MODELING PROTEIN–CARBOHYDRATE COMPLEXES IN ROSETTA

by
Morgan Lynsey Nance

A dissertation submitted to The Johns Hopkins University
in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland, U.S.A
August 2023

Abstract

Carbohydrates are of fundamental importance in biology. These molecules are essential to life, serving to mediate diverse biological functions. Unraveling the biophysical mechanisms by which carbohydrates operate not only helps complete our understanding of life on Earth but enables rational engineering to fine-tune or even modify their roles in biology. However, carbohydrate molecules are complex, with their conformational diversity and chemical heterogeneity making it notoriously difficult to elucidate their structures experimentally. Yet these models are vital to our mechanistic understanding of carbohydrate-mediated biological functions, making scientific advancements challenging to achieve. Computational methods serve to fill this gap by generating native-like models of protein–carbohydrate systems.

In this dissertation, I describe my advancements to the field of computational modeling with the development of GlycanDock, a protein–carbohydrate docking refinement method in Rosetta. I detail the extensive benchmark I developed to evaluate the effectiveness of the GlycanDock protocol to generate native-like protein–carbohydrate models. Further, I provide residue-level analyses of these models to demonstrate the utility of the protocol toward developing a biophysical understanding of protein–carbohydrate complexes. Finally, I describe an approach utilizing GlycanDock and other computational tools to address the more realistic “blind” docking scenarios.

The development of GlycanDock enabled my work computationally modeling the structures of *Fp*GalNAcDeAc and *Fp*GalNase, two enzymes that together convert A-type blood to the universal O-type. I identified *Fp*GalNAcDeAc residues likely to govern the binding of terminal LacNAc motifs present on the surface of red blood cells, offering mutational sites to modify targeting to the cell surface. Additionally, I identified the *Fp*GalNAcDeAc binding site residues most important for A-antigen recognition, providing a guide to understanding and controlling its enzymatic activity. For *Fp*GalNase, I proposed a sequence- and structure-driven hypothesis regarding its active-site and unique specificity to the terminal α-GalN carbohydrate. My work serves as a blueprint for future experimental studies, including rational engineering toward modifying *Fp*GalNase's specificity to the B-antigen, which, if achieved, would mean complete conversion of all A, B, and AB blood types to the universal O-type.

In sum, my work advanced our ability to model and dissect protein–carbohydrate systems.

Thesis Committee

Jeffrey J. Gray (Primary Advisor)
Professor
Department of Chemical and Biomolecular Engineering
Johns Hopkins Whiting School of Engineering

Albert Y. Lau (Reader)
Associate Professor
Department of Biophysics and Biophysical Chemistry
Johns Hopkins School of Medicine

Doug Barrick
Professor and Chair
T.C. Jenkins Department of Biophysics
Johns Hopkins Krieger School of Arts and Sciences

Steven E. Rokita
Professor
Department of Chemistry
Johns Hopkins Krieger School of Arts and Sciences

Marc Ostermeier
Professor
Department of Chemical and Biomolecular Engineering
Johns Hopkins Whiting School of Engineering

Dedication

This dissertation is dedicated to the people who made this achievement possible:

My parents, Jaclyn and John Gates,
whose lifelong support and encouragement has been total and unwavering
and for whom I am eternally appreciative and grateful

My grandma, Betty Ball,
who never let me forget all the work I put in to reach this achievement
and ensured I remembered so with every phone call

My sister, Taylor Nunes,
who, like me, generally had no idea what I was doing during graduate school,
but was nevertheless supportive of me and my journey

My dear friends, Keila Voortman-Sheetz, Rahel Frick, and Kelly Karl
who truly carried me every step of the way
by guiding me out of my lows and celebrating with me at my highs

In loving memory of

My father, Gregory Alvin Nance

It’s dangerous to go alone! Take this.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ABO: A-, B-, and O-type blood

CBM: Carbohydrate-Binding Module

DoFs: Degrees of Freedom

Fp: *Flavonifractor plautii*

*Fp*GalNAcDeAc: A-type *N*-acetyl-alpha-D-galactosamine deacetylase

*Fp*CBM32: the CBM32 domain of *Fp*GalNAcDeAc

*Fp*DeAc: the catalytic domain of *Fp*GalNAcDeAc

*Fp*GalNase: A-type alpha-D-galactosamine galactosaminidase

GBP: Glycan Binding Protein

GH: Glycosyl Hydrolase

LacNAc: *N*-acetyllactosamine

MD: Molecular Dynamics

(M)MCM: (Metropolis) Monte Carlo-plus-Minimization

MSA: Multiple Sequence Alignment

PDB: Protein Databank

PROSS: Protein Repair One Stop Shop

RBC: Red Blood Cell

REF2015: Rosetta Energy Function 2015

REU: Rosetta Energy Unit

RMSD: Root-Mean Square Deviation

SRMSD: Shape Root-Mean Square Deviation

Chapter 1

1. Introduction

1.1 Carbohydrate structure and biological function

The *Essentials of Glycobiology* textbook 4th edition (the most recent edition at the time of writing) provides an excellent, thorough description of carbohydrate structure and function. I highly encourage readers who are interested in delving deeper into the realm of glycobiology (the study of carbohydrates in biological systems) to read the free e-book of the *Essentials of Glycobiology* 4th edition at: https://www.ncbi.nlm.nih.gov/books/n/glyco4/.

Carbohydrates are the most structurally and chemically diverse biomolecule and are found in all living organisms on Earth. Carbohydrates are highly hydrated carbon-based molecules (chemical formula $C_x(H_2O)_n$). They primarily exist in a cyclic (*i.e.*, ring) form and contain at least one asymmetric carbon. Carbohydrates are therefore chiral molecules that exist in either D or L configuration, with D being the primary stereoisomer in vertebrates[1]. The structural diversity of monosaccharides (single carbohydrate units) is due to (1) ring size (*e.g.*, a five-membered pyranose or a six-membered hexose), (2) the conformation of the ring itself (*e.g.*, type $^4C_1$ chair *versus* $^1C_4$ chair *versus* “envelope” *versus* “twist” conformations of a pyranose), and (3) the configuration of the anomeric carbon (*i.e.*, α *versus* β stereoisomers)[1]. The chemical diversity of monosaccharides is due to (1) epimerization of the hydroxyl groups (α *versus* β for the anomeric carbon, axial *versus* equatorial otherwise) and (2) chemical modification (*e.g.*, methylation and acetylation).

A comparison of the ten most abundant monosaccharides in mammals provides a visualization of these components of carbohydrate structural and chemical diversity (**Figure 1.1**). For example, three of the ten monosaccharides are epimers of each other (D-glucose, D-galactose, and D-mannose), three have additional chemical modifications (*N*-acetyl-D-glucosamine, *N*-acetyl-D-galactosamine, and D-glucuronic acid), and the final four are otherwise modified (D-xylose, L-fucose, iduronic acid, and *N*-acetylneuraminic acid (a type of sialic acid))[1] (see **Figure 1.1**).

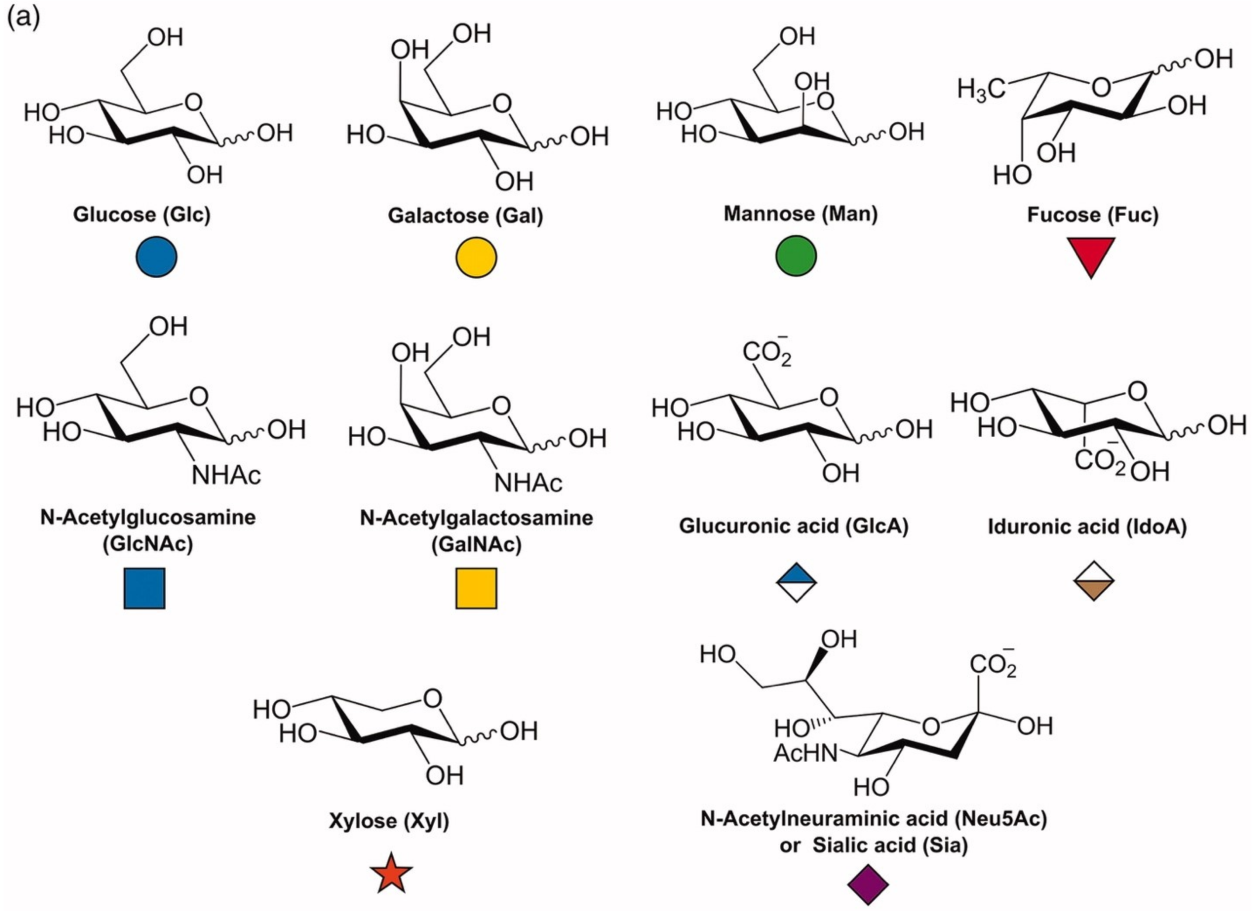
(a)
Glucose (Glc)
Galactose (Gal)
Mannose (Man)
Fucose (Fuc)
N-Acetylglucosamine (GlcNAc)
N-Acetylgalactosamine (GalNAc)
Glucuronic acid (GlcA)
Iduronic acid (IdoA)
Xylose (Xyl)
N-Acetylneuraminic acid (Neu5Ac) or Sialic acid (Sia)

Figure 1.1: The ten most common monosaccharides in mammals and their corresponding Symbol Nomenclature for Glycans (SNFG) notation. Reprinted with permission from Figure 1A of Chen S., Qin R, and Mahal L, Crit. Rev. Biochem. Mol. Biol. 2021.

The complexity of carbohydrates extends beyond the monosaccharide unit. In living organisms, carbohydrates are most often found covalently linked together as oligo- and polysaccharide chains (generally defined as chains of < 12 and ≥ 12 monosaccharide units, respectively). **These carbohydrate chains are often referred to as “glycans”.** Monosaccharides are joined together *via* glycosidic bonds which create a flexible glycosidic linkage between the two units. A glycosidic bond is formed between the anomeric carbon of one carbohydrate and a hydroxyl group of the other carbohydrate, meaning different regioisomers are possible depending on which hydroxyl group is used for the bond. Typically, using an intracyclic hydroxyl group results in a glycosidic linkage with ϕ and ψ dihedrals whereas using an exocyclic hydroxyl group results in a glycosidic linkage with ϕ, ψ, and ω dihedrals. A single glycosidic torsion can contain either zero, one, or even two ω dihedral angles, depending on where (*i.e.*, through which atoms) the glycosidic bond is made. Unlike protein ω dihedral angles which are generally fixed given its double bonded characteristic, carbohydrate ω angles are very flexible. See Section 1.4 for a visual comparison of protein *versus* carbohydrate ϕ, ψ, and ω dihedral angles.

In biological systems, glycan chains (consisting of various monosaccharide units and glycosidic linkages as described above) are primarily observed covalently attached to other macromolecular structures such as proteins (*i.e.*, glycoproteins) and lipids (*i.e.*, glycolipids; frequently those at the cell surface)[2]. In glycoproteins, the side-chain nitrogen of an asparagine or the side-chain oxygen of a serine or threonine serves as the attachment point, resulting in an *N*-linked or *O*-linked glycan, respectively. Glycans are enzymatically installed at a given attachment point starting with a pre-assembled “core” structure (*e.g.*, for *N*-glycans, a “core” pentasaccharide consisting of two *N*-acetylglucosamines and three mannoses). Once attached, the “core” structure can then be further enzymatically modified to create a more complex glycan, such as the highly branched variants characteristic of mammalian cell-surface glycans. Glycans are considered “branched” when more than one glycosidic bond is made to a single monosaccharide unit. **Figure 1.2** depicts the three most common *N*-linked glycans in mammals, each of which vary in size, complexity, and number of branch points.

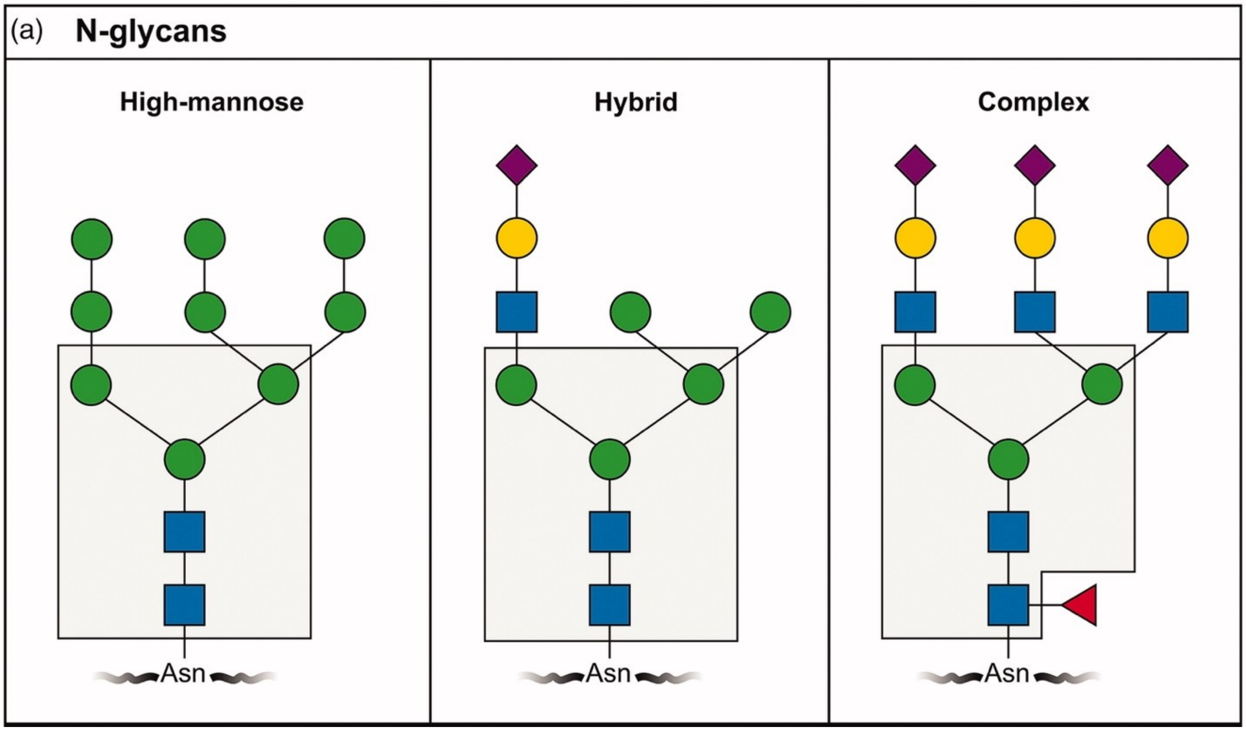

Figure 1.2: Illustration of the three most common *N*-linked glycans in mammals, which differ in chain size, chain length, and number of branch points. Reprinted with permission from Figure 2A of Chen S., Qin R, and Mahal L, *Crit. Rev. Biochem. Mol. Biol.* 2021.

Glycans play many diverse, important roles in biology. These biological roles can be broadly classified as structural (*e.g.*, rigidity and curvature of cellular membranes), metabolic (*e.g.*, sources of energy and energy storage), modulatory (*e.g.*, protein folding, misfolding, and degradation), intrinsic recognition (*e.g.*, cell–cell interactions), and extrinsic recognition (*e.g.*, pathogen detection)[3]. Many of these biological roles are modulated by the identity of the terminal carbohydrate (*i.e.*, the carbohydrate at the non-reducing end(s)). For example, human ABO blood typing is determined by the presence or absence of specific carbohydrate blood-antigens on the termini of red-blood cell associated glycolipids[4]. “Foreign” carbohydrate blood antigens (*i.e.*, non-compatible blood types) are recognized by antibodies (important proteins in our immune system) and can potentially trigger a lethal immune response. The ABO carbohydrate blood antigens are an important component of my doctoral work and are discussed further in Chapter 3.

Our final introductory topic of this section covers how proteins serve as the primary mediators of carbohydrate-based biological function thanks to their ability to bind to the diversity of carbohydrates with varying specificity[5]. These types of proteins include antibodies, enzymes, lectins, and otherwise general glycan-binding proteins (GBPs). Lectins are glycan-binding proteins that are often highly specific for particular carbohydrates (hence the name lectin, the Latin word for “select”). One example of how protein–glycan binding and recognition mediates biological function is the mechanism by which humans are infected by the influenza virus (*i.e.*, how we catch the flu). GBPs present on the surface of the virus are specific for α1-6 linked sialic acids, which are abundant on the surface of our cells. This protein–glycan binding event brings the virus close to the cell surface, facilitating its fusion with the cell membrane and the resulting infection[5]. The following section goes deeper into the biophysical mechanisms by which proteins bind to and recognize carbohydrates.

1.2 Mechanisms of protein–glycan binding and recognition

Experimental structural characterization provides residue-level resolution of protein–glycan complexes, enabling our understanding of the biophysical mechanisms behind carbohydrate binding and recognition. Structural models allow us to visualize many of the interactions that drive protein–glycan complex formation: van der Waals contacts, CH–π and electrostatic interactions, and hydrogen bonds (both direct and water mediated). Two important (and measurable) components of complex formation are affinity (*i.e.*, the strength of binding) and specificity (*i.e.*, the ability to bind the target of interest while discriminating against all others). I will provide the reader with additional description of hydrogen bonding given its foremost importance to both glycan binding affinity and specificity.

Hydroxyl groups can serve as both hydrogen bond acceptors and donors (specifically, they can donate one hydrogen bond and accept two). Carbohydrates have many hydroxyl groups, making them polar molecules that prefer to be solvated to maintain satisfaction of these hydrogen bond donor/acceptor groups. It is therefore crucial that hydrogen bonds are made at a protein–glycan interface to ensure sufficient binding affinity to drive complex formation. The directionality of the hydroxyl groups is another relevant component of protein–glycan complex formation. Depending on the given carbohydrate epimer, different hydroxyl groups are oriented in different directions. For example, in D-glucose the O3 hydroxyl is axial whereas it is equatorial in D-galactose. With the innate sensitivity of hydrogen bonding interactions to even slight distortions in geometry, proteins must not only have compatible amino acids at the interface to drive glycan binding affinity, but these amino acids must also be in the appropriate position and orientation to enable glycan binding specificity.

As stated in the beginning of this section, structural models allow us to visualize hydrogen bonding and other important biophysical interactions at a protein–glycan interface. Characterizing these interfacial interactions is crucial to understanding the mechanism behind a given protein’s glycan-binding affinity and specificity. Elucidating an experimental structure of a protein–glycan complex is, however, not an easy task. Computational techniques serve to fill this gap by providing a means of generating structural models of sufficient resolution in relatively short timeframes. In the next section, I provide a description of Rosetta, the computational “toolkit” I utilized extensively throughout my doctoral work and beyond.

1.3 Rosetta macromolecular modeling and design software suite

Rosetta is a comprehensive software suite for performing theoretical macromolecular modeling, structure and complex prediction, and design simulations[6]. Rosetta has been utilized time and time again to address diverse scientific challenges[7]. Virtually all of Rosetta’s modeling and design protocols aim to accomplish two overarching goals: given some biomolecular input, (1) quickly but broadly sample the structure and sequence space relevant to one’s scientific question (*e.g.*, sampling a binding-competent conformation of a given glycan-binding protein), and (2) accurately distinguish the most relevant (and physically realistic) conformation(s) from the pool of all sampled conformations. The former goal is primarily accomplished through a Metropolis Monte Carlo-plus-minimization sampling approach, while the latter goal is accomplished by applying a scoring function with physical, empirical, and statistical terms that approximate the energy of a biomolecule given its current conformation.

1.3.1 Conformational sampling in Rosetta

Rosetta primarily uses internal coordinates to represent and manipulate (*i.e.*, sample) biomolecular structure. In each simulation, bond lengths are kept fixed, and the degrees-of-freedom (DoFs) sampled are limited to the relevant ϕ, ψ, ω, and χ dihedral angles of each residue (and to rigid-body transformations if there are two or more independent bodies in the system). In 2017, Labonte *et al.* introduced the first framework for residue-centric modeling of carbohydrates in Rosetta[8]. Carbohydrates are modeled using the same principle of limiting DoF sampling to the ϕ, ψ, and any ω glycosidic torsion angles, the many side-chain χ angles of the hydroxyl groups, and the χ angles of any chemical modifications. Users can, however, also enable sampling of carbohydrate ring conformations by including the internal ν angles in the DoFs. **Figure 1.3** compares the ϕ, ψ, ω, and χ DoFs of a peptide bond *versus* a glycosidic bond and depicts the carbohydrate ν angles that determine ring conformation.

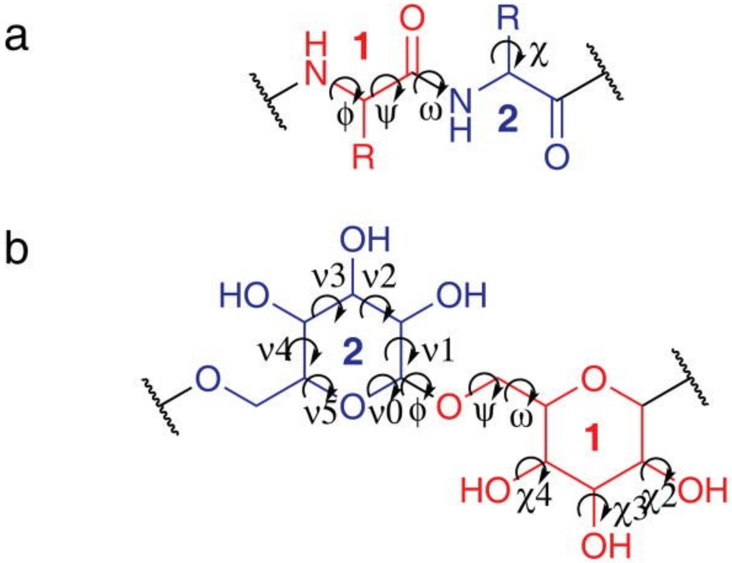

Figure 1.3: A comparison of the degrees-of-freedom (DoFs) found in polypeptide (A) and polysaccharide (B) chains. The first and second residue are labeled and colored red and blue, respectively. Torsion angles are indicated by arrows and labeled. Reprinted with permission from Labonte JW *et al.*, *J Comput. Chem.*, 2017.

The DoFs of a biomolecular structure are sampled in Rosetta using the Metropolis Monte Carlo-plus-minimization approach (MMCM)[9]. In MMCM sampling, a subset of DoFs is randomly perturbed (in Rosetta, this is referred to as a "move"), the DoFs of the entire system are energetically minimized, and the changes to the system are accepted if the system's energy is lower (*i.e.*, more negative; better) than it was previously. If the energy of the system is higher (*i.e.*, more positive; worse), then the Metropolis criterion is applied. In the Metropolis criterion, the probability of accepting the system's new conformational state is sampled from a Boltzmann distribution ($Probability_{new\ state} = \frac{E_{new\ state} - E_{old\ state}}{kT}$ where k = Boltzmann constant, T = temperature, E = energy). If $Probability_{new\ state} \geq U[0,1]$ (where $U[0,1]$ is a uniform random number between 0 and 1, inclusive), then the "move" is accepted; otherwise, the "move" is rejected and the system returns to its previous conformation. The MMCM steps are then repeated a pre-specified number of times.

Rosetta's modeling philosophy follows Anfinsen's dogma–that the native conformation of a biomolecular system (*i.e.*, the most biologically relevant conformation) is a unique and stable conformation accessible at the global minimum of the energy landscape[10]. In computational simulations, the energy landscape is determined by the system's DoFs and any modeled environmental conditions (*e.g.*, temperature, pH, solvation). Given the ruggedness of any biomolecular system's energy landscape, often this global energy minimum conformation is "trapped" behind various high-energy conformations. The MMCM approach allows for the occasional acceptance of "moves" that make the energy of the system higher, thus enabling better traversal of the energy landscape to find the low-energy, native conformation of interest.

1.3.2 The Rosetta scoring function

The Rosetta scoring function consists of multiple physical, empirical, and statistical terms that approximate the energy of a biomolecular system (*e.g.*, a protein, a carbohydrate, a protein–carbohydrate complex) in a given conformation. The Rosetta scoring function can be represented in the functional form $E_{total} = \sum w_i \, E_i(DoF, aa)$, where the total energy of the system (E_{total}) is the sum of each energy term (E_i) calculated as a function of the system's degrees of freedom (DoF) and chemical identities (aa) and scaled by a pre-determined weight (w_i).

Throughout my doctoral work, I solely employed the default (as of post-2016) Rosetta scoring function called $\mathrm{REF2015}$[11]. The energy terms of the $\mathrm{REF2015}$ scoring function cover (1) physical laws such as Lennard–Jones attraction and repulsion, solvation (addressed implicitly), and Coulombic electrostatics, (2) empirical observations such as the orientation and geometric dependence of hydrogen bonding and disulfide bonds, and (3) statistical potentials for amino acid identities (given the backbone dihedral angles), backbone dihedral angles (given the amino acid identity), side-chain rotamers (given backbone dihedral angles), penalizing non-planar backbone ω dihedral angles (given *cis* or *trans* ω configuration), penalizing the open conformation of proline rings, penalizing non-planar conformations of the side-chain hydroxyl group of tyrosine, and reference energies for the 20 canonical amino acid types. Rosetta's energy terms are captured as one- and two-body components, allowing the calculation of the system's total Rosetta energy to be relatively fast and intuitively decomposable.

As of 2017 and upon specification by the user at the command line, any Rosetta scoring function can include a statistical energy term that captures the preferences of the ϕ and ψ angles of glycosidic linkages[8,12,13]. The ϕ angle preferences depend on the stereochemistry (*i.e.*, α or β) of the anomeric carbon through which the glycosidic bond is made, and the ψ angle preferences depend on whether the connecting oxygen is axial or equatorial. This glycosidic linkage-based scoring term (called "sugar_bb") ensures Rosetta preferably samples and favorably scores biologically relevant glycan conformations.

1.4 Dissertation Outline

Prior to my doctoral work, there was no established method in Rosetta to predict the bound conformation of protein–carbohydrate complexes. Meaning whenever elucidating an experimental structure of a protein–carbohydrate complex was challenging or infeasible, researchers could not utilize Rosetta to generate a structural model to address that gap in important information. **Therefore, the primary goal of my doctoral work was to develop and benchmark a Rosetta-based tool for protein–carbohydrate modeling, and accordingly apply this protein–carbohydrate modeling tool on an interesting and practical use case.**

Chapter 1 introduced readers to the general scientific concepts behind my doctoral work. In Chapter 2, I describe my development of a Rosetta protocol for modeling protein–carbohydrate complexes. I also describe how I evaluated the protocol's effectiveness in sampling and identifying native-like bound conformations. In Chapter 3, I detail my efforts modeling two enzymes that in conjunction covert A-type whole blood to O-type. My analyses resulted in enzyme–carbohydrate blood antigen models that facilitated understanding of the enzymatic mechanisms behind this conversion and will inform future protein engineering experiments. Finally, in Chapter 4 I summarize and map back my contributions to the field of computational protein–carbohydrate modeling. I also describe remaining challenges in the field and provide a roadmap for brave, future researchers to follow in their quest to continue and expand upon my doctoral work.

Chapter 2

2. GlycanDock

Adapted from Nance ML, Labonte JW, Adolf-Bryfogle, J, and Gray JJ, “Development and evaluation of GlycanDock: a protein–glycoligand docking algorithm in Rosetta” *J. Phys. Chem. B* 2021, 125, 25, 6807–6820, with permission from the publisher. Full text including all Supplemental Information referred to within this chapter is available online at https://pubs.acs.org/doi/10.1021/acs.jpcb.1c00910. In this Chapter, all material that is stated to be available online can be found at this web address.

2.1 Introduction

Carbohydrates are the most abundant and diverse biomolecules found on Earth[14,15]. Finite chains of carbohydrates known as glycans play numerous functional roles in all three domains of life[3,16–21] as well as viruses[22,23]. Three-dimensional structures of protein–glycan complexes provide insight into how carbohydrates are recognized by proteins and mediate biological functions. For example, extensive structural analysis of lectins (carbohydrate-binding proteins) via X-ray crystallography uncovered the role of carbohydrate recognition in cell–cell interactions and pathogenic invasion[24–27]. The Protein Databank (PDB) serves as a global repository for experimentally determined three-dimensional structures[28]. Recent estimates indicate that entries containing carbohydrates make up less than 10% of the PDB[29]—of which only a few thousand represent a high-quality, true protein–glycoligand complex (a non-covalently attached glycan bound to a protein receptor)[30]. Consequently, resolved structures of protein–glycoligand complexes are relatively underrepresented compared to their ubiquity in biology. Despite technological improvements in experimental structural glycobiology[31,32], the innate flexibility and chemical heterogeneity of carbohydrate chains continues to hinder high-throughput collection of high-quality structures[33–36]. Therefore, computational docking tools that accurately predict the conformation and interfacial interactions of protein–glycoligand complexes are needed to fill the gap in structural characterization and enable further scientific and engineering advancements[37].

Computational simulations have long demonstrated utility in supplementing and deciphering experimental data on the structure of glycans and protein–glycan complexes[34–36,38–47]. Molecular dynamics (MD) simulations in particular are able to sample oligosaccharide conformations that are consistent with experimental data[48,49] and estimate the binding free energy of protein–glycoligand complexes[34,50–52]. However, MD becomes too costly when simulating large systems with many atoms and degrees of freedom[53], making faster (though less rigorous) computational tools for docking more practical. Protein–ligand docking software including AutoDock[54], AutoDock Vina[55], DOCK[56], FlexX[57], Glide[58], and GOLD[59] have all been applied to protein–glycoligand systems[60–65]. However, these tools work best on rigid, small-molecule (*i.e.* drug-like) ligands with few rotatable bonds. Accordingly, modeling and docking tools that account for the size and flexibility of glycoligands, such as AutoDock Vina-Carb[12,13] and the fragment-based approach developed by Samsonov and colleagues[66], are necessary.

The Rosetta macromolecular modeling and design software suite[6] has been used to address diverse scientific challenges[7,67,68]. Rosetta's protein–small molecule docking algorithm *RosettaLigand*[69,70] has been applied to a protein–glycoligand system to capture the effects of mutations on glycoligand binding energetics[71]. However, like other protein–ligand docking software, *RosettaLigand* is only able to treat a ligand as a single residue with discrete, pre-computed conformations—not as flexible oligomers. The recently developed *RosettaCarbohydrate* framework enabled modeling and design of glycans and glycoconjugate systems in a residue-centric (*i.e.*, oligomeric) approach[8]. To that end, we sought to develop a docking refinement algorithm that leverages the *RosettaCarbohydrate* framework and rapid conformational sampling and optimization techniques to predict native-like, biophysically accurate models of protein–glycoligand complexes.

Here, we introduce *GlycanDock*—a residue-centric protein–glycoligand docking refinement algorithm available within the Rosetta software suite. In this work, we assess *GlycanDock*'s ability to sample and discriminate bound, native-like conformations of protein–glycoligand complexes using a benchmark target set of 109 high-resolution structures from the PDB. Targets represent protein binders of broad scientific interest, including 11 antibodies, 33 lectins, 22 enzymes, 24 carbohydrate-binding modules, and 19 viral glycan binders. These 109 proteins are bound to glycoligands of various lengths, including 24 di-, 32 tri-, 28 tetra-, 14 penta-, 7 hexa-, 3 heptasaccharides, and 1 undecasaccharide. 81 are linear oligosaccharides and 28 have one or more branched connections. 17 have one or more exocyclic linkages. We also use 62 experimentally determined unbound protein structures to evaluate the effect of pre-configuration of the protein backbone on docking performance. To assess whether *GlycanDock* captures high-resolution structural details, we examine the counts and recovery of native-like biophysical features such as interfacial residue–residue contacts and hydrogen bonds. As a case study, we probe *GlycanDock*'s ability to recapitulate known glycoligand binding preferences of a carbohydrate-binding module (*Ct*CBM6). Finally, we report on the results of a pipeline for performing "blind" glycoligand docking when only the unbound protein structure and glycoligand sequence are known. The results of the benchmark assessment presented in this work demonstrate the effectiveness and overall utility of the *GlycanDock* protein–glycoligand docking refinement algorithm.

2.2 Materials and Methods

2.2.1 GlycanDock: the Rosetta protein–glycoligand docking refinement algorithm

GlycanDock is a Monte Carlo-plus-minimization docking refinement algorithm that features a high-resolution (all-atom) sampling and refinement strategy to locally optimize a glycoligand's conformation within a putative protein receptor pocket. *GlycanDock*'s sampling algorithm leverages data mined from the Protein Data Bank (PDB)[28] and extracted from quantum-mechanics calculations[12,13] to ensure carbohydrate-specific degrees of freedom (DoFs) fall within energetically-favorable, native-like conformational space[8]. The output of a *GlycanDock* trajectory includes the coordinates of the predicted model in PDB format and a breakdown of the model's total Rosetta score written to a Rosetta score file. The score file additionally reports some of the docking performance metrics described in this work, such as interface score, ring-RMSD, and ring-SRMSD.

2.2.1.1 Initial protein–glycoligand complex for use as input to GlycanDock

The *GlycanDock* algorithm requires a pre-packed (see Stage 0 below) putative protein–glycoligand complex as input, where the protein receptor and glycoligand each have their own unique chain identifiers (*e.g.*, protein chain A and glycoligand chain X). *GlycanDock* performs local docking only, meaning the input structure must have the glycoligand physically placed within the predicted the binding site and, for larger glycoligands, placed in approximately the correct rigid-body orientation. It is assumed that the protein receptor backbone is approximately correct, as it is kept fixed throughout the docking trajectory. In contrast, the initial glycosidic torsion angles can be arbitrary, as they are sampled and energetically minimized during docking; however, it is recommended that the initial glycoligand conformation provided is low energy. Protein side-chain rotamers and carbohydrate side-chain rotamers (*e.g.*, hydroxyl and *N*-acetyl groups) at the protein–glycoligand interface are optimized throughout the docking trajectory to minimize clashes while searching for productive interfacial interactions. By default, carbohydrate ring conformations are not sampled, but may be included as a DoF using the command-line interface. Additional information describing the preparation of input structures is detailed in "Benchmarking and evaluation of the *GlycanDock* algorithm".

2.2.1.2 Stage 0: Pre-packing the initial, putative protein–glycoligand complex

In Stage 0 of the *GlycanDock* algorithm, protein and carbohydrate side chains are pre-packed[72] to ensure compatibility of the input complex with the Rosetta scoring function. This is performed by removing any internal clashes within the initial structure and thus establishing a low-energy conformation at non-interface regions of the protein receptor.

In the case of the bound targets employed in this study, pre-packing additionally serves to erase the pre-configuration of the protein side chains at the interface to bind the glycoligand, thus reducing bias during docking. Hydrogen atoms are added, the glycoligand is separated by 1,000 Å from the protein receptor, and all non-disulfide bridge side-chain conformations are optimized by rotamer (*i.e.*, packed) and energy-minimized. The glycoligand is then translated back to its starting position. The Stage 0 pre-packing procedure should be performed on the initial, putative protein–glycoligand complex prior to running the *GlycanDock* docking algorithm. Pre-packing can be performed using the following example command-line flags:

```
/Rosetta/main/source/bin/./GlycanDock.linuxgccrelease -database
/Rosetta/main/database -include_sugars -alternate_3_letter_codes
pdb_sugar -auto_detect_glycan_connections -in:file:s target.pdb -
in:file:native crystal.pdb -nstruct 1 -ex1 -ex2 -ex3 -ex4 -ex1aro -
ex2aro -docking:partners A_X -out:pdb_gz -
carbohydrates:glycan_dock:prepack_only true
```

Here, `-docking:partners  A_X` informs the GlycanDock algorithm that the upstream protein receptor is identified by chain A and the downstream glycoligand is chain X.

2.2.1.3 Stage 1: Initialize docking trajectory by applying random perturbations to the input glycoligand configuration

Each *GlycanDock* trajectory begins by applying a small, random perturbation to the glycoligand's rigid-body orientation and to all glycosidic torsion angles. The objective of Stage 1 is to increase glycoligand sampling coverage within the protein binding site by promoting additional conformational diversity to the input putative complex. A Gaussian translational perturbation centered around 0.5 Å and a rotational perturbation centered around 7.5° is applied to the center-of-mass of the glycoligand. A uniform perturbation of ± 12.5° is applied to each glycosidic torsion angle. We note that many Rosetta docking protocols typically employ an initial low-resolution (centroid) search stage in lieu of the Stage 1 initial perturbation procedure described here, but the functionality required to model carbohydrates as centroid representations has not yet been incorporated into the *RosettaCarbohydrate* framework[8] at the time of writing.

2.2.1.4 Stage 2: Docking and refinement of the input protein–glycoligand complex

Stage 2 of the *GlycanDock* algorithm focuses on exploring the local conformational space of the glycoligand through rigid-body and glycosidic torsion angle sampling and refinement. This stage consists of two sets of eight inner cycles of Monte Carlo sampling and optimization of the glycoligand conformation at the putative binding site. The inner refinement cycles are wrapped by ten outer cycles that ramp the weights of the attractive and repulsive terms in the Rosetta scoring function, similar to the approach taken in the Rosetta *FlexPepDock* protein–peptide docking algorithm[72,73]. In the first outer cycle, the weight of the repulsive Lennard–Jones term (`fa_rep`) is reduced to 45% of its default magnitude, and the attractive van der Waals term (`fa_atr`) is increased by 325%. The weights are returned to their original magnitudes incrementally over the course of the proceeding outer cycles so that the final outer cycle uses the starting weights for these two score terms.

The inner cycles perform the sampling and optimization procedures on the glycoligand. The inner cycles consist of a set of eight rigid-body perturbations and a set of eight glycosidic torsion angle perturbations (performed in either order every inner cycle). Every perturbation is followed by interfacial side-chain rotamer optimization (packing), and every other perturbation is followed by full-structure energy minimization. Rigid-body sampling consists of uniform perturbations to the glycoligand's center-of-mass as well as occasional translation of the glycoligand toward the protein receptor's center-of-mass. This latter "sliding" step is occasionally necessary when clashes cause large gradients that during minimization jump the glycoligand far away from the protein. Glycosidic-linkage sampling includes performing uniform and non-uniform perturbations of various magnitudes on randomly selected glycosidic torsion angles. Sampling may also include occasionally flipping an entire carbohydrate ring around with respect to the rest of the carbohydrate chain (without changing the internal conformation of the carbohydrate ring itself) for glycoligands that satisfy specific glycosidic dihedral topology requirements (see Supplemental). Further details on the sampling and optimization procedures performed in the *GlycanDock* algorithm are available in the Supplementary Information.

GlycanDock trajectory-level information, such as the number of inner sampling and optimization cycles performed and accepted, are reported at the bottom of the output structure file.

2.2.1.5 Command-line usage of GlycanDock

GlycanDock-specific flags are described in the Table S1 (online[74]). An example of the flags used for the benchmark assessment of *GlycanDock* is shown below:

```
./GlycanDock.macosclangrelease -include_sugars -maintain_links
-in:file:s target-prepacked.pdb -in:file:native crystal.pdb
-cst_fa_file interface.cst -nstruct 50 -n_cycles 10 -ex1 -ex2
-docking:partners A_X  -out:pdb_gz
```

Here, `-maintain_links` is used rather than `-auto_detect_glycan_connections` because the input structure (`target-prepacked.pdb`) has already been processed by Rosetta and therefore has the appropriate `LINK` records defining the glycoligand's carbohydrate connectivity[8,75].

2.2.2 Benchmarking and evaluation of the *GlycanDock* algorithm

2.2.2.1 Metrics for evaluation of model accuracy and ranking of models

The ring-RMSD metric is used to evaluate the structural accuracy of *GlycanDock* models. Ring-RMSD is the root-mean-squared deviation (RMSD) of all ring atoms of the glycoligand in its predicted conformation to its native bound conformation after superposition of the protein receptor onto the native protein backbone. Ring-RMSD captures both the deviation in the shape and the orientation with respect to the binding site of the glycoligand. Models below 2 Å ring-RMSD were classified as a near-native (*i.e.*, sufficiently representative of the native bound conformation). Ring-SRMSD was also calculated to evaluate the structural accuracy of the shape of the glycoligand irrespective of its orientation in the binding site (*i.e.*, this metric can be calculated irrespective of the protein receptor). Ring-SRMSD is the RMSD of all ring atoms of the glycoligand after superposition of its predicted conformation onto its native bound conformation.

GlycanDock models were ranked by interface score. The interface score is calculated by taking the total Rosetta score (the weighted sum of all the terms in the Rosetta scoring function) of the model and subtracting the total score of the separated model where the glycoligand is translated 1,000 Å away from the protein receptor. The interface score approximates the binding free energy of the complex in units of REU (Rosetta Energy Units). The N_5 metric is used to quantify the effectiveness of *GlycanDock* sampling and the discriminatory power of the calculated interface score. N_5 is the count of near-native models ranked among the top-5-scoring of all predicted models for a given target.

2.2.2.2 Bootstrap statistical analysis to determine effective docking range

Bootstrap case resampling was used as described previously[76,77] to determine *GlycanDock*'s effective docking range. Briefly, for a given target, 5,000 sets of resampled models were generated by randomly selecting 1,000 models with replacement from the original set of *GlycanDock* models. The observed N_5 of each target from each randomly resampled set was then averaged and reported as $\langle N_5 \rangle$ (standard deviation $\sigma\langle N_5 \rangle$). Targets that resulted in $\langle N_5 \rangle \geq 1.0$ were considered a docking success. The effective docking range of *GlycanDock* was then defined as the maximum initial ring-RMSD from which 50% or more of tested protein–glycoligand targets achieved $\langle N_5 \rangle \geq 1.0$. Bootstrap case resampling was performed utilizing the pandas[78] data analysis tool (see Supplemental Information online[74] for pseudo-code example).

2.2.3 Analysis of biophysical features of *GlycanDock* models

2.2.3.1 Definition of the protein–glycoligand interface and biophysical features

A protein and carbohydrate residue are making an interfacial residue–residue contact if at least one non-hydrogen atom of a residue on one side of the interface (*e.g.* a protein residue) is within 5 Å of at least one non-hydrogen atom on the other side of the interface. A single protein or carbohydrate residue can make multiple unique interfacial residue–residue contacts (*e.g.* carbohydrate residue 1 contacts protein residue 12; carbohydrate residue 2 contacts protein residues 12 and 19). Interface residues are defined by the unique set of all residues making interfacial contacts (*e.g.* carbohydrate and protein residues 1, 2, 12, and 19 of the previous example are interfacial residues). The counts of interfacial residue–residue contacts and interface residues are reported at the bottom of the output structure file. In addition, the set of interface residues is reported as a PyMOL[79]-based residue selection at the bottom of the output structure file.

Hydrogen bonds are identified geometrically as those interactions contributing at least -0.5 energy units to the total "hbond" term of the Rosetta scoring function[11,80]. To be considered an interfacial hydrogen bond, a hydrogen bond must be between a carbohydrate residue of the glycoligand and a protein residue of the receptor. Interfacial hydrogen bond that pass the score cutoff are reported per carbohydrate residue in the form of a PyMOL-based residue selection at the bottom of the output structure file. The energetic filtering is performed by the HBondSelector at the end of a *GlycanDock* trajectory. Counting of interfacial protein–glycoligand hydrogen bonds is performed by parsing the corresponding PyMOL-based residue selections using an in-house Python script.

2.2.3.2 Measuring recovery of biophysical features

We consider an interfacial residue–residue contact or interfacial hydrogen bond recovered if the pair of interacting residues is the same pair observed in the native crystal complex. Similarly, an interfacial residue is recovered if it is present at the interface in the native crystal complex. Recovery is given as a fraction of the native biophysical feature recovered. Recovery ranges from 0.0 to 1.0, where 1.0 indicates complete native recovery. Recovery of interfacial residue–residue contacts and interfacial residues is calculated at the end of the *GlycanDock* trajectory and reported at the bottom of the output structure file. Recovery of interfacial hydrogen bonds is calculated using the corresponding data after parsing with the in-house Python script.

2.2.4 Local docking refinement of the crystal complex as a reference for *GlycanDock* performance

All bound protein–glycoligand crystal structures employed in this benchmark were subject to *GlycanDock* local docking refinement to serve as a reference for the measures of docking performance (*i.e.* N_5, $\langle N_5 \rangle$, and biophysical feature counts and recoveries). An example of the flags used to perform crystal refinement is shown below:

```
./GlycanDock.macosclangrelease -include_sugars -in:file:s crystal.pdb
-in:file:native crystal.pdb -cst_fa_file interface.cst -nstruct 50
-n_cycles 10 -out:pdb_gz -maintain_links -ex1 -ex2 -docking:partners A_X
-carbohydrates:glycan_dock:refine_only true
```

Here, the `-refine_only` flag makes the *GlycanDock* algorithm skip Stage 1 and apply only a modified version of Stage 2 in which smaller perturbations are made to the glycosidic torsion angles (see Supplementary Information online[74]).

2.2.5 Selection and preparation of benchmark set of protein–glycoligand complexes

A total of 109 experimentally determined bound protein–glycoligand structures were collected from the PDB[28] to create the bound target benchmark set. Thirty-three of these targets were selected from the AutoDock Vina-Carb benchmark set[13] while the rest were selected from protein–carbohydrate databases[30,81]. Some targets contain the same protein receptor sequence; however, no two proteins of the same sequence are bound to identical glycoligands. For example, *Streptococcus pneumoniae* endo-β-1,4-galactosidase binds three different glycoligands in PDB structures 2J1T, 2J1U, and 2J1V. We collected unbound protein structures for 62 of the bound targets to create the unbound target benchmark set. Unbound protein backbones were aligned onto the backbone of their corresponding bound protein structure. Only the coordinates of the aligned unbound protein and the glycoligand from the bound complex were kept. Alignment was performed using the `align` command in PyMOL and excluded hydrogens and non-protein atoms (`remove hydrogens; align <unbound> and !organic, <bound> and !organic`). Protein Cα-RMSD was calculated also using PyMOL (`align <unbound> and name CA, <bound> and name CA, cycles=0`). All benchmark targets were resolved using X-ray crystallography with a resolution of ≤ 2.0 Å. Further details on the selection and preparation of the bound and unbound protein–glycoligand benchmark sets can be found in the Supplementary Information online[74].

2.2.6 Generation of increasingly perturbed starting structures used as input to *GlycanDock*

After the preparation procedure described above, bound and unbound target structures were pre-packed (see Stage 0 of the *GlycanDock* algorithm). The glycoligand was then systematically perturbed in both rigid-body and glycosidic torsion angle conformational space to generate increasingly deviated input starting structures. The glycoligand's center-of-mass was perturbed using uniform translational perturbations of 0.25 Å, 0.5 Å, 1.0 Å, 2.0 Å, and 3.0 Å and uniform rotational perturbations of 3.75°, 7.5°, 15.0°, 30.0°, and 45.0°. Glycosidic torsion angles were perturbed using uniform perturbations of 6.25°, 15.0°, 30.0°, 60.0°, and 90.0°. Perturbed structures were binned on increasing magnitude of deviation measured using ring-RMSD (1.0 ± 0.1 Å, 2.0 ± 0.1 Å, up to 10.0 ± 0.1 Å ring-RMSD; Figure S1, online[74]). For unbound targets, ring-RMSD was calculated in reference to aligned starting structures. Ten perturbed starting structures per bound and unbound target for each ring-RMSD bin were generated. This process resulted in 10,900 perturbed starting structures for the bound benchmark set and 6,200 for the unbound.

2.2.7 Docking constraints employed during benchmarking

All *GlycanDock* benchmark docking trajectories employed a constraint that used a flat harmonic potential to bias the anomeric carbon atom of a specified carbohydrate residue to remain within a distance of 7.5 Å ± 2.5 Å to the Cα atom of any protein receptor residue.

Constraints were split evenly among the carbohydrate residues of the glycoligand to avoid bias to the known bound conformation. For example, a tetrasaccharide glycoligand would result in 25% of all *GlycanDock* models in which the first carbohydrate residue was constrained, 25% in which the second carbohydrate residue was constrained, *etc*. The biasing effect of the docking constraint is enforced *via* the Rosetta scoring function.

2.2.8 GlycanDock benchmark run time

A single *GlycanDock* protein–glycoligand docking trajectory resulting in one output model took on average 316 ± 154 seconds to complete (936 ± 292 seconds for targets containing neuraminic acid). A single *GlycanDock* receptor-free glycoligand conformational sampling trajectory resulting in one output model took on average 68 ± 41 seconds to complete.

2.2.9 Rosetta Technical Details

2.2.9.1 Modeling and sampling of carbohydrates in Rosetta

Carbohydrate oligomers and their DoFs (*e.g.*, main-chain and branch glycosidic torsion angles, internal ring torsions, side-chain torsions) are defined and modeled using the *RosettaCarbohydrate* framework[8]. In this benchmark, the internal ring torsion angles (ν) are held rigid across each docking experiment (*i.e.*, predicted models retain the same ν values as the input starting structures). Glycosidic torsions angles (ϕ, ψ, and, if present, ω) and carbohydrate side-chain torsions (χ) are sampled and optimized throughout the *GlycanDock* algorithm. See Labonte *et al.* for more information on carbohydrate modeling in Rosetta[8].

2.2.9.2 Rosetta scoring function and the glycosidic linkage scoring term

We employed the standard Rosetta Energy Function 2015 (`REF15`)[11] with an additional score term specific to the energetics of glycosidic torsion angles for this work. `REF15` is a scoring function that includes terms for physically derived potentials, such as van der Waals attraction and Lennard–Jones repulsion, Coulombic electrostatics, and a Gaussian exclusion implicit solvation term. It also includes empirical potentials such as orientation-dependent hydrogen-bonding terms and statistically derived terms to capture the energetic preferences of backbone and side-chain torsions in proteins. All scores are expressed as a unitless Rosetta Energy Unit (REU), with negative REU values representing favorable conformations.

The additional Rosetta score term used to capture the energetic preferences of glycosidic torsion angles, deemed `sugar_bb`, is derived from the quantum-mechanics-based Carbohydrate Intrinsic (CHI) energy functions[8,12,13]. The CHI energy functions capture the energetic preferences of ϕ and ψ glycosidic torsion angles between pyranose residues. It depends on the stereochemistry of the anomeric carbon and the upstream connecting oxygen atom and not the chemical identity of the carbohydrate residue.

Recently, the `sugar_bb` score term was expanded to include scoring of additional glycosidic linkage types. New parameters for the CHI energy function were added for the ψ torsion of α6 and β6 linkages. Previously, parameters were only available for ψ torsions of linkages that did not include exocyclic carbons. Additionally, a new energy function was added to represent the preferences for the ω torsions of glycosidic linkages by capturing the “gauche effect”. The “gauche effect” occurs when an ω torsion angle prefers one of the two gauche orientations, instead of the expected anti configuration, when the hydroxyl group of the carbon atom two carbons previous to the exocyclic carbon is equatorial[36]. For example, the preferred ω angle for a residue attached in a (1→6) linkage to glucose, where O4 is equatorial, is not 180° as might be expected but rather 60° or −60°. The new energy function is essentially a set of three harmonic energy wells centered over the gauche and anti torsion angles. It effectively adds a scoring penalty to any glycosidic conformation that does not demonstrate this “gauche effect”.

In this benchmark assessment, the scoring function used included the `REF15` score terms and weights, the updated `sugar_bb` score term with a weight of 0.5, and the `fa_intra_rep_nonprotein` score term with a weight of 0.55.

2.2.9.3 Rosetta version number and documentation

The *GlycanDock* algorithm is available as of version 61659 (weekly release #283) of the Rosetta macromolecular modeling and design software suite. See the online documentation for more information:

https://rosettacommons.org/docs/latest/application_documentation/carbohydrates/GlycanDock

2.3 Results

Figure 2.1 outlines the Rosetta *GlycanDock* Monte Carlo-plus-Minimization (MCM) algorithm for docking flexible glycoligands to protein receptors. Briefly, the *GlycanDock* algorithm takes a pre-packed (Stage 0) putative protein–glycoligand complex as an input structure. During Stage 1 of *GlycanDock*, the initial glycoligand conformation is randomly perturbed in both rigid-body and glycosidic torsion angle space. Stage 1 serves to promote conformational diversity in each independent docking trajectory; therefore, this initial, random perturbation is not subject to the Metropolis criterion. During Stage 2, a set of inner refinement cycles alternates between rigid-body and glycosidic torsion angle sampling followed by protein and carbohydrate side chain optimization at the interface and full-complex energy minimization. To promote thorough sampling of local conformational space, the inner refinement cycles are wrapped in a set of outer cycles that ramp down the van der Waals attractive weight and ramp up the Lennard–Jones repulsive weight of the scoring function[11]. Thus, the early cycles of Stage 2 refinement allow clashes and promote diversification, while later cycles enforce rigid sterics—a strategy shown to be effective in protein–peptide docking[72,73]. Stages 0, 1, and 2 of the *GlycanDock* algorithm are performed with implicit solvent in Rosetta's high-resolution, all-atom mode.

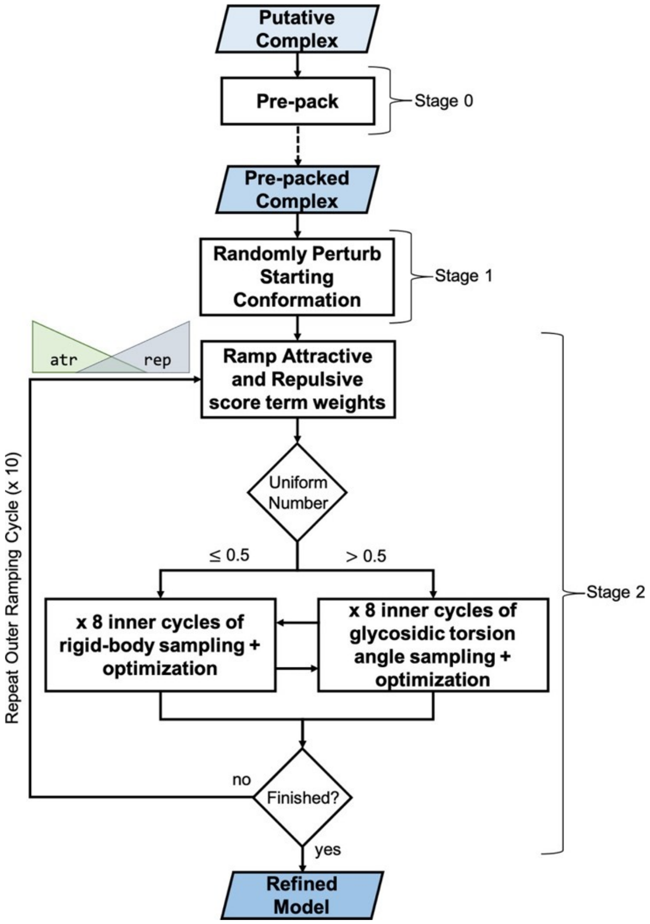
Putative Complex
Pre-pack
Stage 0
Pre-packed Complex
Randomly Perturb Starting Conformation
Stage 1
atr
rep
Ramp Attractive and Repulsive score term weights
Repeat Outer Ramping Cycle (x 10)
Uniform Number
≤ 0.5
> 0.5
x 8 inner cycles of rigid-body sampling + optimization
x 8 inner cycles of glycosidic torsion angle sampling + optimization
Stage 2
no
Finished?
yes
Refined Model

Figure 2.1: Overview of the GlycanDock algorithm. Stage 0 prepacks the initial, putative protein–glycoligand complex. The prepacked output structure is then given to the GlycanDock algorithm as input (indicated by the dashed arrow) to the sampling and optimization stages. Stage 1 applies a random perturbation to the glycoligand in rigid-body and glycosidic torsion angle space (without employing the Metropolis criterion). Stage 2 performs inner cycles of high-resolution rigid-body and glycosidic torsion angle sampling and optimization, where optimization includes packing and energy minimization at specific intervals. Outer cycles wrap the two sets of inner sampling cycles and control the incremental adjustment of the weights of the attractive (fa_atr) and repulsive (fa_rep) Rosetta score terms.

In this benchmark assessment, we apply an ambiguous atom-pair constraint to enforce one random carbohydrate residue of the glycoligand to remain physically close to the protein receptor throughout each independent docking trajectory. That is, this constraint ensures that the final set of models includes, for each carbohydrate residue, a portion of models where that residue contacts the protein. Bound protein–glycoligand models are ranked by a calculated interface score in REU (Rosetta Energy Units) that approximates the binding free energy of the complex. Model quality (*i.e.*, structural accuracy) is measured by calculating the root-mean-squared deviation (RMSD) of the heavy-atoms that compose each carbohydrate ring of the glycoligand in its predicted conformation compared to its native bound state after superposition of the protein receptor backbone (ring-RMSD). We consider models under 2 Å ring-RMSD (a standard model quality cutoff in the field of molecular docking[13,82]) to be sufficiently representative of the native bound conformation and are thus referred to as near-native models.

2.3.1 Determination of effective glycoligand docking range for bound and unbound protein backbones

The effectiveness of a local docking algorithm such as *GlycanDock* depends on the initial quality of the putative input complex. Raveh, London, and Schueler-Furman defined an algorithm's effective docking range as the maximum deviation of a given ligand from which near-native models can be sampled and correctly ranked[73]. To identify the effective docking range of the *GlycanDock* algorithm, we assessed docking performance on 109 bound and 62 unbound protein–glycoligand targets (Table S2, online[74]) of increasing initial deviation. We generated a benchmark set of starting structures by systematically perturbing the 109 bound and 62 unbound protein–glycoligand complexes and binning the resulting conformations based on increasing ring-RMSD (1–10 Å ring-RMSD; Figure S1, online[74]). We generated ten unique starting structures per target for each ring-RMSD bin to ensure diversity of input conformations. Prior to input to *GlycanDock*, all perturbed starting structures underwent an independent optimization of all side-chain rotamers. In the case of bound targets, this procedure erased pre-configuration of the interfacial side chains to bind the glycoligand. We then used *GlycanDock* to generate 2,000 models per target per ring-RMSD bin (10 input starting structures per target × 200 models each = 2,000 models). See Materials and Methods.

Figure 2.2 shows how *GlycanDock* sampled and discriminated near-native models of a branched xyloglucan oligomer bound to its receptor starting from input structures with initial glycoligand deviation of 2.0, 4.0, and 6.0 Å ring-RMSD (± 0.1 Å). For the 2.0 and 4.0 Å inputs (panels A and B, respectively), *GlycanDock* generated multiple low-scoring models similar to the experimental (native) crystal complex. However, in the 6.0 Å case (panel C), the top-5 lowest-scoring structures (blue diamonds) are scattered from 2.5–13 Å ring-RMSD from the native. These non-native-like models score worse (*i.e.*, more positive REU) than the refined native models (maroon), suggesting that the scoring function is sufficient to discriminate near-native models for this target, but the sampling failed to find those low-scoring conformations.

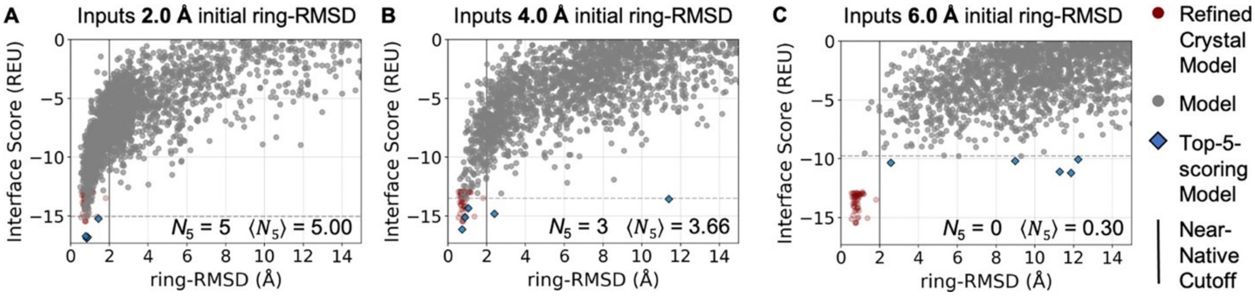

Figure 2.2: Example "funnel" plots depicting GlycanDock N5 and ⟨N5⟩ performance on bound target 4BJ0 at different initial ring-RMSD values. "Funnel" plots depict the relationship between interface score and ring-RMSD of a set of models (gray circles). While ⟨N5⟩ must be calculated, N5 can be determined directly by counting the number of near-native models (models below 2 Å ring-RMSD) within the top-5-scoring models (blue diamonds). (A) GlycanDock models predicted from 2.0 ± 0.1 Å initial ring-RMSD input structures demonstrate unambiguous (⟨N5⟩ = 5.00) funneling toward the refined native crystal structure (maroon circles). (B) Models from 4.0 ± 0.1 Å initial ring-RMSD input structures demonstrate acceptable funneling (⟨N5⟩ = 3.66) toward the refined native. (C) Models from 6.0 ± 0.1 Å initial ring-RMSD input structures demonstrate no funneling (⟨N5⟩ = 0.30) toward the refined native. Accordingly, the funnel plots in panels A and B demonstrate docking success while the funnel plot in panel C demonstrates docking failure.

From data like those shown in Figure 2.2, we quantify docking success using N_5—the count of near-native models ranked among the 5-top-scoring of all predicted models. Due to the stochastic nature of any MCM sampling algorithm, we performed a bootstrap statistical analysis to quantify the variability within each benchmark docking run[68,76]. For each bound and unbound protein–glycoligand target across all initial ring-RMSD bins, we performed bootstrap case resampling to calculate $\langle N_5 \rangle$—the bootstrap average of N_5 and a statistical measure of the reliability of observed docking success[76,77]. We defined $\langle N_5 \rangle \geq 1.0$ (*i.e.*, sampling and discriminating at least one near-native model among the 5-top-scoring with statistical reliability) as the threshold indicating docking success. For example, $\langle N_5 \rangle = 5.00$ and $\langle N_5 \rangle = 3.66$ for the two successful docking cases presented in Figure 2.2, whereas $\langle N_5 \rangle = 0.30$ for the failure case. Finally, we define the effective docking range of the *GlycanDock* algorithm as the maximum ring-RMSD from which 50% or more of the 109 bound and 62 unbound protein–glycoligand targets achieve $\langle N_5 \rangle \geq 1.0$.

***2.3.2 GlycanDock*'s effective docking range is 8 Å ring-RMSD for bound protein backbones and 7 Å for unbound**

Figure 2.3A summarizes *GlycanDock*'s $\langle N_5 \rangle$ performance on the bound benchmark set as a function of the ring-RMSD of the input structures. More than 50% of the 109 bound targets achieved $\langle N_5 \rangle \geq 1.0$ (green bars) up to 8 Å ring-RMSD, suggesting an effective docking range of 8 Å initial ring-RMSD with bound protein backbones. **Figure 2.3B** summarizes the same data for the unbound benchmark set (average protein Cα-RMSD to bound 0.49 Å ± 0.48 Å, minimum 0.05 Å, maximum 3.13 Å). More than 50% of the 62 unbound targets achieved $\langle N_5 \rangle \geq 1.0$ up to 7 Å ring-RMSD, suggesting an effective docking range of 7 Å initial ring-RMSD with unbound protein backbones. To illustrate docking success, **Figures 2.3C and D** depict the near-native conformations of a top-5- and top-1-scoring model, respectively, while **Figure 2.3E** depicts a top-1-scoring, sub-angstrom model. Table S3 (online[74]) reports all observed N5 and bootstrap ensemble averages ⟨N5⟩ and standard deviations (σ⟨N5⟩) for the bound benchmark set; Table S4 (online[74]) reports docking results for the unbound benchmark set.

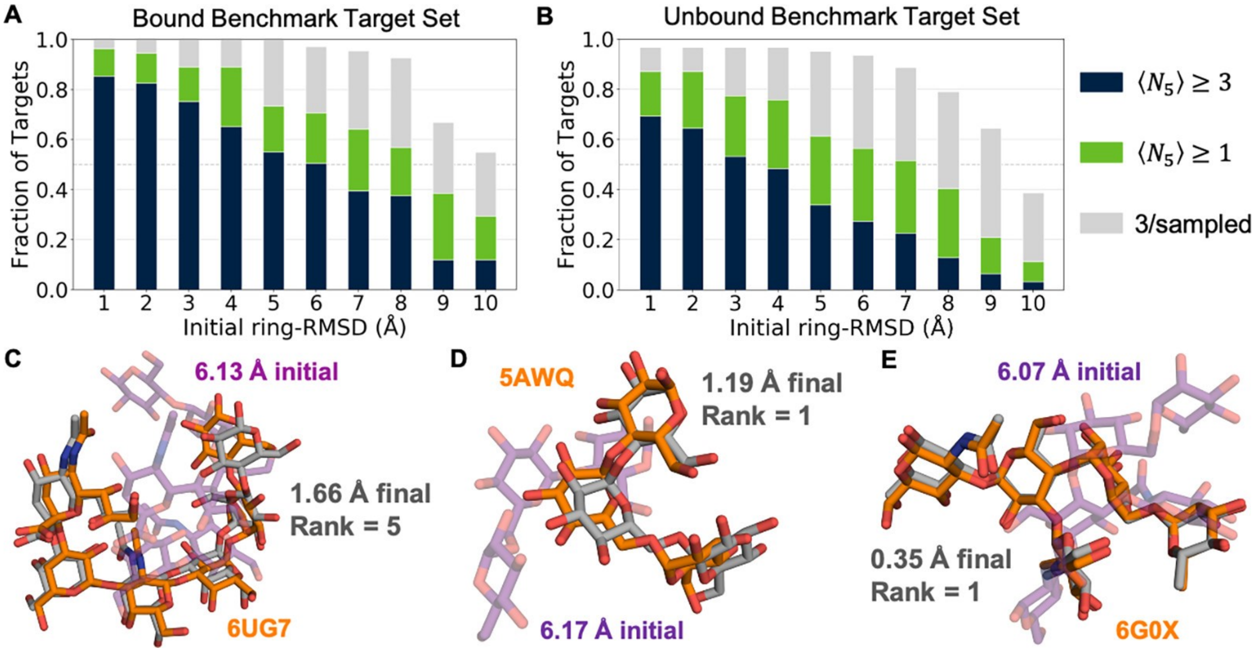

Figure 2.3: Summary of GlycanDock docking benchmark performance: (A) GlycanDock ⟨N5⟩ performance on 109 bound protein–glycoligand targets as a function of initial ring-RMSD bin. Blue bars represent the fraction of targets that achieved ⟨N5⟩ ≥ 3. Green bars represent the fraction of targets that achieved ⟨N5⟩ ≥ 1.0 (the threshold for docking success). Gray bars represent the fraction of targets that sampled at least 3 near-native models overall but failed to rank them among the 5-top-scoring. (B) Same as panel A, but on the 62 targets of the unbound benchmark set. (C–E) Glycoligand conformation from the bound crystal structure (orange) compared to conformation after the stage 1 random perturbation (purple, transparent) and the final conformation (gray) after GlycanDock sampling for three example targets. Protein backbones omitted for clarity.

While achieving $\langle N_5 \rangle \geq 1.0$ indicates docking success, the fraction of targets surpassing the $\langle N_5 \rangle \geq 3$ threshold (*i.e.*, sampling and discriminating at least three near-native models among the 5-top-scoring with statistical reliability) underscores the robustness of the *GlycanDock* algorithm. More than 50% of bound and unbound targets achieved $\langle N_5 \rangle \geq 3$ from input structures up to 5 Å and 3 Å initial ring-RMSD, respectively (Figure 2.3A and 2.3B, blue bars). For both the bound and unbound benchmark target set, we found no significant difference in the average count of near-native models among the 50-top-scoring models based on either the length of the glycoligand or whether the glycoligand was linear or branched (Figure S2, online[74]).

Docking failure cases ($\langle N_5 \rangle < 1.0$) can be either attributed to not having sampled near-native models (*i.e.*, sampling failure), or to the scoring function not correctly discriminating near-native models as top-scoring (*i.e.*, scoring failure). More than 50% of bound and unbound targets sampled three or more near-native models overall from inputs up to 10 Å and 9 Å initial ring-RMSD, respectively (Figure 2.3A and B, gray bars). Accordingly, most *GlycanDock* failure cases in this benchmark assessment can be attributed to scoring failures. Similarly, the greater fraction of unbound targets exhibiting scoring failures (gray bars) compared to bound targets highlights the sensitivity of the scoring function to changes in the protein backbone induced by glycoligand binding. While additional sampling (*e.g.*, generating more models or using more refinement cycles, re-refining top-scoring output models[73], using employing alternative protein receptor backbone conformations[77,83]) may be sufficient to overcome cases of sampling failure, faithful discrimination of near-native models depends on the efficacy of the scoring function.

2.3.3 Analysis of biophysical feature recovery of top-scoring *GlycanDock* models

Glycan-binding proteins have evolved a variety of sequences and structures to recognize the great diversity of carbohydrates, often with impressive selectivity[5,84]. For example, different influenza haemagglutinin subtypes can discriminate between α2,3- and α2,6-linked terminal sialic acids, which determines if a given influenza strain can infect animals or humans or both[85]. To understand how proteins selectively recognize the chemical and structural diversity of carbohydrate chains, the structural model must reveal key interfacial protein–glycoligand interactions. To this end, we examined the counts and recovery rates of protein–glycoligand interfacial residue–residue contacts and hydrogen bonds by the 50-top-scoring models per target from the bound benchmark set. Explicit water molecules (involved in water-mediated hydrogen bonding or otherwise) are not modeled and thus not considered in this analysis. See Materials and Methods.

The 50-top-scoring *GlycanDock* models exhibited a distribution of counts of interfacial residue–residue contacts similar to that of the 50-top-scoring refined crystal structures across the initial ring-RMSD bins examined (**Figure 2.4A**). However, as initial ring-RMSD increased, more of these contacts were made between non-native pairs of residues (*i.e.*, low contact recovery; **Figure 2.4B**). The distributions of the counts and recovery of interface residues followed a similar pattern (Figure S3, online[74]). Further, top-scoring models of targets bound to short glycoligands (*i.e.*, di- and trisaccharides) resulted in similar distributions (*i.e.*, shifted left toward lower counts as the initial ring-RMSD bin increased) compared to targets bound to long glycoligands (Figure S4, online[74]). On the other hand, top-scoring *GlycanDock* models did not make as many interfacial hydrogen bonds as compared to the refined crystal structures, with the recovery of native interfacial hydrogen bonds decreasing more drastically across increasing initial ring-RMSD bins (**Figure 2.4C & 2.4D**). Hydrogen bonds are more challenging to sample with high fidelity because they require precise atom-pair geometries to form[86]. For instance, native interfacial hydrogen bonding networks were difficult to identify and maintain even when the glycoligand was refined starting from its crystal conformation (Figure 2.4D, maroon).

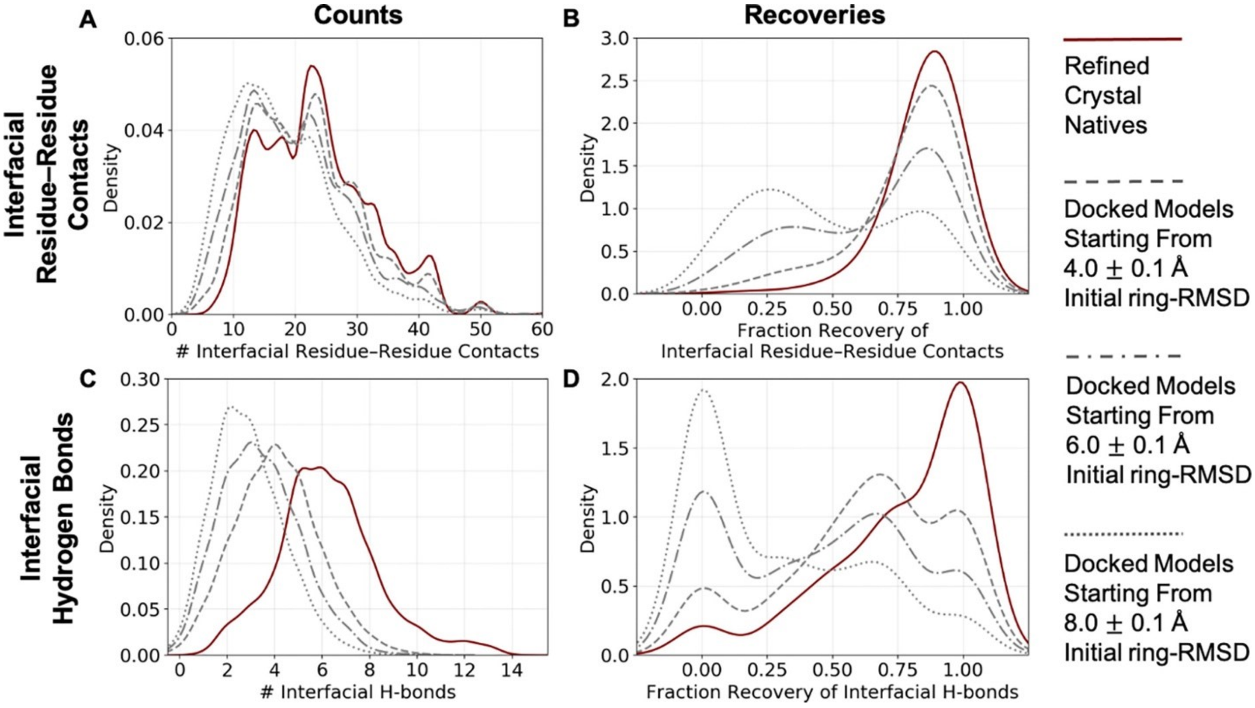

Figure 2.4: Counts and recovery of biophysical features by the 50-top-scoring GlycanDock models of each target from the bound benchmark set. (A) Distributions of the count of interfacial residue–residue contacts by the 50-top-scoring GlycanDock models per target predicted starting from input structures of 4.0, 6.0, and 8.0 Å initial ring-RMSD (gray dashed, dash–dot, and dotted lines, respectively). The distribution of the count of interfacial residue–residue contacts after GlycanDock crystal refinement (solid, maroon) serves as a reference. (B) Distributions of the recovery of native interfacial residue–residue contacts by the 50-top-scoring GlycanDock models per target. (C) Same as panel A, but for the count of interfacial hydrogen bonds. (D) Same as panel B, but for the recovery of interfacial hydrogen bonds. Discrete data are smoothed using kernel density fits using Seaborn[87] (kdeplot), resulting in some curves extending below fractions of 0.0 and above 1.0. Bin widths of 1.0 and 0.5 were used to fit the counts of interfacial residue–residue contacts and hydrogen bonds, respectively, and a bin width of 0.1 was used to fit the recoveries.

GlycanDock's top-scoring models sampled productive protein–glycoligand interfaces as measured by the counts of biophysical features such as interface residues and interfacial residue–residue contacts (Figures 2.4 and S3, online[74]). However, interfacial hydrogen bonds—whether seen in the native bound structure or otherwise—were especially difficult to make (Figure 2.4). This analysis provides a deeper evaluation of the *GlycanDock* algorithm's ability to sample biophysically realistic protein–glycoligand interfaces and highlights the challenge the Rosetta scoring function faces in correctly discriminating true, native-like interfaces. Detailed results and biophysical features for all of the 50-top-scoring models per target per initial ring-RMSD bin are available in Table S5 and S6 online[74].

2.3.4 GlycanDock Refinement Qualitatively Recapitulates Glycoligand Specificity of CtCBM6

Carbohydrate-binding modules (CBMs) are discretely folded, non-catalytic, sugar-binding proteins[88]. CBMs are typically found linked to carbohydrate-active enzymes, serving to enhance catalytic efficiency by binding to carbohydrate ligands and directing the enzyme to its substrate[89]. While some CBMs bind to a range of different carbohydrate ligands, others display distinct binding specificities[90]. For example, the CBM from xylanase 10A of *Clostridium thermocellum* (*Ct*CBM6) binds xylohexaose with a 100-fold higher affinity over cellohexaose[91]. A fast and accurate computational docking tool capable of discriminating glycoligand binders from non-binders would aid in the identification of key interfacial residues that inform protein design efforts to engineer new or improved binding behavior. Here, as a case study, we tested whether *GlycanDock* local docking refinement can capture the structural and energetic factors that determine the glycoligand specificity of *Ct*CBM6.

A synthetic *Ct*CBM6–cellopentaose starting structure was created by manually adding the atoms of glucose's exocyclic hydroxymethyl moiety to each carbohydrate residue of the native *Ct*CBM6–xylopentaose crystal structure (PDB 1UXX). A synthetic *Ct*CBM6^{U}–cellopentaose starting structure (where *Ct*CBM6^{U} distinguishes the unbound protein backbone, PDB 1GMM) was created using the same approach on the aligned unbound *Ct*CBM6^{U}–xylopentaose structure from the unbound benchmark target set. RMSD calculations for cellopentaose-bound models used the input starting structure as the reference structure, whereas xylopentaose-bound models used the native bound crystal structure (PDB 1UXX). We then applied the *GlycanDock* algorithm as described earlier on all four complexes, including the pre-packing in Stage 0 and the random perturbations in Stage 1. Despite these perturbations, we might expect the bound *Ct*CBM6–xylopentaose case to have some memory of the crystal interface (PDB 1UXX) that favors it enough to provide a lower score. But the unbound docking case (starting with PDB 1GMM) will not have this memory and will provide a balanced comparison when docking different substrates.

Figure 2.5A depicts the resulting funnel plot for the 50-top-scoring *Ct*CBM6–xylopentaose *GlycanDock* models (orange circles) *versus* the 50-top-scoring *Ct*CBM6^{U}–xylopentaose models (blue triangles) *versus* the 2,000 *Ct*CBM6–cellopentaose models (gray circles) *versus* the 2,000 *Ct*CBM6^{U}–cellopentaose models (gray triangles). As expected, the Rosetta scoring function clearly favored the native xylopentaose glycoligand docked to the bound conformation of *Ct*CBM6 (orange). More importantly, the docking of xylopentaose was favored in the *unbound* docking case (blue) over the 100-fold weaker cellopentaose binder docked to either conformation of *Ct*CBM6 (gray).

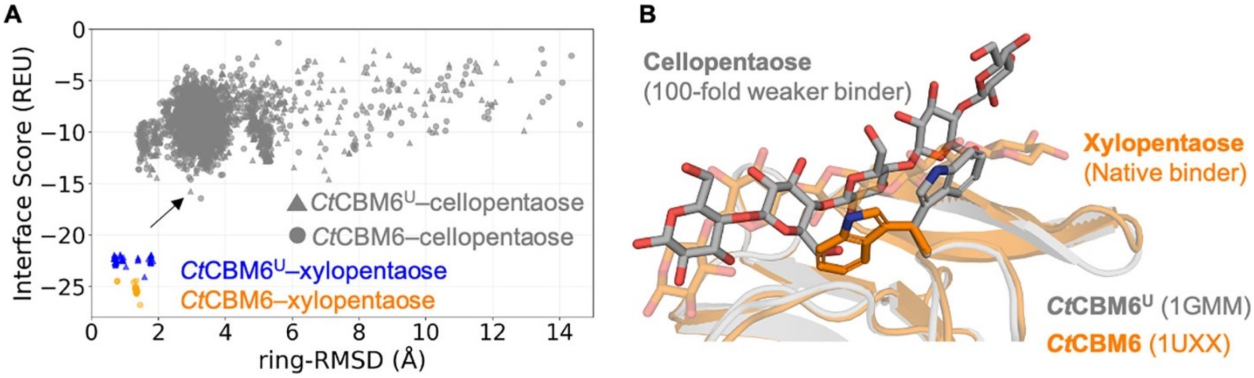

Figure 2.5: Results of GlycanDock local docking refinement qualitatively discriminate between a native glycoligand binder versus a 100-fold weaker binder to CtCBM6. (A) Funnel plots depicting results of GlycanDock local docking refinement of the native CtCBM6–xylopentaose complex (orange circles) versus CtCBM6U–xylopentaose (blue triangles, where CtCBM6U distinguishes the unbound protein backbone) versus CtCBM6–cellopentaose (gray circles) versus CtCBM6U–cellopentaose (gray triangles). The top-scoring CtCBM6U–cellopentaose model containing the 100-fold weaker cellopentaose binder is marked with an arrow. (B) Comparison of the conformation of the native bound CtCBM6–xylohexaose crystal structure (orange, transparent; PDB code 1UXX) to the top-scoring CtCBM6U–cellopentaose model marked in panel A (gray). The rearranged receptor tryptophan is shown in sticks.

It has been suggested that the striking difference in binding affinity between xylopentaose and cellopentaose is due to steric clashes with the distinguishing exocyclic hydroxymethyl moiety of glucose in two subsites of *Ct*CBM6's binding site[91]. The effect of these proposed steric clashes can be seen in **Figure 2.5B** where *GlycanDock*'s top-scoring *Ct*CBM6^{U}–cellopentaose model (gray) is unable to bury as deeply in the binding site as the native xylopentaose glycoligand (orange). Further, these steric clashes led to a rearrangement of a receptor tryptophan in the binding site, disrupting an important CH–π "stacking" interaction[92] with a carbohydrate residue of the glycoligand (Figure 2.5B). These results suggest that cellopentaose would not bind as well as xylopentaose; accordingly, the results of *GlycanDock* local docking refinement qualitatively match the experimentally determined glycoligand specificity of *Ct*CBM6.

2.3.5 Combination of FTMap and RosettaLigand Produces Putative Protein–Glycoligand Models within GlycanDock's Effective Docking Range Starting from "Blind"-like Inputs

In real molecular docking cases, the bound conformation of the target complex is unknown. In fully "blind" docking cases without experimental data, the location of the binding site on the protein receptor is also unknown. In this work, we demonstrated that the *GlycanDock* algorithm's effective glycoligand docking range is up to 7 Å initial ring-RMSD when using unbound protein backbones (Figure 2.3). Accordingly, *GlycanDock* requires as input a putative protein–glycoligand complex where the glycoligand is placed near the binding site and in approximately the correct orientation. When only the unbound protein structure and glycoligand sequence is known, an approach to effectively generate this putative complex is necessary. Here, we report the results of a pipeline for "blind" glycoligand docking using FTMap[93] and *RosettaLigand*[69,70]. We used FTMap to predict the glycoligand binding site and *RosettaLigand* to generate an initial protein–glycoligand structure. Details on the setup and usage of FTMap and *RosettaLigand* can be found in the Supplemental Information.

While various ligand binding site prediction software exist[94], including some specific to carbohydrate ligands[95,96], we chose FTMap for its speed and ease of use *via* its online webserver[97]. FTMap predicts ligand-binding "hot spots" by extensively sampling the surface of a macromolecular receptor using various small organic molecules as probes. Among the probes used is cyclohexane (CHX)—a compound structurally similar to that of a carbohydrate. We hypothesized that the FTMap "hot spots" predicted with a CHX probe would map to glycoligand binding sites. In 36 of the 62 unbound targets, the FTMap server resulted in a CHX probe within the known glycoligand binding site as determined by visual inspection (**Figure 2.6A**; see Table S2 online for list of the 36 targets). Accordingly, we used the 36 unbound protein structures and the Cartesian coordinate of the center-of-mass of the CHX probe to produce putative protein–glycoligand models using *RosettaLigand*.

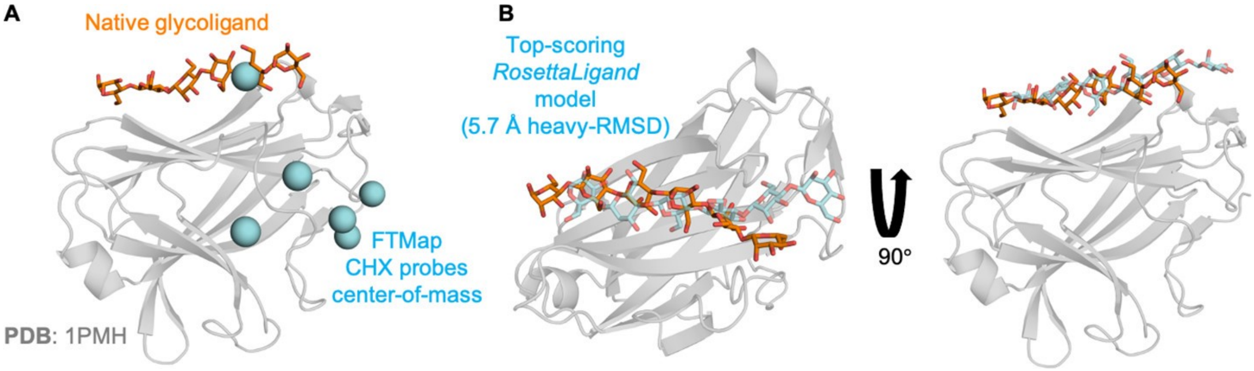

Figure 2.6: Example results of FTMap binding site prediction and RosettaLigand docking. (A) Results of FTMap ligand binding "hot spot" prediction using the unbound structure of CsCBM27-1 (PDB 1PMJ) as input. Cyan spheres represent the center-of-mass of each site prediction identified by the cyclohexane (CHX) probe. The native bound crystal structure (PDB 1PMH) shows that one carbohydrate residue of the glycoligand (orange sticks) overlaps directly on one of the FTMap CHX probe predictions. (B) A 5.7 Å heavy-RMSD RosettaLigand model ranked among the 5-top-scoring models. The same native bound crystal structure from panel A is shown. The predicted conformation of the glycoligand (transparent cyan sticks) overlaps reasonably well with the native conformation (orange sticks) in the binding site but is shifted over by one carbohydrate unit. This RosettaLigand predicted conformation (transparent cyan sticks) is within the effective docking range of the GlycanDock algorithm.

RosettaLigand requires pre-generated conformations of ligands that have multiple rotatable bonds. We used the Rosetta *GlycanSampler* algorithm (Jared Adolf-Bryfogle, unpublished) to generate 200 initial conformations for each of the 36 unbound target glycoligands (see Supplemental Information online[74] for details and for other methods of generating initial glycoligand conformations). *GlycanSampler* glycoligand models had an average heavy-SRMSD of 1.48 Å ± 0.65 Å (minimum 0.73 Å, maximum 4.98 Å), where heavy-SRMSD is calculated using all glycoligand heavy-atoms after superposition of the model glycoligand onto the native bound glycoligand. Using the center-of-mass of the CHX probe predicted within the known glycoligand binding site as the starting coordinates, we generated 2,000 docked models per target using *RosettaLigand*. Models were ranked by interface score, and glycoligand RMSD was calculated using all glycoligand heavy-atoms after alignment of the protein receptor (heavy-RMSD). *RosettaLigand* sampled one or more models below 7 Å heavy-RMSD within the 5-top-scoring models for 25 of the 36 unbound targets (**Figure 2.6B**). Three of the eleven failure targets were carbohydrate binding modules that resulted in top-scoring *RosettaLigand* models where the glycoligand was docked in the reverse direction compared to the native structure (*i.e.*, the non-reducing-end carbohydrate of the model aligned with the reducing end of the native structure). *RosettaLigand* sampled near-native models (here, below 2 Å heavy-RMSD) among the 5-top-scoring for only seven of the 36 unbound targets. Taken together, we have shown that the combination of FTMap and *RosettaLigand* can produce putative protein–glycoligand models within the effective docking range of the *GlycanDock* algorithm starting from "blind"-like docking conditions.

2.4 Discussion

We have developed and evaluated *GlycanDock* (Figure 2.1), a new, residue-centric protein–glycoligand docking refinement algorithm within the Rosetta macromolecular modeling and design software suite. *GlycanDock* treats carbohydrate chains as flexible oligomers, allowing for extensive conformational sampling of the glycoligand. Further, conformations with glycosidic linkages that fall within pre-determined, energetically-favorable torsion space are rewarded during sampling to ensure biophysically-realistic carbohydrate structures[8,12,13]. The *RosettaCarbohydrate* framework[8,75] enables the ability to handle both simple and complex glycoligands including variations in length, composition, ring shape, glycosidic connectivity, and branching, making *GlycanDock* a robust carbohydrate modeling tool. With continued efforts to improve the capabilities of the *RosettaCarbohydrate* framework and expand the Rosetta chemical database[6] with natural and non-natural monosaccharides and chemical modifications, *GlycanDock* will be able to simulate systems that represent the carbohydrate diversity observed in nature[98] and beyond.

In this work, we described the results of a benchmark assessment of the Rosetta *GlycanDock* protein–glycoligand docking refinement algorithm. We evaluated *GlycanDock* performance on 109 bound and 62 unbound protein–glycoligand targets using input structures of systematically increasing initial ring-RMSD. Docking performance was measured per target using N_5—the count of near-native models ranked among the five top-scoring models (Figure 2.2). We used bootstrap case resampling to calculate $\langle N_5 \rangle$, setting a threshold of $\langle N_5 \rangle \geq 1.0$ to define docking success. Bootstrap statistical analysis indicated that *GlycanDock*'s effective docking range is 8 Å ring-RMSD with bound protein backbones and 7 Å for unbound (Figure 2.3). Notably, benchmarking of *GlycanDock* included modeling side-chain flexibility at the protein–glycoligand interface and examination of glycoligand docking performance on 62 unbound protein structures (~58% of the 109 bound benchmark targets).

We sought to measure the *GlycanDock* algorithm's ability to recapitulate some of the biophysical features that drive glycoligand binding. Hydrogen bonding, for instance, is an important interaction at protein–glycoligand interfaces[36], but forming productive hydrogen bonds requires precise alignment of local atomic geometries[86]. Sampling these exact interactions can be challenging, as evidenced by the relatively poor interfacial hydrogen bonding recovery of top-scoring *GlycanDock* models (Figures 2.4C & D). Future work on the *RosettaCarbohydrate* framework will address additional carbohydrate modeling considerations such as CH–π stacking interactions[92,99–104] and water-mediated hydrogen bonding[105–107], which may improve *GlycanDock* sampling of native-like biophysical features at protein–glycoligand interfaces.

We found that the results of *GlycanDock* local docking refinement qualitatively corresponded with the experimentally determined binding specificity of *Ct*CBM6 to xylopentaose and cellopentaose glycoligands (Figure 2.5). While further study is needed, initial results indicate *GlycanDock* can be used to predict binder from non-binder glycoligands.

The performance of other tools for protein–glycoligand docking such as AutoDock Vina-Carb[12,13] and the fragment-based approach by Samsonov and colleagues[66] has also been published. However, both assessments employed different benchmark target sets and docking success definitions, making direct performance comparison difficult. Further, both AutoDock Vina-Carb and the method by Samsonov *et al.* include full rigid-body rotations and translations using a grid box sampling approach, whereas our *GlycanDock* docking refinement algorithm does not perform such extensive sampling of rigid-body space. Accordingly, we presented a possible "blind" docking pipeline that utilizes FTMap[93,97] and *RosettaLigand*[69,70] (the latter of which performs grid box sampling) to generate putative protein–glycoligand complexes that are within the effective docking range of the *GlycanDock* docking refinement algorithm (Figure 2.6). Models from this FTMap-*RosettaLigand* docking pipeline, AutoDock Vina-Carb, or the method by Samsonov *et al.* could then be refined by the *GlycanDock* algorithm for further evaluation. The scientific community therefore has a selection of useful tools to address a variety of modeling and prediction challenges in glycoscience research.

Protein–carbohydrate interactions modulate many cellular and molecular processes that are fundamental to all life. High-resolution models of protein–glycoligand complexes help us understand how proteins recognize carbohydrates and how glycan structure can bring about such diversity in observable function. While experimental limitations continue to hinder high-quality protein–glycoligand structure determination, computational modeling tools have served to fill this gap. We developed the GlycanDock docking refinement algorithm to model, dock, and refine protein–glycoligand complexes and serve as a tool to reveal the atomic details behind the molecular roles carbohydrates play. Also, GlycanDock refinement can be combined with the design algorithms of the Rosetta software suite. With the ubiquity of glycans in biology and biotechnology, the expanded suite of computational tools for glycans has the potential to aid in innovations in human health and disease, glycomimetic drug design, pathogen detection and defense, plant-based renewable bioenergy, and more.

Chapter 3

3. Carbohydrate Blood-antigen Enzyme Modeling and Design

3.1 Introduction

Carbohydrates play an integral role in cell–cell interactions and recognition. One key example of this role is the human ABO blood group system. The surface of our red blood cells (RBCs) is covered in glycoproteins and glycolipids containing carbohydrate chains that terminate in the different ABO carbohydrate blood antigens[108]. The A, B, AB, and O blood types are all differentiated by the identity of (or lack of) the terminal carbohydrate of the blood antigen structure (**Figure 3.1**).

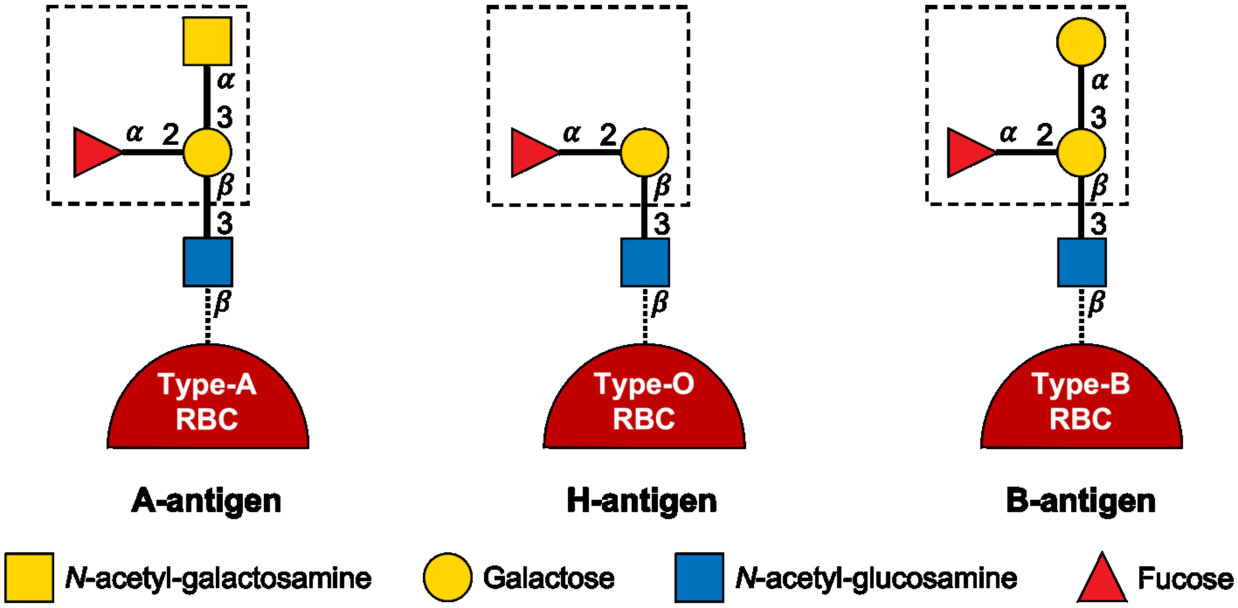

Figure 3.1: The human ABO carbohydrate blood antigens. Carbohydrates are depicted using the SNFG notation[109], and the red blood cell (RBC) surface is depicted as a red half circle. The AB blood type is defined by the presence of both the A- and B-antigens on the RBC surface.

In blood transfusions, it is critical to provide compatible blood types to avoid severe medical complications. For instance, individuals with Type A blood naturally make antibodies against the Type B blood antigen; meaning if a Type A patient were to receive Type B blood, those antibodies would activate a potentially lethal immune response. Providing Type O blood avoids this situation entirely as it contains neither the Type A nor B antigen (Figure 3.1) to activate that immune response, making Type O the “universal” blood group. Notably, patients must have the same or compatible Rhesus type (*e.g.*, Type O- *versus* O+), but the impact of Rhesus type is otherwise not discussed in this Chapter for clarity.

Given its “universal” nature, a sufficient supply of Type O blood in medical and emergency situations is essential, but unfortunately difficult to maintain. To address this limitation, Rahfeld *et al.* uncovered and characterized an enzymatic pathway in the human gut microbiome that converts Type A blood to the “universal” Type O[110,111]. They found two enzymes from *Flavonifractor plautii* (*Fp*) that catalyze this two-step conversion–an A type blood *N*-acetyl-alpha-D-galactosamine deacetylase (*Fp*GalNAcDeAc) and an A type blood alpha-D-galactosamine galactosaminidase (*Fp*GalNase) (**Figure 3.2**).

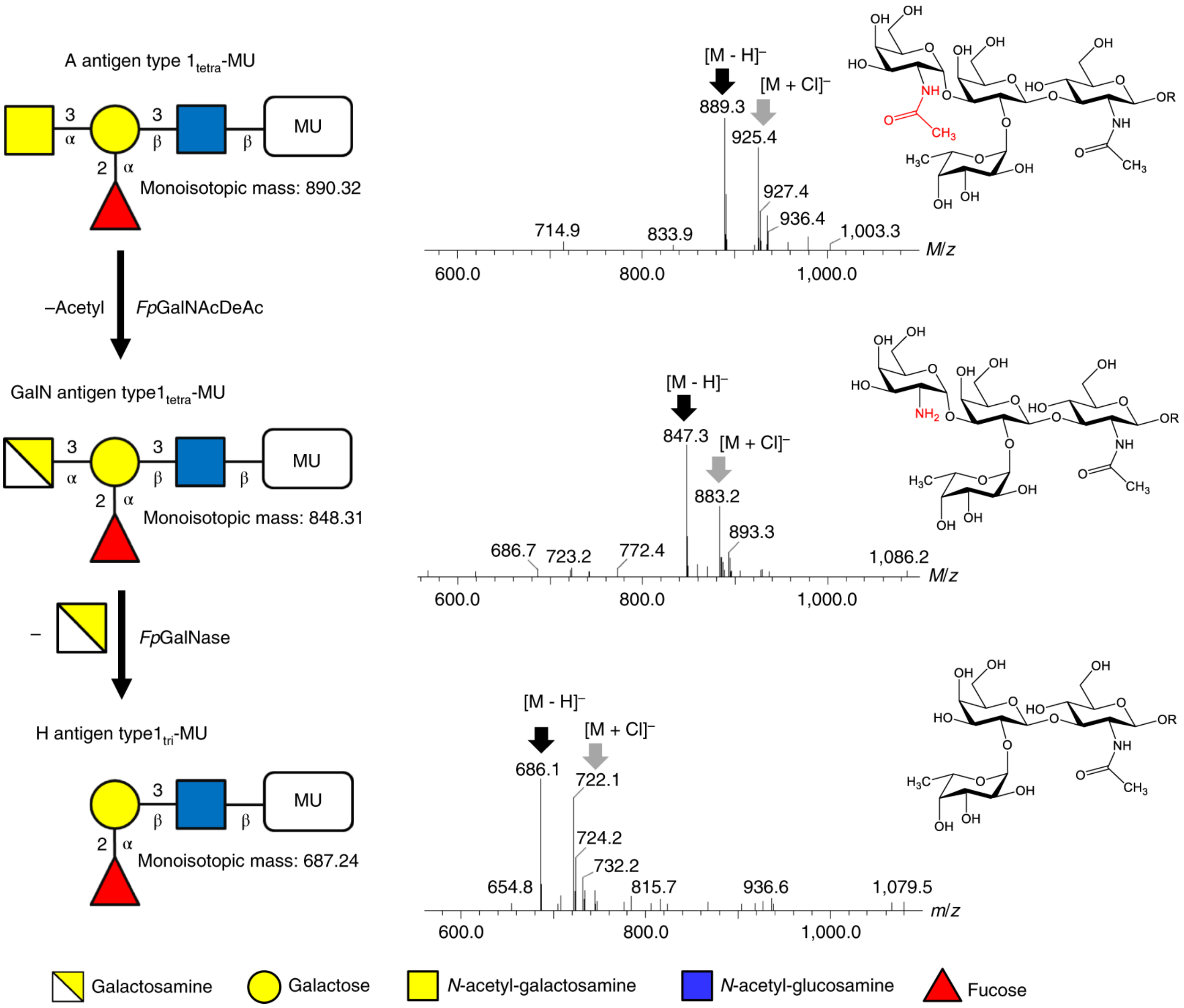

Figure 3.2: The A antigen type1tetra-MU is deacetylated by *Fp*GalNAcDeAc, followed by cleavage of galactosamine by *Fp*GalNase, yielding H antigen. Mass spectra show the mass loss of 42 on deacetylation by *Fp*GalNAcDeAc and 161 mass loss following cleavage of the galactosamine linkage. The black arrow indicates the deprotonated species and the grey arrow indicates the chloride adduct of the product. Sugars are presented as chemical structures (red-labelled part of the chemical structure is the functional group being converted by *Fp*GalNAcDeAc) and symbols using the SNFG notation[109]. Reprinted with permission from Rahfeld *et al. Nat. Microbiol.* 4, 1475–1485, 2019 (DOI: 10.1038/s41564-019-0469-7).

I wanted to develop a deeper understanding of how *Fp*GalNAcDeAc and *Fp*GalNase catalyze the conversion of A type blood to universal O type. This Chapter is therefore broken up into two sections: Section 1 focuses on my modeling and analyses of *Fp*GalNAcDeAc and Section 2 on *Fp*GalNase.

3.2 Section 1 – *Fp*GalNAcDeAc computational modeling and design

3.2.1 Goal 1A – Model the full-length *Fp*GalNAcDeAc enzyme structure

At the time of this study, only the deacetylase domain of *Fp*GalNAcDeAc had been experimentally determined (PDB IDs 6N1A and 6N1B). I wanted to generate a full-length structural model to inform relevant computational simulations (discussed in this Chapter) and future biochemical experiments (suggested throughout this Chapter). To accomplish this goal, I used AlphaFold to predict a structural model of all four domains of *Fp*GalNAcDeAc. I then used this *Fp*GalNAcDeAc model to conduct further computational studies including glycoligand docking (Goals 1B and 1C) and protein design (Goal 1D).

3.2.1.1 AlphaFold predicts the full-length FpGalNAcDeAc enzyme structure with high estimated accuracy

The full-length sequence of *Fp*GalNAcDeAc (772 residues, including the signal peptide) is published under UniProt accession number P0DTR4. From this sequence, I generated a full-length structural model of *Fp*GalNAcDeAc using AlphaFold[112]. *Fp*GalNAcDeAc consists of four distinct protein domains (as published in Rahfeld *et al.* 2019): the N-terminal catalytic deacetylase domain (referred to here on out as *Fp*DeAc), an invasin linker domain, a CBM32 domain (referred to here on out as *Fp*CBM32), and a C-terminal CBM of unknown family. During the time of writing this thesis, an interactive model of the full-length *Fp*GalNAcDeAc structure automatically generated by AlphaFold became available online through the AlphaFold Protein Structure Database at https://alphafold.ebi.ac.uk/entry/P0DTR4[113].

Figure 3.3 depicts the AlphaFold structural model of the full-length *Fp*GalNAcDeAc colored by protein domain. The mean pLDDT of the *Fp*GalNAcDeAc model when excluding the signal peptide is 95.48 (specifically, excluding residues 1–27). The mean pLDDT of the model when also excluding the *Fp*DeAc domain, which has two available crystal structures, is 94.07 (specifically, excluding residues 1–418). Accordingly, the predicted accuracy of the AlphaFold-predicted full-length *Fp*GalNAcDeAc structural model is on average very high. See Section 3.7 Detailed Methods for more information on pLDDT predicted confidence. While the two *Fp*DeAc crystal structures (PDB IDs 6N1A and 6N1B) would likely have been included in AlphaFold's training set, this fact does not appreciably affect my work presented here.

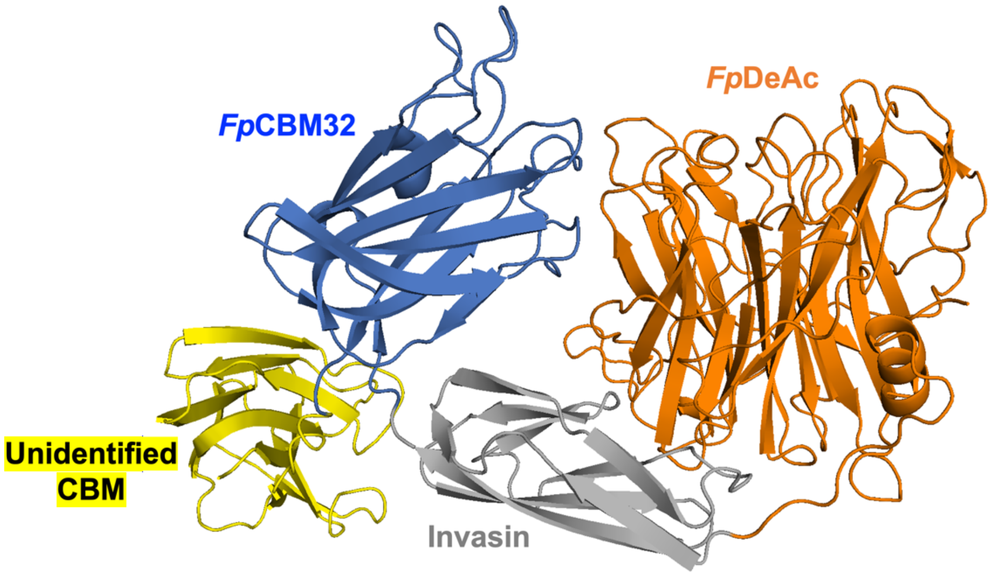

Figure 3.3: AlphaFold structural model of full-length *Fp*GalNAcDeAc colored by domain. The N-terminal *Fp*DeAc domain is in orange, the invasin linker domain is in gray, the *Fp*CBM32 domain is in blue, and the C-terminal CBM domain of unknown family is in yellow.

Readers may be interested in comparing *Fp*GalNAcDeAc to experimentally resolved structures of other carbohydrate-active enzymes with multiple domains, specifically, those with a catalytic domain and a carbohydrate-binding domain linked together *via* a rigid protein domain like invasin. Examples include bacterial sialidases such as PDB IDs 1EUT, 1W8N, and 2BZD.

3.2.1.2 AlphaFold structural modeling provides more specific residue boundaries for the four domains of FpGalNAcDeAc

As published in Rahfeld *et al*. 2019, *Fp*GalNAcDeAc is predicted to have a ~145-residue, C-terminal carbohydrate binding module (CBM) of family 32 (*i.e.*, *Fp*CBM32; UniProt P0DTR4 residues 504–648)[110]. InterPro (a sequence-based protein domain prediction software)[114] initially identified this as a much longer ~265-residue, C-terminal CBM32 domain (UniProt P0DTR4 residues 504–765; see Rahfeld *et al*. 2019 Supplementary

Figure 7[110]). The AlphaFold-predicted model of full length *Fp*GalNAcDeAc divided this ~265-residue region into two distinct domains: the originally identified *Fp*CBM32 domain and an additional CBM-like domain (see Figure 3.3). Sequence analysis of this additional CBM-like domain (UniProt P0DTR4 residues 658–765) failed to classify it into any previously reported CBM family (data not shown), meaning this domain is either a new family of CBM or is potentially not a CBM at all. **Future biochemical experiments should probe the role, if any, of this additional, C-terminal CBM-like domain.**

In summary, the AlphaFold predicted full length structural model of *Fp*GalNAcDeAc provides clearer residue-level boundaries for the four domains of this enzyme. **Reported as UniProt P0DTR4 residue numbers, the four domains of *Fp*GalNAcDeAc are as follows: residues 32–422 the catalytic deacetylase domain, residues 423–503 the invasin domain, residues 504–648 the CBM32, and residues 655–772 the unidentified CBM-like domain (Figure 3.3).** This residue-level information on the domain boundaries of *Fp*GalNAcDeAc should be helpful for guiding future truncation and biochemical analyses.

3.2.2 Goal 1B – Predict the bound conformation of the *Fp*CBM32–LacNAc complex

As discussed in the previous Section 3.2.1, *Fp*GalNAcDeAc contains a *Fp*CBM32 domain toward the C-terminus. CBM32s have predominantly been reported bind a variety of galactose-containing glycoligands such as lactose and *N*-acetyllactosamine (LacNAc)[115]. Rahfeld *et al*. reported glycan array data showing *Fp*CBM32 is specific to glycans with repeating LacNAc structures. The authors hypothesized that *Fp*CBM32 anchors the

*Fp*GalNAcDeAc enzyme to the red blood surface *via* non-competitive binding to LacNAc units (a repeating, terminal component of cell surface glycolipids)[110].

Given this interesting anchoring hypothesis, I wanted to generate a structural model of the *Fp*CBM32–LacNAc bound complex. To accomplish this, I first identified the relevant binding site residues of *Fp*CBM32 using the ligand binding site prediction server RaptorX[116] in combination with a structural comparison to an experimentally resolved CBM32–galactose complex. I then utilized a Rosetta all-atom refinement protocol to generate diverse, low-energy conformations of *Fp*CBM32. To improve glycoligand docking refinement outcomes, I filtered out *Fp*CBM32 conformations that were unlikely to be able to accommodate LacNAc at the predicted binding site using the P2Rank binding pocket prediction tool[117]. Finally, using these selected *Fp*CBM32 conformations, I used my GlycanDock protein–glycoligand docking refinement algorithm (Chapter 2) to generate *Fp*CBM32–LacNAc complex models.

3.2.2.1 Predicting the LacNAc binding site residues of FpCBM32 using RaptorX and structural comparison to an experimental structure of a homologous CBM32 bound to galactose

Given a protein sequence, RaptorX predicts which ligand(s) the protein binds (*e.g.*, calcium, glycerol, galactose) and through which residues this binding occurs (*e.g.*, galactose binding *via* protein residues 4, 13, 23, and 26)[116]. The RaptorX server can be accessed at http://raptorx.uchicago.edu/StructPredV2/predict/. **Providing the sequence of *Fp*CBM32 (UniProt P0DTR4 residues 504–648), RaptorX predicted that**

***Fp*GalNAcDeAc residues Glu527, His543, Tyr546, Arg574, Asn579, Phe635 bind to galactose**.

The composition of a galactose binding site of a homologous bacterial CBM32 (*Cp*CBM32–α-galactose PDB ID 2V72; 46% amino acid identity with 95% coverage) is nearly identical to predicted binding site residues of *Fp*CBM32 (the only difference being a tryptophan instead of the equivalent Tyr546 in *Fp*CBM32). Structural alignment reveals a high degree of overlap between the predicted galactose binding residues of *Fp*CBM32 with the experimentally resolved galactose binding residues of PDB 2V72 (**Figure 3.4**). *Cp*CBM32's specificity for galactose is reported to be driven by direct hydrogen bonds to the equatorial O3 and axial O4 atoms by the binding-site Arg68 and His37 residues, respectively (using residue numbers as reported in PDB 2V72; see Figure 3.4)[118]. Given the presence and predicted orientation of the corresponding Arg574 and His543 residues in my *Fp*CBM32 model, I hypothesized that these two residues played the same important role in recognizing the galactose residue of LacNAc. I therefore moved forward with high confidence in (1) the accuracy of the predicted binding site residues of *Fp*CBM32 and (2) the importance of hydrogen bonds made by *Fp*CBM32's His543 and Arg574 residues for LacNAc docking refinement.

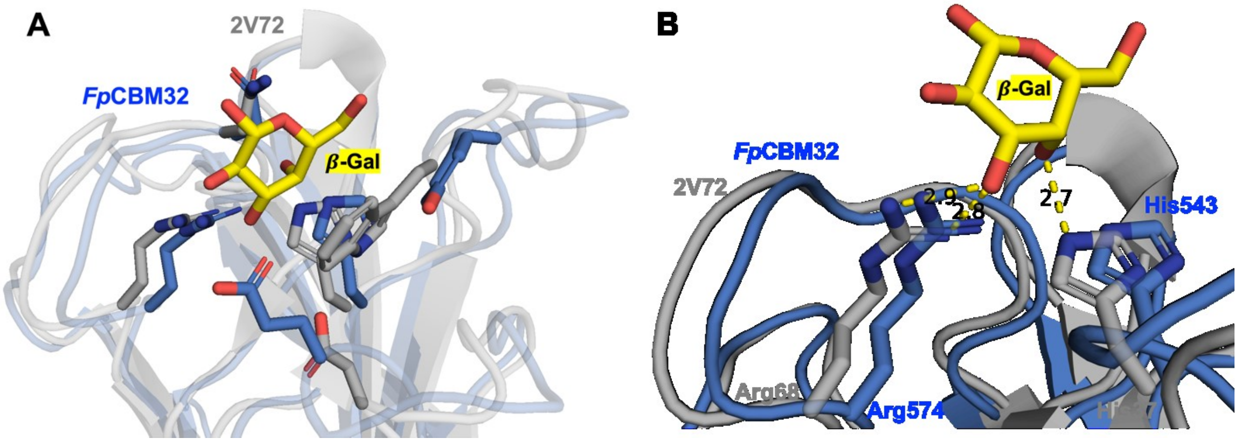

Figure 3.4: Illustration of the similarities between the β-Gal binding site of the experimental *Cp*CBM32–β-Gal crystal structure (PDB 2V72) *versus* the predicted LacNAc binding site of the *Fp*CBM32 AlphaFold model. (A) Overall comparison of the carbohydrate binding sites of *Cp*CBM32 (gray, experimental) and *Fp*CBM32 (blue, predicted). *Cp*CBM32 was resolved bound to β-Gal (yellow, sticks). There is significant structural overlap between the β-Gal binding site residues of *Cp*CBM32 and the predicted LacNAc binding residues of *Fp*CBM32 (gray *versus* blue sticks). (B) A zoom-in on the two primary residues of *Cp*CBM32 that contribute to β-Gal specificity. *Cp*CBM32 residues Arg68 and His37 (gray, opaque sticks) hydrogen bond specifically to the equatorial O3 and axial O4 atoms of β-Gal, respectively (yellow, dashed lines; atom-pair distances are in reported Ångstroms). *Fp*CBM32 has the equivalent Arg574 and His543 residues (blue, transparent sticks) predicted to be in the same orientation as the corresponding residues in the *Cp*CBM32–β-Gal crystal structure.

3.2.2.2 Pre-sampling conformations of FpCBM32 that are best poised to accommodate LacNAc at the predicted binding site

Given the uncertainty as to whether AlphaFold predicted a bound or unbound conformation of *Fp*CBM32, I wanted to increase my chances of generating an accurate *Fp*CBM32–LacNAc complex model by pre-sampling conformations of *Fp*CBM32 that could realistically accommodate LacNAc at the predicted binding site prior to performing glycoligand docking. Using the Rosetta FastRelax protocol in dual-space (*i.e.*, all-atom refinement in both dihedral and Cartesian space)[119,120], I generated 250 low-energy conformations of *Fp*CBM32 starting from the predicted AlphaFold model. Since Rosetta score alone cannot distinguish a bound *versus* unbound conformation in the absence of the given binder, I turned to assessing the accessibility of the predicted binding site using

P2Rank: a machine learning-based, protein template-free tool for predicting ligand binding sites using the calculated solvent accessibility of local chemical neighborhoods[117]. The P2Rank server is accessible at https://prankweb.cz/. By assessing each relaxed *Fp*CBM32 model using P2Rank, I identified 52 (among the original 250) refined *Fp*CBM32 models that I expected to be the best poised to accommodate and bind LacNAc at the predicted binding site. Selection criteria for P2Rank-assessed models were as follows: the top-ranked ligand binding pocket must contain no more than eight residues total and must contain all six predicted LacNAc binding residues (reported in bold in the previous section).

3.2.2.3 Using GlycanDock to refine a model of the FpCBM32–LacNAc complex structure

To run GlycanDock protein–glycoligand refinement, I needed to generate an initial *Fp*CBM32–LacNAc complex to serve as the input starting structure for each of the 52 relaxed *Fp*CBM32 conformations selected by P2Rank. To do so, I first re-aligned these 52 relaxed *Fp*CBM32 conformations and PDB 2V72 onto the original *Fp*CBM32 AlphaFold model. I then grafted in a LacNAc disaccharide by aligning its galactose residue onto the bound galactose residue of PDB 2V72. Finally, I grafted this aligned conformation of LacNAc onto each of the P2Rank-selected 52 relaxed *Fp*CBM32 conformations, resulting in 52 initial *Fp*CBM32–LacNAc input starting structures to be refined using GlycanDock.

In **Figure 3.5**, the blue circles depict the “funnel” plot results of GlycanDock refinement on all the 52 initial *Fp*CBM32–LacNAc models (where the x-axis is the ring-atom RMSD of the refined LacNAc conformation taken to its starting conformation, which is identical in all 52 inputs). Similarly, the orange diamonds depict the “funnel” plot results of GlycanDock refinement when employing atom-pair distance constraints to enforce direct hydrogen bonds between the equatorial O3 and axial O4 atoms of LacNAc’s terminal galactose residue with *Fp*CBM32’s Arg574 and His543 residues, respectively. While GlycanDock sampled more conformational space when no constraints were employed (as evident by the wider spread of blue circles all along the x-axis of Figure 3.5), the atom-pair distance constraints guided sampling toward more realistic docked models that better matched experimental information.

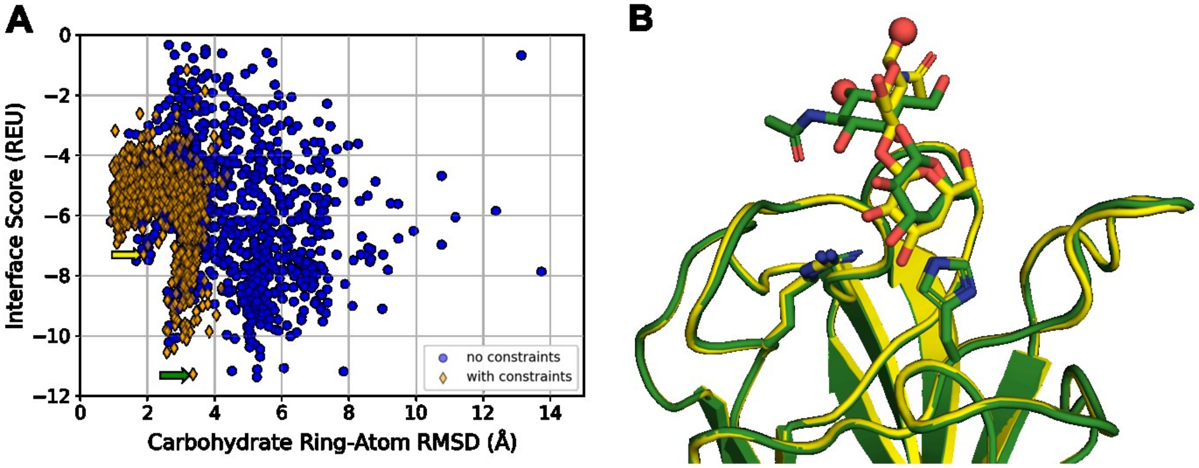

Figure 3.5: Results of GlycanDock refinement of *Fp*CBM32–LacNAc complex models. (A) Results of GlycanDock refinement depicted as a “funnel” plot where the x-axis is the RMSD (Ångstroms) of the carbohydrate ring atoms of the refined model taken to the starting conformation, and the y-axis is Rosetta interface score (REU) of the refined model. Blue circles represent models that did not have any atom-pair distance constraints guiding the system; orange diamonds represent models that did. (B) Two representative GlycanDock refined models of *Fp*CBM32–LacNAc as indicated by the arrows of corresponding color in Panel A. Both models came from a GlycanDock refinement simulation that employed atom-pair distance constraints. The red spheres highlight the reducing-end oxygen atoms (connected to the C1 anomeric carbons) of LacNAc from which, in the relevant biological setting, additional LacNAc units would extend.

The top-scoring *Fp*CBM32–LacNAc model from GlycanDock refinement using atom-pair distance constraints is indicated with a green arrow in Figure 3.5A and shown in green in Figure 3.5B. While the galactose unit of LacNAc may be appropriately hydrogen bonding with the *Fp*CBM32 Arg574 and His 543 residues (hydrogen bonds not indicated), the *N*-acetyl-glucosamine unit of LacNAc appears to be lying flat against the protein. This explains this docked model's highly favorable interface score (-11.2 REU) since more interfacial interactions generally leads to more favorable interface score. However, this conformation where the *N*-acetyl-glucosamine is lying flat against the protein is likely not a biologically relevant conformation despite its favorable Rosetta score. A more biologically relevant conformation is shown in yellow in Figure 3.5B and indicated with a yellow arrow in Figure 3.5A. This docked model still has a favorable interface score of -7.3 REU, but importantly has the *N*-acetyl-glucosamine unit of LacNAc oriented toward the solvent (*i.e.*, away from the protein). This docked conformation is more relevant because, in the predicted biological setting where *Fp*CBM32 is binding to the terminal LacNAc component of a cell surface glycolipid, there would be additional, repeating LacNAc units extending off the C1 atom of the *N*-acetyl-glucosamine (shown as a sphere). Therefore, the C1 atom of the reducing end *N*-acetyl-glucosamine should be oriented toward the solvent like it is in the yellow model.

This *Fp*CBM32–LacNAc docking study serves as a reminder that all known or predicted biological contexts and any experimental information should be considered when selecting a final model to represent the docked complex, rather than selecting a final model based on Rosetta score alone.

3.2.3 Goal 1C – Predict the bound conformation of the *Fp*DeAc–A-antigen complex

As mentioned in Section 3.2.1, Rahfeld *et al*. published a crystal structure of *Fp*DeAc (the deacetylase catalytic domain of *Fp*GalNAcDeAc) in both an unbound state (PDB ID 6N1A) and in a state bound to the blood carbohydrate B-antigen trisaccharide (PDB ID 6N1B; *Fp*DeAc–B-antigen). While the actual substrate of *Fp*DeAc is the A-antigen, the B-antigen mimics *Fp*DeAc's reaction product, the GalN-antigen (deacetylating the A-antigen trisaccharide results in the GalN-antigen trisaccharide). As a reminder, the differences between the A-, B-, and GalN-antigen trisaccharides are as follows: the A-antigen is characterized by containing a terminal *N*-acetyl-galactosamine residue at the non-reducing end whereas the B-antigen contains a galactose and the GalN-antigen contains a galactosamine (in all three cases the carbohydrate is α-linked to its parent). In any case, the structures of the *Fp*DeAc–A-antigen and *Fp*DeAc–GalN-antigen complexes (*i.e.*, the enzyme–substrate and enzyme–product complexes, respectively) would better inform *Fp*DeAc's catalytic mechanism and serve as more relevant starting models for future enzyme engineering endeavors (*e.g.*, improving *Fp*DeAc's catalytic activity on the A-antigen). In this Section 3.2.3, I report my steps to generate a Rosetta-refined model of the *Fp*DeAc–A-antigen and *Fp*DeAc–GalN-antigen complexes using the available *Fp*DeAc–B-antigen crystal structure as a template (PDB ID 6N1B). I then detail my findings after performing a closer structural comparison between all three refined *Fp*DeAc–blood-antigen trisaccharide complex models.

3.2.3.1 Using the experimental FpDeAc–B-antigen structure to generate refined models of the FpDeAc–A-antigen and FpDeAc–GalN-antigen complexes

To manually generate an initial model of the *Fp*DeAc–A-antigen complex, I first aligned an α-*N*-acetyl-galactosamine monosaccharide (PDB ID A2G) onto the terminal, non-reducing α-galactose of the B-antigen trisaccharide from the *Fp*DeAc–B-antigen experimental structure (PDB ID 6N1B) using PyMOL. I then replaced the coordinates of the α-galactose's O2 atom with the coordinates of the *N*-acetyl group from the aligned A2G molecule, thus manually generating a starting model of the *Fp*DeAc–A-antigen complex. By performing these same steps but replacing the α-galactose's O2 atom with the coordinates of only the nitrogen atom of the aligned A2G molecule, I also manually generated a starting model of the *Fp*DeAc–GalN-antigen complex. Finally, I refined both initial complex models using the Rosetta FastRelax protocol (*i.e.*, all-atom refinement in dihedral space)[119]. Throughout this process I retained two of the four resolved calcium ions from PDB 6N1B: the active-site calcium and the calcium ion likely important for enzyme structure (calcium ion residues 501 and 502, respectively). Therefore, relaxing the *Fp*DeAc–blood-antigen models included any non-water-mediated effects of calcium's (a divalent metal ion) reported role of coordinating the active site for catalysis[110].

Figure 3.6 compares the results of subjecting the experimentally resolved *Fp*DeAc–B-antigen structure (gray circles) and the two manually generated *Fp*DeAc–A-antigen (orange diamonds) *Fp*DeAc–GalN-antigen (blue squares) models to Rosetta's FastRelax all-atom refinement protocol. We can clearly observe that the *Fp*DeAc–A-antigen model has the most favorable Rosetta interface score (average -19.65 ± 0.86 REU) compared to *Fp*DeAc–B-antigen (-14.17 ± 0.97) and *Fp*DeAc–GalN-antigen (-13.04 ± 0.99). This result agrees with experimental information given that the *Fp*DeAc–A-antigen complex is the true enzyme–substrate complex, which requires favorable binding for catalysis to proceed, and that the *Fp*DeAc–GalN-antigen complex is the true enzyme–product complex, which requires less favorable binding for the deacetylated product to be released.

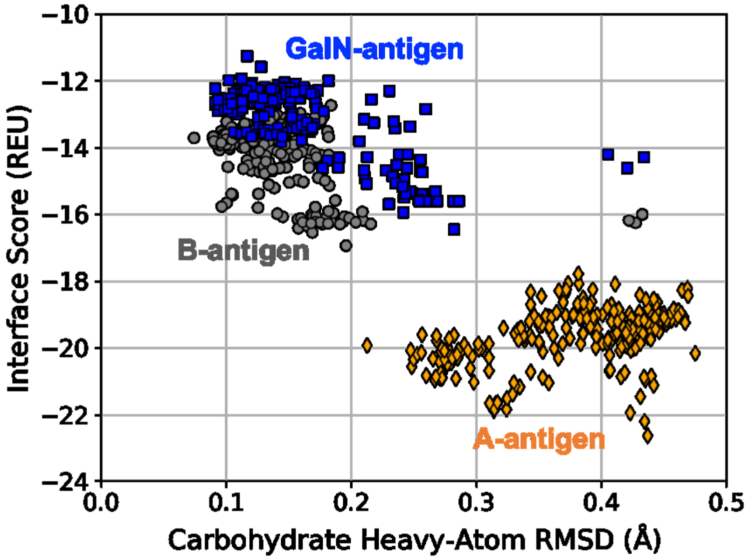

Figure 3.6: “Funnel” plot depicting the relationship between the carbohydrate heavy-atom RMSD (x-axis; Ångstroms) and the Rosetta interface score (y-axis; Rosetta Energy Units) after FastRelaxing the experimental *Fp*DeAc–B-antigen complex (gray circles) and the manually generated *Fp*DeAc–A-antigen (orange diamonds) and *Fp*DeAc–GalN-antigen (blue squares) complex models (*nstruct* = 250 each). The carbohydrate heavy-atom RMSD is calculated using the corresponding input structure as the reference.

3.2.3.2 Analysis of the refined FpDeAc–A-antigen complex models reveals the active site residues most important for binding the terminal N-acetyl-galactosamine residue

Given the agreement between the Rosetta FastRelax simulation results and experimental information, I used my refined *Fp*DeAc–A-antigen complex models to identify the active site residues that contribute the most (and the least) to binding the terminal *N*-acetyl-galactosamine residue characteristic of the A-antigen. Here, contribution to binding is approximated by the two-body Rosetta energy score between the terminal *N*-acetyl-galactosamine and a given *Fp*DeAc interfacial residue.

Figure 3.7 depicts the average Rosetta two-body energy scores between the terminal *N*-acetyl-galactosamine of the A-antigen and the specified *Fp*DeAc residues (including the active site calcium ion, residue 501A) at the interface after the FastRelax refinement described in the previous section. Error bars represent the standard deviation. Only two-body energies where the given *Fp*DeAc residue was present in at least 10% of the refined models generated (*i.e.*, at least 25 of the 250) are shown. The count of *Fp*DeAc residues that contributed to the interface in fewer than the 250 the refined models generated are noted at the top of the figure. All two-body energy scores are calculated from direct interactions; there are no explicit water molecules or water-mediated interactions modeled in the system.

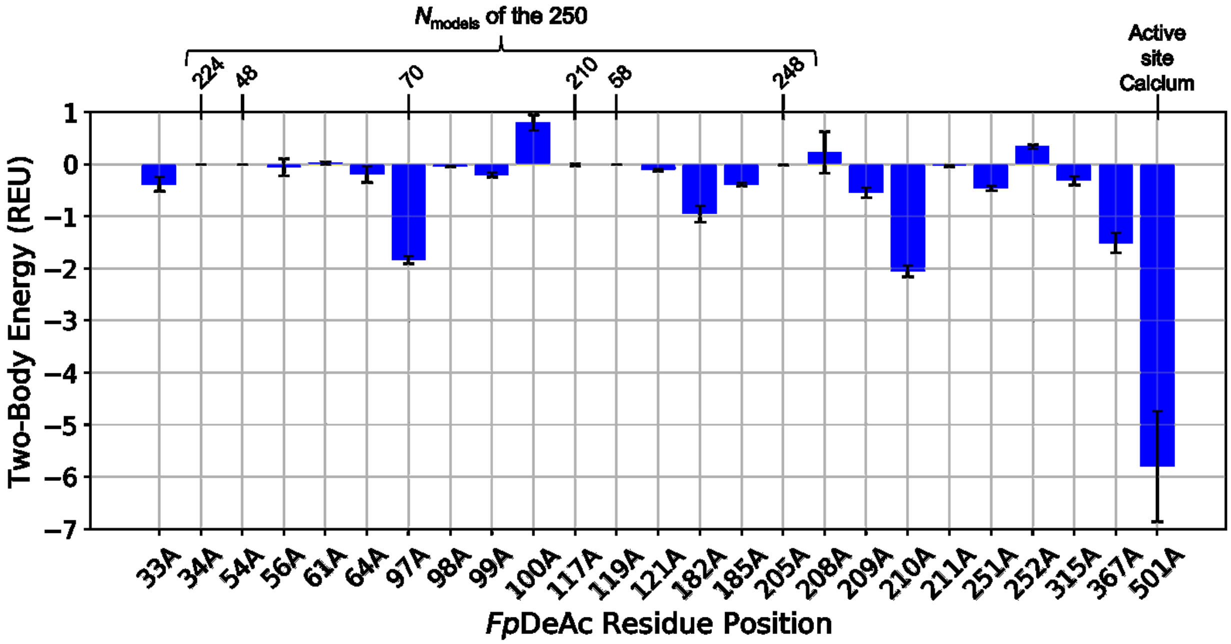

Figure 3.7: Breakdown of average Rosetta two-body energy scores (y-axis; REU) between the *N*-acetyl-galactosamine of the A-antigen to the *Fp*DeAc interfacial residues (x-axis) across refined *Fp*DeAc–A-antigen complex models (*nstruct* = 250). The more negative the two-body energy, the more favorably the given residue interacts with the *N*-acetyl-galactosamine. If a given *Fp*DeAc residue was not present at the *Fp*DeAc–A-antigen interface in all 250 of the refined complex models, the count of times it was present is noted at the top of the chart.

It is clear from **Figure 3.7** that the active site calcium (*Fp*DeAc residue 501A) energetically contributes the most to binding the terminal *N*-acetyl-galactosamine residue of the A-antigen. Rahfeld *et al.* already noted the importance of this calcium (a divalent metal ion) for *Fp*DeAc catalysis to occur. Active site coordinated calcium ions are also commonly observed in carbohydrate deacetylase enzymes[121].

*Fp*DeAc protein residues 97A (His), 182A (Tyr), 210A (Tyr), and 367A (Trp) also contribute significantly to terminal *N*-acetyl-galactosamine binding. Interestingly, His97A only contributed to the interface in 70 of the 250 refined *Fp*DeAc–A-antigen complex models. In those 70 models, His97A is making a hydrogen bond to terminal *N*-acetyl-galactosamine's C3 hydroxyl, whereas His97A is oriented away from the carbohydrate in the remaining 180 refined models and is therefore not contributing this interfacial hydrogen bond (data not shown). The aromatic residues Tyr182A, Tyr210A, and Trp367A appear to be stacking against the bound A-antigen (data not shown), though not necessarily in the canonical orientations observed for CH–π interactions (see Fig 2 of Spiwok 2017[92]). Nevertheless, multiple carbohydrate–aromatic interactions likely serve to make desolvation upon complex formation less unfavorable (thanks to the electron density in the aromatic π orbitals) and are therefore still important for *Fp*DeAc–A-antigen binding[122]. Finally, due to the implicit solvation model Rosetta employs, my analysis does not include the effect of water-mediated hydrogen bonding interactions, which have long been identified as key determinants of productive protein–carbohydrate binding[107]. **Future computational work should model the *Fp*DeAc–A-antigen system with explicit waters and analyze their effects on interfacial Rosetta two-body energies.**

3.2.4 Goal 1D – Predict mutations that improve the stability of the *Fp*DeAc catalytic domain

Commercialization typically necessitates an increase in the scale of enzyme production and/or an increase in enzyme stability to ensure it remains catalytically active throughout the conversion process. In short, enzyme stability and ease of expression positively correlates with its commercial effectiveness. While Rahfeld *et al*. did not provide explicit details on the stability (*e.g.*, melting temperature) or ease of expression of *Fp*GalNAcDeAc (*e.g.*, average protein yield in milligrams per milliliter of host organism expression)[110], any enzyme mutant that increases stability and/or expression without affecting its catalytic activity is a generally welcomed modification. Accordingly, I wanted to utilize my refined *Fp*DeAc–A-antigen complex model (see previous Section 3.2.3), server-generated multiple sequence alignments, and automated Rosetta protein design protocols to identify mutations that are predicted to improve the stability and expression of *Fp*DeAc (and, by association, the *Fp*GalNAcDeAc enzyme as a whole).

3.2.4.1 DeepMSA server generates a multiple sequence alignment created from multiple genome databases

Some protein design protocols can utilize a multiple sequence alignment (MSA) to guide the sequence space explored during the design process (see PROSS in next Section 3.2.4.2). I used DeepMSA[123] (automated server available at https://seq2fun.dcmb.med.umich.edu/DeepMSA/) to generate an MSA for *Fp*DeAc (UniProt P0DTR4 residues 33–417). DeepMSA is an "open-source method for sensitive MSA construction, which has homologous sequences and alignments created from multi-sources of whole-genome and metagenome databases through complementary hidden Markov model algorithms"[123]. DeepMSA generated an alignment with 14,634 sequences (including *Fp*DeAc) with an alignment depth of 355.873 (see Zhang *et al*. 2019 for more information on alignment depth). I then employed HHfilter[124] *via* its implementation into the MPI Bioinformatics Toolkit server (https://toolkit.tuebingen.mpg.de/tools/hhfilter) to extract a representative set of sequences from the first 10,000 sequences of the MSA (HHfilter has an upper sequence limit of 10,000). HHfilter filtered representative sequences with a maximum of 90% sequence identity, a minimum of 30% sequence identity, and a minimum of 75% sequence coverage with respect to the *Fp*DeAc query sequence. The resulting representative MSA contained 77 sequences[b], including *Fp*DeAc.

3.2.4.2 PROSS server identifies mutations predicted to improve FpDeAc stability and bacterial expression

PROSS (the Protein Repair One-Stop Shop server) takes an enzyme structure and uses Rosetta protein design protocols combined with MSA-derived evolutionary sequence constraints to identify mutations predicted to improve enzyme stability and bacterial expression[125]. For my work, I provided PROSS the bound *Fp*DeAc–B-antigen crystal structure (PDB ID 6N1B) as input (removing only two calcium ions, CA 503 and 504), and the MSA generated by DeepMSA as described in the above Section 3.2.4.1. I also provided a list of seventeen protein residues to keep fixed during the design process to minimize the risk of sampling mutations that would impact enzyme activity: *Fp*DeAc positions 36, 59, 61, 64, 97, 99, 100, 102, 121, 182, 185, 209, 210, 251, 252, 254, 315 (using the residue numbering as in 6N1B). Finally, I specified that PROSS use the REF2015 Rosetta scoring function during the protein design simulation (the “sugar_bb” energy term for glycosidic linkages cannot be included *via* the server, but this limitation does not appreciably affect my results). PROSS outputs nine designs it predicts to be more stable and/or easily expressed (in bacteria) than the wildtype, where each design generally contains more mutations than the previous. **Table 3.1 provides a summary of**

[b] I do not fully know how the MSA with >14,000 sequences was filtered down to a representative set of only 77 sequences. DeepMSA’s output did not include identifiers for the sequences (*e.g.*, a UniProt ID) so I could not perform an appropriate follow-up investigation.

seven of the suggested *Fp*DeAc mutations identified by PROSS for improvement of stability and expression.

Table 3.1: All *Fp*DeAc PROSS designs predicted to improve the enzyme's expression with < 10% of a mutational load to the native sequence. *Fp*DeAc residue numbering follows that of the 6N1B crystal structure.

Count	Position	WT	Des1	Des2	Des3	Des4	Des5	Des6	Des7
7	72	S	N	N	N	N	N	N	N
7	60	I	D	D	D	D	D	D	D
7	237	G	N	N	N	N	N	N	N
7	88	E	V	V	V	V	V	V	V
7	51	A	M	M	M	M	M	M	M
7	114	A	C	C	C	C	C	C	C
7	101	A	P	P	P	P	P	P	P
6	360	K		G	G	G	G	G	G
6	327	A		P	P	P	P	P	P
6	107	A		P	P	P	P	P	P
5	274	R			E	E	E	E	E
5	253	Q			L	L	L	L	L
5	200	L			M	M	M	M	M
4	233	V				S	S	S	S
4	302	S				T	T	T	T
4	81	E				P	P	P	P
3	130	V					I	I	I
3	268	S					I	I	I
3	194	F					T	T	T
3	215	A					N	N	N
3	166	A					K	K	K
2	108	E					K	K	K
2	16	Q						L	L
2	112	M						L	L
2	202	A						Y	Y
1	126	V							S
1	282	H							Y
1	230	G							S
1	329	D							N
1	349	A							Q

3.2.4.3 Rosetta Cartesian ΔΔG protocol identifies mutations predicted to improve FpDeAc stability

To supplement the stabilizing *Fp*DeAc mutations predicted by PROSS, I also used the Rosetta Cartesian ΔΔG protocol (cart_ddg) to identify mutations that stabilize the protein's fold[126,127]. The cart_ddg protocol uses Cartesian-space refinement restricted to the neighborhood of a specified mutational site to allow for local side chain and backbone movement to accommodate the change in residue identity. The resulting total energy change due to the mutation (ΔG_{mutant}) is compared to the resulting total energy change due to performing the same mutation-plus-refinement process when "mutating" to the wildtype residue ($\Delta G_{\text{wildtype}}$). If the mutation is energetically favorable (*i.e.*, $\Delta G_{\text{mutant}} - \Delta G_{\text{wildtype}} = \Delta\Delta G_{\text{mutation}} < 0$), then the mutation is predicted to stabilize the protein's fold.

The cart_ddg protocol requires a relaxed starting structure as input. I took the bound conformation of *Fp*DeAc from the 6N1B crystal structure (keeping only the protein residues) and relaxed it in Cartesian space using the parameters suggested by the online cart_ddg documentation (https://www.rosettacommons.org/docs/latest/cartesian-ddG). I kept only the protein residues because metals/ions are generally incompatible with any form of Cartesian-space sampling in Rosetta, and cart_ddg is not well tested with protein interfaces. I performed the initial relaxation preparation step 50 independent times and used the lowest scoring model by total Rosetta score as input to cart_ddg.

Using the cart_ddg protocol, I sampled all point mutations of *Fp*DeAc excluding the same seventeen positions that were held fixed when running PROSS. In other words, I used cart_ddg to independently sample each of the twenty amino acids at all but seventeen positions of *Fp*DeAc (385 residues in length as resolved in 6N1B). The command line arguments to run a cart_ddg simulation for a single mutation are provided in Section 3.6. **Table 3.2 lists all single-point cart_ddg designs with a $\Delta\Delta G_{\text{mutation}} \leq -7.0$ REU.**

Table 3.2: All *Fp*DeAc single-point Rosetta cart_ddg designs predicted to stabilize the enzyme's fold. *Fp*DeAc residue numbering follows that of the 6N1B crystal structure.

Position	WT	Mutant	Position	WT	Mutant
35	P	I / V	117	N	F
56	N	I / V	203	N	V
88	E	V	253	Q	F
			268	S	V

Each design identified by cart_ddg appears to install an additional hydrophobic residue to increase hydrophobic burial (a major driver of protein folding). While these types of mutations were not explicitly sought, mutations that increase the amount of stabilizing van der Waals contacts with neighboring residues (as captured by the "fa_atr" Rosetta score term) are the simplest to identify since these types of residue-pair interactions are independent of orientation and direction (unlike geometry-sensitive interactions like hydrogen bonds and salt bridges).

3.3 Section 1 Summary

In this Section, I described the structural model of full-length *Fp*GalNAcDeAc that I generated using AlphaFold. I detailed how I leveraged binding site prediction software, homologous structural information, and my Rosetta protein–glycoligand docking refinement protocol to generate a model of the *Fp*CBM32–LacNAc complex. Similarly, I detailed my steps to generate a refined model of the *Fp*DeAc–A-antigen complex and the *Fp*DeAc–GalN-antigen complex, leading to a description of the residues important for terminal *N*-acetyl-galactosamine binding. Finally, I reported the mutations I identified using two different protein design approaches that are predicted to stabilize *Fp*DeAc and improve enzyme expression. Designed *Fp*DeAc sequences were shared with our experimental collaborators (Peter Rahfeld and Charlotte Olagnon from the lab of Stephen Withers at the University of British Columbia, Vancouver, Canada) for testing.

3.4 Section 2 – *Fp*GalNase computational modeling and design

3.4.1 Goal 2A – Model the full-length *Fp*GalNase enzyme structure

Unlike *Fp*GalNAcDeAc, there were no experimentally determined structures for any domain of *Fp*GalNase at the time of my study. Accordingly, I used AlphaFold to predict the full-length structure of *Fp*GalNase (Goal 2A). This predicted *Fp*GalNase model was then used to conduct structure- and sequence-based analyses to probe the origin of *Fp*GalNase's unique specificity to α-D-galactosamine (Goal 2B).

3.4.1.1 AlphaFold predicts full-length FpGalNase enzyme structure with high predicted accuracy

The full-length sequence of *Fp*GalNase (1078 residues, including the signal peptide) is published under UniProt accession number P0DTR5. From this sequence, I generated a full-length structural model of *Fp*GalNase using AlphaFold[112]. *Fp*GalNase also consists of four distinct protein domains: the N-terminal α-galactosaminidase catalytic domain (*Fp*GH36; a glycosyl hydrolase (GH) of family 36) and three consecutive carbohydrate binding modules (CBMs) of unknown family[110]. During the time of writing this thesis, an interactive model of the full-length *Fp*GalNase structure automatically generated by AlphaFold became available online through the AlphaFold Protein Structure Database at https://alphafold.ebi.ac.uk/entry/P0DTR5[113].

Figure 3.8 depicts the full-length *Fp*GalNase AlphaFold model colored by the four distinct protein domains. About 96% of the *Fp*GalNase residues (excluding the signal peptide) had a pLDDT estimate ≥ 70 (80% of residues were ≥ 90), meaning the overall confidence in the accuracy of the AlphaFold model is high and can therefore be used in downstream analyses such as those described in the following sections.

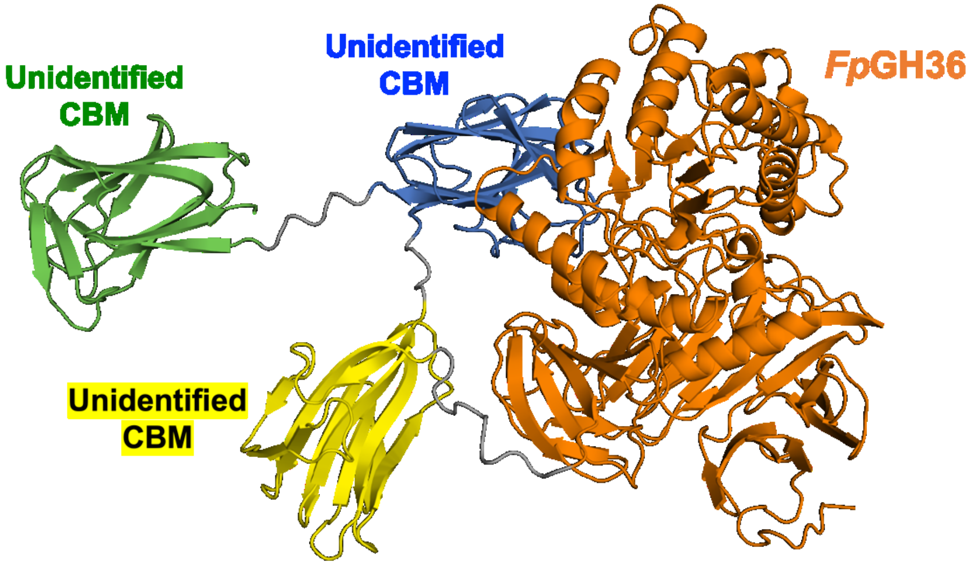

Figure 3.8: AlphaFold structural model of full-length *Fp*GalNase colored by domain. The N-terminal, catalytic *Fp*GH36 domain is in orange and the three CBMs of unknown family are in yellow, blue, and green. Linker regions between domains are colored in gray.

3.4.1.2 Predicted AlphaFold model provides more specific residue-level boundaries for the four domains of FpGalNase

Often, enzymes are truncated down to their individual domains to perform biochemical experiments that can then identify that domain's function. In these truncation experiments, it is important for the experimentalist to have accurate start- and end-residue cutoffs to ensure only the protein domain of interest is biochemically interrogated and that the given domain is in its most biologically relevant, stable form.

AlphaFold's predicted model of full-length *Fp*GalNase provides clearer residue-level boundaries for the four domains of this enzyme. **Reported as UniProt P0DTR5 residue numbers, the four domains of *Fp*GalNase are as follows: residues 28–697 the catalytic α-*N*-acetyl-galactosamine galactosaminidase (GH36) domain, residues 710–827 the first CBM of unknown family, residues 833–950 the second CBM of unknown family, and residues 958–1078 the third and final CBM of unknown family (see Figure 3.8).** Future *Fp*GalNase truncation experiments–in particular, those that probe the carbohydrate specificities of the three unknown CBMs–should be based on these reported start- and end-residue domain cutoffs.

3.4.2 Goal 2B – Predict the *Fp*GH36 residues important for GalN-antigen binding and specificity

Rahfeld *et al*. identified the catalytic domain of *Fp*GalNase as a glycosyl hydrolase (GH) of family 36 (*Fp*GH36)[110]. The GH36 family is a member of the glycosyl hydrolase clan GH-D superfamily, which primarily consists of α-galactosidases and α-*N*-acetyl-galactosaminidases[128]. Enzymes in this superfamily typically share a common catalytic mechanism and structural topology[129]. While α-galactosidases (EC 3.2.1.22) hydrolyze terminal, non-reducing α-D-galactose residues[130], *Fp*GH36 is specific to terminal, non-reducing α-D-galactos<u>amine</u> (α-GalN) residues (and not α-D-galactose (α-Gal) nor α-D-*N*-acetyl-galactosamine (α-GalNAc) residues)[110]. Rahfeld *et al*. claimed that, to the best of their knowledge, there is no other previously reported GH36 enzyme specific to α-D-galactosamine.

My goal was to perform a structure- and sequence-based interrogation toward developing a biophysical understanding of *Fp*GH36's unique specificity to terminal, non-reducing α-GalN sugars. To identify *Fp*GH36 residues important for carbohydrate binding, specificity, and catalysis, I compared the active sites of the predicted *Fp*GH36 AlphaFold model to experimental structures of other GH36 enzymes as well as known residue motifs reported in the GH-D superfamily literature. **Figure 3.9** illustrates the sequence and predicted structure of each residue motif as it is found in the *Fp*GH36 model. **Table 3.3** reports additional information about these conserved residue motifs and the experimental structures used in this analysis. Both Figure 3.9 and Table 3.3 are referred to throughout this Section 3.4.2. Ultimately, the results of my analysis should guide future biochemical experiments toward targeted mutational investigation of the *Fp*GH36 active site.

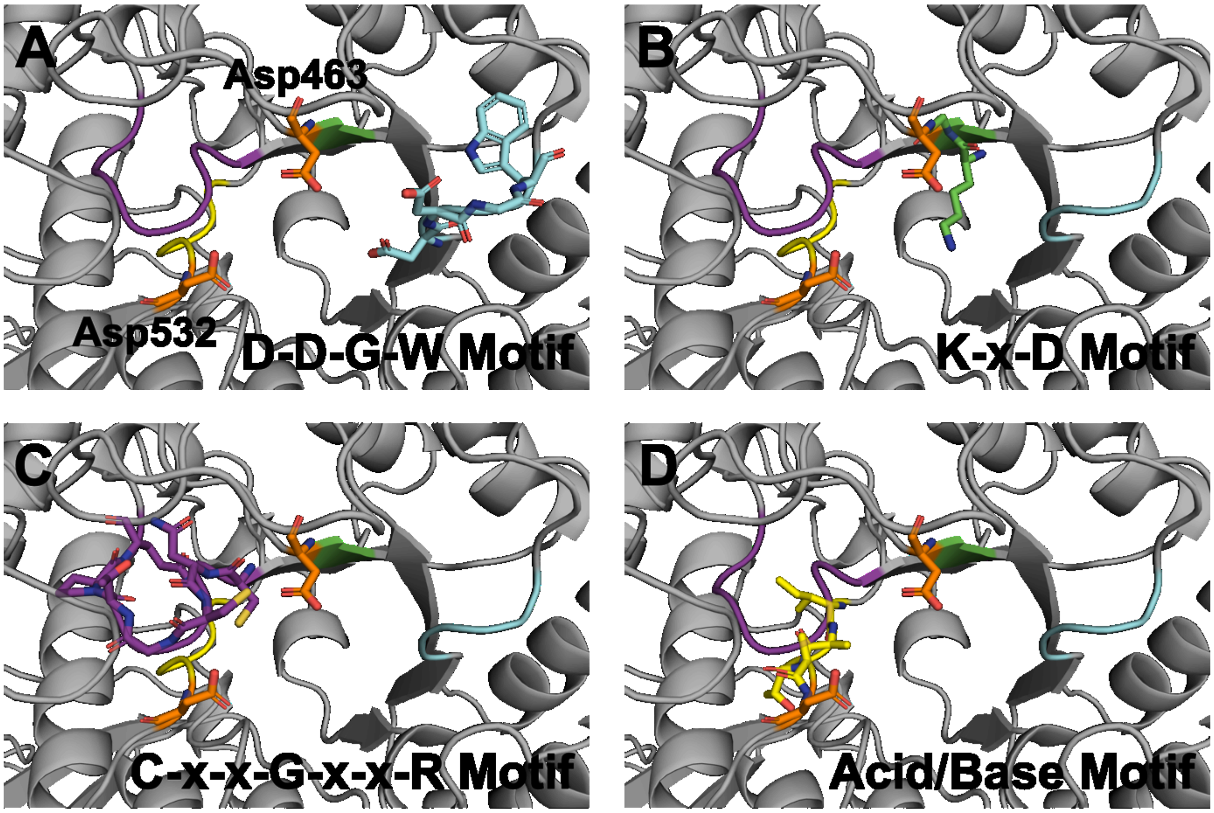

Figure 3.9: Predicted conformations of *Fp*GH36 active site residues that are the equivalent to known GH-D superfamily motifs. (A) The D-D-G-W motif (cyan sticks) with respect to Asp463/Asp532 (orange sticks; the two catalytic residues of *Fp*GH36, gray cartoon). (B) The equivalent K-x-D motif (green sticks) with respect to Asp463/Asp532 (orange sticks). (C) The equivalent C-x-x-G-x-x-R motif (purple sticks) with respect to Asp463/Asp532 (orange sticks). (D) The equivalent acid/base motif (yellow sticks) with respect to Asp463/Asp532 (orange sticks). In all panels, the orientation of *Fp*GH36 does not change.

Table 3.3: The residue motifs of *Fp*GH36 compared to those present in five experimental GH36 structures. * indicates that the Asp residue was mutated to Asn for crystallization purposes. The A chain was utilized for analysis unless otherwise stated.

Target	***Fp*GalNase**	**2YFO**	**4FNT** (chain B)	**6PI0**	**6JHP**	**6LCK**
Resolution	—	1.35 Å	2.6 Å	2.09 Å	2.56 Å	2.85 Å
% Identity (% Coverage)	—	31.73% (14%)	27.16% (22%)	34.74% (13%)	24.30% (49%)	30.77% (25%)
Bound Substrate	—	α-Galactose	Raffinose	Linear Blood Group B Type 2 Trisaccharide	—	p-nitrophenyl α-D-galactopyranoside
Catalytic Residues (nucleophile / acid/base)	Asp463 / Asp532	Asp478 / Asp540	Asp478 / Asp548*	Asp472* / Asp541	Asp509 / Asp571	Asp301 / Asp355
D-D-G-W Motif	^{346}D-D-G-W^{349}	D-D-G-W	D-D-G-W	D-D-G-W	D-D-G-W	D-D-G-W
K-x-D Motif	^{461}K-G-**D^{463}**	K-W-**D**	K-W-**D**	K-W-**D**	K-W-**D**	K-L-**D**
C-x-x-G-x-x-R Motif	Partly; ^{511}C-N-C-G-T-P-Q^{517}	Yes; C-S-G-G-G-G-R	Yes; C-S-G-G-G-G-R	Yes; C-S-G-G-G-G-R	Yes; C-A-S-G-G-G-R	No; Active site Trp50
Acid/Base Motif	Not Present *[see resi. 528–532]*	"W-x-x-**D**"	"W-x-x-**D**"	"W-x-x-**D**"	"W-x-x-**D**"	"R-x-x-x-**D**"
Oligomeric State	Monomer *[personal correspondence]*	Tetramer	Tetramer	Tetramer	Tetramer	Hexamer
Other Related PDBs	—	2YFN	4FNP, 4FNQ, 4FNR, 4FNS, 4FNU	6PHU, 6PHV, 6PHW, 6PHX, 6PHY, 6PQL, 6PRE, 6PRG	—	6LCJ, 6LCL

3.4.2.1 The invariant catalytic residues of GH36 enzymes are two aspartic acid residues

Rahfeld *et al.* identified *Fp*GH36's two catalytic aspartic acid residues before I began this project. Reported in UniProt P0DTR5 numbering, these residues are: Asp463 (a catalytic nucleophile that forms a covalent intermediate with α-GalN) and Asp532 (a general acid/base residue)[110]. Readers should visit the CAZypedia (an online encyclopedia for carbohydrate-active enzymes[90]) for details on the GH family's catalytic mechanism: https://www.cazypedia.org/index.php/Glycoside_hydrolases.

3.4.2.2 Conserved active-site residue motifs contribute to GH36 enzymatic activity

There are at least four evolutionary-derived sequence motifs that contain catalytic and/or non-catalytic residues important for GH36 activity (including binding, specificity, and stabilizing the reaction pathway). Here, I describe my sequence- and structure-based comparison of the four residue motifs as they appear (or not) in *Fp*GH36.

3.4.2.2.A The D-D-G-W substrate recognition motif recognizes the hydroxyl groups at the 4- and 6-position of the bound, terminal carbohydrate

The D-D-G-W substrate recognition motif is nearly invariant across all identified GH36 enzymes[131–133]. This motif appears in an active-site loop where the two aspartic acid residues contribute to enzyme activity by directly hydrogen bonding with the hydroxyl groups at the 4- and 6-position of the terminal carbohydrate substrate[133,134]. *Fp*GH36 contains the D-D-G-W motif, consisting of residues ^{346}D-D-G-W^{349} (Figure 3.9A). AlphaFold predicted Asp346 and Asp347 to be in an appropriate conformation in my

*Fp*GH36 structural model to make the same direct hydrogen bonds important for carbohydrate binding specificity. The D-D-G-W motif does not contribute to α-GalN specificity over α-Gal in *Fp*GH36, though, since the 4- and 6-positions are both invariant between the two carbohydrates.

3.4.2.2.B The K-x-D catalytic nucleophile motif recognizes the hydroxyl groups at the 3- and 4-position of the bound, terminal carbohydrate

The K-x-**D** catalytic nucleophile motif is also nearly invariant across GH36 enzymes[133,134]. Here, the X represents a hydrophobic residue (most often Trp, Leu, or Val) and **D** is the catalytic nucleophile (**Asp463** in *Fp*GH36). The K-x-**D** motif appears in a β-strand where the lysine makes conserved hydrogen bonds to the hydroxyl groups at the 3- and 4-position of the terminal carbohydrate substrate. The aspartic acid is the catalytic nucleophile that attacks the C1 atom of the terminal carbohydrate substrate to form the covalent intermediate during catalysis. In *Fp*GH36, the K-x-**D** motif consists of residues ^{461}K-G-D^{463} where Asp463 is the catalytic nucleophile and Gly462 is, notably, not a hydrophobic residue (Figure 3.9B). AlphaFold predicted Lys461 to be in an appropriate conformation in my *Fp*GH36 model to make the same hydrogen bonds to the 3- and 4-position carbohydrate hydroxyl. Though like the D-D-G-W motif, the K-x-**D** motif does not directly contribute to *Fp*GH36's specificity for terminal α-GalN over α-Gal since the 3- and 4-positions are invariant between the two carbohydrates.

I wondered if the unique presence of Gly462 as residue X of the K-x-**D** active-site motif (rather than a bulky hydrophobic residue) played any role in *Fp*GH36's unique α-GalN specificity. However, no matter its amino acid identity, the sidechain of residue X points away from the active site (due to the alternating nature of a β-strand) and the backbone atoms of residue X are too far away to make any meaningful interactions with a carbohydrate residue. For these reasons, I ruled out the possibility of the K-x-D active-site residue motif directly contributing to *Fp*GH36's specificity for terminal α-GalN over α-Gal. **It remains to be explored, however, whether the bulky tryptophan residue or the β-branched leucine or valine residues at the hydrophobic X position of the K-x-D motif have any allosteric or rigidifying effect that indirectly contributes to *Fp*GH36's terminal carbohydrate specificity.**

3.4.2.2.C Either a C-x-x-G-x-x-R active-site residue motif or an active-site tryptophan recognizes the carbohydrate hydroxyl group at the 2-position, depending on the GH36 enzyme

Fredslund *et al.* identified the C-x-x-G-x-x-R motif as a characteristic of a subgroup of GH36 enzymes' active site [see publication's Supplementary Figure 1, bottom left column][133]. In this motif, the glycine directly hydrogen bonds to the hydroxyl group at the 2-position of a bound, terminal carbohydrate *via* its backbone nitrogen atom; the cysteine is a highly conserved active-site residue in GH-D enzymes that appears to also interact with the 2-position of the carbohydrate; and the arginine makes a structural salt bridge. In a different GH36 subgroup, this motif is absent and instead an active-site tryptophan residue (rather than a glycine) directly hydrogen bonds to the 2-position hydroxyl group *via* its cyclic nitrogen atom (Fredslund *et al*. did not report any corresponding active-site residue motif that included this tryptophan)[133]. The existence of these GH36 motifs suggests the importance of a (nitrogen atom-mediated) hydrogen bond for terminal carbohydrate recognition at the 2-position.

At first inspection, *Fp*GH36 contains most of the residues of the C-x-x-G-x-x-R motif: [511]C-N-C-G-T-P-Q[517], where *Fp*GH36 has a glutamine (Gln517) instead of the salt-bridging arginine (Figure 3.9C). However, AlphaFold predicted the Gly514 to have its backbone nitrogen atom oriented *away* from the active site (pLDDT estimate of 77.91), meaning it is unlikely that Gly514 plays the same role characteristic of the motif. Instead, AlphaFold predicted the thiol group of Cys513 to be oriented toward the active site (pLDDT estimate of 79.71). **Taken together, I hypothesized that the active site-oriented thiol group of Cys513 in part contributes to *Fp*GH36's unique specificity to terminal α-GalN targets while the lack of a properly oriented Gly514 backbone nitrogen atom denies specificity to terminal α-Gal.** Understanding exactly how the thiol group of Cys513 interacts with the $-NH_2$ group of α-GalN, if at all, would greatly benefit from the determination of an experimental structure of a *Fp*GH36–α-GalN complex. **Until then, I recommend experimental studies to observe the effects, if any, of mutating *Fp*GH36 Cys513 to other residues such as glycine, alanine, and serine.**

3.4.2.2.D Neither of the two GH36 type-dependent acid/base residue motifs that help recognize the hydroxyl group at the 2-position of α-galactose are present in *Fp*GH36

There are two reported motifs that contain the catalytic acid/base residue required by GH36 enzymes (Asp532 of *Fp*GH36): W-x-S-D[133] and R-x-x-x-D[134]. In both cases, the acid/base residue motif is found on an active-site loop. While these two motifs may seem rather distinct in sequence space (*i.e.*, tryptophan is bulky and aromatic whereas arginine is long and positively charged), they contribute essentially the same atoms in nearly identical geometries to the GH36 active site. In the W-x-S-D active-site residue motif, the cyclic nitrogen of the tryptophan makes a coordinating hydrogen bond with the side-chain carboxyl group of the acid/base residue and the 2-position hydroxyl of the terminal carbohydrate substrate[133]. Similarly in the R-x-x-x-D motif, the arginine uses its terminal nitrogen atoms to coordinate the same acid/base residue with the 2-position carbohydrate hydroxyl[134].

Neither the W-x-S-D nor the R-x-x-x-D acid/base residue motif is present in *Fp*GH36. Instead, the sequence of this active-site loop region is ^{528}I-A-T-A-D^{532}, where Asp532 is the catalytic acid/base residue (Figure 3.9D). None of these residues (Ile, Ala, Thr) can contribute the same coordinating hydrogen bonds observed by the Trp and Arg residues of the W-x-S-D and R-x-x-x-D motifs, respectively. Further, none of the five other GH36 crystal structure I examined have this ^{528}I-A-T-A-D^{532} (or similar) sequence; they instead have either the W-x-S-D or R-x-x-x-D motif (see Table 3.3).

The lack of either conserved motif in *Fp*GH36 led me to develop two potential theories, that *Fp*GH36 either (1) uses a different mechanism (*e.g.*, non-conserved residue(s)) to coordinate the acid/base residue with the terminal α-GalN substrate for catalysis or (2) does not need to make this type of residue-level coordination for catalysis to occur. Given the unique presence of an additional thiol group (provided by Cys513 of the C-x-x-G-x-x-

R motif discussed in the previous paragrapah) predicted to be oriented toward the active site, I believe the former theory is the most relevant. **In fact, it is possible that these two active site cysteines (Cys513 and the highly conserved Cys511)—given their proximity to each other and predicted orientation—bind a divalent metal (*e.g.*, zinc) that then helps coordinate *Fp*GH36 catalysis**[135,136]**.** A chelation assay (such as the one Rahfeld *et al.* performed on *Fp*GalNAcDeAc[110]) should provide a quick assessment of the influence of divalent metals such as zinc on *Fp*GalNase activity. If *Fp*GalNase shows sensitivity to metal chelation, it would be interesting to test the effects of mutating Ile528, Ala529, Thr530, and Ala531 to alanine and/or glycine on *Fp*GH36 activity and carbohydrate specificity.

3.5 Section 2 Summary

In this Section, I described the structural model of full-length *Fp*GalNase that I generated using AlphaFold. I reported in detail four conserved residue-level motifs and their structural and functional implications in the GH36 family of enzymes. I then identified the equivalent motifs present in *Fp*GH36 (the catalytic domain of *Fp*GalNase) and provided my hypotheses on the functional role of the non-conserved residues present in the active site. Finally, I suggested mutations to test these hypotheses experimentally toward developing a complete understanding of *Fp*GH36's unique specificity to terminal α-GalN carbohydrate substrates (summarized in **Table 3.4**). Ultimately, a complete picture of *Fp*GH36 will enable rational design efforts to alter its specificity to the B-antigen, thus achieving complete conversion of A, B, and AB blood types to the universal O type.

Table 3.4: Summary of all suggested mutations to test the hypotheses regarding *Fp*GalNase catalytic function and carbohydrate specificity.

Design	Motif	Rationale	Section
Gly462I/L/W	K-x-D	X is supposed to be a bulky hydrophobic residue in this motif	3.4.2.2.B
Cys513G/A/S	C-x-x-G-x-x-R	Probe the role, if any, of Cys513 on catalysis and/or specificity	3.4.2.2.C
Ile528A Ala529G Thr530G/A Ala531G	W-x-S-D and R-x-x-x-D	Since Ile, Ala, and Thr cannot contribute the same hydrogen bonding coordination as W or R of these motifs, it would be interesting to probe the impact of increasing flexibility in this active-site region	3.4.2.2.D

3.6 Detailed Methods

AlphaFold

Open-source AlphaFold and the corresponding databases was downloaded from its Github repository (https://github.com/deepmind/alphafold) and installed on Rockfish. Rockfish is computing cluster provided by the Johns Hopkins University and maintained by the Advanced Research Computing (ARCH) in Baltimore, Maryland (https://www.arch.jhu.edu/). All AlphaFold modeling was executed on Rockfish.

Command

While logged into Rockfish, first access a compute node with a GPU partition using slurm:

```
srun --account=<account_name> --nodes=1 --ntasks-per-node=6 --
partition=<GPU_partition> --gres=gpu:1 --time=6:00:00 --pty bash
```

AlphaFold was run with the following command:

```
python run_alphafold.py --fasta_paths=</path/to/*.fasta> --gpus 1 --
cpus 8
```

pLDDT

AlphaFold output includes a per-residue estimate of its confidence of model accuracy on a scale of 0–100 called pLDDT (predicted local distance difference test). Briefly, residues with pLDDT ≥ 90 are modeled with very high confidence sufficient for further structural characterization; residues with pLDDT values between 70–89 are confident but should generally be treated with some caution; and residues with pLDDT values < 70 should generally be considered as incorrectly placed. Refer the associated AlphaFold publications for more information[112,137].

Rosetta all-atom refinement

FastRelax

```
/Rosetta/main/source/bin/relax.linuxgccrelease -database
</path/to/database> -ignore_unrecognized_res -flip_HNQ -no_optH false
-ex1 -ex2 -out:pdb_gz true -
multiple_processes_writing_to_one_directory -in:file:s
</path/to/input.pdb> -in:file:native </path/to/native.pdb>
```

DualspaceRelax

```
/Rosetta/main/source/bin/relax.linuxgccrelease -database
</path/to/database> -ignore_unrecognized_res -flip_HNQ -no_optH false
-ex1 -ex2 -out:pdb_gz true -
multiple_processes_writing_to_one_directory -in:file:s
</path/to/input.pdb> -in:file:native </path/to/native.pdb> -
relax:dualspace true
```

Rosetta Cartesian ΔΔ*G*

Relaxing starting structure in Cartesian space:

```
/Rosetta/main/source/bin/relax.linuxgccrelease -database
</path/to/database> -ignore_unrecognized_res -flip_HNQ -no_optH false
-ex1 -ex2 -out:pdb_gz true -
multiple_processes_writing_to_one_directory -in:file:s
</path/to/input.pdb> -in:file:native </path/to/native.pdb> -fa_max_dis
9.0 -relax:script cart2.script
```

Where the cart2.script file contains:

```
switch:cartesian
repeat 2
ramp_repack_min 0.02  0.01     1.0  50
ramp_repack_min 0.250 0.01     0.5  50
ramp_repack_min 0.550 0.01     0.0 100
ramp_repack_min 1     0.00001  0.0 200
accept_to_best
endrepeat
```

And the cart_ddg protocol is run using:

```
/Rosetta/main/source/bin/cartesian_ddg.linuxgccrelease -database
</path/to/database> -flip_HNQ -no_optH false -ex1 -ex2 -out:pdb_gz
true -multiple_processes_writing_to_one_directory -in:file:s
</path/to/input.pdb> -in:file:native </path/to/native.pdb> -fa_max_dis
9.0 -ddg:iterations 3 -ddg:force_iterations false -ddg:bbnbrs 1 -
ddg:frag_nbrs 4 -ddg:score_cutoff 1.0 -ddg:cartesian -ddg:json false -
ddg:legacy false -ddg:flex_bb false -ddg:dump_pdbs false -
score:weights ref2015_cart -ddg:mut_file example_mutfile.txt
```

Where example_mutfile.txt contains (as a single example case):

```
total 1
1
E 76 L
```

Here, “total 1” means we are testing only a single point mutant design. This example cart_ddg design is Glu (E) at Rosetta position number 76 to Leu (L).

Rosetta GlycanDock

```
/Rosetta/main/source/bin/GlycanDock.linuxgccrelease -database
</path/to/database> -flip_HNQ -no_optH false -ex1 -ex2 -out:pdb_gz
true -multiple_processes_writing_to_one_directory -in:file:s
</path/to/input.pdb> -in:file:native </path/to/native.pdb> -
include_sugars -maintain_links -lock_rings -
alternate_3_letter_pdb_codes pdb_sugar -docking_partners A_X -
carbohydrates:glycan_dock:stage1_perturb_glycan_com_rot_mag 0.0 -
carbohydrates:glycan_dock:stage1_perturb_glycan_com_trans_mag 0.0 -
carbohydrates:glycan_dock:stage1_rotate_glycan_about_com false -
carbohydrates:glycan_dock:stage1_torsion_uniform_pert_mag 0.0
```

Chapter 4

4. Conclusion and Future Directions

Carbohydrates are an integral component of all biological life as we know it. Carbohydrates play important roles in protein structure and function, cellular recognition and signaling, metabolism and regulation, health and disease, and much more. Understanding the residue-level interactions between carbohydrates and their protein binding partners is key to unraveling the mechanisms underlying their diverse functional roles in biology. However, elucidating experimental, three-dimensional structures of protein-carbohydrate complexes remains challenging, leaving researchers to primarily rely on computational tools to generate these necessary models. In early 2016, the development of AutoDock Vina-Carb marked a significant step toward accurate, *in silico* prediction of protein–carbohydrate complex structures[12,13]. However, the AutoDock tool is–as the name suggests–limited to docking only, meaning researchers who were interested in performing downstream analyses or protein design must turn to another computational tool. In this dissertation, I have further advanced the field's computational modeling capabilities by integrating a new protein–carbohydrate docking tool into the much functionally broader Rosetta macromolecular modeling and design software suite.

4.1 Summary of my contributions to the field

I dedicated most of my doctoral career developing and benchmarking GlycanDock, a Rosetta tool for modeling and docking protein–carbohydrate complexes[74]. To rigorously evaluate its accuracy, I curated a set of diverse, biologically relevant experimental complex structures. Ultimately, I used this benchmark set to demonstrate GlycanDock's ability to sample and discriminate native-like docked protein–carbohydrate models from starting structures of up to 7 Å root-mean-square deviation in the carbohydrate ring atoms. Hoping to set a new, higher standard for evaluating the accuracy of a protein–carbohydrate docking tool, I also provided an analysis of the important biophysical features of GlycanDock complex models, such as interfacial residue-residue contacts and hydrogen bonds. To illustrate the practical usage of GlycanDock by future researchers, I detailed an example case where I found GlycanDock modeling results to qualitatively correlate with experimental binding data. This observation served as a first step toward demonstrating GlycanDock's ability to discriminate between binding and non-binding glycoligands. I also provided a detailed guide to address the more realistic "blind" docking scenarios using GlycanDock and the FTMap solvent mapping tool. Overall, my work has resulted in a new computational tool that can contribute to advancing our understanding of protein-carbohydrate interactions.

In the final phase of my doctoral research, I employed multiple computational tools to gain insights into an enzymatic system with the potential for significant scientific and societal impact. I began by generating full-length structural models of *Fp*GalNAcDeAc and *Fp*GalNase, two multi-domain enzymes identified by Rahfeld *et al*. that together convert A-type blood to the universal O-type[110,111]. Using GlycanDock, I generated a docked model of *Fp*CBM32–LacNAc, identifying the *Fp*GalNAcDeAc residues that likely govern carbohydrate binding and thus offering targeted mutational sites to potentially modify or strengthen *Fp*GalNAcDeAc's cell-surface interactions. Similarly, I used GlycanDock to generate a model of the *Fp*DeAc–A-antigen complex, providing a more relevant structural model to further investigate *Fp*GalNAcDeAc's enzymatic mechanism. In the case of *Fp*GalNase, which did not have an experimental model available at the time of my doctoral work, I developed a hypothesis supported by structure- and sequence-based analysis regarding the active-site residues important for *Fp*GH36's unique carbohydrate specificity. If future experimental studies were to result in support of my hypothesis, then the detailed analyses reported in this dissertation will serve as a blueprint toward modifying *Fp*GalNase's carbohydrate specificity toward the B-antigen, thus enabling complete A- and B-type blood conversion to the universal O-type.

4.2 Future directions

There is much left to do in the field of computational glycoscience. While improved conformational sampling techniques are valuable, the field would benefit most from technical advancements that enhance the ability of scoring functions (Rosetta-based or otherwise) to faithfully distinguish the biologically relevant protein–carbohydrate conformations among all docked models. Most current scoring functions have been optimized on proteins, meaning they are likely not adequately capturing the complexities of protein–carbohydrate interactions. Water molecules play a critical role in carbohydrate binding[36,46,107,138], from general solvent effects to mediating interfacial hydrogen bonding, yet the Rosetta scoring function only implicitly considers solvent. Similarly, CH–π stacking[92,139] (an aromatic-mediated interaction at the electron orbital level) is an important interaction that is essentially all but ignored at this computational scale.

The primary focus of continued computational work on the *Fp*GalNAcDeAc and *Fp*GalNase systems should be on design. At the time of writing this dissertation, an experimental model of *Fp*GalNase was resolved through X-ray crystallography, but not publicly released. Upon availability of this structure, the first step should be to employ GlycanDock to predict a model of the catalytic *Fp*GH36 domain bound to the GalN- and B-antigens. Understanding the residue-level interactions that drive GalN-antigen specificity will inform rational design toward favoring the B-antigen. However, it is possible, given the uniqueness of the *Fp*GH36 binding site, that an entire active-site redesign may be necessary to achieve B-type specific conversion. Ultimately, Rahfeld and Withers seek to employ *Fp*GalNAcDeAc and *Fp*GalNase toward producing universally accepted organs (as the ABO carbohydrate blood antigens, among others, are also present on cell surface of organs such as the heart and liver and thus contribute to organ transplant rejection)[110,111,140–142]. Again, protein design will play a vital role in achieving this ambitious yet exciting goal. Future work should start by combining GlycanDock and Rosetta design tools to engineer the enzymes' carbohydrate-binding domains to attach to surface glycans that are more prevalent on the desired cell type[143].

4.3 Parting thoughts

Undoubtedly, molecular modeling tools have been and continue to be instrumental in driving our understanding of complex biomolecular systems. Recently, the rise of employing machine learning techniques on biological data has been quick and the results are promising. AlphaFold, for example, is a modern-day revolution–enabling anyone with access to sufficient computing power to generate a protein structural model from sequence with impressive accuracy. While the necessary data to train an equivalently powerful tool for carbohydrate or protein–carbohydrate modeling is far from available, it is an important direction to consider and strive toward. Already, my colleagues in the Gray lab have developed a machine learning based tool for carbohydrate binding site prediction given a protein structure[144]. These strides forward are only made possible by the diligent work of interdisciplinary scientists and, as always, the generation and open sharing of more experimental data.

References

1. Seeberger, P. H. Monosaccharide Diversity. *Essentials Glycobiol.* (2022) doi:10.1101/GLYCOBIOLOGY.4E.2.
2. Lebrilla, C. B., Liu, J., Widmalm, G. & Prestegard, J. H. Oligosaccharides and Polysaccharides. (2022) doi:10.1101/GLYCOBIOLOGY.4E.3.
3. Varki, A. Biological roles of glycans. *Glycobiology* **27**, 3–49 (2017).
4. Dean, L. Blood Groups and Red Cell Antigens. in *The ABO blood group* (2005). doi:10.1160/TH04-04-0251.
5. Taylor, M. E. *et al. Discovery and Classification of Glycan-Binding Proteins*. *Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, 2015). doi:10.1101/GLYCOBIOLOGY.4E.28.
6. Leaver-Fay, A. *et al.* Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. in *Methods in Enzymology* vol. 487 545 (NIH Public Access, 2011).
7. Leman, J. K. *et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks*. *Nature Methods* vol. 17 665 (NIH Public Access, 2020).
8. Labonte, J. W., Adolf-Bryfogle, J., Schief, W. R. & Gray, J. J. Residue-centric modeling and design of saccharide and glycoconjugate structures. *J. Comput. Chem.* **38**, 276–287 (2017).
9. Li, Z. & Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 6611 (1987).
10. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science (80-.).*

181, 223–230 (1973).

11. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

12. Nivedha, A. K., Makeneni, S., Foley, B. L., Tessier, M. B. & Woods, R. J. Importance of ligand conformational energies in carbohydrate docking: Sorting the wheat from the chaff. *J. Comput. Chem.* **35**, 526–39 (2014).

13. Nivedha, A. K., Thieker, D. F., Makeneni, S., Hu, H. & Woods, R. J. Vina-Carb: Improving Glycosidic Angles during Carbohydrate Docking. *J. Chem. Theory Comput.* **12**, 892–901 (2016).

14. Werz, D. B. *et al.* Exploring the structural diversity of mammalian carbohydrates ('glycospace') by statistical databank analysis. *ACS Chem. Biol.* **2**, 685–691 (2007).

15. Ajit, V. & Sharon, N. Chapter 1 Historical Background and Overview. *Essentials Glycobiol.* 1–21 (2009) doi:10.1101/glycobiology.3e.001.

16. Lauc, G., Krištić, J. & Zoldoš, V. Glycans - the third revolution in evolution. *Front. Genet.* (2014) doi:10.3389/fgene.2014.00145.

17. Corfield, A. P. & Berry, M. Glycan variation and evolution in the eukaryotes. *Trends in Biochemical Sciences* at https://doi.org/10.1016/j.tibs.2015.04.004 (2015).

18. Eichler, J. & Koomey, M. Sweet New Roles for Protein Glycosylation in Prokaryotes. *Trends in Microbiology* at https://doi.org/10.1016/j.tim.2017.03.001 (2017).

19. Schmidt, M. A., Riley, L. W. & Benz, I. Sweet new world: Glycoproteins in bacterial pathogens. *Trends in Microbiology* at https://doi.org/10.1016/j.tim.2003.10.004 (2003).

20. Jarrell, K. F. *et al.* N-Linked Glycosylation in Archaea: a Structural, Functional, and Genetic Analysis. *Microbiol. Mol. Biol. Rev.* (2014) doi:10.1128/mmbr.00052-13.

21. van Kooyk, Y. & Rabinovich, G. A. Protein-glycan interactions in the control of innate and adaptive immune responses. *Nature Immunology* at https://doi.org/10.1038/ni.f.203 (2008).

22. Watanabe, Y., Bowden, T. A., Wilson, I. A. & Crispin, M. Exploitation of glycosylation in enveloped virus pathobiology. *Biochimica et Biophysica Acta - General Subjects* at https://doi.org/10.1016/j.bbagen.2019.05.012 (2019).

23. Vigerust, D. J. & Shepherd, V. L. Virus glycosylation: role in virulence and immune interactions. *Trends in Microbiology* vol. 15 211–218 at https://doi.org/10.1016/j.tim.2007.03.003 (2007).

24. Weis, W. I. & Drickamer, K. Structural basis of lectin-carbohydrate recognition. *Annual Review of Biochemistry* at https://doi.org/10.1146/annurev.bi.65.070196.002301 (1996).

25. Sharon, N. & Lis, H. History of lectins: From hemagglutinins to biological recognition molecules. *Glycobiology* at https://doi.org/10.1093/glycob/cwh122 (2004).

26. Collins, B. E. & Paulson, J. C. Cell surface biology mediated by low affinity multivalent protein-glycan interactions. *Current Opinion in Chemical Biology* at https://doi.org/10.1016/j.cbpa.2004.10.004 (2004).

27. Imberty, A., Mitchell, E. P. & Wimmerová, M. Structural basis of high-affinity glycan recognition by bacterial and fungal lectins. *Current Opinion in Structural Biology* at https://doi.org/10.1016/j.sbi.2005.08.003 (2005).

28. Berman, H. M. *et al.* The Protein Data Bank (www.rcsb.org). *Nucleic Acids Res.*

(2000) doi:10.1093/nar/28.1.235.

29. De Meirelles, J. L. *et al.* Current Status of Carbohydrates Information in the Protein Data Bank. *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.9b00874.

30. Copoiu, L., Torres, P. H. M., Ascher, D. B., Blundell, T. L. & Malhotra, S. ProCarbDB: A database of carbohydrate-binding proteins. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkz860.

31. Buda, S., Nawój, M. & Mlynarski, J. Recent Advances in NMR Studies of Carbohydrates. in *Annual Reports on NMR Spectroscopy* (2016). doi:10.1016/bs.arnmr.2016.04.002.

32. Malhotra, S. & Ramsland, P. A. Editorial overview: Carbohydrates – structural glycobiology catches the wave of rapid progress. *Current Opinion in Structural Biology* at https://doi.org/10.1016/j.sbi.2020.04.004 (2020).

33. Imberty, A. Oligosaccharide structures: Theory versus experiment. *Curr. Opin. Struct. Biol.* (1997) doi:10.1016/S0959-440X(97)80069-3.

34. DeMarco, M. L. & Woods, R. J. Structural glycobiology: A game of snakes and ladders. *Glycobiology* at https://doi.org/10.1093/glycob/cwn026 (2008).

35. Woods, R. J. & Tessier, M. B. Computational glycoscience: characterizing the spatial and temporal properties of glycans and glycan-protein complexes. *Current Opinion in Structural Biology* vol. 20 575–583 at https://doi.org/10.1016/j.sbi.2010.07.005 (2010).

36. Woods, R. J. *Predicting the Structures of Glycans, Glycoproteins, and Their Complexes*. *Chemical Reviews* vol. 118 8005–8024 (American Chemical Society, 2018).

37. National Research Council (US). *Transforming glycoscience: A roadmap for the future*. (National Academies Press, 2012). doi:10.17226/13446.

38. Imberty, A., H. Prestegard, J. & Prestegard, J. H. *Structural Biology of Glycan Recognition*. *Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, 2015). doi:10.1101/GLYCOBIOLOGY.3E.030.

39. Xiong, X. *et al.* Force fields and scoring functions for carbohydrate simulation. *Carbohydrate Research* vol. 401 73–81 at https://doi.org/10.1016/j.carres.2014.10.028 (2015).

40. Copoiu, L. & Malhotra, S. The current structural glycome landscape and emerging technologies. *Current Opinion in Structural Biology* vol. 62 132–139 at https://doi.org/10.1016/j.sbi.2019.12.020 (2020).

41. Imberty, A., Hardman, K. D., Carver, J. P. & Perez, S. Molecular modelling of protein-carbohydrate interactions. Docking of monosaccharides in the binding site of concanavalin A. *Glycobiology* (1991) doi:10.1093/glycob/1.6.631.

42. Pérez, S., Meyer, C. & Imberty, A. Practical tools for molecular modeling of complex carbohydrates and their interactions with proteins. *Mol. Eng.* (1995) doi:10.1007/BF00999595.

43. Woods, R. J. Computational carbohydrate chemistry: What theoretical methods can tell us. *Glycoconjugate Journal* at https://doi.org/10.1023/A:1006984709892 (1998).

44. Rockey, W. M., Laederach, A. & Reilly, P. J. Automated docking of α-(1→4)- and α-(1→6)-linked glucosyl trisaccharides and maltopentaose into the soybean β-amylase active site. *Proteins Struct. Funct. Genet.* (2000) doi:10.1002/(SICI)1097-

0134(20000801)40:2<299::AID-PROT100>3.0.CO;2-G.

45. Fadda, E. & Woods, R. J. Molecular simulations of carbohydrates and protein-carbohydrate interactions: motivation, issues and prospects. *Drug Discov. Today* **15**, 596–609 (2010).

46. Sapay, N., Nurisso, A. & Imberty, A. Simulation of Carbohydrates, from Molecular Docking to Dynamics in Water. in 469–483 (Humana Press, Totowa, NJ, 2013). doi:10.1007/978-1-62703-017-5_18.

47. Grant, O. C. & Woods, R. J. *Recent advances in employing molecular modelling to determine the specificity of glycan-binding proteins*. *Current Opinion in Structural Biology* vol. 0 47 (NIH Public Access, 2014).

48. Vliegenhardt, J. F. G. *NMR Spectroscopy and Computer Modeling of Carbohydrates: Recent Advances*. *NMR Spectroscopy and Computer Modeling of Carbohydrates: Recent Advances* (2006).

49. Kirschner, K. N. *et al.* GLYCAM06: A generalizable biomolecular force field. carbohydrates. *J. Comput. Chem.* (2008) doi:10.1002/jcc.20820.

50. Bryce, R. A., Hillier, I. H. & Naismith, J. H. Carbohydrate-protein recognition: Molecular dynamics simulations and free energy analysis of oligosaccharide binding to Concanavalin A. *Biophys. J.* (2001) doi:10.1016/S0006-3495(01)75793-1.

51. Hadden, J. A., Tessier, M. B., Fadda, E. & Woods, R. J. Calculating binding free energies for protein–carbohydrate complexes. *Methods Mol. Biol.* (2015) doi:10.1007/978-1-4939-2343-4_26.

52. Koppisetty, C. A. K., Frank, M., Lyubartsev, A. P. & Nyholm, P. G. Binding energy

calculations for hevein-carbohydrate interactions using expanded ensemble molecular dynamics simulations. *J. Comput. Aided. Mol. Des.* (2015) doi:10.1007/s10822-014-9792-5.

53. Dror, R. O., Jensen, M. Ø., Borhani, D. W. & Shaw, D. E. Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *J. Gen. Physiol.* (2010) doi:10.1085/jgp.200910373.

54. Garrett M. Morris *et al.* AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* (2009) doi:10.1002/jcc.

55. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* (2009) doi:10.1002/jcc.21334.

56. Lang, P. T. *et al.* DOCK 6: Combining techniques to model RNA-small molecule complexes. *RNA* (2009) doi:10.1261/rna.1563609.

57. Claussen, H. *et al.* The FlexX Database Docking Environment - Rational Extraction of Receptor Based Pharmacophores. *Curr. Drug Discov. Technol.* (2006) doi:10.2174/1570163043484815.

58. Friesner, R. A. *et al.* Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* (2006) doi:10.1021/jm051256o.

59. Hartshorn, M. J. *et al.* Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* (2007) doi:10.1021/jm061277y.

60. Nurisso, A., Kozmon, S. & Imberty, A. Comparison of docking methods for carbohydrate binding in calcium-dependent lectins and prediction of the

carbohydrate binding mode to sea cucumber lectin CEL-III. *Mol. Simul.* **34**, 469–479 (2008).

61. Adam, J., Kříž, Z., Prokop, M., Wimmerová, M. & Koča, J. In silico mutagenesis and docking studies of Pseudomonas aeruginosa PA-IIL lectin - Predicting binding modes and energies. *J. Chem. Inf. Model.* **48**, 2234–2242 (2008).

62. Agostillo, M. *et al.* Molecular docking of carbohydrate ligands to antibodies: Structural validation against crystal structures. *J. Chem. Inf. Model.* **49**, 2749–2760 (2009).

63. Agostino, M., Yuriev, E. & Ramsland, P. A. A computational approach for exploring carbohydrate recognition by lectins in innate immunity. *Front. Immunol.* **2**, (2011).

64. Mishra, S. K., Adam, J., Wimmerová, M. & Koča, J. In silico mutagenesis and docking study of Ralstonia solanacearum RSL lectin: Performance of docking software to predict saccharide binding. *J. Chem. Inf. Model.* **52**, 1250–1261 (2012).

65. Agostino, M., Ramsland, P. A. & Yuriev, E. Docking of carbohydrates into protein binding sites. *Struct. Glycobiol.* 111–138 (2012).

66. Samsonov, S. A., Zacharias, M. & Chauvot de Beauchene, I. Modeling large protein–glycosaminoglycan complexes using a fragment-based approach. *J. Comput. Chem.* **40**, 1429–1439 (2019).

67. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H. & Meiler, J. Practically Useful: What the R OSETTA Protein Modeling Suite Can Do for You. *Biochemistry* (2010).

68. Bender, B. J. *et al.* Protocols for Molecular Modeling with Rosetta3 and RosettaScripts. *Biochemistry* (2016) doi:10.1021/acs.biochem.6b00444.

69. Combs, S. A. *et al.* Small-molecule ligand docking into comparative models with Rosetta. *Nat. Protoc.* **8**, 1277–1298 (2013).

70. DeLuca, S., Khar, K. & Meiler, J. Fully flexible docking of medium sized ligand libraries with rosettaligand. *PLoS One* **10**, 1–19 (2015).

71. Bolia, A. *et al.* A flexible docking scheme efficiently captures the energetics of glycan-cyanovirin binding. *Biophys. J.* **106**, 1142–51 (2014).

72. Raveh, B., London, N., Zimmerman, L. & Schueler-Furman, O. Rosetta FlexPepDock ab-initio: Simultaneous Folding, Docking and Refinement of Peptides onto Their Receptors. *PLoS One* **6**, e18934 (2011).

73. Raveh, B., London, N. & Schueler-Furman, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. *Proteins Struct. Funct. Bioinforma.* **78**, NA-NA (2010).

74. Nance, M. L., Labonte, J. W., Adolf-Bryfogle, J. & Gray, J. J. Development and Evaluation of GlycanDock: A Protein-Glycoligand Docking Refinement Algorithm in Rosetta. *J. Phys. Chem. B* **125**, 6807–6820 (2021).

75. Frenz, B. *et al.* Automatically Fixing Errors in Glycoprotein Structures with Rosetta. *Structure* **27**, 134-139.e3 (2019).

76. Chaudhury, S. *et al.* Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2. *PLoS One* **6**, e22477 (2011).

77. Marze, N. A., Roy Burman, S. S., Sheffler, W. & Gray, J. J. Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics* **34**, 3461–3469 (2018).

78. Reback, J. *et al.* pandas-dev/pandas: Pandas 1.0.3. at

https://doi.org/10.5281/zenodo.3715232 (2020).

79. *The PyMOL Molecular Graphics System, Version 2.3.5 Schrödinger, LLC.*

80. Kortemme, T., Morozov, A. V. & Baker, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* (2003) doi:10.1016/S0022-2836(03)00021-4.

81. Bonnardel, F. *et al.* Unilectin3d, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. *Nucleic Acids Res.* **47**, D1236–D1244 (2019).

82. Yuriev, E., Agostino, M. & Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.* **24**, 149–164 (2011).

83. Harmalkar, A. & Gray, J. J. Advances to tackle backbone flexibility in protein docking. *Current Opinion in Structural Biology* vol. 67 178–186 at https://doi.org/10.1016/j.sbi.2020.11.011 (2021).

84. Cummings, R. D., Schnaar, R. L., Esko, J. D., Drickamer, K. & Taylor, M. E. *Principles of Glycan Recognition*. *Essentials of Glycobiology* (2015).

85. Kosik, I. & Yewdell, J. W. Influenza hemagglutinin and neuraminidase: Yin–yang proteins coevolving to thwart immunity. *Viruses* at https://doi.org/10.3390/v11040346 (2019).

86. O'Meara, M. J. *et al.* Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J. Chem. Theory Comput.* (2015) doi:10.1021/ct500864r.

87. Waskom, M. & the seaborn development team. mwaskom/seaborn. at

https://doi.org/10.5281/zenodo.592845 (2020).

88. Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. Carbohydrate-binding modules: Fine-tuning polysaccharide recognition. *Biochemical Journal* vol. 382 769–781 at https://doi.org/10.1042/BJ20040892 (2004).

89. Shoseyov, O., Shani, Z. & Levy, I. Carbohydrate Binding Modules: Biochemical Properties and Novel Applications. *Microbiol. Mol. Biol. Rev.* (2006) doi:10.1128/mmbr.00028-05.

90. The CAZypedia Consortium. Ten years of CAZypedia: A living encyclopedia of carbohydrate-active enzymes. *Glycobiology* (2018) doi:10.1093/glycob/cwx089.

91. Pires, V. M. R. *et al.* The crystal structure of the family 6 carbohydrate binding module from Cellvibrio mixtus endoglucanase 5A in complex with oligosaccharides reveals two distinct binding sites with different ligand specificities. *J. Biol. Chem.* (2004) doi:10.1074/jbc.M401599200.

92. Spiwok, V. CH/π Interactions in Carbohydrate Recognition. *Molecules* **22**, (2017).

93. Brenke, R. *et al.* Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **25**, 621–627 (2009).

94. Zhao, J., Cao, Y. & Zhang, L. Exploring the computational methods for protein-ligand binding site prediction. *Computational and Structural Biotechnology Journal* vol. 18 417–426 at https://doi.org/10.1016/j.csbj.2020.02.008 (2020).

95. Zhao, H., Taherzadeh, G., Zhou, Y. & Yang, Y. Computational Prediction of Carbohydrate-Binding Proteins and Binding Sites. *Curr. Protoc. Protein Sci.* **94**, e75 (2018).

96. Gattani, S., Mishra, A. & Hoque, M. T. StackCBPred: A stacking based prediction

of protein-carbohydrate binding sites from sequence. *Carbohydr. Res.* (2019) doi:10.1016/j.carres.2019.107857.

97. Kozakov, D. *et al.* The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* **10**, 733–755 (2015).

98. Gagneux, P., Aebi, M. & Varki, A. Evolution of Glycan Diversity. in *Essentials of Glycobiology* (2015). doi:10.1101/glycobiology.3e.020.

99. Muraki, M., Morikawa, M., Jigami, Y. & Tanaka, H. The roles of conserved aromatic amino-acid residues in the active site of human lysozyme: a site-specific mutagenesis study. *Biochim. Biophys. Acta (BBA)/Protein Struct. Mol.* (1987) doi:10.1016/0167-4838(87)90211-1.

100. Fernández-Alonso, M. D. C., Cañada, F. J., Jiménez-Barbero, J. & Cuevas, G. Molecular recognition of saccharides by proteins. Insights on the origin of the carbohydrate-aromatic interactions. *J. Am. Chem. Soc.* (2005) doi:10.1021/ja051020+.

101. Kerzmann, A., Neumann, D. & Kohlbacher, O. SLICK - Scoring and energy functions for protein - Carbohydrate interactions. *J. Chem. Inf. Model.* (2006) doi:10.1021/ci050422y.

102. Kerzmann, A., Fuhrmann, J., Kohlbacher, O. & Neumann, D. BALLDock/SLICK: A new method for protein-carbohydrate docking. *J. Chem. Inf. Model.* **48**, 1616–1625 (2008).

103. Hsu, C.-H. *et al.* The Dependence of Carbohydrate–Aromatic Interaction Strengths on the Structure of the Carbohydrate. *J. Am. Chem. Soc.* **138**, 7636–7648 (2016).

104. Stanković, I. M., Blagojević Filipović, J. P. & Zarić, S. D. Carbohydrate – Protein aromatic ring interactions beyond CH/π interactions: A Protein Data Bank survey and quantum chemical calculations. *Int. J. Biol. Macromol.* **157**, 1–9 (2020).

105. Tschampel, S. M. & Woods, R. J. Quantifying the Role of Water in Protein−Carbohydrate Interactions. (2003) doi:10.1021/JP035027U.

106. Nurisso, A. *et al.* Role of water molecules in structure and energetics of Pseudomonas aeruginosa lectin I interacting with disaccharides. *J. Biol. Chem.* (2010) doi:10.1074/jbc.M110.108340.

107. Ruvinsky, A. M. *et al.* The Role of Bridging Water and Hydrogen Bonding as Key Determinants of Noncovalent Protein–Carbohydrate Recognition. *ChemMedChem* **13**, 2684–2693 (2018).

108. Daniels, G. The molecular definition of red cell antigens. *ISBT Sci. Ser.* **5**, 300–302 (2010).

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

110. Rahfeld, P. *et al.* An enzymatic pathway in the human gut microbiome that converts A to universal O type blood. *Nat. Microbiol.* **4**, 1475–1485 (2019).

111. Rahfeld, P. & Withers, S. G. Toward universal donor blood: Enzymatic conversion of A and B to O type. *Journal of Biological Chemistry* vol. 295 325–334 at https://doi.org/10.1074/jbc.REV119.008164 (2020).

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

113. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the

structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

114. Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).

115. Abbott, D. W., Eirín-López, J. M. & Boraston, A. B. Insight into Ligand Diversity and Novel Biological Roles for Family 32 Carbohydrate-Binding Modules. *Mol. Biol. Evol.* **25**, 155–167 (2008).

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

117. Krivák, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminform.* **10**, 1–12 (2018).

118. Boraston, A. B., Ficko-Blean, E. & Healey, M. Carbohydrate recognition by a large sialidase toxin from Clostridium perfringens. *Biochemistry* **46**, 11352–11360 (2007).

119. Tyka, M. D. *et al.* Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405**, 607 (2011).

120. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55 (2014).

121. Bürger, M. & Chory, J. Structural and chemical biology of deacetylases for carbohydrates, proteins, small molecules and histones. *Commun. Biol.* **1**, (2018).

122. Asensio, J. L., Ardá, A., Cañada, F. J. & Jiménez-Barbero, J. Carbohydrate-

aromatic interactions. *Acc. Chem. Res.* (2013) doi:10.1021/ar300024d.

123. Zhang, C., Zheng, W., Mortuza, S. M., Li, Y. & Zhang, Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **36**, 2105–2112 (2020).

124. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 1–15 (2019).

125. Goldenzweig, A. *et al.* Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **63**, 337 (2016).

126. Park, H. *et al.* Simultaneous optimization of biomolecular energy function on features from small molecules and macromolecules. *J. Chem. Theory Comput.* **12**, 6201 (2016).

127. Frenz, B. *et al.* Prediction of protein mutational free energy: benchmark and sampling improvements increase classification accuracy. *Front Bioeng Biotechnol.* **8**, (2020).

128. Glycoside Hydrolase Family 36 - CAZypedia. https://www.cazypedia.org/index.php/Glycoside_Hydrolase_Family_36.

129. Comfort, D. A. *et al.* Biochemical analysis of Thermotoga maritima GH36 α-galactosidase (TmGalA) confirms the mechanistic commonality of clan GH-D glycoside hydrolases. *Biochemistry* **46**, 3319–3330 (2007).

130. Information on EC 3.2.1.22 - alpha-galactosidase - BRENDA Enzyme Database. https://www.brenda-enzymes.org/enzyme.php?ecno=3.2.1.22.

131. Fujimoto, Z., Kaneko, S., Momma, M., Kobayashi, H. & Mizuno, H. Crystal structure of rice alpha-galactosidase complexed with D-galactose. *J. Biol. Chem.* **278**,

20313–20318 (2003).

132. Golubev, A. M. *et al.* Crystal structure of α-galactosidase from Trichoderma reesei and its complex with galactose: Implications for catalytic mechanism. *J. Mol. Biol.* **339**, 413–422 (2004).

133. Fredslund, F. *et al.* Crystal Structure of α-Galactosidase from Lactobacillus acidophilus NCFM: Insight into Tetramer Formation and Substrate Binding. *J. Mol. Biol.* **412**, 466–480 (2011).

134. Brouns, S. J. J. *et al.* Identification of a novel α-galactosidase from the hyperthermophilic archaeon Sulfolobus solfataricus. *J. Bacteriol.* **188**, (2006).

135. Pace, N. J. & Weerapana, E. Zinc-Binding Cysteines: Diverse Functions and Structural Motifs. *Biomolecules* **4**, 419 (2014).

136. McGregor, N. G. S. *et al.* Cysteine Nucleophiles in Glycosidase Catalysis: Application of a Covalent β-l-Arabinofuranosidase Inhibitor. *Angew. Chemie Int. Ed.* **60**, 5754–5758 (2021).

137. Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nat. 2021 5967873* **596**, 590–596 (2021).

138. Kadirvelraj, R., Foley, B. L., Dyekjær, J. D. & Woods, R. J. Involvement of water in carbohydrate-protein binding: Concanavalin A revisited. *J. Am. Chem. Soc.* (2008) doi:10.1021/ja8039663.

139. Hudson, K. L. *et al.* Carbohydrate–Aromatic Interactions in Proteins. *J. Am. Chem. Soc.* **137**, 15152–15160 (2015).

140. Wang, A. *et al.* Developing Universal Blood Type Donor Lungs Using Ex Vivo ABO Enzymatic Treatment. *J. Heart Lung Transplant.* **39**, S69–S70 (2020).

141. MacMillan, S. *et al.* O042 Enzymatic conversion of human blood group A kidneys to universal blood group O. *Br. J. Surg.* **110**, (2023).

142. Erickson, T. *et al.* (731) Enzymatic Removal of A-Antigen in a Mouse Model of ABO-Incompatible (ABOi) Transplantation. *J. Hear. Lung Transplant.* **42**, S323 (2023).

143. Itakura, Y., Sasaki, N. & Toyoda, M. Glycan characteristics of human heart constituent cells maintaining organ function: relatively stable glycan profiles in cellular senescence. *Biogerontology* **22**, 623 (2021).

144. Canner, S. W., Shanker, S. & Gray, J. J. Structure-based neural network protein–carbohydrate interaction predictions at the residue level. *Front. Bioinforma.* **3**, 1186531 (2023).