Carbohydrate structure: the rocky road to automation

Jon Agirre*, Gideon J. Davies, Keith S. Wilson and Kevin D. Cowtan*

York Structural Biology Laboratory, Department of Chemistry, University of York, York, YO10 5DD, UK

Correspondence: jon.agirre@york.ac.uk and kevin.cowtan@york.ac.uk

Abstract

With the introduction of intuitive graphical software, structural biologists who are not experts in crystallography are now able to build complete protein or nucleic acid models rapidly. In contrast, carbohydrates are in a wholly different situation: scant automation exists, with manual building attempts being sometimes toppled by incorrect dictionaries or refinement problems. Sugars are the most stereochemically-complex family of biomolecules and, as pyranose rings, have clear conformational preferences. Despite this, all refinement programs may produce high-energy conformations at medium to low resolution, without any support from the electron density. This problem renders the affected structures unusable in glyco-chemical terms.

Bringing structural glycobiology up to "protein standards" will require a total overhaul of the methodology. Time is of the essence, as the community is steadily increasing the production rate of glycoproteins, and electron cryo-microscopy has just started to image them in precisely that resolution range where crystallographic methods falter most.

Introduction

Carbohydrates are among the most stereochemically-complex biomolecules. This complexity leads to highly specialised and selective interactions that can play key roles in folding, stabilisation and recognition, overall dynamic processes, and at the same time form the basis for the generation of enduring, static structures such as plant tissue. Such specialisation has traditionally posed many obstacles to advances in glyco-chemistry: bacteria, responsible for most of the current production of recombinant proteins, are largely incapable of glycosylating nascent polypeptides - a process by which certain amino acids conforming to a consensus sequence (sequon) have an oligosaccharide (glycans) covalently attached to a nitrogen (N-glycans) or oxygen atom (O-glycans) of their side chain by a transference enzyme. Furthermore, glycans may not be amenable to structural analysis due to their flexibility, or may even preclude crystallographic studies if their inherent flexibility on the protein surface hinders the formation of crystal contacts, which usually prompts their enzymatic removal as part of the preparation of the sample for crystallisation.

The last decade saw the introduction of new experimental techniques that have almost doubled the structural throughput of glycoproteins. About 10% of the structures deposited annually contain carbohydrates but, while those covalently-linked accounted for ~2.5% of the total in the early 2000's, this number has increased to ~5% since 2010 [1]. It is apparent that the structural biology community has been caught off-guard; a number of communications have raised issues on the way carbohydrates are represented in structural databases [2-4], with numerous problems affecting nomenclature, structure and conformation which, in combination, may affect more than 30% of the glyco-related structural data deposited in the Protein Data Bank (PDB).

Problems arise from a variety of sources. The first and most obvious is the insufficient knowledge of glycan composition and structure when they are not resolved in the electron density maps - such was the case with Vp54 from *Paramecium bursaria* chlorella virus 1, for which a structure (PDB code 1M3Y) was originally determined at 2.0Å resolution containing wrong N-glycans [5]. A recent study

[6] has highlighted just how unexpected the composition of the glycans was at that time, opening the door to the deposition of a corrected version of the structure.

Aside from scientific reasons, a number of technical difficulties have also stacked up over the years: the PDB is not IUPAC-compliant, and therefore creates a gap between how chemistry is defined and the way it is currently represented - e.g. D-mannopyranose is encoded as two three-letter codes, MAN (α-anomer) and BMA (β-anomer). Some dictionary generation programs may produce an improbable high-energy conformer as starting coordinates, or create torsion restraints that lock it into that, or other high-energy conformation [1]. Model building and refinement programs do not take conformational preferences into account, which has a deleterious knock-on effect on other aspects of the model, from interactions to linkage torsions. Finally, while the conformation, geometry, structure and interactions of amino acids have been analysed and reviewed frequently and regularly, the glyco-related structural literature has been largely restricted to the biochemical field, with very little impact in structural biology. This, hopefully, has started to change [1,4,7,8].

Some advances have been made towards bringing carbohydrates up to the high standards set by the protein ecosystem (*vide infra*). Nevertheless, producing meaningful carbohydrate models at medium to low resolution remains a big and underestimated challenge that must be addressed, as cryoEM will soon be routinely delivering many new structures of great biological importance in this resolution range, thanks to advances in direct-electron detector technology and image processing software [9]. In this review, we revisit the recent developments that have come to alleviate some of the difficulties traditionally associated with carbohydrate structure and conformations. Although it entails a deviation from the proposed review format, the text also includes practical pieces of advice that will hopefully help users avoid the most common mistakes.

Conformational anomalies in the PDB

Most sugars find their most stable form in a six-membered saturated ring, which we call a pyranose – pyranoside if the anomeric OH has been converted into OR - in analogy to the oxygen heterocycle (tetrahydro)pyran. Six-membered saturated rings have clear conformational preferences determined by a minimisation of angle and trans-annular strains, which are caused by repulsion between the substituents at each stereocentre. Because it generally provides the optimal conformation - *i.e.* staggered instead of eclipsed - of substituents across each torsionable bond in the ring, their most stable form is a chair ($^4C_1$ or $^1C_4$ in IUPAC nomenclature - see Fig. 1 for a complete description - depending on which carbon atoms lie above and below the main ring plane), with conformational transitions being forced upon catalytic events played on by a carbohydrate-active enzyme [10], or CAZy (for a classification of enzymes, refer to [11]). In protein structural terms, this would equate to an amino acid having a very strong preference for one particular rotamer. Higher energy conformations comprise envelopes - *e.g.* ^3E, with carbon 3 on the upper side of the ring and the rest of the atoms being roughly coplanar - half-chairs - *e.g.* $^2H_1$, with carbon 2 on the upper side of the ring and carbon 1 on the lower side of the ring - boats - *e.g.* $^{2,5}B$ - and skew-boats - *e.g.* $^1S_5$.

Although other conventions are available for determining or even naming conformation [12,13], the most widely used is the Cremer-Pople algorithm [14], which generates two angles ($0^\circ < \phi < 360^\circ$; $0^\circ < \theta < 180^\circ$) and one puckering amplitude (measured in Å). These parameters, illustrated in Fig. 1, have been recently popularised through their use as collective variables for the calculation of free-energy landscapes of the reaction coordinate of several CAZys by the metadynamics approach [15,16] and their adoption by the Privateer software [17] from the CCP4 suite [18]. The Cremer-Pople sphere and its Mercator projection, both depicted in Fig. 1, also offer a clear view of the different conformational itineraries - *i.e.* how a pyranose sugar must go through a half-chair or envelope conformation in order to interconvert between chair and boat conformations.

Despite carbohydrate-centric literature describing chair conformations as relatively rigid [19], a recent view of the conformational landscape of N-glycan forming D-pyranosides in the PDB has revealed that almost 30% are modelled in conformations other than chairs [4]. What would be otherwise a very surprising result in chemical terms – those monosaccharides in N-glycans are not expected to be distorted, as conformational distortion is typically only expected to happen in the active site of carbohydrate-active enzymes – was shown to be a consequence of a combination of many factors, including low real space correlation to electron density, wrongly-modelled sugar structures and under-parameterised refinement. These are in addition to other reported problems [2,3], making the PDB a polluted database in glyco-chemical terms.

Many anomalies start from a dictionary

This very first step in building a structure is usually an invisible one due to refinement programs reading restraints for most common monomers automatically from a library. This would not represent a problem if the library was correct. However, at least 60 entries of the CCP4 monomer library corresponding to carbohydrates have wrong torsion restraints which, if used, will lead to high-energy conformations [1]. Torsion restraints are just one way of simulating torsional strain, which is one of the main drivers behind conformational transitions in ring systems. Sugar entries in the CCP4 monomer library use the now outdated Engh & Huber geometric target [20], although plans are underway to replace these entries with a more accurate target using the new CCP4 dictionary-generation program (AceDRG, F Long *et al.*, unpublished; URL: https://fg.oisin.rc-harwell.ac.uk/projects/acedrg). When used with pyranose sugars, AceDRG has been shown to produce geometric targets which show close agreement with state-of-the-art small molecule data mining applications [1], such as CCDC Mogul [21].

Additionally, the choice of starting coordinates can have an adverse impact on the final, refined coordinates. Since this set of coordinates should reflect the most probable conformer, and most of these are computationally energy-minimised instead of experimentally determined [1], any errors will simply propagate into the model building program – which will use these coordinates as the initial form of the sugar, before any refinement is done – and then into the PDB, should the new ligand be deposited [1,22].

Anomalies created during refinement

When fitting and refining a structure in real space - e.g. in COOT [23] - crystallographers are able to adjust a weighting term that balances the information coming from geometric restraints against that from the experiment itself. This decision is informed purely by visual assessment of the local features of an electron density map, thus different parts of a model can be refined differently depending on a subjective interpretation of the map's quality. In reciprocal space, this formulation is not currently possible, as weights are determined globally - typically by aiming for the best fit-to-data that maintains deviation from ideal geometry within a certain threshold (e.g. 0.020Å for bond lengths). In general, those sections of an electron density map corresponding to the most mobile parts of a macromolecule will be of poorer quality than the rest. This issue is often augmented when working with sugars, as solvent-exposed glycans – including those potentially involved in recognition processes in the form of glycosylation – can show a high degree of conformational variability in comparison to those linked to the core of a glycoprotein (see Fig. 1, top panel). Therefore, globally-set weights will have a negative impact on the geometry of the most flexible parts, which ought to be restrained more tightly.

The most commonly-used way of introducing conformational preferences is by imposing torsion restraints, although other hybrid approaches are showing promise (*vide infra*). These restraints contain, along a tolerance factor, an angular value that is typically measured from the initial, minimal-energy conformation – the one that should be reflected in the dictionaries' starting coordinates – and a periodicity index, which should reflect how many staggered conformations are found in a full, 360º-degree rotation around the torsionable bond. For an endocyclic bond between two sp^3-hybridised carbons, this index would adopt a value of 3. While this periodicity implies that multiple conformations can be contemplated, reducing this value to 1 results in torsion-capable refinement programs restraining a sugar's conformation to the minimal-energy one. As these torsion restraints restraint just one angular value, they are termed 'aperiodic' [1].

Detecting, correcting and reporting conformational anomalies

The Privateer software can process most carbohydrate entries from the PDB Chemical Component Dictionary [24], and in addition offers the possibility of defining new sugars through a graphical user interface (CCP4i2, included in the CCP4 suite [18]). As conformational anomalies may also appear after a wrong choice of sugar or linkage, the glycan structures reported by Privateer should ideally be checked against existing databases (recently reviewed in [25]). Upon detection of high-energy conformations, the software creates dictionaries containing aperiodic torsion restraints in standard CIF format, which can be read by most model building and refinement programs [17]. Privateer has been used successfully to detect and prevent conformational anomalies on medium [26,27] (also shown in Fig. 1) and low resolution [28] crystallographic structures, and more recently on cryoEM data [29]. The software in addition provides statistics for the crystallographic Table 1: the number and percentage of pyranose sugars in the lowest and higher energy conformations. Reporting these is strongly advised, as the presence of conformational anomalies can hint at problems during refinement or errors in model building [17].

Other alternatives to using torsion restraints do exist: the recent inclusion of the AMBER molecular mechanics package [30] into the PHENIX crystallographic suite is expected to produce good results, as one of the strengths of AMBER is precisely its carefully-calibrated torsion potentials.

As conformational distortions may occur in an enzyme's active site, rigidly enforcing one conformation on a ligand sugar near the catalytic residues is not advisable unless the sugar's conformation is not clearly discernible in the map.

Impact on glycosidic bond torsions

Glycosidic bonds have conformational preferences too, and can currently be analysed in terms of torsions and compared to deposited data with the CARP server [7]. However, these torsional data are affected by conformational anomalies, which may not be detected by screening modelling errors out, especially at lower resolution. In order to integrate torsional validation of linkages into a process that by nature is iterative (build, refine, validate, repeat), a new set of torsional data must be derived by excluding conformational anomalies, and be made available through graphics programs (*e.g.* COOT) for iterative validation. For completeness, the influence of interactions (H-bond [31], stacking [32]) on the selection of alternative link energy minima [26,33] should also be reflected. Such is the case of the GlcNAc-Asn bond in N-glycosylation (Fig. 1, bottom panel): a secondary energy minimum can be found at $45^{\circ} < \varphi_N < 110^{\circ}$, which may be selected based on neighbouring interactions [26].

Building carbohydrates automatically

Building carbohydrates manually involves the iterative repetition of many simple steps as the individual monosaccharides are read from a dictionary, fitted to density and linked together. This procedure is clumsy and prone to error. A commonly used alternative involves producing idealised structures of whole glycans which then undergo a minimisation of energy. This can be done for instance with the GLYCAM carbohydrate builder [34]. Other shortcuts may involve the use of pre-defined oligosaccharides - *e.g.* cellohexaose, which can be obtained through the three-letter code CE6, although this approach is not well suited to modelling glycosylation.

N-glycosylation, which may show a predictable core structure [35], can now be modelled semi-automatically using the COOT software [7], but in contrast still there are no mature tools that can match the functionality that ARP/WARP [36], BUCCANEER [37] or PHENIX.autobuild [38] offer for protein. By analogy, the first challenge sugar modelling software will have to overcome is the specification of a sequence – more precisely, structure, as carbohydrates often contain branches – that will not be known with certainty unless very conclusive mass spectra are available. Therefore, programs have to rely on common expected features, such as trying to build the core of a glycan linked to asparagine, leaving the decorations to the user (*vide infra*).

A semi-automated module for building N-glycans (COOT)

This tool, available from the ‘Modules’ submenu under the ‘Extensions’ menu, offers individual addition of monosaccharides, to be chosen from a selection of sugars and linkages (*e.g.* ‘ALPHA1-2 MAN’ or ‘BETA1-4 NAG’) that contain the most common building blocks for glycoproteins, with some notorious exceptions such as β1-2 xylose, α1-3 fucose (both typical of plant glycans) or the different sialic acids. It also offers a scripted addition of an oligomannose glycan, which stops automatically based on fit to electron density. The tool simply requires the user to focus on an asparagine residue, select the relevant entry in the menu, and COOT will then start fitting the monosaccharides into the density, creating LINK records (*i.e.* not REFMAC5's richer LINKR definition) as appropriate. On the negative side, as this tools relies on the same dictionaries as COOT, it is not exempt from creating conformational anomalies at medium to low resolution (Fig. 2).

Despite its present shortcomings, this tool is already able to save a substantial amount of unexciting routine model building time, and will become very powerful once it is perfected and combined with an interactive validation tool.

Automated sugar identification and model building

Although general automated tools for detecting ligands are available [39,40], currently the only carbohydrate-specific one is the CCP4 Sails program (J Agirre and K Cowtan, unpublished; URL: https://fg.oisin.rc-harwell.ac.uk/projects/sails). This software relies on deposited data for generating fingerprints of sugars which are then matched to the experimental map in a fast six-dimensional search, similarly to how the NAUTILUS program builds nucleic acid [41]. The applicability of this detection technique, which may be used for interactive or offline identification of sugars, has been severely limited by how flawed deposited sugar data are, thus making it evident that very strict validation criteria had to be set and implemented first. In addition, as available data are scarce for most sugars other than GlcNAc, unmodified hexoses (glucose, mannose, galactose, for instance) and fructose, it is clear that less frequently-deposited sugars will have to be matched to the existing fingerprints (*e.g.* heparan sulphate: detect glucuronic acid by matching glucose stereochemistry first, then detect GlcNAc). Initial tests with the reduced set of sugars currently built into the program indicate that the program is able to identify most cases (ligand and glycosylation sugars) where complete density is available, and can even deal in some cases with mutarotation at the reducing end of a polysaccharide (e.g. PDB code 5AGD, chain A, MAN:A/BMA:B with ID 500 [42,43]).

Visualisation

New advances are being made towards having a simplified 3D representation of sugars and their interactions: SweetUnityMol [44] transfers and extends the colour code used in the familiar 'Essentials of Glycobiology' nomenclature [45] into texture hexagonal shapes, which are then annotated with the position of the endocyclic oxygen atom in order to confer a notion of orientation. The recently-introduced Glycoblocks representation [46] uses identical shapes and colours to those of the Essentials nomenclature on irregular polyhedra, and simplifies H-bonds and stacking interactions as black and red dashed lines respectively, depicted between each monosaccharide block and the C_α atom of the participating residue (Fig. 3). The Glycoblocks view incorporates the 'Essentials of Glycobiology 3rd edition' style [45] vector glycan diagrams generated with Privateer, which show validation information (conformation, mean B-factor, anomeric and absolute configuration) as a tooltip.

Conclusions and perspectives

Despite the latest developments, carbohydrates are still a long way from their protein counterparts in terms of structural methodology. For example, a glycosylation-equivalent of the Conformation-Dependent Library, which has set a new standard for protein geometry [47], is probably not yet even under investigation.

While many of the criteria recently set by the PDB and its ligand validation task force [48] will have a positive impact on carbohydrates, several aspects particular to sugars, such as ring conformation or branching structure, have been left out of the discussion and may have to be dealt with in the future. We should like to emphasise that correctly annotating ring conformation is of great importance not only for monosaccharides, but for all ligands containing saturated rings.

Due to the incremental availability of methods that are capable of determining and correcting errors in carbohydrate structures [7,17], it is just a matter of time that a successful re-refinement project such as

PDB_REDO [49] catches up with them and produces better carbohydrate models for existing PDB entries, something that is already a reality for proteins [50]. Most of the required functionality exists already, and except for the case of sugars in active sites (which may require further validation), action can be taken upon detecting higher energy conformations. As PDB_REDO uses REFMAC5 for refinement, conformational preferences may be introduced specifically for each monosaccharide using the refinement program's interface for external restraints. This is an exciting potential future development that, if successful, could provide for instance much cleaner torsional statistics of glycosidic links, or a better understanding of protein-glycan and glycan-glycan contacts, something that is of critical importance to the design of antibodies.

Fortunately, the macromolecular crystallographic community is becoming increasingly aware of the need to prevent and report conformational anomalies [4], in the same way that Ramachandran outliers are treated in the protein ecosystem. Since 2015, several high-profile structural studies have acknowledged successfully using Privateer to this effect [28,29,51-54]. Finally, reporting pyranose conformations for glycoprotein structures in the crystallographic information table (known as Table 1) is becoming increasingly common in publications [26-28], leading to a better understanding of how model building and refinement were carried out.

Acknowledgements

The authors would like to express their gratitude to Paul Emsley, Garib N. Murshudov (MRC-LMB, Cambridge, UK) and the rest of the CCP4 Ligands initiative for promoting discussion on ligand chemistry and restraints in an open and inclusive manner. This work was partly supported by the Biotechnology and Biological Sciences Research Council [grant number BB/K008153/1].

References

1. Agirre J: **Strategies for carbohydrate model building, refinement and validation**. *Acta Crystallogr D Struct Biol* 2016, **D73, DOI: 10.1107/S2059798316016910**.

•• This article describes the state of the art of manual carbohydrate model building, from the generation of a dictionary for a new sugar to the validation of conformations, linkages and structures of glycans.

2. Lütteke T, Frank M, von der Lieth CW: **Data mining the protein data bank: automatic detection and assignment of carbohydrate structures**. *Carbohydr Res* 2004, **339**:1015-1020.

3. Crispin M, Stuart DI, Jones EY: **Building meaningful models of glycoproteins**. *Nat Struct Mol Biol* 2007, **14**:354; discussion 354-355.

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

•• Reported a conformational analysis of N-glycan forming D-pyranosides in the PDB, showing that conformational preferences are not accounted for in macromolecular crystallographic refinement programs.

5. Nandhagopal N, Simpson AA, Gurnon JR, Yan X, Baker TS, Graves MV, Van Etten JL, Rossmann MG: **The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus**. *Proc Natl Acad Sci U S A* 2002, **99**:14758-14763.

6. De Castro C, Molinaro A, Piacente F, Gurnon JR, Sturiale L, Palmigiano A, Lanzetta R, Parrilli M, Garozzo D, Tonetti MG, et al.: **Structure of N-linked oligosaccharides attached to chlorovirus PBCV-1 major capsid protein reveals unusual class of complex N-glycans**. *Proc Natl Acad Sci U S A* 2013, **110**:13956-13960.

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

• A summary of the available tools for structural glycobiologists within the COOT and CNS software packages, with a section on validation.

8. Joosten RP, Lütteke T: **Carbohydrate 3D Structure Validation**. *Current Opinion in Structural Biology* 2017, **---- This issue ----**.

•• A review of all the classic and recent developments in carbohydrate structure validation, including practical examples.

9. Kühlbrandt W: **The resolution revolution**. *Science* 2014, **343**:1443-1444.

10. Davies GJ, Henrissat B: **Cracking the code, slowly: the state of carbohydrate-active enzymes in 2013 (Editorial)**. *Curr Opin Struct Biol* 2013, **23**:649–651.

11. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B: **The carbohydrate-active enzymes database (CAZy) in 2013**. *Nucleic Acids Res* 2014, **42**:D490-495.

12. Makeneni S, Foley BL, Woods RJ: **BFMP: a method for discretizing and visualizing pyranose conformations**. *J Chem Inf Model* 2014, **54**:2744-2750.

13. Strauss HL, Pickett HM: **Conformational structure, energy, and inversion rates of cyclohexane and some related oxanes**. *Journal of the American Chemical Society* 1970, **92**:7281-7290.

14. Cremer D, Pople JA: **General definition of ring puckering coordinates**. *Journal of the American Chemical Society* 1975, **97**:1354-1358.

15. Ardèvol A, Iglesias-Fernández J, Rojas-Cervellera V, Rovira C: **The reaction mechanism of retaining glycosyltransferases**. *Biochem Soc Trans* 2016, **44**:51-60.

• An overview of the latest advances on the metadynamics simulation of the controversial catalytic mechanism of retaining glycosyltransferases by quantum mechanics/molecular dynamics (QM/MM) approach.

16. Iglesias-Fernández J, Raich L, Ardèvol A, Rovira C: **The complete conformational free energy landscape of [small beta]-xylose reveals a two-fold catalytic itinerary for [small beta]-xylanases**. *Chemical Science* 2015, **6**:1167-1177.

• Contains the first view of the complete conformational landscape of cyclohexane as calculated by ab initio metadynamics using the Cremer-Pople puckering coordinates as collective variables.

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

•• Describes the Privateer software, which can be used for detecting and correcting conformational anomalies in monosaccharides as well as generating vector diagrams of glycans using the "Essentials of Glycobiology" nomenclature.

18. 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 Crystallographica Section D-Biological Crystallography* 2011, **67**:235-242.

19. Bertozzi CR, Rabuka D: **Structural Basis of Glycan Diversity**. In *Essentials of Glycobiology*, edn 2nd. Edited by Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, Hart GW, Etzler ME; 2009.

20. Engh RA, Huber R: **Accurate bond and angle parameters for X-ray protein structure refinement**. *Acta Crystallographica Section A* 1991, **47**:392-400.

21. Bruno IJ, Cole JC, Kessler M, Luo J, Motherwell WDS, Purkis LH, Smith BR, Taylor R, Cooper RI, Harris SE, et al.: **Retrieval of Crystallographically-Derived Molecular Geometry Information**. *Journal of Chemical Information and Computer Sciences* 2004, **44**:2133-2144.

22. Liu YC, Li YS, Lyu SY, Hsu LJ, Chen YH, Huang YT, Chan HC, Huang CJ, Chen GH, Chou CC, et al.: **Addendum: Interception of teicoplanin oxidation intermediates yields new antimicrobial scaffolds**. *Nat Chem Biol* 2015, **11**:361.

23. Emsley P, Lohkamp B, Scott WG, Cowtan K: **Features and development of Coot**. *Acta Crystallographica Section D-Biological Crystallography* 2010, **66**:486-501.

24. Westbrook JD, 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**:1274-1278.

25. Yuriev E, Ramsland PA: **Carbohydrates in Cyberspace**. *Front Immunol* 2015, **6**:300.

• A summary of the databases currently available to the structural glycobiologist, with examples.

26. 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 beta-D-glucosidases**. *Acta Crystallogr D Struct Biol* 2016, **72**:254-265.

•• Contains a protocol for preventing conformational anomalies during crystallographic refinement with medium to low resolution data, using CCP4 software (ACEDRG, COOT and REFMAC5)

27. Gudmundsson M, Hansson H, Karkehabadi S, Larsson A, Stals I, Kim S, Sunux S, Fujdala M, Larenas E, Kaper T: **Structural and functional studies of the glycoside hydrolase family 3 β-glucosidase Cel3A from the moderately thermophilic fungus Rasamsonia emersonii**. *Acta Crystallographica Section D: Structural Biology* 2016, **72**:860-870.

• An example of how a re-refinement of a glycoprotein structure with the introduction of conformational preferences has led to lower final R-factors.

28. Stewart-Jones GB, Soto C, Lemmin T, Chuang GY, Druz A, Kong R, Thomas PV, Wagh K, Zhou T, Behrens AJ, et al.: **Trimeric HIV-1-Env Structures Define Glycan Shields from Clades A, B, and G**. *Cell* 2016, **165**:813-826.

•• A low resolution (3.4-3.7Å) crystallographic work of the glycan shield of HIV-1, which comprises more than 90 N-linked high-mannose glycans, refined and validated with the aid of Privateer.

29. Pallesen J, Murin CD, de Val N, Cottrell CA, Hastie KM, Turner HL, Fusco ML, Flyak AI, Zeitlin L, Crowe JE, et al.: **Structures of Ebola virus GP and sGP in complex with therapeutic antibodies**. *Nature Microbiology* 2016, **1**:16128.

•• This article reports the interaction of therapeutic antibodies with Ebola virus GP and sGP. The structural work was done by cryoEM, refined and validated with the aid of Privateer.

30. Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, Merz KM, Jr., Onufriev A, Simmerling C, Wang B, Woods RJ: **The Amber biomolecular simulation programs**. *J Comput Chem* 2005, **26**:1668-1688.

31. Fernández-Alonso MdC, Díaz D, Berbis MÁ, Marcelo F, Cañada J, Jiménez-Barbero J: **Protein-Carbohydrate Interactions Studied by NMR: From Molecular Recognition to Drug Design**. *Current Protein & Peptide Science* 2012, **13**:816-830.

32. Hudson KL, Bartlett GJ, Diehl RC, Agirre J, Gallagher T, Kiessling LL, Woolfson DN: **Carbohydrate–Aromatic Interactions in Proteins**. *Journal of the American Chemical Society* 2015, **137**:15152-15160.

•• A multidisciplinary analysis of the main drivers behind protein-carbohydrate complexation, and the first data mining study to filter out carbohydrate models based on conformation and correlation to electron density.

33. Topin J, Lelimousin M, Arnaud J, Audfray A, Perez S, Varrot A, Imberty A: **The Hidden Conformation of Lewis x, a Human Histo-Blood Group Antigen, Is a Determinant for Recognition by Pathogen Lectins**. *ACS Chem Biol* 2016, **11**:2011-2020.

• A very good example of how stacking interactions can favour alternate linkage energy minima

34. **Woods Group. (2005-2017) GLYCAM Web. Complex Carbohydrate Research Center, University of Georgia, Athens, GA. (**http://www.glycam.com/**)**. Edited by.

35. Stanley P, Cummings RD: **Structures Common to Different Glycans**. In *Essentials of Glycobiology*, edn 2nd. Edited by Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, Hart GW, Etzler ME; 2009.

36. Langer G, Cohen SX, Lamzin VS, Perrakis A: **Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7**. *Nat Protoc* 2008, **3**:1171-1179.

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

38. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Moriarty NW, Zwart PH, Hung LW, Read RJ, Adams PD: **Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard**. *Acta Crystallogr D Biol Crystallogr* 2008, **64**:61-69.

39. Carolan CG, Lamzin VS: **Automated identification of crystallographic ligands using sparse-density representations**. *Acta Crystallogr D Biol Crystallogr* 2014, **70**:1844-1853.

40. Echols N, Moriarty NW, Klei HE, Afonine PV, Bunkoczi G, Headd JJ, McCoy AJ, Oeffner RD, Read RJ, Terwilliger TC, et al.: **Automating crystallographic structure solution and refinement of protein-ligand complexes**. *Acta Crystallogr D Biol Crystallogr* 2014, **70**:144-154.

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

• This article offers a detailed description of the algorithm at the core of SAILS, the upcoming automated sugar modelling program from the CCP4 suite.

42. Thompson AJ, Speciale G, Iglesias-Fernández J, Hakki Z, Belz T, Cartmell A, Spears RJ, Chandler E, Temple MJ, Stepper J, et al.: **Evidence for a boat conformation at the transition state of GH76 alpha-1,6-mannanases--key enzymes in bacterial and fungal mannoprotein metabolism**. *Angew Chem Int Ed Engl* 2015, **54**:5378-5382.

43. Thompson AJ, Speciale G, Iglesias-Fernández J, Hakki Z, Belz T, Cartmell A, Spears RJ, Chandler E, Temple MJ, Stepper J, et al.: **Corrigendum: Evidence for a Boat Conformation at the Transition State of GH76 alpha-1,6-Mannanases--Key Enzymes in Bacterial and Fungal Mannoprotein Metabolism**. *Angew Chem Int Ed Engl* 2016, **55**:1949.

44. Perez 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.

45. 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.

46. McNicholas S, Agirre J: **Glycoblocks: a schematic 3D representation for glycans and their interactions**. *Acta Crystallogr D Struct Biol* 2016, **D73, DOI: 10.1107/S2059798316013553**.

• Describes the Glycoblocks representation within the CCP4 Molecular Graphics project. This representation uses the "Essentials of Glycobiology" convention, extended to work in 3D, and is able to depict H-bond and stacking interactions schematically.

47. 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.

48. Adams PD, Aertgeerts K, Bauer C, Bell JA, Berman HM, Bhat TN, Blaney JM, Bolton E, Bricogne G, Brown D, et al.: **Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop**. *Structure* 2016, **24**:502-508.

49. Joosten RP, Long F, Murshudov GN, Perrakis A: **The PDB_REDO server for macromolecular structure model optimization**. *IUCRj* 2014, **1**:7.

• Describes how structures can be improved automatically using the latest computational methods and well-established protocols.

50. Touw WG, Joosten RP, Vriend G: **New Biological Insights from Better Structure Models**. *J Mol Biol* 2016, **428**:1375-1393.

• A detailed view of the improved chemistry of re-refined macromolecular models, showing the potential for a remediation of distorted carbohydrate structures.

51. Attia M, Stepper J, Davies GJ, Brumer H: **Functional and structural characterization of a potent GH74 endo-xyloglucanase from the soil saprophyte Cellvibrio japonicus unravels the first step of xyloglucan degradation**. *FEBS Journal* 2016, **283**:1701-1719.

52. Yin DT, Urresti S, Lafond M, Johnston EM, Derikvand F, Ciano L, Berrin J-G, Henrissat B, Walton PH, Davies GJ: **Structure-function characterization reveals new catalytic diversity in the galactose oxidase and glyoxal oxidase family**. *Nature communications* 2015, **6**.

53. Wu L, Viola CM, Brzozowski AM, Davies GJ: **Structural characterization of human heparanase reveals insights into substrate recognition**. *Nature structural & molecular biology* 2015, **22**:1016-1022.

54. Bokhove M, Nishimura K, Brunati M, Han L, De Sanctis D, Rampoldi L, Jovine L: **A structured interdomain linker directs self-polymerization of human uromodulin**. *Proceedings of the National Academy of Sciences* 2016, **113**:1552-1557.

55. Yu TF, Maestre-Reyna M, Ko CY, Ko TP, Sun YJ, Lin TY, Shaw JF, Wang AH: **Structural insights of the ssDNA binding site in the multifunctional endonuclease AtBFN2 from Arabidopsis thaliana**. *PLoS One* 2014, **9**:e105821.

56. Collins PJ, Vachieri SG, Haire LF, Ogrodowicz RW, Martin SR, Walker PA, Xiong X, Gamblin SJ, Skehel JJ: **Recent evolution of equine influenza and the origin of canine influenza**. *Proc Natl Acad Sci U S A* 2014, **111**:11175-11180.

57. Imberty A, Perez S: **Stereochemistry of the N-glycosylation sites in glycoproteins**. *Protein Engineering* 1995, **8**:10.

58. Marrero A, Duquerroy S, Trapani S, Goulas T, Guevara T, Andersen GR, Navaza J, Sottrup-Jensen L, Gomis-Ruth FX: **The crystal structure of human alpha2-macroglobulin reveals a unique molecular cage**. *Angew Chem Int Ed Engl* 2012, **51**:3340-3344.

59. Nnamchi CI, Parkin G, Efimov I, Basran J, Kwon H, Svistunenko DA, Agirre J, Okolo BN, Moneke A, Nwanguma BC, et al.: **Structural and spectroscopic characterisation of a heme peroxidase from sorghum**. *J Biol Inorg Chem* 2016, **21**:63-70.

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

61. Lee JH, Ozorowski G, Ward AB: **Cryo-EM structure of a native, fully glycosylated, cleaved HIV-1 envelope trimer**. *Science* 2016, **351**:1043-1048.

•• An example of what cryoEM can do at present with a heavily-glycosylated structure, such as the HIV-1 envelope trimer. This structure, natively glycosylated with complex glycans, was modelled using idealised glycan structures, which were manually adjusted in a graphics program.

Figure 1, top panel. The Cremer-Pople sphere and its Mercator projection. Those IUPAC conformation denominations illustrated by a diagram have been highlighted in bold (C: chair; H: half-chair; E: envelope; B: boat; S: skew-boat). Wavy lines identify those atoms defining the main ring plane - *e.g.* in a $^4C_1$ chair, the second, third and fifth carbons, and the endocyclic oxygen are roughly coplanar, while the anomeric carbon (C1) and the fourth carbon lie on the lower and upper side of the ring respectively. **Blue pentagons:** monosaccharides from N-glycans in an endonuclease from *Arabidopsis thaliana* (PDB code 4CXP [55]), determined at 1.2Å resolution. As atomic positions are usually clearly established in electron density maps at this resolution, no torsion restraints were used. As expected, all D-pyranosides show a $^4C_1$ chair conformation, while the α1-3 core-linked L-fucose instead shows a $^1C_4$ chair conformation due to its inverted absolute configuration. **Light blue squares:** a re-refined deposition of a β-glucosidase from the moderately thermophilic fungus *Rasamsonia emersonii* (PDB code 5JU6, supersedes 4D0J [27]). For reasons of clarity, only D-pyranosides from chain A are shown – glycosylation in this enzyme is composed of high-mannose oligosaccharides exclusively. Despite the low resolution, the authors were able to maintain the expected conformations ($^4C_1$ chair) in the re-refined entry (5JU6) by activating torsion restraints, arriving at lower R-factors than in the original deposition (4D0J): 0.173/0.228 vs 0.184/0.235 (R/R_{free}). **Yellow circles:** D-pyranosides from canine haemagglutinin (PDB code 4UO4 [56]). Without torsion restraints nor any other mechanism preserving the initial conformation, many of the monosaccharides show high-energy conformations that cannot be ascertained from the electron density. **Bottom panel.** This diagram, annotated with the energy minima calculated by Imberty and Perez [57], shows the torsions in terms of which the conformation of the GlcNAc-Asn link is typically expressed, φ_N and ψ_N. In a similar way to what is shown in the previous example, both high resolution and lower resolution but well-restrained structures show linkages in the energy minima, whereas a low resolution structure that was refined without torsion restraints for sugars (PDB code 4ACQ [58]) shows implausible bond conformations.

Figure 2. Typical results obtained with the semi-automated N-glycosylation modelling tool within COOT. Those monosaccharides that were skipped by the software have been depicted in greyscale in the 2D diagram, and those built without good experimental support (signified by a real space correlation coefficient – or RSCC – of less than 0.8) have been annotated with a red number. 2mFo-DFc electron density maps were contoured at 1.5σ. **a. Modelling plant glycans in a 1.3 Å-resolution electron density map (natively-glycosylated haem peroxidase from sorghum [59], PDB code 5AOG).** Due to an apparent lack of support for α1-3 core-linked fucose and β1-2 linked xylose (2 and 5 in the figure), the program skipped them without trying to fit anything else in an otherwise clear electron density. Strangely, mannose 6 was fitted in a wrong orientation, thus ended up completely distorted and disconnected from the rest. An additional mannose (7) was added, which the depositors had decided not to model due to unclear electron density, and this achieved a low RSCC of 0.64. **b. Modelling a high-mannose glycan in a 1.8 Å-resolution map (natively-glycosylated fungal GH3 glycosyl hydrolase [26], PDB code 5FJI).** The program was able to trace a glycan that matched the deposited one very closely, although it modelled an additional mannose (5) for which density was scarce. Figure prepared with CCP4mg [60] and Privateer [17] following the 'Essentials of Glycobiology 3rd edition' notation [45].

Figure 3. Advances towards a streamlined 3D visualisation of carbohydrate structures and their interactions. a. Original view from N611, one of the glycan structures on the envelope glycoprotein (Env) trimer of HIV-1 clade 2 (PDB code 5FUU [61], Figure 3A from the original manuscript reprinted with permission from AAAS). The figure shows two views, rotated by 60º along the vertical axis, of the N611 tri-antennary glycan and the neighbouring CDRH2 domain.
b. *Glycoblocks* cartoon representation of the same scenario. This panel offers a clear sketch of the interaction scenario by matching the atomic models of the monosaccharides to 3D extensions of the geometric shapes and colours proposed by the 'Essentials of Glycobiology' nomenclature [45] and plotting glycan-protein contacts between the blocks and the C_{α} of the interacting residue, making it possible to omit side-chains from the picture. The N611 tri-antennary glycan establishes two hydrogen bonds with Ser 70 and Arg 19 from the adjacent FWRH3 domain (coloured in orange), with one galactose molecule (yellow circle) in close proximity of the neighbouring CDRH2 (coloured in pink) domain [61]. Besides the reported contacts, the α1-6 core-linked fucose (red triangle) is also within H-bond distance of Asn 100 in a neighbouring chain. **c. Vector view of the same glycan in 2D.** Diagram produced with Privateer [17] following the 'Essentials of Glycobiology 3rd edition' notation [45].