Title: Enhancing the interoperability of glycan data flow between ChEBI, PubChem, and GlyGen.

Rahi Navelkar[1*], Gareth Owen[2], Venkatesh Mutherkrishnan[2], Paul Thiessen[3], Tiejun Cheng[3], Evan Bolton[3], Nathan Edwards[4], Michael Tiemeyer[5], Matthew P Campbell[6], Maria Martin[7], Jeet Vora[1], Robel Kahsay[1], Raja Mazumder[1]

*corresponding author

[1]The Department of Biochemistry and Molecular Biology, George Washington University Medical Center, Washington, DC 20037

[2]Cheminformatics and Metabolism, European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

[3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

[4]Department of Biochemistry and Cellular Biology, Georgetown University, Washington, DC 20007, USA.

[5]Complex Carbohydrate Research Center, Department of Biochemistry and Molecular Biology, and Department of Chemistry, University of Georgia, Athens, GA 30602, USA.

[6]Institute for Glycomics, Griffith University, Gold Coast, Australia, 4215, Australia

[7]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom

Email address:

Rahi Navelkar: rsn13@gwu.edu

Gareth Owen: gowen@ebi.ac.uk

Venkatesh Mutherkrishnan: venkat@ebi.ac.uk

Paul Thiessen: thiessen@ncbi.nlm.nih.gov

Evan Bolton: bolton@ncbi.nlm.nih.gov

Tiejun Cheng: chengt2@ncbi.nlm.nih.gov

Nathan Edwards: nje5@georgetown.edu

Michael Tiemeyer: mtiemeyer@ccrc.uga.edu

Matthew Campbell: m.campbell2@griffith.edu.au

Maria Martin: martin@ebi.ac.uk

Jeet Vora: jeetvora@gwu.edu

Robel Kahsay: rykahsay@gwu.edu

Raja Mazumder: mazumder@gwu.edu

Mailing address (corresponding author)

1021 Arlington Blvd, Apt 642, Arlington, VA 22209, USA

ORCID IDs:

Rahi Navelkar 0000-0002-4200-7409

Gareth Owen 0000-0001-7625-5776

Venkatesh Mutherkrishnan 0000-0002-6850-9888

Evan Bolton 0000-0002-5959-6190

Paul Thiessen 0000-0002-1992-2086

Tiejun Cheng 0000-0002-4486-3356

Nathan Edwards 0000-0001-5168-3196

Michael Tiemeyer 0000-0002-8704-9143

Matthew P Campbell 0000-0002-9525-792X

Maria Martin 0000-0001-5454-2815

Jeet Vora 0000-0002-5317-1458x

Raja Mazumder 0000-0001-8823-9945

Running Head: Glycan data integration in bioinformatics databases

ABSTRACT

Glycans have a critical role in health and disease. As a direct result, there is keen interest to identify and increase glycan data in bioinformatics databases like ChEBI and PubChem, and in connecting them to biomedical resources at the EMBL-EBI and NCBI. GlyTouCan is a comprehensive archival database that contains over 118,000 glycans, obtained primarily through batch upload from glycan repositories, submissions from glycoprotein databases, and individual laboratories. In many instances, the glycan structures deposited in GlyTouCan may not be fully defined or have supporting experimental evidence or biological source content. Databases like ChEBI and PubChem were designed to accommodate complete atomistic structures with well-defined chemical linkages. As a result, they cannot easily accommodate the structural ambiguity inherent in many glycan database representations. Consequently, there is a need to organize glycan data coherently to improve the connectivity across the major NCBI, EMBL-EBI, and glycoscience databases.

This paper outlines a workflow developed in collaboration between GlyGen, ChEBI, and PubChem to improve the visibility and connectivity of glycan data across these resources. GlyGen hosts a subset of glycans (~29,000) from the GlyTouCan database and has submitted valuable glycan annotations to the PubChem database and registered or mapped over 10,500 (including ambiguously defined) glycans into the ChEBI database. The integrated glycans were prioritized based on links to PubChem and connectivity to glycoprotein data. The pipeline provides a blueprint for how glycan

data can be harmonized between different resources. The current PubChem, ChEBI, and GlyTouCan mappings can be downloaded from GlyGen (https://data.glygen.org).

INTRODUCTION

1.1 Background

Glycosylation is one of the most abundant and bio-medically impactful post-translational modifications of proteins, but the immense diversity of structures generated through this process often hinders the efforts to link structure with essential functions, such as transcription, translation, cell division, cellular differentiation, tissue development, cell adhesion, cell-cell interactions, protein folding and stability, inflammatory responses, and many others(Cummings and Pierce 2014, Varki and Gagneux 2015, Varki 2017). One of the characteristics of protein glycosylation is microheterogeneity, which refers to the fact that an individual glycoprotein can exist in many glycoforms. These glycoforms differ from each other based on the exact glycan structures found at each glycosylation site. Microheterogeneity not only increases the difficulty of describing protein glycosylation, but also contributes to the modulation of protein activities, making it crucial to wholistically understand the key aspects of the glycosylation process (such as the expressing cell type, the specific glycosylation, site on a protein, and the detected glycan heterogeneity) together as inter-related features(Varki and Gagneux 2015). Thus, it is important to extract the glycosylation information from the available literature, translate it accurately as value-added data, and make it publicly available to the biomedical community to facilitate further research. One of the most common problems currently facing the bioinformatics community is the availability of fully-determined glycan structures (where there is complete information of the arrangement of the monosaccharides, linkages/anomeric configuration and glycosidic bonds with the exception where ambiguity regarding the first linkage between glycan and conjugate is tolerated) in literature. Even though it is expected that the glycan analysis provides topology information, it is difficult to transfer such data from literature to bioinformatics databases without a curator's effort to evaluate composition and other supplementary information to generate fully determined structures. As a result the knowledge transfer from literature to bioinformatics resources is lengthened significantly.

Among the available glycan and glycoconjugate database resources available, GlyTouCan(Tiemeyer, Aoki et al. 2017) (designed to serve a function analogous to what GenBank serves for nucleotide sequences) provides a repository of glycans which are fully-defined or with structural ambiguity at multiple levels which can described by the Glycan naming and subsumption ontology or GNOme ((http://obofoundry.org/ontology/gno, http://gnome.glyomics.org) (See Figure 1). Depositions made in GlyTouCan may also include monosaccharide compositions without any structural context or structural representations without complete description of the linkages between their monosaccharide building blocks. In addition to providing accessions for individual glycans (which are adopted as primary identifiers by glycoprotein-centric databases such as GlyGen(York, Mazumder et al. 2020), GlyConnect(Alocci, Mariethoz et al. 2019) and UniCarbKB(Campbell, Peterson et al. 2014)), GlyTouCan also provides cross-references to contributing databases and includes other annotations such as sequences, literature evidence, and Symbol Nomenclature for Glycans (SNFG) images. Similarly, chemical databases like

ChEBI(Degtyarenko, de Matos et al. 2008) and PubChem(Kim, Chen et al. 2021) provide unique identifiers for chemical entities (glycans) and host text-string representation (such as SMILES, InChI, LINUCS, etc.), 2-D structure, and literature evidence for compounds. These databases also host annotation and cross-references to reaction databases Rhea(Lombardot, Morgat et al. 2019)) and pathway database Reactome(Jassal, Matthews et al. 2020) that show the individual compounds that take part in any given reaction (i.e. the reactants and products). Additionally, ChEBI also provides ontology (http://www.obofoundry.org/ontology/chebi.html) for chemical structures which is used by a number of research groups (including Gene Ontology(Ashburner, Ball et al. 2000, Gene Ontology 2021) and model species group).

Additionally, GlyGen(York, Mazumder et al. 2020) is a glycoinformatics resource with extensive glycoprotein and glycan data integrated from various databases. GlyGen presents a dedicated entry page for proteins/glycoproteins (e.g. https://www.glygen.org/protein/P14210-1) and glycans (e.g. https://www.glygen.org/glycan/G17689DH). To facilitate interoperability, GlyGen uses accessions from UniProtKB(UniProt 2019) and GlyTouCan as primary accessions to identify proteins and glycans respectively. Glycan pages present cross-references to other glycoprotein databases (like GlyConnect, UniCarbKB, etc.) and other glycan or chemical databases (such as PubChem, MatrixDB(Chautard, Ballut et al. 2009), Glycosciences.DB(Bohm, Bohne-Lang et al. 2019), etc.), unique digital representation (such as GlycoCT(Herget, Ranzinger et al. 2008), WURCS(Matsubara, Aoki-Kinoshita et al. 2017), and IUPAC), harmonized symbolic nomenclature representations (SNFG(Varki, Kannagi et al. 2015)), and additional annotations which are of importance to glycan-focused researchers, including glycan classifications (N-linked, O-linked, complex core 1, etc), relevant supporting publications, as well as annotations related to disease, expression, mutation, etc.

It is important to enhance the connectivity across GlyTouCan, ChEBI and PubChem databases by promoting interoperability of glycan data available at GlyGen, provided by its project partners. This will allow GlyGen to connect glycans to glycoproteins, reaction and pathway annotations and ontologies provided by GNOme and ChEBI. This will allow users to navigate the same glycan and its associated annotations across multiple resources through one portal. Previously, UniCarbKB demonstrated the importance of connecting glyco-centric databases with PubChem with submission of ~1,700 GlycoSuiteDB entries.

1.2 Data flow of glycans across different bioinformatics databases:

GlyTouCan and ChEBI databases routinely submit entities to the PubChem Substance database as MDL SDF files. These contain the chemical structures (as embedded molfiles), names, synonyms and other information for glycans. PubChem assigns a unique Substance Identifier (SID) to each submitted glycan entity that are standardized to the PubChem Compound database (identified by CID) such that each SID is mapped to a single CID. Multiple SIDs can map to the same CID (see Figure 2).

PubChem only accepts glycans where the arrangement of the monosaccharides, glycosidic bonds and linkages or anomeric positions are known. However, ambiguity regarding the first linkage (between the glycan and conjugate) and stereochemistry of any particular atom is tolerated within PubChem. As a result, not all GlyTouCan accessions are included in the PubChem database. Additionally, before the start of the collaboration only a small portion of the available glycans in PubChem were mapped to the ChEBI database resulting in low connectivity across GlyTouCan, PubChem and ChEBI databases (See Figure 3). This dataflow is utilized by GlyGen to submit glycan annotations to PubChem, connect GlyTouCan accessions to ChEBI IDs, and increase the PubChem CID to ChEBI ID mapping.

RESULTS

2.1 Expanding glycan data coverage in ChEBI

Out of the total GlyGen glycan set (of 29,290 glycans) with links to PubChem or glycans with structural ambiguity (reported on large number of glycoproteins) were prioritized for the integration process in ChEBI. For this release ChEBI has ~10,500 glycans registered or mapped semi-automatically by GlyGen (see Figure 4) that includes 10,121 fully-determined and 25 high-value (i.e reported on large number of glycoproteins, see Table 1) ambiguously defined glycans. The integrated glycans belong to different subsumption categories defined by the Glycan Naming Ontology (GNOme) with the majority of glycans (10,394) belonging to the saccharide and few to other categories like topologies, compositions and base compositions (refer to Figure 1 for a description of these categories). The remaining glycans from the GlyGen set which represent predominately ambiguously defined structures will be manually registered in batches in the ChEBI database routinely. The integrated glycans are annotated with structure data (for glycans with a PubChem CID cross-reference), formula, charge, mass values, InChI, InChIKey and SMILES. Additionally, new ChEBI names (while retaining the original name as a synonym), more precise ontology terms (such as “amino tetrasaccharide” or acetamides”) are added to generate 3-star curated status entries. The corresponding GlyGen, GlyTouCan cross-references are added under “Manual X-ref” whereas PubChem IDs are listed under “Automatic X-ref” sections of the individual ChEBI entry pages respectively (e.g.

https://www.ebi.ac.uk/chebi/searchId.do?chebiId=70967). The mappings are available to download from GlyGen (https://data.glygen.org/GLY_000296) or ChEBI (ftp://ftp.ebi.ac.uk/pub/databases/chebi/Flat_file_tab_delimited/database_accession.tsv) or can be viewed through individual GlyGen pages (e.g. https://glygen.org/glycan/G78059CC#Cross-References).

2.2 Submitting Glycan annotations to PubChem:

GlyGen submitted important glycan annotations such as glycan motif (e.g. Lewis X, Type 2 LN), composition (Hex5 HexNAc4 dHex1), per-methylated mass (2221.13523366), classification (e.g. N-glycan (complex)) to PubChem Compound database using the GlyTouCan accession to PubChem CID mapping (e.g. https://pubchem.ncbi.nlm.nih.gov/compound/25098607#section=Biologic-Description) Additionally, GlyGen also submitted glycosylation annotation such as the glycosylation site, glycosylation type (e.g. N-linked), GlyTouCan accession, literature evidence and a glycosylation summary which list all glycosylation sites from GlyGen with or without glycan information. For example, for PubChem protein entry for the Epidermal growth factor receptor (UniProt: P00533), the summary describes total of 18 sites out of which 15 sites have 101 N glycans, 1 site has 1 O-glycan and 2 sites do not have any reported glycan information. The table only displays glycosylation sites where the corresponding glycan has a PubChem CID mapping. (e.g. https://pubchem.ncbi.nlm.nih.gov/protein/P00533#section=Glycosylation). Additionally, the GlyTouCan accession mapped to PubChem CID promotes interoperability within

PubChem Protein and Compound pages. All the annotations submitted by GlyGen can be found on PubChem's data sources page (https://pubchem.ncbi.nlm.nih.gov/source/23201)

2.3 Increasing ChEBI cross-reference in PubChem

Initially, only a small portion of glycans available in the PubChem database were mapped to ChEBI IDs. Due to this collaboration effort as the glycan data coverage increased in ChEBI more glycans were submitted to the PubChem database through ChEBI's routine submissions. As a result, the number of PubChem compounds with a ChEBI cross-reference increased over time (see Figure 5). However, this process does not include submission of ambiguously-defined glycans into PubChem.

2.4 Identifying problematic PubChem CID to CHEBI ID mapping:

During the integration process, it was observed that limited number of ChEBI IDs were mapped to the same PubChem CID, resulting in one-to-many PubChem CID to ChEBI ID mappings (see Figure 6). Such mappings were observed for polymeric structures in ChEBI where the repeating part of the monomeric structure was represented in square brackets with an "n" multiplier outside of the brackets. However, such polymeric features were ignored when such entries were processed by PubChem. As a result, polymeric entries in ChEBI were mapped to the same PubChem CID as the monomeric entry. It was observed that the polymeric entries in ChEBI did not have any automatically generated “text string” identifiers (such as SMILES, InChIs or InChIKeys) therefore, these PubChem CIDs were mapped to the correct ChEBI ID by matching the InChI strings associated with both PubChem CID and ChEBI ID. The resolved mappings are currently available through GlyGen (e.g.: https://glygen.org/glycan/G19577LJ)

2.5 Registration of high-value ambiguously defined glycans:

As a proof of concept twenty-five high-value ambiguously defined glycans were registered into ChEBI manually (see Table 1). These glycans are selected based on the rules described in method section 4.2 (*Manual registration of glycans into ChEBI*). The ChEBI entries are registered without images (due to high number of possible linkages) and are assigned ChEBI names with a prefix of its source database i.e. “GlyTouCan” (e.g. GlyTouCan G06110VR) in this case (e.g. https://www.ebi.ac.uk/chebi/searchId.do?chebiId=156559). Other related details (such as definition, ontology and WURCS) are manually added to convert such entries from registered 2-star status to 3-star status.

2.6 Generating new ontology classes within ChEBI:

To accommodate ambiguously defined glycans within ChEBI, new ontology classes were generated using the Glycan Naming Ontology (or GNOme). The new ontology classes are equivalent to the base-composition (https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:167481), composition (https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:167502) and topology (https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:167503) instances or subsumption categories described by GNOme. Additionally, the new ontology classes include the appropriate corresponding cross-references to GNOme.

DISCUSSION

3.1 Novelty

GlyTouCan, PubChem and ChEBI are major bioinformatics resources that are routinely used by many different research groups. However, the ChEBI and PubChem databases have significantly less glycan data compared to GlyTouCan. This novel pipeline allows integration of valuable glycan and glycan-related annotations hosted by GlyGen into PubChem and ChEBI databases while maintaining the links to GlyTouCan database. Additionally, ChEBI's routine submissions to PubChem allows the newly integrated glycans to be further integrated into the PubChem database. As a result, this pipeline is not only able to identify and increase the glycan data coverage in ChEBI and PubChem databases but is also able to synchronize and map these entries ensuring connectivity across these major resources. Additionally, GlyGen allows users to access such cross-references under one portal while connecting the glycans to glycoprotein data.

3.2 Applications

Submitting glycan annotations to UniProt database

Currently, the glycosylation data available in the UniProt database describes the protein, glycosylation site, associated literature (for reported glycosylation) and in some cases additional description (such as high mannose, etc) about the glycan or glycosylation. However it lacks the information on the specific glycan that may be reported on a particular site on the protein. This information will be provided by GlyGen (sourced from multiple resources) where specific glycans will be added to the protein and site data already present in the UniProt database. The glycans along with a 2-D image will be identified by a ChEBI ID and will be cross-referenced to GlyGen protein pages. The data will be accessed through the ProtVista viewer in UniProt (see Figure 7). ProtVista is a protein browser providing a compact representation and access to the functional amino acid residue annotation in UniProt such as domains, sites, post-translational modifications and variants. Protvista is available through the UniProt web site and the source code is publicly available through as a web component (https://github.com/ebi-webcomponents/protvista-uniprot). UniProt users will also be able to query for glycans connecting them to knowledge on proteins e.g., proteins that bind them and enzymes that synthesize or degrade them.

Connecting to other important resources

Mapping of GlyTouCan accession to ChEBI ID allows further extension to other major databases such as Rhea (enzyme-centric) or Reactome (pathway-centric) or UniCarbKB, GlyConnect, (or other glycoprotein-centric) resources which use ChEBI or GlyTouCan accessions as their primary identifiers (See Figure 8). This will allow `direct mapping. The mapped cross-references to Rhea and Reactome databases can be viewed through GlyGen glycan pages (currently on beta: https://beta.glygen.org/glycan/G96881BQ#Cross-References). Additionally, as GNOme also provides glycan-glycan relationships (using GlyTouCan accession as the primary identifier), the ambiguously defined glycans can be potentially linked to possible fully-determined, structurally-related glycans within the ChEBI database.

3.3 Future plans

In order to sustain the growing GlyGen glycan collection, one of the major goals is to automate ~90% of the submission pipeline and reduce the time required to manually register ambiguously defined glycans. We also plan to document this pipeline so that resources like GlyTouCan, GlyConnect, and others can utilize this pipeline to integrate their glycans into the ChEBI and the PubChem databases. Additional plans include the generation of tutorials so that users can directly register or map ambiguously-defined glycans into the ChEBI database. This can be achieved by using the PubMed ID or other publication or database identifier as a ChEBI name for the glycan entries in order to comply with the ChEBI registration protocols. Just as the usefulness of genomic and proteomic resources grew as their coverage of genes and proteins expanded, so too will the usefulness of glycan and glycosylation databases increase as the number of well-mapped glycan structures approaches comprehensive coverage. For glycans, reaching this goal has added difficulty due to the chemical nature of the compounds and the inherent limitations of the approaches for their structural analysis. Therefore, well developed submission pipelines that capture and cross-link as much information as possible are essential for growth in the glycosciences and for connecting glycan data to broader bioinformatics domains.

MATERIALS AND METHODS

The GlyTouCan database hosts over 118,000 glycans, each identified by a GlyTouCan accession. The GlyGen glycan set (v1.5.13) comprises only a subset (~24%) of the total GlyTouCan glycan set.

4.1 Rules to generate GlyGen glycan set

The GlyGen glycan set is seeded with GlyTouCan glycan accessions annotated as human (NCBI Tax ID: 9606), mouse (NCBI Tax ID: 10090), rat (NCBI Tax ID: 10116), HCV1a (NCIB TaxID: 63746), SARS CoV 2 (NCBI TaxID:2697049) by GlyTouCan, UniCarbKB, or GlyConnect glycan data-resources; glycans representing motifs from the GlycoMotif (http://glycomotif.glyomics.org) glycan motif data-resource; GlyGen synthetic glycans; and glycans observed in human contexts in the GPTwiki (http://gptwiki.glyomics.org) glycopeptide transition data-resource. The GlyGen glycan set is then extended to include any GlyTouCan glycan accession that subsumes or shares a monosaccharide composition with a seed-glycan using the GNOme glycan naming and subsumption ontology (http://obofoundry.org/ontology/gno, http://gnome.glyomics.org). (See Figure 9). GlyGen data processing and integration framework is utilized to quality control and standardized this data which is supplied to GlyGen portal(Kahsay, Vora et al. 2020).

As an attempt to harmonize glycans between GlyTouCan and ChEBI, the GlyGen glycan set (v.1.5.13) comprising 29,290 glycans was identified as the primary list (see Figure 10). An integration pipeline was developed to map and register all GlyGen glycans into the ChEBI database. (see Figure 11).

4.2 Integration of glycans in ChEBI

Using PubChem as an anchor database.

PubChem Compound was used as an anchor database to identify the existing mapping between the GlyTouCan and ChEBI databases. Out of the GlyGen glycan list of 29,290 GlyTouCan accessions 10,496 were already cross-referenced to PubChem compound identifiers (CIDs). These mostly included glycans which are fully-determined with or without known first linkage (between the glycan and conjugate). This set also includes glycans which consists of a single monosaccharide as well glycans with one or more undefined atomic stereocenter. This mapping was utilized to retrieve the corresponding PubChem CID to ChEBI ID mapping. As a result, over 1,700 glycans were mapped between the GlyTouCan and ChEBI databases (See Figure 12). However, this method provided only about 5% coverage for the GlyGen glycan set.

As for the remaining PubChem CIDs where a ChEBI ID mapping was not directly available from the PubChem database, the integration pipeline was modified so that the structure data for individual entries (downloaded from PubChem) could be uploaded into applications like KNIME (Fillbrunn, Dietz et al. 2017) and ClassyFire (Djoumbou Feunang, Eisner et al. 2016) to generate a suitable SD file for ChEBI.

In this process, the downloaded PubChem structure data for individual entries were uploaded to the KNIME application where the data was first cleaned (e.g. remove explicit hydrogens while maintaining the stereochemical integrity of the structure) and the additional information (such as associated names, synonyms and WURCS sequence) was extracted. The data was further uploaded to the ClassyFire application (at 12 requests per minute rate) in order to generate ontology terms. The generated ontology terms were then converted to the best matching ChEBI ontology terms and an SD file for ChEBI upload was generated. A default classification value (such as carbohydrates and carbohydrates-derivatives) was added in cases where ClassyFire was unable to generate an ontology term for a record. The files generated by the KNIME application were submitted to the ChEBI loader tool (in batches of 250 records) and the GlyTouCan accession (via PubChem CID) was either mapped to an existing ChEBI ID (if the structure was already present in ChEBI) or registered as a new entry with 2-star status. This method allowed integration of any GlyTouCan accession (within GlyGen) with a PubChem CID into the ChEBI database (see Figure 13).

Manual registration of glycans into ChEBI

The PubChem Compound database facilitated mapping or registration of roughly 30% of the GlyGen glycans (GlyTouCan accessions) into the ChEBI database. The remaining set of 18,794 glycans lacked a PubChem CID, likely as they did not fit the

PubChem requirements. This includes (but not restricted to) glycans that strictly compositions (e.g. G00058XR or G00057LF) or glycans with undefined stereochemistry (e.g G00318AV or G00012RZ). However, this set also includes about 3256 fully-determined glycans which lack a PubChem CID probably due to a technical error or a database synchronization issue between PubChem and the GlyTouCan database.

To establish and optimize a protocol, twenty-five high-value glycans (GlyTouCan accession) were selected to be manually registered into ChEBI (see Table 1) as a proof of concept. The subset was identified based on the following conditions: 1) Top-ten composition glycans with the largest number of associated proteins in GlyGen (v1.5.36). 2) Top-ten ambiguously-defined glycans with the largest number of associated proteins in GlyGen (v1.5.36) and. 3) Five composition glycans associated to the HCV1a protein in GlyGen (v1.5.36). Based on the following approaches the remaining glycans will be registered manually: 1) ***Fully determined glycans:*** Such glycans can be treated as ordinary chemical entities in the same manner as most of the other entities in the ChEBI database. Entries can be added by a ChEBI curator using the Curator tool, or by individual submitters using the ChEBI submissions tool (in which a submitter provides a name, structure, etc.). Each submission is immediately visible on the public web site as a 2-star (unchecked) entry and is upgraded to 3-star (checked) status after being checked by a ChEBI curator. After being indexed overnight, submissions can be fully searched, for example, by name, synonyms, InChI, InChIKey, SMILES, and structure, the following day. 2) ***Glycans with unknown stereochemistry:*** In principle, all of the methods described for fully determined structures can be used in cases where some (or all!) of the stereochemistry is unknown. If the alpha/beta configuration of one or more glycosidic bonds in a glycan is not known, then such bonds are depicted in the chemical structure as normal single bonds. For double bonds of unknown configuration, 'wavy' bonds can be used to link the double bond to one or more of its substituents. 3) ***Glycans with unknown position of attachment:*** If the position of attachment of a glycosyl group to a glycan is not known, then the structure is drawn using "R" atoms to indicate the possible sites of attachment. In the Definition field, a note is included to indicate the possible values of R at each position 4) ***Compositions:*** In cases where only the composition such as the type (hexose, pentose, etc.) and number of monosaccharides is known then such glycans can be registered using the source (in this case GlyTouCan) accession as a ChEBI name and the composition of a glycan (e.g. Composition: Hex2 HexNAc5.) can be added as part of the description. However, due to high number of possible linkages, such entries would lack a ChEBI image.

4.3 Integration of glycan annotations to PubChem

A sustainable approach was developed to submit glycan and glycoprotein annotations from GlyGen to PubChem Compound and protein pages. The approach uses stable unique URLs of annotation specific GlyGen datasets (e.g. https://data.glygen.org/GLY_000283) which are regularly and automatically consumed by the PubChem database.

PubChem Compound (Glycan) Pages

Each compound in PubChem has a dedicated page and is identified with a unique compound identifier (or CID) which represents a standardized entry from multiple substances submitted to PubChem. Utilizing the existing GlyTouCan accession to PubChem CID mapping and GlyGen's annotation specific dataset, annotations such as glycan motif (https://data.glygen.org/GLY_000283), mass (https://data.glygen.org/GLY_000281) and composition (https://data.glygen.org/GLY_000286) were added to the respective PubChem Compound page. The annotations were added under the 'Biologic Description' section on the PubChem Compound page along with the URL for the GlyGen glycan page. The individual monosaccharides within the glycan composition are further linked to respective entries within PubChem Compound pages.

PubChem Protein Pages

Similar to the Compound pages, PubChem has dedicated pages for the proteins identified by the NCBI Protein accessions. Utilizing the UniProtKB accession, GlyGen provided the respective glycoprotein annotations (using dataset https://data.glygen.org/GLY_000499) to the PubChem Protein pages. The annotations are added under the 'Glycobiology' section along with a link to GlyGen Protein pages. The section also contains a table that shows glycosylation information including glycosylation site, glycosylation type, GlyTouCan ID, glycan image, PubChem CID, and PMID evidence for only a subset of glycoprotein annotations where the associated glycan (GlyTouCan accession) has a corresponding PubChem CID for a given glycoprotein.

FUNDING

This work was supported by the GlyGen project, a National Institutes of Health Common Fund grant [grant number U01 GM125267-01 to R.N, G.O, V.M, N.E, M.T, M.C, M.M, J.V, R.M]; and the Intramural Research Program of the National Library of Medicine, National Institutes of Health ‘to [E.B, P.T, T.C]’.

ACKNOWLEDGEMENTS

Kiyoko Aoki-Kinoshita, Terence Murphy and Andrew Leach

ABBREVIATIONS

EMBL-EBI: EMBL-European Bioinformatics Institute

SIB: Swiss Institute of Bioinformatics

ChEBI: Chemical Entities of Biological Interest

CFG: Consortium for Functional Glycomics

GNOme: Glycan Naming Ontology

CID: PubChem Compound Identifier

SID: PubChem Substance Identifier

WURCS: Web3 Unique Representation of Carbohydrate Structures

SMILES: Simplified Molecular-Input Line-Entry System

InChI: International Chemical Identifier

IUPAC: International Union of Pure and Applied Chemistry

HCV: Hepatitis C virus

X-ref: Cross-reference

DATA AVAILABILITY STATEMENT

The data can be openly accessed through the GlyGen database at

https://www.glygen.org/ and https://data.glygen.org/

REFERENCES

Alocci, D., J. Mariethoz, A. Gastaldello, E. Gasteiger, N. G. Karlsson, D. Kolarich, N. H. Packer and F. Lisacek (2019). "GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical." J Proteome Res **18**(2): 664-677.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Bohm, M., A. Bohne-Lang, M. Frank, A. Loss, M. A. Rojas-Macias and T. Lutteke (2019). "Glycosciences.DB: an annotated data collection linking glycomics and proteomics data (2018 update)." Nucleic Acids Res **47**(D1): D1195-D1201.

Campbell, M. P., R. Peterson, J. Mariethoz, E. Gasteiger, Y. Akune, K. F. Aoki-Kinoshita, F. Lisacek and N. H. Packer (2014). "UniCarbKB: building a knowledge platform for glycoproteomics." Nucleic Acids Res **42**(Database issue): D215-221.

Chautard, E., L. Ballut, N. Thierry-Mieg and S. Ricard-Blum (2009). "MatrixDB, a database focused on extracellular protein-protein and protein-carbohydrate interactions." Bioinformatics **25**(5): 690-691.

Cummings, R. D. and J. M. Pierce (2014). "The challenge and promise of glycomics." Chem Biol **21**(1): 1-15.

Degtyarenko, K., P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj and M. Ashburner (2008). "ChEBI: a database and ontology for chemical entities of biological interest." Nucleic Acids Res **36**(Database issue): D344-350.

Djoumbou Feunang, Y., R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner and D. S. Wishart (2016). "ClassyFire: automated chemical classification with a comprehensive, computable taxonomy." J Cheminform **8**: 61.

Fillbrunn, A., C. Dietz, J. Pfeuffer, R. Rahn, G. A. Landrum and M. R. Berthold (2017). "KNIME for reproducible cross-domain analysis of life science data." J Biotechnol **261**: 149-156.

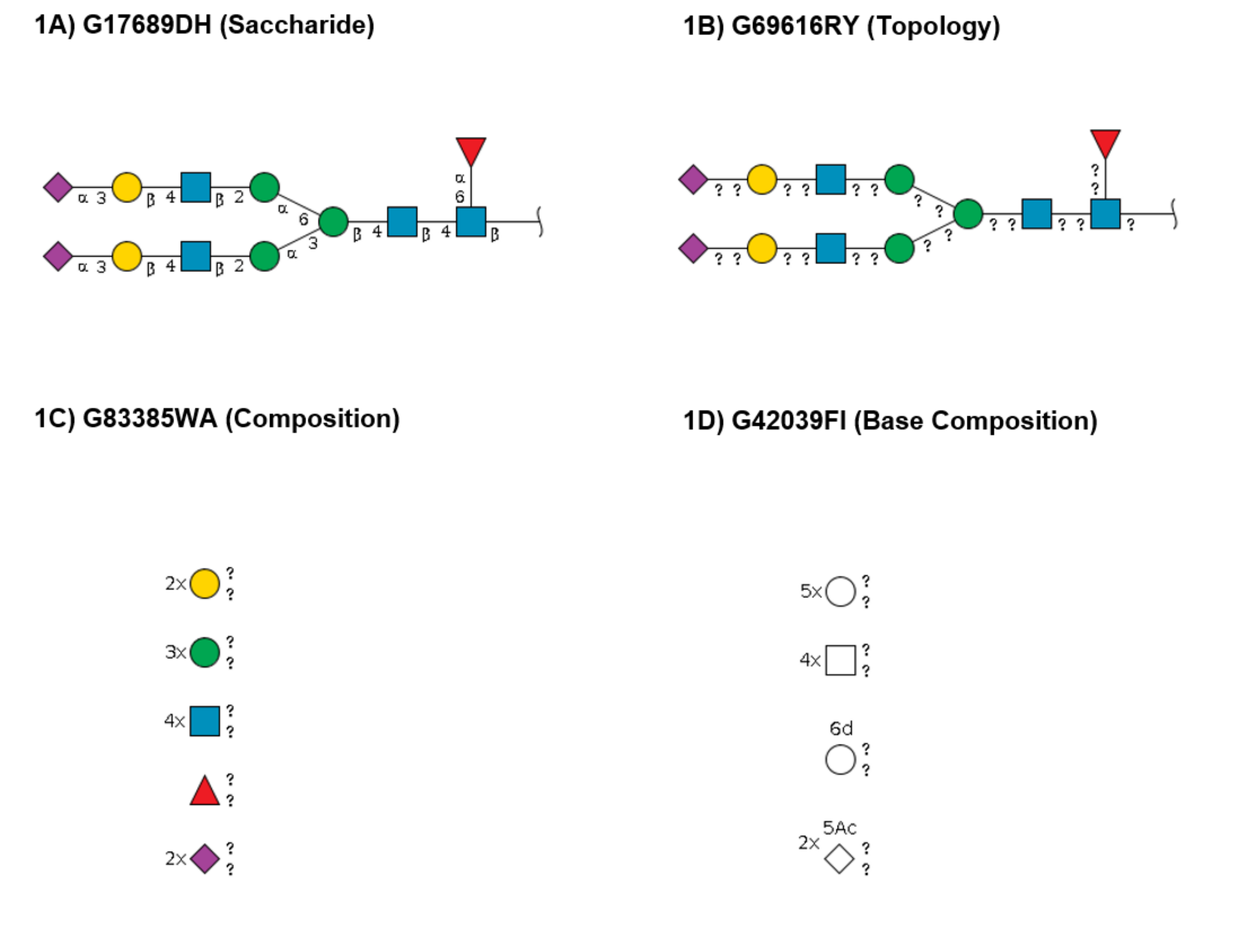
Gene Ontology, C. (2021). "The Gene Ontology resource: enriching a GOld mine." Nucleic Acids Res **49**(D1): D325-D334.

Herget, S., R. Ranzinger, K. Maass and C. W. Lieth (2008). "GlycoCT-a unifying sequence format for carbohydrates." Carbohydr Res **343**(12): 2162-2171.

Jassal, B., L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob and P.

D'Eustachio (2020). "The reactome pathway knowledgebase." Nucleic Acids Res **48**(D1): D498-D503.
Kahsay, R., J. Vora, R. Navelkar, R. Mousavi, B. C. Fochtman, X. Holmes, N. Pattabiraman, R. Ranzinger, R. Mahadik, T. Williamson, S. Kulkarni, G. Agarwal, M. Martin, P. Vasudev, L. Garcia, N. Edwards, W. Zhang, D. A. Natale, K. Ross, K. F. Aoki-Kinoshita, M. P. Campbell, W. S. York and R. Mazumder (2020). "GlyGen data model and processing workflow." Bioinformatics **36**(12): 3941-3943.
Kim, S., J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton (2021). "PubChem in 2021: new data content and improved web interfaces." Nucleic Acids Res **49**(D1): D1388-D1395.
Lombardot, T., A. Morgat, K. B. Axelsen, L. Aimo, N. Hyka-Nouspikel, A. Niknejad, A. Ignatchenko, I. Xenarios, E. Coudert, N. Redaschi and A. Bridge (2019). "Updates in Rhea: SPARQLing biochemical reaction data." Nucleic Acids Res **47**(D1): D596-D600.
Matsubara, M., K. F. Aoki-Kinoshita, N. P. Aoki, I. Yamada and H. Narimatsu (2017). "WURCS 2.0 Update To Encapsulate Ambiguous Carbohydrate Structures." J Chem Inf Model **57**(4): 632-637.
Tiemeyer, M., K. Aoki, J. Paulson, R. D. Cummings, W. S. York, N. G. Karlsson, F. Lisacek, N. H. Packer, M. P. Campbell, N. P. Aoki, A. Fujita, M. Matsubara, D. Shinmachi, S. Tsuchiya, I. Yamada, M. Pierce, R. Ranzinger, H. Narimatsu and K. F. Aoki-Kinoshita (2017). "GlyTouCan: an accessible glycan structure repository." Glycobiology **27**(10): 915-919.
UniProt, C. (2019). "UniProt: a worldwide hub of protein knowledge." Nucleic Acids Res **47**(D1): D506-D515.
Varki, A. (2017). "Biological roles of glycans." Glycobiology **27**(1): 3-49.
Varki, A. and P. Gagneux (2015). Biological Functions of Glycans. Essentials of Glycobiology. rd, A. Varki, R. D. Cummings et al. Cold Spring Harbor (NY)**:** 77-88.
Varki, A., R. Kannagi, B. Toole and P. Stanley (2015). Glycosylation Changes in Cancer. Essentials of Glycobiology. rd, A. Varki, R. D. Cummings et al. Cold Spring Harbor (NY)**:** 597-609.
York, W. S., R. Mazumder, R. Ranzinger, N. Edwards, R. Kahsay, K. F. Aoki-Kinoshita, M. P. Campbell, R. D. Cummings, T. Feizi and M. Martin (2020). "GlyGen: computational and informatics resources for glycoscience." Glycobiology **30**(2): 72-73.
York, W. S., R. Mazumder, R. Ranzinger, N. Edwards, R. Kahsay, K. F. Aoki-Kinoshita, M. P. Campbell, R. D. Cummings, T. Feizi, M. Martin, D. A. Natale, N. H. Packer, R. J. Woods, G. Agarwal, S. Arpinar, S. Bhat, J. Blake, L. J. G. Castro, B. Fochtman, J. Gildersleeve, R. Goldman, X. Holmes, V. Jain, S. Kulkarni, R. Mahadik, A. Mehta, R. Mousavi, S. Nakarakommula, R. Navelkar, N. Pattabiraman, M. J. Pierce, K. Ross, P. Vasudev, J. Vora, T. Williamson and W. Zhang (2020). "GlyGen: Computational and Informatics Resources for Glycoscience." Glycobiology **30**(2): 72-73.

FIGURES:



1A) G17689DH (Saccharide)

1B) G69616RY (Topology)

1C) G83385WA (Composition)

1D) G42039FI (Base Composition)

Figure 1: Different types of glycans (identified by a GlyTouCan accession) represented in an SNFG format with varying levels of knowledge about its stereochemistry, glycosidic bonds, anomeric configuration, etc. The types of glycan shown in the figure also describe different instances or subsumption categories in the Glycan Naming Ontology (http://www.obofoundry.org/ontology/gno.html) or GNOme. 1a) Shows a type of glycan (G17689DH) where the arrangement (type, number, and stereochemistry) of monosaccharides is known along with the glyosidic bonds with partial or complete information about the linkage or anomeric configuration. Such glycans are categorized

under “Saccharide” (http://purl.obolibrary.org/obo/GNO_00000016) instance in GNOme. 1b) Shows a type of glycan (G69616RY) where there is no information about the linkage or anomeric configuration, but the arrangement of monosaccharides and glycosidic bonds is known. Such glycans belong to the “Topology” (http://purl.obolibrary.org/obo/GNO_00000015) instance in GNOme. 1c) Shows a type of glycan (G83385WA) where only the information about type and number of monosaccharides is known in addition to complete or partial knowledge of the stereochemistry of the involved monosaccharides. Such glycans belong to the “composition” (http://purl.obolibrary.org/obo/GNO_00000014) instance in GNOme. Certain glycans from these three instances are also identified as fully-determined glycans where there is complete knowledge of the arrangement of the monosaccharides, glycosidic bonds, and linkage/anomeric configuration with the exception of first linkage (between glycan and conjugate) where the ambiguity is tolerated. 1d) Shows a type of glycan (G42039FI) where only the information about the type and number of monosaccharides is known. Such glycans are categorized under the “basecomposition” (http://purl.obolibrary.org/obo/GNO_00000013) instance in GNOme.

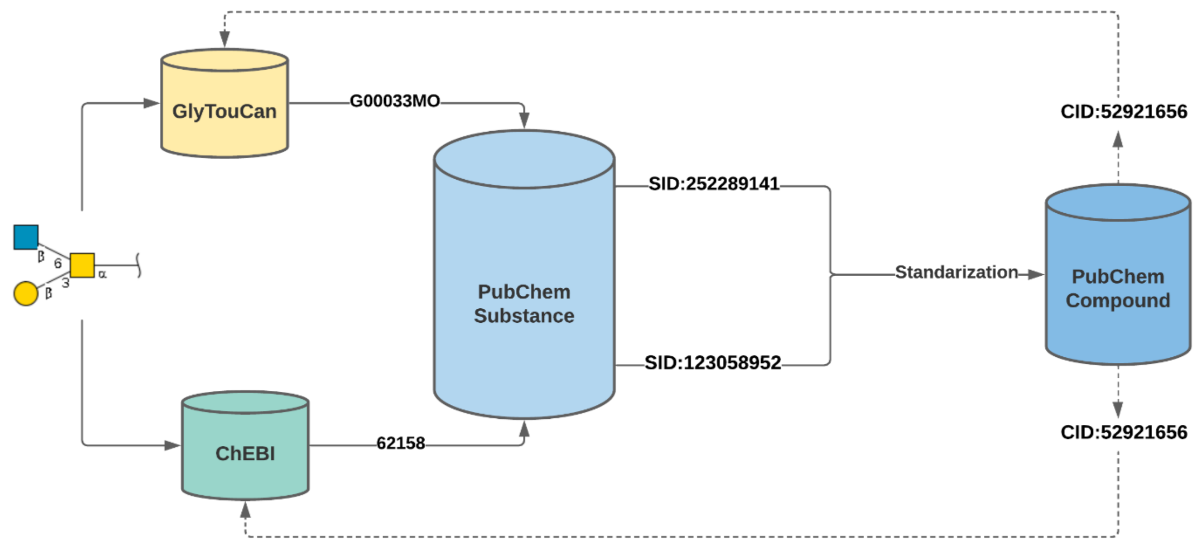


Figure 2: Outlines the data flow of glycans across GlyTouCan, ChEBI, and PubChem databases. The figure shows an example of the same glycan (beta-D-Galp-(1->3)-[beta-D-GlcpNAc-(1->6)]-alpha-D-GalpNAc) present in GlyTouCan (G00033MO) and ChEBI (62158) under respective database identifiers. The GlyTouCan accession and ChEBI ID is mapped to unique PubChem Substance identifiers (G00033MO to SID:252289141; CHEBI:62158 to SID:123058952) when submitted to the PubChem database. PubChem's standardization process maps both the SID's to a single compound identifier (CID:52921656). The same CID is utilized as a cross-reference by both GlyTouCan and ChEBI databases.

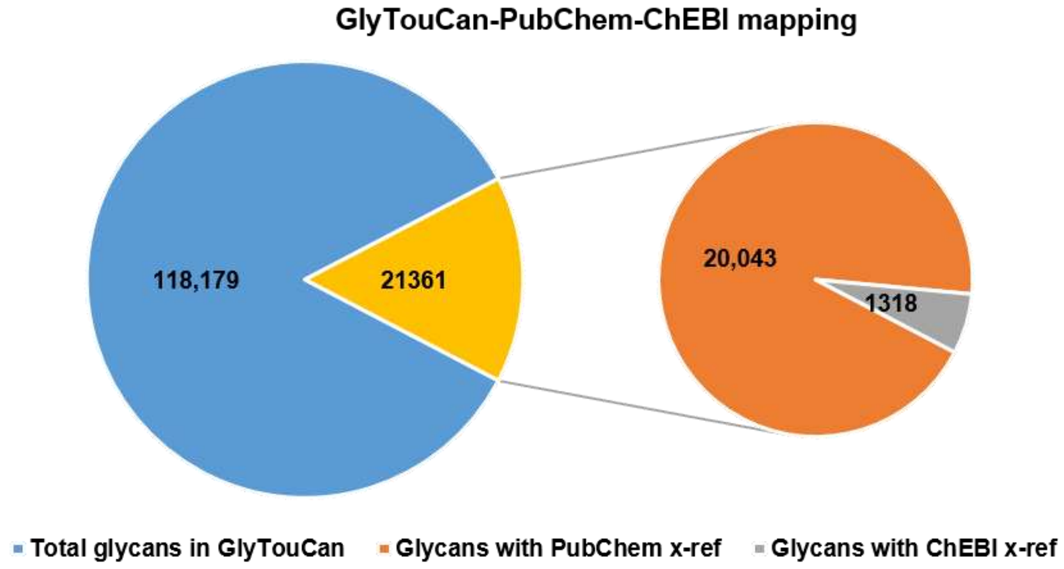


Figure 3: Cross-linking of glycans between GlyTouCan, PubChem, and ChEBI databases. Based on the data collected in April-2020, the total number of glycans in the GlyTouCan database was 118,179 glycans. Of these, 21,519 were submitted as substances (SID) to the PubChem database which mapped to 21,359 compounds (CIDs). Furthermore, of the 21,359 only 2168 CIDs had a ChEBI ID cross-reference.

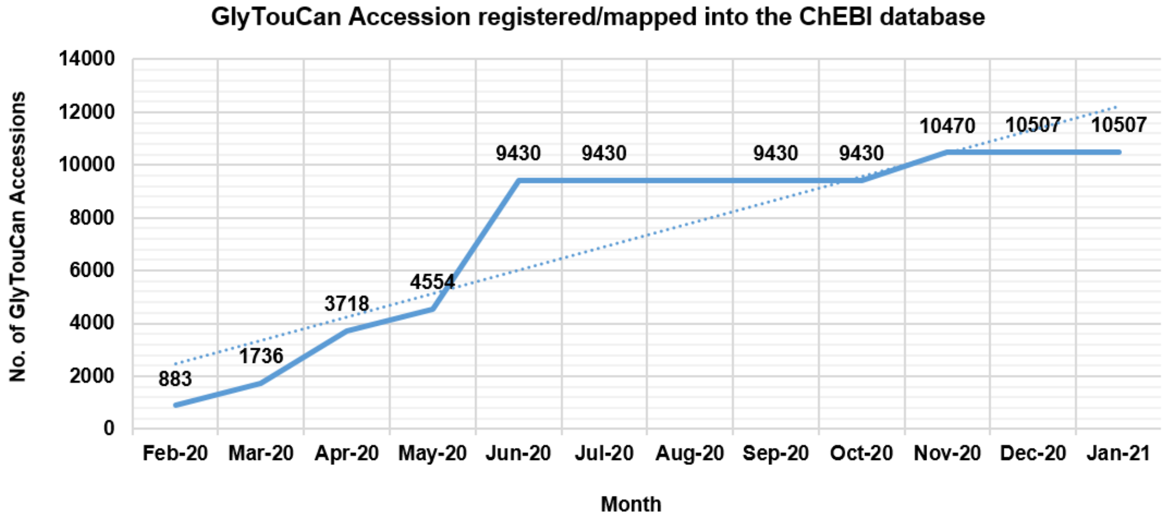


Figure 4: Increase in the number of GlyGen glycans mapped or registered into the ChEBI database. This graph is generated based on monthly versioned files (database_accession.tsv) downloaded from the ChEBI FTP site. These numbers reflect GlyTouCan accessions present in the GlyGen set (v.1.5.3). The file for the month of Aug-20 was not produced by the ChEBI database.

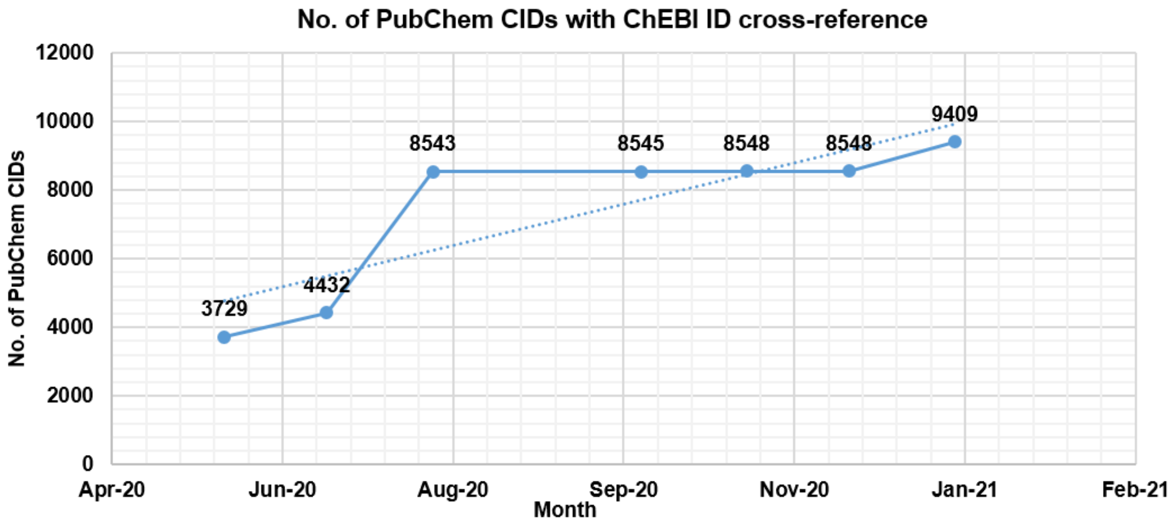


Figure 5: Increase in the number of ChEBI cross-references in the PubChem Compound database through ChEBI's routine submissions. The numbers refer to only those PubChem CID entries which are mapped to the GlyGen GlyTouCan accessions. This graph was generated based on the monthly versioned files (CID-Synonym-filtered.gz) downloaded from the PubChem FTP site.

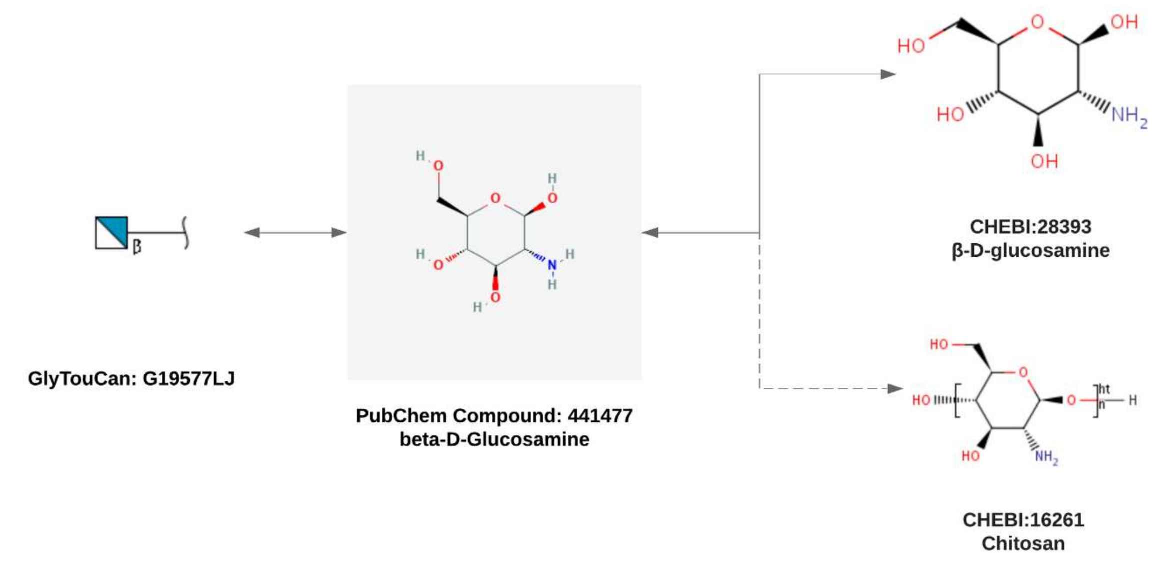
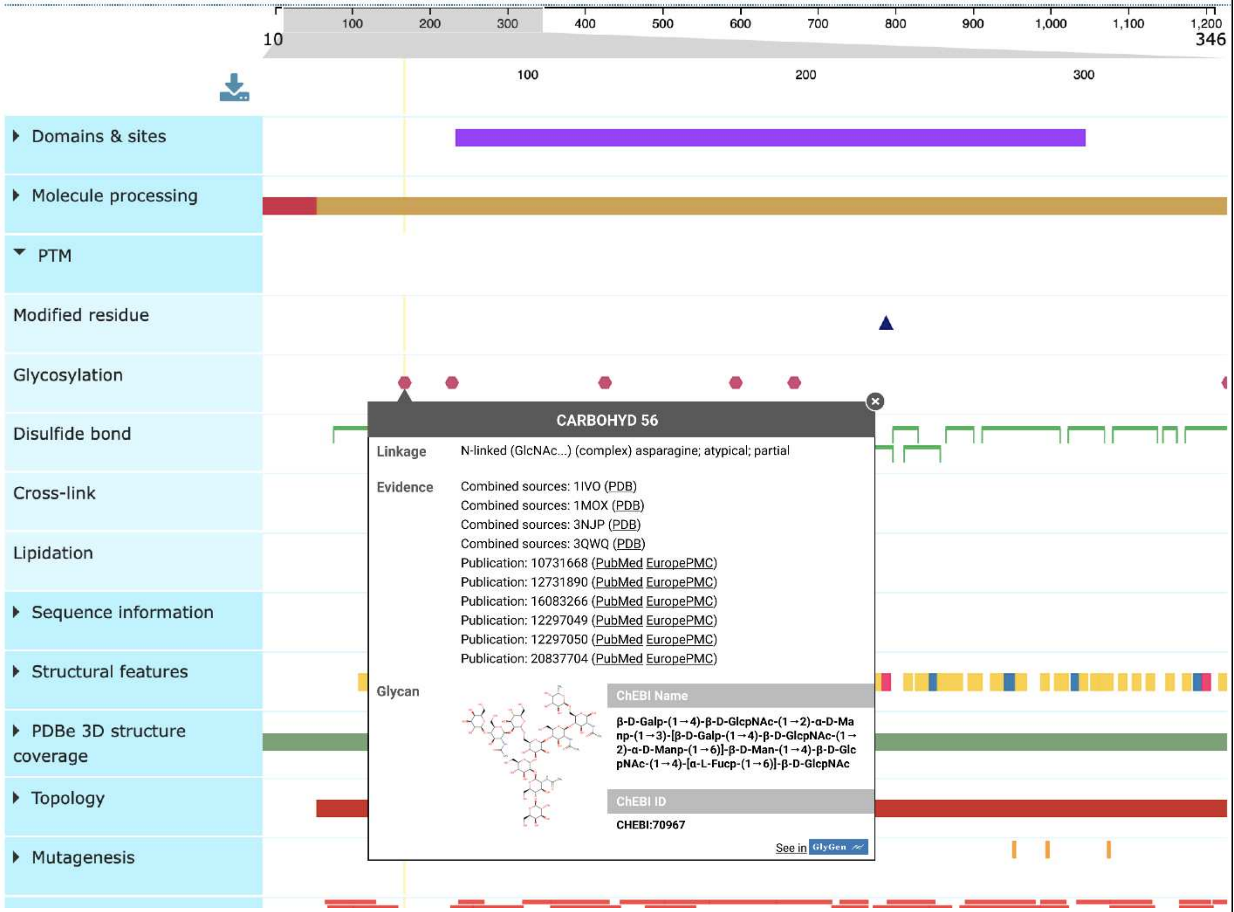


Figure 6: One-to-many PubChem CID to ChEBI ID mapping. The compound, beta-D-Glucosamine, is mapped to both its polymeric (chitosan) and monomeric counterparts in ChEBI. The GlyTouCan accession (G19577LJ) is mapped to the correct ChEBI ID (28393) by matching the InChI strings associated with both ChEBI entries. The correct mapping can be seen on the GlyGen glycan page: https://glygen.org/glycan/G19577LJ



CARBOHYD 56

Linkage N-linked (GlcNAc...) (complex) asparagine; atypical; partial

Evidence
Combined sources: 1IVO (PDB)
Combined sources: 1MOX (PDB)
Combined sources: 3NJP (PDB)
Combined sources: 3QWQ (PDB)
Publication: 10731668 (PubMed EuropePMC)
Publication: 12731890 (PubMed EuropePMC)
Publication: 16083266 (PubMed EuropePMC)
Publication: 12297049 (PubMed EuropePMC)
Publication: 12297050 (PubMed EuropePMC)
Publication: 20837704 (PubMed EuropePMC)

Glycan

ChEBI Name

β-D-Galp-(1→4)-β-D-GlcpNAc-(1→2)-α-D-Manp-(1→3)-[β-D-Galp-(1→4)-β-D-GlcpNAc-(1→2)-α-D-Manp-(1→6)]-β-D-Man-(1→4)-β-D-GlcpNAc-(1→4)-[α-L-Fucp-(1→6)]-β-D-GlcpNAc

ChEBI ID

CHEBI:70967

Figure 7: Glycan data representation in the ProtVista protein viewer. Glycans annotated in ChEBI by GlyGen will be displayed in the UniProt resource through the ProtVista viewer. This example shows the glycosylation site (at amino acid 56) of the human epidermal growth factor receptor (EGFR, UniProt accession P00533). Glycosylation sites are annotated in UniProt with the type of linkage which binds to the carbohydrate chain i.e N-linked and the reducing sugar. The evidence for the annotation i.e. literature references are also provided. The glycan annotation to the corresponding ChEBI identifier and structure will be imported from the GlyGen and ChEBI.

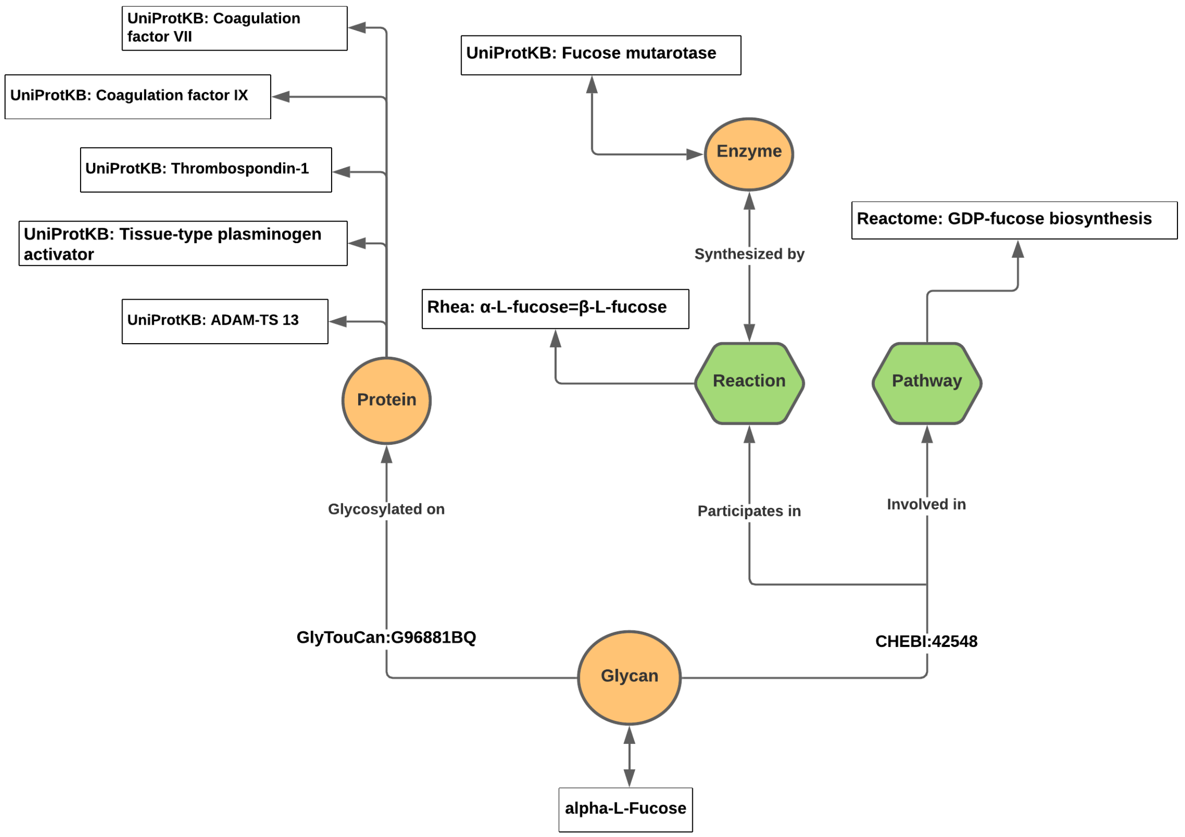

UniProtKB: Coagulation factor VII
UniProtKB: Coagulation factor IX
UniProtKB: Thrombospondin-1
UniProtKB: Tissue-type plasminogen activator
UniProtKB: ADAM-TS 13
UniProtKB: Fucose mutarotase
Enzyme
Synthesized by
Reactome: GDP-fucose biosynthesis
Rhea: α-L-fucose=β-L-fucose
Protein
Reaction
Pathway
Glycosylated on
Participates in
Involved in
GlyTouCan:G96881BQ
CHEBI:42548
Glycan
alpha-L-Fucose

Figure 8: Network of the glycan generated based on the GlyTouCan Accession to ChEBI ID mapping. This enables mapping across multiple databases connecting glycan with information on glycosylation, reaction, and pathway. The network is restricted to human proteins. (e.g https://beta.glygen.org/glycan/G96881BQ#Cross-References)

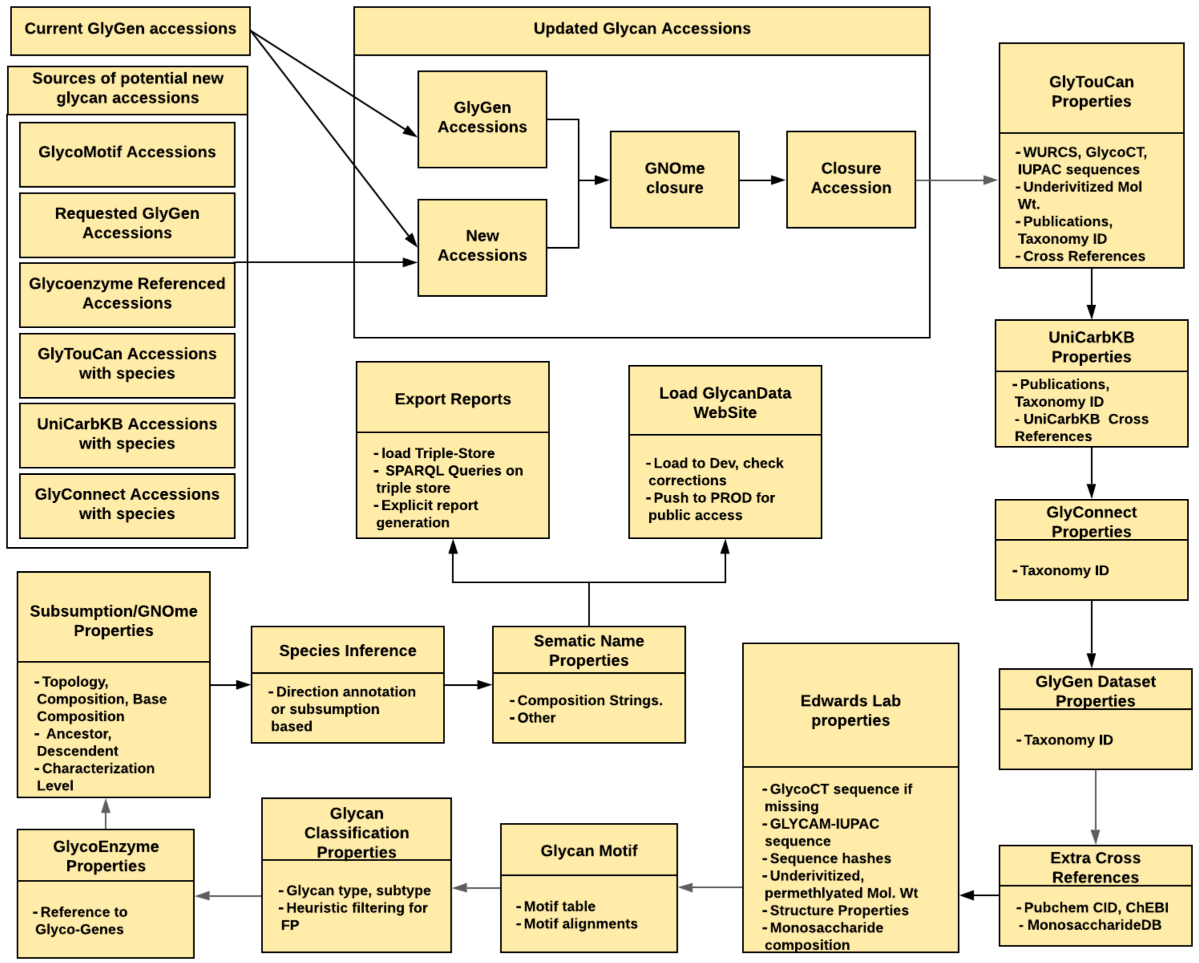

Current GlyGen accessions
Sources of potential new glycan accessions
GlycoMotif Accessions
Requested GlyGen Accessions
Glycoenzyme Referenced Accessions
GlyTouCan Accessions with species
UniCarbKB Accessions with species
GlyConnect Accessions with species
Updated Glycan Accessions
GlyGen Accessions
New Accessions
GNOme closure
Closure Accession
GlyTouCan Properties
-WURCS, GlycoCT, IUPAC sequences
-Underivitized Mol Wt.
-Publications, Taxonomy ID
-Cross References
UniCarbKB Properties
-Publications, Taxonomy ID
- UniCarbKB Cross References
GlyConnect Properties
-Taxonomy ID
GlyGen Dataset Properties
-Taxonomy ID
Extra Cross References
-Pubchem CID, ChEBI
- MonosaccharideDB
Export Reports
-load Triple-Store
- SPARQL Queries on triple store
-Explicit report generation
Load GlycanData WebSite
-Load to Dev, check corrections
-Push to PROD for public access
Subsumption/GNOme Properties
-Topology, Composition, Base Composition
- Ancestor, Descendent
-Characterization Level
Species Inference
-Direction annotation or subsumption based
Sematic Name Properties
-Composition Strings.
-Other
Edwards Lab properties
-GlycoCT sequence if missing
-GLYCAM-IUPAC sequence
-Sequence hashes
-Underivitized, permethlyated Mol. Wt
-Structure Properties
-Monosaccharide composition
GlycoEnzyme Properties
-Reference to Glyco-Genes
Glycan Classification Properties
-Glycan type, subtype
-Heuristic filtering for FP
Glycan Motif
-Motif table
-Motif alignments

Figure 9: Glycan data integration from multiple resources to generate a GlyGen glycan set. Provides an overview of the steps and workflow on collection and integration of glycans from multiple glycans and glycoprotein resources as well as Glycan Naming Ontology to generate a comprehensive glycan dataset for GlyGen.

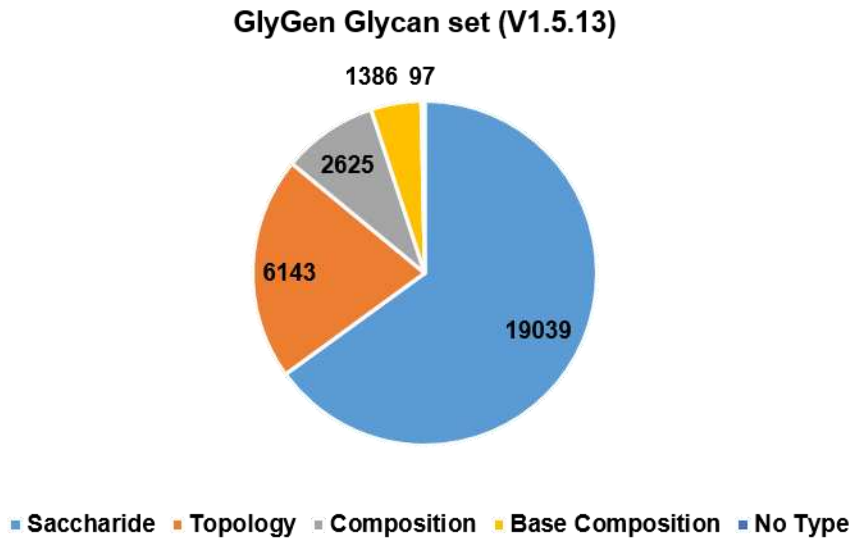


Figure 10: GlyGen glycan set (v1.5.13) and distribution of glycans based on the instances or subsumption categories defined in the Glycan Naming Ontology (GNOme). Out of the entire set of 29,290 glycans, about 13,389 are fully-determined glycans that belong to either saccharide, topology or composition instances, or subsumption categories.

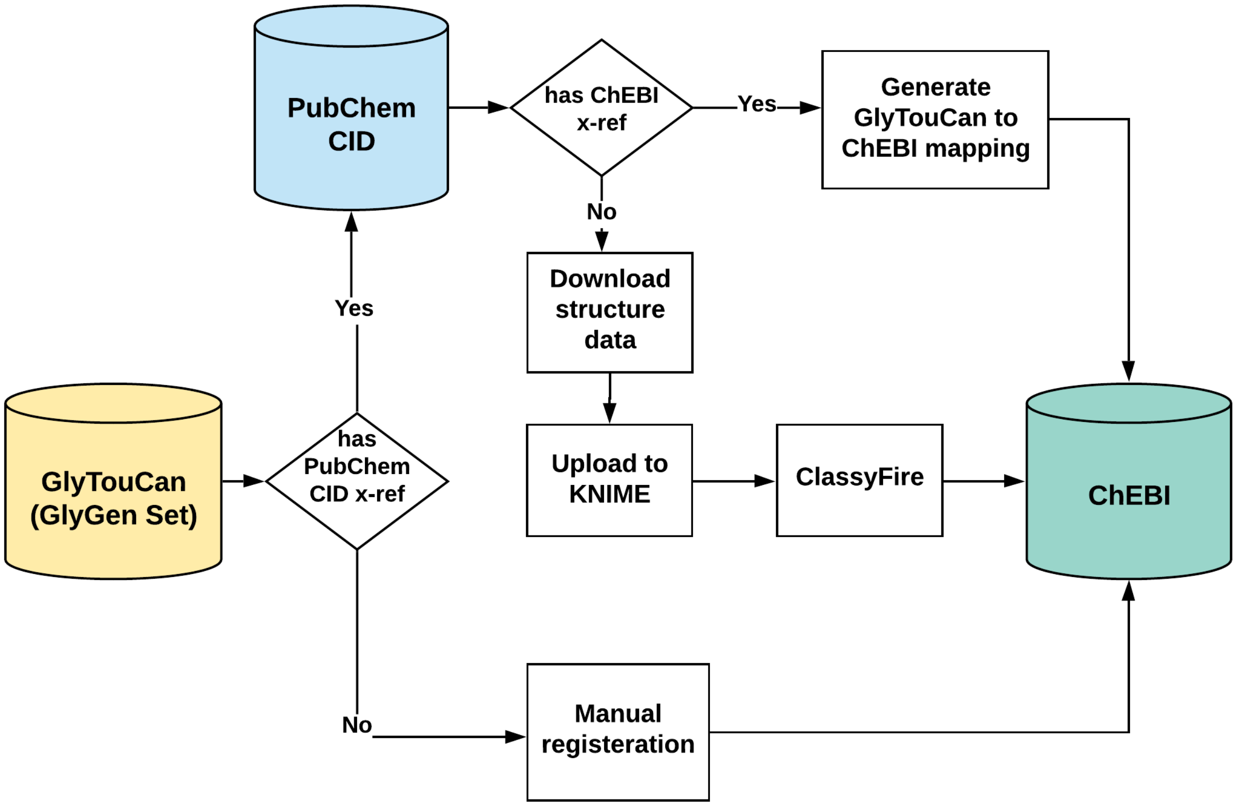


Figure 11: Overview of the data integration pipeline to map or register the GlyGen glycan set of 29,290 GlyTouCan accessions into the ChEBI database. If the GlyTouCan accession had a PubChem CID and a corresponding ChEBI ID, then a cross-reference mapping was generated and added to the ChEBI database where the corresponding GlyTouCan accession was added as a cross-reference. GlyTouCan accessions with a PubChem CID but without a ChEBI ID were uploaded to ChEBI using applications like KNIME and ClassyFire. The remaining GlyTouCan accessions where a PubChem CID mapping was missing were manually registered in the ChEBI database.

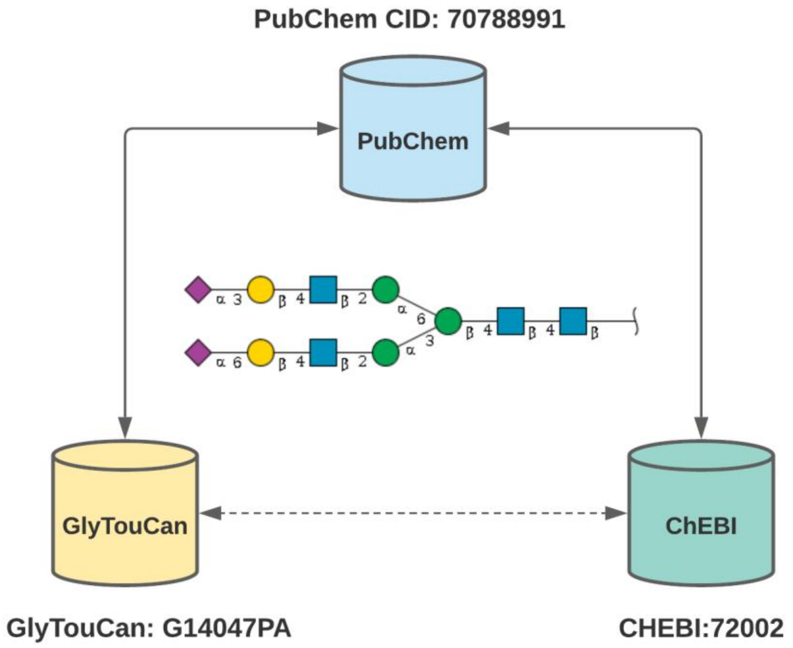
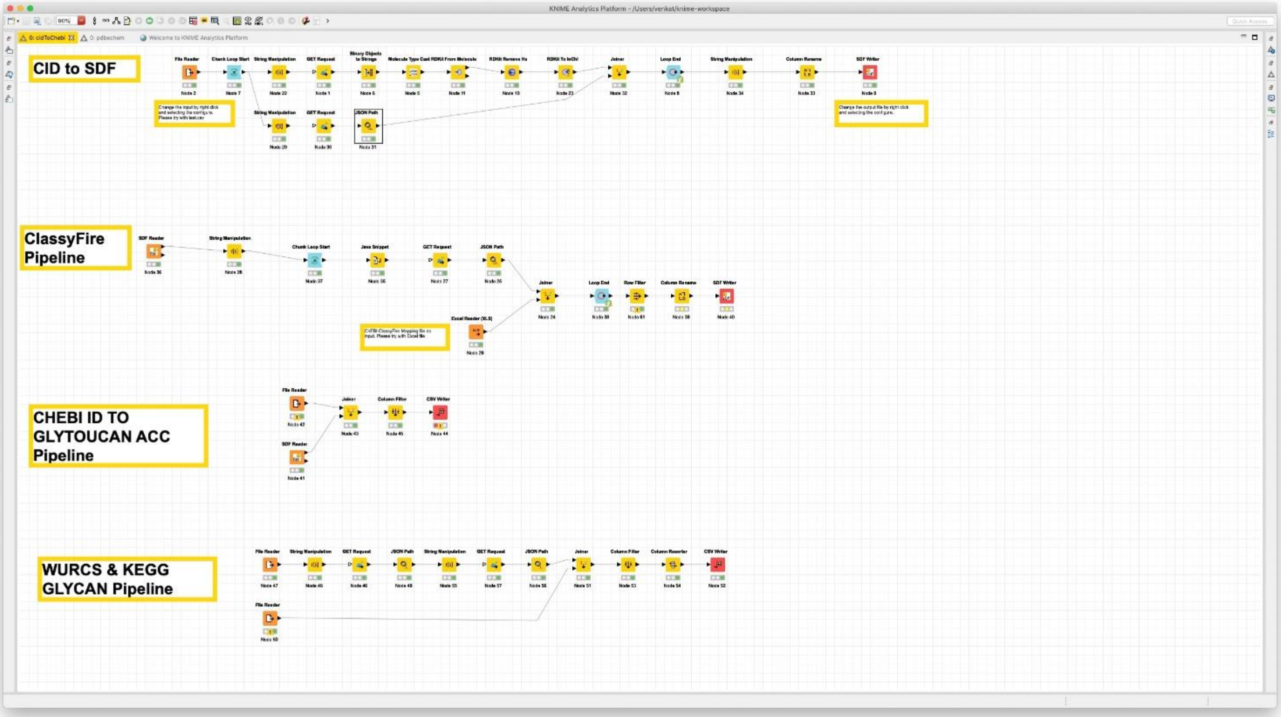


Figure 12: Mapping GlyTouCan accessions to ChEBI IDs using the PubChem CID as an anchor identifier. The figure shows an example of a fully-determined glycan mapped from GlyTouCan to ChEBI ID through a PubChem CID.



Change the input by right click and selecting the configure. Please try with test.csv

Change the output file by right click and selecting the configure.

CHEBI ClassyFire Mapping file as input. Please try with Excel file

Figure 13: The overview process of utilizing the KNIME application to convert structure data associated with PubChem CID into an SDFile for ChEBI database upload. It starts by downloading the structure data associated to a PubChem CID (cross-reference to GlyGen glycan set) and ends with an SDFile generated and ready for ChEBI upload.

LEGENDS TO FIGURES

Figure 1: Different types of glycans (identified by a GlyTouCan accession) represented in an SNFG format with varying levels of knowledge about its stereochemistry, glycosidic bonds, anomeric configuration, etc. The types of glycan shown in the figure also describe different instances or subsumption categories in the Glycan Naming Ontology (http://www.obofoundry.org/ontology/gno.html) or GNOme. 1a) Shows a type of glycan (G17689DH) where the arrangement (type, number, and stereochemistry) of monosaccharides is known along with the glyosidic bonds with partial or complete information about the linkage or anomeric configuration. Such glycans are categorized under "Saccharide" (http://purl.obolibrary.org/obo/GNO_00000016) instance in GNOme. 1b) Shows a type of glycan (G69616RY) where there is no information about the linkage or anomeric configuration, but the arrangement of monosaccharides and glycosidic bonds is known. Such glycans belong to the "Topology" (http://purl.obolibrary.org/obo/GNO_00000015) instance in GNOme. 1c) Shows a type of glycan (G83385WA) where only the information about type and number of monosaccharides is known in addition to complete or partial knowledge of the stereochemistry of the involved monosaccharides. Such glycans belong to the "composition" (http://purl.obolibrary.org/obo/GNO_00000014) instance in GNOme. Certain glycans from these three instances are also identified as fully-determined glycans where there is complete knowledge of the arrangement of the monosaccharides, glycosidic bonds, and linkage/anomeric configuration with the

exception of first linkage (between glycan and conjugate) where the ambiguity is tolerated. 1d) Shows a type of glycan (G42039FI) where only the information about the type and number of monosaccharides is known. Such glycans are categorized under the "basecomposition" (http://purl.obolibrary.org/obo/GNO_00000013) instance in GNOme.

Figure 2: Outlines the data flow of glycans across GlyTouCan, ChEBI, and PubChem databases. The figure shows an example of the same glycan (beta-D-Galp-(1->3)-[beta-D-GlcpNAc-(1->6)]-alpha-D-GalpNAc) present in GlyTouCan (G00033MO) and ChEBI (62158) under respective database identifiers. The GlyTouCan accession and ChEBI ID is mapped to unique PubChem Substance identifiers (G00033MO to SID:252289141; CHEBI:62158 to SID:123058952) when submitted to the PubChem database. PubChem's standardization process maps both the SID's to a single compound identifier (CID:52921656). The same CID is utilized as a cross-reference by both GlyTouCan and ChEBI databases.

Figure 3: Cross-linking of glycans between GlyTouCan, PubChem, and ChEBI databases. Based on the data collected in April-2020, the total number of glycans in the GlyTouCan database was 118,179 glycans. Of these, 21,519 were submitted as substances (SID) to the PubChem database which mapped to 21,359 compounds (CIDs). Furthermore, of the 21,359 only 2168 CIDs had a ChEBI ID cross-reference.

Figure 4: Increase in the number of GlyGen glycans mapped or registered into the ChEBI database. This graph is generated based on monthly versioned files (database_accession.tsv) downloaded from the ChEBI FTP site. These numbers reflect GlyTouCan accessions present in the GlyGen set (v.1.5.3). The file for the month of Aug-20 was not produced by the ChEBI database.

Figure 5: Increase in the number of ChEBI cross-references in the PubChem Compound database through ChEBI's routine submissions. The numbers refer to only those PubChem CID entries which are mapped to the GlyGen GlyTouCan accessions. This graph was generated based on the monthly versioned files (CID-Synonym-filtered.gz) downloaded from the PubChem FTP site.

Figure 6: One-to-many PubChem CID to ChEBI ID mapping. The compound, beta-D-Glucosamine, is mapped to both its polymeric (chitosan) and monomeric counterparts in ChEBI. The GlyTouCan accession (G19577LJ) is mapped to the correct ChEBI ID (28393) by matching the InChI strings associated with both ChEBI entries. The correct mapping can be seen on the GlyGen glycan page: https://glygen.org/glycan/G19577LJ

Figure 7: Glycan data representation in the ProtVista protein viewer. Glycans annotated in ChEBI by GlyGen will be displayed in the UniProt resource through the ProtVista viewer. This example shows the glycosylation site (at amino acid 56) of the human epidermal growth factor receptor (EGFR, UniProt accession P00533). Glycosylation sites are annotated in UniProt with the type of linkage which binds to the carbohydrate

chain i.e N-linked and the reducing sugar. The evidence for the annotation i.e. literature references are also provided. The glycan annotation to the corresponding ChEBI identifier and structure will be imported from the GlyGen and ChEBI.

Figure 8: Network of the glycan generated based on the GlyTouCan Accession to ChEBI ID mapping. This enables mapping across multiple databases connecting glycan with information on glycosylation, reaction, and pathway. The network is restricted to human proteins. (e.g https://beta.glygen.org/glycan/G96881BQ#Cross-References)

Figure 9: Glycan data integration from multiple resources to generate a GlyGen glycan set. Provides an overview of the steps and workflow on collection and integration of glycans from multiple glycans and glycoprotein resources as well as Glycan Naming Ontology to generate a comprehensive glycan dataset for GlyGen.

Figure 10: GlyGen glycan set (v1.5.13) and distribution of glycans based on the instances or subsumption categories defined in the Glycan Naming Ontology (GNOme). Out of the entire set of 29,290 glycans, about 13,389 are fully-determined glycans that belong to either saccharide, topology or composition instances, or subsumption categories.

Figure 11: Overview of the data integration pipeline to map or register the GlyGen glycan set of 29,290 GlyTouCan accessions into the ChEBI database. If the GlyTouCan accession had a PubChem CID and a corresponding ChEBI ID, then a cross-reference

mapping was generated and added to the ChEBI database where the corresponding GlyTouCan accession was added as a cross-reference. GlyTouCan accessions with a PubChem CID but without a ChEBI ID were uploaded to ChEBI using applications like KNIME and ClassyFire. The remaining GlyTouCan accessions where a PubChem CID mapping was missing were manually registered in the ChEBI database.

Figure 12: Mapping GlyTouCan accessions to ChEBI IDs using the PubChem CID as an anchor identifier. The figure shows an example of a fully-determined glycan mapped from GlyTouCan to ChEBI ID through a PubChem CID.

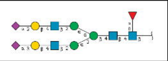
Figure 13: The overview process of utilizing the KNIME application to convert structure data associated with PubChem CID into an SDFile for ChEBI database upload. It starts by downloading the structure data associated to a PubChem CID (cross-reference to GlyGen glycan set) and ends with an SDFile generated and ready for ChEBI upload.

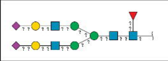
TABLES

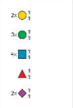
Table 1. Depicts the high-value, ambiguously defined glycans manually registered into the ChEBI database as a proof of concept.

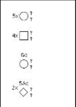
GlyTouCan Accession	Subsumption category (GNOme)	CHEBI ID	No of associated glycoproteins
G06110VR	Base Composition	CHEBI:156559	762
G43417UB	Base Composition	CHEBI:156560	608
G70101JE	Base Composition	CHEBI:156561	489
G84452RH	Composition	CHEBI:156562	360
G06356OH	Composition	CHEBI:156563	280
G74724QE	Base Composition	CHEBI:156564	221
G29068FM	Base Composition	CHEBI:156565	215
G23863VK	Base Composition	CHEBI:156566	200
G23294PN	Composition	CHEBI:156568	190
G14669DU	Base Composition	CHEBI:156569	190
G72291OX	Composition	CHEBI:156570	100
G14548ZL	Saccharide	CHEBI:156581	40
G42358LZ	Saccharide	CHEBI:156582	35
G22140GZ	Saccharide	CHEBI:156583	30
G40608DF	Saccharide	CHEBI:156584	19
G70994MS	Topology	CHEBI:18133	17

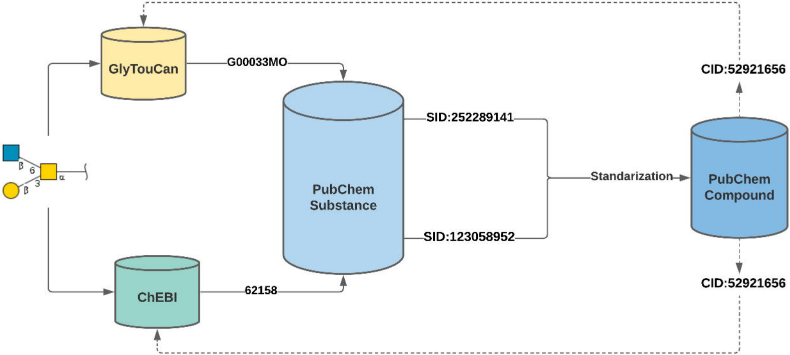
G57818FI	Saccharide	CHEBI:156585	13
G39397SW	Saccharide	CHEBI:157599	12
G93860XO	Saccharide	CHEBI:157638	10
G48975GN	Saccharide	CHEBI:157639	10
G62765YT	Base Composition	CHEBI:157633	9
G80920RR	Base Composition	CHEBI:157634	9
G31852PQ	Base Composition	CHEBI:157635	8
G41247ZX	Base Composition	CHEBI:157636	8
G02815KT	Base Composition	CHEBI:157637	7

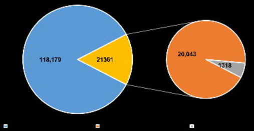

α 3
β 4
β 2
α 6
α 3
β 4
β 2
α 3
β 4
β 4
α 6
β

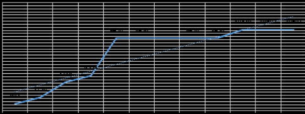

? ?
? ?
? ?
? ?
?
?

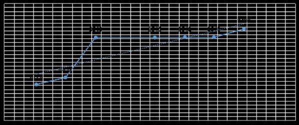

2×
3×
4×
2×
?
?

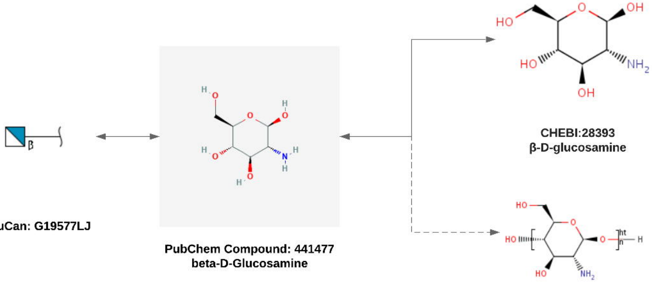

5×
4×
6d
5Ac
2×
?
?


GlyTouCan
G00033MO
PubChem
Substance
SID:252289141
SID:123058952
Standarization
PubChem
Compound
CID:52921656
CID:52921656
β
6
3
α
β
ChEBI
62158


118,179
21361
20,043
1318





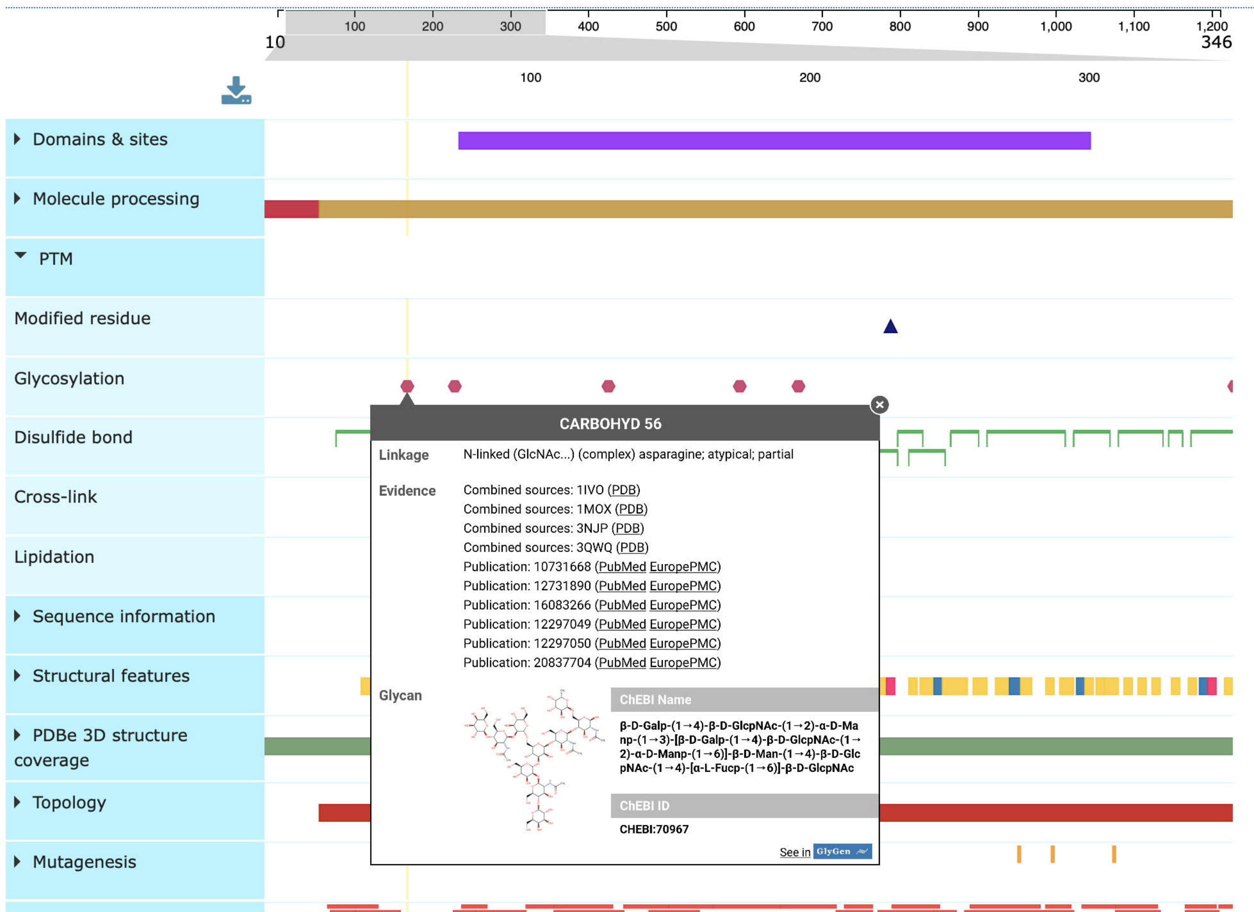


GlyTouCan: G19577LJ

PubChem Compound: 441477
beta-D-Glucosamine

CHEBI:28393
β-D-glucosamine

CHEBI:16261
Chitosan



CARBOHYD 56

Linkage N-linked (GlcNAc...) (complex) asparagine; atypical; partial

Evidence
Combined sources: 1IVO (PDB)
Combined sources: 1MOX (PDB)
Combined sources: 3NJP (PDB)
Combined sources: 3QWQ (PDB)
Publication: 10731668 (PubMed EuropePMC)
Publication: 12731890 (PubMed EuropePMC)
Publication: 16083266 (PubMed EuropePMC)
Publication: 12297049 (PubMed EuropePMC)
Publication: 12297050 (PubMed EuropePMC)
Publication: 20837704 (PubMed EuropePMC)

Glycan

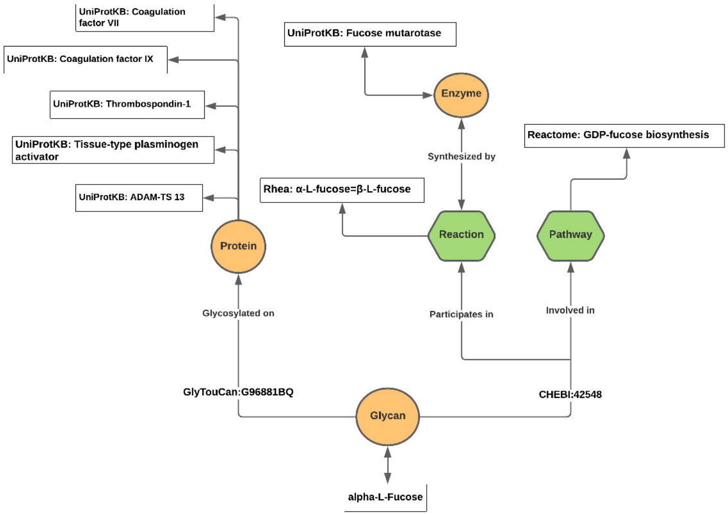
ChEBI Name

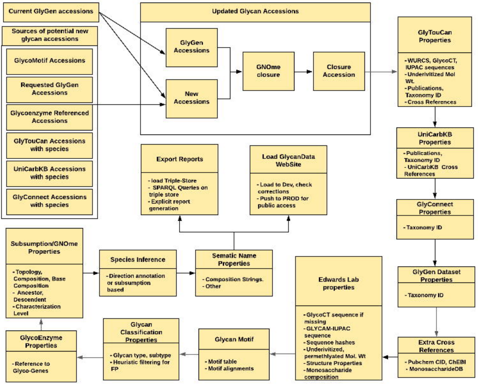
β-D-Galp-(1→4)-β-D-GlcpNAc-(1→2)-α-D-Manp-(1→3)-[β-D-Galp-(1→4)-β-D-GlcpNAc-(1→2)-α-D-Manp-(1→6)]-β-D-Man-(1→4)-β-D-GlcpNAc-(1→4)-[α-L-Fucp-(1→6)]-β-D-GlcpNAc

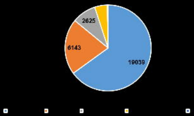
ChEBI ID

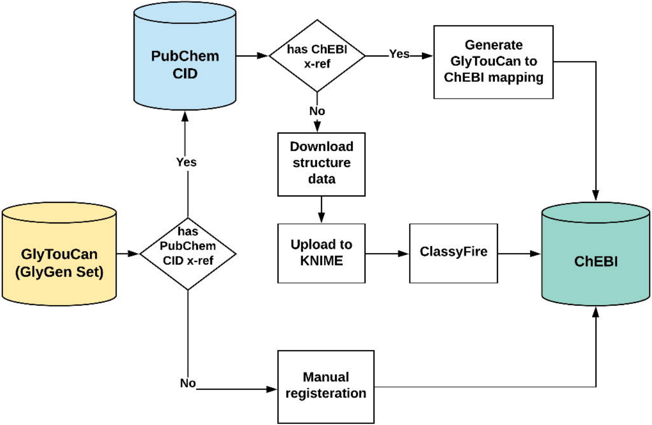
CHEBI:70967

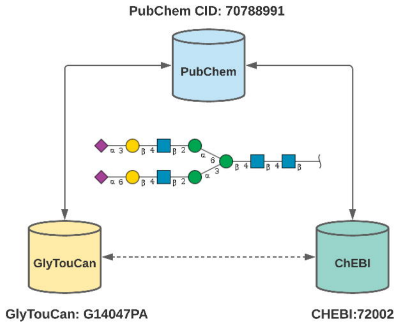
See in GlyGen


UniProtKB: Coagulation factor VII
UniProtKB: Coagulation factor IX
UniProtKB: Thrombospondin-1
UniProtKB: Tissue-type plasminogen activator
UniProtKB: ADAM-TS 13
Protein
Glycosylated on
GlyTouCan:G96881BQ
UniProtKB: Fucose mutarotase
Enzyme
Synthesized by
Rhea: α-L-fucose=β-L-fucose
Reaction
Reactome: GDP-fucose biosynthesis
Pathway
Participates in
Involved in
CHEBI:42548
Glycan
alpha-L-Fucose


Current GlyGen accessions
Sources of potential new glycan accessions
GlycoMotif Accessions
Requested GlyGen Accessions
Glycoenzyme Referenced Accessions
GlyTouCan Accessions with species
UniCarbKB Accessions with species
GlyConnect Accessions with species
Updated Glycan Accessions
GlyGen Accessions
New Accessions
GNOme closure
Closure Accession
GlyTouCan Properties
- WURCS, GlycoCT, IUPAC sequences
- Underivitized Mol Wt.
- Publications, Taxonomy ID
- Cross References
UniCarbKB Properties
- Publications, Taxonomy ID
- UniCarbKB Cross References
GlyConnect Properties
- Taxonomy ID
GlyGen Dataset Properties
- Taxonomy ID
Extra Cross References
- Pubchem CID, ChEBI
- MonosaccharideDB
Edwards Lab properties
- GlycoCT sequence if missing
- GLYCAM-IUPAC sequence
- Sequence hashes
- Underivitized, permethlyated Mol. Wt
- Structure Properties
- Monosaccharide composition
Glycan Motif
- Motif table
- Motif alignments
Glycan Classification Properties
- Glycan type, subtype
- Heuristic filtering for FP
GlycoEnzyme Properties
- Reference to Glyco-Genes
Subsumption/GNOme Properties
- Topology, Composition, Base Composition
- Ancestor, Descendent
- Characterization Level
Species Inference
- Direction annotation or subsumption based
Sematic Name Properties
- Composition Strings.
- Other
Export Reports
- load Triple-Store
- SPARQL Queries on triple store
- Explicit report generation
Load GlycanData WebSite
- Load to Dev, check corrections
- Push to PROD for public access


2625
6143
19039


PubChem CID
has ChEBI x-ref
Yes
Generate GlyTouCan to ChEBI mapping
No
Download structure data
Yes
Upload to KNIME
ClassyFire
GlyTouCan (GlyGen Set)
has PubChem CID x-ref
ChEBI
No
Manual registeration


PubChem CID: 70788991
PubChem
α 3
β 4
β 2
α
6
α 6
β 4
β 2
α
3
β 4
β 4
β
GlyTouCan
ChEBI
GlyTouCan: G14047PA
CHEBI:72002

KNIME Analytics Platform - /Users/venkat/knime-workspace

90%
Quick Access
0: cidToChebi
0: pdbechem
Welcome to KNIME Analytics Platform

CID to SDF

File Reader
Node 2
Chunk Loop Start
Node 7
String Manipulation
Node 22
GET Request
Node 1
Binary Objects to Strings
Node 6
Molecule Type Cast
Node 5
RDKit From Molecule
Node 11
RDKit Remove Hs
Node 10
RDKit To InChI
Node 23
Joiner
Node 32
Loop End
Node 8
String Manipulation
Node 34
Column Rename
Node 33
SDF Writer
Node 9
String Manipulation
Node 29
GET Request
Node 30
JSON Path
Node 31

Change the input by right click
and selecting the configure.
Please try with test.csv

Change the output file by right click
and selecting the configure.

ClassyFire Pipeline

SDF Reader
Node 36
String Manipulation
Node 28
Chunk Loop Start
Node 37
Java Snippet
Node 35
GET Request
Node 27
JSON Path
Node 25
Joiner
Node 24
Loop End
Node 38
Row Filter
Node 61
Column Rename
Node 39
SDF Writer
Node 40
Excel Reader (XLS)
Node 26

ChEBI ClassyFire Mapping file as
input. Please try with Excel file

CHEBI ID TO GLYTOUCAN ACC Pipeline

File Reader
Node 42
SDF Reader
Node 41
Joiner
Node 43
Column Filter
Node 45
CSV Writer
Node 44

WURCS & KEGG GLYCAN Pipeline