

This is a repository copy of *Functional implications of glycans and their curation:insights from the workshop held at the 16th Annual International Biocuration Conference in Padua, Italy*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/216325/>

Version: Published Version

Article:

Martinez, Karina, Agirre, Jon orcid.org/0000-0002-1086-0253, Akune, Yukie et al. (33 more authors) (2024) Functional implications of glycans and their curation:insights from the workshop held at the 16th Annual International Biocuration Conference in Padua, Italy. Database. baae073. ISSN 1758-0463

<https://doi.org/10.1093/database/baae073>

Reuse





































This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Functional implications of glycans and their curation: insights from the workshop held at the 16th Annual International Biocuration Conference in Padua, Italy

Karina Martinez ¹, Jon Agirre ², Yukie Akune ³, Kiyoko F. Aoki-Kinoshita ⁴, Cecilia Arighi ⁵, Kristian B. Axelsen ⁶, Evan Bolton ⁷, Emily Bordeleau ⁸, Nathan J. Edwards ⁹, Elisa Fadda ¹⁰, Ten Feizi ³, Catherine Hayes ¹¹, Callum M. Ives ¹⁰, Hiren J. Joshi ¹², Khakurel Krishna Prasad ¹³, Sofia Kossida ¹⁴, Frederique Lisacek ¹¹, Yan Liu ³, Thomas Lütteke ¹⁵, Junfeng Ma ¹⁶, Adnan Malik ¹⁷, Maria Martin ¹⁷, Akul Y. Mehta ¹⁸, Sriram Neelamegham ¹⁹, Kalpana Panneerselvam ¹⁷, René Ranzinger ²⁰, Sylvie Ricard-Blum ²¹, Gaoussou Sanou ¹⁴, Vijay Shanker ⁵, Paul D. Thomas ²², Michael Tiemeyer ²⁰, James Urban ²³, Randi Vita ²⁴, Jeet Vora ¹, Yasunori Yamamoto ²⁵, Raja Mazumder ^{1,*}

¹Department of Biochemistry & Molecular Medicine, The George Washington University School of Medicine and Health Sciences, 2300 I St. NW, Washington, DC 20052, United States

²York Structural Biology Laboratory, Department of Chemistry, University of York, Wentworth Way, York YO10 5DD, United Kingdom

³The Glycosciences Laboratory, Imperial College London, Hammersmith Campus, Du Cane Road, London W12 0NN, United Kingdom

⁴Glycan and Life Systems Integration Center (GaLSIC), Soka University, 1-236 Tangi-machi, Hachioji, Tokyo 192-8577, Japan

⁵Department of Computer and Information Sciences, University of Delaware, 18 Amstel Ave, Newark, DE 19716, United States

⁶Swiss-Prot Group, Swiss Institute of Bioinformatics (SIB), CMU, 1 rue Michel Servet, Geneva 4 1211, Switzerland

⁷National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, United States

⁸Michael Smith Laboratories, The University of British Columbia, 2185 East Mall, Vancouver, British Columbia V6T 1Z4, Canada

⁹Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, 2115 Wisconsin Ave NW, Washington, DC 20007, United States

¹⁰Department of Chemistry and Hamilton Institute, Maynooth University, Kilcock Road, Maynooth, Co. Kildare W23 AH3Y, Ireland

¹¹Proteome Informatics Group, Swiss Institute of Bioinformatics (SIB), route de Drize 7, Geneva CH-1227, Switzerland

¹²Copenhagen Center for Glycomics, Department of Cellular and Molecular Medicine, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3, Copenhagen DK-2200, Denmark

¹³ELI Beamlines Facility, The Extreme Light Infrastructure ERIC, Za Radnicí 835, Dolní Břežany 25241, Czech Republic

¹⁴IMGT, The International ImMunoGeneTics Information System, National Center for Scientific Research (CNRS), Institute of Human Genetics (IGH), University of Montpellier (UM), 141 rue de la Cardonille, Montpellier 34 090, France

¹⁵Institute of Veterinary Physiology and Biochemistry, Justus-Liebig-University Gießen, Frankfurter Str. 100, Gießen 35392, Germany

¹⁶Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, 3900 Reservoir Road NW, Washington, DC 20007, United States

¹⁷European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

¹⁸Department of Surgery, Beth Israel Deaconess Medical Center, National Center for Functional Glycomics, Harvard Medical School, 330 Brookline Avenue, Boston, MA 02215, United States

¹⁹Departments of Chemical & Biological Engineering, Biomedical Engineering and Medicine, University at Buffalo, State University of New York, 906 Furnas Hall, Buffalo, NY 14260, United States

²⁰Complex Carbohydrate Research Center, University of Georgia, 315 Riverbend Rd, Athens, GA 30602, United States

²¹Institute of Molecular and Supramolecular Chemistry and Biochemistry (ICBMS), UMR 5246, University Lyon 1, CNRS, 43 Boulevard du 11 novembre 1918, Villeurbanne cedex F-69622, France

²²Department of Population and Public Health Sciences, University of Southern California, 2001 N Soto Street, Los Angeles, CA 90032, United States

Received 1 April 2024; Revised 24 June 2024; Accepted 10 July 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

²³Department of Chemistry and Molecular Biology, University of Gothenburg, Medicinaregatan 7 B, Gothenburg 41390, Sweden

²⁴Immune Epitope Database and Analysis Project, La Jolla Institute for Allergy & Immunology, 9420 Athena Circle, La Jolla, CA 92037, United States

²⁵Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 178-4-4 Wakashiba, Kashiwa, Chiba 277-0871, Japan

*Corresponding author. Department of Biochemistry and Molecular Medicine, School of Medicine & Health Sciences, The George Washington University, Ross Hall, Room 540 2300 I Street N.W., Washington, DC 20037, United States. E-mail: mazumder@gwu.edu

Citation details: Martinez, K., Agirre, J., Akune, Y. *et al.* Functional implications of glycans and their curation: insights from the workshop held at the 16th Annual International Biocuration Conference in Padua, Italy. *Database* (2024) Vol. 2024: article ID baae073; DOI: <https://doi.org/10.1093/database/baae073>

Abstract

Dynamic changes in protein glycosylation impact human health and disease progression. However, current resources that capture disease and phenotype information focus primarily on the macromolecules within the central dogma of molecular biology (DNA, RNA, proteins). To gain a better understanding of organisms, there is a need to capture the functional impact of glycans and glycosylation on biological processes. A workshop titled “Functional impact of glycans and their curation” was held in conjunction with the 16th Annual International Biocuration Conference to discuss ongoing worldwide activities related to glycan function curation. This workshop brought together subject matter experts, tool developers, and biocurators from over 20 projects and bioinformatics resources. Participants discussed four key topics for each of their resources: (i) how they curate glycan function-related data from publications and other sources, (ii) what type of data they would like to acquire, (iii) what data they currently have, and (iv) what standards they use. Their answers contributed input that provided a comprehensive overview of state-of-the-art glycan function curation and annotations. This report summarizes the outcome of discussions, including potential solutions and areas where curators, data wranglers, and text mining experts can collaborate to address current gaps in glycan and glycosylation annotations, leveraging each other’s work to improve their respective resources and encourage impactful data sharing among resources.

Database URL: https://wiki.glygen.org/Glycan_Function_Workshop_2023

Introduction

Biocuration has been foundational and an integral part of bioinformatics since Margaret O. Dayhoff demonstrated how collecting, analyzing, and annotating protein sequences can lead to a better understanding of the nature and function of the protein [1]. She and her co-worker’s effort also demonstrated the value of collection, standardization, and curation in an era that had already started producing significant amounts of data from experimental approaches to the study of life and function of biomolecules [2]. Defining the “function” of biomolecules, the building blocks of life, such as proteins, nucleic acids, lipids, and carbohydrates, has been elusive due to their contextual nature. Often biomolecules exist as conjugates or they interact with each other resulting in processes that are tightly coupled with the conditions they are exposed to. Therefore, it requires a collaborative effort from all spheres of biocuration to come together and discuss how glycan function is perceived within each knowledge domain. Such discussions can lead to common standards and definitions, which can accelerate our ability to leverage the different aspects of biocuration activities worldwide and can help us to better represent glycan function in databases and knowledgebases.

Glycan function and how it relates to protein and gene function

The central dogma of molecular biology elegantly defined the relationship between gene, transcript, and protein as a highly regulated, but largely linear process that decodes template molecules into functional products. Many post-translational protein modifications lie outside of this dogma because they are not encoded by a template and usually reflect the convergence of various signaling pathways or other cellular responses to tissue microenvironments, disease, or developmental progression. The glycosylation process,

defined as the synthesis, conjugation, and remodeling of glycans attached to other molecules (primarily proteins and lipids [3, 4], but recently also on RNA [5]), is a form of such nontemplate-driven biomolecule production. The glycans attached to proteins frequently modulate their folding, intrinsic stability, activity, and lifetime in circulation or at the cell surface [6]. They also influence viral- and microbial-host interactions, immune responses to pathogens, as well as pro- and anti-inflammatory status [7]. Detailed studies of individual glycans, their recognition by receptors, and their influence on carrier proteins can provide clues to their function. For each of these functions, the fine structure of the glycan itself and the specific amino acid residue to which the glycan is attached (the glycosylation site) combine to achieve observable biological responses. Expanding knowledge of glycosylation pathways, coupled with rapid advances in analytic technology over the past decade, have made it ever easier to describe the range of glycans expressed by specific cells or tissues and the diversity of glycans linked to specific glycosylation sites on identified proteins. These datasets have provided the raw material to develop knowledgebases and new opportunities to map glycosylation data to other types of data that together have vast potential to contribute to new understandings of biological functions.

Challenges in curating glycan function

Many resources include glycan-related genes and proteins, such as biosynthetic and degradative enzymes, within their scope. However, function and disease annotations are usually associated with the gene or gene product rather than with glycosylation or glycan structural changes that might characterize cellular or tissue states. Despite the immense volume of glycoproteomic and glycomic data generated over the past decades, glycan function annotations in databases and knowledgebases have lagged behind by fundamental challenges in data communication, standards, and curation.

Data communication

Bench scientists are primarily trained to use their laboratory methods to answer specific questions. Their main objectives are to generate data and publish their findings based on accepted practices within their fields. However, these practices are designed to prioritize data presentation for publication and not to facilitate appropriate cataloging for effective curation. Often, information required and expected by curators is not immediately at hand or it takes significant effort for scientists to re-catalog data into a format accessible by curators. In addition, data tables in publications can be prone to errors associated with manual entry, and sometimes the author's assertions are not verifiable based on the data presented in the manuscript.

Standards

Existing standard identifiers provided by various sources (e.g. GlyTouCan [8], ChEBI [9], PubChem [10]) represent glycans, but the adoption of these IDs by researchers has been slow. Currently, no journal requires the use of any of these standard glycan IDs. Some journals, including *Glycobiology*, recommend authors to follow the MIRAGE (Minimum Information Required for A Glycomics Experiment) guidelines [11] (<https://www.beilstein-institut.de/en/projects/mirage/guidelines/>). MIRAGE also recommends various glycan-related databases such as GlyTouCan, GlycoPOST, and UniCarb-DR as repositories to obtain accession numbers for glycans and analytical data (<https://www.beilstein-institut.de/en/projects/mirage/recommendations/>). Beyond standard glycan identifiers, several formats exist to describe glycan sequences, making it difficult for curators to map published sequence information to database IDs. In publications, human-readable International Union of Pure and Applied Chemistry (IUPAC) formats or the Symbol Nomenclature for Glycans (SNFG) graphics predominate [12], whereas databases store glycans using machine-readable formats such as GlycoCT [13] or WURCS [14, 15]. While converters have been developed (GlycanFormatConverter 2.7.0 <https://glyconavi.org/Tools/tool/gfc.php>, MolWURCS <https://github.com/glycoinfo/Executable/tree/master/MolWURCS>), they might not be compatible with all file formats and interconversion between formats used in publications to those applied by databases is often difficult. For instance, SNFG is quickly learned and easily read by humans, but difficult to parse programmatically because of its graphical nature. While IUPAC notation can be displayed as graphics or text, even the textual representations frequently fail to explicitly state information on absolute configuration, ring type, and anomericity. Thus, structural details have to be added by curators, who often need to infer the context to correctly assign these data. This need for additional interpretation also applies to SNFG representations, where anomers or linkage positions are not always explicitly specified but mentioned in the text.

Curation

Although funding bodies and journals have broadly encouraged scientists to deposit all types of data into public repositories, curation always requires significant effort and care. Curating glycan data takes even more effort due to the variable nature of the primary data and the intrinsic complexity of glycan structures. Coordination among resources

to curate biomolecules within the context of glycan function has been lacking, especially in comparison to the coordination between resources specialized for curating gene and protein functions. This disparity may not be broadly appreciated but needs to be addressed in order to help build robust connections between glycoscience data and more mature bioinformatic domains. Resources and funds are not readily available to support such efforts, especially to curate legacy data, resulting in an environment where curation (or even presentation of data in a curatable manner) is not prioritized. Ideally, publication should not be the only incentive to curate and deposit data, as there is a considerable amount of data that never gets published or deposited in an accessible data resource (e.g. data orphaned from abandoned projects or data supporting negative hypotheses).

To address these challenges and improve glycan functional annotations and standards, a whole-day workshop on the functional impact of glycans and their curation was held in conjunction with the 16th Annual International Biocuration Conference in Padua, Italy, on 23 April 2023.

Workshop summary

Bioinformaticians, curators, and glycobiologists from over 20 projects and bioinformatics resources that contain glycan-related data participated in presentations and discussions (Table 1 and Supplementary File 1). Participants were invited to share their perspectives, curation workflows, bottlenecks, and needs for extracting and disseminating knowledge about glycans and their function. With a focus on biocuration, data sharing, and data harmonization, the workshop was organized around four questions (Fig. 1), each presented here along with a summary of the discussion, findings, and proposals for future improvements.

Q1. How would your resource extract glycan-related data from literature or other sources?

Advances in experimental techniques and growth in the field of glycobiology have led to an influx of scientific publications and data related to glycans and glycosylation over the past decade. As described in the previous section, several challenges exist in extracting and curating this data. To gain consensus and better understand how resources handle these challenges, the first session addressed how resources extract glycan-related data from the literature and other sources. Participants focused on identifying strategies for data extraction, needs and bottlenecks, curation workflow, common solutions, and areas of improvement.

Curation

In most cases, information extraction is performed by expert human curators who read the publication and extract relevant data. Literature and text mining tools are often used to assist in identifying relevant publications, such as LitSuggest [16], PubTator [17], and those created by the resources themselves.

Rhea [18] and collaborators at US National Center for Biotechnology Information (NCBI) are developing an expert curated corpus to train and benchmark methods for extracting enzyme functions from text. This corpus was curated using the TeamTat framework [19] and using procedures developed

Table 1. Resources and projects represented at the workshop

Resource/project	Resource type	URL	Latest publication PMID/URL	Participant/s
BioCreative	Standard	https://biocreative.bioinformatics.udel.edu/	36197453	Cecilia Arighi
ChEBI	Knowledgebase, Standard	https://www.ebi.ac.uk/chebi/	26467479	Adnan Malik
Glyco.me	Knowledgebase	https://glyco.me/	29267884	Hiren Joshi
GlycodomainViewer	Tool	https://glycoproteome.expasy.org/glycomics-expasy/	30097532	Frederique Lisacek
Glyco@Expasy	Tool	https://github.com/neel-lab/GlycoEnzOnto	36282863	Ted Groth, Sriram Neelamegham (remote)
GlycoEnzOnto	Standard	http://glycomotif.glycomics.org/	N/A	Nathan Edwards
Glycomotif	Knowledgebase	http://sandbox.glycomics.org/	30574787	Frederique Lisacek
GlyGen Sandbox	Knowledgebase	https://glyconnect.expasy.org/	30357361	Thomas Lütteke (remote)
GlyConnect	Knowledgebase	www.glycosciences.de		
MonosaccharideDB	Knowledgebase	www.monosaccharidedb.org		
GlycoShape	Knowledgebase	https://glycoshape.org/	https://doi.org/10.1101/2023.12.11.571101	Callum Ives, Elisa Fadda (remote)
GlyCosmos	Knowledgebase	https://glycosmos.org	32572234	Kiyoko F. Aoki-Kinoshita, Yasunori Yamamoto
Glycowork	Knowledgebase	https://github.com/BojarLab/glycowork/	34192308	James Urban
GlyGen	Knowledgebase	https://www.glygen.org	31616925	Karina Martinez, René Ranzinger, Vijay Shanker, Nathan Edwards, Michael Tiemeyer, Raja Mazumder, Jeet Vora (remote)
GNOME	Standard	http://gnome.glycomics.org/	https://ceur-ws.org/Vol-3073/paper11.pdf	Nathan Edwards
GO	Standard	https://geneontology.org/	36866529	Paul Thomas
ICL Glycan Microarrays	Knowledgebase	https://www.imperial.ac.uk/glycosciences/ https://glycosciences.med.ic.ac.uk/glycanLibraryIndex.html	N/A	Ten Feizi, Yan Liu, Yukie Akune
IMGT	Knowledgebase	https://www.imgt.org/	34875068	Taciana Manso (remote), Sofia Kossida (remote)
Immune Epitope Database and Analysis Project (IEDB)	Knowledgebase	https://www.iedb.org/	30357391	Randi Vita
IntAct	Knowledgebase	https://www.ebi.ac.uk/intact	34761267	Kalpna Panneerselvam
iPTMnet	Knowledgebase	https://research.bioinformatics.udel.edu/iptmnet/	29145615	Cecilia Arighi
MatrixDB	Knowledgebase	http://matrixdb.univ-lyon1.fr/	30371822	Sylvie Ricard-Blum (post-workshop)
NCFG—Glybrary	Tool	https://www.glybrary.com	N/A	Akul Mehta (remote)
O-GlcNAcAtlas OGT-PIN	Knowledgebase	https://oglnac.org/atlas/ https://oglnac.org/ogt-pin	33442735	Junfeng Ma (remote)
PRIDE	Archive	https://www.ebi.ac.uk/pride/	34723319	Deepti Jaiswal Kundu
Privateer	Tool, Knowledgebase	https://github.com/glycojones/privateer , https://privateer.york.ac.uk/	26581513, 38711584, 38265073	Jon Agirre (remote)
PubChem	Archive, Knowledgebase	https://pubchem.ncbi.nlm.nih.gov	36305812	Evan Bolton (remote)
Rhea	Knowledgebase	https://www.rhea-db.org/	34755880	Kristian Axelsen, Alan Bridge (remote)
SNFG	Standard	https://www.ncbi.nlm.nih.gov/glycans/snfg.html	31184695	Sriram Neelamegham (remote)
TogoID	Tool	https://togoid.dbcls.jp/	35801937	Yasunori Yamamoto
UniLectin	Knowledgebase	https://unilectin.unige.ch/	33174598	Frederique Lisacek, Anne Imberty
UniProtKB	Knowledgebase	https://www.uniprot.org	36408920	Maria Martin, Cecilia Arighi, Kristian Axelsen, Alan Bridge (remote)

N/A indicates data are not available.

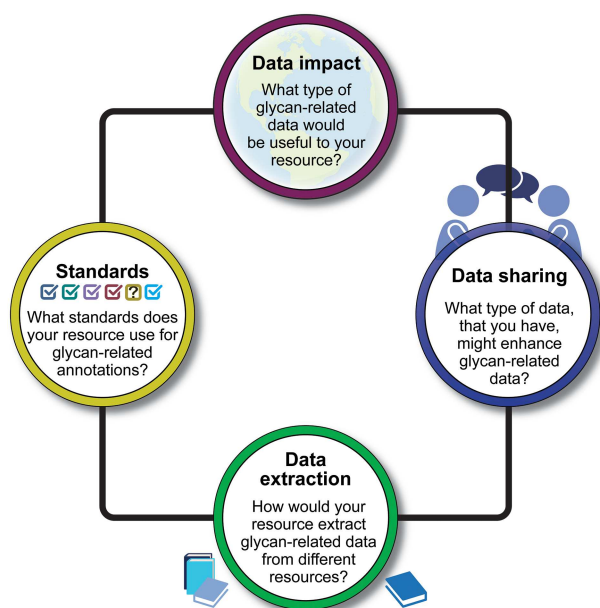


Figure 1. A visual summary of four key questions posed during the workshop. The responses to these questions provide insights into the diverse approaches and strategies employed by the workshop participants in addressing crucial aspects of glycan function annotations.

for the BioRED corpus; it has shown promising results when used to fine-tune models such as BioREx [20]. The Imperial College London (ICL) Glycosciences Lab is creating a novel curated glyco-interactome resource that aims to integrate knowledge on glycan-mediated biomolecular interactions from glycan microarray datasets. They are collaborating with Joram Posma at ICL to establish a text mining approach based on Auto-CORPus [21], a tool originally developed for metabolites. GlyGen [22], in collaboration with Vijay Shankar at the University of Delaware, developed text mining tools to create a dictionary of glycan structure terms [23]. Plans include using this dictionary to extract information related to glycan function from the literature. Automated text mining results from the mining of abstracts are available in the glycosylation section of the GlyGen protein pages. IntAct [24] uses text mining-based screening in PubMed and Europe PubMed Central (EPMC) [25]. They are working with EPMC literature services to develop a screening portal that aims to use a combinatorial text mining and machine learning-based strategy. Glyco@Expasy [26] uses a combination of text mining tools and manual opportunistic selection to identify sources. Depending on the format of the [supplementary material](#) in publications, data extraction ranges from manual to semi-automatic. UniProtKB [27] captures data from the literature by expert annotation. Proteins are annotated when a relevant publication is identified, or if experimental information about glycosylation sites is present. N-glycosylation sites are mainly curated based on prediction tools, but the annotations are not propagated to close homologs. O-glycosylation sites for which there is information about the function may be propagated.

Many curation workflows also include the integration of data from other resources. IntAct partners with MatrixDB [28], which is actively involved in the manual curation of glycosaminoglycan (GAG)–protein interactions. iPTMnet [29] captures sites of protein glycosylation from external resources

including UniProtKB, GlyGen, SIGNOR [30], and others. VirtualGlycome (<https://VirtualGlycome.org>) aims to manually collect data from sources including textbooks, handbooks, research papers, and published datasets from UniProtKB, GlyGen, NCBI, KEGG [31], and others. Data are processed in a field/topic-specific manner to obtain results that are curated by VirtualGlycome. GlyGen retrieves and harmonizes data related to glycoscience from both individual collaborators and public databases, such as UniProtKB, GlyTouCan, GlyConnect [32], O-GlcNAcAtlas [33] and the O-GlcNAc Database [34]. At ChEBI, data are submitted to the resource by the user community or other resources such as Rhea, GlyGen, MetaboLights [35], and others.

Validation and verification

Participants described various ways to validate and verify the quality of annotations. In some cases, this involves a two-step curation process in which annotations are checked by a second biocurator. For others, this mechanism is too costly or creates a bottleneck, and systematic automated checks based on manually created rules are put in place to detect common errors. Some resources try to correct errors when they encounter them, such as inconsistencies in describing alpha versus beta linkages. Others rely on assertions made by the original author and follow consistent rule sets when annotating publications.

The two-step protocol described above is employed by IntAct, MatrixDB, and Rhea, where reactions are created by one curator and verified by a second. SugarBase [36] manually annotates their entries with species, tissue, and disease associations, including glycan structure dysregulation in disease. Programmatic verification is subsequently used to check entries before merging with the SugarBase database. Glycan microarray data curation at ICL will be facilitated by machine learning-based semiautomatic extraction of experimental data that are compliant with the MIRAGE glycan microarray guidelines and followed by manual curation by specialists. GlyGen uses a combination of manual and automated checks that include verifying that glycosylation sites match the correct amino acid position in the UniProtKB canonical sequence and mapping nonstandard identifiers to their primary protein or glycan keys. When data are submitted to ChEBI, a stable and unique identifier is assigned to each submission, and the submitted metadata is subsequently checked and manually annotated by the ChEBI team.

Q1 conclusions

Overall, extraction of glycan-related data from the literature is largely a manual process and requires expert curators with domain knowledge in glycobiology. While text mining tools are used to identify papers of interest, information extraction is still performed by a human curator in most resources. Current tools are limited in scope and focus on a specific type of data, such as array data or glycosylation site annotation. Once data extraction and standardization are complete, programmatic tools are often used to assist curators in verifying the information. However, these verification tools or pipelines are custom-made by individual groups and there is no generalized approach allowing reuse of these tools for more than one curation initiative. It appears that close collaboration among the participating groups could lead to joint curation of publications which in turn can help in text mining efforts.

Better coordination between curation initiatives would benefit the entire ecosystem of resources in this space by creating tools and workflows that can be easily adapted to multiple purposes.

Q2. What type of glycan-related data would be useful to your resource, that you do not currently have access to?

Glycan data exist in various resources but may not be accessible due to inconsistencies in standards and interpretation. In addition to the resources involved in this effort, glycan-related data exist in various omics data resources (e.g. metabolomics and lipidomics) which need to be parsed and annotated. In this session, participants discussed the types of glycan-related data they would like to have access to and would enhance their resources. Resources compared data of interest and challenges to find common ground and opportunities to improve data sharing and accessibility.

Knowledge representation

The Immune Epitope Database (IEDB) [37] is interested in the connection between structures with epitopes, organisms by which the epitopes are produced, structures and organisms that contain the epitopes, and similarities between epitope structures across all species. They expressed that immunologists would benefit from systematic classifications of glycan-related data based on biology. Emphasis was placed on the need for effective linking to glycan resources and the availability of glycan-related annotations and relationships between structures and organisms in an Open Biological and Biomedical Ontology (OBO) [38] compatible format to allow reasoning and interoperability. GlycoMotif (<https://glycomotif.glyomics.org>) and GlyGen Sandbox (<https://sandbox.glyomics.org>) would benefit from glycan-related functional annotation terms from data sources outside the glycobiology domain. Challenges include indirect, non-glyco function for glyco-enzymes (e.g. glycan does not modulate the protein function) and lack of glyco-function terms, such as Gene Ontology (GO) Molecular Function terms for proteins. GlyGen suggested the development of a glycan function ontology describing the glycan field on the model of GO and how that can be useful to better explore and contextualize glycomic datasets. Glycan biology is currently captured in the GO [39, 40] as a function of proteins and ncRNAs. GO collaborates with experts in specific areas of biology to revise the ontology and add protein annotations, and would require a community effort to drive utility in glycobiology. To initiate a discussion in this area, they suggested that the community consider how they would like to model glycan functions. For example, are they conceptualized as molecular machines like proteins, mediators of specific molecular recognition by proteins, and/or post-translational modifications that modulate protein functions?

Interactions

A multitude of processes in health and most diseases critically involve glycans, directly or indirectly. With the increasing amounts of glycan microarray data in scientific publications, there is a great need to organize published and curated data; this together with other forms of glycan interaction data will serve as a much-needed knowledgebase for the broad scientific community. The ICL Glycosciences Lab emphasized

that it would be beneficial to integrate the curated glycan interaction data with the expression of the glycan ligands and their carrier molecules, ideally within the microenvironment where the biorecognition event occurs. This integration would lead to a deeper understanding of glycan-mediated interactions and elucidation of biological pathways in relation to other disciplines. IntAct described the integration in MatrixDB of curated GAG interaction datasets collected by several techniques, including affinity proteomics to identify GAG-binding proteins in various cell types, subcellular compartments, extracellular matrices, and biological fluids in health and disease, to generate context-dependent GAG interactomes, and to contextualize the global GAG interactome 2.0 [41]. The integration of the composition, and ideally of the sequences, of GAGs interacting with proteins would be useful to decipher the molecular recognition mechanisms of GAGs by proteins, and to link them to GAG and protein functions.

Structural representation

Visual representations of glycans and reactions, and tools to map cheminformatic representations to graphical representations (e.g. SMILES [42] to SNFG), would be of interest to several resources. Rhea works closely with ChEBI and requires the structures of substrates and products to construct reactions; however, it is not always easy to find the precise structures in the papers. ChEBI does not currently display glycan structures using the SNFG system; however, the process and tools to curate such data into ChEBI would be welcomed. GlycoEnzOnto [43] would like to more tightly integrate glycan structural data with their ontology to semantically link glycoenzyme function to the glycan structures they produce. Integration of GNOme, GlycoEpitope, GlycoTree, and GlyGen's glycan list could make this connection.

New data

Most groups expressed interest in collecting new data and annotations to enrich their collections, as well as more contextual information on the glycan sources, such as species, tissues, cell lines, instrumentation, disease, age, and sex. Attending glycobiologists commented that more data on site occupancy, glycan structure, and abundances (even if just relative) would be very useful. UniProtKB is focused on sites of protein glycosylation and capturing types of attachment using an internally controlled vocabulary, especially those with functional impact. As part of the ongoing standardization of all small molecule data in UniProtKB, the PTM vocabulary list (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/ptmlist.txt) is being progressively mapped to ChEBI. iPTMnet would like to include text mining results with protein, site, and glycan information. Other resources such as PRIDE [44] are interested in well-annotated glycoproteomic data, and Privateer [45] would benefit from having access to in-depth glycomics data, particularly for uncharacterized systems (e.g. extremophiles), as glycomics results are being used to validate or support 3D structures of glycans for scientists who may not necessarily be experts in glycobiology.

X-ray crystallography is the preferred tool to determine the high-resolution structures of glycan and glycan-protein complex structures [46, 47]. Although these systems are understood as not being easy targets, recent progress in cryo-EM and cryo-electron diffraction has significantly increased the

number of glycan–protein complex structures solved [48, 49]. The evolution of these methods carries the potential to generate a significant amount of structural data of the glycan and glycan–protein complexes. This advance would enrich the experimentally obtained 3D structures of glycan/glycoproteins instead of solely relying on glycan structures from molecular dynamics simulations.

Q2 conclusion

The resources agreed on the importance of knowledge exchange and the need for curation guidelines when considering glycan functions. Machine-readable annotations are lacking for glycan-related data and terms similar to those in GO. Such terms would facilitate linking across resources and connect knowledge of glycan structures, glyco-enzymes, organisms, and other types of information. Assignment of glycan structures to wet lab experimental data (e.g. MS spectra, lectin arrays, etc.) is variable across the scientific community. This inconsistency is partially due to the lack of reporting standards, but largely due to idiosyncrasies of data interpretation. Currently, “reliable” glycan-recognition data that can be curated into databases is somewhat limited and its interpretation can vary depending on the details of the wet lab experiment and interpretation of the data and the publication. Overlapping interests and opportunities to leverage existing frameworks and ontologies warrant further discussion and collaboration.

Q3. What type of data, that you have, might enhance glycan-related data?

The primary aim of a bioinformatic resource is to provide a service or knowledge that will be beneficial to the research community. As the importance of glycans gains visibility outside the glycobiology community, the type and format of the data must be tailored for accessibility to both glycobiologists and non-glycobiologists alike. In the third session, participants focused on opportunities to share data and enhance their glycan-related data by discussing what type of data each resource has that is unique and might be useful to other resources. In addition to what is described below, details of what individual resources have to offer are available in [Supplementary Table S1](#).

GlyGen provides text-mining data from manuscripts on experimentally determined glycosylation sites and has collected additional curated data through manual curation and collaboration with various other groups. Plans include adding ISOglyP-predicted O-glycosylation sites [50]. O-GlcNAcAtlas [33] contains O-GlcNAc sites, peptides, and proteins. UniProtKB annotates N-linked glycosylation, O-linked glycosylation, C-linked glycosylation, S-linked glycosylation, and glycation using an internal controlled vocabulary. The identity of the reducing monosaccharide is indicated in the feature lines, while the monosaccharide composition is indicated in the PTM comments. The PTM keyword “Glycoprotein” is added to the protein entry when these annotations are present. In terms of leveraging different resources to gain knowledge, examples such as GlycoEnzOnto mappings to GO and UniProtKB were presented. Such mappings can contextualize glycoenzyme knowledge and IUPAC-based reaction rules describing glycoenzyme functions, which could be used to construct glycosylation reaction networks in the future.

Several tools and resources were mentioned that biocurators might find useful. Examples include publication annotations from PubAnnotation [51] that provide links to glycan structure images and GlyYouCan IDs from PubMed abstracts. The tools GlycoSim [52] and GlycoMaple [53] allow users to predict and visualize glycan biosynthesis pathways. MicroGlycoDB [54] is a new database containing glycoenes, reactions, pathways, and glycans in microbial species. ChEBI contains a chemical ontology, where glycans are classified based on chemical structure and roles. The resource is widely used and provides an infrastructure for connecting glycans to ChEBI ontology. PubChem contains a wealth of content that helps to define the boundaries of “what is a glycan” from a structural perspective. Glycowork [55] has a large number of tissue, disease, and species associations with glycans. Many come from older papers and other data sources, such as the Consortium for Functional Glycomics (CFG). Associations of this kind can be particularly useful to study glycan biosynthesis during evolution, tissue-specific expression, and to identify disease-specific markers.

In terms of interaction data, MatrixDB has built an automated pipeline to standardize the format of GAG sequences interacting with proteins manually curated from the literature. The pipeline then translates the sequences into machine-readable GlycoCT or SNFG images and converts them into a format processed by a builder. The builder generates 3D structures of GAGs based on a repertoire of conformations experimentally validated by data extracted from crystallized GAG–protein complexes [56]. In addition to the structural data listed above, MatrixDB provides glycosaminoglycan interaction data collected by manual curation of the literature following the curation rules of the International Molecular Exchange consortium [57, 58]. The ICL Glycosciences Lab proposed CarbArrayKB as a new resource centered on the curated glycan microarray database, which will serve as a knowledgebase for the broad scientific community, focusing on glycan-mediated interactions from high-quality publications conforming to MIRAGE guidelines. They are in discussion with EMBL-EBI with the hope of collaborating to integrate the glycan interaction data as an extension of IntAct (molecular interactions) and Reactome (biomolecular pathways). IMGT-KG is the pioneer of the immunogenetics knowledge graph [59]. It integrates and establishes connectivity with other standards, including Protein Data Bank (PDB) [60], IEDB, PubMed, and most recently, the GlyConnect platform. In IMGT-KG, a crystal structure is linked to a glycan through the relation “hasGlyConnectLink” and includes direct links to IMGT and PDB resources. The GlycoShape [61] platform provides a comprehensive open-access database of glycan structures from extensive molecular dynamics simulations. This platform provides a wealth of structural information that can be integrated and coordinated with other glycan-related resources. GlycoShape also provides users with many helpful structural tools, such as ReGlyco, a bespoke algorithm in GlycoShape designed to rapidly restore the natural glycosylation to protein 3D structures and to predict N-glycosylation occupancy where unknown. Privateer validates data of 3D structures, torsional analysis, and cross-linking with some glycomics databases. One area for the future direction that was discussed is to annotate structures with matched entries across GlySpace Alliance resources [62].

Q3 conclusion

All attendees expressed interest in getting feedback on what data, tools, resources, and information will be useful for glycan function curation and allow for easier data sharing with other resources. It was agreed that closer collaboration with participating resources would increase data accessibility and create connections between different fields of research. Communication between researchers and the resources that desire access to their data remains a bottleneck. Creation of a clearing house for questions and recommendations regarding the deposition and curation of glycan-related data would be beneficial. Proper evidence attribution was emphasized as an essential feature for enhancing confidence in the usefulness of a resource when accessed by investigators less familiar with glycoscience data.

Q4. What standards does your resource use for glycan-related annotations?

Standards improve interoperability and interconnectivity between resources and across domains. While some standards such as gene symbols, protein accession numbers, and glycan symbol nomenclature have converged to a few types, others such as glycan IDs and glycan structure text representation are still in flux. Resources use varying identifiers for glycans and their annotations depending on their biomolecules of interest and intended audience. This session focused on standards used by each resource and the challenges in identifying an ideal standard. A summary of standards currently used by the participating resources is presented in [Table 2](#).

Gene, protein, and taxonomy

Gene, protein, and taxonomic information associated with glycan data, such as genes involved in glycan biosynthesis, glycan-binding proteins, and species annotations, can be found across resources. There is broad consensus in using HUGO Gene Nomenclature Committee [63] for gene symbols for human genes, UniProtKB or RefSeq [64] for protein accession numbers, and PDB for 3D structures. NCBI taxonomy ID [65] is used across all participating resources as the primary taxonomy identifier.

Glycan images, identifiers, and descriptions

The SNFG was developed as a community effort to standardize graphical representations of glycan structures and has achieved wide acceptance in the glycobiology community. Almost all glycoinformatics resources present at the workshop, including PubChem, support the SNFG format. MOL format is used to represent glycan structures in chemical-centric resources such as ChEBI, Rhea, and PubChem. Some resources such as GlyConnect support Oxford notation [66] as well.

The majority of resources use either GlyTouCan or ChEBI identifiers for glycans. All GlySpace Alliance members use GlyTouCan identifiers and, whenever available, map to PubChem and ChEBI. GlyGen also assigns unique identifiers to their list of motifs and a dictionary of glycan structure terms [23], where the biocuration community can add new terms. ChEBI identifiers and annotations are used by Rhea, IntAct, IEDB, and others. ChEBI is aligned with the IUPAC Blue Book, IUPAC Gold Book, and NC-IUBMB Enzyme Commission numbers and uses standard practices for assigning

chemical names, special characters, chemical structures, cross-reference links, ontologies, enzymes, and species. Reactions in Rhea, including glycans, are formally described using ChEBI vocabulary. However, glycan representation in Rhea is limited and work is ongoing with GlyGen and ChEBI to add glycans with publication references to ChEBI. IntAct and MatrixDB use ChEBI identifiers for GAG and oligosaccharides in agreement with the IMEx/IntAct curation rules. GAG-protein interactions are curated using Proteomics Standards Initiative Molecular Interaction and MI Ontology Lookup Service (<https://www.ebi.ac.uk/ols4/ontologies/mi>) controlled vocabularies. In general, many identifiers are used in biomedical databases, and ID conversion is needed to ensure interoperability. Tools such as TogoID [67] provide ID mapping between resources (e.g. PubChem and GlyTouCan) and can be accessed through the Web interface and Application Programming Interface.

Glycan sequences can be represented by glycan-specific notations such as WURCS and GlycoCT or by traditional chemical notations such as IUPAC, SMILES, InChI [68], and others. GlySpace Alliance members support WURCS and GlycoCT encoding while PubChem and ChEBI use InChI, SMILES, and IUPAC as the primary structural descriptors. Sequence converters allow resources to display both chemical and glycan-specific notations. For example, PubChem compounds include WURCS and LINUCS [69] sequences for biologics consisting of saccharides. Other participating resources use combinations of the aforementioned notations, except Glycowork, where a slightly modified version of IUPAC-condensed is used for glycan sequences, and MonosaccharideDB [70], which uses residue names (core monosaccharides + substituents) in various notations.

Ontology

Ontologies are used to describe relationships between terms or concepts in a way that is both human and machine-readable, enabling the organization of knowledge and interoperability between resources. There is consensus in the use of ontologies to describe protein biology, disease, cell, tissue, and anatomy, while glycan-centric ontologies have yet to proliferate into the bioinformatics space. The GO project creates terms that are the standard for describing the biological function, cellular component, and biological process of gene products, including glycosylated proteins. Several of the resources use GO terms to annotate protein entries including UniProtKB, Rhea, GlyConnect, and GlyGen. GlyGen also uses Protein Ontology (PRO) [71] annotations to describe proteoforms and glycoforms. GlycoEnzOnto describes the organization of enzymes participating in cellular glycosylation. It is currently focused on ~400 human glycosylating enzymes and related components. GlycoEnzOnto is based on the framework developed by GO and can be used for overrepresentation and pathway analysis.

A few glycan structure-specific ontologies and standards are currently in use. ChEBI has a manually curated ontology that is subdivided into three separate subontologies: Molecular structure, in which glycans are classified according to composition and structure; Role, in which glycans are classified based on their role within either a chemical, biological, and/or application context; and Subatomic Particle, which classifies particles smaller than atoms. ChEBI places carbohydrate structures into a hierarchy by making statements such as

Table 2. Standards relevant to the curation of glycans that are currently in use by different projects and resources

Resource/Project	Primary standards and annotation target								
	Glycan ID	Glycan image	Glycan text representation	Protein ID/AC	Tissue	Disease	Cell lines	Taxonomy	Annotation target
ChEBI	ChEBI, PubChem CID, GlyYouCan, KEGG GLYCAN	MOL	IUPAC, WURCS	UniProtKB, PDB	Uberon, BRENDA	N/A	N/A	NCBI	Small molecules
Glyco.me	N/A	SNFG	N/A	UniProtKB	Uberon, BRENDA	N/A	Cellosaurus	NCBI	Glycosites and glycoproteins
Glyco@Expasy	GlyYouCan	SNFG	GlySTreeM, GlycoCT	UniProtKB	Uberon, BRENDA	DO	Cellosaurus, CLO	NCBI	Glycoproteins, glycan-binding proteins, glycans
GlycoEnzOnto	IUPAC-condensed	N/A	N/A	NCBI Accession, UniProtKB, EC number	N/A	N/A	N/A	NCBI	Glycosylation pathways
GlycoMotif	GlyYouCan	SNFG	WURCS, GlycoCT	UniProtKB, Mouse Genome Informatics (MGI), Gene name	N/A	MP	N/A	N/A	Glycans
Glyco-sciences.de	LINUXS ID	SNFG	LINUXS	PDB	N/A	N/A	N/A	NCBI	Glycans, proteins (PDB entries)
GlycoShape	GlyYouCan	SNFG	IUPAC-condensed, GLYCAM, WURCS	UniProtKB	N/A	N/A	N/A	N/A	Glycans
GlyCosmos	GlyYouCan	SNFG	WURCS	UniProtKB	Uberon	DO	CLO	NCBI	Proteins, glycans, glycoconjugates, glycoconjugates, glycolipids
Glycwork	GlyYouCan	SNFG	IUPAC-Condensed	N/A	Uberon	DO	CL	NCBI	Glycans
GlyGen	GlyYouCan	SNFG	GNOME	UniProtKB	Uberon	DO	Cellosaurus	NCBI	Proteins, glycans, glycosylation sites
Sandbox	GlyYouCan	SNFG	GlycoCT	UniProtKB, Gene name	N/A	N/A	N/A	N/A	Glycans, glycosylation enzymes
GNOME	GlyYouCan	SNFG	N/A	N/A	N/A	N/A	N/A	N/A	Glycans
GO/PANTHER	N/A	N/A	N/A	UniProtKB	Uberon, Plant Ontology (PO), Fungal Anatomy Ontology (FAO)	N/A	CL	NCBI	Gene products (proteins, ncRNAs)

(continued)

Table 2. (Continued)

Resource/Project	Primary standards and annotation target								
	Glycan ID	Glycan image	Glycan text representation	Protein ID/AC	Tissue	Disease	Cell lines	Taxonomy	Annotation target
IEDB	ChEBI	N/A	N/A	UniProtKB	Uberon	DO	CLO	NCBI	Peptides, small molecules (<5000 Da)
IMGT	GlyConnect	N/A	N/A	PDB	N/A	Mondo	N/A	NCBI	Protein structural data
IntAct	ChEBI	N/A	N/A	UniProtKB	Uberon, BRENDA	Mondo	CLO Cellosaurus, EFO	NCBI	Mostly proteins, DNA, RNA, small molecules, and glycans
iPTMnet	N/A	N/A	N/A	UniProtKB	N/A	DO	N/A	NCBI	Proteins
MatrixDB	ChEBI, cross-reference to GlyTouCan when available	SNFG	GlycoCT	UniProtKB	Uberon, BRENDA	Mondo	CLO Cellosaurus, EFO	NCBI	Proteins and glycosaminoglycans
MIRAGE Glycan array guidelines	GlyTouCan (recommended)	SNFG (recommended)	ICL 2D Text, GlycoCT, (recommended)	Public database IDs (if available)	N/A	N/A	N/A	N/A	Glycan microarrays, glycans, glycan binding samples (including proteins and microorganisms)
NCFG—Glybrary	GlyTouCan, CFG Linear Nomenclature	SNFG	CFG Linear Nomenclature	UniProtKB	UniProtKB controlled vocabulary	Mondo	Cellosaurus	NCBI	Glycan microarrays and glycomics
oglcnac.org	O-GlcNAcAtlas, OGT-PIN	N/A	N/A	UniProtKB	N/A	N/A	N/A	NCBI	O-GlcNAcylated sites/peptides/proteins
Privateer	GlyTouCan, GlyConnect	SNFG	N/A	N/A	N/A	N/A	N/A	N/A	Proteins, glycans
PubChem	CID, ChEBI, GlyTouCan	MOL, SNFG	IUPAC-condensed, LINUCS, WURCS, IUPAC	EC Number, NCBI Accession, Ref-Seq Accession, UniProtKB, PRO	Uberon	Medical Subject Headings (MeSH), International Classification of Diseases (ICD-11)	Cellosaurus, ChEMBL, Library of Integrated Network-based Cellular Signatures (LINCS), EFO, CLO, CL, BRENDA	NCBI, Integrated Taxonomic Information System (ITIS), Catalogue of Life (COL)	Small molecules, nucleotides, carbohydrates, lipids, peptides, and chemically-modified macromolecules
Rhea	ChEBI	MOL	IUPAC-extended	UniProtKB	N/A	N/A	N/A	N/A	Biochemical and transport reactions
SNFG	IUPAC	SNFG	N/A	N/A	N/A	N/A	N/A	N/A	Monosaccharides, glycans, glycoconjugates
UniProtKB	N/A	N/A	N/A	UniProtKB	Controlled vocabulary	Controlled vocabulary	N/A	NCBI	Proteins

Note: Acronyms not spelled out in this table are available in the manuscript text. N/A indicates data are not available.

“ β -D-glucose *is_a* D-glucose, which *is_a* D-aldohexose, which *is_a* monosaccharide, which *is_a* carbohydrate.” GNOme is an OBO foundry ontology for GlyTouCan identifiers covering the complete glycan space, including defined and undefined aspects. GlyGen uses the GNOme ontology for subsumption exploration and propagating annotations (including species and glycan classifications) from more characterized structures to less characterized structures. The PRO ontology uses the GNOme ontology to describe the glycosylation of protein isoforms. While there still exists variability among resources in glycan motif nomenclature, the GlycoMotif glycan determinant and motif resource unifies a variety of glycan motif lists in one place and provides precomputed alignments of all motifs with GlyTouCan accessions. Additional motifs are used to help classify glycan structures into types and subtypes. Recently, the enzyme annotations on glycan structures from the GlyGen Sandbox have been combined with motif alignments to structures to associate glycoenzymes with motif residues. Through this mechanism, mouse knockout results from the International Mouse Phenotype Consortium [72] have been associated with glycoenzymes and glycomotifs. This association connects GlycoMotif to the Mammalian Phenotype (MP) Ontology [73], and to the Human Phenotype Ontology [74] by association.

The majority of the resources use Disease Ontology (DO) [75] and Mondo Disease Ontology (Mondo) [76] for disease annotations, Uberon [77] for anatomy, BRENDA Tissue Ontology [78] for tissue, and Cellosaurus [79], Cell Line Ontology (CLO) [80], Cell Ontology (CL) [81], and Experimental Factor Ontology (EFO) [82] for cell lines.

Q4 conclusion

Participants agreed that it is necessary to harmonize or at the very least provide robust mapping tools for the standards used for glycan-related annotations while maintaining a mechanism to represent ambiguity in glycan structures. Although GlyTouCan provides a unique reference for glycan structures, it is not used by all resources as the primary glycan identifier. Many groups use ChEBI IDs instead. However, ChEBI relies on GlySpace group members and others to provide the GlyTouCan cross-references and improve ChEBI's carbohydrate ontology. SNFG work has catalyzed the glycoinformatics community to use a common graphical representation of glycans. Close collaboration between NCBI, SNFG, and GlySpace Alliance is expected to further improve SNFG usage in publications. The GlycoMotif use case has established that glycan structures can be connected to phenotype ontologies, but these ontologies are still lacking terms relevant to glycosylation. Glycan-specific ontologies can provide a better understanding of glycan function within larger biological contexts. A vast majority of glycoinformatics resources provide Application Programming Interfaces and the option of using SPARQL queries (<https://www.w3.org/TR/sparql11-query/>). These options could be better utilized as common standards emerge from workshops such as the one described here.

Other relevant resources

It is important to note there are several resources that were not represented at the workshop that contribute directly or indirectly to glycan function annotation. For instance,

the PDB recently completed a Carbohydrate Remediation Project (<https://www.wwpdb.org/documentation/carbohydrate-remediation>), resulting in the development of a standard representation and validation framework for carbohydrate 3D structures within the PDB Core Archive [83, 84]. The Carbohydrate Structure Databases include data on manually curated natural carbohydrates from prokaryotes, plants, and fungi [85]. The O-GlcNAc Database provides an inventory of human O-GlcNAcylated proteins, their O-GlcNAc sites, identification methods, and corresponding references [34]. KEGG GLYCAN offers a collection of glycan structures, integrated with other KEGG resources, that allows for the examination of glycan-related pathways and networks [86]. The GlycoGene Database is a manually curated database of genes related to glycan synthesis [87]. Several other resources provide contextual information related to glycan function such as MetaboLights, a database for metabolomics studies and derived information [88], and Reactome, a manually curated and peer-reviewed pathway database [89]. A comprehensive list of glycoinformatics resources can be found in the Glycoinformatics chapter of Essentials of Glycobiology [90].

Workshop conclusions and future directions

This inaugural workshop on annotating glycan functions significantly improved our understanding of data collection and curation strategies across resources and also highlighted current deficits in the organization of annotations related to glycans and glycoconjugate function. It became evident that a harmonized, international effort is needed for the comprehensive capture of information in this domain. Several collaborative initiatives were proposed, such as the co-curation of research papers by multiple groups and the establishment of dedicated communication channels to facilitate user queries on glycan function curation and expert responses to those queries. Various groups expressed interest in forging targeted curation efforts. For instance, the GO Group expressed its willingness to collaborate with glycan biology experts to ensure the accuracy and completeness of glycan-related content within GO. This community is committed to soliciting feedback that can make their functional annotations more beneficial to the broader scientific community.

To achieve a cultural shift in data collection practices, it is vital for researchers to actively document glycan information, experimental details, and sample metadata at the data generation stage, paving the way for effective curation. Additionally, there is a recognized need to foster awareness and provide training within the scientific community to underscore the value of curation, which may not currently be apparent to all. The tremendous potential of Large Language Models (LLMs) cannot be ignored. LLMs potentially could be used to retrieve and summarize data from publications and, in the future, help in standardization, such as finding all different names or representations of a given glycan. However, there is still an effort needed from different stakeholders (e.g. authors, publishers, resource/tool developers, and funding agencies) to better structure and standardize the glycan information in the literature to help drive these developments. Two related and potentially parallel strategies were discussed at the workshop. One strategy would have the author/researcher bring data to the resources which would

provide tools to help with standardization for annotation. The other would enlist the community to help curate publications to establish glycan function-related annotations. An incentive for both approaches could be to give contributors acknowledgment through ORCID (<https://orcid.org>), allowing them to cite their contributions. Both models involve significant outreach and applying guidelines at strategic levels, ideally beginning before, during, or soon after publication. Thus, journal cooperation is also critical. Funding agencies should play a role in this effort as well, by supporting bioinformatic efforts that target increased data connectivity and, perhaps, by requesting that grant application reviewers evaluate whether applicants have followed community data standards in their recent publications.

The participants in this workshop were unanimous in their support for future opportunities to gather, evaluate progress, and propose strategies to collect and standardize annotations related to glycan function. They suggested that experts in glycan and glycoconjugate functions should meet next to propose useful annotation terms and concepts that can facilitate progress by informaticists. To that end, a second workshop on glycan function annotation titled “Defining Glycan Functions” was organized and held in conjunction with the 2023 Society for Glycobiology meeting in Hawaii (https://wiki.glygen.org/Glycan_Function_Workshop_at_SfG_2023). The second workshop enlisted experts in various domains of glycobiology to participate in discussions and contribute their insights toward the development of a robust framework comprising terms and concepts that link glycan functions with glycan structural features, motifs, and patterns. The discussion and outcomes of the second workshop, like this workshop, will be published for public comment and will serve as a framework for a future workshop that will bring together the glycoscience and bioinformatics participants for an event that promises to be a significant milestone in advancing the standardization of glycan function annotations and their representation across major biology and biomedical resources. We encourage the scientific community to provide input that can guide curation. The best way to provide feedback to improve and guide the glycan function curation efforts is to contact the GlySpace Alliance members via their contact page (<http://www.glyspace.org/contact.php>). Additionally, community members can also contact the Society for Glycobiology at SFG@glycobiology.org to reach out to a larger group of glycobiology researchers.

Acknowledgements

The following are the workshop organizing committee members: R.M. (GlySpace, GlyGen), K.M. (GlyGen), M.T. (GlySpace, GlyGen), R.R. (GlyGen), M.M. (UniProt, GlyGen), K.A.-K. (GlySpace, GlyCosmos, GlyTouCan), F.L. (GlySpace, GlyConnect), C.A. (PIR, BioCreative, UniProt), R.V. (IEDB), J.V. (GlyGen). We wish to thank Susan Bello, Alan Bridge, Rick Cummings, Patrice Duroux, Ted Groth, Deepti Jaiswal Kundu, and Taciana Manso for their valuable feedback. We also want to thank the International Society of Biocuration and the 16th Annual International Biocuration Conference team for organizing the satellite workshop, where the workshop findings were presented. Figure 1 was created with BioRender.com.

Supplementary data

Supplementary data is available at *Database* online.

Conflict of interest

None declared.

Funding

The workshop was sponsored by GlyGen National Institutes of Health (grant 1R24GM146616). The Society for Glycobiology sponsored Glycobiology Ambassador Ten Feizi’s participation.

Data Availability

No new data were generated or analyzed in support of this research. The workshop agenda can be accessed at https://wiki.glygen.org/Glycan_Function_Workshop_2023.

References

- Dayhoff MO, Eck RV, Chang MA *et al.* *Atlas of Protein Sequence and Structure*. Silver Spring, Maryland: National Biomedical Research Foundation, 1965.
- Strasser BJ. Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff’s Atlas of protein sequence and structure, 1954-1965. *J Hist Biol* 2010;43:623–60.
- Gagneux P, Panin V, Hennes T *et al.* Evolution of glycan diversity. In: Varki A, Cummings RD, Esko JD *et al.* (eds.), *Essentials of Glycobiology*. 4th edn. Cold Spring Harbor NY: Cold Spring Harbor Laboratory Press, 2022, 265–78.
- Schnaar RL, Sandhoff R, Tiemeyer M *et al.* Glycosphingolipids. In: Varki A, Cummings RD, Esko JD *et al.* (eds.), *Essentials of Glycobiology*. 4th edn. Cold Spring Harbor NY: Cold Spring Harbor Laboratory Press, 2022, 129–40.
- Flynn RA, Pedram K, Malaker SA *et al.* Small RNAs are modified with N-glycans and displayed on the surface of living cells. *Cell* 2021;184:3109–3124e3122. <https://doi.org/10.1016/j.cell.2021.04.023>
- Suzuki T, Cummings RD, Aebi M *et al.* Glycans in glycoprotein quality control. In: Varki A, Cummings RD, Esko JD *et al.* (eds.), *Essentials of Glycobiology*. 4th edn. Cold Spring Harbor NY: Cold Spring Harbor Laboratory Press, 2022, 529–38.
- Varki A. Biological roles of glycans. *Glycobiology* 2017;27:3–49. <https://doi.org/10.1093/glycob/cww086>
- Fujita A, Aoki NP, Shinmachi D *et al.* The international glycan repository GlyTouCan version 3.0. *Nucleic Acids Res* 2021;49:D1529–D1533. <https://doi.org/10.1093/nar/gkaa947>
- Hastings J, Owen G, Dekker A *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016;44:D1214–1219. <https://doi.org/10.1093/nar/gkv1031>
- Kim S, Chen J, Cheng T *et al.* PubChem 2023 update. *Nucleic Acids Res* 2023;51:D1373–D1380. <https://doi.org/10.1093/nar/gkac956>
- York WS, Agravat S, Aoki-Kinoshita KF *et al.* MIRAGE: the minimum information required for a glycomics experiment. *Glycobiology* 2014;24:402–06. <https://doi.org/10.1093/glycob/cwu018>
- Neelamegham S, Aoki-Kinoshita K, Bolton E *et al.* Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology* 2019;29:620–24. <https://doi.org/10.1093/glycob/cwz045>
- Herget S, Ranzinger R, Maass K *et al.* GlycoCT—a unifying sequence format for carbohydrates. *Carbohydr Res* 2008;343:2162–71. <https://doi.org/10.1016/j.carres.2008.03.011>

14. Tanaka K, Aoki-Kinoshita KF, Kotera M *et al.* WURCS: the Web3 unique representation of carbohydrate structures. *J Chem Inf Model* 2014;54:1558–66. <https://doi.org/10.1021/ci400571e>
15. Matsubara M, Aoki-Kinoshita KF, Aoki NP *et al.* WURCS 2.0 update to encapsulate ambiguous carbohydrate structures. *J Chem Inf Model* 2017;57:632–37. <https://doi.org/10.1021/acs.jcim.6b00650>
16. Allot A, Lee K, Chen Q *et al.* LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res* 2021;49:W352–W358. <https://doi.org/10.1093/nar/gkab326>
17. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res* 2013;41:W518–522. <https://doi.org/10.1093/nar/gkt441>
18. Bansal P, Morgat A, Axelsen KB *et al.* Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res* 2022;50:D693–D700.
19. Islamaj R, Kwon D, Kim S *et al.* TeamTat: a collaborative text annotation tool. *Nucleic Acids Res* 2020;48:W5–W11. <https://doi.org/10.1093/nar/gkaa333>
20. Lai PT, Wei CH, Luo L *et al.* BioREx: Improving biomedical relation extraction by leveraging heterogeneous datasets. *J Biomed Informat* 2023;146:ArXiv.
21. Beck T, Shorter T, Hu Y *et al.* Auto-CORPus: a natural language processing tool for standardizing and reusing biomedical literature. *Front Digit Health* 2022;4:788124. <https://doi.org/10.3389/fgdth.2022.788124>
22. York WS, Mazumder R, Ranzinger R *et al.* GlyGen: computational and informatics resources for glycoscience. *Glycobiology* 2020;30:72–73. <https://doi.org/10.1093/glycob/cwz080>
23. Vora J, Navelkar R, Vijay-Shanker K *et al.* The Glycan Structure Dictionary—a dictionary describing commonly used glycan structure terms. *Glycobiology* 2023;33:354–57. <https://doi.org/10.1093/glycob/cwad014>
24. Del Toro N, Shrivastava A, Ragueneau E *et al.* The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Res* 2022;50:D648–D653. <https://doi.org/10.1093/nar/gkab1006>
25. Rosonovski S, Levchenko M, Bhatnagar R *et al.* Europe PMC in 2023. *Nucleic Acids Res* 2023;52:D1668–D1676.
26. Mariethoz J, Alocci D, Gastaldello A *et al.* Glycomics@ExPASy: bridging the gap. *Mol Cell Proteomics* 2018;17:2164–76. <https://doi.org/10.1074/mcp.RA118.000799>
27. Bateman A, Martin M-J, Orchard S. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 2023;51:D523–D531.
28. Clerc O, Deniaud M, Vallet SD *et al.* MatrixDB: integration of new data with a focus on glycosaminoglycan interactions. *Nucleic Acids Res* 2019;47:D376–D381. <https://doi.org/10.1093/nar/gky1035>
29. Huang H, Arighi CN, Ross KE *et al.* iPTMnet: an integrated resource for protein post-translational modification network discovery. *Nucleic Acids Res* 2018;46:D542–D550. <https://doi.org/10.1093/nar/gkx1104>
30. Lo Surdo P, Iannuccelli M, Contino S *et al.* SIGNOR 3.0, the Signaling network open resource 3.0: 2022 update. *Nucleic Acids Res* 2023;51:D631–D637. <https://doi.org/10.1093/nar/gkac883>
31. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30. <https://doi.org/10.1093/nar/28.1.27>
32. Alocci D, Mariethoz J, Gastaldello A *et al.* GlyConnect: glycoproteomics goes visual, interactive, and analytical. *J Proteome Res* 2019;18:664–77. <https://doi.org/10.1021/acs.jproteome.8b00766>
33. Ma J, Li Y, Hou C *et al.* O-GlcNAcAtlas: a database of experimentally identified O-GlcNAc sites and proteins. *Glycobiology* 2021;31:719–23. <https://doi.org/10.1093/glycob/cwab003>
34. Wulff-Fuentes E, Berendt RR, Massman L *et al.* The human O-GlcNAcome database and meta-analysis. *Sci Data* 2021;8:25. <https://doi.org/10.1038/s41597-021-00810-4>
35. Kale NS, Haug K, Conesa P *et al.* MetaboLights: an open-access database repository for metabolomics data. *Curr Protoc Bioinform* 2016;53:14–3. <https://doi.org/10.1002/0471250953.bi1413s53>
36. Bojar D, Powers RK, Camacho DM *et al.* Deep-learning resources for studying glycan-mediated host-microbe interactions. *Cell Host Microbe* 2021;29:132–144e133. <https://doi.org/10.1016/j.chom.2020.10.004>
37. Vita R, Mahajan S, Overton JA *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* 2019;47:D339–D343. <https://doi.org/10.1093/nar/gky1006>
38. Jackson R, Matentzoglou N, Overton JA *et al.* OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database (Oxford)* 2021;2021. <https://doi.org/10.1093/database/baab069>
39. Mi H, Muruganujan A, Ebert D *et al.* PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* 2019;47:D419–D426. <https://doi.org/10.1093/nar/gky1038>
40. Gene_Ontology_Consortium. Aleksander SA, Balhoff J, Carbon S *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* 2023;224:iyad031. <https://doi.org/10.1093/genetics/iyad031>
41. Vallet SD, Berthollier C, Ricard-Blum S. The glycosaminoglycan interactome 2.0. *Am J Physiol Cell Physiol* 2022;322:C1271–C1278. <https://doi.org/10.1152/ajpcell.00095.2022>
42. Weininger D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;28:31–36. <https://doi.org/10.1021/ci00057a005>
43. Groth T, Diehl AD, Gunawan R *et al.* GlycoEnzOnto: a GlycoEnzyme pathway and molecular function ontology. *Bioinformatics* 2022;38:5413–20. <https://doi.org/10.1093/bioinformatics/btac704>
44. Perez-Riverol Y, Bai J, Bandla C *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022;50:D543–D552. <https://doi.org/10.1093/nar/gkab1038>
45. Agirre J, Iglesias-Fernandez J, Rovira C *et al.* Privateer: software for the conformational validation of carbohydrate structures. *Nat Struct Mol Biol* 2015;22:833–34. <https://doi.org/10.1038/nsmb.3115>
46. Wormald MR, Petrescu AJ, Pao YL *et al.* Conformational studies of oligosaccharides and glycopeptides: complementarity of NMR, X-ray crystallography, and molecular modelling. *Chem Rev* 2002;102:371–86. <https://doi.org/10.1021/cr990368i>
47. Nagae M, Yamaguchi Y. Function and 3D structure of the N-glycans on glycoproteins. *Int J Mol Sci* 2012;13:8398–429. <https://doi.org/10.3390/ijms13078398>
48. Atanasova M, Bagdonas H, Agirre J. Structural glycobiochemistry in the age of electron cryo-microscopy. *Curr Opin Struct Biol* 2020;62:70–78. <https://doi.org/10.1016/j.sbi.2019.12.003>
49. Agirre J, Davies GJ, Wilson KS *et al.* Carbohydrate structure: the rocky road to automation. *Curr Opin Struct Biol* 2017;44:39–47. <https://doi.org/10.1016/j.sbi.2016.11.011>
50. Mohl JE, Gerken TA, Leung MY. ISOGLYP: de novo prediction of isoform-specific mucin-type O-glycosylation. *Glycobiology* 2021;31:168–72. <https://doi.org/10.1093/glycob/cwaa067>
51. Nam HJ, Yamada R, Park HS. Using the PubAnnotation ecosystem to perform agile text mining on Genomics & Informatics: a tutorial review. *Genomics Inform* 2020;18:e13. <https://doi.org/10.5808/GI.2020.18.2.e13>
52. Kouka T, Akase S, Sogabe I *et al.* Computational modeling of O-linked glycan biosynthesis in CHO cells. *Molecules* 2022;27:1766. <https://doi.org/10.3390/molecules27061766>
53. Huang YF, Aoki K, Akase S *et al.* Global mapping of glycosylation pathways in human-derived cells. *Dev Cell* 2021;56:1195–1209e1197. <https://doi.org/10.1016/j.devcel.2021.02.023>

54. Hosoda M, Aoki K, Guerardel Y *et al.* Meeting report on the international symposium on microbial Glycoconjugates and the GlySpace alliance: from micro- to macroglycoscience (MiGGA symposium). *Glycobiology* 2022;32:1066–67. <https://doi.org/10.1093/glycob/cwab062>
55. Thomes L, Burkholz R, Bojar D. Glycowork: a Python package for glycan data science and machine learning. *Glycobiology* 2021;31:1240–44. <https://doi.org/10.1093/glycob/cwab067>
56. Clerc O, Mariethoz J, Rivet A *et al.* A pipeline to translate glycosaminoglycan sequences into 3D models. Application to the exploration of glycosaminoglycan conformational space. *Glycobiology* 2019;29:36–44. <https://doi.org/10.1093/glycob/cwy084>
57. Porras P, Barrera E, Bridge A *et al.* Towards a unified open access dataset of molecular interactions. *Nat Commun* 2020;11:6144. <https://doi.org/10.1038/s41467-020-19942-z>
58. Orchard S, Kerrien S, Abbani S *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat Methods* 2012;9:345–50. <https://doi.org/10.1038/nmeth.1931>
59. Sanou G, Giudicelli V, Abdollahi N *et al.* IMGT-KG: A Knowledge Graph for Immunogenetics. In: Sattler U *et al.* *The Semantic Web – ISWC 2022. ISWC 2022. Lecture Notes in Computer Science*. Vol. 13489, pp.628–42. Cham: Springer, 2022.
60. Berman HM, Westbrook J, Feng Z *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–42. <https://doi.org/10.1093/nar/28.1.235>
61. Ives CM, Singh O, D’Andrea S *et al.* Restoring protein glycosylation with GlycoShape. *bioRxiv* 2023. <https://doi.org/10.1101/2023.12.11.571101>
62. Lisacek F, Tiemeyer M, Mazumder R *et al.* Worldwide glycoscience informatics infrastructure: the GlySpace Alliance. *JACS Au* 2023;3:4–12.
63. Seal RL, Braschi B, Gray K *et al.* Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res* 2023;51:D1003–D1009. <https://doi.org/10.1093/nar/gkac888>
64. O’Leary NA, Wright MW, Brister JR *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;44:D733–745.
65. Schoch CL, Ciufu S, Domrachev M *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020;2020:1–21.
66. Harvey DJ, Merry AH, Royle L *et al.* Proposal for a standard system for drawing structural diagrams of N- and O-linked carbohydrates and related compounds. *Proteomics* 2009;9:3796–801.
67. Ikeda S, Ono H, Ohta T *et al.* TogoID: an exploratory ID converter to bridge biological datasets. *Bioinformatics* 2022;38:4194–99.
68. Heller SR, McNaught A, Pletnev I *et al.* InChI, the IUPAC International Chemical Identifier. *J Cheminform* 2015;7:23. <https://doi.org/10.1186/s13321-015-0068-4>
69. Bohne-Lang A, Lang E, Forster T *et al.* LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr Res* 2001;336:1–11. [https://doi.org/10.1016/S0008-6215\(01\)00230-0](https://doi.org/10.1016/S0008-6215(01)00230-0)
70. Bohm M, Bohne-Lang A, Frank M *et al.* Glycosciences.DB: an annotated data collection linking glycomics and proteomics data (2018 update). *Nucleic Acids Res* 2019;47:D1195–D1201. <https://doi.org/10.1093/nar/gky994>
71. Natale DA, Arighi CN, Blake JA *et al.* Protein Ontology: a controlled structured network of protein entities. *Nucleic Acids Res* 2014;42:D415–421.
72. Munoz-Fuentes V, Cacheiro P, Meehan TF *et al.* The International Mouse Phenotyping Consortium (IMPC): a functional catalogue of the mammalian genome that informs conservation. *Conserv Genet* 2018;19:995–1005.
73. Smith CL, Goldsmith CA, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol* 2005;6:R7.
74. Kohler S, Gargano M, Matentzoglou N *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res* 2021;49:D1207–D1217.
75. Schriml LM, Munro JB, Schor M *et al.* The Human Disease Ontology 2022 update. *Nucleic Acids Res* 2022;50:D1255–D1261.
76. Vasilevsky NA, Matentzoglou NA, Toro S *et al.* Mondo: unifying diseases for the world, by the world. *medRxiv* 2022. <https://doi.org/10.1101/2022.04.13.22273750>
77. Mungall CJ, Torniai C, Gkoutos GV *et al.* Uberon, an integrative multi-species anatomy ontology. *Genome Biol* 2012;13:R5. <https://doi.org/10.1186/gb-2012-13-1-r5>
78. Gremse M, Chang A, Schomburg I *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res* 2011;39:D507–513.
79. Bairoch A. The Cellosaurus, a cell-line knowledge resource. *J Biomol Tech* 2018;29:25–38. <https://doi.org/10.7171/jbt.18-2902-002>
80. Sarntivijai S, Lin Y, Xiang Z *et al.* CLO: the cell line ontology. *J Biomed Semantics* 2014;5:37. [10.1186/2041-1480-5-37](https://doi.org/10.1186/2041-1480-5-37)
81. Diehl AD, Meehan TF, Bradford YM *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J Biomed Semantics* 2016;7:44.
82. Malone J, Holloway E, Adamusiak T *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 2010;26:1112–18. <https://doi.org/10.1093/bioinformatics/btq099>
83. Shao C, Feng Z, Westbrook JD *et al.* Modernized uniform representation of carbohydrate molecules in the Protein Data Bank. *Glycobiology* 2021;31:1204–18. <https://doi.org/10.1093/glycob/cwab039>
84. Feng Z, Westbrook JD, Sala R *et al.* Enhanced validation of small-molecule ligands and carbohydrates in the Protein Data Bank. *Structure* 2021;29:393–400e391. <https://doi.org/10.1016/j.str.2021.02.004>
85. Toukach PV, Egorova KS. Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts. *Nucleic Acids Res* 2016;44:D1229–1236. <https://doi.org/10.1093/nar/gkv840>
86. Hashimoto K, Goto S, Kawano S *et al.* KEGG as a glycome informatics resource. *Glycobiology* 2006;16:63R–70R. [10.1093/glycob/cwj010](https://doi.org/10.1093/glycob/cwj010)
87. Togayachi A, Dae K, Shikanai T *et al.* A database system for glycogenes (GGDB). 2008. In: Taniguchi N, Suzuki A, Ito Y, Narimatsu H, Kawasaki T, Hase S (eds.), *Experimental Glycoscience Glycobiology*. Japan: Springer. 2008, 423–25.
88. Yurekten O, Payne T, Tejera N *et al.* MetaboLights: open data repository for metabolomics. *Nucleic Acids Res* 2024;52:D640–D646. <https://doi.org/10.1093/nar/gkad1045>
89. Milacic M, Beavers D, Conley P *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res* 2024;52:D672–D678. <https://doi.org/10.1093/nar/gkad1025>
90. Aoki-Kinoshita KF, Campbell MP, Lisacek F *et al.* Glycoinformatics. Chapter 52. In: Varki A, Cummings RD, Esko JD *et al.* (eds.), *Essentials of Glycobiology*. 4th edn. Cold Spring Harbor NY: Cold Spring Harbor Laboratory Press, 2022, 705–18.