BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Chemical glycobiology

Edited by Elisa Fadda, Rachel Hevey,Benjamin Schumannand Ulrika Westerlind

Imprint

Beilstein Journal of Organic Chemistry
www.bjoc.org
ISSN 1860-5397
Email: journals-support@beilstein-institut.de

The *Beilstein Journal of Organic Chemistry* is published by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften.

Beilstein-Institut zur Förderung der Chemischen Wissenschaften
Trakehner Straße 7–9
60487 Frankfurt am Main
Germany
www.beilstein-institut.de

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Chemical glycobiology

Elisa Fadda[*1], Rachel Hevey[*2], Benjamin Schumann[*3,4] and Ulrika Westerlind[*5]

Editorial

Open Access

Address:
[1]School of Biological Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom, [2]Department of Pharmaceutical Sciences, University of Basel, 4056 Basel, Switzerland, [3]Department of Chemistry, Imperial College London, London W12 0BZ, United Kingdom, [4]Chemical Glycobiology Laboratory, The Francis Crick Institute, London NW1 1AT, United Kingdom and [5]Department of Chemistry, Umeå University, 90736 Umeå, Sweden

Email:
Elisa Fadda[*] - elisa.fadda@soton.ac.uk;
Rachel Hevey[*] - rachel.hevey@unibas.ch;
Benjamin Schumann[*] - ben.schumann@crick.ac.uk;
Ulrika Westerlind[*] - ulrika.westerlind@umu.se

* Corresponding author

Beilstein J. Org. Chem. **2025,** *21,* 8–9.
https://doi.org/10.3762/bjoc.21.2

Received: 12 November 2024
Accepted: 29 November 2024
Published: 03 January 2025

This article is part of the thematic issue "Chemical glycobiology".

Guest Editors: E. Fadda, R. Hevey and B. Schumann; Associate Editor: U. Westerlind

As glycoscientists, we are standing on the shoulders of giants. Research on carbohydrates is as old as on any other biomolecule, dating back to the time of Emil Fischer and the elucidation of monosaccharide structures [1]. Later, foundational contributions came in the form of the first glycoconjugate vaccines [2,3], the elucidation of the blood group system [4], and many others. Among these, we dare to include the DNA double helix, featuring deoxyribose as a key structural element of its twisting ladder [5]. A century of innovation, some of the most prestigious awards and highest honours later, one aspect is immediately clear: chemistry and glycobiology are intricately intertwined. This is certainly by choice, but also by necessity. It is difficult to convey to non-glycoscientists how we still struggle with challenges that have been solved years or decades ago for proteins and nucleic acids. When molecular cloning and recombinant protein production became routine, these technologies were not applicable to glycans. Today, the most amazing tools in genome engineering are used to great effect to disrupt or alter the glycan biosynthetic machinery, but they still cannot be used to, for instance, mutate one glycan into another in the same manner as nucleic acids can be mutated. Methods in molecular biology are facile and quantitative. But they do not tell us the function of a particular glycoform on a specific glycoprotein. To put glycans on the map, chemists needed to be inventive.

At the time of writing this Editorial article, we are all early- and mid-career investigators who have learned from the best. We look in awe at the achievements in the field to date, some of those appearing in the previous thematic issues “GlycoBioinformatics” [6] and “Synthesis in the glycosciences” I and II [7,8]. We look ahead, asking the question how we can implement new chemistry, new molecules, and new methods to make the glycosciences even more palatable to generalists. And we see a field that innovates.

This thematic issue seeks to highlight the amazing breadth of contemporary chemical glycobiology. Dal Colle et al. investigate the determinants that influence the oligosaccharide yield in automated glycan assembly [9]. Target-directed synthetic strate-

gies are being developed by Reihill et al. [10] and Karak et al. [11], exploring the syntheses of the linker-displaying, sulfated TF disaccharide and lipid II analogues, respectively. The direct application of synthetic glycans is shown by Fan et al. [12] in the context of photoswitchable ligands to the lectin LecA. Staying in the theme of lectin characterization, Lundstrøm et al. study the glycan binding profile of CMA1 originating from melon [13].

A time that sees great opportunities in computational biology also breeds innovative applications in the glycosciences. A key aspect is the modelling of protein–glycan interactions. Marcisz et al. study the power of umbrella sampling in distinguishing the interactions between different glycosaminoglycans and their receptors [14]. Nieto-Fabregat et al. provide a detailed overview on computational methods that underlie modern glycobioinformatics approaches [15]. Validation of glycoprotein structure is an important aspect of contemporary structural biology, and Dialpuri et al. present the Privateer database to allow for facile quality control of such structures [16]. Finally, Barillot et al. bridge experimental and computational efforts, developing a neural-network-based approach for the interpretation of glycan structures from their vibrational fingerprints [17].

We anticipate that this diverse collection of reports across the entire spectrum of the chemical sciences cements the readers' understanding of chemistry as being a catalyst to more than a century of glycobiology, with a profound and exciting vision for the future.

Elisa Fadda, Rachel Hevey, Benjamin Schumann and Ulrika Westerlind

Southampton, Basel, London, Umeå, November 2024

ORCID® iDs

Elisa Fadda - https://orcid.org/0000-0002-2898-7770
Rachel Hevey - https://orcid.org/0000-0002-2649-3427
Ulrika Westerlind - https://orcid.org/0000-0002-4841-6238

Data Availability Statement

Data sharing is not applicable as no new data was generated or analyzed in this study.

References

1. Fischer, E. *Ber. Dtsch. Chem. Ges.* **1891,** *24,* 1836–1845. doi:10.1002/cber.189102401311
2. Goebel, W. F.; Avery, O. T. *J. Exp. Med.* **1929,** *50,* 521–531. doi:10.1084/jem.50.4.521
3. Avery, O. T.; Goebel, W. F. *J. Exp. Med.* **1929,** *50,* 533–550. doi:10.1084/jem.50.4.533
4. Morgan, W. T. J.; Watkins, W. M. *Glycoconjugate J.* **2000,** *17,* 501–530. doi:10.1023/a:1011014307683
5. Watson, J. D.; Crick, F. H. C. *Nature* **1953,** *171,* 737–738. doi:10.1038/171737a0
6. Aoki-Kinoshita, K. F.; Lisacek, F.; Karlsson, N.; Kolarich, D.; Packer, N. H. *Beilstein J. Org. Chem.* **2021,** *17,* 2726–2728. doi:10.3762/bjoc.17.184
7. Lindhorst, T. K. *Beilstein J. Org. Chem.* **2010,** *6,* No. 16. doi:10.3762/bjoc.6.16
8. Lindhorst, T. K. *Beilstein J. Org. Chem.* **2012,** *8,* 411–412. doi:10.3762/bjoc.8.45
9. Dal Colle, M. C. S.; Ricardo, M. G.; Hribernik, N.; Danglad-Flores, J.; Seeberger, P. H.; Delbianco, M. *Beilstein J. Org. Chem.* **2023,** *19,* 1015–1020. doi:10.3762/bjoc.19.77
10. Reihill, M.; Ma, H.; Bengtsson, D.; Oscarson, S. *Beilstein J. Org. Chem.* **2024,** *20,* 173–180. doi:10.3762/bjoc.20.17
11. Karak, M.; Cloonan, C. R.; Baker, B. R.; Cochrane, R. V. K.; Cochrane, S. A. *Beilstein J. Org. Chem.* **2024,** *20,* 220–227. doi:10.3762/bjoc.20.22
12. Fan, Y.; El Rhaz, A.; Maisonneuve, S.; Gillon, E.; Fatthalla, M.; Le Bideau, F.; Laurent, G.; Messaoudi, S.; Imberty, A.; Xie, J. *Beilstein J. Org. Chem.* **2024,** *20,* 1486–1496. doi:10.3762/bjoc.20.132
13. Lundstrøm, J.; Gillon, E.; Chazalet, V.; Kerekes, N.; Di Maio, A.; Feizi, T.; Liu, Y.; Varrot, A.; Bojar, D. *Beilstein J. Org. Chem.* **2024,** *20,* 306–320. doi:10.3762/bjoc.20.31
14. Marcisz, M.; Anila, S.; Gaardløs, M.; Zacharias, M.; Samsonov, S. A. *Beilstein J. Org. Chem.* **2023,** *19,* 1933–1946. doi:10.3762/bjoc.19.144
15. Nieto-Fabregat, F.; Lenza, M. P.; Marseglia, A.; Di Carluccio, C.; Molinaro, A.; Silipo, A.; Marchetti, R. *Beilstein J. Org. Chem.* **2024,** *20,* 2084–2107. doi:10.3762/bjoc.20.180
16. Dialpuri, J. S.; Bagdonas, H.; Schofield, L. C.; Pham, P. T.; Holland, L.; Agirre, J. *Beilstein J. Org. Chem.* **2024,** *20,* 931–939. doi:10.3762/bjoc.20.83
17. Barillot, T.; Schindler, B.; Moge, B.; Fadda, E.; Lépine, F.; Compagnon, I. *Beilstein J. Org. Chem.* **2023,** *19,* 1825–1831. doi:10.3762/bjoc.19.134

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Linker, loading, and reaction scale influence automated glycan assembly

Marlene C. S. Dal Colle[1,2], Manuel G. Ricardo[1], Nives Hribernik[1], José Danglad-Flores[1], Peter H. Seeberger[1,2] and Martina Delbianco[*1]

Letter

Open Access

Address:
[1]Department of Biomolecular Systems, Max Planck Institute of Colloids and Interfaces, Am Mühlenberg 1, 14476 Potsdam, Germany and [2]Department of Chemistry and Biochemistry, Freie Universität Berlin, Arnimallee 22, 14195 Berlin, Germany

Email:
Martina Delbianco[*] - martina.delbianco@mpikg.mpg.de

* Corresponding author

Keywords:
automated glycan assembly; photocleavable linker; polysaccharides; solid-phase synthesis

Beilstein J. Org. Chem. **2023,** *19,* 1015–1020.
https://doi.org/10.3762/bjoc.19.77

Received: 03 May 2023
Accepted: 28 June 2023
Published: 06 July 2023

This article is part of the thematic issue "Chemical glycobiology".

Guest Editor: R. Hevey

Abstract

Automated glycan assembly (AGA) affords collections of well-defined glycans in a short amount of time. We systematically analyzed how parameters connected to the solid support affect the AGA outcome for three different glycan sequences. We showed that, while loading and reaction scale did not significantly influence the AGA outcome, the chemical nature of the linker dramatically altered the isolated yields. We identified that the major determinants of AGA yields are cleavage from the solid support and post-AGA purification steps.

Introduction

Automated glycan assembly (AGA) is a solid-phase method that enables the rapid synthesis of complex oligo- and polysaccharides from protected monosaccharide building blocks (BBs) [1,2]. Iterative cycles of glycosylation, capping, and selective deprotection afford the support-bound glycan with a programmable sequence (Figure 1A). The protected glycan is then cleaved from the solid support and subjected to post-AGA deprotection steps to reveal the target glycan. AGA is mostly performed on cross-linked polystyrene resins equipped with photocleavable linkers [3], offering orthogonality to all the synthetic steps of the assembly, while selectively releasing the glycan at the end of the synthesis.

In recent years, the implementation of new synthetic strategies [4-7] as well as technological improvements [8,9] permitted access to highly complex carbohydrates [10]. Still, variations in yields are not always ascribable to the AGA process [11-16]. Dissimilar structures are assembled in high purity as indicated by HPLC analysis of the crude products, but isolated in relatively low yields. The optimization procedures are focused on glycan elongation (i.e., glycosylation and deprotection steps), whereas less attention is given to variables associated with the solid support [17]. In contrast, substantial knowledge exists on how loading [18], reaction scale [19], and linkers [20,21] affect the overall yield of solid phase peptide synthesis (SPPS). In the

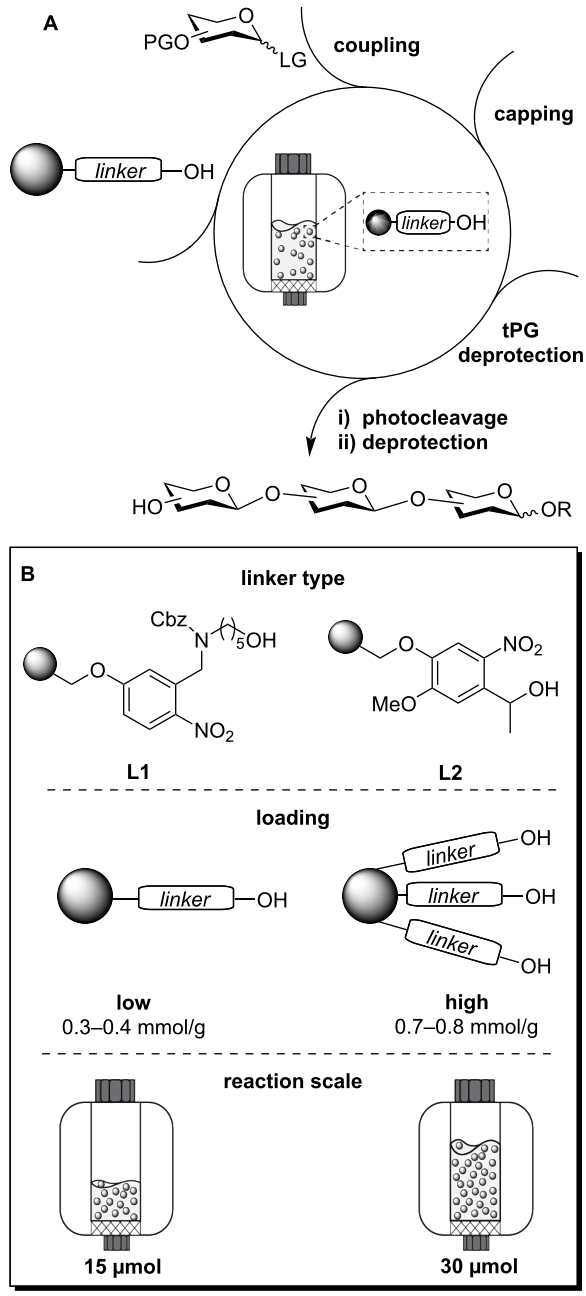


Figure 1: Schematic representation of the AGA process (A). Variables that can affect the AGA outcome investigated in this study (B).

past decades, several supports and linkers have been developed and commercialized for SPPS, enabling a wide range of applications. Solid supports are available with different linker loadings, with low loading (0.1–0.2 mmol/g) being beneficial to avoid aggregation of long peptide sequences, and high loadings (0.4–0.5 mmol/g) advantageous for more efficient syntheses [21].

Herein, we systematically investigate how variations in linker type, resin loading, and reaction scale influence the productivity of AGA.

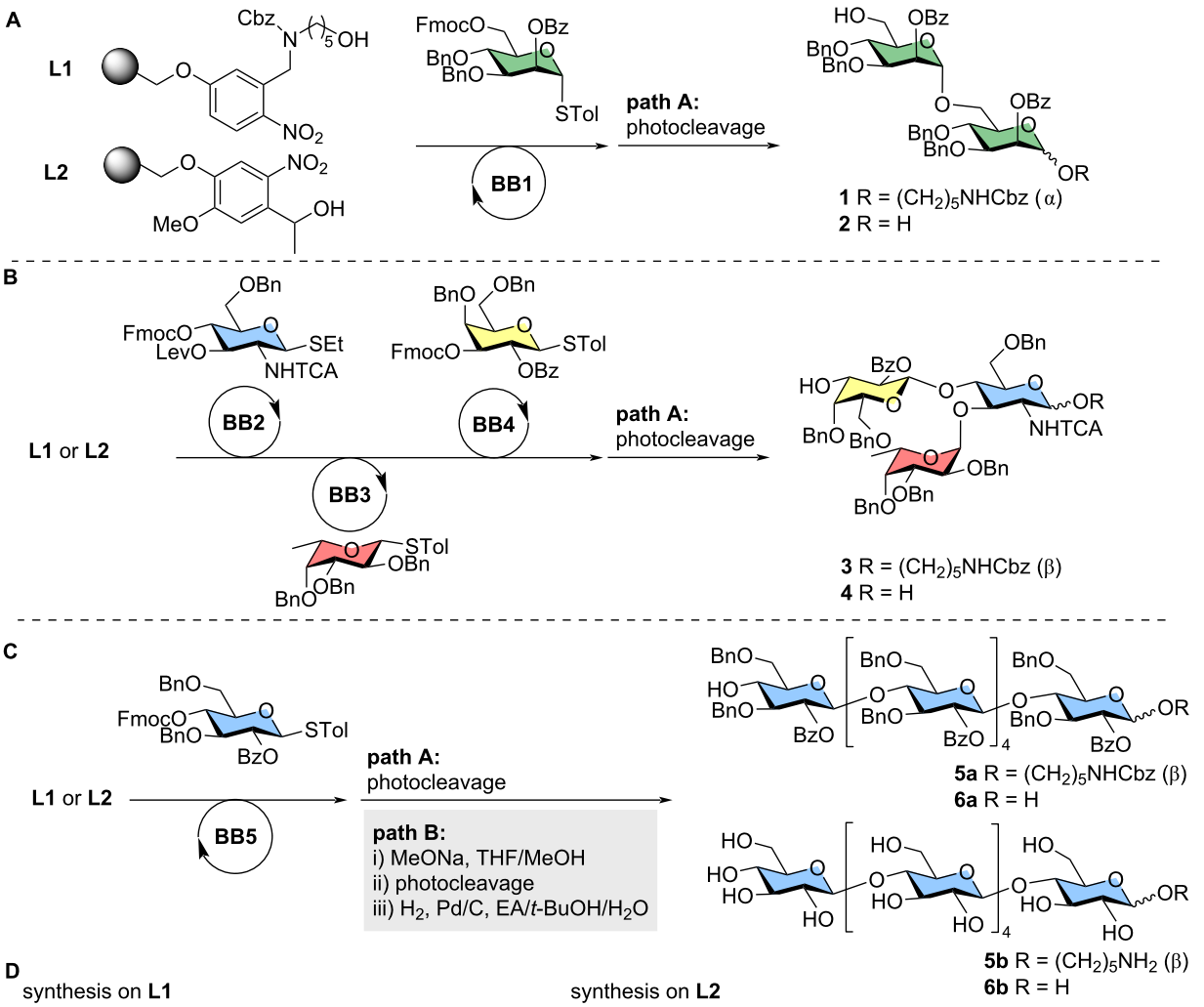
Results and Discussion

We selected three glycan sequences as models to analyze the effect of different parameters on the AGA outcome. Each sequence was prepared on four batches of Merrifield resin functionalized with two photolabile linkers (**L1** [22] vs **L2** [3]), at two linker loadings (low vs high) (Figure 1B). Each AGA experiment was performed at two different reaction scales (15 vs 30 µmol). All AGA runs were performed adjusting the resin amount to the desired reaction scale, while keeping the concentration of all other reagents constant (Figure 1B).

The photolabile linkers **L1** [22] and **L2** [3] are based on the *o*-nitrobenzyl scaffold [23,24] and expose a hydroxy group that serves as glycosyl acceptor in the first AGA cycle (Figure 1B). While **L1** displays a flexible aliphatic chain terminating with a primary alcohol, **L2** carries a secondary benzylic alcohol. Upon irradiation with UV light (λ = 360 nm), **L1** releases the glycan equipped with an aminoalkyl spacer at the reducing end, whereas **L2** affords the free reducing sugar (α/β mixture). Previous data suggested that UV cleavage of **L1** and **L2** was equally efficient, permitting the isolation of a tetramannoside in around 60% yield [3]. We wondered whether different glycan sequences were more sensitive to the linker structure. Less reactive donors might highlight differences in the linker nucleophilicity [25]. The aggregation of the growing glycan chains is conceivable to be connected to linker flexibility [18]. The efficiency of UV cleavage is probably influenced by glycan structure, solubility, and aggregation tendency [26]. Lastly, purification of the protected glycan upon cleavage could be affected by the presence or absence of a linker.

L1 or **L2** were conjugated to Merrifield resins with initial loadings of 0.5 mmol/g and 1.0 mmol/g to yield supports with low (0.3–0.4 mmol/g) or high (0.7–0.8 mmol/g) loadings (see Supporting Information File 1, section 2.3, module A). The latter allows for a larger synthesis scale, but steric hindrance and chain–chain interactions could negatively influence the AGA outcome, as observed for some peptide sequences [18]. Moreover, high-loading supports might result in inefficient UV cleavage due to quenching. These four supports were studied in AGA experiments performed at 15 and 30 µmol reaction scales. While AGA is commonly performed at a 15 µmol reaction scale, a larger reaction scale is attractive to produce more material in a single AGA run, but might suffer from insufficient mixing [27,28], causing slower kinetics [29], temperature gradients [30], and precipitation [31].

We set off to study the effect of these parameters on the AGA of three different glycan sequences (Figure 2). In an increasing order of complexity, we prepared α-1,6-linked dimannosides (**1**,**2**) [32], branched trisaccharides (**3**,**4**) [12], and linear α-1,4-linked hexaglucosides (**5**,**6**) [15,33]. Each synthesis was performed with 6.5 equivalents of BB per glycosylation cycle



D

synthesis on **L1**

Compound	Scale (mmol)	Loading (mmol/g)	HPLC purity	Yield (%) Path A	Yield (%) Path B
1	0.015	0.35	100	57	-
	0.030	0.35	100	49	-
	0.015	0.72	100	48	-
	0.030	0.72	100	59	-
3	0.015	0.35	77	14	-
	0.030	0.39	70	27	-
	0.015	0.72	87	32	-
	0.030	0.72	75	24	-
5	0.015	0.35	100	28	-
	0.030	0.39	100	36	15
	0.015	0.72	100	35	-
	0.030	0.72	97	21	-

synthesis on **L2**

Compound	Scale (mmol)	Loading (mmol/g)	HPLC purity*	Yield (%) Path A	Yield (%) Path B
2	0.015	0.35	100	51	-
	0.030	0.35	100	-	-
	0.015	0.72	100	-	-
	0.030	0.72	100	-	-
4	0.015	0.38	100	38	-
	0.030	0.38	89	29	-
	0.015	0.76	92	44	-
	0.030	0.76	84	36	-
6	0.015	0.38	100	-	57
	0.030	0.38	100	46	48
	0.015	0.76	100	-	44
	0.030	0.76	79	-	30

0 - 20
21 - 30
31 - 40
41 - 50
51 - 60
61 - 70
71 - 80
81 - 100

* no capped linker visible

Figure 2: AGA of model glycan sequences analyzed in this study: α-1,6-linked dimannosides **1**, **2** (A), branched trisaccharides **3**, **4** (B), and linear α-1,4-linked hexaglucosides **5**, **6** (C). Tables summarizing the results obtained for the AGA experiments performed in different conditions (D). The HPLC purity is estimated based on the ELSD profile. This value should be used to compare results within each series of experiments (i.e. same glycan sequence).

using previously reported AGA conditions (see Supporting Information File 1, section 2.3, module C). The outcome of each AGA experiment was analyzed in terms of: i) HPLC purity based on the chromatogram of the crude sample after AGA and UV cleavage, ii) isolated yield of the target compound after photocleavage and HPLC purification (path A).

The syntheses of the α-1,6-linked dimannosides **1** and **2** (Figure 2A) were successful on all resins tested, affording the desired product in complete purity regardless of linker type, loading or reaction scale (Figure 3A, and Figures S2 and S3 in Supporting Information File 1). Isolated yields of 49–59% were obtained in all experiments (Figure 2D), after cleavage of the photolabile unit.

The syntheses of the branched trisaccharides **3** and **4** (Figure 2B, and Figures S4 and S6 in Supporting Information File 1) were less efficient. Even though the target compound was the major product in all experiments, deletion sequences were observed in the chromatograms of the crudes (Figure 3B). MS analysis showed the presence of capped linker (*), capped dimer (†), and Lev-containing dimer (‡) (see Figures S5 and S7 in Supporting Information File 1). No significant variations were noticed within each series of experiments, with slightly better purities obtained for AGA performed on **L2** (to note: for experiments on **L2** no capped linker was detectable by HPLC; see Supporting Information File 1). Isolated yields were relatively low for all experiments (14–32% on **L1** and 29–44% on **L2**, Figure 2D). These values are quite low even considering the presence of deletion sequences, suggesting that cleavage and purification are more challenging for these structures. Overall, a slightly better performance of **L2** resulting in higher purities and better yields was noticed.

HPLC analysis showed that the β-1,4-hexaglucosides **5a** and **6a** were produced in excellent purity in all experiments (Figures 2D, 3C, and Figures S8 and S9 in Supporting Information File 1). For these compounds, we explored two different post-AGA procedures: the standard path A based on photocleavage and HPLC purification, and path B involving on resin methanolysis of the ester groups, photocleavage, hydrogenolysis of the remaining PGs, and purification (Figure 2C). The latter is commonly employed for compounds synthesized on **L2** because of the poor stability of free-reducing glycans in basic conditions needed for the methanolysis step [33].

The isolated yields of the fully protected compound **5a** synthesized on **L1** were significantly lower than expected (21–36%, Figure 2D, path A), with little variation within the series. Isolated yields for the linker-free compound **6a** prepared on **L2** were around 10% higher (46%). The absence of deletion sequences in the HPLC of the crude compounds indicated that cleavage and/or purification are the major bottlenecks of these syntheses.

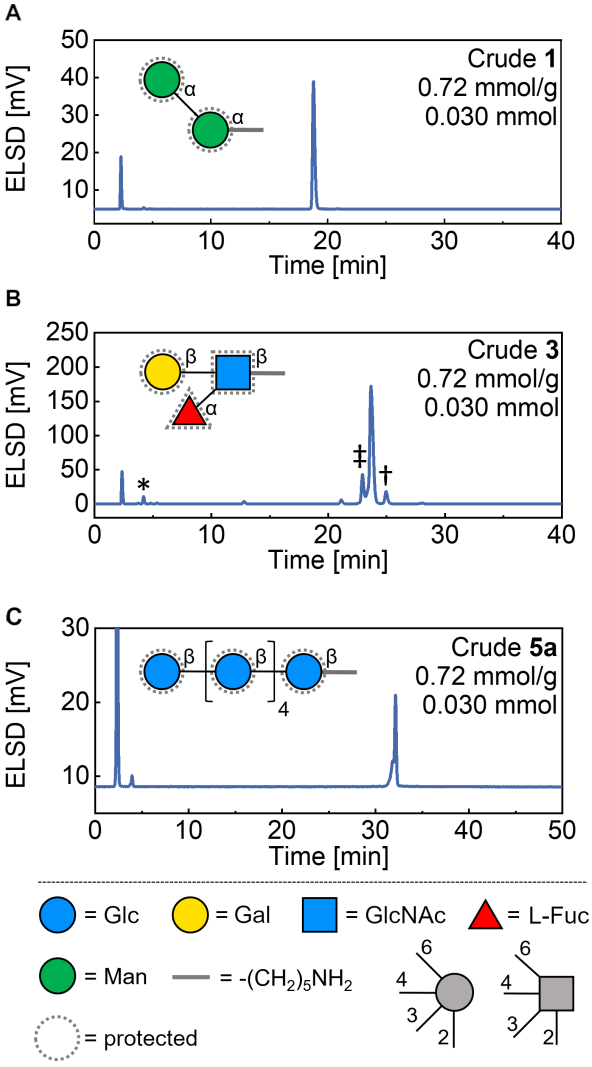


Figure 3: Representative HPLC traces for the crude compounds **1** (A), **3** (B), and **5a** (C) after cleavage from the solid support. HPLC conditions are reported in Supporting Information File 1 (Figures S2, S4, S8, A). In B, MS analysis showed the presence of capped linker (*), capped dimer (†), and Lev-containing dimer (‡). The monosaccharides are represented following the symbol nomenclature for glycans (SNFG).

Higher yields (30–57%) were obtained for compound **6b**, isolated after the post-AGA procedure path B (Figure 2D). This is surprising since the path B procedure involved additional deprotection steps. Therefore, we wondered whether methanoloysis on resin could improve photocleavage efficiency. However, when we tested the same procedure on **L1**, target compound **5b** was isolated in only 15% yield. These results strongly suggest that the two linkers perform differently depending on the glycan sequences.

Conclusion

Taken together, the results showed minimal variation within each series of experiments, indicating that loading and reaction scale are not significantly affecting AGA of those sequences within the range of conditions explored here. This is a promising observation from the perspective of scaling up AGA. No differences were observed for the AGA of simple disaccharides **1**,**2** performed on **L1** and **L2** with an apparent maximal yield of around 60%, in agreement with previous reports [3]. In contrast, other sequences constructed on **L2** were isolated in slightly better yields. This result could be connected to more efficient cleavage of **L2** in the presence of complex glycan sequences, easier purification of linker-free compounds, or a combination of both.

Our systematic study identified that the major determinants of AGA yields are cleavage from the solid support and purification steps. These two aspects are strongly connected to the glycan structure, with minimal variations such as presence or absence of a linker playing an important role in the post-AGA process. In some cases, performing post-AGA manipulations on resin dramatically improved the overall yield of the process. Future efforts need to focus on the development of new linkers, more efficient cleaving procedures [34], and the implementation of post-AGA manipulation steps on resin.

Supporting Information

Supporting Information File 1

Experimental procedures and characterization data.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-19-77-S1.pdf]

Funding

We thank the Max Planck Society, the German Federal Ministry of Education and Research (BMBF, grant number 13XP5114), and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation – SFB 1449 – 431232613; sub-project C2) for generous financial support.

Conflict of interest statement

The authors declare no conflict of interest.

ORCID® iDs

Marlene C. S. Dal Colle - https://orcid.org/0009-0006-1062-0395
Manuel G. Ricardo - https://orcid.org/0000-0003-2365-2864
Nives Hribernik - https://orcid.org/0000-0003-1347-3192
Peter H. Seeberger - https://orcid.org/0000-0003-3394-8466
Martina Delbianco - https://orcid.org/0000-0002-4580-9597

References

1. Huang, J. Y.; Delbianco, M. *Synthesis* **2023,** *55,* 1337–1354. doi:10.1055/a-1938-2293
2. Guberman, M.; Seeberger, P. H. *J. Am. Chem. Soc.* **2019,** *141,* 5581–5592. doi:10.1021/jacs.9b00638
3. Le Mai Hoang, K.; Pardo-Vargas, A.; Zhu, Y.; Yu, Y.; Loria, M.; Delbianco, M.; Seeberger, P. H. *J. Am. Chem. Soc.* **2019,** *141,* 9079–9086. doi:10.1021/jacs.9b03769
4. Sletten, E. T.; Danglad-Flores, J.; Leichnitz, S.; Abragam Joseph, A.; Seeberger, P. H. *Carbohydr. Res.* **2022,** *511,* 108489. doi:10.1016/j.carres.2021.108489
5. Tyrikos-Ergas, T.; Sletten, E. T.; Huang, J.-Y.; Seeberger, P. H.; Delbianco, M. *Chem. Sci.* **2022,** *13,* 2115–2120. doi:10.1039/d1sc06063e
6. Yu, Y.; Kononov, A.; Delbianco, M.; Seeberger, P. H. *Chem. – Eur. J.* **2018,** *24,* 6075–6078. doi:10.1002/chem.201801023
7. Zhu, Y.; Delbianco, M.; Seeberger, P. H. *J. Am. Chem. Soc.* **2021,** *143,* 9758–9768. doi:10.1021/jacs.1c02188
8. Panza, M.; Stine, K. J.; Demchenko, A. V. *Chem. Commun.* **2020,** *56,* 1333–1336. doi:10.1039/c9cc08876h
9. Danglad-Flores, J.; Leichnitz, S.; Sletten, E. T.; Abragam Joseph, A.; Bienert, K.; Le Mai Hoang, K.; Seeberger, P. H. *J. Am. Chem. Soc.* **2021,** *143,* 8893–8901. doi:10.1021/jacs.1c03851
10. Joseph, A. A.; Pardo-Vargas, A.; Seeberger, P. H. *J. Am. Chem. Soc.* **2020,** *142,* 8561–8564. doi:10.1021/jacs.0c00751
11. Fittolani, G.; Shanina, E.; Guberman, M.; Seeberger, P. H.; Rademacher, C.; Delbianco, M. *Angew. Chem., Int. Ed.* **2021,** *60,* 13302–13309. doi:10.1002/anie.202102690
12. Guberman, M.; Bräutigam, M.; Seeberger, P. H. *Chem. Sci.* **2019,** *10,* 5634–5640. doi:10.1039/c9sc00768g
13. Ricardo, M. G.; Reuber, E. E.; Yao, L.; Danglad-Flores, J.; Delbianco, M.; Seeberger, P. H. *J. Am. Chem. Soc.* **2022,** *144,* 18429–18434. doi:10.1021/jacs.2c06882
14. Chaube, M. A.; Trattnig, N.; Lee, D.-H.; Belkhadir, Y.; Pfrengle, F. *Eur. J. Org. Chem.* **2022,** e202200313. doi:10.1002/ejoc.202200313
15. Delbianco, M.; Kononov, A.; Poveda, A.; Yu, Y.; Diercks, T.; Jiménez-Barbero, J.; Seeberger, P. H. *J. Am. Chem. Soc.* **2018,** *140,* 5421–5426. doi:10.1021/jacs.8b00254
16. Pardo-Vargas, A.; Bharate, P.; Delbianco, M.; Seeberger, P. H. *Beilstein J. Org. Chem.* **2019,** *15,* 2936–2940. doi:10.3762/bjoc.15.288
17. Panza, M.; Neupane, D.; Stine, K. J.; Demchenko, A. V. *Chem. Commun.* **2020,** *56,* 10568–10571. doi:10.1039/d0cc03885g
18. Nakaie, C. R.; Oliveira, E.; Vicente, E. F.; Jubilut, G. N.; Souza, S. E. G.; Marchetto, R.; Cilli, E. M. *Bioorg. Chem.* **2011,** *39,* 101–109. doi:10.1016/j.bioorg.2011.01.001
19. Edelstein, M.; Scott, P. E.; Sherlund, M.; Hansen, A. L.; Hughes, J. L. *Chem. Eng. Sci.* **1986,** *41,* 617–624. doi:10.1016/0009-2509(86)87138-x
20. Shelton, P. T.; Jensen, K. J. *Methods Mol. Biol. (N. Y., NY, U. S.)* **2013,** *1047,* 23–41. doi:10.1007/978-1-62703-544-6_2
21. Moss, J. A. *Curr. Protoc. Protein Sci.* **2005,** *40,* 18.7.1–18.7.19. doi:10.1002/0471140864.ps1807s40
22. Eller, S.; Collot, M.; Yin, J.; Hahm, H. S.; Seeberger, P. H. *Angew. Chem., Int. Ed.* **2013,** *52,* 5858–5861. doi:10.1002/anie.201210132
23. Rich, D. H.; Gurwara, S. K. *J. Chem. Soc., Chem. Commun.* **1973,** 610–611. doi:10.1039/c39730000610
24. Guillier, F.; Orain, D.; Bradley, M. *Chem. Rev.* **2000,** *100,* 2091–2158. doi:10.1021/cr980040+

25. van der Vorm, S.; Hansen, T.; van Hengst, J. M. A.; Overkleeft, H. S.; van der Marel, G. A.; Codée, J. D. C. *Chem. Soc. Rev.* **2019,** *48,* 4688–4706. doi:10.1039/c8cs00369f
26. Holmes, C. P. *J. Org. Chem.* **1997,** *62,* 2370–2380. doi:10.1021/jo961602x
27. Kraume, M. *Chem. Eng. Technol.* **1992,** *15,* 313–318. doi:10.1002/ceat.270150505
28. Deen, N. G.; Mudde, R. F.; Kuipers, J. A. M.; Zehner, P.; Kraume, M. Bubble Columns. *Ullmann's Encyclopedia of Industrial Chemistry;* Wiley-VCH Verlag GmbH: Weinheim, Germany, 2010. doi:10.1002/14356007.b04_275.pub2
29. D'Ercole, A.; Pacini, L.; Sabatino, G.; Zini, M.; Nuti, F.; Ribecai, A.; Paio, A.; Rovero, P.; Papini, A. M. *Org. Process Res. Dev.* **2021,** *25,* 2754–2771. doi:10.1021/acs.oprd.1c00368
30. Collins, J. M. Solid Phase Peptide Synthesis. U.S. Patent 10,125,163, Nov 13, 2018.
31. Bray, B. L. *Nat. Rev. Drug Discovery* **2003,** *2,* 587–593. doi:10.1038/nrd1133
32. Calin, O.; Eller, S.; Seeberger, P. H. *Angew. Chem., Int. Ed.* **2013,** *52,* 5862–5865. doi:10.1002/anie.201210176
33. Yu, Y.; Tyrikos-Ergas, T.; Zhu, Y.; Fittolani, G.; Bordoni, V.; Singhal, A.; Fair, R. J.; Grafmüller, A.; Seeberger, P. H.; Delbianco, M. *Angew. Chem., Int. Ed.* **2019,** *58,* 13127–13132. doi:10.1002/anie.201906577
34. Bakhatan, Y.; Alshanski, I.; Grunhaus, D.; Hurevich, M. *Org. Biomol. Chem.* **2020,** *18,* 4183–4188. doi:10.1039/d0ob00821d

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

GlAIcomics: a deep neural network classifier for spectroscopy-augmented mass spectrometric glycans data

Thomas Barillot[1], Baptiste Schindler[1], Baptiste Moge[1], Elisa Fadda[2], Franck Lépine[1] and Isabelle Compagnon[*1]

Full Research Paper

Open Access

Address:
[1]Univ Claude Bernard Lyon 1, CNRS, Institut Lumière Matière, F-69622 Villeurbanne, France and [2]Department of Chemistry and Hamilton Institute, Maynooth University, Maynooth W23 F2H6, Ireland

Email:
Isabelle Compagnon[*] - isabelle.compagnon@univ-lyon1.fr

* Corresponding author

Keywords:
Bayesian neural network; deep learning; glycomics; IR; spectroscopy

Beilstein J. Org. Chem. **2023,** *19,* 1825–1831.
https://doi.org/10.3762/bjoc.19.134

Received: 27 March 2023
Accepted: 29 September 2023
Published: 05 December 2023

This article is part of the thematic issue "Chemical glycobiology".

Associate Editor: P. Schreiner

Abstract

Carbohydrate sequencing is a formidable task identified as a strategic goal in modern biochemistry. It relies on identifying a large number of isomers and their connectivity with high accuracy. Recently, gas phase vibrational laser spectroscopy combined with mass spectrometry tools have been proposed as a very promising sequencing approach. However, its use as a generic analytical tool relies on the development of recognition techniques that can analyse complex vibrational fingerprints for a large number of monomers. In this study, we used a Bayesian deep neural network model to automatically identify and classify vibrational fingerprints of several monosaccharides. We report high performances of the obtained trained algorithm (GlAIcomics), that can be used to discriminate contamination and identify a molecule with a high degree of confidence. It opens the possibility to use artificial intelligence in combination with spectroscopy-augmented mass spectrometry for carbohydrates sequencing and glycomics applications.

Introduction

DNA and protein sequencing technologies that aim at determining the structure of a biopolymer have been established decades ago and are commonly used in a routine and automated manner. However, the development of such technology for the sequencing of the third class of biological polymer – glycans, also known as carbohydrates, saccharides, or "sugars" – lags far behind. This lack of dedicated analytical tools (glycomics) is clearly identified as a critical bottleneck, impeding the full development of glycosciences despite their relevance for various strategic fields such as pharmaceutical and food industry; bio-based materials and renewable energy, and their considerable potential impact for the society in regard to the United Nations sustainable development goal [1].

The major roadblock to carbohydrate sequencing is intrinsically due to their unique molecular properties, among biopoly-

mers. In contrast with proteins and DNA, which are linear polymers made of a limited number of building blocks with distinct molecular structures, carbohydrates feature hundreds of building blocks – many of them coming in groups of closely related isomers with ambiguous molecular structures – and they form complex, branched arrangements due to the versatility of the glycosidic bond (position and anomericity). In this context, designing generic carbohydrate sequencing methods is both a major scientific challenge and a strategic priority [2,3].

Few years ago we proposed an original solution by bringing together the best of both sides of the analytical chemistry world: Spectroscopy and mass spectrometry (MS). In short, our technology is based on a mass spectrometric analysis – which is particularly powerful for the analysis of complex biological samples but does not readily elucidate isomers which have the same molecular mass – augmented with a infrared laser-based spectroscopic dimension (MS–IR), thus providing valuable additional isomer resolution [4].

We demonstrated that this multidimensional MS–IR molecular fingerprint is unique to each carbohydrate building block and can be used to resolve their full sequence, including their monosaccharide content and the detail of their linkages (position and anomericity). Based on this basic principle, the identification of an unknown carbohydrate proceeds as follows: the polymer is fragmented in monomers, yet maintaining information on the initial structure and the spectroscopic fingerprint (frequency and intensity of the vibrational modes) of each monosaccharide unit is measured, and subsequently identified by comparison with a library of reference spectra of synthetic monosaccharide standards. In the early days of MS–IR spectroscopy, ca. one hour was necessary to record the IR fingerprint of a single molecule and the identification was made by visual inspection, which was shortly automated by introducing a score derived from the convolution between the spectrum of the analyte of interest and the library of reference spectra. Despite the advantage of being automated, this later approach remains biased: for each molecular species, a single spectrum is arbitrarily chosen by the operator and serves as reference for all future analyses.

The latest MS–IR developments brought the data collection down to few seconds [5]. This is a considerable step towards high throughput carbohydrate analysis, which must be accompanied by fast data analysis, thus excluding manual interpretation. Besides, in the prospective of deploying the technology beyond the molecular spectroscopy community, it is essential to develop an automated, reliable, and robust strategy for the analysis of the spectroscopic data. Machine learning methods appear to be appealing candidates to address this challenge. They have been used for mass spectrometry data analysis since the 2000's [6] and the idea of using them on vibrational spectra goes back to the early 90's [7]. Support vector machines (SVM) and decision tree ensemble methods were benchmarked on infrared spectra for cancer classification [8] and many research groups focused their efforts on using machine learning for simulating molecular structures; generating vibrational spectra; and classifying chemical groups based on vibrational features [9,10]. In a recent publication, the random forest approach was proposed to identify the presence of structural features in oligosaccharides based on their gas-phase IR spectra [11]. To the best of our knowledge, machine learning classification studies have not been reported to identify saccharides using MS–IR carbohydrate analysis.

Here, we report a study of a probabilistic deep neural network (Bayesian deep neural networks [12]) to support automated monosaccharide recognition for carbohydrate sequencing. We obtained a highly performing algorithm that we called "GlAIcomics", specifically trained on carbohydrates.

Methodology

Data production

Our carbohydrate analysis approach is based on the IRMPD spectroscopic scheme (infrared multiple photon dissociation), which is the combination of mass spectrometry and IR spectroscopy. IRMPD is an action spectroscopy method that allows recording IR absorption spectra of isolated gas-phase ions, based on the measurement of the wavelength-dependent laser-induced fragmentation yield. When the frequency of the laser is resonant with a vibrational mode of the molecule, the molecule absorbs the radiation and accumulates internal energy until fragmentation [13]. In previous works we have demonstrated that the monosaccharides or oligosaccharides resulting from the fragmentation of a larger precursor possess a very specific IR fingerprint in the 2–4 microns spectral range, that is highly valuable to resolve all types of isomers [4]. Typical experimental IR fingerprint data are shown in Figure 1: they feature the intensities of the vibrational resonances as a function of their frequency in the mid-IR range. After measuring its mass and its IR fingerprint, an unknown analyte (Figure 1a) is readily identified as "GlcNAc" (for *N*-acetylglucosamine) by comparison with the reference IR spectra of several candidates of identical mass (Figure 1b, featuring three stereoisomers of $C_8H_{15}NO_6$). With the rapid development of our approach, such method now reached a high data output since a single IR fingerprint can be obtained in few seconds. The fast and automatic identification and classification of the data becomes compulsory, which motivates the present study.

For this study, a first set of 33 labelled experimental spectra obtained as described previously [4] were collected for training

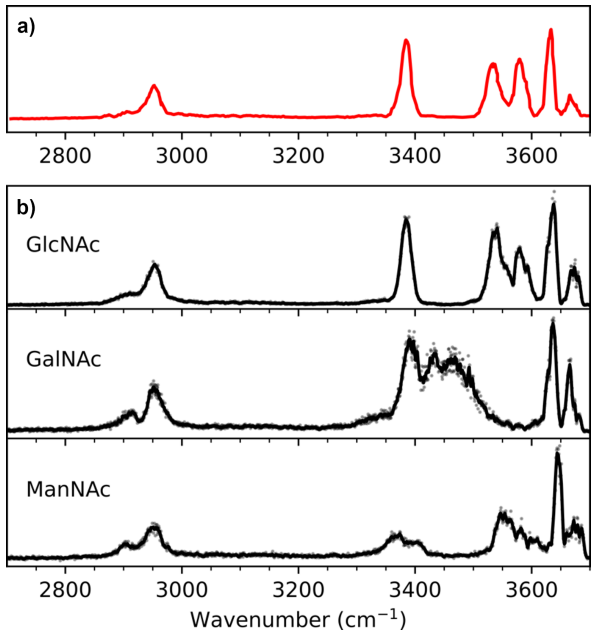


Figure 1: (a) Fingerprint of an unknown monosaccharide. (b) Labelled reference spectra of monosaccharide standards.

and validation of the model. The standard instrumental conditions for recording MS–IR data consist in a laser-enabled mass spectrometer equipped with a 3D ion trap mass analyzer. The following monosaccharides were analyzed: three stereoisomers of hexosamine of chemical formula $C_6H_{13}NO_5$, namely glucosamine (GlcN), galactosamine (GalN), mannosamine (ManN); and *N*-acetyl glucosamine (GlcNAc, chemical formula $C_8H_{15}NO_6$). One typical spectrum of each of the four monomers is shown in Figure 2. Note that both α and β-anomers coexist in the experimental conditions.

The second set of experimental MS–IR spectra was acquired using different instrumental conditions on a different experimental set-up: it consists of the coupling of an alternative design of mass spectrometer (equipped with a 2D ion-trap mass analyzer) with a higher repetition rate laser and a larger spectral bandwidth [5]. New GlcN spectra were acquired in these conditions. One of them is shown in Figure 3 (orange trace) for comparison with an experimental spectrum of GlcN acquired in standard conditions. Due to the larger spectral bandwidth, the spectrum from set 2 looks significantly different: the peaks are broader and less resolved than in the spectrum from set 1. This set is referred to as exogenous and was not use for training: it is used to illustrate the robustness of the method across significantly variable experimental conditions and instrumental performance.

The third set of experimental IRMPD spectra was acquired in standard conditions and includes 5 new spectra from the monomers GlcN, GalN, and ManN as in sets 1 and 2; as well as 7 spectra from species that do not belong in the training set categories (out of distribution, OOD), including disaccharides, a sulfated monosaccharide, and paracetamol. The outlying molecules represent potential "pollutions" in the analysis. This set of data is referred to as endogenous as it was measured on the same apparatus as the training set.

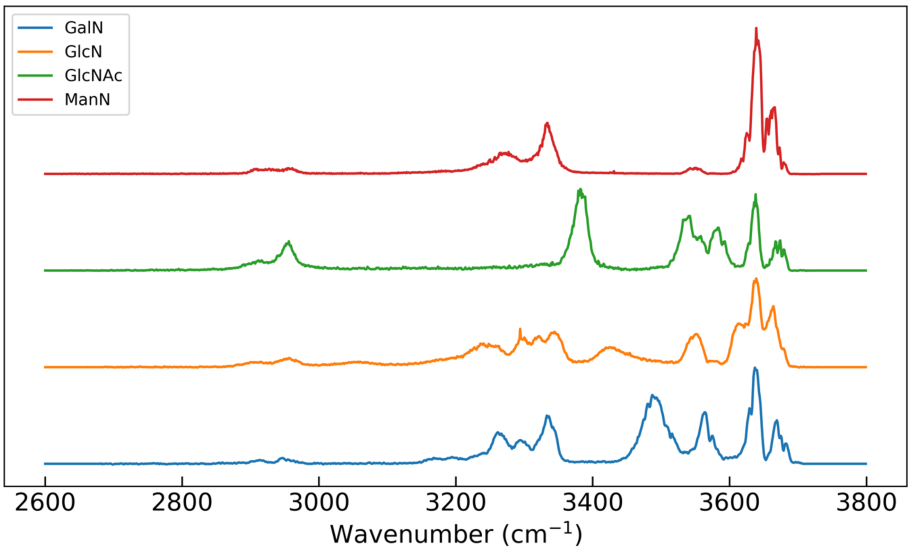


Figure 2: Typical experimental MS–IR spectra of the four categories of monosaccharides included in the first dataset. Blue: GalN; orange: GlcN; green: GlcNAc; red: ManN.

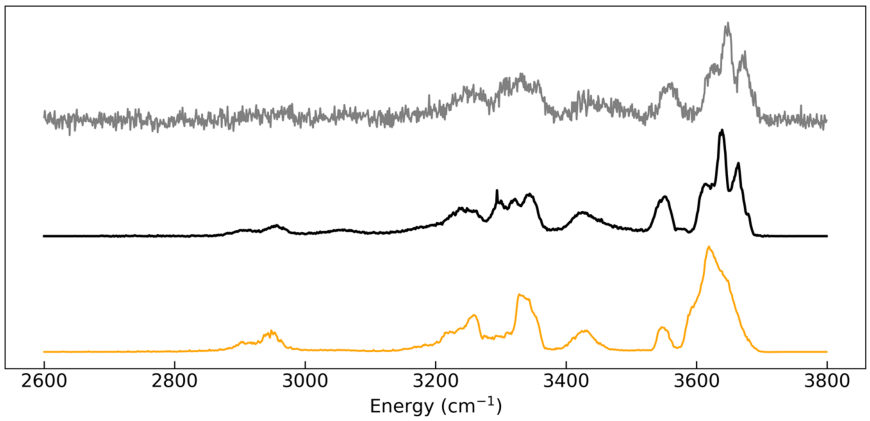


Figure 3: Synthetic IRMPD spectrum (grey trace) generated on the basis of a high resolution endogeneous experimental spectrum of GlcN (black trace) from dataset 1 using additional white noise: 10%; linear signal amplitude modulation: 5%; downsampling coefficient: 2; wavenumber shift: +9 cm^{-1}. The orange trace corresponds to a low-resolution exogeneous GlcN spectrum from dataset 2.

For efficient training of the algorithms, all three experimental datasets were augmented by producing synthetic variants. These synthetic spectra were generated by modulating the experimental ones with the following relevant sources of experimental fluctuations:

- The signal to noise ratio may vary from one measurement to another as it can emerge from a low amount of molecules. This was simulated by adding a Gaussian white noise with a randomly distributed standard deviation between 0 and 5% of the peak signal.
- The overall intensity of the laser can fluctuate from day to day or thorough the entire spectral range, which results in modulated peaks amplitudes. This was simulated as a linear variation of the signal amplitude across the spectral range. The variation was contained in a uniform distribution bounded by ±10%.
- Spectra can be recorded at increased speed for rapid analytical diagnostics, which traduces into a change in binning. To take this into account, data were binned with downgraded resolution then re-binned with 1 cm^{-1} step. The down sampling factor was randomly picked in a range from 1 to 5.
- Small variations of the calibration of the laser wavenumber may occur from day to day, leading to a shift of few wavenumbers of the vibrational spectrum. This was simulated with a maximum random shift per spectrum of ±10 cm^{-1}.

Finally, the synthetic spectra were normalized by z-score and interpolated over 1200 bins in the 2600–3800 cm^{-1} spectral range (1 cm^{-1} step) as input vector for the neural network. An example of a synthetic spectrum generated from an experimental spectrum is shown in Figure 3.

A total of 8000 synthetic spectra were randomly produced (2000 for each monomer category) out of the experimental spectra of set 1. They were shuffled to avoid training batches composed of a unique category of molecules. Finally, 70% of them were used for training of the models, and 30% were used for validation. The composition of the datasets used for training, validation and tests is summarized in Table 1.

Model architecture

In this study we opted for a fully connected feed-forward network based on the multi-layer perceptron architecture [14]

Table 1: Composition of the three datasets.

	Dataset 1	Dataset 2	Dataset 3
	training : 70% validation : 30%	classification tests	discrimination tests
categories	4	1	10
acquisition	standard	low res.	standard
exp. MS–IR spectra	33	4	12
augmented set	8000	8000	1300

with probabilistic approach (Bayesian deep neural network, DNN), which allows quantifying the model uncertainty for the classification results. It is composed of 3 hidden layers of 300, 225, and 100 neurons, respectively, and ReLu (rectified linear unit) activation functions for each layer. Two dropout layers are interleaved after the first and second hidden layers with a dropout setting of 25% to avoid over-fitting issues. The training objective is a classification task between the 4 monomer categories with a cross-entropy loss function.

To account for the probabilistic nature of the deep neural network, we used the variational inference technique. Each deterministic weight parameter was replaced by normal distributions defined by a mean value μ and a standard deviation σ which were optimized using the Bayes-by-Backprop method [15]. We chose this method that constrains the weights posterior distribution to normal distributions instead of the more accurate Markov chain Monte-Carlo (MCMC) method for calculation efficiency. With this approach, a quantitative uncertainty of the model predictions can be achieved by inferring each spectrum category several times with the trained model.

Results and Discussion

Model classification accuracy

Our GlAIcomics model shows a classification accuracy of 100% on the validation set and 99.98% on the test set (S.M : dataset 2 in Table 1). The 8000 synthetic spectra of set 2 were sorted by noise level, amplitude modulation, energy shift, and downsampling. The mean accuracy of the model as a function of these four parameters is shown in Figure 4. Note that all parameters have a uniform distribution over the 8000 samples and can be studied independently. The amplitude modulation and downsampling do not play a major role, with a maximum accuracy variation of 0.5%.

We demonstrated that the neural network is suitable for MS–IR classification in experimental conditions with variable resolution, noise or energy jitter.

The question remains on how to discriminate unknown molecules or to identify problematic spectra, such as the few misclassification events in the discussion above. In order to address these points, we further assessed the precision of the model and discussed its epistemic uncertainty in the next section.

Model precision and uncertainty

In the context of analytical chemistry where the fraction of "known molecules" (that is, previously referenced in databases) is expected to be significant compared to unknown ones, it is important to make sure that the model is discriminative and we want to maximize the precision of the model at this task. Indeed, the large amount of positive results would make it difficult to identify false positives. However, a small number of negative results is expected, which makes it doable to assess them systematically. False negative could be identified manually, labelled correctly, and injected back to improve the model.

The third dataset was used to evaluate the model discriminative power. It consists of 1300 spectra produced by augmentation of 12 original experimental spectra that were acquired on the standard instrumental setup and were never used by the models during the training and validation phases. This set contains 3 of the 4 known monosaccharides: ManN, GlcN, and GalN as well as 8 other molecules. For benchmarking purposes, all spectra were annotated with true labels.

By running the model inference for one spectrum multiple times we can measure the variability of its prediction probability for each category. If the model gives consistently a high probability for one category after each inference, then its uncertainty is low, and the spectrum likely belongs to the said category of molecules. On the other hand, if the model predicts a category with highly variable probability, then the uncertainty is high, and the spectrum likely does not belong to any of the classification categories. We ran model inference 200 times on each sample and obtained the mean prediction probability for every cate-

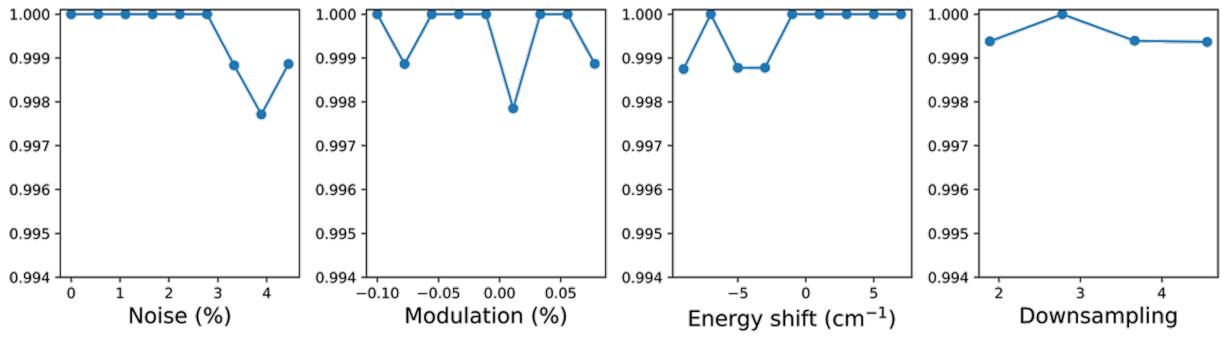


Figure 4: Model accuracy dependance with experimental conditions, represented by the dataset augmentation parameters.

gory as its variability represented by the interpercentile range 5 to 95%. The results are shown in Figure 5. As an example: the spectrum of CS-C is predicted as GlcNAc with 95% probability in average but for 10 inferences out of 200 (the lowest 5% percentile) the prediction probability is below 60%. In this example, by thresholding on the interpercentile range below 0.35 for the most likely prediction of each spectrum one can obtain a precision of 100%.

Most known molecules are assigned to the right category with a very sharp probability distribution that can be used as the prediction distribution under the null hypothesis that the model is reliable. For most of the "unknown" molecules the model prediction oscillates between two categories, but the probability distributions are extremely broad which means that the neural network uncertainty is important, and the corresponding results should be considered as unclassified and put aside for manual evaluation.

Finally, the performance of the GlAIcomics deep neural network model was compared with two different off-the-shelf techniques based on decision trees: Random forest (RF), an XGBoost (XGB). The evaluation methods are detailed in Supporting Information File 1. The classification accuracy for the validation subset (30% of set 1) is 100%, 99.95% and 100% for RF, XGBoost and GlAIcomics, respectively. For the test set (dataset 2), the accuracy is 99.91%, 99.61%, and 99.98%, respectively. When the accuracy of the prediction is further investigated as a function of the data augmentation parameters used to model experimental fluctuations, an advantage is found for GlAIcomics and RF over XGBoost. Lastly, the three methods were compared for the discrimination of molecules outside of the known categorie. GlAIcomics appears to discriminate samples more efficiently than the two other methods with true and false positive rates above 80% (70% and 50% for RF and XGBoost, respectively).

Conclusion

We have evaluated the performances of a Bayesian deep neural network for automatic analysis and classification tasks on glycans MS–IR fingerprints. It showed robust prediction accuracies on an exogeneous dataset. We observed that it is capable to generalize as it could categorize more noisy and distorted spectra. We then benchmarked its discrimination capabilities with a mixture of hexosamines and other molecular spectra: the Bayesian neural network architecture offers an access to the model reliability (through its epistemic error) when it comes to classify the spectra and could be used to discriminate outlying molecules or experimental issues when run on new data samples. Therefore, we conclude that a relatively small Bayesian deep neural network is a suitable solution for analysis and classification of saccharides in the context of MS–IR based carbohydrate sequencing. It can be easily integrated in an experimental data pipeline between the experiment raw spectra recording and the sequencing algorithm. Rejected spectra would be manually reviewed and fed back to the model as new training samples which in turn would reduce the epistemic error. It will therefore speed up the construction of glycans spectroscopic fingerprints database. In MS–IR experiments, the IR data as well as the mass of the molecule are simultaneously acquired,

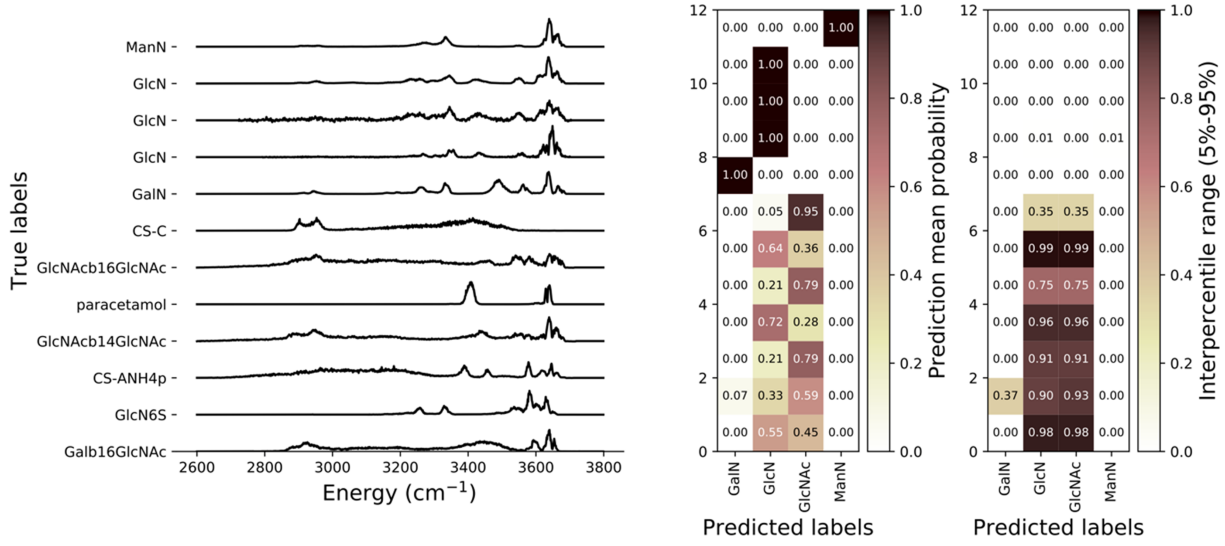


Figure 5: DNN Prediction results for third endogenous dataset (5 hexosamine samples and 7 other molecules). The middle map shows the mean prediction probabilities for each category and the right hand side map shows the 5% to 95% interpercentile range for the prediction probability distributions of each category.

therefore the mass could readily be used as a prefilter. More generally, all experimental data obtained in a glycomics workflow – such as MS/MS; HPLC; ion mobility; … – could ultimately be included in the algorithm for an optimal coverage of complex carbohydrates.

Supporting Information

Supporting Information File 1

Evaluation of the deep neural network model against two different techniques based on decision trees: Random forest (RF) and XGBoost (XGB).
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-19-134-S1.pdf]

Funding

This work was supported by Region Auvergne Rhône Alpes (IROLIGO grant), ANR Algaims (ANR-18-CE29-0006-02) and LABEX IMUST (ANR-10-LABX0064).

Author Contributions

B.S. and I.C. ran the spectroscopy experiments. T.R.B. proposed, trained and evaluated the machine learning model. T.R.B., B.S., B.M., E.F., F.L. and I.C. participated to the writing of the article

ORCID® iDs

Baptiste Schindler - https://orcid.org/0000-0002-7376-4154
Baptiste Moge - https://orcid.org/0000-0002-4932-6357
Isabelle Compagnon - https://orcid.org/0000-0003-2994-3961

References

1. United Nations; Department of Economic and Social Affairs; Sustainable Development Goals. https://sdgs.un.org/goals.
2. National Research Council. *Transforming Glycoscience: A Roadmap for the Future;* The National Academies Press: Washington, D.C., USA, 2012. doi:10.17226/13446
3. Gray, C. J.; Migas, L. G.; Barran, P. E.; Pagel, K.; Seeberger, P. H.; Eyers, C. E.; Boons, G.-J.; Pohl, N. L. B.; Compagnon, I.; Widmalm, G.; Flitsch, S. L. *J. Am. Chem. Soc.* **2019,** *141,* 14463–14479. doi:10.1021/jacs.9b06406
4. Schindler, B.; Barnes, L.; Renois, G.; Gray, C.; Chambert, S.; Fort, S.; Flitsch, S.; Loison, C.; Allouche, A.-R.; Compagnon, I. *Nat. Commun.* **2017,** *8,* 973. doi:10.1038/s41467-017-01179-y
5. Yeni, O.; Schindler, B.; Moge, B.; Compagnon, I. *Analyst* **2022,** *147,* 312–317. doi:10.1039/d1an01870a
6. Hilario, M.; Kalousis, A.; Pellegrini, C.; Müller, M. *Mass Spectrom. Rev.* **2006,** *25,* 409–449. doi:10.1002/mas.20072
7. Luinge, H. J. *Vib. Spectrosc.* **1990,** *1,* 3–18. doi:10.1016/0924-2031(90)80002-I
8. Sattlecker, M. Optimisation of Machine Learning Methods for Cancer Diagnostics using Vibrational Spectroscopy. Ph.D. Thesis, Cranfield University, Cranfield, U.K., 2011.
9. Fu, W.; Hopkins, W. S. *J. Phys. Chem. A* **2018,** *122,* 167–171. doi:10.1021/acs.jpca.7b10303
10. Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. *Adv. Sci.* **2019,** *6,* 1801367. doi:10.1002/advs.201801367
11. Riedel, J.; Lettow, M.; Grabarics, M.; Götze, M.; Miller, R. L.; Boons, G.-J.; Meijer, G.; von Helden, G.; Szekeres, G. P.; Pagel, K. *J. Am. Chem. Soc.* **2023,** *145,* 7859–7868. doi:10.1021/jacs.2c12762
12. Bishop, C. M. *J. Braz. Comput. Soc.* **1997,** *4,* 61–68. doi:10.1590/s0104-65001997000200006
13. Polfer, N. C. *Chem. Soc. Rev.* **2011,** *40,* 2211–2221. doi:10.1039/c0cs00171f
14. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning;* MIT Press: Cambridge, MA, USA, 2016.
15. Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight Uncertainty in Neural Network. In *Proceedings of the 32nd International Conference on Machine Learning, Vol. 37,* July 7–9, 2015; Lille, France; pp 1613–1622.

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Studying specificity in protein–glycosaminoglycan recognition with umbrella sampling

Mateusz Marcisz[‡1,2], Sebastian Anila[‡1], Margrethe Gaardløs[1], Martin Zacharias[3] and Sergey A. Samsonov[*1]

Full Research Paper

Open Access

Address:
[1]Faculty of Chemistry, University of Gdańsk, Gdańsk, Poland, [2]Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, Gdańsk, Poland and [3]Physics Department, Technical University of Munich, Garching, Germany

Email:
Sergey A. Samsonov[*] - sergey.samsonov@ug.edu.pl

* Corresponding author ‡ Equal contributors

Keywords:
glycosaminoglycan; molecular docking; protein–glycosaminoglycan interaction specificity; RS-REMD; umbrella sampling

Beilstein J. Org. Chem. **2023,** *19,* 1933–1946.
https://doi.org/10.3762/bjoc.19.144

Received: 27 October 2023
Accepted: 07 December 2023
Published: 19 December 2023

This article is part of the thematic issue "Chemical glycobiology".

Guest Editor: E. Fadda

Abstract

In the past few decades, glycosaminoglycan (GAG) research has been crucial for gaining insights into various physiological, pathological, and therapeutic aspects mediated by the direct interactions between the GAG molecules and diverse proteins. The structural and functional heterogeneities of GAGs as well as their ability to bind specific proteins are determined by the sugar composition of the GAG, the size of the GAG chains, and the degree and pattern of sulfation. A deep understanding of the interactions in protein–GAG complexes is essential to explain their biological functions. In this study, the umbrella sampling (US) approach is used to pull away a GAG ligand from the binding site and then pull it back in. We analyze the binding interactions between GAGs of three types (heparin, desulfated heparan sulfate, and chondroitin sulfate) with three different proteins (basic fibroblast growth factor, acidic fibroblast growth factor, and cathepsin K). The main focus of our study was to evaluate whether the US approach is able to reproduce experimentally obtained structures, and how useful it can be for getting a deeper understanding of GAG properties, especially protein recognition specificity and multipose binding. We found that the binding free energy landscape in the proximity of the GAG native binding pose is complex and implies the co-existence of several binding poses. The sliding of a GAG chain along a protein surface could be a potential mechanism of GAG particular sequence recognition by proteins.

Introduction

Glycosaminoglycans (GAGs) are long linear periodic anionic polydisperse polysaccharides, with repeating disaccharide units comprised of a hexuronic acid (or galactose in keratan sulfate) and a hexosamine (*N*-acetylglycosamide, GlcNAc or *N*-acetylgalactososamide, GalNAc) throughout a regular alternation of 1→4 and 1→3-glycosidic linkages [1-3]. GAGs are mainly located on the cell surface and in the extracellular matrix [4]. Due to their charged nature, they bind a large amount of water [5]. Although GAGs were previously considered just an inert glue surrounding the cell, GAG research in the past few decades has

illustrated the crucial role in cell signaling processes, including regulation of cell growth, proliferation and promotion of cell adhesion, anticoagulation, and wound repair [6-9]. All these processes are mediated through their direct interactions with diverse protein targets such as collagens, chemokines [10,11], and growth factors [12-14], which makes them essential in the cell biology [15,16]. In addition, GAGs also facilitate cell migration, act as shock-absorbers in joints and as a sieve in extracellular matrices and are important in maintaining the compressibility of the cartilage. The participation of GAGs in physiological, pathological, and therapeutic functions results principally from their unique physicochemical and structural features, including high negative charge, high viscosity and lubrication propensities, unbranched polysaccharide structures, low compressibility as well as the ability to attract and imbibe large amounts of water [17].

Unlike proteins or nucleic acids, GAGs are constantly altered by processing enzymes and thus they vary greatly in molecular mass, disaccharide unit composition, and sulfation. Based on their core structure they are categorized into six different classes, viz. heparan sulfate (HS), heparin (HP), hyaluronic acid (HA), chondroitin sulfate (CS), dermatan sulfate (DS), and keratan sulfate (KS). The structural and functional diversities of GAGs are regulated by their sequence, size of the chains, degree of sulfation, and the ability to bind proteins [1,18-21]. This structural diversity of GAGs translates into highly heterogeneous functions and allows them to modulate interactions with various protein molecules in respective biological processes [4]. Most of these interactions are driven by electrostatics and are non-specific in nature, however, some of them are highly specific or selective [22-26].

The structural analysis of GAGs improves the understanding of their biological functions and helps in the development of structure–activity relationships for these important biopolymers [27,28]. Although the composition of the individual saccharide components of GAGs is simple, the structural analysis of GAGs is extremely difficult due to their complex pattern of modification such as epimerization and sulfation [29]. In addition, GAGs' high flexibility and periodicity render these molecules profoundly challenging to analyze using experimental techniques only [30,31]. Thus, computational approaches could be efficiently used to gain insight into protein–GAG interactions that take place at single-molecule levels [32]. More than a complementary tool, computational approaches provide a better understanding of the role of individual interaction partners (including GAGs, solvent, and ions) by bringing often new and experimentally inaccessible details [33,34]. However, for computational researchers, there are still many challenges to overcome that originate from the physicochemical properties of GAGs, viz. their highly polarized (anionic) nature, their periodicity, and the complexity in decoding their sulfation pattern. Their charged nature necessitates the application of appropriate methods for electrostatics, ions, and solvent, particularly given their abundance in protein–GAG interfaces compared to complexes involving other classes of biomolecules. The periodicity can lead to multipose binding, wherein various configurations of the protein–GAG complex may exhibit similar free binding energies, allowing them to co-exist. Interpreting the "sulfation code", the amount (net sulfation) and particular positions of the sulfation group (sulfation pattern), could assist in the explanation and prediction of GAG specificity [35]. Computational methodologies like molecular docking and molecular dynamics (MD) have proven to be successful in modelling protein–GAG interactions, particularly examining the fundamental questions related to these interactions such as their specificity, the multipose character of GAG binding and the polarity of the binding poses of these periodic molecules.

In the present work, all-atom MD simulations are conducted to study the dynamics of the protein–GAG complexes, and are complemented by free energy analysis. The free energy analysis of the protein–GAG interactions is important in understanding the nature of the interactions and the stability of the binding pose, including the scenario when several co-existing binding poses are identified. We analyze the binding interactions between the GAGs heparin, heparan sulfate, and chondroitin sulfate, and the proteins basic fibroblast growth factor (PDB ID: 1BFC, https://doi.org/10.2210/pdb1BFC/pdb, [12]), acidic fibroblast growth factor (PDB ID: 2AXM, https://doi.org/10.2210/pdb2AXM/pdb, [13]), and cathepsin K (PDB ID: 3C9E, https://doi.org/10.2210/pdb3C9E/pdb, [36], and PDB ID: 4N8W, https://doi.org/10.2210/pdb4N8W/pdb, [37]). The third complex is known to exist in two different binding poses which are experimentally well established. In this study, the umbrella sampling (US) approach is used to pull away a GAG ligand from the binding site and then pull it back in. The main focus of our study is to evaluate whether the application of the US approach is able to reproduce experimentally obtained structures, and how useful it is for understanding GAG properties as protein recognition specificity and multipose binding. We also check for any trace of transition from the 3C9E to the 4N8W structure by pulling the ligand from its bound position and allowing the ligand to approach the protein from a very distant position to the binding sites.

Materials and Methods

Structures and parameters

Ligand preparation

GAG structures used in the study consist of two parts: 1. the part from the experimental structure (heparin in the 1BFC [12]

and 2AXM [13] complexes and chondroitin sulfate-4 in the case of 3C9E [36]/4N8W [37]), where the length is dp6 (dp stands for degree of polymerization) and 2. an additional part with different degree of sulfation or sulfation pattern (in case of ligands 1 and 2 for 1BFC and 2AXM dp6 desulfated heparan sulfate was added to the reducing end and non-reducing end of the GAG, respectively; in case of ligand 3 for the 1BFC and 2AXM dp6 desulfated heparan sulfate was added both to the reducing and non-reducing end of the GAG; in case of the ligand 4 for the 3C9E/4N8W complex dp6 chondroitin sulfate-6 was added to the reducing end of the GAG. The starting binding mode for the cathepsin K complex with chondroitin sulfate corresponded to the 3CE9 complex. Literature data for the sulfate groups [38] and GLYCAM06 [39] force field parameters were used for GAGs in the subsequent MD simulations. A $^1C_4$ conformation for the IdoA2S ring was chosen as it was shown to be the essentially dominant conformation in the microsecond scale simulations performed by Sattelle et al. as it is energetically more favorable than the ^2S_O conformation [40].

Complex preparation

The obtained ligands were docked using RS-REMD (replica exchange with repulsive scaling), an MD-based docking method [41], to assure proper binding poses of the whole ligand and ring puckering and to be consistent with further simulations. The docked ligands cover the binding site the same way as ligands in the experimental structures. Additionally, since the ligands used in the study are longer, they expand over the binding site and interact with the other parts of the protein as well. Experimental structures cover only a small part of the actual GAG molecule that interacts with the protein (as GAGs are built of tens to thousands of sugar units), therefore using longer ligands does not represent artificial behavior and may provide details of additional naturally occurring interactions. Comparison of the docked poses and PDB structures are presented in Supporting Information File 1, Figure S1.

MD simulations

All the MD simulations of the complexes obtained by RS-REMD docking were performed in AMBER20 package [42]. A TIP3P truncated octahedron water box with a distance of 20 Å from the solute to the box's border was used to solvate complexes. Na^+ counterions were used to neutralize the charge of the system. Energy minimization was performed preceding the production US runs (described in the next paragraph). 500 steepest descent cycles and 10^3 conjugate gradient cycles with 100 kcal/mol/Å^2 harmonic force restraint on solute atoms were performed. It was followed by 3×10^3 steepest descent cycles and 3×10^3 conjugate gradient cycles without any restraints and continued with heating up the system to 300 K for 10 ps with harmonic force restraints of 100 kcal/mol/Å^2 on solute atoms. Then, the system was equilibrated for each window at 300 K and 10^5 Pa in an isothermal, isobaric ensemble for 100 ps.

US production runs were performed for all of the complexes to pull away ligands from the binding site and then to bring them back to the binding site. US simulations consisted of 40 windows where in each the distance between ligand and the binding site was increased by 1 Å using harmonic restraints with a force constant of 10 kcal/mol/Å^2. Each window consists of 100 ns of US simulation, therefore each US simulation is 4 μs. Distances between the following atoms were chosen as a reaction coordinate in the corresponding complexes: Cα@Leu225-O5@12IdoA(2S) (the GAG sequence numbering is according to the AMBER order, from reducing to non-reducing end and @ means that a particular atom belongs to a particular residue) for basic FGF-ligand 1; Cα@Leu225-O5@1GlcNS(6S) for basic FGF-ligand 2; Cα@Gly275-O5@6IdoA(2S) for basic FGF-ligand 3; Cα@Gly5-O5@12IdoA(2S) for acidic FGF-ligand 1; Cα@Gly5-O5@1GlcNS(6S) for acidic FGF-ligand 2; Cα@Gly5-O5@4IdoA(2S) for acidic FGF-ligand 3; Cα@Arg296-C3@12GlcA for cathepsin K-ligand 4. The reaction coordinate values increased in each subsequent window, with the starting point for each window taken from the previous one.

The overlap between the probability distributions in adjacent windows was analyzed both using bootstrap error analysis and visually for equilibration and production runs. WHAM (weighted histogram analysis method [43]) was performed using Grossfield's WHAM program [44] to calculate the potential of mean force (PMF). For bootstrap analysis, 0.001 iteration tolerance, 300 K as temperature, and 1000 as number of Monte Carlo trials were used.

After completing the last window of US simulation, 500 ns unrestrained MD runs were carried out in the same isothermal isobaric ensemble to relax the system. A time step of 2 fs and a cut-off of 8 Å for electrostatics were used. The particle mesh Ewald method for treating electrostatics [45] and SHAKE algorithm for all the covalent bonds containing hydrogen atoms [46] were implemented in the MD simulations. The cpptraj program of AMBER was used for the analysis of the trajectories [47]. In particular, native contacts command with default parameters was used for the analysis of the contacts between protein and GAG molecules established in the course of the simulation.

Binding free energy calculations

MM/GBSA (molecular mechanics generalized born surface area) model igb = 2 [48] from AMBER20 was used for free

energy calculations on the trajectories obtained from RS-REMD simulations.

GAG binding pose accuracy evaluation

For the evaluation of the binding pose accuracy RMSD and RMSatd values were used. RMSD stands for root mean square deviation and it is defined as the average distance between the atoms of superimposed molecules. RMSatd (root mean square atom-type distance) is very similar to the widely used RMSD but instead of using specific atoms it compares atom types (e.g., any carbon atom to any carbon atom instead of specifically numbered carbon atom to the carbon atom with the same number). RMSatd is more appropriate when used for long and periodic molecules (such as GAGs), when a shift by one periodic unit yields the same pose but would result in high RMSD. The similar issue happens when GAG is rotated by 180°: although it occupies the same binding site and the pose is similar, the RMSD value would be expressed in tens of angstroms, while the RMSatd value would be significantly smaller.

Data analysis and its graphical representation were done with the R-package [49] and VMD [50].

Results

In total, 14 US simulations were performed to investigate the specificity of GAG–protein interactions, capabilities of US simulations to dissociate and reassociate protein–GAG complexes in these systems, and potential use of the US simulations in docking of GAG molecules to proteins. In order to do so, six different heparin systems (3 for basic FGF and 3 for acidic FGF) and one chondroitin sulfate system (with cathepsin K) were prepared. For each of the systems 2 US simulations were set up. First, hybrid GAGs (Figure 1) were prepared and docked using RS-REMD to find the pose in the binding site with the lowest interaction energy. Then, the GAG was pulled away from the binding site until it was shifted 40 Å from the starting position. Afterwards, the GAG was pulled in towards the binding site to observe if it reproduces a pose similar to the starting pose. To describe these unbinding and rebinding processes, analyses of RMSD, binding energy, contacts, and hydrogen bonds were performed. Additionally, after the final pulling step, a short MD run of 500 ns was performed to relax the system and to check if the final pose was energetically stable or if it changed during the relaxation step. The data depicted in the graphs result from the analysis of merged US trajectories. While this representation is not entirely physically sound, as the outcomes for each US window reflect the system's state under particular conditions with explicitly defined reaction coordinate values, the visualization of these continuous data potentially offers a more comprehensive insight into the complexity of the system related to its dynamic behavior within each window.

Basic FGF

Ligand 1. The RMSD increased gradually up to values of around 40 Å during the unbinding process, and then decreased slowly when it was pulled in. After about the 20th window RMSD stabilized between 15 and 20 Å, suggesting that the GAG did not find the initial pose and was trapped in a different minimum (Figure 2). The same scenario was observed in terms of the binding energy (Supporting Information File 1, Figure S2). When the ligand was pulled away the energy increased and when it was pulled in the energy slowly decreased and converged after about 20 windows. The number of native contacts when the ligand was pulled away rapidly dropped from 1500 to 0 and remained 0 for the rest of the US run (Supporting Information File 1, Figure S3). When pulled in, between 20 and 30% native contacts are restored after the 25th window but not to the original level. Even the additional relaxation MD run did not restore any native contacts. This suggests that the GAG gets close to the binding site but does not return to a similar conformation as the initial (experimental) pose. A similar trend is observed with hydrogen bonds where the number of H-bonds drops when the ligand is pulled away but never gets fully restored after being pulled in to the initial pose. Visual analysis supports the observation that only a small part of the GAG chain from the final pose overlaps with its starting position. The final pose is perpendicular to the initial one (Figure 3).

Ligand 2. RMSD slowly increased when pulled away and then when pulled in it gradually decreased to between 6 and 8 Å. During the additional relaxation step, RMSD was further reduced to 3 to 4 Å suggesting that the GAG finds a pose similar to the starting one (Figure 2). The binding energy gradually increased when the ligand was pulled away (from around −150 kcal/mol to around −30 kcal/mol) (Supporting Information File 1, Figure S1). Then, when pulled in, the energy almost did not decrease at the beginning. Only after the 21st window the energy started to decrease more rapidly but it did not go back to the values of −150 kcal/mol corresponding to the initial pose and oscillated around −120 to −100 kcal/mol. During additional relaxation the energies decreased to the range from −130 to −100 kcal/mol. This shows that after an additional MD run, the binding pose did not only become closer to the original structure but also was stabilized energetically in comparison to the pre-relaxation step. The number of native contacts significantly dropped after the first part of US (from ≈2000 contacts to below ≈500) and then stabilized at around 200 to 300 contacts in the last windows (Supporting Information File 1, Figure S3). When the ligand was pulled back to the binding site, only some native contacts were restored (≈500), but during the subsequent relaxation the number of restored native contacts increased to more than 1000. In case of H-bonds, at the end of the US 70 to

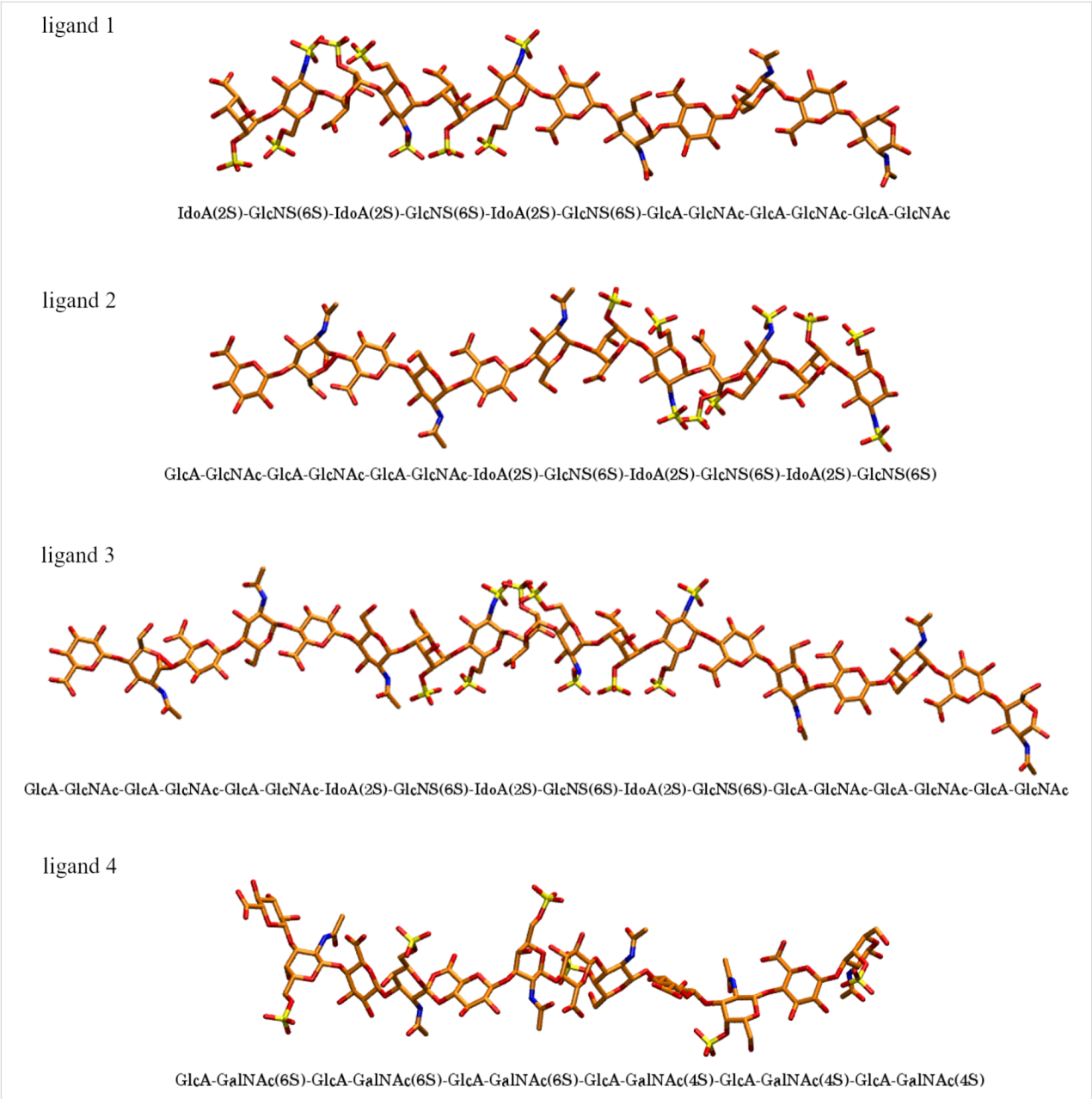


Figure 1: Graphical representation of ligands (in licorice, hydrogen atoms not shown) used in this study. Ligands 1, 2, and 3 were used in complexes with basic fibroblast factor (as modifications of the 1BFC PDB structure) and acidic fibroblast factor (as modifications of the 2AXM PDB structure) while ligand 4 was used with cathepsin K.

90% of them were restored. Visually, both the final and the initial poses look very similar (Figure 3), and this is also reflected in very low RMSD values (3 to 4 Å for such long and flexible molecules is considered to reflect high structural similarity).

Ligand 3. Similar to the other ligands, when pulled away from the binding site the RMSD of ligand 3 gradually increased and during pulling in it slowly decreased but did not return to the initial pose which is represented in the RMSD value of 12 Å at the end of the US simulation. Additional relaxation MD also did not result in any significant decrease of RMSD (Figure 2). The initial binding energy of −160 kcal/mol increased very fast at the start of pulling away and finished below −30 kcal/mol at the end (Supporting Information File 1, Figure S2). During the pulling in of the GAG, the energy decreased slowly and reached −70 to −50 kcal/mol at the end of US. However, after the 37th window the binding energy drops below −120 kcal/mol suggesting a more favorable novel ligand conformation. During relaxation, MD energies only improved slightly which is in

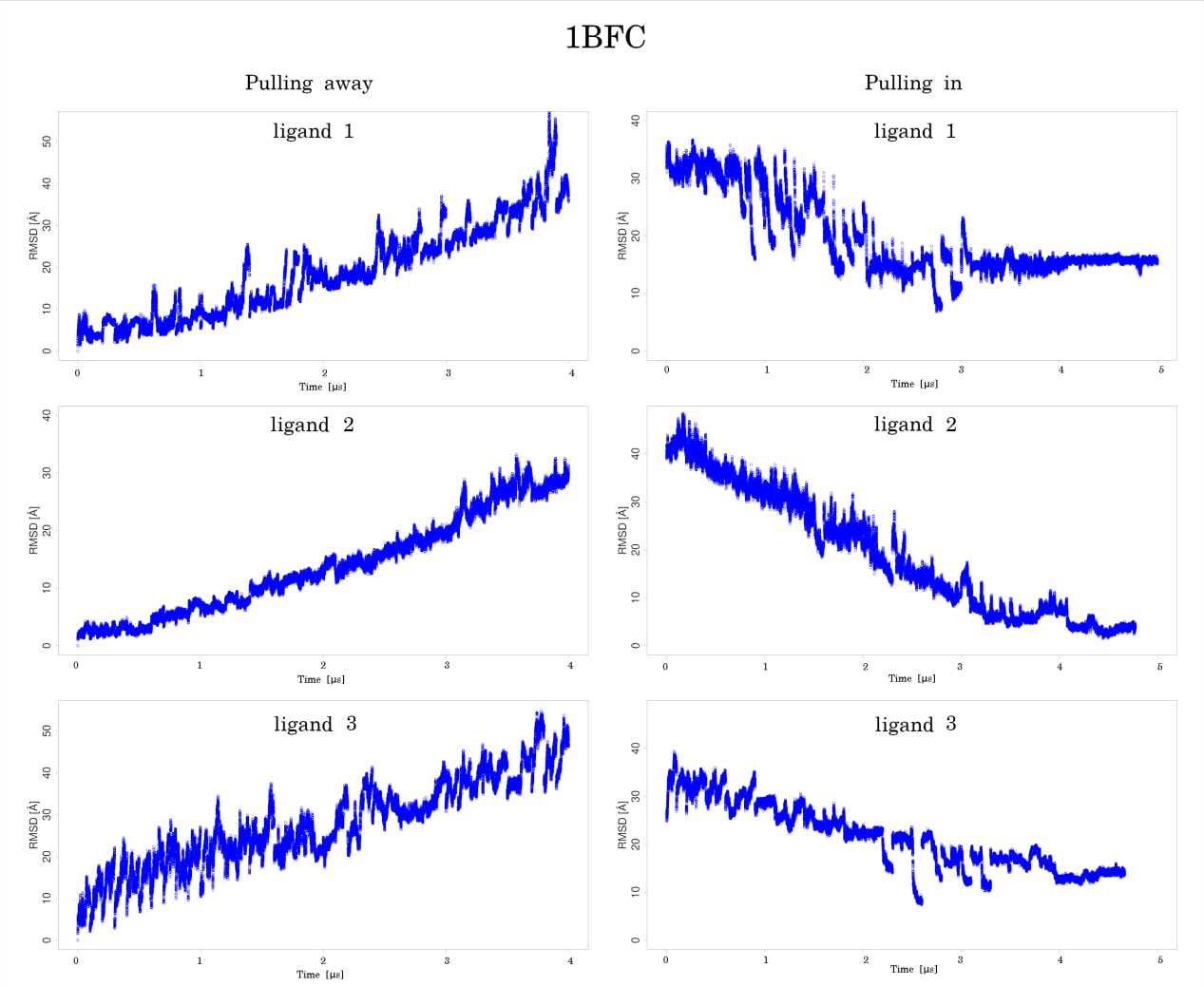


Figure 2: RMSD values obtained using the cpptraj script from AMBER suite for 1BFC (basic fibroblast factor) complexes with 3 different ligands: the ligand is pulled away from the binding site (left panel) and is pulled in towards the binding site (right panel).

agreement with the high RMSD that suggests that GAG did not return to the initial binding pose. The number of native contacts decreased drastically during the first windows of US from 1800 to 0 in the 13th window (Supporting Information File 1, Figure S2). During pulling in ligand towards the binding site only a small percentage of the native contacts were restored (50 to 150 native contacts in the last windows). After the relaxation MD the number of contacts went up to 200 to 250, but it never reached levels close to the initial ones which also suggests that the GAG did not get close to the binding site. The number of H-bonds at the end of pulling in was similar to the start of pulling away. However, none of the H-bonds at the end of the US simulation were established between the same atoms as at the start. Visually, only a part of the GAG's final pose overlaps with the initial one. The final pose adapts a perpendicular conformation to the starting GAG chain orientation (Figure 3).

Acidic FGF

Ligand 1. RMSD slowly increased during the first phase (windows from 1 to 8) of pulling away and afterwards with the pace similar to the other systems analyzed in this work. On the way back, RMSD of the ligand steadily decreased reaching values 4 to 5 Å at the end of the pulling in (Supporting Information File 1, Figure S4). During the relaxation step, RMSD remained around the same level and did not decrease further. Binding energy started at −140 kcal/mol and increased fast during the first 30 windows (Supporting Information File 1, Figure S5). Afterwards, it oscillated between −20 and 0 kcal/mol. During pulling in of the ligand the energy did not change before window 25 when it started to decrease reaching −90 kcal/mol at the last window. During the relaxation, the energy remained at a similar level. Interestingly, despite the low RMSD at the end, the final energy is less favorable (−90 kcal/mol) than the one observed at the beginning of the US

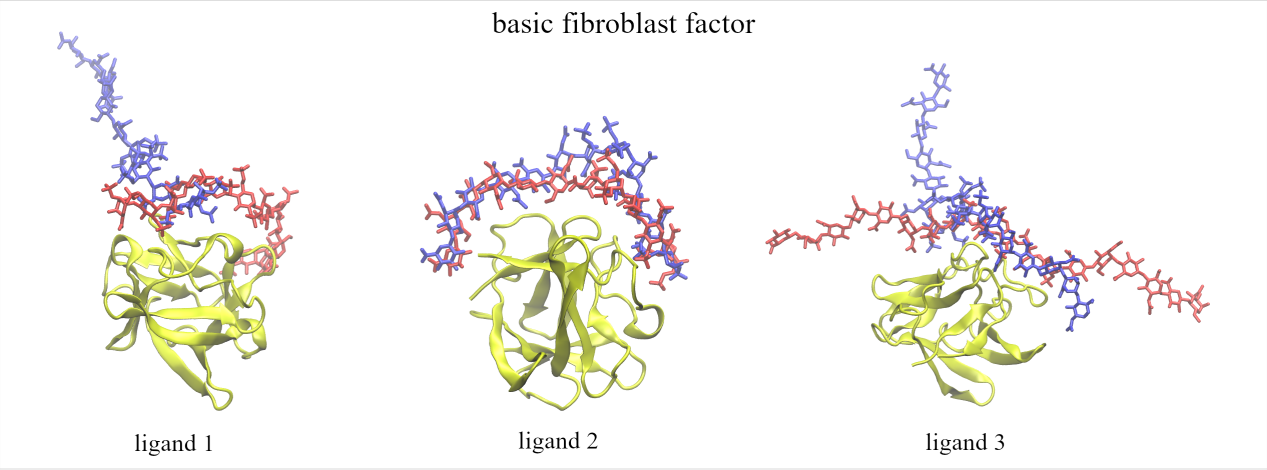


Figure 3: Graphical representation of the ligands' starting (in red, licorice) and final (in blue, licorice) positions in regard of the binding site of basic fibroblast factor (in yellow, new cartoon).

(−140 kcal/mol). The number of native contacts dropped to zero around the 15th window and remained 0 for the rest of the pulling away (Supporting Information File 1, Figure S6). During pulling in, no restoration of native contacts was observed. Also the number of H-bonds when the ligand was pulled all the way in was slightly lower (70 to 80%) than before it was pulled away. The number of H-bonds and native contacts suggest an overall smaller amount of interactions between the ligand and the receptor, but also the establishment of non-native contacts. Visually, the poses from the start and at the end of the US simulations look very similar (Figure 4), with major differences observed around the part of the GAG that is not bound to the protein.

Ligand 2. RMSD increased slowly until the 7th window where it started to increase more rapidly. During pulling in, RMSD did not decrease significantly (although visually the ligand is getting close to the initial binding pose) suggesting a drastically different pose of the ligand (Supporting Information File 1, Figure S4). Additional relaxation also did not improve RMSD significantly. In terms of energy of the system it started around −140 kcal/mol and it dropped to the level between −20 and 0 kcal/mol after the 27th window (Supporting Information File 1, Figure S5). During pulling in the energy did not improve significantly. The number of native contacts dropped from ≈1000 to 0 after the 15th window of pulling away (Supporting Information File 1, Figure S6). Only very few native contacts were restored during pulling in at maximum showing 200 of them. The number of H-bonds during pulling in was slightly lower than during pulling away suggesting less interactions between the ligand and the receptor on the way back than at the start of the US simulation. Visually, the major part of the GAG at the end of the US simulation overlapped with its starting pose. However, the final structure is more bent and shifted by

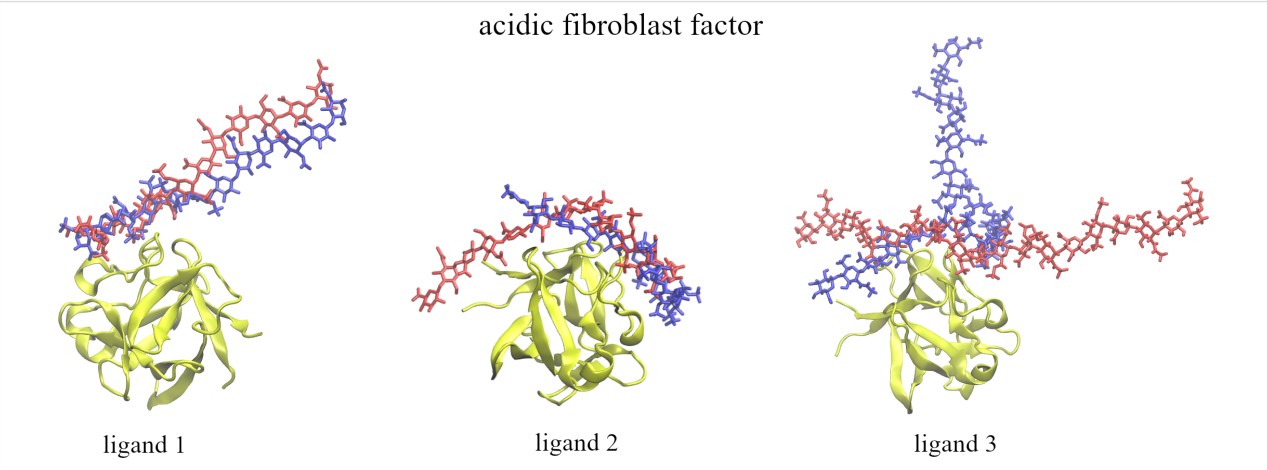


Figure 4: Graphical representation of the ligand's starting (in red, licorice) and final (in blue, licorice) poses in regard of the binding site of acidic fibroblast factor (in yellow, new cartoon).

about 3 rings relative to the initial one (Figure 4). This is also confirmed by relatively high RMSD values that did not improve much during the course of the pulling in US.

Ligand 3. RMSD increased slowly during the first few windows but unlike the other ligands in this particular case the scenario for RMSD did not change significantly afterwards. During pulling in the ligand back to the binding site only a low RMSD decrease was observed (Supporting Information File 1, Figure S4). During the relaxation, again, only a minor decrease in RMSD was observed suggesting that a slightly more favorable pose was achieved. In terms of the energy evolution during pulling away, it started around −140 kcal/mol and then it increased up to the 30th window where it stabilized below 0 kcal/mol (Supporting Information File 1, Figure S5). On the way back, we observe only a partial improvement of the binding energy as it reached values from −80 to −70 kcal/mol at the end of pulling in. However, during relaxation the energy lowered to values from −130 to −120 kcal/mol suggesting binding almost as strong as at the start of the US. The relatively high RMSD and low energy can be justified by the fact that the obtained pose of the ligand was very different from the initial one but there is a small overlapping part that interacts with the ligand around the binding site which can serve as basis for this strong binding. The number of native contacts at the beginning was 1300 and decreased slowly (Supporting Information File 1, Figure S6). In the second part of pulling ligand away changes in number of native contacts were sudden and drastic but they never went completely to 0. The number of contacts oscillated between 50 and 500. On the way back of the ligand, changes are much more subtle and the number of contacts remained between 200 and 400. During the relaxation no significant changes in the number of native contacts was observed. More H-bonds were present at the beginning of pulling the ligand away than at the end of the pulling in suggesting more interactions between the ligand and the protein at the start than at the end of the US. Visually, the final pose of the GAG is much different than the initial one. It is significantly bent and adapts a perpendicular conformation with regard to the starting pose (Figure 4). However, the sulfated part of the GAG overlaps with its initial position.

Additionally, the correlation between the ligand's RMSD and MM/GBSA per frame was analyzed (Table 1). In all cases positive correlations between analyzed values was observed. However, in some cases this correlation was below 0.5. This is in agreement with the data described above, which showed that despite a significantly different binding pose, sometimes the GAG maintained a relatively strong binding to the protein. This is particularly true for ligand 3 of acidic fibroblast growth factor, which when pulled back into the binding site led to low binding energies but a drastically different pose (partially perpendicular) of the ligand.

Energy contributions of sulfated and unsulfated parts of the GAG were investigated from per residue decomposition of MM/GBSA analysis (Table 2). In every case, sulfated parts were always contributing more to the receptor binding than unsulfated ones. Usually, the sulfated part contributed 3–5 times stronger than the unsulfated part. However, during pulling in of ligand 3 for basic fibroblast growth factor the unsuflated part contributed significantly (−7.6 kcal/mol for the unsulfated part in comparison to −10 kcal/mol for the sulfated part, respectively). More interestingly in this case during the pulling in process the contribution of the sulfated part decreased while the one of the unsulfated part increased. This could be interpreted as that the binding of the unsulfated residues can partially compensate the energy loss due to unbinding of the sulfated residues, suggesting rather non-specific interactions between the protein and the ligand.

Cathepsin K

During the pulling away of the GAG RMSD slowly and steadily increased. During pulling in RMSD only lowered slightly reaching 20 Å which suggests that at the end of US the GAG did not return to a pose similar to the starting one. Relaxation MD neither improved the final conformation. The energy of the system increased from −120 kcal/mol to values between −50

Table 1: Pearson correlation coefficients between energies obtained from MM/GBSA analysis and RMSD values of the ligand for all frames of the merged MD trajectories.

basic fibroblast growth factor (1BFC)					
ligand 1 (away)	ligand 1 (in)	ligand 2 (away)	ligand 2 (in)	ligand 3 (away)	ligand 3 (in)
0.77	0.60	0.84	0.90	0.80	0.58
acidic fibroblast growth factor (2AXM)					
ligand 1 (away)	ligand 1 (in)	ligand 2 (away)	ligand 2 (in)	ligand 3 (away)	ligand 3 (in)
0.81	0.78	0.63	0.35	0.50	0.25

Table 2: Energy contributions in kcal/mol of the sulfated and unsulfated parts of GAGs obtained from MM/GBSA per residue decomposition.

basic fibroblast growth factor (1BFC)											
ligand 1 (away)		ligand 1 (in)		ligand 2 (away)		ligand 2 (in)		ligand 3 (away)		ligand 3 (in)	
sulfated	not sulfated	sulfated	not sulfated	sulfated	not sulfated	sulfated	not sulfated	sulfated	not sulfated	sulfated	not sulfated
−13.1	−4.7	−12.8	−4.6	−16.1	−8.3	−11.9	−6.1	−12.1	−4.7	−10.0	−7.6
acidic fibroblast growth factor (2AXM)											
ligand 1 (away)		ligand 1 (in)		ligand 2 (away)		ligand 2 (in)		ligand 3 (away)		ligand 3 (in)	
sulfated	not sulfated	sulfated	not sulfated	sulfated	not sulfated	sulfated	not sulfated	sulfated	not sulfated	sulfated	not sulfated
−11.4	−1.6	−4.4	−0.7	−11.1	−2.5	−14.0	−6.8	−14.8	−3.6	−7.8	−0.8

and −40 kcal/mol around the 22nd window and then remained at this level to the end of pulling away. On the way back of the GAG to the binding site the energy slowly decreased and reached −80 kcal/mol at the end of pulling in. During the relaxation MD the energy decreased further to −110 kcal/mol which is almost the same value as observed at the starting point suggesting that this significantly different pose is almost as stable as the initial one. The number of native contacts lowers from ≈1500 to 0 after the 25th window of the US simulation. During pulling in some native contacts are being restored but the number greatly varies and never surpassed 500 contacts. The number of H-bonds during pulling in are also lower than in the initial pose. Visually, the final pose is significantly different than the starting one (Figure 5 and Figure 6). In the binding site the part with the 4-sulfation of the final GAG conformation is perpendicular to the starting one, and the part of GAG with the 6-sulfation is close to the second GAG binding site of the cathepsin K. The final pose of the GAG partially overlapped with both experimentally known binding sites. This is most likely the reason why the energy at the start and at the end of US is similar to the one of the initial pose despite the fact that much a smaller part of the GAG is located at the first binding site. Hence a comparison of the binding position of the ligand with both crystal structures (3C9E and 4N8W) were carried out to reveal which binding site is preferred upon the reassociation of the ligand (Figure 6). It can be seen that the binding of the ligand to the protein at the end of the sliding in represents a combination of both the binding positions from the crystal structures and the RMSatd score obtained for the two different crystal structures are 4.1 Å and 8.3 Å for 3C9E and 4N8W complexes, respectively. The dodecamer ligand docked to the protein in such a way that the hexameric part with the 4-sulfation (as observed in the crystal structure) occupied the 4N8W site and the hexameric part with the 6-sulfation bound to the site observed in the 3C9E structure. The comparison of the final structure and that obtained after the docking yielded RMSatd of 10.2 Å, which shows that the pulling back results in the structure more similar to that of the crystal structure than to the initial docked pose. A similar comparison of the final structure obtained at the end of the sliding in process was done for the other complexes with their corresponding crystal structures and the one obtained after docking. The ligands' RMSatd values (Tables S1 and S2 in Supporting Information File 1) show that in the majority of the complexes the final structure is more similar to the crystal structure than to the initial docked structure. However, the goal of the study was to compare the final structures to the starting positions rather than to the experimental

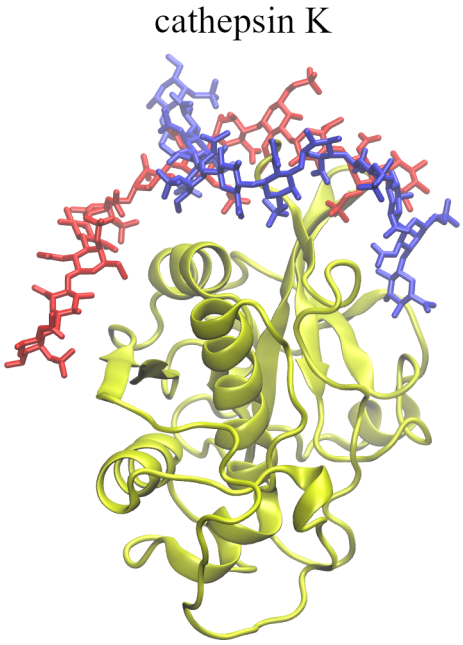


Figure 5: Graphical representation of the ligand 4 starting (in red, licorice) and final (in blue, licorice) position in regard of binding site of cathepsin K (in yellow, new cartoon).

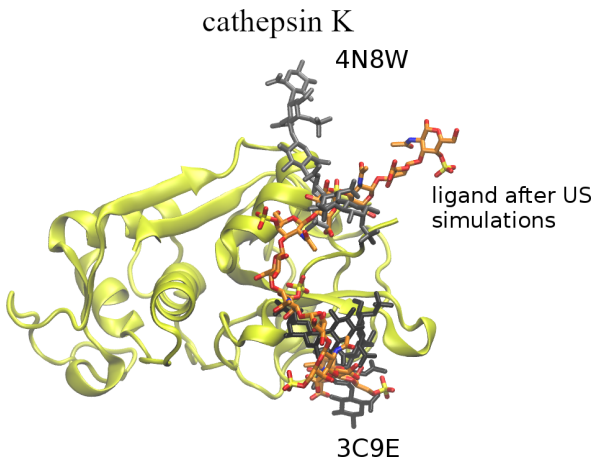


Figure 6: Comparison of the last frame of the US simulation (orange, licorice) with the chondroitin sulfate ligands from the original crystal structures: 3C9E (ligand in black licorice, at the bottom of the figure) and 4N8W (ligand in grey licorice, at the top of the figure).

ones (although both docked and experimental poses are close to each other, see Figure S1 in Supporting Information File 1) to check specificity of the GAG interactions and evaluate the quality of the information obtained from the US simulations.

Investigation of the protein–GAG recognition in the proximity of the native binding pose

Next, we analyzed if the US approach is able to reproduce the native binding pose when pulling away a ligand by just a disaccharide unit and returning it back to the binding site. These simulations involved approximately a 10 Å shift from the native complex of ligand 1 from the basic FGF and allowed to investigate the near-native free energy landscape and the respective atomistic details of the protein–GAG recognition. In the forced dissociation process, the RMSD curve looks similar to the ones from the previously described, longer pulling away trajectories: the RMSD values increase gradually as the ligand pose gets closer to the native pose within the shift of a monomeric unit (first part of the pulling away step, 0.5 μs), yielding a rugged shape of the curve (Figure 7). Interestingly, on the way back, the RMSD values reach minimal values at windows 5, 7, and 8, corresponding to the reaction coordinate values of 5 Å, 3 Å, and 2 Å, respectively (as well as corresponding to the 0.5, 0.7 and 0.8 μs of US, respectively), but then go up at the end of the pulling in process.

This is also reflected by the MM/GBSA binding energy analysis (Supporting Information File 1, Figure S7), where the energy gradually increased in the process of dissociation with

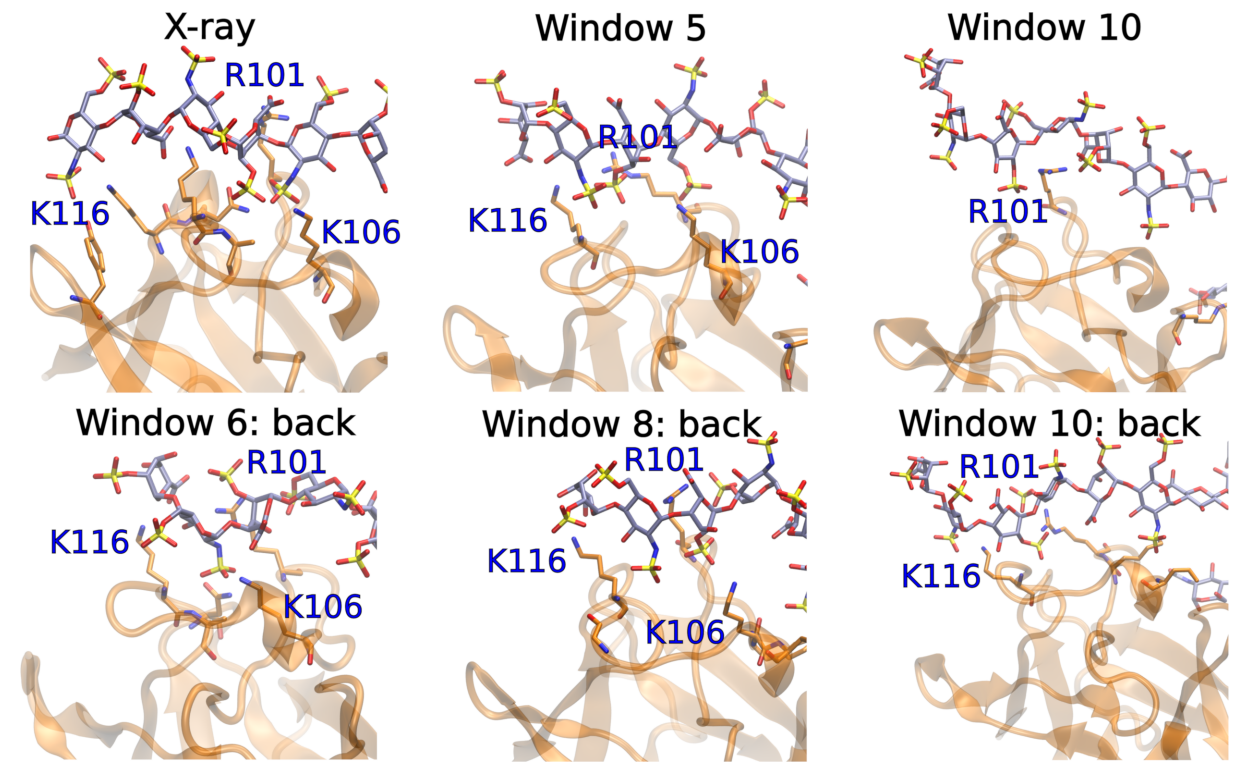


Figure 7: The X-ray conformation of bFGF-HP dp6 (PDB ID: 1BFC) and windows 5 and 10 and windows 6, 8, 10 in the forward and reverse processes for bFGF and ligand 1, respectively. The protein is in cartoon, the GAG and the protein residues establishing H-bonds with the GAG with the occupancy higher than 20% are in sticks. R101, K106, and K116 are labeled.

the exception of the stabilized conformation at window 5 (0.5 μs), where the binding strength is energetically comparable with the one of the native binding pose. The MM/GBSA free energy landscape is very rugged on the way of pulling the ligand in. However, the most favorable energies of very comparable values are observed in windows 6 and 8 (0.6 and 0.8 μs, respectively) suggesting that while for window 8 the proximity to the native pose was energetically favorable, the interaction free energy minimum in window 6 corresponds to a distinct non-native binding pose. This suggests a high heterogeneity of the free energy landscape in the proximity of the native binding pose and a high propensity for multipose binding in the system. The number of native contacts and total H-bonds gradually decrease in the dissociation process (Figure 7), while there is a clear peak of the non-native contacts number and additional H-bonds at window 5 suggesting a partial stabilization of the binding by H-bonds, also in agreement with the MM/GBSA binding energy trend. On the way back, at window 8, the ligand establishes most of the native contacts which also correspond to the increase of the number of H-bonds established. This points out that several stabilized and energetically comparable binding poses co-exist in the system. This is also supported by the absence of significant correlations between the MM/GBSA energy and RMSD to the initial pose (Supporting Information File 1, Figure S8).

Another way to calculate the free energy landscape with less details but in a manner more sound for the applied protocol, is to estimate PMF along the reaction coordinate. Such calculations also support the conclusion that the applied protocol did not succeed in returning the system to the original binding pose (Supporting Information File 1, Figures S9–S11). In comparison to the data from the MM-GBSA approach, the PMF data show even larger differences between the starting and the final poses obtained in the US trajectories.

In turn, results of the analysis of pairwise correlations between the number of established H-bonds, MM/GBSA free energy, native, non-native and total number of contacts differ for pulling out and pulling in processes (Table 3). For the dissociation, as expected the number of native contacts and H-bonds correlate very well with the MM/GBSA energies (Pearson correlation coefficients obtained for all frames of the trajectories are 0.81 and 0.76, respectively), while on the way back, the still high correlation with the H-bond number (0.58) and an essentially decreased one with the native contact number (0.36) mean that hydrogen binding dominates the binding energetics of the system and is an origin of multipose binding.

When correlating the values for MM/GBSA and the number of H-bonds averaged per each US window, the correlation coefficients for the pulling away and pulling back processes are 0.97 and 0.56, respectively. Despite these significant differences in the correlations, implying a more complex free energy landscape topology when the ligand is pulled in, the energies per H-bond calculated from the linear regression model are very similar: −10.1 ± 0.6 kcal/mol and −10.5 ± 1.0 kcal/mol for pulling away and in, respectively. These differences in the correlations, however, can be partially attributed to the arbitrary choice of the US reaction coordinate which can affect the pulling away and pulling in processes and, therefore, the data described here.

Furthermore, we analyzed in detail the most representative H-bonds (with the occupancy higher than 20%) established at different US windows that were the most distinguishable in the pulling away and pulling in processes. In particular, we analyzed the X-ray conformation (PDB ID: 1BFC) based MD trajectory and windows 5 and 10 and windows 6, 8, 10 in the forward and reverse processes, respectively (Figure 7). Interestingly, in all these windows with more favorable binding energies, particular three positively charged residues, R101, K106, and K116, maintained strong H-bonds that have been also established in the X-ray structure-based MD simulation [51]. Some of these residues are absent as the most contributing to H-bonding in the less stable complexes (both last windows of the pulling away and pulling in processes). At the same time, the essential difference between the H-bonding pattern ob-

Table 3: Pearson correlation coefficients obtained for all frames of the pulling away and pulling in MD trajectories between the protein–GAG recognition parameters. Native: native contacts; non-native: non-native contacts; total: the sum of native and non-native contacts.

	Pulling away					Pulling in				
	N_{native}	$N_{non\text{-}native}$	N_{total}	$N_{H\text{-}bonds}$	$\Delta G_{MM/GBSA}$	N_{native}	$N_{non\text{-}native}$	N_{total}	$N_{H\text{-}bonds}$	$\Delta G_{MM/GBSA}$
N_{native}	–	–	–	0.65	0.81	–	–	–	0.18	0.36
$N_{non\text{-}native}$	–	–	–	0.02	0.08	–	–	–	0.25	0.34
N_{total}	–	–	–	0.54	0.72	–	–	–	0.36	0.58
$N_{H\text{-}bonds}$	0.65	0.02	0.54	–	0.76	0.18	0.25	0.36	–	0.57
$\Delta G_{MM/GBSA}$	0.81	–0.08	0.72	0.76	–	0.36	0.34	0.58	0.57	–

served in the unrestrained MD simulation of the X-ray structure is that there were several non-charged residues (N8, A17, Y84) among the top residues contributing to H-bonds, while almost exclusively positively charged residues were observed to be substantial H-bond contributors in the US windows. In a microsecond-scale MD simulation of the same X-ray structure, three non-charged residues were identified as the top MM/GBSA free energy contributors [52]. This suggests that despite a very complex free energy landscape in the proximity of the native pose, the native pose can be potentially distinguished by the essential contributions of the non-charged residues to the GAG recognition. Further, this implies a certain degree of specificity and not simply electrostatics-driven interactions in this particular molecular complex. Estimation of the free energy barriers in the completed analysis suggests that the sliding of a long GAG on the protein surface is a feasible process that could underline the natural recognition of the specific GAG patterns by a protein target.

Conclusion

In this study, the US approach was used to pull away a GAG ligand from the binding site and then to pull it back in to the binding site. The goal was to analyze if US is able to reproduce experimentally obtained structures, and if it can contribute to a deeper understanding of GAG properties as protein recognition specificity and multipose binding. Although the US is a powerful method it was shown not to be able to accurately reproduce experimental structures or the most energetically favorable binding poses in the majority of the investigated systems with the particular protocols we applied in this study. The limitations in our study can be attributed to two main factors. Firstly, the relatively short sampling times (100 ns) may have been insufficient to adequately equilibrate the systems, especially given their complex free energy landscape in each US window. Secondly, the selected reaction coordinates for pulling away and pulling in may not inherently suggest unique reversible pathways. To improve the convergence and sampling of the free energy landscape, advanced sampling protocols can be employed or the US simulations can be repeated multiple times. Therefore, the data obtained in this study and the conclusions related to these data are rather qualitative. In the next steps, we plan to apply more advanced sampling protocols. However, even when using the described protocol in some of the systems it was able to bring the ligand back to the binding site (in two cases with comparable accuracy to one of the most powerful GAG docking tools (RS-REMD), which corresponds to RMSD values <4 Å). Additionally, it allowed to observe multipose binding phenomena manifesting other energetically favorable binding poses of the GAG in the binding site. In these cases, although the RMSD values with reference to the experimental structures were high as only a very small part of the final GAG binding pose overlapped with the initial structure, binding energies remained almost at the same level as the ones corresponding to the experimental binding poses. Regarding the specificity, in most cases a partial overlap between the GAG parts in the experimental and the pulled in structures corresponding to the same sulfation pattern/amount was observed. Nevertheless, in one of the simulations of the basic fibroblast growth factor system a less sulfated part contributed comparably to the sulfated one suggesting a potential of non-purely electrostatics dominance in the protein–GAG interactions. The more detailed analysis of the GAG recognition in this system in near-native states points out to the complexity of free energy landscape but at the same identifies the key charged H-bonding contributors to the GAG binding that together with several non-charged residues in the binding interface potentially determine the specificity of the interactions in this complex. The analysis of free energy landscapes in the studied systems suggests that sliding of a GAG along a binding site in a protein target could occur naturally and, therefore, could be a way for a protein to effectively sample different particular GAG recognition patterns. The findings in this work should contribute to the broadening of the knowledge regarding the specificity of protein–GAG interactions and the limitations of the computational tools employed to analyze them.

Supporting Information

Supporting Information File 1

Additional information and graphical representations.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-19-144-S1.pdf]

Funding

This research was funded by the National Science Centre of Poland (grant numbers UMO-2018/30/E/ST4/00037 and UMO-2018/31/G/ST4/00246).

ORCID® iDs

Sergey A. Samsonov - https://orcid.org/0000-0002-5166-4849

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2023.46.v1

References

1. Bu, C.; Jin, L. *Front. Mol. Biosci.* **2021,** *8,* 646808. doi:10.3389/fmolb.2021.646808
2. Kogut, M. M.; Marcisz, M.; Samsonov, S. A. *Curr. Opin. Struct. Biol.* **2022,** *73,* 102332. doi:10.1016/j.sbi.2022.102332

3. Perez, S.; Makshakova, O.; Angulo, J.; Bedini, E.; Bisio, A.; de Paz, J. L.; Fadda, E.; Guerrini, M.; Hricovini, M.; Hricovini, M.; Lisacek, F.; Nieto, P. M.; Pagel, K.; Paiardi, G.; Richter, R.; Samsonov, S. A.; Vivès, R. R.; Nikitovic, D.; Ricard Blum, S. *JACS Au* **2023,** *3,* 628–656. doi:10.1021/jacsau.2c00569
4. Habuchi, H.; Habuchi, O.; Kimata, K. *Glycoconjugate J.* **2004,** *21,* 47–52. doi:10.1023/b:glyc.0000043747.87325.5e
5. Ruiz Hernandez, S. E.; Streeter, I.; de Leeuw, N. H. *Phys. Chem. Chem. Phys.* **2015,** *17,* 22377–22388. doi:10.1039/c5cp02630j
6. Moustakas, A.; Souchelnytskyi, S.; Heldin, C.-H. *J. Cell Sci.* **2001,** *114,* 4359–4369. doi:10.1242/jcs.114.24.4359
7. Salbach, J.; Rachner, T. D.; Rauner, M.; Hempel, U.; Anderegg, U.; Franz, S.; Simon, J.-C.; Hofbauer, L. C. *J. Mol. Med. (Cham, Switz.)* **2012,** *90,* 625–635. doi:10.1007/s00109-011-0843-2
8. Theocharis, A. D.; Skandalis, S. S.; Gialeli, C.; Karamanos, N. K. *Adv. Drug Delivery Rev.* **2016,** *97,* 4–27. doi:10.1016/j.addr.2015.11.001
9. Karamanos, N. K.; Piperigkou, Z.; Theocharis, A. D.; Watanabe, H.; Franchi, M.; Baud, S.; Brézillon, S.; Götte, M.; Passi, A.; Vigetti, D.; Ricard-Blum, S.; Sanderson, R. D.; Neill, T.; Iozzo, R. V. *Chem. Rev.* **2018,** *118,* 9152–9232. doi:10.1021/acs.chemrev.8b00354
10. Derler, R.; Gesslbauer, B.; Weber, C.; Strutzmann, E.; Miller, I.; Kungl, A. *Int. J. Mol. Sci.* **2017,** *18,* 2605. doi:10.3390/ijms18122605
11. Penk, A.; Baumann, L.; Huster, D.; Samsonov, S. A. *Glycobiology* **2019,** *29,* 715–725. doi:10.1093/glycob/cwz047
12. Faham, S.; Hileman, R. E.; Fromm, J. R.; Linhardt, R. J.; Rees, D. C. *Science* **1996,** *271,* 1116–1120. doi:10.1126/science.271.5252.1116
13. DiGabriele, A. D.; Lax, I.; Chen, D. I.; Svahn, C. M.; Jaye, M.; Schlessinger, J.; Hendrickson, W. A. *Nature* **1998,** *393,* 812–817. doi:10.1038/31741
14. Uciechowska-Kaczmarzyk, U.; Babik, S.; Zsila, F.; Bojarski, K. K.; Beke-Somfai, T.; Samsonov, S. A. *J. Mol. Graphics Modell.* **2018,** *82,* 157–166. doi:10.1016/j.jmgm.2018.04.015
15. Xu, D.; Esko, J. D. *Annu. Rev. Biochem.* **2014,** *83,* 129–157. doi:10.1146/annurev-biochem-060713-035314
16. Wigén, J.; Elowsson-Rendin, L.; Karlsson, L.; Tykesson, E.; Westergren-Thorsson, G. *Stem Cells Dev.* **2019,** *28,* 823–832. doi:10.1089/scd.2019.0009
17. Samantray, S.; Olubiyi, O. O.; Strodel, B. *Int. J. Mol. Sci.* **2021,** *22,* 11529. doi:10.3390/ijms222111529
18. Funderburgh, J. L. *Glycobiology* **2000,** *10,* 951–958. doi:10.1093/glycob/10.10.951
19. Sasisekharan, R.; Venkataraman, G. *Curr. Opin. Chem. Biol.* **2000,** *4,* 626–631. doi:10.1016/s1367-5931(00)00145-9
20. Prydz, K. *Biomolecules* **2015,** *5,* 2003–2022. doi:10.3390/biom5032003
21. Pomin, V. H.; Mulloy, B. *Pharmaceuticals* **2018,** *11,* 27. doi:10.3390/ph11010027
22. Imberty, A.; Lortat-Jacob, H.; Pérez, S. *Carbohydr. Res.* **2007,** *342,* 430–439. doi:10.1016/j.carres.2006.12.019
23. Nagarajan, B.; Holmes, S. G.; Sankaranarayanan, N. V.; Desai, U. R. *Curr. Opin. Struct. Biol.* **2022,** *74,* 102356. doi:10.1016/j.sbi.2022.102356
24. Petitou, M.; Casu, B.; Lindahl, U. *Biochimie* **2003,** *85,* 83–89. doi:10.1016/s0300-9084(03)00078-6
25. Rudd, T. R.; Preston, M. D.; Yates, E. A. *Mol. BioSyst.* **2017,** *13,* 852–865. doi:10.1039/c6mb00857g
26. Sepuru, K. M.; Nagarajan, B.; Desai, U. R.; Rajarathnam, K. *J. Biol. Chem.* **2018,** *293,* 17817–17828. doi:10.1074/jbc.ra118.004866
27. Gandhi, N. S.; Mancera, R. L. *Chem. Biol. Drug Des.* **2008,** *72,* 455–482. doi:10.1111/j.1747-0285.2008.00741.x
28. Palhares, L. C. G. F.; London, J. A.; Kozlowski, A. M.; Esposito, E.; Chavante, S. F.; Ni, M.; Yates, E. A. *Molecules* **2021,** *26,* 5211. doi:10.3390/molecules26175211
29. Raman, R.; Sasisekharan, V.; Sasisekharan, R. *Chem. Biol.* **2005,** *12,* 267–277. doi:10.1016/j.chembiol.2004.11.020
30. Zappe, A.; Miller, R. L.; Struwe, W. B.; Pagel, K. *Mass Spectrom. Rev.* **2022,** *41,* 1040–1071. doi:10.1002/mas.21737
31. Szekeres, G. P.; Pagel, K.; Heiner, Z. *Anal. Bioanal. Chem.* **2022,** *414,* 85–93. doi:10.1007/s00216-021-03705-w
32. Nikitovic, D.; Pérez, S. *Biomolecules* **2021,** *11,* 1630. doi:10.3390/biom11111630
33. Sankaranarayanan, N. V.; Nagarajan, B.; Desai, U. R. *Curr. Opin. Struct. Biol.* **2018,** *50,* 91–100. doi:10.1016/j.sbi.2017.12.004
34. Künze, G.; Huster, D.; Samsonov, S. A. *Biol. Chem.* **2021,** *402,* 1337–1355. doi:10.1515/hsz-2021-0119
35. Pągielska, M.; Samsonov, S. A. *Biomolecules* **2023,** *13,* 247. doi:10.3390/biom13020247
36. Li, Z.; Kienetz, M.; Cherney, M. M.; James, M. N. G.; Brömme, D. *J. Mol. Biol.* **2008,** *383,* 78–91. doi:10.1016/j.jmb.2008.07.038
37. Aguda, A. H.; Panwar, P.; Du, X.; Nguyen, N. T.; Brayer, G. D.; Brömme, D. *Proc. Natl. Acad. Sci. U. S. A.* **2014,** *111,* 17474–17479. doi:10.1073/pnas.1414126111
38. Huige, C. J. M.; Altona, C. *J. Comput. Chem.* **1995,** *16,* 56–79. doi:10.1002/jcc.540160106
39. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008,** *29,* 622–655. doi:10.1002/jcc.20820
40. Sattelle, B. M.; Shakeri, J.; Almond, A. *Biomacromolecules* **2013,** *14,* 1149–1159. doi:10.1021/bm400067g
41. Marcisz, M.; Gaardløs, M.; Bojarski, K. K.; Siebenmorgen, T.; Zacharias, M.; Samsonov, S. A. *J. Comput. Chem.* **2022,** *43,* 1633–1640. doi:10.1002/jcc.26965
42. *Amber 2021;* University of California: San Francisco, CA, USA, 2021.
43. Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1995,** *16,* 1339–1350. doi:10.1002/jcc.540161104
44. Grossfield, A. *WHAM: the weighted histogram analysis method*, Version 2.0.11. http://membrane.urmc.rochester.edu/wordpress/?page_id=126.
45. Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993,** *98,* 10089–10092. doi:10.1063/1.464397
46. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977,** *23,* 327–341. doi:10.1016/0021-9991(77)90098-5
47. Roe, D. R.; Cheatham, T. E., III. *J. Chem. Theory Comput.* **2013,** *9,* 3084–3095. doi:10.1021/ct400341p
48. Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002,** *23,* 1297–1304. doi:10.1002/jcc.10126
49. *R: A Language and Environment for Statistical Computing;* Foundation for Statistical Computing: Vienna,Austria, 2022.
50. Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996,** *14,* 33–38. doi:10.1016/0263-7855(96)00018-5
51. Samsonov, S. A.; Pisabarro, M. T. *Glycobiology* **2016,** *26,* 850–861. doi:10.1093/glycob/cww055
52. Bojarski, K. K.; Sieradzan, A. K.; Samsonov, S. A. *Biopolymers* **2019,** *110,* e23252. doi:10.1002/bip.23252

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Synthesis of the 3'-*O*-sulfated TF antigen with a TEG-N_3 linker for glycodendrimersomes preparation to study lectin binding

Mark Reihill, Hanyue Ma, Dennis Bengtsson and Stefan Oscarson*

Full Research Paper

Open Access

Address:
Centre for Synthesis and Chemical Biology, University College Dublin, Belfield, Dublin 4, Ireland

Email:
Stefan Oscarson* - stefan.oscarson@ucd.ie

* Corresponding author

Keywords:
regioselective sulfation; thioglycoside donors; Thomsen–Friedenreich antigen

Beilstein J. Org. Chem. **2024,** *20,* 173–180.
https://doi.org/10.3762/bjoc.20.17

Received: 07 November 2023
Accepted: 16 January 2024
Published: 30 January 2024

This article is part of the thematic issue "Chemical glycobiology".

Associate Editor: U. Westerlind

Abstract

The synthesis of gram quantities of the TF antigen (β-D-Gal-(1→3)-α-D-GalNAc) and its 3'-sulfated analogue with a TEG-N_3 spacer attached is described. The synthesis of the TF antigen comprises seven steps, from a known *N*-Troc-protected galactosamine donor, with an overall yield of 31%. Both the spacer (85%) and the galactose moiety (79%) were introduced using thioglycoside donors in NIS/AgOTf-promoted glycosylation reactions. The 3'-sulfate was finally introduced through tin activation in benzene/DMF followed by treatment with a sulfur trioxide–trimethylamine complex in a 66% yield.

Introduction

In a collaboration project with groups from Universities in Munich and Pennsylvania we are investigating carbohydrate–lectin interactions using programmable glycodendrimersomes based on synthetic glycans. We have earlier synthesized 2-[2-(2-azidoethoxy)ethoxy]ethyl (TEG-N_3) glycosides of lactose, 3'-Su-lactose and LacdiNAc (β-D-GalNAc-(1→4)-β-D-GlcNAc), which have then been used for production of the glycodenrimersomes and interaction studies with various galectins [1,2]. In the continuation of this collaboration, to investigate the binding of siglec-1 and the chimera of 3'-SuTF-binding siglecs and TF-binding galectin-3, TEG-N_3 glycosides of the TF antigen (β-D-Gal-(1→3)-α-D-GalNAc, **1**) and its 3'-*O*-sulfated analogue (**2**, Figure 1) were required on a gram scale to allow efficient synthesis of the glycodendrisomes. The TF antigen is presented on the surface of most human cancer cell types and its interaction with galectins 1 and 3 leads to tumour cell aggregation and promotes cancer metastasis [3-5]. The 3'-*O*-sulfated analogue is known to bind to siglecs 1, 4, and 8 [6] as well as galectin 4 [7,8], but its biological role is not that well investigated.

Compound **2** is a new compound but two syntheses of compound **1** have recently been reported, one using an enzymatic approach and a commercial α-TEG-N_3 GalNAc acceptor [9]

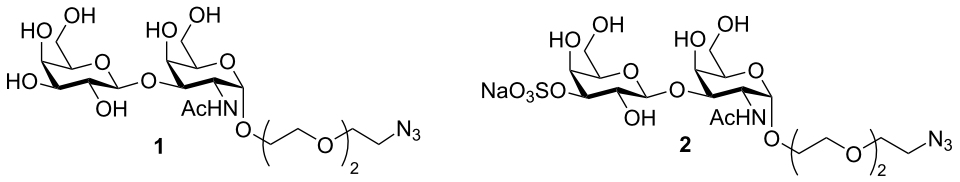


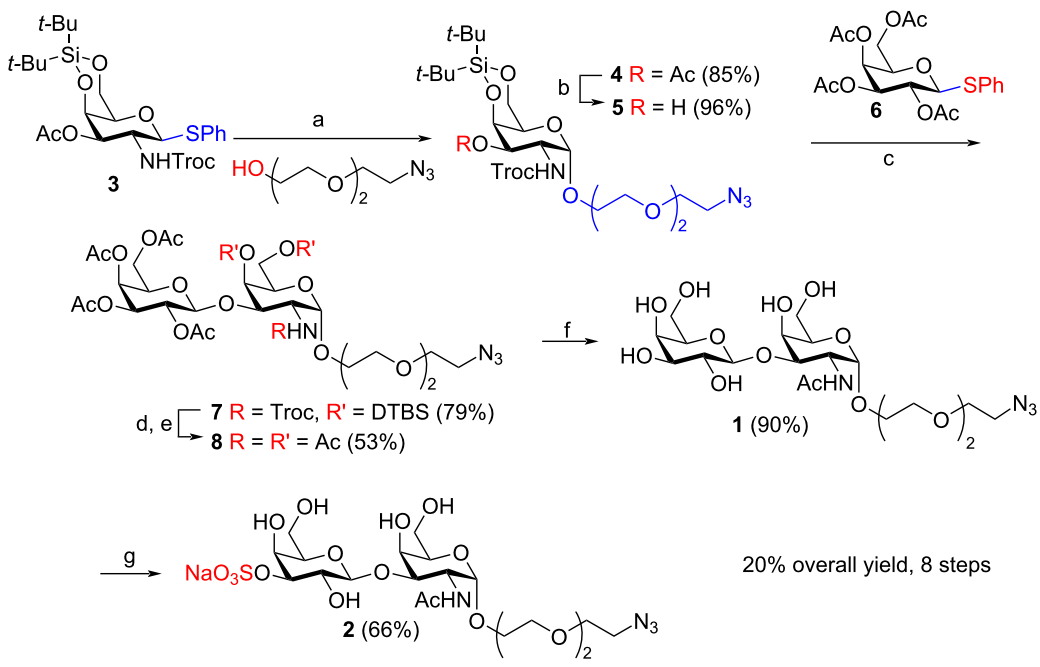
Figure 1: Structure of target compounds **1** and **2**.

and one using glycosyl bromide donors and silver salt-promoted glycosylations [10].

Results and Discussion

To introduce the 2-[2-(2-chloroethoxy)ethoxy]ethyl (TEG-Cl) spacer both a Fischer synthesis starting from unprotected *N*-acetylgalactosamine and a Lewis acid-promoted reaction starting from per-acetylated galactosamine were initially tested. As reported [11], the Fischer synthesis gives low yields and α-selectivity. The Lewis acid-promoted reaction, which had worked well to produce β-linked TEG-spacer glycosides with per-acetylated lactose and 2-phthalimidoglucosamine [1,2] worked well with 2-chloroethanol as a spacer (68%, pure α) but failed with the TEG-Cl spacer [12], why we instead decided to use a thioglycoside donor to introduce the spacer. To ensure α-selectivity a di-*tert*-butylsilyl-4,6-acetal-protected donor, as developed by the Kiso group [13,14], was chosen. After some initial testing the known *N*-Troc-protected donor **3** [15,16] (Scheme 1) was selected [17].

Since donor **3** possessed a Troc group, which contains 3 chlorine atoms, nucleophilic introduction of an azido group at this stage was predicted to be problematic. Therefore, the azido functionality was installed in the spacer before the glycosylation. Donor **3** underwent an NIS/AgOTf-promoted glycosylation with the TEG-N_3 acceptor [18], furnishing α-linked **4** in an 85% yield (Scheme 1). H-1 appeared as a doublet at 4.95 ppm with a *J* value of 3.6 Hz in the ^{1}H NMR spectrum proving the anomeric α-configuration. The presence of Troc-rotamers was also apparent, with a ratio of 19:6 being observed by ^{1}H NMR in $CDCl_3$ at 25 °C. Catalytic amounts of NaOMe (0.005 M) in MeOH were used to remove the acetate from compound **4**, taking care not to affect the Troc group, to afford acceptor **5** in a 96% yield.



Scheme 1: Synthesis of target compounds **1** and **2**. Key: a) NIS, AgOTf (20 mol %), 4 Å molecular sieves, CH_2Cl_2, rt, 40 min, 85%; b) NaOMe (10 mol %), MeOH, rt, 4 h, 96%; c) NIS, AgOTf (19 mol %), AW-300 4 Å molecular sieves, CH_2Cl_2, rt, 1 h, 79%; d) 1 M Bu_4NF/THF, THF, rt, 2 h, then HF·Py, rt, 3 h; e) Ac_2O/Py (1:2, v/v), rt, 16 h, 53% over 3 steps; f) 1 M NaOMe/MeOH, MeOH, pH 10, rt, 1 h, 90%; g) Bu_2SnO, benzene/DMF (5:1, v/v), 125 °C, 24 h, then SO_3-NMe_3, DMF, rt, 72 h, then flash chromatography, then Dowex® 50WX4 (Na^+ form) resin, H_2O, rt, 16 h, 66%.

Earlier optimizations of the introduction of the β-linked galactose moiety using 2-azidoethyl 2-acetamido-4,6-*O*-benzylidene-2-deoxy-α-D-galactopyranoside as acceptor showed an acetylated thioglycoside donor to be the best choice [12], surprisingly better than a benzoylated donor [19], why this donor was the first one tested also with the quite different acceptor **5**. An NIS/AgOTf-promoted glycosylation with donor **6** [20] yielded 79% of disaccharide **7**. Due to the presence of rotamers, NMR spectra of **7** proved to be difficult to analyse when data were recorded in $CDCl_3$. Changing the NMR solvent to CD_3OD greatly reduced the complexity of the spectra [21-23].

Since **7** possessed an azido group as part of the linker, removal of the Troc group under reductive conditions was ruled out due to probable chemoselectivity issues [24,25]. Interestingly, Jacquemard et al. outlined a useful, mild method for removing a range of carbamates using Bu_4NF in an article from 2004 [26]. As **7** contained a DTBS group, the possibility of removing both Troc and DTBS groups in a one-pot procedure was tested. Disaccharide **7** was therefore treated with 1 M Bu_4NF/THF and after 2 hours, full consumption of the starting material was observed by TLC. However, MALDI–TOF mass spectrometry (super-DHB matrix) revealed that only the Troc group had been removed, with the DTBS substituent proving to be stable under these conditions. Addition of a large excess of HF·Py (40 equiv) proved to be necessary to remove the bulky silyl group. After concentration, the crude product was acetylated (Ac_2O/Py, 1:2, v/v), furnishing per-acetylated compound **8** in a 53% yield over the 3 steps. Deacetylation of **8** with freshly prepared 1 M NaOMe/MeOH in MeOH at pH 10 furnished target **1** in a 90% yield.

Formation of a stannylidene acetal via tin-activation was employed to achieve selective 3'-*O*-sulfation of compound **1** [27], with a variety of conditions being attempted (Table 1). With a TEG-N_3 lactose compound, tin-activation was performed with Bu_2SnO in refluxing MeOH, followed by stirring with $SO_3{\cdot}NMe_3$ in 1,4-dioxane to afford the 3'-*O*-sulfate in 65% yield [1]. Here, however, this choice of solvent in the sulfation step led to the material being insoluble and no reaction was observable by TLC. Changing the solvent of the sulfation reaction to DMF resulted in formation of a homogenous solution, but still no conversion to the sulfated product, even when the temperature was raised to 80 °C [28,29]. Switching the sulfating reagent to $SO_3{\cdot}Py$ or performing the reaction at 150 °C in a microwave did not improve the outcome [30,31].

Since there was no observable sulfation taking place, the tin-activation step was suspected to be the root of the problem. To rectify this, similar to Malleron et al., **1** was refluxed, in a Dean–Stark set-up, with Bu_2SnO in benzene/DMF (5:1, v/v) [32]. The solvent in the receiver was drained after 24 hours and the benzene was removed from the reaction mixture in vacuo. Sulfation was then performed through addition of $SO_3{\cdot}NMe_3$ to the DMF solution. Consumption of **1** was observed by TLC after 72 hours and stirring with Dowex® 50WX4 (Na^+ form) resin resulted in formation of target **2**. Purification by flash chromatography, however, led to isolation of a mixture of **2** and a tin-related impurity (*n*-butyl chain evident by NMR). Acetylation of this material followed by flash chromatography proved ineffective in removing the unwanted entity. To overcome this problem, flash chromatography was performed before stirring with the ion-exchange resin, with no apparent presence of tin impurities by NMR when the sequence was executed in this order and sulfated target **2** was obtained in a 66% yield on a one-gram scale. Comparing the ^1H,^{13}C HSQC spectra of compounds **1** and **2**, there is a clear downfield shift of the H-3'/C-3' signal from **1** to sulfated **2** (Figure 2). This showed that regioselective 3'-*O*-sulfation had been achieved, with HRMS also indicating that only one sulfate group was present.

Table 1: Summary of conditions attempted to achieve regioselective 3'-*O*-sulfation.

Tin-activation[a]			Sulfation[a]			
Solvent(s)	Temperature	Set-up	Sulfating reagent	Solvent	Temperature/conditions	Result/ yield
MeOH	95 °C	reflux	$SO_3{\cdot}NMe_3$	1,4-dioxane	rt	material insoluble, no reaction
MeOH	95 °C	reflux	$SO_3{\cdot}NMe_3$	DMF	rt–80 °C	no reaction
MeOH	95 °C	reflux	$SO_3{\cdot}Py$	DMF	80 °C	no reaction
MeOH	95 °C	reflux	$SO_3{\cdot}Py$	DMF	150 °C, microwave	no reaction
benzene/DMF (5:1, v/v)	125 °C	reflux, Dean–Stark	$SO_3{\cdot}NMe_3$	DMF	rt	66%

[a]Tin-activation was performed with 1.2 equiv of Bu_2SnO in all cases for 16–24 h and sulfation reactions proceeded for 24–72 h.

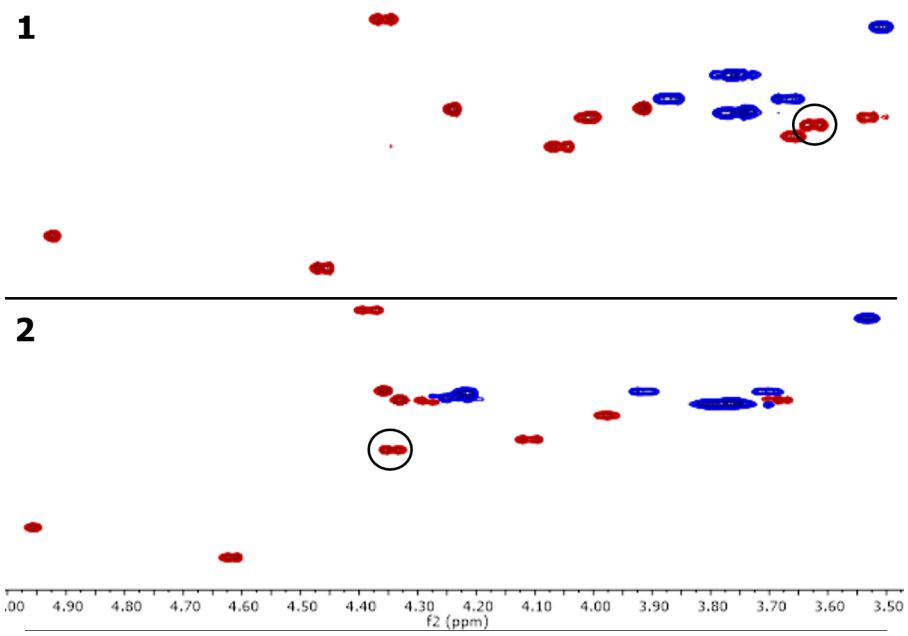


Figure 2: Comparison of the $^{1}H,^{13}C$ HSQC spectra of **1** (top) and 3'-*O*-sulfated **2** (bottom), with circles highlighting the signals corresponding to H-3'/C-3'.

Conclusion

An efficient synthesis of the important TF and 3'-Su-TF antigens equipped with a TEG-N_3 linker to allow formation of various conjugates has been developed for further interaction studies with lectins (galectins and siglecs). The synthesis of the 3'-Su-TF antigen **2** comprises eight steps from the known *N*-galactosamine donor **3**, where two of the steps, removal of the Troc- and DTBS protecting groups are performed in the same pot and the following acetylation without purification of the intermediate, why the synthesis is high-yielding (20% overall yield) and easily scalable (9 g of protected disaccharide **7** and 1 gram of target **2** were synthesized).

Experimental

General methods

All reactions containing air- and moisture-sensitive reagents were carried out under an inert atmosphere of nitrogen in oven-dried glassware with magnetic stirring. N_2-flushed plastic syringes were used to transfer air- and moisture-sensitive reagents. All reactions were monitored by thin-layer chromatography (TLC) on Merck® DC-Alufolien plates precoated with silica gel 60 F_{254}. Visualisation was performed with UV-light (254 nm) fluorescence quenching, and/or by staining with an 8% H_2SO_4 dip (stock solution: 8 mL conc. H_2SO_4, 92 mL EtOH), ninhydrin dip (stock solution: 0.3 g ninhydrin, 3 mL AcOH, 100 mL EtOH) and/or ceric ammonium molybdate dip (stock solution: 25 g ammonium molybdate tetrahydrate, 0.5 g $Ce(SO_4)_2$, 50 mL H_2SO_4, 450 mL EtOH).

Chromatography

Silica gel flash chromatography was carried out using Davisil® LC60A (40–63 µm) silica gel or with automated flash chromatography systems, Buchi Reveleris® X2 (UV 200–500 nm and ELSD detection, Reveleris® silica cartiges 40 µm, Büchi Labortechnik AG®) and Biotage® SP4 HPFC (UV 200–500 nm, Biotage® SNAP KP-Sil 50 µm irregular silica, Biotage® AB).

Instrumentation

^{1}H NMR and ^{13}C NMR spectra were recorded on Varian Inova spectrometers at 25 °C in chloroform-*d* ($CDCl_3$), methanol-d_4 (CD_3OD), deuterium oxide (D_2O) or DMSO-d_6 ($(CD_3)_2SO$). ^{1}H NMR spectra were standardised against the residual solvent peak ($CDCl_3$, δ = 7.26 ppm; CD_3OD, δ = 3.31 ppm; D_2O, δ = 4.79 ppm; $(CD_3)_2SO$ δ = 2.50 ppm); or internal trimethylsilane, δ = 0.00 ppm). ^{13}C NMR spectra were standardised against the residual solvent peak ($CDCl_3$, δ = 77.16 ppm; CD_3OD, δ = 49.00 ppm; $(CD_3)_2SO$ δ = 39.52 ppm and ^{13}C NMR spectra recorded in D_2O are unreferenced. All ^{13}C NMR spectra are ^{1}H decoupled. All NMR data are represented as follows: chemical shift (δ ppm), multiplicity (s = singlet, br s = broad singlet, d = doublet, app d = apparent doublet, t = triplet, q = quartet, dd = doublet of doublets, dt = doublet of triplets, m = multiplet), coupling constant in Hz, integration. Assignments were aided by homonuclear $^{1}H,^{1}H$ (COSY, TOCSY) and $^{1}H,^{13}C$ heteronuclear (HSQC, HMBC) two-dimensional correlation spectroscopies. ^{13}C chem-

ical shifts were reported to one decimal point unless an additional digit was required to distinguish overlapping peaks. Software for data processing: MestReNova, version 11.0.0–17609 (MestReLab Research S.L.). High-resolution mass spectrometry (HRMS) data were recorded on a Waters Micromass LCT LC–TOF instrument using electrospray ionisation (ESI) in positive mode. MALDI–TOF mass spectrometry data were recorded on a Scientific Analysis Instruments MALDI–TOF mass spectrometer in reflectron mode for oligosaccharides and in linear mode for glycoconjugates. Samples were prepared by pre-mixing 1 μL of a solution containing the analyte with 20 μL of a matrix solution (10 mg/mL, MeCN/H_2O, 1:1, v/v + 1% TFA), pipetting 1 μL of the mixture onto the sample plate and drying under gentle heat from a heat gun. Optical rotations were recorded in a Perkin-Elmer polarimeter (Model 343) at the sodium D-line (589 nm) at 20 °C using a 1 dm cell. Samples were prepared at the concentration (g/100 mL) in the solvent indicated. Deprotected glycans were lyophilised using a freeze-dryer Alpha 1-2 LDplus (Christ Ltd): pressure: 0.055 mbar; ice-condenser temperature: −55 °C.

2-[2-(2-Azidoethoxy)ethoxy]ethyl 3-*O*-acetyl-2-deoxy-4,6-*O*-di-*tert*-butylsilylene-2-(2'2'2'-trichloroethoxycarbonylamino)-α-D-galactopyranoside (4): Donor **3** [9,10] (9.3 g, 15 mmol) and the TEG-N_3 acceptor (synthesized as described in reference [12], but also commercially available, 3.9 g, 22 mmol) were placed under N_2 together and dissolved in dry CH_2Cl_2 (300 mL). 4 Å molecular sieves (10.2 g) were added and the resulting suspension was stirred at room temperature for 16 hours. NIS (6.66 g, 29.6 mmol) and AgOTf (760 mg, 2.96 mmol) were then added, and the reaction was stirred at room temperature for 40 minutes. The reaction was then quenched with Et_3N, filtered through Celite® and concentrated in vacuo. The resulting residue was taken up in EtOAc (700 mL) and washed with 10% aq $Na_2S_2O_3$ (700 mL), water (700 mL) and brine (700 mL). The organic phase was then dried over $MgSO_4$, filtered and reduced to dryness. Flash chromatography on silica gel (toluene→toluene/EtOAc, 3:2) yielded **4** as an orange syrup (8.74 g, 85%). R_f = 0.4 (toluene/EtOAc, 7:3); $[\alpha]_D$ +92 (*c* 1.0, $CHCl_3$); ^{1}H NMR (500 MHz, $CDCl_3$) δ 5.44 (d, *J* = 10.2 Hz, 1H, NH), 4.98 (dd, *J* = 11.1, 2.9 Hz, 1H, H-3), 4.95 (d, *J* = 3.6 Hz, 1H, H-1), 4.87 (d, *J* = 12.1 Hz, 1H, $CH_{2(A)Troc}$), 4.69–4.58 (m, 2H, H-4, $CH_{2(B)Troc}$), 4.49 (td, *J* = 10.6, 3.6 Hz, 1H, H-2), 4.26 (dd, *J* = 12.6, 2.2 Hz, 1H, H-$6_{(A)}$), 4.15 (dd, *J* = 12.5, 1.7 Hz, 1H, H-$6_{(B)}$), 3.90–3.77 (m, 2H, H-5, $CH_{2(A)Linker}$), 3.76–3.59 (m, 9H, $CH_{2(B)Linker}$, 4 × $CH_{2(Linker)}$), 3.39 (t, *J* = 5.1 Hz, 2H, $CH_{2(Linker)}$), 2.07 (s, 3H, $CH_{3(OAc)}$), 1.08 (s, 9H, $C(CH_3)_{3(DTBS)}$), 1.02 (s, 9H, $C(CH_3)_{3(DTBS)}$); ^{13}C NMR (126 MHz, $CDCl_3$) δ 171.3 ($C{=}O_{(OAc)}$), 154.6 ($C{=}O_{(Troc)}$), 98.5 (C-1), 95.8 ($CCl_{3(Troc)}$), 74.7 ($CH_{2(Troc)}$), 71.7 (C-3), 70.9 ($CH_{2(Linker)}$), 70.8 ($CH_{2(Linker)}$), 70.5 (C-4), 70.27 ($CH_{2(Linker)}$), 70.25 ($CH_{2(Linker)}$), 67.63 ($CH_{2(Linker)}$), 67.57 (C-5), 67.1 (C-6), 50.8 ($CH_{2(Linker)}$), 49.3 (C-2), 27.7 ($C(CH_3)_{3(DTBS)}$), 27.4 ($C(CH_3)_{3(DTBS)}$), 23.4 ($C(CH_3)_{3(DTBS)}$), 21.1 ($CH_{3(OAc)}$), 20.9 ($C(CH_3)_{3(DTBS)}$); HRESIMS *m*/*z*: $[M + NH_4]^+$ calcd for $C_{25}H_{43}Cl_3N_4O_{10}Si$, 710.2158; found, 710.2158.

2-[2-(2-Azidoethoxy)ethoxy]ethyl 2-deoxy-4,6-*O*-di-*tert*-butylsilylene-2-(2'2'2'-trichloroethoxycarbonylamino)-α-D-galactopyranoside (5): Compound **4** (8.65 g, 12.5 mmol) was placed under N_2 and dissolved in dry MeOH (250 mL). NaOMe (68 mg, 1.3 mmol) was added, and the reaction was stirred at room temperature for 4 hours. The solution was then neutralised with Amberlite® IR120 (H^+ form) resin, filtered and concentrated under reduced pressure. Flash chromatography on silica gel (toluene→toluene/acetone, 7:3) yielded **5** as a gold-coloured syrup (7.83 g, 96%). R_f = 0.4 (toluene/EtOAc, 3:2); $[\alpha]_D$ +67 (*c* 1.0, $CHCl_3$); ^{1}H NMR (500 MHz, $CDCl_3$) δ 5.58 (d, *J* = 9.8 Hz, 1H, NH), 4.93 (d, *J* = 3.6 Hz, 1H, H-1), 4.77 (d, *J* = 12.0 Hz, 1H, $CH_{2(A)Troc}$), 4.72 (d, *J* = 12.0 Hz, 1H, $CH_{2(B)Troc}$), 4.43 (d, *J* = 3.0 Hz, 1H, H-4), 4.28 (dd, *J* = 12.5, 2.2 Hz, 1H, H-$6_{(A)}$), 4.16 (m, 1H, H-$6_{(B)}$), 4.11 (dd, *J* = 10.1, 3.6 Hz, 1H, H-2), 3.87–3.79 (m, 2H, H-5, $CH_{2(A)Linker}$), 3.74 (dd, *J* = 11.4, 3.2 Hz, 1H, H-3), 3.70–3.63 (m, 9H, $CH_{2(B)Linker}$, 4 × $CH_{2(Linker)}$), 3.43–3.34 (m, 2H, $CH_{2(Linker)}$), 2.53 (d, *J* = 11.8 Hz, 1H, OH), 1.07 (s, 9H, $C(CH_3)_{3(DTBS)}$), 1.05 (s, 9H, $C(CH_3)_{3(DTBS)}$); ^{13}C NMR (126 MHz, $CDCl_3$) δ 155.4 ($C{=}O_{(Troc)}$), 98.6 (C-1), 95.7 ($CCl_{3(Troc)}$), 74.9 ($CH_{2(Troc)}$), 73.0 (C-4), 70.9 ($CH_{2(Linker)}$), 70.7 ($CH_{2(Linker)}$), 70.27 ($CH_{2(Linker)}$), 70.24 ($CH_{2(Linker)}$), 70.14 (C-3), 67.8 (C-5), 67.7 ($CH_{2(Linker)}$), 67.2 (C-6), 52.5 (C-2), 50.8 ($CH_{2(Linker)}$), 27.7 ($C(CH_3)_{3(DTBS)}$), 27.5 ($C(CH_3)_{3(DTBS)}$), 23.5 ($C(CH_3)_{3(DTBS)}$), 20.9 ($C(CH_3)_{3(DTBS)}$); HRESIMS *m*/*z*: $[M + Na]^+$ calcd for $C_{23}H_{41}Cl_3N_4O_9Si$, 673.1606; found, 673.1605.

2-[2-(2-Azidoethoxy)ethoxy]ethyl 2,3,4,6-tetra-*O*-acetyl-β-D-galactopyranosyl-(1→3)-2-deoxy-4,6-*O*-di-*tert*-butylsilylene-2-(2'2'2'-trichloroethoxycarbonylamino)-α-D-galactopyranoside (7): Donor **6** [16] (7.67 g, 17.4 mmol) and acceptor **5** (7.57 g, 11.6 mmol) were placed under N_2 together and dissolved in dry CH_2Cl_2 (230 mL). AW-300 4 Å molecular sieves (5.45 g) were added, and the resulting suspension was stirred at room temperature for 23 hours. NIS (5.23 g, 23.2 mmol) and AgOTf (577 mg, 2.25 mmol) were added, and the reaction was stirred at room temperature for 1 hour. Et_3N was then added until the pH became neutral, and the suspension was filtered through Celite®. The filtrate was then washed with water (400 mL), brine (400 mL), dried over $MgSO_4$, filtered and concentrated in vacuo. Compound **7** was isolated by flash chromatography on silica gel (toluene→toluene/EtOAc, 1:4) as an orange foam (8.97 g, 79%). R_f = 0.5 (toluene/EtOAc, 3:7); $[\alpha]_D$

+71 (*c* 1.0, CH_3OH); ^{1}H NMR (500 MHz, CD_3OD) δ 5.40 (d, *J* = 3.4 Hz, 1H, H-4_{Gal}), 5.21 (m, 1H, H-2_{Gal}), 5.11 (dd, *J* = 10.4, 3.4 Hz, 1H, H-3_{Gal}), 5.05 (d, *J* = 12.2 Hz, 1H, $CH_{2(A)Troc}$), 4.93 (d, *J* = 3.6 Hz, 1H, H-$1_{GalNTroc}$), 4.88 (d, *J* = 7.8 Hz, 1H, H-1_{Gal}), 4.79 (d, *J* = 2.8 Hz, 1H, H-$4_{GalNTroc}$), 4.58 (d, *J* = 12.2 Hz, 1H, $CH_{2(B)Troc}$), 4.38 (dd, *J* = 11.1, 3.6 Hz, 1H, H-$2_{GalNTroc}$), 4.30 (m, 1H, H-$6_{(A)}$), 4.22–4.05 (m, 4H, H-5_{Gal}, H-$6_{(B)}$, H-$6_{(A+B)}$), 3.96 (dd, *J* = 11.1, 2.8 Hz, 1H, H-$3_{GalNTroc}$), 3.90–3.79 (m, 2H, H-$5_{GalNTroc}$, $CH_{2(A)Linker}$), 3.77–3.63 (m, 9H, $CH_{2(B)Linker}$, 4 × $CH_{2(Linker)}$), 3.46–3.39 (m, 2H, $CH_{2(Linker)}$), 2.16 (s, 3H, $CH_{3(OAc)}$), 2.12 (s, 3H, $CH_{3(OAc)}$), 2.05 (s, 3H, $CH_{3(OAc)}$), 1.97 (s, 3H, $CH_{3(OAc)}$), 1.13–1.09 (m, 18H, 2 × $C(CH_3)_{3(DTBS)}$); ^{13}C NMR (126 MHz, CD_3OD) δ 171.99 ($C{=}O_{(OAc)}$), 171.93 ($C{=}O_{(OAc)}$), 171.6 ($C{=}O_{(OAc)}$), 171.4 ($C{=}O_{(OAc)}$), 156.6 ($C{=}O_{(Troc)}$), 103.8 (C-1_{Gal}), 99.6 (C-$1_{GalNTroc}$), 97.2 ($CCl_{3(Troc)}$), 79.3 (C-$3_{GalNTroc}$), 75.6 ($CH_{2(Troc)}$), 74.1 (C-$4_{GalNTroc}$), 72.6 (C-3_{Gal}), 71.8 (C-5_{Gal}), 71.52 ($CH_{2(Linker)}$), 71.49 ($CH_{2(Linker)}$), 71.1 ($CH_{2(Linker)}$), 70.5 (C-2_{Gal}), 69.0 (C-$5_{GalNTroc}$), 68.6 (C-4_{Gal}), 68.4 (C-6), 68.1 ($CH_{2(Linker)}$), 62.9 (C-6), 51.8 ($CH_{2(Linker)}$), 51.4 (C-$2_{GalNTroc}$), 28.2 ($C(CH_3)_{3(DTBS)}$), 28.0 ($C(CH_3)_{3(DTBS)}$), 24.3 ($C(CH_3)_{3(DTBS)}$), 21.7 ($C(CH_3)_{3(DTBS)}$), 21.0 ($CH_{3(OAc)}$), 20.6 ($CH_{3(OAc)}$), 20.51 ($CH_{3(OAc)}$), 20.48 ($CH_{3(OAc)}$). HRESIMS *m*/*z*: $[M + NH_4]^+$ calcd for $C_{37}H_{59}Cl_3N_4O_{18}Si$, 998.3003; found, 998.3003.

2-[2-(2-Azidoethoxy)ethoxy]ethyl 2,3,4,6-tetra-*O*-acetyl-β-D-galactopyranosyl-(1→3)-2-acetamido-4,6-di-*O*-acetyl-2-deoxy-α-D-galactopyranoside [3] **(8):** Compound **7** (8.87 g, 9.03 mmol) was placed under N_2 and dissolved in dry THF (180 mL). 1 M Bu_4NF/THF (32 mL, 32 mmol) was added, and the reaction was stirred at room temperature. After 2 hours, the starting material had been consumed (judged by TLC) and HF·Py (70% HF, 9.5 mL, 370 mmol) was added. Stirring was continued at room temperature for a further 3 hours and the solution was then concentrated.

The crude was placed under N_2 and stirred at room temperature in Ac_2O/Py (180 mL, 1:2, v/v) for 16 hours. The solution was then reduced to dryness and purification by flash chromatography on silica gel (EtOAc→EtOAc/MeOH, 17:3) yielded **8** as a dark orange/brown syrup (3.82 g, 53% over 3 steps). R_f = 0.4 (EtOAc/MeOH, 19:1); $[\alpha]_D$ +76 (*c* 1.0, CH_3OH); ^{1}H NMR (500 MHz, CD_3OD) δ 5.42 (d, *J* = 3.1 Hz, 1H, H-4_{GalNAc}), 5.36 (dd, *J* = 3.4, 1.2 Hz, 1H, H-4_{Gal}), 5.06 (dd, *J* = 10.5, 3.4 Hz, 1H, H-3_{Gal}), 5.00 (dd, *J* = 10.5, 7.6 Hz, 1H, H-2_{Gal}), 4.83 (d, *J* = 3.8 Hz, 1H, H-1_{GalNAc}), 4.78 (d, *J* = 7.6 Hz, 1H, H-1_{Gal}), 4.43 (dd, *J* = 11.1, 3.6 Hz, 1H, H-2_{GalNAc}), 4.26 (m, 1H, H-5_{GalNAc}), 4.21–4.11 (m, 3H, H-$6_{(A+B)Gal}$, H-$6_{(A)GalNAc}$), 4.08 (dd, *J* = 11.1, 3.4 Hz, 1H, H-3_{GalNAc}), 4.04 (m, 1H, H-5_{Gal}), 3.97 (dd, *J* = 11.3, 7.3 Hz, 1H, H-$6_{(B)GalNAc}$), 3.81 (m, 1H, $CH_{2(A)Linker}$), 3.73–3.71 (m, 2H, $CH_{2(Linker)}$), 3.70–3.62 (m, 7H, 3 × $CH_{2(Linker)}$, $CH_{2(B)Linker}$), 3.42–3.36 (m, 2H, $CH_{2(Linker)}$), 2.14 (s, 3H, $CH_{3(Ac)}$), 2.11 (s, 3H, $CH_{3(Ac)}$), 2.06–2.02 (m, 9H, 3 × $CH_{3(Ac)}$), 1.99 (s, 3H, $CH_{3(Ac)}$), 1.93 (s, 3H, $CH_{3(Ac)}$); ^{13}C NMR (126 MHz, CD_3OD) δ 173.1, 172.3, 172.08, 172.06, 172.04, 171.5, 171.2 ($C{=}O_{(Ac)}$), 102.4 (C-1_{Gal}), 99.3 (C-1_{GalNAc}), 74.6 (C-3_{GalNAc}), 72.2 (C-3_{Gal}), 71.8 ($CH_{2(Linker)}$), 71.54 (C-5_{Gal}), 71.49 ($CH_{2(Linker)}$), 71.36 ($CH_{2(Linker)}$), 71.3 ($CH_{2(Linker)}$), 71.2 (C-4_{GalNAc}), 70.2 (C-2_{Gal}), 68.6 (C-4_{Gal}), 68.5 (C-5_{GalNAc}), 68.2 ($CH_{2(Linker)}$), 63.9 (C-6_{GalNAc}), 62.4 (C-6_{Gal}), 51.7 ($CH_{2(Linker)}$), 50.3 (C-2_{GalNAc}), 22.89, 20.82, 20.77, 20.73, 20.67, 20.50, 20.47 ($CH_{3(Ac)}$). As NMR spectra in the literature were recorded in $CDCl_3$ [3], NMR data are not comparable. HRESIMS *m*/*z*: $[M + Na]^+$ calcd for $C_{32}H_{48}N_4O_{19}$; 815.2810; found; 815.2806.

2-[2-(2-Azidoethoxy)ethoxy]ethyl β-D-galactopyranosyl-(1→3)-2-acetamido-2-deoxy-α-D-galactopyranoside [3,4] **(1):** Compound **8** (1.47 g, 1.85 mmol) was dissolved in MeOH (100 mL) and freshly prepared 1 M NaOMe/MeOH was added until the solution reached pH 10. The reaction was stirred at room temperature for 1 hour, then neutralised with Amberlite® IR120 (H^+ form) resin. The resin was filtered off, washed with MeOH and the filtrate was concentrated in vacuo. After lyophilisation, **1** was obtained as a light brown/orange solid (900 mg, 90%) and required no further purification. R_f = 0.6 (EtOAc/MeOH, 2:3); $[\alpha]_D$ +76 (*c* 1.0, H_2O); ^{1}H NMR (500 MHz, D_2O) δ 4.92 (d, *J* = 3.7 Hz, 1H, H-1_{GalNAc}), 4.46 (d, *J* = 7.8 Hz, 1H, H-1_{Gal}), 4.36 (dd, *J* = 11.0, 3.7 Hz, 1H, H-2_{GalNAc}), 4.24 (d, *J* = 3.0 Hz, 1H, H-4_{GalNAc}), 4.06 (dd, *J* = 11.1, 3.1 Hz, 1H, H-3_{GalNAc}), 4.01 (m, 1H, H-5), 3.92 (d, *J* = 3.4 Hz, 1H, H-4_{Gal}), 3.87 (m, 1H, $CH_{2(A)Linker}$), 3.81–3.71 (m, 12H, 2 × H-$6_{(A+B)}$, 4 × $CH_{2(Linker)}$), 3.70–3.59 (m, 3H, H-3_{Gal}, H-5, $CH_{2(B)Linker}$), 3.56–3.48 (m, 3H, H-2_{Gal}, $CH_{2(Linker)}$), 2.04 (s, 3H, $CH_{3(NHAc)}$); ^{13}C NMR (126 MHz, D_2O) δ 174.5 ($C{=}O_{(NHAc)}$), 104.7 (C-1_{Gal}), 97.4 (C-1_{GalNAc}), 77.3 (C-3_{GalNAc}), 74.9 (C-5), 72.5 (C-3_{Gal}), 70.64 (C-5), 70.54 (C-2_{Gal}), 69.7 ($CH_{2(Linker)}$), 69.51 ($CH_{2(Linker)}$), 69.46 ($CH_{2(Linker)}$), 69.2 ($CH_{2(Linker)}$), 68.7 (C-4_{GalNAc}), 68.5 (C-4_{Gal}), 66.5 ($CH_{2(Linker)}$), 61.1 (C-6), 60.9 (C-6), 50.1 ($CH_{2(Linker)}$), 48.5 (C-2_{GalNAc}), 22.0 ($CH_{3(NHAc)}$). NMR data match those reported in the literature [3,4]. HRESIMS *m*/*z*: $[M + H]^+$ calcd for $C_{20}H_{36}N_4O_{13}$, 541.2357; found, 541.2354.

2-[2-(2-Azidoethoxy)ethoxy]ethyl 3-*O*-sulfo-β-D-galactopyranosyl-(1→3)-2-acetamido-2-deoxy-α-D-galactopyranoside sodium salt (2): Compound **1** (1.24 g, 2.29 mmol) and Bu_2SnO (645 mg, 2.75 mmol) were placed under N_2 together. Dry benzene/DMF (380 mL, 5:1, v/v) was added and the reaction was refluxed at 125 °C using a Dean–Stark apparatus. After 24 hours, the solvent in the receiver was drained, and the

benzene was removed from the reaction mixture in vacuo. $SO_3{\cdot}NMe_3$ (642 mg, 4.61 mmol) was then added to the DMF solution, and the reaction was stirred at room temperature. After 24 hours, an additional portion of $SO_3{\cdot}NMe_3$ (950 mg, 6.83 mmol) was added and stirring was continued at room temperature for a further 48 hours. The reaction mixture was then concentrated and flash chromatography on silica gel (EtOAc/MeOH, 1:0→0:1) yielded a yellow syrup, which was re-dissolved in H_2O (30 mL). Dowex® 50WX4 (Na^+ form) resin (1.28 g) was added, and the resulting suspension was stirred at room temperature for 16 hours. Filtration followed by concentration and lyophilisation of the filtrate yielded **2** as a pale-yellow foam (972 mg, 66%). R_f = 0.3 (EtOAc/MeOH, 3:2); $[\alpha]_D$ +76 (*c* 1.0, H_2O); ^{1}H NMR (500 MHz, D_2O) δ 4.95 (d, *J* = 3.8 Hz, 1H, H-1_{GalNAc}), 4.62 (d, *J* = 7.9 Hz, 1H, H-1_{Gal}), 4.41–4.31 (m, 4H, H-2_{GalNAc}, H-3_{Gal}, H-4_{Gal}, H-4_{GalNAc}), 4.30–4.16 (m, 5H, H-5, 2 × H-$6_{(A+B)}$), 4.11 (dd, *J* = 11.1, 3.1 Hz, 1H, H-3_{GalNAc}), 3.97 (t, *J* = 6.2 Hz, 1H, H-5), 3.91 (m, 1H, $CH_{2(A)Linker}$), 3.85–3.62 (m, 10H, H-2_{Gal}, $CH_{2(B)Linker}$, 4 × $CH_{2(Linker)}$), 3.56–3.50 (m, 2H, $CH_{2(Linker)}$), 2.05 (s, 3H, $CH_{3(NHAc)}$); ^{13}C NMR (126 MHz, D_2O) δ 174.5 ($C{=}O_{(NHAc)}$), 104.2 (C-1_{Gal}), 97.4 (C-1_{GalNAc}), 79.9 (C-3_{Gal}), 77.5 (C-3_{GalNAc}), 72.1 (C-5), 69.6 ($CH_{2(Linker)}$), 69.47 ($CH_{2(Linker)}$), 69.38 ($CH_{2(Linker)}$), 69.2 ($CH_{2(Linker)}$), 68.8 (C-5), 68.55 (C-2_{Gal}), 68.52 (C-4_{GalNAc}), 68.3 (C-6), 66.9 (C-6), 66.7 ($CH_{2(Linker)}$), 66.5 (C-4_{Gal}), 50.1 ($CH_{2(Linker)}$), 48.3 (C-2_{GalNAc}), 22.0 ($CH_{3(NHAc)}$). HRESIMS *m*/*z*: $[M - Na + 2H]^+$ calcd for $C_{20}H_{37}N_4O_{16}S$, 621.1925; found, 621.1920.

Supporting Information

Supporting Information File 1

NMR spectra of compounds **1**–**5**, **7** and **8**.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-17-S1.pdf]

Acknowledgements

We thank Dr Yannick Ortin and Dr Jimmy Muldoon for NMR and MS support.

Funding

The authors kindly acknowledge Science Foundation Ireland awards 13/IA/1959 and 20/FFP-P/884 (to S.O) and a China Scholarship Council Ph.D. scholarship (to H. M).

ORCID® iDs

Mark Reihill - https://orcid.org/0000-0003-3896-9346
Hanyue Ma - https://orcid.org/0009-0009-7307-0388
Stefan Oscarson - https://orcid.org/0000-0002-8273-4918

References

1. Xiao, Q.; Ludwig, A.-K.; Romanò, C.; Buzzacchera, I.; Sherman, S. E.; Vetro, M.; Vértesy, S.; Kaltner, H.; Reed, E. H.; Möller, M.; Wilson, C. J.; Hammer, D. A.; Oscarson, S.; Klein, M. L.; Gabius, H.-J.; Percec, V. *Proc. Natl. Acad. Sci. U. S. A.* **2018,** *115,* E2509–E2518. doi:10.1073/pnas.1720055115
2. Ludwig, A.-K.; Michalak, M.; Xiao, Q.; Gilles, U.; Medrano, F. J.; Ma, H.; FitzGerald, F. G.; Hasley, W. D.; Melendez-Davila, A.; Liu, M.; Rahimi, K.; Kostina, N. Y.; Rodriguez-Emmenegger, C.; Möller, M.; Lindner, I.; Kaltner, H.; Cudic, M.; Reusch, D.; Kopitz, J.; Romero, A.; Oscarson, S.; Klein, M. L.; Gabius, H.-J.; Percec, V. *Proc. Natl. Acad. Sci. U. S. A.* **2019,** *116,* 2837–2842. doi:10.1073/pnas.1813515116
3. Ju, T.; Otto, V. I.; Cummings, R. D. *Angew. Chem., Int. Ed.* **2011,** *50,* 1770–1791. doi:10.1002/anie.201002313
4. Yu, L.-G. *Glycoconjugate J.* **2007,** *24,* 411–420. doi:10.1007/s10719-007-9034-3
5. Zhao, Q.; Guo, X.; Nash, G. B.; Stone, P. C.; Hilkens, J.; Rhodes, J. M.; Yu, L.-G. *Cancer Res.* **2009,** *69,* 6799–6806. doi:10.1158/0008-5472.can-09-1096
6. Rapoport, E. M.; Pazynina, G. V.; Sablina, M. A.; Crocker, P. R.; Bovin, N. V. *Biochemistry (Moscow)* **2006,** *71,* 496–504. doi:10.1134/s0006297906050051
7. Ideo, H.; Seko, A.; Ohkura, T.; Matta, K. L.; Yamashita, K. *Glycobiology* **2002,** *12,* 199–208. doi:10.1093/glycob/12.3.199
8. Vokhmyanina, O. A.; Rapoport, E. M.; André, S.; Severov, V. V.; Ryzhov, I.; Pazynina, G. V.; Korchagina, E.; Gabius, H.-J.; Bovin, N. V. *Glycobiology* **2012,** *22,* 1207–1217. doi:10.1093/glycob/cws079
9. Hoffmann, M.; Hayes, M. R.; Pietruszka, J.; Elling, L. *Glycoconjugate J.* **2020,** *37,* 457–470. doi:10.1007/s10719-020-09926-y
10. Ou, C.; Li, C.; Feng, C.; Tong, X.; Vasta, G. R.; Wang, L.-X. *Bioorg. Med. Chem.* **2022,** *72,* 116974. doi:10.1016/j.bmc.2022.116974
11. Sanki, A. K.; Mahal, L. K. *Synlett* **2006,** 455–459. doi:10.1055/s-2006-926264
12. Ma, H. Synthesis of human Mucin-type glycans for interaction studies with bacterial and human lectins. Ph.D. Thesis, University College Dublin, Dublin, Ireland, 2019.
13. Imamura, A.; Ando, H.; Korogi, S.; Tanabe, G.; Muraoka, O.; Ishida, H.; Kiso, M. *Tetrahedron Lett.* **2003,** *44,* 6725–6728. doi:10.1016/s0040-4039(03)01647-2
14. Imamura, A.; Matsuzawa, N.; Sakai, S.; Udagawa, T.; Nakashima, S.; Ando, H.; Ishida, H.; Kiso, M. *J. Org. Chem.* **2016,** *81,* 9086–9104. doi:10.1021/acs.joc.6b01685
15. Imamura, A.; Kimura, A.; Ando, H.; Ishida, H.; Kiso, M. *Chem. – Eur. J.* **2006,** *12,* 8862–8870. doi:10.1002/chem.200600832
16. Komori, T.; Ando, T.; Imamura, A.; Li, Y.-T.; Ishida, H.; Kiso, M. *Glycoconjugate J.* **2008,** *25,* 647–661. doi:10.1007/s10719-008-9117-9
17. Reihill, M. Synthesis of Glycans and Glycoconjugates to Assess Binding Properties of Bacterial and Human Lectins. Ph.D. Thesis, University College Dublin, Dublin, Ireland, 2020.
18. Dakanali, M.; Do, T. H.; Horn, A.; Chongchivivat, A.; Jarusreni, T.; Lichlyter, D.; Guizzunti, G.; Haidekker, M. A.; Theodorakis, E. A. *Bioorg. Med. Chem.* **2012,** *20,* 4443–4450. doi:10.1016/j.bmc.2012.05.026
19. Garegg, P. J.; Konradsson, P.; Kvarnström, I.; Norberg, T.; Svensson, S. C. T.; Wigilius, B. *Acta Chem. Scand., Ser. B* **1985,** *39,* 569–577. doi:10.3891/acta.chem.scand.39b-0569
20. Ohlsson, J.; Magnusson, G. *Carbohydr. Res.* **2000,** *329,* 49–55. doi:10.1016/s0008-6215(00)00154-3

21. Modarresi-Alam, A. R.; Najafi, P.; Rostamizadeh, M.; Keykha, H.; Bijanzadeh, H.-R.; Kleinpeter, E. *J. Org. Chem.* **2007,** *72,* 2208–2211. doi:10.1021/jo061301f
22. Rablen, P. R.; Miller, D. A.; Bullock, V. R.; Hutchinson, P. H.; Gorman, J. A. *J. Am. Chem. Soc.* **1999,** *121,* 218–226. doi:10.1021/ja982304f
23. Gamov, G. A.; Aleksandriiskii, V. V.; Sharnin, V. A. *J. Mol. Liq.* **2017,** *231,* 238–241. doi:10.1016/j.molliq.2017.01.078
24. Lin, W.; Zhang, X.; He, Z.; Jin, Y.; Gong, L.; Mi, A. *Synth. Commun.* **2002,** *32,* 3279–3284. doi:10.1081/scc-120014032
25. Tanimoto, H.; Kakiuchi, K. *Nat. Prod. Commun.* **2013,** *8,* 1021–1034.
26. Jacquemard, U.; Bénéteau, V.; Lefoix, M.; Routier, S.; Mérour, J.-Y.; Coudert, G. *Tetrahedron* **2004,** *60,* 10039–10047. doi:10.1016/j.tet.2004.07.071
27. Guilbert, B.; Davis, N. J.; Flitsch, S. L. *Tetrahedron Lett.* **1994,** *35,* 6563–6566. doi:10.1016/s0040-4039(00)78273-6
28. Lubineau, A.; Alais, J.; Lemoine, R. *J. Carbohydr. Chem.* **2000,** *19,* 151–169. doi:10.1080/07328300008544072
29. Roy, R.; Cao, Y.; Kaltner, H.; Kottari, N.; Shiao, T. C.; Belkhadem, K.; André, S.; Manning, J. C.; Murphy, P. V.; Gabius, H.-J. *Histochem. Cell Biol.* **2017,** *147,* 285–301. doi:10.1007/s00418-016-1525-5
30. Xing, G.-W.; Wu, D.; Poles, M. A.; Horowitz, A.; Tsuji, M.; Ho, D. D.; Wong, C.-H. *Bioorg. Med. Chem.* **2005,** *13,* 2907–2916. doi:10.1016/j.bmc.2005.02.018
31. Raghuraman, A.; Riaz, M.; Hindle, M.; Desai, U. R. *Tetrahedron Lett.* **2007,** *48,* 6754–6758. doi:10.1016/j.tetlet.2007.07.100
32. Malleron, A.; Hersant, Y.; Narvor, C. L. *Carbohydr. Res.* **2006,** *341,* 29–34. doi:10.1016/j.carres.2005.10.004

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Optimizations of lipid II synthesis: an essential glycolipid precursor in bacterial cell wall synthesis and a validated antibiotic target

Milandip Karak, Cian R. Cloonan, Brad R. Baker, Rachel V. K. Cochrane and Stephen A. Cochrane*

Full Research Paper

Open Access

Address:
School of Chemistry and Chemical Engineering, Queen's University Belfast, David Keir Building, Stranmillis Road, Belfast, BT9 5AG, UK

Email:
Stephen A. Cochrane* - s.cochrane@qub.ac.uk

* Corresponding author

Keywords:
chemical glycosylation; lipid II; peptidoglycan; polyprenyls; total synthesis

Beilstein J. Org. Chem. **2024,** *20,* 220–227.
https://doi.org/10.3762/bjoc.20.22

Received: 17 November 2023
Accepted: 26 January 2024
Published: 06 February 2024

This article is part of the thematic issue "Chemical glycobiology".

Guest Editor: B. Schumann

Abstract

Lipid II is an essential glycolipid found in bacteria. Accessing this valuable cell wall precursor is important both for studying cell wall synthesis and for studying/identifying novel antimicrobial compounds. Herein, we describe optimizations to the modular chemical synthesis of lipid II and unnatural analogues. In particular, the glycosylation step, a critical step in the formation of the central disaccharide unit (GlcNAc-MurNAc), was optimized. This was achieved by employing the use of glycosyl donors with diverse leaving groups. The key advantage of this approach lies in its adaptability, allowing for the generation of a wide array of analogues through the incorporation of alternative building blocks at different stages of synthesis.

Introduction

Lipid II (Figure 1) is an essential bacterial glycolipid involved in peptidoglycan biosynthesis [1]. It is synthesized on the inner leaflet of the cytoplasmic membrane, before translocation to the outer leaflet, where it is then used as the monomeric building block of peptidoglycan biosynthesis. Lipid II is a validated antibiotic target for clinically prescribed antibiotics including vancomycin and ramoplanin [2]. It is also the target for a host of other antimicrobials (mostly non-ribosomal peptides), including the tridecaptins [3], nisin [4], teixobactin [5], clovibactin [6], malacidin [7], and cilagicin [8].

Despite significant progress in the chemical synthesis of lipid II and its analogues, the scarcity of these compounds and their limited structural diversity present significant obstacles to in-depth explorations of their intricate structural and functional characteristics. This scarcity issue is further exacerbated by an

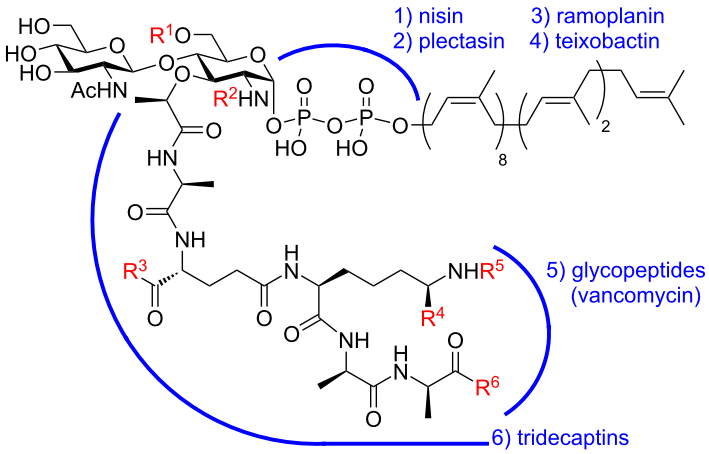


Figure 1: Structure of lipid II, with variable positions shown in red and antimicrobial-binding motifs highlighted with blue arcs. R^1 = H or Ac; R^2 = H or Ac; R^3 = OH, OMe or NH_2; R^4 = H or COOH; R^5 = Gly_5, Ala_2, Ala-Ser/Ala or D-Asp; R^6 = OH, OMe or NH_2. These structural modifications are described in detail by Münch and co-workers [9]. For more details on lipid II-binding antimicrobials, see recent review by Buijs and co-workers [2].

overwhelming demand that far exceeds existing supply capacities. To date, the chemical, chemoenzymatic, or biochemical synthesis of lipid II and its variants has been achieved by several research groups [10-27]. Nonetheless, considering the current state of knowledge, the chemical synthesis approach emerges as a more viable strategy in contrast to other methodologies, as it offers the potential to generate ample quantities of lipid II analogues suitable for high-throughput screening endeavors. In recent years, a major focus of the Cochrane lab has been the chemical synthesis of bacterial polyprenyls to study the mechanism of action of antimicrobial peptides that kill bacteria through binding to these polyprenyls [21,28-34]. Lipid II has been of particular interest, and during our synthesis of multiple different lipid II analogues, we have developed several optimizations, which we describe herein. The base lipid II syntheses upon which optimizations were made are our previously reported syntheses of Gram-negative lipid II in 2016 [20] and Gram-positive lipid II (**11**) in 2018 [23]. Building upon these synthetic strategies we have achieved noteworthy enhancements in glycosylation conditions, including improvements in reaction time and yields. This approach enables the systematic assembly of lipid II and analogues that contain shorter polyprenyl chains, specifically farnesyl (C_{15}), geranylgeranyl (C_{20}), and solanesyl (C_{45}). Such short chain analogues are valuable in several applications due to their improved solubility in aqueous systems. Assembly is achieved by integrating distinct carbohydrate, peptide, and polyprenyl phosphate building blocks. This modular synthetic method allows for the strategic substitution of constituent building blocks at different synthetic stages and provides a practical avenue for producing substantial amounts of lipid II analogues. Consequently, this approach offers a more feasible means of addressing the demands associated with biophysical screening pursuits.

Prior research in the field of total synthesis of lipid II has elucidated that specific combinations of protecting groups on glycosyl acceptors and donors, as represented by compounds **1a** and **2a** in Figure 2, are proficient in the efficient generation of lipid II disaccharide [35,36]. Subsequently, significant endeavors have been directed towards the exploration of glycosyl donors, such as *N*-phthaloyl 3,4,6-*O*-triacetyl-2-deoxy-2-amino-D-glucopyranosyl-1-bromide, *N*-2,2,2-trichloroethoxycarbonyl-3,4,6-*O*-triacetyl-2-deoxy-2-amino-D-glucopyranosyl-1-bromide, and *N*-phthaloyl-2-deoxy-2-amino-3,4,6-*O*-triacetate-D-glucopyranosyl-1-(2,2,2-trichloroacetoimidate), all of which have proven successful in disaccharide synthesis alongside C6-protected acceptors (**2a** or **2b** in Figure 2) [10,11,14,15,37,38]. More recently, an innovative one-pot glycosylation approach using a (2,6-dichloro-4-methoxyphenyl)(2,4-dichlorophenyl)-protected glycosyl acceptor has been developed, demonstrating satisfactory stability under Schmidt glycosylation conditions [18]. In general, the outcome of glycosylation hinges on the specific pairing of glycosyl donors and glycosyl acceptors employed in the reaction. Notably, when glycosyl donors such as **1e**–**g**, featuring acyl group protection at the C2 position, are combined with acceptors like **2b**, which have acyl groups protecting the C6 position, the reaction kinetics become sluggish, resulting in low conversion rates or no conversion [36,39].

Results and Discussion

In our studies, the initial glycosyl donors and acceptors (Figure 2; compounds **1a**–**g** and **2a**,**b**) were synthesized using established procedures from the literature, commencing with D-glucosamine and benzyl 2-acetamido-4,6-*O*-benzylidene-2-deoxy-α-D-glucopyranoside as the starting materials, respectively [40-43]. Imidate donors **1a** and **1e** were obtained exclusively

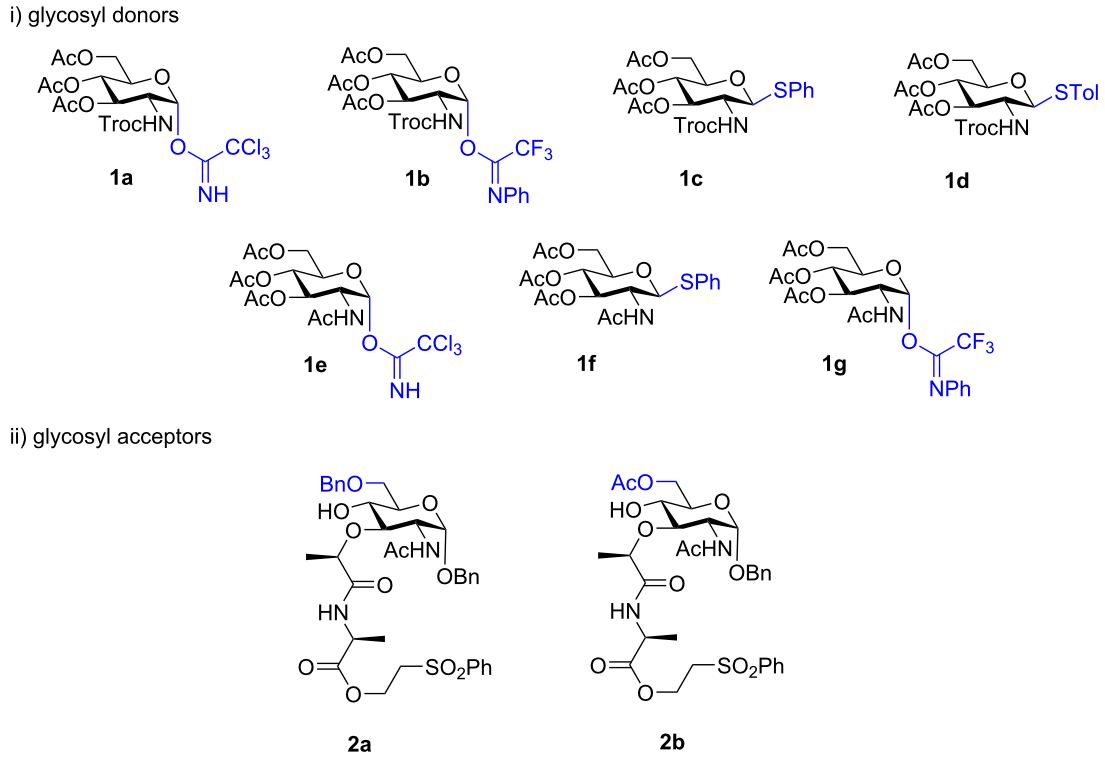
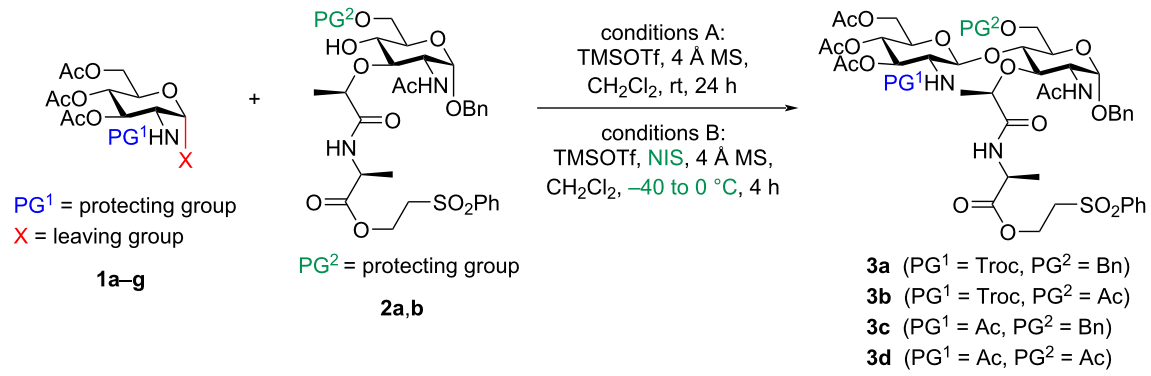


Figure 2: List of i) glycosyl donors and ii) glycosyl acceptors used in this study.

as α-anomers, and **1b** and **1g** as a 1:1 α:β mixture which were then purified to give the desired α-anomers. Thioglycosides **1c**, **1d**, and **1f** were isolated purely as β-anomers due to anchimeric assistance from the C2 *N*-acetyl or *N*-Troc groups. In glycosyl acceptors, the first amino acid of the lipid II pentapeptide, Ala, was incorporated as a 2-(phenylsulfonyl)ethyl ester, as previously reported by Saha and co-workers [44]. This modification prevents a deleterious side reaction occurring, wherein during glycosylation, muramic acid esters undergo a 6-*exo*-*trig* cyclization with the 4-OH group. Comprehensive experimental protocols detailing the preparation of these glycosyl donors can be found in Supporting Information File 1.

Next, we conducted an extended investigation into glycosylation, employing a diverse range of glycosyl donors (**1a**–**g**) and acceptors (**2a** and **2b**), and the comprehensive results are presented in Table 1. Initially, our approach was guided by the established protocols of Kurosu et al., which had previously demonstrated effectiveness in glycosylating glycosyl trichloroacetimidate **1a** and C6-benzylated MurNAc derivative **2a** [18]. Despite our efforts to optimize the yield of the target product **3a**, involving modifications to reaction conditions such as transitioning from 0 °C to room temperature and extending the reaction duration from 3 to 24 hours, we did not observe the anticipated enhancements (51% yield, entry 1, Table 1). This trend persisted when we attempted glycosylation between C6-acetylated MurNAc derivative **2b** and **1a**, where the desired product **3b** remained elusive (Table 1, entry 2). In fact, glycosyl acceptor **2b** failed to yield the desired glycosylation product **3d** under the conditions tested (Table 1, entries 7 and 8). Moderate yields of **3a** were achieved when using glycosyl donors such as **1b**–**d** under standard conditions A or B (Table 1, entries 3–5). Notably, both Troc-protected thio-donors **1c**,**d** exhibited similar behavior in terms of yield. Unfortunately, no target product **3c** was obtained under standard glycosylation conditions A or B when C2-acetamido glycosyl donors (e.g., **1e**–**g**) were subjected to the glycosylation reaction (Table 1, entries 6, 8, and 9). A slight improvement in the yield of **3a** was observed when switching from TMSOTf to TfOH as the activator (Table 1, entry 5 vs entry 10). However, substituting TMSOTf with $BF_3{\cdot}OEt_2$ did not yield any target product **3a** (Table 1, entry 3 vs entry 12). In our observations, we initially noted that at room temperature, the degradation rate of glycosyl donor **1a** exceeded the rate of product formation. This led to a complex mixture consisting of the target product **3a**, acceptor **2a**, and various degraded products of donor **1a**. This situation posed challenges, as even prolonged reaction times did not enhance the product yield, and the subsequent purification of the target product became a difficult task. However, when we conducted the reaction at lower temperatures, the degradation of glycosyl donor **1a**

Table 1: Optimization of the glycosylation conditions.[a]

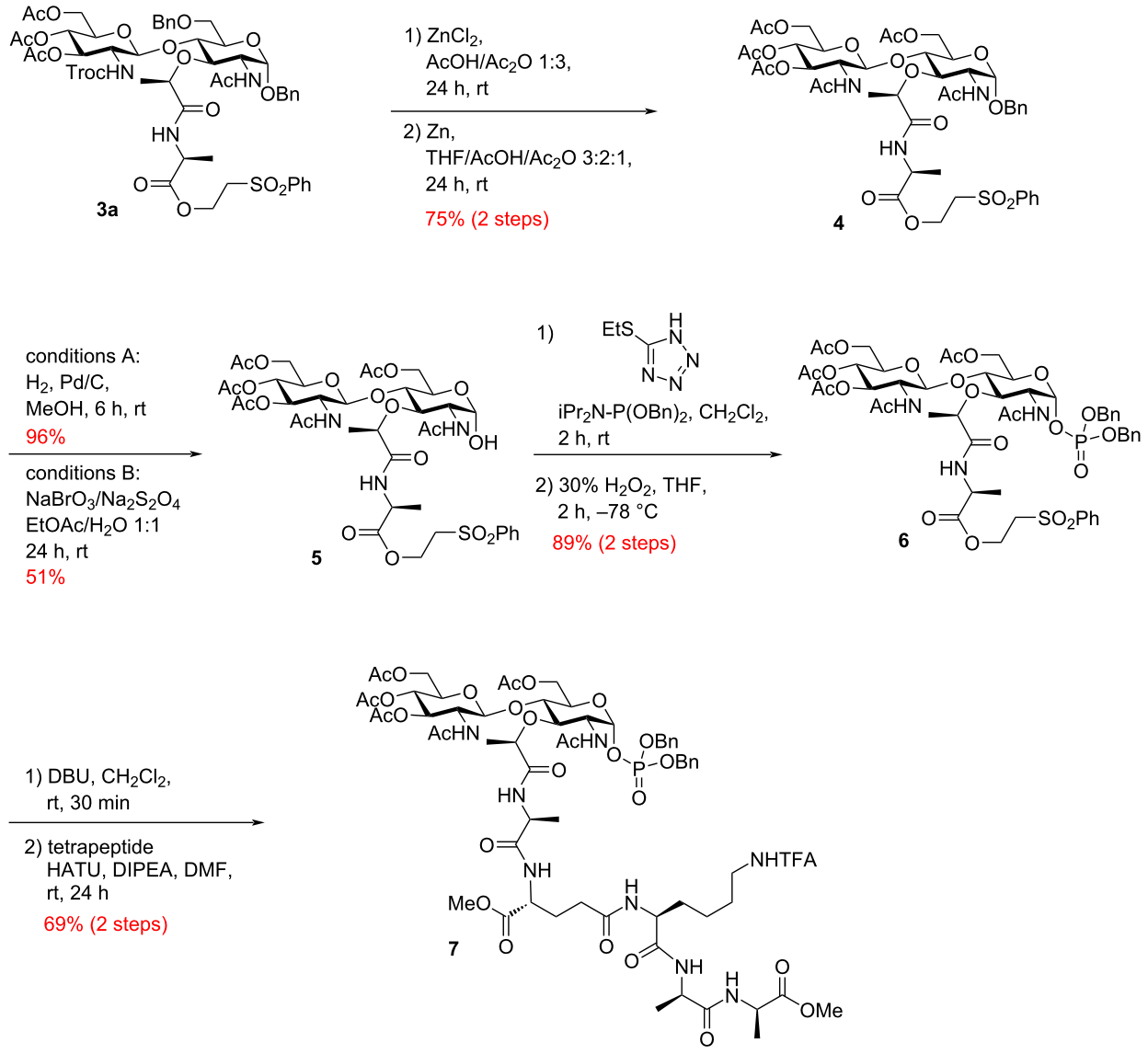


Entry	Donor	Acceptor	Deviation from std. conditions	Product	Yield (%)
1	**1a**	**2a**	conditions A	**3a**	51
2	**1a**	**2b**	conditions A	**3b**	0
3	**1b**	**2a**	conditions A	**3a**	29
4	**1c**	**2a**	conditions B	**3a**	46
5	**1d**	**2a**	conditions B	**3a**	43
6	**1e**	**2a**	conditions A	**3c**	0
7	**1e**	**2b**	conditions A	**3d**	0
8	**1f**	**2b**	conditions B	**3d**	0
9	**1g**	**2a**	conditions A	**3c**	0
10	**1d**	**2a**	**TfOH**, NIS, 4 Å MS, CH_2Cl_2, –40 to 0 °C, 4 h	**3a**	50
11	**1a**	**2a**	TMSOTf, 4 Å MS, CH_2Cl_2, 0 °C, 3 h; **then, added 2 equiv 1a, 1 equiv TMSOTf, 0 °C, 4 h**	**3a**	68
12	**1b**	**2a**	**BF_3·OEt_2**, 4 Å MS, CH_2Cl_2, 0 °C to rt, 24 h	**3a**	0

[a]TMSOTf: trimethylsilyl trifluoromethanesulfonate, MS: molecular sieves, NIS: *N*-iodosuccinimide, Ac: acetyl, Bn: benzyl, Troc: 2,2,2-trichloroethoxycarbonyl.

slowed down, and the reaction proceeded at a moderate rate. Eventually, we found that the utilization of extra equivalents of **1a** and activators, following conditions akin to those employed by Kurosu, resulted in a significant boost in the yield of the target product to 68% (Table 1, entry 11).

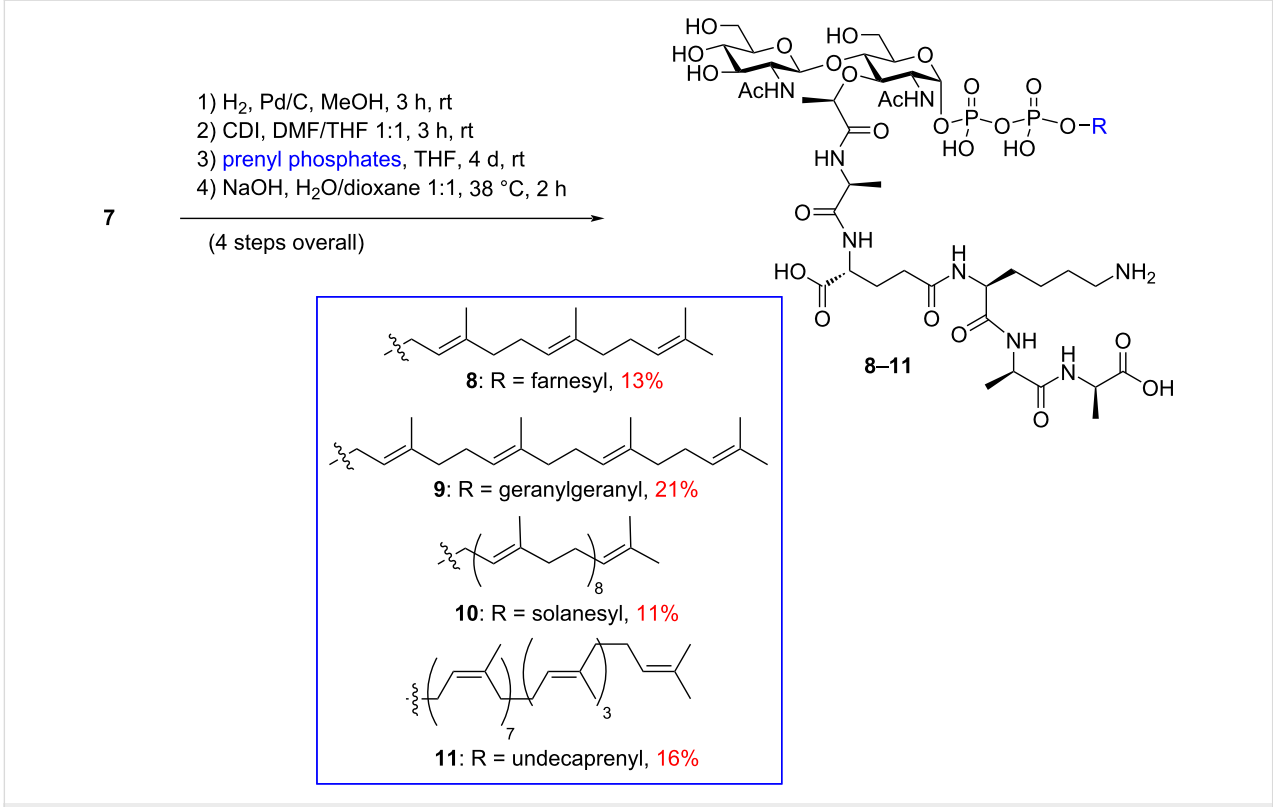
Next, a comprehensive synthetic strategy for the preparation of α-phosphoryl GlcNAc-MurNAc-pentapeptide **7**, based on established protocols with minor adjustments was completed (Scheme 1) [10,11]. After the successful glycosylation reaction, disaccharide **3a**, protected with C2-Troc and C6-benzyl groups, was efficiently deprotected under acidic conditions using $ZnCl_2$/Zn, followed by in situ re-acetylation of the C2-amino group and C6-alcohol with acetic anhydride, resulting in the formation of disaccharide **4** in a one-pot fashion. The anomeric benzyl protecting group in disaccharide **4** was then removed via a Pd/C-catalyzed hydrogenation reaction, producing a mixture of α/β-anomers of compound **5**. It is noteworthy to mention that the benzyl ether in compound **4** exhibited successful cleavage upon treatment with sodium bromate/sodium dithionite in ethyl acetate/water, while other protecting functionalities like acetyl and phenylsulfonylethyl ester groups remained intact [45]. The ratio of α/β-anomers in compound **5** was found to be influenced by the reaction conditions, consistently favoring the β-anomer. Further transformation of compound **5** involved α-selective phosphite formation using dibenzyl *N*,*N*-diisopropylphosphoramidite and 5-(ethylthio)-1*H*-tetrazole. The resulting α-phosphite intermediate was then oxidized with hydrogen peroxide to yield dibenzyl α-phosphate **6**, achieving an overall yield of 89% for these two steps. Removal of the 2-(phenylsulfonyl)ethanol protecting group in compound **6** was successfully achieved through treatment with 1,8-diazabicyclo[5.4.0]undec-7-ene, leading to the formation of the α-phosphoryl GlcNAc-MurNAc-monopeptide derivative. Subsequently, coupling this intermediate with tetrapeptide, TFA·H-L-Ala-γ-D-Glu(OMe)-L-Lys($COCF_3$)-D-Ala-D-Ala-OMe under mild conditions resulted in the synthesis of dibenzyl α-phosphoryl GlcNAc-MurNAc-pentapeptide **7** (see Supporting Information File 1 for compre-



Scheme 1: Synthesis of disaccharide pentapeptide core **7**.

hensive information on the synthesis details of the tetrapeptide). To avoid loss of valuable material through HPLC purification, crude **7** is used directly in the next step, and purification performed after the final prenyl phosphate coupling and global deprotection.

Finally, the benzyl-protecting groups in compound **7** were cleaved via hydrogenolysis, followed by co-evaporation of the resulting crude product in pyridine. This yielded a monopyridyl salt, setting the stage for the final lipid coupling and deprotection sequence. To establish the vital lipid diphosphate linkage, we employed the phosphoroimidazolidate method, as previously utilized in other lipid II total syntheses [10,11]. The monopyridyl α-phosphoryl GlcNAc-MurNAc-pentapeptide was activated with CDI, with excess CDI being neutralized using anhydrous methanol. The resulting phosphoroimidazolidate mixture underwent a cross-coupling reaction with prenyl monophosphates [46] in DMF/THF over a four-day period, yielding fully protected versions of lipid II and its analogues. Subsequent global deprotection reactions, using aqueous NaOH, led to the formation of lipid II (**11**), with an overall yield of 16% (from compound **7**) following reversed-phase HPLC purification (Scheme 2). Similarly, farnesyl, geranylgeranyl, and solanesyl-lipid II analogues **8**–**10** were synthesized with overall yields of 13%, 21%, and 11%, respectively, using the corresponding prenyl phosphates (Scheme 2).



Scheme 2: Synthesis of lipid II (**11**) and its analogues **8**–**10**.

Conclusion

In conclusion, we have successfully optimized a modular approach for the synthesis of lipid II and its analogues, including variants with distinct prenyl-chain lengths. The key to this methodology lies in the optimization of glycosylation conditions, utilizing readily available glycosyl donors, which is a pivotal step in constructing the central disaccharide unit. The adaptability of our method is showcased through the generation of new lipid II analogues, such as geranylgeranyl and solanesyl lipid II analogues, which involve the incorporation of distinct prenyl monophosphates during the final phases of the synthesis. Thus, this strategy holds considerable promise for advancing the synthesis of a diverse range of lipid II analogues, opening avenues for further exploration into their biophysical characteristics, as well as their interactions with antibiotics.

Supporting Information

Supporting Information File 1

Experimental procedures, characterization data, and selected copies of ^{1}H, ^{13}C, and ^{31}P NMR spectra.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-22-S1.pdf]

Acknowledgements

We extend our gratitude to Professor Alethea Tabor and Professor Stefan Howorka from University College London for their valuable collaboration on this project. We also express our appreciation to the dedicated technical team at CCE-QUB for their unwavering technical support throughout this project.

Funding

We thank the Engineering and Physical Sciences Research Council for financial support of this project (Grant No EP/V032860/1).

Conflict of Interest

The authors declare no conflict of interest.

ORCID® iDs

Milandip Karak - https://orcid.org/0000-0001-9998-5994
Cian R. Cloonan - https://orcid.org/0009-0001-0600-4374
Rachel V. K. Cochrane - https://orcid.org/0000-0002-3876-0561
Stephen A. Cochrane - https://orcid.org/0000-0002-6239-6915

Data Availability Statement

The data that supports the findings of this study is available from the corresponding author upon reasonable request.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2023.49.v1

References

1. Egan, A. J. F.; Errington, J.; Vollmer, W. *Nat. Rev. Microbiol.* **2020,** *18,* 446–460. doi:10.1038/s41579-020-0366-3
2. Buijs, N. P.; Matheson, E. J.; Cochrane, S. A.; Martin, N. I. *Chem. Commun.* **2023,** *59,* 7685–7703. doi:10.1039/d3cc01070h
3. Bann, S. J.; Ballantine, R. D.; Cochrane, S. A. *RSC Med. Chem.* **2021,** *12,* 538–551. doi:10.1039/d0md00413h
4. Medeiros-Silva, J.; Jekhmane, S.; Paioni, A. L.; Gawarecka, K.; Baldus, M.; Swiezewska, E.; Breukink, E.; Weingarth, M. *Nat. Commun.* **2018,** *9,* 3963. doi:10.1038/s41467-018-06314-x
5. Medeiros-Silva, J.; Jekhmane, S.; Breukink, E.; Weingarth, M. *ChemBioChem* **2019,** *20,* 1731–1738. doi:10.1002/cbic.201800796
6. Shukla, R.; Peoples, A. J.; Ludwig, K. C.; Maity, S.; Derks, M. G. N.; De Benedetti, S.; Krueger, A. M.; Vermeulen, B. J. A.; Harbig, T.; Lavore, F.; Kumar, R.; Honorato, R. V.; Grein, F.; Nieselt, K.; Liu, Y.; Bonvin, A. M. J. J.; Baldus, M.; Kubitscheck, U.; Breukink, E.; Achorn, C.; Nitti, A.; Schwalen, C. J.; Spoering, A. L.; Ling, L. L.; Hughes, D.; Lelli, M.; Roos, W. H.; Lewis, K.; Schneider, T.; Weingarth, M. *Cell* **2023,** *186,* 4059–4073. doi:10.1016/j.cell.2023.07.038
7. Hover, B. M.; Kim, S.-H.; Katz, M.; Charlop-Powers, Z.; Owen, J. G.; Ternei, M. A.; Maniko, J.; Estrela, A. B.; Molina, H.; Park, S.; Perlin, D. S.; Brady, S. F. *Nat. Microbiol.* **2018,** *3,* 415–422. doi:10.1038/s41564-018-0110-1
8. Wang, Z.; Koirala, B.; Hernandez, Y.; Zimmerman, M.; Brady, S. F. *Science* **2022,** *376,* 991–996. doi:10.1126/science.abn4213
9. Münch, D.; Sahl, H.-G. *Biochim. Biophys. Acta, Biomembr.* **2015,** *1848,* 3062–3071. doi:10.1016/j.bbamem.2015.04.014
10. Schwartz, B.; Markwalder, J. A.; Wang, Y. *J. Am. Chem. Soc.* **2001,** *123,* 11638–11643. doi:10.1021/ja0166848
11. VanNieuwenhze, M. S.; Mauldin, S. C.; Zia-Ebrahimi, M.; Winger, B. E.; Hornback, W. J.; Saha, S. L.; Aikins, J. A.; Blaszczak, L. C. *J. Am. Chem. Soc.* **2002,** *124,* 3656–3660. doi:10.1021/ja017386d
12. Liu, H.; Wong, C.-H. *Bioorg. Med. Chem.* **2006,** *14,* 7187–7195. doi:10.1016/j.bmc.2006.06.058
13. Liu, C.-Y.; Guo, C.-W.; Chang, Y.-F.; Wang, J.-T.; Shih, H.-W.; Hsu, Y.-F.; Chen, C.-W.; Chen, S.-K.; Wang, Y.-C.; Cheng, T.-J. R.; Ma, C.; Wong, C.-H.; Fang, J.-M.; Cheng, W.-C. *Org. Lett.* **2010,** *12,* 1608–1611. doi:10.1021/ol100338v
14. Shih, H.-W.; Chen, K.-T.; Cheng, T.-J. R.; Wong, C.-H.; Cheng, W.-C. *Org. Lett.* **2011,** *13,* 4600–4603. doi:10.1021/ol201806d
15. Meng, F.-C.; Chen, K.-T.; Huang, L.-Y.; Shih, H.-W.; Chang, H.-H.; Nien, F.-Y.; Liang, P.-H.; Cheng, T.-J. R.; Wong, C.-H.; Cheng, W.-C. *Org. Lett.* **2011,** *13,* 5306–5309. doi:10.1021/ol2021687
16. Chen, K.-T.; Kuan, Y.-C.; Fu, W.-C.; Liang, P.-H.; Cheng, T.-J. R.; Wong, C.-H.; Cheng, W.-C. *Chem. – Eur. J.* **2013,** *19,* 834–838. doi:10.1002/chem.201203251
17. Huang, L.-Y.; Huang, S.-H.; Chang, Y.-C.; Cheng, W.-C.; Cheng, T.-J. R.; Wong, C.-H. *Angew. Chem., Int. Ed.* **2014,** *53,* 8060–8065. doi:10.1002/anie.201402313
18. Mitachi, K.; Mohan, P.; Siricilla, S.; Kurosu, M. *Chem. – Eur. J.* **2014,** *20,* 4554–4558. doi:10.1002/chem.201400307
19. Lin, C.-K.; Chen, K.-T.; Hu, C.-M.; Yun, W.-Y.; Cheng, W.-C. *Chem. – Eur. J.* **2015,** *21,* 7511–7519. doi:10.1002/chem.201406629
20. Cochrane, S. A.; Findlay, B.; Bakhtiary, A.; Acedo, J. Z.; Rodriguez-Lopez, E. M.; Mercier, P.; Vederas, J. C. *Proc. Natl. Acad. Sci. U. S. A.* **2016,** *113,* 11561–11566. doi:10.1073/pnas.1608623113
21. Bakhtiary, A.; Cochrane, S. A.; Mercier, P.; McKay, R. T.; Miskolzie, M.; Sit, C. S.; Vederas, J. C. *J. Am. Chem. Soc.* **2017,** *139,* 17803–17810. doi:10.1021/jacs.7b04728
22. Katsuyama, A.; Sato, K.; Yakushiji, F.; Matsumaru, T.; Ichikawa, S. *Chem. Pharm. Bull.* **2018,** *66,* 84–95. doi:10.1248/cpb.c17-00828
23. Dong, Y. Y.; Wang, H.; Pike, A. C. W.; Cochrane, S. A.; Hamedzadeh, S.; Wyszyński, F. J.; Bushell, S. R.; Royer, S. F.; Widdick, D. A.; Sajid, A.; Boshoff, H. I.; Park, Y.; Lucas, R.; Liu, W.-M.; Lee, S. S.; Machida, T.; Minall, L.; Mehmood, S.; Belaya, K.; Liu, W.-W.; Chu, A.; Shrestha, L.; Mukhopadhyay, S. M. M.; Strain-Damerell, C.; Chalk, R.; Burgess-Brown, N. A.; Bibb, M. J.; Barry, C. E., III; Robinson, C. V.; Beeson, D.; Davis, B. G.; Carpenter, E. P. *Cell* **2018,** *175,* 1045–1058. doi:10.1016/j.cell.2018.10.037
24. Wingen, L. M.; Rausch, M.; Schneider, T.; Menche, D. *J. Org. Chem.* **2020,** *85,* 10206–10215. doi:10.1021/acs.joc.0c01004
25. Cochrane, S. A.; Lohans, C. T. *Eur. J. Med. Chem.* **2020,** *194,* 112262. doi:10.1016/j.ejmech.2020.112262
26. Hsu, T.-W.; Fang, J.-M. *Analyst (Cambridge, U. K.)* **2021,** *146,* 5843–5847. doi:10.1039/d1an00941a
27. Katsuyama, A.; Yakushiji, F.; Ichikawa, S. *Tetrahedron Lett.* **2021,** *73,* 153101. doi:10.1016/j.tetlet.2021.153101
28. Kotsogianni, I.; Wood, T. M.; Alexander, F. M.; Cochrane, S. A.; Martin, N. I. *ACS Infect. Dis.* **2021,** *7,* 2612–2619. doi:10.1021/acsinfecdis.1c00316
29. Baker, B. R.; Ives, C. M.; Bray, A.; Caffrey, M.; Cochrane, S. A. *Eur. J. Med. Chem.* **2021,** *210,* 113062. doi:10.1016/j.ejmech.2020.113062
30. Cochrane, R. V. K.; Alexander, F. M.; Boland, C.; Fetics, S. K.; Caffrey, M.; Cochrane, S. A. *Chem. Commun.* **2020,** *56,* 8603–8606. doi:10.1039/d0cc03388j
31. Chiorean, S.; Antwi, I.; Carney, D. W.; Kotsogianni, I.; Giltrap, A. M.; Alexander, F. M.; Cochrane, S. A.; Payne, R. J.; Martin, N. I.; Henninot, A.; Vederas, J. C. *ChemBioChem* **2020,** *21,* 789–792. doi:10.1002/cbic.201900504
32. Bann, S. J.; Ballantine, R. D.; McCallion, C. E.; Qian, P.-Y.; Li, Y.-X.; Cochrane, S. A. *J. Med. Chem.* **2019,** *62,* 10466–10472. doi:10.1021/acs.jmedchem.9b01078
33. Ballantine, R. D.; McCallion, C. E.; Nassour, E.; Tokajian, S.; Cochrane, S. A. *Med. Chem. Commun.* **2019,** *10,* 484–487. doi:10.1039/c9md00031c
34. Ballantine, R. D.; Li, Y.-X.; Qian, P.-Y.; Cochrane, S. A. *Chem. Commun.* **2018,** *54,* 10634–10637. doi:10.1039/c8cc05790g
35. Termin, A.; Schmidt, R. R. *Liebigs Ann. Chem.* **1992,** 527–533. doi:10.1002/jlac.199219920191
36. Zhang, Z.; Ollmann, I. R.; Ye, X.-S.; Wischnat, R.; Baasov, T.; Wong, C.-H. *J. Am. Chem. Soc.* **1999,** *121,* 734–753. doi:10.1021/ja982232s
37. Hitchcock, S. A.; Eid, C. N.; Aikins, J. A.; Zia-Ebrahimi, M.; Blaszczak, L. C. *J. Am. Chem. Soc.* **1998,** *120,* 1916–1917. doi:10.1021/ja973172d
38. Zhang, Y.; Fechter, E. J.; Wang, T.-S. A.; Barrett, D.; Walker, S.; Kahne, D. E. *J. Am. Chem. Soc.* **2007,** *129,* 3080–3081. doi:10.1021/ja069060g

39. Ritter, T. K.; Mong, K.-K. T.; Liu, H.; Nakatani, T.; Wong, C.-H. *Angew. Chem., Int. Ed.* **2003,** *42,* 4657–4660. doi:10.1002/anie.200351534
40. Lioux, T.; Busson, R.; Rozenski, J.; Nguyen-Distèche, M.; Frère, J.-M.; Herdewijn, P. *Collect. Czech. Chem. Commun.* **2005,** *70,* 1615–1641. doi:10.1135/cccc20051615
41. Grann Hansen, S.; Skrydstrup, T. *Eur. J. Org. Chem.* **2007,** 3392–3401. doi:10.1002/ejoc.200700048
42. Subramanian, V.; Moumé-Pymbock, M.; Hu, T.; Crich, D. *J. Org. Chem.* **2011,** *76,* 3691–3709. doi:10.1021/jo102411j
43. Xu, C.; Liu, H.; Li, X. *Carbohydr. Res.* **2011,** *346,* 1149–1153. doi:10.1016/j.carres.2011.03.033
44. Saha, S. L.; Van Nieuwenhze, M. S.; Hornback, W. J.; Aikins, J. A.; Blaszczak, L. C. *Org. Lett.* **2001,** *3,* 3575–3577. doi:10.1021/ol016692t
45. Adinolfi, M.; Barone, G.; Guariniello, L.; Iadonisi, A. *Tetrahedron Lett.* **1999,** *40,* 8439–8441. doi:10.1016/s0040-4039(99)01756-6
46. Lira, L. M.; Vasilev, D.; Pilli, R. A.; Wessjohann, L. A. *Tetrahedron Lett.* **2013,** *54,* 1690–1692. doi:10.1016/j.tetlet.2013.01.059

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Elucidating the glycan-binding specificity and structure of *Cucumis melo* agglutinin, a new R-type lectin

Jon Lundstrøm[1,2], Emilie Gillon[3], Valérie Chazalet[3], Nicole Kerekes[1,2], Antonio Di Maio[4], Ten Feizi[4], Yan Liu[4], Annabelle Varrot[3] and Daniel Bojar[*1,2]

Full Research Paper

Open Access

Address:
[1]Department of Chemistry and Molecular Biology, University of Gothenburg, Medicinaregatan 7B, 413 90 Gothenburg, Sweden, [2]Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, 413 90 Gothenburg, Sweden, [3]Univ. Grenoble Alpes, CNRS, CERMAV, 601 Rue de la Chimie, 38610 Gières, France and [4]Glycosciences Laboratory, Faculty of Medicine, Imperial College London, Du Cane Rd, London W12 0NN, United Kingdom

Email:
Daniel Bojar[*] - daniel.bojar@gu.se

* Corresponding author

Keywords:
carbohydrate; glycan array; melon; plant lectin; R-type

Beilstein J. Org. Chem. **2024,** *20,* 306–320.
https://doi.org/10.3762/bjoc.20.31

Received: 01 December 2023
Accepted: 09 February 2024
Published: 19 February 2024

This article is part of the thematic issue "Chemical glycobiology".

Guest Editor: E. Fadda

Abstract

Plant lectins have garnered attention for their roles as laboratory probes and potential therapeutics. Here, we report the discovery and characterization of *Cucumis melo* agglutinin (CMA1), a new R-type lectin from melon. Our findings reveal CMA1's unique glycan-binding profile, mechanistically explained by its 3D structure, augmenting our understanding of R-type lectins. We expressed CMA1 recombinantly and assessed its binding specificity using multiple glycan arrays, covering 1,046 unique sequences. This resulted in a complex binding profile, strongly preferring C2-substituted, beta-linked galactose (both GalNAc and Fuca1-2Gal), which we contrasted with the established R-type lectin *Ricinus communis* agglutinin 1 (RCA1). We also report binding of specific glycosaminoglycan subtypes and a general enhancement of binding by sulfation. Further validation using agglutination, thermal shift assays, and surface plasmon resonance confirmed and quantified this binding specificity in solution. Finally, we solved the high-resolution structure of the CMA1 N-terminal domain using X-ray crystallography, supporting our functional findings at the molecular level. Our study provides a comprehensive understanding of CMA1, laying the groundwork for further exploration of its biological and therapeutic potential.

Introduction

Lectins have long been the subject of intense scientific scrutiny, serving as molecular bridges that span the realms of biochemistry, cellular biology, and biomedicine. These carbohydrate-binding proteins boast a range of functions, acting as recognition modules in cell–molecule and cell–cell interactions, thereby playing vital roles in immune defense, regulation of

growth, and apoptosis [1]. In plants, they serve as essential components in development, immunity, and stress signaling [2,3].

In light of the burgeoning interest in the intersection of glycobiology and biomedicine, the characterization of new lectins has carved out a significant niche in scientific research. Specifically, lectins have emerged as invaluable tools for staining cells and tissues, thereby offering insights into cellular heterogeneity and function. For instance, the use of wheat germ agglutinin (WGA) and concanavalin A (ConA) has been instrumental in selectively staining cells based on their glycan expression [4], including single-cell approaches [5,6]. In the realm of therapeutics, lectins such as mistletoe lectins have shown promise in cancer therapy, by virtue of their ability to induce apoptosis in malignant cells [7]. Further, the creation of lectin arrays [8,9], which employ a diverse set of characterized lectins, has enabled high-throughput glycan profiling, thereby advancing both diagnostic methods and biomarker discovery. Examples include arrays that can rapidly profile alterations in glycosylation patterns, pivotal in many diseases and inflammatory changes [10,11].

Traditionally, lectins are divided into classes based on structural similarity and, by extension, common folds [12]. Still, shared binding specificity does not always follow from structural similarity, exemplified by divergent evolution within lectin families as well as independent emergence of similar binding patterns [13]. Many of the most commonly used lectins for the abovementioned applications are R-type lectins, especially those derived from plants. Examples include SNA (from *Sambucus nigra*, binding Neu5Acα2-6 [14]) or RCA1 (from *Ricinus communis*, binding terminal β-linked galactose [15]).

Yet, despite the extensive studies on plant lectins, particularly R-type lectins, there are still significant gaps in our understanding. Further, in general, few melon lectins have been studied in detail. Some reports indicate the presence of chitooligosaccharide-binding (i.e., β1-4 GlcNAc oligomers) lectins from phloem exudates of melons [16,17], as well as R-type lectins in bitter melon [18], yet not much else is known about binding specificities exhibited by lectins derived from melons. In particular, existing research in this area often lacks a comprehensive characterization that includes both functional and structural analysis of these lectins.

Here, we introduce a novel member of characterized melon lectins, namely the *Cucumis melo* agglutinin (CMA1), an R-type lectin derived from melon. Prior to our study, CMA1 was only a predicted protein from genomic sequencing, with moderate certainty scores on lectin-specific databases. Our comprehensive analysis using glycan array experiments, thermal shift assays, and high-resolution X-ray crystallography not only confirms its classification as a functional R-type lectin but also provides a deep dive into its unique glycan-binding profile and high-resolution 3D structure. Overall, we present a deeply characterized new lectin with a unique binding profile of specifically recognizing C2-substituted galactose in the context of glycans.

Results and Discussion

Identification and production of a new lectin from the melon *Cucumis melo*

CMA1 is a predicted protein from whole-genome shotgun sequencing of leaves from the melon plant *Cucumis melo* (variant *makuwa*, taxon ID: 1194695) [19] and has, to our knowledge, never been studied before. With prediction scores of 0.453 on LectomeXplore [12] and 0.251 on TrefLec [20] (from 0, lowest, to 1, highest), CMA1 is moderately certain in its prior classification as a lectin. CMA1 comprises 291 amino acids and is predicted to fold into two linked β-trefoil domains belonging to carbohydrate-binding module family 13 (CBM13) and placing it into the group of R-type lectins. Both CBM13 domains are likely to exhibit carbohydrate-binding activity due to the conservation of key amino acids in at least one of the three potential binding sites. In contrast to other R-type lectins such as ricin, it lacks a catalytic domain.

As R-type lectins are both a well-investigated family of lectins and widely used in research and beyond, we first wanted to analyze where CMA1 would be situated in the broader context of R-type lectins. A multiple sequence alignment of binding domains of representative R-type lectins (Figure 1a) showed that CMA1 exhibited a binding domain with a sequence relatively similar to those of the plant lectins SNA and ricin. However, we note that, in general, the substantial heterogeneity of binding motifs of even closely related lectins (SNA: Neu5Acα2-6, ricin: Gal/GalNAc) does not allow for a strong a priori hypothesis of what CMA1 would bind, even though R-type lectins in general are thought to prefer the Gal/GalNAc type motif mentioned in the context of ricin [21].

We next aligned the individual units of the tandem repeat CBM13 domains, indicated by the N-terminal (34-158) and C-terminal units (162-286) and compared those to the domains of ricin (Figure 1b). R-type lectins have a characteristic Q-x-W structural motif close to their binding site, which is highly conserved [21]. We report that CMA1 largely follows this trend, with three such binding sites in both N- and C-terminal domain, albeit with imperfect overlap. Based on the location of

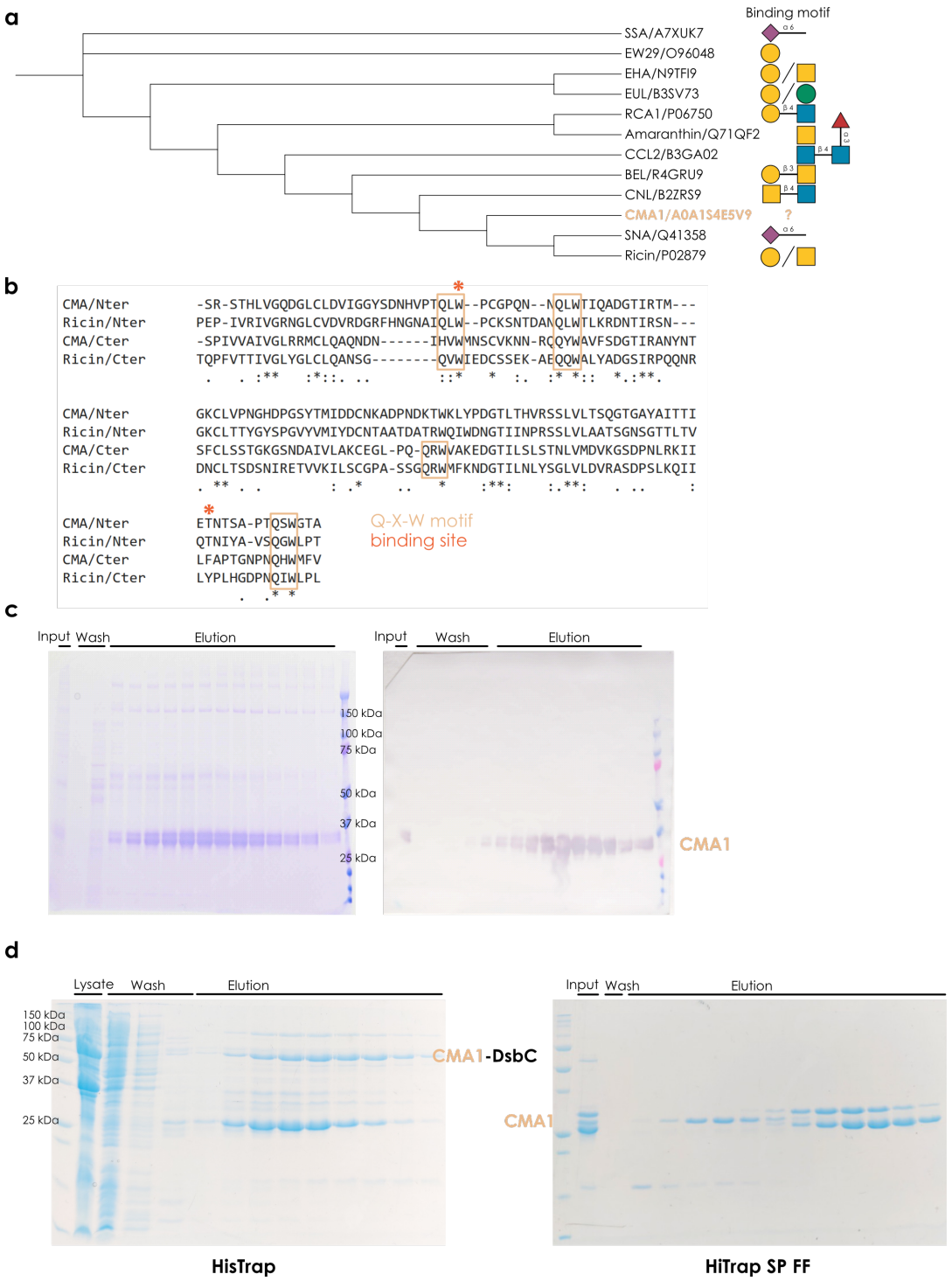


Figure 1: Characterizing a new lectin from the melon *Cucumis melo*. (a) Evolutionary relationships of common R-type lectins. For a range of representative R-type lectins, we aligned their protein sequences via MUSCLE [22] and built a neighbor-joining tree with the resulting alignment distances, which is shown as a cladogram. For each protein, we only used the lectin domain, as annotated by UniProt or InterPro. For each protein, a representative binding specificity, based on literature reports, is provided. (b) Similarity of the two CBM13 domains in CMA1. Using MUSCLE to align the N-terminal (34–158) and C-terminal domains (162–286) of CMA1 and ricin (321–448 and 451–575), we indicated the position of the conserved Q-x-W motif in R-type lectins. (c) Recombinant expression of CMA1 in mammalian cells. SDS-PAGE and anti-His-tag Western blot of fractions from the expression of CMA1 protein in CHO-S cells. Note the smeared band indicating the presence of glycosylation. (d) Recombinant expression of CMA1 in bacteria. SDS-PAGE gels of the His-tag affinity chromatography and cation exchange chromatography from the expression of CMA1 protein in *E. coli* BL21* cells.

the known binding pocket of the R-type lectin ricin and the respective sequence conservation in CMA1, we postulate binding sites around W^{63} for the N-terminal domain and F^{273} for the C-terminal domain of CMA1.

As binding specificities of melon lectins in general (beyond chitooligosaccharides), and CMA1 in particular, are still unknown, we set out to measure, quantify, and understand the glycan-binding properties of CMA1 in depth, as an archetypal example of melon lectins. For this, we needed to express the lectin recombinantly. As it is a secreted plant protein, we elected to express it in mammalian cell lines, to maximize the chances of a functional protein, because of post-translational modifications that would be lacking in bacteria as well as the oxidative environment of the secretory pathway, as CMA1 exhibits predicted disulfide bridges. A single step of His-tag affinity chromatography was sufficient to yield protein of adequate purity and good yield (≈15 mg of eluted protein from 800 mL of cell culture, Figure 1c).

In parallel, we also expressed CMA1 in a bacterial expression system, which allowed us to ascertain whether binding was influenced by lectin glycosylation. The full-length mature protein (6–264) and individual N- or C-terminal domains were expressed using a N-terminal fusion comprising DsbC and a hexa-His tag, cleavable by TEV (*Tobacco etch virus*) protease. Despite the presence of the DsbC signal peptide, we did not observe periplasmic localization, and all proteins were instead purified from the cytoplasm. Ni-NTA affinity chromatography followed by TEV protease cleavage of the fusion construct and subsequent reverse Ni-NTA affinity chromatography resulted in significant co-purification of *E. coli* contaminants, necessitating an extra purification step, where cation exchange chromatography allowed us to obtain pure fractions of $CMA1^{6–291}$. Of note, this additional purification step was not necessary for the purification of the CMA1 N-terminal domain (Figure 1d). Expression of the CMA1 C-terminal domain did not yield sufficiently pure and monodisperse protein for further biochemical and structural analyses.

Cucumis melo agglutinin binds C2-substituted, beta-linked galactose

We then set out to answer the question whether CMA1 was a functional lectin and, if yes, what its binding specificity was. The standard approach to elucidate lectin binding specificity is via glycan array experiments. Here, tagged soluble lectin is added to, often, immobilized glycans and bound lectin is quantified via fluorescence scanners, which can be paired with glycan information due to the known arrangements of immobilized glycans on the plate. To cover the broadest possible sequence space, we tested our eukaryotically produced CMA1 protein against the two largest glycan arrays at the National Center for Functional Glycomics (NCFG, Figure 2a) and the Glycosciences Laboratory at Imperial College London (ICL, Figure 2b). We note that, together, this encompasses 1,046 unique glycan sequences, spanning all major glycan classes and substantial taxonomic diversity. Next to these unique sequences, even more effects stem from a variety of linkers with which these molecules are immobilized.

In general, we observed two binding preferences that were strongly enriched among bound sequences, namely glycans containing Fucα1-2Gal epitopes and glycans containing terminal GalNAc residues (Figure 2c). Amongst the bound sequences, these substructures occurred in many different contexts, such as blood group H, LacdiNAc, or the Sd^a motif, and particularly in sequences resembling *O*-glycans, milk oligosaccharides, and glycosphingolipids. At first glance, these two binding specificities may seem unconnected, indicating a rather broadly binding lectin. However, we noticed that the commonality of these two epitopes is hidden in the IUPAC-condensed nomenclature: Both substructures exhibited a bulky substituent on C2 of galactose, either a fucosyl (Fucα1-2Gal) or *N*-acetyl (GalNAc) moiety (Figure 2d). We thus conclude that CMA1 is highly specific for C2-substituted galactose. We further argue for a preference for a beta-linked epitope as, while we do observe binding to structures containing α-linked GalNAc, the binding to their β-linked counterparts was generally stronger (e.g., GalNAcα: 1.57 vs GalNAcβ: 2.21, in *z*-scores (see Experimental section)). In part, this is reminiscent to the LacdiNAc binding specificity of *Clitocybe nebularis* lectin (CNL; Figure 1a) [27].

An important finding from the ICL array was that CMA1 exhibited robust binding to glycosaminoglycans (GAGs; Figure 2e; Supporting Information File 1, table “imperial”), in particular chondroitin sulfate (CS) C and A. Given the preference for terminal binding epitopes described above, the question naturally arose how the binding to these longer-chain glycans works. On the ICL array, CS sequences are typically capped with 4,5-unsaturated hexuronic acid derivatives on their non-reducing end and, thus, do not provide terminal GalNAc epitopes for binding. Further, while CMA1 did also bind to GalNAc-terminated GAGs (e.g., CSC-5, CSA-5), we measured higher binding to similar GAGs without the terminal GalNAc in several cases (Figure 2d,e). While some of the GAG probes varied in their immobilization amounts, we confirmed these results in a GAG-focused array (data not shown). We thus posit a binding to internal GalNAc epitopes for the case of GAG binding, potentially mediated by several binding sites.

This argument is strengthened by the observation that the highest observed binding to CSC and CSA was not with the

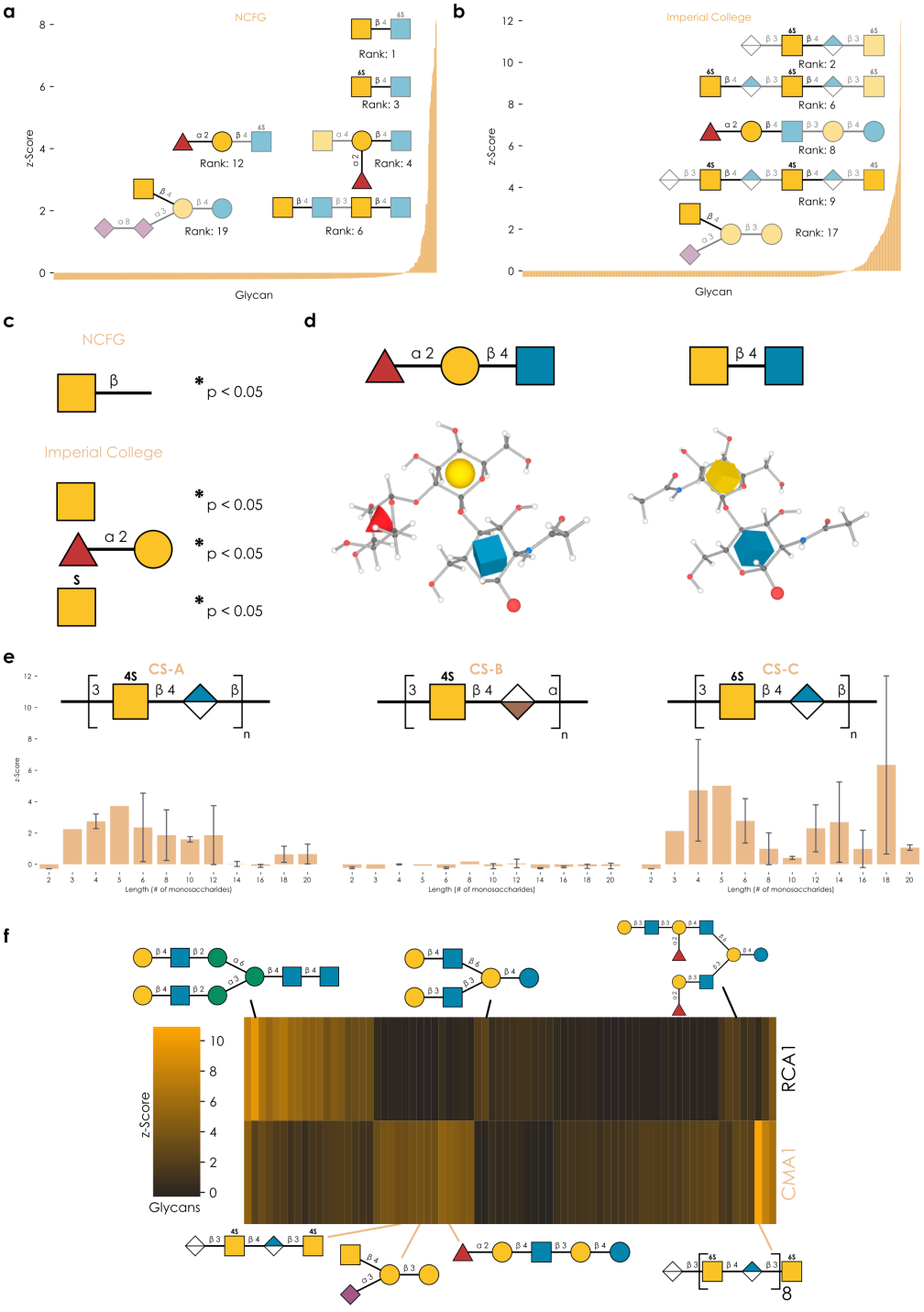


Figure 2: Characterizing the binding specificity of CMA1. (a, b) Lectin produced in mammalian cells was analyzed on the NCFG array (a) and the ICL array (b). Representative structures bound by CMA1 are shown via the "Symbol Nomenclature For Glycans" (SNFG), drawn with GlycoDraw [23]. Everything except the assigned binding motif is shown with added transparency. Full array data are available in Supporting Information File 1, tables "cfg" and "imperial". (c) Enrichment analysis of glycan array data. For both NCFG and ICL array data, we used the *get_pvals_motif* function from glycowork [24] (version 0.8.1) with the keywords 'terminal' and 'exhaustive', to obtain significantly enriched motifs. *$p < 0.05$. (d) Common binding motif on the atomic level. Glycan 3D structures for the binding motifs were obtained from the GLYCAM web server [25,26]. (e) Binding of CMA1 to glycosaminoglycans. We grouped chondroitin sulfate (CS) types (A, B, and C) and plotted CMA1 binding against CS chain length. Shown are mean values with their 95% confidence interval. (f) Comparison of CMA1 and RCA1 binding. Glycans with a *z*-score of at least 0.5 in at least one lectin were retained and plotted as a hierarchically clustered heatmap via the *get_heatmap* function of glycowork. Representative glycans are shown.

shortest sequences and required at least three repeats, with longer sequences such as CSC-18 even exhibiting the highest binding on the entire array (although we note that the longest GAG sequences were not generally the best binders, potentially hinting at steric clashes or density effects). Another supporting finding can be seen in the fact that CSB (exhibiting iduronic acid in α-configuration, rather than its epimer, glucuronic acid, in β-configuration) showed virtually no binding to CMA1, further arguing for contacts of the GAG chain with the binding site. Lastly, we note that both CSC and CSA contain sulfated GalNAc, which, together with the observation of GalNAc6Sβ1-4GlcNAc as one of the highest binders on the NCFG array, leads us to speculate that sulfation further enhances CMA1 binding, a pattern that has been observed for several lectins [28].

Overall, this characterized binding specificity seemed distinct from other R-type lectins and we thus further compared it to a typical R-type lectin, *Ricinus communis* agglutinin (RCA1), on the ICL array. Canonically, RCA1 binds β-linked terminal galactose residues, which is generally what we also found in our array experiments, with Galβ in various substructures and glycan types, particularly in those with multiple branches (Figure 2f). At best, the same sequences showed weak binding to CMA1, as they lacked a C2-substitution (Figure S1, Supporting Information File 2). Conversely, CMA1-favored sequences, containing Fucα1-2Gal or GalNAc epitopes, were on average not bound by RCA1 (the exception being sequences in which there was an additional free Galβ terminus). Similarly, most chondroitin sulfate probes were not bound by RCA1. This gives rise to the conclusion that CMA1 does not merely tolerate but rather actively and strongly prefers C2-substituted Gal, while RCA1 does not even tolerate these substitutions. Interestingly, we also find that fucosylation of the GlcNAc residue (as in Lewis antigen motifs) completely abrogates CMA1 binding (Figure S1, Supporting Information File 2), despite the presence of Fucα1-2Gal, likely due to steric clashes in the binding pocket. We thus conclude that the binding profile of CMA1 is distinct from that of the typical R-type lectin RCA1 and unusual for a R-type lectin in general. We also note that the flexibility of accommodated C2 substituents (from *N*-acetyl moieties to whole monosaccharides), could make CMA1 an interesting candidate for probing synthetically produced glycans with novel substituents.

It is of course interesting to speculate about the physiological role of CMA1 in melons, yet this is hard to probe. It is noteworthy, however, that the glycan types in which its preferred binding motifs occur (*O*-glycans, milk glycans, GAGs) are absent from most plants, including melons. We thus hypothesize that the role of this lectin might be to recognize non-self epitopes, such as for protection against pathogens, which is a common function in plant lectins [3].

Validating binding in solution and assessing binding affinity

As CMA1 both exhibited multiple binding sites and robust binding to blood group epitopes (H-antigen), we hypothesized that it would be capable of agglutinating red blood cells, justifying its new name. When testing the protein recombinantly produced in mammalian cells, incubation with rabbit erythrocytes indeed resulted in moderate agglutination (Figure 3a), which also demonstrated the binding to these glycan substructures in a physiological context.

To further strengthen the case for CMA1 binding glycans in solution, and corroborate its binding specificity with orthogonal methods, we used a thermal shift assay. Herein, the binding of ligands is assessed by the stabilization of the protein, measured by a denaturation curve. Both the protein produced in mammalian and in bacterial cells exhibited similar melting temperatures here, of approximately 42 °C (Figure 3b). Then, we tested the binding of CMA1 to GlcNAc, GalNAc, and H type 2 blood group antigen (BGH_{T2}; Fucα1-2Galβ1-4GlcNAcβ1-3Gal; Figure 3c). This resulted in clear melting points shifts for both GalNAc and BGH_{T2} to up to 50 °C, yet importantly not for GlcNAc, demonstrating both binding in solution and a further confirmation of the binding specificity obtained by the array experiments. We note that the functional activity of bacterially produced CMA1 indicates that potential modification by glycosylation is not required for ligand binding.

Next, we set out to quantify the binding affinity of CMA1 to its ligands. Lectins often only exhibit weak to moderate binding affinities, which is somewhat ameliorated by an increased avidity on the side of the lectin but also a dense presentation of the bound glycan epitope on the cell surface. We therefore used surface plasmon resonance (SPR) spectroscopy to derive binding constants for the interaction between CMA1 and GalNAc. A single cycle kinetics approach was applied, resulting in a measured K_D of 1.66 ± 0.08 μM (Figure 3d,e). Inhibiting binding of CMA1 to the GalNAc chip through a dilution series of *N*-acetyllactosamine (LacNAc) via multicycle kinetics allowed us to derive an IC_{50} of 1.4 μM (Figure S2a,b; Supporting Information File 2). No inhibition was observed with chondroitin 6-sulfate tetrasaccharide (CSC), and only very weak inhibition for BGH_{T2} but no IC_{50} could be determined as we could not increase the concentration to reach the plateau. For the recombinant CMA1-Nter, no binding could be observed on the GalNAc chip. This suggests either avidity effects in conjunction with the C-terminal domain or a high-affinity site on the C-terminal domain, giving rise to the measured K_D of the

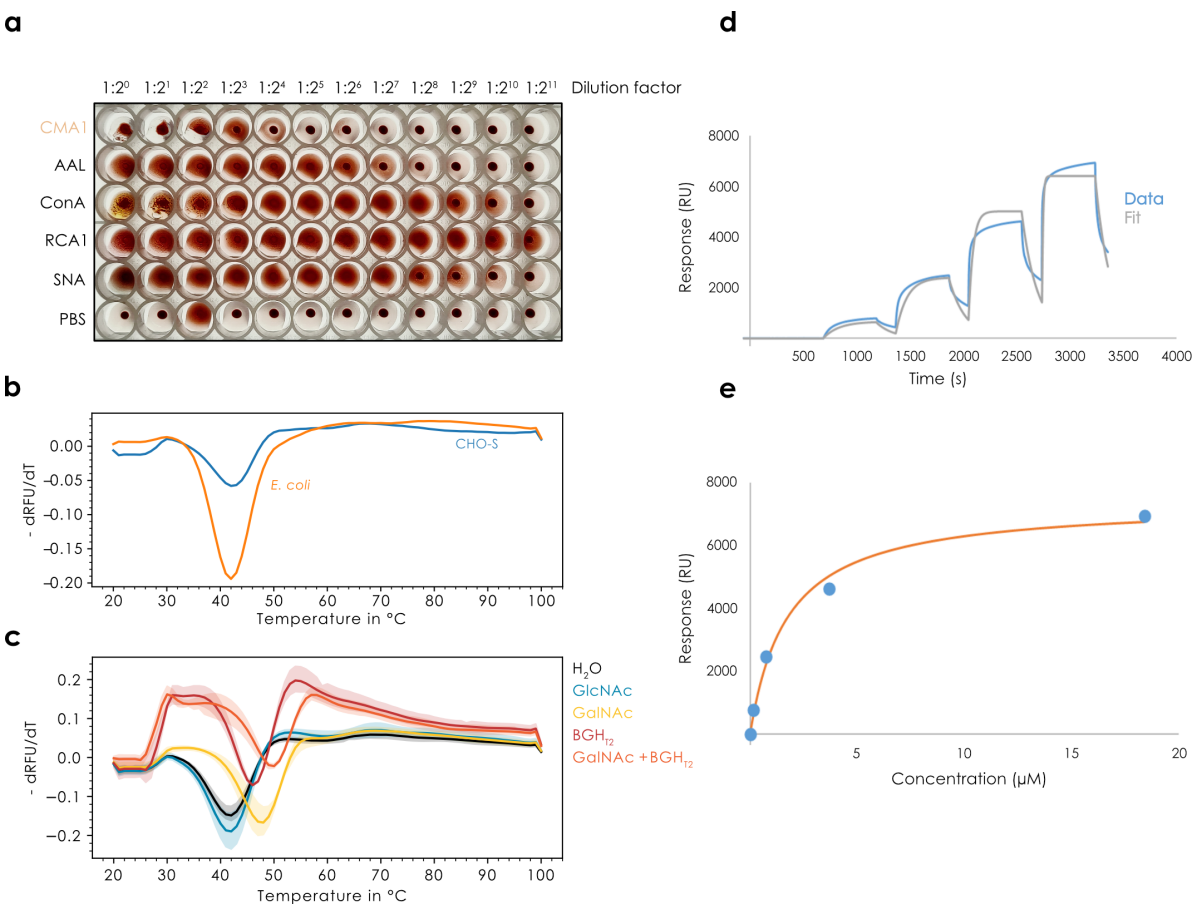


Figure 3: Assessing and quantifying in-solution binding of CMA1. (a) Erythrocyte agglutination assay. Using rabbit red blood cells, CMA1 protein recombinantly produced in mammalian cells was used in a two-fold dilution series to measure its ability to agglutinate erythrocytes, compared to other lectins, such as AAL, ConA, RCA1, and SNA-I, as well as a PBS negative control. (b, c) Thermal shift assay. After comparing the melting curves of CMA1 produced in mammalian cells (CHO-S) and bacteria (*E. coli*), we incubated the bacterially produced CMA1 with GlcNAc, GalNAc, and H type 2 blood group antigen (BGH_{T2}; Fucα1-2Galβ1-4GlcNAcβ1-3Gal) and measured a denaturation curve to assess shifts in melting temperature, $n = 3$ (c). (d, e) SPR analysis of CMA1 binding to a GalNAc chip with single-cycle kinetics and affinity measurement at the equilibrium, $n = 2$.

full-length protein. Still, we were able to measure the affinity of CMA1-Nter to GalNAc in solution by isothermal calorimetry (ITC), obtaining a K_D of 940 µM, confirming the low affinity (Figure S2c,d; Supporting Information File 2).

Structural insights from the N-terminal domain of CMA1

Given the unusual binding specificity exhibited by CMA1, we were intrigued to elucidate the molecular mechanism that would enable the specific binding of C2-substituted galactose. The natural hypothesis here would be the creation of an additional pocket in the 3D structure of the binding site, accommodating the additional substituent at C2. However, as we observed little to no binding to unsubstituted galactose, we rather hypothesized the existence of specific interactions made with the C2-substituents, that did not exist in other R-type lectins such as RCA1. To determine this, we set out to resolve the detailed three-dimensional structure of CMA1 via X-ray crystallography.

We obtained several hits for the full-length protein after sparse screening using a crystallization robot at the HTX platform, EMBL, Grenoble. Pill-shaped crystals obtained under conditions of a high salt concentration, in particular ammonium sulfate (Figure S3, Supporting Information File 2), did not give rise to any diffraction. Multiple layer plate or needles clusters were obtained in the presence of PEGs, but only showed weak diffraction (≈3.5 Å). Finally, in the presence of 20% PEG 8K, 0.2 M $MgCl_2$, and 0.1 M Tris HCl pH 8.5, single diamond-shaped crystals were obtained after 1–2 days for the N-terminal domain (Figure S3, Supporting Information File 2). High-resolution diffraction of the crystals allowed us to solve the CMA1-Nter structure in complex with LacNAc at 1.3 Å and GalNAc at 1.55 Å (see data and refinement statistics in Table 1). All

Table 1: Data collection and refinement statistics.

Complex	CMA1-Nter-LacNAc	CMA1-Nter-GalNAc
Data collection		
beamline	Soleil PX1	Soleil PX2
wavelength (Å)	0.97856	0.98011
space group	*I*2	*I*2
cell parameters *a*, *b*, *c* (Å) α, β, γ (°)	36.70 36.78 94.79 90.00 99.24 90.00	36.61 36.86 94.81 90.00 99.17 90.00
protein chains in a.u.	1	1
resolution (Å)[a]	46.78–1.32 (1.34–1.32)	35.68–1.55 (15.8–1.55)
CC1/2 (%)[a]	99.9 (96.9)	99.8 (85.7)
R_{merge} (within I+/I–)[a]	0.055 (0.369)	0.052 (0.496)
R_{meas} (within I+/I–)[a]	0.059 (0.400)	0.064 (0.618)
R_{pim} (within I+/I–)[a]	0.022 (0.153)	0.037 (0.364)
mean I/σ (I)[a]	25.2 (5.7)	14.4 (2.9)
completeness (%)[a]	99.8 (96.0)	99.7 (99.9)
number reflections[a]	399970 (18410)	95115 (4581)
number of unique reflections[a]	29695 (1434)	18279 (911)
multiplicity[a]	13.5 (12.8)	5.2 (5.0)
Wilson ***B***-factor (Å^2)	14.1	19
Refinement		
resolution (Å)	46.78–1.32	35.69–1.55
no. reflections/no. free reflections	28192/1503	17373/905
R_{work}/R_{free} (%)	14.35/18.58	16.3/20.4
R.m.s. bond lengths (Å)	0.0130	0.0127
Rmsd bond angles (°)	1.721	1.893
Rmsd chiral (Å^3)	0.097	0.092
no. atoms / Bfac (Å^2)		
protein	1029/15.1	985/19.95
ligand	26/20.3	30/22.3
cadmium	3/21.9	3/27.0
water	248/28.7	176/31.8
Ramachandran allowed (%)	100	100
favored (%)	99	100
outliers	0	0

[a]Values in parenthesis refer to the highest-resolution shell.

residues of the N-terminal construct (Val^6 to Asp^{132}) could be modelled, and unambiguous electron density permitted us to locate and model four cation binding sites (three in each structure) and one sugar binding site (Figure 4a,b and Figure S4, Supporting Information File 2).

The complexed structures allowed us to shed light on the arrangement of the ligand in the binding site (Figure 4c,d). While lectins such as CMA1 typically can present three binding pockets in their CBM13 domain, we hypothesized that the N-terminal half of CMA1 would in fact only exhibit two functional binding sites. However, only the alpha site was found occupied with a carbohydrate here. It is found in a shallow groove, supporting our data on the lack of a distinct distal binding specificity. We report a tight coordination of the O3 and O4 hydroxy groups of the galactose residue involving Asp^{21}, Asn^{43}, and Gln^{41} side chains, as well as the Gly^{24} main chain nitrogen. CH–π stacking and hydrophobic interactions occur between the aromatic ring of Trp^{36} and the alpha face of the ring as well as the hydroxymethyl moiety of the galactose residue, additionally ensuring specificity for galactoside over glucoside as an equatorial conformation of the O4 hydroxy

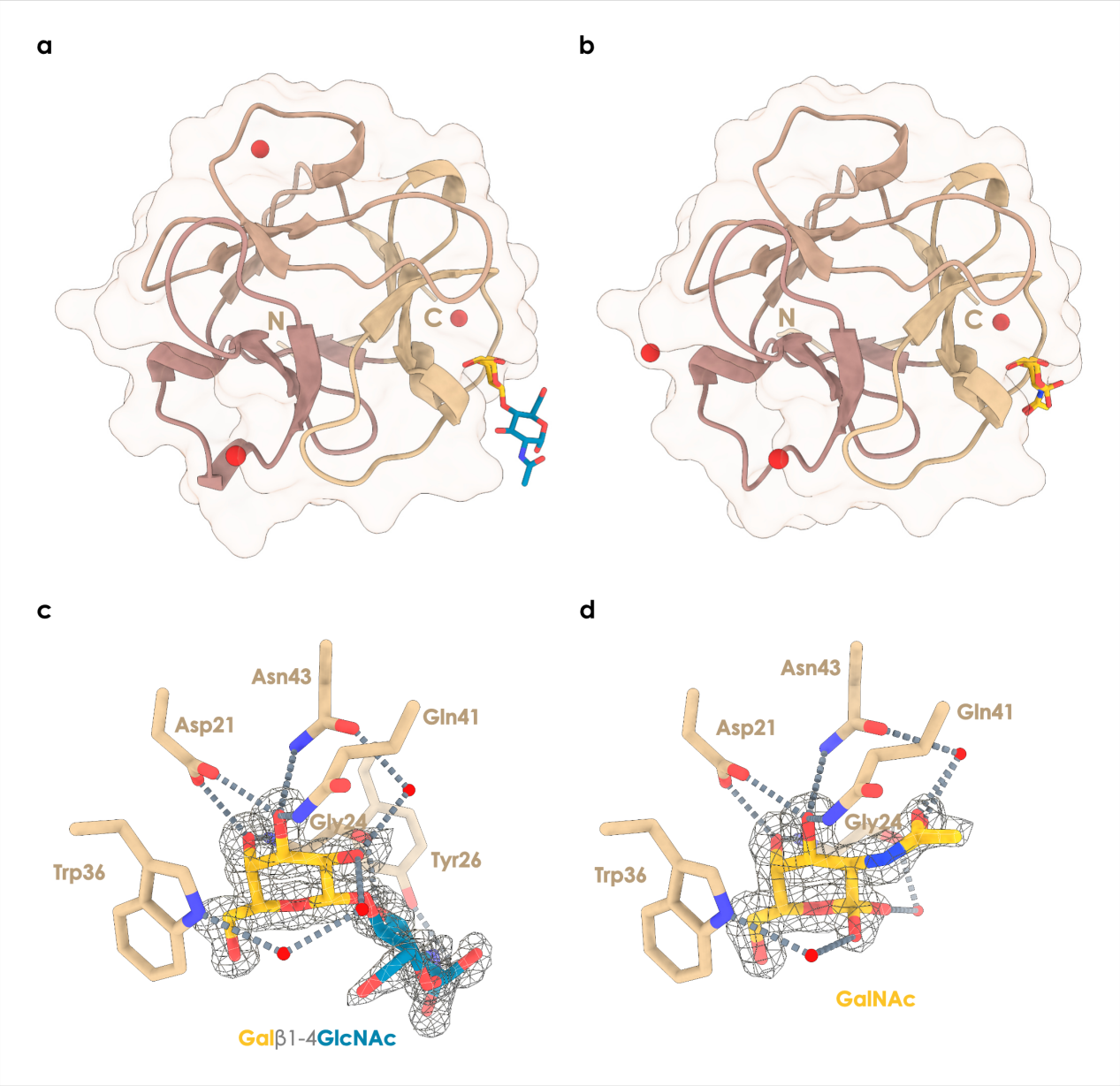


Figure 4: Structural insights into the binding mechanism of CMA1. (a, b) Overall representation of the N-terminal domain of CMA1 in complex with (a) LacNAc (Galβ1-4GlcNAc) [29] or (b) GalNAc [30]. Trefoil repeats are colored differently, and cadmium ions are represented as red spheres. (c, d) Close-up on the interactions between CMA1 and LacNAc (c) or GalNAc (d), with the 2mFo-DFc electron density map displayed around the sugar ligands at 1 sigma (LacNAc: 0.47 e·Å^{-3}, GalNAc: 0.415 e·Å^{-3}). Water molecules are indicated by red spheres and interactions by proximal residues are indicated by broken lines. The figures were prepared using UCSF ChimeraX 1.6 [31].

group would lead to steric clashes and loss of strong hydrogen bonding.

In the LacNAc-complexed structure (PDB ID 8R8A) [29], the GlcNAc residue did not seem to engage in extensive interactions, with only a hydrogen bond between the *N*-acetyl moiety and the main chain oxygen of Gly^{24} and hydrophobic interaction with the aromatic ring of Tyr^{26} (Figure 4c). Further, beyond the C2 position of galactose, a cavity filled with coordinated water molecules hinted at the binding mode for C2-substituted galactose. Notably, the seemingly inactive beta site was found to be occupied by a cadmium ion (Figure S4, Supporting Information File 2), supporting our ITC and SPR data where no multivalent binding effects were observed for the single-domain N-terminal construct.

In the GalNAc-complexed structure (PDB ID 8R8C) [30], the *N*-acetyl group of GalNAc extended beyond C2 into the cavity noted in the LacNAc complex. While no direct interactions with the protein backbone were observed, we found one water mole-

cule to mediate hydrogen bonding between the oxygen of the *N*-acetyl group and the Asn[43] side chain oxygen (Figure 4d). Both GalNAc anomers could be observed, showing interactions through water molecule coordination with the Trp[36] ring nitrogen (alpha anomer) or the Gly[24] main chain oxygen (beta anomer).

Conclusion

Our work presents a substantial exploration of the binding specificity and mechanism of the hitherto uncharacterized lectin CMA1 from melons. The binding specificity of CMA1, C2-substituted galactose that is preferentially presented in a β-configuration, enables it to bind to a range of biologically relevant epitopes, such as LacdiNAc, Sd^a, blood group H, and chondroitin sulfate motifs. Further, the inhibition of binding by the presence of Lewis antigen motifs additionally narrows it binding specificity. Our binding data and structural information lead us to the conclusion that crucially positioned asparagine residues facilitate this unusual binding specificity that delineates CMA1 from typical R-type lectins such as RCA1. Together, these results advance our knowledge of R-type lectins in general and the range of their binding specificities, but also our knowledge of melon lectins in particular, which has remained limited so far. Further experiments are still required to determine the role of the C-terminal domain, as well as the physiological function of the full-length CMA1 protein.

Experimental

Recombinant protein expression

For mammalian expression, the gene for CMA1 (A0A1S4E5V9) was synthesized with human-optimized codons and a C-terminal hexa-His tag (GSHHHHHH). We then cloned this gene into a pCI backbone (U47119; Promega GmbH) for expression in mammalian cells under a constitutive cytomegalovirus (CMV) promoter. Then, the Mammalian Protein Expression core facility at the University of Gothenburg transfected this plasmid into FreeStyle™ CHO-S cells (Cat nr R80007, ThermoFisher Scientific). Cells were cultured in Freestyle™ CHO medium at 37 °C in 5% CO_2 in Optimum Growth™ flasks (Thomson instrument company) at 130 rpm in a Multitron 4 incubator (Infors) and transfected at 2×10^6 cells/mL using FectoPro transfection reagent (Polyplus). Protein-containing culture supernatant (0.8 L) was harvested after 120 h, filtered using Polydisc AS 0.45 µm (Whatman, Cytiva) and loaded onto a 5 mL HisExcel column (GE healthcare) at 5 mL/min. The column was washed with 10 mM phosphate-buffered saline (Medicago), 500 mM NaCl and 50 mM imidazole before elution of the protein using the same buffer with a gradient from 50 mM to 500 mM imidazole (G-Biosciences) over 15 column volumes. Pooled fractions were concentrated using Vivaspin concentrators (MWCO 10 kDa, Sartorius Stedim), passed over a HiPrep 26/10 desalting column (GE Healthcare) in phosphate-buffered saline (Medicago), and finally concentrated again.

For bacterial expression, the gene of CMA1 (33–291, corresponding to residues 6–291 of the mature protein) with optimized codons for *Escherichia coli* was synthesized flanked by *Nco*I and *Xho*I restriction sites where L^6 was mutated to valine. The gene was inserted in the homemade plasmid pET40b-TEV where the enterokinase cleaving site was replaced by a TEV cleavage site by site directed mutagenesis. This plasmid was obtained by PCR using pET-40b(+) (Novagen, Merck, #70091) as template and the following primers: forward (gcccagatctgggtaccGAAAACCTGTATTTTCAGGGCGccatggcgatatcgg) and reverse (GGTACCCAGATCTGGGCTGTCCATGTGCTGGC) with complementary sequence underlined. PCR was performed using PrimeSTAR DNA polymerase (Takara #TAKR045A); then the product was digested by *Dpn*I and finally transformed in NEB5α strain (New England Biolabs, #C2992H). Both gene and vector were digested by *Nco*I and *Xho*I restriction enzymes (New England Biolabs) prior to purification on agarose gel using Monarch Gel extraction kit and supplier instructions (New England Biolabs, #T1020S) and ligation using the DNA ligation kit, Mighty Mix (Ozyme, Takara, #TAK6023Z), at room temperature to form the pET40b-TEV-CMA11 plasmid.

The N-terminal domain of CMA1 (6–132 in mature protein) was amplified by PCR using the following primers: forward (ACGCCATGGTGAGCCGTTCTACGC) and reverse (ATATCTCGAGTTAATCTG CCGTACCCCAGGATTGTGTAGG) and pET40b-TEV-CMA1 plasmid as template. Similarly, the C-terminal domain of CMA1 (136–264 in mature protein) was amplified by PCR using the subsequent primers: forward (ATTCCATGGGTCCGATTGTGGTTGCCATTGTTGG) and reverse (ACACCTCGAGTTAGGGTTTGTACTGTGTCACGAACATCC). The primers contained the restriction sites (underlined) *Nco*I (sense) and *Xho*I (antisense) on their 5′-ends for further sub-cloning. PCR was performed using PrimeSTAR DNA polymerase. The purified PCR fragment of 395 bp was digested by *Nco*I and *Xho*I restriction enzymes, then ligated into pET40b-TEV vector, and finally transformed in NEB5α strain to form the pET40b-TEV-CMA1-Nter and pET40b-TEV-CMA1-Cter plasmids. All plasmids and new vectors were verified by sequencing (Eurofins Genomics, Ebersberg, Germany). Primers were purchased from Eurofins Genomics (Ebersberg, Germany).

E. coli BL21*(DE3) [Invitrogen, #C601003] cells were transformed by heat shock at 42 °C with pET40b-TEV-CMA1 and Tuner(DE3) [Novagen, #70623] cells with pET40b-TEV-CMA1Nter prior pre-culturing in lysogeny broth (LB) [Invit-

rogen, #12780052] media containing 25 µg/mL kanamycin [Euromedex, #UK0015-A] at 37 °C, 180 rpm overnight. Then, 1 L LB medium supplemented with 25 µg/mL kanamycin was inoculated with 25 mL of the pre-culture and incubated at 37 °C, 180 rpm. When OD_{600nm} reached 0.4, the temperature was lowered to 16 °C, and when OD_{600nm} reached 0.8, protein expression was induced by the addition of 0.1 mM isopropyl β-D-thiogalactoside (IPTG) [Euromedex, #EU0008-C]. After 20 h, the cells were harvested by centrifugation at 5,000*g* for 10 min at 4 °C.

For purification of bacterial recombinant CMA1, each gram of cell pellet was resuspended with 5 mL of buffer A (20 mM Tris-HCl pH 7.5, 500 mM NaCl). After addition of 1 µL of Denarase® (C-LEcta GmbH, #20804) and moderate agitation on a rotating wheel for a period of 30 min at room temperature, cells were lysed using a cell disruptor (Constant Systems Ltd, UK) under a pressure of 2.5 kbar. The lysate was cleared by centrifugation at 24,000*g* for 30 min at 4 °C and passed through a 0.45 µm syringe filter prior to affinity chromatography purification using 1 mL HisTrap™ HP column (Cytiva) preequilibrated with buffer A and an NGC chromatography system (Bio-Rad). After loading the cleared lysate, the column was washed with buffer A + 50 mM imidazole (Sigma-Aldrich, Merck, #56749) to remove all contaminants and unbound proteins. CMA1 was eluted by a 20 mL linear gradient from 50 mM to 500 mM imidazole in buffer A. The fractions were analyzed by SDS-PAGE with 15% gel and those containing CMA1 were collected and deprived of imidazole by buffer exchange in buffer A using a Macro and Microsep Advance Spin 3 kDa MWCO centrifugal filter (Pall). The N-terminal His-tag was removed by TEV cleavage in the presence of 1 mM EDTA (Euromedex, #EU0084.B) overnight at 10 °C, using a TEV/CMA1 ratio of 1:50. TEV was prepared in-house. The protein mixture was then purified on a 1 mL HisTrap column, where pure CMA1 protein was collected in the flowthrough and column wash. Full-length CMA1 (6-291) was purified from remaining *E. coli* contaminants using a 1 mL HiTrap™ SP Sepharose FF column (Cytiva) preequilibrated with 50 mM sodium acetate pH 5.5. After loading, the column was washed, and CMA1 was eluted by a 20 mL linear gradient from 0 to 700 mM NaCl in 50 mM sodium acetate pH 5.5. The protein was concentrated and the buffer exchange to 20 mM HEPES pH 8, 100 mM NaCl using a 3 kDa MWCO centrifugal filter and stored at 4 °C.

For CMA1-Nter, the same protocol was followed, with the following changes: Purification was carried out by exploiting gravity using 1 mL of Ni Sepharose High Performance resin (Cytiva, #17.5268.01) and an Econo-Pac® Chromatography Column (Bio-Rad, #7321010). Buffer A was exchanged to buffer B (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 500 mM urea, and 5 mM imidazole). Washing steps were performed using buffer B and buffer B containing 50 mM imidazole. Elution was performed using buffer B plus 250 mM imidazole. The buffer was exchanged with 20 mM HEPES pH 7.5, 100 mM NaCl by three times 10× dilution and the sample was concentrated to at least 1 mg/mL using a 3 kDa MWCO centrifugal filter prior to TEV cleavage.

Glycan array experiments

NCFG array

For the NCFG array, data was collected by the National Center for Functional Glycomics (NCFG) at Beth Israel Deaconess Medical Center, Harvard Medical School. For experiments, a standard binding buffer (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 2 mM $CaCl_2$, 2 mM $MgCl_2$, 0.05% Tween 20, 1% BSA) was used. CMA1 binding was probed by incubation with a penta-His-488 antibody (5 µg/mL). CMA1 was tested in two concentrations (5 and 50 µg/mL) on Version 5.4 of the printed CFG array, consisting of 585 printed glycans in replicates of six. Results from replicates were combined as average RFU (raw fluorescence unit). For this average, the highest and lowest value was removed for each glycan, mitigating the effects of outliers. The results can be found in Supporting Information File 1, table “cfg”.

ICL array

For experiments, a standard binding buffer (10 mM HEPES, 150 mM NaCl, 1% BSA, 0.02% casein blocker (Pierce), 5 mM $CaCl_2$) was used. CMA1 was tested at 100 µg/mL for 1 h on the broad spectrum screening array (in house designation ‘Array Sets 42–56’) of the Glycoscience Laboratory at Imperial College London, consisting of 866 lipid-linked glycans. Then the detecting solution composed of anti-polyHistidine (Sigma-Aldrich, Merck, SAB4200620) and biotin anti-mouse IgG (Sigma-Aldrich, Merck, B7264) antibodies (10 µg/mL, precomplexed in a ratio of 1:1) was overlaid onto the arrays for 1 h. The final detection was with a 30 min overlay of streptavidin-Alexa Fluor 647 (Molecular Probes) at 1 µg/mL. The microarray slides were scanned with GenePix 4300A scanner instrument (50% laser power at PMT 350), and the image analysis (quantitation) was performed with GenePix® Pro 7 software. The results can be found in Supporting Information File 1, table “imperial” and “rca_imperial”, with the array generation in Supporting Information File 3 according to the MIRAGE guidelines (Minimum Information Required for A Glycomics Experiment) [32].

For both array types, data were transformed into *z*-scores by subtracting the mean value across the array and dividing the results by the standard deviation.

Agglutination assay

The hemagglutinating activity of CMA1 was determined in V-bottom 96-well plates by a twofold serial dilution procedure in PBS using rabbit red blood cells (Atlantis France). 25 µL of 4% erythrocyte suspension was added to an equal volume of the sample, and the mixture was incubated for 60 min at room temperature. Starting concentrations were: CMA1 0.6 mg/mL, AAL 0.5 mg/mL, ConA 2.5 mg/mL, RCA1 2.5 mg/mL, and SNA 0.5 mg/mL.

Thermal shift assay

Thermal shift assays were performed using a Mini Opticon Real Time PCR machine (BioRad). 0.6 mg/mL protein in PBS was mixed with SYPRO Orange (Sigma-Aldrich, Merck, #S5692) and glycan ligand (10 mM GalNAc; Carbosynth, #MA04390; 10 mM GlcNAc, Carbosynth, #MA00834; 10 mM blood group H type-2 tetrasaccharide; Elicityl, GLY032-2) in a total reaction volume of 25 µL. The temperature was raised by 1 °C/min from 25 to 100 °C, and fluorescence readings were taken at each step.

Surface plasmon resonance spectroscopy

Experiments were performed using a Biacore X100 instrument (Cytiva) at 25 °C in HBS-T running buffer (10 mM HEPES pH 7.4, 150 mM NaCl and 0.05% Tween 20). Biotinylated PAA-GalNAc (Lectinity, GlycoNZ, #0031-BP) was immobilized on CM5 chips (Cytiva #BR100012) that were coated previously with streptavidin (Sigma-Aldrich, Merck, #S4762), following standard protocol. Biotinylated GalNAc was diluted to 2 µg/mL in HBS-T before being injected into one of the flow cells of the chip. An immobilization level of 900 response units (RU) was obtained. A reference surface was always present in flow cell 1, allowing for the subtraction of bulk effects and non-specific interactions with streptavidin. The mammalian-produced CMA1 was injected in single cycle kinetic over the flow cell surface at 10 µL/min at increasing concentrations with a contact time of 500 s. Dissociation was achieved by passing running buffer for 2 min. Surfaces were regenerated with four consecutive 30 s injections of 50 mM NaOH and 1 M NaCl. Binding affinity (K_D) was measured after subtracting the channel 1 reference (streptavidin only) and subtracting a blank injection (running buffer – zero analyte concentration). Data evaluation and curve fitting was performed using the provided BIACORE X100 evaluation software (version 2.0). Measurements were at least done in duplicate.

Then, to perform competition experiments, nine concentrations of LacNAc (Elicityl, #GLY008) from 10 to 0 mM with a dilution coefficient of two supplemented with a fixed concentration of 0.8 µM was injected into the cell surface in multiple cycle kinetic with an association time of 500 s and a dissociation time of 12 s at a flow rate of 10 µL/min. Surfaces were regenerated with 30 s injections of 50 mM NaOH and 1 M NaCl. IC_{50} was measured using the response at equilibrium for each concentration of competitive sugar that were translated in percentage of inhibition, then plotted against the molar concentration of competitive sugar using the free software "data entry". The IC_{50} was calculated using https://www.aatbio.com/tools/ic50-calculator.

X-ray crystallography

All consumables for crystallization and crystal handling were purchased at Molecular Dimensions, Calibre Scientific, Rotherham, UK, unless stated otherwise. CMA1 concentrated at 5.7 or 3.5 mg/mL in 20 mM HEPES pH 8, 100 mM NaCl, and 14 mM GalNAc was subjected to crystallization screening using the robotized HTXlab platform (EMBL, Grenoble, France) with 200 nL sitting drops at 20 °C using a 1:1 ratio. Wizard I and II screen (Rigaku) and SaltRX (Hampton Research) screens were used and led to more than 30 hits after one to three days. Pill-like crystals were obtained with high salt concentration that could be reproduced by hand in the laboratory. Plates and needles clusters were obtained with PEG containing solutions. For CMA1-Nter, protein at a concentration of 2.9–3.5 mg/mL was crystallized using hanging drop and vapor diffusion methods with a 2 µL drop in 1:1 ratio at 20 °C. Bipyramidal single crystals were obtained after one or two days in a solution containing 10–12% PEG Smear Medium, 0.1 M MES pH 6.5, 1× divalent (5 mM of $CaCl_2$, $MgCl_2$, $CsCl_2$, $CdCl_2$, $NiCl_2$, and zinc acetate), or 5 mM $CdCl_2$, and in the presence or not of 5 mM GalNAc. Cocrystals of CMA1-Nter in complex with LacNAc (Galβ1-4GlcNAc, Elicityl, #GLY008) were obtained by the addition of 5 mM LacNAc to the protein solution and incubation at room temperature for 30 min prior to crystallization. For both complexes, single crystals were mounted in a cryoloop after transfer in a cryoprotectant solution, composed of 30% PEG Smear Medium and 5 mM $CdCl_2$, and flash-cooled in liquid nitrogen. Crystal diffraction was evaluated, and data were collected on the Proxima 1 and 2 beamlines at the synchrotron SOLEIL, Saint Aubin, France using an Eiger 16M or 9M detector (Table 1) for LacNAc and GalNAc complexed structures, respectively. XDS and XDSME were used to process the data and all further steps were performed using programs of the CCP4 suite version 8.25–27 [33-35]. The model coordinates predicted by Alphafold [36] Monomer v2.0 for the monomer of CMA1 (A0A1S4E5V9) were trimmed to only include the N-terminal domain (residues 33–159), with all B-factors reset to 15 $Å^2$, to be subsequently used as a search model to solve the structure of CMA1-Nter by molecular replacement using PHASER [37]. Multiple iterations of anisotropic restrained maximum likelihood refinement using REFMAC 5.8 [38] and manual building using Coot [39] were performed.

Hydrogen atoms were added in their riding positions during refinement and 5% of the observations were set aside for cross-validation analysis. Upon inspection of the electron density maps, carbohydrate moieties were introduced and checked using Privateer [40]. The final model was validated using the wwPDB validation server (https://validate-rcsb-1.wwpdb.org). Structure figures were made using PyMol 2.5.7 and ChimeraX 1.6 [31]. The parameters for CH−π interactions were defined as previously reported [41,42].

Supporting Information

Supporting Information File 1

Full array data regarding the binding specificity of CMA1.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-31-S1.xlsx]

Supporting Information File 2

Additional figures.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-31-S2.pdf]

Supporting Information File 3

Supplementary glycan microarray document (MIRAGE) for the ICL glycan arrays.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-31-S3.pdf]

Acknowledgements

We acknowledge support from the Mammalian Protein Expression core facility at the University of Gothenburg via the Protein Production Sweden (PPS) framework as well as the synchrotron SOLEIL (Saint Aubin, France) for access to beamline Proxima 1 and 2 (Proposal Number 20210859) and for the technical support of Pierre Legrand and Martin Savko, respectively. The authors would like to thank Iris Lopez and Federico Musso for their technical help in the expression assays of CMA1-Nter and CMA1 purification trials, respectively, and Wengang Chai for preparing the GAG probes on the ICL array.

Funding

This work was funded by a Branco Weiss Fellowship – Society in Science awarded to D.B., by the Knut and Alice Wallenberg Foundation, and the University of Gothenburg, Sweden as well as support from the GLYCONanoPROBES (CA18132) and INNOGLY (CA18103) COST actions awarded to J.L. This work was further supported by the Protein-Glycan Interaction Resource of the CFG and the National Center for Functional Glycomics (NCFG) at Beth Israel Deaconess Medical Center, Harvard Medical School (supporting grant R24 GM137763). The glycan microarray studies were performed in the Carbohydrate Microarray Facility at the ICL Glycosciences Laboratory, which is supported by Wellcome Trust biomedical resource grants (099197/Z/12/Z, 108430/Z/15/Z, and 218304/Z/19/Z) and partially by the March of Dimes Prematurity research centre grant (22-FY18-82). The sequence-defined glycan microarrays contain many saccharides provided by collaborators whom we thank, as well as members of the Glycosciences Laboratory for their contribution in the establishment of the NGL-based microarray system. This work benefited from access to EMBL HTX lab, which has been supported by iNEXT-Discovery, project number 871037, funded by the Horizon 2020 program of the European Commission.

ORCID® iDs

Nicole Kerekes - https://orcid.org/0000-0001-7065-8092
Antonio Di Maio - https://orcid.org/0000-0002-2740-9098
Ten Feizi - https://orcid.org/0000-0001-6495-0329
Annabelle Varrot - https://orcid.org/0000-0001-6667-8162
Daniel Bojar - https://orcid.org/0000-0002-3008-7851

Data Availability Statement

All generated data here can be found in the Supporting Information. The coordinates of CMA1 in complex with LacNAc (PDB ID 8R8A), https://doi.org/10.2210/pdb8R8A/pdb, and GalNAc (PDB ID 8R8C) https://doi.org/10.2210/pdb8R8C/pdb, have been deposited in the Protein Data Bank (PDB).

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.1101/2023.11.30.569503

References

1. Sharon, N. *Glycobiology* **2004,** *14,* 53R–62R. doi:10.1093/glycob/cwh122
2. Damme, E. J. M. V.; Peumans, W. J.; Barre, A.; Rougé, P. *Crit. Rev. Plant Sci.* **1998,** *17,* 575–692. doi:10.1080/07352689891304276
3. Lannoo, N.; Van Damme, E. J. M. *Front. Plant Sci.* **2014,** *5,* 397. doi:10.3389/fpls.2014.00397
4. Keller, L.-A.; Niedermeier, S.; Claassen, L.; Popp, A. *Acta Histochem.* **2022,** *124,* 151877. doi:10.1016/j.acthis.2022.151877
5. Kearney, C. J.; Vervoort, S. J.; Ramsbottom, K. M.; Todorovski, I.; Lelliott, E. J.; Zethoven, M.; Pijpers, L.; Martin, B. P.; Semple, T.; Martelotto, L.; Trapani, J. A.; Parish, I. A.; Scott, N. E.; Oliaro, J.; Johnstone, R. W. *Sci. Adv.* **2021,** *7,* eabe3610. doi:10.1126/sciadv.abe3610
6. Minoshima, F.; Ozaki, H.; Odaka, H.; Tateno, H. *iScience* **2021,** *24,* 102882. doi:10.1016/j.isci.2021.102882
7. Choi, S. H.; Lyu, S. Y.; Park, W. B. *Arch. Pharmacal Res.* **2004,** *27,* 68. doi:10.1007/bf02980049
8. Hirabayashi, J.; Yamada, M.; Kuno, A.; Tateno, H. *Chem. Soc. Rev.* **2013,** *42,* 4443. doi:10.1039/c3cs35419a

9. Pilobello, K. T.; Slawek, D. E.; Mahal, L. K. *Proc. Natl. Acad. Sci. U. S. A.* **2007,** *104,* 11534–11539. doi:10.1073/pnas.0704954104
10. Qin, R.; Meng, G.; Pushalkar, S.; Carlock, M. A.; Ross, T. M.; Vogel, C.; Mahal, L. K. *J. Proteome Res.* **2022,** *21,* 1974–1985. doi:10.1021/acs.jproteome.2c00251
11. Heindel, D. W.; Chen, S.; Aziz, P. V.; Chung, J. Y.; Marth, J. D.; Mahal, L. K. *ACS Infect. Dis.* **2022,** *8,* 1075–1085. doi:10.1021/acsinfecdis.2c00082
12. Bonnardel, F.; Mariethoz, J.; Pérez, S.; Imberty, A.; Lisacek, F. *Nucleic Acids Res.* **2021,** *49,* D1548–D1554. doi:10.1093/nar/gkaa1019
13. Taylor, M. E.; Drickamer, K. *Curr. Opin. Struct. Biol.* **2014,** *28,* 14–22. doi:10.1016/j.sbi.2014.07.003
14. Bojar, D.; Meche, L.; Meng, G.; Eng, W.; Smith, D. F.; Cummings, R. D.; Mahal, L. K. *ACS Chem. Biol.* **2022,** *17,* 2993–3012. doi:10.1021/acschembio.1c00689
15. Wu, A. M.; Wu, J. H.; Singh, T.; Lai, L.-J.; Yang, Z.; Herp, A. *Mol. Immunol.* **2006,** *43,* 1700–1715. doi:10.1016/j.molimm.2005.09.008
16. Swamy, M. J.; Bobbili, K. B.; Mondal, S.; Narahari, A.; Datta, D. *Phytochemistry* **2022,** *201,* 113251. doi:10.1016/j.phytochem.2022.113251
17. Allen, A. K. *Biochem. J.* **1979,** *183,* 133–137. doi:10.1042/bj1830133
18. Wang, H.; Ng, T. B. *Biochem. Biophys. Res. Commun.* **1998,** *253,* 143–146. doi:10.1006/bbrc.1998.9765
19. Shin, A.-Y.; Koo, N.; Kim, S.; Sim, Y. M.; Choi, D.; Kim, Y.-M.; Kwon, S.-Y. *Sci. Data* **2019,** *6,* 220. doi:10.1038/s41597-019-0244-x
20. Notova, S.; Bonnardel, F.; Rosato, F.; Siukstaite, L.; Schwaiger, J.; Lim, J. H.; Bovin, N.; Varrot, A.; Ogawa, Y.; Römer, W.; Lisacek, F.; Imberty, A. *Commun. Biol.* **2022,** *5,* 954. doi:10.1038/s42003-022-03869-w
21. Cummings, R. D.; Schnaar, R. L.; Ozeki, Y. R-Type Lectins. In *Essentials of Glycobiology;* Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Mohnen, D.; Kinoshita, T.; Packer, N. H.; Prestegard, J. J.; Schnaar, R. L.; Seeberger, P. H., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2022.
22. Edgar, R. C. *Nucleic Acids Res.* **2004,** *32,* 1792–1797. doi:10.1093/nar/gkh340
23. Lundstrøm, J.; Urban, J.; Thomès, L.; Bojar, D. *Glycobiology* **2023,** *33,* 927–934. doi:10.1093/glycob/cwad063
24. Thomès, L.; Burkholz, R.; Bojar, D. *Glycobiology* **2021,** *31,* 1240–1244. doi:10.1093/glycob/cwab067
25. GLYCAM-Web | Utilities for molecular modeling of carbohydrates. https://glycam.org/ (accessed Feb 8, 2024).
26. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008,** *29,* 622–655. doi:10.1002/jcc.20820
27. Pohleven, J.; Obermajer, N.; Sabotič, J.; Anžlovar, S.; Sepčić, K.; Kos, J.; Kralj, B.; Štrukelj, B.; Brzin, J. *Biochim. Biophys. Acta, Gen. Subj.* **2009,** *1790,* 173–181. doi:10.1016/j.bbagen.2008.11.006
28. Jung, J.; Enterina, J. R.; Bui, D. T.; Mozaneh, F.; Lin, P.-H.; Nitin; Kuo, C.-W.; Rodrigues, E.; Bhattacherjee, A.; Raeisimakiani, P.; Daskhan, G. C.; St. Laurent, C. D.; Khoo, K.-H.; Mahal, L. K.; Zandberg, W. F.; Huang, X.; Klassen, J. S.; Macauley, M. S. *ACS Chem. Biol.* **2021,** *16,* 2673–2689. doi:10.1021/acschembio.1c00501
29. Lundstrøm, J.; Varrot, A. Structure of the N-terminal domain of CMA in complex with N-acetyllactosamine. https://www.wwpdb.org/pdb?id=pdb_00008r8a (accessed Feb 12, 2024). doi:10.2210/pdb8r8a/pdb
30. Varrot, A. Structure of the N-terminal domain of CMA from *Cucumis melo* in complex with N-acetylgalactosamine. https://www.wwpdb.org/pdb?id=pdb_00008r8c (accessed Feb 12, 2024). doi:10.2210/pdb8r8c/pdb
31. Meng, E. C.; Goddard, T. D.; Pettersen, E. F.; Couch, G. S.; Pearson, Z. J.; Morris, J. H.; Ferrin, T. E. *Protein Sci.* **2023,** *32,* e4792. doi:10.1002/pro.4792
32. Liu, Y.; McBride, R.; Stoll, M.; Palma, A. S.; Silva, L.; Agravat, S.; Aoki-Kinoshita, K. F.; Campbell, M. P.; Costello, C. E.; Dell, A.; Haslam, S. M.; Karlsson, N. G.; Khoo, K.-H.; Kolarich, D.; Novotny, M. V.; Packer, N. H.; Ranzinger, R.; Rapp, E.; Rudd, P. M.; Struwe, W. B.; Tiemeyer, M.; Wells, L.; York, W. S.; Zaia, J.; Kettner, C.; Paulson, J. C.; Feizi, T.; Smith, D. F. *Glycobiology* **2017,** *27,* 280–284. doi:10.1093/glycob/cww118
33. Kabsch, W. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010,** *66,* 125–132. doi:10.1107/s0907444909047337
34. legrandp/xdsme: March 2019 version working with the latest XDS version (Jan 26, 2018). https://zenodo.org/records/2613389 (accessed Feb 8, 2024). doi:10.5281/zenodo.837885
35. Agirre, J.; Atanasova, M.; Bagdonas, H.; Ballard, C. B.; Baslé, A.; Beilsten-Edmands, J.; Borges, R. J.; Brown, D. G.; Burgos-Mármol, J. J.; Berrisford, J. M.; Bond, P. S.; Caballero, I.; Catapano, L.; Chojnowski, G.; Cook, A. G.; Cowtan, K. D.; Croll, T. I.; Debreczeni, J. É.; Devenish, N. E.; Dodson, E. J.; Drevon, T. R.; Emsley, P.; Evans, G.; Evans, P. R.; Fando, M.; Foadi, J.; Fuentes-Montero, L.; Garman, E. F.; Gerstel, M.; Gildea, R. J.; Hatti, K.; Hekkelman, M. L.; Heuser, P.; Hoh, S. W.; Hough, M. A.; Jenkins, H. T.; Jiménez, E.; Joosten, R. P.; Keegan, R. M.; Keep, N.; Krissinel, E. B.; Kolenko, P.; Kovalevskiy, O.; Lamzin, V. S.; Lawson, D. M.; Lebedev, A. A.; Leslie, A. G. W.; Lohkamp, B.; Long, F.; Malý, M.; McCoy, A. J.; McNicholas, S. J.; Medina, A.; Millán, C.; Murray, J. W.; Murshudov, G. N.; Nicholls, R. A.; Noble, M. E. M.; Oeffner, R.; Pannu, N. S.; Parkhurst, J. M.; Pearce, N.; Pereira, J.; Perrakis, A.; Powell, H. R.; Read, R. J.; Rigden, D. J.; Rochira, W.; Sammito, M.; Sánchez Rodríguez, F.; Sheldrick, G. M.; Shelley, K. L.; Simkovic, F.; Simpkin, A. J.; Skubak, P.; Sobolev, E.; Steiner, R. A.; Stevenson, K.; Tews, I.; Thomas, J. M. H.; Thorn, A.; Valls, J. T.; Uski, V.; Usón, I.; Vagin, A.; Velankar, S.; Vollmar, M.; Walden, H.; Waterman, D.; Wilson, K. S.; Winn, M. D.; Winter, G.; Wojdyr, M.; Yamashita, K. *Acta Crystallogr., Sect. D: Struct. Biol.* **2023,** *79,* 449–461. doi:10.1107/s2059798323003595
36. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. *Nature* **2021,** *596,* 583–589. doi:10.1038/s41586-021-03819-2
37. McCoy, A. J. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2007,** *63,* 32–41. doi:10.1107/s0907444906045975

38. Murshudov, G. N.; Skubák, P.; Lebedev, A. A.; Pannu, N. S.; Steiner, R. A.; Nicholls, R. A.; Winn, M. D.; Long, F.; Vagin, A. A. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2011,** *67,* 355–367. doi:10.1107/s0907444911001314
39. Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010,** *66,* 486–501. doi:10.1107/s0907444910007493
40. Agirre, J.; Iglesias-Fernández, J.; Rovira, C.; Davies, G. J.; Wilson, K. S.; Cowtan, K. D. *Nat. Struct. Mol. Biol.* **2015,** *22,* 833–834. doi:10.1038/nsmb.3115
41. Hudson, K. L.; Bartlett, G. J.; Diehl, R. C.; Agirre, J.; Gallagher, T.; Kiessling, L. L.; Woolfson, D. N. *J. Am. Chem. Soc.* **2015,** *137,* 15152–15160. doi:10.1021/jacs.5b08424
42. Brandl, M.; Weiss, M. S.; Jabs, A.; Sühnel, J.; Hilgenfeld, R. *J. Mol. Biol.* **2001,** *307,* 357–377. doi:10.1006/jmbi.2000.4473

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Monitoring carbohydrate 3D structure quality with the *Privateer* database

Jordan S. Dialpuri, Haroldas Bagdonas, Lucy C. Schofield, Phuong Thao Pham, Lou Holland and Jon Agirre*

Full Research Paper

Open Access

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

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

* Corresponding author

Keywords:
carbohydrates; database; *N*-glycans; *N*-glycosylation; polysaccharides; validation; website

Beilstein J. Org. Chem. **2024,** *20,* 931–939.
https://doi.org/10.3762/bjoc.20.83

Received: 30 January 2024
Accepted: 10 April 2024
Published: 24 April 2024

This article is part of the thematic issue "Chemical glycobiology".

Guest Editor: E. Fadda

Abstract

The remediation of the carbohydrate data of the Protein Data Bank (PDB) has brought numerous enhancements to the findability and interpretability of deposited glycan structures, yet crucial quality indicators are either missing or hard to find on the PDB pages. Without a way to access wider glycochemical context, problematic structures may be taken as fact by keen but inexperienced scientists. The *Privateer* software is a validation and analysis tool that provides access to a number of metrics and links to external experimental resources, allowing users to evaluate structures using carbohydrate-specific methods. Here, we present the *Privateer* database, a free resource that aims to complement the growing glycan content of the PDB.

Introduction

Carbohydrate modelling is an important but often cumbersome stage in the macromolecular X-ray structure solution workflow. The accurate modelling of glycoproteins and protein–carbohydrate complexes is pivotal in understanding the complex biochemical interactions that affect the physiological function of cells [1]. Any mechanistic analysis done with finely grained approaches such as QM/MM [2] relies heavily on the correctness of the starting coordinates. Despite this, carbohydrate models often contain modelling inconsistencies that cannot easily be attributed to known biochemical principles [3]. These inconsistencies cannot solely be attributed to model-building inexperience, as carbohydrate model building is an inherently difficult task, which in the past has been plagued with software related problems from incorrect libraries to incomplete support [4]. Carbohydrates are mobile, highly branched additions to the comparatively rigid protein framework; in macromolecular crystallography, this causes heterogeneity throughout the crystal lattice and, therefore, poorly resolved density regions, whereas

in electron cryo-microscopy different conformations and compositions are averaged out during image classification and volume reconstruction [5].

Owing to these difficulties, it is not uncommon to find problematic carbohydrate structures in the Protein Data Bank (PDB), from the initial works of Lütteke, Frank and von der Lieth [6,7], who identified numerous issues affecting nomenclature and linkages (estimated to affect 30% of the structures at the time), to the reports of surprising – or indeed glyco-chemically impossible – linkages in a glycoprotein as pointed out by Crispin and collaborators [8], and more recently the realisation that high-energy ring conformations, a rare event in six-membered pyranosides, were present in ca. 15% of the *N*-glycan components of glycoproteins in the PDB [3]. Many of these findings originated the development of new resources, including services and databases [9-13], and standalone software [14-18]. Among these, the *Privateer* software package has been a key tool for glycoprotein and protein–carbohydrate complex validation: *Privateer* analyses the conformational plausibility of each sugar model [3], checks that structures match the nomenclature used for deposition in the PDB [14], compares glycan compositions to known structures as reported by glycomics (e.g., GlyConnect [19]) and glyco-informatics (e.g., GlyTouCan [20]) databases and repositories [15], and checks how close the overall conformation of *N*-glycans comes to that of validated deposited structures [16].

The PDB-REDO [21] database is a separate resource, albeit linked to the PDB in that the entries that compound PDB-REDO are those original PDB crystallographic entries that included experimental data (i.e., reflection intensities or amplitudes); each entry includes a re-refined, sometimes even re-built to some extent, copy of the original model. These newer versions are produced with state-of-the-art methods, many of which were probably not available at the time of deposition; hence, the quality of the models is expected to improve. Because the methodology included in PDB-REDO had been affected by the lack of automatic support that plagued general purpose crystallographic model building and refinement software [4], carbohydrate-specific methods have been gradually introduced over the years [22,23].

Whilst *Privateer* has been a staple tool in carbohydrate validation, the results of *Privateer* have not been collated in such a way that allows for easy judgement of carbohydrate model quality in the PDB [24]. Providing users with metrics that allow them to make chemically sound conclusions about the model is an important facility, especially for novice users. To allow this to happen readily on PDB distribution sites, we present the *Privateer* database, a freely available, up-to-date collection of validation information for both the PDB and PDB-REDO [21] archives.

Results and Discussion

Format of the validation report

The JSON file deposited for each PDB entry follows a consistent format, as shown in Figure 1. At the top level, the file contains metadata about the validation report. This metadata provides the date that the validation report was generated as well as the availability of experimental data. It is helpful to have this information easily accessible as *Privateer* cannot calculate the real space correlation coefficient without experimental data; therefore, programmatic access to further validation metrics could be streamlined, knowing the information is not present.

Also at the top level of the validation report is the beginning of the carbohydrate information, listed as ‘glycans’ in the JSON format. Within this ‘glycan’ scope, information is segmented into glycan types, that is, ‘n-glycan’, ‘o-glycan’, ‘s-glycan’, ‘c-glycan’, and 'ligand'. Each of these glycan types contains an array of individual glycans of that type, and the format of the data inside each of these glycan types is identical.

The data contained in each glycan entry is shown in Table 1. Each entry contains information about the protein chain attachment, the number of sugars in the glycan, the WURCS2.0 code [25], the standard nomenclature for glycan SVG, and an array of sugar entries. The validation data calculated by *Privateer* for each sugar entry is shown in Table 2, and that for each linkage is shown in Table 3.

Visualising a validation report

While the database is available on GitHub for programmatic access, viewing a validation report entry in plaintext can be difficult, time-consuming and would certainly be a poor experience for the end user. To improve the utility of this database, we have provided a visualisation of the information contained within the validation report for both PDB and PDB-REDO databases, which is available alongside the *Privateer Web App* [26], https://privateer.york.ac.uk/database.

The first section of this visual report displays a global outlook on the validity of the model through two graphs. The first graph shows the conformational landscape for the pyranose sugars. For a sugar model to be deemed valid, the ring must be in the $^4C_1$ chair conformation. This can be measured through the Cremer–Pople parameters θ and ψ [27]. Theta angles of $0° < \theta < 360°$ indicate that the sugar may be in a higher-energy confirmation; therefore, caution should be placed on any conclusions drawn from the molecular model of the sugar. Also

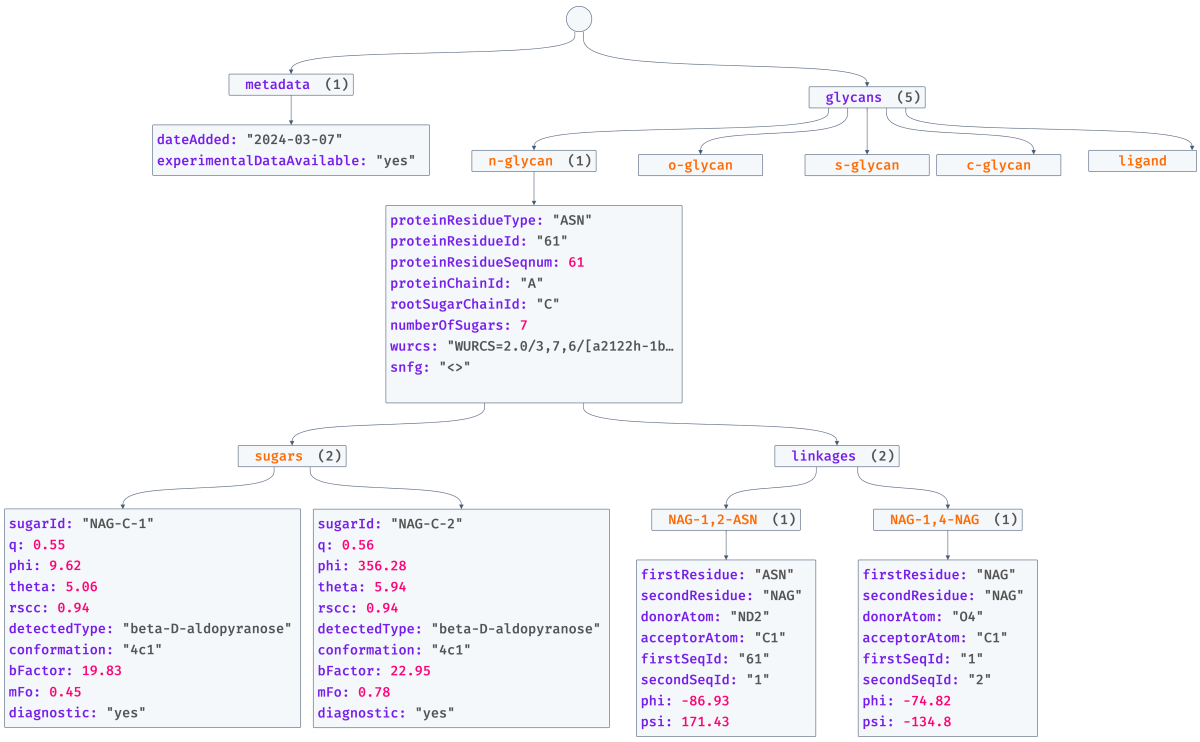


Figure 1: Format of a validation report in JSON format. At the top level of the tree, the report contains metadata about itself, such as the date the entry was added to the database and if experimental data is available. Also at the top level of the tree is the glycan information, separated into glycan types. Each glycan also contains a list of sugars, with a range of validation information and a list of linkage with torsion angle information. Tree visualisation was created with jsoncrack.com.

Table 1: Data contained within each glycan entry.

Key	Example	Type
proteinResidueType	ASN	string
proteinResidueId	61	string
proteinResidueSeqnum	61	number
proteinChainId	A	string
rootSugarChainId	C	string
numberOfSugars	7	number
wurcs	WURCS=2.0/3,7,6/…	string
snfg	<svg> … </svg>	string
sugars	see Table 2	array

Table 2: Data contained within each sugar entry.

Key	Example	Type
sugarID	NAG-D-1	string
q	0.54	number
phi	303.44	number
theta	6.45	number

Table 2: Data contained within each sugar entry. (continued)

rscc	0.922	number
detectedType	beta-D-aldopyranose	string
conformation	4c1	string
bFactor	22.367	number
mFo	0.421	number
diagnostic	yes	string

Table 3: Data contained within each linkage entry.

Key	Example	Type
firstResidue	NAG	string
secondResidue	NAG	string
donorAtom	O4	string
acceptorAtom	C1	string
firstSeqId	1	string
secondSeqId	2	string
phi	−54.91	number
psi	−108.47	number

in the first section of the visual validation report is a plot of the B-factor (temperature factor) versus the real space correlation coefficient (RSCC) (Figure 2). A well-refined, well-built model would be expected to have a B-factor that increases somewhat linearly as the RSCC decreases. Over-refined models may deviate from this trend and would be trivial to identify.

The validation report also displays a table (Figure 3) representing two-dimensional descriptions of each glycan in the

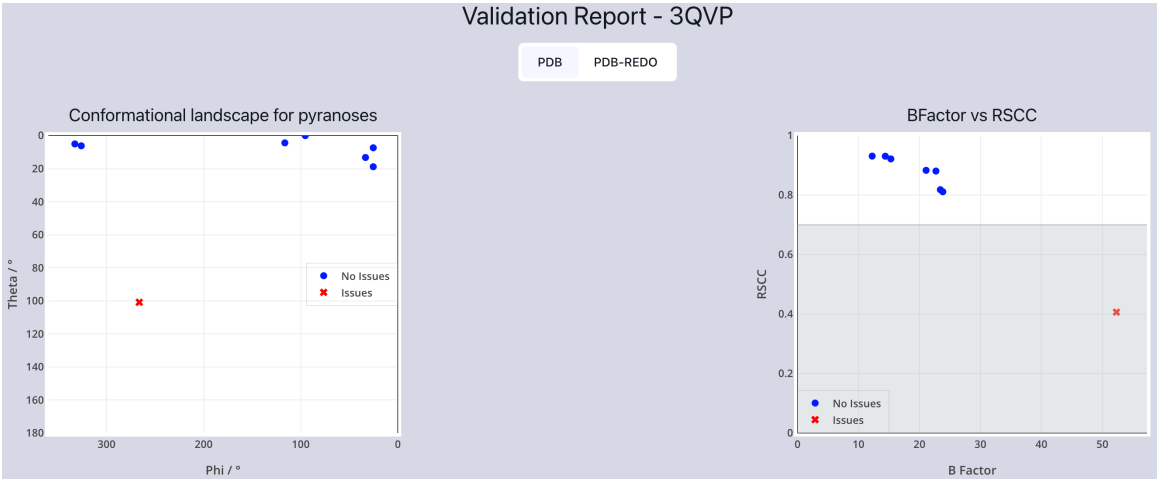


Figure 2: Left: Graphical representation of the conformational landscape of pyranose sugars. A well-modelled D-sugar would be expected to be in the lowest-energy conformation and have a theta angle close to 0° and would be indicated by a blue point; deviations from the ideal conformation are highlighted with a red cross. Right: Real space correlation coefficient plotted against the B-factor, which enables the refinement of the sugars to be assessed. A slight negative correlation would be expected for a well-refined model. Results taken from the *Privateer* database report for 3QVP [28].

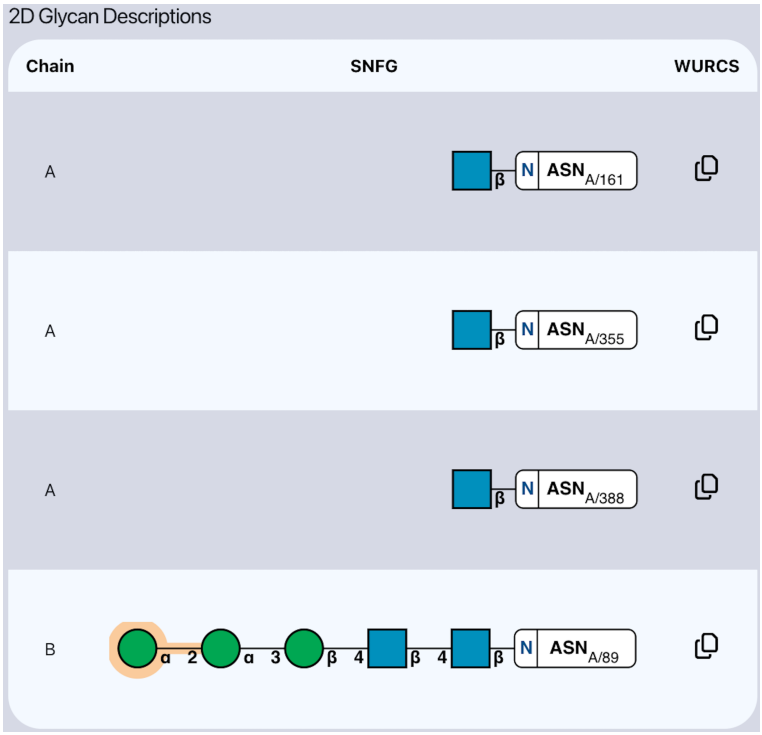


Figure 3: Table of two-dimensional Symbol Nomenclature for Glycan (SNFG) visualisations, which can allow for easy oversight of the validity of a particular glycan. Sugars that have issues identified by *Privateer* are highlighted in orange, and linkages that have unusual torsion angles are also highlighted in orange. The WURCS codes for each glycan are also available to copy to the clipboard. Table taken from the *Privateer* database report for 3QVP.

model. Each row in the table represents a unique glycan and includes the chain identifier, standard Symbol Nomenclature for Glycans (SNFG [29]) visualisation, and copyable WURCS [25] identifier. The SNFG displayed for each glycan paints a picture of how well built the glycan model is, as the metrics and validity conclusions calculated by *Privateer* are embedded within each shape and linkage of the diagram. For example, a shape with an orange highlight indicates something is abnormal about the ring's conformation, puckering, or monosaccharide nomenclature [30]. Similarly, a linkage with an orange highlight indicates that the torsion angles between the linkages are unexpected and require further inspection [16].

In addition to the SNFG, also displayed for each table entry is a copyable WURCS link, which encodes the complete glycan format in a linear code. The decision to present this information as a copyable link, as opposed to as plaintext is due to the inherent difficulty and unlikeliness for a human to read and understand the WURCS code. It is much more likely that the WURCS code would be copied and searched for in a glycomics database, hence we provide that functionality in a streamlined way.

The final section of the validation report includes all of the validation metrics calculated by *Privateer* and, most importantly, the diagnostic provided by *Privateer* (Figure 4). A 'yes' diagnostic indicates the conformation is correct for the glycosylation type (e.g., $^4C_1$ for GlcNAc in an *N*-glycan, $^1C_4$ for mannose in a *C*-glycan), has the correct anomer, and has an acceptable fit to density. This diagnostic indicates that the sugar is valid, whereas a diagnostic of 'check' indicates that *Privateer* has detected a potential inconsistency affecting ring conformation, which requires manual inspection. Finally, a 'no' diagnostic indicates that the sugar needs a more detailed manual inspection to correct any conformational issues, anomeric issues, or fitting issues.

Searching for entries in the *Privateer* database

Another interesting application of the collection of data available in the *Privateer* database is to visualise aggregated carbohydrate data from the PDB. Using the search interface on the *Privateer* database homepage, carbohydrate-containing PDB entries can easily be found and filtered. *Privateer* database entries for specific glycosylation types, namely, *N*-glycosylation, *O*-glycosylation, *S*-glycosylation, or *C*-glycosylation can be filtered quickly and easily. Additional filtering by linkage type is also possible, allowing niche glycosylation targets to be obtained. For example, filtering for *C*-glycans with a 'BMA-1,1-TRP' (the correct pair would be 'MAN-1,1-TRP', as the linkage in the modification is an alpha linkage) returns nine instances of incorrect sugar conformations in *C*-mannosylation found within the *Privateer* database in a table containing the frequency of the target linkage as well as a link to the *Privateer* database report page for target entry (Figure 5). This table view is also keyword or range-filterable at every data column, which allows for trivial searches of potentially interesting models.

Detailed monosaccharide validation data

Sugar ID	Conformation	Q	Phi	Theta	RSCC	B Factor	Detected Type	Type	Diagnostic
NAG-A-603	$^4c_1$	0.57	25.25	7.38	0.93	12.25	beta-D-aldopyranose	n-glycan	yes
NAG-A-604	$^4c_1$	0.56	326.04	6.21	0.88	21.09	beta-D-aldopyranose	n-glycan	yes
NAG-A-605	$^4c_1$	0.56	332.64	5.06	0.88	22.70	beta-D-aldopyranose	n-glycan	yes
NAG-B-1	$^4c_1$	0.55	116.33	4.41	0.93	14.40	beta-D-aldopyranose	n-glycan	yes
NAG-B-2	$^4c_1$	0.55	33.27	13.26	0.92	15.31	beta-D-aldopyranose	n-glycan	yes
BMA-B-3	$^4c_1$	0.52	25.27	18.87	0.81	23.83	beta-D-aldopyranose	n-glycan	yes
MAN-B-4	$^4c_1$	0.59	95.31	0.03	0.82	23.43	alpha-D-aldopyranose	n-glycan	yes
MAN-B-5	$^1s_5$	0.80	266.38	100.86	0.41	52.31	alpha-D-aldopyranose	n-glycan	check

Figure 4: Table of validation data for each sugar residue within PDB code 3QVP available in the visual validation report. The table contains all validation metrics calculated by *Privateer* including the Cremer–Pople puckering parameters, correlation coefficient, and, importantly, *Privateer* diagnostic, which can be used to identify the validity of each sugar. Table taken from the Privateer database report for 3QVP.

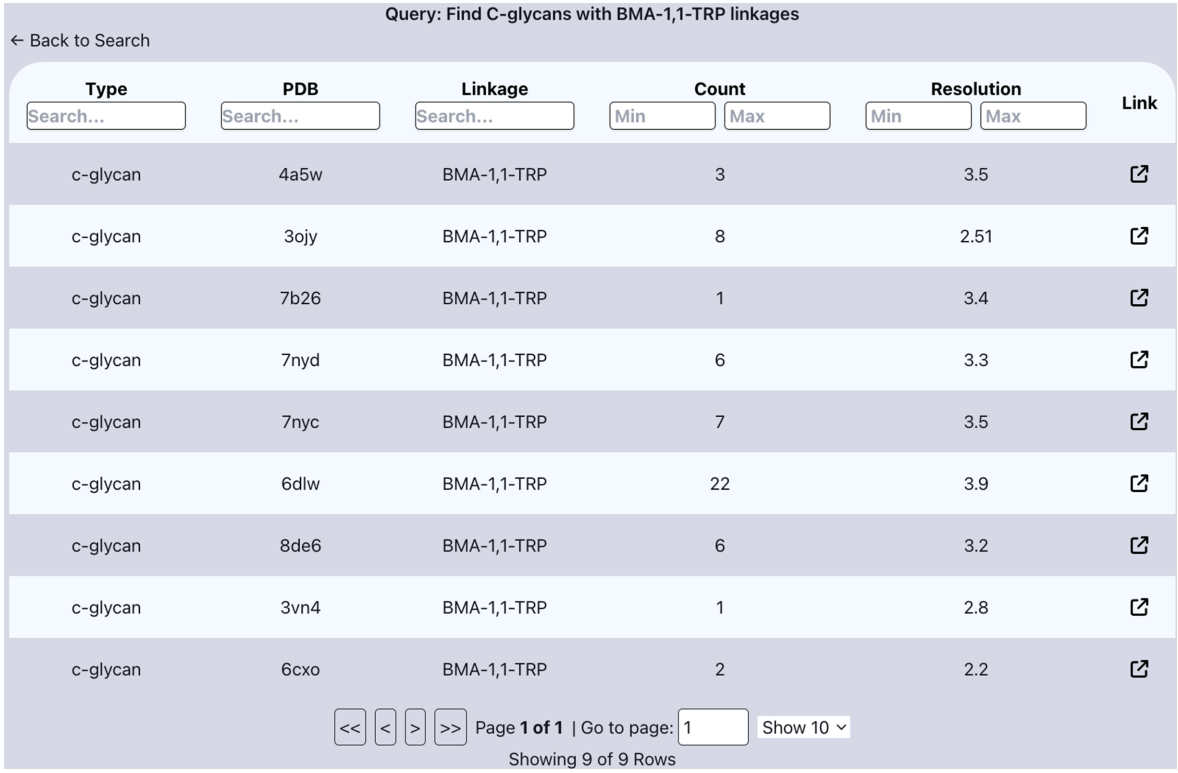


Figure 5: Table of available *Privateer* reports for the BMA-1,1-TRP linkage in *C*-glycans (*C*-mannosylation) sorted by the frequency (count) of the linkage in the deposited model. The table contains information of the carbohydrate type, PDB code, linkage, frequency, and resolution, as well as a link to the *Privateer* database report for each PDB entry.

Trends in the *Privateer* database

Using the *Privateer* database, global statistics throughout the PDB and PDB-REDO can be calculated with ease. Observing deposition trends in the PDB is often interesting as it can provide insight into the kinds of structures that are experimentally obtainable over time. With the *Privateer* database, trends in glycosylation deposition in the PDB over time can be measured, as shown in Figure 6. Importantly, as the *Privateer* database is completely recompiled every week, these trends remain consistent with the PDB. To allow for easy and up-to-date observation for anyone, compiled statistics are freely available alongside the *Privateer Web App*, https://privateer.york.ac.uk/statistics.

While simply looking at glycosylation over time using the *Privateer* database is possible, the validation reports calculated by *Privateer* contain a whole host of other interesting pieces of information. In an analogous way to looking at glycosylation over time, the type and validity of carbohydrates in the PDB can also be observed over time. The statistics page available alongside the *Privateer Web App* contains up-to-date plots of validation and conformational errors over time and resolution.

Conclusion

In conclusion, the new *Privateer* database encompasses the carbohydrate validation capabilities of *Privateer* in an easily accessible pre-prepared form. The database contains all validation metrics calculated by *Privateer* as well as highlighted SNFG diagrams in SVG format for easy third-party web use. Statistics are automatically computed weekly and are available alongside the database both on GitHub and the interactive web page.

Materials and Methods

The *Privateer* software package [14] was used to compute metrics and statistics for each entry in the PDB [24] or in PDB-REDO [21]. For each structure in the PDB, the carbohydrate-containing chains are first identified before being validated using the suite of validation tools available within *Privateer*. Using the Python bindings available within the latest versions of *Privateer*, a validation report can be generated for each carbohydrate in the molecular model. This report is put out in JSON format for easy consumption by web-based database frontends. The initial report generation was completed in parallel over 64 CPU cores in around 5 h. After the initial surveys through PDB and PDB-REDO, this process only needs

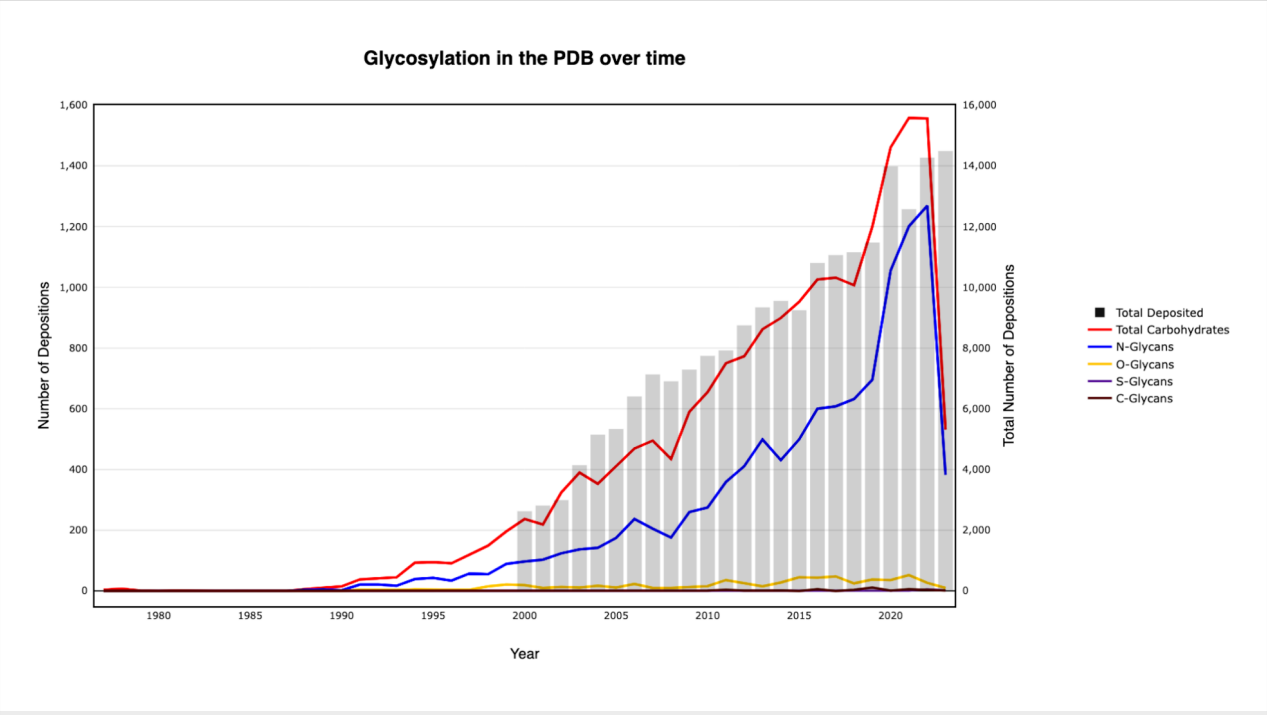


Figure 6: Plot showing trends in deposition in the PDB over time from 1975 to the present. Grey bars show the total deposited models into the PDB for all structural determination methods. Lines show glycosylation in the PDB over time, split into *N*-glycans, *O*-glycans, *S*-glycans, and *C*-glycans.

to be completed when new molecular models are deposited into the PDB, which occurs weekly. Although compiling validation reports for only new structures would be more efficient, this would fail to encompass changes in structures in historical entries, therefore the *Privateer* database is recompiled weekly.

The database, which receives any updates to the reports after recompilation is hosted on GitHub. The database is separated into PDB and PDB-REDO sections, which are in turn structured in the same format as the PDB archive, separated into folders by the middle two characters of the PDB four-letter code. For convenience, the presentation of the database is hosted alongside the *Privateer Web App* [26]; the database part can be accessed at https://privateer.york.ac.uk/database or by navigating to the database icon on the top right of the screen. The website is dynamic and compatible with desktop and laptop computers, plus tablets and smartphones.

Acknowledgements

We are grateful to the University of York IT Services and Darren Miller in particular for accommodating our needs and offering timely and excellent technical support. Lastly, we should like to acknowledge and highlight the contributions of Thomas Lütteke, Martin Frank, and the late Willy von der Lieth, pioneers of carbohydrate structure validation, whose research informed some of the methods showcased in the *Privateer* database.

Funding

Jordan Dialpuri is funded by the Biotechnology and Biological Sciences Research Council (BBSRC; grant No. BB/T0072221). Haroldas Bagdonas is funded by The Royal Society (grant No. RGF/R1/181006). Lucy Schofield is funded by STFC/CCP4 PhD studentship agreement 4462290 (York) / S2 2024 012 (STFC) awarded to Jon Agirre. Phuong Thao Pham is a self-funded PhD student. Lou Holland is funded by The Royal Society (URF\R\221006). Jon Agirre is a Royal Society University Research Fellow (awards UF160039 and URF\R\221006).

Author Contributions

Jordan S. Dialpuri: conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; software; validation; visualization; writing – original draft; writing – review & editing. Haroldas Bagdonas: software. Lucy C. Schofield: conceptualization; software; visualization. Phuong Thao Pham: data curation. Lou Holland: software; validation; visualization. Jon Agirre: conceptualization; data curation; funding acquisition; investigation; project administration; software; supervision; validation; writing – original draft; writing – review & editing.

ORCID® iDs

Jordan S. Dialpuri - https://orcid.org/0000-0002-6205-2661
Haroldas Bagdonas - https://orcid.org/0000-0001-5028-4847
Lucy C. Schofield - https://orcid.org/0009-0001-2069-878X

Phuong Thao Pham - https://orcid.org/0000-0002-6205-1298
Lou Holland - https://orcid.org/0000-0002-3867-1833
Jon Agirre - https://orcid.org/0000-0002-1086-0253

Data Availability Statement

All source code is publicly available on GitHub (https://github.com/glyco-jones/privateer and https://github.com/Dialpuri/PrivateerDatabase). The *Privateer* database is available at https://privateer.york.ac.uk/database and calculated statistics are available at https://privateer.york.ac.uk/statistics. Both pages will remain automatically updated with respect to the source code on GitHub.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2024.6.v1

References

1. Brockhausen, I.; Schutzbach, J.; Kuhns, W. *Acta Anat.* **1998,** *161,* 36–78. doi:10.1159/000046450
2. Calvelo, M.; Males, A.; Alteen, M. G.; Willems, L. I.; Vocadlo, D. J.; Davies, G. J.; Rovira, C. *ACS Catal.* **2023,** *13,* 13672–13678. doi:10.1021/acscatal.3c02378
3. Agirre, J.; Davies, G.; Wilson, K.; Cowtan, K. *Nat. Chem. Biol.* **2015,** *11,* 303. doi:10.1038/nchembio.1798
4. Agirre, J. *Acta Crystallogr., Sect. D: Struct. Biol.* **2017,** *73,* 171–186. doi:10.1107/s2059798316016910
5. Atanasova, M.; Bagdonas, H.; Agirre, J. *Curr. Opin. Struct. Biol.* **2020,** *62,* 70–78. doi:10.1016/j.sbi.2019.12.003
6. Lütteke, T.; Frank, M.; von der Lieth, C.-W. *Nucleic Acids Res.* **2005,** *33,* D242–D246. doi:10.1093/nar/gki013
7. Lütteke, T.; Frank, M.; von der Lieth, C.-W. *Carbohydr. Res.* **2004,** *339,* 1015–1020. doi:10.1016/j.carres.2003.09.038
8. Crispin, M.; Stuart, D. I.; Jones, E. Y. *Nat. Struct. Mol. Biol.* **2007,** *14,* 354. doi:10.1038/nsmb0507-354a
9. Frank, M.; Lütteke, T.; von der Lieth, C.-W. *Nucleic Acids Res.* **2007,** *35,* 287–290. doi:10.1093/nar/gkl907
10. von der Lieth, C.-W.; Freire, A. A.; Blank, D.; Campbell, M. P.; Ceroni, A.; Damerell, D. R.; Dell, A.; Dwek, R. A.; Ernst, B.; Fogh, R.; Frank, M.; Geyer, H.; Geyer, R.; Harrison, M. J.; Henrick, K.; Herget, S.; Hull, W. E.; Ionides, J.; Joshi, H. J.; Kamerling, J. P.; Leeflang, B. R.; Lütteke, T.; Lundborg, M.; Maass, K.; Merry, A.; Ranzinger, R.; Rosen, J.; Royle, L.; Rudd, P. M.; Schloissnig, S.; Stenutz, R.; Vranken, W. F.; Widmalm, G.; Haslam, S. M. *Glycobiology* **2011,** *21,* 493–502. doi:10.1093/glycob/cwq188
11. Lütteke, T.; Bohne-Lang, A.; Loss, A.; Goetz, T.; Frank, M.; von der Lieth, C.-W. *Glycobiology* **2006,** *16,* 71R–81R. doi:10.1093/glycob/cwj049
12. Toukach, P. V.; Egorova, K. S. *Nucleic Acids Res.* **2016,** *44,* D1229–D1236. doi:10.1093/nar/gkv840
13. Böhm, M.; Bohne-Lang, A.; Frank, M.; Loss, A.; Rojas-Macias, M. A.; Lütteke, T. *Nucleic Acids Res.* **2019,** *47,* D1195–D1201. doi:10.1093/nar/gky994
14. Agirre, J.; Iglesias-Fernández, J.; Rovira, C.; Davies, G. J.; Wilson, K. S.; Cowtan, K. D. *Nat. Struct. Mol. Biol.* **2015,** *22,* 833–834. doi:10.1038/nsmb.3115
15. Bagdonas, H.; Ungar, D.; Agirre, J. *Beilstein J. Org. Chem.* **2020,** *16,* 2523–2533. doi:10.3762/bjoc.16.204
16. Dialpuri, J. S.; Bagdonas, H.; Atanasova, M.; Schofield, L. C.; Hekkelman, M. L.; Joosten, R. P.; Agirre, J. *Acta Crystallogr., Sect. D: Struct. Biol.* **2023,** *79,* 462–472. doi:10.1107/s2059798323003510
17. Emsley, P.; Crispin, M. *Acta Crystallogr., Sect. D: Struct. Biol.* **2018,** *74,* 256–263. doi:10.1107/s2059798318005119
18. Atanasova, M.; Nicholls, R. A.; Joosten, R. P.; Agirre, J. *Acta Crystallogr., Sect. D: Struct. Biol.* **2022,** *78,* 455–465. doi:10.1107/s2059798322001103
19. Alocci, D.; Mariethoz, J.; Gastaldello, A.; Gasteiger, E.; Karlsson, N. G.; Kolarich, D.; Packer, N. H.; Lisacek, F. *J. Proteome Res.* **2019,** *18,* 664–677. doi:10.1021/acs.jproteome.8b00766
20. Fujita, A.; Aoki, N. P.; Shinmachi, D.; Matsubara, M.; Tsuchiya, S.; Shiota, M.; Ono, T.; Yamada, I.; Aoki-Kinoshita, K. F. *Nucleic Acids Res.* **2021,** *49,* D1529–D1533. doi:10.1093/nar/gkaa947
21. Joosten, R. P.; Long, F.; Murshudov, G. N.; Perrakis, A. *IUCrJ* **2014,** *1,* 213–220. doi:10.1107/s2052252514009324
22. van Beusekom, B.; Lütteke, T.; Joosten, R. P. *Acta Crystallogr., Sect. F: Struct. Biol. Commun.* **2018,** *74,* 463–472. doi:10.1107/s2053230x18004016
23. van Beusekom, B.; Wezel, N.; Hekkelman, M. L.; Perrakis, A.; Emsley, P.; Joosten, R. P. *Acta Crystallogr., Sect. D: Struct. Biol.* **2019,** *75,* 416–425. doi:10.1107/s2059798319003875
24. Berman, H.; Henrick, K.; Nakamura, H.; Markley, J. L. *Nucleic Acids Res.* **2007,** *35,* D301–D303. doi:10.1093/nar/gkl971
25. Matsubara, M.; Aoki-Kinoshita, K. F.; Aoki, N. P.; Yamada, I.; Narimatsu, H. *J. Chem. Inf. Model.* **2017,** *57,* 632–637. doi:10.1021/acs.jcim.6b00650
26. Dialpuri, J. S.; Bagdonas, H.; Schofield, L. C.; Pham, P. T.; Holland, L.; Bond, P. S.; Sánchez Rodríguez, F.; McNicholas, S. J.; Agirre, J. *Acta Crystallogr., Sect. F: Struct. Biol. Commun.* **2024,** *80,* 30–35. doi:10.1107/s2053230x24000359
27. Cremer, D.; Pople, J. A. *J. Am. Chem. Soc.* **1975,** *97,* 1354–1358. doi:10.1021/ja00839a011
28. Kommoju, P.-R.; Chen, Z.-w.; Bruckner, R. C.; Mathews, F. S.; Jorns, M. S. *Biochemistry* **2011,** *50,* 5521–5534. doi:10.1021/bi200388g
29. Neelamegham, S.; Aoki-Kinoshita, K.; Bolton, E.; Frank, M.; Lisacek, F.; Lütteke, T.; O'Boyle, N.; Packer, N. H.; Stanley, P.; Toukach, P.; Varki, A.; Woods, R. J.; The SNFG Discussion Group. *Glycobiology* **2019,** *29,* 620–624. doi:10.1093/glycob/cwz045
30. Agirre, J.; Davies, G. J.; Wilson, K. S.; Cowtan, K. D. *Curr. Opin. Struct. Biol.* **2017,** *44,* 39–47. doi:10.1016/j.sbi.2016.11.011

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Photoswitchable glycoligands targeting *Pseudomonas aeruginosa* LecA

Yu Fan[1], Ahmed El Rhaz[2], Stéphane Maisonneuve[1], Emilie Gillon[3], Maha Fatthalla[2], Franck Le Bideau[2], Guillaume Laurent[1], Samir Messaoudi[4], Anne Imberty[*3] and Juan Xie[*1]

Full Research Paper

Open Access

Address:
[1]Université Paris-Saclay, ENS Paris-Saclay, Institut d'Alembert, CNRS, Photophysique et Photochimie Supramoléculaires et Macromoléculaires, 91190, Gif-sur-Yvette, France, [2]Université Paris-Saclay, CNRS, BioCIS, 92290, Orsay, France, [3]Université Grenoble Alpes, CNRS, CERMAV, 38000 Grenoble, France and [4]Laboratoire de Synthèse Organique, Ecole Polytechnique, CNRS, ENSTA, Institut Polytechnique de Paris, 91128 Palaiseau, France

Email:
Anne Imberty* - anne.imberty@cermav.cnrs.fr; Juan Xie* - joanne.xie@ens-paris-saclay.fr

* Corresponding author

Keywords:
carbohydrates; glycosyl azobenzenes; lectin A; photoswitchable ligands

Beilstein J. Org. Chem. **2024,** *20,* 1486–1496.
https://doi.org/10.3762/bjoc.20.132

Received: 19 March 2024
Accepted: 21 June 2024
Published: 03 July 2024

This article is part of the thematic issue "Chemical glycobiology".

Guest Editor: E. Fadda

Abstract

Biofilm formation is one of main causes of bacterial antimicrobial resistance infections. It is known that the soluble lectins LecA and LecB, produced by *Pseudomonas aeruginosa*, play a key role in biofilm formation and lung infection. Bacterial lectins are therefore attractive targets for the development of new antibiotic-sparing anti-infective drugs. Building synthetic glycoconjugates for the inhibition and modulation of bacterial lectins have shown promising results. Light-sensitive lectin ligands could allow the modulation of lectins activity with precise spatiotemporal control. Despite the potential of photoswitchable tools, few photochromic lectin ligands have been developed. We have designed and synthesized several *O*- and *S*-galactosyl azobenzenes as photoswitchable ligands of LecA and evaluated their binding affinity with isothermal titration calorimetry. We show that the synthesized monovalent glycoligands possess excellent photophysical properties and strong affinity for targeted LecA with K_d values in the micromolar range. Analysis of the thermodynamic contribution indicates that the *Z*-azobenzene isomers have a systematically stronger favorable enthalpy contribution than the corresponding *E*-isomers, but due to stronger unfavorable entropy, they are in general of lower affinity. The validation of this proof-of-concept and the dissection of thermodynamics of binding will help for the further development of lectin ligands that can be controlled by light.

Introduction

Bacterial infection is a growing health problem due to antimicrobial resistance (AMR) among others. AMR causes approximately 33,000 deaths per annum in Europe only [1], and costs between €1.5 and €9 billion in healthcare and associated activities. Many bacterial infections occur by adhesion to host tissues through receptor–ligand interaction between bacterial carbohydrate-binding proteins (lectins) and oligosaccharides at the host cell surface. *Pseudomonas aeruginosa* (PA), a Gram-negative, opportunistic and ubiquitous environmental bacterium, is known as the leading cause of morbidity and mortality in cystic fibrosis and immunocompromised patients and as one of the leading causes of nosocomial infections [1]. Due to the existence of numerous molecular mechanisms conferring resistance to multiple classes of antibiotics, therapeutic options are increasingly limited for treatment of infections. PA has been classified as a priority 1 pathogen by the WHO [2,3]. Various approaches to treating PA, in addition to traditional antibiotics, have been developed including inhibition of quorum sensing, biofilm formation, iron chelation, and interfering with biosynthetic pathways of the bacterium [2,3]. The soluble lectins LecA and LecB produced by PA play a key role in the infection [4]. PA LecA is demonstrated to be crucial for biofilm formation and internalization, while the extracellular LecB plays a key role in bacterial adhesion to the host and biofilm formation [5-8]. Building synthetic glycoconjugates for the inhibition and modulation of bacterial lectins responsible for biofilm formation have shown promising results [9,10]. Unlike antibiotics, lectin inhibitors could prevent pathogenicity by interfering with virulence factors instead of killing the bacteria. Bacterial lectins are therefore attractive targets for the development of new antibiotic-sparing anti-infective drugs. For example, some *Escherichia coli* fimbrial lectin FimH inhibitors are currently in clinical development to treat and prevent urinary tract infections [9,10]. A large number of glycomimetic inhibitors of PA LecA and LecB have also been reported, with antibiofilm formation activity for some of them [5-8].

Photochromic molecules, which may be reversibly converted between different isomers upon illumination, offer numerous opportunities for reversibly photomodulating chemical, biological or pharmacological activities or properties [11,12]. Light is generally noninvasive and orthogonal toward most elements of living systems. It can be easily and precisely controlled in time, location, wavelength, and intensity, thus enabling the precise activation and deactivation of biological function. It also offers the potential to change the properties of defined molecules in biological systems with minimal disturbance to the rest of the system. Photoswitchable ligands, i.e., the incorporation of light-responsive moieties into a drug-like molecular structure, allow reversible light modulation of their activity since each isomer shows distinct structural and electronic properties [13]. Photoisomerization-induced conformational and polarity changes may allow to increase or decrease the interaction with the target protein or receptors, then modulate the drug potency on/off or from low to high. This strategy can be used for specific targeting or local drug activation to reduce its toxicity [14]. There is an increasing use of the photoisomerization to control the conformation as well as the activities of various biomolecules with the development of photopharmacology [11-18]. The group of Lindhorst has reported a series of mannosyl azobenzenes targeting *E. coli* lectin FimH, demonstrating the possibility to control the type 1 fimbriae-mediated bacterial adhesion to a self-assembled monolayer of mannosyl azobenzene on a gold surface [19,20] or to mannosyl azobenzene-modified human cells [21] through photoswitching the orientation of the attached mannoside [22]. Photoswitchable glycooligomers [23] or glycodendrimers [24] have been investigated for the inhibition of PA lectin PA-IL or LecA and LecB. A variation of the IC_{50} value by a factor up to 1.6 has been observed for the divalent ligand [23]; while almost no difference of inhibition was observed for LecA and LecB upon irradiation, probably due to the low photoisomerization of glycodendrimers [24]. Very recently, the group of Wittmann reported an arylazopyrazole-linked divalent *N*-acetylglucosamine targeting lectin wheat germ agglutinin [25]. The binding affinity K_d evaluated by isothermal titration calorimetry (ITC) showed a variation by a factor of 12.5 upon photoisomerization. However, a direct photomodulation of a monovalent lectin ligand has not been achieved up to date. Based on our experiences in photoswitchable glycosides and bacterial lectins [4,6-8,26-31], we have designed, synthesized, and characterized the first generation of *O*- and *S*-galactosyl azobenzenes as photoswitchable monovalent ligands targeting PA LecA. Their binding affinity with LecA evaluated by ITC showed K_d values in the micromolar range with significant thermodynamic differences between *E*- and *Z*-azobenzene isomers, demonstrating the proof-of-concept of photomodulation of the ligand–lectin interactions.

Results and Discussion

Design of LecA photoswitchable ligands

The cytotoxic LecA which has a tetrameric structure, displays a high affinity for D-galactose (D-Gal, with K_d = 34 μM) and galactosides. The 3- and 4-hydroxy function on the D-Gal unit are involved in the coordination of Ca^{2+} in the binding site [5-8,32]. A large range of galactosyl conjugates have been synthetized, with K_d values from micromolar (for monovalent galactosides) to nanomolar range (for di- and multivalent derivatives) [5-8]. For the monovalent system, it has been shown that aromatic aglycons favored "T-shaped" CH···π interactions with the protons of the His50 imidazole in the carbohydrate-binding

pocket, with the β-linked aromatic aglycons having five-fold higher affinity compared to aliphatic analogues [33,34]. Beside β-*O*-aryl galactosides, enzymatically more stable β-*S*-aryl galactosides have also been successfully developed as monovalent LecA ligands (Figure 1A) [30,35]. Since different sizes and substituents are tolerated on the aryl aglycon, we decided to replace the aryl aglycon by photoswitchable azobenzene in both *O*- and *S*-galactosides (Figure 1B) to investigate their binding affinity and the influence of the photoisomerization on the lectin interaction. The ammonium group is introduced on the azobenzene to increase the water solubility. The influence of *ortho*, *meta*, and *para*-substitution patterns of the azobenzene on the lectin binding has also been studied.

Synthesis

The β-*O*-galactosyl *p*,*p'*-bis-substituted azobenzene derivative **1** was prepared from galactose and commercially available *p*,*p'*-dihydroxyazobenzene (**6**), by using our recently developed DMC (2-chloro-1,3-dimethylimidazolinium chloride)-mediated one-pot glycosylation method in water [28], followed by *O*-alkylation of the remaining hydroxy group with $BrCH_2CH_2NHBoc$ and acidic deprotection (Scheme 1). Three equivalents of dihydroxyazobenzene **6** were used for the selective monoglycosylation step, with the excess of azobenene being recovered after column chromatography. Under these conditions, no bisglycosylated azobenzene was observed [28]. The observed 1,2-*trans* glycosylation could be explained either by the formation of the 1,2-anhydro sugar through intramolecular attack of the 2-hydroxy group of the DMC-activated β-intermediate, followed by dihydroxyazobenzene attacking the anomeric center in an S_N2 manner, or by direct nucleophilic S_N2 attack on the DMC-activated α-intermediate, to produce the corresponding β-*O*-galactoside [36]. The same strategy was applied for the *m*,*m'*-substituted derivative **2**, starting from the glycosylation of *m*,*m'*-dihydroxyazobenzene (**9**) [37], followed by *O*-alkylation and Boc deprotection to afford the galacoside **2** in 19% total yield. Unfortunately, all our attempts to synthesize the *o*,*o'*-bis-substituted derivative failed. For the β-*S*-galactosyl azobenzene derivatives which are accessible by our previously reported Pd-catalyzed cross-coupling methodology between glycosyl thiols and iodoaryl partners [30,38], the required *p*-, *m*- or *o*-iodo-*p'*-hydroxyazobenzenes **12**, **17**, and **21** were prepared by the diazonium coupling method according to a reported procedure [39,40]. Then the coupling with tetra-*O*-acetylated β-galactosylthiol **13** catalyzed by Xantphos Pd-G_3 [38] as precatalyst followed by post-functionalization furnished the desired β-*S*-galactosyl azobenzenes **3**, **4**, and **5** in respectively 71%, 41%, and 37% total yields (Scheme 1).

Photophysical characterization

The photoswitching properties of galactosyl azobenzenes **1**–**5** were realized in water or in Tris buffer containing 5 to 10% DMSO, in accordance with the biophysical evaluation conditions by using ITC. All these compounds undergo reversible photoisomerization under UV–vis irradiation in aqueous solution. The *O*-galactosyl azobenzene **1** shows reversible photoisomerization under UV (370 nm) and visible (485 nm) irradiations in water, with a high fatigue resistance as no degradation

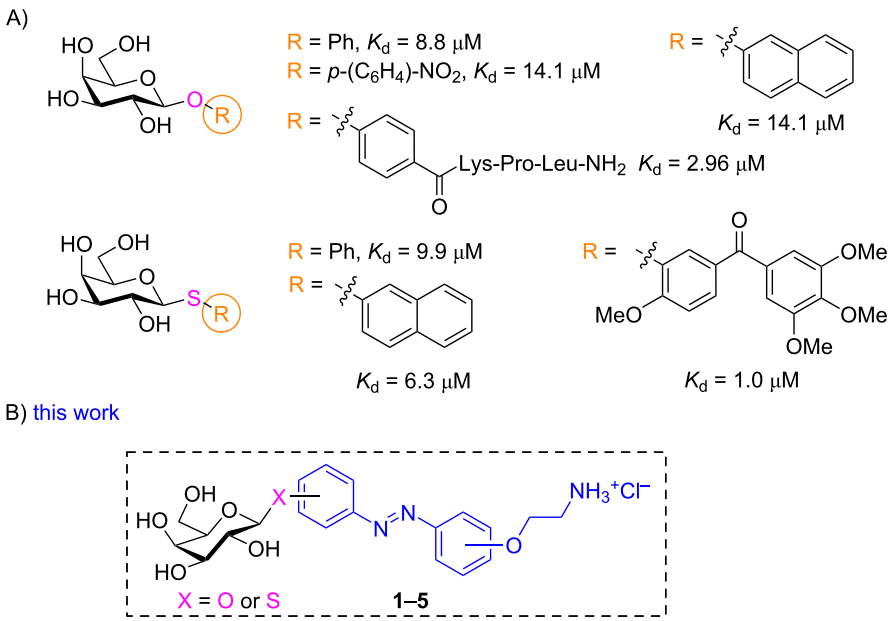
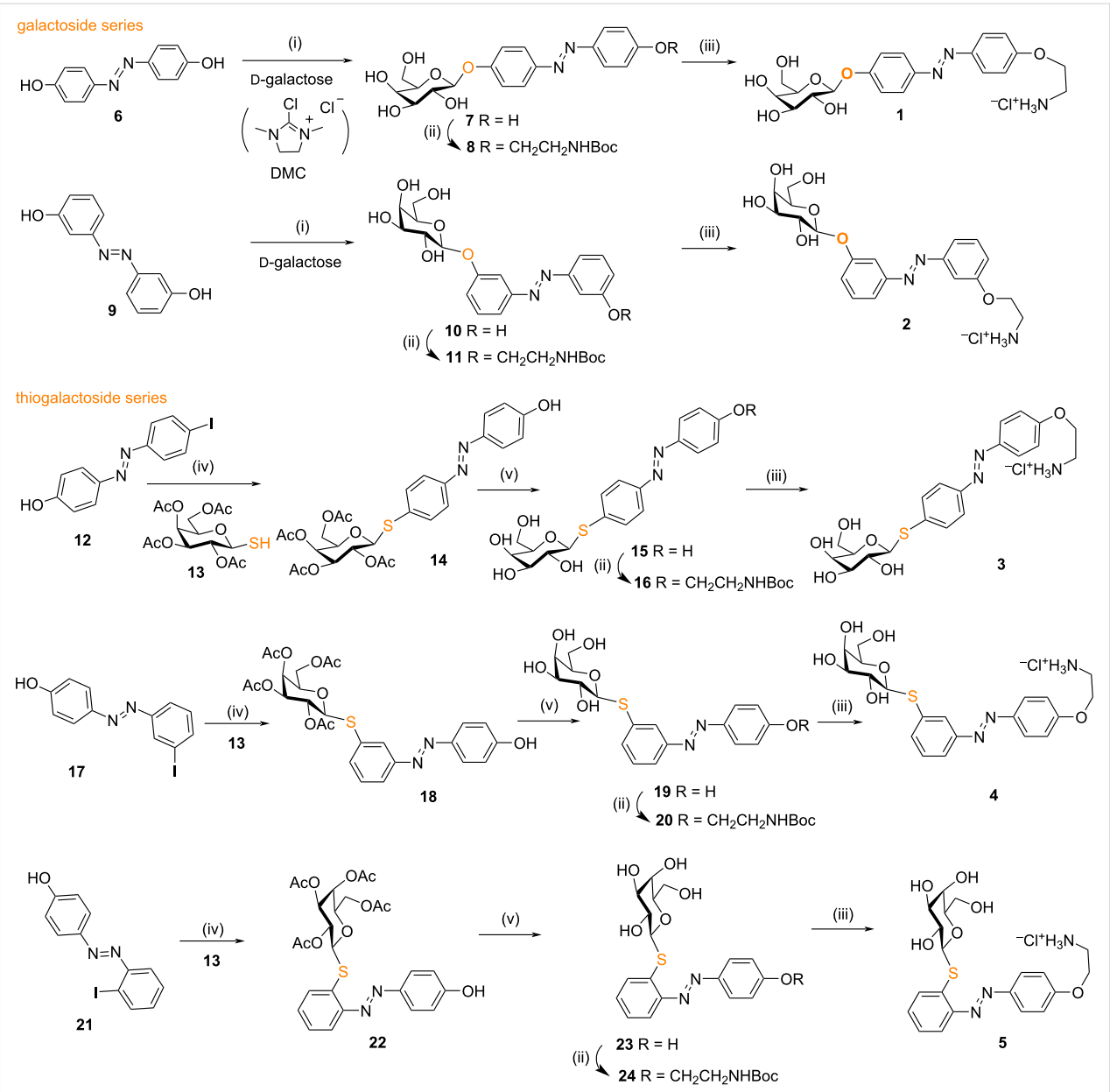


Figure 1: (A) Selected monovalent inhibitors for PA LecA and (B) designed general structure of photoswitchable ligands **1**–**5** targeting LecA.



Scheme 1: Synthesis of photoswitchable LecA inhibitors. Reagents and conditions: (i) DMC, Et_3N, H_2O, –10 °C to rt, 8 h, 50% for **7**, 40% for **10**; (ii) $BrCH_2CH_2NHBoc$, K_2CO_3, DMF, 60 °C, 15 h, 91% for **8**, 80% for **11**, 88% for **16**, 50% for **20**, 88% for **24**; (iii) AcCl, MeOH, 0 °C to rt, 15 h, 58% for **1**, 46% for **2**, 90% for **3**, 85% for **4**, 77% for **5**; (iv) Xantphos Pd-G_3 (5 mol %), Et_3N, THF, 6–8 h, rt, 90% for **14**, 98% for **18**, 54% for **22**; (v) MeONa/MeOH, 30 min–2 h, rt, compounds **15**, **19** and **23** were used without further purification.

has been observed after more than 10 UV–vis irradiation cycles (Figure 2). According to the absorption spectra (Figure 2, black line), the *E*-isomer shows a relatively strong π→π* transition (λ_{max} = 353 nm) and a weaker forbidden n→π* transition (λ_{max} ≈ 440 nm). After irradiation at 370 nm to induce the *E*-to-*Z* photoisomerization, the band at 353 nm decreases concomitantly to the appearance of two new bands at 312 and 438 nm (Figure 2, blue line). Two isosbestic points can also be observed at 310 and 429 nm. The back *Z*→*E* photoisomerization can be achieved by illumination at 485 nm (Figure 2, red line). ^{1}H NMR spectroscopy has been used to determine the *Z*/*E* ratios during irradiation, showing an excellent photoconversion yield of *Z*/*E* = 99:1 at PSS_{370}, and *E*/*Z* = 87:13 at PSS_{485} in D_2O/5% DMSO (Figure 3). As the *Z*-isomer is metastable, its half-life has been determined to be 44.4 h in water at room temperature (Figure S9 in Supporting Information File 1). All the photophysical properties of compounds **1**–**5** are summarized in Table 1 (spectra are shown in Figure S1–S24 in Supporting Information File 1). Concerning the *meta*-substituted azobenzene **2**, a 30 nm blue shift is observed for the π→π* transition

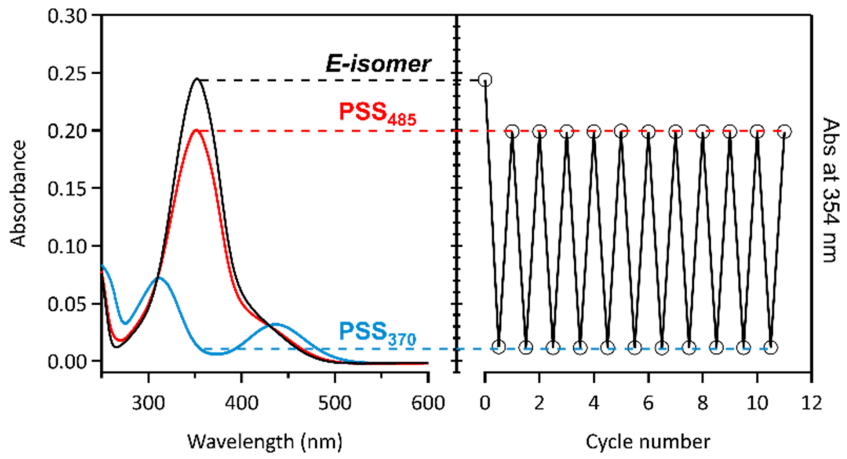


Figure 2: (Left) Absorption spectra and (right) fatigue resistance of **1** under alternated 370/485 nm irradiations in Tris buffer/DMSO 95:5 at rt: *E*-**1** (black line), PSS_{370} (blue line), PSS_{485} (red line). Irradiation conditions at 370 nm: 12.8 mW·cm^{-2}, 20 s; at 485 nm: 1.5 mW·cm^{-2}, 480 s.

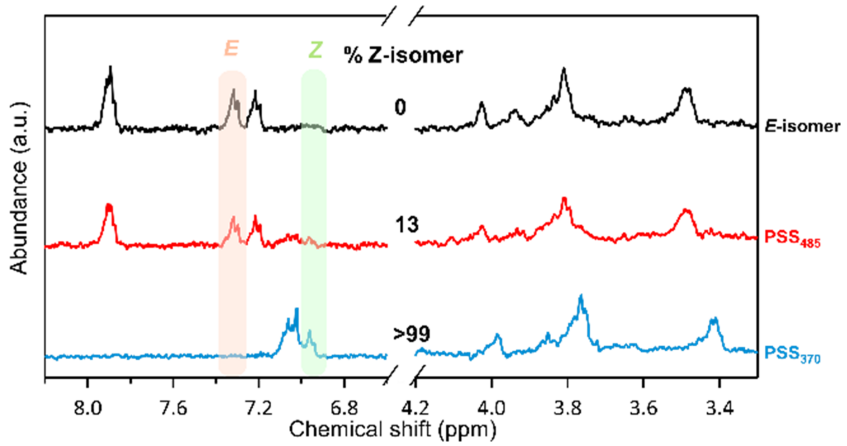


Figure 3: ^{1}H NMR (400 MHz) spectra of *E*-**1** (black line), PSS_{370} (red line), PSS_{485} (blue line) in D_2O/DMSO-d_6 95:5.

Table 1: Steady-state absorption, photostationary state composition, and half-life of *Z*-isomers of **1**–**5**.

Entry	Compound	Solvent	ε [M^{-1}cm^{-1}]	λ_{max} [nm]	*Z*/*E* PSS_{370}	*E*/*Z* PSS_{485}	$t_{1/2}$
1	**1**	H_2O	25632	353	99/1	87/13	44.4 h
2		Tris[a]/DMSO 5%	24400	354	99/1	87/13	n.d.[b]
3	**2**	Tris/DMSO 5%	14155	321	87/13	71/29	29.1 d[c]
4		Tris/DMSO 10%	15288	321	87/13	74/26[d]	n.d.
5	**3**	H_2O	18111	362	99/1	73/27	30.4 h
6		Tris/DMSO 10%	16991	364	99/1	71/29	25.9 h
7	**4**	Tris/DMSO 10%	22358	348	99/1	72/28	9.0 d
8	**5**	Tris/DMSO 10%	17336	348	92/8	60/40	73.3 h

[a]Tris buffer: Tris 20 mM (pH 7.5), NaCl 100 mM, $CaCl_2$ 100 μM; [b]not determined; [c]days; [d]PSS_{438} for **2**.

(λ_{max} = 321 nm) as well as a lower absorption coefficient compared to the *para*-derivative **1** (Table 1, entries 3 and 4 vs entries 1 and 2), probably due to less-conjugated azobenzene. Compared to compound **1**, a better $Z \rightarrow E$ photoconversion was achieved with irradiation at 438 nm instead of 485 nm. Moreover, the thermostability is increased ($t_{1/2}$ = 29 days). The *S*-galactosyl azobenzenes **3**–**5** also displayed excellent photoswitching properties, with a red shift for the $\pi \rightarrow \pi^*$ transition (λ_{max} = 348–364 nm) compared to the *O*-galactosyl derivatives (Table 1, entries 5–8). However, the absorption coefficient and the thermostability of the *Z*-isomers are increased for the *meta*-derivative **4**, compared to the *ortho*- (**5**) and *para*-substituted **3**.

Biophysical evaluation by ITC

The interaction of compounds **1**–**5** with LecA was characterized by ITC analysis for both the *E*- and *Z*-isomers. As the initial isomer state of the galactosyl azobenzenes is the *E*-form, ITC measurements made on *E*-isomers correspond to 100% purity of them. After 370 nm irradiation to induce the photoizomerisation process, a photostationary state is reached between *E*- and *Z*-isomers. For ITC measurements made on *Z*-isomers, the percentage of isomers is shown in the column *Z*/*E* (PSS_{370}) of Table 1. Depending on the corresponding galactosyl azobenzenes, the *Z*-isomer is pure from 87 to 99%. Spectroscopy measurements were performed on ligand solution just before each experiment to check the efficiency of the isomerization, with results as indicated in Table 2. In all experiments, strong exothermic peaks were observed for the first injection, followed by titration corresponding to stoichiometry of 1, in agreement with known structure (Figure 4). Control experiment with injection of compounds in buffer only did not show significant heat of dilution.

Affinity values, as well as thermodynamics contribution could be extracted through fitting procedure with a one site model and the data are reported in Table 2. All compounds have a strong affinity for LecA with K_d values ranging from 1.9 µM to 13.6 µM. These values are in the range of those observed previously for aromatic galactoside derivatives [33,34], confirming the favorable interaction of the aryl group with the protein surface. For all compounds, no significant differences of affinities are observed between the *E*- and *Z*-isomer, with the exception of compounds **1** and **3** with *para*-orientation between the two aryl groups. The affinity of the *E*-isomer is twice better than for its *Z*-counterpart for the *S*-linked compound **3** and three times better for the *O*-glycoside **1**. Even though the other compounds do not exhibit significant variations of affinities between *E*- and *Z*-isomers, a closer look at the thermodynamic values indicates that the mechanisms of binding display significant variations (Table 2). All of the *Z*-isomers display stronger favorable enthalpy of binding, i.e., a more negative ΔH contribution (ΔH varying from −41.1 to −49.4 kJ/mol) than their *E*-counterpart (ΔH from −38.0 to −43.5 kJ/mol). This is fully counterbalanced by a stronger unfavorable entropy barrier, i.e. a more positive entropy contribution ($-T\Delta S$), varying from 10.3 to 18.5 kJ/mol for the *Z*-isomers, and from 8.8 to 13.1 kJ/mol for the *E*-isomers. As displayed in Figure 4, this enthalpy–entropy compensation results in a limited variation of ΔG and therefore in the observed rather similar K_d values.

Table 2: Microcalorimetry data and thermodynamics contribution for binding to LecA. The experiments were realized in duplicate at 298 K unless otherwise stated.

Ligand	K_d [µM]	*n*	$-\Delta G$ [kJ/mol]	$-\Delta H$ [kJ/mol]	$T\Delta S$ [kJ/mol]
***E*-1**	4.8 ± 0.3	0.98 ± 0.01	30.3	40.8 ± 0.5	−10.5
***Z*-1**	13.6 ± 1.2	1.04 ± 0.04	27.8	41.3 ± 0.5	−13.5
***E*-2**	5.1 ± 0.7	1.01 ± 0.05	30.2	43.3 ± 1.0	−13.1
***Z*-2**[a]	5.1 ± 0.7	0.97 ± 0.04	30.2	45.8 ± 0.6	−15.6
***E*-3**	1.9 ± 0.1	1[b]	32.6	43.5 ± 0.4	−10.9
***Z*-3**	4.1 ± 0.02	1[b]	30.7	49.4 ± 0.4	−18.7
***E*-4**	7.7 ± 1.3	0.96 ± 0.07	29.2	38.0 ± 1.3	−8.8
***Z*-4**	5.1 ± 0.1	0.96 ± 0.02	30.2	47.1 ± 0.2	−16.9
***E*-5**	4.3 ± 0.2	1.02 ± 0.05	30.6	40.1 ± 0.6	−9.5
***Z*-5**[a]	4.1 ± 0	0.96 ± 0.03	30.8	41.1 ± 0.4	−10.3

[a]*Z*-isomer of compound **2** is mixed with 13% *E*-isomer and compound **5** is mixed with 8% *E*-isomer as established by PSS_{370}. This contamination is less than 2% for the other compounds. [b]Concentration of compound **3** could not be determined from weight products due to aggregation. Active concentration was determined fitting ITC data to stoichiometry of 1, value confirmed from other compounds. For all other compounds, the concentration was calculated from weighted compound and confirmed by spectroscopy (see Table S1 in Supporting Information File 1).

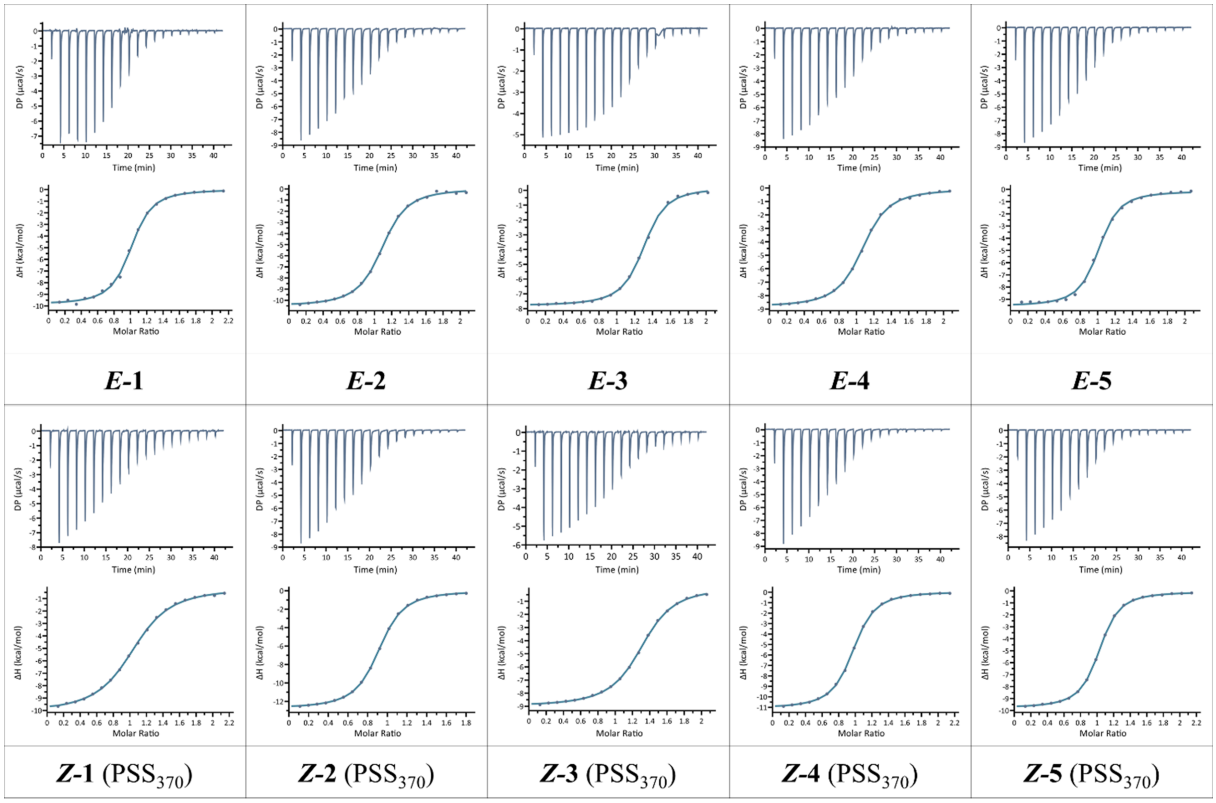


Figure 4: ITC titration of LecA with *E*- (up) and *Z*-isomers (bottom) of compounds **1**–**5** in Tris buffer containing 5 to 10% DMSO. The plot in the lower panel shows the total heat released as a function of total ligand concentration for the titration shown in the upper panel. The solid line represents the best least-square fit to experimental data using a one site model.

In order to rationalize this difference in binding mechanism, molecular models were obtained for selected low-energy conformations of *E*- and *Z*-isomers of a "model" scaffold of the *para*-azobenzene derivative in the binding site of LecA (Figure 5), by simple superpositioning of the known crystal structure. The extended *E*-isomer establishes contact through galactoside and the first aryl ring only, while the bent *Z*-isomer has proper conformation to wrap around the central His53 residue and to establish a more extended interaction with the protein surface. This would be in agreement with a stronger

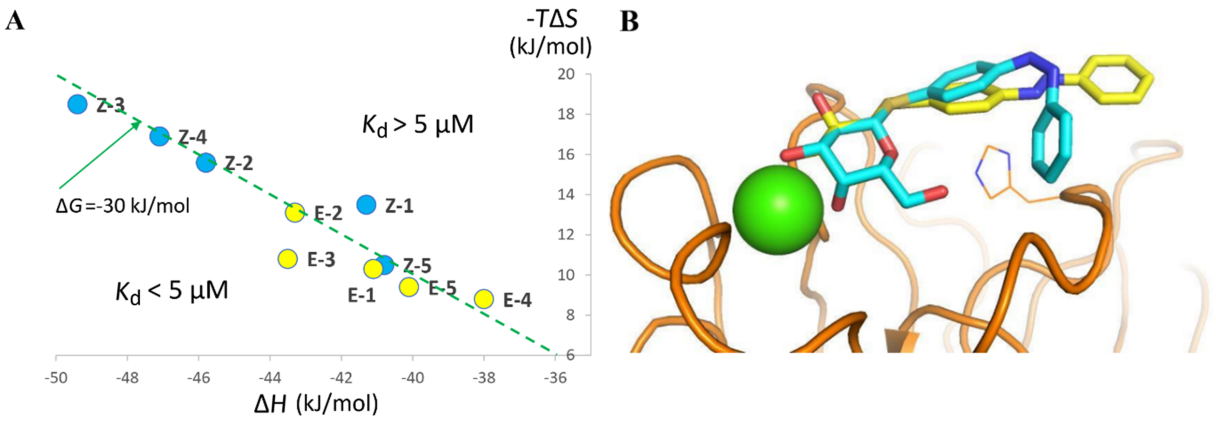


Figure 5: (A) Enthalpy–entropy compensation plot of compounds **1**–**5** from ITC analysis. The dotted green line represents a Δ*G* value of –30 kJ/mol, corresponding to a K_d of approx 5 μM in the experimental conditions. (B) Manual docking of scaffold for compound **3** with selected low energy conformations of the *E*-isomer (yellow sticks) and *Z*-isomer (cyan sticks) superimposed on conserved position of galactose in all LecA crystalline complexes. The protein is represented by orange ribbon, His53 by lines, and calcium by green sphere.

enthalpy of interaction, while the entropy barrier could arise from a limitation of flexibility and/or blocking of water molecules at the new interface.

Conclusion

We have designed and synthesized in three to five steps *O*- and *S*-galactosyl azobenzenes targeting the *Pseudomonas aeruginosa* lectin LecA. The five synthesized glycoconjugates can be reversibly photoconverted between *E*- and *Z*-isomers under UV and vis irradiation, with good to excellent photoconversion yields and high fatigue resistance in aqueous media. Furthermore, all the Z-isomers displayed good thermostability, with the half-live varied from 26 h to 29 days at room temperature depending on the type of glycosidic linkage and the substitution pattern on the azobenzene moiety. The bistability of the azobenzene derivatives is suitable for the investigation of azobenzene isomers on the binding affinity with LecA. All the galactosyl azobenzenes bound to LecA in the low micromolar range. Interestingly, the *para*-substituted *O*- (**1**) and *S*- (**3**) galactosides displayed 2 to 3-fold difference in affinity between *E*- and *Z*-isomers (3-fold difference for **1** and 2-fold for **3**), demonstrating the proof-of-concept of tuning the LecA binding by light. Few differences were observed for the *meta*- (**2** and **4**) and *ortho*-substituted azobenzenes (**5**). Thermodynamics contributions exhibit larger variations with stronger enthalpy of binding for the *Z*-isomer, probably in relation with a folded conformation generating additional contact with the surface. Due to enthalpy–entropy compensation, that is a general effect in protein–carbohydrate interactions [41], this does not reflect in differences in affinity. However, these observations, together with future modeling studies, will help in designing new compounds with more selective binding of one isomer only.

Experimental

General experimental details. Commercially available solvents and reagents were used without further purification. Reactions carried out under anhydrous conditions are performed under argon in glassware previously dried in an oven. DMF and THF were previously dried through alumina or molecular sieves-containing cartridges using a solvent purificator MBRAUN SPS-800. All the reactions with azobenzene-containing substrates were carried out in the dark. Reactions were monitored by TLC on Silica Gel 60F-254 plates with detection by UV (254 nm or 365 nm) or by spraying with 10% H_2SO_4 in EtOH and heating about 30 s at 400–600 °C. Column chromatography purification was performed on CombiFlash® Rf+ and RediSep® RF or RF Gold normal phase silica columns (with UV detection at 254 and 350 nm for all azobenzene derivatives), or by flash column chromatography employing silica gel (60 Å pore size, 40–63 μm). ^{1}H and ^{13}C NMR spectra were recorded on a JEOL ECS-400 spectrometer or on Bruker Avance 300 and 400 spectrometers. Structural assignments were made with additional information from gCOSY, HMBC, and gHMQC experiments. High-resolution mass spectra (HRMS) were performed on a Bruker maXis mass spectrometer by the SALSA platform from ICOA laboratory or on an Agilent 1260 Infinity system with a quadrupole time-of-light (Q-TOF) mass analyzer. Melting points were measured with a Köfler bench previously calibrated using the usual standard references or on a digital melting point capillary apparatus. Specific optical rotations were measured in solution using sodium light at 589 nm where no absorption occurred for all compounds. Absorption spectra were recorded on a Cary-5000 spectrophotometer from Agilent Technologies. Photochromic reactions were induced in situ by a continuous irradiation Hg/Xe lamp (Hamamatsu, LC6- or LC8-Lightningcure, 200 W) equipped with narrow band interference filters of appropriate wavelengths: Semrock BP-370/36 for λ_{irr} = 370 nm, Semrock FF01-438/24-25 for λ_{irr} = 438 nm, Semrock FF01-485/20-25 for λ_{irr} = 485 nm. The irradiation power was measured using a photodiode from Ophir (PD300-UV) and corrected after a measurement with an additional Schott long pass filter (LP-545) to measure NIR contribution (P_{LP}) that is let through the Semrock filter (P_{Total}), considering a 90% transmittance: $P_{\lambda irr} = P_{Total} - (10/9 \times P_{LP})$. The photoconversion reaction was followed by a combination of ^{1}H NMR and UV–vis absorption spectra, realized by successive irradiations at 370 nm (438 or 485 nm). The *E*/*Z* ratios were determined by integration of the azobenzene proton signals of each isomer. A quartz cell of 10 mm path length has been used for solution measurement.

The photoconversion yields were measured from a solution of the compounds in deuterated solvent and monitored by ^{1}H NMR and UV–vis absorption, after successive irradiations at 370 nm (438 nm or 485 nm) in the case of the PSS. The *E*/*Z* ratios were determined by integration of characteristic of each isomer.

Data processing of spectroscopic measurements was realized with the help of Microsoft® Excel® and Igor Pro from WaveMetrics, Inc (versions 7 to 9).

Isothermal titration calorimetry: LecA was expressed and purified as previously described [42]. All experiments were performed at 25 °C with an ITC200 isothermal titration calorimeter (Microcal-Malvern Panalytical, Grenoble, France). The lyophilized LecA protein was dissolved in a buffer composed of 20 mM Tris·HCl (pH 7.5), 100 mM NaCl and 100 μM $CaCl_2$ with 5% or 10% DMSO final. All compounds were first dissolved in DMSO then in same buffer for a final concentration of 5% or 10% DMSO. The 200 μL sample cell containing LecA

(concentrations ranging from 200 to 300 μM) was subjected to injections of ligand solution: 20 injections of 2 μL (2–3 mM, depending on the ligand) at intervals of 120 s while stirring at 850 rpm. Control experiments were performed by repeating the same protocol, but injecting the ligand into buffer solution. The supplied software Origin 7 or MicroCal PEAQ-ITC was used to fit the experimental data to a theoretical titration curve allowing the determination of affinity (i.e., dissociation constant, K_d), binding enthalpy (ΔH), and stoichiometry (n). Values for free energy change (ΔG) and entropy contributions ($T\Delta S$) were derived from the equation $\Delta G = \Delta H - T\Delta S = -RT \ln K_d$ (with T = 298.15 K and R = 8.314 J $mol^{-1}K^{-1}$).

General procedure I for the *O*-alkylation with $BrCH_2CH_2NHBoc$: A solution of glycosyl azobenzene (1.0 equiv) in anhydrous DMF (≈3.5 mL per mmol) was added K_2CO_3 (2.0–4 equiv) and $BocNHCH_2CH_2Br$ (1.5–4 equiv), then stirred overnight at 60 °C. After the reaction was completed (TLC monitoring), the mixture was evaporated to dryness under reduced pressure. The residue was dissolved in EtOAc, neutralized with HCl (1 M), and extracted with EtOAc (3 times). The organic phase was washed with brine, dried over anhydrous Na_2SO_4, evaporated under reduced pressure in vacuo, and purified by CombiFlash Rf+ (CH_2Cl_2/MeOH 15:1).

General procedure II for the Boc deprotection: To a solution of the Boc-protected compound in anhydrous MeOH (≈10 mL per mmol) was added dropwise AcCl (1.0–3.0 equiv) at 0 °C, slowly warmed to rt, and stirred overnight. After the reaction was completed (TLC monitoring), the mixture was evaporated to dryness under reduced pressure. The residue was dissolved in MeOH, acetone was added, and a precipitate was obtained, which was washed with CH_2Cl_2 and *n*-pentane successively to give a pure compound.

General procedure III for the syhthesis of *S*-galactosyl azobenzenes: A round-bottomed flask was charged with Xantphos Pd-G_3 (5 mol %), acetylated β-thiogalactoside **13** [38] (1.1 equiv), and iodinated azobenzene (1 equiv). After Ar flushing, dry THF (0.25 M) was added and the mixture stirred for 5 min before NEt_3 (1.1 equiv) was added. The reaction mixture was stirred at rt under Ar for 6–8 h, diluted with EtOAc, filtered over celite, and washed with EtOAc. The collected organic layers were concentrated under reduced pressure and purified by flash chromatography (cyclohexane/EtOAc 7:3) to give the thioglycoside.

General procedure IV for the Zemplén deacetylation: To a seal tube containing the galactose derivatives in dry MeOH (0.15 M), NaOMe (30 mol %, 0.5 M sol. in MeOH) was added. The mixture was stirred at room temperature until total deprotection. The solution was neutralized using Amberlite IR-120 (H), filtered, concentrated and the crude material used without further purification to give the desired product in quantitative yield.

Supporting Information

Supporting Information File 1

Compound characterization, detailed photochemical and photophysical procedures and copies of spectra.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-132-S1.pdf]

Acknowledgements

A. Imberty and E. Gillon acknowledge support from Glyco@Alps (ANR-15-IDEX-02) and Labex Arcane/CBH-EUR-GS (ANR-17-EURE-0003). The authors would like to thank also Cyril Colas from the "Fédération de Recherche" ICOA/CBM (FR2708)" for HRMS analysis.

Funding

Y. Fan gratefully acknowledges China Scholarship Council (CSC) for a doctoral scholarship. The authors thank the Centre National de la Recherche Scientifique (CNRS) for support of this research and MRES for a doctoral fellowship to A. El Rhaz.

Author Contributions

Yu Fan: data curation; formal analysis; investigation; visualization; writing – original draft. Ahmed El Rhaz: data curation; formal analysis; investigation. Stéphane Maisonneuve: data curation; formal analysis; investigation; software; supervision; validation. Emilie Gillon: investigation. Maha Fatthalla: investigation. Franck Le Bideau: formal analysis; validation. Guillaume Laurent: supervision; validation. Samir Messaoudi: conceptualization; funding acquisition; methodology; resources; validation; visualization; writing – review & editing. Anne Imberty: conceptualization; data curation; formal analysis; funding acquisition; resources; supervision; validation; writing – original draft; writing – review & editing. Juan Xie: conceptualization; funding acquisition; methodology; project administration; resources; supervision; validation; visualization; writing – original draft; writing – review & editing.

ORCID® iDs

Ahmed El Rhaz - https://orcid.org/0009-0003-1772-1699
Stéphane Maisonneuve - https://orcid.org/0000-0003-0459-3459
Franck Le Bideau - https://orcid.org/0000-0003-4365-4525
Samir Messaoudi - https://orcid.org/0000-0002-4994-9001

Anne Imberty - https://orcid.org/0000-0001-6825-9527
Juan Xie - https://orcid.org/0000-0001-7664-5532

Data Availability Statement

All data that supports the findings of this study is available in the published article and/or the supporting information to this article.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2024.15.v1

References

1. Cassini, A.; Högberg, L. D.; Plachouras, D.; Quattrocchi, A.; Hoxha, A.; Skov, S. G.; Colomb-Cotinat, M.; Kretzschmar, M. E.; Devleesschauwer, B.; Cecchini, M.; Ait, D.; Cravo, T.; Struelens, M. J.; Suetens, C.; Monnet, D. L.; Burden of AMR Collaborative Group. *Lancet Infect. Dis.* **2019,** *19,* 56–66. doi:10.1016/s1473-3099(18)30605-4
2. Pelegrin, A. C.; Palmieri, M.; Mirande, C.; Oliver, A.; Moons, P.; Goossens, H.; van Belkum, A. *FEMS Microbiol. Rev.* **2021,** *45,* fuab026. doi:10.1093/femsre/fuab026
3. Tacconelli, E.; Carrara, E.; Savoldi, A.; Harbarth, S.; Mendelson, M.; Monnet, D. L.; Pulcini, C.; Kahlmeter, G.; Kluytmans, J.; Carmeli, Y.; Ouellette, M.; Outterson, K.; Patel, J.; Cavaleri, M.; Cox, E. M.; Houchens, C. R.; Grayson, M. L.; Hansen, P.; Singh, N.; Theuretzbacher, U.; Magrini, N.; WHO Pathogens Priority List Working Group. *Lancet Infect. Dis.* **2018,** *18,* 318–327. doi:10.1016/s1473-3099(17)30753-3
4. Chemani, C.; Imberty, A.; de Bentzmann, S.; Pierre, M.; Wimmerova, M.; Guery, B. P.; Faure, K. *Infect. Immun.* **2009,** *77,* 2065–2075. doi:10.1128/iai.01204-08
5. Wojtczak, K.; Byrne, J. P. *ChemMedChem* **2022,** *17,* e202200081. doi:10.1002/cmdc.202200081
6. Zahorska, E.; Rosato, F.; Stober, K.; Kuhaudomlarp, S.; Meiers, J.; Hauck, D.; Reith, D.; Gillon, E.; Rox, K.; Imberty, A.; Römer, W.; Titz, A. *Angew. Chem., Int. Ed.* **2023,** *62,* e202215535. doi:10.1002/anie.202215535
7. Siebs, E.; Shanina, E.; Kuhaudomlarp, S.; da Silva Figueiredo Celestino Gomes, P.; Fortin, C.; Seeberger, P. H.; Rognan, D.; Rademacher, C.; Imberty, A.; Titz, A. *ChemBioChem* **2022,** *23,* e202100563. doi:10.1002/cbic.202100563
8. Mała, P.; Siebs, E.; Meiers, J.; Rox, K.; Varrot, A.; Imberty, A.; Titz, A. *J. Med. Chem.* **2022,** *65,* 14180–14200. doi:10.1021/acs.jmedchem.2c01373
9. Behren, S.; Westerlind, U. *Eur. J. Org. Chem.* **2023,** *26,* e202200795. doi:10.1002/ejoc.202200795
10. Leusmann, S.; Ménová, P.; Shanin, E.; Titz, A.; Rademacher, C. *Chem. Soc. Rev.* **2023,** *52,* 3663–3740. doi:10.1039/d2cs00954d
11. Lerch, M. M.; Hansen, M. J.; van Dam, G. M.; Szymanski, W.; Feringa, B. L. *Angew. Chem., Int. Ed.* **2016,** *55,* 10978–10999. doi:10.1002/anie.201601931
12. Yu, Z.; Hecht, S. *Chem. Commun.* **2016,** *52,* 6639–6653. doi:10.1039/c6cc01423b
13. Kobauri, P.; Dekker, F. J.; Szymanski, W.; Feringa, B. L. *Angew. Chem., Int. Ed.* **2023,** *62,* e202300681. doi:10.1002/anie.202300681
14. Volarić, J.; van der Heide, N. J.; Mutter, N. L.; Samplonius, D. F.; Helfrich, W.; Maglia, G.; Szymanski, W.; Feringa, B. L. *ACS Chem. Biol.* **2024,** *19,* 451–461. doi:10.1021/acschembio.3c00640
15. Velema, W. A.; Szymanski, W.; Feringa, B. L. *J. Am. Chem. Soc.* **2014,** *136,* 2178–2191. doi:10.1021/ja413063e
16. Broichhagen, J.; Frank, J. A.; Trauner, D. *Acc. Chem. Res.* **2015,** *48,* 1947–1960. doi:10.1021/acs.accounts.5b00129
17. Hüll, K.; Morstein, J.; Trauner, D. *Chem. Rev.* **2018,** *118,* 10710–10747. doi:10.1021/acs.chemrev.8b00037
18. Fuchter, M. J. *J. Med. Chem.* **2020,** *63,* 11436–11447. doi:10.1021/acs.jmedchem.0c00629
19. Weber, T.; Chandrasekaran, V.; Stamer, I.; Thygesen, M. B.; Terfort, A.; Lindhorst, T. K. *Angew. Chem., Int. Ed.* **2014,** *53,* 14583–14586. doi:10.1002/anie.201409808
20. Chandrasekaran, V.; Jacob, H.; Petersen, F.; Kathirvel, K.; Tuczek, F.; Lindhorst, T. K. *Chem. – Eur. J.* **2014,** *20,* 8744–8752. doi:10.1002/chem.201402075
21. Möckl, L.; Müller, A.; Bräuchle, C.; Lindhorst, T. K. *Chem. Commun.* **2016,** *52,* 1254–1257. doi:10.1039/c5cc08884d
22. Despras, G.; Möckl, L.; Heitmann, A.; Stamer, I.; Bräuchle, C.; Lindhorst, T. K. *ChemBioChem* **2019,** *20,* 2373–2382. doi:10.1002/cbic.201900269
23. Ponader, D.; Igde, S.; Wehle, M.; Märker, K.; Santer, M.; Bléger, D.; Hartmann, L. *Beilstein J. Org. Chem.* **2014,** *10,* 1603–1612. doi:10.3762/bjoc.10.166
24. Hu, Y.; Beshr, G.; Garvey, C. J.; Tabor, R. F.; Titz, A.; Wilkinson, B. L. *Colloids Surf., B* **2017,** *159,* 605–612. doi:10.1016/j.colsurfb.2017.08.016
25. Osswald, U.; Boneberg, J.; Wittmann, V. *Chem. – Eur. J.* **2022,** *28,* e202200267. doi:10.1002/chem.202200267
26. Lin, C.; Maisonneuve, S.; Métivier, R.; Xie, J. *Chem. – Eur. J.* **2017,** *23,* 14996–15001. doi:10.1002/chem.201703461
27. Lin, C.; Jiao, J.; Maisonneuve, S.; Mallétroit, J.; Xie, J. *Chem. Commun.* **2020,** *56,* 3261–3264. doi:10.1039/c9cc09853d
28. Wang, Z.; Maisonneuve, S.; Xie, J. *J. Org. Chem.* **2022,** *87,* 16165–16174. doi:10.1021/acs.joc.2c01511
29. Jiao, J.; Maisonneuve, S.; Xie, J. *J. Org. Chem.* **2022,** *87,* 8534–8543. doi:10.1021/acs.joc.2c00652
30. Bruneau, A.; Gillon, E.; Furiga, A.; Brachet, E.; Alami, M.; Roques, C.; Varrot, A.; Imberty, A.; Messaoudi, S. *Eur. J. Med. Chem.* **2023,** *247,* 115025. doi:10.1016/j.ejmech.2022.115025
31. Gajdos, L.; Blakeley, M. P.; Haertlein, M.; Forsyth, V. T.; Devos, J. M.; Imberty, A. *Nat. Commun.* **2022,** *13,* 194. doi:10.1038/s41467-021-27871-8
32. Cioci, G.; Mitchell, E. P.; Gautier, C.; Wimmerová, M.; Sudakevitz, D.; Pérez, S.; Gilboa-Garber, N.; Imberty, A. *FEBS Lett.* **2003,** *555,* 297–301. doi:10.1016/s0014-5793(03)01249-3
33. Kadam, R. U.; Garg, D.; Schwartz, J.; Visini, R.; Sattler, M.; Stocker, A.; Darbre, T.; Reymond, J.-L. *ACS Chem. Biol.* **2013,** *8,* 1925–1930. doi:10.1021/cb400303w
34. Sommer, R.; Wagner, S.; Rox, K.; Varrot, A.; Hauck, D.; Wamhoff, E.-C.; Schreiber, J.; Ryckmans, T.; Brunner, T.; Rademacher, C.; Hartmann, R. W.; Brönstrup, M.; Imberty, A.; Titz, A. *J. Am. Chem. Soc.* **2018,** *140,* 2537–2545. doi:10.1021/jacs.7b11133
35. Rodrigue, J.; Ganne, G.; Blanchard, B.; Saucier, C.; Giguère, D.; Shiao, T. C.; Varrot, A.; Imberty, A.; Roy, R. *Org. Biomol. Chem.* **2013,** *11,* 6906–6918. doi:10.1039/c3ob41422a
36. Fairbanks, A. J. *Carbohydr. Res.* **2021,** *499,* 108197. doi:10.1016/j.carres.2020.108197

37. Gund, S. H.; Shelkar, R. S.; Nagarkar, J. M. *RSC Adv.* **2014,** *4,* 42947–42951. doi:10.1039/c4ra06027j
38. Bruneau, A.; Roche, M.; Hamze, A.; Brion, J.-D.; Alami, M.; Messaoudi, S. *Chem. – Eur. J.* **2015,** *21,* 8375–8379. doi:10.1002/chem.201501050
39. Schultzke, S.; Walther, M.; Staubitz, A. *Molecules* **2021,** *26,* 3916. doi:10.3390/molecules26133916
40. Ngaini, Z.; Hissam, M. A.; Mortadza, N. A.; Abd Halim, A. N.; Daud, A. I. *Nat. Prod. Res.* **2023,** 1–11. doi:10.1080/14786419.2023.2262713
41. Dam, T. K.; Brewer, C. F. *Chem. Rev.* **2002,** *102,* 387–430. doi:10.1021/cr000401x
42. Kuhaudomlarp, S.; Gillon, E.; Varrot, A.; Imberty, A. LecA (PA-IL): A Galactose-Binding Lectin from Pseudomonas aeruginosa. In *Lectin Purification and Analysis;* Hirabayashi, J., Ed.; Methods in Molecular Biology, Vol. 2132; Humana Press: New York, NY, USA, 2020. doi:10.1007/978-1-0716-0430-4_25

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Computational toolbox for the analysis of protein–glycan interactions

Ferran Nieto-Fabregat, Maria Pia Lenza, Angela Marseglia, Cristina Di Carluccio, Antonio Molinaro, Alba Silipo and Roberta Marchetti*

Review

Open Access

Address:
Department of Chemical Sciences, University of Naples Federico II, Via Cinthia 4, 80126, Italy

Email:
Roberta Marchetti* - roberta.marchetti@unina.it

* Corresponding author

Keywords:
computational tools; glycan–protein interactions; MD; molecular recognition

Beilstein J. Org. Chem. **2024,** *20,* 2084–2107.
https://doi.org/10.3762/bjoc.20.180

Received: 20 December 2023
Accepted: 01 August 2024
Published: 22 August 2024

This article is part of the thematic issue "Chemical glycobiology".

Guest Editor: E. Fadda

Abstract

Protein–glycan interactions play pivotal roles in numerous biological processes, ranging from cellular recognition to immune response modulation. Understanding the intricate details of these interactions is crucial for deciphering the molecular mechanisms underlying various physiological and pathological conditions. Computational techniques have emerged as powerful tools that can help in drawing, building and visualising complex biomolecules and provide insights into their dynamic behaviour at atomic and molecular levels. This review provides an overview of the main computational tools useful for studying biomolecular systems, particularly glycans, both in free state and in complex with proteins, also with reference to the principles, methodologies, and applications of all-atom molecular dynamics simulations. Herein, we focused on the programs that are generally employed for preparing protein and glycan input files to execute molecular dynamics simulations and analyse the corresponding results. The presented computational toolbox represents a valuable resource for researchers studying protein–glycan interactions and incorporates advanced computational methods for building, visualising and predicting protein/glycan structures, modelling protein–ligand complexes, and analyse MD outcomes. Moreover, selected case studies have been reported to highlight the importance of computational tools in studying protein–glycan systems, revealing the capability of these tools to provide valuable insights into the binding kinetics, energetics, and structural determinants that govern specific molecular interactions.

Review

Introduction

Carbohydrates also referred to as saccharides, sugars, or glycans, constitute one of the main building blocks of biomolecules, alongside lipids, proteins, and nucleic acids. In humans and animals, they form the so-called glycocalyx, a protecting sugar coat decorating the cell surface and modulating a myriad of cell–cell interactions [1]. It is composed of branched or elon-

gated glycan chains covalently linked to proteins or lipids, hereby constituting glycoproteins or glycolipids, respectively. Recently, glycan structures exposed on the cellular membrane have also been found to be associated with tRNA [2]. In other species, such as prokaryotes, plants or fungi, glycoconjugates comprise the cell wall, playing critical metabolic, structural and physical functions [3].

Glycoscience encompasses the comprehensive study of glycans focusing on their structural, biosynthetic, biological and evolutionary aspects [4], thus playing a central role in the identification and characterisation of the glycome' structure and function, and in unveiling its interaction with host proteins [5,6]. Notably, the complexity of the glycome far surpasses that of the genome, transcriptome, and proteome, not only due to the structural and conformational diversity of glycans, whose synthesis is not template driven, but also due to their dynamic nature [5,6]. Although mammalian glycans rely on a group of "only" 10 monosaccharide units, they can be assembled, in linear or branched chains, through different glycosidic linkages and diverse spatial orientations, which can also undergo modifications, such as methylation, sulfation, and phosphorylation, resulting in a plethora of different and particular structures [7,8]. Additionally, glycans can adopt a wide variety of different shapes; five-membered ring sugars can exhibit envelope and twist conformations usually represented on a pseudo-rotational wheel; while six-membered ring structures can adopt chair (C), boat (B), skew (S), and half-chair (H) conformations (Figure 1). Among them, chair' shapes typically have the lowest energy and are thus preferred, except few cases in which different conformations can exist in a dynamic equilibrium, as for the iduronic acid that can adopt three low-energy solution conformations (Figure 1): $^{1}C_{4}$, $^{4}C_{1}$ (chair forms) and an additional skew-boat shape ($^{2}S_{0}$) [9]. The glycosidic torsion angles Φ (H_1–C_1–O_x–C_x) and Ψ (C_1–O_x–C_x–H_x) describe the relative orientation of two connected monosaccharide units; moreover, when dealing with monosaccharides containing an exocyclic hydroxymethyl group, such as in the case of 1-6 linked sugars, an additional torsion, namely ω (O_6–C_6–C_5–O_5), must be defined and three staggered conformers, denoted as gg/tg/gt (ω angles of −60°/180°/60°, respectively), should be considered (Figure 1).

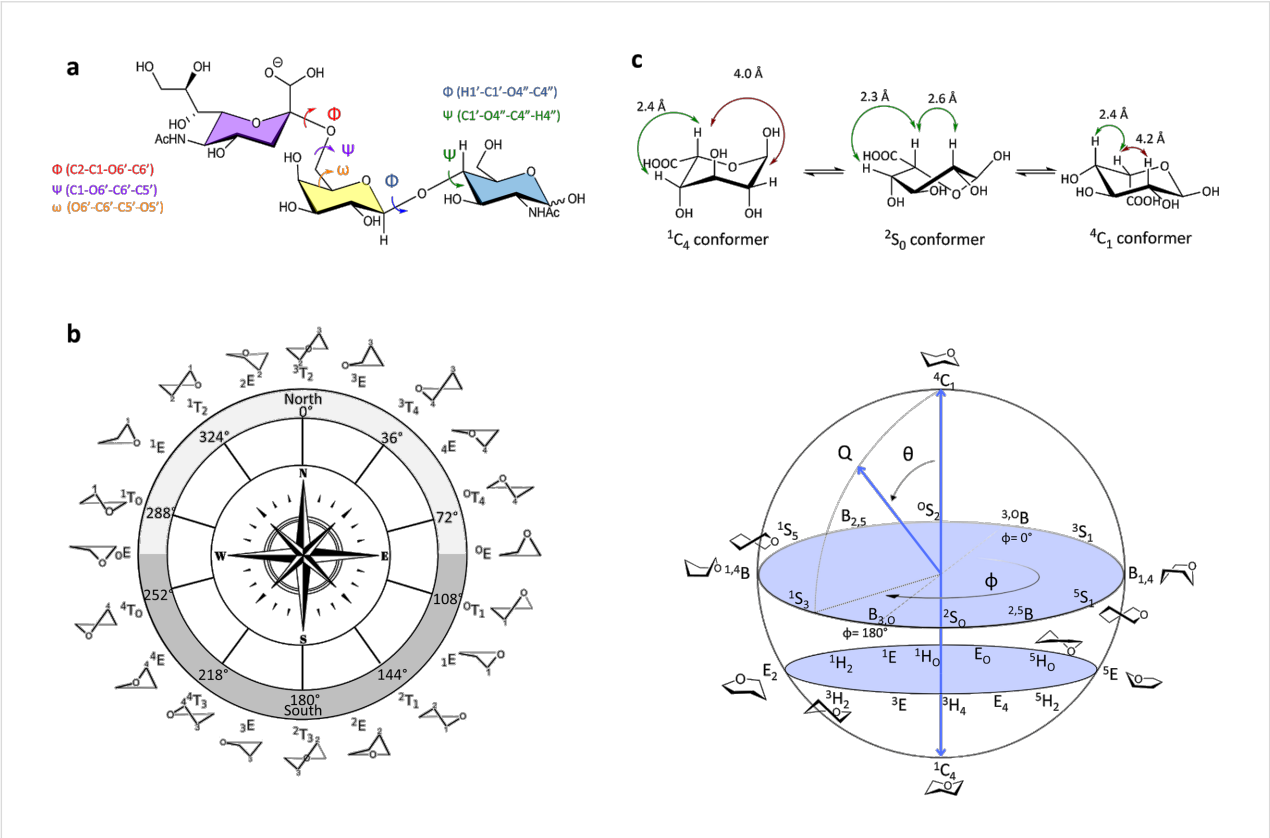


Figure 1: Carbohydrate conformational variability. a) Illustration of Φ, Ψ and ω dihedral angles for a representative trisaccharide coloured according to the symbol nomenclature for glycans (SNFG). b) On the left: Pseudo-rotational wheel depiction of five-membered ring structures showcasing envelope (E) and twist (T) conformations. On the right: Glove representation illustrating the puckering of six-membered pyranoside ring conformations. c) Equilibrium between the low-energy solution conformations of the iduronic acid. The exclusive (Nuclear Overhauser Effect) ^{1}H–^{1}H NOE contacts characteristic of each conformation, $^{1}C_{4}$, $^{4}C_{1}$ and $^{2}S_{0}$, are also depicted.

Longer and branched glycans exhibit heightened structural dynamics, depending on the values adopted by the torsional angles around the glycosidic linkages [10].

The high variability of linkages type, branching, stoichiometry, anomeric configuration (alpha and beta), and conformation contributes to the intricate nature of glycans. The complexity of the glycome is even higher in bacteria, which are able to use most of the mammalian sugar units to construct their glycoconjugates but, in addition, can also use a wide variety of particular, and potentially endless, monosaccharides that are instead not present in eukaryotes (Figure 2).

This huge diversity and complexity, especially in bacterial glycans, makes the structural and conformational analysis of glycans extremely difficult, posing a considerable challenge when employing conventional structural biology methods for glycan analysis [10,12]. Nevertheless, understanding the three-dimensional structure of glycans is crucial for comprehending their roles and biological activities and for correlating their structural features with their activity [3,13]. Given the plethora of remarkable biological roles played by complex glycans, this knowledge is essential for their potential applications in promoting health benefits for humans, animals and plants, including drug design [14,15], vaccine development [15,16] and

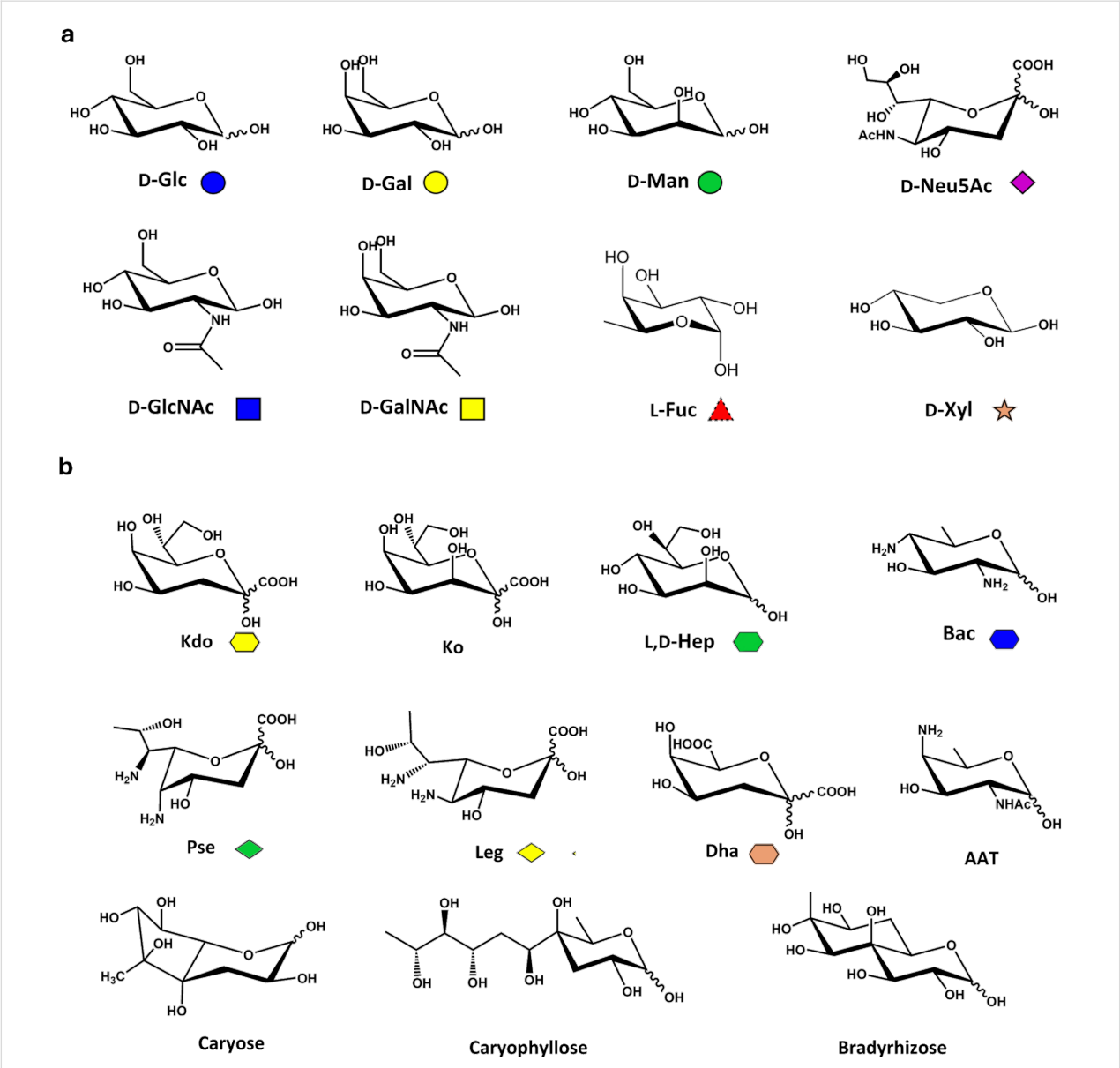


Figure 2: Monosaccharides diversity in eukaryotes and bacteria. a) Eukaryotic monosaccharides. b) Examples of some peculiar bacterial monosaccharides, including hexuronic acids, heptoses, or octulosonic acids. The SNFG symbol [11] of each monosaccharide is also reported (if any).

numerous other possibilities in the field of carbohydrate chemistry and biology.

Notably, the regulation of the host immune response is often mediated by glycans, particularly through their recognition by a wide array of glycan-binding proteins (GBP) [17], which have a unique capability to specifically interact with endogenous and/or exogenous glycans [18,19]. Thus, disclosing the molecular basis of protein–glycan interactions has a unique potential to help modulate a myriad of complex biological events affecting the health and well-being of living organisms and the natural environment. Being key participants in the molecular dialogue, glycan binding proteins emerge as fascinating and critical components of molecular events that regulate life at its core. Their functions span from the catalysis of chemical processes [20,21], transporting and storing molecules [22], transducing and integrating information [23] providing structural and mechanical support [24], and generating movement [25], among other functions [26]. To fold and carry out their function properly, proteins often need post-translational modifications, including glycosylation, in which a carbohydrate chain is directly attached to a specific amino acid to generate glycoproteins and proteoglycans [27]. Based on the amino acid involved in the link with the carbohydrates chain, it is possible to classify different types of glycosylation: i) *N*-glycosylation, where a *N*-acetylglucosamine (Glc*N*Ac) is linked to the nitrogen atom of an asparagine side chain [28]; ii) *O*-glycosylation, where a Glc*N*Ac or *N*-acetylgalactosamine (Gal*N*Ac) is linked to the hydroxy group of a serine or threonine residue [29]; iii) *C*-glycosylation, where a mannose (Man) directly binds a tryptophan residue [30]; iv) the covalent attachment to core protein of glycosaminoglycans (GAGs), anchored to a Ser, or at lesser extent to Thr or Asn, forms proteoglycans. GAGs are complex negatively charged polysaccharides composed by disaccharide repeats of Glc*N*Ac or Gal*N*Ac combined with uronic acid (glucuronic or iduronic acid) or galactose residues, forming chains which can also be partially sulfated. GAGs family includes heparan sulphate (HS), dermatan sulphate (DS), chondroitin sulphate (CS), keratan sulphate (KS), and hyaluronic acid (HA) [31]. The extraordinary proteins versatility places them at the core of almost every biological event, including cell–cell communication and regulation of immune responses. In the majority of cases, these mechanisms are significantly influenced by the molecular interactions occurring between glycans and receptor proteins.

A well-known family of GBPs is constituted by the lectins, ubiquitous receptors that exhibit the ability to specifically recognise different carbohydrates through their well-defined binding pocket and they conserved three-dimensional structure similarities [32]. On the other hand, GAG-binding proteins, which are able to recognise carboxylic acid and sulphate groups along glycosaminoglycan chains using clusters of positively charged amino acids [33], also mediate a wide variety of cell–cell and cell–pathogen communication, controlling immune cell functions, and overseeing cellular trafficking [34]. Another class of GBP is represented by anti-carbohydrate antibodies, that are generally produced by the host organism for example against bacterial, fungal, and viral carbohydrates [35].

Given the wide variety of biological processes influenced by the protein–glycan interplay, an increasing attention has been focused in the last decades on the development of new techniques and technologies for the systematic analysis of complex glycans and the study of their interactions with proteins. A multidisciplinary approach, spanning from wet laboratory experiments and biophysical techniques to bioinformatics methods is needed to deeply investigate the multifaceted aspects of protein–glycan interactions. To date, advanced and versatile NMR, X-ray crystallography, and MS methods [36-38] above all, have been developed to reach extensive information on the structural and conformational features of glycans and proteins. The experimental techniques employed for the analysis of these complex biomolecules are not discussed here; for a more in-depth understanding on this topic, the reader is referred to some comprehensive reviews [7,36,39-41]. Here, we focus instead on different computational and bioinformatic tools, designed to guide the structural and conformational elucidation process, and on the application of molecular dynamic simulations to the study of proteins and glycans in free and bound states. Detailed protocols and methods for protein and glycan modelling are extensively described and links to web servers and downloadable software, which can help researchers in designing the workflow to study a glycan–protein system, are also reported.

Computational tools to study glycans in the free state

Since the first molecular dynamics simulations performed in the late 1980s on oligomannose type glycans [42] and in the early 1990s on complex type glycans [43], great steps forward have been made in the computational analysis of complex carbohydrates. The advancement of computing power, the emergence of GPUs, and specialised processors accelerated MD simulations making it a key scientific tool to explore complex systems, including glycans, with ever-increasing accuracy and efficiency [44].

Tools for building structural models of carbohydrates

Before going into details of the computational tools that can be used to dissect the 3D conformational features of glycans, an

overview of the most useful web services and software to build 2D and 3D models of carbohydrate structures is reported here.

Notably, despite the existence of several encoding formats for glycans (Figure 3), significant efforts have been made in the years to enable a simple and standardised glycan representation, which would simplify the transmission and efficiency of the communication within the scientific community. This led to the extensive use of the symbol nomenclature for glycans (SNFG) representation that is used in all the tools described below (Figure 3) [11,45].

Numerous computer applications have been developed to allow manual drawing and sketching of carbohydrates, as reviewed by Lal et al. [47]. Here, we list free tools useful not only for sketching and drawing but also for building the glycan structure of interest (last accessed date: May 2024) [47,48].

1. doGlycans [49]: Free desktop software package in the python framework that allows users to prepare carbohydrate structures for atomistic simulations of complex glycoproteins, glycolipids and carbohydrate polymers in the GROMACS force field format (see below). Polysaccharides can be prepared by using the *prepreader.py* tool, glycoproteins and glycolipids by using the *doglycans.py* tool. (https://bitbucket.org/biophys-uh/doglycans/).

2. Glycam-Web carbohydrate builder [50]: Free online webservice that gives the possibility to model the 3D structures of molecules and complexes containing carbohydrates starting from monosaccharide building blocks, being also able to add branching points and some sugar derivatisation, including methylation and acetylation. The user has also the possibility to choose the ring type and the anomeric configuration of each monosaccharide. Once the structure is complete, it is possible to download not only the generated .pdb files of the minimised resulting structures but also files for input to an AMBER simulation (https://glycam.org/). Notably, among the currently available interfaces for modelling oligosaccharide conformations on glycam website, one is dedicated to GAG modelling [51] (https://glycam.org/gag/).

3. CHARMM-GUI [52]: Online free web-service that offers a great variety of possibilities for reading and modelling .pdb files. It is a versatile program for atomic-level simulations, which can be run directly in the webserver. It has a special focus on macromolecules of biological interest; indeed, this platform contains a number of different modules designed to construct complex glycans, glycoconjugates as lipopolysaccharides (LPS), or even building a membrane system or solvating a

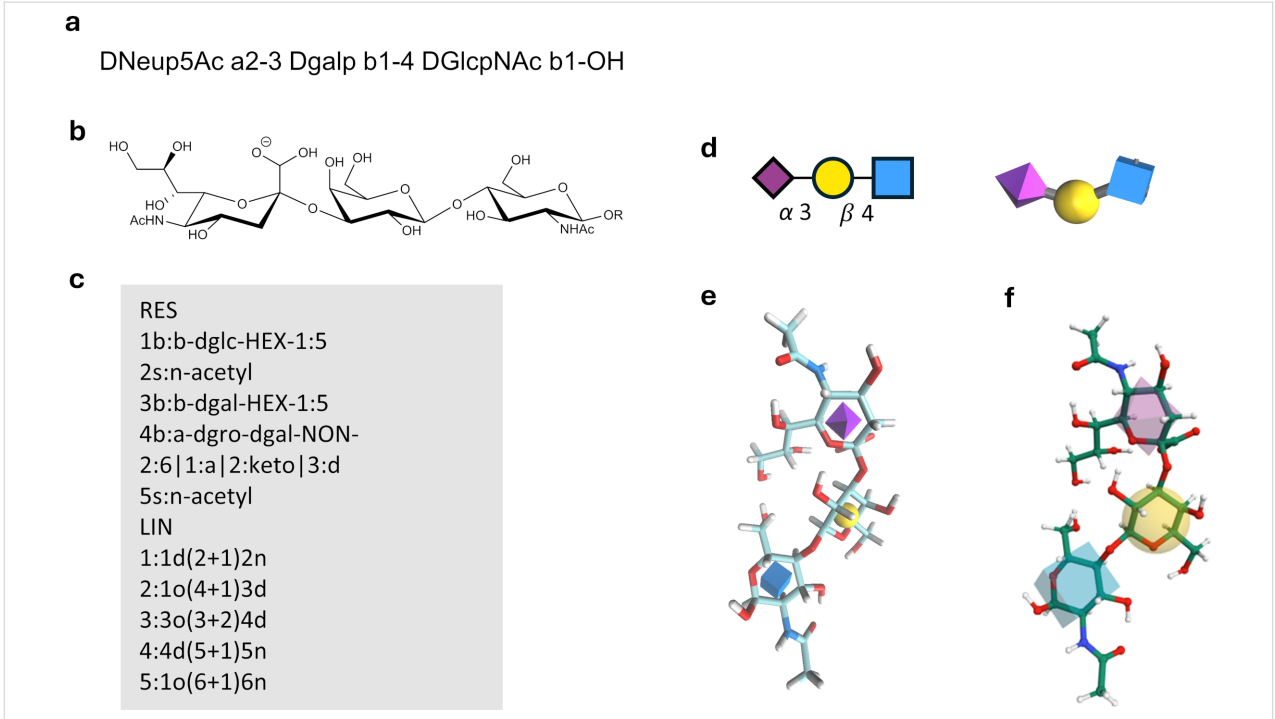


Figure 3: Different glycan representations. The 3'-sialyllactosamine is depicted according to the a) IUPAC nomenclature b) chemical representation as implemented in ChemDraw suite [46] c) GlycoCT nomenclature d) SNFG and 3D SNFG nomenclature e) VMD 3D-SNFG plugin f) MolStar representation.

protein. To use all these features, grouped in the *Input Generator* tab (https://www.charmm-gui.org/), it is necessary to be registered to the web site.

4. Azahar [53]: freely available python-based plugin that permits to visualise, analyse and model glycans and glycoconjugates (https://pymolwiki.org/index.php/Azahar).

5. 3D-SNFG: It is a script integrated in the visual molecular dynamics (VMD) program [54] (see below) that allows a cartoon representation of glycans according to the symbol nomenclature for glycans (see Figure 3) (https://glycam.org/docs/othertoolsservice/downloads/downloads-software/index.html).

Choosing the most appropriate simulation software package

With the aim to accurately prepare glycans for MD simulations, it is fundamental not only to build their 3D structure but also to choose the appropriate force field, that is a set of empirical energy functions and parameters used to calculate the potential energy of a system as a function of the molecular coordinates. The collection of equations and associated constants designed to reproduce the molecular geometry and selected properties of a system, as well as the naming and labelling of the system atoms, vary from one force field to another; therefore, it is important to ensure compatibility between the input file of the tested structure and the chosen force field. The Automated Topology Builder (ATB) and repository [55] (https://atb.uq.edu.au/) is a free web server providing topologies and parameters for a wide range of molecules. It provides access to classical force fields in formats compatible with different simulation packages, including GROMACS (see below), even offering a GROMOS to AMBER topology file converter. In the years, different MD simulation software packages have been developed and designed to simulate the movements and interactions of atoms and molecules over time; the three described below are currently widely used in the field of computational chemistry and biochemistry:

1. AMBER [56]: AMBER (https://ambermd.org/) is the acronym for "Assisted Model Building with Energy Refinement", and it is an open-source software widely employed for molecular modelling and simulation. It is known for its stability, user-friendly interface, and a wide range of analysis tools for studying complex biomolecular systems. AMBER provides various force fields, specifically optimised for simulating biological molecules, as lipids (lipids21) [57], proteins (ff14SB) [58], water molecules (TIP3P) [59], general organic molecules (gaff2) [60], and sugars (GLYCAM_06j) [61]. Notably, as mentioned above, on the GLYCAM-web site, it is possible to easily construct glycans with GLYCAM force field nomenclature; however, it is worth to note that only a few bacterial monosaccharides are available in the GLYCAM-web carbohydrate builder. For most bacterial sugars, the parametrisation of each building block is needed and requires the use of ab initio methodologies, including several steps of charges and electron density calculations, optimization and minimization, making the computational study of bacterial glycans difficult and time-consuming.

2. CHARMM [62]: CHARMM, acronym of "Chemistry at HARvard Molecular Mechanics", is a free extensively utilised molecular modelling and simulation software package. Its force field is at the core of CHARMM's capabilities, which serves as a comprehensive set of parameters and mathematical functions to describe the potential energy and interatomic interactions within a molecular system. The CHARMM force field includes parameters for various types of atoms, bonds, angles, dihedrals and non-bonded interactions, encompassing van der Waals forces and electrostatic interactions. CHARMM19 (united atom), CHARMM22, CHARMM27, and CHARMM36 (all atom) are some of the popular force fields available in the program. (https://www.academiccharmm.org/)

3. GROMACS [63]: GROMACS is the acronym for "Groningen Machine for Chemical Simulations"; it is a powerful open-source software package for molecular dynamics simulations in the field of computational chemistry and bioinformatics. It is extensively used to model and simulate the dynamic behaviour of various molecular systems, including proteins, nucleic acids, lipids, and small molecules. GROMACS provides various tools for system preparation, simulation setup, and post-simulation analysis. It is possible to use different force fields that include GROMOS96, GROMOS53A6, and GROMOS54A7, which are suitable for simulations of biomolecules, organic compounds, and a wide range of solvents (https://www.gromacs.org/index.html).

Despite several carbohydrate-specific force fields have been developed over the years [64-68], to date, the most widely used force field for carbohydrates is GLYCAM, which is continuously updated and improved to accurately describe their peculiar and complex set of conformational and energetic properties [61,69]. For specific studies involving unusual ligands, useful tools can be employed to provide the parameters needed for running MD simulations. Charge calculations and electron density computations for glycan units can be performed using tools like the online RED Server [70]. Although information on force fields is usually available, modifications can sometimes be required and can be achieved through ab initio calculations or programs like for example VFFDT [71].

Tools for the conformational analysis of glycans in the free state

The structure and biological functions of glycans are closely intertwined; the roles they play are influenced not only by their chemical composition but also by their conformation. As mentioned above, glycans are characterised by a huge conformational diversity (see Figure 1): even individual furanoid or pyranoid monosaccharides can assume various shapes and in longer glycans the relative orientation of the different monosaccharide building blocks is dictated by the values of different glycosidic torsion angles.

MM calculations: Investigating the energetically favourable conformations of carbohydrate disaccharide units composing the molecule of interest represents a pivotal step for generating reliable 3D glycan structure. A first analysis of glycan conformational features can be done by means of molecular mechanic calculations that allow to build the adiabatic energy maps, represented as a function of Φ and Ψ torsion angles, in which the energetic minima that can be populated by a specific disaccharide are reported [72-74]. Currently, different databases, which are described below, collect adiabatic energy maps facilitating the construction of glycan 3D models by enabling the selection only of permitted low energy conformations:

1. CSDB [75-77]: Carbohydrate Structure Database is a publicly accessible platform for multiple glycoinformatic studies and web tools, which among the other services allows the users to locate adiabatic maps for specific glycosidic linkages. Generally, CSDB offers a wealth of valuable features; it provides structural, bibliographic, taxonomic, NMR spectroscopic and other information on glycan and glycoconjugate structures of prokaryotic, plant and fungal origin. The retrospective literature analysis is the main source of structural data, which are then manually curated and approved. Besides structures, the database includes bibliography, abstracts, keywords, biological source data up to strains, methods used to elucidate structures, NMR signal assignment and other information (http://csdb.glycoscience.ru/database/).

2. Disac3DB: free annotated database that contains the 3D structural information of about 120 entries of disaccharides. For each disaccharide, an exhaustive search was performed using the MM3 molecular mechanics force field [66], giving a complete sampling of the conformational space and yielding the construction of relaxed adiabatic energy maps (https://glyco3d.cermav.cnrs.fr/disac3db/). It is worth to note that the presence of additional residues in the neighborhood of the studied glycosidic linkage may cause shifts in the values of the favored torsional angles. Thus, to evaluate if the presence of further residues results in limitations of the possible conformations of an individual glycosidic linkage, and/or if the adiabatic map of interest is not present in the aforementioned databases, the Schrodinger Suite of programs through the Maestro graphical interface can be exploited to generate the maps by using the MM3 force field [66]. The Schrodinger platform (https://www.schrodinger.com/) offers several services for molecular design and discovery providing access to physics-based molecular modelling tools and machine learning technologies from a single modelling environment, however, it is not free-accessible.

Further valuable insights into the structure and conformation of saccharides, determined by experiment and simulation, are available on the Stenutz's website (https://www.stenutz.eu). In particular, information on the preferred conformation of glycosidic linkages and the favoured dihedral angles for the OH group at position 6 in hexoses are reported. This website also provides a compilation of standardised procedures, providing practical guidelines for carbohydrate structural analysis, spanning from the purification to the structural analysis of polysaccharides.

MD simulations: Once the 3D glycan structure is built, taking into account the energetically favourable conformations of each constituent disaccharide unit, and the appropriate force field/simulation package is chosen, molecular dynamics simulations can be performed to gain insights into the glycan conformational behaviour. MD simulation generates an ensemble of conformations by applying the laws of motion to the atoms of the molecule [48], allowing to: i) sample the glycan conformational space; ii) investigate how the glycan behaves in a solution (if the MD is performed in explicit solvent), describing carbohydrate–water interactions; iii) monitoring the intramolecular interactions. Usually, this information has to be further validated by performing experimental studies (primarily nuclear Overhauser effect (NOE) and residual dipolar coupling-based experiments) to get accurate information on the glycan conformational behaviour and eventually apply some experimentally derived constraints.

The calculation and analysis of MD simulations of glycans in the free state can be performed with the same tools described below in the protein–ligand interactions section.

Computational tools to study proteins in the free state

Knowledge on the three-dimensional structure of a protein is essential for understanding the functions and the dynamics of protein interactions. Several experimental techniques, including NMR, X-ray crystallography and Cryo-EM, can provide critical information for the characterisation of protein structure and

conformation. Their widespread and wisely use permitted to experimentally determine the structures of around 200,000 proteins [78], all organized in the Protein Data Bank (PDB), that is a freely and publicly available central archive of macromolecular structural data, established in 1971. However, the three-dimensional shape of billions of known protein sequences is not available yet. In this scenario, bioinformatic tools can come to the aid of predicting protein three-dimensional structure with high accuracy, as outlined below [79-85]. Notably, generated models have also occasionally helped solve protein structures [86], further highlighting the great potential of the integrated use of bioinformatic tools and experimental data.

Tools for protein structure prediction

Due to the vast conformational space and a complex energy function, protein structure prediction (PSP) is a computationally challenging task. Homology modelling is a template-based PSP that may be used to predict the 3D structure of a protein based on its amino acid sequence and the structure of a related protein that is already known. However, also template-free PSP has obtained significant progress recently via machine learning and search-based optimisation approaches [87]. There are several software programs and tools available for homology modelling, and some of the most popular include:

1. AlphaFold2 [88]: It is an open-access protein structure prediction system based on artificial intelligence and machine learning. It is based on a neural network that can predict the 3D protein structure at a high accuracy level. The AlphaFold solution is composed of two steps. First, given a protein sequence, it generates multiple alignments with sequences from all the species, including evolutionary profiles from different sources. In the second step, a model refinement is generated based on structural refinement (where the network optimises the torsion angles, bond length and bond angles), distance constraints (according to laws of physics) and gives an output with the structure with the minimised energy (https://alphafold.ebi.ac.uk/).

2. I-TASSER [89]: Iterative Threading ASSEmbly Refinement is a free online server that combines ab initio protein structure prediction with template-based modelling. It is known for its ability to predict both the structure and function of a protein. It is based on identifying structural templates from the PDB by several threading methodologies with full-length atomic models (https://zhanggroup.org/I-TASSER/).

3. Modeller [90]: It is an open-access program used for homology or comparative modelling of proteins. The user inputs an alignment of a sequence to be modelled with known related structures, and the computer generates a model of all non-hydrogen atoms. It can do de-novo modelling of protein loops and apply spatial constraints (https://salilab.org/modeller/).

4. Rosetta [91]: It was developed for de novo protein structure prediction in a free version. Homology modelling is also applied in this instance by using several protein templates that hybridise the most homologous sections of various templates into a single model while modelling missing residues de novo. Advances in the scoring function, which is a mix of physics-based and knowledge-based potentials fitted against known structures and thermodynamic observables, have increased the accuracy of predictions. Incorporating experimental data into models has been made more accessible. The same research group also developed RoseTTAFold, which uses deep learning to quickly and accurately predict protein structures based on limited information [92]. However, very accurate structures for complex proteins are yet to be achieved at a level suitable for effective drug design. Moreover, ab initio prediction of a protein's structure only from its amino acid sequence remains unsolved. Accessing Rosetta molecular modelling software tools (https://www.rosettacommons.org/software) has traditionally required expertise in the Unix command line environment, limiting their use. A web server called ROSIE [93] was created to provide a more accessible environment for selected Rosetta protocols. Academic users can access ROSIE freely (https://rosie.rosettacommons.org/).

5. SWISS-MODEL [94]: It is a web-based integrated free service dedicated to protein structure homology modelling. It guides the user in building protein homology models at different levels of complexity. This program builds a homology model by employing four main steps: (i) identification of structural template(s), (ii) alignment of target sequence and template structure(s), (iii) model-building, (iv) model quality evaluation. Each of the above processes may be repeated interactively until a satisfactory model is produced (https://swiss-model.expasy.org/).

6. UniLectin [95]: It is an interactive, publicly accessible platform that provides curated and predicted lectin data, not only including structural information on lectins and their interactions with carbohydrate ligands, but also predicting the occurrence of lectins in genomes. UniLectin3d is one of the modules integrated in UniLectin, which provides curated information on 3D structures of lectins [94-96] a classification system based on both taxonomic origin and structural fold (https://unilectin.unige.ch/).

7. GlycoShape3D [97]: It is a freely available database for academic user that enriches the landscape of glycobiology

resources. It offers structural insights into glycoproteins, addressing challenges posed by glycan complexity, flexibility, and heterogeneity. In particular, the Re-glyco tool allows the user to restore the missing glycosylation on glycoproteins deposited in the RCSB PDB or in the EBI-EMBL AlphaFold protein structure database (https://glycoshape.org/).

The quality of the generated protein model is contingent on elements such as the chosen template structure (if any), the sequence alignment, and the choice of the modelling algorithm. To ensure a high accuracy of the predicted model, which should be at least comparable to that of experimental structures, several programs can be employed for model validation and refinement. Among them, PROCHEK [98] is an open-source program that permits to check the quality of the protein structure by analysing the Ramachandran plots, the planarity of peptide bonds, the bad non-bonded interactions, the distortions of the geometry around the Cα atoms, the energies of hydrogen bonds, and the departure of the side chain χ torsion angles from expected values. Improvement and/or validation of modelled or experimentally solved structures can be also obtained by using CASP (critical assessment of protein structure prediction) [99], which consists of a free platform established in 1994 to help advance the methods of identifying protein structure from sequence. The Protein Structure Prediction Center (https://www.predictioncenter.org/) has been organized to allow researchers to objectively test their structure prediction methods. Some of the best performing methods (including among the others AlphaFold, RosettaFold and I-TASSER) are implemented as fully automated servers, which can be used by public for protein structure modelling.

MD simulations: Generating an accurate protein model or choosing the appropriate published 3D structure of a protein is essential to obtain reliable and precise results from MD simulation. It is also worth to note that, to generate the input files for MD, some modifications on the .pdb file of the protein are required. For instance, capping the protein termini with non-charged groups and replacing original hydrogens to guarantee compatibility with the selected force field is required before running MD simulations. Other modifications can include adding a disulfide bond between specific cysteines and filling the missing side chains and missing loops (if any) to restore the integrity of the protein. Here, we list a series of software tools and packages which are commonly employed to generate the protein input files for MD:

1. Molprobity [100]: It is a widely used web-based software suite for evaluating and enhancing the quality of protein structures, especially those intended for molecular dynamics simulations, available in a free version. Specifically, it is possible to check the H atoms, the quality of the structure, evaluate some steric clashes and visualise in a friendly manner the full structure (http://molprobity.biochem.duke.edu/).

2. PDBtools [101]: It is a freely accessible software that allows the manipulation and modification of a PDB file. Different tools are available, such as deleting atoms, renaming the polypeptide chain, calculating disulphide bonds, adding missing atoms and mutating residues (https://wenmr.science.uu.nl/pdbtools/submit).

3. ProteinPrepare [102]: This application enables users to modify PDB files and create input files for molecular dynamics by adding missing atoms, removing H atoms, and analysing the proton state of amino acids. The registration of the user to the web-site is required to access these tools (https://playmolecule.com/proteinPrepare/).

Several MD simulation packages, including CHARMM-GUI [52], AMBER [56], and GROMACS [63], offer built-in utilities for preparing input files. These tools also provide extensive documentation and tutorials to help users effectively create MD input files for proteins. Once the protein input files are generated, MD simulations can be run.

Computational tools to study protein–ligand complexes

Detailed investigations of protein–ligand interactions, combining experimental and computational methods, provide an indispensable basis to depict holistic pictures of molecular complexes allowing to modulate them at will. The computational approach involves i) predicting/building the protein and the ligand in their optimal conformation (as discussed above), ii) predicting the protein binding site; iii) modelling the ligand into the protein binding site, iv) assessing binding affinity through sampling and scoring, as discussed in the following paragraphs [103].

Prediction of the protein binding site

Over the years, structure-, sequence-, and homology/template-based methods have been employed to identify and predict carbohydrate-binding sites starting from the protein structure [104]. Recently, thanks to the fast development of machine learning techniques, new computational tools have been developed to facilitate the prediction of protein binding sites.

We report here only the applications related to the protein interaction with glycans:

1. PeSTo-Carbs [105]: it is an extension of Protein Structure Transformer (PeSTo) [106], a deep learning method to

predict protein interaction interfaces with other proteins, nucleic acids, lipids, small molecules, and ions, starting from a protein structure. PeSTo-Carbs is specifically trained to predict carbohydrate and cyclodextrin binding interfaces on proteins. Two different modules are available: a general model PS-G for a wide range of carbohydrates, their derivatives and cyclodextrins, and a specific model PS-S for important carbohydrate monomers. All of these features are available for free without registration as online tools (https://pesto.epfl.ch/).

2. GlyNet [107]: it is a free deep learning algorithm, based on neural networks (NN), that allows the user to predict protein-glycan binding. Taking a glycan structure as input, this model is able to predict the strength of the interaction based on the relative fluorescence units (RFUs) measured in the Consortium for Functional Glycomics glycan arrays and extrapolating these to RFUs from untested glycans (https://github.com/shauseth/glynet).

3. LectinOracle [108]: it is a freely available deep learning-based model that combines transformer-based representations for proteins and graph convolutional neural networks for glycans to predict their interaction (https://github.com/BojarLab/LectinOracle).

4. CAPSIF [109]: CArbohydrate-Protein Site IdentiFier is a convolutional neural network able to predict protein–carbohydrate binding interface from a protein structure. In contrast to other DN algorithms, as GlyNet and LectinOracle, which predict lectin-carbohydrate binding on a protein level, it provides residue-level information for non-covalently bound carbohydrates either from an experimental or generated-model protein structure. It includes two modules: CAPSIF-Voxel that predicts the protein binding residues and CAPCIF-Graph that predicts which residues bind sugars. It is freely available for use, and the code for CAPSIF can be accessed on GitHub (https://github.com/Graylab/CAPSIF).

To identify potential binding sites on the protein's surface, docking calculations can also be performed (see below).

Docking calculation tools for interaction studies

Molecular docking plays a crucial role in computer-aided drug development, allowing systematic evaluation of compound libraries to identify high-affinity lead compounds for specific targets. Bio-algorithms enable modelling protein tertiary structures, predicting ligand binding pockets, and supporting drug discovery through molecular docking [110]. Advances in information technology and improved computational efficiency have made computational methods integral to modern biological research, and large-scale structure-based docking screens have become common, facilitating the exploration of vast chemical spaces and identifying potential target hits from extensive compound libraries [103]. While docking programs and servers may exhibit variations in their operational methods, they generally adhere to a common workflow comprising two primary phases. The first phase involves a conformational search aimed at predicting potential ligand conformations. This is followed by the second phase, which focuses on scoring the binding poses obtained during the conformational search. In this phase, the generated ligand–receptor complexes are assessed and ranked based on their binding energy thanks to the use of scoring functions [111].

Docking calculations can be conducted in two distinct ways: blind dockings, which explore the entire protein surface [112], and directed docking, typically employed when prior knowledge of the binding pocket exists and performed within a predefined box. Blind dockings are performed using cavity detection programs and online servers, as follows:

1. CB-Dock2 [113]: Cavity-detection guided Blind Docking 2 (https://cadd.labshare.cn/cb-dock2/index.php) is an online protein–ligand docking program designed to perform blind docking at predicted sites instead of the entire surface of a protein. Thus CB-Dock automatically recognises putative binding sites to determine their centre and size, with the aim to adjust the docking box to suit specific query ligands. Finally, molecular docking calculations are performed with Autodock Vina (see below).

2. Fpocket [114]: It is an open-source pocket detection package based on Voronoi tessellation and alpha spheres. It consists of three main programs: Fpocket for pocket identification, Tpocket for benchmarking pocket detection, and Dpocket for collecting pocket descriptor values. Written in C, Fpocket is well-suited for developing new scoring functions and extracting various pocket descriptors on a large scale. Fpocket 1.0 outperforms industry standards by detecting a high percentage of pockets within the best-ranked ones and offers a fast, open-source solution for protein pocket detection (https://github.com/Discngine/fpocket).

3. GRID [115]: It is a computational tool used to identify energetically favourable binding sites, known as molecular interaction fields (MIFs), on molecules with known structures. GRID has various applications, including ADME prediction, site of metabolism prediction, ligand-based and structure-based design, pharmacophore elucidation, water network prediction, and 3D-QSAR. GRID 2021 introduced a new interface aimed at structure-based design. It enables users to explore binding sites

using classic GRID MIFs, encompassing 74 different chemical types. Additionally, it offers a new molecular probe for generating MIFs specific to fragments of interest. GRID 2021 includes a 3D sketcher for visualising ligand modifications, and its Designer mode assists in finding optimal chemical moieties for specific sites (https://www.moldiscovery.com/software/grid/#:~:text=GRID%202021%20is%20a%20new,MIFs%20for%20fragments%20of%20interest).

When there is prior knowledge of the protein binding pocket, it is time saving to define an optimal docking search space or box and study specific binding pockets, improving docking accuracy and efficiency. Customising the box size for individual ligands, based on their size and the relationship with the search space, can be done by comparing the target protein to related proteins or those co-crystallised with ligands [103] and manually superimposing the new ligand to the reference structure. Some payment software like Glide [116,117], GOLD [118] and Molecular Operating Environment (MOE) dock [119] can be used for this purpose, although here are listed free docking tools:

1. Autodock [120,121]: It is a suite of advanced docking tools used for predicting how small molecules, such as drug candidates or substrates, interact with known 3D protein structures. It offers two generations of software, namely AutoDock 4 and AutoDock Vina, and a user-friendly graphical interface called AutoDockTools (ADT) to assist in configuring ligand rotatable bonds and analysing docking results. Additionally, the accelerated AutoDock-GPU is designed for faster performance, surpassing the original single-CPU docking code by hundreds of times. AutoDock 4 comprises two main programs: *autodock* handles ligand docking by aligning it with precomputed protein grids, while *autogrid* generates these grids. The grids can also assist organic chemists in designing better binding molecules (https://autodock.scripps.edu/).

2. Autodock Vina [122]: It is the open-source improved successor of Autodock. Vina is improved in terms of accuracy and performance as simplifies the process by instantly calculating grids internally, eliminating the need for manual grid map selection and atom type assignments (https://vina.scripps.edu/#).

3. FlexAID [123]: It is a molecular docking software capable of setting small molecules and peptides as ligands, and proteins and nucleic acids serve as docking targets. Notably, FlexAID shows support for full ligand flexibility and the flexibility of side chains in the target. It achieves this by employing a soft scoring function that assesses the complementarity between the surfaces of the ligand and the target. Thus, FlexAID has demonstrated superior performance compared to well-established software like AutoDock Vina, particularly when target flexibility plays a pivotal role, as is often the case when working with homology models (http://biophys.umontreal.ca/nrg/resources.html).

4. HADDOCK [124]: High Ambiguity Driven protein–protein DOCKing is an advanced computational approach used for modelling interactions in biomolecular complexes. Noteworthy, HADDOCK incorporates information from known or predicted protein interfaces into the docking process through ambiguous interaction restraints and allows the specification of precise distance restraints (e.g., based on MS cross-links). It also supports a range of experimental data, including NMR residual dipolar couplings, pseudo contact shifts, and cryo-EM maps, positioning HADDOCK as a versatile tool capable of handling various modelling scenarios, such as protein–protein, protein–nucleic acids, and protein–ligand interactions.

The majority of existing docking software was originally designed for small, rigid, drug-like molecules, therefore, limiting their effectiveness in studying protein–carbohydrate interactions [125,126]. The development of specialized programs has been crucial in enhancing the accuracy of docking calculations [125,126]. We here listed a series of programs for running docking calculations with a special focus on those specifically designed to address the unique challenges posed by glycans [127,128].

1. Vina-Carb [129,130]: Vina-Carb is a module of AutoDock Vina (downloadable with a free version at https://glycam.org/docs/othertoolsservice/downloads/downloads-software/index.html), proven to be a valuable tool for studying carbohydrates. It incorporates carbohydrate intrinsic (CHI) energy functions and explicit water to better handle glycosidic linkages and improve docking accuracy. When Vina-Carb was applied to antibodies, lectins, and carbohydrate binding modules (CBM), the success rates in predicting accurate binding modes reached 86%, 50%, and 42%, respectively, compared to 70%, 50%, and 0% for AutoDock Vina. Although Vina-Carb generally performed slightly better over AutoDock Vina when docking glycans to proteins, it does not always rank the best docking pose as the top scoring pose.

2. BALLDock/SLICK [131]: It is a molecular docking method specifically designed to accommodate carbohydrate-like compounds, employing a genetic algorithm that allows for ligand and receptor side-chain flexibility. Designed specifically for protein–carbohydrate interactions, SLICK includes terms that consider CH–π interactions, hydrogen bonds, smoothed van der Waals interactions, and electrostatic interactions. The SLICK

scoring function, tailored for carbohydrates, enhances the accuracy of predicting binding modes and free binding energies. Compared to other programs such as AutoDock and FlexX (see below), BALLDock/SLICK demonstrates superior performance in structural and energetic precision. This method is particularly valuable in drug design involving protein–carbohydrate interactions, addressing weaknesses such as the CH–π interactions that are challenging for other programs like Vina-Carb.

3. FlexX [132,133]: It is a molecular docking software (unfortunately not free) designed to predict the conformations of small molecules in protein binding sites, thus facilitating the discovery of new drugs. Within SeeSAR, its functionality allows ligands to be placed in binding sites using an incremental construction algorithm that splits ligands into fragments, places, and scores them quickly in the binding site. The best fragments are then assembled to form the complete ligand, optimizing the generated conformations. FlexX's strengths include rapid and efficient exploration of the conformational space, handling ligand flexibility, a precise scoring function, and smooth integration with other molecular modelling programs. Additionally, FlexX excels in processing large libraries at high speed and is user-friendly, requiring no prior receptor preparation.

4. ROSETTA [134]: The development of GlycanDock [134], a protein−glycoligand docking refinement algorithm integrated in the RosettaCarbohydrate framework [135], allowed the use of Rosetta macromolecular modelling and design software suite to perform docking calculations on glycans bound to proteins with a higher accuracy with respect to previous Rosetta's protein−small molecule docking algorithms. Unlike other docking programs such as AutoDock, AutoDock Vina, DOCK, FlexX, Glide, and GOLD, which are primarily designed for small, rigid ligands, GlycanDock is specifically optimized to address the flexibility and complex structural features of glycans. The carbohydrate chains are treated as flexible oligomers, allowing extensive conformational sampling of the glycoligand while maintaining glycosidic linkages within predetermined, energetically favorable minima to ensure biophysically realistic carbohydrate structures. GlycanDock handles the flexibility and complexity of glycans better than other docking programs and can be downloaded as part of the Rosetta package from the Rosetta Commons (https://www.rosettacommons.org).

5. GlycoTorch Vina [136]: GTV is a free molecular docking tool specifically designed for GAGs. Based on Vina-Carb, it enhances the accuracy of modeling these carbohydrates by including parameters for sugars in the 2SO conformation and glycosidic linkages specific to GAGs. GlycoTorch Vina also allows the integration of experimental data, such as NMR, and considers water-mediated interactions, providing more accurate predictions in the formation of GAG-protein complexes.

6. DOCK [137]: It is a molecular docking program (free for academic research) that predicts the orientation and conformation of ligands within the binding site of proteins or nucleic acids. It uses an incremental construction approach ("anchor-and-grow") to handle ligand flexibility and employs a scoring function based on the AMBER force field to evaluate the stability of the complex. DOCK is particularly useful for GAGs due to its ability to accurately model conformational flexibility and enhance sampling, allowing for more precise predictions of ligand–receptor interactions.

7. ATTRACT [138]: It is a docking (not free) program that models interactions between proteins and other biomolecules such as DNA, RNA, and small ligands. Originally designed for protein docking, it has been successfully adapted for GAGs due to its coarse-grained force field approach, which allows for protein flexibility and the simultaneous handling of multiple protein bodies [139]. This adaptability makes it particularly useful for large and dynamic complexes. Although not initially intended for GAGs, researchers have modified its protocols to account for the unique features of these molecules, such as their high flexibility and electrostatic charge. This has enabled ATTRACT to effectively predict binding poses and rank GAG-protein complexes, demonstrating its utility and versatility in advanced biological interaction studies.

The key distinctions among various docking programs stem from the specific computational search algorithms they employ and the characteristics of the scoring functions utilised to order the docked poses. Over the years a plethora of different scoring functions have been developed, in particular thanks to the evolution of machine learning and collection of high-resolution structural information [140-142]. Recently, some of them have been also optimised for evaluating the binding affinities between proteins and carbohydrates [143,144]. Among them, the CSM (cutoff scanning matrix)-carbohydrate outperforms previous methods and scoring functions, also providing a freely accessible and user-friendly web interface and an application programming interface (API) (http://biosig.unimelb.edu.au/csm_carbohydrate/).

Unravelling complex molecular interactions: tools for molecular dynamics studies

Once the conformational space accessible to the ligand has been studied, the protein binding pocket has been identified, and a

model of protein–ligand complex has been obtained, MD simulations can be performed with the aim to accurately describe the conformational and dynamic properties of the bound state. All-atom MD simulations involve 4 steps: i) energy minimisation, ii) heating, iii) equilibration and iv) production. MD simulations are often run by using explicit solvent box to account for molecular interactions, and if needed, Cl^- or Na^+ ions are added to neutralise the system. Although different programs for running MD simulations, including CP2K [145], DESMOND [146], LAMMPS [147], TINKER [148], YASARA [149], and NAMD [150] are available, the most widely used packages to run all-atom MD simulations on protein glycan complexes are the already described AMBER [56], CHARM-GUI [52] and GROMACS [63].

Tools for the analysis of computational data

Once the protein in the apo-form has been analysed, the ligand in the free-state has been studied, and the protein–ligand complex has been extensively subjected to MD simulations, in-depth insights into the structural and conformational features governing the molecular interactions can be achieved by analysing and visualising the obtained data. The critical step of post-processing analysis involves examining the binding poses of the ligands, evaluating the stability and dynamics of the complexes, investigating the electronic structure and interactions and calculating binding energies or free energy profiles within the system, thus helping in understanding the energetics and thermodynamics of the interactions [151]. To facilitate these tasks, various software tools have been developed; generally, the simulation packages used to run all-atom MD simulations offer built-in utilities and scripts for post-processing the MD data. For example, AmberTools, released within the AMBER suite of biomolecular simulation programs, includes cpptraj and ptraj codes for analysing structure and dynamics in trajectories [152]. Additionally, other programs as VMD (see below) and PLUMED [153], an open-source library compatible with popular MD engines like Amber and GROMACS. (https://www.plumed.org/), can support data analysis for molecular dynamics simulations.

Moreover, several custom scripts have been developed within computational chemistry laboratories (see for example: https://github.com/roviralab/utils) enabling the tracking of glycan conformational changes throughout the dynamics, monitoring dihedral angles, distances, and other parameters. Additional programs allow for the combination and comparison of experimental and theoretical data, enhancing the reliability and accuracy of the simulations. As example, the software package MD2NOE [154] permits to properly simulate NOE effects also of flexible molecules sampling multiple conformational states directly from molecular dynamics (MD) trajectories. With the advent of GPU-based simulation code, indeed, MD simulations have been extended into the microsecond regime, allowing to sample glycan conformational space sufficiently and enabling the computation of key NMR properties [154].

Different visualisation programs, as those described below, play a crucial role in rendering complex 3D structures, visualising molecular interactions, and generating high-quality images for publications or presentations (Figure 4). These user-friendly tools are indispensable for researchers in the fields of structural biology, biochemistry, and computational chemistry, making it easier to comprehend and communicate the results of sophisticated simulations.

1. VMD [54]: Visual Molecular Dynamics is a popular and freely accessible molecular modelling program designed to display, animate, and analyse biomolecular systems using 3D graphics and built-in scripting. It provides tools for simulation preparation, visualisation, and analysis of molecular dynamics. (https://www.ks.uiuc.edu/Research/vmd/).

2. PyMOL [157]: It is an open-source molecular visualisation system developed by Schrödinger. It is one of the most used programs for the visualisation of the 3D structure of the protein alone or in a complex with a ligand. Several tools are available to create, manipulate and visualise the 3D structures. Other tools consent to generate the surface of a protein and highlight the electrostatic potentials or hydrophobicity. PyMOL is used for various applications, such as protein structure analysis, molecular docking studies, and drug design.

3. UCSF Chimera [158]: It is a highly versatile and widely used free molecular visualisation and analysis program developed by the University of California, San Francisco (UCSF). It is a powerful software tool for visualising and analysing the 3D structures of biological macromolecules, such as proteins, nucleic acids, and other complex molecular assemblies. UCSF Chimera provides a user-friendly interface for exploring and manipulating molecular structures, offering a wide range of features for tasks like molecular modelling, molecular dynamics analysis, structural biology, and more. UCSF Chimera is commonly used to gain insights into the structure and function of biomolecules. It supports various file formats, offers diverse visualisation options, and allows for the creation of stunning images and animations of molecular structures, making it an invaluable resource (https://www.cgl.ucsf.edu/chimera/).

4. BIOVIA Discovery Studio [159]: It is a freely downloadable suite of science applications designed for life sciences discovery research, which includes addressing multiple optimisation objectives in drug discovery. This comprehensive software

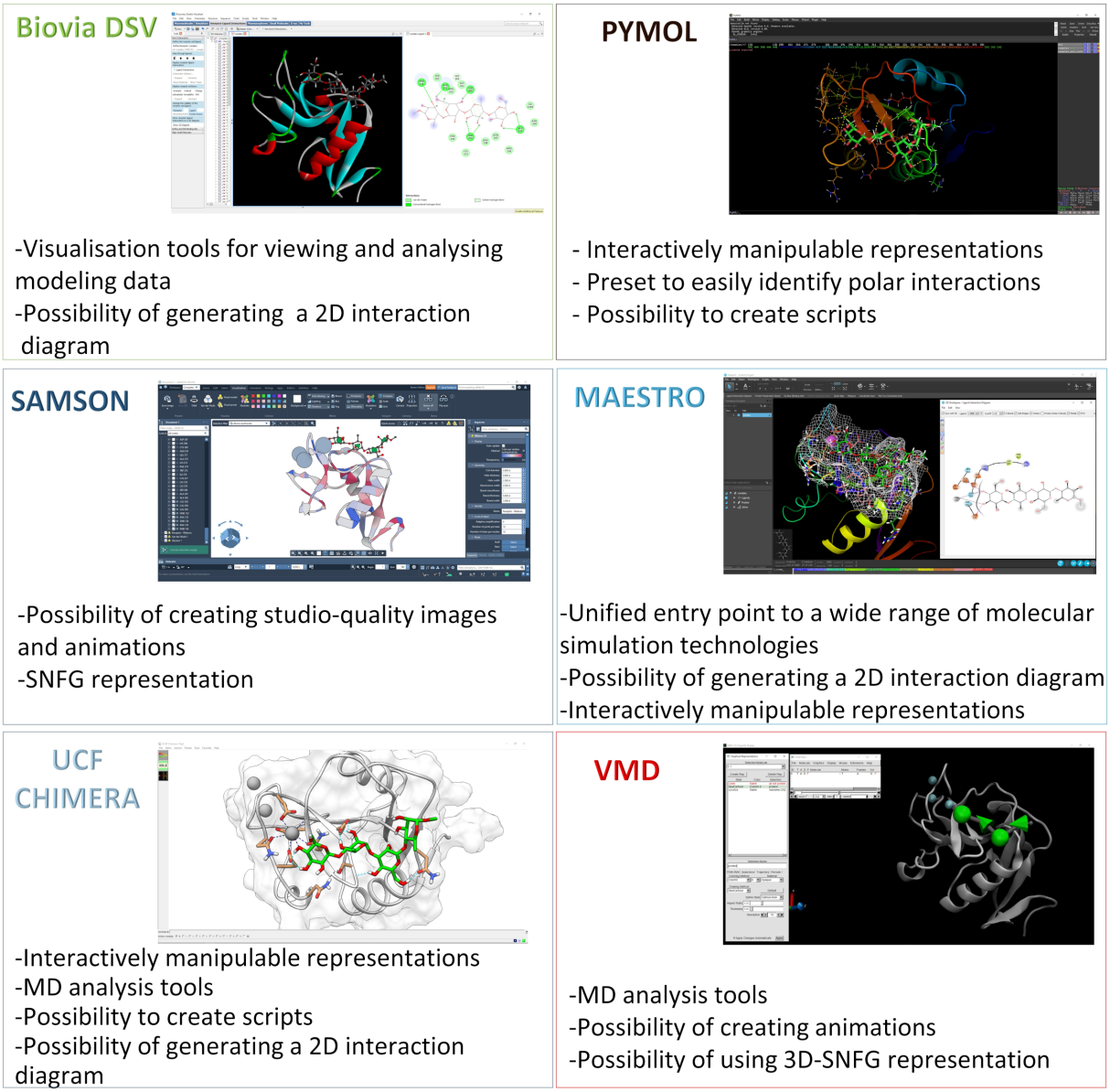


Figure 4: Visualisation programs. Different representation of a protein–ligand complex by using the most used visualisation programs reported in this review. The previously published complex [155] between the Ca^{2+} dependant C-type lectin, DC-SIGN (PDB: 1SL4) [156], and a tetrasaccharide composed of mannose and rhamnose residues, has been used to highlight the main advantages of each visualisation program applied to protein–glycan complexes.

suite, built on BIOVIA Pipeline Pilot, provides a wide range of validated applications. It offers a scalable and collaborative research environment, making it a valuable tool for life sciences discovery research (https://discover.3ds.com/discovery-studio-visualizer-download).

5. Schrödinger Maestro [160]: It is a comprehensive molecular modelling and computational chemistry software suite designed for researcher fields like drug discovery, materials science, and structural biology. Schrödinger Maestro provides tools for molecular visualisation, ligand-receptor docking (with Glide [117]), molecular dynamics simulations, quantum mechanics calculations, and more. It is widely used in the pharmaceutical and biotechnology industries for drug design and discovery, as well as in academic research and other scientific applications that involve the study of molecular structures and interactions (https://www.schrodinger.com/products/maestro).

6. SAMSON [161]: Software for Adaptive Modelling and Simulation Of Nanosystems is a computer software platform for molecular design, unfortunately not freely avaiable. Its modular

architecture enables a wide range of tasks, including model creation, calculations, interactive or offline simulations, and result visualisation and interpretation. Notably, SAMSON offers modules related to glycans and glycans visual models, facilitating the use of the SNFG nomenclature for ligand design and visualisation (https://www.samson-connect.net/).

All the computational tools here reported, summarised in Figure 5, constitute a unique kit for the analysis of protein–glycan interactions.

Computational tools applied to the study of glycans in the free state

In the years, the architectural and conformational features of different mammalian glycans, including oligomannose [162] and complex-type *N*-glycans, have been unravelled by employing computational approaches. As example, the work conducted by A. M. Harbison et al. [163] reported the molecular dynamics of complex biantennary IgG Fc *N*-glycans and their implications for the structural integrity and functionality of human immunoglobulins G (IgGs).

MD methods have also been employed, in combination with experimental methods, as NMR, to explore the three-dimensional features of bacterial glycoconjugates, as the exopolysaccharides [164], and the rough-type lipopolysaccharide [165] isolated from *Methylobacterium extorquens* or the lipopolysaccharide isolated from *Herbaspirillum* Root189 [164]. In these studies, once built the parameters for non-standard bacterial monosaccharides, which were not included in the GLYCAM-website, the overall conformation and properties of the saccharide chain has been accurately described and compared to the experimental NOE data. The tight integration between computational and experimental results allowed to highlight how modifications of the saccharidic backbone, as example with *O*-methyl and *O*-acetyl groups, can affect the polysaccharide biophysical properties tuning its ability to interact with other polymers and/or receptor proteins.

Another example of the application of simulation methods to the analysis of complex glycans is presented by Makshakova et al., who analysed the three-dimensional structure of the exopolysaccharide isolated from *Alteromonas infernus* GY785 [166]. The main chain of the so-called “Infernan” polysaccharide includes glucose, galacturonic acid and galactose, with branches composed of uronic acids and sulphate groups that contribute to modulate its unique properties. Specifically, the authors used molecular mechanics and dynamics calculations to describe the helical structure of the polysaccharide chain and the role of its side chains in the creation of Ca^{2+} chelating sites in the region of the polysaccharide branching points.

A further application of computational tools to the analysis of complex glycans is the study of the conformational behaviour of the naturally cationic polysaccharide, carboxymethyl chitosan (CMCS), reported by Zhang et al. [167]. Due to its non-toxic, biodegradable, biocompatible, and versatile features, chitosan has been widely used in various fields such as biomedicine, cosmetics, agriculture and food. However, its insolubility in neutral or alkaline pH conditions largely limits chitosan's applications. MD simulations were thus employed to mimic the behaviour of CMCS in water under different pH values and different degrees of deacetylation and substitutions in order to study its aggregation pattern.

Computational tools applied to the study of proteins in the free state

Understanding the dynamics of proteins in their free state is key to investigate how they can interact with other biomolecules. MD simulations can be used to study the conformational changes that occur in proteins as they move from one state to another, which is important for understanding their function; additionally, MD simulations can also be used to study the thermodynamics of protein folding, which is important for understanding how proteins fold into their native state.

Recently, extensive MD studies have been performed to investigate the structural and functional features of the SARS-CoV-2 spike glycoprotein, allowing to reveal the critical role of glycans attached to the viral protein in the infection. In particular, the full spike receptor consists of a trimer of S protein, and each monomer comprises 22 *N*-glycan sites. Recent studies suggest that the structure and occupancy of the SARS-CoV-2 S glycans affect the structural integrity of the trimer [168]. Specifically, different Spike protein models with varying glycan compositions in positions N234, N165 and N343 were created by homology modelling using SWISS-MODEL. Then *N*-glycans were added by aligning conformationally equilibrated *N*-glycan structures from a Glyco Shape library to the GlcNAc residues resolved in the cryo-EM structure, with adjustments of the torsion angles to resolve steric clashes with the surrounding protein. By using AMBER, the MD simulation was performed and suggested that diminishing the size of *N*-glycans at position N234 results in destabilising the "wide-open" conformation of the receptor-binding domain (RBD). This destabilisation leads to increased RBD dynamics. Furthermore, the composition of *N*-glycans at positions N165 and N343 influenced the stability of the open RBD, where shorter structures exhibited reduced effectiveness in interacting with the disordered loop within the receptor-binding motif.

Moreover, several groups have employed MD simulation to design new proteins. For example, the group of Mayo [169] em-

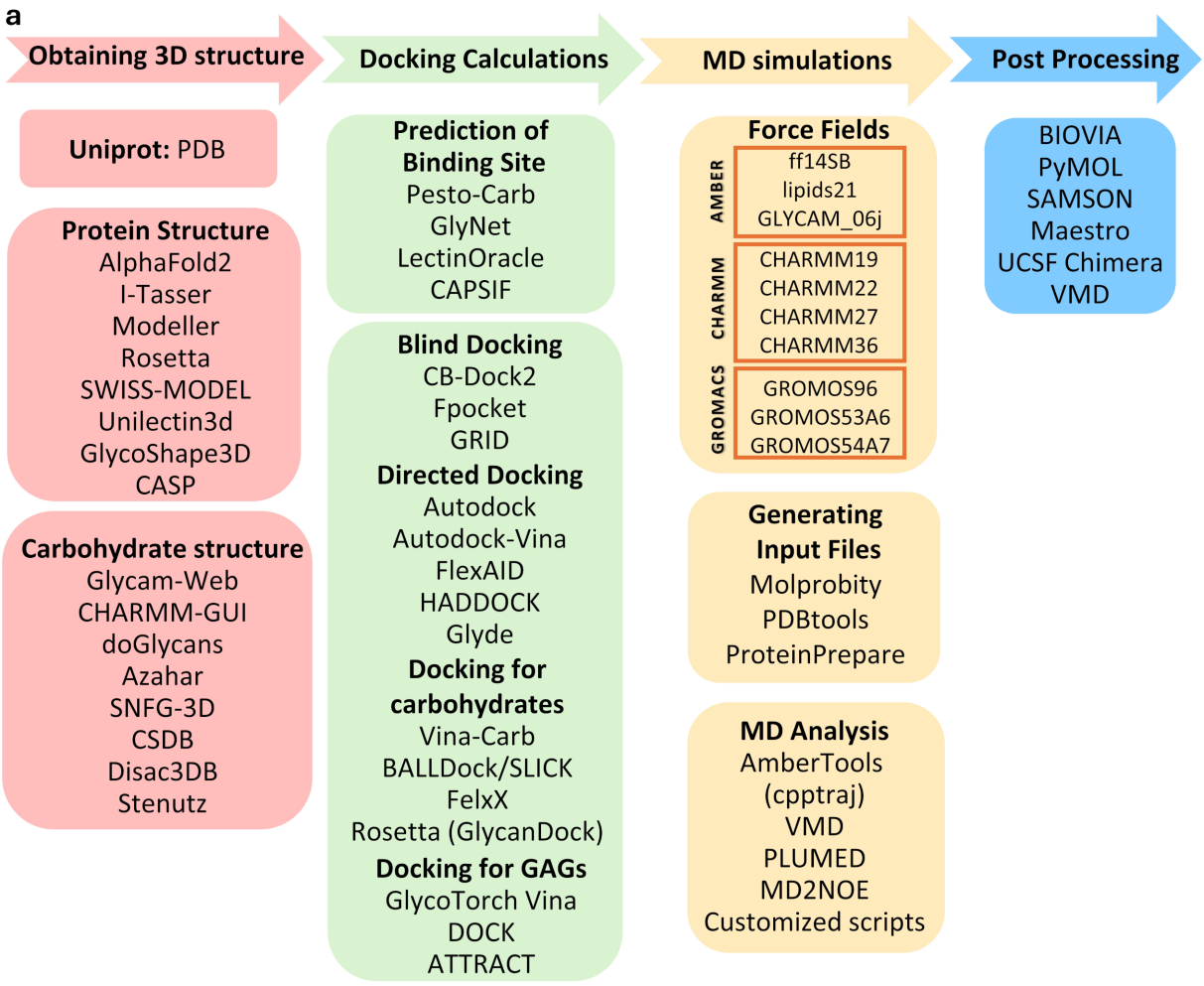


Figure 5: A schematic representation of useful computational methods to study protein–glycan interactions. a) Workflow including different key steps needed to analyse protein–glycan interactions: 1. Building/choosing the appropriate glycan/protein 3D structure; 2. Modelling protein–glycan complex; 3. Running and analysing MD simulations; 4. Processing and visualising the results. b) Summary of the presented tools for building structural model and/or generating topology files of glycan/protein structures.

ployed this approach to engineer a de novo homodimer from a monomeric protein. The integration of computational protein design (CPD) and MD simulation allowed to refine the structural and dynamic aspects of the designed proteins overcoming. CPD inherent limitations, including constraints on side chain rotamers, fixed protein backbones, and a lack of consideration for solvent interactions. Thus, the use of MD simulation allowed to provide a more precise depiction of the protein's behaviour, shedding light on its dynamics and stability.

Another example of the relevance of all-atom molecular dynamics simulations is the study performed by X. Cao et al. [170], which reports the structural dynamics of GH33 sialidases. The computational analysis revealed significant conformational rearrangements within the enzyme active sites leading to the formation of a new cleft to accommodate glycosyl acceptors. Furthermore, the simulations shed light on the role of specific residues within the enzyme's active site, such as the arginine triad and other key residues, which adjusted their conformations to interact with sialic acid and facilitate the opening of a new cleft. Computational tools, including GROMACS and AutoDock, were pivotal in uncovering key insights into the catalytic mechanisms of GH33 sialidases offering a promising avenue for the rational design of improved biocatalysts.

Computational tools applied to the study of protein–glycan interactions

Molecular dynamics simulation has proven to be a powerful tool in understanding and elucidating the intricate dynamics of glycan interactions with biomolecules. This computational technique allows researchers to delve deep into the molecular-level details of how glycans bind to their respective target proteins, providing valuable insights into binding mechanisms, thermodynamics, and the overall stability of protein–glycan complexes.

MD methods have been extensively employed by different groups to explore glycan recognition by host receptors, including mammalian and bacterial proteins. For example, several studies have been published on the recognition of sialic acids by different classes of proteins. It is known that sialic acid plays an essential role in the modulation of immune response through the binding with sialic acid-binding immunoglobulin-like lectins (SIGLEC). In the study reported by Martin Frank et al. [125], MD were used to analyse the binding modes of several glycomimetics for Siglec-7 and describe their molecular interactions at atomic level. The conformational characteristics of both natural, unmodified, and synthetic, modified α-sialoside glycerol sidechains of sialic acid were investigated. The applied computational tools allowed to discover a new modification in the sialic acid glycerol chain that binds to Siglec-7, providing a basis for designing next-generation Siglec-7 ligands.

Sialic acid can also be recognised from some bacterial proteins which exploit this interaction to adhere to host cells during the first stages of infection. An example is given by the sialic acid-binding serine-rich repeat adhesins from *Streptococci,* which contain a sialic acid-binding region (SLBR) and are known as Siglec-like adhesins. Di Carluccio et al. [171] described the interactions between two different siglec-like adhesins with natural glycans by using a combination of NMR and MD simulations. This integrated approach allowed to accurately describe the different selectivity and flexibility of the proteins towards sialoglycans recognition and binding, providing a privileged starting point for the design and development of novel compounds to counteract streptococcal infections by inhibiting bacterial adherence to host tissues.

Another example of the application of MD simulations in studying bacterial proteins in the interaction with glycans comes from Bernardi's group [172], whose focus was on examining the interaction between different glycomimetic antagonists and BC2L-C lectin derived from *B. cenocepacia*. The MD results showed that the binding site at the interface of two BC2L-C-Nt monomers is pre-organised to host the bifunctional ligands. Additionally, the simulation with the water molecules highlights the importance of two of these molecules in the binding site, establishing an interaction network.

Bacterial glycoconjugates, as lipopolysaccharides-related systems, have also been dissected, gaining critical information about the ability of LPS to both stimulate the host immune system, mainly by interacting with TLR-4/MD-2 complex, and interact with several molecules. The Martin-Santamaria group [173] extensively contributed to increase the knowledge on this topic, analysing the conformational changes of the TLR4/MD2 complex when interacting either with small and LPS-like molecules

Notably, the comparison of free and bound state MD results can allow to determine critical differences in the glycan conformational behaviour upon binding with selected proteins, paving the way for the design of tailored synthetic inhibitors and therapeutics. As example, in the study of L. Pirone et al. [174], computational techniques were combined with biophysical and spectroscopic methods to investigate the interaction between a selenoglycoside (SeDG) and galectins Gal-1 or Gal-3CRD. The integration of data from NMR, CD, and ITC provided valuable insights into designing selective inhibitors. The computational studies uncovered two different binding modes: when bound to Gal-1, SeDG adopted a V-shaped conformation driven by van

der Waals interactions; on the contrary, when in complex with Gal-3CRD, it assumed an extended conformation. Comparing these modes identified specific interaction sites, guiding the design of selective inhibitors that can differentiate between the two galectins.

Noteworthy several computational studies have been conducted also for exploring protein-GAG interactions [175,176]. The study conducted by U. Uciechowska-Kaczmarzyk et al. [139] reports an extensive evaluation of protein–GAG complexes using a dataset of 28 complexes where the GAG length exceeded DP3 [139]. Through various statistical analyses to differentiate and highlight the docking programs with superior performance, valuable insights were provided into the most effective tools for studying these biologically relevant systems [177]. The interaction between the chemokine CXCL8/IL-8 and heparin-derived oligosaccharides was investigated by applying these docking procedures together with NMR spectroscopic techniques demonstrating the that higher affinity of the CXCL8 dimer for GAGs compared to the monomer and highlighting the structural plasticity that allows multiple binding modes. The use of HADDOCK in this context underscored its capability to model complex protein-GAG interactions accurately, providing a detailed understanding of the binding mechanisms at play [178].

Conclusion

In structural biology, the investigation of protein–glycan interactions often relies on applying various structural techniques, including NMR, X-ray crystallography, and cryo-EM. Each of these methodologies comes with distinct advantages and limitations. NMR is particularly valuable for its ability to dynamically study molecules at the atomic level while preserving sample integrity. This makes it especially suitable for studying carbohydrates, offering insights into their 3D structures and conformations, but it generates a huge amount of data, which can be challenging to interpret effectively. X-ray crystallography provides high-resolution structural information, but unfortunately, this technique often fails when investigating carbohydrate–protein interactions due to the intrinsic flexibility of sugars, rendering them invisible in the density maps. In recent years, cryo-EM has seen widespread adoption in solving protein structures and glycoconjugates, thanks to significant advancements in instrumentation. Nevertheless, a notable limitation of cryo-EM lies in its capacity to handle large, intricate complexes. In this context, computational approaches can be valuable allies to develop accurate models helping in integrating and rationalizing data obtained from different methods and bridging the gap between the insights obtained from experimental data and the detailed understanding of complex biological systems. As example, models of protein and ligand, both in the free and bound states, can assist not only the interpretation of NMR spectra but also the building of structures that satisfy experimentally derived distance and angle restraints. Moreover, in X-ray crystallography and cryo-EM, protein models can be used to provide accurate templates for molecular replacement in the crystal cell or for backbone tracing and fitting sequence into a map, respectively.

We provided here an overview of computational tools available for ligand and protein building as well as the analysis of their molecular interactions, with a special focus on carbohydrates (Figure 5). Generally, to allow the prediction of an accurate 3D model of protein–glycan complexes, the combined use of different tools is highly recommended. A typical workflow could include firstly research to investigate the favoured carbohydrates bound to a protein (i.e., by using Glynet or LectinOracle), then other tools (such as CASPIF or PESTO) can be employed to predict the binding location. Subsequently, appropriate docking software (i.e., AutoDock Vina-Carb) can be used to provide a model of protein–glycan complex, which can be further refined (as example thanks to GlycanDock) and explored by molecular dynamic simulations (i.e., by using AMBER). Finally, the detailed analysis of the trajectory (i.e., by using AmberTools) provides unique vision of the 3D structure and real dynamics of glycan motifs in the bound state. Notably, the recent fusion of cutting-edge technologies, such as virtual reality, with interactive molecular simulations also allows to create an immersive environment, offering an opportunity without precedents to explore and manipulate molecular systems in real-time [179].

However, it is worth to note that, despite the continuous improvement of computational techniques and force fields development, there are still some limitations in the application of bioinformatic methods to the analysis of biomolecular interactions, especially in the case of complex carbohydrates, not to speak about bacterial glycans. Step forwards have been done in the improvement of docking programs dedicated to carbohydrates, however, the available software performed better for smaller glycans, while additional glycosidic linkages still remain a big challenge for docking calculations, and there is still room for improvement in ranking the best sugar docking pose. Additionally, over the years, different carbohydrate-specific force fields have been developed, the choice of which varies depending on the preferred simulation conditions, however, only few parameters have been defined for peculiar bacterial monosaccharides hampering a user-friendly and not time-consuming analysis of the system via MD simulations. A step change in this direction will permit the integrated use of valuable bioinformatic tools tailored on carbohydrates and would be of great help in unveiling critical structural and

conformational features, at the atomic level, of complex glycans in the free and bound state, that can serve as essential resources for structural glycomics research to both experts and non-experts in glycobiology.

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement no 851356 to R.M.

The work described in this paper was supported by Italian MUR, PRIN 2020, Project SEA-WAVE 2020BKK3W9. FSE, PON Ricerca e Innovazione Azione I.1 ''Dottorati Innovativi con caratterizzazione Industriale'' is acknowledged for funding the Ph.D. grant to A.M.

Author Contributions

Ferran Nieto-Fabregat: conceptualization; writing – original draft; writing – review & editing. Maria Pia Lenza: conceptualization; writing – original draft; writing – review & editing. Angela Marseglia: writing – review & editing. Cristina Di Carluccio: writing – review & editing. Antonio Molinaro: writing – review & editing. Alba Silipo: writing – review & editing. Roberta Marchetti: conceptualization; writing – original draft; writing – review & editing.

ORCID® iDs

Ferran Nieto-Fabregat - https://orcid.org/0000-0001-9847-3030
Maria Pia Lenza - https://orcid.org/0000-0002-9733-5020
Angela Marseglia - https://orcid.org/0000-0003-1831-6831
Cristina Di Carluccio - https://orcid.org/0000-0001-5895-9829
Antonio Molinaro - https://orcid.org/0000-0002-3456-7369
Alba Silipo - https://orcid.org/0000-0002-5394-6532

Data Availability Statement

Data sharing is not applicable as no new data was generated or analyzed in this study.

References

1. Kuo, J. C.-H.; Paszek, M. J. *Annu. Rev. Cell Dev. Biol.* **2021,** *37,* 257–283. doi:10.1146/annurev-cellbio-120219-054401
2. Flynn, R. A.; Pedram, K.; Malaker, S. A.; Batista, P. J.; Smith, B. A. H.; Johnson, A. G.; George, B. M.; Majzoub, K.; Villalta, P. W.; Carette, J. E.; Bertozzi, C. R. *Cell* **2021,** *184,* 3109–3124.e22. doi:10.1016/j.cell.2021.04.023
3. Varki, A. *Glycobiology* **2017,** *27,* 3–49. doi:10.1093/glycob/cww086
4. Varki, A.; Kornfeld, S. Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Mohnen, D.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H., Eds.; Cold Spring Harbor: New York, NY, USA, 2022; pp 61–91.
5. Hart, G. W.; Copeland, R. J. *Cell* **2010,** *143,* 672–676. doi:10.1016/j.cell.2010.11.008
6. Rudd, P. M.; Karlsson, N. G.; Khoo, K.-H.; Thaysen-Andersen, M.; Wells, L.; Packer, N. H. Glycomics and Glycoproteomics. *Essentials of Glycobiology;* Cold Spring Harbor: New York, NY, USA, 2022; pp 1188–1213.
7. Marchetti, R.; Forgione, R. E.; Fabregat, F. N.; Di Carluccio, C.; Molinaro, A.; Silipo, A. *Curr. Opin. Struct. Biol.* **2021,** *68,* 74–83. doi:10.1016/j.sbi.2020.12.003
8. Sarkar, A.; Drouillard, S.; Rivet, A.; Perez, S. *Glycobiology* **2015,** *25,* 1480–1490. doi:10.1093/glycob/cwv054
9. Plazinski, W.; Plazinska, A. *Pure Appl. Chem.* **2017,** *89,* 1283–1294. doi:10.1515/pac-2016-0922
10. Fadda, E. *Curr. Opin. Chem. Biol.* **2022,** *69,* 102175. doi:10.1016/j.cbpa.2022.102175
11. Neelamegham, S.; Aoki-Kinoshita, K.; Bolton, E.; Frank, M.; Lisacek, F.; Lütteke, T.; O'Boyle, N.; Packer, N. H.; Stanley, P.; Toukach, P.; Varki, A.; Woods, R. J.; The SNFG Discussion Group. *Glycobiology* **2019,** *29,* 620–624. doi:10.1093/glycob/cwz045
12. Turnbull, J. E.; Field, R. A. *Nat. Chem. Biol.* **2007,** *3,* 74–77. doi:10.1038/nchembio0207-74
13. Hricovini, M. *Curr. Med. Chem.* **2004,** *11,* 2565–2583. doi:10.2174/0929867043364414
14. Frank, M. Computational Docking as a Tool for the Rational Design of Carbohydrate-Based Drugs. In *Carbohydrates as drugs;* Seeberger, P. H.; Rademacher, C., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp 53–72. doi:10.1007/7355_2014_42
15. Mallik, B.; Morikis, D. *Curr. Proteomics* **2006,** *3,* 259–270. doi:10.2174/157016406780655568
16. Kuttel, M. M.; Ravenscroft, N. The Role of Molecular Modeling in Predicting Carbohydrate Antigen Conformation and Understanding Vaccine Immunogenicity. *Carbohydrate-based vaccines: from concept to clinic;* American Chemical Society: Washington, DC, USA, 2018; pp 139–173. doi:10.1021/bk-2018-1290.ch007
17. Baum, L. G.; Cobb, B. A. *Glycobiology* **2017,** *27,* 619–624. doi:10.1093/glycob/cwx036
18. Wexler, A. G.; Goodman, A. L. *Nat. Microbiol.* **2017,** *2,* 17026. doi:10.1038/nmicrobiol.2017.26
19. Di Lorenzo, F.; Duda, K. A.; Lanzetta, R.; Silipo, A.; De Castro, C.; Molinaro, A. *Chem. Rev.* **2022,** *122,* 15767–15821. doi:10.1021/acs.chemrev.0c01321
20. Rauch, J.; Volinsky, N.; Romano, D.; Kolch, W. *Cell Commun. Signaling* **2011,** *9,* 23. doi:10.1186/1478-811x-9-23
21. Wang, Z.; Cole, P. A. *Methods Enzymol.* **2014,** *548,* 1–21. doi:10.1016/b978-0-12-397918-6.00001-x
22. Yang, N. J.; Hinner, M. J. Getting Across the Cell Membrane: An Overview for Small Molecules, Peptides, and Proteins. In *Site-Specific Protein Labeling. Methods in Molecular Biology;* Gautier, A.; Hinner, M., Eds.; Humana Press: New York, NY, USA, 2015; Vol. 1266, pp 29–53. doi:10.1007/978-1-4939-2272-7_3
23. Ford, K. G.; Souberbielle, B. E.; Darling, D.; Farzaneh, F. *Gene Ther.* **2001,** *8,* 1–4. doi:10.1038/sj.gt.3301383
24. Litvinov, R. I.; Weisel, J. W. *Matrix Biol.* **2017,** *60–61,* 110–123. doi:10.1016/j.matbio.2016.08.003
25. Pollard, T. D. *Cold Spring Harbor Perspect. Biol.* **2016,** *8,* a018226. doi:10.1101/cshperspect.a018226
26. Watford, M.; Wu, G. *Adv. Nutr.* **2018,** *9,* 651–653. doi:10.1093/advances/nmy027

27. Reily, C.; Stewart, T. J.; Renfrow, M. B.; Novak, J. *Nat. Rev. Nephrol.* **2019,** *15,* 346–366. doi:10.1038/s41581-019-0129-4
28. Schwarz, F.; Aebi, M. *Curr. Opin. Struct. Biol.* **2011,** *21,* 576–582. doi:10.1016/j.sbi.2011.08.005
29. Magalhães, A.; Duarte, H. O.; Reis, C. A. *Mol. Aspects Med.* **2021,** *79,* 100964. doi:10.1016/j.mam.2021.100964
30. Bloch, J. S.; John, A.; Mao, R.; Mukherjee, S.; Boilevin, J.; Irobalieva, R. N.; Darbre, T.; Scott, N. E.; Reymond, J.-L.; Kossiakoff, A. A.; Goddard-Borger, E. D.; Locher, K. P. *Nat. Chem. Biol.* **2023,** *19,* 575–584. doi:10.1038/s41589-022-01219-9
31. Noborn, F.; Nilsson, J.; Larson, G. *Matrix Biol.* **2022,** *111,* 289–306. doi:10.1016/j.matbio.2022.07.002
32. Fujimoto, Z.; Tateno, H.; Hirabayashi, J. Lectin Structures: Classification Based on the 3-D Structures. In *Lectins: Methods and Protocols;* Hirabayashi, J., Ed.; Springer: New York, NY, USA, 2014; pp 579–606. doi:10.1007/978-1-4939-1292-6_46
33. Varki, A.; Etzler, M. E.; Cummings, R. D.; Esko, J. D. *Glycobiology* **2009,** *4,* 28.
34. Napoletano, C.; Zizzari, I. G.; Rughetti, A.; Rahimi, H.; Irimura, T.; Clausen, H.; Wandall, H. H.; Belleudi, F.; Bellati, F.; Pierelli, L.; Frati, L.; Nuti, M. *Eur. J. Immunol.* **2012,** *42,* 936–945. doi:10.1002/eji.201142086
35. Temme, J. S.; Butler, D. L.; Gildersleeve, J. C. *Biochem. J.* **2021,** *478,* 1485–1509. doi:10.1042/bcj20200610
36. Gimeno, A.; Valverde, P.; Ardá, A.; Jiménez-Barbero, J. *Curr. Opin. Struct. Biol.* **2020,** *62,* 22–30. doi:10.1016/j.sbi.2019.11.004
37. Maldonado-Hernández, R.; Quesada, O.; González-Feliciano, J. A.; Baerga-Ortiz, A.; Lasalde-Dominicci, J. A. *Proteomics* **2024,** *24,* 2300151. doi:10.1002/pmic.202300151
38. Pandey, B.; S., S.; Chatterjee, A.; Mangala Prasad, V. *Proteins: Struct., Funct., Bioinf.* **2024,** in press. doi:10.1002/prot.26636
39. Fontana, C.; Widmalm, G. *Chem. Rev.* **2023,** *123,* 1040–1102. doi:10.1021/acs.chemrev.2c00580
40. Wilkinson, H.; Saldova, R. *J. Proteome Res.* **2020,** *19,* 3890–3905. doi:10.1021/acs.jproteome.0c00435
41. Angulo, J.; Zimmer, J.; Imberty, A.; Prestegard, J. In *Essentials of Glycobiology;* Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G. W.; Aebi, M.; Mohnen, D.; Kinoshita, T.; Packer, N. H.; Prestegard, J. H.; Schnaar, R. L.; Seeberger, P. H., Eds.; Cold Spring Harbor: New York, NY, USA, 2022; pp 61–91.
42. Homans, S. W.; Pastore, A.; Dwek, R. A.; Rademacher, T. W. *Biochemistry* **1987,** *26,* 6649–6655. doi:10.1021/bi00395a014
43. Dauchez, M.; Mazurier, J.; Montreuil, J.; Spik, G.; Vergoten, G. *Biochimie* **1992,** *74,* 63–74. doi:10.1016/0300-9084(92)90185-h
44. Sinha, S.; Tam, B.; Wang, S. M. *Membranes* **2022,** *12,* 844. doi:10.3390/membranes12090844
45. Varki, A.; Cummings, R. D.; Aebi, M.; Packer, N. H.; Seeberger, P. H.; Esko, J. D.; Stanley, P.; Hart, G.; Darvill, A.; Kinoshita, T.; Prestegard, J. J.; Schnaar, R. L.; Freeze, H. H.; Marth, J. D.; Bertozzi, C. R.; Etzler, M. E.; Frank, M.; Vliegenthart, J. F.; Lütteke, T.; Perez, S.; Bolton, E.; Rudd, P.; Paulson, J.; Kanehisa, M.; Toukach, P.; Aoki-Kinoshita, K. F.; Dell, A.; Narimatsu, H.; York, W.; Taniguchi, N.; Kornfeld, S. *Glycobiology* **2015,** *25,* 1323–1324. doi:10.1093/glycob/cwv091
46. Cousins, K. R. *J. Am. Chem. Soc.* **2005,** *127,* 4115–4116. doi:10.1021/ja0410237
47. Lal, K.; Bermeo, R.; Perez, S. *Beilstein J. Org. Chem.* **2020,** *16,* 2448–2468. doi:10.3762/bjoc.16.199
48. Perez, S.; Fadda, E.; Makshakova, O. Computational Modeling in Glycoscience. *Comprehensive Glycoscience,* 2nd ed.; Elsevier: Amsterdam, Netherlands, 2021; pp 374–404. doi:10.1016/b978-0-12-819475-1.00004-3
49. Danne, R.; Poojari, C.; Martinez-Seara, H.; Rissanen, S.; Lolicato, F.; Róg, T.; Vattulainen, I. *J. Chem. Inf. Model.* **2017,** *57,* 2401–2406. doi:10.1021/acs.jcim.7b00237
50. Woods Group (2005-2024;) GLYCAM Web.
51. Singh, A.; Montgomery, D.; Xue, X.; Foley, B. L.; Woods, R. J. *Glycobiology* **2019,** *29,* 515–518. doi:10.1093/glycob/cwz027
52. Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *J. Comput. Chem.* **2008,** *29,* 1859–1865. doi:10.1002/jcc.20945
53. Arroyuelo, A.; Vila, J. A.; Martin, O. A. *J. Comput.-Aided Mol. Des.* **2016,** *30,* 619–624. doi:10.1007/s10822-016-9944-x
54. Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996,** *14,* 33–38. doi:10.1016/0263-7855(96)00018-5
55. Malde, A. K.; Zuo, L.; Breeze, M.; Stroet, M.; Poger, D.; Nair, P. C.; Oostenbrink, C.; Mark, A. E. *J. Chem. Theory Comput.* **2011,** *7,* 4026–4037. doi:10.1021/ct200196m
56. *AMBER;* University of California: San Francisco, CA, USA, 2018.
57. Dickson, C. J.; Walker, R. C.; Gould, I. R. *J. Chem. Theory Comput.* **2022,** *18,* 1726–1736. doi:10.1021/acs.jctc.1c01217
58. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. *J. Chem. Theory Comput.* **2015,** *11,* 3696–3713. doi:10.1021/acs.jctc.5b00255
59. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983,** *79,* 926–935. doi:10.1063/1.445869
60. He, X.; Man, V. H.; Yang, W.; Lee, T.-S.; Wang, J. *J. Chem. Phys.* **2020,** *153,* 114502. doi:10.1063/5.0019056
61. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. *J. Comput. Chem.* **2008,** *29,* 622–655. doi:10.1002/jcc.20820
62. Guvench, O.; Mallajosyula, S. S.; Raman, E. P.; Hatcher, E.; Vanommeslaeghe, K.; Foster, T. J.; Jamison, F. W., II; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2011,** *7,* 3162–3180. doi:10.1021/ct200328p
63. Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. *SoftwareX* **2015,** *1–2,* 19–25. doi:10.1016/j.softx.2015.06.001
64. Guvench, O.; Hatcher, E.; Venable, R. M.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2009,** *5,* 2353–2370. doi:10.1021/ct900242e
65. Pol-Fachin, L.; Verli, H.; Lins, R. D. *J. Comput. Chem.* **2014,** *35,* 2087–2095. doi:10.1002/jcc.23721
66. Allinger, N. L.; Rahman, M.; Lii, J. H. *J. Am. Chem. Soc.* **1990,** *112,* 8293–8307. doi:10.1021/ja00179a012
67. Dauchez, M.; Derreumaux, P.; Lagant, P.; Vergoten, G. *J. Comput. Chem.* **1995,** *16,* 188–199. doi:10.1002/jcc.540160206
68. Hatcher, E.; Guvench, O.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2009,** *113,* 12466–12476. doi:10.1021/jp905496e
69. Woods, R. J.; Dwek, R. A.; Edge, C. J.; Fraser-Reid, B. *J. Phys. Chem.* **1995,** *99,* 3832–3846. doi:10.1021/j100011a061
70. Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F.-Y. *Nucleic Acids Res.* **2011,** *39* (Suppl. 2), W511–W517. doi:10.1093/nar/gkr288
71. Zheng, S.; Tang, Q.; He, J.; Du, S.; Xu, S.; Wang, C.; Xu, Y.; Lin, F. *J. Chem. Inf. Model.* **2016,** *56,* 811–818. doi:10.1021/acs.jcim.5b00687

72. Kamerling, J. P. Basics Concepts and Nomenclature Recommendations in Carbohydrate Chemistry. In *Comprehensive Glycoscience;* Kamerling, H., Ed.; Elsevier: Oxford, UK, 2007; Vol. 1, pp 1–38. doi:10.1016/b978-044451967-2/00001-5
73. Yu, Y.; Delbianco, M. *Chem. – Eur. J.* **2020,** *26,* 9814–9825. doi:10.1002/chem.202001370
74. Xu, B.; Unione, L.; Sardinha, J.; Wu, S.; Ethève-Quelquejeu, M.; Pilar Rauter, A.; Blériot, Y.; Zhang, Y.; Martín-Santamaría, S.; Díaz, D.; Jiménez-Barbero, J.; Sollogoub, M. *Angew. Chem.* **2014,** *126,* 9751–9756. doi:10.1002/ange.201405008
75. Toukach, P. V.; Egorova, K. S. *Nucleic Acids Res.* **2016,** *44,* D1229–D1236. doi:10.1093/nar/gkv840
76. Toukach, P. V.; Egorova, K. S. Bacterial, Plant, and Fungal Carbohydrate Structure Databases: Daily Usage. In *Glycoinformatics;* Lütteke, T.; Frank, M., Eds.; Methods in Molecular Biology, Vol. 1273; Humana Press: New York, NY, USA, 2015; pp 55–85. doi:10.1007/978-1-4939-2343-4_5
77. Egorova, K. S.; Toukach, P. V. Carbohydrate Structure Database (CSDB): Examples of Usage. In *A Practical Guide to Using Glycomics Databases;* Aoki-Kinoshita, K., Ed.; Springer: Tokyo, Japan, 2017; pp 75–113. doi:10.1007/978-4-431-56454-6_5
78. wwPDB consortium. *Nucleic Acids Res.* **2019,** *47,* D520–D528. doi:10.1093/nar/gky949
79. Cheng, H.; Abed, A. M.; Alizadeh, A.; Ghabra, A. A.; Altalbawy, F. M. A.; Sabetvand, R.; Smaisim, G. F.; Yadav, A.; Toghraie, D.; Riadi, Y. *J. Mol. Liq.* **2023,** *369,* 120893. doi:10.1016/j.molliq.2022.120893
80. Yu, Y.; Xu, S.; He, R.; Liang, G. *J. Agric. Food Chem.* **2023,** *71,* 2684–2703. doi:10.1021/acs.jafc.2c06789
81. Wang, Y.; Dong, B.; Wang, D.; Jia, X.; Zhang, Q.; Liu, W.; Zhou, H. *Appl. Sci.* **2023,** *13,* 7287. doi:10.3390/app13127287
82. Sabe, V. T.; Ntombela, T.; Jhamba, L. A.; Maguire, G. E. M.; Govender, T.; Naicker, T.; Kruger, H. G. *Eur. J. Med. Chem.* **2021,** *224,* 113705. doi:10.1016/j.ejmech.2021.113705
83. Lokhande, K. B.; Pawar, S. V.; Madkaiker, S.; Nawani, N.; Venkateswara, S. K.; Ghosh, P. *J. Biomol. Struct. Dyn.* **2023,** *41,* 2698–2712. doi:10.1080/07391102.2022.2038271
84. Hisama, K.; Valadez Huerta, G.; Koyama, M. *Comput. Mater. Sci.* **2023,** *218,* 111955. doi:10.1016/j.commatsci.2022.111955
85. Günay, M. G.; Kemerli, U.; Karaman, C.; Karaman, O.; Güngör, A.; Karimi-Maleh, H. *Environ. Res.* **2023,** *217,* 114785. doi:10.1016/j.envres.2022.114785
86. Garcia-Alai, M. M.; Heidemann, J.; Skruzny, M.; Gieras, A.; Mertens, H. D. T.; Svergun, D. I.; Kaksonen, M.; Uetrecht, C.; Meijers, R. *Nat. Commun.* **2018,** *9,* 328. doi:10.1038/s41467-017-02443-x
87. Mufassirin, M. M. M.; Newton, M. A. H.; Sattar, A. *Artif. Intell. Rev.* **2023,** *56,* 7665–7732. doi:10.1007/s10462-022-10350-x
88. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. *Nature* **2021,** *596,* 583–589. doi:10.1038/s41586-021-03819-2
89. Zhou, X.; Zheng, W.; Li, Y.; Pearce, R.; Zhang, C.; Bell, E. W.; Zhang, G.; Zhang, Y. *Nat. Protoc.* **2022,** *17,* 2326–2353. doi:10.1038/s41596-022-00728-0
90. Webb, B.; Sali, A. *Curr. Protoc. Bioinf.* **2016,** *54,* 5.6.1–5.6.37. doi:10.1002/cpbi.3
91. Leman, J. K.; Weitzner, B. D.; Lewis, S. M.; Adolf-Bryfogle, J.; Alam, N.; Alford, R.; Aprahamian, M.; Baker, D.; Barlow, K. A.; Barth, P.; Basanta, B.; Bender, B. J.; Blacklock, K.; Bonet, J.; Boyken, S. E.; Bradley, P.; Bystroff, C.; Conway, P.; Cooper, S.; Correia, B. E.; Coventry, B.; Das, R.; De Jong, R. M.; DiMaio, F.; Dsilva, L.; Dunbrack, R.; Ford, A. S.; Frenz, B.; Fu, D. Y.; Geniesse, C.; Goldschmidt, L.; Gowthaman, R.; Gray, J. J.; Gront, D.; Guffy, S.; Horowitz, S.; Huang, P.-S.; Huber, T.; Jacobs, T. M.; Jeliazkov, J. R.; Johnson, D. K.; Kappel, K.; Karanicolas, J.; Khakzad, H.; Khar, K. R.; Khare, S. D.; Khatib, F.; Khramushin, A.; King, I. C.; Kleffner, R.; Koepnick, B.; Kortemme, T.; Kuenze, G.; Kuhlman, B.; Kuroda, D.; Labonte, J. W.; Lai, J. K.; Lapidoth, G.; Leaver-Fay, A.; Lindert, S.; Linsky, T.; London, N.; Lubin, J. H.; Lyskov, S.; Maguire, J.; Malmström, L.; Marcos, E.; Marcu, O.; Marze, N. A.; Meiler, J.; Moretti, R.; Mulligan, V. K.; Nerli, S.; Norn, C.; Ó'Conchúir, S.; Ollikainen, N.; Ovchinnikov, S.; Pacella, M. S.; Pan, X.; Park, H.; Pavlovicz, R. E.; Pethe, M.; Pierce, B. G.; Pilla, K. B.; Raveh, B.; Renfrew, P. D.; Burman, S. S. R.; Rubenstein, A.; Sauer, M. F.; Scheck, A.; Schief, W.; Schueler-Furman, O.; Sedan, Y.; Sevy, A. M.; Sgourakis, N. G.; Shi, L.; Siegel, J. B.; Silva, D.-A.; Smith, S.; Song, Y.; Stein, A.; Szegedy, M.; Teets, F. D.; Thyme, S. B.; Wang, Y.-R.; Watkins, A.; Zimmerman, L.; Bonneau, R. *Methods* **2020,** *17,* 665–680. doi:10.1038/s41592-020-0848-2
92. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. *Science* **2021,** *373,* 871–876. doi:10.1126/science.abj8754
93. Moretti, R.; Lyskov, S.; Das, R.; Meiler, J.; Gray, J. J. *Protein Sci.* **2018,** *27,* 259–268. doi:10.1002/pro.3313
94. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. *Nucleic Acids Res.* **2018,** *46,* W296–W303. doi:10.1093/nar/gky427
95. Bonnardel, F.; Mariethoz, J.; Salentin, S.; Robin, X.; Schroeder, M.; Perez, S.; Lisacek, F.; Imberty, A. *Nucleic Acids Res.* **2019,** *47,* D1236–D1244. doi:10.1093/nar/gky832
96. Bonnardel, F.; Perez, S.; Lisacek, F.; Imberty, A. Structural Database for Lectins and the UniLectin Web Platform. In *Lectin Purification and Analysis: Methods and Protocols;* Hirabayashi, J., Ed.; Springer: New York, NY, USA, 2020; pp 1–14. doi:10.1007/978-1-0716-0430-4_1
97. Ives, C. M.; Singh, O.; D'Andrea, S.; Fogarty, C. A.; Harbison, A. M.; Satheesan, A.; Tropea, B.; Fadda, E. *bioRxiv* **2023**. doi:10.1101/2023.12.11.571101
98. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993,** *26,* 283–291. doi:10.1107/s0021889892009944
99. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. *Proteins: Struct., Funct., Bioinf.* **2023,** *91,* 1539–1549. doi:10.1002/prot.26617

100. Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; Jain, S.; Lewis, S. M.; Arendall, W. B., III; Snoeyink, J.; Adams, P. D.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. *Protein Sci.* **2018,** *27,* 293–315. doi:10.1002/pro.3330
101. Rodrigues, J. P. G. L. M.; Teixeira, J. M. C.; Trellet, M.; Bonvin, A. M. J. J. *F1000Research* **2018,** *7,* 1961. doi:10.12688/f1000research.17456.1
102. Martínez-Rosell, G.; Giorgino, T.; De Fabritiis, G. *J. Chem. Inf. Model.* **2017,** *57,* 1511–1516. doi:10.1021/acs.jcim.7b00190
103. Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. *Curr. Comput.-Aided Drug Des.* **2011,** *7,* 146–157. doi:10.2174/157340911795677602
104. Zhao, H.; Taherzadeh, G.; Zhou, Y.; Yang, Y. *Curr. Protoc. Protein Sci.* **2018,** *94,* e75. doi:10.1002/cpps.75
105. Bibekar, P.; Krapp, L.; Peraro, M. D. *J. Chem. Theory Comput.* **2024,** *20,* 2985–2991. doi:10.1021/acs.jctc.3c01145
106. Krapp, L. F.; Abriata, L. A.; Cortés Rodriguez, F.; Dal Peraro, M. *Nat. Commun.* **2023,** *14,* 2175. doi:10.1038/s41467-023-37701-8
107. Carpenter, E. J.; Seth, S.; Yue, N.; Greiner, R.; Derda, R. *Chem. Sci.* **2022,** *13,* 6669–6686. doi:10.1039/d1sc05681f
108. Lundstrøm, J.; Korhonen, E.; Lisacek, F.; Bojar, D. *Adv. Sci.* **2022,** *9,* 2103807. doi:10.1002/advs.202103807
109. Canner, S. W.; Shanker, S.; Gray, J. J. *Front. Bioinform.* **2023,** *3,* 1186531. doi:10.3389/fbinf.2023.1186531
110. Feinstein, W. P.; Brylinski, M. *J. Cheminf.* **2015,** *7,* 18. doi:10.1186/s13321-015-0067-5
111. Yang, C.; Chen, E. A.; Zhang, Y. *Molecules* **2022,** *27,* 4568. doi:10.3390/molecules27144568
112. Hassan, N. M.; Alhossary, A. A.; Mu, Y.; Kwoh, C.-K. *Sci. Rep.* **2017,** *7,* 15451. doi:10.1038/s41598-017-15571-7
113. Liu, Y.; Grimm, M.; Dai, W.-t.; Hou, M.-c.; Xiao, Z.-X.; Cao, Y. *Acta Pharmacol. Sin.* **2020,** *41,* 138–144. doi:10.1038/s41401-019-0228-6
114. Le Guilloux, V.; Schmidtke, P.; Tuffery, P. *BMC Bioinf.* **2009,** *10,* 168. doi:10.1186/1471-2105-10-168
115. Goodford, P. J. *J. Med. Chem.* **1985,** *28,* 849–857. doi:10.1021/jm00145a002
116. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. *J. Med. Chem.* **2006,** *49,* 6177–6196. doi:10.1021/jm051256o
117. *Schrödinger,* Release 2024-2; Schrödinger, LLC: New York, NY, USA, 2024.
118. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997,** *267,* 727–748. doi:10.1006/jmbi.1996.0897
119. *Molecular Operating Environment (MOE),* 2022.02; Chemical Computing Group ULC: 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7, Canada, 2023.
120. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. *J. Comput. Chem.* **2009,** *30,* 2785–2791. doi:10.1002/jcc.21256
121. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* **1998,** *19,* 1639–1662. doi:10.1002/(sici)1096-987x(19981115)19:14<1639::aid-jcc10>3.0.co;2-b
122. Trott, O.; Olson, A. J. *J. Comput. Chem.* **2010,** *31,* 455–461. doi:10.1002/jcc.21334
123. Gaudreault, F.; Najmanovich, R. J. *J. Chem. Inf. Model.* **2015,** *55,* 1323–1336. doi:10.1021/acs.jcim.5b00078
124. van Zundert, G. C. P.; Rodrigues, J. P. G. L. M.; Trellet, M.; Schmitz, C.; Kastritis, P. L.; Karaca, E.; Melquiond, A. S. J.; van Dijk, M.; de Vries, S. J.; Bonvin, A. M. J. J. *J. Mol. Biol.* **2016,** *428,* 720–725. doi:10.1016/j.jmb.2015.09.014
125. Frank, M.; Kuhfeldt, E.; Cramer, J.; Watzl, C.; Prescher, H. *J. Med. Chem.* **2023,** *66,* 14315–14334. doi:10.1021/acs.jmedchem.3c01349
126. Bender, B. J.; Gahbauer, S.; Luttens, A.; Lyu, J.; Webb, C. M.; Stein, R. M.; Fink, E. A.; Balius, T. E.; Carlsson, J.; Irwin, J. J.; Shoichet, B. K. *Nat. Protoc.* **2021,** *16,* 4799–4832. doi:10.1038/s41596-021-00597-z
127. Zhang, S.; Chen, K. Y.; Zou, X. *Commun. Inf. Syst.* **2021,** *21,* 147–163. doi:10.4310/cis.2021.v21.n1.a7
128. Agu, P. C.; Afiukwa, C. A.; Orji, O. U.; Ezeh, E. M.; Ofoke, I. H.; Ogbu, C. O.; Ugwuja, E. I.; Aja, P. M. *Sci. Rep.* **2023,** *13,* 13398. doi:10.1038/s41598-023-40160-2
129. Nivedha, A. K.; Thieker, D. F.; Makeneni, S.; Hu, H.; Woods, R. J. *J. Chem. Theory Comput.* **2016,** *12,* 892–901. doi:10.1021/acs.jctc.5b00834
130. Nivedha, A. K.; Makeneni, S.; Foley, B. L.; Tessier, M. B.; Woods, R. J. *J. Comput. Chem.* **2014,** *35,* 526–539. doi:10.1002/jcc.23517
131. Kerzmann, A.; Fuhrmann, J.; Kohlbacher, O.; Neumann, D. *J. Chem. Inf. Model.* **2008,** *48,* 1616–1625. doi:10.1021/ci800103u
132. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.* **1996,** *261,* 470–489. doi:10.1006/jmbi.1996.0477
133. Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. *J. Med. Chem.* **2006,** *49,* 5912–5931. doi:10.1021/jm050362n
134. Nance, M. L.; Labonte, J. W.; Adolf-Bryfogle, J.; Gray, J. J. *J. Phys. Chem. B* **2021,** *125,* 6807–6820. doi:10.1021/acs.jpcb.1c00910
135. Labonte, J. W.; Adolf-Bryfogle, J.; Schief, W. R.; Gray, J. J. *J. Comput. Chem.* **2017,** *38,* 276–287. doi:10.1002/jcc.24679
136. Boittier, E. D.; Burns, J. M.; Gandhi, N. S.; Ferro, V. *J. Chem. Inf. Model.* **2020,** *60,* 6328–6343. doi:10.1021/acs.jcim.0c00373
137. Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. *J. Comput.-Aided Mol. Des.* **2006,** *20,* 601–619. doi:10.1007/s10822-006-9060-4
138. de Vries, S. J.; Schindler, C. E. M.; Chauvot de Beauchêne, I.; Zacharias, M. *Biophys. J.* **2015,** *108,* 462–465. doi:10.1016/j.bpj.2014.12.015
139. Uciechowska-Kaczmarzyk, U.; Chauvot de Beauchene, I.; Samsonov, S. A. *J. Mol. Graphics Mod.* **2019,** *90,* 42–50. doi:10.1016/j.jmgm.2019.04.001
140. Ballester, P. J.; Mitchell, J. B. O. *Bioinformatics* **2010,** *26,* 1169–1175. doi:10.1093/bioinformatics/btq112
141. Pires, D. E. V.; Ascher, D. B. *Nucleic Acids Res.* **2016,** *44,* W557–W561. doi:10.1093/nar/gkw390
142. Li, H.; Sze, K.-H.; Lu, G.; Ballester, P. J. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021,** *11,* e1478. doi:10.1002/wcms.1478
143. Chalmers, G.; Glushka, J. N.; Foley, B. L.; Woods, R. J.; Prestegard, J. H. *J. Magn. Reson.* **2016,** *265,* 1–9. doi:10.1016/j.jmr.2016.01.006
144. Nguyen, T. B.; Pires, D. E. V.; Ascher, D. B. *Briefings Bioinf.* **2022,** *23,* bbab512. doi:10.1093/bib/bbab512

145. Kühne, T. D.; Iannuzzi, M.; Del Ben, M.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F.; Golze, D.; Wilhelm, J.; Chulkov, S.; Bani-Hashemian, M. H.; Weber, V.; Borštnik, U.; Taillefumier, M.; Jakobovits, A. S.; Lazzaro, A.; Pabst, H.; Müller, T.; Schade, R.; Guidon, M.; Andermatt, S.; Holmberg, N.; Schenter, G. K.; Hehn, A.; Bussy, A.; Belleflamme, F.; Tabacchi, G.; Glöß, A.; Lass, M.; Bethune, I.; Mundy, C. J.; Plessl, C.; Watkins, M.; VandeVondele, J.; Krack, M.; Hutter, J. *J. Chem. Phys.* **2020,** *152*. doi:10.1063/5.0007045
146. Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing,* 2006; pp 84–ee. doi:10.1145/1188455.1188544
147. Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. *Comput. Phys. Commun.* **2022,** *271,* 108171. doi:10.1016/j.cpc.2021.108171
148. Lagardère, L.; Jolly, L.-H.; Lipparini, F.; Aviat, F.; Stamm, B.; Jing, Z. F.; Harger, M.; Torabifard, H.; Cisneros, G. A.; Schnieders, M. J.; Gresh, N.; Maday, Y.; Ren, P. Y.; Ponder, J. W.; Piquemal, J.-P. *Chem. Sci.* **2018,** *9,* 956–972. doi:10.1039/c7sc04531j
149. Bowers, K.; Chow, E.; Xu, H.; Dror, R.; Eastwood, M.; Gregersen, B.; Klepeis, J.; Kolossváry, I.; Moraes, M.; Sacerdoti, F.; Salmon, J.; Shan, Y.; Shaw, D. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *SC '06: Proceedings of the 2006 ACM/IEEE conference on Supercomputing,* Association for Computing Machinery: New York, NY, USA, 2006. doi:10.1145/1188455.1188544
150. Kalé, L. V.; Bhatele, A.; Bohm, E. J.; Phillips, J. C. NAMD (NAnoscale Molecular Dynamics). In *Encyclopedia of Parallel Computing;* Padua, D., Ed.; Springer: Boston, MA, USA, 2011; pp 1249–1254. doi:10.1007/978-0-387-09766-4_505
151. Du, X.; Li, Y.; Xia, Y.-L.; Ai, S.-M.; Liang, J.; Sang, P.; Ji, X.-L.; Liu, S.-Q. *Int. J. Mol. Sci.* **2016,** *17,* 144. doi:10.3390/ijms17020144
152. Roe, D. R.; Cheatham, T. E., III. *J. Chem. Theory Comput.* **2013,** *9,* 3084–3095. doi:10.1021/ct400341p
153. The PLUMED consortium. *Nat. Methods* **2019,** *16,* 670–673. doi:10.1038/s41592-019-0506-8
154. Kuprov, I.; Morris, L. C.; Glushka, J. N.; Prestegard, J. H. *J. Magn. Reson.* **2021,** *323,* 106891. doi:10.1016/j.jmr.2020.106891
155. Nieto-Fabregat, F.; Zhu, Q.; Vivès, C.; Zhang, Y.; Marseglia, A.; Chiodo, F.; Thépaut, M.; Rai, D.; Kulkarni, S. S.; Di Lorenzo, F.; Molinaro, A.; Marchetti, R.; Fieschi, F.; Xiao, G.; Yu, B.; Silipo, A. *JACS Au* **2024,** *4,* 697–712. doi:10.1021/jacsau.3c00748
156. Guo, Y.; Feinberg, H.; Conroy, E.; Mitchell, D. A.; Alvarez, R.; Blixt, O.; Taylor, M. E.; Weis, W. I.; Drickamer, K. *Nat. Struct. Mol. Biol.* **2004,** *11,* 591–598. doi:10.1038/nsmb784
157. *The PyMOL Molecular Graphics System,* Version~1.8; Schrödinger LLC: New York, NY, USA, 2015.
158. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004,** *25,* 1605–1612. doi:10.1002/jcc.20084
159. *Design, L.;* BIOVIA: California, 2014.
160. *Schrödinger Release 2022-3,* 2021; Schrödinger, LLC: New York, NY, USA.
161. *SAMSON: Software for Adaptive Modeling and Simulation of Nanosystems;* Samson: Grenoble, France, 2016.
162. Fogarty, C. A.; Fadda, E. *J. Phys. Chem. B* **2021,** *125,* 2607–2616. doi:10.1021/acs.jpcb.1c00304
163. Harbison, A. M.; Brosnan, L. P.; Fenlon, K.; Fadda, E. *Glycobiology* **2019,** *29,* 94–103. doi:10.1093/glycob/cwy097
164. Vanacore, A.; Vitiello, G.; Wanke, A.; Cavasso, D.; Clifton, L. A.; Mahdi, L.; Campanero-Rhodes, M. A.; Solís, D.; Wuhrer, M.; Nicolardi, S.; Molinaro, A.; Marchetti, R.; Zuccaro, A.; Paduano, L.; Silipo, A. *Carbohydr. Polym.* **2022,** *277,* 118839. doi:10.1016/j.carbpol.2021.118839
165. Di Lorenzo, F.; Nicolardi, S.; Marchetti, R.; Vanacore, A.; Gallucci, N.; Duda, K.; Nieto Fabregat, F.; Nguyen, H. N. A.; Gully, D.; Saenz, J.; Giraud, E.; Paduano, L.; Molinaro, A.; D'Errico, G.; Silipo, A. *JACS Au* **2023,** *3,* 929–942. doi:10.1021/jacsau.3c00025
166. Makshakova, O.; Zykwinska, A.; Cuenot, S.; Colliec-Jouault, S.; Perez, S. *Carbohydr. Polym.* **2022,** *276,* 118732. doi:10.1016/j.carbpol.2021.118732
167. Lei, M.; Huang, W.; Jin, Z.; Sun, J.; Zhang, M.; Zhao, S. *Carbohydr. Polym.* **2022,** *297,* 119993. doi:10.1016/j.carbpol.2022.119993
168. Harbison, A. M.; Fogarty, C. A.; Phung, T. K.; Satheesan, A.; Schulz, B. L.; Fadda, E. *Chem. Sci.* **2022,** *13,* 386–395. doi:10.1039/d1sc04832e
169. Mou, Y.; Huang, P.-S.; Thomas, L. M.; Mayo, S. L. *J. Mol. Biol.* **2015,** *427,* 2697–2706. doi:10.1016/j.jmb.2015.06.006
170. Cao, X.; Yang, X.; Xiao, M.; Jiang, X. *Int. J. Mol. Sci.* **2023,** *24,* 6830. doi:10.3390/ijms24076830
171. Di Carluccio, C.; Forgione, R. E.; Bosso, A.; Yokoyama, S.; Manabe, Y.; Pizzo, E.; Molinaro, A.; Fukase, K.; Fragai, M.; Bensing, B. A.; Marchetti, R.; Silipo, A. *RSC Chem. Biol.* **2021,** *2,* 1618–1630. doi:10.1039/d1cb00173f
172. Antonini, G.; Civera, M.; Lal, K.; Mazzotta, S.; Varrot, A.; Bernardi, A.; Belvisi, L. *Front. Mol. Biosci.* **2023,** *10,* 1201630. doi:10.3389/fmolb.2023.1201630
173. Matamoros-Recio, A.; Franco-Gonzalez, J. F.; Perez-Regidor, L.; Billod, J.-M.; Guzman-Caldentey, J.; Martin-Santamaria, S. *Chem. – Eur. J.* **2021,** *27,* 15406–15425. doi:10.1002/chem.202102995
174. Pirone, L.; Nieto-Fabregat, F.; Di Gaetano, S.; Capasso, D.; Russo, R.; Traboni, S.; Molinaro, A.; Iadonisi, A.; Saviano, M.; Marchetti, R.; Silipo, A.; Pedone, E. *Int. J. Mol. Sci.* **2022,** *23,* 8273. doi:10.3390/ijms23158273
175. Zhang, S.; Trinh, C.; Schweitzer-Stenner, R.; Urbanc, B. *Biophys. J.* **2019,** *116,* 61a. doi:10.1016/j.bpj.2018.11.374
176. Köhling, S.; Blaszkiewicz, J.; Ruiz-Gómez, G.; Fernández-Bachiller, M. I.; Lemmnitzer, K.; Panitz, N.; Beck-Sickinger, A. G.; Schiller, J.; Pisabarro, M. T.; Rademann, J. *Chem. Sci.* **2019,** *10,* 866–878. doi:10.1039/c8sc03649g
177. Sankaranarayanan, N. V.; Nagarajan, B.; Desai, U. R. *Curr. Opin. Struct. Biol.* **2018,** *50,* 91–100. doi:10.1016/j.sbi.2017.12.004
178. Joseph, P. R. B.; Mosier, P. D.; Desai, U. R.; Rajarathnam, K. *Biochem. J.* **2015,** *472,* 121–133. doi:10.1042/bj20150059
179. Perez, S.; Makshakova, O.; Angulo, J.; Bedini, E.; Bisio, A.; de Paz, J. L.; Fadda, E.; Guerrini, M.; Hricovini, M.; Hricovini, M.; Lisacek, F.; Nieto, P. M.; Pagel, K.; Paiardi, G.; Richter, R.; Samsonov, S. A.; Vivès, R. R.; Nikitovic, D.; Ricard Blum, S. *JACS Au* **2023,** *3,* 628–656. doi:10.1021/jacsau.2c00569