GlycoData: A GlycoMIP initiative glycomaterials database

Fangxi Wang[1], Lakshmi K. Kunche[1], Hunter Irving[2], David Montgomery[3] and Sanket A. Deshmukh[1*]

1) Department of Chemical Engineering, Virginia Tech, Blacksburg, VA
2) Advanced Research Computing, Virginia Tech, Blacksburg, VA
3) Complex Carbohydrate Research Center, University of Georgia, Athens, GA

Motivation

The application of data-driven methodology and artificial intelligence has sparked significant interest in the development of novel materials, particularly glycomaterials. With the scientific community actively working on glycomaterials, a wealth of theoretical and experimental data is being generated worldwide, and the amount of data is increasing exponentially over time. However, the majority of this data remains difficult to use for data-driven materials exploration, owing to the lack of an appropriate data management framework that would enable the data to be findable, accessible, interoperable, and reusable. To overcome these challenges, GlycoMIP, a Materials Innovation Platform funded by the National Science Foundation (NSF), is actively working to establishing a novel data storage and management platform, named **GlycoData** (https://www.glycodata.org). This platform will allow researchers within and beyond the GlycoMIP community to share both structured and unstructured data for scientific inquiry.

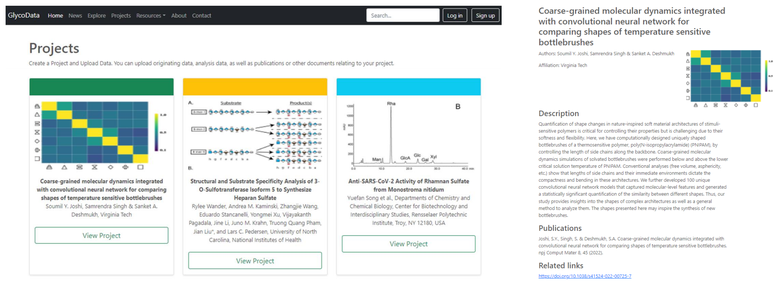
Objectives

- To create a platform that contains glycomaterials data and machine learning models and tools developed *via* GlycoMIP to study glycomaterials to the scientific community.
- To enhance the visibility and impact of users' research. By making their data accessible and discoverable, researchers can reach a broader audience and improve their research reputation.
- To facilitate new discoveries and innovations. Researchers can use existing data to validate their hypotheses, identify knowledge gaps, and generate novel insights.
- To encourage collaborations. Researchers can collaborate more effectively by making data available via GlycoDATA and create new interdisciplinary research opportunities.
- To make education and training resources available to glycomaterials community, which can help them acquire the necessary skills and knowledge to work with complex data types.
- To accelerate the design of glycomaterials by providing a centralized repository of data that can be used to train and validate machine learning models.

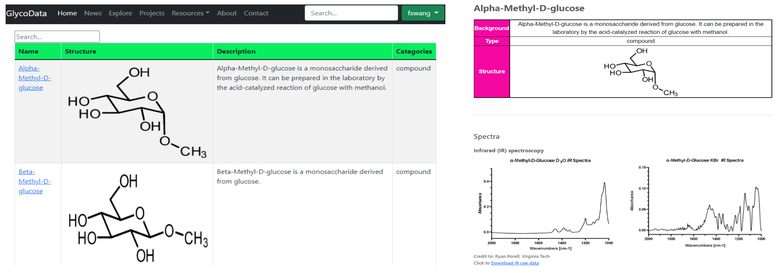
Data Collection and Management

Currently, the **GlycoData** platform consists of three tables (database objects that contain all the data in a database) for storing data.

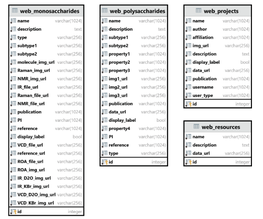
1) Projects table: Users can create projects that link to their published work or create a new, open-access project. A project description is required and a graphical abstract is recommended. To accommodate projects with large datasets, users can also share a link to raw data housed in Google Drive or on GitHub.



2) Monosaccharides table: The experimental data of each monosaccharide characterized using GlycoMIP facilities is provided, including Infrared (IR), Vibrational Circular Dichroism (VCD), Raman Optical Activity (ROA), and Nuclear Magnetic Resonance (NMR) spectra. This experimental raw data can be download by registered users. We believe that this data holds considerable value for those who wish to utilize machine learning approaches in big data projects.



3) Polysaccharides table: This table currently stores data from published, in-house GlycoMIP projects. We are reaching out to the Principal Investigators to obtain their experimental and/or computational raw data. The table schema we use to catalog and present this information will evolve based on the data formats.

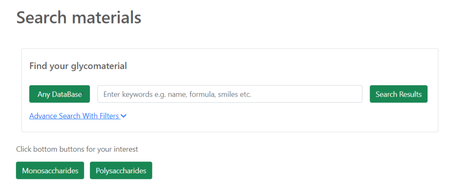


4) Resources table: This table summarizes many existing glycomaterials databases for users' convenience.

We will constantly update our table schemas based on the available data and build up the relationships between tables.

User Interface for Search

Users can utilize advanced filters to search for information on specified databases.



Future Work

- We are closely engaging with the GlycoMIP communities to gather data and update existing or create new tables for different categories of glycomaterials.
- The simulation tools, machine learning scripts, and analysis tools are under development. They will be delivered soon.
- We will build a file management module for internal short-term data sharing.
- We welcome feedback from **Glycodata** users. We will use their comments to enhance the experience and improve our database.

Join Today: glycodata.org/signup

Acknowledgement

This work was supported by GlycoMIP, a National Science Foundation Materials Innovation Platform funded through Cooperative Agreement DMR-1933525.

Advanced Research Computing

Computational Design of Hybrid Materials (CDHM)