GNOme – A glycan naming and subsumption ontology

Nathan Edwards

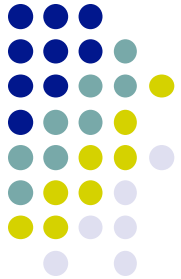
Department of Biochemistry and Molecular & Cellular Biology

Georgetown University Medical Center

About me…

- Celera Genomics + Applied Biosystems
  - Proteomics pipelines, peptide identification software, informatics of new MS technologies
- NCI Clinical Proteomics Tumor Analysis Consortium (CPTAC Phases I,II,III):
  - Novel peptides using genomic/transcripts data
  - Data Coordinating Center (DCC/Portal)
  - Common Data Analysis Pipeline (AWS/Galaxy)
- Intact glycopeptide MS/MS / SWATH tools
  - w/ Rado Goldman (also Georgetown)

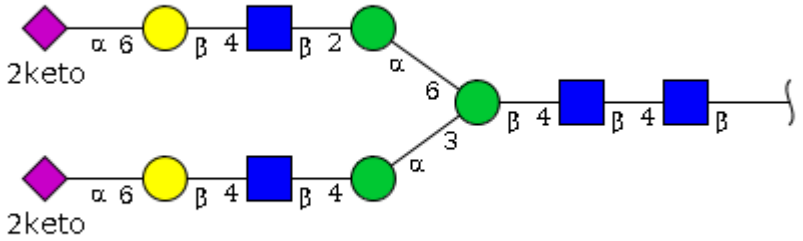
Glycans are complex molecules



- Difficult to completely characterize
  - Mass spectrometry, etc.
- Partial characterization is common
  - Prior (biological) knowledge can fill in some gaps
  - Sample manipulation (glycosidases, etc.), too…
- Difficult to describe extent of characterization
  - Complex “sequence” formats precisely describe known and missing details
- Accessions provide a stable identifier
  - Link to explicit sequence formats

Glycan Structure Descriptions

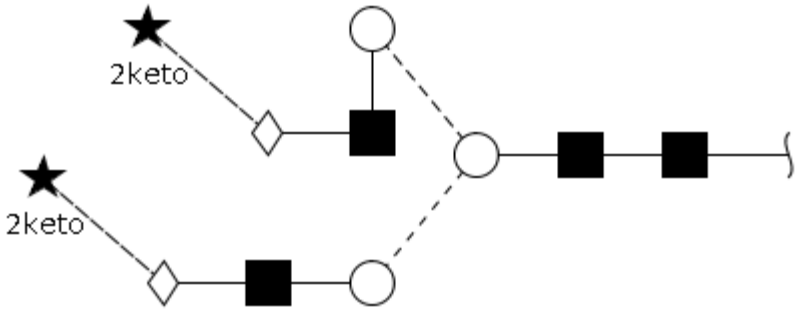
GlyTouCan: G39764AC
PubChem: SID252281000
GlycomeDb: 9088



Neu5Ac(a2-6)Gal(b1-4)GlcNAc(b1-4)Man(a1-3)[Neu5Ac(a2-6)Gal(b1-4)GlcNAc(b1-2)Man(a1-6)]Man(b1-4)GlcNAc(b1-4)GlcNAc(b1-

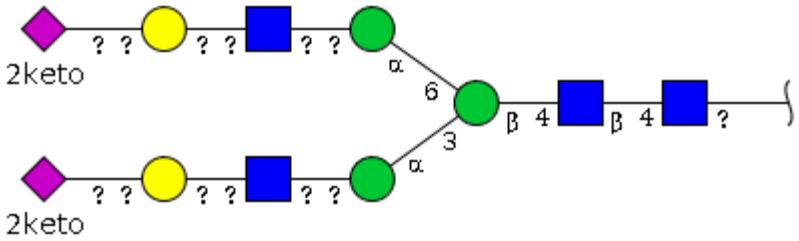
WURCS=2.0/5,11,10/[a2122h-1b_1-5_2*NCC/3=O][a1122h-1b_1-5][a1122h-1a_1-5][a2112h-1b_1-5][Aad21122h-2a_2-6_5*NCC/3=O]/
1-1-2-3-1-4-5-3-1-4-5/a4-b1_b4-c1_c3-d1_c6-h1_d4-e1_e4-f1_f6-g2_h2-i1_i4-j1_j6-k2

RES	LIN
1b:b-dglc-HEX-1:5	1:1d(2+1)2n
2s:n-acetyl	2:1o(4+1)3d
3b:b-dglc-HEX-1:5	3:3d(2+1)4n
4s:n-acetyl	4:3o(4+1)5d
5b:b-dman-HEX-1:5	5:5o(3+1)6d
6b:a-dman-HEX-1:5	6:6o(4+1)7d
7b:b-dglc-HEX-1:5	7:7d(2+1)8n
8s:n-acetyl	8:7o(4+1)9d
9b:b-dgal-HEX-1:5	9:9o(6+2)10d
10b:a-dgro-dgal-NON-2:6|1:a|2:keto|3:d	10:10d(5+1)11n
11s:n-acetyl	11:5o(6+1)12d
12b:a-dman-HEX-1:5	12:12o(2+1)13d
13b:b-dglc-HEX-1:5	13:13d(2+1)14n
14s:n-acetyl	14:13o(4+1)15d
15b:b-dgal-HEX-1:5	15:15o(6+2)16d
16b:a-dgro-dgal-NON-2:6|1:a|2:keto|3:d	16:16d(5+1)17n
17s:n-acetyl	



Glycan (Partial) Structure Descriptions

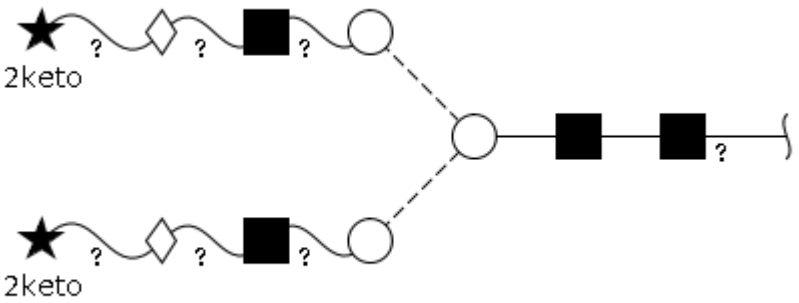
GlyTouCan: G27817LK
UniCarbKB: 4923
GlycomeDb: 39722



Neu5Ac(?2-?)Gal(?1-?)GlcNAc(?1-?)Man(a1-3)[Neu5Ac(?2-?)Gal(?1-?)GlcNAc(?1-?)Man(a1-6)]Man(b1-4)GlcNAc(b1-4)GlcNAc(?1-

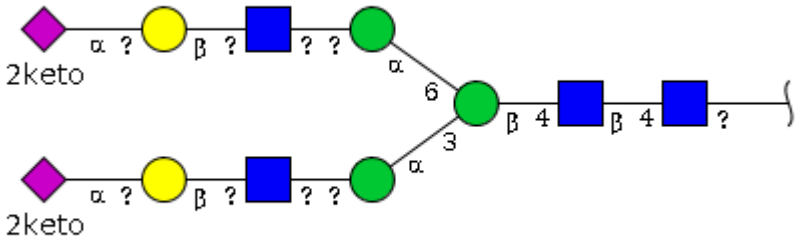
WURCS=2.0/6,11,10/[a2122h-1x_1-5_2*NCC/3=O][a2122h-1b_1-5_2*NCC/3=O][a1122h-1b_1-5][a1122h-1a_1-5][a2112h-1x_1-5]
[Aad21122h-2x_2-6_5*NCC/3=O]/1-2-3-4-1-5-6-4-1-5-6/a4-b1_b4-c1_c3-d1_c6-h1_d?-e1_e?-f1_f?-g2_h?-i1_i?-j1_j?-k2

RES	LIN
1b:x-dglc-HEX-1:5	1:1d(2+1)2n
2s:n-acetyl	2:1o(4+1)3d
3b:b-dglc-HEX-1:5	3:3d(2+1)4n
4s:n-acetyl	4:3o(4+1)5d
5b:b-dman-HEX-1:5	5:5o(3+1)6d
6b:a-dman-HEX-1:5	6:6o(-1+1)7d
7b:x-dglc-HEX-1:5	7:7o(-1+1)8d
8b:x-dgal-HEX-1:5	8:8o(-1+2)9d
9b:x-dgro-dgal-NON-2:6|1:a|2:keto|3:d	9:9d(5+1)10n
10s:n-acetyl	10:7d(2+1)11n
11s:n-acetyl	11:5o(6+1)12d
12b:a-dman-HEX-1:5	12:12o(-1+1)13d
13b:x-dglc-HEX-1:5	13:13o(-1+1)14d
14b:x-dgal-HEX-1:5	14:14o(-1+2)15d
15b:x-dgro-dgal-NON-2:6|1:a|2:keto|3:d	15:15d(5+1)16n
16s:n-acetyl	16:13d(2+1)17n
17s:n-acetyl	



Glycan (Partial) Structure Descriptions

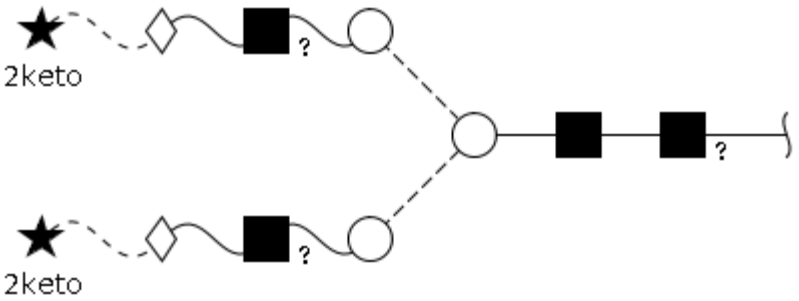
GlyTouCan: G97545TB
UniCarbKB: 4227
GlycomeDb: 39610



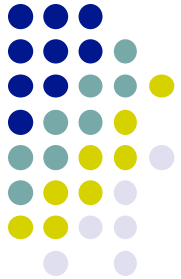
Neu5Ac(a2-?)Gal(b1-?)GlcNAc(?1-?)Man(a1-3)[Neu5Ac(a2-?)Gal(b1-?)GlcNAc(?1-?)Man(a1-6)]Man(b1-4)GlcNAc(b1-4)GlcNAc(?1-

WURCS=2.0/6,11,10/[a2122h-1x_1-5_2*NCC/3=O][a2122h-1b_1-5_2*NCC/3=O][a1122h-1b_1-5][a1122h-1a_1-5][a2112h-1b_1-5]
[Aad21122h-2a_2-6_5*NCC/3=O]/1-2-3-4-1-5-6-4-1-5-6/a4-b1_b4-c1_c3-d1_c6-h1_d?-e1_e?-f1_f?-g2_h?-i1_i?-j1_j?-k2

RES	LIN
1b:x-dglc-HEX-1:5	1:1d(2+1)2n
2s:n-acetyl	2:1o(4+1)3d
3b:b-dglc-HEX-1:5	3:3d(2+1)4n
4s:n-acetyl	4:3o(4+1)5d
5b:b-dman-HEX-1:5	5:5o(3+1)6d
6b:a-dman-HEX-1:5	6:6o(-1+1)7d
7b:x-dglc-HEX-1:5	7:7o(-1+1)8d
8b:b-dgal-HEX-1:5	8:8o(-1+2)9d
9b:a-dgro-dgal-NON-2:6|1:a|2:keto|3:d	9:9d(5+1)10n
10s:n-acetyl	10:7d(2+1)11n
11s:n-acetyl	11:5o(6+1)12d
12b:a-dman-HEX-1:5	12:12o(-1+1)13d
13b:x-dglc-HEX-1:5	13:13o(-1+1)14d
14b:b-dgal-HEX-1:5	14:14o(-1+2)15d
15b:a-dgro-dgal-NON-2:6|1:a|2:keto|3:d	15:15d(5+1)16n
16s:n-acetyl	16:13d(2+1)17n
17s:n-acetyl	

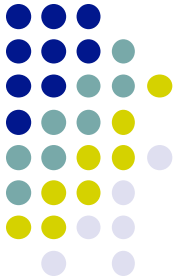


Glycans are complex molecules



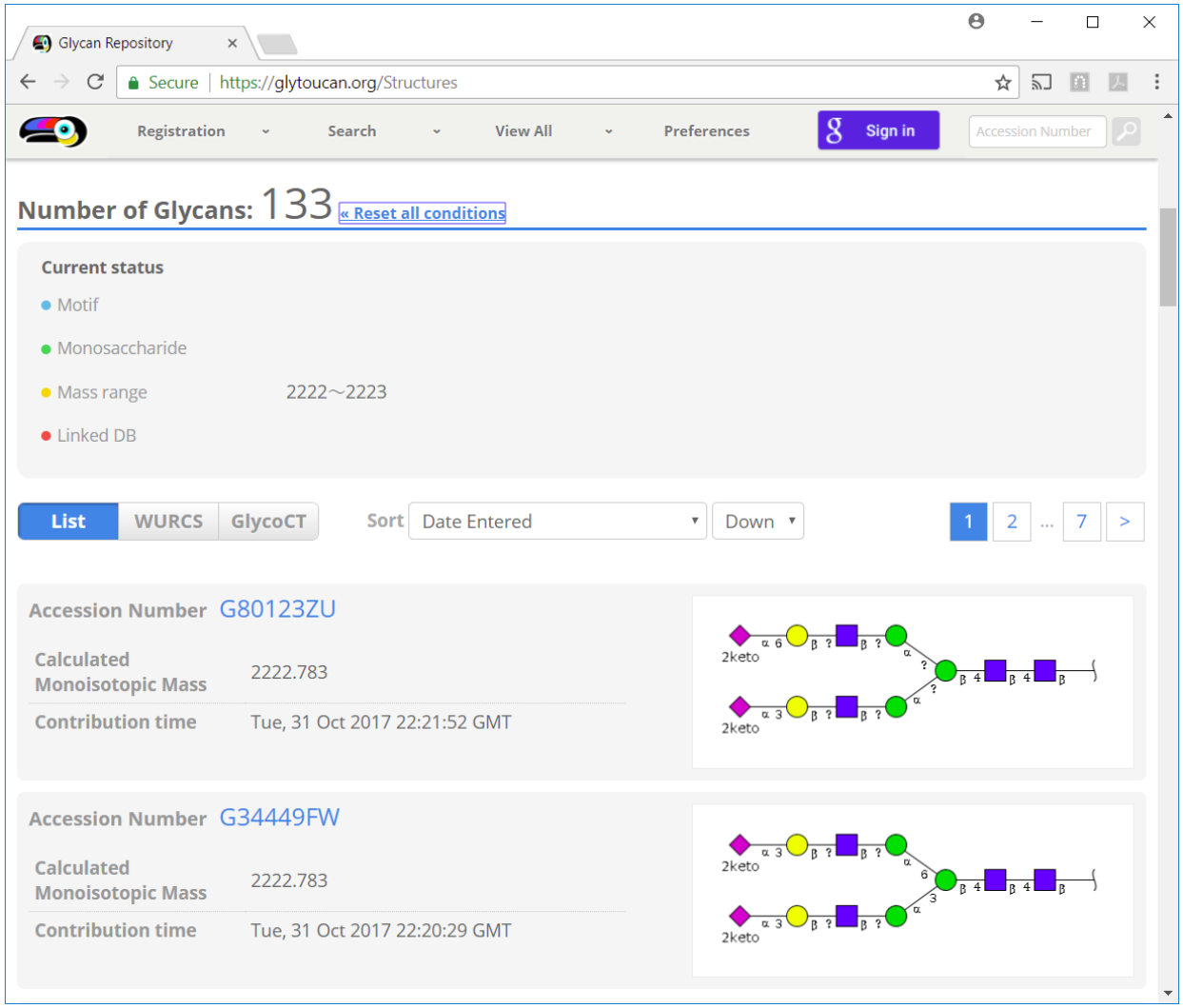
- Manuscripts and tools often use a short-hand
  - Monosaccharide composition
    - 4 GlcNAc, 3 Man, 2 Gal, 2 Neu5Ac (4-3-2-2-0)
    - 4 HexNAc, 5 Hex, 2 Neu5Ac (4-5-2-0)
  - Words (!)
    - Bi-antennary di-sialated complex N-glycan
  - Abbreviations
    - A2G2S2, 2A2SA, ….
  - Pictures
- Linking with accessions is difficult…

Finding GlyTouCan Structure Descriptions

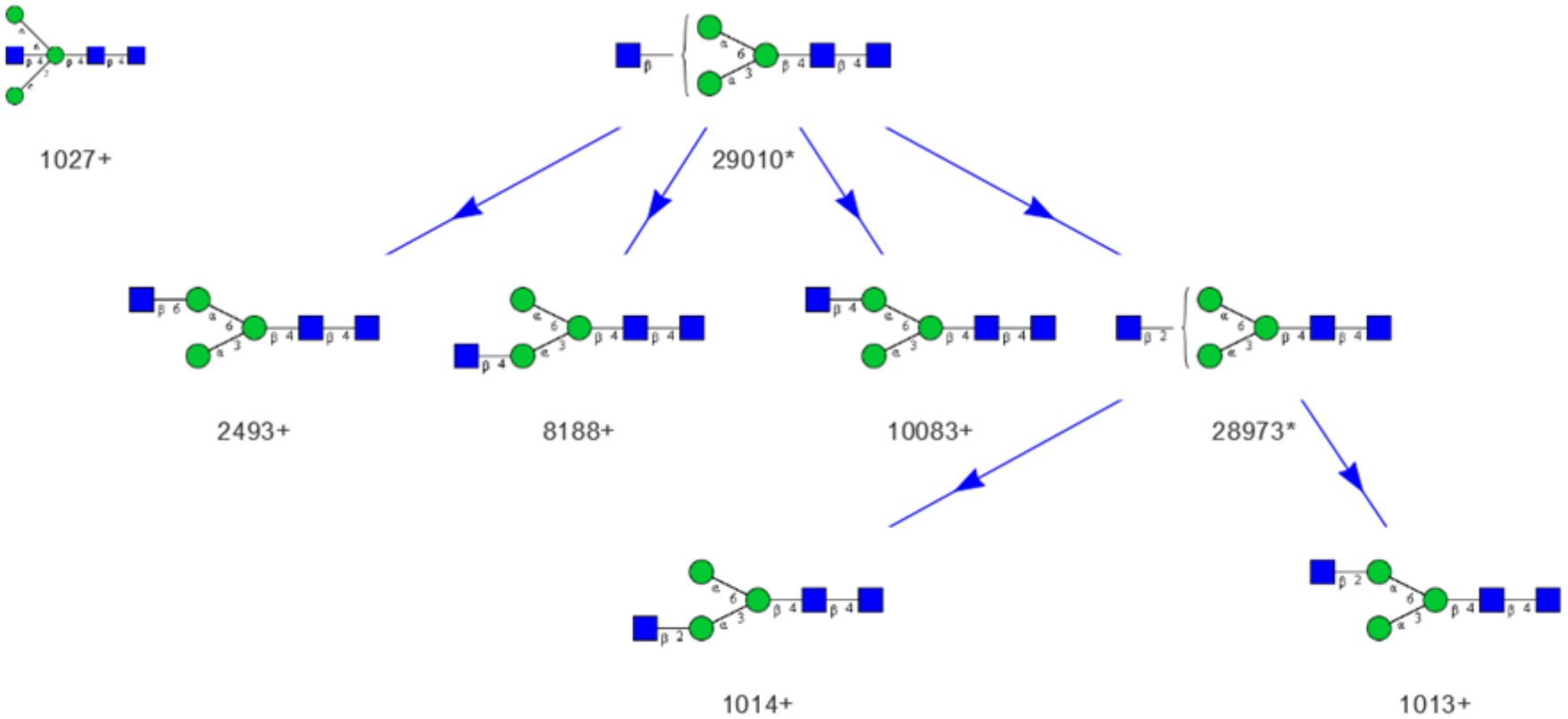


- 49 GlyTouCan glycans by composition
- 68 GlyTouCan glycans by mol. weight
- 95 GlyTouCan glycans by int. mol wt

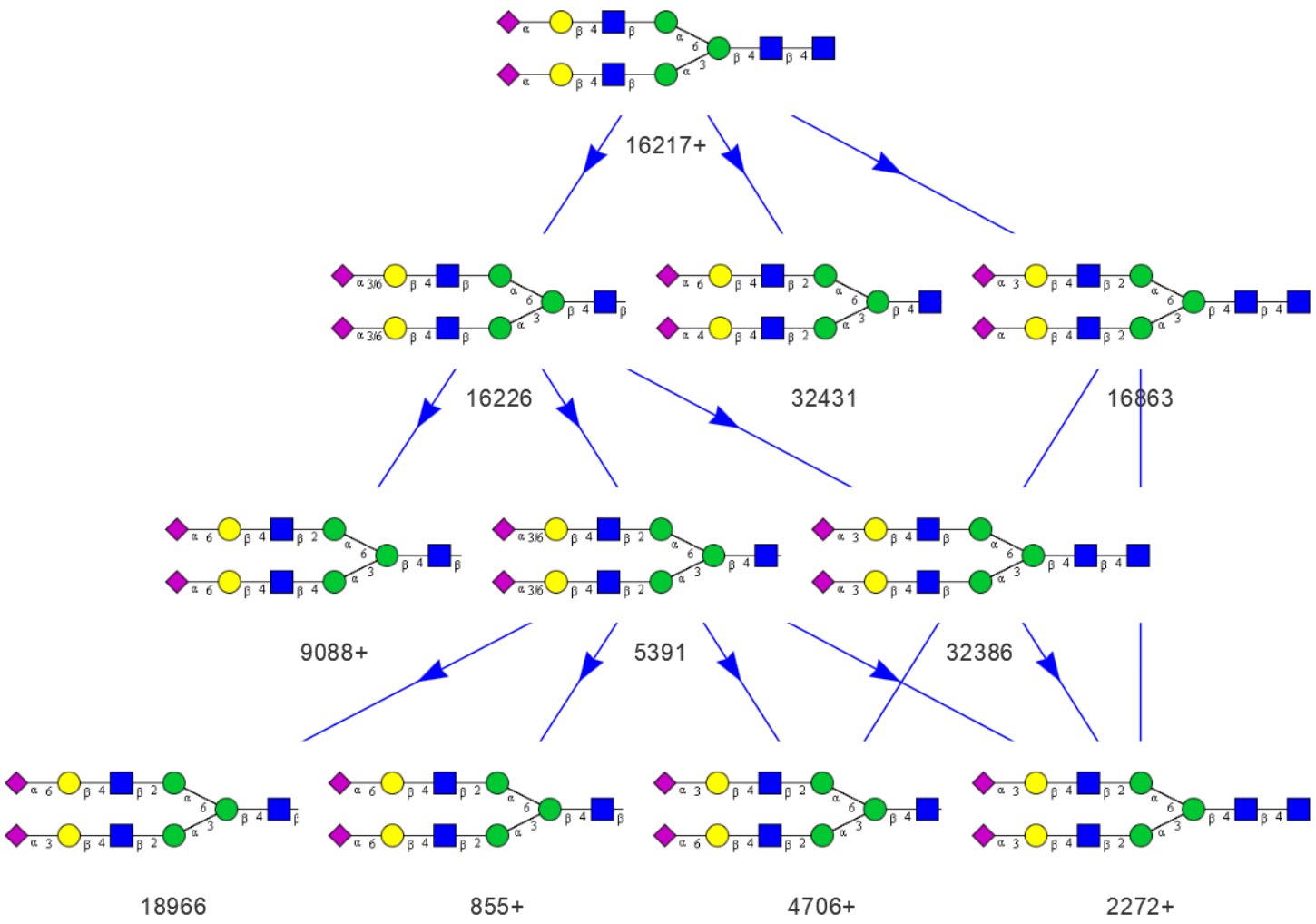
Finding GlyTouCan Structure Descriptions



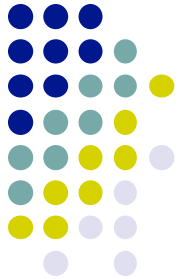
Glycan (Partial) Structure Subsumption



Glycan (Partial) Structure Subsumption

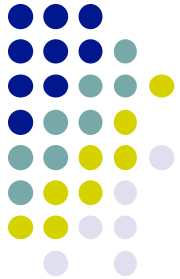


GlyGen Glycans



- Glycans from GlyTouCan
  - Primary accessions
  - Complete WURCS coverage - GlycoCT partial
  - Species/Motif annotations
- Glycans from UniCarbKB
  - Connection to UniProt proteins (vital!)
  - Species / Protein-based species annotations
  - Accessions linked from GlyTouCan

GlyGen Glycans (Year 1)

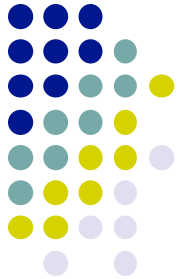


- Require human and mouse glycans, but…
  - Species annotations are sparse
- Data-resources do not annotate consistent (partial) structure descriptions
  - Human curators extract structure descriptions from manuscripts inconsistently,
  - Different resources see different manuscripts
- Human Glycans:
  - GlyTouCan: 2261; UniCarbKB: 1854 (1471);
  - In common: 486

GlyGen Glycans

- Which compositions, topologies are human?
  - Use subsumption relationships to propagate annotation
  - Semantics: …at least one glycan structure…
- Merged human annotation:
  - (Partial?) Structures: 3455 (+621)
  - Topologies: 1720 (+1707)
  - Compositions: 0 (+1062)

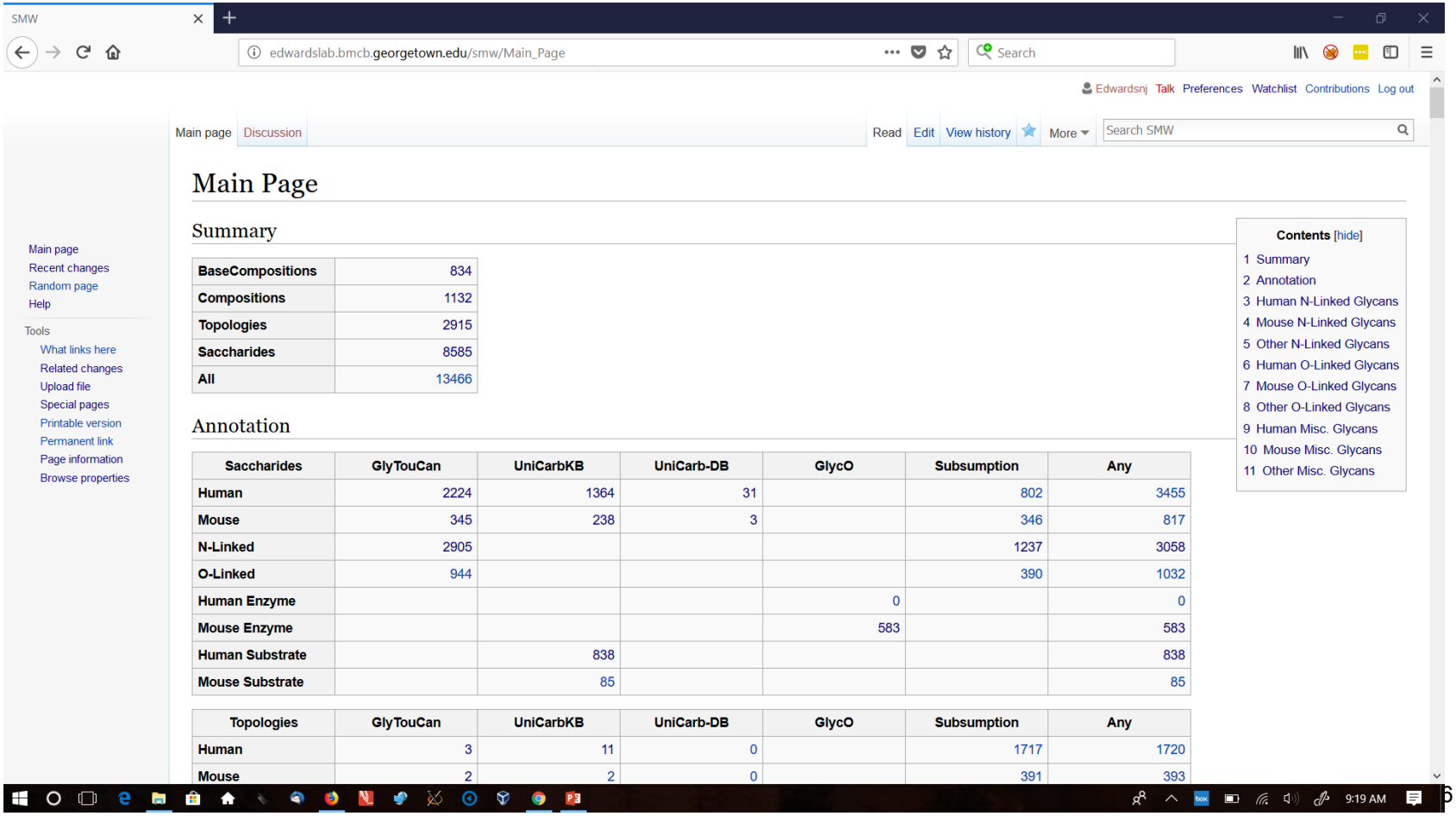
GlyGen Glycoenzymes



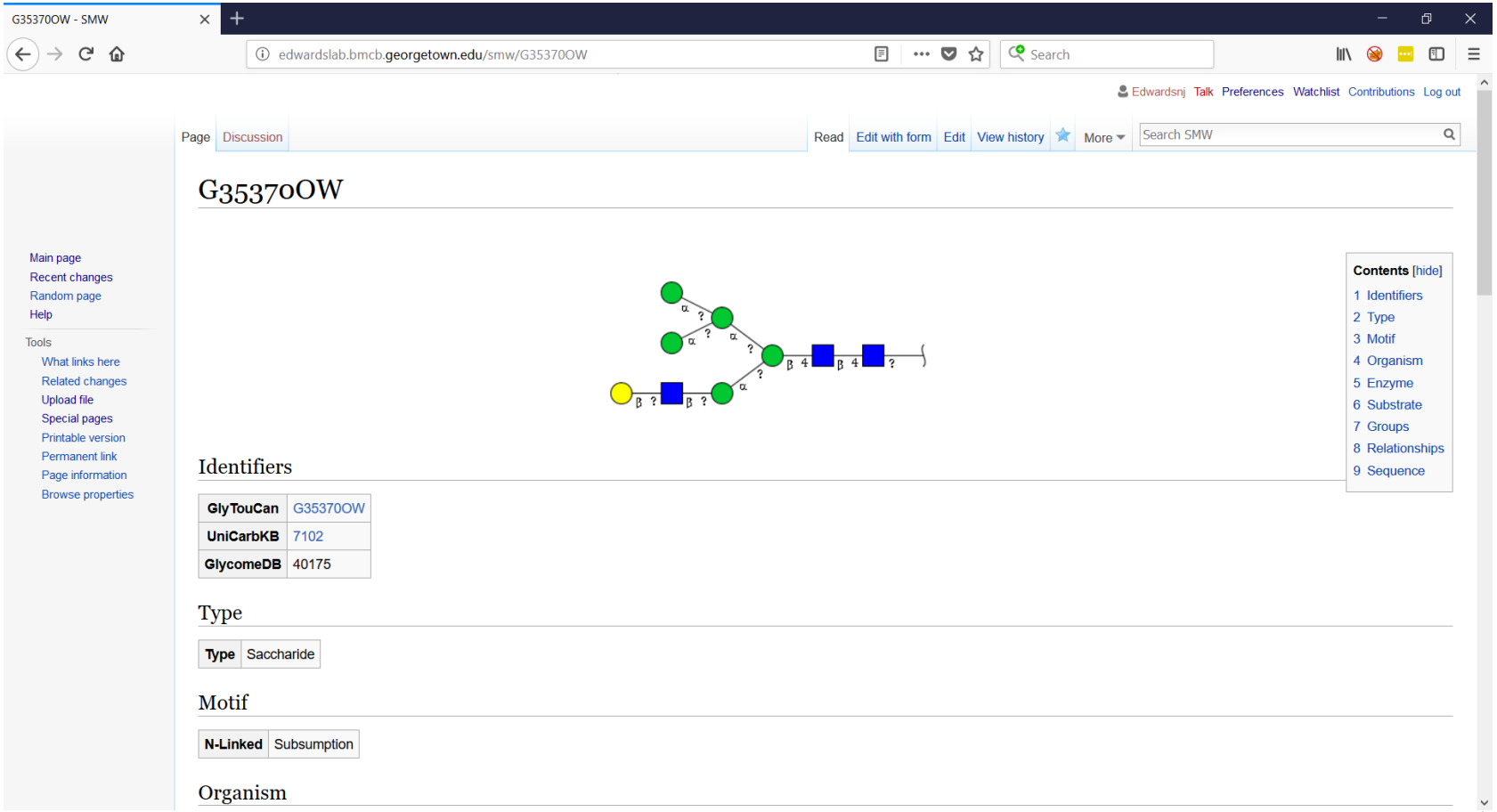
- Glycosylation enzyme annotations
  - ID mapping: GlycO → GlycomeDb → GlyTouCan
- GlycO “tree” of potential structures
  - w/ species specific glycotransferase annotations
  - Glycan monosaccharide indexing

- Will York, Alison Nairn, Kelley Moremen (CCRC)

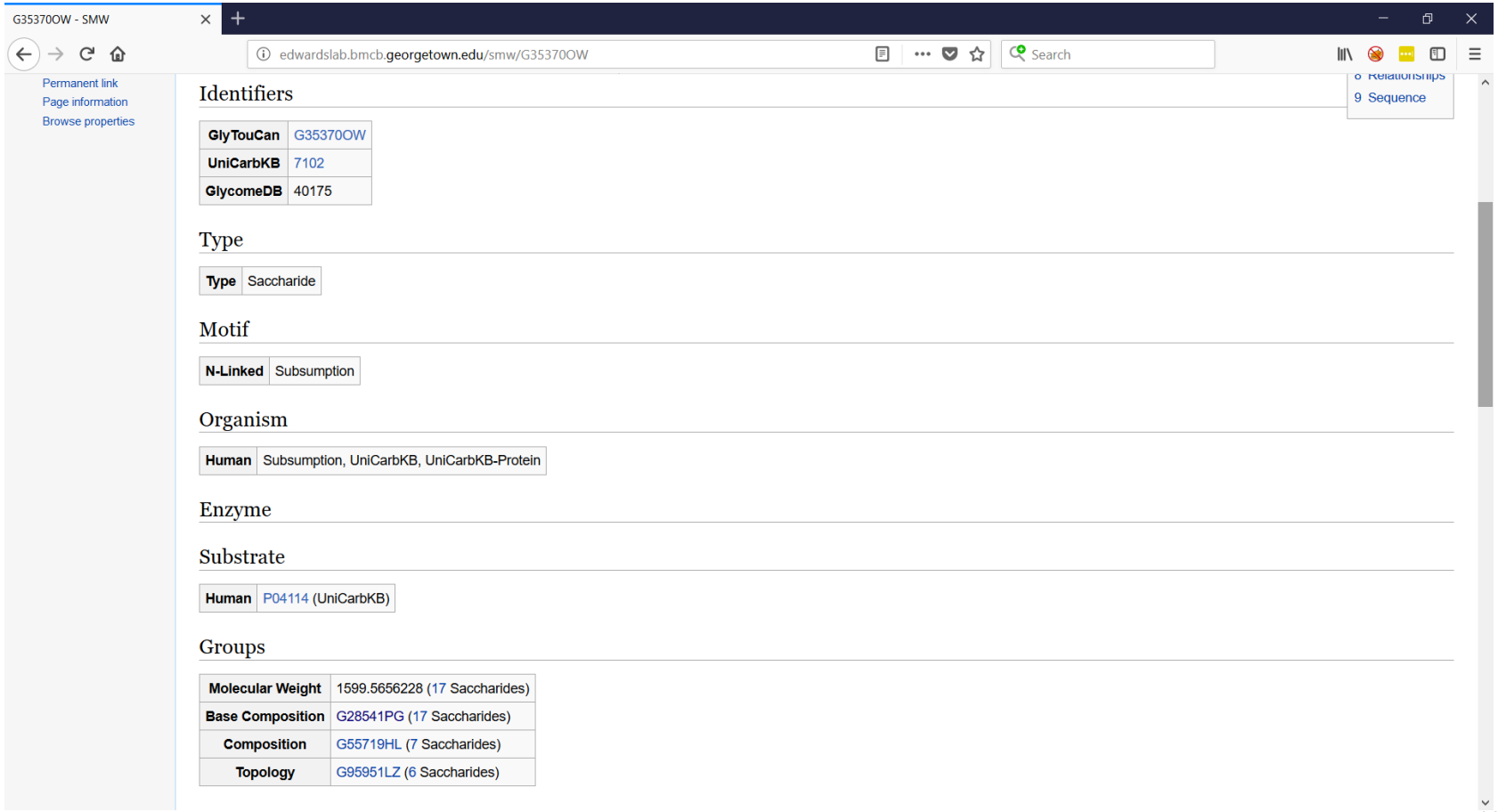
GlyGen Glycan Data Integration



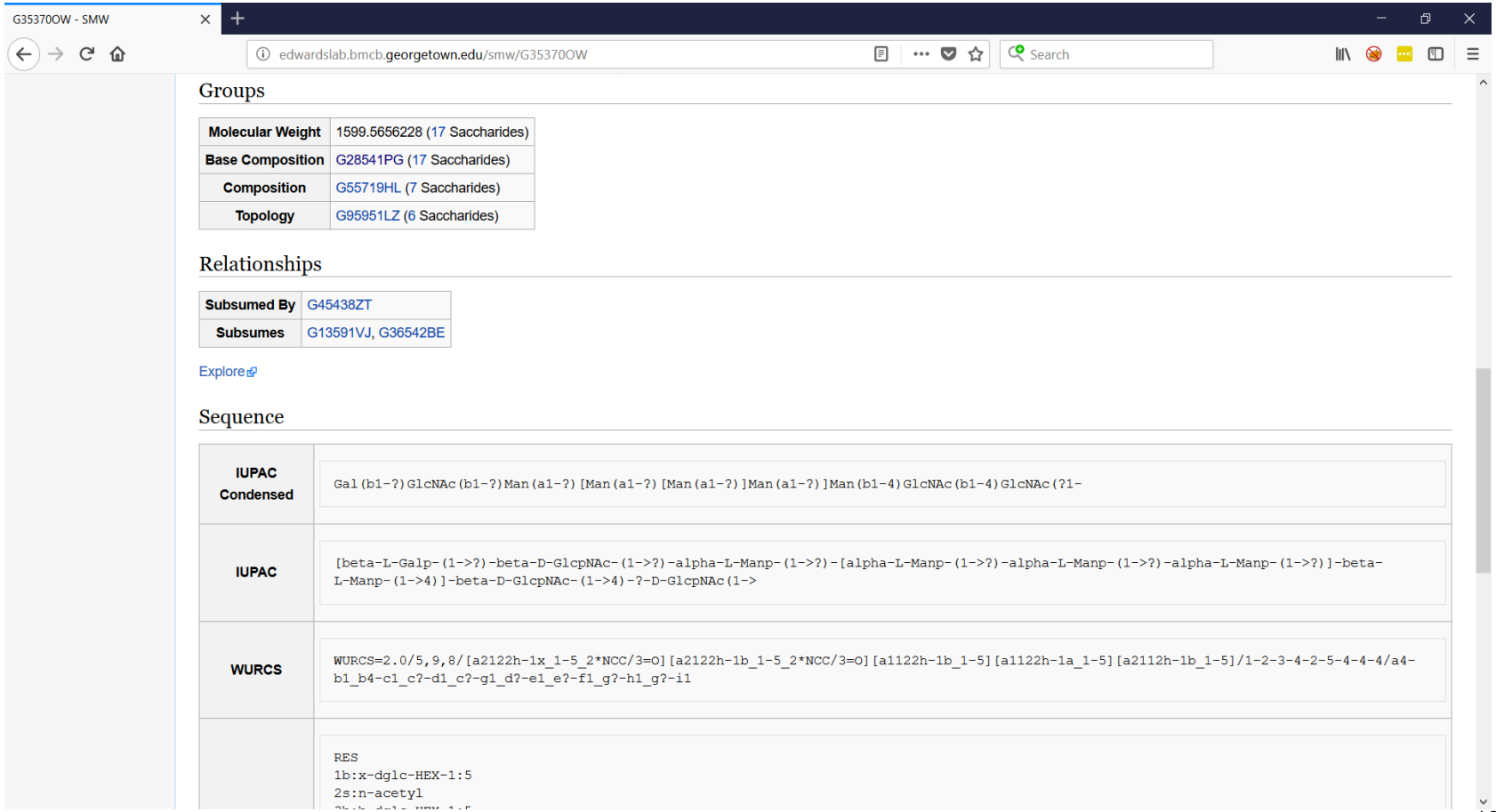
GlyGen Glycan Data Integration



GlyGen Glycan Data Integration



GlyGen Glycan Data Integration



Groups

Molecular Weight	1599.5656228 (17 Saccharides)
Base Composition	G28541PG (17 Saccharides)
Composition	G55719HL (7 Saccharides)
Topology	G95951LZ (6 Saccharides)

Relationships

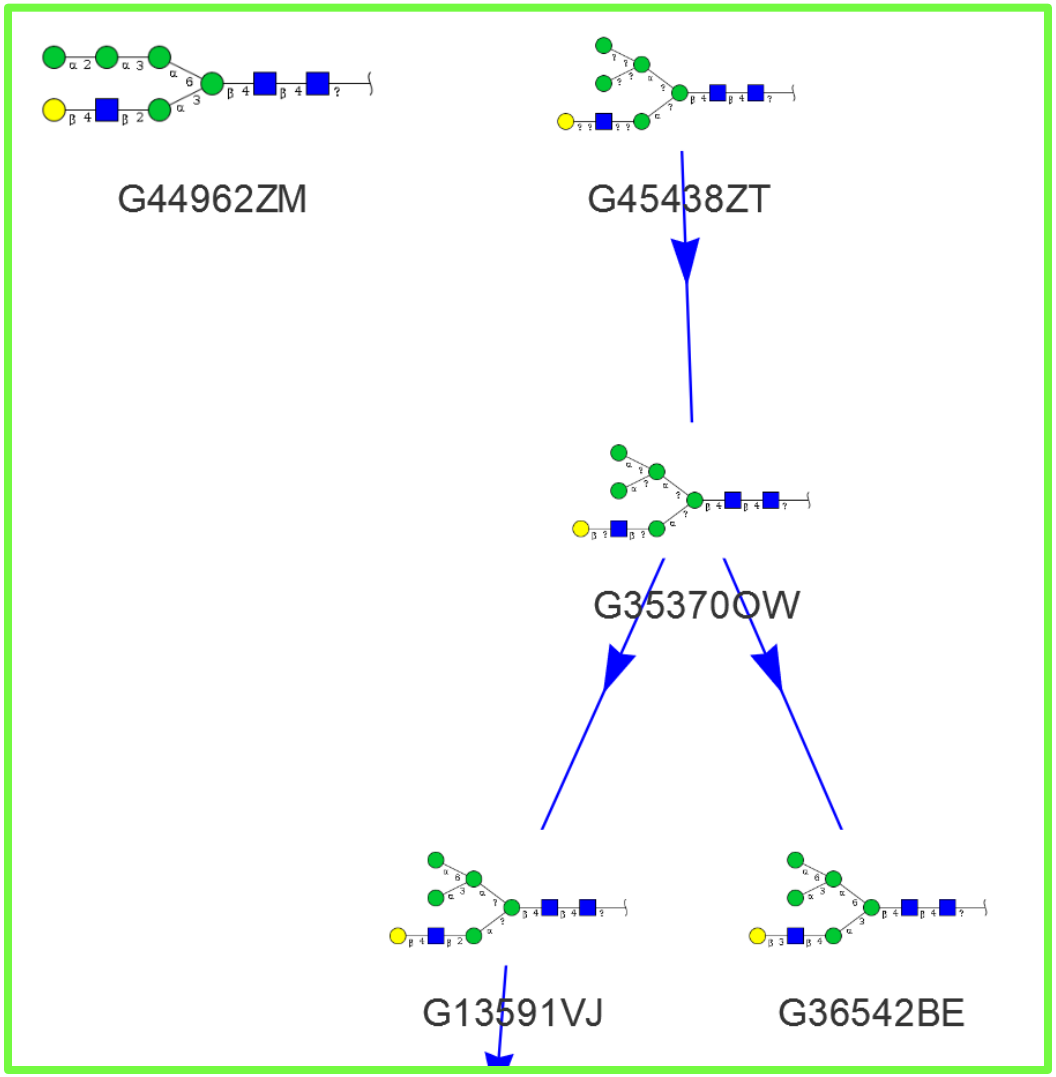
Subsumed By	G45438ZT
Subsumes	G13591VJ, G36542BE

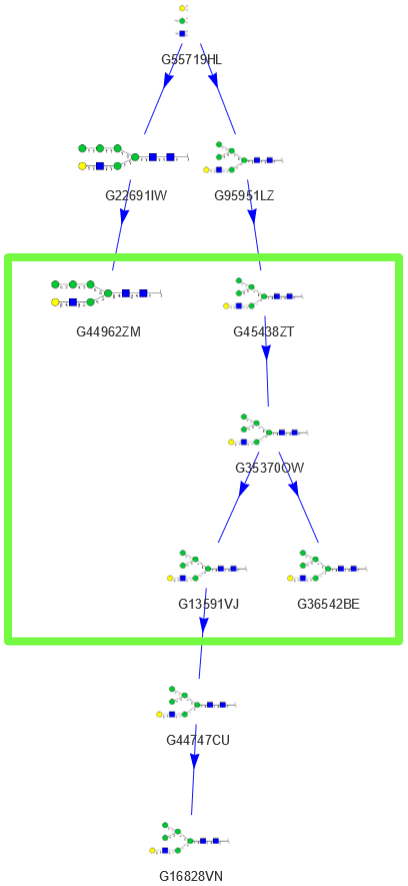
Explore

Sequence

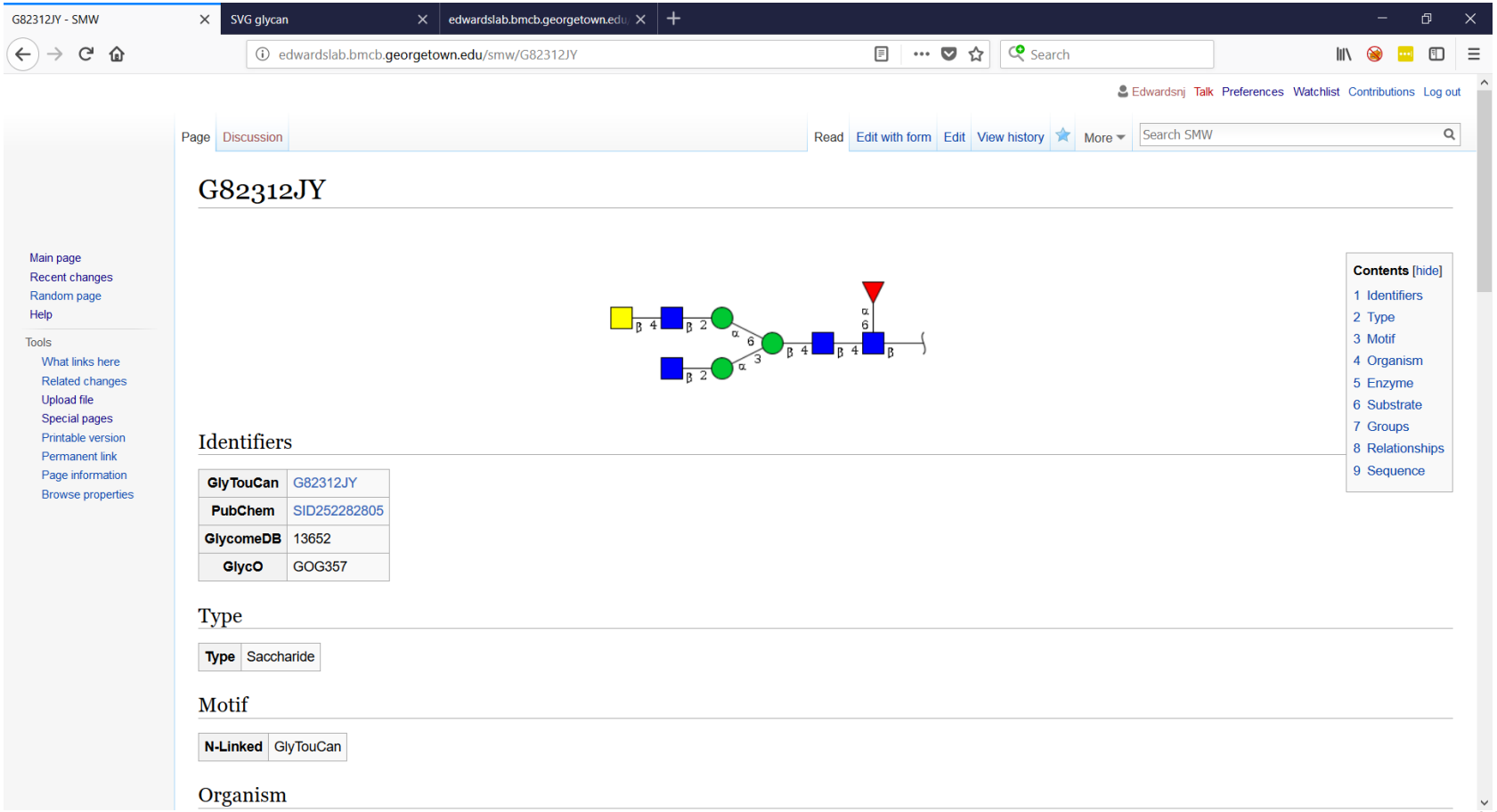
IUPAC Condensed	Gal(b1-?)GlcNAc(b1-?)Man(a1-?)[Man(a1-?)[Man(a1-?)]Man(a1-?)]Man(b1-4)GlcNAc(b1-4)GlcNAc(?1-
IUPAC	[beta-L-Galp-(1->?)-beta-D-GlcpNAc-(1->?)-alpha-L-Manp-(1->?)-[alpha-L-Manp-(1->?)-alpha-L-Manp-(1->?)-alpha-L-Manp-(1->?)]-beta-L-Manp-(1->4)]-beta-D-GlcpNAc-(1->4)-?-D-GlcpNAc(1->
WURCS	WURCS=2.0/5,9,8/[a2122h-1x_1-5_2*NCC/3=O][a2122h-1b_1-5_2*NCC/3=O][a1122h-1b_1-5][a1122h-1a_1-5][a2112h-1b_1-5]/1-2-3-4-2-5-4-4-4/a4-b1_b4-c1_c?-d1_c?-g1_d?-e1_e?-f1_g?-h1_g?-i1
	RES 1b:x-dglc-HEX-1:5 2s:n-acetyl

Explore subsumption relationships

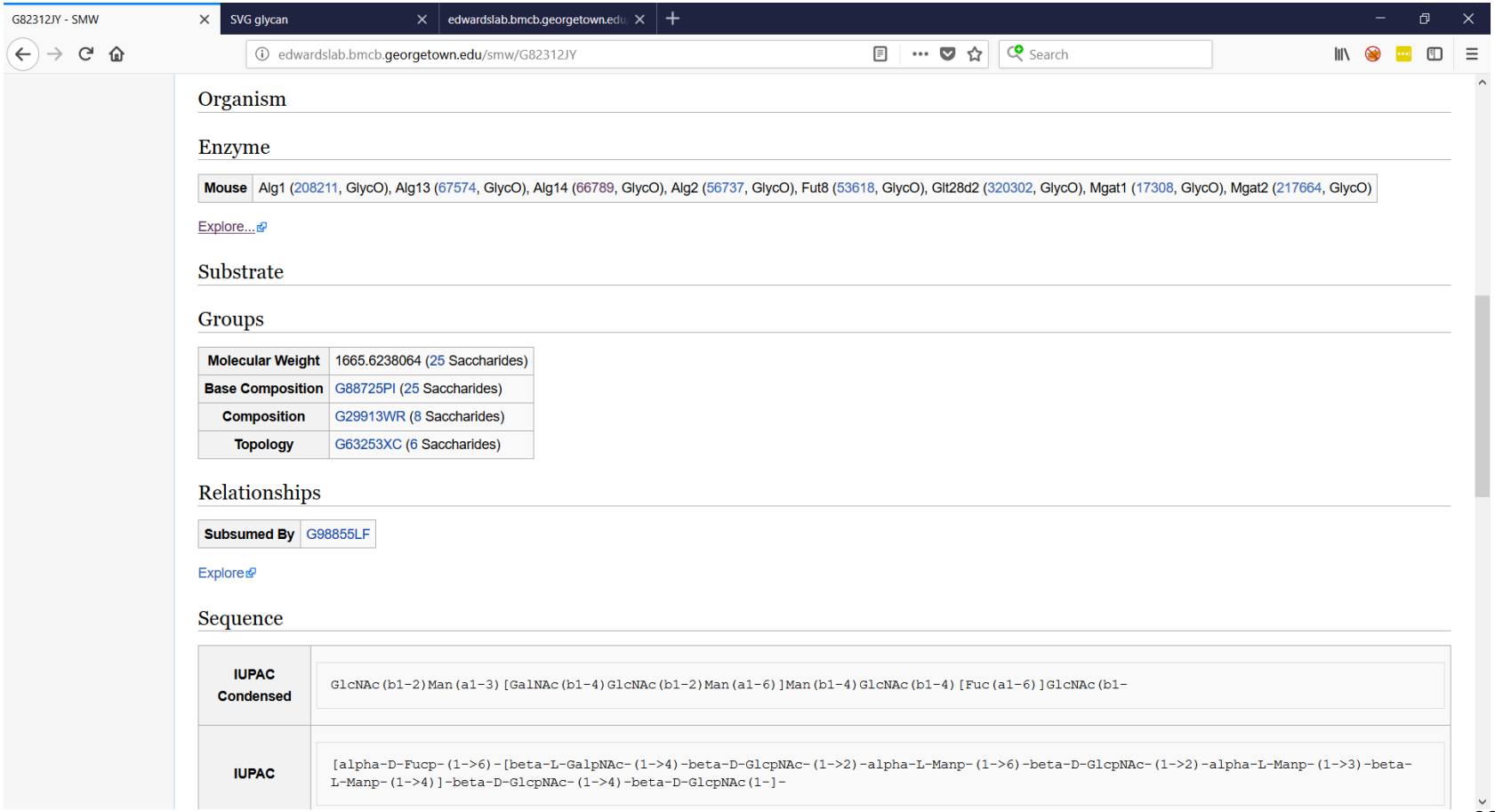




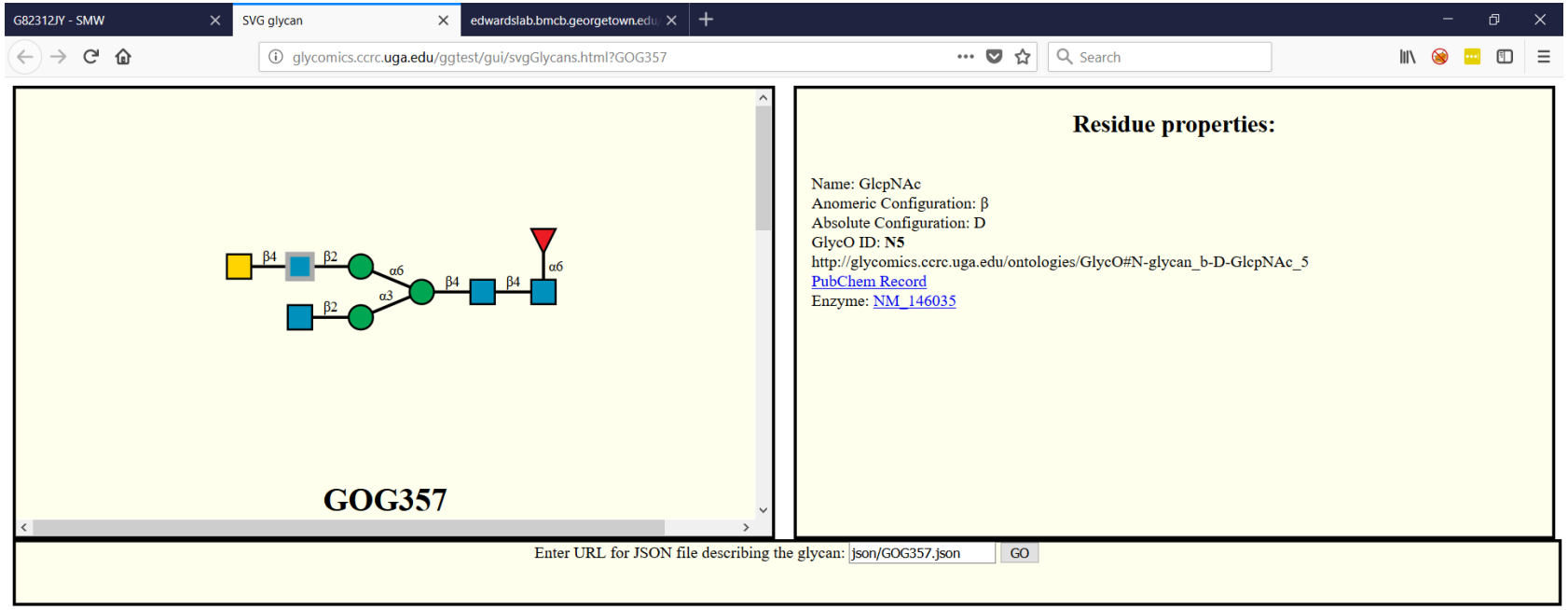
GlyGen Glycan Data Integration



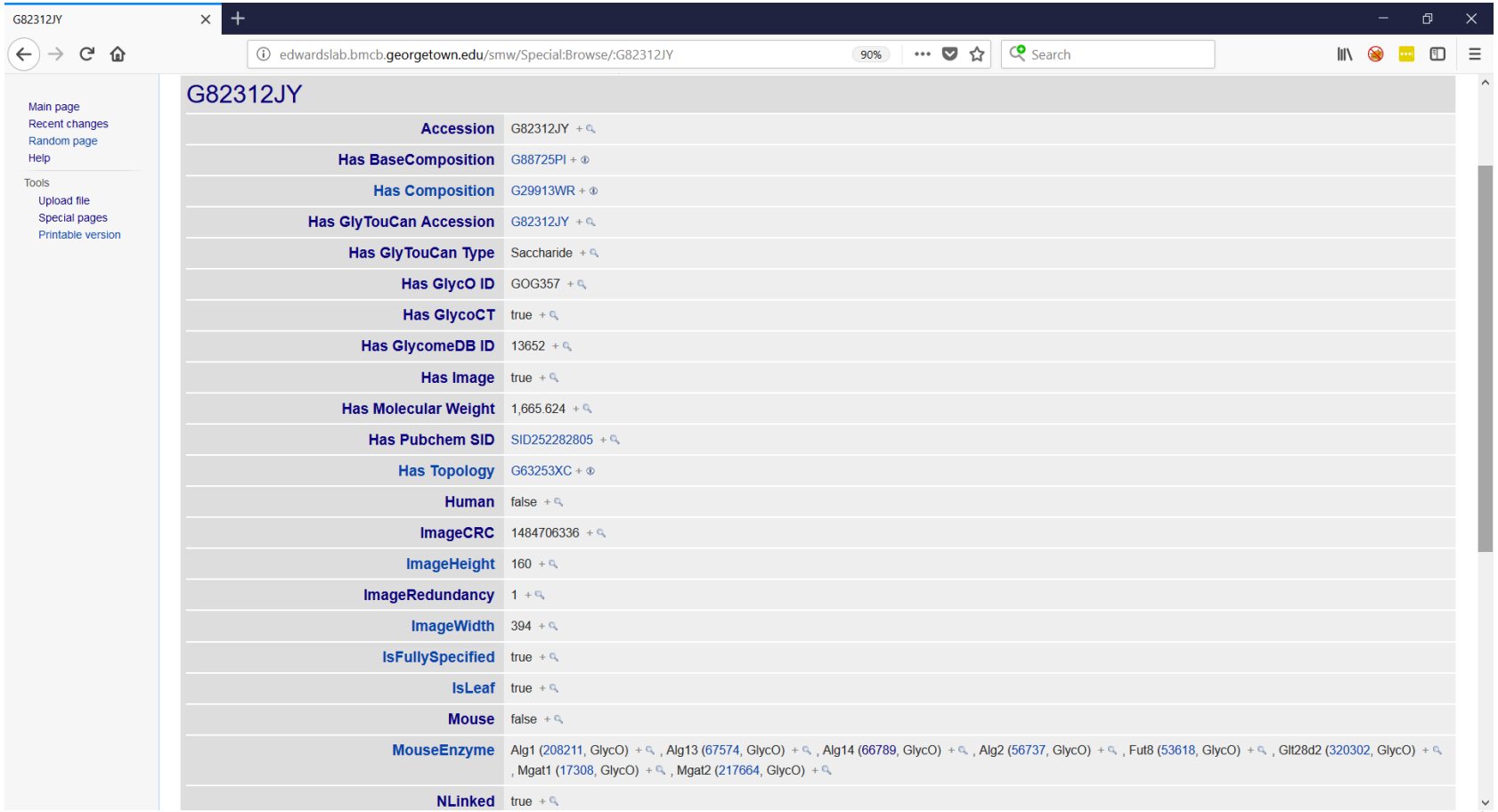
GlyGen Glycan Data Integration



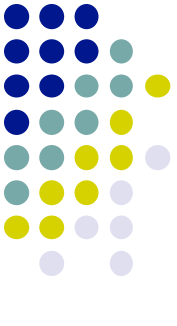
Explore Glycoenzymes

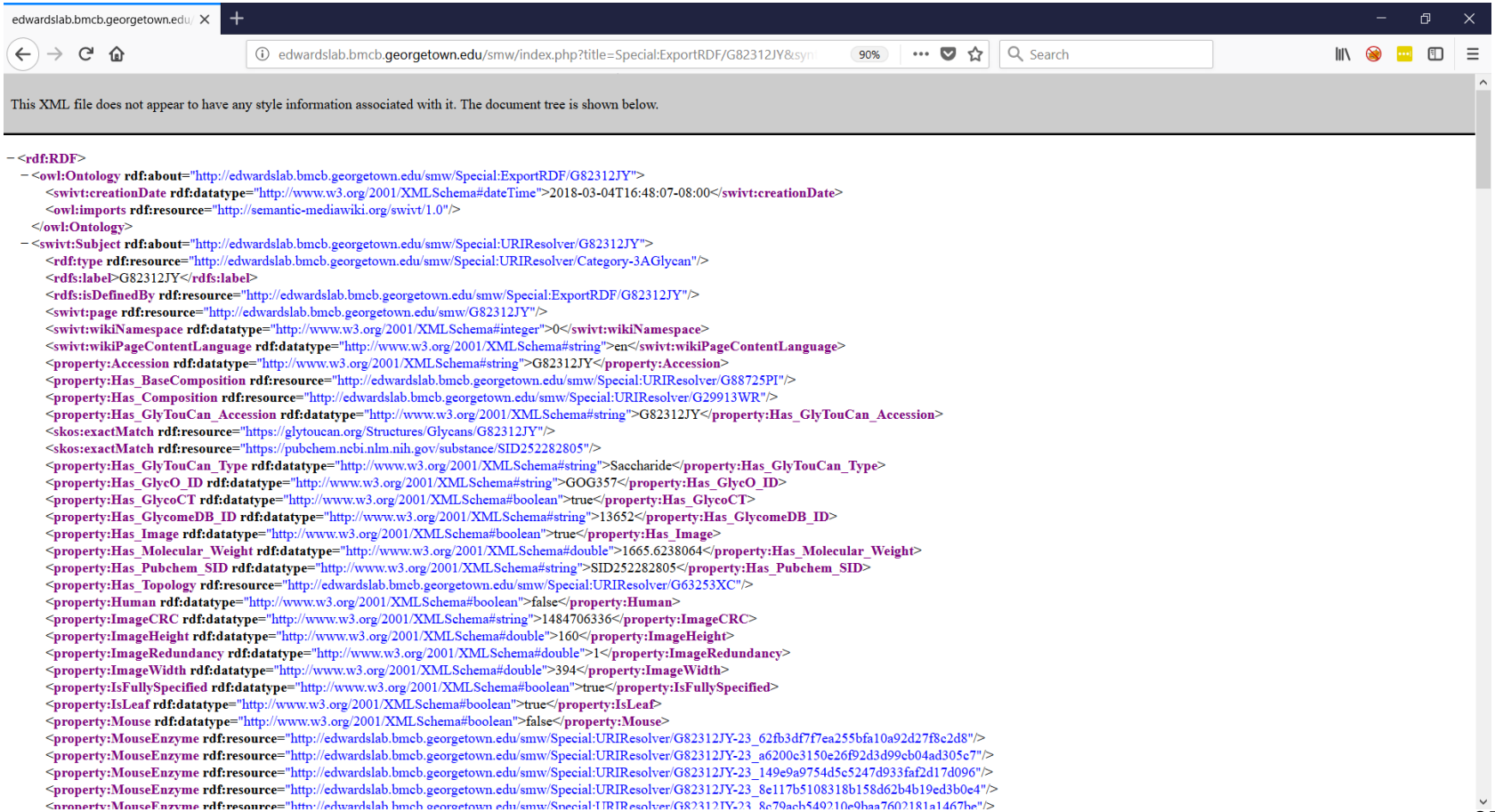


Semantic Media-Wiki

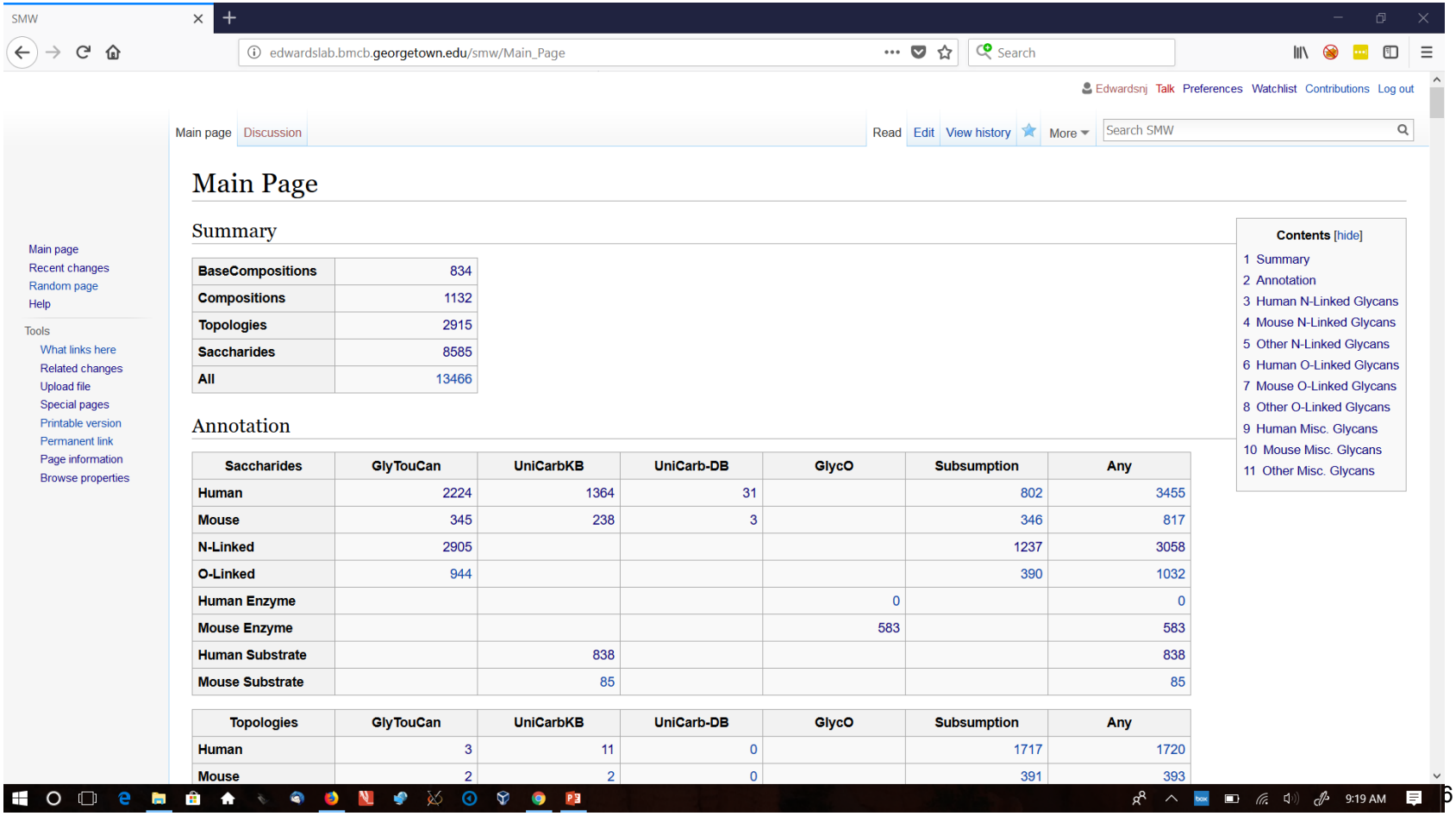


Semantic Media-Wiki





GlyGen Glycan Data Integration



GNOme

- Glycan Naming (and Subsumption) Ontology
- Pre-compute GlyTouCan subsumption relationships
  - Publish as OBO Foundry Ontology
- Ontologies good for structured terms
- Name/identifier resolution
  - Human terminology, synonyms, ID mapping

GNOme

- Provide browsing tools for interactive glycan (partial) structure description discovery
  - Help humans navigate glycan descriptions
- Well defined (partial) structure levels
- Glycan annotation propagation and inference

Acknowledgments

- GlyGen Glycoscience Portal Team
  - Will York (CCRC), Rene Ranzinger (CCRC, Glycome-DB), Raja Mazumder (GWU)
  - Kiyoko F Aoki-Kinoshita (GlyTouCan)
  - Matthew Campbell (UniCarbKB)
- NIH Common Fund: U01 - GlyGen