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

FY22 NOT-OD-22-065

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

University at Buffalo
The State University of New York

AI Supplement Program PI meeting
Breakout 3/Day 1
[2:45-3:45]

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.

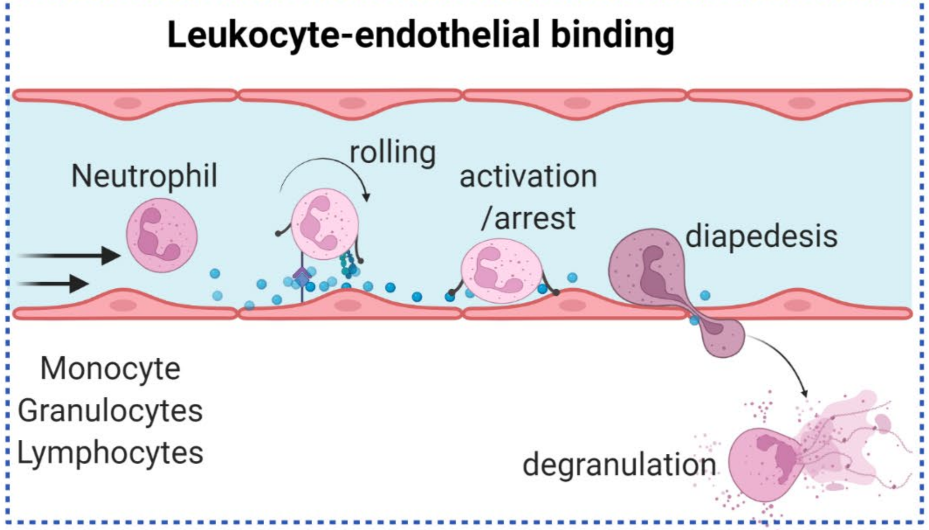
Aim 3. To demonstrate the use of the transformed data in a DL/ML application.

Highlights: Research outputs and shared data

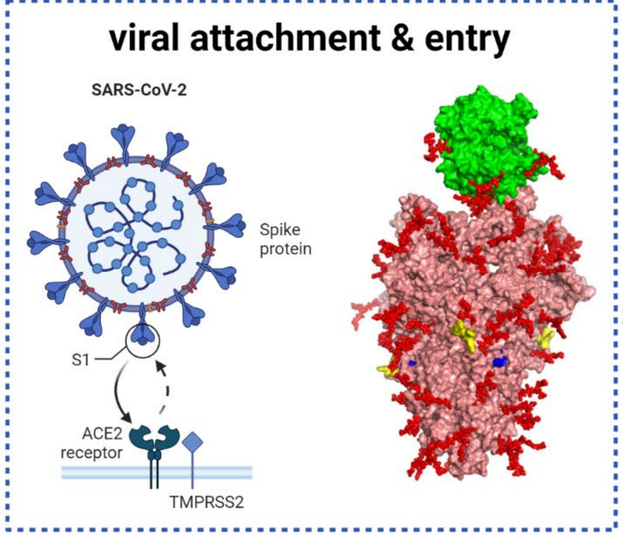
- Papers (other manuscripts are in preparation):
  a. P. Chrysinas, C. Chen, and R. Gunawan. CrossTx: Cross-cell-line transcriptomic signature predictions. *Processes*, 12:332, 2024.
  b. Cell and tissue-specific glycosylation pathways and transcriptional regulation informed by single-cell transcriptomics*. 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/

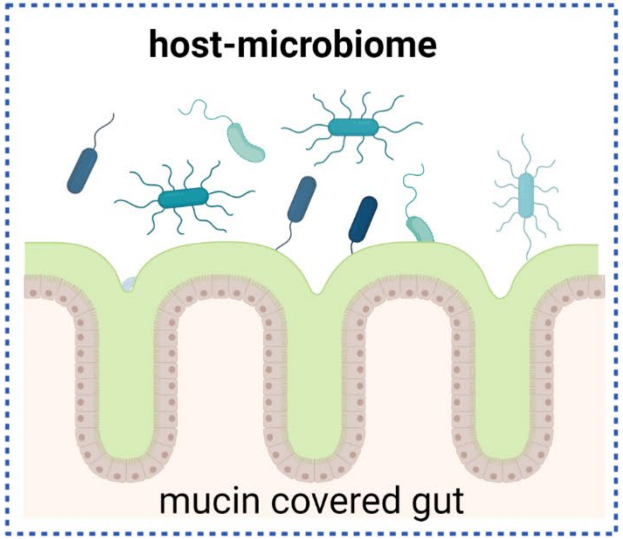
Ubiquitous in nature and relevant to biotechnology

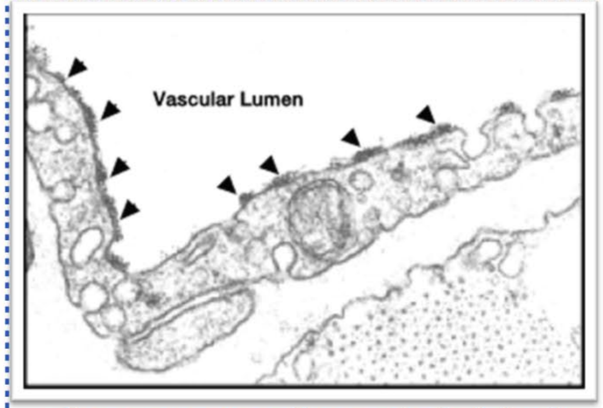
Role of selectins in leukocyte rolling



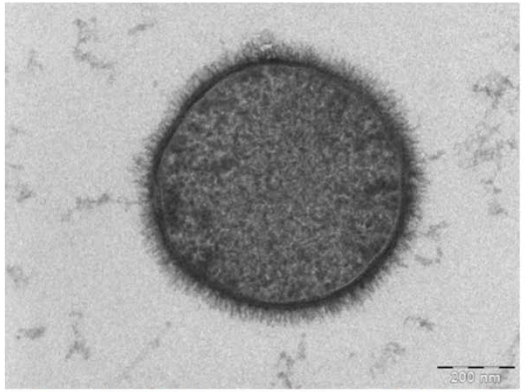
O- and N-glycans regulating SARS-CoV-2



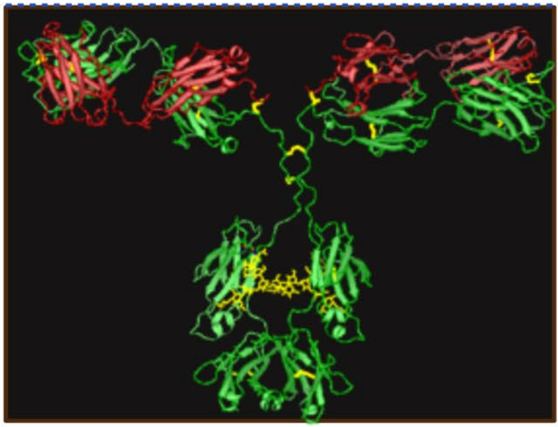




Glycans on rat capillary endothelial cells (Essentials of Glycobiology; George Palade)



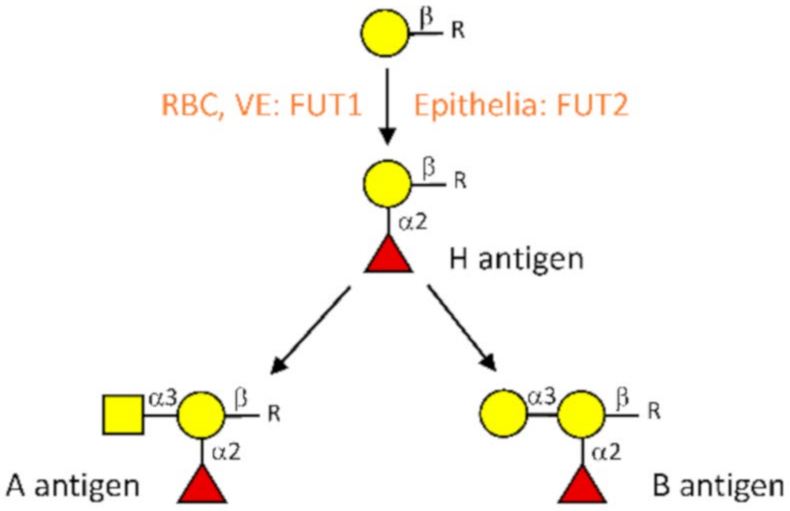
B. subtilis bacterium, with hair-like glycocalyx (Wikipedia)



biotechnology

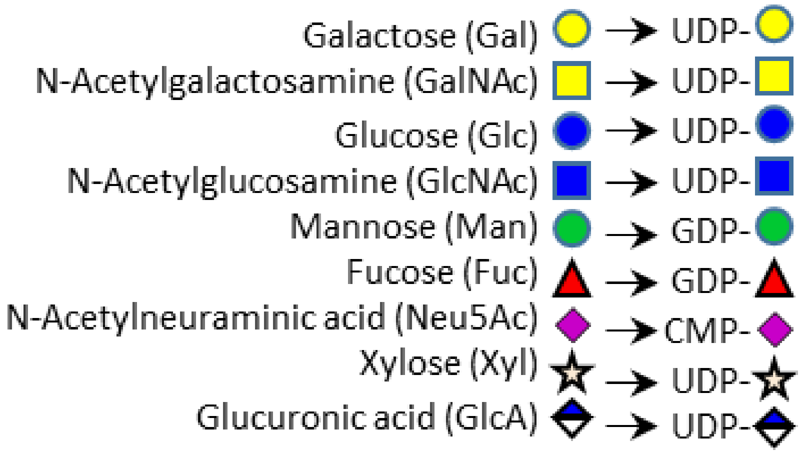
Most common glycans are blood group antigen:

Expressed on RBCs and critical for transfusion medicine



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

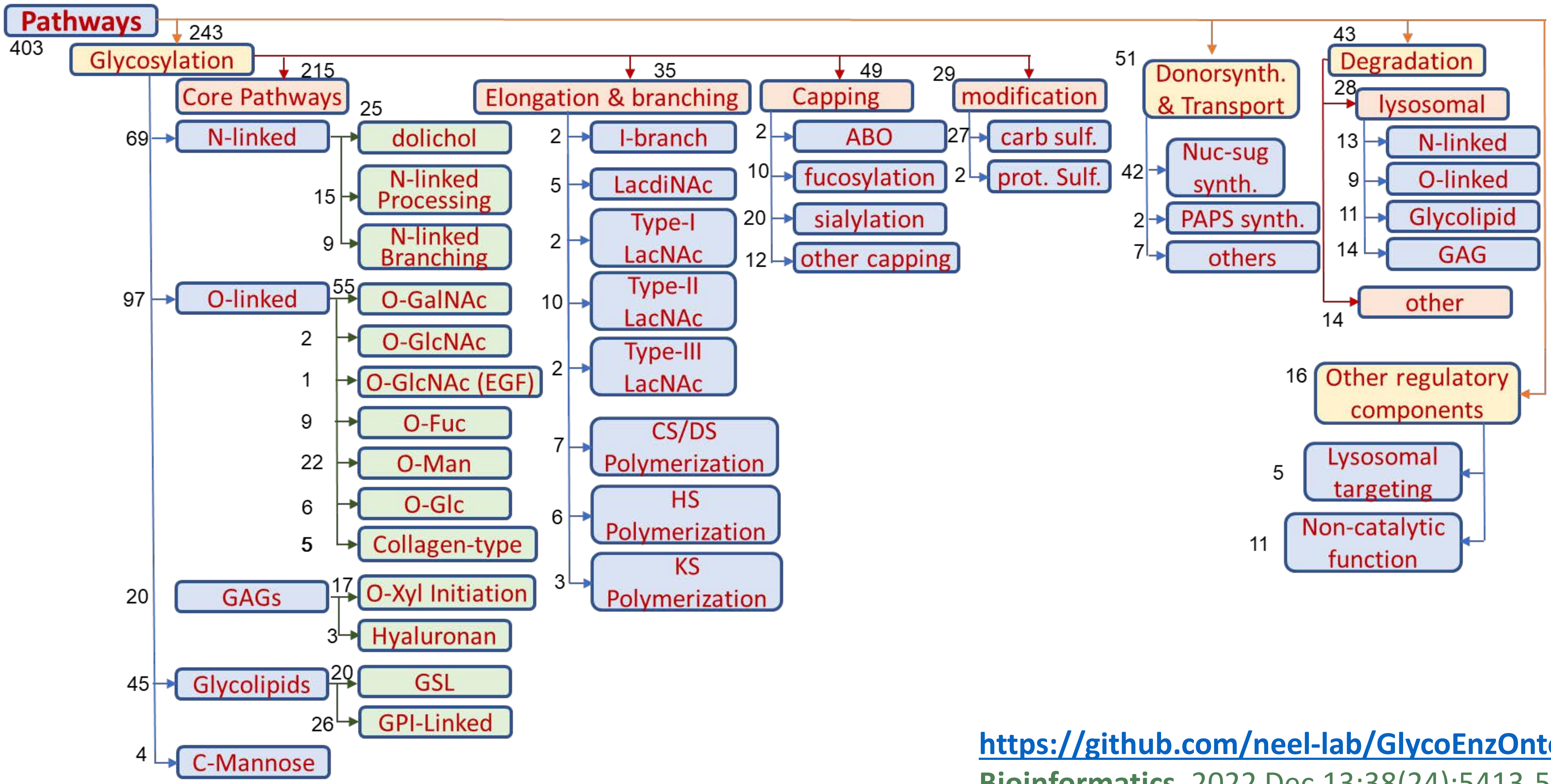
9 human monosaccharides



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

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

GlycoEnzOnto: An ontology for human glycosylating enzymes

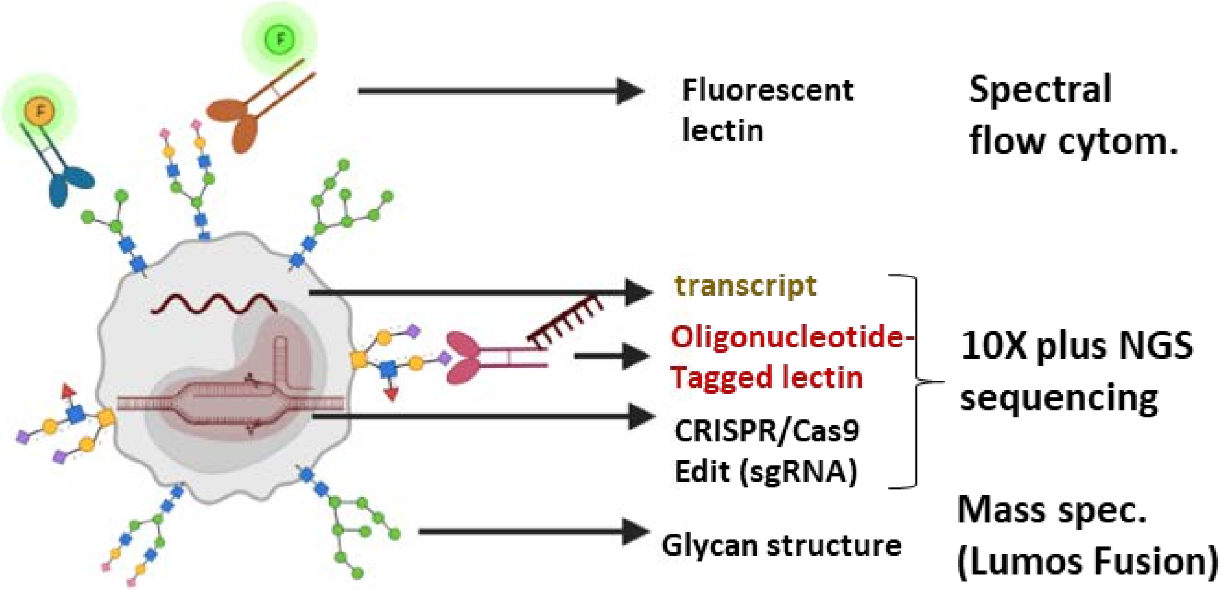


https://github.com/neel-lab/GlycoEnzOnto

Bioinformatics. 2022 Dec 13;38(24):5413-5420

Multimodal measurements and their data integration

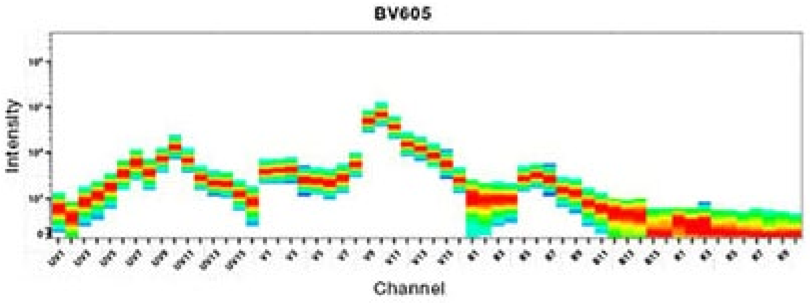
Single-cell multiOMICs



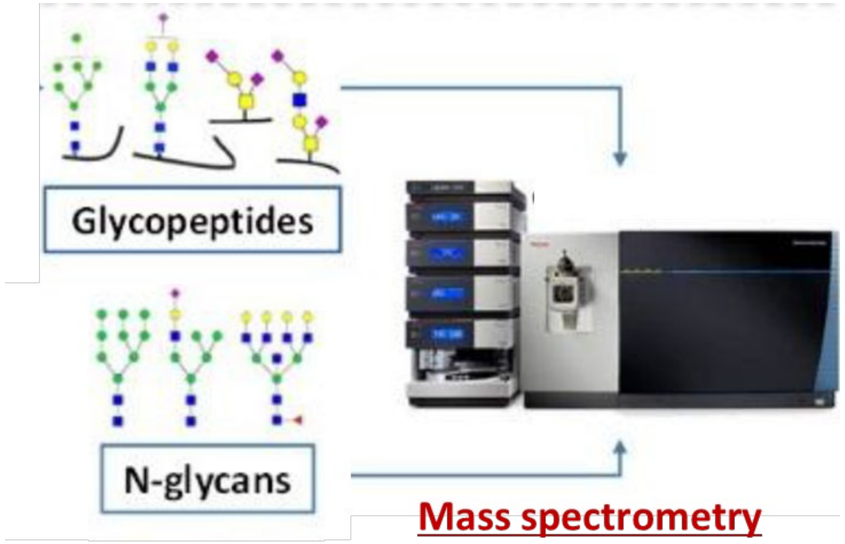
Fluorescent lectin → Spectral flow cytom.

transcript
Oligonucleotide-Tagged lectin
CRISPR/Cas9 Edit (sgRNA)
→ 10X plus NGS sequencing

Glycan structure → Mass spec. (Lumos Fusion)



Spectral flow cytometry

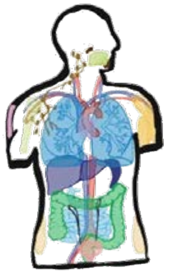


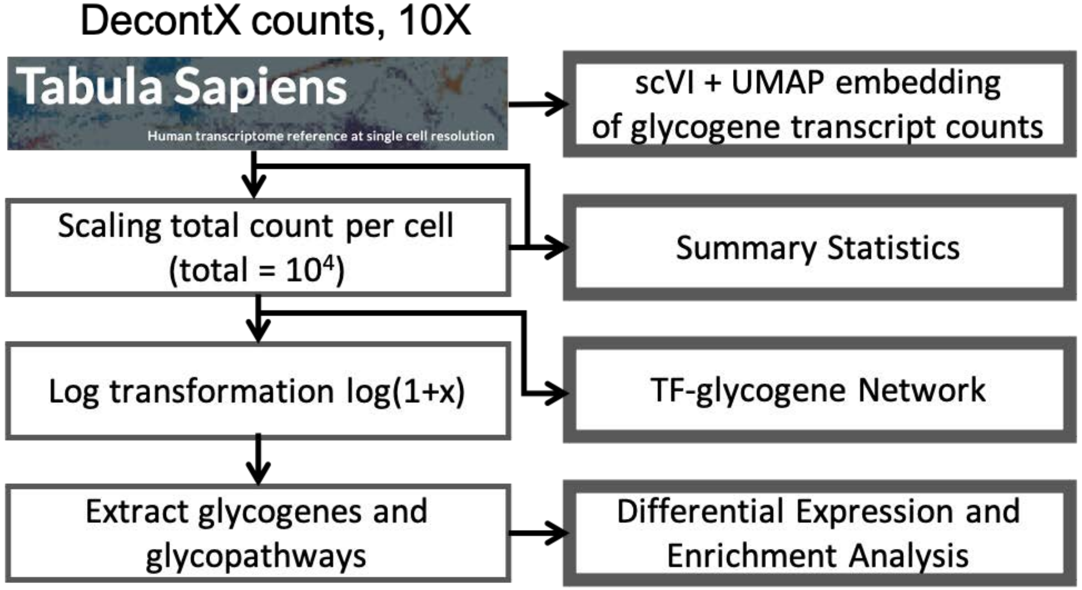
Mass spectrometry

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

Tabula Sapiens

Human transcriptome reference at single cell resolution

scVI + UMAP embedding of glycogene transcript counts

Scaling total count per cell (total = 10^4)

Summary Statistics

Log transformation log(1+x)

TF-glycogene Network

Extract glycogenes and glycopathways

Differential Expression and Enrichment Analysis

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

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

How prevalent are the glycogenes?

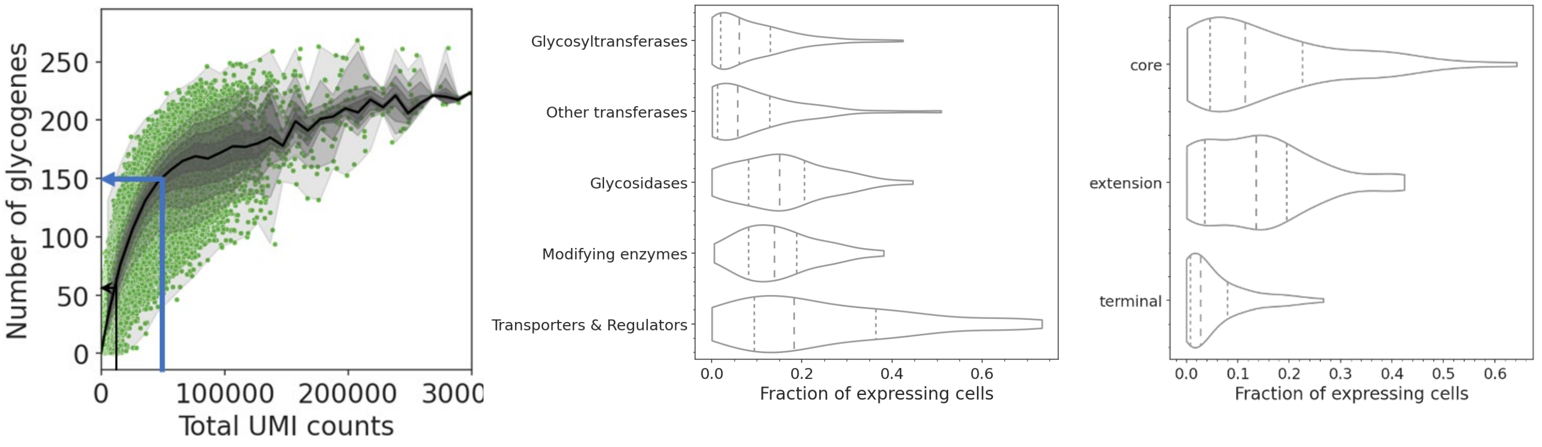
Computational prediction of TFs of Glycosylation?

Glycosylation pathway variation across cell type and tissues ?

https://tabula-sapiens-portal.ds.czbiohub.org/

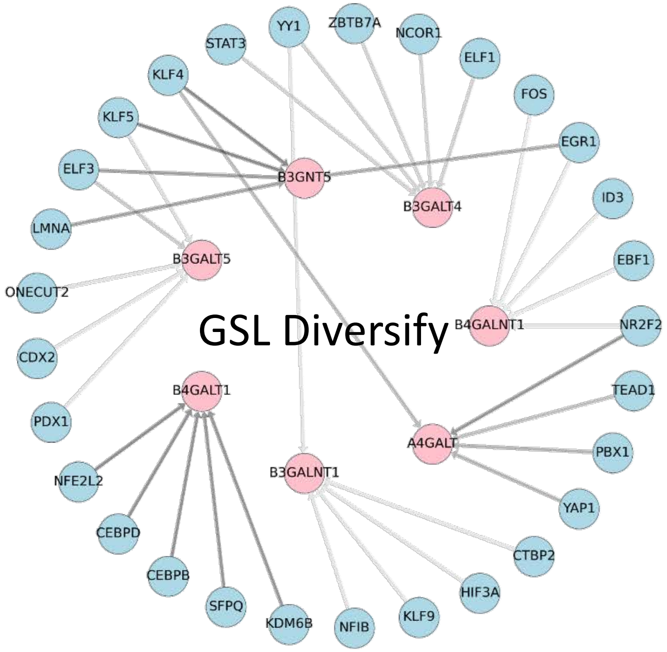
Science 376(6594):eabl4896 2022.

Single-cell Analysis of Glycosylation

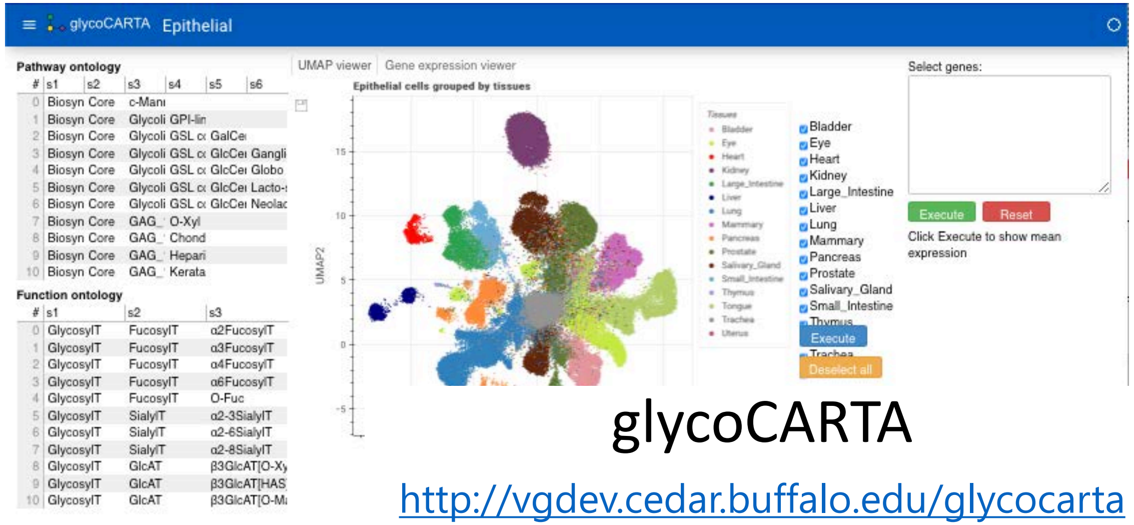


GlycoTF

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



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



Pathway ontology

#	s1	s2	s3	s4	s5	s6
0	Biosyn	Core	c-Mani			
1	Biosyn	Core	Glycoli	GPI-lin		
2	Biosyn	Core	Glycoli	GSL α	GalCe	
3	Biosyn	Core	Glycoli	GSL α	GlcCe	Gangli
4	Biosyn	Core	Glycoli	GSL α	GlcCe	Globo
5	Biosyn	Core	Glycoli	GSL α	GlcCe	Lacto-:
6	Biosyn	Core	Glycoli	GSL α	GlcCe	Neolac
7	Biosyn	Core	GAG_'	O-Xyl		
8	Biosyn	Core	GAG_'	Chond		
9	Biosyn	Core	GAG_'	Hepari		
10	Biosyn	Core	GAG_'	Kerata		

Function ontology

#	s1	s2	s3
0	GlycosylT	FucosylT	α2FucosylT
1	GlycosylT	FucosylT	α3FucosylT
2	GlycosylT	FucosylT	α4FucosylT
3	GlycosylT	FucosylT	α6FucosylT
4	GlycosylT	FucosylT	O-Fuc
5	GlycosylT	SialylT	α2-3SialylT
6	GlycosylT	SialylT	α2-6SialylT
7	GlycosylT	SialylT	α2-8SialylT
8	GlycosylT	GlcAT	β3GlcAT[O-Xy
9	GlycosylT	GlcAT	β3GlcAT[HAS
10	GlycosylT	GlcAT	β3GlcAT[O-Ma

☑ Bladder
☑ Eye
☑ Heart
☑ Kidney
☑ Large_Intestine
☑ Liver
☑ Lung
☑ Mammary
☑ Pancreas
☑ Prostate
☑ Salivary_Gland
☑ Small_Intestine
☑ Thymus
☑ Trachea

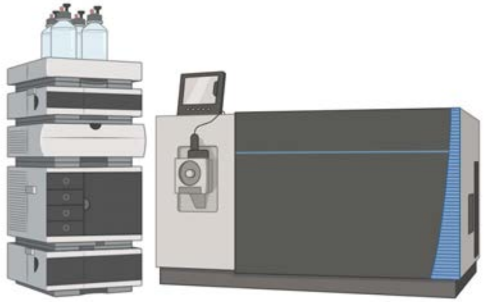
Select genes:

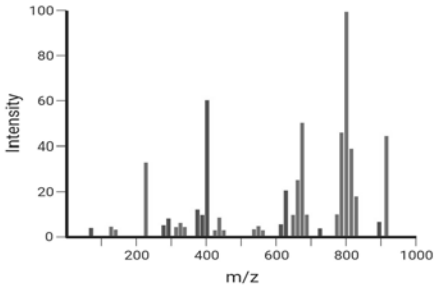
Click Execute to show mean expression

glycoCARTA

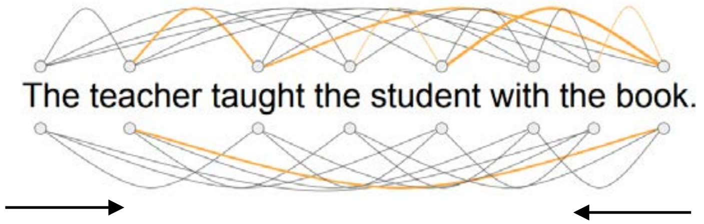
http://vgdev.cedar.buffalo.edu/glycocarta

Large Language Model for Mass Spectrometry

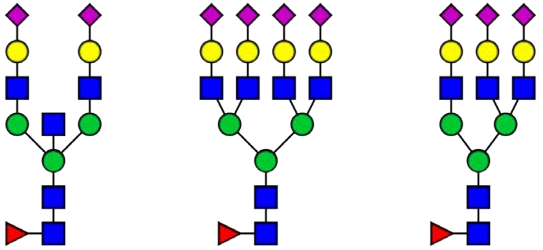




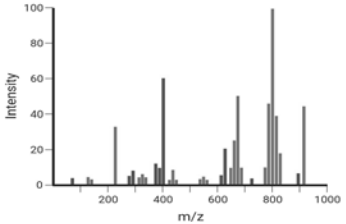
Bidirectional **E**ncoder **R**epresentations from **T**ransformers (**BERT**)



Attending context from both directions



MS2 spectrum



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

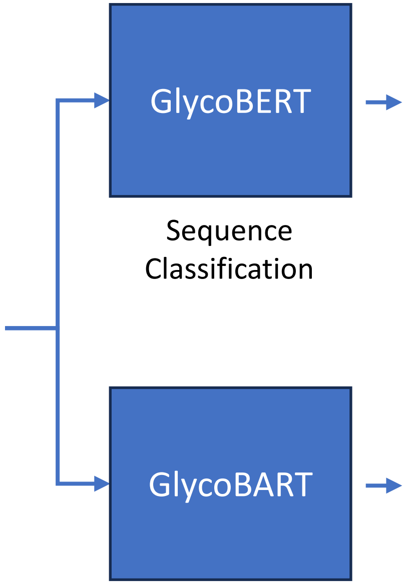
peak m/z : peak intensity

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



Sequence Classification

Sequence Generation

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

----

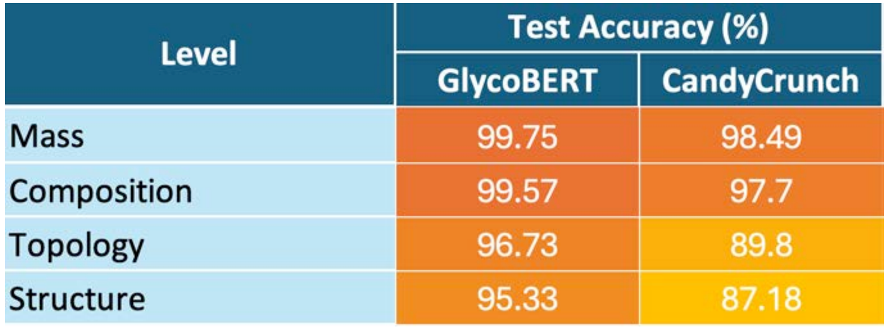
Inference : Neu5Gc(a2-3)Gal(b1-4)Glc
Ground truth: Neu5Gc(a2-3)Gal(b1-4)Glc

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

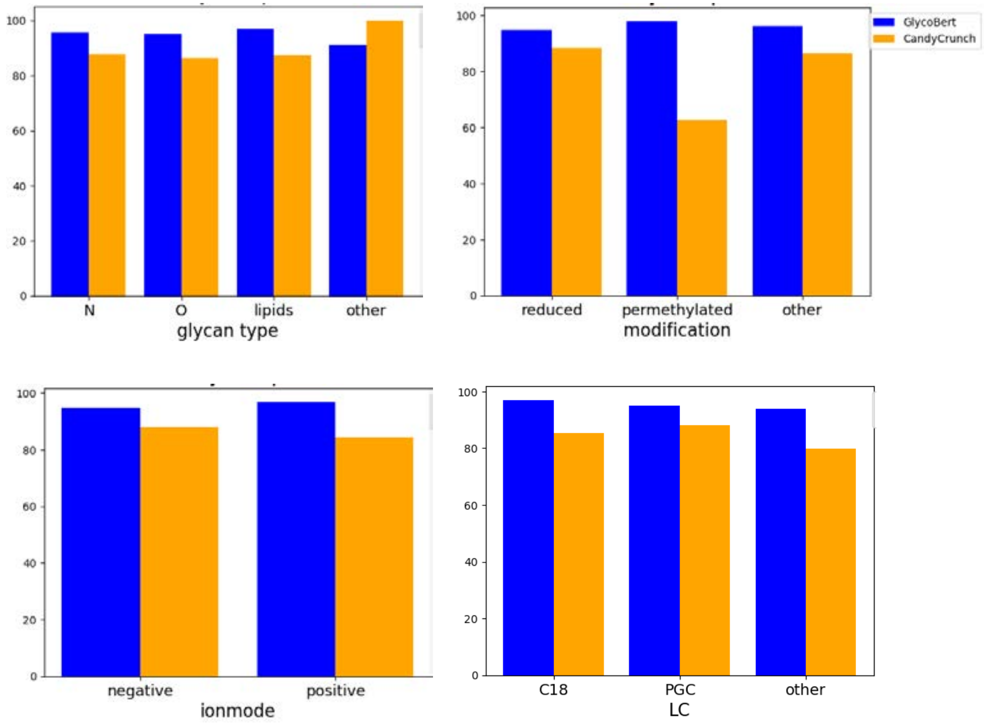
Inference: Gal(.1-.)GalNAc(a1-3)[Neu5Gc(a2-6)]GalNAc
Ground truth: Neu5Gc(a2-3)Gal(b1-3)[GlcNAc(b1-6)]GalNAc

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 glycogenes
  - 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