GlyGen

Computational and Informatics Resources for Glycosciences

René Ranzinger

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

PI: Raja Mazumder
The George Washington University

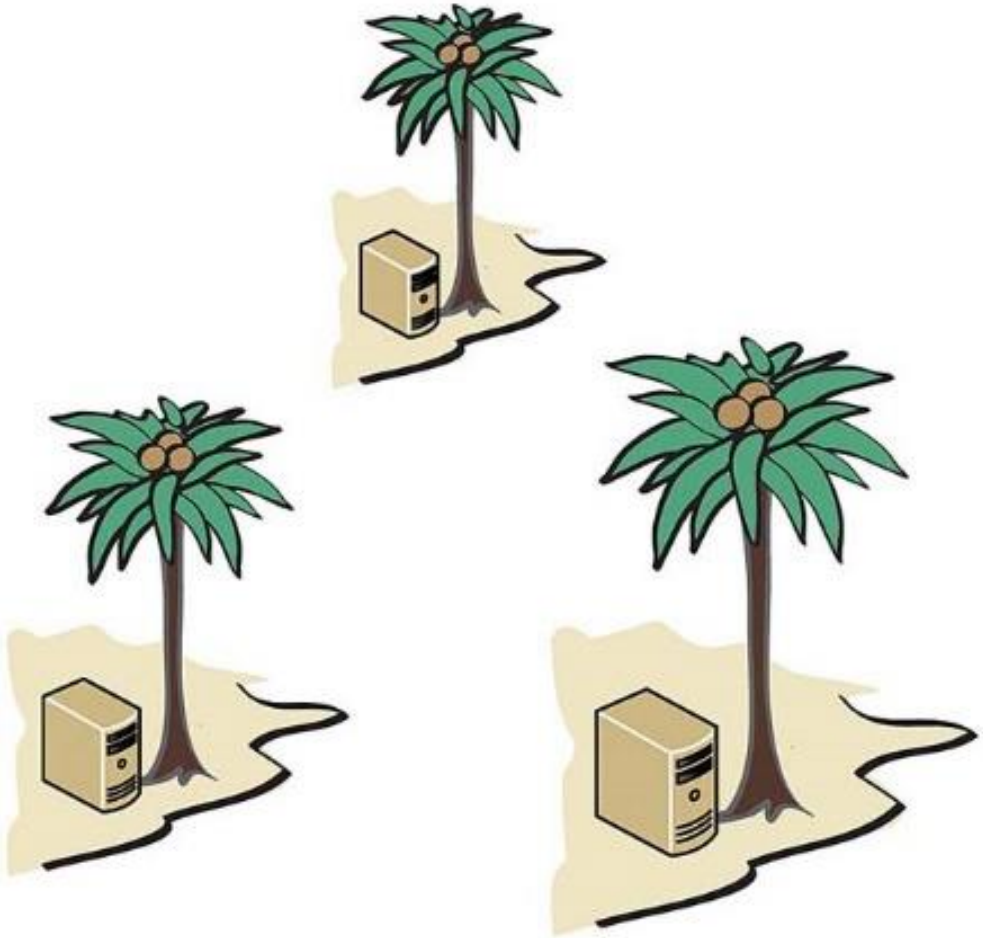
Glycoscience

NIH Common Fund Glycoscience Program
1U01GM125267-01

UNIVERSITY OF GEORGIA

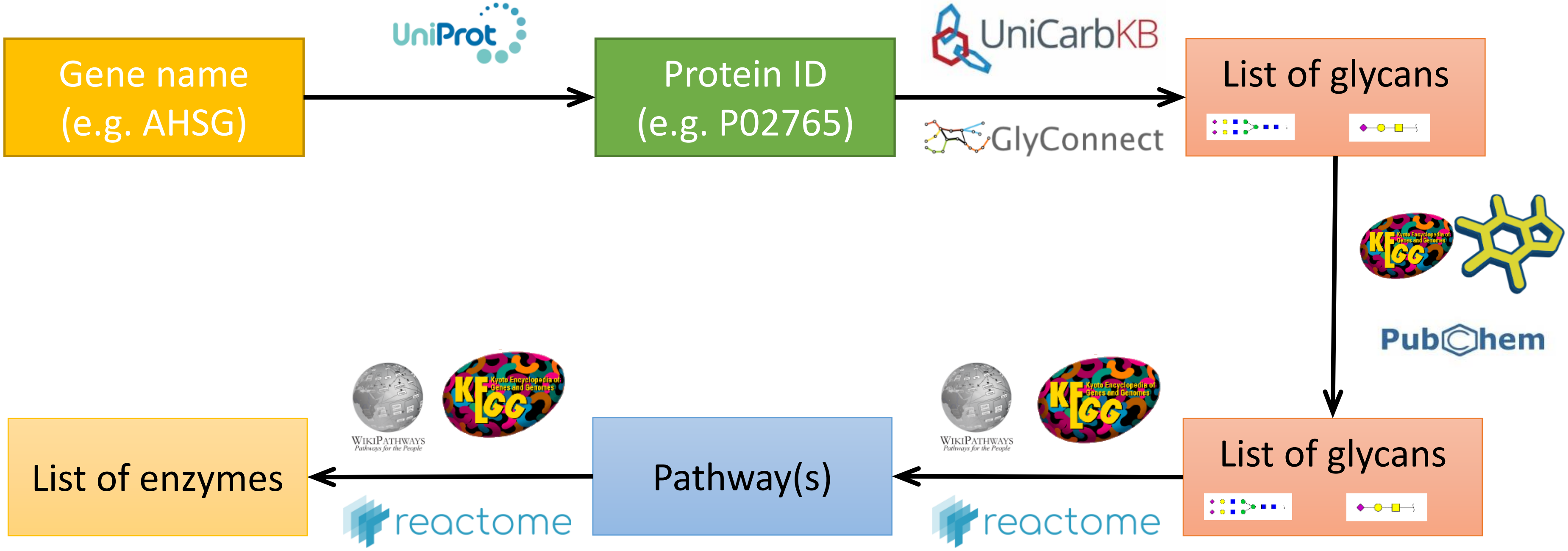
PI: William York
CCRC, *University of Georgia*

Islands of information



A complex query

- Get all **enzymes** that my of been involved in *synthesis of the glycans* on protein X.



Gene name (e.g. AHSG)

Protein ID (e.g. P02765)

List of glycans

List of glycans

Pathway(s)

List of enzymes

Aim of the project

- **Integration of glycobiology-related information** from diverse domains/databases
- **Creation of an intuitive web portal** to browse and search for knowledge in glycobiology
- Provide **free and standardized access** to the integrated datasets
- **Developing essential new information resources**, including:
  - An open, comprehensive **Glycan microarray data repository**
  - A **Glycan Naming Ontology (GNOme)** that facilitates interpretation of incomplete structural information in the context of biological function

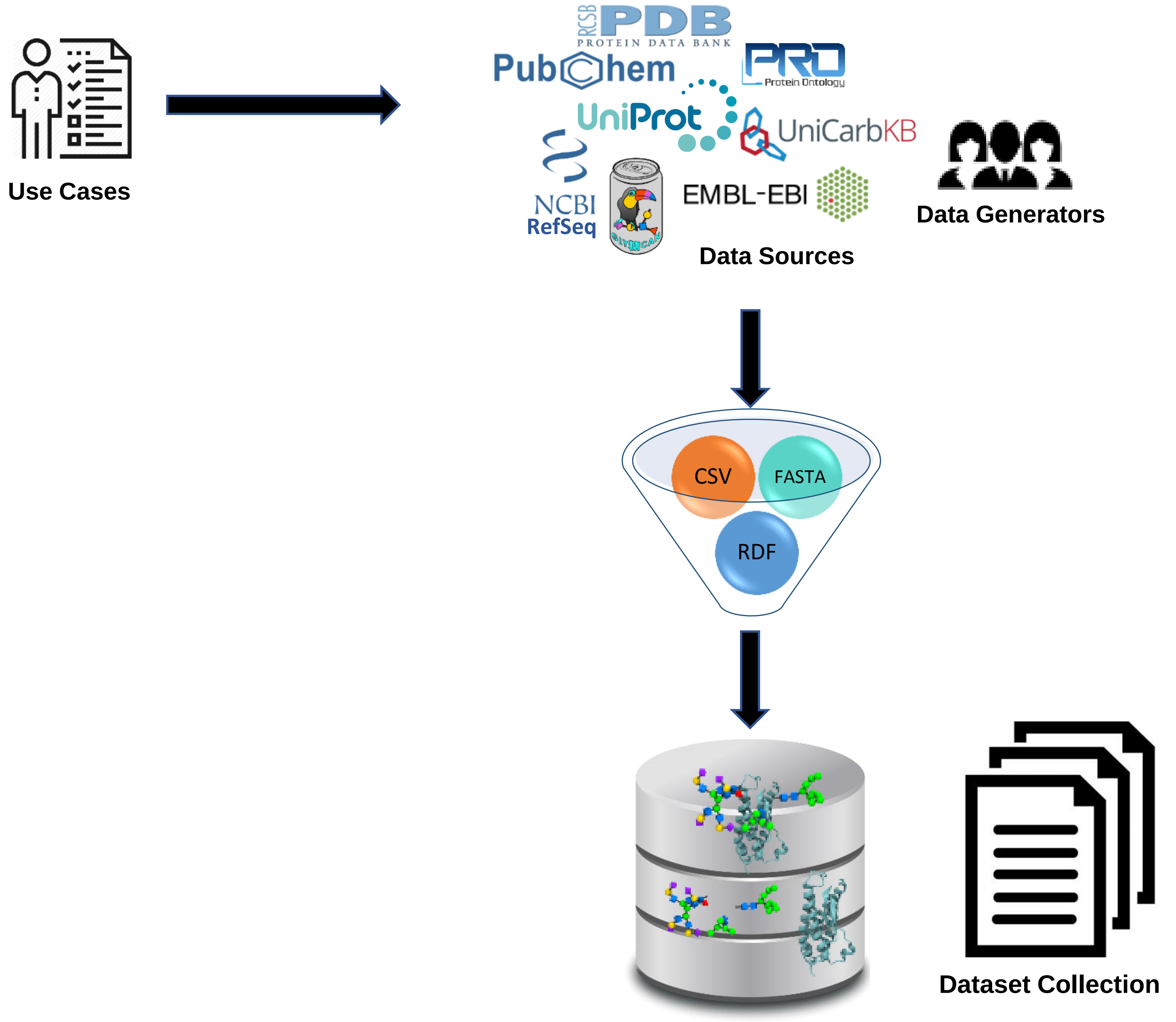
Table of content

- Data integration

- Data access

- Data sharing

Data Integration



Use Cases

Data Generators

Data Sources

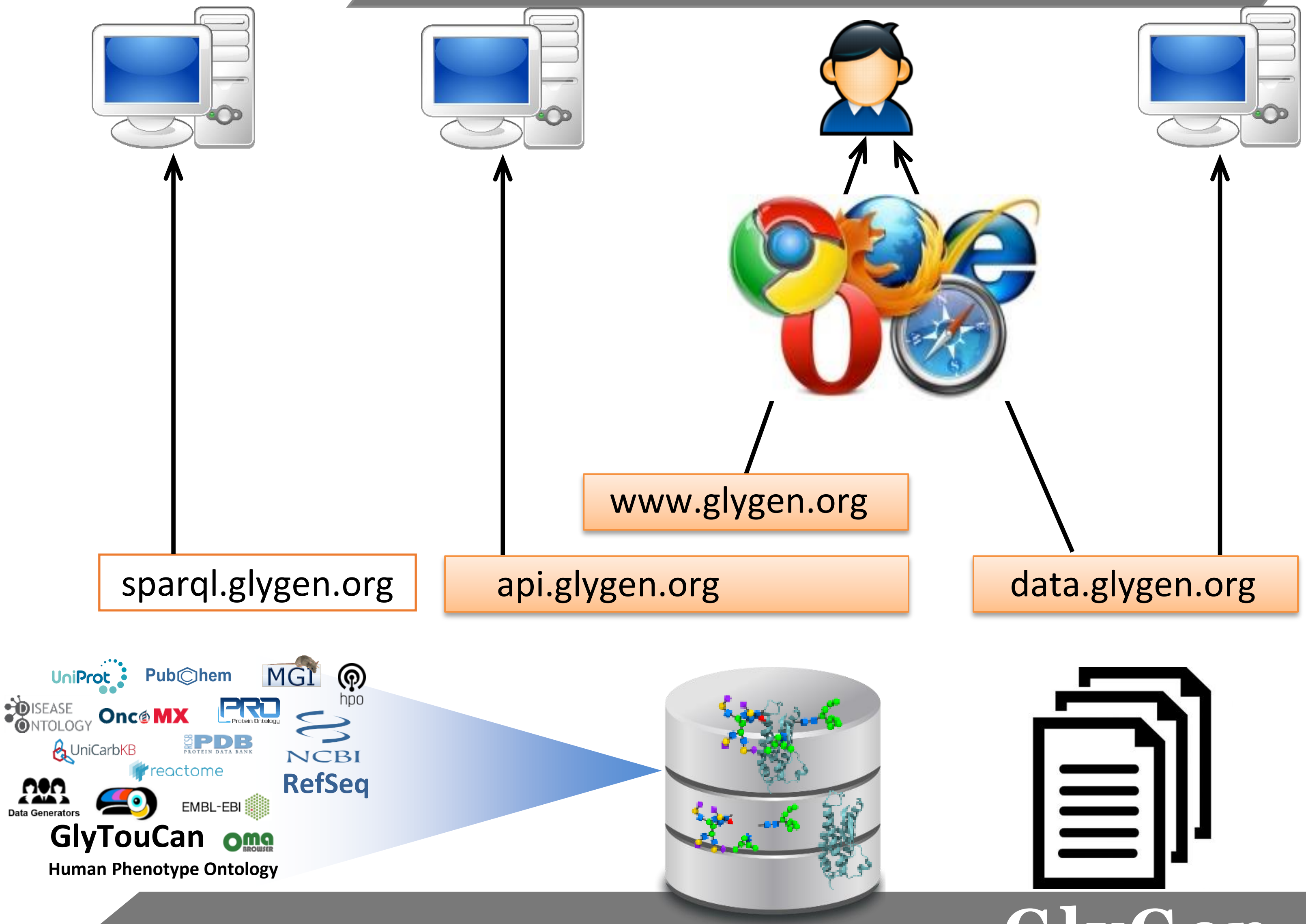
CSV

FASTA

RDF

Dataset Collection

Architecture



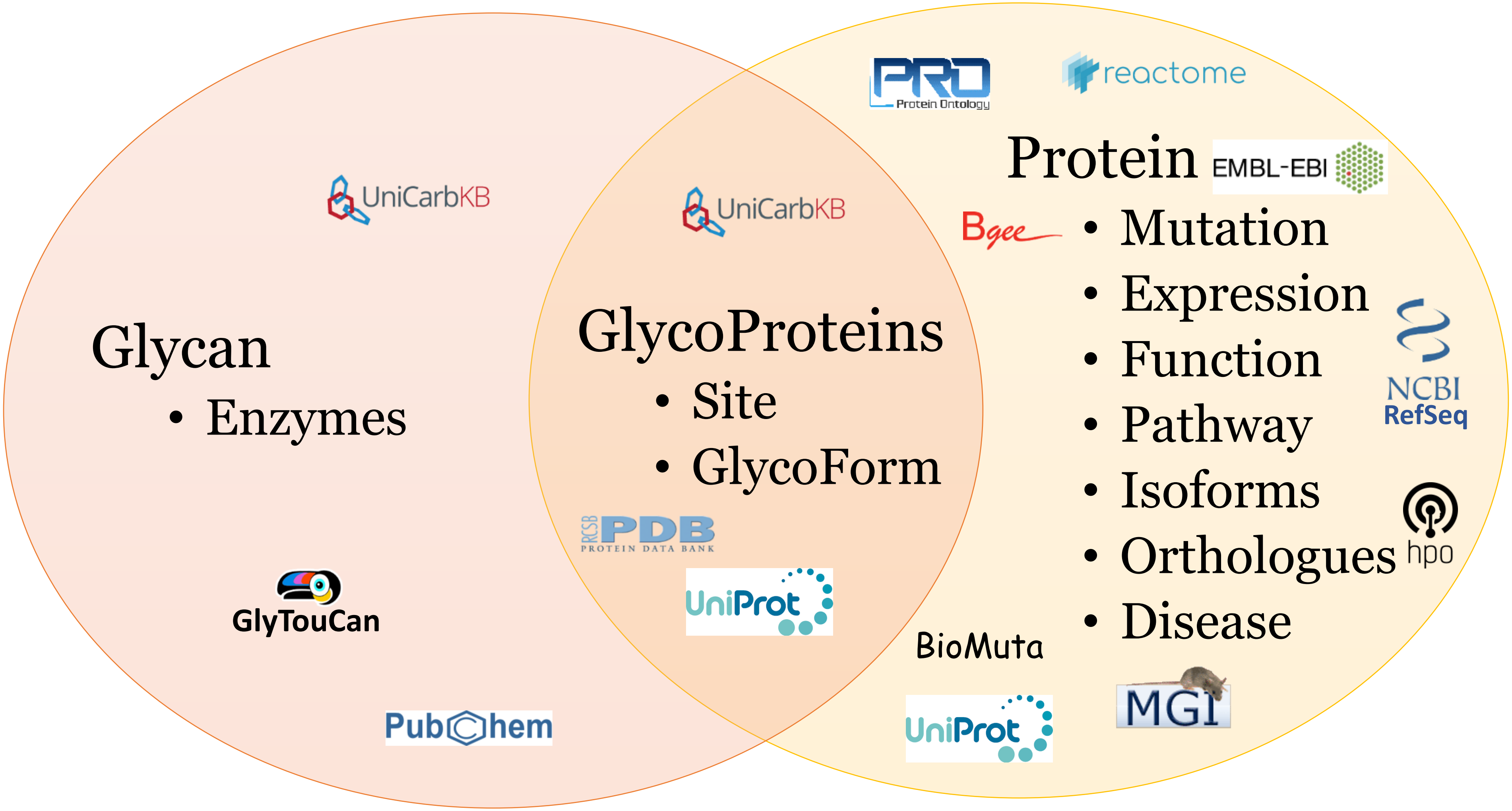
www.glygen.org

sparql.glygen.org

api.glygen.org

data.glygen.org

Data in GlyGen



Glycan

- Enzymes

GlycoProteins

- Site
- GlycoForm

Protein

- Mutation
- Expression
- Function
- Pathway
- Isoforms
- Orthologues
- Disease

BioCompute Objects

- BioCompute is a community driven project to build a framework to standardize bioinformatics computations and analyses communication.

- **The dataset readme provides the following info** –
  - Provenance, Authoring and Versioning details
  - Usability and description of the dataset
  - Detailed steps required to create the dataset
  - Input and Output files and their formats
  - Software and computational platform requirements
  - Tools, scripts and codes required for creating the dataset
  - Content description of the dataset
  - Dataset statistics

BioCompute Object example

BioCompute Objects

Provenance Domain

ID: http://data.glygen.org/DSBCO_000038
Name: human_proteoform_glycosylation_sites_unicarbkb_glytoucan.csv
Title: Glycosylation Sites [UniCarbKB]
Version: 1.0
Created: 2018-02-21T14:46:55-5:00
Created by: Rahi Navelkar [rsn13@gwu.edu; Jeet Vora [jeetvora@gwu.edu]
Digital Signature: RYFNNKE22594E007JKV457
Review status: Approved
Contribution: Matthew Campbell[contributedBy], Robel Kahsay [curatedBy], Rahi Navelkar [createdBy]
License: Data – Attribution 4.0 International CC BY 4.0 [https://creativecommons.org/licenses/by/4.0/]
Scripts – GNU General Public License v3.0 [https://www.gnu.org/licenses/gpl-3.0.en.html]

Provenance, Authoring, Versioning and License Information

Usability Domain

List of human [taxid:9606] proteins with information on glycosylation sites from UniCarbKB database. The file also includes GlyTouCan accessions and UniCarbKB structure ids for associated glycan structures. https://academic.oup.com/nar/article/42/D1/D215/1052197 ,https://doi.org/10.1093/nar/gkt1128]

Describes Intended Use Cases

Description Domain

Keywords: protein, canonical, glycosylation, glycan
Pipeline Steps:
Step 1: The input file was retrieved directly from source.
Step 2: The UniProtKB protein accessions in the input file were mapped to UniProt canonical accessions …..
Step 3: The glycosylation type [linkage type] was retrieved through UniCarbKB structure webpage using scripts [make-proteoform_glycosylation_sites_unicarbkb_glytoucan-csv-step2a.py,……..
Step 4: The file was processed for quality check using a python Records which fall under one or more following criteria's are flagged and eliminated [eliminated records can be accessed using log file]

Human Readable Description of Process

Execution Domain:

Script Access Type: Text
Scripts: make-proteoform_glycosylation_sites_unicarbkb_glytoucan-csv-step2a.py, make-proteoform_glycosylation_sites_unicarbkb_glytoucan-csv-step2b.py,
Script Location: https://github.com/glygener/glygen-backend/blob/master/integration/
Script Driver: manual
Platform: CentOS7

Explicit Computational Inputs and Processes

Software Prerequisites:

Name: Python
Version: 2.7.13

Software Requirements

(……………)

BioCompute Object example

```
I/O Domain
  Input Subdomain:
    uri: https://data.glygen.org/ln2wwwdata/source/human_glytoucan_140918_2018_10_31_02_17_32.txt
    filename: human_glytoucan_140918_2018_10_31_02_17_32.txt

    uri: https://data.glygen.org/ln2wwwdata/reviewed/human_protein_all.fasta
    filename : human_protein_all.fasta
```

Information on Input files required in the pipeline steps to create the dataset

```
Output Subdomain:
    mediatype: csv
    uri: https://data.glygen.org/ln2wwwdata/reviewed/human_proteoform_glycosylation_sites_unicarbkb_glytoucan.csv
    filename: human_proteoform_glycosylation_sites_unicarbkb_glytoucan.csv

    mediatype: csv
    uri: https://data.glygen.org/.../reviewed/human_proteoform_glycosylation_sites_unicarbkb_glytoucan.stat.csv
    filename: human_proteoform_glycosylation_sites_unicarbkb_glytoucan.csv

    Content:
      Column Headers:
        uniprotkb_canonical_ac: Accession assigned to the protei
                                UniProtKB database
        glycosylation_site: Site on the protein sequence where g
        evidence: NCBI PubMed Id (PMID) as evidence for the entry
        unicarbkb_id: UnicarbKB database identifier
        glytoucan_ac: Unique accession assigned to the registered glycan structure in GlyTouCan database
        amino_acid: Three letter abbreviation code of the amino acid
        glycosylation_type: Type of glycosylation (N/O/C/S linked glycosylation)

    Statistics [Unique Value]:
                uniprotkb_canonical_ac: 92
                glycosylation_site: 223
                evidence: 163
                uckb_id: 984
                glytoucan_acc: 824
                amino_acid: 3
                glycosylation_type: 3
```

Information on the output files Log files provide information on discarded/obsolete entries. The stat csv provides description of the column headers and the unique value statistics of the dataset.

Integration results

- Homo sapiens

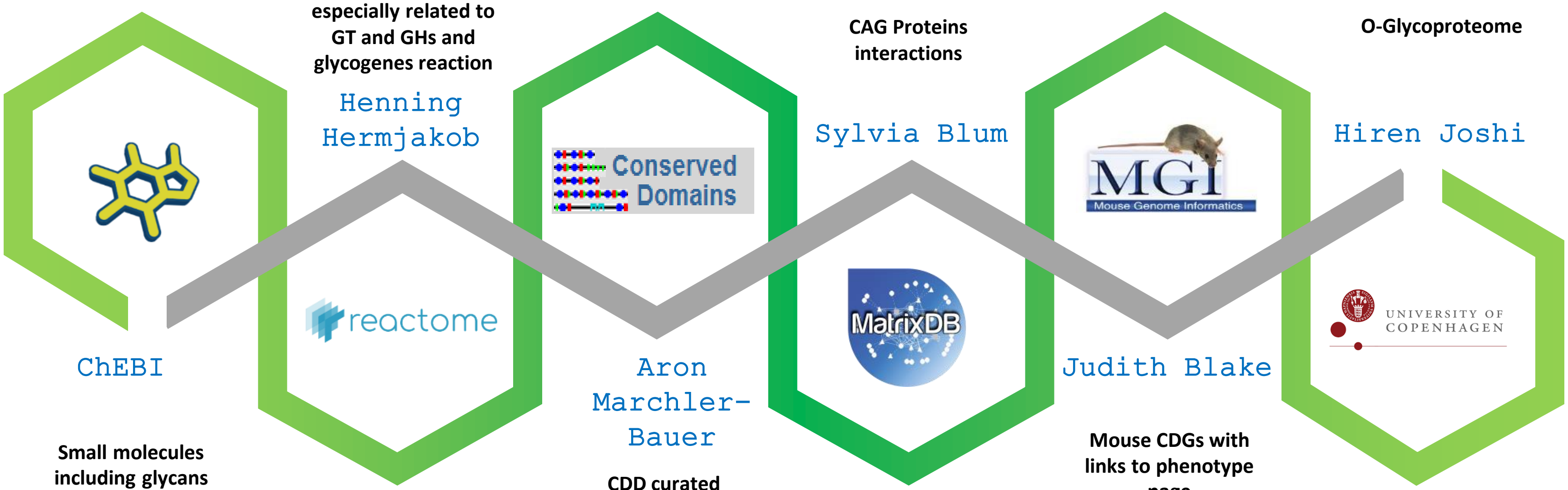
Glycans	3302
Proteins	21080
Glycoproteins	4530

- Mus musculus

Glycans	517
Proteins	22289
Glycoproteins	3792

https://data.glygen.org/

New Collaboration



ChEBI

Small molecules including glycans

Henning Hermjakob

Reactome data especially related to GT and GHs and glycogenes reaction

Aron Marchler-Bauer

CDD curated alignments for glycosyltransferase and glycosylhydrolase

Sylvia Blum

CAG Proteins interactions

Judith Blake

Mouse CDGs with links to phenotype page

Hiren Joshi

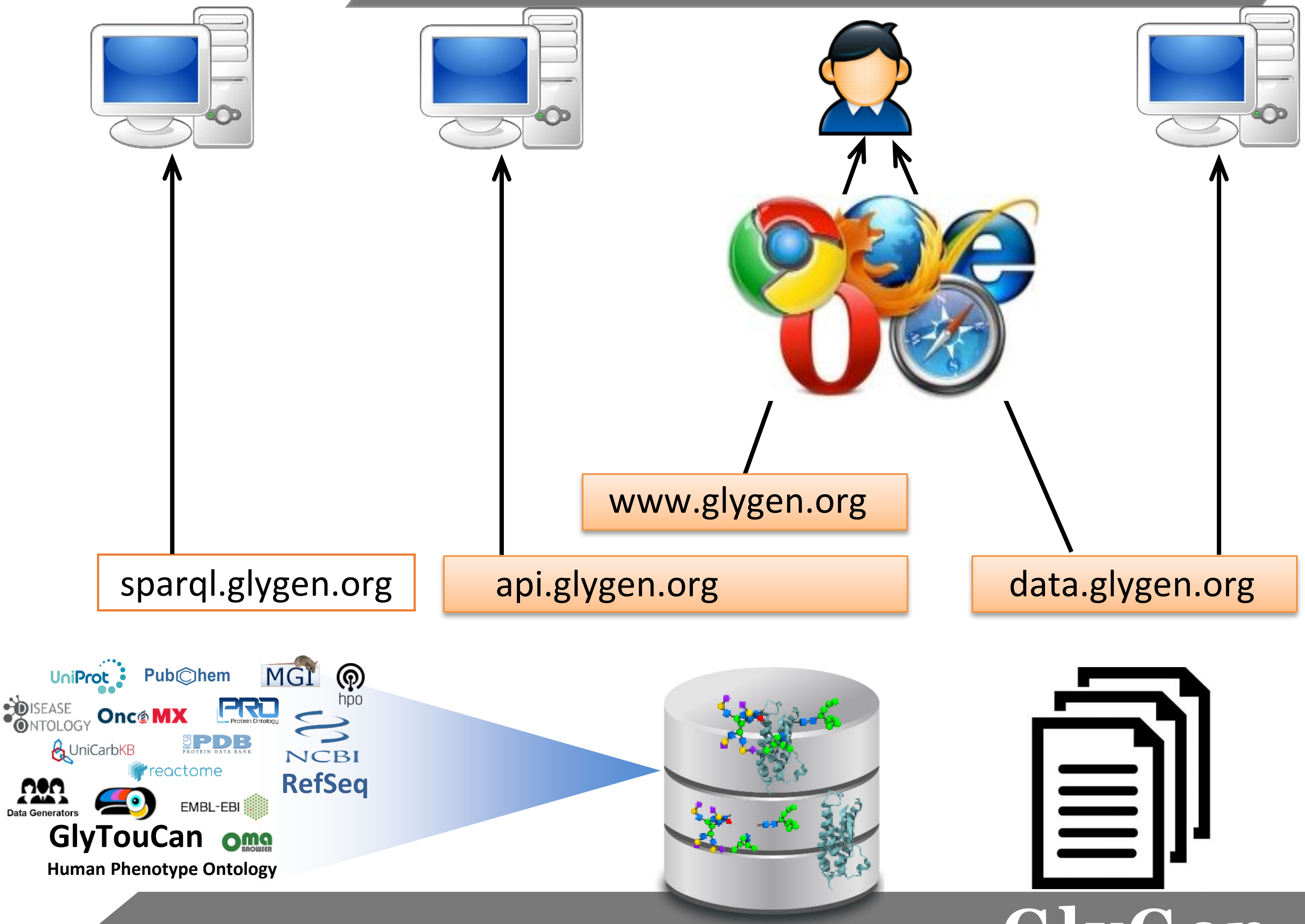
O-Glycoproteome

Table of content

- Data integration
- Data access
- Data sharing

GlyGen

Architecture



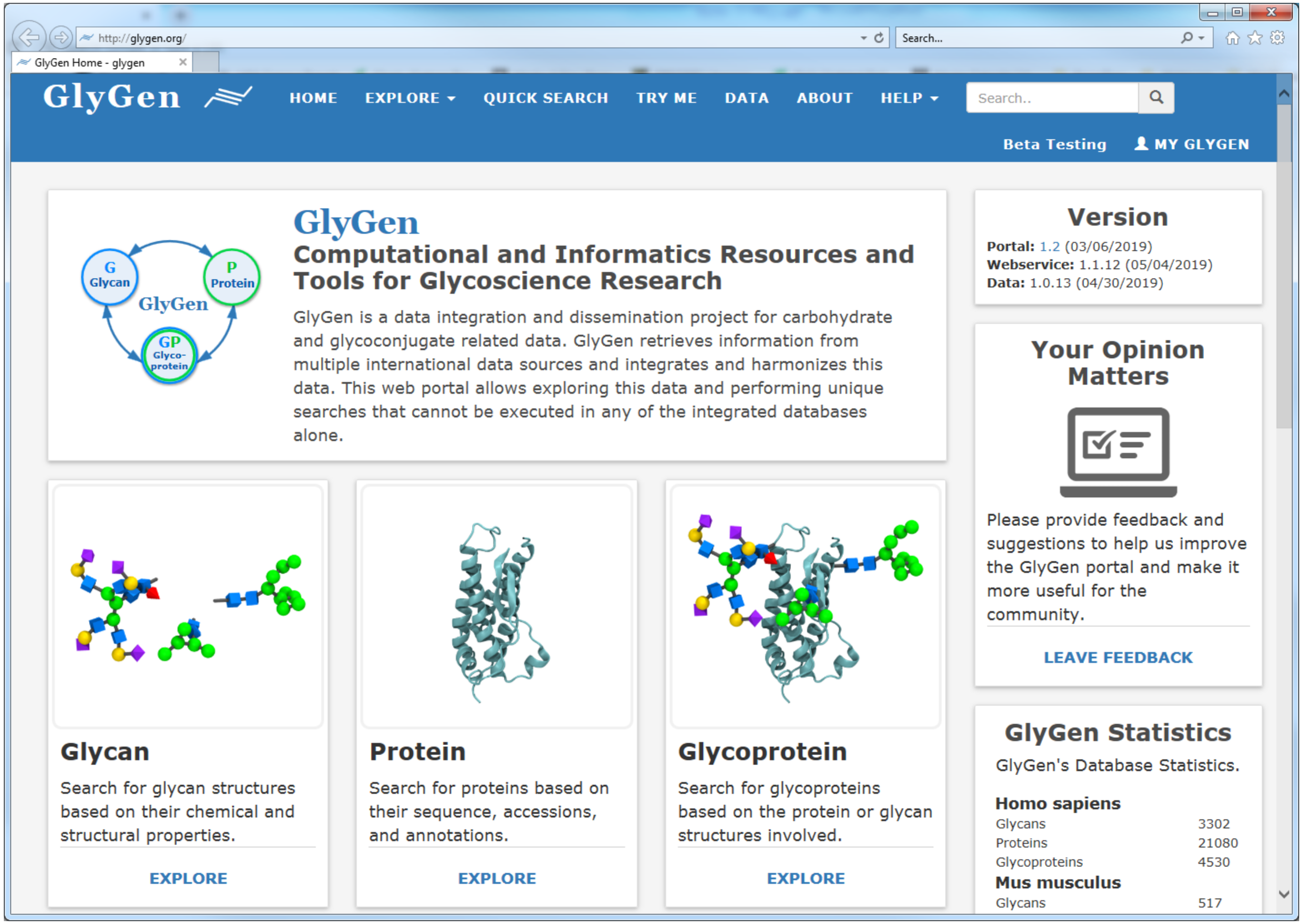
www.glygen.org

sparql.glygen.org

api.glygen.org

data.glygen.org

GlyGen http://www.glygen.org



GlyGen

Computational and Informatics Resources and Tools for Glycoscience Research

GlyGen is a data integration and dissemination project for carbohydrate and glycoconjugate related data. GlyGen retrieves information from multiple international data sources and integrates and harmonizes this data. This web portal allows exploring this data and performing unique searches that cannot be executed in any of the integrated databases alone.

Version

Portal: 1.2 (03/06/2019)
Webservice: 1.1.12 (05/04/2019)
Data: 1.0.13 (04/30/2019)

Your Opinion Matters

Please provide feedback and suggestions to help us improve the GlyGen portal and make it more useful for the community.

LEAVE FEEDBACK

Glycan

Search for glycan structures based on their chemical and structural properties.

EXPLORE

Protein

Search for proteins based on their sequence, accessions, and annotations.

EXPLORE

Glycoprotein

Search for glycoproteins based on the protein or glycan structures involved.

EXPLORE

GlyGen Statistics

GlyGen's Database Statistics.

Homo sapiens	
Glycans	3302
Proteins	21080
Glycoproteins	4530
Mus musculus	
Glycans	517

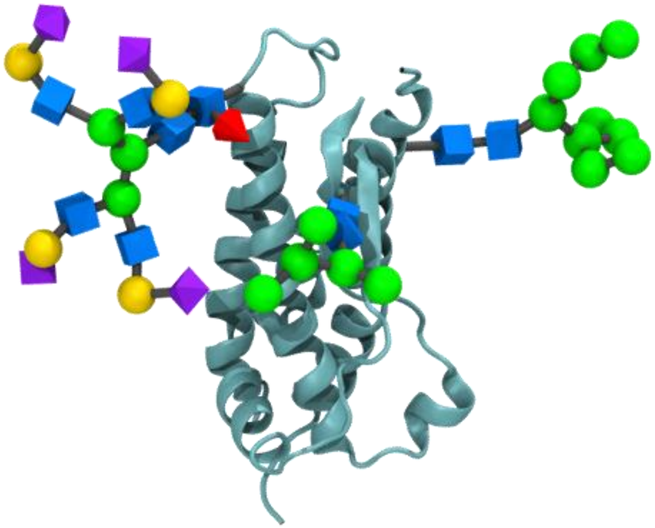
GlyGen.org

Search glycans by …

- GlyTouCan ID
- Mass
- Number of Monosaccharides
- Organism
- Glycan Type
- Protein Id / name
- Motif
- Enzyme

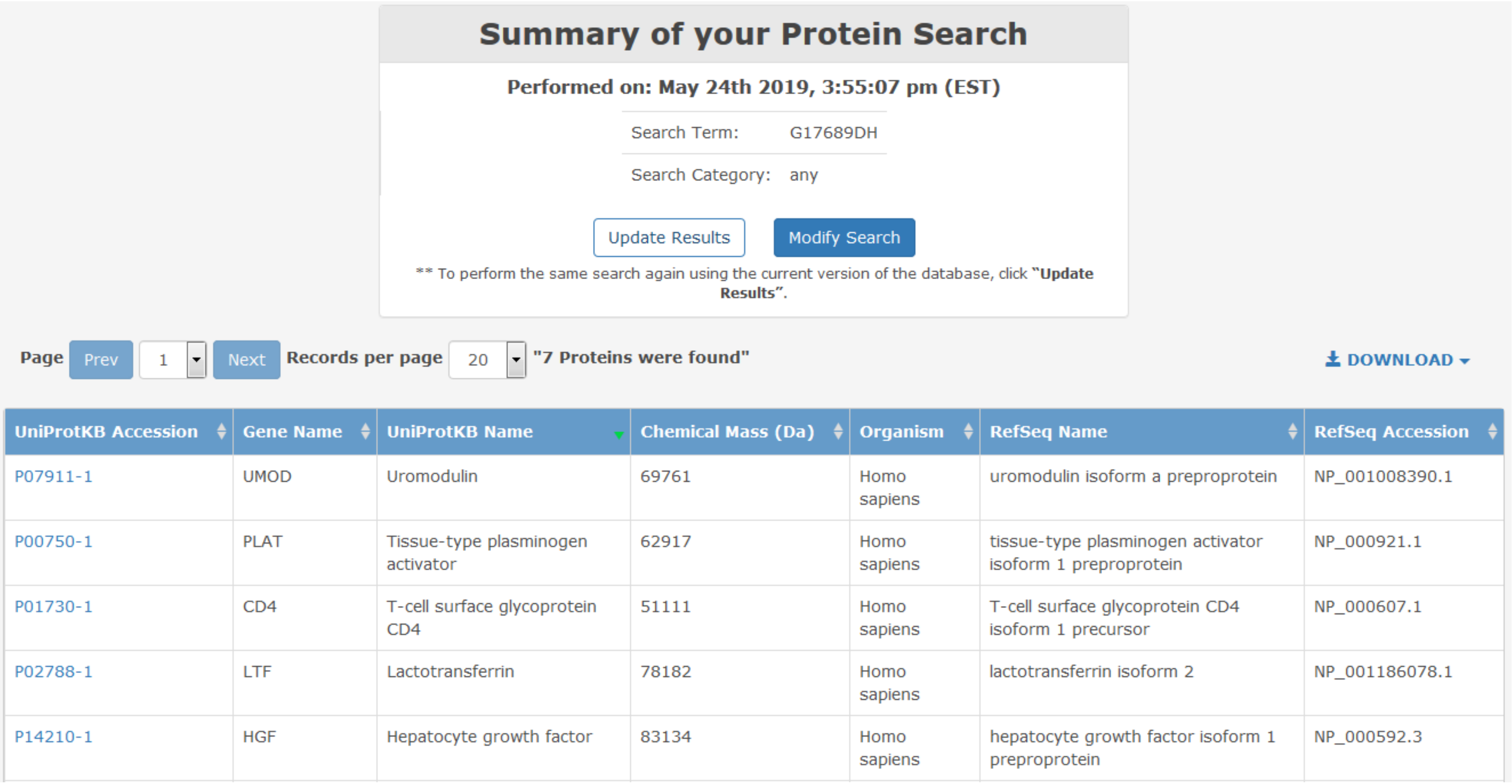
Search proteins by ...

- Accession (UniProt, RefSeq)
- Mass
- Organism
- Protein name
- Gene name
- Attached glycan
- Glycosylated amino acid
- Sequence
- Pathway



GlyGen

Search results



Summary of your Protein Search

Performed on: May 24th 2019, 3:55:07 pm (EST)

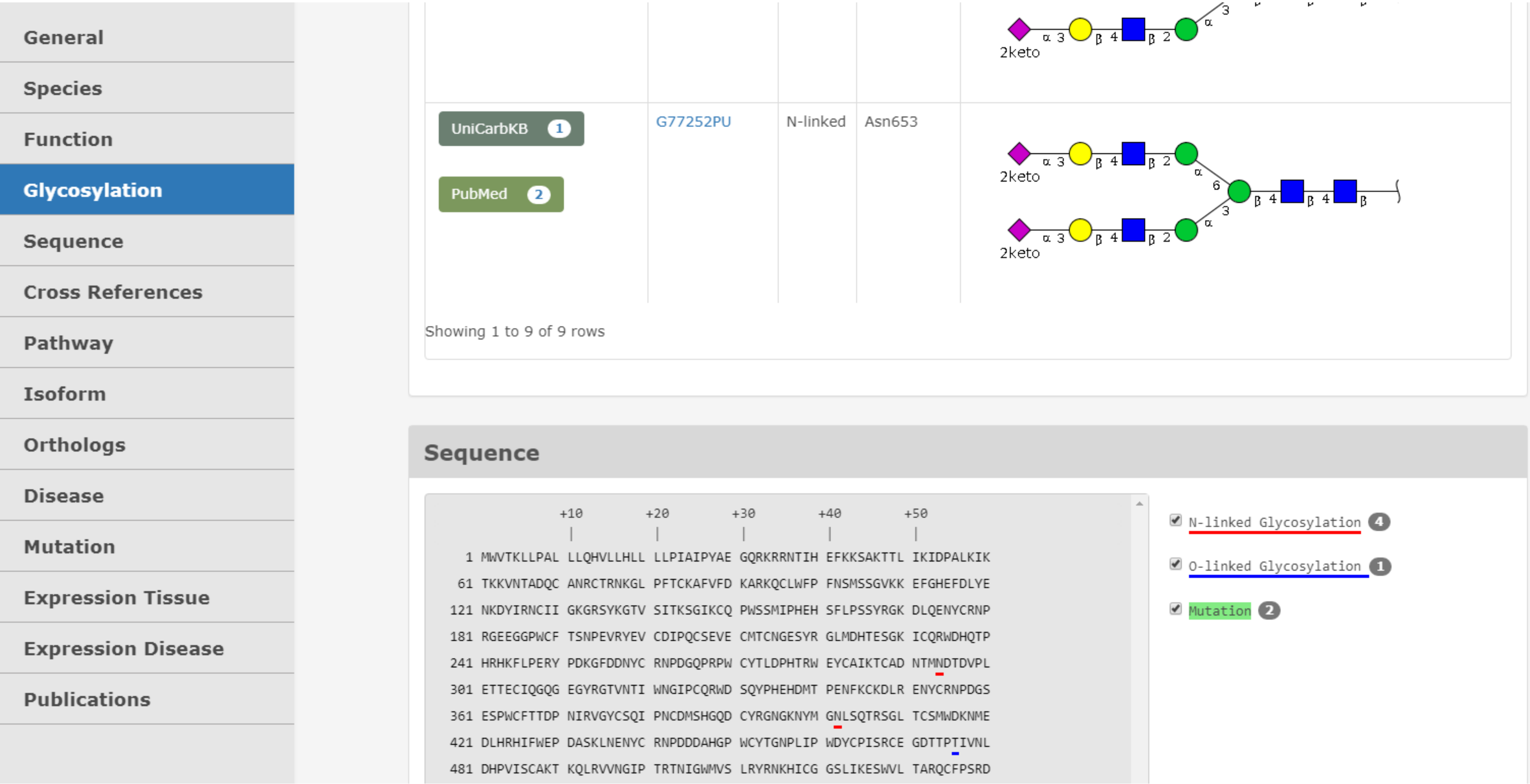
Search Term:	G17689DH
Search Category:	any

** To perform the same search again using the current version of the database, click **"Update Results"**.

"7 Proteins were found"

UniProtKB Accession	Gene Name	UniProtKB Name	Chemical Mass (Da)	Organism	RefSeq Name	RefSeq Accession
P07911-1	UMOD	Uromodulin	69761	Homo sapiens	uromodulin isoform a preproprotein	NP_001008390.1
P00750-1	PLAT	Tissue-type plasminogen activator	62917	Homo sapiens	tissue-type plasminogen activator isoform 1 preproprotein	NP_000921.1
P01730-1	CD4	T-cell surface glycoprotein CD4	51111	Homo sapiens	T-cell surface glycoprotein CD4 isoform 1 precursor	NP_000607.1
P02788-1	LTF	Lactotransferrin	78182	Homo sapiens	lactotransferrin isoform 2	NP_001186078.1
P14210-1	HGF	Hepatocyte growth factor	83134	Homo sapiens	hepatocyte growth factor isoform 1 preproprotein	NP_000592.3

Protein details



UniCarbKB 1 PubMed 2	G77252PU	N-linked	Asn653	

Showing 1 to 9 of 9 rows

Sequence

```
          +10        +20        +30        +40        +50
           |          |          |          |          |
  1 MWVTKLLPAL LLQHVLLHLL LLPIAIPYAE GQRKRRNTIH EFKKSAKTTL IKIDPALKIK
 61 TKKVNTADQC ANRCTRNKGL PFTCKAFVFD KARKQCLWFP FNSMSSGVKK EFGHEFDLYE
121 NKDYIRNCII GKGRSYKGTV SITKSGIKCQ PWSSMIPHEH SFLPSSYRGK DLQENYCRNP
181 RGEEGGPWCF TSNPEVRYEV CDIPQCSEVE CMTCNGESYR GLMDHTESGK ICQRWDHQTP
241 HRHKFLPERY PDKGFDDNYC RNPDGQPRPW CYTLDPHTRW EYCAIKTCAD NTMNDTDVPL
301 ETTECIQGQG EGYRGTVNTI WNGIPCQRWD SQYPHEHDMT PENFKCKDLR ENYCRNPDGS
361 ESPWCFTTDP NIRVGYCSQI PNCDMSHGQD CYRGNGKNYM GNLSQTRSGL TCSMWDKNME
421 DLHRHIFWEP DASKLNENYC RNPDDDAHGP WCYTGNPLIP WDYCPISRCE GDTTPTIVNL
481 DHPVISCAKT KQLRVVNGIP TRTNIGWMVS LRYRNKHICG GSLIKESWVL TARQCFPSRD
```

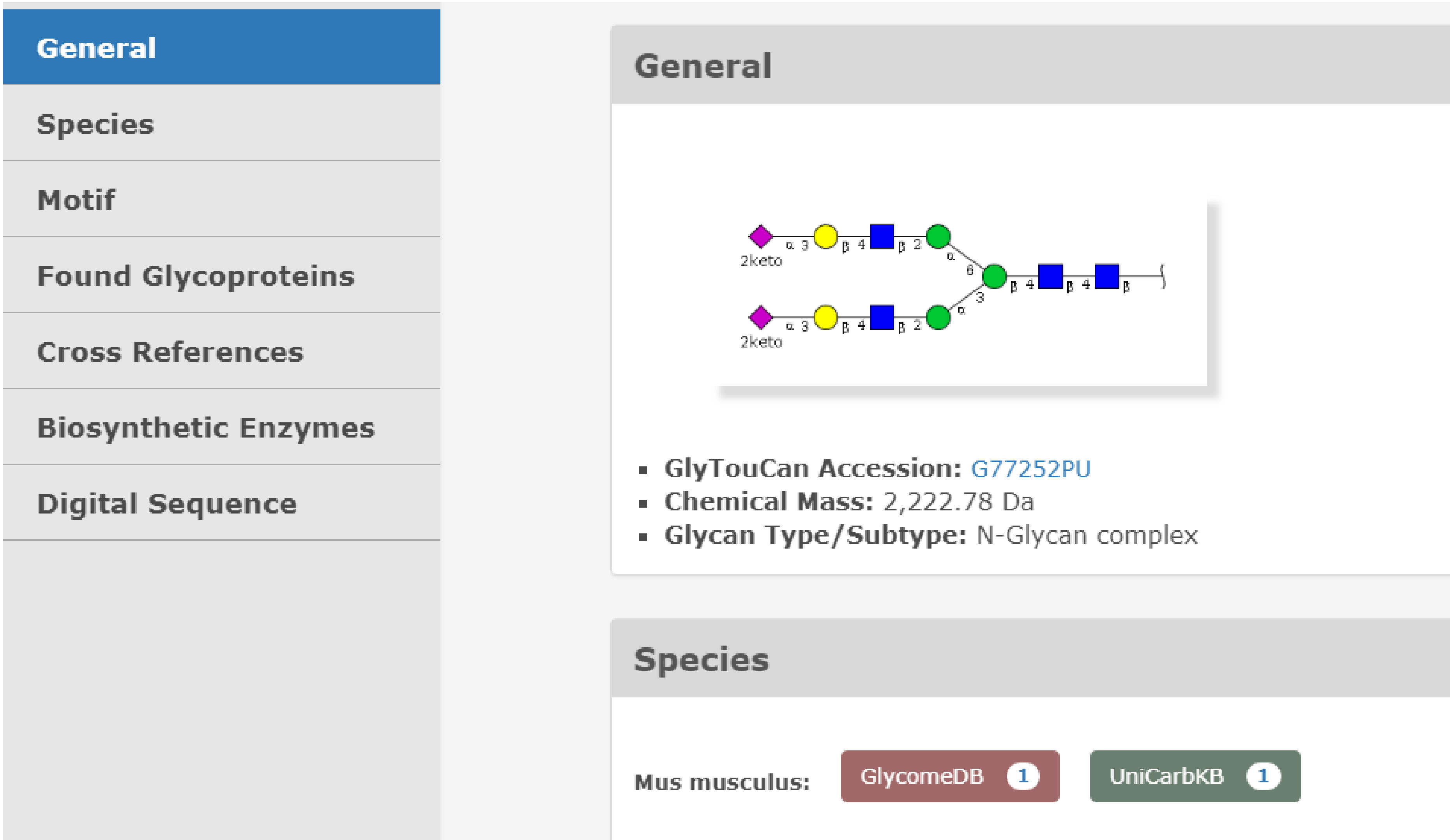
☑ N-linked Glycosylation 4

☑ O-linked Glycosylation 1

☑ Mutation 2

GlyGen

Glycan details



- General
- Species
- Motif
- Found Glycoproteins
- Cross References
- Biosynthetic Enzymes
- Digital Sequence

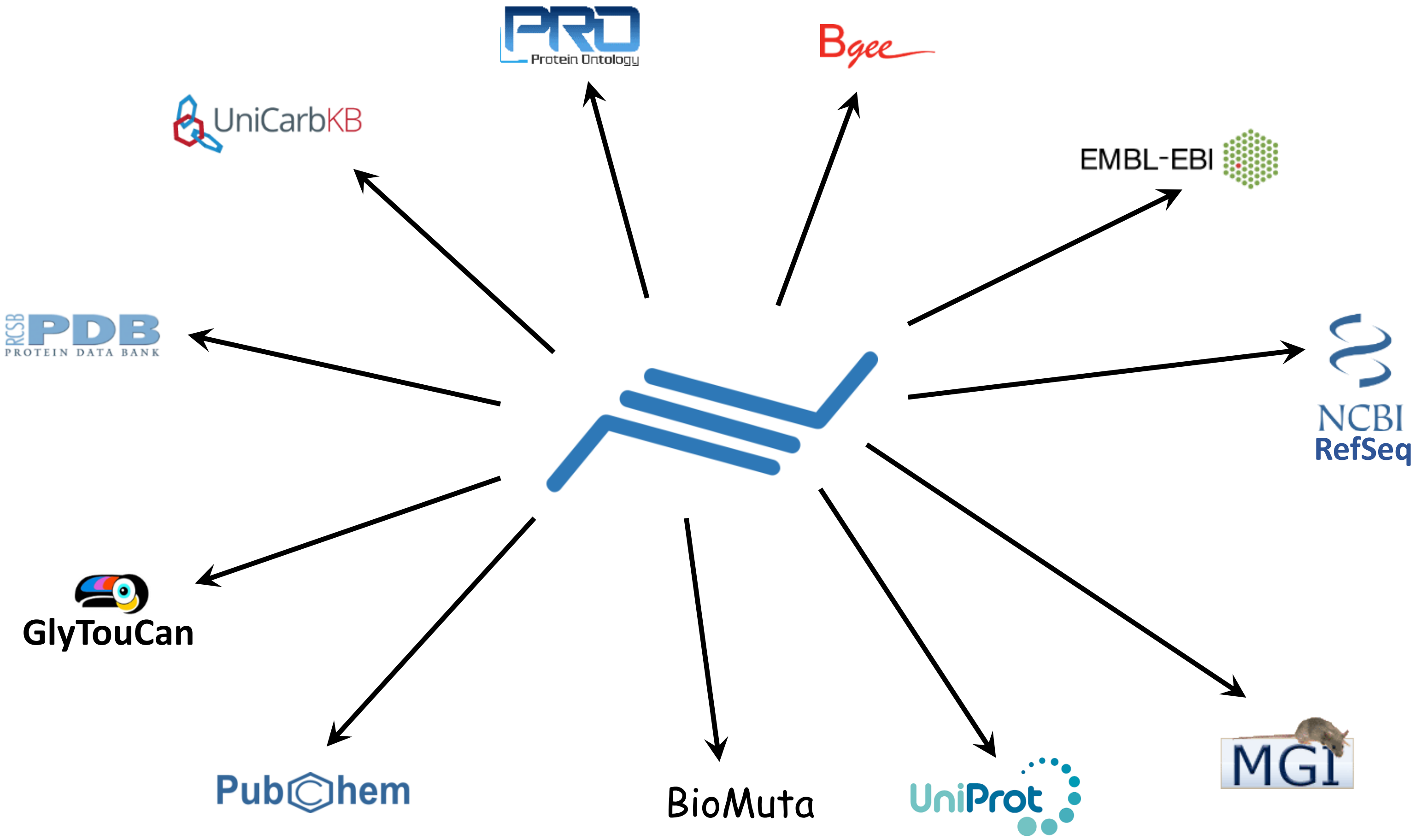
General

- **GlyTouCan Accession:** G77252PU
- **Chemical Mass:** 2,222.78 Da
- **Glycan Type/Subtype:** N-Glycan complex

Species

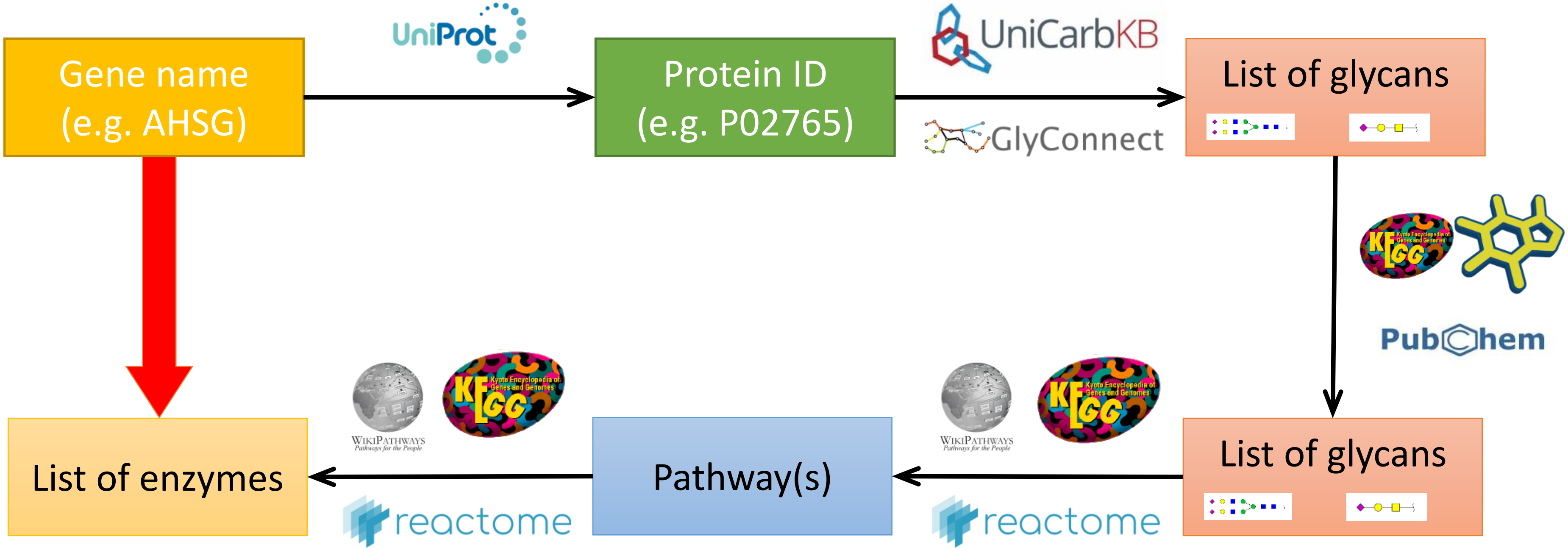
Mus musculus: GlycomeDB 1 UniCarbKB 1

GlyGen as a Hub

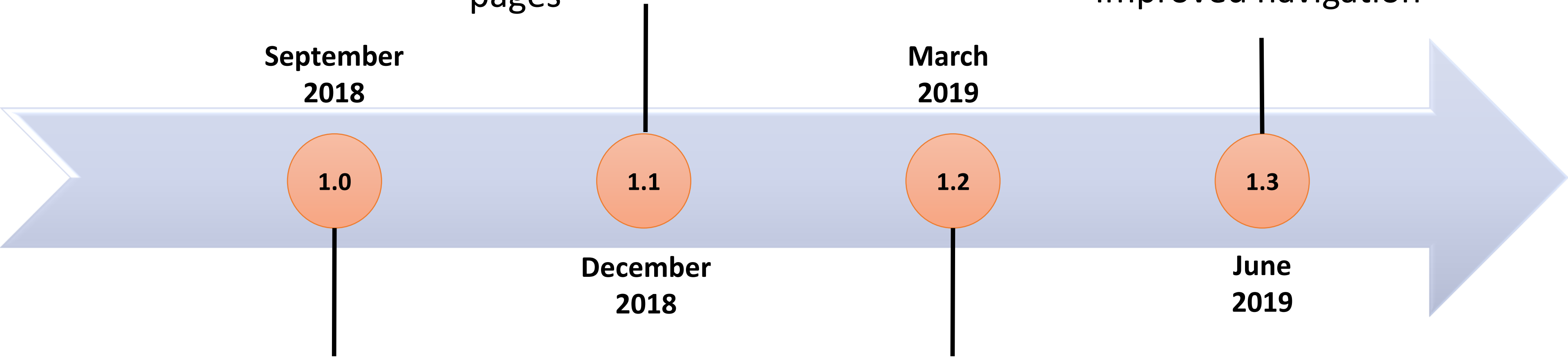


Quick Search

- Get all **enzymes** that my of been involved in *synthesis of the glycans* on protein X.



Releases



September 2018

1.0

- Human and Mouse data
- Basic search for glycan, proteins and glycoproteins
- Information pages for glycan, proteins and glycoproteins

December 2018

1.1

- Simple and Advance search options
- Example input for searches
- Download of data on all pages

March 2019

1.2

- Improved Evidence display
- Motif pages
- New help feature
- Global search

June 2019

1.3

- Feedback feature on all pages
- Orthologous for proteins
- Improved sequence display
- Improved navigation

What's Next

- Data from new sources
- New species (Rat, HIV)
- Improved help system for users
- More use cases
- 3D structures for glycans, proteins and glycoproteins
- Better widgets to search for and display data

Feedback

- Try the webpage – give feedback
- What are the questions you would like to ask
- Enriching the database with own data

Table of content

- Data integration
- Data access
- Data sharing

License

www.glygen.org
data.glygen.org
api.glygen.org

https://github.com/glygener

creative commons

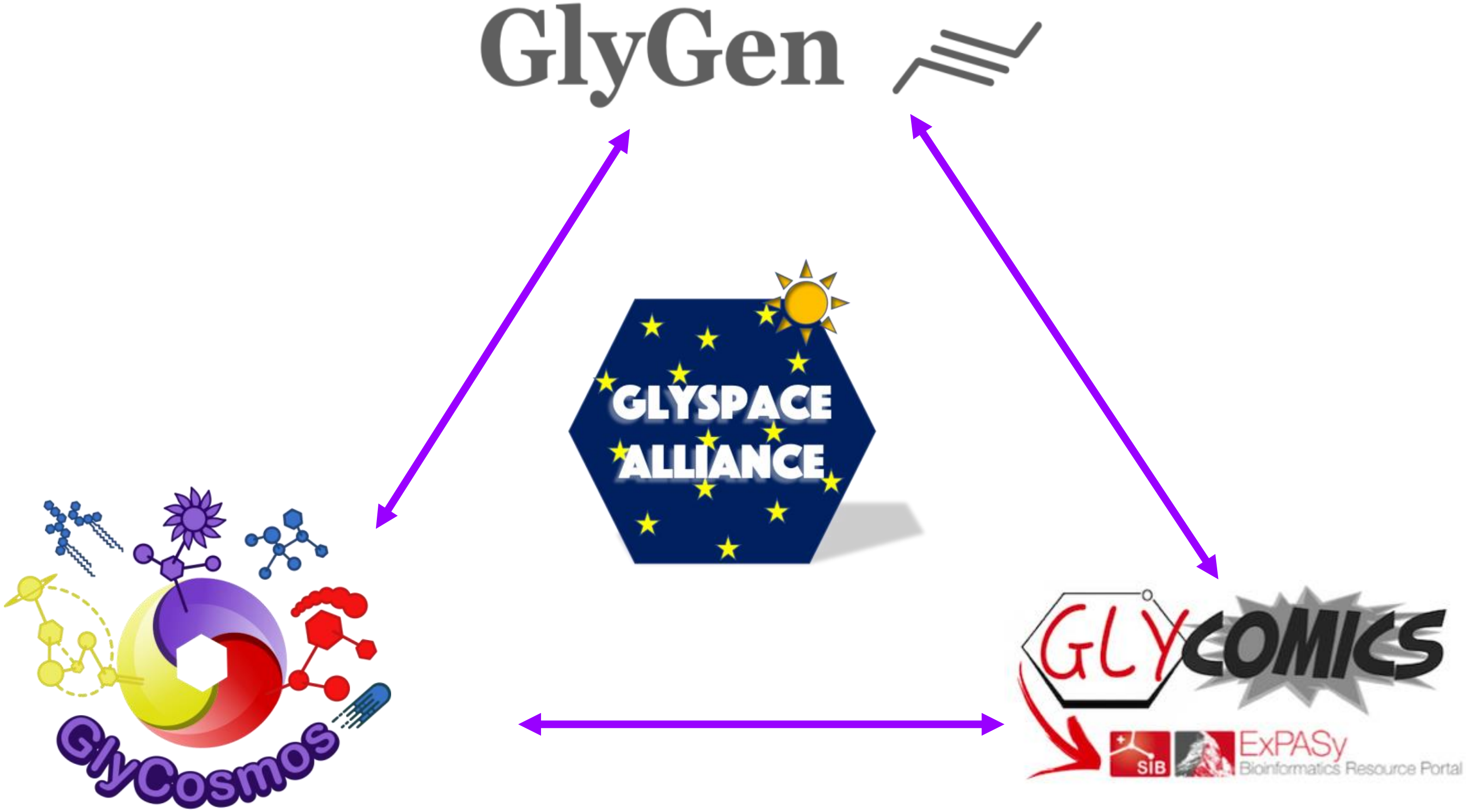
Attribution 4.0 International

GPL 3 Free Software

Free as in Freedom

GitHub

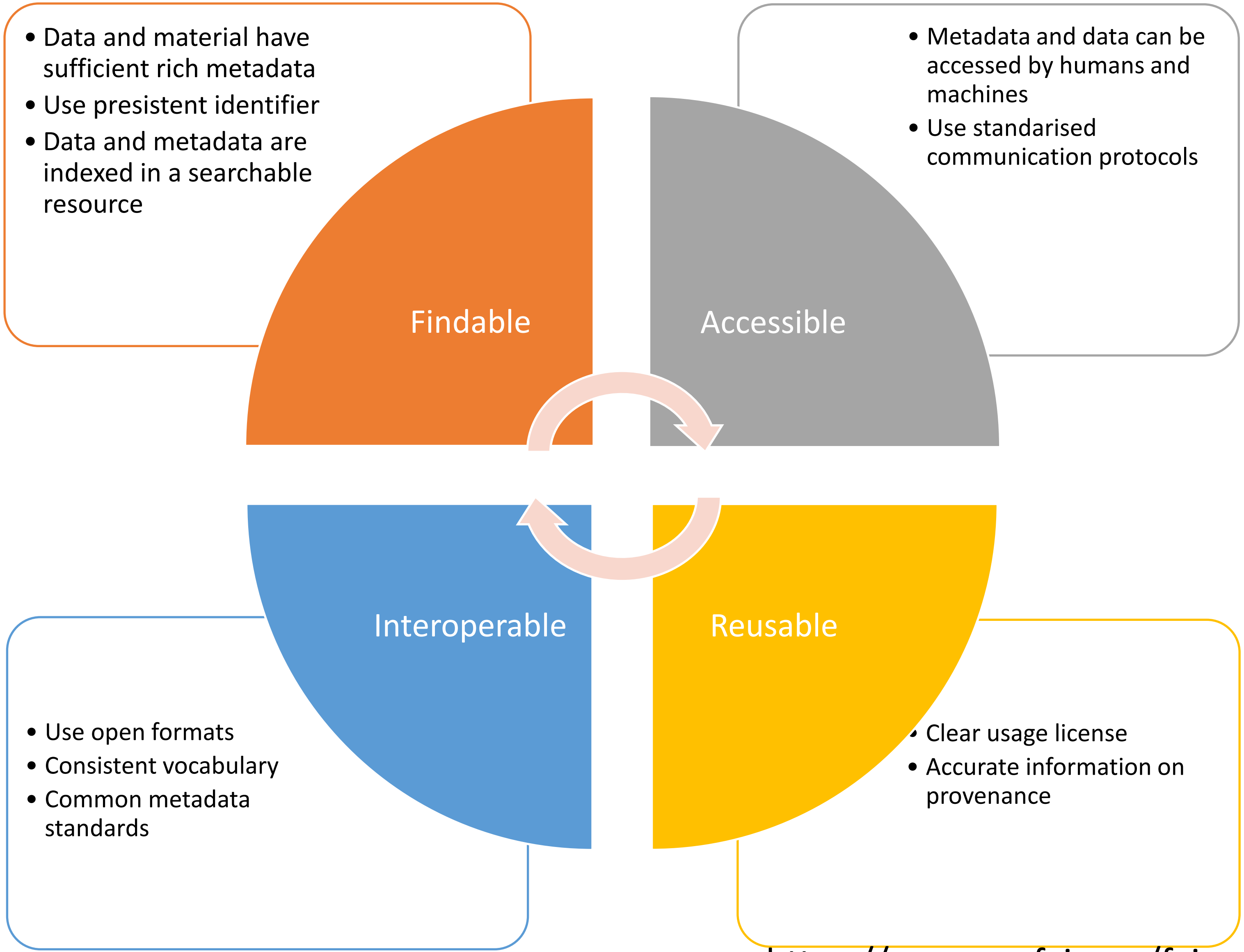
Glyspace Alliance



Established August, 2018 @ Warren Workshop

http://glyspace.org

FAIR data



- Data and material have sufficient rich metadata
- Use presistent identifier
- Data and metadata are indexed in a searchable resource

- Metadata and data can be accessed by humans and machines
- Use standarised communication protocols

- Use open formats
- Consistent vocabulary
- Common metadata standards

- Clear usage license
- Accurate information on provenance

https://www.go-fair.org/fair-principles/

GlyGen

Acknowledgements

University of Georgia
René Ranzinger
Michael Pierce
Robert Woods
Rupali Mahadik
Tatiana Williamson
Sena Arpinar
Sanath Bhatt
Sujeet Kulkarni
Sandeep Nakarakommula

EMBL-EBI
Maria Martin
Leyla Jael Garcia Castro
Preethi Vasudev

NCBI
Kim Pruitt
Evan Bolton

The George Washington University
Raja Mazumder
Robel Kahsay
Jeet Vora
Rahi Navelkar
Reza Mousavi
Nagarajan Pattabiraman
Xavier Holmes
Brian Fochtman

Georgetown University
Nathan Edwards
Radoslav Goldman
Darren Natale
Karen Ross
Wenjin Zhang

Harvard University
Richard Cummings

The Jackson Laboratory
Judith Blake

Soka University
Kiyoko Aoki-Kinoshita

The Griffith University
Matthew Campbell

Imperial College London
Ten Feizi

Macquarie University
Nicki Packer

NIH-NCI
Jefferey Gildersleeve

NIH Grant - U01 GM125267-01

UNIVERSITY OF GEORGIA · THE GEORGE WASHINGTON UNIVERSITY WASHINGTON, DC · GEORGETOWN UNIVERSITY · SOKA UNIVERSITY · Griffith UNIVERSITY · HARVARD UNIVERSITY · NCBI · Imperial College London · EMBL-EBI

Web portal: *https://glygen.org*
Data store: *https://data.glygen.org*
WS API: *https://api.glygen.org*

CONTACT
Will York will@ccrc.uga.edu
Raja Mazumder mazumder@gwu.edu

GlyGen.org