Glycoinformatics tools to analyze and curate large scale experimental datasets

Sriram Neelamegham

Departments of Chemical & Biological Engineering,
Biomedical Engineering and Medicine

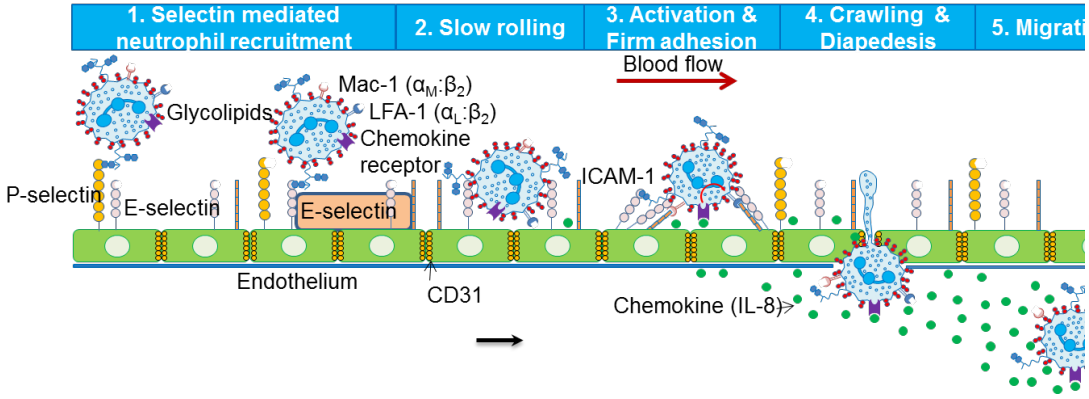
State University of New York, Buffalo, NY

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

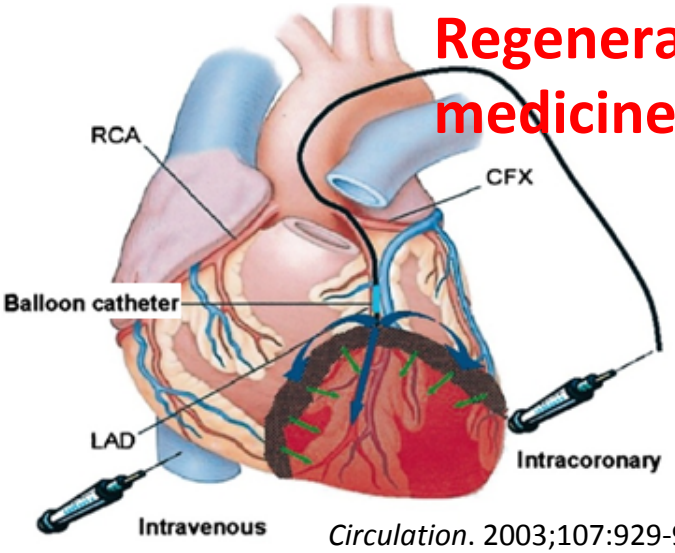
University at Buffalo
The State University of New York

Overview of research interests

Inflammation

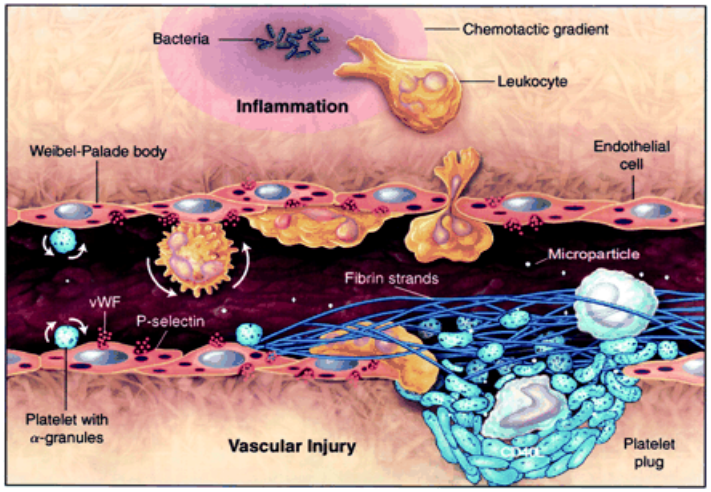


Regenerative medicine



Circulation. 2003;107:929-934

Thrombosis



ATVB, 25:1321, 2005.

Systems Biology

Input-output relationship

Input-output response

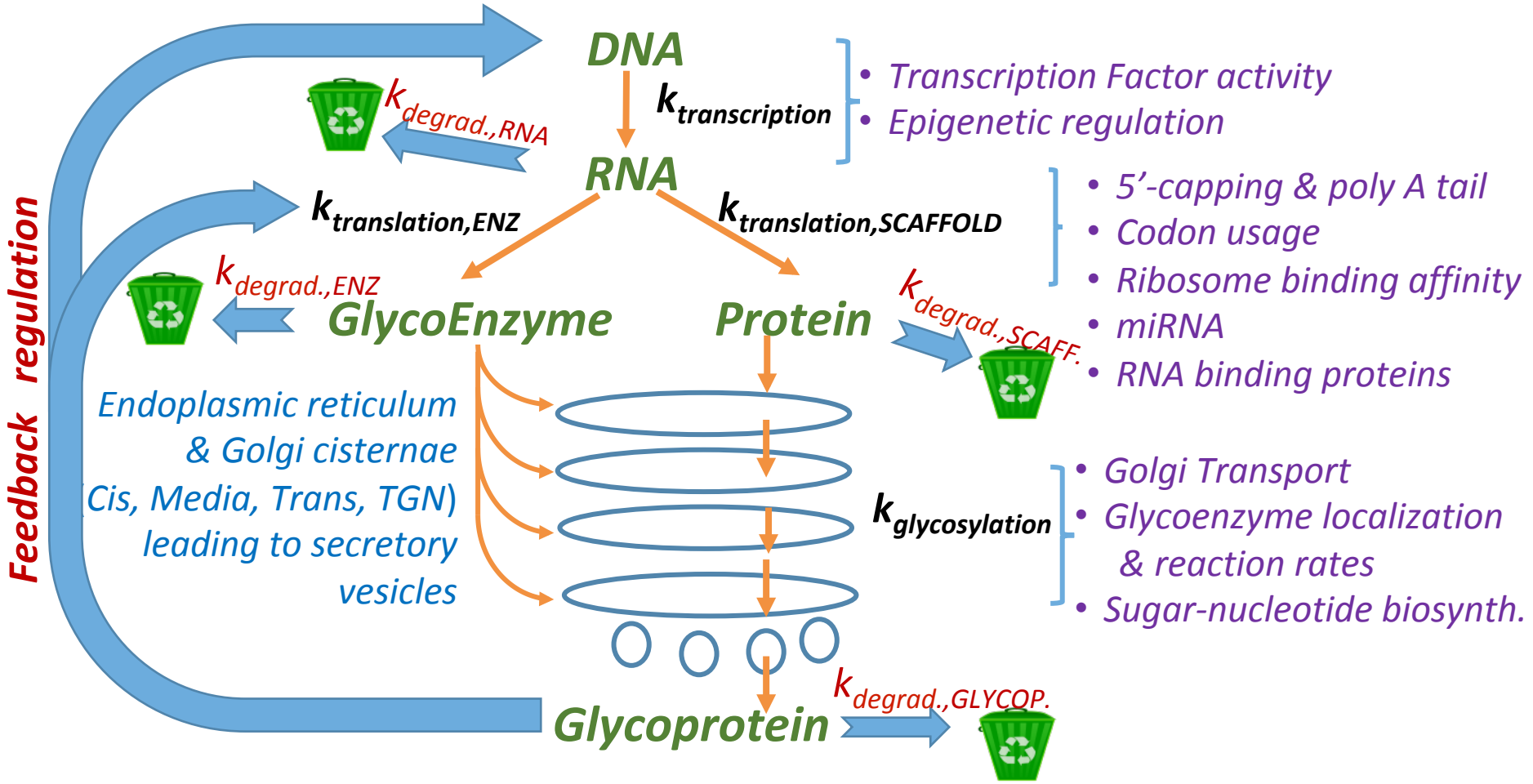
- **Their generation**

  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



Neelamegham & Mahal
Curr Opin Struct Biol. 40:145-152, 2016

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-MS^n data, with focus on glycoProteomics

1. GNAT-Web

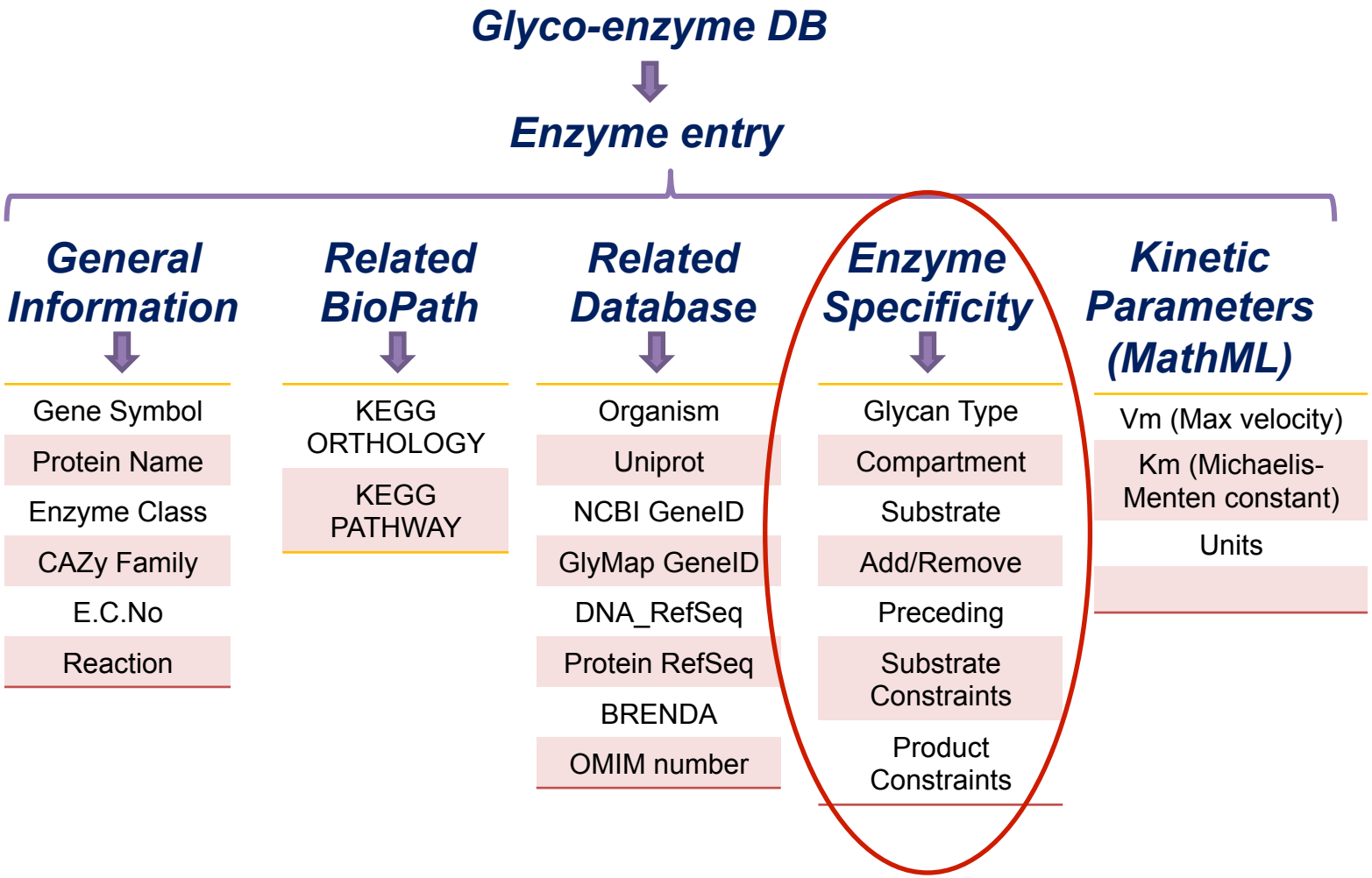
Glycosylation Network Analysis Toolbox

Yusen Zhou

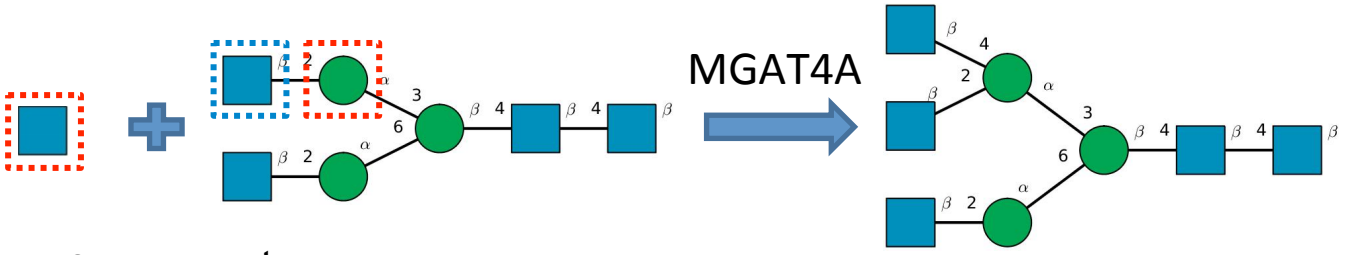
- Define glycoEnzymes *in silico*
- Develop reaction network from RNA-Seq amd MS data processing
- Eventually, simulate reaction networks to bridge data across scales

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)



add GlcNAc(β1-4) to **substrate** '^Man(a1-3)', provided it is **preceded** by GlcNAc(b1-2)

^: Caret is space for inclusion of arbitrary branches

Constraint: MGAT4 acts before addition of:

a. Galactose (Gal), i.e. Gal cannot exist in string or **Gal#0**

b. Bisecting MGAT3, i.e. **GlcNAc(b1-4)^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	^Man(a1-3)
Preceding	GlcNAc(b1-2)

References:

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

SubstConstraints	Value
MaxSubsubst	Gal#0&GlcNAc(b1-4)^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>
```

```
        <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>
```

```
- <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

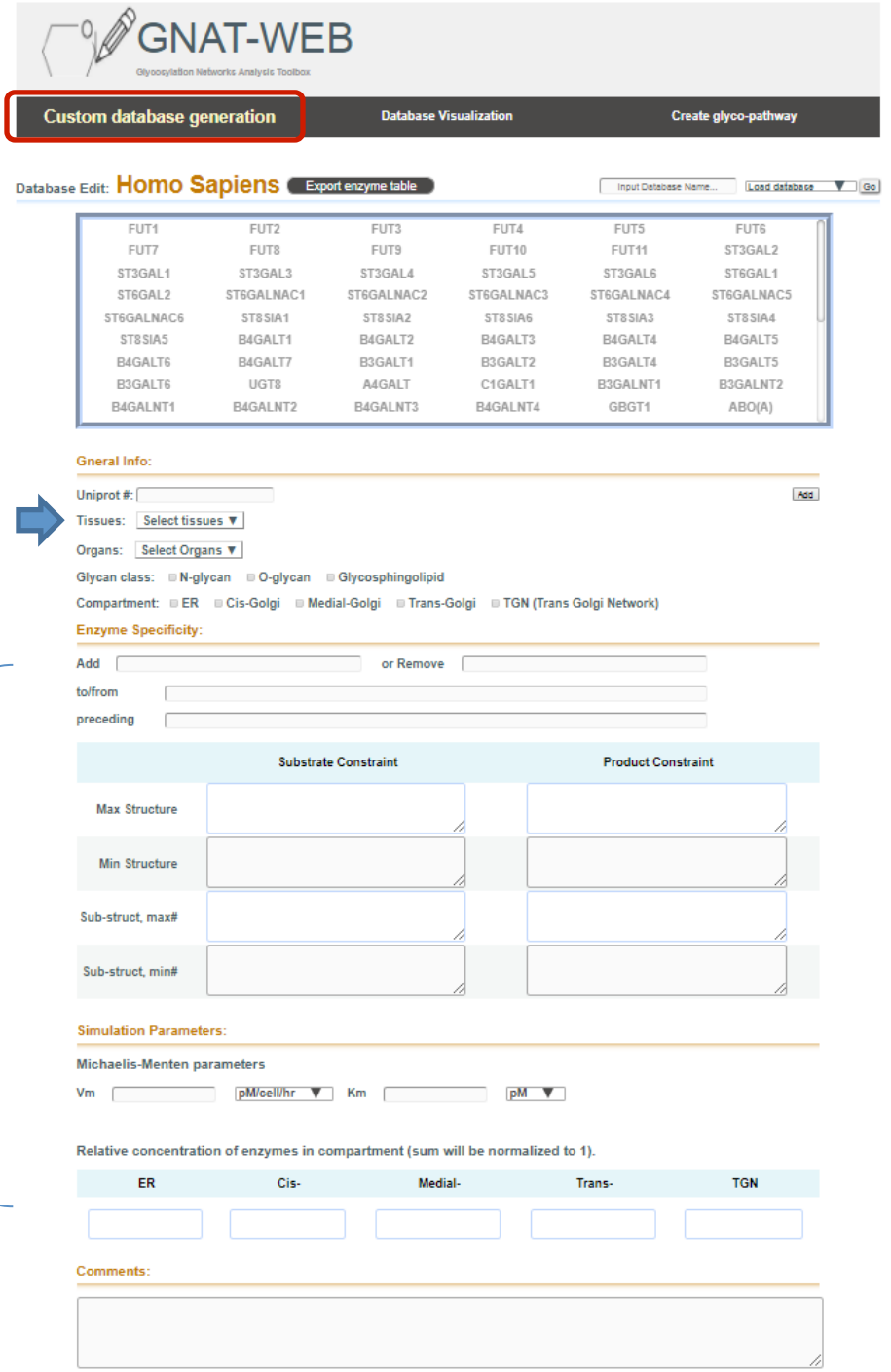
XML data structure for glycoenzyme

```
<?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
      <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)Glc
        <MinSubst>Man(a1-3)[Man(a1-6)[Man(a1-3)]Man(a1-6)]Man(b1-4)GlcNAc(b1-4)GlcN
        <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>
```

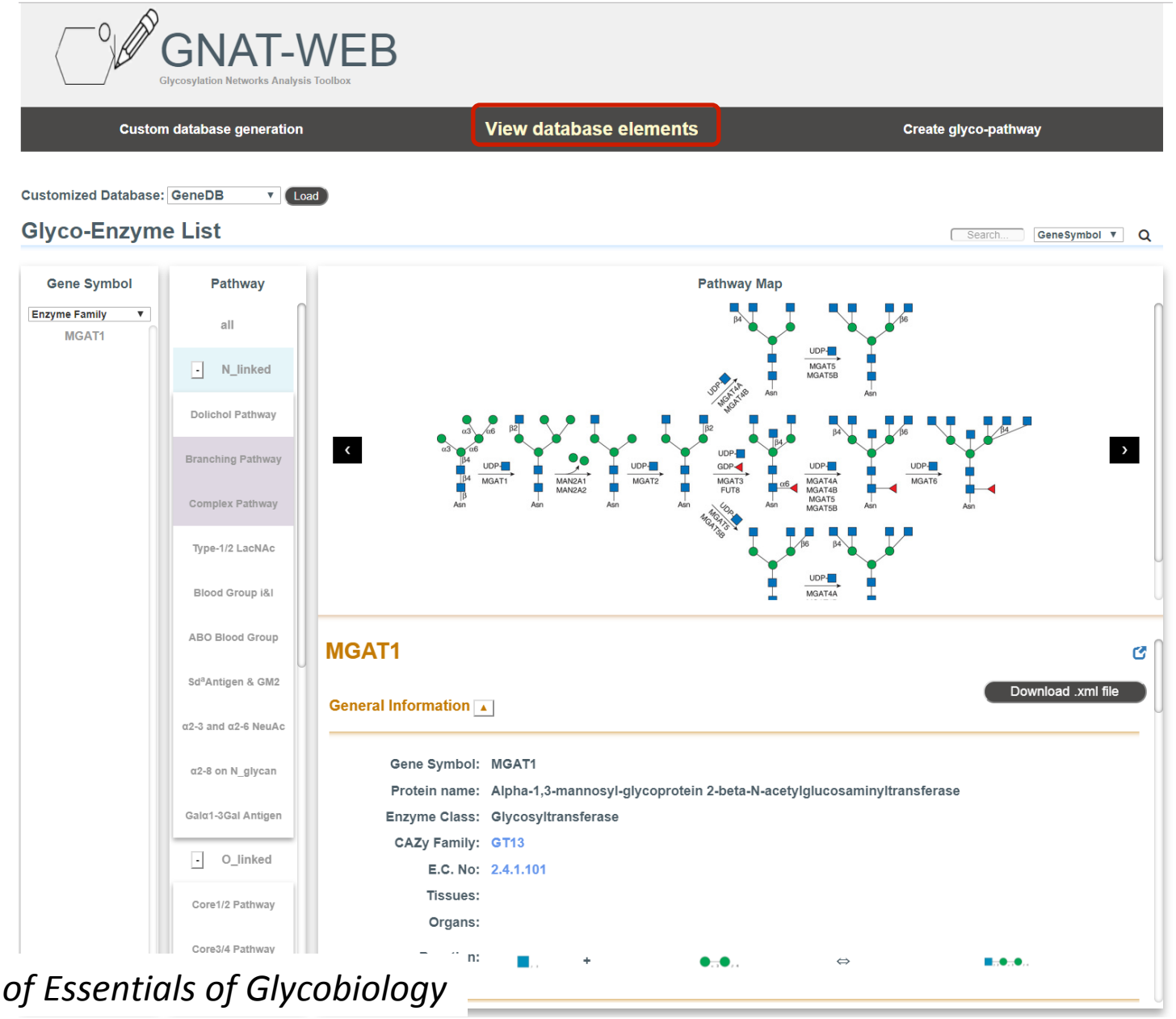
Custom database generation



Obtain from existing databases

Enzyme specificity & kinetics

View database elements

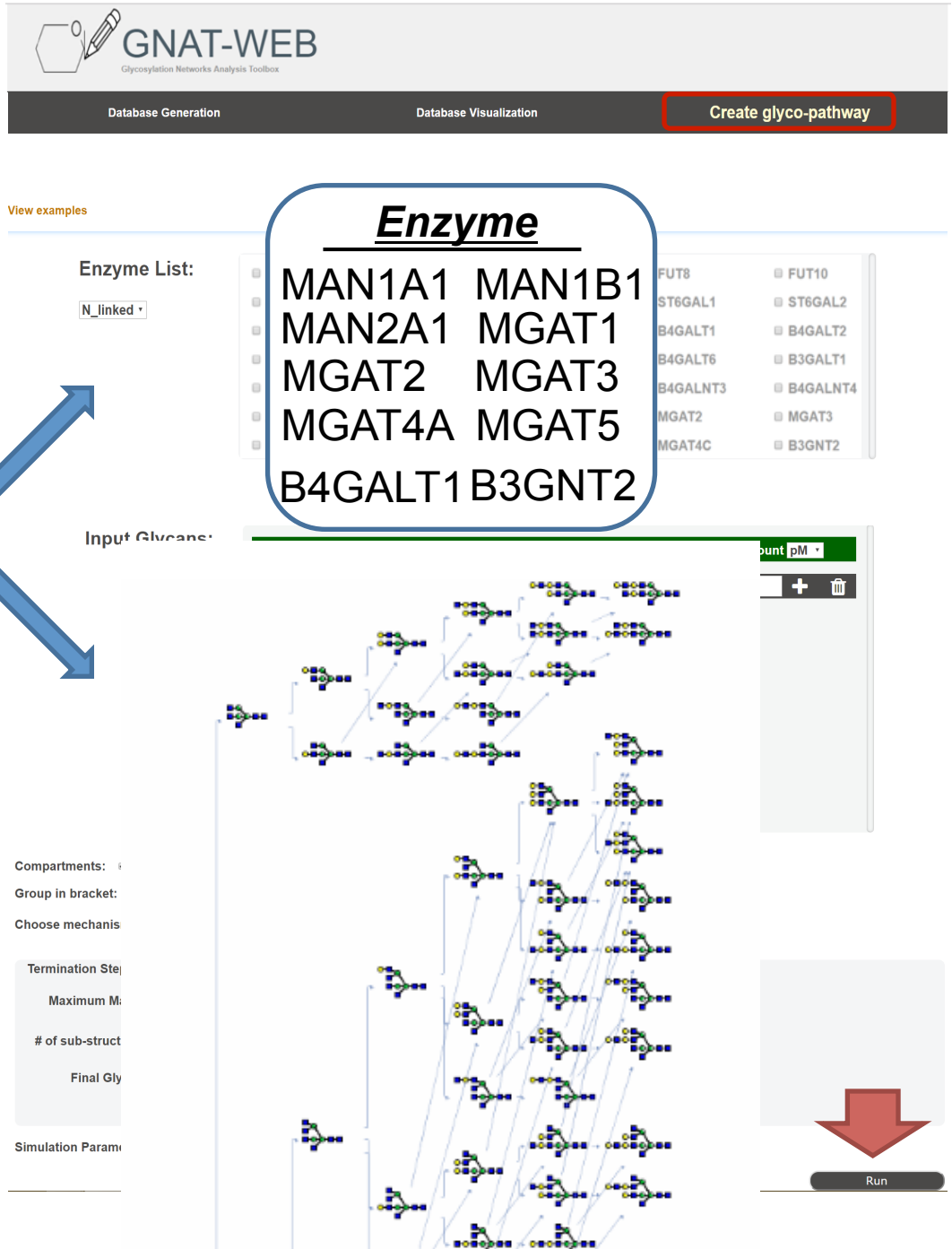


Ref: 3rd Ed. of Essentials of Glycobiology

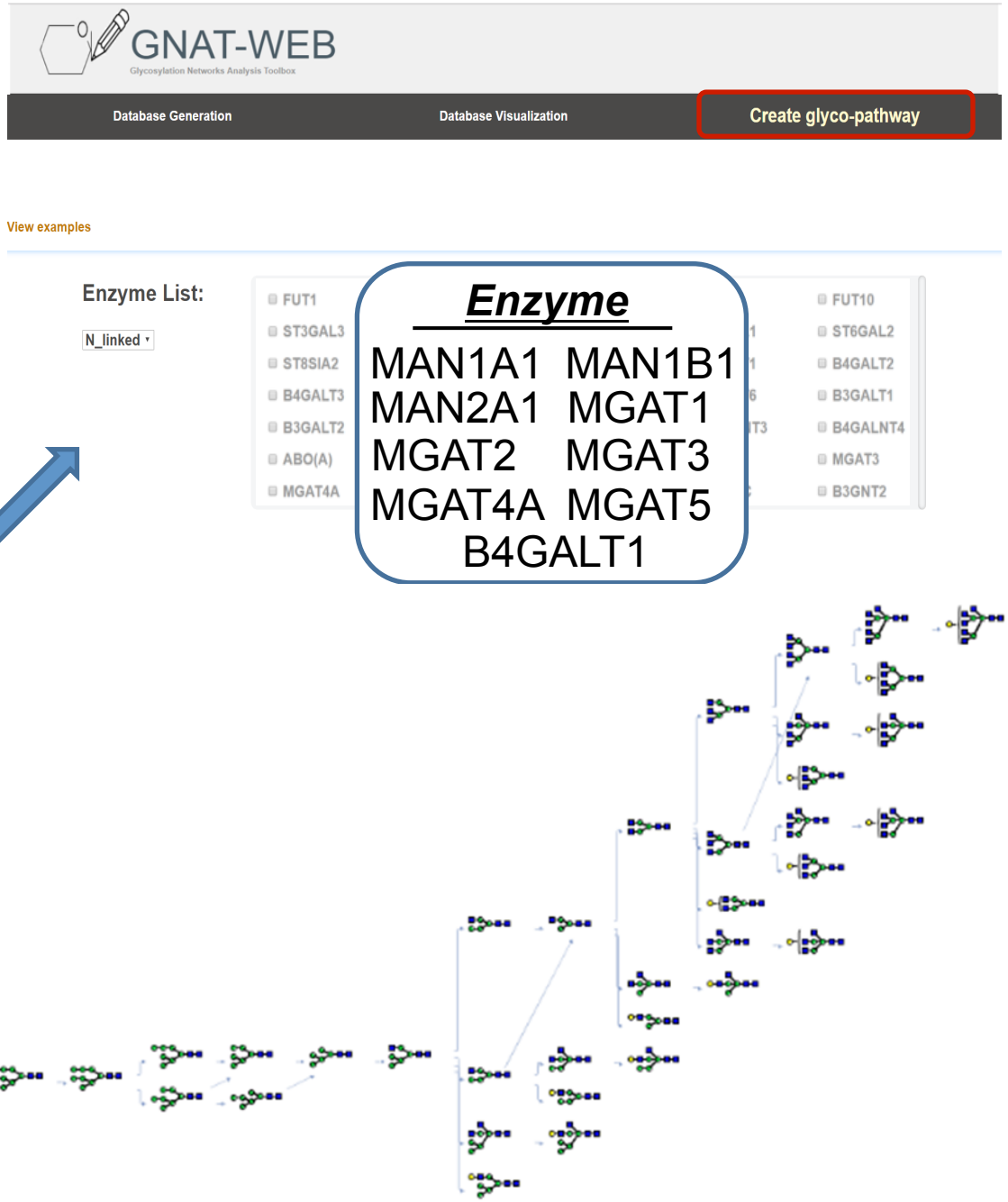
Create pathways : forward

Specify starting material and enzymes

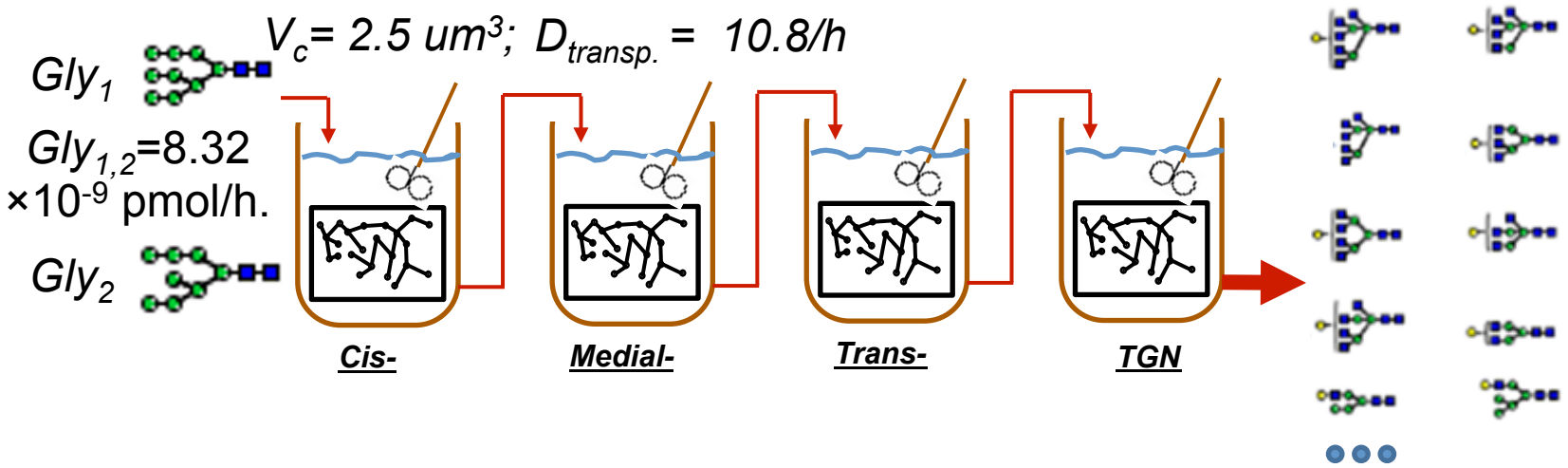
Constraints to limit network size



Create pathways :reverse



Pathway: 4 compartment CSTR



Species balance equation:

$$d[Gly_{i,j}]/dt = D_{transp.} \times [Gly_{i,j-1}] - V_{i,j} \times [Gly_{i,j}]/Vc \times Km_{i,j} \times (1+\sum_k [Gly_{k,j}]/(Km_{k,j})) - D_{transp.} \times [Gly_{i,j}]$$

0 (CSTR) — in — Reaction — out

i=36 glycans
j=4 compartment

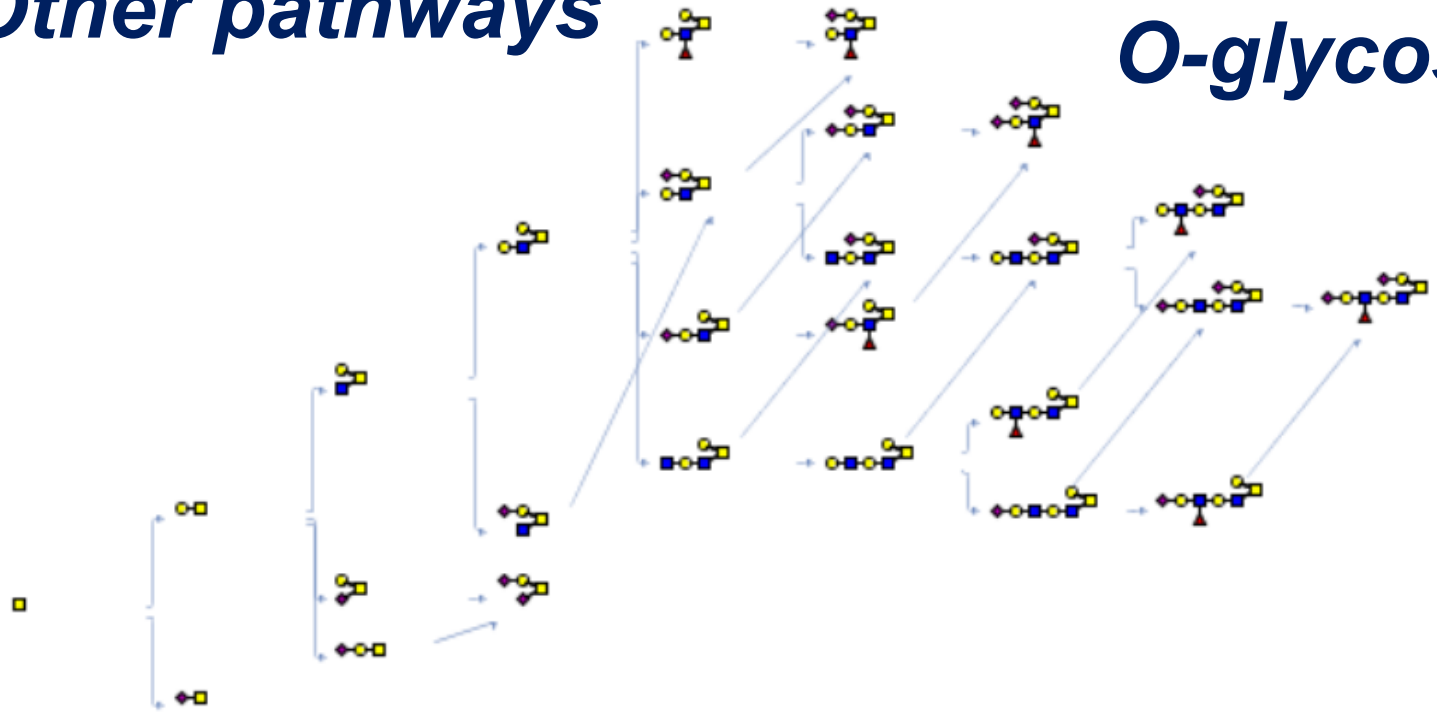
In silico simulation

- ***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



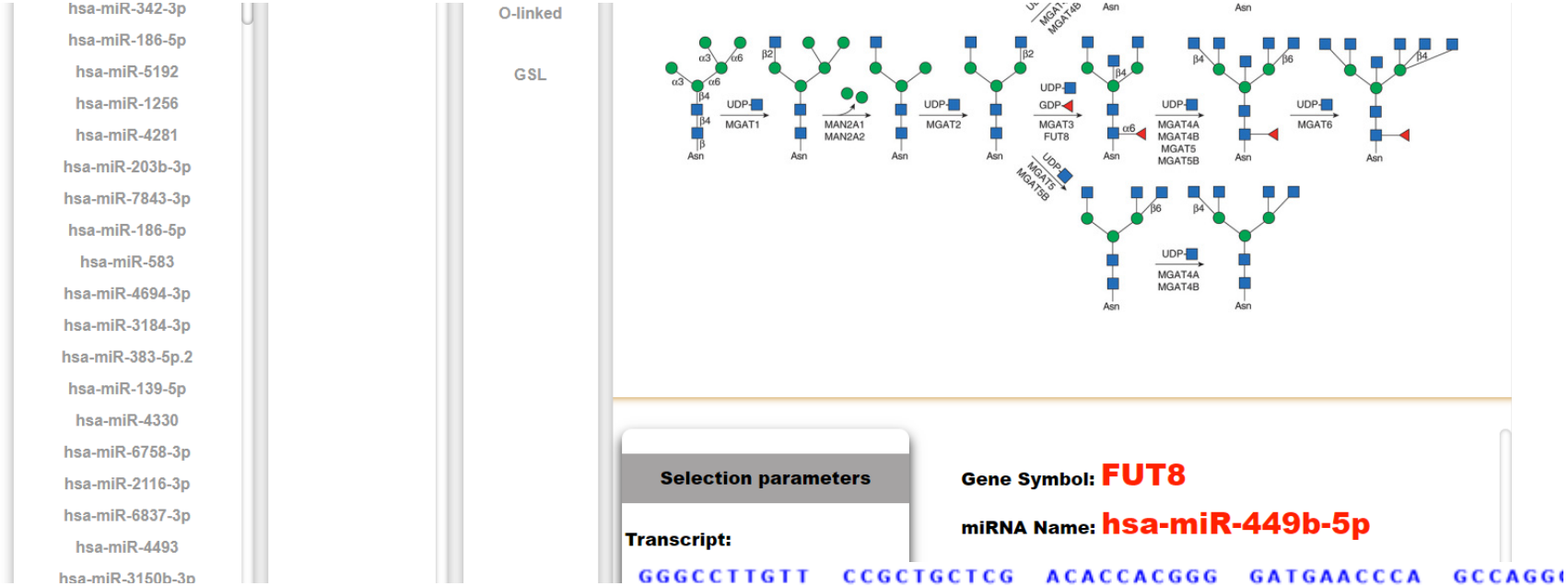
Glycolipid biosynth.

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 contruct pathway maps.**
- **GlycoMir: The glycogene microRNA targets**
- ***Functional integration with other databases***



* Collaboration with Lara Mahal (NYU)

Kai Cheng

2. DrawGlycan-SNFG

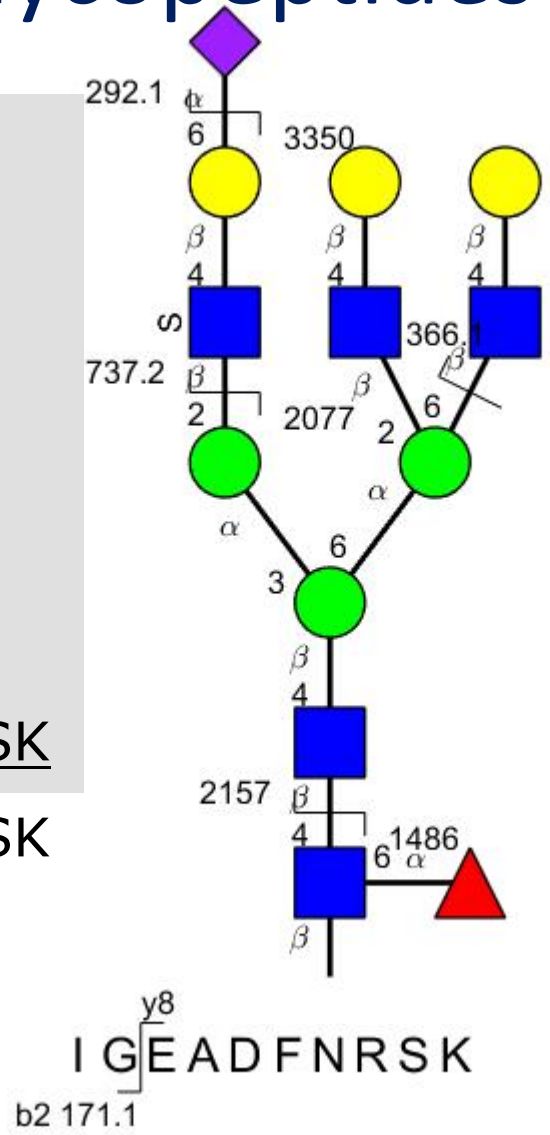
- From IUPAC to Symbolic Nomenclature for Glycans (SNFG)
- Draw glycopeptides
- Draw glycan and peptide fragmentation
- *Other features…*

Glycobiology. 2017 Mar 15;27(3):200-205

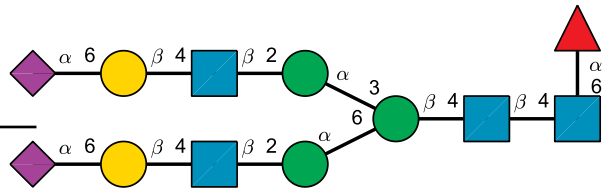
DrawGlycan-SNFG:
Render glycans and glycopeptides

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

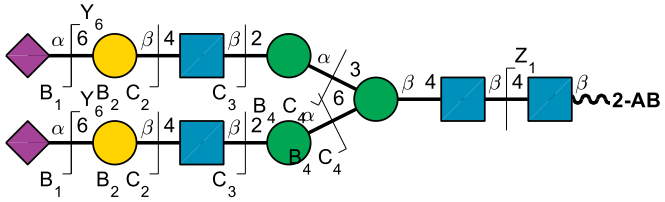
Advantages: straightforward, easy to read & write, adequate for common use

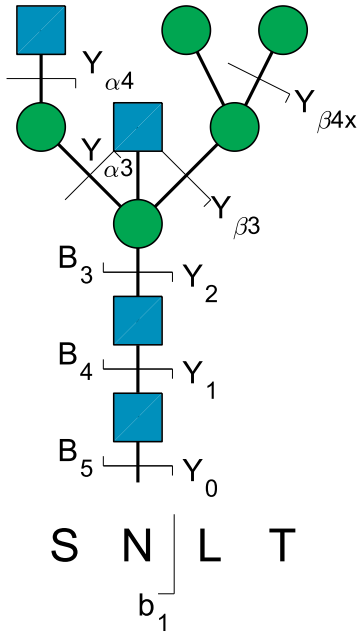


Fragmentation options

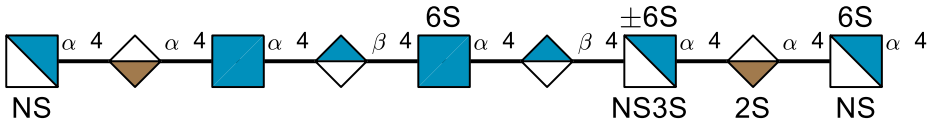


	-Option	Representation
1	-R	Glycan reducing end
2	-NR	Glycan non-reducing end
3	-N	Peptide backbone N-terminus
4	-C	Peptide backbone C-terminus

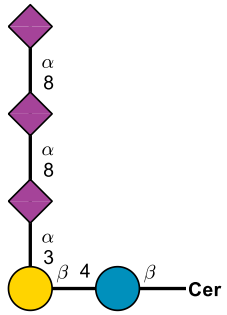


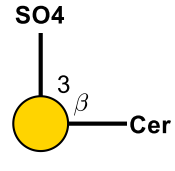


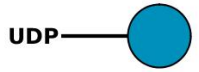
Monosaccharide options

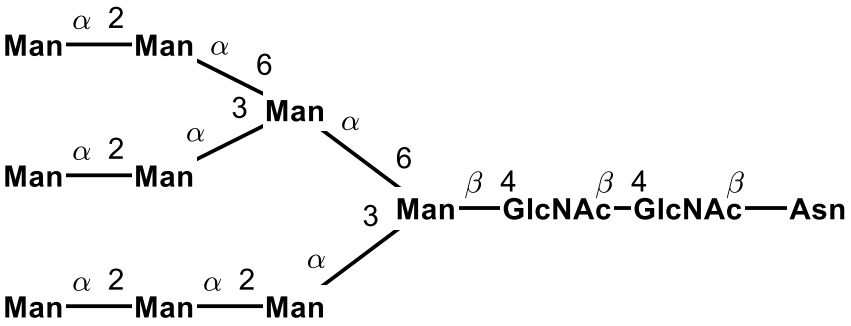


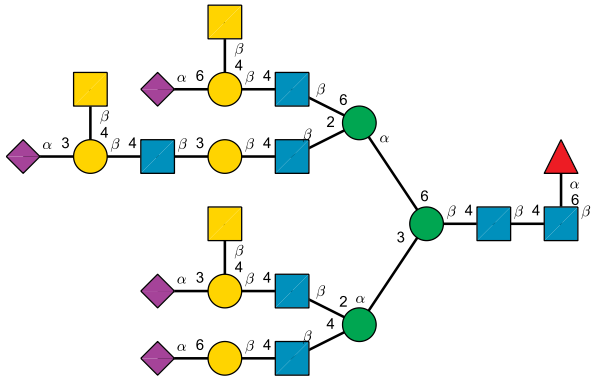
	-Option	Representation
1	-U	Annotate above monosac.
2	-D	Annotate below monosac.
3	-P	Identify a perpendicular monosac.
4	-CHAR	Introduce arbitrary text or present monosac. in text form





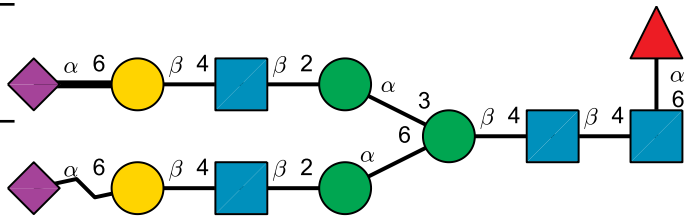


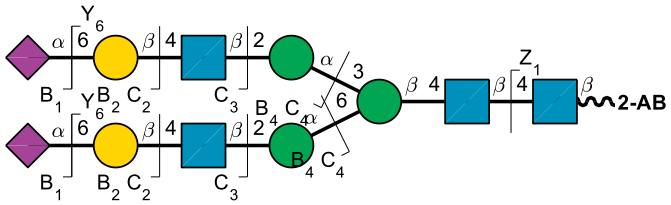


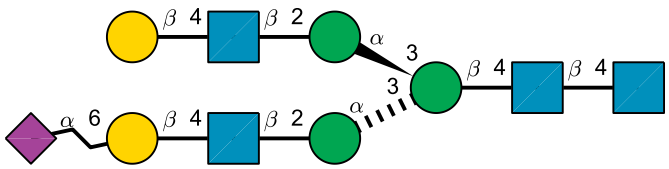


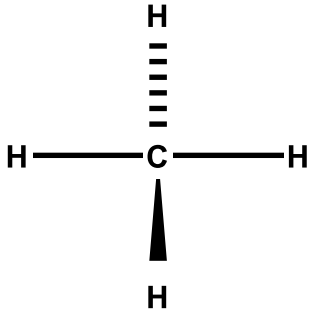
Bond options

	-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

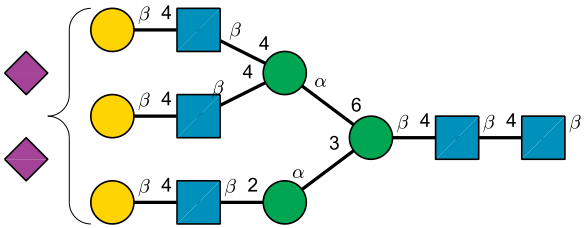


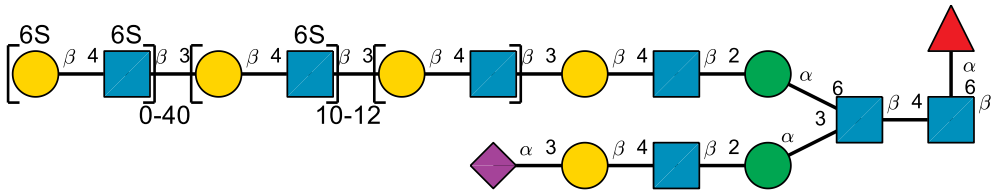




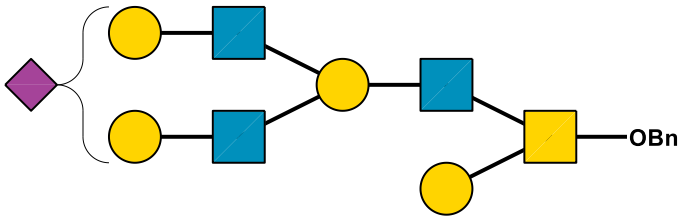


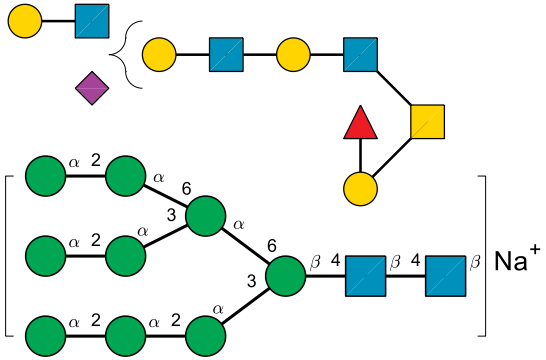
Repeats, adducts and fuzzy options

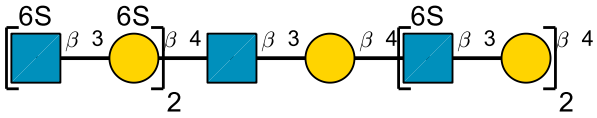


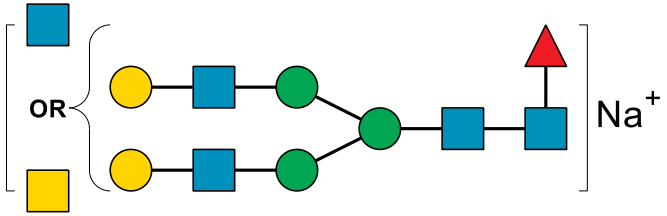


	-Option	Representation
1	-RS	Repeating unit start
2	-RE	Repeating unit ends
3	-ADDUCT	Add glycan adduct
4	-CURLY	Ambiguous assignments/fuzzy structures









DrawGlycan: Web, GUI & Command-line version

VirtualGlycome.org/DrawGlycan

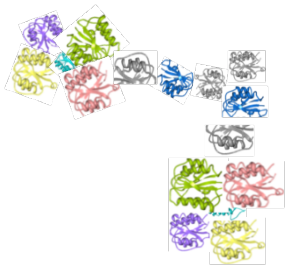
3. GlycoPAT

Kai Cheng

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

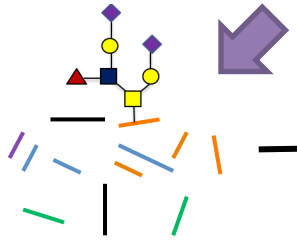
*Mol Cell Proteomics.*16:2032-47, Nov 2017.

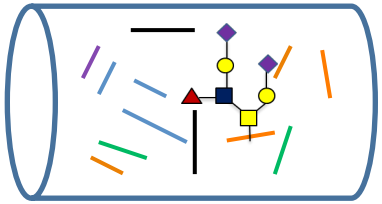
GlycoPAT: High-throughput glycoproteomics analysis



Protease Digestion

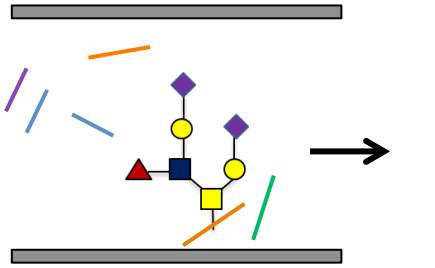
(Trypsin, Glu-C etc.)





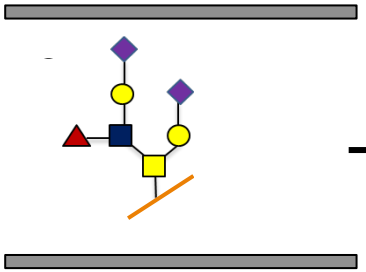
nanoLC

Tandem mass spectrometry in Orbitrap



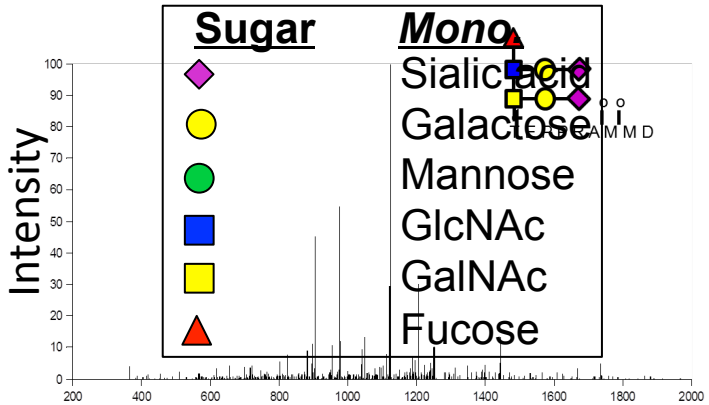
MS^1

Separation based on precursor m/z

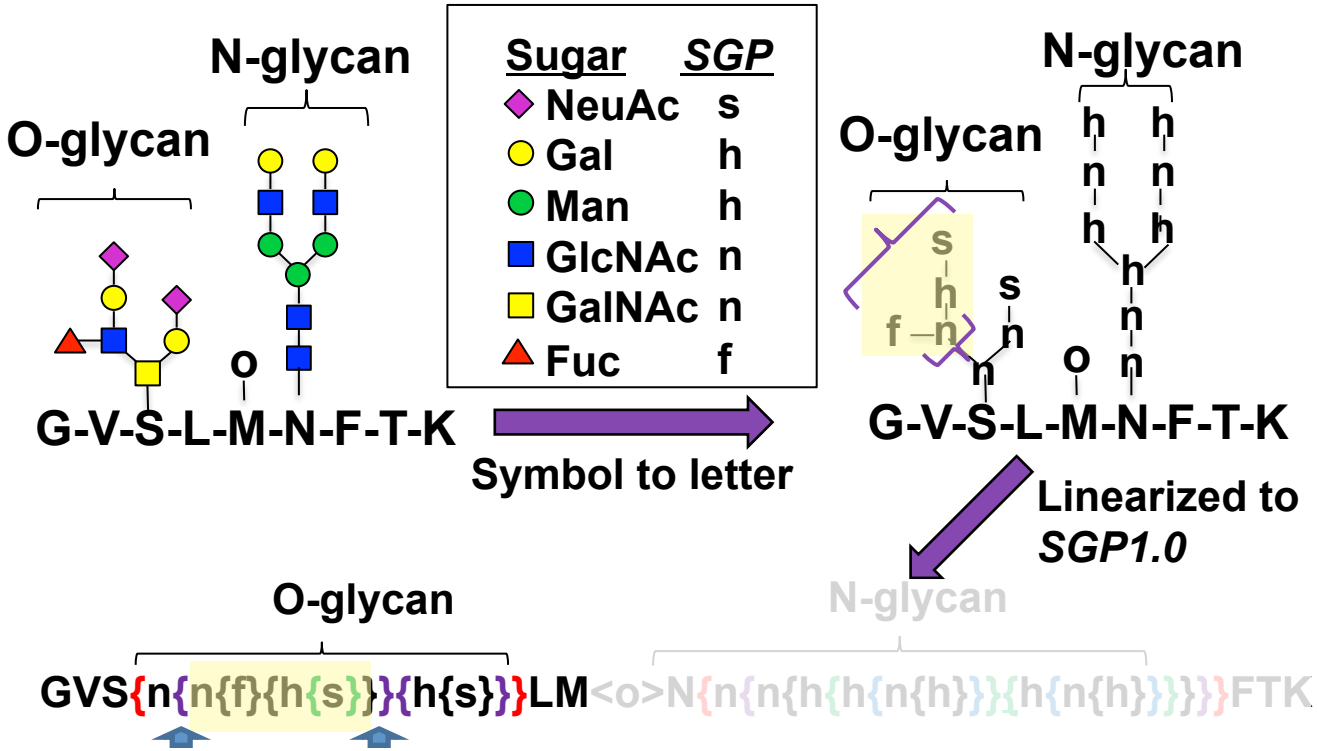


MS^2

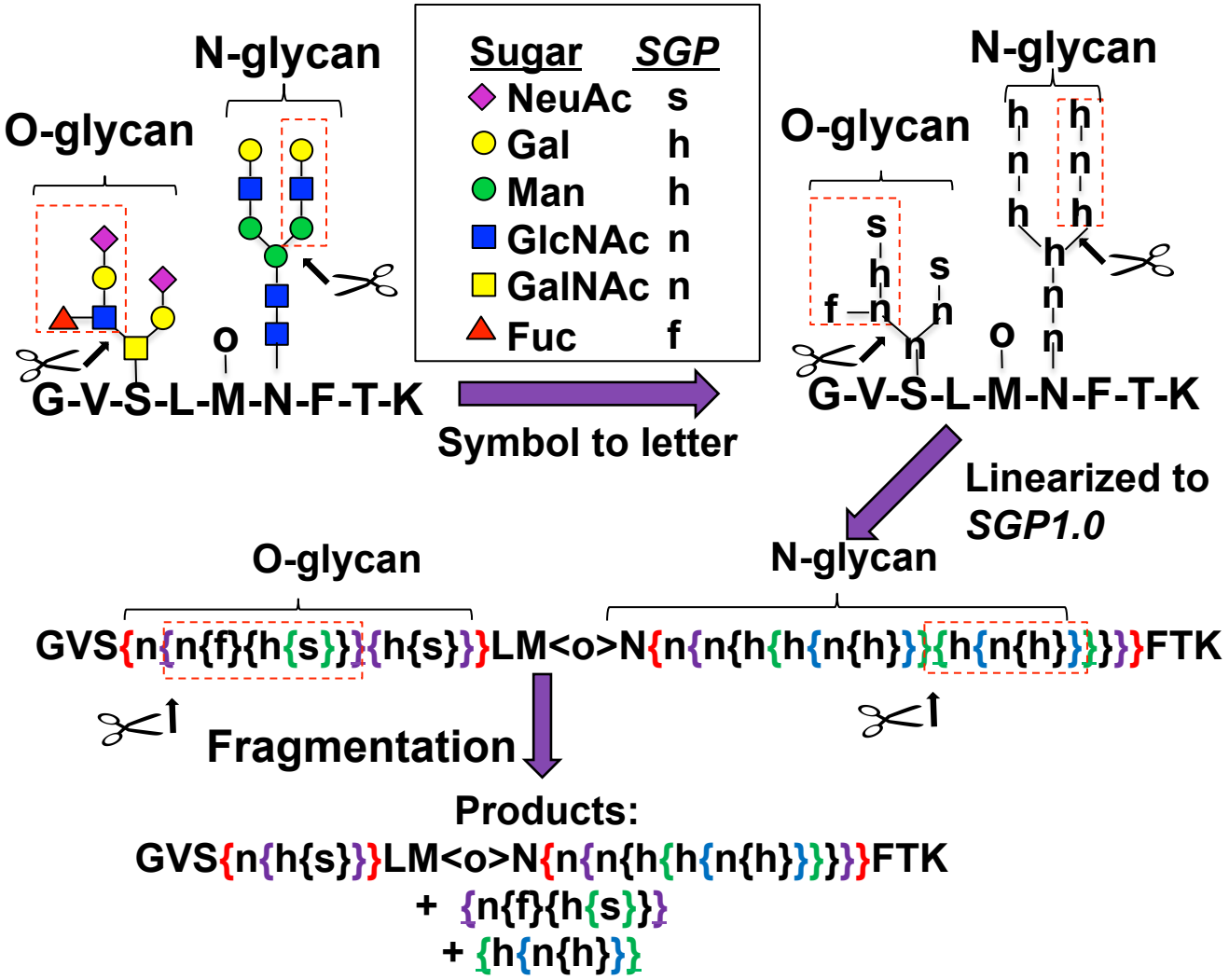
fragmentation



SmallGlyPep : The *minimal* representation of glycopeptide for MS

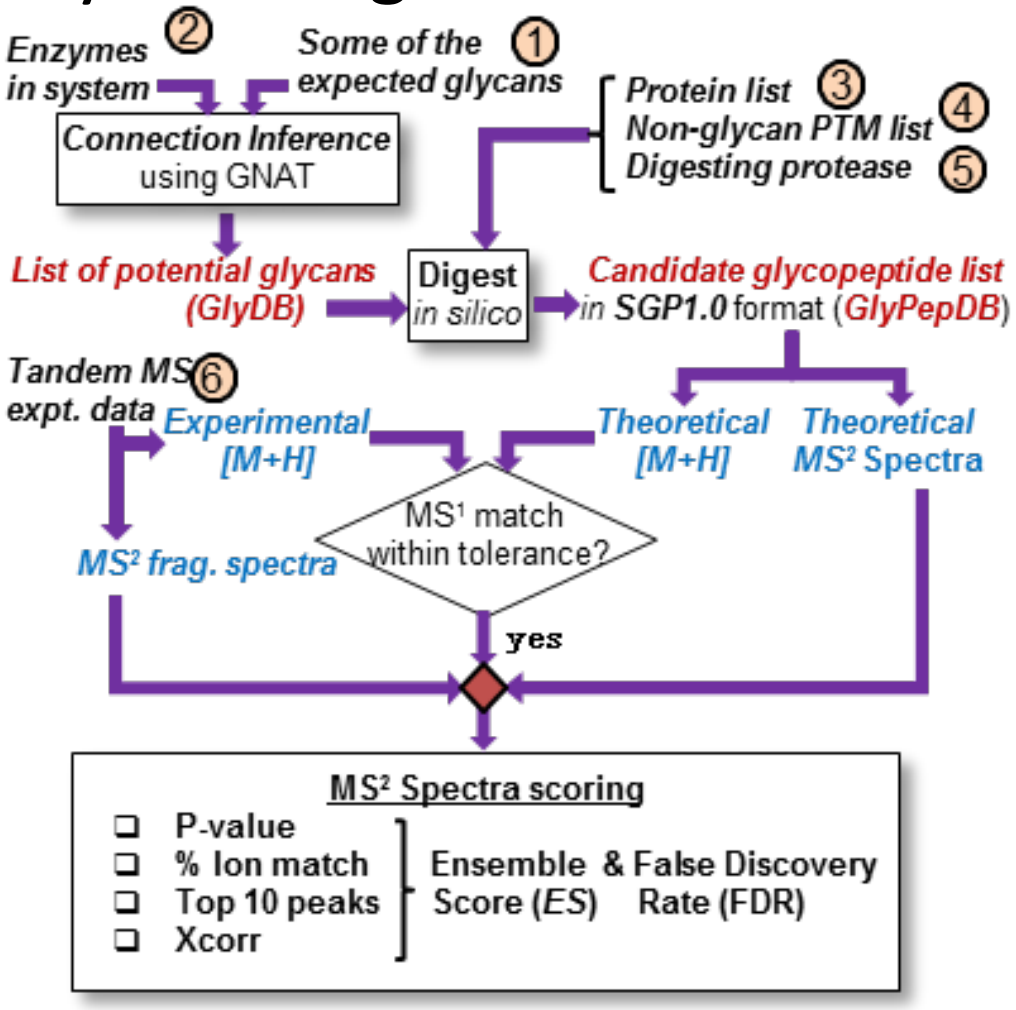


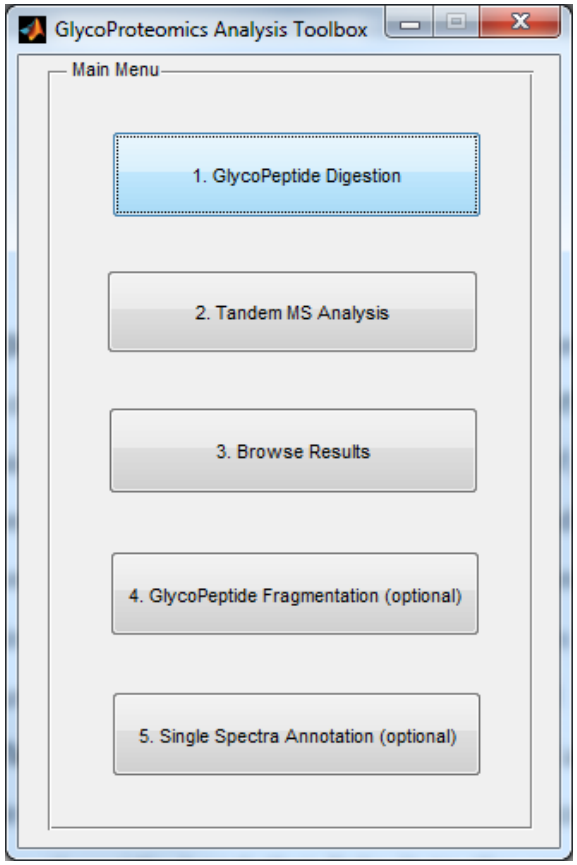
SmallGlyPep : The *minimal* representation of glycopeptide for MS



Methods

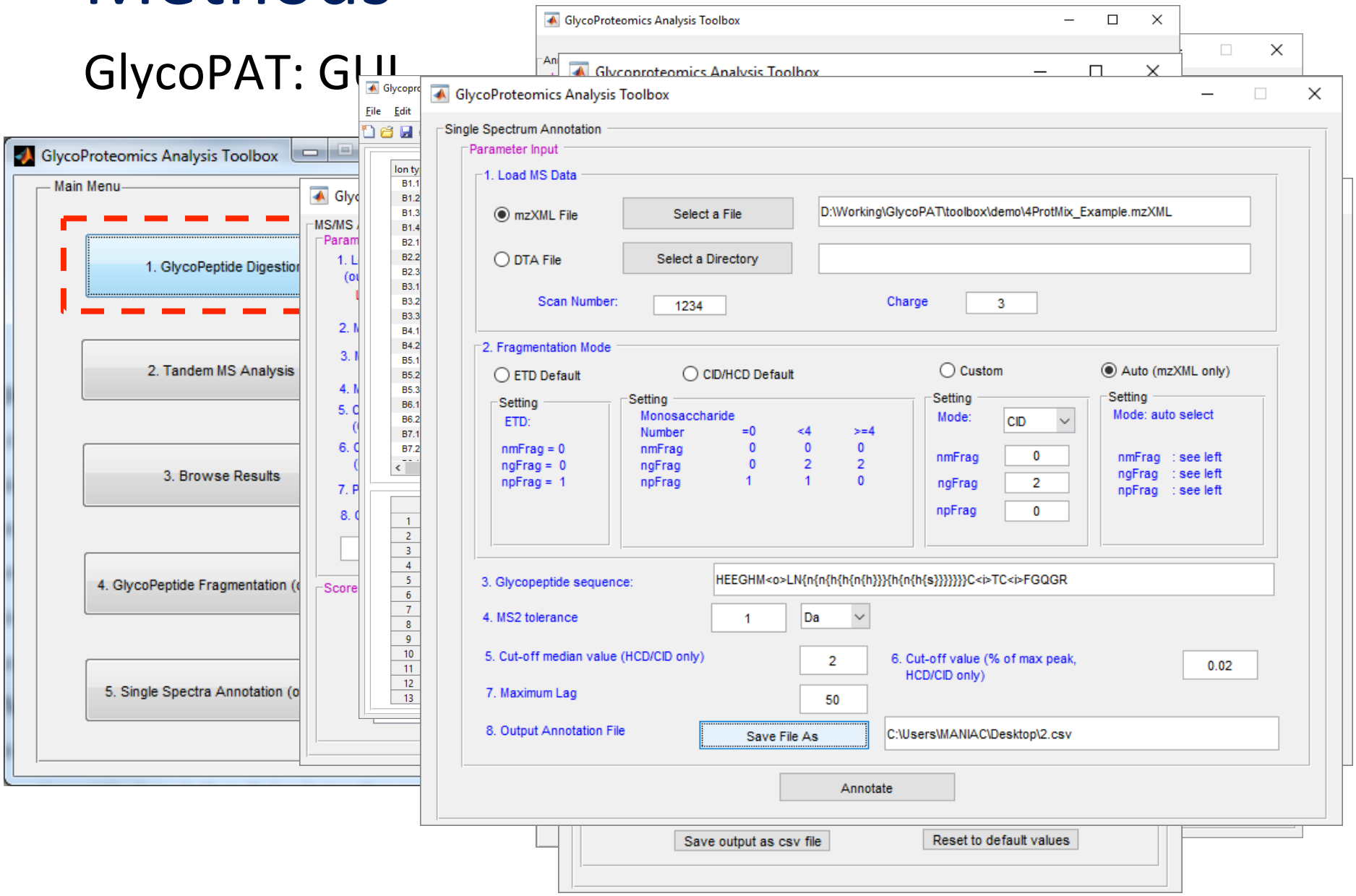
GlycoPAT: general workflow



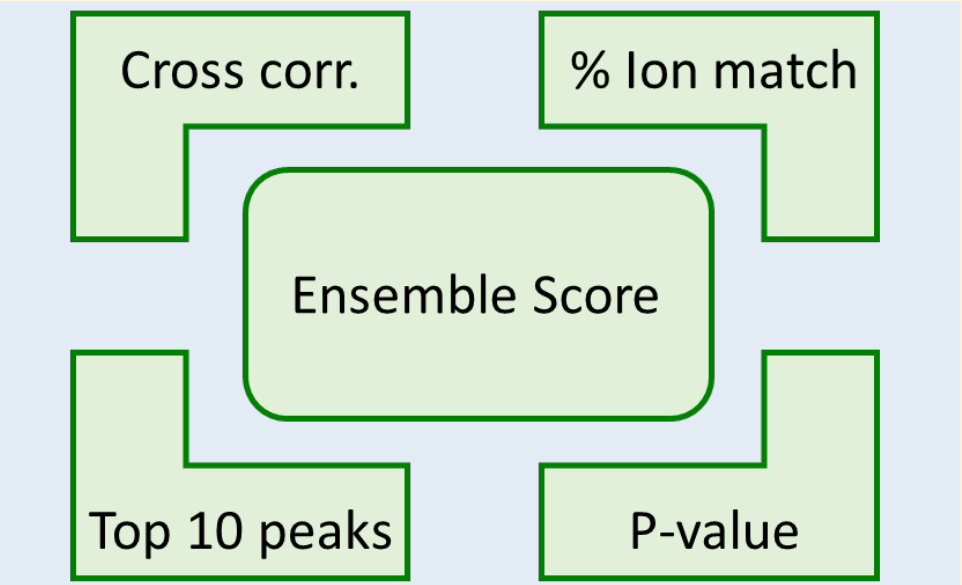


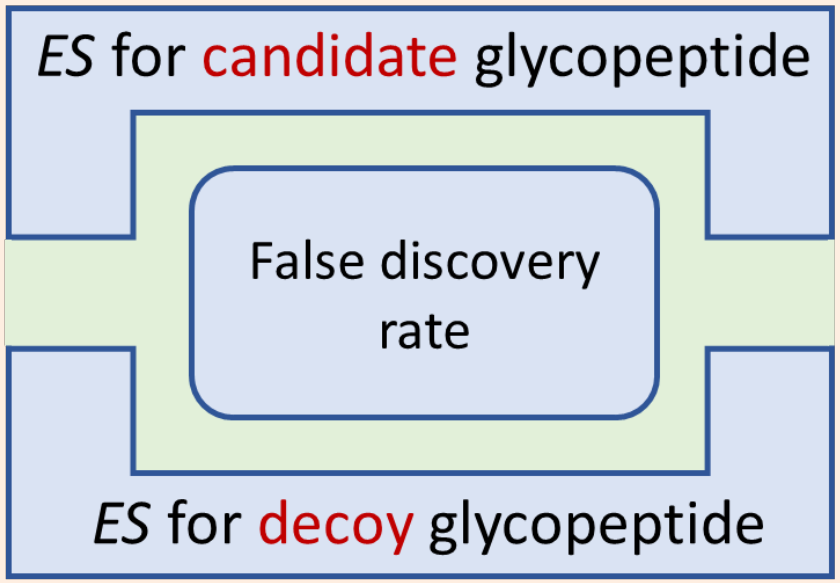
Methods

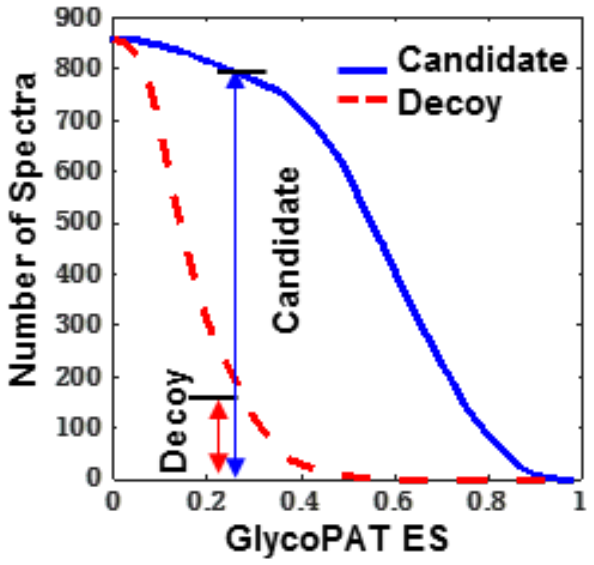
GlycoPAT: GUI

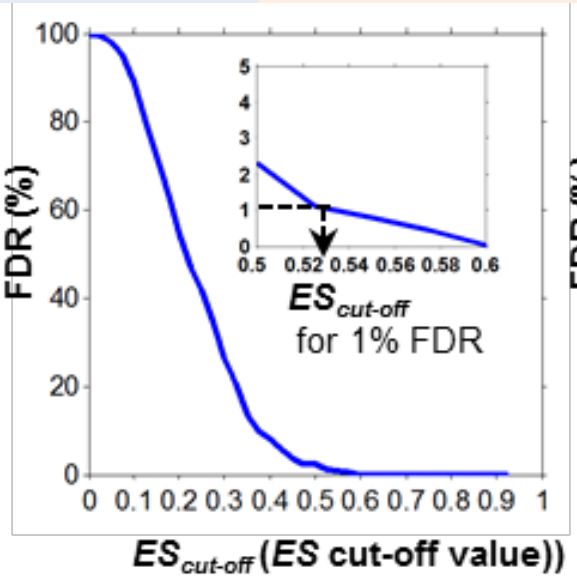


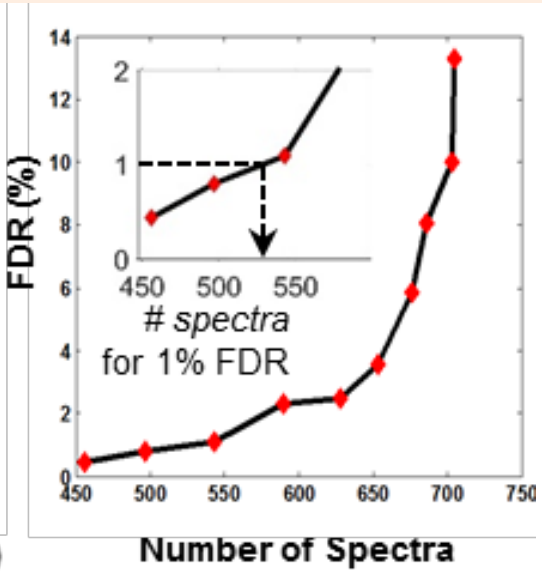
GlycoPAT: ensemble score (*ES*) and false discovery rate (*FDR*)









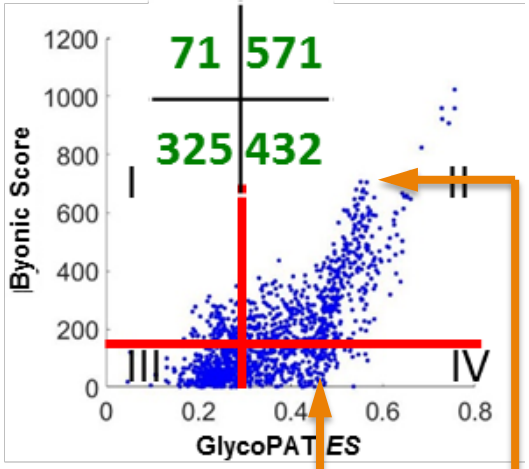


GlycoPAT: scoring of HCD spectra

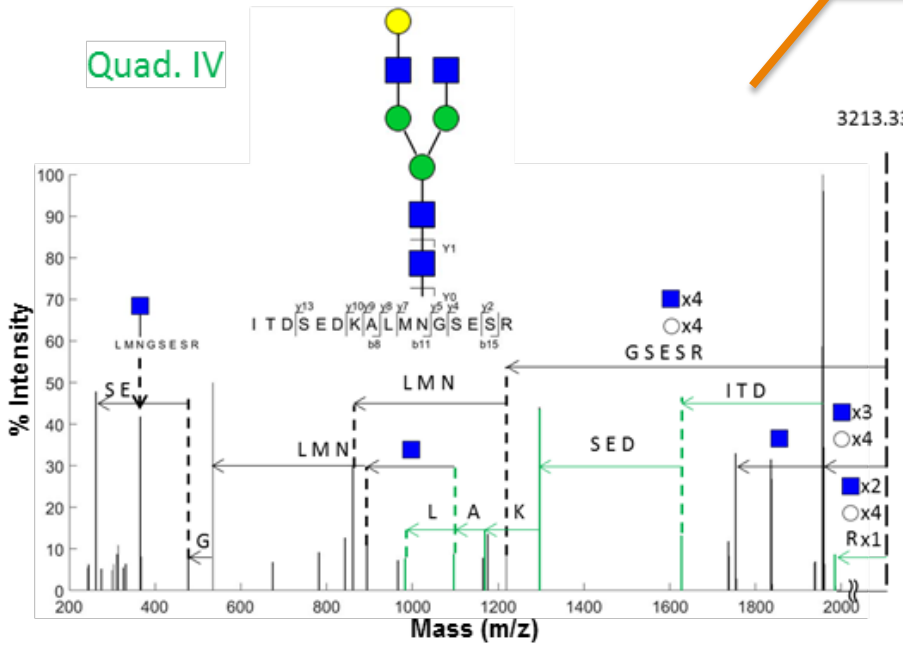
Basigen|HCD

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

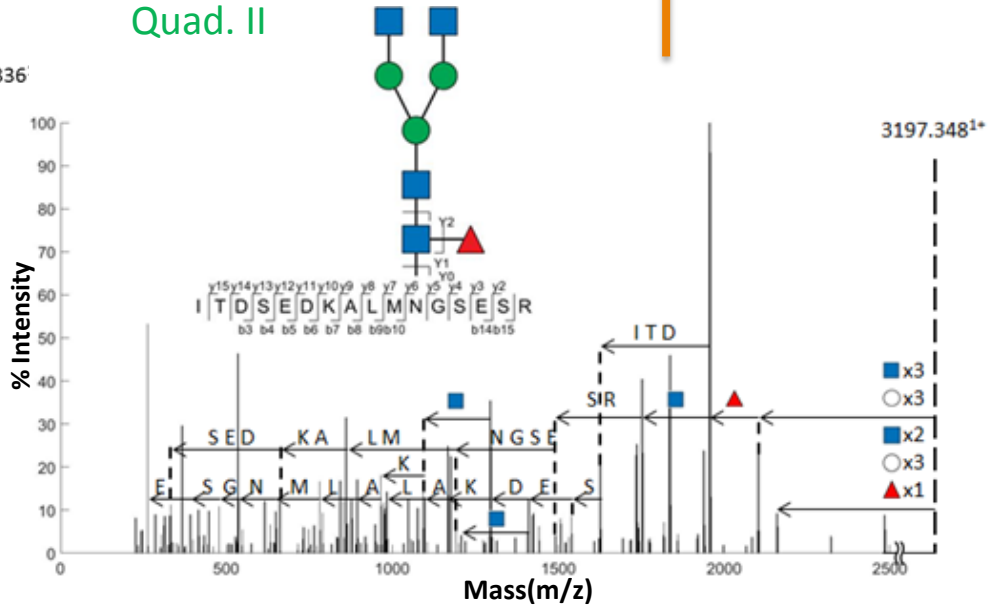
** Simultaneous breakage of peptide backbone and glycan structures*



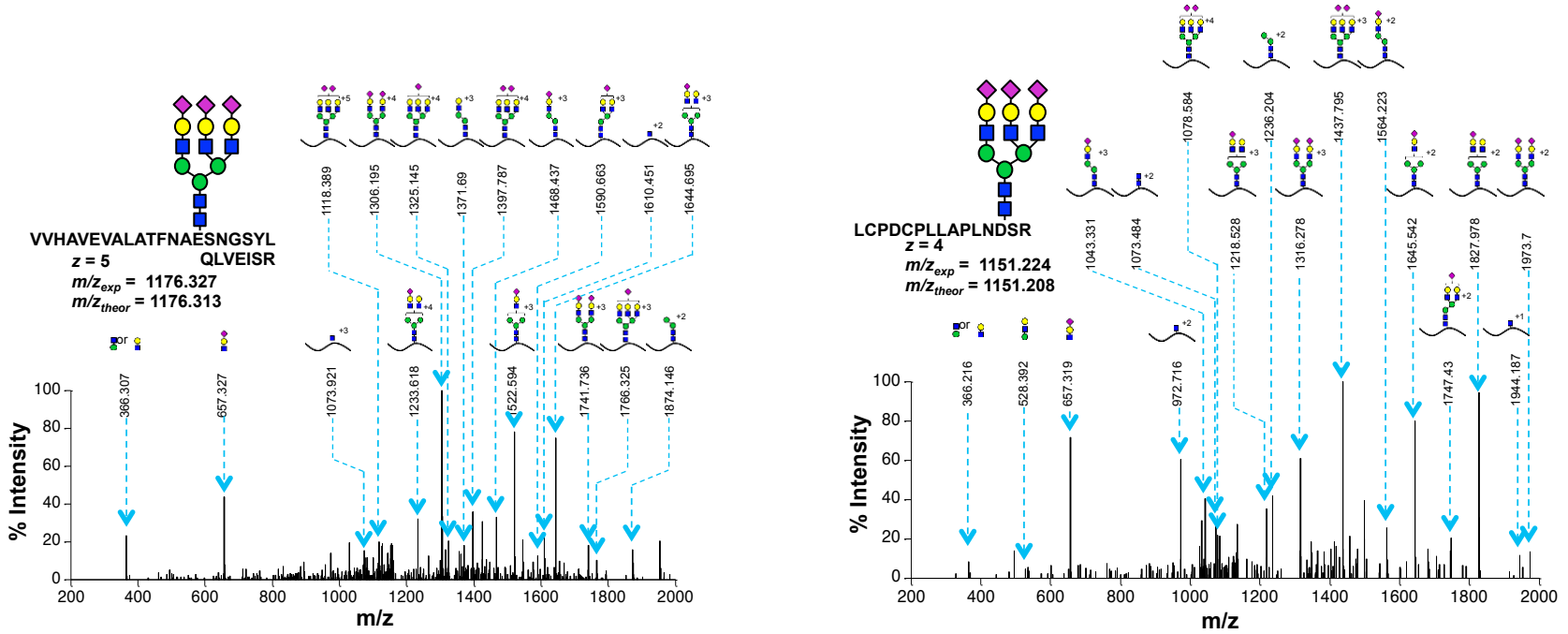
Quad. IV



Quad. II

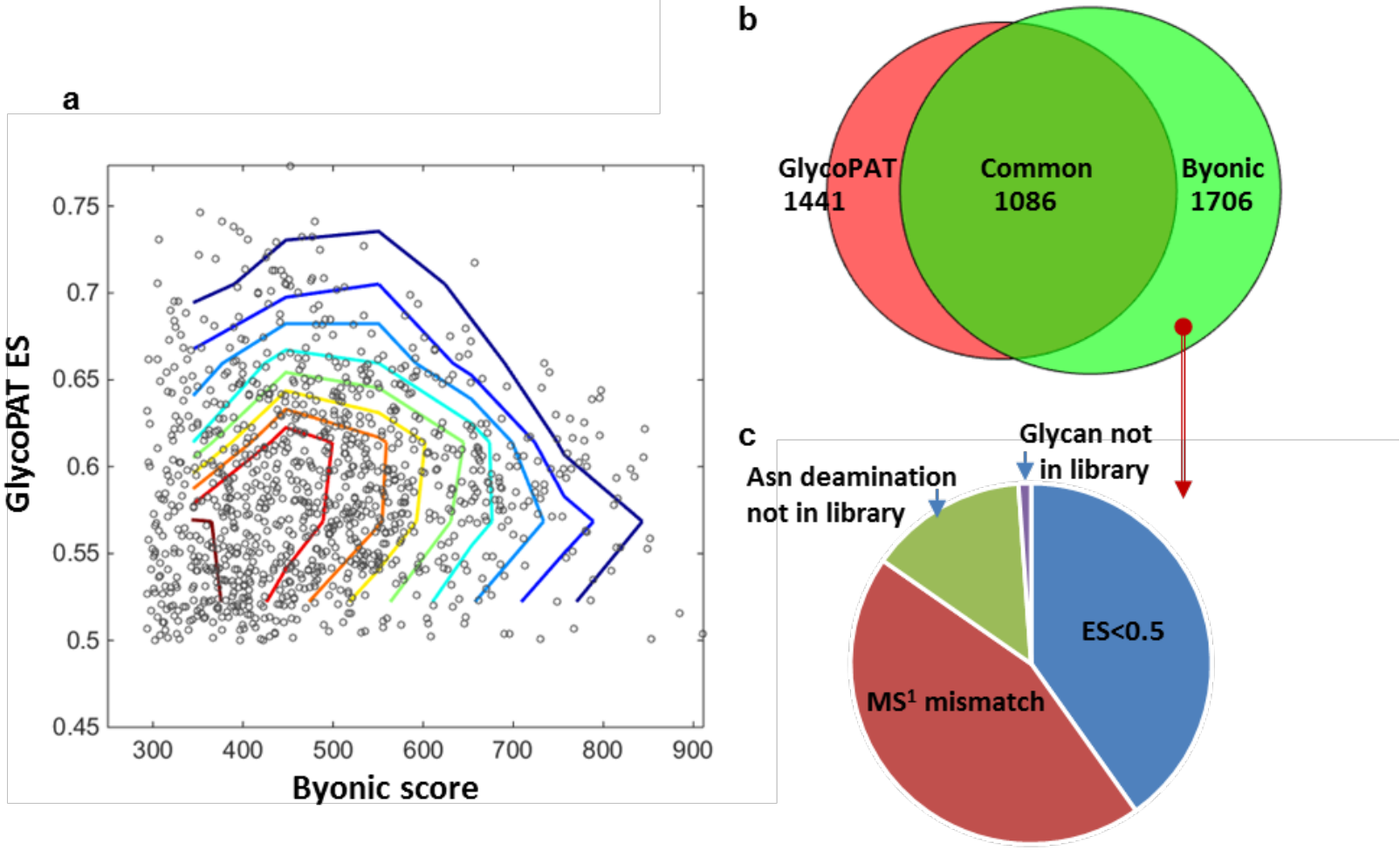


GlycoPAT: scoring of CID spectra



** Analysis of ladder-like breakdown of glycans*

Analysis of glycopeptides in whole prostate cancer lysates



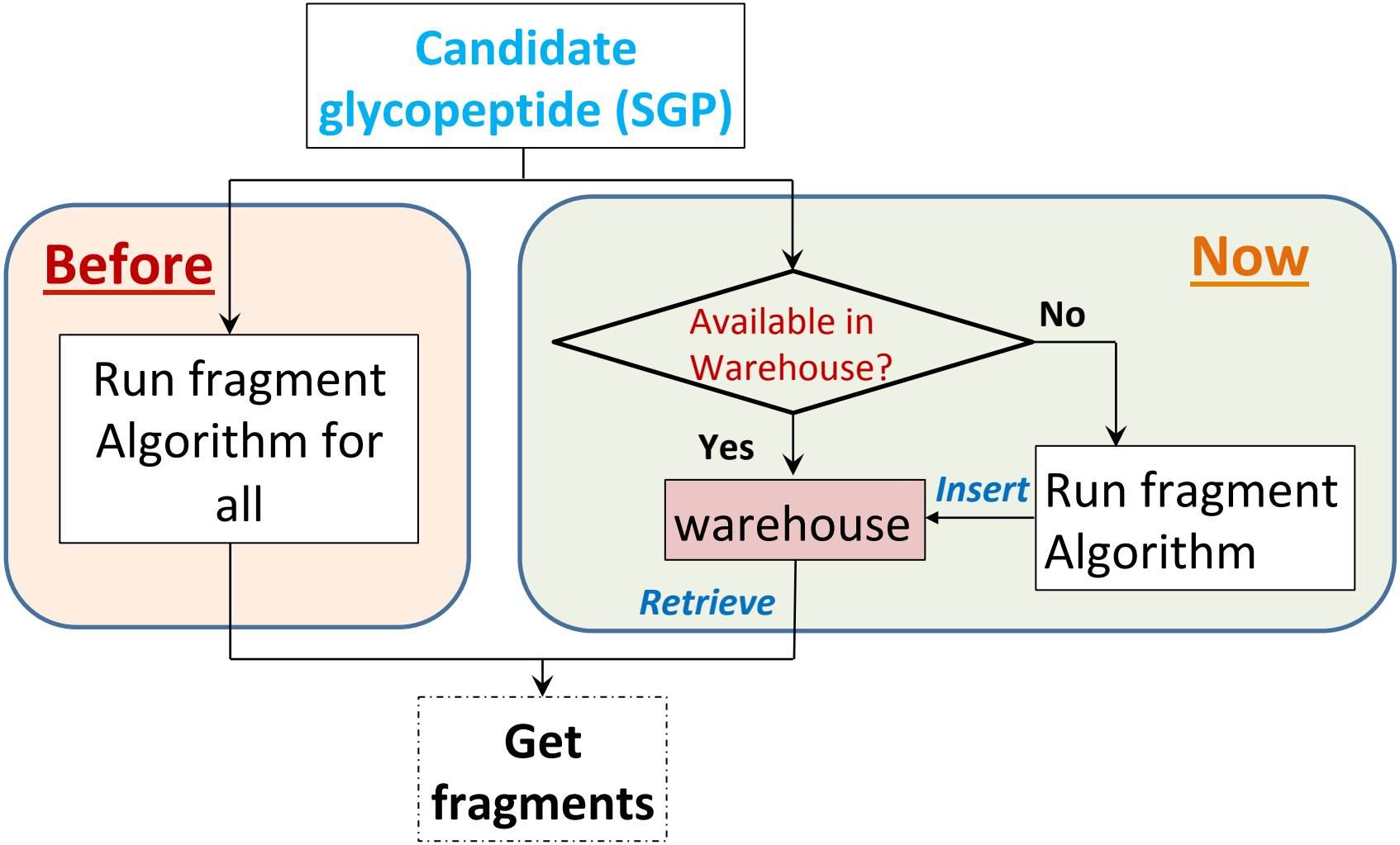
Mol Cell Proteomics. 2015 Oct;14(10):2753-63

Work in progress

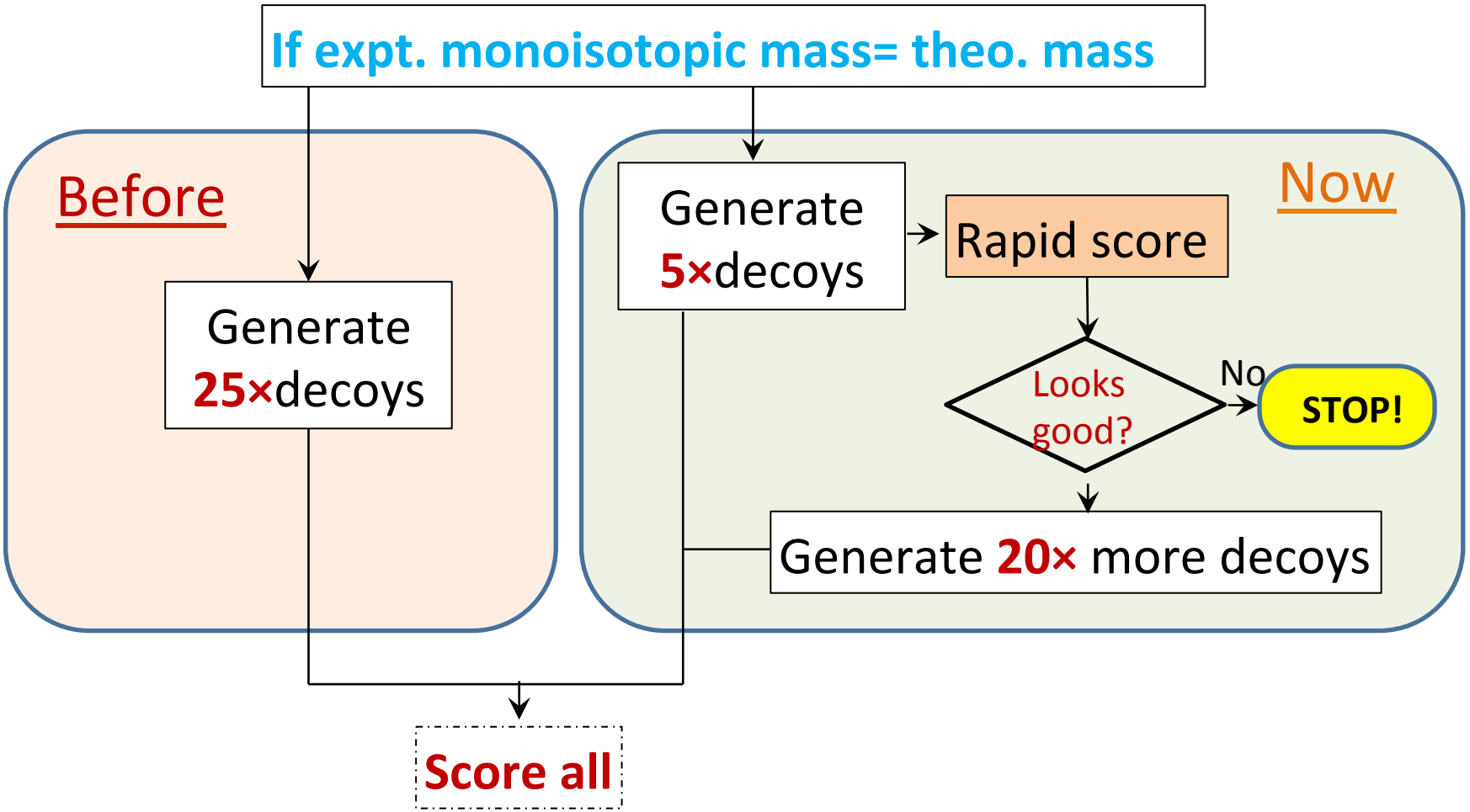
- Improving calculation speed:

  FOCUS!! Don’t do everything for everyone!

Improve speed: *The fragment warehouse*



Improved speed: *Selective scoring*



Work in progress

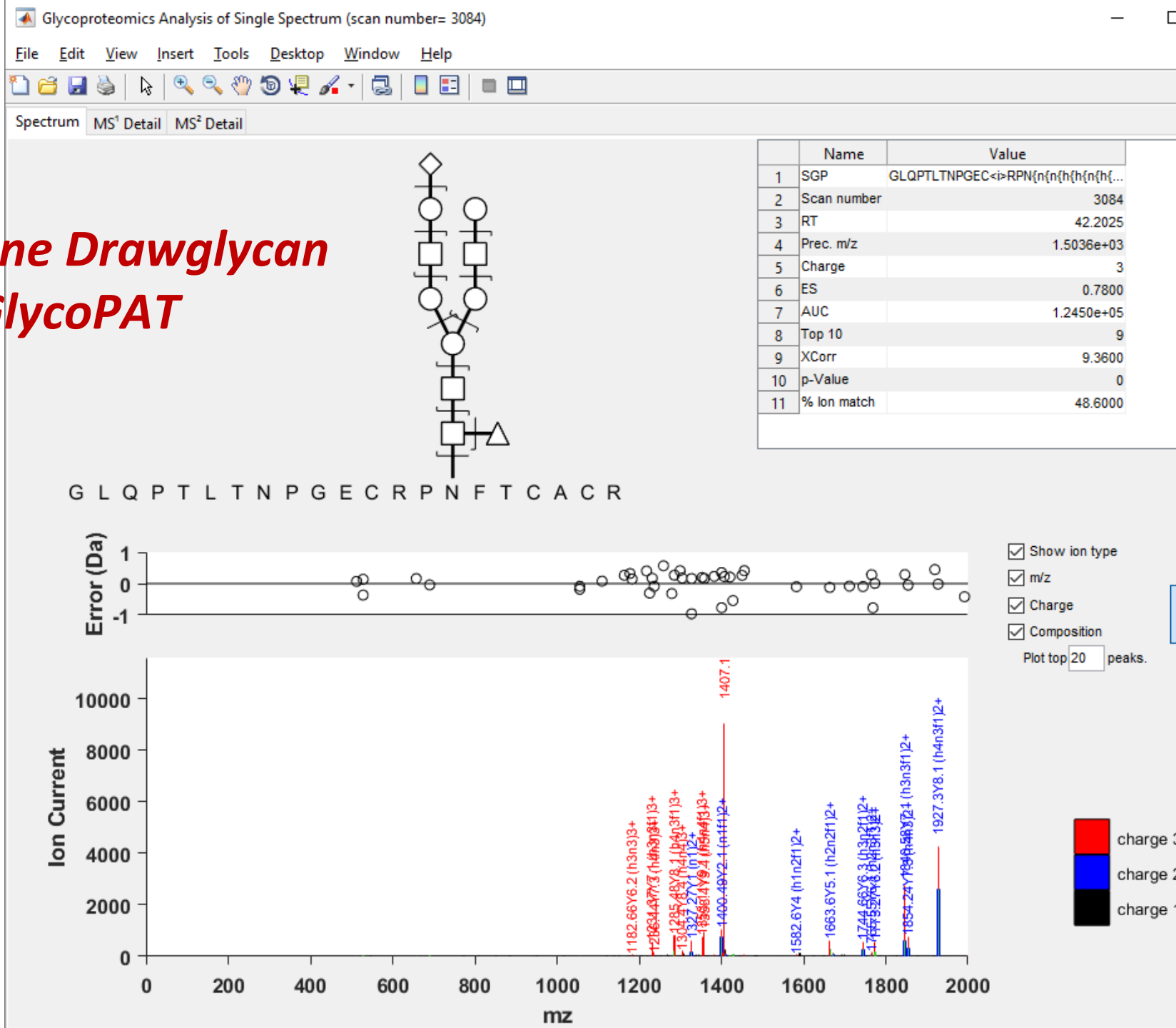
- Improving calculation speed:

  FOCUS!! Don't do everything for everyone!

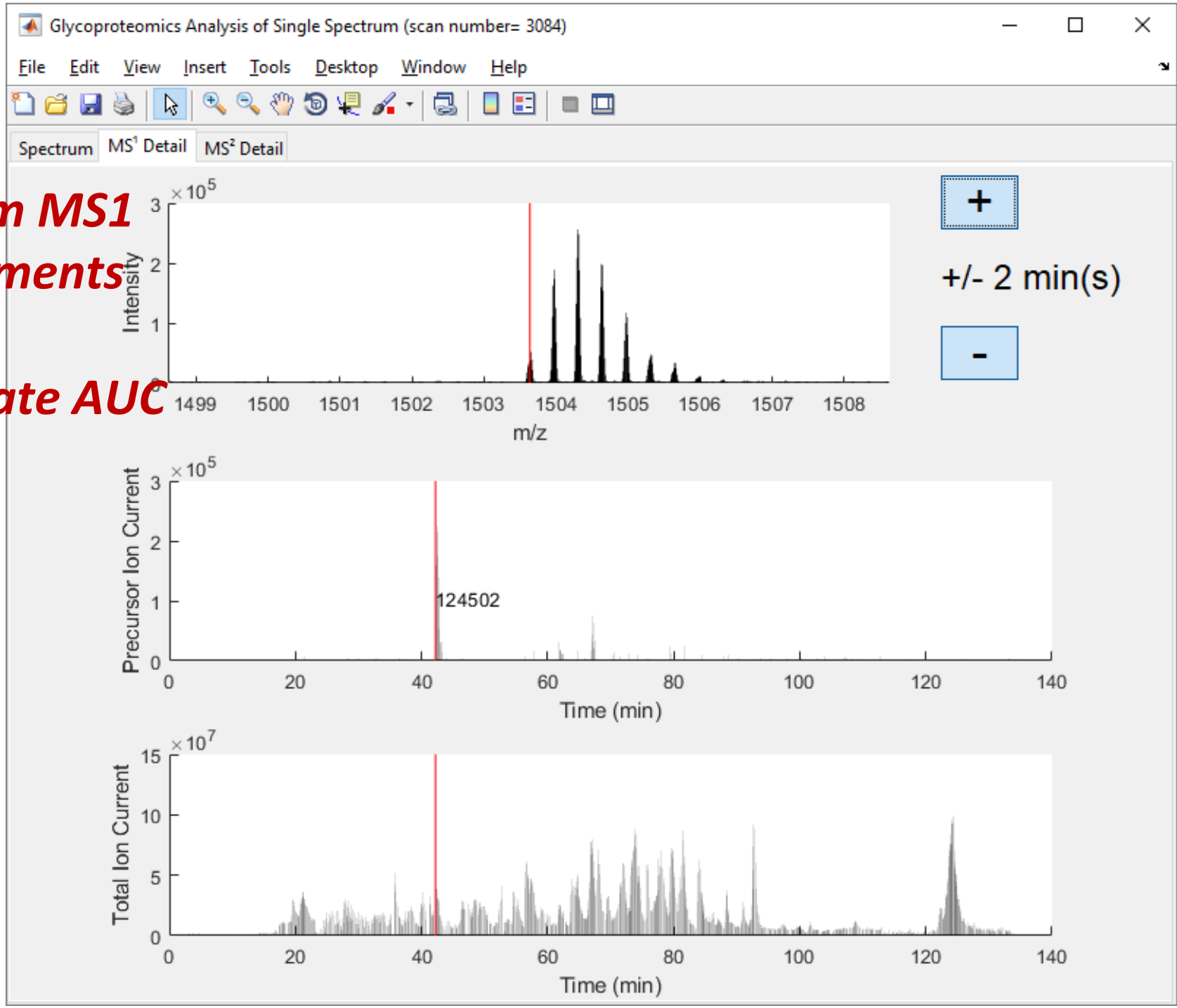
- Streamlining result visualization:

  SIMPLIFY!! Tell the story in pictures!

Combine Drawglycan with GlycoPAT



Confirm MS1 assignments & calculate AUC

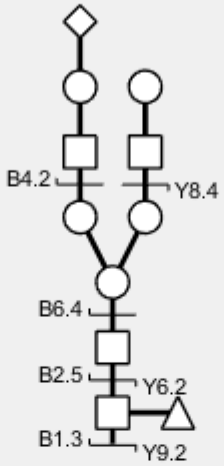


Provide more details

Glycoproteomics Analysis of Single Spectrum (scan number= 3079)

File Edit View Insert Tools Desktop Window Help

Spectrum | MS¹ Detail | MS² Detail



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/ngFrag
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	GLQPTLTNPGEC<i>RPNFTC<i>AC<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{h{n{h{s}}}}}}{f}}	3	512.19	1745.53	0.77	1947	B8.3	0 0 2
9	GLQPTLTNPGEC<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**

NIH **NIGMS**: General Medicine

NYSTEM
NEW YORK STATE STEM CELL SCIENCE

American Heart Association®