Towards prediction of N-glycan compositions from atomic structural data

Haroldas Bagdonas

Doctor of Philosophy

University of York

Chemistry

September 2023

Abstract

Glycobiology, the study of saccharides and their biological significance, delves into understanding glycans, oligosaccharides that form essential structures in various living organisms. However, these glycans, covalently linked to proteins or lipids, possess a structural complexity that exceeds that of nucleic acids and proteins, attributed to their non-templated assembly. This complexity, characterised by diverse linkage positions, degrees of branching, and isomerism, facilitates glycans' multifaceted roles, including cell-cell recognition, immune response, and protein function optimization.

Structural Biology is one of the fields concerned with the study of glycobiology, however current model-building software leans heavily towards proteins. A major hurdle is the absence of upfront knowledge of glycan compositions at glycosylation sites. While protein sequences are easily derived from DNA, glycan sequences are not directly encoded in genomes. As a result of these challenges, many modelled *N*-glycan chains in glycoproteins show errors as featured in numerous communications and remediation efforts. Therefore, part of the thesis was devoted to implementing a software solution that would enable scientists building atomic models of glycoproteins to easily access information retrieved from glycoproteomic studies. The new code, implemented as part of the Privateer carbohydrate model validation and analysis software, was demonstrated to be useful in validation of modelled *N*-glycan compositions during iterative model building.

Following the successful bridging of atomic coordinates and glycoproteomic data, the research pivoted to assess the interplay between amino acid identities and *N*-glycan composition. Limited data indicated a potential relationship, especially with aromatic amino acids. Thankfully, the advent of AlphaFold motivated the implementation of a grafting algorithm in the Privateer software, responsible for transplanting *N*-glycan atomic coordinates, therefore enabling the expansion of *N*-glycan atomic structure data. The development of new software tools enabled the discovery of potentially meaningful discriminatory relationships in terms of neighbouring amino acid chemical properties and the *N*-glycan processing products.

List of Contents

List of Tables

List of Figures

medial and *trans*. In Golgi, the nascently processed glycoprotein may be acted upon numerous glycosyltransferases to produce a variety of potential glycoforms due to nascent glycoprotein exit from Golgi at any time during processing. Solid arrows represent the potential biosynthesis product of complex, biantennary, fucosylated glycan - $GlcNAc_2Man_3GlcNAc_2Gal_2Neu5Ac_2$. Dashed arrows represent alternative biosynthesis pathways that could either lead to Pauci-mannose glycan (Hex) or complex, triantennary, fucosylated glycan (GnT3).

Figure 1.6: Comparison of *N*-glycan features in density maps over a range of resolutions. (a)–(c) Electron density maps obtained with X-ray crystallography (MX). (d)–(f) Electronic potential maps obtained with cryo-EM; PDB codes and data resolution have been annotated directly on the figure. In the MX cases (a)–(c), at high resolution (a) it is possible to identify monosaccharides and their ring conformation from the density map; at medium resolution (b), ring conformation becomes difficult to determine, whereas at low resolution (c), and indeed with many cryo-EM maps (d)–(f), density associated with *N*-glycans have poorly defined discernible features of individual carbohydrates. 50

Figure 1.7: Visualisation of modelled glycans at ASN396 and ASN513 in human uromodulin (PDB: 7PFP) against the associated EM density map (brown chicken wire mesh). Some modelled carbohydrates are modelled outside the density map, due to lack of associated signal. The EM density map in the figure is rendered at 0.008V (2.5σ) contour level. Contour level recommended by the authors is 0.006V (2.0σ), according to the metadata deposited in EMDB entry: EMD-13378[109]. 51

Figure 2.1: Comparison of glycan features in electron density maps over a range of resolutions from select glycoprotein structures (PDB entries: 6RI6[7]; 6MZK[8]; 4O5I[9]) Electron Density maps obtained with X-Ray crystallography. Data resolution and PDB entry IDs associated with structures have been directly annotated on the figure. Left - depicts a high resolution example, where monosaccharides and their conformations can be elucidated; centre – a medium resolution example, where identification starts to become difficult; right – a low resolution example, for which all prior knowledge must be used. Despite coming from different glycoprotein structures, the glycan has the same composition and thus is assigned a unique GlyTouCan ID of G15407YE. 56

glycan and the glycoprotein were automatically eliminated from consideration in this study, due to the following reasons: 1) Glycoprotein was expressed in a fungal expression system, 2) modelled *N*-glycan likely contains potential modelling mistakes, specifically terminal mannose sugars being modelled as β-Man anomers, thus failing to return a match on GlyConnect database.

Acknowledgments

I would like to express my eternal gratitude for the opportunities that have been provided to me by multiple parties over the years. I would like to thank the University of York for providing a safe and engaging environment, where I was enabled to discover my passion for software engineering outside a Computer Science degree. I would like to thank Emma Rand for planting a seed in my pursuit to independently learn software development skills during my undergraduate studies through modules related to data science. Following the developing interest in computational structural bioinformatics, life-changing opportunities were provided by K Cowtan and Jon Agirre, who provided me with necessary infrastructure to further develop my skills.

I would like to thank my supervisors, Jon Agirre and Dani Ungar for excellent mentorship, guidance and supervision that allowed me to establish myself as a scientific researcher. The guidance of my mentors has allowed me to overcome numerous challenges that I encountered during my degree, both work and outside work.

Additionally, I would like to thank my fellow peers from Agirre, Ungar and Cowtan groups for effective collaborative efforts and fruitful discussions - Mihaela, Paul, Mateusz, Manal, Ben, Jordan, Thao, Lucy and Ali.

Finally, I would like to thank my family - my mum, Loreta, my dad, Auridas and my significant other, Ingrid, who have supported me through thick and thin, through my deep lows and highs to get to where I currently am.

Ačiū!
Haroldas

Declaration

I declare that this thesis is a presentation of original work, and I am the sole author, with the exception of the collaborative work listed below.

- Chapter 1, Figure 1.6: The figure is also published in Atanasova, M., Bagdonas, H. & Agirre, J. Structural glycobiology in the age of electron cryo-microscopy. *Curr. Opin. Struct. Biol.* **62**, 70–78 (2019).
    - HB produced the figure under the guidance of JA.
- Bagdonas, H., Ungar, D. & Agirre, J. Leveraging glycomics data in glycoprotein 3D structure validation with Privateer. *Beilstein J. Org. Chem.* **16**, 2523–2533 (2020).

  The paper was published as part of this PhD project and is reproduced in Chapter 2.

    - HB produced all the figures.
    - HB designed and implemented the software solution.
    - HB collected and analysed the data.
    - HB wrote the majority of the text with guidance and feedback from JA and DA.
- Bagdonas, H., Fogarty, C. A., Fadda, E. & Agirre, J. The case for post-predictional modifications in the AlphaFold Protein Structure Database. *Nat. Struct. Mol. Biol.* **28**, 869–870 (2021). The paper was published as part of this PhD project and is reproduced in Chapter 4, Section 4.4.1.1.

    - JA and EF supervised the collaborative effort.
    - CAF provided the template MD-equilibrated *N*-glycan structures.
    - HB designed and implemented the software solution to convert MD-equilibrated *N*-glycan structures into wwPDB compliant format.
    - HB designed and implemented the software solution to graft *N*-glycans onto input models.
    - EF produced Figure 4.1.
    - JA and EF wrote and edited the manuscript.
- The following papers include work produced during this PhD and are included in "Publications Released as part of this work" section:
    - Dialpuri, J. S. *et al.* Analysis and validation of overall N-glycan conformation in Privateer. *Acta Crystallogr D Struct Biol* **79**, 462–472 (2023).
        - HB provided mentorship to JSD in terms of inner workings of the Privateer software.
        - HB co-implemented torsional database into the Privateer software.

- HB implemented the display of torsional validation results in Symbol Nomenclature for Glycans (SNFG) generated by the Privateer software.
- JSD wrote the majority of the text with guidance and feedback from all co-authors.
- JSD generated and analysed the torsional linkage data.

- Agirre, J. *et al.* The CCP4 suite: integrative software for macromolecular crystallography. *Acta Crystallogr D Struct Biol* **79**, 449–461 (2023).
  - HB contributed by publishing the Privateer update to CCP4 and accompanying modifications to the Privateer GUI in CCP4i2.

COVID19 Pandemic Impact Statement

The COVID19 pandemic had a tangible impact on my PhD work. Even though I was able to successfully transition to working from home and carry out my research remotely, my mental health wellbeing was affected by lack of in-person interactions. This has resulted in difficulties in staying motivated and focused to an extent.

Nevertheless, the impact on my research progression was not as pronounced as that of my peers, due to not having to rely on wet lab experiments to begin with. This gave me a relative advantage in maintaining continuity in my work during the lockdowns and restrictions.

Introduction

1.1 Glycobiology

Glycobiology, the study of the structure, biosynthesis, and biology of saccharides (oligosaccharides), is an essential yet often underappreciated aspect of biochemistry and molecular biology. At the core of glycobiology are glycans, oligosaccharides that are found covalently linked to proteins (forming glycoproteins) or lipids (forming glycolipids). Glycans can be found on the surface of all cells and many proteins in all living organisms. They are incredibly diverse, far surpassing nucleic acids and proteins in terms of structural complexity. This is because, unlike proteins and nucleic acids, which are linear polymers produced through a template-driven process, glycans are assembled in a non-templated manner. This absence of a template allows for the creation of numerous configurations that differ in terms of linkage positions, degree of branching and isomerism, giving rise to a multitude of possible structures[26].

Glycans perform a variety of roles essential to biological processes, and this diversity of function is due, in large part, to their structural complexity. Glycans are central to cell-cell communication, protein folding, immune response, and microbial pathogenesis, among many other functions. For instance, the glycans present on cell surfaces are critical for cell-cell recognition and adhesion[27], playing significant roles in embryonic development[28], immune response[29], and disease processes[30]. Additionally, glycans attached to proteins (glycoproteins) can influence the protein's folding, stability, and are critical for proper protein function[31,32].

Glycans are introduced into products through a co- or post-translational modification known as glycosylation, a biological mechanism where an oligosaccharide (glycan) is covalently attached to a functional group of another molecule, such as protein or lipid, in an en-bloc manner, particularly in the case of *N*-glycosylation[33]. Likewise, the glycosylation modification can also occur by covalent attachment of a single sugar residue to another molecule, such as protein or lipid, which may further be extended by additional sugar residues, particularly in the case of *O*-glycosylation. It occurs in specific regions of the cell, most commonly in the endoplasmic reticulum (ER) and the Golgi apparatus. The glycosylation modification is universally facilitated by glycan processing enzymes. On the other hand, covalent oligosaccharide attachment to proteins is also possible without glycan processing enzymes, in the process defined as glycation, usually occurring in the bloodstream, creating glycated

proteins[34]. The rate of glycation, in principle, is dependent upon the extent of ageing and disease progression in individuals, thus being less prevalent than glycosylation[35,36]. Therefore, glycosylation describes the process of oligosaccharide attachment to molecules such as protein or lipids through a covalent linkage, driven by glycan-processing enzymes.

1.2 Chemical properties of monosaccharides

The most basic building block of oligosaccharides are monosaccharides, known as sugars. Monosaccharides are defined as molecules, consisting predominantly of carbon, oxygen and hydrogen atoms, although there are examples of sugars containing other types of atoms. The number of carbon atoms in the backbone of the monosaccharide is used to classify sugars, such as trioses (three carbons), tetroses (four carbons), pentoses (five carbons), hexoses (six carbons), heptoses (seven carbons), octoses (eight carbons) and nonoses (nine carbons). Oligosaccharides involved in protein glycosylation are predominantly composed of hexose monosaccharides, with some notable exceptions such as xylose sugars being pentoses and sialic acid sugars being nonoses[37]. All monosaccharides contain a carbonyl functional group (C=O). If the carbonyl forms an aldehyde group at the end of the carbon backbone of the sugar, then it is named aldose. If the carbonyl is located inside of the carbon backbone of the sugar forming a ketone group, then it is named ketose.

In addition to the carbonyl functional group, sugars also contain multiple hydroxyl functional groups (-OH), typically one per each carbon in the backbone that does not already contain the carbonyl group. As a result, sugars have at least one or more chiral centres. The varying arrangement of hydroxyl groups around the chiral carbons leads to different stereoisomers of the same monosaccharide (Figure 1.1). As a result, Symbol Nomenclature for Glycans (SNFG) notation uses a defined colour mapping for specific sugar stereoisomers (Table 1.1)[38]. It is common to name sugar stereoisomers based on their stereodescriptor, "*D*-" and "*L*-". The "*D*-" and "*L*-" stereodescriptors rely on the chiral carbon most distant from the carbonyl group[39]. Visually, this can be described using Fischer projection, which is a two-dimensional representation of a molecule that shows the stereochemistry at each of the chiral centres. In the Fischer projection, the chiral carbon used to assign "*D*-" and "*L*-" configuration is the penultimate carbon. If the hydroxyl group on the penultimate carbon is on the right, then the sugar is labelled as "*D*-". Otherwise, the hydroxyl group on the left of configurational chiral carbon is labelled as "*L*-" (Figure 1.1). The two different configurations are enantiomers (mirror images) of each other, but *D*-sugars are the form that are most

commonly found in nature and most readily utilised in cellular metabolism. L-sugar metabolism is the exception, done in some bacteria and malignant tumour cells[40,41].

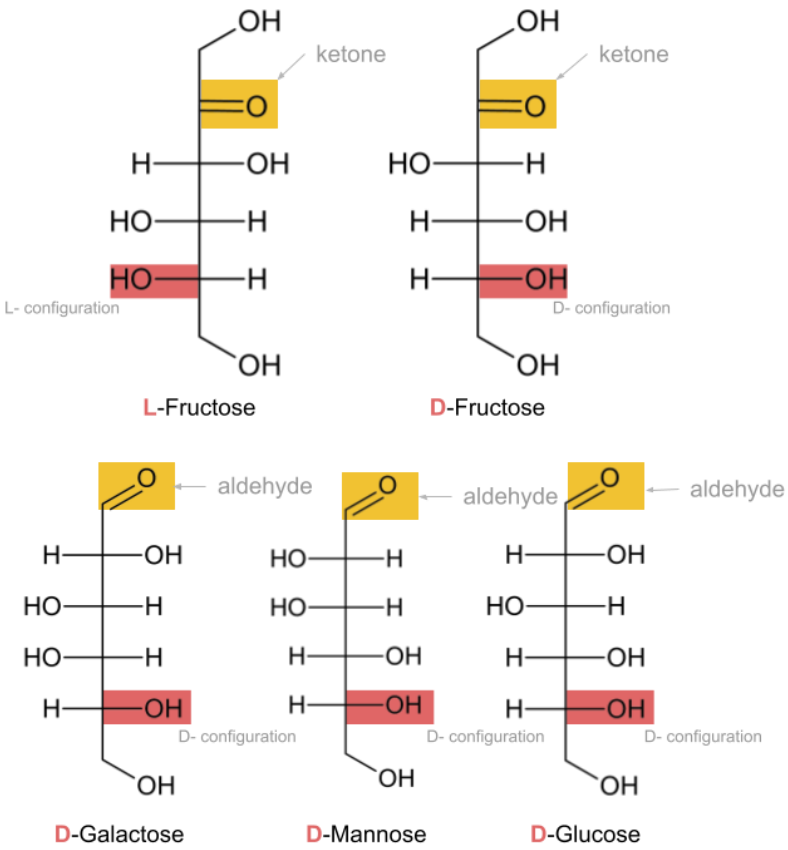


Figure 1.1: A selection of six-carbon (hexose) sugars, represented in Fischer projection. Ketose sugars have ketone functional groups, aldose sugars have aldehyde functional groups. Hydroxyl group position on the penultimate carbon (*L*- left, *D*- right) determines the configuration of the sugar.

While monosaccharides can exist in linear form, they often naturally occur in a cyclic form in aqueous solution due to the reversible reaction between the carbonyl group and one of the hydroxyl groups[42]. The reaction results in a cyclic ring structure, where the sugar either becomes a pyranose (six-membered ring) or a furanose (five-membered ring). Hexose sugars can be both pyranose and furanose, depending on which hydroxyl group performs the nucleophilic attack on the carbonyl. As shown in Figure 1.2, if the hydroxyl group on the fourth carbon (C4) of D-glucose in Fischer projection performs a nucleophilic attack on the carbonyl of aldehyde group at the first carbon (C1), then a *D*-glucofuranose product occurs. If the hydroxyl group on the penultimate carbon (C5) performs a nucleophilic attack on the carbonyl of the aldehyde group at the first carbon (C1), then a *D*-glucopyranose product occurs.

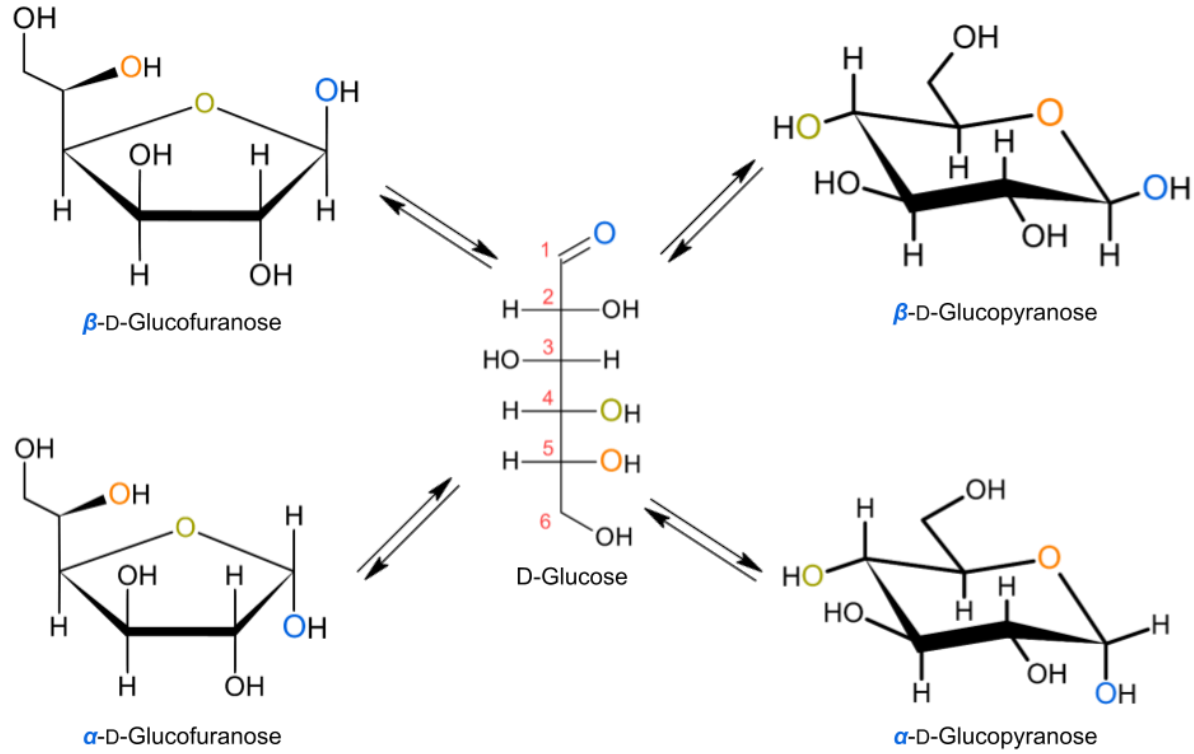


Figure 1.2: *D*-Glucose in linear and cyclic forms (pyranose and furanose). The oxygen shown in olive at C4 position performs a nucleophilic attack on the aldehyde at C1 to form *D*-Glucofuranose in either α or β anomeric configurations. The oxygen shown in orange at C5 position performs a nucleophilic attack on the aldehyde at C1 to form *D*-Glucopyranose in either α or β anomeric configurations.

In the process of forming the ring, as the carbonyl on either ketone or aldehyde group is planar, the nucleophilic attack can occur either from above or below. When the ring is formed, the first carbon (C1) becomes a new chiral centre, known as the anomeric carbon. The stereochemistry at the anomeric carbon creates an additional description of the sugar - whether it is the alpha (α) or beta (β) form of the monosaccharide, known as anomeric configuration (Figure 1.2). In the alpha form, the hydroxyl group on the anomeric carbon is on the opposite side (*trans*) of the ring from the CH_2OH group on fifth carbon (C5). In the beta form, the hydroxyl group on the anomeric carbon is on the same side (*cis*) of the ring as the CH_2OH group on penultimate carbon (C5).

1.3 Chemical properties of oligosaccharides

A minimum of two units of monosaccharides form an oligosaccharide. Monosaccharides within an oligosaccharide are linked together by glycosidic bonds. Glycosidic bonds are formed through an enzymatically-catalysed dehydration reaction where a hydrogen atom from the hydroxyl group of the anomeric carbon is removed from one monosaccharide and a hydroxyl group (OH) is removed from another sugar, allowing a bond to form between the two monosaccharides through the release of a water molecule. When the reaction is complete, a nascent disaccharide (an oligosaccharide composed of two monosaccharides) possesses the following properties: the disaccharides' non-reducing end is where the anomeric carbon is involved in a glycosidic bond and the disaccharides' reducing end is where the anomeric carbon of the terminal monosaccharide is free to form a pristine glycosidic bond. The formed glycosidic bond can be described in terms of its: stereochemistry - configuration of the anomeric carbons involved; regioselectivity - numerical identifiers of specific carbon atoms in proximity of the linkage; and the types of monosaccharides forming the bond.

A nascent oligosaccharide molecule may be significantly extended through the addition of more monosaccharides[43]. Oligosaccharides can be linear, with each monosaccharide connected to at most two other monosaccharides. However, each monosaccharide has multiple hydroxyl groups that can participate in glycosidic bond formation. As shown in Figure 1.3, unlike nucleic acid or protein biomolecules, oligosaccharides can be and are often branched, where a single monosaccharide is connected to three or four other monosaccharides. The ability to form branched structures significantly contributes to the complexity and diversity of oligosaccharide structures. Because each branch can be of different length and can be made up of different types of monosaccharides, there is a huge degree of diversity in glycan structures that can be constructed from a set of basic building blocks (summarised in Table 1.1).

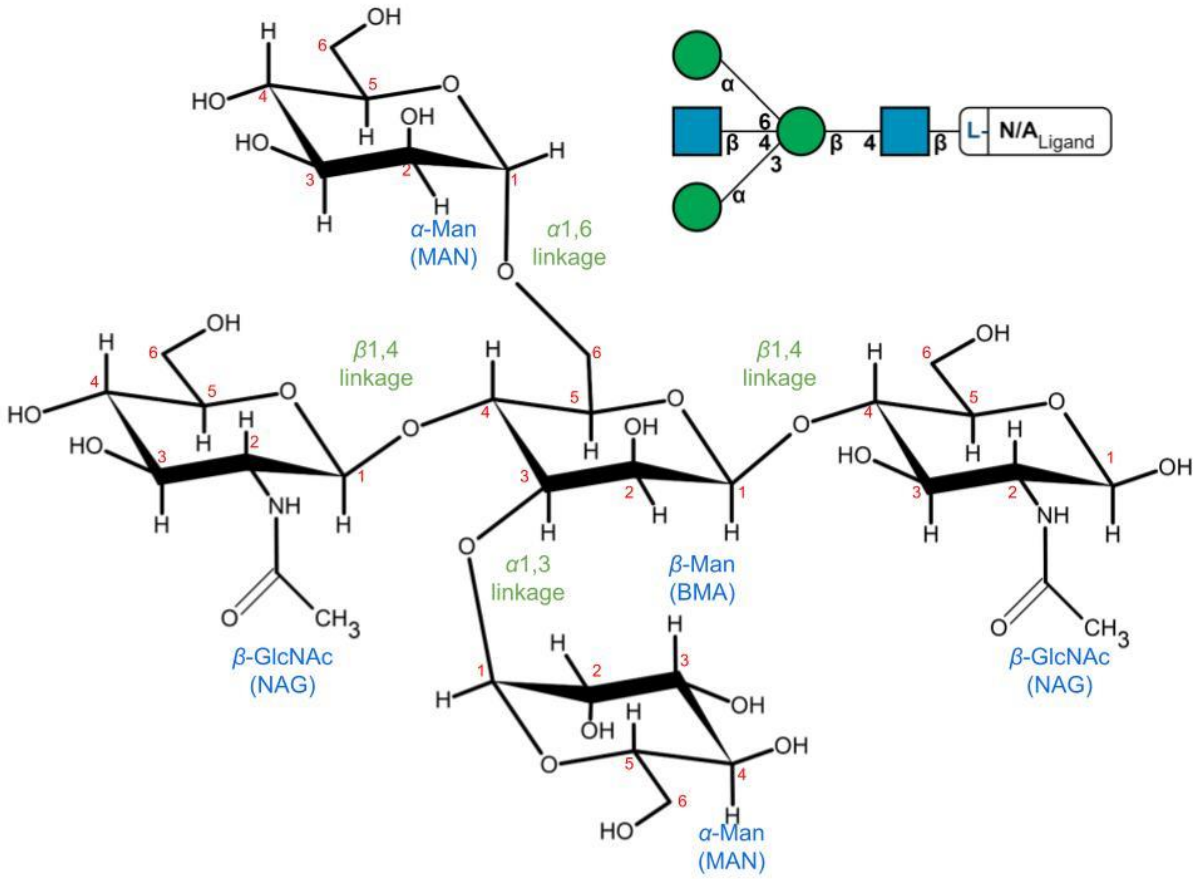


Figure 1.3: Branched oligosaccharide ligand composed of five monosaccharides representing a partial fragment of a core *N*-glycan with a bisecting GlcNAc. Upper right corner demonstrates the oligosaccharide description in Symbol Nomenclature for Glycans (SNFG) notation produced using Privateer[2]. Blue coloured labels represent anomeric configuration and monosaccharide names with their associated PDB three-letter residue identifier, green labels represent glycosidic linkage description. Red coloured labels represent carbon atom positions within each monosaccharide.

Table 1.1: Basic *N*-glycan building blocks of monosaccharides encountered in this thesis. Every row contains monosaccharide's common name and its abbreviation, PDB residue code associated with the monosaccharide, complete International Union of Pure and Applied Chemistry (IUPAC) name and SNFG symbol.

Name (abbreviation)	PDB residue code	IUPAC name	SNFG symbol
N-acetyl-beta-*D*-glucosamine (β-GlcNAc)	NAG	2-acetamido-2-deoxy-β-*D*-gluco-hexopyranose	
alpha-*D*-mannose (α-Man)	MAN	α-*D*-manno-hexopyranose	
beta-*D*-mannose (β-Man)	BMA	β-*D*-manno-hexopyranose	
beta-*D*-galactopyranose (β-Gal)	GAL	β-*D*-galacto-hexopyranose	
alpha-*D*-glucopyranose (α-Glc)	GLC	α-*D*-gluco-hexopyranose	
alpha-*L*-fucopyranose (α-Fuc)	FUC	6-deoxy-α-*L*-galacto-hexopyranose	
N-acetyl-alpha-neuraminic acid (α-Neu5Ac)	SIA	5-acetamido-3,5-dideoxy-*D*-glycero-α-*D*-galacto-non-2-ulopyranosonic acid	

1.4 Protein Glycosylation

Oligosaccharides can form a covalent bond with other biomolecules such as nucleic acids[44], protein and lipids[45]. A protein-oligosaccharide covalent bond, which is also referred to as glycosidic bond, is formed between a sugar and amino acid molecule. Depending on the identity of the amino acid, there can be multiple types of glycosidic bonds, which therefore directly correlate with the type of glycosylation. As a result, there are several types of protein glycosylation: *N*-linked glycosylation - where the oligosaccharide through an *N*-glycosidic linkage is attached to the nitrogen atom of the amide group of an asparagine residue in a protein; *O*-linked glycosylation - where the oligosaccharide through an *O*-glycosidic linkage is attached to the oxygen atom of hydroxyl group of serine, threonine, tyrosine, hydroxylysine, or hydroxyproline amino acid residues; *C*-linked glycosylation - where an α-mannose sugar through a *C*-glycosidic linkage is attached to the carbon, specifically CD1 atom, of a tryptophan amino acid[46]; and *P*-linked glycosylation - where the glycan is attached through a *P*-glycosidic linkage to the phosphorus atom of a phosphoserine residue[47]. After the formation of a glycosidic linkage, the oligosaccharide can alternatively be referred to as a glycan. Covalently-linked glycans are then typically modified either by extending the tree via the action of glycosyltransferases, or in the case of *N*-glycosylation, a combination of extension and fragment trimming is employed through the additional action of glycosidase enzymes. The final composition of a glycan is determined by a competition between numerous glycan processing enzymes that may compete for the same substrate in a non-templated process[48]. As a result, glycosylation can be described as a highly heterogeneous process. Therefore, factors such as the glycosylation site accessibility and expression levels of glycan processing enzymes, availability of sugar donors and the localization of enzymes within cellular organelles are significant determinants of glycan compositional homogeneity on separate molecules of the same protein.

N-linked and *O*-linked are the most prevalent types of glycosylation in eukaryotes. Both types of post-translational modifications are similar in the fact that they occur in the secretory pathway of the cell, which involve the translocation of nascently-synthesised glycoprotein through endoplasmic reticulum and Golgi apparatus organelles.

Nevertheless, there are significant differences between the two types of glycosylations. Besides *O*-linked glycosylation forming an *O*-glycosidic linkage, it primarily is initiated by enzymes residing in the Golgi apparatus, where sugar addition to the nascent *O*-glycosylation site occurs on a one at a time basis exclusively facilitated by glycosyltransferases[49]. On the other hand, *N*-linked glycosylation is facilitated by the action

of glycosidases and glycosyltransferases. A hallmark of *N*-linked glycan synthesis is the partial fragment trimming by glycosidase enzymes to enable action of glycosyltransferases later in the non-templated pathway[50]. The *N*-linked glycosylation process starts in the endoplasmic reticulum, where a 14-sugar-long intermediate glycan is attached to the nascent glycoprotein that is partially trimmed by mannosidases, with an eventual transfer to the Golgi apparatus for further and more elaborate processing[51]. As a result, *N*-linked glycans have a common pentasaccharide core that is not affected by glycosidase enzymes, unlike *O*-linked glycans that may have multiple core structures, due to the exclusive action of glycosyltransferases[52,53]. Finally, *O*-glycosylation does not have a consensus protein sequence, unlike *N*-glycosylation which has a consensus sequence of Asn-X-Ser/Thr (where X is any amino acid except proline)[54].

1.5 *N*-glycan biosynthesis pathway in mammalian expression systems

The biosynthesis of *N*-glycosylated products in mammalian cells starts with the synthesis of a dolichol-linked precursor oligosaccharide. The synthesis of the glycan is initiated by the activation of a dolichol phosphate intermediate, where *N*-acetylglucosamine (GlcNAc) is transferred from Uridine diphosphate *N*-acetylglucosamine (UDP-GlcNAc) to dolichol phosphate (Dol-P) embedded in the ER membrane[55]. This reaction is catalysed by UDP-GlcNAc:dolichol-P GlcNAc-1-P transferase (DPAGT1), resulting in the Dol-PP-GlcNAc intermediate[56]. Another addition of GlcNAc from UDP-GlcNAc is catalysed by ALG13/14 transferase to form the Dol-PP-$GlcNAc_2$ intermediate[57]. Next, multiple mannosyltransferase enzymes are involved to catalyse the addition of singular β-mannose and multiple α-mannose residues from GDP-Man donors to Dol-PP-$GlcNAc_2$ to eventually form Dol-PP-$GlcNAc_2Man_5$ intermediate[58]. The Dol-PP-$GlcNAc_2Man_5$ intermediate initiates the flipping of the precursor into the inside of ER lumen, which is catalysed by an ATP-independent flippase enzyme[59]. Inside the ER lumen, a further four α-mannose and three glucose (Glc) residues are added by mannosyltransferases and glucosyltransferases, respectively, to form Dol-PP-$GlcNAc_2Man_9Glc_3$ product. The necessary sugar donors, Dol-P-Man and Dol-P-Glc, are transferred to ER lumen by various other flippase enzymes[60].

Simultaneously, nascent protein being synthesised at the ribosome gets transferred to the ER lumen through the protein translocation channels, composed of membrane protein complexes[61]. When the potential glycosylation site is past the membrane barrier, the nascent $GlcNAc_2Man_9Glc_3$ glycan is transferred from Dol-PP to the nitrogen atom in the amide group

of the asparagine sidechain forming a *N*-glycosidic covalent bond in a reaction catalysed by oligosaccharyltransferase (OST) complex, where either of STT3A or STT3B subunit playing the role of a catalytic subunit[62]. STT3A subunit is involved in mediating the translocation of nascent polypeptide across the protein translocation channel and therefore is involved in co-translational *N*-glycosylation of the nascent polypeptide[63]. STT3B subunit on the other hand mediates the *N*-glycosylation of sites that are skipped by STT3A, in a post-translational process[33]. The nascent glycan plays an important role in aiding protein folding. Passively, the nascent glycan aids protein folding via the thermodynamic destabilisation of the unfolded state of the protein[31]. Furthermore, the sugar molecules increase the solubility of the glycoprotein, thereby preventing the hydrophobic regions of protein from aggregating in the ER[64]. Besides the passive mechanism, the ER in addition has an active quality control mechanism to monitor protein folding that relies on the glycosylation status of the nascent protein. Glucosidase enzymes, responsible for removing two residues of glucose from the nascent glycan, produce glycoproteins containing the $GlcNAc_2Man_9Glc_1$ intermediate. The intermediate glycan is recognized and bound by the molecular chaperones calnexin and calreticulin through their carbohydrate-binding domains. The chaperones recruit other proteins that facilitate protein folding through the formation of disulfide bridges. After a round of calnexin-calreticulin association, the terminal glucose residue is removed by Glucosidase-II. If the nascent glycoprotein failed to achieve proper folding, the hydrophobic patches and molten globule-like states initiate the recruitment of UDP-glucose glycoprotein glucosyltransferase (UGGT), which catalyses the re-addition of terminal glucose. Re-addition of terminal glucoses initiates an additional cycle of calnexin-calreticulin association[65]. If the protein fails to fold correctly after multiple cycles of calnexin-calreticulin association, then reglucosylation by UGGT is inhibited to initiate ER α1,2-mannosidase action[65,66]. The enzyme removes a terminal α-mannose residue from the middle antenna of nascent glycoprotein, producing $GlcNAc_2Man_8$ isomer that becomes a target for the ER associated degradation (ERAD) pathway[66].

After successful protein folding, glycoproteins exit the ER to travel to the Golgi apparatus, with attached *N*-glycans of high-mannose composition that undergo further processing in the Golgi. Final products of glycosylation machinery can be categorised into the following products: high-mannose, hybrid, complex and pauci-mannose, which are shown in Figure 1.4. All these products share a common $GlcNAc_2Man_3$ core. High-mannose *N*-glycans are exclusively composed of two GlcNAc and a varying number of α-mannose sugars in the range of 5-9. Hybrid *N*-glycans retain two α-mannose residues on the α1,6 arm while α1,3 at minimum contains a GlcNAc that can be further decorated with additional galactose and neuraminic acid residues. Complex *N*-glycan at minimum has the composition of

$GlcNAc_2Man_3GlcNAc_1$, where two α-mannose residues are removed from α1,6 arm and replaced with GlcNAc sugars that can be further decorated with additional galactose, neuraminic acid and fucose residues. The number of GlcNAc residues outside of the $GlcNAc_2Man_3$ core indicates the degree of branching, i.e., if there are three GlcNAc then the complex glycan is triantennary.

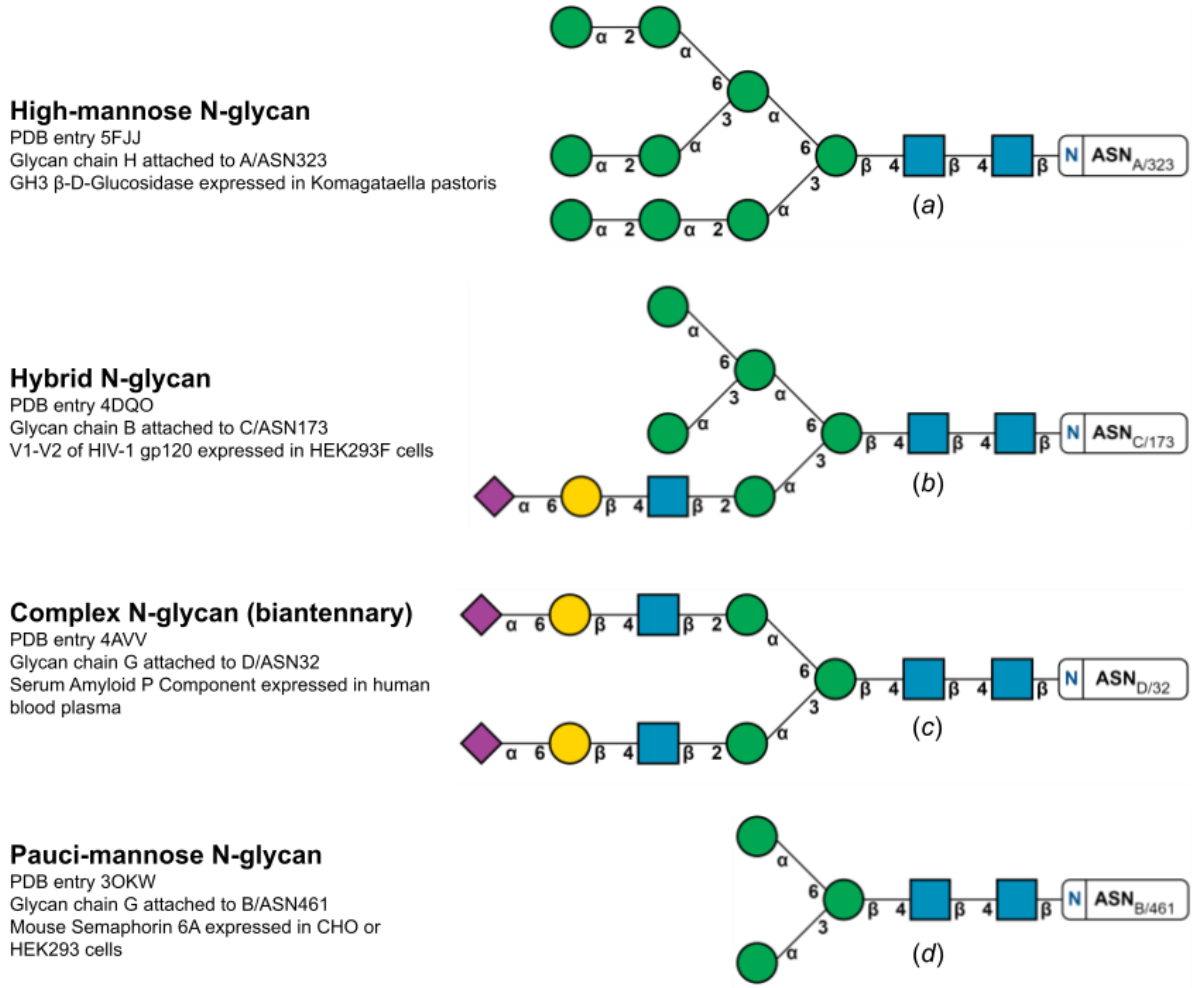


Figure 1.4: Classification of *N*-glycosylation machinery products with selected examples displayed in Symbol Nomenclature for Glycans (SNFG) notation. (a) High-mannose *N*-glycan modelled in 5FJJ[3]. (b) Hybrid *N*-glycan modelled in 4DQO[4]. (c) Complex biantennary *N*-glycan modelled in 4AVV[5]. (d) Pauci-mannose *N*-glycan modelled in 3OKW[6]. Even though the Pauci-mannose product is naturally occurring in the Golgi mammalian cells, it does appear as a product in other expression systems and is visualised as a separate class due to prevalence of Pauci-mannose *N*-glycans in structural biology experiments. The pictures of *N*-glycans in SNFG notation were generated by Privateer[2].

The Golgi apparatus organelle consists of multiple cisternae that can be broadly simplified into three compartments: *cis*, *medial* and *trans*[51]. A folded protein containing an immature high-mannose glycan from the endoplasmic reticulum enters the Golgi at the *cis* end. At the

cis-Golgi the glycan is acted upon by the mannosidase I (ManI) enzyme, which recursively trims immature high-mannose down to the $GlcNAc_2Man_5$ glycan. The nascent glycoprotein product can exit the Golgi at any point during recursive mannose trimming without any follow-up processing to simply contain a *N*-glycan of high-mannose composition. If the nascent high-mannose glycoprotein does not exit the Golgi, a GlcNAc residue can be added to the α1,3 linked α-mannose by alpha-1,3-mannosyl-glycoprotein 2-beta-*N*-acetylglucosaminyltransferase (GnTI) enzyme, resulting in the formation of a hybrid glycan, with the composition of $GlcNAc_2Man_5GlcNAc_1$. In order to convert hybrid to complex *N*-glycan, two additional α-mannose residues are trimmed in a reaction catalysed by Mannosidase II (ManII) to obtain the composition of $GlcNAc_2Man_3GlcNAc_1$. The simplest complex *N*-glycan product either exits the Golgi entirely or enters medial-Golgi for further processing[51]. Recent evidence suggests that in human cells the simplest complex glycan can be converted to a Pauci-mannose by *N*-acetyl-β-hexosaminidase (Hex) isoenzymes, which catalyse the release of capping GlcNAc residue to generate the $GlcNAc_2Man_3$ chitobiose core[67]. Alternatively, the simplest complex *N*-glycan can be further processed by the addition of a second GlcNAc onto α1,6 linked α-mannose by alpha-1,6-mannosyl-glycoprotein 2-beta-*N*-acetylglucosaminyltransferase (GnTII) enzyme to produce an *N*-glycan with the composition of $GlcNAc_2Man_3GlcNAc_2$. Complex *N*-glycans can further be processed by the addition of GlcNAc branches catalysed by respective GnTIV, GnTV enzymes to produce tri- and tetra- antennary glycans. A bisecting complex *N*-glycan can be produced by the addition of β1,4 linked GlcNAc to β1,4 linked β-mannose in the chitobiose core catalysed by GnTIII enzyme (partial fragment shown in Figure 1.3). Finally, in medial-Golgi, a fucose residue can be added to the first GlcNAc catalysed by fucosyltransferase 8 (FUT8) enzyme. Following the addition of branches, the nascent glycoprotein can either exit the Golgi entirely or be translocated to trans-Golgi, where GlcNAc caps could be further modified via the addition of galactose, neuraminic acid and fucose residues, catalysed by galactosyltransferase (GalT), sialyltransferase (SiaT), and other fucosyltransferase (FUT) enzymes[51]. A visual summary of the pathway is shown in Figure 1.5.

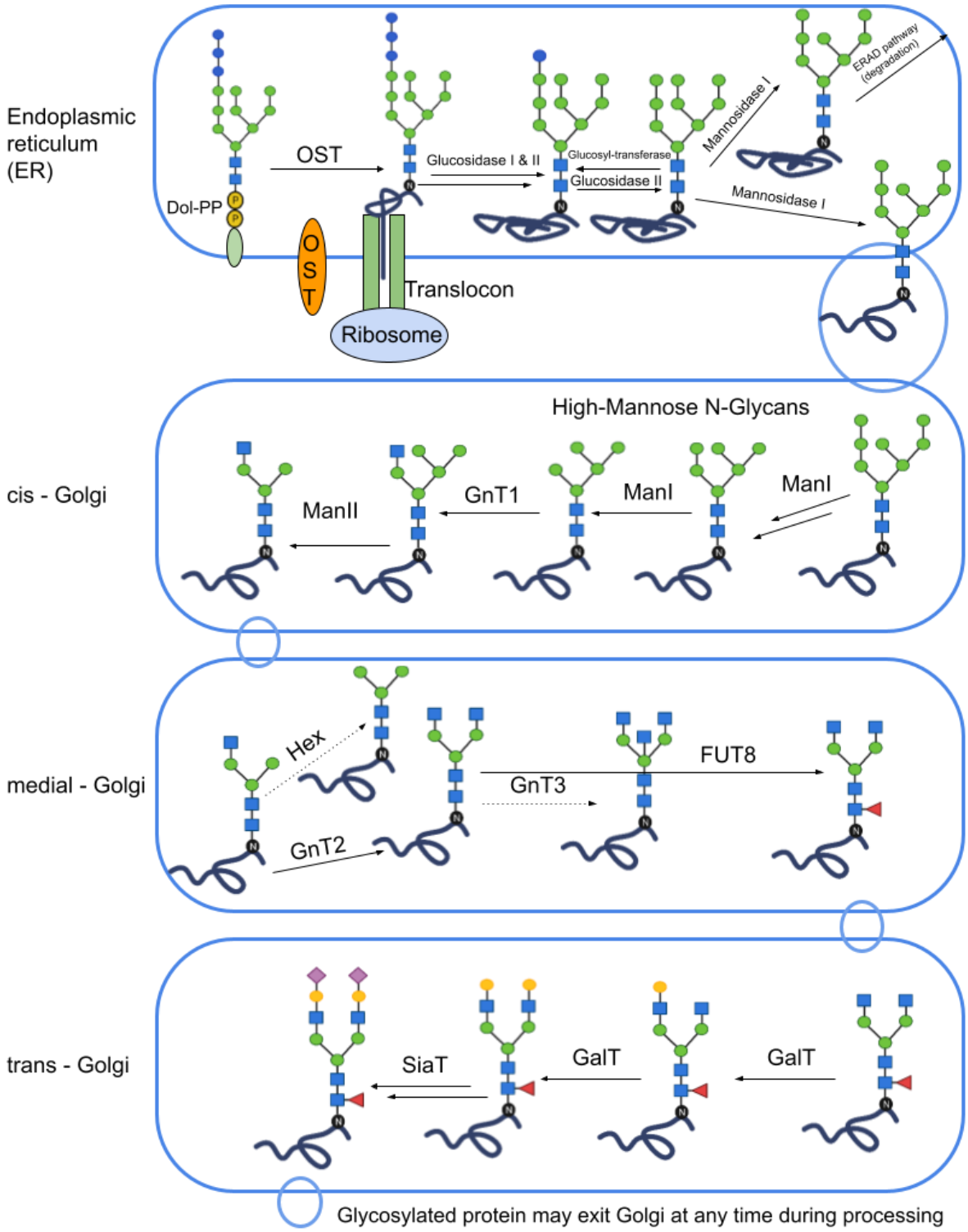


Figure 1.5: Schematic summary of *N*-Glycosylation pathway in mammalian expression systems. The simplified diagram represents oligosaccharide transfer to nascently synthesised protein in the ER, with glycan processing steps carried out in Golgi apparatus, organised into three cisternae: *cis*, *medial* and *trans*. In Golgi, the nascently processed glycoprotein may be acted upon numerous glycosyltransferases to produce a

variety of potential glycoforms due to nascent glycoprotein exit from Golgi at any time during processing. Solid arrows represent the potential biosynthesis product of complex, biantennary, fucosylated glycan - $GlcNAc_2Man_3GlcNAc_2Gal_2Neu5Ac_2$. Dashed arrows represent alternative biosynthesis pathways that could either lead to Pauci-mannose glycan (Hex) or complex, triantennary, fucosylated glycan (GnT3).

In summary, *N*-Glycosylation is a highly heterogeneous process largely due to the numerous glycan processing enzymes that may compete for an identical substrate with similar affinities. Therefore, it is typical for specific glycosylation sites across identical protein subunits to contain varying glycoforms. The outcome of *N*-glycan processing machinery would be completely random if it were not for the Golgi apparatus being organised into three different cisternae that distribute *N*-glycan processing enzymes in a sequential assembly line. In other words, the potential randomness is minimised, although *N*-glycosylation is still a relatively heterogeneous process. This thesis largely focuses on mammalian *N*-glycan biosynthesis as there are non-marginal differences in the pathway across different eukaryotic kingdoms. For example, *N*-glycosylation complexity in fungal expression systems trends towards producing mannan *N*-glycans versus *N*-glycosylation complexity in mammalian expression systems trending towards addition of capping motifs composed of GlcNAc, sialic acid, galactose and fucose sugars instead.

1.6 Glycoproteomics to study the effects of glycosylation at the cellular scale

Glycomics and glycoproteomics are two critical fields of study that focus on the identity of glycans and glycosylated proteins, respectively. The study of glycosylation at a cellular scale is necessary to understand the complex effects of protein glycosylation on cellular phenotypes. Glycomics focuses on understanding the complete picture of all glycans in a cell, tissue, or organism (the glycome). This includes identifying the types of glycans, their location, their relevant structural details, and how they vary between different phenotype states. Advances in this field have improved our ability to profile glycomes, leading to an enhanced understanding of the role of glycans in health and disease. Glycoproteomics, on the other hand, specifically looks at proteins that have been glycosylated. This involves identifying which proteins are glycosylated, where the glycosylation sites are on the protein

(site-specific glycosylation), and the structure of the attached glycans. It provides detailed information about the variety and abundance of glycoproteins within a cell or tissue.

Both glycomics and glycoproteomics rely on common techniques - predominantly mass spectrometry and various types of chromatography - to elucidate compositional details about oligosaccharide structures that decorate glycosylated proteins. Detailed analysis of glycan compositions requires the determination of individual monosaccharide units and glycosidic linkage descriptions. In practice, this is accomplished via the use of mass spectrometry (MS), where molecules and their fragments are ionised in the sample to obtain mass-to-charge (m/z) ratios that can be cross-referenced to the theoretical weight of fragmentation products. MS offers several advantages in the analysis of glycan compositions. MS is a sensitive technique that allows the detection and identification of even trace amounts of compounds[68]. This makes it possible to determine glycan composition in complex mixtures, where the concentrations of individual glycans might be very low. In combination with techniques such as tandem mass spectrometry (MS/MS), it can reveal the sequence, branching patterns, and linkages of monosaccharide residues in a glycan molecule[69]. Therefore, MS can provide detailed structural information about glycans. Fundamentally, MS is not a quantitative technique, but under certain configurations it can be used for relative quantification of specific fragment identities within samples, allowing for the analysis of the relative abundance of a target in the sample within the obtained profile. Compared to other techniques like nuclear magnetic resonance (NMR) spectroscopy, MS is relatively fast and allows for the analysis of many samples in a comparatively short time. This makes it particularly useful for large-scale glycomics studies. However, even for MS, glycans in unpurified cellular extracts are of low abundance and have poor ionisation efficiency. To counteract this, various chromatographic separation techniques are employed to isolate glycans in the sample. Both to improve chromatographic separation efficiency and ionisation efficiency during mass spectrometry, monosaccharides in glycans can be chemically modified or, in other words, derivatized. Therefore, to ensure good signal-to-noise ratio, MS is often combined with separation techniques like liquid chromatography (LC) or capillary electrophoresis (CE) that separate glycans based on their different physical and chemical properties. This further improves the resolution and sensitivity of glycan analysis. As a result, detailed analysis of glycan compositions in cellular extracts requires significant considerations of analytical techniques revolving around improving signal-to-noise ratio in mass spectrometry analysis. The aim of the following sections is to introduce a broad overview of state-of-the-art workflows that enable the elucidation of glycan compositions from cellular extracts, ultimately resulting in glycan composition descriptions of associated glycoproteins potentially used in glycan composition validation of glycoprotein 3D structures.

1.6.1 Sample preparation

The preparation of glycoprotein samples for mass spectrometry (MS) analysis is an important step that can significantly influence the quality of the results. Glycoproteins can be extracted from a plethora of samples, which could be biological fluids like blood or tissue extracts. Nevertheless, the most comprehensive results are achieved when glycoproteins are extracted from cellular samples. Cell cultures are amenable to be controlled under different conditions that manipulate specific variables, such as nutrients, growth factors, temperature, pH and targeted mutations to enable the investigations of specific factors altering glycan profiles of whole cells or individual glycoproteins. Moreover, cell cultures are generally more homogeneous than biological tissues, which often contain a mixture of different cell types. The diversity of cell types in tissue samples can sometimes make it difficult to attribute observed glycan patterns to specific cells or biological processes. As a result, analyses carried out in cell cultures offer more reproducible results than tissue samples. Finally, cell lines can be cultured in large quantities in the lab, which is an important factor in the quality of MS analysis, while tissue samples are less abundant and may require invasive procedures that are constrained by ethical and legal concerns. However, cell cultures do not fully reflect the complexity and diversity of biological tissues and as a result the obtained results might not fully translate to *in vivo* conditions, therefore the choice of sample source depends on context and specific research question.

After the glycoprotein has been purified in the sample, it is often necessary to denature the protein to make protein digestion more efficient for sample preparation into MS. This can be achieved by heating the soluble sample to break all of the intra-protein hydrogen bonds or by using denaturing agents such as urea. If the glycoprotein has a high content of disulfide bonds, then disulfide bonds can be broken down using addition of DTT and alkylation with iodoacetamide to fully unfold the glycoprotein target[70]. After the glycoprotein has been linearized, the protein is fragmented into smaller peptides using digestion enzymes. Most commonly, trypsin digestion is used which cleaves proteins at the carboxyl side of lysine and arginine residues, unless they are followed by proline. This generates smaller peptide fragments that are of a more manageable size for most configurations of MS devices.

In order to improve glycan signal-to-noise ratio, glycans may be released from peptides using enzymes or chemical reagents. For *N*-glycan release from peptide, there is peptide *N*-glycosidase F (PNGase F) enzyme that cleaves between the nitrogen atom, which is covalently bound to the GlcNAc sugar and the neighbouring carbon atom of amide functional

group of the asparagine residue. As a result of this reaction, the entire glycan is detached from the peptide, while asparagine residue is converted to aspartate residue. In addition, glycans can also be cleaved using endoglycosidase enzymes that cleave the glycosidic linkage between the first and second GlcNAc sugar residues of the glycan containing chitobiosyl core, leaving a single GlcNAc residue attached to the asparagine. PNGase F enzyme is considered to have broad specificity, as it can cleave almost all *N*-glycans, while specific endoglycosidases show preferentially for specific *N*-glycan compositions[71]. Unfortunately, such enzymes do not exist for *O*-glycan glycopeptides, and the release of *O*-glycans utilises hydrazinolysis or β-elimination which require the use of chemical reagents to catalyse the *O*-glycosidic linkage cleavage[72]. While the signal-to-noise ratio is improved upon release of glycans from their respective peptides, the information to which peptide fragment a specific glycan was covalently linked to is lost. In order to preserve glycosylation site information, the glycan release needs to be skipped, resulting in glycopeptides with attached glycans. The glycopeptides are then analysed directly, allowing to preserve the glycosylation site information at a cost of reduced signal-to-noise ratio. Therefore, both approaches are combined in glycoproteomics to not only get high resolution data of released glycan compositions, but also protein sequence information describing glycosylation sites[70].

Released glycans in their native states have poor ionisation efficiency, which has a direct impact on the signal detection in MS analysis. Moreover, if released glycans contain monosaccharides that have labile functional groups, such as neuraminic acids, the identity of such sugars can be lost during MS analysis. Therefore, released glycans are chemically modified (derivatized) to not only boost signal-to-noise ratio, but also preserve crucial information about the components of the oligosaccharide. Additionally, derivatization may also enhance the detection of glycans in chromatography-based techniques prior to MS analysis. There are two broad categories of derivatization techniques: chemical modifications that target the labile hydroxyl groups of monosaccharide units, example being permethylation and chemical modifications that target reducing position of released *N*-glycans, an example being fluorescent labelling.

Permethylation is a well-established derivatization technique, where acidic protons of all carbohydrates in glycans are substituted for methyl groups. The main reagent used for permethylation is iodomethane (CH_3I), which serves as a source of methyl groups. The reaction is usually carried out in an organic solvent such as dimethyl sulfoxide (DMSO), where sodium hydroxide (NaOH) is used as a base catalyst that deprotonates the acidic protons of monosaccharides to make them more nucleophilic. The deprotonated groups initiate a nucleophilic attack on the methyl iodide, where deprotonated acidic protons

become methoxy groups. After the reaction, the permethylated glycans are usually purified to remove excess reagents and byproducts. As a result, permethylation significantly enhances glycan ionisation and chromatographic efficiency via the addition of methyl groups that make the glycan more hydrophobic. Moreover, for monosaccharides that have labile functional groups, such as neuraminic acids, permethylation esterifies carboxylic acid groups that otherwise might get lost during MS analysis. Finally, permethylation partially enables the discrimination of isomeric structures in MS/MS analysis by changing the fragmentation patterns of glycans. Specifically, permethylation can make it easier to distinguish 1,3-linkages from 1,4-linkages and 1,6-linkages, which are a hallmark feature of glycan branching[73].

Reducing end labelling is a technique used to attach a fluorescent or UV-absorbing label to the reducing end of a released glycan molecule. The labelled glycans can then be detected and quantified by fluorescence or UV absorption during separation stages. After the glycan release using PNGase F, released glycans have a free reducing terminus at the GlcNAc that was previously covalently linked to asparagine. The specific GlcNAc is in equilibrium between its closed ring (pyranose) form and straight chain (aldehyde) form. The free carbonyl group on the aldehyde form can be reacted with a label molecule that contains an amine group to form a Schiff base intermediate. Schiff base intermediates containing the fluorescent label can then be reduced to a secondary amine using a reducing agent like sodium cyanoborohydride ($NaBH_3CN$) to form a stable bond between the reducing end of the glycan and the molecule containing the fluorescent label. Examples of labels include 2-aminobenzamide (2-AB), 2-aminopyridine (2-AP), and anthranilic acid (2-AA). The fluorescent or UV-absorbing label greatly enhances the detection of the glycans, as they can be detected and quantified by fluorescence or UV absorption. Finally, some labels can also improve the ionisation efficiency of the glycans in MS, increasing sensitivity[74].

Prior to MS analysis, chromatographic separation of derivatized or non-derivatized glycans can be carried out. Fundamentally, chromatography is a physical separation technique used to divide the components of a mixture. The principle behind chromatography is that different substances in a mixture have different affinities for two different phases - a mobile phase, which can be a liquid (LC), and a stationary phase. As the mobile phase moves over or through the stationary phase, the different components of the mixture interact with the two phases to different extents. Some molecules have a higher affinity for the stationary phase and get slowed down as they interact with it, while others have a higher affinity for the mobile phase and move more quickly. This difference in migration rates leads to the separation of the components as they move through the chromatographic system. Depending on the type

of glycan derivatization or whether released glycans were derivatized at all, there are a number of chromatographic separation techniques that can be used. Chromatographic separation is a powerful tool in glycan analysis, but it is also challenging. Glycans often exhibit complex, overlapping elution patterns, which can make it difficult to fully resolve different structures. Furthermore, different glycans can sometimes have very similar chromatographic properties, making them hard to separate. The choice of separation method, as well as the use of different sample preparation and derivatization techniques, can have a significant impact on the effectiveness of the separation and therefore mass spectrometry output.

1.6.2 Mass Spectrometry

If the sample containing glycoproteins was previously purified using chromatographic techniques, then it is directly injected into the Mass Spectrometer. Otherwise, Mass spectrometry analysis can be directly coupled to liquid chromatography (LC) instruments. In state-of-the-art mass spectrometry setup for glycoproteomics, Ion Mobility-tandem Mass spectrometry (IM-MS/MS) is used, which is capable of separating ions based on their size, charge and conformation[75]. Even though IM-MS/MS does not require derivatized glycans, derivatization can enhance the distinction of isomers by providing more distinct fragmentation patterns.

Once the sample is introduced to IM-MS/MS machinery, the glycans are ionised to produce charged particles that can be manipulated and detected in the mass spectrometer. The most common ionisation techniques for glycan analysis are electrospray ionisation (ESI) and matrix-assisted laser desorption/ionisation (MALDI). In ESI, the sample is sprayed through a needle at high voltage to produce a fine mist of charged droplets. As these droplets evaporate, they leave behind charged molecules or ions. In MALDI, the sample is mixed with a matrix compound and irradiated with a laser. The matrix absorbs the laser energy and assists in the desorption and ionisation of the sample.

Following ionisation, the charged ions are then separated based on their mass-to-charge ratio (m/z) in the first mass analyzer. Several types of mass analyzers are commonly used in glycan analysis, including time-of-flight (TOF), ion trap analyzers. In a TOF analyzer, ions are accelerated by an electric field and then allowed to fly down a long tube. Ions with a lower m/z will reach the detector faster than ions with a higher m/z. In an ion trap, ions are trapped in an electric field and their oscillations are measured to determine their m/z.

The selected ions from the first mass analyzer are then subjected to ion mobility separation (IMS). IMS enables differentiation of ions based either on space or time, with time-dispersion methods being more commonly used[75]. Time-dispersion IMS works by applying an electric field to ions in a gas-filled chamber. The electric field causes ions to exhibit different behaviour based on their size. Ions with a smaller surface area will move faster through the field, while ions with a larger surface area will move slower. Because glycans often exhibit overlapping m/z values, IMS enables the separation of glycans based on their 3D structure, as different patterns of glycan branching, or glycosidic linkages may correlate with changes in surface area of the glycan ion.

Following ion mobility separation, glycan ions are fragmented using techniques such as collision-induced dissociation (CID) or electron transfer dissociation. CID is often a preferred choice, as it can produce more predictable fragmentation patterns. CID is essentially a collision cell, where accelerated high kinetic energy ions are collided with neutral gas molecules, such as helium, to induce the transfer of kinetic energy into internal energy. Internal energy of ions may cause covalent bond breakage, which results in fragmentation of the parent ion.

Smaller fragments from the parent ion are then analysed in the second mass analyser. The m/z of these product ions are then measured to derive the composition of the parent ion. The fragmentation of derivatized glycans enables the derivation of structural information, such as the sequence of monosaccharide units and the type of linkages.

The output from an IM-MS/MS experiment includes multiple layers of data that provide information about the sample being analysed. The output from the first mass spectroscopy analysis is a spectrum showing the intensity of ions (y-axis) versus their mass-charge ratio (x-axis). However, to assign glycan composition details to observed individual peaks, the output of the second mass spectroscopy analysis is used. The selected ion peaks that were further fragmented are expressed as m/z values with relative abundances of fragments observed in the parent ion. The relative abundance relationship enables the derivation of glycan composition associated with the parent ion. Therefore, IM-MS/MS is capable of elucidating what kind of glycans are detected in the sample with the ability to measure their abundances.

1.7 Structural Biology to study glycoproteins at atomic scale

Structural biology is a branch of molecular biology, biochemistry, and biophysics concerned with the molecular structure of biological macromolecules, how they acquire the structures they have and how alterations in their structures affect their function. In order to obtain an atomistic description of glycoproteins, several experimental techniques are used: Nuclear Magnetic Resonance Spectroscopy (NMR), Macromolecular X-Ray Crystallography (MX) and Cryogenic Electron Microscopy (Cryo-EM). Of these three techniques, a significant majority of the glycoprotein structures have been elucidated using MX.

1.7.1 Macromolecular X-ray Crystallography (MX)

X-ray crystallography is an experimental pipeline that utilises diffraction of X-rays from atoms of the target embedded in a crystal to resolve macromolecular structures. It provides a 3D image of the density of electrons within the macromolecule, allowing the position of every atom in the macromolecule to be described. In the context of glycoproteins, it can reveal how the oligosaccharide moieties are attached to the protein backbone and how these attachments influence the overall protein structure.

X-rays have several properties that make them ideal for use in crystallographic studies of macromolecules like proteins, nucleic acids, and complex carbohydrates. The wavelengths of X-rays are in the range of 0.1-10 Angstroms, comparable to the size of atoms. Visible light, in comparison, has a much longer wavelength, making it unsuitable for studying atomic structures. Additionally, X-rays have the ability to penetrate matter, enabling the study of internal features of molecular structures. This also presents the most significant drawback for this technique; in that it must rely on collecting diffraction patterns. This is due to a physical inability to re-focus X-rays through a lens to obtain an image in real space: as X-rays penetrate most materials, their X-ray refraction index is close to 1.0, making the design of X-ray lenses problematic. Real space represents the Cartesian positions of atoms in a crystal lattice; reciprocal space mathematically represents the directions and wavelengths of the diffracted beams. This results in the “phase problem”. While the amplitude of the diffracted X-rays can be calculated from the measured intensity of the spots, the phase, which corresponds to the specific position within the cycle of each wave at the moment it was measured, cannot be directly measured. Nevertheless, workable solutions have been developed to solve “the phase problem”, allowing to achieve interconversion between real space and reciprocal space using the Fourier transform.

The first step and arguably the most significant bottleneck in the pipeline is the growth of a crystal containing multiple copies of the target glycoprotein. The quality of the grown crystal has significant implications for the quality of the electron density map reconstruction. In an ideal scenario, a single, highly homogenous crystal containing sufficient amounts of identically ordered copies of target glycoprotein would be obtained. Nevertheless, realistically, crystal quality is bottlenecked by factors such as thermal motion, heterogeneously ordered contents of the unit cell and various physical defects to the crystal itself. Specifically for glycoprotein crystals, the heterogeneity emerging from varying glycoforms has a direct impact on the homogeneity requirement, leading to noisy electron density maps.

Once a glycoprotein crystal is successfully produced, it is commonly subjected to a highly intense X-ray beam, usually sourced from synchrotrons or Free Electron Lasers (FELs). Synchrotrons, which are sophisticated particle accelerators, produce incredibly bright X-ray beams by accelerating electrons to relativistic speeds and then deflecting them with powerful magnetic fields. As these electrons navigate through the magnetic fields, X-rays are emitted. This process of synchrotron radiation allows for the generation of X-rays that span a broad spectrum of energies.

Given that the resultant X-rays span a range of energies, a monochromator is necessary to filter the X-rays to a single wavelength. Typically, in molecular crystallography, wavelengths between 0.5 Å and 1.6 Å are used, aligning with expected covalent bond lengths in molecular structures. The monochromatic X-ray beam then targets the crystal. The electron clouds of atoms within the crystal diffract the X-rays, and the regular atomic arrangement in the crystal lattice causes scattered X-rays to interfere constructively and destructively, as governed by Bragg's law:

$$n\lambda = 2d\,sin\theta$$

where λ is the wavelength of the X-ray, n is a positive integer, d is the spacing between the crystal planes, θ is the incident angle. The constructive interferences produce a diffraction pattern — essentially a snapshot of the crystal — which is efficiently captured by state-of-the-art detectors. For comprehensive electron density mapping, the crystal undergoes incremental rotation to gather diffraction patterns from varied angles. Many crystals are flash cooled in liquid nitrogen to reduce radiation damage during data collection. Although X-rays emit ionising radiation that can harm the target molecule in the crystal, the multiple identical copies of the molecule within the crystal typically yield sufficient data before radiation damage becomes significant.

Upon data collection, the data are then is processed to generate an electron density map. This process is initiated by data integration, where the intensity of individual spots is measured. Each spot is indexed using Miller indices, which describe the directions and planes in crystal lattice that the diffracted X-ray beam took to create each spot. The assigned Miller indices are inherent to the crystal structure and are consistent regardless of the crystal orientation during data collection. Therefore, the assignment of Miller indices enables the scaling and integration of individual diffraction spots across multiple diffraction patterns to obtain singular intensity values, optimised for signal-to-noise ratio. Once singular intensity values are obtained, the "phase problem" is addressed. Traditionally, molecular replacement (MR) would be used, where the model of a previously-solved structure that is similar to the target molecule (homologous) would be used in a six-dimensional search to inform the orientation and position of the target in the crystal lattice. The comparison would then be used to calculate phase estimates. If no homologous structure was available, then multiple isomorphous replacement (MIR) could be used, where heavy atoms would be introduced to the target molecule to create derivative crystals. The differences in the diffraction patterns between the native and the derivative crystals are then analysed to estimate the positions of heavy atoms and therefore calculate the phase information for the diffracted X-rays. Nevertheless, the use of this technique has significant health and safety considerations, as the introduction of heavy atoms to target molecules requires handling highly toxic metal salts[76]. It has been recently reported that the limitation of homologous replacement models not being available can be overcome thanks to AlphaFold2 predictions[77,78]. Therefore, it is likely that in the longer term not only AlphaFold2, but also RoseTTAFold, ESMfold powered molecular replacement will be completely sufficient to resolve the "phase problem"[79–81].

Once the phases have been estimated and the intensity values of diffraction pattern spots have been measured, the electron density map is calculated. The electron density map represents an image in 3D space that corresponds to the contours of the atoms in the crystal. This is done using a Fourier transform, which transforms the processed diffraction data from reciprocal space into an electron density map in real space. The conversion is bi-directional, meaning that electron density maps in real space can also be converted into diffraction spots in reciprocal space. This mathematical property is what enables techniques such as molecular replacement to work fundamentally.

1.7.2 Cryogenic Electron Microscopy (Cryo-EM)

Cryo-EM is a type of transmission electron microscopy (TEM) where the sample is studied at cryogenic temperatures. TEM in principle works in a similar manner as a light microscope except, instead of using light, a beam of electrons is used to visualise the sample. Given that electrons used in TEM typically have wavelengths on the order of subangstrom range, compared to visible light which has wavelengths ranging from about 400 nm to 700 nm, TEM can achieve much higher imaging resolution than light microscopy[82]. However, the electron beam used in TEM is of high energy, which may cause structural damage to the sample by breaking covalent bonds between atoms, therefore there are significant considerations in terms of limiting exposure. The beam of electrons transmitted through a thin layer of sample is scattered and collected by the detector to reconstruct the signal into a magnified image. The magnified image is a projection of the 3D structure of the sample on a flat plane[83]. One significant advantage of Cryo-EM over X-ray crystallography is that the crystallisation of the sample is not performed, therefore removing a significant bottleneck in the pipeline. Instead, samples in Cryo-EM are vitrified - where the biological sample is cooled down to 93 K or below using liquid ethane and maintained during data collection. Vitrification, which is likewise used in MX, enables the preservation of the native hydrated structure of biological samples by preventing the formation of ordered structures by water molecules which can damage the sample and instead solidifying the aqueous solvent into an amorphous state[84]. However, Cryo-EM requires biological samples to have a molecular weight greater than 38 kDa, unlike X-Ray crystallography which is capable of handling samples of lower molecular weight[85].

The 3D density map of the target in the sample is reconstructed from multiple flat plane projections of 3D structure obtained from a variety of angles. Usually, this process generates terabytes of 2D images. The processing of 2D images is initiated by particle picking, where snapshots associated with the target are identified and extracted to isolate it for further processing. Afterwards, two-dimensional classification of picked particles is carried out which are grouped into classes based on similarity to produce class averages with enriched signal-to-noise ratio. Next, follows three-dimensional classification of 2D classes which maps flat plane images onto potential 3D structure. Following the mapping of flat plane images, 3D refinement follows, which seeks to optimise the alignment of the individual particle images and the overall model to get the best possible resolution. After additional post-processing steps, a 3D density map of the target is obtained, which corresponds to a collection of voxels (the 3D equivalent of a pixel), with individual values representing the estimated electron potential.

1.7.3 Nuclear Magnetic Resonance Spectroscopy (NMR)

Protein NMR relies on the quantum mechanical properties of atomic nuclei to produce a dynamic spectrum that allows for the mapping of atomic linkages, distances between atoms, and changes in their positions within glycoproteins. The latter property of being able to map change in atom position in real time is a particular advantage over X-ray crystallography and cryo-EM. Therefore, protein NMR is oftentimes used to study protein interactions in their native environments of aqueous solutions. In X-ray crystallography, on the other hand, the sample environment is modified to aid crystallisation. Moreover, the captured states of samples within the crystal are just a single snapshot of the protein at lowest energy states, as additional presence of entropy would disrupt the crystallisation progress. Therefore, since NMR can deal with dynamic states of glycoproteins, it could be a viable alternative used to obtain structural descriptions of oligosaccharide regions. Regardless of these advantages, NMR experiments only make up a small portion of elucidated structures that have been deposited in the Protein Data Bank (PDB). The reasons for this are twofold. Firstly, depending on the biological question, but certainly for structural biology, multiple types of NMR experiments have to be conducted until discerning features of atoms within the glycoprotein would become obvious. In comparison to X-ray crystallography, while crystallisation runs are a time-consuming process with some elements of luck involved, the diffraction and electron density map reconstruction of the contents of the crystal are relatively quick and straightforward. Most importantly, however, NMR is only able to handle molecular structures that are relatively small. With bigger samples, the resolution decreases significantly too and NMR experiment outputs become uninterpretable. Because of these reasons there is a size limit on biological structure that could be elucidated via NMR and in the grand scheme that tends to be relatively small proteins, including glycoproteins. Most of the intact glycoproteins are indeed significantly bigger than the size limit of NMR[86]. Secondly, significantly large amounts of protein are necessary for data acquisition in NMR experiments. Moreover, for NMR to discern useful features of the sample, the glycoprotein typically has to be expressed with 2H, 13C, 15N isotopes added to the media of the expression system. The isotopes are extraordinarily expensive. Adding to the fact that multiple NMR experiments need to be executed, this may quickly end up putting a huge dent in the finances of the project, purely due to the need for radiolabelled isotopes[87]. Therefore, NMR is typically used if there is a specific need to capture and describe conformational changes that are crucial in protein-protein interactions. Nevertheless, X-ray crystallography or cryo-EM could be used in conjunction with NMR in elucidating glycoprotein structures. NMR could be used to exclusively deal with highly mobile oligosaccharide regions of the

protein, while other experiments could target describing the structural features of the rest of the protein[88,89].

1.7.4 Fitting density maps to reconstruct atomic descriptions of glycoproteins

Unlike multiple iterations of NMR experiments, cryo-EM and X-ray methods do not immediately produce atomic descriptions of glycoprotein models. Processing the raw data of the two methods produces density maps that need to be interpreted, refined and atoms fitted to produce a complete glycoprotein model.

1.7.4.1 Protein model building, refinement and validation

If the resolution of the reconstructed density map is sufficient then automatic software tools, such as phenix.autobuild (builds protein and nucleic acids)[90], ARP/wARP (builds protein, nucleic acids and ligands)[91], BUCCANEER (builds protein)[92], Nautilus (builds nucleic acids)[93], ModelAngelo (builds protein)[94] and ModelCraft (builds protein and nucleic acids)[95] are capable of reconstructing atomic models with varying levels of model completeness. However, typically the initial versions of calculated density maps are insufficient to model complete atomic descriptions of target molecules in the sample. Therefore, in order to improve the density map and atomic reconstruction of the protein model, refinement of the initial density map against the initial atomic model is carried out. There are a number of available refinement softwares, most notably phenix.refine and REFMAC-Servalcat, that are capable of refining both types of density maps[96,97]. The goal of refinement is to optimise the fit between experimental density map and calculated density map from the fitted atomic coordinates of the model. As a result, this process is iterative - as signal-to-noise ratio is improved in the working electron density map, model completeness improves, further improving successive electron density maps' signal-to-noise ratio in the next cycles of refinement. However, overinterpretation of weak features in electron density maps or refinement without or insufficient restraints may lead to overfitting. In order to avoid overfitting, refinement softwares utilises geometric restraints as prior knowledge to maintain chemically plausible geometries of biomolecules by preventing models from adopting unrealistic conformations just to fit noise in the data. The prior knowledge of geometric restraints is described in monomer libraries, such as CCP4-ML, that contain lowest energy parameters for individual units, such as amino acids, carbohydrates, ligands and nucleic acids[13,98,99]. Therefore, cycles of refinement and model building are carried out recursively, until the most optimal solution is found. In order to determine the most optimal solution, it is

imperative to have objectively quantifiable metrics that can evaluate the improvements between refinement cycles. In X-ray crystallography, the improvements of individual cycles of refinement of electron density maps are evaluated using the R-factor metric, which measures the agreement between experimental and calculated X-ray diffraction data. In order to measure the degree of overfitting, reflection data is randomly split into R_{work} and R_{free}, where refinement is carried out on R_{work}, while R_{free} contains a small set of reflections set aside that were entirely excluded from refinement. In the case of cryo-EM, the R-factor metric equivalent is not available. Instead, cryo-EM density map refinement relies on calculating FSC scores between two half-maps obtained during the 3D reconstruction procedure. FSC scores are defined as correlation coefficient between two half-maps as a function of spatial frequency. One half-map is used for refinement, while another is preserved as a test map for validation. If the model is overfit to the density map used in refinement, the FSC against the test set will drop sharply at high resolutions, indicating that the model does not generalise to randomly excluded data[100]. Once the refined experimental density map is obtained and if automatic model building softwares failed to generate a complete description of the model, then missing parts of the model are manually built using model building software, such as Coot.

After the model has been built, it is then validated before deposition using a variety of softwares that target specific aspects of the model. For example, MolProbity can be used to detect amino acid chemistry geometric errors, in terms of main-chain (using the Ramachandran criterion) and sidechain (checks of conformation against a rotamer library) outliers and atomic clashes[101]. Tortoize is another software that can be used to compute a model's Ramachandran Z-score, which describes how “normal” a model is compared to a reference set of high-resolution structures, detecting any potential outliers in terms of backbone geometry[102]. Both tools can be used to validate models obtained from both cryo-EM and X-ray crystallography density maps as experimental density input is not required. For X-ray crystallography, EDSTATS software can be used to calculate real-space metrics to evaluate reconstructed models fit to a refined density map[103]. The equivalent of EDSTATS in cryo-EM is TEMPy which provides a variety of different scoring functions to evaluate goodness-of-fit between a built model and density map or between two maps[104]. It is typical for validation software to flag up potential issues, therefore model building software such as Coot (which also provides numerous validation functionalities) enables to not only qualitatively assess the model, but also fix potential issues in a more manual manner, greatly aided by built-in refinement capability accessible through real space refine feature[105].

1.7.4.2 Challenges associated with building *N*-glycans

Most model building software is protein centric. There are ongoing efforts to also automate the reconstruction of carbohydrates into density maps using Sails software[106]. In addition, Coot offers functionality to add *N*-linked glycans into refined density maps in a semi-automated manner[107]. However, for a significant amount of time, little to no software attempted to target automated oligosaccharide building and there were multiple justified reasons, ranging from the proportion of non-glycosylated protein versus glycosylated protein structures being successfully crystallised to the modelling challenges raised by the complexity of carbohydrate chemistry. One of the primary challenges associated with the modelling of oligosaccharides is indirectly apparent in the following statistic: the median resolution for glycoproteins (2.4 Å) is lower than that of non-glycosylated proteins (2.0 Å) when X-ray crystallography cases are considered[12]. If electron density associated with the potential glycan is observed at all, typically at this lower resolution range there are little discernible features in terms of monosaccharide identity (Figure 1.6). In cryo-EM electron potential maps, the stated resolution value is a mere average, with significant local variations in resolution across the model. Indeed, density regions associated with potential glycans tend to have significantly lower resolution values.

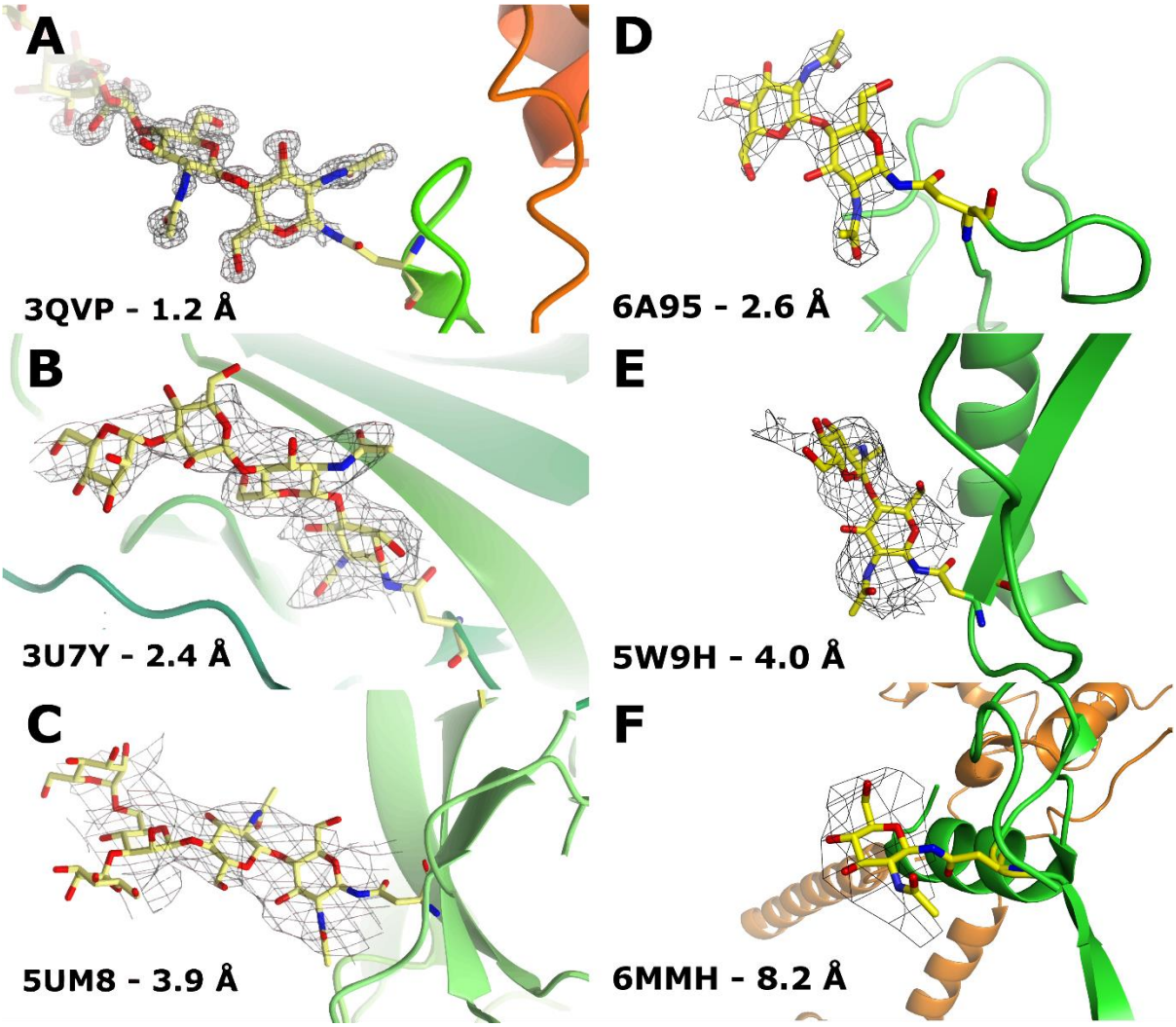


Figure 1.6: Comparison of *N*-glycan features in density maps over a range of resolutions. (a)–(c) Electron density maps obtained with X-ray crystallography (MX). (d)–(f) Electronic potential maps obtained with cryo-EM; PDB codes and data resolution have been annotated directly on the figure. In the MX cases (a)–(c), at high resolution (a) it is possible to identify monosaccharides and their ring conformation from the density map; at medium resolution (b), ring conformation becomes difficult to determine, whereas at low resolution (c), and indeed with many cryo-EM maps (d)–(f), density associated with *N*-glycans have poorly defined discernible features of individual carbohydrates.

Another significant contributing factor to difficulties associated with modelling oligosaccharides into electron density maps is not having prior knowledge for the glycan compositions present in specific glycosylation sites if glycoproteomics was not an integral part of the project. It is much easier to fit electron density for protein backbones, as there exists easily accessible prior knowledge for the amino acid sequence. While protein sequences are derived from DNA sequences and can be found in databases, unfortunately, the contents of glycan sequence compositions are not directly encoded in the genomes of

expression systems. Since *N*-glycosylation is carried out by specific glycosyltransferases and glycosidases, this information enables an informed guess as to what categories of *N*-glycan products are to be expected, as many enzymes related to glycosylation are well classified and investigated. For example, if an expression system does not have its genome encoded for GnT enzymes, then it is likely that complex *N*-glycans are unlikely to be synthesised. The lack of prior knowledge for glycan compositions in specific glycosylation sites can be overcome by integration of glycoproteomics into structural biology pipelines. This approach has proven to be successful in elucidation of native human uromodulin (PDB: 7PFP), where low-resolution cryo-EM density map was combined with glycoproteomics to model glycans beyond the available signal in density map (Figure 1.7)[108].

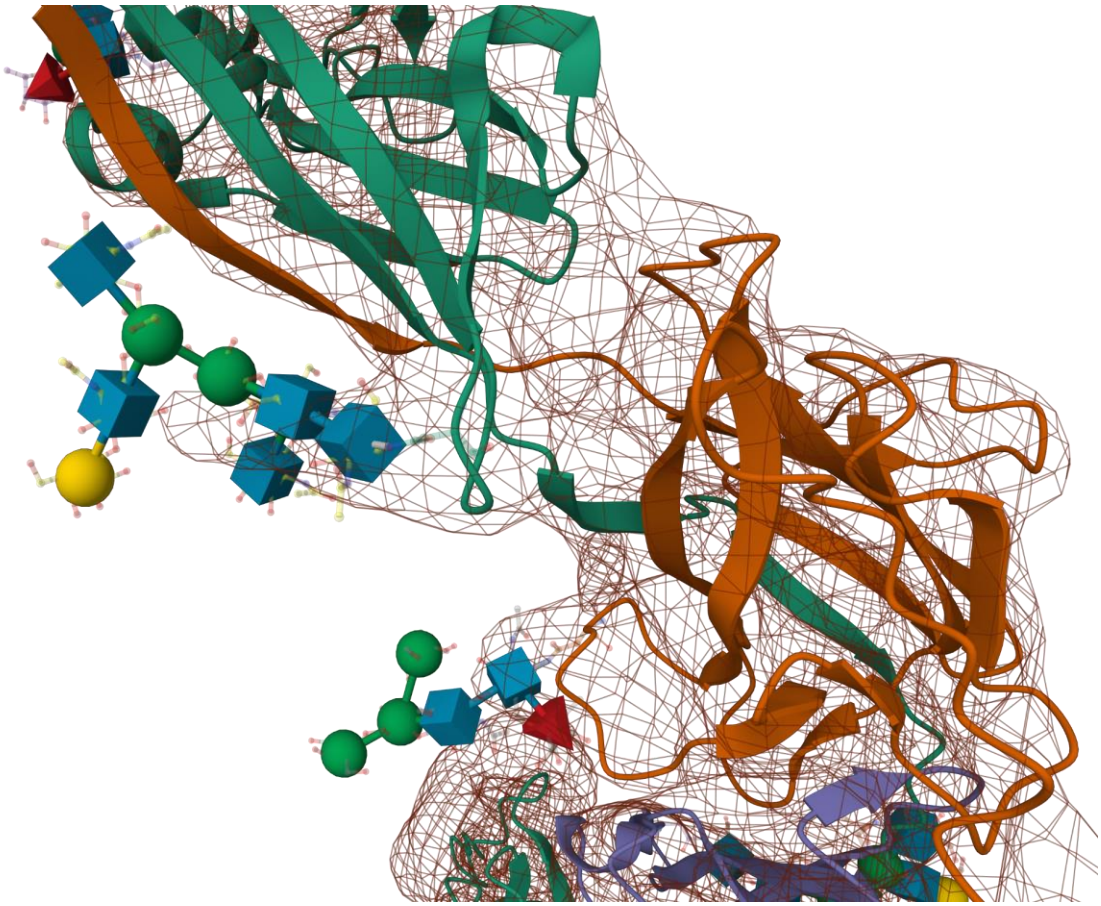


Figure 1.7: Visualisation of modelled glycans at ASN396 and ASN513 in human uromodulin (PDB: 7PFP) against the associated EM density map (brown chicken wire mesh). Some modelled carbohydrates are modelled outside the density map, due to lack of associated signal. The EM density map in the figure is rendered at 0.008V (2.5σ) contour level. Contour level recommended by the authors is 0.006V (2.0σ), according to the metadata deposited in EMDB entry: EMD-13378[109].

Because of these factors, glycan chains that are modelled in glycoproteins, tend to contain a significant number of modelling errors. As a demonstrable example of potentially significant

modelling errors, studies into monosaccharide ring conformations of *N*-glycans deposited into wwPDB reveal a stark picture. Typically, six membered ring carbohydrates that participate in protein glycosylation, should be modelled in lowest energy conformation state - chair and not high energy conformation state such as skew-boat, half-chair or envelopes. Monosaccharide conformational analysis using data obtained from the PDB, reveals that a significant amount of carbohydrate units in PDB exist in high-energy conformational states without experimental explanation for why it was modelled as such, suggesting that modelling mistakes were likely made[110]. Active development of software tools such as Privateer aims to overcome aforementioned issues associated with modelling of *N*-glycosylation.

1.8 In-silico predictions of glycoprotein structures at atomic scale.

The integration of experimental glycoproteomics into structural biology pipelines necessitates significant investment in both expertise and capital costs. According to estimates, even though 70% of the human proteome is glycosylated, glycosylated proteins make up only a small minority of structures deposited to PDB[111,112]. Therefore, one potential avenue for bridging the gap could potentially be in-silico modelling of glycoprotein structures.

Over past decades, there have been numerous attempts to develop predictive models that can predict 3D protein structures from amino acid sequence input. The developed tools would be benchmarked at the Critical Assessment of protein Structure Prediction (CASP) experiment. During the 14th edition of CASP, AlphaFold2 emerged which revolutionised the field with its remarkable accuracy. AlphaFold's performance in the competition demonstrated that it could predict some protein structures with an accuracy comparable to experimental methods like X-ray crystallography and cryo-EM[113]. AlphaFold utilises deep learning techniques, including convolutional neural networks and attention mechanisms, to model the relationships between amino acids and predict their arrangement in 3D space. Crucially, AlphaFold can self-evaluate its own prediction by providing various confidence scores[81].

AlphaFold is only capable of predicting spatial arrangement of amino acid residues, but not spatial arrangement of post-translational modifications such as *N*-glycosylation, nucleic acids or ligands affecting protein folding. Nevertheless, research shown in this thesis (Chapter 4) and other works, such as AlphaFill, practically demonstrate that AlphaFold predictions can be enriched with prior knowledge by simply grafting missing non-amino acid residues[114].

However, in order to enrich predictions produced by AlphaFold on a massive scale, predictive tools targeted at glycosylation composition prediction are required.

One of the requirements to glycosylate AlphaFold predictions is predicting the locations of potential glycosylation sites. Thankfully, such tools already exist, such as NetNGlyc, which uses artificial neural networks to analyse the primary protein sequence and predict the potential glycosylation sites and GlycoMine which is another machine learning approach that considers not only the amino acid sequence but also the 3D structure of the protein[115,116]. However, to date there is no predictive tool that would be capable of predicting potential compositions of glycans harbouring the potential glycosylation sites based on 3D structure of the protein. Therefore, the aim of this thesis is to address this gap and pioneer the development of a predictive tool that could be used to aid the modelling of specific glycan compositions in target glycosites.

1.9 GlyTouCan and GlyConnect: Datastores of Glycomics Research.

In the evolving field of glycomics, GlyTouCan and GlyConnect datastores emerge as pivotal repositories, facilitating the study of complex glycan structures and their biological implications. This subsection aims to describe the roles and features of these databases, which are recurrently referenced throughout the thesis.

1.9.1 GlyTouCan

GlyTouCan acts as an international registry for all known glycan compositions, offering researchers a centralized platform for the deposition and retrieval of known glycan structures[159, 160]. The deposited glycan compositions are described using both GlycoCT and WURCS notation languages, enabling the integration of diverse glycan data[160]. The database assigns unique identification tags to each glycan sequence, a crucial step for standardizing glycan representation and facilitating cross-database searches[160]. This universal identifier system not only promotes interoperability among glycomics databases but also simplifies the tracking of glycan data across different projects and platforms.

The GlyTouCan project's ambition extends beyond mere data aggregation; it aims to support the glycoscience community by providing tools like the GlycanFormatConverter[161]. This tool converts complex WURCS notations into more interpretable formats, thereby enhancing data usability for researchers[161]. Through its comprehensive approach to data standardization and accessibility, GlyTouCan plays a critical role in advancing glycomics research, enabling scientists to share, discover, and analyze glycan sequences using a common notation in the form of GlyTouCan identifiers and WURCS notation.

1.9.2 GlyConnect

While GlyTouCan focuses on glycan sequences, GlyConnect extends the research horizon by connecting glycan data with glycoproteomics information[15]. This database specializes in the curation and dissemination of data concerning glycan structures and their associated proteins, offering insights into glycosylation processes and glycan function within biological systems. Approximately 70% of its data is manually curated by experts, ensuring high-quality and reliable information on glycan compositions, linkage patterns, and their biological contexts[15]. Most crucially, GlyConnect has deep integration with the GlyTouCan repository.

GlyConnect's utility is exemplified by its ability to provide detailed metadata, notably protein backbone associations and N-glycosylation compositions. By integrating glycoproteomic data, GlyConnect facilitates a more comprehensive understanding of glycan functionality, shedding light on the roles of glycans in health and disease. The database's interoperability with other platforms, such as UniProt, puts it at the forefefront of glycomics research, enabling researchers to cross-reference glycan compositions with protein structures and functions seamlessly[15].

Integrative structural glycobiology

2.1 Published Article: Leveraging glycomics data in glycoprotein 3D structure validation with Privateer

Most of the content in this chapter is taken word for word from an already published peer-reviewed paper in “Beilstein Journal of Organic Chemistry” under the title of “Leveraging glycomics data in glycoprotein 3D structure validation with Privateer” by Bagdonas, Ungar and Agirre[2]. The chapter contains an addendum that was not originally published in the paper.

2.2 Introduction

Glycosylation-related processes are prevalent in life. The attachment of carbohydrates to macromolecules extends the capabilities of cells to convey significantly more information than what is available through protein synthesis and expression of genetic code alone. For example, glycosylation is used as a switch to modulate protein activity[117]; glycosylation plays a crucial part in folding/unfolding pathways of some proteins in cells[118,119]; the level of *N*-glycan expression regulates adhesiveness of a cell[120]; glycosylation also plays a role in immune function[121] and cellular signalling[121,122]. At the forefront, glycosylation plays a significant role in influencing protein-protein interactions[119]. For example, influenza virus uses haemagglutinin glycoprotein to recognise and bind sialic acid decorations of human cells in the respiratory tract[123]. Glycosylation is also used by pathogens to evade the host’s immune system via glycan shields[124–126] and thereby delay an immune response[127]. The structural study of these glycan-mediated interactions can provide unique insight into the molecular interplay governing these processes. In addition, it can provide structural snapshots in atomistic detail that can be used to generate molecular dynamics simulations describing a wider picture underpinning glycan and protein interactions[128]. Unfortunately, significant challenges have affected the determination of glycoprotein structures for decades and have had a detrimental impact on the quality and reliability of the produced models. Anomalies have been reported regarding carbohydrate nomenclature[129], glycosidic linkage stereochemistry[130] and torsion[131,132], and most recently, ring conformation[133]. Most of these issues have now been addressed as part of ongoing efforts to provide better software tools

for structure determination of glycoproteins, although the most difficult cases remain hard to solve. Chiefly among these is the scenario where the experimentally resolved electron density map provides evidence of glycosylation, without enough resolution to derive definite and comprehensive details about structural composition of the oligosaccharides (Figure 2.1). Glycan microheterogeneity and the lack of carbohydrate-specific modelling tools have often been named as principal causes for these issues[110].

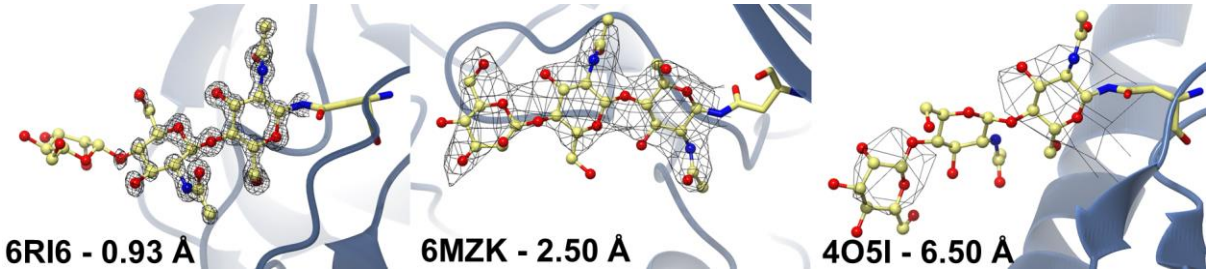


Figure 2.1: Comparison of glycan features in electron density maps over a range of resolutions from select glycoprotein structures (PDB entries: 6RI6[7]; 6MZK[8]; 4O5I[9]) Electron Density maps obtained with X-Ray crystallography. Data resolution and PDB entry IDs associated with structures have been directly annotated on the figure. Left - depicts a high-resolution example, where monosaccharides and their conformations can be elucidated; centre – a medium resolution example, where identification starts to become difficult; right – a low resolution example, for which all prior knowledge must be used. Despite coming from different glycoprotein structures, the glycan has the same composition and thus is assigned a unique GlyTouCan ID of G15407YE.

2.2.1 Heterogeneity of glycoproteins

Unlike protein synthesis, which is encoded in the genome and follows a clear template, glycan biosynthesis is not template-directed. A single glycoprotein will exist in multiple possibilities of products that can emerge from the glycan biosynthesis pathways, these are known as glycoforms[134]. More specifically, the variation can appear in terms of which potential glycosylation sites are occupied at any time – macroheterogeneity – or variations in compositions of the glycans added to specific glycosylation sites – microheterogeneity. This variation in microheterogeneous composition patterns arises due to competition of glycan processing enzymes in biosynthesis pathways[135].

2.2.2 Implications for structure determination of glycoproteins

Several experimental techniques can be used to obtain 3D structures of glycoproteins: X-Ray Crystallography (MX, which stands for macromolecular crystallography), Nuclear Magnetic Resonance Spectroscopy (NMR) and Electron Cryo-microscopy (Cryo-EM). As of publication date, the overwhelming majority of glycoprotein structures have been solved using MX[24,26].

The biggest bottleneck in MX is the formation of crystals of the target macromolecule or complex. The quality of the crystal directly determines the resolution – a measure of the detail in the electron density map. Homogenous samples at high concentrations are required to produce well-diffracting crystals[136]. Samples containing glycoprotein molecules do not usually fulfil those criteria. More often than not, MX falls short at elucidating carbohydrate features in glycoproteins due to glycosylated proteins being inherently mobile and heterogeneous[134], moreover oligosaccharides often significantly interfere in the formation of crystal contacts that allow formation of well-diffracting crystals. Because of this, glycans are often truncated in MX samples to aid crystal formation[137].

In Cryo-EM, samples of glycoproteins are vitrified at extremely low temperatures, rather than crystallised as in MX. The rapid cooling of the sample allows to capture snapshots of molecules at their various conformational states, thus potentially maintaining glycoprotein states more closely to their native environments in comparison to crystallography[138]. Nevertheless, Cryo-EM is still not an end-all solution to solving glycoprotein structures: the flexible and heterogeneous nature of glycans still has an adverse effect on the quality of the data, affecting image reconstruction[139]. Moreover, due to the low signal-to-noise ratio, the technique works more easily with samples of high molecular weight; this situation, however, is evolving rapidly, with reports of sub-100 kDa structures becoming more frequent lately[140,141]. Crucially, MX and Cryo-EM can complement each other to counteract issues that both face individually[142].

The two techniques produce different information – electron density (MX) or electron potential (Cryo-EM) maps – but the practical considerations in terms of atomistic interpretation hold true for both: provided that at least secondary structural features can be resolved in a 3D map, a more or less complete atomic model will be expected as the final result of the study. Modelling of carbohydrates into 3D maps can be more complex than

modelling proteins[112], although recent advances in software are closing the gap[107,143,144]. However, to date it remains true that most model building software is protein-centric[131]. As a consequence, the glycan chains in glycoprotein models that have been elucidated before recent developments in carbohydrate validation and modelling software, tend to contain a significant amount of errors: wrong carbohydrate nomenclature[129], biologically implausible glycosidic linkage stereochemistry[130], incorrect torsion[131,132], and unlikely high-energy ring conformations[133]. Early efforts in the validation of carbohydrate structures saw the introduction of online tools such as PDB-CARE[145] and CARP[132]; more recently, we released the Privateer software[24], which was the first carbohydrate validation tool available as part of the CCP4i2 crystallographic structure solution pipeline[146]. In its first release, Privateer was able to perform stereochemical and conformational validation of pyranosides, analyze the glycan fit to electron density map, and offered tools for restraining a monosaccharide's minimal-energy conformation.

While these features were recognised to address some long-standing needs in carbohydrate structure determination[147,148], significant challenges remain, particularly in the scenario where glycan composition cannot be ascertained solely from the three-dimensional map. Unfortunately, this problematic situation happens frequently, especially in view of the fact that the median resolution for glycoproteins (2.4 Å) is lower than that of non-glycosylated – potentially including fully deglycosylated – proteins (2.0 Å)[12]. To date, only one publicly-available model building tool has attacked this issue: the *Coot* software offers a module that will build some of the most common *N*-linked glycans in a semi-automated fashion[107]. Indeed, the *Coot* module was built around the suggestion that only the most-probable glycoforms should be modelled unless prior knowledge of an alternative glycan composition exists, in the form of *e.g.* mass spectrometry data[130].

2.2.3 Harnessing glycomics and glycoproteomics results to inform glycan model building

Current methods used to obtain accurate atomistic descriptions of molecules fall short in dealing with the heterogeneity of glycoproteins. However, there are other methods that have been proven to successfully tackle challenges posed by glycan heterogeneity, with mass spectrometry emerging as the one with most relevance due to its ability to elucidate complete composition descriptions of individual oligosaccharide chains on glycoproteins[149].

Mass spectrometric analysis of glycosylated proteins can be with (glycomics) or without (glycoproteomics) release of oligosaccharides from the glycoprotein. Usually, glycomics and glycoproteomics experiments are carried out together to obtain a complete description of the glycoprotein profile. Glycomics experiments are required to distinguish stereoisomers and linkage information in order to obtain full structural description about a glycan, whereas glycoproteomics are required to establish glycan variability and glycan occupancy at the glycosylation sites of the protein[150]. Typically, these analyses are based on Mass Spectrometry techniques such as electrospray ionization-mass spectrometry (ESI-MS) and matrix-assisted laser desorption ionization MS (MALDI-MS)[150]. Mass spectrometry techniques are best suited for determination of composition of monosaccharide classes and chain length, however in-depth analysis of glycan typically requires integration of complementary analytic techniques, such as nuclear magnetic resonance (NMR) and capillary electrophoresis (CE). Nevertheless, depending on the sample, advanced Mass Spectrometry techniques can be used to counteract the need for complementary analytic techniques. One of the examples of this is tandem mass spectrometry, where glycan fragmentation is controlled to obtain identification of the glycosylation sites and complete description of glycan structure compositions, including linkage and sequence information[151]. Moreover, recent advances in ion mobility mass spectrometry can now also be used for complete glycan analysis[152].

The analysis and interpretation of mass spectrometry spectra produced by glycans is a challenge. Most significantly, in MS outputs, glycans appear in their generalized composition classes, i.e., Hex, HexNAc, dHex, NeuAc, etc. Identity elucidation of generalized unit classes into specific monosaccharide units (such as Glc, Gal, Man, GalNA, etc) requires prior knowledge of glycan biosynthetic pathways[153]. Additional sources of prior knowledge are bioinformatics databases that have been curated through deposition of experimental data. Bioinformatics databases contain detailed descriptions of glycan compositions and m/z values of specific glycans, therefore aiding the process of glycan annotation[154]. Such bioinformatics databases can usually be interrogated using textual or graphical notations that describe glycan sequence. However, due to glycan complexity and the incremental nature of the different glycomics projects numerous notations have been developed over the years – *e.g.* CarbBank[155] utilized CCSD[155], EuroCarbDB[156] and GlycomeDB[157] used GlycoCT[158] (Table 2.1).

Table 2.1: A comparison of the structural information storage capabilities of different sequence formats used in glycobioinformatics. “+” denotes that information can be stored directly without any significant issues, “(+)” denotes that information can be stored indirectly, or there are some issues and “-” denotes that information description in particular sequence format is unavailable. This table is a simplified version of the one originally published by Matsubara *et al.*[1].

Notation	Multiple Connections	Repeating Units	Alternative Residues	Linear Notation	Atomic Ambiguity
CCSD (CarbBank)	-	+	-	+	-
LINUCS	-	+	-	+	-
GlycoSuite	-	-	+	+	-
BCSDB	(+)	(+)	+	+	-
LinearCode	-	-	+	+	-
KCF	+	+	-	-	-
GlycoCT	+	+	+	-	-
Glyde-II	+	+	-	-	-
WURCS 2.0	+	+	+	+	+

Thankfully, data from discontinued glycomics projects are not lost but were integrated into newer platforms, often with novel notations. One such example is GlyTouCan[159], which uses both GlycoCT[160] and WURCS[159] as notation languages. As a result, tools that interconvert between notations were developed to successfully integrate old data onto new platforms. Additionally, the introduction of tools such as GlycanFormatConverter[161] to convert WURCS notations into more human-readable formats has eased the interpretation of glycan databases.

Significantly, the GlyTouCan project aims to create a public repository of known glycan sequences by assigning them unique identification tags. Each identification tag describes a glycan sequence in WURCS notation, and this allows to link specific glycans to other databases, such as GlyConnect[15], UniCarb-DB[162] and others, any of which are tailored to specific flavours of glycomics and glycoproteomics experiments. Ideally, this implementation ends up requiring the user to be familiar with a single notation – WURCS – used to represent sequences of glycans.

2.2.4 From glycomics/glycoproteomics to carbohydrate 3D model building and validation in Privateer

Many fields, for example pharmaceutical design & engineering[163], molecular dynamics simulations[164] and protein interaction studies[165], rely upon structural biology to produce accurate atomistic descriptions of glycoproteins. However, due to clear limitations of elucidating carbohydrate features in MX/Cryo-EM electron density maps, structural biologists are likely to make mistakes. This introduces the possibility of modelling wrong glycan compositions in glycoprotein models, going as far as not conforming with general glycan biosynthesis knowledge. Model building pipelines would therefore greatly benefit from the ability to validate against the knowledge of glycan compositions elucidated via glycomics/glycoproteomics experiments. This warrants the need for new tools that are able link these methodologies, through an intermediate - inter-conversion library.

A foundation for such inter-conversion libraries exists in the form of the carbohydrate validation software Privateer. The program is able to compute individual monosaccharide conformations from a glycoprotein model, check whether the modelled carbohydrates' atomistic definitions match dictionary standards, as well as output multiple helper tools to aid the processes of refinement and model building[24]. Most importantly, Privateer already contains methods that allow extraction of carbohydrate's atomistic definitions to create abstract definitions of glycans in memory, thus already laying a foundation for the generation of unique WURCS notations and providing a straightforward access to bioinformatics databases that are integrated in the GlyTouCan project.

2.3 Methods and results

The algorithm used to generate WURCS notation in Privateer is based on the description published in Tanaka *et al*.[166], with required updates applied from Matsubara *et al*.[1] WURCS was designed to deal with the incomplete descriptions of Glycan sequences emerging from Glycomics/Glycoproteomics experiments (i.e., undefined linkages, undefined residues and ambiguous structures in general). However, the lack of this detail is unlikely to be supported in "pdb" or "mmCIF" format files, which are a standard in structural biology. As a result, "atomic ambiguity" capability (Table 2.1) is not supported in Privateer's implementation. Moreover, Privateer's implementation of WURCS relies on a manually compiled dictionary that translates PDB Chemical Component Dictionary[167] three-letter codes of carbohydrate

monomer definitions found in structure files into WURCS definitions of unique monomers (described as “UniqueRES”[1]).

The WURCS notations are generated for all detected glycans that are linked to protein backbones in the input glycoprotein model. For every glycan chain in the model, the algorithm computes a list of all detected monosaccharides that are unique and stores that information internally in memory. Then, the algorithm calculates unit counts in a glycan chain - how many unique monosaccharides are modelled in the glycan chain, total length of the glycan chain and computes the total number linkages between monosaccharides. After composition calculations are carried out, the algorithm begins the generation of the notation by printing out the unit counts. Then, the list of unique monosaccharide definitions in the glycan chain are printed out by converting the three-letter PDB codes into WURCS-compliant definitions. Afterwards, each individual monosaccharide of the glycan is assigned a numerical ID according to its occurrence in the list of unique monosaccharides. Finally, linkage information between pair monosaccharides is generated by assigning individual monosaccharides a unique letter ID according to their position in the glycan chain. Alongside a unique letter ID, a numerical term is added that describes a carbon position from which the bond is formed to another carbohydrate unit. Crucially, linkage detection in Privateer does not rely at all on metadata present in the structure file. Instead, linkages are identified based on the perceived chemistry of the input model: which atoms are close enough – but not too close – to be plausibly linked.

The generated WURCS string can then be used to search whether an individual glycan chain has been deposited in GlyTouCan. The scan of the repository occurs internally within the Privateer software, as all the data is stored in a single structured data file written in JSON format that is distributed together with Privateer. If the existence of a glycan in the database is confirmed, then the software can attempt to find records about the sequence on other, more specialised databases (currently only GlyConnect) to obtain information such as the source organism, type of glycosylation and glycan core to carry out further checks in the glycoprotein model (Figure 2.2).

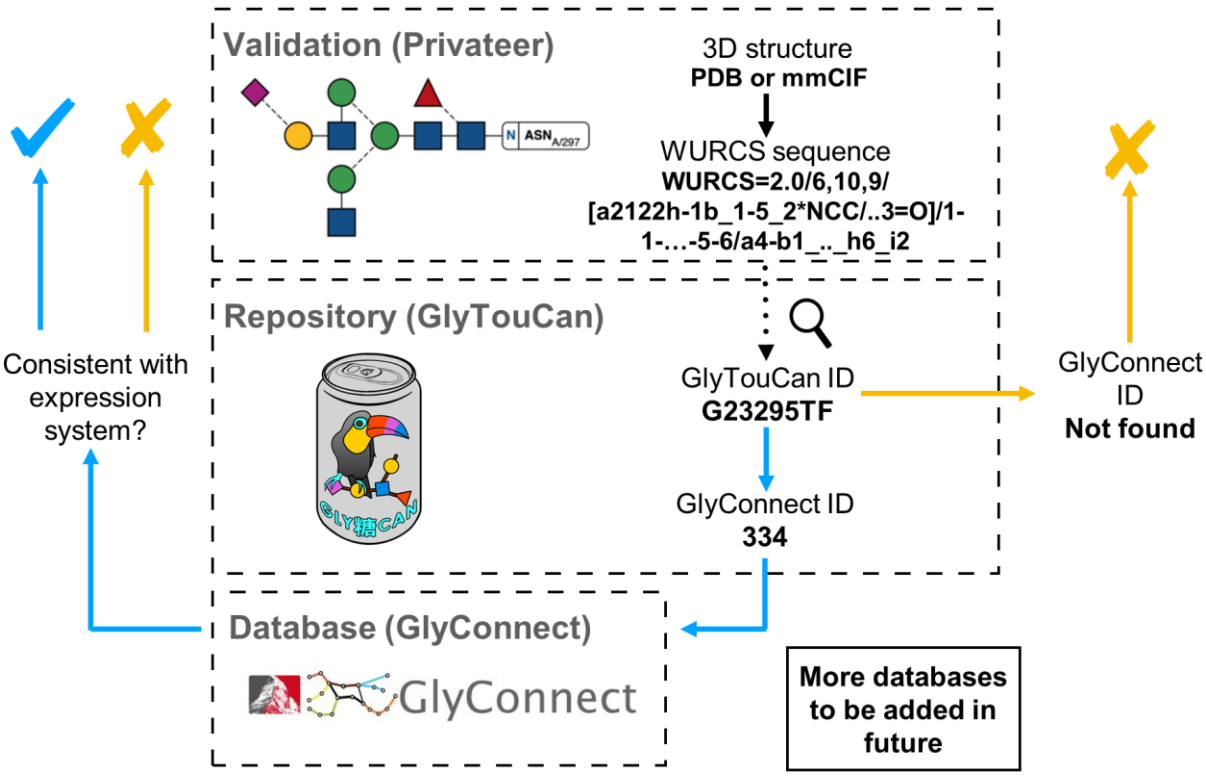


Figure 2.2: A roadmap of the software development project that allows Structural Biologists to quickly obtain detailed information about specific glycans in Glycoprotein models from Glycomics/Glycoproteomics databases. The GlyTouCan (https://glytoucan.org/) and GlyConnect (https://glyconnect.expasy.org/) logos have been reproduced here under explicit permission from their respective authors.

2.3.1 Availability and performance of the algorithm

This new version of Privateer (MKIV) will be released as an update to CCP4 7.1 as soon as the suite starts shipping with Python 3.7 (Privateer is no longer compatible with Python 2.7 due to its recent discontinuation). To demonstrate the capabilities of the computational bridge integrated in the newest version of Privateer (now officially released as a part of CCP4 and CCP-EM suites), it was run on all *N*-glycosylated structures in the PDB solved using MX and cryo-EM[13,14]. The list of structures used in this demonstration was obtained from Atanasova *et al*.[110]. The computational analysis of the demonstration revealed a relatively small proportion of deposited glycoprotein models containing glycan chains that do not have a unique GlyTouCan accession ID assigned, raising questions about the provenance of their structures. Importantly, the majority of the glycan chains that do have a unique GlyTouCan accession ID assigned (except for single residues linked to protein backbones), have also been successfully matched on GlyConnect database (Table 2.2).

Table 2.2: Comparison of successful glycan matches detected by Privateer in GlyTouCan and GlyConnect database. Glycans obtained from glycoprotein models elucidated by X-Ray crystallography and Cryo-EM.

Experimental Technique	Glycan chain length	GlyTouCan ID found	GlyTouCan ID not found	% of GlyTouCan in GlyConnect	Total glycan chains
MX	1	16,797	0	1%	16,797
MX	2	5,870	5	90%	5,875
MX	3	2,550	17	71%	2,567
MX	4	1,012	21	80%	1,033
MX	5	834	72	74%	906
MX	6	460	85	69%	545
MX	7	345	55	77%	400
MX	8	235	25	85%	260
MX	9	164	16	81%	180
MX	10	118	5	92%	123
MX	11	20	5	85%	25
MX	12	8	4	75%	12
MX	13	0	1	0%	1
MX	14	0	0	0%	0
MX	15	2	0	0%	2
MX	16	0	1	0%	1
Cryo-EM	1	2,080	0	3%	2,080
Cryo-EM	2	1,081	0	98%	1,081
Cryo-EM	3	439	0	96%	439
Cryo-EM	4	143	0	93%	143
Cryo-EM	5	146	2	85%	148
Cryo-EM	6	70	1	97%	71
Cryo-EM	7	45	0	100%	45
Cryo-EM	8	26	0	88%	26
Cryo-EM	9	15	1	100%	16
Cryo-EM	10	16	0	100%	16
Cryo-EM	11	4	0	100%	4
Cryo-EM	12	1	0	100%	1
Cryo-EM	13	1	0	0%	1

2.3.2 Examples of use

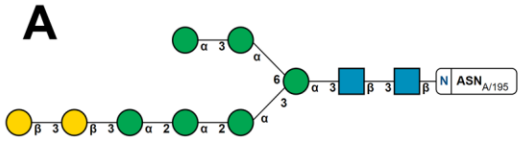
As observed in previous studies, glycoprotein models deposited in PDB feature flaws ranging from minor irregularities to gross modelling errors[12,130,133,168]. Automated validation of minor irregularities was already possible with automated tools such as pdb-care[145], CARP[169], and Privateer[24]. However, automated detection of gross modelling errors is currently a challenge due to the lack of publicly available tools. Our newly developed computational bridge between structural biology and glycomics databases makes detection of gross modelling errors easier, as demonstrated by the following examples.

2.3.2.1 Example 1 - 2H6O:

The glycoprotein model (PDB code 2H6O) proposed by Szakonyi *et al*.[10] contains 12 glycans as detected by Privateer. The model became infamous after it sparked submission of a critical correspondence published by Crispin *et al*.[130] The article contained a discussion about the proposed model containing glycans that were previously unreported and inconsistent with glycan biosynthetic pathways. In particular, the model contained oligosaccharide chains with Man-(1→3)-GlcNAc and GlcNAc- (1→3)-GlcNAc linkages, β-galactosyl motifs capping oligomannose-type glycans and hybrid-type glycans containing terminal Man-(1→3)-GlcNAc[130]. Moreover, the proposed model contained systematic errors in anomer annotations and carbohydrate stereochemistry. To this day, there is still no experimental evidence reported for these types of linkages and capping in an identical context.

The new version of Privateer was run on the proposed model. WURCS notations were successfully generated for all glycans, with only 1 glycan chain out of 12 successfully returning a GlyTouCan ID. Under further manual review of the one glycan, and with help from other validation tools contained in Privateer, it was found to contain anomer mismatch errors (the three-letter code denoting one anomeric form does not match the anomeric form reflected in the atomic coordinates). After the anomer mismatch errors were corrected, the oligosaccharide chain also failed to return GlyTouCan and GlyConnect IDs. The other 11 chains that failed to return a GlyTouCan ID also contained flaws as described previously (Figure 2.3).

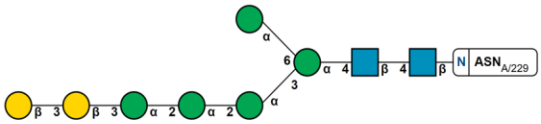
A



WURCS=2.0/3,10,9/[a2122h-1b_1-5_2*NCC/3=O][a1122h-1a_1-5][a2112h-1b_1-5]/1-1-2-2-2-3-3-2-2/a3-b1_b3-c1_c3-d1_c6-i1_d2-e1_e2-f1_f3-g1_g3-h1_i3-j1

GlyTouCan ID: Not Found.

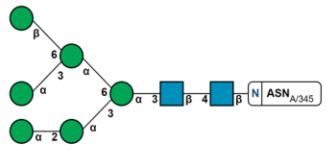
GlyConnect ID: Not Found.



WURCS=2.0/3,9,8/[a2122h-1b_1-5_2*NCC/3=O][a1122h-1a_1-5][a2112h-1b_1-5]/1-1-2-2-2-3-3-2/a4-b1_b4-c1_c3-d1_c6-i1_d2-e1_e2-f1_f3-g1_g3-h1

GlyTouCan ID: Not Found.

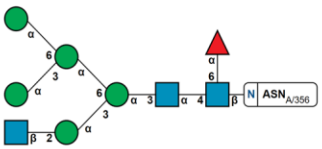
GlyConnect ID: Not Found.



WURCS=2.0/3,8,7/[a2122h-1b_1-5_2*NCC/3=O][a1122h-1a_1-5][a1122h-1b_1-5]/1-1-2-2-2-2-3/a4-b1_b3-c1_c3-d1_c6-f1_d2-e1_f3-g1_f6-h1

GlyTouCan ID: Not Found.

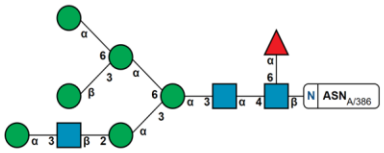
GlyConnect ID: Not Found.



WURCS=2.0/4,9,8/[a2122h-1b_1-5_2*NCC/3=O][a2122h-1a_1-5_2*NCC/3=O][a1122h-1a_1-5][a1221m-1a_1-5]/1-2-3-3-1-3-3-3-4/a4-b1_a6-i1_b3-c1_c3-d1_c6-f1_d2-e1_f3-g1_f6-h1

GlyTouCan ID: Not Found.

GlyConnect ID: Not Found.

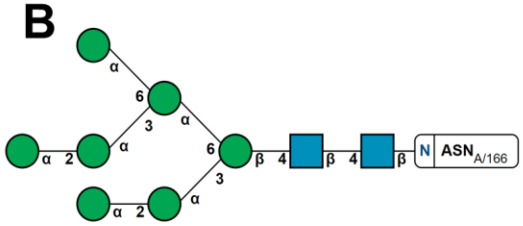


WURCS=2.0/5,10,9/[a2122h-1b_1-5_2*NCC/3=O][a2122h-1a_1-5_2*NCC/3=O][a1122h-1a_1-5][a1122h-1b_1-5][a1221m-1a_1-5]/1-2-3-3-1-3-3-4-3-5/a4-b1_a6-j1_b3-c1_c3-d1_c6-g1_d2-e1_e3-f1_g3-h1_g6-i1

GlyTouCan ID: Not Found.

GlyConnect ID: Not Found.

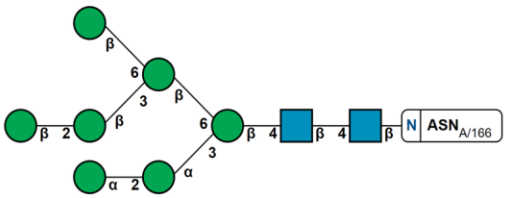
B



WURCS=2.0/3,9,8/[a2122h-1b_1-5_2*NCC/3=O][a1122h-1b_1-5][a1122h-1a_1-5]/1-1-2-3-3-2-2-2-2/a4-b1_b4-c1_c3-d1_c6-f1_d2-e1_f3-g1_f6-i1_g2-h1

GlyTouCan ID: Not Found.

GlyConnect ID: Not Found.



WURCS=2.0/3,9,8/[a2122h-1b_1-5_2*NCC/3=O][a1122h-1b_1-5][a1122h-1a_1-5]/1-1-2-3-3-3-3-3-3/a4-b1_b4-c1_c3-d1_c6-f1_d2-e1_f3-g1_f6-i1_g2-h1

GlyTouCan ID: G63337SS

GlyConnect ID: 1334

Figure 2.3: *N*-linked glycans detected by Privateer in Epstein Barr Virus Major Envelope Glycoprotein (PDB entry: 2H6O[10]). A) Depicts a selection of detected glycan chains that failed to return GlyTouCan and GlyConnect IDs with their WURCS sequences generated by Privateer (graphics taken directly from Privateer's CCP4i2 report). B) Depicts a glycan chain (right) for which a GlyTouCan and GlyConnect ID have successfully been matched with the modelling errors present in the model. After manual rectification of modelling errors (left), the generated WURCS sequence for the glycan fails to return GlyTouCan and GlyConnect IDs. Highlighting in red shows the locations in WURCS notation where both glycans differ.

The analysis of this PDB entry highlights the kind of cross-checks that could be done by Protein Data Bank annotators upon validation and deposition of a new glycoprotein entry. It should be recognised that PDB annotators might not necessarily be experts in structural glycobiology. The fact that these glycans could not be matched to standard database entries should be enough to raise the question with depositors, and at the very least write a caveat on a deposited entry where glycans could not be correctly identified. Furthermore, despite the example showing just *N*-glycosylation, other kinds of glycosylation are searchable as well, and therefore this tool could shed much needed light on the validity of models representing more obscure types of modifications.

2.3.2.2 Example 2 - 2Z62:

Successfully matching WURCS string to a GlyTouCan ID, should not be a sole measure of a structure's validity. GlyTouCan is a repository of all potential glycans collected from a set of databases, its entries often representing glycans. Therefore, the correctness of composition should be critically validated against information provided in specialized and high-quality databases such as GlyConnect[15] and UniCarbKB[170]. The computational bridge provides direct search of entries stored in GlyConnect, with plans to expand this to more databases in the near future.

An example, where sole reliance on detection of a glycan in GlyTouCan would not be sufficient is rebuilding of the **2Z62** glycoprotein structure[11] to improve model quality[12] (Figure 2.4). Analysis of the original model generated the GlyTouCan ID **G28454KX**, which could not be detected in GlyConnect. The automated tools used by PDB-REDO slightly improved the model by renaming one of the fucose residues from FUL to FUC, due to an anomer mismatch between the three-letter code and actual coordinates of the monomer. The new model thus generated the GlyTouCan ID **G21290RB**, which in turn could be matched to the GlyConnect ID **54**. Under further manual review of mFo-DFc difference density map, a (1–3)-linked fucose was added, along with additional corrections to the coordinates of the molecule[12]. The newly generated WURCS notation for the model returned a GlyTouCan ID of **G63564LA**, with a GlyConnect ID of **145**. The iterative steps taken to rebuild the glycoprotein model have been demonstrated in Figure 2.4. Because the data in GlyConnect is approximately 70% manually curated by experts in the field[15], a match of a specific glycan in this database is likely a valid confirmation of a specific oligosaccharide composition and linkage pattern found in nature.

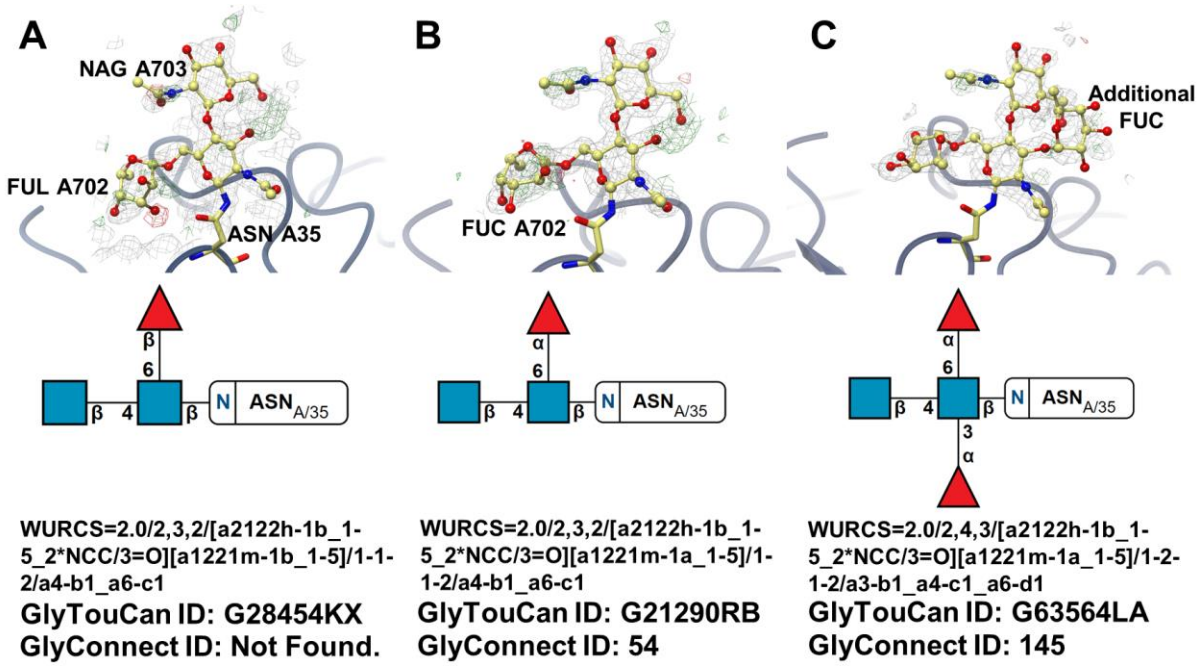


Figure 2.4: An *N*-linked glycan attached to Asn35 of human Toll-like receptor 4 (A: PDB entry 2Z62[11]). Model iteratively rebuilt by PDB-Redo as shown in steps B and C[12]. Pictures at the top depict glycoprotein models of the region of interest and electron density maps of the glycan chain (grey - 2mFo DFc map, green and red - mFo DFc difference density map), pictures at the bottom depict SNFG representations of glycan chains, their WURCS sequence and accession IDs to relevant databases (taken directly from Privateer's CCP4i2 report).

2.4 Conclusions and future work

The mirrors of GlyConnect and GlyTouCan were obtained thanks to the public access to the API commands which allowed the creation of scripts that automated the query of the entries stored in the databases with relative ease. However, integration of additional databases might require support from the developers of those databases. Support for lipo-polysaccharides and polysaccharides may be added in future too, owing to the general purpose of the integrated databases – i.e., they are not limited to protein glycosylation.

Currently, the generated WURCS strings are matched against an identical sequence in the database. This means that, if a glycan model has a single modelling mistake, for example at one end of the chain, but is correct elsewhere, the current version of the software would still fail to return a match. This issue has been solved in the development version by the incorporation of a subtree matching algorithm, which will reveal modelling mistakes at specific positions of the glycans, and report these to the user.

Currently all the developments outlined in this work are accessible through Privateer's command line interface and general releases of CCP4 and CCP-EM software suites[13,14].

2.5 Addendum

2.5.1 Permutation search algorithm

Next, further mechanisms were investigated to support the iterative glycan building process. As a result, a permutation algorithm was developed to highlight potential glycan compositional errors in terms of linkage, anomer and monosaccharide type descriptors based on data available in the GlyConnect database.

2.5.2 Current implementation

The algorithm is initiated for every glycan chain in the input structure file only if originally modelled glycan composition is not detected on the GlyConnect database. Three types of permutations are carried out to modify input glycan structures: anomer permutations, residue permutations and residue deletions. Anomer permutations in various combinations are carried out on all monosaccharide units, where for example α-Man (MAN) is replaced with β-Man (BMA). When all anomer permutation combinations are exhausted, the algorithm in various combinations replaces specific monosaccharide units for similar sugars, for example α-Man (MAN) sugars for α-Glc (GLC) or α-Gal (GLA). For every combination of residue permutation, anomer permutations are re-computed. After all possible anomer and residue replacement permutations are explored for a glycan of specific length, it then is trimmed by a single sugar, where a shorter glycan's anomer and residue permutations are entirely explored. This process is recursively repeated, until the permuted glycan becomes too short for further computations of potential permutations.

The algorithm returns a list of generated permutations that were successfully located in the GlyConnect database. Qualitative assessment of the returned list of permutations detected on GlyConnect database enables the determination of potentially mismodelled fragments of modelled glycan. The resulting implementation has been integrated into the most recent update of Privateer and distributed through CCP4i2[146] and CCP-EM[14] graphical user interfaces (Figure 2.5).

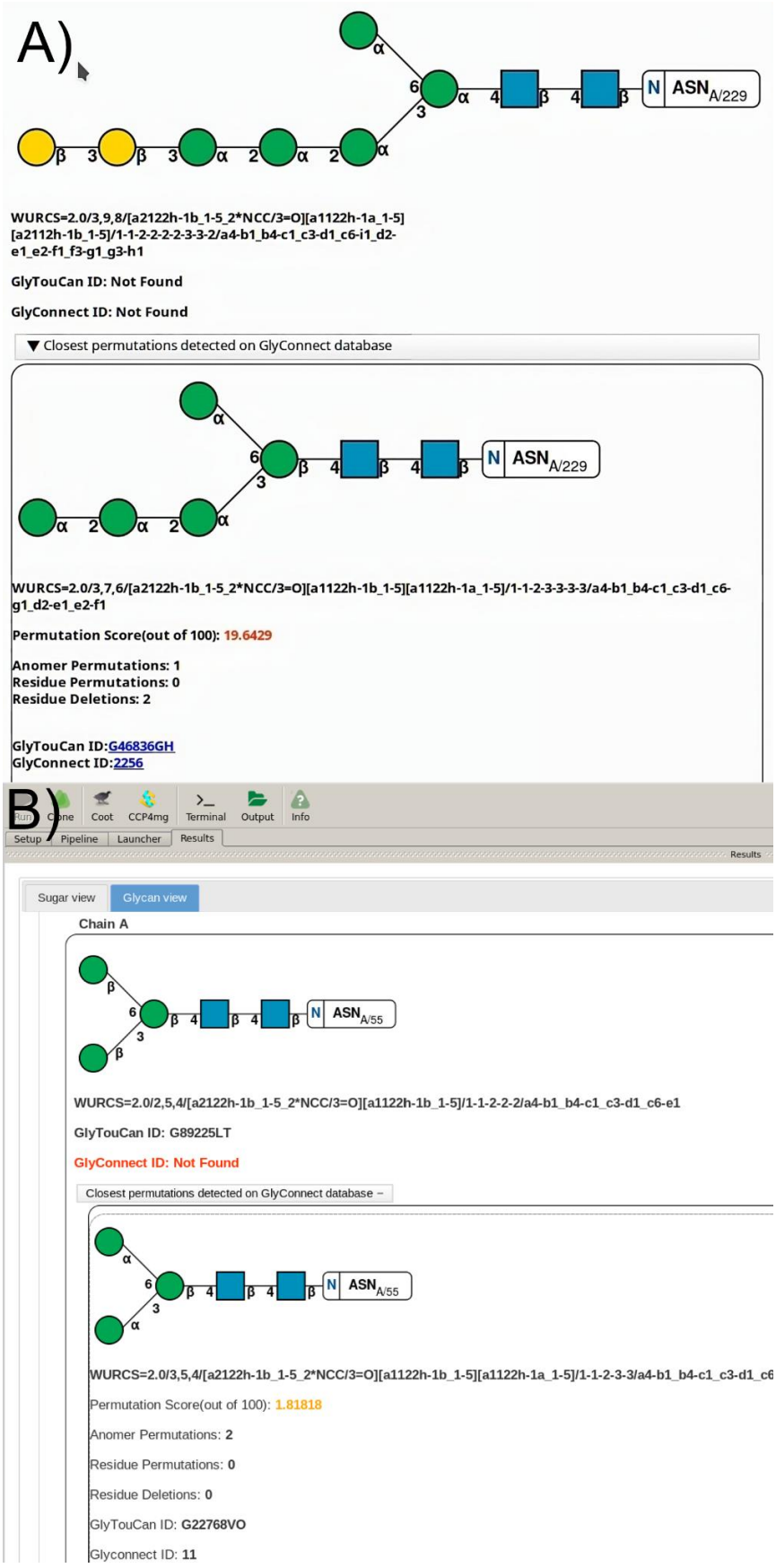


Figure 2.5: Implementation of Privateer's glycan permutation algorithm in A) CCP4i2 and B) CCP-EM graphical user interfaces[13,14]. A) Screenshot of partial permutation algorithm output

for a modelled glycan in Epstein Barr Virus Major Envelope Glycoprotein attached to ASN229 (PDB entry: 2H6O[10]). The permutation algorithm qualitatively demonstrates that β-Gal capping at β1-3 linkage configuration is inconsistent with data deposited in GlyConnect[15], leading to β-Gal capping elimination. In addition, the permutation algorithm reveals a potential modelling mistake, where a branching mannose should be modelled as β-Man rather than α-Man as indicated by an anomeric permutation. B) Screenshot of partial permutation algorithm output for a modelled glycan in human gamma-secretase complex attached to ASN55 (PDB entry: 5A63[16]). The algorithm indicates that terminal β-Man sugars should be modelled as α-Man sugars instead, according to the anomeric permutations.

2.5.3 Potential improvements

Currently, the permutation algorithm is only capable of trimming the modelled glycan. The search space and performance of the algorithm could significantly be improved by considering the addition of monosaccharide units at various linkage configurations to make the search algorithm more performant in terms of potential glycan compositions that could be modelled. Such a solution outside of Privateer has already been implemented in the GlySTreeM knowledgebase[171].

2.6 Algorithmic implementation to integrate GlyTouCan and GlyConnect data in Privateer software.

The implementation of the 'generate_wurcs' function within the Privateer software is a fundamental implementation responsible for translating all modelled glycan structures into a standardized WURCS notations, thereby enabling seamless integration with glycan databases such as GlyTouCan and GlyConnect.

At the outset, the function initiates the WURCS string construction by declaring the version of the notation used. The algorithm dynamically calculates the length of the glycan and identifies unique monosaccharide residues by parsing the glycan structure, which is internally represented as a linked list. This process is crucial for understanding the glycan's complexity and ensuring that the notation accurately reflects its monomeric composition.

Following this initial setup, the function iteratively constructs a description of each unique residue, appending it to the WURCS string within brackets.

The algorithm concludes in the assembly of the complete WURCS string, which includes all necessary linkage information. For glycans consisting of a single monomer, the function is concluded early, as linkage information becomes irrelevant. The generated WURCS string then can be used to search the internal Privateer database to find an associated GlyTouCan identifier if it at all exists.

Upon generating the WURCS notation with the 'generate_wurcs' function, the Privateer software integrates with glycan databases using a NoSQL approach through a JSON file for database queries. This step enables the identification of corresponding GlyTouCan identifiers for modeled glycan structures.

The process involves using the WURCS string to search the internal Privateer database for a matching GlyTouCan ID. The print_output_from_database function is central to this process, retrieving the GlyTouCan ID from the JSON-formatted database. When a match is identified, the function outputs the GlyTouCan Accession ID and a direct link to its entry on the GlyTouCan website.

Additionally, the function checks for a corresponding GlyConnect ID. If found, it provides the GlyConnect ID and a link to its entry, facilitating access to further information about the glycan. If a GlyTouCan ID does not have a corresponding entry on GlyConnect, the software notes the absence of a GlyConnect deposition and initiates the aforementioned permutation algorithm to find the closest match.

This functionality allows for the linking of computational glycan models with their entries in public glycan composition datastores, providing a practical tool for researchers to validate and explore glycan structures further.

Investigation of *N*-glycan processing using Protein Data Bank data

This chapter describes the efforts to investigate the potential influence of the protein environment on the glycan processing machinery using deposited structures of glycoproteins as snapshots. In order to gain insight into the determinants of glycan processing within protein structures, structural bioinformatics algorithms were developed to extract contextual glycosylation information from glycoprotein structures deposited to Protein Data Bank (PDB). Valuable clues about *N*-glycan processing were uncovered by investigating the neighbouring amino acid context in the vicinity of modelled *N*-glycans.

3.1 Introduction

N-glycosylation is often required for the efficient transport of proteins through the secretory pathway, and alterations in the status of *N*-glycosylation can lead to impaired trafficking of proteins[172]. Likewise, investigation of factors affecting glycan maturation is as important due to specific *N*-glycans modulating various aspects of protein properties, particularly protein stability and interaction with other molecules. Different types of *N*-glycans can determine the specificity and affinity of protein-protein interactions, affecting cellular processes and signalling pathways[173]. For example, the presence of specific *N*-glycans can either activate or inhibit immune cell receptors, modulating immune responses and contributing to the regulation of inflammation and other immune processes[174]. As a result, aberrations in the glycan processing machinery can lead to various pathologies[175].

In order to study the effects of aberrations in *N*-glycosylation machinery at the molecular level, detailed and complete descriptions of glycoprotein structures are required. Nevertheless, due to *N*-glycan heterogeneity and flexibility, it currently is a significant challenge to obtain detailed enough information about *N*-glycan compositions at specific glycosylation sites. To address this shortfall, several computational tools have been developed to predict, analyse, and annotate glycosylation sites in protein sequences, as well as tools to predict glycan profiles within cells[176]. However, no tools exist to predict *N*-glycan compositions on a site-specific basis. The influence of protein structure may be one of the factors influencing glycan processing. For this reason, it is hoped that the investigation into

the final products of glycan processing machinery deposited to Protein Data Bank (PDB) can reveal any potential clues that could be used to build predictive models.

The idea of protein structure playing a crucial role in determining the products of glycosylation machinery has been investigated in the past. Particularly, in the investigation carried out by Hang *et al*., it was experimentally demonstrated that changes to protein structure alters the glycan profiles of specific glycosylation sites[177]. One specific glycosylation site (labelled as S4 in the study) contained less processed *N*-glycan relative to other glycosylation sites. The molecular dynamics studies attributed terminal mannose residues of the *N*-glycan making long-lasting contacts with specific amino acids of two different domains (a and b) of the protein. The interactions were hypothesised to reduce accessibility of the glycan to processing enzymes. Upon *in vivo* removal of the a-domain, the site-specific glycan profile changed to more processed glycan, in line with glycan profiles from other glycosylation sites[177]. The study of protein playing a role in glycan processing was further built upon by Suga *et al*., who analysed multiple glycoprotein structures deposited to PDB, concluding that *N*-glycan processing is largely explained by the solvent accessibility of glycosylated Asn residue in combination with solvent accessibility of nascently attached oligosaccharide[178]. These studies build a foundation of understanding how protein structure may have a direct impact on specific glycosylation sites containing *N*-glycans processed to a varying degree. Building upon the foundation of understanding further, this chapter is an attempt to find specific amino acid associations as features of protein structure having an impact on glycan maturation through a novel analysis approach described in this chapter.

3.1.1 Rationale for a novel analysis approach

In order to gain meaningful insight into structural determinants of glycan processing, it was imperative to gather a high-quality dataset from the PDB that would be representative of products that are likely to result from actions of the glycosylation machinery. Throughout the PhD, new features and improvements have been made to the Privateer software that allow it to interact with other databases and software. In particular, the implementation of WURCS notation generation to query the GlyConnect database to enable automated validation of *N*-glycan compositions. As a result of these contributions, Privateer became a powerful tool to collect and assess the quality of all instances of modelled *N*-glycans that could be utilised to analyse the *N*-glycosylation machinery. In addition, new features were integrated into Privateer to analyse the structural neighbourhood context of glycan moieties.

Mammalian *N*-glycans universally contain a $Man_3GlcNAc_2$ core, regardless of the extent of processing, as shown in Figure 1.5. The characteristic features that define different types of glycans are located towards the terminal ends of oligosaccharides. As a result, enzymes associated with the glycan processing machinery are only acting on sugars that are located beyond the $Man_3GlcNAc_2$ core. Therefore, the investigation of the glycan processing machinery on glycan moieties should only consider the structural context of sugars at the terminal ends of modelled *N*-glycans. In contrast to the investigation carried out by Suga *et al.*, this approach is much more local and is therefore likely to uncover specific amino acid type identities that may have an impact on glycan processing.

3.2 Aims

The aim of the work presented in this chapter was to analyse the structural context in the vicinity of *N*-glycan termini to potentially reveal determinants of glycan processing and its associated products. In order to achieve this aim, a dataset of high-quality model instances of *N*-Glycosylation were retrieved from PDB. A successful curation of the dataset enabled neighbourhood context computations of modelled *N*-glycans at the terminal ends, to analyse the potential relationship of neighbouring amino acids and the associated *n*-glycan type.

3.3 Methods

3.3.1 Glycosylation data accumulation from PDB

The PDB contains structures of various glycoproteins, including *N*-linked and *O*-linked glycosylated proteins. These structures have been determined using techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM).

As previously discussed, the challenges associated with determining glycoprotein structures are multi-faceted. Due to inherent flexibility and heterogeneity of glycans, associated experimental density regions are often missing or poorly resolved. As a result, the structural data for glycoproteins in the PDB is often incorrect. In some instances, the PDB contains glycan compositions incompatible with the glycan biosynthesis machinery of the expression system[130]. While the potential solution to this aspect has been presented in Chapter 2, during the research additional considerations for modelled *N*-glycosylations emerged. Specific PDB depositions may model glycan-lectin recognition, rather than glycosylation.

Automatic discrimination of glycan-lectin binding from glycosylation on a massive scale is not straightforward based on metadata provided by PDB files. Therefore, the methods presented in this chapter describe an automated procedure that was used to curate a dataset of modelled *N*-glycosylations in glycoprotein structures deposited to PDB to enable biologically meaningful interpretation of glycan processing at scale.

3.3.2 Oligosaccharide instance accumulation from PDB

A local copy of atomic coordinates in PDB Format were downloaded from RCSB PDB on 21st of July 2022 using the RSYNC protocol[179]. The atomic coordinates were downloaded as compressed '.ent.gz' tarball files.

For every file in the local mirror of PDB database, a bespoke Python script would temporarily extract a PDB file from tarball and pass the generated path to a function that would execute glycan detection features of Privateer. If Privateer had detected at least one glycan instance in atomic coordinate file, relevant information would be collected and stored, such as: glycosylation type; glycan composition in WURCS notation as generated by Privateer; GlyTouCan and GlyConnect identifiers; if available, glycosylation site information in terms of amino acid type, sequence number and chain identifier; amino acid sequence of protein chain containing target glycosylation site.

The algorithm to generate WURCS notation in Privateer is different from the implementation in PDB2Glycan, which is used by wwPDB[180]. Because of the differences in implementation, some modelled glycans may produce different WURCS notations between Privateer and PDB2Glycan. In this instance, it was decided to treat WURCS notations generated by PDB2Glycan as ground truth due to the software being implemented by the creators of WURCS notation and endorsed by the WURCS working group, as well as PDB2Glycan being used to carry out carbohydrate remediations for the wwPDB[181,182]. Therefore, to ensure agreement between PDB2Glycan and Privateer, every PDB identifier in the local mirror of PDB database, was queried through the API of RCSB PDB to extract oligosaccharide entries, in terms of WURCS notation and GlyTouCan identifier, as generated by PDB2Glycan software[180].

3.3.3 Addition of information for evaluation of redundancy and experimental quality in glycosylated PDB depositions

For every instance of "*N*-glycan" oligosaccharide detected by Privateer in the accumulated dataset, an algorithm queried RCSB PDB GraphQL API to relate modelled protein chains to UniProt identifiers and other biologically relevant data[183,184]. The following arguments were used: PDB identifier; chain identifier and amino acid sequence number of the glycosylation site. The query would return the following output: if available, UniProt identifier of glycosylated protein chain; target organism of the glycoprotein; expression system of the experiment used to obtain the structure; common name of glycosylated protein chain as described in PDB file; common name of glycosylated protein chain as described in UniProt; a binary descriptor whether the glycosylation site location was successfully aligned within UniProt entry sequence; a binary descriptor whether glycosylated protein chain is a fusion of multiple proteins.

In addition, every PDB identifier was queried for experimental method (X-Ray Crystallography, Cryo-EM, NMR); resolution of the associated data, if available (X-Ray Crystallography and Cryo-EM); and EMDB identifier if experimental method was Cryo-EM.

3.3.4 Enrichment of *N*-Glycosylations in PDB depositions

The Privateer software uses the identity of amino acid forming the glycosidic bond between an amino acid residue and sugar to assign the type of glycosylation. During manual review of the collected data, it was found that some cases of "*N*-glycans" assigned by Privateer and PDB2Glycan were modelling recognition interactions between instances of lectin and isolated glycopeptides. The detection and removal of such cases was complicated by the fact that the modelled glycopeptide fragments contained a singular Asn (asparagine) amino acid residue with an identical chain label identifier to one of the modelled instances of lectin protein chains. Therefore, to ensure that "*N*-glycan" oligosaccharide assignments were indeed modelling *N*-Glycosylations, rather than lectin-glycopeptide interactions, an algorithm was implemented to compute amino acid neighbours of the sugar engaged in the *N*-glycosidic linkage with an Asn residue. If the algorithm found that the Asn residue engaged in *N*-glycosidic linkage had no other amino acid neighbours with an identical chain identifier and sequence number within the offset of range -5 to +5, then it was deemed to be an instance of glycopeptide engaged in lectin recognition. If at least one amino acid neighbour

was detected containing an identical chain identifier and having an offset of range -5 or +5 in sequence number to the Asn residue engaged in *N*-glycosidic linkage, then it was deemed to be an instance of actual *N*-glycosylation. The remaining oligosaccharides representing *N*-Glycosylation instances were queried through GlyConnect API using GlyConnect identifiers to retrieve *N*-glycan composition categories. *N*-glycan processing can be characterised by the processing of High-Mannose *N*-glycans to Processed *N*-glycans of the following types: Hybrid, Pauci-Mannose and Complex. Due to challenges of glycan moiety resolvability in structural biology, an absolute majority of *N*-Glycosylation instances retrieved from PDB did not extend beyond the $Man_3GlcNAc_2$ core which would enable the assignment of an *N*-glycan type. Due to the identical circumstances, modelled instances of Pauci-Mannose could not be confidently attributed to *N*-glycan biosynthesis machinery products and were not considered for further analysis. Therefore, the investigation exclusively considered modelled *N*-glycan instances that at minimum contained 6 monosaccharide units in total, with the shortest glycan in the analysis having the composition of $GlcNAc_1Man_3GlcNAc_2$. The only types of *N*-glycans that extend beyond the $Man_3GlcNAc_2$ core are: High-Mannose, Hybrid and Complex. However, due to a severe under-representation of Hybrid *N*-glycans, a decision was made to group Hybrid and Complex *N*-glycans, under a common label of "Processed". Therefore, the investigation of *N*-glycan processing based on data retrieved from PDB, only considered the following types of *N*-glycans: High-Mannose and Processed (composed of Complex and Hybrid *N*-glycan instances).

3.3.5 Compilation of a non-redundant glycoprotein dataset

The sampling of glycoprotein space in PDB is uneven, with some proteins such as IgG having multiple associated depositions in thousands, and some proteins only having a single associated deposition. Therefore, to ensure non-redundancy, enriched glycoproteins containing *N*-glycans were clustered by the UniProt's "recommended name". If a cluster contained multiple associated PDB depositions, the entries were sorted by resolution, selecting the representative structure with the highest resolution value. Some clusters contained multiple glycosylation sites in terms of the Asn residue involved in the *N*-glycosidic linkage, with a varying degree of representation in associated PDB depositions. Therefore, the representative glycosylation site was selected by picking the most popular glycosylation site across multiple structures associated with the cluster.

3.3.6 Analysis of terminal *N*-glycan neighbourhood

To analyse how the 3D structure of a protein chain might impact glycan processing, neighbourhood information was computed up to 10 Å radius originating from the C4 atom of terminal sugars. Due to glycan branching, multiple origin points associated with terminal sugars occur. Unique instances of amino acid neighbour detections were ensured by comparing the sequence numbers in a temporary list and if a detection was made at a closer distance by another origin point, the calculated distance between sugar and closest amino acid atom would be updated by the closer hit. In addition, it was also ensured that the only contacts that were considered, were the ones that were detected in the same protein chain identifier as the asparagine residue engaged in *N*-glycosidic bond with glycan under consideration. If the terminal contact list was found to contain less than 75% of detected amino acid neighbourhoods from an identical protein chain identifier as asparagine residue engaged in *N*-glycosidic bond, then the entry would be removed from further analysis. This step aims to isolate modelled glycans that are not potentially affected by factors such as crystal contacts or glycan being located at the interface of a dimer.

Due to redundancy of chemical features of residues, in some analyses specific amino acids were clustered according to the following labels: **Sulphuric** (Cysteine); **Featureless** (Glycine); **Positive** (Histidine, Lysine, Arginine); **Negative** (Glutamate, Aspartate), **Polar uncharged** (Serine, Threonine, Glutamine, Asparagine); **Aromatic** (Phenylalanine, Tryptophan, Tyrosine); **Hydrophobic** (Alanine, Valine, Leucine, Isoleucine, Proline, Methionine).

Enrichment analyses relied upon the calculation of relative occurrence of each amino acid or amino acid type up to a selected distance radius, allowing for derivation of the frequency of amino acids close to the terminal end of modelled glycans. Furthermore, the frequency of amino acids or amino acid type occurrences within the associated protein chain sequences was additionally calculated to provide a basis for comparison. This enabled the computation of the enrichment ratios of amino acids or amino acid clusters in the terminal regions of modelled *N*-glycans relative to their overall abundance in the protein sequences. The enrichment ratios offer insights into the propensity of specific amino acids to be located near *N*-glycan termini. The equation to calculate enrichment ratio is described below:

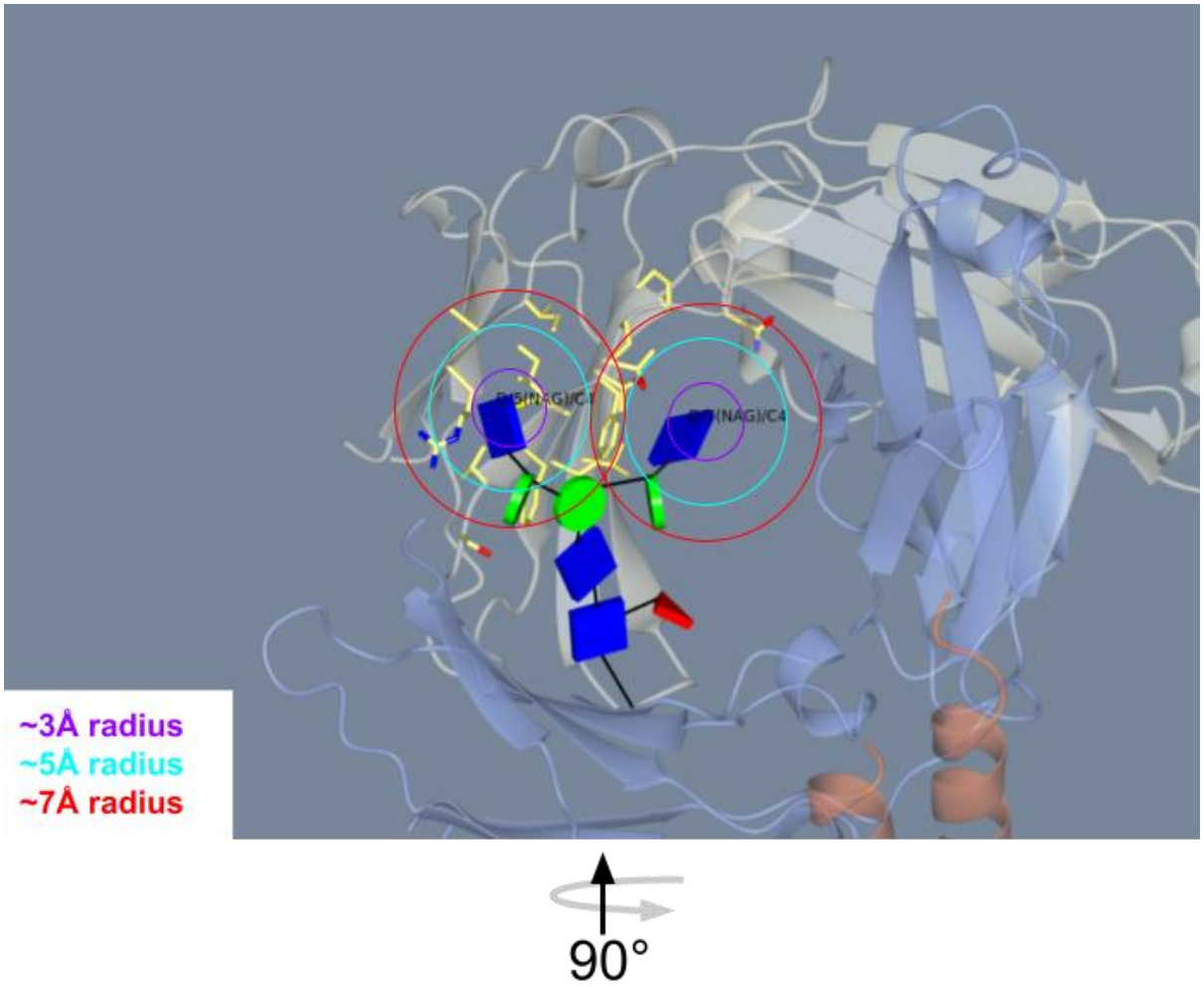
$$(1)\ Freq\ in\ Neighbourhood\ \% = \frac{Total\ instances\ of\ amino\ acid\ in\ neighbourhood}{Total\ amino\ acids\ in\ analysed\ neighbourhoods} \times 100$$

$$(2)\ Freq\ in\ Sequence\ \% = \frac{Total\ instances\ of\ amino\ acid\ in\ analysed\ sequences}{Total\ amino\ acids\ in\ analysed\ sequences} \times 100$$

$$(3)\ Enrichment\ Ratio = \frac{Freq\ in\ Neighbourhood\ \%}{Freq\ in\ Sequence\ \%}$$

In some analyses, the 5 Å, 7 Å and 9 Å radius distance thresholds are applied in order to greatly aid the interpretability of results. The 5 Å threshold cutoff can be thought of in the context of detecting amino acid residues that are separated by a single monosaccharide unit, as the shortest distance from O1 atom to O4 atom in idealised α-Man (MAN) obtained from CCP4-ML is 5.70 Å[185]. Therefore, neighbours at this distance are most likely to be located at the interface of *N*-glycan processing.

In visual terms, the neighbourhood scan at the terminal ends of modelled *N*-glycans is illustrated in Figure 3.1.



90°

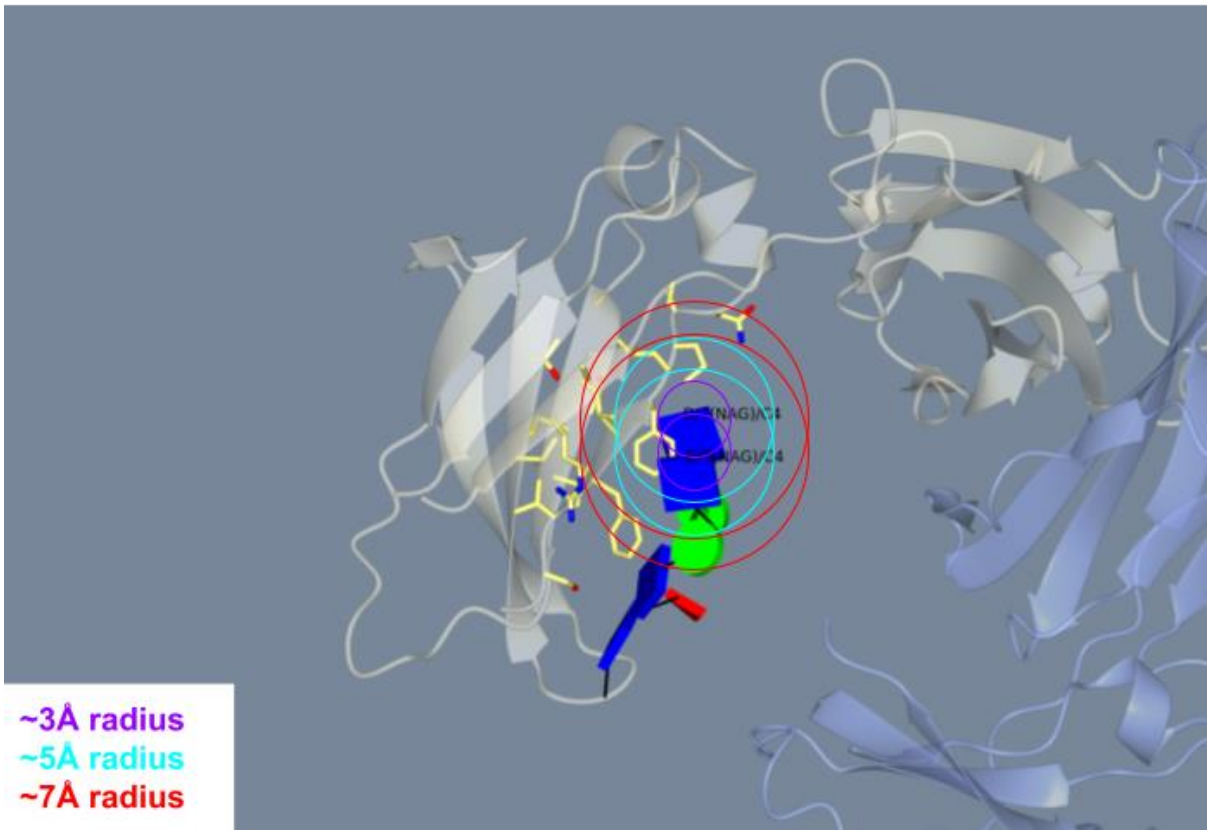


Figure 3.1: Approximate visual illustration of neighbourhood scan at the terminal ends of modelled glycans demonstrated in crystal structure of a mutant mIgG2b Fc heterodimer in complex with Protein A peptide analog Z34C (PDB ID: 5UBX[17]). The amino acid residues displayed in stick and ball representation on chain B of the model were detected within 10 Å distance radius from the origin points of C4 atoms of two terminal GlcNAc sugars. Coloured circles are approximate representations of various distance cutoff thresholds.

3.4 Results

3.4.1 Automated *N*-glycosylation dataset curation

In order to investigate the *N*-glycan processing machinery using structures deposited on the PDB, a validated and representative dataset needed to be curated. Privateer was successfully used to parse 189,255 PDB files to retrieve 98,945 detected instances of oligosaccharides. Based on Privateer's automatic assignment of oligosaccharide type, most of the oligosaccharides modelled in the PDB are "*N*-glycans" as shown in Table 3.1

Table 3.1: Summary of oligosaccharide instances detected in Protein Data Bank (PDB). Column legend: 'Oligosaccharide type' is the type of glycosylation assigned by Privateer, determined by the amino acid character in the vicinity of the glycan root; 'Total structures' is the number of unique PDB identifiers associated with specific oligosaccharide type; 'Unique compositions' is the number of unique oligosaccharide compositions determined by the WURCS notation generated by Privateer associated with specific oligosaccharide type.

Oligosaccharide type	**Instances**	**Total structures**	**Unique compositions**
N-glycan	74,178	9,323	648
O-glycan	1,971	613	126
C-mannose	211	39	8
S-glycan	12	12	6
Ligand	22,573	8,702	1,540
Total	989,945	18,689	2,328

The initial inspection of retrieved oligosaccharide compositions by Privateer revealed some differences between oligosaccharide compositions defined by wwPDB, when judged by differences in generated WURCS notations. Therefore, oligosaccharide detections

generated by Privateer and RCSB PDB oligosaccharide instances were cross-referenced. The comparison revealed that 622 structures had at least one oligosaccharide instance, where WURCS notation generated by PDB2Glycan was not detected by Privateer. As a result, 851 "*N*-glycan" oligosaccharide instances were eliminated from the aggregated dataset, with the remaining 73,327 "*N*-glycan" oligosaccharide instances used for further filtering procedure.

The possible factors resulting in different WURCS notations were manually investigated. The manual investigation revealed that different "*N*-glycan" composition definitions can be generated, depending on differences in internal parameters. In the example, shown in Figure 3.2, PDB2Glycan generated three different oligosaccharide entities, while Privateer generated a single oligosaccharide entity, most likely due to the differences in the definition of glycosidic bond distance threshold used in generating oligosaccharide representations.

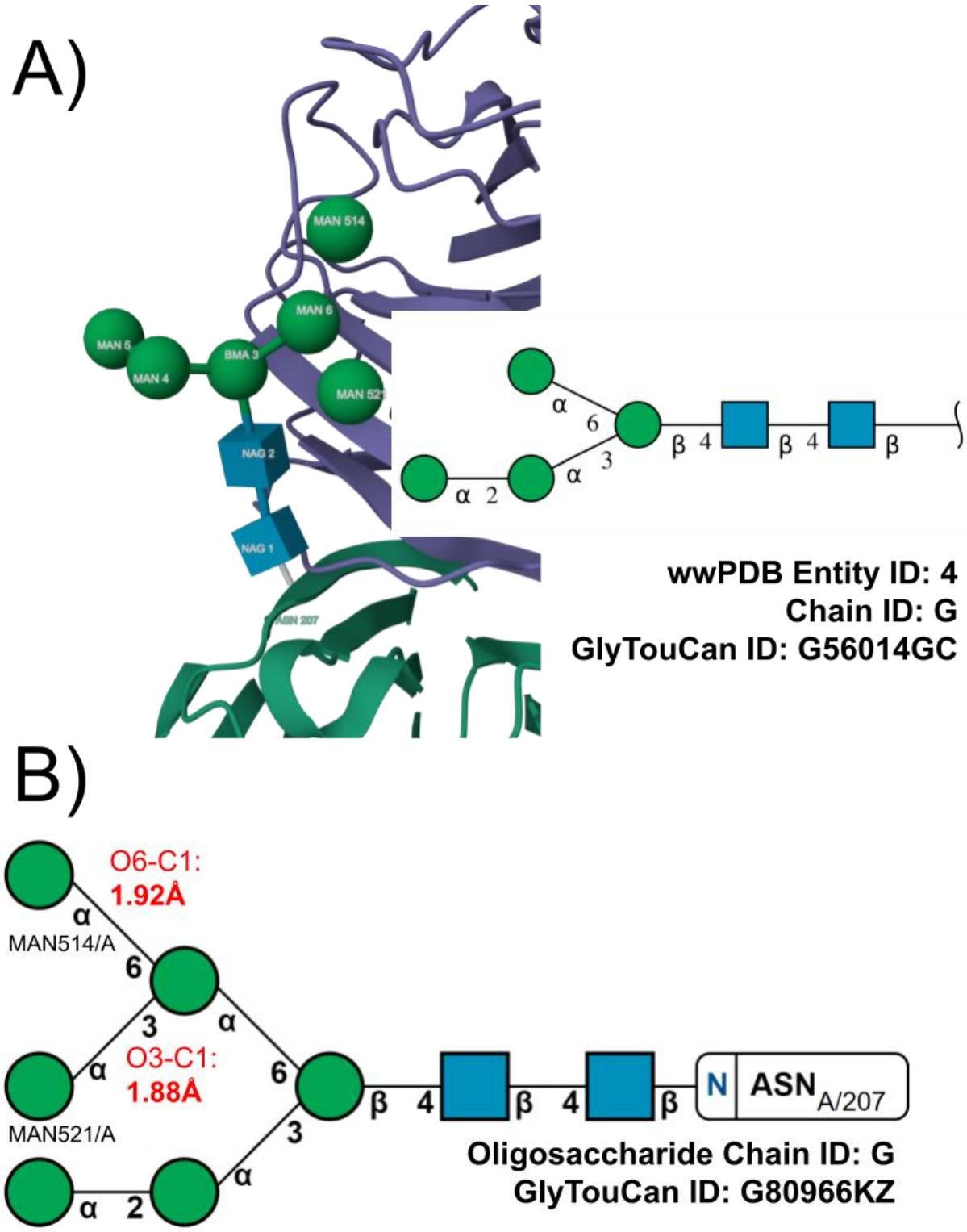


Figure 3.2: Comparison of oligosaccharide entity detection between Privateer and wwPDB in Neuraminidase structure from English duck subtype N6 (PDB ID: 1V0Z[18]). The differences in detection can most likely be explained by potential glycosidic bond distances being inconsistent with expected glycosidic bond linkage distances in the deposited structure (highlighted in red labels), as the internal parameter in Privateer – itself a validation software

that needs to deal with user-introduced problems, including bonds that are too long – was more relaxed in comparison to PDB2Glycan. A) 3D-SNFG and 2D-SNFG representations displayed in wwPDB. B) 2D-SNFG representation generated by Privateer.

On the other hand, Privateer was unable to successfully generate WURCS notation for oligosaccharide entities containing unusual monosaccharides or linkages, mostly associated with “ligands”. The principal cause for the failure is Privateer’s internal database responsible for conversion of PDB three letter code to WURCS UniqueRES being out of date at the time of data collection. Thankfully, this issue had a relatively minor impact on collected “*N*-glycan” structures, as the number of monosaccharide building blocks and linkage configurations associated with *N*-glycosylation is finite and well defined. This result demonstrates that Privateer has potential to be used as a cross-validation tool for agreement with PDB2Glycan in isolation of glycoprotein structures for potential remediation of atomic coordinates to ensure chemical property consistency. Finally, the manual investigation affirmed the decision to use oligosaccharide entities generated by PDB2Glycan as ground truth.

The remaining 73,327 “*N*-glycan” oligosaccharide instances were further filtered to eliminate entries that did not have a GlyConnect identifier, resulting in removal of 4,499 instances. This step ensured that “*N*-glycan” oligosaccharide compositions were compliant with the known products of glycan biosynthesis machinery. To ensure that “*N*-glycan” oligosaccharide instances were indeed modelling *N*-Glycosylations, rather than lectin binding, remaining 68,828 entries were scanned for potential representation of lectin-glycopeptide interaction, resulting in elimination of further 1,693 entries. An example of such an instance and the need for an elaborate algorithm is demonstrated in Figure 3.3. In summary, after exhaustive filtering 67,135 instances of protein *N*-Glycosylation were retrieved from PDB, as shown in Table 3.2.

Table 3.2: Summary of filtering steps for enrichment of *N*-Glycosylation instances in PDB.

Filtering step	**Instances Removed**	**Remaining Instances**
Initial "*N*-glycan" oligosaccharide instances	-	74,178
Privateer/PDB2Glycan mismatch	851	73,327
No GlyConnect identifier	4,499	68,828
Elimination of *N*-glycan substrates	1,693	67,135

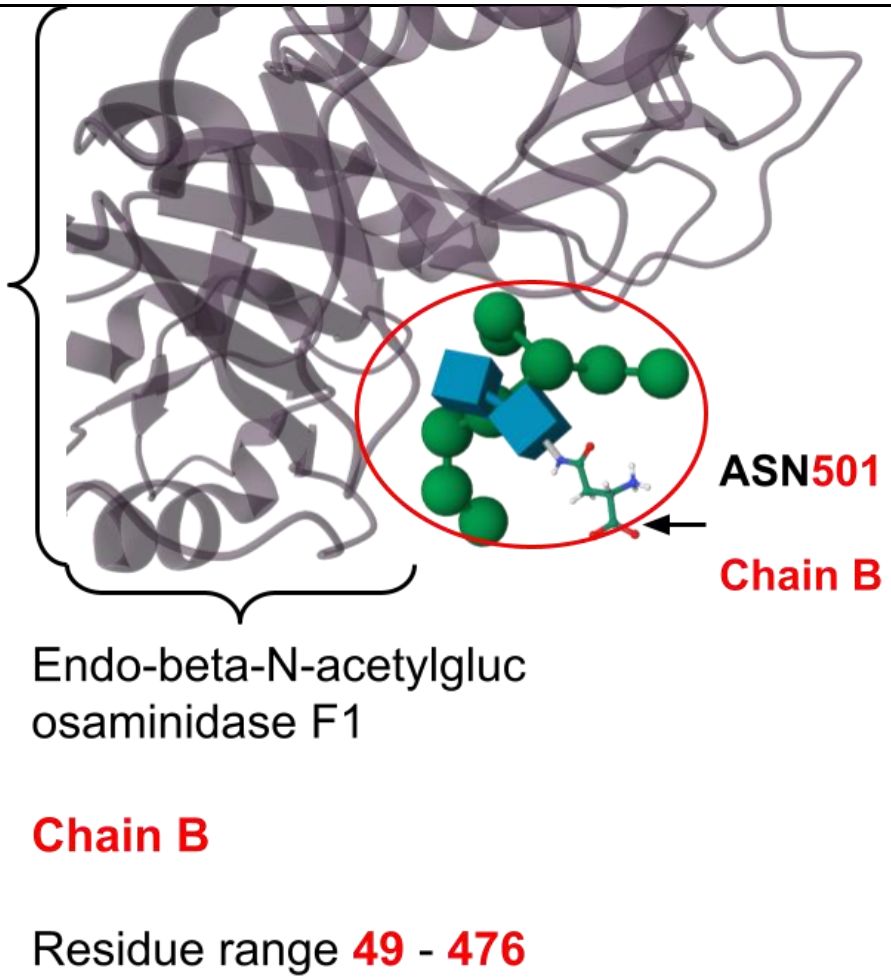


Figure 3.3: An example of PDB deposition modelling *N*-glycan binding, rather than *N*-glycosylation in a crystal structure of Bacteroides thetaiotamicron EndoBT-3987 in complex with $Man_9GlcNAc_2Asn$ substrate (PDB ID: 6TCV[19]). The example demonstrates the need for an elaborate algorithm as metadata in terms of chain identifiers is not sufficient to automatically recognize instances of *N*-glycan binding, rather than *N*-glycosylation. In principle, the *N*-glycosylation filtering algorithm is searching for at least one amino acid residue assigned to Chain B in the depicted search area (red circle), in combination with an

assigned sequence number from the following list: 496, 497, 498, 499, 500, 502, 503, 504, 505, 506. If no amino acid is found to fulfil the criteria in the vicinity of modelled *N*-glycan, then the PDB structure is deemed to be modelling *N*-glycan recognition, rather than *N*-glycosylation.

The remaining *n*-glycans representing *N*-Glycosylation instances were queried through GlyConnect API using GlyConnect identifiers to retrieve *N*-glycan composition types, as summarised in Table 3.3. Due to lack of Hybrid *N*-glycans that were retrieved from PDB and addition of GlcNAc sugars at the terminal ends being the key event in *N*-glycan type assignment, instances of Complex and Hybrid *N*-glycans were grouped into a category labelled “Processed”, which was used in subsequent analyses throughout the chapter.

Table 3.3: Summary of *N*-glycan types retrieved from Protein Data Bank (PDB). Column legend: ‘*N*-glycan type’ is *N*-glycan composition type classified using the GlyConnect identifier. ‘Count’ is the total number of instances of particular *N*-glycan composition types retrieved from PDB. Bolded *N*-glycan types were used for subsequent analysis as they extend beyond the $Man_3GlcNAc_2$ core.

***N*-glycan Type**	**Count**
single-GlcNAc	42,706
No-core	21,844
Pauci-Mannose	1,186
High-Mannose	1,087
Complex	310
Hybrid	2

Upon examining the outcome of the filtering process, it is noteworthy that the amount of usable data is significantly limited. Putting numbers into perspective, only 2.08% of modelled structures containing *N*-glycosylations extend beyond the $Man_3GlcNAc_2$ core with an assigned GlyConnect identifier.

Finally, non-redundant individual representatives of glycoproteins were computed, based on grouping using UniProt’s “recommended name” label. Additionally, due care was taken to only filter for glycoprotein models expressed in non-fungal expression systems. The fungal expression systems were filtered out due to glycan processing being radically different from

other expression systems, specifically *N*-glycan maturation using oligomannose extension, rather than branching via the addition of terminal GlcNAc sugars.
The non-redundant glycoprotein clustering resulted in 104 clusters, composed of 86 High-Mannose *N*-glycan representatives and 18 Processed *N*-glycan representatives.

In an attempt to clarify the degree to which unique glycoprotein data, potentially useful for study, was rendered unusable due to incomplete or erroneous *N*-glycan modelling, a comparative study was carried out. This study compared the number of unique glycoprotein representatives derived from the filtering process utilised in this chapter to those that were specifically modelled for *N*-glycosylation. The comparison revealed that 639 representatives of unique glycoproteins, each containing at least one instance of *N*-glycosylation, were deemed unusable due to their removal by the filtering algorithm outlined in this chapter. Notably, these significant findings can mostly be attributed to the absence of any cluster representatives that incorporated an *N*-glycan modelled beyond the $Man_3GlcNAc_2$ core. This result strongly suggests that while there is a wealth of data available for identifying the positions of glycosylation sites in relation to specific amino acid residues in the sequence, the structural details of *n*-glycans associated with specific glycosylation sites in the Protein Data Bank (PDB) are currently severely lacking. This shortcoming highlights the need for more community driven efforts to comprehensively model glycans in the area of glycoprotein research.

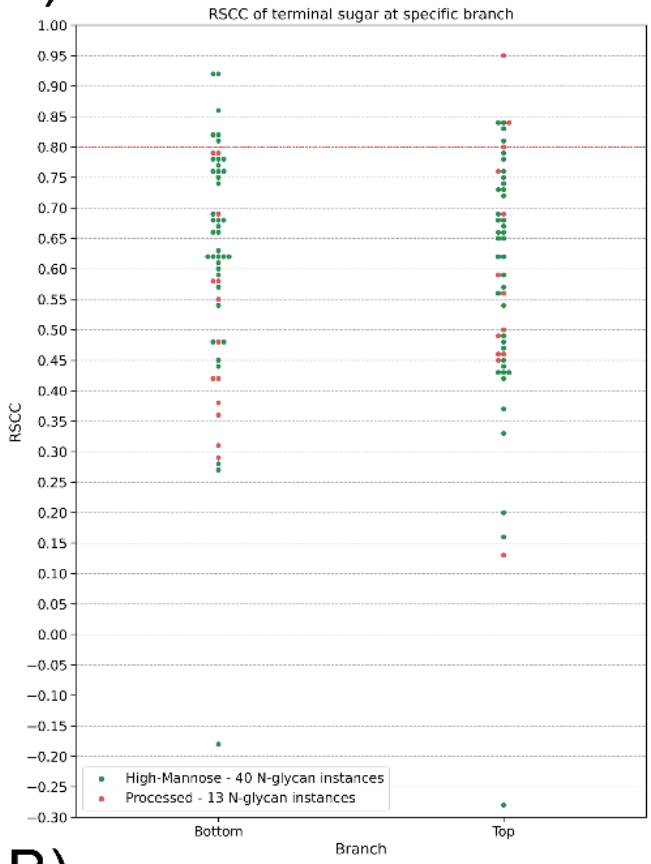
Finally, the fit of terminal sugars from extracted *N*-glycans within associated PDB files to experimental density in 104 clusters were evaluated using Real Space Correlation Coefficient (RSCC) metric. The equation to calculate Real Space Correlation Coefficient (RSCC) is described below:

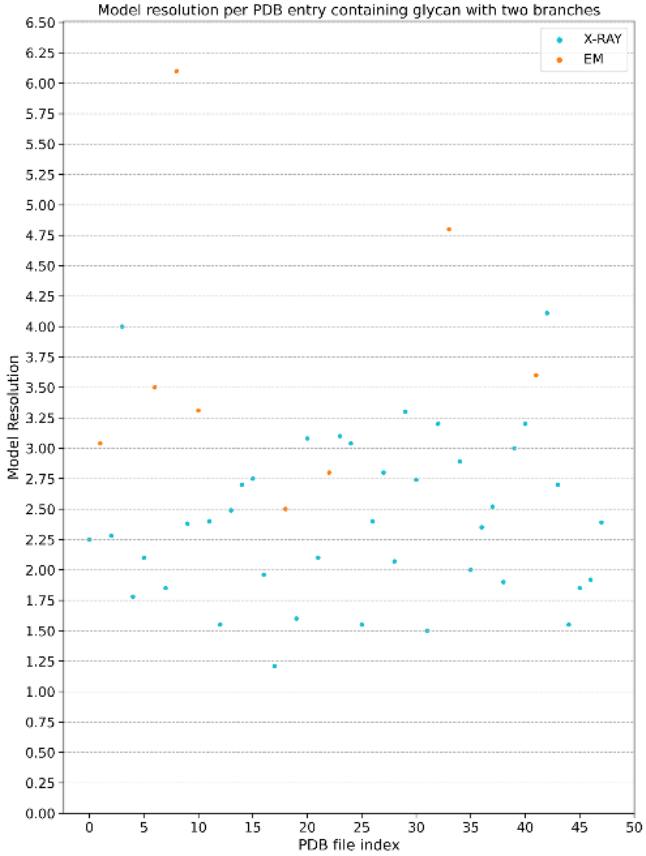
$$RSCC = \frac{\Sigma(\rho_{obs} - \langle\rho_{obs}\rangle)(\rho_{calc} - \langle\rho_{calc}\rangle)}{[\Sigma(\rho_{obs} - \langle\rho_{obs}\rangle)^2 \Sigma(\rho_{calc} - \langle\rho_{calc}\rangle)^2]^{1/2}}$$

The RSCC metric in quantitative terms describes how well a modelled residue fits its associated experimental density, by calculating the difference between observed (experimental) structural factors and calculated structural factors associated with the fitted model. For monosaccharides that are modelled as components of glycans, RSCC values of above 0.80 are considered to signify a good model fit to its associated experimental density.

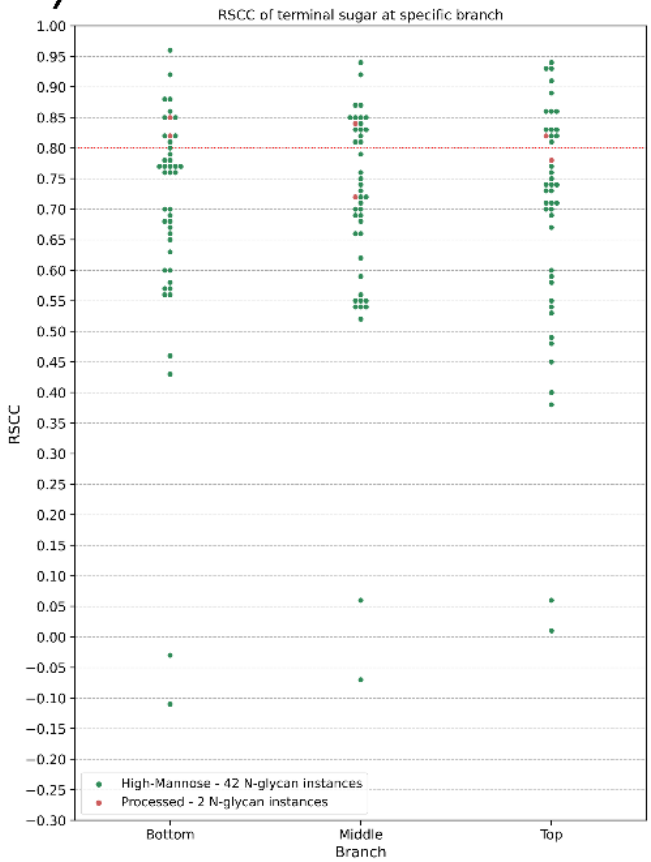
The computation of the RSCC metric was successfully carried out for 97 cluster representatives out of a total of 104 cluster representatives. Unfortunately, for 7 cluster representatives it was impossible to calculate RSCC scores for modelled sugars in glycans due to reasons such as: PDB entry deposition not containing files associated with experimental density or a glycoprotein model being solved using NMR, rather than X-ray crystallography or cryo-EM. The computed RSCC scores for terminal sugars in modelled *N*-glycans from 97 cluster representatives is shown in Figure 3.4.

A)


RSCC of terminal sugar at specific branch
RSCC
Branch
Bottom
Top
High-Mannose - 40 N-glycan instances
Processed - 13 N-glycan instances


Model resolution per PDB entry containing glycan with two branches
Model Resolution
PDB file index
X-RAY
EM

B)


RSCC of terminal sugar at specific branch
RSCC
Branch
Bottom
Middle
Top
High-Mannose - 42 N-glycan instances
Processed - 2 N-glycan instances

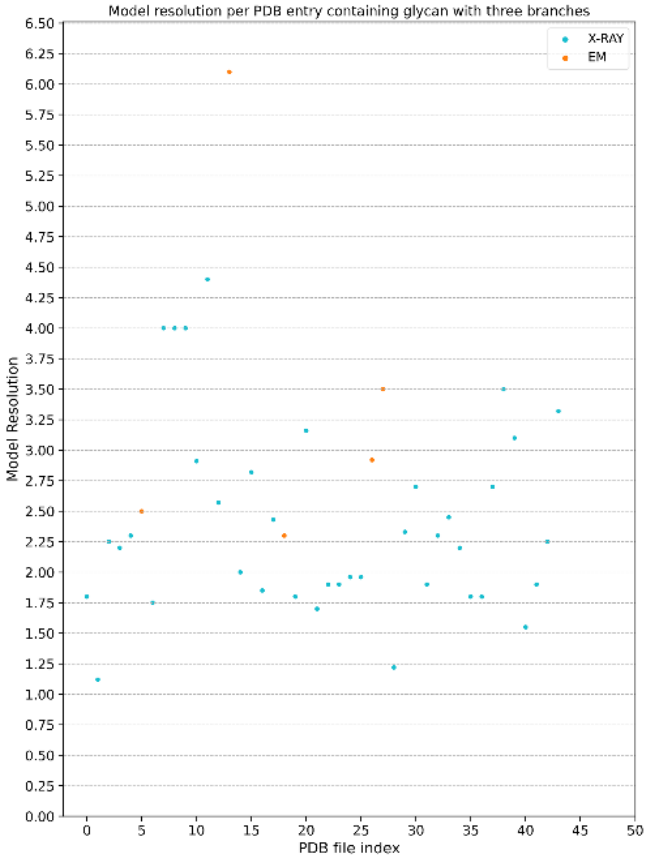

Model resolution per PDB entry containing glycan with three branches
Model Resolution
PDB file index
X-RAY
EM

Figure 3.4: Summary of computed RSCC scores for terminal sugars modelled within 97 cluster representatives of glycoproteins. A) (left) RSCC scores for terminal sugars of modelled *N*-glycans that are composed of two branches, grouped by *N*-glycan type. Red dashed line denotes the cutoff of 0.80, which is considered to demonstrate a good fit between modelled monosaccharide and its associated experimental density (right). Scatterplot of resolution values of PDB depositions that contain the modelled two branch *N*-glycans, grouped by experimental method, where X-Ray is X-Ray crystallography and EM is cryo-EM. B) (left) RSCC scores for terminal sugars of modelled *N*-glycans that are composed of three branches, grouped by *N*-glycan type. Red dashed line denotes the threshold value of 0.80, which is considered to demonstrate a good fit between modelled monosaccharide and its associated experimental density. (right) Scatterplot of resolution values of PDB depositions that contain the modelled three branch *N*-glycans, grouped by experimental method, where X-Ray is X-Ray crystallography and EM is cryo-EM.

The RSCC computation results are testament that modelling *N*-glycans beyond the $Man_3GlcNAc_2$ core is a significant challenge in structural biology. The RSCC values show that terminal sugars of modelled *N*-glycans in general tend to have low RSCC values, indicating a poor fit between modelled sugar and its associated experimental density, most often due to simply there being a lack of experimental density to begin with. High RSCC values of terminal sugars is an exception, rather than a regular occurrence. There also appears to be a tendency for *N*-glycans containing three branches to have more instances of terminal sugars passing the 0.80 RSCC value threshold than two branch *N*-glycans, especially taking into the account sample size being similar. This tendency can likely be attributed to a higher number of models containing three branch *N*-glycans being resolved at higher resolutions, therefore associated experimental density associated with terminal sugars of three branch *N*-glycans having well defined experimental density. The majority of associated PDB files tend to fall in the range between 1.75 Å and 3.00 Å in terms of model resolution. There are a considerable number of structures having significantly worse resolution than 3.00 Å and ideally the filtering algorithm should have excluded such cases. However, the analysis prioritised having as many representative clusters as possible and due to the already severe lack of unique representatives, a decision was made to forgo exclusion of representatives based on experimental density metrics.

Finally, the resulting clusters were manually inspected for non-redundancy. During the inspection it became apparent that, even though clustering by UniProt common name descriptor resulted in four different IgG structures, i.e., Immunoglobulin heavy constant gamma 1 (IgG1), Immunoglobulin heavy constant gamma 2 (IgG2), Immunoglobulin heavy

constant gamma 3 (IgG3) and Immunoglobulin heavy constant gamma 4 (IgG4), upon structural superposition using GESAMT algorithm it was revealed that the relevant chains containing *N*-glycosylation sites were highly redundant. The GESAMT algorithm successfully superposed the four structures with an RMSD score of less than 2 Å, indicating near perfect similarity[186]. The superposition is visualised in Figure 3.5.

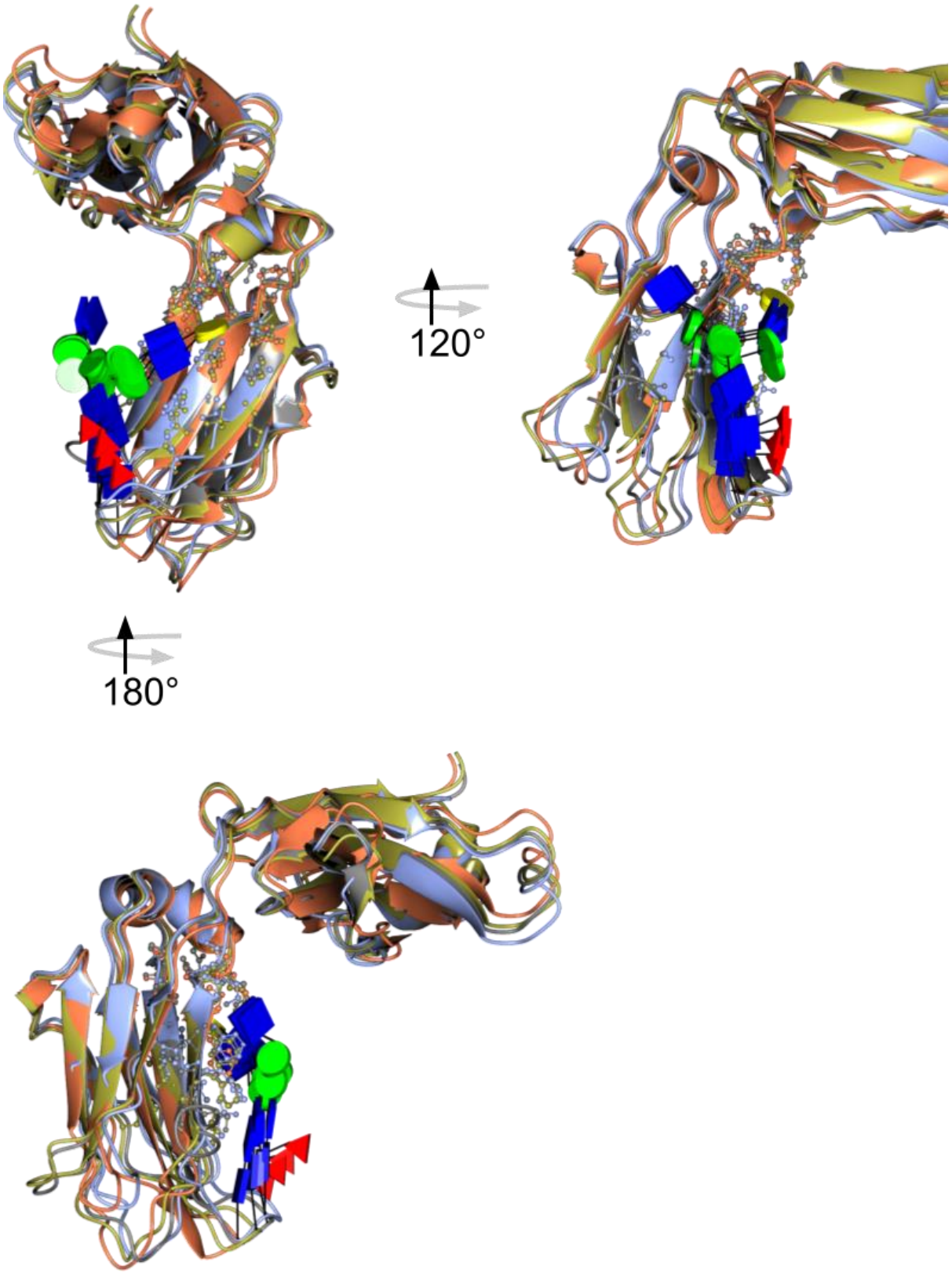


Figure 3.5: Superposition of IgG1 (PDB ID: 6YT7[20]) – coloured in orange, IgG2 (PDB ID: 4L4J[21]) – coloured in yellow, IgG3 (6D58[22]) – coloured in light blue, IgG4 (5W5N[23]) – coloured in grey. The amino acid residues displayed in “stick and ball” representation are

neighbours up to 7 Å distance away from terminal sugars of the modelled biantennary complex *N*-glycans.

As a result, entries of IgG2, IgG3 and IgG4 were manually removed from the clustering output, leaving only IgG1 as a global representative of Immunoglobulin heavy constant gamma glycoproteins. The manual adjustment resulted in 101 clusters, composed of 86 High-Mannose *N*-glycan representatives and 15 Processed *N*-glycan representatives and were consistently used throughout neighbourhood analysis.

3.4.2 *N*-glycosylation terminal neighbourhood analysis

In order to investigate protein structure influence on *N*-glycan processing machinery, the terminal neighbourhoods of modelled *N*-glycans were analysed. The analysis is predicated upon an understanding of the conversion of high-mannose *N*-glycans into more processed *N*-glycans occurring beyond the $Man_3GlcNAc_2$ core, therefore only the terminal area of the modelled *N*-glycan being relevant in the analysis.

Initial analysis considered individual associations of all 20 standard amino acids with the type of *N*-glycan product modelled in the glycoprotein across 101 *N*-glycosylation site representatives. The output of the analysis is depicted in Figure 3.6.

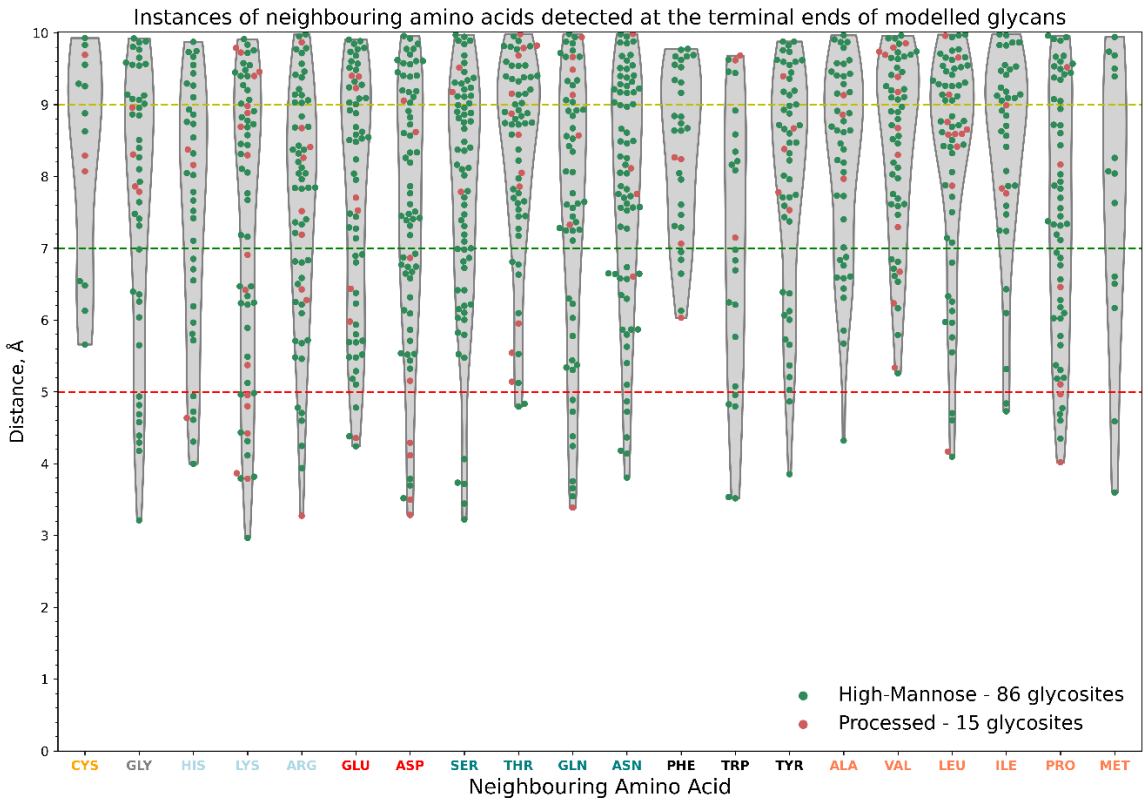
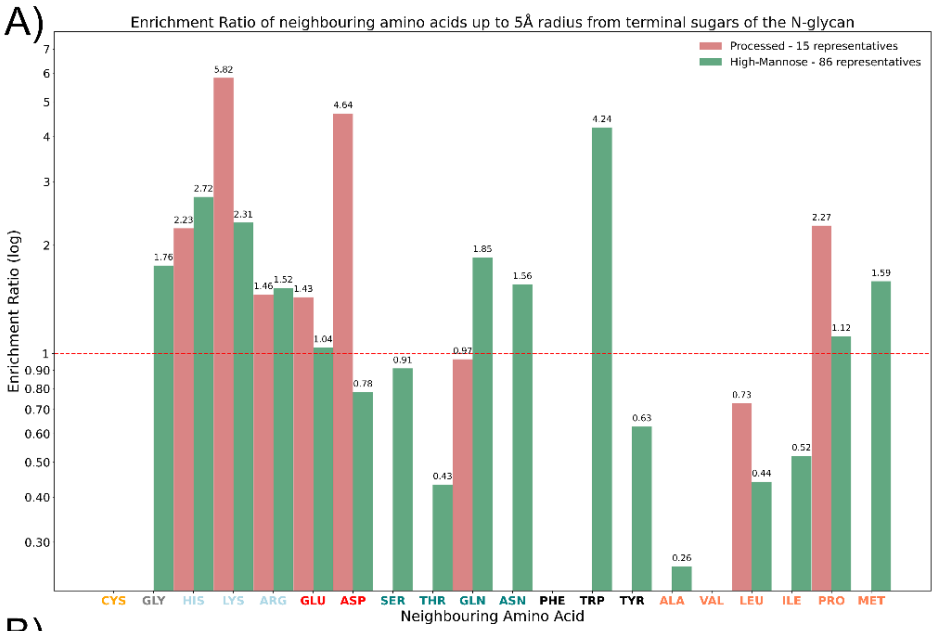


Figure 3.6: Individual neighbouring amino acid detections at first detected radius distance in the vicinity of terminal sugars of modelled *N*-glycans. Green circles denote an individual neighbour amino acid at the terminal end associated with high-mannose *N*-glycan product across 86 representatives, red circles denote a processed *N*-glycan product across 18 representatives. Red, green and yellow dashed lines represent radius distance cutoff thresholds used in subsequent enrichment analyses. The violin plot background denotes the distribution density of amino acid distances. Individual amino acids are represented in their three letter codes and are colour coded according to the assigned grouping in terms of redundant chemical features used in subsequent analyses (orange – sulphuric, grey – featureless, light blue – positive/basic, red – negative/acidic, teal – polar, black – aromatic, coral – hydrophobic).

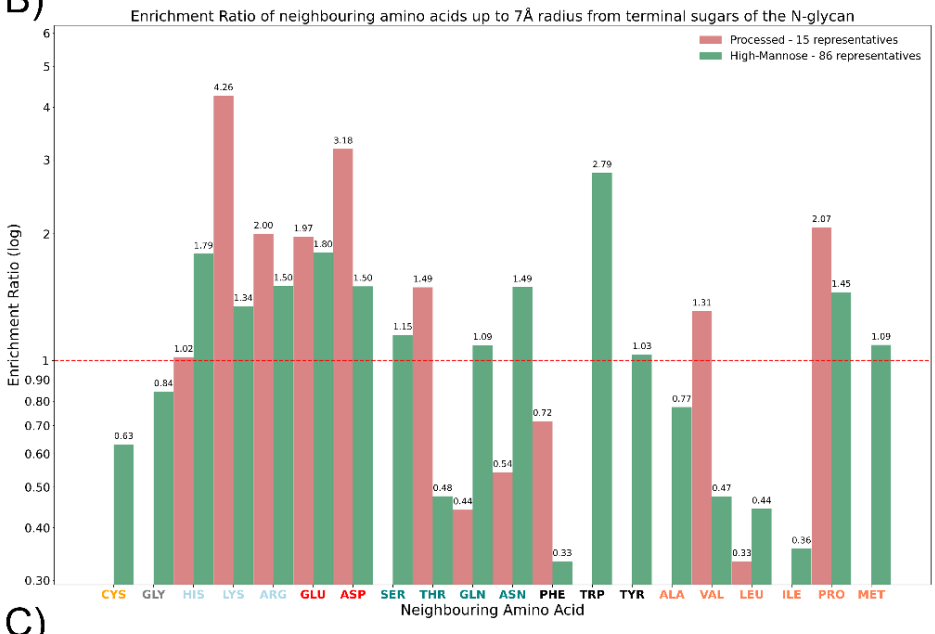
There are several notable associations that appear in the analysis. The closest neighbouring amino acids at the terminal ends of modelled *N*-glycans usually occur at around ~3-4 Å distance radius. Therefore, to ensure a good amount of hits for enrichment analysis, a 5 Å distance radius threshold was selected as an initial cutoff, representing detections of the closest neighbours. It appears that some amino acids are preferred as direct neighbours of terminal sugars and there is some degree of discrimination associated with certain *N*-glycan types. Particularly, it appears that some of the modelled high-mannose *N*-glycans display a clear preference for Glycine, Serine, Glutamine, Asparagine, Tryptophan, Tyrosine, Alanine, Leucine, Isoleucine and Methionine amino acid neighbours, as these amino acids are not only detected as being located in the immediate vicinity of the termini, but also at further distances too. On the other hand, it appears processed *N*-glycans display a preference for lysine and aspartate amino acid neighbours, at least in the immediate vicinity of the termini area. It is notable that some amino acid neighbours are excluded from the immediate vicinity of terminal sugars from both types of *N*-glycans, particularly Cysteine and Phenylalanine. It is possible that those amino acids could be located in the vicinity of other sugars, rather than terminal sugars specifically. Finally, it is also notable, that within the assigned groups of amino acids by redundant chemical features, there appears to be a preference for specific identities of amino acid neighbours at terminal ends, i.e., Serine neighbours being preferred by terminal sugars of high-mannose *N*-glycans versus Threonines being preferred by terminal sugars of processed *N*-glycans. Similarly, aspartate being preferred by terminal sugars of processed *N*-glycans versus terminal sugars of high-mannose *N*-glycans preferring glutamate, as well as Lysines being preferred by terminal sugars of processed *N*-glycans far more than Histidine and Arginine neighbours.

To ensure that discovered associations are due to specific amino acids having propensity to be located near the *N*-glycan termini, rather than analysed glycoproteins simply being enriched in specific amino acids, an enrichment analysis of terminal neighbourhood context was performed. The analysis was performed over multiple radius distance cutoff thresholds, with three selected thresholds shown in Figure 3.7.

A)


Enrichment Ratio of neighbouring amino acids up to 5Å radius from terminal sugars of the N-glycan
Processed - 15 representatives
High-Mannose - 86 representatives
Enrichment Ratio (log)
Neighbouring Amino Acid
CYS GLY HIS LYS ARG GLU ASP SER THR GLN ASN PHE TRP TYR ALA VAL LEU ILE PRO MET

B)


Enrichment Ratio of neighbouring amino acids up to 7Å radius from terminal sugars of the N-glycan
Processed - 15 representatives
High-Mannose - 86 representatives
Enrichment Ratio (log)
Neighbouring Amino Acid
CYS GLY HIS LYS ARG GLU ASP SER THR GLN ASN PHE TRP TYR ALA VAL LEU ILE PRO MET

C)

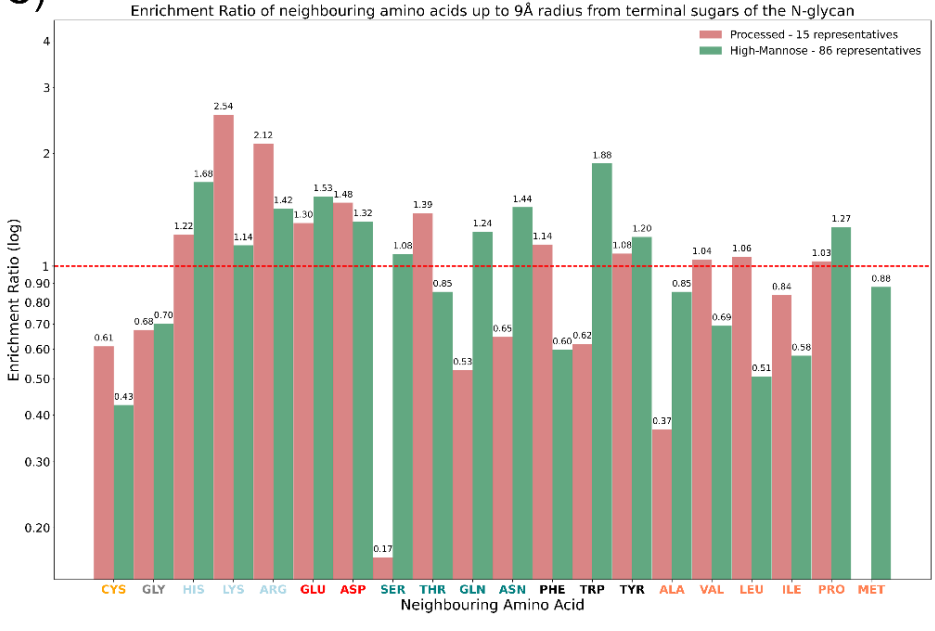

Enrichment Ratio of neighbouring amino acids up to 9Å radius from terminal sugars of the N-glycan
Processed - 15 representatives
High-Mannose - 86 representatives
Enrichment Ratio (log)
Neighbouring Amino Acid
CYS GLY HIS LYS ARG GLU ASP SER THR GLN ASN PHE TRP TYR ALA VAL LEU ILE PRO MET

Figure 3.7: Amino acid enrichment ratios over a variety of distance radius threshold cutoffs in vicinity of terminal sugars of modelled *N*-glycans. A) 5 Å radius threshold, B) 7 Å radius threshold, C) 9 Å radius threshold. The labels above bars represent the enrichment ratio in the dataset for individual amino acids.

The enrichment analysis enables the interpretation of data in two dimensions: 1) enrichment ratios of above 1 signify that specific amino acids have propensity to be concentrated near the *N*-glycan termini, while ratios of less than 1 signify that specific amino acids are depleted near modelled *N*-glycan termini and 2) difference in enrichment ratios between high-mannose *N*-glycans and processed *N*-glycans allows to measure the degree of preference between two types of *N*-glycans under consideration. Moreover, smaller radius distance thresholds enable focus on the immediate vicinity of *N*-glycan termini, versus higher radius distance threshold being more likely to capture the neighbourhood of the entire modelled *N*-glycan.

Within the context of the enrichment analysis, certain amino acids demonstrate distinct tendencies in proximity to *N*-glycan termini. Cysteines, for instance, are infrequently found adjacent to both processed and high-mannose *N*-glycan termini. Immediate glycine neighbours predominantly associate with high-mannose *N*-glycans, yet their enrichment ratio declines when larger search area is considered. Histidines maintain consistent enrichment across different search areas to both *N*-glycan types. Lysines exhibit a consistent pattern, being enriched around both *N*-glycan types across multiple thresholds, but with an apparent degree of preference for processed *N*-glycans. Arginines and Glutamates are enriched for both types of *N*-glycans in a similar pattern. Aspartates display a marked enrichment as immediate neighbours of processed *N*-glycans, however, with the expansion of threshold distance, the pronounced preference is decreased. Serines demonstrate a stark preference by high-mannose *N*-glycans through an insignificant enrichment ratio. On the other hand, Threonines demonstrate a stark preference relationship for processed *N*-glycans that is diminished as search radius is increased. Glutamines and Asparagines demonstrate a stark preference for high-mannose *N*-glycan termini that is consistent across multiple search radius distances. Phenylalanines demonstrate a clear preference for processed *N*-glycans with enrichment becoming prevalent as search radius distance is increased. Tryptophans demonstrate a clear preference for high-mannose *N*-glycans with consistent enrichment across all search radius distances. To an extent, Tyrosines also demonstrate preference for high-mannose *N*-glycans within immediate vicinities, although to a lesser degree of enrichment. Alanines, Leucine and Isoleucine amino acids do not demonstrate a particular preference for specific *N*-glycan types together with lack of enrichment for either type of *N*-

glycan type. Valine and Proline appear to demonstrate a preference for processed *N*-glycans, although enrichment and preference are diminished as search radius is increased. Finally, Methionines appear to demonstrate a consistent and obvious preference for high-mannose *N*-glycans. The most notable findings of the enrichment analysis are summarised in Table 3.4.

Table 3.4: A summary of the most significant amino acid preference relationship near *N*-glycan Termini. Label designation – **N+**: High-mannose *N*-glycan positive enrichment, **N-**: High-mannose *N*-glycan negative enrichment, **P+**: Processed *N*-glycan positive enrichment, **P-**: Processed *N*-glycan negative enrichment. The colouring on the amino acid labels corresponds to a direct mapping described in Figure 3.6.

Amino acid	5 Å radius threshold	7 Å radius threshold	9 Å radius threshold
Glycine	**N+, P-**	**N-, P-**	**N-, P-**
Lysine	**N+, P+**	**N+, P+**	**N+, P+**
Aspartate	**N-, P+**	**N+, P+**	**N+, P+**
Serine	**N-, P-**	**N+, P-**	**N+, P-**
Threonine	**N-, P-**	**N-, P+**	**N-, P+**
Glutamine	**N+, P-**	**N+, P-**	**N+, P-**
Asparagine	**N+, P-**	**N+, P-**	**N+, P-**
Phenylalanine	**N-, P-**	**N-, P-,**	**N-, P+**
Tryptophan	**N+, P-**	**N+, P-**	**N+, P-**
Tyrosine	**N-, P-**	**N+, P-**	**N+, P-**
Methionine	**N+, P-**	**N+, P-**	**N-, P-**

Following the analysis of individual amino acid neighbours of terminal sugars within modelled *N*-glycans, an attempt was made to group 20 amino acids by redundant chemical features to simplify interpretation. The output of the analysis of terminal amino acid neighbours grouped by redundant chemical features is shown in Figure 3.8.

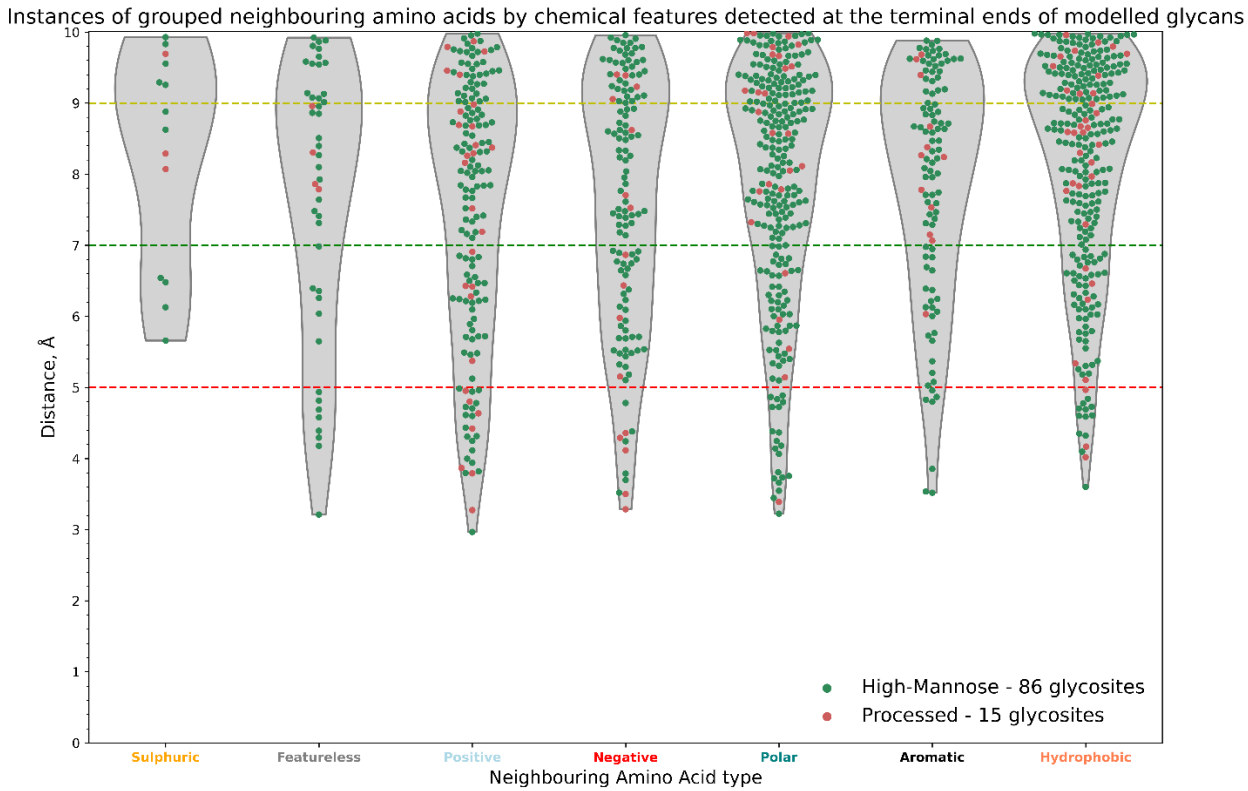


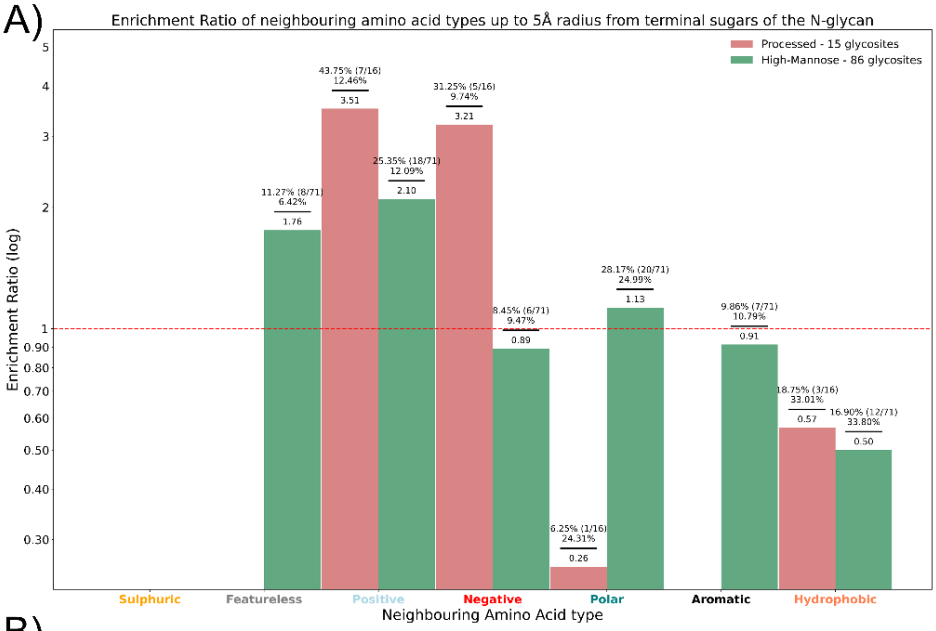
Figure 3.8: Detections of individual neighbouring amino acids grouped by redundant chemical features in the vicinity of terminal sugars of modelled *N*-glycans. Green circles denote an individual neighbour amino acid at the terminal end associated with high-mannose *N*-glycan product across 86 representatives, red circles denote a processed *N*-glycan product across 18 representatives. Red, green and yellow dashed lines represent radius distance cutoff thresholds used in subsequent enrichment analyses. The violin plot background denotes the distribution density of amino acid type distances. Individual amino acid are grouped into redundant clusters and are represented by colour coded labels which correspond to a direct mapping described in Figure 3.6 (orange – Cys; grey – Gly; light blue – His, Lys, Arg; red - Glu, Asp, teal – Ser, The, Gln, Asn; black – Phe, Trp, Tyr; coral – Ala, Val, Leu, Ile, Pro, Met).

From the output, it appears that terminal sugars from modelled processed *N*-glycans are most closely situated to positive (basic) and negative (acidic) amino acids, while terminal sugars from high-mannose *N*-glycans appears to demonstrate preference for polar, aromatic and hydrophobic amino acid neighbours.

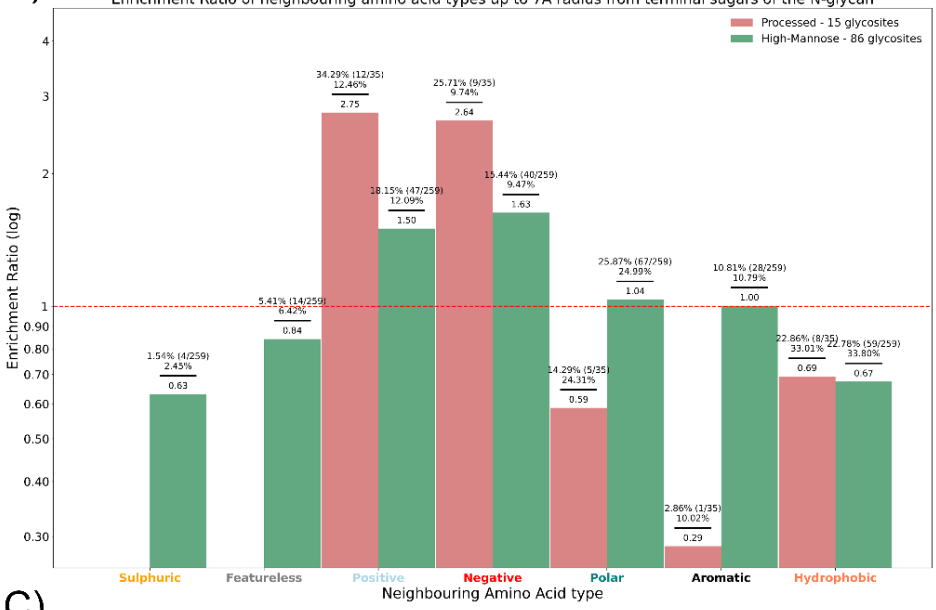
To ensure that discovered associations are due to specific amino acids groupings having propensity to be located near the *N*-glycan non-reducing ends (here addressed as termini), rather than analysed glycoproteins simply being enriched in specific amino acids groupings,

an enrichment analysis on terminal amino acid neighbour grouping was replicated. The analysis was performed over multiple radius distance cutoff thresholds, with select three thresholds shown in Figure 3.9.

A)


Enrichment Ratio of neighbouring amino acid types up to 5Å radius from terminal sugars of the N-glycan
Processed - 15 glycosites
High-Mannose - 86 glycosites
Enrichment Ratio (log)
11.27% (8/71) 6.42% 1.76
43.75% (7/16) 12.46% 3.51
25.35% (18/71) 12.09% 2.10
31.25% (5/16) 9.74% 3.21
8.45% (6/71) 9.47% 0.89
6.25% (1/16) 24.31% 0.26
28.17% (20/71) 24.99% 1.13
9.86% (7/71) 10.79% 0.91
18.75% (3/16) 33.01% 0.57
16.90% (12/71) 33.80% 0.50
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic
Neighbouring Amino Acid type

B)


Enrichment Ratio of neighbouring amino acid types up to 7Å radius from terminal sugars of the N-glycan
Processed - 15 glycosites
High-Mannose - 86 glycosites
Enrichment Ratio (log)
1.54% (4/259) 2.45% 0.63
5.41% (14/259) 6.42% 0.84
34.29% (12/35) 12.46% 2.75
18.15% (47/259) 12.09% 1.50
25.71% (9/35) 9.74% 2.64
15.44% (40/259) 9.47% 1.63
14.29% (5/35) 24.31% 0.59
25.87% (67/259) 24.99% 1.04
2.86% (1/35) 10.02% 0.29
10.81% (28/259) 10.79% 1.00
22.86% (8/35) 33.01% 0.69
22.78% (59/259) 33.80% 0.67
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic
Neighbouring Amino Acid type

C)

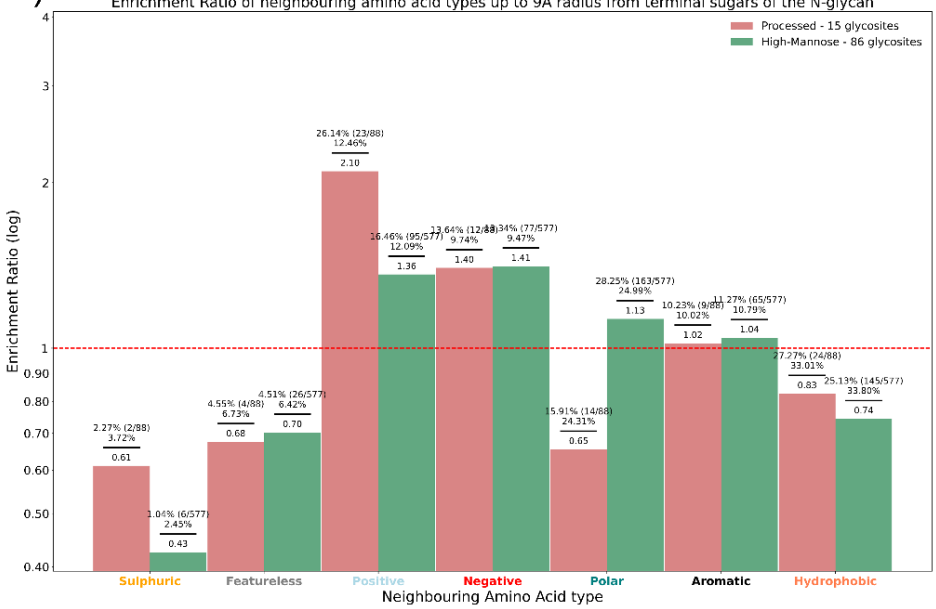

Enrichment Ratio of neighbouring amino acid types up to 9Å radius from terminal sugars of the N-glycan
Processed - 15 glycosites
High-Mannose - 86 glycosites
Enrichment Ratio (log)
2.27% (2/88) 3.72% 0.61
1.04% (6/577) 2.45% 0.43
4.55% (4/88) 6.73% 0.68
4.51% (26/577) 6.42% 0.70
26.14% (23/88) 12.46% 2.10
16.46% (95/577) 12.09% 1.36
13.64% (12/88) 9.74% 1.40
13.34% (77/577) 9.47% 1.41
15.91% (14/88) 24.31% 0.65
28.25% (163/577) 24.99% 1.13
10.23% (9/88) 10.02% 1.02
11.27% (65/577) 10.79% 1.04
27.27% (24/88) 33.01% 0.83
25.13% (145/577) 33.80% 0.74
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic
Neighbouring Amino Acid type

Figure 3.9: Amino acid type enrichment ratios over a variety of distance radius threshold cutoffs in vicinity of terminal sugars of modelled *N*-glycans. A) 5 Å radius threshold, B) 7 Å radius threshold, C) 9 Å radius threshold. The labels above bars represent the exact values that were used to calculate enrichment ratio for individual Amino Acid types in the dataset. Values in the bracket on the numerator represent the number of Amino Acid type detections out of total neighbouring amino acids detected.

In this adjusted approach, analysis is shifted from individual amino acids to groups categorised by their chemical characteristics. The shifted focus enables the analysis of localizations of chemically similar amino acid clusters around *N*-glycan structures, rather than the behaviour of each individual amino acid. This alternative viewpoint allows for the discernment of patterns and trends on a broader, chemically unified scale. No new insights can be obtained from “sulphuric” and “featureless” groups of neighbouring amino acid types, as they are composed of singular amino acids that had been covered in the analysis before. Nevertheless, other groupings demonstrate that only a few groups of amino acids are enriched in the vicinity of *N*-glycan termini over multiple thresholds. The enrichment analysis appears to suggest the following about selection of amino acid groups:

Positive amino acids, characterised by their basic nature, are consistently enriched in the vicinity of both high-mannose and processed *N*-glycan termini throughout various radius distance thresholds. Furthermore, it appears there is a slight preference relationship for positive amino acid neighbours in comparison to processed *N*-glycans throughout numerous search radius distances. Negative amino acids, characterised by their acidic character, prominently align with the terminal sugars of processed *N*-glycans. However, this marked association with processed *N*-glycans is diminished as the search radius expands. Polar uncharged amino acids present a tendency to be excluded by processed *N*-glycans, though this exclusionary trend becomes less pronounced with wider search radii. High-mannose *N*-glycans, meanwhile, maintain a consistent enrichment ratio that surpasses threshold across multiple search distance criteria. Aromatic amino acids distinctly disassociate from the terminal sugars of processed *N*-glycans when situated in close proximity. High-mannose *N*-glycans neither reject nor prefer aromatic amino acids with any notable emphasis, holding an enrichment ratio that hovers just above one throughout the different search radius thresholds. Hydrophobic amino acids, in contrast, maintain an equivalent stance, revealing no conspicuous preference or exclusion towards either high-mannose or processed *N*-glycans across the selected search radius threshold distance limits. The most notable findings of the enrichment analysis are summarised in Table 3.5.

Table 3.5: A summary of the most significant amino acid type preference relationship near *N*-glycan Termini. Label designation – **N+**: High-mannose *N*-glycan positive enrichment, **N-**: High-mannose *N*-glycan negative enrichment, **P+**: Processed *N*-glycan positive enrichment, **P-**: Processed *N*-glycan negative enrichment. The colouring on the amino acid labels corresponds to a direct mapping described in Figure 3.6.

Amino acid type	5 Å radius threshold	7 Å radius threshold	9 Å radius threshold
Positive	**N+, P+**	**N+, P+**	**N+, P+**
Negative	**N+, P-**	**N+, P+**	**N+, P+**
Polar uncharged	**N+, P-**	**N+, P-**	**N+, P-**
Aromatic	**N-, P-**	**N+, P-,**	**N+, P+**
Hydrophobic	**N-, P-**	**N-, P-**	**N-, P-**

In summary, these results demonstrate that there are indeed discriminatory features in terms of neighbouring amino acid identities and chemical characteristics of amino acids that might have an impact on *N*-glycosylation processing machinery.

3.5 Discussion & Conclusion

In this study, glycoproteins containing *N*-glycans modelled beyond the $Man_3GlcNAc_2$ core were obtained from PDB and analysed in a high-throughput manner to find specific amino acid associations at the terminal sugars that could potentially explain the differences in *N*-glycan products associated with specific glycosylation sites. The analysis revealed potentially meaningful associations of individual amino acids and their associated chemical characteristics that could explain differences in *N*-glycan processing between different *N*-glycosylation sites. Particularly, it appears that amino acids classified as positive (basic), negative (acidic), polar uncharged and aromatic, as well as glycine specifically, display discriminatory associations between the considered *N*-glycan types in the study.

A particularly meaningful finding is aromatic amino acid association with exclusion by processed *N*-glycans, as there exists *in vivo* evidence of aromatic amino acids influencing *N*-glycan processing determined by an increase in the homogeneity of *N*-glycan profiles for two glycoproteins. In the study published by Murray *et al.*, two glycoproteins (CD2ad and FGF9) were engineered to contain an enhanced aromatic sequon that was expressed in HEK293 cell lines. One of the findings in the study was that aromatic amino acids two residues before the consensus sequon of *N*-glycosylation decreased the degree of *N*-glycan processing into complex *N*-glycans, thereby increasing glycoform homogeneity[187]. This finding is in support of aromatic amino acids potentially being excluded by terminal sugars of processed *N*-

glycans presented in this chapter, in the sense that aromatic amino acids may have an association with the outcome of *N*-glycan processing. Even though Murray *et al.* found aromatic amino acid enhancement was engineered outside the terminal region and more towards the *N*-glycosylated asparagine of the *N*-glycan, it could potentially be hypothesised that multiple aromatic amino acids, both at the terminal end of glycan and in vicinity of glycosylated asparagine are potential determinants of *N*-glycan processing from high-mannose *N*-glycans to processed *N*-glycans. This hypothesis is especially supported by the fact that aromatic amino acids can form CH−π interactions between the aromatic rings of amino acids and carbohydrate rings, potentially creating a stereoelectronic barrier for *N*-glycan processing enzymes that convert high-mannoses into more processed *N*-glycans[188]. Therefore, these circumstances are favourable to expand the study conducted by Murray *et al.* as future work in attempts to engineer glycoproteins with enhanced aromatic amino acids in the vicinity of terminal sugars of high-mannose *N*-glycans to assess *in vivo* the relationship between aromatic amino acid enrichment and potential rational control of *N*-glycan processing.

The analysis methodology of this study is most directly comparable to the work carried out by Suga *et al*. The study had shown that *N*-glycosylation sites with higher accessibility to the solvent were more frequently associated with immature glycans (equivalent to “high-mannose” N-glycan type) than mature glycans (equivalent to “processed” N-glycan type) over multiple surface area cutoff thresholds. The authors also attempted to investigate the potential bias of amino acid residues surrounding nascent oligosaccharides. According to the analysis results, solvent exposed Asn residues were significantly populated in proteins with immature glycans across multiple surface area of the glycosylation site thresholds, while Ile was more frequently associated with immature glycans up to 500 $Å^2$ threshold and Tyr more frequently associated with immature glycans up to 3,000 $Å^2$ threshold. The work carried out in this chapter of thesis agrees in terms of Asn residues being associated with immature *N*-glycans, which is the equivalent of high-mannose *N*-glycans in this study. The work carried out in this chapter of thesis is also in agreement in terms of Tyr residues being in association with immature *N*-glycans. However, authors concluded that there are no strong correlations between *N*-glycan type and amino acid residue types in terms of their chemical features, such as hydrophilicity/hydrophobicity or positive/negative charges of the side chains[178]. The authors’ conclusion is a contradiction to the results obtained in this study. The contradictory results can potentially be explained by the difference in methodology between the two studies, as shown in Figure 3.10.

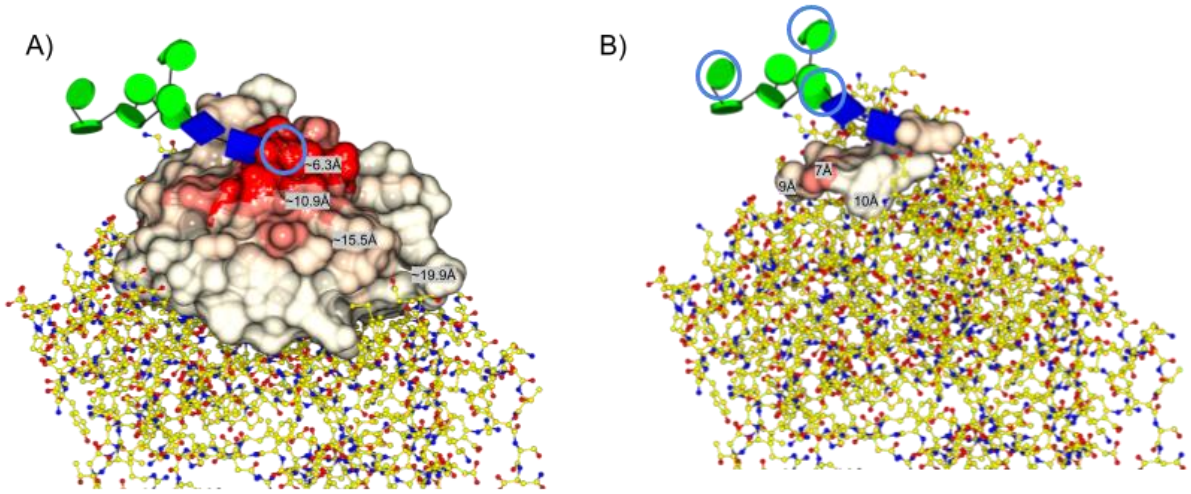


Figure 3.10: Comparison of origin points and their detected neighbouring amino acid outputs between two analyses using an identical view of a glycoprotein structure (PDB ID: 1H4P). A) Origin of the probe point (ASN165, chain A) circled in blue and its associated amino acid neighbour outputs visualised using protein surface representation with different colours and their labels denoting converted radius distance threshold criteria by Suga *et al*. The conversion of sphere surface area to radius distance was converted using: $r = \sqrt{\frac{A}{4\pi}}$, where r is radius and A is sphere's surface area. The neighbouring amino acid output is an approximation and not a direct conversion, as the authors employed a more elaborate method to compute surface area to detect neighbouring amino acids. The purpose of the approximation is to serve as a visual aid in the comparison to the study presented in this chapter. B) Origin of multiple probe points (BMA6, MAN8, BMA10, chain C) circled in blue and its associated amino acid neighbour outputs visualised using protein surface representation with different colours and their labels denoting radius distance threshold used in this study. The modelled *N*-glycan and the glycoprotein were automatically eliminated from consideration in this study, due to the following reasons: 1) Glycoprotein was expressed in a fungal expression system, 2) modelled *N*-glycan likely contains potential modelling mistakes, specifically terminal mannose sugars being modelled as β-Man anomers, thus failing to return a match on GlyConnect database.

The primary difference between the two studies is the location of the probe point used to compute the neighbouring amino acids of the modelled *N*-glycan. In the Suga *et al.* analysis, the computed surface area is significantly higher, and the probe point is biased away from terminal sugars, where *N*-glycan processing is likely to occur. As demonstrated by the enrichment analysis presented in this chapter, as the radius distance threshold is increased, the distribution of detected neighbouring amino acids trends towards distribution of amino acids of analysed protein sequences. Therefore, it is likely that authors were not able to

make strong conclusions in terms of correlation between *N*-glycan type and amino acid features because of the search area being too vast. Additionally, the study by Suga *et al.*, provided a detailed list of glycoprotein structures considered in the analysis. One observation in terms of obtained representative structures between two studies was that Suga *et al.* were much more successful in obtaining a balanced dataset, i.e., equivalent number of representatives between immature (equivalent of high-mannose) and mature (equivalent of processed) *N*-glycans. A significant number of representative *N*-glycans were filtered out from the study in this chapter, which were used by Suga *et al.*, in principle due to *N*-glycan compositions containing potential modelling mistakes and being expressed in fungal expression systems, with a particular example highlighted in Figure 3.10 B).

Indeed, the most significant drawback of this study is the small number of available non-redundant glycoproteins containing processed *N*-glycan products in the PDB. One of the biggest contributing factors to the issue was the elimination of processed *N*-glycan instances due to failure to return an assigned GlyConnect identifier. Therefore, this highlights the need for investment of resources to engineer a platform that would automatically modify the atomic coordinates associated with flawed models of *N*-glycans to enable the repetition of this study in ensuring that presented findings can be presented with a higher degree of confidence in terms of their validity – a task that seems to fall within the remit of PDB-REDO[12,189].

Another point to consider is that models deposited to PDB are mere snapshots of the dynamical nature of glycoproteins. The glycan parts of glycoproteins are often flexible and mobile, unlike the more rigid structure of the peptide backbone. This is partly due to the fact that glycans are typically attached to the protein via a single point, allowing the rest of the glycan structure to move freely. Indeed, some of the glycoprotein models included in the dataset did not have any neighbours within 5 Å radius at the terminal sugar positions. A snapshot of a dynamical nature is unable to reveal whether the oligosaccharide chain ever moves to the vicinity of neighbouring amino acids. Therefore, as a potential improvement to this study, molecular dynamic simulations could be utilised to reveal a clearer picture. In addition, the heterogeneity of glycans oftentimes results in incomplete glycan descriptions due to signal becoming dominated by noise during model building from density maps. As a result of these factors, the currently revealed patterns are likely to become altered once a larger set of representative glycoprotein structures becomes available.

Investigation of *N*-glycan processing using predictional data

4.1 Introduction

In the prior investigations into *N*-glycan processing using atomic structural data, discernible features in the neighbouring amino acids in proximity to terminal sugars were demonstrated. Unfortunately, the analysis has also revealed that there was a significant deficiency in availability of unique representatives of glycoprotein models that would contain modelled *N*-glycans beyond the $Man_3GlcNAc_2$ core. The analysis demonstrated that scarcity in modelled *N*-glycans can most likely be attributed to the inherent difficulties encountered in obtaining sufficient experimental density associated with oligosaccharide regions. The very few instances of deposited glycoprotein structures containing modelled *N*-glycans, based on RSCC analysis, demonstrate that sugars at the terminal ends are primarily modelled by relying on prior knowledge of *N*-glycan biosynthesis pathways if experimental density is insufficient. It is unlikely that there will emerge a generalizable technique in structural biology that would enable to overcome information loss caused by sample heterogeneity. Therefore, community driven efforts might consider developing and improving techniques that enable complete modelling of glycoproteins without the requirement of associated experimental density maps.

As efforts to overcome these challenges were investigated, a collaborative project was concurrently initiated. The collaboration with Dr Elisa Fadda group considered the feasibility of potentially implementing a grafting algorithm to complete the *N*-glycans on SARS-CoV-2 spike glycoproteins using template *N*-glycan structures. This endeavour was seen as pivotal given the global urgency surrounding the pandemic[190]. However, with the introduction of the AlphaFold database in the summer of 2021, the collaborative nature of work had shifted due to the recognition of the transformative potential of structures predicted by AlphaFold 2[191].

During the PhD a prototype software solution was developed that combines protein structures predicted by AlphaFold 2 with the post-predictional addition of *N*-glycans using a grafting algorithm. The development of a prototype grafting algorithm enabled the investigation of neighbouring amino acid contexts in the vicinity of grafted *N*-glycans for glycoprotein structures that were unavailable in the Chapter 3 of the thesis.

4.1.1 Glycan grafting to protein structures

Systematic studies have been carried out to show that *N*-glycan modelling can be carried out without available experimental data using foreign *N*-glycan structures from different glycoproteins, where experimental data is available as a template[192]. Particularly, in the study by Jo *et al.*, it has been demonstrated through a statistical analysis that *N*-glycan structures on homologous glycoproteins are significantly conserved compared to the random background in terms of oligosaccharide conformation[192]. This indicates that *N*-glycans with similar parent glycoprotein structure in terms of sequence similarity, can be confidently transplanted as is. On the other hand, identical *N*-glycan structures on non-homologous parents do not display similarity in terms of oligosaccharide conformation. Nevertheless, part of *N*-glycan oligosaccharide closer to the protein backbone, specifically $Man_1GlcNAc_2$ fragment has been demonstrated to be conformationally conserved, indicating that partial *N*-glycan fragments can indeed be used as templates for *N*-glycan modelling when no experimental density is available[192].

Indeed, as the analysis had shown in Chapter 3 of this thesis, most of the *N*-glycans retrieved from glycoprotein structures deposited to wwPDB are indeed modelled as partial fragments with the composition of $Man_1GlcNAc_2$ or $GlcNAc_2$ and are labelled as “No-core” in Table 3.3. The lack of complete *N*-glycan compositions has not prevented computational glycobiologists from studying glycan-protein interactions using molecular dynamics (MD) methods. The community has devised several computational approaches to modify and re-model incomplete *N*-glycan structures without associated experimental density. Of note are tools, such as Glycosylator, GLYCAM-WEB that are routinely used by computational glycobiologists to modify *N*-glycan structures on input glycoproteins to prepare the models for molecular dynamic studies[193,194]. In principle, these tools work by replacing incomplete *N*-glycans in input structures with more complete *N*-glycans. The replacement procedure of *N*-glycan structures tends to result in steric clashes or atomic overlaps between the grafted *N*-glycan and protein backbone. In order to overcome resulting clashes, the aforementioned tools employ sampling of alternate sugar-sugar torsional angles of the *N*-glycan to find a configuration that completely eliminates clashes. However, optimizations to eliminate clashes and overlaps between *N*-glycan and protein backbone are likely to result in an overall *N*-glycan conformation that is inconsistent with glycan-protein interactions[195]. For molecular dynamic based workflows, this is not a major concern, as the processed glycoprotein complex would then undergo actual molecular dynamic simulations using appropriate force fields that would eventually capture feasible overall *N*-glycan conformations during sampling procedures. However, executing molecular dynamic

simulations for glycoprotein complexes is computationally expensive both in computational resources and time[196]. Molecular dynamic simulation of a typical glycoprotein complex is likely to require days if not weeks of computational time, therefore scalability in terms of replicating the procedure for the number of glycoproteins needed for amino acid neighbour analysis is a major concern.

In order to counteract scalability concerns, there are currently ongoing efforts to produce a library of MD-equilibrated *N*-glycan structures fine-tuned for glycoprotein structures by Dr Elisa Fadda's group. In the MD equilibration study of high-mannose *N*-glycans carried out by Fogarty and Fadda it has been demonstrated that protein backbone constrains the high-mannose conformational ensemble with well-defined conformational hotspots for specific linkages in order to satisfy steric and hydrogen bonding requirements demanded by the protein's surface[196]. Therefore, such an equilibrated library of *N*-glycans in near future could theoretically be used as a set of templates by structural biologists in modelling novel glycoprotein structures.

4.1.2 AlphaFold 1 & 2

Proteins are composed of linear chains of amino acids, which undergo a process called "protein folding", where they spontaneously establish defined three-dimensional configurations, largely driven by the thermodynamics of interatomic forces. For the past century, experimental techniques such as X-ray crystallography, cryo-electron microscopy and nuclear magnetic resonance were used to elucidate three-dimensional configurations of about 170 thousand proteins, which is a small fraction of over 200 million known protein structures across multiple life forms[197]. The relatively small number of elucidated three-dimensional structures of known proteins can be attributed to labour, capital and time costs associated with aforementioned experimental techniques. Therefore, with the increases in computing power available to the general population, highly correlating with Moore's law, community driven efforts were invested into addressing the gap between resolved three dimensional structures of proteins and known protein structures by creating predictive tools of three-dimensional protein structures given an input of amino acid sequence[197,198]. Community driven efforts dedicated to solving the "protein folding problem" since 1994 would regularly be benchmarked biennially at the Critical Assessment of Protein Structure Prediction (CASP) competition.

The foundations of AlphaFold can be traced back to work developed by various teams in the 2010s, which coincided with the next generation sequencing revolution in the field of

genomics. The innovations in sequencing techniques that enabled massively parallel sequencing has significantly improved the cost and scale of living organism genome characterizations, leading to an abundance of large databases of related DNA sequences available from many different organisms[199]. The community aiming to solve the “protein folding problem” embraced the developments in genome sequencing, leading to development of techniques, where multiple sequence alignments (MSA) of homologously related proteins were analysed to find changes at different residues that appeared to be correlated, despite residues not being consecutive in the protein chain in combination with the available 3D structures of resolved proteins[200]. Due to protein function being driven by its three-dimensional conformation, rather than directly by its primary sequence, such correlations suggested that residues may be close to each other physically, despite not having complete alignment in terms of sequence. These insights enabled the development of techniques targeted at predicting contact maps, describing the distance between all possible amino acid residue pairs of a three-dimensional structure in a binary two-dimensional matrix[201], validated against a contact map output by the already existing 3D descriptions of protein structures, deposited in the wwPDB.

DeepMind, the company behind AlphaFold, extended the approach of predicting contact maps further, by developing a predictive model that can estimate a probability distribution, rather than binary output, for how close the inter-residue distances might likely be, essentially predicting a distance map between amino acid residues given a protein sequence input. This development was benchmarked in 2018 during the 13th edition of CASP, for what is now commonly referred to as AlphaFold 1. The initial version of AlphaFold employed multiple neural networks, each trained separately to process different types of available data. The combined outputs of these modules were then assessed using physics-based energy simulations for most likely protein folding predictions[202]. Thus, alongside the prediction of distance maps, AlphaFold 1 was engineered to predict Φ (phi) and Ψ (psi) angles for each amino acid residue, enabling for the prediction of 3D structure of a protein given its sequence. While a breakthrough in its own right, AlphaFold 1 had a tendency to overemphasise potential interactions between amino acid residues that were nearby in the sequence in comparison to interactions between residues at longer distances in the primary sequence, resulting in predicted 3D structures by AlphaFold 1 that demonstrated bias towards predicting models with an overrepresentation of secondary structure features such as alpha helices and beta sheets, otherwise known as overfitting[203].

To decrease the degree of overfitting via overrepresentation of secondary structures features in final predictions, AlphaFold 2 was developed to supersede the initial version of

AlphaFold, benchmarked in the 14th edition of CASP in 2020[81]. In principle, the most significant difference between the two versions of AlphaFold is the replacement of multiple separately trained modules in the initial version with a system of integrated sub-networks coupled together into a single differentiable end-to-end model, which is trained as a unified architecture. In addition, AlphaFold 2 employed an “attention” mechanism through its two key modules, which are based on a transformer design[81]. The transformer attention model through the attention mechanism enables AlphaFold to weight the relevance of input elements, rather than considering all elements equally. It calculates an 'attention score' for each element in a sequence, assigning higher scores to more relevant elements[204]. The scores are used to weight the contribution of each element when producing an output, thus allowing the model to focus more on relevant information, by progressively refining the residue/residue and residue/sequence information[205]. In principle, this implementation addressed the tendency of the previous version of AlphaFold to overemphasise potential interactions between amino acid residues that were nearby in the sequence and instead tend towards capturing distant interactions as well. The subsequent version of AlphaFold was significantly more accurate, albeit the bias towards over-predictions of secondary structure elements such as α-helices and β-strands was not completely eliminated[203]. Shortly after the public release of AlphaFold 2, AlphaFold Protein Structure Database (AlphaFoldDB) was created to store 100 million protein sequence predictions across multiple model-organism proteomes, which are all indexed using unique UniProt identifiers[191].

Contrary to popular belief in the media, AlphaFold did not “solve” the “protein folding problem”. While AlphaFold is quite capable at predicting snapshots of most stable conformations for each protein, the model is not able to predict changes in protein structure under different pH conditions or temperature[206]. Therefore, it is reasonable to treat AlphaFold predictions as an accelerator towards resolving experimental structures obtained from structural biology experiments, particularly as a substitute for homology based molecular replacement techniques to resolve the crystallographic phase problem[207].

4.1.3 AlphaFill

Protein structures resolved using structural biology experiments extend beyond amino acid sequence and its 3D conformation. The 3D conformation of the protein is also affected by factors, such as ligands or co-factors[208]. The predicted protein models by AlphaFold, lack atomic coordinates for such molecules, which are directly implicated in molecular structure or function of the prediction target, for example: haemoglobin lacking bound haem. Curiously, multiple independent observations successfully demonstrate that, even though

AlphaFold does not directly consider ligands, the model is capable of predicting protein structures in 3D conformations consistent with presence of potential ligands indirectly taken into account. In other words, AlphaFold has successfully learned aspects of protein folding with cofactor binding taken into the account, based on the thermodynamic snapshot profile of protein structures deposited in wwPDB. Notably, there are active efforts to expand the independent observations into practical and applicable tools to enrich AlphaFold predictions in a post-predictional manner and that has been successfully demonstrated through the development of AlphaFill. AlphaFill has successfully enriched 99,541 AlphaFold protein models with a pool of 2694 unique compounds through 12,029,789 transplantation operation, as of 24th November 2022[114].

AlphaFill implementation for a protein structure predicted by AlphaFold searches for sequence homologs in PDB-REDO databank, which has an alignment of at least 85 residues as hits with an identity higher than 25%. Next, an algorithm determines whether any of the returned PDB-REDO structures from sequence homologs contain compounds of interest. If compounds of interest are detected in sequence homologs in PDB-REDO models, then selection of structures containing the compounds of interest are structurally aligned with AlphaFold model based on Cα-atom positions. The structural alignments are evaluated using root-mean-square deviation (r.m.s.d) at the global level to sort for the most similar protein structures in terms of overall 3D conformation. Starting from the most similar homolog, all protein backbone atoms within 6 Å from every atom of compound of interest are selected as input for local structural alignment of a current PDB-REDO model to the current AlphaFold model. Local structural alignment is evaluated using root-mean-square deviation (r.m.s.d) at the local level. The local structural alignment enables to trivially transplant compound of interest from PDB-REDO model to AlphaFold model, as the two different models are effectively overlapping, and their coordinate systems have been relatively transformed to each other.

The authors of the software concluded that the transplantation of AlphaFold models with a missing pool of select compounds is successful to an extent. The successful aspect of the transplantation procedure is that the approach enables the production of sensible depictions of ligand binding sites at a qualitative level. However, the enriched AlphaFold structures through transplantation should not be used as sources for quantitative measurements as, for example, in depictions of zinc binding sites, the atomic distances between the zinc ion and surrounding amino acids deviate from previously established target values[209]. AlphaFill is not capable of handling polymer ligands, such as peptides, nucleic acids or oligosaccharides[114]. Nevertheless, this development is strong evidence that protein structure predictions

produced by AlphaFold are accurate to an extent, where there is little variation between key individual amino acid residue positions when structurally aligned at a local level with experimentally-resolved structures.

4.1.4 GlyConnect - source of glycoproteomic data

The advent of AlphaFold has opened access to investigating *N*-glycosylation processing machinery on glycoproteins that have either not yet been resolved and deposited to wwPDB, or deposited glycoproteins that did not have complete *N*-glycan compositions. However, a source of information is required to determine which glycosylation sites for specific glycoproteins harbour high-mannose *N*-glycans or are processed into more complex *N*-glycans via the addition of terminal GlcNAc sugars. Thankfully, the necessary details of information are available on GlyConnect[15]. Besides GlyConnect containing a repository of *N*-glycosylation compositions, it also contains additional metadata about protein backbone associations, particularly UniProt identifiers and sequence numbers of asparagine residues forming *N*-glycosidic linkage. Because both AlphaFoldDB and GlyConnect are cross-referenced with the UniProt database, this enables trivial cross-referencing of *N*-glycan compositions from GlyConnect with predicted structures from AlphaFoldDB.

4.2 Aims

One of the aims presented in this chapter was to analyse the feasibility of using models predicted by AlphaFold to graft *N*-glycans in a post-predictional manner. After establishing the feasibility, together with independent observations from other groups, the aim of the chapter shifts towards analysing the structural contexts in the vicinity of grafted *N*-glycans to potentially reveal determinants of glycan processing and its associated products. In order to achieve this aim, a glycoproteomic dataset was retrieved from GlyConnect to graft *N*-glycan structures from MD-equilibrated *N*-glycan library. The presented work in later sections of the chapter can also be alternatively thought of as an effort to significantly expand the glycosylation site representative dataset. The significant expansion of the dataset enabled the establishment of whether associations discovered in Chapter 3 of the thesis, could be replicated in the context of a bigger sample size.

4.3 Methods

4.3.1 Grafting algorithm implementation in Privateer

The grafting algorithm in Privateer was specifically designed to integrate MD-equilibrated *N*-glycan structures, developed by Dr Elisa Fadda group, onto targeted PDB files. The MD-equilibrated *N*-glycan PDB files exclusively contain the structure of an *N*-glycan, which is then grafted onto a target glycoprotein.

The algorithm as input requires: a donor PDB file containing target glycan coordinates, receiver amino acid numeric identifier and its associated letter chain identifier. The donor PDB file containing a glycan for transplantation has to contain the following: O1 atom at the reducing end of target glycan or be in close proximity of appropriate amino acid residues to derive the potential O1 position using the glycosidic linkage, for example ND2 atom from ASN residue acting as a substitute for O1. Once the glycan is analysed from donor PDB file, the algorithm looks up required atoms for transplantation procedure, for example for *N*-glycosidic linkage with ASN residue: ND2, CB, CG from ASN and C1, O1, O5 from the first sugar of glycan to be transplanted. Once the required atom positions are located a translation matrix is calculated to translate the entire donor glycan in proximity of the glycosylation site by overlaying the O1 atom of the first sugar with the ND2 atom of ASN residue, using the RT-operator[210]. At this point, the grafted glycan is likely to contain a significant number of clashes/atomic overlaps with protein backbone. Initially the algorithm rotates the translated glycan structure to have the following torsion angles for resulting *N*-glycosidic linkage: -97.5 phi and 178 psi. If any clashes are detected, the algorithm attempts to carry out further rotation of the entire grafted glycan structure around resulting glycosidic linkage with the aim to find the best combination of Phi and Psi torsion angles that result in the least amount or no clashes between the glycan and protein backbone. The degree of rotation in current implementation is limited to ±25° for both phi and psi torsion angles. After the most optimal torsional angle combination is found for a glycosidic linkage, the resulting glycoprotein structure is written to file on disk.

4.3.2 Curation of site-specific glycoprotein data from UniProt and GlyConnect

The UniProt database contains a comprehensive list of proteins at various levels of annotation quality. Some UniProt entries contain information about glycosylation as post-translational modification with varying levels of detail. In order to obtain protein identifiers

associated with *N*-glycosylation, the following query "uniprot.org/uniprotkb?query=(ft_carbohyd:asparagine)" was used to obtain UniProt identifiers from the UniProtKB web service. Additional UniProt identifiers associated with *N*-glycosylation were extracted using GlyConnect's Resource Description Framework (RDF) API endpoint, specifically "protein_ref_isoform.rq" sample file[15,211]. Upon curation of target UniProt identifiers, every UniProt ID was queried through GlyConnect SPARQL API endpoint to retrieve relevant information about *N*-glycosylation sites within the protein, particularly associated *N*-glycan structure used to determine *N*-glycan type and amino acid residue. There were multiple instances, where a UniProt identifier query through GlyConnect SPARQL API would fail to return information about *N*-glycan structure, but GlyConnect would still have information about a particular *N*-glycosylation site expressed as composition of monosaccharides that make up the *N*-glycan structure. Therefore, for such cases, retrieved composition of monosaccharides was additionally queried through the API to determine *N*-glycan type. Specifically for processed *N*-glycan representatives, an additional API query to GlyConnect was made to determine the degree of processing for a processed *N*-glycan representative, i.e., whether it was biantennary, triantennary, tetra-antennary or over-tetra-antennary. Finally, the retrieved list of UniPort identifiers and their associated "common names" were cross-referenced with UniProt identifiers obtained from the analysis in Chapter 3 of the thesis. UniProt identifiers and their associated "common names" that were used in Chapter 3 were eliminated from the dataset used in this chapter to gain access into context that was not investigated with the data obtained from wwPDB.

4.3.3 Grafting experiments

Every glycosylation site representative in the dataset underwent associated *N*-glycan grafting. A particular *N*-glycosylation site representative could undergo multiple attempts at *N*-glycan grafting by varying different cluster representatives of donor *N*-glycans. Cluster representatives were sorted by population size, derived from the studies of *N*-glycan 3D architecture in response to the FcγRIIIa glycoprotein structural landscape, with cluster 1 label representing the most populous conformational representative[25]. If no cluster representative was found to produce an *N*-glycan graft without any clashes or atomic overlaps, then the glycosylation site would be removed from the particular dataset for further consideration.

4.3.3.1 Associated *N*-glycan type grafting on predicted AlphaFold structures at scale

This experiment is a direct equivalent of the *N*-glycosylation terminal neighbourhood vicinity analysis presented in Chapter 3 of the thesis but applied to glycoprotein structures obtained

from AlphaFold predictions in combination with *N*-glycan grafting procedure. The retrieved list of UniProt identifiers with their associated *N*-glycan compositions at particular glycosylation sites were cross-checked with available *N*-glycan structures from MD-equilibrated library. If a specific *N*-glycan composition was not available, then for High-Mannose *N*-glycan representatives, Man9 structure was grafted by default and a biantennary non fucosylated complex *N*-glycan (a2g2) was used for processed *N*-glycan representatives. Typically, every *N*-glycan structure in the MD library would be represented by at least 5 conformational representatives that were obtained through MD clustering sampling. If no cluster representative from MD-equilibrated *N*-glycan library successfully produced a grafted glycoprotein structure without any clashes, then the representative was eliminated from further consideration. The computation of amino acid neighbours in the vicinity of terminal sugars of the grafted *N*-glycan were computed using an identical approach, described in detail in Chapter 3.

4.3.3.2 Man9 grafting on predicted AlphaFold structures at scale

Another experiment grafts a universal Man9 *N*-glycan structure on all glycosylation site representatives and compares the difference in amino acid profile between two different outcomes of *N*-glycan processing. The Man9 representative was chosen, as all final *N*-glycan products contained a Man9 *N*-glycan representative at some point in the processing pathway. Seven cluster representatives of a Man9 structure were available. If no cluster representative successfully produced a grafted structure without any clashes, then the representative was eliminated from further consideration. The computation of amino acid neighbours in the vicinity of terminal sugars of the grafted *N*-glycan were computed using an identical approach, described in detail in Chapter 3 of the thesis.

4.3.3.3 Biantennary (a2g2) versus Tri- and Tetrantennary (a3g3) processed *N*-glycan grafting on predicted AlphaFold structures at scale

The final experiment investigates differences in amino acid neighbour profiles, when considering the processing of more processed *N*-glycans. For processed *N*-glycans that were processed beyond the biantennary state towards more antennas, the triantennary complex *N*-glycan representative was chosen (a3g3), with less processed state being represented by a biantennary complex *N*-glycan (a2g2). Five cluster representatives were available for triantennary structures, while six cluster representatives were available for biantennary structures. If no cluster representative successfully produced a grafted structure

without any clashes, then the glycosylation site representative was eliminated from further consideration. The computation of amino acid neighbours in the vicinity of terminal sugars of the grafted *N*-glycan were computed using an identical approach, described in detail in Chapter 3 of the thesis.

4.4 Results

4.4.1 Preliminary investigation of post-predictional modifications in models predicted by AlphaFold

Most of the content in this section is taken word for word from an already published and peer-reviewed article in “Nature Structural & Molecular Biology” under the title of “The case for post-predictional modifications in the AlphaFold Protein Structure Database” by Bagdonas, Fogarty, Fadda and Agirre[212].

4.4.1.1 Published article: The case for post-predictional modifications in the AlphaFold Protein Structure Database

AlphaFold2 has arrived to change workflows in structural biology, for good. However, the algorithm does not account for essential modifications that affect protein structure and function, giving us only part of the picture. Here we discuss how this omission can be addressed in a relatively straightforward manner, leading to a complete structural prediction of complex biomolecular systems.

The recent release of the AlphaFold Protein Structure Database[213] by DeepMind and EMBL-EBI marks a breakthrough in structural biology, making available to the scientific community worldwide, highly accurate structural predictions for 20,000 human proteins and from 20 other biologically relevant organisms, including *E. coli*. Like many scientists working on macromolecular structure, we are genuinely excited about this development, yet we feel that there is a non-negligible potential for misinterpretation of its content in its current form. In particular, the protein-only predictions in the AlphaFold database means that cofactors and, most importantly, co- and post-translational modifications, are understandably – due to the scope of the technique – excluded. Among the most relevant co- and post-translational modifications is protein glycosylation – relevant and very visible, as recent studies of the

dynamics of a fully glycosylated SARS-CoV-2 spike illustrate[190,214]. Indeed, between 50% and 70% of those 20,000 predicted human proteins are believed to be glycosylated[111], but none of this is yet visibly highlighted on the database. Detailed information on the likelihood of modifications is readily available through their links to Uniprot (https://www.uniprot.org), and thus we strongly encourage the users of this fantastic new resource to check the information available on Uniprot before downloading a model.

Within this framework, we believe that the absence of cofactors and of co-/post-translational modifications in the models in the AlphaFold Protein Structure Database might be remediated through the use of sequence and structure-based comparative studies. Indeed, in the specific case of glycosylation, the algorithms implemented by DeepMind have digested inter-residue distances from the Protein Data Bank[215], where glycosylated proteins often exhibit either full or partial glycan structures; therefore, the space where unmodeled modifications, such as protein glycosylation, should be somehow preserved in AlphaFold models, allowing for these structural features to be directly grafted onto a model. To demonstrate the potential of this approach, we have developed proof-of-concept functionality that grafts protein glycosylation from a library of structurally equilibrated glycan blocks, obtained from molecular dynamics (MD)[25], into an AlphaFold model. This task has been automated and integrated into the new Python interface of the carbohydrate-specific Privateer software[24] and is available to all on its GitHub repository (https://github.com/glycojones/privateer.git). Figure 4.1 shows AlphaFold model P29016 (depicted in magenta) of a human T-cell surface glycoprotein Cd1b, superposed onto the protein's crystal structure PDB 5WL1. The latter was expressed in an insect cell line and shows a characteristic double core-fucosylation of the *N*-glycans, which were omitted in Figure 4.1 for clarity. The *N*-glycan our tool grafted onto the AlphaFold model is not just compatible with the available space, but shows a high complementarity to the protein surface, where the Man6 core is involved with Trp 23 in a CH-π interaction[188], as seen in the crystal structure.

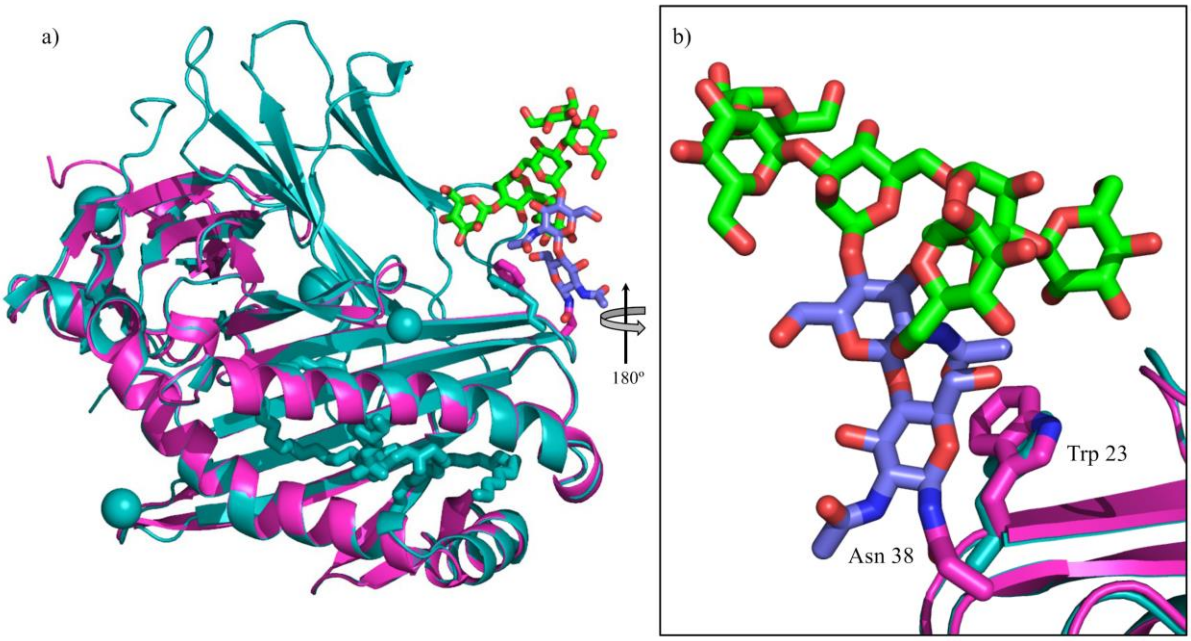


Figure 4.1: Panel a) Structural alignment of the crystal structure of human CD1b in complex with phosphatidylglycerol (PDB 5WL1), shown in cyan, onto the model predicted by AlphaFold (accession code P29016), shown in magenta. The *N*-glycosylation at position N38 was reconstructed with Privateer[24], where the linked Man6 structure was selected from a library of highly populated conformers at equilibrium, obtained from molecular dynamics simulations at 300 K[25]. Panel b) Close-up view of the grafted Man6, with the structure rotated around the z-axis by 180°, represented in sticks with colouring compliant to the SNFG scheme. The relative positions of the Trp 23 sidechain stacking the Man6 core is highlighted in sticks in both the crystal structure (cyan) and in the AlphaFold model (magenta).

We would like to emphasise that this approach may be also useful to complete the AlphaFold models in the database with other types of modifications. For example, the AlphaFold model P6887, a Haemoglobin subunit beta, contains a heme binding site with just enough space for a heme cofactor. Certain structure completions will only be feasible via automated comparative analyses against available structural information – e.g. co-translational modifications such as myristoylation[216], or *O*-GlcNAcylation[217] – while others such as *N*-glycosylation or tryptophan mannosylation, which rely on consensus sequences, will be more amenable to prediction. As comparative studies would have to rely on experimental structural information, positional uncertainty (e.g. a pLDDT-like score[81]) may be estimated by comparing the placed coordinates to a superposition of the available structural information. However, in the particular case of protein glycosylation, we see more of a compositional problem; indeed, the biggest challenge would be to get a good estimation of

what glycoform is linked to each sequon. Experimental structures offer only partial information due to limiting factors such as mobility and micro-heterogeneity[110], so other sources of knowledge (e.g. glycomics, molecular dynamics simulations) ought to be used, especially when attempting to model full-length glycans, which is something we are sure the glycobiology community will appreciate. We are expanding the Privateer software to address these cases, by harnessing the rich information available in glycomics databases[2].

To conclude, we think that these early results are highly encouraging to serve as a rallying point for the developers' community to complete and enrich the predicted protein models with likely modifications, to bring them to their fullest potential and correctly inform the next generation of structural biology studies.

4.4.2 Grafting experiments to investigate *N*-glycan processing

This section expands upon the foundational work introduced in the previous section, concentrating on the analysis of *N*-glycan processing through grafting experiments. To explore glycoprotein structures not included in the dataset from Chapter 3, a combination of AlphaFold predictions and glycoproteomic data from the GlyConnect database were utilized. This approach allowed for detailed analysis of the terminal neighbourhoods of grafted *N*-glycans. Re-modelling of structures predicted by AlphaFold, enables the capture of potential snapshots of amino acid neighbourhood contexts in the vicinity of terminal sugars, which could potentially inform the dynamics of *N*-glycan processing, as demonstrated in Figure 4.2. Most crucially, the analysis provided in this chapter enables the comparison of findings discovered in Chapter 3 of the thesis, in principle by validating whether the findings based on analysis of glycoprotein models deposited to wwPDB could be replicated by alternative methodology.

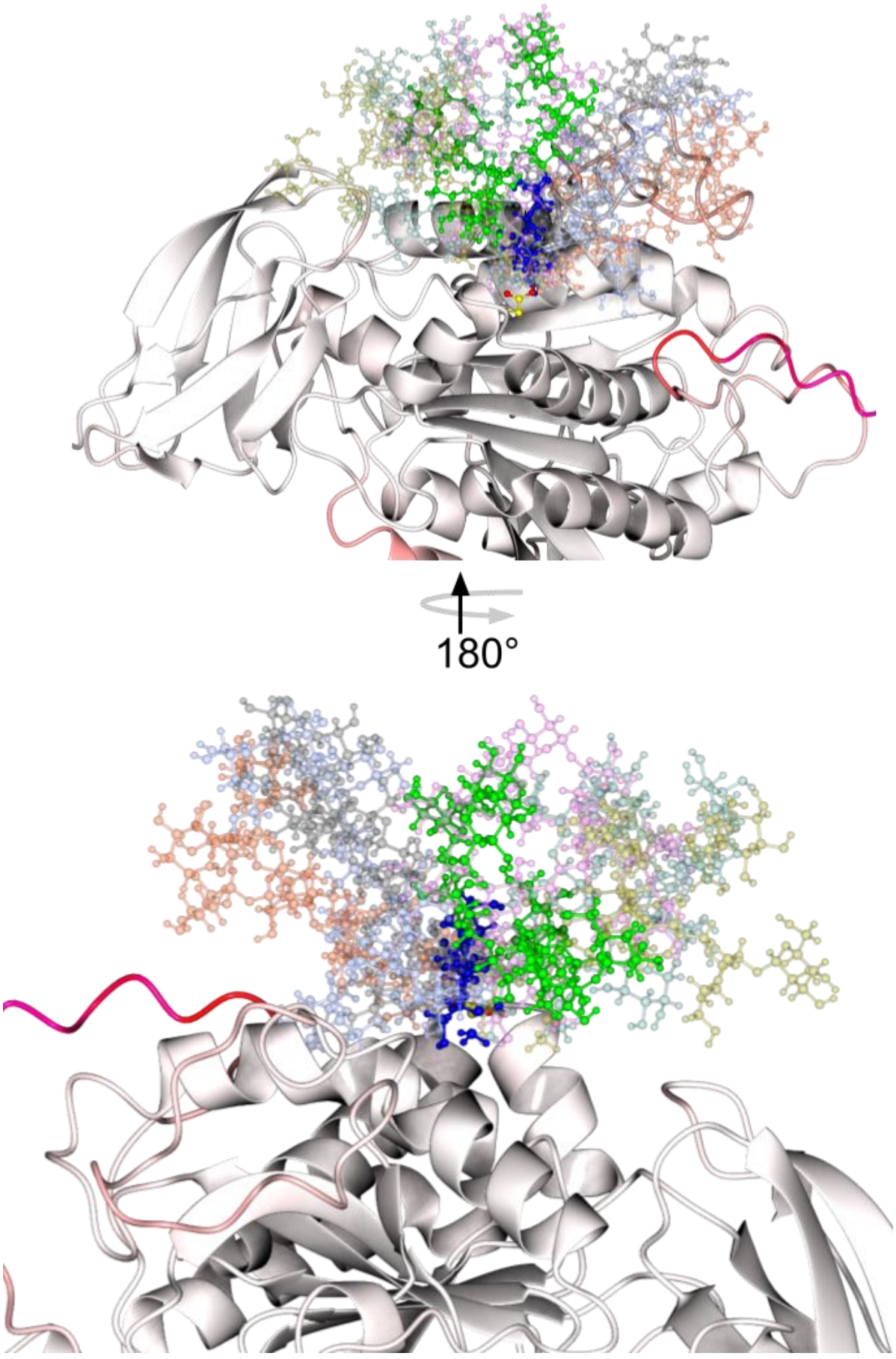

180°

Figure 4.2: Visualisation of Man9 *N*-glycan grafting attempt at ASN139 of P16870 predicted by AlphaFold. Protein backbone depicted in ribbon representation, with the colour scheme portraying pLDDT score (residues coloured in white represent high confidence). Six cluster representatives of Man9 *N*-glycan were grafted that produced clashes with protein backbone (coloured in transparent colours), with the seventh and only cluster representative producing a graft that did not produce any clashes (coloured in non-transparent colours, according to the SNFG colour scheme for individual monosaccharides, i.e., blue - GlcNAc sugar, green - Man sugar)

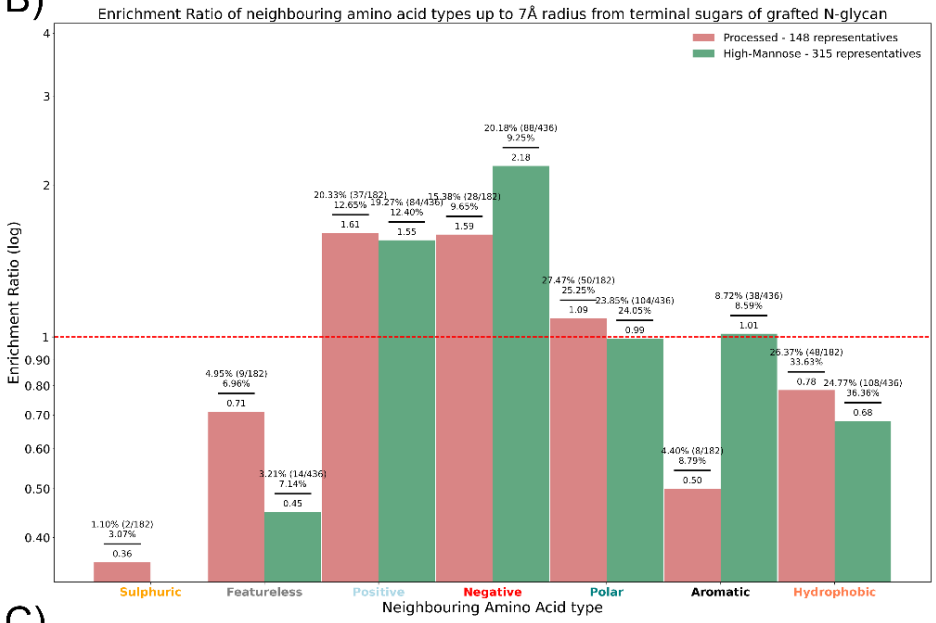
The data curation and association procedure yielded 1,174 representative *N*-glycosylation sites in 896 unique proteins, based on UniProt identifier and name label. After cross-reference with glycoproteins analysed in Chapter 3 of the thesis, 1,129 representative *N*-glycosylation sites from 865 unique proteins remained that were not analysed in the previous analysis.

4.4.2.1 Associated *N*-glycan structure grafting on predicted AlphaFold structures at scale

The initial experiment of associated *N*-glycan type grafting on predicted AlphaFold structures at scale attempted to replicate the analysis of Chapter 3, with the difference being that it would represent unseen glycoprotein structures. The grafting procedure resulted in successfully producing 643 unique glycoproteins containing grafted *N*-glycans without any instances of atomic overlaps/clashes between grafted *N*-glycan and protein backbone. Unfortunately, 486 instances out of 1,129 total extracted representatives from GlyConnect and UniProt, failed to produce a usable structure for neighbourhood vicinity analysis due to containing at least one instance of atomic overlap/clashing between grafted *N*-glycan structure and any atom from any amino acid residue in the protein backbone. Upon further analysis, 180 out 643 successful grafting events produced glycoprotein structures containing grafted *N*-glycans that had no amino acid neighbourhoods in the vicinity of terminal sugars, therefore being removed from further consideration. Therefore, the terminal sugar neighbourhood analysis considered 463 *N*-glycosylation site representatives, composed of 397 unique glycoproteins, which had at least a single amino acid neighbour at the *N*-glycan termini. The total 463 *N*-glycosylation site representatives were composed of 315 high-mannose *N*-glycan representatives, 126 representatives of complex biantennary *N*-glycosylations and 22 complex triantennary *N*-glycosylations. The terminal neighbourhood vicinity analysis for associated *N*-glycan type grafting on predicted AlphaFold structures at

scale are represented by the amino acid type enrichment ratios over varying distance radius thresholds, depicted in Figure 4.3.


A)
Enrichment Ratio of neighbouring amino acid types up to 5Å radius from terminal sugars of grafted N-glycan
Processed - 148 representatives
High-Mannose - 315 representatives
Enrichment Ratio (log)
4.55% (2/44)
6.96%
0.65
3.57% (3/84)
7.14%
0.50
11.36% (5/44)
12.65%
0.90
19.05% (16/84)
12.40%
1.54
18.18% (8/44)
9.65%
1.88
20.24% (17/84)
9.25%
2.19
27.27% (12/44)
25.25%
1.08
20.24% (17/84)
24.05%
0.84
4.55% (2/44)
8.79%
0.52
10.71% (9/84)
8.59%
1.25
34.09% (15/44)
33.63%
1.01
26.19% (22/84)
36.36%
0.72
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic
Neighbouring Amino Acid type


B)
Enrichment Ratio of neighbouring amino acid types up to 7Å radius from terminal sugars of grafted N-glycan
Processed - 148 representatives
High-Mannose - 315 representatives
Enrichment Ratio (log)
1.10% (2/182)
3.07%
0.36
4.95% (9/182)
6.96%
0.71
3.21% (14/436)
7.14%
0.45
20.33% (37/182)
12.65%
1.61
19.27% (84/436)
12.40%
1.55
15.38% (28/182)
9.65%
1.59
20.18% (88/436)
9.25%
2.18
27.47% (50/182)
25.25%
1.09
23.85% (104/436)
24.05%
0.99
4.40% (8/182)
8.79%
0.50
8.72% (38/436)
8.59%
1.01
26.37% (48/182)
33.63%
0.78
24.77% (108/436)
36.36%
0.68
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic
Neighbouring Amino Acid type


C)
Enrichment Ratio of neighbouring amino acid types up to 9Å radius from terminal sugars of grafted N-glycan
Processed - 148 representatives
High-Mannose - 315 representatives
Enrichment Ratio (log)
1.33% (7/527)
3.07%
0.43
0.74% (9/1209)
2.22%
0.34
4.74% (25/527)
6.96%
0.68
6.29% (76/1209)
7.14%
0.88
17.46% (92/527)
12.65%
1.38
18.20% (220/1209)
12.40%
1.47
14.18% (75/527)
9.65%
1.48
15.38% (186/1209)
9.25%
1.66
32.07% (169/527)
25.25%
1.27
26.96% (326/1209)
24.05%
1.12
4.74% (25/527)
8.79%
0.54
8.11% (98/1209)
8.59%
0.94
25.43% (134/527)
33.63%
0.76
24.32% (294/1209)
36.36%
0.67
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic
Neighbouring Amino Acid type

Figure 4.3: Amino acid type enrichment ratios over a variety of distance radius threshold cutoffs in vicinity of terminal sugars of associated *N*-glycan type grafts on predicted AlphaFold structures at scale. A) 5 Å radius threshold, B) 7 Å radius threshold, C) 9 Å radius threshold. The labels above bars represent the exact values that were used to calculate enrichment ratio for individual Amino Acid types in the dataset. Values in the bracket on the numerator represent the number of Amino Acid type detections out of total neighbouring amino acids detected.

The enrichment analysis of neighbouring amino acids at the terminal sugars of associated *N*-glycan type grafts on predicted AlphaFold structures at scale appears to contain a number of discriminatory features between different types of *N*-glycans. Positive amino acids, characterised by their basic nature, are enriched and preferred as immediate neighbours of high-mannose termini, nevertheless the relationship is diminished as search radius is increased with eventual enrichment for both types of *N*-glycans. Negative amino acids, characterised by their acidic nature, display no discriminatory signal, as the enrichment is consistent for both types of *N*-glycan products across all different search criteria. Polar uncharged amino acids present a tendency to be preferred by processed *N*-glycan termini as immediate neighbours, although the trend is diminished as search radius distance criteria is increased through enrichment for both types of *N*-glycan products. Aromatic amino acids demonstrate an obvious discriminatory relationship, as high-mannose *N*-glycan termini display preference and enrichment throughout multiple search radius distances. Hydrophobic amino acids demonstrate slight preference by processed *N*-glycan termini in the immediate vicinity, although the discriminatory relationship is diminished as search radius distance is increased. The most notable findings of the enrichment analysis are summarised in Table 4.1.

Table 4.1: A summary of the most significant amino acid type preference relationships near *N*-glycan Termini for grafted *N*-glycans with their associated structures. Label designation - **N+**: High-mannose *N*-glycan positive enrichment, **N-**: High-mannose *N*-glycan negative enrichment, **P+**: Processed *N*-glycan positive enrichment, **P-**: Processed *N*-glycan negative enrichment. The colouring on the amino acid labels corresponds to a direct mapping described in Figure 4.3.

Amino acid type	5 Å radius threshold	7 Å radius threshold	9 Å radius threshold
Positive	**N+, P-**	**N+, P+**	**N+, P+**
Negative	**N+, P+**	**N+, P+**	**N+, P+**
Polar uncharged	**N-, P+**	**N-, P+**	**N+, P-**
Aromatic	**N+, P-**	**N+, P-,**	**N-, P-**
Hydrophobic	**N-, P+**	**N-, P-**	**N-, P-**

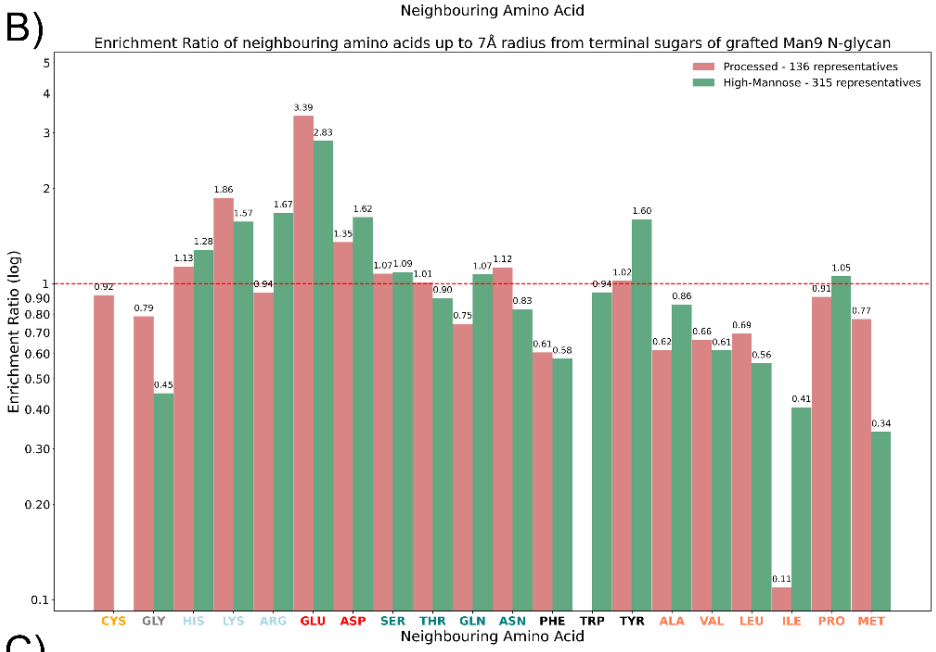
The results obtained from this experiment can also be cross-referenced with results obtained in Chapter 3. The pattern in terms of Sulphuric, Featureless and Hydrophobic amino acids being consistently unenriched between the two types of *N*-glycans agrees with analysis results based on data obtained from wwPDB. Although, the neighbourhood data from wwPDB suggests that there is a degree of discrimination between the two types of *N*-glycans when considering closest neighbours, the pattern for the aforementioned amino acid types could not be replicated in this experiment. Disappointingly, an identical pattern could neither be replicated for Positive, Negative and Polar amino acid types when compared to results in Chapter 3. The results in the previous chapter suggested that there was a degree of discrimination in terms of Positive and Negative amino acid neighbour types being preferred by more processed *N*-glycans. Unfortunately, the results from this experiment suggest that Positive and Negative amino acid neighbours are preferred by high-mannose *N*-glycans as immediate neighbours instead, with the preference relationship being eliminated at higher distance threshold. Nevertheless, the two experiments are indeed in agreement in terms of enrichment of these specific amino acid types as neighbours of grafted *N*-glycans at the terminal ends. Most crucially, the results from Chapter 3 demonstrated a clear discriminatory relationship between the two *N*-glycan types in terms of Polar uncharged amino acid type neighbours being preferred by high-mannose *N*-glycans, however, this association was not successfully reproduced in this experiment, with both types of *N*-glycans demonstrating near similar preference and enrichment patterns. Nevertheless, the grafting experiment was able to maintain a similar discriminatory

relationship of Aromatic amino acids neighbours being located in the vicinity of high-mannose *N*-glycans.

4.4.2.2 Man9 grafting on predicted AlphaFold structures at scale

Upon the analysis of grafted *N*-glycans by their associated structures, it was noted that most of the grafted *N*-glycan structures were defaults, due to unavailable *N*-glycan structure in the MD-equilibrated *N*-glycan library. Moreover, the results obtained in Chapter 3 relied on *N*-glycan structures that were shorter in terms of oligosaccharide length, especially in high-mannose *N*-glycan representatives. This is an important consideration, as the output from the vicinity scans relies on the positioning of the probe point that is directly dependent on the length of an oligosaccharide. Therefore, to improve consistency in terms of probe positioning, the second experiment of associated *N*-glycan type grafting on predicted AlphaFold structures relied on grafting universal *N*-glycan structure (Man9) on all representatives and using it as a basis to compute neighbourhood profile. Man9 was chosen, as this specific *N*-glycan structure is the predecessor to all *N*-glycan products. Arguably, the design of this experiment is superior in terms of having a uniform *N*-glycan representative, enabling the placement of probe points to be relatively consistent in 3D space between different glycosylation site representatives. After selecting for structures that had successfully produced a graft without any clashes and more than one amino acid neighbours in the vicinity of terminal sugars, 452 representatives were successfully produced. The terminal neighbourhood analysis for associated *N*-glycan type determined by the profile generated by the grafted Man9 *N*-glycan on predicted AlphaFold structures is first represented by the individual amino acid enrichment ratios over varying distance radius thresholds, depicted in Figure 4.4.


A)
Enrichment Ratio of neighbouring amino acids up to 5Å radius from terminal sugars of grafted Man9 N-glycan
Processed - 136 representatives
High-Mannose - 315 representatives
Enrichment Ratio (log)
CYS GLY HIS LYS ARG GLU ASP SER THR GLN ASN PHE TRP TYR ALA VAL LEU ILE PRO MET
Neighbouring Amino Acid


B)
Enrichment Ratio of neighbouring amino acids up to 7Å radius from terminal sugars of grafted Man9 N-glycan
Processed - 136 representatives
High-Mannose - 315 representatives
Enrichment Ratio (log)
CYS GLY HIS LYS ARG GLU ASP SER THR GLN ASN PHE TRP TYR ALA VAL LEU ILE PRO MET
Neighbouring Amino Acid

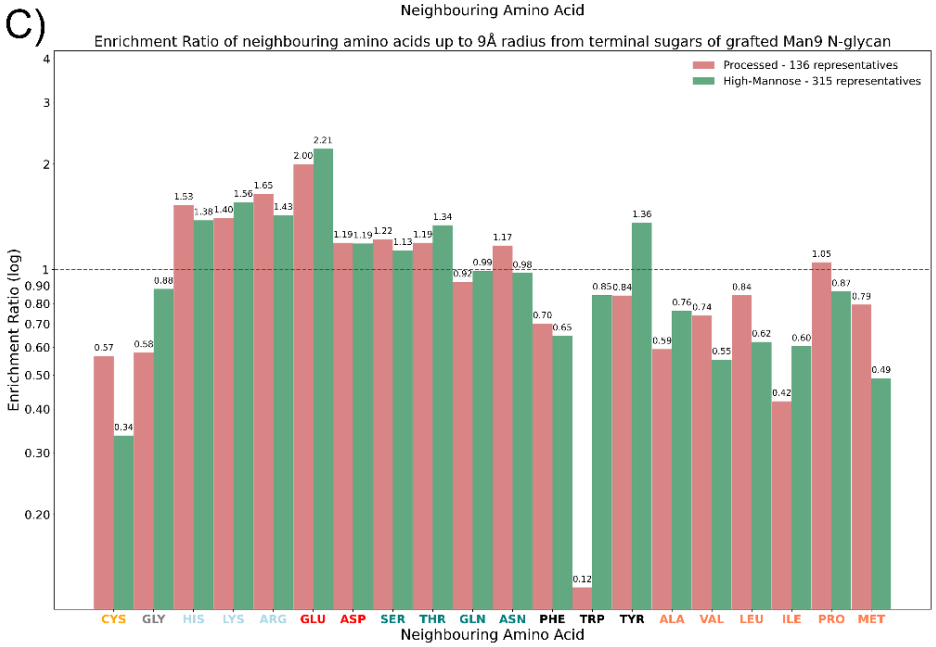

C)
Enrichment Ratio of neighbouring amino acids up to 9Å radius from terminal sugars of grafted Man9 N-glycan
Processed - 136 representatives
High-Mannose - 315 representatives
Enrichment Ratio (log)
CYS GLY HIS LYS ARG GLU ASP SER THR GLN ASN PHE TRP TYR ALA VAL LEU ILE PRO MET
Neighbouring Amino Acid

Figure 4.4: Amino acid enrichment ratios over a variety of distance radius threshold cutoffs in vicinity of terminal sugars of Man9 *N*-glycan grafts on predicted AlphaFold structures at scale. A) 5 Å radius threshold, B) 7 Å radius threshold, C) 9 Å radius threshold. The labels above bars represent the enrichment ratio in the dataset for individual amino acids.

Upon the grafting of a universal Man9 *N*-glycan, certain amino acids demonstrate distinct tendencies in proximity to *N*-glycan termini. A small number of Cysteine instances are found adjacent to processed *N*-glycan termini, although enrichment threshold is never exceeded at multiple search distance criterion. Even though many more Glycine neighbour instances are detected in the vicinity of both processed and high-mannose *N*-glycans, likewise the enrichment threshold is not exceeded across multiple search distance criteria. Histidines appear to demonstrate discrimination between *N*-glycan types in the immediate vicinity, with enrichment for high-mannose *N*-glycans, although the trend is diminished as search radius distance criteria is increased. Lysines exhibit preference together with enrichment for processed *N*-glycans in the immediate vicinity with diminishing trend as search radius distance is increased. Arginines, Glutamates and Aspartates appear to display a similar relationship, in that there is a similar degree of enrichment for both types of *N*-glycans without apparent discrimination across multiple search distance thresholds. Serines demonstrate a stark preference by high-mannose *N*-glycans in the immediate vicinity through an enrichment that gets diminished as the search distance radius threshold is increased. Likewise, some degree of preference for Threonines and Glutamines is demonstrated by high-mannose *N*-glycans in the immediate vicinity, although these amino acids are not enriched. The discrimination is diminished as the search radius distance threshold is increased, eventually demonstrating an insignificant degree of enrichment for both types of *N*-glycans. On the other hand, Asparagines demonstrate preference for processed *N*-glycans in the immediate vicinity, which gets diminished as the search radius distance threshold is increased. Phenylalanines demonstrate a small degree of preference for high-mannose *N*-glycans in the immediate vicinity, although enrichment is never exceeded as the search radius distance threshold is increased. Tryptophans demonstrate the starkest preference for high-mannose *N*-glycans consistently across multiple search distance criteria, although likewise the enrichment threshold is not exceeded across any of the search distance criteria. Tyrosines, on the other hand, demonstrate consistent enrichment for both types of *N*-glycans, with a small degree of preference for high-mannose *N*-glycans. Finally, hydrophobic amino acids, apart from Proline, demonstrate a similar pattern of lack of enrichment without a notable discriminatory relationship. Prolines, on the other hand, appear to demonstrate a notable preference for high-mannose *N*-glycans that are slightly enriched, although the relationship is diminished as the search radius distance

threshold is increased. The most notable findings of the enrichment analysis are summarised in Table 4.2.

Table 4.2: A summary of the most significant amino acid preference relationship near *N*-glycan Termini. Label designation - **N+**: High-mannose *N*-glycan positive enrichment, **N-**: High-mannose *N*-glycan negative enrichment, **P+**: Processed *N*-glycan positive enrichment, **P-**: Processed *N*-glycan negative enrichment. The colouring on the amino acid labels corresponds to a direct mapping described in Figure 4.3.

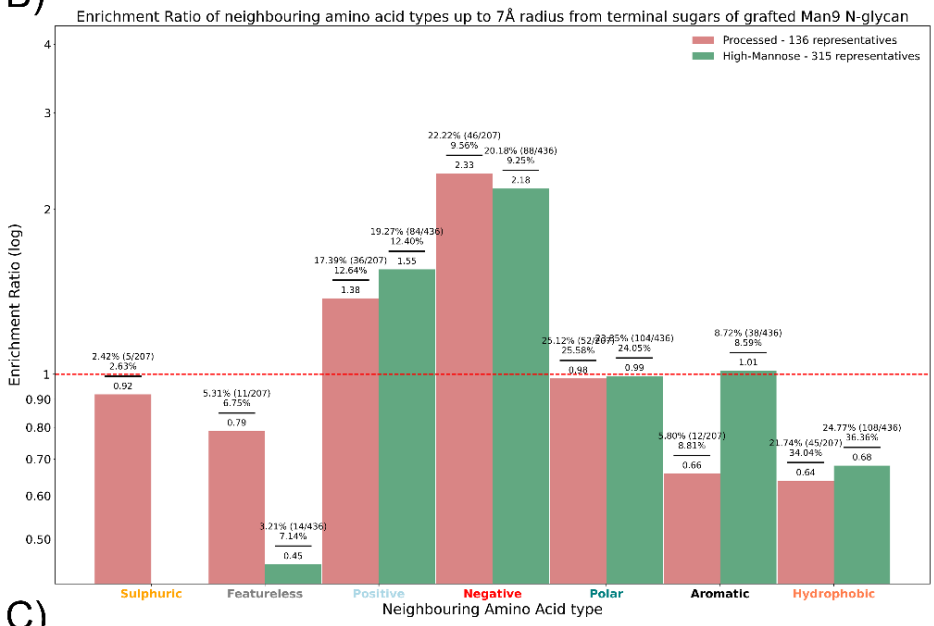
Amino acid	5 Å radius threshold	7 Å radius threshold	9 Å radius threshold
Histidine	**N+, P-**	**N+, P+**	**N+, P+**
Lysine	**N+, P+**	**N+, P+**	**N+, P+**
Serine	**N+, P-**	**N+, P+**	**N+, P+**
Tryptophan	**N-, P-**	**N-, P-**	**N-, P-**
Tyrosine	**N+, P+**	**N+, P+**	**N+, P-**
Proline	**N+, P-**	**N+, P-**	**N+, P-**

Analysis focusing on detected individual amino acids within the vicinity of grafted Man9 *N*-glycans has demonstrated to be challenging due to lack of obvious discriminatory relationships. The analysis has especially proven difficult due to instances of single or low number observations of specific amino acids within immediate vicinity search area distance threshold, therefore not allowing to draw strong conclusions. Therefore, following the analysis of individual amino acid neighbours of terminal sugars within modelled *N*-glycans, an attempt was made to group 20 amino acids by redundant chemical features to simplify interpretation. The output of the analysis of terminal amino acid neighbours grouped by redundant chemical features is shown in Figure 4.5.

A)


Enrichment Ratio of neighbouring amino acid types up to 5Å radius from terminal sugars of grafted Man9 N-glycan
Processed - 136 representatives
High-Mannose - 315 representatives
Enrichment Ratio (log)
Neighbouring Amino Acid type
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic

B)


Enrichment Ratio of neighbouring amino acid types up to 7Å radius from terminal sugars of grafted Man9 N-glycan
Processed - 136 representatives
High-Mannose - 315 representatives
Enrichment Ratio (log)
Neighbouring Amino Acid type
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic

C)

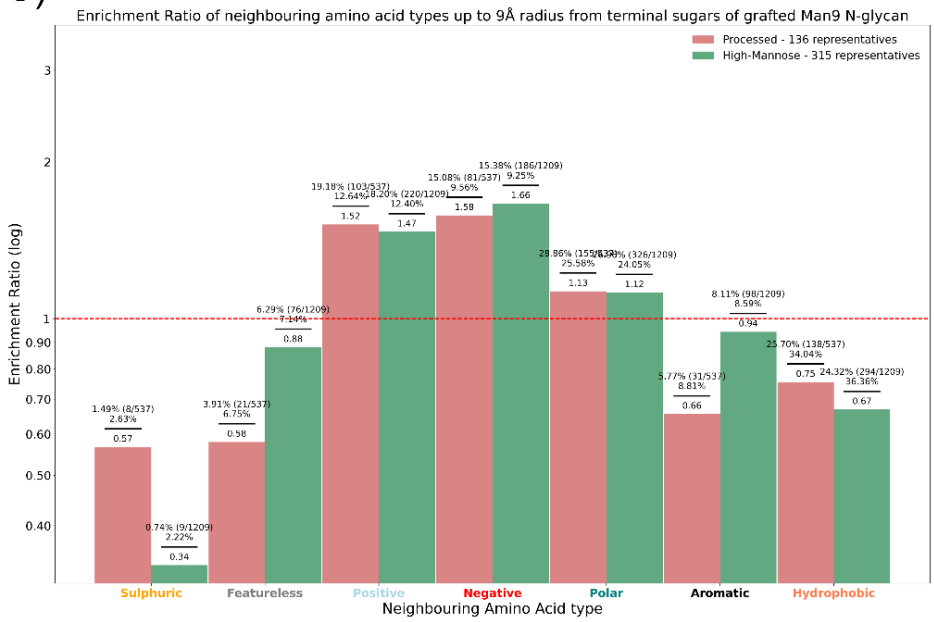

Enrichment Ratio of neighbouring amino acid types up to 9Å radius from terminal sugars of grafted Man9 N-glycan
Processed - 136 representatives
High-Mannose - 315 representatives
Enrichment Ratio (log)
Neighbouring Amino Acid type
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic

Figure 4.5: Amino acid type enrichment ratios over a variety of distance radius threshold cutoffs in vicinity of terminal sugars of Man9 *N*-glycan grafts on predicted AlphaFold structures at scale. A) 5 Å radius threshold, B) 7 Å radius threshold, C) 9 Å radius threshold. The labels above bars represent the exact values that were used to calculate enrichment ratio for individual Amino Acid types in the dataset. Values in the bracket on the numerator represent the number of Amino Acid type detections out of total neighbouring amino acids detected.

The enrichment analysis of neighbouring amino acids at the terminal sugars of Man9 *N*-glycan grafts on predicted AlphaFold structures at scale appears to contain two discriminatory observations, concerning Polar uncharged and Aromatic amino acids. When a Man9 *N*-glycan is grafted, polar uncharged amino acids present a tendency to be preferred by high-mannose *N*-glycan termini as immediate neighbours, although the trend is diminished as search radius distance criteria is increased through enrichment for both types of *N*-glycan products. Aromatic amino acids demonstrate an obvious discriminatory relationship via preference for high-mannose *N*-glycan termini with consistent enrichment throughout multiple search radius distances. The most notable findings in terms of discriminatory relationships of the enrichment analysis are summarised in Table 4.3.

Table 4.3: A summary of the most significant amino acid type preference relationships near *N*-glycan Termini for grafted *N*-glycans with their associated structures. Label designation - **N+**: High-mannose *N*-glycan positive enrichment, **N-**: High-mannose *N*-glycan negative enrichment, **P+**: Processed *N*-glycan positive enrichment, **P-**: Processed *N*-glycan negative enrichment. The colouring on the amino acid labels corresponds to a direct mapping described in Figure 4.3.

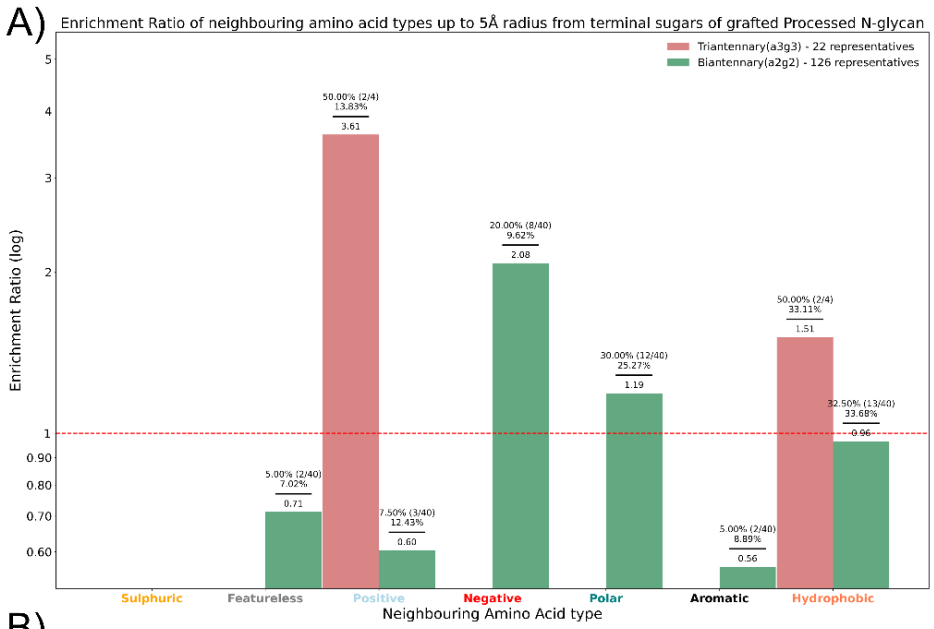
Amino acid type	5 Å radius threshold	7 Å radius threshold	9 Å radius threshold
Polar uncharged	**N-, P+**	**N-, P+**	**N+, P-**
Aromatic	**N+, P-**	**N+, P-,**	**N-, P-**

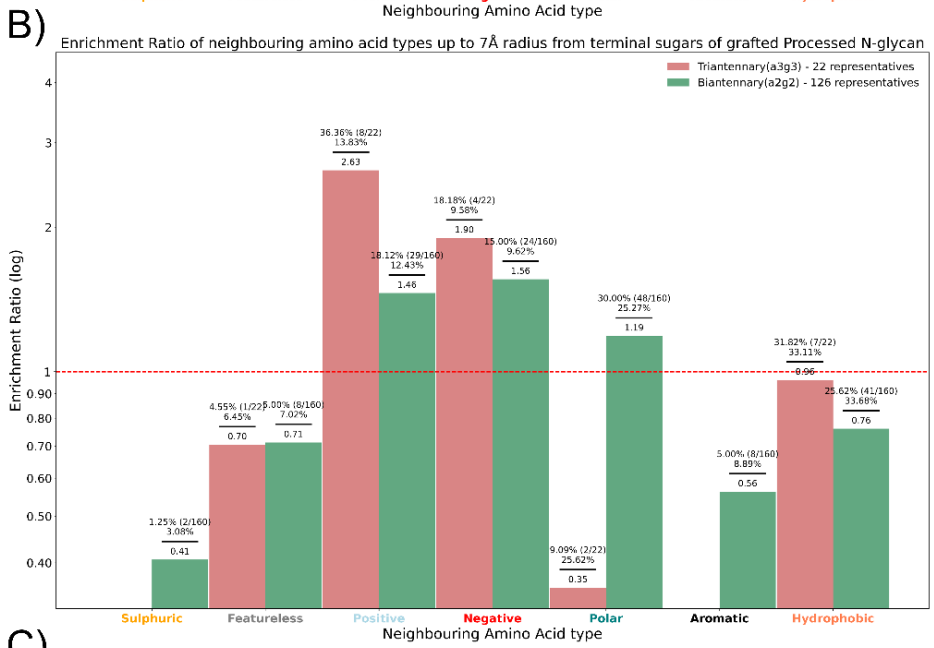
Grafting a universal Man9 *N*-glycan structure appears to reproduce a more similar pattern to the results observed in Chapter 3. The pattern in terms of Sulphuric, Featureless and Hydrophobic amino acids being consistently unenriched between the two types of *N*-glycans is in agreement with analysis results based on data obtained from wwPDB. When Positive and Negative amino acids are considered, the enrichment patterns could also be replicated, although without obvious discriminatory relationship patterns, in terms of *N*-glycan type

preference for specific amino acid groups as observed in Chapter 3. Nevertheless, when Polar uncharged amino acids are considered, the pattern of discrimination between different *N*-glycan types could somewhat be replicated at lowest radius threshold distances, with the relationship being lost at higher radius threshold distances. Most significantly, the Man9 grafting experiment was able to maintain a similar discriminatory relationship of Aromatic amino acids neighbours being located in the vicinity of high-mannose *N*-glycans.

4.4.2.3 Biantennary (a2g2) versus Tri- and Tetrantennary (a3g3) processed *N*-glycan grafting on predicted AlphaFold structures at scale

The final experiment analysed the degree of *N*-glycan processing within the representatives containing processed *N*-glycans. After selecting for structures that had successfully produced a graft without any clashes and more than one amino acid neighbours in the vicinity of terminal sugars, 148 representatives were successfully produced. The terminal neighbourhood analysis for processed *N*-glycans was determined by either grafting biantennary non-fucosylated complex (a2g2) *N*-glycan for representatives that contained the least drastic degree of processing or triantennary non-fucosylated complex (a3g3) *N*-glycan for representatives that contained the more drastic degree of processing (triantennary, tetra-antennary) complex *N*-glycan. Therefore, 148 processed representatives were composed of 126 biantennary (a2g2) and 22 triantennary (a3g3) triantennary representatives. Finally, the analysis is summarised by the amino acid type enrichment ratios over varying distance radius thresholds, depicted in Figure 4.6.


A)
Enrichment Ratio of neighbouring amino acid types up to 5Å radius from terminal sugars of grafted Processed N-glycan
Triantennary(a3g3) - 22 representatives
Biantennary(a2g2) - 126 representatives
Enrichment Ratio (log)
5.00% (2/40) 7.02% 0.71
50.00% (2/4) 13.83% 3.61
7.50% (3/40) 12.43% 0.60
20.00% (8/40) 9.62% 2.08
30.00% (12/40) 25.27% 1.19
5.00% (2/40) 8.89% 0.56
50.00% (2/4) 33.11% 1.51
32.50% (13/40) 33.68% 0.96
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic
Neighbouring Amino Acid type


B)
Enrichment Ratio of neighbouring amino acid types up to 7Å radius from terminal sugars of grafted Processed N-glycan
Triantennary(a3g3) - 22 representatives
Biantennary(a2g2) - 126 representatives
Enrichment Ratio (log)
1.25% (2/160) 3.08% 0.41
4.55% (1/22) 6.45% 0.70
5.00% (8/160) 7.02% 0.71
36.36% (8/22) 13.83% 2.63
18.12% (29/160) 12.43% 1.46
18.18% (4/22) 9.58% 1.90
15.00% (24/160) 9.62% 1.56
9.09% (2/22) 25.62% 0.35
30.00% (48/160) 25.27% 1.19
5.00% (8/160) 8.89% 0.56
31.82% (7/22) 33.11% 0.96
25.62% (41/160) 33.68% 0.76
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic
Neighbouring Amino Acid type

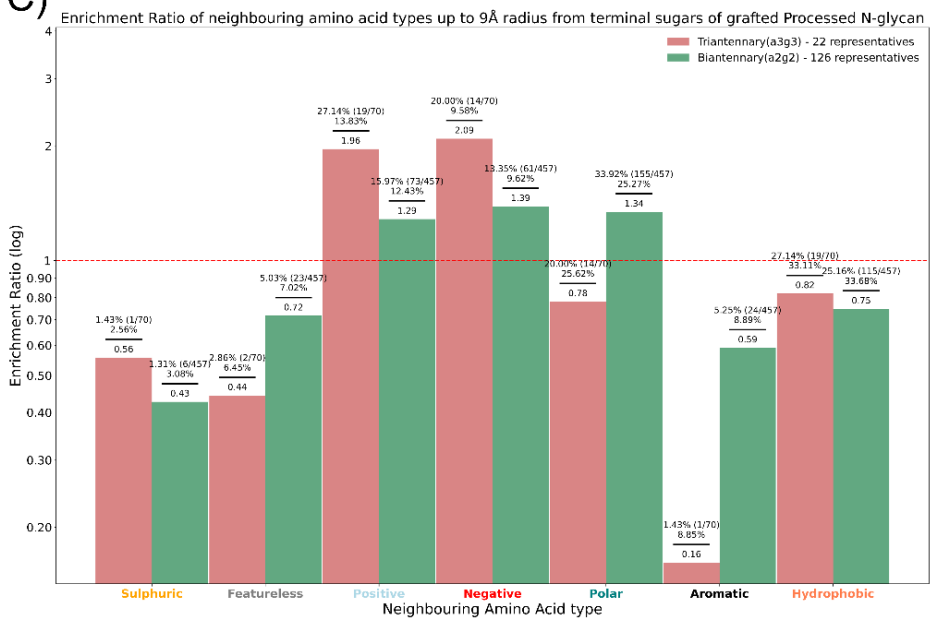

C)
Enrichment Ratio of neighbouring amino acid types up to 9Å radius from terminal sugars of grafted Processed N-glycan
Triantennary(a3g3) - 22 representatives
Biantennary(a2g2) - 126 representatives
Enrichment Ratio (log)
1.43% (1/70) 2.56% 0.56
1.31% (6/457) 3.08% 0.43
2.86% (2/70) 6.45% 0.44
5.03% (23/457) 7.02% 0.72
27.14% (19/70) 13.83% 1.96
15.97% (73/457) 12.43% 1.29
20.00% (14/70) 9.58% 2.09
13.35% (61/457) 9.62% 1.39
20.00% (14/70) 25.62% 0.78
33.92% (155/457) 25.27% 1.34
1.43% (1/70) 8.85% 0.16
5.25% (24/457) 8.89% 0.59
27.14% (19/70) 33.11% 0.82
25.16% (115/457) 33.68% 0.75
Sulphuric
Featureless
Positive
Negative
Polar
Aromatic
Hydrophobic
Neighbouring Amino Acid type

Figure 4.6: Amino acid type enrichment ratios over a variety of distance radius threshold cutoffs in vicinity of terminal sugars of processed *N*-glycan grafts on predicted AlphaFold structures at scale. A) 5 Å radius threshold, B) 7 Å radius threshold, C) 9 Å radius threshold. The labels above bars represent the exact values that were used to calculate enrichment ratio for individual Amino Acid types in the dataset. Values in the bracket on the numerator represent the number of Amino Acid type detections out of total neighbouring amino acids detected.

The enrichment analysis of neighbouring amino acids at the terminal sugars of associated processed *N*-glycan grafts on predicted AlphaFold structures at scale appears to suggest a few notable discriminatory relationships. Amino acids with basic characteristics, termed as positive amino acids, reveal a tendency to be preferred by the more processed *N*-glycans, especially as immediate neighbours when a triantennary complex *N*-glycan is grafted. As the search radius distance is increased, both biantennary and triantennary complex *N*-glycans show enrichment. Negative amino acids, known for their acidic nature, are preferred and enriched near the terminal sugars of less processed complex *N*-glycans. Similarly, polar uncharged and aromatic amino acids display a similar relationship, being preferred and enriched close to the terminal sugars of the less processed complex *N*-glycans. Notably, as the search radius distance grows, both types of complex *N*-glycans, biantennary and triantennary, consistently demonstrate enrichment. Lastly, hydrophobic amino acids seem to have an affinity for the terminal sugars of more processed *N*-glycans, although this association dwindles as the radius distance threshold increases. The most notable findings of the enrichment analysis are summarised in Table 4.4.

Table 4.4: A summary of the most significant amino acid type preference relationships near processed *N*-glycan Termini for grafted *N*-glycans. Label designation - **N+**: Less processed *N*-glycan (a2g2) positive enrichment, **N-**: Less processed *N*-glycan (a2g2) negative enrichment, **P+**: More processed *N*-glycan (a3g3) positive enrichment, **P-**: More processed *N*-glycan (a3g3) negative enrichment. The colouring on the amino acid labels corresponds to a direct mapping described in Figure 4.3.

Amino acid type	5 Å radius threshold	7 Å radius threshold	9 Å radius threshold
Positive	**N-, P+**	**N+, P+**	**N+, P+**
Negative	**N-, P+**	**N+, P+**	**N+, P+**
Polar uncharged	**N-, P+**	**N-, P+**	**N-, P+**
Aromatic	**N-, P-**	**N-, P-,**	**N-, P-**
Hydrophobic	**N-, P+**	**N-, P-**	**N-, P-**

In summary, the results of the three experiments demonstrate that the observations made in the Chapter 3 could partially be replicated through in-silico *N*-glycan grafting on protein structures predicted by AlphaFold. Therefore, it can be concluded that there may indeed discriminatory features in terms of neighbouring amino acid identities and chemical characteristics of amino acids that might have an impact on *N*-glycosylation processing machinery.

4.5 Discussion & Conclusion

In this chapter, glycoprotein representations were generated by grafting MD-equilibrated *N*-glycan structures onto AlphaFold predictions. The initial prototype versions of the grafting algorithm were successful at demonstrating that AlphaFold prediction outputs are capable of taking into the account amino acid rotamer configurations that would support trivial transplantation of *N*-glycans from donor template models. Building upon the initial success of the prototype, an attempt was made to expand the dataset of glycoprotein representatives as the analysis in Chapter 3 reveals severe under-representation of deposited glycoprotein structures containing *N*-glycans extending beyond the $Man_3GlcNAc_2$ core in wwPDB. The expansion efforts of accessing unseen non-redundant glycoprotein representatives were proven to be mostly successful through the grafting approach. Unfortunately, there was a notable minority of glycoprotein structures generated through the grafting algorithm that were rendered unusable due to resulting clashes. Nevertheless, all of the computational grafting experiments demonstrate that the extent of *N*-glycan processing is largely driven by the enrichment ratio of the aromatic amino acids in the vicinity of terminal sugars of grafted *N*-glycans. This result is in significant agreement with the analysis presented in Chapter 3 of the thesis. The grafting experiments were also successful at replicating the pattern in terms of polar uncharged amino acids being associated with less-processed *N*-glycans, as demonstrated in Man9 grafting and a2g2 versus a3g3 experiments, most notably in the immediate vicinity of terminal sugars. Finally, the investigation into further levels of processing, *i.e.*, a2g2 versus a3g3 experiments, appear to indicate notable levels of discrimination in terms of positive and negative amino acid types. The results suggest that positive amino acids are promoting further processing of complex *N*-glycans, unlike negative amino acids, at least when immediate vicinities are considered. Moreover, the occurrence of polar uncharged and aromatic amino acids being associated with the degree of processing is also notable in this experiment. Therefore, based on these results, a postulation can be considered that the degree of N-glycan processing is dependent on the predominant amino

acid environment surrounding the glycosylation site. Further studies, especially *in vivo*, are desired to validate this hypothesis and understand the mechanistic insights driving these observations.

Although the representative glycoproteins containing *N*-glycans were tripled thanks to the grafting algorithm in comparison to glycoprotein representatives extracted from wwPDB, the methodology does somewhat demonstrate lack of glycoproteomic data being publicly available. Ideally, the glycoproteomic data would be easily accessible through UniProt entry annotations or at least GlyConnect containing more specific *N*-glycan associations with specific glycosylation sites on specific glycoproteins. Unfortunately the methodology employed in this chapter to retrieve data had to rely on optimising data extraction by associating compositions of monosaccharide units with specific *N*-glycan types. Therefore, it is unclear whether the chosen representative *N*-glycans truly reflect the actual glycoprotein products that emerge from the synthesis pathway. Nevertheless, this nuance may be of minimal consequence as the grafted *N*-glycans served predominantly as spatial probe points within a three-dimensional Cartesian space, with the biosynthetic outcomes being abstracted to specific *N*-glycan types.

Some of the extracted representatives from relevant databases were lost due to failure to produce grafted structures without any clashes, even after exhausting multiple *N*-glycan conformational representatives. Therefore, this issue demonstrates the need for improvements in future versions of the grafting algorithm. After multiple rounds of qualitative analysis, it is apparent that the grafting algorithm implementation in Privateer could be improved by engineering a solution that would manipulate multiple torsion angles between sugar-sugar linkages, rather than being limited to manipulating *N*-glycosidic linkage between ASN residue and the first sugar of the *N*-glycan. However, further engineering of the grafting algorithm implementation could lead to losing the essence of using MD-equilibrated for gains in computational speeds that enable it to carry out grafting at scale. Nevertheless, the experiments could be considered a success in demonstrating that crude grafting of *N*-glycan templates enables significant expansion of representatives to investigate *N*-glycan processing. In essence, the results presented in this chapter allude towards the idea that *N*-glycan modelling on glycoprotein structures could be achieved without having to rely on experimental density maps and that potential remediation of *N*-glycan containing glycoproteins in PDB could rely upon this approach, especially when taking into the account the success of AlphaFill[114].

While the results appear to be appealing, it is important to remember that structures obtained from AlphaFold, despite their groundbreaking capabilities, generate predictions rather than empirically-resolved structures. AlphaFold's predictions, though often of high accuracy, are algorithmically inferred from available data[206,218]. Therefore, yet again, the importance of additional *in vivo* experiments to this research cannot be understated.

Finally, the engineering solutions that were developed as part of this thesis to graft *N*-glycan structures onto glycoproteins and compute neighbouring amino acids given specific probe points, could be re-used in other approaches such as machine learning engineering to train models that could predict *N*-glycan types based on the profiles of neighbouring amino acids that surround the *N*-glycosylation site.

Conclusions and Future Work

The initial aim of this PhD was to investigate ways of harnessing existing structural and glycomics data towards the development of a method that is able to predict what *N*-glycans to model given an input protein structure. The feasibility of such implementation was supported by glycoproteomics projects able to provide necessary data to elucidate at which specific amino acid residue an *N*-glycan attaches in protein sequence and what the composition of the attached *N*-glycan is. Unfortunately, at the time the project was formulated, the public databases used to deposit and store data from various glycoproteomics projects were independent to the extent that there were significant challenges involved in cross-referencing data between multiple sources, particularly due to different notations used to describe *N*-glycan compositions. Therefore, the initial stage of this PhD was to develop a software solution that would enable inter-conversion between different notations used to describe *N*-glycan compositions. Coincidentally and thankfully, at a similar time when the PhD project commenced, this initial hurdle had pretty much been taken care of thanks to the efforts of the GlyCosmos project. The GlyCosmos project developed the WURCS standard notation, the GlyTouCan repository and the GlycanFormatConverter to convert WURCS notation into other notations used to describe *N*-glycan compositions. All of these developments provided the necessary infrastructure to enable Privateer to query databases such as GlyConnect with relative ease, as demonstrated in Chapter 2 of the thesis. Enabling Privateer to retrieve data from an offline mirror of GlyConnect (for faster access, especially for frequent, repetitive queries) and similar databases allows users to quickly cross-reference *N*-glycan composition in an input structure file with publicly-available data describing particular *N*-glycosylation compositions. This is particularly useful during iterative model building, where potentially minor adjustments in individual atom coordinates of target *N*-glycans can significantly alter the chemical interpretation, which can therefore result in inconsistencies with the known *N*-glycan biosynthesis pathways for specific expression systems.

Chapter 3 was initiated as an effort to extract glycoprotein data from wwPDB to generate a training data set for potential machine learning implementations to predict what *N*-glycans should be modelled given an input structure. The extracted data from wwPDB demonstrate that there is an insufficient amount of non-redundant glycoprotein models containing complete *N*-glycan compositions that could be used to train machine learning models. Therefore, faced with this challenge, the project pivoted towards analysing the neighbourhood contexts of terminal sugars in terms of amino acid residues to determine

whether a relationship between amino acid identity and *N*-glycan composition could be established. Based on a limited number of non-redundant glycosylation site representatives, a relationship was indeed established, especially when considering *in vivo* experiments in regard to aromatic amino acids promoting *N*-glycosylation homogeneity carried out by other groups.

During the research, the public release of AlphaFold, following its noteworthy performance at CASP14, garnered significant attention in the Structural Biology community[219]. The impact of AlphaFold cannot be understated for this PhD research as well. The released predictive tool enabled a solution to potentially overcome a severe shortage of non-redundant glycoprotein samples in the analysis of *N*-glycosylation termini structural contexts. Therefore, Chapter 4 was a crude attempt to expand significantly the number of glycoprotein representatives *via* the implementation of a grafting algorithm in Privateer. The implementation enables Privateer to transplant atomic coordinates associated with a particular *N*-glycan from one structure file onto an acceptor protein backbone described in another structure file by establishing a theoretical *N*-glycosidic linkage between specific amino acid residues and the first sugar of donor *N*-glycan. Even though numerous molecule transplantation algorithms implementations exist, few tools specialised for glycosylation transplantation were available. The very few tools that are concerned with transplanting *N*-glycans are manual to a significant extent or have scalability concerns due to utilisation of complex biochemical system molecular dynamic (MD) simulations. Therefore, through collaboration with Dr Elisa Fadda group, a grafting algorithm was implemented in the Privateer software. The development of the grafting algorithm enabled for significant expansion of non-redundant glycoprotein representatives thanks to the utilisation of AlphaFold predictions. Coincidentally, further development of the current grafting algorithm implementation in Privateer could be used for *N*-glycan remediation in projects such as PDB-REDO without having to rely on experimental density as more signal is unlikely to be extracted from experimental data associated with glycan regions in deposited structures to wwPDB[12]. With a significantly expanded dataset of non-redundant representatives, through a variation of different computational experiments, the relationship between amino acid identity and *N*-glycan type was further strengthened, particularly in terms of aromatic amino acids. Therefore, the associations discovered in Chapters 3 and 4 are likely substantial enough to justify further mutagenic *in vivo* experiments to determine whether *N*-glycosylation homogeneity could be promoted by modifying target amino acid residues located in the vicinity of terminal sugars.

After an arguably successful implementation to significantly expand the dataset containing non-redundant glycoprotein models, in the long term the research could be continued in the

machine learning direction. It is likely that with the significant expansion of non-redundant data a machine learning model could be trained to predict *N*-glycan compositions given an input structure. Nevertheless, before actual machine learning implementations are considered, some potential areas of improvement also deserve attention. Particularly, as demonstrated in Chapter 4, the implementation of the grafting algorithm could be improved. There are numerous cases where the highly scalable grafting algorithm fails to produce a good representative model due to the resulting clashes and atomic overlaps between the grafted *N*-glycan and protein backbone. Additional work carried out in the group has extracted glycosidic linkage torsional preferences between pairs of monosaccharides based on high resolution X-ray crystallography data[220]. Therefore, a potential avenue in terms of improving the grafting algorithm implementation in Privateer could be to model *N*-glycan trees on a monosaccharide-by-monosaccharide basis, rather than transplanting whole *N*-glycans as different conformational representatives from MD-equilibrated libraries until a suitable match is found that produces no atomic overlaps. After the grafting algorithm implementation in Privateer is in a more robust state, model processing procedures could be implemented where *N*-glycans are re-modelled with a high degree of confidence. Therefore, this would most likely enable generation of non-redundant datasets in sufficient quantities to successfully train a model with highly capable predictive features.

A potential prototype for a machine learning pipeline to predict what *N*-glycans to model given an input protein structure could be implemented based on a Graph Neural Network (GNN) architecture. Such choice of the architecture is informed by the fact that atomic models are natural representations of a graph, where every atom is a node, and every linkage is an edge. In addition, the algorithms used to detect neighbouring amino acids around modelled *N*-glycans could also be reused to flag more relevant amino acid residues in terms of their vicinity near a potential *N*-glycan. Therefore in the training pipeline, for every representative, Privateer could be used to graft a representative *N*-glycan structure to assign flags to specific amino acid residues that are likely to be involved in modulating *N*-glycan processing. Then, using tools such as Graphein that can streamline conversion of structure files into graph representations required for deep learning libraries, a graph representation could be provided with necessary attributes of flagged relevant amino acid residues into the network during the training process[221]. At prediction stage on unseen data, to predict what *N*-glycan to model given an input protein structure, a given structure file would undergo grafting at designated glycosylation site to generate a graph representation with flagged amino acid residues, which in turn would ideally return probabilities of what kind of *N*-glycan a particular glycosylation site is most likely biased towards. It is unlikely that such an approach would be able to provide a definite *N*-glycan structure prediction, therefore the prediction outputs

should most likely be coarse-grained by predicting key *N*-glycosylation processing events rather than specific structures. Therefore, the prediction pipeline is most likely to be composed of multiple neural network models, trained to predict specific events as summarised in Table 5.1.

Table 5.1: Summary of a potentially identical neural network architecture trained for different tasks to obtain in-silico predictions of *N*-glycan processing based on an input protein structure.

N-Glycosylation event	N-glycan transplant structure	Purpose
High-mannose conversion into processed *N*-glycans	Man9	Determine if processing ceases at High-mannose stage
Number of terminal GlcNAc additions	a2g2 or $GlcNAc_2Man_3GlcNAc_1$	Determine the number of antennae
Core fucosylation	Fucosylated fragment	Determine the status of core fucosylation
Galactosylation	Galactosylated complex *N*-glycan	Determine the status of galactosylation
Sialylation	Sialylated complex *N*-glycan	Determine the status of sialylation

While the analysis presented in the thesis primarily focused on the conversion of high-mannose structures into processed *N*-glycans due to more data being available, the medical and therapeutic implications of *N*-glycan processing extend beyond just this transition. It Is imperative to delve deeper into the nuances of complex *N*-glycan modifications, especially when considering therapeutic proteins and potential implications for immunogenicity. More complex *N*-glycans, in their mature forms, may carry additional sugar moieties such as fucose and galactose. These sugars are not merely decorative; they can have profound effects on how the immune system perceives and interacts with the glycoprotein in question. For example, a reduction in core fucosylation of antibodies can significantly enhance their antibody-dependent cell-mediated cytotoxicity (ADCC) potential, a vital mechanism through which therapeutic antibodies exert anticancer effects[222]. In the example of galactosylation,

alterations in levels of galactosylation of specific glycoproteins has been correlated with increased inflammatory responses due to the recognition of non-galactosylated glycoforms by specific antibodies[223]. Therefore, from a therapeutic standpoint, understanding and predicting the degree of processing within complex *N*-glycans is not just a matter of academic interest. It holds the key to developing more effective and safer biopharmaceuticals. With the surge in the development and use of therapeutic proteins, particularly monoclonal antibodies, it is essential to have a precise knowledge of their glycosylation status. This knowledge aids in fine-tuning their efficacy, pharmacokinetics, and immunogenicity[224]. Therefore, the uncovered clues about aromatic and polar uncharged amino acid impact to the degree of complex *N*-glycan processing presented in Chapter 4, could provide a good basis for *in vivo* experiments to test the emerging hypotheses for specific glycoproteins.

An area that remains unaddressed in this thesis is *N*-glycan processing being highly dependent on the configuration of cellular expression systems, in terms of environment temperature, pH, availability of donor sugars and enzymes executing *N*-glycan processing[225,226]. Other areas of glycobiology research are investigating whether it is possible to predict determinants of changes in *N*-glycan processing based on alterations to cellular states directly affecting the action of *N*-glycan processing enzymes[227,228]. Perhaps it is likely that efforts to address glycosylation site amino acid residue neighbour influence on *N*-glycan processing, could be combined with such work to obtain a complete model and understanding of why specific *N*-glycans of specific glycoproteins are processed to different extents, thereby creating the properties of *N*-glycosylation microheterogeneity.

While this thesis has laid a speculative foundation and offered potentially meaningful insights, it is but a stepping stone in the complex world of protein glycosylation. It is the hope that the methodologies, findings, and hypotheses presented herein serve as a catalyst for further research in the field of Structural Glycobiology.

References

1. Matsubara, M., Aoki-Kinoshita, K. F., Aoki, N. P., Yamada, I. & Narimatsu, H. WURCS 2.0 Update To Encapsulate Ambiguous Carbohydrate Structures. *J. Chem. Inf. Model.* **57**, 632–637 (2017).

2. Bagdonas, H., Ungar, D. & Agirre, J. Leveraging glycomics data in glycoprotein 3D structure validation with Privateer. *Beilstein J. Org. Chem.* **16**, 2523–2533 (2020).

3. Agirre, J. *et al.* Three-dimensional structures of two heavily N-glycosylated Aspergillus sp. family GH3 β-D-glucosidases. *Acta Crystallogr D Struct Biol* **72**, 254–265 (2016).

4. Pancera, M. *et al.* Structural basis for diverse N-glycan recognition by HIV-1-neutralizing V1-V2-directed antibody PG16. *Nat. Struct. Mol. Biol.* **20**, 804–813 (2013).

5. Kolstoe, S. E. *et al.* Interaction of serum amyloid P component with hexanoyl bis(D-proline) (CPHPC). *Acta Crystallogr. D Biol. Crystallogr.* **70**, 2232–2240 (2014).

6. Janssen, B. J. C. *et al.* Structural basis of semaphorin-plexin signalling. *Nature* **467**, 1118–1122 (2010).

7. Polyakov, K. M., Gavryushov, S., Fedorova, T. V., Glazunova, O. A. & Popov, A. N. The subatomic resolution study of laccase inhibition by chloride and fluoride anions using single-crystal serial crystallography: insights into the enzymatic reaction mechanism. *Acta Crystallogr D Struct Biol* **75**, 804–816 (2019).

8. Dai, Y. N., Fremont, D. H. & Center for Structural Genomics of Infectious Diseases (CSGID). Crystal structure of hemagglutinin from influenza virus A/Pennsylvania/14/2010 (H3N2). Preprint at https://doi.org/10.2210/pdb6mzk/pdb (2019).

9. Lee, P. S. *et al.* Receptor mimicry by antibody F045-092 facilitates universal binding to the H3 subtype of influenza virus. *Nat. Commun.* **5**, 3614 (2014).

10. Szakonyi, G. *et al.* Structure of the Epstein-Barr virus major envelope glycoprotein. *Nat. Struct. Mol. Biol.* **13**, 996–1001 (2006).

11. Kim, H. M. *et al.* Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist Eritoran. *Cell* **130**, 906–917 (2007).

12. van Beusekom, B., Lütteke, T. & Joosten, R. P. Making glycoproteins a little bit sweeter with PDB-REDO. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **74**, 463–472 (2018).

13. Agirre, J. *et al.* The CCP4 suite: integrative software for macromolecular crystallography. *Acta Crystallogr D Struct Biol* **79**, 449–461 (2023).

14. Burnley, T., Palmer, C. M. & Winn, M. Recent developments in the CCP-EM software suite. *Acta Crystallogr D Struct Biol* **73**, 469–477 (2017).

15. Alocci, D. *et al.* GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. *J. Proteome Res.* **18**, 664–677 (2019).

16. Bai, X.-C. *et al.* An atomic structure of human γ-secretase. *Nature* **525**, 212–217 (2015).

17. Zwolak, A. *et al.* Modulation of protein A binding allows single-step purification of mouse bispecific antibodies that retain FcRn binding. *MAbs* **9**, 1306–1316 (2017).

18. wwPDB: 1V0Z. https://doi.org/10.2210/pdb1V0Z/pdb.

19. Trastoy, B. *et al.* Structural basis of mammalian high-mannose N-glycan processing by human gut Bacteroides. *Nat. Commun.* **11**, 899 (2020).

20. Dengl, S. *et al.* Format chain exchange (FORCE) for high-throughput generation of bispecific antibodies in combinatorial binder-format matrices. *Nat. Commun.* **11**, 1–11 (2020).

21. An engineered Fc variant of an IgG eliminates all immune effector functions via structural perturbations. *Methods* **65**, 114–126 (2014).

22. wwPDB: 6D58. https://doi.org/10.2210/pdb6d58/pdb.

23. Tam, S. H. *et al.* Functional, Biophysical, and Structural Characterization of Human IgG1 and IgG4 Fc Variants with Ablated Immune Functionality. *Antibodies* **6**, 12 (2017).

24. Agirre, J. *et al.* Privateer: software for the conformational validation of

carbohydrate structures. *Nat. Struct. Mol. Biol.* **22**, 833–834 (2015).

25. Fogarty, C. A. & Fadda, E. The oligomannose N-glycans 3D architecture and its response to the FcγRIIIa structural landscape. Preprint at https://doi.org/10.1101/2021.01.11.426234.

26. *Essentials of Glycobiology*. (Cold Spring Harbor Laboratory Press, 2016).

27. Varki, A. Biological roles of oligosaccharides: all of the theories are correct. *Glycobiology* **3**, 97–130 (1993).

28. Hiyama, G. *et al.* Changes in post-translational modifications of prolactin during development and reproductive cycles in the chicken. *Gen. Comp. Endocrinol.* **161**, 238–245 (2009).

29. Moran, A. P., Gupta, A. & Joshi, L. Sweet-talk: role of host glycosylation in bacterial pathogenesis of the gastrointestinal tract. *Gut* **60**, 1412–1425 (2011).

30. Chang, I. J., He, M. & Lam, C. T. Congenital disorders of glycosylation. *Ann Transl Med* **6**, 477 (2018).

31. Shental-Bechor, D. & Levy, Y. Effect of glycosylation on protein folding: a close look at thermodynamic stabilization. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 8256–8261 (2008).

32. Ishino, T. *et al.* A protein engineering approach differentiates the functional importance of carbohydrate moieties of interleukin-5 receptor α. *Biochemistry* **50**, 7546–7556 (2011).

33. Shrimal, S., Cherepanova, N. A. & Gilmore, R. Cotranslational and posttranslocational N-glycosylation of proteins in the endoplasmic reticulum. *Semin. Cell Dev. Biol.* **41**, 71–78 (2015).

34. Ulrich, P. & Cerami, A. Protein glycation, diabetes, and aging. *Recent Prog. Horm. Res.* **56**, 1–21 (2001).

35. Glenn, J. V. & Stitt, A. W. The role of advanced glycation end products in retinal ageing and disease. *Biochim. Biophys. Acta* **1790**, 1109–1116 (2009).

36. Omsland, T. K., Bangstad, H.-J., Berg, T. J. & Kolset, S. O. [Advanced

glycation end products and hyperglycaemia]. *Tidsskr. Nor. Laegeforen.* **126**, 155–158 (2006).

37. Freeze, H. H., Boyce, M., Zachara, N. E., Hart, G. W. & Schnaar, R. L. Glycosylation Precursors. in *Essentials of Glycobiology* (eds. Varki, A. et al.) (Cold Spring Harbor Laboratory Press).

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

39. Shallenberger, R. S., Acree, T. E. & Lee, C. Y. Sweet taste of D and L-sugars and amino-acids and the steric nature of their chemo-receptor site. *Nature* **221**, 555–556 (1969).

40. Ono, K., Takigawa, S. & Yamada, K. L-Glucose: Another Path to Cancer Cells. *Cancers* **12**, (2020).

41. Shimizu, T., Takaya, N. & Nakamura, A. An L-glucose catabolic pathway in Paracoccus species 43P. *J. Biol. Chem.* **287**, 40448–40456 (2012).

42. Seeberger, P. H. Monosaccharide Diversity. in *Essentials of Glycobiology* (eds. Varki, A. et al.) (Cold Spring Harbor Laboratory Press).

43. Rini, J. M., Moremen, K. W., Davis, B. G. & Esko, J. D. Glycosyltransferases and Glycan-Processing Enzymes. in *Essentials of Glycobiology* (eds. Varki, A. et al.) (Cold Spring Harbor Laboratory Press).

44. Flynn, R. A. *et al.* Small RNAs are modified with N-glycans and displayed on the surface of living cells. *Cell* **184**, 3109–3124.e22 (2021).

45. Grant, C. W. Model membranes bearing glycolipids and glycoproteins. *Chem. Phys. Lipids* **40**, 285–302 (1986).

46. Crine, S. L. & Acharya, K. R. Molecular basis of C-mannosylation - a structural perspective. *FEBS J.* **289**, 7670–7687 (2022).

47. McNaught, A. D. Nomenclature of carbohydrates (recommendations 1996). *Adv. Carbohydr. Chem. Biochem.* **52**, 43–177 (1997).

48. Meuris, L. *et al.* GlycoDelete engineering of mammalian cells simplifies N-

glycosylation of recombinant proteins. *Nat. Biotechnol.* **32**, 485–489 (2014).

49. Gupta, R., Leon, F., Rauth, S., Batra, S. K. & Ponnusamy, M. P. A Systematic Review on the Implications of O-linked Glycan Branching and Truncating Enzymes on Cancer Progression and Metastasis. *Cells* **9**, (2020).

50. Elbein, A. D. Glycosidase inhibitors: inhibitors of N-linked oligosaccharide processing. *FASEB J.* **5**, 3055–3063 (1991).

51. Kornfeld, R. & Kornfeld, S. Assembly of asparagine-linked oligosaccharides. *Annu. Rev. Biochem.* **54**, 631–664 (1985).

52. Varki, A. & Chrispeels, M. J. *Essentials of Glycobiology*. (CSHL Press, 1999).

53. Wang, N., Li, S.-T., Lu, T.-T., Nakanishi, H. & Gao, X.-D. Approaches towards the core pentasaccharide in N- linked glycans. *Chin. Chem. Lett.* **29**, 35–39 (2018).

54. Hunt, L. T. & Dayhoff, M. O. The occurrence in proteins of the tripeptides Asn-X-Ser and Asn-X-Thr and of bound carbohydrate. *Biochem. Biophys. Res. Commun.* **39**, 757–765 (1970).

55. Hubbard, S. C. & Ivatt, R. J. Synthesis and processing of asparagine-linked oligosaccharides. *Annu. Rev. Biochem.* **50**, 555–583 (1981).

56. Lehrman, M. A. Biosynthesis of N-acetylglucosamine-P-P-dolichol, the committed step of asparagine-linked oligosaccharide assembly. *Glycobiology* **1**, 553–562 (1991).

57. Schwarz, F. & Aebi, M. Mechanisms and principles of N-linked protein glycosylation. *Curr. Opin. Struct. Biol.* **21**, 576–582 (2011).

58. Absmanner, B., Schmeiser, V., Kämpf, M. & Lehle, L. Biochemical characterization, membrane association and identification of amino acids essential for the function of Alg11 from Saccharomyces cerevisiae, an alpha1,2-mannosyltransferase catalysing two sequential glycosylation steps in the formation of the lipid-linked core oligosaccharide. *Biochem. J* **426**, 205–217 (2010).

59. Sanyal, S. & Menon, A. K. Specific transbilayer translocation of dolichol-linked oligosaccharides by an endoplasmic reticulum flippase. *Proc. Natl. Acad. Sci. U. S. A.*

106, 767–772 (2009).

60. Rush, J. S. Role of Flippases in Protein Glycosylation in the Endoplasmic Reticulum. *Lipid Insights* **8**, 45–53 (2015).

61. Mandon, E. C., Trueman, S. F. & Gilmore, R. Protein translocation across the rough endoplasmic reticulum. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).

62. Kelleher, D. J. & Gilmore, R. An evolving view of the eukaryotic oligosaccharyltransferase. *Glycobiology* **16**, 47R–62R (2006).

63. Cherepanova, N. A., Venev, S. V., Leszyk, J. D., Shaffer, S. A. & Gilmore, R. Quantitative glycoproteomics reveals new classes of STT3A- and STT3B-dependent N-glycosylation sites. *J. Cell Biol.* **218**, 2782–2796 (2019).

64. Høiberg-Nielsen, R., Westh, P., Skov, L. K. & Arleth, L. Interrelationship of steric stabilization and self-crowding of a glycosylated protein. *Biophys. J.* **97**, 1445–1453 (2009).

65. Vembar, S. S. & Brodsky, J. L. One step at a time: endoplasmic reticulum-associated degradation. *Nat. Rev. Mol. Cell Biol.* **9**, 944–957 (2008).

66. Jakob, C. A., Burda, P., Roth, J. & Aebi, M. Degradation of misfolded endoplasmic reticulum glycoproteins in Saccharomyces cerevisiae is determined by a specific oligosaccharide structure. *J. Cell Biol.* **142**, 1223–1233 (1998).

67. Tjondro, H. C., Loke, I., Chatterjee, S. & Thaysen-Andersen, M. Human protein paucimannosylation: cues from the eukaryotic kingdoms. *Biol. Rev. Camb. Philos. Soc.* **94**, 2068–2100 (2019).

68. Aggarwal, S. K., Kinter, M., Fitzgerald, R. L. & Herold, D. A. Mass spectrometry of trace elements in biological samples. *Crit. Rev. Clin. Lab. Sci.* **31**, 35–87 (1994).

69. Zaia, J. Mass spectrometry and glycomics. *OMICS* **14**, 401–418 (2010).

70. Li, Q., Xie, Y., Wong, M., Barboza, M. & Lebrilla, C. B. Comprehensive structural glycomic characterization of the glycocalyxes of cells and tissues. *Nat. Protoc.* **15**, 2668–2704 (2020).

71. Freeze, H. H. Endoglycosidase and glycoamidase release of N-linked oligosaccharides. *Curr. Protoc. Mol. Biol.* **Chapter 17**, Unit17.13A (2001).

72. Wilkinson, H. & Saldova, R. Current Methods for the Characterization of -Glycans. *J. Proteome Res.* **19**, 3890–3905 (2020).

73. Zhou, S., Dong, X., Veillon, L., Huang, Y. & Mechref, Y. LC-MS/MS analysis of permethylated N-glycans facilitating isomeric characterization. *Anal. Bioanal. Chem.* **409**, 453–466 (2017).

74. Ruhaak, L. R. *et al.* Glycan labeling strategies and their use in identification and quantification. *Anal. Bioanal. Chem.* **397**, 3457–3481 (2010).

75. Fenn, L. S. & McLean, J. A. Structural separations by ion mobility-MS for glycomics and glycoproteomics. *Methods Mol. Biol.* **951**, 171–194 (2013).

76. Blundell, T. L., Johnson, L. & Johnson, L. N. *Protein Crystallography*. (Academic Press, 1976).

77. McCoy, A. J., Sammito, M. D. & Read, R. J. Implications of AlphaFold2 for crystallographic phasing by molecular replacement. *Acta Crystallogr D Struct Biol* **78**, 1–13 (2022).

78. Flower, T. G. & Hurley, J. H. Crystallographic molecular replacement using an in silico-generated search model of SARS-CoV-2 ORF8. *Protein Sci.* **30**, 728–734 (2021).

79. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

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

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

82. Scherzer, O. The Theoretical Resolution Limit of the Electron Microscope. *J. Appl. Phys.* **20**, 20–29 (2004).

83. Franken, L. E., Grünewald, K., Boekema, E. J. & Stuart, M. C. A. A Technical

Introduction to Transmission Electron Microscopy for Soft-Matter: Imaging, Possibilities, Choices, and Technical Developments. *Small* **16**, e1906198 (2020).

84. Passmore, L. A. & Russo, C. J. Specimen Preparation for High-Resolution Cryo-EM. *Methods Enzymol.* **579**, 51–86 (2016).

85. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**, 171–193 (1995).

86. Yu, H. Extending the size limit of protein nuclear magnetic resonance. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 332–334 (1999).

87. Govil, G. Introduction to biological NMR spectroscopy. in *Biomedical Magnetic Resonance: Proceedings of the International Workshop* 12–12 (Jaypee Brothers Medical Publishers (P) Ltd., 2005).

88. Quintana, J. I., Atxabal, U., Unione, L., Ardá, A. & Jiménez-Barbero, J. Exploring multivalent carbohydrate-protein interactions by NMR. *Chem. Soc. Rev.* **52**, 1591–1613 (2023).

89. Gimeno, A., Valverde, P., Ardá, A. & Jiménez-Barbero, J. Glycan structures and their interactions with proteins. A NMR view. *Curr. Opin. Struct. Biol.* **62**, 22–30 (2020).

90. Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).

91. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **3**, 1171–1179 (2008).

92. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 1002–1011 (2006).

93. Cowtan, K. Automated nucleic acid chain tracing in real time. *IUCrJ* **1**, 387–392 (2014).

94. Jamali, K. *et al.* Automated model building and protein identification in cryo-EM maps. (2023) doi:10.1101/2023.05.16.541002.

95. Bond, P. S. & Cowtan, K. D. ModelCraft: an advanced automated model-building pipeline using Buccaneer. *Acta Crystallogr D Struct Biol* **78**, 1090–1098 (2022).

96. Yamashita, K., Wojdyr, M., Long, F., Nicholls, R. A. & Murshudov, G. N. GEMMI and Servalcat restrain REFMAC5. *Acta Crystallogr D Struct Biol* **79**, 368–373 (2023).

97. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).

98. Nicholls, R. A. *et al.* The missing link: covalent linkages in structural models. *Acta Crystallogr D Struct Biol* **77**, 727–745 (2021).

99. Atanasova, M., Nicholls, R. A., Joosten, R. P. & Agirre, J. Updated restraint dictionaries for carbohydrates in the pyranose form. *Acta Crystallogr D Struct Biol* **78**, 455–465 (2022).

100. Beckers, M., Mann, D. & Sachse, C. Structural interpretation of cryo-EM image reconstructions. *Prog. Biophys. Mol. Biol.* **160**, 26–36 (2021).

101. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).

102. Sobolev, O. V. *et al.* A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry. *Structure* **28**, 1249–1258.e2 (2020).

103. Tickle, I. J. Statistical quality indicators for electron-density maps. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 454–467 (2012).

104. Cragnolini, T. *et al.* TEMPy2: a Python library with improved 3D electron microscopy density-fitting and validation workflows. *Acta Crystallogr D Struct Biol* **77**, 41–47 (2021).

105. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).

106. Atanasova, M., Agirre, J. & IUCr. SAILS: automated model building of

carbohydrates. *Acta Crystallographica Section A: Foundations and Advances* **77**, C775–C775 (2021).

107. Emsley, P. & Crispin, M. Structural analysis of glycoproteins: building N-linked glycans with Coot. *Acta Crystallogr D Struct Biol* **74**, 256–263 (2018).

108. Stsiapanava, A. *et al.* Structure of the decoy module of human glycoprotein 2 and uromodulin and its interaction with bacterial adhesin FimH. *Nat. Struct. Mol. Biol.* **29**, 190–193 (2022).

109. EMDB. Electron Microscopy Data Bank. *Electron Microscopy Data Bank* https://www.ebi.ac.uk/emdb/EMD-13378.

110. Atanasova, M., Bagdonas, H. & Agirre, J. Structural glycobiology in the age of electron cryo-microscopy. *Curr. Opin. Struct. Biol.* **62**, 70–78 (2019).

111. An, H. J., Froehlich, J. W. & Lebrilla, C. B. Determination of glycosylation sites and site-specific heterogeneity in glycoproteins. *Curr. Opin. Chem. Biol.* **13**, 421–426 (2009).

112. Agirre, J. Strategies for carbohydrate model building, refinement and validation. *Acta Crystallogr D Struct Biol* **73**, 171–186 (2017).

113. Simpkin, A. J., Sánchez Rodríguez, F., Mesdaghi, S., Kryshtafovych, A. & Rigden, D. J. Evaluation of model refinement in CASP14. *Proteins* **89**, 1852–1869 (2021).

114. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023).

115. Gupta, R. & Brunak, S. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symp. Biocomput.* 310–322 (2002).

116. Li, F. *et al.* GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics* **31**, 1411–1419 (2015).

117. Rohne, P., Prochnow, H., Wolf, S., Renner, B. & Koch-Brandt, C. The chaperone activity of clusterin is dependent on glycosylation and redox environment. *Cell. Physiol. Biochem.* **34**, 1626–1639 (2014).

118. Wyss, D. F. *et al.* Conformation and function of the N-linked glycan in the adhesion domain of human CD2. *Science* **269**, 1273–1278 (1995).

119. Mitra, N., Sharon, N. & Surolia, A. Role of N-linked glycan in the unfolding pathway of Erythrina corallodendron lectin. *Biochemistry* **42**, 12208–12216 (2003).

120. Gu, J. *et al.* Potential roles of N-glycosylation in cell adhesion. *Glycoconj. J.* **29**, 599–607 (2012).

121. Lyons, J. J., Milner, J. D. & Rosenzweig, S. D. Glycans Instructing Immunity: The Emerging Role of Altered Glycosylation in Clinical Immunology. *Front Pediatr* **3**, 54 (2015).

122. Boscher, C., Dennis, J. W. & Nabi, I. R. Glycosylation, galectins and cellular signaling. *Current Opinion in Cell Biology* vol. 23 383–392 Preprint at https://doi.org/10.1016/j.ceb.2011.05.001 (2011).

123. Russell, R. J. *et al.* Structure of influenza hemagglutinin in complex with an inhibitor of membrane fusion. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 17736–17741 (2008).

124. Crispin, M., Ward, A. B. & Wilson, I. A. Structure and Immune Recognition of the HIV Glycan Shield. *Annu. Rev. Biophys.* **47**, 499–523 (2018).

125. Watanabe, Y. *et al.* Structure of the Lassa virus glycan shield provides a model for immunological resistance. *Proceedings of the National Academy of Sciences* vol. 115 7320–7325 Preprint at https://doi.org/10.1073/pnas.1803990115 (2018).

126. Pinger, J. *et al.* African trypanosomes evade immune clearance by O-glycosylation of the VSG surface coat. *Nat Microbiol* **3**, 932–938 (2018).

127. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281–292.e6 (2020).

128. Wood, N. T. *et al.* The influence of N-linked glycans on the molecular dynamics of the HIV-1 gp120 V3 loop. *PLoS One* **8**, e80301 (2013).

129. Lütteke, T. & von der Lieth, C. W. Data mining the PDB for glyco-related data. *Methods Mol. Biol.* **534**, 293–310 (2009).

130. Crispin, M., Stuart, D. I. & Jones, E. Y. Building meaningful models of glycoproteins. *Nature structural & molecular biology* vol. 14 354; discussion 354–5 (2007).

131. Agirre, J., Davies, G. J., Wilson, K. S. & Cowtan, K. D. Carbohydrate structure: the rocky road to automation. *Curr. Opin. Struct. Biol.* **44**, 39–47 (2017).

132. Frank, M., Lutteke, T. & von der Lieth, C.-W. GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Research* vol. 35 287–290 Preprint at https://doi.org/10.1093/nar/gkl907 (2007).

133. Agirre, J., Davies, G., Wilson, K. & Cowtan, K. Carbohydrate anomalies in the PDB. *Nat. Chem. Biol.* **11**, 303 (2015).

134. Rudd, P. M. & Dwek, R. A. Glycosylation: heterogeneity and the 3D structure of proteins. *Crit. Rev. Biochem. Mol. Biol.* **32**, 1–100 (1997).

135. Fisher, P., Thomas-Oates, J., Jamie Wood, A. & Ungar, D. The N-Glycosylation Processing Potential of the Mammalian Golgi Apparatus. *Frontiers in Cell and Developmental Biology* vol. 7 Preprint at https://doi.org/10.3389/fcell.2019.00157 (2019).

136. Geerlof, A. *et al.* The impact of protein characterization in structural proteomics. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 1125–1136 (2006).

137. Stura, E. A., Nemerow, G. R. & Wilson, I. A. Strategies in the crystallization of glycoproteins and protein complexes. *Journal of Crystal Growth* vol. 122 273–285 Preprint at https://doi.org/10.1016/0022-0248(92)90256-i (1992).

138. Cheng, Y., Grigorieff, N., Penczek, P. A. & Walz, T. A primer to single-particle cryo-electron microscopy. *Cell* **161**, 438–449 (2015).

139. Serna, M. Hands on Methods for High Resolution Cryo-Electron Microscopy Structures of Heterogeneous Macromolecular Complexes. *Front Mol Biosci* **6**, 33 (2019).

140. Fan, X. *et al.* Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Angstrom resolution. *Nat. Commun.* **10**, 2386 (2019).

141. Herzik, M. A., Jr, Wu, M. & Lander, G. C. High-resolution structure determination of sub-100 kDa complexes using conventional cryo-EM. *Nat. Commun.* **10**, 1032 (2019).

142. Wang, H.-W. & Wang, J.-W. How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Sci.* **26**, 32–39 (2017).

143. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr D Struct Biol* **74**, 519–530 (2018).

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

145. Lütteke, T. & von der Lieth, C.-W. pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics* **5**, 69 (2004).

146. Potterton, L. *et al.* CCP4i2: the new graphical user interface to the CCP4 program suite. *Acta Crystallogr D Struct Biol* **74**, 68–84 (2018).

147. Gristick, H. B., Wang, H. & Bjorkman, P. J. X-ray and EM structures of a natively glycosylated HIV-1 envelope trimer. *Acta Crystallographica Section D Structural Biology* vol. 73 822–828 Preprint at https://doi.org/10.1107/s2059798317013353 (2017).

148. Joosten, R. P. & Lütteke, T. Carbohydrate 3D structure validation. *Curr. Opin. Struct. Biol.* **44**, 9–17 (2017).

149. Nakahara, Y. *et al.* Amino acid sequence and carbohydrate structure of a recombinant human tissue factor pathway inhibitor expressed in Chinese hamster ovary cells: one N-and two O-linked carbohydrate chains are located between Kunitz domains 2 and 3 and one N-linked carbohydrate chain is in Kunitz domain 2. *Biochemistry* **35**, 6450–6459 (1996).

150. Shajahan, A., Heiss, C., Ishihara, M. & Azadi, P. Glycomic and glycoproteomic analysis of glycoproteins—a tutorial. *Analytical and Bioanalytical*

Chemistry vol. 409 4483–4505 Preprint at https://doi.org/10.1007/s00216-017-0406-7 (2017).

151. Liu, H. *et al.* Mass spectrometry-based analysis of glycoproteins and its clinical applications in cancer biomarker discovery. *Clin. Proteomics* **11**, 14 (2014).

152. Hofmann, J. & Pagel, K. Glycan Analysis by Ion Mobility-Mass Spectrometry. *Angewandte Chemie International Edition* vol. 56 8342–8349 Preprint at https://doi.org/10.1002/anie.201701309 (2017).

153. Leymarie, N. & Zaia, J. Effective use of mass spectrometry for glycan and glycopeptide structural analysis. *Anal. Chem.* **84**, 3040–3048 (2012).

154. Ceroni, A. *et al.* GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J. Proteome Res.* **7**, 1650–1659 (2008).

155. Albersheim, P. CarbBank: A structural and bibliographic data base. Preprint at https://doi.org/10.2172/5715461 (1989).

156. von der Lieth, C.-W. *et al.* EUROCarbDB: An open-access platform for glycoinformatics. *Glycobiology* **21**, 493–502 (2011).

157. Ranzinger, R., Herget, S., Wetter, T. & von der Lieth, C.-W. GlycomeDB - integration of open-access carbohydrate structure databases. *BMC Bioinformatics* **9**, 384 (2008).

158. Herget, S., Ranzinger, R., Maass, K. & Lieth, C.-W. V. D. GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr. Res.* **343**, 2162–2171 (2008).

159. Tiemeyer, M. *et al.* GlyTouCan: an accessible glycan structure repository. *Glycobiology* **27**, 915–919 (2017).

160. Fujita, A. *et al.* The international glycan repository GlyTouCan version 3.0. *Nucleic Acids Research* **49**(D1), D1529–D1533 https://doi.org/10.1093/nar/gkaa947 (2021).

161. Tsuchiya, S., Yamada, I. & Aoki-Kinoshita, K. F. GlycanFormatConverter: a conversion tool for translating the complexities of glycans. *Bioinformatics* **35**, 2434–2440 (2019).

162. Hayes, C. A. *et al.* UniCarb-DB: a database resource for glycomic discovery. *Bioinformatics* **27**, 1343–1344 (2011).

163. Congreve, M., Murray, C. W. & Blundell, T. L. Keynote review: Structural biology and drug discovery. *Drug Discovery Today* vol. 10 895–907 Preprint at https://doi.org/10.1016/s1359-6446(05)03484-7 (2005).

164. Jong, D. de, de Jong, D., Periole, X. & Marrink, S. J. Towards Molecular Dynamics Simulations of Large Protein Complexes. *Biophysical Journal* vol. 98 57a Preprint at https://doi.org/10.1016/j.bpj.2009.12.329 (2010).

165. Aloy, P. & Russell, R. B. Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 5896–5901 (2002).

166. Tanaka, K. *et al.* WURCS: the Web3 unique representation of carbohydrate structures. *J. Chem. Inf. Model.* **54**, 1558–1566 (2014).

167. Westbrook, J. D. *et al.* The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics* **31**, 1274–1278 (2015).

168. Lütteke, T., Frank, M. & von der Lieth, C.-W. Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr. Res.* **339**, 1015–1020 (2004).

169. Lütteke, T., Frank, M. & von der Lieth, C.-W. Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res.* **33**, D242–6 (2005).

170. Campbell, M. P. *et al.* UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res.* **42**, D215–21 (2014).

171. Daponte, V., Hayes, C., Mariethoz, J. & Lisacek, F. Dealing with the Ambiguity of Glycan Substructure Search. *Molecules* **27**, (2021).

172. Goto, Y. *et al.* N-glycosylation is required for secretion and enzymatic activity of human hyaluronidase1. *FEBS Open Bio* **4**, 554–559 (2014).

173. Zhou, Q. & Qiu, H. The Mechanistic Impact of N-Glycosylation on Stability,

Pharmacokinetics, and Immunogenicity of Therapeutic Proteins. *J. Pharm. Sci.* **108**, 1366–1377 (2019).

174. Green, R. S. *et al.* Mammalian N-glycan branching protects against innate immune self-recognition and inflammation in autoimmune disease pathogenesis. *Immunity* **27**, 308–320 (2007).

175. Linders, P. T. A., Peters, E., Ter Beest, M., Lefeber, D. J. & van den Bogaart, G. Sugary Logistics Gone Wrong: Membrane Trafficking and Congenital Disorders of Glycosylation. *Int. J. Mol. Sci.* **21**, (2020).

176. West, B., Wood, A. J. & Ungar, D. Computational Modeling of Glycan Processing in the Golgi for Investigating Changes in the Arrangements of Biosynthetic Enzymes. *Methods Mol. Biol.* **2370**, 209–222 (2022).

177. Hang, I. *et al.* Analysis of site-specific N-glycan remodeling in the endoplasmic reticulum and the Golgi. *Glycobiology* **25**, 1335–1349 (2015).

178. Suga, A., Nagae, M. & Yamaguchi, Y. Analysis of protein landscapes around N-glycosylation sites from the PDB repository for understanding the structural basis of N-glycoprotein processing and maturation. *Glycobiology* **28**, 774–785 (2018).

179. PDB Archive Downloads. https://www.wwpdb.org/ftp/pdb-ftp-sites#rcsbpdb.

180. Feng, Z. *et al.* Enhanced validation of small-molecule ligands and carbohydrates in the Protein Data Bank. *Structure* **29**, 393–400.e1 (2021).

181. WURCS Tools and Databases. *WorkingGroup* https://www.wurcs-wg.org/tools/.

182. wwPDB: https://www.wwpdb.org/documentation/carbohydrate-remediation.

183. Rose, Y. *et al.* RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive. *J. Mol. Biol.* **433**, 166704 (2021).

184. Velankar, S. *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* **41**, D483–9 (2013).

185. Vagin, A. A. *et al.* REFMAC5 dictionary: organization of prior chemical

knowledge and guidelines for its use. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2184–2195 (2004).

186. Krissinel, E. Enhanced fold recognition using efficient short fragment clustering. *J Mol Biochem* **1**, 76–85 (2012).

187. Murray, A. N. *et al.* Enhanced Aromatic Sequons Increase Oligosaccharyltransferase Glycosylation Efficiency and Glycan Homogeneity. *Chem. Biol.* **22**, 1052–1062 (2015).

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

189. Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213–220 (2014).

190. Casalino, L. *et al.* Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent Sci* **6**, 1722–1734 (2020).

191. 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).

192. Jo, S., Lee, H. S., Skolnick, J. & Im, W. Restricted N-glycan conformational space in the PDB and its implication in glycan structure modeling. *PLoS Comput. Biol.* **9**, e1002946 (2013).

193. Lemmin, T. & Soto, C. Glycosylator: a Python framework for the rapid modeling of glycans. *BMC Bioinformatics* **20**, 513 (2019).

194. gmml/internalPrograms/GlycoproteinBuilder at 15f746a8c386342edbfe74d9f9d6588e80c33a22 · GLYCAM-Web/gmml. *GitHub* https://github.com/GLYCAM-Web/gmml/tree/15f746a8c386342edbfe74d9f9d6588e80c33a22/internalPrograms/GlycoproteinBuilder.

195. gmml/internalPrograms/GlycoproteinBuilder at actual · GLYCAM-Web/gmml. *GitHub* https://github.com/GLYCAM-

Web/gmml/tree/actual/internalPrograms/GlycoproteinBuilder.

196. Fogarty, C. A. & Fadda, E. Oligomannose -Glycans 3D Architecture and Its Response to the FcγRIIIa Structural Landscape. *J. Phys. Chem. B* **125**, 2607–2616 (2021).

197. Robert Service. 'The game has changed.' AI triumphs at solving protein structures. *Science* (2020) doi:10.1126/science.abf9367.

198. Keyes, R. W. The impact of moore's law. *IEEE Solid-State Circuits Soc. Newsl.* **11**, 25–27 (2006).

199. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38 (2013).

200. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **20**, 681–697 (2019).

201. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **13**, e1005324 (2017).

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

203. Stevens, A. O. & He, Y. Benchmarking the Accuracy of AlphaFold 2 in Loop Structure Prediction. *Biomolecules* **12**, (2022).

204. AlphaFold 2 is here: what's behind the structure prediction miracle. https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/.

205. Vaswani, A. *et al.* Attention is all you need. (2017) doi:10.48550/ARXIV.1706.03762.

206. Read, R. J., Baker, E. N., Bond, C. S., Garman, E. F. & van Raaij, M. J. AlphaFold and the future of structural biology. *Acta Crystallogr D Struct Biol* **79**, 556–558 (2023).

207. Terwilliger, T. C. *et al.* Accelerating crystal structure determination with iterative AlphaFold prediction. *Acta Crystallogr D Struct Biol* **79**, 234–244 (2023).

208. Wankowicz, S. A., de Oliveira, S. H., Hogan, D. W., van den Bedem, H. & Fraser, J. S. Ligand binding remodels protein side-chain conformational heterogeneity. *Elife* **11**, (2022).

209. Touw, W. G., van Beusekom, B., Evers, J. M. G., Vriend, G. & Joosten, R. P. Validation and correction of Zn-CysHis complexes. *Acta Crystallogr D Struct Biol* **72**, 1110–1118 (2016).

210. Kearsley, S. K. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr. A* **45**, 208–210 (1989).

211. GlyConnect RDF. https://glyconnect.expasy.org/rdf.

212. Bagdonas, H., Fogarty, C. A., Fadda, E. & Agirre, J. The case for post-predictional modifications in the AlphaFold Protein Structure Database. *Nat. Struct. Mol. Biol.* **28**, 869–870 (2021).

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

214. Turoňová, B. *et al.* In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science* **370**, 203–208 (2020).

215. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).

216. Udenwobele, D. I. *et al.* Myristoylation: An Important Protein Modification in the Immune Response. *Front. Immunol.* **8**, 751 (2017).

217. Zhu, Y. *et al.* O-GlcNAc occurs cotranslationally to stabilize nascent polypeptide chains. *Nat. Chem. Biol.* **11**, 319–325 (2015).

218. Terwilliger, T. C. *et al.* AlphaFold predictions: great hypotheses but no match for experiment. *bioRxiv* (2022) doi:10.1101/2022.11.21.517405.

219. Pereira, J. *et al.* High-accuracy protein structure prediction in CASP14. *Proteins* **89**, 1687–1699 (2021).

220. Dialpuri, J. S. *et al.* Analysis and validation of overall N-glycan conformation in Privateer. *Acta Crystallogr D Struct Biol* **79**, 462–472 (2023).

221. Jamasb, A. R. *et al.* Graphein - a Python library for geometric deep learning and network analysis on protein structures and interaction networks. *bioRxiv* (2020) doi:10.1101/2020.07.15.204701.

222. Shields, R. L. *et al.* Lack of fucose on human IgG1 N-linked oligosaccharide improves binding to human Fcgamma RIII and antibody-dependent cellular toxicity. *J. Biol. Chem.* **277**, 26733–26740 (2002).

223. Parekh, R. B. *et al.* Association of rheumatoid arthritis and primary osteoarthritis with changes in the glycosylation pattern of total serum IgG. *Nature* **316**, 452–457 (1985).

224. Van Landuyt, L., Lonigro, C., Meuris, L. & Callewaert, N. Customized protein glycosylation to improve biopharmaceutical function and targeting. *Curr. Opin. Biotechnol.* **60**, 17–28 (2019).

225. Goochee, C. F. & Monica, T. Environmental effects on protein glycosylation. *Biotechnology* **8**, 421–427 (1990).

226. Rivinoja, A., Hassinen, A., Kokkonen, N., Kauppila, A. & Kellokumpu, S. Elevated Golgi pH impairs terminal N-glycosylation by inducing mislocalization of Golgi glycosyltransferases. *J. Cell. Physiol.* **220**, 144–154 (2009).

227. Fisher, P., Spencer, H., Thomas-Oates, J., Wood, A. J. & Ungar, D. Modeling Glycan Processing Reveals Golgi-Enzyme Homeostasis upon Trafficking Defects and Cellular Differentiation. *Cell Rep.* **27**, 1231–1243.e6 (2019).

228. Fisher, P., Thomas-Oates, J., Wood, A. J. & Ungar, D. The -Glycosylation Processing Potential of the Mammalian Golgi Apparatus. *Front Cell Dev Biol* **7**, 157 (2019).

Supplementary Data

Supplementary Table 1.1: Representative *N*-Glycosylation site data used throughout Chapter 3 to produce the following Figures: **Figure 3.6**, **Figure 3.7**, **Figure 3.8**, **Figure 3.9**.

PDB ID	GlyTouCan ID	GlyConnect ID	Glycosylation site	Glycan info	UniProt ID	UniProt Common name	Expression System	Target Organism	Method	Resolution	EMDB ID	Total terminal amino acid neighbours	Glycan Type
3u2s	G55220VL	1443	G/ASN-160	D/NAG-1	Q6TCP8	Envelope glycoprotein gp160	Homo sapiens	Human immunodeficiency virus 1	X-RAY	1.8	nan	3	High-Mannose
6pzd	G91704UR	759	A/ASN-200	B/NAG-1	R4NFR6	Neuraminidase	Spodoptera frugiperda	Influenza A virus (A/Shanghai/02/2013(H7N9))	X-RAY	1.12	nan	1	High-Mannose
5ffg	G80966KZ	2039	A/ASN-525	G/NAG-1	P06756	Integrin alpha-V	Escherichia coli	Homo sapiens	X-RAY	2.25	nan	4	High-Mannose
5ffg	G09724ZC	313	A/ASN-266	E/NAG-1	P06756	Integrin alpha-V	Escherichia coli	Homo sapiens	X-RAY	2.25	nan	4	High-Mannose
6fyw	G55220VL	1443	A/ASN-301	E/NAG-1	C0LT35	Hemagglutinin	Spodoptera frugiperda	Influenza B virus (B/Brisbane/60/2008)	X-RAY	2.2	nan	13	High-Mannose
6pxh	G80966KZ	2039	A/ASN-222	J/NAG-1	K9N5Q8	Spike glycoprotein	Homo sapiens	Middle East respiratory syndrome-related coronavirus	X-RAY	2.3	nan	5	High-Mannose
6huj	G09724ZC	313	E/ASN-149	M/NAG-1	P28472	Gamma-aminobutyric acid receptor subunit beta-3	Homo sapiens	Homo sapiens	EM	3.04	EMD-0279	3	High-Mannose
7jpi	G09724ZC	313	B/ASN-563	C/NAG-1	Q05320	Envelope glycoprotein	Homo sapiens	Ebola virus - Mayinga, Zaire, 1976	X-RAY	2.28	nan	1	High-Mannose

7ql5	G55220VL	1443	A/ASN-141	F/NAG-1	P02710	Acetylcholine receptor subunit alpha	nan	Tetronarce californica	EM	2.5	EMD-14064	10	High-Mannose
5mol	G55220VL	1443	B/ASN-394	D/NAG-1	P01854	Immunoglobulin heavy constant epsilon	Mus musculus	Homo sapiens	X-RAY	1.75	nan	25	High-Mannose
5l56	G55220VL	1443	A/ASN-1096	J/NAG-1	P70206	Plexin-A1	Homo sapiens	Mus musculus	X-RAY	4	nan	13	High-Mannose
5l56	G55220VL	1443	A/ASN-658	E/NAG-1	P70206	Plexin-A1	Homo sapiens	Mus musculus	X-RAY	4	nan	12	High-Mannose
5l56	G09724ZC	313	A/ASN-1041	I/NAG-1	P70206	Plexin-A1	Homo sapiens	Mus musculus	X-RAY	4	nan	2	High-Mannose
5l56	G55220VL	1443	A/ASN-670	F/NAG-1	P70206	Plexin-A1	Homo sapiens	Mus musculus	X-RAY	4	nan	1	High-Mannose
4wk0	G09724ZC	313	A/ASN-275	D/NAG-1	P08648	Integrin alpha-5	Homo sapiens	Homo sapiens	X-RAY	1.78	nan	11	High-Mannose
4mj2	G23799GS	354	B/ASN-372	F/NAG-1	P35475	Alpha-L-iduronidase	Arabidopsis thaliana	Homo sapiens	X-RAY	2.1	nan	25	High-Mannose
6bfu	G56014GC	2638	A/ASN-74	D/NAG-1	A0A140ESF1	Spike protein	Drosophila melanogaster	Porcine deltacoronavirus	EM	3.5	EMD-7094	17	High-Mannose
6bfu	G56014GC	2638	B/ASN-914	b/NAG-1	A0A140ESF1	Spike protein	Drosophila melanogaster	Porcine deltacoronavirus	EM	3.5	EMD-7094	7	High-Mannose
6bfu	G56014GC	2638	A/ASN-241	F/NAG-1	A0A140ESF1	Spike protein	Drosophila melanogaster	Porcine deltacoronavirus	EM	3.5	EMD-7094	2	High-Mannose
3rg1	G61846BY	3329	A/ASN-402	R/NAG-1	A6QNK7	CD180 molecule	Spodoptera frugiperda	Bos taurus	X-RAY	2.91	nan	18	High-Mannose
5nuz	G09724ZC	313	C/ASN-178	E/NAG-1	C1K9J9	Pre-glycoprotein polyprotein GP complex	Homo sapiens	Argentinian mammarenavirus	X-RAY	1.85	nan	5	High-Mannose
4adf	G55220VL	1443	C/ASN-95	a/NAG-1	P03228	Secreted protein BARF1	HOMO SAPIENS	Human gammaherpesvirus 4	X-RAY	4.4	nan	6	High-Mannose
6crd	G68668TB	3373	D/ASN-200	Q/NAG-1	P03472	Tetrabrachion	Insect cell expression vector pTIE1	Staphylothermus marinus	X-RAY	2.57	nan	1	High-Mannose
7pfp	G80966KZ	2039	C/ASN-275	O/NAG-1	P07911	Uromodulin	nan	Homo sapiens	EM	6.1	EMD-13378	12	High-Mannose

7pfp	G36191CD	3233	C/ASN-232	N/NAG-1	P07911	Uromodulin	nan	Homo sapiens	EM	6.1	EMD-13378	11	Complex
5dlv	G37135JQ	3265	A/ASN-524	C/NAG-1	Q64610	Ectonucleotide pyrophosphatase/ phosphodiesterase family member 2	Homo sapiens	Rattus norvegicus	X-RAY	2	nan	32	High-Mannose
7nrh	G07617FP	1964	A/ASN-134	F/NAG-1	A0A077D153	Envelopment polyprotein	Homo sapiens	Hantaan orthohantavirus	EM	19	EMD-12544	1	High-Mannose
4q4b	G60230HH	2472	A/ASN-325	F/NAG-1	Q14108	Lysosome membrane protein 2	homo sapiens	Homo sapiens	X-RAY	2.82	nan	7	High-Mannose
4gwm	G37135JQ	3265	A/ASN-547	F/NAG-1	Q16820	Meprin A subunit beta	Trichoplusia ni	Homo sapiens	X-RAY	1.85	nan	11	High-Mannose
6fb3	G83582BK	27	C/ASN-2365	Q/NAG-1	Q9DER5	Teneurin-2	Homo sapiens	Gallus gallus	X-RAY	2.38	nan	4	High-Mannose
3gwj	G19958IL	2784	B/ASN-196	I/NAG-1	Q7Z1F8	Arylphorin	nan	Antheraea pernyi	X-RAY	2.43	nan	8	High-Mannose
7d3f	G40702WU	362	B/ASN-109	E/NAG-1	Q1HG43	Dual oxidase maturation factor 1	Homo sapiens	Homo sapiens	EM	2.3	EMD-30556	22	High-Mannose
6jx7	G07617FP	1964	B/ASN-1218	I/NAG-1	C6GHB7	Fibritin	Homo sapiens	Feline infectious peritonitis virus	EM	3.31	EMD-9891	12	High-Mannose
6jx7	G23799GS	354	C/ASN-357	s/NAG-1	C6GHB7	Fibritin	Homo sapiens	Feline infectious peritonitis virus	EM	3.31	EMD-9891	7	High-Mannose
3i26	G55220VL	1443	D/ASN-313	P/NAG-1	P0C0V9	Hemagglutinin-esterase	Homo sapiens	Breda virus serotype 1	X-RAY	1.8	nan	18	High-Mannose
6sff	G09724ZC	313	A/ASN-220	B/NAG-1	P43432	Interleukin-12 subunit beta	Homo sapiens	Mus musculus	X-RAY	2.4	nan	16	High-Mannose
5fuk	G34442SS	2767	B/ASN-122	D/NAG-1	A0A023H437	Aromatic peroxygenase	nan	Marasmius rotula	X-RAY	1.55	nan	4	High-Mannose
5j67	G89864BN	1387	B/ASN-732	F/NAG-1	O75129	Astrotactin-2	Homo sapiens	Homo sapiens	X-RAY	3.16	nan	21	High-Mannose
6zjz	G56014GC	2638	A/ASN-293	C/NAG-1	Q9NR97	Toll-like receptor 8	Drosophila falleni	Homo sapiens	X-RAY	2.49	nan	12	High-Mannose

6zjz	G56014GC	2638	A/ASN-590	E/NAG-1	Q9NR97	Toll-like receptor 8	Drosophila falleni	Homo sapiens	X-RAY	2.49	nan	3	High-Mannose
3hn3	G89864BN	1387	D/ASN-173	G/NAG-1	P08236	Beta-glucuronidase	Mus musculus	Homo sapiens	X-RAY	1.7	nan	17	High-Mannose
3t6q	G37135JQ	3265	A/ASN-402	E/NAG-1	Q62192	CD180 antigen	Drosophila	Mus musculus	X-RAY	1.9	nan	22	High-Mannose
4gtw	G09724ZC	313	A/ASN-567	C/NAG-1	P06802	Ectonucleotide pyrophosphatase/ phosphodiesterase family member 1	Homo sapiens	Mus musculus	X-RAY	2.7	nan	4	High-Mannose
2hr7	G56014GC	2638	A/ASN-111	D/NAG-1	P06213	Insulin receptor	Cricetulus griseus	Homo sapiens	X-RAY	2.32	nan	7	High-Mannose
4neh	G07617FP	1964	A/ASN-373	E/NAG-1	P20702	Integrin alpha-X	Homo sapiens	Homo sapiens	X-RAY	2.75	nan	17	High-Mannose
6mjo	G55220VL	1443	C/ASN-45	A/NAG-1	A3RFZ7	Low affinity immunoglobulin gamma Fc region receptor III-A	Homo sapiens	Macaca mulatta	X-RAY	1.9	nan	14	High-Mannose
6qp7	G55220VL	1443	B/ASN-314	J/NAG-1	Q24323	Semaphorin-2A	Homo sapiens	Drosophila melanogaster	X-RAY	1.96	nan	12	High-Mannose
6qp7	G34442SS	2767	A/ASN-190	E/NAG-1	Q24323	Semaphorin-2A	Homo sapiens	Drosophila melanogaster	X-RAY	1.96	nan	4	High-Mannose
6qp7	G61846BY	3329	B/ASN-163	H/NAG-1	Q24323	Semaphorin-2A	Homo sapiens	Drosophila melanogaster	X-RAY	1.96	nan	1	High-Mannose
6z7a	G56014GC	2638	A/ASN-130	B/NAG-1	A0A291L8F4	Variant surface glycoprotein Sur	Trypanosoma brucei brucei	Trypanosoma brucei rhodesiense	X-RAY	1.21	nan	11	High-Mannose
7ql5	G34442SS	2767	E/ASN-141	L/NAG-1	P02714	Acetylcholine receptor subunit gamma	nan	Tetronarce californica	EM	2.5	EMD-14064	10	High-Mannose
7drc	G61846BY	3329	C/ASN-143	F/NAG-1	A0A2I8B6R1	Membrane-localized LRR receptor-like protein	Trichoplusia ni	Nicotiana benthamiana	EM	2.92	EMD-30826	30	High-Mannose

6f9t	G80858MF	2967	A/ASN-109	C/NAG-1	P12821	Angiotensin-converting enzyme	Cricetulus griseus	Homo sapiens	X-RAY	1.6	nan	23	Complex
7sn0	G61751GZ	629	B/ASN-322	K/NAG-1	Q9BYF1	Angiotensin-converting enzyme 2	Homo sapiens	Homo sapiens	X-RAY	3.08	nan	9	Complex
7sn0	G29905OR	1695	B/ASN-546	M/NAG-1	Q9BYF1	Angiotensin-converting enzyme 2	Homo sapiens	Homo sapiens	X-RAY	3.08	nan	1	Complex
6i01	G61751GZ	629	A/ASN-225	D/NAG-1	O94923	D-glucuronyl C5-epimerase	Homo sapiens	Homo sapiens	X-RAY	2.1	nan	7	Complex
6s7t	G80966KZ	2039	E/ASN-299	N/NAG-1	P04843	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 1	nan	Homo sapiens	EM	3.5	EMD-10112	9	High-Mannose
4cxp	G80966KZ	2039	A/ASN-91	B/NAG-1	Q9C9G4	Endonuclease 2	ARABIDOPSIS THALIANA	Arabidopsis thaliana	X-RAY	1.22	nan	3	High-Mannose
7lbf	G07617FP	1964	C/ASN-160	K/NAG-1	Q8BCU3	Envelope glycoprotein O	Homo sapiens	Human betaherpesvirus 5	EM	2.8	EMD-23253	26	High-Mannose
3wo3	G34442SS	2767	F/ASN-197	W/NAG-1	Q13478	Interleukin-18 receptor 1	Spodoptera frugiperda	Homo sapiens	X-RAY	3.1	nan	2	High-Mannose
7s69	G09724ZC	313	A/ASN-87	C/NAG-1	Q68F17	LOC446283 protein	Spodoptera frugiperda	Xenopus laevis	X-RAY	3.04	nan	3	High-Mannose
7de5	G09724ZC	313	A/ASN-205	B/NAG-1	L8ICE9	Lactoperoxidase	nan	Bos grunniens	X-RAY	1.55	nan	4	High-Mannose
1ppf	G61334IA	3077	E/ASN-159	A/NAG-1	P08246	Neutrophil elastase	nan	Homo sapiens	X-RAY	1.8	nan	1	Complex
1cvi	G80858MF	2967	A/ASN-1188	E/NAG-1	P15309	Prostatic acid phosphatase	nan	Homo sapiens	X-RAY	3.2	nan	10	Complex
1cvi	G34442SS	2767	B/ASN-2301	I/NAG-1	P15309	Prostatic acid phosphatase	nan	Homo sapiens	X-RAY	3.2	nan	5	High-Mannose
7rgf	G12398HZ	910	A/ASN-236	C/NAG-1	Q91XX0	Protocadherin gamma C4	Homo sapiens	Mus musculus	X-RAY	2.4	nan	2	Complex

6qp8	G83161QT	654	A/ASN-321	F/NAG-1	A0A0B4KG38	Semaphorin 2b, isoform D	Homo sapiens	Drosophila melanogaster	X-RAY	2.33	nan	19	High-Mannose
3wjm	G23799GS	354	D/ASN-208	J/NAG-1	H9JHM9	Silkworm storage protein	nan	Bombyx mori	X-RAY	2.8	nan	7	High-Mannose
2dw2	G30159WR	3465	B/ASN-371	D/NAG-1	Q90282	Zinc metalloproteinase -disintegrin-like VAP2B	nan	Crotalus atrox	X-RAY	2.7	nan	1	Complex
1zag	G90725ZC	693	B/ASN-239	F/NAG-1	P25311	Zinc-alpha-2-glycoprotein	nan	Homo sapiens	X-RAY	2.8	nan	12	Complex
6ibm	G56014GC	2638	B/ASN-192	F/NAG-1	P06280	Alpha-galactosidase A	Cricetulus griseus	Homo sapiens	X-RAY	2.07	nan	2	High-Mannose
2gd4	G56014GC	2638	I/ASN-155	D/NAG-1	P01008	Antithrombin-III	nan	Homo sapiens	X-RAY	3.3	nan	7	High-Mannose
7syy	G09724ZC	313	A/ASN-481	C/NAG-1	F4YH71	Attachment glycoprotein	Trichoplusia	Hendra henipavirus	X-RAY	2.74	nan	12	High-Mannose
4cvu	G60230HH	2472	A/ASN-255	E/NAG-1	A0A075B5H6	Beta-mannosidase	nan	Trichoderma harzianum	X-RAY	1.9	nan	39	High-Mannose
6z30	G09724ZC	313	A/ASN-1312	B/NAG-1	P11717	Cation-independent mannose-6-phosphate receptor	Spodoptera frugiperda	Homo sapiens	X-RAY	1.5	nan	3	High-Mannose
6mf0	G56014GC	2638	B/ASN-2118	H/NAG-1	P12263	Coagulation factor VIII	Cricetulus griseus	Sus scrofa	X-RAY	3.2	nan	7	High-Mannose
2ok5	G55220VL	1443	A/ASN-97	B/NAG-1	P00751	Complement factor B	Homo sapiens	Homo sapiens	X-RAY	2.3	nan	17	High-Mannose
6vlk	G55220VL	1443	A/ASN-257	C/NAG-1	Q4JR05	Envelope glycoprotein B	Homo sapiens	Human herpesvirus 3 strain Oka vaccine	X-RAY	2.45	nan	9	High-Mannose
6c5v	G56014GC	2638	A/ASN-60	D/NAG-1	K9US75	Envelope glycoprotein H	Homo sapiens	Human gammaherpesvirus 4	EM	4.8	EMD-7344	3	High-Mannose
5xwd	G39213VZ	3258	A/ASN-328	B/NAG-1	P00533	Epidermal growth factor receptor	Homo sapiens	Homo sapiens	X-RAY	2.89	nan	5	Complex

6urh	G55220VL	1443	C/ASN-532	E/NAG-1	A0A2P0NE26	Genome polyprotein	Homo sapiens	Hepacivirus C	X-RAY	2.2	nan	5	High-Mannose
4mze	G80966KZ	2039	A/ASN-351	C/NAG-1	P08492	Hemagglutinin-neuraminidase	Trichoplusia ni	Human parainfluenza 3 virus (strain NIH 47885)	X-RAY	1.8	nan	3	High-Mannose
2pe4	G09724ZC	313	A/ASN-350	B/NAG-1	Q12794	Hyaluronidase-1	Drosophila melanogaster	Homo sapiens	X-RAY	2	nan	4	High-Mannose
3ze2	G09724ZC	313	B/ASN-320	G/NAG-1	P05106	Integrin beta-3	CRICETULUS GRISEUS	Homo sapiens	X-RAY	2.35	nan	1	High-Mannose
6dg5	G27389SR	2099	B/ASN-43	E/NAG-1	P16297	Interleukin-2 receptor subunit beta	Trichoplusia ni	Mus musculus	X-RAY	2.52	nan	24	High-Mannose
5mgr	G55220VL	1443	A/ASN-169	C/NAG-1	Q12918	Killer cell lectin-like receptor subfamily B member 1	Homo sapiens	Homo sapiens	X-RAY	1.8	nan	7	High-Mannose
1o7d	G83161QT	654	C/ASN-497	F/NAG-1	Q29451	Lysosomal alpha-mannosidase	nan	Bos taurus	X-RAY	2.7	nan	17	High-Mannose
2h6o	G63337SS	1334	A/ASN-166	J/NAG-1	Q9QP87	Major outer envelope glycoprotein gp350	Spodoptera frugiperda	Human gammaherpesvirus 4	X-RAY	3.5	nan	12	High-Mannose
1ckl	G61846BY	3329	E/ASN-80	P/NAG-1	P15529	Membrane cofactor protein	Cricetulus griseus	Homo sapiens	X-RAY	3.1	nan	8	High-Mannose
6bbe	G39213VZ	3258	A/ASN-303	B/NAG-1	Q9N1X4	Pulmonary surfactant-associated protein D	Homo sapiens	Sus scrofa	X-RAY	1.9	nan	1	Complex
7lyu	G09724ZC	313	A/ASN-3412	C/NAG-1	Q60841	Reelin	Homo sapiens	Mus musculus	X-RAY	3	nan	6	High-Mannose
1ioo	G32473MI	571	B/ASN-28	D/NAG-1	Q7SID5	Ribonuclease S-F11	nan	Nicotiana alata	X-RAY	1.55	nan	13	Complex
1o7v	G55220VL	1443	A/ASN-105	B/NAG-1	Q9Y286	Sialic acid-binding Ig-like lectin 7	Cricetulus griseus	Homo sapiens	X-RAY	1.9	nan	14	High-Mannose

1rer	G07483YN	678	B/ASN-141	E/NAG-1	P03315	Structural polyprotein	nan	Semliki Forest virus	X-RAY	3.2	nan	5	Complex
7oix	G23799GS	354	A/ASN-312	D/NAG-1	P60508	Syncytin-2	Homo sapiens	Homo sapiens	EM	3.6	EMD-12935	19	High-Mannose
1gya	G83161QT	654	A/ASN-65	B/NAG-1	P06729	T-cell surface antigen CD2	Cricetulus griseus	Homo sapiens	NMR	nan	nan	12	High-Mannose
6tp5	G40702WU	362	A/ASN-382	D/NAG-1	Q9NXG6	Transmembrane prolyl 4-hydroxylase	Spodoptera frugiperda	Homo sapiens	X-RAY	2.25	nan	2	High-Mannose
6trf	G09724ZC	313	A/ASN-56	B/NAG-1	G0SB58	UDP-glucose-glycoprotein glucosyltransferase-like protein	Homo sapiens	Chaetomium thermophilum var. thermophilum DSM 1495	X-RAY	4.11	nan	16	High-Mannose
4bxs	G80966KZ	2039	V/ASN-212	C/NAG-1	Q7SZN0	Venom prothrombin activator pseutarin-C non-catalytic subunit	CRICETULUS GRISEUS	Pseudonaja textilis	X-RAY	3.32	nan	21	High-Mannose
5uem	G07617FP	1964	G/ASN-289	C/NAG-1	A0A0M3KKW9	clade A/E 93TH057 HIV-1 gp120 core	Homo sapiens	Human immunodeficiency virus 1	X-RAY	2.7	nan	1	High-Mannose
6yt7	G61334IA	3077	B/ASN-297	D/NAG-1	P01857	Immunoglobulin heavy constant gamma 1	Homo sapiens	Homo sapiens	X-RAY	1.55	nan	23	Complex
5w5n	G27919IH	2198	B/ASN-297	D/NAG-1	P01861	Immunoglobulin heavy constant gamma 4	Homo sapiens	Homo sapiens	X-RAY	1.85	nan	30	Complex
4l4j	G45889JQ	2705	A/ASN-297	C/NAG-1	P01859	Immunoglobulin heavy constant gamma 2	Homo sapiens	Homo sapiens	X-RAY	1.92	nan	21	Complex
6d58	G80858MF	2967	B/ASN-297	D/NAG-1	P01860	Immunoglobulin heavy constant gamma 3	Homo sapiens	Homo sapiens	X-RAY	2.39	nan	21	Complex

Supplementary Table 1.2: Representative *N*-Glycosylation site data of predicted protein models by AlphaFold that did not result in any clashes between grafted *N*-glycan and protein backbone after the grafting procedure. Data was extracted from the UniProt and GlyConnect. The following models (UniProt ID) were used throughout Chapter 4 to produce the following Figures: **Figure 4.3**, **Figure 4.4**, **Figure 4.5**, **Figure 4.6**.

UniProt ID	Protein Name	Species	Glycosylation Site	Glycan Type	GlyConnect ID	Template exists for GlyConnect ID	Branching	Total terminal amino acid neighbours
O08795	Glucosidase 2 subunit beta	Mus musculus	469	high-mannose	3115	FALSE	omannose	11
O14524	Nuclear envelope integral membrane protein 1	Homo sapiens	125	high-mannose	1710	FALSE	omannose	8
O14657	Torsin-1B	Homo sapiens	64	high-mannose	1530	FALSE	omannose	7
O15342	V-type proton ATPase subunit e 1	Homo sapiens	70	high-mannose	1710	FALSE	omannose	1
P00742	Coagulation factor X	Homo sapiens	221	complex	3233	TRUE	triantennary	9
P00747	Plasminogen	Homo sapiens	308	complex	3233	TRUE	biantennary	11
P00749	Urokinase-type plasminogen activator	Homo sapiens	322	complex	2611	FALSE	biantennary	5
P00750	Tissue-type plasminogen activator	Homo sapiens	483	complex	1017	FALSE	biantennary	1
P01033	Metalloproteinase inhibiantennarytor 1	Homo sapiens	101	complex	1391	FALSE	triantennary	1
P01127	Platelet-derived growth factor subunit B	Homo sapiens	63	high-mannose	1710	FALSE	omannose	3
P01190	Pro-opiomelanocortin	Bos taurus	91	complex	1112	FALSE	biantennary	2
P01215	Glycoprotein hormones alpha chain	Homo sapiens	76	complex	1641	FALSE	biantennary	1

P01229	Lutropin subunit beta	Homo sapiens	50	complex	1277	FALSE	biantennary	2
P01231	Lutropin subunit beta	Ovis aries	33	complex	1844	FALSE	biantennary	5
P01588	Erythropoietin	Homo sapiens	65	complex	1152	FALSE	triantennary	2
P01592	Immunoglobulin J chain	Mus musculus	70	complex	3649	FALSE	biantennary	8
P01730	T-cell surface glycoprotein CD4	Homo sapiens	296	complex	3233	TRUE	biantennary	7
P01871	Immunoglobulin heavy constant mu	Homo sapiens	440	high-mannose	1334	TRUE	omannose	1
P01876	Immunoglobulin heavy constant alpha 1	Homo sapiens	144	complex	1478	FALSE	biantennary	16
P01903	HLA class II histocompatibiantennarylity antigen, DR alpha chain	Homo sapiens	103	high-mannose	1163	FALSE	omannose	4
P02458	Collagen alpha-1(II) chain	Homo sapiens	1388	high-mannose	3206	FALSE	omannose	6
P02679	Fibrinogen gamma chain	Homo sapiens	78	complex	1021	FALSE	biantennary	3
P02752	Riboflavin-biantennarynding protein	Gallus gallus	164	complex	1378	FALSE	biantennary	3
P02771	Alpha-fetoprotein	Homo sapiens	251	complex	3233	TRUE	biantennary	7
P04035	3-hydroxy-3-methylglutaryl-coenzyme A reductase	Homo sapiens	281	high-mannose	1710	FALSE	omannose	1
P07214	SPARC	Mus musculus	115	high-mannose	1710	FALSE	omannose	6
P07288	Prostate-specific antigen	Homo sapiens	69	complex	1850	FALSE	triantennary	4
P08123	Collagen alpha-2(I) chain	Homo sapiens	1267	high-mannose	1530	FALSE	omannose	10
P08138	Tumor necrosis factor receptor superfamily member 16	Homo sapiens	60	complex	1951	FALSE	biantennary	8

P08861	Chymotrypsin-like elastase family member 3B	Homo sapiens	114	complex	107	FALSE	biantennary	1
P09466	Glycodelin	Homo sapiens	46	complex	1080	FALSE	biantennary	20
P0DN86	Choriogonadotropin subunit beta 3	Homo sapiens	33	complex	1077	FALSE	biantennary	4
P0DN86	Choriogonadotropin subunit beta 3	Homo sapiens	50	complex	1077	FALSE	biantennary	2
P11087	Collagen alpha-1(I) chain	Mus musculus	1354	high-mannose	3206	FALSE	omannose	13
P13987	CD59 glycoprotein	Homo sapiens	43	complex	3225	FALSE	biantennary	5
P14210	Hepatocyte growth factor	Homo sapiens	294	complex	1077	FALSE	triantennary	10
P14210	Hepatocyte growth factor	Homo sapiens	402	complex	1077	FALSE	triantennary	6
P14210	Hepatocyte growth factor	Homo sapiens	566	complex	1077	FALSE	triantennary	6
P16870	Carboxypeptidase E	Homo sapiens	139	high-mannose	3115	FALSE	omannose	4
P18242	Cathepsin D	Mus musculus	134	high-mannose	1530	FALSE	omannose	3
P18632	Pectate lyase 1	Cryptomeria japonica	191	complex	1347	FALSE	biantennary	13
P18632	Pectate lyase 1	Cryptomeria japonica	354	complex	1347	FALSE	biantennary	2
P19224	UDP-glucuronosyltransferase 1-6	Homo sapiens	346	high-mannose	1530	FALSE	omannose	13
P19823	Inter-alpha-trypsin inhibiantennarytor heavy chain H2	Homo sapiens	118	complex	799	FALSE	biantennary	11
P19827	Inter-alpha-trypsin inhibiantennarytor heavy chain H1	Homo sapiens	285	complex	799	FALSE	biantennary	8

P19827	Inter-alpha-trypsin inhibiantennarytor heavy chain H1	Homo sapiens	588	complex	3366	FALSE	biantennary	6
P20933	N(4)-(beta-N-acetylglucosaminyl)-L-asparaginase	Homo sapiens	38	high-mannose	1530	FALSE	omannose	9
P23141	Liver carboxylesterase 1	Homo sapiens	79	high-mannose	1710	FALSE	omannose	7
P23280	Carbonic anhydrase 6	Homo sapiens	67	complex	1519	FALSE	biantennary	6
P23280	Carbonic anhydrase 6	Homo sapiens	256	complex	1519	FALSE	biantennary	6
P25063	Signal transducer CD24	Homo sapiens	52	high-mannose	1710	FALSE	omannose	1
P26048	Gamma-aminobutyric acid receptor subunit alpha-2	Mus musculus	138	high-mannose	1530	FALSE	omannose	5
P30542	Adenosine receptor A1	Homo sapiens	159	high-mannose	3206	FALSE	omannose	1
P36955	Pigment epithelium-derived factor	Homo sapiens	285	complex	3507	FALSE	biantennary	5
P42098	Zona pellucida sperm-biantennarynding protein 3	Sus scrofa	146	complex	1819	FALSE	biantennary	1
P48199	C-reactive protein	Rattus norvegicus	147	complex	1080	FALSE	biantennary	2
P50454	Serpin H1	Homo sapiens	120	high-mannose	1530	FALSE	omannose	9
P51910	Apolipoprotein D	Mus musculus	98	high-mannose	3115	FALSE	omannose	7
P55083	Microfibril-associated glycoprotein 4	Homo sapiens	87	high-mannose	1530	FALSE	omannose	2
P61823	Ribonuclease pancreatic	Bos taurus	60	high-mannose	1334	TRUE	omannose	2
P62812	Gamma-aminobutyric acid receptor subunit alpha-1	Mus musculus	137	high-mannose	1530	FALSE	omannose	1

P78410	Butyrophilin subfamily 3 member A2	Homo sapiens	115	high-mannose	1163	FALSE	omannose	2
Q00493	Carboxypeptidase E	Mus musculus	139	high-mannose	1530	FALSE	omannose	1
Q05769	Prostaglandin G/H synthase 2	Mus musculus	53	high-mannose	1539	FALSE	omannose	25
Q05769	Prostaglandin G/H synthase 2	Mus musculus	396	high-mannose	263	FALSE	omannose	13
Q05769	Prostaglandin G/H synthase 2	Mus musculus	580	high-mannose	263	FALSE	omannose	1
Q09163	Protein delta homolog 1	Mus musculus	100	complex	1003	FALSE	biantennary	13
Q13410	Butyrophilin subfamily 1 member A1	Homo sapiens	55	complex	1519	FALSE	triantennary	7
Q13410	Butyrophilin subfamily 1 member A1	Homo sapiens	215	complex	1523	FALSE	biantennary	9
Q13438	Protein OS-9	Homo sapiens	177	high-mannose	1530	FALSE	omannose	1
Q14508	WAP four-disulfide core domain protein 2	Homo sapiens	44	high-mannose	3115	FALSE	omannose	1
Q14766	Latent-transforming growth factor beta-biantennarynding protein 1	Homo sapiens	1366	complex	1118	FALSE	biantennary	21
Q16842	CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 2	Homo sapiens	211	high-mannose	3115	FALSE	omannose	11
Q3TRM4	Patatin-like phospholipase domain-containing protein 6	Mus musculus	9	high-mannose	1530	FALSE	omannose	2
Q61288	Serine/threonine-protein kinase receptor R3	Mus musculus	32	high-mannose	3206	FALSE	omannose	4
Q62313	Trans-Golgi network integral membrane protein 1	Mus musculus	110	high-mannose	3206	FALSE	omannose	5

Q6PIX5	Inactive rhomboid protein 1	Mus musculus	584	high-mannose	3115	FALSE	omannose	19
Q6X4U4	Sclerostin domain-containing protein 1	Homo sapiens	47	complex	1523	FALSE	biantennary	1
Q6ZMG9	Ceramide synthase 6	Homo sapiens	18	high-mannose	1530	FALSE	omannose	1
Q7Z4H8	Protein O-glucosyltransferase 3	Homo sapiens	61	high-mannose	1710	FALSE	omannose	1
Q7Z7H5	Transmembrane emp24 domain-containing protein 4	Homo sapiens	117	high-mannose	1530	FALSE	omannose	4
Q8BJI1	Sodium-dependent neutral amino acid transporter SLC6A17	Mus musculus	186	high-mannose	3206	FALSE	omannose	1
Q8BJS4	SUN domain-containing protein 2	Mus musculus	650	high-mannose	3206	FALSE	omannose	3
Q8BW41	Protein O-linked-mannose beta-1,4-N-acetylglucosaminyltransferase 2	Mus musculus	276	high-mannose	1530	FALSE	omannose	4
Q8CIV2	Membralin	Mus musculus	180	high-mannose	1530	FALSE	omannose	3
Q8K2B0	Endoplasmic reticulum protein SC65	Mus musculus	367	high-mannose	3206	FALSE	omannose	2
Q8R2Y2	Cell surface glycoprotein MUC18	Mus musculus	58	high-mannose	3115	FALSE	omannose	4
Q91ZW2	GDP-fucose protein O-fucosyltransferase 1	Mus musculus	165	high-mannose	1530	FALSE	omannose	4
Q921I1	Serotransferrin	Mus musculus	513	complex	3649	FALSE	biantennary	3
Q921T2	Torsin-1A-interacting protein 1	Mus musculus	411	high-mannose	1530	FALSE	omannose	9
Q924Z4	Ceramide synthase 2	Mus musculus	19	high-mannose	1530	FALSE	omannose	1
Q92765	Secreted frizzled-related protein 3	Homo sapiens	49	high-mannose	1163	FALSE	omannose	7

Q92876	Kallikrein-6	Homo sapiens	134	complex	1523	FALSE	biantennary	1
Q96JB6	Lysyl oxidase homolog 4	Homo sapiens	629	high-mannose	1163	FALSE	omannose	9
Q99JY8	Phospholipid phosphatase 3	Mus musculus	171	high-mannose	3115	FALSE	omannose	2
Q9BY76	Angiopoietin-related protein 4	Homo sapiens	177	complex	1021	FALSE	biantennary	3
Q9D8B7	Junctional adhesion molecule C	Mus musculus	192	high-mannose	3115	FALSE	omannose	3
Q9DBV4	Matrix remodeling-associated protein 8	Mus musculus	118	high-mannose	3115	FALSE	omannose	6
Q9ER41	Torsin-1B	Mus musculus	64	high-mannose	1530	FALSE	omannose	8
Q9GZX9	Twisted gastrulation protein homolog 1	Homo sapiens	81	high-mannose	1530	FALSE	omannose	2
Q9H488	GDP-fucose protein O-fucosyltransferase 1	Homo sapiens	160	high-mannose	3115	FALSE	omannose	4
Q9QYK5	Heparan-sulfate 6-O-sulfotransferase 1	Mus musculus	320	high-mannose	1710	FALSE	omannose	3
Q9WU62	Inner centromere protein	Mus musculus	450	complex	1667	FALSE	triantennary	1
Q9WVT6	Carbonic anhydrase 14	Mus musculus	213	high-mannose	3115	FALSE	omannose	3
Q9YGP1	C-type lectin TsL	Trimeresurus stejnegeri	28	high-mannose	263	FALSE	omannose	1
Q9Z2E9	Seipin	Mus musculus	88	high-mannose	1530	FALSE	omannose	7
A2A690	Protein TANC2	Mus musculus	1932	complex	3649	FALSE	biantennary	3
O88668	Protein CREG1	Mus musculus	160	high-mannose	1530	FALSE	omannose	17
P01218	Glycoprotein hormones alpha chain	Ovis aries	80	complex	1844	FALSE	biantennary	1
P04651	Lutropin subunit beta	Bos taurus	33	complex	2232	FALSE	biantennary	4

P13087	Miraculin	Synsepalum dulcificum	71	complex	2647	FALSE	biantennary	1
P26792	Beta-fructofuranosidase, insoluble isoenzyme 1	Daucus carota	170	complex	1768	FALSE	biantennary	8
P26792	Beta-fructofuranosidase, insoluble isoenzyme 1	Daucus carota	311	complex	1768	FALSE	biantennary	7
Q3TEW6	Myelin protein zero-like protein 1	Mus musculus	50	high-mannose	3115	FALSE	omannose	3
Q61704	Inter-alpha-trypsin inhibiantennarytor heavy chain H3	Mus musculus	580	complex	430	FALSE	biantennary	7
Q8BYU6	Torsin-1A-interacting protein 2	Mus musculus	318	high-mannose	3206	FALSE	omannose	6
Q8BYW9	EGF domain-specific O-linked N-acetylglucosamine transferase	Mus musculus	354	high-mannose	1530	FALSE	omannose	7
Q8C7X2	ER membrane protein complex subunit 1	Mus musculus	917	high-mannose	1710	FALSE	omannose	3
Q8R1V4	Transmembrane emp24 domain-containing protein 4	Mus musculus	117	high-mannose	1530	FALSE	omannose	1
Q91VF5	EMI domain-containing protein 1	Mus musculus	136	high-mannose	3115	FALSE	omannose	6
Q91XD7	Protein disulfide isomerase Creld1	Mus musculus	205	high-mannose	1530	FALSE	omannose	1
Q96DA0	Zymogen granule protein 16 homolog B	Homo sapiens	197	high-mannose	1530	FALSE	omannose	1
Q99MR3	Solute carrier family 12 member 9	Mus musculus	228	high-mannose	3206	FALSE	omannose	13
Q9BU23	Lipase maturation factor 2	Homo sapiens	616	high-mannose	1530	FALSE	omannose	10

Q9CY50	Translocon-associated protein subunit alpha	Mus musculus	136	high-mannose	1530	FALSE	omannose	5
Q9EPK6	Nucleotide exchange factor SIL1	Mus musculus	197	high-mannose	1530	FALSE	omannose	3
Q9ER38	Torsin-3A	Mus musculus	110	high-mannose	1530	FALSE	omannose	14
Q9UN71	Protocadherin gamma-B4	Homo sapiens	543	high-mannose	1163	FALSE	omannose	7
Q9Y5F8	Protocadherin gamma-B7	Homo sapiens	545	high-mannose	1163	FALSE	omannose	4
Q9Y5G0	Protocadherin gamma-B5	Homo sapiens	541	high-mannose	1163	FALSE	omannose	4
P26334	Variant surface glycoprotein MITAT 1.6	Trypanosoma brucei brucei	456	high-mannose	2472	TRUE	omannose	13
P59024	Peptidyl-prolyl cis-trans isomerase FKBP14	Mus musculus	176	high-mannose	1530	FALSE	omannose	4
P81191	Relaxin-like protein AGF	Hypanus sabiantennaryn us	37	complex	2646	FALSE	biantennary	5
Q6F5E0	Transmembrane protein 158	Mus musculus	73	high-mannose	1710	FALSE	omannose	16
Q8BU25	Inactive serine protease PAMR1	Mus musculus	614	high-mannose	3206	FALSE	omannose	2
Q8BXA5	Lipid scramblase CLPTM1L	Mus musculus	91	high-mannose	1530	FALSE	omannose	20
Q9CPW5	Translocon-associated protein subunit beta	Mus musculus	88	high-mannose	1530	FALSE	omannose	7
Q9U8R2	Androgenic gland hormone	Armadillidium vulgare	133	complex	436	FALSE	biantennary	11
Q32M26	Uncharacterized protein C11orf87 homolog	Mus musculus	19	high-mannose	3206	FALSE	omannose	14
Q80VP8	Transmembrane protein 106C	Mus musculus	184	high-mannose	998	FALSE	omannose	18
A2BH40	AT-rich interactive domain-containing protein 1A	Mus musculus	1598	high-mannose	3115	FALSE	omannose	4

A2APX8	Sodium channel protein type 1 subunit alpha	Mus musculus	1403	high-mannose	1710	FALSE	omannose	1
E9PVB5	Tetratricopeptide repeat protein 17	Mus musculus	815	high-mannose	3206	FALSE	omannose	4
B1AZA5	Transmembrane protein 245	Mus musculus	548	high-mannose	3115	FALSE	omannose	2
B1AWJ4	Transmembrane protein 8B	Mus musculus	348	high-mannose	1710	FALSE	omannose	15
A6X935	Inter alpha-trypsin inhibiantennarytor, heavy chain 4	Mus musculus	517	complex	3649	FALSE	biantennary	3
O00115	Deoxyribonuclease-2-alpha	Homo sapiens	86	high-mannose	1710	FALSE	omannose	1
B0FP48	Uroplakin-3b-like protein	Homo sapiens	110	high-mannose	1710	FALSE	omannose	5
G5E8Q8	Adhesion G protein-coupled receptor F5	Mus musculus	86	high-mannose	1710	FALSE	omannose	2
E9Q7X7	Neurexin-2	Mus musculus	1236	high-mannose	3115	FALSE	omannose	3
B2RXS4	Plexin-B2	Mus musculus	1005	high-mannose	998	FALSE	omannose	7
A2AJQ3	Probable C-mannosyltransferase DPY19L4	Mus musculus	122	high-mannose	1530	FALSE	omannose	7
E9PXF0	Protocadherin-17	Mus musculus	451	high-mannose	3115	FALSE	omannose	11
O95502	Neuronal pentraxin receptor	Homo sapiens	42	high-mannose	3206	FALSE	omannose	5
O89026	Roundabout homolog 1	Mus musculus	788	high-mannose	3115	FALSE	omannose	13
O96005	Cleft lip and palate transmembrane protein 1	Homo sapiens	295	high-mannose	1530	FALSE	omannose	4
O95302	Peptidyl-prolyl cis-trans isomerase FKBP9	Homo sapiens	174	high-mannose	1530	FALSE	omannose	6
O88325	Alpha-N-acetylglucosaminidase	Mus musculus	501	high-mannose	1530	FALSE	omannose	1

O88829	Lactosylceramide alpha-2,3-sialyltransferase	Mus musculus	235	high-mannose	1530	FALSE	omannose	16
P00450	Ceruloplasmin	Homo sapiens	138	complex	1523	FALSE	biantennary	15
P00450	Ceruloplasmin	Homo sapiens	397	complex	1519	FALSE	triantennary	4
P00450	Ceruloplasmin	Homo sapiens	762	complex	1519	FALSE	biantennary	2
P02751	Fibronectin	Homo sapiens	1007	complex	3353	FALSE	biantennary	6
P02750	Leucine-rich alpha-2-glycoprotein	Homo sapiens	186	complex	3507	FALSE	biantennary	10
P02750	Leucine-rich alpha-2-glycoprotein	Homo sapiens	79	complex	3534	FALSE	triantennary	11
P02750	Leucine-rich alpha-2-glycoprotein	Homo sapiens	325	complex	1519	FALSE	biantennary	1
P02749	Beta-2-glycoprotein 1	Homo sapiens	253	complex	3353	FALSE	biantennary	3
O09118	Netrin-1	Mus musculus	95	high-mannose	3115	FALSE	omannose	3
O00592	Podocalyxin	Homo sapiens	104	high-mannose	1710	FALSE	omannose	1
O00391	Sulfhydryl oxidase 1	Homo sapiens	130	complex	1519	FALSE	biantennary	10
O00391	Sulfhydryl oxidase 1	Homo sapiens	575	complex	1523	FALSE	biantennary	3
O00391	Sulfhydryl oxidase 1	Homo sapiens	591	complex	1523	FALSE	biantennary	2
O09126	Semaphorin-4D	Mus musculus	139	high-mannose	3115	FALSE	omannose	2
O00300	Tumor necrosis factor receptor superfamily member 11B	Homo sapiens	289	complex	1519	FALSE	biantennary	6
P08607	C4b-biantennarynding protein	Mus musculus	74	complex	3649	FALSE	biantennary	1
P08582	Melanotransferrin	Homo sapiens	135	high-mannose	3115	FALSE	omannose	4
P0C0L4	Complement c4-a	Homo sapiens	1328	complex	1021	FALSE	biantennary	6

P08603	Complement factor h	Homo sapiens	1029	complex	473	FALSE	biantennary	7
P08603	Complement factor h	Homo sapiens	1095	complex	3041	FALSE	triantennary	4
O75197	Low-density lipoprotein receptor-related protein 5	Homo sapiens	138	high-mannose	1710	FALSE	omannose	4
O70309	Integrin beta-5	Mus musculus	586	high-mannose	3115	FALSE	omannose	5
O75355	Ectonucleoside triphosphate diphosphohydrolase 3	Homo sapiens	81	high-mannose	2636	FALSE	omannose	1
O75462	Cytokine receptor-like factor 1	Homo sapiens	292	high-mannose	1710	FALSE	omannose	7
O75054	Immunoglobulin superfamily member 3	Homo sapiens	700	high-mannose	3115	FALSE	omannose	3
O60938	Keratocan	Homo sapiens	167	high-mannose	1530	FALSE	omannose	1
O75629	Protein CREG1	Homo sapiens	160	high-mannose	1530	FALSE	omannose	10
O60656	UDP-glucuronosyltransferase 1-9	Homo sapiens	344	high-mannose	1530	FALSE	omannose	13
O70472	Transmembrane protein 131	Mus musculus	288	high-mannose	1530	FALSE	omannose	21
O70362	Phosphatidylinositol-glycan-specific phospholipase d	Mus musculus	496	complex	3649	FALSE	biantennary	8
P01877	Immunoglobulin alpha (triantennary secretory)	Homo sapiens	205	complex	2643	FALSE	biantennary	5
P01833	Polymeric immunoglobulin receptor	Homo sapiens	186	complex	1523	FALSE	biantennary	9
P01833	Polymeric immunoglobulin receptor	Homo sapiens	469	complex	1519	FALSE	biantennary	4
P01833	Polymeric immunoglobulin receptor	Homo sapiens	499	complex	1519	FALSE	biantennary	10
P01833	Polymeric immunoglobulin receptor	Homo sapiens	90	complex	3428	FALSE	biantennary	1

O43166	Signal-induced proliferation-associated 1-like protein 1	Homo sapiens	1411	high-mannose	1710	FALSE	omannose	5
O60512	Beta-1,4-galactosyltransferase 3	Homo sapiens	166	high-mannose	1163	FALSE	omannose	1
O35409	Glutamate carboxypeptidase 2	Mus musculus	123	high-mannose	3115	FALSE	omannose	2
O35448	Lysosomal thioesterase PPT2	Mus musculus	289	high-mannose	1530	FALSE	omannose	1
O35604	Niemann-Pick C1 protein	Mus musculus	1063	high-mannose	1710	FALSE	omannose	2
O43493	Trans-Golgi network integral membrane protein 2	Homo sapiens	39	complex	1519	FALSE	biantennary	4
P12259	Coagulation factor V	Homo sapiens	1559	complex	11203	FALSE	triantennary	11
P12259	Coagulation factor V	Homo sapiens	468	complex	3471	FALSE	biantennary	10
P12259	Coagulation factor V	Homo sapiens	938	complex	11193	FALSE	triantennary	4
P12259	Coagulation factor V	Homo sapiens	1221	complex	11244	FALSE	biantennary	2
P12259	Coagulation factor V	Homo sapiens	1703	complex	11244	FALSE	biantennary	5
P11835	Integrin beta-2	Mus musculus	626	high-mannose	1710	FALSE	omannose	3
P11609	Antigen-presenting glycoprotein CD1d1	Mus musculus	38	high-mannose	3115	FALSE	omannose	13
P11678	Eosinophil peroxidase	Homo sapiens	363	high-mannose	1530	FALSE	omannose	7
P11362	Fibroblast growth factor receptor 1	Homo sapiens	317	high-mannose	1710	FALSE	omannose	1
P10721	Mast/stem cell growth factor receptor Kit	Homo sapiens	130	high-mannose	998	FALSE	omannose	5
P11627	Neural cell adhesion molecule L1	Mus musculus	1073	high-mannose	1530	FALSE	omannose	4
P10909	Clusterin	Homo sapiens	103	complex	1523	FALSE	biantennary	6

P06909	Complement factor h	Mus musculus	1061	complex	3649	FALSE	biantennary	16
P06731	Carcinoembryonic antigen-related cell adhesion molecule 5	Homo sapiens	197	high-mannose	1530	FALSE	omannose	15
P05555	Integrin alpha-M	Mus musculus	1022	high-mannose	3206	FALSE	omannose	7
P05546	Heparin cofactor 2	Homo sapiens	49	complex	473	FALSE	biantennary	2
P49961	Ectonucleoside triphosphate diphosphohydrolase 1	Homo sapiens	292	high-mannose	1163	FALSE	omannose	4
P50897	Palmitoyl-protein thioesterase 1	Homo sapiens	197	complex	1061	FALSE	biantennary	3
Q60767	Lymphocyte antigen 75	Mus musculus	529	high-mannose	3115	FALSE	omannose	7
Q5W064	Lipase member J	Homo sapiens	240	high-mannose	1710	FALSE	omannose	1
Q5VW38	Protein GPR107	Homo sapiens	169	high-mannose	1163	FALSE	omannose	4
Q61220	Protein kinase C-biantennarynding protein NELL2	Mus musculus	228	high-mannose	3115	FALSE	omannose	1
Q5FWI3	Transmembrane protein 2	Mus musculus	1234	high-mannose	1530	FALSE	omannose	1
Q61147	Ceruloplasmin	Mus musculus	138	complex	3649	FALSE	biantennary	1
Q61503	5'-nucleotidase	Mus musculus	313	high-mannose	3115	FALSE	omannose	1
Q64687	Alpha-N-acetylneuraminide alpha-2,8-sialyltransferase	Mus musculus	244	high-mannose	3115	FALSE	omannose	2
Q64237	Dopamine beta-hydroxylase	Mus musculus	570	high-mannose	998	FALSE	omannose	4
Q62469	Integrin alpha-2	Mus musculus	472	high-mannose	3115	FALSE	omannose	5
Q64487	Receptor-type tyrosine-protein phosphatase delta	Mus musculus	724	high-mannose	1530	FALSE	omannose	5

Q66PY1	Signal peptide, CUB and EGF-like domain-containing protein 3	Mus musculus	756	high-mannose	3115	FALSE	omannose	4
Q61838	Pregnancy zone protein	Mus musculus	157	complex	3649	FALSE	biantennary	4
Q61646	Haptoglobiantennaryn	Mus musculus	256	complex	3649	FALSE	biantennary	8
Q61490	CD166 antigen	Mus musculus	167	high-mannose	3115	FALSE	omannose	6
Q64449	C-type mannose receptor 2	Mus musculus	1133	high-mannose	1530	FALSE	omannose	2
Q61851	Fibroblast growth factor receptor	Mus musculus	96	high-mannose	3206	FALSE	omannose	12
Q61625	Glutamate receptor ionotropic, delta-2	Mus musculus	426	high-mannose	3115	FALSE	omannose	3
Q640R3	Hepatocyte cell adhesion molecule	Mus musculus	138	high-mannose	3115	FALSE	omannose	1
Q61739	Integrin alpha-6	Mus musculus	284	high-mannose	1530	FALSE	omannose	1
Q61576	Peptidyl-prolyl cis-trans isomerase FKBP10	Mus musculus	69	high-mannose	1530	FALSE	omannose	9
Q64455	Receptor-type tyrosine-protein phosphatase eta	Mus musculus	145	high-mannose	3115	FALSE	omannose	6
Q03137	Ephrin type-A receptor 4	Mus musculus	408	high-mannose	3115	FALSE	omannose	11
Q03173	Protein enabled homolog	Mus musculus	43	high-mannose	3115	FALSE	omannose	1
Q07456	Protein ambp	Mus musculus	233	complex	3649	FALSE	biantennary	1
P39060	Collagen alpha-1(XVIII) chain / Endostatin	Homo sapiens	129	high-mannose	3206	FALSE	omannose	2
P40238	Thrombopoietin receptor	Homo sapiens	117	high-mannose	8542	FALSE	omannose	10
P43652	Afamin	Homo sapiens	109	complex	3471	FALSE	biantennary	8

P43652	Afamin	Homo sapiens	383	complex	3471	FALSE	biantennary	15
P43251	biantennaryotinidase	Homo sapiens	150	complex	1523	FALSE	biantennary	1
P42892	Endothelin-converting enzyme 1	Homo sapiens	166	high-mannose	3115	FALSE	omannose	14
P41234	ATP-biantennarynding cassette sub-family A member 2	Mus musculus	1678	high-mannose	1710	FALSE	omannose	11
P39038	Cadherin-4	Mus musculus	658	high-mannose	1530	FALSE	omannose	9
P39087	Glutamate receptor ionotropic, kainate 2	Mus musculus	378	high-mannose	1530	FALSE	omannose	1
Q14515	Sparc-like protein 1	Homo sapiens	176	complex	1519	FALSE	biantennary	9
Q13753	Laminin subunit gamma-2	Homo sapiens	342	high-mannose	3115	FALSE	omannose	14
Q13201	Multimerin-1	Homo sapiens	1020	high-mannose	242	FALSE	omannose	10
Q14624	Inter-alpha-trypsin inhibiantennarytor heavy chain h4	Homo sapiens	517	complex	3353	FALSE	triantennary	1
P82349	Beta-sarcoglycan	Mus musculus	213	high-mannose	3115	FALSE	omannose	2
P82347	Delta-sarcoglycan	Mus musculus	108	high-mannose	3115	FALSE	omannose	6
Q00651	Integrin alpha-4	Mus musculus	145	high-mannose	3115	FALSE	omannose	1
Q01339	Beta-2-glycoprotein 1	Mus musculus	162	complex	430	FALSE	biantennary	1
Q02083	N-acylethanolamine-hydrolyzing acid amidase	Homo sapiens	37	high-mannose	1530	FALSE	omannose	12
Q02809	Procollagen-lysine,2-oxoglutarate 5-dioxygenase 1	Homo sapiens	197	high-mannose	1530	FALSE	omannose	8
P97449	Aminopeptidase N	Mus musculus	106	high-mannose	1530	FALSE	omannose	11

Q01097	Glutamate receptor ionotropic, NMDA 2B	Mus musculus	341	high-mannose	1530	FALSE	omannose	12
P97952	Sodium channel subunit beta-1	Mus musculus	135	high-mannose	1530	FALSE	omannose	7
P33146	Cadherin-15	Mus musculus	575	high-mannose	3115	FALSE	omannose	1
P35436	Glutamate receptor ionotropic, NMDA 2A	Mus musculus	340	high-mannose	3115	FALSE	omannose	3
P35613	Basigin	Homo sapiens	160	high-mannose	1163	FALSE	omannose	2
P35503	UDP-glucuronosyltransferase 1-3	Homo sapiens	348	high-mannose	1530	FALSE	omannose	1
P35504	UDP-glucuronosyltransferase 1-5	Homo sapiens	348	high-mannose	1530	FALSE	omannose	13
P35438	Glutamate receptor ionotropic, NMDA 1	Mus musculus	203	high-mannose	1530	FALSE	omannose	6
P23219	Prostaglandin G/H synthase 1	Homo sapiens	67	high-mannose	1710	FALSE	omannose	1
P23229	Integrin alpha-6	Homo sapiens	323	high-mannose	1530	FALSE	omannose	19
P24821	Tenascin	Homo sapiens	1093	complex	1523	FALSE	biantennary	7
P24821	Tenascin	Homo sapiens	1392	complex	1519	FALSE	biantennary	6
P24821	Tenascin	Homo sapiens	166	complex	1523	FALSE	biantennary	2
P24821	Tenascin	Homo sapiens	184	complex	1519	FALSE	biantennary	5
P24821	Tenascin	Homo sapiens	327	complex	1523	FALSE	biantennary	4
P22309	UDP-glucuronosyltransferase 1-1	Homo sapiens	347	high-mannose	1530	FALSE	omannose	13
P22897	Macrophage mannose receptor 1	Homo sapiens	104	complex	357	FALSE	biantennary	4

P22897	Macrophage mannose receptor 1	Homo sapiens	1205	complex	1519	FALSE	biantennary	2
P22897	Macrophage mannose receptor 1	Homo sapiens	344	complex	1519	FALSE	biantennary	3
P56974	Pro-neuregulin-2, membrane-bound isoform	Mus musculus	186	high-mannose	998	FALSE	omannose	15
P55058	Phospholipid transfer protein	Homo sapiens	143	complex	1523	FALSE	biantennary	14
P55058	Phospholipid transfer protein	Homo sapiens	64	complex	1523	FALSE	biantennary	2
P59823	Interleukin-1 receptor accessory protein-like 1	Mus musculus	122	high-mannose	3115	FALSE	omannose	6
P15088	Mast cell carboxypeptidase A	Homo sapiens	255	high-mannose	3115	FALSE	omannose	5
P13688	Carcinoembryonic antigen-related cell adhesion molecule 1	Homo sapiens	197	high-mannose	1530	FALSE	omannose	4
P14625	Endoplasmin	Homo sapiens	217	high-mannose	1443	TRUE	omannose	1
P15209	BDNF/NT-3 growth factors receptor	Mus musculus	121	high-mannose	1530	FALSE	omannose	8
P15116	Cadherin-2	Mus musculus	190	high-mannose	3115	FALSE	omannose	1
P14151	L-selectin	Homo sapiens	104	high-mannose	1163	FALSE	omannose	8
P15144	Aminopeptidase n	Homo sapiens	128	high-mannose	3115	FALSE	omannose	8
Q8BGQ6	EF-hand calcium-biantennarynding domain-containing protein 14	Mus musculus	294	high-mannose	3115	FALSE	omannose	4
Q8BS35	Alkylglycerol monooxygenase	Mus musculus	9	high-mannose	1530	FALSE	omannose	2
Q8BQ86	Carbohydrate sulfotransferase 8	Mus musculus	360	high-mannose	1710	FALSE	omannose	1

Q8BHJ7	Gamma-aminobutyric acid receptor subunit alpha-5	Mus musculus	145	high-mannose	1530	FALSE	omannose	4
Q810U4	Neuronal cell adhesion molecule	Mus musculus	1003	high-mannose	8542	FALSE	omannose	5
Q8BM13	Noelin-2	Mus musculus	269	high-mannose	1530	FALSE	omannose	5
Q8BG19	Transmembrane and TPR repeat-containing protein 4	Mus musculus	497	high-mannose	1530	FALSE	omannose	10
P16144	Integrin beta-4	Homo sapiens	695	high-mannose	1163	FALSE	omannose	4
P18052	Receptor-type tyrosine-protein phosphatase alpha	Mus musculus	47	high-mannose	3206	FALSE	omannose	2
P15848	Arylsulfatase B	Homo sapiens	188	high-mannose	1530	FALSE	omannose	1
P16671	Platelet glycoprotein 4	Homo sapiens	321	complex	1063	FALSE	biantennary	7
P16301	Phosphatidylcholine-sterol acyltransferase	Mus musculus	108	high-mannose	1530	FALSE	omannose	5
P16546	Spectrin alpha chain, triantennary-erythrocytic 1	Mus musculus	1051	high-mannose	1530	FALSE	omannose	4
P61812	Transforming growth factor beta-2	Homo sapiens	140	high-mannose	3115	FALSE	omannose	2
P68500	Contactin-5	Mus musculus	539	high-mannose	3115	FALSE	omannose	1
P70208	Plexin-A3	Mus musculus	1074	high-mannose	3115	FALSE	omannose	3
P61315	Galactose-3-O-sulfotransferase 3	Mus musculus	110	high-mannose	1710	FALSE	omannose	10
P63080	Gamma-aminobutyric acid receptor subunit beta-3	Mus musculus	105	high-mannose	3115	FALSE	omannose	12
P70207	Plexin-A2	Mus musculus	163	high-mannose	1530	FALSE	omannose	9
P70665	Sialate O-acetylesterase	Mus musculus	356	high-mannose	3115	FALSE	omannose	15

P70663	SPARC-like protein 1	Mus musculus	168	high-mannose	1710	FALSE	omannose	3
P78324	Tyrosine-protein phosphatase triantennary-receptor type substrate 1	Homo sapiens	292	high-mannose	242	FALSE	omannose	6
Q3UMW8	Ceroid-lipofuscinosis neuronal protein 5 homolog	Mus musculus	254	high-mannose	3115	FALSE	omannose	10
Q58D62	Fetuin-b	Bos taurus	137	complex	3353	FALSE	biantennary	7
Q3V3R4	Integrin alpha-1	Mus musculus	1102	high-mannose	1530	FALSE	omannose	10
Q3UH93	Plexin-D1	Mus musculus	1019	high-mannose	1530	FALSE	omannose	2
Q3TVI8	Pre-B-cell leukemia transcription factor-interacting protein 1	Mus musculus	455	high-mannose	1530	FALSE	omannose	6
Q3UNZ8	Quitriantennarye oxidoreductase-like protein 2	Mus musculus	281	high-mannose	1163	FALSE	omannose	2
Q3UTY6	Thrombospondin type-1 domain-containing protein 4	Mus musculus	441	high-mannose	3115	FALSE	omannose	4
Q3U3W2	Transmembrane protein 181a	Mus musculus	140	high-mannose	1710	FALSE	omannose	9
P04062	Glucosylceramidase	Homo sapiens	185	complex	1061	FALSE	biantennary	5
P04062	Glucosylceramidase	Homo sapiens	309	complex	1519	FALSE	biantennary	1
P04066	Tissue alpha-L-fucosidase	Homo sapiens	268	complex	1061	FALSE	biantennary	2
P03952	Plasma kallikrein	Homo sapiens	396	complex	2643	FALSE	biantennary	2
P02790	Hemopexin	Homo sapiens	64	complex	3041	FALSE	triantennary	2
P02790	Hemopexin	Homo sapiens	246	complex	32	FALSE	biantennary	1
P03951	Coagulation factor XI	Homo sapiens	126	complex	1519	FALSE	triantennary	16

Q14biantennary1	Sodium/potassium/calcium exchanger 2	Mus musculus	112	high-mannose	3206	FALSE	omannose	6
Q15417	Calponin-3	Homo sapiens	240	high-mannose	1163	FALSE	omannose	6
Q15262	Receptor-type tyrosine-protein phosphatase kappa	Homo sapiens	462	complex	1523	FALSE	biantennary	4
Q15904	V-type proton ATPase subunit S1	Homo sapiens	170	high-mannose	1530	FALSE	omannose	1
Q16880	2-hydroxyacylsphingosine 1-beta-galactosyltransferase	Homo sapiens	333	high-mannose	1530	FALSE	omannose	8
Q6PKC3	Thioredoxin domain-containing protein 11	Homo sapiens	595	high-mannose	1530	FALSE	omannose	7
Q6P179	Endoplasmic reticulum aminopeptidase 2	Homo sapiens	103	high-mannose	1530	FALSE	omannose	11
Q6P4A8	Phospholipase B-like 1	Homo sapiens	366	high-mannose	3115	FALSE	omannose	4
Q6P4E1	Protein CASC4	Homo sapiens	150	high-mannose	1163	FALSE	omannose	13
Q6P5F7	Protein tweety homolog 3	Mus musculus	144	high-mannose	3115	FALSE	omannose	12
Q6KAS7	Zinc finger protein 521	Mus musculus	1048	high-mannose	3115	FALSE	omannose	7
Q6P5F6	Zinc transporter ZIP10	Mus musculus	191	high-mannose	3206	FALSE	omannose	4
Q08380	Galectin-3-biantennarynding protein	Homo sapiens	551	complex	1519	FALSE	biantennary	3
Q08380	Galectin-3-biantennarynding protein	Homo sapiens	69	complex	1519	FALSE	biantennary	9
Q08431	Lactadherin	Homo sapiens	325	complex	1519	FALSE	biantennary	3
Q0V8T7	Contactin-associated protein like 5-3	Mus musculus	282	high-mannose	3115	FALSE	omannose	6
Q08857	Platelet glycoprotein 4	Mus musculus	220	high-mannose	1710	FALSE	omannose	12

P28654	Decorin	Mus musculus	206	high-mannose	3115	FALSE	omannose	2
P27090	Transforming growth factor beta-2	Mus musculus	241	high-mannose	3206	FALSE	omannose	3
P28665	Murinoglobulin-1	Mus musculus	313	complex	3649	FALSE	biantennary	10
P28907	ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1	Homo sapiens	100	high-mannose	1163	FALSE	omannose	1
P27701	CD82 antigen	Homo sapiens	198	high-mannose	1710	FALSE	omannose	1
P28799	Progranulin	Homo sapiens	530	high-mannose	1710	FALSE	omannose	16
P26049	Gamma-aminobutyric acid receptor subunit alpha-3	Mus musculus	163	high-mannose	1530	FALSE	omannose	4
Q80ZF8	Adhesion G protein-coupled receptor B3	Mus musculus	779	high-mannose	3115	FALSE	omannose	22
Q7TSK2	Seizure protein 6	Mus musculus	396	high-mannose	3115	FALSE	omannose	6
Q7Z388	Probable C-mannosyltransferase DPY19L4	Homo sapiens	123	high-mannose	1530	FALSE	omannose	7
Q7Z739	YTH domain-containing family protein 3	Homo sapiens	175	high-mannose	1163	FALSE	omannose	5
Q7TNS7	Acid-sensing ion channel 4	Mus musculus	138	high-mannose	3206	FALSE	omannose	5
Q7TT36	Adhesion G protein-coupled receptor A3	Mus musculus	676	high-mannose	1530	FALSE	omannose	8
Q80TR1	Adhesion G protein-coupled receptor L1	Mus musculus	526	high-mannose	1530	FALSE	omannose	7
Q7TPD3	Roundabout homolog 2	Mus musculus	786	high-mannose	3115	FALSE	omannose	3
Q80YX1	Tenascin	Mus musculus	1018	high-mannose	3115	FALSE	omannose	3
Q80X71	Transmembrane protein 106B	Mus musculus	184	high-mannose	1530	FALSE	omannose	10

Q9CR23	Transmembrane protein 9	Mus musculus	38	high-mannose	1530	FALSE	omannose	4
Q9C0H2	Protein tweety homolog 3	Homo sapiens	144	high-mannose	1163	FALSE	omannose	3
Q9CZ42	ATP-dependent (S)-NAD(P)H-hydrate dehydratase	Mus musculus	236	high-mannose	1530	FALSE	omannose	14
Q9ER65	Calsyntenin-2	Mus musculus	100	high-mannose	1710	FALSE	omannose	3
Q9D6F4	Gamma-aminobutyric acid receptor subunit alpha-4	Mus musculus	144	high-mannose	1530	FALSE	omannose	2
Q9D906	Ubiantennaryquitin-like modifier-activating enzyme ATG7	Mus musculus	314	high-mannose	1710	FALSE	omannose	1
Q8C031	Leucine-rich repeat-containing protein 4C	Mus musculus	278	high-mannose	3115	FALSE	omannose	3
Q8CBC6	Leucine-rich repeat neuronal protein 3	Mus musculus	579	high-mannose	3206	FALSE	omannose	4
Q8BY89	Choline transporter-like protein 2	Mus musculus	200	high-mannose	3115	FALSE	omannose	4
Q8C129	Leucyl-cystinyl aminopeptidase	Mus musculus	145	high-mannose	1530	FALSE	omannose	1
Q8C8T7	Protein ELFN1	Mus musculus	349	high-mannose	3115	FALSE	omannose	2
Q8C145	Zinc transporter ZIP6	Mus musculus	275	high-mannose	3115	FALSE	omannose	4
P01019	Angiotensinogen	Homo sapiens	170	complex	3353	FALSE	biantennary	9
P01019	Angiotensinogen	Homo sapiens	304	complex	2642	FALSE	biantennary	2
P01019	Angiotensinogen	Homo sapiens	47	complex	3507	FALSE	biantennary	2
P01042	Kininogen-1	Homo sapiens	294	complex	3353	FALSE	triantennary	1
P01011	Alpha-1-antichymotrypsin	Homo sapiens	271	complex	1519	FALSE	triantennary	1
P01009	Alpha-1-antitrypsin	Homo sapiens	70	complex	2011	FALSE	biantennary	4

P01009	Alpha-1-antitrypsin	Homo sapiens	107	complex	3471	FALSE	triantennary	5
P00736	Complement c1r subcomponent	Homo sapiens	514	complex	3353	FALSE	biantennary	1
P01023	Alpha-2-macroglobulin	Homo sapiens	1424	complex	2575	FALSE	triantennary	1
P01023	Alpha-2-macroglobulin	Homo sapiens	869	complex	231	FALSE	biantennary	12
P00738	Haptoglobiantennaryn	Homo sapiens	211	complex	3353	FALSE	triantennary	2
Q8IYK4	Procollagen galactosyltransferase 2	Homo sapiens	382	high-mannose	1530	FALSE	omannose	6
Q8K298	Actin-biantennarynding protein anillin	Mus musculus	555	high-mannose	3115	FALSE	omannose	2
Q8NBL1	Protein O-glucosyltransferase 1	Homo sapiens	204	high-mannose	1530	FALSE	omannose	2
Q8K209	G-protein coupled receptor 56	Mus musculus	148	high-mannose	242	FALSE	omannose	11
Q8K297	Procollagen galactosyltransferase 1	Mus musculus	91	high-mannose	1530	FALSE	omannose	7
Q96FE5	Leucine-rich repeat and immunoglobulin-like domain-containing nogo receptor-interacting protein 1	Homo sapiens	144	high-mannose	1163	FALSE	omannose	7
Q96AQ6	Pre-B-cell leukemia transcription factor-interacting protein 1	Homo sapiens	455	high-mannose	3206	FALSE	omannose	2
Q96PD5	N-acetylmuramoyl-l-alanine amidase	Homo sapiens	485	complex	3353	FALSE	biantennary	9
Q9NRJ7	Protocadherin beta-16	Homo sapiens	567	high-mannose	1710	FALSE	omannose	5
Q9NZC9	SWI/SNF-related matrix-associated actin-dependent	Homo sapiens	203	complex	8883	FALSE	biantennary	5

	regulator of chromatin subfamily A-like protein 1							
Q9NZP8	Complement C1r subcomponent-like protein	Homo sapiens	242	complex	1523	FALSE	biantennary	2
Q9NZ08	Endoplasmic reticulum aminopeptidase 1	Homo sapiens	414	high-mannose	1530	FALSE	omannose	6
Q9NUN5	Probable lysosomal cobalamin transporter	Homo sapiens	78	high-mannose	1530	FALSE	omannose	5
Q9JKR6	Hypoxia up-regulated protein 1	Mus musculus	515	high-mannose	1530	FALSE	omannose	11
Q9JJX6	P2X purinoceptor 4	Mus musculus	208	high-mannose	1710	FALSE	omannose	14
Q9JIS5	Synaptic vesicle glycoprotein 2A	Mus musculus	498	high-mannose	3206	FALSE	omannose	5
Q925F2	Endothelial cell-selective adhesion molecule	Mus musculus	111	high-mannose	3115	FALSE	omannose	16
Q920V1	UDP-GalNAc:beta-1,3-N-acetylgalactosaminyltransferase 1	Mus musculus	198	high-mannose	1710	FALSE	omannose	1
Q92859	Neogenin	Homo sapiens	470	high-mannose	1163	FALSE	omannose	9
Q92626	Peroxidasin homolog	Homo sapiens	1178	high-mannose	1710	FALSE	omannose	4
Q92854	Semaphorin-4D	Homo sapiens	419	high-mannose	1710	FALSE	omannose	3
Q9Z2A9	Gamma-glutamyltransferase 5	Mus musculus	303	high-mannose	3115	FALSE	omannose	2
Q9Z1M0	P2X purinoceptor 7	Mus musculus	202	high-mannose	1710	FALSE	omannose	4
Q9UJ14	Gamma-glutamyltransferase 7	Homo sapiens	394	high-mannose	3206	FALSE	omannose	6
Q9UN67	Protocadherin beta-10	Homo sapiens	567	high-mannose	1710	FALSE	omannose	1
Q9UKY4	Protein O-mannosyl-transferase 2	Homo sapiens	583	high-mannose	1710	FALSE	omannose	1

Q6W4X9	Mucin-6	Homo sapiens	2366	high-mannose	1163	FALSE	omannose	7
Q6UX06	Olfactomedin-4	Homo sapiens	72	complex	1523	FALSE	biantennary	6
Q6UX06	Olfactomedin-4	Homo sapiens	136	complex	1523	FALSE	biantennary	8
Q8WVJ2	NudC domain-containing protein 2	Homo sapiens	144	high-mannose	1163	FALSE	omannose	21
Q8WTV0	Scavenger receptor class B member 1	Homo sapiens	227	high-mannose	3206	FALSE	omannose	6
Q8TDL5	BPI fold-containing family B member 1	Homo sapiens	401	high-mannose	1163	FALSE	omannose	3
Q8TEM1	Nuclear pore membrane glycoprotein 210	Homo sapiens	1441	high-mannose	1530	FALSE	omannose	8
Q8TCT8	Signal peptide peptidase-like 2A	Homo sapiens	149	high-mannose	1530	FALSE	omannose	5
Q91YY2	Beta-1,4-galactosyltransferase 3	Mus musculus	387	high-mannose	1530	FALSE	omannose	3
Q8VEM1	E3 ubiantennaryquitin-protein ligase RNF130	Mus musculus	135	high-mannose	1530	FALSE	omannose	1
Q8VI51	VPS10 domain-containing receptor SorCS3	Mus musculus	797	high-mannose	3115	FALSE	omannose	10
Q9P2W7	Galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase 1	Homo sapiens	140	high-mannose	3206	FALSE	omannose	8
Q9R0Q6	Actin-related protein 2/3 complex subunit 1A	Mus musculus	296	high-mannose	1530	FALSE	omannose	6
Q9R0A1	Chloride channel protein 2	Mus musculus	411	high-mannose	3115	FALSE	omannose	5
Q9QY40	Plexin-B3	Mus musculus	469	high-mannose	3115	FALSE	omannose	4
Q9R0B9	Procollagen-lysine,2-oxoglutarate 5-dioxygenase 2	Mus musculus	725	high-mannose	3206	FALSE	omannose	12

Q9UBX1	Cathepsin F	Homo sapiens	195	high-mannose	1163	FALSE	omannose	2
Q9UBV2	Protein sel-1 homolog 1	Homo sapiens	195	high-mannose	1710	FALSE	omannose	2
Q9P2J2	Protein turtle homolog A	Homo sapiens	188	high-mannose	1710	FALSE	omannose	1
Q9UBS9	SUN domain-containing ossification factor	Homo sapiens	202	high-mannose	1710	FALSE	omannose	4
Q9QUN9	Dickkopf-related protein 3	Mus musculus	106	high-mannose	1530	FALSE	omannose	7
Q9QY81	Nuclear pore membrane glycoprotein 210	Mus musculus	1135	high-mannose	1530	FALSE	omannose	4
P20645	Cation-dependent mannose-6-phosphate receptor	Homo sapiens	57	high-mannose	1530	FALSE	omannose	3
P21661	Neuroendocrine convertase 2	Mus musculus	374	high-mannose	1530	FALSE	omannose	2
Q99PI8	Reticulon-4 receptor	Mus musculus	179	high-mannose	3115	FALSE	omannose	2
Q96RQ9	L-amino-acid oxidase	Homo sapiens	134	high-mannose	1530	FALSE	omannose	8
Q99523	Sortilin	Homo sapiens	98	complex	9425	FALSE	biantennary	7
Q99PH1	Leucine-rich repeat-containing protein 4	Mus musculus	362	high-mannose	3115	FALSE	omannose	5
Q99102	Recombiantennarynant Mucin-1 Muc1f/4tr	Homo sapiens	1647	complex	1523	FALSE	biantennary	5
Q99102	Recombiantennarynant Mucin-1 Muc1f/4tr	Homo sapiens	2049	complex	1519	FALSE	biantennary	10
P04217	Alpha-1b-glycoprotein	Homo sapiens	179	complex	3353	FALSE	biantennary	10
P04217	Alpha-1b-glycoprotein	Homo sapiens	371	complex	3353	FALSE	biantennary	4
P04217	Alpha-1b-glycoprotein	Homo sapiens	363	complex	3507	FALSE	biantennary	23
P05156	Complement factor i	Homo sapiens	177	complex	1523	FALSE	biantennary	1

P05155	Plasma protease c1 inhibiantennarytor	Homo sapiens	352	complex	3471	FALSE	biantennary	6
P05155	Plasma protease c1 inhibiantennarytor	Homo sapiens	253	complex	3353	FALSE	biantennary	3
P05543	Thyroxine-biantennarynding globulin	Homo sapiens	36	complex	3471	FALSE	biantennary	5
P07996	Thrombospondin-1	Homo sapiens	360	complex	1519	FALSE	biantennary	18
Q9Y337	Kallikrein-5	Homo sapiens	208	high-mannose	3206	FALSE	omannose	2
Q9Y5F2	Protocadherin beta-11	Homo sapiens	567	high-mannose	1710	FALSE	omannose	1
Q9Y5E9	Protocadherin beta-14	Homo sapiens	567	high-mannose	1710	FALSE	omannose	9
Q9Y5E6	Protocadherin beta-3	Homo sapiens	567	high-mannose	1710	FALSE	omannose	18
Q9Y5E4	Protocadherin beta-5	Homo sapiens	566	high-mannose	1710	FALSE	omannose	3
Q9Y5E1	Protocadherin beta-9	Homo sapiens	567	high-mannose	1710	FALSE	omannose	11
Q9Y5L3	Ectonucleoside triphosphate diphosphohydrolase 2	Homo sapiens	129	high-mannose	1530	FALSE	omannose	5
Q9Y639	Neuroplastin	Homo sapiens	197	high-mannose	1530	FALSE	omannose	4
Q9Y5G3	Protocadherin gamma-B1	Homo sapiens	541	high-mannose	1163	FALSE	omannose	1
Q9HCN3	Transmembrane protein 8A	Homo sapiens	535	high-mannose	1710	FALSE	omannose	6
Q9JHJ3	Glycosylated lysosomal membrane protein	Mus musculus	133	high-mannose	3115	FALSE	omannose	2
Q9HAW7	UDP-glucuronosyltransferase 1-7	Homo sapiens	344	high-mannose	1530	FALSE	omannose	1

Publications Released as part of the PhD

The following pages list all publications in their original '.pdf' format that were released as part of this PhD, in the order in which they were published.

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Structural glycobiology in the age of electron cryo-microscopy

Mihaela Atanasova, Haroldas Bagdonas and Jon Agirre

The methodology underpinning the construction, refinement, validation and analysis of atomic models of glycoproteins and protein-carbohydrate complexes has received a long-overdue boost in the last five years. This is a very timely development, as the resolution revolution in electron cryo-microscopy is now routinely delivering structures of key glycomedical importance, with a three-dimensional precision where X-ray crystallographic methods have traditionally floundered. This review will focus on the new software developments that have been introduced in the past two years, and their impact on the field of structural glycobiology in terms of published structures.

Address
York Structural Biology Laboratory, Department of Chemistry, University of York, UK

Corresponding author: Agirre, Jon (jon.agirre@york.ac.uk)

Current Opinion in Structural Biology 2020, **62**:70–78

This review comes from a themed issue on **Carbohydrates**

Edited by **Sony Malhotra** and **Paul A Ramsland**

For a complete overview see the Issue and the Editorial

Available online 23rd December 2019

https://doi.org/10.1016/j.sbi.2019.12.003

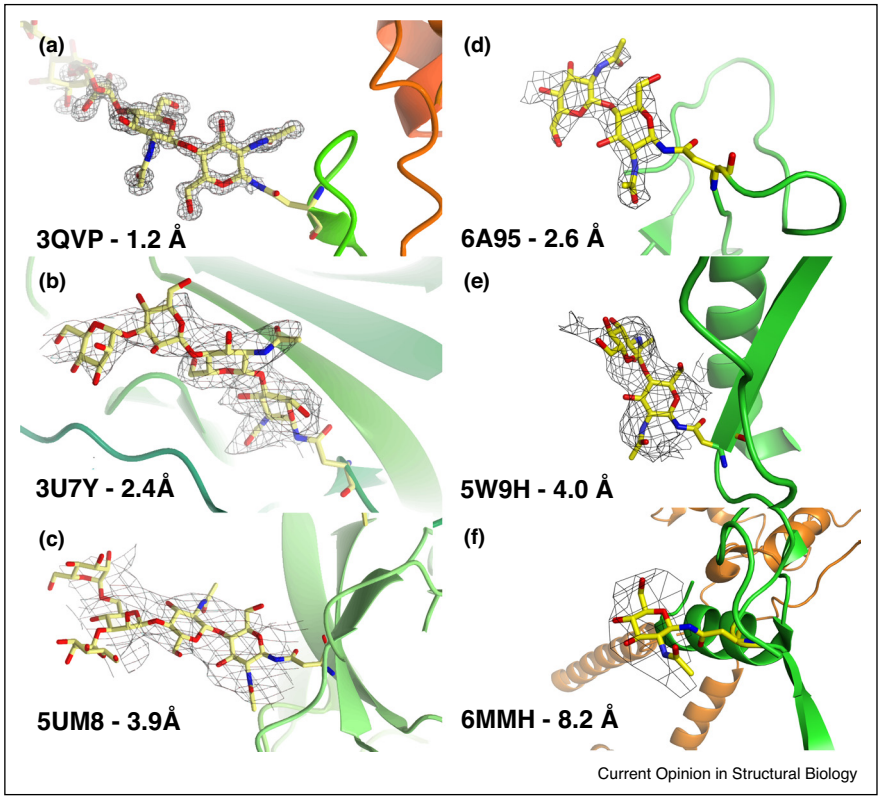
Introduction

Protein glycosylation plays a crucial role in recognition processes in, for example, viral infection, cancer, fertilisation, immunity and inflammation [1]. In this role, glycans are expected to provide stabilising contacts within the buried surface of a glycoprotein, while additionally playing a role as interaction partners on the surface, via hydrogen bonds or CH–π interactions. As independent entities, carbohydrates also have promising biotechnological applications, being a staple in the production of more eco-friendly second-generation biofuels from previously untractable crop waste. Assisting in this task, carbohydrate-active enzymes recognise, transfer and cut saccharide building blocks, often distorting individual rings to achieve catalysis.

Complicated stereochemistry, branching and unpredictable sequence/structure make protein glycosylation in particular harder to work with than pure protein, or even nucleic acid. Perhaps unsurprisingly, the software for handling structures of carbohydrate moieties are not yet as feature-rich as that for other biomolecules. This gap in capabilities becomes evident in both macromolecular crystallography (MX) and electron cryo-microscopy (cryo-EM) whenever the model fitting problem deviates from standard propositions. Indeed, at high-resolution it is possible to identify a monosaccharide and ascertain its ring conformation (Figure 1a) — to date, this has only been possible with X-ray crystallography. Nevertheless, we fully expect cryo-EM to reach this level of precision in the near future. As resolution decreases, it becomes increasingly difficult to determine its ring conformation — thus requiring additional restraints for idealising ring puckering (Figure 1b–f) [2]. Finally, at low resolution, usually neither the monosaccharide nor its conformation can be identified (Figure 1c–f). It is in this particular case where the articulation of prior glycochemical knowledge must cross boundaries from the realm of validation, and play a central role in the structure building process: lowest energy ring conformations, a constant in pyranosides except in rare cases (catalysis is one of them), can be enforced using unimodal torsion restraints; the most probable linkage types, which should match the expression system's available glycosyltransferases, can be modelled using automated tools (*vide infra*); low energy glycosidic linkage orientations can be encouraged by using information from homologous structures via external restraints. As with protein methodology, whatever prior information is useful for validation at high resolution – for example, the Ramachandran criterion – can be turned into restraints for refinement at low resolution – for example, Ramachandran restraints. In becoming a target for refinement, validation metrics lose independence; yet as part of a balance between experimental and geometric terms, they are still useful as validation criteria – for example, ideal bond lengths and angles are also used both as restraints in refinement, and as a measure of distortion particularly for ligands. It is ultimately the structural biologist's choice whether they want to produce the best possible structure, or have a measure of how correct it is.

Experimentally, it is clear that the mobility of the glycans poses a problem for both MX and cryo-EM, with Nuclear Magnetic Resonance (NMR) providing much of the insight into protein–carbohydrate interactions due to the degrading resolvability of the sugars down the glycans' branches [3••] typically found with the two former techniques. In contrast, most of the challenges present in software spring from the

Figure 1



Comparison of N-glycan features in electron density maps over a range of resolutions. **(a)–(c)** Electron density maps obtained with X-ray crystallography (MX). **(d)–(f)** Electronic potential maps obtained with cryo-EM; PDB codes and data resolution have been annotated directly on the figure. In the MX cases (a)–(c), at high resolution it is possible to identify monosaccharides and their ring conformation from the density map; at medium resolution, ring conformation becomes difficult to determine, whereas at low resolution, and indeed with many cryo-EM maps (d)–(f), a modelled N-glycan should always be backed by prior glyco-chemical knowledge: lowest energy ring conformations, most probable linkage types considering the expression system's available glycosyltransferases, and low energy glycosidic linkage orientations.

particularities of carbohydrate chemistry. Upon cyclisation, there are two choices for the orientation of the anomeric hydroxy group, which leads to two anomeric forms – alpha or beta (refer to Ref. [4•] for a graphical description). Most D-sugar pyranoses adopt the $^{4}C_{1}$ conformation, while most L-sugar pyranoses adopt the $^{1}C_{4}$ conformation. Interconversion of pyranose rings between different conformations requires an itinerary, which can be described using the Cremer–Pople sphere [5]. The two chair conformations, $^{4}C_{1}$ and $^{1}C_{4}$ are optimal because of the 60/−60 degree torsion angle between substituents, leaving them staggered instead of eclipsed. Conversion from $^{4}C_{1}$ to $^{1}C_{4}$ and vice versa requires jumping over a very high energy barrier, and normally would involve catalysis, which can be achieved with the help of a carbohydrate active enzyme [4•,6].

Carbohydrate residue nomenclature is challenging for several reasons, including the two different types of glycosidic linkages (alpha or beta), branching and ring contortions. Lutteke *et al.* [7] first reported that about 30% of the deposited carbohydrate structures contain one or more nomenclature errors, a finding that gave rise to carbohydrate validation software, recently reviewed in Refs. [8••,9••]. A few years later, Crispin *et al.* also criticised the lack of methodological support for carbohydrates, singling out a deposited structure with a glycosidic linkage for which there were no available glycosyltransferases along its biosynthetic pathway [10,11]. More recently, Agirre *et al.* [2] performed an analysis on all N-glycan forming D-pyranosides found in the PDB using the Privateer software (CCP4 suite [12]): as data resolution decreases, more and more sugar monomers appear in high-energy conformations and/or have low real-space correlation. This indicated the need for using appropriate restraints during refinement.

In this review, we shall go through the latest software developments and their application to solving real-world structures, placing an emphasis on their impact on the recent evolution of electron cryo-microscopy into an all-around player in the structural glycobiology field. Aside from the growing access to automated, integrative model

building and validation tools, a number of online support resources are available to the structural glycobiologist too: see Refs. [13,14] for a review of online resources, and Perez and De Sanctis [15•] for a recent summary of the resources and techniques available where a synchrotron light source is available.

Dictionaries: the book of chemical knowledge

The model building process involves macromolecular refinement programs deriving geometric restraints from libraries of dictionaries, at least for most commonly occurring monomers. Dictionaries are used to store prior chemical knowledge about compounds, including their composition, connectivity and stereochemistry. The CCP4 Monomer Library, one of the first examples of its kind, was based on the geometry proposed by Engh and Huber [16], which is now outdated particularly concerning sugars [4•]. If a chemical compound does not have a library entry, or if it is incorrect, a new one needs to be generated. There are several programs that can be used for this, with irregular results for carbohydrates [4•]. The CCP4 program ACEDRG [17,18] works by mining databases such as the Crystallography Open Database (COD) [18] to generate dictionaries from the data available there. It then uses RDKit (open source chemoinformatics; http://www.rdkit.org) to generate conformers which are ranked by free energy, and the minimal-energy one is chosen. ACEDRG/COD produces similar results to GRADE (Global Phasing Ltd.) and Phenix.eLBOW [19], which derive their restraints from Mogul [20], a tool that in turn mines the Cambridge Structural Database (CSD). Mogul is currently in use for geometry validation upon deposition with the Protein Data Bank, meaning that the use of old dictionaries during refinement with tight geometry targets – for example, when refining against a cryo-EM map – can produce a disproportionate number of bond length and angle outliers. A modernisation effort is currently underway in CCP4, with hundreds of carbohydrate entries being marked for update through the combination of ACEDRG and Privateer [21]. The new dictionaries have an expected release date of 2020.

Model building

The improved *N*-glycosylation building module for Coot

Coot [22] has a carbohydrate-building tool [23••] – earlier version reviewed in Ref. [9••] – that can be used to build *N*-glycosylation into both crystallographic and cryo-EM maps. The module has three modes: manual, semi-automated and automated. The manual mode allows the user to choose a monosaccharide and a bond type from a selection of commonly available glycoforms. Coot chooses the best position, orientation and conformation for the selected monosaccharide, and refines the structure. In the semi-automated mode the user selects a glycan type and Coot returns possible options for the monosaccharide and the glycosidic bond. The automated mode requires the user to simply choose the starting point and the glycosylation tree type, and Coot builds it automatically, interrupting the process when no more sugars can be built into clear density. An overview of results is presented in Figure 2 (adapted with permission from IUCr Journals). The tool has received positive adoption by the community, as shown by its use on several high-profile X-ray and cryo-EM structures with abundant protein *N*-glycosylation [24–27].

Its main limitation is the relatively narrow selection of glycoforms available. This is clearly a design decision rather than an oversight, as these represent the most common forms that can usually be resolved experimentally. Moreover, *Coot* does not include temperature-factor refinement, as all atoms are set to a fixed value. The authors suggested integrating the model-free B-factor refinement procedure described by Cowtan and Agirre [28] as an improvement.

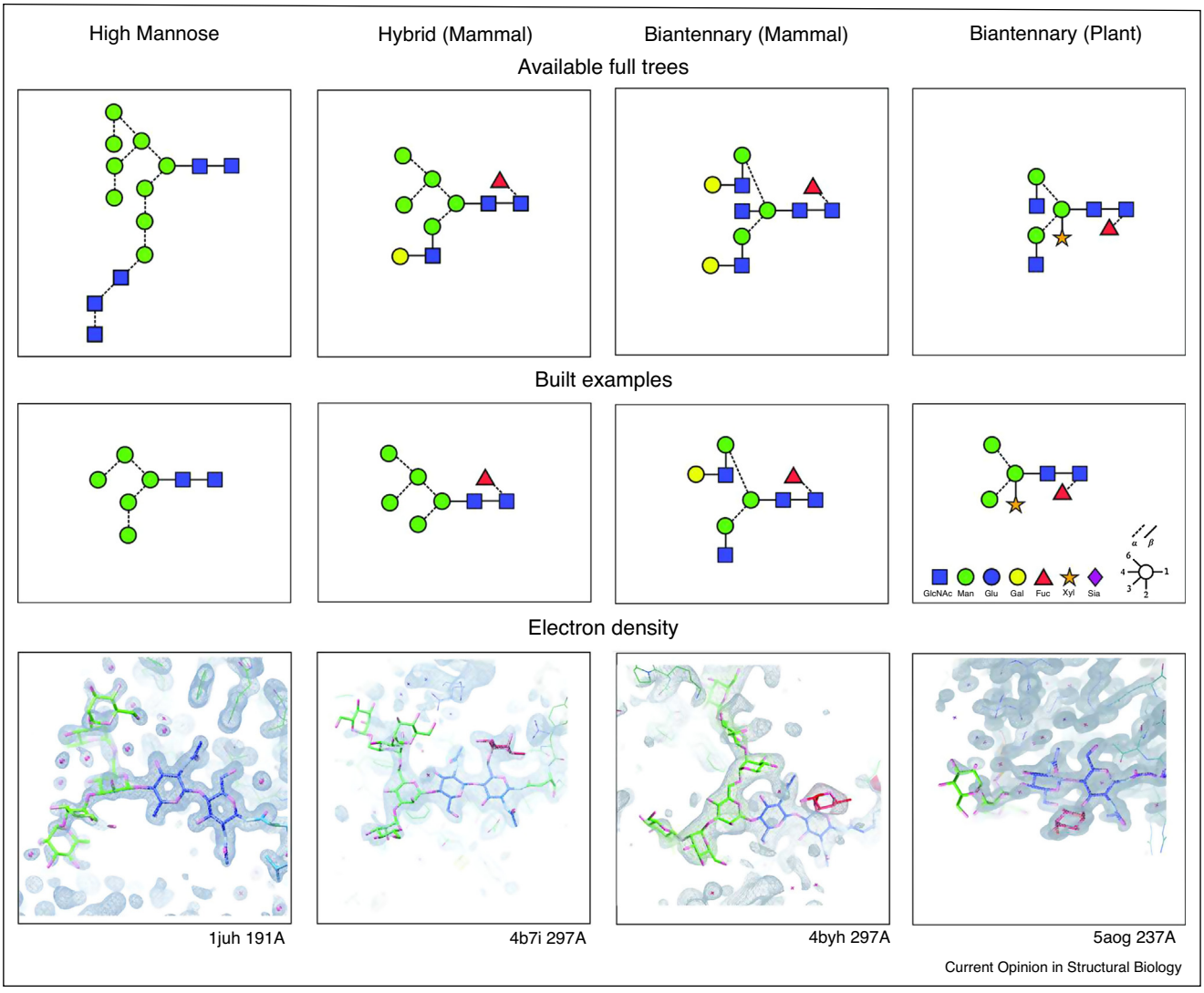
PDB-REDO: Carbivore and Carbonanza

Van Beusekom *et al.* [29•] presented a set of tools that build on the Coot *N*-glycosylation building module to achieve a more automated behaviour; indeed, the software is meant to be part of their PDB-REDO [30] rebuilding and re-refinement pipeline. The first tool they presented is *Carbivore*, which can be used to rebuild and extend existing *N*-glycosylation trees automatically, or add new trees where they are missing. For the case glycosylation was not detected due to C1 not facing the asparagine side-chain, the authors introduced another program, named Carbonanza, to generate link records. The whole-tree addition method of Coot was extended to allow for building partial trees, that is, extending existing trees. Moreover, a feature that finds *N*-glycosylation sites based on the consensus sequence Asn-X-Ser/Thr was implemented in *Carbivore*. In addition, an option for finding *N*-glycosylation sites based on homologous models was also presented; however, this is not used by default as the search is likely to be slow.

ISOLDE

The ISOLDE plugin [31] for ChimeraX [32] offers a refreshing way of dealing with protein glycosylation, and supports both electron cryo-microscopy and X-ray crystallographic data. The graphical frontend connects to an interactive, GPU-accelerated molecular mechanics simulation, updating the model – and electron density maps, if working on crystallographic data – based on both the user's push-pull movements and the results of running the simulation on the updated coordinates. Technology-wise, this new tool makes use of the OpenMM toolkit [33] for simulations, and the Clipper-python module [34] for electron density calculations, which is heavily CPU-parallelised – using C++11-style threads – in the latest version available from the *ChimeraX toolshed* at the time of publication. Protein glycosylation is handled by an adapted version of the GLYCAM force field [35]. Although at present some unwanted effects such as ring

Figure 2



Results from a test of the *N*-glycosylation building tool in Coot [23••]. The diagrams in SNFG format show the expected glycoforms and the subsets Coot was able to build automatically, while the third row of pictures shows how the maps looked like in each example. Reproduced from Ref. [23••] with permission of the International Union of Crystallography.

inversions might appear as a result of the unrealistically high temperatures simulated by the user's push–pull movements, it is clear that this tool will be of great assistance when multiple overall glycan conformations need to be evaluated in a low resolution map; a combination with real-time validation at both the monosaccharide and glycan levels could further inform the fitting process and prevent errors too. The capabilities of ISOLDE are most effectively demonstrated in the supplementary video of [31].

Sails

Sails [36] can be used to build sugars automatically, either covalently linked to protein or as ligands. The software is currently in the middle of a major infrastructural change but is slated for general release in 2020 (with, or through an update to CCP4 7.1). It uses a method similar to that of Nautilus [37] and Buccaneer [38,39], using fingerprint-based detection of fragments, which account for both the target and its environment. The correlation function behind Sails has been proven to work with electron cryo-microscopy data, although adjustments may be needed if, for example, the scale of the EM map is not accurate or different map sharpening or blurring is required. Privateer and Refmac will be integrated with Sails in a pipeline for iterative building, refinement and validation.

Refinement and validation

Privateer

Privateer [21] is a carbohydrate-specific validation tool that can determine ring conformation of furanose and

pyranose rings, anomeric form, absolute stereochemistry, real space correlation between model and omit density. In addition, Privateer generates other output such as SVG glycan diagrams in the Symbol Nomenclature For Glycans (SNFG) notation, and scripts for both Refmac5 [40] and Coot [22]. Like Sails, it is undergoing a change in infrastructure in order to future-proof its architecture.

Among the different checks that Privateer will do on carbohydrate models, a comparison of ring conformation and the ideal, minimal-energy conformation for each monosaccharide provides the fastest and most useful indication of potential mistakes in modelling and/or refinement: at high resolution, unjustified high-energy conformations – those without support of clear electron density – can reveal problems in the glycosidic bond (wrong anomer used, for instance) or wrong restraints (e.g. inverted chiralities). At low resolution, the problem can appear if the model is allowed to deviate from the ideal geometry due to providing insufficient restraints during refinement. Privateer generates dictionaries containing unimodal restraints upon detecting unjustified high-energy conformations. The validation and re-refinement process via these dictionaries is now completely automated via the CCP4*i2* interface [41]. These developments were spearheaded after it was revealed that the PDB contained an unrealistically high number of non-chairs as part of *N*-glycosylation [2].

Many newer cryo-EM structures of glycoproteins are in the 2 Å–6 Å resolution range due to improvement in electron sources, detectors, and image processing and 3D reconstruction algorithms. But the software for structure solution and validation has also improved, and perhaps as a result of that, high-resolution cryo-EM structures display fewer sugars in high-energy conformations than crystallographic ones. To illustrate this point, Privateer was run on all *N*-glycosylated structures in the PDB, solved with X-ray crystallography and cryo-EM. The decoupled results are shown in Figure 3. D-sugars are shown in blue, L-sugars are shown in yellow. Ideally, in the particular case of *N*-glycosylation all D-sugars should be in $^{4}C_{1}$ conformation, and all L-sugars in $^{1}C_{4}$ conformation.

As previously highlighted elsewhere [4•], pyranose higher-energy conformations are even more unusual than Ramachandran outliers, and should be reported alongside them in the refinement summary table.

Phenix, Rosetta and AMBER

Phenix uses a conformation-dependent library of restraints for the protein backbone [42] and homology refinement [43] for protein modeling. Rosetta can be used for carbohydrate refinement of both X-ray and cryo-EM structures using parameterisation derived from X-ray structures to approximate conformational energy [44]. Frenz *et al.* [45•] developed a protocol that can use either low-resolution crystallographic data, through Phenix-Rosetta integration [46] or cryo-EM data.

The RosettaCarbohydrate framework includes torsion-space refinement for glycans, which assumes ideal bond lengths and angles [47]. Frenz *et al.* [45•] build on previous work by expanding Rosetta's geometry term to include bond geometry deviations. These were derived from Phenix using eLBOW with AM1 optimisation and added to the Rosetta database. Currently the sugar monomers included are alpha and beta glucose, *N*-acetyl glucosamine, alpha and beta mannose, and alpha and beta fucose.

The authors recommend using Privateer [21] before and after refinement to detect errors in the structure. For refinement of crystallography data, Rosetta's integration with Phenix can be used [48]. The protocols were modified to account for glycans, including steps for minimisation, increasing repulsive weights, and idealisation of anomeric hydrogen.

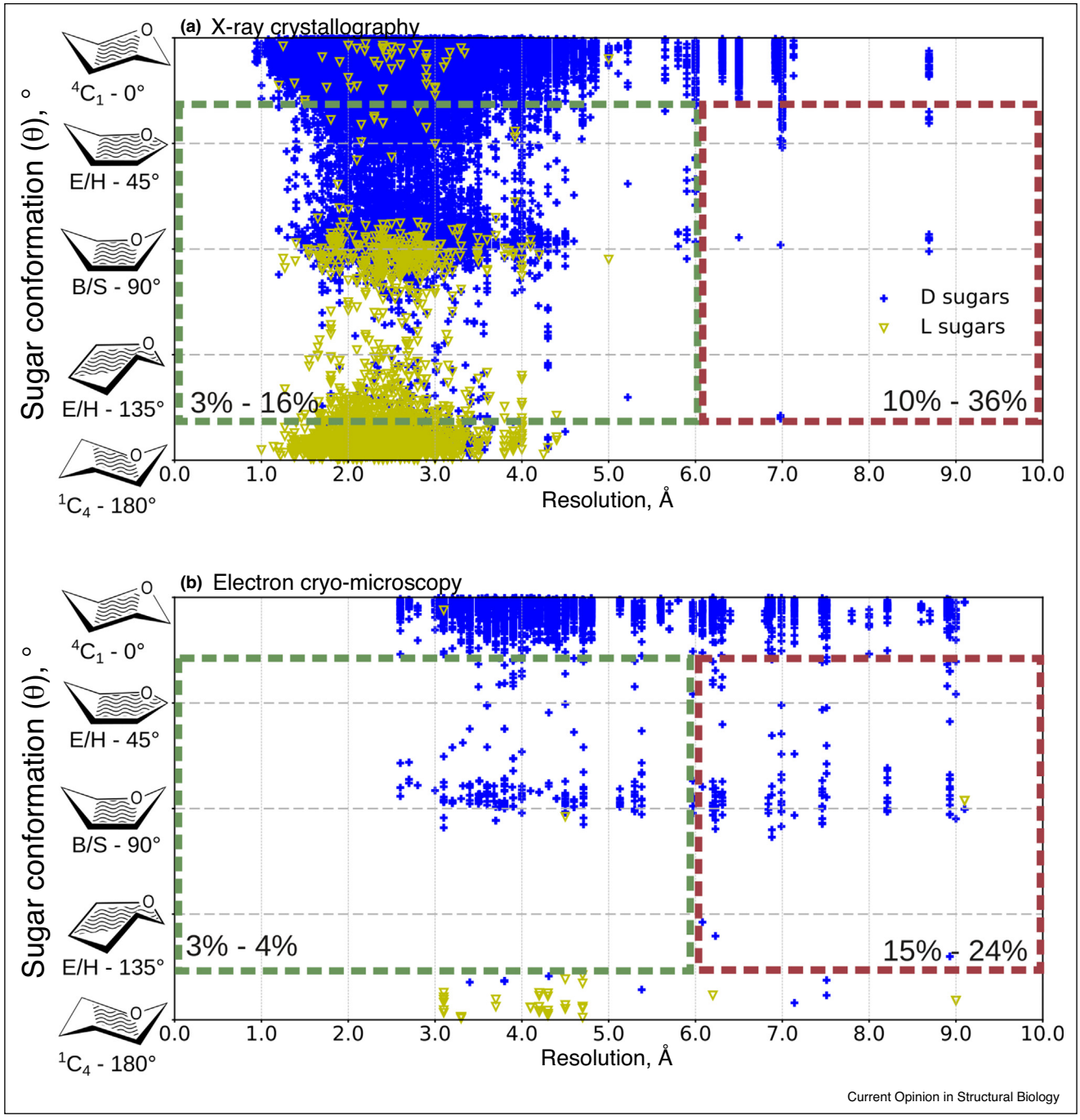
Phenix also offers integration with the AMBER molecular mechanics package, which is known for calculating torsion potentials accurately [49].

A word on legacy validation tools

While the tools outlined in this section are now sadly unsupported, it is worth mentioning them not just for the sake of completeness, but because there is no substitute tool yet for some of the key functions they provide. PDB-CARE (PDB CArbohydrate REsidue check; [50,51]) is a tool that can be used for bond and nomenclature validation. It is based on pdb2linucs, which is a software for carbohydrate detection based on atom types and their coordinates. The LINUCS notation [52] is used to normalise carbohydrate structures. This is done by comparing the carbohydrate structures' LINUCS notation to the PDB HET Group Dictionary, which contains sugar residues present in the coordinate file [50]. If a structure contains multiple anomers due to mutarotation at the reducing end of a saccharide, both forms need to have the correct PDB three-letter codes.

CARP (CArbohydrate Ramachandran Plot) is a tool that can be used to evaluate glycosidic linkage torsions. CARP also uses the pdb2linucs algorithm to analyse data, and compares it to data in GlyTorsionDB or GlycoMapsDB (for less common linkages). For each pair of monosaccharides and linkage combination, a separate torsional plot is created [7]. While these tools have been used mainly for validation purposes, they are a nice complement when examining the different linkage conformations in disaccharides [53].

Figure 3



Pyranose ring conformations versus resolution for all sugars part of N-linked glycoproteins solved with **(a)** X-ray crystallography or **(b)** electron cryo-microscopy in the PDB by April 2019. E/H: Envelopes and Half-chairs, B/S: Boats and Skew-boats. Wavy lines denote the main ring plane. For reasons of clarity, half-chair, skew-boat and envelope were omitted from the axes at $\theta = 45°$, $\theta = 90°$ and $\theta = 135°$ respectively. Percentage of sugars in non-chair conformations is shown for resolution ranges 0.0–6.0 Å and 6.01–10.0 Å.

Representation

While all-atom representations are the way to go for showing the interactions between protein and carbohydrate ligands, there is a case for using a simplified representation for glycans taking part in protein glycosylation; indeed, the sheer number of potential interactions occurring due to the size of the glycans – in optimal cases, nine or more linked monosaccharides could be visible – and the particular relevance of their composition make all-atom figures difficult or near-impossible to

follow. McNicholas and Agirre [54] introduced a representation (*Glycoblocks* for CCP4mg [55]) that, building on a 3D extension of the now standard Symbol Nomenclature For Glycans (SNFG) [6,56], added minimalistic dashed lines for hydrogen bonds and CH–π interactions.

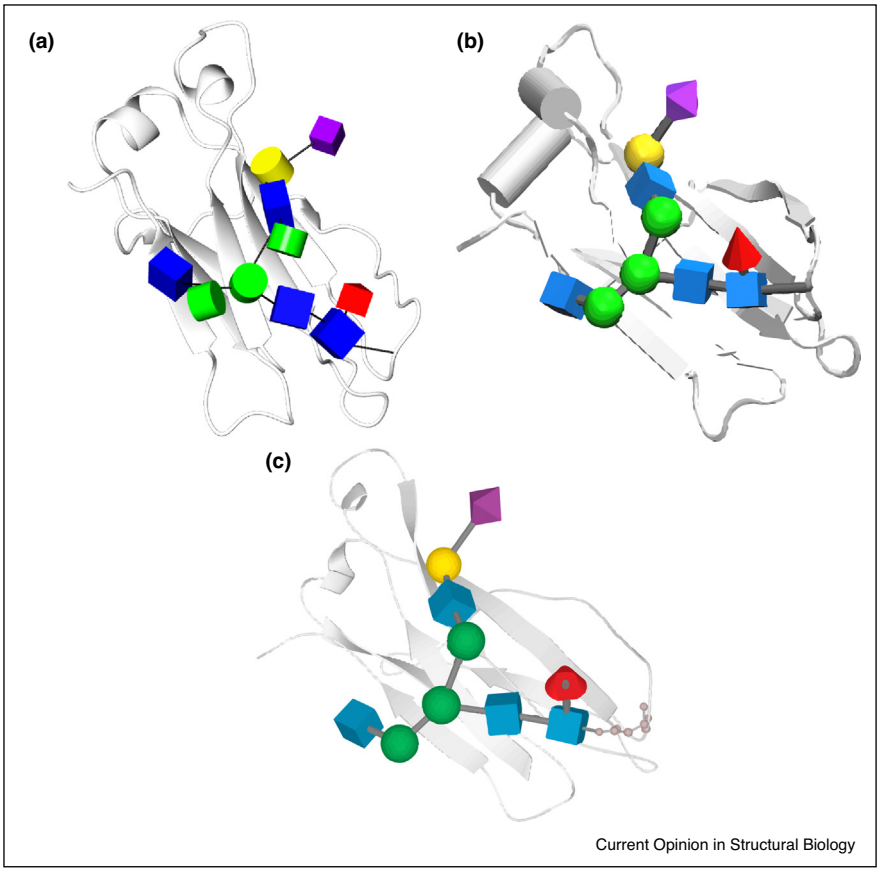
Not focusing on interactions, many 3D SNFG representations exist now either as plugins or as an integral part of wider-purpose graphics software, for example, VMD [57], LiteMol [58], and UCSF Chimera [59] via the Tangram plugin [60]. These provide stand-out depictions of protein glycosylation using big regular polyhedra. A side-by-side comparison is shown in Figure 4. Finally, other software such as SweetUnityMol [61] and Pymol [62] combine the familiar colouring scheme with a more atomistic representation.

Future perspectives

It appears the gears are finally turning in the methodological machine towards implementing better support for carbohydrates. However, software still require expert knowledge of carbohydrate structure or very high resolution to work automatically. Work is currently being done on the Sails program to be able to overcome many of these limitations. In addition, based on encouraging early results [63,64•,65,66], new carbohydrate dictionaries with more faithful model geometry and accurate torsion restraints will improve refinement, particularly for cryo-EM. Finally, sugars in active sites of enzymes might be distorted into high energy conformations, and thus may require further validation; work will need to be done in this respect in order to give users a confidence level on their conformational assignment.

We should like to emphasise that model building, refinement and validation will need to be further integrated together for maximum benefit of users. Recently, Van Beusekom, Lutteke and Joosten [8••] used a set of tools, including PDB-REDO [30], Privateer [21] and CARP [51] to analyse 8114 glycoproteins from the PDB. They succeeded in correctly re-annotating 3620 carbohydrate residues, which were then re-refined and are now available for the community to use. Incorporating prior glycochemical knowledge into the structure solution process will, as exemplified by the aforementioned authors, extend the limits of resolvability further down our glycans.

Figure 4



3D SNFG glycan representation comparison of PDB code 4BYH in selected software: **(a)** CCP4mg [53] with Glycoblocks [54], **(b)** VMD [56] and **(c)** LiteMol [57].

Conflict of interest statement

Nothing declared.

Acknowledgements

Mihaela Atanasova is funded by the UK Engineering and Physical Sciences Research Council [EPSRC, grant number EP/R513386/1]. Haroldas Bagdonas is funded by The Royal Society [grant number RGF/R1/181006]. Jon Agirre is a Royal Society University Research Fellow [award number UF160039]. We should also like to acknowledge the support – by no means limited to financial backing – of the Department of Chemistry and the University of York.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Schnaar RL: **Glycobiology simplified: diverse roles of glycan recognition in inflammation**. *J Leukoc Biol* 2016, **99**:825-838.

2. Agirre J, Davies G, Wilson K, Cowtan K: **Carbohydrate anomalies in the PDB**. *Nat Chem Biol* 2015, **11**:303.

3. •• Valverde P, Quintana JI, Santos JI, Ardá A, Jiménez-Barbero J: **Novel NMR avenues to explore the conformation and interactions of glycans**. *ACS Omega* 2019, **4**:13618-13630.
An overview of the recent advances for analysing protein–glycan interactions with Nuclear Magnetic Resonance spectroscopy, a great alternative technique when neither crystallography nor electron cryo-microscopy can resolve them.

4. • Agirre J: **Strategies for carbohydrate model building, refinement and validation**. *Acta Crystallogr Sect D Struct Biol* 2017, **73**:171-186.
An overview of the manual model building process for carbohydrates, including dictionary generation, refinement and validation.

5. Cremer D, Pople JA: **A general definition of ring puckering coordinates**. *J Am Chem Soc* 1975, **97**:1354-1358.

6. Varki A, Cummings RD, Aebi M, Packer NH, Seeberger PH, Esko JD, Stanley P, Hart G, Darvill A, Kinoshita T *et al.*: **Symbol nomenclature for graphical representations of glycans**. *Glycobiology* 2015, **25**:1323-1324.

7. Lütteke T, Frank M, Von Der Lieth CW: **Data mining the protein data bank: automatic detection and assignment of carbohydrate structures**. *Carbohydr Res* 2004:1015-1020.

8. •• Van Beusekom B, Lütteke T, Joosten RP: **Making glycoproteins a little bit sweeter with PDB-REDO**. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2018, **74**:463-472.
A study of how re-annotation and re-refinement of carbohydrate residues improve carbohydrate models.

9. •• Agirre J, Davies GJ, Wilson KS, Cowtan KD: **Carbohydrate structure: the rocky road to automation**. *Curr Opin Struct Biol* 2017, **44**:39-47.
A review of the recent software developments for carbohydrate structure solution.

10. Crispin M, Stuart DI, Yvonne Jones E: **Building meaningful models of glycoproteins**. *Nat Struct Mol Biol* 2007, **14**:354.

11. Berman HM, Henrick K, Nakamura H, Markley J: **Reply to: building meaningful models of glycoproteins**. *Nat Struct Mol Biol* 2007, **14**:354-355.

12. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A *et al.*: **Overview of the CCP4 suite and current developments**. *Acta Crystallogr Sect D Biol Crystallogr* 2011, **67**:235-242.

13. Yuriev E, Ramsland PA: **Carbohydrates in cyberspace**. *Front Immunol* 2015, **6**:300.

14. Emsley P, Brunger AT, Lütteke T: **Tools to assist determination and validation of carbohydrate 3D structure data**. *Methods Mol Biol* 2015, **1273**:229-240.

15. • Pérez S, de Sanctis D: **Glycoscience@Synchrotron: synchrotron radiation applied to structural glycoscience**. *Beilstein J Org Chem* 2017, **13**:1145-1167.
Review of the use of synchrotron radiation experiments for structure determination of glycan-interacting proteins.

16. Engh RA, Huber R: **Accurate bond and angle parameters for X-ray protein structure refinement**. *Acta Crystallogr Sect A Found Crystallogr* 1991, **47**:392-400.

17. Long F, Nicholls RA, Emsley P, Gražulis S, Merkys A, Vaitkus A, Murshudov GN: **AceDRG: a stereochemical description generator for ligands**. *Acta Crystallogr Sec D Struct Biol* 2017, **73**:112-122.

18. Long F, Nicholls RA, Emsley P, Gražulis S, Merkys A, Vaitkus A, Murshudov GN: **Validation and extraction of molecular-geometry information from small-molecule databases**. *Acta Crystallogr Sect D Struct Biol* 2017, **73**:103-111.

19. Moriarty NW, Grosse-Kunstleve RW, Adams PD: **Electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation**. *Acta Crystallogr D Biol Crystallogr* 2009, **65**:1074-1080.

20. Bruno IJ, Cole JC, Kessler M, Luo J, Momerwell WDS, Purkis LH, Smith BR, Taylor R, Cooper RI, Harris SE *et al.*: **Retrieval of crystallographically-derived molecular geometry information**. *J Chem Inf Comput Sci* 2004, **44**:2133-2144.

21. Agirre J, Iglesias-Fernández J, Rovira C, Davies GJ, Wilson KS, Cowtan KD: **Privateer: software for the conformational validation of carbohydrate structures**. *Nat Struct Mollar Biol* 2015, **22**:833-834.

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

23. •• Emsley P, Crispin M: **Structural analysis of glycoproteins: building N-linked glycans with coot**. *Acta Crystallogr Sect D Struct Biol* 2018, **74**:256-263.
A summary of the tools for building N-linked glycans available within Coot, with examples.

24. Zhu S, Noviello CM, Teng J, Walsh RM, Kim JJ, Hibbs RE: **Structure of a human synaptic GABAA receptor**. *Nature* 2018, **559**:67-72.

25. Zhang B, Wang KB, Wang W, Wang X, Liu F, Zhu J, Shi J, Li LY, Han H, Xu K *et al.*: **Enzyme-catalysed [6+4] cycloadditions in the biosynthesis of natural products**. *Nature* 2019, **568**:122-126.

26. Klünemann T, Preuß A, Adamczack J, Rosa LFM, Harnisch F, Layer G, Blankenfeldt W: **Crystal structure of Dihydro-Heme d1 dehydrogenase NirN from pseudomonas aeruginosa reveals amino acid residues essential for catalysis**. *J Mol Biol* 2019, **431**:3246-3260.

27. Lee Y, Wiriyasermkul P, Jin C, Quan L, Ohgaki R, Okuda S, Kusakizako T, Nishizawa T, Oda K, Ishitani R *et al.*: **Cryo-EM structure of the human L-type amino acid transporter 1 in complex with glycoprotein CD98hc**. *Nat Struct Mol Biol* 2019, **26**:510-517.

28. Cowtan K, Agirre J: **Macromolecular refinement by model morphing using non-atomic parameterizations**. *Acta Crystallogr D Struct Biol* 2018, **74**:125-131.

29. • van Beusekom B, Wezel N, Hekkelman ML, Perrakis A, Emsley P, Joosten RP: **Building and rebuilding N-glycans in protein structure models**. *Acta Crystallogr Sect D Struct Biol* 2019, **75**:416-425.
A set of tools incorporated in the PDB-REDO pipeline for building, rebuilding and extending glycosylation trees.

30. Joosten RP, Lütteke T: **Carbohydrate 3D structure validation**. *Curr Opin Struct Biol* 2017, **44**:9-17.

31. Croll TI: **ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps**. *Acta Crystallogr D Struct Biol* 2018, **74**:519-530.

32. Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE: **UCSF ChimeraX: meeting modern challenges in visualization and analysis**. *Protein Sci* 2018, **27**:14-25.

33. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, Wang L-P, Simmonett AC, Harrigan MP, Stern CD *et al.*: **OpenMM 7: rapid development of high performance algorithms for molecular dynamics**. *PLoS Comput Biol* 2017, **13**: e1005659.

34. McNicholas S, Croll T, Burnley T, Palmer CM, Hoh SW, Jenkins HT, Dodson E, Cowtan K, Agirre J: **Automating tasks in protein structure determination with the clipper python module**. *Protein Sci* 2018, **27**:207-216.

35. Kirschner KN, Yongye AB, Tschampel SM, González-Outeiriño J, Daniels CR, Foley BL, Woods RJ: **GLYCAM06: a generalizable biomolecular force field. Carbohydrates**. *J Comput Chem* 2008, **29**:622-655.

36. Sails (software). Available from GitHub at https://github.com/glycojones/sails.

37. Cowtan K: **Automated nucleic acid chain tracing in real time**. *IUCrJ* 2014, **1**:387-392.

38. Cowtan K: **The Buccaneer software for automated model building. 1. Tracing protein chains**. *Acta Crystallogr D Biol Crystallogr* 2006, **62**:1002-1011.

39. Cowtan K: **Completion of autobuilt protein models using a database of protein fragments**.. IUCr *Acta Crystallogr D Biol Crystallogr* 2012, **68**:328-335.

40. Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA: **REFMAC5 for the refinement of macromolecular crystal structures**. *Acta Crystallogr D Biol Crystallogr* 2011, **67**:355-367.

41. Potterton L, Agirre J, Ballard C, Cowtan K, Dodson E, Evans PR, Jenkins HT, Keegan R, Krissinel E, Stevenson K *et al.*: **CCP4i2: the new graphical user interface to the CCP4 program suite**. *Acta Crystallogr D Struct Biol* 2018, **74**:68-84.

42. Moriarty NW, Tronrud DE, Adams PD, Karplus PA: **Conformation-dependent backbone geometry restraints set a new standard for protein crystallographic refinement**. *FEBS J* 2014, **281**:4061-4071.

43. Park H, Ovchinnikov S, Kim DE, DiMaio F, Baker D: **Protein homology model refinement by large-scale energy optimization**. *Proc Natl Acad Sci U S A* 2018, **115**:3054-3059.

44. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K *et al.*: **The Rosetta all-atom energy function for macromolecular modeling and design**. *J Chem Theory Comput* 2017, **13**:3031-3048.

45. • Frenz B, Rämisch S, Borst AJ, Walls AC, Adolf-Bryfogle J, Schief WR, Veesler D, DiMaio F: **Automatically fixing errors in glycoprotein structures with Rosetta**. *Structure* 2019, **27**:134-139.e3.
A Rosetta-based protocol for carbohydrate refinement of X-ray and cryoEM structures.

46. Dimaio F, Echols N, Headd JJ, Terwilliger TC, Adams PD, Baker D: **Improved low-resolution crystallographic refinement with Phenix and Rosetta**. *Nat Methods* 2013, **10**:1102-1106.

47. Labonte JW, Adolf-Bryfogle J, Schief WR, Gray JJ: **Residue-centric modeling and design of saccharide and glycoconjugate structures**. *J Comput Chem* 2017, **38**:276-287.

48. Terwilliger TC, Di Maio F, Read RJ, Baker D, Bunkóczi G, Adams PD, Grosse-Kunstleve RW, Afonine PV, Echols N: **Phenix. mr-rosetta: molecular replacement and model rebuilding with Phenix and Rosetta**. *J Struct Funct Genomics* 2012, **13**:81-90.

49. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ: **The amber biomolecular simulation programs**. *J Comput Chem* 2005, **26**:1668-1688.

50. Lütteke T, Von Der Lieth C-W: **pdb-care (PDB CArbohydrate REsidue check): a program to support annotation of complex carbohydrate structures in PDB files**. *BMC Bioinformatics* 2004, **5**.

51. Lutteke T: **Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB**. *Nucleic Acids Res* 2004, **33**:D242-D246.

52. Bohne-Lang A, Lang E, Förster T, Von der Lieth CW: **LINUCS: linear notation for unique description of carbohydrate sequences**. *Carbohydr Res* 2001, **336**:1-11.

53. Fushinobu S: **Conformations of the type-1 lacto-N-biose I unit in protein complex structures**. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 2018, **74**:473-479.

54. McNicholas S, Agirre J: **Glycoblocks: a schematic three-dimensional representation for glycans and their interactions**. *Acta Crystallogr Sect D Struct Biol* 2017, **73**:187-194.

55. McNicholas S, Potterton E, Wilson KS, Noble MEM: **Presenting your structures: the CCP4mg molecular-graphics software**. *Acta Crystallogr D Biol Crystallogr* 2011, **67**:386-394.

56. Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Marth JD, Bertozzi CR, Hart GW, Etzler ME: **Symbol nomenclature for glycan representation**. *Proteomics* 2009, **9**:5398-5399.

57. Thieker DF, Hadden JA, Schulten K, Woods RJ: **3D implementation of the symbol nomenclature for graphical representation of glycans**. *Glycobiology* 2016, **26**:786-787.

58. Sehnal D, Grant OC: **Rapidly display glycan symbols in 3D structures: 3D-SNFG in LiteMol**. *J Proteome Res* 2019, **18**:770-774.

59. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF chimera—a visualization system for exploratory research and analysis**. *J Comput Chem* 2004, **25**:1605-1612.

60. Tangram-snfg (software). Available from GitHub at https://github.com/insilichem/tangram_snfg.

61. Pérez S, Tubiana T, Imberty A, Baaden M: **Three-dimensional representations of complex carbohydrates and polysaccharides—SweetUnityMol: a video game-based computer graphic software**. *Glycobiology* 2015, **25**:483-491.

62. Arroyuelo A, Vila JA, Martin OA: **Azahar: a PyMOL plugin for construction, visualization and analysis of glycan molecules**. *J Comput Aided Mol Des* 2016, **30**:619-624.

63. Agirre J, Ariza A, Offen WA, Turkenburg JP, Roberts SM, McNicholas S, Harris PV, McBrayer B, Dohnalek J, Cowtan KD *et al.*: **Three-dimensional structures of two heavily N-glycosylated *Aspergillus* sp. family GH3 β-D-glucosidases**. *Acta Crystallogr D Struct Biol* 2016, **72**:254-265.

64. • Agirre J, Moroz O, Meier S, Brask J, Munch A, Hoff T, Andersen C, Wilson KS, Davies GJ: **The structure of the AliC GH13 α-amylase from *Alicyclobacillus* sp. reveals the accommodation of starch branching points in the α-amylase family**. *Acta Crystallogr D Struct Biol* 2019, **75**:1-7.
A description of the structure solution of AliC GH13 alpha-amylase, including refinement and validation. Ad-hoc dictionaries were needed to refine two coexisting anomeric forms that were rotated with respect to each other.

65. Schumann B, Malaker SA, Wisnovsky SP, Debets MF, Agbay AJ, Fernandez D, Wagner LJS, Lin L, Choi J, Fox DM *et al.*: **Chemical precision glyco-mutagenesis by glycosyltransferase engineering in living cells**. *bioRxiv* 2019 http://dx.doi.org/10.1101/669861.

66. Ji S, Dix SR, Aziz AA, Sedelnikova SE, Baker PJ, Rafferty JB, Bullough PA, Tzokov SB, Agirre J, Li F-L *et al.*: **The molecular basis of endolytic activity of a multidomain alginate lyase from Defluviitalea phaphyphila, a representative of a new lyase family, PLxx**. *J Biol Chem* 2019, **294**:18077-18091 http://dx.doi.org/10.1074/jbc.RA119.010716.

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Leveraging glycomics data in glycoprotein 3D structure validation with Privateer

Haroldas Bagdonas[1], Daniel Ungar[2] and Jon Agirre[*1]

Full Research Paper

Open Access

Address:
[1]York Structural Biology Laboratory, Department of Chemistry, University of York, Wentworth Way, York, YO10 5DD, UK and [2]Department of Biology, University of York, Wentworth Way, York, YO10 5DD, UK

Email:
Jon Agirre[*] - jon.agirre@york.ac.uk

* Corresponding author

Keywords:
electron cryomicroscopy; glycoinformatics; glycomics; Privateer; X-ray crystallography

Beilstein J. Org. Chem. **2020,** *16,* 2523–2533.
https://doi.org/10.3762/bjoc.16.204

Received: 18 July 2020
Accepted: 06 October 2020
Published: 09 October 2020

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

Guest Editor: N. H. Packer

Abstract

The heterogeneity, mobility and complexity of glycans in glycoproteins have been, and currently remain, significant challenges in structural biology. These aspects present unique problems to the two most prolific techniques: X-ray crystallography and cryo-electron microscopy. At the same time, advances in mass spectrometry have made it possible to get deeper insights on precisely the information that is most difficult to recover by structure solution methods: the full-length glycan composition, including linkage details for the glycosidic bonds. The developments have given rise to glycomics. Thankfully, several large scale glycomics initiatives have stored results in publicly available databases, some of which can be accessed through API interfaces. In the present work, we will describe how the Privateer carbohydrate structure validation software has been extended to harness results from glycomics projects, and its use to greatly improve the validation of 3D glycoprotein structures.

Introduction

Glycosylation-related processes are prevalent in life. The attachment of carbohydrates to macromolecules extends the capabilities of cells to convey significantly more information than what is available through protein synthesis and the expression of the genetic code alone. For example, glycosylation is used as a switch to modulate protein activity [1]; glycosylation plays a crucial part in folding/unfolding pathways of some proteins in cells [2,3]; the level of *N*-glycan expression regulates the adhesiveness of a cell [4]; glycosylation also plays a role in immune function [5] and cellular signalling [5,6]. At the forefront, glycosylation plays a significant role in influencing protein–protein interactions. For example, the influenza virus uses the haemagglutinin glycoprotein to recognise and bind sialic acid decorations of human cells in the respiratory tract [7]. Glycosylation is also used by pathogens to evade the host's immune system via glycan shields [8-10], and thereby to delay

an immune response [11]. The structural study of these glycan-mediated interactions can provide unique insight into the molecular interplay governing these processes. In addition, it can provide structural snapshots in atomistic detail that can be used to generate molecular dynamics simulations describing a wider picture underpinning glycan and protein interactions [12]. Unfortunately, significant challenges have affected the determination of glycoprotein structures for decades and have had a detrimental impact on the quality and reliability of the produced models. Anomalies have been reported regarding carbohydrate nomenclature [13], glycosidic linkage stereochemistry [14] and torsion [15,16], and most recently, ring conformation [17]. Most of these issues have now been addressed as part of ongoing efforts to provide better software tools for structure determinations of glycoproteins, although the most difficult cases remain hard to solve. Chiefly among these is the scenario where the experimentally resolved electron density map provides evidence of glycosylation, without enough resolution to derive definite and comprehensive details about the structural composition of the oligosaccharides (Figure 1). Glycan microheterogeneity and the lack of carbohydrate-specific modelling tools have often been named as the principal causes for these issues [18].

Heterogeneity of glycoproteins

Unlike protein synthesis, which is encoded in the genome and follows a clear template, glycan biosynthesis is not template-directed. A single glycoprotein will exist in multiple possibilities of products that can emerge from the glycan biosynthesis pathways, and these are known as glycoforms [22]. More specifically, the variation can appear in terms of which potential glycosylation sites are occupied at any time – macroheterogeneity – or variations in the compositions of the glycans added to specific glycosylation sites – microheterogeneity. This variation in the microheterogeneous composition patterns arises due to the competition of glycan-processing enzymes in biosynthesis pathways [23].

Implications for the structure determination of glycoproteins

Several experimental techniques can be used to obtain 3D structures of glycoproteins: X-ray crystallography (MX, which stands for macromolecular crystallography), nuclear magnetic resonance spectroscopy (NMR) and electron cryomicroscopy (cryo-EM). As of publication date, the overwhelming majority of glycoprotein structures have been solved using MX [24,25].

The biggest bottleneck in MX is the formation of crystals of the target macromolecule or complex. The quality of the crystal directly determines the resolution – a measure of the detail in the electron density map. Homogenous samples at high concentrations are required to produce well-diffracting crystals [26]. Samples containing glycoprotein molecules do not usually fulfill this criterion. More often than not, MX falls short at elucidating carbohydrate features in glycoproteins due to glycosylated proteins being inherently mobile and heterogeneous [22]. Moreover, oligosaccharides often significantly interfere with the formation of crystal contacts that allow the formation of well-diffracting crystals. Because of this, glycans are often truncated in MX samples to aid crystal formation [27].

In cryo-EM, samples of glycoproteins are vitrified at extremely low temperatures rather than crystallised, as in MX. The rapid cooling of the sample allows to capture snapshots of the molecules at their various conformational states, and thus potentially maintaining glycoprotein states more closely to their native environments in comparison to crystallography [28]. Nevertheless, cryo-EM is still not an end-all solution to solving glycoprotein structures: the flexible and heterogeneous nature of glycans still has an adverse effect on the quality of the data,

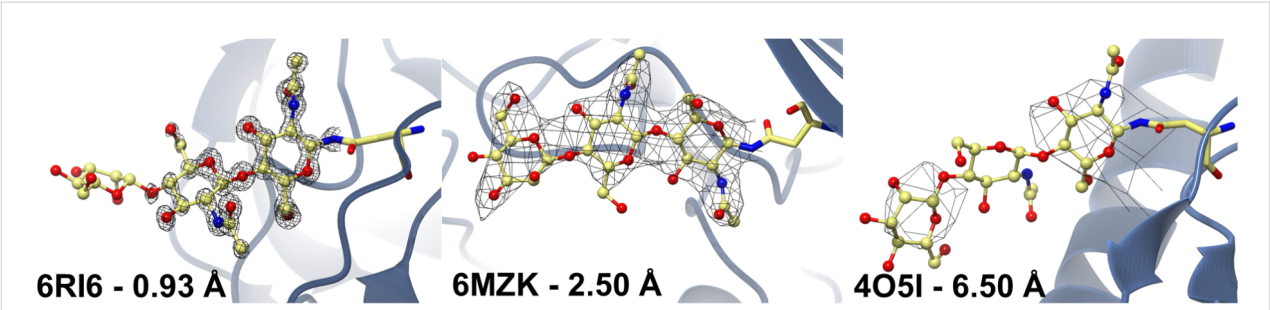


Figure 1: Comparison of the glycan features in electron density maps over a range of resolutions from selected glycoprotein structures (PDB entries: 6RI6 [19]; 6MZK [20]; 4O5I [21]). The electron density maps were obtained with X-ray crystallography. The data resolution and PDB entry IDs associated with the structures have been directly annotated on the structure. Left: A high-resolution example where monosaccharides and the conformations can be elucidated; middle: A medium resolution example where the identification starts to become difficult; right: A low-resolution example for which all prior knowledge must be used. Despite coming from different glycoprotein structures, the glycan has the same composition, and thus is assigned a unique GlyTouCan ID of G15407YE.

affecting the image reconstruction [29]. Moreover, due to the low signal-to-noise ratio, the technique works more easily with samples of a high molecular weight; this situation, however, is evolving rapidly, with reports of sub-100 kDa structures becoming more frequent lately [30,31]. Crucially, MX and cryo-EM can complement each other to counteract issues that both face individually [32].

The two techniques produce different information – electron density (MX) or electron potential (cryo-EM) maps – but the practical considerations in terms of the atomistic interpretation hold true for both: provided that at least the secondary structural features can be resolved in a 3D map, a more or less complete atomic model will be expected as the final result of the study. Modelling of carbohydrates into 3D maps can be more complex than modelling proteins [33], although recent advances in software are closing the gap [34-36]. However, to date it remains true that most model building software is protein-centric [15]. As a consequence, the glycan chains in glycoprotein models that have been elucidated before recent developments in carbohydrate validation and modelling software tend to contain a significant amount of errors: wrong carbohydrate nomenclature [13], biologically implausible glycosidic linkage stereochemistry [14], incorrect torsion [15,16], and unlikely high-energy ring conformations [17]. Early efforts in the validation of carbohydrate structures saw the introduction of online tools such as PDB-CARE [37] and CARP [16]; more recently, we released the Privateer software [24], which was the first carbohydrate validation tool available as part of the CCP4i2 crystallographic structure solution pipeline [38]. In its first release, Privateer was able to perform stereochemical and conformational validation of pyranosides, analyse the glycan fit to electron density map and offered tools for restraining a monosaccharide minimal-energy conformation.

While these features were recognised to address some long-standing needs in carbohydrate structure determination [39,40], significant challenges remain, particularly in the scenario where the glycan composition cannot be ascertained solely from the three-dimensional map. Unfortunately, this problematic situation happens frequently, especially in view of the fact that the median resolution for glycoproteins (2.4 Å) is lower than that of non-glycosylated – potentially including fully deglycosylated – proteins (2.0 Å) [41]. To date, only one publicly available model building tool has attacked this issue: the Coot software offers a module that will build some of the most common *N*-linked glycans in a semiautomated fashion [34]. Indeed, the Coot module was built around the suggestion that only the most probable glycoforms should be modelled unless prior knowledge of an alternative glycan composition exists in the form of, e.g., mass spectrometry data [14].

Harnessing glycomics and glycoproteomics results to inform glycan model building

Current methods used to obtain accurate atomistic descriptions of molecules fall short in dealing with the heterogeneity of glycoproteins. However, there are other methods that have been proven to successfully tackle the challenges posed by glycan heterogeneity, with mass spectrometry emerging as the one with the most relevance due to the ability to elucidate the complete composition descriptions of individual oligosaccharide chains on glycoproteins [42].

The mass spectrometric analysis of glycosylated proteins can be with (glycomics) or without (glycoproteomics) the release of oligosaccharides from the glycoprotein. Usually, glycomics and glycoproteomics experiments are carried out together to obtain a complete description of the glycoprotein profile. Glycomics experiments are required to distinguish stereoisomers and the linkage information in order to obtain a full structural description about a glycan, whereas glycoproteomics are required to establish the glycan variability and occupancy at the glycosylation sites of the protein [43]. Typically, these analyses are based on mass spectrometry techniques, such as electrospray ionization mass spectrometry (ESIMS) and matrix-assisted laser desorption ionization MS (MALDIMS) [43]. Mass spectrometry techniques are best suited for the determination of the composition of monosaccharide classes and the chain length. However, the in-depth analysis of a glycan typically requires the integration of complementary analytic techniques, such as nuclear magnetic resonance (NMR) and capillary electrophoresis (CE). Nevertheless, depending on the sample, advanced mass spectrometry techniques can be used to counteract the need for complementary analytic techniques. One of the examples of this is tandem mass spectrometry, where the glycan fragmentation is controlled to obtain the identification of the glycosylation sites and a complete description of the glycan structure compositions, including linkage and sequence information [44]. Moreover, recent advances in ion mobility mass spectrometry can now also be used for a complete glycan analysis [45].

The analysis and interpretation of mass spectrometry spectra produced by glycans is a challenge. Most significantly, in MS outputs, glycans appear in their generalized composition classes, i.e., Hex, HexNAc, dHex, NeuAc, etc. The identity elucidation of generalized unit classes into specific monosaccharide units (such as Glc, Gal, Man, GalNAc, etc.) requires prior knowledge of the glycan biosynthetic pathways [46]. Additional sources of prior knowledge are bioinformatics databases that have been curated through the deposition of experimental data. Bioinformatics databases contain detailed descriptions of the glycan compositions and

m/*z* values of specific glycans, and therefore aiding the process of glycan annotation [47]. Such bioinformatics databases can usually be interrogated using textual or graphical notations that describe the glycan sequence. However, due to the glycan complexity and the incremental nature of the different glycomics projects, numerous notations have been developed over the years – e.g., CarbBank [48] utilized CCSD [48] and EuroCarbDB [49] and GlycomeDB [50] used GlycoCT [51] (Table 1).

Thankfully, data from discontinued glycomics projects are not lost but were integrated into newer platforms, often with novel notations. One such example is GlyTouCan [53], which uses both GlycoCT [54] and WURCS [53] as notation languages. As a result, tools that interconvert between notations were developed to successfully integrate old data into new platforms. Additionally, the introduction of tools such as GlycanFormatConverter [55] to convert WURCS notations into more human-readable formats has eased the interpretation of glycan databases.

Significantly, the GlyTouCan project aims to create a public repository of known glycan sequences by assigning them unique identification tags. Each identification tag describes a glycan sequence in the WURCS notation, and this allows to link specific glycans to other databases, such as GlyConnect [56], UniCarb-DB [57] and others, any of which are tailored to specific flavours of glycomics and glycoproteomics experiments. Ideally, this implementation ends up requiring the user to be familiar with a single notation – WURCS – used to represent sequences of glycans.

From glycomics/glycoproteomics to carbohydrate 3D model building and validation in Privateer

Many fields, for example pharmaceutical design and engineering [58], molecular dynamics simulations [59] and protein interaction studies [60], rely upon structural biology to produce accurate atomistic descriptions of glycoproteins. However, due to clear limitations of elucidating carbohydrate features in MX/cryo-EM electron-density maps, structural biologists are likely to make mistakes. This introduces the possibility of modelling wrong glycan compositions in glycoprotein models, going as far as not conforming with general glycan biosynthesis knowledge. Model building pipelines would therefore greatly benefit from the ability to validate against the knowledge of glycan compositions elucidated via glycomics/glycoproteomics experiments. This warrants the need for new tools that are able to link these methodologies, through an intermediate interconversion library.

A foundation for such interconversion libraries exists in the form of the carbohydrate validation software Privateer. The program is able to compute individual monosaccharide conformations from a glycoprotein model, check whether the modelled carbohydrates atomistic definitions match dictionary standards as well as output multiple helper tools to aid the processes of refinement and model building [24]. Most importantly, Privateer already contains methods that allow the extraction of carbohydrate atomistic definitions to create abstract definitions of glycans in memory, and thus already laying a foundation for the generation of unique WURCS notations and providing a straightforward access to bioinformatics databases that are integrated in the GlyTouCan project.

Table 1: A comparison of the structural information storage capabilities of different sequence formats used in glycobioinformatics.[a]

notation	multiple connections	repeating units	alternative residues	linear notation	atomic ambiguity
CCSD(CarbBank)	–	+	–	+	–
LINUCS	–	+	–	+	–
GlycoSuite	–	–	+	+	–
BCSDB	(+)	(+)	+	+	–
LinearCode	–	–	+	+	–
KCF	+	+	–	–	–
GlycoCT	+	+	+	–	–
Glyde-II	+	+	–	–	–
WURCS 2.0	+	+	+	+	+

[a]"+" Denotes that information can be stored directly without any significant issues, "(+)" denotes that information can be stored indirectly, or that there are some issues and "–" denotes that information description in the particular sequence format is unavailable. This table is a simplified version of the one originally published by Matsubara et al. [52].

Methods

The algorithm used to generate the WURCS notation in Privateer is based on the description published in Tanaka et al. [61], with required updates applied from Matsubara et al. [52]. WURCS was designed to deal with the incomplete descriptions of glycan sequences emerging from glycomics/glycoproteomics experiments (i.e., undefined linkages, undefined residues and ambiguous structures in general). However, the lack of this detail is unlikely to be supported in "pdb" or "mmCIF'" format files, which are a standard in structural biology. As a result, the "atomic ambiguity" capability (Table 1) is not supported in Privateer's implementation. Moreover, Privateer's implementation of WURCS relies on a manually compiled dictionary that translates the PDB Chemical Component Dictionary [62] three-letter codes of carbohydrate monomer definitions found in the structure files into WURCS definitions of unique monomers (described as "UniqueRES" [52]).

The WURCS notations are generated for all detected glycans that are linked to protein backbones in the input glycoprotein model. For every glycan chain in the model, the algorithm computes a list of all detected monosaccharides that are unique and stores that information internally in memory. Then, the algorithm calculates the unit counts in a glycan chain – how many unique monosaccharides are modelled in the glycan chain, the total length of the glycan chain and computes the total number linkages between monosaccharides. After the composition calculations are carried out, the algorithm begins the generation of the notation by printing out the unit counts. Then, the list of unique monosaccharide definitions in the glycan chain are printed out by converting the three-letter PDB codes into WURCS-compliant definitions. Afterwards, each individual monosaccharide of the glycan is assigned a numerical ID according to its occurrence in the list of unique monosaccharides. Finally, the linkage information between monosaccharide pairs are generated by assigning individual monosaccharides a unique letter ID according to their position in the glycan chain. Alongside a unique letter ID, a numerical term is added that describes a carbon position from which the bond is formed to another carbohydrate unit. Crucially, the linkage detection in Privateer does not rely at all on metadata present in the structure file. Instead, linkages are identified based on the perceived chemistry of the input model: which atoms are close enough – but not too close – to be plausibly linked.

The generated WURCS string can then be used to search whether an individual glycan chain has been deposited in GlyTouCan. The scan of the repository occurs internally within the Privateer software, as all the data is stored in a single structured data file written in JSON format that is distributed together with Privateer. If the existence of a glycan in the database is confirmed, then the software can attempt to find records about the sequence on other, more specialised databases (currently only GlyConnect) to obtain information such as the source organism, the type of glycosylation and the glycan core to carry out further checks in the glycoprotein model (Figure 2).

Availability and performance of the algorithm

This new version of Privateer (MKIV) will be released as an update to CCP4 7.1. To demonstrate the capabilities of the computational bridge integrated in the newest version of Privateer (for standalone bundles, please refer to privateer branch "privateerMKIV_noccp4" of GitHub repository with the installation instructions provided in the README.md file [63]), it was run on all *N*-glycosylated structures in the PDB solved using MX and cryo-EM. The list of structures used in this demonstration was obtained from Atanasova et al. [18]. The computational analysis of the demonstration revealed a relatively small proportion of deposited glycoprotein models containing glycan chains that do not have a unique GlyTouCan accession ID assigned, raising questions about the provenance of their structures. Importantly, the majority of the glycan chains that do have a unique GlyTouCan accession ID assigned (except for single residues linked to protein backbones), have also been successfully matched on the GlyConnect database (Table 2).

Results

Examples of use

As observed in previous studies, glycoprotein models deposited in the PDB feature flaws ranging from minor irregularities to gross modelling errors [14,17,41,64]. The automated validation of minor irregularities was already possible with automated tools such as pdb-care [37], CARP [65], and Privateer [24]. However, the automated detection of gross modelling errors is currently a challenge due to the lack of publicly available tools. Our newly developed computational bridge between structural biology and glycomics databases makes the detection of gross modelling errors easier, as demonstrated by the following examples.

Example 1 – 2H6O

The glycoprotein model (PDB code 2H6O) proposed by Szakonyi et al. [66] contains 12 glycans, as detected by Privateer. The model became infamous after it sparked the submission of a critical correspondence published by Crispin et al. [14]. The article contained a discussion about the proposed model containing glycans that were previously unreported and inconsistent with glycan biosynthetic pathways. In particular, the model contained oligosaccharide chains with Man-(1→3)-GlcNAc and GlcNAc-(1→3)-GlcNAc linkages, β-galactosyl

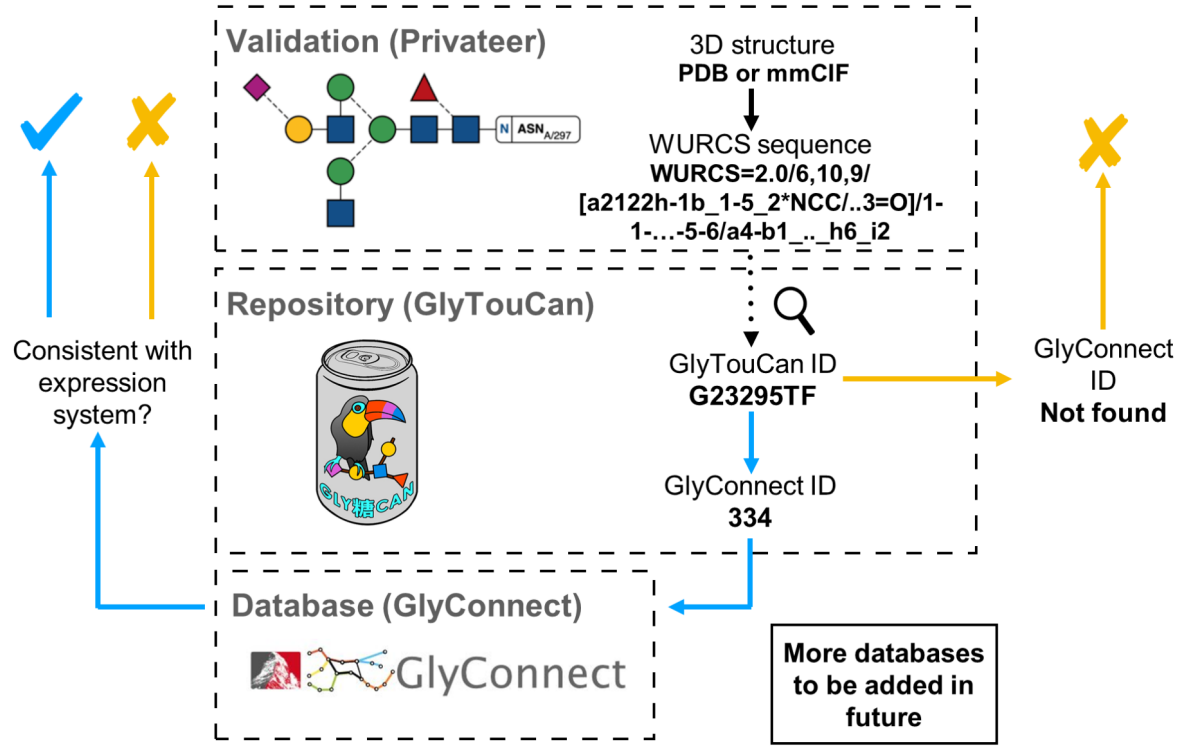


Figure 2: A roadmap of the software development project that allows structural biologists to quickly obtain detailed information about specific glycans in glycoprotein models from glycomics/glycoproteomics databases. The GlyTouCan (https://glytoucan.org/) and GlyConnect (https://glyconnect.expasy.org/) logos have been reproduced here under explicit permission from their respective authors.

Table 2: Comparison of the successful glycan matches detected by Privateer in the GlyTouCan and the GlyConnect database.[a]

experimental technique	glycan chain length	GlyTouCan ID found	GlyTouCan ID not found	% of GlyTouCan in GlyConnect	total glycan chains
MX	1	16797	0	1%	16797
MX	2	5870	5	90%	5875
MX	3	2550	17	71%	2567
MX	4	1012	21	80%	1033
MX	5	834	72	74%	906
MX	6	460	85	69%	545
MX	7	345	55	77%	400
MX	8	235	25	85%	260
MX	9	164	16	81%	180
MX	10	118	5	92%	123
MX	11	20	5	85%	25
MX	12	8	4	75%	12
MX	13	0	1	0%	1
MX	14	0	0	0%	0
MX	15	2	0	0%	2
MX	16	0	1	0%	1
cryo-EM	1	2080	0	3%	2080
cryo-EM	2	1081	0	98%	1081
cryo-EM	3	439	0	96%	439
cryo-EM	4	143	0	93%	143

Table 2: Comparison of the successful glycan matches detected by Privateer in the GlyTouCan and the GlyConnect database.[a] (continued)

cryo-EM	5	146	2	85%	148
cryo-EM	6	70	1	97%	71
cryo-EM	7	45	0	100%	45
cryo-EM	8	26	0	88%	26
cryo-EM	9	15	1	100%	16
cryo-EM	10	16	0	100%	16
cryo-EM	11	4	0	100%	4
cryo-EM	12	1	0	100%	1
cryo-EM	13	1	0	0%	1

[a]Glycans obtained from the glycoprotein models were elucidated by X-ray crystallography and cryo-EM.

motifs capping oligomannose-type glycans and hybrid-type glycans containing terminal Man-(1→3)-GlcNAc [14]. Moreover, the proposed model contained systematic errors in the anomer annotations and carbohydrate stereochemistry. To this day, there is still no experimental evidence reported for these types of linkages and capping in an identical context.

The new version of Privateer was run on the proposed model. WURCS notations were successfully generated for all glycans, with only 1 glycan chain out of 12 successfully returning a GlyTouCan ID. Under further manual review of the one glycan and with help from other validation tools contained in Privateer, it was found to contain anomer mismatch errors (the three letter code denoting one anomeric form did not match the anomeric form reflected in the atomic coordinates). After the anomer mismatch errors were corrected, the oligosaccharide chain also failed to return GlyTouCan and GlyConnect IDs. The other 11 chains that failed to return a GlyTouCan ID also contained flaws, as described previously (Figure 3).

The analysis of this PDB entry highlights the kind of cross-checks that could be done by Protein Data Bank annotators upon validation and deposition of a new glycoprotein entry. It should be recognised that PDB annotators might not necessarily be experts in structural glycobiology. The fact that these glycans could not be matched to standard database entries should be enough to raise the question with depositors, and at the very least write a caveat on a deposited entry where glycans could not be correctly identified. Furthermore, despite the example showing just *N*-glycosylation, other kinds of glycosylation are searchable as well, and therefore this tool could shed much needed light on the validity of models representing more obscure types of modifications.

Example 2 – 2Z62

Successfully matching the WURCS string to a GlyTouCan ID, should not be a sole measure of a structure validity. GlyTouCan is a repository of all potential glycans collected from a set of databases, with the entries often representing glycans. Therefore, the correctness of the composition should be critically validated against the information provided in specialized and high-quality databases such as GlyConnect [56] and UniCarbKB [67]. The computational bridge provides direct search of entries stored in GlyConnect, with plans to expand this to more databases in the near future.

An example where the sole reliance on the detection of a glycan in GlyTouCan would not be sufficient is rebuilding of the 2Z62 glycoprotein structure [68] to improve the model quality [41] (Figure 4). The analysis of the original model generated the GlyTouCan ID G28454KX, which could not be detected in GlyConnect. The automated tools used by PDB-REDO slightly improved the model by renaming one of the fucose residues from FUL to FUC due to an anomer mismatch between the three letter code and the actual coordinates of the monomer. The new model thus generated the GlyTouCan ID G21290RB, which in turn could be matched to the GlyConnect ID 54. Under further manual review of mFo-DFc difference density map, a (1→3)-linked fucose was added, along with additional corrections to the coordinates of the molecule [41]. The newly generated WURCS notation for the model returned a GlyTouCan ID of G63564LA, with a GlyConnect ID of 145. The iterative steps taken to rebuild the glycoprotein model have been portrayed (Figure 4). Because the data in GlyConnect is approximately 70% manually curated by experts in the field [56], a match of a specific glycan in this database is likely a valid confirmation of a specific oligosaccharide composition and linkage pattern found in nature.

Conclusion

The mirrors of GlyConnect and GlyTouCan were obtained thanks to the public access to the API commands, which allowed to create scripts that automated the query of the entries stored in the databases with relative ease. However, the integration of additional databases might require support from the developers of those databases. Support for lipopolysaccharides

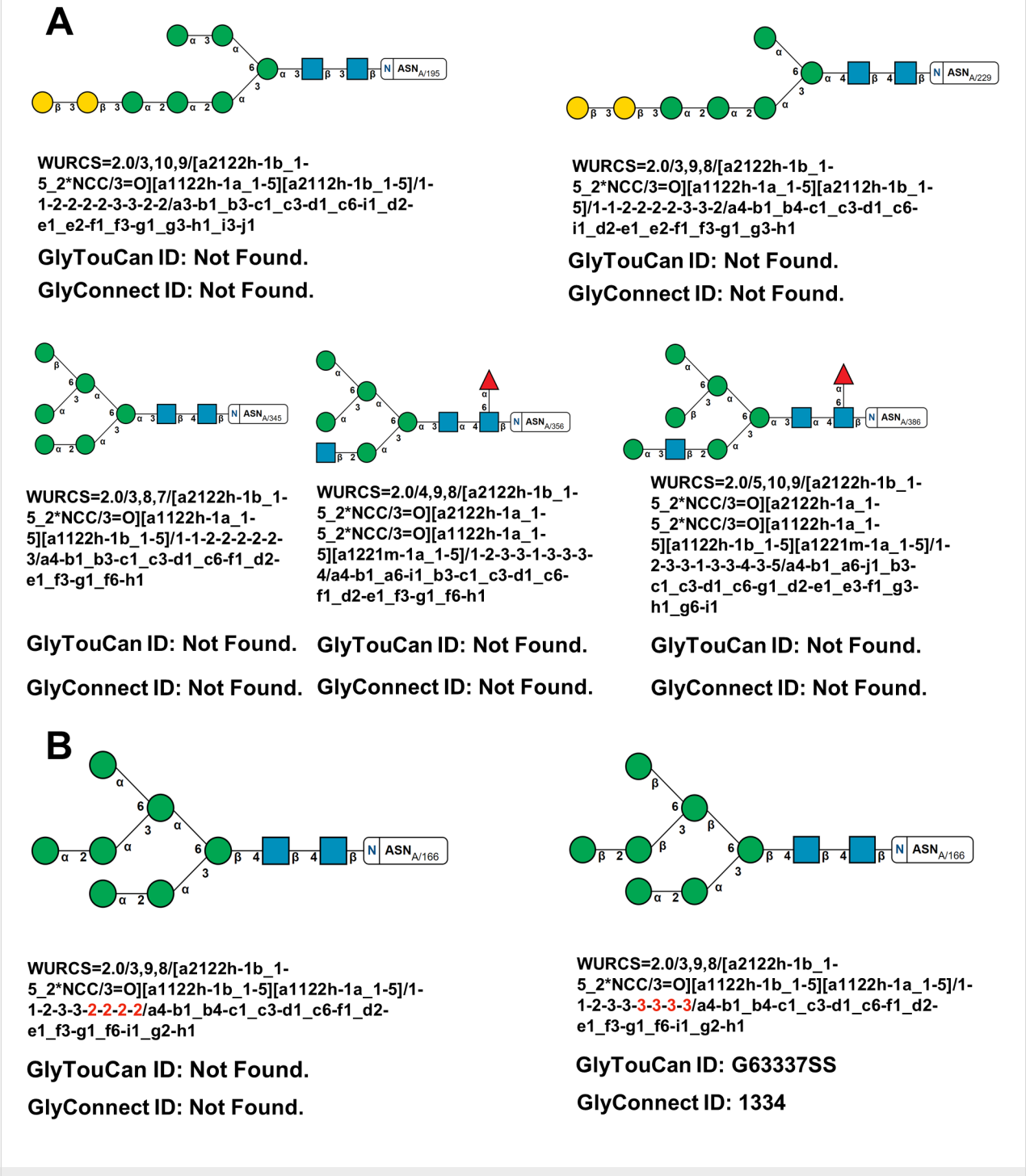


Figure 3: N-Linked glycans in Epstein Barr virus major envelope glycoprotein (PDB entry: 2H6O [66]). A) A selection of the glycan chains that failed to return database IDs with their WURCS sequences extracted from the Privateer CCP4i2 report. B) Glycan chain (right) for which a GlyTouCan and GlyConnect ID have successfully been matched with the modelling errors present in the model. After manual fixing (left), the WURCS sequence for the glycan failed to return database IDs. Highlighting in red depicts the locations in WURCS notation where both glycans differ.

and polysaccharides may be added in future, too, owing to the general purpose of the integrated databases – i.e., they are not limited to protein glycosylation.

Currently, the generated WURCS strings are matched against an identical sequence in the database. This means that if a glycan model has a single modelling mistake, for example, at

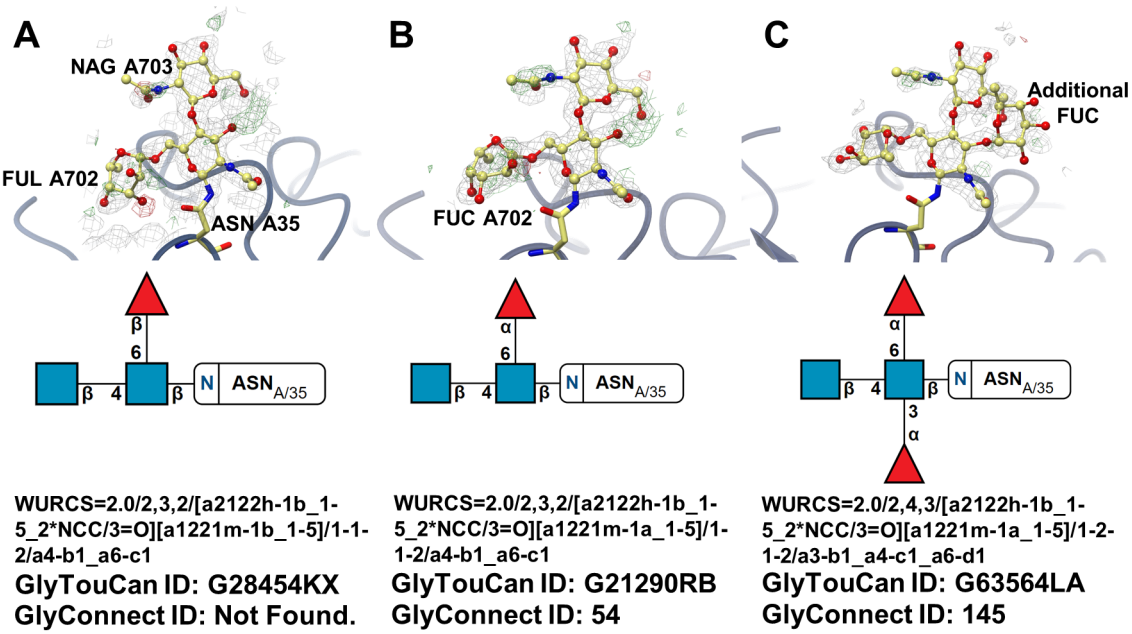


Figure 4: An *N*-linked glycan attached to Asn35 of human Toll-like receptor 4 (A: PDB entry 2z62 [68]). Model iteratively rebuilt by PDB-Redo as shown in steps B and C [41]. Pictures at the top depict glycoprotein models of the region of interest and electron-density maps of the glycan chain (grey: 2mFo DFc map, green and red: mFo DFc difference density map). Pictures at the bottom depict the SNFG representations of glycan chains, their WURCS sequence and accession IDs to relevant databases (taken directly from Privateer's CCP4i2 report).

one end of the chain but is correct elsewhere, the current version of the software would still fail to return a match. This issue has been solved in the development version by the incorporation of a subtree matching algorithm, which will reveal modelling mistakes at specific positions of the glycans, and report these to the user.

Currently, all the developments outlined in this work are accessible exclusively through the Privateer command line interface and through Coot scripts. In order to facilitate the interaction with users, a graphical interface to the new functionality will be provided through the CCP4i2 [38] framework. This new version of the interface is at the testing stage at the time of publication.

Acknowledgements

We would also like to acknowledge the support of the Departments of Chemistry and Biology at the University of York.

Funding

Haroldas Bagdonas is funded by The Royal Society [grant number RGF/R1/181006]. Jon Agirre is the Royal Society Olga Kennard Research Fellow [award number UF160039]. The work in Daniel Ungar's group is supported by the BBSRC [grant number BB/M018237/1].

ORCID® iDs

Haroldas Bagdonas - https://orcid.org/0000-0001-5028-4847
Daniel Ungar - https://orcid.org/0000-0002-9852-6160
Jon Agirre - https://orcid.org/0000-0002-1086-0253

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2020.83.v1

References

1. Rohne, P.; Prochnow, H.; Wolf, S.; Renner, B.; Koch-Brandt, C. *Cell. Physiol. Biochem.* **2014,** *34,* 1626–1639. doi:10.1159/000366365
2. Wyss, D. F.; Choi, J. S.; Li, J.; Knoppers, M. H.; Willis, K. J.; Arulanandam, A. R.; Smolyar, A.; Reinherz, E. L.; Wagner, G. *Science* **1995,** *269,* 1273–1278. doi:10.1126/science.7544493
3. Mitra, N.; Sharon, N.; Surolia, A. *Biochemistry* **2003,** *42,* 12208–12216. doi:10.1021/bi035169e
4. Gu, J.; Isaji, T.; Xu, Q.; Kariya, Y.; Gu, W.; Fukuda, T.; Du, Y. *Glycoconjugate J.* **2012,** *29,* 599–607. doi:10.1007/s10719-012-9386-1
5. Lyons, J. J.; Milner, J. D.; Rosenzweig, S. D. *Front. Pediatr.* **2015,** *3,* 54. doi:10.3389/fped.2015.00054
6. Boscher, C.; Dennis, J. W.; Nabi, I. R. *Curr. Opin. Cell Biol.* **2011,** *23,* 383–392. doi:10.1016/j.ceb.2011.05.001
7. Russell, R. J.; Kerry, P. S.; Stevens, D. J.; Steinhauer, D. A.; Martin, S. R.; Gamblin, S. J.; Skehel, J. J. *Proc. Natl. Acad. Sci. U. S. A.* **2008,** *105,* 17736–17741. doi:10.1073/pnas.0807142105

8. Crispin, M.; Ward, A. B.; Wilson, I. A. *Annu. Rev. Biophys.* **2018,** *47,* 499–523. doi:10.1146/annurev-biophys-060414-034156
9. Watanabe, Y.; Raghwani, J.; Allen, J. D.; Seabright, G. E.; Li, S.; Moser, F.; Huiskonen, J. T.; Strecker, T.; Bowden, T. A.; Crispin, M. *Proc. Natl. Acad. Sci. U. S. A.* **2018,** *115,* 7320–7325. doi:10.1073/pnas.1803990115
10. Pinger, J.; Nešić, D.; Ali, L.; Aresta-Branco, F.; Lilic, M.; Chowdhury, S.; Kim, H.-S.; Verdi, J.; Raper, J.; Ferguson, M. A. J.; Papavasiliou, F. N.; Stebbins, C. E. *Nat. Microbiol.* **2018,** *3,* 932–938. doi:10.1038/s41564-018-0187-6
11. Walls, A. C.; Park, Y.-J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. *Cell* **2020,** *181,* 281–292. doi:10.1016/j.cell.2020.02.058
12. Wood, N. T.; Fadda, E.; Davis, R.; Grant, O. C.; Martin, J. C.; Woods, R. J.; Travers, S. A. *PLoS One* **2013,** *8,* e80301. doi:10.1371/journal.pone.0080301
13. Lütteke, T.; von der Lieth, C. W. Data mining the PDB for Glyco-related data. In *Glycomics. Methods in Molecular Biology;* Packer, N. H.; Karlsson, N. G., Eds.; Humana Press: Totowa, NJ, USA, 2009; Vol. 534, pp 293–310. doi:10.1007/978-1-59745-022-5_21
14. Crispin, M.; Stuart, D. I.; Jones, E. Y. *Nat. Struct. Mol. Biol.* **2007,** *14,* 354–355. doi:10.1038/nsmb0507-354a
15. Agirre, J.; Davies, G. J.; Wilson, K. S.; Cowtan, K. D. *Curr. Opin. Struct. Biol.* **2017,** *44,* 39–47. doi:10.1016/j.sbi.2016.11.011
16. Frank, M.; Lutteke, T.; von der Lieth, C.-W. *Nucleic Acids Res.* **2007,** *35,* 287–290. doi:10.1093/nar/gkl907
17. Agirre, J.; Davies, G.; Wilson, K.; Cowtan, K. *Nat. Chem. Biol.* **2015,** *11,* 303. doi:10.1038/nchembio.1798
18. Atanasova, M.; Bagdonas, H.; Agirre, J. *Curr. Opin. Struct. Biol.* **2020,** *62,* 70–78. doi:10.1016/j.sbi.2019.12.003
19. Polyakov, K. M.; Gavryushov, S.; Fedorova, T. V.; Glazunova, O. A.; Popov, A. N. *Acta Crystallogr., Sect. D: Struct. Biol.* **2019,** *75,* 804–816. doi:10.1107/s2059798319010684
20. Dai, Y. N.; Fremont, D. H. PDB ID 6MZK; Crystal structure of hemagglutinin from influenza virus A/Pennsylvania/14/2010 (H3N2). https://www.wwpdb.org/pdb?id=pdb_00006mzk (accessed Oct 5, 2020). doi:10.2210/pdb6mzk/pdb
21. Lee, P. S.; Ohshima, N.; Stanfield, R. L.; Yu, W.; Iba, Y.; Okuno, Y.; Kurosawa, Y.; Wilson, I. A. *Nat. Commun.* **2014,** *5,* 3614. doi:10.1038/ncomms4614
22. Rudd, P. M.; Dwek, R. A. *Crit. Rev. Biochem. Mol. Biol.* **1997,** *32,* 1–100. doi:10.3109/10409239709085144
23. Fisher, P.; Thomas-Oates, J.; Wood, A. J.; Ungar, D. *Front. Cell Dev. Biol.* **2019,** *7,* 157. doi:10.3389/fcell.2019.00157
24. Agirre, J.; Iglesias-Fernández, J.; Rovira, C.; Davies, G. J.; Wilson, K. S.; Cowtan, K. D. *Nat. Struct. Mol. Biol.* **2015,** *22,* 833–834. doi:10.1038/nsmb.3115
25. Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Darvill, A. G.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H. *Essentials of Glycobiology,* 3rd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2016.
26. Geerlof, A.; Brown, J.; Coutard, B.; Egloff, M.-P.; Enguita, F. J.; Fogg, M. J.; Gilbert, R. J. C.; Groves, M. R.; Haouz, A.; Nettleship, J. E.; Nordlund, P.; Owens, R. J.; Ruff, M.; Sainsbury, S.; Svergun, D. I.; Wilmanns, M. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006,** *62,* 1125–1136. doi:10.1107/s0907444906030307
27. Stura, E. A.; Nemerow, G. R.; Wilson, I. A. *J. Cryst. Growth* **1992,** *122,* 273–285. doi:10.1016/0022-0248(92)90256-i
28. Cheng, Y.; Grigorieff, N.; Penczek, P. A.; Walz, T. *Cell* **2015,** *161,* 438–449. doi:10.1016/j.cell.2015.03.050
29. Serna, M. *Front. Mol. Biosci.* **2019,** *6,* 33. doi:10.3389/fmolb.2019.00033
30. Fan, X.; Wang, J.; Zhang, X.; Yang, Z.; Zhang, J.-C.; Zhao, L.; Peng, H.-L.; Lei, J.; Wang, H.-W. *Nat. Commun.* **2019,** *10,* 2386. doi:10.1038/s41467-019-10368-w
31. Herzik, M. A., Jr.; Wu, M.; Lander, G. C. *Nat. Commun.* **2019,** *10,* 1032. doi:10.1038/s41467-019-08991-8
32. Wang, H.-W.; Wang, J.-W. *Protein Sci.* **2017,** *26,* 32–39. doi:10.1002/pro.3022
33. Agirre, J. *Acta Crystallogr., Sect. D: Struct. Biol.* **2017,** *73,* 171–186. doi:10.1107/s2059798316016910
34. Emsley, P.; Crispin, M. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018,** *74,* 256–263. doi:10.1107/s2059798318005119
35. Croll, T. I. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018,** *74,* 519–530. doi:10.1107/s2059798318002425
36. Frenz, B.; Rämisch, S.; Borst, A. J.; Walls, A. C.; Adolf-Bryfogle, J.; Schief, W. R.; Veesler, D.; DiMaio, F. *Structure* **2019,** *27,* 134–139. doi:10.1016/j.str.2018.09.006
37. Lütteke, T.; von der Lieth, C.-W. *BMC Bioinf.* **2004,** *5,* 69. doi:10.1186/1471-2105-5-69
38. Potterton, L.; Agirre, J.; Ballard, C.; Cowtan, K.; Dodson, E.; Evans, P. R.; Jenkins, H. T.; Keegan, R.; Krissinel, E.; Stevenson, K.; Lebedev, A.; McNicholas, S. J.; Nicholls, R. A.; Noble, M.; Pannu, N. S.; Roth, C.; Sheldrick, G.; Skubak, P.; Turkenburg, J.; Uski, V.; von Delft, F.; Waterman, D.; Wilson, K.; Winn, M.; Wojdyr, M. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018,** *74,* 68–84. doi:10.1107/s2059798317016035
39. Gristick, H. B.; Wang, H.; Bjorkman, P. J. *Acta Crystallogr., Sect. D: Struct. Biol.* **2017,** *73,* 822–828. doi:10.1107/s2059798317013353
40. Joosten, R. P.; Lütteke, T. *Curr. Opin. Struct. Biol.* **2017,** *44,* 9–17. doi:10.1016/j.sbi.2016.10.010
41. van Beusekom, B.; Lütteke, T.; Joosten, R. P. *Acta Crystallogr., Sect. F: Struct. Biol. Commun.* **2018,** *74,* 463–472. doi:10.1107/s2053230x18004016
42. Nakahara, Y.; Miyata, T.; Hamuro, T.; Funatsu, A.; Miyagi, M.; Tsunasawa, S.; Kato, H. *Biochemistry* **1996,** *35,* 6450–6459. doi:10.1021/bi9524880
43. Shajahan, A.; Heiss, C.; Ishihara, M.; Azadi, P. *Anal. Bioanal. Chem.* **2017,** *409,* 4483–4505. doi:10.1007/s00216-017-0406-7
44. Liu, H.; Zhang, N.; Wan, D.; Cui, M.; Liu, Z.; Liu, S. *Clin. Proteomics* **2014,** *11,* 14. doi:10.1186/1559-0275-11-14
45. Hofmann, J.; Pagel, K. *Angew. Chem., Int. Ed.* **2017,** *56,* 8342–8349. doi:10.1002/anie.201701309
46. Leymarie, N.; Zaia, J. *Anal. Chem. (Washington, DC, U. S.)* **2012,** *84,* 3040–3048. doi:10.1021/ac3000573
47. Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. *J. Proteome Res.* **2008,** *7,* 1650–1659. doi:10.1021/pr7008252
48. Albersheim, P. Technical Report of CarbBank: A structural and bibliographic data base. USA, 1989; https://www.osti.gov/biblio/5715461-m7GJFJ/ (accessed Oct 5, 2020). doi:10.2172/5715461

49. von der Lieth, C.-W.; Freire, A. A.; Blank, D.; Campbell, M. P.; Ceroni, A.; Damerell, D. R.; Dell, A.; Dwek, R. A.; Ernst, B.; Fogh, R.; Frank, M.; Geyer, H.; Geyer, R.; Harrison, M. J.; Henrick, K.; Herget, S.; Hull, W. E.; Ionides, J.; Joshi, H. J.; Kamerling, J. P.; Leeflang, B. R.; Lütteke, T.; Lundborg, M.; Maass, K.; Merry, A.; Ranzinger, R.; Rosen, J.; Royle, L.; Rudd, P. M.; Schloissnig, S.; Stenutz, R.; Vranken, W. F.; Widmalm, G.; Haslam, S. M. *Glycobiology* **2011,** *21,* 493–502. doi:10.1093/glycob/cwq188
50. Ranzinger, R.; Herget, S.; Wetter, T.; von der Lieth, C.-W. *BMC Bioinf.* **2008,** *9,* 384. doi:10.1186/1471-2105-9-384
51. Herget, S.; Ranzinger, R.; Maass, K.; Lieth, C.-W. v. d. *Carbohydr. Res.* **2008,** *343,* 2162–2171. doi:10.1016/j.carres.2008.03.011
52. Matsubara, M.; Aoki-Kinoshita, K. F.; Aoki, N. P.; Yamada, I.; Narimatsu, H. *J. Chem. Inf. Model.* **2017,** *57,* 632–637. doi:10.1021/acs.jcim.6b00650
53. Tiemeyer, M.; Aoki, K.; Paulson, J.; Cummings, R. D.; York, W. S.; Karlsson, N. G.; Lisacek, F.; Packer, N. H.; Campbell, M. P.; Aoki, N. P.; Fujita, A.; Matsubara, M.; Shinmachi, D.; Tsuchiya, S.; Yamada, I.; Pierce, M.; Ranzinger, R.; Narimatsu, H.; Aoki-Kinoshita, K. F. *Glycobiology* **2017,** *27,* 915–919. doi:10.1093/glycob/cwx066
54. Aoki-Kinoshita, K.; Agravat, S.; Aoki, N. P.; Arpinar, S.; Cummings, R. D.; Fujita, A.; Fujita, N.; Hart, G. M.; Haslam, S. M.; Kawasaki, T.; Matsubara, M.; Moreman, K. W.; Okuda, S.; Pierce, M.; Ranzinger, R.; Shikanai, T.; Shinmachi, D.; Solovieva, E.; Suzuki, Y.; Tsuchiya, S.; Yamada, I.; York, W. S.; Zaia, J.; Narimatsu, H. *Nucleic Acids Res.* **2016,** *44,* D1237–D1242. doi:10.1093/nar/gkv1041
55. Tsuchiya, S.; Yamada, I.; Aoki-Kinoshita, K. F. *Bioinformatics* **2019,** *35,* 2434–2440. doi:10.1093/bioinformatics/bty990
56. Alocci, D.; Mariethoz, J.; Gastaldello, A.; Gasteiger, E.; Karlsson, N. G.; Kolarich, D.; Packer, N. H.; Lisacek, F. *J. Proteome Res.* **2019,** *18,* 664–677. doi:10.1021/acs.jproteome.8b00766
57. Hayes, C. A.; Karlsson, N. G.; Struwe, W. B.; Lisacek, F.; Rudd, P. M.; Packer, N. H.; Campbell, M. P. *Bioinformatics* **2011,** *27,* 1343–1344. doi:10.1093/bioinformatics/btr137
58. Congreve, M.; Murray, C. W.; Blundell, T. L. *Drug Discovery Today* **2005,** *10,* 895–907. doi:10.1016/s1359-6446(05)03484-7
59. Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. *Curr. Opin. Struct. Biol.* **2015,** *31,* 64–74. doi:10.1016/j.sbi.2015.03.007
60. Aloy, P.; Russell, R. B. *Proc. Natl. Acad. Sci. U. S. A.* **2002,** *99,* 5896–5901. doi:10.1073/pnas.092147999
61. Tanaka, K.; Aoki-Kinoshita, K. F.; Kotera, M.; Sawaki, H.; Tsuchiya, S.; Fujita, N.; Shikanai, T.; Kato, M.; Kawano, S.; Yamada, I.; Narimatsu, H. *J. Chem. Inf. Model.* **2014,** *54,* 1558–1566. doi:10.1021/ci400571e
62. Westbrook, J. D.; Shao, C.; Feng, Z.; Zhuravleva, M.; Velankar, S.; Young, J. *Bioinformatics* **2015,** *31,* 1274–1278. doi:10.1093/bioinformatics/btu789
63. GitHub repository of Privateer. United Kingdom, 2020; https://github.com/glycojones/privateer (accessed Oct 5, 2020).
64. Lütteke, T.; Frank, M.; von der Lieth, C.-W. *Carbohydr. Res.* **2004,** *339,* 1015–1020. doi:10.1016/j.carres.2003.09.038
65. Lütteke, T.; Frank, M.; von der Lieth, C.-W. *Nucleic Acids Res.* **2005,** *33* (Suppl. 1), D242–D246. doi:10.1093/nar/gki013
66. Szakonyi, G.; Klein, M. G.; Hannan, J. P.; Young, K. A.; Ma, R. Z.; Asokan, R.; Holers, V. M.; Chen, X. S. *Nat. Struct. Mol. Biol.* **2006,** *13,* 996–1001. doi:10.1038/nsmb1161
67. Campbell, M. P.; Peterson, R.; Mariethoz, J.; Gasteiger, E.; Akune, Y.; Aoki-Kinoshita, K. F.; Lisacek, F.; Packer, N. H. *Nucleic Acids Res.* **2014,** *42,* D215–D221. doi:10.1093/nar/gkt1128
68. Kim, H. M.; Park, B. S.; Kim, J.-I.; Kim, S. E.; Lee, J.; Oh, S. C.; Enkhbayar, P.; Matsushima, N.; Lee, H.; Yoo, O. J.; Lee, J.-O. *Cell* **2007,** *130,* 906–917. doi:10.1016/j.cell.2007.08.002

License and Terms

The case for post-predictional modifications in the AlphaFold Protein Structure Database

To the editor — AlphaFold2 has arrived to change workflows in structural biology, for good. However, the algorithm does not account for essential modifications that affect protein structure and function, which gives us only part of the picture. Here we discuss how this omission can be addressed in a relatively straightforward manner, which leads to a complete structural prediction of complex biomolecular systems.

The recent release of the AlphaFold Protein Structure Database[1] by DeepMind and EMBL-EBI marks a major breakthrough in structural biology, as it makes available to the scientific community worldwide highly accurate structural predictions for 20,000 proteins from humans and proteins from 20 other biologically relevant organisms that include *Escherichia coli*. Like many scientists that work on macromolecular structure, we are genuinely excited about this development, yet we feel that there is a non-negligible potential for misinterpretation of its content in its current form. In particular, the protein-only predictions in the AlphaFold database means that cofactors and, most importantly, co- and post-translational modifications are understandably — owing to the scope of the technique — excluded. Among the most relevant co- and post-translational modifications is protein glycosylation — relevant and very visible, as recent studies of the dynamics of a fully glycosylated SARS-CoV-2 spike protein illustrate[2,3]. Indeed, between 50% and 70% of those 20,000 predicted human proteins are believed to be glycosylated[4], but none of this is yet visibly highlighted on the database. Detailed information on the likelihood of modifications is readily available through AlphaFolds's links to Uniprot (https://www.uniprot.org), and thus we strongly encourage the users of this fantastic new resource to check the information available on Uniprot before downloading a model.

Within this framework, we believe that the absence of cofactors and of co- or post-translational modifications in the models in the AlphaFold Protein Structure Database might be remediated through the use of sequence and structure-based comparative studies. Indeed, in the specific case of glycosylation, the algorithms that are implemented by DeepMind have digested inter-residue distances from the Protein Data Bank (PDB)[5], where glycosylated proteins often exhibit either full or partial glycan structures; therefore, the space where unmodeled modifications, such as protein glycosylation, should have appeared is somehow preserved in AlphaFold models, which allows for these structural features to be directly grafted onto a model. To demonstrate the potential of this approach, we have developed proof-of-concept functionality that grafts protein glycosylation from a library of structurally equilibrated glycan blocks, obtained from molecular dynamics[6], onto an AlphaFold model.

This task has been automated and integrated into the new Python interface of the carbohydrate-specific Privateer software[7] and is available to all on its GitHub repository (https://github.com/glycojones/privateer.git). Figure 1 shows AlphaFold model P29016 (depicted in magenta) of a human T cell surface glycoprotein Cd1b, superposed onto the protein's crystal structure PDB 5WL1. The latter was expressed in an insect cell line and it shows a characteristic double core-fucosylation of the N-glycans, which were omitted in Fig. 1 for clarity. The N-glycan our tool grafted onto the AlphaFold model is not just compatible with the available space, but it shows a high complementarity to the protein surface, where the Man6 core is involved with Trp 23 in a CH-π interaction[8], as seen in the crystal structure.

We would like to emphasize that this approach may also be useful to complete the AlphaFold models in the database with other types of modifications. For example, the AlphaFold model P6887, a hemoglobin subunit beta, contains a heme binding site with just enough space for a heme cofactor. Certain structure completions will only be feasible via automated comparative analyses against available structural information — for example, co-translational modifications such as myristoylation[9], or O-GlcNAcylation[10] — while others such as N-glycosylation or tryptophan mannosylation, which rely on consensus sequences, will be more

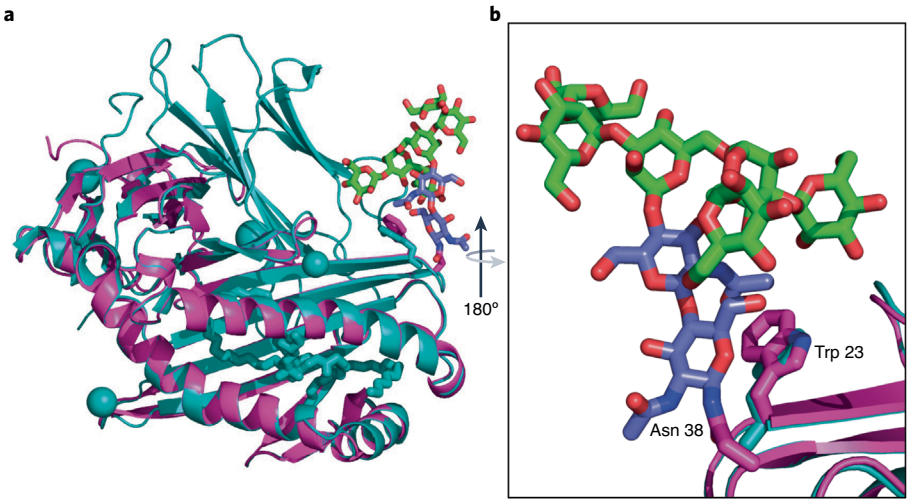


Fig. 1 | Grafting an N-glycan onto an AlphaFold model. a, Structural alignment of the crystal structure of human CD1b in complex with phosphatidylglycerol (PDB 5WL1), shown in cyan, onto the model predicted by AlphaFold (accession code P29016), shown in magenta. The N-glycosylation at position N38 was reconstructed with Privateer[7], where the linked Man6 structure was selected from a library of highly populated conformers at equilibrium, obtained from molecular dynamics simulations at 300 K[6]. **b**, Close-up view of the grafted Man6, with the structure rotated around the *z*-axis by 180°, represented in sticks with colouring compliant to the Symbol Nomenclature for Glycans scheme. The relative positions of the Trp 23 sidechain stacking the Man6 core are highlighted in sticks in both the crystal structure (cyan) and in the AlphaFold model (magenta).

amenable to prediction. As comparative studies would have to rely on experimental structural information, positional uncertainty (for example, a pLDDT-like score[11]) may be estimated by comparing the placed coordinates to a superposition of the available structural information. However, in the particular case of protein glycosylation, we see more of a compositional problem; indeed, the biggest challenge would be to get a good estimation of what glycoform is linked to each sequon. Experimental structures offer only partial information owing to limiting factors such as mobility and micro-heterogeneity[12], so other sources of knowledge (for example, glycomics and molecular dynamics simulations) ought to be used, especially when attempting to model full-length glycans, which is something we are sure the glycobiology community will appreciate. We are expanding the Privateer software to address these cases, by harnessing the rich information available in glycomics databases[13].

To conclude, we think that these early results are highly encouraging to serve as a rallying point for the developers' community to complete and enrich the predicted protein models with likely modifications, to bring them to their fullest potential and to correctly inform the next generation of structural biology studies. ❐

Haroldas Bagdonas[1], Carl A. Fogarty[2], Elisa Fadda[2] ✉ and Jon Agirre[1] ✉

[1]*York Structural Biology Laboratory, Department of Chemistry, University of York, York, UK.* [2]*Department of Chemistry and Hamilton Institute, Maynooth University, Maynooth, Ireland.*
✉*e-mail: elisa.fadda@mu.ie; jon.agirre@york.ac.uk*

Published online: 29 October 2021
https://doi.org/10.1038/s41594-021-00680-9

References

1. Tunyasuvunakool, K. et al. *Nature* **596**, 590–596 (2021).
2. Casalino, L. et al. *ACS Cent. Sci.* **6**, 1722–1734 (2020).
3. Turoňová, B. et al. *Science* **370**, 203–208 (2020).
4. An, H. J., Froehlich, J. W. & Lebrilla, C. B. *Curr. Opin. Chem. Biol.* **13**, 421–426 (2009).
5. Berman, H., Henrick, K. & Nakamura, H. *Nat. Struct. Biol.* **10**, 980 (2003).
6. Fogarty, C. A. & Fadda, E. *J. Phys. Chem. B* **125**, 2607–2616 (2021).
7. Agirre, J. et al. *Nat. Struct. Mol. Biol.* **22**, 833–834 (2015).
8. Hudson, K. L. et al. *J. Am. Chem. Soc.* **137**, 15152–15160 (2015).
9. Udenwobele, D. I. et al. *Front. Immunol.* **8**, 751 (2017).
10. Zhu, Y. et al. *Chem. Biol.* **11**, 319–325 (2015).
11. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
12. Atanasova, M., Bagdonas, H. & Agirre, J. *Curr. Opin. Struct. Biol.* **62**, 70–78 (2020).
13. Bagdonas, H., Ungar, D. & Agirre, J. *Beilstein J. Org. Chem.* **16**, 2523–2533 (2020).

Acknowledgements

H.B. is funded by The Royal Society grant RGF/R1/181006. J.A. is the Royal Society Olga Kennard Research Fellow award ref. UF160039. C.A.F. is funded by the Irish Research Council (IRC) Government of Ireland Postgraduate Scholarship Programme. Data and methods are available at https://doi.org/10.5281/zenodo.5290624

Competing interests

The authors declare no competing interests.

STRUCTURAL BIOLOGY

ISSN 2059-7983

Analysis and validation of overall *N*-glycan conformation in *Privateer*

Jordan S. Dialpuri,[a] Haroldas Bagdonas,[a] Mihaela Atanasova,[a] Lucy C. Schofield,[a] Maarten L. Hekkelman,[b] Robbie P. Joosten[b]* and Jon Agirre[a]*

[a]York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, United Kingdom, and [b]Oncode Institute and Division of Biochemistry, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. *Correspondence e-mail: r.joosten@nki.nl, jon.agirre@york.ac.uk

Received 25 November 2022
Accepted 17 April 2023

Edited by M. Vollmar, Diamond Light Source, United Kingdom

Keywords: glycobiology; validation; *Privateer*; *N*-glycans; torsion angles.

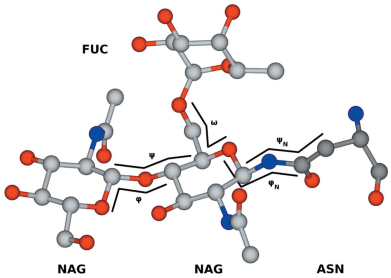
Supporting information: this article has supporting information at journals.iucr.org/d

The oligosaccharides in *N*-glycosylation provide key structural and functional contributions to a glycoprotein. These contributions are dependent on the composition and overall conformation of the glycans. The *Privateer* software allows structural biologists to evaluate and improve the atomic structures of carbohydrates, including *N*-glycans; this software has recently been extended to check glycan composition through the use of glycomics data. Here, a broadening of the scope of the software to analyse and validate the overall conformation of *N*-glycans is presented, focusing on a newly compiled set of glycosidic linkage torsional preferences harvested from a curated set of glycoprotein models.

1. Introduction

Post-translational modifications (PTMs) are covalent modifications of proteins that occur after the nascent polypeptide has left the ribosome. PTMs may induce significant changes in the structure and function of the protein (Xin & Radivojac, 2012). A fundamental and abundant PTM is *N*-glycosylation, in which an oligosaccharide moiety is attached to the N atom of an asparagine side chain in the target protein. The oligosaccharide is subsequently trimmed and modified according to the available cellular enzymes: glycoside hydrolases, glycosyl and oligosaccharyl transferases. The resulting oligosaccharide, or *N*-glycan, may end up having anything from a complex to a minimal composition, leading to a specific 3D conformation of the mature glycoprotein (Shental-Bechor & Levy, 2009). *N*-Glycosylation is key to all sorts of interactions, including those with cell-surface receptors (Petrescu *et al.*, 2006; Rudd *et al.*, 2004) or even other parts of the same glycoprotein, as shown in studies of the dynamics of SARS-CoV-2 spike, where conformational changes in the Asn165 glycan push up the receptor-binding domain of the spike (Casalino *et al.*, 2020).

Understanding the complex structure of carbohydrates is challenging due to the various stereochemical and regiochemical possibilities exhibited by *N*-glycans. Producing a correct 3D structure of a glycoprotein at a good enough resolution can be vital in understanding how some biological processes unfold. Alas, working with glycans in software for X-ray crystallography and electron cryo-microscopy has historically been all but straightforward: many carbohydrate modelling, refinement and validation processes relied on software written primarily for proteins and nucleic acids (Atanasova *et al.*, 2020), and libraries of restraints had become outdated or were incorrect (Agirre, 2017). While recent efforts have aimed to address this situation (Atanasova *et al.*, 2022; Joosten *et al.*, 2022), carbohydrate methodology still trails that designed for proteins.



Obtaining a glycoprotein structure at a high enough resolution can generally be considered to be more difficult than with a glycan-free protein. Two main issues are routinely identified as problematic when it comes to obtaining higher resolutions: heterogeneity and mobility, both of which translate into poorer experimental data. Owing to these complications, the Protein Data Bank (PDB; Berman *et al.*, 2000) contains models that include incorrect nomenclature (Lütteke *et al.*, 2005), impossible linkages (Crispin *et al.*, 2007) and improbably high-energy conformations of carbohydrates that deviate from the low-energy chair conformation of six-membered rings (Agirre, Davies *et al.*, 2015): in general, a $^4C_1$ chair for D-pyranosides and a $^1C_4$ chair for L-pyranosides. Ring conformations (Cremer & Pople, 1975) and their energetics (Davies *et al.*, 2012) are discussed in detail elsewhere. Using models with incorrect glycochemistry in downstream analyses or molecular simulations will cause misrepresentation and misinterpretation, while also perpetuating these errors. Software packages such as *pdb-care* and *CARP* (Lütteke *et al.*, 2005), and more recently *Privateer* (Agirre, Iglesias-Fernández *et al.*, 2015; Bagdonas *et al.*, 2020), can be utilized for the identification and rectification of these model errors, therefore allowing future refinement data libraries to be as accurate and representative as possible.

In this study, torsion angles (dihedral angles) in curated structures of *N*-glycan-forming pyranosides were collected in order to create accurate torsional libraries for use in the *Privateer* validation software. Previous torsional databases such as GlyTorsionDB (Lütteke *et al.*, 2005) and its associated link-checking tool (*CARP*) incorporate potentially flawed models from the PDB, as they pre-dated the introduction of ring conformation into the routine validation of glycan structures (Agirre, Iglesias-Fernández *et al.*, 2015); therefore, a survey of the PDB was completed with each PDB entry being analysed and validated using *Privateer* to ensure that the *N*-glycans were well fitted to the electron density without any conformational errors. Also, in order to avoid the presentation of data on multiple torsional plots and to allow the easy identification of standout (outlier) linkage conformations, a *Z*-score is calculated for each linkage, with standout linkages being highlighted in orange on glycan diagrams that follow the third edition of the Standard Symbol Nomenclature for Glycans (SNFG; Varki *et al.*, 2015). Furthermore, in recognition that not every standout linkage conformation will be the consequence of a modelling mistake, a collection of verified cases where the interaction between glycan and protein residues has caused an unusual conformation is presented. Finally, a similar study was completed using *PDB-REDO* (van Beusekom, Touw *et al.*, 2018) to analyse whether modern refinement techniques can lead to less frequent errors in the *N*-glycan models.

2. Materials and methods

2.1. Data-set collection and validation

A local PDB mirror (August 2021) was created for this study. The PDB mirror was then scanned for proteins containing glycosylated amino-acid residues. Of the monosaccharides contained within these chains, the conformations of the six-membered rings (pyranosides) were validated using *Privateer*: the software calculates ring conformation using the Cremer–Pople algorithm (Cremer & Pople, 1975) and then compares the detected ring conformation with the minimal energy conformation stored in an internal database. The data set was filtered to include only monosaccharides with a real-space correlation coefficient (RSCC) higher than 0.80 [RSCC (equation 1) is a measure of the local agreement between a portion of an atomic model and the observed electron-density map that surrounds it] and which had been deemed diagnostically correct by *Privateer*, *i.e.* no nomenclature errors, no unphysical puckering amplitude and all pyranosides in their minimal energy conformations (a chair in all analysed cases). *Privateer* checks that the anomeric and absolute stereochemistry in the structure matches that encoded in the three-letter code (for example that a monosaccharide modelled as MAN is perceived to be α-D-mannose), that the ring conformation matches the lowest energy pucker, which is a $^4C_1$ chair for most D-pyranosides, with special cases such as $^1C_4$ for the mannose moiety in tryptophan mannosylation (Akkermans *et al.*, 2022; Frank *et al.*, 2020), including puckering amplitude (Cremer & Pople, 1975).

$$\mathrm{RSCC} = \mathrm{corr}(\rho_{\mathrm{obs}}, \rho_{\mathrm{calc}}) = \frac{\mathrm{cov}(\rho_{\mathrm{obs}}, \rho_{\mathrm{calc}})}{[\mathrm{var}(\rho_{\mathrm{obs}})\mathrm{var}(\rho_{\mathrm{calc}})]^{1/2}}. \quad (1)$$

No resolution cutoffs were explicitly applied, although some filtering is implicit in requiring a minimum RSCC, as the accumulation of model-error components at low resolutions makes it harder to obtain high RSCC values. A total of 68 541 monosaccharides were analysed, 57 569 of which *Privateer* deemed correct; only these were used in the study. A further 8511 showed a high-energy ring conformation, which normally requires manual assessment. A total of 2421 monosaccharides showed geometry and/or nomenclature errors.

For the *PDB-REDO* comparison, the equivalent monosaccharides were taken from the so-called 'conservatively optimized' models in the PDB-REDO databank (van Beusekom, Touw *et al.*, 2018), *i.e.* models that were re-refined without any torsional restraints for carbohydrates but were not subjected to *N*-glycan rebuilding procedures (van Beusekom *et al.*, 2019).

Example linkages present in diverse glycans are shown in Fig. 1 using the third edition of the SNFG (Varki *et al.*, 2015), which *Privateer* implements. The definition of φ and ψ for *N*-acetyl-β-D-glucosamine (GlcNAc, or NAG in the PDB Chemical Component Dictionary) linked to asparagine, plus all 1–2, 1–3 and 1–4 glycosidic bonds, and additionally ω, which covers 1–6 bonds such as in α-D-mannose–1,6–α-D-mannose and α-L-fucose–1,6–*N*-acetyl β-D-glucosamine, is shown in Fig. 2. While completing this study, a large array of different linkages were identified; however, only a small number had enough independent observations to enable meaningful data extraction. Indeed, only approximately 10% of protein models deposited in the PDB contain one or more carbohydrate

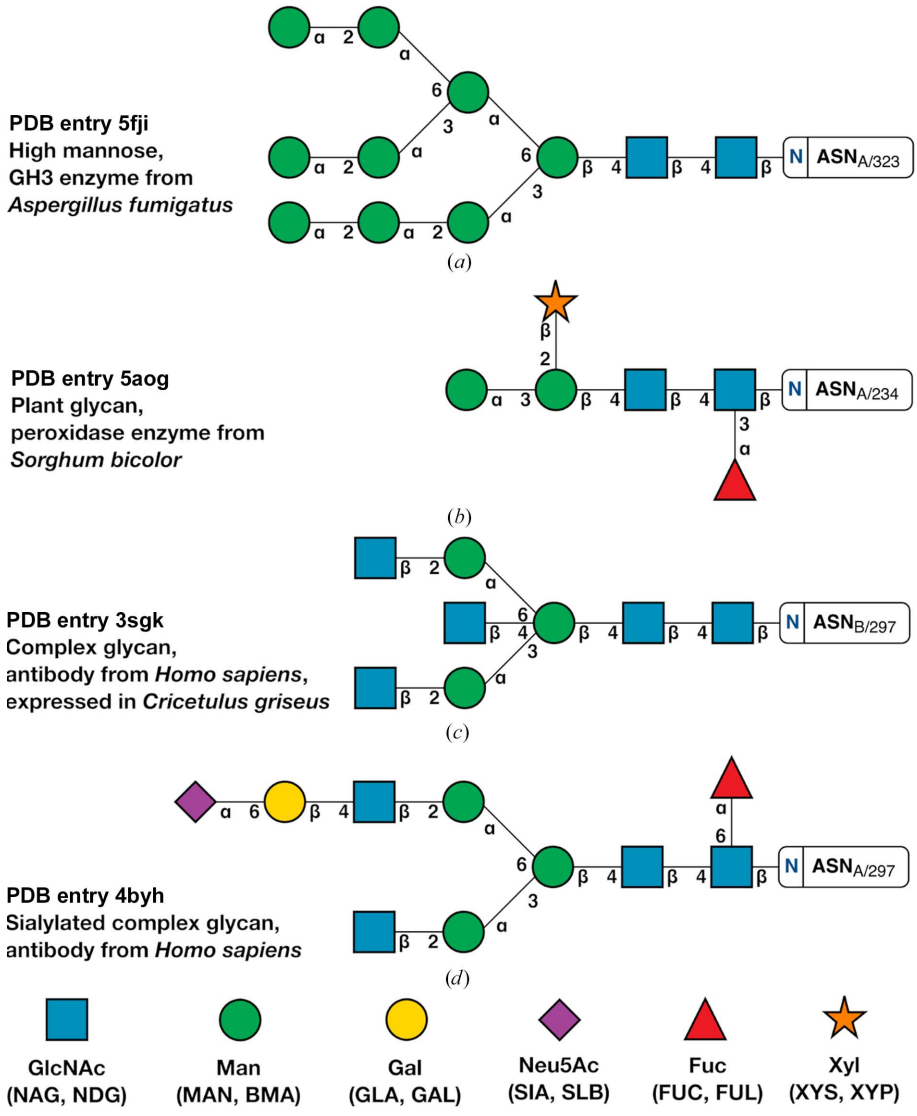


Figure 1
Examples of different types of *N*-glycans shown using the Symbol Nomenclature for Glycans (SNFG). The Greek letters and numbers show the *N*-glycan linkage naming. (*a*) High mannose from PDB entry 5fji, a GH3 glucosidase from *Aspergillus fumigatus* (Agirre *et al.*, 2016). (*b*) Plant glycan from PDB entry 5aog, a sorghum peroxidase (Nnamchi *et al.*, 2016). (*c*) PDB entry 3sgk (Ferrara *et al.*, 2011) shows a complex glycan from an Fc fragment of a human antibody, which was in turn expressed in CHO cells. (*d*) A sialylated complex glycan from PDB entry 4byh (Crispin *et al.*, 2013) expressed in *Homo sapiens*. This figure was produced with *Privateer*, which follows SNFG version 3 (Varki *et al.*, 2015).

groups, while around 6% are *N*-glycosylated (Agirre, 2017). We set the minimum number of required observations to 50 and introduced a mechanism for *Privateer* to report which linkages could not be validated due to insufficient data (see below). A table of the linkages investigated in this study is given as Table 1, as well as the commonly used abbreviations associated with them.

2.2. Implementation in *Privateer*

To assess the normality of the torsion angles between monosaccharides in *N*-glycans, a *Z*-score system was implemented using similar methods to the *Tortoize* (van Beusekom, Joosten *et al.*, 2018) and *WHAT_CHECK* (Hooft *et al.*, 1997) software. The *Z*-score is based on how common a certain (φ, ψ) combination is compared with a reference set of the same glycosidic linkages calculated from high-quality structure models. To calculate the *Z*-scores, torsional data from each linkage were split into two-dimensional bins with a 2° bin spacing and formed into a database. The *Z*-score is calculated as described by Hooft *et al.* (1997) and shown in equation (2).

$$z_k = \frac{c_k^l - \langle c^l \rangle}{\sigma(c^l)}. \qquad (2)$$

Let k be a particular glycosidic linkage, for example BMA402–NAG401 in a PDB file, under scrutiny and z_k be its *Z*-score for the φ/ψ torsion pair measured on the structure; l is the linkage type (Man–β1,4–GlcNAc in this case), c_k^l is the number of data points of that linkage (where c is a count) in the 2° × 2° bin corresponding to the φ/ψ torsion pair in the database, $\langle c_l \rangle$ is the average number of data points for that linkage across all bins and $\sigma\langle c_l \rangle$ is the corresponding standard deviation of the

Table 1
Full names, linkage abbreviations and shorthand notations with PDB Chemical Component Dictionary (CCD) codes for those linkages with sufficient data.

No anomeric data are displayed for CCD codes, as this information is integrated into the codes themselves; for example MAN is α-D-mannose and BMA is β-D-mannose.

Full linkage denomination	Abbreviation	CCD code
N-Acetyl-β-D-glucosamine–asparagine	GlcNAc–β–Asn	NAG-ASN
N-Acetyl-β-D-glucosamine–1,4–*N*-acetyl-β-D-glucosamine	GlcNAc–β–GlcNAc	NAG-1,4-NAG
β-D-Mannose–1,4–*N*-acetyl-β-D-glucosamine	Man–β1,4–GlcNAc	BMA-1,4-NAG
α-D-Mannose–1,3–β-D-mannose	Man–α1,3–Man	MAN-1,3-BMA
α-D-Mannose–1,6–β-D-mannose	Man–α1,6–Man	MAN-1,6-BMA
α-D-Mannose–1,2–α-D-mannose	Man–α1,2–Man	MAN-1,2-MAN
α-D-Mannose–1,3–α-D-mannose	Man–α1,3–Man	MAN-1,3-MAN
α-D-Mannose–1,6–α-D-mannose	Man–α1,6–Man	MAN-1,6-MAN
α-L-Fucose–1,3–*N*-acetyl-β-D-glucosamine	Fuc–α1,3–GlcNAc	FUC-1,3-NAG
α-L-Fucose–1,6–*N*-acetyl-β-D-glucosamine	Fuc–α1,6–GlcNAc	FUC-1,6-NAG
N-Acetyl-β-d-glucosamine–1,2–α-D-mannose	GlcNAc–β1,2–Man	NAG-1,2-MAN
β-D-Galactose–1,4–*N*-acetyl-β-D-glucosamine	Gal–β1,4–GlcNAc	GAL-1,4-NAG
α-Sialic acid–2,6–β-D-galactose	Sia–α2,6–Gal	SIA-2,6-GAL

number of data points for linkage l in the database, again across all bins. As derived from the formula, positive Z-scores indicate that the φ/ψ torsion pair is well represented in the database and thus normal, whereas negative Z-scores indicate the opposite. Also, the scores are normalized to make the results comparable between different linkages. Detailed results and their interpretation are discussed in the next section.

After scoring every glycosidic linkage, a global Z-score may be calculated by simply averaging the Z-scores of all N-glycan linkages. In addition to this, comparison to a reference set of PDB entries with N-glycans allowed the calculation of a relative 'quality Z-score', which is an additional parameter that can be used as a measure of glycan normality. The reference set was chosen following a set of criteria: crystallographic structures and reflections from the wwPDB with $R_{\rm free} < 0.25$ and reported resolution ≤ 2.50 Å, with glycans longer than four pyranosides and with a composition backed up by a GlyConnect ID (Alocci *et al.*, 2019). As a result, 510 structures were chosen containing 59 unique glycan structures. The resolution range covered by the data set was 1.12–2.50 Å, and the $R_{\rm work}$ and $R_{\rm free}$ values were in the ranges 0.10–0.23 and 0.12–0.25, respectively.

To provide a visual means of highlighting those linkages with an unusual Z-score, the SNFG (Varki *et al.*, 2015) vector engine within *Privateer* (McNicholas & Agirre, 2017) was modified to create an orange background behind the linkages. Linkages for which insufficient data could be collected for validation are marked with a grey background. This representation was used in the figures presented in this study. The representation was also extended to cover the monosaccharides in glycans, so that interesting or problematic models can quickly be identified. We note that an orange background does not automatically mean that there is a modelling mistake, but rather that the linkage is worth inspecting.

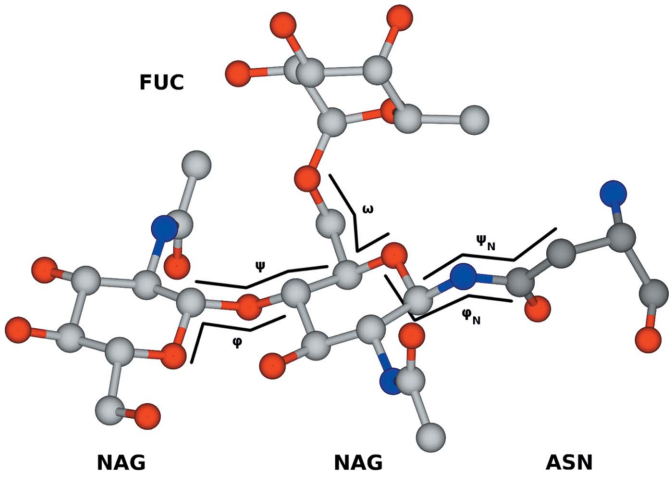


Figure 2
Visual representation of φ and ψ in both sugar–sugar linkages and the NAG-ASN linkage. This figure was generated from PDB entry 4byh (Crispin *et al.*, 2013).

3. Results and discussion

The number of N-glycosylated structures in the PDB is growing steadily (Scherbinina & Toukach, 2020; Agirre, 2017), supported by the introduction of carbohydrate structure modelling and validation tools such as *pdb-care* (Lütteke & von der Lieth, 2004), the N-glycan building module in *Coot* (Emsley & Crispin, 2018) and *Privateer* (Agirre, Iglesias-Fernández *et al.*, 2015). However, as the resolvability of pyranosides in N-glycans decreases the further the monosaccharides are from the asparagine residue (Atanasova *et al.*, 2020), the abundance of the data collected here dwindles for linkages that form the antennae of the glycans. As stated previously, we set a cutoff of 50 data points in order to guarantee the reliability of the Z-score calculation, and this necessarily means that some glycosidic linkages are not yet included in the analysis performed by the *Privateer* software. Scripts for reproducing and extending this work are included in the relevant section here, meaning that the torsion library can be regenerated in future when more data are available.

The torsional data that we harvested are plotted in Fig. 3. A first close inspection of the graphs reveals a straightforward correspondence between the most frequent linkage conformations for every link type and their calculated energy minimum or minima in the Disac3-DB section of the Glyco3D 2.0 database (Pérez *et al.*, 2015) and GlycoMapsDB (Frank *et al.*, 2007). The mean linkage torsion angles and respective standard deviations of this PDB survey are shown in

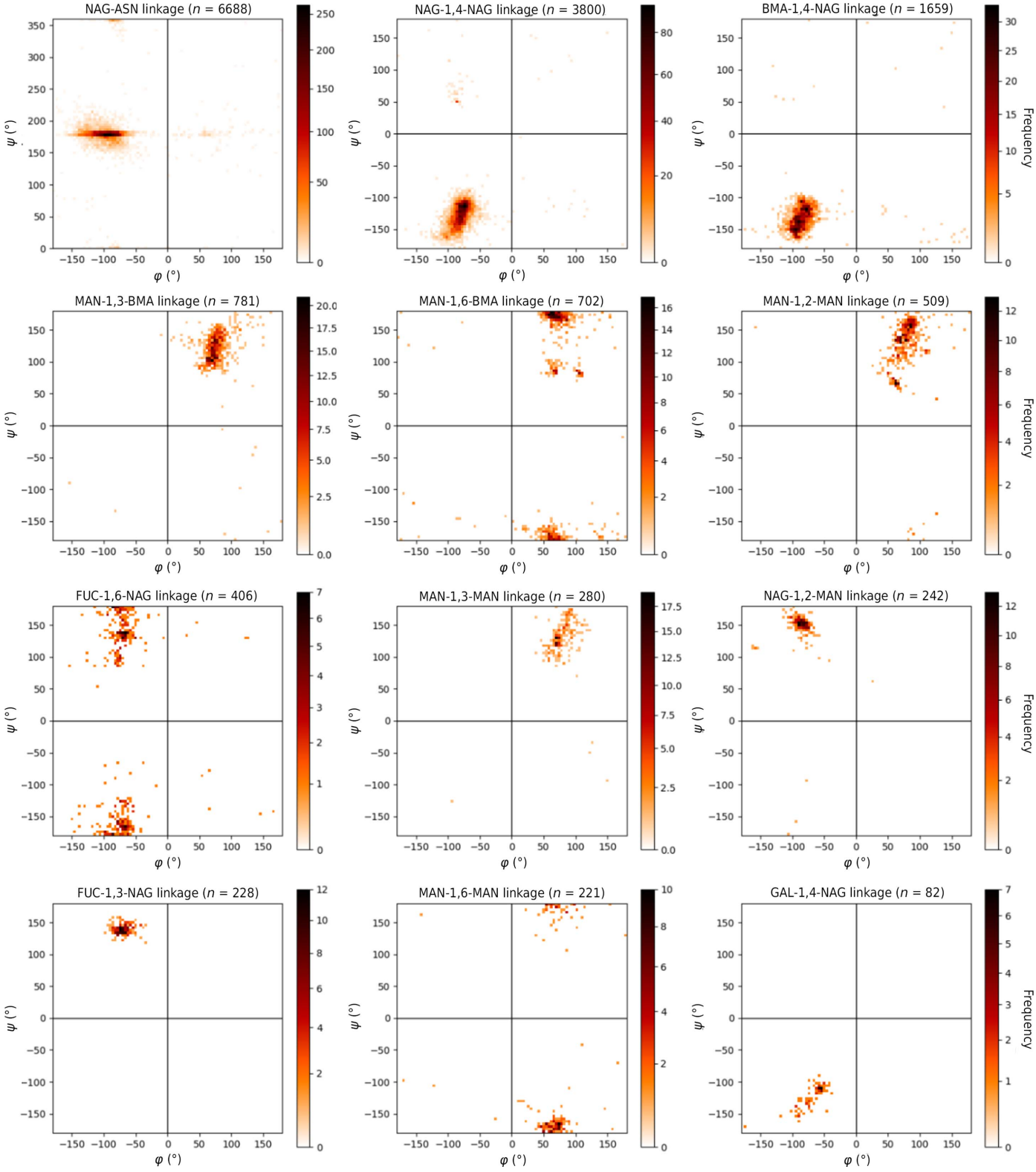


Figure 3
Plots of φ and ψ values for all linkages collected with over 50 data points. Colour bars are plotted using the power-law distribution (Clauset *et al.*, 2009) to highlight outliers visually. Plots allow visualization of the energy-minima values.

Supplementary Tables S1 and S2. Supplementary Table S1 shows the values implemented into *Privateer*. A comparative plot of quality *Z*-scores for the curated data set versus the rest of the PDB is available in Supplementary Fig. S1. Low-quality *Z*-scores ($Z < -2$) may indicate serious problems with the overall quality of glycans in the structure model. High-quality *Z*-scores ($Z > 2$), particularly in low-resolution structure models, may indicate over-restraining of torsions in model refinement and may warrant further inspection, as previously shown for proteins (Sobolev *et al.*, 2020).

3.1. GlcNAc–asparagine bond

Investigations of the torsion-angle data set between the asparagine (ASN) amino-acid side chain and GlcNAc (NAG) highlight a perhaps unsurprising trend. The φ torsion-angle data set has a greater standard deviation ($\sigma = 25.3°$) when compared with the ψ torsion angle ($\sigma = 22.1°$). This is most likely due to the ψ torsion angle referring to a C—N bond which has a bond order of greater than one, analogous to a peptide bond. Indeed, the mean value of ψ is 178.5°, which is very similar to the 180° torsion angle expected for a peptide bond. Such a bond has limited torsional freedom. The φ torsion angle refers to a single bond which has more rotational freedom, leading to the increased spread of torsional data for φ.

Correct modelling of the protein–sugar linkage torsion angle is particularly important to establish a good basis for other monosaccharides to be modelled further down the *N*-glycan tree. Two main conformations for NAG-ASN exist (Fig. 4), in which the conformation with a negative φ angle (Fig. 4*a*) is the most abundant and the other conformation (Fig. 4*b*), which is much more infrequent due to the additional CH–π interaction (Trp431) that is required to stabilize it, is flagged up as an outlier by *Privateer*. The arrangement shown in Fig. 4(*b*), found in a fungal GH3 β-glucosidase, is conserved across homologous structures.

3.2. Glycosidic linkages between pyranosides

N-Glycans exhibit common structures, as shown in Fig. 1. The similarity of these conformations explains the consistency in the types of linkages seen in various glycoproteins and allows this quantitative study. *N*-Glycosylated chains attach to the residue with a NAG sugar through a β-linkage. Attached to this initial NAG sugar through a β-1,4 linkage is an additional NAG sugar. This initial NAG-1,4-NAG linkage is

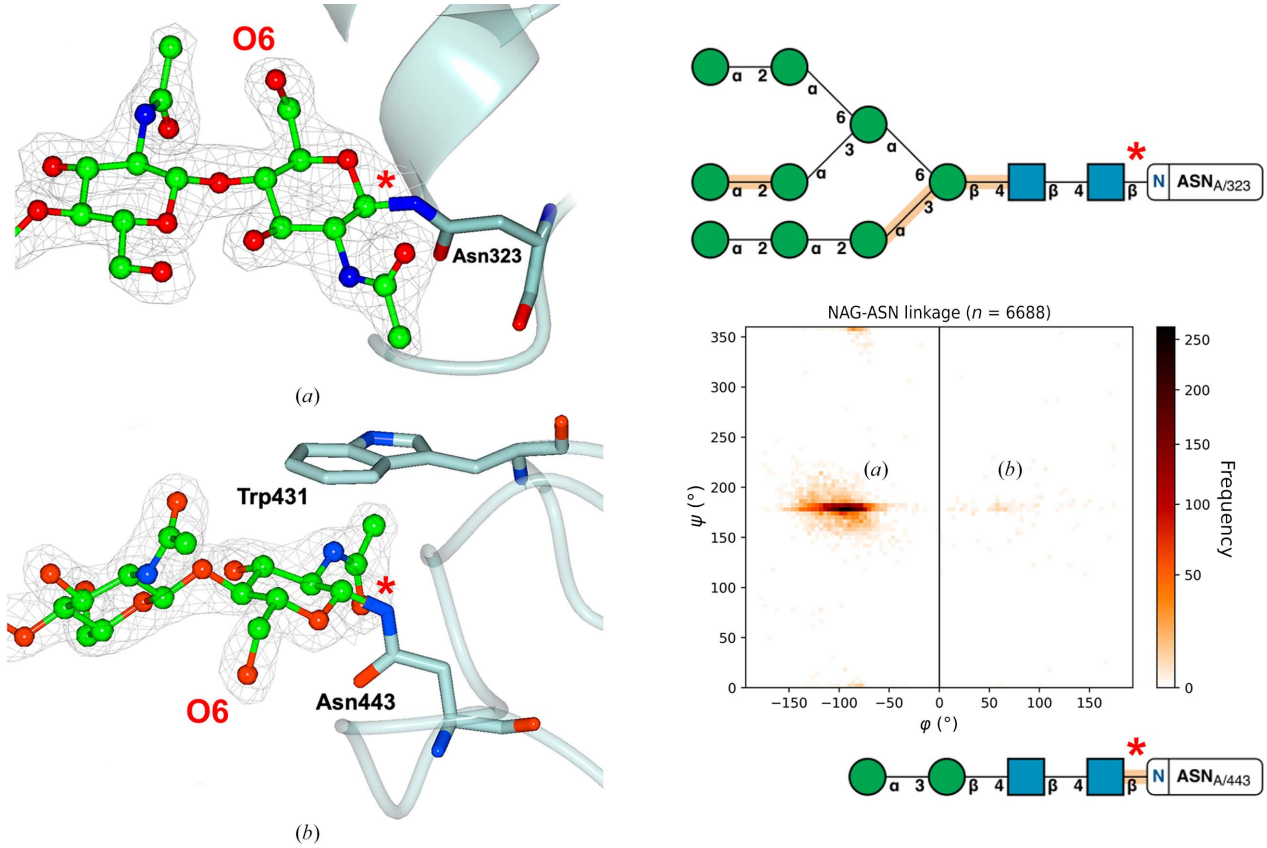


Figure 4
Two main conformations for the NAG-ASN bond are detected in our data set, as previously shown in the literature (Imberty & Perez, 1995). (*a*) shows the most frequent conformation (see the graph on the right for an idea of the numbers), with (*b*) showing a secondary and much more infrequent preference. In (*b*) the GlcNAc appears flipped with respect to the orientation it has in (*a*); this can be spotted easily by looking at O6 of GlcNAc (annotated in the figure), which appears on opposite sides of the asparagine side chain. Both shown conformations are from PDB entry 5fji (Agirre *et al.*, 2016); $2mF_o - DF_c$ electron density is shown at 1σ for the glycans, but is omitted for the asparagine side chains for reasons of clarity; the positions of the asparagine side chains showed a good fit to the electron density.

Table 2
Comparison between the PDB and PDB-REDO torsional data.

Values have been rounded to the nearest integer due to the large deviations that were encountered.

		φ (°)		ψ (°)		
Resolution (Å)	Linkage	PDB	PDB-REDO	PDB	PDB-REDO	No. of entries
$x < 1.50$	NAG-1,4-NAG	−79 ± 8	−79 ± 24	−127 ± 18	−126 ± 26	132
$1.50 < x < 3.00$	NAG-1,4-NAG	−80 ± 13	−74 ± 24	−127 ± 23	−125 ± 24	3190
$x > 3.00$	NAG-1,4-NAG	−83 ± 24	−67 ± 36	−130 ± 23	−135 ± 27	472
All	**NAG-1,4-NAG**	**−80 ± 14**	**−73 ± 26**	**−127 ± 23**	**−126 ± 25**	**3800**
$x < 1.50$	BMA-1,4-NAG	−82 ± 10	−84 ± 10	−125 ± 14	−122 ± 13	37
$1.50 < x < 3.00$	BMA-1,4-NAG	−87 ± 16	−79 ± 29	−133 ± 18	−136 ± 23	1369
$x > 3.00$	BMA-1,4-NAG	−85 ± 26	−65 ± 47	−134 ± 21	−142 ± 26	250
All	**BMA-1,4-NAG**	**−87 ± 18**	**−77 ± 32**	**−133 ± 18**	**−137 ± 24**	**1659**
$x < 1.50$	MAN-1,6-BMA	69 ± 6	70 ± 5	150 ± 45	149 ± 45	17
$1.50 < x < 3.00$	MAN-1,6-BMA	72 ± 24	67 ± 24	167 ± 33	167 ± 34	606
$x > 3.00$	MAN-1,6-BMA	79 ± 42	66 ± 41	177 ± 31	179 ± 34	75
All	**MAN-1,6-BMA**	**72 ± 27**	**66 ± 26**	**168 ± 33**	**168 ± 35**	**702**
$x < 1.50$	MAN-1,3-BMA	77 ± 14	76 ± 14	122 ± 21	122 ± 21	23
$1.50 < x < 3.00$	MAN-1,3-BMA	75 ± 16	69 ± 20	121 ± 21	126 ± 23	602
$x > 3.00$	MAN-1,3-BMA	82 ± 21	68 ± 26	125 ± 30	135 ± 34	130
All	**MAN-1,3-BMA**	**76 ± 17**	**69 ± 21**	**121 ± 23**	**127 ± 26**	**777**
$x < 1.50$	MAN-1,6-MAN	60 ± 6	60 ± 3	−179 ± 6	−177 ± 4	8
$1.50 < x < 3.00$	MAN-1,6-MAN	67 ± 19	65 ± 20	−173 ± 16	−171 ± 16	175
$x > 3.00$	MAN-1,6-MAN	83 ± 45	68 ± 46	−174 ± 34	−180 ± 49	38
All	**MAN-1,6-MAN**	**68 ± 25**	**65 ± 25**	**−173 ± 20**	**−173 ± 24**	**221**
$x < 1.50$	MAN-1,2-MAN	73 ± 12	72 ± 12	126 ± 37	125 ± 37	23
$1.50 < x < 3.00$	MAN-1,2-MAN	77 ± 16	70 ± 16	134 ± 33	139 ± 35	387
$x > 3.00$	MAN-1,2-MAN	82 ± 25	71 ± 28	125 ± 26	130 ± 30	94
All	**MAN-1,2-MAN**	**78 ± 18**	**71 ± 19**	**132 ± 32**	**137 ± 35**	**507**
$x < 1.50$	MAN-1,3-MAN	74 ± 5	73 ± 6	118 ± 17	118 ± 18	9
$1.50 < x < 3.00$	MAN-1,3-MAN	77 ± 16	75 ± 16	133 ± 22	135 ± 24	234
$x > 3.00$	MAN-1,3-MAN	89 ± 18	83 ± 19	129 ± 34	130 ± 33	36
All	**MAN-1,3-MAN**	**78 ± 17**	**76 ± 16**	**132 ± 24**	**134 ± 25**	**280**

abundant in the PDB and hence contains a large number (n = 3800) of validated data points. As evident by the two-dimensional histogram (Fig. 3), most NAG-1,4-NAG linkages contain torsion angles around $\varphi \simeq -80°$ and $\psi \simeq -130°$.

Often, a BMA sugar is attached to the second NAG sugar through a β-1,4 linkage. This BMA-1,4-NAG linkage may theoretically have slightly more conformational variability than NAG-1,4-NAG due to its position further down the glycan tree; however, the spread of data (standard deviation) is similar for both NAG-1,4-NAG and BMA-1,4-NAG. In addition to this, in the complex tree a FUC sugar can be attached to the initial NAG through an α-1,6 linkage. The FUC-1,6-NAG linkage exhibits a large standard deviation around both torsion angles, particularly around the ψ angle. This could partially be the result of FUC being a terminal residue at this position in the glycan, but the FUC-1,3-NAG linkage, in which the FUC is also a terminal residue connected to the same NAG, has less spread in the observed torsion angles. A key difference, however, is the presence of a third torsion angle, ω, that gives more flexibility to the FUC-1,6-NAG linkage. This additional flexibility also leads to less well defined experimental data and thus more room for modelling errors.

Attachment of additional mannose sugars onto the *N*-glycan chain can often increase the amount of branching and the size of the chain (see Fig. 1*a*). The most common attachment onto the currently terminal BMA sugar is MAN-1,3-BMA; indeed, this is shown in our data set of validated glycans (n = 781), with the positional isomer MAN-1,6-BMA being almost as frequent (n = 702). Interestingly, the MAN-1,3-BMA linkage exhibits standard deviations (ψ: σ = 22.6°) which are similar to those of NAG-1,4-NAG (ψ: σ = 22.8°). However, the MAN-1,6-BMA linkage torsion angles do not exist in a singular cluster and hence exhibit a larger standard deviation (ψ: σ = 33.3°). Again, this additional spread may be caused by the presence of a third torsion angle in the linkage.

Certain glycoproteins have further monosaccharide attachments such as a variety of MAN-MAN, NAG-MAN and SIA-GAL linkages. Interestingly, the torsion-angle spread for all MAN-MAN linkages (1,2, 1,3 and 1,6) is far greater than the torsion-angle spread for NAG-MAN torsion-angle data, despite having a similar data-set size and existing in a similar area of the protein. A reason for this may be the *N*-acetyl group in NAG, which makes the placement of the monomer in relatively poor density less error-prone. The large standard deviation of MAN-MAN linkages causes similar challenges to

Table 3
Results of *t*-tests between the PDB and PDB-REDO data sets at all resolutions.

Values that are not significantly different ($p > 0.05$) are shown in bold.

Linkage	Resolution range (Å)	*t*-test result: φ	*t*-test result: ψ
NAG-1,4-NAG	0.93–6.92	Significantly different ($p \leq 0.05$)	Significantly different ($p \leq 0.05$)
BMA-1,4-NAG	1.20–8.69	Significantly different ($p \leq 0.05$)	Significantly different ($p \leq 0.05$)
MAN-1,6-BMA	1.20–6.92	Significantly different ($p \leq 0.05$)	**Not significantly different ($p = 0.34$)**
MAN-1,3-BMA	1.20–6.92	Significantly different ($p \leq 0.05$)	**Not significantly different ($p = 0.39$)**
MAN-1,6-MAN	1.12–6.31	**Not significantly different ($p = 0.14$)**	**Not significantly different ($p = 0.35$)**
MAN-1,2-MAN	1.20–6.92	Significantly different ($p \leq 0.05$)	**Not significantly different ($p = 0.18$)**
MAN-1,3-MAN	1.20–6.31	**Not significantly different ($p = 0.12$)**	**Not significantly different ($p = 0.56$)**

MAN-BMA linkages in torsional restraint application. As well as this, no apparent cluster was observed for the SIA-GAL linkage, most likely due to the very low number of deposited and curated linkages available in the data set. The values that φ can adopt appear to be determined by the anomeric form involved in the glycosidic linkage: for D-pyranosides this means $-180° < \varphi < 0°$ for β-anomers and $0° < \varphi < 180°$ for α-anomers. The inverse is true for L-pyranosides.

Using this large torsion-angle data set, an investigation of torsion-angle spread with glycan chain length and branching was conducted, although no meaningful trend was identified between glycan chain length and torsion-angle standard deviation. Despite this, this large data set can be incorporated into software packages such as *Privateer* to improve the accuracy of glycoprotein models.

3.3. *PDB-REDO* analysis

With the increasingly commonplace solution of protein complexes with high-resolution data, it is imperative that model-building software can depict the conformation and position of *N*-glycans accurately. Through the comparison of *N*-glycan torsion angles of proteins deposited in the PDB and the PDB-REDO databank, the applicability and necessity of modern refinement techniques can be assessed. Comparisons between torsion angles in *N*-glycans deposited in the PDB and the PDB-REDO databank highlight an interesting relationship between structure resolution and torsion-angle accuracy, as shown in Table 2.

The PDB-REDO models used in this study had no torsional restraints applied during refinement. Therefore, the torsion angles calculated by *PDB-REDO* are not influenced by the potentially flawed torsional restraints applied before the model was initially deposited in the PDB. This application of consistent refinement techniques without torsional restraints leads to a data set which naturally has a larger spread than the PDB. To assess whether the PDB and PDB-REDO data sets are significantly different, a series of *t*-tests were performed and are summarized in Table 3.

For the NAG-1,4-NAG and BMA-1,4-NAG linkages, both mean torsion angles were deemed to be significantly different ($p < 0.05$) in the PDB and PDB-REDO data sets by the *t*-test. For the MAN-1,6-BMA linkage, while the φ angle was deemed to be significantly different, the ψ angle was not significantly different. Interestingly, both data sets showed no significant difference between both torsion angles for MAN-1,6-MAN and MAN-1,3-MAN linkages. While the PDB-REDO models had many occurrences in which the torsion angles were not statistically similar to those in the PDB data set, the torsion angles in both data sets are within one standard deviation of each other for every linkage. While it is impossible to automatically determine whether the glycosidic linkages in a deposited structure were restrained to certain values, we know

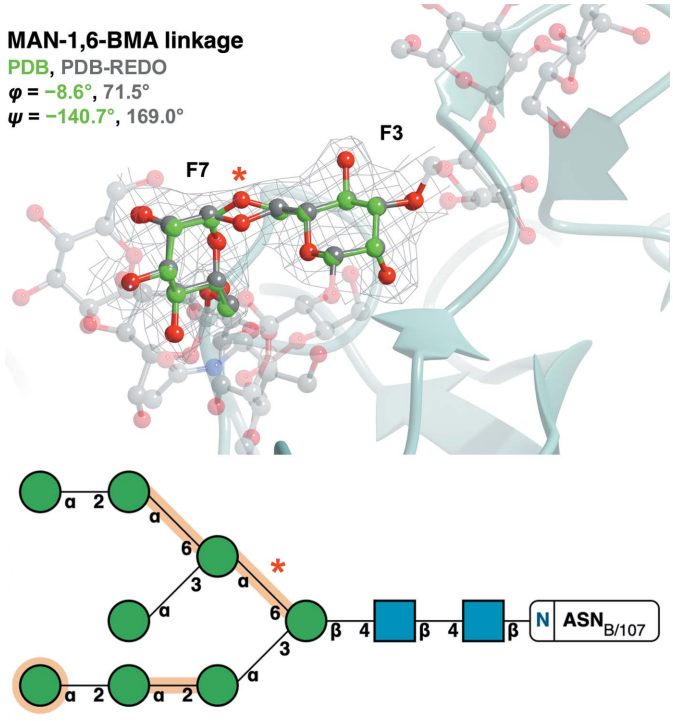


Figure 5
Refinement of PDB entry 6s2g (Ramirez-Escudero *et al.*, 2019) in *PDB-REDO* changes the torsion angle from an outlier in the PDB to an inlier in the PDB-REDO databank. The MAN (chain ID and sequence number F7)-1,6-BMA (chain ID and sequence number F3) linkage (red asterisk in the bottom panel) of PDB entry 6s2g (green) is identified as an outlier in the PDB ($\varphi = -8.6°$, $\psi = -140.7°$) but as an inlier in the PDB-REDO databank ($\varphi = 71.5°$, $\psi = 169.0°$): $\Delta\varphi = 80.1°$, $\Delta\psi = 50.3°$. The change is brought on by moving the O6 atom (red asterisk in the top panel). BMA(F3) and MAN(F7) are represented by ball-and-stick models [C atoms in green (PDB model) or grey (PDB-REDO model)], whilst the rest of the attached glycan (PDB-REDO model) is represented in a faded grey ball-and-stick representation. $2F_o - F_c$ electron density (grey) is displayed for the linkage contoured to 1σ. The *Z*-scores for this linkage is −1.03 in the PDB model and 1.53 in the PDB-REDO model. The top image was produced using *CCP4mg*. Bottom: SNFG notation output from *Privateer*.

that *PDB-REDO* does not apply torsional restraints. Hence, in the absence of potential bias towards torsion restraint targets, it is likely that the PDB-REDO databank represents a more realistic distribution of *N*-glycan glycosidic torsion angles and could be used as an alternative data source for validation in *Privateer*. A future update of *Privateer* will allow users to analyse their structures against either the PDB or PDB-REDO torsional sets.

The application of consistent refinement techniques was also shown to improve outliers which had no physical basis for occurring (little clear interaction with residues or other ligands). Fig. 5 highlights the correction that *PDB-REDO* applies to the initially skewed MAN-1,6-BMA linkage. The data set of linkages originating from the PDB has numerous instances like this in which *PDB-REDO* corrects the torsion angles to more reasonable values. This powerful correction is another interesting and useful feature that *PDB-REDO* facilitates.

3.4. Outlier analysis

This analysis of *N*-glycan torsion angles deposited in the PDB reveals clusters of abundant torsion angles, as shown in Fig. 3. Perhaps due to the inherent variability in the environment surrounding monosaccharides in *N*-glycans, these torsion-angle clusters are spread over a large range in most cases. Outliers were quantified as any linkage which had a *Z*-score which was lower than −1. The *Z*-score reported here depends on the number of φ/ψ pairs relative to the database (Fig. 3) and not the deviation from the mean. The limit of −1 was chosen to highlight linkages that are uncommon in the database. Examining these linkages in further detail may highlight the cause of this. As always, surprising cases may either be chemically interesting to look at or be wrong. Here, we present one example of each.

3.4.1. Electrostatic interactions. Repulsive and attractive electrostatic interactions are crucial for the functionality and stability of proteins (Law *et al.*, 2006). These interactions are facilitated by both positively charged (lysine and arginine) and negatively charged (glutamic acid and aspartic acid) amino-acid side chains. Similarly, these amino acids can affect the

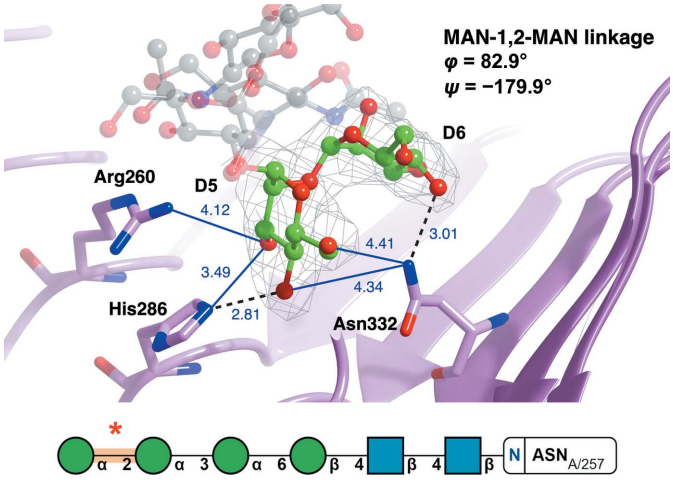


Figure 6
An unusual pair of MAN-1,2-MAN torsions in PDB entry 4j0m (She *et al.*, 2013). The mannose–mannose pair is well supported by the electron density, indicating that the unusual conformation of the linkage (red asterisk in the bottom panel) may be stabilized by interactions, electrostatic in this case, with surrounding side chains. The MAN (chain ID and sequence number D5)–MAN (chain ID and sequence number D6) linkage of PDB entry 4j0m (pink) is identified as an outlier ($\varphi = 82.9°$, $\psi = -179.9°$). The carbohydrate linkage is represented by a ball-and-stick model (C, green; O, red; N, blue). Residues identified as interacting with the linkage are represented by a cylindrical model (C, pink). Hydrogen bonds (black dashed line) and electrostatic interactions (within 4.5 Å, blue line) are shown with the distance between atoms in Å. $2F_o - F_c$ electron density (grey) is displayed for the linkage contoured to 1σ. Possible electrostatic interactions were identified for residues within 4.5 Å of the linkage and can be seen between Arg260 NH1 and MAN5 O3, His286 NE2 and MAN5 O3, Asn332 ND2 and MAN5 O4, and Asn332 ND2 and MAN5 O6. This linkage has a *Z*-score of −1.06. The top image was produced using *CCP4mg*. Bottom: SNFG notation output from *Privateer*.

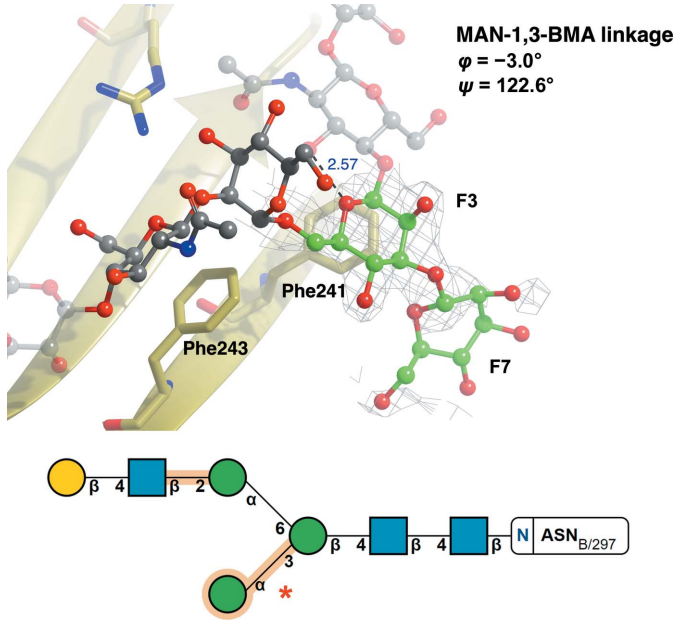


Figure 7
High-energy ring conformations may cause glycosidic link anomalies. The MAN(F7)–BMA(F3) linkage (red asterisk in the bottom panel) of PDB entry 5gsq (Chen *et al.*, 2017; gold), which was not part of the curated torsion-angle data set because the MAN residue has a poor RSCC, is identified as an outlier ($\varphi = -3.0°$, $\psi = 122.6°$). BMA (chain ID and sequence number F3) and MAN (chain ID and sequence number F7) are represented by a ball-and-stick model (C, green; O, red), whilst the rest of the attached glycan is shown in a faded grey ball-and-stick representation. Residues identified as interacting with the linkage are represented in stick form (C, gold; O, red; N, blue). Hydrogen bonds (black dashed lines) are shown with the distance between atoms in Å. $2F_o - F_c$ electron density (grey) is displayed for the linkage contoured to 1σ. Possible CH–π interactions were identified and can be seen between Phe243 and NAG(F5) and between Phe241 and BMA(F3). This linkage has a *Z*-score of −1.32, and presumably became distorted because the terminal mannose, MAN(F7), is in a $^1S_3$ skew-boat ring conformation (high energy; for further reading on conformational anomalies, please refer to Agirre, Davies *et al.*, 2015), as also highlighted in orange in the figure, due to the absence of well defined electron density. Both the linkage and ring conformations are unsupported by the electron density and should be either removed or corrected before deposition to reflect the most probable, low-energy conformations. The top image was produced using *CCP4mg*. Bottom: SNFG notation output from *Privateer*.

positions of monosaccharides contained within *N*-glycans via varying degrees of electrostatic interactions.

Fig. 6 depicts an *N*-glycan (PDB entry 4j0m; She *et al.*, 2013) with MAN-1,2-MAN torsion angles that are highly deviated from the mean. Since this glycan has been validated using *Privateer* (all monosaccharides, including those involved in the linkage, were in low-energy chair conformations) and has an RSCC of greater than 0.80, indicating a good fit to electron density, it can be assumed that these torsion angles are a direct result of external factors. Upon examination of the area surrounding the glycan, it becomes evident that a network of electrostatic interactions could be affecting the conformation of the *N*-glycan chain. The proximity of the linkage to arginine, histidine and asparagine side chains may cause the observed deviation. Furthermore, this highlights how linkages further down a glycan tree can also be subject to interactions with protein residues. These interactions may also explain why MAN-MAN linkage torsion angles are less concentrated on one pair of values than the more constrained NAG-NAG linkage.

3.4.2. High-energy ring-conformation anomalies may distort a linkage. Fig. 7 shows a glycan stabilized by CH–π interactions with phenylalanine side chains (PDB entry 5gsq; Chen *et al.*, 2017). While the fit to electron density is reasonable for the first few pyranosides (which show no issues in the validation report), the MAN-1,3-BMA and the terminal MAN residue are highlighted in orange in the *Privateer* SNFG representation: the link has a *Z*-score of −1.32, indicating a large deviation, and the ring of the terminal mannose is in a $^{1}S_{3}$ conformation, which is wholly unexpected for a pyranoside that is part of an *N*-glycan and therefore is marked as worthy of inspection (orange). Examination of the electron-density map around the MAN-1,3-BMA pair reveals that the fit to the observed data is poor for the MAN residue; refinement against incomplete density usually results in high-energy ring conformations without the inclusion of torsion restraints (Agirre, 2017). The distortion of the ring conformation in pyranosides has been reported to have a knock-on effect on linkages (Agirre *et al.*, 2017); hence, we believe this is the most probable explanation for this outlier.

4. Conclusions

In this study, a large number and range of *N*-glycan linkage torsion angles were collected from both the PDB and the PDB-REDO databank after being curated using *Privateer*. The collected data, released and articulated through the *Privateer* software, will provide a strong foundation for future model building, refinement and validation software. The comparisons between the PDB and PDB-REDO models presented here assessed the importance of modern refinement techniques. The differences in the torsion angles between the validated PDB and PDB-REDO data sets are minimal. However, in certain cases the application of a consistent refinement technique can alleviate errors in the model-building process. Furthermore, the absence of torsional restraints in PDB-REDO perhaps allows a more realistic spread of torsional values to be observed. It is also important to note valid rationalizations for linkage torsion angles deviating from the calculated mean. Electrostatic and steric interactions play a large role in protein folding in general and can cause or stabilize the skewed *N*-glycan linkage torsions exhibited in certain glycoproteins. Therefore, it is highly likely that these electrostatically charged or sterically bulky amino acids play a role in overall *N*-glycan conformation.

5. Availability and open research data

All scripts, data and graphics associated with this work have been uploaded to Zenodo (https://doi.org/10.5281/zenodo.7356467). The *Privateer* source code is available from GitHub (https://github.com/glycojones/privateer). Binaries will be released as an update to *CCP*4 8.0.

Acknowledgements

This work is based on a proof of concept as part of the MChem course of the Department of Chemistry at the University of York. The authors would like to thank Alex Ascham, Sarah Hargan, Eleanor Charnley, Alex Muriel, Charles Kingston, Conor MacDonald, Eve Tipple, Andrew Harvey, Thomas Hartshorn, Alex Bentley, Jordan Wilson, Rachel Napier, Luke Julyan, Catrin Ellis, Will Ashwood and James Jones for their work in the previous study.

Funding information

Jordan Dialpuri is funded by the Biotechnology and Biological Sciences Research Council (BBSRC; grant No. BB/T0072221). Haroldas Bagdonas is funded by The Royal Society (grant No. RGF/R1/181006). Mihaela Atanasova is funded by the UK Engineering and Physical Sciences Research Council (EPSRC; grant No. EP/R513386/1). Lucy Schofield's work was funded by The Royal Society through a summer studentship. Robbie Joosten is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871037 (iNEXT-Discovery) and by CCP4. Jon Agirre is a Royal Society University Research Fellow (award No. UF160039).

References

Agirre, J. (2017). *Acta Cryst.* D**73**, 171–186.
Agirre, J., Ariza, A., Offen, W. A., Turkenburg, J. P., Roberts, S. M., McNicholas, S., Harris, P. V., McBrayer, B., Dohnalek, J., Cowtan, K. D., Davies, G. J. & Wilson, K. S. (2016). *Acta Cryst.* D**72**, 254–265.
Agirre, J., Davies, G., Wilson, K. & Cowtan, K. (2015). *Nat. Chem. Biol.* **11**, 303.
Agirre, J., Davies, G. J., Wilson, K. S. & Cowtan, K. D. (2017). *Curr. Opin. Struct. Biol.* **44**, 39–47.
Agirre, J., Iglesias-Fernández, J., Rovira, C., Davies, G. J., Wilson, K. S. & Cowtan, K. D. (2015). *Nat. Struct. Mol. Biol.* **22**, 833–834.
Akkermans, O., Delloye-Bourgeois, C., Peregrina, C., Carrasquero-Ordaz, M., Kokolaki, M., Berbeira-Santana, M., Chavent, M., Reynaud, F., Raj, R., Agirre, J., Aksu, M., White, E. S., Lowe, E., Ben Amar, D., Zaballa, S., Huo, J., Pakos, I., McCubbin, P. T. N.,

Comoletti, D., Owens, R. J., Robinson, C. V., Castellani, V., del Toro, D. & Seiradake, E. (2022). *Cell*, **185**, 3931–3949.
Alocci, D., Mariethoz, J., Gastaldello, A., Gasteiger, E., Karlsson, N. G., Kolarich, D., Packer, N. H. & Lisacek, F. (2019). *J. Proteome Res.* **18**, 664–677.
Atanasova, M., Bagdonas, H. & Agirre, J. (2020). *Curr. Opin. Struct. Biol.* **62**, 70–78.
Atanasova, M., Nicholls, R. A., Joosten, R. P. & Agirre, J. (2022). *Acta Cryst.* D**78**, 455–465.
Bagdonas, H., Ungar, D. & Agirre, J. (2020). *Beilstein J. Org. Chem.* **16**, 2523–2533.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Beusekom, B. van, Joosten, K., Hekkelman, M. L., Joosten, R. P. & Perrakis, A. (2018). *IUCrJ*, **5**, 585–594.
Beusekom, B. van, Touw, W. G., Tatineni, M., Somani, S., Rajagopal, G., Luo, J., Gilliland, G. L., Perrakis, A. & Joosten, R. P. (2018). *Protein Sci.* **27**, 798–808.
Beusekom, B. van, Wezel, N., Hekkelman, M. L., Perrakis, A., Emsley, P. & Joosten, R. P. (2019). *Acta Cryst.* D**75**, 416–425.
Casalino, L., Gaieb, Z., Goldsmith, J. A., Hjorth, C. K., Dommer, A. C., Harbison, A. M., Fogarty, C. A., Barros, E. P., Taylor, B. C., McLellan, J. S., Fadda, E. & Amaro, R. E. (2020). *ACS Cent. Sci.* **6**, 1722–1734.
Chen, C.-L., Hsu, J.-C., Lin, C.-W., Wang, C.-H., Tsai, M.-H., Wu, C.-Y., Wong, C.-H. & Ma, C. (2017). *ACS Chem. Biol.* **12**, 1335–1345.
Clauset, A., Shalizi, C. R. & Newman, M. E. J. (2009). *SIAM Rev.* **51**, 661–703.
Cremer, D. & Pople, J. A. (1975). *J. Am. Chem. Soc.* **97**, 1354–1358.
Crispin, M., Stuart, D. I. & Jones, E. Y. (2007). *Nat. Struct. Mol. Biol.* **14**, 354.
Crispin, M., Yu, X. & Bowden, T. A. (2013). *Proc. Natl Acad. Sci. USA*, **110**, E3544–E3546.
Davies, G. J., Planas, A. & Rovira, C. (2012). *Acc. Chem. Res.* **45**, 308–316.
Emsley, P. & Crispin, M. (2018). *Acta Cryst.* D**74**, 256–263.
Ferrara, C., Grau, S., Jäger, C., Sondermann, P., Brünker, P., Waldhauer, I., Hennig, M., Ruf, A., Rufer, A. C., Stihle, M., Umaña, P. & Benz, J. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 12669–12674.
Frank, M., Beccati, D., Leeflang, B. R. & Vliegenthart, J. F. G. (2020). *iScience*, **23**, 101371.
Frank, M., Lütteke, T. & von der Lieth, C.-W. (2007). *Nucleic Acids Res.* **35**, 287–290.
Hooft, R. W., Sander, C. & Vriend, G. (1997). *Comput. Appl. Biosci.* **13**, 425–430.
Imberty, A. & Perez, S. (1995). *Protein Eng. Des. Sel.* **8**, 699–709.
Joosten, R. P., Nicholls, R. A. & Agirre, J. (2022). *Curr. Med. Chem.* **29**, 1193–1207.
Law, M. J., Linde, M. E., Chambers, E. J., Oubridge, C., Katsamba, P. S., Nilsson, L., Haworth, I. S. & Laird-Offringa, I. A. (2006). *Nucleic Acids Res.* **34**, 275–285.
Lütteke, T., Frank, M. & von der Lieth, C.-W. (2005). *Nucleic Acids Res.* **33**, D242–D246.
Lütteke, T. & von der Lieth, C.-W. (2004). *BMC Bioinformatics*, **5**, 69.
McNicholas, S. & Agirre, J. (2017). *Acta Cryst.* D**73**, 187–194.
Nnamchi, C. I., Parkin, G., Efimov, I., Basran, J., Kwon, H., Svistunenko, D. A., Agirre, J., Okolo, B. N., Moneke, A., Nwanguma, B. C., Moody, P. C. E. & Raven, E. L. (2016). *J. Biol. Inorg. Chem.* **21**, 63–70.
Pérez, S., Sarkar, A., Rivet, A., Breton, C. & Imberty, A. (2015). *Methods Mol. Biol.* **1273**, 241–258.
Petrescu, A.-J., Wormald, M. R. & Dwek, R. A. (2006). *Curr. Opin. Struct. Biol.* **16**, 600–607.
Ramirez-Escudero, M., Miguez, N., Gimeno-Perez, M., Ballesteros, A. O., Fernandez-Lobato, M., Plou, F. J. & Sanz-Aparicio, J. (2019). *Sci. Rep.* **9**, 17441.
Rudd, P. M., Wormald, M. R. & Dwek, R. A. (2004). *Trends Biotechnol.* **22**, 524–530.
Scherbinina, S. I. & Toukach, P. V. (2020). *Int. J. Mol. Sci.* **21**, 7702.
She, J., Han, Z., Zhou, B. & Chai, J. (2013). *Protein Cell*, **4**, 475–482.
Shental-Bechor, D. & Levy, Y. (2009). *Curr. Opin. Struct. Biol.* **19**, 524–533.
Sobolev, O. V., Afonine, P. V., Moriarty, N. W., Hekkelman, M. L., Joosten, R. P., Perrakis, A. & Adams, P. D. (2020). *Structure*, **28**, 1249–1258.e2.
Varki, A., Cummings, R. D., Aebi, M., Packer, N. H., Seeberger, P. H., Esko, J. D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., Prestegard, J. J., Schnaar, R. L., Freeze, H. H., Marth, J. D., Bertozzi, C. R., Etzler, M. E., Frank, M., Vliegenthart, J. F., Lütteke, T., Perez, S., Bolton, E., Rudd, P., Paulson, J., Kanehisa, M., Toukach, P., Aoki-Kinoshita, K. F., Dell, A., Narimatsu, H., York, W., Taniguchi, N. & Kornfeld, S. (2015). *Glycobiology*, **25**, 1323–1324.
Xin, F. & Radivojac, P. (2012). *Bioinformatics*, **28**, 2905–2913.

STRUCTURAL BIOLOGY

ISSN 2059-7983

Received 29 March 2023
Accepted 19 April 2023

Edited by S. Antonyuk, Institute of Integrative Biology, University of Liverpool, United Kingdom

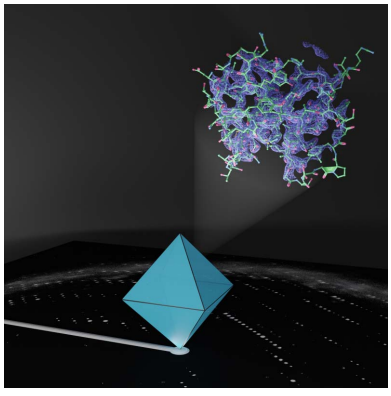
‡ Alexei Vagin passed away on 25 March 2023.

Keywords: Collaborative Computational Project No. 4; *CCP*4; crystallography software; macromolecular crystallography

The *CCP*4 suite: integrative software for macromolecular crystallography

Jon Agirre,[a]* Mihaela Atanasova,[a] Haroldas Bagdonas,[a] Charles B. Ballard,[b,c] Arnaud Baslé,[d] James Beilsten-Edmands,[e] Rafael J. Borges,[f] David G. Brown,[g] J. Javier Burgos-Mármol,[h] John M. Berrisford,[i] Paul S. Bond,[a] Iracema Caballero,[j] Lucrezia Catapano,[k,l] Grzegorz Chojnowski,[m] Atlanta G. Cook,[n] Kevin D. Cowtan,[a] Tristan I. Croll,[o,p] Judit É. Debreczeni,[q] Nicholas E. Devenish,[e] Eleanor J. Dodson,[a] Tarik R. Drevon,[b,c] Paul Emsley,[k] Gwyndaf Evans,[e,r] Phil R. Evans,[k] Maria Fando,[b,c] James Foadi,[s] Luis Fuentes-Montero,[e] Elspeth F. Garman,[t] Markus Gerstel,[e] Richard J. Gildea,[e] Kaushik Hatti,[o] Maarten L. Hekkelman,[u] Philipp Heuser,[v] Soon Wen Hoh,[a] Michael A. Hough,[e,w] Huw T. Jenkins,[a] Elisabet Jiménez,[j] Robbie P. Joosten,[u] Ronan M. Keegan,[b,c,h] Nicholas Keep,[x] Eugene B. Krissinel,[b,c] Petr Kolenko,[y,z] Oleg Kovalevskiy,[b,c] Victor S. Lamzin,[m] David M. Lawson,[aa] Andrey A. Lebedev,[b,c] Andrew G. W. Leslie,[k] Bernhard Lohkamp,[bb] Fei Long,[k] Martin Malý,[y,z,cc] Airlie J. McCoy,[o] Stuart J. McNicholas,[a] Ana Medina,[j] Claudia Millán,[o] James W. Murray,[dd] Garib N. Murshudov,[k] Robert A. Nicholls,[k] Martin E. M. Noble,[ee] Robert Oeffner,[o] Navraj S. Pannu,[ff] James M. Parkhurst,[e,r] Nicholas Pearce,[gg] Joana Pereira,[hh] Anastassis Perrakis,[u] Harold R. Powell,[dd] Randy J. Read,[o] Daniel J. Rigden,[h] William Rochira,[a] Massimo Sammito,[o,ii] Filomeno Sánchez Rodríguez,[a,e,h] George M. Sheldrick,[jj] Kathryn L. Shelley,[kk] Felix Simkovic,[h] Adam J. Simpkin,[g] Pavol Skubak,[ff] Egor Sobolev,[v] Roberto A. Steiner,[l,ll] Kyle Stevenson,[b] Ivo Tews,[cc] Jens M. H. Thomas,[h] Andrea Thorn,[mm] Josep Triviño Valls,[j] Ville Uski,[b,c] Isabel Usón,[j,nn] Alexei Vagin,[a]‡ Sameer Velankar,[i] Melanie Vollmar,[i] Helen Walden,[oo] David Waterman,[b,c] Keith S. Wilson,[a] Martyn D. Winn,[pp] Graeme Winter,[e] Marcin Wojdyr[qq] and Keitaro Yamashita[k]

[a]York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, United Kingdom, [b]STFC, Rutherford Appleton Laboratory, Didcot OX11 0FA, United Kingdom, [c]CCP4, Research Complex at Harwell, Rutherford Appleton Laboratory, Didcot OX11 0FA, United Kingdom, [d]Biosciences Institute, Newcastle University, Newcastle upon Tyne NE2 4HH, United Kingdom, [e]Diamond Light Source, Harwell Science and Innovation Campus, Didcot OX11 0DE, United Kingdom, [f]The Center of Medicinal Chemistry (CQMED), Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Av. Dr. André Tosello 550, 13083-886 Campinas, Brazil, [g]Laboratoires Servier SAS Institut de Recherches, Croissy-sur-Seine, France, [h]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom, [i]Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL–EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom, [j]Crystallographic Methods, Institute of Molecular Biology of Barcelona (IBMB–CSIC), Barcelona Science Park, Helix Building, Baldiri Reixac 15, 08028 Barcelona, Spain, [k]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, United Kingdom, [l]Randall Centre for Cell and Molecular Biophysics, Faculty of Life Sciences and Medicine, King's College London, London SE1 9RT, United Kingdom, [m]European Molecular Biology Laboratory, Hamburg Unit, Notkestrasse 85, 22607 Hamburg, Germany, [n]The Wellcome Centre for Cell Biology, University of Edinburgh, Michael Swann Building, Max Born Crescent, The King's Buildings, Edinburgh EH9 3BF, United Kingdom, [o]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge CB2 0XY, United Kingdom, [p]Altos Labs, Portway Building, Granta Park, Great Abington, Cambridge CB21 6GP, United Kingdom, [q]Discovery Sciences, R&D BioPharmaceuticals, AstraZeneca, Darwin Building, Cambridge Science Park, Milton Road, Cambridge CB4 0WG, United Kingdom, [r]Rosalind Franklin Institute, Harwell Science and Innovation Campus, Didcot OX11 0QS, United Kingdom, [s]Department of Mathematical Sciences, University of Bath, Bath, United Kingdom, [t]Department of Biochemistry, University of Oxford, Dorothy Crowfoot Hodgkin Building, Oxford OX1 3QU, United Kingdom, [u]Oncode Institute and Department of Biochemistry, Netherlands Cancer Institute, Amsterdam, The Netherlands, [v]European Molecular Biology Laboratory, c/o DESY, Notkestrasse 85, 22607 Hamburg, Germany, [w]School of Life Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom, [x]Department of Biological Sciences, Institute of Structural and Molecular Biology, Birkbeck College, London WC1E 7HX, United Kingdom, [y]Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Břehová 7, 115 19 Prague 1, Czech Republic, [z]Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, Průmyslová 55, 252 50 Vestec, Czech Republic, [aa]Department of Biochemistry and Metabolism, John Innes Centre, Norwich NR4 7UH, United Kingdom, [bb]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-171 77 Stockholm, Sweden, [cc]Biological Sciences, Institute for Life Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom, [dd]Department of Life Sciences, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom, [ee]Translational and Clinical Research Institute, Newcastle University,



Paul O'Gorman Building, Medical School, Framlington Place, Newcastle upon Tyne NE2 4HH, United Kingdom, [ff]Department of Infectious Diseases, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands, [gg]Department of Physics, Chemistry and Biology (IFM), Linköping University, SE-581 83 Linköping, Sweden, [hh]Biozentrum and SIB Swiss Institute of Bioinformatics, University of Basel, 4056 Basel, Switzerland, [ii]Discovery Centre, Biologics Engineering, AstraZeneca, Biomedical Campus, 1 Francis Crick Avenue, Trumpington, Cambridge CB2 0AA, United Kingdom, [jj]Department of Structural Chemistry, Georg-August-Universität Göttingen, Tammannstrasse 4, 37077 Göttingen, Germany, [kk]Institute for Protein Design, University of Washington, Seattle, WA 98195, USA, [ll]Department of Biomedical Sciences, University of Padova, Italy, [mm]Institute for Nanostructure and Solid State Physics, Universität Hamburg, 22761 Hamburg, Germany, [nn]ICREA, Institució Catalana de Recerca i Estudis Avançats, Passeig Lluís Companys 23, 08003 Barcelona, Spain, [oo]School of Molecular Biosciences, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom, [pp]Scientific Computing Department, Science and Technology Facilities Council, Didcot OX11 0FA, United Kingdom, and [qq]Global Phasing Limited (United Kingdom), Sheraton House, Castle Park, Cambridge CB3 0AX, United Kingdom. *Correspondence e-mail: jon.agirre@york.ac.uk

The Collaborative Computational Project No. 4 (CCP4) is a UK-led international collective with a mission to develop, test, distribute and promote software for macromolecular crystallography. The *CCP*4 suite is a multiplatform collection of programs brought together by familiar execution routines, a set of common libraries and graphical interfaces. The *CCP*4 suite has experienced several considerable changes since its last reference article, involving new infrastructure, original programs and graphical interfaces. This article, which is intended as a general literature citation for the use of the *CCP*4 software suite in structure determination, will guide the reader through such transformations, offering a general overview of the new features and outlining future developments. As such, it aims to highlight the individual programs that comprise the suite and to provide the latest references to them for perusal by crystallographers around the world.

1. Introduction

As a technique, macromolecular crystallography (MX) relies heavily on computational methods, built on top of a strict set of conventions and common formats. Most conventions follow the lead of the International Union of Crystallography (IUCr), while MX software development is undertaken by both academic and private sector initiatives, such as the Phenix Consortium (Liebschner *et al.*, 2019) and Global Phasing Ltd (Cambridge, United Kingdom). Based in the UK, MX software tools find a common distribution and maintenance channel under the umbrella of the Collaborative Computational Project No. 4, best known as CCP4. This consortium was established by the UK Science Research Council in 1979, almost 45 years ago, to facilitate the coordination and collaboration of MX software developers (Agirre & Dodson, 2018). Aside from coordinating and distributing software, CCP4 has a mission of promoting the teaching of MX, with an annual didactic CCP4 Study Weekend and numerous online and in-person annual workshops around the world. Forums, which originally took the shape of email lists – the CCP4 bulletin board (or CCP4bb) for general users' questions and ccp4-dev for developer discussions – are an evolving aspect of the CCP4 community, with social media taking a more prominent role in hosting other kinds of exchanges, for example paper or event announcements (Twitter: @ccp4_mx) or parallel discussions at conferences (Slack channels). The CCP4 website (https://www.ccp4.ac.uk) is the primary

mechanism for reference and asynchronous communication but, most importantly, provides a central distribution point for software downloads. A minimal installer package can be obtained from the site, and this will proceed to install the latest version of the suite. Updates are then distributed via a non-disruptive mechanism that was first introduced with *CCP*4 version 6.3.0 in 2012. Update reminders are generated automatically, although the update mechanism itself is, by design, initiated manually. As an indication of update frequency, the 7.0 series, which was originally released in 2016, saw more than 70 updates until the 7.1 series was released in 2020. Updates are not a one-way road: they may be rolled back if problems are encountered. Whilst every effort has been made to keep the suite streamlined and maintainable, the inclusion of large databases and toolkits has driven space requirements steadily upwards (Fig. 1).

The last decade has seen some large transformations in the field of MX: new workflows have been created (for example phasing with *AlphaFold*2 models) and some old workflows have been optimized, while some others are on the verge of disappearing; this has often been the result of cross-pollination with other techniques in structural biology, for example electron cryo-microscopy (cryo-EM) in particular, through a synergistic collaboration with CCP-EM (Burnley *et al.*, 2017), the Collaborative Computational Project for Cryo-EM, which repurposes some *CCP*4 code for the cryo-EM community. For example, owing to the deep-learning revolution in computational structure prediction (Jumper *et al.*, 2021), it is now possible to phase most structures using large predicted fragments or, owing to the accuracy of the method, even to rigid-body fit an initial predicted model into electron density (Oeffner *et al.*, 2022; McCoy *et al.*, 2022; Medina *et al.*, 2022). As a side effect of the creation of these new workflows, experimental phasing is now losing importance in the everyday activities of an MX laboratory, with derivatives only being created as a last resort after all of the now conventional methods have failed. Data acquisition and processing, greatly bolstered by both software and hardware developments *in situ* at synchrotrons, is now performed almost instantaneously after data collection, presenting the user with the results of applying different processing strategies. Although seemingly unconnected, most of these newer developments have one thing in common: the Python programming language as a platform for pipelining and program communication.

While some Python scripts were already part of the *CCP*4 suite even before the time of the last general publication (Winn *et al.*, 2011), most of the recent source code committed to the CCP4 repositories involves Python in one way or another; for example, both the data-integration tool *DIALS* (Winter *et al.*, 2018) and its *CCP*4 graphical user interface *DUI* (Fuentes-Montero *et al.*, 2016) are Python-heavy software. Other *CCP*4 programs, encoded in a different language such as C++ for performance reasons, may also offer Python bindings; examples include *Coot* (Emsley *et al.*, 2010), *Privateer* (Agirre *et al.*, 2015) and *GEMMI* (Wojdyr, 2022), which is a crystallographic toolkit developed in collaboration with Global Phasing Ltd. Both the Python language and its interpreter are now at the core of the *CCP*4 suite. Importantly, both new graphical user interfaces to the *CCP*4 suite (see below) make substantial use of the Python language.

On the subject of graphical user interfaces, a large paradigm shift is also under way, with both *CCP*4*i*2 and *CCP*4 Cloud making extensive use of web technologies: HTML, CSS and JavaScript are used for both interface design and result presentation, with *CCP*4 Cloud making a strong case for the transformation of existing interactive model-building and illustration applications, for example *Coot* and *CCP*4*mg*, into apps that can be run within a web browser.

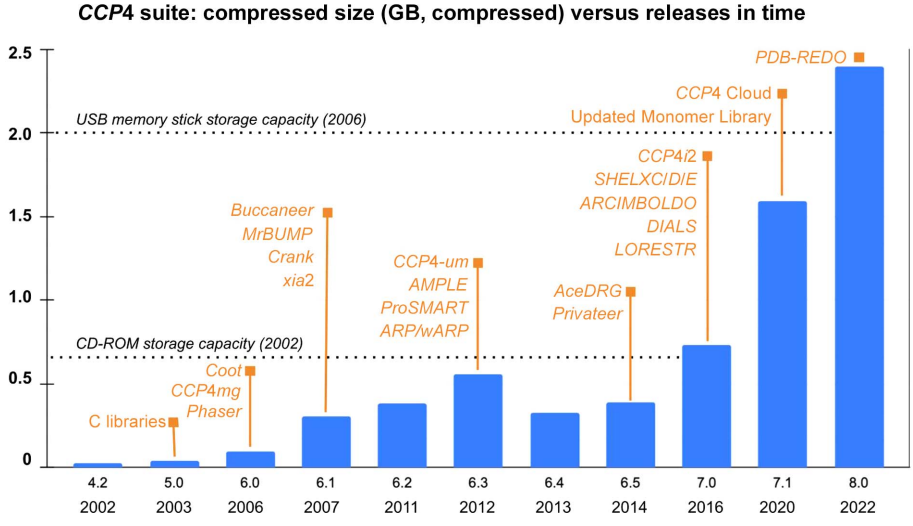


Figure 1
Evolution in the size of the *CCP*4 suite from version 4.2 (2002) through to version 8.0 (2022). Some representative programs included in the releases are highlighted in orange. The update mechanism (*CCP*4-*um*) was first used in version 6.3. New graphical interfaces were introduced in versions 7.0 (*CCP*4*i*2) and 7.1 (*CCP*4 Cloud). *Coot* and *CCP*4*mg* were originally distributed separately, but were bundled with the suite from version 6.5. For reference, the sizes of two popular contemporary storage devices are shown as dotted lines; please note that these were never targeted as distribution media.

2. Overview of the newest developments

2.1. Graphical user interfaces

The long-serving *CCP*4*i* interface (developed in Tcl/Tk) has recently been deprecated and replaced by a more modern, QT/PySide graphical user interface (GUI) named *CCP*4*i*2 (Potterton *et al.*, 2018). The *CCP*4*i*2 GUI, the main purpose of which is to provide a desktop-based experience, has introduced a number of architectural differences with respect to the first iteration. (i) A real database system, as opposed to a directory structure, provides traceability of files and jobs, and allows the automatic population of inputs to follow-on jobs with outputs from previous jobs. (ii) Large MTZ files are separated into important column sets defining particular data types and with predictable names, for example Miller indices (H, K and L columns) plus amplitudes and estimated standard deviations or e.s.d.s (F and SIGF columns) define an 'Amplitudes' data type. (iii) Individual programs are wrapped in Python for their incorporation into tasks, which in many cases will be pipelines themselves; for example 'Data reduction' is a pipeline that involves use of the programs *POINTLESS*, *AIMLESS*, *CTRUNCATE* and *FREER*. (iv) Communication of results between individual programs is consolidated in structured data (XML) files. In addition, task reports aim to present only fundamental results and, where possible, provide expert diagnostics in a natural human-readable language, for example 'No evidence of possible translational noncrystallographic symmetry'. Other utilities include a multiplatform project import and export mechanism, instant job search by keywords, the use of task-specific key performance indicators, for example $R_{\rm work}/R_{\rm free}$, and context-dependent follow-on jobs with automatic selection of input files and default options. Outside the graphical user interface but very much within its infrastructure, the *i*2*run* module provides a command-line mechanism for running *CCP*4*i*2 pipelines, opening the door to batch processing using interface-level decision making.

*CCP*4 Cloud (Krissinel *et al.*, 2022) is a complete reimagination of what an interface should look like in the context of macromolecular crystallography. Technology-wise, it provides a server-side JavaScript implementation (based on Node.js) designed to work with high-performance computing (HPC) facilities (clusters and generic clouds) but which can also be run on a user's PC. This implementation also enables secure web access by a browser via HTML5, CSS and JavaScript (jQuery), and allows *CCP*4 Cloud to look consistent across different browsers and platforms, making it possible to run jobs and manage projects from, for example, mobile devices. The interface provides a general file-import function, which allows it to decide what kind of jobs can be run: for example, automated model building can only be performed if at least reflections and a sequence have been imported. The system features task interfaces for many *CCP*4 programs and some newly introduced pipelines. One such example is *CCP*4*build*, which combines *Parrot* for density modification (Cowtan, 2010), *Buccaneer* for model building (Cowtan, 2006), *REFMAC* for refinement (Murshudov *et al.*, 2011), *Coot* for model editing (Emsley *et al.*, 2010) and *EDSTATS* (Tickle, 2012) for model accuracy analysis; using these tools, *CCP*4*build* is able to make expert decisions depending on the phasing approach and model completeness. High-level progress indicators are available in both *CCP*4 Cloud and *CCP*4*i*2; one such example is the 'verdict' functionality, which provides a score for model completion and fit to the experimental data. *CCP*4*i*2 and *CCP*4 Cloud have a conceptually similar set of tasks, although their graphical presentation differs (Fig. 2).

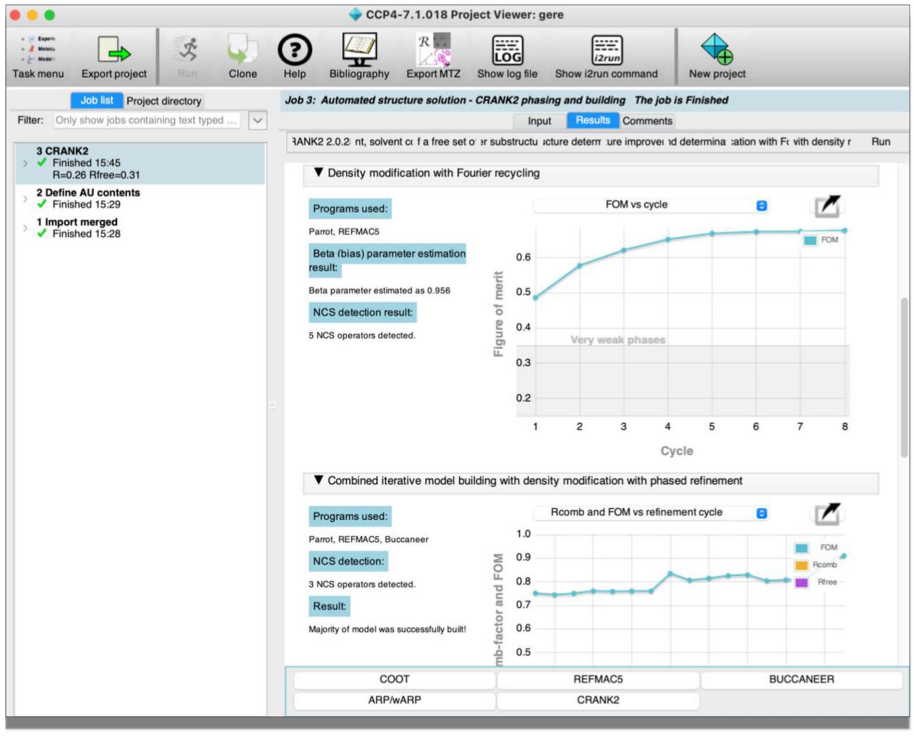
2.2. Data processing

Developed in collaboration with Diamond Light Source and the Lawrence Berkeley National Laboratory, the *DIALS* project (Winter *et al.*, 2018) is the *CCP*4 suite's main diffraction image processing toolkit; it is modular and hackable by design, so experienced crystallographers can tweak, extend or add new algorithms. Regardless of this specialist component-based approach, complete *DIALS* workflows are provided in the *xia*2 pipeline (Winter, 2010), which incorporates expert decision making (Winter *et al.*, 2013). More recently, a graphical user interface (*DIALS User Interface* or *DUI*) has also been introduced (Fuentes-Montero *et al.*, 2016). The *xia*2 pipeline is run automatically at the end of data collections at Diamond Light Source (Oxfordshire, United Kingdom), providing the results of applying multiple data-processing strategies: users are expected to look at the metrics provided and decide which is better suited to their diffraction data set. Newcomer users wanting to learn more about *DIALS* are advised to use *DUI*, which provides a guided step-by-step execution of the whole process, although command-line use through simple scripts is designed to be accessible to the non-expert user.

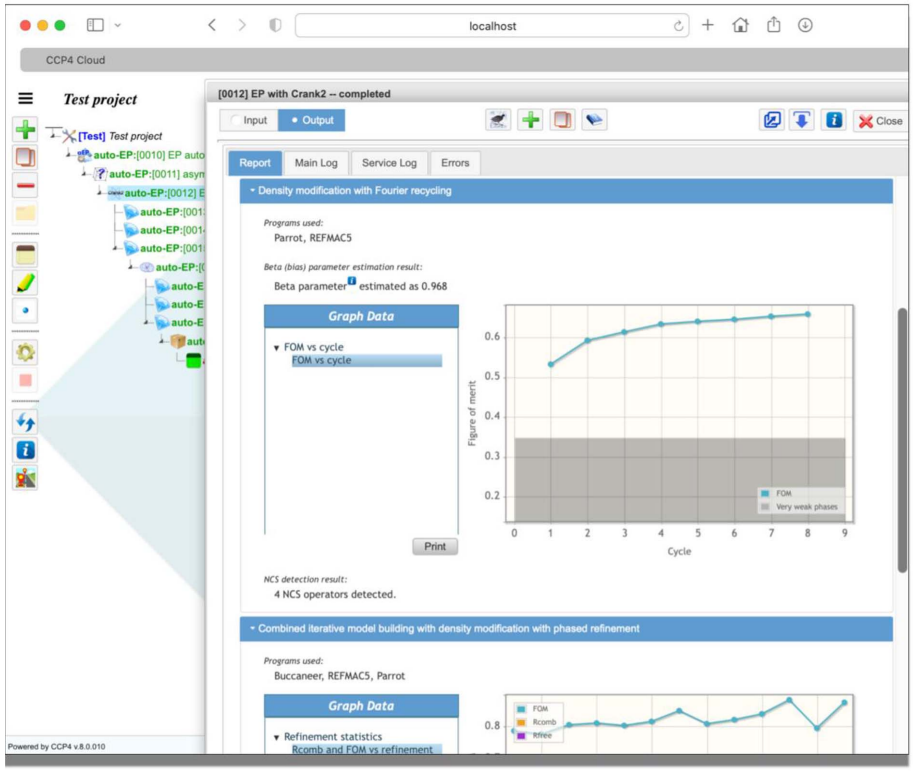
DIALS is able to natively process data obtained at X-ray free-electron laser (XFEL) facilities (Ginn *et al.*, 2015; Uervirojnangkoorn *et al.*, 2015) and supports multi-crystal scaling (Beilsten-Edmands *et al.*, 2020) and analysis via *xia*2.*multiplex* (Gildea *et al.*, 2022), serial crystallography (Brewster *et al.*, 2018; Parkhurst, 2020) and electron diffraction such as that obtained with standard field emission gun (FEG) cryo-microscopes (Clabbers *et al.*, 2018). Data from multiple crystals may be scaled and merged together with *BLEND* (Mylona *et al.*, 2017). Ice rings and further pathologies in measured data can be identified by a separate standalone tool named *AUSPEX*, which provides visual and automatic diagnostics based on statistics (Thorn *et al.*, 2017) and, more recently, machine learning (Nolte *et al.*, 2022). Alternatively, the *iMosflm* software (Powell *et al.*, 2017) provides an easy-to-use interface to the *MOSFLM* image-processing program; while the software is no longer under active development, it contains many useful features and remains popular with users.

Once the data have been processed, Laue group determination and data scaling and reduction can be performed directly with *DIALS*, although *POINTLESS* and *AIMLESS* are also offered as a fallback mechanism (Evans & Murshudov, 2013); indeed, the latter two programs form the basis of the *CCP*4*i*2 'data reduction' task. Further diagnostics

can be obtained by running *CTRUNCATE*, which was originally an implementation of French and Wilson's algorithm (French & Wilson, 1978), to obtain structure-factor amplitudes from intensities; it will scan data sets for signs of anisotropic diffraction, twinning and translational non-crystallographic symmetry (tNCS) among other critical issues



(*a*)



(*b*)

Figure 2
Comparison of the new *CCP*4 graphical user interface offerings: (*a*) desktop (*CCP*4*i*2) and (*b*) online (*CCP*4 Cloud). The same pipeline (*Crank*-2) has been run on both interfaces. The reports show equivalent graphs due to the use of a compatibility layer that allows the same report code to run on both platforms.

that could complicate or even compromise the downstream structure-determination process. This set of programs has graphical interfaces in both *CCP*4*i*2 and *CCP*4 Cloud, producing colour-coded reports that flag up potential problems. Importantly, detailed reports are generated whenever merged intensities or amplitudes are imported into the graphical interfaces, providing a sanity check and metadata tracking.

2.3. Phasing

The *CCP*4 suite provides software for all phasing methods, although they mainly fall within one of the following categories: molecular replacement (MR), *ab initio* phasing with ideal fragments (a special case of molecular replacement) and experimental phasing. In the coming years, and due to the recent improvement in protein structure-prediction methods, the line between the former two is expected to become blurred or even disappear.

2.3.1. Molecular replacement and *ab initio* phasing, including bioinformatics. While the ever-growing area of bioinformatics is outside the remit of *CCP*4, the search for suitable molecular-replacement templates is primarily driven by protein homology analysis and therefore exploits bioinformatics methods. Various third-party tools have been incorporated into the suite to give support to the *CCP*4 model-preparation tools and automated structure-solution pipelines. *MrBUMP* is an automated tool that will perform searches for templates and attempt molecular replacement with them, displaying comprehensive results that can be taken forward provided that the *R* factors are low enough. It can find structures of homologues using *PHMMER* (Eddy, 2011) or *HHpred* (Söding, 2005) and place them using either *Phaser* (McCoy *et al.*, 2007) or *MOLREP* (Vagin & Teplyakov, 2010). The template search code of *MrBUMP* can also be harnessed interactively in *CCP*4*mg*, allowing users to create composite models and ensembles for subsequent MR searches; this tool can be accessed from both *CCP*4*i*2 and *CCP*4 Cloud. *MrParse* (Simpkin, Thomas *et al.*, 2022) provides a convenient visualization of potential search models from the PDB and databases of new generation models such as the AlphaFold Protein Structure Database (Varadi *et al.*, 2022). Designed to slice predicted models as well as homologs into domains that may differ in relative orientation from the crystal structure, *Slice'N'Dice* (Simpkin, Elliott *et al.*, 2022) is an automated molecular-replacement pipeline that facilitates the placement of these domains in molecular replacement. By processing and slicing the models, it simplifies the task of placing these domains. *CCP*4*mg* (McNicholas *et al.*, 2011) can also be used to visualize the slicing of the input models.

*CCP*4 has a number of efficient molecular-replacement packages: *AMoRe* (Trapani & Navaza, 2008), *MOLREP* (Vagin & Teplyakov, 2010) and *Phaser* (McCoy *et al.*, 2007) all have different strengths, although only the latter is under active development.

Phaser uses a maximum-likelihood approach to the phasing problem; it is the only molecular-replacement software that uses intensities natively, *i.e.* without turning them into amplitudes first, and can also use SAD data (for SAD and MR-SAD phasing). The *voyager* (Sammito *et al.*, 2019) automated procedure within *Phaser* presents a new architecture that allows more flexibility, guiding user decisions in creating ensembles. It also provides, alongside a plethora of new and reimplemented algorithms, code to make the best use of *AlphaFold* (Jumper *et al.*, 2021) and *RoseTTAFold* (Baek *et al.*, 2021) structure predictions, or high-confidence subsets of them, including the transformation of model confidence metrics (for example the *AlphaFold* pLDDT) into estimated *B* factors. Owing to the flexibility of the new design, tools for fitting models into cryo-EM maps have been included. An *ad hoc* graphical user interface is under development; this will allow easier navigation of the different solutions calculated during the search strategy, presenting the user with essential plots such as the self-rotation function.

*CCP*4 also has fragment-based *ab initio* phasing packages: *ARCIMBOLDO* (Rodríguez *et al.*, 2009) and *Fragon* (Jenkins, 2018), which use ideal fragments of proteins (mainly helices) in targeted molecular-replacement searches. The use of these programs was initially confined to high-resolution data, but they have recently enjoyed success at resolutions lower than 2.3 Å, a threshold beyond which it becomes difficult to ascertain the direction of helical fragments, owing to their improved search strategies (Medina *et al.*, 2022), phase combination (Millán *et al.*, 2020) and the use of available structural information, including *AlphaFold* predictions. *ARCIMBOLDO* (Rodríguez *et al.*, 2009) can use fragments of homologous models and phase previously intractable coiled-coil structures (Caballero *et al.*, 2018). It should be noted that part of the success of these methods is down to the ability of *Phaser* to place single amino acids or even atoms with great accuracy (McCoy *et al.*, 2017) and the ability of the density-modification and autotracing algorithms in *SHELXE* (Usón & Sheldrick, 2018) to bootstrap solutions from poor starting phase sets with average errors as high as 70° (Millán *et al.*, 2015). Also in alternative MR territory is *AMPLE* (Bibby *et al.*, 2012), which majors on editing search-model ensembles, particularly *ab initio* predictions and distant homologues.

SIMBAD (Simpkin *et al.*, 2018, 2020) provides a sequence-independent phasing pipeline that may be used for phasing crystals of unknown contaminants (Simpkin *et al.*, 2018). Other MR pipelines use larger fragments or domains as their source of phasing information: *BALBES* (Long *et al.*, 2008) and *MoRDA* (Vagin & Lebedev, 2015) are automated pipelines that use *MOLREP* to place matches from curated databases containing fragments, domains and homo- and hetero-oligomers. *Dimple* (Wojdyr *et al.*, 2013) is an automated procedure that aims to quickly arrive at a solved structure of a protein–ligand complex starting from an isomorphous crystal; the software will phase the data and produce preliminary maps, including a difference density map where omit density for a ligand might be found.

2.3.2. Experimental phasing. The steady increase in unique new domains deposited every year in the PDB, the availability of millions of predicted models in the AlphaFold Protein

Structure Database (Varadi *et al.*, 2022) and the continuous improvement of fragment-based *ab initio* phasing methods mean that experimental phasing is increasingly becoming a last-resort approach to recovering phases; it also means that software will have to deal with the most difficult cases. New since the time of the last *CCP*4 general publication (Winn *et al.*, 2011) is the inclusion of the *SHELXC*/*D*/*E* (Sheldrick, 2008) programs, which can be run individually or in a pipeline through the *Crank*-2 (Skubák & Pannu, 2013) frontend, which is available in both the *CCP*4*i*2 and *CCP*4 Cloud interfaces. *Crank*-2 itself incorporates a number of different algorithms that can deal with SAD, SIRAS, MAD and MR-SAD. As stated in the previous section, the *Phaser* software (McCoy *et al.*, 2007) is also able to perform both SAD and MR-SAD phasing.

2.4. Model building and refinement

2.4.1. Interactive model building. The *CCP*4 suite ships with the *de facto* industry-standard interactive model-building program *Coot* (Emsley *et al.*, 2010). After two decades under constant development, the *Coot* software package has now reached version 1.0, which incorporates a major rework of the graphical architecture, interface, tools and components of the program. Aside from all of the well known tools for manual model building, the software has a built-in ligand building tool *Lidia*, which can use *AceDRG* (see below) for restraint generation, the ability to create covalent linkages between protein and ligand or between molecular components (Nicholls, Joosten *et al.*, 2021), a semi-automatic *N*-glycan building tool, which is able to build entire oligosaccharides that are consistent with the most common biosynthetic pathways (Emsley & Crispin, 2018), a real-space, accelerated refinement tool that is able to process whole macromolecules, in contrast to the manual localized real-space refinement that users typically perform when fitting or tweaking parts of a model (Casañal *et al.*, 2020), and validation tools that run the most common checks on protein models (Ramachandran plots, rotamer propensities, planarity of the peptide bond, per-residue *B* factors and density-fit analysis, amongst others), plus tools to facilitate ligand fitting (Nicholls, 2017) and validation (Emsley, 2017), for example deviation from ideal geometry values in dictionaries, clashes and interaction maps. *Coot* makes use of the CCP4 Monomer Library to obtain restraints for the most common biomolecule monomers (amino acids, carbohydrates, nucleic acids) and most ligands defined in the PDB Chemical Component Dictionary (Westbrook *et al.*, 2015).

At present, *Coot* is tied to desktop machines due to its reliance on the GTK toolkit (Emsley *et al.*, 2010). This means that users of *CCP*4 Cloud (Krissinel *et al.*, 2022) need to have a local installation of the *CCP*4 suite in order to perform manual model building. However, there is an ongoing effort to produce a web-based interface, which will use the *Coot* engine in the same manner that the GTK version does but without requiring a local *CCP*4 installation.

2.4.2. Automated model building. While *Coot* has incrementally added a wealth of automatic procedures over the years, the *CCP*4 suite includes several fully automated pipelines that combine automated model-building software [*Buccaneer* (Cowtan, 2006) and *Nautilus* (Cowtan, 2014), *ARP*/*wARP* 8.0 (Lamzin *et al.*, 2012) or the chain-tracing code in *SHELXE* (Usón & Sheldrick, 2018)] with reciprocal-space refinement (see Section 2.4.4) and validation [*EDSTATS* (Tickle, 2012) and *MolProbity* (Williams *et al.*, 2018)] to produce protein and nucleic acid models that are completed iteratively. These pipelines, for example *Modelcraft* (Bond & Cowtan, 2022) in *CCP*4*i*2 and *CCP*4*build* in *CCP*4 Cloud, are available from both modern graphical user interfaces (*CCP*4*i*2 and *CCP*4 Cloud) and are completed by either graphical or textual summaries of the completeness of the built model. Outside the protein realm, *AlphaFold* (Jumper *et al.*, 2021) and *RoseTTAfold* (Baek *et al.*, 2021) models can be glycosylated using the glycan library and tools in the *Privateer* software (Bagdonas *et al.*, 2021). *PanDDA* (Pearce *et al.*, 2017) allows users to increase the signal-to-noise ratio of their ligand maps by combining several data sets from ligand-free and ligand-bound forms of the protein; the program has algorithms for combining different crystal forms. The current automated model-building offerings in the suite are completed by *ARP*/*wARP* 8.0 (Lamzin *et al.*, 2012), which was jointly released with *CCP*4 version 7.0 for the first time in 2018; this software pioneered the iterative combination of model building and refinement (Perrakis *et al.*, 1999), a feature that is now present in all modern model-building pipelines, and the automated addition of ligands (Langer *et al.*, 2008). Modern versions of *ARP*/*wARP* may also be used with cryo-EM data (Chojnowski *et al.*, 2021). At a higher level, the *PDB-REDO* pipeline has been integrated into *CCP*4 through graphical interfaces in *CCP*4*i*2 and *CCP*4 Cloud, with API calls to the *PDB-REDO* web server (Joosten *et al.*, 2014).

2.4.3. Restraint dictionaries: the CCP4 Monomer Library. The dictionaries in the CCP4 Monomer Library (Vagin *et al.*, 2004) have been improved by the introduction of *AceDRG* (Long *et al.*, 2017), which since version 7.0 of the suite can also generate restraint dictionaries for covalent linkages (Nicholls, Wojdyr *et al.*, 2021; Nicholls, Joosten *et al.*, 2021). New dictionaries are now routinely generated for many compounds, although pyranose sugars have received a separate treatment to account for their conformational preferences (Atanasova *et al.*, 2022; Joosten *et al.*, 2022). H atoms have been modelled and restrained in their nuclear positions in the CCP4 Monomer Library (Catapano *et al.*, 2021), as informed by neutron diffraction data (Allen & Bruno, 2010).

2.4.4. Refinement. The main tool for full-model reciprocal-space refinement in *CCP*4 is *REFMAC*5 (Murshudov *et al.*, 2011). The program uses the sparse-matrix approximation of the Fisher's information matrix (Steiner *et al.*, 2003) and is designed to be fast and flexible, with a number of refinement methods built into the engine, including restrained, unrestrained and rigid-body refinement. Jelly-body restraints are particularly useful for stabilizing refinement, for example, after molecular replacement, where larger parts of a structure might need to move into place. In addition to controlling model parameterization and performing macromolecular

refinement, *REFMAC*5 also performs map calculation. A variety of types of weighted maps are produced, which allow visualization, subsequent analyses and validation.

*REFMAC*5 allows the addition of case-specific structural knowledge to be utilized during refinement through the external restraints mechanism (Nicholls *et al.*, 2012; Kovalevskiy *et al.*, 2018). These external restraints, which are most useful when only low-resolution data are available, can for instance be generated by *ProSMART* (Nicholls *et al.*, 2014) for proteins and nucleic acids using homologues or backbone hydrogen-bonding patterns, *LibG* (Brown *et al.*, 2015) for nucleic acid base-pairing and stacking, and *Platonyzer* (Touw *et al.*, 2016) for zinc, sodium and magnesium sites. The automated pipeline *LORESTR* (Kovalevskiy *et al.*, 2016) can be used to optimize the refinement protocol at low resolution, expediting the process and easing manual user effort. New developments and the next generation of structure-refinement tools are being implemented in *Servalcat* utilizing the *GEMMI* library (Yamashita *et al.*, 2021, 2023).

The *PAIREF* program (Malý *et al.*, 2020), which has recently been introduced into *CCP*4*i*2, performs automatic paired refinement (Karplus & Diederichs, 2012) using the *REFMAC*5 refinement engine. It analyses the impact of weak reflections beyond the traditional high-resolution diffraction-limit cutoff on the quality of the refined model. The program monitors model and data indicators and model-to-data agreement metrics and implements a decision-suggesting routine for the high-resolution cutoff that may result in the best model. Outside *REFMAC*5 and associated tools, the *SHEETBEND* software (Cowtan *et al.*, 2020) allows a very fast preliminary refinement of the atomic coordinates and, optionally, isotropic or anisotropic *B* factors (Cowtan & Agirre, 2018). It is based on a novel approach in which a shift field, and not atoms, is refined to update and morph models. This approach is particularly indicated to correct large shifts in secondary-structure elements after molecular replacement and is run by default as part of the *Modelcraft* pipeline (Bond & Cowtan, 2022).

2.5. Validation and deposition

Both the *CCP*4*i*2 and *CCP*4 Cloud interfaces include a validation and deposition interface developed in collaboration with the PDBe (the Protein Data Bank in Europe; wwPDB Consortium, 2019; Armstrong *et al.*, 2020). The purpose of this tool is to prepare mmCIF files for deposition; additionally, it provides the convenience of letting users see what their preliminary wwPDB validation report (Gore *et al.*, 2012, 2017) would look like and allowing them to fix errors and notice interesting chemical features of a model before going through the actual deposition process. Also, in preparation for deposition, the model and structure factors are converted into an mmCIF, which in turn allows the wwPDB to pre-populate many of the required metadata for deposition, such as refinement statistics.

Further validation tools exist in *CCP*4 outside this online validation process. Protein model validation can be performed with a variety of tools. *MolProbity* analyses backbone geometry, rotamers and clashes, and produces a script file that will generate a menu within *Coot* containing lists of outliers. *Coot* itself contains a plethora of interactive and live-updated validation tools, ranging from *MolProbity*-equivalent metrics to other less frequently quoted metrics, for example the Kleywegt Plot, which can be of great value depending on the problem. The *EDSTATS* software (Tickle, 2012) provides a unique analysis of model-to-data fit, separating results by main chain and side chain and looking at difference density, with the results being able to point out common modelling problems, such as poorly fitting regions requiring a peptide flip. Version 8.0 of *CCP*4 has seen the gradual inclusion of *PDB-REDO* (Joosten *et al.*, 2012) functionality into the *CCP*4 interfaces; for example *Tortoize* (Sobolev *et al.*, 2020), a tool that analyses main-chain and side-chain geometry and reports *Z*-scores for every amino acid, is now integrated into the *CCP*4 validation tasks. The visual output of *PDB-REDO* calculations is displayed consistently across *CCP*4*i*2, *CCP*4-Cloud and the *PDB-REDO* website by encapsulating various interactive plots and tables in a self-contained single web component. Detection of errors, particularly sequence-register errors, by analysing the agreement between observed contacts and inter-residue distances with the predictions from software such as *AlphaFold*2 (Sánchez Rodríguez *et al.*, 2022) is available in *ConKit* (Simkovic *et al.*, 2017). The *findMySequence* software (Chojnowski *et al.*, 2022) uses machine learning for the identification of unknown proteins in X-ray crystallography and cryo-EM data, with the added benefit of detecting elusive register errors, which may have a detrimental effect on the quality of the rest of the structure. The *Iris* validation framework (Rochira & Agirre, 2021) is a standalone tool that displays a variety of validation metrics as concentric circles, with modelling errors becoming visible as ripples in successive circles. Carbohydrate model validation, including protein glycosylation, can be carried out with the *Privateer* software (Agirre *et al.*, 2015), which in the MKIV version incorporates checks of glycan composition against offline mirrors of several glycomics databases (Bagdonas *et al.*, 2020) and overall glycan conformation using *Z*-scores (Dialpuri *et al.*, 2023). Specific structural radiation-damage sites in structures derived from cryocooled crystals can be identified with *RABDAM* through the $B_{\rm damage}$ (Shelley *et al.*, 2018) and $B_{\rm net}$ (Shelley & Garman, 2022) metrics, and space-group and origin ambiguity may be determined and resolved using *Zanuda* (Lebedev & Isupov, 2014).

2.6. Analysis and representation

PISA (Krissinel & Henrick, 2007) allows the analysis of molecular interfaces, calculating likely assemblies, intramolecular and intermolecular contacts, and accessible areas, offering insight into crystal packing. Intramolecular (predicted) contact maps and other related representations can be visualized with *ConKit* (Simkovic *et al.*, 2017) or online at the *ConPlot* server (Sánchez Rodríguez *et al.*, 2021).

On the representation side, the main tool in *CCP*4 is the *CCP*4 *Molecular Graphics* project (*CCP*4*mg*). Since the last

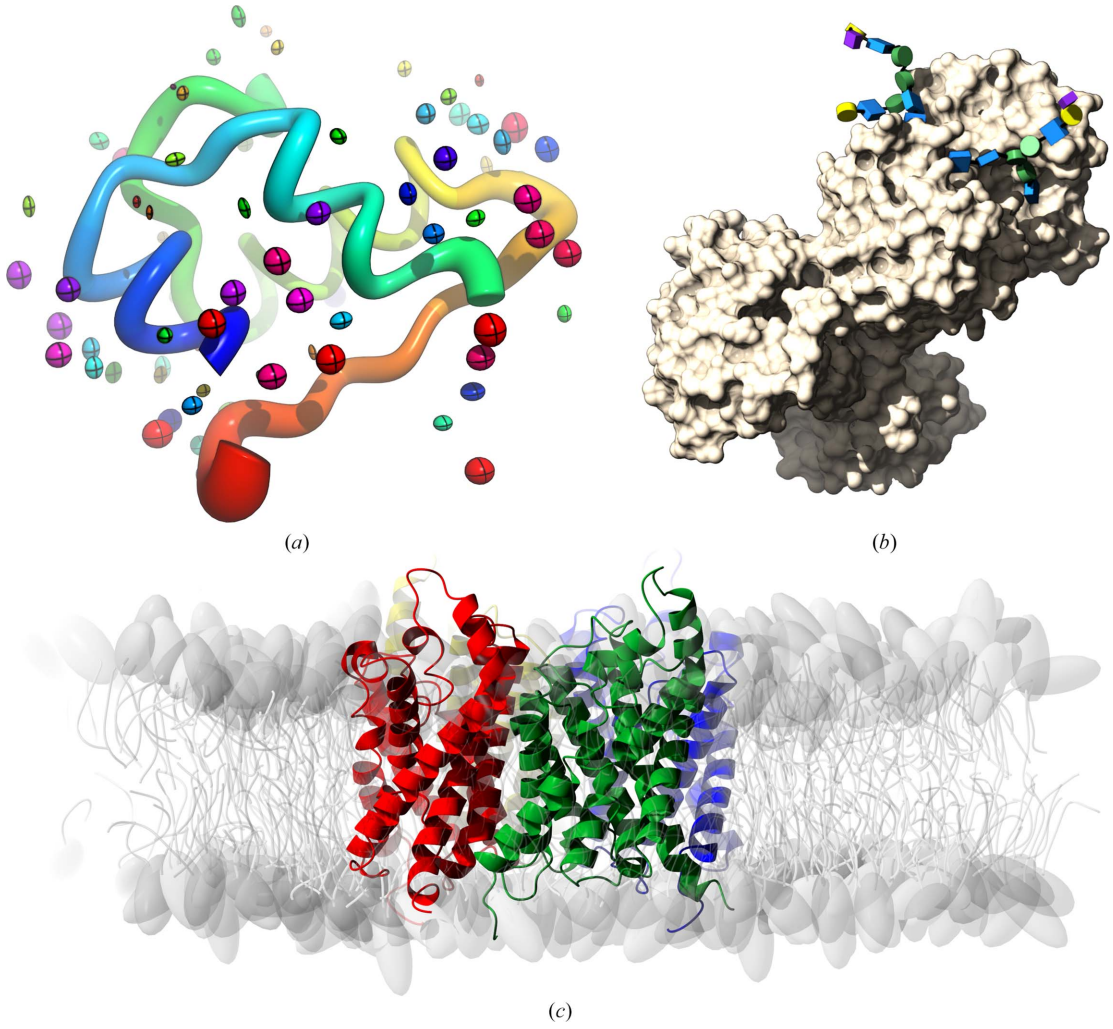


Figure 3
A collection of newer representations included in the *CCP*4 *Molecular Graphics* project (*CCP*4*mg*). (*a*) PDB entry 2bn3 is a high-resolution model of insulin (Nanao *et al.*, 2005); it is shown here as worms, with water molecules drawn as ellipsoids, both coloured and scaled by the anisotropic *B* factors of the model. (*b*) PDB entry 3v8x (Noinaj *et al.*, 2012) is a structure of human transferrin (chain *B*), drawn here as a solvent-accessible surface with *N*-glycans shown as Glycoblocks (McNicholas & Agirre, 2017). (*c*) PDB entry 3c02, a structure of aquaglyceroporin from *Plasmodium falciparum* (Newby *et al.*, 2008), embedded in a lipid bilayer by *CHARMM-GUI* (Jo *et al.*, 2008); lipids are shown as cartoons.

*CCP*4*mg* general publication (McNicholas *et al.*, 2011), the main updates have involved new functionalities for handling cryo-EM maps, 3D representation of *N*-glycans (McNicholas & Agirre, 2017) and the addition of a new interactive interface to the functionality of *MrBUMP* (Keegan *et al.*, 2018). Some newer representations from *CCP*4*mg* can be seen in Fig. 3.

2.7. Under the bonnet

The *dxtbx* toolkit for *DIALS* (Parkhurst *et al.*, 2014) is included as part of the *cctbx* (Grosse-Kunstleve *et al.*, 2002) distribution; the *clipper-python* module (McNicholas *et al.*, 2018), a SWIG wrapper around the original C++ Clipper library, is also included and supports a number of functions of the *CCP*4*i*2 interface, including the *Iris* validation framework (Rochira & Agirre, 2021). At a higher level, *CCP*4*i*2 (Potterton *et al.*, 2018) provides code reusability via the command line, offering a mechanism for executing Python-only pipelines without a running instance of the graphical user interface (headless mode). *CCP*4 Cloud projects and automatic structure-solution workflows can also be initiated from the command line using the 'cloudrun' utility; this is useful for performing serial computations for selected targets. The *Coot* model-building software (Emsley & Cowtan, 2004), originally conceived as a C++ object-oriented toolkit, is now exposed as an importable Python module to allow code reuse in new applications, and is also able to run in headless mode, suppressing all graphical output. Finally, *CCP*4*mg* (McNicholas *et al.*, 2011) is also able to run without graphics, generating images from a scene-description file in XML format; this functionality is used in *CCP*4*i*2 to generate molecular graphics of, for instance, autobuilt structures.

3. Future plans

The transition towards web technologies, which is already under way with the introduction of *CCP*4 Cloud, will be

completed in the near future by the introduction of fully fledged model-building, visualization and figure-preparation web-browser interfaces to the existing *Coot* and *CCP*4*mg* engines. We also foresee an increase in the number of connections to theoretical modelling packages such as *AlphaFold* (Jumper *et al.*, 2021) and *RoseTTAfold* (Baek *et al.*, 2021), as well as deeper harnessing of the AlphaFold Protein Structure Database (Varadi *et al.*, 2022).

4. Software availability and data-access statement

The *CCP*4 software suite can be obtained from https://www.ccp4.ac.uk/download. CCP4 maintains a public instance of *CCP*4 Cloud at https://cloud.ccp4.ac.uk available to both academic and licenced commercial users. No data were generated in the context of the present publication.

5. Individual author contributions

Jon Agirre wrote the majority of the manuscript, coordinated the authors and contributed to *Privateer*, *clipper-python*, *clipper-progs*, *CCP*4*i*2, *CCP*4 Cloud, *Iris*, the CCP4 Monomer Library and other software. Haroldas Bagdonas contributed to *Privateer* MKIV. James Beilsten-Edmands, Luis Fuentes-Montero, Markus Gerstel, Richard J. Gildea, James M. Parkhurst, Nicholas E. Devenish, Melanie Vollmar, David Waterman, Graeme Winter and Gwyndaf Evans contributed to *xia*2 (Winter) and *DIALS*. James Foadi and Gwyndaf Evans developed *BLEND*. Rafael J. Borges, Claudia Millán, Iracema Caballero, Elisabet Jiménez, Josep Triviño Valls and Isabel Usón developed the *ARCIMBOLDO* package, with Massimo Sammito and Ana Medina contributing to *ALEPH*. George Sheldrick is the lead developer of *SHELXC*/*D*/*E*; Isabel Usón is now the main contributor to and maintainer of the *SHELXC*/*D*/*E* suite. Maarten L. Hekkelman, Robbie P. Joosten and Anastassis Perrakis developed the *PDB-REDO* software package. Paul Bond, Soon Wen Hoh and Kevin D. Cowtan contributed to *Modelcraft* and *Buccaneer* (Bond, Hoh and Cowtan), *Nautilus* (Hoh and Cowtan) and the Clipper libraries (Cowtan). Tristan I. Croll, Soon Wen Hoh, Stuart McNicholas and Jon Agirre led the development of the released *clipper-python* module. J. Javier Burgos-Mármol, Ronan M. Keegan, Filomeno Sánchez Rodríguez, Felix Simkovic, Adam J. Simpkin, Jens M. H. Thomas and Daniel J. Rigden developed *SIMBAD*, *MrBUMP*, *ConKit*, *Slice'N'Dice* and *AMPLE*. Stuart J. McNicholas, Kyle Stevenson, Huw T. Jenkins, Eleanor J. Dodson, Keith S. Wilson and Martin E. M. Noble contributed to the development and testing of the *CCP*4*i*2 graphical user interface. John Berrisford and Sameer Velankar contributed towards the development of a validation and deposition task in the *CCP*4 graphical user interfaces. Paul Emsley is the lead developer of *Coot* and associated programs, to which Bernhard Lohkamp has contributed. William Rochira developed *Iris* under Jon Agirre's supervision. Nicholas Pearce contributed *PanDDA* to the suite. Philipp Heuser, Joana Pereira, Egor Sobolev, Grzegorz Chojnowski and Victor S. Lamzin contributed to *ARP*/*wARP* 8.0. Pavol Skubak and Navraj S. Pannu developed *Crank*-2. Oleg Kovalevskiy is the lead developer of *LORESTR*. Fei Long is the lead developer of *AceDRG*, *BALBES* and *LibG*. Garib N. Murshudov is the lead developer of *REFMAC*5. Robert A. Nicholls is the lead developer of *ProSMART*. Mihaela Atanasova, Lucrezia Catapano, Robbie P. Joosten, Andrey A. Lebedev, Fei Long, Stuart J. McNicholas, Garib N. Murshudov, Robert A. Nicholls, Roberto A. Steiner and Keitaro Yamashita contributed to *REFMAC*5 and/or the CCP4 Monomer Library. Andrew G. W. Leslie and Harold R. (Harry) Powell led the development of *MOSFLM* and *iMosflm*, respectively. Andrea Thorn is the lead developer of *AUSPEX*. Phil R. Evans is the developer of *POINTLESS* and *AIMLESS*. Alexei Vagin was the lead developer of *MoRDA*. Airlie J. McCoy, Kaushik Hatti, Robert Oeffner, Massimo Sammito, Claudia Millán and Randy J. Read developed *Phaser* and the associated tools. Eugene Krissinel developed *PISA*, *SSM*, *Gesamt* and, with Andrey A. Lebedev and others, the *CCP*4 Cloud software. Martin Malý and Petr Kolenko designed and implemented the *PAIREF* software. Kathryn L. Shelley and Elspeth F. Garman led the development of *RABDAM*. Maria Fando developed a new documentation architecture for *CCP*4*i*2 and *CCP*4 Cloud and converted, with help from others, old documentation to the new system. Gregorz Chojnowski developed the *findMySequence* software. Martyn Winn wrote the original implementation of TLS refinement in *REFMAC* and contributed to the development of the core C libraries and to *MrBUMP*.

At the time of writing, the CCP4 Executive Committee was composed of David G. Brown, Helen Walden, Kevin D. Cowtan, Judit Debreczeni, Gwyndaf Evans, Michael A. Hough, Dave Lawson, James Murray, Martyn D. Winn, Garib N. Murshudov, Martin E. M. Noble, Randy J. Read, Dan J. Rigden, Ivo Tews, Eugene Krissinel and Keith S. Wilson. Jon Agirre and Arnaud Baslé were subsequently elected as co-chairs of CCP4 Working Group 2 and took seats on the CCP4 Executive Committee, of which Ivo Tews was elected as chair. Charles B. Ballard, Ronan M. Keegan, Andrey A. Lebedev, Maria Fando, Tarik R. Drevon, David Waterman, Ville Uski and Eugene B. Krissinel were the members of the CCP4 Core Team responsible for the maintenance and distribution of the *CCP*4 software suite, *CCP*4 Cloud and website.

Acknowledgements

The *CCP*4 program authors are grateful for the support of more than 150 industrial licensees. CCP4 project members are indebted to Karen McIntyre for her continuous support, dedication and her contribution as CCP4 Equity, Diversity and Inclusion Champion.

Funding information

Jon Agirre is a Royal Society University Research Fellow (UF160039 and URF\R\221006). Mihaela Atanasova is funded by the UK Engineering and Physical Sciences Research Council (EPSRC; EP/R513386/1). Haroldas Bagdonas is

funded by The Royal Society (RGF/R1/181006). José Javier Burgos-Mármol and Daniel J. Rigden are supported by the BBSRC (BB/S007105/1). Robbie P. Joosten is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871037 (iNEXT-Discovery) and by CCP4. This work was supported by the Medical Research Council as part of United Kingdom Research and Innovation, also known as UK Research and Innovation: MRC file reference No. MC_UP_A025_1012 to Garib N. Murshudov, which also funded Keitaro Yamashita, Paul Emsley and Fei Long. Robert A. Nicholls is funded by the BBSRC (BB/S007083/1). Soon Wen Hoh is funded by the BBSRC (BB/T012935/1). Kevin D. Cowtan and Paul S. Bond are funded in part by the BBSRC (BB/S005099/1). John Berrisford and Sameer Velankar thank the European Molecular Biology Laboratory–European Bioinformatics Institute, who supported this work. Andrea Thorn was supported in the development of *AUSPEX* by the German Federal Ministry of Education and Research (05K19WWA and 05K22GU5) and by Deutsche Forschungsgemeinschaft (TH2135/2-1). Petr Kolenko and Martin Malý are funded by the MEYS CR (CZ.02.1.01/0.0/0.0/16_019/0000778). Martin Malý is funded by the Czech Academy of Sciences (86652036) and CCP4/STFC (521862101). Anastassis Perrakis acknowledges funding from iNEXT (grant No. 653706), iNEXT-Discovery (grant No. 871037), West-Life (grant No. 675858) and EOSC-Life (grant No. 824087) funded by the Horizon 2020 program of the European Commission. Robbie P. Joosten has been the recipient of a Veni grant (722.011.011) and a Vidi grant (723.013.003) from the Netherlands Organization for Scientific Research (NWO). Maarten L. Hekkelman, Robbie P. Joosten and Anastassis Perrakis thank the Research High Performance Computing facility of the Netherlands Cancer Institute for providing and maintaining computation resources and acknowledge the institutional grant from the Dutch Cancer Society and the Dutch Ministry of Health, Welfare and Sport. Tarik R. Drevon is funded by the BBSRC (BB/S007040/1). Randy J. Read is supported by a Principal Research Fellowship from the Wellcome Trust (grant 209407/Z/17/Z). Atlanta G. Cook is supported by a Wellcome Trust SRF (200898) and a Wellcome Centre for Cell Biology core grant (203149). Isabel Usón acknowledges support from STFC-UK/CCP4: 'Agreement for the integration of methods into the CCP4 software distribution, ARCIMBOLDO_LOW' and Spanish MICINN/AEI/FEDER/UE (PID2021-128751NB-I00). Pavol Skubak and Navraj Pannu were funded by the NWO Applied Sciences and Engineering Domain and CCP4 (grant Nos. 13337 and 16219). Bernhard Lohkamp was supported by the Röntgen Ångström Cluster (grant 349-2013-597). Nicholas Pearce is currently funded by the SciLifeLab and Wallenberg Data Driven Life Science Program (grant KAW 2020.0239) and has previously been funded by a Veni Fellowship (VI.Veni.192.143) from the Dutch Research Council (NWO), a Long-term EMBO fellowship (ALTF 609-2017) and EPSRC grant EP/G037280/1. David M. Lawson received funding from BBSRC Institute Strategic Programme Grants (BB/P012523/1 and BB/P012574/1). Lucrezia Catapano is the recipient of an STFC/CCP4-funded PhD studentship (Agreement No: 7920 S2 2020 007).

References

Agirre, J. & Dodson, E. (2018). *Protein Sci.* **27**, 202–206.
Agirre, J., Iglesias-Fernández, J., Rovira, C., Davies, G. J., Wilson, K. S. & Cowtan, K. D. (2015). *Nat. Struct. Mol. Biol.* **22**, 833–834.
Allen, F. H. & Bruno, I. J. (2010). *Acta Cryst.* B**66**, 380–386.
Armstrong, D. R., Berrisford, J. M., Conroy, M. J., Gutmanas, A., Anyango, S., Choudhary, P., Clark, A. R., Dana, J. M., Deshpande, M., Dunlop, R., Gane, P., Gáborová, R., Gupta, D., Haslam, P., Koča, J., Mak, L., Mir, S., Mukhopadhyay, A., Nadzirin, N., Nair, S., Paysan-Lafosse, T., Pravda, L., Sehnal, D., Salih, O., Smart, O., Tolchard, J., Varadi, M., Svobodova-Vařeková, R., Zaki, H., Kleywegt, G. J. & Velankar, S. (2020). *Nucleic Acids Res.* **48**, D335–D343.
Atanasova, M., Nicholls, R. A., Joosten, R. P. & Agirre, J. (2022). *Acta Cryst.* D**78**, 455–465.
Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J. & Baker, D. (2021). *Science*, **373**, 871–876.
Bagdonas, H., Fogarty, C. A., Fadda, E. & Agirre, J. (2021). *Nat. Struct. Mol. Biol.* **28**, 869–870.
Bagdonas, H., Ungar, D. & Agirre, J. (2020). *Beilstein J. Org. Chem.* **16**, 2523–2533.
Beilsten-Edmands, J., Winter, G., Gildea, R., Parkhurst, J., Waterman, D. & Evans, G. (2020). *Acta Cryst.* D**76**, 385–399.
Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2012). *Acta Cryst.* D**68**, 1622–1631.
Bond, P. S. & Cowtan, K. D. (2022). *Acta Cryst.* D**78**, 1090–1098.
Brewster, A. S., Waterman, D. G., Parkhurst, J. M., Gildea, R. J., Young, I. D., O'Riordan, L. J., Yano, J., Winter, G., Evans, G. & Sauter, N. K. (2018). *Acta Cryst.* D**74**, 877–894.
Brown, A., Long, F., Nicholls, R. A., Toots, J., Emsley, P. & Murshudov, G. (2015). *Acta Cryst.* D**71**, 136–153.
Burnley, T., Palmer, C. M. & Winn, M. (2017). *Acta Cryst.* D**73**, 469–477.
Caballero, I., Sammito, M., Millán, C., Lebedev, A., Soler, N. & Usón, I. (2018). *Acta Cryst.* D**74**, 194–204.
Casañal, A., Lohkamp, B. & Emsley, P. (2020). *Protein Sci.* **29**, 1069–1078.
Catapano, L., Steiner, R. A. & Murshudov, G. N. (2021). *Acta Cryst.* A**77**, C381.
Chojnowski, G., Simpkin, A. J., Leonardo, D. A., Seifert-Davila, W., Vivas-Ruiz, D. E., Keegan, R. M. & Rigden, D. J. (2022). *IUCrJ*, **9**, 86–97.
Chojnowski, G., Sobolev, E., Heuser, P. & Lamzin, V. S. (2021). *Acta Cryst.* D**77**, 142–150.
Clabbers, M. T. B., Gruene, T., Parkhurst, J. M., Abrahams, J. P. & Waterman, D. G. (2018). *Acta Cryst.* D**74**, 506–518.
Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.
Cowtan, K. (2010). *Acta Cryst.* D**66**, 470–478.
Cowtan, K. (2014). *IUCrJ*, **1**, 387–392.
Cowtan, K. & Agirre, J. (2018). *Acta Cryst.* D**74**, 125–131.
Cowtan, K., Metcalfe, S. & Bond, P. (2020). *Acta Cryst.* D**76**, 1192–1200.
Dialpuri, J. S., Bagdonas, H., Atanasova, M., Schofield, L. C., Hekkelman, M. L., Joosten, R. P. & Agirre, J. (2023). *Acta Cryst.* D**79**, 462–472.
Eddy, S. R. (2011). *PLoS Comput. Biol.* **7**, e1002195.
Emsley, P. (2017). *Acta Cryst.* D**73**, 203–210.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Emsley, P. & Crispin, M. (2018). *Acta Cryst.* D**74**, 256–263.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* D**69**, 1204–1214.
French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517–525.
Fuentes-Montero, L., Parkhurst, J., Gerstel, M., Gildea, R., Winter, G., Vollmar, M., Waterman, D. & Evans, G. (2016). *Acta Cryst.* A**72**, s189–s189.
Gildea, R. J., Beilsten-Edmands, J., Axford, D., Horrell, S., Aller, P., Sandy, J., Sanchez-Weatherby, J., Owen, C. D., Lukacik, P., Strain-Damerell, C., Owen, R. L., Walsh, M. A. & Winter, G. (2022). *Acta Cryst.* D**78**, 752–769.
Ginn, H. M., Brewster, A. S., Hattne, J., Evans, G., Wagner, A., Grimes, J. M., Sauter, N. K., Sutton, G. & Stuart, D. I. (2015). *Acta Cryst.* D**71**, 1400–1410.
Gore, S., Sanz García, E., Hendrickx, P. M. S., Gutmanas, A., Westbrook, J. D., Yang, H., Feng, Z., Baskaran, K., Berrisford, J. M., Hudson, B. P., Ikegawa, Y., Kobayashi, N., Lawson, C. L., Mading, S., Mak, L., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Peisach, E., Sahni, G., Sekharan, M. R., Sen, S., Shao, C., Smart, O. S., Ulrich, E. L., Yamashita, R., Quesada, M., Young, J. Y., Nakamura, H., Markley, J. L., Berman, H. M., Burley, S. K., Velankar, S. & Kleywegt, G. J. (2017). *Structure*, **25**, 1916–1927.
Gore, S., Velankar, S. & Kleywegt, G. J. (2012). *Acta Cryst.* D**68**, 478–483.
Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.
Jenkins, H. T. (2018). *Acta Cryst.* D**74**, 205–214.
Jo, S., Kim, T., Iyer, V. G. & Im, W. (2008). *J. Comput. Chem.* **29**, 1859–1865.
Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. (2012). *Acta Cryst.* D**68**, 484–496.
Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. (2014). *IUCrJ*, **1**, 213–220.
Joosten, R. P., Nicholls, R. A. & Agirre, J. (2022). *Curr. Med. Chem.* **29**, 1193–1207.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature*, **596**, 583–589.
Karplus, P. A. & Diederichs, K. (2012). *Science*, **336**, 1030–1033.
Keegan, R. M., McNicholas, S. J., Thomas, J. M. H., Simpkin, A. J., Simkovic, F., Uski, V., Ballard, C. C., Winn, M. D., Wilson, K. S. & Rigden, D. J. (2018). *Acta Cryst.* D**74**, 167–182.
Kovalevskiy, O., Nicholls, R. A., Long, F., Carlon, A. & Murshudov, G. N. (2018). *Acta Cryst.* D**74**, 215–227.
Kovalevskiy, O., Nicholls, R. A. & Murshudov, G. N. (2016). *Acta Cryst.* D**72**, 1149–1161.
Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
Krissinel, E., Lebedev, A. A., Uski, V., Ballard, C. B., Keegan, R. M., Kovalevskiy, O., Nicholls, R. A., Pannu, N. S., Skubák, P., Berrisford, J., Fando, M., Lohkamp, B., Wojdyr, M., Simpkin, A. J., Thomas, J. M. H., Oliver, C., Vonrhein, C., Chojnowski, G., Basle, A., Purkiss, A., Isupov, M. N., McNicholas, S., Lowe, E., Triviño, J., Cowtan, K., Agirre, J., Rigden, D. J., Uson, I., Lamzin, V., Tews, I., Bricogne, G., Leslie, A. G. W. & Brown, D. G. (2022). *Acta Cryst.* D**78**, 1079–1089.
Lamzin, V. S., Perrakis, A. & Wilson, K. S. (2012). *International Tables for Crystallography*, Vol. F, 2nd online ed., edited by E. Arnold, D. M. Himmel & M. G. Rossmann, pp. 525–528. Chester: International Union of Crystallography.
Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nat. Protoc.* **3**, 1171–1179.
Lebedev, A. A. & Isupov, M. N. (2014). *Acta Cryst.* D**70**, 2430–2443.
Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Cryst.* D**75**, 861–877.
Long, F., Nicholls, R. A., Emsley, P., Gražulis, S., Merkys, A., Vaitkus, A. & Murshudov, G. N. (2017). *Acta Cryst.* D**73**, 112–122.
Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 125–132.
Malý, M., Diederichs, K., Dohnálek, J. & Kolenko, P. (2020). *IUCrJ*, **7**, 681–692.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
McCoy, A. J., Oeffner, R. D., Wrobel, A. G., Ojala, J. R. M., Tryggvason, K., Lohkamp, B. & Read, R. J. (2017). *Proc. Natl Acad. Sci. USA*, **114**, 3637–3641.
McCoy, A. J., Sammito, M. D. & Read, R. J. (2022). *Acta Cryst.* D**78**, 1–13.
McNicholas, S. & Agirre, J. (2017). *Acta Cryst.* D**73**, 187–194.
McNicholas, S., Croll, T., Burnley, T., Palmer, C. M., Hoh, S. W., Jenkins, H. T., Dodson, E., Cowtan, K. & Agirre, J. (2018). *Protein Sci.* **27**, 207–216.
McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. M. (2011). *Acta Cryst.* D**67**, 386–394.
Medina, A., Jiménez, E., Caballero, I., Castellví, A., Triviño Valls, J., Alcorlo, M., Molina, R., Hermoso, J. A., Sammito, M. D., Borges, R. & Usón, I. (2022). *Acta Cryst.* D**78**, 1283–1293.
Millán, C., Jiménez, E., Schuster, A., Diederichs, K. & Usón, I. (2020). *Acta Cryst.* D**76**, 209–220.
Millán, C., Sammito, M. & Usón, I. (2015). *IUCrJ*, **2**, 95–105.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Mylona, A., Carr, S., Aller, P., Moraes, I., Treisman, R., Evans, G. & Foadi, J. (2017). *Crystals*, **7**, 242.
Nanao, M. H., Sheldrick, G. M. & Ravelli, R. B. G. (2005). *Acta Cryst.* D**61**, 1227–1237.
Newby, Z. E. R., O'Connell, J., Robles-Colmenares, Y., Khademi, S., Miercke, L. J. & Stroud, R. M. (2008). *Nat. Struct. Mol. Biol.* **15**, 619–625.
Nicholls, R. A. (2017). *Acta Cryst.* D**73**, 158–170.
Nicholls, R. A., Fischer, M., McNicholas, S. & Murshudov, G. N. (2014). *Acta Cryst.* D**70**, 2487–2499.
Nicholls, R. A., Joosten, R. P., Long, F., Wojdyr, M., Lebedev, A., Krissinel, E., Catapano, L., Fischer, M., Emsley, P. & Murshudov, G. N. (2021). *Acta Cryst.* D**77**, 712–726.
Nicholls, R. A., Long, F. & Murshudov, G. N. (2012). *Acta Cryst.* D**68**, 404–417.
Nicholls, R. A., Wojdyr, M., Joosten, R. P., Catapano, L., Long, F., Fischer, M., Emsley, P. & Murshudov, G. N. (2021). *Acta Cryst.* D**77**, 727–745.
Noinaj, N., Easley, N. C., Oke, M., Mizuno, N., Gumbart, J., Boura, E., Steere, A. N., Zak, O., Aisen, P., Tajkhorshid, E., Evans, R. W., Gorringe, A. R., Mason, A. B., Steven, A. C. & Buchanan, S. K. (2012). *Nature*, **483**, 53–58.
Nolte, K., Gao, Y., Stäb, S., Kollmannsberger, P. & Thorn, A. (2022). *Acta Cryst.* D**78**, 187–195.
Oeffner, R. D., Croll, T. I., Millán, C., Poon, B. K., Schlicksup, C. J., Read, R. J. & Terwilliger, T. C. (2022). *Acta Cryst.* D**78**, 1303–1314.
Parkhurst, J. M. (2020). PhD thesis. University of Cambridge, United Kingdom. https://doi.org/10.17863/CAM.46755.
Parkhurst, J. M., Brewster, A. S., Fuentes-Montero, L., Waterman, D. G., Hattne, J., Ashton, A. W., Echols, N., Evans, G., Sauter, N. K. & Winter, G. (2014). *J. Appl. Cryst.* **47**, 1459–1465.

Pearce, N. M., Krojer, T., Bradley, A. R., Collins, P., Nowak, R. P., Talon, R., Marsden, B. D., Kelm, S., Shi, J., Deane, C. M. & von Delft, F. (2017). *Nat. Commun.* **8**, 15123.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nat. Struct. Biol.* **6**, 458–463.
Potterton, L., Agirre, J., Ballard, C., Cowtan, K., Dodson, E., Evans, P. R., Jenkins, H. T., Keegan, R., Krissinel, E., Stevenson, K., Lebedev, A., McNicholas, S. J., Nicholls, R. A., Noble, M., Pannu, N. S., Roth, C., Sheldrick, G., Skubak, P., Turkenburg, J., Uski, V., von Delft, F., Waterman, D., Wilson, K., Winn, M. & Wojdyr, M. (2018). *Acta Cryst.* D**74**, 68–84.
Powell, H. R., Battye, T. G. G., Kontogiannis, L., Johnson, O. & Leslie, A. G. W. (2017). *Nat. Protoc.* **12**, 1310–1325.
Rochira, W. & Agirre, J. (2021). *Protein Sci.* **30**, 93–107.
Rodríguez, D. D., Grosse, C., Himmel, S., González, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Nat. Methods*, **6**, 651–653.
Sammito, M. D., McCoy, A. J., Hatti, K., Oeffner, R. D., Stockwell, D. H., Croll, T. I. & Read, R. J. (2019). *Acta Cryst.* A**75**, e182.
Sánchez Rodríguez, F., Chojnowski, G., Keegan, R. M. & Rigden, D. J. (2022). *Acta Cryst.* D**78**, 1412–1427.
Sánchez Rodríguez, F., Mesdaghi, S., Simpkin, A. J., Burgos-Mármol, J. J., Murphy, D. L., Uski, V., Keegan, R. M. & Rigden, D. J. (2021). *Bioinformatics*, **37**, 2763–2765.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.
Shelley, K. L., Dixon, T. P. E., Brooks-Bartlett, J. C. & Garman, E. F. (2018). *J. Appl. Cryst.* **51**, 552–559.
Shelley, K. L. & Garman, E. F. (2022). *Nat. Commun.* **13**, 1314.
Simkovic, F., Thomas, J. M. H. & Rigden, D. J. (2017). *Bioinformatics*, **33**, 2209–2211.
Simpkin, A., Simkovic, F., Thomas, J., Savko, M., Ballard, C., Wojdyr, M., Shepard, W., Rigden, D. & Keegan, R. (2018). *Acta Cryst.* A**74**, e173.
Simpkin, A. J., Elliott, L. G., Stevenson, K., Krissinel, E., Rigden, D. J. & Keegan, R. M. (2022). *bioRxiv*, 2022.06.30.497974.
Simpkin, A. J., Simkovic, F., Thomas, J. M. H., Savko, M., Lebedev, A., Uski, V., Ballard, C. C., Wojdyr, M., Shepard, W., Rigden, D. J. & Keegan, R. M. (2020). *Acta Cryst.* D**76**, 1–8.
Simpkin, A. J., Thomas, J. M. H., Keegan, R. M. & Rigden, D. J. (2022). *Acta Cryst.* D**78**, 553–559.
Skubák, P. & Pannu, N. S. (2013). *Nat. Commun.* **4**, 2777.
Sobolev, O. V., Afonine, P. V., Moriarty, N. W., Hekkelman, M. L., Joosten, R. P., Perrakis, A. & Adams, P. D. (2020). *Structure*, **28**, 1249–1258.
Söding, J. (2005). *Bioinformatics*, **21**, 951–960.
Steiner, R. A., Lebedev, A. A. & Murshudov, G. N. (2003). *Acta Cryst.* D**59**, 2114–2124.
Thorn, A., Parkhurst, J., Emsley, P., Nicholls, R. A., Vollmar, M., Evans, G. & Murshudov, G. N. (2017). *Acta Cryst.* D**73**, 729–737.
Tickle, I. J. (2012). *Acta Cryst.* D**68**, 454–467.
Touw, W. G., van Beusekom, B., Evers, J. M. G., Vriend, G. & Joosten, R. P. (2016). *Acta Cryst.* D**72**, 1110–1118.
Trapani, S. & Navaza, J. (2008). *Acta Cryst.* D**64**, 11–16.
Uervirojnangkoorn, M., Zeldin, O. B., Lyubimov, A. Y., Hattne, J., Brewster, A. S., Sauter, N. K., Brunger, A. T. & Weis, W. I. (2015). *eLife*, **4**, e05421.
Usón, I. & Sheldrick, G. M. (2018). *Acta Cryst.* D**74**, 106–116.
Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. & Murshudov, G. N. (2004). *Acta Cryst.* D**60**, 2184–2195.
Vagin, A. & Lebedev, A. (2015). *Acta Cryst.* A**71**, s19.
Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.
Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. & Velankar, S. (2022). *Nucleic Acids Res.* **50**, D439–D444.
Westbrook, J. D., Shao, C., Feng, Z., Zhuravleva, M., Velankar, S. & Young, J. (2015). *Bioinformatics*, **31**, 1274–1278.
Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall, W. B. III, Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (2018). *Protein Sci.* **27**, 293–315.
Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Cryst.* D**67**, 235–242.
Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.
Winter, G., Lobley, C. M. C. & Prince, S. M. (2013). *Acta Cryst.* D**69**, 1260–1273.
Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). *Acta Cryst.* D**74**, 85–97.
Wojdyr, M. (2022). *J. Open Source Softw.* **7**, 4200.
Wojdyr, M., Keegan, R., Winter, G. & Ashton, A. (2013). *Acta Cryst.* A**69**, s299.
wwPDB Consortium (2019). *Nucleic Acids Res.* **47**, D520–D528.
Yamashita, K., Palmer, C. M., Burnley, T. & Murshudov, G. N. (2021). *Acta Cryst.* D**77**, 1282–1291.
Yamashita, K., Wojdyr, M., Long, F., Nicholls, R. A. & Murshudov, G. N. (2023). *Acta Cryst.* D**79**, 368–373.