Preprint Title	Leveraging glycomics data in glycoprotein 3D structure validation with Privateer
Authors	Haroldas Bagdonas, Daniel Ungar and Jon Agirre
Publication Date	20 Jul 2020
Article Type	Full Research Paper
Supporting Information File 1	Supplementary_Figure_1.docx; 555.8 KB
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

Leveraging glycomics data in glycoprotein 3D structure validation with Privateer

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

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

Department of Biology, University of York, UK

[*]Correspondence: jon.agirre@york.ac.uk

Abstract

The heterogeneity, mobility and complexity of glycans in glycoproteins have been, and currently remain, significant challenges in structural biology. Those 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: full-length glycan composition, including linkage details for the glycosidic bonds. These 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.

Keywords

Glycomics; Privateer; Glycoinformatics; X-ray Crystallography; Electron cryo-microscopy.

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[1]; glycosylation plays a crucial part in folding/unfolding pathways of some proteins in cells[2,3]; the level of N-glycan expression regulates 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, influenza virus uses 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 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 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 1). Glycan microheterogeneity and the lack of carbohydrate-specific modelling tools have often been named as 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, these are known as glycoforms[19]. 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 arise due to competition of glycan processing enzymes in biosynthesis pathways[20].

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[21,22].

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[23]. Samples containing glycoprotein molecules do not usually fulfill that criteria. More often than not, MX falls short at elucidating carbohydrate features in glycoproteins due to glycosylated proteins being inherently mobile and heterogeneous[19], 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[24].

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[25]. 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[26]. 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[27,28]. Crucially, MX and Cryo-EM can complement each other to counteract issues that both face individually[29].

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[30], although recent advances in software are closing the gap[31–33]. 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[34] and CARP[16]; more recently, we released the Privateer software[21], which was the first carbohydrate validation tool available as part of the CCP4i2 crystallographic structure solution pipeline[35]. 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[36,37], 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)[38]. 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[31]. 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 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[39].

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[40]. 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)[40]. 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 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[41]. Moreover, recent advances in ion mobility mass spectroscopy can now also be used for complete glycan analysis[42].

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) require prior knowledge of glycan biosynthetic pathways[43]. 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[44]. 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[45] utilized CCSD[45], EuroCarbDB[46] and GlycomeDB[47] used GlycoCT[48] (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[49], which uses both GlycoCT[50] and WURCS[49] 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[51] 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[52], UniCarb-DB[53] 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 & engineering[54], molecular dynamics simulations[55] and protein interaction studies[56], 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[21]. 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.

Methods and results

The algorithm used to generate WURCS notation in Privateer is based on the description published in *Tanaka et al*[57], with required updates applied from *Matsubara et al*[58]. 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 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[59] three-letter codes of carbohydrate monomer definitions found in structure files into WURCS definitions of unique monomers (described as “UniqueRES”[58]).

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 monosaccharide 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 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, 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).

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 (for standalone bundles, please refer to https://github.com/glycojones/privateer/tree/privateerMKIV_noccp4 – installation instructions are provided in README.md in the repository), 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 GlyConnect database (Table 2a and 2b).

Examples of use

As observed in previous studies, glycoprotein models deposited in PDB feature flaws ranging from minor irregularities to gross modelling errors[14,17,60,61]. Automated validation of minor irregularities was already possible with automated tools such as pdb-care[34], CARP[62], and Privateer[21]. 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.

Example 1 - 2H6O:

The glycoprotein model (PDB code 2H6O) proposed by Szakonyi et al63 contains 12 glycans as detected by Privateer. The model became infamous after it sparked submission of a critical correspondence published by Crispin et al14. The article contained a discussion about the proposed model containing glycan 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)-GlcNAc14. 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 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 its 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[52] and UniCarbKB[64]. 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[65] to improve model quality[61] (Figure 5). 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[61]. 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 5). Because the data in GlyConnect is approximately 70% manually curated by experts in the field[52], 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.

Conclusions and future work

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, integration of additional databases might require support from the developers of those databases.

Currently, the generated WURCS strings are matched against an identical sequence in the database. This means that, if the glycoprotein model has a single modelling mistake, for example at one end of the chain, but is correct elsewhere, the current version of software would still fail to return a match. This issue will be solved by subtrees rather than only an exact match. This development 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 Privateer's command line interface and through Coot scripts. In order to facilitate interaction with users, a graphical interface to the new functionality will be provided through the CCP4i2[35] framework in the near future.

Acknowledgements

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]. The work in Daniel Ungar’s group is supported by the BBSRC [grant

number BB/M018237/1]. We would also like to acknowledge the support of the Departments of Chemistry and Biology at the University of York.

References

(1) 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.* **2014**, *34* (5), 1626–1639.
(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. Conformation and Function of the N-Linked Glycan in the Adhesion Domain of Human CD2. *Science* **1995**, *269* (5228), 1273–1278.
(3) Mitra, N.; Sharon, N.; Surolia, A. Role of N-Linked Glycan in the Unfolding Pathway of Erythrina Corallodendron Lectin. *Biochemistry* **2003**, *42* (42), 12208–12216.
(4) Gu, J.; Isaji, T.; Xu, Q.; Kariya, Y.; Gu, W.; Fukuda, T.; Du, Y. Potential Roles of N-Glycosylation in Cell Adhesion. *Glycoconj. J.* **2012**, *29* (8-9), 599–607.
(5) Lyons, J. J.; Milner, J. D.; Rosenzweig, S. D. Glycans Instructing Immunity: The Emerging Role of Altered Glycosylation in Clinical Immunology. *Front Pediatr* **2015**, *3*, 54.
(6) Boscher, C.; Dennis, J. W.; Nabi, I. R. Glycosylation, Galectins and Cellular Signaling. *Current Opinion in Cell Biology*. 2011, pp 383–392. https://doi.org/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. Structure of Influenza Hemagglutinin in Complex with an Inhibitor of Membrane Fusion. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (46), 17736–17741.
(8) Crispin, M.; Ward, A. B.; Wilson, I. A. Structure and Immune Recognition of the HIV Glycan Shield. *Annu. Rev. Biophys.* **2018**, *47*, 499–523.
(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. Structure of the Lassa Virus Glycan Shield Provides a Model for Immunological Resistance. *Proceedings of the National Academy of Sciences*. 2018, pp 7320–7325. https://doi.org/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. African Trypanosomes Evade Immune Clearance by O-Glycosylation of the VSG Surface Coat. *Nat Microbiol* **2018**, *3* (8), 932–938.
(11) Walls, A. C.; Park, Y.-J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181* (2), 281–292.e6.
(12) Wood, N. T.; Fadda, E.; Davis, R.; Grant, O. C.; Martin, J. C.; Woods, R. J.; Travers, S. A. The Influence of N-Linked Glycans on the Molecular Dynamics of the HIV-1 gp120 V3 Loop. *PLoS One* **2013**, *8* (11), e80301.
(13) Lütteke, T.; von der Lieth, C. W. Data Mining the PDB for Glyco-Related Data. *Methods Mol. Biol.* **2009**, *534*, 293–310.
(14) Crispin, M.; Stuart, D. I.; Jones, E. Y. Building Meaningful Models of Glycoproteins. *Nat. Struct. Mol. Biol.* **2007**, *14* (5), 354; discussion 354–355.
(15) Agirre, J.; Davies, G. J.; Wilson, K. S.; Cowtan, K. D. Carbohydrate Structure: The Rocky Road to Automation. *Curr. Opin. Struct. Biol.* **2017**, *44*, 39–47.
(16) Frank, M.; Lutteke, T.; von der Lieth, C.-W. GlycoMapsDB: A Database of the Accessible Conformational Space of Glycosidic Linkages. *Nucleic Acids Research*. 2007, pp 287–290. https://doi.org/10.1093/nar/gkl907.
(17) Agirre, J.; Davies, G.; Wilson, K.; Cowtan, K. Carbohydrate Anomalies in the PDB.

Nat. Chem. Biol. **2015**, *11* (5), 303.
(18) Atanasova, M.; Bagdonas, H.; Agirre, J. Structural Glycobiology in the Age of Electron Cryo-Microscopy. *Curr. Opin. Struct. Biol.* **2019**, *62*, 70–78.
(19) Rudd, P. M.; Dwek, R. A. Glycosylation: Heterogeneity and the 3D Structure of Proteins. *Crit. Rev. Biochem. Mol. Biol.* **1997**, *32* (1), 1–100.
(20) 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*. 2019. https://doi.org/10.3389/fcell.2019.00157.
(21) Agirre, J.; Iglesias-Fernández, J.; Rovira, C.; Davies, G. J.; Wilson, K. S.; Cowtan, K. D. Privateer: Software for the Conformational Validation of Carbohydrate Structures. *Nat. Struct. Mol. Biol.* **2015**, *22* (11), 833–834.
(22) *Essentials of Glycobiology*; 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., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2016.
(23) 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. The Impact of Protein Characterization in Structural Proteomics. *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62* (Pt 10), 1125–1136.
(24) Stura, E. A.; Nemerow, G. R.; Wilson, I. A. Strategies in the Crystallization of Glycoproteins and Protein Complexes. *Journal of Crystal Growth*. 1992, pp 273–285. https://doi.org/10.1016/0022-0248(92)90256-i.
(25) Cheng, Y.; Grigorieff, N.; Penczek, P. A.; Walz, T. A Primer to Single-Particle Cryo-Electron Microscopy. *Cell* **2015**, *161* (3), 438–449.
(26) Serna, M. Hands on Methods for High Resolution Cryo-Electron Microscopy Structures of Heterogeneous Macromolecular Complexes. *Front Mol Biosci* **2019**, *6*, 33.
(27) Fan, X.; Wang, J.; Zhang, X.; Yang, Z.; Zhang, J.-C.; Zhao, L.; Peng, H.-L.; Lei, J.; Wang, H.-W. Single Particle Cryo-EM Reconstruction of 52 kDa Streptavidin at 3.2 Angstrom Resolution. *Nat. Commun.* **2019**, *10* (1), 2386.
(28) Herzik, M. A., Jr; Wu, M.; Lander, G. C. High-Resolution Structure Determination of Sub-100 kDa Complexes Using Conventional Cryo-EM. *Nat. Commun.* **2019**, *10* (1), 1032.
(29) Wang, H.-W.; Wang, J.-W. How Cryo-Electron Microscopy and X-Ray Crystallography Complement Each Other. *Protein Sci.* **2017**, *26* (1), 32–39.
(30) Agirre, J. Strategies for Carbohydrate Model Building, Refinement and Validation. *Acta Crystallogr D Struct Biol* **2017**, *73* (Pt 2), 171–186.
(31) Emsley, P.; Crispin, M. Structural Analysis of Glycoproteins: Building N-Linked Glycans with Coot. *Acta Crystallogr D Struct Biol* **2018**, *74* (Pt 4), 256–263.
(32) Croll, T. I. ISOLDE: A Physically Realistic Environment for Model Building into Low-Resolution Electron-Density Maps. *Acta Crystallogr D Struct Biol* **2018**, *74* (Pt 6), 519–530.
(33) Frenz, B.; Rämisch, S.; Borst, A. J.; Walls, A. C.; Adolf-Bryfogle, J.; Schief, W. R.; Veesler, D.; DiMaio, F. Automatically Fixing Errors in Glycoprotein Structures with Rosetta. *Structure* **2019**, *27* (1), 134–139.e3.
(34) 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*, 69.
(35) 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. CCP4i2: The New Graphical User Interface to the CCP4 Program Suite. *Acta Crystallogr D Struct Biol* **2018**, *74* (Pt 2), 68–84.
(36) 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*. 2017, pp 822–828. https://doi.org/10.1107/s2059798317013353.
(37) Joosten, R. P.; Lütteke, T. Carbohydrate 3D Structure Validation. *Curr. Opin. Struct. Biol.* **2017**, *44*, 9–17.
(38) van Beusekom, B.; Lütteke, T.; Joosten, R. P. Making Glycoproteins a Little Bit Sweeter withPDB-REDO. *Acta Crystallographica Section F Structural Biology Communications*. 2018, pp 463–472. https://doi.org/10.1107/s2053230x18004016.
(39) Nakahara, Y.; Miyata, T.; Hamuro, T.; Funatsu, A.; Miyagi, M.; Tsunasawa, S.; Kato, H. 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* **1996**, *35* (20), 6450–6459.
(40) Shajahan, A.; Heiss, C.; Ishihara, M.; Azadi, P. Glycomic and Glycoproteomic Analysis of Glycoproteins—a Tutorial. *Analytical and Bioanalytical Chemistry*. 2017, pp 4483–4505. https://doi.org/10.1007/s00216-017-0406-7.
(41) Liu, H.; Zhang, N.; Wan, D.; Cui, M.; Liu, Z.; Liu, S. Mass Spectrometry-Based Analysis of Glycoproteins and Its Clinical Applications in Cancer Biomarker Discovery. *Clin. Proteomics* **2014**, *11* (1), 14.
(42) Hofmann, J.; Pagel, K. Glycan Analysis by Ion Mobility-Mass Spectrometry. *Angewandte Chemie International Edition*. 2017, pp 8342–8349. https://doi.org/10.1002/anie.201701309.
(43) Leymarie, N.; Zaia, J. Effective Use of Mass Spectrometry for Glycan and Glycopeptide Structural Analysis. *Anal. Chem.* **2012**, *84* (7), 3040–3048.
(44) Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. GlycoWorkbench: A Tool for the Computer-Assisted Annotation of Mass Spectra of Glycans. *J. Proteome Res.* **2008**, *7* (4), 1650–1659.
(45) Albersheim, P. CarbBank: A Structural and Bibliographic Data Base. 1989. https://doi.org/10.2172/5715461.
(46) 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. EUROCarbDB: An Open-Access Platform for Glycoinformatics. *Glycobiology* **2011**, *21* (4), 493–502.
(47) Ranzinger, R.; Herget, S.; Wetter, T.; von der Lieth, C.-W. GlycomeDB - Integration of Open-Access Carbohydrate Structure Databases. *BMC Bioinformatics* **2008**, *9*, 384.
(48) Herget, S.; Ranzinger, R.; Maass, K.; Lieth, C.-W. V. D. GlycoCT-a Unifying Sequence Format for Carbohydrates. *Carbohydr. Res.* **2008**, *343* (12), 2162–2171.
(49) 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. GlyTouCan: An Accessible Glycan Structure Repository. *Glycobiology* **2017**, *27* (10), 915–919.
(50) 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. GlyTouCan 1.0 – The International Glycan Structure Repository. *Nucleic Acids Research*. 2016, pp D1237–D1242. https://doi.org/10.1093/nar/gkv1041.
(51) Tsuchiya, S.; Yamada, I.; Aoki-Kinoshita, K. F. GlycanFormatConverter: A Conversion Tool for Translating the Complexities of Glycans. *Bioinformatics* **2019**, *35* (14), 2434–2440.
(52) Alocci, D.; Mariethoz, J.; Gastaldello, A.; Gasteiger, E.; Karlsson, N. G.; Kolarich, D.;

Packer, N. H.; Lisacek, F. GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. *J. Proteome Res.* **2019**, *18* (2), 664–677.
(53) Hayes, C. A.; Karlsson, N. G.; Struwe, W. B.; Lisacek, F.; Rudd, P. M.; Packer, N. H.; Campbell, M. P. UniCarb-DB: A Database Resource for Glycomic Discovery. *Bioinformatics* **2011**, *27* (9), 1343–1344.
(54) Congreve, M.; Murray, C. W.; Blundell, T. L. Keynote Review: Structural Biology and Drug Discovery. *Drug Discovery Today*. 2005, pp 895–907. https://doi.org/10.1016/s1359-6446(05)03484-7.
(55) Jong, D. de; de Jong, D.; Periole, X.; Marrink, S. J. Towards Molecular Dynamics Simulations of Large Protein Complexes. *Biophysical Journal*. 2010, p 57a. https://doi.org/10.1016/j.bpj.2009.12.329.
(56) Aloy, P.; Russell, R. B. Interrogating Protein Interaction Networks through Structural Biology. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (9), 5896–5901.
(57) Tanaka, K.; Aoki-Kinoshita, K. F.; Kotera, M.; Sawaki, H.; Tsuchiya, S.; Fujita, N.; Shikanai, T.; Kato, M.; Kawano, S.; Yamada, I.; Narimatsu, H. WURCS: The Web3 Unique Representation of Carbohydrate Structures. *J. Chem. Inf. Model.* **2014**, *54* (6), 1558–1566.
(58) 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.* **2017**, *57* (4), 632–637.
(59) Westbrook, J. D.; Shao, C.; Feng, Z.; Zhuravleva, M.; Velankar, S.; Young, J. The Chemical Component Dictionary: Complete Descriptions of Constituent Molecules in Experimentally Determined 3D Macromolecules in the Protein Data Bank. *Bioinformatics* **2015**, *31* (8), 1274–1278.
(60) Lütteke, T.; Frank, M.; von der Lieth, C.-W. Data Mining the Protein Data Bank: Automatic Detection and Assignment of Carbohydrate Structures. *Carbohydr. Res.* **2004**, *339* (5), 1015–1020.
(61) 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.* **2018**, *74* (Pt 8), 463–472.
(62) 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.* **2005**, *33* (Database issue), D242–D246.
(63) Szakonyi, G.; Klein, M. G.; Hannan, J. P.; Young, K. A.; Ma, R. Z.; Asokan, R.; Holers, V. M.; Chen, X. S. Structure of the Epstein-Barr Virus Major Envelope Glycoprotein. *Nat. Struct. Mol. Biol.* **2006**, *13* (11), 996–1001.
(64) Campbell, M. P.; Peterson, R.; Mariethoz, J.; Gasteiger, E.; Akune, Y.; Aoki-Kinoshita, K. F.; Lisacek, F.; Packer, N. H. UniCarbKB: Building a Knowledge Platform for Glycoproteomics. *Nucleic Acids Res.* **2014**, *42* (Database issue), D215–D221.
(65) 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. Crystal Structure of the TLR4-MD-2 Complex with Bound Endotoxin Antagonist Eritoran. *Cell* **2007**, *130* (5), 906–917.
(66) 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* **2019**, *75* (Pt 9), 804–816.
(67) 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). 2019. https://doi.org/10.2210/pdb6mzk/pdb.
(68) Lee, P. S.; Ohshima, N.; Stanfield, R. L.; Yu, W.; Iba, Y.; Okuno, Y.; Kurosawa, Y.; Wilson, I. A. Receptor Mimicry by Antibody F045-092 Facilitates Universal Binding to the H3 Subtype of Influenza Virus. *Nat. Commun.* **2014**, *5*, 3614.

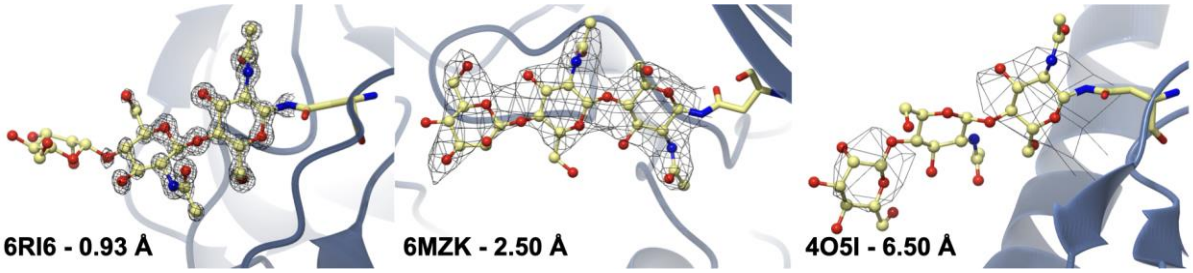

Figure 1: Comparison of glycan features in electron density maps over a range of resolutions from select glycoprotein structures (PDB entries: 6RI6[66]; 6MZK[67]; 4O5I[68]) Electron Density maps obtained with X-Ray crystallography. Data resolution and PDB entry IDs associated with structures have been 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.

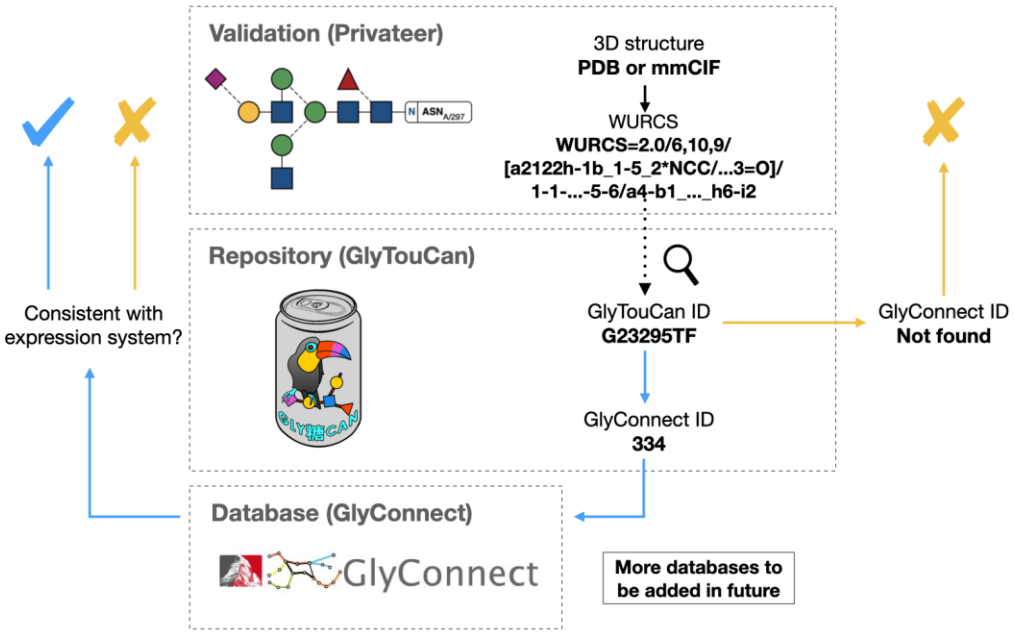

Figure 2: A roadmap of the software development project that allows Structural Biologists to quickly obtain detailed information about specific glycans in Glycoprotein models from Glycomics/Glycoproteomics databases. The GlyTouCan (https://glytoucan.org/) and GlyConnect (https://glyconnect.expasy.org/) logos have been reproduced here under explicit permission from their respective authors.

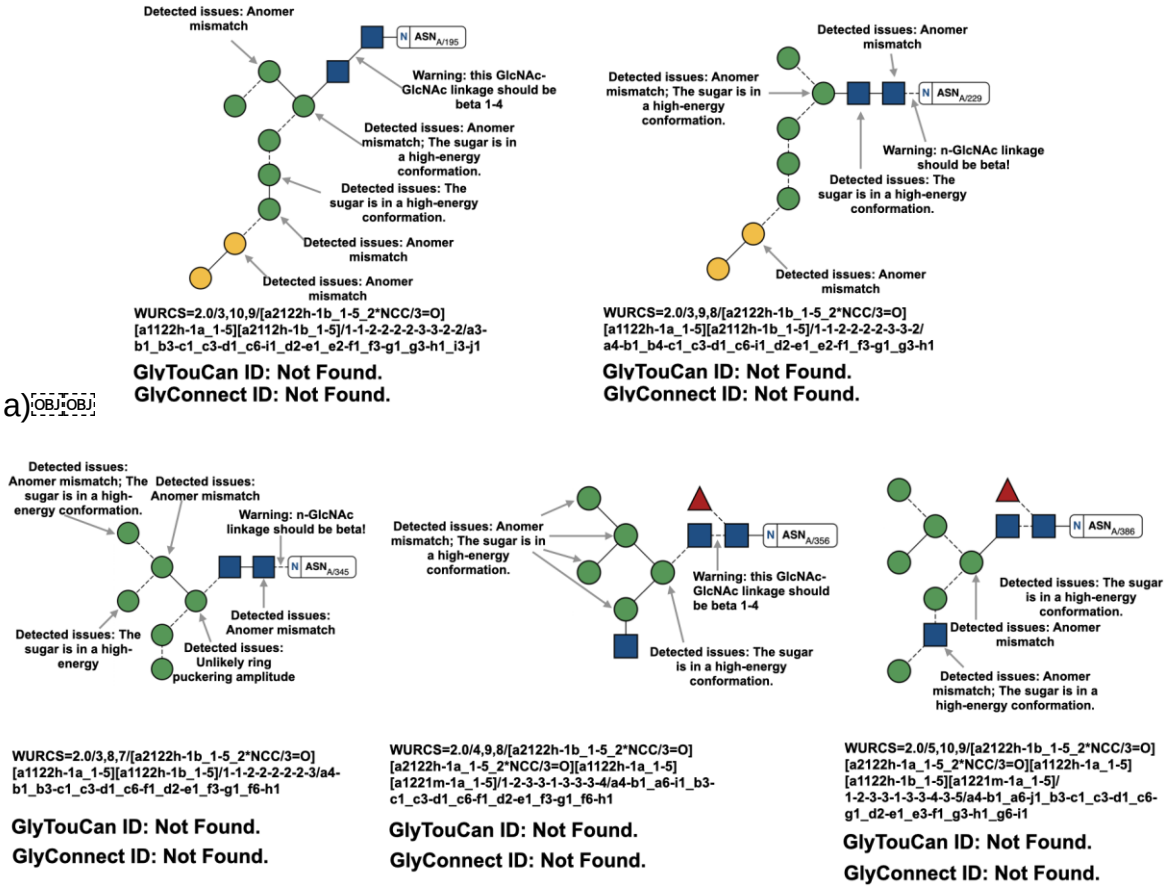

Figure 3: N-linked glycans detected by Privateer in Epstein Barr Virus Major Envelope Glycoprotein (PDB entry: 2H6O[63]). a) Depicts all the detected glycan chains that failed to return GlyTouCan and GlyConnect IDs, with their WURCS sequences generated and modelling errors detected by Privateer. b) Depicts 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 depicts the locations in WURCS notation where both glycans differ.

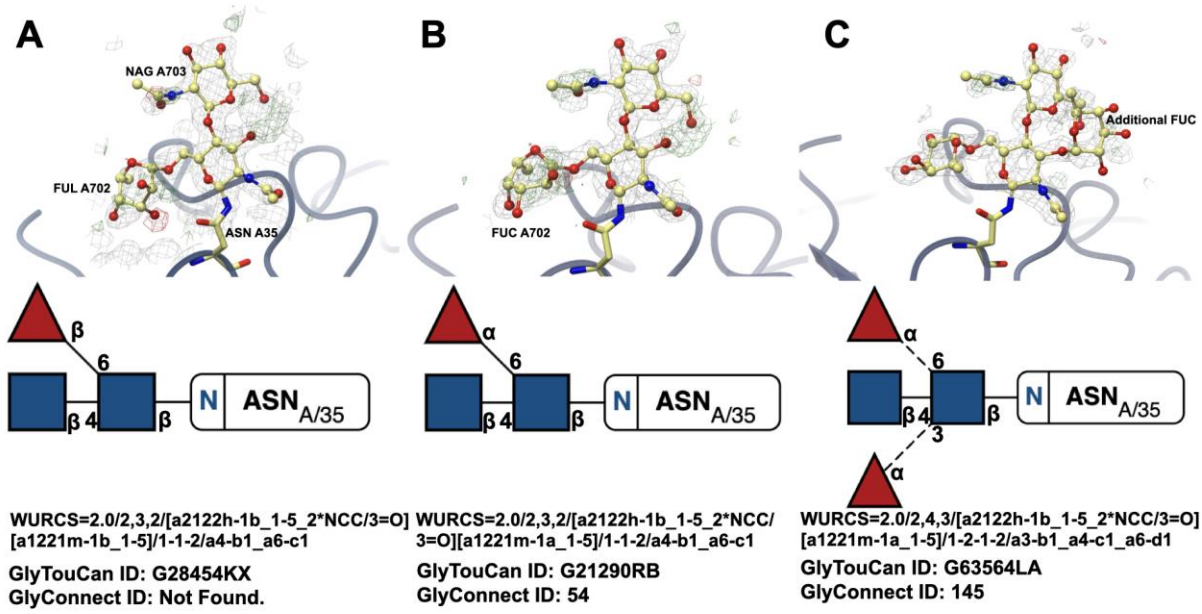

Figure 4: An N-linked glycan attached to Asn35 of human Toll-like receptor 4 (A: PDB entry 2z62[65]). Model iteratively rebuilt by van Beusekom et al. and PDB-Redo as shown in steps B and C. Pictures at the top depict glycoprotein models of the region of interest and electron density maps of the glycan 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.

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

Table 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[58].

a)

Glycan chain length	GlyTouCan ID found	GlyTouCan ID not found	% of GlyTouCan in GlyConnect	Total glycan chains
1	16797	0	1%	16797
2	5870	5	90%	5875
3	2550	17	71%	2567
4	1012	21	80%	1033
5	834	72	74%	906
6	460	85	69%	545
7	345	55	77%	400
8	235	25	85%	260
9	164	16	81%	180
10	118	5	92%	123
11	20	5	85%	25
12	8	4	75%	12
13	0	1	0%	1
14	0	0	0%	0
15	2	0	0%	2
16	0	1	0%	1

b)

Glycan chain length	GlyTouCan ID found	GlyTouCan ID not found	% of GlyTouCan in GlyConnect	Total glycan chains
1	2080	0	3%	2080
2	1081	0	98%	1081
3	439	0	96%	439
4	143	0	93%	143
5	146	2	85%	148
6	70	1	97%	71
7	45	0	100%	45
8	26	0	88%	26
9	15	1	100%	16
10	16	0	100%	16
11	4	0	100%	4
12	1	0	100%	1
13	1	0	0%	1

Table 2: Comparison of successful glycan matches detected by Privateer in GlyTouCan and GlyConnect database. **a)** Glycan obtained from glycoprotein models elucidated by X-Ray crystallography. **b)** Glycan obtained from glycoprotein models elucidated by Cryo-EM.