Updated restraint dictionaries and automated model building of pyranose carbohydrates

Mihaela Atanasova

Doctor of Philosophy

University of York
Chemistry

September 2022

Abstract

Carbohydrates are essential biomolecules, which facilitate biological processes involving other biomolecules, notably proteins. But, as opposed to proteins, carbohydrates can have complex stereochemistry, multiple native forms, polymeric branching and tight conformational preferences. The Protein Data Bank (PDB) has been shown to contain numerous errors in carbohydrate structures determined with X-ray crystallography or electron cryo-microscopy. This is partly because the software aimed at solving carbohydrate structures have yet to become as featureful as their protein counterparts. The aim of this thesis is to better understand the problems affecting carbohydrate structures in the PDB and develop software methods to address those. Following the analysis in the first chapter (also in Atanasova et al., 2020), two areas were targeted for improvement - model refinement and automated model building.

Refinement software uses dictionaries with chemical geometry data about molecules, such as bond lengths and angles, that can be used when the X-ray crystallography or electron cryo-microscopy data are unclear. However, carbohydrate dictionaries have been reported to contain errors that have led to incorrect structures. The aim of the work reported on the second chapter was to introduce a completely new set of restraint dictionaries that correct these errors and add new unimodal torsion restraints. These allowed for carbohydrate conformation to be fixed automatically, as evidenced by the results presented here (also in Atanasova et al., 2022).

Finally, a new method for building N-glycans into electron density/potential maps was explored. This new software, called Sails, uses fingerprint-assisted detection: it relies on a database of monosaccharide fingerprints, which it uses to scan a map and locate sugars. The method was found to be moderately successful at detecting monosaccharides at medium to high resolution, showing most hits at the beginning of N-glycans – this opens the door to the extension of the glycan chain by other approaches.

List of Contents

List of Tables

ring shape (all pyranose in this study); 'Ok?' presents the result of Privateer's tri-state validation diagnosis, as introduced in the main text; Lastly, 'name' is the full IUPAC name of the monosaccharide.

List of Figures

the number of analysed sugars after refinement with the CCP4-ML dictionaries (red); after refinement with the new updated dictionaries (blue); and after refinement with the new dictionaries with activated unimodal torsion restraints (yellow). (A) shows all analysed pyranosides; (B) only includes pyranosides that were diagnosed with 'check' or 'no' for at least one protocol.

Figure 2.4. Refinement with the new dictionaries and unimodal torsion restraints leads to fewer unlikely carbohydrate conformations. (A) Sugars part of N/O-glycosylation; (B) Other sugars. θ vs ϕ plot for D-sugars (blue circles) and L-sugars (yellow triangles) – see Discussion for a description of θ and the Cremer-Pople parameters. D-sugars usually adopt the $^4C_1$ conformation with $\theta \approx 0°$; L-sugars normally adopt $^1C_4$ conformation with $\theta \approx 180°$. Use of the new unimodal torsion restraints (top) shows fewer deviations from these values. The PDB codes corresponding to entries discussed in Figures 5 and 6 are labelled. The number of sugars in high energy conformations (according to Privateer) is shown in the bottom right corner of each plot. Resolution ranges contain equal numbers of sugars (1,668 each). High resolution is 0.9 to 1.8 Å, medium resolution is 1.8 to 1.9 Å and low resolution is 1.9 to 2 Å. 63

Figure 2.5. Sugars in unusual conformations after refinement with the new dictionaries with unimodal torsion restraints. (A) BMA-B-3 from PDB ID 5JUG (Y. Jin et al. 2016); (B) SIA-A-522 from PDB ID 6HG0 (Salinger, M.T., Hobbs, J.R., Murray, J.W., Laver, W.G., Kuhn, P., Garman, E.F., n.d.); (C) NAG-E-1 from PDB ID 5O7U (Tobola et al. 2018); (D) GLC-C-1 from PDB ID 5UPM (Pluvinage et al. 2017). These sugars appear as outliers in Figure 4(B). They remain in high-energy conformations after refinement, but have high RSCC. This figure was produced with CCP4mg (S. McNicholas et al. 2011). Map types are 2Fo-Fc, displayed at 1σ contour level with a sampling rate of 0.5. 64

Figure 2.6. Change in conformation and real-space correlation coefficient (RSCC) after refinement. (A) Sugar in a $^1S_5$ conformation after refinement with its old CCP4-ML restraint dictionary entry (Figure 2.4A, bottom middle panel). (B) The conformation of the sugar has been changed to the minimal energy conformation after refinement with the updated restraint dictionary entry and unimodal torsion restraints and the RSCC has increased (Figure 2.4A, top middle panel); the sugar in (A) and (B) is MAN-Q-4 from PDB ID 4UO0 (Devi et al. 2015) at 1.90 Å resolution, mean B value 34 Å^2. (C) Sugar in a ^2S_O conformation after refinement 65

Acknowledgements

First of all, I would like to thank my supervisors Jon Agirre and K Cowtan for giving me the opportunity to work on this project and for supporting me throughout this PhD. I would also like to thank Rob Nicholls and Robbie Joosten for their contributions to some of the research presented here and the invaluable discussions. Additionally, I would like to thank my fellow group members Harold and Jordan for collaborating with me. I would like to thank Paul for all the help and support throughout my PhD. I would like to thank my colleagues at the York Structural Biology Laboratory for making my time during my PhD very enjoyable. Finally, I would like to thank my family for always believing in me and my partner Charles for being my rock.

Declaration

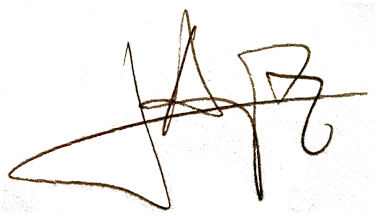
I declare that this thesis is a presentation of original work and I am the sole author, with the exception of the published or collaborative work listed below.

- Atanasova, Mihaela, Haroldas Bagdonas, and Jon Agirre. 2019. “Structural Glycobiology in the Age of Electron Cryo-Microscopy.” *Current Opinion in Structural Biology* 62 (December): 70–78 was published as part of this PhD project and is reproduced in Section 1.4, Chapter 1.
    - MA conducted the literature review.
    - MA collected and analysed the data.
    - MA wrote the majority of the text with guidance and feedback from JA.
    - MA produced all figures except for:
        - Figure 1.4 (produced by JA and HB).
        - Figure 1.7 is reproduced from Emsley, Paul, and Max Crispin. 2018. “Structural Analysis of Glycoproteins: Building N-Linked Glycans with Coot.” *Acta Crystallographica Section D: Structural Biology* 74 (4): 256–63.

Signed:

- Atanasova, Mihaela, Robert A. Nicholls, Robbie P. Joosten, and Jon Agirre. 2022. “Updated Restraint Dictionaries for Carbohydrates in the Pyranose Form.” *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 4): 455–6 was published as part of this PhD project and is reproduced in Chapter 2.
    - MA produced the new dictionaries together with RN.
    - MA validated all new dictionaries.
    - JA wrote the Privateer code for producing the unimodal torsion restraints.
    - MA added the unimodal torsion restraints to the new dictionaries.
    - RJ carried out the testing of the new dictionaries.
    - MA analysed the outcome of the testing, adjusted dictionary generation procedure or fed back modifications to other software as required, and produced all figures except Figure 2.1 (produced by JA).
    - MA wrote the majority of the text with guidance and feedback from all co-authors.
- The supplementary material from Atanasova, Mihaela, Robert A. Nicholls, Robbie P. Joosten, and Jon Agirre. 2022. “Updated Restraint Dictionaries for Carbohydrates in the Pyranose Form.” *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 4): 455–6 is reproduced in Appendix A.

Signed:



- The software presented in Chapter 3, Sails, was initially written by K Cowtan and Jon Agirre. MA calculated and iteratively improved the fingerprints, and put them in a database, made improvements to the software algorithm and architecture, carried out the testing and analysis presented in this thesis.
- Figures 1.1-1.3 are reproduced from Agirre, Jon. 2017. “Strategies for Carbohydrate Model Building, Refinement and Validation.” *Acta Crystallographica Section D: Structural Biology* 73 (2): 171–86.

The following papers include work produced during this PhD and are included in Appendices B and C:

- Dialpuri, Jordan S., Haroldas Bagdonas, Mihaela Atanasova, Lucy Schofield, Maarten L. Hekkelman, Robbie P. Joosten, and Jon Agirre. “Analysis and validation of overall N-glycan conformation in Privateer” (under review).

- Agirre, Jon, Mihaela Atanasova, Haroldas Bagdonas, Ben Bax, Atlanta G. Cook, James Beilstein-Edmands, Rafael J. Borges, José Javier Burgos-Mármol, Charles Ballard, John M. Berrisford, Paul S. Bond, Lucrezia Catapano, Grzegorz Chojnowsky, Kevin D. Cowtan, Tristan I. Croll, Judit É. Debreczeni, Eleanor J. Dodson, Tarik R. Drevon, Paul Emsley, Gwyndaf Evans, Phil R. Evans, Maria Fando, James Foadi, Luis Fuentes-Montero, Elspeth F. Garman, Markus Gerstel, Richard Gildea, Kaushik Hatti, Maarten L. Hekkelman, Soon Wen Hoh, Huw T. Jenkins, Robbie P. Joosten, Ronan M. Keegan, Nicholas Keep, Eugene B. Krissinel, Petr Kolenko, Oleg Kovalevskiy, Victor Lamzin, Dave Lawson, Andrey Lebedev, Andrew Leslie, Bernhard Lohkamp, Fei Long, Airlie J. McCoy, Stuart J. McNicholas, Claudia Millán, Garib N. Murshudov, Robert A. Nicholls, Martin E.M. Noble, Robert Oeffner, Navraj S. Pannu, James Parkhurst, Nicholas Pearce, Arwen Pearson, Joana Pereira, Anastassis Perrakis, Liz A. Potterton, Harry Powell, Randy J. Read, Daniel Rigden, William Rochira, Massimo Sammito, Paula Salgado, Filomeno Sánchez-Rodríguez, George Sheldrick, Kathryn Shelley, Felix Simkovic, Adam Simpkin, Pavol Skubak, Egor Sobolev, Roberto Steiner, Kyle Stevenson, Ivo Tews, Jens M.H. Thomas, Andrea Thorn, Ian Tickle, Josep Triviño, Ville Uski, Isabel Usón, Alexei Vagin, Sameer Velankar, Melanie Vollmar, Frank von Delft, David Waterman, Keith S. Wilson, Martyn Winn, Graeme Winter, Marcin Wojdyr, Keitaro Yamashita, David Brown. “The CCP4 suite: integrative software for macromolecular crystallography” (in preparation).

This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

COVID19 Pandemic Impact Statement

The COVID19 pandemic had a significant impact on my PhD work, despite the fact that I was able to continue with my research fully remotely. The transition to remote work was initially challenging, as it required adapting to a new way of collaborating with my colleagues and accessing resources. The lack of in-person interactions also made it more difficult to stay motivated and focused.

On the positive side, I also had the opportunity to attend virtual conferences and seminars that I might not have been able to attend otherwise, which broadened my exposure to different scientific perspectives.

Chapter 1

Introduction

1.1 Aims and Overview

The general aim of this thesis was to identify problems with deposited carbohydrate structures, relate those to shortcomings in existing methodology, and target certain areas to improve the software tools available for carbohydrate model building and refinement. Currently there are few options for automated model building, and they have limitations. However, protein and nucleic acid model building has been very well developed and can be used as inspiration for carbohydrate model building.

Chapter 1 is an introduction into the scientific background of this thesis. It also includes the published article “Structural glycobiology in the age of electron cryo-microscopy”. This paper contains quality analysis of the carbohydrates found in the Protein Data Bank (PDB). Most potential issues can be detected with the Privateer carbohydrate validation software, developed in our team. The structures in the PDB were shown to contain a great proportion of carbohydrates in high-energy conformation, in defiance of glyco-chemical knowledge (also in Atanasova et al., 2020). This shows the necessity of improving carbohydrate software tools.

Chapter 2 consists of the published article: “Updated restraint dictionaries for carbohydrates in the pyranose form”. The aim of the paper is to correct the errors found in multiple restraint dictionaries of carbohydrates. After generating correct ones, the dictionaries can be improved by adding the possibility of using unimodal torsion restraints to enforce sugar conformation. It is also vital to test the new dictionaries extensively before they are added to the CCP4 Monomer Library. The new dictionaries should improve carbohydrate refinement, which is necessary for the model building presented in the third chapter (also in Atanasova et al., 2022).

Chapter 3 is on automated carbohydrate model building. It is based on a method previously described by (Kevin Cowtan 2014) for use in nucleic acid model building. However, the method heavily relies on the availability of high quality re-refined structures of carbohydrates, which is why the studies outlined in Chapter 2 are necessary. Chapter 3 is currently unpublished.

Appendix A contains the supplementary information for the article presented in Chapter 2, “Updated restraint dictionaries for carbohydrates in the pyranose form”. The full names for all monosaccharides in this thesis can be found in Table A.1 in Appendix A described by their three-letter CCD component IDs

Appendix B contains the draft article “Analysis and validation of overall N-glycan conformation in Privateer”. Appendix C contains the draft article “The CCP4 suite: integrative software for macromolecular crystallography”. Both of these articles include contributions made during this PhD.

1.2 Scientific background

Determining the 3D structure of proteins has been of scientific interest for decades since it provides insight into many biological processes. There are numerous techniques used for that, but the most common ones are X-ray crystallography and electron cryo-microscopy (cryo-EM). X-ray crystallography is generally used for the solution of protein structures in the 50-100 kDa range, while cryo-EM is usually used for bigger proteins or protein complexes of ~100 kDa or more.

Both of these techniques have been used to determine multiple structures that have contributed to understanding the mechanisms of diseases such as cancer, diabetes, neurodegenerative disease and viral infection, in addition to determining structures with various biotechnological applications. For example, X-ray crystallography has been used to determine the structure of the HIV-1 protease, an enzyme involved in viral maturation. The crystal structure of the HIV-1 protease has contributed to the development of protease inhibitors, a class of drugs that has revolutionised the treatment of HIV/AIDS (Navia et al. 1989). Moreover, X-ray crystallography has improved our understanding of melanoma. Melanoma patients often have mutations that enhance BRAF kinase activity, which increases cancer cells proliferation and invasion. The structure of the BRAF kinase has been used to design small molecule inhibitors for the treatment of melanoma (Xie et al. 2009). Furthermore, determining the 3D structure of insulin gave insight into how it binds to its receptor and activates signalling pathways that regulate glucose metabolism. This allowed for the development of insulin analogues with various pharmacokinetic and pharmacodynamic properties (Blundell et al. 1971).

This chapter discusses the scientific background of X-ray crystallography and cryo-EM. Section 1.2.1 describes scattering, which is the underlying physical phenomenon of both

x-ray crystallography and cryo-EM. Section 1.2.2 discusses X-ray crystallography. An overview is given of the structure solution process, starting with data collection and ending with deposition of the protein structure model in the PDB. Section 1.2.3 discusses cryo-EM, including data collection and data processing. The large availability of protein structure models has caused a surge in attempts to use artificial intelligence methods for protein structure prediction from sequence data. Section 1.2.4 presents the most successful one, called AlphaFold2.

However, despite the developments in protein structure determination techniques, the structural methods for studying post-translational modifications still have a long way to go. The interest in carbohydrates and glycosylation has been growing steadily, as there is now evidence that they take part in just about every biological process in the human body and many others. As such, the number of carbohydrate-containing structures has been increasing in the PDB too. Carbohydrate structures can be determined with both X-ray crystallography and cryo-EM and often the same software tools can be used with both techniques. A brief overview of carbohydrates and their biological roles is given in Section 1.2.5. The software tools specific to carbohydrates are discussed in Section 1.2.6.

1.2.1 Scattering

Scattering is a physical phenomenon that describes the interaction of electromagnetic radiation with matter and is the basis of multiple biophysical techniques. When a beam, such as X-rays or electrons, is directed towards a sample it interacts with the atoms in the sample and is scattered in a random direction. The scattering can be detected and analysed to gain insights into the sample. Scattering can be coherent, which means that the scattered wave has the same phase as the incident wave, or incoherent - where the phase is different. Scattering can also be elastic, where the wavelength and kinetic energy of the wave remain the same, or inelastic - where they are different. There are different combinations of the types of scattering possible, eg. elastic coherent, inelastic coherent, etc. and they all have various technological applications. The techniques described here will focus mainly on elastic and coherent scattering.

1.2.2 X-ray crystallography

X-ray crystallography is an experimental technique that uses the diffraction of X-rays from atoms in a sample to determine the structure of a molecule. It can be used on small molecules, as well as large biological molecules like proteins and nucleic acids. It has contributed to the solution of close to 170,000 protein structures. This gives insight into biological mechanisms and has led to the development of pharmaceuticals, as well as multiple biotechnological applications. In order for this technique to be used on molecules, the sample needs to be produced in the form of a single crystal. Ideally, the crystal structure presents multiple perfectly ordered copies of the molecule, which can serve to amplify the signal and thus give better information about the structure. When the waves are scattered by atoms in adjacent crystal planes of the lattice, they interfere in phase which results in diffraction. The molecule is also immobilised, which allows for a high resolution snapshot to be obtained. In practice, crystal order is limited by several factors, such as crystal defects, thermal motion of molecules and residual solvent. These lead to a reduced long-range order in the crystal, which decreases the diffraction pattern's resolution. Also, having multiple crystals attached to each other leads to noisy and difficult to interpret data.

When an X-ray beam is directed towards a crystal, the photons can excite the electrons into vibrating with the frequency of the X-ray. The excited electrons then release the photon causing coherent elastic scattering. The scattered waves from the multiple perfectly ordered copies of the molecule interfere constructively or destructively, which is described by Bragg's Law (1) (Brindley 1933).

$$n\lambda = 2d\,sin\theta \quad (1)$$

where λ is the wavelength of the X-ray, n is a positive integer, d is the spacing between the crystal planes, θ is the incident angle. This results in a diffraction pattern, which can be detected and analysed. However, there is also incoherent inelastic scattering which causes damage to the molecule. This could lead to the atom becoming ionised or excited, which in turn can lead to the creation of very reactive free radicals. Having many copies of the molecule in a crystal ensures that there is still enough elastic coherent scattering to compensate for the damage. The diffraction pattern produced is in reciprocal space, as X-rays cannot be focused through a lens to invert them into real space.

1.2.2.1 Data collection

X-rays are normally produced by accelerating electrons towards a metal (usually copper or molybdenum) in a vacuum. The wavelengths used in crystallography are normally between 0.5 and 1.6 Å. This is commensurate with bond lengths, so it can be used to study molecules at resolutions where the individual atoms can be seen, typically with the exception of hydrogens. Commonly, data are collected at synchrotron facilities, which produce very intense X-rays, thus allowing for better data resolution. In synchrotrons, an electron beam is accelerated around a circular track in a vacuum at over 99% of the speed of light. The electron bunches are directed through magnetic fields, such as bending magnets or multipole wigglers, which cause the electrons to change their angular momentum. This change in angular momentum results in the emission of X-rays. However, once the X-rays have been generated, they cannot be directly manipulated by magnetic fields. The X-ray spectrum produced is continuous, therefore the whole range between 0.5 and 1.6 Å can be used for experiments, giving rise to a range of X-ray techniques ranging from small molecular and macromolecular crystallography, collection of small-angle scattering, and microscopy.

1.2.2.2 Data processing

While X-rays and visible light both propagate in space in the same way because they are electromagnetic waves, X-rays interact with matter quite differently due to their much higher frequency and photon intensity. Unlike visible light, which can be easily redirected using lenses and mirrors, X-rays tend to initially enter and eventually get absorbed in most materials without altering direction substantially. These beams produce a diffraction pattern of dots when they touch a piece of film or another detector.

From the diffraction pattern, through the use of numerous software tools, an electron density map of the molecule can be produced, and from there - a structural model. The two major software suites for macromolecular crystallography are CCP4 (Winn et al. 2011a) and Phenix (Liebschner et al. 2019). Each of these provides tools for the entire structure solution process - from image processing, indexing, integration, scaling, density modification, through building the molecular model, to refining and validating it.

In order to start processing the data, first, it is vital to determine the crystal symmetry. The symmetry describes the way in which the molecules are packed in the crystal. It is defined by the space group, which is the complete set of symmetry operations, eg. translations, in a

crystal lattice. The resulting model must fit the determined symmetry, which must fit the scattering. To determine the symmetry, each diffraction spot is indexed so that the Miller indices and the unit cell can be determined.

While having correct symmetry is important for indexing X-ray diffraction images, errors in indexing can still lead to inaccurate location of diffraction spots during integration. To obtain accurate values, least-squares refinement is used, with two commonly used approaches. The first, called positional refinement, optimises the fit between observed and calculated spot coordinates on the image and is used in all integration programs. Positional refinement can be done iteratively by comparing the experimental reflections with calculated reflections. The second approach is post-refinement, which analyses the relative intensity of partially recorded reflections across multiple images. Post-refinement can only be performed after the intensities of the spots have been measured.

After refinement, each pixel, part of a reflection, is integrated. Integration is the process of measuring the intensities of the diffraction spots. Integration can be done using two basic methods: summation integration and profile fitting. Summation integration is when the intensity is simply obtained by adding together the values for each pixel that forms a spot on the image. However, accurately locating all the spots, identifying which pixels correspond to which spots and accounting for non-zero background on the image due to detector noise can be challenging and requires correct indexing. Profile fitting involves modelling the diffraction profile of each reflection using a mathematical function, such as a Gaussian function, to describe the intensity of the spot (Enzo et al. 1988). Also, the background noise is removed which gives a clearer data set that simplifies the determination of the unit cell and from there - the crystal symmetry.

The data obtained from integration programs are not uniform in scale due to various physical errors. These factors include changes in incident radiation intensity, the volume of the crystal being illuminated, anisotropic absorption of X-rays, radiation damage to the crystal and non-uniformity of the detector response. To address these issues, scaling and merging are necessary to model the changes that occur during the diffraction experiment (Powell 2017). After scaling and merging, the amplitudes are calculated.

1.2.2.3 Phasing

X-ray detectors measure the intensity of the scattered photons, but they cannot measure the phase. This is referred to as *the phase problem*. There are experimental and theoretical

ways of solving it. The Patterson function can be used to model small molecule structures by avoiding the phase problem. It uses the intensities to calculate the Patterson density, which can then be used to elucidate the structure. For bigger molecules, such as proteins, the phasing process normally starts with Molecular Replacement (MR). This is a technique that uses a comparison between the molecule to be modelled with a molecule for which the structure is already available. Current MR software allows for structures with as little as 30% sequence similarity to be used as templates. Resolution of both the template and the data is also important for MR. When MR attempts are unsuccessful, an experimental approach needs to be considered. Usually a technique based on anomalous scattering is used. It involves finding changes to the diffraction pattern caused by heavy atoms present in the structure. The X-ray needs to be close in energy to the absorption edge of the heavy atom, ie. the energy needed to promote a core electron. According to Friedel's law, opposite reflections have the same amplitude, but opposite phases. In anomalous scattering, however, the phase is at 90° to normal scattering. This creates a difference in amplitude that can be detected with fluorescence and used to find the positions of heavy atoms in a structure. Commonly used software for both molecular replacement and experimental phasing is Phaser (McCoy et al. 2007) and SHELXE (Sheldrick 2008b; Thorn and Sheldrick 2013).

1.2.2.4 Density modification

Often phasing does not provide a perfectly clear electron density map. In such cases it is possible to use density modification techniques to avoid having to collect further experimental data. For example, solvent flattening is used to "remove" the parts of the data that represent the solvent. An "improved" electron-density map is then calculated for the protein (Wang 1985). Another routinely used density modification approach is histogram matching. It uses a histogram of the electron density values of a protein and compares it to the predicted values. Differences between the two histograms are attributed to errors during phasing. However, this approach can only be used on proteins, but not on nucleic acids (K. Y. Zhang, Cowtan, and Main 1997). These are the two most commonly used approaches, but there are many others for more complicated cases. Commonly used software for density modification is PARROT (Kevin Cowtan 2010b).

1.2.2.5 Model Building

Once the amplitudes and the phases have been calculated, the full electron density map can be analysed. The next step is to build a model consistent with the data and the symmetry determined earlier. When the resolution is good, it is generally straightforward to do this. For lower resolution, it is necessary to use chemical and biological knowledge to build a sensible protein structure. There are a number of programs (BUCCANEER (Kevin Cowtan 2006, 2012), phenix.autobuild (Terwilliger et al. 2007), ARP/wARP (Lamzin, Perrakis, and Wilson 2012b; Chojnowski, Pereira, and Lamzin 2019)) and pipelines (ModelCraft (Bond and Cowtan 2022)) available and when it comes to proteins, this step has been highly automated.

1.2.2.6 Refinement

The purpose of refinement is to improve the model so it fits the experimental data as close as possible. This is measured by the R-factor which is a statistical measure that assesses the agreement between observed and calculated X-ray diffraction data. It is calculated as the ratio of the sum of the difference between the observed and calculated intensity to the sum of the observed intensity (2):

$$R = \frac{\Sigma\left|\left|F_{obs}\right| - \left|F_{calc}\right|\right|}{\Sigma\left|F_{obs}\right|} \quad (2)$$

where F_{obs} is the observed structure factor, F_{calc} is the calculated structure factor; the structure factor is related to the intensity of the reflection it represents.

During refinement, the aim is to minimise the difference between the observed experimental data and the calculated values from the model, ie. to lower the R-factor. This is typically achieved through a process called least-squares refinement, where the model parameters are adjusted to minimise the sum of the squared differences between the observed and the calculated data.

Refinement is done in a series of iterative cycles in which the model is refined and the R-factor is re-calculated. There are multiple software tools for refinement with a large number of settings and parameters. Two of the most commonly used refinement programs are REFMAC5 (Murshudov et al. 2011) and phenix.refine (Afonine et al. 2012). The software tools often combine model building and refinement iteratively. At each step, an electron density map is calculated from the model and compared to the experimental map. This

process is repeated until the two maps are sufficiently similar. Refinement is discussed in more detail in Chapter 2.

1.2.2.7 Carbohydrate model building

Ligands, glycosylation and other post-translational modifications are normally built after the main protein is built and refined. It is generally more difficult to build carbohydrates which are part of glycosylation and doing this on a mostly complete model provides much fewer possibilities for errors. The model building tools used for carbohydrates are not as featureful or as automated as the ones for protein discussed above. However, when it comes to refinement, the same tools can generally be used on both the protein backbone and the carbohydrates.

1.2.2.8 Validation

After a model has been built and refined, it generally goes through validation. Model validation is the process of evaluating the quality and accuracy of a model to make sure it is biologically and chemically meaningful and to identify potential errors or inconsistencies. This step uses suites, such as MolProbity (Williams et al. 2018), to identify amino acids that do not fit well the electron density or have geometrical errors, clashes and rotamer outliers. These can then be rebuilt if necessary until the quality is sufficient. MolProbity also provides useful statistics, such as Ramachandran plot analysis. Another useful tool is EDSTATS (Tickle 2012) which is used for analysing the quality of electron density maps calculated from experimental data. PISA (Evgeny Krissinel and Henrick 2007) is a tool for analysing the quaternary structure of proteins. It can be used to identify protein-protein interfaces and to calculate the buried surface area of protein complexes. PISA can also be used to identify potential errors or inconsistencies in the quaternary structure of the protein. RABDAM (Shelley et al. 2018) is a tool that estimates the degree of radiation damage to the protein crystal. It is used to estimate the quality of the data. It is important to use both validation metrics and visual inspection to ensure that the model is as correct as possible. For this reason, the model is often visualised in a molecular graphics program, such as Coot, which allows for the structure to be visually examined, as well as improved using a range of built-in refinement and validation features. In the case of carbohydrates, the Privateer software (Agirre, Iglesias-Fernández, et al. 2015b) provides a simple and straightforward way to validate carbohydrates. This is discussed in more detail in Section 1.2.6.

After the model is completed to a sufficient standard, it is uploaded to the PDB (Burley et al. 2019), along with some of the experimental data and the validation metrics.

1.2.3 Electron cryo-microscopy (cryo-EM)

An electron microscope uses a beam of electrons, which is re-focused using magnetic lenses and is scattered by the sample. The sample can be rotated so that multiple images can be collected at various angles. The electrons are scattered by the sample in two ways - elastic and inelastic. In transmission electron microscopy (TEM) the electron beam passes through the sample, the electrons are scattered and detected to create a projection image of the sample in real space. The images are projection images, rather than just 2D images, as they are obtained by projecting a 3D specimen onto a flat image plane. The resulting image represents a cross-section of the specimen. In TEM elastic scattering is used to study the sample, while inelastic scattering is avoided, because the released energy is released when electrons from the beam clash into electrons from the sample and cause them to change orbitals. This causes damage to the sample.

Cryo-EM is a type of TEM which involves vitrification of a biological sample by cooling it to below 140 K and performing the entire data collection process at that temperature. This is the temperature at which water becomes vitreous, ie. the sample is preserved in its native state (Dubochet et al. 1988) and is protected from radiation damage. This differs from X-ray crystallography, where the sample needs to be in the form of a pure and highly-ordered crystal. This is a major advantage of cryo-EM, as growing crystals can take time and is often impossible for large biological complexes or disordered proteins. However, cryo-EM sample preparation can still be challenging, as when the sample is mounted on a grid, this poses the risk of degradation. The sample needs to be thin, because this minimises noise from inelastic scattering. In addition, the resolution achieved with cryo-EM is in general lower than that of X-ray crystallography, because of the wavelength of electrons (2-3 pm for typical accelerating voltages of 200-300 kV in a TEM) and also because exposure of the sample to the electron beam often needs to be limited to minimise inelastic scattering. Furthermore, cryo-EM is overall very expensive, considering the cost of a microscope, the cost per run and the specialised laboratories required to house it.

1.2.3.1 Data processing

As mentioned above, cryo-EM involves taking 2-D images of specimens from different angles. The model is then reconstructed from these images. This creates a major challenge of data processing, as a data set can consist of thousands of images. The images are divided into classes by image type, then each class is averaged. The final model of the 3-D electron potential map is constructed from the average of each class. The construction of the 3-D map model is based on the Fourier slice theorem, which states that from a group of 2-D images at various angles, a 3-D image can be constructed. This is done iteratively until the map does not change. The model can be improved using a maximum-likelihood approach which aims to maximise the probability of the angle of an image of the particle being the correct one relative to the final model.

This stage has been greatly facilitated through the use of various software tools, such as the image processing software RELION (Scheres 2012), which has also been integrated with the CCP-EM suite, the CCP4 equivalent for cryo-EM. RELION uses an empirical Bayesian approach to build the model and a range of statistical methods to assess whether the particle position has been determined correctly. Furthermore, artificial intelligence approaches are also being developed for use in image reconstruction (Punjani and Fleet 2021; Zhong et al. 2021).

Once an electron density map has been obtained, the rest of the structure solution process does not differ vastly from the structure solution process of X-ray crystallography. Many of the software tools used in crystallography can also be used in cryo-EM. This includes, BUCCANEER, phenix.autobuild, REFMAC5, phenix.refine, Coot, Privateer, etc. In fact some of these have been integrated into CCP-EM. However, the lower resolution necessitates the use of pipelines specifically designed for cryo-EM (Hoh, Burnley, and Cowtan 2020; Liebschner et al. 2019). Also, it is important to note that refinement software used for crystallography normally uses reciprocal space refinement. Cryo-EM maps can be converted into Fourrier coefficients and hence refined in the reciprocal space just like X-ray crystallography models (Afonine et al. 2018).

X-ray data can often be used to generate an atomic-resolution 3D model, as opposed to cryo-EM data, for which this is currently uncommon. However, X-ray maps often result in an average representation that may not accurately reflect conformational heterogeneity. Cryo-EM maps, on the other hand, sometimes only provide information on the overall shape, internal structure and relative position of the macromolecules as they typically have lower resolution compared to X-ray maps. Nevertheless, cryo-EM data has the potential to model conformational heterogeneity, as it can average information from many different molecules, each of which may have a slightly different conformation. This can lead to a more accurate

representation of the protein's conformational variability, especially in cases where the protein has intrinsic flexibility or exists in multiple conformations.

In the case of carbohydrates, cryo-EM has the additional advantage of being able to study protein complexes with the carbohydrates, without them potentially interfering with the crystallisation process. The lower resolution is a hurdle, especially since carbohydrates tend to have even lower resolution than the rest of the protein. However, the restraint dictionaries in the CCP4 Monomer Library, including the carbohydrate ones discussed in Chapter 2, are used in CCP-EM. These offer a solution to building and refining carbohydrates at low resolution, as they offer the option of using external restraints to enforce the correct structure.

1.2.4 AlphaFold2

With the advancements in computing and artificial intelligence, there have been efforts to predict protein structures. Currently, the most successful tool for this is AlphaFold2 (Jumper et al. 2021), which can predict protein structures with over 80% accuracy. It involves a complex neural network that uses genetic sequence alignment, pairing of amino acids in proximity and protein templates. AlphaFold2 also outputs various metrics to indicate how good the predicted model is. AlphaFold2 has been used to predict the structures of the entire human proteome (Tunyasuvunakool et al. 2021), which has been deposited into a database (Varadi et al. 2022). Predicted models can be used for molecular replacement, essentially solving the phase problem. Moreover, the predicted model scoring is a good indicator of how disordered a protein is. Furthermore, AlphaFold2 predicted models can be used when building a structure at low resolution, which is often the case with cryo-EM. However, AlphaFold2 does not predict ligands and post-translational modifications, including carbohydrates and glycosylation, although it is often possible to see the space in the model where the ligands or glycans should be built.

Overall, the impact of AlphaFold2 has been significant. Its success has demonstrated the potential of artificial intelligence approaches for solving complex biological problems and has paved the way for further developments in the field. It has increased the speed and accuracy of protein structure prediction, which has a wide range of applications in drug discovery and biotechnology. It has led to the development of machine learning models to predict protein-protein interactions, ligand binding sites, protein stability, etc.

AlphaFold2, like all machine learning models, has the potential to be biased, ie. its predictions may be influenced by systematic errors that are not based on the underlying biology. In particular, the potential sources of bias include the training data used to develop the model and the way AlphaFold2 processes the data. If the training data are not representative of the full range of protein structures, the model may be biased towards certain types of structures. Additionally, if the training data contain systematic errors, such as incorrect protein structures, these errors may be reflected in the model's predictions. In order to minimise potential bias, training data have been carefully curated and AlphaFold2 has been subjected to rigorous testing and validation to assess its accuracy.

1.2.5 Carbohydrates

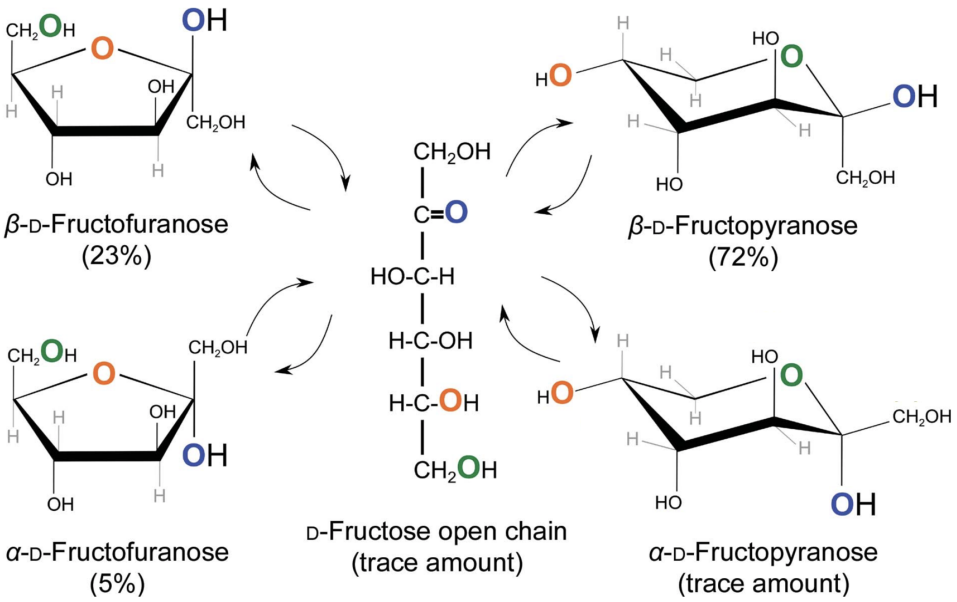


Figure 1.1. D-Fructose in linear and cyclic forms. The oxygen shown in orange performs a nucleophilic attack on the ketone to form the furanose forms shown on the left. The oxygen shown in green does the same to form the pyranose forms shown on the right. As the ketone group is planar, the nucleophilic attack can happen from either above or below. This allows for the formation of the different anomeric forms. Adapted from (Agirre 2017).

Monosaccharides generally have the formula $C_x(H_2O)_n$, where *n* is an integer in the range 3-9, and also have an aldehyde or ketone group. Monosaccharides are formed as a chain of chiral hydroxymethylene groups with an aldehyde or an α-hydroxy ketone group at the end. These chiral groups give rise to a number of stereoisomeric forms. Monosaccharides are described as D or L depending on the configuration of the stereogenic centre furthest away from the aldehyde or ketone group. In Fisher projection (seen in Figure 1.1, centre, for fructose), if the OH group is on the left, the monosaccharide has L configuration, if the OH group is on the right - D configuration. Monosaccharides in D configuration are more commonly found in nature. Monosaccharides which only differ from each other by the configuration of a single chiral carbon are called epimers.

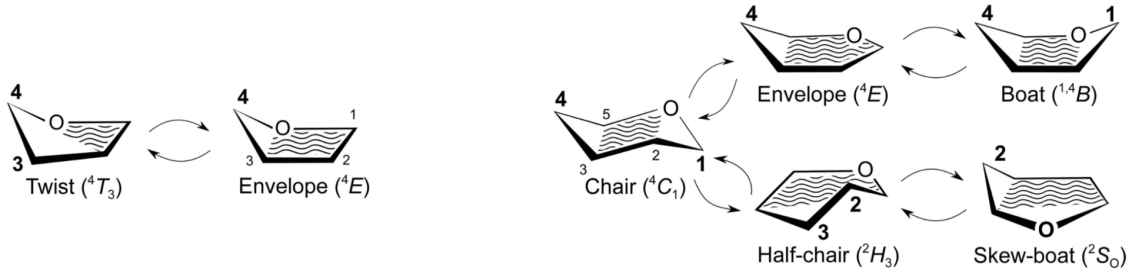


Figure 1.2. Possible conformation itineraries for furanoses (left) and pyranoses (right). Furanoses always exhibit some strain, while pyranoses can have very low energy when in chair form. Adapted from (Agirre 2017).

In solution, 5 or 6-carbon sugars usually exist as a mixture of linear, 5-membered ring (furanose) and 6-membered ring (pyranose) forms (Figure 1.2). Both pyranoses and furanoses can adopt multiple conformations. For pyranoses, the most stable conformations are $^1C_4$ and $^4C_1$ chair. For furanoses, conformation is more complex as there are multiple lower-energy conformations, but they all exhibit some degree of flexibility and puckering and hence are not as stable as chairs (Figure 1.2). In the cyclic form, monosaccharides have another asymmetric centre at the carbonyl carbon atom, which is known as the anomeric carbon. The group at that carbon can only point in two directions - equatorial or axial. The anomeric configuration of a monosaccharides is determined by whether the configuration at the anomeric carbon is the same as the configuration at the chiral centre furthest away from the anomeric carbon. If the two configurations are the same, the anomer is α, if they are different - β. The conversion from one anomeric form to another in solution is called mutarotation. Anomeric configurations further expand the diversity of monosaccharides.

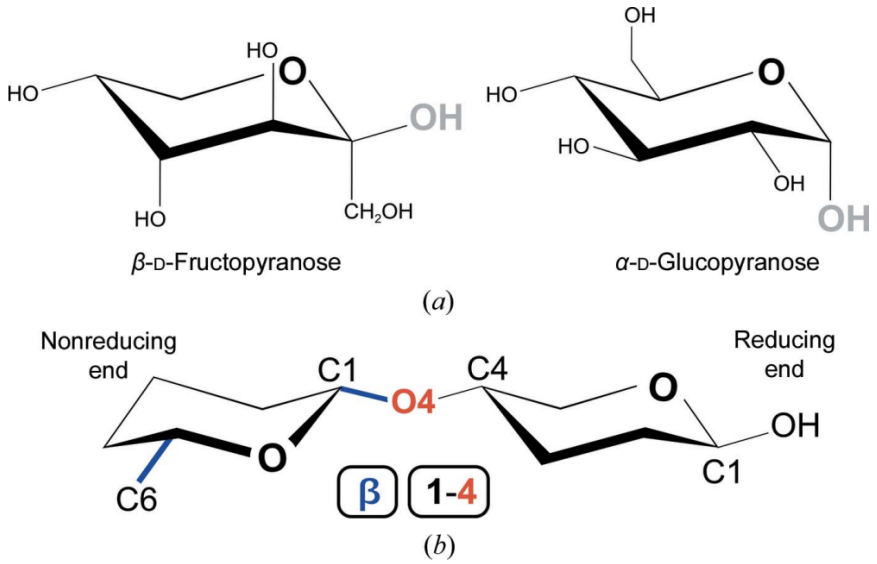


Figure 1.3. Example of a glycosidic bond between a β-D-fructopyranose and an α-D-glucopyranose. (a) The leaving group on the reducing end is shown in grey. (b). The bond between the anomeric carbon with β configuration and the hydroxyl group on the fourth carbon (C4) is called a β1-4 bond. Reproduced from (Agirre 2017).

Monosaccharides are linked into oligosaccharides (usually up to 10-12 monosaccharides) and polysaccharides (usually more than 10-12 monosaccharides) via glycosidic bonds. This is a bond between the anomeric carbon, also known as the reducing end, of one monosaccharide with one of the hydroxyl groups of another, known as the non-reducing end. The reaction is acetal formation from a hemi-acetal on the first monosaccharides (involving the anomeric carbon) and an alcohol (on the second monosaccharides). Just like with monosaccharides, the glycosidic bond has an anomeric configuration which once formed remains the same. There are, however, multiple regioisomers possible, due to the number of hydroxyl groups available. In fact, a monosaccharides can have multiple other monosaccharides attached to different hydroxyl groups and thus can serve as a branching point. The nomenclature for glycosidic bonds includes the anomeric configuration, the number of the first carbon and the number of the second carbon as shown in Figure 1.3. In addition, the glycosidic bond is quite flexible with three torsion angles. These factors further increase the complexity of carbohydrates and complicate studying them.

Sugars can be linked to other biological molecules, such as proteins, lipids and they are part of nucleic acid bases. They are also involved in one of the most common post-translational modifications - glycosylation. There are three major types of glycosylation: N-glycosylation, which involves covalently-linked carbohydrates to the N atom of an asparagine (Asn), O-glycosylation, where the carbohydrate is attached to the O atom of a Thr (via acetal

formation as described above for monosaccharides) and glypiation, where glycophosphatidylinositol is added to a protein to serve as a membrane anchor. Glycosylation is primarily found in eukaryotes, where it plays a crucial role in a variety of biological processes, such as protein folding, stability, trafficking and signalling. However, there are some notable examples of glycosylation in certain bacterial and archaeal species, which often involve simpler glycan structures and use distinct biosynthetic pathways compared to those in eukaryotes.

1.2.5.1 N-glycosylation

N-glycans are often attached to the amide nitrogen of an Asn part of a Asn-X-Ser/Thr sequon, where X is any amino acid, except for Pro. The oligosaccharide chain normally starts with 2 GlcNAc monosaccharides. N-glycosylation involves multiple complex steps. This can lead to having different glycoforms attached to different copies of the same protein, which further complicates studying them. First, an oligosaccharide precursor $Man_5GlcNAc_2$ is assembled on a dolichylphosphate on the cytosolic surface of the ER. It is then flipped to the lumenal side of the membrane and extended to $Glc_3Man_9GlcNAc_2$. After that, the oligosaccharide is transferred to the Asn by an oligosaccharyltransferase at the endoplasmic reticulum (ER). Next, three of the glucoses are trimmed. After ensuring that the protein is correctly folded, further trimming is done in the ER to $Man_8GlcNAc_2$. This is then transferred to the Golgi apparatus and further trimmed to $Man_5GlcNAc_2$. Afterwards, this glycoform can be directly extended or trimmed further and then extended to form a diverse range of glycans.

Glycosylation can have a major effect on a protein's properties, such as folding kinetics, thermodynamic stability, etc. In addition, glycoproteins are involved in diverse biological processes, such as cell signalling, host-pathogen interactions, immunity, cell differentiation, cancer metastasis, protein degradation signalling. Deletion of the *Mgat1* gene was found to interfere with glycan synthesis in mice. This results in death during embryonic development (Stanley 2016). In cancer cells, N-glycans are often longer and more branched. This is a result of hypersialation by sialyltransferases, which are upregulated in cancer. The therapeutic potential of the inhibition of sialyltransferases is currently investigated (Dobie and Skropeta 2021). Glycans are also the underlying features in the ABO blood group categorization, which needs to be considered for successful blood transfusions (Lee-Sundlov, Stowell, and Hoffmeister 2020). N-glycans can also be involved in thyroid hormone production (Ząbczyńska, Kozłowska, and Pocheć 2018). Missing triantennary

N-glycans in mice lead them to develop type II diabetes when fed a high-fat diet (Rudman, Gornik, and Lauc 2019). Also, pathogens have evolved to target the sialic acids on the surfaces of cells. This has been exploited in the development of therapeutics, for example competitive inhibitors of viral sialidases can hinder infection with certain strains of influenza (Glanz et al. 2018).

1.2.6 Published Article: "Structural glycobiology in the age of electron cryo-microscopy"

Carbohydrate structure solution has been challenging because of the many specific issues outlined above. This has led to errors in nomenclature (Lütteke and von der Lieth 2009), conformation (Agirre, Davies, et al. 2015b), monosaccharide ring torsions, glycosidic linkage stereochemistry (Crispin, Stuart, and Jones 2007) and torsions (M. Frank, Lütteke, and von der Lieth 2007; Agirre et al. 2017b). The software tools addressing these issues have been described in this section. Moreover, a study was carried out to detect conformation errors in monosaccharides part of N-glycans in the PDB. The content below is taken from "Structural glycobiology in the age of electron cryo-microscopy" by Atanasova *et al* (Atanasova, Bagdonas, and Agirre 2020).

1.2.6.1 Introduction

Protein glycosylation plays a crucial role in recognition processes in e.g. viral infection, cancer, fertilisation, immunity and inflammation (Schnaar 2016). In this role, glycans are expected to provide stabilising contacts within the buried surface of a glycoprotein, while additionally playing a role as interaction partners on the surface, via hydrogen bonds or CH-π interactions. As independent entities, carbohydrates also have promising biotechnological applications, being a staple in the production of more eco-friendly second-generation biofuels from previously untractable crop waste. Assisting in this task, carbohydrate-active enzymes recognise, transfer and cut saccharide building blocks, often distorting individual rings to achieve catalysis.

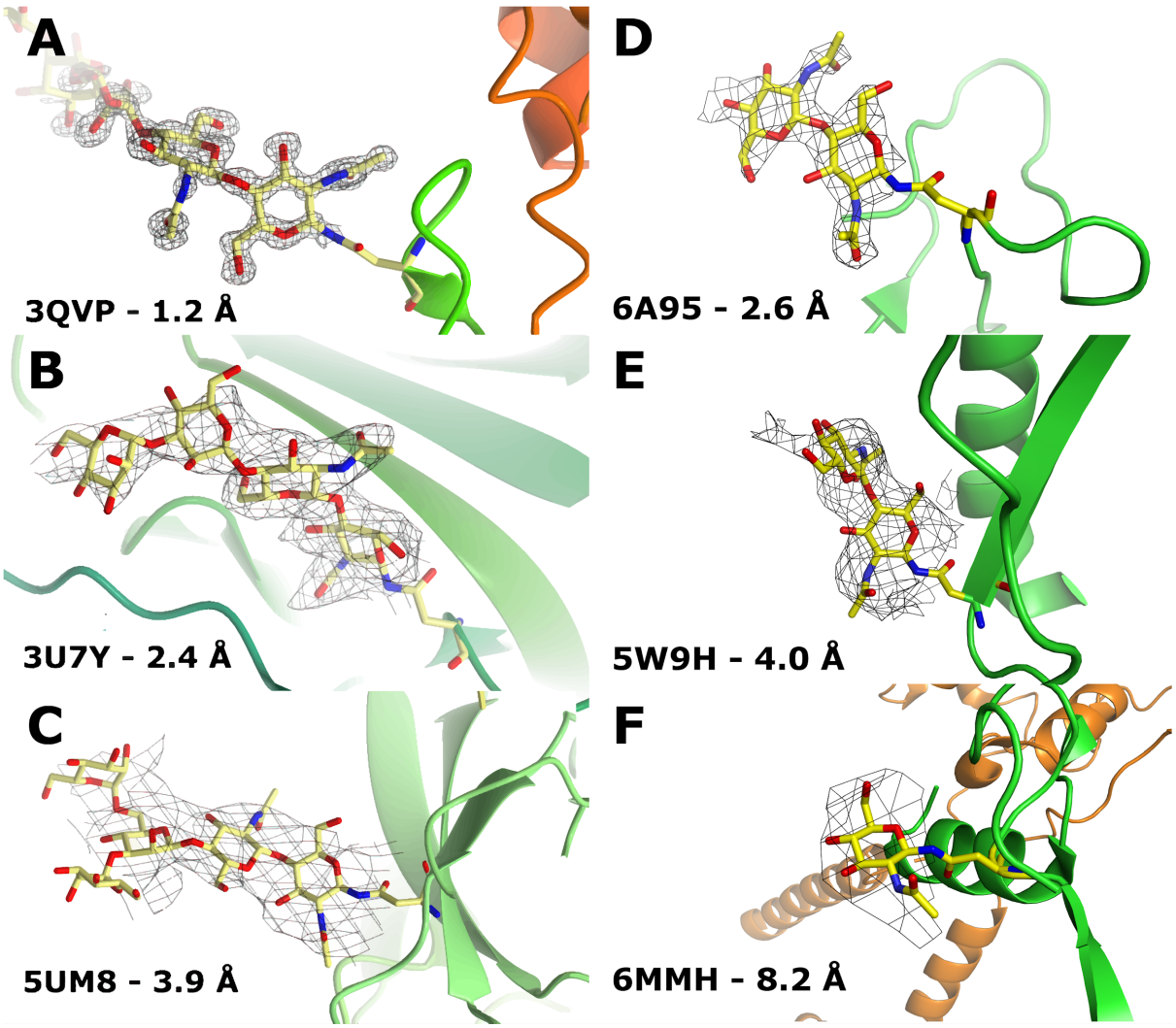


Figure 1.4. Comparison of glycan features in electron density maps over a range of resolutions. (A-C) Electron Density maps obtained with X-Ray crystallography, (D-F) Electronic potential maps obtained with cryo-EM; PDB codes and data resolution have been annotated directly on the figure. In the MX cases (A-C), at high resolution it is possible to identify monosaccharides and their ring conformation from the density map; at medium resolution, ring conformation becomes difficult to determine, whereas at low resolution, and indeed with many cryo-EM maps (D-F), a modelled glycan should always be backed by prior glyco-chemical knowledge. X-ray map types are 2mFo-DFc, the maps were generated with Privateer; sigma levels: (A) 1.5 σ; (B) 1.0 σ; (C) 1.0 σ. Cryo-EM contour levels as suggested by depositors: (D) 0.08; (E) 0.015; (F) 5.0.

Complicated stereochemistry, branching and unpredictable sequence/structure make protein glycosylation in particular harder to work with than pure protein, or even nucleic acid. Perhaps unsurprisingly, the software for handling structures of carbohydrate moieties is not yet as featureful as that for other biomolecules. This gap in capabilities becomes evident in both macromolecular crystallography (MX) and electron cryo-microscopy (cryo-EM)

whenever the model fitting problem deviates from standard propositions. Indeed, at high-resolution it is possible to identify a monosaccharide and ascertain its ring conformation (Figure 1.4). As resolution decreases, it becomes increasingly difficult to determine its ring conformation - thus requiring additional restraints for idealising ring puckering (Agirre, Davies, et al. 2015b). Finally, at low resolution, usually neither the monosaccharide nor its conformation can be identified. It is in this particular case where the articulation of prior glyco-chemical knowledge must cross boundaries from the realm of validation and play a central role in the structure building process.

Experimentally, it is clear that the mobility of the glycans poses a problem for both MX and cryo-EM, with Nuclear Magnetic Resonance (NMR) providing much of the insight into protein-carbohydrate interactions due to the degrading resolvability of the sugars down the glycans' branches (Valverde et al. 2019) typically found with the two other techniques. On the other hand, most of the challenges present in software spring from the particularities of carbohydrate chemistry. Upon cyclisation, there are two choices for the orientation of the anomeric hydroxy group, which leads to two anomeric forms – α or β (refer to (Agirre 2017) for a graphical description). Most D-sugar pyranoses adopt the $^4C_1$ conformation, while most L-sugar pyranoses adopt the $^1C_4$ conformation. Interconversion of pyranose rings between different conformations requires an itinerary, which can be described using the Cremer-Pople sphere (Cremer and Pople 1975a). The two chair conformations, $^4C_1$ and $^1C_4$ are optimal because of the 60/-60 degree torsion angle between substituents, leaving them staggered instead of eclipsed. Conversion from $^4C_1$ to $^1C_4$ and vice versa requires jumping over a very high energy barrier, and normally would involve catalysis, which can be achieved with the help of a carbohydrate active enzyme (Agirre 2017; Varki et al. 2015).

Carbohydrate residue nomenclature is challenging for several reasons, including the two different types of glycosidic linkages (α or β), branching and ring contortions. Lutteke et al., 2004 (Lütteke, Frank, and Von Der Lieth 2004) first reported that about 30% of the deposited carbohydrate structures contain one or more nomenclature errors, a finding that gave rise to carbohydrate validation software, recently reviewed in (van Beusekom, Lütteke, and Joosten 2018; Agirre et al. 2017b). A few years later, Crispin et al. also criticised the lack of methodological support for carbohydrates, singling out a deposited structure with a glycosidic linkage for which there were no available glycosyltransferases along its biosynthetic pathway (Crispin, Stuart, and Jones 2007; Helen M. Berman et al. 2007). More recently, Agirre et al. (Agirre, Davies, et al. 2015b) performed an analysis on all N-glycan forming D-pyranosides found in the PDB using the Privateer software (CCP4 suite (Winn et al. 2011a)): as data resolution decreases, more and more sugar monomers appear in

high-energy conformations and/or have low real-space correlation. This indicated the need for using appropriate restraints during refinement.

In this review, we shall go through the latest software developments and their application to solving real-world structures, placing an emphasis on their impact on the recent evolution of electron cryo-microscopy into an all-around player in the structural glycobiology field. Aside from the growing access to automated, integrative model building and validation tools, a number of online support resources are available to the structural glycobiologist too: see (Yuriev and Ramsland 2015; Paul Emsley, Brunger, and Lütteke 2015) for a review of online resources, and Perez and De Sanctis (Pérez and de Sanctis 2017) for a recent summary of the resources and techniques available where a synchrotron light source is available.

1.2.6.2 Dictionaries: the book of chemical knowledge

The model building process involves macromolecular refinement programs deriving geometric restraints from libraries of dictionaries, at least for most commonly occurring monomers. Dictionaries are used to store prior chemical knowledge about compounds, including their composition, connectivity and stereochemistry. The CCP4 Monomer Library, one of the first examples of its kind, was based on the geometry proposed by Engh and Huber (Engh and Huber 1991), which is now outdated particularly concerning sugars (Agirre 2017). If a chemical compound does not have a library entry, or if it is incorrect, a new one needs to be generated. There are several programs that can be used for this, with irregular results for carbohydrates (Agirre 2017). The CCP4 program ACEDRG (Long et al. 2017a, [b] 2017) works by mining databases such as the Crystallography Open Database (COD) (Long et al. 2017b) to generate dictionaries from the data available there. It then uses RDKit (open source chemoinformatics; http://www.rdkit.org) to generate conformers which are ranked by free energy, and the minimal-energy one is chosen. ACEDRG/COD produces similar results to GRADE (Global Phasing Ltd.) and Phenix.eLBOW (Moriarty, Grosse-Kunstleve, and Adams 2009), which derive their restraints from Mogul (Bruno et al. 2004), a tool that in turn mines the Cambridge Structural Database (CSD). Mogul is currently in use for geometry validation upon deposition with the Protein Data Bank, meaning that the use of old dictionaries during refinement with tight geometry targets – e.g. when refining against a cryo-EM map – can produce a disproportionate number of bond length and angle outliers. A modernisation effort has been carried out in CCP4, with hundreds of carbohydrate entries updated through the combination of ACEDRG and Privateer (Agirre, Iglesias-Fernández, et al. 2015b). The new dictionaries were released in 2022.

1.2.6.3 Model building

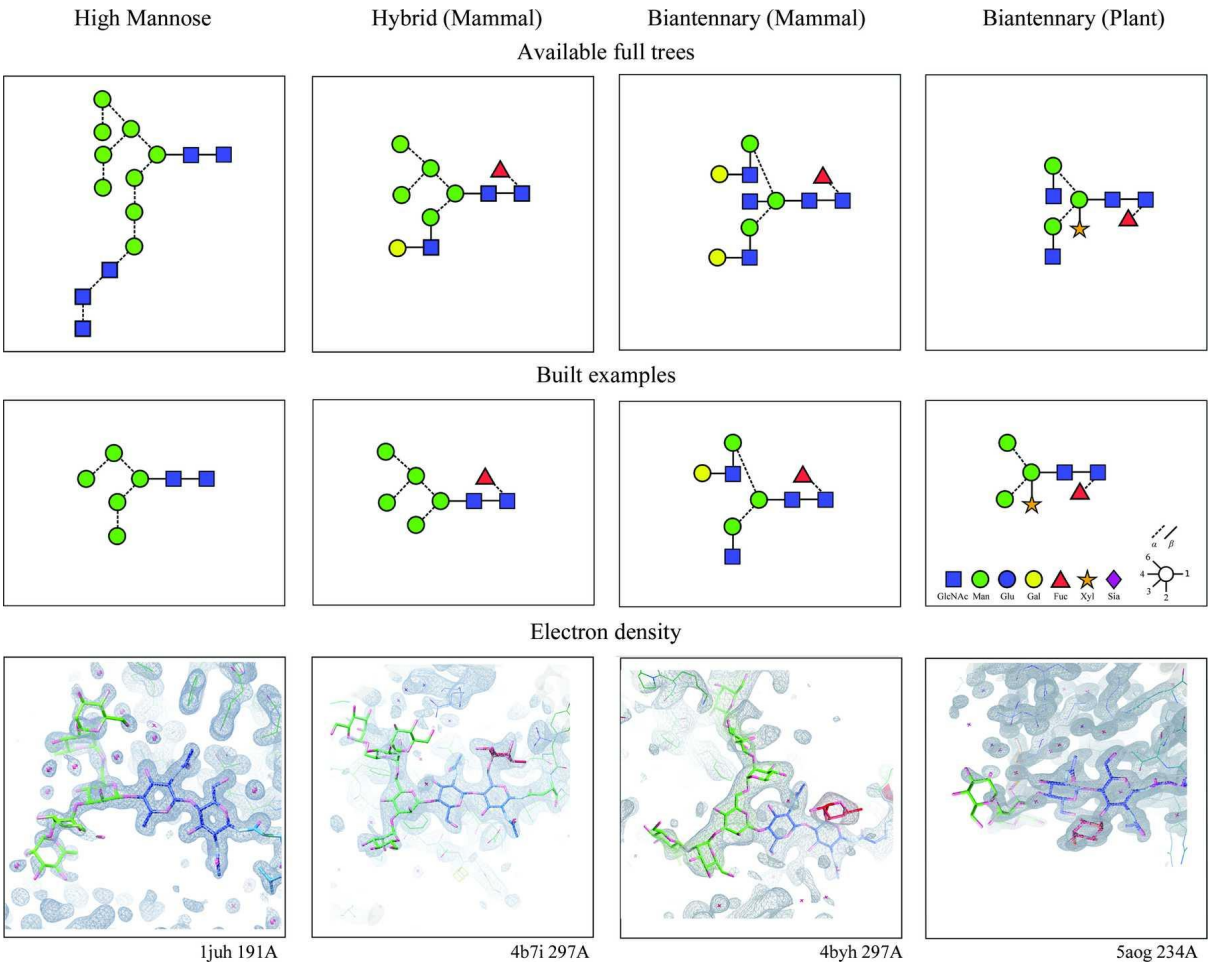


Figure 1.5. Results from a test of the N-glycosylation building tool in Coot (Paul Emsley and Crispin 2018). The diagrams in SNFG format show the expected glycoforms and the subsets Coot was able to build automatically, while the third row of pictures shows how the maps looked like in each example. Reproduced from (Paul Emsley and Crispin 2018) with permission of the International Union of Crystallography.

The improved N-glycosylation building module for Coot

Coot (P. Emsley et al. 2010) has a carbohydrate-building tool (Paul Emsley and Crispin 2018) – earlier version reviewed in (Agirre et al. 2017b) – that can be used to build N-glycosylation into both crystallographic and cryo-EM maps. The module has three modes: manual, semi-automated and automated. The manual mode allows the user to choose a monosaccharide and a bond type from a selection of commonly available glycoforms. Coot chooses the best position, orientation and conformation for the selected monosaccharide, and refines the structure. In the semi-automated mode the user selects a glycan type and Coot returns possible options for the monosaccharide and the glycosidic bond. The automated mode requires the user to simply choose the starting point and the glycosylation tree type, and Coot builds it automatically, interrupting the process when no more sugars can be built into clear density. An overview of results is presented in Figure 1.5 (adapted with permission from IUCr Journals). The tool has received positive adoption by the community, as evidenced by its use on several high-profile X-ray and cryo-EM structures (Zhu et al. 2018; B. Zhang et al. 2019; Klünemann et al. 2019; Lee et al. 2019).

Its main limitations are the relatively narrow selection of glycoforms available – clearly a design decision rather than an oversight, as these represent the most common forms that can usually be determined experimentally – and the fact that Coot does not include temperature-factor refinement, as all atoms are set to a fixed value. Indeed, Coot provides a fixed B-factor value (20.0 $Å^2$) for each new atom in the structure, which can be fed forward to other refinement programs. The authors suggested integrating the model-free B-factor refinement procedure described by Cowtan and Agirre (Kevin Cowtan and Agirre 2018) as an improvement.

PDB-REDO: Carbivore and carbonanza

Van Beusekom et al. (van Beusekom et al. 2019) presented a set of tools that build on the Coot N-glycosylation building module to achieve a more automated behaviour; indeed, the software is meant to be part of their PDB-REDO (Joosten and Lütteke 2017b) rebuilding and re-refinement pipeline. The first tool they presented is *Carbivore*, which can be used to rebuild and extend existing N-glycosylation trees automatically, or add new trees where they are missing. For the case glycosylation was not detected due to C1 not facing the asparagine side-chain, the authors introduced another program, named Carbonanza, to generate link records. The whole-tree addition method of Coot was extended to allow for building partial trees, i.e. extending existing trees. Moreover, a feature that finds N-glycosylation sites based on the consensus sequence Asn-X-Ser/Thr was implemented in

Carbivore. In addition, an option for finding N-glycosylation sites based on homologous models was also presented, however this is not used by default as the search is likely to be slow.

ISOLDE

The ISOLDE plugin (Tristan Ian Croll 2018) for ChimeraX (Goddard et al. 2018) offers a refreshing way of dealing with protein glycosylation, and supports both electron cryo-microscopy and X-ray crystallographic data. The graphical frontend connects to an interactive, GPU-accelerated molecular mechanics simulation, updating the model – and electron density maps, if working on crystallographic data – based on both the user's push-pull movements and the results of running the simulation on the updated coordinates. Technology-wise, this new tool makes use of the OpenMM toolkit (Eastman et al. 2017) for simulations, and the Clipper-python module (Stuart McNicholas et al. 2018) for electron density calculations, which is heavily CPU-parallelised – using C++11-style threads – in the latest version available from the *ChimeraX toolshed* at the time of publication. Protein glycosylation is handled by an adapted version of the GLYCAM force field (Kirschner et al. 2008). Although at present some unwanted effects such as ring inversions might appear as a result of the unrealistically high temperatures simulated by the user's push-pull movements, it is clear that this tool will be of great assistance when multiple overall glycan conformations need to be evaluated in a low resolution map; a combination with real-time validation at both the monosaccharide and glycan levels could further inform the fitting process and prevent errors too.

Sails

Sails (glycojones n.d.) can be used to build sugars automatically, either covalently linked to protein or as ligands. The software is currently in the middle of a major infrastructural change but is slated for general release in 2020 (with, or through an update to CCP4 7.1). It uses a method similar to that of Nautilus (Kevin Cowtan 2014) and Buccaneer (Kevin Cowtan 2006, 2012), using fingerprint-based detection of fragments, which account for both the target and its environment. The correlation function behind Sails has been proven to work with electron cryo-microscopy data, although adjustments may be needed if *e.g.* the scale of the EM map is not accurate or different map sharpening or blurring is required. Privateer and Refmac will be integrated with Sails in a pipeline for iterative building, refinement and validation.

1.2.6.4 Refinement and validation

Privateer

Privateer (Agirre, Iglesias-Fernández, et al. 2015b) is a carbohydrate-specific validation tool that can determine ring conformation of furanose and pyranose rings, anomeric form, absolute stereochemistry, real space correlation between model and omit density, and will generate other output such as SVG glycan diagrams in the Symbol Nomenclature For Glycans (SNFG) notation, and scripts for both Refmac5 (Murshudov et al. 2011) and Coot (P. Emsley et al. 2010). Like Sails, it is undergoing a change in infrastructure in order to future-proof its architecture.

Among the different checks that Privateer will do on carbohydrate models, a comparison of ring conformation and the ideal, minimal-energy conformation for each monosaccharide provides the fastest and most useful indication of potential mistakes in modelling and/or refinement: at high resolution, unjustified high-energy conformations - those without support of clear electron density - can reveal problems in the glycosidic bond (wrong anomer used, for instance) or wrong restraints (e.g. inverted chiralities). At low resolution, the problem can appear if the model is allowed to deviate from the ideal geometry due to providing insufficient restraints during refinement. Privateer generates dictionaries containing unimodal restraints upon detecting unjustified high-energy conformations. The validation and re-refinement process via these dictionaries is now completely automated via the CCP4i2 interface (Potterton et al. 2018a). These developments were spearheaded after it was revealed that the PDB contained an unrealistically high number of non-chairs as part of N-glycosylation (Agirre, Davies, et al. 2015b).

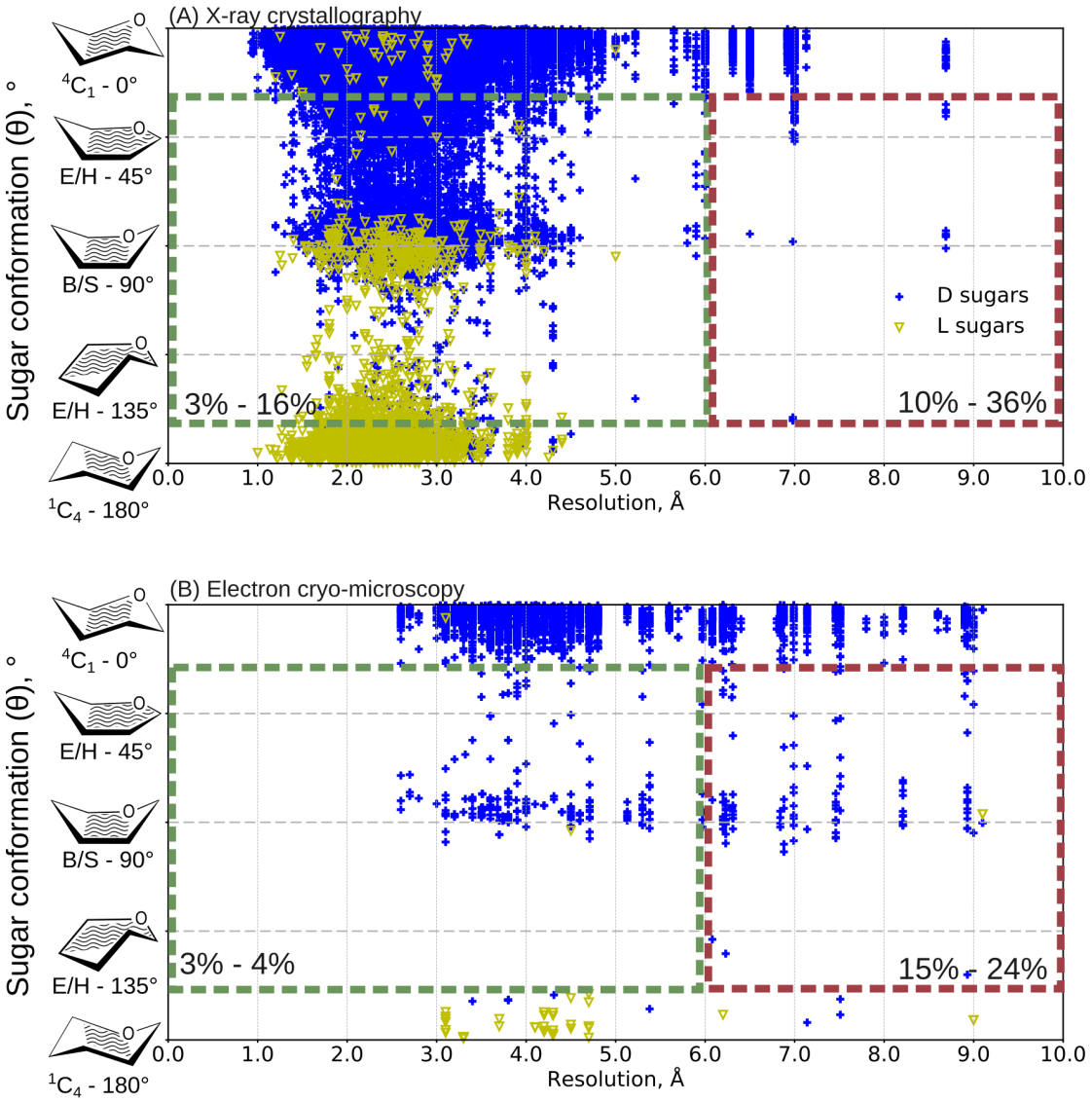


Figure 1.6. Pyranose ring conformations vs resolution for all sugars part of N-linked glycoproteins determined with (A) X-ray crystallography or (B) electron cryo-microscopy in the PDB by April 2019. E/H: Envelopes and Half-chairs, B/S: Boats and Skew-boats. Wavy lines denote the main ring plane. For reasons of clarity, half-chair, skew-boat and envelope were omitted from the axes at θ=45°, θ=90° and θ=135° respectively. Percentage of sugars in non-chair conformations is shown for resolution ranges 0.0-6.0 Å and 6.01-10.0 Å.

Many newer cryo-EM structures of glycoproteins are in the 2 Å to 6 Å resolution range due to improvement in electron sources, detectors, and image processing and 3D reconstruction algorithms. But the software for structure solution and validation have also improved, and perhaps as a result of that, high-resolution cryo-EM structures display fewer sugars in high-energy conformations than crystallographic ones. To illustrate this point, Privateer was

run on all N-glycosylated structures in the PDB, determined with X-ray crystallography and cryo-EM. The decoupled results are shown in Figure 1.6. D-sugars are shown in blue, L-sugars are shown in yellow. Ideally, in the particular case of N-glycosylation all D-sugars should be in $^4C_1$ conformation, and all L-sugars in $^1C_4$ conformation.

As previously highlighted elsewhere (Agirre 2017), pyranose higher-energy conformations are unusual as Ramachandran outliers, and should be reported alongside them in the refinement summary table.

Cryo-EM models of SARS-CoV2 have generally received a lot of attention due to the urgent need for better understanding of the virus. As a result, there has been a significant effort to achieve the highest possible resolution and accuracy in the resulting models, both pre- and post-deposition (Tristan I. Croll et al. 2021). For example, several studies have reported cryo-EM structures of the SARS-CoV2 spike protein at resolutions up to 2.5 Å , which is higher than many other cryo-EM structures (Wrapp et al. 2020; Lan et al. 2020). In addition, the SARS-CoV2 models have been subjected to careful validation and scrutiny and may contain fewer errors than other cryo-EM models.

Phenix, Rosetta and AMBER

Phenix uses a conformation-dependent library of restraints for the protein backbone (Moriarty et al. 2014) and homology refinement (Park et al. 2018) for protein modeling. Rosetta can be used for carbohydrate refinement of both X-ray and cryo-EM structures using parameterisation derived from X-ray structures to approximate conformational energy (Alford et al. 2017). Frenz et al., (Frenz et al. 2019) developed a protocol that can use either low-resolution crystallographic data, through Phenix-Rosetta integration (DiMaio et al. 2013) or cryo-EM data.

The RosettaCarbohydrate framework includes torsion-space refinement for glycans, which assumes ideal bond lengths and angles (Labonte et al. 2017). Frenz et al., (Frenz et al. 2019) build on previous work by expanding Rosetta's geometry term to include bond geometry deviations. These were derived from Phenix using eLBOW with AM1 optimization and added to the Rosetta database. Currently the sugar monomers included are α and β glucose, N-acetyl glucosamine, α and β mannose, and α and β fucose.

The authors recommend using Privateer (Agirre, Iglesias-Fernández, et al. 2015b) before and after refinement to detect errors in the structure. For refinement of crystallography data, Rosetta's integration with Phenix can be used (Terwilliger et al. 2012). The protocols were

modified to account for glycans, including steps for minimisation, increasing repulsive weights, and idealisation of anomeric hydrogens.

Phenix also offers integration with the AMBER molecular mechanics package, which is known for calculating torsion potentials accurately (Case et al. 2005).

A word on legacy validation tools

While the tools outlined in this section are now sadly unsupported, it is worth mentioning them not just for the sake of completeness, but because there is no substitute tool yet for some of the key functions they provide. PDB-CARE (PDB CArbohydrate REsidue check; (Lütteke and Von Der Lieth 2004; Lutteke 2004)) is a tool that can be used for bond and nomenclature validation. It is based on pdb2linucs, which is a software for carbohydrate detection based on atom types and their coordinates. The LINUCS notation (Bohne-Lang et al. 2001) is used to normalise carbohydrate structures. This is done by comparing the carbohydrate structures' LINUCS notation to the PDB HET Group Dictionary, which contains sugar residues present in the coordinate file (Lütteke and Von Der Lieth 2004). If a structure contains multiple anomers due to mutarotation at the reducing end of a saccharide, both forms need to have the correct PDB three-letter codes.

CARP (CArbohydrate Ramachandran Plot) is a tool that can be used to evaluate glycosidic linkage torsions. CARP also uses the pdb2linucs algorithm to analyse data, and compares it to data in GlyTorsionDB or GlycoMapsDB (for less common linkages). For each pair of monosaccharides and linkage combination, a separate torsional plot is created (Lütteke, Frank, and Von Der Lieth 2004). While these tools have been used mainly for validation purposes, they are a nice complement when examining the different linkage conformations in disaccharides (Fushinobu 2018).

1.2.6.5 Representation

While all-atom representations are the way to go for showing the interactions between protein and carbohydrate ligands, there is a case for using a simplified representation for glycans taking part in protein glycosylation; indeed, the sheer number of potential interactions occurring due to the size of the glycans – in optimal cases, 9 or more linked monosaccharides could be visible – and the particular relevance of their composition make

all-atom figures difficult or near-impossible to follow. McNicholas and Agirre (Stuart McNicholas and Agirre 2017) introduced a representation (*Glycoblocks* for CCP4mg (S. McNicholas et al. 2011)) that, building on a 3D extension of the now standard Symbol Nomenclature For Glycans (SNFG) (Varki et al. 2009, 2015), added minimalistic dashed lines for hydrogen bonds and CH-π interactions.

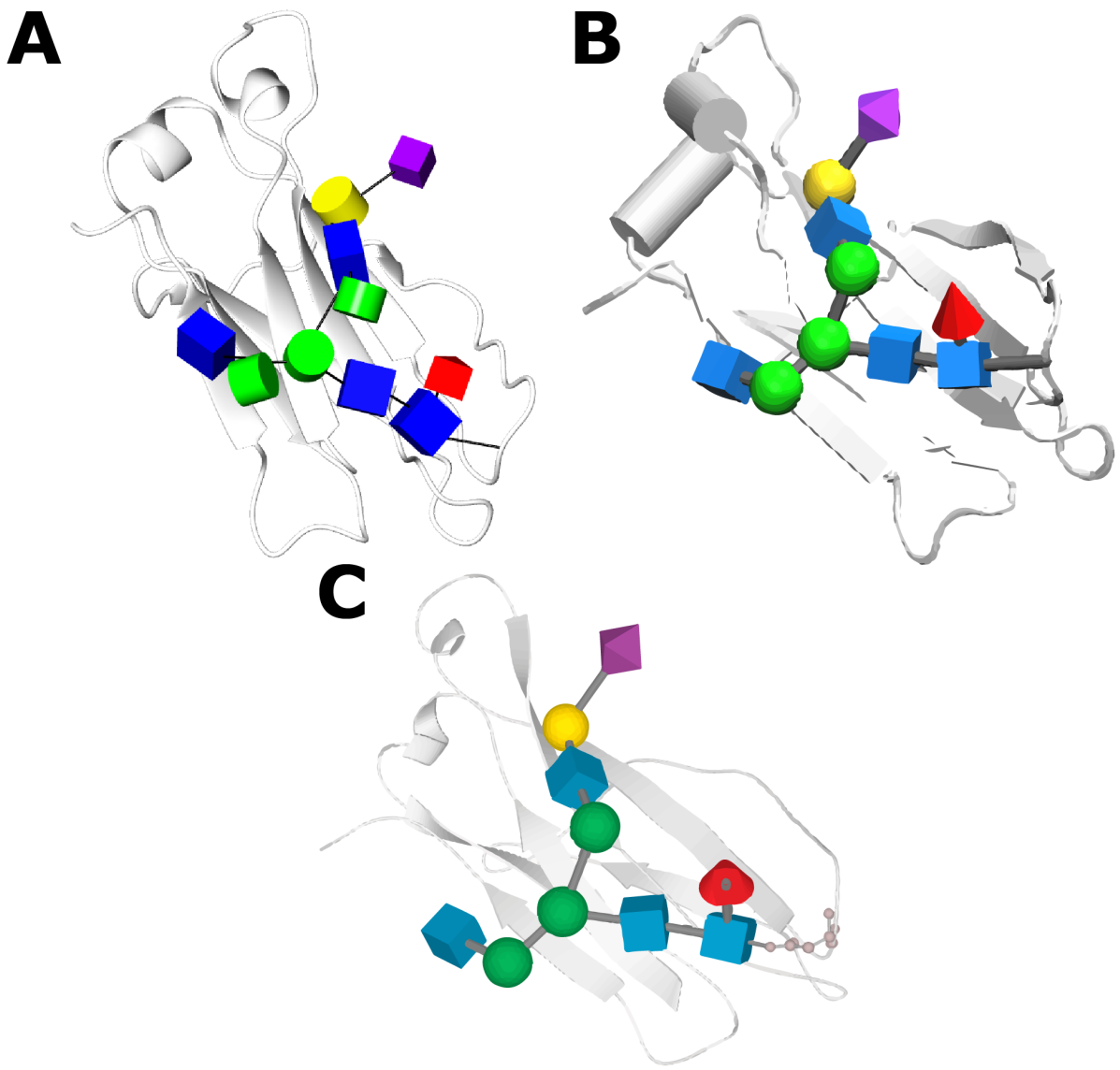


Figure 1.7. 3D SNFG glycan representation comparison of PDB code 4BYH in selected software: (A) CCP4mg (Stuart McNicholas et al. 2018) with Glycoblocks (Stuart McNicholas and Agirre 2017), (B) VMD (Thieker et al. 2016) and (C) LiteMol (Sehnal and Grant 2019).

Not focusing on interactions, many 3D SNFG representations exist now either as plugins or as an integral part of wider-purpose graphics software, e.g. VMD (Thieker et al. 2016),

LiteMol (Sehnal and Grant 2019), and UCSF Chimera (Pettersen et al. 2004) via the Tangram plugin (insilichem n.d.). These provide stand-out depictions of protein glycosylation using big regular polyhedra. A side-by-side comparison is shown in Figure 1.7. Finally, other software such as SweetUnityMol (Pérez, Tubiana, et al. 2015) and Pymol (Arroyuelo, Vila, and Martin 2016) combine the familiar colouring scheme with a more atomistic representation.

1.2.6.6 Future perspectives

It appears the gears are finally turning in the methodological machine towards implementing better support for carbohydrates. However, software still require expert knowledge of carbohydrate structure or very high resolution to work automatically. Work is currently being done on the Sails program to be able to overcome many of these limitations. In addition, based on encouraging early results (Agirre et al. 2016, 2019; Schumann et al. 2019), new carbohydrate dictionaries with more faithful model geometry and accurate torsion restraints will improve refinement, particularly for cryo-EM. Finally, sugars in active sites of enzymes might be distorted into high energy conformations, and thus may require further validation; work will need to be done in this respect in order to give users a confidence level on their conformational assignment.

We should like to emphasise that model building, refinement and validation will need to be further integrated together for maximum benefit of users. Recently, Van Beusekom, Lutteke and Joosten (van Beusekom, Lütteke, and Joosten 2018) used a set of tools, including PDB-REDO (Joosten and Lütteke 2017b), Privateer (Agirre, Iglesias-Fernández, et al. 2015b) and CARP (Lutteke 2004) to analyse 8,114 glycoproteins from the PDB. They succeeded in correctly re-annotating 3,620 carbohydrate residues, which were then re-refined and are now available for the community to use. Incorporating prior glyco-chemical knowledge into the structure solution process will, as exemplified by the aforementioned authors, extend the limits of resolvability further down our glycans.

1.2.6.7 Acknowledgements

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

means limited to financial backing – of the Department of Chemistry and the University of York.

1.2.7 Summary

A general introduction to the background of x-ray crystallography and cryo-EM, along with a review of the software tools aimed at carbohydrate 3D structure solution was presented in this section. In addition, a study was carried out to detect conformation errors in monosaccharides part of N-glycans in the PDB. Up to 36% of these monosaccharides show high energy conformations that are unlikely to be found in N-glycans and are therefore considered incorrect. Since its publication, this article has been cited in multiple publications as a source of information for carbohydrate software tools.

The spike protein of SARS-CoV2 is the viral protein responsible for binding to human cells and initiating infection. It is heavily glycosylated with N-glycans, including a glycan shield used to circumvent the host immunity. Cryo-EM has been crucial in determining the structure of the spike protein and its interactions with human cells (Wrapp et al. 2020; Gui et al. 2017). Stagnoli et al. (Stagnoli et al. 2022) use a combination of cryo-EM and molecular dynamics simulations to study the geometries of the N-glycans in the SARS-CoV2 spike protein's glycan shield, citing the article presented in Section 1.2.6. In addition, X-ray crystallography has been used to study the structure and function of other SARS-CoV2 proteins, such as the main protease (Z. Jin et al. 2020) and the RNA-dependent RNA polymerase (Gao et al. 2020).

Chapter 2

Carbohydrate Dictionaries

Using restraint dictionaries is important when doing structure refinement at low resolution for both X-ray crystallography and electron cryo-microscopy. They form a database used by refinement software. They provide prior knowledge of chemistry and structure to aid with producing a realistic model. For example, they can contain values for bond lengths and angles that can inform the bond lengths and angles in the structure to be determined. This is especially important for carbohydrates, because of the inherent difficulties associated with their structure discussed in Chapter 1. In fact, a large number of errors were identified in carbohydrate structures in the PDB determined by both X-ray crystallography and electron cryo-microscopy (Atanasova, Bagdonas, and Agirre 2020). A contributing factor for this was likely that over 60 of the old set of carbohydrate dictionaries in the CCP4 Monomer Library contained errors, such as monosaccharides having torsion angles set at 0° (Agirre 2017).

This chapter is an attempt to correct these errors and to add additional restraints that can enforce carbohydrate conformation. As discussed in Chapter 1, conformation can be distorted at low resolution. The first step to correcting the errors in the dictionaries was to generate new dictionaries with a dictionary-generation program (done by Rob Nicholls and Mihaela Atanasova). Then, these dictionaries were validated with Privateer and Coot to ensure they are correct (done by Mihaela Atanasova). Furthermore, new unimodal ring torsions were added with Privateer (the specific code in Privateer was written by Jon Agirre; the addition of the unimodal ring torsions to the dictionaries was done by Mihaela Atanasova).

The second aim of this chapter is to provide a comparison between the new carbohydrate dictionaries and the old ones. This also provides insights into the use of unimodal torsion restraints when refining carbohydrates. To do this, the dictionaries were tested on a large dataset using the PDB-REDO pipeline (done by Robbie Joosten). The output was analysed (done by Mihaela Atanasova). The main conclusion is that the new dictionaries lead to fewer errors during refinement, especially if the unimodal torsion restraints are activated. However, care needs to be taken when using the unimodal torsion restraints at high resolution, as the data should already give a good idea of what the monosaccharide conformation should be.

2.1 Published Article: “Updated restraint dictionaries for carbohydrates in the pyranose form”

The content in this section is taken from “Updated restraint dictionaries for carbohydrates in the pyranose form” by Atanasova *et al.* (Atanasova et al. 2022b). Supplementary data are available in Appendix A.

2.1.1 Introduction

Macromolecular refinement is a computational procedure that lies among the final steps in protein structure solution. Provided that a suitable strategy is selected, restrained refinement iteratively improves the agreement between a macromolecular model and experimental data. Describing the position and properties of each atom in a macromolecule requires many parameters, but often only a limited amount of experimental observations is available. This can lead to a problem of overfitting, where the model fits too closely to the available data, even if the model is not an accurate representation of the macromolecule. To avoid this problem, prior chemical knowledge about the molecules involved is usually required in order to maintain refinement stability. In macromolecular crystallography, such prior knowledge is stored in dictionaries, typically in the form of Crystallographic Information Files (Hall, Allen, and Brown 1991; I. D. Brown and McMahon 2002) or MOL(2) files (MDL Information Systems Inc., San Leandro, California, USA). It is often the case for each molecular component (or residue, including carbohydrate monomers) to be represented in a separate CIF file. The restraint dictionary entries used by software in the CCP4 suite (Winn et al. 2011a) are collected in the CCP4 Monomer Library (A. A. Vagin et al. 2004b). Such prior chemical knowledge usually consists of atom names, a description of stereochemical properties such as connectivity, bond lengths, angles, chirality, torsion angles and, if applicable, a list of groups of four or more atoms in planar co-arrangements.

This is especially important when modelling carbohydrates, which tend to be less well resolved due to flexibility, microheterogeneity, disorder (Joosten and Lütteke 2017a; Atanasova, Bagdonas, and Agirre 2020) and relatively low data resolution (van Beusekom, Lütteke, and Joosten 2018). In addition, pyranosides – monosaccharides that form a 6-membered saturated ring – can exhibit a range of different ring conformations. However, pyranosides are most frequently found in their lowest-energy conformation: a chair – $^{4}C_{1}$ for D-sugars and $^{1}C_{4}$ for L-sugars; this is particularly true of N-glycosylation (Agirre, Davies, et al. 2015a). Chair conformations have minimal repulsions and strain due to the substituents

being staggered rather than eclipsed, resulting in the dihedral angles between consecutive ring atoms being +/-60 degrees. Non-chair conformations and conformational transitions are costly in terms of energy, and occur most frequently in enzymatic reactions (Davies, Planas, and Rovira 2012). Therefore, atomic models showing high-energy ring conformations need to be supported clearly by experimental data, and evidenced in electron density/potential maps (Agirre 2017).

The CCP4 Monomer Library (CCP4-ML) was originally generated using the LIBCHECK software (A. A. Vagin et al. 2004b), which derived ideal values for saccharides from nucleic acid studies (Saenger 1984). The CCP4-ML has seen recent expansion, and many component entries and linkages have now been replaced with AceDRG dictionaries (Nicholls, Joosten, et al. 2021a). However, monosaccharides in pyranose form were set aside to be treated separately due to the particularities concerning ring conformation, which have been recently reviewed (Joosten, Nicholls, and Agirre 2022). Revision of these entries is thus overdue, not just as an effort to modernise geometric estimates, but as a way of correcting issues flagged up in the past (Agirre 2017).

In general, restraint dictionary generation programs use methods based either on data derived from small molecule databases or on quantum chemical calculations. Small molecule-based dictionary generation programs extract high-resolution geometric information from small molecule databases in order to produce restraints for use in macromolecular crystallography. Examples of such small molecule databases, which contain structural models that were derived using small molecule X-ray crystallography, include the Cambridge Structural Database (CSD; paid access, 1,136,493 deposited structure models at the time of writing) (Groom et al. 2016) and the Crystallography Open Database (COD; free access, 473,816 deposited structure models at the time of writing) (Gražulis et al. 2012). Whilst both databases are curated, a recent study showed that *post hoc* validation checks need to be in place if intending to use derived information to make reliable inferences regarding stereochemical geometries (Long et al. 2017b). Mogul (Bruno et al. 2004) and AceDRG (Long et al. 2017a) utilise molecular geometry information extracted from the CSD and the COD, respectively.

There are multiple contemporary restraint dictionary generation programs, which use different combinations of databases and mining tools. AceDRG mines the COD, validating entries. Then, it compiles “AceDRG tables” containing atom types, bond types and other information included in restraint dictionaries. These tables are distributed as a part of the CCP4 software suite, and are used during the restraint dictionary generation procedure. AceDRG uses RDKit (http://www.rdkit.org) for internal molecular representation, from which

it identifies atom types. Combined with the data from the aforementioned tables, this produces a restraint dictionary entry. Finally, AceDRG uses RDKit to generate multiple possible conformers, and chooses the one with the lowest free energy. GRADE (Global Phasing Ltd.), Pyrogen from Coot (P. Emsley et al. 2010) and Phenix.eLBOW (Moriarty, Grosse-Kunstleve, and Adams 2009) can use Mogul (Bruno et al. 2004) to mine the CSD. Pyrogen can also use the CCP4-distributed tables created by AceDRG. In addition to mining small molecule databases, eLBOW can also use force fields to utilise quantum chemical calculations. A default simple force field and the semi-empirical RM1/AM1 method (Dewar et al. 1985; Rocha et al. 2006) are both implemented internally and do not rely on external software or third-party resources. Full quantum chemical calculations with a number of third party quantum chemistry packages are also available for use. These are useful when insufficient data about a particular chemistry are present in small molecule databases.

Carbohydrate model validation software such as the Privateer software (Agirre, Iglesias-Fernández, et al. 2015b) use a combination of established metrics (RSCC, average B-factors) and carbohydrate-specific ones (puckering coordinates, nomenclature checks) to identify problematic models. Further approaches to general ligand validation available in CCP4 are discussed by Nicholls (Nicholls 2017). Coot includes ligand validation features that allow visual assessment of multiple metrics alongside associated percentile ranks relative to all X-ray structural models (Paul Emsley 2017). These include RSCC (Equation 1) of the ligand-omitted $2mF_o$-DF_c map, RSCC of the difference map, the number of atom-pairs with unlikely contacts between them and the *Mogul Z*-worst score (comparing the value of a geometric parameter to data collected from the CSD). Flatland Ligand Environment View (Paul Emsley 2017) is a Lidia feature that shows the ligand in 2D for an alternative visualisation of a ligand in its structural context, highlighting intermolecular interactions. Map Sharpening can uncover missing features and is especially useful for flexible regions of the model. Finally, inspecting refined B factors may also provide a useful insight into model reliability, especially when comparing the B-factors of proximal atoms (Masmaliyeva, Babai, and Murshudov 2020).

$$RSCC = \frac{\Sigma(\rho_{obs} - \langle\rho_{obs}\rangle)(\rho_{calc} - \langle\rho_{calc}\rangle)}{[\Sigma(\rho_{obs} - \langle\rho_{obs}\rangle)^2 \Sigma(\rho_{calc} - \langle\rho_{calc}\rangle)^2]^{1/2}} \quad \text{(Equation 2.1)}$$

New dictionaries with improved ring torsion restraints, coordinates reflecting the lowest-energy ring pucker and updated geometry, have been produced and evaluated using

some of the metrics mentioned above. The new dictionaries, now part of the CCP4 Monomer Library, will be released with CCP4 8.0.

2.1.2 Materials and Methods

2.1.2.1 Design guidelines

When using a restraint dictionary generation program, the user needs to specify the molecular component for which a dictionary entry is to be generated. Typically, a preexisting file that specifies chemical composition, connectivity and atomic nomenclature is used as input (e.g. using CIF, MOL or MOL2 format). In cases where such a file is unavailable, the chemistry can be specified using a Simplified Molecular-Input Line-Entry System (SMILES) string as input. SMILES is a linear notation that can represent 3D molecules as strings of characters (Weininger 1988). Another option is to use a 2D sketcher, such as Coot's Lidia (Paul Emsley 2017) and ChemDraw (PerkinElmer Informatics) or a 3D sketcher, such as JLigand (Lebedev et al. 2012), to manually draw the molecule to either produce a SMILES string or otherwise a file that can be used as input.

Agirre (Agirre 2017) analysed multiple restraint dictionary generation programs by comparing the bond lengths and angles of the output for α-D-glucopyranose from a SMILES string to the ideal geometry described in the CCP4-ML. It was concluded that the dictionaries produced by AceDRG, GRADE, and eLBOW using Mogul were roughly in agreement – meaning that they showed similar deviations from the targets proposed in the CCP4-ML. Recently, this observation was confirmed in a wider study, which showed that modern restraint dictionary generation programs now show consistent results for carbohydrates in the pyranose form (Joosten, Nicholls, and Agirre 2022). Since AceDRG is the CCP4 program that is already being used to update dictionaries for other peptide, nucleotide, and non-polymeric chemical compounds in the CCP4-ML (Nicholls, Wojdyr et al., 2021), as well as for chemical linkages between components, it was chosen for generating the restraint dictionary entries reported herein.

AceDRG allows different input options for the molecule to be generated. As mentioned above, a SMILES string is commonly used as input when there is no restraint dictionary entry already available for a given molecular component – a common scenario during drug discovery. However, pyranose sugars follow the atomic nomenclature established by the Worldwide Protein Data Bank (wwPDB, (Burley et al. 2019) in its Chemical Component

Dictionary (CCD, (Westbrook et al. 2015)), which in turn now mirrors IUPAC nomenclature following the recent remediation of carbohydrate entries. In contrast, restraint dictionary entries produced from SMILES strings do not follow this convention (as atom nomenclature is not encoded in SMILES) and thus may end up causing issues during model building and refinement. For this reason, existing component definitions from the CCD were used as a starting point. These CIF files contain a description of the compound in terms of atom names, types and connectivity. Additionally, many component definitions contain idealised atomic coordinates from QM calculations. However, these coordinates do not provide enough information to construct restraints; any derived restraint target would lack an estimated standard deviation, which would be needed for relative weighting during refinement. Moreover, it has been found that restraints derived from QM-based calculations are at present inconsistent with those mined from high-quality small molecule X-ray structures (Joosten, Nicholls & Agirre, 2021).

During the restraint dictionary generation process for pyranose sugar entries, it is necessary to ensure that the anomeric configuration and the stereochemistry of the substituents are correct, and that the Cremer-Pople puckering coordinates that define the conformation of the pyranose ring in the conformer (Cremer and Pople 1975a) reflect a minimal-energy ring pucker, representative of the majority use case due to the rigid conformational preferences that saturated rings exhibit. As per the recommendations proposed by Joosten, Nicholls & Agirre (Joosten, Nicholls, and Agirre 2022), restraint dictionary entries should present coordinates as close as possible to the most probable conformer. Torsion restraints, if present, should match those coordinates, allowing refinement software to restrain ring conformation to a minimal-energy pucker at low resolution. The Privateer software (Agirre, Iglesias-Fernández, et al. 2015b) was used to analyse the produced coordinates – this ensured that stereochemistry, anomeric configuration and puckering parameters met the expectations for each particular compound. As a secondary sanity check, all the produced sugar monomers were visually inspected in Coot (P. Emsley et al. 2010) to confirm that the stereochemistry of the substituents, the anomeric configuration and the conformation meet expectations. Furthermore, for those sugars where a protein crystal structural model derived using data extending to 1.5 Å resolution or better was available, the conformer presented in the new restraint dictionary entry was visually compared to the crystal structural model for further validation.

It is important to note that the atomic coordinates listed in the restraint dictionary are only used when initially placing the sugar into the model. Having reasonable starting coordinates - i.e. a low-energy conformer - is important in order to ensure a sensible starting point from which further model building and refinement can proceed. Any large deviation of the initial

coordinates from the restraint targets causes imbalance in the refinement target function causing slow and possibly suboptimal refinement. Once the model is under refinement, these initial coordinates are discarded and it is the restraints themselves that continue to ensure reasonable conformation. Consequently, it is of primary importance for the restraints to adequately reflect the component's allowable geometric landscape. Inclusion of unimodal ring torsion restraints helps to ensure maintenance of low-energy ring conformation throughout the refinement procedure, except in cases where there is strong evidence to the contrary.

2.1.2.2 Protocol for generating new dictionaries

A set of pyranosides was obtained from the list of monosaccharides supported by the Privateer software (obtainable by running 'privateer -list' on the command line); this set comprises the most frequently modelled pyranosides. CCD CIF files containing existing pyranoside definitions were downloaded from the PDB. These files were provided as input to AceDRG (v231). Additionally, different AceDRG options were explored to avoid unexpected results such as distorted conformers. AceDRG samples many potential conformers; those with the lowest energy according to RDKit are optimised using the idealisation mode of REFMAC5 (Murshudov et al. 2011), and ultimately the one with the lowest energy is selected (Long et al. 2017a). Specifying for a greater number of conformers to be sampled results in noticeably increased computation time. The new restraint dictionary entry for each monosaccharide was output as a CIF file, along with a PDB file containing coordinates corresponding to a low-energy conformer. The PDB files for all entries were provided as input to Privateer for validation (Table A.1), and Coot for further visual inspection. A compilation of all Privateer validation data is presented in Table A.1. Furthermore, in order to allow use of predefined restraints for glycosidic linkages from the CCP4-ML, the component types were set to 'pyranose' for aldopyranoses and 'ketopyranose' for the ketopyranoses (Nicholls, Wojdyr, et al. 2021). Adding the correct type is necessary as the glycosidic linkage restraints assume standard atom naming conventions for the atoms involved, which are different for aldo- and ketopyranoses.

Patched _chem_comp_tor section of a restraint dictionary separating **ring torsion angles** from the **rest**, as specified by **four atoms**, **target value**, **uncertainty** and **periodicity**

NAG	ring_1	C5	O5	C1	C2	-59.675385	3.0	1
NAG	ring_2	O5	C1	C2	C3	53.650513	3.0	1
NAG	ring_3	C1	C2	C3	C4	-52.014420	3.0	1
NAG	ring_4	C2	C3	C4	C5	54.096725	3.0	1
NAG	ring_5	C3	C4	C5	O5	-56.921230	3.0	1
NAG	ring_6	C4	C5	O5	C1	61.200516	3.0	1
NAG	tors_1	C8	C7	N2	C2	-175.114227	10.0	2
NAG	tors_2	N2	C7	C8	H81	-13.703261	10.0	6
NAG	tors_3	C5	C6	O6	HO6	-177.520996	10.0	3
NAG	tors_4	C4	C5	C6	O6	61.135471	10.0	3
NAG	tors_5	C6	C5	O5	C1	-175.561295	10.0	3
NAG	tors_6	O4	C4	C5	C6	63.707928	10.0	3
NAG	tors_7	C3	C4	O4	HO4	-61.268230	10.0	3
NAG	tors_8	O3	C3	C4	O4	-63.830528	10.0	3
NAG	tors_9	C2	C3	O3	HO3	-169.485916	10.0	3
NAG	tors_10	C7	N2	C2	C1	124.894669	10.0	6
NAG	tors_11	N2	C2	C3	O3	61.918137	10.0	3
NAG	tors_12	C2	C1	O1	HO1	163.115189	10.0	3
NAG	tors_13	O1	C1	O5	C5	179.557251	10.0	3
NAG	tors_14	O1	C1	C2	N2	-62.214077	10.0	3

Figure 2.1. A view of the patched torsion section in a CIF restraint restraint dictionary entry. This is an extract of the new CCP4 restraint dictionary entry for N-acetyl-β-D-glucosamine, GlcNAc, which is represented in the PDB database as 'NAG'. The new dictionaries distinguish ring torsion angles (prepended by 'ring_') from the rest ('tors_') so they can be activated separately to keep a low-energy ring pucker. Older CCP4 dictionaries had no separation between ring (unimodal) and rest of the torsions (periodicity 2, 3 or 6), and had a uniform uncertainty of 20.0º. Six decimal places – completely beyond the precision of even the highest-resolution structures – are used for the value of the torsion angle for reasons of compatibility with existing software. This should be changed in future, as a single decimal place would be enough.

The default torsion restraints generated by AceDRG do not exactly match the conformer's coordinates – e.g. a generic 60º *versus* 53.65º as measured along O5-C1-C2-C3 on an energy-minimised conformer of N-acetyl β-D-glucosamine. While 60º may be appropriate for a carbon-only saturated ring such as cyclohexane, the presence of an endocyclic oxygen in pyranosides means not all bond lengths are the same, and therefore torsion angles will reflect those differences. In order to address this potential shortcoming, the Privateer software has been recently extended to patch any restraint dictionary entry with torsion restraints that are measured from the cartesian coordinates, writing separate names for ring torsions and other torsions so they can be used separately or together (Figure 2.1). This

functionality was used to patch the torsion restraints in the new dictionaries. The restraints called “ring_1” to “ring_6” (shown in bold) are the ring torsion restraints responsible for enforcing the monosaccharide's minimum-energy ring conformation. Different sigmas for the ring torsions were tested (3.0º, 6.0º, 10.0º), selecting 3.0º as the value that yielded the fewest conformational outliers without having a detrimental impact on R-free. Outliers at higher sigma levels were manually inspected and found to be unsupported by the electron density, meaning that they should have been corrected by the torsion restraints. All ring torsions were therefore set to 3.0º, with the rest of the restraints left at AceDRG's default value of 10.0º. For reference, the sigmas in the LIBCHECK-generated CCP4-ML dictionaries were all 20.0º – indeed, most restraints in the new dictionaries now show smaller sigmas than the ones in the LIBCHECK-generated CCP4-ML dictionaries.

2.1.2.3 Testing the new dictionaries

As has been described elsewhere (Agirre et al. 2017a), errors in carbohydrate models may be caused by incorrect model building. For example, if a monosaccharide is wrongly identified and ends up being distorted into the electron density/potential map, or when the restraints used are insufficient to ensure reasonable geometry during refinement. Improved restraint dictionaries are expected to help prevent issues with refinement. On the other hand, they will in no way avoid modelling mistakes. Such mistakes may be corrected either manually using available prior glyco-chemical knowledge (Bagdonas, Ungar, and Agirre 2020b), or automatically using specialised tools (van Beusekom, Lütteke, and Joosten 2018; van Beusekom et al. 2019). Previous conformational analyses of PDB glycan data showed that the proportion of distorted pyranosides increased with worsening resolution, spiking significantly in the 1.8 Å - 2.0 Å region (Agirre, Davies, et al. 2015a). Keeping in mind that the frequency of modelling mistakes is much higher at low resolution (Kleywegt and Jones 1995), a decision was made to limit the test dataset to include only entries with nominal resolutions better than 2.0 Å. Many pyranosides in our list are present in a very limited set of published structures and were not featured in the test data set. Therefore, a decision was made to choose representatives from the most frequently modelled pyranosides: NAG, MAN, BMA, GLC, BGC, BOG, FUL, GAL, GLA and SIA. A test data set was then assembled from the 100 PDB models with the highest numbers of the aforementioned pyranosides, under the 2.0 Å resolution limit. The new dictionaries were then tested by refining the selected structural models with REFMAC5 (Murshudov et al. 2011), using previously optimised refinement settings (restraint weights, B-factor models, and solvent mask

parameters) extracted from the PDB-REDO databank (van Beusekom, Touw, et al. 2018). Three separate refinement protocols were devised: refinement with the current (referred to hereafter as 'old') dictionaries, refinement with the new dictionaries without torsion restraints, and finally refinement with the new dictionaries with activated unimodal torsion restraints for pyranoside ring bonds. The output structural models and maps were then analysed using Privateer and Coot. The resultant dataset was divided into 2 parts - sugars part of N-/O-glycosylation, and ligands. The sugars part of N-/O- glycosylation were further filtered by excluding monomers marked as “wrong anomer” (mismatch between the anomeric form specified by the three-letter code and what is on the structure), and glycans that cannot be found in GlyConnect, one of the glycomics databases supported by Privateer (Bagdonas, Ungar, and Agirre 2020b).

2.1.3 Results

A set of 243 new carbohydrate dictionaries has been produced with updated torsion restraints that encourage refinement software to retain the minimal energy ring pucker. Figure 1 shows an updated torsion section, taken from the new CCP4-ML entry for N-acetyl-β-D-glucosamine, or GlcNAc (CCD component ID: NAG). These new torsion restraints are especially important when the experimental data extend to only low resolution; enforcing torsion restraints from the dictionaries forces the sugar ring into the most likely conformation. As expected, introducing additional restraints – the torsional kind in this case – may occasionally lead to a lower Real Space Correlation Coefficient (RSCC) between model and map. This simply reflects the fact that refinement software is no longer able to (inappropriately) improve model-to-map correlation at the expense of stereochemical geometry, *e.g.* unfavourable bond lengths or angles, or inverted ring conformations that would require massive energy expenditure.

The test dataset was composed of 955 structural models containing 11,291 sugar residues; 5,620 of these sugars were covalently bound to protein as part of N/O-glycosylated structures, and 5,671 were ligands. Obsolete CCD entries (e.g. all disaccharides, which following the PDB's remediation of carbohydrate entries are now described as linked monosaccharides) were not included in the set.

Privateer was run on all structures in the test dataset, and the results were analysed. The sugars in the test dataset were split into categories “old dictionaries”, “new dictionaries” or “new dictionaries and torsions” based on which dictionaries were used during their refinement, and labelled as “yes”, “no” or “check” according to Privateer’s validation report.

Privateer assigns a “yes” diagnosis when a sugar’s anomeric configuration, chirality, Cremer-Pople puckering parameters and ring conformation are what is expected for the sugar's lowest-energy conformer. The use of ring conformation as a validation metric for pyranosides is attractive because it cannot be targeted directly by a minimisation of bond length and angle distortion – indeed, a boat conformation, which Privateer would show as an outlier, may have close-to-ideal bond lengths and angles. If Privateer detects any problems, the sugar is marked as “no”. However, if the only issue detected is in the ring conformation, Privateer instead marks the sugar as “check", in which case the user should check whether the high-energy conformation is supported by the electron density. Privateer contains a database of puckering parameters calculated from a manually curated set of conformers obtained from the PDB CCD and compared against CSD, COD and high resolution PDB structures. Ring conformation is a useful validation metric for pyranosides, however it needs to be used in combination with other metrics, particularly density-based ones, whenever unimodal restraints have been used due to bias towards one conformation.

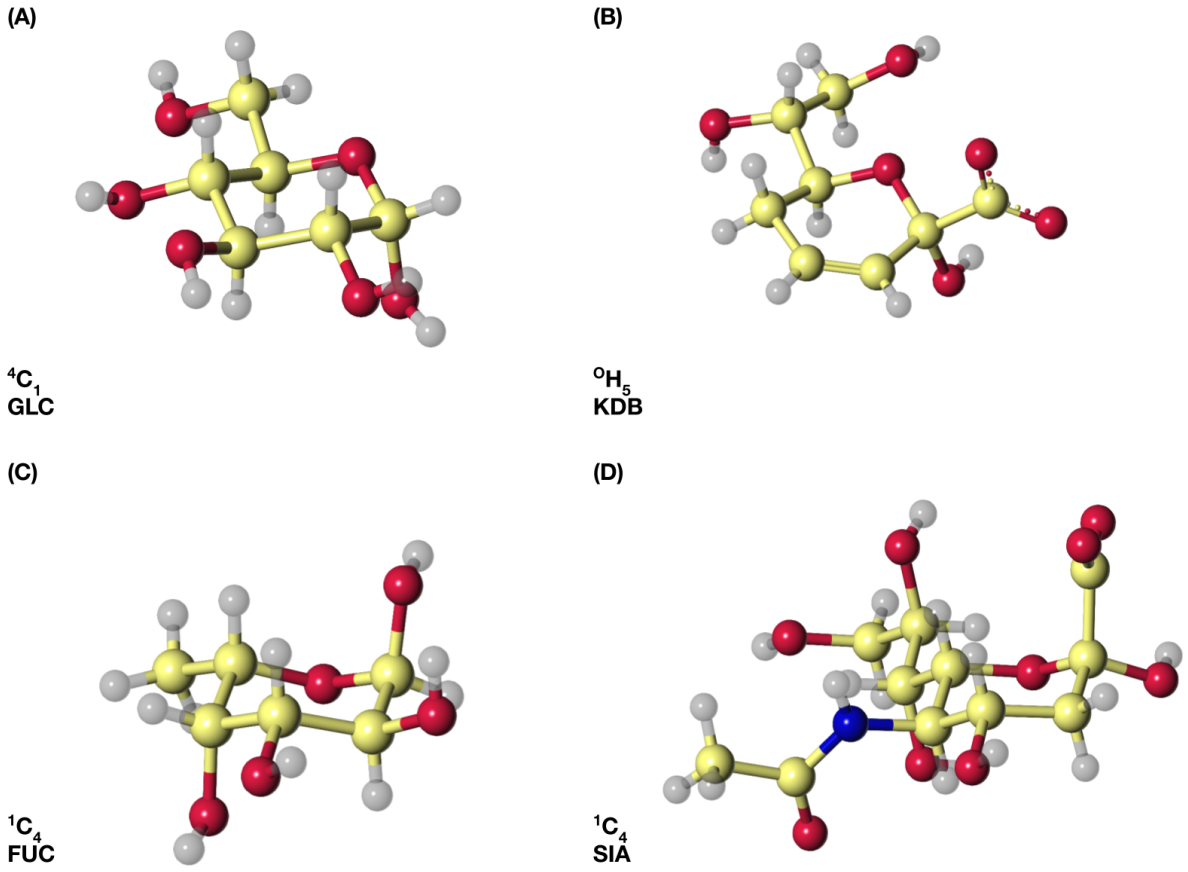
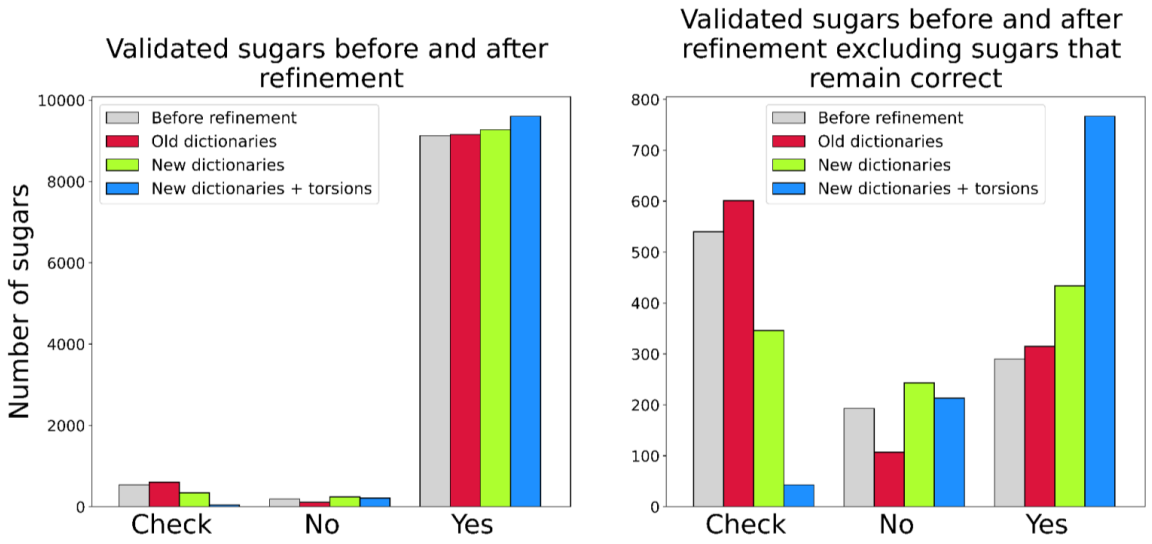


Figure 2.2. Carbohydrate restraint dictionary entries generated with AceDRG. (A) α-D-glucose in $^{4}C_{1}$ conformation, (B) 3,4,5-trideoxy-α-D-erythro-oct-3-en-2-ulopyranosonic acid in $^{O}H_{5}$ conformation, (C) α-L-fucose in $^{1}C_{4}$ conformation and (D) N-acetyl-α-neuraminic

acid (sialic acid) in $^{1}C_{4}$ conformation. This figure was produced with CCP4mg (S. McNicholas et al. 2011).

The puckering parameters of the conformers stored in the dictionaries were also analysed using Privateer. All dictionaries show the expected puckering for their particular chemistry. For example, saturated rings show a chair conformation ($^{4}C_{1}$ for D-pyranosides and $^{1}C_{4}$ for L-pyranosides) and pyranosides with a double bond in the ring, e.g. 3,4,5-trideoxy-α-D-erythro-oct-3-en-2-ulopyranosonic acid (CCD component ID: KDB), show four coplanar atoms in the ring (see Figure 2.2).



(A) (B)

Figure 2.3. Numbers of sugars diagnosed by Privateer as 'check', 'no' and 'yes' before and after refinement. A set of structures from the Protein Data Bank were refined with the CCP4-ML dictionaries, new dictionaries generated by AceDRG, and the new dictionaries with unimodal torsion restraints activated. From left to right, the coloured bars represent the number of sugars before refinement (grey); the number of analysed sugars after refinement with the CCP4-ML dictionaries (red); after refinement with the new updated dictionaries (blue); and after refinement with the new dictionaries with activated unimodal torsion restraints (yellow). (A) shows all analysed pyranosides; (B) only includes pyranosides that were diagnosed with 'check' or 'no' for at least one protocol.

The number of pyranosides in each category was counted, the incorrect entries were excluded as described in Methods and the results are presented in Figure 2.3. Figure 2.3A includes all 9,863 pyranosides from the test dataset. In order to focus on the sugars where

the new dictionaries have led to a change in behaviour, Figure 2.3B only includes pyranosides validated as “check” or “no” for at least one of the test runs (1,023 sugars). Sugars validated as correct for all three runs are deemed to be well supported by the experimental data and relatively easy to interpret. As expected, we registered a slight decrease in RSCC, whereas both refinement protocols involving the new dictionaries – with and without torsion restraints activated – managed to reduce the gap between R_{work} and R_{free}, meaning a reduction in overfitting (Figures A.1-A.3). In addition there was a slight reduction in mean atomic B-factors (Figure A.4)

2.1.4 Discussion

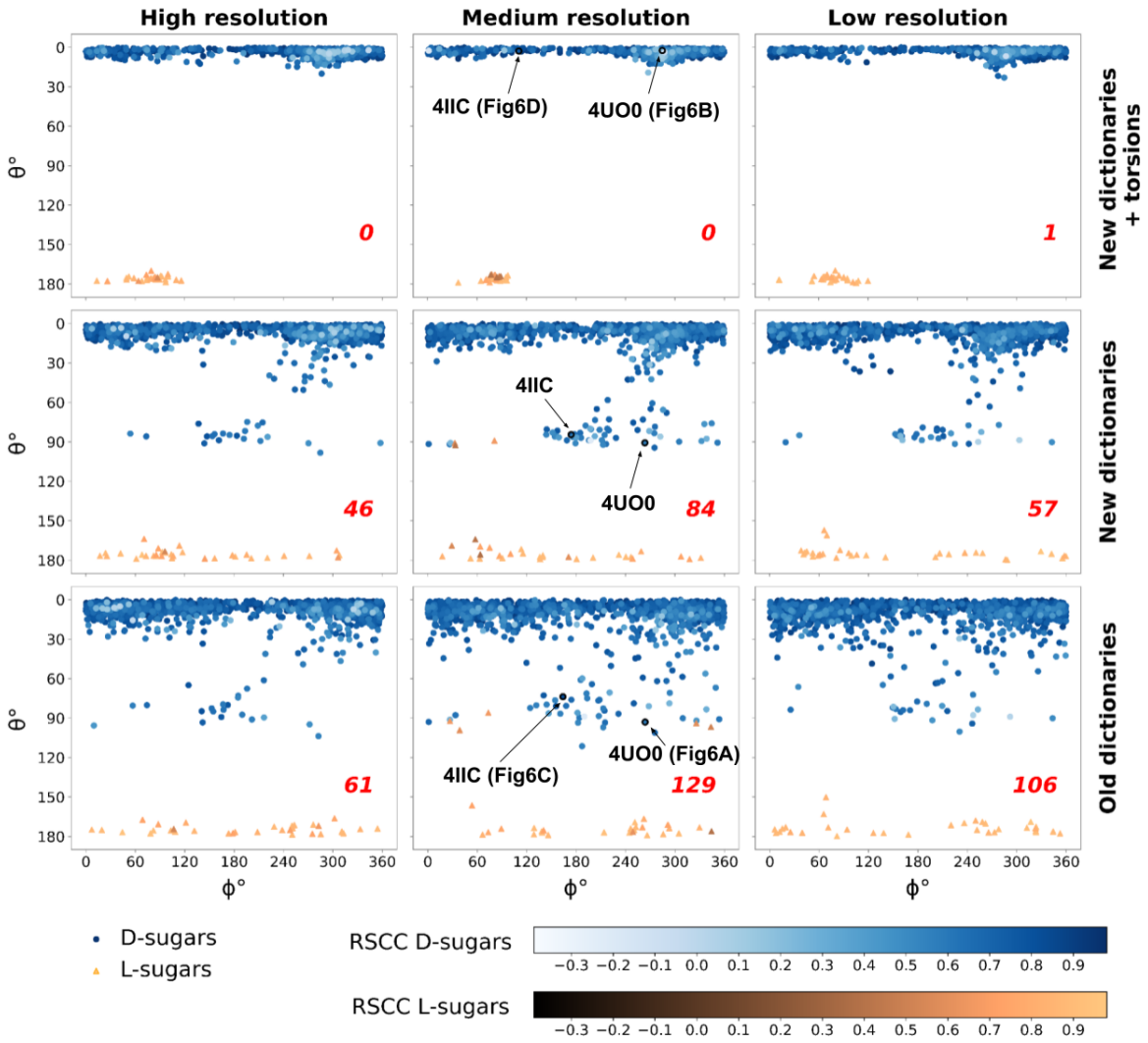
Three protocols were considered: using old dictionaries from the CCP4 Monomer Library, using new dictionaries generated using AceDRG, and using new dictionaries with the addition of unimodal torsion restraints. Using Privateer for validation, we have sought to gain insight into how the quality of pyranoside models differs when using different restraint dictionaries during refinement. It should be noted however that the power of ring conformation as a useful validation metric for pyranosides is hindered by the imposition of torsion restraints during refinement, as there is an obvious bias towards the lowest-energy conformation – other indicators (*e.g.* fit to density, average B-factor, linkage geometry) should be monitored instead in a production environment.

Figure 2.3A shows that the great majority of pyranosides in the test set were correct before and after all three refinement protocols. This was expected, as previous research has shown that modelling errors increase greatly with decreasing resolution, and particularly at resolutions lower than 2.0 Å (Atanasova, Bagdonas, and Agirre 2020). Figure 2.3A also shows that refinement with the old dictionaries produces very similar validation results to the original PDB models. This is somewhat surprising, as the original structures were produced using a variety of dictionaries and refinement software. Figure 3B eliminates from the picture all the structures that were correct in the original PDB models, and continue to remain correct when using all three refinement protocols. The remaining cases indicate that use of the old dictionaries reduces the number of residues validated as “no” by Privateer, and moves the majority of these to the “check” class (high-energy ring pucker, but no other pathologies). In contrast, use of the new dictionaries, use of the new dictionaries (without activating torsion restraints) resulted in a slight decrease in the number of sugars diagnosed with “check” , indicating that the updated geometric estimates in the new dictionaries are enough to sway some models from a high-energy ring pucker into a chair conformation. This

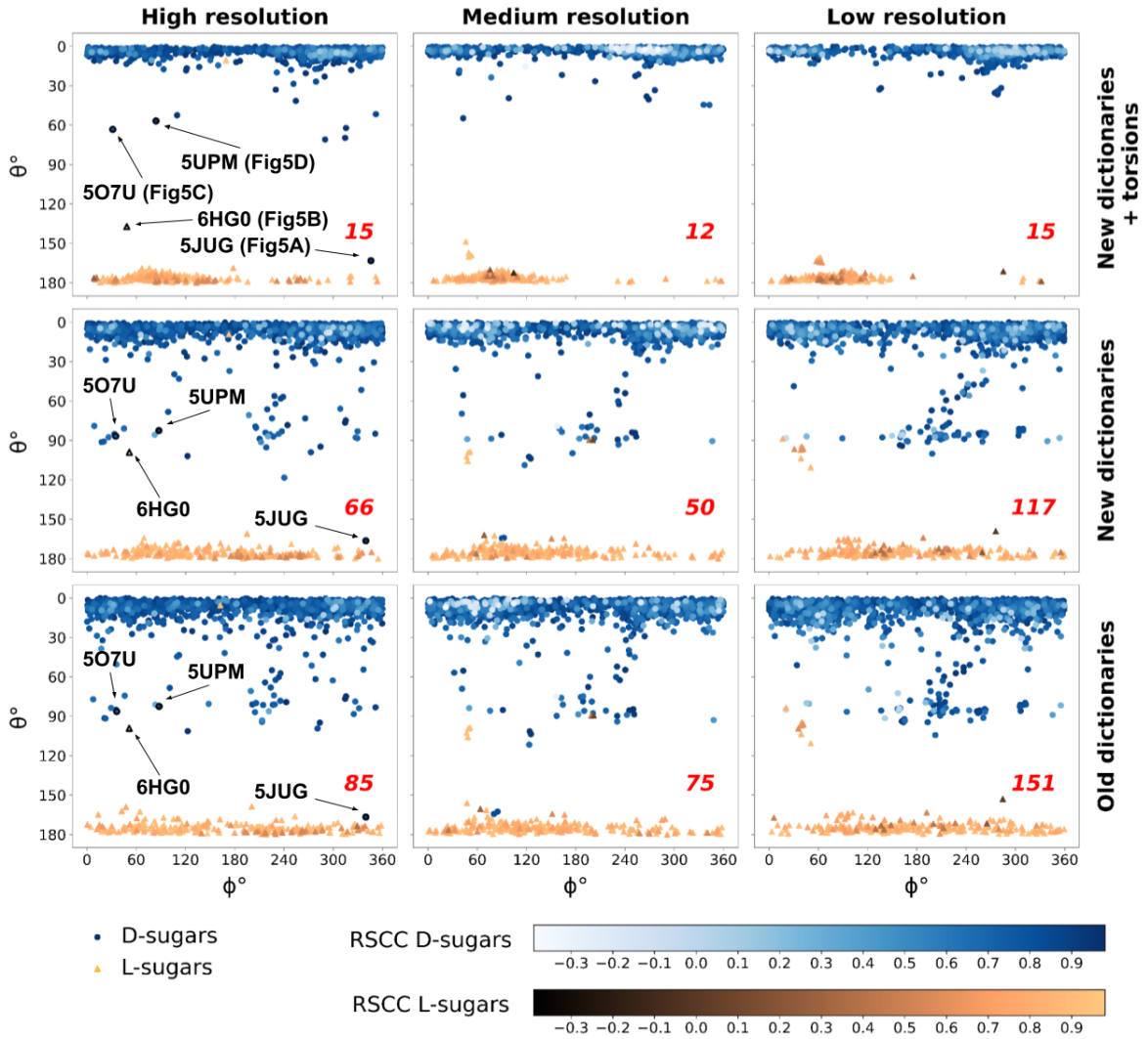
effect is greatly amplified when the new unimodal torsion restraints are activated; the number of pyranosides that change from showing one or more problems to being fully validated (a “yes” diagnostic) almost doubles.

Some pyranosides remain incorrect after refinement with all three protocols. Upon closer inspection, most of these are cases where the electron density is difficult to interpret, and often involve a modelling error. These cannot not be fixed by refinement alone, and in such cases, additional intervention – *e.g.* interactive real space refinement, or running advanced rebuilding protocols such as those in Rosetta (Frenz et al. 2019) or Coot (Paul Emsley and Crispin 2018), also automated in PDB-REDO (van Beusekom et al. 2019) – would be required in order to model the sugar correctly.

In addition to diagnosing each pyranoside as “yes”, “no” or “check”, Privateer also calculates the RSCC between the sugar-omitted “observed” ($2mF_o$-DF_c) electron density map (ρ_{obs}) and that calculated from the model (ρ_{calc}) in the vicinity of the sugar (Equation 2.1). Figure 2.4 shows the change in RSCC after refinement. The general trend is that RSCC remains high overall. The new unimodal torsion restraints lead to an increase in the number of sugars validated as “yes”, and a decrease in the number of sugars diagnosed with “no” or “check”. The average RSCC after refinement with the old dictionaries was 0.793; with the new dictionaries, it decreased to 0.791; finally, with the new dictionaries and unimodal torsion restraints, it went further down to 0.789 (Figures A.2-A.3). As already discussed, restraining a sugar to the most likely conformation could lead to a small decrease in RSCC. This is due to the sugar being encouraged to adopt a sensible conformation, rather than being allowed to sink into the electron density, however faint or incomplete, at the expense of unphysical geometric distortions. Such avoidance of overfitting is generally the appropriate course of action (in the absence of clear evidence to the contrary). Indeed, a modest reduction of RSCC should be seen as an acceptable trade-off when the electron density map does not unambiguously demonstrate evidence for a high-energy conformation. Consistently, we also found a small but significant increase in ΔR_{work} while ΔR_{free} remained essentially the same with both refinement protocols involving the new dictionaries. This reduction of the gap between R_{work} and R_{free} (Figure A.1) provides further evidence that using the new dictionaries can help prevent overfitting.


High resolution
Medium resolution
Low resolution
New dictionaries + torsions
New dictionaries
Old dictionaries
θ°
ϕ°
4IIC (Fig6D)
4UO0 (Fig6B)
4IIC
4UO0
4IIC (Fig6C)
4UO0 (Fig6A)
0
1
46
84
57
61
129
106
D-sugars
L-sugars
RSCC D-sugars
RSCC L-sugars
−0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9

(A)



(B)

Figure 2.4. Refinement with the new dictionaries and unimodal torsion restraints leads to fewer unlikely carbohydrate conformations. (A) Sugars part of N/O-glycosylation; (B) Other sugars. θ vs ϕ plot for D-sugars (blue circles) and L-sugars (yellow triangles) – see Discussion for a description of θ and the Cremer-Pople parameters. D-sugars usually adopt the $^4C_1$ conformation with $\theta \approx 0°$; L-sugars normally adopt $^1C_4$ conformation with $\theta \approx 180°$. Use of the new unimodal torsion restraints (top) shows fewer deviations from these values. The PDB codes corresponding to entries discussed in Figures 5 and 6 are labelled. The number of sugars in high energy conformations (according to Privateer) is shown in the bottom right corner of each plot. Resolution ranges contain equal numbers of sugars (1,668 each). High resolution is 0.9 to 1.8 Å, medium resolution is 1.8 to 1.9 Å and low resolution is 1.9 to 2 Å.

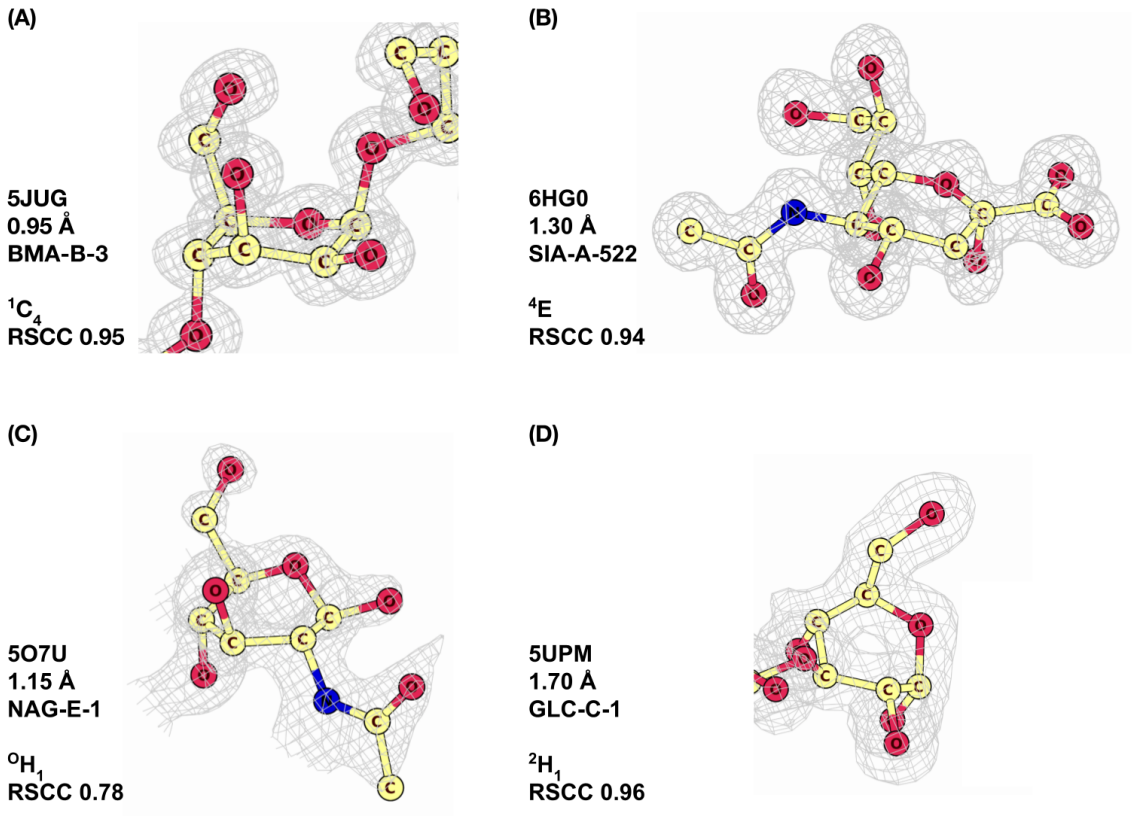


Figure 2.5. Sugars in unusual conformations after refinement with the new dictionaries with unimodal torsion restraints. (A) BMA-B-3 from PDB ID 5JUG (Y. Jin et al. 2016); (B) SIA-A-522 from PDB ID 6HG0 (Salinger, M.T., Hobbs, J.R., Murray, J.W., Laver, W.G., Kuhn, P., Garman, E.F., n.d.); (C) NAG-E-1 from PDB ID 5O7U (Tobola et al. 2018); (D) GLC-C-1 from PDB ID 5UPM (Pluvinage et al. 2017). These sugars appear as outliers in Figure 4(B). They remain in high-energy conformations after refinement, but have high RSCC. This figure was produced with CCP4mg (S. McNicholas et al. 2011). Map types are 2Fo-Fc, displayed at 1σ contour level with a sampling rate of 0.5.

The θ angle of the Cremer-Pople parameters for pyranose rings (Cremer and Pople 1975a) is a useful tool in conformational analysis, as it helps monitor the transition from chairs ($\theta \approx 0°$ for $^{4}C_{1}$ chairs, $\theta \approx 180°$ for $^{1}C_{4}$ chairs) into envelopes and half-chairs ($\theta \approx 45°$ and $\theta \approx 135°$) and then into boats and skew-boats ($\theta \approx 90°$). As these transitions involve eclipsing of substituents and thus energy penalties, θ may be seen as a simple summary of the deviation of a sugar's geometric parameters from ideal values. Examining the θ angle distribution provides further evidence to support the assertion that the unimodal torsion restraints decrease the number of unlikely conformations. Figure 2.4 (A and B) presents a conformational analysis of all pyranosides in the test data set, binned into three resolution

ranges (left: 0.9 Å to 1.8 Å; middle: 1.8 Å to 1.9 Å; and right: 1.9 Å to 2.0 Å). The three bins were chosen to contain the same number of pyranosides. As seen in both panels, the number of sugars with unusual θ angle values decreases significantly when the ring is restrained using the new unimodal torsion restraints. Even the new dictionaries without torsions activated seemed to have a beneficial impact on ring conformation. Interestingly, the ligand pyranosides that remain in unusual conformations (Figure 2.4B) generally exhibit a high RSCC. A closer inspection of these outliers (Figure 2.5) revealed that these pyranosides' conformations are retained due to being supported by the data, as evidenced by strong and featureful electron density maps. The relative weighting between the data and geometric components in REFMAC5 makes this possible, allowing for torsion restraints to be down-weighted in favour of strong observations (Murshudov et al. 2011). High-energy conformations such as these are usually adopted by ligands – bound, or trapped, covalently linked in the middle of a reaction within an enzyme (Davies, Planas, and Rovira 2012).

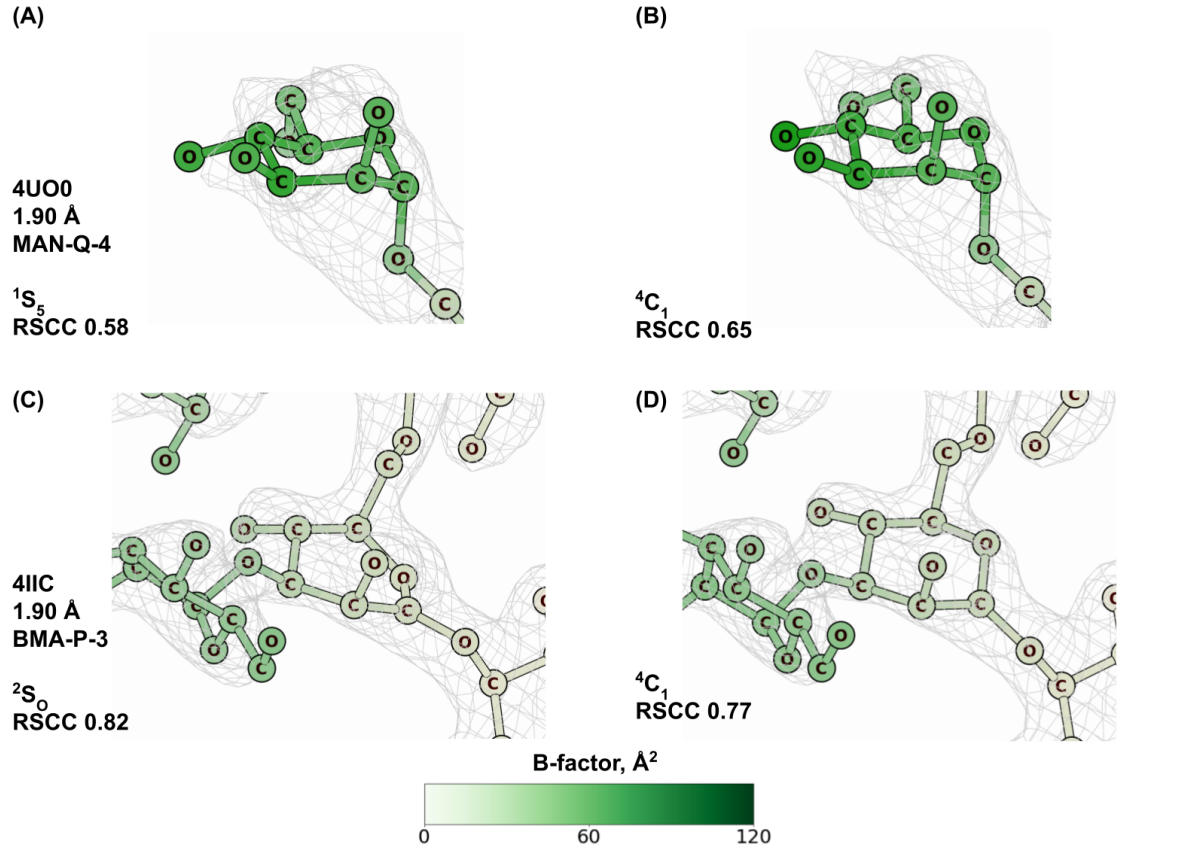


Figure 2.6. Change in conformation and real-space correlation coefficient (RSCC) after refinement. (A) Sugar in a $^{1}S_{5}$ conformation after refinement with its old CCP4-ML restraint dictionary entry (Figure 2.4A, bottom middle panel). (B) The conformation of the sugar has been changed to the minimal energy conformation after refinement with the updated restraint dictionary entry and unimodal torsion restraints and the RSCC has increased (Figure 2.4A,

top middle panel); the sugar in (A) and (B) is MAN-Q-4 from PDB ID 4UO0 (Devi et al. 2015) at 1.90 Å resolution, mean B value 34 Å^2. (C) Sugar in a ^2S_O conformation after refinement with its old CCP4-ML restraint dictionary entry (Figure 2.4A, bottom middle panel). (D) The minimal energy conformation of the sugar after refinement with the new restraint dictionary entry and unimodal torsion restraints; RSCC has decreased (Figure 2.4A, top middle panel). The sugar in (C) and (D) is BMA-P-3 from PDB ID 4IIC (Suzuki et al. 2013) at 1.90 Å resolution, mean B value 18 Å^2. This figure was produced with CCP4mg (S. McNicholas et al. 2011). Map types are 2Fo-Fc, displayed at 1σ contour level with a sampling rate of 0.5.

Figure 2.6 demonstrates the change in RSCC when restraining pyranosides to their most likely ring conformations using the new dictionaries with unimodal torsion restraints. A pyranoside refined with the old CCP4-ML dictionaries is shown in Figure 2.6A, adopting an unlikely high-energy conformation ($^1S_5$) with an RSCC of 0.58. When the conformation is moved to the more probable $^4C_1$ after refinement with the new dictionaries and unimodal torsion restraints (Figure 2.6B), the RSCC increases to 0.65, representing better agreement between the sugar and the electron density map. The sugar shown in Figure 2.6C is also in a high-energy conformation ($^2S_0$)) after refinement with the old dictionaries, which is corrected to $^4C_1$ after refinement with the new dictionaries and torsion restraints. However, in this case the RSCC decreases from 0.82 to 0.77. This once again demonstrates how restraining a sugar to the most likely conformation can have either an incremental or decremental effect on RSCC, and that RSCC is not always a helpful metric for assessing local model reliability. Finally, we should like to emphasise that while unimodal torsion restraints seem like a good tool for the refinement of pyranosides in general, they may mask out other problems that can be detected by Privateer when they are not in use. Indeed, an unexpected high-energy ring pucker is considered a good indication of other modelling problems (Agirre 2017).

2.1.5 Conclusion

As part of a recent overhaul of the CCP4 Monomer Library, in which existing dictionaries have been replaced with those generated by AceDRG, we have augmented the dictionaries for pyranose entries by patching them with unimodal torsion restraints generated by Privateer. This development has the potential to dramatically reduce the number of conformational anomalies in refined structures Users still have to be mindful of the need to activate torsion restraints in their respective refinement programs should they want to use them – they are currently deactivated by default in CCP4 software, though they may or may

not be used automatically in other suites. To the best of our knowledge, this is the first time torsional sets for pyranoses have been tested extensively in this manner and the results should give confidence that the torsion restraints in these dictionaries can lead to a chemically sensible result in the absence of serious modelling mistakes.

2.1.6 Open research data: availability and reproducibility

Our dictionaries will be released as part of the CCP4 suite with the 8.0 release. The latest pre-release version of the CCP4-ML can be accessed by anonymous checkout (command: `bzr checkout https://ccp4serv6.rc-harwell.ac.uk/anonscm/bzr/monomers/trunk mon_lib`). In addition to the new carbohydrate dictionaries, the previous CCP4 carbohydrate dictionaries (referred to in the text as 'old') at the time of publication and the results from all the refinements can be downloaded from: https://zenodo.org/record/5764924.

2.1.7 Acknowledgements

The authors are grateful to Jake Grimmett and Toby Darling from the MRC-LMB Scientific Computing Department and the Netherlands Cancer Institute's Research High-Performance Computing facility for support with computing resources, and to Garib Murshudov and Fei Long for valuable discussion and providing AceDRG, which was used for generating the dictionaries discussed herein. This work was supported by the Medical Research Council, as part of United Kingdom Research and Innovation (also known as UK Research and Innovation) [MRC file reference number MC_UP_A025_1012], by the European Union's Horizon 2020 research and innovation programme under grant agreement No 871037 (iNEXT-Discovery) and by CCP4. Mihaela Atanasova is funded by the UK Engineering and Physical Sciences Research Council [EPSRC, grant number EP/R513386/1]. Robert Nicholls is funded by the BBSRC [ref: BB/S007083/1]. Jon Agirre is the Royal Society Olga Kennard Research Fellow [award number UF160039].

2.2 Summary

The work presented in this chapter is an attempt to produce a set of updated carbohydrate dictionaries with improved torsion restraints for use in refinement software. The new dictionaries corrected errors present in the torsion restraints of the old dictionaries and also

introduced a new set of unimodal torsion restraints. These can be used to enforce a $^4C_1$ (for D-sugars) or $^1C_4$ (for L-sugars) conformation when refining carbohydrates. This is often necessary when working on an N-glycosylated model, as the resolution tends to be lower than that of non-glycosylated proteins. In the case of electron cryo-microscopy, the resolution is often even lower. The dictionaries presented have led to much lower numbers of monosaccharides with wrong conformations when used to refine a test set of carbohydrate structures. However, a small decrease in RSCC and increase in R-factor was observed, which is likely due to monosaccharides not being forced into a wrong conformation by inconclusive electron density. As per the publication referees' suggestions, glycans containing major modelling errors were excluded from the dataset, as these would not be fixed by refinement and as such are not of interest for this study.

The updated carbohydrate dictionaries have now been released as part of the CCP4 8.0 Monomer Library. Since then, they have been used to refine the N-glycans on human myeloperoxidase (MPO) at 2.6 Å resolution (Krawczyk et al. 2022). MPO oxidises organic compounds and as such has multiple roles in the body, such as in building cellular components, metabolism, immunity, inflammation and the N-glycans play a role in its enzymatic activity. Inhibiting MPO is of therapeutic interest for the treatment of major depressive disorder with inflammatory syndrome. Moreover, since release of the updated carbohydrate dictionaries, the crystallographic community has not reported any issues with monosaccharide restraints.

Chapter 3

Model Building

This chapter discusses a proof of concept piece of software, Sails, that applies a similar automated model building approach to that described for amino acids in proteins, and nucleic acids in the Buccaneer and Nautilus computer programs respectively. While the methodology for the identification of features is similar, additional considerations have been put in place for the particular case of pyranoses, such as the notion of ring conformation – $^4C_1$ for D-pyranosides, and $^1C_4$ for L-pyranosides – and added support for the different linkages.

Sails was originally written by K Cowtan and Jon Agirre. The aim of this chapter is to assess whether this methodology can be applied to carbohydrates as effectively as to proteins and nucleic acids. In order to do this, the database of fingerprints that Sails uses was extended to include all sugars commonly found in N-glycans (done by Mihaela Atanasova). The program was then tested on a set of structures at a range of resolutions to evaluate its performance (done by Mihaela Atanasova). Sails was shown to perform well at high resolution. At lower resolution it can only detect the well-resolved monosaccharides at the beginning of a glycosylation tree.

3.1 Introduction

Free glycans are used as signals in biological processes such as embryo development, defence responses and interactions between organisms, in addition to being used as structural components and for energy storage. For example, oligogalacturonides are free glycans that induce flower formation, but inhibit root growth in plants and also protect the plant from pathogen infections (Ferrari et al. 2013). In embryonic development, glycosaminoglycans (GAGs) play a significant role because the absence of certain enzymes necessary for GAG synthesis affects the ability of cells to migrate and tissues to form (Smock and Meijers 2018). Furthermore, chitin, a free glycan and a polymer of GlcNAc, is a structural component of cell walls of fungi, insect exoskeletons and some hard structures in fish (Moussian 2019). Also, glycans are part of the antigens that cover the surface of erythrocytes. These antigens are responsible for the categorization of the different blood

types as they need to be compatible for successful blood transfusions (Lee-Sundlov, Stowell, and Hoffmeister 2020). Moreover, cancer cells can be distinguished from non-cancerous cells by their glycan composition - malignant cells often have much more branched N-glycans on their surface. The reason for this is that sialyltransferases can be upregulated in cancer, which causes hyper-sialylation. The inhibition of this pathway is currently explored as potential cancer therapy (Dobie and Skropeta 2021). Additionally, the sialic acids on the surface of healthy cells can be targeted by pathogens. This has therapeutic potential too, as competitive inhibitors of viral sialidases can prevent infection with influenza (Glanz et al. 2018).

The complex biological roles of glycans, their therapeutic potential and industrial applications emphasise the need to have detailed knowledge of their structure. This is complicated by the fact that carbohydrates have complex structure and historically they have been overlooked by scientists. Unlike the protein backbone, carbohydrate chains can be branched and the linkages between monomers are more varied than the linkages between amino acids. The reason for this is that glycosidic linkages are normally formed between the anomeric carbon of one monosaccharide and an unmodified hydroxyl group on another one. There are multiple possibilities for which hydroxyl group on the second monosaccharide can be linked. Also, the anomeric carbon is a stereogenic centre, which means that different anomeric configurations are possible for the glycosidic bond. Branching arises when multiple monosaccharides or oligosaccharides are linked onto the hydroxyl groups of the same monosaccharide. Furthermore, the complex biosynthesis of glycans and the availability of glycan-processing enzymes can lead to different glycan products formed on the same protein core known as glycoforms (P. M. Rudd and Dwek 1997; Fisher et al. 2019). This is referred to as heterogeneity. Microheterogeneity is when on the same glycosylation site there are different glycans attached. Macroheterogeneity is when different glycosylation sites are glycosylated. Moreover, monosaccharides have complex stereochemistry and conformation. Long carbohydrate chains, such as those found in protein N-glycosylation, tend to be flexible and, since they are usually on the surface of proteins, they are in contact with the water molecules surrounding the protein. All of these factors make carbohydrates difficult to resolve with X-ray crystallography, which involves taking a snapshot of the protein. As a result, the mean resolution of glycosylated proteins (2.4 Å) tends to be lower than that of non-glycosylated proteins (2 Å) (van Beusekom, Lütteke, and Joosten 2018).

In practice, monosaccharides, especially those part of N-glycans, tend to adopt the lowest-energy conformation possible, which for pyranoses is chair - $^{4}C_{1}$ for D-monosaccharides and $^{1}C_{4}$ for L-monosaccharides. In chair conformations the substituents are staggered, the angles between them are 60/-60 degree, which minimises repulsions and

strain making the conformation low energy. Going from $^4C_1$ to $^1C_4$ or vice versa requires high activation energy, which usually involves catalysis with a carbohydrate-active enzyme. Deviations from the lowest-energy conformation are usually in the active sites of enzymes and need to be clearly supported by data.

Several software tools that can help when building carbohydrate models are currently available. Coot (P. Emsley et al. 2010), a software commonly used in structural biology for the display and manipulation of models, has an N-glycan building feature (Paul Emsley and Crispin 2018). It allows the user to build an N-glycan model in three modes - manual, semi-automatic and automatic. The manual mode requires the user to select a carbohydrate monomer and a type of linkage from a range of possible options; Coot places it in the map in the correct orientation and conformation. The semi-automatic mode involves the user choosing just the glycan type and Coot suggesting possible monomer and linkage options. The automatic mode is the easiest to use, with the user just selecting the starting point and the glycosylation tree type. The limitations of the N-glycan building feature of Coot is the narrow selection of glycoforms available and the trees built in the automatic mode are often incomplete if the resolution is low.

Van Beusekom et al. (van Beusekom et al. 2019) have built on this work by creating two software tools that automate further the N-glycan building process. The first tool, Carbivore, allows for the automated rebuilding and/or extension of existing carbohydrate trees and building new trees through the use of Coot's N-glycan building feature. The second tool, Carbonanza, allows for the creation of missing links between Asn and N-acetyl-D-glucosamine when N-glycosylation was not detected. Both of these programs are currently available through the PDB-REDO re-building and re-refinement pipeline.

The ISOLDE (Tristan Ian Croll 2018) plugin for ChimeraX (Pettersen et al. 2021) is an immersive environment that facilitates model building into low to medium resolution maps by providing real-time validation. While ISOLDE does not provide model building itself, it uses molecular dynamics force fields to fix errors automatically as the user makes changes to the model. It works with carbohydrates through the use of the GLYCAM force field (Kirschner et al. 2008).

Overall, the software available for building carbohydrates is not as featureful as the software available for building proteins. For this reason, we are proposing a new piece of software, Sails, that allows for the automated building of N-glycans and ligands in X-ray crystallography or electron cryo-microscopy maps.

3.2 Method

Sails (Software for Automatic Identification of Linked Sugars) comprises a set of programs for building carbohydrates into electron density (MX) or electron potential (cryo-EM) maps. Its algorithm is based on fingerprint detection, similar to the algorithm of Nautilus, which can build models of nucleic acids (Kevin Cowtan 2014). The detection engine relies on a database of fingerprints, which are used to detect monosaccharides on the map. These fingerprints are generated by superimposing monosaccharide from structures in the PDB in the same conformation, usually their minimal-energy conformation. High probe points (peaks) are then placed where electron density/potential is high (i.e. on each atom), and low probe points (voids) are placed where the electron density/potential is empty (i.e. in the surroundings). Examples of fingerprints generated by Sails can be seen in Figure 1. After a six-dimensional search, each fragment from the database is then scored against the electron density/potential map. The high probe points of the fingerprint need to match places of high electron density/potential on the map. The low probe points on the fingerprint need to match the empty electron density/potential on the map. A score is then calculated based on how well the fingerprint matches the electron density/potential for each scanned position. The fingerprints are then placed on the map starting from those with the highest score until reaching a threshold score.

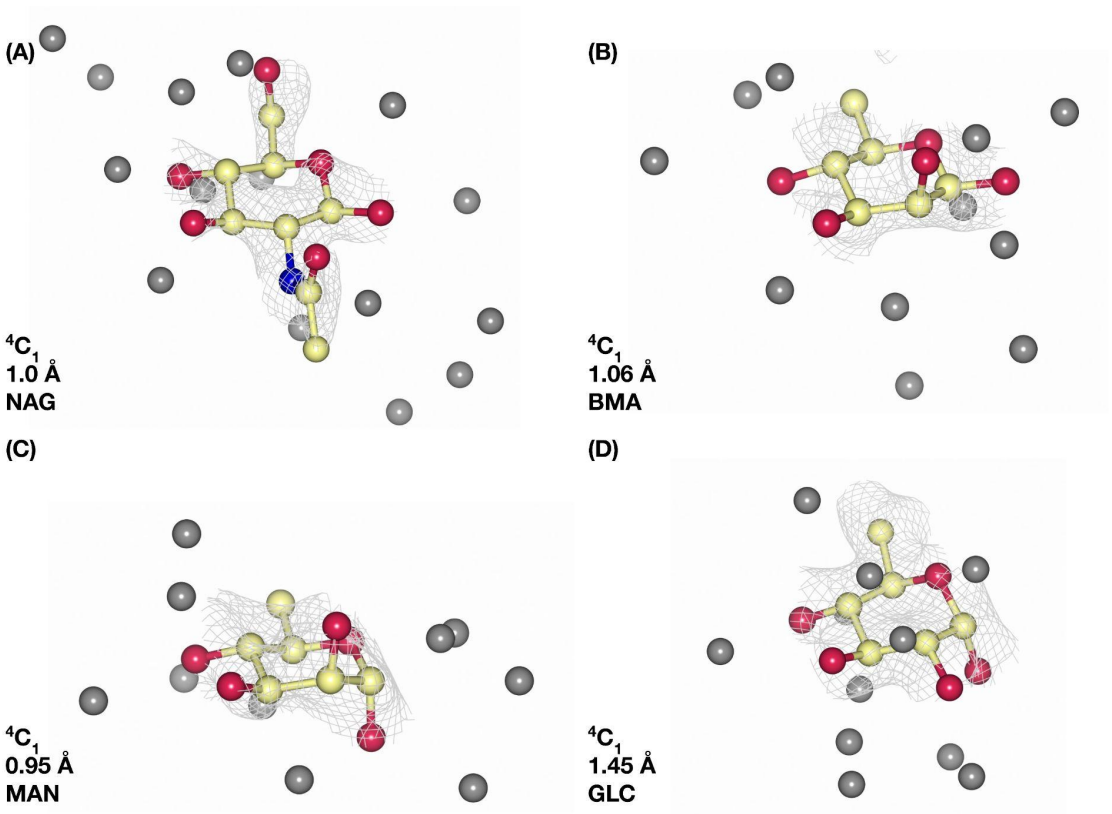


Figure 3.1. A set of fingerprints generated with Sails. Positive probe points are placed on the atoms and negative probe points are placed in the voids. (A) N-acetyl-β-D-glucosamine (NAG), (B) β-D-mannose (BMA), (C) α-D-mannose (MAN), (D) α-D-glucose (GLC). All monosaccharides depicted are in $^4C_1$ conformation. Map types are 2Fo-Fc, displayed at 1σ contour level with a sampling rate of 0.5. The maps were calculated with the Sails minimax tool. This figure was produced with CCP4mg (S. McNicholas et al. 2011).

3.2.1 Fingerprints creation

Design guidelines

The fingerprint database is a major component of Sails and determines how well it works. For this reason it was important to ensure that its quality is good. There are several considerations that need to be made. First of all, the structures selected for fingerprinting need to be high resolution and the monosaccharides contained in them need to be correct. Moreover, the number of monosaccharides used for the creation of one fingerprint should be optimal for best results. In fact, often one high resolution structure with a few copies of the

same monosaccharides is sufficient to produce a satisfactory fingerprint for that monosaccharide.

Structure selection

In order to find structures suitable for fingerprint generation, Privateer was run on all of the .pdb files of a mirror of the Protein Data Bank. This produced a list of all protein structures which contain monosaccharides along with validation information for each monosaccharide, including its anomeric configuration, conformation, θ and φ angles, and Privateer's assessment of whether the monosaccharide is correct. This data was used to select structures for fingerprinting based on several criteria: number of a given monosaccharide present in the structure, correctness of these monosaccharides, resolution. Ideally to be selected for fingerprinting, a protein should contain a few copies of a particular monosaccharide, all validated as correct by Privateer, and the data should be high resolution. Having too many copies of the same monosaccharide leads to the fingerprint becoming less defined, often resulting in unresolved atoms which can then lead to false positive monosaccharide detection. Having too few monosaccharides in a structure often leads to the fingerprint being too specific to a particular structure and not being able to generalise well. If a structure does not contain enough monosaccharides, multiple structures can be combined using one of Sails' tools specifically designed for this. This is especially useful for monosaccharides for which less data are available. A good fingerprint created at high resolution should work well even at lower resolution.

Data preparation for fingerprinting

To prepare a structure for fingerprinting it was refined to ensure that the agreement between the model and the data is as close as possible. The .mtz and .pdb files were input into REFMAC5 which was then run for 20 cycles with the new carbohydrate dictionaries described in Chapter 2. The output .mtz and .pdb files were input into the Sails minimax tool. This tool uses the .mtz and .pdb data to produce a minimum and maximum map for a specified monosaccharide in a particular conformation, along with a .pdb file containing just the monosaccharide. The minimum and maximum electron density/potential maps are calculated. Sails also includes a tool that allows combining maps calculated from multiple structures. The user first inputs the map to be used as the base, then the map to be added to it. The new map, calculated as the average of the two maps, is output and can be used as input for the fingerprinting tool.

Fingerprinting

The fingerprinting tool takes the .pdb file, minimum and maximum maps of the specified monosaccharide output by the minimax tool. It also allows the user to change the mask radius and map radius. The mask radius controls how spread out the negative probe points are. The map radius sets the radius of the map used. The mask radius and map radius that give the best results depend on the monosaccharide being fingerprinted and on the data used for fingerprinting. Generally, values of 2-3 Å for the mask radius and about 9 Å for the map radius seem to give the best results. However, these values should be optimised for a particular fingerprint. In order to avoid putting a void where one already exists, Sails calculates where they should be, rather than placing them randomly. First, a dummy atom at a random position is placed and a place of empty electron density/potential is found near it. A void is placed there and then the next place of empty electron density/potential is found. The calculated fingerprint is given as an output to the console, from where it can be copied into the Sails fingerprint database. The fingerprint also contains information about what type of carbohydrate it should be used for, ie. ligand, N-glycan, etc. The protocols used for fingerprint generation are given in Table 3.1, including PDB codes for the structures used for each monosaccharide, map and mask radius in Å.

Table 3.1. Protocols for generation of the fingerprints in Sails' internal database. Column legend: 'Sugar' is the three-letter CCD component ID assigned to each monosaccharide; 'PDB' is the PDB codes for the structures used to generate the fingerprints; 'Resolution' is the resolution in Å; 'Number of monosaccharides' is the number of copies of the monosaccharide being fingerprinted in the structure; 'Mask radius' and 'Map radius' are the values for these properties used as input for Sails' fingerprinting tool; mask radius controls how spread out the negative probe points are and map radius controls the radius of the map used for fingerprinting. All structures used for fingerprinting were determined using X-ray crystallography.

Sugar	PDB	Resolution, Å	Number of monosaccharides	Mask radius, Å	Map radius, Å
NAG	3bwh	1.00	2	2.3	9.5
MAN	5o2x	0.95	15	2.4	9
BMA	1pmh	1.06	5	2.4	9.5
GLC	3weo	1.45	3	2.3	6
NDG	1q4g	2.00	4	2.5	9.5
XYP	5lal	1.40	3	2	8
	1e6s	1.35	2		
	6rs9	1.40	2		
	3bwh	1.00	1		
	5aog	1.27	1		
	6idn	1.50	1		
FUC	7c38	1.20	4	3	9.5
FUL	7c38	1.20	4	2.5	11
GAL	5elb	1.08	18	2.5	9.5
BGC	3pfz	1.10	6	2.5	9.5

Running Sails

Sails can work with two different inputs. The first one is a .pdb and an .mtz file for a structure, the second is a .map file. .pdb and an .mtz files are output by REFMAC5 after the protein is refined. Map files can be computed by Privateer and REFMAC5, or they can be electron cryo-microscopy maps. Sails performs a search for each monosaccharide across the map and calculates a score for each position checked. It then removes all the monosaccharides below a certain score threshold. Afterwards, Sails places the remaining monosaccharides starting from the highest-scoring ones first. For each of these monosaccharides, Sails retrieves its dictionary definition from the CCP4 Monomer Library and positions it in the detected location. Sails also makes sure that if a monosaccharide has already been placed in a certain location, no other monosaccharide is being placed on top. Then, Sails outputs a .pdb file containing the monosaccharides found. A difference map can also be used. This speeds up the search, because there is less electron density for Sails to search across. In addition, the step at which the angular sampling is carried out can be changed by the user (the default is 15°). Having a very high step leads to fewer monosaccharides being found, but a very low step can significantly slow down the process.

3.2.2 Testing

Sails was tested on structures that were not used for fingerprint generation. A test set of 10 high resolution (0 - 1.50 Å), 10 medium resolution (1.51 - 2.50 Å) and 10 low resolution (2.51 - 4.00 Å) X-ray crystallography structures was chosen based on having the highest number of the monosaccharides for which fingerprints were available. In addition, 10 structures determined with cryo-EM were also selected with resolution of 2.50 - 4.00 Å and high numbers of monosaccharides. The resolution ranges were chosen to reflect the overall lower resolution for carbohydrate-containing proteins. The PDB codes for the structures are given in Table 3.2. The test set is small, because its purpose is to show the proof-of-concept and to highlight what improvements need to be made to Sails. Each structure was first run through REFMAC5. Then, the pdb and mtz files output by REFMAC5 were input into Sails. In addition, the PDB model was validated with Privateer. The monosaccharides built by Sails were visually inspected in Coot.

Table 3.2. PDB codes of structures used for testing Sails. Column legend: 'the X-ray crystallography' columns contain PDB codes of structures originally determined with X-ray crystallography; 'High resolution' contains structures in the 0-1.50 Å resolution range, 'Medium resolution' - in the 1.51 - 2.50 Å resolution range and 'Low resolution' - in the 2.51 - 4.00 resolution range; the 'Electron cryo-microscopy' column contains PDB codes of structures originally determined with cryo-EM in the 2.50 - 4.00 Å resolution range; 'PDB' refers to the pdb code of the structure being modelled; 'Rsln' is the resolution in Å; 'No.' is number of monosaccharides the structure contains.

X-ray crystallography									**Electron cryo-microscopy**		
High resolution			**Medium resolution**			**Low resolution**					
PDB	**Rsln**	**No.**	**PDB**	**Rsln**	**No.**	**PDB**	**Rsln**	**No.**	**PDB**	**Rsln**	**No.**
6hgb	1.50	41	5fjj	1.95	157	5d9q	4.40	216	5szs	3.40	222
1e6q	1.35	20	5fji	1.95	91	5ju6	2.20	184	6cde	3.80	219
4h53	1.50	35	4iih	2.00	90	1zpu	2.80	166	6bfu	3.50	168
2q9o	1.30	34	4iib	1.80	86	5fyj	3.11	162	6nb3	3.50	160
6gsz	1.38	26	4iic	1.90	84	4nco	4.70	147	6myy	3.80	156
3og2	1.20	23	4iie	2.00	85	5fyk	3.11	130	6nqd	3.90	141
3ogr	1.50	21	4dl1	2.00	68	5i8h	4.30	116	6dcq	3.10	133
2hox	1.40	21	5fk8	1.88	68	5fyl	3.10	114	5vn3	3.70	129
1e4m	1.20	20	4uo0	1.90	70	6crd	2.57	112	6dg7	3.32	35
1e73	1.50	20	5fmd	1.78	66	5es4	3.30	112	6hug	3.10	32

3.3 Results

A set of 10 monosaccharide fingerprints were produced covering monosaccharides frequently involved in N-glycosylation, ie. NAG, MAN, BMA, NDG, GLC, GAL, BGC, FUC, FUL, XYP. The fingerprints were produced at their lowest energy conformation, as this is the

most common conformation observed in N-glycosylation. Higher energy conformations are more frequently observed in ligands. Enforcing the lowest energy conformation is especially important when building monosaccharides at low resolution, as it is often impossible to distinguish the conformation. This could sometimes come at the price of higher R-factor and B-factor, and lower RSCC, but unless the electron density/potential clearly indicates that the monosaccharide is in a higher energy conformation, it should be kept at a low energy conformation.

The fingerprints were input into Sails' internal database. Sails was then run on a set of test structures, which includes 10 high resolution, 10 medium resolution and 10 low resolution X-ray crystallography structures, in addition to 10 cryo-EM structures.

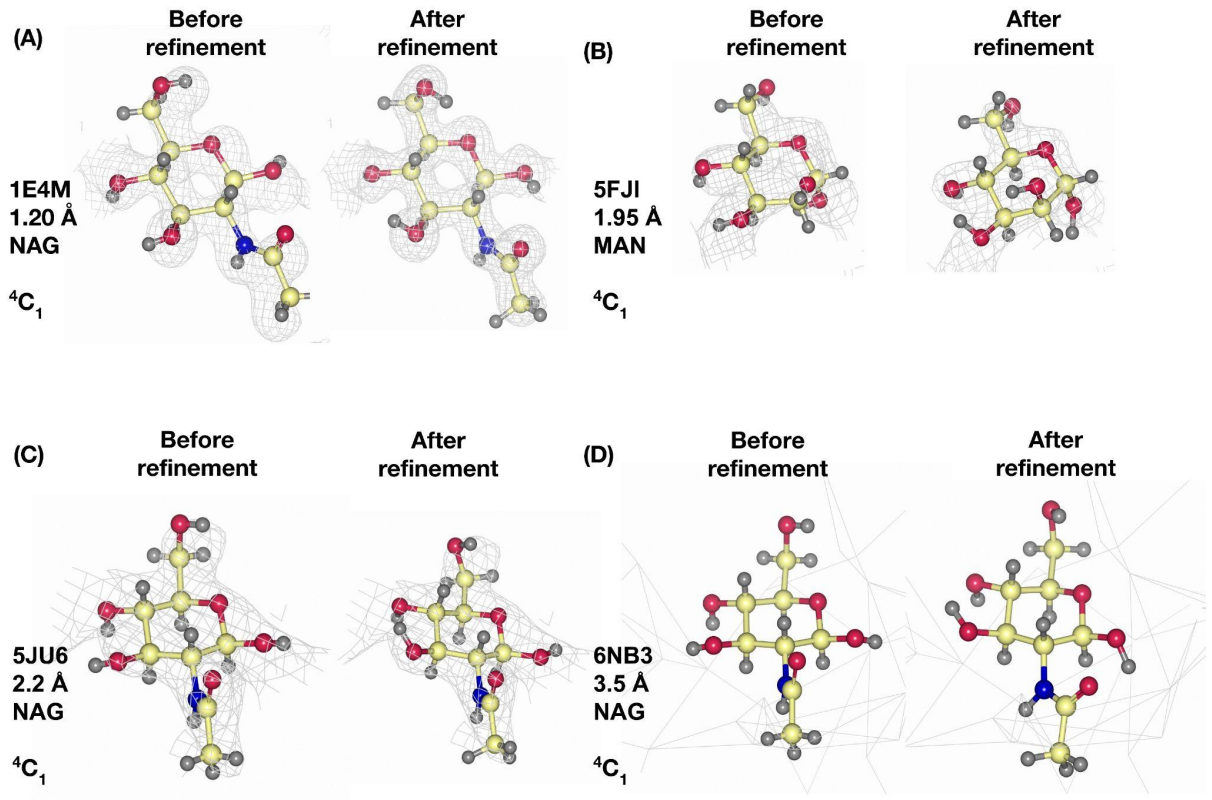


Figure 3.2. Monosaccharides built by Sails at different resolutions before refinement (left) and after refinement with Coot's Real Space Refinement using the new carbohydrate dictionaries with unimodal torsion restraints (right). (A) N-acetyl-β-D-glucosamine (NAG) built at 1.20 Å resolution; PDB entry 1E4M (Burmeister et al. 2000) determined with X-ray crystallography. (B) α-D-mannose (MAN) built at 1.95 Å resolution; PDB entry 5FJI (Agirre et

al. 2016) determined with X-ray crystallography. (C) NAG built at 2.2 Å resolution; PDB entry 5JU6 (Gudmundsson et al. 2016) determined with X-ray crystallography. (D) NAG built at 3.5 Å resolution; PDB entry 6NB3 (Walls et al. 2019) determined with electron cryo-microscopy. Map types for the X-ray crystallography structures (A-C) are 2Fo-Fc, displayed at 1σ contour level with a sampling rate of 0.5. Cryo-EM maps in (D) displayed at contour level 1. This figure was produced with CCP4mg (S. McNicholas et al. 2011).

The test set structures built by Sails were inspected in Coot, looking at how well the monosaccharide fits the electron density/potential and comparing it to the PDB model. The structures deposited in the PDB were run through Privateer in order to obtain the total number of monosaccharides in the deposited structures. The results of the models built by Sails and the models from the PDB were compared and the percentages of correctly built monosaccharides for different resolution ranges are presented in Table 3.3. Table 3.3 also shows data on the first NAGs, ie. NAGs at the beginning of a glycosylation tree, built correctly by Sails. The results showed that Sails can correctly detect up to 40% of the monosaccharides present in a structure at sufficiently high resolution. Examples of monosaccharides built correctly by Sails at a range of resolutions are shown in Figure 3.2. These monosaccharides were successfully refined with Coot's Real Space Refinement.

Table 3.3. Percentages of correctly identified monosaccharides by Sails at different resolution ranges. 'High resolution' are structures in the 0-1.50 Å resolution range, 'Medium resolution' - in the 1.51 - 2.50 Å resolution range and 'Low resolution' - in the 2.51 - 4.00 resolution range; the 'Electron cryo-microscopy' structures are in the 2.50 - 4.00 Å resolution range.

		Percentage of correctly identified monosaccharides	**Percentage of correctly identified first NAGs**
X-ray crystallography	**High resolution**	39.21%	71.42%
	Medium resolution	11.37%	37.50%
	Low resolution	6.84%	1.63%
Electron cryo-microscopy		6.15%	0.28%

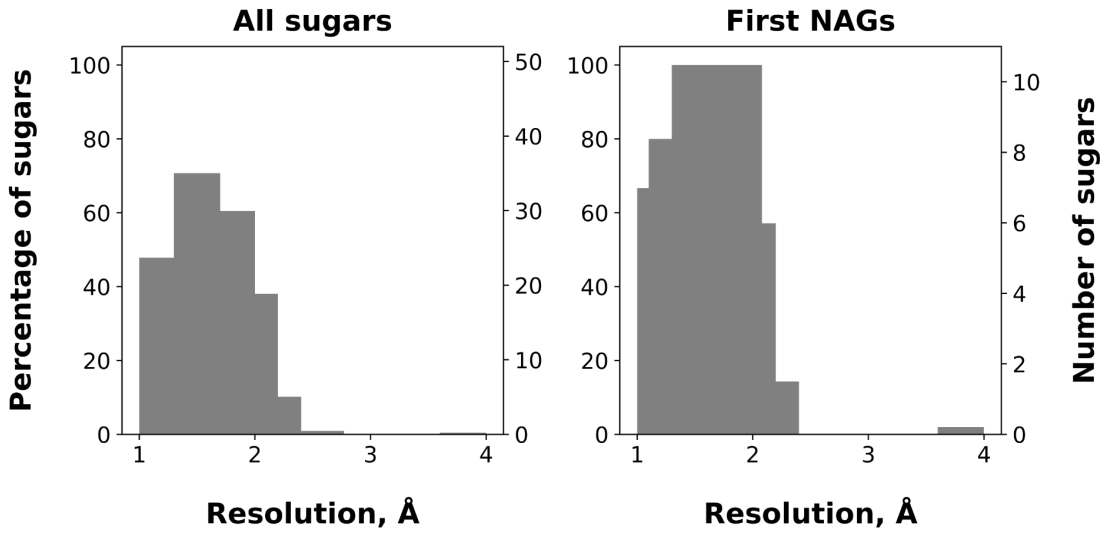


Figure 3.3. Percentage of monosaccharides part of N-glycan trees detected correctly by the Sails software vs resolution (Å); structures deposited in the PDB were used for comparison. The number of monosaccharides detected by Sails was obtained by inspecting the structures visually in Coot. The number of monosaccharides in PDB structures was obtained with Privateer. Left: all monosaccharides detected in N-glycan trees; right: N-acetyl-D-glucosamine residues detected at the beginning of N-glycan trees.

3.4 Discussion

The possibility to build carbohydrate models into electron density/potential maps automatically using a similar approach to the nucleic acid model building software Nautilus was assessed. This approach differs from the approach often used for proteins, where the amino acid sequence is available and can be traced on the electron density/potential map. The lack of sequence data for carbohydrates means that building carbohydrates can be significantly more complicated. Unlike with protein, nucleic acids and ligands, when a structure is being built, often there is no prior knowledge on what carbohydrates it should contain. The fingerprinting approach offers a solution to this problem. However, the main factor for the success of this approach is the quality of the fingerprints.

The main hurdle for producing high quality fingerprints is the availability of data. The generation of a good fingerprint requires a few copies of the monosaccharide at high resolution. These can be all on the same protein, or combined from multiple structures. The monosaccharides used for the fingerprint creation also all need to be correct and in the same conformation, ie. the lowest energy conformation for N-glycosylation. Moreover, each monosaccharide needs to be well resolved. Structures with long glycan chains can be quite flexible, which causes the structures at the end of the glycan chains to be poorly resolved. These monosaccharides can make the fingerprint less defined by having fewer peaks than non-hydrogen atoms. This causes the fingerprint to be able to fit into more positions on the map, which leads to more false positive results.

The fingerprints produced cover the most commonly found monosaccharides in N-glycosylated structures in the PDB. The availability of data means that the fingerprints for NAG, MAN and BMA are much better than the rest of them. In fact, a large number of the monosaccharides built by Sails incorrectly (Figure 3.3) are erroneously placed small monosaccharides like GLC and XYP. Despite this, Sails is able to build a number of monosaccharides correctly, in particular NAG, MAN, BMA, at high and medium resolution. The monosaccharides Sails is able to build successfully tend to be well resolved. Less resolved monosaccharides, where the electron density/potential is partly missing are unlikely to be built by Sails, but these are usually cases that require manual building often with additional knowledge. Moreover, Sails can sometimes place smaller monosaccharides, like XYP, in place of larger ones, like NAG. This happens because the small monosaccharides can have a higher score for that particular position. The reason for this is that part of the score is based on how well the peaks and the voids fit into and around the electron density/potential of the monosaccharide. The more peaks and voids a monosaccharide fingerprint has, the lower its overall score tends to be, ie. larger monosaccharides tend to have lower scores even when placed correctly. However, this issue is overcome simply by building the larger monosaccharides first and then moving on to the smaller ones. Furthermore, sometimes Sails places the wrong anomer. This is not unexpected, as the difference between two anomers is quite small.

Figure 3.2 shows examples of monosaccharides built by Sails that fit the electron density/potential well at a range of resolutions. In Figure 3.2(A) it is clear that the NAG fits well into the electron density map and the individual features of the monosaccharide are well resolved. In Figure 3.2(B-C) the resolution of the electron density map is lower, but it is still possible to distinguish the monosaccharides well. In Figure 3.2(D) the low resolution of the electron potential map makes identifying anything challenging. Also, after building them with Sails, these monosaccharides were refined with Coot's Real Space Refinement. The

refinement makes the monosaccharides fit the electron density/potential even better, which is seen in the higher resolution examples A-C.

Another current limitation of Sails is that the monosaccharides built are not linked into chains. This is a feature that will be added in the future, but currently it means that the models cannot be fully refined and validated. As the monosaccharides built into the model have idealised coordinates and have not been put through refinement software, their puckering amplitude, torsion angles and bonds are all correct, ie, they are the ones defined in the monosaccharide dictionaries. Moreover, the conformation is always the low-energy chair conformation usually found in N-glycosylated structures.

Another thing to consider when using Sails is the angle of the step used to scan the electron density/potential map. The lower the angle is, the more potential positions of monosaccharides are scanned. A step higher than 20° is likely to cause Sails to miss monosaccharides. However, a lower step, usually about 5°, often leads to lots of monosaccharides being built incorrectly across the map. Moreover, lowering the step can make Sails much slower. This also depends on the size of the map - a bigger map takes much longer to be scanned than a smaller one. A step of 7.5 deg seems to work well for a number of structures without slowing down the carbohydrate model building process too much.

Sails works with both X-ray crystallography and cryo-EM maps as input. However, because of the challenges of resolution, Sails does not perform well on cryo-EM maps. In fact, it does comparably to when building carbohydrates into low resolution X-ray crystallography maps. For both low resolution X-ray crystallography and cryo-EM very few monosaccharides were detected correctly and a few were flipped upside down. The few that were detected correctly were found in higher resolution maps. This could be because the low resolution means that it is difficult to distinguish any of the features of the monosaccharide, even longer substituents like the amine on NAG. The large blobs of electron density/potential at low resolution mean that it is difficult to fit the voids onto most of the map. Some insight into this could be gained if more detailed studies are carried out on the mistakes that Sails makes at different resolution ranges. Data could be collected on how likely each monosaccharide is to be built correctly, built in a wrong location or built in a correct location but rotated the wrong way. For the majority of low resolution structures, no monosaccharides were detected using the current set of fingerprints.

Generally, when building a biological macromolecule model, carbohydrate modelling would take place after the protein has been built and refined. For this reason, Sails would have to be combined iteratively with Refmac5 in a similar manner as in protein model building

pipelines. After that, validation of the model would take place. For the carbohydrate part, this would be done with Privateer.

3.5 Conclusion and future work

Multiple software tools have been developed for building protein or nucleic acid models into electron density/potential maps. Two of these tools, BUCCANEER for building proteins and Nautilus for building nucleic acids, make use of a method based on fingerprint detection. This method has been evaluated for the use for carbohydrate model building. The software presented, Sails, uses a database of monosaccharide fingerprints to build N-glycan chains. The successful detection of monosaccharides by Sails greatly depends on resolution. At high resolution Sails can detect up to 70% of the monosaccharides in a structure. However, at low resolution it is not uncommon for Sails not to detect any monosaccharides.

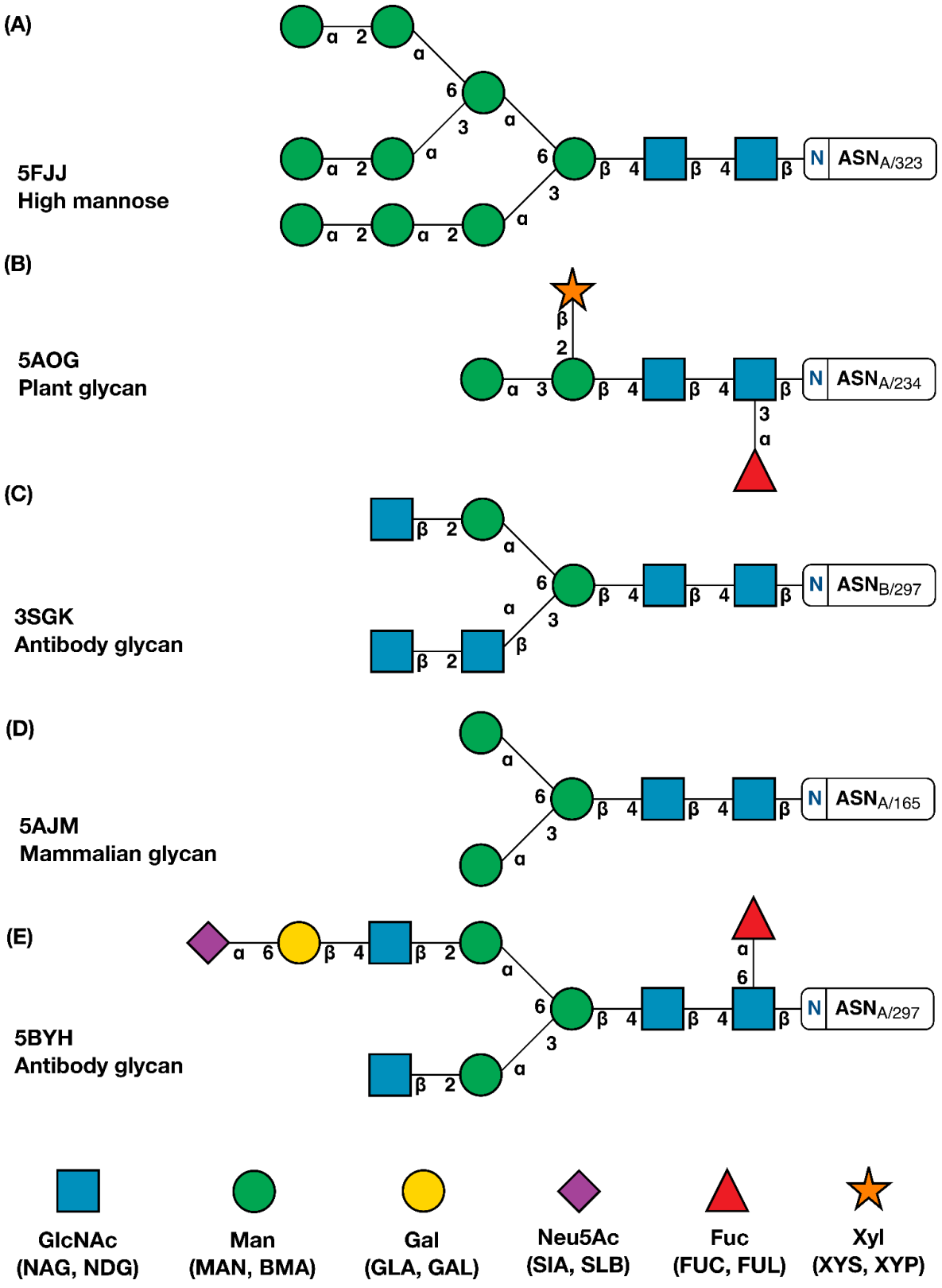


Figure 3.4. Examples of different types of N-glycans shown using the Symbol Nomenclature for Glycans. The greek letters and numbers show the N-glycan linkages naming. (A) High mannose from PDB entry 5FJJ (Agirre et al. 2016). (B) Plant glycan from PDB entry 5AOG

(Nnamchi et al. 2016). (C) Antibody glycan from PDB entry 3SGK (Ferrara et al. 2011). (D) Mammalian glycan from PDB entry 5AJM (Xiong et al. 2014). (E) Antibody glycan from PDB entry 5BYH (Yang et al. 2015). This figure was produced with Privateer.

To improve N-glycan tree building, it is possible to exploit the fact that Sails can detect the first GlcNAc of a tree generally more successfully than it can detect the monosaccharides further down the tree. The reason for this is that the first GlcNAc is usually better resolved than the more flexible monosaccharides towards the end of the tree. As seen in Figure 3.4, different types of N-glycans all start with a GlcNAc. If that GlcNAc is detected successfully, it can serve as a starting point for building the rest of the tree using idealised monosaccharides. The composition of the tree can be obtained from glycomics databases using Privateer (Bagdonas, Ungar, and Agirre 2020b). This information could be used in a similar manner as the amino acid sequence when building proteins or the nucleotide sequence when building nucleic acids. The electron density/potential should still inform the specific three dimensional structure of the tree.

Another issue of Sails that should be addressed in the future is that it currently produces models containing unlinked monosaccharides. These monosaccharides should be linked into N-glycan trees. On the electron density/potential maps the linkages can be flexible and badly resolved, but information about them will be available in Privateer (Dialpuri et al., in preparation). This data will contain linkage torsion angles collected from a large set of validated structures of N-glycan trees. This information will be especially useful when building trees based on glycomics data when the electron density/potential is unclear.

3.6 Availability

Sails is available on GitHub at:
https://github.com/glycojones/sails

The results from testing Sails can be downloaded from:

https://zenodo.org/record/7116530

3.7 Summary

This chapter introduces a software tool, Sails, for building N-glycosylation trees into electron density/potential maps. Sails is moderately successful at detecting monosaccharides at high-medium resolution and not very good at low resolution. However, Sails is very successful at detecting the first GlcNAc of N-glycan trees, because it normally has higher resolution than the rest of the protein. This offers the potential for future development of Sails into being able to build idealised trees using the detected first GlcNAc hit and information from glycomic databases collected with Privateer. Furthermore, Sails should be extended to als build the linkages between the monosaccharides. Data on the torsion angles of these linkages can be obtained with Privateer.

Chapter 4

Conclusions and Future Work

The general aim of this PhD was to make improvements to software tools for carbohydrate structure solution with X-ray crystallography and electron cryo-microscopy, with a focus on refinement and model building. There are numerous errors in carbohydrate structures in the PDB. The software tools currently available mostly focus on the protein structure. The carbohydrate-specific tools are in general less developed. The reason for this is that the carbohydrate structure solution process is more difficult. Also, in the case of X-ray crystallography, they are often trimmed from the protein before crystallisation as they make crystal formation more difficult. The specifics of carbohydrate structure need to be considered when designing software tools, both on monosaccharide level and polysaccharide level. For monosaccharides it is important to consider anomeric configuration, stereochemistry, conformation, bond lengths and angles. For polysaccharides, in addition to having the correct geometry for each monosaccharide, it is important to ensure that they are linked correctly, with accurate values for bond lengths and angles of the linkage. Moreover, common issues encountered when building carbohydrate trees with the current software tools are incomplete or missing trees and wrong types of trees being built.

In Chapter 1, an analysis was carried out of the monosaccharides part of N-glycosylation trees in the PDB focusing specifically on conformation errors. Data on the Cremer-Pople puckering parameters were collected for monosaccharides in N-glycosylated structures at a range of resolutions from the PDB. The θ angle, in particular, indicates the conformation of a sugar and can be found in the output of Privateer. How much the value of θ deviates from ideal values indicates how distorted the conformation is. Up to 36% were discovered to have a value for the θ angle indicative of a high energy conformation. The number of monosaccharides in high energy conformations increases when the resolution decreases. If the conformation of a sugar is distorted, this usually indicates an error unless the electron density/potential clearly supports this. However, in N-glycosylated structures, monosaccharides in high energy conformations are rare. The lower the resolution is, the more common conformation errors are, which is also noticeable in cryo-EM structures. These conformation errors can arise during both the model building and refinement steps of structure solution. The aims of Chapters 2 and 3 were to address the errors arising during carbohydrate refinement and model building respectively.

The software used for carbohydrate refinement is the same as the software used for protein refinement, but making use of monosaccharide specific dictionaries. A contributing factor for the errors arising during refinement are the issues with the monosaccharide dictionaries in the CCP4 Monomer Library. Dictionaries generally contain, among other things, torsion restraints informing the angles in a molecule. For some monosaccharides, these angles were incorrectly set to 0° in the old CCP4 Monomer Library. When used with refinement software, these dictionaries cause distorted geometry. New monosaccharide dictionaries were generated with the most current version of AceDRG, which led to a clear improvement when used for refinement. Moreover, refinement software can sometimes distort a monosaccharide from its lowest energy conformation to fit the electron density/potential better. This has been addressed by introducing a new set of unimodal torsion restraints, which contain values for the angles of the ring of a monosaccharide. When activated, they allow the sugar to retain its lowest energy conformation. These restraints should, however, be used with care when working with monosaccharides where the electron density/potential supports a higher energy conformation, which, as mentioned above, is not common for monosaccharides part of N-glycosylation.

The subsequent testing of these dictionaries indicated that the new unimodal torsion restraints clearly help monosaccharides adopt the lowest energy conformation in addition to generally leading to fewer errors during refinement. This could sometimes lead to a higher RSCC and R-factor. This may look like worsening at first, but it actually indicates that the model is not being wrongly distorted. When working on carbohydrates, especially monosaccharides at the end of a carbohydrate tree and especially at lower resolution, the electron density/potential is often not completely clear, because the chain tends to move. This commonly causes monosaccharides to be wrongly modelled with high energy conformations. The new unimodal torsion restraints prevent this by keeping the ring in a low energy chair conformation. This could lead to a slight increase in RSCC and R-factors, but in fact prevents overfitting and leads to a better model overall. These restraints are deactivated by default in CCP4 and it is up to the user to decide if they are useful for the particular case. A potential continuation of this work would be having unimodal torsion restraints available for a variety of conformations in order to cover cases where the monosaccharide is clearly in a non-chair conformation. This would, however, have to come with a feature in refinement software that allows the user to select the correct unimodal torsion restraints for these cases.

The next aim was to improve automated carbohydrate model building from electron density/potential maps. Currently, there is only one software tool for this, the carbohydrate module in Coot. It works by assigning atom positions and then refining the monosaccharide iteratively until it arrives at a solution of a sufficient quality. However, it has a limited selection

of glycoforms and also for best results requires a certain level of user expertise. Moreover, the carbohydrate module of Coot can sometimes lead to incomplete or missing trees and incorrectly built residues when used in the fully automated mode. This tool has been improved by van Beusekom et al by allowing partial trees to be extended. This allows for the completion or correction of trees already present in a structure. The main issue remaining after these improvements, is still that sometimes the wrong type of tree is built, often the tree is not fully built and residues towards the end of the tree can be built incorrectly (van Beusekom et al. 2019). However, this module represents an example of how carbohydrate tree extension can be performed.

Chapter 3 is an attempt at using a different approach to build carbohydrates. The method used by Sails is inspired by protein and nucleic acid building software. This method involves a database of fingerprints that are scored against the electron density/potential map. Each fingerprint is a collection of coordinates that indicates where a residue, in this case a monosaccharide, is expected to present with points of high electron density/potential or low electron density/potential. These points have to match the presence or lack of electron density/potential on the map. This method is underlying BUCCANEER for building proteins and NAUTILUS for building nucleic acids. The aim of Chapter 3 was to find out if it is transferable to carbohydrates. The main conclusion is that, while it is possible to detect a range of monosaccharides, the success is highly dependent on resolution. Detection of monosaccharides is significantly better at high resolution than at low resolution. This means that highly glycosylated structures, which rarely have very high resolution, can be challenging to model. This, unfortunately, includes structures determined with electron cryo-microscopy. Cryo-EM is commonly used on large structures like antibodies and viruses, which often contain long and branched carbohydrates. The fingerprint approach also greatly depends on the quality of the fingerprints available in the database, which in turn depends on the availability of data for the generation of these fingerprints. The availability of high resolution N-glycosylated structures in the PDB is limited, especially structures containing sugars different from NAG. For this reason, good quality fingerprints for less common monosaccharides are challenging to generate. Bad quality fingerprints generated from low resolution data cause monosaccharides to be missed or wrongly detected.

Despite its limitations, the fingerprinting method when used for carbohydrates is good at detecting the well resolved GlcNAc residues at the beginning of a glycosylation tree. NAG is the easiest to detect monosaccharide residue, because it is a very common one with plenty of high resolution structures available and also because the root of the N-glycan tree normally has better resolution. The initial hit can then be extended based on data from glycomics databases, which can be retrieved with Privateer. Privateer allows for information

on what types of trees a glycoprotein can contain, tree length and composition. This, combined with Sails detecting where each tree starts, can be used to build the entire glycan. The tree extension from the GlcNAc detected at the base of the tree can be done using a method similar to the one used by Coot's carbohydrate module. This approach would be able to circumvent the issues of the Coot carbohydrate building module of building incorrect types of trees. Moreover, having data on the glycan content of a structure allows for speeding up the model building step, as it renders it unnecessary to scan the map for each monosaccharide fingerprint individually. To summarise, the process would involve the following steps:

- Retrieve data from glycomics databases about the glycan composition of a protein
- Use the fingerprinting method to find the Asn-NAG at the root of the glycosylation tree.
- Extend the carbohydrate tree in a manner similar to the way the carbohydrate module in Coot works

The resultant glycans can then be refined using REFMAC5 and the monosaccharide dictionaries mentioned above before being validated with Privateer. For best results, in the future this could be implemented as part of a pipeline with iterative refinement and validation until the best possible result is obtained. Furthermore, the Privateer validation metrics, such as B-factor, RSCC, Cremer-Pople parameters could even be used to inform the model building process in real time. These, combined with the current score, which is based on how well a monosaccharide fits into the electron density/potential, can give an improved method for scoring monosaccharides. This could greatly decrease the number of major errors that Sails makes. Moreover, this could be further extended into a machine learning-based approach similar to the one used by Bond et al. (Bond, Wilson, and Cowtan 2020) to predict if a protein model is correct. This method uses a neural network trained on a test set consisting of a number of high quality protein structures and a collection of validation metrics for these structures. This method has been included as part of the ModelCraft model building pipeline. A similar approach could be attempted for carbohydrates. A set of suitable validation metrics would have to be selected, in addition to a training set for the neural network. However, a potential issue would be the insufficient availability of high quality glycosylation data in the PDB. Other machine learning approaches that require less data for training could also be explored.

Another potential area to explore would be making use of information obtained from molecular dynamics simulations. Molecular dynamics simulations can provide insight into

how carbohydrates interact with the protein. This approach has been implemented as a proof of concept by Bagdonas et al. (Bagdonas et al. 2021) to show that glycan blocks produced by molecular dynamics simulations can be built into protein structures predicted by AlphaFold2. This functionality is currently under development as a part of the Privateer software. This feature can also be implemented into Sails to give glycans of better quality with more realistic structures.

An area that remains unaddressed in both Chapters 2 and 3 is furanose sugars. Their conformation will always exhibit more strain than a pyranose chair conformation and care needs to be taken when enforcing it. Furthermore, Sails does not currently handle different types of glycosylation, such as O- and C-glycosylation, and carbohydrate ligands. O- and C-glycosylation are even more challenging to model than N-glycosylation, because the glycans are even less resolved and there is generally less data that can be used for fingerprinting. Nevertheless, the combined approach suggested above could potentially be successful at addressing them. The challenge with carbohydrate ligands is that they can have more variable conformations. However, the fingerprint database can be extended to contain monosaccharides in different conformations to be used only when looking for ligands. It is important to consider the most common scenarios, as expanding the fingerprint database too much could lead to a significant decrease in speed.

Appendix A.

Supplementary Data for “Updated restraint dictionaries for carbohydrates in the pyranose form”

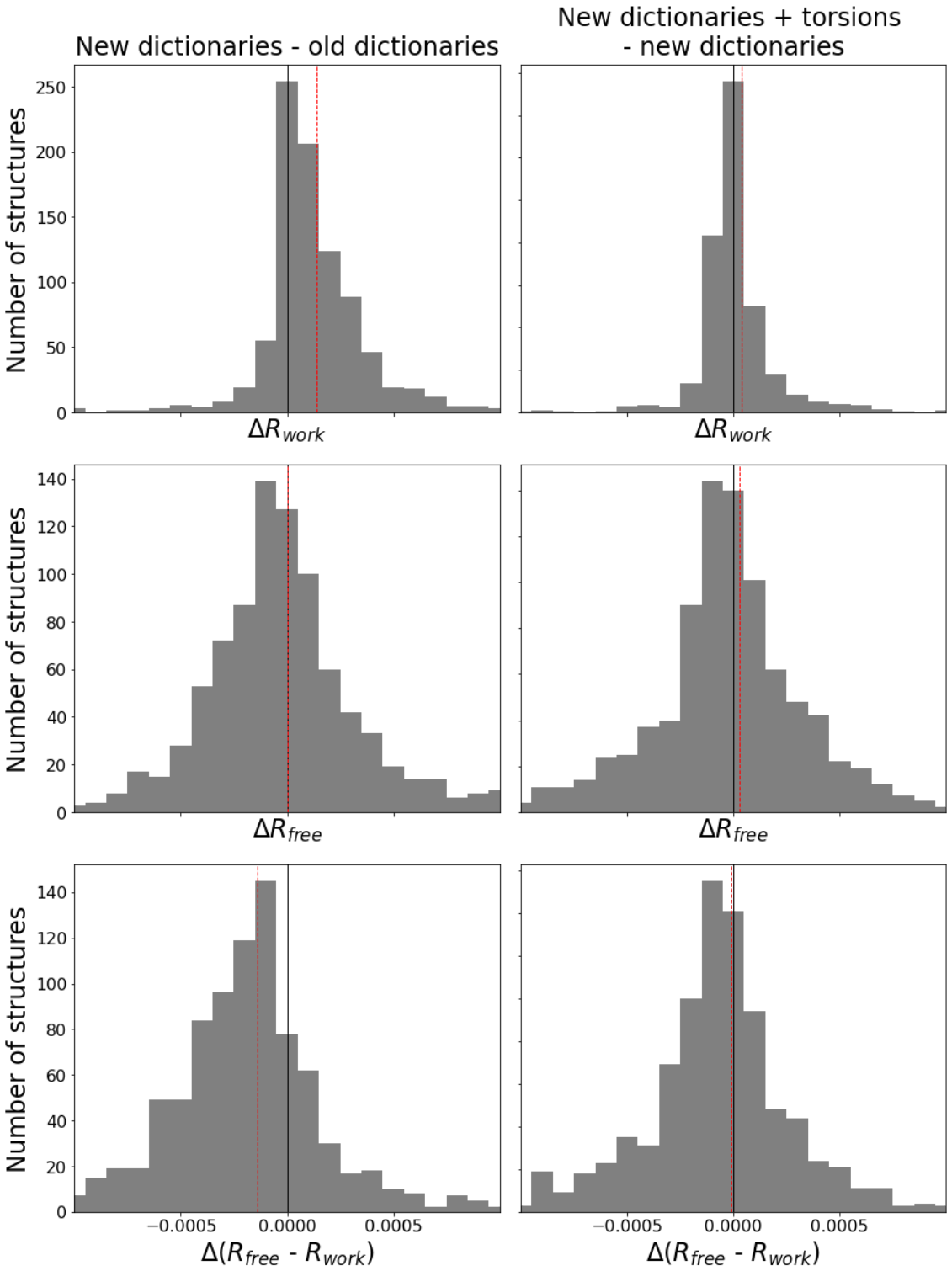


Figure A.1. Distribution of changes in R_{work}, R_{free}, and the R-factor gap R_{work}-R_{free}. Black lines indicate the origin, and red dashed lines the median. The horizontal axis is truncated to the region of interest. Using the new dictionaries increases R_{work} while R_{free} is unchanged. As a result the R-factor gap is reduced indicating less overfitting in refinement. This effect is

further strengthened, if to a lesser degree, when also including additional unimodal torsion restraints.

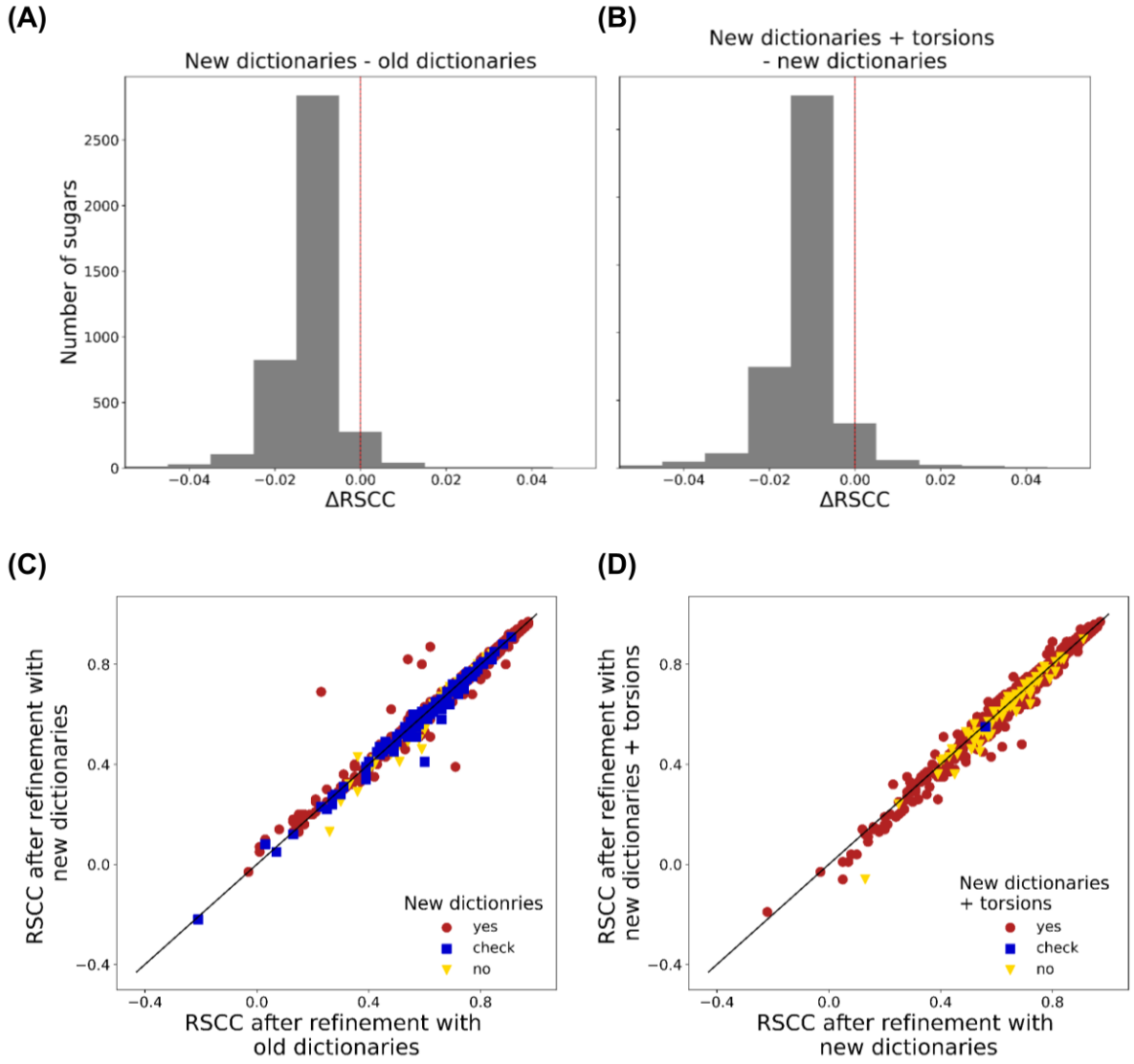


Figure A.2. There is a slight decrease in RSCC for sugars involved in N/O glycosylation after refinement with the new dictionaries, regardless of the use of unimodal torsion restraints. The sugars are marked as “yes”, “no”, and “check” based on their validation after refinement with the protocol on the vertical axis. The horizontal axis is truncated to the region of interest. In (A) and (B) black lines indicate the origin, and red dashed lines the median.

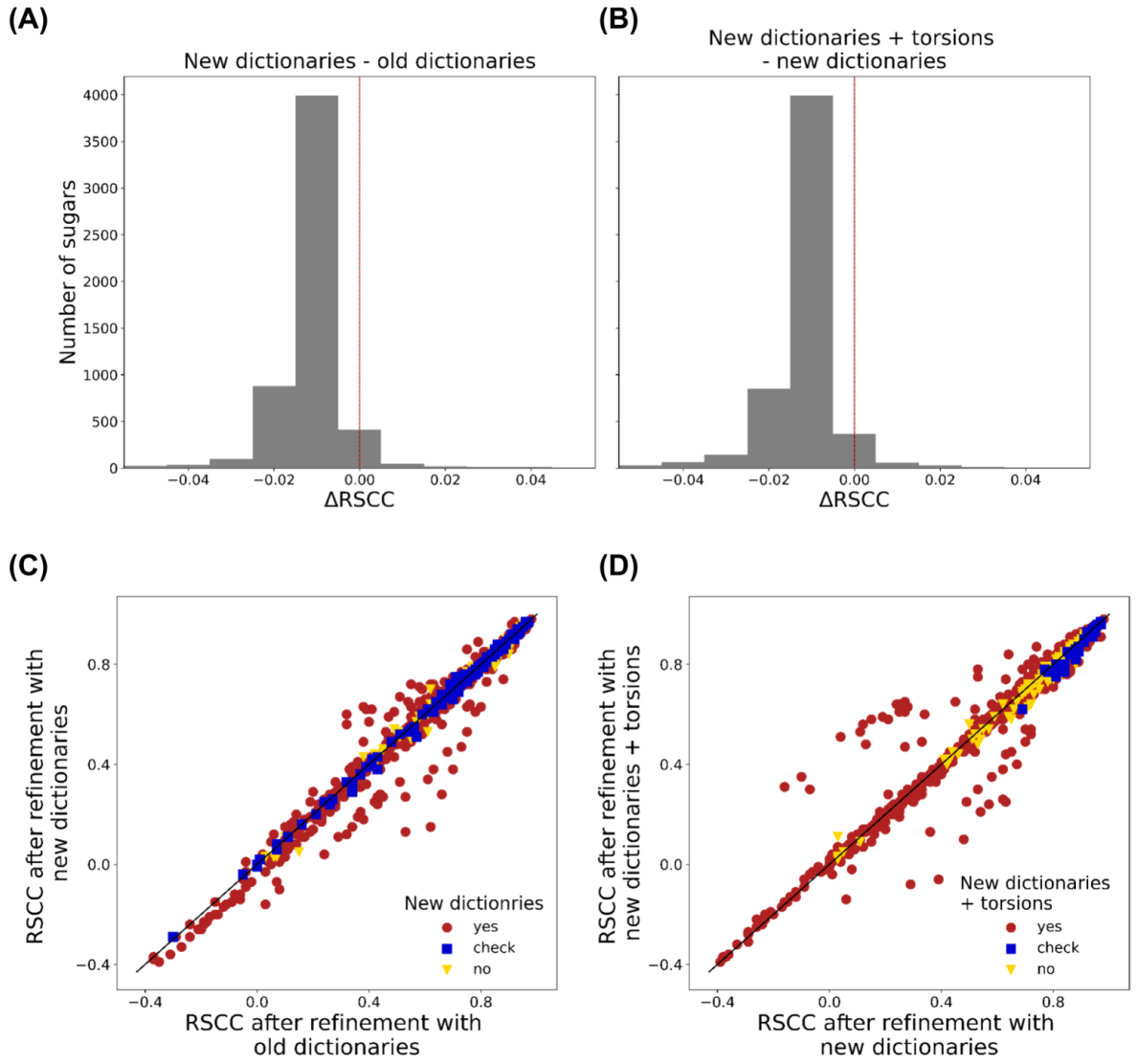
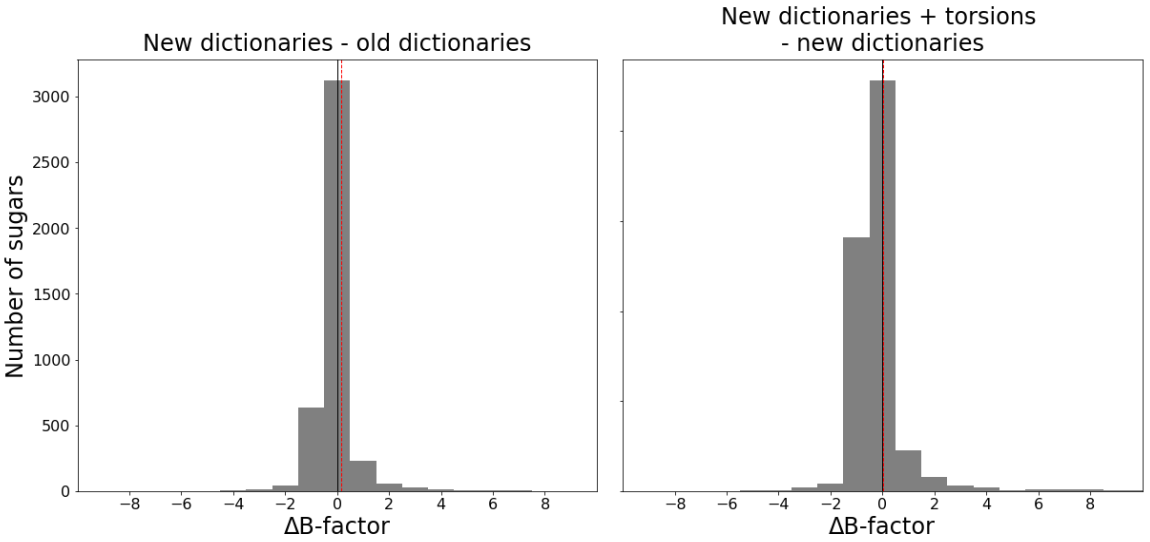
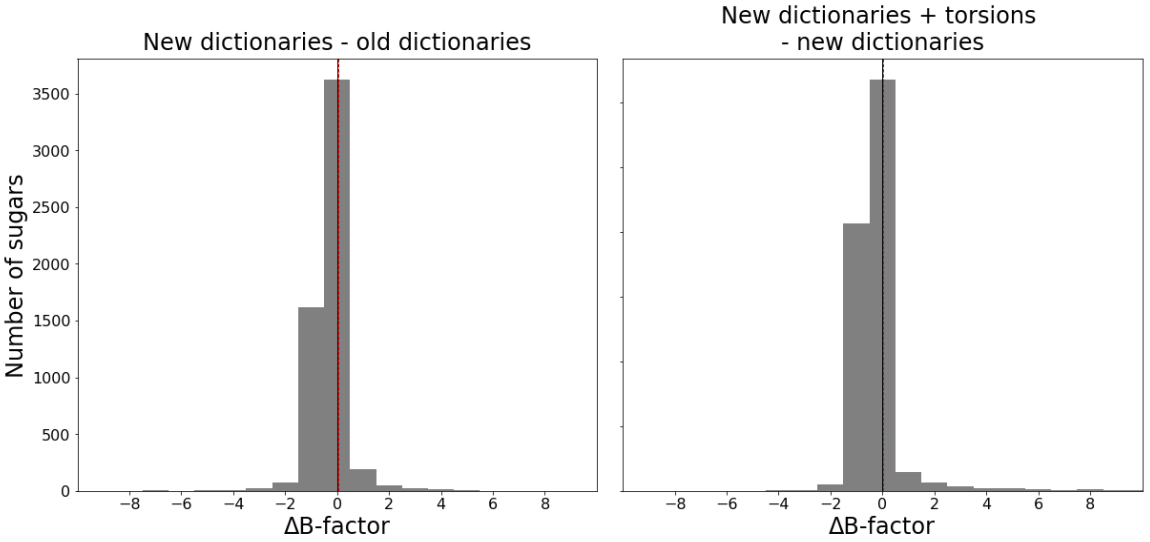


Figure A.3. There is a slight decrease in RSCC for ligand sugars after refinement with the new dictionaries, regardless of the use of unimodal torsion restraints. We observe a bias in the distribution of outliers: a few isolated cases appear to achieve higher RSCC with the old dictionaries (C), while the same is true for the new dictionaries with torsions versus those without (D). These outliers are cases similar to that discussed in Figure 5. The sugars are marked as “yes”, “no”, and “check” based on their validation after refinement with the protocol on the vertical axis. The horizontal axis is truncated to the region of interest. In (A) and (B) black lines indicate the origin, and red dashed lines the median.



(A)



(B)

Figure A.4. Distribution of changes in B-factor. The horizontal axis is truncated to the region of interest. Black lines indicate the origin, and red dashed lines the median. (A) Sugars that are part of N/O-glycosylation; (B) Ligands. There is a slight decrease in B-factor after refinement with the new dictionaries, and a further decrease when unimodal torsion restraints are used.

CCD	Q	Phi	Theta	Detected type	Cnf	Ok?	Name
145	0.574	344.33	3.93	beta-D-aldopyranose	4c1	yes	2-nitrophenyl beta-D-galactopyranoside
147	0.574	351.33	3.59	beta-D-aldopyranose	4c1	yes	4-nitrophenyl beta-D-galactopyranoside
16G	0.552	223.59	3.31	alpha-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-6-O-phosphono-alpha-D-glucopyranose
18D	0.542	53.26	175.51	alpha-L-ketopyranose	1c4	yes	3,5-dideoxy-5-(propanoylamino)-D-glycero-alpha-D-galacto-non-2-ulopyranosonic acid
1GL	0.541	259.05	4.43	alpha-D-aldopyranose	4c1	yes	2,6-dideoxy-4-O-methyl-alpha-D-galactopyranose
1GN	0.562	293.42	5.96	beta-D-aldopyranose	4c1	yes	2-amino-2-deoxy-beta-D-galactopyranose
289	0.549	271.25	5.57	beta-D-aldopyranose	4c1	yes	D-glycero-alpha-D-manno-heptopyranose
291	0.550	233.39	2.37	alpha-D-aldopyranose	4c1	yes	prop-2-en-1-yl 7-O-carbamoyl-L-glycero-alpha-D-manno-heptopyranoside
293	0.544	266.78	4.27	alpha-D-aldopyranose	4c1	yes	2-deoxy-beta-L-galacto-heptopyranose
2DG	0.547	268.84	5.72	alpha-D-aldopyranose	4c1	yes	2-deoxy-alpha-D-galactopyranose
2FG	0.562	290.35	4.65	beta-D-aldopyranose	4c1	yes	2-deoxy-2-fluoro-beta-D-galactopyranose
2GS	0.559	261.69	3.69	alpha-D-aldopyranose	4c1	yes	2-O-methyl-alpha-D-galactopyranose
3FM	0.538	333.45	1.10	alpha-D-aldopyranose	4c1	yes	3-O-carbamoyl-alpha-D-mannopyranose
3HD	0.574	328.99	2.30	beta-D-aldopyranose	4c1	yes	1,5-anhydro-3-O-methyl-D-mannitol
3MG	0.557	319.74	3.81	beta-D-aldopyranose	4c1	yes	3-O-methyl-beta-D-glucopyranose
42D	0.545	46.01	175.50	alpha-L-ketopyranose	1c4	yes	3,5-dideoxy-5-[(methoxycarbonyl)amino]-D-glycero-alpha-D-galacto-non-2-ulopyranosonic acid
445	0.562	10.77	4.37	beta-D-aldopyranose	4c1	yes	N-[oxo(phenylamino)acetyl]-beta-D-glucopyranosylamine
46M	0.567	295.18	5.81	beta-D-aldopyranose	4c1	yes	(4AR,6R,7S,8R,8AS)-hexahydro-6,7,8-trihydroxy-2-methylpyrano[3,2-D][1,3]dioxine-2-carboxylic acid
475	0.565	6.97	4.10	beta-D-aldopyranose	4c1	yes	N-[oxo(pyridin-2-ylamino)acetyl]-beta-D-glucopyranosylamine
49A	0.441	81.24	128.26	beta-L-ketopyranose	5h4	yes	4,9-amino-2,4-deoxy-2,3-dehydro-n-acetyl-neuraminic acid
4AM	0.443	82.25	128.10	beta-L-ketopyranose	5h4	yes	4-amino-2-deoxy-2,3-dehydro-N-neuraminic acid
4GP	0.564	9.60	4.25	beta-D-aldopyranose	4c1	yes	N-(carboxycarbonyl)-beta-D-glucopyranosylamine
6GP	0.567	27.78	4.98	beta-D-aldopyranose	4c1	yes	N-[methoxy(oxo)acetyl]-beta-D-glucopyranosylamine
6MN	0.556	239.61	6.04	alpha-D-aldopyranose	4c1	yes	2-amino-2-deoxy-6-O-phosphono-alpha-D-mannopyranose
7JZ	0.554	315.73	5.29	beta-D-aldopyranose	4c1	yes	2-deoxy-2,2-difluoro-beta-D-lyxo-hexopyranose
8GP	0.566	19.44	4.88	beta-D-aldopyranose	4c1	yes	N-[(cyclopropylamino)(oxo)acetyl]-beta-D-glucopyranosylamine
A2G	0.541	261.50	9.77	alpha-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-alpha-D-galactopyranose
A6P	0.552	255.48	3.24	alpha-D-aldopyranose	4c1	yes	6-O-phosphono-alpha-D-allopyranose

ABE	0.552	305.94	2.11	alpha-D-aldopyranose	4c1	yes	alpha-D-Abequopyranose
ADA	0.546	270.30	4.39	alpha-D-aldopyranose	4c1	yes	alpha-D-galactopyranuronic acid
AGL	0.551	292.99	3.10	alpha-D-aldopyranose	4c1	yes	4-amino-4,6-dideoxy-alpha-D-glucopyranose
AMN	0.539	33.50	176.30	alpha-L-ketopyranose	1c4	yes	methyl 5-acetamido-9-amino-3,5,9-trideoxy-D-glycero-alpha-D-galacto-non-2-ulopyranosidonic acid
AMU	0.562	305.63	2.62	beta-D-aldopyranose	4c1	yes	N-acetyl-beta-muramic acid
AMV	0.575	13.82	3.13	beta-D-aldopyranose	4c1	yes	methyl 2-acetamido-3-O-[(1R)-1-carboxyethyl]-2-deoxy-beta-D-glucopyranoside
ANA	0.542	40.04	175.76	alpha-L-ketopyranose	1c4	yes	methyl 4-O-acetyl-5-acetamido-3,5-dideoxy-D-glycero-alpha-D-galacto-non-2-ulopyranosidonic acid
ARA	0.557	336.74	4.85	alpha-N-aldopyranose	4c1	yes	alpha-L-arabinopyranose
ARW	0.570	15.06	2.28	beta-N-aldopyranose	4c1	yes	methyl beta-D-arabinopyranoside
ASG	0.556	323.57	3.82	beta-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-4-O-sulfo-beta-D-galactopyranose
ASO	0.568	351.87	2.88	beta-D-aldopyranose	4c1	yes	1,5-anhydro-D-glucitol
B16	0.566	329.64	3.43	beta-D-aldopyranose	4c1	yes	1,6-di-O-phosphono-beta-D-glucopyranose
B7G	0.567	15.10	3.50	beta-D-aldopyranose	4c1	yes	heptyl beta-D-glucopyranoside
BDG	0.550	238.39	4.87	alpha-D-aldopyranose	4c1	yes	2,6-diamino-2,6-dideoxy-alpha-D-glucopyranose
BDP	0.554	327.33	4.34	beta-D-aldopyranose	4c1	yes	beta-D-glucopyranuronic acid
BEM	0.564	333.78	2.56	beta-D-aldopyranose	4c1	yes	beta-D-mannopyranuronic acid
BG6	0.554	315.30	4.13	beta-D-aldopyranose	4c1	yes	6-O-phosphono-beta-D-glucopyranose
BGC	0.554	316.40	3.92	beta-D-aldopyranose	4c1	yes	beta-D-glucopyranose
BGL	0.567	345.41	1.08	beta-D-aldopyranose	4c1	yes	2-O-octyl-beta-D-glucopyranose
BGN	0.565	305.91	2.69	beta-D-aldopyranose	4c1	yes	2-(butanoylamino)-2-deoxy-beta-D-glucopyranose
BGP	0.560	289.86	7.25	beta-D-aldopyranose	4c1	yes	6-O-phosphono-beta-D-galactopyranose
BGS	0.563	346.40	4.17	beta-D-aldopyranose	4c1	yes	(1S)-1,5-anhydro-1-(ethylsulfonyl)-D-glucitol
BHG	0.567	338.47	3.60	beta-D-aldopyranose	4c1	yes	hexyl beta-D-galactopyranoside
BM3	0.538	256.00	9.53	alpha-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-alpha-D-mannopyranose
BMA	0.557	349.91	3.55	beta-D-aldopyranose	4c1	yes	beta-D-mannopyranose
BNG	0.563	353.62	3.07	beta-D-aldopyranose	4c1	yes	nonyl beta-D-glucopyranoside
BOG	0.565	358.18	3.46	beta-D-aldopyranose	4c1	yes	octyl beta-D-glucopyranoside
C3X	0.567	3.31	2.43	beta-N-aldopyranose	4c1	yes	(2R)-oxiran-2-ylmethyl beta-D-xylopyranoside
C4X	0.568	13.85	2.58	beta-N-aldopyranose	4c1	yes	3,4-epoxybutyl-beta-D-xyloside
C5X	0.567	3.18	2.54	beta-N-aldopyranose	4c1	yes	3-[(2R)-oxiran-2-yl]propyl beta-D-xylopyranoside
CBF	0.533	253.47	6.92	alpha-D-aldopyranose	4c1	yes	(2R,3R,4S,5S,6R)-2,3,4,5-tetrahydroxy-6-(hydroxymethyl)oxane-2-carboxamide

CDG	0.566	351.87	3.15	beta-D-aldopyranose	4c1	yes	methyl 4,6-O-[(1R)-1-carboxyethylidene]-beta-D-galactopyranoside
CEG	0.565	288.24	6.01	beta-D-aldopyranose	4c1	yes	4,6-O-[(1S)-1-carboxyethylidene]-beta-D-glucopyranose
CNP	0.540	52.29	174.60	alpha-L-ketopyranose	1c4	yes	2-propenyl-N-acetyl-neuraminic acid
CR1	0.562	7.78	4.75	beta-D-aldopyranose	4c1	yes	N-(methoxycarbonyl)-beta-D-glucopyranosylamine
CR6	0.536	246.33	8.55	beta-D-aldopyranose	4c1	yes	1-deoxy-1-acetylamino-beta-D-gluco-2-heptulopyranosonamide
CRA	0.543	204.12	6.97	beta-D-aldopyranose	4c1	yes	1-deoxy-1-methoxycarbamido-beta-D-gluco-2-heptulopyranosonamide
D6G	0.543	263.84	4.07	alpha-D-aldopyranose	4c1	yes	2-deoxy-6-O-phosphono-alpha-D-glucopyranose
DAG	0.558	319.16	5.82	beta-D-aldopyranose	4c1	yes	4-amino-4,6-dideoxy-beta-D-glucopyranose
DDA	0.552	316.07	5.27	beta-D-aldopyranose	4c1	yes	beta-D-Olivopyranose
DDL	0.555	306.71	5.79	beta-D-aldopyranose	4c1	yes	2,6-dideoxy-beta-D-galactopyranose
DEG	0.548	254.68	2.66	alpha-D-aldopyranose	4c1	yes	butyl alpha-D-mannopyranoside
DK4	0.547	350.24	5.15	beta-D-aldopyranose	4c1	yes	1-(3-deoxy-3-fluoro-beta-D-glucopyranosyl)-5-fluoropyrimidine-2,4(1H,3H)-dione
DK5	0.535	335.00	6.19	beta-D-aldopyranose	4c1	yes	1-(2,3-dideoxy-3-fluoro-beta-D-arabino-hexopyranosyl)-4-[(phenylcarbonyl)amino]pyrimidin-2(1H)-one
DKX	0.547	354.37	5.06	beta-D-aldopyranose	4c1	yes	1-(3-deoxy-3-fluoro-beta-D-glucopyranosyl)pyrimidine-2,4(1H,3H)-dione
DKY	0.551	3.52	5.88	beta-D-aldopyranose	4c1	yes	1-(3-deoxy-3-fluoro-beta-D-glucopyranosyl)-4-[(phenylcarbonyl)amino]pyrimidin-2(1H)-one
DKZ	0.546	356.14	5.14	beta-D-aldopyranose	4c1	yes	4-amino-1-(3-deoxy-3-fluoro-beta-D-glucopyranosyl)pyrimidin-2(1H)-one
DL6	0.564	17.07	4.57	beta-D-aldopyranose	4c1	yes	N-(azidoacetyl)-beta-D-glucopyranosylamine
DLF	0.544	90.52	173.63	alpha-L-aldopyranose	1c4	yes	2-deoxy-alpha-L-fucopyranose
DO8	0.544	243.34	6.62	alpha-D-ketopyranose	4c1	yes	3-deoxy-8-O-phosphono-alpha-D-manno-oct-2-ulopyranosonic acid
DRI	0.547	326.88	3.85	beta-D-aldopyranose	4c1	yes	2,6-dideoxy-4-O-methyl-beta-D-glucopyranose
DSR	0.552	318.97	4.12	beta-D-aldopyranose	4c1	yes	2,6-dideoxy-4-thio-beta-D-allopyranose
DVC	0.520	117.97	172.73	alpha-L-aldopyranose	1c4	yes	(2R,4S,6S)-4-azanyl-4,6-dimethyl-oxane-2,5,5-triol
EAG	0.575	350.87	3.58	beta-D-aldopyranose	4c1	yes	2-aminoethyl 2-acetamido-2-deoxy-beta-D-glucopyranoside
EBG	0.553	233.65	2.46	alpha-D-aldopyranose	4c1	yes	2-[(2S)-oxiran-2-yl]ethyl alpha-D-glucopyranoside
EMP	0.548	350.26	4.58	alpha-N-aldopyranose	4c1	yes	2,4-dideoxy-4-(ethylamino)-3-O-methyl-alpha-L-threo-pentopyranose
EPG	0.555	250.89	1.76	alpha-D-aldopyranose	4c1	yes	(2R)-oxiran-2-ylmethyl alpha-D-glucopyranoside
EQP	0.560	184.81	175.36	alpha-L-ketopyranose	1c4	yes	(1R)-4-acetamido-1,5-anhydro-2,4-dideoxy-1-phosphono-D-glycero-D-galacto-octitol
F1P	0.547	297.95	177.05	beta-N-ketopyranose	1c4	yes	1-O-phosphono-beta-D-fructopyranose

FCA	0.554	270.33	5.45	alpha-D-aldopyranose	4c1	yes	alpha-D-fucopyranose
FCB	0.561	300.51	6.18	beta-D-aldopyranose	4c1	yes	beta-D-fucopyranose
FUC	0.553	89.18	175.16	alpha-L-aldopyranose	1c4	yes	alpha-L-fucopyranose
FUL	0.560	122.71	174.09	beta-L-aldopyranose	1c4	yes	beta-L-fucopyranose
G0S	0.566	352.88	4.84	beta-D-aldopyranose	4c1	yes	3-(beta-D-galactopyranosylthio)propanoic acid
G16	0.559	261.53	3.94	alpha-D-aldopyranose	4c1	yes	1,6-di-O-phosphono-alpha-D-glucopyranose
G1P	0.555	241.59	2.06	alpha-D-aldopyranose	4c1	yes	1-O-phosphono-alpha-D-glucopyranose
G2F	0.555	230.43	4.48	alpha-D-aldopyranose	4c1	yes	2-deoxy-2-fluoro-alpha-D-glucopyranose
G3F	0.552	5.55	4.25	beta-D-aldopyranose	4c1	yes	3-deoxy-3-fluoro-beta-D-glucopyranose
G4D	0.537	274.87	5.93	alpha-D-aldopyranose	4c1	yes	4-deoxy-alpha-D-glucopyranose
G4S	0.554	318.99	4.37	beta-D-aldopyranose	4c1	yes	4-O-sulfo-beta-D-galactopyranose
G6P	0.552	263.79	4.57	alpha-D-aldopyranose	4c1	yes	6-O-phosphono-alpha-D-glucopyranose
G6S	0.558	311.67	5.22	beta-D-aldopyranose	4c1	yes	6-O-sulfo-beta-D-galactopyranose
G7P	0.558	304.76	4.80	beta-D-aldopyranose	4c1	yes	6,7-dideoxy-7-phosphono-beta-D-gluco-heptopyranose
GAA	0.561	280.37	3.92	alpha-D-aldopyranose	4c1	yes	3-nitrophenyl alpha-D-galactopyranoside
GAF	0.553	244.64	4.46	alpha-D-aldopyranose	4c1	yes	2-deoxy-2-fluoro-alpha-D-galactopyranose
GAL	0.559	292.83	5.65	beta-D-aldopyranose	4c1	yes	beta-D-galactopyranose
GAT	0.559	167.72	1.14	alpha-D-aldopyranose	4c1	yes	4-aminophenyl alpha-D-galactopyranoside
GC4	0.544	344.07	4.54	beta-D-aldopyranose	4c1	yes	4-deoxy-beta-D-glucopyranuronic acid
GCN	0.535	278.16	4.66	alpha-D-aldopyranose	4c1	yes	2-amino-2,3-dideoxy-alpha-D-glucoyranose
GCS	0.565	289.39	3.79	beta-D-aldopyranose	4c1	yes	2-amino-2-deoxy-beta-D-glucopyranose
GCU	0.545	250.13	2.20	alpha-D-aldopyranose	4c1	yes	alpha-D-glucopyranuronic acid
GCV	0.534	271.22	5.98	alpha-D-aldopyranose	4c1	yes	4-O-methyl-alpha-D-glucopyranuronic acid
GCW	0.555	322.41	4.17	beta-D-aldopyranose	4c1	yes	4-O-methyl-beta-D-glucopyranuronic acid
GDA	0.559	330.06	5.09	beta-D-aldopyranose	4c1	yes	4-amino-4-deoxy-beta-D-glucopyranose
GFP	0.554	236.24	4.46	alpha-D-aldopyranose	4c1	yes	2-deoxy-2-fluoro-1-O-phosphono-alpha-D-glucopyranose
GL0	0.554	321.70	4.75	beta-D-aldopyranose	4c1	yes	beta-D-gulopyranose
GL1	0.556	260.38	2.41	alpha-D-aldopyranose	4c1	yes	1-O-phosphono-alpha-D-galactopyranose
GL2	0.532	194.81	5.60	beta-D-aldopyranose	4c1	yes	(5S,7R,8S,9S,10R)-3-amino-8,9,10-trihydroxy-7-(hydroxy methyl)-6-oxa-1,3-diazaspiro[4.5]decane-2,4-dione
GLA	0.551	263.76	4.39	alpha-D-aldopyranose	4c1	yes	alpha-D-galactopyranose
GLC	0.548	245.03	1.71	alpha-D-aldopyranose	4c1	yes	alpha-D-glucopyranose
GLD	0.546	279.14	3.04	alpha-D-aldopyranose	4c1	yes	4,6-dideoxy-alpha-D-xylo-hexopyranose
GLF	0.550	264.44	2.89	alpha-D-aldopyranose	4c1	yes	alpha-D-glucopyranosyl fluoride
GLG	0.544	262.48	8.47	alpha-D-aldopyranose	4c1	yes	Alpha-D-glucopyranosyl-2-carboxylic acid amide
GLP	0.556	244.30	6.21	alpha-D-aldopyranose	4c1	yes	2-amino-2-deoxy-6-O-phosphono-alpha-D-glucopyranos e

GLS	0.534	211.43	5.10	beta-D-aldopyranose	4c1	yes	Beta-D-glucopyranose spirohydantoin
GMB	0.570	30.82	2.69	beta-D-aldopyranose	4c1	yes	1,7-di-O-phosphono-L-glycero-beta-D-manno-heptopyranose
GMH	0.545	278.34	3.06	alpha-D-aldopyranose	4c1	yes	L-glycero-alpha-D-manno-heptopyranose
GP4	0.550	238.87	4.40	alpha-D-aldopyranose	4c1	yes	2-amino-2-deoxy-4-O-phosphono-alpha-D-glucopyranose
GS1	0.563	352.72	4.33	beta-D-aldopyranose	4c1	yes	1-thio-beta-D-glucopyranose
GTM	0.554	2.72	4.09	beta-D-aldopyranose	4c1	yes	methyl 4-thio-beta-D-glucopyranoside
GTR	0.557	309.98	5.98	beta-D-aldopyranose	4c1	yes	beta-D-galactopyranuronic acid
GU0	0.568	305.59	1.59	beta-D-aldopyranose	4c1	yes	2,3,6-tri-O-sulfonato-beta-D-glucopyranose
GU3	0.562	251.91	1.48	alpha-D-aldopyranose	4c1	yes	methyl 3-O-methyl-2,6-di-O-sulfo-alpha-D-glucopyranoside
GU4	0.577	219.06	6.21	alpha-D-aldopyranose	4c1	yes	2,3,4,6-tetra-O-sulfonato-alpha-D-glucopyranose
GU8	0.574	61.68	3.43	beta-D-aldopyranose	4c1	yes	2,3,6-tri-O-methyl-beta-D-glucopyranose
GU9	0.503	262.07	167.03	alpha-D-aldopyranose	1c4	yes	2,3,6-tri-O-methyl-alpha-D-glucopyranose
GUP	0.537	119.59	175.27	alpha-L-aldopyranose	1c4	yes	alpha-L-gulopyranose
GXL	0.552	85.34	175.36	alpha-L-aldopyranose	1c4	yes	Alpha-L-galactopyranose
H1M	0.555	225.05	2.59	alpha-D-aldopyranose	4c1	yes	methyl 2-deoxy-2-(2-hydroxyethyl)-alpha-D-mannopyranoside
H2P	0.555	151.08	1.46	alpha-D-aldopyranose	4c1	yes	1-deoxy-2-O-phosphono-alpha-D-gluco-hept-2-ulopyranose
IDG	0.564	142.69	177.18	beta-L-aldopyranose	1c4	yes	2,6-diamino-2,6-dideoxy-beta-L-idopyranose
IDR	0.540	102.30	175.54	alpha-L-aldopyranose	1c4	yes	alpha-L-idopyranuronic acid
IMK	0.560	319.96	3.48	beta-D-aldopyranose	4c1	yes	2-(beta-D-glucopyranosyl)-5-methyl-1-benzimidazole
IPT	0.567	346.35	4.37	beta-D-aldopyranose	4c1	yes	1-methylethyl 1-thio-beta-D-galactopyranoside
JHM	0.536	272.67	9.83	alpha-D-aldopyranose	4c1	yes	2-deoxy-6-O-sulfo-alpha-D-glucopyranose
JZR	0.565	339.54	2.98	beta-D-aldopyranose	4c1	yes	hexyl beta-D-glucopyranoside
KDA	0.552	267.18	3.87	alpha-D-ketopyranose	4c1	yes	prop-2-en-1-yl 3-deoxy-alpha-D-manno-oct-2-ulopyranosidonic acid
KDB	0.445	315.41	48.84	beta-D-ketopyranose	Oh5	yes	3,4,5-trideoxy-alpha-D-erythro-oct-3-en-2-ulopyranosonic acid
KDO	0.543	238.68	5.25	beta-D-ketopyranose	4c1	yes	3-deoxy-alpha-D-manno-oct-2-ulopyranosonic acid
KDR	0.554	268.11	3.59	beta-D-ketopyranose	4c1	yes	prop-2-en-1-yl 3-deoxy-alpha-D-manno-octos-2-ulopyranoside
KME	0.554	284.38	2.61	alpha-D-ketopyranose	4c1	yes	(1E)-prop-1-en-1-yl 3-deoxy-7-O-methyl-alpha-D-manno-oct-2-ulopyranosidonic acid
KOT	0.571	36.10	1.86	beta-D-aldopyranose	4c1	yes	1-beta-D-glucopyranosyl-4-phenyl-1H-1,2,3-triazole
L6S	0.550	88.07	177.81	alpha-L-aldopyranose	1c4	yes	6-O-sulfo-alpha-L-galactopyranose
LGU	0.536	93.17	174.07	alpha-L-aldopyranose	1c4	yes	alpha-L-gulopyranuronic acid

LXB	0.561	312.87	4.14	beta-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-beta-D-gulopyranose
LXC	0.561	151.16	176.06	beta-N-aldopyranose	1c4	yes	Beta-L-xylopyranose
LXZ	0.541	270.80	6.85	alpha-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-alpha-D-idopyranose
LZ0	0.559	75.44	175.73	alpha-L-aldopyranose	1c4	yes	[1-(2-oxoethyl)-1H-1,2,3-triazol-5-yl]methyl 6-deoxy-alpha-L-galactopyranoside
M07	0.537	201.48	3.79	alpha-D-aldopyranose	4c1	yes	(5R,7R,8S,9S,10R)-7-(hydroxymethyl)-3-(4-methoxyphenyl)-1,6-dioxa-2-azaspiro[4.5]dec-2-ene-8,9,10-triol
M08	0.538	201.20	3.79	alpha-D-aldopyranose	4c1	yes	(5R,7R,8S,9S,10R)-7-(hydroxymethyl)-3-phenyl-1,6-dioxa-2-azaspiro[4.5]dec-2-ene-8,9,10-triol
M09	0.539	227.02	3.64	alpha-D-aldopyranose	4c1	yes	(3S,5R,7R,8S,9S,10R)-7-(hydroxymethyl)-3-(4-nitrophenyl)-1,6-dioxa-2-azaspiro[4.5]decane-8,9,10-triol
M1P	0.549	263.28	1.69	alpha-D-aldopyranose	4c1	yes	1-O-phosphono-alpha-D-mannopyranose
M6D	0.562	355.95	1.99	beta-D-aldopyranose	4c1	yes	6-O-phosphono-beta-D-mannopyranose
M7P	0.553	269.62	6.46	beta-D-aldopyranose	4c1	yes	7-O-phosphono-D-glycero-alpha-D-manno-heptopyranose
M8C	0.549	266.30	5.30	alpha-D-aldopyranose	4c1	yes	methyl alpha-D-galactopyranuronate
MA1	0.550	257.24	1.68	alpha-D-aldopyranose	4c1	yes	1,4-dithio-alpha-D-glucopyranose
MA2	0.549	272.99	7.66	alpha-D-aldopyranose	4c1	yes	4-S-methyl-4-thio-alpha-D-glucopyranose
MA3	0.550	272.33	0.99	alpha-D-aldopyranose	4c1	yes	methyl 4-thio-alpha-D-glucopyranoside
MAG	0.573	346.83	3.41	beta-D-aldopyranose	4c1	yes	methyl 2-acetamido-2-deoxy-beta-D-glucopyranoside
MAN	0.538	260.65	5.99	alpha-D-aldopyranose	4c1	yes	alpha-D-mannopyranose
MAT	0.524	112.14	172.74	alpha-N-aldopyranose	1c4	yes	2,4-dideoxy-3-O-methyl-4-(propan-2-ylamino)-alpha-L-threo-pentopyranose
MAV	0.545	237.87	2.46	alpha-D-aldopyranose	4c1	yes	alpha-D-mannopyranuronic acid
MBF	0.561	297.37	2.58	beta-D-aldopyranose	4c1	yes	2-deoxy-2-fluoro-beta-D-mannopyranose
MBG	0.567	334.98	3.74	beta-D-aldopyranose	4c1	yes	methyl beta-D-galactopyranoside
MDA	0.551	335.67	4.63	beta-D-aldopyranose	4c1	yes	2,6-dideoxy-3-C-methyl-beta-D-ribo-hexopyranose
MDP	0.563	322.01	3.58	beta-D-aldopyranose	4c1	yes	N-carboxyl-N-methyl-beta-muramic acid
MFB	0.570	120.95	176.50	beta-L-aldopyranose	1c4	yes	methyl beta-L-fucopyranoside
MGC	0.565	279.53	1.57	alpha-D-aldopyranose	4c1	yes	methyl 2-acetamido-2-deoxy-alpha-D-galactopyranoside
MGL	0.565	348.90	3.08	beta-D-aldopyranose	4c1	yes	methyl beta-D-glucopyranoside
MGS	0.551	292.07	3.16	alpha-D-aldopyranose	4c1	yes	methyl 4,6-dideoxy-4-{[(2R)-2,4-dihydroxybutanoyl]amino}-2-O-methyl-alpha-D-mannopyranoside
MMA	0.557	259.64	2.21	alpha-D-aldopyranose	4c1	yes	methyl alpha-D-mannopyranoside
MNA	0.541	33.61	176.01	alpha-L-ketopyranose	1c4	yes	2-O-methyl-5-N-acetyl-alpha-D-neuraminic acid
MQT	0.578	322.92	2.79	beta-D-aldopyranose	4c1	yes	methyl 2-O-acetyl-3-O-(4-methylbenzoyl)-beta-D-talopyranoside
MRP	0.551	92.30	177.66	alpha-L-aldopyranose	1c4	yes	3-O-methyl-alpha-L-rhamnopyranose
MUR	0.562	293.29	4.65	beta-D-aldopyranose	4c1	yes	beta-muramic acid

MXZ	0.578	302.54	176.41	alpha-L-aldopyranose	1c4	yes	2-O-methyl-alpha-L-fucopyranose
NAA	0.565	299.76	2.55	beta-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-beta-D-allopyranose
NAG	0.564	300.23	2.93	beta-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-beta-D-glucopyranose
NBG	0.564	9.27	4.48	beta-D-aldopyranose	4c1	yes	N-acetyl-beta-D-glucopyranosylamine
NDG	0.550	244.22	5.24	alpha-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-alpha-D-glucopyranose
NG1	0.550	263.29	4.96	alpha-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-1-O-phosphono-alpha-D-galactopyranose
NG6	0.564	329.01	2.94	beta-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-6-O-sulfo-beta-D-galactopyranose
NGA	0.562	307.41	4.41	beta-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-beta-D-galactopyranose
NGK	0.546	263.09	6.19	alpha-D-aldopyranose	4c1	yes	2-acetamido-2-deoxy-4-O-sulfo-alpha-D-galactopyranose
NGZ	0.551	55.73	175.71	alpha-L-aldopyranose	1c4	yes	2-acetamido-2-deoxy-alpha-L-glucopyranose
NNG	0.557	229.00	4.63	alpha-D-aldopyranose	4c1	yes	2-deoxy-2-{[(S)-hydroxy(methyl)phosphoryl]amino}-6-O-phosphono-alpha-D-glucopyranose
NOK	0.569	71.73	3.73	beta-D-aldopyranose	4c1	yes	2-acetamido-1,2-dideoxynojirmycin
NTF	0.564	8.02	4.30	beta-D-aldopyranose	4c1	yes	N-(trifluoroacetyl)-beta-D-glucopyranosylamine
NXD	0.545	49.28	175.95	beta-L-ketopyranose	1c4	yes	methyl 5-acetamido-9-{[amino(oxo)acetyl]amino}-3,5,9-trideoxy-D-glycero-alpha-D-galacto-non-2-ulopyranosidonic acid
OAK	0.567	31.24	5.66	beta-D-aldopyranose	4c1	yes	N-(phenylcarbonyl)-beta-D-glucopyranosylamine
OPM	0.549	257.55	2.17	alpha-D-aldopyranose	4c1	yes	pentyl alpha-D-mannopyranoside
OTG	0.539	252.54	7.21	alpha-D-aldopyranose	4c1	yes	2-deoxy-2-{[(2-methylphenyl)carbonyl]amino}-alpha-D-glucopyranose
OX2	0.566	332.14	2.73	beta-D-aldopyranose	4c1	yes	(1R)-1,5-anhydro-1-(5-methyl-1,3,4-oxadiazol-2-yl)-D-glucitol
PA1	0.554	234.78	4.63	alpha-D-aldopyranose	4c1	yes	2-amino-2-deoxy-alpha-D-glucopyranose
PDX	0.555	196.44	1.26	alpha-D-aldopyranose	4c1	yes	2,3-di-O-sulfo-alpha-D-glucopyranose
PH5	0.541	32.39	175.80	alpha-L-ketopyranose	1c4	yes	benzyl 3,5-dideoxy-5-(propanoylamino)-D-glycero-alpha-D-galacto-non-2-ulopyranosidonic acid
PNA	0.558	251.39	3.29	alpha-D-aldopyranose	4c1	yes	4-nitrophenyl alpha-D-mannopyranoside
PNG	0.558	247.45	1.78	alpha-D-aldopyranose	4c1	yes	4-nitrophenyl alpha-D-glucopyranoside
PNJ	0.566	319.97	4.34	beta-D-aldopyranose	4c1	yes	4-nitrophenyl 2-amino-2-deoxy-beta-D-glucopyranoside
PNW	0.573	357.34	3.30	beta-D-aldopyranose	4c1	yes	4-nitrophenyl beta-D-glucopyranoside
PSG	0.560	3.32	3.74	beta-D-aldopyranose	4c1	yes	4-nitrophenyl 1-thio-beta-D-glucopyranoside
RAM	0.551	93.67	175.46	alpha-L-aldopyranose	1c4	yes	alpha-L-rhamnopyranose
RAO	0.560	74.22	177.17	alpha-L-aldopyranose	1c4	yes	methyl 6-deoxy-alpha-L-rhamnopyranoside
RER	0.508	92.81	172.03	alpha-L-aldopyranose	1c4	yes	vancosamine
RGG	0.567	343.74	3.59	beta-D-aldopyranose	4c1	yes	(2R)-2,3-dihydroxypropyl beta-D-galactopyranoside
RIP	0.564	333.38	2.88	beta-N-aldopyranose	4c1	yes	beta-D-ribopyranose

RM4	0.570	134.59	176.74	beta-L-aldopyranose	1c4	yes	Beta-L-rhamnopyranose
RUG	0.573	46.59	2.13	beta-D-aldopyranose	4c1	yes	1-beta-D-glucopyranosyl-4-(hydroxymethyl)-1H-1,2,3-triazole
S06	0.539	227.03	3.51	alpha-D-aldopyranose	4c1	yes	(3S,5R,7R,8S,9S,10R)-7-(hydroxymethyl)-3-(2-naphthyl)-1,6-dioxa-2-azaspiro[4.5]decane-8,9,10-triol
S13	0.540	227.54	3.66	alpha-D-aldopyranose	4c1	yes	(3S,5R,7R,8S,9S,10R)-7-(hydroxymethyl)-3-(4-methylphenyl)-1,6-dioxa-2-azaspiro[4.5]decane-8,9,10-triol
SFU	0.570	176.94	177.84	alpha-L-aldopyranose	1c4	yes	methyl 1-seleno-alpha-L-fucopyranoside
SGA	0.564	295.07	3.87	beta-D-aldopyranose	4c1	yes	3-O-sulfo-beta-D-galactopyranose
SGC	0.553	346.52	4.24	beta-D-aldopyranose	4c1	yes	4-thio-beta-D-glucopyranose
SGN	0.543	255.36	9.21	alpha-D-aldopyranose	4c1	yes	2-deoxy-6-O-sulfo-2-(sulfoamino)-alpha-D-glucopyranose
SHG	0.562	286.68	3.36	beta-D-aldopyranose	4c1	yes	2-deoxy-2-fluoro-beta-D-glucopyranose
SIA	0.539	61.92	175.13	alpha-L-ketopyranose	1c4	yes	N-acetyl-alpha-neuraminic acid
SID	0.542	41.16	176.14	alpha-L-ketopyranose	1c4	yes	methyl 9-S-acetyl-5-acetamido-3,5-dideoxy-9-thio-D-glycero-alpha-D-galacto-non-2-ulopyranosidonic acid
SLB	0.539	69.21	175.93	beta-L-ketopyranose	1c4	yes	N-acetyl-beta-neuraminic acid
SLM	0.543	65.58	175.84	beta-L-ketopyranose	1c4	yes	(2S,4S,5R,6R)-5-acetamido-2,4-dihydroxy-6-[(1R,2R)-1,2,3-trihydroxypropyl]oxane-2-carboxamide
SN5	0.563	303.83	3.63	beta-D-aldopyranose	4c1	yes	2-deoxy-2-(ethanethioylamino)-beta-D-glucopyranose
SOE	0.551	321.70	176.81	alpha-N-ketopyranose	1c4	yes	alpha-L-sorbopyranose
SOG	0.564	4.35	4.65	beta-D-aldopyranose	4c1	yes	octyl 1-thio-beta-D-glucopyranoside
SSG	0.559	5.99	4.84	beta-D-aldopyranose	4c1	yes	1,4-dithio-beta-D-glucopyranose
STZ	0.563	298.49	3.44	beta-D-aldopyranose	4c1	yes	streptozotocin
SUS	0.550	194.66	2.29	alpha-D-aldopyranose	4c1	yes	2-deoxy-3,6-di-O-sulfo-2-(sulfoamino)-alpha-D-glucopyranose
TGA	0.567	331.97	3.56	beta-D-aldopyranose	4c1	yes	2-sulfanylethyl beta-D-galactopyranoside
TMR	0.555	313.66	3.21	beta-D-aldopyranose	4c1	yes	2,6-dideoxy-4-S-methyl-4-thio-beta-D-ribo-hexopyranose
TOA	0.534	286.20	4.28	alpha-D-aldopyranose	4c1	yes	3-ammonio-3-deoxy-alpha-D-glucopyranose
TOC	0.538	299.00	3.14	alpha-D-aldopyranose	4c1	yes	2,6-diammonio-2,3,6-trideoxy-alpha-D-glucopyranose
TYV	0.550	304.50	1.43	alpha-D-aldopyranose	4c1	yes	alpha-D-Tyvelopyranose
X1P	0.555	228.40	1.07	alpha-N-aldopyranose	4c1	yes	1-O-phosphono-alpha-D-xylopyranose
XYP	0.560	329.52	4.07	beta-N-aldopyranose	4c1	yes	Beta-D-xylopyranose
XYS	0.554	259.31	1.86	alpha-N-aldopyranose	4c1	yes	alpha-D-xylopyranose
YX0	0.558	83.60	177.05	alpha-L-aldopyranose	1c4	yes	[(3E)-3-(1-hydroxyethylidene)-2,3-dihydroisoxazol-5-yl]methyl alpha-L-fucopyranoside
YX1	0.569	291.93	5.27	beta-D-aldopyranose	4c1	yes	2-deoxy-2-{[(2-hydroxy-1-methylhydrazino)carbonyl]amino}-beta-D-glucopyranose

Table A.1. validation results for all the produced conformers as calculated by Privateer. Column legend: 'CCD' is the three-letter code assigned by the CCD; the Cremer-Pople parameters (Cremer and Pople 1975a), named 'Q', 'Phi' and 'Theta' describe pyranose ring conformation (denoted as 'Cnf' here), with 'Phi' and 'Theta' describing what atoms move away from the average ring plane, and 'Q' (termed 'total puckering amplitude' by Cremer & Pople, measured in $Å^2$) dictating by how much; 'Detected type' describes the monosaccharide in terms of anomeric form, absolute stereochemistry, position of the carbonyl group (aldose or ketose) and ring shape (all pyranose in this study); 'Ok?' presents the result of Privateer's tri-state validation diagnosis, as introduced in the main text; Lastly, 'name' is the full IUPAC name of the monosaccharide.

Appendix B.

Analysis and validation of overall N-glycan conformation in Privateer

Jordan S. Dialpuri[1], Haroldas Bagdonas[1], Mihaela Atanasova[1], Lucy Schofield[1], Maarten L. Hekkelman[2], Robbie P. Joosten[2,*] and Jon Agirre[1,*]

[1] York Structural Biology Laboratory, Department of Chemistry, University of York, UK.

[2]Oncode Institute and Division of Biochemistry, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.

[*]Correspondence: r.joosten@nki.nl or jon.agirre@york.ac.uk

Abstract

The oligosaccharides in N-glycosylation provide key structural and functional contributions to a glycoprotein. These contributions are dependent on the glycans' composition and overall conformation. The Privateer software allows structural biologists to evaluate and improve atomic structures of carbohydrates, including N-glycans; this software was recently extended to check glycan composition through the use of glycomics data. Here, we present a broadening of the software's scope to analyse and validate the overall conformation of N-glycans, focusing on a newly compiled set of glycosidic linkage torsional preferences harvested from a curated set of glycoprotein models.

1. Introduction

Post-translational modifications (PTMs) are covalent modifications of proteins that occur after the nascent polypeptide has left the ribosome. PTMs may induce significant changes to the protein's structure and function (Xin and Radivojac 2012). A fundamental and abundant PTM is N-glycosylation, where an oligosaccharide moiety is attached to the nitrogen atom of an asparagine side-chain in the target protein. The oligosaccharide is subsequently trimmed and modified according to the available cellular enzymes – glycoside hydrolases, glycosyl and oligosaccharyl transferases. The resulting oligosaccharide, or N-glycan, may end up having anything from a complex to a minimal composition, leading to a specific 3D conformation of the mature glycoprotein (Shental-Bechor and Levy 2009). N-glycosylation is key in all sorts of interactions, including with cell surface receptors (Petrescu, Wormald, and

Dwek 2006; Pauline M. Rudd, Wormald, and Dwek 2004) or even other parts of the same glycoprotein, as evidenced in studies of the dynamics of the SARS-CoV-2 spike where conformational changes in the N165 glycan push up the spike's own receptor binding domain (Casalino et al. 2020).

Understanding the complex structure of carbohydrates is challenging due to various stereochemical and regiochemical possibilities exhibited in N-glycans. Producing a correct 3D structure of a glycoprotein at good enough resolution can be vital in understanding how some biological processes unfold. Alas, working with glycans in software for X-ray crystallography and Electron Cryo-microscopy has historically been all but straightforward: many carbohydrate modelling, refinement, and validation processes relied on software written primarily for proteins and nucleic acids (Atanasova, Bagdonas, and Agirre 2020), and libraries of restraints had become outdated or were incorrect (Agirre 2017). While recent efforts have aimed to address this situation (Atanasova et al. 2022b; Joosten, Nicholls, and Agirre 2022), carbohydrate methodology still trails those designed for proteins.

Obtaining a glycoprotein structure at a high enough resolution can generally be considered more difficult than doing it with a glycan-free protein. Two main issues are routinely identified as problematic when it comes to obtaining higher resolutions: heterogeneity and mobility, both of which translate into poorer experimental data. Owing to these complications, the Protein Data Bank (PDB) (H. M. Berman et al. 2000) contains models that include wrong nomenclature (Lütteke, Frank, and von der Lieth 2005), impossible linkages (Crispin, Stuart, and Jones 2007), and improbably high energy conformations of carbohydrates that deviate from the low energy chair conformation of six-membered rings (Agirre, Davies, et al. 2015a) – in general, a $^{4}C_{1}$ chair for d-pyranosides and $^{1}C_{4}$ chair for l-pyranosides; ring conformations (Cremer and Pople 1975b) and their energetics (Davies, Planas, and Rovira 2012) are discussed in detail elsewhere. Using models with wrong glycochemistry in downstream analyses or molecular simulations will cause misrepresentation and misinterpretation, while also perpetuating these errors. Software packages such as PDB-CARE and CARP (Lütteke, Frank, and von der Lieth 2005), and more recently Privateer (Agirre, Iglesias-Fernández, et al. 2015a; Bagdonas, Ungar, and Agirre 2020a) can be utilised for the identification and rectification of these model errors, therefore allowing future refinement data libraries to be as accurate and representative as possible.

In this study, torsion angles (dihedral angles) between curated structures of N-glycan forming pyranosides were collected in order to create accurate torsional libraries for use in the Privateer validation software. Previous torsional databases such as GlyTorsionDB (Lütteke, Frank, and von der Lieth 2005) and its associated link checking tool (CARP)

incorporate potentially flawed models from the PDB, as they pre-dated the introduction of ring conformation in the routine validation of glycan structures (Agirre, Iglesias-Fernández, et al. 2015a); therefore, a survey of the PDB was completed, with each PDB entry being analysed and validated using Privateer to ensure the N-glycans were well fit to the electron density and without any conformational errors. Also, in order to avoid the presentation of data on multiple torsional plots and to allow the easy identification of standout (outlier) linkage conformations, a Z score is calculated for each linkage, with standout linkages being highlighted in orange on glycan diagrams that follow the Standard Symbol Nomenclature for Glycans (SNFG), third edition (Varki et al. 2015). Furthermore, in recognition that not every standout linkage conformation will be the consequence of a modelling mistake, a collection of verified cases where the interaction between glycan and protein residues has caused an unusual conformation is presented. Finally, a similar study was completed using PDB-REDO (van Beusekom, Touw, et al. 2018), to analyse whether modern refinement techniques can achieve less frequent errors in the N-glycan models.

2. Materials and Methods

2.1 Dataset collection and validation

A local PDB mirror (August 2021) was created for this study. The PDB mirror was then scanned for proteins containing glycosylated amino acid residues. Of the monosaccharides contained within these chains, the conformation of the 6-membered rings (pyranosides) were validated using Privateer: the software calculates ring conformation using the Cremer-Pople algorithm (Cremer and Pople 1975b), and then compares the detected ring conformation to the minimal energy one it stores in an internal database. The dataset was filtered to include only monosaccharides with a Real Space Correlation Coefficient higher than 0.80 – RSCC (Equation 1) is a measure of local agreement between a portion of an atomic model and the observed electron density map that surrounds it – and which had been deemed diagnostically correct by Privateer, i.e. no nomenclature errors, no unphysical puckering amplitude and all pyranosides in their minimal energy conformations (a chair in all analysed cases). Privateer checks that the anomeric and absolute stereochemistry in the structure matches the one encoded in the three-letter code (*e.g.* that a monosaccharide modelled as **MAN** is perceived to be **α**-**D**-mannose), that the ring conformation matches the lowest energy pucker – a $^{4}C_{1}$ chair for most D-pyranosides, with special cases like $^{1}C_{4}$ for the mannose moiety in tryptophan mannosylation (Akkermans et al. 2022; Martin Frank et al. 2020) – including puckering amplitude (Cremer and Pople 1975b).

$$RSCC = corr(\rho_{obs}, \rho_{calc}) = \frac{cov(\rho_{obs}, \rho_{calc})}{[var(\rho_{obs})\, var(\rho_{calc})]^{1/2}} \quad (1)$$

No resolution cut-offs were explicitly applied, although some filtering is implicit in requiring a minimum RSCC, as the accumulation of model-error components at low resolutions makes it harder to obtain high RSCC values. In total, 68,541 monosaccharides were analysed, 57,569 of which Privateer deemed correct – only these were used in the study. A further 8,511 showed a high-energy ring conformation, which normally requires manual assessment. A total of 2,421 monosaccharides showed geometry and/or nomenclature errors.

For the PDB-REDO comparison, the equivalent monosaccharides were taken from the so-called “conservatively optimised” models in the PDB-REDO databank (van Beusekom, Touw, et al. 2018), *i.e.* models that were re-refined without any torsional restraints for carbohydrates, but were not subjected to N-glycan rebuilding procedures (van Beusekom et al. 2019).

Example linkages present in diverse glycans are shown in Figure 1, using the Standard Symbol Nomenclature for Glycans (SNFG), third edition (Varki et al. 2015), which Privateer implements. The definition of – for N-acetyl β-D-glucosamine (GlcNAc, or NAG in the PDB Chemical Component Dictionary) to asparagine, plus all 1-2, 1-3 and 1-4 glycosidic bonds – and additionally ω – covering those 1,6 bonds such as α-D-mannose–1,6–α-D-mannose or α-L-fucose–1,6–N-acetyl β-D-glucosamine – is shown in Figure 2. While completing this study, a large array of different linkages were identified, however, only a small number had enough independent observations to enable meaningful data extraction – indeed, only approximately 10% of protein models deposited in the PDB contain one or more carbohydrate groups, while around 6% are N-glycosylated (Agirre 2017). We set the minimum number of required observations to 50, and introduced a mechanism for Privateer to report what linkages could not be validated due to insufficient data – *vide infra*. A table of the linkages investigated in this study are given in Table 1 as well as the commonly used abbreviations associated with them.

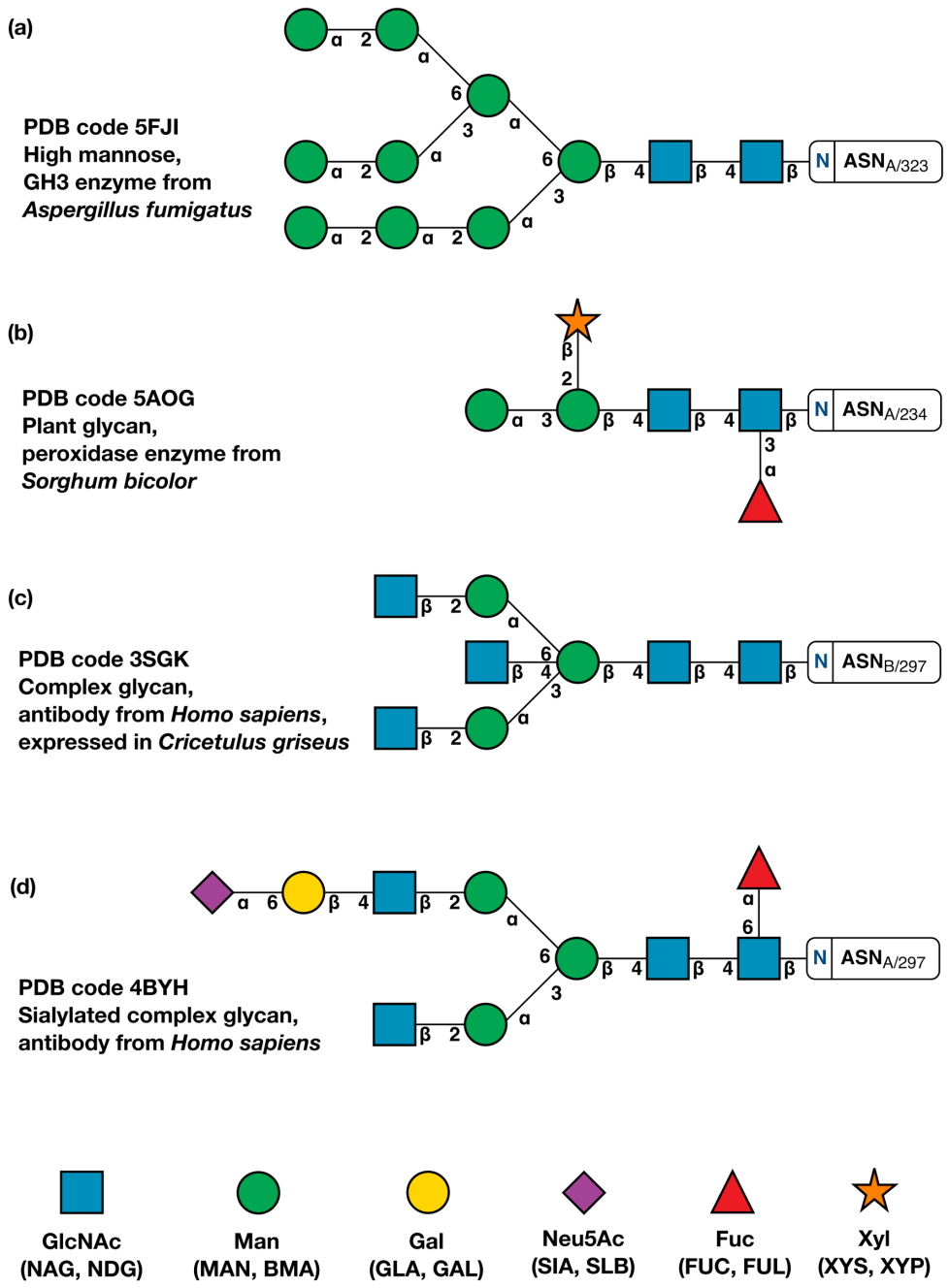


Figure 1

Examples of different types of N-glycans shown using the Symbol Nomenclature for Glycans (SNFG). The greek letters and numbers show the N-glycan linkages naming. (a) High mannose from PDB entry 5FJI, a GH3 glucosidase from *Aspergillus fumigatus* (Agirre et al. 2016). (b) Plant glycan from PDB entry 5AOG, a sorghum peroxidase (Nnamchi et al. 2016). (c) PDB entry 3SGK (Ferrara et al. 2011) shows a complex glycan from an Fc fragment of a human antibody, which was in turn expressed in CHO cells. (d) A sialylated complex glycan from PDB entry 4BYH (Crispin, Yu, and Bowden 2013), expressed in *Homo sapiens*. This

figure was produced with Privateer, which follows the Standard Symbol Nomenclature For Glycans – or SNFG – version 3 (Varki et al. 2015).

Full Linkage denomination	Abbreviation	CCD codes
N-acetyl β-D-glucosamine–asparagine	GlcNAc–β–Asn	NAG-ASN
N-acetyl β-D-glucosamine–1,4–N-acetyl β-D-glucosamine	GlcNAc–β–GlcNAc	NAG-1,4-NAG
β-D-mannose–1,4–N-acetyl β-D-glucosamine	Man–β1,4–Man	BMA-1,4-NAG
α-D-mannose–1,3–β-D-mannose α-D-mannose–1,6–β-D-mannose	Man–α1,3–Man Man–α1,6–Man	MAN-1,3-BMA MAN-1,6-BMA
α-D-mannose–1,2–α-D-mannose α-D-mannose–1,3–α-D-mannose α-D-mannose–1,6–α-D-mannose	Man–α1,2–Man Man–α1,3–Man Man–α1,6–Man	MAN-1,2-MAN MAN-1,3-MAN MAN-1,6-MAN
α-L-fucose–1,3–N-acetyl β-D-glucosamine α-L-fucose–1,6–N-acetyl β-D-glucosamine	Fuc–α1,3–GlcNAc Fuc–α1,6–GlcNAc	FUC-1,3-NAG FUC-1,6-NAG
N-acetyl β-D-glucosamine–1,2–α-D-mannose	GlcNAc–β1,2–Man	NAG-1,2-MAN
β-D-galactose–1,4–N-acetyl β-D-glucosamine	Gal–β1,4–GlcNAc	GAL-1,4-NAG
α-sialic acid–2,6–β-D-galactose	Sia–α2,6–Gal	SIA-2,6-GAL

Table 1

Full name, linkage abbreviations and shorthand notation with PDB Chemical Component Dictionary (CCD) codes for those linkages with enough data. No anomeric data is displayed for CCD codes, as they integrate that information in the codes themselves – *e.g.* MAN is α-D-mannose, BMA is β-D-mannose.

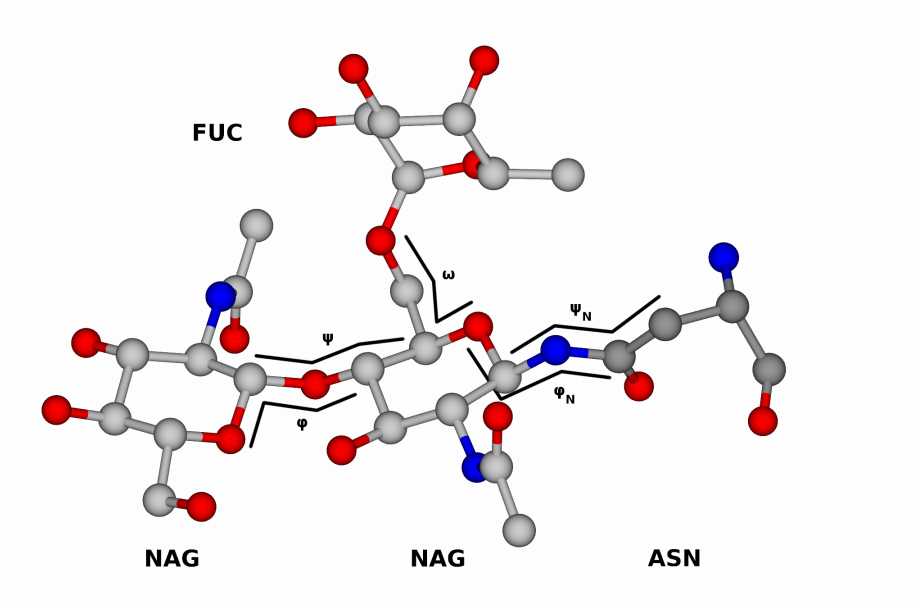


Figure 2
Visual representation of the φ and ψ in both the sugar-sugar linkages and the NAG-ASN linkage. Figure generated from PDB 4BYH (Crispin, Yu, and Bowden 2013)

2.2 Implementation in Privateer

To assess the normality of torsion angle between monosaccharides in N-glycans, a Z score system was implemented using similar methods to software programs *Tortoize* (van Beusekom, Joosten, et al. 2018) and WHAT_CHECK (Hooft, Sander, and Vriend 1997). The Z-score is based on how common a certain φ,ψ-combination is compared to a reference set of the same glycosidic linkages from high quality structure models. To calculate the Z scores, torsional data from each linkage was split into two dimensional bins with a 2 ° bin spacing and formed into a database.

From this, the average and standard deviation of the counts over all bins can be calculated and used to evaluate a new set of torsion angles. The Z score is calculated as described by Hooft *et al* (1997) and shown in equation 2. The correct bin corresponding to the torsion angles in the linkage under survey is selected and the count of that bin in the database is used to calculate the Z score. Therefore, it is important to note that Z scores depend on the amount of data in each bin, not the relative deviation of the torsion angles from the mean.

$$z_k = \frac{c_k^{\,l} - \langle c^l \rangle}{\sigma(c^l)} \quad (2)$$

Where z_k is the z score for the k[th] linkage, $c_k^{\,l}$ is the count in the appropriate bin of that linkage (*l*), $\langle c^{\,l} \rangle$ is the database average for that linkage and $\sigma(c^l)$ is the corresponding standard deviation of the database.

After individual scoring for each linkage type, a global Z score can be calculated by simply averaging the Z scores of all N-glycan linkages. In addition to this, comparison to a reference set of PDB entries with N-glycans allowed the calculation of a relative 'quality Z score' which is an additional parameter that can be used as a measure for glycan normality. The reference set was chosen following a set of criteria: crystallographic structures and reflections from the wwPDB with Rfree < 0.25, reported resolution ≤ 2.50 Å, with glycans longer than 4 pyranosides and composition backed up by a GlyConnect ID (Alocci et al. 2019). As a result, 510 structures were chosen, containing 59 unique glycan structures. The resolution range covered by the dataset was 1.12 - 2.50 Å, and Rwork/Rfree values were 0.10 - 0.23 / 0.12 - 0.25 respectively.

To provide a visual means of highlighting those linkages having an unusual Z score, the SNFG (Varki et al. 2015) vector engine within Privateer (Stuart McNicholas and Agirre 2017) was modified to create an orange background behind the linkages. Linkages for which not enough data could be collected for validation are marked with a grey background. This representation was used in the figures presented in this study. The representation was also extended to cover the monosaccharides in glycans, so that interesting or problematic models can be quickly identified. We note that an orange background does not automatically mean there is a modelling mistake involved, rather that the linkage is worth inspecting.

3. Results and discussion

The number of N-glycosylated structures in the PDB is growing steadily (Scherbinina and Toukach 2020; Agirre 2017), supported by the introduction of carbohydrate structure modelling and validation tools such as pdb-care (Lütteke and von der Lieth 2004), the N-glycan building module in *Coot* (Paul Emsley and Crispin 2018) and Privateer (Agirre, Iglesias-Fernández, et al. 2015a). However, as the resolvability of pyranosides in N-glycans decreases the further the monosaccharides are from the asparagine residue (Atanasova, Bagdonas, and Agirre 2020), the abundance of the data collected here dwindles for linkages

that form the glycans' antennae. As stated previously, we set a cut-off of 50 data points in order to guarantee the reliability of the Z score calculation, and this necessarily means that some glycosidic linkages are not yet included in the analysis the Privateer software does. Scripts for reproducing and extending this work are included in the relevant section here, meaning that the torsion library can be re-generated in future when more data are available.

The torsional data we harvested are plotted in Figure 3. A first close inspection of the graphs reveals a straightforward correspondence between the most frequent linkage conformations for every link type and their calculated energy minimum or minima in the Disac3-DB section of the Glyco3D 2.0 database (Pérez, Sarkar, et al. 2015) and GlycoMapsDB (M. Frank, Lütteke, and von der Lieth 2007). The mean linkage torsion angles and respective standard deviations of this PDB survey are shown in Supplementary Tables 1 and 2. Supplementary Table 1 shows the values implemented into Privateer. A comparative plot of quality Z scores for the curated dataset *versus* the rest of the PDB is available in Supplementary Figure 1. Low quality Z scores ($Z < -2$) indicate serious problems with the overall quality of glycans in the structure model. High quality Z scores ($Z > 2$), particularly in low resolution structure models may indicate over-restraining of torsions in model refinement and warrant further inspection as previously shown for proteins (Sobolev et al. 2020).

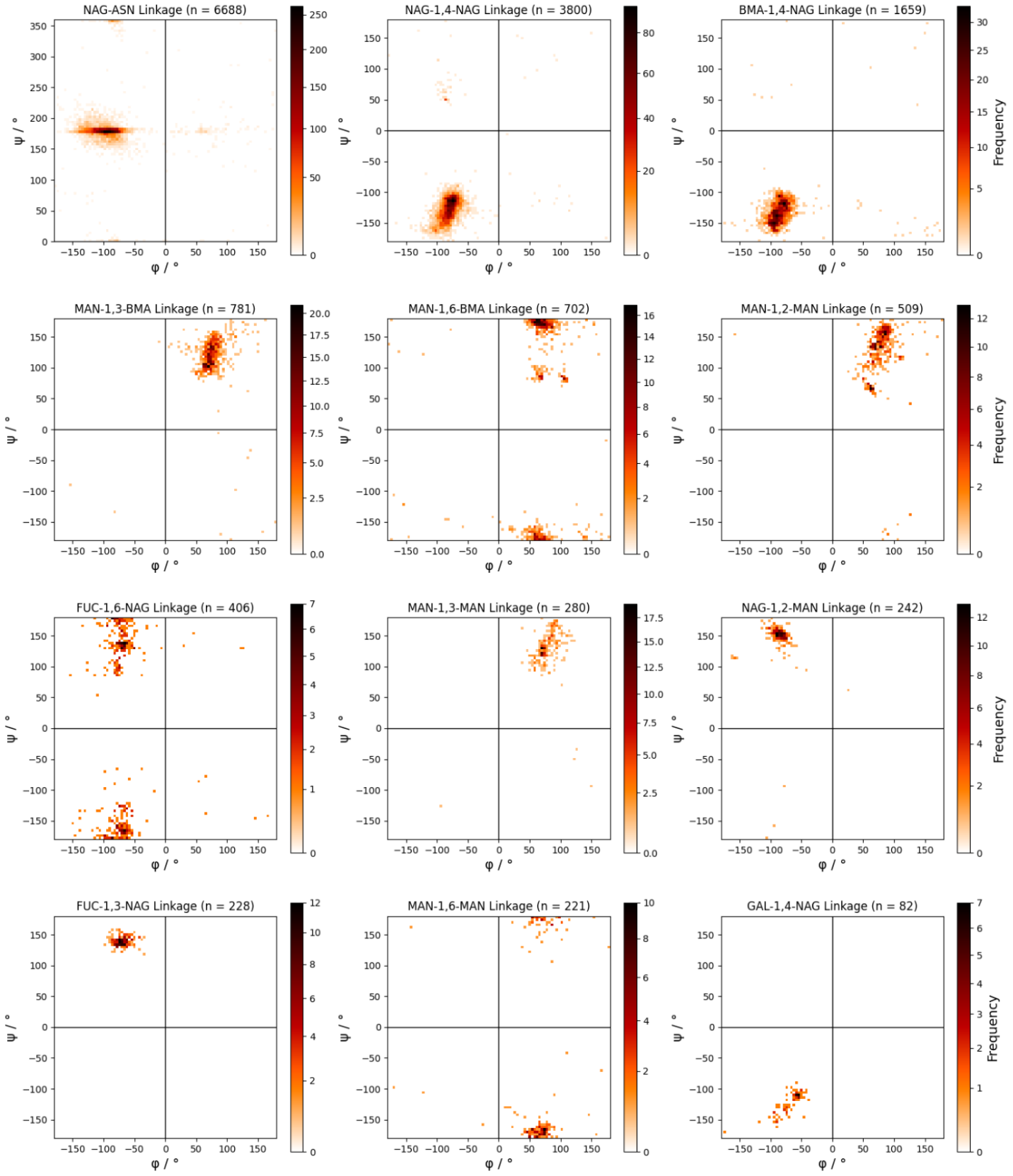


Figure 3

Plots of φ and ψ values for all linkages collected with over 50 data points. Colour bars are

plotted using the Power-Law Distribution (Clauset, Shalizi, and Newman 2009) to highlight outliers visually. Plots allow visualisation of the energy minima values.

3.1 GlcNAc - asparagine bond

Investigations on the torsion angle dataset between the asparagine (ASN) amino acid side-chain and GlcNAc (NAG) highlight a perhaps unsurprising trend. The ϕ torsion angle dataset has a greater standard deviation (σ = 25.3 °) when compared to the ψ torsion angle (σ = 22.1 °). This is most likely due to ψ torsion angle referring to a C-N bond which has a bond order higher than unit, analogous to a peptide bond. Indeed the mean value of the ψ is 178.5 ° which is very similar to the 180 ° torsion angle expected for a peptide bond. Such a bond has limited torsional freedom. The ϕ torsion angle refers to a single bond which has more rotational freedom, leading to the increased spread of torsional data for ϕ.

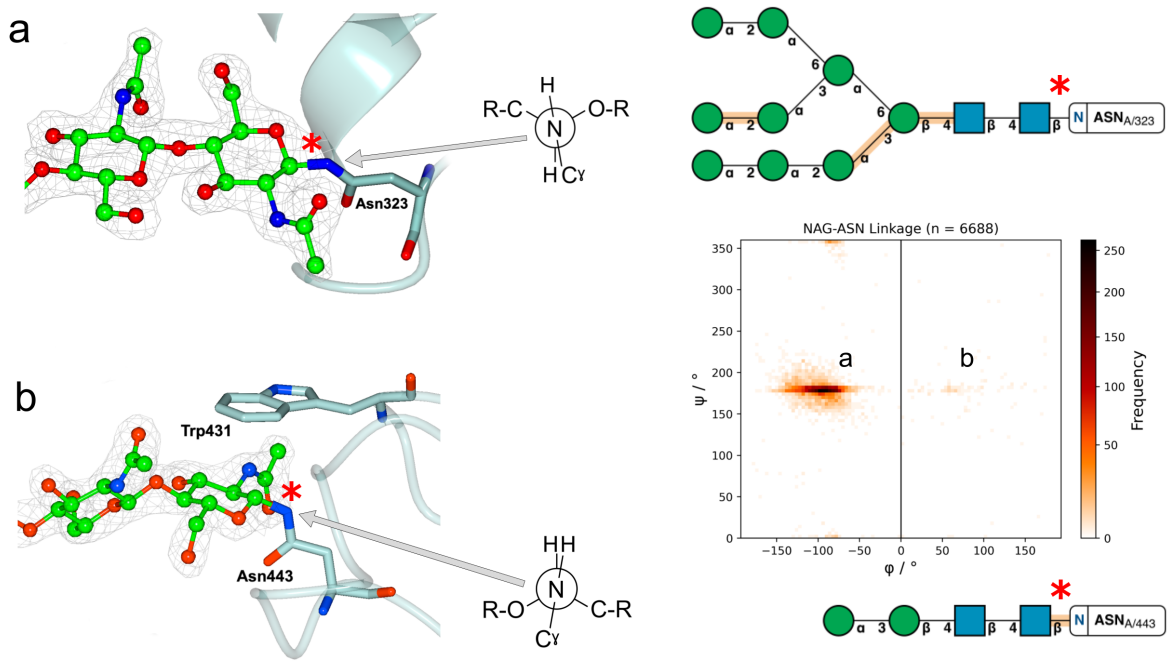


Figure 4

Two main conformations for the NAG-ASN bond are detected in our dataset, as previously shown in the literature (Imberty and Perez 1995). Both shown conformations are from PDB 5FJI (Agirre et al. 2016); 2mFo-DFc electron density is shown at 1σ for the glycans, but omitted for the asparagine side-chains for reasons of clarity – the asparagine side-chains' positions showed a good fit to the electron density. Newman projections for the GlcNAc-Asn linkage are included, with arrows pointing in the direction of the link being inspected, which have also been marked with a red asterisk on both the 3D views and 2D diagrams.

The correct modelling of the protein-sugar linkage torsion angle is particularly important to establish a good basis for other monosaccharides to be modelled further down the N-glycan tree. Two main conformations for the NAG-ASN exist (Figure 4) in which the conformation with a negative ϕ angle (a) is the most abundant and the other one (b), much more infrequent due to the additional CH-π interaction (Trp431) required to stabilise it, is flagged up as an outlier by Privateer. The arrangement shown in Figure 4b, found in a fungal GH3 beta-glucosidase, is conserved across homologous structures.

3.2 Glycosidic linkages between pyranosides

N-glycans exhibit common structures as shown in Figure 1. The similarity of these conformations explains the consistency in the types of linkages seen in various glycoproteins and allows this quantitative study. N-glycosylated chains attach to the residue with a NAG sugar through a beta linkage. Attached to this initial NAG sugar, through a beta-1,4 linkage is an additional NAG sugar. This initial NAG-1,4-NAG linkage is abundant in the PDB and hence contains a large number (n = 3800) of validated data points. As evident by the two dimensional histogram (Figure 3), most NAG-1,4-NAG linkages contain torsion angles around $\varphi \approx -80\,^{\circ}$ and $\psi \approx -130\,^{\circ}$.

Often, a BMA sugar is attached to the second NAG sugar through a beta-1,4 linkage. This BMA-1,4-NAG linkage may theoretically have slightly more conformational variability than NAG-1,4-NAG due to its position further down the glycan tree, however the spread of data (standard deviation) is similar for both NAG-1,4-NAG and BMA-1,4-NAG. In addition to this, in the complex tree, a FUC sugar can be attached to the initial NAG through an alpha-1,6 linkage. The FUC-1,6-NAG linkage exhibits a large standard deviation around both torsion angles, particularly around the ψ angle. This could partially be the result of FUC being a terminal residue at this position in the glycan, but the FUC-1,3-NAG linkage in which the FUC is also a terminal residue connected to the same NAG has less spread in the observed torsion angles. A key difference however is the presence of a third torsion angle, ω that gives more flexibility to the FUC-1,6-NAG linkage. This additional flexibility also leads to less well-defined experimental data and thus more room for modelling errors.

Attachment of additional mannose sugars onto the N-glycan chain can often increase the amount of branching and size of the chain (see Figure 1a). The most common attachment onto the so-far terminal BMA sugar is MAN-1,3-BMA, indeed this is shown in our dataset of validated glycans (n = 781) with the positional isomer MAN-1,6-BMA being almost as

frequent (n = 702). Interestingly, the MAN-1,3-BMA linkage exhibits standard deviations (ψ: σ = 22.6 °) which are similar to NAG-1,4-NAG (ψ: σ = 22.8 °). However, the MAN-1,6-BMA linkage torsion angles do not exist in a singular cluster and hence exhibit a larger standard deviation (ψ: σ = 33.3 °). Again this additional spread may be caused by the presence of a third torsion angle in the linkage.

Certain glycoproteins have further monosaccharide attachments such as a variety of MAN-MAN, NAG-MAN and SIA-GAL linkages. Interestingly, the torsion angle spread for all MAN-MAN linkages (1,2, 1,3 and 1,6) is far greater than the torsion angle spread of NAG-MAN torsion data despite having similar data set size and existing in a similar area of the protein. A reason for this may be the N-acetyl group in NAG which makes placing the monomer in relatively poor density less error prone. The large standard deviation of MAN-MAN linkages causes similar challenges to MAN-BMA linkages with torsional restraint application. As well as this, no apparent cluster was observed for the SIA-GAL linkage, most likely due to the very low number of deposited and curated linkages available in the dataset. The values φ can adopt appear to be determined by the anomeric form involved in the glycosidic linkage: for D-pyranosides, that means -180 º < φ < 0 º for beta anomers, and 0 º < φ < 180 º for alpha anomers. The inverse is true for L-pyranosides.

Using this large torsion angle dataset, an investigation of torsion angle spread with glycan chain length and branching was conducted, although no meaningful trend was identified between glycan chain length and torsion angle standard deviation. Despite this, this large dataset can be incorporated into software packages like Privateer to improve the accuracy of glycoprotein models.

3.3 PDB-REDO Analysis

With the increasingly commonplace solution of protein complexes with high resolution data, it is imperative that model building software can depict the conformation and position of N-glycans accurately. Through the comparison of N-glycan torsion angles of proteins deposited in the PDB and PDB-REDO, the applicability and necessity of modern refinement techniques can be assessed. Comparisons between torsion angles in N-glycans deposited on the PDB and PDB-REDO highlight an interesting relationship between structure resolution and torsion angle accuracy, shown in Table 2.

Resolution	Linkage	PDB : φ / °	PDB-REDO : φ / °	PDB : ψ / °	PDB-REDO : ψ / °	No. of entries
x < 1.50	NAG-1,4-NAG	-79.4 ± 7.8	-78.9 ± 24.2	-126.6 ± 18.0	-126.3 ± 25.7	132
1.50 < x < 3.00	NAG-1,4-NAG	-79.9 ± 12.6	-74.0 ± 24.5	-126.7 ± 22.7	-125.3 ± 24.4	3190
x > 3.00	NAG-1,4-NAG	-83.4 ± 24.2	-67.2 ± 36.0	-130.3 ± 23.5	-134.6 ± 26.9	472
ALL	**NAG-1,4-NAG**	**-80.3 ± 14.5**	**-73.4 ± 26.2**	**-127.2 ± 22.8**	**-126.5 ± 24.9**	**3800**
x < 1.50	BMA-1,4-NAG	-81.9 ± 9.6	-83.6 ± 10.2	-124.9 ± 14.1	-121.8 ± 13.5	37
1.50 < x < 3.00	BMA-1,4-NAG	-87.2 ± 16.2	-78.7 ± 29.0	-133.4 ± 17.6	-136.0 ± 23.1	1369
x > 3.00	BMA-1,4-NAG	-85.0 ± 26.2	-64.8 ± 46.8	-133.8 ± 21.0	-142.4 ± 26.0	250
ALL	**BMA-1,4-NAG**	**-86.8 ± 17.9**	**-77.1 ± 32.2**	**-133.3 ± 18.2**	**-136.9 ± 23.6**	**1659**
x < 1.50	MAN-1,6-BMA	68.7 ± 6.0	70.0 ± 4.8	150.2 ± 45.0	148.8 ± 44.9	17
1.50 < x < 3.00	MAN-1,6-BMA	71.7 ± 24.5	66.6 ± 24.0	167.0 ± 33.0	167.2 ± 33.7	606
x > 3.00	MAN-1,6-BMA	79.3 ± 42.3	66.4 ± 40.7	176.6 ± 30.7	179.4 ± 34.4	75
ALL	**MAN-1,6-BMA**	**72.4 ± 26.6**	**66.4 ± 26.3**	**167.7 ± 33.3**	**168.2 ± 34.8**	**702**
x < 1.50	MAN-1,3-BMA	77.1 ± 14.4	76.4 ± 14.4	121.7 ± 20.6	122.5 ± 21.1	23
1.50 < x < 3.00	MAN-1,3-BMA	74.6 ± 16.5	68.8 ± 20.0	120.8 ± 20.6	125.9 ± 23.5	602
x > 3.00	MAN-1,3-BMA	81.5 ± 21.5	67.9 ± 26.2	125.2 ± 30.4	134.9 ± 33.6	130
ALL	**MAN-1,3-BMA**	**75.8 ± 16.8**	**68.9 ± 21.1**	**121.5 ± 22.6**	**127.2 ± 25.6**	**777**
x < 1.50	MAN-1,6-MAN	59.6 ± 5.6	60.0 ± 3.5	-178.6 ± 5.9	-177.0 ± 4.3	8
1.50 < x < 3.00	MAN-1,6-MAN	66.6 ± 18.6	65.0 ± 19.6	-172.7 ± 16.1	-171.4 ± 16.0	175

x > 3.00	MAN-1,6-MAN	82.7 ± 44.8	67.8 ± 46.1	-174.4 ± 34.4	-179.9 ± 48.9	38
ALL	**MAN-1,6-MAN**	**68.5 ± 25.0**	**65.2 ± 25.3**	**-173.3 ± 20.0**	**-172.9 ± 24.0**	**221**

x < 1.50	MAN-1,2-MAN	73.0 ± 12.5	72.5 ± 12.1	126.2 ± 37.1	124.8 ± 37.5	23
1.50 < x < 3.00	MAN-1,2-MAN	77.3 ± 15.7	70.4 ± 15.9	134.5 ± 33.0	139.1 ± 35.5	387
x > 3.00	MAN-1,2-MAN	82.1 ± 25.3	71.0 ± 28.3	125.1 ± 25.6	130.2 ± 30.1	94
ALL	**MAN-1,2-MAN**	**77.9 ± 17.8**	**70.6 ± 18.6**	**132.3 ± 32.1**	**136.8 ± 34.9**	**507**

x < 1.50	MAN-1,3-MAN	74.2 ± 5.5	73.0 ± 5.9	118.4 ± 17.2	118.3 ± 17.8	9
1.50 < x < 3.00	MAN-1,3-MAN	77.0 ± 16.1	75.2 ± 16.0	133.2 ± 22.5	135.0 ± 24.0	234
x > 3.00	MAN-1,3-MAN	89.1 ± 18.7	82.6 ± 19.4	129.2 ± 33.8	130.2 ± 33.1	36
ALL	**MAN-1,3-MAN**	**78.5 ± 16.8**	**76.0 ± 16.4**	**132.3 ± 24.2**	**133.9 ± 25.3**	**280**

Table 2
Comparison between the PDB and PDB-REDO torsional data.

The PDB-REDO models used in this study had no torsional restraints applied during refinement. Therefore, torsion angles calculated from PDB-REDO are not influenced by the potentially flawed torsional restraints applied before the model was deposited to the PDB initially. This application of consistent refinement techniques without torsional restraints leads to a dataset which naturally has a larger spread than the PDB. To assess whether the datasets from the PDB and PDB-REDO are significantly different, a series of t-tests were performed and summarised in Table 3.

For linkages NAG-1,4-NAG and BMA-1,4-NAG, the PDB and PDB-REDO both mean torsion angles were deemed significantly different ($p < 0.05$) by the t-test. For the MAN-1,6-BMA linkage, while the φ angle was deemed significantly different, the ψ angle was not significantly different. Interestingly, both datasets showed no significant difference between both torsion angles for linkages MAN-1,6-MAN and MAN-1,3-MAN. While PDB-REDO had many occurrences whereby the torsion angles were not statistically similar to the PDB, both datasets torsion angles exist within 1 standard deviation of each other for every linkage.

These repetitive small differences could be attributed to any torsional restraints applied before deposition to the PDB, hence providing a strong case for using the torsion angles collected using PDB-REDO. With the absence of the aforementioned torsion restraints, it is likely that PDB-REDO represents a more realistic distribution of N-glycan torsion angles which may be useful to incorporate during modelling in the future. It is possible or even likely that using these updated torsion angle libraries will allow for more accurate N-glycosylated chains in future models. A future update of Privateer will allow switching between both PDB and PDB-REDO torsional sets.

Linkage	**Resolution Range**	**t-test result : φ**	**t-test result : ψ**
NAG-1,4-NAG	0.93 - 6.92	Significantly Different (p = 0)	Significantly Different (p = 0)
BMA-1,4-NAG	1.20 - 8.69	Significantly Different (p = 0)	Significantly Different (p = 0)
MAN-1,6-BMA	1.20 - 6.92	Significantly Different (p = 0.0022)	**Not Significantly Different (p = 0.34)**
MAN-1,3-BMA	1.20 - 6.92	Significantly Different (p = 0)	**Not Significantly Different (p = 0.39)**
MAN-1,6-MAN	1.12 - 6.31	**Not Significantly Different (p = 0.14)**	**Not Significantly Different (p = 0.35)**
MAN-1,2-MAN	1.20 - 6.92	Significantly Different (p = 0)	**Not Significantly Different (p = 0.18)**
MAN-1,3-MAN	1.20 - 6.31	**Not Significantly Different (p = 0.12)**	**Not Significantly Different (p = 0.56)**

Table 3

Results of t-tests between PDB and PDB-REDO dataset with all resolutions. Not significantly different values ($p > 0.05$) are shown in bold.

The application of consistent refinement techniques was also shown to improve outliers which had no physical basis for occurring (little clear interaction with residue or other ligands). Figure 5 highlights the correction PDB-REDO applies to the initially skewed MAN-1,6-BMA linkage. The dataset of linkages originating from the PDB has numerous instances similar to this in which PDB-REDO corrects to the torsion angles to more reasonable values. This powerful correction is another interesting and useful feature that PDB-REDO facilitates.

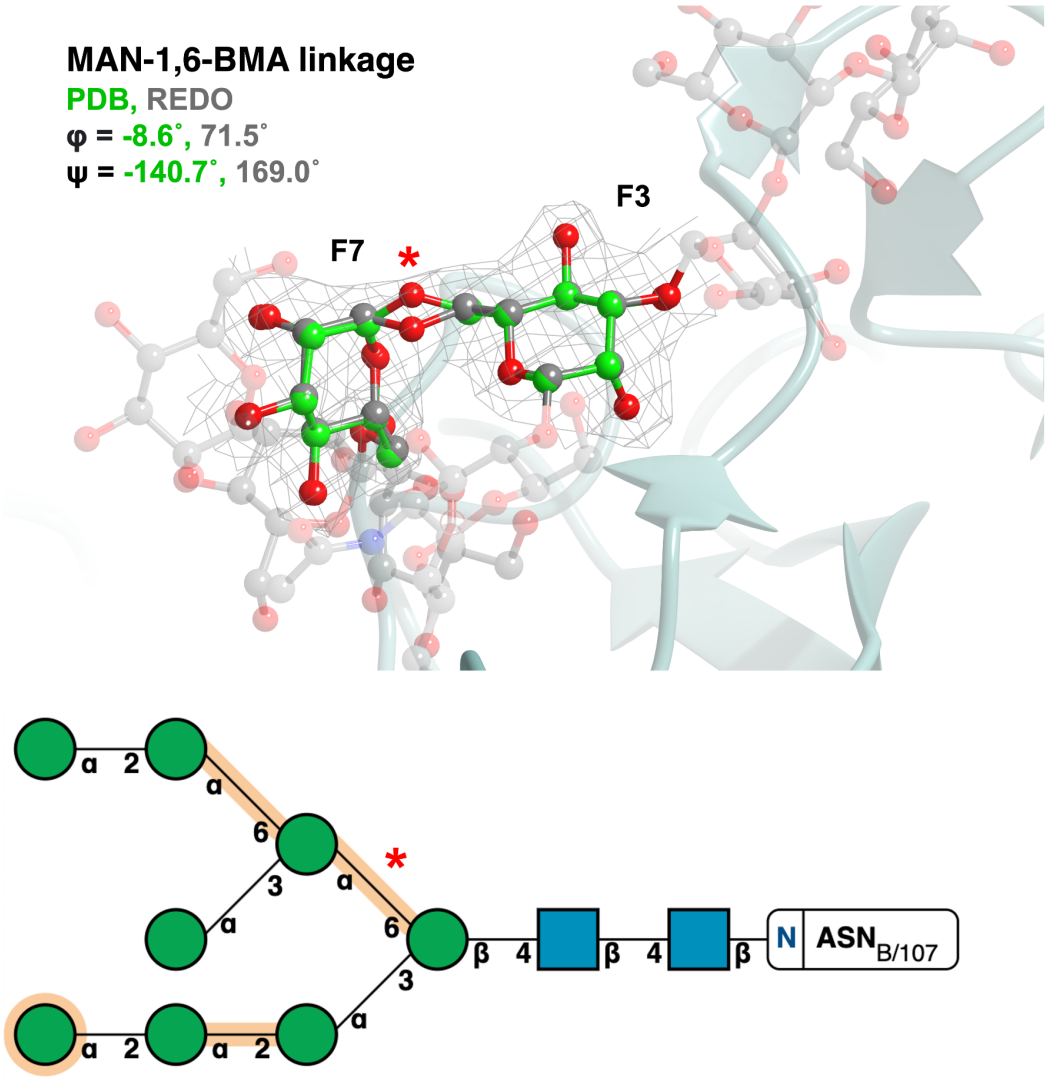


Figure 5

Refinement of PDB 6S2G (Ramirez-Escudero et al. 2019) in PDB-REDO changes the torsion angle from an outlier in PDB to inlier in REDO. The MAN (chain ID and sequence number: F7) -1,6- BMA (chain ID and sequence number: F3) linkage (red asterisk on the bottom panel) of 6S2G (green) identified as an outlier in the PDB (φ = -8.6 ˚, ψ = -140.7 ˚), but an inlier in PDB-REDO (φ = 71.5 ˚, ψ = 169.0 ˚): Δφ = 80.1 ˚, Δψ = 50.3 ˚. The change is brought on by moving the O6 atom (red asterisk on the top panel). BMA(F3) and MAN(F7) are represented by a ball-and-stick model (carbon atoms in green (PDB) or grey (PDB-REDO)), whilst the rest of the attached glycan (PDB-REDO) is represented in faded grey ball and stick. $2F_o$-F_c electron density (grey) is displayed for the linkage contoured to 1σ. The Z scores for this linkage in the PDB model is -1.03 and 1.53 in the PDB-REDO model. Top - produced using CCP4mg. Bottom - SNFG notation output from Privateer.

3.4 Outlier analysis

This analysis of N-glycan torsion angles deposited in the PDB reveals clusters of abundant torsion angles, as shown in Figure 3. Perhaps due to the inherent variability in the environment surrounding monosaccharides in N-glycans, these torsion angle clusters are spread over a large range in most cases. Outliers were quantified as any linkage which had a Z score which was lower than -1. The Z score reported here depends on the amount of φ/ψ pairs relative to the database (Figure 3) and not the deviation from the mean. The limit of -1 was chosen to highlight linkages that are uncommon in the database. Examining these linkages in further detail may highlight the cause of this. As always, surprising cases may either be chemically interesting to look at, or wrong. Here we present one example of each.

3.4.1 Electrostatic Interactions

Repulsive and attractive electrostatic interactions are crucial for the functionality and stability of proteins (Law et al. 2006). These interactions are facilitated by both positively charged (lysine and arginine) and negatively charged (glutamic acid and aspartic acid) amino acid side chains. Similarly, these amino acids can affect the positions of monosaccharides contained within N-glycans via varying degrees of electrostatic interactions.

Figure 6 depicts an N-glycan with MAN-1,2-MAN (PDB code: 4J0M) torsion angles that are highly deviated from the mean. Since this glycan has been validated using Privateer (all monosaccharides including those involved in the linkage were in low-energy chair conformations) and has a RSCC of greater than 0.80 – indicating good fit to electron density – it can be assumed that these torsion angles are a direct result of external factors. Upon examination of the area surrounding the glycan, it becomes evident that a network of electrostatic interactions could be affecting the conformation of the N-glycan chain. The proximity of the linkage to arginine, histidine and asparagine side-chains may cause the observed deviation. Furthermore, this highlights how linkages further down a glycan tree can also be subject to interactions with protein residues. These interactions may also explain why MAN-MAN linkage torsion angles are less concentrated to one pair of values than the more constrained NAG-NAG linkage.

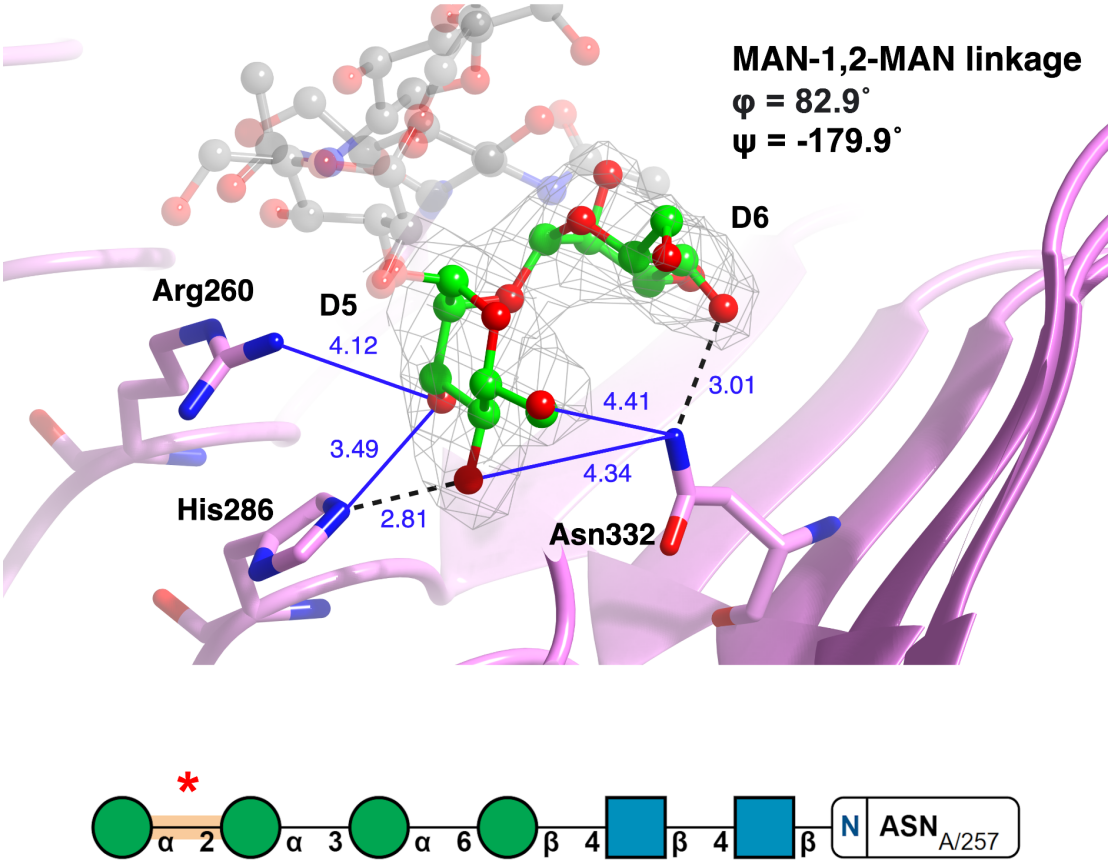


Figure 6

An unusual pair of MAN-1,2-MAN torsions in PDB 4J0M (She et al. 2013). The mannose-mannose pair is well supported by the electron density, indicating that the unusual conformation of the linkage (red asterisk on the bottom panel) may be stabilised by interactions – electrostatic in this case – with surrounding side-chains. The MAN(chain ID and sequence number: D5)-MAN(chain ID and sequence number: D6) linkage of 4J0M (pink) is identified as an outlier (φ = 82.9˚, ψ = -179.9˚). The carbohydrate linkage is represented by a ball-and-stick model (C = green, O = red, N = blue). Residues identified as interacting with the linkage are represented by a cylindrical model (C = pink). Hydrogen bonds (black dashed line) and electrostatic interactions (within 4.5Å in COOT, blue line) with distance between atoms (Å) shown. $2F_o$-F_c electron density (grey) is displayed for the linkage contoured to 1σ. Possible electrostatic interactions were identified in COOT for residues within 4.5Å of the linkage, and can be seen between Arg260/NH1 and MAN5/O3, His286/NE2 and MAN5/O3, Asn332/ND2 and MAN5/O4, and Asn332/ND2 and MAN5/O6. This linkage has a Z score of -1.06. Top - produced using CCP4mg. Bottom - SNFG notation output from Privateer.

3.4.2 High-energy ring conformation anomalies may distort a linkage

Figure 7 shows a glycan stabilised by CH-π interactions with phenylalanine side-chains (PDB code: 5GSQ). While the fit to electron density is reasonable for the first few pyranosides (which show no issues on the validation report), the MAN-1,3-BMA and the terminal MAN residue are highlighted in orange on Privateer's SNFG representation: the link has a Z score of -1.32, indicating a large deviation, and the terminal mannose's ring is in a $^{1}S_{3}$ conformation, which is wholly unexpected for a pyranoside that is part of an N-glycan, and therefore it is marked as worthy of inspection (orange). Examination of the electron density map around the MAN-1,3-BMA pair reveals that the fit to the observed data is poor for the MAN residue – refinement against incomplete density usually results in high-energy ring conformations without the inclusion of torsion restraints (Agirre 2017). The distortion of the ring conformation in pyranosides has been reported to have a knock-on effect on linkages (Agirre et al. 2017a), hence we believe this is the most probable explanation for this outlier.

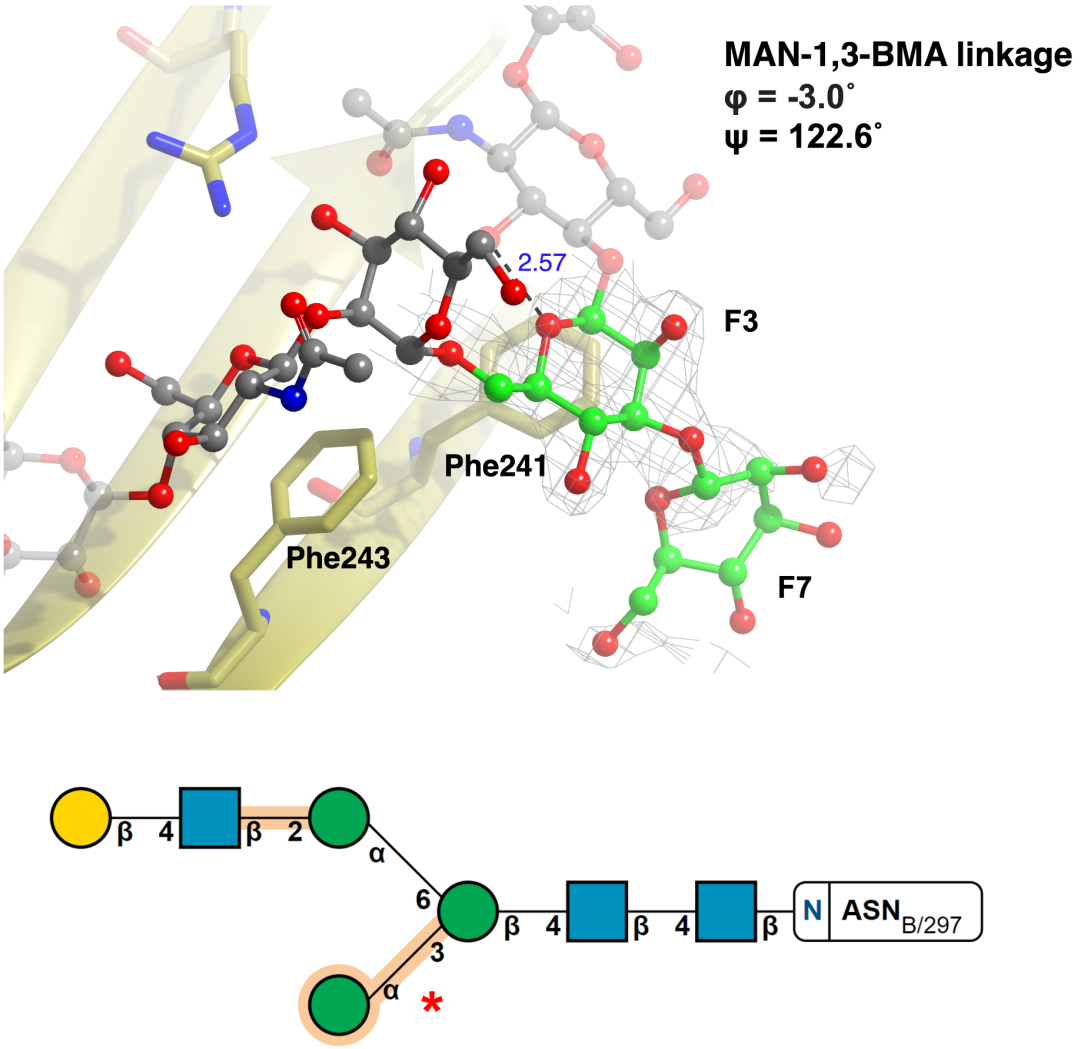


Figure 7

High-energy ring conformations may cause glycosidic link anomalies. The MAN(F7)-BMA(F3) linkage (red asterisk on the bottom panel) of 5GSQ (Chen et al. 2017) (gold) – which was not part of the curated torsions dataset because the MAN residue has poor RSCC – is identified as an outlier (φ = -3.0˚, ψ = 122.6˚). BMA(chain ID and sequence number: F3) and MAN(chain ID and sequence number: F7) are represented by a ball-and-stick model (C = green, O = red), whilst the rest of the attached glycan is represented in faded grey ball and stick. Residues identified as interacting with the linkage are represented in stick form (C = gold, O = red, N = blue). Hydrogen bonds (black dashed line) with distance between atoms (Å) shown. $2F_o$-F_c electron density (grey) is displayed for the linkage contoured to 1σ. Possible CH-π interactions were identified, and can be seen between Phe243 and NAG(F5), and Phe241 and BMA(F3). This linkage has a Z score of -1.32, and presumably got distorted because the terminal mannose, MAN(F7), is in a $^1S_3$

skew-boat (high-energy, please refer to (Agirre, Davies, et al. 2015a) for further reading on conformational anomalies) ring conformation – also highlighted in orange on the figure – due to the absence of well defined electron density. Both linkage and ring conformations are unsupported by the electron density and should be either removed, or corrected before deposition to reflect the most probable, low-energy conformations. Top - produced using CCP4mg. Bottom - SNFG notation output from Privateer.

4. Conclusions

In this study, a large number and range of N-glycan linkage torsion angles were collected from both the PDB and PDB-REDO after being curated using Privateer. The collected data, released and articulated through the Privateer software, will provide a strong foundation for future model building, refinement and validation software. The comparisons between the PDB and PDB-REDO presented here assessed the importance of modern refinement techniques. The difference in the torsion angles between the validated PDB and PDB-REDO datasets are minimal. However, in certain cases, the application of a consistent refinement technique can alleviate errors in the model building process. Furthemore, the absence of torsional restraints in PDB-REDO perhaps allows a more realistic spread of torsional values to be observed. It is also important to note valid rationalisations for linkage torsion angles deviating from the calculated mean. Electrostatic and steric interactions play a large role in protein folding in general and can cause or stabilise the skewed N-glycan linkage torsions exhibited in certain glycoproteins. Therefore, it is highly likely that these electrostatically charged or sterically bulky amino acids play a role in overall N-glycan conformation.

Availability and open research data

All scripts, data and graphics associated with this work are uploaded to Zenodo (10.5281/zenodo.7356467) (Dialpuri et al. 2022). The Privateer source code is available from GitHub (https://github.com/glycojones/privateer.git). Binaries will be released as an update to CCP4 8.0.

Acknowledgements

Jordan Dialpuri is funded by Biotechnology and Biological Sciences Research Council [BBSRC, grant number BB/T0072221]. Haroldas Bagdonas is funded by The Royal Society [grant number RGF/R1/181006]. Mihaela Atanasova is funded by the UK Engineering and Physical Sciences Research Council [EPSRC, grant number EP/R513386/1]. Lucy Schofield's work was funded by The Royal Society through a summer studentship. Robbie Joosten is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 871037 (iNEXT-Discovery) and by CCP4. Jon Agirre is a Royal Society University Research Fellow [award number UF160039]. This work is based on a proof of concept as part of the MChem course of the Department of Chemistry at the University of York. The authors would like to thank Alex Ascham, Sarah Hargan, Eleanor Charnley, Alex Muriel, Charles Kingston, Conor MacDonald, Eve Tipple, Andrew Harvey, Thomas Hartshorn, Alex Bentley, Jordan Wilson, Rachel Napier, Luke Julyan, Catrin Ellis, Will Ashwood and James Jones for their work in the previous study.

Supplementary Information

Linkage	Number in data set	φ (°)	σ φ (°)	ψ (°)	σ ψ (°)
ASN-NAG	6688	-97.5	25.3	178.7	22.1
NAG-1,4-NAG	3800	-80.3	14.5	-127.1	22.8
BMA-1,4-NAG	1659	-86.8	17.9	-133.3	18.2
MAN-1,3-BMA	781	75.8	17.6	121.5	22.6
MAN-1,6-BMA	702	72.4	26.6	167.7	33.3
MAN-1,2-MAN	509	78.0	17.9	132.3	32.1
MAN-1,3-MAN	280	78.5	16.8	132.3	24.2
MAN-1,6-MAN	221	68.5	25.0	-173.3	20.0
FUC-1,6-NAG	406	-73.6	24.2	167.2	41.0
FUC-1,3-NAG	228	-69.7	10.5	139.2	6.8
NAG-1,2-MAN	242	-86.2	15.0	150.7	14.6
GAL-1,4-NAG	84	-67.5	21.8	-122.2	23.7

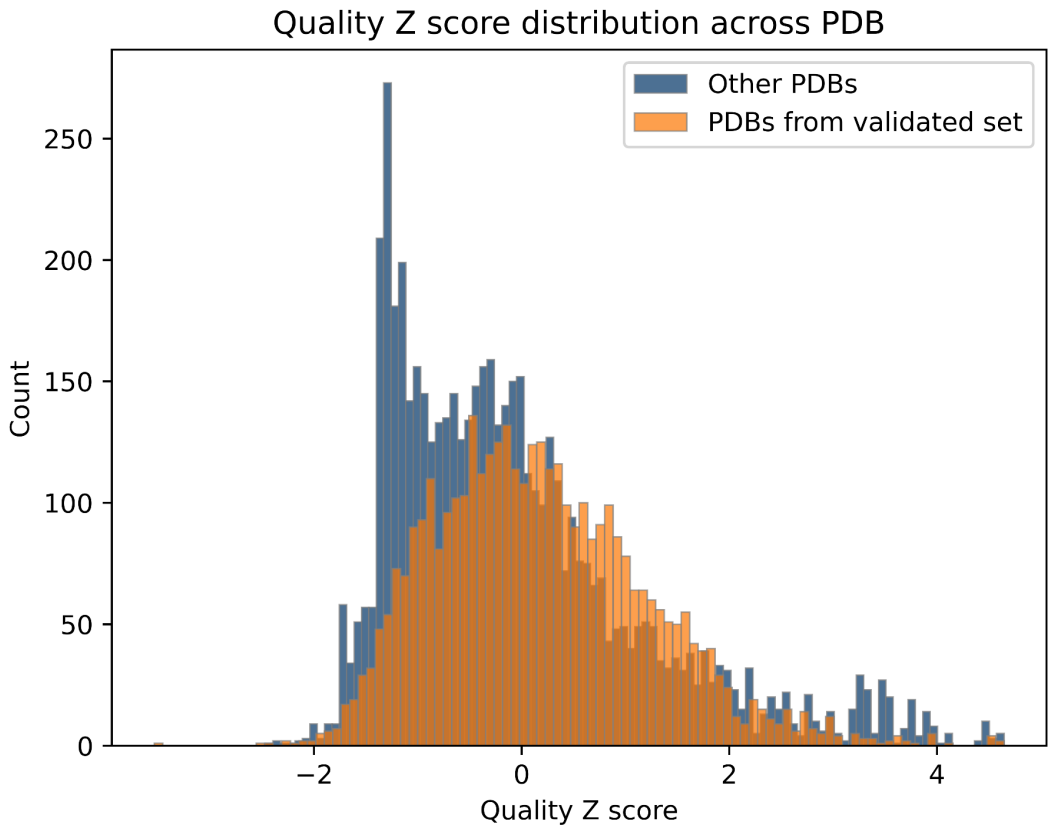
Supplementary Table 1

Raw data from the survey for linkages containing more than 50 data points. The number of data points in each data set is shown as well as the mean and standard deviation for each torsion angle value.

Linkage	Number in data set	ω (°)	σω (°)
NAG-1,6-FUC	405	-58.1	42.9
BMA-1,6-MAN	701	18.4	69.9
MAN-1,6-MAN	221	-44.7	54.7

Supplementary Table 2

Raw data from the survey for 1,6 linkages that have an omega torsion angle.



Supplementary Figure 1

The distribution of quality Z scores across the PDB, the orange bars represent models which had at least one entry (glycan linkage) in our dataset of curated linkages. The blue bars represent the rest of the models in the PDB that contain N-glycans.

Appendix C.

The CCP4 suite: integrative software for macromolecular crystallography

Jon Agirre[1,*], Mihaela Atanasova[1], Haroldas Bagdonas[1], Charles B. Ballard[2,3], Arnaud Baslé[4], James Beilsten-Edmands[5], Rafael J. Borges[6], David G. Brown[7], J. Javier Burgos-Mármol[8], John M. Berrisford[9], Paul S. Bond[1], Iracema Caballero[10], Lucrezia Catapano[11,12], Grzegorz Chojnowski[13], Atlanta G. Cook[14], Kevin D. Cowtan[1], Tristan I. Croll[15,16], Judit É. Debreczeni[17], Nicholas E. Devenish[5], Eleanor J. Dodson[1], Tarik R. Drevon[2,3], Paul Emsley[11], Gwyndaf Evans[5,18], Phil R. Evans[11], Maria Fando[2,3], James Foadi[42], Luis Fuentes-Montero[5], Elspeth F. Garman[16], Markus Gerstel[5], Richard J. Gildea[5], Kaushik Hatti[15], Maarten L. Hekkelman[17], Soon Wen Hoh[1], Michael A. Hough[5,18], Huw T. Jenkins[1], Robbie P. Joosten[17], Ronan M. Keegan[2,3,8], Nicholas Keep[19], Eugene B. Krissinel[2,3], Petr Kolenko[20,21], Oleg Kovalevskiy[2,3], Victor S. Lamzin[13], David M. Lawson[23], Andrey A. Lebedev[2,3], Andrew G.W. Leslie[11], Bernhard Lohkamp[24], Fei Long[11], Martin Malý[20,21,25], Airlie J. McCoy[15], Stuart J. McNicholas[1], Ana Medina[10], Claudia Millán[15], James W. Murray[29], Garib N. Murshudov[11], Robert A. Nicholls[11], Martin E.M. Noble[4], Robert Oeffner[15], Navraj S. Pannu[26], James M. Parkhurst[5,18], Nicholas Pearce[27], Joana Pereira[28], Anastassis Perrakis[17], Harold R. Powell[29], Randy J. Read[15], Daniel J. Rigden[8], William Rochira[1], Massimo Sammito[15,41], Filomeno Sánchez Rodríguez[1,5,8], George M. Sheldrick[30], Kathryn L. Shelley[31], Felix Simkovic[8], Adam J. Simpkin[8], Pavol Skubak[26], Egor Sobolev[32], Roberto A. Steiner[12,33], Kyle Stevenson[2], Ivo Tews[25], Jens M.H. Thomas[8], Andrea Thorn[34], Josep Triviño Valls[10], Ville Uski[2,3], Isabel Usón[10,35], Alexei Vagin, Sameer Velankar[9], Melanie Vollmar[36], Helen Walden[38], David Waterman[2,3], Keith S. Wilson[1], Martyn D. Winn[39], Graeme Winter[5], Marcin Wojdyr[40], Keitaro Yamashita[11].

Correspondence: jon.agirre@york.ac.uk

[1]York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, UK.

[2]STFC, Rutherford Appleton Laboratory, Didcot, OX11 0FA, UK.

[3]CCP4, Research Complex at Harwell, Rutherford Appleton Laboratory, Didcot OX11 0FA, UK.

[4]Biosciences Institute, Newcastle University, Newcastle upon Tyne NE2 4HH, UK.

[5]Diamond Light Source, Harwell Science and Innovation Campus, Didcot OX11 0QX, UK.

[6]Center for Molecular Biology and Genetic Engineering, University of Campinas (UNICAMP), Av. Dr. André Tosello, 550, 13083-886, Campinas, Brazil.
[7]Laboratoires Servier SAS Institut de recherches Croissy-sur-Seine, France.
[8]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK.
[9]Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
[10]Crystallographic Methods, Institute of Molecular Biology of Barcelona (IBMB-CSIC), Barcelona Science Park, Helix Building, Baldiri Reixac 15, 08028 Barcelona, Spain.
[11]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK.
[12]Randall Centre for Cell and Molecular Biophysics, Faculty of Life Sciences and Medicine, King's College London, London SE1 9RT, UK.
[13]European Molecular Biology Laboratory, Hamburg Unit, Notkestrasse 85, 22607 Hamburg, Germany.
[14]The Wellcome Centre for Cell Biology, University of Edinburgh, Michael Swann Building, Max Born Crescent, The King's Buildings, Edinburgh EH9 3BF, UK.
[15]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge CB2 0XY, UK.
[16]Altos Labs, Portway Building, Granta Park, Great Abington, Cambridge CB21 6GP, UK.
[17]Discovery Sciences, R&D BioPharmaceuticals, AstraZeneca, Darwin Building, Cambridge Science Park, Milton Road, Cambridge CB4 0WG, UK.
[18]Rosalind Franklin Institute, Harwell Science and Innovation Campus, Didcot OX11 0QS, UK.
[16]Department of Biochemistry, Dorothy Crowfoot Hodgkin Building, University of Oxford, Oxford OX1 3QU, UK.
[17]Oncode Institute and Department of Biochemistry, Netherlands Cancer Institute, Amsterdam, the Netherlands.
[18]School of life sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK.
[19]Department of Biological Sciences, Institute of Structural and Molecular Biology, Birkbeck College, London WC1E 7HX, UK.
[20]Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Břehová 7, Prague 1, 115 19 Czech Republic.
[21]Institute of Biotechnology of the Czech Academy of Sciences, BIOCEV, Průmyslová 595, Vestec, 252 50 Czech Republic.
[23]Department of Biochemistry and Metabolism, John Innes Centre, Norwich, NR4 7UH UK.

[24]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-171 77 Stockholm, Sweden.
[25]Biological Sciences, Institute for Life Sciences, University of Southampton, Southampton SO17 1BJ, UK.
[26]Department of Infectious Diseases, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands.
[27]Department of Physics, Chemistry and Biology (IFM), Linköping University, SE-581 83 Linköping, Sweden.
[28]Biozentrum and SIB Swiss Institute of Bioinformatics, University of Basel, 4056 Basel, Switzerland.
[29]Department of Life Sciences, Imperial College London, South Kensington Campus, London SW7 2AZ, UK.
[30]Department of Structural Chemistry, Georg-August-Universität Göttingen, Tammannstrasse 4, 37077 Göttingen, Germany.
[31]Institute for Protein Design, University of Washington, Seattle WA 98195, USA.
[32]European Molecular Biology Laboratory, c/o DESY, Notkestr. 85, 22607 Hamburg, Germany.
[33]Department of Biomedical Sciences, University of Padova, Italy.
[34]Institute for Nanostructure and Solid State Physics, Universität Hamburg, 22761 Hamburg, Germany.
[35]ICREA, Institució Catalana de Recerca i Estudis Avançats, Passeig Lluís Companys, 23, Barcelona, E-08003, Spain.
[36]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
[38]School of Molecular Biosciences, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK.
[39]Scientific Computing Department, Science and Technology Facilities Council, Didcot OX11 0FA, UK.
[40]Global Phasing Limited, Sheraton House, Castle Park, Cambridge CB3 0AX, UK.
[41]AstraZeneca, Discovery Centre, Biologics Engineering, Biomedical Campus, 1 Francis Crick Ave, Trumpington, Cambridge CB2 0AA, UK.
[42]University of Bath, Bath, UK.

Abstract

The Collaborative Computational Project 4 (CCP4) is a UK-led international collective with a mission to develop, test, distribute and promote software for macromolecular crystallography. The CCP4 suite is a multiplatform collection of programs brought together by familiar execution routines, a set of common libraries, and graphical interfaces. The CCP4 suite has experienced several considerable changes since its last reference article, involving new infrastructure, original programs and graphical interfaces. This article, which is intended as a general literature citation for the use of the CCP4 software suite in structure determination, will guide the reader through such transformations, offering a general overview of the new features and outlining future developments. As such, it aims to highlight the individual programs that compose the suite, and to provide the latest references to them for perusal by crystallographers around the world.

1. Introduction

As a technique, macromolecular crystallography (MX) relies heavily on computational methods, built on top of a strict set of conventions and common formats. Most conventions follow the lead of the International Union of Crystallography (IUCr), while MX software development is undertaken by both academic and private sector initiatives. Based in the UK, MX software tools find a common distribution and maintenance channel under the umbrella of the Collaborative Computational Project No. 4, best known as CCP4. This consortium was established by the UK Science Research Council in 1979, almost 45 years ago, to facilitate the coordination and collaboration of MX software developers (Agirre and Dodson 2018). Aside from coordinating and distributing software, CCP4 has a mission of promoting the teaching of MX, with an annual didactic CCP4 Study Weekend in January, and numerous annual workshops, both online and in person, worldwide. Forums, which originally took the shape of email lists – the CCP4 bulletin board (or CCP4bb) for general users' questions, and ccp4-dev for developer discussions – are an evolving aspect of the CCP4 community, with social media taking a more prominent role in hosting other kinds of exchanges, e.g. paper or event announcements (Twitter: @ccp4_mx) or parallel discussions at conferences (Slack channels). The CCP4 website (https://www.ccp4.ac.uk) is the primary mechanism for reference and asynchronous communication but, most importantly, provides a central distribution point for software downloads. A minimal installer package can be obtained from the site, and this will proceed to install the latest version of the suite. Updates are then distributed via a non-disruptive mechanism that was first introduced with CCP4 6.3.0 in

2012. Update reminders are generated automatically, although the update mechanism itself is, by design, initiated manually. As an indication of update frequency, the 7.0 series – originally released in 2016 – saw more than 70 updates until the 7.1 series was released in 2020. Updates are not a one-way road: they may be rolled back if problems are encountered. Whilst every effort has been made to keep the suite streamlined and maintainable, the inclusion of big databases and toolkits has driven space requirements steadily upwards (Figure 1).

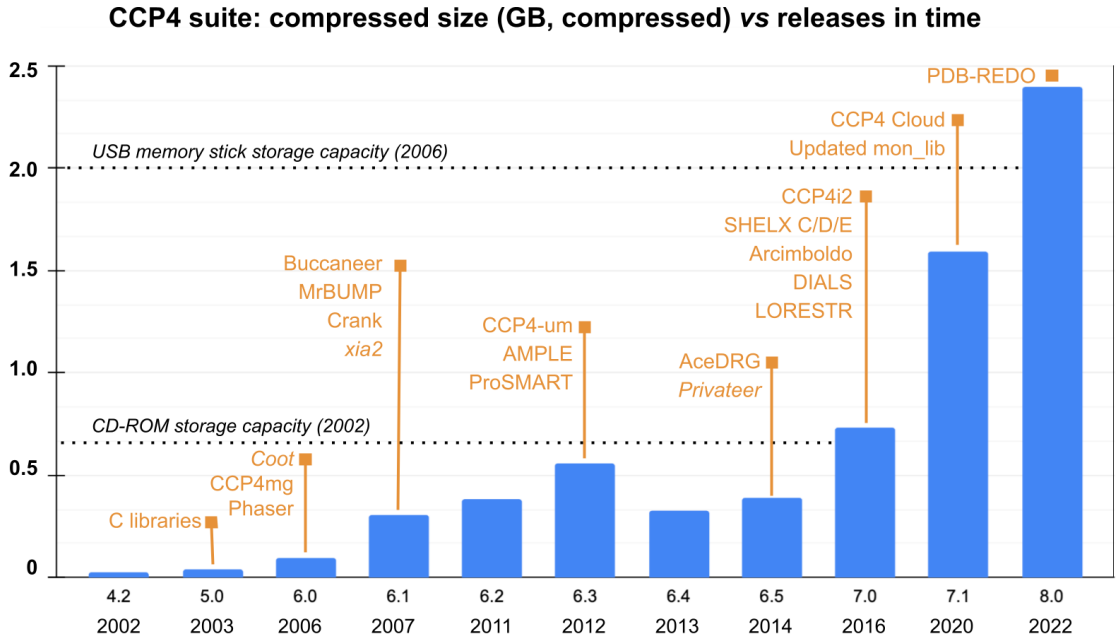


Figure 1. Evolution in size of the CCP4 suite since version 4.2 (2002) through to version 8.0 (2022). Some representative programs included in the releases are highlighted in orange. The update mechanism (CCP4-um) was first used in version 6.3. New graphical interfaces were introduced with 7.0 (CCP4i2) and 7.1 (CCP4 Cloud). Coot and CCP4mg were originally distributed separately, but were bundled in with the suite from CCP4 version 6.5. For reference, the size of two popular contemporary storage devices is shown as dotted lines; please note that these were never targeted as distribution media.

The last decade has seen some big transformations in the field of MX: new workflows have been created (*e.g.* phasing with AlphaFold2 models), some old workflows have been optimised, while some others are on the verge of disappearing; this has often been the result of cross-pollination with other techniques in structural biology – *e.g.* electron cryo-microscopy (Cryo-EM) particularly, through a synergistic collaboration with CCP-EM

(Burnley, Palmer, and Winn 2017), the Collaborative Computational Project for Cryo-EM, which repurposes some CCP4 code for the Cryo-EM community. For example, owing to the deep-learning revolution in computational structure prediction (Jumper et al. 2021), it is now possible to phase most structures using large predicted fragments, or, owing to the accuracy of the method, even to rigid-body-fit an initial predicted model into electron density (Oeffner et al. 2022; McCoy, Sammito, and Read 2022; Medina et al. 2022). As a side effect of the creation of these new workflows, experimental phasing is now losing weight in the everyday activities of an MX laboratory, with derivatives only being created as a last resort after all the now-conventional methods have failed. Data acquisition and processing, greatly bolstered by both software and hardware developments led *in situ* at synchrotrons, is now done almost instantaneously after collection, presenting the user with the results from applying different processing strategies. Though seemingly unconnected, most of these newer developments have one thing in common: the Python programming language as a platform for pipelining and program communication.

While some Python scripts were already part of the CCP4 suite even before the time of the last general publication (Winn et al. 2011b), most of the recent source code committed to the CCP4 repositories involves Python in one way or another – for example, both data integration tools DIALS (Graeme Winter et al. 2018) and its graphical user interface, DUI (https://github.com/ccp4/DUI), are Python-heavy software; other CCP4 programs, encoded in a different language such as C++ for performance reasons, may also offer Python bindings – examples include *Coot* (P. Emsley et al. 2010), *Privateer* (Agirre, Iglesias-Fernández, et al. 2015a), or GEMMI (Wojdyr 2022), which is a crystallographic toolkit developed in collaboration with Global Phasing Ltd. Both the Python language and its interpreter are now at the core of the CCP4 suite. Importantly, both new graphical user interfaces to the CCP4 suite (*vide infra*) make substantial use of the Python language.

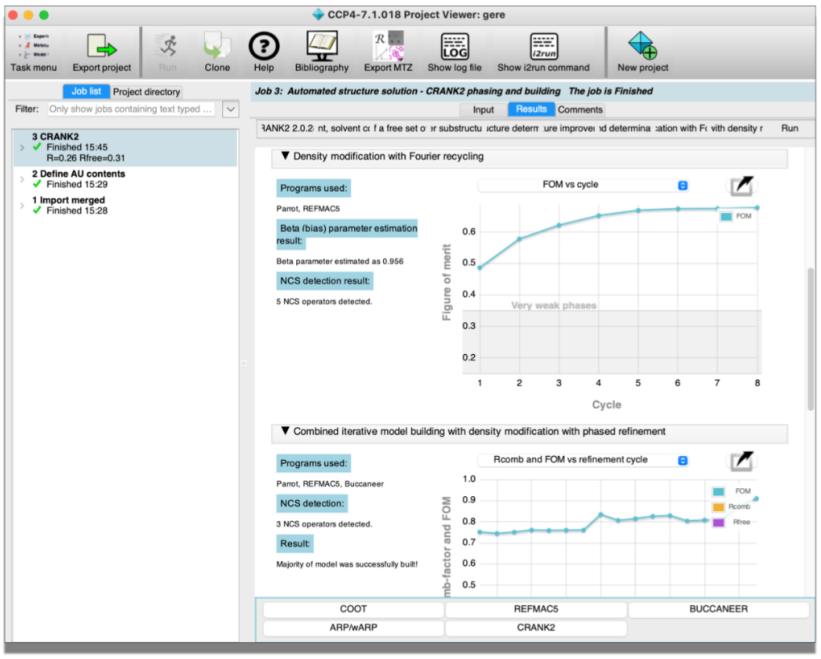
On the subject of graphical user interfaces, a large paradigm shift is underway too, with both CCP4i2 and CCP4 Cloud making extensive use of web technologies: HTML, CSS and Javascript are used for both interface design and result presentation, with CCP4 Cloud making a strong case for the transformation of existing interactive model building and illustration applications – *e.g.* Coot and CCP4mg – into apps that can be run within a web browser.

2. Overview of the newest developments

2.1. Graphical user interfaces

The long-serving CCP4i interface (developed in Tcl/Tk) has recently been deprecated and replaced by a more modern, QT/PySide graphical user interface (GUI), named CCP4i2 (Potterton et al. 2018b). The CCP4i2 GUI, whose main purpose is to provide a desktop-based experience, has introduced a number of architectural differences with respect to the first iteration: i) a real database system – as opposed to a directory structure – provides traceability of files and jobs, and allows for the automatic population of inputs on follow-on jobs with outputs from previous jobs; ii) large MTZ files are separated into important column sets defining particular data types and with predictable names, *e.g.* Miller indices (H, K and L columns) plus amplitudes and estimated standard deviations – or e.s.d.s – (F and SIGF columns) define an '*Amplitudes'* data type; iii) individual programs are wrapped in Python for their incorporation into tasks, which in many cases will be pipelines themselves – e.g. 'Data reduction' is a pipeline that involves use of the programs POINTLESS, AIMLESS, CTRUNCATE and FREER; iv) communication of results between individual programs is consolidated in structured data (XML) files. In addition, task reports aim to present only fundamental results and, where possible, provide expert diagnostics in a natural human-readable language – *e.g.* 'No evidence of possible translational non-crystallographic symmetry'. Other utilities include a multiplatform project import and export mechanism, instant job search by keywords, the use of task-specific key performance indicators – *e.g.* R-work/R-free – and context-dependent follow-on jobs with automatic selection of input files and default options. Outside of the graphical user interface but very much within its infrastructure, the *i2run* module provides a command line mechanism for running CCP4i2 pipelines, opening the door to batch processing using interface-level decision making.

(a)


CCP4-7.1.018 Project Viewer: gere
Task menu
Export project
Clone
Help
Bibliography
Export MTZ
Show log file
Show i2run command
New project
Job list
Project directory
Filter:
Only show jobs containing text typed ...
3 CRANK2
Finished 15:45
R=0.26 Rfree=0.31
2 Define AU contents
Finished 15:29
1 Import merged
Finished 15:28
Job 3: Automated structure solution - CRANK2 phasing and building The job is Finished
Input
Results
Comments
Run
Density modification with Fourier recycling
Programs used:
Parrot, REFMAC5
Beta (bias) parameter estimation result:
Beta parameter estimated as 0.956
NCS detection result:
5 NCS operators detected.
FOM vs cycle
FOM
Figure of merit
Very weak phases
Cycle
Combined iterative model building with density modification with phased refinement
Programs used:
Parrot, REFMAC5, Buccaneer
NCS detection:
3 NCS operators detected.
Result:
Majority of model was successfully built!
Rcomb and FOM vs refinement cycle
FOM
Rcomb
Rfree
Rcomb-factor and FOM
COOT
REFMAC5
BUCCANEER
ARP/wARP
CRANK2

(b)

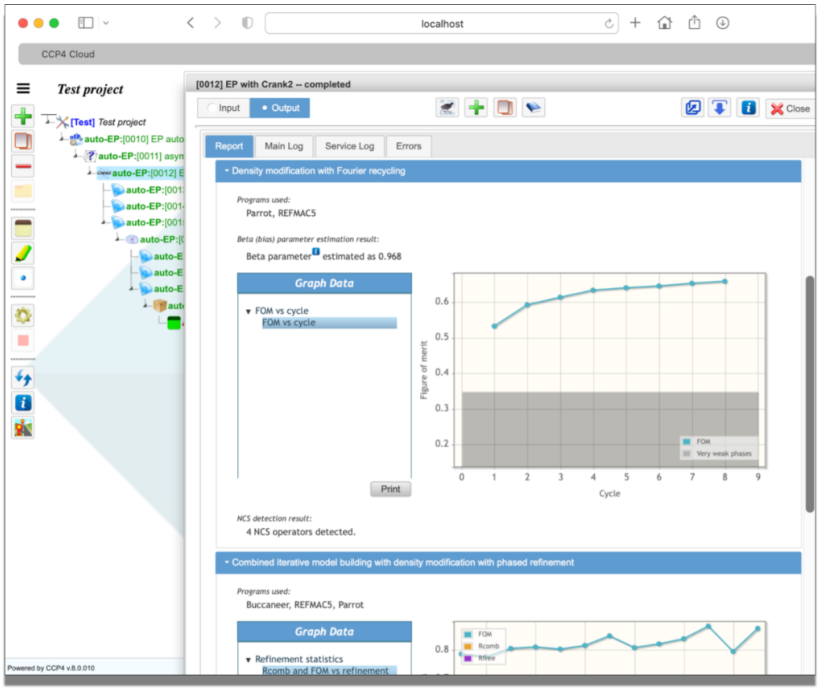

localhost
CCP4 Cloud
Test project
[Test] Test project
auto-EP:[0010] EP auto
auto-EP:[0011] asyn
auto-EP:[0012] E
[0012] EP with Crank2 -- completed
Input
Output
Close
Report
Main Log
Service Log
Errors
Density modification with Fourier recycling
Programs used:
Parrot, REFMAC5
Beta (bias) parameter estimation result:
Beta parameter estimated as 0.968
Graph Data
FOM vs cycle
FOM vs cycle
Print
Figure of merit
Very weak phases
Cycle
NCS detection result:
4 NCS operators detected.
Combined iterative model building with density modification with phased refinement
Programs used:
Buccaneer, REFMAC5, Parrot
Graph Data
Refinement statistics
Rcomb and FOM vs refinement
Powered by CCP4 v.8.0.010

Figure 2. Comparison between the new CCP4 graphical user interface offerings: desktop (CCP4i2, panel a) and online (CCP4 Cloud, panel b). The same pipeline (Crank-2) has been run on both interfaces. The reports show equivalent graphs due to the use of a compatibility layer that allows for the same report code to run on both platforms.

CCP4 Cloud (Eugene Krissinel et al. 2022) is a complete reimagination of what an interface should look like in the context of macromolecular crystallography. Technology-wise, it provides a server-side Javascript implementation (based on node.JS) designed to work with High Performance Computing (HPC) facilities – clusters and generic clouds – but which can also be run on a user's PC. This implementation also enables secure web access by a browser via HTML5, CSS and Javascript (jQuery), and allows CCP4 Cloud to look consistent across different browsers and platforms, making it possible to run jobs and manage projects from, for example, mobile devices. The interface provides a general file import function, which allows it to decide what kind of jobs can be run: for example, automated model building can only be done if at least reflections and a sequence have been imported. The system features task interfaces for many CCP4 programs, and some newly introduced pipelines. One such example is CCP4build, which combines Parrot for density modification (Kevin Cowtan 2010a), Buccaneer for model building (Kevin Cowtan 2006), Refmac for refinement (Murshudov et al. 2011), *Coot* for model editing (P. Emsley et al. 2010), and EDSTATS (Tickle 2012) for model accuracy analysis; with these tools, CCP4build is able to make expert decisions depending on the phasing approach and model completeness. High level progress indicators are available on both CCP4 Cloud and CCP4i2; one such example is the 'verdict' functionality, which provides a score for model completion and fit to the experimental data. CCP4i2 and CCP4 Cloud have a conceptually similar set of tasks, albeit their graphical presentation differs.

2.2. Data processing

Developed in collaboration with scientists at the Lawrence Berkeley National Laboratory, the DIALS project (Graeme Winter et al. 2018) is the CCP4 suite's main diffraction image processing *toolkit* – modular and hackable by design, so experienced crystallographers can tweak, extend, or add new algorithms. Regardless of this specialist component-based approach, complete DIALS workflows are provided in the *xia2* pipeline (G. Winter 2010), which incorporates expert decision making (Graeme Winter, Lobley, and Prince 2013). More recently, a graphical user interface (DIALS User Interface, or DUI) has been introduced as

well (Fuentes-Montero et al. 2016). The *xia2* pipeline is run automatically at the end of data collections at Diamond Light Source (Oxfordshire, UK), providing the results of applying multiple data processing strategies: users are expected to look at the metrics provided, and decide which is better suited to their diffraction dataset. Newcomer users wanting to learn more about DIALS are advised to use DUI, which provides a guided step-by-step execution of the whole process, though command line use through simple scripts is designed to be accessible to the non-expert user.

DIALS is able to natively process data obtained at X-ray Free-Electron Laser (XFEL) facilities (Ginn et al. 2015; Uervirojnangkoorn et al. 2015), and supports multi-crystal scaling (Beilsten-Edmands et al. 2020) and analysis via xia2.multiplex (Gildea et al. 2022), serial crystallography (Brewster et al. 2018; James Michael Parkhurst 2020), and electron diffraction such as that obtained with standard Field Emission GUN (FEG) cryo-microscopes (Clabbers et al. 2018). Data from multiple crystals may be scaled and merged together with BLEND (Mylona et al. 2017). Ice rings and further pathologies in measured data can be identified by a separate standalone tool named AUSPEX, which provides visual and automatic diagnostics based on statistics (Thorn et al. 2017) and, more recently, machine learning (Nolte et al. 2022). Alternatively, the iMosflm software (Powell et al. 2017) provides an easy to use interface to the MOSFLM image processing program; while the software is no longer actively developed, it contains many useful features and remains popular with users.

Once the data are processed, Laue group determination, and data scaling and reduction can be done with DIALS directly, although POINTLESS and AIMLESS are also offered as a fallback mechanism (Evans and Murshudov 2013) – indeed, these programs form the basis for the CCP4i2 'data reduction' task. Further diagnostics can be obtained by running CTRUNCATE – originally an implementation of French and Wilson's algorithm (French and Wilson 1978) to obtain structure factor amplitudes from intensities – which will scan datasets for signs of anisotropic diffraction, twinning, and translational non-crystallographic symmetry (tNCS) among other pathologies, in order to identify critical issues that could complicate or even compromise the downstream structure determination process. This set of programs have graphical interfaces on both CCP4i2 and CCP4 Cloud, producing colour-coded reports that flag up potential problems. Importantly, detailed reports are generated whenever merged intensities or amplitudes are imported into the graphical interfaces, providing a sanity check and metadata tracking.

2.3. Phasing

The CCP4 suite provides software for all phasing methods, although they mainly fall within one of the following categories: molecular replacement (MR), *ab initio* phasing with ideal fragments (a special case of molecular replacement), and experimental phasing. In the coming years, and due to the recent improvement in protein structure prediction methods, the line between the former two is expected to become blurred or even disappear.

2.3.1. Molecular replacement and *ab initio* phasing, including bioinformatics

While the ever-growing area of bioinformatics is outside the remit of CCP4, the search for suitable molecular replacement templates is primarily driven by protein homology analysis and therefore exploits bioinformatics methods. Various third-party tools have been incorporated into the suite to give support to CCP4's model preparation tools and automated structure solution pipelines. MrBUMP is an automated tool that will perform searches for templates and attempt molecular replacement with them, displaying comprehensive results that can be taken forward provided that the R-factors are low enough. It can find structures of homologues using PHMMER (Eddy 2011) or HHPRED (Söding 2005), and place them using either Phaser (McCoy et al. 2007) or MOLREP (Alexei Vagin and Teplyakov 2010). The template search code of MrBUMP can also be harnessed interactively in CCP4mg, allowing users to create composite models and ensembles for subsequent MR searches; this tool can be accessed from both CCP4i2 and CCP4 Cloud. MR-Parse (A. J. Simpkin, Thomas, et al. 2022) provides a convenient visualisation of potential search models from the PDB and databases of new generation models such as the AlphaFold Protein Structure Database (Varadi et al. 2022). Designed to slice predicted models as well as homologs into domains that may differ in relative orientation from the crystal structure, Slice'N'Dice (A. J. Simpkin, Elliott, et al. 2022) is an automated molecular replacement pipeline that facilitates the placement of these domains in molecular replacement. By processing and slicing the models, it simplifies the task of placing these domains. CCP4mg (S. McNicholas et al. 2011) can also be used to visualise the slicing of the input models.

The CCP4 has a number of efficient molecular replacement packages: AMoRe (Trapani and Navaza 2008), MOLREP (Alexei Vagin and Teplyakov 2010) and Phaser (McCoy et al. 2007) all have different strengths, although only the latter is under active development.

Phaser uses a maximum likelihood approach to the phasing problem; it is the only molecular replacement software that uses intensities natively – *i.e.,* without turning them into amplitudes first – and can also use SAD data (for SAD and MR-SAD phasing). The *Voyager* (Sammito et al. 2019) automated procedure within Phaser presents a new architecture that allows for more flexibility, guiding user decisions in creating ensembles. It also provides, alongside a plethora of new and reimplemented algorithms, code to make best use of AlphaFold (Jumper et al. 2021) and RoseTTAFold (Baek et al. 2021) structure predictions – or high-confidence subsets of them – including transformation of model confidence metrics (*e.g.,* AlphaFold's pLDDT) into estimated B-factors. Owing to the flexibility of the new design, tools for fitting models into Cryo-EM maps have been included. An *ad hoc* graphical user interface is under development; this will allow for easier navigation of the different solutions calculated along the search strategy, presenting the user essential plots such as the self-rotation function.

CCP4 also has fragment-based *ab initio* phasing packages: Arcimboldo (Rodríguez et al. 2009) and Fragon (Jenkins 2018), which use ideal fragments of proteins (mainly helices) in targeted molecular replacement searches. The use of these programs was initially confined to high resolution data, but they have recently enjoyed success at resolutions lower than 2.3 Å – a threshold beyond which it becomes difficult to ascertain the direction of helical fragments – owing to their improved search strategies (Medina et al. 2022), phase combination (Millán et al. 2020) and use of available structural information, including AlphaFold predictions. Arcimboldo (Rodríguez et al. 2009) can use fragments of homologous models and phase previously intractable coiled coil structures (Caballero et al. 2018). It should be noted that part of the success of these methods is down to Phaser's ability to place single amino acids or even atoms with great accuracy (McCoy et al. 2017). Also in alternative MR territory is AMPLE (Bibby et al. 2012) which majors on editing search model ensembles, particularly *ab initio* predictions and distant homologues.

SIMBAD (A. Simpkin et al. 2018; A. J. Simpkin et al. 2020) provides a sequence-independent phasing pipeline that may be used for phasing crystals of unknown contaminants (A. Simpkin et al. 2018). Other MR pipelines use larger fragments or domains as their source of phasing information: BALBES (Long et al. 2008) and MoRDA (Alexey Vagin and Lebedev 2015) are automated pipelines that, using MOLREP, place matches from curated databases containing fragments, domains, and homo- and hetero-oligomers. *Dimple* (Wojdyr et al. 2013) is an automated procedure aimed at quickly arriving at a solved structure of a protein-ligand complex starting from an isomorphous crystal; the software will phase the data and produce preliminary maps including a difference density map where omit density for a ligand might be found.

2.3.2. Experimental phasing

The steady increase of unique new domains deposited every year in the PDB, the availability of millions of models in the AlphaFold Protein Structure Database (Varadi et al. 2022), and the continuous improvement of fragment-based *ab initio* phasing methods mean that experimental phasing is increasingly becoming a last-resort approach to recovering phases – it also means that software will have to deal with the most difficult cases. New since the time of the last CCP4 general publication (2011) are the inclusion of the SHELX C/D/E (Sheldrick 2008a) programs, which can be run individually or in a pipeline through the Crank-2 (Skubák and Pannu 2013) frontend, available in both CCP4i2 and CCP4 Cloud interfaces. Crank-2 itself incorporates a number of different algorithms that can deal with SAD, SIRAS, MAD and MR-SAD. As stated in the previous section, the Phaser software (McCoy et al. 2007) is able to do both SAD and MR-SAD phasing as well.

2.4. Model building and refinement

2.4.1. Interactive model building

The CCP4 suite ships with the *de facto* industry standard interactive model building program: *Coot* (P. Emsley et al. 2010). After two decades under constant development, the *Coot* software package has now reached version 1.0, which incorporates a major rework of the program's graphical architecture, interface, tools and components. Aside from all the well-known tools for manual model building, the software has: a built-in ligand building tool Lidia, which can use AceDRG (*vide infra*) for restraint generation; the ability to create covalent linkages between protein and ligand or between molecular components (Nicholls, Joosten, et al. 2021b); a semi-automatic N-glycan building tool, which is able to build entire oligosaccharides that are consistent with the most common biosynthetic pathways (Paul Emsley and Crispin 2018); a real-space, accelerated refinement tool that is able to process whole macromolecules, in contrast with the manual localised real-space refinement that users typically perform when fitting or tweaking parts of a model (Casañal, Lohkamp, and Emsley 2020); validation tools that run the most common checks on protein models (Ramachandran plots, rotamer propensities, planarity of the peptide bond, per-residue B-factors and density fit analysis, amongst others), plus tools to facilitate ligand fitting (Nicholls 2017) and validation (Paul Emsley 2017) – *e.g.* deviation from ideal geometry values in dictionaries, clashes, interaction maps. *Coot* makes use of the CCP4 monomer library to obtain restraints for the most common biomolecule monomers (amino acids,

carbohydrates, nucleic acids) and most ligands defined in the PDB's Chemical Component Dictionary (Westbrook et al. 2015).

At present, *Coot* is tied to desktop machines due to its reliance on the GTK toolkit (P. Emsley et al. 2010). This means that users of CCP4 Cloud (Eugene Krissinel et al. 2022) need to have a local installation of the CCP4 suite in order to do manual model building. However, there is an ongoing effort to produce a web-based interface, which will use the *Coot* engine in the same manner that the GTK one does but without requiring a local CCP4 installation.

2.4.2. Automated model building

While *Coot* has incrementally added a wealth of automatic procedures over the years, the CCP4 suite includes several fully automated pipelines that combine automated model building software – BUCCANEER (Kevin Cowtan 2006) and NAUTILUS (Kevin Cowtan 2014), ARP/wARP 8.0 (Lamzin, Perrakis, and Wilson 2012a) or the chain tracing code in SHELXE (Usón and Sheldrick 2018) – with reciprocal space refinement (see next section for more details) and validation (EDSTATS (Tickle 2012), MolProbity (Williams et al. 2018)) to produce protein and nucleic acid models that are completed iteratively. These pipelines – *e.g. Modelcraft* (Bond and Cowtan 2022) in CCP4i2, and CCP4build in CCP4 Cloud – are available from both modern graphical user interfaces (CCP4i2 and CCP4 Cloud) and are completed by either graphical or textual summaries about the completeness of the built model. Outside of the protein realm, AlphaFold (Jumper et al. 2021) and RoseTTAfold (Baek et al. 2021) models can be glycosylated using the glycan library and tools in the *Privateer* software (Bagdonas et al. 2021). *PanDDA (Pearce et al. 2017)* allows users to increase the signal-to-noise ratio of their ligand maps by combining several datasets of ligand-free and ligand-bound forms of the protein; the program has algorithms for combining different crystal forms. The suite's current automated model building offerings are completed with ARP/wARP 8.0 (Lamzin, Perrakis, and Wilson 2012a); this software pioneered the iterative combination of model building and refinement (Perrakis, Morris, and Lamzin 1999) – a feature now present in all modern model building pipelines – and automated addition of ligands (Langer et al. 2008). Modern versions of ARP/wARP may also be used with Cryo-EM data (Chojnowski et al. 2021).

At a higher level, the PDB-REDO pipeline has been integrated into CCP4 through graphical interfaces in CCP4i2 and CCP4 Cloud, and with API calls to the PDB-REDO web server (Joosten et al. 2014).

2.4.3. Restraint dictionaries: the CCP4 monomer library

The dictionaries in the CCP4 monomer library (A. A. Vagin et al. 2004a) have been improved with the introduction of AceDRG (Long et al. 2017a), which since version 7.0 of the suite can also generate restraint dictionaries for covalent linkages (Nicholls, Wojdyr, et al. 2021; Nicholls, Joosten, et al. 2021b). New dictionaries are now routinely generated for many compounds, although pyranose sugars have received separate treatment to account for their conformational preferences (Atanasova et al. 2022a; Joosten, Nicholls, and Agirre 2022). Hydrogen atoms have been modelled and restrained in their nuclear positions in the CCP4 Monomer Library (Catapano, Steiner, and Murshudov 2021), as informed by neutron diffraction data (Allen and Bruno 2010).

2.4.4 Refinement

The main tool for full-model reciprocal-space refinement in CCP4 is REFMAC5 (Murshudov et al. 2011). The program uses the sparse-matrix approximation of the Fisher's information matrix (Steiner, Lebedev, and Murshudov 2003) and is designed to be fast and flexible, with a number of refinement methods built into the engine, including restrained, unrestrained, and rigid body refinement. Jelly body restraints are particularly useful for stabilising refinement *e.g.* after molecular replacement, where larger parts of a structure might need to move into place. In addition to controlling model parameterisation and performing macromolecular refinement, REFMAC5 also performs map calculation. A variety of types of weighted maps are produced, which allow visualisation, subsequent analyses and validation.

REFMAC5 allows the addition of case-specific structural knowledge to be utilised during refinement through the external restraints mechanism (Nicholls, Long, and Murshudov 2012; Kovalevskiy et al. 2018). These external restraints, most useful when only low-resolution data are available, can for instance be generated by ProSMART (Nicholls et al. 2014) for proteins and nucleic acids using homologues or backbone hydrogen bonding patterns, LibG (A. Brown et al. 2015) for nucleic acid base-pairing and stacking, and Platonyzer (Touw et al. 2016) for zinc, sodium and magnesium sites. The automated pipeline LORESTR (Kovalevskiy, Nicholls, and Murshudov 2016) can be used to optimise the refinement protocol at low resolution, expediting the process and easing manual user effort. New developments and the next generation of structure refinement tools are being implemented in Servalcat utilising the GEMMI library (Yamashita et al. 2021).

The PAIREF program (Malý et al. 2020), recently introduced into CCP4i2, performs automatic paired refinement (Karplus and Diederichs 2012) using the REFMAC5 refinement engine. It analyses the impact of weak reflections beyond the traditional high-resolution diffraction limit cut-off on the quality of the refined model. The program monitors model and data indicators, model-to-data agreement metrics, and implements a decision-suggesting routine for the high-resolution cutoff that may result in the best model.

Outside of REFMAC5 and associated tools, the SHEETBEND software (K. Cowtan, Metcalfe, and Bond 2020) allows for a very fast preliminary refinement of the atomic coordinates and, optionally, isotropic or anisotropic B-factors (Kevin Cowtan and Agirre 2018). It is based on a novel approach, in which a shift field – and not atoms – is refined to update and morph models. This approach is particularly indicated to correct large shifts in secondary structure elements after molecular replacement, and is run by default as part of the *Modelcraft* pipeline (Bond and Cowtan 2022).

2.5. Validation, deposition, analysis and representation

Both CCP4i2 and CCP4 Cloud interfaces include a validation and deposition interface developed in collaboration with the PDBe, the Protein Data Bank in Europe (wwPDB consortium 2019; Armstrong et al. 2020). The purpose of this tool is to prepare mmCIF files for deposition; additionally, it provide the convenience of letting users see what their preliminary wwPDB validation report (Gore, Velankar, and Kleywegt 2012; Gore et al. 2017) would look like, and allow them to fix errors and notice interesting chemical features of a model before going through the actual deposition process. Also, in preparation for deposition, model and structure factors are converted into mmCIF, which in turn allows the wwPDB to pre-populate many of the required metadata for deposition, *e.g.* refinement statistics.

Further validation tools exist in CCP4 outside of this online validation process. Protein model validation can be done with a variety of tools: MolProbity analyses backbone geometry, rotamers and clashes, and produces a script file that will generate a menu within Coot containing lists of outliers; Coot itself contains a plethora of interactive and live-updated validation tools, ranging from MolProbity-equivalent metrics to other less frequently quoted ones – *e.g.* the Kleywegt Plot – but which can be of great value depending on the problem. The EDSTATS software (Tickle 2012) provides a unique analysis of model-to-data fit, separating results by main-chain and side-chain and looking at difference density, with the results being able to point at common modelling problems, such as poorly fitting regions

requiring a peptide flip. The 8.0 version of CCP4 has seen the gradual inclusion of PDB-REDO (Joosten et al. 2012) functionality into CCP4's interfaces; for example Tortoize (Sobolev et al. 2020), a tool that analyses main-chain and side-chain geometry and reports Z scores for every amino acid, is now integrated into CCP4's validation tasks. The visual output of PDB-REDO calculations is displayed consistently on CCP4i2, CCP4-cloud and the PDB-REDO website by encapsulating various interactive plots and tables in a self-contained, single web component. The findMySequence software (Chojnowski et al. 2022) uses machine learning for the identification of unknown proteins in X-ray crystallography and cryo-EM data, with the added benefit of detecting elusive register errors, which may have a detrimental effect on the quality of the rest of the structure. The Iris validation framework (Rochira and Agirre 2021) is a standalone tool that displays a variety of validation metrics as concentric circles, modelling errors becoming visible as ripples on successive circles. PISA allows for the analysis of molecular interfaces, calculating likely assemblies (Evgeny Krissinel and Henrick 2007). Carbohydrate model validation, including protein glycosylation, can be carried out with the Privateer software (Agirre, Iglesias-Fernández, et al. 2015a), which in the MKIV version incorporates checks of glycan composition against offline mirrors of several glycomics databases (Bagdonas, Ungar, and Agirre 2020a). Specific structural radiation damage sites in structures derived from cryo-cooled crystals can be identified with RABDAM through the B_{Damage} (Shelley et al. 2018) and the B_{Net} (Shelley and Garman 2022) metrics, and space-group and origin ambiguity may be determined and resolved by using Zanuda (Lebedev and Isupov 2014).

On the representation side, CCP4's main tool is the CCP4 Molecular Graphics Project (CCP4mg). Since the last CCP4mg general publication (S. McNicholas et al. 2011), the main updates have involved new functionalities for handling Cryo-EM maps, 3D representation of N-glycans (Stuart McNicholas and Agirre 2017), and the addition of a new interactive interface to the functionality of Mr BUMP (Keegan et al. 2018). Some of CCP4mg's newer representations can be seen in Fig. 3.

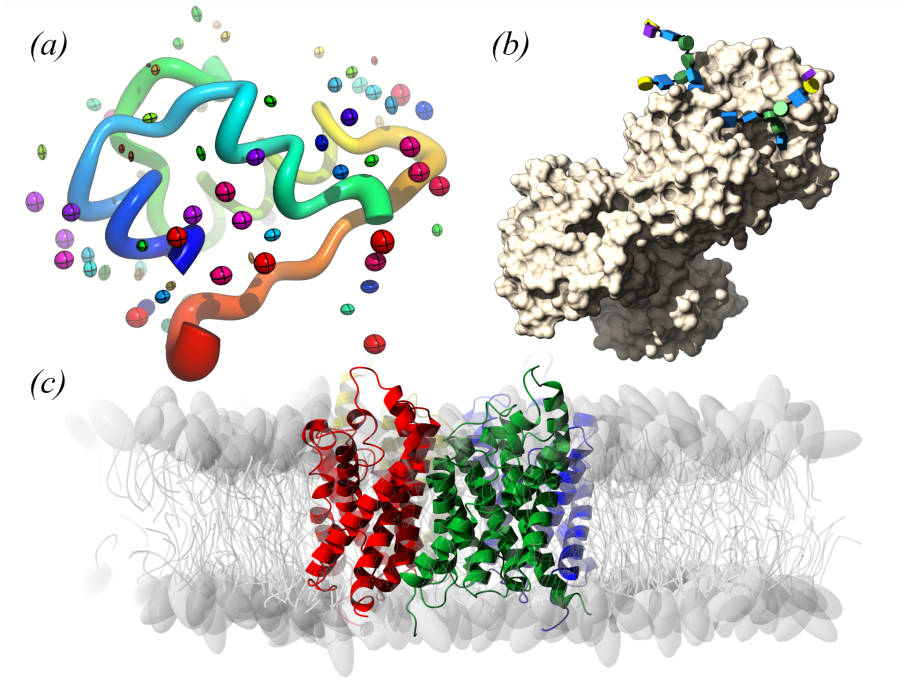


Figure 3. A collection of newer representations included in the CCP4 Molecular Graphics project (CCP4mg). (a) PDB 2BN3 is a high resolution model of insulin (Nanao, Sheldrick, and Ravelli 2005); it is shown here as *worms*, with water molecules drawn as ellipsoids, both coloured and scaled by the model's anisotropic B-factors; (b) PDB 3V8X (Noinaj et al. 2012) has a structure of human transferrin (chain B), drawn here as a solvent-accessible surface with N-glycans shown as Glycoblocks (Stuart McNicholas and Agirre 2017); (c) PDB 3C02, a structure of aquaglyceroporin from *Plasmodium falciparum* (Newby et al. 2008), embedded in a lipid bilayer by CHARMM-GUI (Jo et al. 2008); lipids are shown as *cartoons.*

2.6. Under the bonnet

The *dxtbx* toolkit for DIALS (James M. Parkhurst et al. 2014) is included as part of the *cctbx (Grosse-Kunstleve et al. 2002)* distribution; the clipper-python module (Stuart McNicholas et al. 2018), a SWIG wrapper around the original C++ Clipper library, is also included and supports a number of functions of the CCP4i2 interface, including the *Iris* validation framework (Rochira and Agirre 2021). At a higher level, CCP4i2 (Potterton et al. 2018b) provides code reusability via the command-line, offering a mechanism for executing Python-only pipelines without a running instance of the graphical user interface (*headless* mode). CCP4 Cloud projects and automatic structure solution workflows can also be initiated from the command line, using the “cloudrun” utility – this is useful for performing serial computations for selected targets. The *Coot* model building software (Paul Emsley and Cowtan 2004), originally conceived as a C++ object-oriented toolkit, is now exposed as an

importable Python module to allow for code reuse in new applications, and is also able to run in *headless* mode, suppressing all graphical output. Finally, CCP4mg (S. McNicholas et al. 2011) is also able to run without graphics, generating images from a scene description file in XML format – this functionality is used in CCP4i2 for generating molecular graphics of, for instance, auto-built structures.

3. Future plans

The transition towards web technologies, which is already underway with the introduction of CCP4 Cloud, will be completed in the near future by the introduction of fully fledged model building, visualisation and figure preparation web browser interfaces to the existing *Coot* and CCP4mg engines. We also foresee an increase in the number of connections to theoretical modelling packages such as AlphaFold (Jumper et al. 2021) and RoseTTAfold (Baek et al. 2021), as well as deeper harnessing of the AlphaFold Protein Structure Database (Varadi et al. 2022).

4. Software availability and data access statement

The CCP4 software suite can be obtained from http://www.ccp4.ac.uk/download. CCP4 maintains a public instance of CCP4 Cloud at http://cloud.ccp4.ac.uk available for both academic and licensed commercial users. No data were generated in the context of the present publication.

5. Individual author contributions

Jon Agirre wrote the majority of the manuscript, coordinated the authors, and contributed to *Privateer, clipper-python*, *clipper-progs, CCP4i2, CCP4 cloud, Iris,* the CCP4 Monomer Library and other software. Haroldas Bagdonas contributed to *Privateer MKIV*. James Beilsten-Edmands, Luis Fuentes-Montero, Markus Gerstel, Richard J. Gildea, James M. Parkhurst, Nicholas E. Devenish, Melanie Vollmar, David Waterman, Graeme Winter and Gwyndaf Evans contributed to xia2 (Winter) and DIALS. James Foadi and Gwyndaf Evans developed BLEND. Rafael J. Borges, Claudia Millán, Iracema Caballero, Josep Triviño Valls and Isabel Usón developed the Arcimboldo package, with Massimo Sammito and Ana Medina contributing to ALEPH. George Sheldrick is the lead developer of SHELX C/D/E;

Isabel Usón is now the SHELX C/D/E suite's main contributor and maintainer. Maarten L. Hekkelman, Robbie P. Joosten and Anastassis Perrakis develop the PDB-REDO software package. Paul Bond, Soon Wen Hoh and Kevin D. Cowtan contributed to Modelcraft and Buccaneer (Bond, Hoh & Cowtan), Nautilus (Hoh & Cowtan), the Clipper libraries (Cowtan). Tristan I. Croll, Soon Wen Hoh, Stuart McNicholas and Jon Agirre led the development of the released clipper-python module. José Javier Burgos-Mármol, Ronan M. Keegan, Filomeno Sánchez Rodríguez, Felix Simkovic, Adam J. Simpkin, Jens M.H. Thomas and Daniel J. Rigden developed SIMBAD, MrBUMP, CONKIT, Slice'N'Dice and AMPLE. Stuart J. McNicholas, Kyle Stevenson, Huw T. Jenkins, Eleanor J. Dodson, Keith S. Wilson and Martin E.M. Noble contributed to the development and testing of the CCP4i2 graphical user interface. John Berrisford and Sameer Velankar contributed towards the development of a validation and deposition task in the CCP4 graphical user interfaces. Paul Emsley is the lead developer of *Coot* and associated programs, which Bernhard Lohkamp has contributed to. William Rochira developed Iris under Jon Agirre's supervision. John Berrisford contributed towards the development of a validation and deposition task in the CCP4 graphical user interfaces. Nicholas Pearce contributed PanDDA to the suite. Joana Pereira, Egor Sobolev, Grzegorz Chojnowski and Victor S. Lamzin contributed to ARP/wARP 8.0. Pavol Skubak and Navraj S. Pannu developed Crank-2. Oleg Kovalevskiy is the lead developer of LORESTR. Fei Long is the lead developer of AceDRG, BALBES and LibG. Garib N. Murshudov is the lead developer of REFMAC5. Robert A. Nicholls is the lead developer of ProSMART. Mihaela Atanasova, Lucrezia Catapano, Robbie P. Joosten, Andrey A. Lebedev, Fei Long, Stuart J. McNicholas, Garib N. Murshudov, Robert A. Nicholls, Roberto A. Steiner and Keitaro Yamashita contributed to Refmac5 and/or to the *CCP4 monomer library*. Andrew G.W. Leslie and Harold R. (Harry) Powell led the development of MOSFLM and iMosflm respectively. Andrea Thorn is the lead developer of *AUSPEX*. Phil R. Evans is the developer of Pointless and Aimless. Alexei Vagin is the lead developer of Morda. Airlie J. McCoy, Kaushik Hatti, Robert Oeffner, Massimo Sammito, Claudia Millán and Randy J. Read developed Phaser and associated tools. Eugene Krissinel developed PISA, SSM, Gesamt and, with Andrey A. Lebedev and others, the CCP4 cloud software. Martin Malý and Petr Kolenko designed and implemented the PAIREF software. Kathryn L. Shelley and Elspeth F. Garman led the development of RABDAM. Maria Fando developed a new documentation architecture for CCP4i2 and CCP4 Cloud and converted, with help from others, old documentation to the new system. Gregorz Chojnowski developed the findMySequence software. Martyn Winn wrote the original implementation of TLS refinement in Refmac, and contributed to the development of the core C libraries and to MrBUMP.

At the time of writing, the CCP4 Executive Committee was composed of David G. Brown, Helen Walden, Kevin D. Cowtan, Judit Debreczeni, Gwyndaf Evans, Michael A. Hough, Dave Lawson, James Murray, Martyn D. Winn, Garib N. Murshudov, Martin E.M. Noble, Randy J. Read, Dan J. Rigden, Ivo Tews, Eugene Krissinel and Keith S. Wilson. Jon Agirre and Arnaud Baslé were subsequently elected as co-chairs of CCP4 Working Group 2 and took seats at the CCP4 Executive Committee, of which Ivo Tews was elected as chair. Charles B. Ballard, Ronan M. Keegan, Andrey A. Lebedev, Maria Fando, Tarik R. Drevon, David Waterman, Ville Uski and Eugene B. Krissinel were the members of the CCP4 Core Team, responsible for the maintenance and distribution of the CCP4 Software Suite, CCP4 Cloud and website.

Acknowledgements

The CCP4 program authors are grateful for the support of the more than 150 industrial licensees. CCP4 project members are indebted to Karen McIntyre for her continuous support, dedication, and her contribution as CCP4 Equity, Diversity and Inclusion champion.

Funding information

Jon Agirre is a Royal Society University Research Fellow (refs: UF160039 and URF\R\221006). Mihaela Atanasova is funded by the UK Engineering and Physical Sciences Research Council (EPSRC, ref: EP/R513386/1). Haroldas Bagdonas is funded by The Royal Society (ref: RGF/R1/181006). José Javier Burgos-Mármol and Daniel J. Rigden are supported by the BBSRC (BB/S007105/1). Robbie P. Joosten is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 871037 (iNEXT-Discovery) and from CCP4. This work was supported by the Medical Research Council as part of United Kingdom Research and Innovation, also known as UK Research and Innovation: MRC file reference No. MC_UP_A025_1012 to Garib N. Murshudov also funded Keitaro Yamashita, Paul Emsley and Fei Long. Robert A. Nicholls is funded by the BBSRC (ref: BB/S007083/1). Soon Wen Hoh is funded by the BBSRC (ref: BB/T012935/1). Kevin D. Cowtan and Paul S. Bond are funded in part by the BBSRC (ref: BB/S005099/1). John Berrisford and Sameer Velankar thank the European Molecular Biology Laboratory-European Bioinformatics Institute who supported this work. Andrea Thorn was supported to develop AUSPEX from the German Federal Ministry of Education and Research [ref: 05K19WWA and 05K22GU5] and Deutsche Forschungsgemeinschaft

[ref: TH2135/2-1]. Petr Kolenko and Martin Malý are funded by the MEYS CR (ref: CZ.02.1.01/0.0/0.0/16_019/0000778); Martin Malý is funded by the Czech Academy of Sciences (ref: 86652036) and CCP4/STFC (ref: 521862101). Anastassis Perrakis acknowledges funding from iNEXT (grant number 653706), iNEXT-Discovery (grant number 871037), West-Life (grant number 675858), and EOSC-Life (grant number 824087) funded by the Horizon 2020 program of the European Commission; Robbie P. Joosten has been the recipient of a Veni grant (722.011.011) and a Vidi grant (723.013.003) from the Netherlands Organization for Scientific Research (NWO); Maarten L. Hekkelman, Robbie P. Joosten and Anastassis Perrakis thank the Research High Performance Computing facility of the Netherlands Cancer Institute for providing and maintaining computation resources and acknowledge the institutional grant of the Dutch Cancer Society and of the Dutch Ministry of Health, Welfare and Sport. Tarik R. Drevon is funded by the BBSRC (ref: BB/S007040/1). Randy J. Read is supported by a Principal Research Fellowship from the Wellcome Trust (grant 209407/Z/17/Z). Atlanta G. Cook is supported by a Wellcome Trust SRF: 200898 and a Wellcome Centre for Cell Biology core grant: 203149. Isabel Usón acknowledges support from STFC-UK/CCP4: “Agreement for the integration of methods into the CCP4 software distribution, ARCIMBOLDO_LOW” and Spanish MICINN/AEI/FEDER/UE (PID2021-128751NB-I00). Pavol Skubak and Navraj Pannu were funded by the NWO Applied Sciences and Engineering Domain and CCP4 (grant Nos 13337 and 16219). Bernhard Lohkamp was supported by the Röntgen Ångström Cluster (Grant 349-2013-597). Nicholas Pearce is currently funded by the SciLifeLab & Wallenberg Data Driven Life Science Program (grant: KAW 2020.0239) and has previously been funded by a Veni Fellowship (VI.Veni.192.143) from the Dutch Research Council (NWO), a Long-term EMBO fellowship (ALTF 609-2017) and EPSRC grant EP/G037280/1. David M. Lawson received funding from BBSRC Institute Strategic Programme Grants (BB/P012523/1; BB/P012574/1).

References

Afonine, Pavel V., Ralf W. Grosse-Kunstleve, Nathaniel Echols, Jeffrey J. Headd, Nigel W. Moriarty, Marat Mustyakimov, Thomas C. Terwilliger, Alexandre Urzhumtsev, Peter H. Zwart, and Paul D. Adams. 2012. “Towards Automated Crystallographic Structure Refinement with Phenix.refine.” *Acta Crystallographica. Section D, Biological Crystallography* 68 (Pt 4): 352–67.

Afonine, Pavel V., Billy K. Poon, Randy J. Read, Oleg V. Sobolev, Thomas C. Terwilliger, Alexandre Urzhumtsev, and Paul D. Adams. 2018. “Real-Space Refinement in PHENIX for Cryo-EM and Crystallography.” *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 6): 531–44.

Agirre, Jon. 2017. “Strategies for Carbohydrate Model Building, Refinement and Validation.” *Acta Crystallographica. Section D, Structural Biology* 73 (Pt 2): 171–86.

Agirre, Jon, Antonio Ariza, Wendy A. Offen, Johan P. Turkenburg, Shirley M. Roberts, Stuart McNicholas, Paul V. Harris, et al. 2016. “Three-Dimensional Structures of Two Heavily N-Glycosylated Aspergillus Sp. Family GH3 β-D-Glucosidases.” *Acta Crystallographica. Section D, Structural Biology* 72 (Pt 2): 254–65.

Agirre, Jon, Gideon J. Davies, Keith S. Wilson, and Kevin D. Cowtan. 2017a. “Carbohydrate Structure: The Rocky Road to Automation.” *Current Opinion in Structural Biology* 44 (June): 39–47.

———. 2017b. “Carbohydrate Structure: The Rocky Road to Automation.” *Current Opinion in Structural Biology*. Elsevier Ltd. https://doi.org/10.1016/j.sbi.2016.11.011.

Agirre, Jon, Gideon Davies, Keith Wilson, and Kevin Cowtan. 2015a. “Carbohydrate Anomalies in the PDB.” *Nature Chemical Biology* 11 (5): 303.

———. 2015b. “Carbohydrate Anomalies in the PDB.” *Nature Chemical Biology*. Nature Publishing Group. https://doi.org/10.1038/nchembio.1798.

Agirre, Jon, and Eleanor Dodson. 2018. “Forty Years of Collaborative Computational Crystallography.” *Protein Science: A Publication of the Protein Society* 27 (1): 202–6.

Agirre, Jon, Javier Iglesias-Fernández, Carme Rovira, Gideon J. Davies, Keith S. Wilson, and Kevin D. Cowtan. 2015a. “Privateer: Software for the Conformational Validation of Carbohydrate Structures.” *Nature Structural & Molecular Biology* 22 (11): 833–34.

———. 2015b. “Privateer: Software for the Conformational Validation of Carbohydrate Structures.” *Nature Structural and Molecular Biology*. Nature Publishing Group. https://doi.org/10.1038/nsmb.3115.

Agirre, Jon, Olga Moroz, Sebastian Meier, Jesper Brask, Astrid Munch, Tine Hoff, Carsten Andersen, Keith S. Wilson, and Gideon J. Davies. 2019. “The Structure of the AliC GH13 α-Amylase from Alicyclobacillus Sp. Reveals the Accommodation of Starch Branching Points in the α-Amylase Family.” *Acta Crystallographica. Section D, Structural Biology* 75 (Pt 1): 1–7.

Akkermans, Onno, Céline Delloye-Bourgeois, Claudia Peregrina, Maria Carrasquero-Ordaz, Maria Kokolaki, Miguel Berbeira-Santana, Matthieu Chavent, et al. 2022. “GPC3-Unc5 Receptor Complex Structure and Role in Cell Migration.” *Cell* 185 (21): 3931–49.e26.

Alford, Rebecca F., Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O’Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, et al. 2017. “The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design.” *Journal of Chemical Theory and Computation* 13 (6): 3031–48.

Allen, Frank H., and Ian J. Bruno. 2010. “Bond Lengths in Organic and Metal-Organic Compounds revisited:*X*—H Bond Lengths from Neutron Diffraction Data.” *Acta Crystallographica Section B Structural Science*. https://doi.org/10.1107/s0108768110012048.

Alocci, Davide, Julien Mariethoz, Alessandra Gastaldello, Elisabeth Gasteiger, Niclas G. Karlsson, Daniel Kolarich, Nicolle H. Packer, and Frédérique Lisacek. 2019. “GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical.” *Journal of Proteome Research* 18 (2): 664–77.

Armstrong, David R., John M. Berrisford, Matthew J. Conroy, Aleksandras Gutmanas, Stephen Anyango, Preeti Choudhary, Alice R. Clark, et al. 2020. “PDBe: Improved Findability of Macromolecular Structure Data in the PDB.” *Nucleic Acids Research* 48 (D1): D335–43.

Arroyuelo, Agustina, Jorge A. Vila, and Osvaldo A. Martin. 2016. “Azahar: A PyMOL Plugin for Construction, Visualization and Analysis of Glycan Molecules.” *Journal of Computer-Aided Molecular Design* 30 (8): 619–24.

Atanasova, Mihaela, Haroldas Bagdonas, and Jon Agirre. 2020. “Structural Glycobiology in the Age of Electron Cryo-Microscopy.” *Current Opinion in Structural Biology* 62 (June):

70–78.

Atanasova, Mihaela, Robert A. Nicholls, Robbie P. Joosten, and Jon Agirre. 2022a. “Updated Restraint Dictionaries for Carbohydrates in the Pyranose Form.” *Acta Crystallographica Section D Structural Biology*. https://doi.org/10.1107/s2059798322001103.

———. 2022b. “Updated Restraint Dictionaries for Carbohydrates in the Pyranose Form.” *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 4): 455–65.

Baek, Minkyung, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, et al. 2021. “Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network.” *Science* 373 (6557): 871–76.

Bagdonas, Haroldas, Carl A. Fogarty, Elisa Fadda, and Jon Agirre. 2021. “The Case for Post-Predictional Modifications in the AlphaFold Protein Structure Database.” *Nature Structural & Molecular Biology* 28 (11): 869–70.

Bagdonas, Haroldas, Daniel Ungar, and Jon Agirre. 2020a. “Leveraging Glycomics Data in Glycoprotein 3D Structure Validation with Privateer.” *Beilstein Archives*. https://doi.org/10.3762/bxiv.2020.83.v1.

———. 2020b. “Leveraging Glycomics Data in Glycoprotein 3D Structure Validation with Privateer.” *Beilstein Journal of Organic Chemistry* 16 (October): 2523–33.

Beilsten-Edmands, James, Graeme Winter, Richard Gildea, James Parkhurst, David Waterman, and Gwyndaf Evans. 2020. “Scaling Diffraction Data in the DIALS Software Package: Algorithms and New Approaches for Multi-Crystal Scaling.” *Acta Crystallographica. Section D, Structural Biology* 76 (Pt 4): 385–99.

Berman, Helen M., Kim Henrick, Haruki Nakamura, and John Markley. 2007. “Reply to: Building Meaningful Models of Glycoproteins.” *Nature Structural & Molecular Biology*. https://doi.org/10.1038/nsmb0507-354b.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. “The Protein Data Bank.” *Nucleic Acids Research* 28 (1): 235–42.

Beusekom, Bart van, Krista Joosten, Maarten L. Hekkelman, Robbie P. Joosten, and Anastassis Perrakis. 2018. “Homology-Based Loop Modeling Yields More Complete Crystallographic Protein Structures.” *IUCrJ* 5 (Pt 5): 585–94.

Beusekom, Bart van, Thomas Lütteke, and Robbie P. Joosten. 2018. “Making Glycoproteins

a Little Bit Sweeter with PDB-REDO." *Acta Crystallographica. Section F, Structural Biology and Crystallization Communications* 74 (Pt 8): 463–72.

Beusekom, Bart van, Wouter G. Touw, Mahidhar Tatineni, Sandeep Somani, Gunaretnam Rajagopal, Jinquan Luo, Gary L. Gilliland, Anastassis Perrakis, and Robbie P. Joosten. 2018. "Homology-Based Hydrogen Bond Information Improves Crystallographic Structures in the PDB." *Protein Science: A Publication of the Protein Society* 27 (3): 798–808.

Beusekom, Bart van, Natasja Wezel, Maarten L. Hekkelman, Anastassis Perrakis, Paul Emsley, and Robbie P. Joosten. 2019. "Building and Rebuilding N-Glycans in Protein Structure Models." *Acta Crystallographica. Section D, Structural Biology* 75 (Pt 4): 416–25.

Bibby, Jaclyn, Ronan M. Keegan, Olga Mayans, Martyn D. Winn, and Daniel J. Rigden. 2012. "*AMPLE*: A Cluster-and-Truncate Approach to Solve the Crystal Structures of Small Proteins Using Rapidly Computed*ab Initio*models." *Acta Crystallographica Section D Biological Crystallography*. https://doi.org/10.1107/s0907444912039194.

Blundell, T. L., J. F. Cutfield, S. M. Cutfield, E. J. Dodson, G. G. Dodson, D. C. Hodgkin, D. A. Mercola, and M. Vijayan. 1971. "Atomic Positions in Rhombohedral 2-Zinc Insulin Crystals." *Nature* 231 (5304): 506–11.

Bohne-Lang, A., E. Lang, T. Förster, and C. W. von der Lieth. 2001. "LINUCS: Linear Notation for Unique Description of Carbohydrate Sequences." *Carbohydrate Research* 336 (1): 1–11.

Bond, Paul S., and Kevin D. Cowtan. 2022. "ModelCraft: An Advanced Automated Model-Building Pipeline Using Buccaneer." *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 9): 1090–98.

Bond, Paul S., Keith S. Wilson, and Kevin D. Cowtan. 2020. "Predicting Protein Model Correctness in Coot Using Machine Learning." *Acta Crystallographica. Section D, Structural Biology* 76 (Pt 8): 713–23.

Brewster, Aaron S., David G. Waterman, James M. Parkhurst, Richard J. Gildea, Iris D. Young, Lee J. O'Riordan, Junko Yano, Graeme Winter, Gwyndaf Evans, and Nicholas K. Sauter. 2018. "Improving Signal Strength in Serial Crystallography with DIALS Geometry Refinement." *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 9): 877–94.

Brindley, G. W. 1933. "On the Reflection and Refraction of X-Rays by Perfect Crystals."

Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 140 (841): 301–13.

Brown, Alan, Fei Long, Robert A. Nicholls, Jaan Toots, Paul Emsley, and Garib Murshudov. 2015. “Tools for Macromolecular Model Building and Refinement into Electron Cryo-Microscopy Reconstructions.” *Acta Crystallographica. Section D, Biological Crystallography* 71 (Pt 1): 136–53.

Brown, I. David, and Brian McMahon. 2002. “CIF: The Computer Language of Crystallography.” *Acta Crystallographica. Section B: Structural Crystallography and Crystal Chemistry* 58 (Pt 3 Pt 1): 317–24.

Bruno, Ian J., Jason C. Cole, Magnus Kessler, Jie Luo, W. D. Sam Motherwell, Lucy H. Purkis, Barry R. Smith, et al. 2004. “Retrieval of Crystallographically-Derived Molecular Geometry Information.” *Journal of Chemical Information and Computer Sciences* 44 (6): 2133–44.

Burley, Stephen K., Helen M. Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, et al. 2019. “RCSB Protein Data Bank: Biological Macromolecular Structures Enabling Research and Education in Fundamental Biology, Biomedicine, Biotechnology and Energy.” *Nucleic Acids Research* 47 (D1): D464–74.

Burmeister, W. P., S. Cottaz, P. Rollin, A. Vasella, and B. Henrissat. 2000. “High Resolution X-Ray Crystallography Shows That Ascorbate Is a Cofactor for Myrosinase and Substitutes for the Function of the Catalytic Base.” *The Journal of Biological Chemistry* 275 (50): 39385–93.

Burnley, Tom, Colin M. Palmer, and Martyn Winn. 2017. “Recent Developments in the CCP-EM Software Suite.” *Acta Crystallographica. Section D, Structural Biology* 73 (Pt 6): 469–77.

Caballero, Iracema, Massimo Sammito, Claudia Millán, Andrey Lebedev, Nicolas Soler, and Isabel Usón. 2018. “ARCIMBOLDO on Coiled Coils.” *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 3): 194–204.

Casalino, Lorenzo, Zied Gaieb, Jory A. Goldsmith, Christy K. Hjorth, Abigail C. Dommer, Aoife M. Harbison, Carl A. Fogarty, et al. 2020. “Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein.” *ACS Central Science* 6 (10): 1722–34.

Casañal, Ana, Bernhard Lohkamp, and Paul Emsley. 2020. “Current Developments in Coot for Macromolecular Model Building of Electron Cryo-Microscopy and Crystallographic

Data.” *Protein Science: A Publication of the Protein Society* 29 (4): 1069–78.

Case, David A., Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. 2005. “The Amber Biomolecular Simulation Programs.” *Journal of Computational Chemistry*. https://doi.org/10.1002/jcc.20290.

Catapano, Lucrezia, Roberto A. Steiner, and Garib N. Murshudov. 2021. “Modelling and Refinement of Hydrogens: New Developments in the CCP4 Suite.” *Acta Crystallographica Section A Foundations and Advances*. https://doi.org/10.1107/s0108767321093053.

Chen, Chia-Lin, Jen-Chi Hsu, Chin-Wei Lin, Chia-Hung Wang, Ming-Hung Tsai, Chung-Yi Wu, Chi-Huey Wong, and Che Ma. 2017. “Crystal Structure of a Homogeneous IgG-Fc Glycoform with the N-Glycan Designed to Maximize the Antibody Dependent Cellular Cytotoxicity.” *ACS Chemical Biology* 12 (5): 1335–45.

Chojnowski, Grzegorz, Joana Pereira, and Victor S. Lamzin. 2019. “Sequence Assignment for Low-Resolution Modelling of Protein Crystal Structures.” *Acta Crystallographica. Section D, Structural Biology* 75 (Pt 8): 753–63.

Chojnowski, Grzegorz, Adam J. Simpkin, Diego A. Leonardo, Wolfram Seifert-Davila, Dan E. Vivas-Ruiz, Ronan M. Keegan, and Daniel J. Rigden. 2022. “*findMySequence*: A Neural-Network-Based Approach for Identification of Unknown Proteins in X-Ray Crystallography and Cryo-EM.” *IUCrJ*. https://doi.org/10.1107/s2052252521011088.

Chojnowski, Grzegorz, Egor Sobolev, Philipp Heuser, and Victor S. Lamzin. 2021. “The Accuracy of Protein Models Automatically Built into Cryo-EM Maps with ARP/wARP.” *Acta Crystallographica. Section D, Structural Biology* 77 (Pt 2): 142–50.

Clabbers, Max T. B., Tim Gruene, James M. Parkhurst, Jan Pieter Abrahams, and David G. Waterman. 2018. “Electron Diffraction Data Processing with DIALS.” *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 6): 506–18.

Clauset, Aaron, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. “Power-Law Distributions in Empirical Data.” *SIAM Rev.* 51 (4): 661–703.

Cowtan, Kevin. 2006. “The Buccaneer Software for Automated Model Building. 1. Tracing Protein Chains.” *Acta Crystallographica. Section D, Biological Crystallography* 62 (Pt 9): 1002–11.

———. 2010a. “Recent Developments in Classical Density Modification.” *Acta Crystallographica Section D Biological Crystallography*. https://doi.org/10.1107/s090744490903947x.

———. 2010b. “Recent Developments in Classical Density Modification.” *Acta Crystallographica. Section D, Biological Crystallography* 66 (Pt 4): 470–78.

———. 2012. “Completion of Autobuilt Protein Models Using a Database of Protein Fragments.” *Acta Crystallographica. Section D, Biological Crystallography* 68 (Pt 4): 328–35.

———. 2014. “Automated Nucleic Acid Chain Tracing in Real Time.” *IUCrJ* 1 (Pt 6): 387–92.

Cowtan, Kevin, and Jon Agirre. 2018. “Macromolecular Refinement by Model Morphing Using Non-Atomic Parameterizations.” *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 2): 125–31.

Cowtan, K., S. Metcalfe, and P. Bond. 2020. “Shift-Field Refinement of Macromolecular Atomic Models.” *Acta Crystallographica. Section D, Structural Biology* 76 (Pt 12): 1192–1200.

Cremer, D., and J. A. Pople. 1975a. “A General Definition of Ring Puckering Coordinates.” *Journal of the American Chemical Society* 97: 1354–58.

———. 1975b. “General Definition of Ring Puckering Coordinates.” *J. Am. Chem. Soc.* 97 (6): 1354–58.

Crispin, Max, David I. Stuart, and E. Yvonne Jones. 2007. “Building Meaningful Models of Glycoproteins.” *Nature Structural & Molecular Biology*.

Crispin, Max, Xiaojie Yu, and Thomas A. Bowden. 2013. “Crystal Structure of Sialylated IgG Fc: Implications for the Mechanism of Intravenous Immunoglobulin Therapy.” *Proceedings of the National Academy of Sciences of the United States of America*.

Croll, Tristan Ian. 2018. “ISOLDE: A Physically Realistic Environment for Model Building into Low-Resolution Electron-Density Maps.” *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 6): 519–30.

Croll, Tristan I., Kay Diederichs, Florens Fischer, Cameron D. Fyfe, Yunyun Gao, Sam Horrell, Agnel Praveen Joseph, et al. 2021. “Making the Invisible Enemy Visible.” *Nature Structural & Molecular Biology* 28 (5): 404–8.

Davies, Gideon J., Antoni Planas, and Carme Rovira. 2012. “Conformational Analyses of the Reaction Coordinate of Glycosidases.” *Accounts of Chemical Research* 45 (2): 308–16.

Devi, Seenivasan Karthiga, Vishnu Priyanka Reddy Chichili, J. Jeyakanthan, D. Velmurugan, and J. Sivaraman. 2015. “Structural Basis for the Hydrolysis of ATP by a Nucleotide Binding Subunit of an Amino Acid ABC Transporter from Thermus Thermophilus.” *Journal of Structural Biology* 190 (3): 367–72.

Dewar, Michael J. S., Eve G. Zoebisch, Eamonn F. Healy, and James J. P. Stewart. 1985. “Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model.” *Journal of the American Chemical Society* 107 (13): 3902–9.

Dialpuri, Jordan, Haroldas Bagdonas, Mihaela Atanasova, Lucy Schofield, Maarten Hekkelman, Robbie Joosten, and Jon Agirre. 2022. “Analysis and Validation of Overall N-Glycan Conformation in Privateer.” Zenodo. https://doi.org/10.5281/ZENODO.7356467.

DiMaio, Frank, Nathaniel Echols, Jeffrey J. Headd, Thomas C. Terwilliger, Paul D. Adams, and David Baker. 2013. “Improved Low-Resolution Crystallographic Refinement with Phenix and Rosetta.” *Nature Methods* 10 (11): 1102–4.

Dobie, Christopher, and Danielle Skropeta. 2021. “Insights into the Role of Sialylation in Cancer Progression and Metastasis.” *British Journal of Cancer* 124 (1): 76–90.

Dubochet, J., M. Adrian, J. J. Chang, J. C. Homo, J. Lepault, A. W. McDowall, and P. Schultz. 1988. “Cryo-Electron Microscopy of Vitrified Specimens.” *Quarterly Reviews of Biophysics* 21 (2): 129–228.

Eastman, Peter, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, et al. 2017. “OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics.” *PLoS Computational Biology* 13 (7): e1005659.

Eddy, Sean R. 2011. “Accelerated Profile HMM Searches.” *PLoS Computational Biology* 7 (10): e1002195.

Emsley, Paul. 2017. “Tools for Ligand Validation in Coot.” *Acta Crystallographica. Section D, Structural Biology* 73 (Pt 3): 203–10.

Emsley, Paul, Axel T. Brunger, and Thomas Lütteke. 2015. “Tools to Assist Determination

and Validation of Carbohydrate 3D Structure Data." *Methods in Molecular Biology* 1273: 229–40.

Emsley, Paul, and Kevin Cowtan. 2004. "Coot: Model-Building Tools for Molecular Graphics." *Acta Crystallographica. Section D, Biological Crystallography* 60 (Pt 12 Pt 1): 2126–32.

Emsley, Paul, and Max Crispin. 2018. "Structural Analysis of Glycoproteins: Building N-Linked Glycans with Coot." *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 4): 256–63.

Emsley, P., B. Lohkamp, W. G. Scott, and K. Cowtan. 2010. "Features and Development of Coot." *Acta Crystallographica. Section D, Biological Crystallography* 66 (Pt 4): 486–501.

Engh, R. A., and R. Huber. 1991. "Accurate Bond and Angle Parameters for X-Ray Protein Structure Refinement." *Acta Crystallographica Section A Foundations of Crystallography*. https://doi.org/10.1107/s0108767391001071.

Enzo, S., G. Fagherazzi, A. Benedetti, and S. Polizzi. 1988. "A Profile-Fitting Procedure for Analysis of Broadened X-Ray Diffraction Peaks. I. Methodology." *Journal of Applied Crystallography* 21 (5): 536–42.

Evans, Philip R., and Garib N. Murshudov. 2013. "How Good Are My Data and What Is the Resolution?" *Acta Crystallographica. Section D, Biological Crystallography* 69 (Pt 7): 1204–14.

Ferrara, Claudia, Sandra Grau, Christiane Jäger, Peter Sondermann, Peter Brünker, Inja Waldhauer, Michael Hennig, et al. 2011. "Unique Carbohydrate–carbohydrate Interactions Are Required for High Affinity Binding between FcγRIII and Antibodies Lacking Core Fucose." *Proceedings of the National Academy of Sciences* 108 (31): 12669–74.

Ferrari, Simone, Daniel V. Savatin, Francesca Sicilia, Giovanna Gramegna, Felice Cervone, and Giulia De Lorenzo. 2013. "Oligogalacturonides: Plant Damage-Associated Molecular Patterns and Regulators of Growth and Development." *Frontiers in Plant Science* 4 (March): 49.

Fisher, Peter, Jane Thomas-Oates, A. Jamie Wood, and Daniel Ungar. 2019. "The N-Glycosylation Processing Potential of the Mammalian Golgi Apparatus." *Frontiers in Cell and Developmental Biology* 7 (August): 157.

Frank, Martin, Daniela Beccati, Bas R. Leeflang, and Johannes F. G. Vliegenthart. 2020.

"C-Mannosylation Enhances the Structural Stability of Human RNase 2." *iScience* 23 (8): 101371.

Frank, M., T. Lütteke, and C-W von der Lieth. 2007. "GlycoMapsDB: A Database of the Accessible Conformational Space of Glycosidic Linkages." *Nucleic Acids Research* 35 (Database issue): 287–90.

French, S., and K. Wilson. 1978. "On the Treatment of Negative Intensity Observations." *Acta Crystallographica Section A*. https://doi.org/10.1107/s0567739478001114.

Frenz, Brandon, Sebastian Rämisch, Andrew J. Borst, Alexandra C. Walls, Jared Adolf-Bryfogle, William R. Schief, David Veesler, and Frank DiMaio. 2019. "Automatically Fixing Errors in Glycoprotein Structures with Rosetta." *Structure* 27 (1): 134–39.e3.

Fuentes-Montero, L., J. Parkhurst, M. Gerstel, R. Gildea, G. Winter, M. Vollmar, D. Waterman, G. Evans, and IUCr. 2016. "Introducing DUI, a Graphical Interface for DIALS." *Acta Crystallographica Section A: Foundations and Advances* 72 (August): s189–s189.

Fushinobu, Shinya. 2018. "Conformations of the Type-1 Lacto-N-Biose I Unit in Protein Complex Structures." *Acta Crystallographica. Section F, Structural Biology and Crystallization Communications* 74 (Pt 8): 473–79.

Gao, Yan, Liming Yan, Yucen Huang, Fengjiang Liu, Yao Zhao, Lin Cao, Tao Wang, et al. 2020. "Structure of the RNA-Dependent RNA Polymerase from COVID-19 Virus." *Science* 368 (6492): 779–82.

Gildea, R. J., J. Beilsten-Edmands, D. Axford, S. Horrell, P. Aller, J. Sandy, J. Sanchez-Weatherby, et al. 2022. "xia2.multiplex: A Multi-Crystal Data-Analysis Pipeline." *Acta Crystallographica Section D: Structural Biology* 78 (6): 752–69.

Ginn, Helen Mary, Aaron S. Brewster, Johan Hattne, Gwyndaf Evans, Armin Wagner, Jonathan M. Grimes, Nicholas K. Sauter, Geoff Sutton, and David Ian Stuart. 2015. "A Revised Partiality Model and Post-Refinement Algorithm for X-Ray Free-Electron Laser Data." *Acta Crystallographica. Section D, Biological Crystallography* 71 (Pt 6): 1400–1410.

Glanz, Victor Yu, Veronika A. Myasoedova, Andrey V. Grechko, and Alexander N. Orekhov. 2018. "Inhibition of Sialidase Activity as a Therapeutic Approach." *Drug Design, Development and Therapy* 12 (October): 3431–37.

glycojones. n.d. “Glycojones/sails.” GitHub. Accessed September 20, 2019. https://github.com/glycojones/sails.

Goddard, Thomas D., Conrad C. Huang, Elaine C. Meng, Eric F. Pettersen, Gregory S. Couch, John H. Morris, and Thomas E. Ferrin. 2018. “UCSF ChimeraX: Meeting Modern Challenges in Visualization and Analysis.” *Protein Science: A Publication of the Protein Society* 27 (1): 14–25.

Gore, Swanand, Eduardo Sanz García, Pieter M. S. Hendrickx, Aleksandras Gutmanas, John D. Westbrook, Huanwang Yang, Zukang Feng, et al. 2017. “Validation of Structures in the Protein Data Bank.” *Structure* 25 (12): 1916–27.

Gore, Swanand, Sameer Velankar, and Gerard J. Kleywegt. 2012. “Implementing an X-Ray Validation Pipeline for the Protein Data Bank.” *Acta Crystallographica. Section D, Biological Crystallography* 68 (Pt 4): 478–83.

Gražulis, Saulius, Adriana Daškevič, Andrius Merkys, Daniel Chateigner, Luca Lutterotti, Miguel Quirós, Nadezhda R. Serebryanaya, Peter Moeck, Robert T. Downs, and Armel Le Bail. 2012. “Crystallography Open Database (COD): An Open-Access Collection of Crystal Structures and Platform for World-Wide Collaboration.” *Nucleic Acids Research* 40 (Database issue): D420–27.

Groom, Colin R., Ian J. Bruno, Matthew P. Lightfoot, and Suzanna C. Ward. 2016. “The Cambridge Structural Database.” *Acta Crystallographica Section B, Structural Science, Crystal Engineering and Materials* 72 (Pt 2): 171–79.

Grosse-Kunstleve, Ralf W., Nicholas K. Sauter, Nigel W. Moriarty, and Paul D. Adams. 2002. “TheComputational Crystallography Toolbox: Crystallographic Algorithms in a Reusable Software Framework.” *Journal of Applied Crystallography*. https://doi.org/10.1107/s0021889801017824.

Gudmundsson, Mikael, Henrik Hansson, Saeid Karkehabadi, Anna Larsson, Ingeborg Stals, Steve Kim, Sergio Sunux, et al. 2016. “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* 72 (Pt 7): 860–70.

Gui, Miao, Wenfei Song, Haixia Zhou, Jingwei Xu, Silian Chen, Ye Xiang, and Xinquan Wang. 2017. “Cryo-Electron Microscopy Structures of the SARS-CoV Spike Glycoprotein Reveal a Prerequisite Conformational State for Receptor Binding.” *Cell*

Research 27 (1): 119–29.

Hall, S. R., F. H. Allen, and I. D. Brown. 1991. “The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography.” *Acta Crystallographica. Section A, Foundations of Crystallography* 47 (6): 655–85.

Hoh, Soon Wen, Tom Burnley, and Kevin Cowtan. 2020. “Current Approaches for Automated Model Building into Cryo-EM Maps Using Buccaneer with CCP-EM.” *Acta Crystallographica. Section D, Structural Biology* 76 (Pt 6): 531–41.

Hooft, R. W., C. Sander, and G. Vriend. 1997. “Objectively Judging the Quality of a Protein Structure from a Ramachandran Plot.” *Computer Applications in the Biosciences: CABIOS* 13 (4): 425–30.

Imberty, A., and S. Perez. 1995. “Stereochemistry of the N-Glycosylation Sites in Glycoproteins.” *Protein Eng. Des. Sel.* 8 (7): 699–709.

insilichem. n.d. “Insilichem/tangram_snfg.” GitHub. Accessed September 20, 2019. https://github.com/insilichem/tangram_snfg.

Jenkins, Huw T. 2018. “Fragon: Rapid High-Resolution Structure Determination from Ideal Protein Fragments.” *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 3): 205–14.

Jin, Yi, Marija Petricevic, Alan John, Lluís Raich, Huw Jenkins, Leticia Portela De Souza, Fiona Cuskin, et al. 2016. “A β-Mannanase with a Lysozyme-like Fold and a Novel Molecular Catalytic Mechanism.” *ACS Central Science* 2 (12): 896–903.

Jin, Zhenming, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, et al. 2020. “Structure of Mpro from SARS-CoV-2 and Discovery of Its Inhibitors.” *Nature* 582 (7811): 289–93.

Joosten, Robbie P., Krista Joosten, Garib N. Murshudov, and Anastassis Perrakis. 2012. “PDB_REDO: Constructive Validation, More than Just Looking for Errors.” *Acta Crystallographica. Section D, Biological Crystallography* 68 (Pt 4): 484–96.

Joosten, Robbie P., Fei Long, Garib N. Murshudov, and Anastassis Perrakis. 2014. “The PDB_REDO Server for Macromolecular Structure Model Optimization.” *IUCrJ* 1 (Pt 4): 213–20.

Joosten, Robbie P., and Thomas Lütteke. 2017a. “Carbohydrate 3D Structure Validation.” *Current Opinion in Structural Biology* 44 (June): 9–17.

———. 2017b. “Carbohydrate 3D Structure Validation.” *Current Opinion in Structural Biology*. Elsevier Ltd. https://doi.org/10.1016/j.sbi.2016.10.010.

Joosten, Robbie P., Robert A. Nicholls, and Jon Agirre. 2022. “Towards Consistency in Geometry Restraints for Carbohydrates in the Pyranose Form: Modern Dictionary Generators Reviewed.” *Current Medicinal Chemistry* 29 (7): 1193–1207.

Jo, Sunhwan, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. 2008. “CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM.” *Journal of Computational Chemistry* 29 (11): 1859–65.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.

Karplus, P. Andrew, and Kay Diederichs. 2012. “Linking Crystallographic Model and Data Quality.” *Science* 336 (6084): 1030–33.

Keegan, Ronan M., Stuart J. McNicholas, Jens M. H. Thomas, Adam J. Simpkin, Felix Simkovic, Ville Uski, Charles C. Ballard, Martyn D. Winn, Keith S. Wilson, and Daniel J. Rigden. 2018. “Recent Developments in MrBUMP: Better Search-Model Preparation, Graphical Interaction with Search Models, and Solution Improvement and Assessment.” *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 3): 167–82.

Kirschner, Karl N., Austin B. Yongye, Sarah M. Tschampel, Jorge González-Outeiriño, Charlisa R. Daniels, B. Lachele Foley, and Robert J. Woods. 2008. “GLYCAM06: A Generalizable Biomolecular Force Field. Carbohydrates.” *Journal of Computational Chemistry* 29 (4): 622–55.

Kleywegt, G. J., and T. A. Jones. 1995. “Where Freedom Is Given, Liberties Are Taken.” *Structure* 3 (6): 535–40.

Klünemann, Thomas, Arne Preuß, Julia Adamczack, Luis F. M. Rosa, Falk Harnisch, Gunhild Layer, and Wulf Blankenfeldt. 2019. “Crystal Structure of Dihydro-Heme d1 Dehydrogenase NirN from Pseudomonas Aeruginosa Reveals Amino Acid Residues Essential for Catalysis.” *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2019.05.046.

Kovalevskiy, Oleg, Robert A. Nicholls, Fei Long, Azzurra Carlon, and Garib N. Murshudov. 2018. “Overview of Refinement Procedures within REFMAC5: Utilizing Data from Different Sources.” *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 3):

215–27.

Kovalevskiy, Oleg, Robert A. Nicholls, and Garib N. Murshudov. 2016. “Automated Refinement of Macromolecular Structures at Low Resolution Using Prior Information.” *Acta Crystallographica. Section D, Structural Biology* 72 (Pt 10): 1149–61.

Krawczyk, Lucas, Shubham Semwal, Jalal Soubhye, Salma Lemri Ouadriri, Martin Prévost, Pierre Van Antwerpen, Goedele Roos, and Julie Bouckaert. 2022. “Native Glycosylation and Binding of the Antidepressant Paroxetine in a Low-Resolution Crystal Structure of Human Myeloperoxidase.” *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 9): 1099–1109.

Krissinel, Eugene, Andrey A. Lebedev, Ville Uski, Charles B. Ballard, Ronan M. Keegan, Oleg Kovalevskiy, Robert A. Nicholls, et al. 2022. “CCP4 Cloud for Structure Determination and Project Management in Macromolecular Crystallography.” *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 9): 1079–89.

Krissinel, Evgeny, and Kim Henrick. 2007. “Inference of Macromolecular Assemblies from Crystalline State.” *Journal of Molecular Biology* 372 (3): 774–97.

Labonte, Jason W., Jared Adolf-Bryfogle, William R. Schief, and Jeffrey J. Gray. 2017. “Residue-Centric Modeling and Design of Saccharide and Glycoconjugate Structures.” *Journal of Computational Chemistry* 38 (5): 276–87.

Lamzin, V. S., A. Perrakis, and K. S. Wilson. 2012a. “*ARP/wARP*– Automated Model Building and Refinement.” *International Tables for Crystallography*. https://doi.org/10.1107/97809553602060000862.

———. 2012b. “ARP/wARP– Automated Model Building and Refinement.” In *International Tables for Crystallography*, 525–28. Chester, England: International Union of Crystallography.

Langer, Gerrit, Serge X. Cohen, Victor S. Lamzin, and Anastassis Perrakis. 2008. “Automated Macromolecular Model Building for X-Ray Crystallography Using ARP/wARP Version 7.” *Nature Protocols* 3 (7): 1171–79.

Lan, Jun, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, et al. 2020. “Structure of the SARS-CoV-2 Spike Receptor-Binding Domain Bound to the ACE2 Receptor.” *Nature* 581 (7807): 215–20.

Law, Michael J., Michael E. Linde, Eric J. Chambers, Chris Oubridge, Phinikoula S.

Katsamba, Lennart Nilsson, Ian S. Haworth, and Ite A. Laird-Offringa. 2006. “The Role of Positively Charged Amino Acids and Electrostatic Interactions in the Complex of U1A Protein and U1 Hairpin II RNA.” *Nucleic Acids Research* 34 (1): 275–85.

Lebedev, Andrey A., and Michail N. Isupov. 2014. “Space-Group and Origin Ambiguity in Macromolecular Structures with Pseudo-Symmetry and Its Treatment with the Program Zanuda.” *Acta Crystallographica. Section D, Biological Crystallography* 70 (Pt 9): 2430–43.

Lebedev, Andrey A., Paul Young, Michail N. Isupov, Olga V. Moroz, Alexey A. Vagin, and Garib N. Murshudov. 2012. “JLigand: A Graphical Tool for the CCP4 Template-Restraint Library.” *Acta Crystallographica. Section D, Biological Crystallography* 68 (Pt 4): 431–40.

Lee-Sundlov, Melissa M., Sean R. Stowell, and Karin M. Hoffmeister. 2020. “Multifaceted Role of Glycosylation in Transfusion Medicine, Platelets, and Red Blood Cells.” *Journal of Thrombosis and Haemostasis: JTH* 18 (7): 1535–47.

Lee, Yongchan, Pattama Wiriyasermkul, Chunhuan Jin, Lili Quan, Ryuichi Ohgaki, Suguru Okuda, Tsukasa Kusakizako, et al. 2019. “Cryo-EM Structure of the Human L-Type Amino Acid Transporter 1 in Complex with Glycoprotein CD98hc.” *Nature Structural & Molecular Biology* 26 (6): 510–17.

Liebschner, Dorothee, Pavel V. Afonine, Matthew L. Baker, Gábor Bunkóczi, Vincent B. Chen, Tristan I. Croll, Bradley Hintze, et al. 2019. “Macromolecular Structure Determination Using X-Rays, Neutrons and Electrons: Recent Developments in Phenix.” *Acta Crystallographica. Section D, Structural Biology* 75 (Pt 10): 861–77.

Long, Fei, Robert A. Nicholls, Paul Emsley, Saulius Graæulis, Andrius Merkys, Antanas Vaitkus, and Garib N. Murshudov. 2017a. “AceDRG: A Stereochemical Description Generator for Ligands.” *Acta Crystallographica. Section D, Structural Biology* 73 (Pt 2): 112–22.

———. 2017b. “Validation and Extraction of Molecular-Geometry Information from Small-Molecule Databases.” *Acta Crystallographica. Section D, Structural Biology* 73 (Pt 2): 103–11.

Long, Fei, Alexei A. Vagin, Paul Young, and Garib N. Murshudov. 2008. “BALBES: A Molecular-Replacement Pipeline.” *Acta Crystallographica. Section D, Biological Crystallography* 64 (Pt 1): 125–32.

Lutteke, T. 2004. “Carbohydrate Structure Suite (CSS): Analysis of Carbohydrate 3D Structures Derived from the PDB.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gki013.

Lütteke, Thomas, Martin Frank, and Claus-W von der Lieth. 2005. “Carbohydrate Structure Suite (CSS): Analysis of Carbohydrate 3D Structures Derived from the PDB.” *Nucleic Acids Research* 33 (Database issue): D242–46.

Lütteke, Thomas, Martin Frank, and Claus W. Von Der Lieth. 2004. “Data Mining the Protein Data Bank: Automatic Detection and Assignment of Carbohydrate Structures.” In *Carbohydrate Research*, 339:1015–20.

Lütteke, Thomas, and Claus-W von der Lieth. 2004. “BMC Bioinformatics.” https://doi.org/10.1186/1471-2105-5-69.

Lütteke, Thomas, and Claus W. von der Lieth. 2009. “Data Mining the PDB for Glyco-Related Data.” In *Glycomics: Methods and Protocols*, edited by Nicolle H. Packer and Niclas G. Karlsson, 293–310. Totowa, NJ: Humana Press.

Lütteke, Thomas, and Claus-W Von Der Lieth. 2004. “Pdb-Care (PDB CArbohydrate REsidue Check): A Program to Support Annotation of Complex Carbohydrate Structures in PDB Files.” *BMC Bioinformatics* 5 (69). http://www.biomedcentral.com/1471-2105/5/69.

Malý, Martin, Kay Diederichs, Jan Dohnálek, and Petr Kolenko. 2020. “Paired Refinement under the Control of.” *IUCrJ* 7 (Pt 4): 681–92.

Masmaliyeva, Rafiga C., Kave H. Babai, and Garib N. Murshudov. 2020. “Local and Global Analysis of Macromolecular Atomic Displacement Parameters.” *Acta Crystallographica. Section D, Structural Biology* 76 (Pt 10): 926–37.

McCoy, Airlie J., Ralf W. Grosse-Kunstleve, Paul D. Adams, Martyn D. Winn, Laurent C. Storoni, and Randy J. Read. 2007. “Phaser Crystallographic Software.” *Journal of Applied Crystallography* 40 (Pt 4): 658–74.

McCoy, Airlie J., Robert D. Oeffner, Antoni G. Wrobel, Juha R. M. Ojala, Karl Tryggvason, Bernhard Lohkamp, and Randy J. Read. 2017. “Ab Initio Solution of Macromolecular Crystal Structures without Direct Methods.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (14): 3637–41.

McCoy, Airlie J., Massimo D. Sammito, and Randy J. Read. 2022. “Implications of

AlphaFold2 for Crystallographic Phasing by Molecular Replacement.” *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 1): 1–13.

McNicholas, S., E. Potterton, K. S. Wilson, and M. E. M. Noble. 2011. “Presenting Your Structures: The CCP4mg Molecular-Graphics Software.” *Acta Crystallographica. Section D, Biological Crystallography* 67 (Pt 4): 386–94.

McNicholas, Stuart, and Jon Agirre. 2017. “Glycoblocks: A Schematic Three-Dimensional Representation for Glycans and Their Interactions.” *Acta Crystallographica. Section D, Structural Biology* 73 (Pt 2): 187–94.

McNicholas, Stuart, Tristan Croll, Tom Burnley, Colin M. Palmer, Soon Wen Hoh, Huw T. Jenkins, Eleanor Dodson, Kevin Cowtan, and Jon Agirre. 2018. “Automating Tasks in Protein Structure Determination with the Clipper Python Module.” *Protein Science: A Publication of the Protein Society* 27 (1): 207–16.

Medina, Ana, Elisabet Jiménez, Iracema Caballero, Albert Castellví, Josep Triviño Valls, Martin Alcorlo, Rafael Molina, et al. 2022. “Verification: Model-Free Phasing with Enhanced Predicted Models in ARCIMBOLDO_SHREDDER.” *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 11): 1283–93.

Millán, Claudia, Elisabet Jiménez, Antonia Schuster, Kay Diederichs, and Isabel Usón. 2020. “ALIXE: A Phase-Combination Tool for Fragment-Based Molecular Replacement.” *Acta Crystallographica. Section D, Structural Biology* 76 (Pt 3): 209–20.

Moriarty, Nigel W., Ralf W. Grosse-Kunstleve, and Paul D. Adams. 2009. “Electronic Ligand Builder and Optimization Workbench (eLBOW): A Tool for Ligand Coordinate and Restraint Generation.” *Acta Crystallographica. Section D, Biological Crystallography* 65 (Pt 10): 1074–80.

Moriarty, Nigel W., Dale E. Tronrud, Paul D. Adams, and P. Andrew Karplus. 2014. “Conformation-Dependent Backbone Geometry Restraints Set a New Standard for Protein Crystallographic Refinement.” *The FEBS Journal* 281 (18): 4061–71.

Moussian, Bernard. 2019. “Chitin: Structure, Chemistry and Biology.” *Advances in Experimental Medicine and Biology* 1142: 5–18.

Murshudov, Garib N., Pavol Skubák, Andrey A. Lebedev, Navraj S. Pannu, Roberto A. Steiner, Robert A. Nicholls, Martyn D. Winn, Fei Long, and Alexei A. Vagin. 2011. “REFMAC5 for the Refinement of Macromolecular Crystal Structures.” *Acta Crystallographica. Section D, Biological Crystallography* 67 (Pt 4): 355–67.

Mylona, Anastasia, Stephen Carr, Pierre Aller, Isabel Moraes, Richard Treisman, Gwyndaf Evans, and James Foadi. 2017. “A Novel Approach to Data Collection for Difficult Structures: Data Management for Large Numbers of Crystals with the BLEND Software.” *Crystals* 7 (8): 242.

Nanao, Max H., George M. Sheldrick, and Raimond B. G. Ravelli. 2005. “Improving Radiation-Damage Substructures for RIP.” *Acta Crystallographica. Section D, Biological Crystallography* 61 (Pt 9): 1227–37.

Navia, M. A., P. M. Fitzgerald, B. M. McKeever, C. T. Leu, J. C. Heimbach, W. K. Herber, I. S. Sigal, P. L. Darke, and J. P. Springer. 1989. “Three-Dimensional Structure of Aspartyl Protease from Human Immunodeficiency Virus HIV-1.” *Nature* 337 (6208): 615–20.

Newby, Zachary E. R., Joseph O’Connell 3rd, Yaneth Robles-Colmenares, Shahram Khademi, Larry J. Miercke, and Robert M. Stroud. 2008. “Crystal Structure of the Aquaglyceroporin PfAQP from the Malarial Parasite Plasmodium Falciparum.” *Nature Structural & Molecular Biology* 15 (6): 619–25.

Nicholls, Robert A. 2017. “Ligand Fitting with CCP4.” *Acta Crystallographica. Section D, Structural Biology* 73 (Pt 2): 158–70.

Nicholls, Robert A., Marcus Fischer, Stuart McNicholas, and Garib N. Murshudov. 2014. “Conformation-Independent Structural Comparison of Macromolecules with ProSMART.” *Acta Crystallographica. Section D, Biological Crystallography* 70 (Pt 9): 2487–99.

Nicholls, Robert A., Robbie P. Joosten, Fei Long, Marcin Wojdyr, Andrey Lebedev, Eugene Krissinel, Lucrezia Catapano, Marcus Fischer, Paul Emsley, and Garib N. Murshudov. 2021a. “Modelling Covalent Linkages in CCP4.” *Acta Crystallographica Section D Structural Biology*. https://doi.org/10.1107/s2059798321001753.

———. 2021b. “Modelling Covalent Linkages in CCP4.” *Acta Crystallographica. Section D, Structural Biology* 77 (Pt 6): 712–26.

Nicholls, Robert A., Fei Long, and Garib N. Murshudov. 2012. “Low-Resolution Refinement Tools in REFMAC5.” *Acta Crystallographica. Section D, Biological Crystallography* 68 (Pt 4): 404–17.

Nicholls, Robert A., Marcin Wojdyr, Robbie P. Joosten, Lucrezia Catapano, Fei Long, Marcus Fischer, Paul Emsley, and Garib N. Murshudov. 2021. “The Missing Link: Covalent Linkages in Structural Models.” *Acta Crystallographica. Section D, Structural Biology* 77 (Pt 6): 727–45.

Nnamchi, Chukwudi I., Gary Parkin, Igor Efimov, Jaswir Basran, Hanna Kwon, Dimitri A. Svistunenko, Jon Agirre, et al. 2016. “Structural and Spectroscopic Characterisation of a Heme Peroxidase from Sorghum.” *Journal of Biological Inorganic Chemistry: JBIC: A Publication of the Society of Biological Inorganic Chemistry* 21 (1): 63–70.

Noinaj, Nicholas, Nicole C. Easley, Muse Oke, Naoko Mizuno, James Gumbart, Evzen Boura, Ashley N. Steere, et al. 2012. “Structural Basis for Iron Piracy by Pathogenic Neisseria.” *Nature* 483 (7387): 53–58.

Nolte, Kristopher, Yunyun Gao, Sabrina Stäb, Philip Kollmannsberger, and Andrea Thorn. 2022. “Detecting Ice Artefacts in Processed Macromolecular Diffraction Data with Machine Learning.” *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 2): 187–95.

Oeffner, Robert D., Tristan I. Croll, Claudia Millán, Billy K. Poon, Christopher J. Schlicksup, Randy J. Read, and Tom C. Terwilliger. 2022. “Putting AlphaFold Models to Work with Phenix.process_predicted_model and ISOLDE.” *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 11): 1303–14.

Park, Hahnbeom, Sergey Ovchinnikov, David E. Kim, Frank DiMaio, and David Baker. 2018. “Protein Homology Model Refinement by Large-Scale Energy Optimization.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (12): 3054–59.

Parkhurst, James M., Aaron S. Brewster, Luis Fuentes-Montero, David G. Waterman, Johan Hattne, Alun W. Ashton, Nathaniel Echols, Gwyndaf Evans, Nicholas K. Sauter, and Graeme Winter. 2014. “Dxtbx: The Diffraction Experiment Toolbox.” *Journal of Applied Crystallography*. https://doi.org/10.1107/s1600576714011996.

Parkhurst, James Michael. 2020. “Statistically Robust Methods for the Integration and Analysis of X-Ray Diffraction Data from Pixel Array Detectors.” University of Cambridge. https://doi.org/10.17863/CAM.46755.

Pearce, Nicholas M., Tobias Krojer, Anthony R. Bradley, Patrick Collins, Radosław P. Nowak, Romain Talon, Brian D. Marsden, et al. 2017. “A Multi-Crystal Method for Extracting Obscured Crystallographic States from Conventionally Uninterpretable Electron Density.” *Nature Communications* 8 (April): 15123.

Pérez, Serge, and Daniele de Sanctis. 2017. “Glycoscience@Synchrotron: Synchrotron Radiation Applied to Structural Glycoscience.” *Beilstein Journal of Organic Chemistry*.

https://doi.org/10.3762/bjoc.13.114.

Pérez, Serge, Anita Sarkar, Alain Rivet, Christelle Breton, and Anne Imberty. 2015. “Glyco3D: A Portal for Structural Glycosciences.” *Methods in Molecular Biology* 1273: 241–58.

Pérez, Serge, Thibault Tubiana, Anne Imberty, and Marc Baaden. 2015. “Three-Dimensional Representations of Complex Carbohydrates and Polysaccharides--SweetUnityMol: A Video Game-Based Computer Graphic Software.” *Glycobiology* 25 (5): 483–91.

Perrakis, A., R. Morris, and V. S. Lamzin. 1999. “Automated Protein Model Building Combined with Iterative Structure Refinement.” *Nature Structural Biology* 6 (5): 458–63.

Petrescu, Andrei-José, Mark R. Wormald, and Raymond A. Dwek. 2006. “Structural Aspects of Glycomes with a Focus on N-Glycosylation and Glycoprotein Folding.” *Current Opinion in Structural Biology* 16 (5): 600–607.

Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. 2004. “UCSF Chimera--a Visualization System for Exploratory Research and Analysis.” *Journal of Computational Chemistry* 25 (13): 1605–12.

Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Elaine C. Meng, Gregory S. Couch, Tristan I. Croll, John H. Morris, and Thomas E. Ferrin. 2021. “UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers.” *Protein Science: A Publication of the Protein Society* 30 (1): 70–82.

Pluvinage, Fillo, Massel, and Boraston. 2017. “The Quaternary Structure of Beta-1,3-Glucan Contributes to Its Recognition and Hydrolysis by a Multimodular Family 81 Glycoside Hydrolase.” *Structure* .

Potterton, Liz, Jon Agirre, Charles Ballard, Kevin Cowtan, Eleanor Dodson, Phil R. Evans, Huw T. Jenkins, et al. 2018a. “CCP4i2: The New Graphical User Interface to the CCP4 Program Suite.” *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 2): 68–84.

———. 2018b. “CCP4i2: The New Graphical User Interface to theCCP4 Program Suite.” *Acta Crystallographica. Section D, Structural Biology* 74 (2): 68–84.

Powell, Harold R. 2017. “X-Ray Data Processing.” *Bioscience Reports* 37 (5). https://doi.org/10.1042/BSR20170227.

Powell, Harold R., T. Geoff G. Battye, Luke Kontogiannis, Owen Johnson, and Andrew G. W.

Leslie. 2017. “Integrating Macromolecular X-Ray Diffraction Data with the Graphical User Interface iMosflm.” *Nature Protocols* 12 (7): 1310–25.

Punjani, Ali, and David J. Fleet. 2021. “3D Variability Analysis: Resolving Continuous Flexibility and Discrete Heterogeneity from Single Particle Cryo-EM.” *Journal of Structural Biology* 213 (2): 107702.

Ramirez-Escudero, Mercedes, Noa Miguez, Maria Gimeno-Perez, Antonio O. Ballesteros, Maria Fernandez-Lobato, Francisco J. Plou, and Julia Sanz-Aparicio. 2019. “Deciphering the Molecular Specificity of Phenolic Compounds as Inhibitors or Glycosyl Acceptors of β-Fructofuranosidase from Xanthophyllomyces Dendrorhous.” *Scientific Reports* 9 (1): 17441.

Rocha, Gerd B., Ricardo O. Freire, Alfredo M. Simas, and James J. P. Stewart. 2006. “RM1: A Reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I.” *Journal of Computational Chemistry* 27 (10): 1101–11.

Rochira, William, and Jon Agirre. 2021. “Iris: Interactive All-in-One Graphical Validation of 3D Protein Model Iterations.” *Protein Science: A Publication of the Protein Society* 30 (1): 93–107.

Rodríguez, Dayté D., Christian Grosse, Sebastian Himmel, César González, Iñaki M. de Ilarduya, Stefan Becker, George M. Sheldrick, and Isabel Usón. 2009. “Crystallographic Ab Initio Protein Structure Solution below Atomic Resolution.” *Nature Methods* 6 (9): 651–53.

Rudd, Pauline M., Mark R. Wormald, and Raymond A. Dwek. 2004. “Sugar-Mediated Ligand-Receptor Interactions in the Immune System.” *Trends in Biotechnology* 22 (10): 524–30.

Rudd, P. M., and R. A. Dwek. 1997. “Glycosylation: Heterogeneity and the 3D Structure of Proteins.” *Critical Reviews in Biochemistry and Molecular Biology* 32 (1): 1–100.

Rudman, Najda, Olga Gornik, and Gordan Lauc. 2019. “Altered N-Glycosylation Profiles as Potential Biomarkers and Drug Targets in Diabetes.” *FEBS Letters* 593 (13): 1598–1615.

Saenger, Wolfram. 1984. “Principles of Nucleic Acid Structure.” *Springer Advanced Texts in Chemistry*. https://doi.org/10.1007/978-1-4612-5190-3.

Salinger, M.T., Hobbs, J.R., Murray, J.W., Laver, W.G., Kuhn, P., Garman, E.F. n.d. “High

Resolution Structures of Viral Neuraminidase with Drugs Bound in the Active Site. (In Preparation)."

Sammito, A. J. McCoy, K. Hatti, R. D. Oeffner, D. H. Stockwell, T. I. Croll, R. J. Read, and IUCr. 2019. "Phaser.Voyager: Data-Guided Model Generation and Visualization." *Acta Crystallographica Section A: Foundations and Advances* 75 (August): e182–e182.

Scherbinina, Sofya I., and Philip V. Toukach. 2020. "Three-Dimensional Structures of Carbohydrates and Where to Find Them." *International Journal of Molecular Sciences* 21 (20). https://doi.org/10.3390/ijms21207702.

Scheres, Sjors H. W. 2012. "RELION: Implementation of a Bayesian Approach to Cryo-EM Structure Determination." *Journal of Structural Biology* 180 (3): 519–30.

Schnaar, Ronald L. 2016. "Glycobiology Simplified: Diverse Roles of Glycan Recognition in Inflammation." *Journal of Leukocyte Biology* 99 (6): 825–38.

Schumann, Benjamin, Stacy A. Malaker, Simon P. Wisnovsky, Marjoke F. Debets, Anthony J. Agbay, Daniel Fernandez, Lauren J. S. Wagner, et al. 2019. "Chemical Precision Glyco-Mutagenesis by Glycosyltransferase Engineering in Living Cells." *bioRxiv*. https://doi.org/10.1101/669861.

Sehnal, David, and Oliver C. Grant. 2019. "Rapidly Display Glycan Symbols in 3D Structures: 3D-SNFG in LiteMol." *Journal of Proteome Research* 18 (2): 770–74.

She, Ji, Zhifu Han, Bin Zhou, and Jijie Chai. 2013. "Structural Basis for Differential Recognition of Brassinolide by Its Receptors." *Protein & Cell* 4 (6): 475–82.

Sheldrick, George M. 2008a. "A Short History ofSHELX." *Acta Crystallographica Section A Foundations of Crystallography*. https://doi.org/10.1107/s0108767307043930.

———. 2008b. "A Short History of SHELX." *Acta Crystallographica. Section A, Crystal Physics, Diffraction, Theoretical and General Crystallography* 64 (Pt 1): 112–22.

Shelley, Kathryn L., Thomas P. E. Dixon, Jonathan C. Brooks-Bartlett, and Elspeth F. Garman. 2018. ": Quantifying Specific Radiation Damage in Individual Protein Crystal Structures." *Journal of Applied Crystallography* 51 (Pt 2): 552–59.

Shelley, Kathryn L., and Elspeth F. Garman. 2022. "Quantifying and Comparing Radiation Damage in the Protein Data Bank." *Nature Communications* 13 (1): 1314.

Shental-Bechor, Dalit, and Yaakov Levy. 2009. "Folding of Glycoproteins: Toward

Understanding the Biophysics of the Glycosylation Code." *Curr. Opin. Struct. Biol.* 19 (5): 524–33.

Simpkin, Adam J., Luc G. Elliott, Kyle Stevenson, Eugene Krissinel, Daniel J. Rigden, and Ronan M. Keegan. 2022. "Slice'N'Dice: Maximising the Value of Predicted Models for Structural Biologists." *bioRxiv*. https://doi.org/10.1101/2022.06.30.497974.

Simpkin, Adam J., Felix Simkovic, Jens M. H. Thomas, Martin Savko, Andrey Lebedev, Ville Uski, Charles C. Ballard, et al. 2020. "Using Phaser and Ensembles to Improve the Performance of SIMBAD." *Acta Crystallographica. Section D, Structural Biology* 76 (Pt 1): 1–8.

Simpkin, Adam J., Jens M. H. Thomas, Ronan M. Keegan, and Daniel J. Rigden. 2022. "MrParse: Finding Homologues in the PDB and the EBI AlphaFold Database for Molecular Replacement and More." *Acta Crystallographica. Section D, Structural Biology* 78 (Pt 5): 553–59.

Simpkin, Adam, Felix Simkovic, Jens Thomas, Martin Savko, Charles Ballard, Marcin Wojdyr, William Shepard, Daniel Rigden, and Ronan Keegan. 2018. "Identification of Contaminants with SIMBAD: A Sequence-Independent Molecular Replacement Pipeline." *Acta Crystallographica Section A Foundations and Advances*. https://doi.org/10.1107/s2053273318092641.

Skubák, Pavol, and Navraj S. Pannu. 2013. "Automatic Protein Structure Solution from Weak X-Ray Data." *Nature Communications* 4: 2777.

Smock, Robert G., and Rob Meijers. 2018. "Roles of Glycosaminoglycans as Regulators of Ligand/receptor Complexes." *Open Biology* 8 (10). https://doi.org/10.1098/rsob.180026.

Sobolev, Oleg V., Pavel V. Afonine, Nigel W. Moriarty, Maarten L. Hekkelman, Robbie P. Joosten, Anastassis Perrakis, and Paul D. Adams. 2020. "A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry." *Structure* 28 (11): 1249–58.e2.

Söding, Johannes. 2005. "Protein Homology Detection by HMM-HMM Comparison." *Bioinformatics* 21 (7): 951–60.

Stagnoli, Soledad, Francesca Peccati, Sean R. Connell, Ane Martinez-Castillo, Diego Charro, Oscar Millet, Chiara Bruzzone, et al. 2022. "Assessing the Mobility of Severe Acute Respiratory Syndrome Coronavirus-2 Spike Protein Glycans by Structural and Computational Methods." *Frontiers in Microbiology* 13 (April): 870938.

Stanley, Pamela. 2016. “What Have We Learned from Glycosyltransferase Knockouts in Mice?” *Journal of Molecular Biology* 428 (16): 3166–82.

Steiner, Roberto A., Andrey A. Lebedev, and Garib N. Murshudov. 2003. “Fisher’s Information in Maximum-Likelihood Macromolecular Crystallographic Refinement.” *Acta Crystallographica. Section D, Biological Crystallography* 59 (Pt 12): 2114–24.

Suzuki, Kentaro, Jun-Ichi Sumitani, Young-Woo Nam, Toru Nishimaki, Shuji Tani, Takayoshi Wakagi, Takashi Kawaguchi, and Shinya Fushinobu. 2013. “Crystal Structures of Glycoside Hydrolase Family 3 β-Glucosidase 1 from Aspergillus Aculeatus.” *Biochemical Journal* 452 (2): 211–21.

Terwilliger, Thomas C., Frank Dimaio, Randy J. Read, David Baker, Gábor Bunkóczi, Paul D. Adams, Ralf W. Grosse-Kunstleve, Pavel V. Afonine, and Nathaniel Echols. 2012. “Phenix.mr_rosetta: Molecular Replacement and Model Rebuilding with Phenix and Rosetta.” *Journal of Structural and Functional Genomics* 13 (2): 81–90.

Terwilliger, Thomas C., Ralf W. Grosse-Kunstleve, Pavel V. Afonine, Nigel W. Moriarty, Peter H. Zwart, Li Wei Hung, Randy J. Read, and Paul D. Adams. 2007. “Iterative Model Building, Structure Refinement and Density Modification with the PHENIX AutoBuild Wizard.” In *Acta Crystallographica Section D: Biological Crystallography*, 64:61–69.

Thieker, David F., Jodi A. Hadden, Klaus Schulten, and Robert J. Woods. 2016. “3D Implementation of the Symbol Nomenclature for Graphical Representation of Glycans.” *Glycobiology* 26 (8): 786–87.

Thorn, Andrea, James Parkhurst, Paul Emsley, Robert A. Nicholls, Melanie Vollmar, Gwyndaf Evans, and Garib N. Murshudov. 2017. “AUSPEX: A Graphical Tool for X-Ray Diffraction Data Analysis.” *Acta Crystallographica. Section D, Structural Biology* 73 (Pt 9): 729–37.

Thorn, Andrea, and George M. Sheldrick. 2013. “Extending Molecular-Replacement Solutions with SHELXE.” *Acta Crystallographica. Section D, Biological Crystallography* 69 (Pt 11): 2251–56.

Tickle, Ian J. 2012. “Statistical Quality Indicators for Electron-Density Maps.” *Acta Crystallographica. Section D, Biological Crystallography* 68 (Pt 4): 454–67.

Tobola, Felix, Mickael Lelimousin, Annabelle Varrot, Emilie Gillon, Barbara Darnhofer, Ola Blixt, Ruth Birner-Gruenberger, Anne Imberty, and Birgit Wiltschi. 2018. “Effect of Noncanonical Amino Acids on Protein-Carbohydrate Interactions: Structure, Dynamics,

and Carbohydrate Affinity of a Lectin Engineered with Fluorinated Tryptophan Analogs." *ACS Chemical Biology* 13 (8): 2211–19.

Touw, Wouter G., Bart van Beusekom, Jochem M. G. Evers, Gert Vriend, and Robbie P. Joosten. 2016. "Validation and Correction of Zn-CysHis Complexes." *Acta Crystallographica. Section D, Structural Biology* 72 (Pt 10): 1110–18.

Trapani, Stefano, and Jorge Navaza. 2008. "AMoRe: Classical and Modern." *Acta Crystallographica. Section D, Biological Crystallography* 64 (Pt 1): 11–16.

Tunyasuvunakool, Kathryn, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, et al. 2021. "Highly Accurate Protein Structure Prediction for the Human Proteome." *Nature* 596 (7873): 590–96.

Uervirojnangkoorn, Monarin, Oliver B. Zeldin, Artem Y. Lyubimov, Johan Hattne, Aaron S. Brewster, Nicholas K. Sauter, Axel T. Brunger, and William I. Weis. 2015. "Enabling X-Ray Free Electron Laser Crystallography for Challenging Biological Systems from a Limited Number of Crystals." *eLife* 4 (March). https://doi.org/10.7554/eLife.05421.

Usón, Isabel, and George M. Sheldrick. 2018. "An Introduction to Experimental Phasing of Macromolecules Illustrated by SHELX; New Autotracing Features." *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 2): 106–16.

Vagin, Alexei A., Roberto A. Steiner, Andrey A. Lebedev, Liz Potterton, Stuart McNicholas, Fei Long, and Garib N. Murshudov. 2004a. "REFMAC5 Dictionary: Organization of Prior Chemical Knowledge and Guidelines for Its Use." *Acta Crystallographica. Section D, Biological Crystallography* 60 (12): 2184–95.

———. 2004b. "REFMAC5 Dictionary: Organization of Prior Chemical Knowledge and Guidelines for Its Use." *Acta Crystallographica. Section D, Biological Crystallography* 60 (Pt 12 Pt 1): 2184–95.

Vagin, Alexei, and Alexei Teplyakov. 2010. "Molecular Replacement with MOLREP." *Acta Crystallographica. Section D, Biological Crystallography* 66 (Pt 1): 22–25.

Vagin, Alexey, and Andrey Lebedev. 2015. "*MoRDa*, an Automatic Molecular Replacement Pipeline." *Acta Crystallographica Section A Foundations and Advances*. https://doi.org/10.1107/s2053273315099672.

Valverde, Pablo, Jon I. Quintana, Jose I. Santos, Ana Ardá, and Jesús Jiménez-Barbero. 2019. "Novel NMR Avenues to Explore the Conformation and Interactions of Glycans."

ACS Omega 4 (9): 13618–30.

Varadi, Mihaly, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, et al. 2022. “AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models.” *Nucleic Acids Research* 50 (D1): D439–44.

Varki, Ajit, Richard D. Cummings, Markus Aebi, Nicole H. Packer, Peter H. Seeberger, Jeffrey D. Esko, Pamela Stanley, et al. 2015. “Symbol Nomenclature for Graphical Representations of Glycans.” *Glycobiology* 25 (12): 1323–24.

Varki, Ajit, Richard D. Cummings, Jeffrey D. Esko, Hudson H. Freeze, Pamela Stanley, Jamey D. Marth, Carolyn R. Bertozzi, Gerald W. Hart, and Marilynn E. Etzler. 2009. “Symbol Nomenclature for Glycan Representation.” *Proteomics*. https://doi.org/10.1002/pmic.200900708.

Walls, Alexandra C., Xiaoli Xiong, Young-Jun Park, M. Alejandra Tortorici, Joost Snijder, Joel Quispe, Elisabetta Cameroni, et al. 2019. “Unexpected Receptor Functional Mimicry Elucidates Activation of Coronavirus Fusion.” *Cell* 176 (5): 1026–39.e15.

Wang, Bi-Cheng. 1985. “Resolution of Phase Ambiguity in Macromolecular Crystallography.” In *Methods in Enzymology*, 115:90–112. Academic Press.

Weininger, David. 1988. “SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules.” *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/ci00057a005.

Westbrook, John D., Chenghua Shao, Zukang Feng, Marina Zhuravleva, Sameer Velankar, and Jasmine Young. 2015. “The Chemical Component Dictionary: Complete Descriptions of Constituent Molecules in Experimentally Determined 3D Macromolecules in the Protein Data Bank.” *Bioinformatics* 31 (8): 1274–78.

Williams, Christopher J., Jeffrey J. Headd, Nigel W. Moriarty, Michael G. Prisant, Lizbeth L. Videau, Lindsay N. Deis, Vishal Verma, et al. 2018. “MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation.” *Protein Science*. https://doi.org/10.1002/pro.3330.

Winn, Martyn D., Charles C. Ballard, Kevin D. Cowtan, Eleanor J. Dodson, Paul Emsley, Phil R. Evans, Ronan M. Keegan, et al. 2011a. “Overview of the CCP4 Suite and Current Developments.” *Acta Crystallographica Section D: Biological Crystallography*. https://doi.org/10.1107/S0907444910045749.

———. 2011b. “Overview of the CCP4 Suite and Current Developments.” *Acta Crystallographica. Section D, Biological Crystallography* 67 (Pt 4): 235–42.

Winter, G. 2010. “xia2: An Expert System for Macromolecular Crystallography Data Reduction.” *Journal of Applied Crystallography*. https://doi.org/10.1107/s0021889809045701.

Winter, Graeme, Carina M. C. Lobley, and Stephen M. Prince. 2013. “Decision Making in xia2.” *Acta Crystallographica. Section D, Biological Crystallography* 69 (Pt 7): 1260–73.

Winter, Graeme, David G. Waterman, James M. Parkhurst, Aaron S. Brewster, Richard J. Gildea, Markus Gerstel, Luis Fuentes-Montero, et al. 2018. “DIALS: Implementation and Evaluation of a New Integration Package.” *Acta Crystallographica. Section D, Structural Biology* 74 (Pt 2): 85–97.

Wojdyr, Marcin. 2022. “GEMMI: A Library for Structural Biology.” *Journal of Open Source Software* 7 (73): 4200.

Wojdyr, Marcin, Ronan Keegan, Graeme Winter, and Alun Ashton. 2013. “*DIMPLE*- a Pipeline for the Rapid Generation of Difference Maps from Protein Crystals with Putatively Bound Ligands.” *Acta Crystallographica Section A Foundations of Crystallography*. https://doi.org/10.1107/s0108767313097419.

Wrapp, Daniel, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan. 2020. “Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation.” *Science* 367 (6483): 1260–63.

wwPDB consortium. 2019. “Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data.” *Nucleic Acids Research* 47 (D1): D520–28.

Xie, Peng, Craig Streu, Jie Qin, Howard Bregman, Nicholas Pagano, Eric Meggers, and Ronen Marmorstein. 2009. “The Crystal Structure of BRAF in Complex with an Organoruthenium Inhibitor Reveals a Mechanism for Inhibition of an Active Form of BRAF Kinase.” *Biochemistry* 48 (23): 5187–98.

Xin, Fuxiao, and Predrag Radivojac. 2012. “Post-Translational Modifications Induce Significant yet Not Extreme Changes to Protein Structure.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts541.

Xiong, Xiaoli, Haixia Xiao, Stephen R. Martin, Peter J. Coombs, Junfeng Liu, Patrick J.

Collins, Sebastien G. Vachieri, et al. 2014. “Enhanced Human Receptor Binding by H5 Haemagglutinins.” *Virology* 456-457 (100): 179–87.

Yamashita, K., C. M. Palmer, T. Burnley, and G. N. Murshudov. 2021. “Cryo-EM Single-Particle Structure Refinement and Map Calculation Using Servalcat.” *Acta Crystallographica Section D: Structural Biology* 77 (10): 1282–91.

Yang, Yun, Vidya C. Darbari, Nan Zhang, Duo Lu, Robert Glyde, Yi-Ping Wang, Jared T. Winkelman, et al. 2015. “TRANSCRIPTION. Structures of the RNA Polymerase-σ54 Reveal New and Conserved Regulatory Strategies.” *Science* 349 (6250): 882–85.

Yuriev, Elizabeth, and Paul A. Ramsland. 2015. “Carbohydrates in Cyberspace.” *Frontiers in Immunology* 6 (June): 300.

Ząbczyńska, Marta, Kamila Kozłowska, and Ewa Pocheć. 2018. “Glycosylation in the Thyroid Gland: Vital Aspects of Glycoprotein Function in Thyrocyte Physiology and Thyroid Disorders.” *International Journal of Molecular Sciences* 19 (9): 2792.

Zhang, Bo, Kai Biao Wang, Wen Wang, Xin Wang, Fang Liu, Jiapeng Zhu, Jing Shi, et al. 2019. “Enzyme-Catalysed [6+4] Cycloadditions in the Biosynthesis of Natural Products.” *Nature* 568 (7750): 122–26.

Zhang, K. Y., K. Cowtan, and P. Main. 1997. “Combining Constraints for Electron-Density Modification.” *Methods in Enzymology* 277: 53–64.

Zhong, Ellen D., Tristan Bepler, Bonnie Berger, and Joseph H. Davis. 2021. “CryoDRGN: Reconstruction of Heterogeneous Cryo-EM Structures Using Neural Networks.” *Nature Methods* 18 (2): 176–85.

Zhu, Shaotong, Colleen M. Noviello, Jinfeng Teng, Richard M. Walsh, Jeong Joo Kim, and Ryan E. Hibbs. 2018. “Structure of a Human Synaptic GABAA Receptor.” *Nature*. https://doi.org/10.1038/s41586-018-0255-3.