



## MicroGlycoDB: A database of microbial glycans using Semantic Web technologies

Sunmyoung Lee<sup>a</sup>, Louis-David Leclercq<sup>c</sup>, Yann Guerardel<sup>c</sup>, Christine M. Szymanski<sup>d</sup>, Thomas Hurtaux<sup>e</sup>, Tamara L. Doering<sup>e</sup>, Takane Katayama<sup>f</sup>, Kiyotaka Fujita<sup>g</sup>, Kazuhiro Aoki<sup>h</sup>, Kiyoko F. Aoki-Kinoshita<sup>a,b,\*</sup>

<sup>a</sup> Glycan and Life Systems Integration Center (GaLSIC), Soka University, Hachioji, Tokyo, Japan

<sup>b</sup> Graduate School of Science and Engineering, Soka University, Hachioji, Tokyo, Japan

<sup>c</sup> French National Center for Scientific Research (CNRS), University of Lille, Lille, France

<sup>d</sup> Complex Carbohydrate Research Center (CCRC), University of Georgia, Athens, GA, USA

<sup>e</sup> Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO, USA

<sup>f</sup> Graduate School of Biostudies, Kyoto University, Kyoto, Japan

<sup>g</sup> The United Graduate School of Agricultural Sciences, Kagoshima University, Kagoshima, Japan

<sup>h</sup> Department of Cell Biology, Neurobiology and Anatomy, Medical College of Wisconsin, Milwaukee, WI, USA

### ARTICLE INFO

#### Keywords:

Microbes  
Glycosylation  
Database  
Data integration  
Semantic data

### ABSTRACT

Glycoconjugates are present on microbial surfaces and play critical roles in modulating interactions with the environment and the host. Extensive research on microbial glycans, including elucidating the structural diversity of the glycan moieties of glycoconjugates and polysaccharides, has been carried out to investigate the function of glycans in modulating the interactions between the host and microbes, to explore their potential applications in the therapeutic targeting of pathogenic species, and in the use as probiotics in gut microbiomes. However, glycan-related information is dispersed across numerous databases and a vast amount of literature, which makes it laborious and time-consuming to identify and gather the relevant information about microbial glycosylation. This challenge can be addressed by a comprehensive database, which could offer insight into the fundamental processes underlying glycosylation. We have developed a MicroGlycoDB database to provide integrated glycan information on important model microorganisms. The data is described using Semantic Web Technologies, which allow microbial glycan data to be represented in a structured format accessible by machines, thus facilitating data sharing and integration with other resources that catalog features such as pathways, diseases, or interactions. This semantic data based on ontologies will contribute to the discovery of new knowledge in the field of microbiology, along with the expansion of information on the glycosylation of other microorganisms.

### Introduction

Glycans on the microbial cell surface and glycoenzymes play critical roles in the integrity of the cell [1], permeability of chemical compounds related to drug resistance [2,3], and movement of the cell [4]. Glycosylation in microbes responsible for disease, including *Streptococcus* spp., *Salmonella typhi*, *Pseudomonas aeruginosa*, *Acinetobacter baumannii*, *Neisseria meningitidis* ([5–9]), heavily glycosylated viruses [10], and the encapsulated yeast *Cryptococcus neoformans* [11], plays significant roles in their pathogenesis by aiding in immune evasion. Also, the normal commensal microbiota maintain intestinal health through

glycoenzymes, allowing the metabolism of complex glycans [12]. For example, the lipoglycan of *Mycobacterium tuberculosis* (*M. tuberculosis*), characterized by latent infection, shows the role of glycans in inhibiting the function of immune cells. This occurs via the interaction of lipooarabinomannan (ManLAM) with pattern-recognition receptors (PRRs) on dendritic cells (DCs), ultimately resulting in immune evasion [13]. Some microbial glycans mimic host glycans to avoid host immune surveillance, while bacterial glycoenzymes can modify the host glycans to enable bacteria to adhere to the host cell or use them as nutrient sources [14]. *N*-glycosylation of *Campylobacter jejuni* (*C. jejuni*) proteins protects them from cleavage by host proteases [15] and its mimicry of host

\* Corresponding author at: 1 Chome-236 Tangimachi, Hachioji, Tokyo, Japan.

E-mail address: [kkiyoko@soka.ac.jp](mailto:kkiyoko@soka.ac.jp) (K.F. Aoki-Kinoshita).

<https://doi.org/10.1016/j.bbadv.2024.100126>

Received 20 September 2024; Received in revised form 25 November 2024; Accepted 26 November 2024

Available online 30 November 2024

2667-1603/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

glycans enhances host-pathogen interactions [16]. This can lead to the autoimmune disease Guillain-Barré syndrome [17], which shows the crucial roles of microbial glycans for initial attachment and colonization leading to disease manifestation.

Recent research has also revealed the role of *C. jejuni* N-glycans in multidrug-resistant efflux pump proteins (CmeABC) that actively remove drugs from bacterial cells. N-glycans directly mediate the activity of the efflux pump by affecting the protein conformation [18], which suggests they could be an ideal target to overcome antibacterial drug resistance.

Bacterial organelles for motility, such as flagella and pili [19,20], are frequently decorated with distinctive monosaccharides. These present a greater diversity of structural compositions than their eukaryotic counterparts, for example including pseudaminic acid (Pse) [21], legionaminic acid (Leg) [22], rhamnose (Rha) [23], 3-deoxy-d-manno-oct-2-ulosonic acid (Kdo) [24], and N-acetylglucosamine (GlcNAc) [25]. This diversity may reflect the adaptation of diverse host organisms to glycans, enabled by short generation time and horizontal genetic exchange. Among these monosaccharides, Pse and Leg are exclusively expressed in pathogenic microbes, including *C. jejuni* and *Helicobacter pylori*.

Research into the metabolic pathway of these intriguing glycan structures of *C. jejuni* has revealed the role of novel glycoenzymes that act in forming glycosidic linkages and modifications of biomolecules. The accumulated information on glycosyltransferases, glycosidases, kinases, monosaccharides, and glycoconjugates has prompted the development of metabolic pathway-based recombinant glycans and fermentable products for research and industrial purposes [26,27]. Genetic studies into the gene clusters that encode glycoenzymes have also advanced understanding of the functional and structural diversity of carbohydrates that exists within a species or genus [28].

However, despite the elucidation of the structures and functions of glycans in multiple microorganisms [29,30], crucial information is dispersed across numerous publications and databases. Thus, the details of even a single glycoenzyme are not readily accessible in a single location, either online (in a database) or in the literature. Efforts to acquire specific glycan-related information therefore require substantial amounts of time and effort. For example, to obtain information about the capsular polysaccharides (CPS) of *C. jejuni*, an extensive search across multiple databases would be needed: UniProtKB (UniProt Knowledgebase), a global protein resource encompassing almost all species [31], has glycoprotein information; CAZy (Carbohydrate-Active enZymes) provides amino acid sequences and 3D structures of the carbohydrate-related enzymes [32] organized into families of similar enzymes; Pfam provides the protein families classified by shared structural and functional features [33]; CSDB (the Carbohydrate Structure Database) provides detailed information on the structural characteristics of carbohydrates in bacteria and fungi [34]; PDB (Protein Data Bank) is a repository of three-dimensional structures of proteins, gene, and their complexes [35]; GlyTouCan is the international glycan structure repository [36]; ChEBI (Chemical Entities of Biological Interest) organizes compounds of biological interest based on ontologies [37]; and Rhea (<https://www.rhea-db.org>) provides comprehensive and detailed information about biochemical reactions and their associated enzymes [38].

Sharing or integrating data between experimental databases from different biological disciplines of interest, including glycomics, transcriptomics, proteomics, and genomics, will provide users with a comprehensive view of a research question, which will help researchers gain valuable insights and eventually suggest hypotheses for further study. The most significant challenge in data integration is unifying the data format and the terms (controlled vocabulary) used to describe the same molecules across disparate databases, which originate from variations in naming that are generated by different research groups. For instance, trehalose 6,6'-dimycolate, a pathogenic glycolipid in *M. tuberculosis*, has the recommended symbol name of TDM, but also has

synonyms such as “cord factor” or “trehalose dimycolate”. To address the problem of inconsistent naming, ontologies have been proposed as a potentially effective way to annotate resources in a formal format. They allow a computer to comprehend the semantics of data represented in natural language, whether the data derive from biomolecules, biological processes, phenotypes, or other complex concepts. Numerous ontologies have been developed to handle the semantic annotation of complex molecules and concepts within fields of interest, and thereby guarantee consistent representation of knowledge. With the semantic model for knowledge representation, another prerequisite for data integration is standardization of file format, which makes data exchange or reuse between different databases easier. Prominent databases in the life sciences such as UniProt, Ensembl, PubChem, and the Microbial Genome Database (MBGD) have shown the successful use of Semantic Web technologies. This is a standard that allows data to be described in a structured, standardized format and makes it easier to be integrated with other resources.

Generally, the investigation of bacterial glycans has focused on model organisms that hold medical significance as either pathogens or commensal bacteria, such as *C. jejuni*, *M. tuberculosis*, and *Bifidobacterium bifidum*. Model organisms are essential for discovering the genetic and structural features of glycans in metabolism or pathogenesis because they facilitate the investigation of fastidious organisms in a detailed and rapid way. Thus, we have applied these Semantic Web standards to six model organisms to create our novel MicroGlycoDB database. We first obtained glycan-related data encompassing a wide range of information (e.g., cellular compartments, enzymes, genes, glycans, metabolic pathways, and glycan synthesis pathways) for representative microbes from experts. Then, we transformed this data into a standardized format through a process called RDFization, in which this data is transformed into a standardized format, allowing for the representation of resources in a structured and machine-readable manner (such that computers can read and process the information), thereby facilitating information sharing and reuse among various life science communities. Since the RDFization process entails semantic annotation of resources by applying ontologies that provide a standardized methodology for sharing and reuse of scientific data, the data can be semantically enriched by connecting them to other RDF data in publicly available databases. Ontologies are pre-defined vocabulary terms that are hierarchically organized to represent various concepts in the life sciences; when the same ontologies are referenced by databases and data resources across the Web, data can be linked to indicate that they refer to the same concept. Therefore, as we have done in this work, by RDFizing microbial glycan data, we expect that the semantically interconnected glycan-related data in the MicroGlycoDB will contribute to our understanding of the roles of microbial glycans at a systems level.

## Methods

### Data collection for glycan-related information

We obtained glycan-related data for six microbial species: *B. bifidum*, *Bifidobacterium longum*, *C. jejuni*, *C. neoformans*, *Mycobacterium abscessus*, and *M. tuberculosis*. We conducted a thorough review of the literature using PubMed identifiers to gather details on glycan information related to each glycogene. To assign GlyTouCan IDs to the glycans described in textual formats, such as IUPAC or Linear Code, the glycans were transformed into WURCS (Web3 Unique Representation of Carbohydrate Structures) [39] format using the glycan format converter API (Application Programming Interface) (<https://doc.glycosmos.org/api/glycanformatconverter>) developed by the GlyCosmos project. In the case of the glycan format used in the CSDB database, we used the CSDB/SNFG structure editor (<http://csdb.glycoscience.ru/snfgedit/snfgedit.html>) to obtain the image and CSDB text data of glycan structures, and then converted the glycan linear format to WURCS format. We registered the glycans in WURCS format into the GlyTouCan

repository to obtain glycan IDs. When a gene name is present in the UniProt database, the UniProt ID and Rhea ID were obtained. Using the acquired Rhea ID, the ChEBI ID and PubChem ID, of the substrate and product glycans participating in the enzyme reaction were obtained.

#### RDFization and validation

The MicroGlycoDB database is developed based on Semantic Web Technologies, which consists of Resource Description Framework (RDF) for the data format [40], Web Ontology Language (OWL) for the definition of ontologies [41], and SPARQL (SPARQL Protocol and RDF Query Language) for queries [42]. The RDF data model that we designed for the description of microbial glycosylation-related information was reorganized in a spreadsheet format. RDF statements were represented in turtle format, which is the simplest and most easily understandable format among the major RDF formats. The data was converted into RDF turtle format using RDFLib, a Python library for handling RDF data, based on a previous study for a more specific method to generate RDF triples [43]. We were able to save the RDF data in a compact textual form, in which a long and repeated IRI can be shortened as a prefixed name, which is an arbitrary name created by the researcher. The transformed RDF triples were uploaded into our Virtuoso database server [44], which is a graph database designed to store the RDF data, and the data was utilized to evaluate the accuracy of the semantic data via the SPARQL endpoint (<https://ts.glycosmos.org/sparql>).

To verify the RDF data, SPARQL queries were generated and organized using SPARQLList, the Shape Expressions (ShEx) were coded using PyShEx, version 0.8.1, and then ShEx was carried out by simple Python code. The RDF data was loaded into the ShEx engine and evaluated based on the shape definition to report the object node, object value, and type (<https://github.com/sunmyoung/MicroGlycoDB>).

#### User interface

The development of MicroGlycoDB involved the implementation of an efficient retrieval system of the data from the RDF database; this operates as the backend using Python 3.2. The client-side view for the web interface was implemented using HTML, CSS, JavaScript, jQuery, and Ajax. We used Flask, a Python-based web development framework, to create unique web pages for each microbe. These pages make it easier to visualize microbial images, update web pages to reflect new glycan information, and display extracted data from the RDF storage.

#### Ontologies for the RDF graph

To transform the collected data into RDF format, ontologies pre-existing within the biological domain were inspected using Protégé (<https://protege.stanford.edu/>), an editor for creating, sharing, and visualizing ontologies, and the OLS (Ontology Lookup Service) online service (<https://www.ebi.ac.uk/ols4>). After standardized vocabularies for resources were determined, the terms of the ontologies were employed according to their defined specifications. The glycans, glycoconjugates, and enzyme reactions were described using the GlycoRDF [45] or GlycoConjugate Ontology (GlyCoCoO) [46] that were developed for the semantic description of glycan structures, their modifications, and their core proteins.

## Results

#### Data collection

We collected datasets for the following six microorganisms that include the names or identification numbers of genes that encode glycoenzymes, revealing the structures or roles of glycosylation based on the accumulated results: for *B. bifidum* and *B. longum* information on enzymes that break down glycosidic bonds; for *M. tuberculosis* and

*M. abscessus* enzymes that form glycosidic bonds to elongate or branch glycan structures; for *C. jejuni* motility-related proteins and enzymes, including hydrolases, glycosyltransferases, epimerases, kinases, and ligases; and for *C. neoformans* glycosyltransferases, phosphorylases, and chitin synthases that generate protein glycans, GPI structures, and components of the capsule and cell wall.

*Bifidobacterium*, one of the most prevalent bacterial genera in the intestinal tract, establishes beneficial relationships with the host by contributing to immune homeostasis in the intestinal epithelium, where their metabolites and ability to utilize host carbohydrates play important roles [47–49]. The information on enzymes that are related to the degradation of HMOs (Human Milk Oligosaccharides), such as sialidase (SiaBb2),  $\beta$ -galactosidases (BbgIII), and  $\alpha$ -fucosidases (AfcA, AfcB), was provided with gene names. The essential information for the glycan structures produced by the corresponding enzymes was extracted from the ChEBI, GlyTouCan, and CSDB databases. When the information that was verified by these databases was limited, it was supplemented with information obtained through an extensive search across pertinent literature and databases. For example, for the *cj1641* gene of *C. jejuni*, the relevant enzyme product information was retrieved from the UniProtKB, CAZy, PDB, and Enzyme Commission (EC) databases. The enzyme reaction components, such as the donor and acceptor substrates, were extracted from the Rhea database with their Rhea ID, which was used to retrieve the information of the participant biomolecules taking part in the enzymatic reaction, such as glycans or chemicals; the latter could be represented by a unique identifier provided by the corresponding database such as PubChem or ChEBI. The provided gene names of string datatype were represented by a unique URI (Uniform Resource Identifier), which is essential to standardizing semantic data and allowing a computer to read the meaning and relationship of data resources. The glycan participants were also assigned a GlyTouCan ID, which is the international repository developed to facilitate the referencing of complex glycan structures and thereby avoid the confusion caused by the multiple nomenclatures of glycans, such as IUPAC, Linear Code, or research group-specific naming formats.

The pathogenic bacteria *M. tuberculosis* (MTb) and *M. abscessus*, which feature a unique membrane structure, have been studied as important pulmonary pathogens. The MTb cell envelope, which evolved as a formidable defensive barrier, contains multiple distinct glycoconjugate structures such as the mAGP (mycolyl-arabinogalactan-peptidoglycan) complex, phosphatidylinositol mannosides (PIMs), lipomannans (LMs), and lipoarabinomannans (LAMs). These allow it to survive intracellularly in the phagosome, despite a hostile environment filled with a low pH and reactive oxidative-nitrosative stressors [50,51]. *M. abscessus* has further been observed to transition from a smooth type, characterized by the presence of cell surface-associated glycopeptidolipids (GPL), to a rough type lacking GPL [52]. To represent these glycolipids, which consist of characteristic fatty acids and glycan residues such as l-arabinose and d-mannose, we performed a search on the CSDB database. The unique identifier number, figure format, and CSDB linear text format for the lipoglycan structure were retrieved so that we could describe the glycolipids in MTb and *M. abscessus* with a focus on glycan structures.

*C. neoformans* is an opportunistic pathogen that causes fungal meningoencephalitis. The polysaccharide structures of its cell wall, composed of an inner layer of glucans and chitin and an outer layer of glucans and mannoproteins, and its unique capsule play important roles for its integrity and virulence. We obtained information regarding the enzymes involved in the synthesis of these glycans, such as NCBI protein ID, CAZy family, and conserved protein domain family (CDD), from FungiDB [53], a database that provides tools for functional analysis and data mining against a wide range of integrated datasets.

#### Standardization of microbial glycosylation data

- Graph model for RDFization

As a first step to describe microbial glycosylation data in a standard format, we designed an RDF schema, which provides a way to capture the architecture of data description in a simple manner and to describe complex data concisely (Fig. 1). The RDF graph model was created according to the distinct data for each model microorganism. It is represented by resource nodes (data points) and edges that show the relationships between nodes and their values. Each node represents the relevant resource entities, which include concepts like a biochemical reaction or pathway for the biosynthesis of a glycoconjugate, as well as biomolecules like genes, proteins, species, references, enzymes, and glycans. To serialize the graph model as RDF documents in a text file, we adopted preexisting controlled vocabularies and hierarchical ontologies, which are recommended to improve data interoperability between different knowledge bases and data integration among multi-domain data. We also extensively searched common vocabularies and ontology terms using the OLS and Ontobee (<https://ontobee.org/>), which is a web service to help identify appropriate ontology terms or vocabularies used for annotation of resources in various domains. Additionally, to identify the definition or usage of the identified ontology, the OWL files of these ontologies were examined using Protégé software [54], which is a tool that enables the inspection of ontologies comprised of complex logic and constraints. For example, for glycoenzyme activity, it is necessary to express the concept of an enzyme exerting its activity on the substrate, resulting in the generation of a product. To describe the glycoenzyme activity, we inspected the GlycoRDF ontology using Protégé, and then introduced adequate vocabularies. To simplify the long URI comprised of a namespace and an identifier, the detailed URI can be abbreviated by specifying it with a pre-defined prefix such as *glycan* for 'http://purl.jp/bio/12/glyco/glycan#'. Using this ontology, the substrate and product glycans that take part in an enzyme reaction could be represented using a **Reaction** class node and a '*glycan:catalyzed\_by*' property, which means that the glycans are the participants of the reaction. Also, to represent their roles in the

reaction, such as reactant or product, the properties of '*glycan:has\_product*' and '*glycan:has\_substrate*' were used, respectively. For metadata annotations that provide readability, descriptive vocabulary such as '*rdfs:label*' or '*rdfs:comment*' properties were used. When the enzyme activity was provided with a descriptive explanation that was inferred from *in silico* analysis or binding assays without information about the reaction participants, the Gene Ontology (GO), a structural representation developed through the process of biocuration for the annotation of gene expression and function, was employed to assign and specify semantics to enzyme activity.

The gene information dataset for *C. jejuni* consisted of proteins involved in O-linked glycosylation of flagellin, N-glycosylation in the periplasm and membranes, and biosynthesis of the lipooligosaccharide (LOS) and capsular polysaccharide (CPS). We used GO identifiers to describe the corresponding enzyme activities that lack information about a specific chemical reaction. For instance, "Kdo transferase activity" is denoted by GO identifier GO:0043842; "flagellin subunit A" corresponds to GO:0005198; "flippase activity" is encoded by GO:0140327; and so forth. The GO ontology is useful for describing the predicted activity of a corresponding protein, and the results of advanced biochemical research will provide information about the reaction participants. In addition to enzyme activity containing glycan information, the gene, protein sequence, localization of cellular anatomy, and taxon information were also semantically connected. This allowed us to expand our scope and retrieve specific glycans or common glycans existing between microorganisms.

- Validation of RDF data

The graph model we developed was serialized into RDF sentences, a process that produces triples in the form of subject-predicate-object in accordance with the defined specification of the controlled vocabulary and ontology terms. We validated the triples using ShEx (Shape

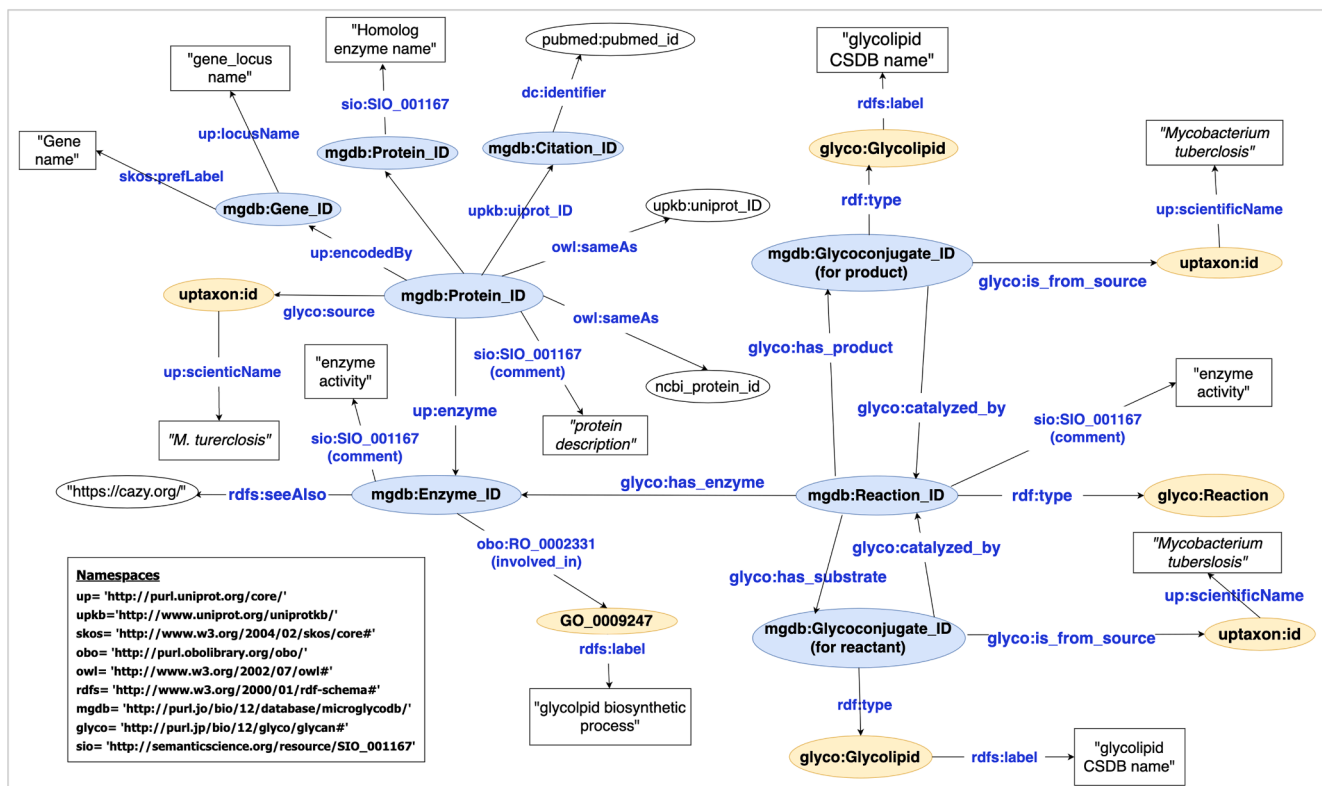


Fig. 1. The RDF schema describes glycan-related data for model organisms.

The GlycoRDF ontology is utilized to depict glycoenzymes that are responsible for catalytic reactions. The protein node is used to connect the enzyme entity to gene resources, and the GO ontology is used to characterize enzyme activity, as long as the appropriate term is available.



Expression), which enables us to resolve interpretation ambiguities raised by the RDF schema diagrams and identify data that does not match the schema [55]. The generated instances of all **Classes** were evaluated through ShEx validation, such as checking for consistency of the resource type or determining the cardinality of the number of objects that the subject instance can possess under a particular property. The validation process was iterated until all errors or discrepancies in the object values were corrected to be consistent with the ShEx outcomes, and the RDF documents were then modified. The verified RDF files were uploaded to our Virtuoso RDF database, and then, for validation, we tested whether the RDF data describing microbial glycosylation adhered to the RDF schema in the endpoint (<https://ts.glycosmos.org/sparql>).

### User interface

MicroGlycoDB was designed to provide detailed and comprehensive information on microbial glycosylation. Each microbial species has its own dedicated webpage with visual representations, which provides users with glycan structures and a wide range of glycan-related information, including the biosynthetic pathway of the glycoconjugates and their localization in the microbe. The individual page for each microorganism is accessible from the sidebar or main body of the home page (<https://microglycodb.glycosmos.org/>), where users can easily navigate or download the curated information about microbes, access useful tools for glycan drawing such as GlycoNAVI (<https://glyconavi.org/Draw/index.php>) or Drawglycan-SNFG (<http://www.virtualglycome.org/DrawGlycan/>), and links to related resources such as GlyTouCan, GlyCosmos, Rhea, and UniProt. The glycan and gene tabs on the right side of the microbe's summary page show a list of glycan structures and glycozymes encoding enzymes and proteins that are responsible for glycosylation, such as glycan synthesis, modification, or regulation. Each resource in the list has a link to a web page showing the details about the selected glycan or glycozyme. The subcategory menu on the Genes and Glycans tab depends on the type of glycoconjugates found, thereby enabling the user to retrieve specific information contained within each category. This information may also be accessed by selecting a particular segment of the membrane structures illustrated in the main figure of the selected microbe. Glycan and glycolipid structures are represented in the SNFG (Symbol Nomenclature for Glycans) format [56], which is supplied by the GlyTouCan repository, and CSDB format, respectively.

As an example of a model organism, information regarding the cellular localization of glycoconjugates in *C. jejuni* or *Mycobacterium* is provided with respect to the overall cell structure, so as to represent published reports describing particular glycostructures exist in specific cell components and are associated with their respective pathological roles. Gram-negative bacterial cell architecture, such as CPS that remain attached to the cell surface together with LOS, *N*-linked glycoproteins in the periplasm and membranes, peptidoglycans in the periplasm, and *O*-linked glycans of flagellin, are displayed on the graphic image, allowing the user to access specific information by clicking on the substructures. When clicking the components on the bacterial image, the relevant glycosylation process is displayed to the user as a graphic representation along with a table containing the genes, enzyme activities, and a link that will take the user to the detailed page of the glycans (Fig. 2A). For instance, when a user clicks on CPS, the CPS biosynthesis pathway is displayed using SNFG symbols (Fig. 2B). The details of the pathway components, such as glycozymes or carbohydrates, are simultaneously displayed on the right side, with information associated with pathogenicity, highlighting a potential target for therapeutic development.

As a second example, lipoproteins, including PIMs, LMs, and LAMs, are anchored to the plasma membrane of *M. tuberculosis*. These lipoproteins play an important role in maintaining cell membrane integrity and regulating interactions between hosts and pathogens. Data related to the location of glycoconjugates at the interface between MTb and the host cell will provide users with insights that may advance vaccine

development in the field of glycoengineering. When a user clicks on a lipoprotein, a pathway diagram representing the enzymatic process is displayed, and clicking on a glycozyme or glycan in the pathway diagram brings users to a page where they can inspect specific information on the glycan structure (Fig. 3A). If information about the glycozyme involved in the biosynthesis of the glycan structure is known, including the results of *in silico* analysis, the glycozyme information is mapped to the corresponding UniProt and Rhea IDs. Thus, the user can identify the relevant enzymatic reaction as well as integrated information about the sequence and function of the protein. The user can also access enzyme information, including the EC enzyme number, CAZy number, and corresponding enzyme reaction, by using the Rhea ID of the table that is displayed for each glycan detail. In addition, the ChEBI ID from the chemical ontology ChEBI, which provides standardized names for chemicals to reduce nomenclature confusion, appears as an external link to the glycans (Fig. 3B).

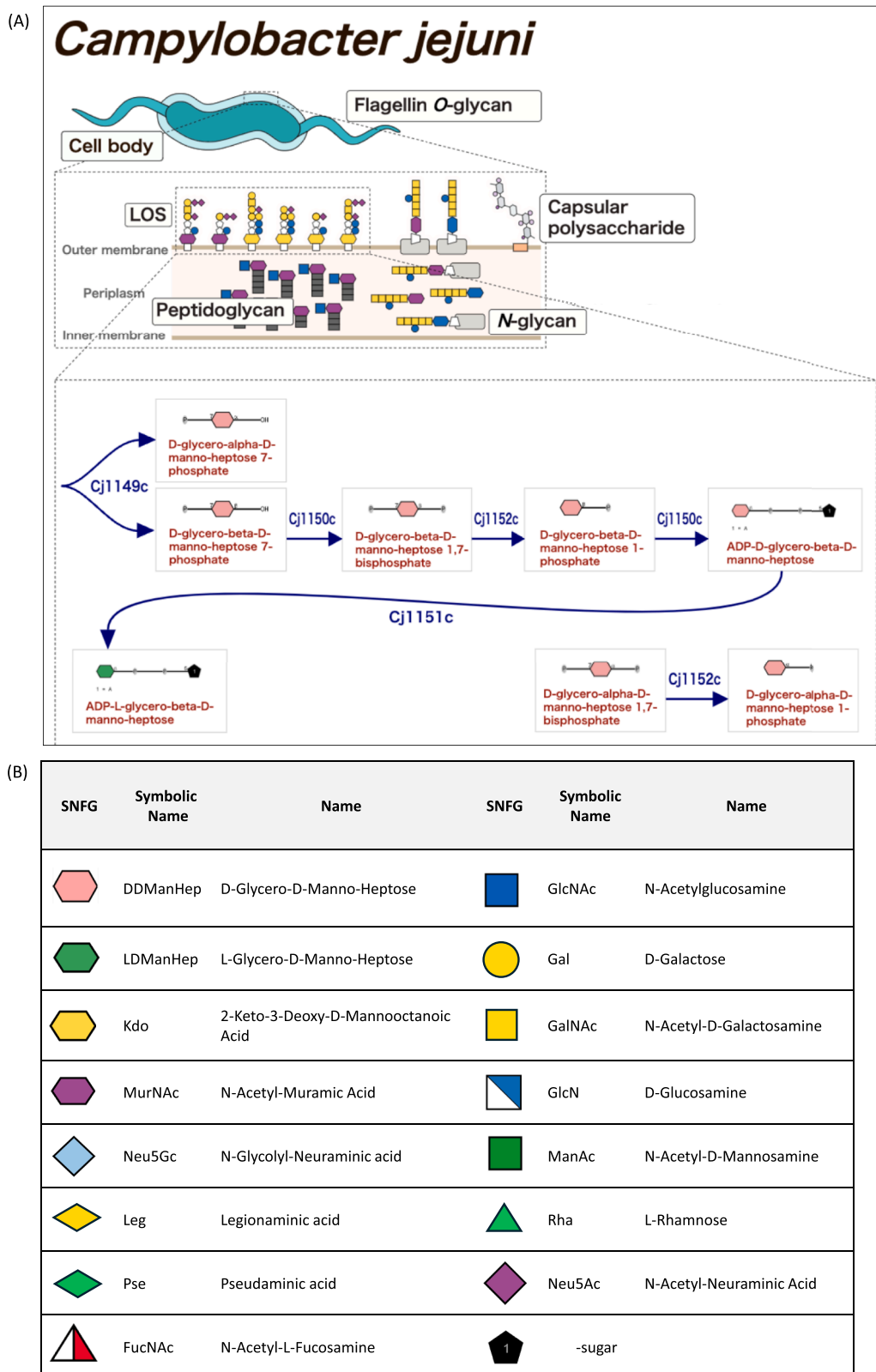
The homepage of MicroGlycoDB offers links to the entry pages of six model microorganisms, as well as a link to download the dump file, tools for searching glycans, and access to external resources. The search page has two options: selecting a species name and a keyword search pertaining to glycan structure. The latter option supports the glycan structure in four formats: CSDB linear, IUPAC condensed format, LINUCS, and WURCS. To retrieve the results using a glycan name in the keyword search, a SPARQL query may be used (Table 1). Through the glycan keyword search results, users can compare the distribution of glycans across the microbes and infer the meaning of shared glycan structures. Clicking the common glycan in the results page transfers the user to the detail page of the glycan (Fig. 4).

### Discussion

The important roles of microbial glycans in their interactions with hosts and the environment have been elucidated through various hypotheses and an extensive range of experimental evidence, including genomics, proteomics, metabolomics, and glycomics. However, the resulting data are fragmented and dispersed in publications and across multiple databases with diverse formats. Thus, the acquisition of relevant data on microbial glycans requires a significant amount of time and effort, which in turn slows down the sharing and expansion of knowledge among researchers. MicroGlycoDB has been developed to provide users with comprehensive information on glycans, specializing in the six important model microorganisms provided by a collaborative research group. Our database allows users to explore essential resources, such as enzymes and glycan structures for their glycosylation pathways, along with graphical representations of microbial structures, which helps users to investigate the functions of microbial glycans and glycan-related information.

Importantly, the glycan-related data of microbes available in MicroGlycoDB is presently restricted when considering pathogenic bacteria and microbial communities present in the human mucosa, including the gastrointestinal tract, oral cavity, skin, nasal passages, and other tissues. For example, the pathogenic *N. meningitidis* and *Neisseria gonorrhoeae* have extensive glycosylation on their capsules, LOS, and pili, which facilitate their masking of surface adhesins, promote adhesion and invasion, and enhance bacterial adherence, respectively [57]. *P. aeruginosa*, a principal pathogen responsible for nosocomial pneumonia, modifies glycosylation on its pili, resulting in enhanced virulence [58]. Also, *H. pylori* requires pseudaminic acid (Pse) glycosylation on flagella for the proper assembly of the filament and motility and its Lewis antigen mimicry within the *O*-antigen region of lipopolysaccharides for colonization, exhibiting variability in type among its strains, akin to human ABO blood group antigens [59–61]. Pse5Ac7Am in flagellin is also important to the attachment and entry of *C. jejuni* into the host intestinal epithelium [62].

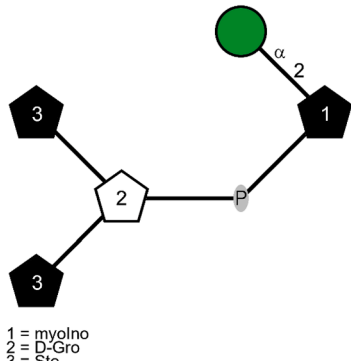
Glycosylation in non-bacterial organisms, including fungi and viruses, has frequently been shown to impact pathogenicity. For a



**Fig. 2.** Graphic depiction of cellular architecture and glycosylation processes in *C. jejuni*. (A) represents the genes and glycan constituents involved in biosynthesis of the lipooligosaccharides (LOS) that are anchored to the outer membrane of *C. jejuni*. The LOS consist of a diverse array of short oligosaccharides typically modified with sialic acid, which vary both among and within strains and contribute to immune evasion through mimicry of a wide range of human gangliosides. The user can see other structural components by clicking on membrane components such as flagellin or capsular polysaccharides, and peptidoglycan in the periplasm. (B) The SNFG format and names of the monosaccharides that take part in LOS synthesis or are found exclusively in bacteria.

### microglycodb\_glycan\_00350

Glycan Structure



1 = myoIno  
2 = D-Gro  
3 = Ste

<b>Glycan Name</b>	
<b>CSDB Linear</b>	aDManp(1-2)[[XSte(1-?)][XSte(1-?)xDGro(1-P-?)]]xXmyoIno
<b>IUPAC Condensed</b>	
<b>LINUCS</b>	
<b>WURCS</b>	
<b>Organism</b>	<i>Mycobacterium tuberculosis</i>

Glycosyltransferases
External Links

Gene ID	Gene No.	Gene Name	UniProt ID	Protein Name
<a href="#">microglycodb_gene_00350</a>	Rv2610c	<i>pimA</i>	<a href="#">P9WMZ5</a>	
<a href="#">microglycodb_gene_00351</a>	Rv2611c	<i>patA</i>	<a href="#">P9WMB5</a>	

A

### microglycodb\_gene\_00365

<b>Gene Number</b>	Rv3726c
<b>Gene Name</b>	<i>aftD</i>
<b>UniProt ID</b>	<a href="#">P96419</a>
<b>Protein Name</b>	
<b>Reviewed</b>	true
<b>Biosynthesis</b>	LAM
<b>Organism</b>	<i>Mycobacterium tuberculosis</i>

functions
sequence
publications
reaction

Involved in the biosynthesis of the arabinogalactan (AG) region of the mycolylarabinogalactan-peptidoglycan (mAGP) complex, an essential component of the mycobacterial cell wall. Catalyzes the addition of an arabinofuranosyl (Araf) residue from the sugar donor decaprenyl-phospho-arabinose (DPA) on the C-3 of an alpha-(1->5)-linked Araf from the arabinan backbone of AG.

B

Fig. 3. The information pages contain details about microbial glycans and glycosyltransferases.

(A) is a page providing comprehensive information on glycan structures, external resources, and references, allowing users to recognize various glycan formats that help in referring to or registering glycans. (B) is a gene-related details showing gene name, UniProt ID, description of the enzyme function, and enzyme reaction.

eukaryotic example, the biofilm matrix of *Candida albicans*, one of the most frequent fungal pathogens, mainly consists of  $\alpha$ -1,2 branched mannans and  $\alpha$ -1,6 mannans. This provides a physical barrier to protect from immune attack such as phagocytosis by neutrophils [63] and contributes to antifungal drug resistance [64]. For a viral example,

severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of coronavirus disease 2019 (COVID-19), shows significant *N*- and *O*-glycosylation on its proteins, including the spike protein, in the envelope which are crucial for host recognition and pathogenesis [65]. The essential role of these glycans indicates that they could be

**Table 1**

Retrieving the species containing a glycan of the keyword search.

```

SPARQL query
PREFIX glycan: <http://purl.jp/bio/12/glyco/glycan#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT DISTINCT ?g ?glycan_url ?species_name ?glycan_name
WHERE { GRAPH ?g { <http://rdf.glycosmos.org/microglycobd>}
?glycan_url rdf:type glycan:Saccharide;
rdfs:label ?species_name;
glycan:has_glycosequence ?glycosequence .
?glycosequence glycan:in_carbohydrate_format glycan:carbohydrate_format_{format};
glycan:has_sequence "{query}"xsd:string .
OPTIONAL { ?glycan_url skos:prefLabel ?glycan_name . }
}

```

potential targets for novel therapeutics, and a database providing the pathway information for glycan synthesis would therefore be extremely valuable.

In the context of disease, the significant roles of glycans in host interactions have been elucidated in infectious diseases caused by bacteria, viruses, fungi, and protozoa. Glycan-binding proteins on the immune cells of the host, including galectins, C-type lectins, and siglecs, constantly monitor surface glycans of microorganisms to discriminate between self and nonself. This is crucial in determining the fate of health or disease, such as chronic inflammation, autoimmune diseases, and cancer [66]. For instance, *C. jejuni* is also enveloped in lipooligosaccharides analogous to human gangliosides discovered in nerve tissue. Consequently, antiganglioside antibodies (AGAs) that respond to *C. jejuni*-associated glycans can give rise to an autoimmune disease known as Guillain-Barré syndrome [67]. *Mtb* responsible for tuberculosis produces various types of O-mannosylated proteins. The interaction of ManLAM with C-type lectin receptors (CLRs), including dendritic cell-specific intercellular adhesion molecule-3-grabbing nonintegrin (DC-SIGN), mannose receptor (MR), and dectins, has been linked to enhanced host immunity against *Mtb* infection [68,69]. The human immunodeficiency virus (HIV) is extensively glycosylated with oligomannose-GlcNAc to protect against neutralizing antibodies, similar to many other viruses, and viral Man5-GlcNAc complexes facilitate HIV attachment to host cells [71]. SARS-CoV-2 has further demonstrated the significance of glycans as critical components in the interface of virus-host interactions [70]. The emerging fungus, *Candida auris*, which exhibits severe multidrug resistance, expresses a cell wall  $\beta$ -glucan that influences the immune response [72]. Increased understanding of the interactions between glycan-binding proteins and glycans is being cataloged in several databases, and we plan to incorporate microbial interaction data into these databases. Overall, this highlights the significance of a comprehensive database that provides information on genes, enzymes, glycans, pathways, associated diseases, and other relevant factors pertaining to the role of glycans in host-microorganism interactions in both health and disease in a standard format.

We describe the resources of MicroGlycoDB using ontology and controlled vocabulary in RDF format, which is appropriate for data description clarity and data integration. Ontologies not only clarify the uncertainty arising from the diverse nomenclature of biological molecules, as seen in the multitude of synonyms, but they also assist the discovery of new knowledge by inferring and representing the semantic relationships between different types of data. We are currently developing a new ontology to clearly define the complex structures of glycoconjugates that consist of many chemical modifications, glycosyl moieties, and backbone molecules in LPS, flagella, pili, and capsules. Thus, MicroGlycoDB will help researchers to discover not only the glycan structures of interest, providing details such as genes, enzymes, and related pathways within the microbial anatomy, but also the connections between glycan structures and their roles, gain insight into complex glycosylation processes, and understand the mechanisms that

underlie interactions between microorganisms and hosts.

Some resources or external references for glycan structures or glycoconjugates in our database are limited. For instance, *C. neoformans* shows only glycoconjugates without the information on the glycans. This is due to the lack of mapping data for glycoenzymes expected at the gene level and the glycan structures that mostly remain in publications. As one way to resolve this issue, we are developing an automated pipeline to inspect and verify the information that is uploaded, via a Web tool called MicroGlycoCurator (<https://microglycorepo.alpha.glycosmos.org/>), an online platform that transforms tabular data into semantic data, and subsequently conveys it to MicroGlycoDB, which is particularly useful for handling large data sets. This system will reduce the effort and errors associated with manual processes while facilitating the integration of individual data within a single microbe into a knowledgebase where all data sets are linked as semantic data by using ontologies. This Web tool will be released as an alpha version in 2025, with public release via GlyCosmos the following year. Just as we have been holding various user workshops and booth exhibits at conferences for GlyCosmos, we will also advertise MicroGlycoCurator and MicroGlycoDB to the community for feedback, which will be incrementally implemented in these respective resources on a periodic basis. The major challenge will be to develop a template format that is compatible with data formats that are currently used by the community. This will require close communications with researchers to develop an extendable and useful template so users can easily upload their information into MicroGlycoCurator and to edit the information online for submission. Another concern is the annotations that users may want to attach to their data and whether it will be straightforward to map them to standardized ontologies. Again, these will be discussed closely with the relevant members of the community. We expect that MicroGlycoDB will encompass additional microbial species and offer more detailed information through our curator system.

## Conclusion

We have developed MicroGlycoDB to provide a comprehensive data resource on microbial glycosylation. Although some resources have missing data that needs to be filled in through additional studies, our database emphasizes the important role of integrated knowledge, which assists users to inspect the consolidated information that comes from diverse domains. This database is structured to facilitate and promote data integration, so that we can provide not only specific information about individual glycans but also pertinent information such as glycan structures in various formats and their standardized identifier numbers for easy reference, glycoconjugates, glycoenzymes, glycosylation pathways, and their graphical representation in cellular architecture. This reduces the effort required for users to search across multiple websites and provides an opportunity to implement an integrated approach to glycosylation systems in microbes. MicroGlycoDB, a comprehensive database on microbial glycosylation, thus lays the foundation for a platform that will provide insights into microbial glycosylation by



## Search for glycan structure

Search

CSDB Linear  
 IUPAC Condensed  
 LINUCS  
 WURCS

CSDB Linear

IUPAC Condensed

LINUCS

WURCS

Search: P-7)aXDDmanHepp(1-P  
Format: CSDB Linear  
1 result found

**microglycodb\_glycan\_00071**

- D-glycero-alpha-D-manno-heptose 1,7-bisphosphate

ChEBI: [60207](#)  
GlyYouCan: [G66933WP](#)

<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 5px;"> <b>Bifidobacterium bifidum</b> <table style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 60%;">Gene</td><td style="text-align: center; border: 1px solid #ccc; border-radius: 50%; width: 20px;">0</td></tr> <tr><td>Glycan</td><td style="text-align: center; border: 1px solid #ccc; border-radius: 50%;">0</td></tr> </table> </div>	Gene	0	Glycan	0	<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 5px;"> <b>Bifidobacterium longum</b> <table style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 60%;">Gene</td><td style="text-align: center; border: 1px solid #ccc; border-radius: 50%; width: 20px;">0</td></tr> <tr><td>Glycan</td><td style="text-align: center; border: 1px solid #ccc; border-radius: 50%;">0</td></tr> </table> </div>	Gene	0	Glycan	0	<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 5px;"> <b>Campylobacter jejuni</b> <table style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 60%;">Gene</td><td style="text-align: center; border: 1px solid #ccc; border-radius: 50%; width: 20px;">0</td></tr> <tr><td style="color: blue;">Glycan</td><td style="text-align: center; border: 1px solid #ccc; border-radius: 50%; color: blue;">1</td></tr> </table> </div>	Gene	0	Glycan	1
Gene	0													
Glycan	0													
Gene	0													
Glycan	0													
Gene	0													
Glycan	1													
<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 5px;"> <b>Cryptococcus neoformans</b> <table style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 60%;">Gene</td><td style="text-align: center; border: 1px solid #ccc; border-radius: 50%; width: 20px;">0</td></tr> <tr><td>Glycan</td><td style="text-align: center; border: 1px solid #ccc; border-radius: 50%;">0</td></tr> </table> </div>	Gene	0	Glycan	0	<div style="border: 1px solid #ccc; padding: 5px; margin-bottom: 5px;"> <b>Mycobacterium abscessus</b> <table style="width: 100%; border-collapse: collapse;"> <tr><td style="width: 60%;">Gene</td><td style="text-align: center; border: 1px solid #ccc; border-radius: 50%; width: 20px;">0</td></tr> <tr><td>Glycan</td><td style="text-align: center; border: 1px solid #ccc; border-radius: 50%;">0</td></tr> </table> </div>	Gene	0	Glycan	0					
Gene	0													
Glycan	0													
Gene	0													
Glycan	0													

**Fig. 4.** Keyword search interface based on SPARQL queries.

The upper box is for the search of relevant glycan information; this allows users to search using a glycan structure as a keyword with options for glycan format, including CSDB Linear, IUPAC, and WURCS. The results are presented in the selection box showing the species harboring the glycan, which is realized by executing SPARQL queries on the database server.

integrating independent information from publications and databases.

#### Funding sources

This work was supported by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST) Grant Number JPMJND2204 and the Soka University International Collaborative Research Grant. Doering lab studies of cryptococcal glycans are supported by National Institutes of Health grant AI135012.

#### CRedit authorship contribution statement

**Sunmyoung Lee:** Writing – review & editing, Writing – original draft, Resources. **Louis-David Leclercq:** Validation, Resources. **Yann**

**Guerardel:** Writing – review & editing, Resources. **Christine M. Szymanski:** Writing – review & editing, Resources. **Thomas Hurtaux:** Validation, Resources. **Tamara L. Doering:** Writing – review & editing, Funding acquisition, Conceptualization. **Takane Katayama:** Writing – review & editing, Investigation. **Kiyotaka Fujita:** Writing – review & editing, Investigation. **Kazuhiro Aoki:** Writing – review & editing, Investigation. **Kiyoko F. Aoki-Kinoshita:** Writing – review & editing, Resources, Funding acquisition, Conceptualization.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

We would like to thank Miyuki Kikuchi and Atsuto Uchino for their initial contributions to the development of MicroGlycoDB. We thank Liza Loza for assistance in depicting the cryptococcal cell wall and Cory Wenzel, Michel Gilbert, Ian Schoenhofen and Frank Raushel for assistance with *C. jejuni* glycan pathway annotations. We would also like to acknowledge Yotsuba Hattori for updating the user interface to its current version.

## Data availability

Data will be made available on request.

## References

- [1] K.C. Huang, R. Mukhopadhyay, B. Wen, Z. Gitai, N.S. Wingreen, Cell shape and cell-wall organization in Gram-negative bacteria, *Proc. Natl. Acad. Sci. U.S.A.* 105 (2008) 19282–19287.
- [2] C. Whitfield, D.M. Williams, S.D. Kelly, Lipopolysaccharide O-antigens—bacterial glycans made to measure, *J. Biol. Chem.* 295 (2020) 10593–10609.
- [3] L. Yakovlieva, J.A. Fülleborn, M.T.C. Walvoort, Opportunities and challenges of bacterial glycosylation for the development of novel antibacterial strategies, *Front. Microbiol.* 12 (2021) 745702.
- [4] H. Sanchez, D. Hopkins, S. Demirdjian, C. Gutierrez, G.A. O'Toole, S. Neelamegham, B. Berwin, Identification of cell-surface glycans that mediate motility-dependent binding and internalization of *Pseudomonas aeruginosa* by phagocytes, *Mol. Immunol.* 131 (2021) 68–77.
- [5] H. Hiyoshi, T. Wangdi, G. Lock, C. Saechao, M. Raffatellu, B.A. Cobb, A.J. Bäuml, Mechanisms to evade the phagocyte respiratory burst arose by convergent evolution in typhoidal salmonella serovars, *Cell Rep.* 22 (2018) 1787–1797.
- [6] T. Wangdi, C.-Y. Lee, A.M. Spees, C. Yu, D.D. Kingsbury, S.E. Winter, C.J. Haste, R.P. Wilson, V. Heinrich, A.J. Bäuml, The Vi capsular polysaccharide enables salmonella enterica serovar typhi to evade microbe-guided neutrophil chemotaxis, *PLoS Pathog.* 10 (2014) e1004306.
- [7] S.M. Huszczyński, C. Coumoundouros, P. Pham, J.S. Lam, C.M. Khursigara, Unique regions of the polysaccharide copolymerase Wzz<sub>2</sub> from *Pseudomonas aeruginosa* are essential for O-specific antigen chain length control, *J. Bacteriol.* (2019) 201.
- [8] T. Fiebig, J.T. Cramer, A. Beth, P. Baruch, U. Curth, J.I. Fühling, F.F.R. Buettner, U. Vogel, M. Schubert, R. Fedorov, M. Mühlhoff, Structural and mechanistic basis of capsule O-acetylation in *Neisseria meningitidis* serogroup A, *Nat. Commun.* 11 (2020) 4723.
- [9] E. Geisinger, W. Huo, J. Hernandez-Bird, R.R. Isberg, *Acinetobacter baumannii*: envelope determinants that control drug resistance, virulence, and surface variability, *Annu. Rev. Microbiol.* 73 (2019) 481–506.
- [10] R. Raman, K. Tharakaraman, V. Sasisekharan, R. Sasisekharan, Glycan–protein interactions in viral pathogenesis, *Curr. Opin. Struct. Biol.* 40 (2016) 153–162.
- [11] E. Barreto-Bergter, R.T. Figueiredo, Fungal glycans and the innate immune recognition, *Front. Cell. Infect. Microbiol.* 4 (2014).
- [12] A.S. Luis, G.C. Hansson, Intestinal mucus and their glycans: a habitat for thriving microbiota, *Cell Host Microb.* 31 (2023) 1087–1100.
- [13] S.I. Gringhuis, J. Den Dunnen, M. Litjens, M. Van Der Vlist, T.B.H. Geijtenbeek, Carbohydrate-specific signaling through the DC-SIGN signalosome tailors immunity to *Mycobacterium tuberculosis*, HIV-1 and *Helicobacter pylori*, *Nat. Immunol.* 10 (2009) 1081–1088.
- [14] J. Poole, C.J. Day, M. Von Itzstein, J.C. Paton, M.P. Jennings, Glycointeractions in bacterial pathogenesis, *Nat. Rev. Microbiol.* 16 (2018) 440–452.
- [15] A. Alemka, H. Nothaft, J. Zheng, C.M. Szymanski, N-glycosylation of campylobacter jejuni surface proteins promotes bacterial fitness, *Infect. Immun.* 81 (2013) 1674–1682.
- [16] R.S. Houlston, E. Vinogradov, M. Dzieciatkowska, J. Li, F. St. Michael, M.-F. Karwaski, D. Brochu, H.C. Jarrell, C.T. Parker, N. Yuki, R.E. Mandrell, M. Gilbert, Lipooligosaccharide of *Campylobacter jejuni*, *J. Biol. Chem.* 286 (2011) 12361–12370.
- [17] C.J. Day, E.A. Semchenko, V. Korolik, Glycoconjugates play a key role in campylobacter jejuni infection: interactions between host and pathogen, *Front. Cell. Inf. Microbiol.* 2 (2012).
- [18] S. Abouelhadid, J. Raynes, T. Bui, J. Cuccui, B.W. Wren, Characterization of posttranslationally modified multidrug efflux pumps reveals an unexpected link between glycosylation and antimicrobial resistance, *mBio* 11 (2020) e02604–e02620.
- [19] M. Schirm, I.C. Schoenhofen, S.M. Logan, K.C. Waldron, P. Thibault, Identification of unusual bacterial glycosylation by tandem mass spectrometry analyses of intact proteins, *Anal. Chem.* 77 (2005) 7774–7782.
- [20] P. Castric, F.J. Cassels, R.W. Carlson, Structural characterization of the *Pseudomonas aeruginosa* 1244 pilin glycan, *J. Biol. Chem.* 276 (2001) 26479–26485.
- [21] M. Schirm, E.C. Soo, A.J. Aubry, J. Austin, P. Thibault, S.M. Logan, Structural, genetic and functional characterization of the flagellin glycosylation process in *Helicobacter pylori*, *Mol. Microbiol.* 48 (2003) 1579–1592.
- [22] Y.A. Knirel, E.Th. Rietschel, R. Marre, U. Zähringer, The structure of the O-specific chain of *Legionella pneumophila* serogroup 1 lipopolysaccharide, *Eur. J. Biochem.* 221 (1994) 239–245.
- [23] M.-Y. Mistou, I.C. Sutcliffe, N.M. van Sorge, Bacterial glycobiology: rhamnose-containing cell wall polysaccharides in Gram-positive bacteria, *FEMS Microbiol. Rev.* 40 (2016) 464–479.
- [24] J. Lodowska, D. Wolny, L. Weglarz, The sugar 3-deoxy-D-manno-oct-2-ulonic acid (Kdo) as a characteristic component of bacterial endotoxin – a review of its biosynthesis, function, and placement in the lipopolysaccharide core, *Can. J. Microbiol.* 59 (2013) 645–655.
- [25] J. Horzempa, T.K. Held, A.S. Cross, D. Furst, M. Qutyan, A.N. Neely, P. Castric, Immunization with a *Pseudomonas aeruginosa* 1244 pilin provides O-antigen-specific protection, *Clin. Vaccine Immunol.* 15 (2008) 590–597.
- [26] A. Singh, S. Bajar, A. Devi, D. Pant, An overview on the recent developments in fungal cellulase production and their industrial applications, *Bioresour. Technol. Rep.* 14 (2021) 100652.
- [27] G. Amaro Bittencourt, L. Porto De Souza Vandenberghe, K. Valladares-Diestra, L. Wedderhoff Herrmann, A. Fátima Murawski De Mello, Z. Sarmiento Vázquez, S. Grace Karp, C. Ricardo Soccol, Soybean hulls as carbohydrate feedstock for medium to high-value biomolecule production in biorefineries: a review, *Bioresour. Technol.* 339 (2021) 125594.
- [28] J.S. Lam, V.L. Taylor, S.T. Islam, Y. Hao, D. Kocincová, Genetic and functional diversity of *Pseudomonas aeruginosa* lipopolysaccharide, *Front. Microbiol.* 2 (2011).
- [29] J. Li, A. Martin, A.D. Cox, E.R. Moxon, J.C. Richards, P. Thibault, Mapping bacterial glycolipid complexity using capillary electrophoresis and electrospray mass spectrometry. *Methods in Enzymology*, Elsevier, 2005, pp. 369–397.
- [30] M.E. Mnich, R. Van Dalen, N.M. Van Sorge, C-type lectin receptors in host defense against bacterial pathogens, *Front. Cell. Infect. Microbiol.* 10 (2020) 309.
- [31] A. Bateman, UniProt: a worldwide hub of protein knowledge, *Nucl. Acid. Res.* 47 (2019) D506–D515.
- [32] V. Lombard, H. Golaconda Ramulu, E. Drula, P.M. Coutinho, B. Henrissat, The carbohydrate-active enzymes database (CAZy) in 2013, *Nucl. Acid. Res.* 42 (2014) D490–D495.
- [33] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L. L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman, Pfam: the protein families database in 2021, *Nucl. Acid. Res.* 49 (2021) D412–D419.
- [34] P.V. Toukach, K.S. Egorova, Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts, *Nucl. Acid. Res.* 44 (2016) D1229–D1236.
- [35] S.K. Burley, H.M. Berman, G.J. Kleywegt, J.L. Markley, H. Nakamura, S. Velankar, Protein Data Bank (PDB): the single global macromolecular structure archive, *Method. Mol. Biol.* 1607 (2017) 627–641.
- [36] A. Fujita, N.P. Aoki, D. Shinmachi, M. Matsubara, S. Tsuchiya, M. Shiota, T. Ono, I. Yamada, K.F. Aoki-Kinoshita, The international glycan repository GlyTouCan version 30, *Nucl. Acid. Res.* 49 (2021) D1529–D1533.
- [37] P. de Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, C. Steinbeck, Chemical Entities of Biological Interest: an update, *Nucl. Acid. Res.* 38 (2010) D249–D254.
- [38] P. Bansal, A. Morgat, K.B. Axelsen, V. Muthukrishnan, E. Coudert, L. Aimò, N. Hyka-Nouspikel, E. Gasteiger, A. Kerhornou, T.B. Neto, M. Pozzato, M.-C. Blatter, A. Ignatchenko, N. Redaschi, A. Bridge, Rhea, the reaction knowledgebase in 2022, *Nucl. Acid. Res.* 50 (2022) D693–D700.
- [39] K. Tanaka, K.F. Aoki-Kinoshita, M. Kotera, H. Sawaki, S. Tsuchiya, N. Fujita, T. Shikanai, M. Kato, S. Kawano, I. Yamada, H. Narimatsu, WURCS: the Web3 unique representation of carbohydrate structures, *J. Chem. Inf. Model.* 54 (2014) 1558–1566.
- [40] S. Decker, P. Mitra, S. Melnik, Framework for the semantic Web: an RDF tutorial, *IEEE Internet Comput.* 4 (2000) 68–73.
- [41] D.L. McGuinness, F. Van Harmelen, OWL web ontology language overview, *W3C Recommendation.* 10 (2004) 2004.
- [42] A. Seaborn, E. Prud'hommeaux, SPARQL query language for RDF W3C Recommendation, *World Wide Web Consort.* 15 (2008). January.
- [43] S. Lee, T. Ono, K. Aoki-Kinoshita, RDFizing the biosynthetic pathway of *E. coli* O-antigen to enable semantic sharing of microbiology data, *BMC Microbiol.* 21 (2021) 325.
- [44] OpenLink Software, *Virtuoso Universal Server*, (2018).
- [45] R. Ranzinger, K.F. Aoki-Kinoshita, M.P. Campbell, S. Kawano, T. Lütke, S. Okuda, D. Shinmachi, T. Shikanai, H. Sawaki, P. Toukach, M. Matsubara, I. Yamada, H. Narimatsu, GlycoRDF: an ontology to standardize glycomics data in RDF, *Bioinformatics* 31 (2015) 919–925.
- [46] I. Yamada, M.P. Campbell, N. Edwards, L.J. Castro, F. Lisacek, J. Mariethoz, T. Ono, R. Ranzinger, D. Shinmachi, K.F. Aoki-Kinoshita, The glycoconjugate ontology (GlycoCoO) for standardizing the annotation of glycoconjugate data and its application, *Glycobiology* 31 (2021) 741–750.
- [47] S. Yao, Z. Zhao, W. Wang, X. Liu, *Bifidobacterium Longum*: protection against Inflammatory Bowel Disease, *J. Immunol. Res.* (2021) 1–11, 2021.
- [48] T. Katoh, C. Yamada, M.D. Wallace, A. Yoshida, A. Gotoh, M. Arai, T. Maeshibu, T. Kashima, A. Hagenbeek, M.N. Ojima, H. Takada, M. Sakanaka, H. Shimizu, K. Nishiyama, H. Ashida, J. Hirose, M. Suarez-Diez, M. Nishiyama, I. Kimura, K. A. Stubbs, et al., A bacterial sulfoglycosidase highlights mucin O-glycan breakdown in the gut ecosystem, *Nat. Chem. Biol.* 19 (2023) 778–789.
- [49] M. Sakanaka, M.E. Hansen, A. Gotoh, T. Katoh, K. Yoshida, T. Odamak, H. Yachi, Y. Sugiyama, S. Kurihara, J. Hirose, T. Urashima, J. Xiao, M. Kitaoka, S. Fukiya, A. Yokota, L. Lo Leggio, M. Abou Hachem, T. Katayama, Evolutionary adaptation

- in fucosylactose uptake systems supports bifidobacteria-infant symbiosis, *Sci. Adv.* 5 (2019) eaaw7696.
- [50] A. Maitra, T. Munshi, J. Healy, L.T. Martin, W. Vollmer, N.H. Keep, S. Bhakta, Cell wall peptidoglycan in *Mycobacterium tuberculosis*: an Achilles' heel for the TB-causing pathogen, *FEMS Microbiol. Rev.* 43 (2019) 548–575.
- [51] M.J. Catalão, S.R. Filipe, M. Pimentel, Revisiting anti-tuberculosis therapeutic strategies that target the peptidoglycan structure and synthesis, *Front. Microbiol.* 10 (2019) 190.
- [52] J.Y. Kam, E. Hortle, E. Krogman, S.E. Warner, K. Wright, K. Luo, T. Cheng, P. Manuneechi Cholan, K. Kikuchi, J.A. Triccas, W.J. Britton, M.D. Johansen, L. Kremer, S.H. Oehlers, Rough and smooth variants of *Mycobacterium abscessus* are differentially controlled by host immunity during chronic infection of adult zebrafish, *Nat. Commun.* 13 (2022) 952.
- [53] E. Basenko, J. Pulman, A. Shanmugasundram, O. Harb, K. Crouch, D. Starns, S. Warrenfeltz, C. Aurrecoechea, C. Stoerckert, J. Kissinger, D. Roos, C. Hertz-Fowler, FungiDB: an integrated bioinformatic resource for fungi and oomycetes, *JoF* 4 (2018) 39.
- [54] T. Tudorache, C. Nyulas, N.F. Noy, M.A. Musen, WebProtégé: a collaborative ontology editor and knowledge acquisition tool for the Web, *Semant. Web* 4 (2013) 89–99.
- [55] K. Thornton, H. Solbrig, G.S. Stupp, J.E. Labra Gayo, D. Mietchen, E. Prud'hommeaux, A. Waagmeester, et al., Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation, in: P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A.J.G. Gray, V. Lopez, et al. (Eds.), *The Semantic Web*, Springer International Publishing, Cham, 2019, pp. 606–620.
- [56] A. Varki, R.D. Cummings, M. Aebi, N.H. Packer, P.H. Seeberger, J.D. Esko, P. Stanley, G. Hart, A. Darvill, T. Kinoshita, J.J. Prestegard, R.L. Schnaar, H. H. Freeze, J.D. Marth, C.R. Bertozzi, M.E. Etzler, M. Frank, J.F. Vliegenthart, T. Lütke, S. Perez, et al., Symbol nomenclature for graphical representations of glycans, *Glycobiology* 25 (2015) 1323–1324.
- [57] T.D. Mubaiwa, E.A. Semchenko, L.E. Hartley-Tassell, C.J. Day, M.P. Jennings, K. L. Seib, The sweet side of the pathogenic *Neisseria*: the role of glycan interactions in colonisation and disease, *Pathog. Dis.* (2017) 75.
- [58] J.G. Smedley, E. Jewell, J. Roguskie, J. Horzempa, A. Syboldt, D.B. Stolz, P. Castric, Influence of pilin glycosylation on *Pseudomonas aeruginosa* 1244 Pilus Function, *Infect. Immun.* 73 (2005) 7922–7931.
- [59] M.A. Heneghan, C.F. McCarthy, A.P. Moran, Relationship of blood group determinants on *Helicobacter pylori* lipopolysaccharide with host lewis phenotype and inflammatory response, *Infect. Immun.* 68 (2000) 937–941.
- [60] M. Chmiela, Structural modifications of *Helicobacter pylori* lipopolysaccharide: an idea for how to live in peace, *WJG* 20 (2014) 9882.
- [61] A.I.M. Salah Ud-Din, A. Roujeinikova, Flagellin glycosylation with pseudaminic acid in *Campylobacter* and *Helicobacter*: prospects for development of novel therapeutics, *Cell. Mol. Life Sci.* 75 (2018) 1163–1178.
- [62] P. Guerry, C.M. Szymanski, *Campylobacter* sugars sticking out, *Trend. Microbiol.* 16 (2008) 428–435.
- [63] D. Sandai, Y.M. Tabana, A.E. Ouweini, I.O. Ayodeji, Resistance of *Candida albicans* biofilms to drugs and the host immune system, *Jundishap. J. Microbiol.* 9 (2016).
- [64] J. Kaur, C.J. Nobile, Antifungal drug-resistance mechanisms in *Candida* biofilms, *Curr. Opin. Microbiol.* 71 (2023) 102237.
- [65] Y. Gong, S. Qin, L. Dai, Z. Tian, The glycosylation in SARS-CoV-2 and its receptor ACE2, *Sig. Transduct. Target Ther.* 6 (2021) 396.
- [66] L.I. Crouch, C.S. Rodrigues, C.R. Bakshani, L. Tavares-Gomes, J. Gaifem, S.S. Pinho, The role of glycans in health and disease: regulators of the interaction between gut microbiota and host immune system, *Semin. Immunol.* 73 (2024) 101891.
- [67] R.K. Yu, S. Usuki, T. Ariga, Ganglioside molecular mimicry and its pathological roles in Guillain-Barré syndrome and related diseases, *Infect. Immun.* 74 (2006) 6517–6527.
- [68] L. Jia, S. Sha, S. Yang, A. Taj, Y. Ma, Effect of protein O-mannosyltransferase (MSMEG\_5447) on *M. smegmatis* and its survival in macrophages, *Front. Microbiol.* 12 (2021) 657726.
- [69] I. Alves, A. Fernandes, B. Santos-Pereira, C.M. Azevedo, S.S. Pinho, Glycans as a key factor in self and nonself discrimination: impact on the breach of immune tolerance, *FEBS Lett.* 596 (2022) 1485–1502.
- [70] A. Breiman, N. Ruvoën-Clouet, M. Deleers, T. Beauvais, N. Jouand, J. Rocher, N. Bovin, N. Labarrière, H. El Kenz, J. Le Pendu, Low levels of natural anti- $\alpha$ -N-acetylgalactosamine (Tn) antibodies are associated with COVID-19, *Front. Microbiol.* 12 (2021) 641460.
- [71] B.L. Spillings, C.J. Day, A. Garcia-Minambres, A. Aggarwal, N.D. Condon, T. Haselhorst, D.F.J. Purcell, S.G. Turville, J.L. Stow, M.P. Jennings, J. Mak, Host glycoalyx captures HIV proximal to the cell surface via oligomannose-GlcNAc glycan-glycan interactions to support viral entry, *Cell Rep.* 38 (2022) 110296.
- [72] S.M.G. Selisana, X. Chen, E. Mahfudhoh, A. Bowolaksono, A. Rozaliyani, K. Orihara, S. Kajiwaru, Alteration of  $\beta$ -glucan in the emerging fungal pathogen *Candida auris* leads to immune evasion and increased virulence, *Med. Microbiol. Immunol.* 213 (2024) 13.