

Glycoinformatics tools to analyze and curate large scale experimental datasets

Sriram Neelamegham

**Departments of Chemical & Biological Engineering,
Biomedical Engineering and Medicine**

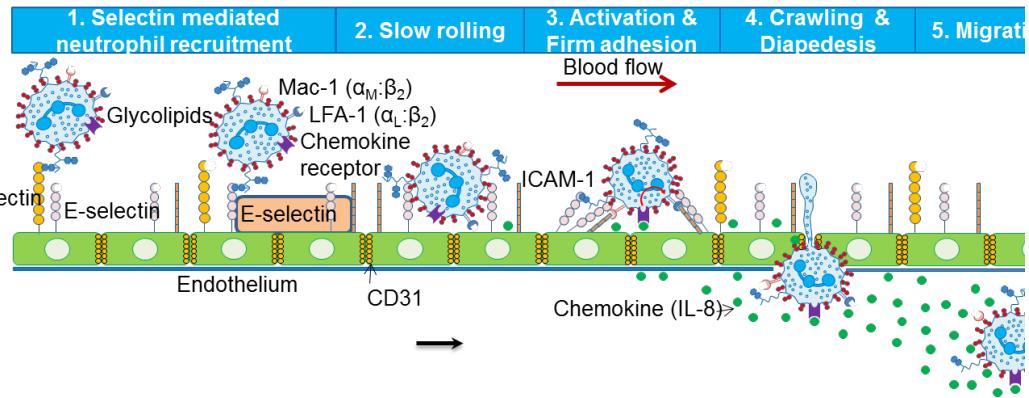
State University of New York, Buffalo, NY

9:10-9:50am March 6, 2018
Tokyo, Japan

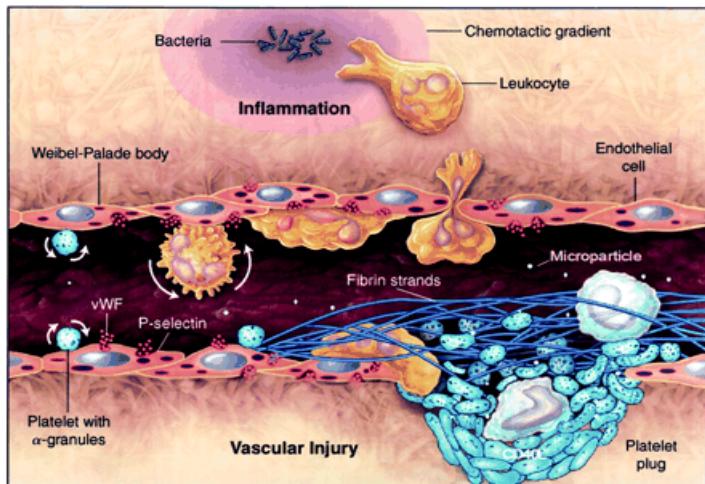


Overview of research interests

Inflammation

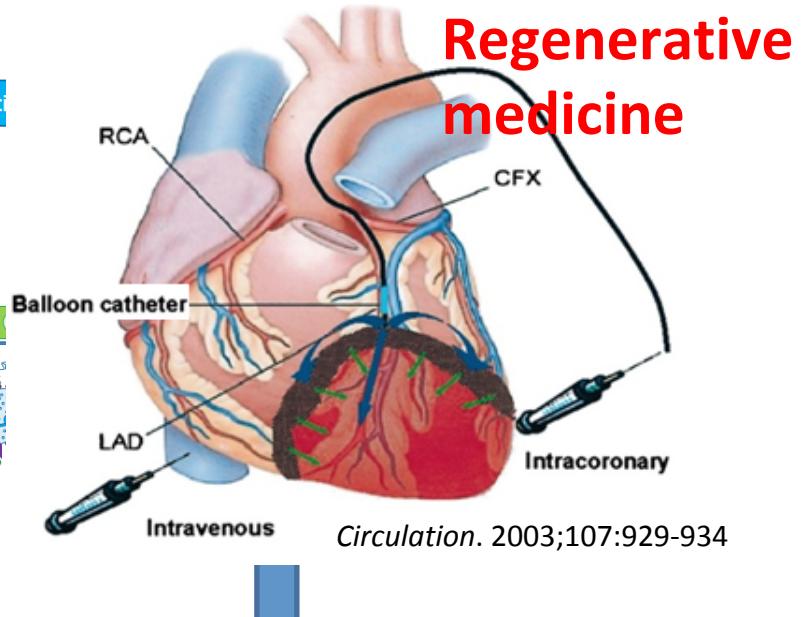


Thrombosis



ATVB, 25:1321, 2005.

Regenerative medicine



Systems Biology
Input-output
relationship

Input-output response

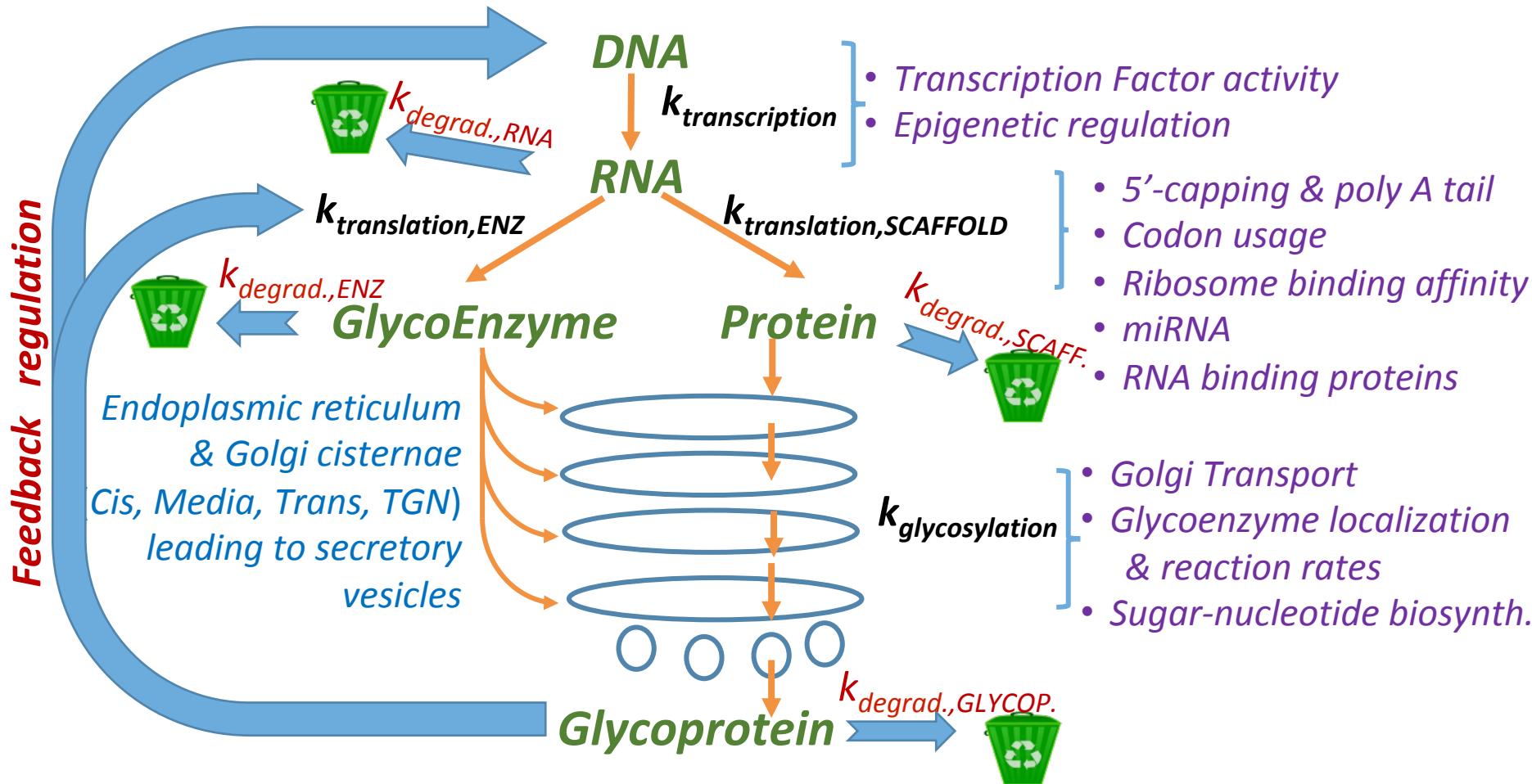
- Their generation

Wet-lab: Next Generation Seq., LC-MS with CRISPR-Cas9 perturbations

- Their visualization, analysis and simulation

Dry lab: LC-MS data analysis programs,
Pathway maps

VirtualGlycome.org: Systems level view of glycosylation



Open-source integration of knowledge across scales

- **GNAT-Web**: Glycosylation network analysis toolbox
- **DrawGlycan-SNFG**: Simple tool to convert IUPAC strings to SNFG sketches
- **GlycoPAT**: High-throughput analysis of LC-MSⁿ data, with focus on glycoProteomics

1. GNAT-Web

Glycosylation Network Analysis Toolbox

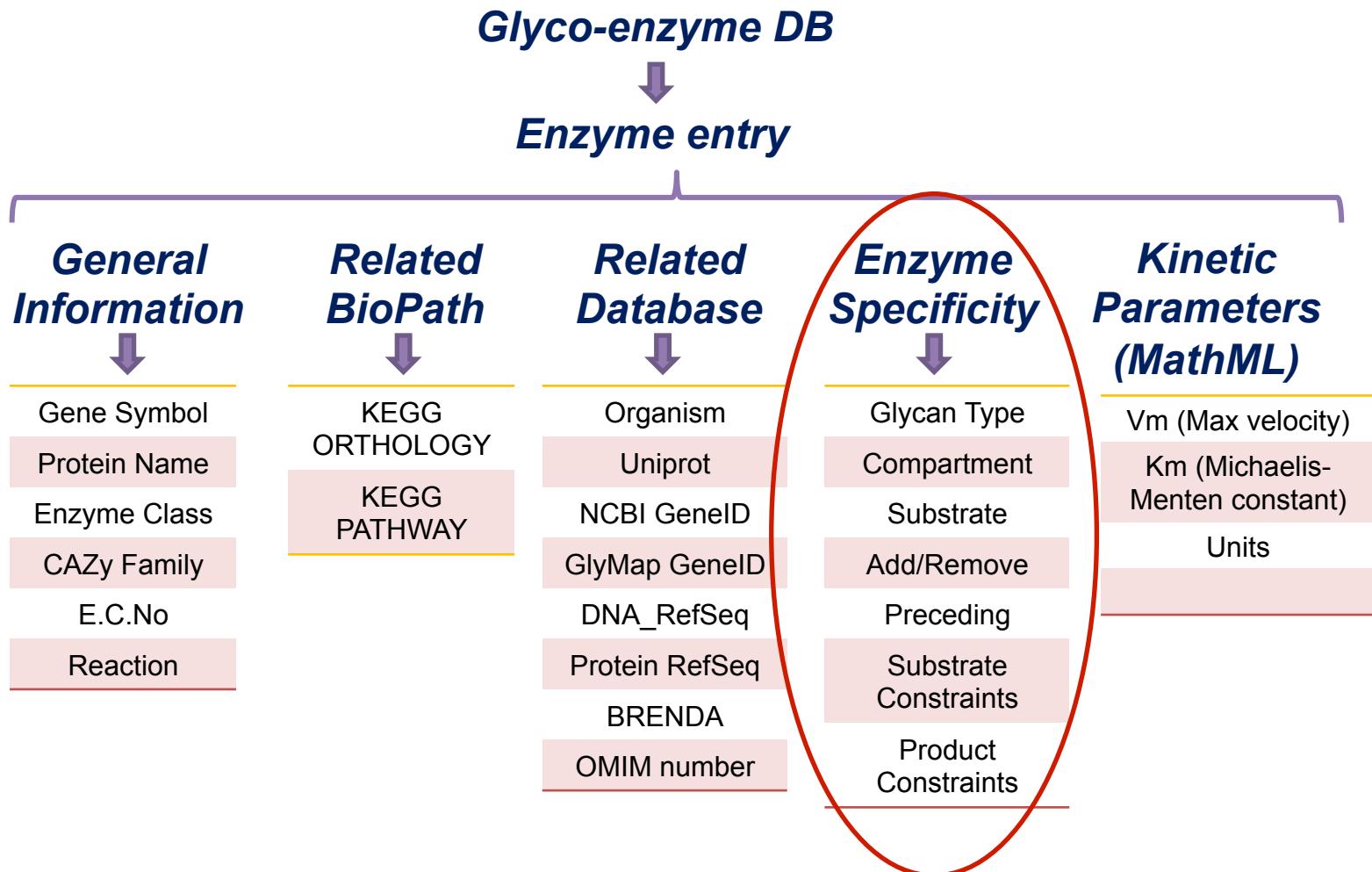


- Define glycoEnzymes *in silico*
- Develop reaction network from RNA-Seq and MS data processing
- Eventually, simulate reaction networks to bridge data across scales

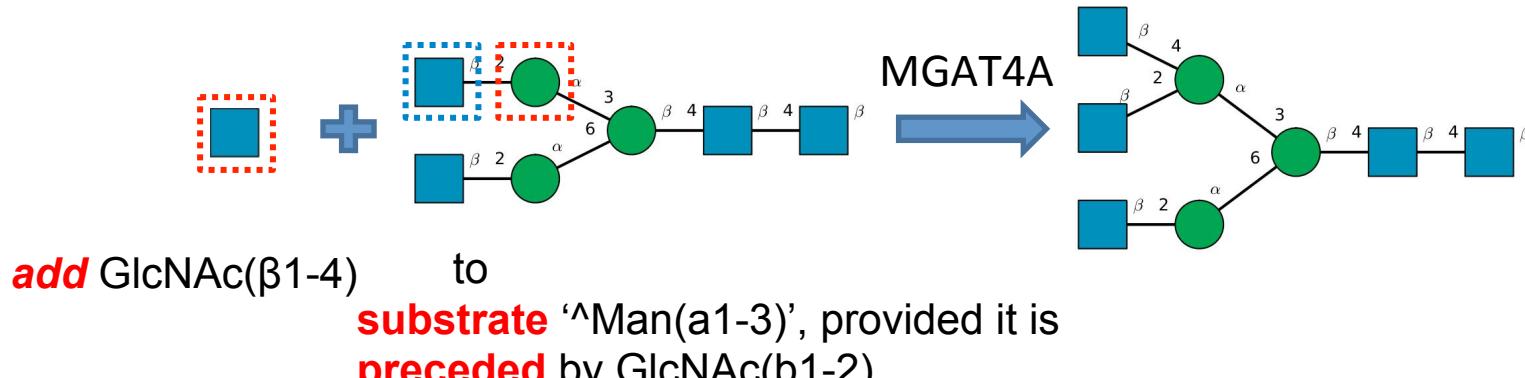
Yusen Zhou

Liu G, et al.
Bioinformatics. 24(23):2740-7, 2008;
Glycobiology. 21(12):1541-53, 2011;
Bioinformatics. 29(3):404-6, 2013;
PLoS One. 9(6):e100939, 2014.

XML based glycoenzyme definition



Enzyme specificity (e.g. MGAT 4)



$^{\wedge}$: Caret is space for inclusion
of arbitrary branches

Constraint: MGAT4 acts before addition of:

- Galactose (Gal), i.e. Gal cannot exist in string or **Gal#0**
- Bisecting MGAT3, i.e. **GlcNAc(b1-4) $^{\wedge}$ Man(b1-4)#0**

Maximum # = 0
indicates **NOT**;
*But it could be
any other
number as
well*

Enzyme rule	Value
Add	GlcNAc(b1-4)
Substrate	$^{\wedge}$ Man(a1-3)
References:	

a. Bennun et al., PLoS Comput Biol. 9(1):e1002813, 2013;

SubstConstraints	Value
MaxSubsubst	Gal#0&GlcNAc(b1-4) $^{\wedge}$ Man(b1-4)#0

Enzyme specificity (e.g. MGAT 4)

```
- <GeneralInfo>
  <GeneSymbol>MGAT4A</GeneSymbol>
  <ProteinName>Alpha-1,3-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase A</ProteinName>
  <EnzymeClass>Glycosyltransferase</EnzymeClass>
  <GlycanType>N_linked</GlycanType>
  <CAZy>GT54</CAZy>
  <ECNo>2.4.1.145</ECNo>
  <TissuesName/>
  <OrgansName/>
</GeneralInfo>
```

^: C
of a
Conc:

```
  <Add>GlcNAc(b1-4)</Add>
  <Remove/>
  <Substrate>^Man(a1-3)</Substrate>
  <Preceding>GlcNAc(b1-2)</Preceding>
```

```
- <BioPath>
  <KO>K00738</KO>
  <Pathway>ko00510/ko00513</Pathway>
</BioPath>
```

```
  </SubstConstraints>
  - <ProdConstraints>
    <MaxProd/>
    <MinProd/>
    <MaxProdsubst/>
    <MinProdsubst/>
  </ProdConstraints>
</EnzSpecificity>
```

(b1-4)^

n # = 0
NOT;
ld be

```
  - <EnzKinetics>
    <Vm>1e-5</Vm>
    <Km>3.4e9</Km>
    <Units>pM</Units>
  - <EnzDistribution>
    <ER/>
    <Cis>0.15</Cis>
    <Medial>0.45</Medial>
    <Trans>0.3</Trans>
    <TGN>0.1</TGN>
  </EnzDistribution>
</EnzKinetics>
```

MaxSubsubst

Gal#0&GlcNAc(b1-4)-Man(b1-4)#0

Integration into enzyme database

XML data structure for glycoenzyme

 <?xml version="1.0" encoding="UTF-8"?>
<EnzymeDB>
 {<Enzyme EnzClass="...">
 <GeneralInfo>...</GeneralInfo>
 <BioPath>...</BioPath>
 <RelatedDB>...<RelatedDB>
 <EnzSpecificity>
 ...
 </EnzSpecificity>
 <EnzKinetics>...</EnzKinetics>
 </Enzyme>
 <Enzyme EnzClass = "...">
 ...
 </Enzyme>
 ...
</EnzymeDB>

Example

```
<?xml version="1.0" encoding="UTF-8"?>
- <GlycoEnzyme>
  - <GeneralInfo>
    <GeneSymbol>MGAT1</GeneSymbol>
    <ProteinName>Alpha-1,3-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase</ProteinName>
    <EnzymeClass>Glycosyltransferase</EnzymeClass>
    <GlycanType>N_linked</GlycanType>
    <CAZy>GT13</CAZy>
    <ECNo>2.4.1.101</ECNo>
    <TissuesName/>
    <OrgansName/>
  </GeneralInfo>
  - <BioPath>
    <KO>K00726</KO>
    <Pathway>ko00510/ko00513</Pathway>
  </BioPath>
  - <RelatedDB>
    - <Organism type="Homo sapiens">
      <Uniprot>P26572</Uniprot>
      <GeneID>4245</GeneID>
      <DNA_RefSeq>NM_001114617.1</DNA_RefSeq>
      <Protein_RefSeq>NP_001108089.1</Protein_RefSeq>
      <OMIM_number>160995</OMIM_number>
    </Organism>
  </RelatedDB>
  - <EnzSpecificity>
    <Add>GlcNAc(b1-2)</Add>
    <Remove/>
    <Substrate>Man(a1-3)^Man(b1-4)</Substrate>
    <Preceding/>
    - <SubstConstraints>
      <MaxSubst>Man(a1-3)[Man(a1-6)[Man(a1-3)]Man(a1-6)]Man(b1-4)GlcNAc(b1-4)GlcNAc(b1-2)</MaxSubst>
      <MinSubst>Man(a1-3)[Man(a1-6)[Man(a1-3)]Man(a1-6)]Man(b1-4)GlcNAc(b1-4)GlcNAc(b1-2)</MinSubst>
      <MaxSubsubst/>
      <MinSubsubst/>
    </SubstConstraints>
    - <ProdConstraints>
      <MaxProd/>
      <MinProd/>
      <MaxProdsubst/>
      <MinProdsubst/>
    </ProdConstraints>
  </EnzSpecificity>
  - <EnzKinetics>
    <Vm>4.5e-4</Vm>
    <Km>2.6e8</Km>
    <Units>pM</Units>
    - <EnzDistribution>
      <ER/>
      <Cis>0.15</Cis>
      <Medial>0.45</Medial>
      <Trans>0.3</Trans>
      <TCM>0.1</TCM>
    </EnzDistribution>
  </EnzKinetics>

```

Custom database generation

Obtain from existing databases

Enzyme specificity & kinetics

GNAT-WEB
Glycosylation Networks Analytic Toolbox

Custom database generation Database Visualization Create glyco-pathway

Database Edit: **Homo Sapiens** Export enzyme table Input Database Name... Load database Go

FUT1	FUT2	FUT3	FUT4	FUT5	FUT6
FUT7	FUT8	FUT9	FUT10	FUT11	ST3GAL2
ST3GAL1	ST3GAL3	ST3GAL4	ST3GAL5	ST3GAL6	ST6GAL1
ST6GAL2	ST6GALNAC1	ST6GALNAC2	ST6GALNAC3	ST6GALNAC4	ST6GALNAC5
ST6GALNAC6	ST8SIA1	ST8SIA2	ST8SIA3	ST8SIA4	
ST8SIA5	B4GALT1	B4GALT2	B4GALT3	B4GALT4	B4GALT5
B4GALT6	B4GALT7	B3GALT1	B3GALT2	B3GALT4	B3GALT5
B3GALT6	UGT8	A4GALT	C1GALT1	B3GALNT1	B3GALNT2
B4GALNT1	B4GALNT2	B4GALNT3	B4GALNT4	GBGT1	ABO(A)

General Info:

Uniprot #: Add

Tissues: Organs:

Glycan class: N-glycan O-glycan Glycosphingolipid

Compartment: ER Cis-Golgi Medial-Golgi Trans-Golgi TGN (Trans Golgi Network)

Enzyme Specificity:

Add or Remove
to/from
preceding

	Substrate Constraint	Product Constraint
Max Structure	<input type="text"/>	<input type="text"/>
Min Structure	<input type="text"/>	<input type="text"/>
Sub-struct, max#	<input type="text"/>	<input type="text"/>
Sub-struct, min#	<input type="text"/>	<input type="text"/>

Simulation Parameters:

Michaelis-Menten parameters

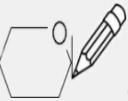
V_m pM/cell/hr K_m pM

Relative concentration of enzymes in compartment (sum will be normalized to 1).

ER	Cis-	Medial-	Trans-	TGN
<input type="text"/>				

Comments:

View database elements

 **GNAT-WEB**
Glycosylation Networks Analysis Toolbox

Custom database generation **View database elements** Create glyco-pathway

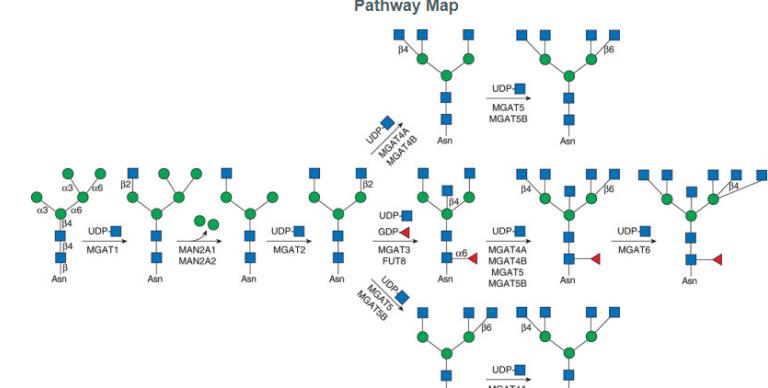
Customized Database: GeneDB

Glyco-Enzyme List

Gene Symbol: MGAT1
Enzyme Family: N_linked

Pathway: all
N_linked
Dolichol Pathway
Branching Pathway
Complex Pathway
Type-1/2 LacNAc
Blood Group i&i
ABO Blood Group
Sd^aAntigen & GM2
 α 2-3 and α 2-6 NeuAc
 α 2-8 on N_glycan
Gal01-3Gal Antigen
O_linked
Core1/2 Pathway
Core3/4 Pathway

Pathway Map



MGAT1

General Information

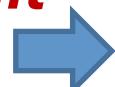
Gene Symbol: MGAT1
Protein name: Alpha-1,3-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase
Enzyme Class: Glycosyltransferase
CAZy Family: GT13
E.C. No: 2.4.1.101
Tissues:
Organs:

n: - + ↔

Ref: 3rd Ed. of Essentials of Glycobiology

Create pathways : forward

Specify starting material and enzymes



GNAT-WEB
Glycosylation Networks Analysis Toolbox

Database Generation Database Visualization Create glyco-pathway

View examples

Enzyme List:
N_linked ▾

Enzyme

MAN1A1	MAN1B1
MAN2A1	MGAT1
MGAT2	MGAT3
MGAT4A	MGAT5
B4GALT1	B3GNT2

FUT8
ST6GAL1
B4GALT1
B4GALT6
B4GALNT3
MGAT2
MGAT4C

FUT10
ST6GAL2
B4GALT2
B3GALT1
B4GALNT4
MGAT3
B3GNT2

Input Glycans: Count pM + -

Compartments: Group in bracket: Choose mechanism: Termination Step: Maximum M: # of sub-structures: Final Glycan: Simulation Parameters: Run

Constraints to limit network size

Create pathways :reverse

GNAT-WEB
Glycosylation Networks Analysis Toolbox

Database Generation Database Visualization Create glyco-pathway

[View examples](#)

Enzyme List:

N_linked ▾

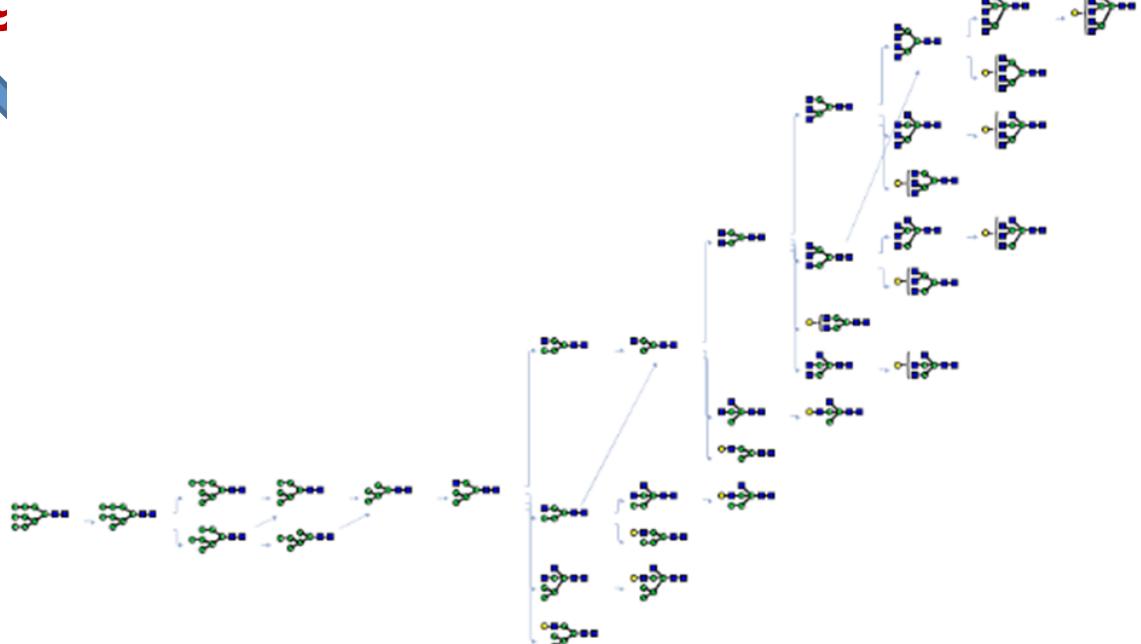
- FUT1
- ST3GAL3
- ST8SIA2
- B4GALT3
- B3GALT2
- ABO(A)
- MGAT4A

Enzyme

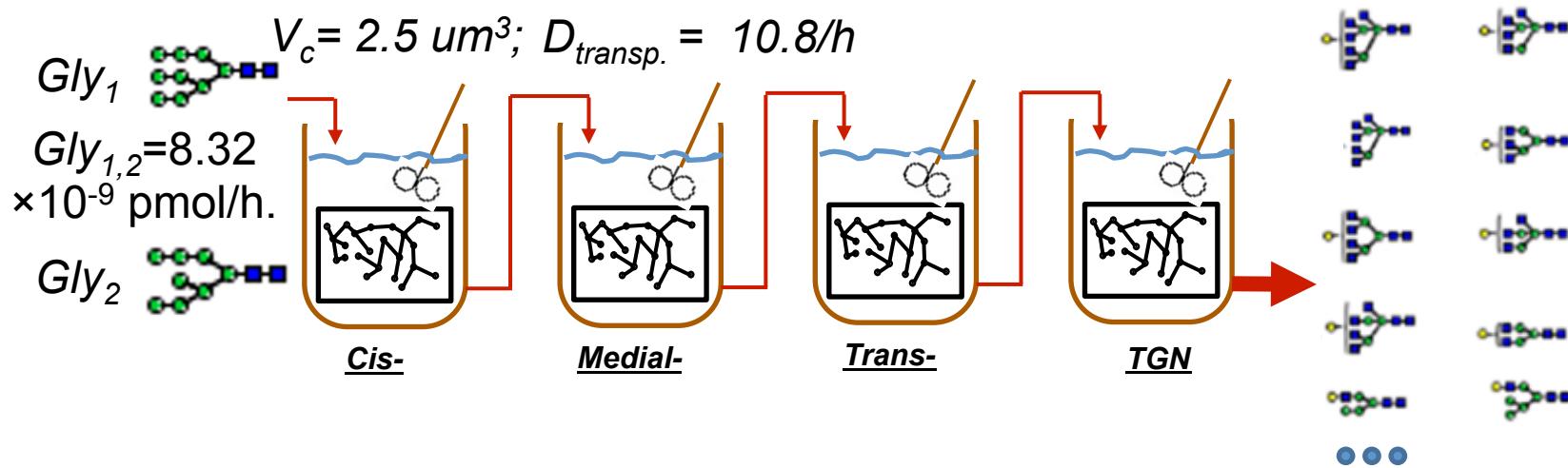
MAN1A1 MAN1B1
MAN2A1 MGAT1
MGAT2 MGAT3
MGAT4A MGAT5
B4GALT1

- FUT10
- ST6GAL2
- B4GALT2
- B3GALT1
- B4GALNT4
- MGAT3
- B3GNT2

Specify ending glycans & enzymes



Pathway: 4 compartment CSTR



Species balance equation:

$$\frac{d[Gly \downarrow i,j]}{dt} = D \downarrow transp. \times [Gly \downarrow i,j-1] - V \downarrow i,j \times [Gly \downarrow i,j] / Vc \times Km \downarrow i,j \\ \times (1 + \sum k \uparrow [Gly \downarrow k,j] / (Km \downarrow k,j)) - D \downarrow transp. \times [Gly \downarrow i,j]$$

0 **in** **Reaction** **out**

(CSTR)

$i=36 \text{ glycans}$
 $j=4 \text{ compartment}$

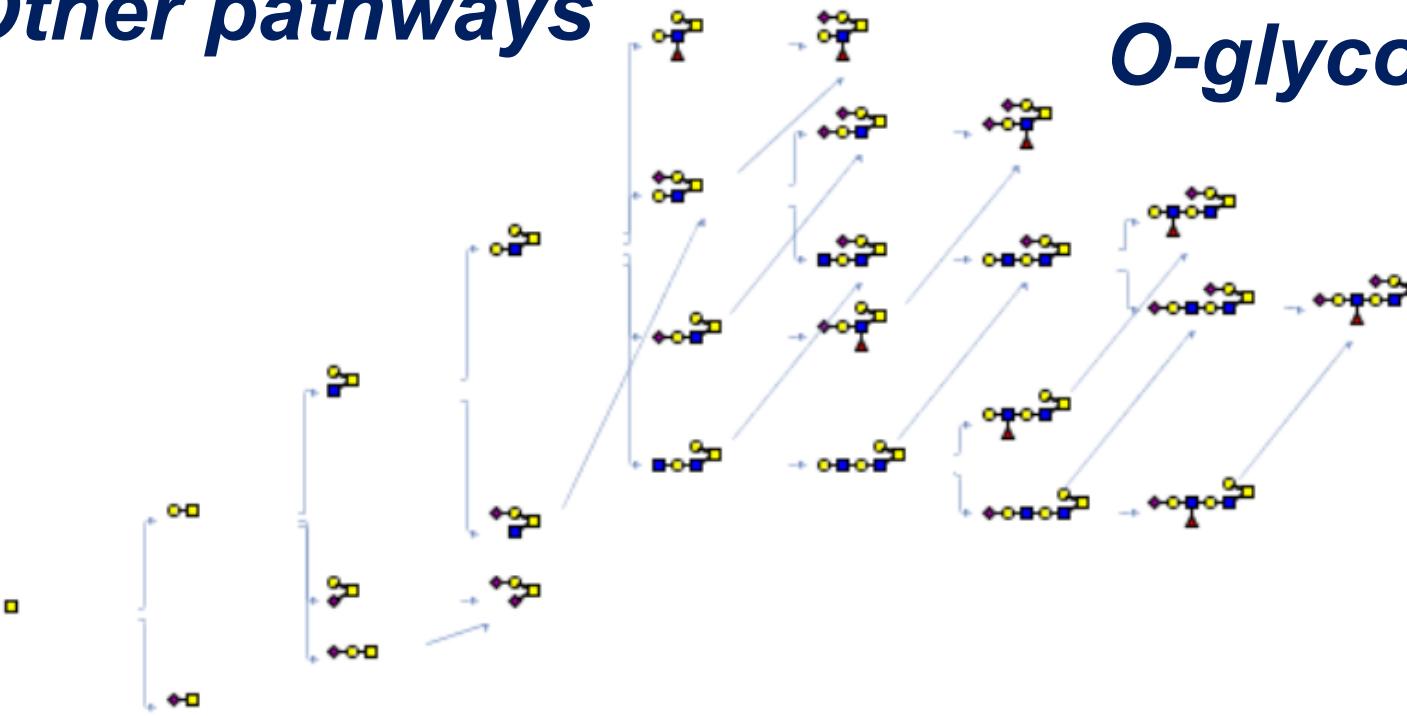
In silico simulation

- *Deterministic*
- *Stochastic*

	High mannose	Bi-	Tri-	Tetra-	Bisecting
Cis-	11.62%	54.64%	4.61%	0.06%	48.38%
Medial-	2.73%	80.26%	8.85%	0.07%	70.33%
Trans-	2.44%	82.61%	9.12%	0.07%	71.11%
TGN	2.43%	82.73%	9.13%	0.07%	71.11%

Other pathways

O-glycosylation



Glycolipid biosynth.

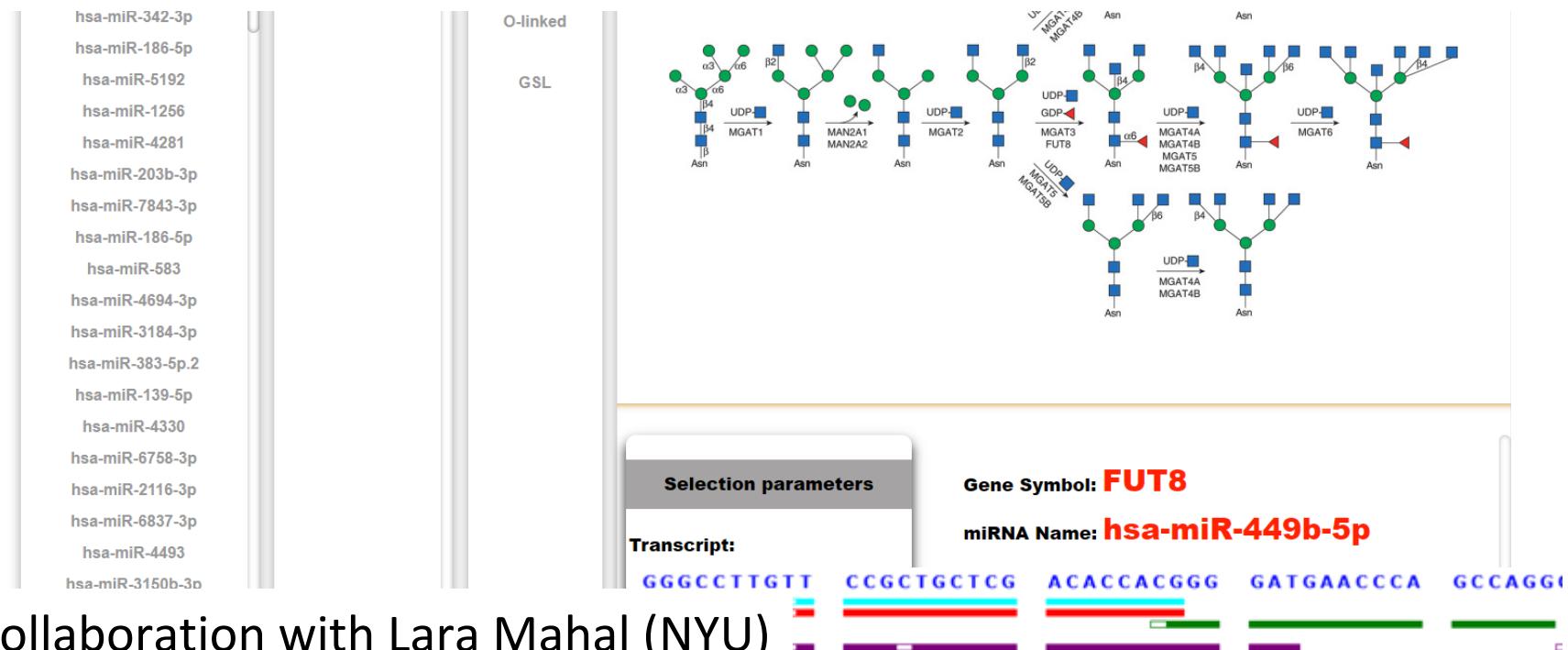


Pathway generation times: short!

Forward inference		
# of species	# of reactions	Time
144	300	~8s
405	692	~30s
916	3330	~70s
166	452	~11s
Reverse inference		
160	526	~12s
181	611	~14s
356	1246	~37s
96	267	~6s

What else could this be useful for?

- Mapping RNA-Seq and Glycomics data to construct pathway maps.
- GlycoMir: The glycogene microRNA targets
- *Functional integration with other databases*

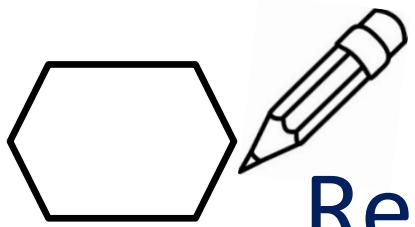




2. DrawGlycan-SNFG

Kai Cheng

- From IUPAC to Symbolic Nomenclature for Glycans (SNFG)
- Draw glycopeptides
- Draw glycan and peptide fragmentation
- *Other features...*



DrawGlycan-SNFG: Render glycans and glycopeptides

IGEADFN[Gal(b1-4)GlcNAc(b1-6 -NR "366.1")

[Gal(b1-4)GlcNAc(b1-2)]Man(a1-6)

[Neu5Ac(a2-6 -NR "292.1" -R

"3350")Gal(b1-4)GlcNAc(b1-2 -NR "737.2" -R

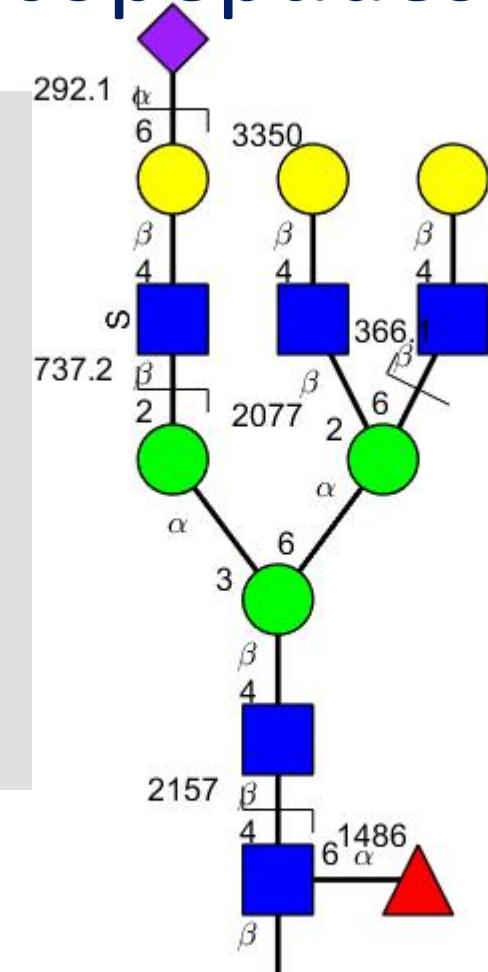
"2077" -U

"S")Man(a1-3)]Man(b1-4)GlcNAc(b1-4 -NR

[~~B157~~(a1-6)[1486")][Fuc(a1-6)GlcNAc(Nac(4?NR)]RSK

"2157" -R "1486")][Fuc(a1-6)GlcNAc(b?-?)]RSK

Advantages: straightforward, easy to read & write, adequate for common use

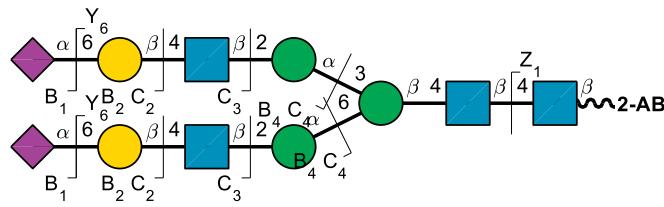
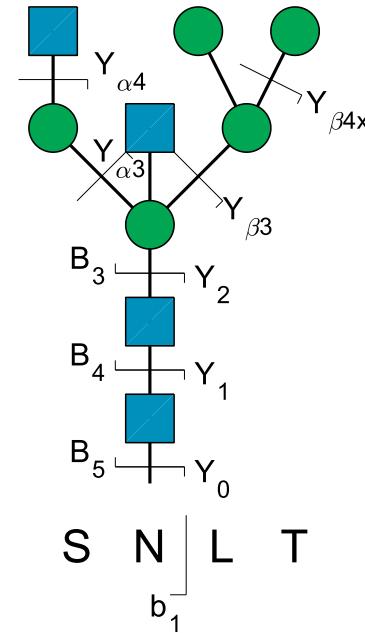
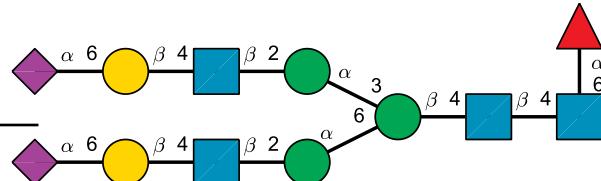


I G**EADFNRSK**
b2 171.1
y8

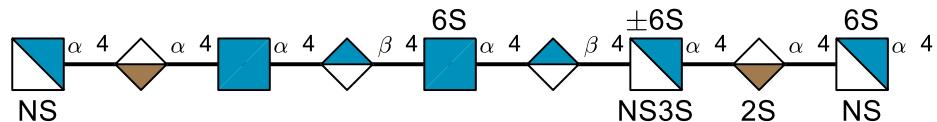
Fragmentation options

-Option	Representation
----------------	-----------------------

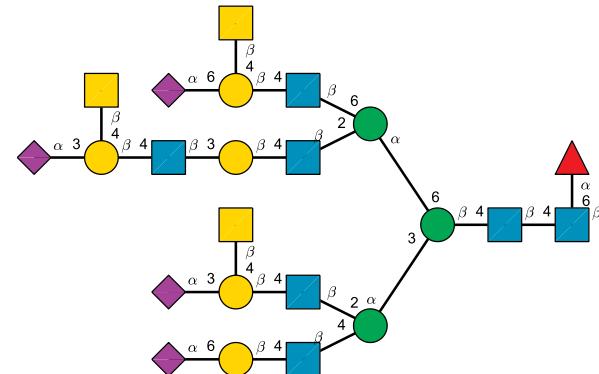
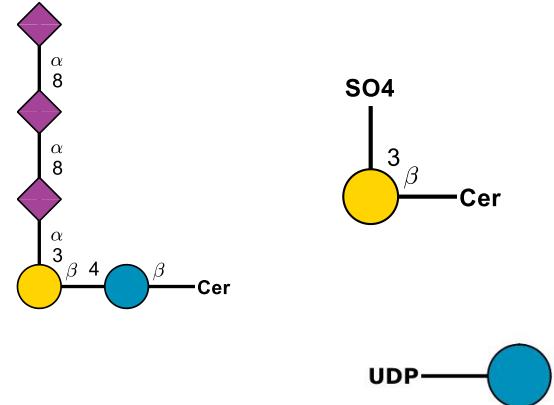
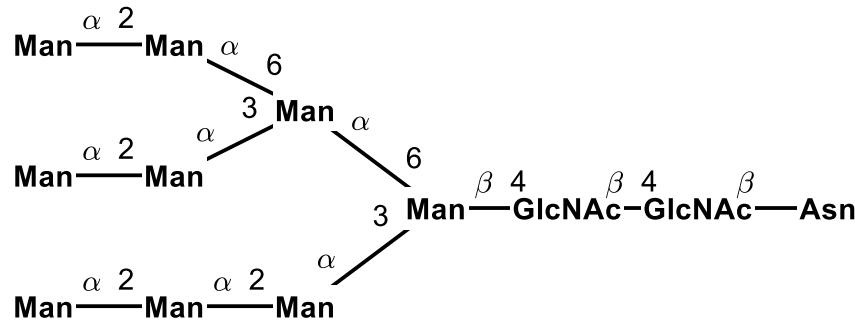
- 1 -R Glycan reducing end
- 2 -NR Glycan non-reducing end
- 3 -N Peptide backbone N-terminus
- 4 -C Peptide backbone C-terminus



Monosaccharide options



-Option	Representation
1 -U	Annotate above monosac.
2 -D	Annotate below monosac.
3 -P	Identify a perpendicular monosac.
4 -CHAR	Introduce arbitrary text or present monosac. in text form



Bond options

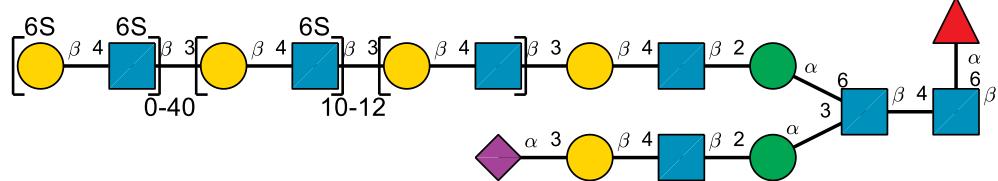
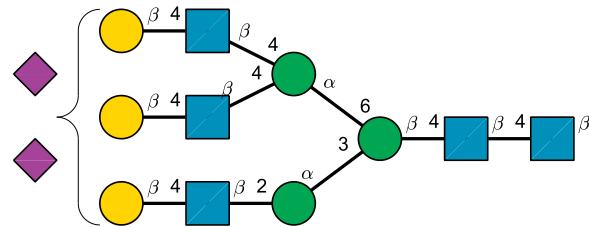
-Option	Representation
1 -BOLD	Paint glycosidic bond bold
2 -ZIG	Paint glycosidic bond zigzag
3 -WAVY	Paint glycosidic bond wavy
4 -DASH	Paint glycosidic bond dashed
5 -WEDGE	Paint glycosidic bond wedge

Chemical structures illustrating the bond options:

- BOLD:** Shows a glycosidic bond between a purple diamond and a yellow circle, represented by a thick black line.
- ZIG:** Shows a glycosidic bond between a purple diamond and a yellow circle, represented by a zigzag black line.
- WAVY:** Shows a glycosidic bond between a purple diamond and a yellow circle, represented by a wavy black line.
- DASH:** Shows a glycosidic bond between a purple diamond and a yellow circle, represented by a dashed black line.
- WEDGE:** Shows a glycosidic bond between a purple diamond and a yellow circle, represented by a wedge bond (dashed line).

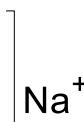
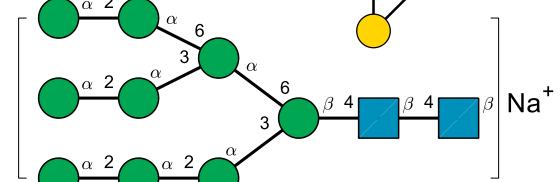
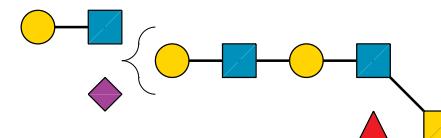
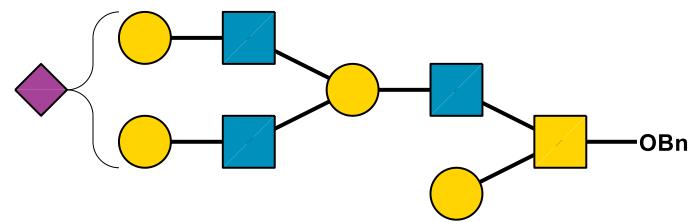
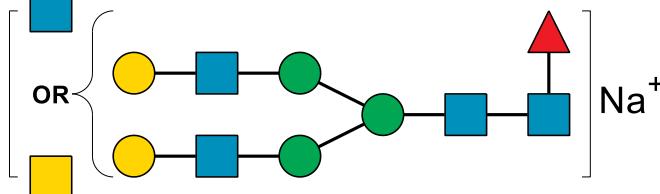
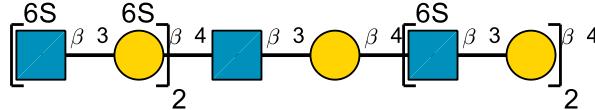
Other chemical structures shown include a branched glycan chain with various substituents and a methyl group (CH_3) with three hydrogens.

Repeats, adducts and fuzzy options



-Option Representation

1	-RS	Repeating unit start
2	-RE	Repeating unit ends
3	-ADDUCT	Add glycan adduct
4	-CURLY	Ambiguous assignments/fuzzy structures



DrawGlycan: Web, GUI & Command-line version



Virtually yours,
Glycan.sugarcane.org/DrawGlycan

3. GlycoPAT

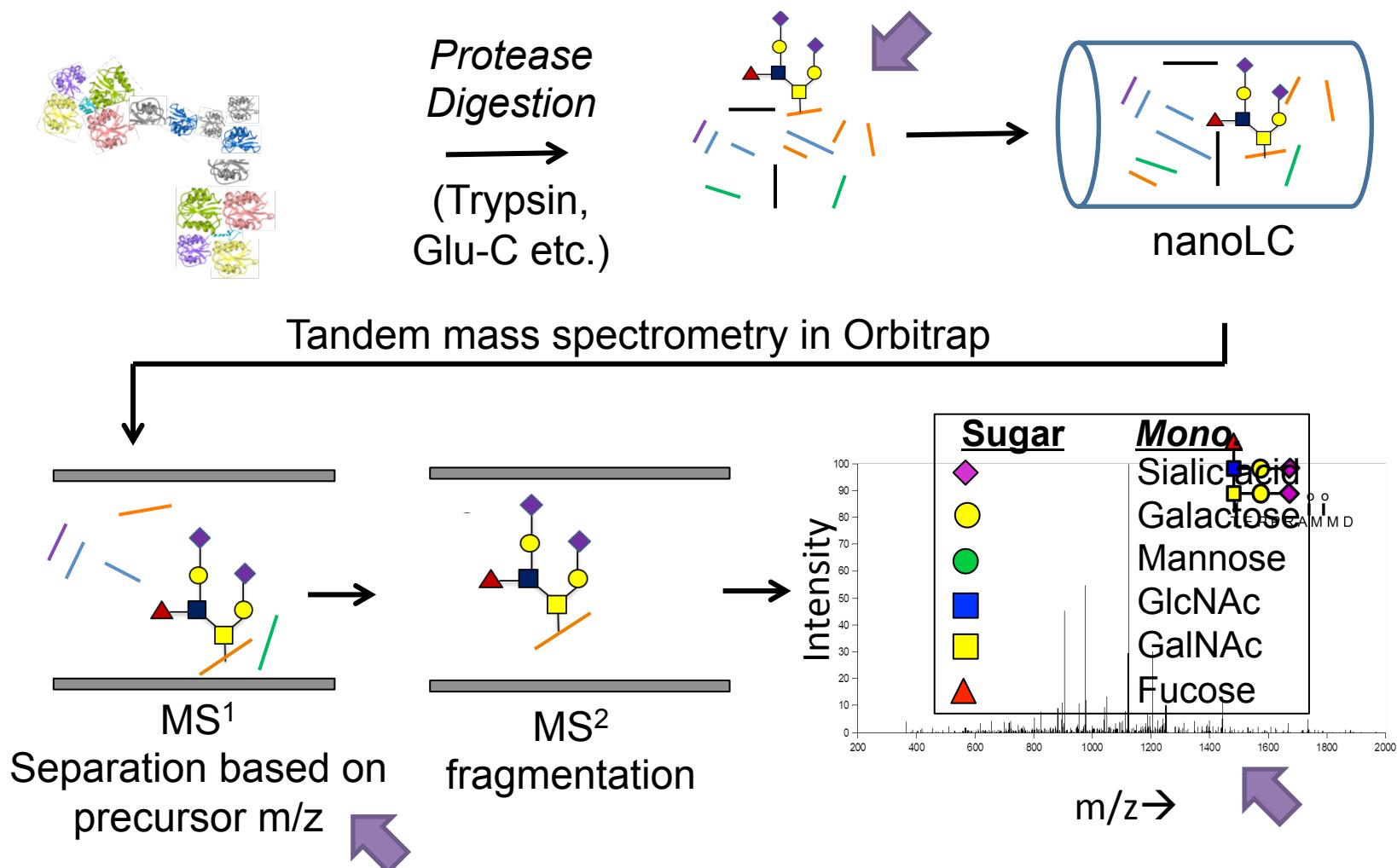


Kai Cheng

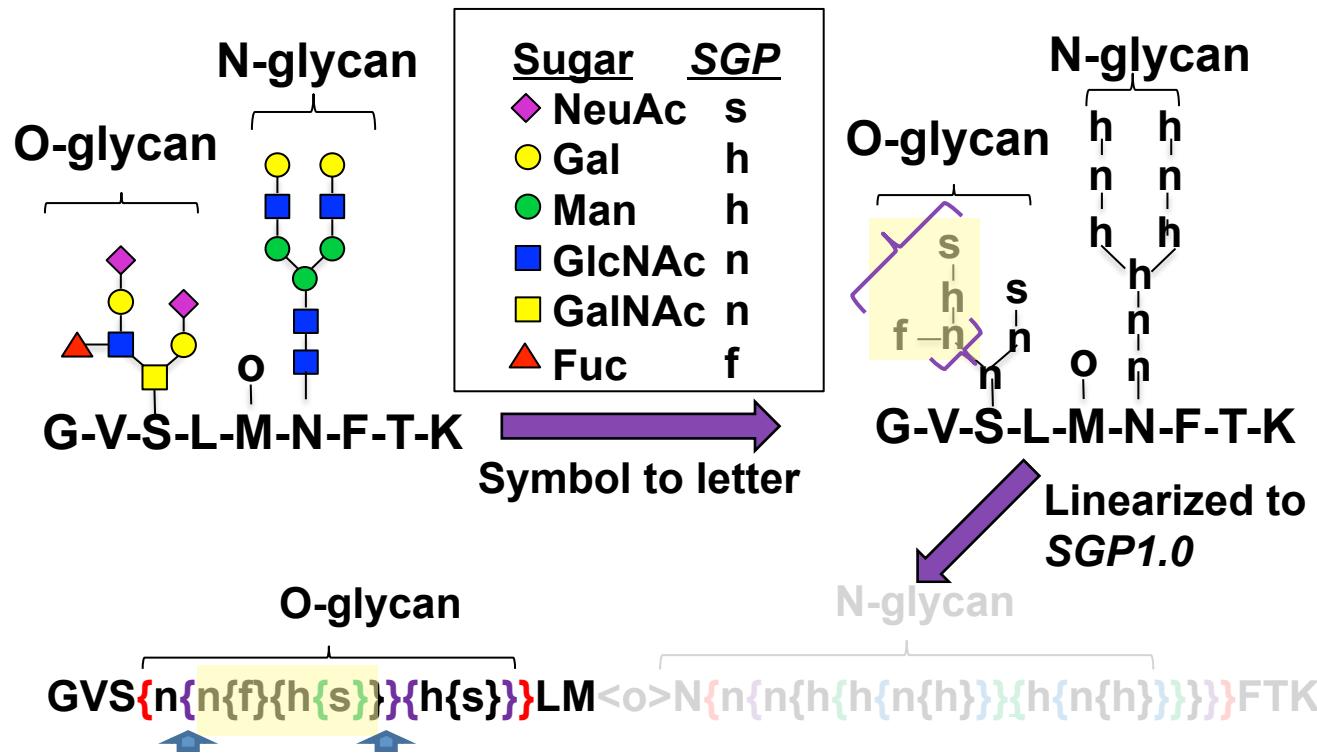
- Analyze high-throughput glycoproteomics data
- Comprehensive scoring and false discovery rate (FDR) calculation algorithm for MSⁿ data analysis

*Mol Cell Proteomics.*16:2032-47, Nov 2017.

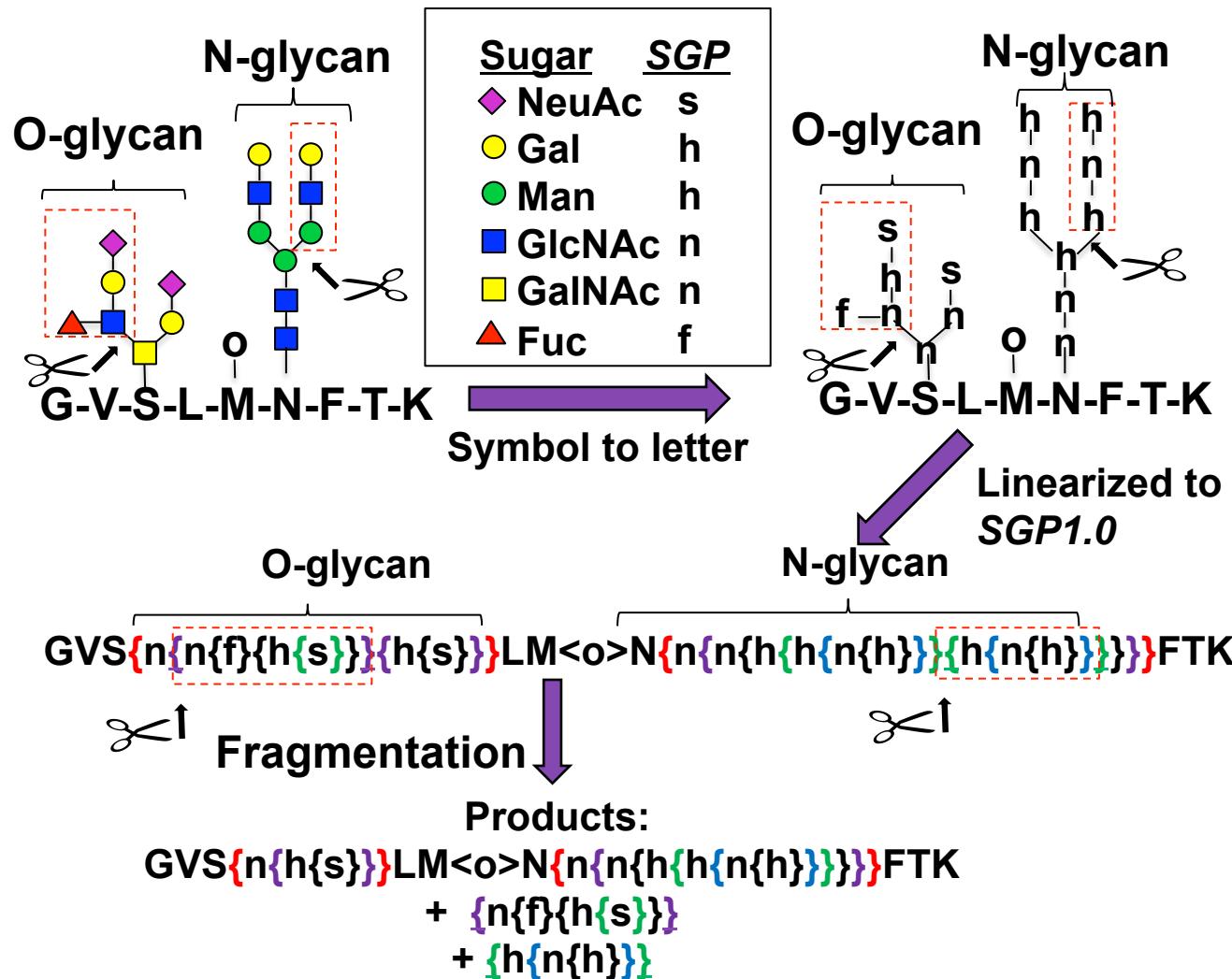
GlycoPAT: High-throughput glycoproteomics analysis



SmallGlyPep : The *minimal* representation of glycopeptide for MS

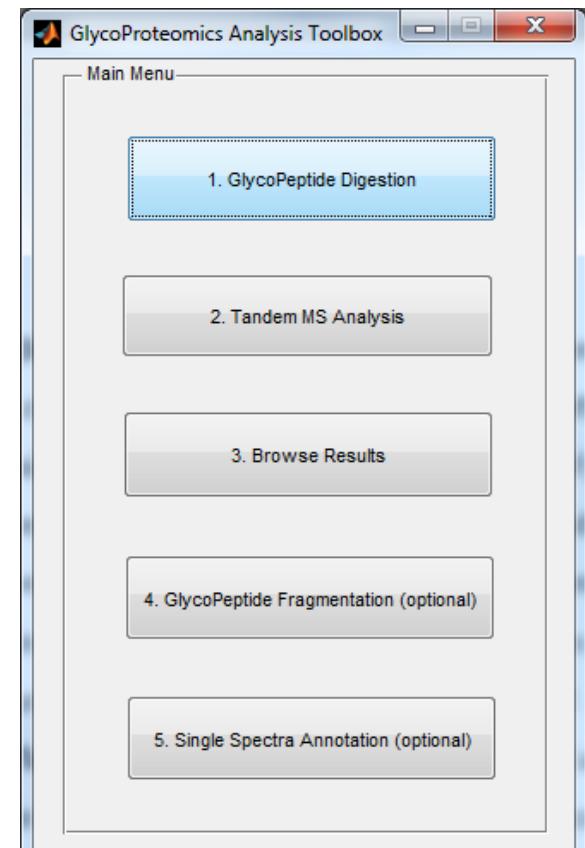
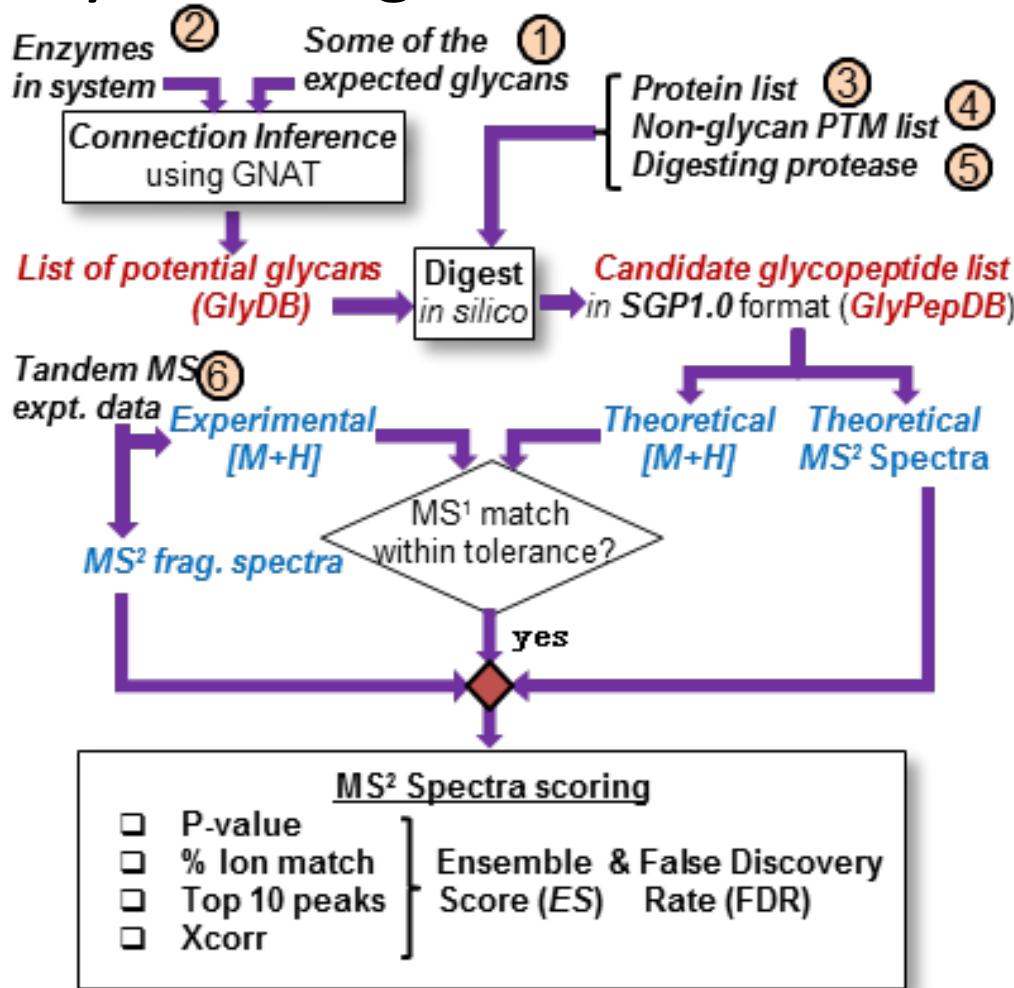


SmallGlyPep : The *minimal* representation of glycopeptide for MS



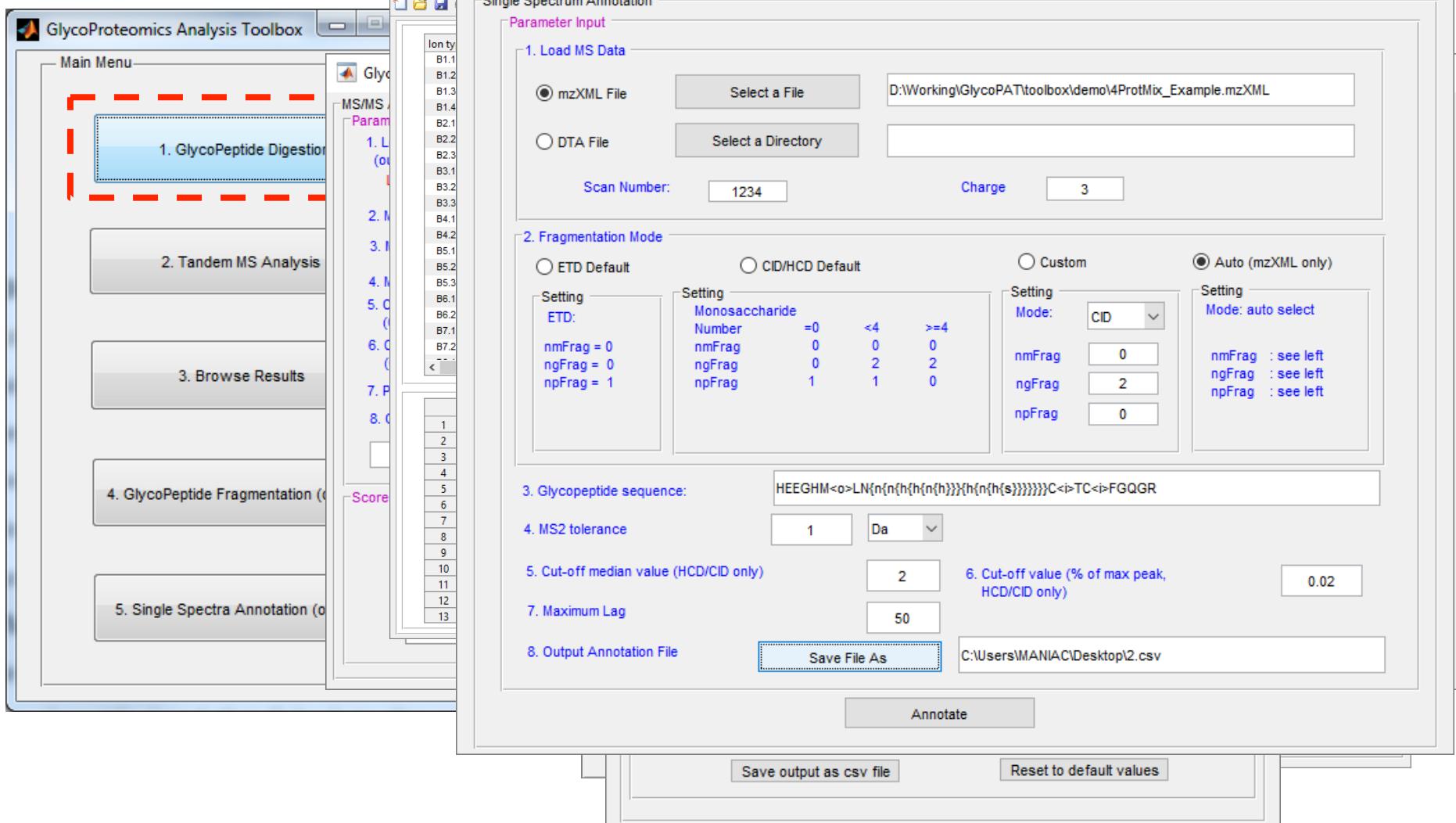
Methods

GlycoPAT: general workflow

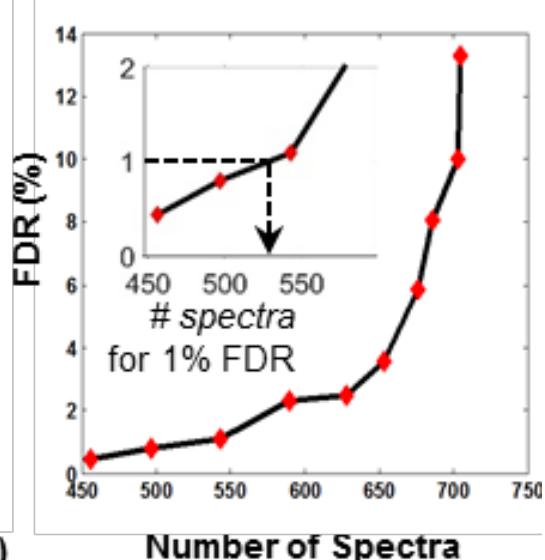
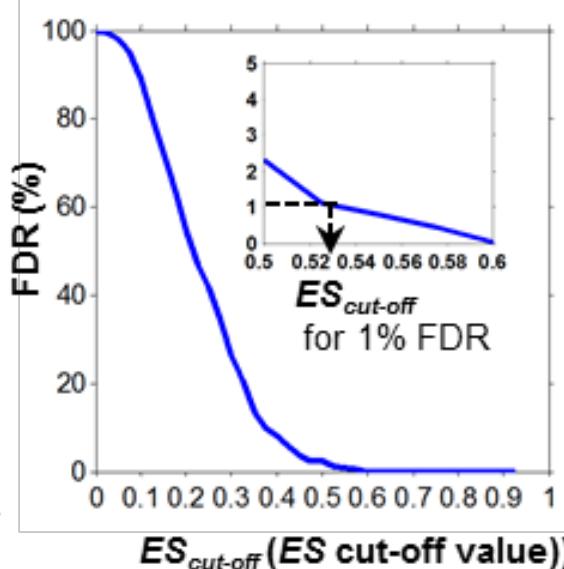
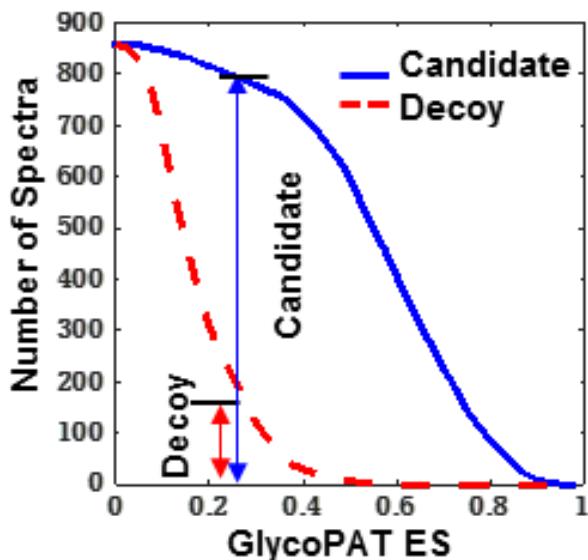
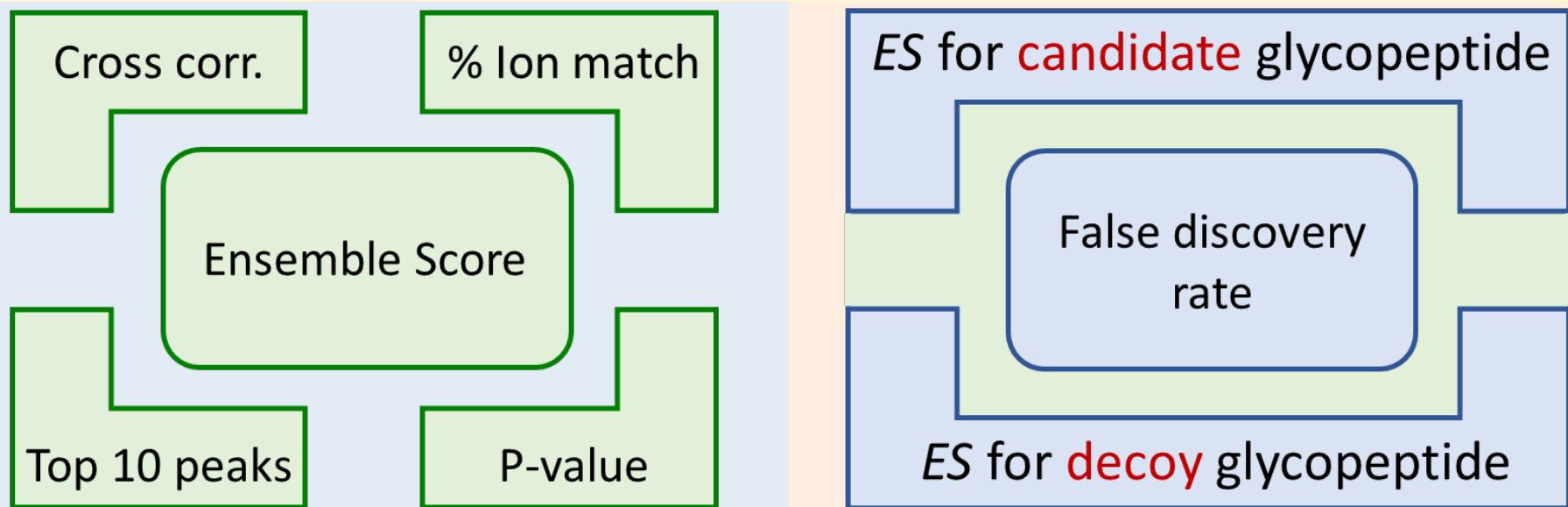


Methods

GlycoPAT: GUI



GlycoPAT: ensemble score (*ES*) and false discovery rate (*FDR*)

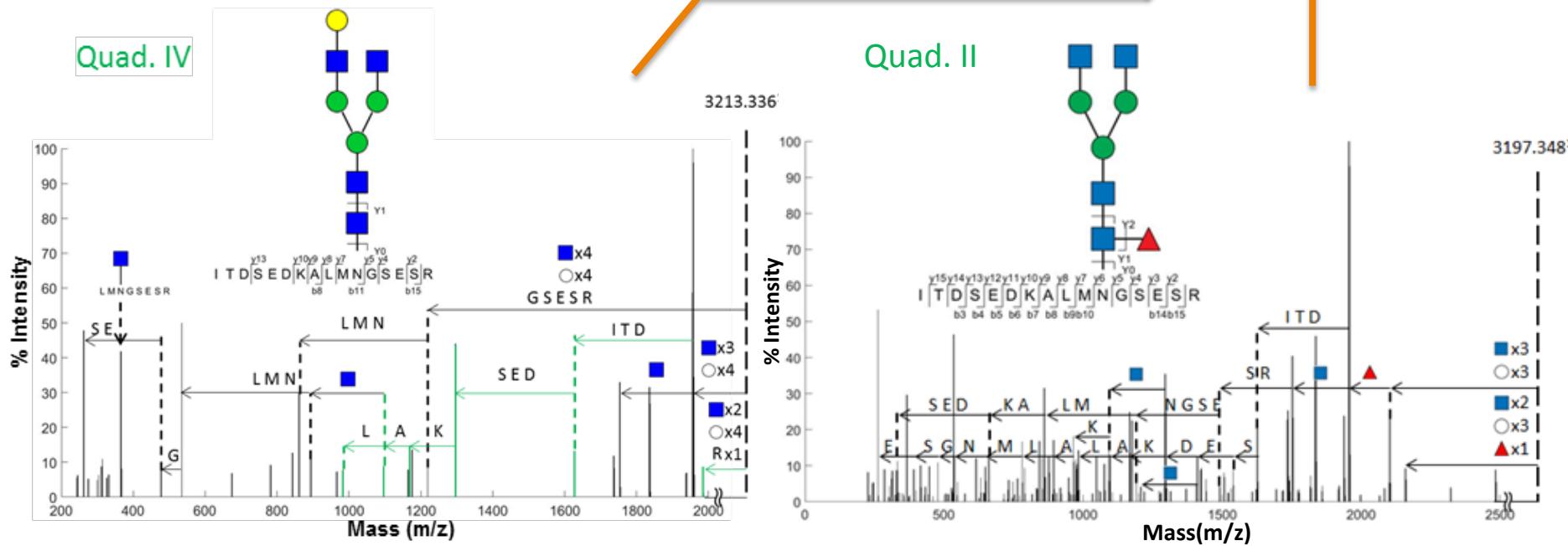


GlycoPAT: scoring of HCD spectra

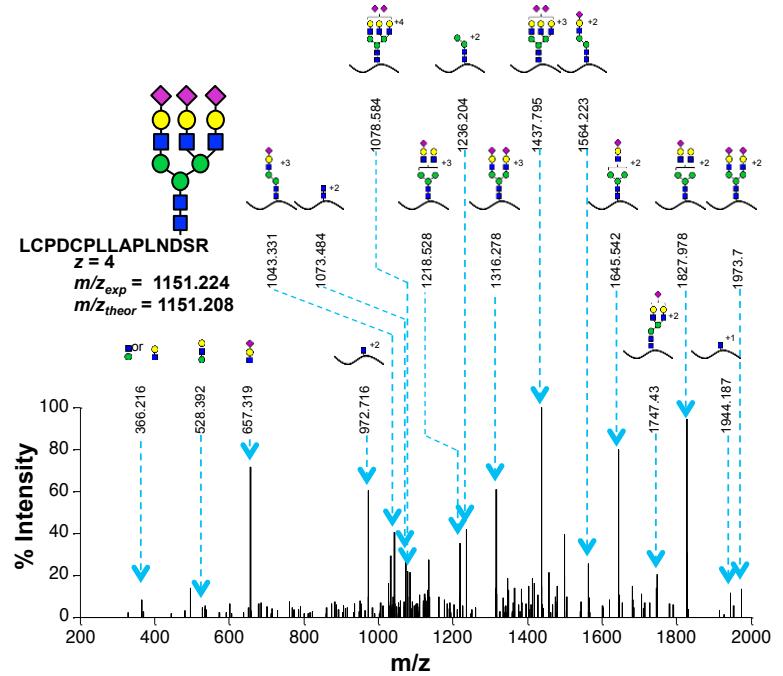
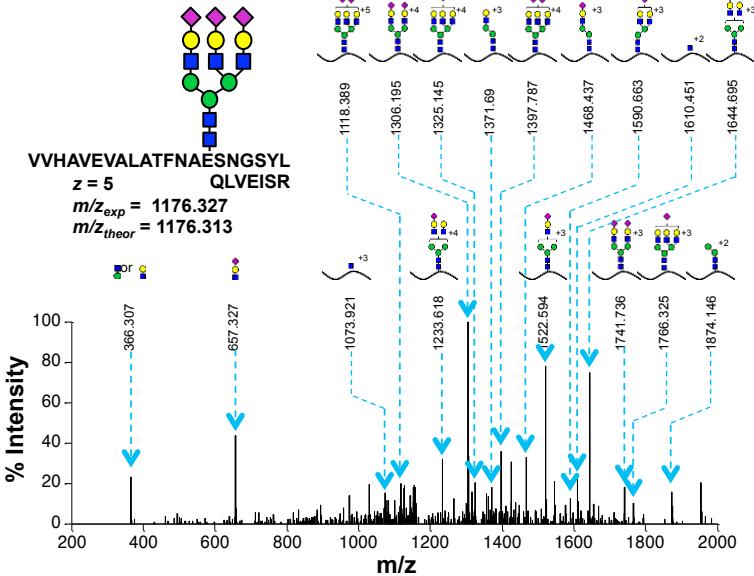
Basigen/HCD

J Proteome Res. 15(10):3904-3915, 2016

* Simultaneous breakage of peptide backbone and glycan structures

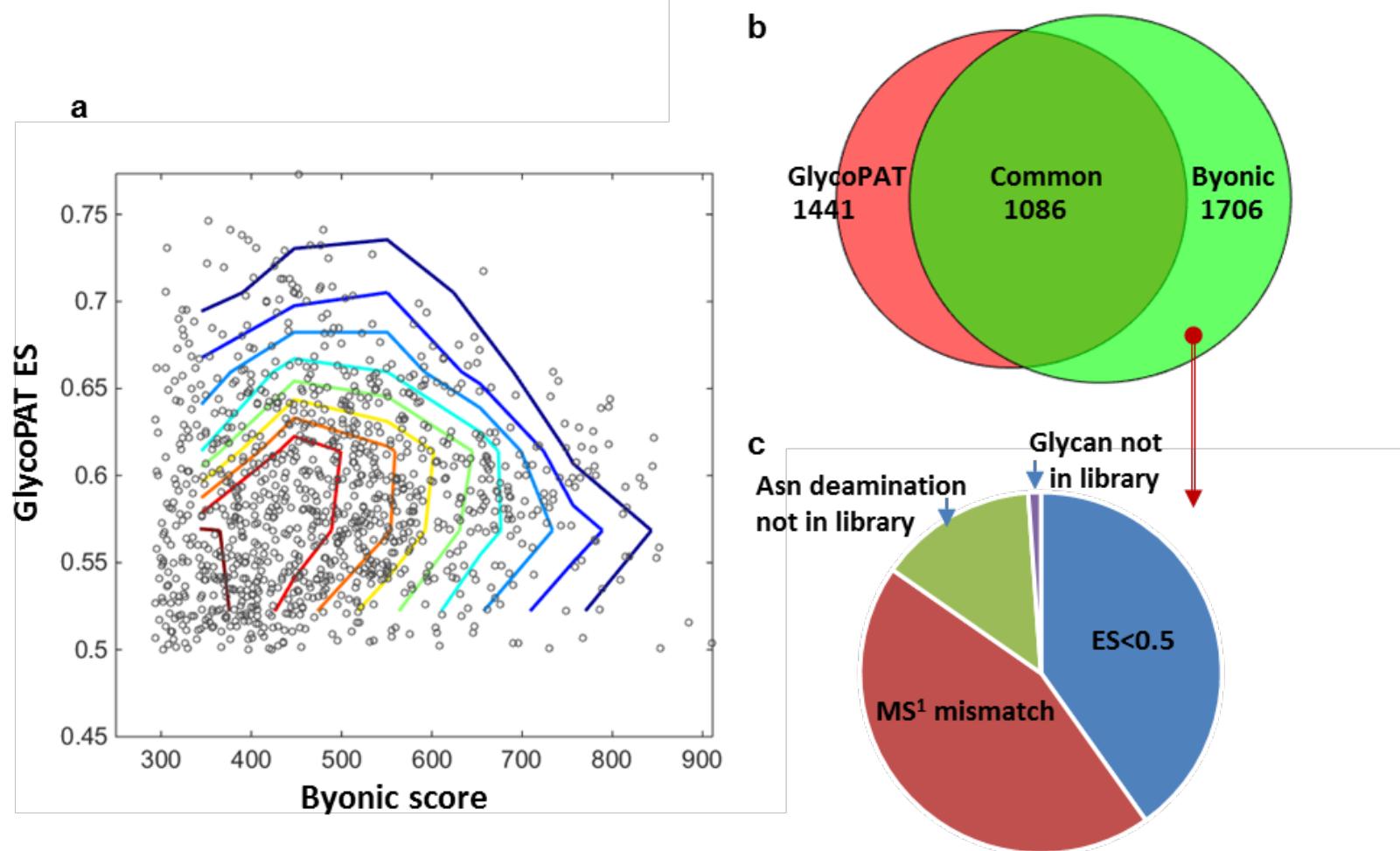


GlycoPAT: scoring of CID spectra



* Analysis of ladder-like breakdown of glycans

Analysis of glycopeptides in whole prostate cancer lysates

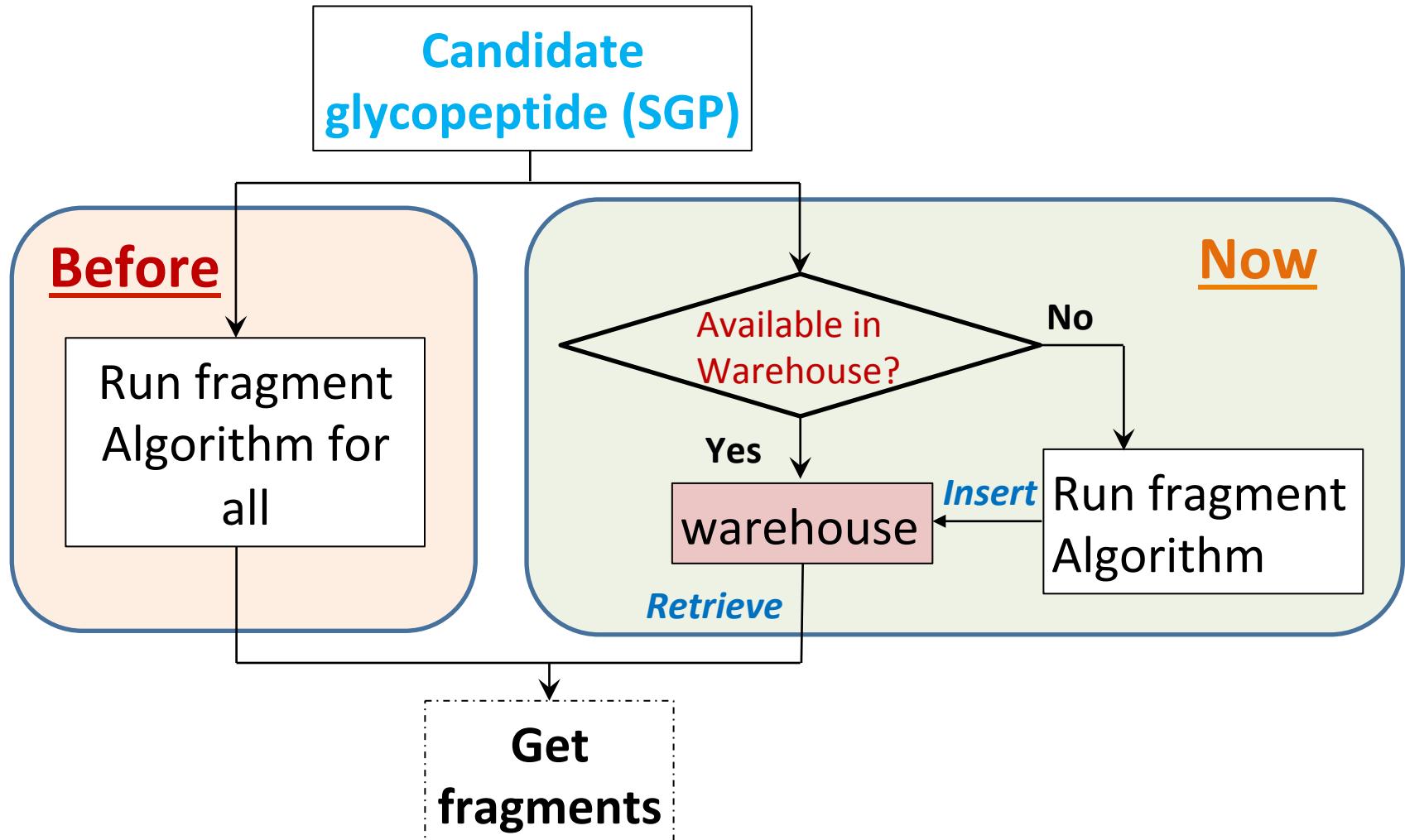


Work in progress

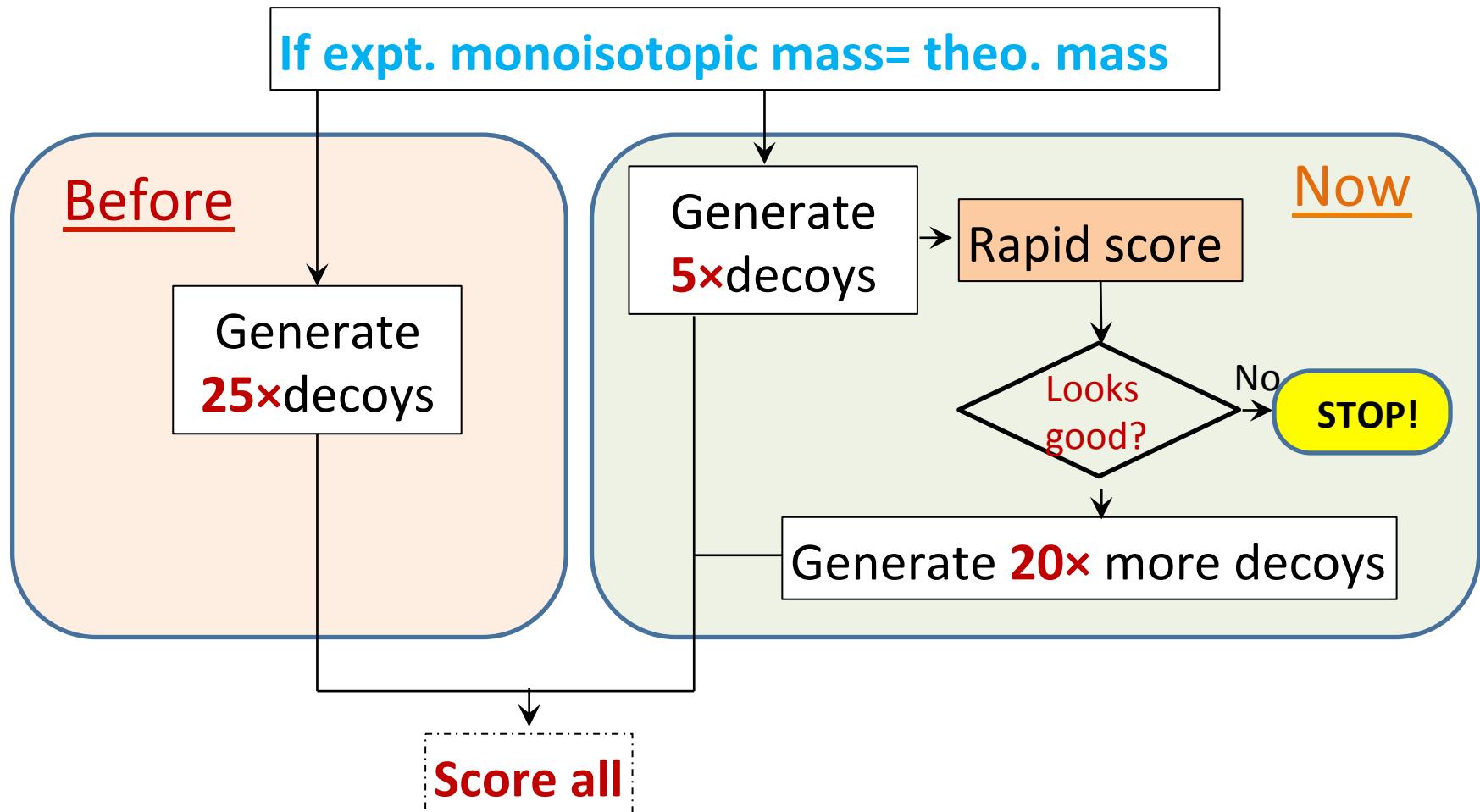
- Improving calculation speed:

FOCUS!! Don't do everything for everyone!

Improve speed: *The fragment warehouse*



Improved speed: *Selective scoring*



Work in progress

- Improving calculation speed:

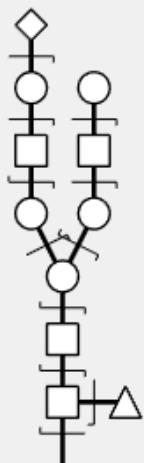
FOCUS!! Don't do everything for everyone!

- Streamlining result visualization:

SIMPLIFY!! Tell the story in pictures!

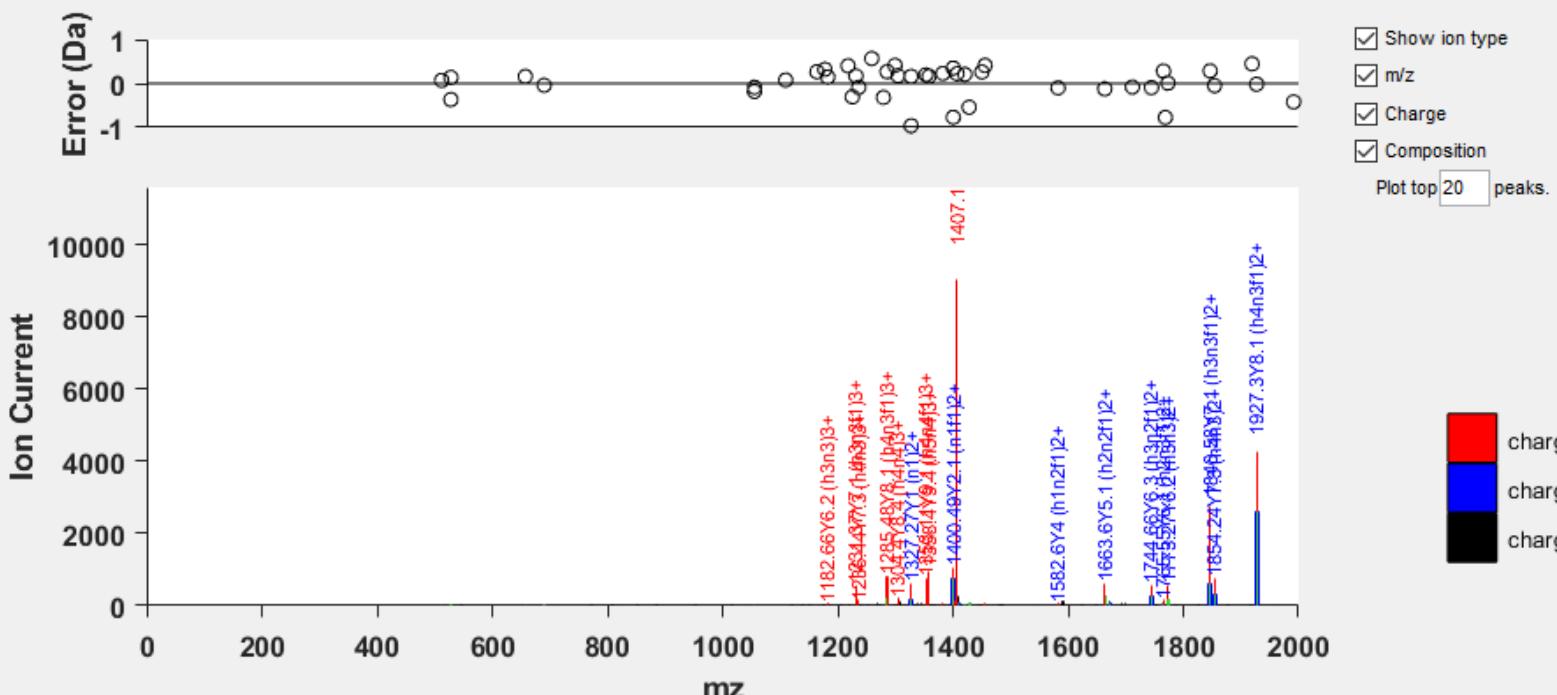
Spectrum MS¹ Detail MS² Detail

Combine Drawglycan with GlycoPAT



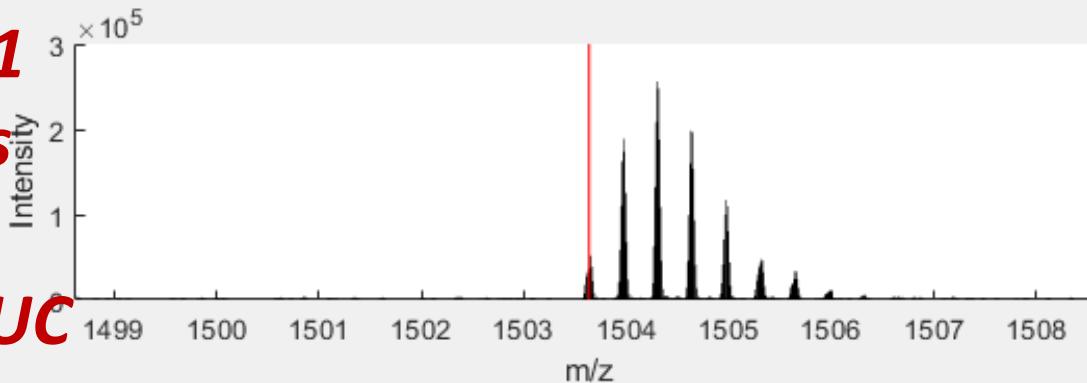
G L Q P T L T N P G E C R P N F T C A C R

	Name	Value
1	SGP	GLQPTLTNPGE<i>RPN[n{n{h{h{n{h{...
2	Scan number	3084
3	RT	42.2025
4	Prec. m/z	1.5036e+03
5	Charge	3
6	ES	0.7800
7	AUC	1.2450e+05
8	Top 10	9
9	XCorr	9.3600
10	p-Value	0
11	% Ion match	48.6000

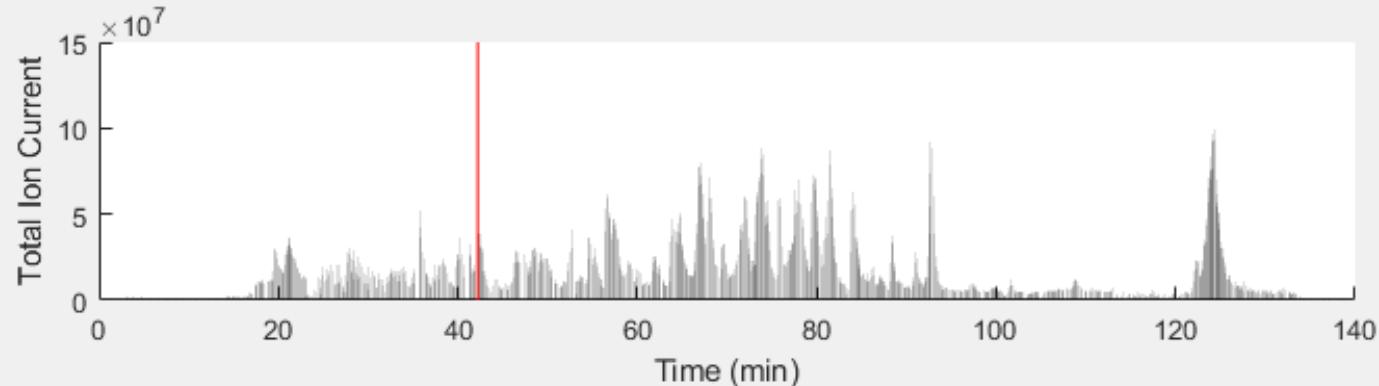
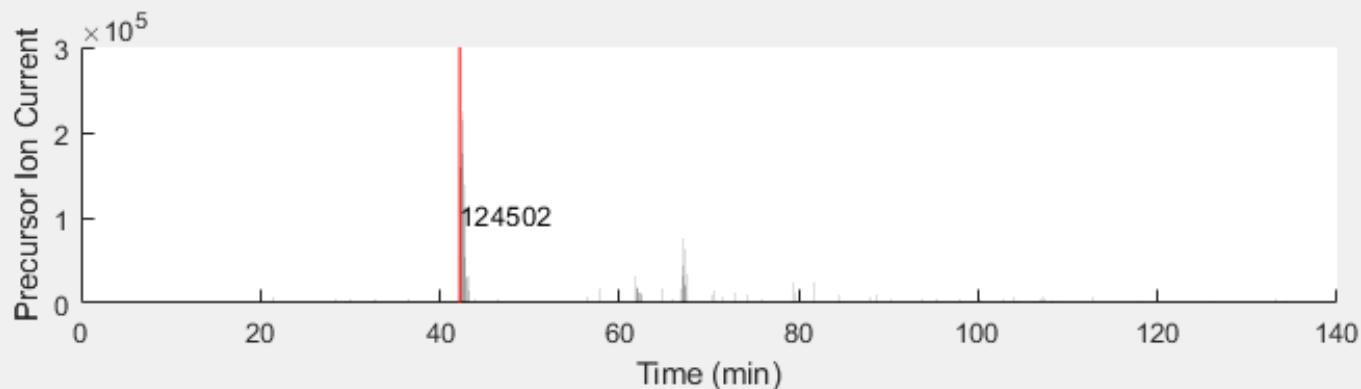


Spectrum MS¹ Detail MS² Detail

Confirm MS1 assignments & calculate AUC



+
+/- 2 min(s)
-



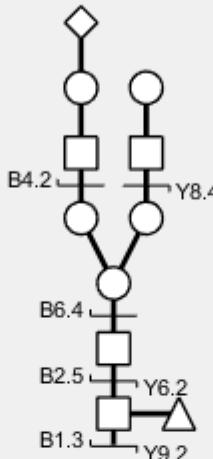
Glycoproteomics Analysis of Single Spectrum (scan number= 3079)

- X

File Edit View Insert Tools Desktop Window Help

Spectrum MS¹ Detail MS² Detail

Provide more details



G L Q P T L T N P G E C R P N F T C A C R

	sgp	charge	mz	mzExpt	DaError	Intensity	iontype	nm/np/ngFrac
1	{h}	3	55.36	1285.55	0.33	28724	B1.8	0 0 2
2	{h{h{n{h}}}}	3	231.09	1407.32	0.39	11106	B4.2	0 0 2
3	{f}	3	50.03	1231.45	0.25	9567	B1.3	0 0 1
4	{h{n{h}}}	3	177.07	1382.63	0.38	7832	B3.2	0 0 1
5	{n}	3	69.04	1236.83	0.3	6917	B1.4	0 0 2
6	GLQPTLTNPGE<i>R	2	1225.57	1182.71	0.2	2458	Y0	0 0 1
7	{h{h{n{h{s}}}}}{h{n}}	3	449.83	1663.96	0.24	2020	B7.5	0 0 2
8	{n{n{h{n{h{s}}}}}}{f}	3	512.19	1745.53	0.77	1947	B8.3	0 0 2
9	GLQPTLTNPGE<i>RPN{n{n{h{h{n}}}}}{h{...}}	1	3909.65	366.28	0.14	1728	Y8.4	0 0 2
10	{h{n}}	3	123.05	1333.78	0.22	1574	B2.4	0 0 2
11	{h{n}}	3	123.05	1353.03	0.12	1403	B2.6	0 0 2
12	{n{h{s}}}	3	220.09	1358.11	-0.13	1219	B3.1	0 0 1
13	{n{h{s}}}	3	220.09	1358.89	0.65	1128	B3.1	0 0 1

Conclusions

- **GNAT-Web**: Build reaction networks efficiently
 - Use this for *in silico* deterministic and stochastic simulations
 - Display of experimental data sets
- **DrawGlycan-SNFG**: Easy and Robust
- **GlycoPAT**: MS data analysis toolbox
 - Improve computational time
 - Integrate analysis from different fragmentation modes
 - Test in more biomedical applications

Acknowledgements

Lab members:

Anju Kelkar, Ph.D.

Virginia del Solar Fernandez, Ph.D.

Graduate students

Ted Groth

Changjie Zhang

Kai Cheng

Xinheng Yu

Yuqi Zhu

Yusen Zhou

Arezoo Momeni

Gabbie Pawlowski

Collaborators:

Alan Friedman and Jun Qu, Buffalo

Anne Dell, Stuart Haslam

Imperial College

Funding support:

NHLBI Systems Biology Collaborations



NIGMS: General Medicine



NEW YORK STATE STEM CELL SCIENCE

