

Breakout Session 3: Track A

Systems Biology of Glycosylation: Extending Mechanistic Analysis Toward AI

Dr. Sriram Neelamegham

Professor/PI, University at Buffalo, State University of New York

Dr. Rudiyanto Gunawan

Associate Professor, State University of New York - Buffalo

Systems Biology of Glycosylation: Extending Mechanistic Analysis Toward ML/DL

Sriram Neelamegham

Rudiyanto Gunawan

Changyou Chen

Chemical & Biological Engineering, and Computer Science and Engineering
State University of New York, Buffalo

Summary of the project and project goals

Overarching goal: To conduct Systems Biology experimental and computational investigations in order to understand how **gene expression** and **cellular epigenetics** regulate **glycan biosynthesis** at the single cell level

Aim 1. To enhance the depth of single-cell multi-omics studies in order to collect sufficient data for ML/DL.

Aim 2. To process and normalize data for ML/DL applications.

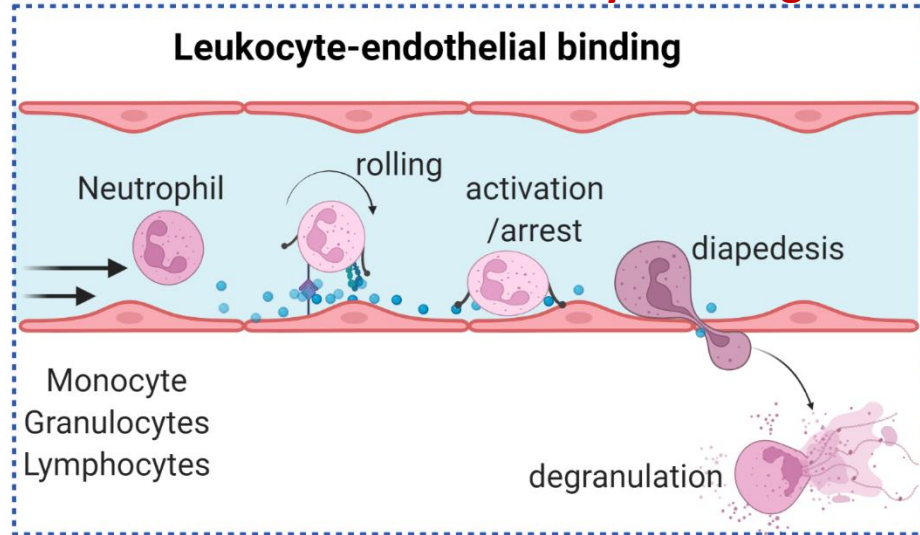
Aim 3. To demonstrate the use of the transformed data in a DL/ML application.

Highlights: Research outputs and shared data

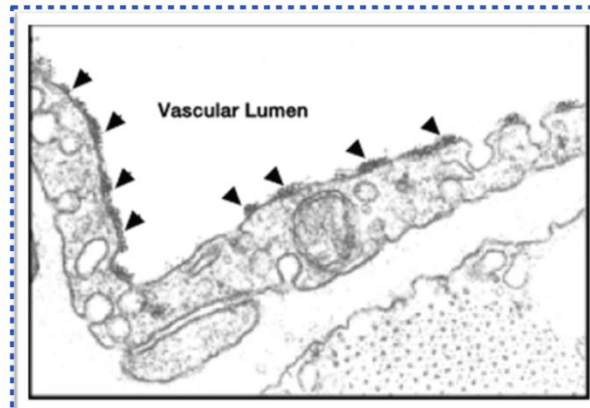
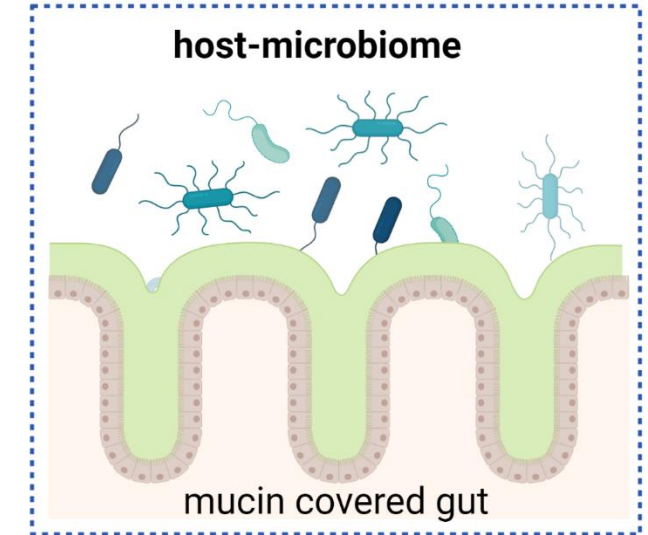
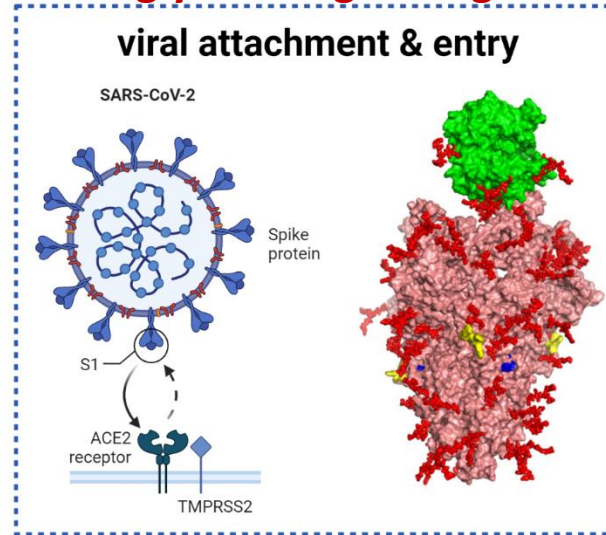
- Papers (other manuscripts are in preparation):
 - a. P. Chrysinas, C. Chen, and R. Gunawan. CrossTx: Cross-cell-line transcriptomic signature predictions. *Processes*, 12:332, 2024.
 - b. Cell and tissue-specific glycosylation pathways and transcriptional regulation informed by single-cell transcriptomics. *bioRxiv*, 559616, 2023.
- Website:
 - a. glycoCARTA: Single-cell transcriptome of glycosylation.
<http://vgdev.cedar.buffalo.edu/glycocarta/>
 - b. glycoTF: Transcriptional factors of glycosylation.
<http://vgdev.cedar.buffalo.edu/glycotf/>

Ubiquitous in nature and relevant to biotechnology

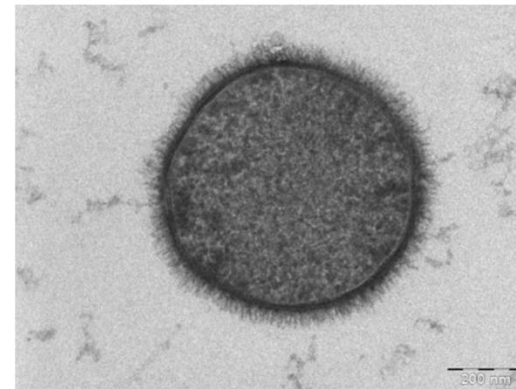
Role of selectins in leukocyte rolling



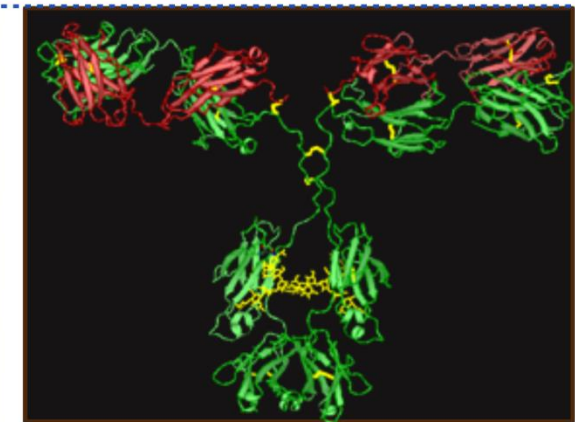
O- and N-glycans regulating SARS-CoV-2



Glycans on rat capillary endothelial cells
(Essentials of Glycobiology; George Palade)

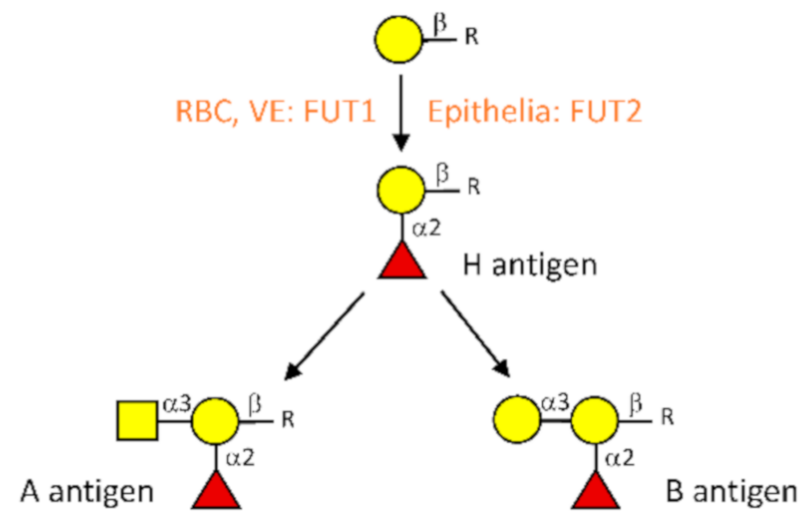


B. subtilis bacterium, with hair-like glycoalyx (Wikipedia)



biotechnology

Most common glycans are blood group antigen: Expressed on RBCs and critical for transfusion medicine



9 human monosaccharides

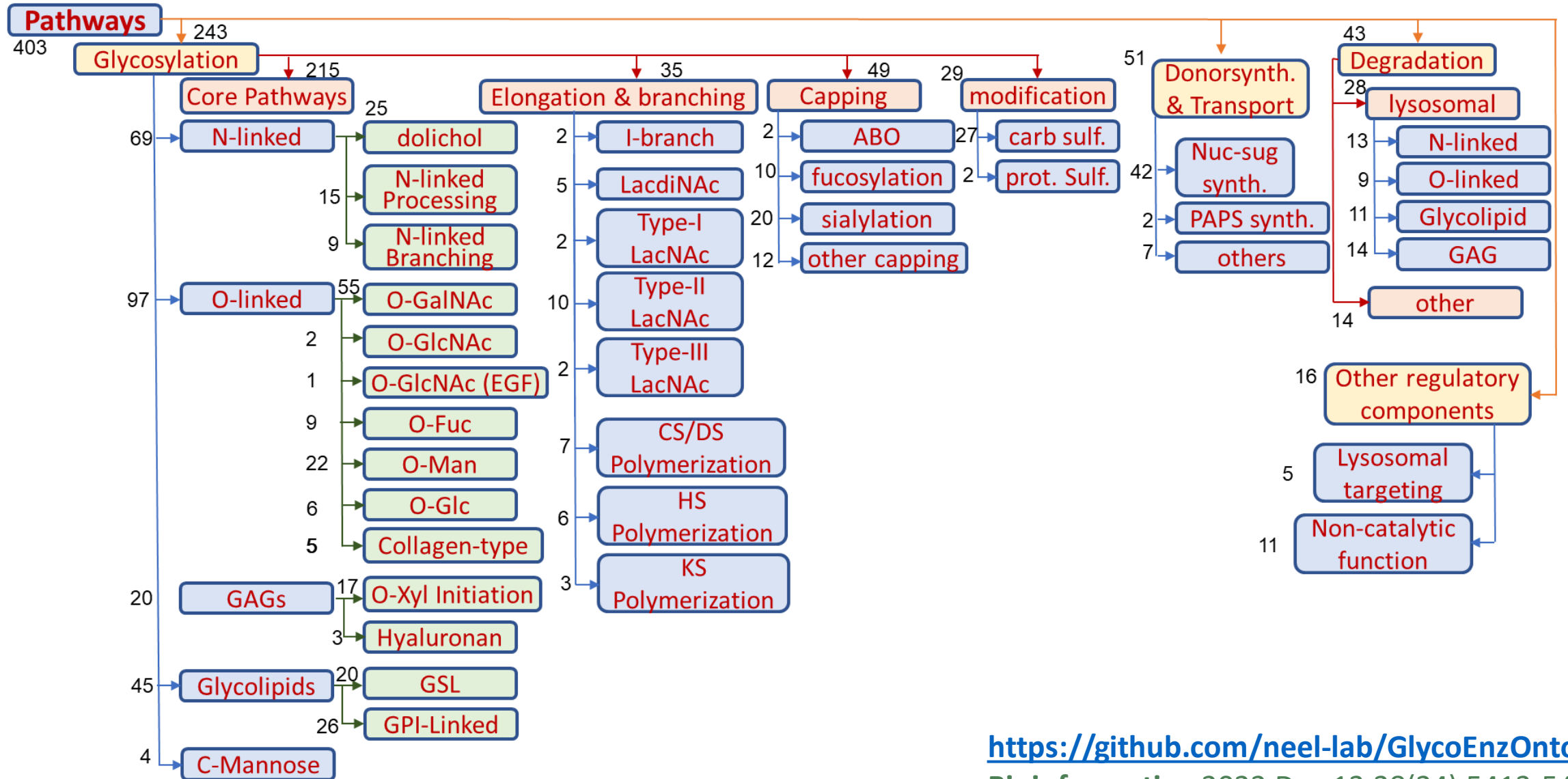
- Galactose (Gal) → UDP-
- N-Acetylgalactosamine (GalNAc) → UDP-
- Glucose (Glc) → UDP-
- N-Acetylglucosamine (GlcNAc) → UDP-
- Mannose (Man) → GDP-
- Fucose (Fuc) → GDP-
- N-Acetylneuraminic acid (Neu5Ac) → CMP-
- Xylose (Xyl) → UDP-
- Glucuronic acid (GlcA) → UDP-

Genotypes	O/O	A/O	A/A	B/O	B/B	A/B
Phenotypes	O	A	A	B	B	AB
Antigens	H	A	A	B	B	AB
Antibodies	anti-A anti-B	anti-B	anti-B	anti-A	anti-A	none

<https://www.ncbi.nlm.nih.gov/glycans/snfg.html>

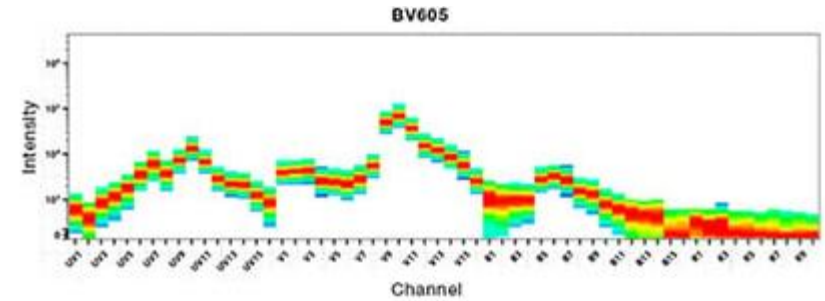
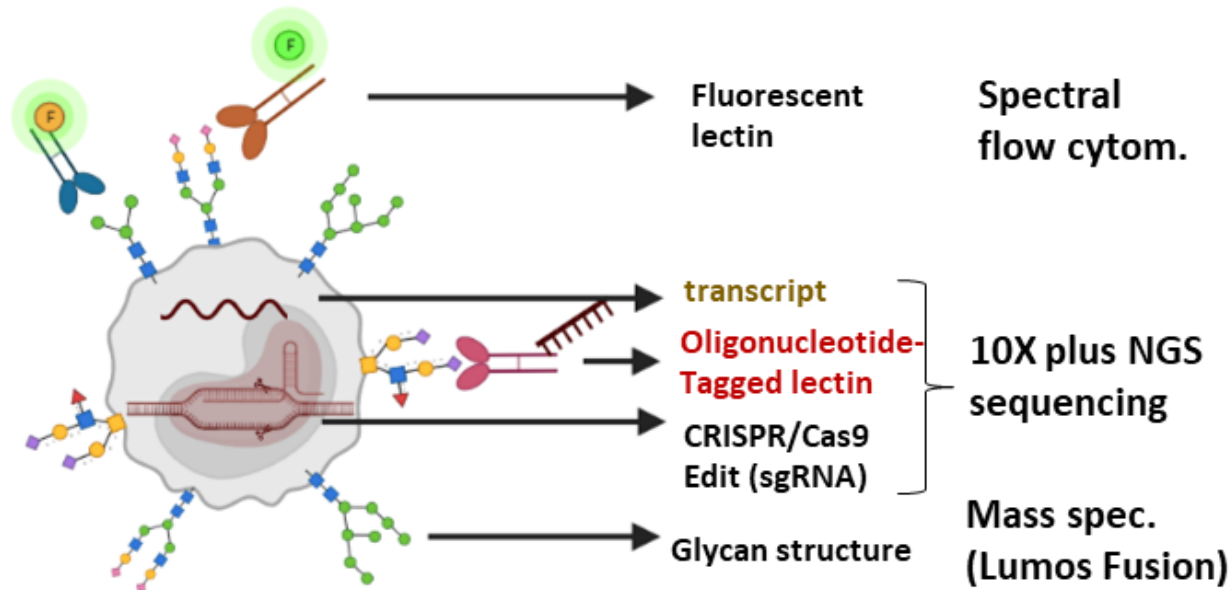
Neelamegham, et al. *Glycobiology*, 29:620-624, 2019

GlycoEnzOnto: An ontology for human glycosylating enzymes

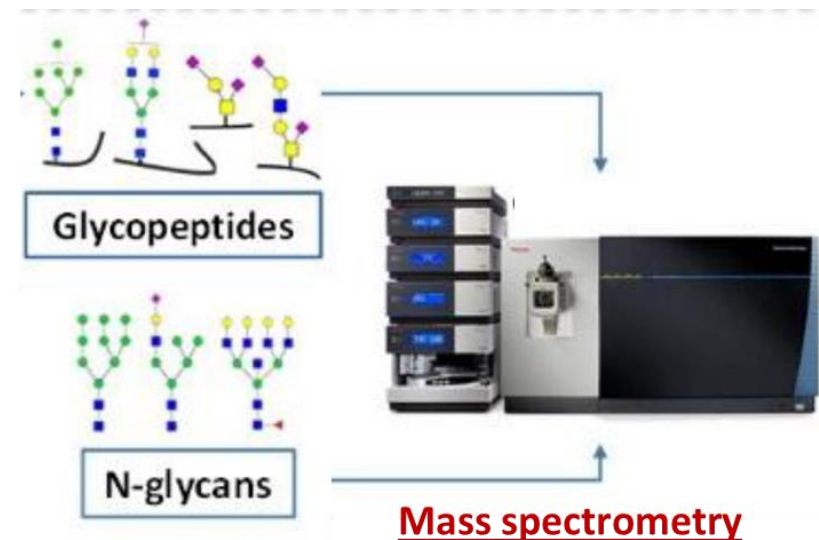


Multimodal measurements and their data integration

Single-cell multiOMICs



Spectral flow cytometry



Mass spectrometry

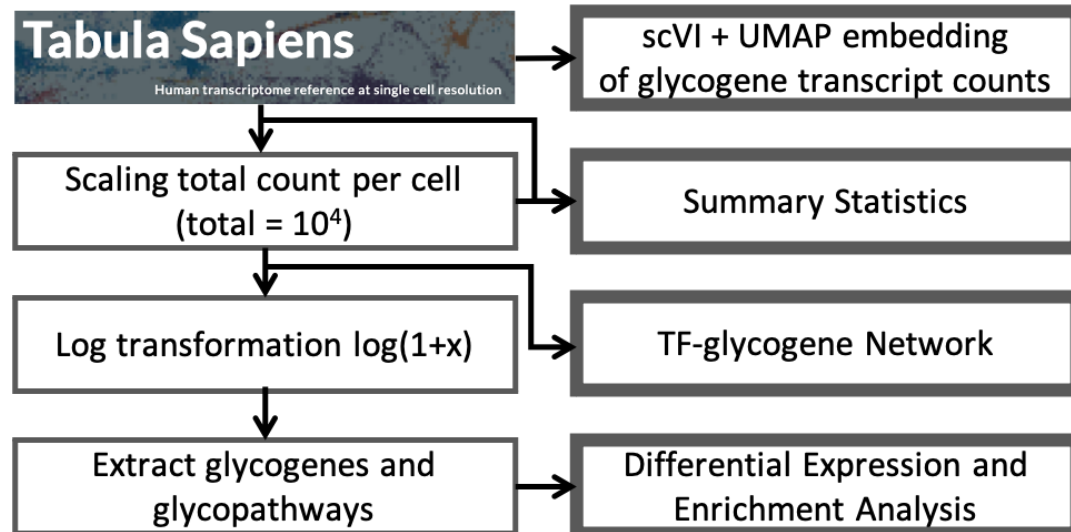
Single-cell Analysis of Glycosylation

Tabula Sapiens Single-cell RNA-seq

- ~500K cells, 400 cell types
- 24 organs, 15 normal subjects
 - Sex: 9 male/6 female
 - Age: 22-74y
 - 6-White/6-Hispanic/2-Black/ 1-Asian



DecontX counts, 10X



- Establish baseline glycoprotein single-cell expressions in human
- Establish data processing pipeline for ML/DL modeling

How does the expression of glycoproteins vary with cell/tissue types?

How prevalent are the glycoproteins?

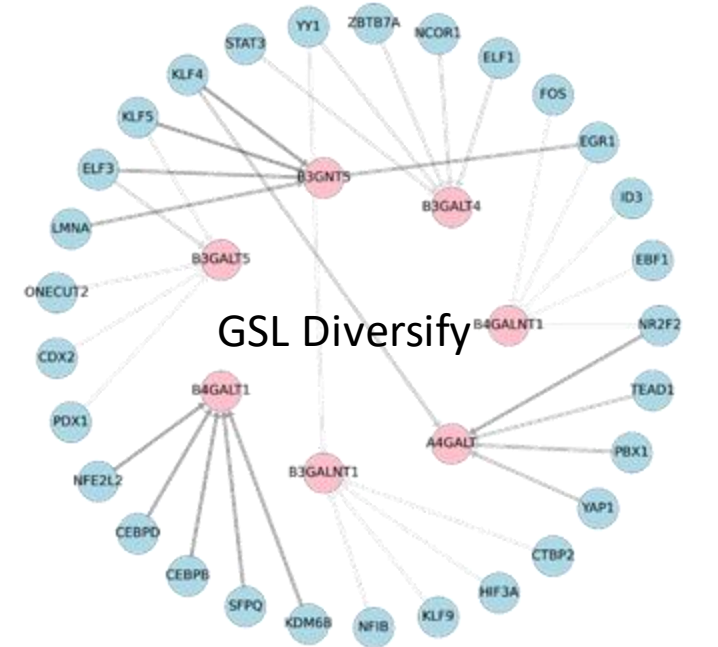
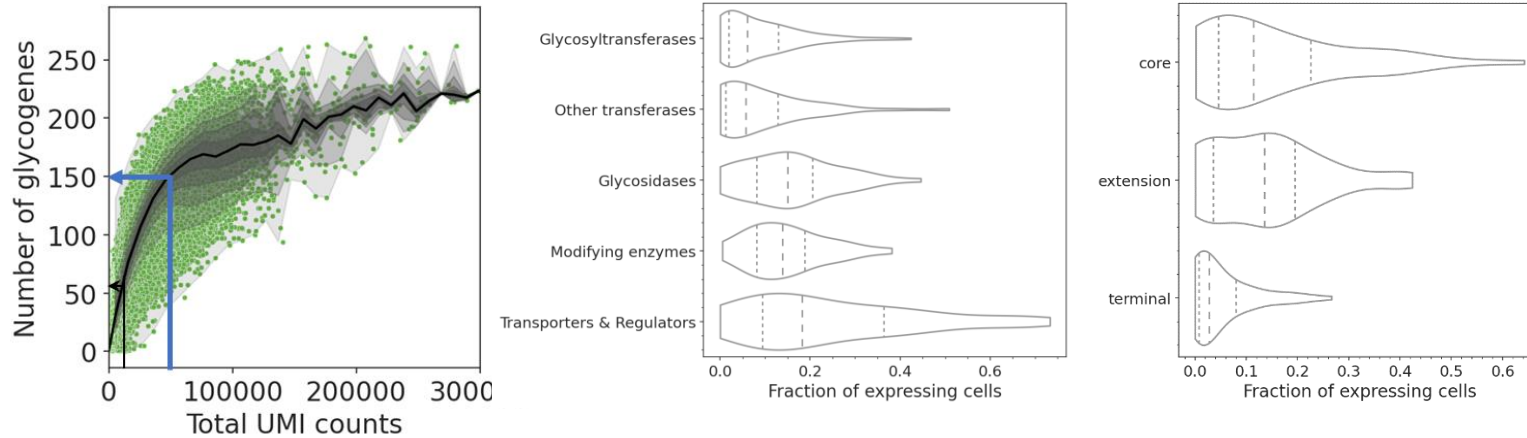
Computational prediction of TFs of Glycosylation?

Glycosylation pathway variation across cell type and tissues ?

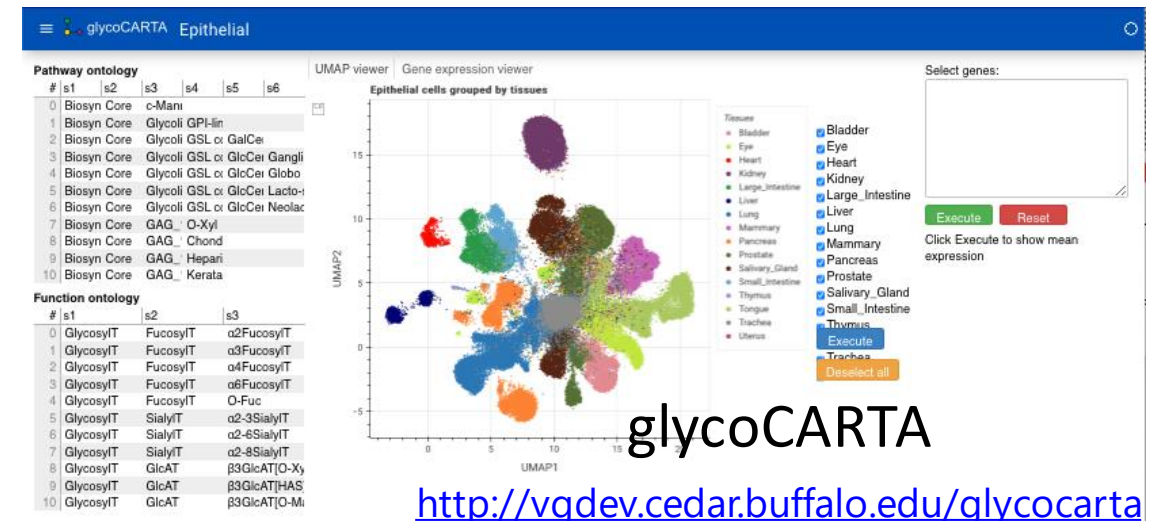
Single-cell Analysis of Glycosylation

GlycoTF

<http://vgdev.cedar.buffalo.edu/glycotf>



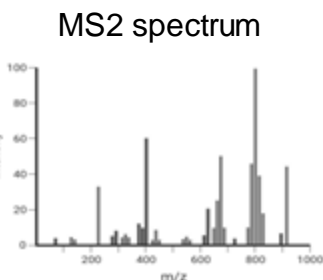
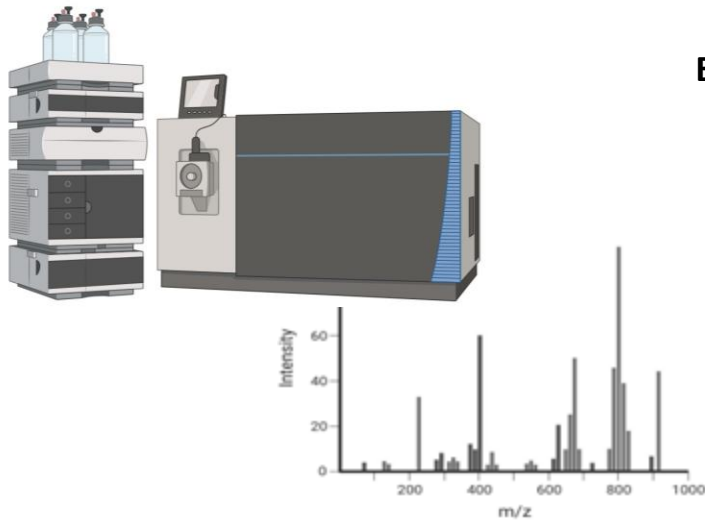
- Glycogenes are as commonly expressed as other protein coding genes.
- At 50K-70K reads/cell, on average ~60 glycogenes are detected (a max. of ~220 genes).
- Core pathways are expressed at higher levels than extension and terminal pathways.



<http://vgdev.cedar.buffalo.edu/glycocarta>

Large Language Model for Mass Spectrometry

Bidirectional Encoder Representations from Transformers (BERT)



{475.1: 1820.0, 576.7: 7323.0, 577.2: 3710.0, 577.7: 1254.0, ...}

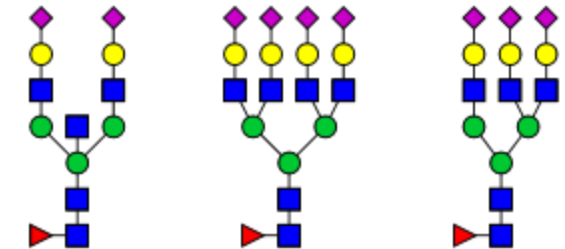
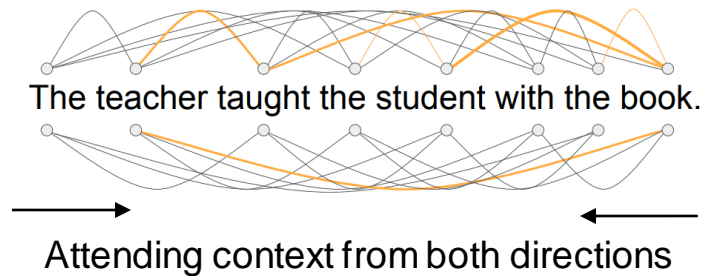
peak m/z : peak intensity

Convert m/z peaks into MS sentence

PGC negative ESI reduced linear
 CID N rt84 mz2368 mz1856
 mz1858 mz1517 mz1859 ...

Sentence = *expt. metadata* + *retention time* + *precursor m/z* + *MS2 m/z peaks*

Tokenize MS sentence



GlycoBERT

Sequence Classification

Inference: Neu5Ac(a2-3)Gal(b1-3)[Neu5Ac(a2-6)]GalNAc
 Ground truth: Neu5Ac(a2-3)Gal(b1-3)[Neu5Ac(a2-6)]GalNAc

Inference: Man(a1-3)[Man(a1-6)]Man(b1-4)GlcNAc(b1-4)GlcNAc
 Ground truth: Man(a1-3)[Man(a1-6)]Man(b1-4)GlcNAc(b1-4)GlcNAc

Inference: Neu5Gc(a2-3)Gal(b1-4)Glc
 Ground truth: Neu5Gc(a2-3)Gal(b1-4)Glc

Inference: HexNAc(.1-3)[Neu5Ac(a2-6)]GalNAc
 Ground truth: HexNAc(.1-3)[Neu5Ac(a2-6)]GalNAc

GlycoBART

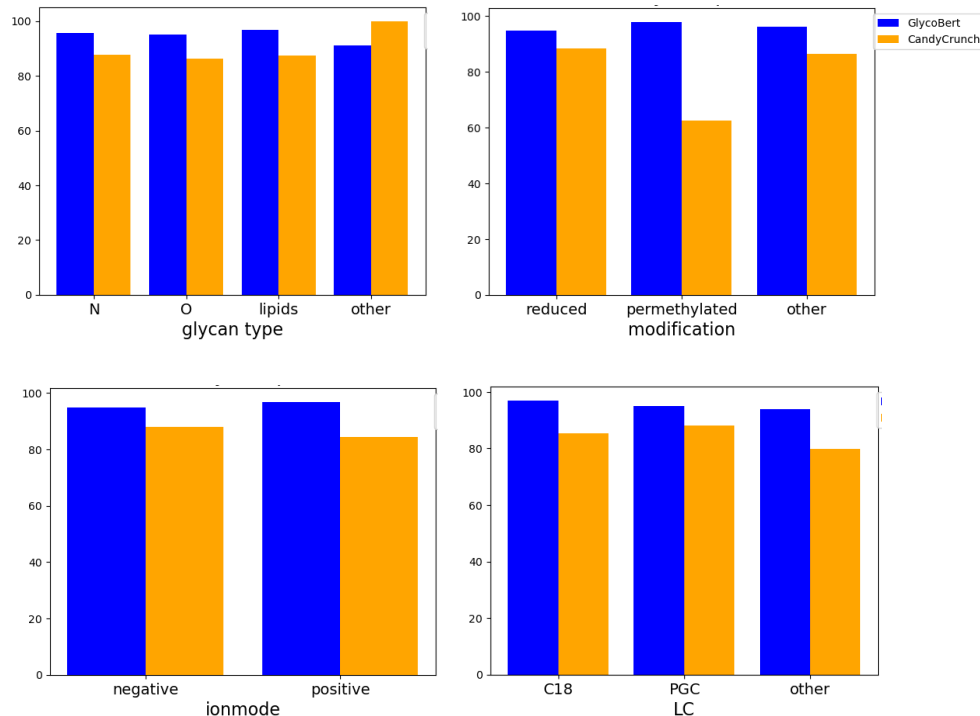
Sequence Generation

Inference: Gal(.1-.)GalNAc(a1-3)[Neu5Gc(a2-6)]GalNAc
 Ground truth: Neu5Gc(a2-3)Gal(b1-3)[GlcNAc(b1-6)]GalNAc

Inference: Neu5Ac(a2-3)Gal(b1-4)GlcNAc(b1-6)[Fuc(a1-2)Gal(b1-3)]GalNAc
 Ground Truth: Fuc(a1-2)Gal(b1-4)GlcNAc(b1-6)[Neu5Ac(a2-3)Gal(b1-3)]GalNAc

Large Language Model for Mass Spectrometry

Level	Test Accuracy (%)	
	GlycoBERT	CandyCrunch
Mass	99.75	98.49
Composition	99.57	97.7
Topology	96.73	89.8
Structure	95.33	87.18



- Trained on MS2 data from glycomics (~480K spectra)
- Transformer-based LLM is a powerful architecture for analyzing MS data of glycomics profiling.
- GlycoBART is capable of generating *de novo* glycan structure prediction.
- Metadata (glycan type, experimental parameters) are highly informative.
- A promising framework for building foundational models of mass spectra

Challenges and future work

- Sparsity of glycosylation-specific data in literature:
 - Develop glycan specific tools, e.g. focused transcriptomics on glyco genes
 - Streamlined quantitative analysis of glycoproteins using MS
- Data: lack of labelled data and imbalanced dataset
 - Employ *in silico* data, self-supervised model
- Generative AI (glycoBART) can hallucinate.
 - Incorporate postprocessing of predictions
- Translation to better patient healthcare and treatment
 - Incorporate constraints / structures informed by biological knowledge