affinity for sTn on microarrays. Additionally, mammalian proteins including antibodies are typically not very stable or easy to express and purify. Other GBP options for probe development are bacterial and viral proteins. However, viral proteins also have low stability and there are no known bacterial GBPs that bind to α2,6 linked sialoglycans.

With limited success in antibody development and without stable glycan binding proteins specific for sTn, probe development for sTn has eluded scientists for decades. Here, I propose using a novel starting scaffold for probe development, a sialyltransferase. Previous research has demonstrated that bacterial sialyltransferases are stable and can be expressed in high quantities. Additionally, several bacteria and different bacterial strains produce a sialyltransferase that specifically catalyzes an α2,6 linkage. Enzymes have a native affinity for both a substrate and product. Since the products of bacterial α2,6 linkage specific sialyltransferases are α2,6 linked sialoglycans, we know that these enzymes have a native affinity for the target. The catalytic activity of these sialyltransferases can be eliminated through mutation of active site residues. This converts the sialyltransferase to a GBP, ready for probe development.

Here, I use a rational approach for probe development. Probe development for glycan binding proteins using an evolutionary approach has been successful for other targets. An evolutionary approach is a preferred approach when there is little known about the starting affinity and selectivity of a protein. However, there is an abundance of knowledge of the structure and biochemistry of bacterial α2,6 linkage specific sialyltransferases (Table i-2). Many of these enzymes were identified and characterized in the 2000s for use in synthesizing polysaccharides. Additionally, the kinetic activity of several sialyltransferases has been measured with multiple acceptor substrates.[44, 45] These structures and kinetic data will aid in the design of mutant sialyltransferases that will knock out catalytic activity and have increased affinity for the target sialoglycan, sTn.

Structure and mechanism of bacterial sialyltransferases

Bacterial glycosyltransferases all have either a glycosyltransferase A (GT-A) or Glycosyltransferase B (GT-B) fold. There are both GT-A and GT-B bacterial sialyltransferases. There is a C-terminal tail, and N-terminal and C-terminal sialyltransferase domains. The donor substrate for all bacterial sialyltransferases is CMP-sialic acid. The acceptor substrate, however, can be a number of different glycans attached to proteins or lipids. The donor and acceptor substrates bind sequentially. First the donor substrate CMP-sialic acid binds, causing the sialyltransferase to change conformation to a closed state. The C-terminal domain has little to no local conformational changes. The N-terminal domain rotates more than 20° towards the C-terminal domain enclosing the CMP-sialic acid donor substrate and creating a tighter pocket for binding of an acceptor substrate.[46] The size and charge of the acceptor substrate binding pocket can vary significantly between sialyltransferases. For example, the acceptor substrate binding pocket of the sialyltransferase from *Pasteurella multocida* is wider and more negatively charged than the pocket of *Vibrionaceae Photobacterium* sp. JT-ISH-224 sialyltransferase.[46, 47] This could lead to differences in acceptor substrate preference. To date, structural studies with sialyltransferases have only included lactose as an acceptor substrate. However, kinetic studies have shown differences in Km values and the amount of sialoglycan produced depending on the acceptor substrate for multiple different sialyltransferases.[44, 45]

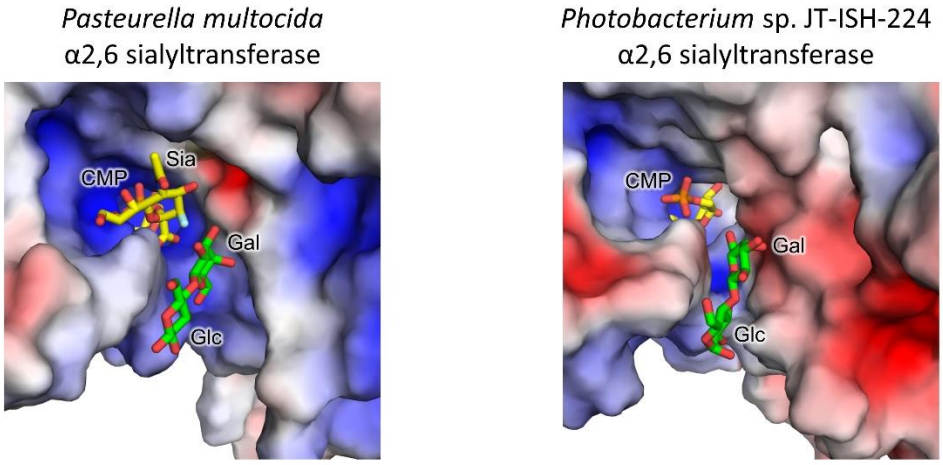

Figure i-1 Donor and acceptor substrate binding pockets of two bacterial sialyltransferases.

A2,6 linkage specific sialyltransferases from *Pasteurella multocida* (left) and *Photobacterium* sp. JT-ISH-224 (right) are shown in a surface representation and colored by electrostatic charge. Red indicates negative charge, blue positive charge, and white a hydrophobic surface. Lactose, an acceptor substrate is shown in green sticks. The donor substrate, CMP or CMP-sialic acid is shown in yellow sticks. Oxygen atoms are shown in red.

Structure of streptococcal serine rich repeat adhesins

While there are not bacterial glycan binding proteins selective for α2,6 linked sialoglycans, some exist for α2,3 sialoglycans. Many streptococcal bacteria express proteins on their cell surface that facilitate adhesion to host surfaces. Fittingly named adhesins, these proteins can bind glycoproteins or glycolipids known as host receptors on the surface of host cells. Streptococci express a specific type of adhesin known as serine rich repeat adhesins or SRRPs, some of which are selective for binding α2,3 sialoglycans. These proteins have five functional domains: an aSec transport signal sequence, a short serine rich repeat sequence, a Siglec-like binding region, a second longer serine rich repeat sequence, and a cell wall anchoring motif. Serine comprises approximately 50% of the amino acids in both serine rich repeat domains. These domains contribute to an overall fibril-like structure, extending the binding region(s) of the protein away from the cell surface where they can more readily access and bind host receptors.[48]

The Siglec-like binding region (SLBR) is named after mammalian sialic acid binding immunoglobulin-like lectins (Siglecs) with which they share a V-set Ig fold. While both bacterial SLBRs and mammalian Siglecs bind sialylated glycans, the glycan binding pockets differ. Mammalian Siglecs recognize sialic acid using a conserved arginine at the N-terminal half of the F strand. Some bacterial SLBRs also have an arginine in the same position, but it is not evolutionarily conserved. While mutagenesis studies show that this residue contributes to SLBR affinity for sialoglycans,[49] x-ray crystallography suggests this is not the primary motif used for sialic acid recognition in bacterial SLBRs. A ϕTRX motif positioned in the C-terminal half of the F strand makes several hydrogen bond contacts to sialic acid. The exact number of contacts depends on the specific sequence of the ϕTRX motif. For example, $SLBR_{Hsa}$ has a more canonical YTRY motif, which forms five hydrogen bonds with Neu5Ac.[50] Contrastingly, $SLBR_{SK1b}$ has a YTFK motif and only forms two hydrogen bonds with sialic acid.[51] While the mammalian homologous arginine residue and ϕTRX motif are on the same

surface of the protein, these binding pockets are not necessarily contiguous. The FG loop of bacterial SLBRs varies significantly in its length and conformation. $SLBR_{GspB}$ and $SLBR_{Hsa}$ have FG loops that effectively wall off the glycan binding pocket to the N-terminal half of the F strand.[51, 52] $SLBR_{SrpA}$ and $SLBR_{SK1b}$, however, have small FG loops. This leaves space for a longer glycan to contact the ϕTRX and the N-terminal half of the F strand.[49, 51] The binding pocket is surrounded by three loops: the CD, EF, and FG loops. Between adhesins, these loops vary considerably in sequence and length. SLBR selectivity for specific glycan features can be attributed to differences in the selectivity loops, hence their name. Only one loop, the EF loop hydrogen bonds with sialoglycans. When bound, a backbone carbonyl of the EF loop hydrogen bonds to O4 of Neu5Ac or Neu5Gc. The CD and FG loops can hydrogen bond with the second and third nonreducing end glycan units, though most contacts are weaker electrostatic ones.

Challenges in assessing affinity and selectivity of glycan binding proteins for glycans.

Measuring the binding strength or affinity of a protein for a specific ligand is critical information for quantitative analysis of biochemical events. For example, if a protein is in the presence of three ligands you will have an environment where competitive binding will occur. Dissociation constants describing the equilibrium between a bound and unbound state of the protein for each ligand can be used to model the likelihood of a protein binding one ligand over another in solution. Many biochemical techniques including Surface Plasmon Resonance, Isothermal Calorimetry, and microscale thermophoresis can be used to detect and measure a dissociation constant between two biomolecules. Notoriously in the field, measurements of glycan affinity for a glycan binding protein (GBP) determined by different biochemical techniques often do not agree. There are several technical and biological reasons for this.

The first consideration is simply the method of detection. Surface plasmon resonance as well as ELISAs require immobilization of either the protein or ligand of interest. In this scenario, one biomolecule is able to freely move through 3D space while the other is stationary. The inability of the immobilized molecule to diffuse will affect the kinetics of the molecule and thus affect the dissociation constant. Another more biologically interesting reason that binding strength measured through different techniques may disagree is due to the

context of the glycan linkage. As described previously, glycans can exist as a free molecule, or as a protein or lipid modification. Many biochemical techniques used to measure glycan affinity are designed using free synthetic glycans. Depending on the biological interaction, a protein may only make contacts with the glycan. In other instances, the protein may bind the glycan and the molecule it is covalently attached to. For O-glycosylated glycoproteins this could include the amino acid the glycan is directly modifying or other adjacent residues. A third possibility is a protein binding a cluster of glycans, creating an avidity effect. This has been shown to be true for several anti-sTn antibodies.[53]

Bacteria	Sialyltransferase	PDB Code	Ligand binding state
Pasteurella multocida	PMST	2EXO	Apo, unliganded
Pasteurella multocida	PMST	2EX1	CMP bound
Pasteurella multocida	PMST	2IIQ	CMP bound, open conformation
Pasteurella multocida	PMST	2ILV	CMP and alpha lactose bound
Pasteurella multocida	PMST	2IHJ	CMP-3F(axial)-Neu5Ac bound
Pasteurella multocida	PMST	2IHK	CMP-3F(equatorial)-Neu5Ac bound
Pasteurella multocida	PMST	2IHZ	CMP-3F(axial)-Neu5Ac and alpha-lactose bound
Pasteurella multocida	sialyltransferase PM0188 with selenomethionine residues	2IY7	CMP-3F(axial)-Neu5Ac bound
Pasteurella multocida	Sialyltransferase PM0188 with selenomethionine residues	2IY8	CMP-3F(axial)-Neu5Ac and alpha-lactose bound
Pasteurella multocida	Sialyltransferase PM0188 with selenomethionine residues	2C83	apo
Pasteurella multocida	Sialyltransferase PM0188 with selenomethionine residues	2C84	CMP bound
Pasteurella multocida	$PMST^{D141N}$	2II6	CMP bound, open conformation
Pasteurella multocida	$PMST^{D141N}$	2IIB	CMP bound, closed conformation
Pasteurella multocida	$PMST^{D141N}$	Not deposited	CMP bound, partially closed conformation
Pasteurella multocida	$PMST^{D141N}$	Not deposited	apo
Pasteurella multocida	$PMST^{M144D}$	3S44	CMP bound
Pasteurella multocida	PMST	4MMP	Apo

Photobacterium sp. JT-ISH-224	JT-ISH-224	2Z4T	CMP and lactose bound
*Pasteurella dagmatis**	$PDST^{P7H, M117H}$	4V39	Apo
*Pasteurella dagmatis**	$PDST^{P7H, M117H}$	4V3C	CMP bound

Table i-2 Publicly available structures of bacterial sialyltransferases with α2,6 activity.

The wildtype sialyltransferases of the bacteria noted with an asterisk are not α2,6 linkage specific.

Expression and Purification of $SK1_{BR}$

DNA encoding residues 252-660 of the full-length SK1 adhesins, termed $SK1_{BR}$, was cloned into the pBG101 vector (Vanderbilt), which encodes an N-terminal His_6-GST affinity tag followed by a cleavage sequence for the 3C precision protease (Figure 1). $SK1_{BR}$ was expressed in *E. coli* BL21(DE3) cells. Cells were grown in LB at 37 °C to an $OD_{600\,nm}$ of ~0.6 and expression induced for 4 hours with 0.5 mM IPTG. The cells were harvested by centrifugation at 9220 *g* for 15 min. The pellets were resuspended in 250 mL lysing buffer (20 mM Tris-HCl, 150 mM NaCl, pH 7.6) supplemented with 2 µg/mL pepstatin, 2 µg/mL leupeptin, 1 µg/mL DNAase, and 1 mg/mL Lysozyme. The cells were lysed by sonication. The lysate was clarified by centrifugation at 38,465 *g* for 1 h, then the supernatant was filtered (0.45 µm) and loaded onto a 5 mL His Trap column. $SK1_{BR}$ was eluted with 75 mM imidazole elution buffer (20 mM Tris-HCl, 150 mM NaCl, 75 mM imidazole, pH 7.6). The N terminal His_6-GST tag was cleaved with 3C precision protease (2 mg/mL $SK1_{BR}$ 20 mM tris, 150 mM NaCl, 75 mM imidazole, pH 7.6). The imidazole was diluted, and the protein was concentrated using a 30K MWCO centrifugal concentrator. Concentrated protein was passed through a His Trap column to remove the cleaved His_6-GST tag. The protein was then purified by Superdex 200 size exclusion column (20 mM Tris-HCl pH 7.6, 150 mM NaCl). Bradford assay was used to determine the final protein concentration.

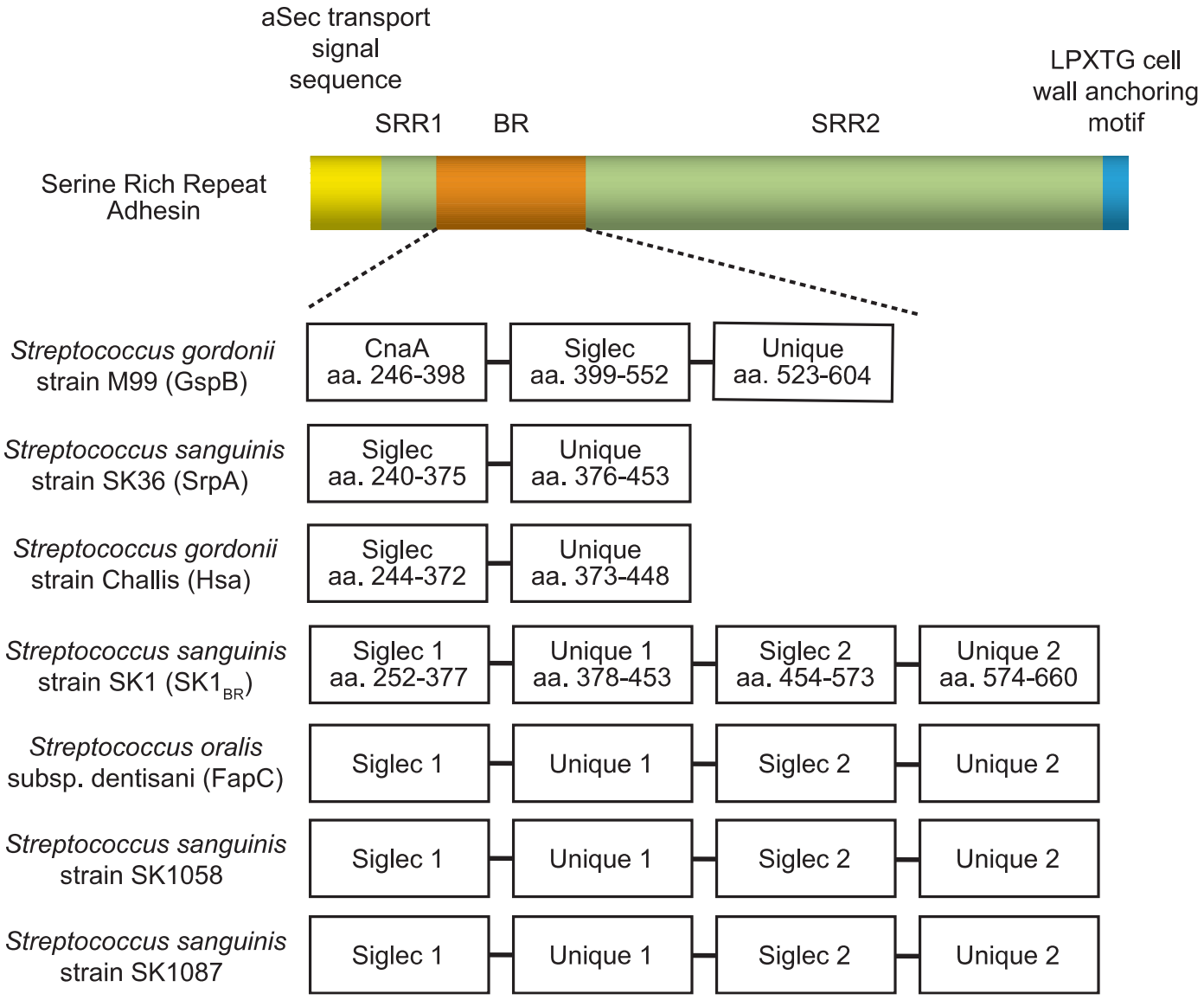

Figure ii-1 Siglec-like serine rich repeat adhesins of Streptococci.

Serine rich repeat proteins follow a conserved functional organization including an N-terminal signal sequence, a short serine rich repeat sequence, a binding region, a second longer serine rich repeat region, and a cell wall anchoring motif. The binding region itself consists of multiple domains and can include domains that contain diverse folds. Depicted here are the organizations of binding regions of Siglec-like SRR adhesins discussed in the text. The bacteria and strain for each binding region are listed, and if the binding region or adhesin has been given a specific name it is listed in parentheses. The SRR adhesins with tandem Siglec-like binding domains are from *S. sanguinis* strains SK160 (WP_080555651.1; 99% identity, 99% similarity), SK1058 (WP_004191732.1; 99% identity, 99% similarity), NCTC 10904 (WP_126436113; 99% identity, 99% similarity), SK1087 (WP_080558715.1; 96% identity, 97% similarity), and BCC39 D8870 (WP_125332456.1; 96% identity, 98% similarity), *S. cristatus* strain 550_SOLI (WP_141640267.1; 94% identity, 96% similarity)(6) and FapC from *S. oralis* subsp. dentisani strain F0392 (25% identity, 39% similarity).[54] Residue numbers for the domains are listed for those binding regions with reported structures.[50, 52, 55]

Synthesis of STa trisaccharide Neu5Acα2–3Galβ1–3GalNAc.

Galβ1–3GalNAc[56] (30 mg, 0.078 mmol), Neu5Ac (37 mg, 0.117 mmol), and CTP (66 mg, 0.117 mmol) were dissolved in a solution containing Tris-HCl buffer (8.0 mL, 100 mM, pH 8.5) and $MgCl_2$ (20 mM). *Neisseria meningitidi*s CMP-sialic acid synthetase[57] (1.0 mg) and *Pasteurella multocida* sialyltransferase 1 M144D[58] (1.5 mg) were then added. The reaction was carried out by incubating the reaction mixture in an incubator shaker at 37 °C for 12 h. The reaction was monitored by thin-layer chromatography (TLC) (EtOAc/MeOH/H_2O/HOAc = 4:2:1:0.1, by volume) with *p*-anisaldehyde sugar staining and mass spectrometry. When an optimal yield was achieved, the same volume (8 mL) of pre-chilled ethanol was added and the mixture was incubated at 4 °C for 30 min. The sample was centrifuged, and the precipitates were removed. The supernatant was concentrated, passed through a BioGel P-2 gel filtration column, and eluted with water to obtain the partially purified product. Further purification was achieved using silica gel chromatography (EtOAc/MeOH/H_2O = 4:2:1, by volume) and a final pass through of a P-2 gel filtration column to produce pure STa (47 mg, 86%). NMR data were in agreement with those reported previously.[59]

Crystallization and Structure Determination

$SK1_{BR}$ (72 mg/mL in 150 mM NaCl, 20 mM tris pH 7.6) was crystallized using the hanging drop vapor diffusion method at 25 °C using a reservoir solution containing 20% (w/v) PEG 3350, 0.2 M $MgSO_4$, 0.01 M $SrCl_2$. Crystals were harvested one week later. To obtain the sialoglycan bound $SK1_{BR}$ structures, crystals of the unliganded $SK1_{BR}$ were grown for one week, removed from the mother liquor, and placed in a new drop containing all of the crystallization components (20% (w/v) PEG 3350, 0.2 M $MgSO_4$, 0.01 M $SrCl_2$) and either 20 mM sTa (synthesized in-house) or 20 mM 3'sLn (Glycotech). The crystals were allowed to incubate with the glycan for one hour at room temperature before harvesting.

Crystals were cryoprotected with 40% (1:1 ethylene glycol: glycerol) and 60% reservoir solution and then cryocooled by plunging in liquid nitrogen. X-ray diffraction data were collected using the Advanced Photon

Source beamline 21-ID-F and a Rayonix MX300 detector. The data were processed using HKL2000.[60] Data collection statistics are in Table ii-1.

ligand	none	sTa	3'sLn
SBGrid Entry	756	754	755
Resolution (Å)	2.00	1.55	2.10
Highest resolution shell (Å)	2.00-2.07	1.55-1.58	2.10-2.18
Data collection			
Beamline	APS 21-ID-F	APS 21-ID-F	APS 21-ID-F
Wavelength (Å)	0.97872	0.97946	0.97946
Space group	$P2_12_12$	$P2_12_12$	$P2_12_12$
Unit cell dimensions			
a (Å)	82.213	83.498	81.549
b (Å)	269.859	271.85	271.063
c (Å)	47.511	47.815	46.849
R_{sym}	0.13 (0.85)	0.067 (0.615)	0.086 (0.514)
R_{pim}	0.045 (0.285)	0.035 (0.402)	0.045 (0.311)
I/σ	25.78 (3.38)	26.6 (1.9)	16.26 (1.51)
Completeness (%)	100 (99.8)	93.1 (74.4)	92.34 (67.30)
Redundancy	10.0 (9.9)	4.3 (2.8)	4.3 (3.3)
$CC_{1/2}$	0.910 (0.706)	0.997 (0.723)	0.999 (0.727)

Table ii-1 Diffraction Data Collection Statistics for the liganded and unliganded SRR binding region from *S. sanguinis* strain SK1.

Values in parentheses are statistics for the highest resolution shell.

The structure of unliganded $SK1_{BR}$ was determined by molecular replacement using the Phaser[61] subroutine in Phenix[62] and the coordinates of the binding region of the unliganded adhesin from *Streptococcus mitis* strain

unliganded NCTC10712 (PDB entry 6EFF)[50] as the search model. All solvent molecules were removed from the coordinates prior to searches. Molecular replacement required two separate steps. The first step used the unliganded Siglec domain (residues 244-370) as the search model and identified four copies of the Siglec domain. The coordinates for the Siglec domain were then fixed, and a second step using the Unique domain as the search model (residues 371-445) identified three copies of the Unique domain. The final Unique domain was manually placed, and the connections between domains were made during refinement.

Structures of the sialoglycan-bound $SK1_{BR}$ were determined by rigid body refinement of the individual domains of the unliganded $SK1_{BR}$ with all solvent molecules removed. Unambiguous electron density for each sialoglycan was visible in the initial maps. The sialoglycans were manually placed into the difference electron density in Coot[63] immediately following structure determination and prior to solvent placement. The ligands were then individually adjusted in Coot[63] prior to refinement in Phenix[62]. To avoid model bias in analysis of the ligand electron density, the ligands were deleted, and the final solvated model was refined in Phenix to produce $|F_o|-|F_c|$ and $2|F_o|-|F_c|$ electron density maps. The $|F_o|-|F_c|$ maps obtained from this protocol are shown in Figure **iii-5** at 3σ. Ligand placements were validated with MotiveValidator[64].

Models were improved using real space refinement in Coot[63] and reciprocal space refinement in Phenix[62]. For unliganded $SK1_{BR}$, 5% of the reflections (totaling 3466 reflections) were randomly selected to use as the free-R and were held separately for the duration of the refinement. For the liganded structures, the equivalent reflections were selected for use as the free-R in the Reflection Editor subroutine in Phenix[62]. The final models of liganded and sialoglycan-bound structures of $SK1_{BR}$ each contain two copies of $SK1_{BR}$ in each asymmetric unit, with a single copy containing all residues of the purified protein, i.e., residues 252-660 of the full-length adhesin. For the sialoglycan-bound structures, one trisaccharide is bound to each Siglec domain, such that a single copy of $SK1_{BR}$ binds two glycans and there are four trisaccharides per asymmetric unit. Refinement statistics and information regarding the content of the models can be found in Table ii-2.

ligand	none	sTa	3sLn
PDB entry	6VS7	6VT2	6VU6
Model Content (per ASU)			
Protein molecules	2	2	2
Glycans	4	4	4
Water molecules	945	1633	317
Ions	14	13	8
Other solvent	13	1	0
Refinement			
R_{cryst}	0.211	0.176	0.225
R_{free}	0.240	0.193	0.248
RMS deviation			
bond lengths (Å)	0.015	0.022	0.018
bond angles (°)	1.25	1.65	1.78
Ramachandran			
Favored (%)	97.05	98.4	97.3
Allowed (%)	2.83	1.6	2.7
Outliers (%)	0.12	0	0
Mean *B* factors			
Protein ($Å^2$)	34.90	15.52	33.57
Glycans ($Å^2$)	-	20.93	49.26

Table ii-2 **Refinement Statistics for the liganded and unliganded SRR binding region from *S. sanguinis* strain SK1**.

Ramachandran statistics were obtained using the MolProbity output of Phenix[62].

Structural Analysis of SK1

$SK1_{Siglec1}$ (residues 252-377) and $SK1_{Siglec2}$ (residues 455-573), were aligned using Pymol (43) which rejected 26 atoms from the alignment and calculated an RMS deviation for the remaining Cα positions of 1.058 Å. $SK1_{Unique1}$ (residues 378-454) and $SK1_{Unique2}$ (residues 574-660) were aligned using the same method and had a

calculated RMS deviation of 0.639 Å. The maximum distances between the loops of $SK1_{Siglec1}$ and $SK1_{Siglec2}$ were measured from the Cα atoms of $SK1^{N229}$ and $SK1^{A497}$ for the CD loop, $SK1^{G342}$ and $SK1^{G543}$ for the EF loop, and $SK1^{D365}$ and $SK1^{K565}$ for the FG loop. The sequence logo (Figure iii-1) was generated using WebLogo 3.7.4[65] from a Clustal Omega[66] multiple sequence alignment. Ligand-protein interactions were analyzed by PDBsumgenerate[67] and LigPlot[68].

Adhesins containing individual Single-Unique Binding Module

Individual binding modules of $SK1_{BR}$, i.e., $SK1_{Siglec1+Unique1}$ (residues 252-455) and $SK1_{Siglec2+Unique2}$ (residues 449-660) were designed based upon the manual evaluation of the end of each folded domain in the crystal structure. $SK1_{Siglec1+Unique1}$ and $SK1_{Siglec2+Unique2}$ were each expressed from pGEX-3X containing an N-terminal GST tag using the protocol detailed above for the full-length $SK1_{BR}$.

Far-western blotting of $SK1_{BR}$

Human blood samples were collected under protocol 11-06207, approved by the UCSF Institutional Review Board and these studies and protocols abide by the Declaration of Helsinki principles. The binding of GST-tagged $SK1_{BR}$, GST-tagged $SK1_{Siglec1+Unique1}$, and GST-tagged $SK1_{Siglec2+Unique2}$ to plasma, platelet and salivary glycoproteins was performed as described.[69, 70]

$SK1_{BR}$ Binding to immobilized platelets.

Human blood samples were collected under protocol 11-06207, approved by the UCSF Institutional Review Board and these studies and protocols abide by the Declaration of Helsinki principles. Binding to formalin-fixed human platelet monolayers was performed as described.[69] In brief, human platelets were freshly prepared, washed and immobilized in 96-well plates. After blocking non-specific binding with 1x Blocking Reagent (Roche) in Dulbecco's phosphate-buffered saline (DPBS), the blocking solution was replaced with 50 µL GST-tagged $SK1_{BR}$ and split binding module proteins diluted to the indicated concentrations into 1x blocking solution, and plates were incubated for 1 h at room temperature (~22 °C) with vigorous rocking. Wells were washed three times with 100 µl DPBS, and the bound GST-tagged proteins were detected by using a rabbit anti-GST (Life Technologies)

diluted 1:500 in 1x blocking solution, followed by a peroxidase-conjugated goat anti-rabbit (Sigma) diluted 1:5000 in DPBS, along with the chromogenic substrate *o*-phenylenediamine dihydrochloride (Sigma).

Binding of biotin-glycans to immobilized SK1 binding regions.

The binding of polyvalent biotinylated glycans (Glycotech) to GST-tagged $SK1_{BR}$ and split binding module proteins (Figure ii-2) immobilized in 96-well plates was performed as described.[69] In brief, wells were coated with the GST-tagged proteins (500 nM in DPBS) by incubating overnight at 4 °C. Wells were washed twice with DPBS, and biotinylated glycans that had been diluted to the indicated concentrations in 1x Blocking Reagent (Roche) in DPBS were added. Plates were incubated for 90 min at room temperature (~22 °C) with vigorous rocking, and the unbound glycans were removed by washing three times with DPBS. The bound glycans were detected by using peroxidase-conjugated streptavidin (Sigma), followed by the chromogenic substrate *o*-phenylenediamine dihydrochloride (Sigma).

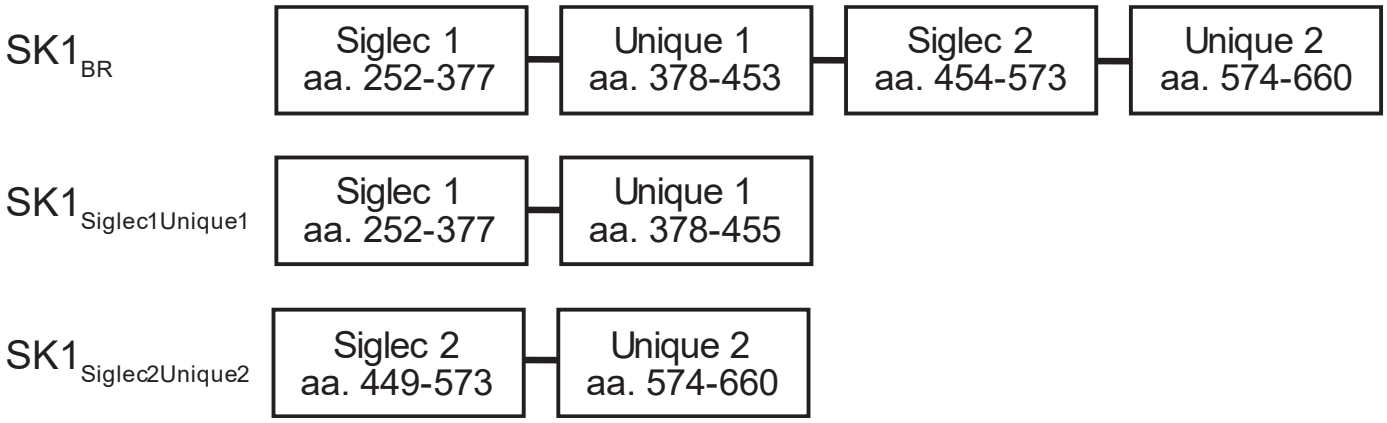

Figure ii-2 Graphical representation of $SK1_{BR}$ split domain proteins.

The residues of the full length adhesin included in the expressed protein are shown. Information regarding the affinity tags (not shown) is included in the Experimental Procedures section.

Sequence analysis.

SLBR sequences were aligned using the MUSCLE[71] subroutine in Geneious Pro 11.1.4[72]. The JTT-G evolution model was selected using the ProtTest server[73], and the phylogenetic tree was built using the MrBayes[74] subroutine.

Cloning, expression, and purification for crystallization.

DNA encoding all SLBRs except $SLBR_{Hsa}$ were cloned into the pBG101 vector (Vanderbilt University), which encodes an N-terminal His_6-GST tag cleavable with 3C protease. $SLBR_{Hsa}$ was cloned into the pSV278 vector (Vanderbilt University), which encodes a thrombin-cleavable His_6-maltose binding protein (MBP) tag. Proteins were expressed in *E. coli* BL21 (*DE3*) with 50 µg/ml kanamycin at 37 °C. Expression was induced with 0.5-1 mM IPTG at 24 °C for 3-7 hrs. Cells were harvested by centrifugation at 5,000 × *g* for 15 min and stored at –20 °C before purification.

Cells were resuspended in 20-50 mM Tris-HCl, pH 7.5, 150-200 mM NaCl, 1mM EDTA, 1 mM PMSF, 2 µg/ml Leupeptin, 2 µg/ml Pepstatin then disrupted by sonication. Lysate was clarified by centrifugation at 38500 × *g* for 35-60 min. Tagged fusion proteins were purified using a Glutathione Sepharose 4B column eluted with 30 mM GSH in 50 mM Tris-HCl, pH 8.0, aNi^{2+} affinity chromatography eluted with 20 mM Tris-HCl, 150 mM NaCl, 250 mM imidazole, pH 7.6, or an MBP-Trap column eluted in 10 mM maltose. Affinity tags were cleaved with 1 U of protease per mg of protein overnight at 4 °C. Protein was separated from the cleaved affinity tag by passing over the relevant affinity column. Protein aggregates were removed using either a Superose-12 column in 50 mM Tris-HCl pH 7.6 and 150 mM NaCl or a Superdex 200 increase 10/30 GL column equilibrated in 20 mM Tris-HCl pH 7.6 or in 20 mM Tris-HCl pH 7.5 and 200 mM NaCl.

Crystallization, data collection, and structure determination.

Crystallizations were performed at room temperature (~23 °C) using the conditions in **Supplementary Table 5**. The $SLBR_{GspB}$-sTa structure used crystals where the ligand was introduced by cocrystallization, and the $SLBR_{Hsa}$-ligand structures used crystals where the ligand was introduced by soaking. Data collection and refinement statistics are listed in **Supplementary Tables 1, 2, and 3**. Structures were determined by molecular replacement using the Phaser[75] subroutine of Phenix 1.18.2[76] using the starting models listed in **Supplementary Table 5**.

All models were improved with iterative rounds of model building in Coot 0.9[77] and refinement in Phenix 1.18.2[76]. Riding hydrogens were included at resolutions better than 1.4 Å. For sialoglycan-bound $SLBR_{Hsa}$, the crystals were isomorphous with unliganded crystals and R_{free} reflections were selected as identical. Ligand occupancies were held at 1.0 during refinement. Representative electron density maps for each structure can be found Figure 4 and Supplementary Figure 22.

Sialoglycan binding assays.

DNA encoding wild-type and variant SLBRs were cloned into pGEX-3X. Individual GST-SLBR fusions were expressed and purified using glutathione-sepharose, and the binding of biotinylated glycans to immobilized GST-SLBRs was performed as described previously[69]. Anti-GST antibody was from Invitrogen (A5800) and peroxidase-conjugated goat anti-rabbit IgG was from Sigma (A0545). The number of replicates of each data point are in each figure legend. Replicates are independent replicates from separately prepared samples.

Far western and lectin blotting of human salivary and plasma proteins.

Far-western blotting of human saliva and plasma proteins using the indicated GST-SLBRs (15 nM) as probes was performed as described [69, 70]. Plasma was purchased from Innovative Research (Novi, MI). De-identified samples of SMSL saliva were provided by S. Fisher (UCSF). Because these specimens were de-identified prior to gifting, our use of this material was exempt from approval by the UCSF Institutional Review Board. Anti-GST antibody was from Invitrogen (A5800) and peroxidase-conjugated goat anti-rabbit IgG was from Sigma (A0545).

MUC7 affinity capture and O-glycan profiling.

A combination of GST-$SLBR_{Hsa}$ and GST-$SLBR_{UB10712}$ immobilized on magnetic glutathione beads was used to capture the total sialylated MUC7 from 300 µl of SMSL saliva. The resin-bound GST-SLBRs and affinity-captured MUC7 were co-eluted into LDS sample buffer (Invitrogen) supplemented with dithiothreitol (100 mM final concentration), separated by electrophoresis in 4-12% polyacrylamide gradient gels, and then stained with

SimplyBlue SafeStain (Invitrogen). The captured proteins, which ranged from 120-160 kDa, were excised from the gel. A portion of the sample was submitted for protein identification by nanoflow LC-MS/MS of tryptic digests (MSBioworks), which confirmed MUC7 as the major component. A second portion of the excised gel slices was minced, treated by four cycles of rinsing with 100mM ammonium bicarbonate and dehydration in 100% acetonitrile, and then dried to completion in a vacuum evaporator. The gel pieces were immersed in a mixture of 100 mM NaOH and 1M $NaBH_4$ and incubated at 45 °C for 18 h to release the O-glycans. The supernatant was collected and placed on ice, and the remaining gel pieces were washed with water and sonicated for 30 min to extract the remaining O-glycans. The initial and secondary extracts were combined and acidified to pH 4-6 by drop-wise addition of 10% acetic acid. The *O*-glycan samples were then enriched using porous graphitized carbon cartridges (Agilent, Santa Clara, CA) and dried prior to analysis by mass spectrometry. Glycan samples were analyzed on an Agilent 6520 Accurate Mass Q-TOF LC/MS equipped with a porous graphitic carbon microfluidic chip. A binary gradient consisting of (A) 0.1% formic acid in 3% acetonitrile, and (B) 1% formic acid in 89% acetonitrile was used to separate the glycans at a flow rate of 0.3 µl/min. Data were processed with Agilent MassHunter B.07 software, using the Find by Molecular Feature algorithm with an in-house library of O-glycan masses and chemical formulae to identify and quantitate the O-glycan signals.

In silico structure predictions and MD analyses.

The model of $SLRB_{SK678}^{Hsa\text{-}loops}$ was calculated using MOE[78]. For MD of $SLBR_{Hsa}$, $SLBR_{GspB}$, $SLBR_{SK678}$, and $SLBR_{SK678}^{Hsa\text{-}loops}$ each set of PDB coordinates was solvated in a 10 Å octahedral box of TIP3P[79] water. The Amber16 ff14SB[80] force field was used for the protein. In the first step of the MD simulation, the backbone and side chains of the protein were restrained using 500 kcal mol^{-1} $Å^{-2}$ harmonic potentials while the system was energy minimized for 500 steps of steepest descent[81] and the conjugate gradient method[82]. Restraints were removed and 1000 steps of steepest descent minimization were performed followed by 1500 steps of conjugate gradient. The system was then subjected to MD at 300 K with the backbone and side chains restrained using 10

kcal mol^{-1} $Å^{-2}$ harmonic potentials for 1000 steps. Bonds were constrained using SHAKE[83]. MD (200 ns) was performed at 300 K in the NPT ensemble and a 2-fs time step. Probability distribution analyses and RMSF calculations were performed on 200 ns of 3 independent runs. Analyses were performed using the cpptraj and pytraj[84] modules of AMBER16. The last snapshot from 20-ns trajectory was used for mapping the interaction between the glycans and $SLRB_{SK678}$ or $SLBR_{SK678}^{Hsa\text{-}loops}$.

Expression and Purification of catalytically inactive α2,6-linkage specific Pasteurella multocida sialyltransferase D141N

Residues 25–412 of Pasteurella multocida α2,6 linkage specific sialyltransferase were expressed with a 3C protease cleavable N-terminal GST and 6x His tag in E. Coli BL21 DE3 Gold cells vector PBG101 (Vanderbilt). Cells were grown until they reached OD_{600nm} of 0.4. The cultures were cooled to 18°C for an hour and expression was induced with 0.1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) for 16 hours. The cells were pelleted by centrifugation at 8,000g at 4°C. Cells were resuspended in 20 mM Tris HCl pH 7.5, 500 mM NaCl and lysed by at 20,000 psi. The lysate was clarified by centrifugation at 18,000g for 1 hour and the supernatant was filtered through a 0.45μ filter. The supernatant was loaded onto a Histrap column and eluted with 20 mM Tris pH 7.5, 100 mM NaCl, 500 mM imidazole. The buffer was exchanged for 20 mM Tris HCl pH 7.5, 100 mM NaCl by gel filtration on a Superdex 200 increase column.

Crystallization, Data Analysis, and Model Refinement of apo and CMP-bound $PMST^{D141N}$

$PMST^{D141N}$ was crystallized similarly to published conditions. Briefly, $PMST^{D141N}$ was crystallized by hanging drop. $PMST^{D141N}$ (10mg/mL) in 20 mM Tris pH 7.5, was combined with reservoir buffer (100mM HEPES pH 7.5, 100mM NaCl, 23% PEG3350, 0.4% TritonX-100) with or without 5mM CMP in a 1:1 ratio to form 3 uL drops. Crystals grew in a week and cryo-cooled by plunging in liquid nitrogen. Crystals were diffracted under a stream of nitrogen at 100K. X-ray diffraction data for apo $PMST^{D141N}$ were collected using the Advanced

Photon Source beamline 21-ID-F and a Rayonix MX300 detector. X-ray diffraction data for CMP-bound PMST^{D141N} were collected using the Advanced Photon Source beamline 21-ID-D and a Dectris Eiger 9M detector. All data were processed using HKL2000[60] and ccp4i[85]. Molecular replacement was performed with PHASER[61] using CMP bound open and closed state PMST^{D141N} models (PDB 2ii6 and 2iib) for apo and CMP-bound PMST^{D141N} structures respectively. Models were refined using Phenix.[62, 76]

Glycan Array analysis of PMST^{D141N} selectivity

Data were obtained from the National Center for Functional Glycomics (NCFG) funded by NIGMS-GM62116. Methods can be found at https://ncfg.hms.harvard.edu/protocols/glycan-binding-assay-fusion-or-epitope-tagged-protein. Briefly, 50 ug PMST^{D141N} (TSM binding buffer: 20 mM Tris-HCl pH 7.4, 150 mM NaCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 0.05% Tween 20, 1% bovine serum albumin) was incubated on a version 5.5 CFG glycan printed array slide for one hour at 23°C in a humidified chamber. The slide was washed with wash buffer (TSM wash buffer: 20 mM Tris-HCl pH 7.4, 150 mM NaCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 0.05% Tween 20) and incubated with rabbit anti-glutathione S transferase conjugated with Alexa Fluor™ 488 (Invitrogen A-11131) and incubate for one hour at room temperature in a humidified chamber. After incubating the antibody, wash the slide three times with TSM wash buffer, TSM (TSM binding buffer: 20 mM Tris-HCl pH 7.4, 150 mM NaCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$), and deionized water respectively. Remove excess water with a stream of nitrogen and scan the slide at 518 nm.

iii. TANDEM SIALOGLYCAN-BINDING MODULES IN A STREPTOCOCCUS SANGUINIS SERINE-RICH REPEAT ADHESIN CREATE TARGET DEPENDENT AVIDITY EFFECTS

Haley E. Stubbs, Barbara A. Bensing, Izumi Yamakawa, Pankaj Sharma, Hai Yu, Xi Chen, Paul M. Sullam, T.M. Iverson. Tandem sialoglycan-binding modules in a Streptococcus sanguinis serine-rich repeat adhesin create target dependent avidity effects. **J Biol Chem**. 2020 Oct 23;295(43):14737-14749.

I expressed, purified, and crystallized $SK1_{BR}$ with the help of Izumi Yamkawa and Dr. Pankaj Sharma. I analyzed the data for unliganded $SK1_{BR}$, refined unliganded, sTa-bound, and 3'sLn-bound $SK1_{BR}$ and made all structural figures with the help of Pankaj Sharma. I wrote the text with guidance and editing from Dr. T. M. Iverson.

Siglec-like domains of streptococcal serine-rich repeat (SRR) adhesins recognize sialylated glycans on human salivary, platelet, and plasma glycoproteins via a "YTRY" sequence motif. The SRR adhesin from *Streptococcus sanguinis* strain SK1 has tandem sialoglycan-binding domains and has previously been shown to bind sialoglycans with high affinity. However, both domains contain substitutions within the canonical "YTRY" motif, making it unclear how they interact with host receptors. To identify how the *S. sanguinis* strain SK1 SRR adhesin affects interactions with sialylated glycans and glycoproteins, we determined high-resolution crystal structures of the binding domains alone and with purified trisaccharides. These structural studies identify that the ligands still bind at the non-canonical binding motif, but with fewer hydrogen-bonding interactions to the protein than is observed in structures of other Siglec-like adhesins. Complementary biochemical studies

identify that each of the two binding domains has a different selectivity profile. Interestingly, the binding of SK1 to platelets and plasma glycoproteins identifies that the interaction to some host targets is dominated by the contribution of one binding domain, while the binding to other host receptors is mediated by both binding domains. These results provide insight into outstanding questions concerning the roles of tandem domains in targeting host receptors and suggest mechanisms for how pathogens can adapt to the availability of a range of related but non-identical host receptors. They further suggest that the definition of the "YTRY" motif should be changed to ϕTRX, a more rigorous description of this sialic acid-recognition motif given recent findings.

Introduction

The serine-rich repeat (SRR) adhesins are a family of bacterial cell-surface glycoproteins containing two sequence motifs where serine constitutes approximately 50% of the sequence (Figure iii-1**,** Figure ii-1). These follow a modular architecture that initiates with an atypical signal peptide, a short N-terminal serine-rich region, a ligand-binding region (frequently termed "adhesin$_{BR}$" or "strain$_{BR}$", *e.g.* the binding region from *S. sanguinis* strain SK1 is termed SK1$_{BR}$), a second serine-rich repeat region that varies in length between several hundred and several thousand amino acids with serine as every other residue, *e.g.* …SVSASTSASTSASTSAS…, and a cell wall anchoring motif. Fiber diffraction studies suggest that these repeat regions form a spring-like linker that tethers the host binding region to the bacterium.[86]

The SRR adhesins are expressed by a variety of Gram-positive commensal and pathogenic bacteria and are broadly distributed.[87] A survey of NCBI GenBank identified over a thousand sequences that may belong to this family. Indeed, all sequenced strains of *Streptococcus gordonii* and *Streptococcus sanguinis* encode serine-rich repeat adhesins[88] and homologs have been found in strains of *S. oralis* and *S. mitis*, as well as other oral streptococci.[54, 69, 87, 89] One known functional role of these SRR adhesins is to mediate attachment to protein or glycoprotein receptors, which allows for adherence to host tissues. Accordingly, SRR adhesins have been linked to a variety of infections, including endocarditis, meningitis, and pneumonia.[90, 91, 92, 93, 94, 95]

Despite the conserved functional organization (Figure iii-1), the ligand binding regions are highly diverse with species-specific trends in the binding region type. Some of the ligand binding regions of SRR adhesins target glycan structures.[96] For example, the SRR adhesins of *S. gordonii* and *S. sanguinis* bind *O*-linked sialoglycans displayed on mucin-like proteins including salivary glycoprotein MUC7 and platelet glycoprotein GPIb.[70, 87, 97] Binding to MUC7 may facilitate oral colonization, while interaction with platelet GPIb can allow streptococci to establish endocardial infections.[52, 97]

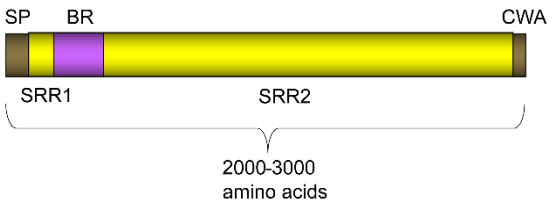

Figure iii-1 General organization of SRR adhesin proteins.

SRR adhesins initiate with a ~90 amino acid N-terminal signal peptide (SP) that facilitates trafficking to a specialized glycoprotein transporter known as the accessory Sec system. The serine-rich repeat regions (SRR1 and SRR2) are extensively O-glycosylated in the bacterial cytoplasm prior to transport. The ligand binding region (BR) varies depending upon the organism and contains structural modules that are highly diverse in sequence, fold, and function, which may provide binding specificity for different bacterial strains (6). The C-terminal cell wall anchor (CWA) includes an LPXTG sequence motif that covalently links the cell wall peptidoglycan.

Interaction of the *S. gordonii* and *S. sanguinis* SRR adhesins with host sialoglycan structures relies upon a domain within the binding region related in fold to mammalian sialic acid-binding immunoglobulin-like lectins (Siglecs); indeed both are organized around a V-set Ig fold.[49, 50, 52, 69] These "Siglec-like" SRR adhesins always contain a second domain immediately following the Siglec domain.[52] Termed the Unique domain, this C-terminal region has no counterpart in mammalian Siglecs, and its function remains unknown.

Despite the conserved fold in the Siglec domain, the binding location for sialoglycans differs between the bacterial Siglec-like adhesins and mammalian Siglecs. The streptococcal Siglec-like adhesins hydrogen-bond with sialic acid via a semi-conserved “YTRY” sequence motif on the F strand of the V-set Ig fold (Figure iii-2).[49, 52, 55, 69] Of these, the first Tyr residue contributes only backbone interactions to the ligand. Here, the aromatic sidechain faces away from the binding site and is involved in packing interactions that likely contribute to the correct presentation of central “Thr-Arg”. The Thr-Arg makes multiple key side-chain hydrogen-bonding contacts to the sialic acid of host sialoglycans and therefore the sequence of these central residues appears to be the most important for binding.[49, 52] Prior mutagenesis of either the Thr or Arg in characterized Siglec-like adhesins substantially reduces binding to defined, synthetic sialoglycans and to platelets.[55, 69, 98] Moreover, isogenic strains

of streptococci containing mutations in the YTRY motif exhibit reduced virulence in an animal model.[99] The final Tyr of the motif contributes a single hydrogen-bond to the sub-terminal galactose of α2-3-sialoglycans and is therefore not involved in sialic acid recognition but may contribute to overall binding affinity of sialoglycans.[52]

Recent structural and engineering studies revealed that sialoglycan binding and selectivity is also affected by three adjacent loops of high sequence diversity.[50] Using nomenclature from the V-set Ig fold identifies these as the CD loop, the EF loop, and the FG loop.[52] Amino acid side chains in these loops directly hydrogen-bond with sialoglycan ligands[49, 52, 55] and the sequence diversity of these loops is proposed as a major determinant of sialoglycan selectivity in the Siglec-like adhesins. As a result, these have been termed "selectivity loops".[50]

The only structurally characterized Siglec-like binding region that differs somewhat in the topology of its binding pocket is found in SrpA from *S. sanguinis* strain SK36. In $SrpA_{BR}$, the YTRY sequence is a non-canonical FTRT but retains the central "Thr-Arg" important for ligand binding.[49, 55] In addition, $SrpA_{BR}$ lacks an appreciable FG selectivity loop. Notably, $SrpA_{BR}$ contains a second Arg residue outside of the canonical binding sequence motif (Figure iii-2) that cooperates with the non-canonical FTRT sequence to promote sialoglycan binding.[49, 55] This residue is not highly conserved in the Siglec-like binding regions (Figure iii-2) and is located too far from the FTRT motif to interact with a bound trisaccharide. However, structures of $SrpA_{BR}$ show that the binding pocket is contiguous with this distal arginine both because of the presence of a Thr versus Tyr in the fourth position of the YTRY motif and because of the absence of the FG loop. These alterations extend the glycan binding site, which may allow the accommodation of either two oriented trisaccharides or larger, branched sialoglycans. Indeed, a disialylated hexasaccharide has been modeled into this site with the distal arginine binding to the second sialic acid of this significantly larger and disialylated ligand.[49]

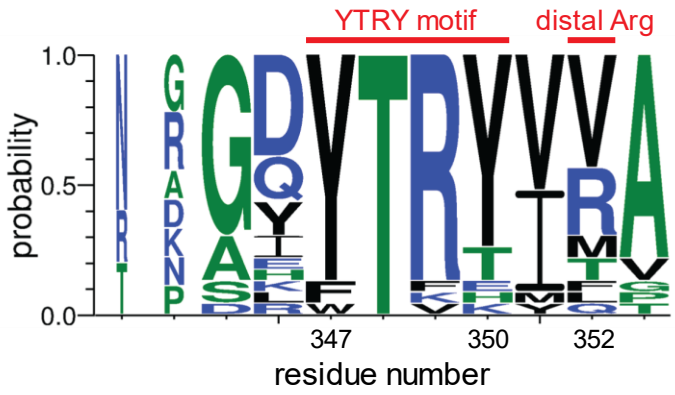

Figure iii-2 The Sialoglycan binding motif of Siglec-like SRR adhesins.

The conservation of residues is indicated by letter size with the larger letters representing a more strongly conserved residue. The positions of the YTRY motif (positions 347-350) and the distal Arg (position 352) are notated above with a red line above the letters. The numbering reflects the residue positions within $SK1_{Siglec1}$. The letters colored blue, green, and black indicate charged, nonpolar, and polar residues, respectively. The adhesins included in the alignment are WP_125444382.1 **from *Streptococcus gordonii* strain M99,** WP_046165954.1 from *Streptococcus gordonii* strain 72-40, WP_080889728.1 from *Streptococcus gordonii* strain G9B, WP_046165954.1 from *Streptococcus sp.* Strain 1236FAA, WP_009659981.1 from *Streptococcus sp.* Strain AS14, WP_002906900.1 from *Streptococcus sanguinis* strain SK115, WP_125439128.1 from *Streptococcus sanguinis* strain SK150, WP_125444035.1 from *Streptococcus sanguinis* strain SK678, WP_081102781.1 from *Streptococcus gordonii* strain Challis, WP_045635027.1 from *Streptococcus gordonii* strain UB10712, WP_080555651.1 from *Streptococcus sanguinis* strain SK1, WP_011836739.1 from *Streptococcus sanguinis* strain SK36, WP_080555852.1 from *Streptococcus sanguinis* strain SK408, WP_000466180.1 from *Streptococcus sanguinis* strain SK140, WP_046165954.1 from *Streptococcus sanguinis* strain PS478, WP_087941957.1 from *Streptococcus sanguinis* strain SK1056, WP_080557024.1 from *Streptococcus sanguinis* strain SK330, WP_080560819.1 from *Streptococcus sanguinis* strain SK355, WP_080555460.1 from *Streptococcus sanguinis* strain SK405, WP_061600538.1 from *Streptococcus gordonii* strain SK49, and WP_000466181.1 from *Streptococcus oralis* strain SF100. Although previously termed the "YTRY" motif, this binding sequence motif is formally defined as ϕTRX, where ϕ represents W, F or Y and X represents Y, T, E, H, or K.

The binding region from *Streptococcus sanguinis* strain SK1 (residues 252-660 and termed $SK1_{BR}$) differs from structurally-characterized Siglec-containing binding regions in two ways.[69] First, it contains two copies of the Siglec and Unique domains in tandem ($SK1_{Siglec1}$-$SK1_{Unique1}$-$SK1_{Siglec2}$-$SK1_{Unique2}$) (**Fig S1**) with 39%/56% sequence identity/similarity between the two Siglec domains and 44%/50% between the two Unique domains. Tandem domains are rarely observed in sequences of the Siglec-like adhesins. Only eight other tandem domain Siglec-like binding regions are identifiable in GenBank. Seven of the eight SRR adhesins have ≥94% identity and

≥96% similarity to $SK1_{BR}$; these adhesins are from various strains of *S. sanguinis* and one from *S. cristatus* (**Fig S1**). The other adhesin containing a tandem domain is FapC from *Streptococcus oralis* subsp. *dentisani* strain F0392 (**Fig S1**)[54], where the $FapC_{BR}$ binding region exhibits 25% identity and 39% similarity to $SK1_{BR}$. This evolutionary relatedness is consistent with evidence that $FapC_{BR}$ may bind sialoglycans and may be important for oral colonization.[54] The functional implications for tandem domains have not been explained in the literature.

The second way that $SK1_{BR}$ differs from other structurally-characterized Siglec-like binding regions is that each of the putative Siglec domains of $SK1_{BR}$ contains a non-canonical YTRY sialic acid-binding sequence motif that lacks the central "Thr-Arg" deemed to be critical for binding in other Siglec-like adhesins.[69] This motif in $SK1_{Siglec1}$ is YTKY and the motif in $SK1_{Siglec2}$ is YTFK.[69] Although it has been shown that $SK1_{BR}$ binds sialoglycans[69], it was unknown if both Siglec domains could bind sialoglycans. If so, what is the relative contribution of each to binding? How do the non-canonical YTRY sequence motifs, particularly in $SK1_{Siglec2}$, maintain the interaction with host receptors?

Here, we present crystal structures of unliganded and sialoglycan-bound $SK1_{BR}$, which show that both $SK1_{Siglec1}$ and $SK1_{Siglec2}$ adopt V-set Ig folds and that both domains interact with sialoglycan ligands at the non-canonical YTRY motifs. We validate these interactions using binding studies of isolated $SK1_{Siglec1+Unique1}$ and $SK1_{Siglec2+Unique2}$ and demonstrate that each domain has a distinct selectivity profile for synthetic sialoglycans and glycoprotein ligands. The tandem domains allow increased binding to host salivary glycoprotein, MUC7, via an avidity effect, possibly by binding simultaneously to two oriented, large glycans. In contrast, platelet binding mainly occurs via $SK1_{Siglec2}$ and binding to the plasma glycoprotein PRG4 (also called lubricin) mainly occurs via $SK1_{Siglec1}$. Taken together, these findings support a mechanism of host interaction where the tandem domains of the *S. sanguinis* strain SK1 adhesin interact most strongly with a patch of oriented large glycans, and that each individual domain differently impacts binding to sialoglycoprotein targets. This adhesin architecture therefore allows for increased flexibility and breadth in the host receptors that are recognized.

Results

Structure of S. sanguinis $SK1_{BR}$

To develop hypotheses for how $SK1_{BR}$ binds to sialoglycans, we began by determining its X-ray crystal structure using molecular replacement methods (Table ii-1**,** Table ii-2). The tandem repeats of unliganded $SK1_{BR}$ fold independently and slight angles between the domains yield an elongated and overall arc-shape. The limits of the domains can be clearly distinguished as $SK1_{Siglec1}$ (residues 252-377), $SK1_{Unique1}$ (residues 378-453), $SK1_{Siglec2}$ (residues 454-573), and $SK1_{Unique2}$ (residues 574-660).

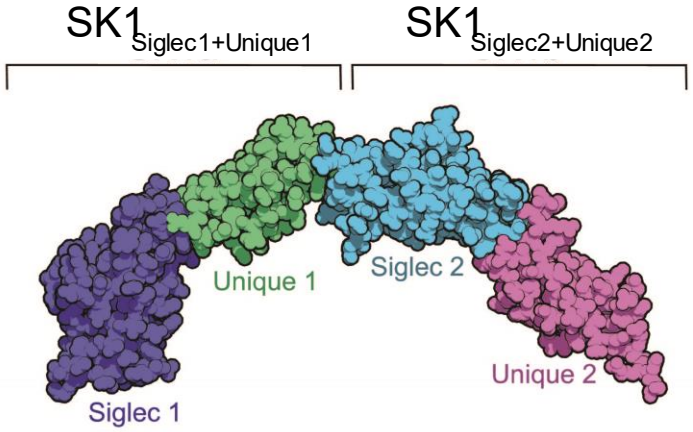

Figure iii-3 Structure of the binding region of unliganded SK1.

$SK1_{BR}$ has four domains in the order $SK1_{Siglec1}$, $SK1_{Unique1}$, $SK1_{Siglec2}$, $SK1_{Unique2}$ depicted from left to right and colored by domain. The domain repeats, named $SK1_{Siglec1+Unique1}$ and $SK1_{Siglec2+Unique2}$ are homologous but not identical.

As is anticipated from the amino acid sequence conservation, the individual Siglec domains and Unique domains exhibit structural similarity, with an RMS deviations in Cα position of 1.058 Å between $SK1_{Siglec1}$ and $SK1_{Siglec2}$ and an RMS deviations in Cα position of 0.639 Å between $SK1_{Unique1}$ and $SK1_{Unique2}$ (Figure iii-4). To evaluate the basis for the higher overall RMS deviations in Cα position of the Siglec domains, we overlaid $SK1_{Siglec1}$ and $SK1_{Siglec2}$ and identified whether this was a global difference or whether the structural difference was localized to specific regions. We identified disproportionately large structural deviations in the CD, EF, and FG selectivity loops that surround the putative sialoglycan-binding pockets (Figure iii-4**a**), which contain

different numbers of residues, with the FG loop of $SK1_{Siglec2}$ being so short that it is effectively absent. The maximal displacement of these loops in overlays is 10.9 Å for the CD loop, 3.2 Å for the EF loop, and 8.4 Å for the FG loop. In the CD loop, this also manifests as a difference in secondary structure where the CD loop of $SK1_{Siglec1}$ lacks secondary structure while the CD loop of $SK1_{Siglec2}$ folds into an α-helix. In addition, the CD loop of $SK1_{Siglec2}$ is displaced from the binding pocket when compared to the CD loop of $SK1_{Siglec1}$. Together, these differences in loop length and structure result in a larger and more open binding site as compared to $SK1_{Siglec1}$. This large binding pocket of $SK1_{Siglec2}$ is somewhat reminiscent of the binding pocket in the SrpA adhesin from *S. sanguinis*, which also lacks the FG loop.[49, 52, 70]

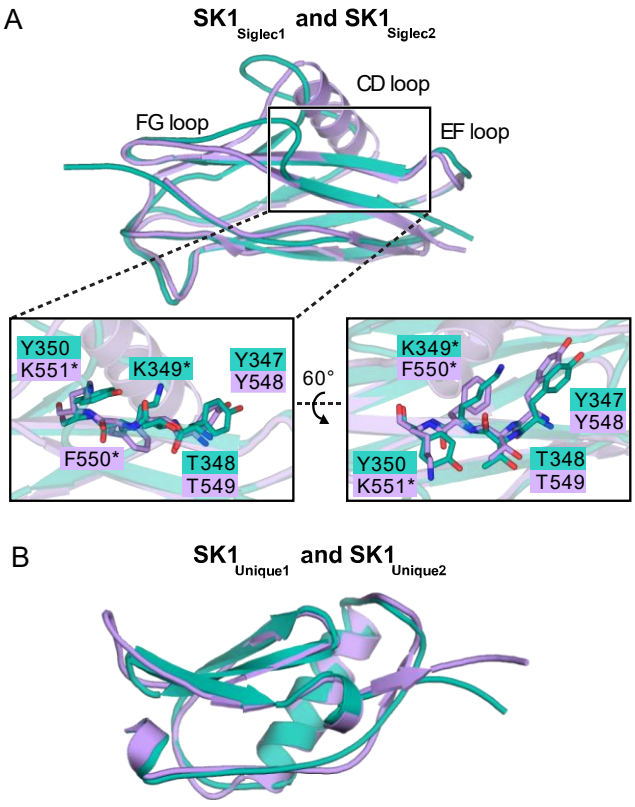

Figure iii-4 Structural comparison of the tandem Siglec and Unique domains.

A) and B) Cartoon diagram of unliganded $SK1_{BR}$. $SK1_{Siglec1}$ and $SK1_{Unique1}$ are colored in teal, and the $SK1_{Siglec2}$ and $SK1_{Unique2}$ are colored in lavender. A) Residues Ninety-eight atoms were aligned and twenty-six were rejected after five cycles to output an RMS deviation of 1.058 Å. The noncanonical binding motifs of $SK1_{Siglec1}$ and $SK1_{Siglec2}$ are shown in sticks, and residues that deviate from the canonical YTRY motif definition are noted

with an asterisk. B) Residues… Seventy-two atoms were aligned and 3 were rejected after two cycles to output an RMS deviation of 0.639 Å.

Structures of S. sanguinis $SK1_{BR}$ bound to sialoglycans.

Prior glycan array analysis identified that $SK1_{BR}$ can bind to a broad range of defined, synthetic sialoglycan ligands. It was not clear, however, whether $SK1_{Siglec1}$ and $SK1_{Siglec2}$ are both important for these interactions, or how the non-canonical YTRY motifs support glycan binding. We therefore determined cocrystal structures of $SK1_{BR}$ soaked with either 10 mM sialyl T-antigen (sTa), a "core 1" glycan that can be conjugated to Ser or Thr residues of glycoproteins, or 10 mM 3'-sialyl-*N*-acetyllactosamine (3'sLn), a trisaccharide that can be a component of larger, branched glycans. For both sTa and 3'sLn, we observed the appearance of unambiguous electron density adjacent to the non-canonical YTRY motif of both $SK1_{Siglec1}$ and $SK1_{Siglec2}$. We were able to model sTa and 3'sLn with confidence into this electron density, and the hydrogen bond networks observed are consistent with specific binding (Figure **iii-5a-d,** Figure **iii-6a-d**). Thus, $SK1_{Siglec1}$ and $SK1_{Siglec2}$ are both capable of binding sialoglycan ligands and can do so simultaneously. Local conformational changes observed upon binding were slight. When each domain is individually aligned to the corresponding unliganded domain, the RMS deviations in the Cα positions is < 0.3 Å (**Table iii-1**). The biological significance of conformational changes of this magnitude cannot be determined.

RMS deviation for Cα atoms of liganded and unliganded domains		
	sTa	**3'sLn**
$SK1_{Siglec1}$	0.237 Å	0.252 Å
$SK1_{Unique1}$	0.214 Å	0.196 Å
$SK1_{Siglec2}$	0.194 Å	0.191 Å
$SK1_{Unique2}$	0.232 Å	0.160 Å

Table iii-1 Supplementary Table 1 RMS deviation values for liganded and unliganded domain alignments. Each Siglec and Unique domain from the liganded structures were individually aligned to the same domain in the unliganded structure in Pymol.

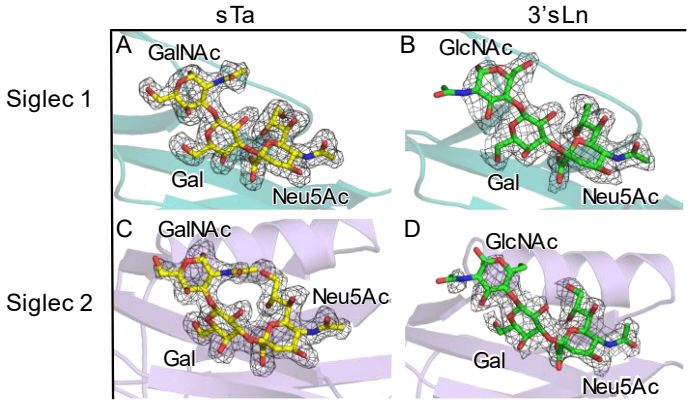

Figure iii-5 **Ligand electron density.**

Cartoon representations of $SK1_{Siglec1}$ and $SK1_{Siglec2}$ are shown in teal and lavender respectively. sTa and 3'sLn are shown in yellow and green sticks respectively. Oxygen atoms are colored in red and nitrogen atoms in blue. Ligands were manually placed in Coot after refinement of the protein and prior to solvent placement. The ligands were then refined with rigid-body and real space refinement in Coot prior to solvent placement and final structure refinements. $|F_o|-|F_c|$ electron density maps were calculated from coordinates that had been refined in Phenix[62] for three rounds after the removal of the sialoglycans from the model. Maps are contoured at 3σ and are shown in dark grey mesh.

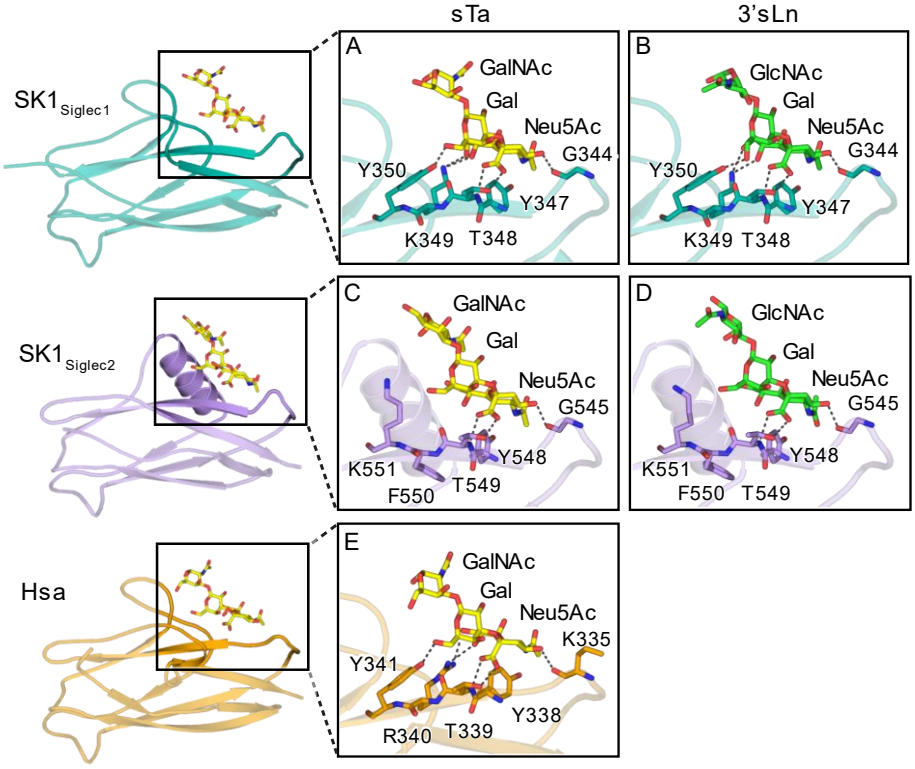

Figure iii-6 **SK1 interactions with ligands.**

The structures of the Siglec domains of $SK1_{Siglec1+Unique1}$, $SK1_{Siglec2+Unique2}$, and Hsa (PDB entry 6EFD) (18) are shown in cartoon in teal, lavender, and orange respectively. The adhesin residues that hydrogen bond with the ligands are shown in sticks. Hydrogen bonds between the adhesins and ligands are shown in dark grey dashed lines. The ligands, sTa and 3'sLn, are shown in yellow and green sticks respectively. Oxygen and nitrogen atoms are colored red and blue respectively.

As compared to adhesins with canonical YTRY motifs[50, 52, 55], the sialic acid of each trisaccharide interacts with the two non-canonical motifs of SK1 via fewer hydrogen-bonding interactions (Figure iii-6a-e). The contacts that are similar between SK1 and structurally characterized adhesins with YTRY motifs include interactions with the backbone of the N-terminal Tyr and sidechain of the Thr residues in both non-canonical YTRY motifs (Figure iii-6). Hydrogen-bonding to sialic acid additionally occurs via a backbone carbonyl in the EF loop ($SK1^{G344}$ and $SK1^{G545}$) (Figure iii-6a-d). While not previously reported for the SRR adhesins, this interaction is conserved across other adhesins ($GspB^{I479}$, $SrpA^{R342}$, and Hsa^{K335}) (Figure iii-6e).[49, 50, 69]

$SK1_{Siglec1}$ hydrogen-bonds to the sialic acid and galactose of both sTa and 3'sLn via the YTKY sequence (Figure iii-6a-b). Here, the Arg → Lys substitution, eliminates one side chain hydrogen-bond (Figure **iii-6a,** Figure iii-7a, Figure iii-8a), but the overall interaction remains similar to that supported by a canonical YTRY motif, like Hsa from *S. gordonii* strain Challis or GspB from *S. gordonii* strain M99 (Figure iii-7**b,** Figure **iii-6e**).[52]

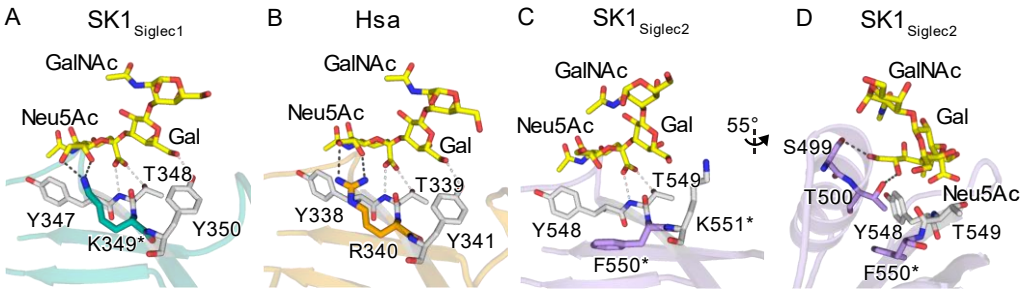

Figure iii-7 Hydrogen bond contacts between the non-canonical binding motif and sTa.

Transparent cartoon representations of $SK1_{Siglec1}$, Hsa[50], and $SK1_{Siglec2}$ are shown in teal, orange, and lavender respectively. sTa is shown in yellow sticks. The residues in the third position of the YTRY motif is colored A) teal, B) orange, and C-D) lavender. The other residues of the YTRY motif are shown in grey sticks. The hydrogen bonds between the third residue of the YTRY motif and sTa are shown in dark grey dashed lines and all other hydrogen bonds between sTa and the YTRY motifs are shown in grey dashed lines. D) Residues in the

CD loop of $SK1_{Siglec2}$ that hydrogen bond with sTa are shown in lavender sticks. Residues that deviate from the canonical YTRY motif definition are noted with an asterisk.

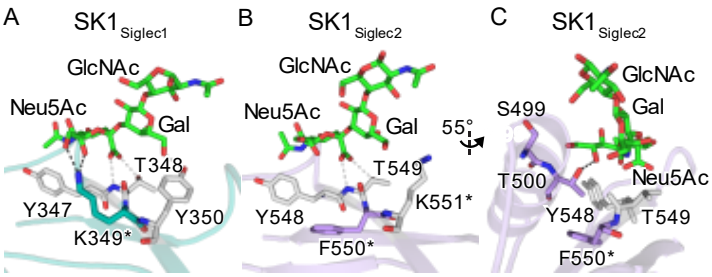

Figure iii-8 Hydrogen Bond Contacts between non-canonical binding motifs of $SK1_{Siglec1}$ and $SK1_{Siglec2}$ and 3'sLn.

3'sLn is shown in green sticks bound to $SK1_{Siglec1}$ (A) and $SK1_{Siglec2}$ (B). $SK1_{Siglec1}$ and $SK1_{Siglec2}$ areshown in transparent teal and lavender cartoon respectively. The YTRY motifs of both are shown in grey sticks, except for the residue aligned with the canonical arginine, which is shown in teal and lavender for $SK1_{Siglec1}$ and $SK1_{Siglec2}$ respectively. (A) and (B) Hydrogen bond contacts between the residue aligned with the arginine are shown in dark grey dashed lines. All other hydrogen bonds between the YTRY motif and 3'sLn are shown in grey dashed line. (C) Residues in the CD loop of $SK1_{Siglec2}$ that hydrogen bond with 3'sLn are shown in lavender sticks. Hydrogen bond contacts between the CD loop and 3'sLn are shown in dark grey dashed lines. Residues that deviate from the canonical YTRY motif definition are noted with an asterisk.

In contrast, the Phe and Lys residues in the $SK1_{Siglec2}$ YTFK sequence do not form hydrogen bond contacts with the ligand. In addition, the YTFK sequence of $SK1_{Siglec2}$ does not hydrogen bond to the galactose or the variable third sugar, GalNAc/GlcNAc, of sTa/3'sLn (Figure iii-6c-d). Compensating for this, sidechains in the helical CD loop make additional hydrogen bond contacts (Figure iii-7c-d, Figure iii-8b-c). Both observations identify that sialic acid recognition by $SK1_{Siglec2}$ differs from characterized ligand interactions in the GspB, Hsa, and SrpA adhesins.[50, 52, 55]

This reduced number of binding contacts is reflected in temperature factor analysis of the ligand. Crystallographic temperature factors can give a rough estimate of inherent mobility (Figure iii-9). It is important to note that the CD, EF, and FG loops, the YTRY motifs, and the ligands do not participate in crystal contacts in any of the three structures; as a result, the temperature factors are not influenced by crystal packing interactions. Here, the low temperature factors of sTa and $SK1_{Siglec1}$ suggest that the ligand is has little mobility when bound, which may be interpreted as strong binding (Figure iii-9a). In contrast, the higher temperature factors of 3'sLn

bound to $SK1_{Siglec1}$ or of either sTa or 3'sLn bound to $SK1_{Siglec2}$ suggest higher mobility, which suggests that these could be lower-affinity ligands (Figure iii-9b-d). Nevertheless, in all cases, the temperature factor is the lowest at the sialic acid and increases over the length of the trisaccharide, consistent with lower ligand mobility at the sialic acid and increased mobility at the reducing end sugar.

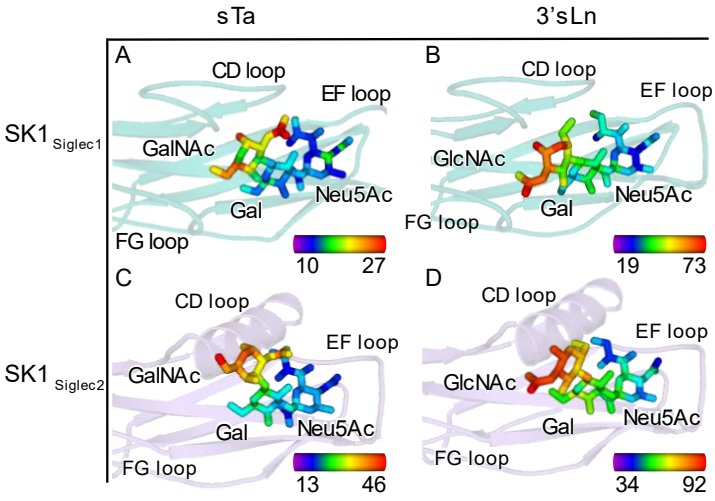

Figure iii-9 Temperature factor analysis of Siglec domains and bound ligands.

A-D) $SK1_{Siglec1}$ is shown in teal cartoon and $SK1_{Siglec2}$ is shown in lavender cartoon. The sTa liganded structures are shown on the left and the 3'sLn liganded structures are shown on the right. Both sTa and 3'sLn are shown in sticks and colored by temperature factor where blue represents a low temperature factor and red represents a high temperature factor as depicted by the scale in the bottom right corner of each panel. The scale values are in $Å^2$.

We then performed the converse analysis, assessing the temperature factors of each Siglec domain near the ligand (Figure iii-10). In the unliganded state, we observe that the selectivity loops have elevated temperature factor values as compared to the V-set Ig fold that forms the core of the Siglec domain, suggesting that they have increased mobility (Figure iii-10a-b). Upon ligand binding, the temperature factors of all loops become more similar to the temperature factors of the Siglec domain core in both $SK1_{Siglec1}$ and $SK1_{Siglec2}$ (compare Figure iii-10a,b with Figure iii-10c-f). Moreover, the temperature factors of the YTKY motif in $SK1_{Siglec1}$ decreased upon ligand binding (compare Figure iii-10a, c, e). The temperature factors of the YTFK motif in $SK1_{Siglec2}$ decrease upon sTa binding but do not decrease upon 3'sLn binding (compare Figure iii-10b, d, f). This stabilization is more

pronounced in the sTa-bound $SK1_{Siglec2}$ than in the 3'sLn-bound $SK1_{Siglec2}$, possibly due to the additional hydrogen bond between sTa and the CD helix. This extra hydrogen bond further links the YTRY region and the CD helix decreasing the mobility of both regions (Figure iii-7c-d). Together, this analysis suggests that the binding of ligands stabilizes the positions of the selectivity loops in both domains. This is consistent with prior molecular dynamics simulations of the Hsa adhesin from *S. gordonii* strain Challis, which suggested that these selectivity loops can adjust in order to optimize the interaction to ligands.[50]

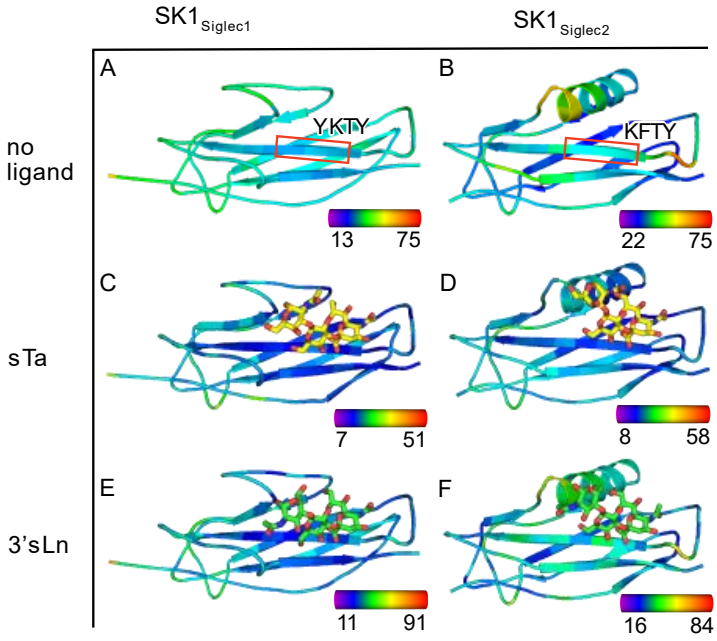

Figure iii-10 Siglec domain colored by temperature factor.

SK1 unliganded, sTa bound, and 3'sLn bound are shown in cartoon and colored by temperature factor where blue represents a low temperature factor and red represents a high temperature factor. Color bars in the bottom right corner of each panel indicate the ranges of B factors in $Å^2$. sTa and 3'sLn are shown in sticks. The oxygen and nitrogen atoms are colored red and blue respectively.

Intriguingly, the interactions between sialic acid of sTa or 3'sLn and the YTRY motif contain structural parallels to staphylococcal superantigen-like protein SSL5 bound to sialyl $Lewis^X$ (Figure iii-11).[100] Prior comparisons of SSL5 with a range of evolutionarily unrelated sialic acid-binding proteins suggested a common sialic acid recognition motif that contains two features: 1) a YYT/S motif on an edge strand of a β sheet and 2)

an arginine distant in sequence but spatially adjacent to the YYT/S motif.[100] Despite the somewhat different sequence elements in the sialic acid binding motif of $SK1_{BR}$ and other SRR adhesins, the hydrogen-bonding pattern to the sialic acid is similar to that of SSL5 (Figure iii-11). The arginine, or lysine in the case of $SK1_{Siglec1}$, associated with sialic acid binding is provided from within the YTRY motif of $SK1_{Siglec1}$ (K349) and other Siglec-like adhesins, while it is outside of the YYT/S motif in SSL5 (R186) and PT (R125) and other sialic acid-binding proteins. The interactions between the binding pocket and sialic acid may be a product of convergent evolution.[100]

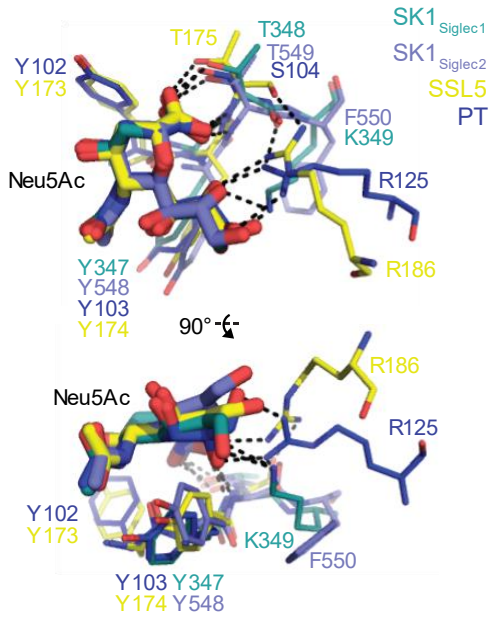

Figure iii-11 Comparison of bacterial sialic acid binding pockets.

Shown here is an overlay of the two Siglec domains of $SK1_{BR}$ with Staphylococcal superantigen-like protein 5 (SSL5; PDB entry 2R61) and Pertussis toxin (PT; PDB entry 1PTO)(48), two proteins where the sialic acid recognition motif has previously been identified (22).

$SK1_{Siglec1}$ and $SK1_{Siglec2}$ have unique selectivity profiles and exhibit synergistic binding.

Based upon our observation that both $SK1_{Siglec1}$ and $SK1_{Siglec2}$ bind sialoglycans in a co-crystal structure, we tested the relevance of these interactions in binding to defined, synthetic glycans. To do this, we developed recombinant glutathione S-transferase (GST)-tagged proteins containing either the first Siglec and Unique

domains, $SK1_{Siglec1+Unique1}$ ($SK1^{252-455}$), or the second Siglec and Unique domains, $SK1_{Siglec2+Unique2}$ ($SK1^{449-660}$) and compared the binding of these isolated binding modules to that of full-length $SK1_{BR}$.

We began by evaluating how each repeat bound to a small library of tri- and tetrasaccharide sialoglycans at a single concentration of ligand (Figure iii-12a). Both $SK1_{Siglec1+Unique1}$ and $SK1_{Siglec2+Unique2}$ bound to at least some of the tested sialoglycans, albeit less strongly than did full-length $SK1_{BR}$. Consistent with the crystallographic temperature factor analysis (Figure iii-9), $SK1_{Siglec1+Unique1}$ showed a statistically significant preference for sTa. In contrast, $SK1_{Siglec2+Unique2}$ appears to be more broadly selective. The latter finding is consistent with the observation that the $SK1_{Siglec2}$ domain does not make hydrogen bonding contacts to the second and third sugars of trisaccharide in the crystal structures (Figure iii-6c-d).

We then performed a more detailed dose-dependent binding to sTa and 3'sLn (Figure iii-12b). This analysis identifies that both $SK1_{Siglec1+Unique1}$ and $SK1_{Siglec2+Unique2}$ bind to sTa about 100-fold less strongly than the full-length $SK1_{BR}$, which contains the two sub-domains in tandem. This suggests the possibility of binding synergy when in the presence of high concentrations of carbohydrate. Physiologically, small portions of large and complex branched glycans could have a similar appearance as high local concentrations in a binding assay. Such high levels of a particular glycan are hypothesized for glycoproteins that contain oriented glycan patches.[101]

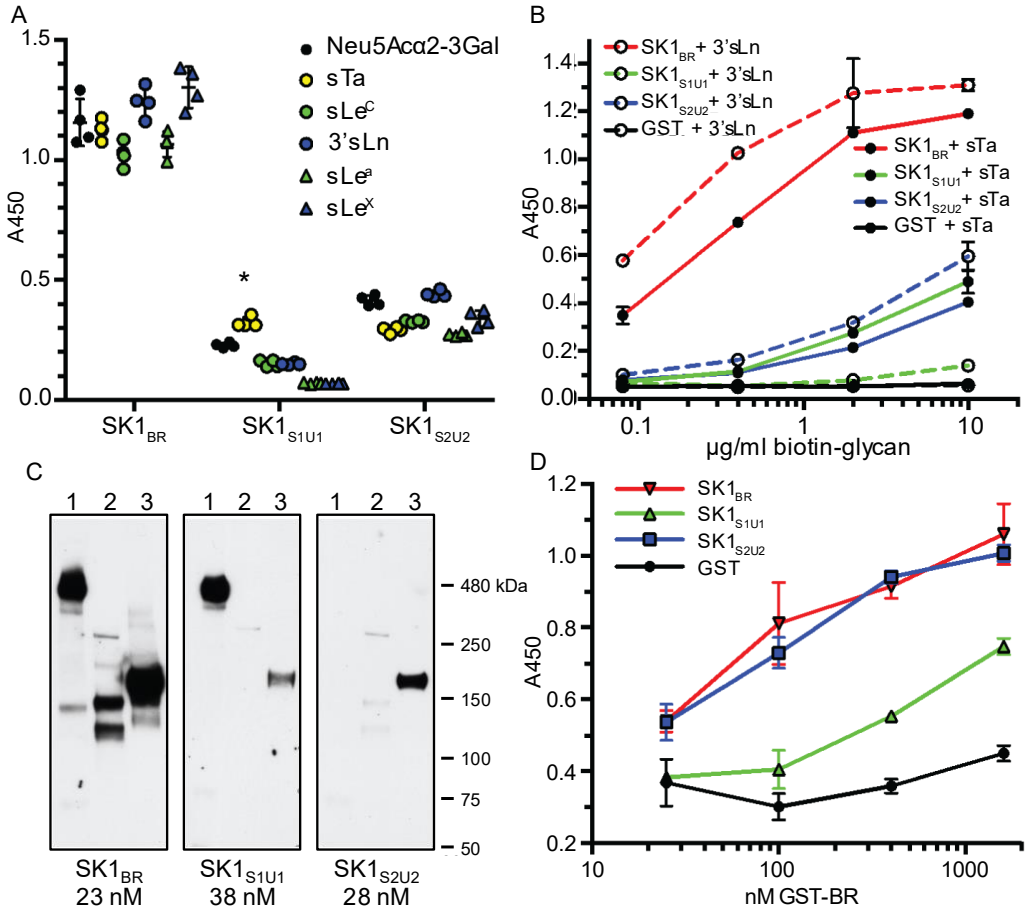

Figure iii-12 Binding of $SK1_{BR}$ and split variants to glycans and glycoproteins.

A) Biotin-glycan binding to immobilized GST-tagged $SK1_{BR}$ and the split binding modules (n = 4 technical replicates). Asterisk indicates binding that was significantly greater than the level of binding to all other glycans in the set of six ($p<0.05$ using a two-way ANOVA with Tukey's correction for multiple comparisons). B) Binding of biotinylayed sTa or 3'sLn to immobilized GST-tagged $SK1_{BR}$ and split constructs (n = 3 technical replicates). C) Binding of GST-tagged $SK1_{BR}$ and split constructs to glycoproteins in human plasma (lane 1), platelet lysate (lane 2) or submandibular sublingual saliva (lane 3). D) Binding of GST-tagged $SK1_{BR}$ and $SK1_{BR}$ deletion constructs to immobilized human platelets (n = 3 technical replicates). In panels A, B and D, mean values ± standard deviation are indicated. In cases where error bars are not evident, the deviations were smaller than the size of the symbol used for the data point. Background values for GST alone were not subtracted, but are shown in panels B and D.

Synergistic properties of $SK1_{BR}$ affect binding to host receptors.

We next evaluated whether binding to the synthetic glycans explains how $SK1_{BR}$ interacts with host glycoproteins. Previously identified ligands for the Siglec-like SRR adhesins are consistent with their biological roles in oral commensalism and the pathogenesis of infective endocarditis. These include salivary MUC7, platelet

GPIbα, and several *O*-glycosylated plasma glycoproteins. We therefore evaluated the interactions of GST-tagged $SK1_{BR,}$ $SK1_{Siglec1+Unique1}$ and $SK1_{Siglec2+Unique2}$ with human salivary, platelet, and plasma glycoprotein targets via Far Western blotting. Isolated binding regions, $SK1_{Siglec1+Unique1}$ and $SK1_{Siglec2+Unique2}$, both bound modestly to MUC7 compared with binding by the tandem domains of $SK1_{BR}$ (Figure iii-12c). $SK1_{Siglec1+Unique1}$ readily bound a 460 kDa plasma protein, whereas $SK1_{Siglec2+Unique2}$ did not. Here, the tandem domains of $SK1_{BR}$ did not increase the binding over what was observed for $SK1_{Siglec1+Unique1}$. Neither $SK1_{Siglec1+Unique1}$ nor $SK1_{Siglec2+Unique2}$ bound appreciably to GPIbα in the platelet lysate.

We also assessed binding to fixed, immobilized platelets (Figure iii-12d). $SK1_{Siglec1+Unique1}$ bound weakly, while $SK1_{Siglec2+Unique2}$ bound more strongly. There was not a cooperative effect of linking these domains, as the binding of $SK1_{BR}$ to platelets could be fully explained by the binding of $SK1_{Siglec2+Unique2}$. These results suggest that there is a high-affinity ligand for $SK1_{BR}$ on intact platelets that is due primarily to binding by $SK1_{Siglec2+Unique2}$. The combined results indicate that $SK1_{BR}$ can bind multiple simple and complex sialoglycan ligands on biological targets via a combination of interactions.

Discussion

All previously determined structures of the binding regions of Siglec-like adhesins have a single Siglec domain and a single Unique domain. This prior work has shown that sialic acid binding affinity largely stems from binding to a YTRY motif with the selectivity tuned via adjacent loop regions of the Siglec domain.[50] Our data are consistent with each repeat of $SK1_{BR}$ following the same principles for binding and selectivity. Specifically, the sialoglycan binds via specific interactions between sialic acid and the non-canonical YTRY motif. Given this finding in conjunction with studies of SrpA, the YTRY motif can be more formally defined as ϕTRX, where ϕ represents W, F or Y and X represents Y, T, E, H, or K (Figure iii-2).[49, 55] This ϕTRX motif represents a sequence modification of the YYS/T motif found in other sialic acid binding proteins, with ϕT of the **ϕT**RX corresponding to the YT of Y**Y**S/**T**. As a result, both sequence motifs interact with sialic acid via a similar pattern of hydrogen-bonds. For the Siglec-like adhesins, the selectivity loops may control the identity of the

preferred sialoglycan. In $SK1_{Siglec1}$ and $SK1_{Siglec2}$, these selectivity loops differentially impact the size of the binding pocket and the orientation of the ligand, resulting in unique selectivity profiles.

While the ϕTRX motif and selectivity loops of $SK1_{BR}$ support our current understanding of sialoglycan binding, two unique features highlight unanswered questions regarding the link between sialoglycan binding and adhesion. For instance, $SK1_{Siglec1+Unique1}$ exhibited selectivity for sTa, but had a lower affinity for glycoproteins than $SK1_{Siglec2+Unique2}$. In contrast, $SK1_{Siglec2+Unique2}$ binds well to host components but poorly to purified trisaccharides. It is possible that the trisaccharides tested do not include the full biological ligand. SrpA similarly has a larger and more open binding pocket that could possibly accommodate a larger ligand (Figure iii-13c). Hypotheses for the biological ligand can be developed by considering parallels to the SrpA adhesin from *S. sanguinis* strain SK36. Like $SK1_{Siglec2+Unique2}$, SrpA binds poorly to purified tri- and tetra-saccharides *in vitro*, but SrpA binds robustly to human platelets. Lacking an FG loop, SrpA has a significantly larger binding site than many other SRR adhesins. This increased size of the binding site opens to a distal arginine residue (Figure iii-13c, f). For this reason, it has been proposed that physiological targets of SrpA may include a disialylated hexasaccharide or a patch of multiple, oriented glycans.[49] Either possibility for the ligand could promote cooperativity that would be expected to increase adhesion to host targets. Like SrpA, $SK1_{Siglec2}$ similarly has a large sialoglycan binding site due to the small size of the FG loop (Figure iii-13b, Figure iii-14a). We propose that $SK1_{Siglec2}$ may therefore also bind a "core 2" sialoglycan or patch of oriented glycans (Figure iii-14b-c). In contrast, the distal arginine residue in $SK1_{Siglec1}$ is occluded from the sialoglycan binding site by a larger FG loop, likely prohibiting interactions with a sialoglycan longer than a trisaccharide (Figure iii-13**a,** Figure iii-14**a**); however, binding of multiple glycans is still a possibility. Alternatively, a large binding site alone may be sufficient for some adhesins to bind larger saccharides. Hsa also has a large and open binding site, but with no distal arginine (Figure iii-13d, f). This is consistent with the ability of Hsa to bind both trisaccharides and larger, branched sialoglycans.[69, 102, 103]

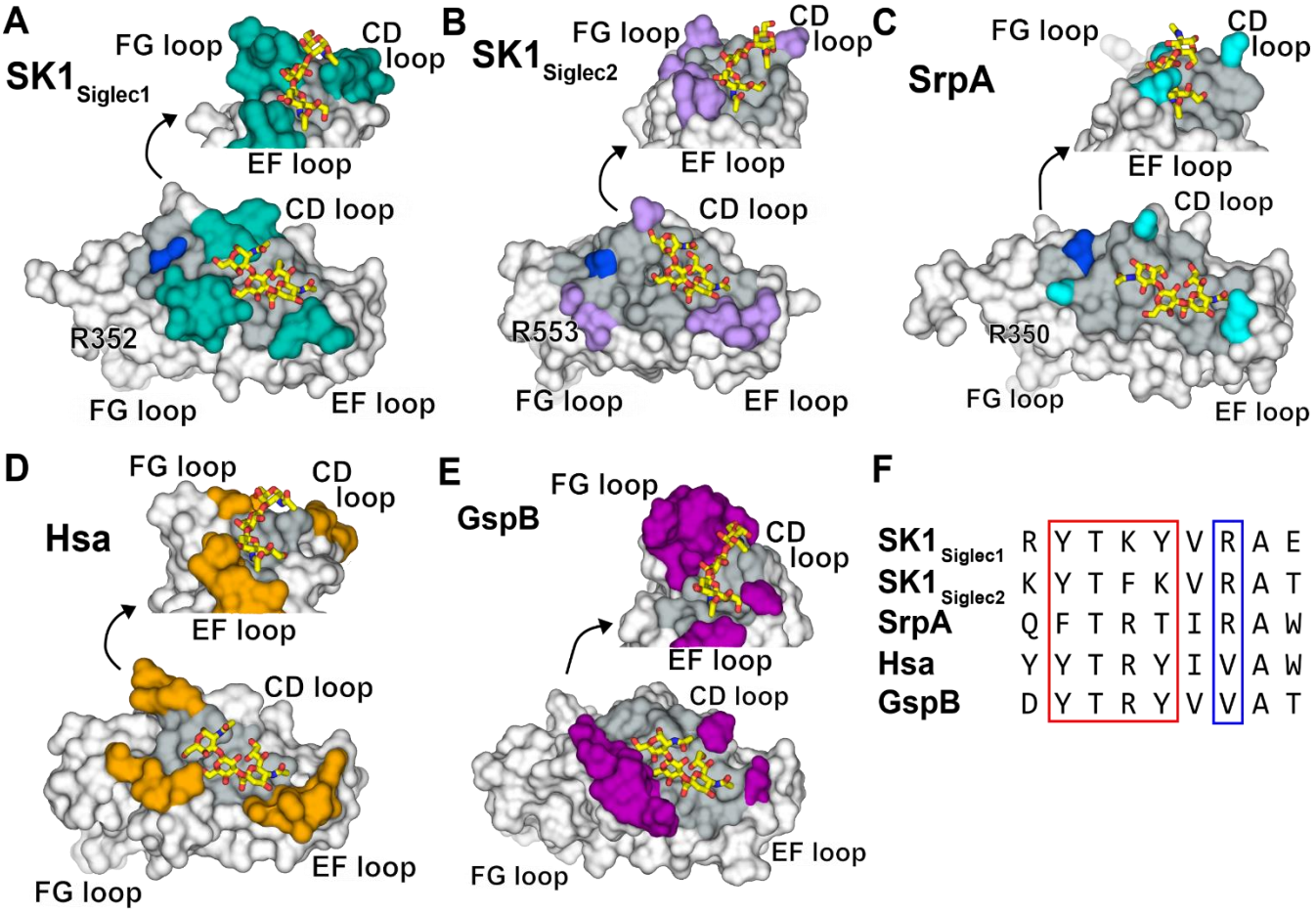

Figure iii-13 SRR adhesins binding pocket size comparison.

A) - E) The adhesins are shown in surface representations (16-18). The top image of the binding site is rotated 80° around the z axis and 70° around the x axis. The binding pocket of each is colored in grey and the portions of the CD, EF, and FG loops that create the walls of the binding pocket are colored in teal, lavender, cyan, orange, and magenta for $SK1_{Siglec1}$, $SK1_{Siglec2}$, SrpA, Hsa, and GspB, respectively. The Arg distal to the YTRY motif is colored blue. Sialyl T antigen is shown in yellow sticks bound to each adhesin. F) Multiple sequence alignment of the above adhesins is shown. The YTRY motif is outlined in red and the distal Arg is outlined in blue.

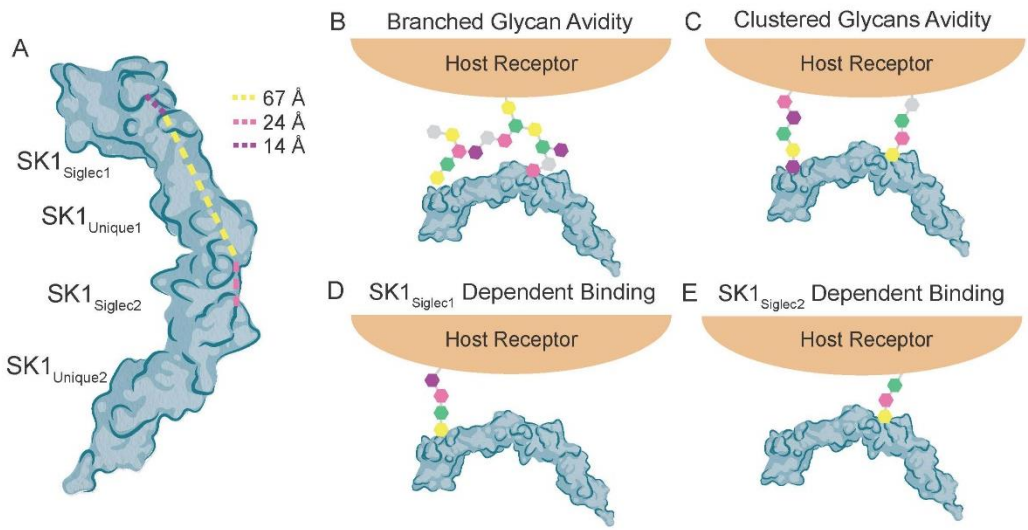

Figure iii-14 Model of target-specific effects of SK1.

A) The distance between the two binding sites is 67Å, measured from the C_α atoms of Y347, the first residue in the YTRY motif of $SK1_{Siglec1}$ and R553, the distal arginine residue of $SK1_{Siglec2}$. The length of the binging site in $SK1_{Siglec1}$ is 14Å, measured from the Cα atoms of G344 and K349. The length of the binding site in $SK1_{Siglec2}$ is 24Å, measured from the Cα atoms of G545 to R553. B-E) Note that the glycans shown here are only meant to serve as a hypothetical glycan structure and are not meant to represent a specific, defined glycan target. B) Multivalent binding of a branched glycan: Given the distance between the two binding sites, $SK1_{BR}$ could bind the same branched glycan with both $SK1_{Siglec1}$ and $SK1_{Siglec2}$. C) Multivalent binding of a patch of clustered glycans: Each binding site of $SK1_{BR}$ binds a separate glycan structure. D) $SK1_{Siglec1}$ dominant binding: E) $SK1_{Siglec2}$ dominant binding given the openness of the binding site and the presence of the distal Arg, $SK1_{Siglec2}$ may bind a short saccharide, two short saccharides, or a longer hexasaccharide.

$SK1_{BR}$ also demonstrates how individual binding domains versus tandem linkage can differentially contribute to the affinity for host targets. Conceptually, tandem linking of these two binding regions would be expected to confer cooperative binding capability in binding to some biological targets.[104] Or to put it another way, if a target protein contains glycan modifications that are correctly spaced and oriented, this adhesin could bind more strongly to host receptors via multivalent binding (Figure iii-14b-c).[105] Synergy was indeed observed between $SK1_{Siglec1+Unique1}$ and $SK1_{Siglec2+Unique2}$ when binding to platelet lysate or to salivary glycoproteins (Figure iii-12c). This suggests that both binding regions contribute to adherence for certain host targets and could be explained by binding to either a large, branched glycan structure or a patch of clustered glycans (Figure iii-14b-c).

On the other hand, binding of $SK1_{Siglec1+Unique1}$ to human plasma is roughly equivalent to that of $SK1_{BR}$ suggesting $SK1_{Siglec1+Unique1}$ has a high affinity sialoglycan target in human plasma, but $SK1_{Siglec2+Unique2}$ does not. $SK1_{Siglec1+Unique1}$ may be responsible for adherence to human plasma, consistent with an $SK1_{Siglec1}$ dependent binding model (Figure iii-12c, Figure iii-14d). $SK1_{Siglec2+Unique2}$ seems to be solely responsible for adherence to immobilized platelets suggesting an $SK1_{Siglec2}$ dependent binding mode (Figure iii-12d, Figure iii-14e). The use of tandem repeats and multivalent binding capabilities could confer two distinct evolutionary advantages. First, tandem repeats can be separately mutated for a faster evolution. For example, if an adhesin contains two binding regions following a gene duplication event[106], the domains could then be individually mutated, with each domain conferring different selectivity for host receptors. Divergent evolution of the two binding domains could effectively double the evolutionary rate, leading to faster adaptation. Tandem linkage of binding domain modules could also allow an individual binding domain to evolve through an intermediate with lower affinity and broader specificity. This could allow for the evolution of larger changes in selectivity.[107, 108, 109]

Second, the combined action of two binding domains could allow binding of this adhesin to a broader range of targets. Linking an sTa-specific domain with a domain of another or broader selectivity could allow adherence to either platelets or to other host targets. Increasing the range of ligands an adhesin can bind could increase tropism and allow bacteria to migrate from one tissue to another. Avidity and affinity optimization of protein scaffolds for recognition of on- and off-target biomolecules can increase specificity of cellular targeting.[110] This hetero-multivalent binding could be important for increasing specificity for a target tissue.[105, 111, 112]

Taken together, the findings reported here get us closer to addressing unanswered questions in the field. The finding that non-canonical motifs of SRR adhesins interact robustly with sialoglycan ligands reveals these as important for host interaction. More importantly, these results provide insight into how adhesive proteins adapt to various biological niches with different host receptors and provide evidence for adhesion to patches of oriented glycans.[103] This feeds into ongoing work that seeks to develop a predictive model for streptococcal pathogenicity.

iv. ORIGINS OF GLYCAN SELECTIVITY IN STREPTOCOCCAL SIGLEC-LIKE ADHESINS SUGGEST MECHANISMS OF RECEPTOR ADAPTATION

Barbara A. Bensing, Haley E. Stubbs, Rupesh Agarwal, Izumi Yamakawa, Kelvin Luong, Kemal Solakyildirim, Hai Yu, Azadeh Hadadianpour, Manuel A Castro, Kevin Fialkowski, Zdzislaw Wawrzak, Xi Chen, Carlito B. Lebrilla, Jerome Baudry, Jeremy C. Smith, Paul M Sullam, T M Iverson. Origins of glycan selectivity in streptococcal Siglec-like adhesins suggest mechanisms of receptor adaptation. 2022. **Nat Commun 13**(1): 2753.

My specific contributions to this work include refinement of sLe^C, sLe^X, 6S'sLe^X, and 3'sLn bound $SLBR_{Hsa}$ structures. I compared these structures to those of $SLBR_{GspB}$, sTa bound $SLBR_{Hsa}$, $SLBR_{SrpA}$, and $SLBR_{SK1}$. I also performed statistical analysis on all the ELISA binding data that was experimentally obtained by Dr. Barbara A. Bensing, co-first author except for supplementary table 4. I made figures 3c-d, 4, 6f, 7d-f 10, and supplementary figures 2-8, 13-14, 17a-b. Additionally, I contributed to or edited figures 1, 2, 3, 7a-c, 8, and supplementary figures 15, 19a-b, d and 22. Lastly, I was responsible for deposition of raw crystallography data to SBGrid and refined structural models to the PDB. I also wrote responses to reviewers and organized and submitted all raw data to the journal including raw far western blots, ELISA absorbance values, and mass spectrometry protein identification data.

Bacterial binding to host receptors underlies both commensalism and pathogenesis. Many streptococci adhere to protein-attached carbohydrates expressed on cell surfaces using Siglec-like binding regions (SLBRs). The precise glycan repertoire recognized may dictate whether the organism is a strict commensal versus a pathogen. However, it is currently not clear what drives receptor selectivity. Here, we use five representative SLBRs and

identify regions of the receptor binding site that are hypervariable in sequence and structure. We show that these regions control the identity of the preferred carbohydrate ligand using chimeragenesis and single amino acid substitutions. We further evaluate how the identity of the preferred ligand affects the interaction with glycoprotein receptors in human saliva and plasma samples. As point mutations can change the preferred human receptor, these studies suggest how streptococci may adapt to changes in the environmental glycan repertoire.

Introduction

Selection between many possible host receptors determines whether a bacterium can adhere to a preferred anatomical niche or can infect a particular host[113, 114]. Streptococci and staphylococci are among the organisms that use host-associated carbohydrates as receptors; these may specifically bind to sialic acid-containing glycans (sialoglycans; Figure iv-1). As an example, human *O*-linked glycosylated proteins commonly contain a terminal α2-3-linked sialic acid-galactose disaccharide, (Neu5Acα2-3Gal). Additional forms of sialic acid and alternative linkages are found in animal sialoglycans[115, 116].

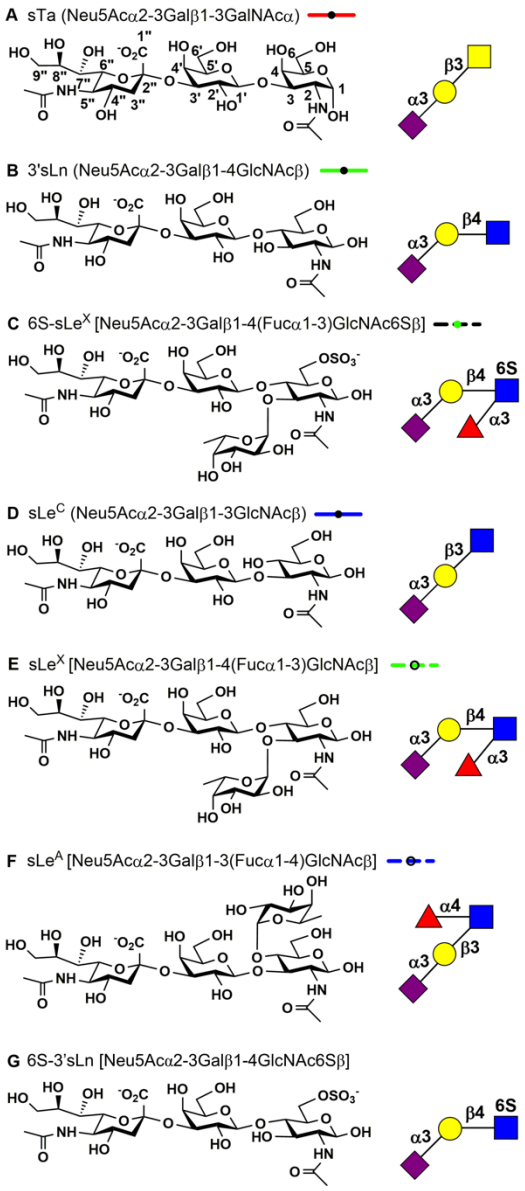

Figure iv-1 Sialoglycans used in this study.

The chemical structure of each indicated sialoglycan is shown on the left with the symbolic representation shown on the right. The line style used for all dose response curves is shown to the right of each name.

Neu5Acα2-3Gal is present on the human salivary mucin MUC7[69, 117, 118], a glycoprotein in blood plasma[70], and surface platelet proteins[119, 120]. Bacterial binding to glycoproteins terminating with Neu5Acα2-3Gal may therefore allow colonization of the oral cavity as a commensal. In animal models, sialoglycan binding is also implicated in the persistence of these organisms in the bloodstream as an endovascular pathogen[52, 121, 122, 123], although it is not known whether all streptococci can act as pathogens.

Siglec-like binding regions (SLBRs) are among the streptococcal adhesive modules that bind sialoglycans. SLBRs are usually found within the context of serine-rich repeat proteins, which form fibrils extending from the bacterial surface. SLBRs contain two adjacent modules: a “Siglec” domain, which shares some features with mammalian Siglecs, and a “Unique” domain[52] with no close homologs outside of the family. The Siglec domain contains a ΦTRX sequence motif[124] that recognizes Neu5Acα2-3Gal in the context of larger glycans. Reported mutagenesis of the ΦTRX motif demonstrates its importance in sialoglycan binding[52, 55, 69] and in endovascular disease in animal models[52]. This has motivated the development of compounds that bind the ΦTRX motif as a potential therapy for human endovascular infections caused by these organisms[125, 126].

SLBRs display a range of selectivity. Some SLBRs bind selectively to the α2-3-linked trisaccharide sialyl-T antigen (sTa, Neu5Acα2-3Galβ1-3GalNAc; Figure iv-1A)[69, 127]. Others have intermediate selectivity and bind to a small number of closely related glycans[69, 127]. Still others can bind to a broad range of sialoglycans and do not distinguish between related structures[69, 127]. The binding profile of these SLBRs is likely adapted to the host display of sialoglycans. In the oral cavity for example, the display of sialylated *O*-glycans on MUC7 varies between individuals, making it possible that the binding preferences of the SLBRs reflect the specific glycosylation display of an individual[69, 117, 118, 128, 129]. The binding profile can also affect virulence; streptococci

containing SLBRs that preferentially bind to sTa *in vitro* exhibit higher pathogenicity in an animal model of endocarditis. [130]

Despite the importance of the sialoglycan binding profile in the interaction between streptococci and host [130], the sequence determinants that underlie glycan selectivity are not currently clear. Here, we determine the molecular basis for glycan selectivity of a phylogenetically informed library of SLBRs. We test our predictions for selectivity by engineering the binding spectrum of selected SLBRs and assessing host receptor switching in human saliva and plasma glycoproteins. Collectively, these studies improve our understanding of the glycan selectivity that underlies commensalism and pathogenesis. In addition, they suggest how these bacteria may adapt to host sialoglycan repertoires.

Results

Selection of SLBRs for study

Starting with SLBRs with at least some previously reported selectivity, we correlated selectivity with phylogeny (Figure iv-2)[49, 69, 70, 127]. Our initial trees contained two major branches. This identified that evolutionary relatedness of SLBRs is a moderate, but not strong, predictor of glycan selectivity. Most SLBRs of the first major branch of the tree (blue in Figure iv-2) are broadly-selective and recognize two or more related tri-, tetra-, or hexasaccharides (see examples in Figure iv-1). However, sequence similarly does not clearly predict the preferred glycan[49, 69, 70, 127]. In contrast, characterized SLBRs of the second major branch (green in Figure iv-2) are selective for sTa (Figure iv-1**A**)[49, 69, 70, 127].

To understand selectivity of these SLBRs for human glycans, we chose comparators from each branch for detailed study. From the first branch of the tree (blue in Figure iv-2), we selected the SLBRs of the Hsa adhesin from *S. gordonii* strain Challis (termed $SLBR_{Hsa}$) and the equivalent SLBRs from *Streptococcus sanguinis* strain SK678 ($SLBR_{SK678}$) and *Streptococcus gordonii* strain UB10712 ($SLBR_{UB10712}$; see footnote). Although these three SLBRs are >80% identical in amino acid sequence, when they were tested with arrays containing 49 sialoglycans, they exhibited distinct binding profiles[69, 127]. $SLBR_{Hsa}$ was quite broadly selective and bound to a

range of α2-3-linked sialoglycans, but not to the corresponding fucosylated derivatives[69, 127]. In comparison, $SLBR_{UB10712}$ and $SLBR_{SK678}$ were more narrowly selective, although both bound to multiple sialoglycan ligands. Specifically, $SLBR_{UB10712}$ bound strongly to 3'-sialyl-*N*-acetyllactosamine (3'sLn; Neu5Acα2-3Galβ1-4GlcNAcβ, Figure iv-1B) and a small range of related structures[69], while $SLBR_{SK678}$ bound to only two of the glycans on this array, 3'sLn and 6-*O*-sulfo-sialyl Lewis X (6S-sLe^X; Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAc6Sβ, Figure iv-1C)[69]. In summary, all three of these SLBRs bind multiple ligands with promiscuity following $SLBR_{Hsa}$ > $SLBR_{UB10712}$ > $SLBR_{SK678}$.

The second major branch of the evolutionary tree (green in Figure iv-2) includes the well-characterized $SLBR_{GspB}$ from *S. gordonii* strain M99[52, 118, 119, 131, 132]. $SLBR_{GspB}$ exhibits narrow specificity for the sTa trisaccharide, as have other previously characterized members of this evolutionary branch[49, 69, 70, 127, 131]. The binding results for GST-$SLBR_{GspB}$ with sTa, 3'sLn, and sialyl $Lewis^C$ (sLe^C, Neu5Acα2-3Galβ1-3GlcNAc) (Figure iv-1D) were recapitulated here by ELISA showing concentration-dependent binding (Figure iv-3A).

In seeking comparators of $SLBR_{GspB}$, we evaluated close homologs for their binding spectrum. We identified that a previously uncharacterized SLBR from *Streptococcus sanguinis* strain SK150 (termed $SLBR_{SK150}$) displays 62% identity to $SLBR_{GspB}$ but exhibits a distinct binding profile (Figure iv-3B). In short, there was modest binding to each of the three trisaccharides, i.e. sTa, 3'sLn, and sLe^C, but little detectable binding to any of the tetrasaccharides (i.e. 6S-sLe^X (Figure iv-1C), sialyl Lewis X (sLe^X; Neu5Acα2-3Galβ1-4(Fucα1-3)GlcNAcβ; Figure iv-1D), and sialyl $Lewis^A$ (sLe^A; Neu5Acα2-3Galβ1-3(Fucα1-4)GlcNAcβ; Figure iv-1E)) (Figure iv-3B). The high sequence similarity and distinct binding properties of $SLBR_{GspB}$ and $SLBR_{SK150}$ make these good comparators for understanding selectivity.

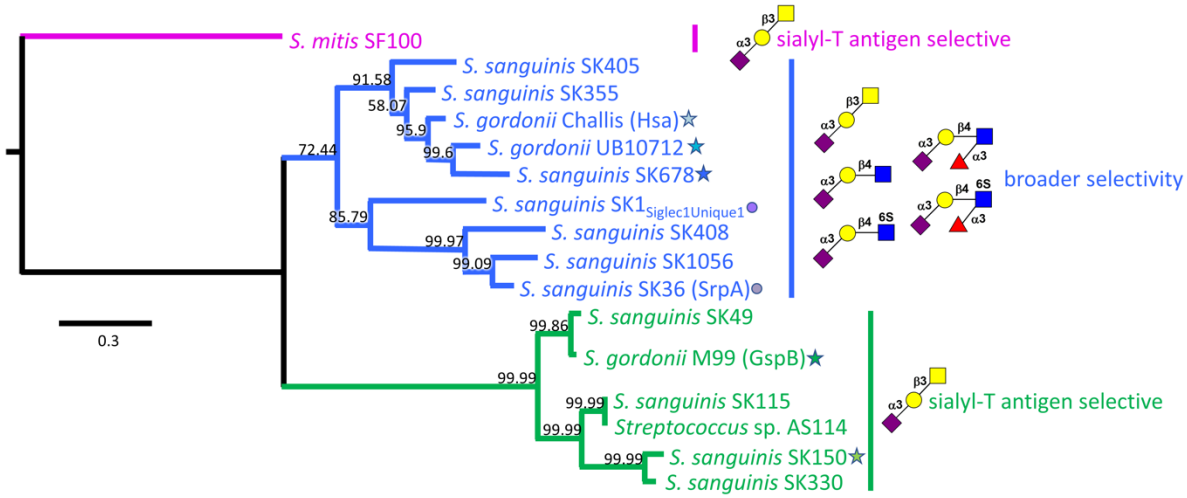

Figure iv-2 Phylogeny of select bacterial SLBRs.

Phylogenetic analysis of the SLBRs from the indicated strains comprising the tandem Siglec and Unique domains reveals three distinct subgroups. Glycans are depicted using standard symbol nomenclature, with linkage designations shown as numbers and the 6S elaborations shown in red. Characterized $SLBR_{Hsa}$-like SLBRs (*blue*) bind to two or more of the indicated sialoglycans; the previously characterized $SLBR_{GspB}$-like SLBRs (*green*) exhibited narrow selectivity for sialyl-T antigen. The tree is rooted using the distantly related *S. mitis* $SLBR_{SF100}$ (*magenta*). SLBRs investigated here are highlighted with a star with the color family reflect the branch of tree; later figure panels comparing properties of these SLBRs follow this coloring. The structure and ligand binding properties of $SLBR_{SrpA}$ and $SLBR_{SK1}$ are highlighted with circles as they have previously been reported[49, 52, 55, 124] and are used as comparators in this report. The scale bar indicates the average number of nucleotide substitutions per site, and the numbers on each branch represent the confidence of inferred tree topology.

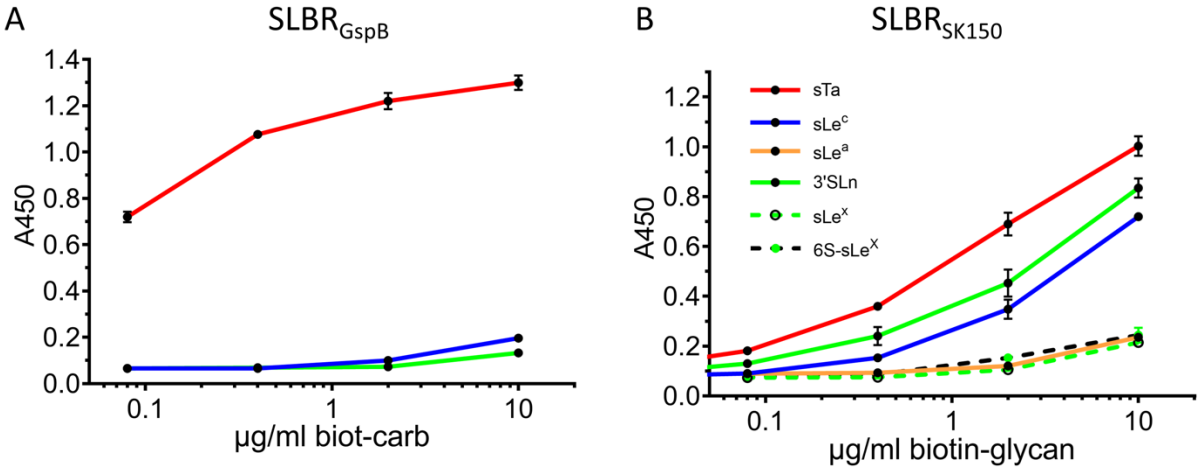

Figure iv-3 Binding properties of $SLBR_{GspB}$ and $SLBR_{SK150}$.

Dose response curves of biotin-glycan binding to immobilized A. GST-$SLBR_{GspB}$ and B. GST-$SLBR_{SK150}$. Measurements were performed by ELISA using 500 nM of immobilized GST-SLBR and the indicated concentrations of each ligand are shown as the mean ± SD (n = 3 independent experiments with a single protein preparation). sTa, sLe^C, and 3'sLn are trisaccharides; sLe^A, sLe^X, and 6S-sLe^X, are tetrasaccharides (Figure iv-1).

Structural basis for inclusion or exclusion of sialoglycan elaborations

To reveal how similar SLBRs could support different selectivity, we determined crystal structures of these five SLBRs at resolutions between 1.0 – 1.7 Å (Table iv-1**,** Table iv-2 Figure iv-4**,** Figure iv-5). This included a structure of $SLBR_{GspB}$ with improved resolution as compared to a previous report[52]. In each structure, the N-terminal Siglec domain is organized around a V-set variation of the Ig fold (Figure iv-4), named for its discovery in antibody variable domains[133]. The C-terminal Unique domain of the SLBRs displays a fold that has only been observed in other members of this family (Figure iv-5).

	$SLBR_{Hsa}$	**$SLBR_{UB10712}$**	**$SLBR_{SK678}$**
PDB entry	6EFC	6EFF	6EFI
DATAID	328	509	510
Resolution	1.4 Å	1.6 Å	1.7 Å
Data collection			
Beamline	APS 21-ID-F	APS 21-ID-F	SSRL 9-2
Wavelength	0.978 Å	0.978 Å	0.979 Å
Space group	$P2_12_12_1$	P1	$P2_1$
Unit cell	a=46.6 Å b=58.1 Å c=76.0 Å	a=39.8 Å b=48.9 Å c=99.8 Å α=101.8° β=91.4° γ=89.9°	a=59.6 Å b=59.58 Å c=61.8 Å β=100.7°
R_{sym}	0.084 (0.650)	0.075 (0.730)	0.099 (0.530)
R_{pim}	0.024 (0.281)	0.047 (0.479)	0.040 (0.213)
I/σ	49.7 (2.3)	22.9 (1.9)	15.0 (4.4)
Completeness (%)	93.3% (60.9%)	92.4% (70.9%)	97.7% (97.3%)
Redundancy	12.6 (5.6)	3.6 (3.4)	7.0 (7.1)
$CC_{1/2}$	0.837	0.648	0.998
Refinement			
R_{cryst}	0.146[&]	0.180	0.177
R_{free}	0.179	0.207	0.210

No. Mol per ASU	1	4	2
RMS deviation			
bond lengths	0.01 Å	0.01 Å	0.01 Å
bond angles	1.6°	0.9°	0.7°
Ramachandran			
favored	97.0%	96.8%	99.0%
allowed	2.5%	3.1%	1.0%
outliers	0.5%*	0.1%	0.0%

Table iv-1 Crystallographic data collection and refinement statistics for unliganded $SLBR_{Hsa}$, $SLBR_{UB10712}$, and $SLBR_{SK678}$.

Values in parentheses are for the highest resolution shell. Raw data are deposited with SBGrid and can be accessed at: data.sbgrid.org/dataset/DATAID. The Ramachandran angles identified as outliers ($SLBR_{Hsa}^{S253}$, $SLBR_{Hsa}^{L363}$, $SLBR_{UB10712}^{S253}$, $SLBR_{UB10712}^{L361}$) are associated with clear electron density. Data collected at the Advanced Photon Source (APS) used the Remote Access, data collected at Stanford Synchrotron Radiation Lightsource (SSRL) used the Blu-Ice software.

	$SLBR_{GspB\text{-}Siglec}$ Form 1	$SLBR_{GspB\text{-}Siglec}$ Form 2	$SLBR_{GspB}$	$SLBR_{SK150}$
PDB entry	6EF7	6EF9	6EFA	6EFB
SBGrid DATAID	812	601	604	508
Resolution	1.03 Å	1.3 Å	1.6 Å	1.90 Å
Data collection				
Beamline	21-ID-F	21-ID-G	21-ID-F	Bruker X8R
Wavelength	0.979 Å	0.979 Å	0.979 Å	1.542 Å
Space group	$P2_12_12$	R32	$P2_12_12_1$	$P2_1$
Unit cell	a= 33.9 Å	a=b=92.1 Å	a=33.0 Å	a=24.3 Å
	b= 46.2 Å		b=47.6 Å	b=62.6 Å
	c= 73.0 Å	c=143.9 Å	c=136.2 Å	c=62.9 Å
				β=98.6°
R_{sym}	0.049 (0.430)	0.061 (0.771)	0.057 (0.610)	0.139 (0.538)
R_{pim}	0.024 (0.233)	0.017 (0.302)	0.019 (0.213)	0.044 (0.295)
I/σ	25.5 (4.4)	59.3 (2.8)	43.8 (3.3)	9.3 (1.9)
Completeness (%)	95.4% (90.6%)	99.9% (98.0%)	88.9% (48.6%)	97.3% (91.6%)
Redundancy	9.3 (8.2)	13.8 (7.1)	9.2 (7.7)	9.0 (3.6)
$CC_{1/2}$	0.911	0.941	0.975	0.996
Refinement				
R_{cryst}	0.125	0.131	0.166	0.172
R_{free}	0.141	0.144	0.209	0.188
No. Mol per ASU	1	2	1	1
RMS deviation				
bond lengths	0.01 Å	0.01 Å	0.02 Å	0.01 Å
bond angles	1.3°	1.1°	1.6°	0.7°
Ramachandran				
favored	100%	97.0%	98.0%	99.0%
allowed	0%	2.2%	1.5%	1.0%

outliers*	0%	0.8%	0.5%	0.0%

Table iv-2 Crystallographic data collection and refinement statistics for unliganded $SLBR_{GspB}$, $SLBR_{GspB\text{-}Siglec}$ and $SLBR_{SK150}$.

Values in parentheses are for the highest resolution shell. Raw data are deposited with SBGrid and can be accessed at: data.sbgrid.org/dataset/DATAID. Data collected at the APS used Remote software, and data collected on the Bruker X8R used the CR-XRD software.

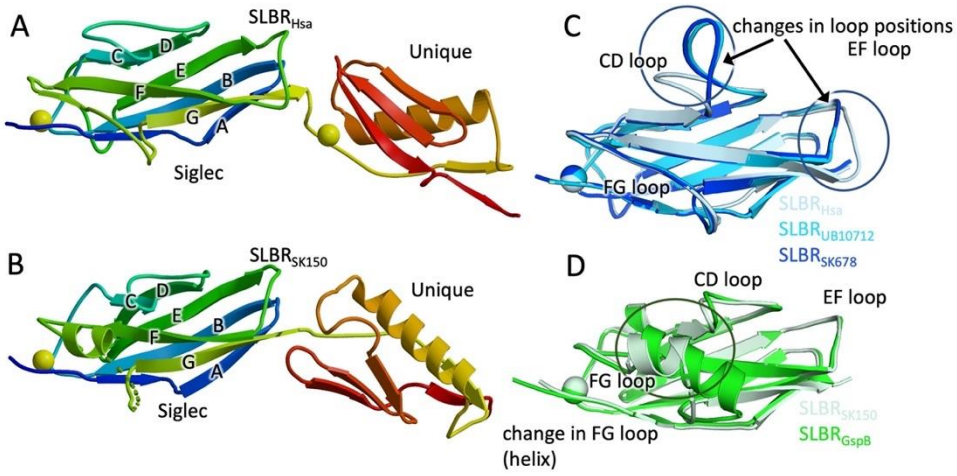

Figure iv-4 Structural differences between $SLBR_{Hsa}$, $SLBR_{UB10712}$, and $SLBR_{SK678}$.

Ribbon diagrams of A. $SLBR_{Hsa}$ and B. $SLBR_{SK150}$ with the N-terminus in *blue* and the C-terminus in *red*. Ions are shown as spheres. C. Overlay of the Siglec domain from the $SLBR_{Hsa}$ (*grey-blue*) $SLBR_{UB10712}$ (*cyan*) and $SLBR_{SK678}$ (*blue*). The color of each SLBR is the same as the color of the stars for each strain in Figure 2. D. Overlay of the Siglec domain in the $SLBR_{GpsB}$ (*green*) and $SLBR_{SK150}$ (*light green*). The CD, EF, and FG loops are highlighted. These are poorly conserved in sequence and length (**Supplementary Figure 4**) and display significant structural variability.

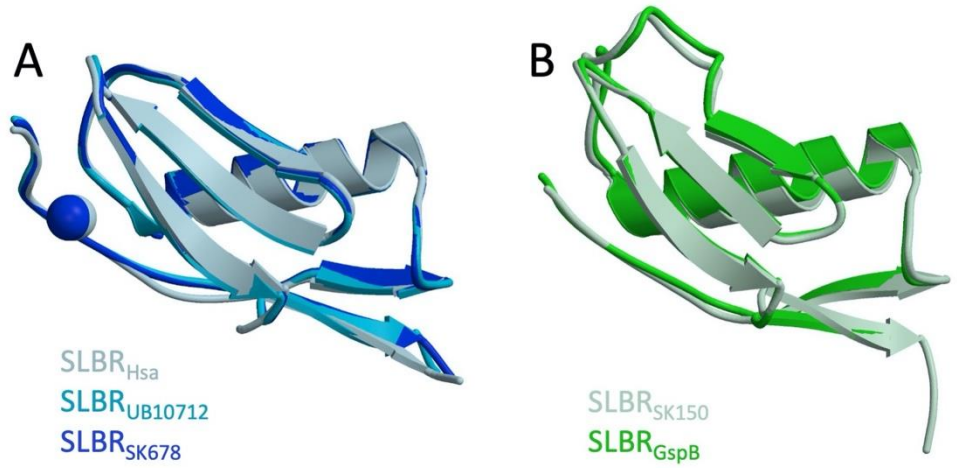

Figure iv-5 Comparison of the Unique domain of bacterial SLBRs.

Overlay of the Unique domains of: A. $SLBR_{Hsa}$, $SLBR_{UB10712}$ and $SLBR_{SK678}$ and B. $SLBR_{GspB}$ and $SLBR_{SK150}$. Ions are shown as spheres in the color corresponding to the SLBR. Based on the composition of the crystallization solution, the ions were tentatively assigned as Na^+ in $SLBR_{Hsa}$, Ca^{2+} in $SLBR_{SK678}$ and Ca^{2+} in $SLBR_{UB10712}$. The view is rotated as compared to Figure iv-4 in order to highlight the structural similarity of this domain across the different branches of the phylogenetic tree.

We next evaluated how these SLBRs interact with preferred versus disfavored ligands. Only the crystallization conditions for $SLBR_{Hsa}$ and the isolated Siglec domain of $SLBR_{GspB}$ ($SLBR_{GspB\text{-}Siglec}$) supported sialoglycan binding (

Table iv-3). For $SLBR_{Hsa}$, this included structures from crystals soaked with the high-affinity ligands sTa (Figure iv-1A, Figure iv-6A, Error! Reference source not found.) and sLe^C (Figure iv-1D, Figure iv-6B, Supplementary Figure 4), the intermediate-affinity ligand 3'sLn (Figure iv-1B, Figure iv-6C, Supplementary Figure 5), and the low-affinity ligand 6S-sLe^X (Figure iv-1C, Figure iv-6D, Supplementary Figure 6). The resolution ranged from 1.3 Å – 2.4 Å and the diffraction quality loosely correlated with ligand affinity (

Table iv-3). Cocrystals of $SLBR_{GspB\text{-}Siglec}$ with sTa diffracted to 1.25 Å resolution and the resultant maps contained unambiguous electron density for the sTa ligand (Figure iv-6E, Supplementary Figure 7,

Table iv-3). This structure is superior to a reported structure of $SLBR_{GspB}$ with sTa, where the low occupancy of the ligand made its assignment ambiguous[52].

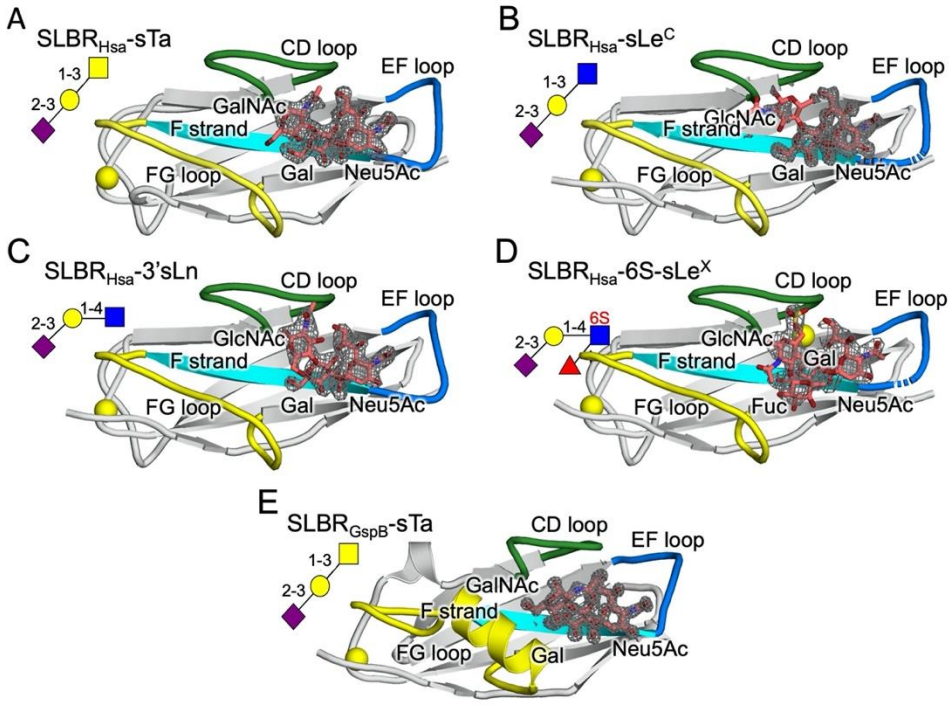

Figure iv-6 **Sialoglycans bound to $SLBR_{Hsa}$ and $SLBR_{GspB}$.**

A – D. $SLBR_{Hsa}$ bound to sialoglycans A. sTa, B. sLe^C, C. 3'sLn and D. 6S-sLe^X. E. $SLBR_{GspB\text{-}Siglec}$ bound to sTa. In each panel, the SLBR is shown as a cartoon with the CD, EF, and FG selectivity loops colored in *green*,

blue, and *yellow* respectively. The F strand contains the conserved YTRY motif and is shown in *cyan*. Ions are shown in *yellow* spheres. Carbon atoms of each sialoglycan are colored *salmon* with nitrogen shown in *blue* and oxygen in *red*. $|F_o| - |F_c|$ difference electron density calculated after removing the sialoglycan and performing three rounds of refinement in Phenix[76] are shown in grey mesh and contoured at 3σ. The standard depiction for each carbohydrate is shown in the upper left, with linkages indicated.

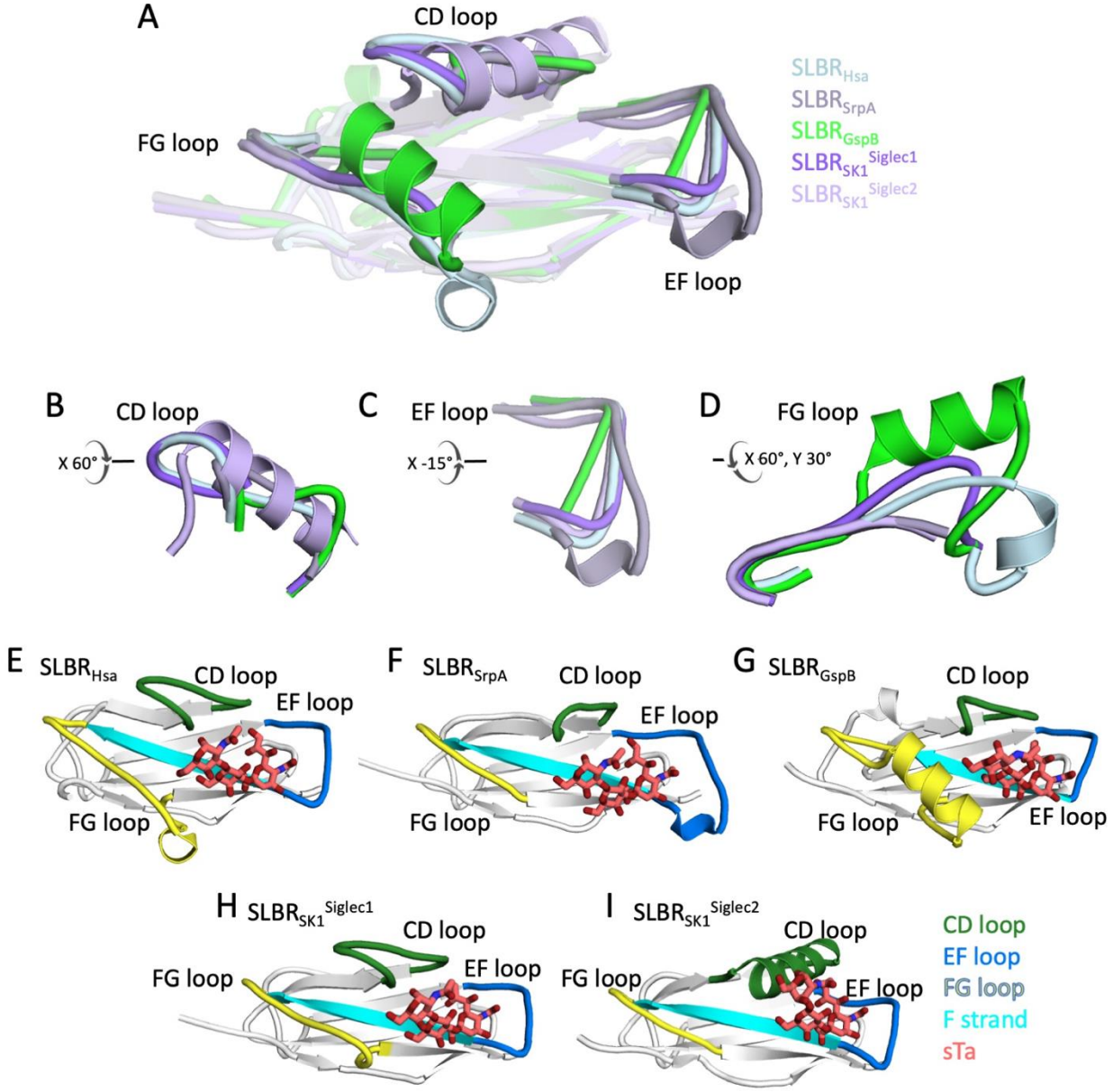

Figure iv-7 Selectivity loops in sTa-bound SLBRs.

Various SLBRs bound to sTa are shown in cartoon, and sTa is shown in *gray* sticks with oxygen colored *red* and nitrogen colored *blue*. A) Overlay of the sTa-bound SLBRs shown in E-F. B-D) overlays of the selectivity loops of the five sTa-bound SLBRs. E-G) SLBR$_{Hsa}$, SLBR$_{SrpA}$[49], and SLBR$_{GspB}$[52] are shown in *blue-gray*, *gray*, and *green* respectively. H-I) The SLBR$_{SK1(a)}$ and SLBR$_{SK1(b)}$[124] are shown in *purple* and *lavender* respectively.

	$SLBR_{Hsa}$ + sTa	**$SLBR_{Hsa}$ + 3'sLn**	**$SLBR_{Hsa}$ + 6S-sLe^X**	**$SLBR_{Hsa}$ + sLe^C**	**$SLBR_{GspB\text{-}Siglec}$ + sTa**
PDB entry	6EFD	6X3Q	6X3K	7KMJ	5IUC
DATAID	329	788	787	813	507
Resolution	1.85 Å	2.2 Å	2.47 Å	1.3 Å	1.25Å
Data collection					
Beamline	APS 21-ID-G	SSRL 9-2	SSRL 9-2	SSRL 9-2	21-ID-F
Wavelength	0.978 Å	0.979 Å	0.979 Å	0.979 Å	0.979 Å
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12$
Unit cell	a=46.7 Å b= 58.0 Å c=76.1 Å	a=44.9 Å b=57.1 Å c=76.3 Å	a=47.7 Å b=57.8 Å c=75.7 Å	a=46.6 Å b=58.1 Å c=76.0 Å	a=67.7 Å b=66.6 Å c=55.9 Å
R_{sym}	0.107 (0.638)	0.126 (0.643)	0.123 (0.740)	0.076 (0.696)	0.066 (0.406)
R_{pim}	0.037 (0.218)	0.055 (0.283)	0.053 (0.318)	0.025 (0.263)	0.018 (0.111)
I/σ	31.7 (2.9)	12.1 (1.5)	15.6 (1.7)	40.5 (1.6)	35.6 (7.9)
Completeness (%)	98.8 % (89.5%)	99.6% (96.4%)	98.7% (99.7%)	99.7% (98.5%)	95.7% (90.5%)
Redundancy	9.5 (8.7)	4.6 (4.8)	4.9 (5.1)	8.0 (6.7)	14.9 (14.3)
$CC_{1/2}$	0.940	0.993	0.989	0.998	0.964
Refinement					
R_{cryst}	0.196[&]	0.206	0.236	0.187	0.156
R_{free}	0.217	0.233	0.250	0.216	0.178
No. Mol per ASU	1	1	1	1	2
RMS deviation					
bond lengths	0.01 Å	0.02 Å	0.03 Å	0.01 Å	0.01 Å
bond angles	0.9°	2.4°	2.1°	1.63°	1.5°
Ramachandran					
favored	97.1%	95.1%	95.5%	97.0%	99.2%
allowed	2.9%	4.4%	4.5%	2.5%	0.8%
outliers	0.0%[*]	0.5%	0.0%	0.5%	0.0%

Table iv-3 Crystallographic data collection and refinement statistics for $SLBR_{Hsa}$ and $SLBR_{GspB\text{-}Siglec}$ bound to sialoglycans.

Values in parentheses are for the highest resolution shell. Raw data are deposited with SBGrid and can be accessed at: data.sbgrid.org/dataset/DATAID. Data collected at the APS used the Remote software, data collected at SSRL used the Blu-Ice software.

The sialoglycan-bound structures of $SLBR_{Hsa}$ and $SLBR_{GspB\text{-}Siglec}$ identifies that the sialic acid of all glycans binds above the ΦTRX motif in a similar way. This suggests that while the ΦTRX motif is important for binding, it does not select between potential ligands. More careful comparison suggests that the distinct

selectivity may originate from three loops of the V-set Ig fold that surround the sialoglycan binding site: the CD loop ($SLBR_{Hsa}^{284-296}$ or $SLBR_{GspB}^{440-453}$), the EF loop ($SLBR_{Hsa}^{330-336}$ or $SLBR_{GspB}^{475-481}$), and the FG loop ($SLBR_{Hsa}^{352-364}$ or $SLBR_{GspB}^{499-511}$) (Figure iv-6**,** Figure iv-7). Variation of both sequence and structure of SLBRs disproportionately maps to these loops (Figure iv-7**,** Figure iv-9). Moreover, temperature factor analysis suggests that these loops have high flexibility in the absence of ligand (Figure iv-10). Finally, molecular dynamics (MD) simulations of unliganded $SLBR_{Hsa}$ and $SLBR_{GspB}$ predict that these loops exhibit considerably more flexibility than other parts of the protein (Figure iv-8A, Figure iv-11). The MD further suggests that these loops sample the ligand-bound form even in the absence of sialoglycan. This supports a conformational selection mechanism, where structural change of the protein precedes binding of ligand[134]. The timing of ligand-associated conformational changes in enzymes affects fidelity[135] and may similarly contribute to ligand selectivity in binding proteins.

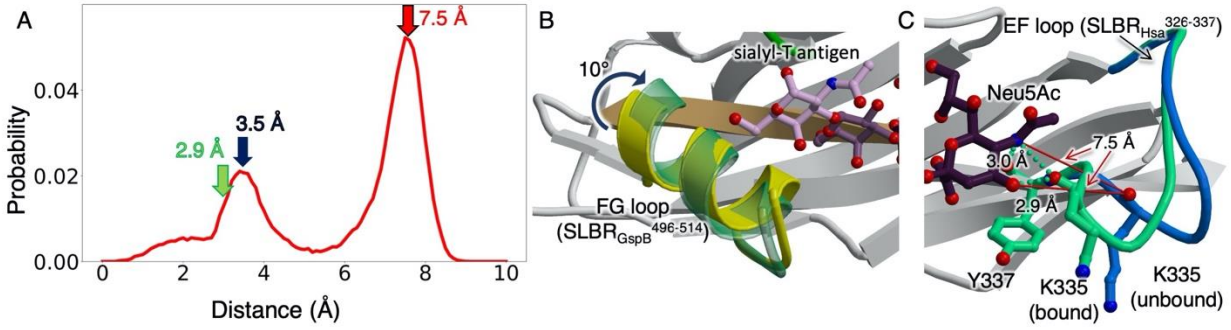

Figure iv-8 Conformations associated with SLBRs bound to sialoglycans.

Probability of distance distribution between the position of the Neu5Ac O4-hydroxyl in sTa and the $SLBR_{Hsa}^{K335}$ backbone carbonyl, as calculated by MD simulations. A bimodal distribution of distances exhibits maxima at 7.5 Å (*red* arrow), which reflects the unliganded crystal structure, and at 3.5 Å (*navy* arrow), which approaches the liganded crystal structure. The formation of the hydrogen-bond between the $SLBR_{Hsa}^{K335}$ carbonyl and Neu5Ac likely shifts the conformational equilibrium to a pose that supports the 2.9 Å distance (*light green* arrow) observed in the bound state. B. The FG loop of $SLBR_{GspB\text{-}Siglec}$ rotates 10° upon sTa binding. The position in the unbound structure is shown in *yellow* and the position in the bound structure is shown in *light green*. C. The EF loop of $SLBR_{Hsa}$ adjusts to promote formation of hydrogen-bonding interactions between $SLBR_{Hsa}^{K335}$ and the Neu5Ac of sTa. The position of this loop in the unbound structure is shown in *blue*, and the position occupied in the bound structure is shown in *light green*. The distance between the $SLBR_{Hsa}^{K335}$ backbone carbonyl and the position of the Neu5Ac O4-hydroxyl of the unliganded state are shown in *red lines* and match the 7.5 Å distance calculated by MD simulations (panel A). The distance between the $SLBR_{Hsa}^{K335}$ backbone carbonyl and the position of the Neu5Ac O4-hydroxyl is shown in *light green dots*.

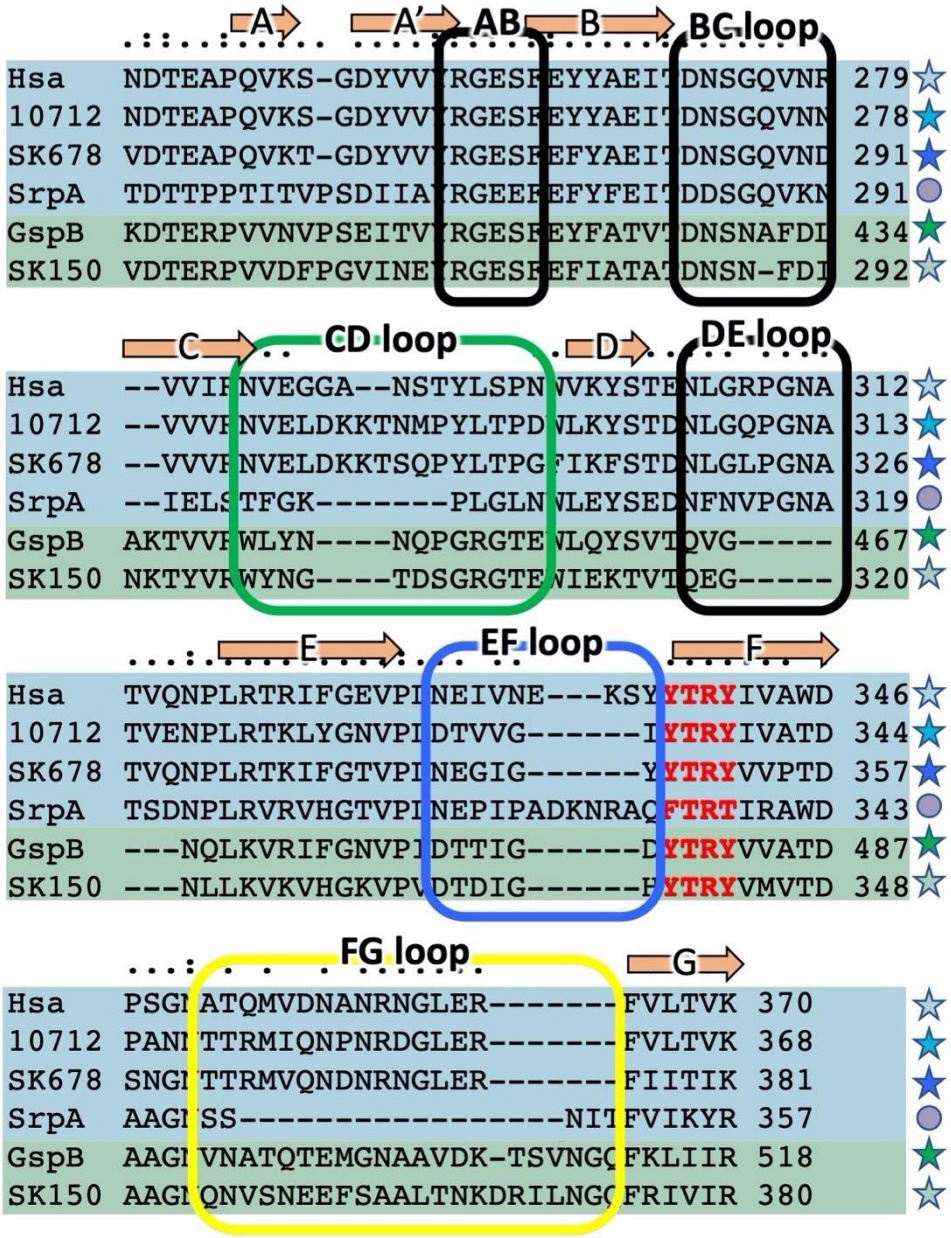

Figure iv-9 Sequence alignment of the Siglec domain of select SLBRs.

Sequences of the evolutionary-grouped SLBRs similar to $SLBR_{Hsa}$ are highlighted with a blue background. As comparators, sequences of $SLBR_{GspB}$ and an additional comparator in the same branch of the evolutionary tree, $SLBR_{SK150}$, are shown with a green background. Strands conserved in the V-set Ig fold are indicated, and residues of the interstrand loops are highlighted with boxes. The CD (*green*), EF (*blue*), and FG (*yellow*) loops disproportionately differ in length and homology. The ΦTRX motif on the F-strand of the Ig fold is highlighted with red text.

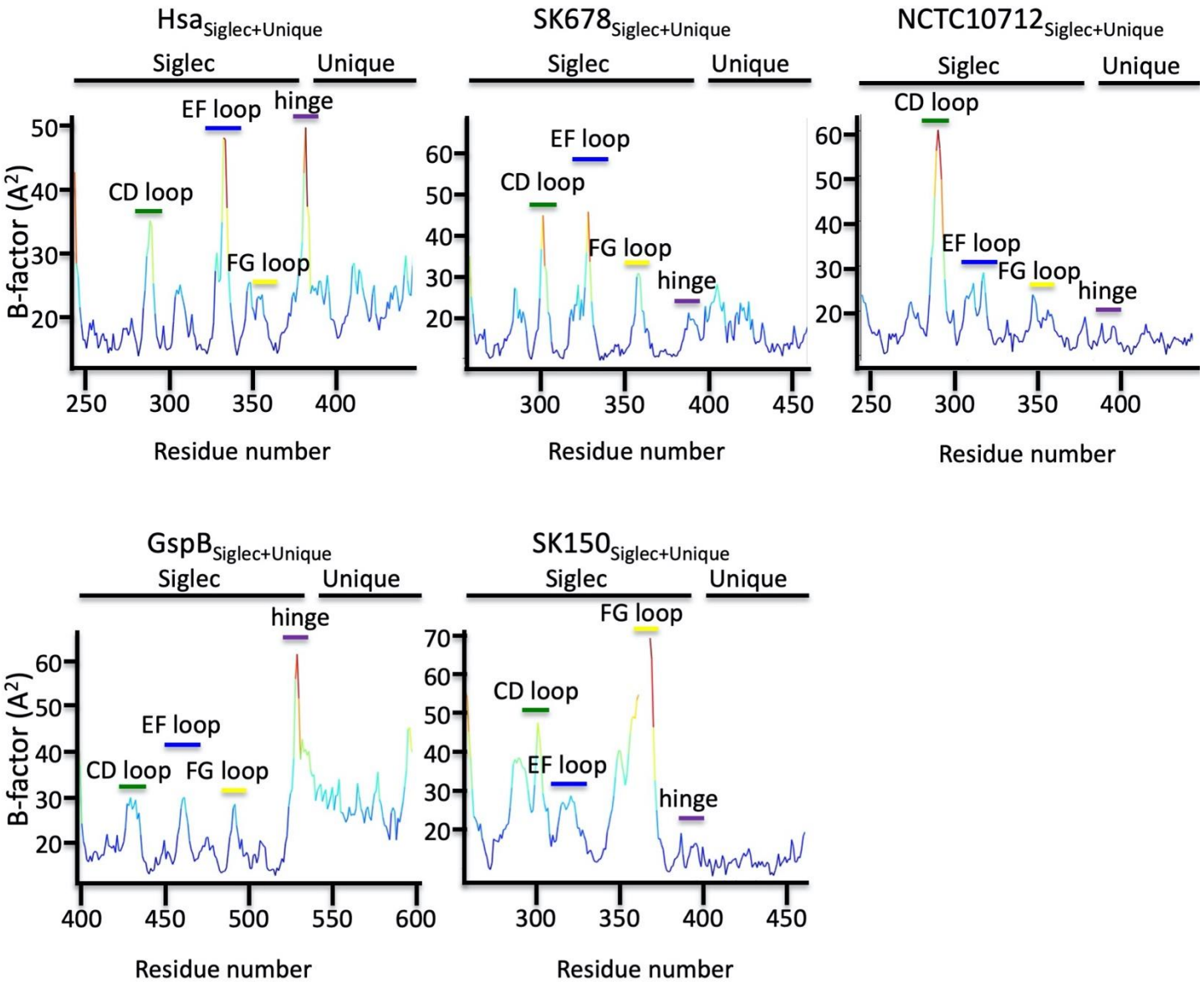

Figure iv-10 Temperature factor analysis of unliganded SLBR structures.

For each graph, the residue number is on the x-axis, and the crystallographic temperature factor (B-factor) is on the y-axis. Coloring is by relative B-factor. Regions with the lowest B-factors are predicted to have the lowest mobility (*dark blue*); regions with the highest B-factors are predicted to have the highest mobility (*red*).

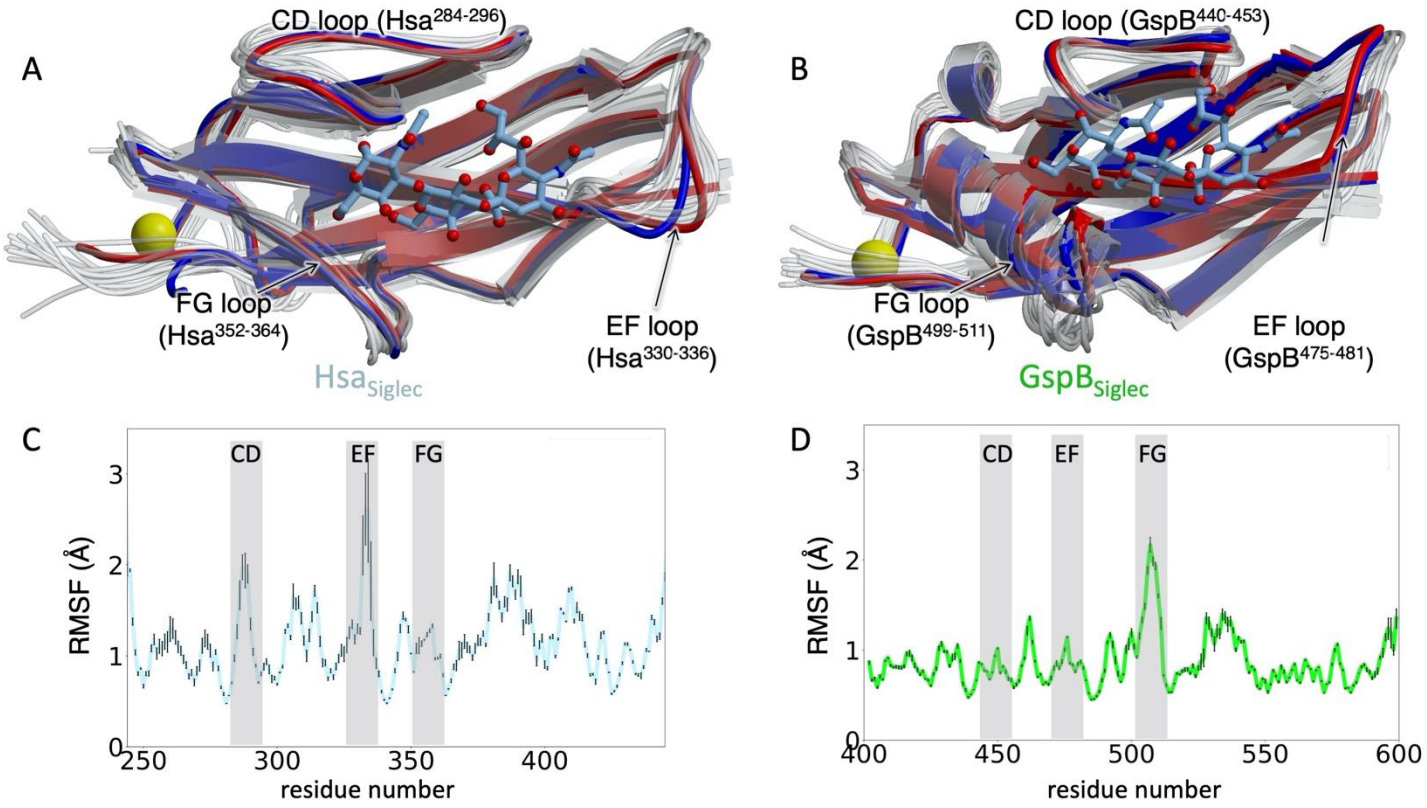

Figure iv-11 MD simulations of SLBRs.

A, B. Superposition of a representative subset of MD simulation snapshots (*translucent*) of A. The Siglec domain of $SLBR_{Hsa}$ and B. $SLBR_{GspB}$ onto the crystal structures determined in the presence (*blue*) and absence (*red*) of the sTa sialoglycan. MD simulations used structures determined in the absence of ligand as a starting point. C, D. Root mean square fluctuations (RMSF) of the Siglec domain of C. $SLBR_{Hsa}$ and D. $SLBR_{GspB}$ are plotted for each residue as compared to the average position of the Cα atoms of each residue. Calculations were performed on the adjacent Siglec and Unique domains, with only the resected Siglec domain shown. Error bars are shown in black lines at each residue and correspond to the standard error over 3 independent simulations. The positions of the CD, EF, and FG loops within the sequence are shown with a grey background. In $SLBR_{Hsa}$, the EF loop exhibits the largest predicted mobility of any region within the adhesin. In contrast, in $SLBR_{GspB}$, the FG loop exhibits the largest mobility.

Distinct loops in $SLBR_{Hsa}$ and $SLBR_{GspB}$ showed the largest conformational differences between the unbound and sialoglycan-bound structures. This provides the first hints into how narrow- versus broad-selectivity is conferred in this family. In the sTa-bound structure of $SLBR_{GspB\text{-}Siglec}$, the helix of the FG loop is rotated 10° as compared to the unliganded conformation. This rotation results in a maximum physical displacement of 1.3 Å (Figure iv-8B), which optimizes contacts to the GalNAc of sTa. Mechanistically, this would be consistent with

the conserved region of the glycan first interacting with a relatively pre-formed binding pocket comprised of the CD and EF loops prior to interaction with the FG loop.

In $SLBR_{Hsa}$, the conformation of the FG loop is similar in the presence and absence of glycan. Instead, comparing costructure determined with sTa with the costructure determined with sLe^C shows that the position of the EF loop differs by 5.9 Å (Figure iv-8C). This allows the $SLBR_{Hsa}{}^{K335}$ carbonyl to form hydrogen-bonding interactions to the invariant portion of the sialoglycans, i.e., the terminal Neu5Acα2-3Gal. In costructures determined with lower-affinity ligands, i.e., 3'sLn or 6S-sLe^X, this loop is not associated with clear electron density. This may result from crystal contacts to the EF loop that stabilize its position in the unliganded pose, resulting in a mixture of open and closed conformations (Figure iv-12). Comparison of the EF loop positions in the various crystal structures (Figure iv-4B**,** Figure iv-6**,** Figure iv-13A) with the positions calculated by the MD simulations (Figure iv-8C, Figure iv-11) suggests that closed conformation of the EF loop in the sTa and 3'sLn-bound crystal structures is likely the lowest energy state (Figure iv-11). Mechanistically, this suggests that for $SLBR_{Hsa}$, the variable, sub-terminal region of a sialoglycan ligand would first interact with the CD and FG loops. The ligand could then adjust in global position to optimize hydrogen-bonding interactions. The flexibility of the EF loop could then adapt to a range of different orientations of bound sialoglycan. This would be expected to promote broad selectivity. Thus, the location of inherent protein flexibility may define whether an SLBR is narrowly- versus broadly-selective.

To further evaluate how the broadly-selective $SLBR_{Hsa}$ could select for particular sialoglycans, we compared the binding positions of strong, intermediate, and weak ligands (Figure iv-13). In the strong and intermediate ligands, the invariant Neu5Acα2-3Gal effectively superimposes (Figure iv-13A, B) and has similar hydrogen bonds. Differences in the SLBR-ligand interactions predominantly map to the variable third sugar of the glycan (Figure iv-13C-F). Biding strength may therefore be related to these interactions. In contrast, the global binding

position of the weak ligand 6S-sLe^X is shifted as compared with all other ligands (Figure iv-6D, Figure iv-13B, F). This affects the hydrogen bonds along the entirety of the ligand.

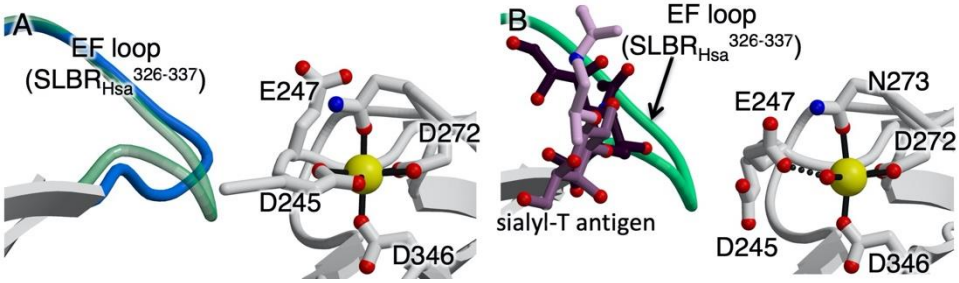

Figure iv-12 Crystal packing and conformational change upon ligand binding in SLBR$_{Hsa}$.

Crystal contact between the EF loop of unliganded SLBR$_{Hsa}$ (*blue*) and the N-terminus of a neighboring molecule. The position of the loop in SLBR$_{Hsa}$ bound to sTa (*transparent green*) would be in steric conflict with the N-terminus of the adjacent molecule in the absence of a conformational change. B. Change in crystal contact following binding to sTa. When the EF loop closes over sTa, the N-terminus undergoes a compensatory conformational adjustment that changes the coordination sphere of a labile cation in the neighboring molecule. Specifically, the main chain of SLBR$_{Hsa}^{D245}$ normally coordinates the ion but would be in steric conflict with ligand-bound position of the EF loop. Following the conformational change, SLBR$_{Hsa}^{E247}$ now coordinates the ion. This crystal contact likely creates an energy minimum and shifts the conformational equilibrium of the EF loop toward the open position, even in the presence of glycan. Adjustment of the EF loop to ligand is observed only in a subset of the costructures, but it is expected to close over ligand when in solution.

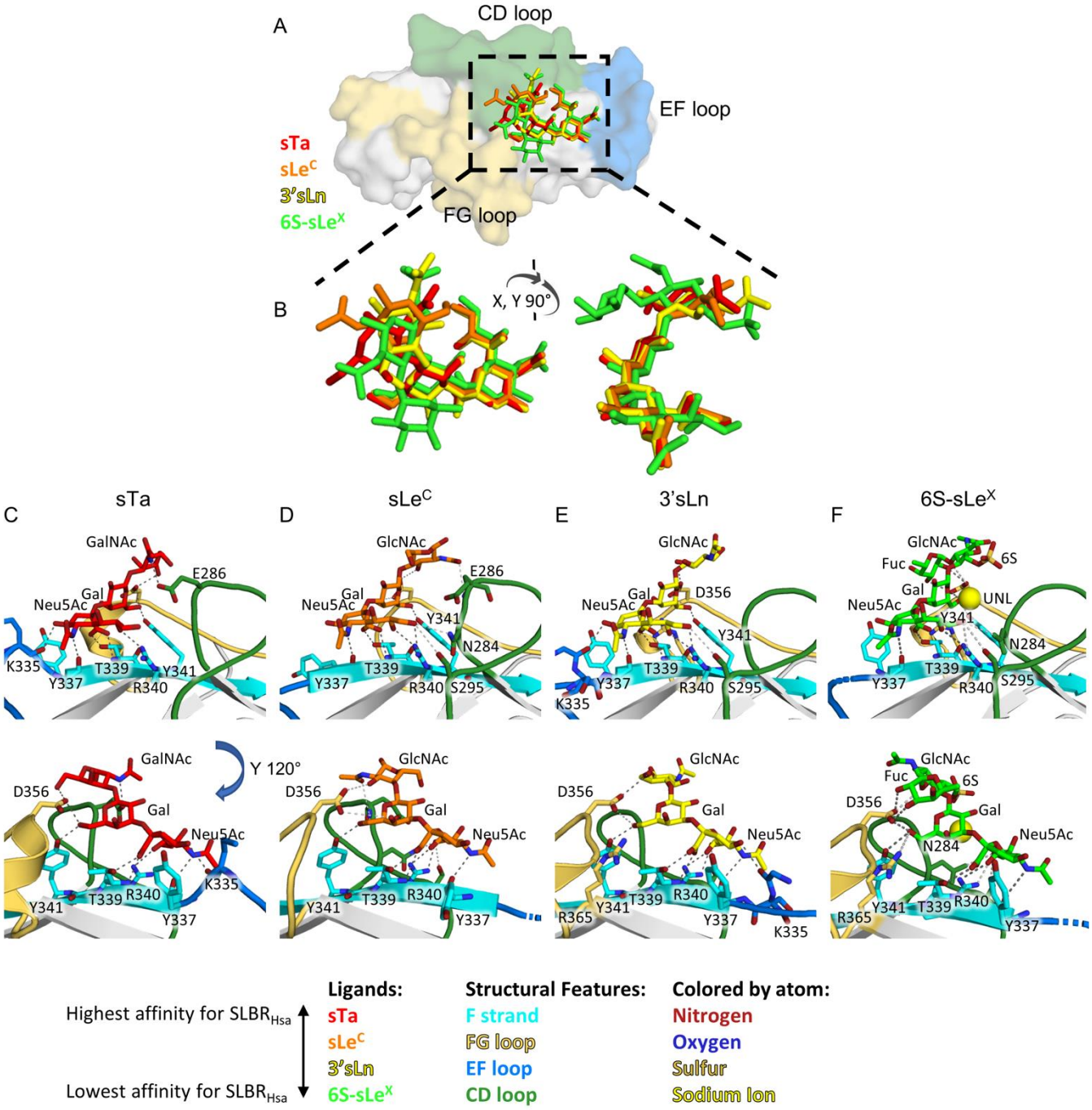

Figure iv-13 Sialoglycan position in the SLBR$_{Hsa}$ binding pocket.

A-F. The sialoglycan ligands sTa, sLe^C, 3'sLn, and 6S-sLe^X are shown in *red*, *orange*, *yellow*, and *green*, where *red* is sTa, which is the highest affinity ligand for SLBR$_{Hsa}$, and *green* is 6S-sLe^X the lowest affinity ligand for SLBR$_{Hsa}$ used in this study. A. sTa-bound SLBR$_{Hsa}$ is shown as a *grey* surface. B. Close-up view of the overlaid sialoglycan ligands. The position of 6S-sLe^X is shifted by ~0.8Å from the position of the highest affinity ligand, sTa. C-F. The binding pocket of SLBR$_{Hsa}$ is shown in cartoon with the CD loop colored *green*, the EF loop colored *blue*, and the FG loop colored *yellow*. The F strand containing the ϕTRX motif is shown in cyan. The unknown ligand (UNL) in the 6S-sLe^X structure is shown as a *yellow sphere*. The sialoglycan ligands and residues that participate in hydrogen bonding with the ligands are shown in sticks with nitrogen shown in *blue*

and oxygen shown in *red*. Hydrogen bonds and electrostatic interactions between $SLBR_{Hsa}$ and the sialoglycans are shown as *dark grey* and *light grey* dashed lines respectively.

6S-sLe^X is both α1,3-fucosylated and *O*-sulfated at the C6 (6S) of the GlcNAc, modifications that are absent in the strong $SLBR_{Hsa}$ ligands (Figure iv-14C). The evaluation of the interactions between these groups and $SLBR_{Hsa}$ suggests how related SLBRs include or exclude these elaborations. In considering how the α1,3-fucose in glycans such as sLe^X and 6S-sLe^X is excluded from $SLBR_{Hsa}$, our analysis suggests that the β-branching of $SLBR_{Hsa}^{D356}$ on the FG loop disfavors the binding of a fucosylated glycan (Figure iv-13C-F). MD simulations also indicate that the FG loop does not sample a position that allows an extra fucose or other large elaboration at this position (Figure iv-11). This is consistent with the crystal structure, which shows that the loop position does not allow 6S-sLe^X to sit optimally in the sialoglycan binding site.

In considering how a 6S group might be included or excluded, the structure reveals that $SLBR_{Hsa}^{E286}$ of the CD loop contacts the sulfate of 6S-sLe^X. This does not exclude a 6S group *per se*, but both are negatively charged. The structure suggests that an unknown ligand, possibly a component of the buffer, binds near this site to bridge the interaction (Figure iv-6D**,** Figure iv-13F). Taken together, these structural and computational analyses show that steric and electrostatic interactions of the broadly selective $SLBR_{Hsa}$ exclude specific structural additions to the glycan ligands.

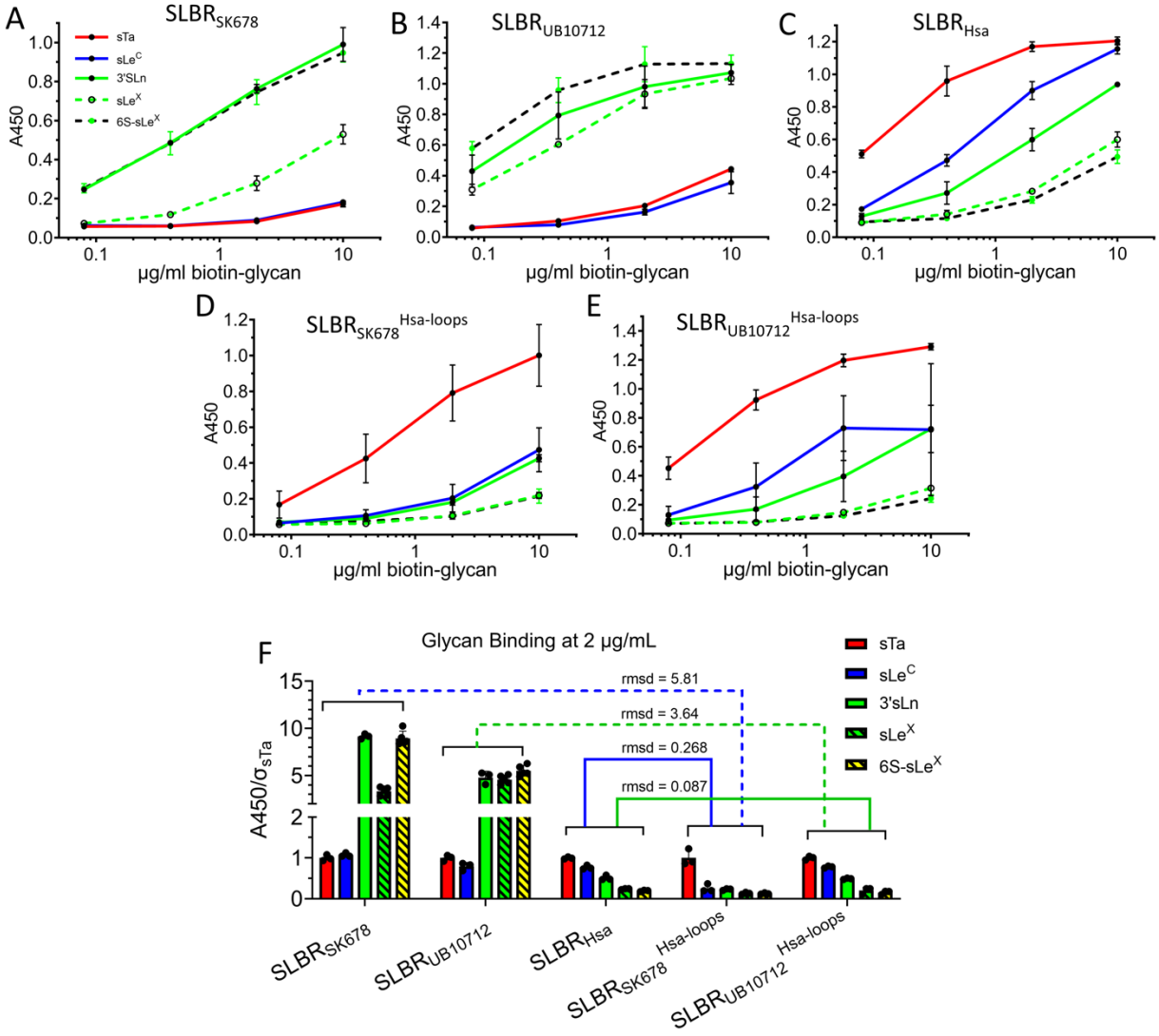

Figure iv-14 Chimeragenesis of SLBR$_{Hsa}$ and its close homologs.

Dose-response curves of A. wild-type GST-SLBR$_{SK678}$, B. wild-type GST-SLBR$_{UB10712}$, and C. wild-type GST-SLBR$_{Hsa}$ to five selected ligands. D and E. Dose-response curves of the chimeras D. GST-SLBR$_{SK678}$$^{Hsa\text{-}loops}$ and E. GST-SLBR$_{UB10712}$$^{Hsa\text{-}loops}$ which contain the CD, EF, and FG loops of SLBR$_{Hsa}$. F. Quantitation of bound glycans at a concentration of 2µg/ml to parent and chimeric SLBRs. Individual datapoints are shown. This identifies that both GST-SLBR$_{SK678}$$^{Hsa\text{-}loops}$ and GST-SLBR$_{UB10712}$$^{Hsa\text{-}loops}$ bind to sTa more strongly than the parent SLBRs. In addition, the chimeric SLBRs now have a preference for glycans more similar to SLBR$_{Hsa}$. Specifically, wild-type SLBR$_{SK678}$ and $_{UB10712}$ bind most strongly to 6S-sLe^X/3'sLn > sLe^X. In contrast, SLBR$_{Hsa}$ and SLBR$_{UB10712}$$^{Hsa\text{-}loops}$ bind sTa > sLe^C > 3'sLn > sLe^X > 6S-sLe^X while SLBR$_{SK678}$$^{Hsa\text{-}loops}$ bound sTa > sLe^C/3'sLn > sLe^X/6S-sLe^X Measurements were performed using 500 nM of immobilized GST-SLBR and the indicated concentrations of each ligand and are shown as the mean ± SD (n = 3 independent experiments with a single protein preparation).

The CD, EF, and FG loops determine SLBR selectivity.

Because structural studies suggest that the combined action of the CD, EF, and FG loops allow SLBRs to select between ligands, we developed chimeras with the backbone of one SLBR and the loops of a closely related SLBR. We first replaced the CD, EF, and FG loops of $SLBR_{SK678}$ and $SLBR_{UB10712}$ with the equivalent loops from $SLBR_{Hsa}$ to create the $SLBR_{SK678}^{Hsa\text{-}loops}$ and $SLBR_{UB10712}^{Hsa\text{-}loops}$ chimeras. MD simulations would suggest that the loops retain the structure found within the parent $SLBR_{Hsa}$ (Figure iv-15). Using physiologically-relevant sialoglycans[69, 70, 127], we measured binding to parent and chimeric SLBRs in ELISAs (Figure iv-14A-E). We found that these chimeras bound glycans strongly and had a sialoglycan-binding preference that closely resembled $SLBR_{Hsa}$ rather than the parent SLBR (Figure iv-14F, Table iv-4). This change in selectivity occurred via both a gain-of-function that promoted binding to sTa and a loss-of-function that decreased binding to α1,3-fucosylated and *O*-sulfated sialoglycans. This change of binding spectrum confirms that a major determinant of selectivity in these SLBRs is the loops that surround the ligand-binding pocket.

	sTa	sLe^c	3'sLn	sLe^x	6S-sLe^x
GST-$SLBR_{Hsa}$[a]	0.12 ± 0.06	0.67 ± 0.18	1.61 ± 1.44	>5	>5
E286R[a]	0.11 ± 0.08	0.06 ± 0.08	0.11 ± 0.10	0.30 ± 0.17	0.01 ± 0.12[c]
D356Q[a]	0.55 ± 0.22	0.81 ± 0.52	0.17 ± 0.05	0.68 ± 0.36	0.00 ± 0.04[c]
D356R[a]	1.93 ± 2.21	0.74 ± 1.40	0.26 ± 0.09	2.62 ± 5.12	0.22 ± 0.05
GST-$SLBR_{UB10712}$[a]	>5	>5	0.13 ± 0.19	0.63 ± 0.18	0.14 ± 0.10
E285R[a]	>5	>5	1.40 ± 0.91	>5	0.00 ± 0.05[c]
Q354D[a]	1.89 ± 1.04	>5	0.58 ± 0.41	>5	>5
+ $SLBR_{Hsa}$ CD loop[b]	+	-	++	+	+++
+ $SLBR_{Hsa}$ EF loop[b]	+	+	+++	+++	+++
+ $SLBR_{Hsa}$ FG loop[b]	++	+	++	+	+
+ all $SLBR_{Hsa}$ loops[a]	0.17 ± 0.08	1.28 ± 2.36	>5	>5	>5
GST-$SLBR_{SK678}$[a]	>5	>5	0.90 ± 0.51	>5	0.63 ± 0.36
E298R	>5	>5	2.43 ± 1.66	>5	0.07 ± 0.09
Q367D[a]	>5	>5	>5	>5	>5
+ $SLBR_{Hsa}$ CD loop[b]	-	-	-	-	-
+ $SLBR_{Hsa}$ EF loop[b]	-	-	++	+	++

+ $SLBR_{Hsa}$ FG loop[b]	-	-	+	-	-
+ all $SLBR_{Hsa}$ loops[a]	0.67 ± 0.78	>5	>5	>5	>5
GST-$SLBR_{GspB}$[a]	0.08 ± 0.05	>5	>5	nd	nd
L442Y/Y443N[a]	0.31 ± 0.17	>5	0.15 ± 0.07	nd	nd
+ $SLBR_{SK150}$ CD loop[b]	-	nd	-	nd	nd
+ $SLBR_{SK150}$ EF loop[b]	+++	nd	-	nd	nd
+ $SLBR_{SK150}$ FG loop[b]	-	nd	-	nd	nd
+ all $SLBR_{SK150}$ loops[b]	-	nd	-	nd	nd
GST-$SLBR_{SK150}$[a]	1.09 ± 0.32	4.74 ± 6.31	2.91 ± 3.59	>5	>5
Y300L/N301Y[a]	>5	nd	>5	nd	nd

Table iv-4 **Summary of binding preferences of wild-type and variant SLBRs.**

Numbers reflect EC_{50} values while +/- designations are indicators of relative binding strength from one-point analysis. Abbreviations: sTa, sialyl T antigen; 3'sLn, 3'-sialyl-*N*-acetyllactosamine; sLe^C, sialyl $Lewis^C$; sLe^X, sialyl $Lewis^X$; 6S-sLe^X, 6-*O*-sulfo-sialyl $Lewis^X$; nd= not determined.

[a]EC_{50} values (µg/ml) ± standard errors were obtained via nonlinear regression of the ELISA curves in Figure iv-3, Figure iv-19, Figure iv-22, Figure iv-14, Figure iv-16 Figure iv-18, and Figure iv-21, using Prism 7 (GraphPad).

[b]Relative binding strengths are based on absorbance values obtained using 1-2 µg/ml biotinylated glycans. +++, A_{450} > 1; ++, A_{450} = 0.7-1; +, A_{450} = 0.3-0.7; -, A_{450} < 0.3.

[c]Approximate value

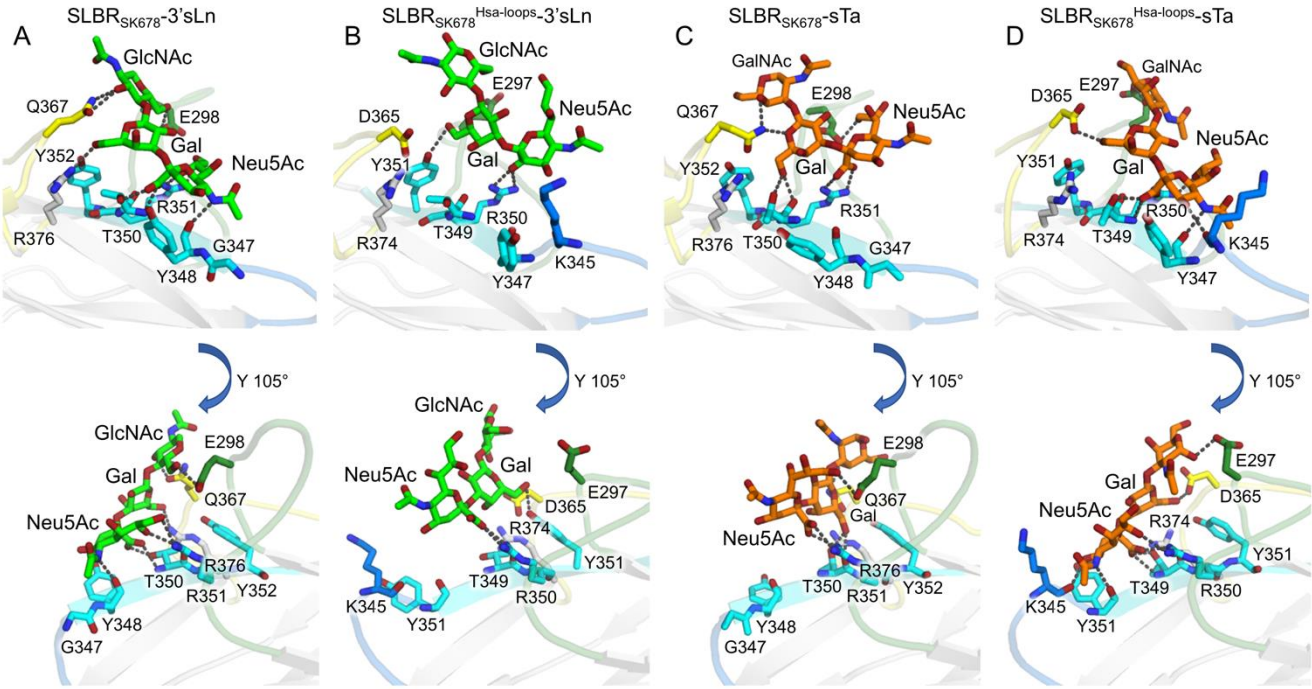

Figure iv-15 Computational analysis of SLBR$_{SK678}$$^{Hsa-loops}$ binding to ligands.

A-D. computational structures of 3'sLn and sTa bound SLBR$_{SK678}$ and the SLBR$_{SK678}$$^{Hsa-loops}$ chimera. A-B. The loops of the SLBR$_{SK678}$$^{Hsa-loops}$ chimera are predicted to make additional interactions between the F strand and sTa when compared to the parent SLBR$_{SK678}$. C-D. The SLBR$_{SK678}$$^{Hsa-loops}$ chimera is predicted to lose interactions between the EF loop, CD loop, and the F strand; and 3'sLn when compared to parent SLBR$_{SK678}$. The lower affinity complexes, A. sTa bound SLBR$_{SK678}$ and D. 3'sLn bound SLBR$_{SK678}$$^{Hsa-loops}$ both lack interactions between the C-terminal end of the F strand and the sialic acid moiety.

We next assessed the contributions of each loop to selectivity (Figure iv-16). Substitution of the EF loop of SLBR$_{SK678}$ with the EF loop from SLBR$_{Hsa}$ resulted in increased binding to sTa, sLe^{C}, sLe^{X}, and 6S-sLe^{X} (Figure iv-16). This result is consistent with the structural prediction that a SLBR with a flexible EF loop can potentially accommodate a greater range of ligands.

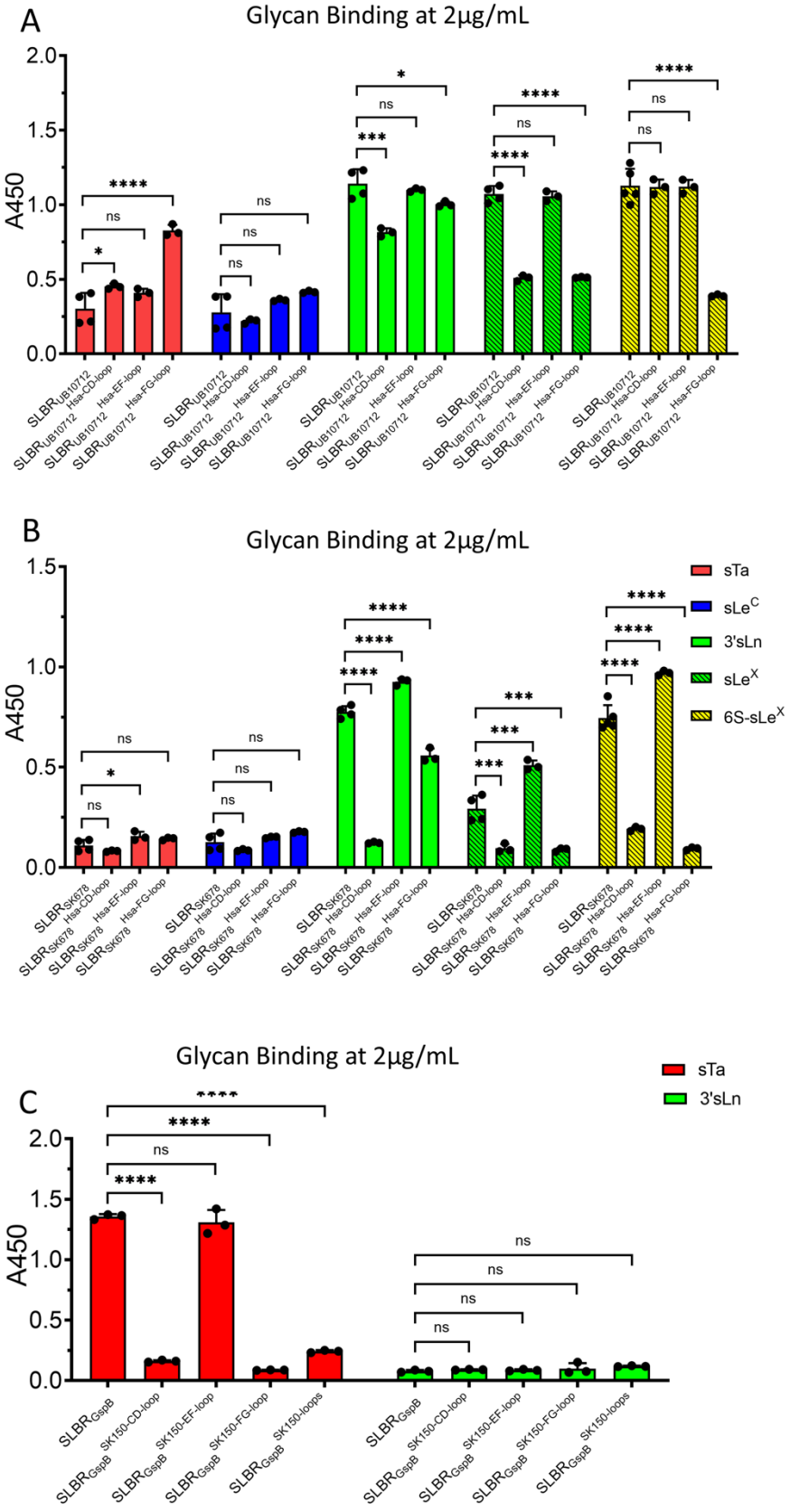

Figure iv-16 Quantitation of glycan binding by the single-loop chimeras of SLBRs.

Binding of biotin-glycans (2 µg/ml) to A. GST-$SLBR_{UB10712}$ containing loops CD, EF, or FG of $SLBR_{Hsa}$, substituted individually; and B. GST-$SLBR_{SK678}$ containing loops CD, EF, or FG of $SLBR_{Hsa}$, substituted individually. A. The p values for the comparison of binding of sTa to $SLBR_{UB10712}$ and variants $SLBR_{UB10712}^{Hsa\text{-}CD\text{-}loop}$, $SLBR_{UB10712}^{Hsa\text{-}EF\text{-}loop}$, and $SLBR_{UB10712}^{Hsa\text{-}FG\text{-}loop}$ are 0.0404, 0.155, and <0.0001 respectively. The p values for the comparison of binding of sLe^C to $SLBR_{UB10712}$ and variants $SLBR_{UB10712}^{Hsa\text{-}CD\text{-}loop}$, $SLBR_{UB10712}^{Hsa\text{-}EF\text{-}loop}$, and $SLBR_{UB10712}^{Hsa\text{-}FG\text{-}loop}$ are 0.599, 0.361, and 0.0811 respectively. The p values for the comparison of binding of 3'sLn to $SLBR_{UB10712}$ and variants $SLBR_{UB10712}^{Hsa\text{-}CD\text{-}loop}$, $SLBR_{UB10712}^{Hsa\text{-}EF\text{-}loop}$, and $SLBR_{UB10712}^{Hsa\text{-}FG\text{-}loop}$ are 0.0001, 0.703, and 0.350 respectively. The p values for the comparison of binding of sLe^X to $SLBR_{UB10712}$ and variants $SLBR_{UB10712}^{Hsa\text{-}CD\text{-}loop}$, $SLBR_{UB10712}^{Hsa\text{-}EF\text{-}loop}$, and $SLBR_{UB10712}^{Hsa\text{-}FG\text{-}loop}$ are <0.0001, 0.924, and <0.0001 respectively. The p values for the comparison of binding of 6S-sLe^X to $SLBR_{UB10712}$ and variants $SLBR_{UB10712}^{Hsa\text{-}CD\text{-}loop}$, $SLBR_{UB10712}^{Hsa\text{-}EF\text{-}loop}$, and $SLBR_{UB10712}^{Hsa\text{-}FG\text{-}loop}$ are 0.998, 0.999, and <0.0001 respectively. B. The p values for the comparison of binding of sTa to $SLBR_{SK678}$ and variants $SK678^{Hsa\text{-}CD\text{-}loop}$, $SK678^{Hsa\text{-}EF\text{-}loop}$, and $SK678^{Hsa\text{-}FG\text{-}loop}$ are 0.226, 0.0351, and 0.131 respectively. The p values for the comparison of binding of sLe^C to $SLBR_{SK678}$ and variants $SLBR_{SK678}^{Hsa\text{-}CD\text{-}loop}$, $SLBR_{SK678}^{Hsa\text{-}EF\text{-}loop}$, and $SLBR_{SK678}^{Hsa\text{-}FG\text{-}loop}$ are 0.158, 0.471, and 0.0626. The p values for the comparison of binding of 3'sLn to $SLBR_{SK678}$ and variants $SLBR_{SK678}^{Hsa\text{-}CD\text{-}loop}$, $SLBR_{SK678}^{Hsa\text{-}EF\text{-}loop}$, and $SLBR_{SK678}^{Hsa\text{-}FG\text{-}loop}$ are all <0.0001. The p values for the comparison of binding of sLe^X to $SLBR_{SK678}$ and variants $SLBR_{SK678}^{Hsa\text{-}CD\text{-}loop}$, $SLBR_{SK678}^{Hsa\text{-}EF\text{-}loop}$, and $SLBR_{SK678}^{Hsa\text{-}FG\text{-}loop}$ are 0.0004, 0.0002, and 0.0003 respectively. The p values for the comparison of binding of 6S-sLe^X to $SLBR_{SK678}$ and variants $SLBR_{SK678}^{Hsa\text{-}CD\text{-}loop}$, $SLBR_{SK678}^{Hsa\text{-}EF\text{-}loop}$, and $SLBR_{SK678}^{Hsa\text{-}FG\text{-}loop}$ are all <0.0001. $SLBR_{SK678}^{Hsa\text{-}CD\text{-}loop}$ exhibited substantially decreased affinity for 3'sLn and 6S-sLe^X, $SLBR_{SK678}^{Hsa\text{-}EF\text{-}loop}$ had an increase in binding to all tested ligands except for sLe^C. $SK678^{Hsa\text{-}FG\text{-}loop}$ had a moderately decreased affinity for 3'sLn, and a substantially decreased affinity for sLe^X and 6S-sLe^X. $SLBR_{UB10712}^{Hsa\text{-}CD\text{-}loop}$ exhibited a moderate loss of affinity to 3'sLn and sLe^X and a moderate gain of affinity to sTa; $SLBR_{UB10712}^{Hsa\text{-}FG\text{-}loop}$ exhibited a decrease in affinity to sLe^X and 6S-sLe^X and a moderate increase of affinity to sTa. C. Binding of biotin-glycans (1 µg/ml) to GST-$SLBR_{GspB}$ containing loops CD, EF, or FG of $SLBR_{SK150}$ substituted either individually or together. The p values comparing binding of sTa to the parent $SLBR_{GspB}$ and the variants $SLBR_{GspB}^{SK150\text{-}CD\text{-}loop}$, $SLBR_{GspB}^{SK150\text{-}EF\text{-}loop}$, $SLBR_{GspB}^{SK150\text{-}FG\text{-}loop}$, and $SLBR_{GspB}^{SK150\text{-}loops}$ are <0.0001, 0.555, <0.0001, and <0.0001. The p values comparing binding of 3'sLn to the parent $SLBR_{GspB}$ and the variants $SLBR_{GspB}^{SK150\text{-}CD\text{-}loop}$, $SLBR_{GspB}^{SK150\text{-}EF\text{-}loop}$, $SLBR_{GspB}^{SK150\text{-}FG\text{-}loop}$, and $SLBR_{GspB}^{SK150\text{-}loops}$ are 0.881, 0.973, 0.623, and 0.122 respectively. With the exception of the $SLBR_{GspB}^{SK150\text{-}EF\text{-}loop}$, all chimeras showed a substantial reduction in binding. Values correspond to the mean ± standard deviation, with n = 3 independent experiments using a single protein preparation. Binding of each glycan to each mutant was statistically compared to binding of the same glycan to the $SLBR^{WT}$ with an ordinary one-way ANOVA corrected for multiple comparisons with Dunnett's test. Statistical significance is indicated by "*" (ns P >0.05; *P <0.05; **P < 0.01; ***P < 0.001; **** < 0.0001)

In contrast, substitution of the CD or FG loops altered the identity of the preferred ligands. The altered selectivity of these chimeras involved a combination of reduced binding to some sialoglycans and increased binding to others, i.e., both a loss-of-function and a gain-of-function. For example, both $SLBR_{SK678}^{Hsa\text{-}FG\text{-}loop}$ and $SLBR_{UB10712}^{Hsa\text{-}FG\text{-}loop}$ exhibited decreased binding to the fucosylated ligands sLe^X and 6S-sLe^X while

$SLBR_{UB10712}^{Hsa\text{-}FG\text{-}loop}$ also increased binding to sTa (Figure iv-15A, B). This is consistent with the crystallographic interpretation that $SLBR_{Hsa}^{D356}$ on the FG loop restricts accommodation of Fucα1-3GlcNAc.

The single-loop chimeras also suggest synergy between these three selectivity loops. For example, the substantial decrease in binding of $SLBR_{SK678}^{Hsa\text{-}CD\text{-}loop}$ to 6S-sLe^X (Figure iv-16B) is consistent with a proposal that the binding of 6S-ligands is controlled by the CD loop. However, the $SLBR_{UB10712}^{Hsa\text{-}CD\text{-}loop}$ chimera retains robust binding to 6S-sLe^X (Figure iv-16A) suggesting that the other loops moderate the effects.

We next turned to $SLBR_{GspB}$ and $SLBR_{SK150}$, which both bind sTa preferentially (Supplementary Figure 1). Here, we substituted the loops of $SLBR_{SK150}$ into $SLBR_{GspB}$ and assessed the binding to sTa and 3'sLn, which are the ligands with the highest affinity for $SLBR_{SK150}$. In contrast to the results observed with $SLBR_{Hsa}$ and its close homologs, substitution of the EF loop of $SLBR_{SK150}$ into $SLBR_{GspB}$ had little impact (Figure iv-16C). In all remaining chimeras, there was little detectable binding to sTa or 3'sLn (Figure iv-16C). To determine whether protein misfolding may be a contributing factor in variants with loss of binding, we used size exclusion chromatography (Figure iv-17A-C), which can distinguish between folded and mis-folded SLBRs[49]. The chromatogram of the $SLBR_{GspB}^{SK150\text{-}loops}$ showed a monodisperse peak with little aggregation, indicating that loss of binding in this case was not due to misfolding. However, the chromatograms of the $SLBR_{GspB}^{SK150\text{-}CD\text{-}loops}$ and $SLBR_{GspB}^{SK15\text{-}FG\text{-}loops}$ chimeras showed significant levels of protein aggregates and break-down products, indicating that misfolding may contribute to loss of binding for these two variants.

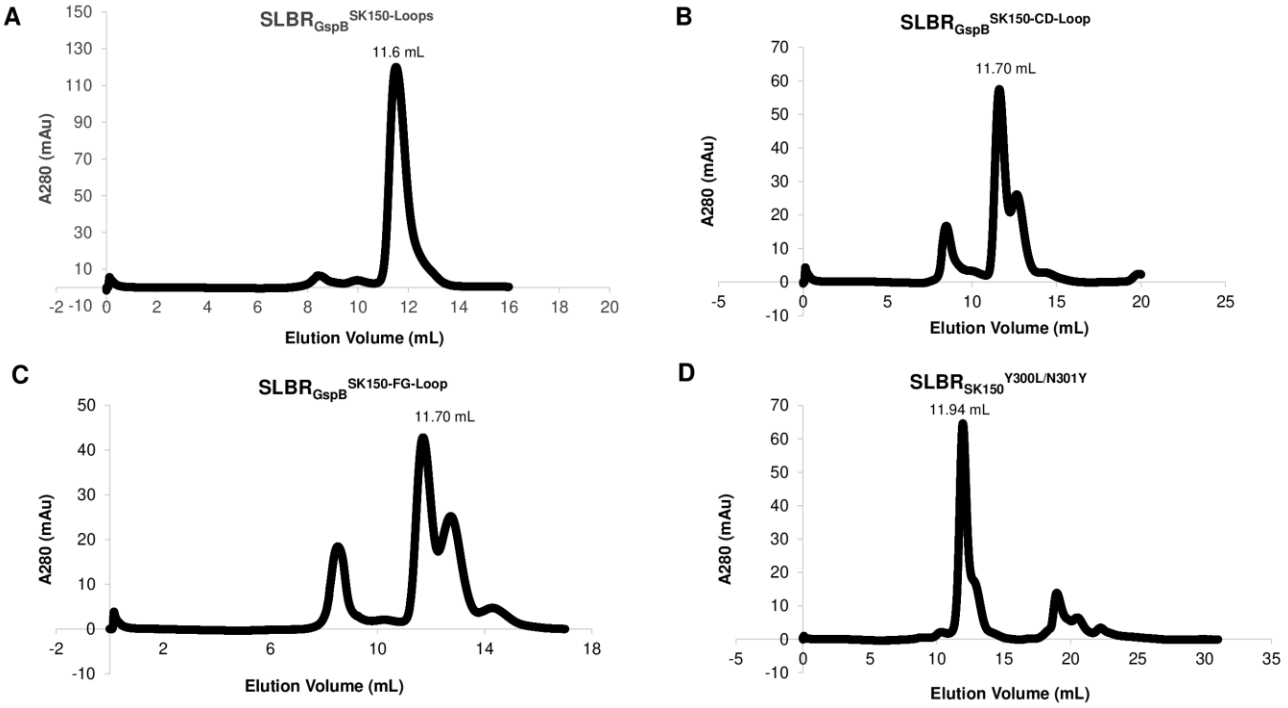

Figure iv-17 Evaluation of protein folding in non-binding variant SLBR$_{GspB}$ and SLBR$_{SK150}$.

For the GST-SLBRs that exhibited no binding to any tested glycan, folding was evaluated by observing the properties of the 0.5 mg non-binding variant SLBR on a 24-mL Superdex S200 size exclusion chromatography column. A. GST-SLBR$_{GspB}^{SK150\text{-}loops}$ appears to be monodisperse and contains a single Gaussian peak with an elution volume of 11.60 mL. This profile is consistent with folded protein. B. GST-SLBR$_{GspB}^{SK150\text{-}CD\text{-}loops}$ is a non-binding variant that exhibits loss of the peak corresponding to the folded protein, the appearance of an aggregate peak at 8.5 mL, and the appearance of a breakdown product at 12.75 mL. This is consistent with a deficit in either folding or stability. C. SLBR$_{GspB}^{SK150\text{-}FG\text{-}loop}$ similarly exhibits a loss in the 11.6 mL peak and the appearance of both an aggregate peak and breakdown products. D. The profile of SLBR$_{SK150}^{Y300L/N301Y}$ exhibits loss of the folded peak and appearance of breakdown products but does not contain aggregates. B-D. Decreases in sialoglycan binding of these variants as compared to wildtype (see Figure iv-16C and Figure iv-18F) may be due to the misfolding and stability loss observed here.

The ability to develop functional chimeras for the three SLBR$_{Hsa}$-like adhesins, but not the two SLBR$_{GspB}$-like adhesins, might be explained in several ways. First, the broadly-selective scaffolds of SLBR$_{Hsa}$, SLBR$_{SK678}$, and SLBR$_{UB10712}$ may have more plasticity, allowing these to better accommodate non-native loops. Conversely, the broadly-selective SLBRs may contain somewhat more flexible loops that more easily adjust to the non-native scaffold. Finally, the sequence identity between SLBR$_{Hsa}$, SLBR$_{SK678}$, and SLBR$_{UB10712}$ is higher than that between SLBR$_{GspB}$ and SLBR$_{SK150}$, allowing a better fit between the scaffold and chimeric loops in the SLBR$_{Hsa}$-like proteins. To better understand why SLBR$_{Hsa}$-like proteins were more mutable, we leveraged our

crystal structure of $SLBR_{GspB}$ in complex with sTa (Figure iv-6E) and identified that $SLBR_{GspB}^{L442}$ and $SLBR_{GspB}^{Y443}$ closely approach the GalNAc (Figure iv-18A,B). We engineered $SLBR_{GspB\text{-}SK150}$ "mini-chimeras" that swapped single amino acids at these positions with the equivalent residues from $SLBR_{SK150}$. We then measured binding to sTa, 3'sLn, and sLe^C (Figure iv-18C-F). The $SLBR_{GspB}^{L442Y/Y443N}$ mini-chimera had increased binding to 3'sLn and sLe^C and was overall more similar in selectivity to $SLBR_{SK150}$ than to $SLBR_{GspB}$ (compare Figure iv-18C and Figure iv-3); however, the converse $SLBR_{SK150}^{Y300L/N301Y}$ mini-chimera still exhibited reduced binding (Figure iv-18D) and a size exclusion profile that suggested the presence of breakdown products, indicating that misfolding likely contributes to loss of binding for this variant (Figure iv-17D). The incomplete success of the mini-chimeras suggests complex origins for the inability to change selectivity in $SLBR_{GspB}$ and $SLBR_{SK150}$ via mutagenesis.

In summary, the SLBRs from the two branches of the evolutionary tree respond differently to chimeragenesis. The parent $SLBR_{GspB}$ and $SLBR_{SK150}$ cannot easily undergo alteration of their binding spectrum and tend to exhibit lower stability (Figure iv-17A-D) and loss of function (Figure iv-18C-F). In contrast, $SLBR_{Hsa}$, $SLBR_{SK678}$, and $SLBR_{UB10712}$ readily tolerate changes in binding spectrum via chimeragenesis to allow strong binding of alternative ligands (Table iv-4).

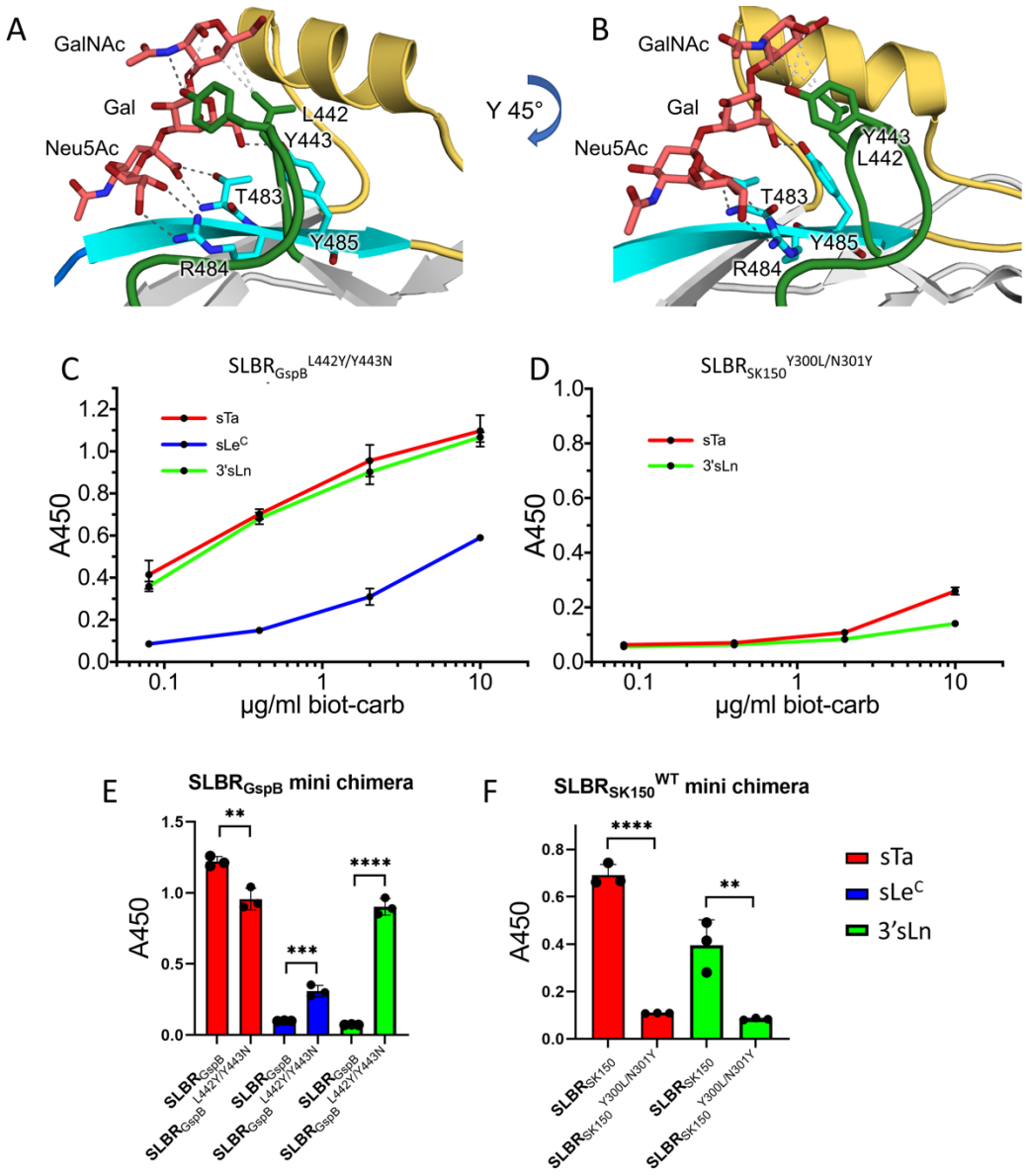

Figure iv-18 Mini-chimeragenesis of $SLBR_{GspB}$ and $SLBR_{SK150}$.

A-B. Design of mini-chimeras was based upon examining the interactions between $SLBR_{GspB}$ and sTa. The F strand is colored *cyan*. The CD and FG loops are colored *green* and *yellow* respectively. sTa is shown in *salmon* with oxygen shown in *red* and nitrogen shown in *blue*. Hydrogen bonds are shown in *dark grey* dashed lines and electrostatic interactions are shown in *light grey* dashed lines. C-D. Dose-response curves of biotin-glycan binding to immobilized binding regions (500 nM) are shown as mean. C. Mini-chimeragenesis with $SLBR_{SK150}$ was accomplished with the GST-$SLBR_{GspB}^{L442Y/Y443N}$ double mutant. The mini-chimera becomes more broadly selective by increasing the affinity for 3'sLn and sLe^C. As a result, it exhibits binding selectivity more similar to wild-type GST-$SLBR_{SK150}$ (see Supplementary Figure 1). D. The converse mini-chimeragenesis of $SLBR_{SK150}$ exhibited reduced binding for sialoglycan ligands that bind most avidly to both wild-type $SLBR_{GspB}$ and $SLBR_{SK150}$. E-F. Comparison of parent and mini-chimera $SLBR_{GspB}$ and $SLBR_{SK150}$. Black circles represent individual data points and bars represent the mean ± SD. Binding of each glycan to the mini-chimera was compared to the parent SLBR using two-tailed parametric t tests. E. The p values comparing binding of sTa, sLeC, and 3'sLn to $SLBR_{GspB}$ and $SLBR_{GspB}^{L442Y/Y443N}$ are 0.0054, 0.0008, and <0.0001 respectively. F. The p

values comparing binding of sTa and 3'sLn to $SLBR_{SK150}$ and $SLBR_{SK150}^{Y300L/N301Y}$ are <0.0001 and 0.0072 respectively. C-F Values correspond to the mean ± standard deviation, with n = 3 independent experiments using a single protein preparation. Statistical significance is indicated by "*" (ns P >0.05; *P <0.05; **P < 0.01; ***P < 0.001; **** < 0.0001).

Site-directed mutagenesis identifies key residues involved in selectivity.

The identification of the CD, EF, and FG loops as the regions that are of largest natural sequence variation (Figure iv-9) and as regions that may control glycan selectivity (Figure iv-14, Figure iv-16) could suggest that these evolved to allow for binding to different host receptors. Natural evolutionary changes in SLBR sequence might involve point mutations rather than substitutions of entire loops. We therefore wanted to test whether point mutations of the loops of $SLBR_{Hsa}$, $SLBR_{SK678}$, and $SLBR_{UB10712}$ could change the selectivity. In $SLBR_{Hsa}$, $SLBR_{SK678}$, and $SLBR_{UB10712}$, we substituted residues at positions equivalent to $SLBR_{Hsa}^{E286}$ of the CD loop and $SLBR_{Hsa}^{D356}$ of the FG loop because our structures show that these residues closely approach the variable region of the ligands (Figure iv-13). We then measured relative binding to five physiologically-relevant ligands via ELISA (Figure iv-19A-C, Table iv-4).

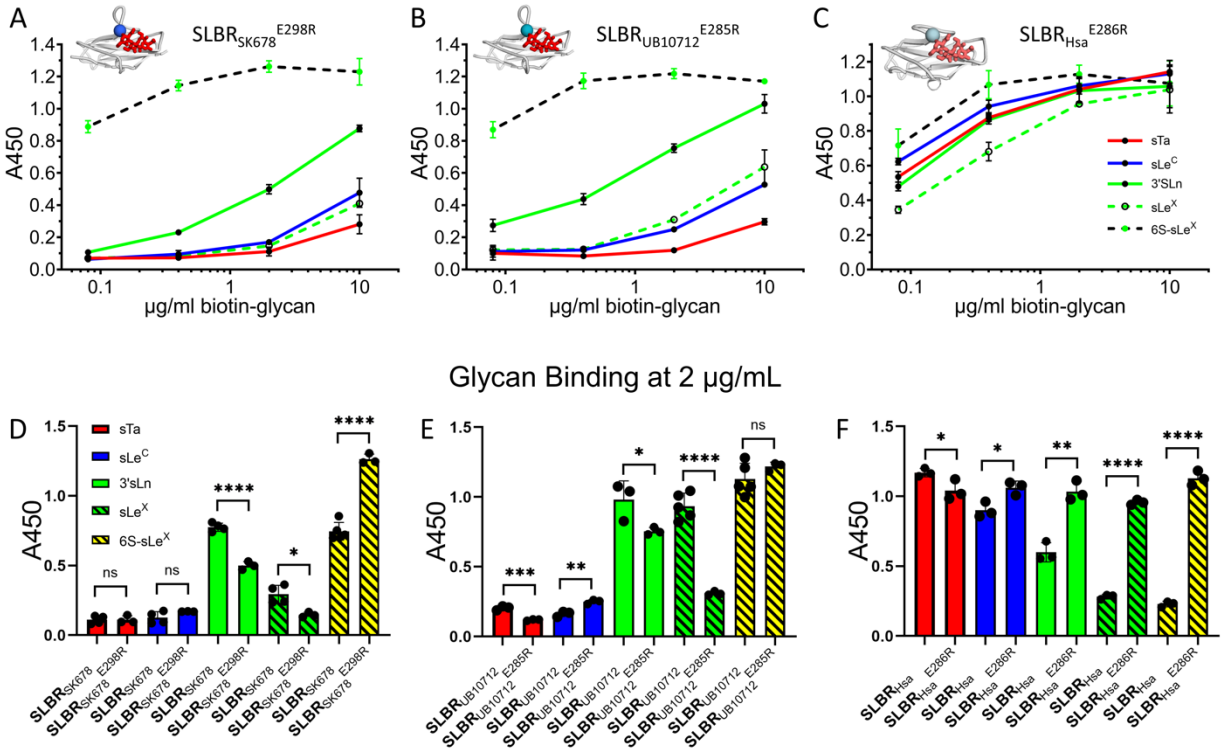

Figure iv-19 Binding selectivity of CD loop variants in $SLBR_{SK678}$, $SLBR_{UB10713}$ and $SLBR_{Hsa}$.

Dose response curves of biotin-glycan binding to immobilized variant GST-SLBRs (500 nM). A. GST-$SLBR_{SK678}^{E298R}$, B. GST-$SLBR_{UB10712}^{E285R}$, C. GST-$SLBR_{Hsa}^{E2866}$. The respective SLBRs are shown in grey cartoon in the top left corner of each panel with the site of mutation represented as a colored sphere. sTa, shown in red sticks, was placed over the binding site by superimposing sTa bound-$SLBR_{Hsa}$. Measurements were performed using 500 nM of immobilized GST-SLBR and the indicated concentrations of each ligand are shown as the mean ± SD (n = 3 independent experiments with a single protein preparation). D-F. Binding of each sugar to each mutant was statistically compared to binding of the same sugar to the $SLBR^{WT}$ with a two-tailed parametric t test. Black circles represent individual data points and bars represent the mean ± SD. Statistical significance is indicated by "*" (ns P >0.05; *P <0.05; ***P* < 0.01; ****P* < 0.001; **** < 0.0001).

In the CD loop (i.e., $SLBR_{Hsa}^{E286}$), our crystallographic analysis suggested that ionic repulsion from the negatively charged side chain excludes the negative charge of a sulfated ligand. We therefore substituted a positive charge at this location in $SLBR_{UB10712}$, $SLBR_{SK678}$, and $SLBR_{Hsa}$. All three of these variants exhibited a substantial increase in binding for 6S-sLe^X (Figure iv-19D-F and Table iv-4). $SLBR_{Hsa}^{E286R}$ retained binding to non-sulfated ligands and this variant became quite promiscuous for the ligands tested by ELISA (Figure iv-16C). To better evaluate the binding spectrum of $SLBR_{UB10712}$ and $SLBR_{SK678}$, we assessed >500 glycans via array analysis as compared to a GST control (Figure iv-20**, Supplementary Data 1**). These studies indicate that the engineered SLBRs are selective for two closely-related glycans: 6S-sLe^X and 6S-3'sialyllactosamine (6S-3'sLn, Neu5Acα2-3Galβ1-4GlcNAc6Sβ, Figure iv-1G) which lacks the fucose.

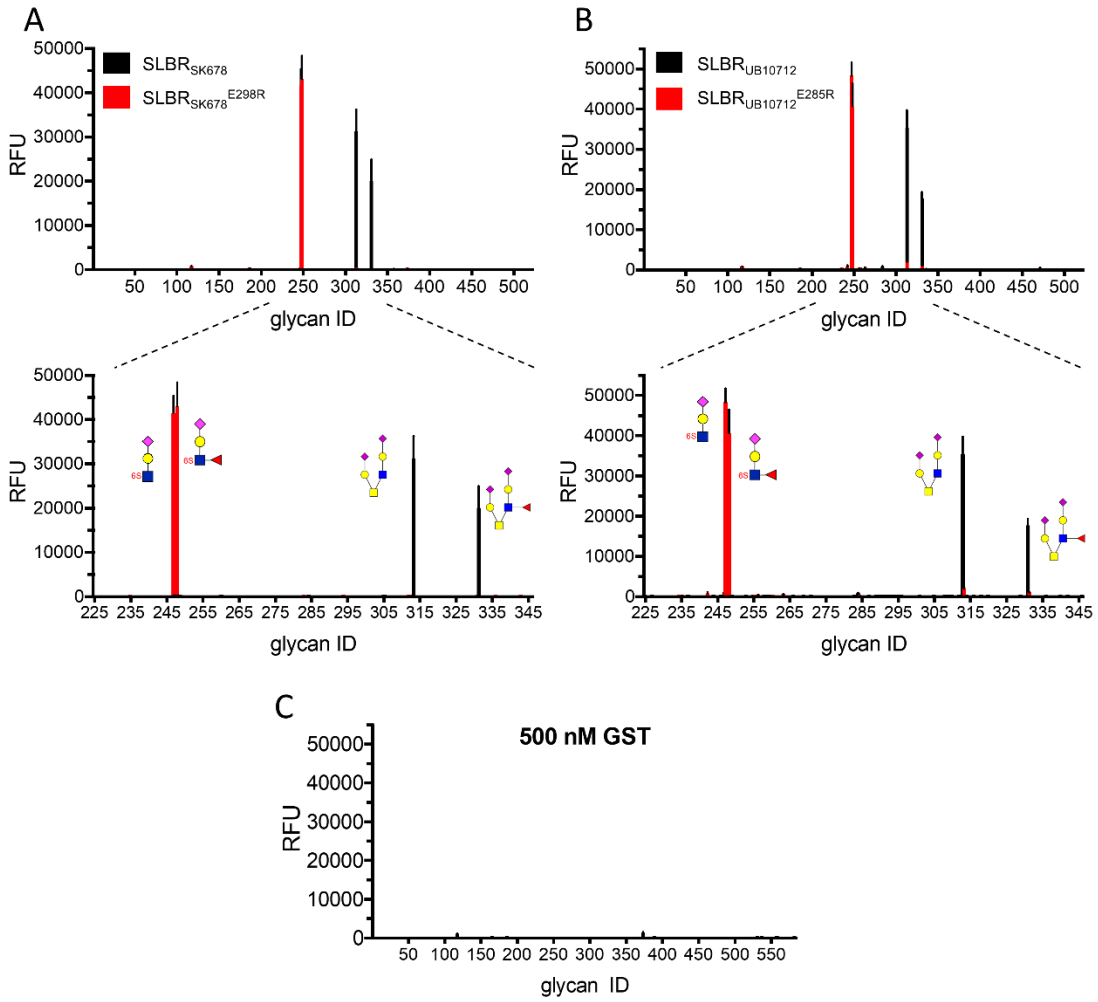

Figure iv-20 Center for Functional Glycomics (CFG) version 5.4 glycan arrays for SLBR$_{SK678}$ and SLBR$_{SK678}^{E298R}$.

Binding to 500 glycans in the CFG array was independently assessed to A. GST-SLBR$_{SK678}$ (500 nM, *black*) and GST-SLBR$_{SK678}^{E298R}$ (500 nM, *red*); B. GST-SLBR$_{UB10712}$ (500 nM, *black*) and GST-SLBR$_{UB10712}^{E285R}$ (500 nM, *red*); and C. GST. The inset in A. and B. highlights the narrow selectivity and the difference in glycans that are robustly recognized by parent versus the engineered SLBRs. Numbers on the X-axis correspond to individual glycans in the arrays. The Y-axis is relative response.

We then evaluated selectivity conferred by the FG loop where crystallographic analysis would suggest that the β-branching of SLBR$_{Hsa}^{D356}$ excludes C3 fucosylation, while the larger, unbranched Gln of SLBR$_{UB10712}$ and SLBR$_{SK678}$ can bind fucosylated ligands. We therefore substituted Asp for Gln in SLBR$_{UB10712}$ and SLBR$_{SK678}$ and conversely substituted Gln for Asp in SLBR$_{Hsa}$. As assessed by ELISA, the SLBR$_{SK678}^{Q354D}$ and SLBR$_{UB10712}^{Q367D}$

variants lost binding to fucosylated ligands (Figure iv-22A-B**,** Figure iv-21A, B). As a result, $SLBR_{UB10712}^{Q354D}$ became more selective for 3'sLn while the $SLBR_{SK678}^{Q367D}$ exhibited low binding to all tested ligands. As assessed by size exclusion chromatography, the $SLBR_{SK678}^{Q367D}$ variant was properly folded such that loss of binding did not result from a folding defect (Figure iv-21C). The observed loss of binding to the fucose-containing sLe^X and 6S-sLe^X by these FG loop variants is consistent with the structural analysis and chimeragenesis showing that the FG loop is particularly important for accommodation of α1,3-fucosylation (Figure iv-14**,** Figure iv-16A, B). The converse $SLBR_{Hsa}^{D356R}$, and $SLBR_{Hsa}^{D356Q}$ remained broadly-selective but with increased binding to the α1,3-fucosylated sLe^X and 6S-sLe^X as compared to parent $SLBR_{Hsa}$ (Figure iii-8C-D, Figure iv-21D).

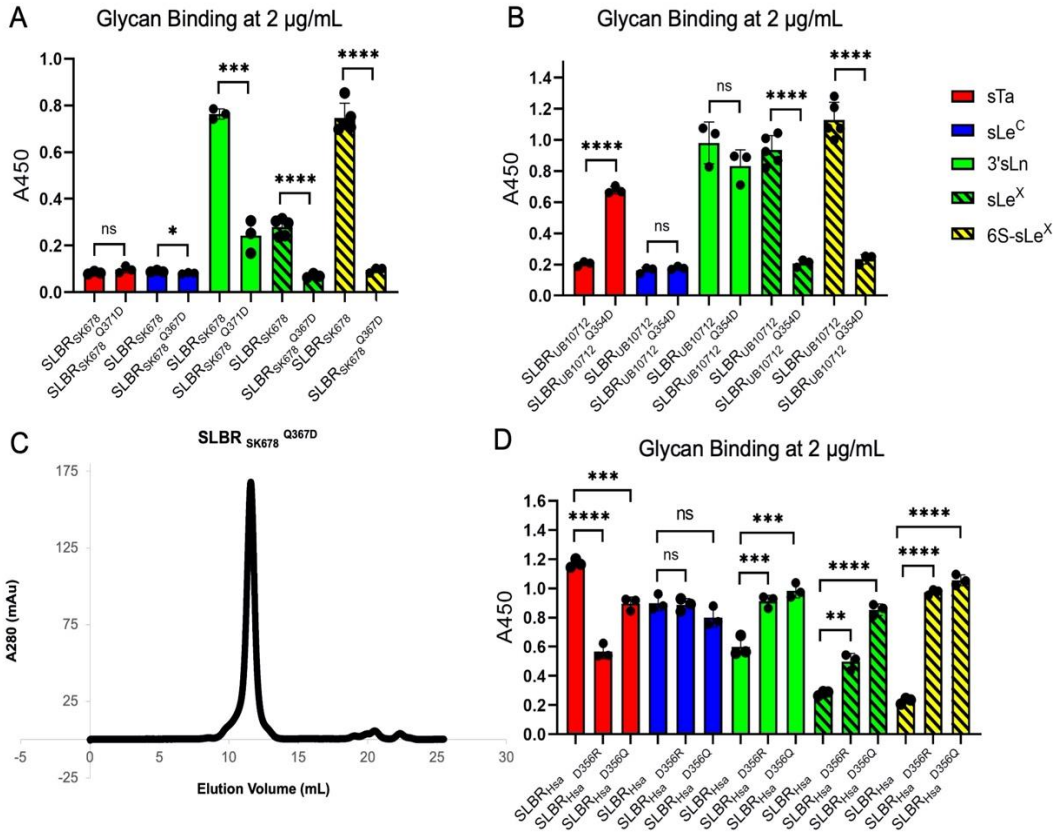

Figure iv-21 Analysis of glycan binding by SLBRs with point mutations in the FG loop.

A-B. Binding of glycans to each engineered variant SLBR was statistically compared to binding of the same glycan to parent SLBR using a two tailed nonparametric t test. A. The p values comparing binding of sTa, sLe^C, 3'sLn, sLe^X, and 6S-sLe^X to $SLBR_{SK678}$ and $SLBR_{SK678}^{Q367D}$ variant are 0.201, 0.0456, 0.0003, <0.0001, and <0.0001 respectively. B. The p values comparing binding of sTa, sLe^C, 3'sLn, sLe^X, and 6S-sLe^X to the parent $SLBR_{UB10712}$ to the variant $SLBR_{UB10712}^{Q354D}$ are <0.0001, 0.326, 0.205, <0.0001, and <0.0001 respectively. C.

Size exclusion chromatography of 0.5 mg of the non-binding $SLBR_{SK678}^{Q367D}$ variant shows a single peak at 11.6 mL, consistent with folded protein. D. Binding of each glycan to $SLBR_{Hsa}$ FG mutants was statistically compared to binding of the same glycan to parent $SLBR_{Hsa}$ with an ordinary one-way ANOVA corrected for multiple comparisons with Dunnett's test. The p values comparing binding of sTa, sLe^C, 3'sLn, sLe^X, and 6S-sLe^X to the parent $SLBR_{Hsa}$ and the variant $SLBR_{Hsa}^{D356R}$ are <0.0001, 0.948, 0.0008, 0.0011, and <0.0001 respectively. The p values comparing binding of sTa, sLe^C, 3'sLn, sLe^X, and 6S-sLe^X to the parent $SLBR_{Hsa}$ and the variant $SLBR_{Hsa}^{D356Q}$ are 0.0004, 0.129, 0.0003, <0.0001, and <0.0001 respectively. A,B,D Black circles represent individual data points and bars represent the mean ± SD with n = 3 independent experiments using a single protein preparation. Statistical significance in each panel is indicated by "*" (ns P >0.05; *P <0.05; **P < 0.01; ***P < 0.001; **** < 0.0001).

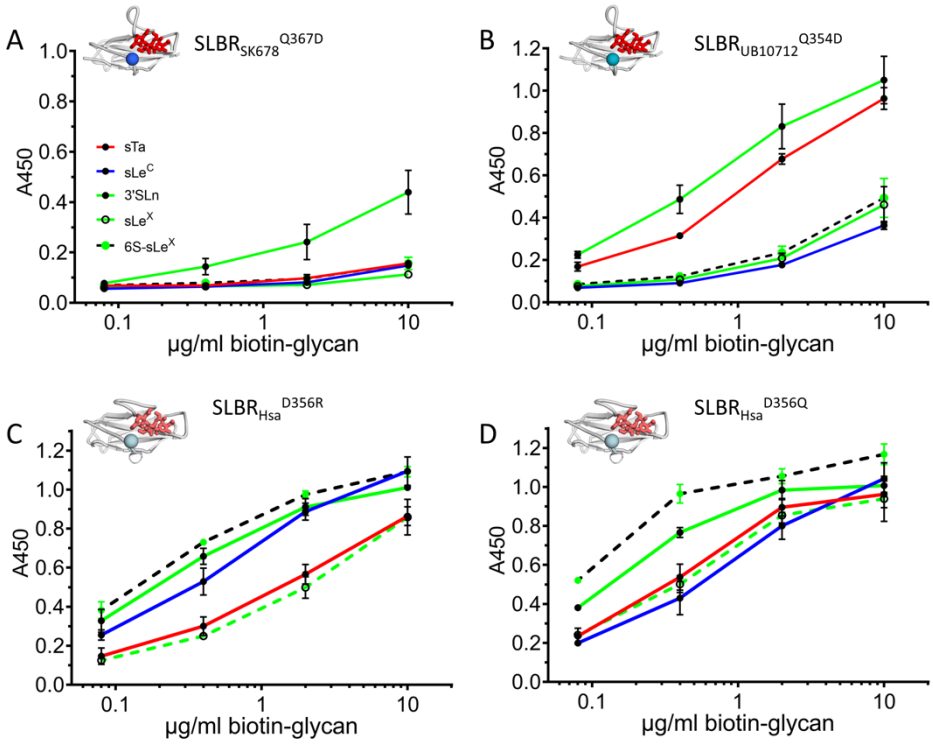

Figure iv-22 Binding selectivity of FG loop variants in $SLBR_{SK678}$, $SLBR_{UB10713}$, and $SLBR_{Hsa}$.

Dose response curves of biotin-glycan binding to immobilized variant SLBRs (500 nM). Both the A. GST-$SLBR_{SK678}^{Q367D}$ and B. GST-$SLBR_{UB10712}^{Q345D}$ variants have substantially reduced binding to the fucosylated ligands sLe^X and 6S-sLe^X. In $SLBR_{Hsa,}$ charge reversal or neutralization at this same position was assessed in C. GST-$SLBR_{Hsa}^{D356R}$ and D. GST-$SLBR_{Hsa}^{D356Q}$. Both variants had increased binding to 6S-sLe^X, 3'sLn, and sLe^X and decreased binding to sTa, albeit to somewhat different extents. Measurements were performed using 500 nM of immobilized GST-SLBR and the indicated concentrations of each ligand are shown as the mean ± SD (n = 3 independent experiments with a single protein preparation). Statistical comparisons of ligand affinity between FG mutants and parent SLBR can be found in Supplementary Figure 18.

Taken together, point mutations in the broadly-selective SLBRs can alter the identity of the preferred ligand, and can bind robustly to the newly-preferred ligand. The EC_{50} values suggest that the binding is strong enough to

make physiologically-relevant adhesive interactions to host receptors. A possible evolutionary rationale for facile alteration in sialoglycan binding spectrum is that this allows a bacterium to adapt to changes in host sialoglycan display.

Binding of selectivity variants to host receptors in saliva and plasma.

To test whether changes in SLBR binding to synthetic glycans had corresponding effects in the interactions of the SLBRs with human ligands, we examined the binding of parent and variant SLBRs to human salivary and plasma glycoproteins using far western analysis. We focused on the chimeras and variants that had narrower selectivity, where changes in binding would be most evident. We first identified the glycoprotein targets of parent and variant SLBRs in submandibular sublingual (SMSL) ductal saliva from four donors. The three parent SLBRs recognized a band consistent with the mobility of MUC7 in all four samples (Figure iv-23A, **Supplementary Data 2**), but the levels of binding differed between samples. In addition, the SLBRs bound MUC7 glycoforms of different apparent mass ranges, likely reflecting differences in the number, size and composition of *O*-glycan structures[128, 129]. $SLBR_{SK678}$ and $SLBR_{UB10712}$ detected glycoforms of ~160 kDa, whereas $SLBR_{Hsa}$ bound more readily to 140-150 kDa glycoforms (Figure iv-23A). $SLBR_{UB10712}$ recognized the band consistent with MUC7 in all four samples nearly equally, whereas $SLBR_{SK678}$ detected this band from donor 3 > donors 1 and 4 > donor 2, and $SLBR_{Hsa}$ detected this band from donor 3 > donors 2 and 4 > donor 1. The recognition pattern of the $SLBR_{SK678}^{Hsa\text{-}loops}$ and $SLBR_{UB10712}^{Hsa\text{-}loops}$ chimeras resembled that of $SLBR_{Hsa}$ rather than that of the parent $SLBR_{SK678}$ and $SLBR_{UB10712}$. These loop exchanges altered both the apparent mass recognized and the avidity of the binding. In contrast, the 6S-sialoglycan-selective point mutants showed preferential binding to the uppermost mass range of MUC7 in samples from donors 1 and 4, and a near loss of binding to samples from donors 2 and 3.

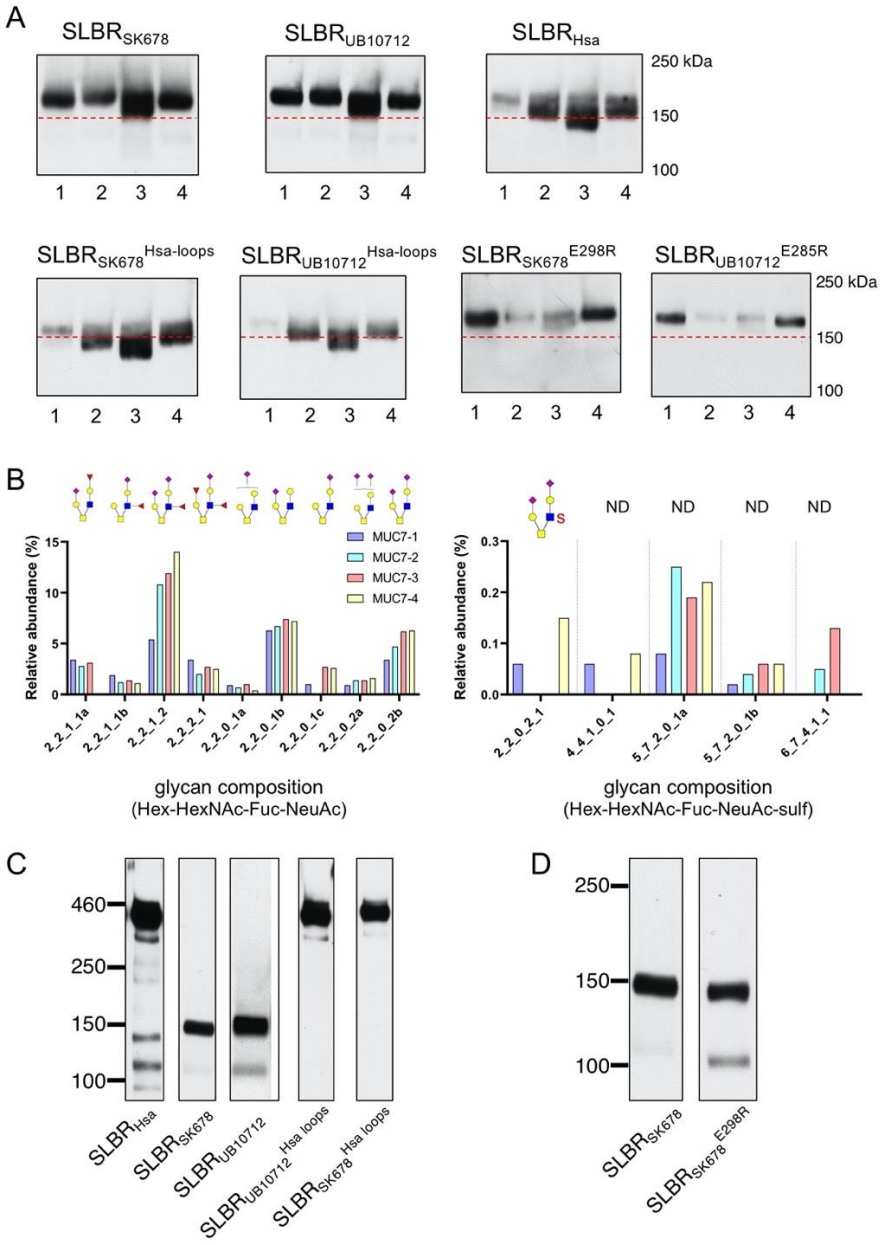

Figure iv-23 MUC7 O-glycans and SLBR recognition of glycoproteins in human saliva and plasma.

A. Representative far-western blot of the SMSL saliva samples with parent and variant GST-tagged SLBRs (n=2). The MUC7 glycoforms range from 120 to 160 kDa. Saliva samples (1µl) were run on the same gel and transferred to the same nitrocellulose membrane. The membrane was subsequently cut in order to separately probe with parent versus variant SLBRs (15 nM). The dashed red line indicates the 150kD molecular weight marker. B. The major non-sulfated (left) and sulfated (right) O-linked glycans from MUC7 in four samples of submandibular sublingual (SMSL) saliva. The x-axis represents glycan compositions Hex-HexNAc-Fuc-Neu5Ac and Hex-HexNAc-Fuc-Neu5Ac-Sulf for the upper left and right panel, respectively. Lower case letters a, b, and c indicate different isomer structures with the same monosaccharide compositions. Putative structures are shown above the graphs (ND, not determined). C. Representative far-western blot of human plasma with parent and chimeric GST-tagged SLBRs (n=2). As previously identified by affinity capture and mass spectrometry[70], the 460 kD band is proteoglycan 4, the 150 kD band is GP1bα, and the 100 kD band is C1-

esterase inhibitor. D. Representative far-western blot of human plasma with parent $SLBR_{SK678}$ and the $SLBR_{SK678}^{E298R}$ point mutant (n=2).

We next determined whether the recognition patterns correlated with the presence of sTa versus 3'sLn (for the loop chimeras) or with the presence of 6-*O*-sulfo structures (for the single residue substitutions), decorating larger physiological glycans. To do this, we used affinity capture and mass spectrometry to characterize the O-glycan composition of the four MUC7 samples (Figure iv-23B, Figure iv-24**, Figure iv-26**). The O-glycan profiles were similar to those seen in two earlier reports[128, 129], in that dozens of different structures were evident in each sample. The most abundant structures were mono- or di-sialylated Core 2 glycans. There were relatively minor amounts of sTa and there were differences in the assortment of other minor structures. The glycans from the four donors differed in the extent of sialylation and fucosylation (Figure iv-24**, Figure iv-26**), the presence or absence of sulfated structures (Figure iv-23B), and the relative abundance of each species. The *O*-glycan profiles are consistent with the ELISA measurements to purified glycans (Figure iv-14C-E). Specifically, $SLBR_{Hsa}$, $SLBR_{SK678}^{Hsa\text{-}loops}$, and $SLBR_{UB10712}^{Hsa\text{-}loops}$ preferred sTa in the ELISA assays with purified glycans (Figure iv-14) and bound Core 2 structures that contain Neu5Ac on the sTa-like Core 1 branch in salivary MUC7 (Figure iv-23B). In addition, $SLBR_{SK678}$ and $SLBR_{UB10712}$ bound to 3'sLn and 6S-sLe^X in ELISA assays (Figure iv-14A,B) and bound to structures that have Neu5Ac on the 3'sLn branch in MUC7 (Figure iv-23B). Finally, the $SLBR_{SK678}^{E298R}$ and $SLBR_{UB10712}^{E285R}$ both strongly preferred 6-*O*-sulfated species over other ligands (Figure iv-19). The presence of a 6S-3'sLn moiety in the samples from donors 1 and 4 (the 2-2-0-2-1 structure) suggests that these variants recognize MUC7 modified with relatively minor amounts of 6S-3'sLn, potentially reflecting high-affinity binding.

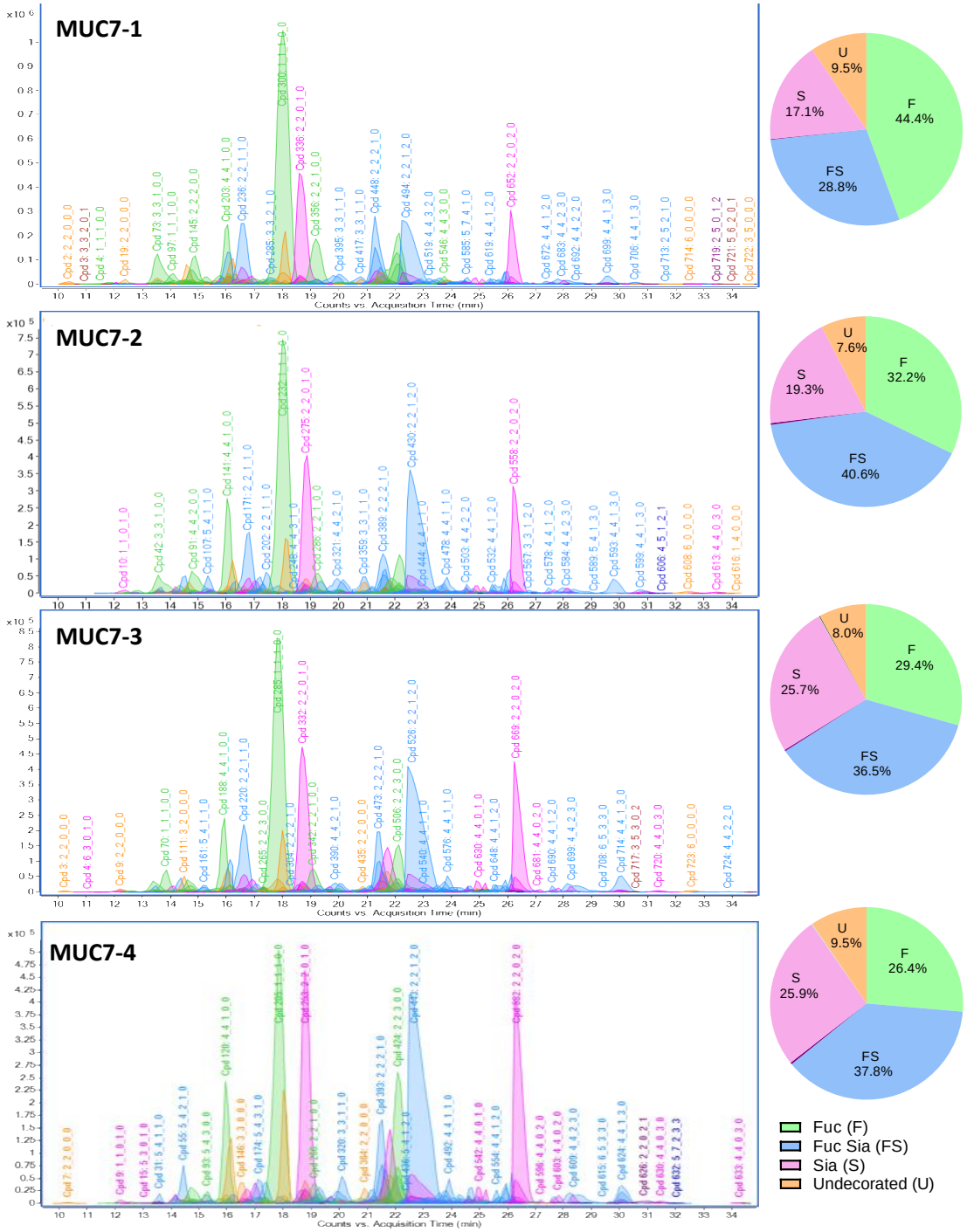

Figure iv-24 Comparison of the identity of *O*-glycans released from four MUC7 samples, and pie charts representing the relative abundances of sub-glycan groups.

The extracted compound chromatograms (ECCs) of *O*-glycans from four different saliva donors were categorized into four different groups: undecorated (U); fucosylated (F); sialylated (S); and fucosialylated (FS).

The monosaccharide compositions (hexose (Hex)-HexNAc-Fuc-NeuAc-Sulf) were inferred from the precise masses determined by LC-MS.

SLBRs may also interact with glycoproteins in the bloodstream, and the binding spectrum may have consequences for pathogenicity. We therefore next evaluated binding to human plasma proteins by far western analysis. Consistent with our prior studies, parent $SLBR_{Hsa}$ preferentially bound proteoglycan 4 (460 kD) from human plasma, while $SLBR_{UB10712}$ bound GPIbα (150 kD). Of note, proteoglycan 4 is a major carrier of sTa in plasma, whereas GPIbα has predominantly di-sialylated Core 2 structures. These SLBRs also bound different glycoforms of the C1-esterase inhibitor (100-120 kDa)[70] (Figure iv-23C). The chimeric $SLBR_{UB10712}^{Hsa\text{-}loops}$ and $SLBR_{SK678}^{Hsa\text{-}loops}$ chimeras now recognized proteoglycan 4 rather than the preferred receptors for parent $SLBR_{SK678}$ and $SLBR_{UB10712}$ (Figure iv-23**C**). We also found that the $SLBR_{SK678}^{E298R}$ variant bound both GPIbα, a receptor associated with infective endocarditis, and the C1-esterase inhibitor (Figure iv-23D). Thus, the preferred plasma ligands for the SLBRs appears to be largely determined by the loop residues, as was the case for the recognition of MUC7 glycoforms.

Discussion

Bacterial attachment to host structures is critical for commensalism and is the first committed step in many types of infection. SLBRs can mediate streptococcal binding to a variety of host glycoproteins[69, 70, 118, 119, 120, 123, 130, 131, 132, 136, 137, 138], and binding to sTa correlates with pathogenesis in an animal model of endovascular infection[130]. But it has not previously been clear how the SLBRs distinguish between the many protein-attached glycans displayed by host. Here, we evaluated how five SLBRs select between sialoglycan receptors. The common element of these glycans, Neu5Acα2-3Gal, interacts with SLBRs via the ΦTRX motif[52, 55, 69, 139] and the EF loop (Figure iv-6**,** Figure iv-9)[49, 55]. The CD and FG loops select for the underlying reducing end (Figure iv-14**,** Figure iv-22**,** Figure iv-16**,** Figure iv-18), which varies in the identities of its individual sugars, the linkage between the sugars, and the elaborations on the sugars (Figure iv-1). This suggests roles for distinct regions of the SLBR structure in glycan selection (Figure iv-25) The substantial sequence and structural variability in the CD, EF, and

FG loops as compared to the core fold of the SLBR (Figure iv-6) suggests that these regions can tolerate more substitutions while avoiding the liability of misfolding. Indeed, modification of these regions via chimeragenesis or mutation allowed some of the SLBRs to bind different glycoforms of MUC7 or interact with different preferred sialoglycans (Figure iv-14**,** Figure iv-19**,** Figure iv-22**,** Figure iv-15**,** Figure iv-16**,** Figure iv-18**,** Figure iv-24) and different host plasma proteins (Figure iv-23).

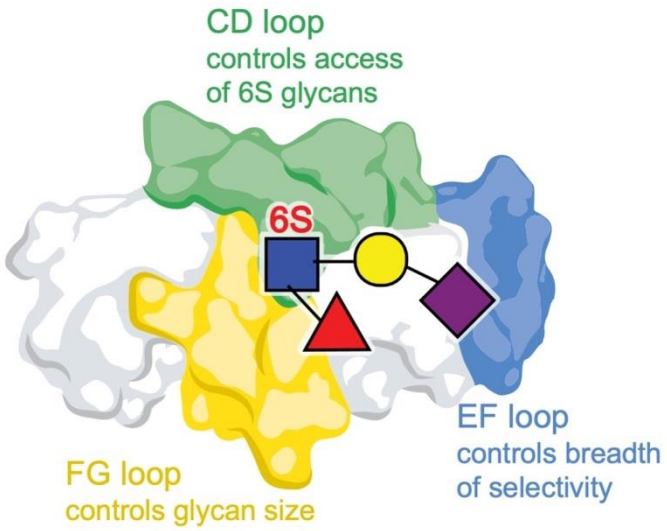

Figure iv-25 Model for how SLBRs control sialoglycan selectivity.

The glycan-binding pocket of SLBRs is organized above a ΦTRX sequence motif on the F-strand of the V-set Ig fold that interacts with sialic acid. Three variable loops surround this sialoglycan binding pocket and affect selectivity. In the broadly-selective SLBRs, flexibility of the EF loop correlates with breadth of selectivity. The CD loop controls specificity for 6-*O*-sulfated glycans, and the FG loop may control whether the SLBR prefers trisaccharides versus larger glycans. Tetrasaccharides containing α1,3-fucosylation were tested here, but past studies of $SLBR_{SrpA}$ identify that a small FG loop correlates with the ability to accommodate larger glycans[49].

Although not previously noted for bacterial SLBRs, the use of loops to control selectivity has been observed in other sialoglycan-binding systems. For example, mammalian Siglec proteins are organized around a V-set Ig-fold but are not detectably related in sequence to the SLBRs[49, 52, 140, 141]. The GG' and CC' loops are adjacent to the sialoglycan binding site and are variable in structure. In Siglec-7, the CC' loop[142] controls sialoglycan selectivity. In Siglec-8, alteration of this same loop allows the binding of 6'S sialoglycans[143]. Thus, changes in loop structure may therefore be a common way to evolve changes in ligand binding selectivity.

The use of loops to control selectivity appears to be a robust way to accommodate a broad range of complex glycans. Indeed, the glycans recognized by SLBRs differ in both the identity of the individual glycans as well as in the linkages between the individual carbohydrates. When bound to these SLBRs (see Figure iv-13B), glycans with different linkages differ in the overall shape as well as in the pattern of hydrogen-bonding donors and acceptors. However, the glycosidic linkage itself does not differ in position with respect to the SLBR binding pocket. Thus, these SLBRs distinguish between glycans with different linkages by changing the steric and electrostatic properties of the region of the pocket that follows the linkage, namely the CD and FG loops. While we focused on SLBRs that recognize tri- and tetrasaccharides, SLBRs can recognize sialoglycans with as few as three and possibly more than six monosaccharide units[49, 69, 70, 127]. For example, $SLBR_{SrpA}$ may biologically recognize a hexasaccharide[70] but can bind to partial ligands with lower affinity[49, 55, 69]. SLBRs that recognize larger sialoglycans appear to contain a modular binding site similar to those studied here, albeit with larger binding pockets and with more independent recognition regions. In the oral cavity, this may assist in colonization through interaction with salivary MUC7, which exhibits heterogeneity of its sialoglycan modifications both within and between human hosts (Figure iv-23A,B, Figure iv-24**, Figure iv-26**). Here, sialoglycans are attached to MUC7 and the SLBR binding pocket can bind glycan receptors that are linked to host proteins. The linkage to the receptor protein could affect binding and could involve additional contacts to the SLBR[144].

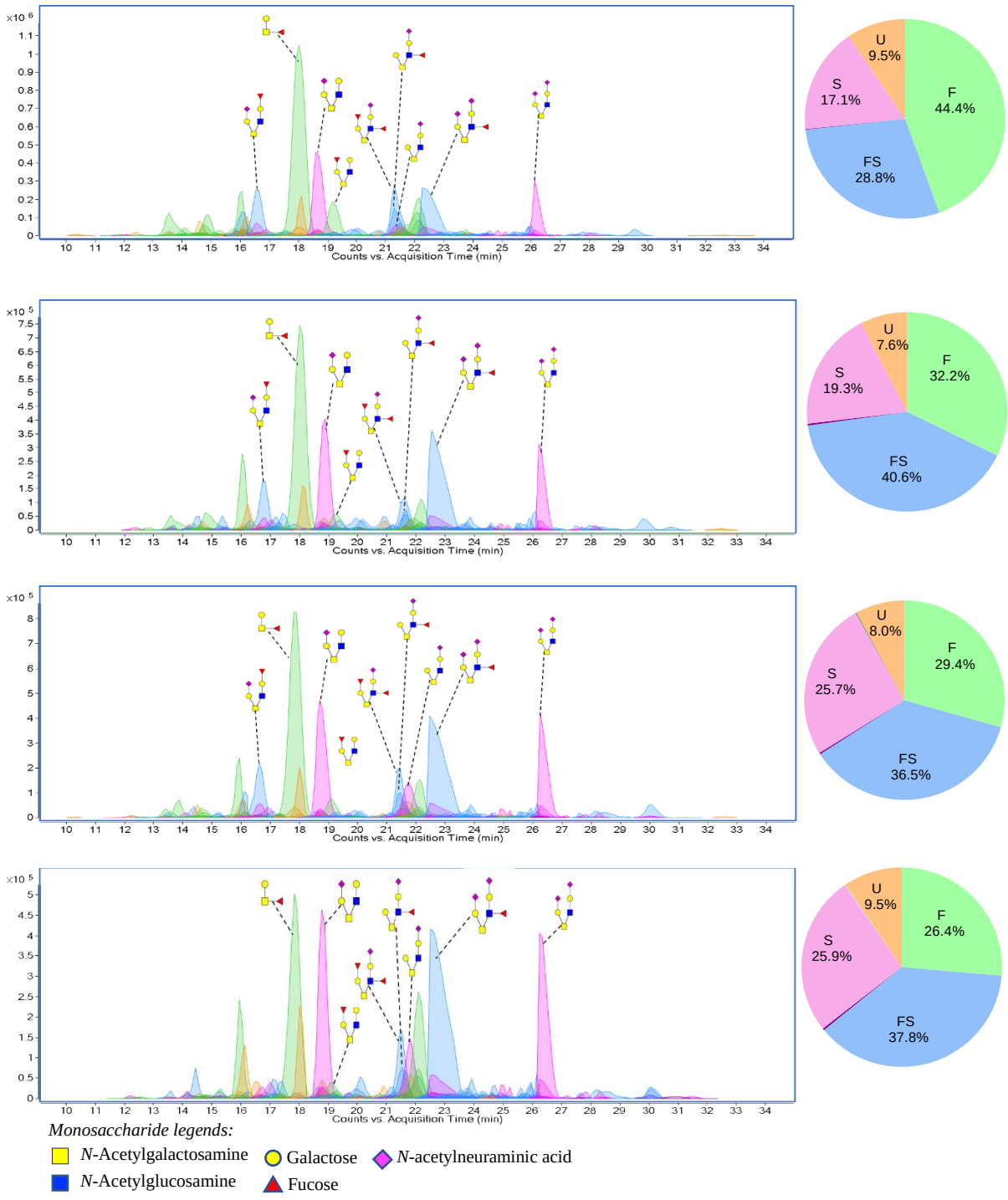

Figure iv-26 P**roposed structures of the major *O*-glycans of MUC7 from four donors.**

The putative structures are based on the precise masses and inferred monosaccharide compositions in addition to the MS/MS fragmentation patterns and literature data.

In this context, mutation of these loops may be advantageous to the bacterium because it allows facile switching of host receptors. While we do not know how the sequences of the SLBRs actually change during

evolution, streptococci compete with numerous other species in the oral cavity[145]. As many of these strains contain SLBRs, genetic recombination is likely, which can allow a bacterium to incorporate or modify a SLBR. The ready toleration of mutations in the loops may allow these regions to disproportionately change their sequences. Some of these changes may enable the bacterium to bind a different sialoglycan structure (Figure iii-7**,** Figure iii-8**,** Figure iii-9, Figure iv-16**,** Figure iv-20**,** Figure iv-21). Within a single human host, this could allow colonization of a region of the oral cavity that displays different glycans, could promote binding to different salivary components, or could allow binding to other oral bacteria that are sialylated. This mutability could also permit improved binding to different individuals in the population or allow the colonization of a preferred host, as animals and humans may differ in their glycosylation.[146] This mechanism mirrors that of polyomavirus and rotavirus, where single amino acid substitution or a very small number of point mutations can change the identity of preferred host sialoglycan receptors[147, 148].

In some of our point mutants, the improvement in affinity and selectivity to alternative ligands exceeds those reported for dedicated engineering studies of glycan-binding lectins[43, 149, 150, 151, 152, 153, 154, 155, 156]. In those past reports, the maximum enhancement in binding to a non-native glycan is ~20-fold[43, 149, 150, 151, 152, 153] and selectivity was often achieved via a decrease in affinity to non-desired ligands in a promiscuous starting lectin[154, 155, 156]. Development to increase the affinity and narrow the selectivity even further could allow the SLBRs to be used as probes to assess glycan identity and abundance. Key aspects of a probe include the ability to detect glycans in cells and in patient samples. The cellular interaction was shown in recent studies that evaluated the binding SLBRs to engineered HEK293 cell lines with altered glycosylation[157], while the ability to recognize glycans in saliva and plasma suggests that these will be useful in other samples (Figure iv-23).

Collectively, our findings give a description for how SLBRs recognize ligands. The conserved sialic acid-recognition motif governs general specificity while sequence diversity in surrounding loop regions allows the SLBR to select between related sialoglycans (Figure iv-25). This binding site architecture may be optimized for facile selectivity changes in related SLBRs. This may further explain how bacterial adhesive proteins have evolved to adapt to host receptors. Finally, this work suggests a route for engineering these SLBRs to use as probes to

detect specific glycosylation, which is a focus of ongoing work. A library of SLBR-based binding proteins could be used for glycome mapping or as diagnostic or therapeutic tools for disease states with aberrant glycosylation.

v. MOLECULAR BASIS FOR SIALIC ACID SELECTIVITY IN A FAMILY OF STREPTOCOCCAL SERINE RICH REPEAT ADHESINS

Introduction

The term sialic acid (Sia) collectively refers to a broad range of neuraminic acid derivatives, with over 50 distinct forms of Sia found in nature. Sia-containing glycans decorate many cell surface glycoproteins. Mammalian proteins known as selectins or Sia-binding immunoglobulin-like lectins (Siglecs) recognize and bind sialic acid. These interactions contribute to self-recognition, which underlies a range of mammalian signaling processes, including inflammation and immunity. However, sialic acid can also be exploited by bacteria and viruses for use as a host receptor, allowing for cell-specific adherence. For example, strains of streptococci that contain adhesins with a Siglec-like domain bind to sialoglycans on salivary mucins and blood plasma during oral commensalism and endocardial pathogenesis respectively.

Interestingly, humans synthesize different forms of Sia than do other mammals. Many mammals have an abundance of the Sia known as Neu5Gc, however, humans cannot synthesize this Sia. Instead, humans have an abundance of Neu5Ac, making up more than ninety percent of our total Sia.[3] Neu5Gc differs from Neu5Ac only by a single hydroxyl group found on the C11 carbon (Figure v-1). Recent glycan array analysis to characterize how related Siglec-like streptococcal adhesins interact with synthetic sialoglycans identified differences in Neu5Ac/Neu5Gc selectivity. Some Siglec-like adhesins, like GspB from *Streptococcus gordonii* strain M99, bind selectively to Neu5Ac-containing sialoglycans.[127] Other adhesins, like SrpA from *S. sanguinis* strain SK36 bind selectively to Neu5Gc-containing sialoglycans.[158] Finally, some adhesins, like Hsa from *S. gordonii* strain Challis, bind with dual specificity to both Neu5Ac and Neu5Gc.[158] This Neu5Ac/Neu5Gc selectivity may have implications for host preference.[159] Indeed, it has been proposed that loss of Neu5Gc biosynthesis in humans

conferred an evolutionary advantage by allowed resistance to ancient forms of malaria that used Neu5Gc as a receptor.[160]

Prior studies of Neu5Ac/Neu5Gc selectivity have focused on the basis for selectivity in high-fidelity Neu5Ac or Neu5Gc recognizing adhesins. In aggregate, these prior studies suggest that specific contacts at the C11 position of Sia dictate selectivity for either Neu5Ac or Neu5Gc. For example, human Siglec-family proteins prefer Neu5Ac while some homologous murine Siglecs prefer Neu5Gc.[161] In bacterial Siglec-like proteins, structural analysis of the Neu5Ac-preferring *S. gordonii* GspB showed the EF loop is likely too close to the Sia binding site to allow for the C11 hydroxyl of Neu5Gc. In contrast, the Neu5Gc-preferring *S. sanguinis* SrpA contained a hydrogen bond to the unique hydroxyl of Neu5Gc. However, site directed mutagenesis at this position destabilized the protein such that this hypothesis could not be experimentally verified.

With the conclusion that specific contacts to the C11 hydroxyl dictate human versus animal Sia selectivity in members of this family of bacterial adhesins, it is difficult to imagine how both a hydroxylated and non-hydroxylated Sia C11 can be recognized with equivalent affinity by some homologs. This leaves a mechanism for promiscuous binding to Neu5Ac/Neu5Gc unclear. Prior structural studies of the Neu5Ac/Neu5Gc promiscuous *S. gordonii* $SLBR_{Hsa}$ identified that binding to a Neu5Ac-based ligand is accompanied by a conformational change near the sialic acid. Thus, one compelling hypothesis for the observed equivalent Neu5Ac/Neu5Gc binding is that distinct conformational changes accompany the binding of Neu5Ac- and Neu5Gc-based sialoglycans to promote different contacts to the C11 position.

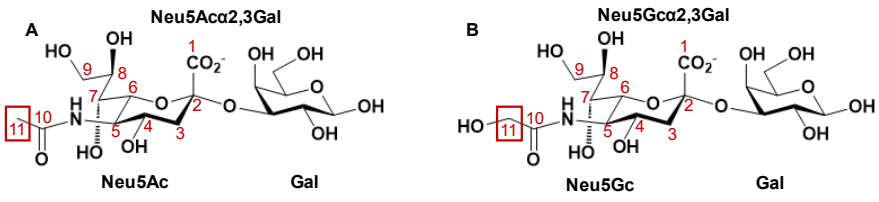

Figure v-1 Structures of Neu5Acα2,3Gal and Neu5Gcα2,3GalβOMe.

The carbons of both A) Neu5Ac and B) Neu5Gc are labeled in red numbers. The C11 of both Sias is outlined in a red box.

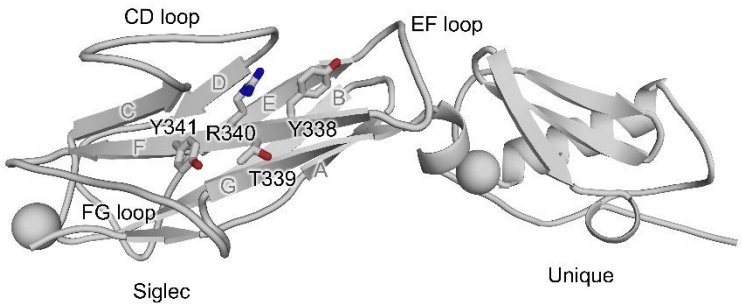

Figure v-2 Structural fold and binding motif overview of Siglec-like binding regions (SLBRs).

The Siglec binding region of SLBRHsa is shown in gray cartoon. The beta strands are labeled in gray. The selectivity loops (CD, EF, and FG) surrounding the sialoglycan binding site are labeled in black. The ΦTRX motif of is shown in grey sticks and the residues are labeled in black font.

Table 1. Crystallographic data collection and refinement statistics for $SLBR_{Hsa}$ bound to Neu5Acα2-3Gal and Neu5Gcα2-3Gal. Values in parentheses are for the highest resolution shell. Structure factors and final model are deposited with the Protein Data Bank (PDB) with accession codes listed. Raw data are deposited with SBGrid and can be accessed at: data.sbgrid.org/dataset/DATAID.

Ligand	**Neu5Ac-Gal**	**Neu5Gc-Gal**
PDB entry	8ST5	8ST6
DATAID	1020	1019
Resolution	50 Å - 1.45 Å	1.30 Å
Highest resolution shell	1.48 Å – 1.45 Å	1.32 Å – 1.30 Å
Data collection		
Beamline	SSRL 9-2	SSRL 9-2
Wavelength	0.97946 Å	0.97946 Å
Space group	$P2_12_12_1$	$P2_12_12_1$
Unit cell	a=46.1 Å	a=46.6 Å
	b=57.7 Å	b=58.1 Å
	c=76.0 Å	c=76.0 Å
R_{sym}	0.086 (0.490)	0.079 (0.465)
R_{pim}	0.026 (0.204)	0.023 (0.216)
I/σ	37.97 (1.81)	48.1 (1.8)

Completeness (%)	90.4% (47.1%)	91.3% (46.5%)
Redundancy	11.3 (5.9)	11.0 (4.9)
$CC_{1/2}$	1.00 (0.964)	1.00 (0.963)
Refinement		
R_{cryst}		0.1710
R_{free}		0.2081
No. Mol per ASU	1	1
RMS deviation		
bond lengths (Å)		0.006
bond angles (°)		1.399
Ramachandran (%)		
Favored		97.52
Allowed		2.48
Outliers		0

Results

Structures of Neu5Acα2,3Gal and Neu5Gcα2,3Gal bound $SLBR_{Hsa.}$

There is not a global conformational change upon ligand binding. Further, the only significant difference in local regions of the protein is within the position of the EF loop (residues 330 – 336), which shifts by 5.6 Å between the structures **(**Figure v-3**a)**. The shift in the position of the EF loop has been observed in structures of homologous ligand-bound bacterial adhesins. However, this conformational difference is unlikely to affect Ac/Gc selectivity. The EF loop of $SLBR_{Hsa}$ does not approach the C11 position of either form of sialic acid. Moreover, past work identified that crystal contacts may affect the position of this loop. Simulations of Neu5Ac-bound SLBRHsa within the crystal lattice demonstrate a large population occupying the open state **(**Figure v-3**b)**. Contrastingly, simulations of a monomer Neu5Ac-bound $SLBR_{Hsa}$ show a large population occupying a closed conformation **(**Figure v-3**c).** The differences in $SLBR_{Hsa}$ conformation between different

environments demonstrates that the open and closed conformations are affected by crystal contacts rather than Sia preference.

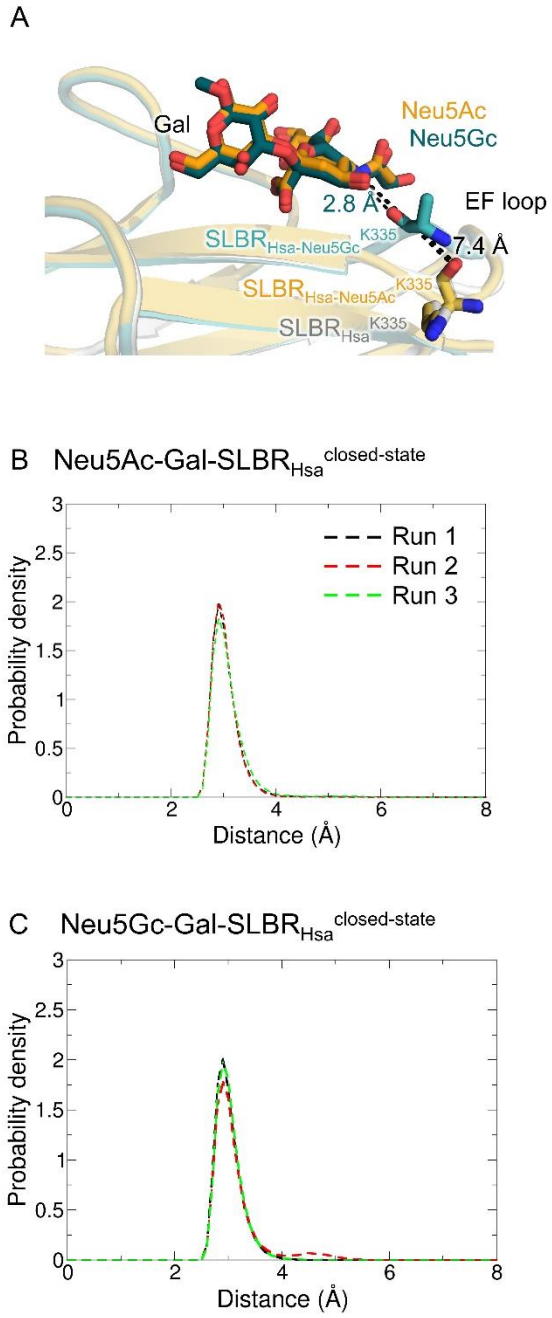

Figure v-3 The EF loop of SLBRHsa can adopt an open or closed conformation.

a) SLBRHsa unliganded, Neu5Acα2,3Gal and NeuGcα2,3Gal-bound structures are shown in *grey*, *yellow*, and *teal* cartoon respectively with $SLBR_{Hsa}^{K335}$ of each structure shown in sticks. The two sialoglycans Neu5Acα2,3Gal and Neu5Gcα2,3Gal are shown in *gold* and *turquoise* sticks respectively. There is no density for the side chain of residue K335 of Neu5Gc-bound $SLBR_{Hsa}$, so the model only includes the backbone and Cβ atoms of K335. The distances (D) between Lys 335 backbone carbonyl and Neu5Ac-O4 atom are shown with a

dark teal and *gold* dotted line c, d) Probability distributions of the distance (D). Simulations of the monomer Neu5Ac-bound SLBRHsa suggest that the protein mostly takes on a closed conformational state. This suggests that the open conformation of SLBRHsa observed in the crystal structure and simulations in the crystal lattice is stabilized by crystal contacts. The open and closed state therefore likely do not affect Sia preference. All simulations were performed in triplicate.

$SLBR_{Hsa}$ does not make direct contacts with the C11 hydroxyl.

It has been hypothesized that a direct hydrogen bond is required to confer Neu5Gc selectivity. It is not known if promiscuous binding SLBRs also contact the C11 hydroxyl. Further, it is unknown whether residues that do contact the C11 hydroxyl are the same across different adhesins. Many streptococcal adhesin SLBRs contain a conserved sialic acid recognition motif, ΦTRX. Mutagenesis studies have shown that changes to this motif can lower affinity for Neu5Ac linked sialoglycans significantly. However, less is known about how this motif interacts with other forms of sialic acid such as Neu5Gc. X-ray crystallographic structures demonstrate that the ΦTRX motif of the Neu5Gc selective $SLBR_{SrpA}$ makes the same hydrogen bond contacts to Neu5Ac and Neu5Gc. This suggests that the ΦTRX motif does not confer Sia preference. The ΦTRX motif of the promiscuous $SLBR_{Hsa}$ (YTRY) varies from the motif of $SLBR_{SrpA}$ (FTRT). Even with these sequence and ligand preference differences, the $SLBR_{Hsa}$ motif makes the same hydrogen bond contacts to Neu5Ac and Neu5Gc, similar to $SLBR_{SrpA}$. Together, evidence from both SLBRs suggest the ΦTRX motif recognizes a range of Sias and does not confer preference within these SLBRs.

While not in the ΦTRX motif, $SLBR_{SrpA}$ does form a hydrogen bond to the C11 hydroxyl group unique to Neu5Gc. Residue Y368, located on the G strand just below the conserved sialic acid recognition motif, interacts with the C11 hydroxyl with a bond length of 2.6 Å (PDBID 5EQ3)**(Fig. 4a)**. As validated by site-directed mutagenesis, this hydrogen-bonding interaction is key for promoting Neu5Gc selectivity. While this may confer selectivity in $SLBR_{SrpA}$, it is unknown whether promiscuous SLBRs directly contact the C11 hydroxyl of Neu5Gc. To determine if $SLBR_{Hsa}$ makes any direct contacts with the C11 hydroxyl of Neu5Gc, we measured the distance of the non-hydrogen atoms surrounding the C11 binding pocket. Intriguingly, the C11-hydroxyl group does not contact any residues in the binding pocket either directly or indirectly **(Fig. 4b)**.

Outside of the FTRX motif, the closest two atoms to the C11 hydroxyl are the D1 oxygen and D2 nitrogen of the N333 side chain of at 4.4 and 4.6 Å respectively **(Fig. 3b)**. These distances are significantly longer than the typical length of a hydrogen bond (2.8-3.2 Å). The residue structurally homologous to $SLBR_{SrpA}^{Y368}$ is $SLBR_{Hsa}^{V370}$. This residue is not surface exposed and is hydrophobic, preventing a hydrogen bond with the C11 hydroxyl. Similarly, the Neu5Ac preferring adhesin, $SLBR_{GspB}$, contains a buried isoleucine in this position.

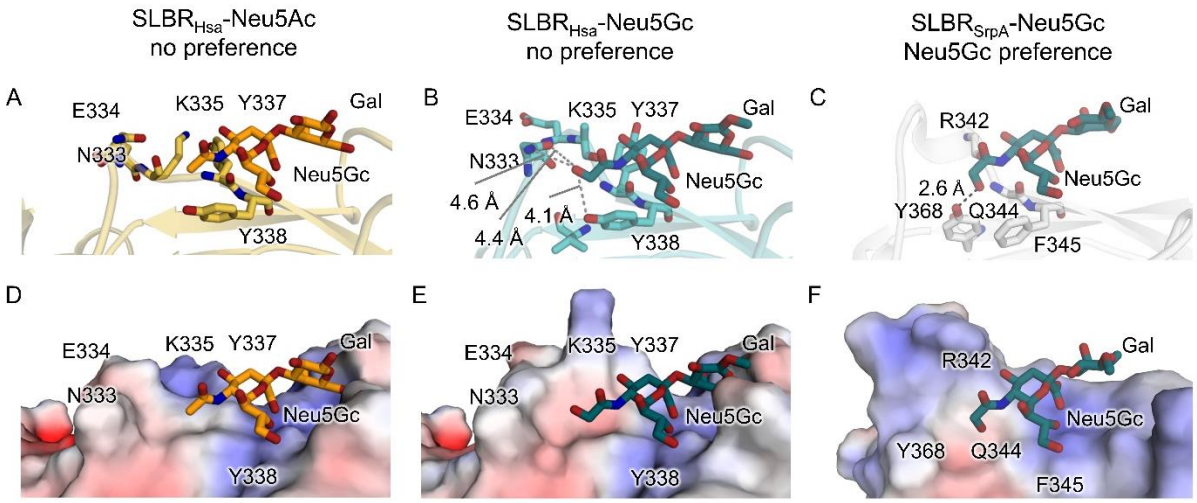

Figure v-4 Binding site comparison of SLBRs with a range of Sia selectivity.

Binding pocket of a) Neu5Ac bound-$SLBR_{Hsa}$ (*yellow*, (PDB 5EQ3)); b) Neu5Gc bound-$SLBR_{Hsa}$ (*teal*); c) $SLBR_{SrpA}$ adhesins are shown in cartoon. Glycans (Neu5Ac shown in gold, Neu5Gc shown in dark teal) and residues surrounding the glycan are shown in sticks. Nitrogens and oxygens are shown in blue and red respectively. d-f) Surface view of the binding pocket bound to Neu5Gc,Gal (turquoise sticks). e-f) The distance between residues surrounding the C11 hydroxyl and the C11 hydroxyl are shown in dashed lines with the distance labeled in Ä. Hydrogen bonds are shown in black and longer distances are shown in gray.

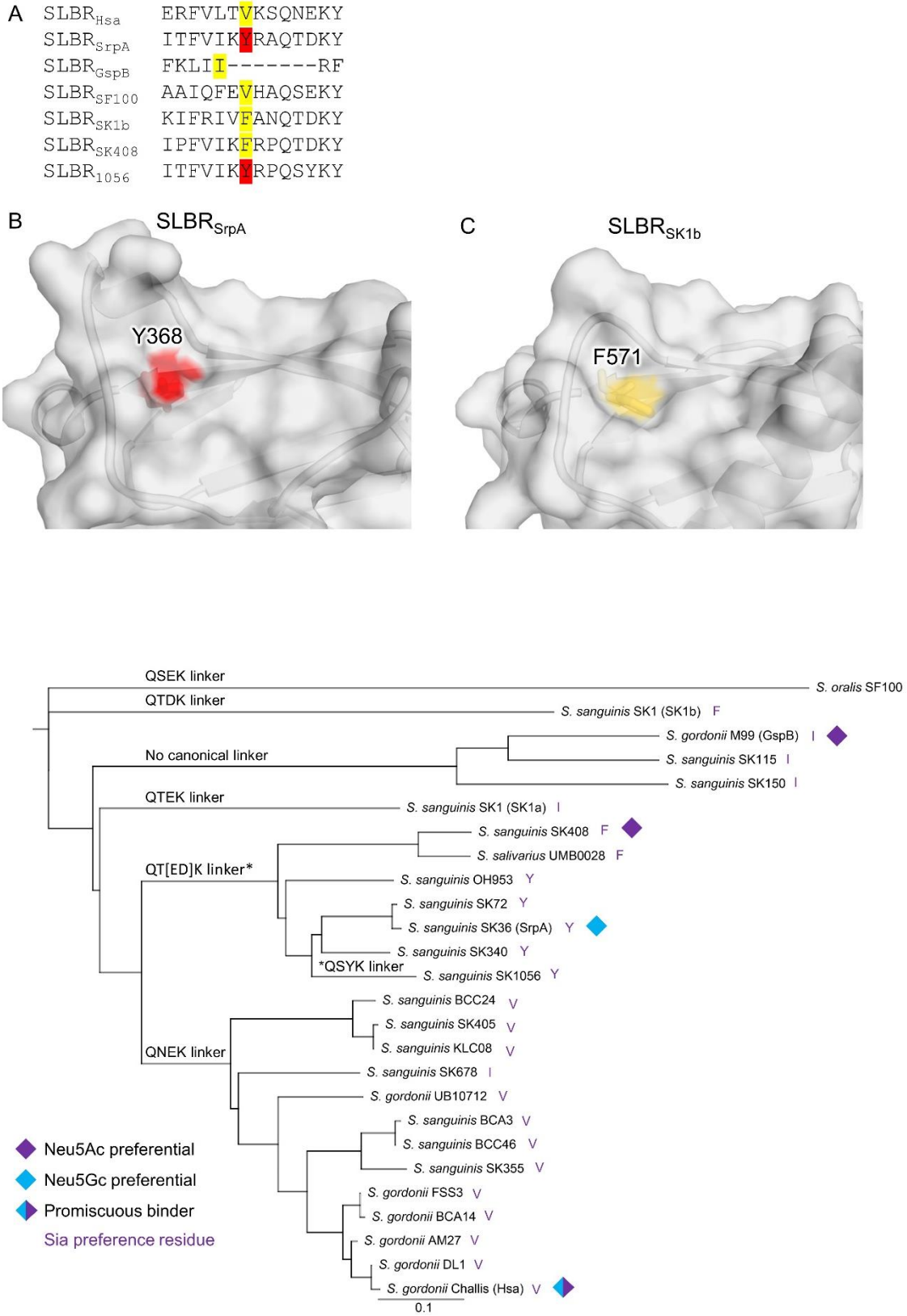

Figure v-5 Sia selectivity may be encoded by different mechanisms in different SLBR subfamilies.

Discussion

Together, the results provide a molecular basis for binding of *S. gordonii* strain Challis to both human and animal receptors. It further suggests how sialic acid-binding proteins support broad selectivity for sialic acid. Bacterial selectivity for Neu5Ac or Neu5Gc can determine which mammalian species bacteria can infect.[159] It has previously been reported that $SLBR_{SrpA}$, the sialic acid binding region of a serine rich repeat adhesin from Streptococcus sanguinis, is selective for Neu5Gc sialylated glycans. A single, pocket accessible tyrosine residue, $SLBR_{SrpA}{}^{Y368}$, hydrogen bonds with the C11 hydroxyl unique to Neu5Gc. Here, we name this residue the Sia preference residue (SPR). While previous mutagenesis work has shown the SPR to be critical for Neu5Gc selectivity in $SLBR_{SrpA}$, it is unclear if this single residue confers Sia selectivity across other adhesins as well. There are not any hydrogen bond donors or acceptors in Neu5Ac that Neu5Gc does not also have. This means that no hydrogen bonds to sialic acid can exist to confer a preference for Neu5Ac over promiscuously binding Neu5Ac and Neu5Gc. The lack of a polar, pocket accessible SPR does not account for the differences between Sia promiscuous SLBRs like $SLBR_{Hsa}$ and Neu5Ac selective SLBRs such as $SLBR_{GspB}$. Preference for Neu5Ac is therefore likely from negative selection.

To see if the Sia preference residue (SPR) has sequential diversity that can support a spectrum of Sia preferences, the SLBRs of several streptococcal serine rich repeat adhesins were aligned (**Fig. 6a**). Interestingly, phylogenetic clades can mostly be differentiated by the sequence of the linker between the Siglec and Unique domains. The clade with QNEK linkers including $SLBR_{Hsa}$ contain either a valine or isoleucine SPR.

The last large phylogenetic clade, the one that holds $SLBR_{SrpA}$, contains SLBRs with a QT[ED]K linker. Many members of this clade have a tyrosine SPR similar to $SLBR_{SrpA}$. These adhesins have recently been shown to be Neu5Gc selective. Some clade members in addition to $SLBR_{SK1b}$, however, have a phenylalanine SPR. This residue in the $SLBR_{SK1b}$ experimental model and AlphaFold predicted models of $SLBR_{SK408}$ and $SLBR_{UMB0028}$ are slightly pocket exposed. These residues may therefore behave more promiscuously or even Neu5Ac preferential.

$SLBR_{Hsa}$ is not the only protein found to bind promiscuously to different sialic acids. Other proteins from bacterial pathogens have binding sites similar to $SLBR_{Hsa}$ that lack a specific interaction to the C11 hydroxyl of Neu5Gc. Previous crystal structures of periplasmic binding protein (SatA) from the pathogenic bacterium *Haemophilus ducreyi*[162] show that both Neu5Ac and Neu5Gc have similar interactions in the binding pocket. The C11-hydroxyl of Neu5Gc does not form specific electrostatic interactions. Similar results are also shown for Neu5Ac and Neu5Gc binding in case of m-CD22 and h-CD22.[163]

vi. SIALYLTRANSFERASES

Introduction

Many epithelial cancers aberrantly express a disaccharide known as sTn on their cell surface. sTn overexpression positively correlates with a poor prognosis and more aggressive cancer characteristics such as local invasion and lymph node metastasis. A mechanism linking sTn overexpression to these biological phenomena has not yet been identified. One function sTn may serve is immune evasion. Sialoglycans similar to sTn are recognized by immune receptors known as Siglecs. Once activated, Siglecs inhibit an immune response from being initiated. In this way, cancer cells with greater amounts of sTn on their cell surface can avoid immune detection. These findings pose sTn as a potent target for both cancer detection and treatment.

First efforts to target sTn were through antibody and vaccine development. Immunization with cancer cells, the disaccharide sTn, and sTn conjugated carrier proteins yielded antibodies that are either not specific enough or failed to produce an efficacious effect in clinical trials. More recent efforts have been focusing on using lectins for glycan detection. A recent study aimed to characterize glycan binding proteins for several different glycan characteristics. While bacterial glycan binding proteins have the potential to serve as probes for many different types of glycans, no known bacterial binding proteins are specific for α2,6 linked Neu5Ac, much less sTn specifically. The most well-known and commonly used α2,6 linked Neu5Ac specific binding protein is a plant lectin from *Sambucas nigra* (elderberry).

Rather than using a less stable and expensive to produce plant protein, we can instead use a bacterial enzyme with a sialoglycan product. Given that a chemical product is bound to an enzyme before dissociating, a sialoglycan enzyme product should have a native affinity for the enzyme. A class of enzymes called sialyltransferases catalyze the transfer of a nine-carbon backbone monosaccharide known as sialic acid, to a

growing glycan chain. They convert the donor substrate, CMP-sialic acid, and acceptor substrate, an asialoglycan, to products CMP and a sialoglycan. Some sialyltransferases specifically produce sialoglycans with certain linkages. These can be used as scaffolds for a binding protein that is specific for a specific linkage of Neu5Ac. Some bacteria express α2,6 linkage specific sialyltransferases making them an ideal scaffold for sTn probe development. Additionally, it has previously been shown that a single mutation of a catalytic aspartate residue completely eliminates catalytic activity. This makes converting bacterial sialyltransferases to binding proteins straight-forward. One bacterium, *Pasteurella multocida*, expresses a sialyltransferase with dual α2,3 and α2,6 activity. While PMST has dual activity it makes an ideal starting scaffold because the wildtype has previously been structurally and biochemically characterized. Structures of apo, CMP bound, CMP-sialic acid inhibitor bound, and CMP and lactose bound are publicly available. Additionally, two structures, CMP-bound open state and CMP-bound closed state $PMST^{D141N}$ are publicly available in the PDB. Here, I show with crystal structures of $PMST^{D141N}$ in an apo and CMP-bound half-open state, that the overall protein structure and dynamics are not likely affected by mutation of the catalytic aspartate. Additionally, with a glycan microarray and ELISAs, I demonstrate that $PMST^{D141N}$ does bind sialoglycans.

Results

$PMST^{D141N}$		
ligand	Apo (8U9I)	CMP (8SV2)
SBGrid Entry	1021	1018
Resolution (Å)	1.55	1.85
Beamline	21-ID-F	21-ID-D
Wavelength (Å)	0.979	
Space group	P2	$P2_12_12_1$
Unit cell		
a (Å)	63.92	67.60
b(Å)	60.95	66.40
c(Å)	96.17	96.38

R_{sym}	0.073 (0.550)	0.081 (0.418)
R_{pim}	0.041 (0.310)	0.024 (0.136)
I/σ	17.63 (2.37)	47.83 (4.15)
Completeness (%)	86.1 (86.0)	99.9 (99.9)
Redundancy	3.9 (3.7)	12.4 (10.2)
CC1/2	1.000 (0.702)	0.999 (0.961)
Refinement		
R_{cryst}	0.179	0.216
R_{free}	0.212	0.254
No. Mol per ASU	2	1
RMS deviation		
bond lengths (Å)	0.009	0.011
bond angles (°)	0.980	2.089
Ramachandran (%)		
Favored	97.56	95.85
Allowed	2.44	3.63
Outliers	0	0.52

Table vi-1 Data Collection Statistics for X-ray crystallographic structures.

The highest resolution shell is 1.98-1.95.

	2EX0 apo PMST	2EX1 PMST CMP-bound, closed state	2iib – $PMST^{D141N}$, CMP bound, closed state	2ii6 – $PMST^{D141N}$, CMP bound, open state
Apo $PMST^{D141N}$	0.15 Å	0.80 Å	0.820 Å	0.707 Å
CMP-bound, half-open $PMST^{D141N}$ structure	0.72 Å	0.48 Å	0.446 Å	1.262 Å

Table vi-2 CMP binding to $PMST^{D141N}$ causes small global conformations.

The RMSD values were calculated using the align function of Pymol.

	2iib – $PMST^{D141N}$, CMP bound, closed state	2ii6 – $PMST^{D141N}$, CMP bound, open state
Apo $PMST^{D141N}$	0.444 Å	0.371 Å

New CMP bound PMST^{D141N} structure	0.265 Å	0.332 Å

N-terminal domain of PMST^{D141N} has little to no global structural changes. These values are within the error of the x-ray crystallographic diffraction data. The RMSD values were calculated using the align function of Pymol.

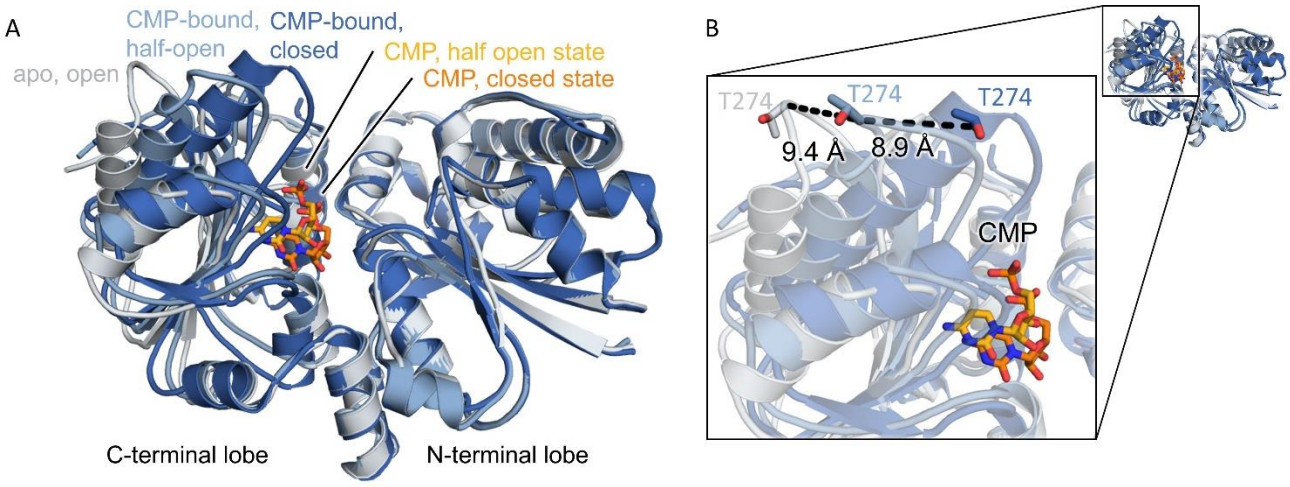

Figure vi-1 PMST^{D141N} closed state equilibrium depends on CMP binding and pH.

PMST^{D141N} apo and open state, CMP-bound half open state, and CMP-closed state (PDB 2IIB) are shown in *light*, *medium*, and *dark periwinkle* cartoon respectively. CMP is shown in sticks in gold and orange for the CMP-bound half open state and CMP-bound closed state respectively.

To understand how mutating the catalytic residue of PMST affects CMP binding, PMST^{D141N} was co-crystallized in the presence and absence of CMP. Using x-ray crystallography, models of apo and CMP-bound PMST^{D141N} were made. Apo PMST and PMST^{D141N} and CMP-bound PMST and PMST^{D141N} are structurally similar with rmsd values of only 0.15 and 0.48 Å respectively (Table vi-2). In contrast to both apo and CMP-bound PMST (PDB 2EX0 and 2EX1) and both previously solved CMP-bound open and closed state PMST^{D141N} structures (PDB 2II6 and 2IIB), the CMP-bound structure reported here is halfway between an open and closed state. When the N-terminal domains are aligned, the distance between Cαs of T274 in helix 16 of the apo open PMST^{D141N} structure and the halfway open CMP-bound PMST^{D141N} structures reported here is 9.4 Å. The Cα of T274 of CMP-bound half-open state PMST^{D141N} is also 8.9Å away from that of CMP-bound closed state PMST^{D141N} (PDB 2IIB). This wide range of motion between an open and fully closed state has been noted for wildtype PMST. With the structures reported here taken together with the CMP-bound closed state PMST^{D141N} (PDB 2IIB), it is likely that the range of motion of helix 16 in the catalytically inactive PMST is similar to that

of wildtype. It has been reported that a more acidic pH of 5.4 is optimal for producing α2,6 sialoglycans while a more basic pH of 6.4 is optimal for producing α2,3 sialoglycans.[164] Differences in pH could explain the differences in conformational stated between the CMP-bound half open state structure reported here and the CMP-bound open and closed state structures available in the PDB, however, all crystals were grown at a similar pH. The apo PMST^{D141N} structure and open state PMST^{D141N} (PDB 2II6) structure grew at pH 7.7. CMP-bound half open state was grown at 7.1 pH. PDB 2IIB is reported to have grown at pH 7.5.

CMP binds PMST^{D141N} similarly in an open and closed state. CMP bound to a closed state conformation of PMST^{D141N} appears to bind deeper in the cleft between the N and C-terminal domains (Figure vi-1). However, hydrogen bond and electrostatic interactions between CMP and PMST^{D141N} in both conformational states remain the same (Figure vi-2). The lengths of these bonds vary slightly, but the differences are within the error of the x-ray crystallographic data.

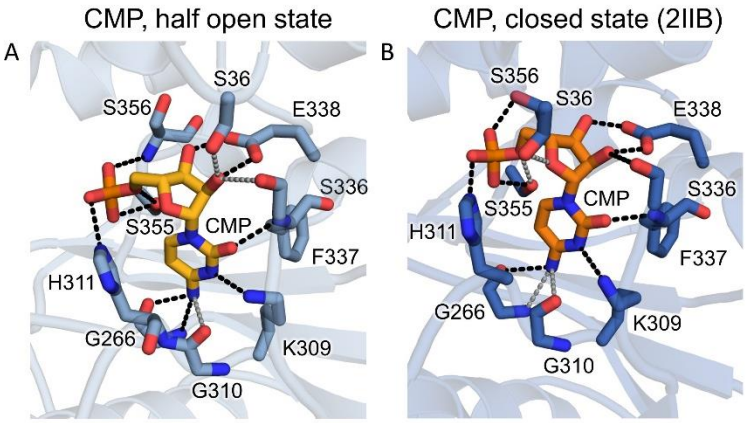

Figure vi-2 CMP hydrogen bonding patterns are the same in half-open and closed states.

A-B) CMP of A) CMP-bound half-open state and B) CMP-bound closed state shown in gold and orange sticks respectively. Hydrogen bond contacts (≤ 3.2Å) are shown in black dashed lines and electrostatic interactions (3.2Å<, ≤3.9Å) are shown in grey dashed lines. Residues that hydrogen bond with CMP are also shown in sticks. Nitrogen atoms are shown in blue and oxygen atoms are shown in red. Hydrogen bonding patterns are nearly identical between the CMP-bound halfway-open and closed states.

Glycan Array Analysis

To determine any acceptor ligand selectivity PMST^{D141N} may have, we screened PMST^{D141N} for binding against a panel of 562 glycans known to be present in mammals. A relatively small number of glycans bind greater than the average glycan. While this could suggest that PMST^{D141N} has narrow selectivity, there is a single clear motif common amongst the glycans that PMST binds. Many of the preferred glycans have a terminal galactose with an alpha 1,3 linkage. Four strongly binding glycans contain a terminal Galα1,3(Fucα1,2)Galβ (**Error! R eference source not found.**, glycans 102, 104, 107, and 456). This indicates that PMST^{D141N} has a large enough acceptor substrate binding pocket to accommodate larger, fucosylated glycans. Four others strongly binding glycans have a terminal Galβ1-4GlcNAc. While not in the top ten preferred glycans, there is evidence that suggests PMST^{D141N} may bind sialylated glycans (**Error! Reference source not found.**, glycans 258, 465, and 467). It is not clear whether these sialoglycans bind with Neu5Ac in the CMP-Neu5Ac binding pocket or only within the acceptor substrate binding pocket. Two of the moderately binding sialoglycans are α2,3 linked while the other is α2,6 linked. PMST^{D141N} did not bind any glycans with a terminal Neu5Acα2,8 including polysialic acids. PMST is known to produce Neu5Ac-α2,3 and -α2,6 linkages, but not α2,8 sialoglycans, which is congruent with the order of binding preference of sialoglycans seen in the array.[164, 165, 166] Additionally, PMST^{D141N} cannot synthesize polysialic acid, therefore it is unlikely that the acceptor substrate binding pocket can accommodate sialic acid. Taken together it is unlikely that a terminal Neu5Ac can bind within the donor substrate binding pocket.

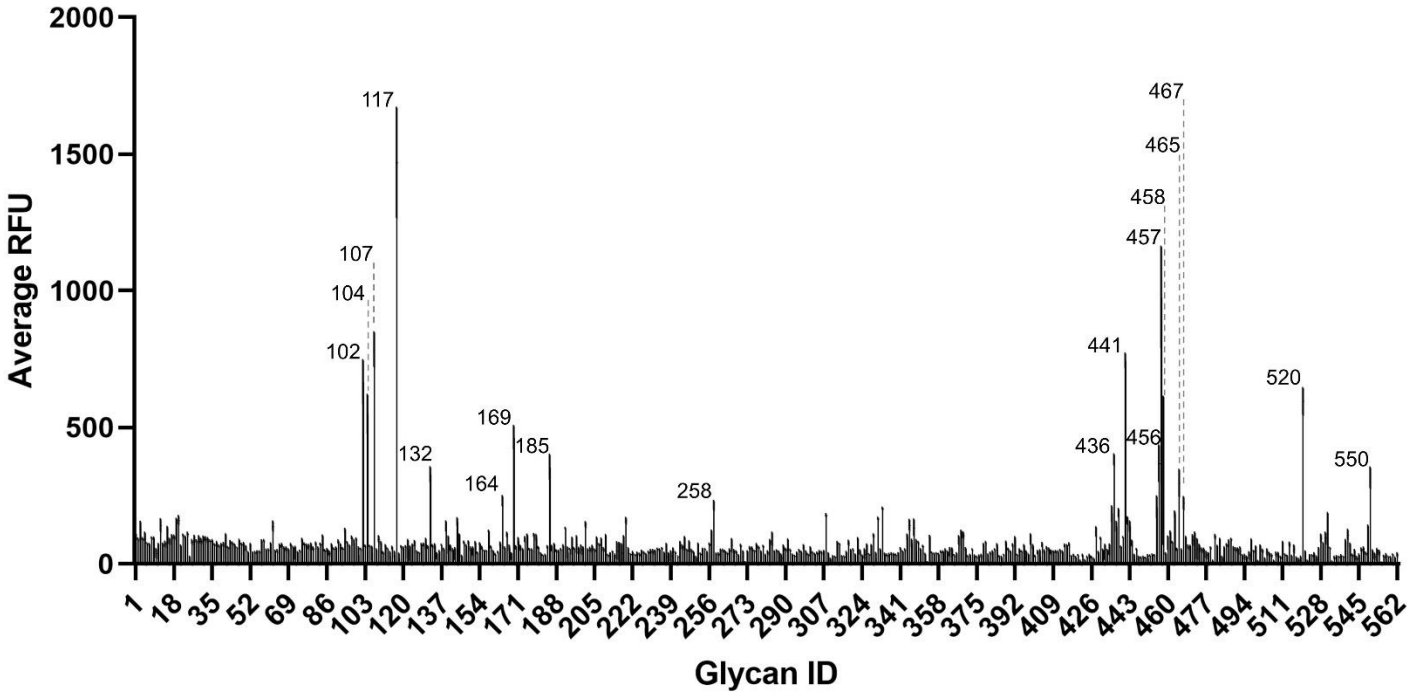

Figure vi-3 Microarray analysis of PMSTD141N ligand preference.

Glycan binding is measured by average fluorescent units. Standard deviation is shown in black bars. Glycan IDs for top binding glycans are shown above their respective bars. Bar height represents an average value of six replicates.

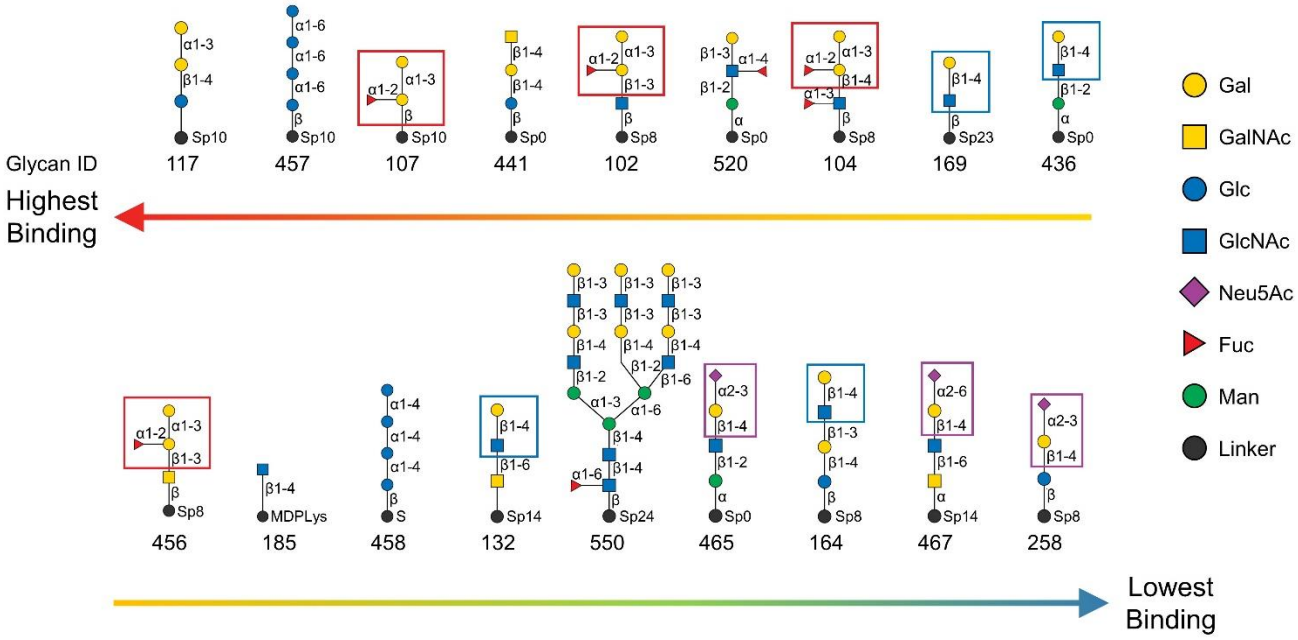

Figure vi-4 PMST^{D141N} prefers ligands that have terminal Neu5Ac, *N*-acetyllactosamine, or fucosylated galactose disaccharides.

Structures for glycans detected as strong binders are depicted using standard symbol nomenclature for glycans. While most strongly binding glycans contain a terminal galactose, a few others contain a terminal sialic acid (Neu5Ac).

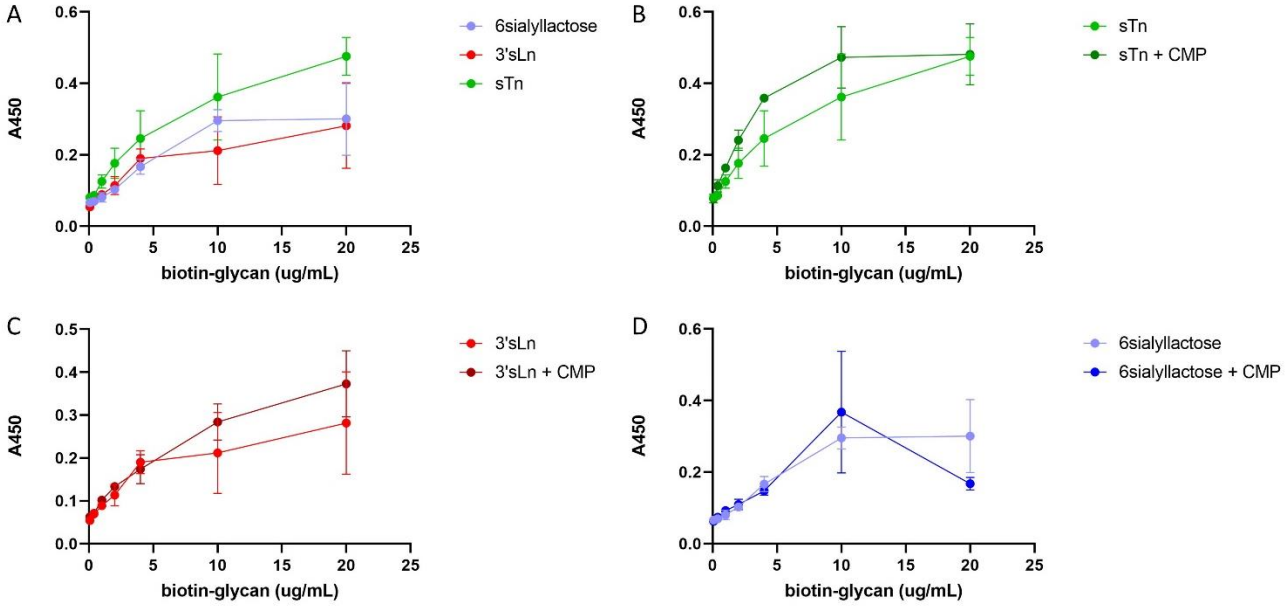

Figure vi-5 PMST^{D141N} binds both α2,3 and α2,6 linked sialoglycans.

A-D) ELISAs of sialoglycans binding PMST^{D141N}. Data points represent the mean of three replicates. Standard deviations are shown in bars. Darker colors indicate the assay was performed in the presence of 100 μM CMP.

The curves in panel A are also shown in panels B-D. 6'sialyllactose is Neu5Acα2,6Galβ1,4Glc. 3'sLn is Neu5Acα2,3Galβ1,4GlcNAc. sTn is Neu5Acα2,6GalNAc.

To confirm that $PMST^{D141N}$ can bind sialoglycans an ELISA was performed. $PMST^{D141N}$ can bind both α2,3 and α2,6 sialoglycans (**Figure vi-5A**). Wildtype PMST binds its substrates in a sequential order. First, CMP-sialic acid binds, inducing a conformational change. Then the acceptor substrate, a glycan, binds in the cleft between the N- and C-terminal domains, just above CMP. CMP alone is also capable of inducing this same conformational change as seen in crystal structures. It is not clear whether sialic acid is also capable of inducing these conformational changes. Without these conformational changes an acceptor substrate won't bind. To test whether $PMST^{D141N}$ would have a lower binding strength for sialoglycans in the absence of CMP, ELISAs for each sialoglycan were performed in the presence and absence of CMP. Here, there are not statistically significant differences in binding for the sialoglycans in the presence or absence of CMP (**Figure vi-5**). It could still be that $PMST^{D141N}$ has a lower affinity for sialoglycans in the absence of CMP, but the dynamic range of the assay is too small to detect these differences. It could also indicate sialoglycans are capable of inducing the same conformational change observed upon binding of CMP. Further assay optimization is required to determine this.

vii. DISCUSSION

Application of bacterial adhesins and sialyltransferases to additional sialoglycans

While the bacterial adhesins and sialyltransferases presented here offer a promising solution for probe development against the sialoglycans discussed here, they will not likely be a one size fits all scaffold for targeting other glycans. Whether natively an enzyme or a binding protein, glycan binding proteins are generally specific for a single or at most two sialic acid linkages. Of the proteins presented here, the bacterial adhesins can be used for α2,3 linked sialoglycans and the sialyltransferases used for α2,6 and α2,3 linked sialoglycans.

However, sialic acid can also be linked with an α2,8 linkage is a sialoglycans. For these sialoglycans, α2,8 linkage specific bacterial sialyltransferases could be adapted. Polysialyltransferases (polySTs) from bacteria such as *Escherichia Coli* K1 and K92[167], *Neisseria meningitidis* serogroup B[168], and *Pasteurella haemolytica* A2[169] catalyze the production of α2,8 linked polysialic acid chains from CMP-Neu5Ac. Bacterial polySTs comprise the glycotransferase family 38 (GT38) for which there are 159 sequences discovered to date.[168] While polysialyltransferases should recognize α2,8 linked sialic acid, they will need to be engineered to recognize reducing end units other than sialic acid to target a larger range of sialoglycans.

Other than the linkage between sialic acid and the nonreducing end glycan unit, it would be beneficial to develop probes for specific types of sialic acid. Most of the work presented here focuses on recognition of the sialic acid Neu5Ac. This sialic acid is a whopping majority of the total sialic acid found in humans. The second most common is Neu5,9Ac_2.[3] In addition to these two sialic acids, Neu5Gc is commonly found in other mammals. While humans cannot synthesize Neu5Gc, this sialic acid has been detected in human tissue. It is likely that any Neu5Gc found in humans is absorbed through their diet. While the α2,3 linkage specific streptococcal serine rich repeat adhesin SLBRs containing a tyrosine for the sialic acid preference residue such as $SLBR_{SrpA}$ prefer binding Neu5Gc, alternatives for sialic acids other than α2,3 linked Neu5Ac and Neu5Gc or α2,6 linked Neu5Ac should be explored. A recent study found bacterial glycan binding proteins that were specific for either all sialic acids, Neu5Ac, α2,3 linked sialic acids ($SLBR_{Hsa}$; discussed in Chapter IV), 4-*O*-acetykated sialic acids, Neu5Gc, α2,6 linked sialic acids, 7-*O*-acetyated sialic acids, α2,8 linked disialic acid, or 9-*O*-acetylated sialic acids.[3] In addition to these proteins, an α2,6 sialyltransferase and mutants with neosialidase activity from *Photobacterium damselae* catalyze reactions with glycans containing Neu5Ac and Neu5Gc.[170] Sialic acid synthetases could be developed as probes but may not be practical due to steric hinderance. Structures of N. meningitis CMP-sialic acid synthetase show that sialic acid binds in a deep pocket.[171] While the C2 atom faces the entryway of the binding site, there is likely not sufficient space for a longer glycan chain to bind. These proteins likely cannot bind a polysaccharide.

Recognizing different lengths of sialoglycans is important for more than just affinity. Glycosylation is a highly regulated cell process. Specific patterns of glycans are displayed on cell surfaces based on the expression level and localization of certain glycosylation enzymes. These expression levels can affect the length of glycans produced, which could indicate an important biological difference in cell behavior. For example, sTn found in abundance on the surface of many human epithelial cancers is a disaccharide with the structure Neu5Acα2,6GalNAc. Common mutations in epithelial cancers lead to termination of glycans at shorter lengths with sialic acid. Noncancerous human cells may also display longer sialoglycans with Neu5Acα2,6GalNAc at the nonreducing terminus. If a probe recognizes Neu5Acα2,6GalNAc as a terminal sialoglycan unit, but not as an entire glycan chain attached to an amino acid, then sTn and longer Neu5Acα2,6GalNac terminated glycans would be detected similarly. Probes that can detect sialoglycans longer than disaccharides are also needed for diagnosing other diseases. For example, the trisaccharide sLe^X has been associated with human diseases such as acute lung injury,[172] triple negative breast cancer,[173] and other immune conditions involving selectins.[174] In these cases, recognition of three nonreducing end glycan units is necessary to distinguish selectively bind sLe^X.

Some of the streptococcal SRRP SLBRs presented here can recognize all three units of some trisaccharides. $SLBR_{Hsa}$ makes contacts to the *N*-acetylated galactose of sTa, the *N*-acetylated glucose of 3'sLn, sLe^C, and 6S'sLe^X (Figure iv-13Figure iv-13 Sialoglycan position in the $SLBR_{Hsa}$ binding pocket.). Some adhesins that have a wider, more open binding pocket and a distal arginine may bind longer glycans as well. Both $SLBR_{SrpA}$ and $SLBR_{SK1b}$ have small CD and FG loops. These smaller loops leave the binding pocket over the ΦTRX motif to be contiguous with the rest of the length of the protein surface. This length leaves for room where additional saccharide units could contact a larger area of the SLBR surface. Some SLBRs with an open binding pocket have an arginine just C-terminal of the ΦTRX motif. When this residue is mutated to a glutamate in $SLBR_{SrpA}$, binding to platelet monolayers is reduced to just a third of wildtype binding. This arginine has been termed the distal arginine. It is conserved strongly in mammalian Siglecs such as Siglec 5, indicating that this residue may facilitate binding of longer saccharide chains to bacterial adhesins. While evolutionary conservation supports this hypothesis, sialoglycans do not contact this arginine in crystal structures of $SLBR_{SrpA}$ and $SLBR_{SK1b}$. It is

therefore unlikely that sialoglycan binding to the area surrounding the distal arginine is a much lower affinity binding pocket than over the ΦTRX motif.

It is unlikely that sialyltransferases recognize sialoglycans further than a third reducing end carbohydrate unit. While binding and kinetic data for trisaccharide acceptor substrates does not exist, crystal structures of lactose bound to bacterial sialyltransferases suggest the sialyltransferase would not bind a fourth reducing end carbohydrate unit. The acceptor substrate binding pocket is roughly the length of a linear disaccharide (Figure **vii-1**). Kinetic data from another bacterial α2,6 sialyltransferase from *Photobacterium leiognathi* strain JT-SHIZ-145 suggest that other sialyltransferases may only recognize the first monosaccharide at the non-reducing end.[44]

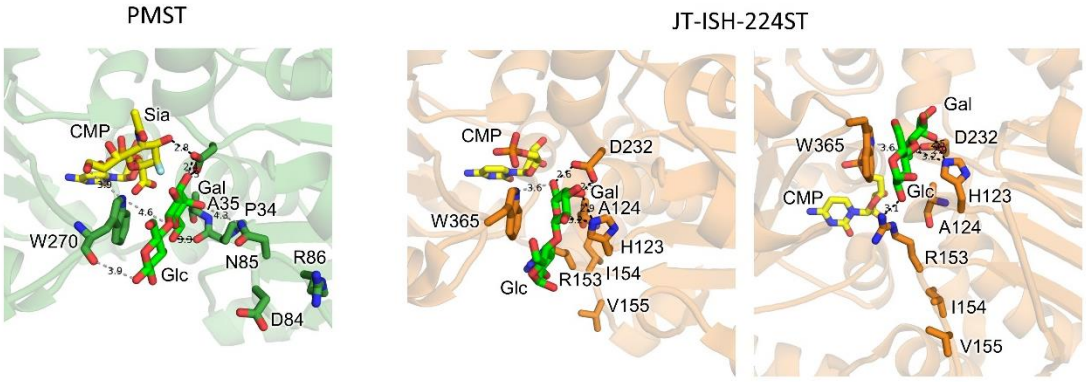

Figure vii-1 α2,6 linkage specific sialyltransferases binding lactose

A2,6 linkage specific sialyltransferases from *Pasteurella multocida* and *Photobacterium* JT-ISH-224 are shown in green and orange cartoon respectively. Amino acids surrounding the lactose ligand are shown in sticks. Lactose is shown in green sticks and CMP and CMP sialic acid are shown in yellow sticks. Oxygen atoms are colored red and nitrogen atoms are colored blue. JT-ISH-224ST makes one hydrogen bond contact with the glucose unit of lactose. PMST only makes one electrostatic interaction with the glucose unit of lactose.

Manipulating avidity to increase probe affinity and selectivity.

Avidity is an important biological principle that can help to develop selective, high affinity probes. Avidity occurs when two multivalent entities bind together. With binding to each successive binding site, the affinity for the next binding site is increased. Cells are multivalent in that they have many—and often branched—glycans

on the cell surface. Glycan binding proteins can be multivalent too if they can bind glycans with multiple binding sites.

Avidity was used successfully to develop antibodies against sTn. When immunized with purified sTn glycans, little to no antibodies targeting sTn are produced. To try and increase the immunogenicity of sTn and to develop antibodies that could bind multiple sTn molecules, a carrier protein with ~3000 sites glycosylated with sTn was used for immunization. This mimics a physiological sample more closely than does a purified glycan because local concentrations on the carrier protein and cell surface are much higher. Additionally, a carrier protein can present multiple sTn antigens in an orientation that could facilitate production of an antibody that can bind multiple sTn molecules.

An example of a bacterial protein with multiple glycan binding sites are AB_5 toxins such as the ones from shiga, typhoid, pertussis, and cholera. These toxins have a glycan receptor binding subunit, also known as the B subunit. that forms a homopentamer. Each B subunit can bind a glycan. Recently, it was discovered that a typhoid toxin pentamer of PltC can bind two glycans per subunit, increasing its valency to ten.[175] While the SRRP SLBRs studied to date do not form oligomers, they can likely be connected to increase valency. Two bacterial adhesins known to have tandem glycan binding regions are $SLBR_{SK1}$ from *Streptococcus sanguinis* and FapC from *Streptococcus oralis*.[51, 54] Glycans can bind both Siglec domains of $SLBR_{SK1}$ simultaneously, indicating that binding a biological target will likely demonstrate avidity. Both FapC and $SLBR_{SK1}$ also provide examples of suitable linkers for connecting these glycan binding regions that will produce functional multivalent probes. In addition to naturally occurring tandem binding domains, a tandem repeat of an engineered lectin selective for α2,6-sialylated N-glycans was developed that increased affinity to its glycan target.[176]

Another way to increase avidity in sialoglycan probes is to design probes that bind the linkage between the glycan and the amino acid residues of a target glycoprotein. Glycan binding proteins typically bind physiological targets with a much greater affinity than to purified sialoglycans. For the bacterial SRRP SLBRs,

this difference can be from tens of nanomolar to a glycoprotein to low micromolar for purified glycans. The difference in affinity is likely due to the presence or absence of a linkage to a protein. In addition to developing antibodies against sTn, efforts have been made to develop antibodies against an sTn glycosylated mucin known to be associated with cancer named MUC1. Binding studies between an anti-MUC1 antibody and a MUC1 glycopeptide library showed that the antibody targets a glycosylated threonine residue in a tandem repeat region. Additionally, neighboring glycosylation sites can increase or decrease affinity based on steric hinderance.[177] While specific linkages to a serine or threonine sidechain were not determined, these experiments do demonstrate that the 3D structure of the glycoprotein affects another proteins ability to bind a glycosylated site.

viii. FUTURE DIRECTIONS

Structural and biochemical studies to test the sialic acid preference residue hypothesis.

$SLBR_{SrpA}$ is known to have a stronger binding strength for Neu5Gc over Neu5Ac. A tyrosine in the G strand beneath the ΦTRX motif makes a hydrogen bond with the C11 hydroxyl of Neu5Gc. I have termed this residue the Sia preference residue (SPR). Other SLBRs in the same phylogenetic clade as $SLBR_{SrpA}$ have a phenylalanine rather than a tyrosine. Phenylalanines in the SPR position seem to be pocket exposed, creating a small hydrophobic patch. These SLBRs likely prefer Neu5Ac due to its lack of a nearby hydrophilic group like the Neu5Gc C11 hydroxyl. An ELISA or other technique that measures binding strength of protein-ligand interactions can be used to look at the relative difference in binding strength of adhesins with a phenylalanine SPR for Neu5Ac and Neu5Gc. The same difference in binding strength can also be measured for mutants with phenylalanine to tyrosine mutations. This could shift the ligand preference to Neu5Gc. For adhesins with a tyrosine SPR, phenylalanine mutants can be made to test if Neu5Gc preferring adhesins can be mutated to have a preference for Neu5Ac. Previously, the structurally homologous residue in $SLBR_{GspB}$ was mutated to a tyrosine, but the protein did not fold properly. $SLBR_{GspB}$ is less tolerant of mutations in general than $SLBR_{Hsa}$

and $SLBR_{SrpA}$. The mutants proposed here for the $SLBR_{SrpA}$ like proteins are either a phenylalanine to a tyrosine or tyrosine to phenylalanine. Since these residues are both bulky and only vary by a single hydroxyl group, these mutant proteins are likely to fold correctly. Gel filtration chromatography can be used to identify any mutants that are prone to aggregation.

Demonstrate binding of glycans identified as preferred ligands in the glycan array.

While the glycan microarray is a powerful tool to interrogate glycan ligand preference for a range of different types of glycan binding proteins, false positive and negative hits may be observed. Previous work from our lab using the sialic acid binding region of a bacterial adhesin $SLBR_{Hsa}$, showed no binding to any ligands, including some previously shown to bind with ELISAs. False negative results may occur due to steric hinderance from the linker attached to the glycan, orientation of the immobilized glycan, or poor protein quality. While unrealistic for the <500 glycans in the array, an ELISA could be used to validate the glycans identified as preferred ligands.

I have ordered some glycans representative of the features of the glycans identified as preferred ligands, namely: Galβ1-4Glcβ (lactose; Sigma 0046a), Galβ1-4GlcNAcβ (N-acetylated lactosamine; Sigma 0047), Galα1-3(Fucα1-2)Galβ1-4GlcNAcβ (Sigma 0007), and Galα1-3(Fucα1-2)Galβ (Sigma 0086). I have already collected some preliminary data with 6'sialyllactosamine, 3'sialyllactosamine, and sTn. Comparing lactose to lactosamine will demonstrate preference for a nonterminal glucose or N-acetylated glucosamine monosaccharide. Lactosamine can also be compared to 6' and 3' sialyl lactosamine to indicate a preference for sialylated versus nonsialylated glycans. Galα1-3(Fucα1-2)Galβ and Galα1-3(Fucα1-2)Galβ1-4GlcNAcβ will be used to confirm preference for an α1,2 fucosylated Galactose chain. Additionally, direct comparison of the two may indicate the number of monosaccharides that PMST can contact upon binding. If a preference for the longer glycan is demonstrated, it is likely that PMST makes favorable interactions with the third monosaccharide in the glycan chain.

There is one compound in the glycan microarray data that I believe is a false negative. As discussed in the results section of Chapter VI, the array data did not demonstrate $PMST^{D141N}$ binding to Neu5Ac. Given the number of hydrogen bond interactions between PMST and CMP-sialic acids inhibitors (~ 17 hydrogen bonds) with six of those contacting the sialic acid moiety, $PMST^{D141N}$ likely has a high affinity for Neu5Ac. Curiously, ELISAs using the conditions described in the Methods section, failed to demonstrate $PMST^{D141N}$ binding to biotinylated CMP-sialic acid. Crystal structures with PMST have shown binding of CMP, two CMP-sialic acid inhibitors, and lactose. One of these three compounds would serve as a great positive control for the ELISA assay. In addition to performing ELISAs with biotinylated lactose, I propose ELISAs with biotinylated Neu5Ac. A few compounds with various linkers are available through Sigma. Together these experiments will validate glycans identified as preferred ligands in the glycan microarray as well as begin to tease out selectivity for specific glycan moieties.

One additional use for an ELISA assay would be to test mutations that could increase selectivity for α2,6 over α2,3 linkages. It has previously been reported that $PMST^{P34H}$ has 980-fold greater production of 6'sialyllactose than wildtype PMST.[166] This mutant likely also has increased selectivity for binding α2,6 sialoglycans when made catalytically inactive. The binding strength of $PMST^{D141N}$ and $PMST^{D141N, P34H}$ for various 3' and 6' sialoglycans could be directly compared with an ELISA.

Determine the structural basis of ligand selectivity of $PMST^{D141N}$

Proceeding binding validation, non-biotinylated versions of any of these glycans can be cocrystallized or soaked with PMST crystals to determine the structural basis of selectivity. While structures are available of lactose bound to CMP-bound PMST using crystallization conditions similar to mine, I was unable to cocrystallize CMP-bound PMST with GalNAc or sTn. GalNAc may not have cocrystallized due to having a low affinity. While ligands are used in molar excess in cocrystallization and crystal soaking, ligands still need to have an affinity for the protein of interest. Lactose has previously and successfully been soaked with PMST crystals. Lactose, however, is a disaccharide and is shown as a preferred ligand in the glycan microarray data. Lactose likely has a higher affinity for PMST than does GalNAc explaining why soaking lactose but not

GalNAc may have been successful. As for sTn, it may not bind PMST when CMP is also bound. The sialic acid and CMP may have electrostatic repulsion within the binding site. If a sialylated glycan and CMP bound the sialyltransferase with a high affinity, product dissociation would proceed slowly making sialyltransferases poor enzymes. Moving forward, cocrystallization with the preferred glycan ligands may be successful with the conditions used in these experiments. If soaking or cocrystallization is not successful, screening for new conditions may be necessary.

Additional bacterial sialyltransferases suited for α2,6 sialoglycan probe development.

To develop a probe against sTn, multiple starting scaffolds will need to be tested. To select another sialyltransferase for engineering, I started with a BLAST search using the amino acid sequence from *Pasteurella multocida* alpha 2,6 sialyltransferase (PMST). Identifying similar sialyltransferases will likely experimentally be easier to work with as they are more likely to express and purify under similar conditions. The sialyltransferase most similar to PMST is an alpha 2,3 linkage specific sialyltransferase from *Pasteurella dagmatis*. This sialyltransferase has previously been characterized and crystallized. A variant from wild type with two mutations, M117A and P7H, has been shown to catalyze solely alpha 2,6 linkages [178]. However, there are a couple of patents that mention using this protein for producing sialoglycans. This could limit patentability of this protein for use as a probe. The next most similar sialyltransferases are from *Pasteurella oralis* (94% coverage, 73.59% identity) and *Mergibacter septicus* (94% coverage, 49.87% identity). After that are two sialyltransferases confirmed to be specific for alpha 2,6 linkages from *Photobacterium leiognathi* strains JT-SHIZ-119 and JT-SHIZ-145[44, 45]. The wildtype form of both these sialyltransferases have previously been expressed and purified. Additionally, some kinetic data for both sialyltransferases with a selection of acceptor ligands exists in the literature. X-ray crystallography, ELISAs, and a glycan array will help to structurally and biochemically characterize the preferred ligands of either of these sialyltransferases. Moving forward, the sialyltransferase with either the highest affinity for, or highest selectivity for sTn will be selected.

Designing mutants to inhibit binding of the donor substrate CMP-sialic acid and the product CMP.

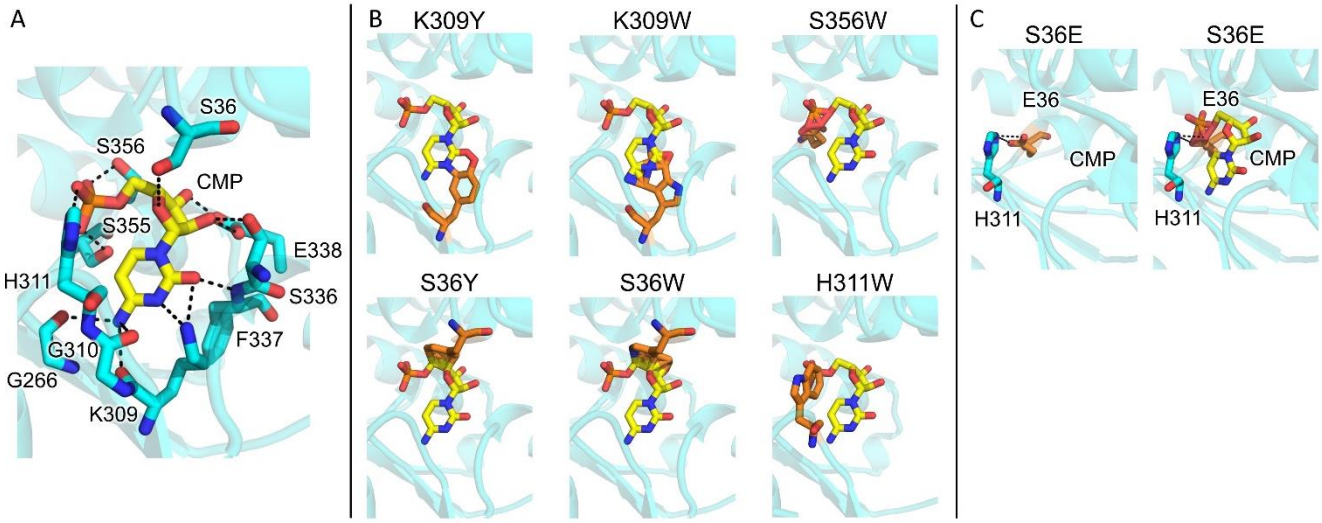

Figure viii-1 Steric and charge repulsion inhibition of donor substrate binding.

CMP binding to PMST. PMST is shown in teal and CMP is shown in yellow. Nitrogen and oxygen are colored in blue and red, respectively. Phosphorus is colored orange. Proposed mutations are colored orange. A) Residues that form hydrogen bonds with CMP are shown in sticks and hydrogen bonds are indicated by dashed black lines. B and C) PMST mutants are modeled using PDB 2IIB. B) mutants that will likely sterically inhibit CMP binding. C) A mutant that will likely inhibit CMP binding by charge repulsion and hydrogen bond with H311 to stabilize the closed conformational state of PMST.

To prepare PMST for development as an α2,6 sialoglycan probe, PMST binding to off target molecules needs to be eliminated or significantly reduced. This includes the donor substrate, CMP-sialic acid and the product, CMP. CMP-sialic acid binds deep in a positively charged pocket in the enzyme. The phosphate group is stabilized by interactions with H311, S355, and S356. S36, S336, and E338 form hydrogen bonds with the ribose ring while G266, K309, G310, and F337 form hydrogen bonds with the cytosine nucleobase (Fig 2). I designed several mutants that I predict will inhibit binding of CMP. The residues mentioned above point inwards towards the CMP binding pocket. Mutation of these residues to one with a bulkier side chain such as tryptophan or tyrosine should sterically inhibit CMP from binding. I have decided against using phenylalanine as a substitute residue as it is hydrophobic and may be less stable in the positively charged binding pocket.

Another mutation that could prevent CMP binding is S36E. The side chain of a glutamate in this position should charge repulse the phosphate group of CMP. Additionally, this mutation could stabilize the closed state of the sialyltransferase by hydrogen bonding with H311 (Figure viii-1). A binding technique will be needed to demonstrate loss of binding of these mutants to CMP and CMP-sialic acid. Unfortunately, an ELISA is not suitable for these experiments. To my knowledge biotinylated CMPs do not exist. Additionally, biotinylated CMP-sialic acid to $PMST^{D141N}$ with an ELISA was not observed. Another binding technique will be needed for these measurements. Other possibilities are NMR, SPR, and ITC. Currently, there are no NMR assignments for PMST or $PMST^{D141N}$. These would need to be made before binding of CMP can be measured. Given the sequence similarity of PMST to other bacterial sialyltransferases and the conservation of the substrate binding pocket, mutations that inhibit CMP and CMP-sialic acid binding to PMST will likely do the same in other bacterial sialyltransferases.

Confirm the presence and activity of bacterial α2,6 sialyltransferases without catalytic aspartates.

Many bacterial α2,6 sialyltransferases are in the GT80 family. Previous studies have shown that these sialyltransferases contain a conserved aspartate residue (*Pasteurella multocida* D141, *Photobacterium leiognathi* strain JT-ISH-224 D232)[47, 179] that when mutated to an alanine or asparagine eliminates catalytic activity. While the vast majority of available sequences contain the catalytic aspartate, a few do not. Some strains of *Pasteurella multocida* contain a tyrosine instead (accession numbers QMT71932, QMT73958, QMT76688, QMT78714, and QMT80740. It is unlikely that these sialyltransferases are catalytically active, however, GT80 sialyltransferases also contain a conserved aspartate one residue N-terminal to the catalytic aspartate. This residue usually faces away from the binding pocket into the center of the N-lobe. To see if this aspartate residue faces the binding pocket and therefore could possibly support catalytic activity, I used AlphaFold to model one of these sialyltransferases containing a tyrosine. The aspartate residues N-terminal of the catalytic aspartate position still face inward towards the core of the protein indicating they cannot be used for catalysis. These proteins could still serve a biological function as a binding protein. In human spermatid development, transcripts for an inactive GalNAc-T encoded by gene GALNTL5 have been detected. Mouse

models suggest that this protein is important in protein loading into the acrosome.[180] Currently, there are no known bacterial α2,6 sialoglycan binding proteins. These likely nonfunctional enzymes could serve another biological purpose in bacteria. These proteins could also serve as good starting scaffolds for α2,6 sialoglycan probe development. A sialyltransferase assay would need to be performed to confirm activity or lack thereof of these proteins. These sialyltransferases could also be knocked out of bacteria. Using microscopy, the cell shape and glycocalyx thickness could be measured and compared to wildtype to see if there are any obvious changes due to loss of these proteins.

Alternative non-bacterial sTn probe scaffold

As elaborated on in the introduction, mammalian and plant lectins have been characterized with the intention to identify a probe for sTn. It is logical that if sTn prevents human epithelial cancers from being recognized as an antigen, we likely produce a glycan binding protein that recognizes sTn. Many human Siglecs have been biochemically characterized and the binding selectivity of some has also been elucidated. A lot of attention has been paid to Siglec2, also known as CD22. A glycan array recently suggested that CD22 does not have a high binding strength for α2,6 linked sialoglycans including sTn. It has even more recently been shown that two human Siglecs, Siglec7 and Siglec15, seem to be specific for sTn. This was demonstrated with a cell-based glycan array.[23] The ELISA assay described above could be used to confirm that Siglec7 and 15 bind to sTn and elucidate its selectivity. While human Siglecs are not particularly stable proteins, stability could be engineered if ELISA results demonstrate Siglec7 or 15 is a promising scaffold for sTn probe development.

Competition binding studies of sTn probe and human Siglecs

Once a specific probe is developed, comparing the affinity of the probe for sTn to the affinity of human Siglecs for sTn will be important when considering therapeutic potential. Siglecs are primarily responsible for recognition of sTn on the surface of cancer cells. Siglec binding to sTn initiates a signaling pathway that inhibits an immune reaction. If an sTn probe is to be used in the context of an immunotherapy, the sTn probe must be able to compete for binding with human Siglecs. Competition binding studies of human Siglecs and an sTn probe will demonstrate the ability of the probe to compete for binding.

ix. APPENDIX A: GENOTYPE-PHENOTYPE CORRELATIONS IN INHERITED RETINAL DISORDERS: A CASE STUDY WITH PERIPHERIN 2

Prevalence and Clinical Characterization of Inherited Retinal Diseases

1 in 3,000 to 1 in 4,000 people are estimated to have an inherited retinal disease adding up to around 1.5 million people worldwide. These diseases are often progressive and many lead to legal blindness by age 40-60. They are described clinically first and foremost by the parts of the retina that degenerate. This can be limited to a geographical area of the retina such as the macula, which are called macular dystrophies (formally called juvenile macular degeneration). It is important to note that this is distinct from macular degeneration, which is not congenital. The degeneration could also be specific to the type of photoreceptor or the order in which either the cones or rods degenerate, leading to general characterizations of rod, cone, rod-cone, or cone-rod dystrophies.

Genotype-Phenotype Correlations

Inherited retinal diseases are notorious for their phenotypic heterogeneity. At the most basic level of clinical characterization, there are rod dystrophies and cone dystrophies. Research on gene and stem cell therapy is progressing quickly and patients with IRDs may well see treatment options in their lifetime. Understanding the disease progression and mechanisms by which mutations lead to retinal degeneration is critical to developing a treatment plan for patients. As new therapies are put into clinical trials, patients who have reached advanced stages of their disease will be enlisted. The reason for this is that these patients have already lost most of their vision. The risk of using new therapies that could still have detrimental effects to their vision farther outweighs the reward they could gain of preserving or improving what vision they have left. Younger patients are typically excluded from clinical trials for this reason, but some IRDs progress more quickly. These younger patients may be excluded from trials even though their disease may be progressing at a rapid rate, making their reward to risk ratio high as well. Understanding disease progression and protein biology also plays a big role in deciding if a stem cell or gene therapy approach is best for any specific patient. Genes that are larger such as ABCA4, a common pathogenic gene for Stargardt disease, are much more difficult to deliver with viral systems, making

gene therapy a less viable option. Additionally, if a patient has lost a majority of their photoreceptors, gene therapy will do little to restore their vision. Stem cell therapy could replace lost photoreceptors and make for a more effective treatment strategy in these cases. Gene therapy thus far has made greater clinical progress than stem cell therapy. Gene therapy used to supplement photoreceptor cells with the wildtype gene may be sufficient to treat patients with certain mutations, while others may additionally require removal or transcription inhibition of the mutated gene. Understanding of the underlying protein biology will inform these decisions to develop successful therapeutic options.

Within each of these distinct dystrophies there lies heterogeneity in the time of onset of disease, progression, and histology of the retina. To confirm clinical diagnoses and prepare patients for possible clinical trials, genetic sequencing is recommended for patients diagnosed with a retinal dystrophy. The National Institutes of Health (NIH) National Eye Institute (NEI) set priorities for better understanding the genetic landscape of IRDs [181]. Much of the progress in this area has been led by the Foundation Fighting Blindness, which provides financial aid and access to genetic counseling by IRD specialists and maintains a patient registry where afflicted patients with the help of their clinicians can log clinical and genetic information regarding their IRD [182, 183]. Today, extensive documentation of pathogenic mutations exists, but even this knowledge is not enough to predict disease progression and outcome. Multiple reports describe heterogeneity between family members with the same mutation [184]. The inability reliably predict disease progression lies in the knowledge gap of disease mechanism—understanding how a specific protein mutation leads to photoreceptor degeneration.

Here, I will focus on a particular gene, PRPH2, and the phenotypic variation clinically observed upon mutations. PRPH2 is representative of the phenotypic variability observed in inherited retinal diseases as at least five different phenotypes have been observed: vitelliform macular dystrophy, butterfly-shaped pattern dystrophy, retinitis punctate albscens, digenic retinitis pigmentosa, and general cone-rod dystrophy. PRPH2 also makes for a good case study gene simply because scientists have dedicated a lot of resources to studying the protein product, peripherin 2. The function of this protein is relevant not only to the field of retinal biology, but also to the fields of membrane proteins and membrane architecture, so resources from many different scientific fields have been

dedicated to understanding this protein. First, I present research investigating the structure and function of peripherin 2. Next, I will summarize each of the five phenotypes caused by mutations in the PRPH2 gene.

Peripherin 2

Peripherin 2, a member of the tetraspanin family

The role of peripherin 2 was first studied in mouse models before scientists had discovered the gene or protein product. The gene was initially called retinal degeneration slow (rds), after it was discovered to cause similar but slower retinal degeneration than previously discovered cGMP-phosphodiesterase mutants [185, 186]. Both homozygous and heterozygous mutant mice showed complete loss of photoreceptors by the age of twelve months as assessed by electroretinography [187]. Additional physiological characterization through immunohistochemistry, and light and electron microscopy, was carried out through the 1980s [187, 188, 189, 190, 191, 192, 193, 194, 195, 196]. It was not until 1989 that the mRNA for the rds gene, now PRPH2, gene was cloned [197]. The protein product of this gene was not realized until 1991 when two groups independently published within one month of each other. Dean Bok and his lab found that the mouse rds gene encoded for a 39 kDa glycoprotein [198, 199]. A month later, Robert Molday, who had been studying a bovine retinal protein, peripherin 2, and his lab identified that the rds gene product was the mice homologue of peripherin 2 [200].

Peripherin 2, also known as tetraspanin 2, is a member of the tetraspanin protein family. Expressed in many different cell types, tetraspanins carry out vital functions including cellular signaling, adhesion, and migration by shaping and organizing membranes. Tetraspanins are structurally highly homologous; they have N-terminal and C-terminal short intracellular loops, four transmembrane domains, and two extracellular loops. Peripherin 2 is localized to the disc rim of both rod and cone photoreceptor outer segments, where it induces membrane curvature of outer segments (OSs), making up 4% of the total bovine rod outer segment membrane protein [201, 202, 203]. Fusion protein studies showed that the C-terminus contains a sorting signal necessary for trafficking of peripherin 2. Residues 317-336 were sufficient for trafficking to discs in the outer segment, but a larger portion of the C-terminus, residues 307-346, is important for specifically targeting disk rims [204]. Later studies showed that a highly

conserved single valine residue within this region, V332, when mutated to alanine stops peripherin 2 from being trafficked to the outer segment [205]. This effect is incredible considering the high similarity between valine and alanine. To date, how this signal sequence contributes to proper trafficking is unknown. Sorting sequences in other retinal proteins including rhodopsin have been identified but are not thought to be related to the sequence found in peripherin 2.

Arguably one of the most important structural regions of tetraspanins is a cysteine rich region in one of their extracellular loops, EC2, that is responsible for many interactions between tetraspanins and other proteins. For peripherin 2 this region is specifically called D2 for intradiscal loop 2 [206]. 70% of pathogenic mutations in peripherin 2 are found within the D2 loop. The loop, misleadingly named, is partially structured. Consisting of approximately 150 residues, it contains three highly conserved alpha helices, spanning residues 121-145, 149-163, and 172-181 [207]. Peripherin 2 has thirteen cysteine residues in this loop, seven of which are highly conserved. Mutation of six of the seven cysteines, C165, C166, C213, C214, C222, and C250, prevent proper protein folding and protein aggregates form, indicating that these residues play an important role in protein folding. Evidence suggests that these cysteines form intramolecular disulfide binds that stabilize the D2 loop [208]. The D2 loop is glycosylated at Asn289. Mass spectrometry studies show that the glycan ranges from a pentose to a decamer and is primarily composed of hexose, manose, N-acetylheoxsamine, N-acetylglucosamine, and N-acetylneuraminic acid [209]. The second functionally important portion of peripherin 2 is the cytoplasmic C-terminal loop. Residues 312-328 are an inducible amphipathic alpha helix. Upon contact with the plasma membrane this portion of the protein folds into an alpha helix [210, 211]. Importantly, the alpha helix is amphipathic, meaning one face of the helix is charged and the other is not, giving it the ability to lie along the surface of the intradiscal surface of the disc membrane.

Complex Formation

The tetraspanin family is known to form organized local membrane environments known as tetraspanin webs. Tetraspanins interact with other tetraspanins and non-tetraspanin proteins through three loosely defined levels of interaction. These interaction levels are determined by the strength of the interaction, which has

traditionally been experimentally determined by stability of the complex in the presence of detergents. The first level of interaction is the strongest; these interactions are specific and resistant to strong detergents. The second level of interaction are typically interactions required for formation of higher order tetraspanin webs. While these interactions are weaker and only resistant to mild detergents, they still maintain a certain specificity and organization. The third level of interactions are only resistant to very mild detergents [212, 213]. These rules can generally be applied to peripherin 2, but a few key differences should be noted. First, the core oligomers of most other tetraspanin webs are dimers, while the core oligomer for peripherin 2 and rom1 is a tetramer. Second, while the second extracellular loop is highly conserved in sequence and three alpha-helical structure, the helices in peripherin 2 are suggested to be organized slightly different than most other family members [207]. Given that this region is flexible in nature, accurate structural comparison of this loop between peripherin 2 and other tetraspanins is difficult. Regardless, this alpha helical organization can result in slightly different dimer patterns and disulfide formation, which could explain point number one.

The first level of interactions is mediated by highly conserved cysteine residues within the second intradiscal loop [214]. While it has been proposed, evidence suggests that the C-terminal amphipathic helix does not affect dimer or tetramer formation [215]. Studies have shown that C150 is responsible for disulfide bond mediated homodimer formation. This has been confirmed with the C150S peripherin 2 mutant, which is not able to form disulfide bonded dimers but is still able to form homo and heterotetramers [208]. Peripherin 2 forms either homo or heterotetramers (with Rom1 as discussed below) after translation in the inner segment of photoreceptors. These tetramers are then trafficked to the outer segment where higher order oligomers form. It has been proposed that mutations that prevent disulfide bond peripherin 2 tetramer formation may keep peripherin 2 from being trafficked to the outer segment [216]. This indicates that oligomerization is not only important to the core nature of peripherin 2 in forming larger oligomer networks but is also important in ensuring that protein is shuttled from the endoplasmic reticulum of the inner segment where it is translated, to the outer segment.

Interactions with ROM1

ROM-1, or rod outer segment membrane protein-1, is another member of the tetraspanin protein family localized to photoreceptors. ROM1 and peripherin 2 have 35% sequence similarity and even greater structural similarity. More than this, both are localized to outer segments disc rims and form noncovalent heterotetramers [202, 217, 218]. Rom1 does not contain the sorting signal identified in peripherin 2, but heterotetramers of peripherin 2 and Rom1 are formed in the inner segment prior to trafficking. Given that peripherin 2 is expressed at levels two and a half times greater than Rom1, it is estimated that approximately 78% of Rom1 containing tetramers would contain at least one sorting signal [204]. Initial evidence for a peripherin 2/ROM-1 tetramer was shown through size exclusion chromatography and velocity sedimentation studies [201, 217]. Later studies of the tetramer surprisingly found that the tetramer was an a1b1/a2b2 complex rather than a dimer of homodimers as was predicted at the time [209]. The individual roles of peripherin 2 and Rom-1 in creating and maintaining outer segment disc curvature remains elusive. It has been known that peripherin 2 alone is necessary and sufficient for rod OS disc formation, but without rom-1 there is a decrease in photo response as well as slow photoreceptor death [219]. Unlike peripherin 2 homomers, peripherin 2/ROM1 heteromers associate through noncovalent interactions [214, 217]. Peripherin 2, but not Rom1 or peripherin 2/ROM1 complexes form larger oligomers [214]. Rom-1 maintains the ability to form tetramers without peripherin 2 [220].

Creating Membrane Curvature

As mentioned above, peripherin 2 is localized to the rim of disc outer segments. This localization already hints at peripherin 2's essential function in creating and maintaining disc morphology: creating membrane curvature. While all membranes have some amount of natural curvature, it is energetically unfavorable to have highly curved membranes because the negatively charged phospholipid head groups will repel each other when they come into close proximity. The energetics of this have been particularly well-studied in the case of soluble N-ethylmaleimide sensitive factor (NSF) attachment protein receptor (SNARE) proteins, which are responsible for pulling two lipid membranes into close contact for fusion. In photoreceptor discs, peripherin 2 extended oligomers form a web that curves the disc rim into a tubular shape with an average diameter of 35nm [221]. Using

a low-resolution 3D negative stain electron microscopy structure of the peripherin-Rom1 tetramer, Kevaney et al developed a model of the disc rim using measurements from electron micrographs of peripherin-Rom1 reconstituted lipid vesicles [209]. Experiments with peptides of the peripherin 2 C-terminal amphipathic helix showed an increase in membrane curvature as the ratio of peptide to lipid increased, suggesting that the C-terminus induces membrane curvature through a wedging mechanism similar to how many other membrane proteins induce membrane curvature [221, 222]. Interestingly, expression of peripherin 2 without the amphipathic alpha helix in the C terminus resulted in higher membrane curvature [215]. This evidence makes for a compelling argument that the amphipathic T-terminal helix does not create membrane curvature. Additionally, since the experiments are done in cell rather than in vitro, the environment is much more physiologically accurate suggesting that these observations are closer to what would be observed in vivo. The role of this amphipathic helix in creating membrane curvature remains in question as in vivo studies have found that the C-terminal amphipathic helix is sufficient for OS formation (discussed below). I hypothesize that without a protein that can form higher order oligomers the amphipathic helix may serve to curve the membrane, however, in the case of peripherin 2 specifically, oligomerization is responsible for creating membrane curvature. The amphipathic helix may hinder curvature or regulate rigidity of the protein network and membrane. Further studies aimed at investigating how the amphipathic helix of one peripherin 2 molecule interacts with other surrounding peripherin 2 molecules would further inform its functional role. Cryo electron tomography could be used to visualize in vitro disc formation and to compare wildtype peripherin 2 with peripherin 2Δαhelix.

Outer Segment Formation

OS formation and morphology is an important qualitative measure of retinal health as disc disorganization levels are positively correlated with photoreceptor degeneration [223]. A number of proteins are responsible for the morphogenesis of photoreceptor cells' outer segments. In rds knockout mice, discs fail to form showing that peripherin 2 is necessary for disc formation [189, 220]. Studies of a C150 mutant peripherin 2 showed small disc formation even with only 10% of wildtype levels [224]. This indicates that while peripherin 2 is necessary for OS formation, very little is needed for initiation. Interestingly, while peripherin 2 is required for disc formation, it is

not responsible for disc elongation, the next step in morphogenesis. Rhodopsin knockout mice exhibited disc formation, but the discs remained small and circular. Furthermore, heterozygous rds and rho mice showed that the relative abundances of peripherin 2 and rhodopsin determine disc and overall OS morphology [220]. Recently, peripherin 2 chimera knock-in studies provide evidence that the cytoplasmic C terminal loop of peripherin 2 may serve to induce membrane curvature [225]. Experiments with mice expressing a chimera consisting of ROM-1 with the C-terminal cytoplasmic loop of peripherin 2 exhibited OS formation. These OSs are spatially less organized and ERGs of these mice show decreased functionality in comparison to prph2 wildtype mice, but the fact that OSs form at all suggests that the C terminus of peripherin 2 plays a role in creating membrane curvature required for disc formation [226]. Peripherin 2 has also been suggested to play a role in disc formation by inhibiting ectosome release. In prph2 knockout mice, ectosome buildup around the connecting cilium are observed [189]. Furthermore, mice expressing rhodopsin fused with the C terminus of peripherin 2 are able to form an outer segment, although it is highly disorganized [227]. Together, this evidence suggests that while the C-terminus of peripherin is sufficient for OS formation, but an organized and fully functional outer segment requires peripherin 2 to form higher order oligomers to create membrane curvature and rhodopsin for elongation of the flattened portion of discs [225].

Peripherin 2 has also been thought to play a role in disc renewal by facilitating membrane fusion, required for disc shedding [228]. This hypothesis stems from the discovery of the amphipathic helix, similar to hypotheses regarding the role of this helix in creating membrane curvature. Purified bovine peripherin 2 increases rates of membrane fusion when reincorporated into lipid vesicles [229]. Peptide models of the C-terminus of peripherin 2 promotes membrane fusion and destabilization [210]. This peptide can also form a tetramer in vitro [230]. While this is unlikely to occur within the full-length protein in vivo, the possibility leaves this protein more structurally similar to other membrane fusion SNARE complexes. Furthermore, a fusion protein with a truncated form of the C-terminus was unable membrane fusion [210]. Two charged residues, E321 and K324, in the C-terminal cytoplasmic loop are critical for fusogenic functionality of peripherin 2. When the charges are neutralized

through mutation to either leucine or alanine, peripherin 2 is no longer able to mediate membrane fusion in lipid-mixing assays [211].

Genotype and Phenotype Characterization

Vitelliform Macular Dystrophy

Vitelliform macular dystrophy is an inherited retinal disease characterized by loss of photoreceptors in the macula. Fundus photos of patients with vitelliform macular dystrophy show a raised yellow area in the macula and is sometimes described as yolk-like. This buildup is waste from the retinal pigment epithelium including proteins, lipids, retinol, and derivatives [231]. Patients with vitelliform macular dystrophy can be subdivided into two categories: Best and adult onset. Best disease has a much earlier onset time than adult onset and is usually caused by mutations in a gene called BEST. Adult onset vitelliform macular dystrophy usually develops in a patient's middle-age years. Decreases in visual acuity through disease progression are usually slight. One mutation reported to cause adult onset vitelliform macular dystrophy had a frameshift mutation starting at residue 33 in peripherin 2 [232]. Another patient with a nonsense mutation, p.Tyr258Ter, has been reported [233].

Pattern Dystrophy

Broadly, pattern dystrophies are inherited retinal diseases that impact the macula. Compared to almost all other inherited retinal diseases, pattern dystrophies have a good prognosis. Typically, patients lose very little vision, and the disease progression is slow. Patients are usually diagnosed during routine eye visits without noticing any changes in their vision. Pattern dystrophies are characterized by the buildup of lipofuscin around the macula and can be more specifically labeled for the pattern that the lipofuscin forms. One patient diagnosed with pattern dystrophy was heterozygous for a nonsense mutation, p.Gln239Ter [234]. Patients with butterfly-shaped pattern dystrophy have three to five arms or wings of lipofuscin projecting out from the macula. This pattern is often bilateral, but the lipofuscin patterns are known to change over time and can vary between the two eyes. C213Y in peripherin 2 is known to cause butterfly shaped pattern dystrophy in patients and cone-rod dystrophies in mice [235]. It is surprising that this mutation does not result in a more severe form of retinal

degeneration as C213 is one of the six cysteine residues critical for proper protein folding. When mutated to serine in mice models the protein is able to form dimers but largely forms aggregates in vitro [208]. Low levels of photoreceptor degeneration in humans with this mutation suggest that enough peripherin 2 is able to properly fold to maintain disc stability. This is likely aided by chaperones in the inner segment. Peripherin 2 needs to fold properly to form dimers and tetramers, which is likely to be a requirement for trafficking to the outer segment. Recently, a case of pattern dystrophy thought to be caused by a null mutation in ROM1 has been reported [236]. While null mutations in rom1 in mice models are not reported to cause any phenotypes, pattern dystrophies are mild, and this mutation may present differently in humans. If this ROM1 mutation is indeed the pathogenic mutation, it will be the first reported ROM1 mutation to cause a monogenic inherited retinal disease.

Retinitis Punctata Albscens

Retinitis punctate albscens is an incredibly rare inherited retinal disease. This disease typically begins with child onset night blindness and progresses until the disease resembles retinitis pigmentosa. A unique characteristic that sets retinitis punctata albscens apart from retinitis pigmentosa is the presence of small white flecks, or punctate, observed upon fundus examination. Many cases of this inherited retinal disease are caused by mutations in the gene encoding for retinal binding protein 1, RLBP1, but cases have also been reported due to null mutations in PRPH2. It has been suggested that retinitis punctate albscens rather than retinitis pigmentosa is more directly comparable to rds mutant mice, but since retinitis pigmentosa is much more common studies typically reference it specifically [237].

Autosomal Dominant Retinitis Pigmentosa

Retinitis Pigmentosa (RP) is the most common retinal dystrophy making up for an estimated one in four thousand cases in the United States and Ireland. Retinitis pigmentosa is characterized by severely reduced rod function, bone spicule pigmentation, thin and pale nerves, and thinning of the peripheral outer nuclear layer. Only about half of patients diagnosed with retinitis pigmentosa have relatives with the disease. The prognosis for retinitis pigmentosa is poor. Many patients will meet the criteria for legal blindness by age sixty to seventy, which is less than 20/200 best corrected visual acuity or less than 20 degrees of peripheral vision. Many

pathogenic mutations in PRPH2 lead to autosomal dominant retinitis pigmentosa. As discussed above, rds heterozygous mice display a significant decrease in rod photoreceptor functionality and a decrease in disc organization. Some evidence suggests that the effect of peripherin 2 is dose dependent [223, 238]. Introducing viral introduction of wildtype rds in rds heterozygous mice rescues the disease phenotype [239, 240]. These data suggest that the abundance of peripherin 2 in the outer segment is critical for proper maintenance of disc morphology. Murine models of two mutants known to cause autosomal dominant retinitis pigmentosa, C214S and P210L, prevent peripherin 2 trafficking to the outer segment. Interestingly, when both peripherin 2 mutants form dimers with wildtype peripherin 2, the dimer is properly trafficked to the outer segment. If wildtype peripherin 2 can rescue trafficking of mutant peripherin 2, future therapeutics may simply need to genetically supplement wildtype peripherin 2 [241, 242].

Digenic Retinitis Pigmentosa

Some mutations in Prph2 have been found that are thought to be non-pathogenic except when in conjunction with mutations in other retinal proteins. These other retinal proteins are thought to be disease "modifiers" and are of clinical interest for two main reasons: 1) mutations in modifier proteins can lead to digenic disease and 2) these modifier proteins may play a role in phenotypic variability. Mutations in ABCA4, a gene encoding for an ATP-binding cassette transporter, and ROM1 are thought to cause or convert certain Prph2 mutations to digenic RP [243]. Many mutations in ABCA4 lead to Stargardt disease. A missense mutation predicted to encode for V2050L ABCA4 was identified in a family in conjunction with R172W PRPH2 and is thought to mildly worsen the observed phenotype [243]. There is not a lot of evidence for an ABCA4 modifier effect, and the evidence that does exists points to worsening a cone-rod dystrophy phenotype in contrast to the retinitis pigmentosa further discussed here. ROM1 has not been found to be pathogenic for any form of retinal degeneration to date. However, heterozygous mutations in conjunction with certain non-monopathogenic mutations in PRPH2, result in retinitis pigmentosa.

Specifically, three families heterozygous for null alleles in ROM1 and heterozygous for a missense mutation, Leu185Pro, in PRPH2 developed RP [244]. This mutation is in the second intradiscal loop and prevents

peripherin 2 from forming tetramers with other mutated peripherin 2 molecules, but Leu185Pro peripherin is able to form higher order oligomers with wildtype peripherin 2 [245, 246]. In the presence of ROM1, larger oligomers are still able to form, leading to stable discs. It is thought that in combination with a null allele in ROM1, higher order tetraspanin oligomers are not able to form leading to retinal degeneration [245]. Digenic RP can also be caused by a missense mutation in ROM1. Leu185Pro in PRPH2 in combination with Gly113Glu in ROM1 has been found to be pathogenic, although expression levels of Gly113Glu rom1 are markedly lower than wildtype suggesting that this mutation effectively acts as a null allele [246, 247]. Mice studies showed that mice heterozygous for a rom1 null mutation and mice heterozygous for Leu185Pro in prph2 each had a combined rom1 and prph2 transcript level above 60% of wildtype. Mice that are heterozygous for both rom1 null mutation and prph2 Leu185Pro mutation only about 50% transcript levels of wildtype, showing that 60% transcript levels may be a critical threshold for the quantity of these proteins needed to ensure disc stability [223]. This clinical and experimental evidence reinforces the importance of having a threshold amount of tetraspanin to ensure web formation and disc stability.

Additionally, Rom1 has been shown to modify the disease phenotype of the Y141C peripherin 2 mutation. Expressed with wildtype Rom1, Y141C peripherin 2 results in a cone-rod dystrophy similar to pattern dystrophies observed in patients, but when expressed without Rom1 ($rom1^{-/-}$) peripherin 2 displays a rod-dominant dystrophy similar to the phenotype of heterozygous rds mice and patients with retinitis pigmentosa. This is because Rom1-Y141C peripherin 2 dimers are trafficked together to the outer segment. Without wildtype Rom1, this particular mutant of peripherin 2 fails to traffic to the OS [248].

Cone-Rod Dystrophy

The other retinal degenerative phenotypes described above primarily affect rod photoreceptors, but cone-rod dystrophies due to pathogenic mutations in peripherin 2 have been described. This is particularly interesting as it suggests that certain mutations in peripherin 2 impact cone photoreceptors more than rod photoreceptors, suggesting that peripherin 2 has slightly different functionality in the two different photoreceptors. Heterozygous rds mice do not display cone function deficits in comparison to wildtype, but rod functionality is decreased [249]. A larger quantity of peripherin 2 may be required for maintaining disc morphology in rod

photoreceptors than for cone photoreceptors. This is justified by the fact that peripherin 2 is localized to both rims in rod discs but is only localized to the rim distal to the axoneme in cone outer segments. One animal model that has been developed to study cone and macular dystrophies caused by mutations in peripherin is a mice strain heterozygous for Y141C. This mutation has been identified as a pathogenic mutation in human cone dystrophies and the mice model recapitulates the disease phenotype and inheritance pattern well [250]. Both R172W and N244H have been identified as pathogenic peripherin 2 mutations that cause macular or cone-rod dystrophies. These residues lie within intradiscal loop 2. While changes in tryptic digestion of R172W peripherin 2 compared to wildtype suggest that this mutation causes structural changes, computational modeling of the loop mutant predicts that R172W does not significantly affect the structure of the loop [207, 249]. R172W rds heterozygous mice showed a rescue of rod function in comparison to heterozygous rds mice but showed a 41% decrease in cone function. Even homozygous R172W rds mice retained rod function [249]. N244H is predicted to disrupt one of the three highly conserved alpha helices [207].

Future Work

One important outstanding question in the field is the differences in the role of peripherin 2 in rod and cone photoreceptors. Mutations in peripherin 2 can cause both rod dominant and cone dominant retinal degeneration, but the mechanisms by which these mutations differentially affect the function of peripherin 2 in rods and cones is unknown. Research on the function of peripherin 2 in cone photoreceptors is complicated by the fact that only approximately four percent of photoreceptor cells are cones. Mice models homozygous for *nrl*$^{-/-}$ have been used to study a few mutations in peripherin 2. This gene knockout converts rod photoreceptors to cone like receptors making the experiments concerning the role of peripherin 2 in cone photoreceptors easier. Some research points towards the glycosylation in the D2 loop being important for cone function but not rod function. Mice with N229S mutations preventing glycosylation showed a 40% decrease in cone function and N229S/*nrl*$^{-/-}$ mice showed a 60% decrease in peripherin 2 and rom1 levels meeting the proposed critical protein levels for disc stability (although this percentage has been proposed for rods and is very likely different in cones) [251].

Further questions include mosaicity patterns in retinal tetraspanin webs. Strong evidence supports the hypothesis that a threshold quantity of peripherin 2 and Rom1 are required for disc stability and retinal homeostasis but understanding how relative differences in expression levels of wildtype peripherin 2 and Rom1 as well as mutants affect this mosaicity and resulting structural stability is still elusive. Tetramers are thought to be composed of an a1b1/a2b2 structure, but the ability of Rom1 to compensate for peripherin 2 in heterozygous mice and the ability of peripherin 2 to largely but not completely compensate for Rom1 knockout mice suggest that this mosaicity is flexible. Furthermore, higher order retina tetraspanin oligomers do not contain Rom1, this suggests that mosaicity may change from the lamellar proximal rim portion of discs to the lamellar distal end. Other proteins have also been suggested to bind to peripherin 2, increasing the complexity of this mosaicity.

In addition to further research on oligomerization, more experimentation is needed to answer questions regarding the possible role of peripherin 2 in membrane fusion events. Some studies have posed that peripherin 2 serves to suppress ectosome release suggesting that peripherin 2 inhibits membrane fusion. Other studies have suggested that peripherin 2 promotes membrane fusion by inducing membrane curvature. Many lipid mixing and membrane fusion assays and even molecular dynamics simulations have been used to study the mechanism by which SNARE proteins facilitate membrane fusion. These assays and methods could easily be adapted to answer the same questions regarding peripherin 2. While evidence that peripherin 2 facilitates membrane fusion is weak, peripherin 2 may play a significant role in supporting membrane fusion. Overcoming energetic barriers to bring membranes into closer contact is the core principle of how SNARE proteins work and is also the function of peripherin 2. Whether peripherin 2 suppresses or supports membrane fusion this could be an important function for determining overall outer segment structure. Knockout and mutant peripherin 2 mice models often have large whorl-like discs that are irregular and unstacked. Peripherin 2 may be responsible for how many and the size of discs that are created. On the other hand, peripherin 2 has also been thought to be important in keeping discs stacked together, acting more as an adhesion than an architectural protein. While this hypothesis is compelling in explaining whirl-like discs in mutant mice, little to no experimental evidence directly supports this claim.

While mice models have greatly informed the effects of peripherin 2 mutations, molecular analyses of mutations are sparse. Limitations in biochemical studies of peripherin 2 were initially limited by the inability to express and purify large amounts of protein. Traditionally, protein needed for studies of many retinal proteins, especially rhodopsin and G proteins, were obtained from tissue extraction of bovine retinas. These purifications are difficult (and smelly) and require large amounts of retinas. Purification of the beta and gamma subunits of the heterotrimer G protein from 200 bovine retinas yields approximately 5mg of protein, and this protein is expressed at much greater levels than peripherin 2. Furthermore, tetraspanins in general have proved difficult to structurally characterize due to difficulties in crystallizing the protein for x-ray crystallography analysis. To date, only one tetraspanin has been crystallized. Cryo electron microscopy is a promising technique for future structural analyses because the protein can remain in solution. To date, only a low-resolution structure of the peripherin 2/Rom1 heterotetramer exists at 18 angstroms. The heterotetramer complex is approximately 150kDa, which is not the smallest molecule that has been modeled with cryo EM but is definitely nearing the forefront of technique development. A recent trend in electron microscopy has been to use negate stain grids to screen for large structural differences between related samples. For example, screening antibody fragments binding to an antigen of interest across different animal sources and number of injections has informed antibody development and convergence in vivo. Recently, members of the Goldberg laboratory have screened different peripherin 2 and Rom1 oligomers using this approach. In this way, average particle length can be quantified to inform oligomerization and if reconstituted in larger lipid structures can be used to quantify membrane curvature induced. More than this, they were able to use cryo electron tomography to develop a 3D model of tubules created from reconstitution of heterotetramers [252]. These techniques present exciting opportunities to screen mutations known to be pathogenic in inherited retinal diseases to probe disc stability at a molecular level.

Largely, there also is a lack of understanding of how disc instability leads to degeneration of the retinal pigment epithelium. This is a critical gap in knowledge as mutations in many genes in addition to PRPH2 also lead to disc instability. This link between disorganized disks and breakdown of the retinal pigment epithelium could explain some of the phenotypic variability observed in patients. Computational systems biological

modeling could be used to model the "noise" in degradation and the role that stochasticity may ultimately play in phenotype. This phenotype variability observed in patients also likely stems in large from the clinician's discretion. Inherited retinal diseases are rare and there are few clinicians trained in differentiating and diagnosing them. Furthermore, patients could be diagnosed with a catch-all dystrophy that doesn't describe a patient's unique phenotypic features. For instance, pattern dystrophies could simply be diagnosed as a cone or macular dystrophy. Patients with retinitis punctate albscens could be diagnosed with retinitis pigmentosa or simply a rod-cone dystrophy. If more direct correlations between genetic mutations and phenotypes could be determined, clinicians would be able to more accurately diagnose patients and develop a treatment plan. The many databases containing mutations linked to inherited retinal diseases and clinical evaluations of patients will play a large role in narrowing this down as the data is analyzed over the next years. Overall, great progress has been made in understanding peripherin 2 and other proteins related to inherited retinal diseases. New therapies are undergoing preclinical and clinical trials and one has already been approved by the Food and Drug Administration.

x. APPENDIX B: ARRESTIN INTERACTIONS WITH SRC FAMILY KINASES

Background

Arrestin is known for its role in transducing signals from G protein coupled receptors (GPCRs) through two biochemical arms. The first role is to arrest signaling through interactions with clathrin that lead to receptor internalization. The second arm is to elicit specific cellular responses both dependently and independently of GPCR activation [253, 254]. Specifically, arrestin, acting as a scaffold, can facilitate activation of enzymes including Mitogen activated protein kinases (MAPKs), leucine zipper kinases, and Src family kinases (SFKs).

The Src family kinases (SFKs) are nonreceptor tyrosine kinases. In humans, there are nine members of the Src kinase family: Src, Lyn, Fyn, Fgr, Blk, Yes, Lck, Hck, and Frk. The SFKs have five different regions: the SH4 domain, the Unique domain, SH3 domain, SH2 domain, and the kinase domain. The SH4 and Unique domains are thought to be intrinsically disordered and are just recently starting to be characterized [255]. The

kinase domain phosphorylates tyrosine residues and the SH3 and SH2 domains regulate the kinase's activity. Structures of the SH3-SH2-KD domains of Hck and Src suggest that in the inactive state, the SH3 and SH2 domains do not block the active site of the kinase domain [256, 257]. Instead, they are thought to decrease catalytic activity by clamping the N and C lobe of the kinase domain, hindering relative movement of the N and C lobes hypothesized to be required for catalysis. Interactions between the SH2-KD linker region and the SH3 domain create a compact, assembled state. Proteins that increase the activity of the SFKs typically do so by binding the SH3 and SH2 domains and releasing the clamp. In this way, proteins that modulate SFK activity must have a higher affinity for the regulatory domains than the regulatory domains have for each other and the kinase domain. Early research into the arrestin-SFK interaction provides strong evidence that the binding interface is between the N lobe of arrestin and the SH3 domain and kinase domain of SFKs [258, 259]. Higher resolution structural information of the interaction between arrestin and SFKs will greatly contribute to our understanding of kinase activation and arrestin-mediated signaling.

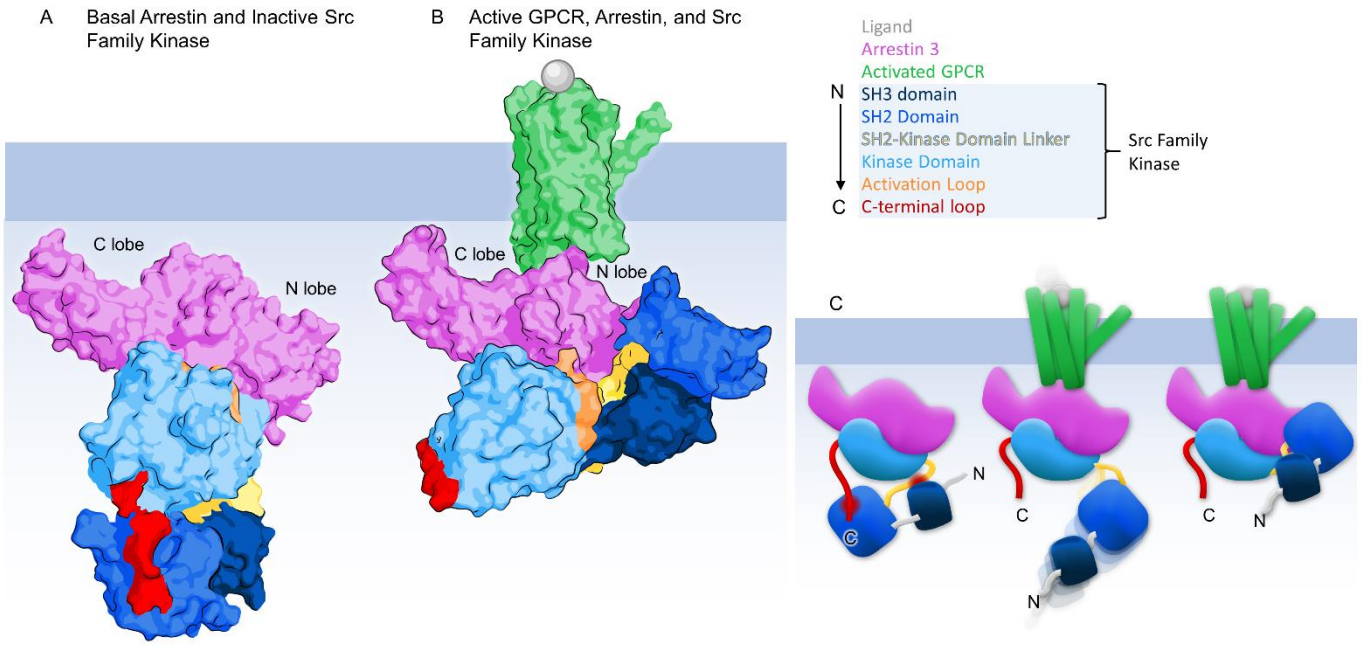

Figure x-1 Computational docking of Arrestin, an activated GPCR, and a Src family kinase.

The computational model in panels A) and B) were made by docking in Coot by Tina Iverson and molecular dynamics simulations by Rupesh Agarwal. C) A cartoon model of the interaction. Two regulatory interactions (red glow) keep the Src family kinases (SFKs) in an assembled autoinhibited state: interactions between the

SH2 domain and the C-terminal tail, and interactions between the SH2-Kinase domain linker and the SH3 domain. When disrupted, the kinase forms an elongated, disassembled, active state. Both the kinase domain and the SH3 domain are known to interact with arrestin [258, 259].

Proteins of Interest Rationale

Humans express four isoforms of arrestin, but only arrestins 2 and 3 are expressed ubiquitously. As this project focuses on mammalian signaling broadly, the visual specific arrestins, arrestin 1 and arrestin 4, will not be used here for analysis. The truncated forms of arrestin 2 and 3 (bovine arrestin 2 truncated residues: 1-382; bovine arrestin 3 truncated residues: 1-393) are tools the Iverson and Gurevich labs have readily used for some time to evaluate the affinity and structure of active conformations of arrestin. The C-terminal tail of arrestin stabilizes the inactive state of arrestin. Without these stabilizing interactions the truncated forms of arrestin more readily sample an active conformation making them more amenable to studying arrestin-mediated signaling complexes. A second way to study different conformations of arrestin is to study them in the presence of IP6, a non-receptor activator that leads to trimerization of arrestin 3 and larger filament-like oligomers of arrestin 2. Interestingly, all protomers in the arrestin 3 trimer maintain a receptor-bound conformation, while arrestin 2 retains a basal conformation in IP6 induced oligomerization.

As for SFKs tools, I have successfully expressed five of the eight SFK kinase domains (Blk, Fgr, Hck, Yes, and Fyn) and purified three of them (Yes, Blk, and Fyn). I verified the protein identity by gel electrophoresis, western blot, and for HCK by mass spectrometry. I am now able to purify the Blk, Yes, and Fyn kinase domains to ~95% purity with a three-step purification: 1) Ni purification, 2) Anion exchange chromatography, and 3) gel filtration. 8L of culture yields ~20 mg of protein for Yes and BLK kinase domains and yields ~40 mg for Fyn kinase domain. These three kinases are more than sufficient to study the interaction between the kinase domains and arrestin as the kinase domains are highly homologous (≥ 65% sequence identity). Additionally, I have expressed and purified a three domain (3D) version of Fyn that contains the SH3 domain, SH2 domain, and kinase domain. The C-terminal tyrosine that regulates kinase activity is not present in this construct.

xi. APPENDIX C: $G_{\beta\gamma}$ INTERACTION WITH THE NEURONAL SNARE COMPLEX

Introduction

The transport of membrane encased biological cargo into, through, and out of a cell, referred to as membrane trafficking, is a vital cellular process that ensures proper localization of proteins and communication between cells. A superfamily of proteins, the soluble NSF factor associated attachment protein receptor (SNARE) proteins, are the core membrane fusion machinery. SNAREs mediate a spatially and functionally diverse set of trafficking events such as Golgi and endosomal trafficking, phagocytosis, and secretion. One protein from each of the four SNARE subclasses (Q_a, Q_b, Q_c, and R) zipper together to form a highly thermostable four helix bundle that provides the energy necessary for membrane fusion[260, 261, 262, 263]. Canonically, a calcium sensing protein, synaptotagmin, binds the assembled SNARE complex to promote rapid vesicle fusion. New studies demonstrated that Gβγ, a dimer of two of the subunits making up the heterotrimeric G protein, facilitates inhibition of secretion events by competing with synaptotagmin to bind the SNARE complex [264, 265, 266, 267, 268, 269]. This regulation is important in maintaining homeostatic intercellular communication. Animal models lacking the Gβγ secretory inhibition mechanism present metabolic disease and impaired motor coordination among other physiological and behavioral phenotypes [270, 271].

Despite the clear physiological importance of the Gβγ—SNARE interaction, the interaction remains biophysically uncharacterized due to experimental challenges. The interaction between Gβγ and the ternary SNARE complex is transient and is therefore difficult to experimentally control. Additionally, minimal sampling of Gβγ dimers and SNARE complexes limits knowledge of the specificity of this interaction. Taking protein homology into account, the complete theoretical human proteome of this interaction consists of 4 β subunits, 12 γ subunits, and 38 human SNARE proteins. To date, only $G\beta_1\gamma_1$ and $G\beta_1\gamma_2$ and the three proteins that make up the ternary SNARE complex (SNAP25, syntaxin1, and synaptobrevin2) have been investigated in the context of the Gβγ—SNARE interaction. The affinity of Gβγ for the SNARE complex depends on the specific β and γ pairing. Relative affinities for specific isoforms of the Gβγ dimer may appreciably affect the dynamics of this interaction in a cell, a phenomenon similar to signal bias. Furthermore, we do not know

whether the Gβγ—SNARE interaction regulates other membrane fusion events such as those occurring between vesicles and other organelles including the endoplasmic reticulum and the Golgi apparatus. These trafficking events are similarly mediated by SNAREs homologous to those found in the ternary SNARE complex. I hypothesize that the Gβγ—SNARE interaction is not limited to regulating secretion events, but rather plays a more global role in regulating membrane trafficking events. This interaction may mediate Gβγ regulation of other cellular processes such as endocytosis, intracellular protein trafficking, and phagocytosis.

Experimental analysis of both Gβγ and SNARE isoform bias will elucidate the molecular basis for the interaction and the breadth of functional roles it may play in regulating membrane trafficking. Given the large theoretical proteome of this interaction, exhaustive screening of specific protein-protein interactions would be economically expensive.

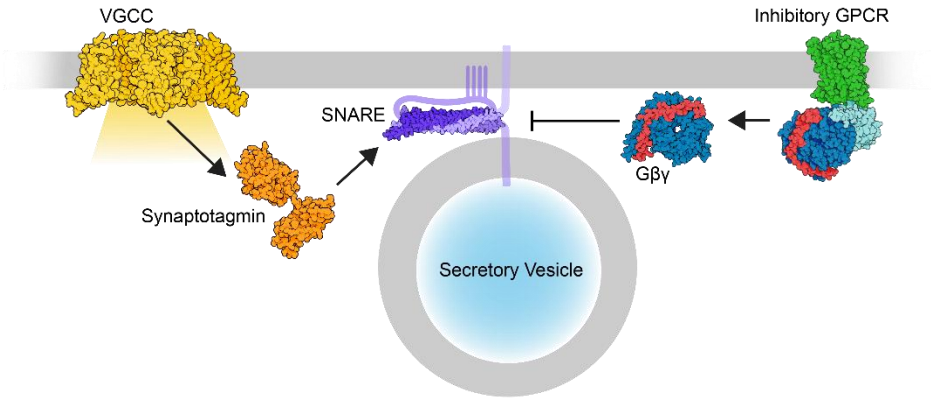

Figure xi-1 Gβγ inhibits calcium mediated neurotransmitter release.

Upon activation of voltage gated calcium channels (VGCCs)[272, 273], Ca^{2+} flows into the presynaptic neuron and binds and activates synaptotagmin [264, 266, 274]. Synaptotagmin then binds the ternary SNARE complex and promotes rapid and synchronous membrane fusion and neurotransmitter release[265, 274, 275, 276]. $G_{i/o}$-coupled Gβγ dimers, upon activation of an inhibitory GPCR and Gβγ dissociation from the heterotrimer—GPCR complex, will compete with synaptotagmin for binding to the SNARE complex and inhibit neurotransmitter release[269, 277, 278, 279, 280, 281, 282, 283, 284].

Structural models of the Gβγ—SNARE interaction and isoform specificity screening will aid in developing drugs that inhibit this interaction with high potency and little to no off-target effects.

Individually, both SNARE proteins and Gβγ dimers fill many functional roles in various subcellular compartments. Gβγ is a signaling effector in numerous pathways that regulate vital cell functions including chemotaxis, Golgi trafficking, and nuclear signaling[285]. SNAREs collectively function to facilitate fusion of membrane-bound compartments; however, these fusion events are diverse and include Golgi and endosomal trafficking, phagocytosis, lysosome digestion, and secretion. The Gβγ—SNARE interaction is novel and has only been studied in the context of secretory events involving the SNARE protein SNAP25. However, SNARE proteins and Gβγ dimers have significant intrafamilial homology and subcellular colocalization. These biochemical characteristics raise questions as to whether the Gβγ—SNARE interaction is specific to the ternary SNARE complex or if this interaction occurs between other non-secretory SNARE proteins as well. The latter indicates that the Gβγ—SNARE interaction plays a more global regulatory role in cell trafficking. Without Gβγ inhibition of secretion events, physiological dysfunction including ataxia and metabolic disease present in mice models [271]. Developing small molecules to modulate the Gβγ—SNARE interaction would provide an invaluable tool for future *in vivo* studies of the interaction as well as a potential therapeutic for clinical use [271, 286]. The aforementioned questions regarding the molecular and functional specificity of this interaction pose significant gaps in knowledge preventing accurate manipulation of this interaction for clinical and basic science research. Small molecule and mutational targeting of the Gβγ—SNARE interaction without understanding related cell processes could result in off-target effects making such studies invalid. Structural models will make structure-based drug design possible. Additionally, information gleaned regarding isoform bias would provide knowledge necessary for tight experimental and clinical control of this interaction.

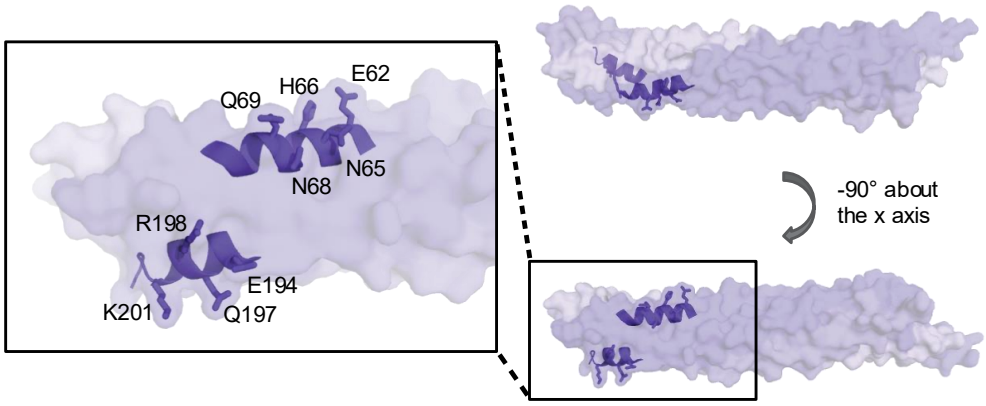

Figure xi-2 SNAP25 peptides with crosslinking unnatural amino acid incorporation.

SNARE peptides with a single unnatural amino acid, p-benzoyl-L-phenylalanine (bzf), incorporated at surface exposed positions will be used to map the specific interactions between $G_{\beta\gamma}$ and the SNARE complex. Peptides are shown in cartoon and native residues, which will have bzf incorporation, are shown in sticks and labeled.

I developed a FLAG tagged SNARE construct and successfully purified it via the FLAG affinity resin. The SNARE complex is more than 90% pure based on a gel, so I did not use any further purification steps.

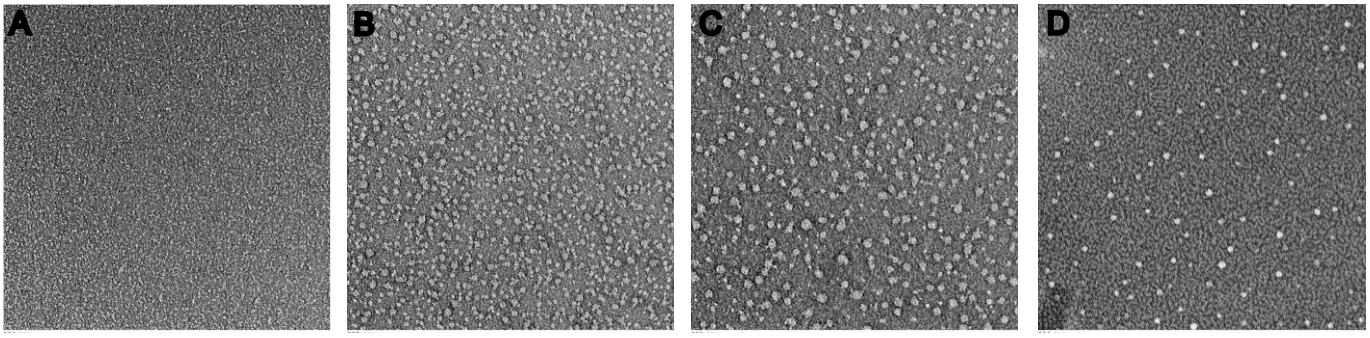

Figure xi-3 Electron Micrographs of $G_{\beta\gamma}$ and SNARE complex.

A) SNARE complex at 25 ug/mL and 28,000X magnification. B) and C) $G_{\beta\gamma}$ at 10 ug/mL at 36,000X and 44,000X magnification respectively. D) SNARE + $G_{\beta\gamma}$ at 0.35 ug/mL at 36,000X magnification.

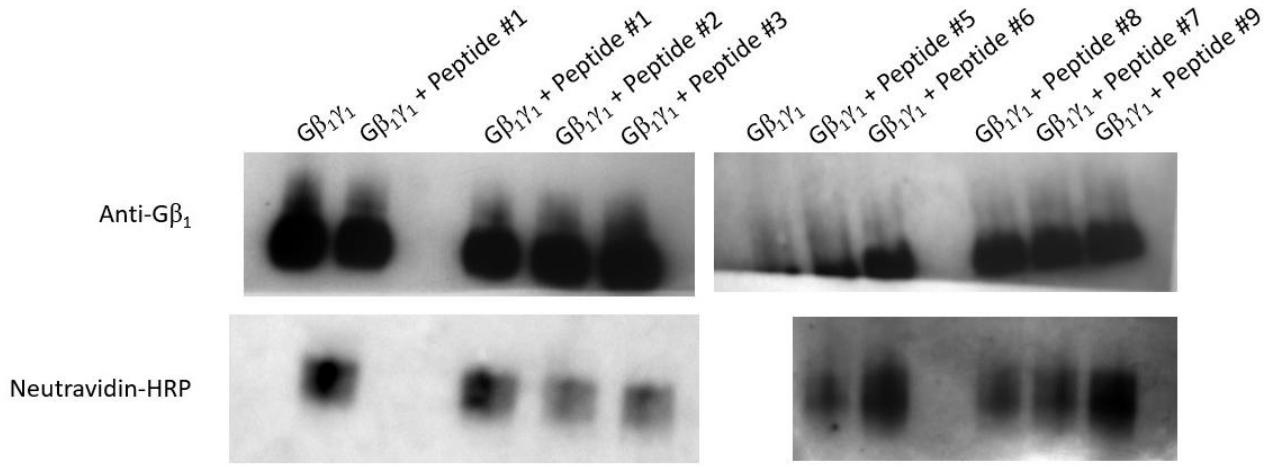

Figure xi-4 Site specific crosslinking of SNAP25 peptides to $G\beta_1\gamma_1$.

Western blots using an anti $G\beta_1\gamma_1$ antibody (upper) and neutravidin HRP (lower). SNAP25 peptides were labeled with a biotin tag. Signal appears at the molecular weight of $G\beta_1\gamma_1$ for all samples incubated with a SNAP25 peptide using neutravidin-HRP. This indicates that UV crosslinking of the peptides to $G\beta_1\gamma_1$ was successful for each SNAP25 peptide.

1. Varki A, *et al.* Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology* **25**, 1323-1324 (2015).

2. Neelamegham S, *et al.* Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology* **29**, 620-624 (2019).

3. Srivastava S, *et al.* Development and applications of sialoglycan-recognizing probes (SGRPs) with defined specificities: exploring the dynamic mammalian sialoglycome. *Glycobiology*, (2022).

4. Barnard KN, *et al.* Modified Sialic Acids on Mucus and Erythrocytes Inhibit Influenza A Virus Hemagglutinin and Neuraminidase Functions. *J Virol* **94**, (2020).

5. Soares CO, Grosso AS, Ereño-Orbea J, Coelho H, Marcelo F. Molecular Recognition Insights of Sialic Acid Glycans by Distinct Receptors Unveiled by NMR and Molecular Modeling. *Front Mol Biosci* **8**, 727847 (2021).

6. Jandus C, Simon HU, von Gunten S. Targeting siglecs--a novel pharmacological strategy for immuno- and glycotherapy. *Biochem Pharmacol* **82**, 323-332 (2011).

7. Lowe JB, Marth JD. A genetic approach to Mammalian glycan function. *Annu Rev Biochem* **72**, 643-691 (2003).

8. Becker JL, Tran DT, Tabak LA. Members of the GalNAc-T family of enzymes utilize distinct Golgi localization mechanisms. *Glycobiology* **28**, 841-848 (2018).

9. Livingston BD, Paulson JC. Polymerase chain reaction cloning of a developmentally regulated member of the sialyltransferase gene family. *J Biol Chem* **268**, 11504-11507 (1993).

10. Geremia RA, Harduin-Lepers A, Delannoy P. Identification of two novel conserved amino acid residues in eukaryotic sialyltransferases: implications for their mechanism of action. *Glycobiology* **7**, v-vii (1997).

11. Jeanneau C, *et al.* Structure-function analysis of the human sialyltransferase ST3Gal I: role of n-glycosylation and a novel conserved sialylmotif. *J Biol Chem* **279**, 13461-13468 (2004).

12. Drickamer K. A conserved disulphide bond in sialyltransferases. *Glycobiology* **3**, 2-3 (1993).

13. Julien S, *et al.* Sialyl-Tn vaccine induces antibody-mediated tumour protection in a relevant murine model. *Br J Cancer* **100**, 1746-1754 (2009).

14. Román-Carrasco P, Hemmer W, Cabezas-Cruz A, Hodžić A, de la Fuente J, Swoboda I. The α-Gal Syndrome and Potential Mechanisms. *Front Allergy* **2**, 783279 (2021).

15. Mita Y, Aoyagi Y, Suda T, Asakura H. Plasma fucosyltransferase activity in patients with hepatocellular carcinoma, with special reference to correlation with fucosylated species of alpha-fetoprotein. *J Hepatol* **32**, 946-954 (2000).

16. Munkley J. The Role of Sialyl-Tn in Cancer. *Int J Mol Sci* **17**, 275 (2016).

17. Varki A, Kannagi R, Toole B, Stanley P. Glycosylation Changes in Cancer. In: *Essentials of Glycobiology* (eds Varki A, *et al.*). Cold Spring Harbor Laboratory Press

18. Silva ML, Gutiérrez E, Rodríguez JA, Gomes C, David L. Construction and validation of a Sambucus nigra biosensor for cancer-associated STn antigen. *Biosens Bioelectron* **57**, 254-261 (2014).

19. Echeverri D, Orozco J. Glycan-Based Electrochemical Biosensors: Promising Tools for the Detection of Infectious Diseases and Cancer Biomarkers. *Molecules* **27**, (2022).

20. Kang JG, Ko JH, Kim YS. Application of cancer-associated glycoforms and glycan-binding probes to an in vitro diagnostic multivariate index assay for precise diagnoses of cancer. *Proteomics* **16**, 3062-3072 (2016).

21. Onitsuka K, *et al.* Prognostic significance of UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase-3 (GalNAc-T3) expression in patients with gastric carcinoma. *Cancer Sci* **94**, 32-36 (2003).

22. Julien S, Videira PA, Delannoy P. Sialyl-tn in cancer: (how) did we miss the target? *Biomolecules* **2**, 435-466 (2012).

23. Büll C, *et al.* Probing the binding specificities of human Siglecs by cell-based glycan arrays. *Proc Natl Acad Sci U S A* **118**, (2021).

24. Springer GF, Desai PR, Banatwala I. Blood group MN specific substances and precursors in normal and malignant human breast tissues. *Naturwissenschaften* **61**, 457-458 (1974).

25. Miles D, *et al.* Phase III multicenter clinical trial of the sialyl-TN (STn)-keyhole limpet hemocyanin (KLH) vaccine for metastatic breast cancer. *Oncologist* **16**, 1092-1100 (2011).

26. Holmberg LA, Sandmaier BM. Vaccination with Theratope (STn-KLH) as treatment for breast cancer. *Expert Rev Vaccines* **3**, 655-663 (2004).

27. MacLean GD, Reddish MA, Koganty RR, Longenecker BM. Antibodies against mucin-associated sialyl-Tn epitopes correlate with survival of metastatic adenocarcinoma patients undergoing active specific immunotherapy with synthetic STn vaccine. *J Immunother Emphasis Tumor Immunol* **19**, 59-68 (1996).

28. Nuti M, Teramoto YA, Mariani-Costantini R, Hand PH, Colcher D, Schlom J. A monoclonal antibody (B72.3) defines patterns of distribution of a novel tumor-associated antigen in human mammary carcinoma cell populations. *Int J Cancer* **29**, 539-545 (1982).

29. Johnson VG, Schlom J, Paterson AJ, Bennett J, Magnani JL, Colcher D. Analysis of a human tumor-associated glycoprotein (TAG-72) identified by monoclonal antibody B72.3. *Cancer Res* **46**, 850-857 (1986).

30. Muraro R, *et al.* Generation and characterization of B72.3 second generation monoclonal antibodies reactive with the tumor-associated glycoprotein 72 antigen. *Cancer Res* **48**, 4588-4596 (1988).

31. Kjeldsen T, Clausen H, Hirohashi S, Ogawa T, Iijima H, Hakomori S. Preparation and characterization of monoclonal antibodies directed to the tumor-associated O-linked sialosyl-2----6 alpha-N-acetylgalactosaminyl (sialosyl-Tn) epitope. *Cancer Res* **48**, 2214-2220 (1988).

32. Colcher D, Hand PH, Nuti M, Schlom J. A spectrum of monoclonal antibodies reactive with human mammary tumor cells. *Proc Natl Acad Sci U S A* **78**, 3199-3203 (1981).

33. Springer GF, Desai PR, Robinson MK, Tegtmeyer H, Scanlon EF. The fundamental and diagnostic role of T and Tn antigens in breast carcinoma at the earliest histologic stage and throughout. *Prog Clin Biol Res* **204**, 47-70 (1986).

34. Kurosaka A, *et al.* Mucin-carbohydrate directed monoclonal antibody. *FEBS Lett* **215**, 137-139 (1987).

35. Kurosaka A, *et al.* A monoclonal antibody that recognizes a cluster of a disaccharide, NeuAc alpha(2----6)GalNAc, in mucin-type glycoproteins. *J Biol Chem* **263**, 8724-8726 (1988).

36. Nozawa S, *et al.* Tumor-associated mucin-type glycoprotein (CA54/61) defined by two monoclonal antibodies (MA54 and MA61) in ovarian cancers. *Cancer Res* **49**, 493-498 (1989).

37. An Y, *et al.* A novel anti-sTn monoclonal antibody 3P9 Inhibits human xenografted colorectal carcinomas. *J Immunother* **36**, 20-28 (2013).

38. Loureiro LR, *et al.* Novel monoclonal antibody L2A5 specifically targeting sialyl-Tn and short glycans terminated by alpha-2-6 sialic acids. *Sci Rep* **8**, 12196 (2018).

39. Gaidzik N, *et al.* Synthetic antitumor vaccines containing MUC1 glycopeptides with two immunodominant domains-induction of a strong immune response against breast tumor tissues. *Angew Chem Int Ed Engl* **50**, 9977-9981 (2011).

40. Kaiser A, *et al.* A synthetic vaccine consisting of a tumor-associated sialyl-T(N)-MUC1 tandem-repeat glycopeptide and tetanus toxoid: induction of a strong and highly selective immune response. *Angew Chem Int Ed Engl* **48**, 7551-7555 (2009).

41. Liu CC, Ye XS. Carbohydrate-based cancer vaccines: target cancer with sugar bullets. *Glycoconj J* **29**, 259-271 (2012).

42. Ragupathi G, *et al.* Vaccines prepared with sialyl-Tn and sialyl-Tn trimers using the 4-(4-maleimidomethyl)cyclohexane-1-carboxyl hydrazide linker group result in optimal antibody titers against ovine submaxillary mucin and sialyl-Tn-positive tumor cells. *Cancer Immunol Immunother* **48**, 1-8 (1999).

43. Yabe R, Suzuki R, Kuno A, Fujimoto Z, Jigami Y, Hirabayashi J. Tailoring a novel sialic acid-binding lectin from a ricin-B chain-like galactose-binding protein by natural evolution-mimicry. *J Biochem* **141**, 389-399 (2007).

44. Yamamoto T, *et al.* A beta-galactoside alpha2,6-sialyltransferase produced by a marine bacterium, Photobacterium leiognathi JT-SHIZ-145, is active at pH 8. *Glycobiology* **17**, 1167-1174 (2007).

45. Mine T, *et al.* An alpha2,6-sialyltransferase cloned from Photobacterium leiognathi strain JT-SHIZ-119 shows both sialyltransferase and neuraminidase activity. *Glycobiology* **20**, 158-165 (2010).

46. Ni L, Sun M, Yu H, Chokhawala H, Chen X, Fisher AJ. Cytidine 5'-monophosphate (CMP)-induced structural changes in a multifunctional sialyltransferase from Pasteurella multocida. *Biochemistry* **45**, 2139-2148 (2006).

47. Kakuta Y, *et al.* Crystal structure of Vibrionaceae Photobacterium sp. JT-ISH-224 alpha2,6-sialyltransferase in a ternary complex with donor product CMP and acceptor substrate lactose: catalytic mechanism and substrate recognition. *Glycobiology* **18**, 66-73 (2008).

48. Ramboarina S, *et al.* Structural insights into serine-rich fimbriae from Gram-positive bacteria. *J Biol Chem* **285**, 32446-32457 (2010).

49. Loukachevitch LV, *et al.* Structures of the Streptococcus sanguinis SrpA Binding Region with Human Sialoglycans Suggest Features of the Physiological Ligand. *Biochemistry* **55**, 5927-5937 (2016).

50. Bensing BA, *et al.* Selectivity and engineering of the sialoglycan-binding spectrum in Siglec-like adhesins.). Cold Spring Harbor Laboratory (2019).

51. Stubbs HE, *et al.* Tandem sialoglycan-binding modules in a Streptococcus sanguinis serine-rich repeat adhesin create target dependent avidity effects. *J Biol Chem* **295**, 14737-14749 (2020).

52. Pyburn TM, *et al.* A structural model for binding of the serine-rich repeat adhesin GspB to host carbohydrate receptors. *PLoS Pathog* **7**, e1002112 (2011).

53. Zhang S, *et al.* Immune sera and monoclonal antibodies define two configurations for the sialyl Tn tumor antigen. *Cancer Res* **55**, 3364-3368 (1995).

54. Ronis A, *et al.* Streptococcus oralis subsp. dentisani Produces Monolateral Serine-Rich Repeat Protein Fibrils, One of Which Contributes to Saliva Binding via Sialic Acid. *Infect Immun* **87**, e00406-00419 (2019).

55. Bensing BA, *et al.* Structural Basis for Sialoglycan Binding by the Streptococcus sanguinis SrpA Adhesin. *J Biol Chem* **291**, 7230-7240 (2016).

56. Yu H, *et al.* Enzymatic and Chemoenzymatic Syntheses of Disialyl Glycans and Their Necrotizing Enterocolitis Preventing Effects. *J Org Chem* **82**, 13152-13160 (2017).

57. Yu H, Yu H, Karpel R, Chen X. Chemoenzymatic synthesis of CMP–sialic acid derivatives by a one-pot two-enzyme system: comparison of substrate flexibility of three microbial CMP–sialic acid synthetases. *Bioorganic & Medicinal Chemistry* **12**, 6427-6435 (2004).

58. Sugiarto G, *et al.* A sialyltransferase mutant with decreased donor hydrolysis and reduced sialidase activities for directly sialylating LewisX. *ACS Chem Biol* **7**, 1232-1240 (2012).

59. Lau K, *et al.* Sequential two-step multienzyme synthesis of tumor-associated sialyl T-antigens and derivatives. *Organic & biomolecular chemistry* **9**, 2784-2789 (2011).

60. Otwinowski Z, Minor W. [20] Processing of X-ray diffraction data collected in oscillation mode. In: *Methods in Enzymology*). Elsevier (1997).

61. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).

62. Liebschner D, *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**, 861-877 (2019).

63. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010).

64. Vařeková RS, *et al.* MotiveValidator: interactive web-based validation of ligand and residue structure in biomolecular complexes. *Nucleic Acids Res* **42**, W227-W233 (2014).

65. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).

66. Sievers F, *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).

67. Laskowski RA. PDBsum new things. *Nucleic Acids Res* **37**, D355-D359 (2009).

68. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *"Protein Engineering, Design and Selection"* **8**, 127-134 (1995).

69. Bensing BA, *et al.* Novel aspects of sialoglycan recognition by the Siglec-like domains of streptococcal SRR glycoproteins. *Glycobiology* **26**, 1222-1234 (2016).

70. Bensing BA, Li Q, Park D, Lebrilla CB, Sullam PM. Streptococcal Siglec-like adhesins recognize different subsets of human plasma glycoproteins: implications for infective endocarditis. *Glycobiology* **28**, 601-611 (2018).

71. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).

72. Kearse M, *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).

73. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105 (2005).

74. Ronquist F, *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* **61**, 539-542 (2012).

75. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).

76. Adams PD, *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica Section D, Biological crystallography* **66**, 213-221 (2010).

77. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta crystallographica Section D, Biological crystallography* **60**, 2126-2132 (2004).

78. Molecular Operating Environment (MOE)). 2016 edn. Chemical Computing Group Inc (2013.08).

79. Jorgensen W, Chandrasekhar J, Madura J, Impey R, Klein M. Comparison of simple potential functions for simulating liquid wate. *J Chem Physics* **79**, 926-935 (1983).

80. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696-3713 (2015).

81. Arfken G, Weber H, Harris F. *Mathematical Methods for Physicists - A comprehensive guide*, 7th edn. Academic Press (2012).

82. Case DA, *et al.* The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668-1688 (2005).

83. Ryckaert J, Ciccotti G, Berendsen H. Numerical integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes *J Computational Phys* **23**, 327-341 (1977).

84. Roe DR, Cheatham TE, 3rd. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **9**, 3084-3095 (2013).

85. Potterton E, Briggs P, Turkenburg M, Dodson E. A graphical user interface to the CCP4 program suite. *Acta Crystallogr D Biol Crystallogr* **59**, 1131-1137 (2003).

86. Ramboarina S, *et al.* Structural insights into serine-rich fimbriae from Gram-positive bacteria. *The Journal of biological chemistry* **285**, 32446-32457 (2010).

87. Deng L, *et al.* Oral streptococci utilize a Siglec-like domain of serine-rich repeat adhesins to preferentially target platelet sialoglycans in human blood. *PLoS pathogens* **10**, e1004540-e1004540 (2014).

88. Zheng W, Tan MF, Old LA, Paterson IC, Jakubovics NS, Choo SW. Distinct Biological Potential of Streptococcus gordonii and Streptococcus sanguinis Revealed by Comparative Genome Analysis. *Scientific reports* **7**, 2949-2949 (2017).

89. Denapaite D, *et al.* The genome of Streptococcus mitis B6--what is a commensal? *PloS one* **5**, e9426-e9426 (2010).

90. Xiong YQ, Bensing BA, Bayer AS, Chambers HF, Sullam PM. Role of the serine-rich surface glycoprotein GspB of Streptococcus gordonii in the pathogenesis of infective endocarditis. *Microb Pathog* **45**, 297-301 (2008).

91. Seo HS, Xiong YQ, Sullam PM. Role of the serine-rich surface glycoprotein Srr1 of Streptococcus agalactiae in the pathogenesis of infective endocarditis. *PloS one* **8**, e64204-e64204 (2013).

92. van Sorge NM, Quach D, Gurney MA, Sullam PM, Nizet V, Doran KS. The group B streptococcal serine-rich repeat 1 glycoprotein mediates penetration of the blood-brain barrier. *J Infect Dis* **199**, 1479-1487 (2009).

93. Takahashi Y, Takashima E, Shimazu K, Yagishita H, Aoba T, Konishi K. Contribution of sialic acid-binding adhesin to pathogenesis of experimental endocarditis caused by Streptococcus gordonii DL1. *Infect Immun* **74**, 740-743 (2006).

94. Shivshankar P, Sanchez C, Rose LF, Orihuela CJ. The Streptococcus pneumoniae adhesin PsrP binds to Keratin 10 on lung cells. *Molecular microbiology* **73**, 663-679 (2009).

95. Mistou M-Y, Dramsi S, Brega S, Poyart C, Trieu-Cuot P. Molecular dissection of the secA2 locus of group B Streptococcus reveals that glycosylation of the Srr1 LPXTG protein is required for full virulence. *J Bacteriol* **191**, 4195-4206 (2009).

96. Singh AK, Woodiga SA, Grau MA, King SJ. Streptococcus oralis Neuraminidase Modulates Adherence to Multiple Carbohydrates on Platelets. *Infect Immun* **85**, e00774-00716 (2017).

97. Takamatsu D, *et al.* Binding of theStreptococcus gordoniisurface glycoproteins GspB and Hsa to specific carbohydrate structures on platelet membrane glycoprotein Ibα. *Molecular Microbiology* **58**, 380-392 (2005).

98. Urano-Tashiro Y, Takahashi Y, Oguchi R, Konishi K. Two Arginine Residues of Streptococcus gordonii Sialic Acid-Binding Adhesin Hsa Are Essential for Interaction to Host Cell Receptors. *PloS one* **11**, e0154098-e0154098 (2016).

99. Bensing BA, *et al.* Recognition of specific sialoglycan structures by oral streptococci impacts the severity of endocardial infection. *PLoS pathogens* **15**, e1007896-e1007896 (2019).

100. Baker HM, Basu I, Chung MC, Caradoc-Davies T, Fraser JD, Baker EN. 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. *Journal of Molecular Biology* **374**, 1298-1308 (2007).

101. Cohen M, Varki A. Modulation of Glycan Recognition by Clustered Saccharide Patches. In: *International Review of Cell and Molecular Biology*). Elsevier (2014).

102. Deng L, *et al.* Host adaptation of a bacterial toxin from the human pathogen Salmonella Typhi. *Cell* **159**, 1290-1299 (2014).

103. Narimatsu Y, *et al.* An Atlas of Human Glycosylation Pathways Enables Display of the Human Glycome by Gene Engineered Cells. *Mol Cell* **75**, 394-407.e395 (2019).

104. Nagashima H, Tezuka T, Tsuchida W, Maeda H, Kohroki J, Masuho Y. Tandemly repeated Fc domain augments binding avidities of antibodies for Fcγ receptors, resulting in enhanced antibody-dependent cellular cytotoxicity. *Molecular Immunology* **45**, 2752-2763 (2008).

105. Moonens K, *et al.* Structural and functional insight into the carbohydrate receptor binding of F4 fimbriae-producing enterotoxigenic Escherichia coli. *The Journal of biological chemistry* **290**, 8409-8419 (2015).

106. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology* **310**, 311-325 (2001).

107. Moonens K, *et al.* Structural Insights into Polymorphic ABO Glycan Binding by Helicobacter pylori. *Cell Host Microbe* **19**, 55-66 (2016).

108. Chen SL, *et al.* Positive selection identifies an in vivo role for FimH during urinary tract infection in addition to mannose binding. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 22439-22444 (2009).

109. Aspholm-Hurtig M, *et al.* Functional Adaptation of BabA, the H. pylori ABO Blood Group Antigen Binding Adhesin. *Science* **305**, 519-522 (2004).

110. Csizmar CM, Petersburg JR, Perry TJ, Rozumalski L, Hackel BJ, Wagner CR. Multivalent Ligand Binding to Cell Membrane Antigens: Defining the Interplay of Affinity, Valency, and Expression Density. *Journal of the American Chemical Society* **141**, 251-261 (2019).

111. Moonens K, Remaut H. Evolution and structural dynamics of bacterial glycan binding adhesins. *Current Opinion in Structural Biology* **44**, 48-58 (2017).

112. Worstell NC, *et al.* Hetero-Multivalency of Pseudomonas aeruginosa Lectin LecA Binding to Model Membranes. *Scientific reports* **8**, 8419-8419 (2018).

113. Langereis MA, *et al.* Complexity and Diversity of the Mammalian Sialome Revealed by Nidovirus Virolectins. *Cell reports* **11**, 1966-1978 (2015).

114. Varki A. Nothing in glycobiology makes sense, except in the light of evolution. *Cell* **126**, 841-845 (2006).

115. Varki NM, Strobert E, Dick EJ, Jr., Benirschke K, Varki A. Biomedical differences between human and nonhuman hominids: potential roles for uniquely human aspects of sialic acid biology. *Annu Rev Pathol* **6**, 365-393 (2011).

116. Varki A. Biological roles of glycans. *Glycobiology* **27**, 3-49 (2017).

117. Thamadilok S, Roche-Hakansson H, Hakansson AP, Ruhl S. Absence of capsule reveals glycan-mediated binding and recognition of salivary mucin MUC7 by Streptococcus pneumoniae. *Mol Oral Microbiol* **31**, 175-188 (2016).

118. Takamatsu D, Bensing BA, Prakobphol A, Fisher SJ, Sullam PM. Binding of the streptococcal surface glycoproteins GspB and Hsa to human salivary proteins. *Infection and immunity* **74**, 1933-1940 (2006).

119. Takamatsu D, *et al.* Binding of the *Streptococcus gordonii* surface glycoproteins GspB and Hsa to specific carbohydrate structures on platelet membrane glycoprotein Ibalpha. *Molecular microbiology* **58**, 380-392 (2005).

120. Plummer C, Wu H, Kerrigan SW, Meade G, Cox D, Ian Douglas CW. A serine-rich glycoprotein of Streptococcus sanguis mediates adhesion to platelets via GPIb. *British journal of haematology* **129**, 101-109 (2005).

121. Bashore TM, Cabell C, Fowler V, Jr. Update on infective endocarditis. *Current problems in cardiology* **31**, 274-352 (2006).

122. Lizcano A, Sanchez CJ, Orihuela CJ. A role for glycosylated serine-rich repeat proteins in gram-positive bacterial pathogenesis. *Mol Oral Microbiol* **27**, 257-269 (2012).

123. Takahashi Y, Takashima E, Shimazu K, Yagishita H, Aoba T, Konishi K. Contribution of sialic acid-binding adhesin to pathogenesis of experimental endocarditis caused by Streptococcus gordonii DL1. *Infection and immunity* **74**, 740-743 (2006).

124. Stubbs HE, *et al.* Tandem sialoglycan-binding modules in a Streptococcus sanguinis serine-rich repeat adhesin create target dependent avidity effects. *The Journal of biological chemistry*, (2020).

125. Agarwal R, *et al.* Structure based virtual screening identifies small molecule effectors for the sialoglycan binding protein Hsa. *Biochem J* **477**, 3695-3707 (2020).

126. Di Carluccio C, *et al.* Molecular recognition of sialoglycans by streptococcal Siglec-like adhesins: toward the shape of specific inhibitors. *RSC Chemical Biology*, (2021).

127. Deng L, *et al.* Oral streptococci utilize a Siglec-like domain of serine-rich repeat adhesins to preferentially target platelet sialoglycans in human blood. *PLoS pathogens* **10**, e1004540 (2014).

128. Prakobphol A, *et al.* Human low-molecular-weight salivary mucin expresses the sialyl lewisx determinant and has L-selectin ligand activity. *Biochemistry* **37**, 4916-4927 (1998).

129. Karlsson NG, Thomsson KA. Salivary MUC7 is a major carrier of blood group I type O-linked oligosaccharides serving as the scaffold for sialyl Lewis x. *Glycobiology* **19**, 288-300 (2009).

130. Bensing BA, *et al.* Recognition of specific sialoglycan structures by oral streptococci impacts the severity of endocardial infection. *PLoS Pathog* **15**, e1007896 (2019).

131. Bensing BA, Lopez JA, Sullam PM. The Streptococcus gordonii surface proteins GspB and Hsa mediate binding to sialylated carbohydrate epitopes on the platelet membrane glycoprotein Ibalpha. *Infection and immunity* **72**, 6528-6537 (2004).

132. Xiong YQ, Bensing BA, Bayer AS, Chambers HF, Sullam PM. Role of the serine-rich surface glycoprotein GspB of *Streptococcus gordonii* in the pathogenesis of infective endocarditis. *Microbial pathogenesis* **45**, 297-301 (2008).

133. Wang JH. The sequence signature of an Ig-fold. *Protein Cell* **4**, 569-572 (2013).

134. Changeux JP, Edelstein S. Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol Rep* **3**, 19 (2011).

135. Johnson KA. Role of induced fit in enzyme specificity: a molecular forward/reverse switch. *J Biol Chem* **283**, 26297-26301 (2008).

136. Bensing BA, Thomas W, Iverson TM, Sullam PM. Recognition of specific sialoglycan structures by viridans group streptococci can impact the severity of endocardial infection. (in preparation).

137. Gaytan MO, *et al.* A novel sialic acid-binding adhesin present in multiple species contributes to the pathogenesis of Infective endocarditis. *PLoS pathogens* **17**, e1009222 (2021).

138. Ronis A, *et al.* Streptococcus oralis subsp. dentisani Produces Monolateral Serine-Rich Repeat Protein Fibrils, One of Which Contributes to Saliva Binding via Sialic Acid. *Infection and immunity* **87**, (2019).

139. Urano-Tashiro Y, Takahashi Y, Oguchi R, Konishi K. Two Arginine Residues of Streptococcus gordonii Sialic Acid-Binding Adhesin Hsa Are Essential for Interaction to Host Cell Receptors. *PloS one* **11**, e0154098 (2016).

140. May AP, Robinson RC, Vinson M, Crocker PR, Jones EY. Crystal structure of the N-terminal domain of sialoadhesin in complex with 3' sialyllactose at 1.85 A resolution. *Molecular cell* **1**, 719-728 (1998).

141. Vinson M, van der Merwe PA, Kelm S, May A, Jones EY, Crocker PR. Characterization of the sialic acid-binding site in sialoadhesin by site-directed mutagenesis. *J Biol Chem* **271**, 9267-9272 (1996).

142. Alphey MS, Attrill H, Crocker PR, van Aalten DM. High resolution crystal structures of Siglec-7. Insights into ligand specificity in the Siglec family. *J Biol Chem* **278**, 3372-3377 (2003).

143. Propster JM, Yang F, Rabbani S, Ernst B, Allain FH, Schubert M. Structural basis for sulfation-dependent self-glycan recognition by the human immune-inhibitory receptor Siglec-8. *Proc Natl Acad Sci U S A* **113**, E4170-4179 (2016).

144. Di Carluccio C, *et al.* Molecular recognition of sialoglycans by streptococcal Siglec-like adhesins: toward the shape of specific inhibitors. *RSC Chem Biol* **2**, 1618-1630 (2021).

145. Kolenbrander PE. Oral microbial communities: biofilms, interactions, and genetic systems. *Annu Rev Microbiol* **54**, 413-437 (2000).

146. Chou HH, *et al.* A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence. *Proc Natl Acad Sci U S A* **95**, 11751-11756 (1998).

147. Liu Y, *et al.* Structural basis of glycan specificity of P[19] VP8*: Implications for rotavirus zoonosis and evolution. *PLoS pathogens* **13**, e1006707 (2017).

148. Stroh LJ, *et al.* Structural Basis and Evolution of Glycan Receptor Specificities within the Polyomavirus Family. *mBio* **11**, (2020).

149. Ielasi FS, Verhaeghe T, Desmet T, Willaert RG. Engineering the carbohydrate-binding site of Epa1p from Candida glabrata: generation of adhesin mutants with different carbohydrate specificity. *Glycobiology* **24**, 1312-1322 (2014).

150. Salomonsson E, *et al.* Mutational tuning of galectin-3 specificity and biological function. *J Biol Chem* **285**, 35079-35091 (2010).

151. Hu D, Tateno H, Kuno A, Yabe R, Hirabayashi J. Directed evolution of lectins with sugar-binding specificity for 6-sulfo-galactose. *J Biol Chem* **287**, 20313-20320 (2012).

152. Hu D, Tateno H, Sato T, Narimatsu H, Hirabayashi J. Tailoring GalNAcalpha1-3Galbeta-specific lectins from a multi-specific fungal galectin: dramatic change of carbohydrate specificity by a single amino-acid substitution. *Biochem J* **453**, 261-270 (2013).

153. Abo H, Soga K, Tanaka A, Tateno H, Hirabayashi J, Yamamoto K. Mutated Leguminous Lectin Containing a Heparin-Binding like Motif in a Carbohydrate-Binding Loop Specifically Binds to Heparin. *PloS one* **10**, e0145834 (2015).

154. Imamura K, Takeuchi H, Yabe R, Tateno H, Hirabayashi J. Engineering of the glycan-binding specificity of Agrocybe cylindracea galectin towards alpha(2,3)-linked sialic acid by saturation mutagenesis. *Journal of biochemistry* **150**, 545-552 (2011).

155. Sato T, *et al.* Engineering of recombinant Wisteria floribunda agglutinin specifically binding to GalNAcbeta1,4GlcNAc (LacdiNAc). *Glycobiology* **27**, 743-754 (2017).

156. Hu D, *et al.* Engineering of a 3'-sulpho-Galbeta1-4GlcNAc-specific probe by a single amino acid substitution of a fungal galectin. *Journal of biochemistry* **157**, 197-200 (2015).

157. Nason R, *et al.* Display of the human mucinome with defined O-glycans by gene engineered cells. *Nat Commun* **12**, 4070 (2021).

158. Deng L, *et al.* Oral streptococci utilize a Siglec-like domain of serine-rich repeat adhesins to preferentially target platelet sialoglycans in human blood. *PLoS Pathog* **10**, e1004540 (2014).

159. Yu X, Dang VT, Fleming FE, von Itzstein M, Coulson BS, Blanchard H. Structural basis of rotavirus strain preference toward N-acetyl- or N-glycolylneuraminic acid-containing receptors. *J Virol* **86**, 13456-13466 (2012).

160. Dankwa S, *et al.* Ancient human sialic acid variant restricts an emerging zoonotic malaria parasite. *Nat Commun* **7**, 11187 (2016).

161. Ereño-Orbea J, *et al.* Molecular basis of human CD22 function and therapeutic targeting. *Nat Commun* **8**, 764 (2017).

162. T GS, *et al.* Molecular characterization of the interaction of sialic acid with the periplasmic binding protein from Haemophilus ducreyi. *The Journal of biological chemistry* **293**, (2018).

163. C DC, *et al.* Behavior of glycolylated sialoglycans in the binding pockets of murine and human CD22. *iScience* **24**, (2020).

164. Guo Y, *et al.* A Pasteurella multocida sialyltransferase displaying dual trans-sialidase activities for production of 3'-sialyl and 6'-sialyl glycans. *J Biotechnol* **170**, 60-67 (2014).

165. Choi YH, *et al.* Protein engineering of α2,3/2,6-sialyltransferase to improve the yield and productivity of in vitro sialyllactose synthesis. *Glycobiology* **24**, 159-169 (2014).

166. Guo Y, Jers C, Meyer AS, Li H, Kirpekar F, Mikkelsen JD. Modulating the regioselectivity of a Pasteurella multocida sialyltransferase for biocatalytic production of 3'- and 6'-sialyllactose. *Enzyme Microb Technol* **78**, 54-62 (2015).

167. Vimr ER, Bergstrom R, Steenbergen SM, Boulnois G, Roberts I. Homology among Escherichia coli K1 and K92 polysialytransferases. *J Bacteriol* **174**, 5127-5131 (1992).

168. Freiberger F, *et al.* Biochemical characterization of a Neisseria meningitidis polysialyltransferase reveals novel functional motifs in bacterial sialyltransferases. *Mol Microbiol* **65**, 1258-1275 (2007).

169. Puente-Polledo L, Reglero A, González-Clemente C, Rodríguez-Aparicio LB, Ferrero MA. Biochemical conditions for the production of polysialic acid by Pasteurella haemolytica A2. *Glycoconj J* **15**, 855-861 (1998).

170. McArthur JB, Yu H, Tasnima N, Lee CM, Fisher AJ, Chen X. α2-6-Neosialidase: A Sialyltransferase Mutant as a Sialyl Linkage-Specific Sialidase. *ACS Chem Biol* **13**, 1228-1234 (2018).

171. Matthews MM, McArthur JB, Li Y, Yu H, Chen X, Fisher AJ. Catalytic Cycle of Neisseria meningitidis CMP-Sialic Acid Synthetase Illustrated by High-Resolution Protein Crystallography. *Biochemistry* **59**, 3157-3168 (2020).

172. Liu W, Xiong W, Liu W, Wei Z, Abo H, Kawashima H. Therapeutic Effects of an Anti-Sialyl Lewis x Antibody in a Murine Model of Acute Lung Injury. *Monoclon Antib Immunodiagn Immunother* **42**, 97-103 (2023).

173. Pascoal C, *et al.* Sialyl Lewis(X/A) and Cytokeratin Crosstalk in Triple Negative Breast Cancer. *Cancers (Basel)* **15**, (2023).

174. Selvaraj C, Abhirami R, Vijayakumar R, Alfaiz FA, Singh SK. Immunological insights of selectins in human disease mechanism. *Adv Protein Chem Struct Biol* **129**, 163-188 (2022).

175. Liu X, *et al.* Molecular Insights into the Assembly and Functional Diversification of Typhoid Toxin. *mBio* **13**, e0191621 (2022).

176. Yabe R, Itakura Y, Nakamura-Tsuruta S, Iwaki J, Kuno A, Hirabayashi J. Engineering a versatile tandem repeat-type alpha2-6sialic acid-binding lectin. *Biochem Biophys Res Commun* **384**, 204-209 (2009).

177. Ohyabu N, *et al.* An essential epitope of anti-MUC1 monoclonal antibody KL-6 revealed by focused glycopeptide library. *J Am Chem Soc* **131**, 17102-17109 (2009).

178. Schmölzer K, *et al.* Complete switch from α-2,3- to α-2,6-regioselectivity in Pasteurella dagmatis β-D-galactoside sialyltransferase by active-site redesign. *Chem Commun (Camb)* **51**, 3083-3086 (2015).

179. Kim DU, Yoo JH, Lee YJ, Kim KS, Cho HS. Structural analysis of sialyltransferase PM0188 from Pasteurella multocida complexed with donor analogue and acceptor sugar. *BMB Rep* **41**, 48-54 (2008).

180. Takasaki N, *et al.* A heterozygous mutation of GALNTL5 affects male infertility with impairment of sperm motility. *Proc Natl Acad Sci U S A* **111**, 1120-1125 (2014).

181. Vision Research Needs, Gaps, and Opportunities.). National Eye Institute (2012).

182. Shaberman B, Durham T. The Foundation Fighting Blindness Plays an Essential and Expansive Role in Driving Genetic Research for Inherited Retinal Diseases. *Genes* **10**, (2019).

183. T. Alastalo KT, B. Shaberman. Introducing an open access genetic testing program for patients with inherited retinal degeneration.). https://www.blueprintgenetics.com (2019).

184. Reiff C, *et al.* The mutation p.E113K in the Schiff base counterion of rhodopsin is associated with two distinct retinal phenotypes within the same family. *Scientific reports* **6**, 36208 (2016).

185. Bowes C, Li T, Danciger M, Baxter LC, Applebury ML, Farber DB. Retinal degeneration in the rd mouse is caused by a defect in the beta subunit of rod cGMP-phosphodiesterase. *Nature* **347**, 677-680 (1990).

186. Sidman RL, Green MC. RETINAL DEGENERATION IN THE MOUSE: LOCATION OF THE RD LOCUS IN LINKAGE GROUP XVII. *J Hered* **56**, 23-29 (1965).

187. Hawkins RK, Jansen HG, Sanyal S. Development and degeneration of retina in rds mutant mice: photoreceptor abnormalities in the heterozygotes. *Exp Eye Res* **41**, 701-720 (1985).

188. Sanyal S, Zeilmaker GH. Development and degeneration of retina in rds mutant mice: light and electron microscopic observations in experimental chimaeras. *Exp Eye Res* **39**, 231-246 (1984).

189. Jansen HG, Sanyal S. Development and degeneration of retina in rds mutant mice: electron microscopy. *J Comp Neurol* **224**, 71-84 (1984).

190. Sanyal S, Dees C, Zeilmaker GH. Development and degeneration of retina in rds mutant mice: observations in chimaeras of heterozygous mutant and normal genotype. *J Embryol Exp Morphol* **98**, 111-121 (1986).

191. Sanyal S, Hawkins RK. Development and degeneration of retina in rds mutant mice: effects of light on the rate of degeneration in albino and pigmented homozygous and heterozygous mutant and normal mice. *Vision Res* **26**, 1177-1185 (1986).

192. Sanyal S, Hawkins RK, Zeilmaker GH. Development and degeneration of retina in rds mutant mice: analysis of interphotoreceptor matrix staining in chimaeric retina. *Curr Eye Res* **7**, 1183-1190 (1988).

193. Sanyal S, Hawkins RK. Development and degeneration of retina in rds mutant mice: altered disc shedding pattern in the albino heterozygotes and its relation to light exposure. *Vision Res* **28**, 1171-1178 (1988).

194. Sanyal S, Hawkins RK. Development and degeneration of retina in rds mutant mice: altered disc shedding pattern in the heterozygotes and its relation to ocular pigmentation. *Curr Eye Res* **8**, 1093-1101 (1989).

195. Schalken JJ, Janssen JJ, Sanyal S, Hawkins RK, de Grip WJ. Development and degeneration of retina in rds mutant mice: immunoassay of the rod visual pigment rhodopsin. *Biochimica et biophysica acta* **1033**, 103-109 (1990).

196. Sanyal S, Jansen HG. Absence of receptor outer segments in the retina of rds mutant mice. *Neurosci Lett* **21**, 23-26 (1981).

197. Travis GH, Brennan MB, Danielson PE, Kozak CA, Sutcliffe JG. Identification of a photoreceptor-specific mRNA encoded by the gene responsible for retinal degeneration slow (rds). *Nature* **338**, 70-73 (1989).

198. Travis GH, *et al.* The human retinal degeneration slow (RDS) gene: chromosome assignment and structure of the mRNA. *Genomics* **10**, 733-739 (1991).

199. Connell GJ, Molday RS. Molecular cloning, primary structure, and orientation of the vertebrate photoreceptor cell protein peripherin in the rod outer segment disk membrane. *Biochemistry* **29**, 4691-4698 (1990).

200. Connell G, Bascom R, Molday L, Reid D, McInnes RR, Molday RS. Photoreceptor peripherin is the normal product of the gene responsible for retinal degeneration in the rds mouse. *Proc Natl Acad Sci U S A* **88**, 723-726 (1991).

201. Goldberg AF, Molday RS. Subunit composition of the peripherin/rds-rom-1 disk rim complex from rod photoreceptors: hydrodynamic evidence for a tetrameric quaternary structure. *Biochemistry* **35**, 6144-6149 (1996).

202. Arikawa K, Molday LL, Molday RS, Williams DS. Localization of peripherin/rds in the disk membranes of cone and rod photoreceptors: relationship to disk membrane morphogenesis and retinal degeneration. *J Cell Biol* **116**, 659-667 (1992).

203. Molday RS, Hicks D, Molday L. Peripherin. A rim-specific membrane protein of rod outer segment discs. *Investigative ophthalmology & visual science* **28**, 50-61 (1987).

204. Tam BM, Moritz OL, Papermaster DS. The C terminus of peripherin/rds participates in rod outer segment targeting and alignment of disk incisures. *Molecular biology of the cell* **15**, 2027-2037 (2004).

205. Salinas RY, Baker SA, Gospe SM, 3rd, Arshavsky VY. A single valine residue plays an essential role in peripherin/rds targeting to photoreceptor outer segments. *PloS one* **8**, e54292 (2013).

206. Ding XQ, Stricker HM, Naash MI. Role of the second intradiscal loop of peripherin/rds in homo and hetero associations. *Biochemistry* **44**, 4897-4904 (2005).

207. Chakraborty D, Rodgers KK, Conley SM, Naash MI. Structural characterization of the second intra-discal loop of the photoreceptor tetraspanin RDS. *The FEBS journal* **280**, 127-138 (2013).

208. Goldberg AF, Loewen CJ, Molday RS. Cysteine residues of photoreceptor peripherin/rds: role in subunit assembly and autosomal dominant retinitis pigmentosa. *Biochemistry* **37**, 680-685 (1998).

209. Kevany BM, Tsybovsky Y, Campuzano ID, Schnier PD, Engel A, Palczewski K. Structural and functional analysis of the native peripherin-ROM1 complex isolated from photoreceptor cells. *The Journal of biological chemistry* **288**, 36272-36284 (2013).

210. Boesze-Battaglia K, Goldberg AF, Dispoto J, Katragadda M, Cesarone G, Albert AD. A soluble peripherin/Rds C-terminal polypeptide promotes membrane fusion and changes conformation upon membrane association. *Exp Eye Res* **77**, 505-514 (2003).

211. Ritter LM, *et al.* Uncoupling of photoreceptor peripherin/rds fusogenic activity from biosynthesis, subunit assembly, and targeting: a potential mechanism for pathogenic effects. *The Journal of biological chemistry* **279**, 39958-39967 (2004).

212. Hemler ME. Tetraspanin functions and associated microdomains. *Nature reviews Molecular cell biology* **6**, 801-811 (2005).

213. van Deventer SJ, Dunlock VE, van Spriel AB. Molecular interactions shaping the tetraspanin web. *Biochemical Society transactions* **45**, 741-750 (2017).

214. Loewen CJ, Molday RS. Disulfide-mediated oligomerization of Peripherin/Rds and Rom-1 in photoreceptor disk membranes. Implications for photoreceptor outer segment morphogenesis and degeneration. *The Journal of biological chemistry* **275**, 5370-5378 (2000).

215. Milstein ML, Kimler VA, Ghatak C, Ladokhin AS, Goldberg AFX. An inducible amphipathic helix within the intrinsically disordered C terminus can participate in membrane curvature generation by peripherin-2/rds. *The Journal of biological chemistry* **292**, 7850-7865 (2017).

216. Chakraborty D, Ding XQ, Fliesler SJ, Naash MI. Outer segment oligomerization of Rds: evidence from mouse models and subcellular fractionation. *Biochemistry* **47**, 1144-1156 (2008).

217. Bascom RA, Manara S, Collins L, Molday RS, Kalnins VI, McInnes RR. Cloning of the cDNA for a novel photoreceptor membrane protein (rom-1) identifies a disk rim protein family implicated in human retinopathies. *Neuron* **8**, 1171-1184 (1992).

218. Moritz OL, Molday RS. Molecular cloning, membrane topology, and localization of bovine rom-1 in rod and cone photoreceptor cells. *Investigative ophthalmology & visual science* **37**, 352-362 (1996).

219. Clarke G, *et al.* Rom-1 is required for rod photoreceptor viability and the regulation of disk morphogenesis. *Nat Genet* **25**, 67-73 (2000).

220. Chakraborty D, Conley SM, Al-Ubaidi MR, Naash MI. Initiation of rod outer segment disc formation requires RDS. *PloS one* **9**, e98939 (2014).

221. Khattree N, Ritter LM, Goldberg AF. Membrane curvature generation by a C-terminal amphipathic helix in peripherin-2/rds, a tetraspanin required for photoreceptor sensory cilium morphogenesis. *Journal of cell science* **126**, 4659-4670 (2013).

222. Drin G, Antonny B. Amphipathic helices and membrane curvature. *FEBS letters* **584**, 1840-1847 (2010).

223. Kedzierski W, *et al.* Deficiency of rds/peripherin causes photoreceptor death in mouse models of digenic and dominant retinitis pigmentosa. *Proc Natl Acad Sci U S A* **98**, 7718-7723 (2001).

224. Zulliger R, Conley SM, Mwoyosvi ML, Al-Ubaidi MR, Naash MI. Oligomerization of Prph2 and Rom1 is essential for photoreceptor outer segment formation. *Hum Mol Genet* **27**, 3507-3518 (2018).

225. Conley SM, Al-Ubaidi MR, Naash MI. The Role of the Prph2 C-Terminus in Outer Segment Morphogenesis. *Adv Exp Med Biol* **1185**, 495-499 (2019).

226. Conley SM, Stuck MW, Watson JN, Zulliger R, Burnett JL, Naash MI. Prph2 initiates outer segment morphogenesis but maturation requires Prph2/Rom1 oligomerization. *Hum Mol Genet* **28**, 459-475 (2019).

227. Salinas RY, Pearring JN, Ding JD, Spencer WJ, Hao Y, Arshavsky VY. Photoreceptor discs form through peripherin-dependent suppression of ciliary ectosome release. *J Cell Biol* **216**, 1489-1499 (2017).

228. Boesze-Battaglia K, Goldberg AF. Photoreceptor renewal: a role for peripherin/rds. *International review of cytology* **217**, 183-225 (2002).

229. Boesze-Battaglia K, Kong F, Lamba OP, Stefano FP, Williams DS. Purification and light-dependent phosphorylation of a candidate fusion protein, the photoreceptor cell peripherin/rds. *Biochemistry* **36**, 6835-6846 (1997).

230. Boesze-Battaglia K, Stefano FP, Fenner M, Napoli AA, Jr. A peptide analogue to a fusion domain within photoreceptor peripherin/rds promotes membrane adhesion and depolarization. *Biochimica et biophysica acta* **1463**, 343-354 (2000).

231. Kennedy CJ, Rakoczy PE, Constable IJ. Lipofuscin of the retinal pigment epithelium: a review. *Eye (Lond)* **9 (Pt 6)**, 763-771 (1995).

232. Palma MMD, Martin D, Salles MV, Motta FLT, Abujamra S, Sallum JMF. Retinal dystrophies and variants in PRPH2. *Arq Bras Oftalmol* **82**, 158-160 (2019).

233. Wells J, *et al.* Mutations in the human retinal degeneration slow (RDS) gene can cause either retinitis pigmentosa or macular dystrophy. *Nat Genet* **3**, 213-218 (1993).

234. Daftarian N, *et al.* PRPH2 mutation as the cause of various clinical manifestations in a family affected with inherited retinal dystrophy. *Ophthalmic Genet* **40**, 436-442 (2019).

235. Chakraborty D, *et al.* Novel molecular mechanisms for Prph2-associated pattern dystrophy. *Faseb j* **34**, 1211-1230 (2020).

236. Ma CJ, *et al.* Late-onset pattern macular dystrophy mimicking ABCA4 and PRPH2 disease is caused by a homozygous frameshift mutation in ROM1. *Cold Spring Harb Mol Case Stud* **5**, (2019).

237. Kajiwara K, Sandberg MA, Berson EL, Dryja TP. A null mutation in the human peripherin/RDS gene in a family with autosomal dominant retinitis punctata albescens. *Nat Genet* **3**, 208-212 (1993).

238. Nour M, Ding XQ, Stricker H, Fliesler SJ, Naash MI. Modulating expression of peripherin/rds in transgenic mice: critical levels and the effect of overexpression. *Investigative ophthalmology & visual science* **45**, 2514-2521 (2004).

239. Conley S, Nour M, Fliesler SJ, Naash MI. Late-onset cone photoreceptor degeneration induced by R172W mutation in Rds and partial rescue by gene supplementation. *Investigative ophthalmology & visual science* **48**, 5397-5407 (2007).

240. Nour M, Fliesler SJ, Naash MI. Genetic supplementation of RDS alleviates a loss-of-function phenotype in C214S model of retinitis pigmentosa. *Adv Exp Med Biol* **613**, 129-138 (2008).

241. Bohm S, *et al.* Peripherin-2 and Rom-1 have opposing effects on rod outer segment targeting of retinitis pigmentosa-linked peripherin-2 mutants. *Scientific reports* **7**, 2321 (2017).

242. Molday RS, Molday LL, Loewen CJ. Role of subunit assembly in autosomal dominant retinitis pigmentosa linked to mutations in peripherin 2. *Novartis Found Symp* **255**, 95-112; discussion 113-116, 177-118 (2004).

243. Poloschek CM, *et al.* ABCA4 and ROM1: implications for modification of the PRPH2-associated macular dystrophy phenotype. *Investigative ophthalmology & visual science* **51**, 4253-4265 (2010).

244. Kajiwara K, Berson EL, Dryja TP. Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science (New York, NY)* **264**, 1604-1608 (1994).

245. Goldberg AF, Molday RS. Defective subunit assembly underlies a digenic form of retinitis pigmentosa linked to mutations in peripherin/rds and rom-1. *Proc Natl Acad Sci U S A* **93**, 13726-13730 (1996).

246. Loewen CJ, Moritz OL, Molday RS. Molecular characterization of peripherin-2 and rom-1 mutants responsible for digenic retinitis pigmentosa. *The Journal of biological chemistry* **276**, 22388-22396 (2001).

247. Dryja TP, Hahn LB, Kajiwara K, Berson EL. Dominant and digenic mutations in the peripherin/RDS and ROM1 genes in retinitis pigmentosa. *Investigative ophthalmology & visual science* **38**, 1972-1982 (1997).

248. Conley SM, Stuck MW, Watson JN, Naash MI. Rom1 converts Y141C-Prph2-associated pattern dystrophy to retinitis pigmentosa. *Hum Mol Genet* **26**, 509-518 (2017).

249. Ding XQ, Nour M, Ritter LM, Goldberg AF, Fliesler SJ, Naash MI. The R172W mutation in peripherin/rds causes a cone-rod dystrophy in transgenic mice. *Hum Mol Genet* **13**, 2075-2087 (2004).

250. Stuck MW, Conley SM, Naash MI. The Y141C knockin mutation in RDS leads to complex phenotypes in the mouse. *Hum Mol Genet* **23**, 6260-6274 (2014).

251. Stuck MW, Conley SM, Naash MI. Retinal Degeneration Slow (RDS) Glycosylation Plays a Role in Cone Function and in the Regulation of RDS.ROM-1 Protein Complex Formation. *The Journal of biological chemistry* **290**, 27901-27913 (2015).

252. Milstein ML, Cavanaugh BL, Roussey NM, Volland S, Williams DS, Goldberg AFX. Multistep peripherin-2/rds self-assembly drives membrane curvature for outer segment disk architecture and photoreceptor viability. *Proc Natl Acad Sci U S A,* (2020).

253. Gurevich VV, Gurevich EV. Plethora of functions packed into 45 kDa arrestins: biological implications and possible therapeutic strategies. *Cellular and molecular life sciences : CMLS* **76**, 4413-4421 (2019).

254. Perry NA, *et al.* Arrestin-3 scaffolding of the JNK3 cascade suggests a mechanism for signal amplification. *Proc Natl Acad Sci U S A* **116**, 810-815 (2019).

255. Ahler E, *et al.* A Combined Approach Reveals a Regulatory Mechanism Coupling Src's Kinase Activity, Localization, and Phosphotransferase-Independent Functions. *Molecular cell* **74**, 393-408.e320 (2019).

256. Sicheri F, Moarefi I, Kuriyan J. Crystal structure of the Src family tyrosine kinase Hck. *Nature* **385**, 602-609 (1997).

257. Xu W, Harrison SC, Eck MJ. Three-dimensional structure of the tyrosine kinase c-Src. *Nature* **385**, 595-602 (1997).

258. Luttrell LM, *et al.* Beta-arrestin-dependent formation of beta2 adrenergic receptor-Src protein kinase complexes. *Science (New York, NY)* **283**, 655-661 (1999).

259. Miller WE, Maudsley S, Ahn S, Khan KD, Luttrell LM, Lefkowitz RJ. beta-arrestin1 interacts with the catalytic domain of the tyrosine kinase c-SRC. Role of beta-arrestin1-dependent targeting of c-SRC in receptor endocytosis. *The Journal of biological chemistry* **275**, 11312-11319 (2000).

260. Aeffner S, Reusch T, Weinhausen B, Salditt T. Energetics of stalk intermediates in membrane fusion are controlled by lipid composition. *Proc Natl Acad Sci U S A* **109**, E1609-1618 (2012).

261. Sharma S, Lindau M. t-SNARE Transmembrane Domain Clustering Modulates Lipid Organization and Membrane Curvature. *Journal of the American Chemical Society* **139**, 18440-18443 (2017).

262. Sharma S, Lindau M. Molecular mechanism of fusion pore formation driven by the neuronal SNARE complex. *Proc Natl Acad Sci U S A* **115**, 12751-12756 (2018).

263. Shi L, *et al.* SNARE proteins: one to fuse and three to keep the nascent fusion pore open. *Science (New York, NY)* **335**, 1355-1359 (2012).

264. Sutton RB, Davletov BA, Berghuis AM, Sudhof TC, Sprang SR. Structure of the first C2 domain of synaptotagmin I: a novel Ca2+/phospholipid-binding fold. *Cell* **80**, 929-938 (1995).

265. Zhou Q, *et al.* The primed SNARE-complexin-synaptotagmin complex for neuronal exocytosis. *Nature* **548**, 420-425 (2017).

266. Geppert M, *et al.* Synaptotagmin I: a major Ca2+ sensor for transmitter release at a central synapse. *Cell* **79**, 717-727 (1994).

267. Chapman ER, Hanson PI, An S, Jahn R. Ca2+ regulates the interaction between synaptotagmin and syntaxin 1. *The Journal of biological chemistry* **270**, 23667-23671 (1995).

268. Sorensen JB. Formation, stabilisation and fusion of the readily releasable pool of secretory vesicles. *Pflugers Archiv : European journal of physiology* **448**, 347-362 (2004).

269. Yoon EJ, Hamm HE, Currie KP. G protein betagamma subunits modulate the number and nature of exocytotic fusion events in adrenal chromaffin cells independent of calcium entry. *Journal of neurophysiology* **100**, 2929-2939 (2008).

270. Zurawski Z, *et al.* Disabling the Gβγ-SNARE interaction disrupts GPCR-mediated presynaptic inhibition, leading to physiological and behavioral phenotypes. **12**, eaat8595 (2019).

271. Zurawski Z, *et al.* Disabling Gβγ SNARE interaction in transgenic mice disrupts GPCR-mediated presynaptic inhibition leading to physiological and behavioral phenotypes. 280347 (2018).

272. Nowycky MC, Fox AP, Tsien RW. Three types of neuronal calcium channel with different calcium agonist sensitivity. *Nature* **316**, 440-443 (1985).

273. Tang L, *et al.* Structural basis for Ca2+ selectivity of a voltage-gated calcium channel. *Nature* **505**, 56-61 (2014).

274. Zhou Q, *et al.* Architecture of the synaptotagmin-SNARE machinery for neuronal exocytosis. *Nature* **525**, 62-67 (2015).

275. Fernandez-Chacon R, *et al.* Synaptotagmin I functions as a calcium regulator of release probability. *Nature* **410**, 41-49 (2001).

276. Sutton RB, Fasshauer D, Jahn R, Brunger AT. Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. *Nature* **395**, 347-353 (1998).

277. Blackmer T, Larsen EC, Takahashi M, Martin TF, Alford S, Hamm HE. G protein betagamma subunit-mediated presynaptic inhibition: regulation of exocytotic fusion downstream of Ca2+ entry. *Science (New York, NY)* **292**, 293-297 (2001).

278. Gerachshenko T, Blackmer T, Yoon EJ, Bartleson C, Hamm HE, Alford S. Gbetagamma acts at the C terminus of SNAP-25 to mediate presynaptic inhibition. *Nature neuroscience* **8**, 597-605 (2005).

279. Blackmer T, *et al.* G protein betagamma directly regulates SNARE protein fusion machinery for secretory granule exocytosis. *Nature neuroscience* **8**, 421-425 (2005).

280. Yoon EJ, Gerachshenko T, Spiegelberg BD, Alford S, Hamm HE. Gbetagamma interferes with Ca2+-dependent binding of synaptotagmin to the soluble N-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) complex. *Molecular pharmacology* **72**, 1210-1219 (2007).

281. Wells CA, *et al.* Gbetagamma inhibits exocytosis via interaction with critical residues on soluble N-ethylmaleimide-sensitive factor attachment protein-25. *Molecular pharmacology* **82**, 1136-1149 (2012).

282. Zurawski Z, Rodriguez S, Hyde K, Alford S, Hamm HE. Gbetagamma Binds to the Extreme C Terminus of SNAP25 to Mediate the Action of Gi/o-Coupled G Protein-Coupled Receptors. *Molecular pharmacology* **89**, 75-83 (2016).

283. Zurawski Z, *et al.* Gbetagamma directly modulates vesicle fusion by competing with synaptotagmin for binding to neuronal SNARE proteins embedded in membranes. *The Journal of biological chemistry* **292**, 12165-12177 (2017).

284. Garcia-Nafria J, Nehme R, Edwards PC, Tate CG. Cryo-EM structure of the serotonin 5-HT1B receptor coupled to heterotrimeric Go. *Nature* **558**, 620-623 (2018).

285. Khan SM, Sung JY, Hebert TE. Gbetagamma subunits-Different spaces, different faces. *Pharmacological research* **111**, 434-441 (2016).

286. Wells CA, Betke KM, Lindsley CW, Hamm HE. Label-free detection of G protein-SNARE interactions and screening for small molecule modulators. *ACS chemical neuroscience* **3**, 69-78 (2012).