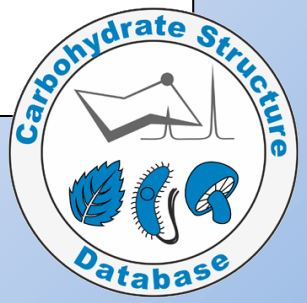
Philip Toukach

Carbohydrate Databases

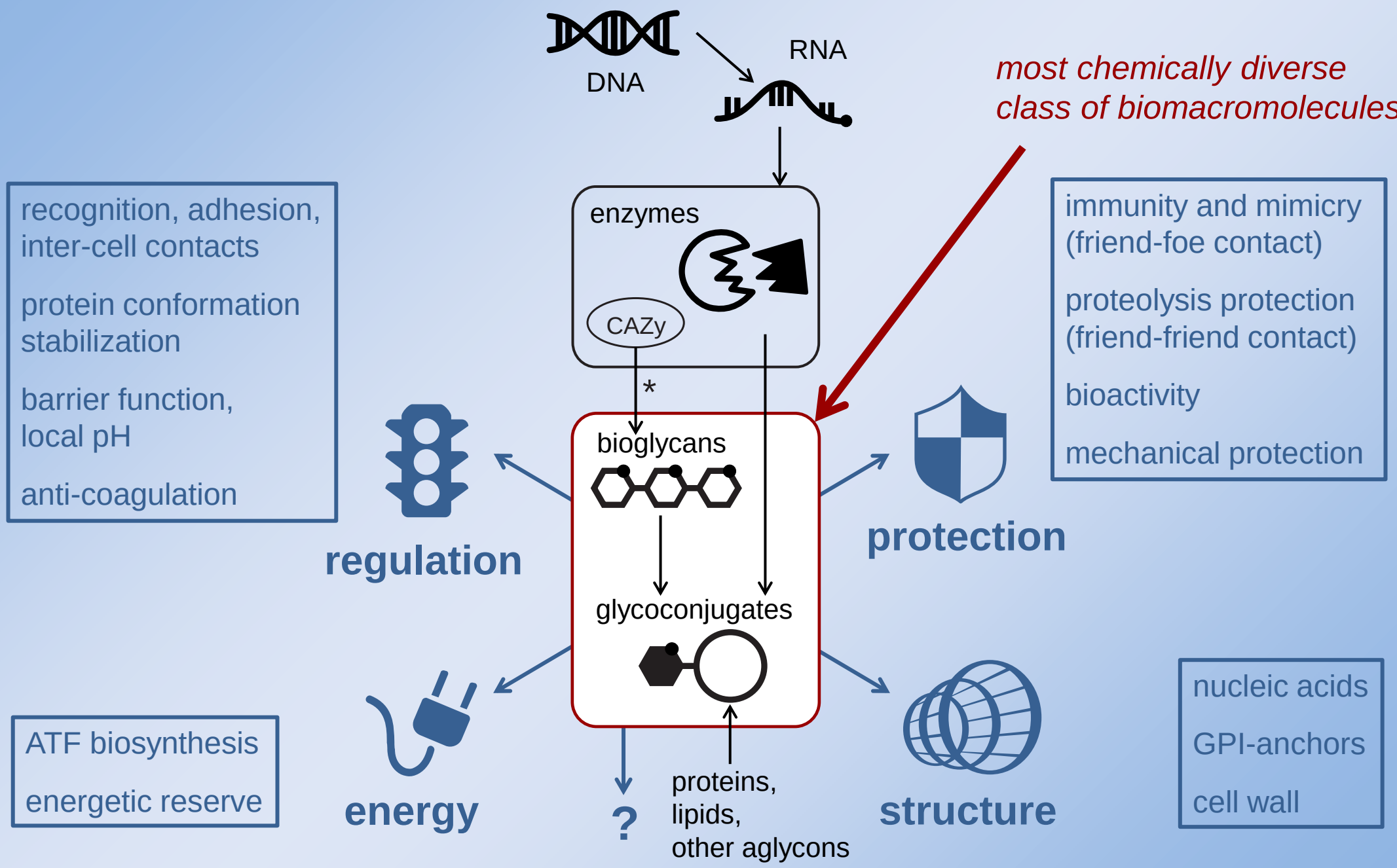
http://toukach.ru/rus/glyco-db.htm

Carbohydrate databases

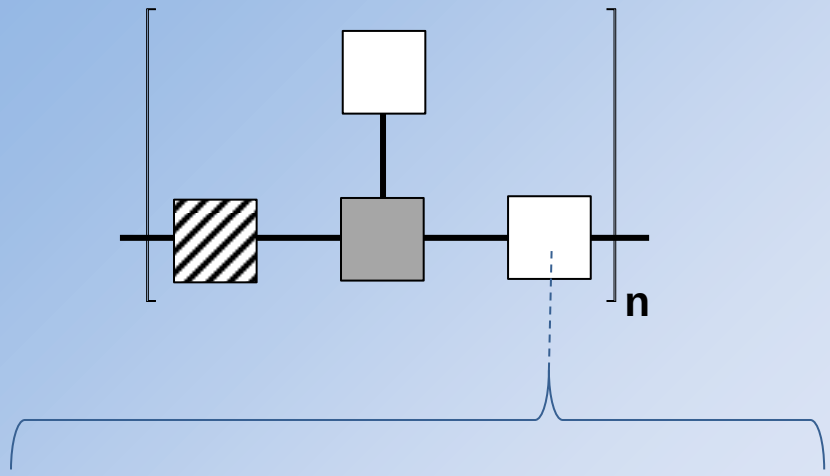
- What are glycans and what are databases?
- What do we expect from them?
- Why do not we get it?
- What is an ideal database?
- What do we have already?



Glycans in cells

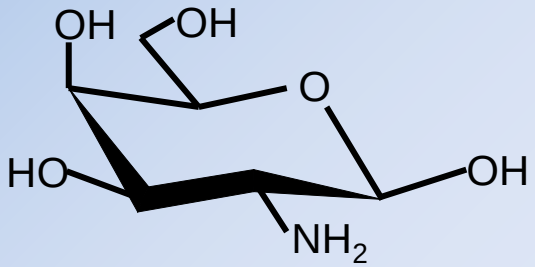


Glycan structure

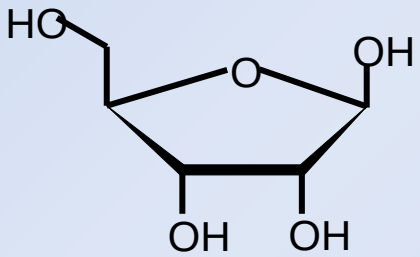


Complete structure

- monomeric residues, incl. non-sugars
- topology & sequence
- linkage positions
- side chain stoichiometry
- polymer size and frame positioning



aldo-pyranose example (β-D-Gal*p*N)



aldo-furanose example (β-D-Rib*f*)

Residue structure

- carbon skeleton size (4-10)
- stereo pattern (monomer identity)
- ring form (*p/f/a, aldo/keto*)
- anomeric configuration (α/β)
- absolute configuration
- modifications (-NH_2,-COOH, deoxy)

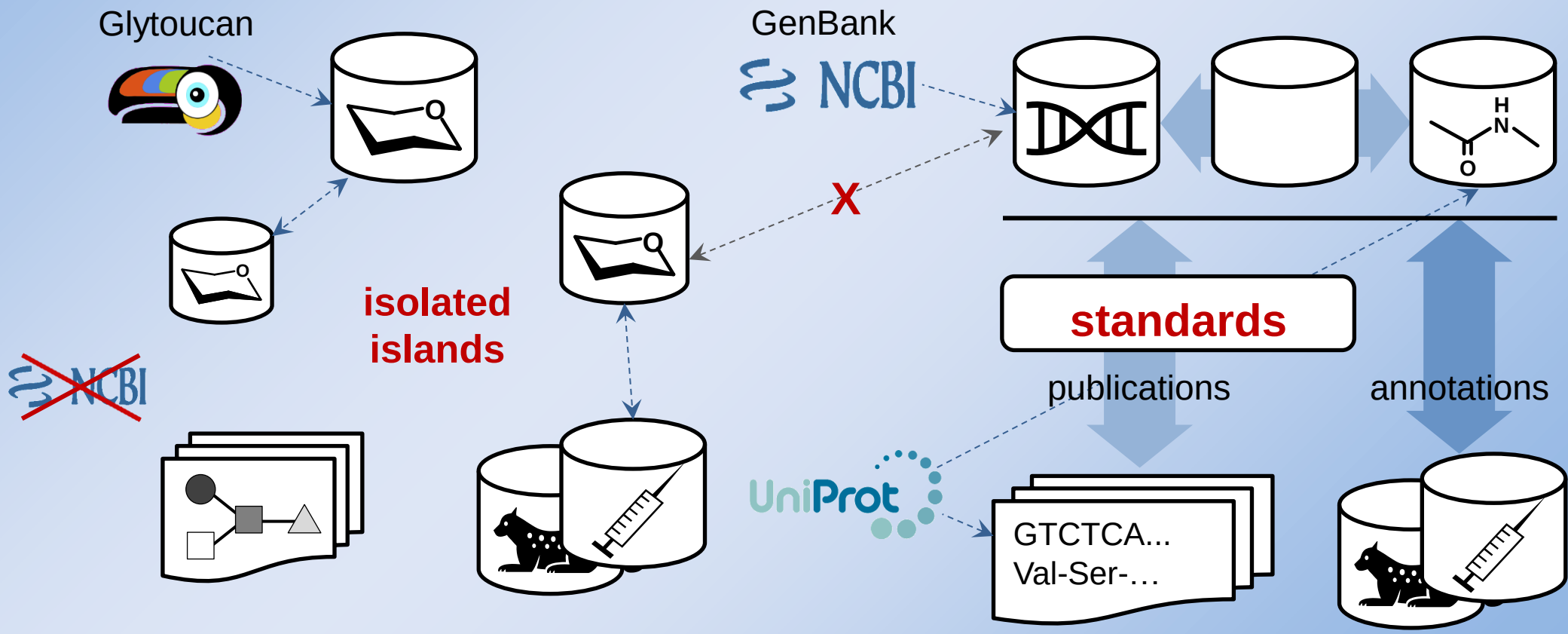
Glycomics in biology

- Structure, diversity, conformation of carbohydrates
- Taxonomy and classification of microorganisms
- Glyco-epitopes and immunospecificity of strains
- Explanation of antigen-antibody interaction
- Carbohydrate vaccines and immunostimulators
- Correlation of microorganism bioactivity to its glycome
- Carbohydrate biosynthesis and turn-over

Glycomics vs. genomics, proteomics

as compared to other –omics:

- similar information scope (>100 000 known structures)
- greater chemical diversity
- poorer IT involvement (databases, services)
- less standardized



Why do we need a database?

- **Easy access to knowledge and research automation**

  Which natural structures look like a model of interest? Which of their fragments are specific to a genus of interest? Where were they published, and in association to which taxa, diseases, and organs? Which enzymes synthesize them and how was it proven? Which glyco-epitopes induce immune response in an organism of interest?

- **Simulation of molecular properties**

  Molecular geometry and energy, MS and NMR spectra, bioactivity, …

- **Structure prediction from experimental data**

- **Prediction of taxon properties**

  Glycome-based clustering, search for similarities and differences in taxa, chemotaxonomic classification

- **Molecule identification and visualization** (in publications as well)

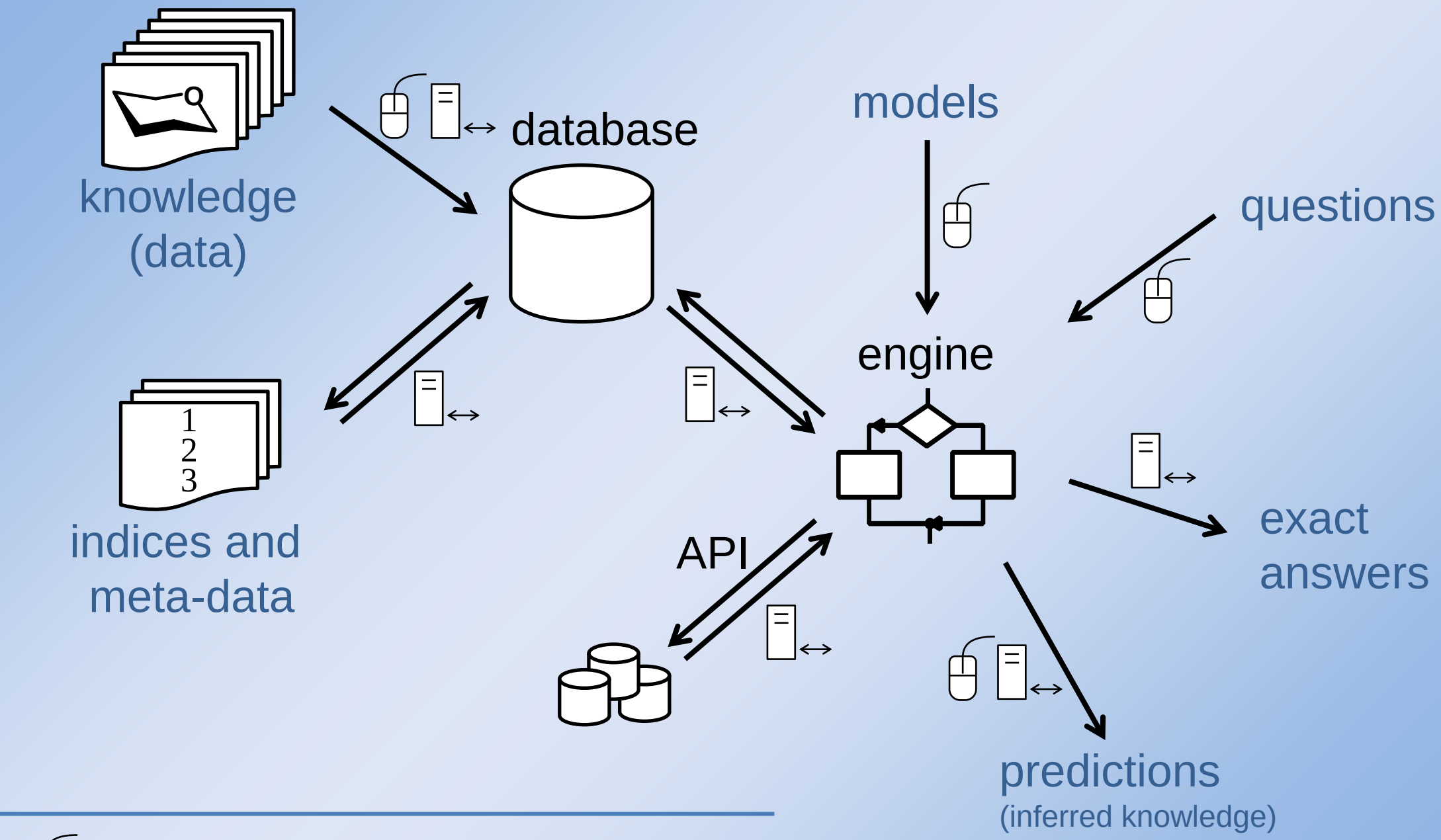
Glycoinformatic challenges

- Diversity and heterogeneity of objects
- Ambiguous structural description
- Difficult input and visualization of complex structures
- Project isolation and lack of standards
- Incompleteness of data and poor data quality
- Resource-greedy algorithms
- Lack of systematic view from users and developers

  (no commonly-agreed services; incompatibility of initiatives)

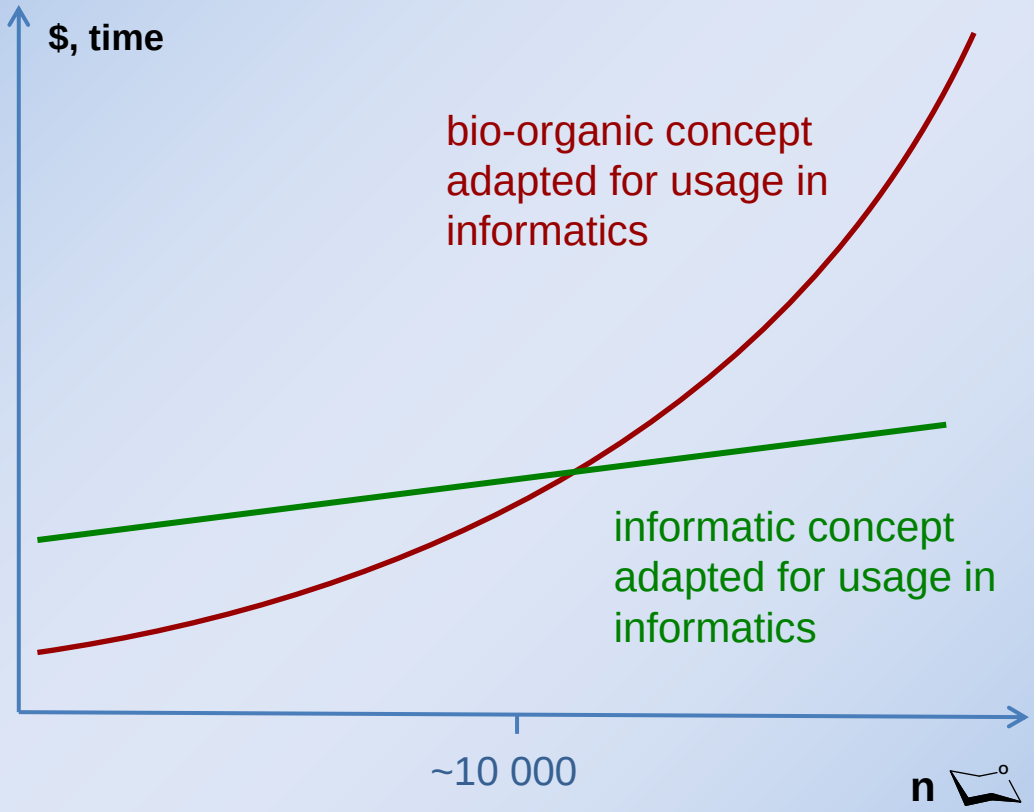
Aoki-Kinoshita K, *et al.* **Introducing glycomics data into Semantic Web** *J Biomed Semantics 2013, 4: id 39*
Egorova K, Toukach Ph **Bridging isolated islands in the sea of data** *Angewandte Chemie Intl Ed 2018, 57:14986-14990*

What is a database?



Development approaches

- A database and a platform should follow the rules of informatics
- These rules should be adapted and specificated for carbohydrates

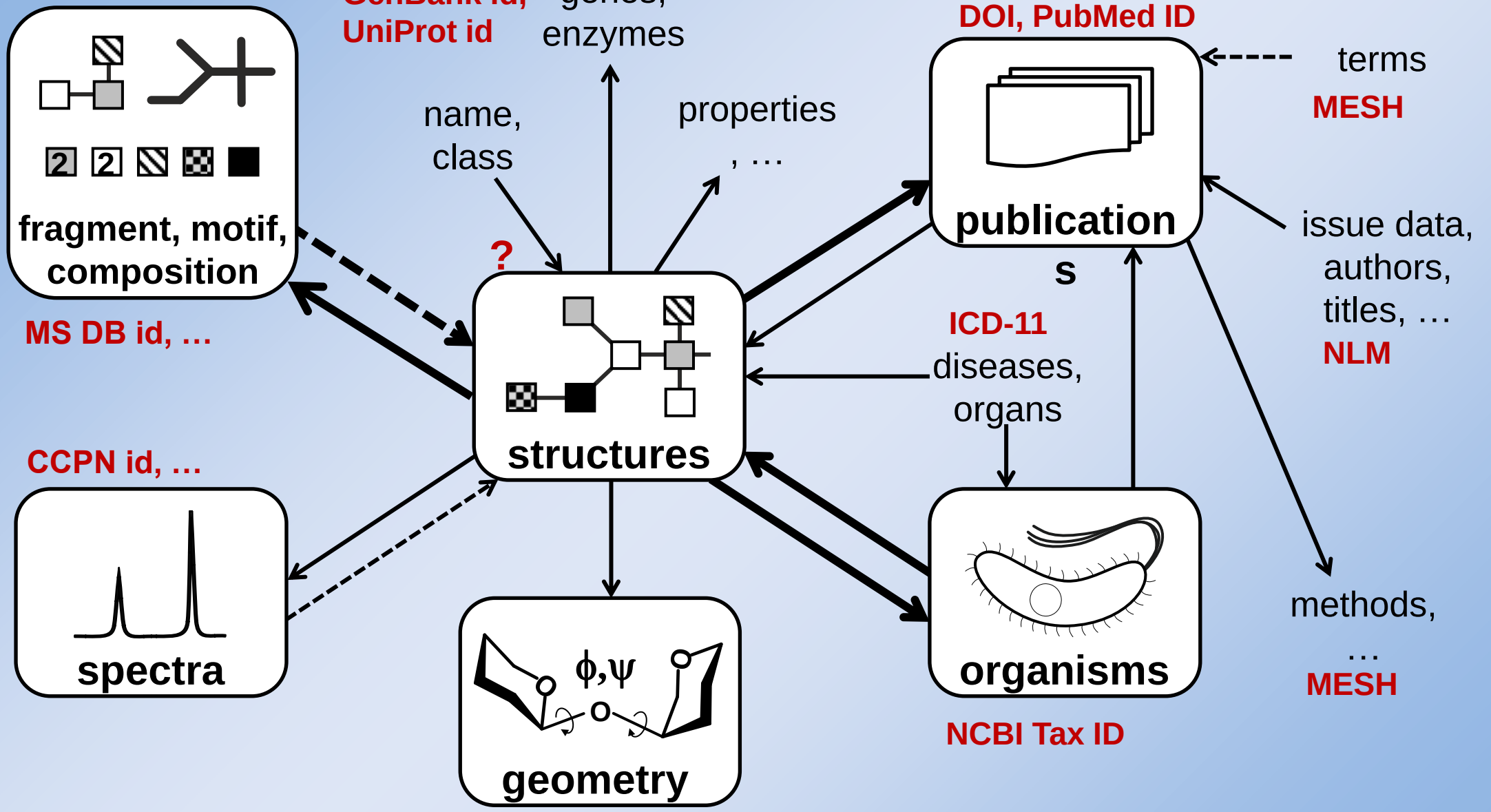


CarbBank

1989-1995

2008-…

Typical queries

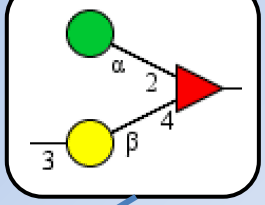


CSDB: complex query

fragment

Lys / Ala
↓ 2
6
SugA

fragment



composition

2 x KDO
1 x HexN
+ …

OR

AND

AND

class / name

AND

AND

AND

taxonomy

Proteus O1

host:
human

bibliography

≥ 2005

keyword:
LPS

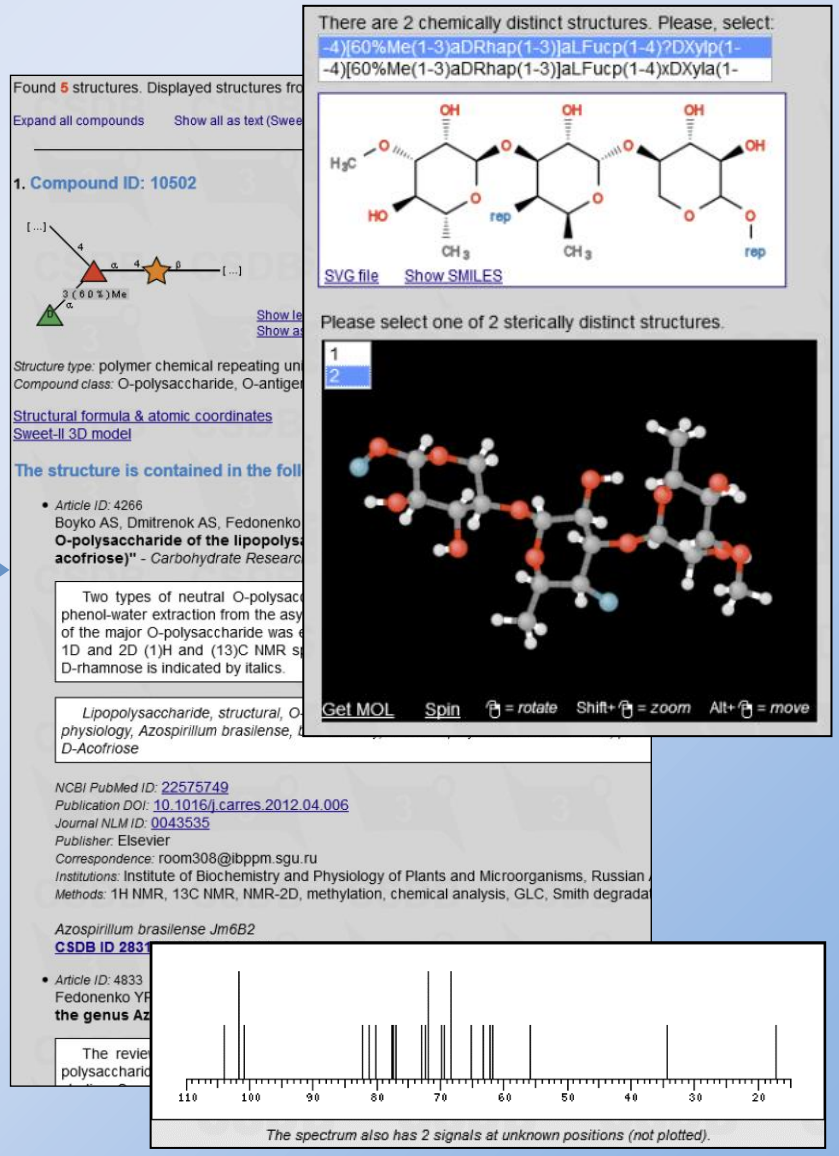
NMR filter

NOT

$\delta^{13}C$:
18.3 ± 0.8

IDs

Data arranged by compound, publication, organism, etc.

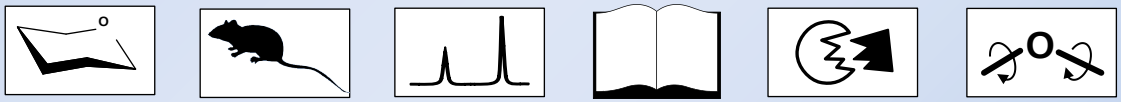


CSDB: Carbohydrate Structure Database

since 2005 | since 2012 | since 2017

Database set + platform for services

Zelinsky Institute
Moscow, Russia



25K | 13K | 14K | 10K | 2K | 3K

CSDB

- regular updates
- extensible architecture
- data analysis tools
- curated content (15% = Carbbank, 85% = literature)
- complete coverage (bacteria and fungi)
- integration with other DBs

Egorova K, Toukach Ph **CSDB_GT: a new curated database on glycosyltransferases** *Glycobiology 2017, 27:285-290*
Toukach Ph, Egorova K **Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts** *Nucl Acid Res 2016, 44:D1229-1236*
Egorova K, Toukach Ph **Expansion of coverage of Carbohydrate Structure Database (CSDB)** *Carbohydr Res 2014, 389:112-114*
Toukach Ph **Bacterial Carbohydrate Structure Database 3: principles and realization** *J Chem Inf Model 2011, 51:159-170*

Structural databases

All DBs

CarbBank — 23
- ☺ complete to 1995
- ☹ architecture, % errors
- ORIG

GlycomeDB — 39 → GLYTOUCAN — 99
- ☺ meta-repository
- ☹ poor annotations
- ☹ no aglycons

CFG glycan
- mammals, > 6

SweetDB, SugaBase → GLYCOSCIENCES.DE
- PDB
- 25 / 19

$ KEGG GLYCAN — 11

Eurocarb DB
- ☺ architecture
- ☹ model only

BCSDB — PFCSDB → Carbohydrate Structure Database
- 13 / 5 (bacteria, archaea)
- 8 / 3 (fungi, plants)
- ☺ complete on prokaryots
- ☺ curated
- ORIG
- GTR

$ GlycoSuite (By Proteome Systems)
- ORIG
- mammals+…
- 10 / 1
- ☺ complete to 2005

> 70 — jCGGDB AIST
- ☺ database collection
- ☹ poor annotations

4 / 1 — $ UniCarbKB
- LC
- ☺ curated

0.7 O- & N- — nibrt GLYCOBASE
- HPLC
- ORIG

44 — Glycoconjugate Data Bank (Organized by Hokkaido University)
- PDB

0.2 *E. coli* — EcoDAB (Escherichia coli O-Antigen Database)
- GTR
- ORIG

0.3 animals — $ GlycoBase
- ORIG

Special content

JCGGDB AIST

human glyco genes — GGdb — ~0.2

N- & O-glycan $MS^{2,3,4}$ — GMDB — ~0.2

glycochemical reactions — GlycoNAVI — ~3.0 (4.4 str)

conjugates & aglycones — GlycosideDB — >70

N-glycoproteins
C. elegans + mice — GlycoProtDB — ~2.5

protocols for
synthesis & analysis — GlycoPOD (GlycoScience Protocol Online Database) — ~0.2 (0.5 sub)

glyco-eitopes
& antibodies — GlycoEpitope — ~0.2 (0.6 ABs)

GlyTOUcan
(id repository) — ~99

binding to pathogens — SugarBindDB (Pathogen Sugar-Binding Database) — ~0.9

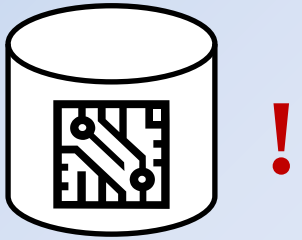
O-glycBase,
O- & C-glycoproteins — CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS — ~0.2

GlycoMaps,
computed conformational maps — ~2.6

MSDB
monosaccharides & notations — MONOSACCHARIDE DB — ~0.8

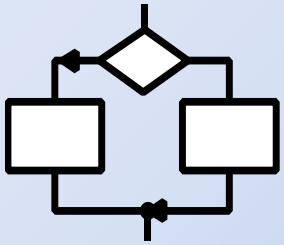
Evaluation criteria

- **Completeness** (+ chosen domain)
- **Data quality** (% of errors, human-readability)
- **Functionality** (data and index types, query processing)
- **Integration** (supported data formats, import & export, API, RDF)
- **Interface** (user-friendliness, stability, performance)



!

architecture



control scripts

data

recoverability

Architecture, functionality

- Relational database
- Indexation + standard indices (DOI, TaxID, ICD-11, PMID, Genbank, …)
- Structures, taxonomy, bibliography (different entries and data types)
- Human-readable dump (organization of data upload)
- Controlled term vocabularies (monomers; MSDB)
- ~~Free text~~
- Connection table

minimum

structure,
taxonomy,
bibliography,
cross-DB references

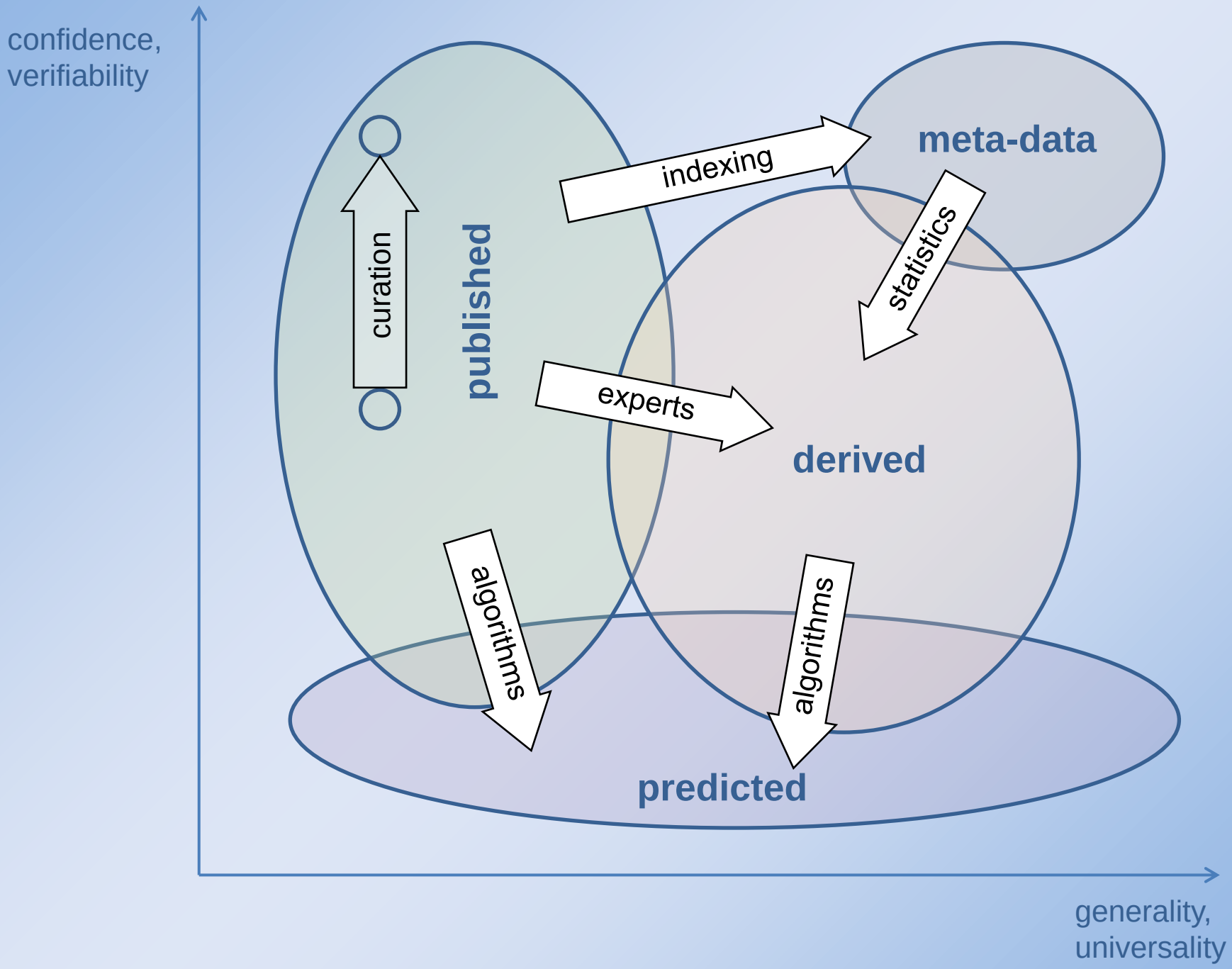
optional

trivial names,
NMR and MS spectra,
spectroscopic conditions,
bio-activity,
genes, enzymes,
conformation

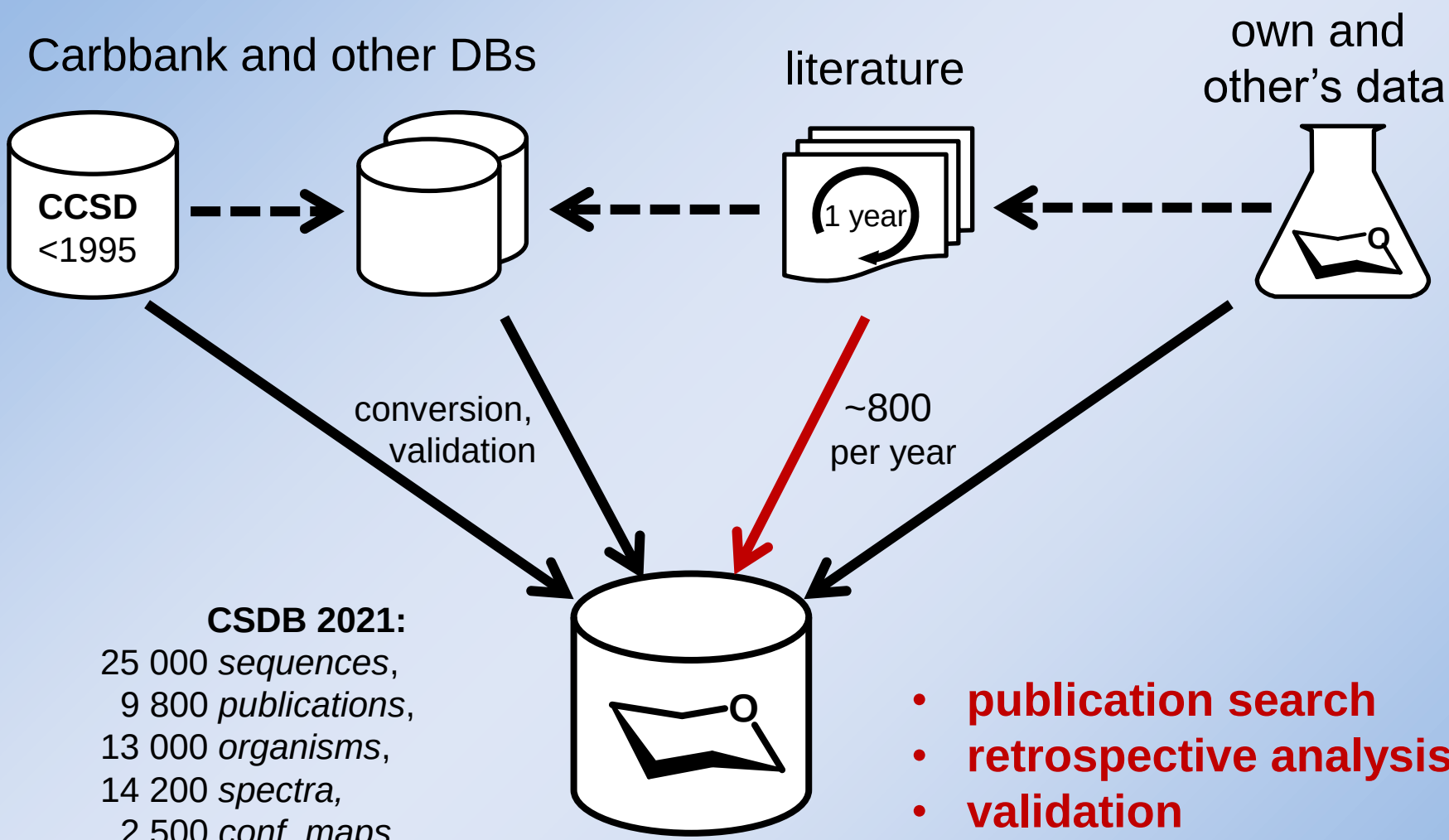
diseases,
organs, tissues,
genotype, life stage

keywords,
abstracts,
affiliations

Data levels


confidence,
verifiability
curation
published
indexing
meta-data
statistics
experts
derived
algorithms
algorithms
predicted
generality,
universality

Data sources



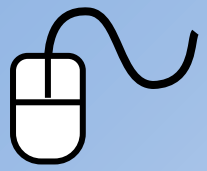
CSDB 2021:
25 000 *sequences*,
9 800 *publications*,
13 000 *organisms*,
14 200 *spectra,*
2 500 *conf. maps*,
2 100 *GT activities*
+interrelations

- **publication search**
- **retrospective analysis**
- **validation**

complete coverage:

negative search result = still valuable scientific information

Data quality

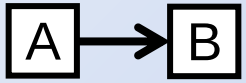


from annotators

from databases

found & corrected

from publications



in scripts

errors, inconsistency

correctable

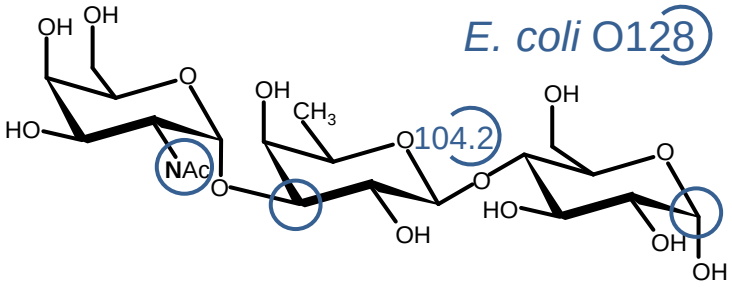
2dGlc → araHex,
α-Rib-ol → Rib-ol,
D-Kdo → Kdo,
1-methyl → 1-Me,
n.m.r. → NMR,
taxid 583 → Proteus,
…

detectable

Glc(1-2)GlcN,
anhydro-Kdo,
D-manHep,
Gal*p*5N,
Ac(1-2)[Glc(1-2)]Gal,
Escherichia sapiens,
Dev Food Sci 2012,
#Ac : 23ppm, 65 ppm,
D-Gcl, …

undetectable

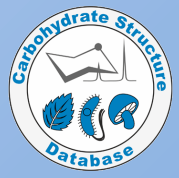
E. coli O127:
aDGal*p*N(1-**4**)bDFuc*p*(1-4)**b**DGlc*p*
Fuc C1 10**3**.2 ppm



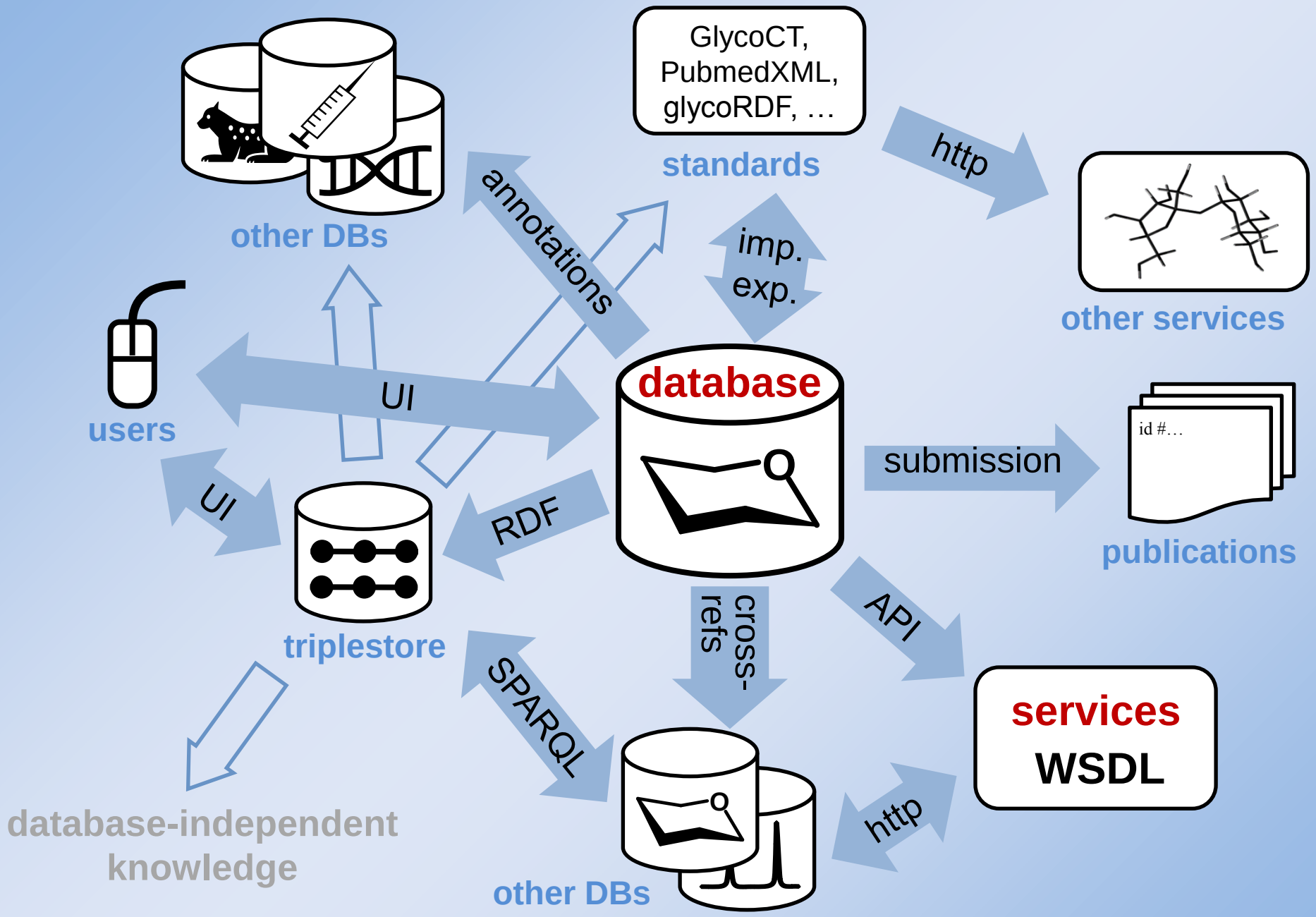
CarbBank ☹ **>50%** (incorrect, missing, falsely present structures, strains, annotations)

→ other DBs

☺ **<10%**



Ideal integration



Aoki-Kinoshita K, *et al.* **Towards an international glycan structure repository** *Glycobiology 2013, 23:1422-1424*
Toukach Ph, Joshi H, Ranzinger R, Knirel Y, von der Lieth C-W **Sharing of worldwide distributed carbohydrate resources** *Nucl Acid Res 2007, 35:D280*

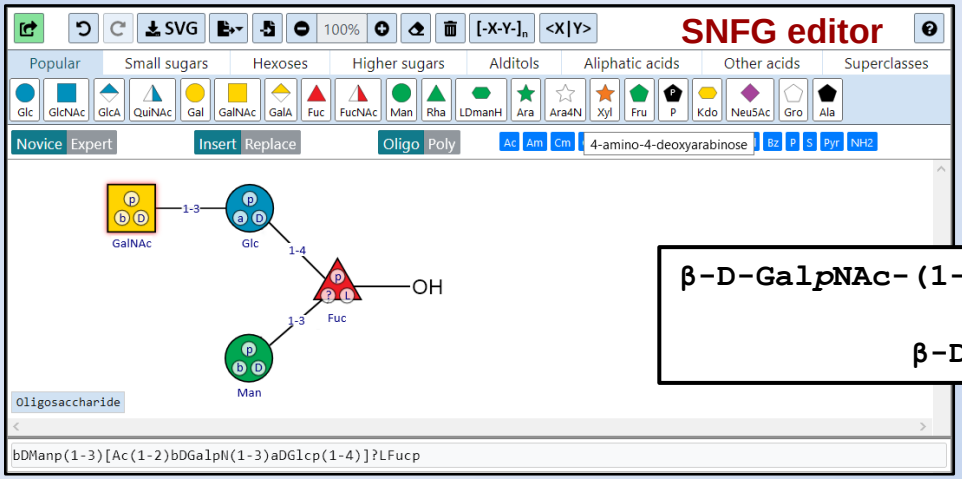
Interfaces

- Data conversion ↔ other formats
- Automated web-services (WSDL)
- Import, export
- Documentation, user's HELP
- User-friendliness, performance
- Links to other projects (queries, indices, data)
- Structure input & output

SNFG,
WURCS,
GlycoCT,
SMILES,
MOL, PDB,
GlydeII,
LinUCS,
Sweet-DB
GLYCAM,
GlycoRDF,
DCI XML, …

NCBI PubMed,
NCBI Taxonomy,
DOI,
Uniprot, Genbank,
Glycosciences.DE,
MonosaccharideDB,
Glytoucan, ICD-11

wizard,
graphic builder,
library,
CSDB Linear,
GlycoCT

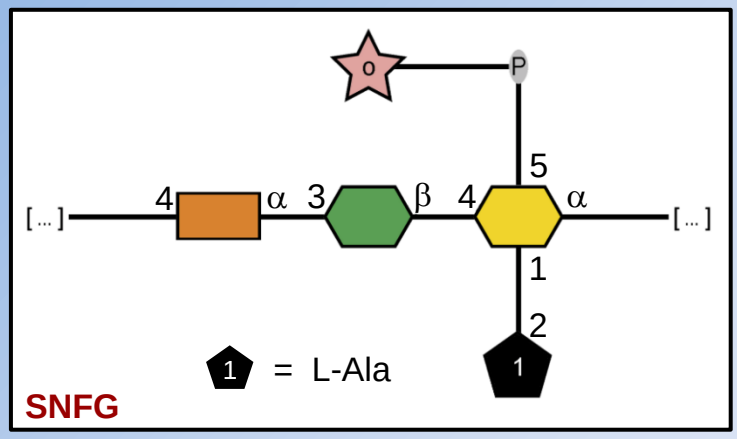


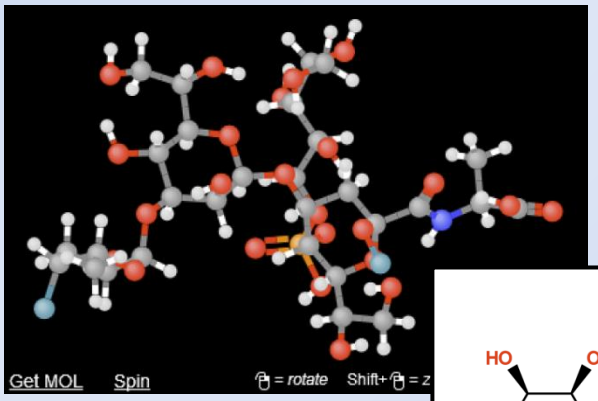
β-D-GalpNAc-(1-3)-α-D-Glcp(1-4)
β-D-Manp-(1-3)-L-Fucp

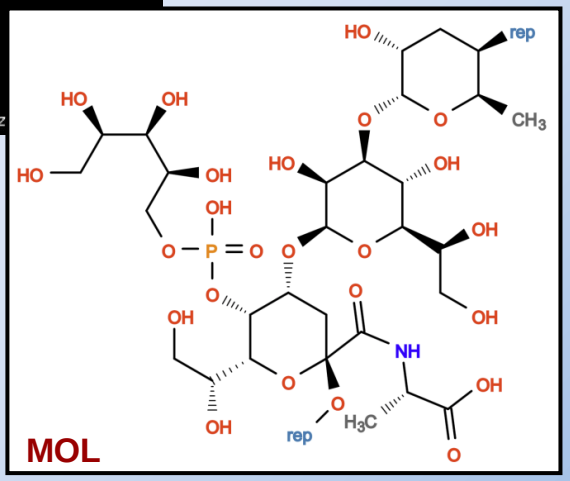
bDManp(1-3)[Ac(1-2)DGalpN(1-?)aDGlcp(1-4)]?LFucp CSDB

Structure output

Visualization in human-readable formats:








```
                                  D-Rib-ol-(1-P-5)┐
                                                  │
-4)-α-Abep-(1-3)-β-L-gro-D-manHepp-(1-4)-α-Kdop-(2-
                                                  │
                                  L-Ala-(2-1)┘
```

Sweet-DB

Export in machine-readable formats:

[*]O[C@]1(C(=O)N[C@@H](C)C(=O)O)C[C@@H](O[C@@H]2O[C@H]([C@@H](O)CO)[C@@H](O)[C@H](O[C@H]3O[C@H](C)[C@H]([*])C[C@H]3O)[C@@H]2O)[C@@H](OP(=O)(O)OC[C@H](O)[C@H](O)[C@H](O)CO)[C@@H]([C@H](O)CO)O1 **SMILES**

2.0/5,5,5/[Aad1122h-2a_2-6][h222h][a11221h-1b_1-5][a2d12m-1a_1-5][A1m_2*N]/1-2-3-4-5/a1-e2_a2-d4~_a4-c1_a5-b1*OPO*/3O/3=O_c3-d1 **WURCS**

```
-4)aXAbep(1-3)bXLDmanHepp(1-4)[xDRib-ol(1-P-5),xLAla?(2-1)]aXKdop(2-
```

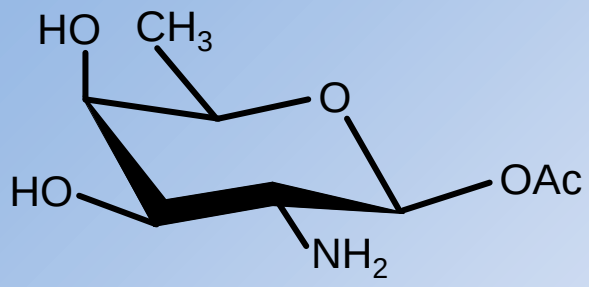
CSDB

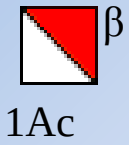
Chernyshov I, Toukach Ph **Automated translation of glycans from residue-based notation to SMILES and coordinates** *Bioinformatics 2017, 34:2679*

Varki A, *et al.* **Symbol nomenclature for graphical representations of glycans** *Glycobiology 2015, 25:1323-1324*

Alocci D, Suchankova P, Costa R, Hory N, Mariethoz J, Toukach Ph, Lisacek F **SugarSketcher: quick online glycan drawing** *Molecules 2018, 23: id 3206*

Nomenclature fuzziness





(CFG)

bDFucpN(1-1)Ac (CSDB) ← unambiguously maps to structure but, nevertheless, is human-readable

D-Fuc*p*N-β1OAc

beta-fucosamine acetate

1-acetoxy-beta-D-fucopyranosamine

2-deoxy-2-amino-β-D-fucopyranosyl acetate (IUPAC)

β-D-fucosamine acetic ester

β-6-deoxy-D-galactosamine acetate

b-dgal-HEX|1:5|2-amino|1-acetate (GlycoCT)

β-D-фукозамин-1-О-ацетат (another human language)

(2S,3R,4R,5R,6R)-3-amino-4,5-dihydroxy-6-methyltetrahydro-2H-pyran-2-yl acetate (IUPAC)

N[C@H]([C@H]([C@H]([C@@H](C)O1)O)O)[C@@H]1OC(C)=O (SMILES)

1S/C8H15NO5/c1-3-6(11)7(12)5(9)8(13-3)14-4(2)10/h3,5-8,11-12H,9H2,1-2H3/t3-,5-,6+,7-,8+/m1/s1 (InChI)

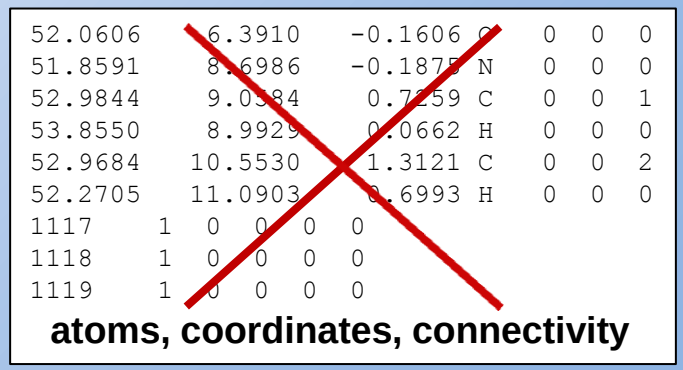
Existing molecular description, MOL?

- A lot of deep stones in translation of coordinates to residue-based notation
  (to publish like α-D-Gal*f*-(1-3)-β-D-Glc*p)*
- Difficult translation from a residue-based notation to coordinates
  => annotation laboriousness
- Cannot describe underdetermined structures
- Dara are not visually linked to knowledge

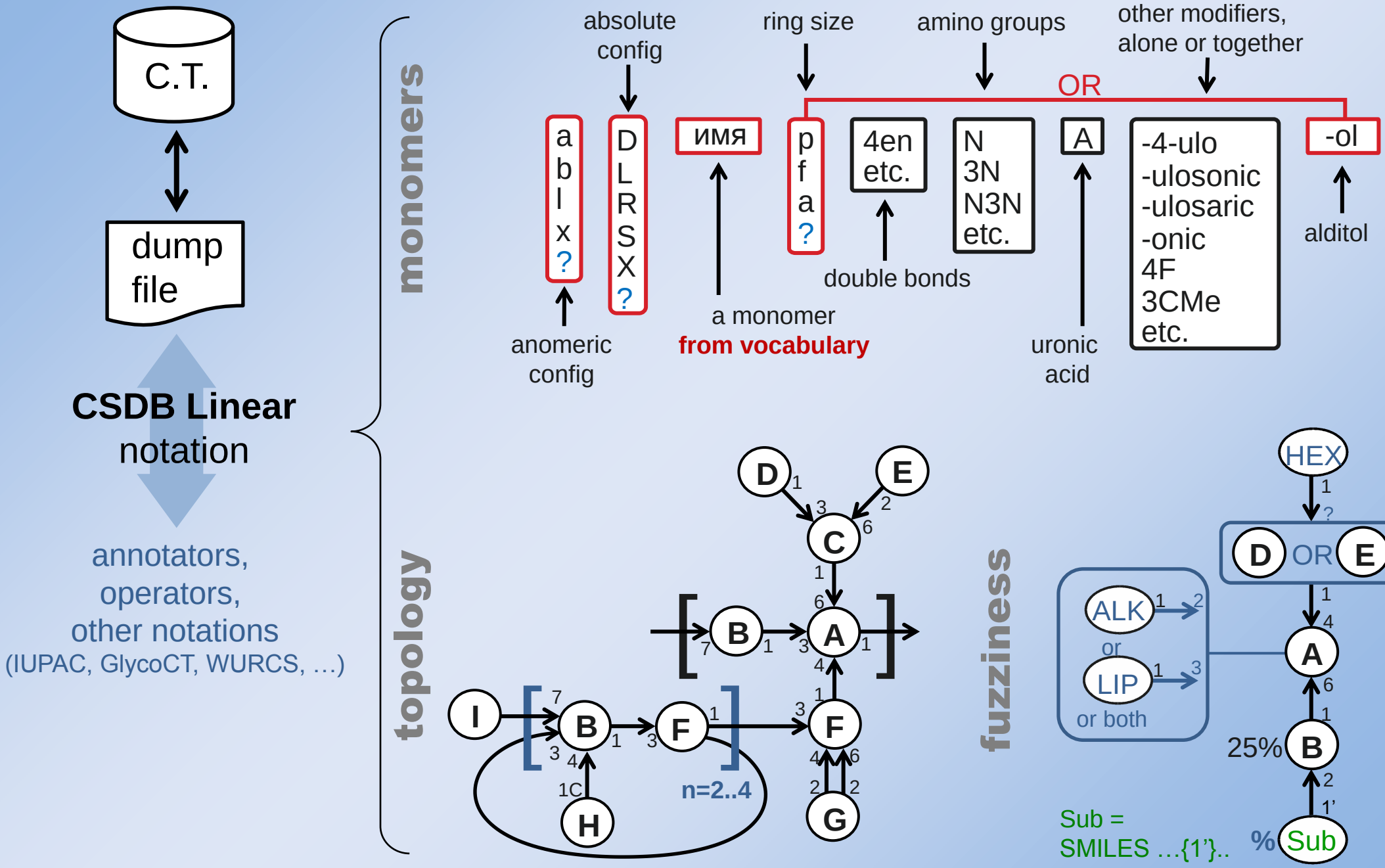
⇩

- Non-human-readable → difficult to curate → data errors
- Atom coordinates are not primary data
  but incomplete MOL (without 3D) has the same format as 3D MOL
- Large in storage and network transfer

  (and cannot be used as an URL part)



Structural features and notation



Structure abstraction levels

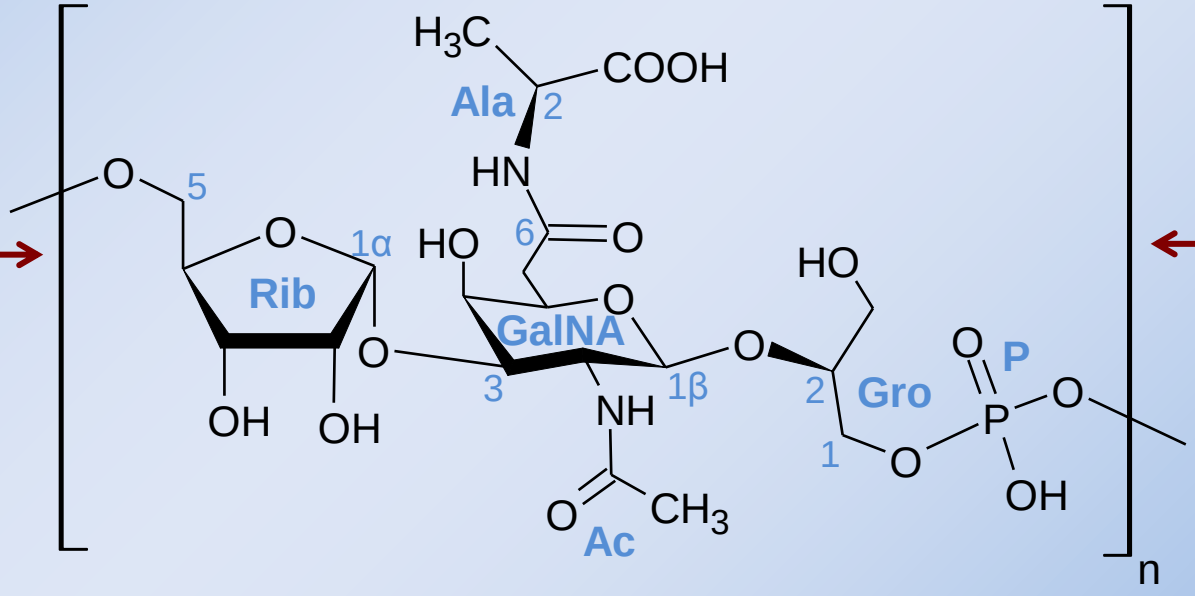
-3)[xLAla(2-6),Ac(1-2)]bDGalpNA(1-

exact fragment and its linkage

xDGro(1-P-5)aDRibf

unspecified topology and capping bonds

L-Ala(2-6)+
|
-5)α-D-Ribf(1-3)β-D-GalpNAcA(1-2)D-Gro(1-P-



-?)?Dhex(1-

only residue class and 1-linkage

approximate motif

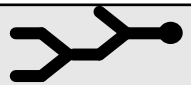
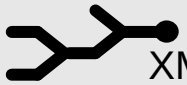
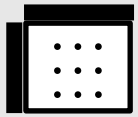
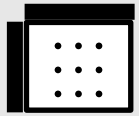
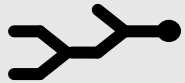
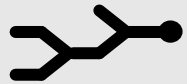
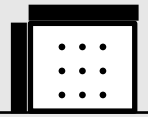
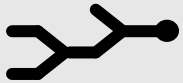
x?Ala(2-?)?DGal?NA

unspecified configurations, substitution positions, ringsize, N-acetylation

HEX, xDRib?, PEP

partial composition

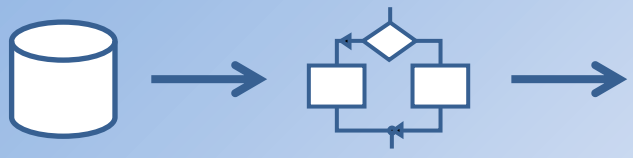
Notation comparison

	approach	**complete**	**unambi-guous**	**human control**	**parseable**	**fuzziness support**
IUPAC ⛔		○	◔	◕	○	◔
IUPAC extended (SweetDB, Carbbank) ⛔	pseudo-graphics	◔	◔	●	○	◔
Glyde I ⛔	 XML	◕	◕	○	● URL	◔
WURCS (JCGGDB, ChEBI, PDB)		◕	●	○	● URL	◕
GlycoCT (Glycome-DB)		●	●	○	●	◕
LinearCode (CFG)		○	◕	◔	● URL	◕
LinUCS (GlycoSCIENCES)		○	●	◕	● URL	◕
KCF (KEGG)		◔	●	○	●	◔
CSDB linear (CSDB)		◕	●	◕	● URL	◕

☹ compatibility

worse ○ ○ ◕ ● better

CSDB: add-ons



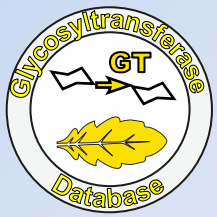
conclusions inferred on ALL data,
including screening and statistics

- Biosynthesis pathway analysis (glycosyltransferase database)
- Oligosaccharide conformation maps
- NMR simulation and spectrum assignment (^{13}C, ^{1}H, 2D)
- Structure prediction from NMR and other data
- Glycome-based taxon clustering
- Feature distribution in structures and taxa
- Monomer classification

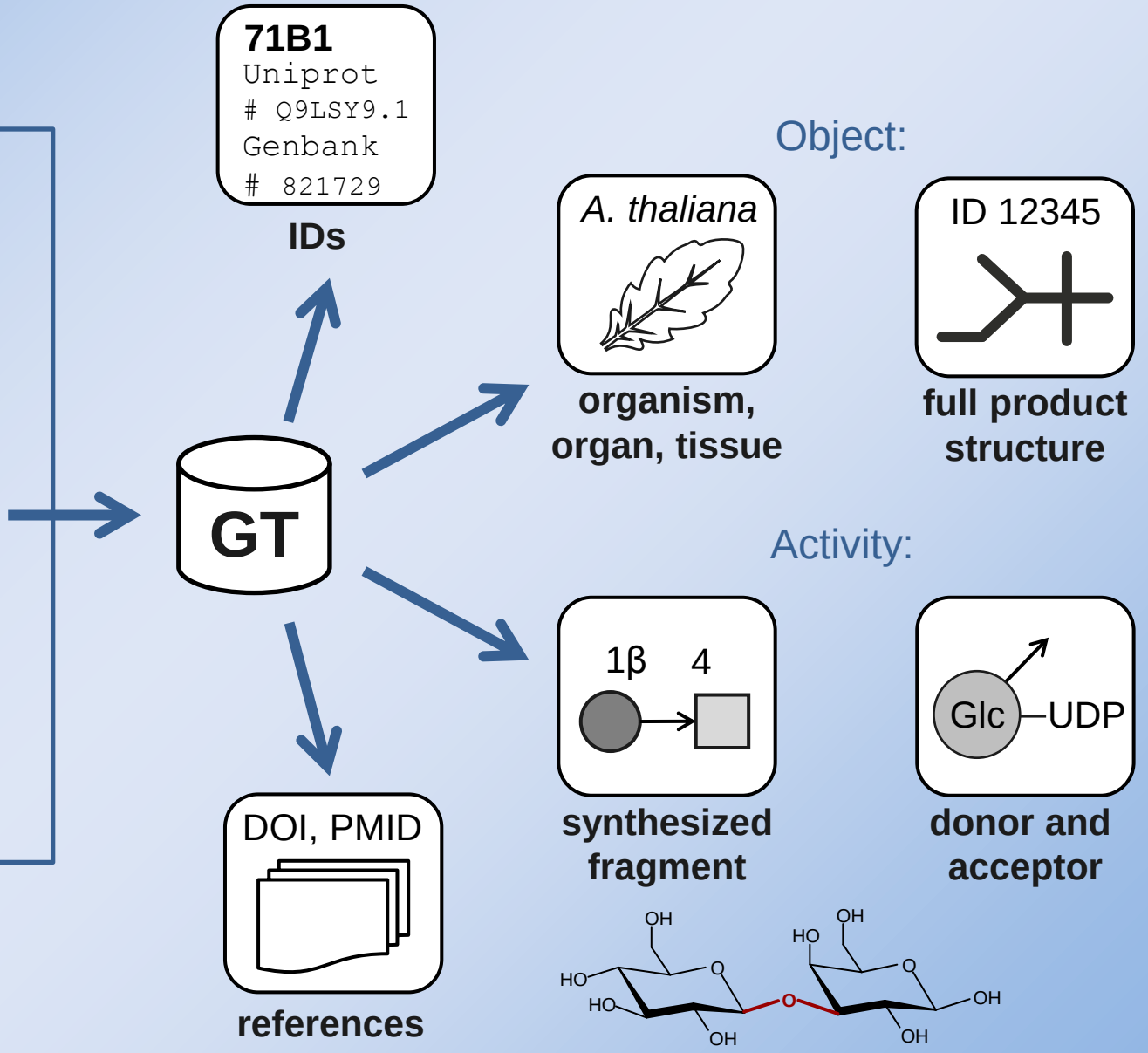
Glycosyltransferases

Criteria:
(in any combination)

- IDs in databases
- enzyme name / group
- gene name / cluster
- CAZy family
- organism (species, strain)
- synthesized linkage
- donor (or its fragment)
- acceptor (or its fragment)
- object cellular role
- trustworthiness level



CSDB GT

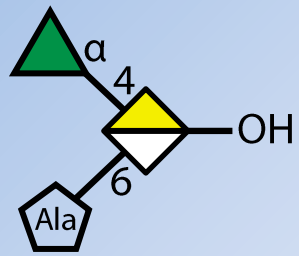


Egorova K, Toukach Ph **CSDB_GT: a new curated database on glycosyltransferases** *Glycobiology 2017, 27:285-290*
Egorova K, Toukach Ph **Expanding CSDB_GT glycosyltransferase database with *Escherichia coli*** *Glycobiology 2019, 29:285-287*
Egorova K, Smirnova NS, Toukach Ph **CSDB_GT, a curated glycosyltransferase database with close-to-full coverage on three most studied non-animal species** *Glycobiology 2020, ePub ahead of print*

Conformation analysis

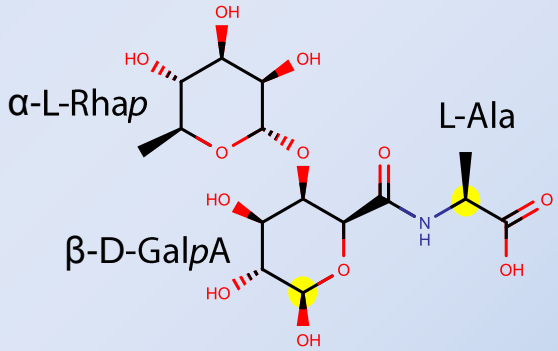
aLRhap(1-4)[x?Ala?(2-6)]?DGalpA

structure,
incl. incomplete

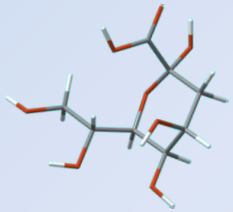


other variants
(α-GalA, D-Ala, etc.)

SMILES

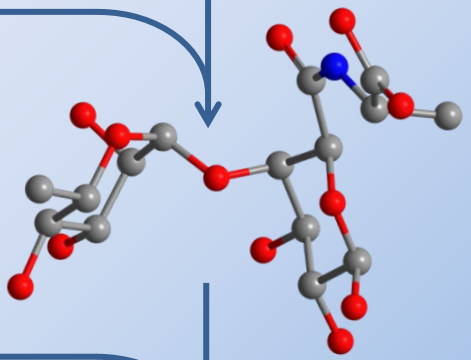


populated ring
conformers
~1000 residues

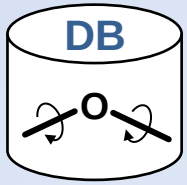


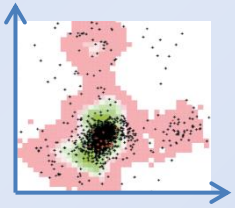
$^4C_1$,$^1C_4$, …

«chairification»



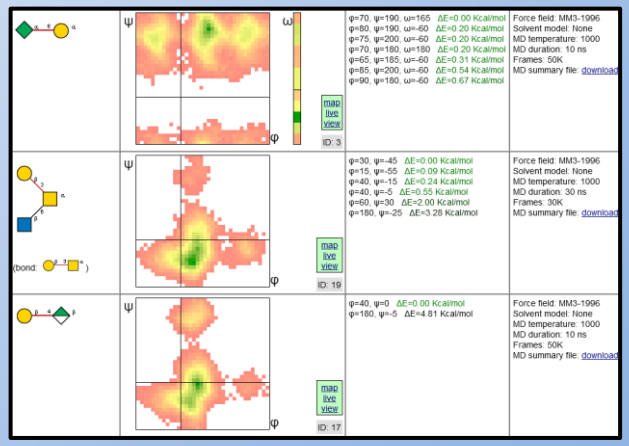
populated
bond torsions





φ,ψ,ω,θ

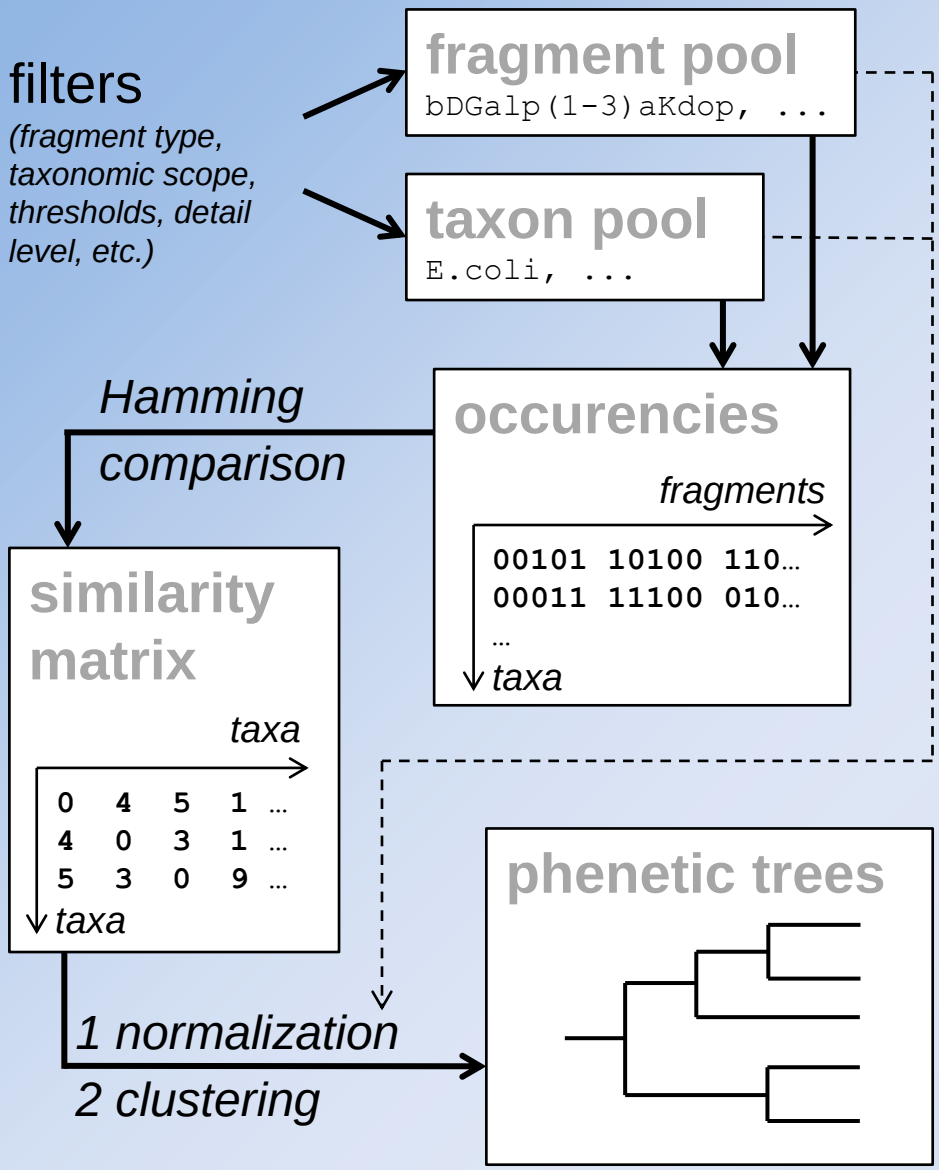
picking of minima
MM-relaxation

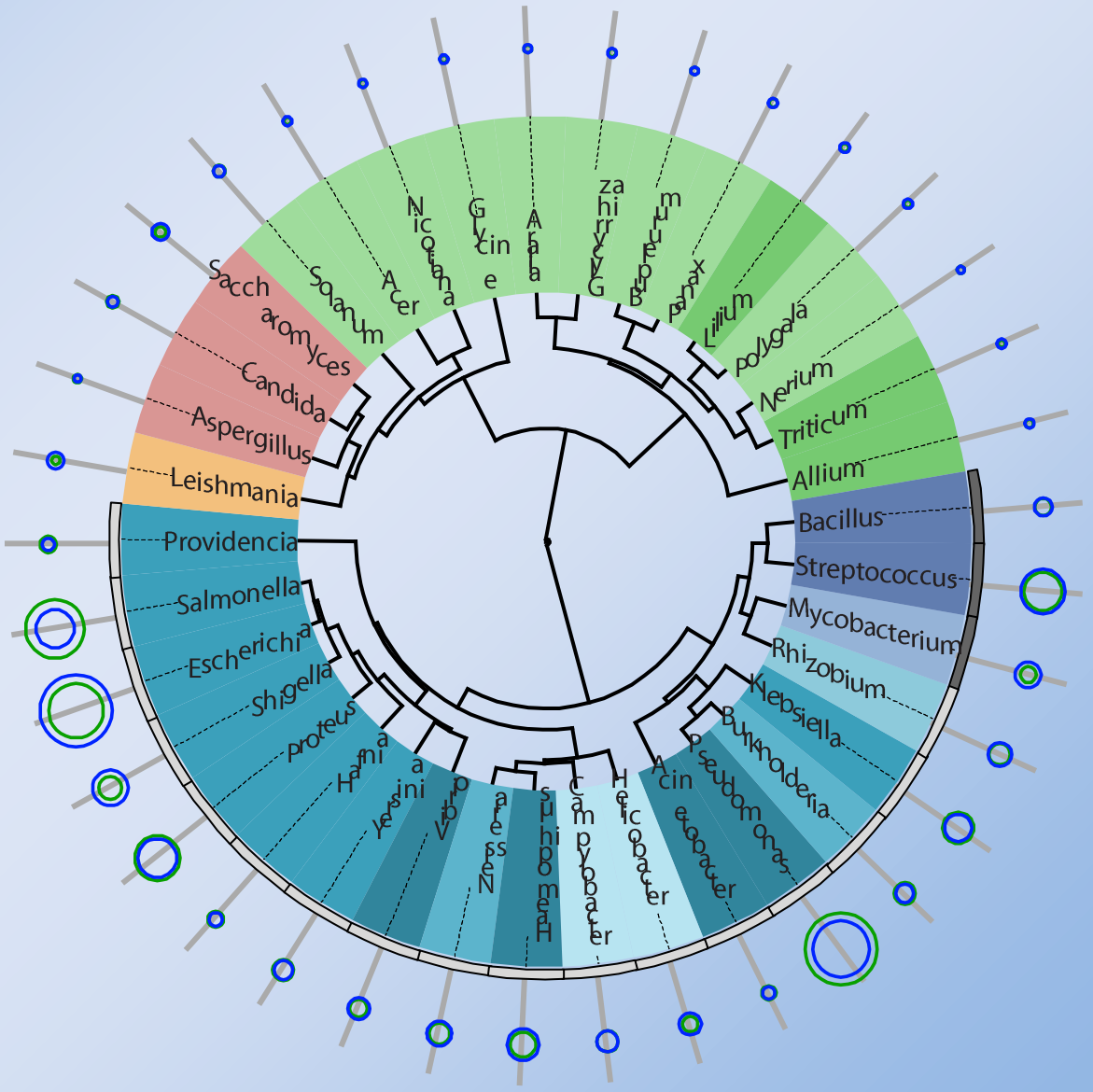


conformers
+ energies

mol. dynamics
300K, 100ns, H_2O

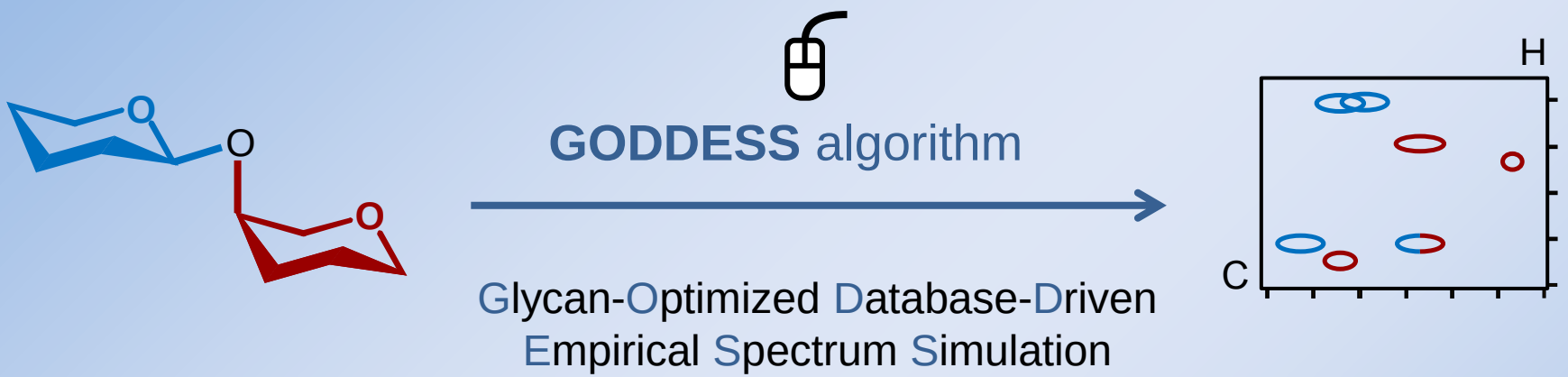
Taxon clustering





NMR simulation

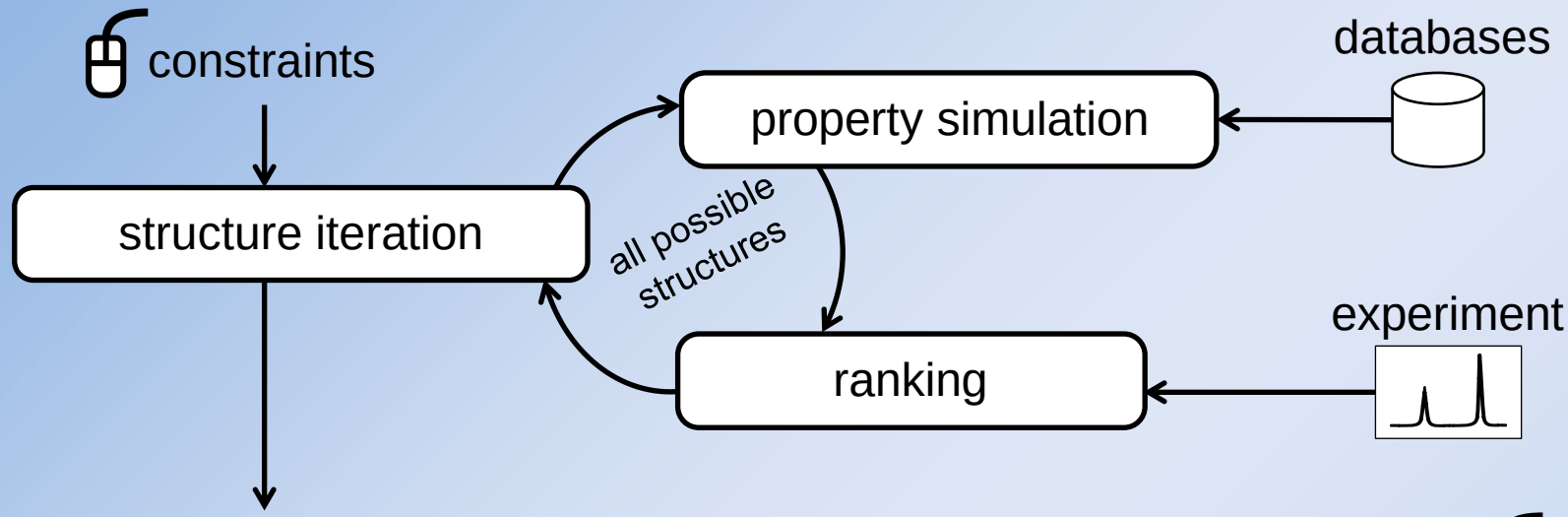
NMR is a main methods for sequence analysis in glycobiology



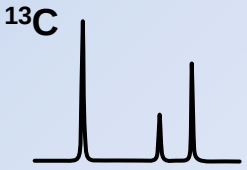
- Chemical shift simulation (^{13}C ~ 0.7 ppm, ^{1}H ~ 0.06 ppm)
- Support for structure elucidation
- Signal assignment and hypothesis validation
- Structure iteration and simulation-to-experiment comparison
- Verification of molecular geometry models

Toukach F, Ananikov V **Computational predictions of NMR parameters for structure elucidation of carbohydrates** *Chem Soc Rev 2013, 42:8376-8415*
Kapaev R, Egorova K, Toukach Ph **Carbohydrate structure generalization for simulation of experimental observables** *J Chem Inf Model 2014, 54:2594*
Kapaev R, Toukach Ph **Improved carbohydrate structure generalization scheme for ^{1}H and ^{13}C NMR simulations** *Analyt Chem 2015, 87:7006-7010*
Kapaev R, Toukach Ph **Simulation of 2D NMR spectra of carbohydrates using GODESS software** *J Chem Inf Model 2016, 56:1100-1104*

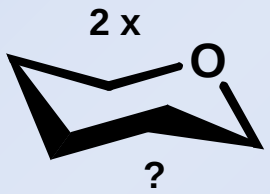
Structure iterator



unassigned NMR spectrum

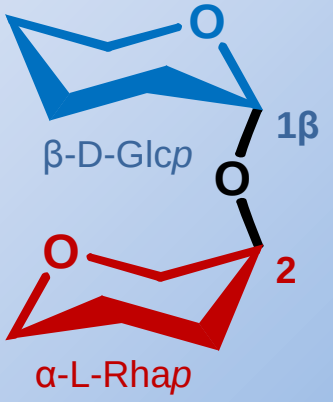


known data about structure



GRASS algorithm

Generation, Ranking and Assignment of Saccharide Structures



full structure (remaining unknowns)

monomers, configurations, modifications, linkage pattern, sequence

Perspective

● done in CSDB ● to do ● almost done

- Recognized human-readable language (SNFG, CSDB Linear, …)
- Cross-project data access (GlycoRDF, GlycoCoO, central triplestore)
- Cross-project services
  (structure input & output , conformational calculations, spectra simulation, …)
- Recognized indices
  (Glytoucan ID, MSDB, PMID, DOI, TaxID, ICD-11, PDB id, Genbank, …)
- Standard models and protocols (API, WSDL, SPARQL, …)
- Ideological replacement of Carbbank 
- Requirement to include IDs in publications (Glytoucan ID?)

(who will remove unpublished / erroneous data?)

Links and further reading

J. Abrahams et al. **Recent advances in glycoinformatic platforms for glycomics and glycoproteomics** (2020) *Curr Opin Struct Biol* **62**, 59-69. doi: 10.1016/j.sbi.2019.11.009

http://glytoucan.org

K. Aoki-Kinoshita **A practical guide to using glycomic databases** (2017) *Springer*. doi: 10.1007/978-4-431-56454-6

http://jcggdb.jp/index_en.html

T. Lütteke **The use of glyco-informatics in glycochemistry** (2012) *Beilstein J Org Chem* **8**, 915-929. doi: 10.3762/bjoc.8.104

http://glycosciences.de

http://www.genome.jp/kegg/glycan/

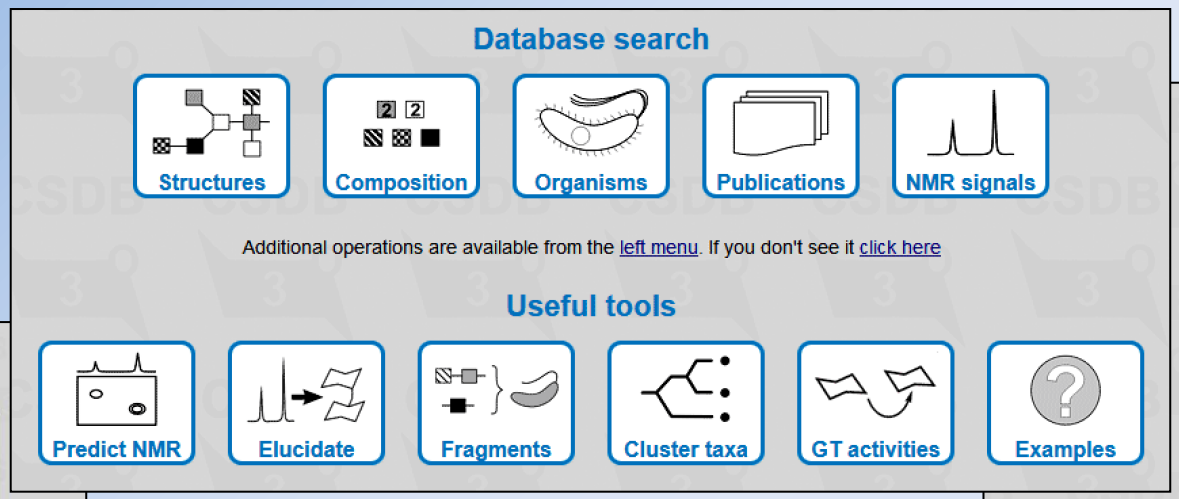
http://www.unicarbkb.org/

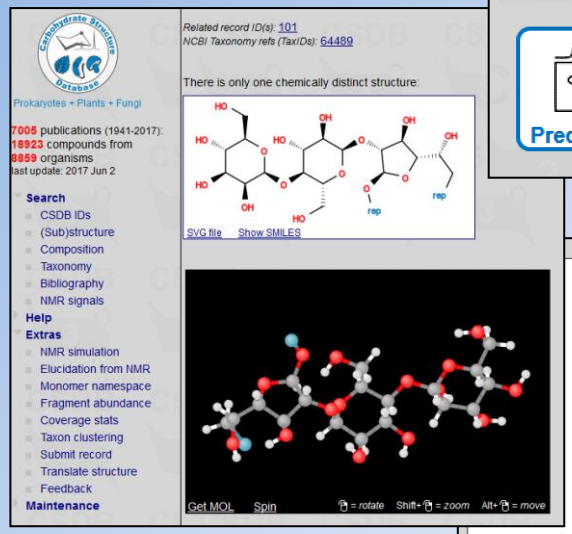
Ph. Toukach, K. Egorova **Carbohydrate Structure Database merged from bacterial, plant and fungal parts** (2016) *Nucl Acid Res* **44**, D1229–D1236. doi: 10.1093/nar/gkv840

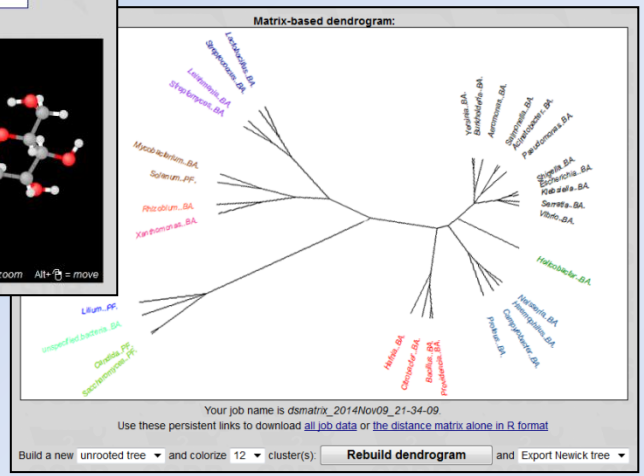
http://csdb.glycoscience.ru

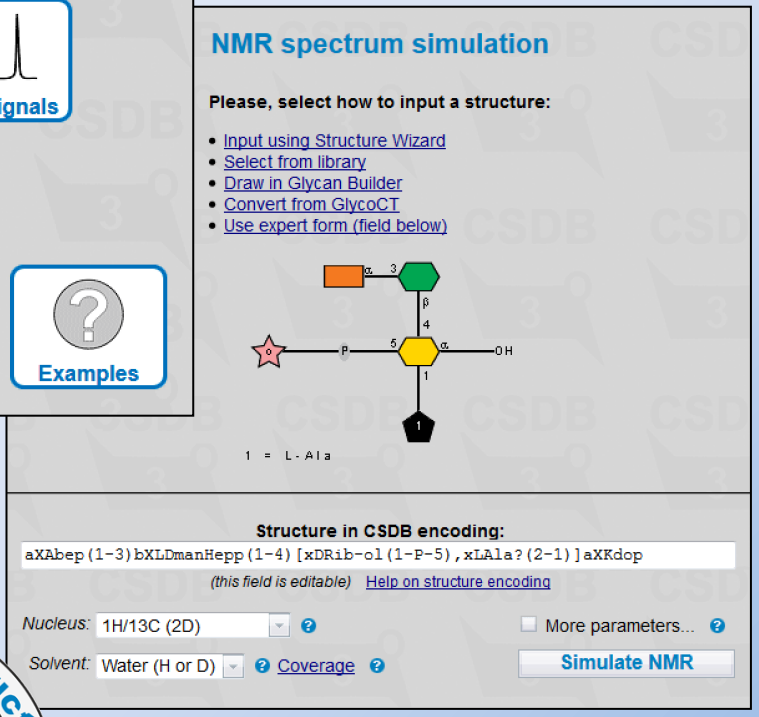
http://toukach.ru/rus/glyco-db.htm

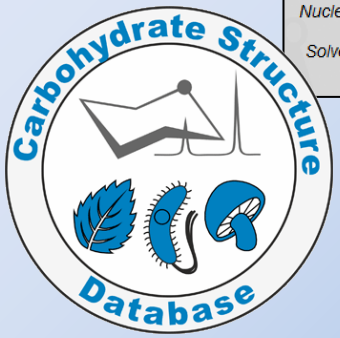
CSDB on the Internet











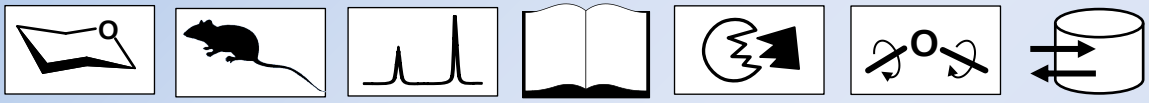
http://csdb.glycoscience.ru

- free access
- detailed manuals
- problem solution examples

Credits

Carbohydrate Structure Database

curated content, close-to-full coverage



Zelinsky Institute
Moscow, Russia

programming	Roman Kapaev, Andrei Bochkov, Ivan Chernyshov, …
literature processing & verification	Ksenia Egorova, Nadezhda Kalinchuk, Kirill Kazantsev, …
general support, data collection	Yuriy Knirel
integration, ontology	René Ranzinger, Kiyoko Aoki-Kinoshita, Thomas Lütteke, …
conformation analysis	Victor Stroylov, Sofya Scherbinina, …
ideas, R&D, notation, programming, interface, supervision	Philip Toukach
partners	GlycomeDB, JCGGDB, UniCarbKB, GLYCOSCIENCES.DE

Russian Foundation for Basic Research	International Science and Technonogy Center	Russian Federation President Grants	Deutsches Krebbs-ForschungsZentrum	Russian Science Support Agency	Russian Science Foundation
2005-2007, 2012-2020 (x3)	2004-2005	2005-2007	2007-2010 (x4)	2008-2009	2018-2020

Supplementary

(used for discussion)

Examples of queries to CSDB

- Study how an introduction of the amino group will affect the NMR chemical shifts of the lactose fragment
- Find bacterial glycans containing a galacturonic acid residue and at least one more hexose, published after 2005 in relation to antigens
- Find all compounds extracted from the plants of the genus *Solanum* which contain a solanidine constituent
- Find all carbohydrate structures having a signal close to 34 ppm in the ^{13}C NMR spectrum, except those containing any octose
- Find all papers by Knirel or Shashkov AS on bacterial structures containing quinovose-4-amine amidated by any N-acetylated amino acid
- Find all bacterial nonose monosaccharide structures (monomers or homopolymers)
- Predict ^{13}C NMR spectrum of 3-O-α-abequosyl-6-deoxy-β-D-mannoheptopyranosyl-(D-ribitol-1)-phosphate in water solution and explore the credibility of chemical shifts simulated with lowest reported trustworthiness
- Rank structural hypotheses for an unelucidated oligomer conforming to an experimental ^{13}C NMR spectrum and containing bacillosamine, galacturonic acid and lysine residues
- Study monomeric composition of two fungal species, *Aspergillus oryzae* and *Aspergillus fumigatus*, and reveal which monomers occupy termini of side chains
- Find which dimeric fragments (including sugars, aglycons and other residues) of higher plant carbohydrates are specific to lupins
- Study coverage statistics of *Proteobacteria*

Resource Description Framework

RDF is a model to store data as *object-predicate-subject* triples.

☺ allows federated queries with minimal knowledge of source DB formats

☹ needs a triplestore and agreed ontology

Question: find carrier protein data for any given glycan in JCGGDB.

Problem: JCGGDB does not have links to protein databases.

Preamble:
JCGGDB entries have links to GlycomeDB IDs.
Both GlycomeDB and UniCarbKB have structures in GlycoCT format.
UniCarbKB entries have links to UniProt IDs.

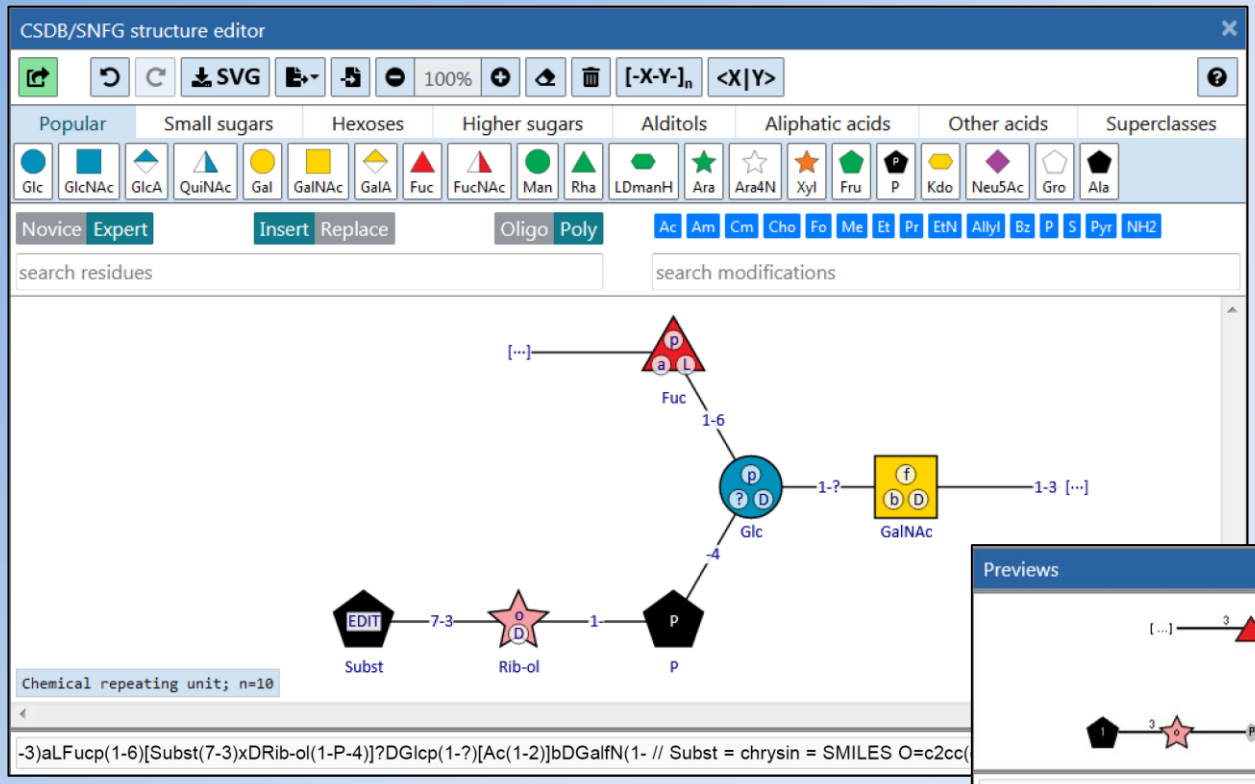
Solution *(9-line SPARQL script)*:
Map JCGGDB IDs to UniCarbKB IDs using GlycomeDB and retrieve the UniProt IDs from UniCarbKB for each JCGGDB ID.

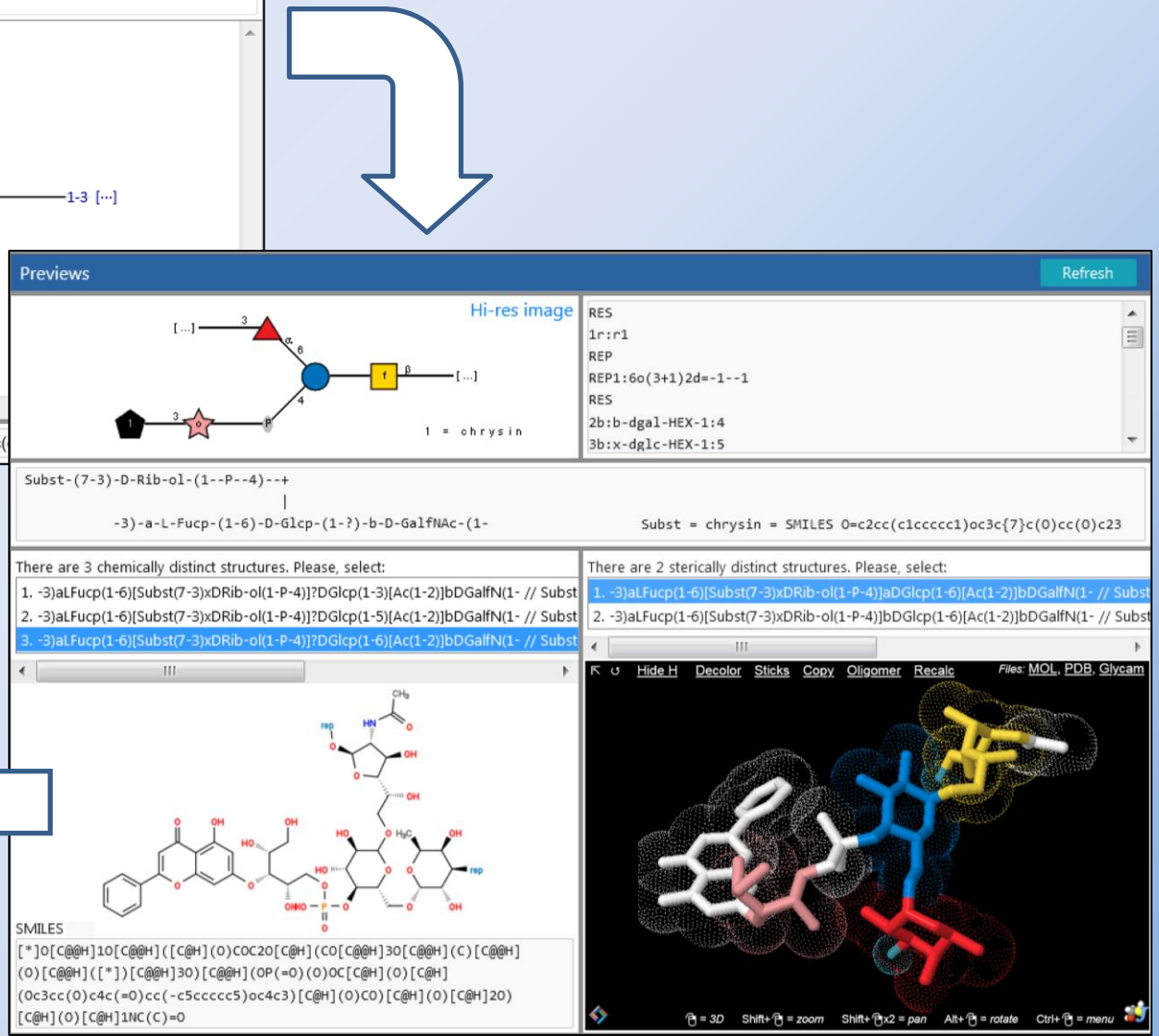
Need:
ontology standard → data exported to RDF → a triplestore → SPARQL endpoint

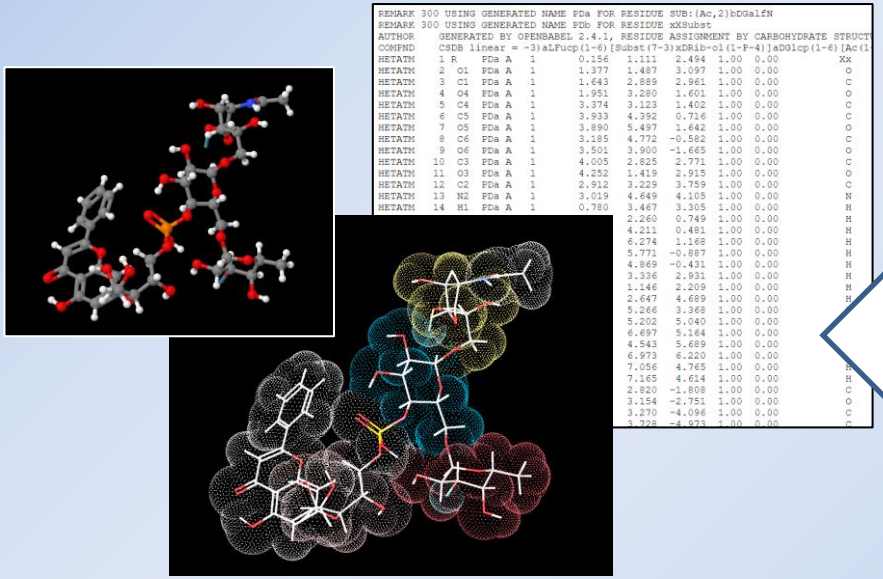
GlycoRDF is a first formal carbohydrate ontology (OWL),
GlycoCoO is its extension for glycoconjugates.

Ranzinger R, *et al.* **GlycoRDF: An ontology to standardize glycomics data in RDF** *Bioinformatics 2015, 31:919-925*
Aoki-Kinoshita K, *et al.* **Introducing glycomics data into the Semantic Web** *J Biomed Semantics 2013, 4:id39*

Structure input and output





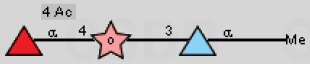


(Sub)structure search

Structure wizard

Topology: 3 residues (linear: A->B->C) Ⓐ→Ⓑ→Ⓒ

Structure: aLFucp4Ac(1-4)xDRib-ol(1-3)a?6dTal?(1-1)Me



Residue Ⓐ: aLFucp4Ac(1-

a L fucose (pyranose)
aLFucp
substitutes C4 of Residue B
is terminal

add substitution
add substituent Acetylated at 4
add substituent
add substituent
add substituent

Residue Ⓑ: xDRib-ol(1-

D ribose (alditol)
DRib-ol
substitutes C3 of Residue C

add substituent
add substituent
add substituent
add substituent

Residue Ⓒ: a?6dTal?(1-1)Me

a ? 6-deoxytalose (?)
a?6dTal?
has aglycon : methyl

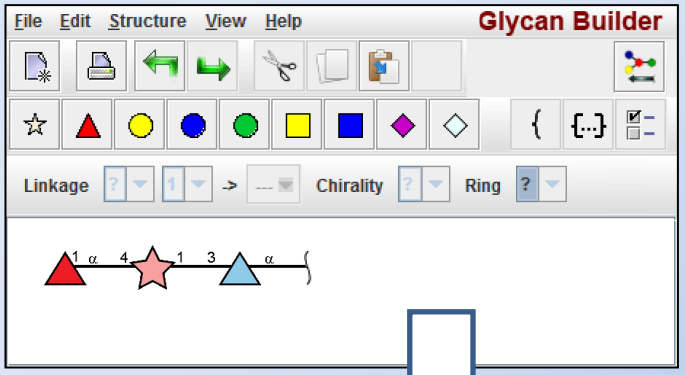
add substituent
add substituent
add substituent
add substituent

Structure in CSDB encoding:

Ac(1-4)aLFucp(1-4)xDRib-ol(1-3)a?6dTal?(1-1)Me

Return the structure to the search page and close this window

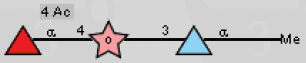
Home Help



Search for (sub)structure

Please, select how to input structure:

- Input using Structure Wizard
- Select from library
- Draw in Glycan Builder
- Convert from GlycoCT
- Copy from the previous query (aDFucp3N)
- Use expert form (field below)



Structural fragment in CSDB encoding:

Ac(1-4)aLFucp(1-4)xDRib-ol(1-3)a?6dTal?(1-1)Me

(this field is editable) Help on structure encoding

Only those containing text: factor-2A' in aglycons, aliases or linear code in trivial names

Search scope:

Search the whole database
Search in the result of the previous query (logical AND)
Combine with the result of the previous query (logical OR)
Negate search (find results NOT matching current query)

Treat search term as a fragment (substructure)
Search for molecule types: All molecule types
Search for structures with published NMR data only
Restrict compound class: inner core component
Restrict taxonomical domain: All prokaryotes

Previous results: 122 structures: <ID list>

Go! & display 30 records per page.

Predict NMR Sweet 3D model GLYCAM model Home Help HELP !!!

Organism search

Found **12** organisms. Displayed organisms from **1** to **12**

Expand all organisms Show all as text (SweetDB notation)

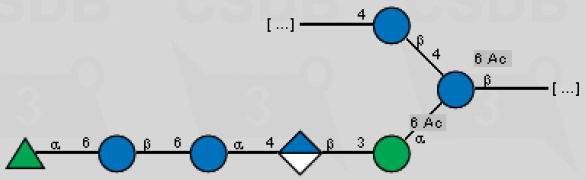
1. (*Organism ID:* 1005)

Acetobacter xylinum
(Ancestor NCBI TaxID 28448, species name lookup)

Later renamed to: *Komagataeibacter xylinus*
Taxonomic group: bacteria
Phylum: Proteobacteria

The following compound(s) are assigned to this organism:

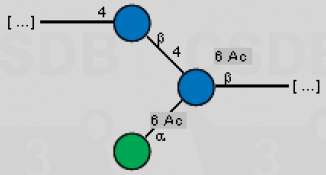
- *Compound ID:* 1717



Show legend
Show as text

Carbohydrate Research 2004, "Synergistic interactions between the genetica
bacterial polysaccharide P2 and carob or konjac mannan"
CSDB ID 9262 (all data & tools)

- *Compound ID:* 1720



Show legend
Show as text

Carbohydrate Research 2004, "Synergistic interactions between the genetically modified
bacterial polysaccharide P2 and carob or konjac mannan"
CSDB ID 9414 (all data & tools)

Search for organism

Display domains: ☑ bacteria ☑ archaea ☑ protista ☑ algae ☑ fungi ☐ plants ☐ animals

Genus: Absidia, **Acetobacter**, Acholeplasma, Acidithiobacillus, Acinetobacter, Acremonium, Actinobacillus

Species: *Any*, sp., diazotrophicus, methanolicus, tropicalis, **xylinum**

Specify: *

Strain / subspecies: *Any*, B42, CKE5, CKEP, CR1/4, IFO 13693

Search scope:

- (•) Search the whole database
- ( ) Search in the result of the previous query (logical AND)
- ( ) Combine with the result of the previous query (logical OR)
- ☐ Negate search (find results NOT matching current query)

☐ Search among HOST organisms
☐ Use NCBI taxID
☑ Include subtaxons

Previous results: 6 structures: <ID list>

Go! & display 30 records per page.

List of organisms Home Help

Process taxonomy in NCBI Taxonomy DB (fields are editable):

Genus: Acetobacter *Species:* xylinum Process

Bibliography search

Found 3 publications. Displayed publications from 1 to 3

Expand all publications Show all as text (SweetDB notation)

1. (*Article ID:* 1525)

Knirel YA, Lindner B, Vinogradov EV, Shaikhutdinova RZ, Senchenkova SN, Kocha

Cold temperature-induced modifications to the composition and str of Yersinia pestis

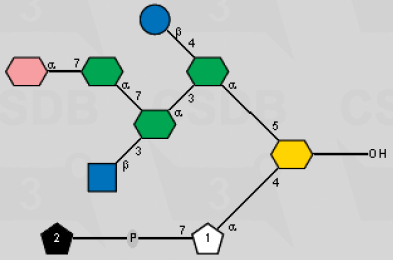
Carbohydrate Research **340(9)** (2005) 1625-1630

Following a report of variations in the lipopolysaccharide (LPS) structure degrees C) and flea (25 degrees C) temperatures, a number of changes to the the bacterium was cultivated at a temperature of winter-hibernating rodents (6 known Y. pestis LPS types, LPS of a new type was isolated from Y. pestis KM21 the latter differs in: (i) replacement of terminal galactose with termin phosphorylation of terminal oct-2-ulosonic acid with phosphoethanolamine; (iii) the absence of glycine; lipid A differs in the lack of any 4-amino-4-deox (di)oxygenation of a fatty acid(s). The data obtained suggest that cold temp mechanism of control of the synthesis of Y. pestis LPS.

Lipopolysaccharide, structure, core, modification, agent, composition, Yersinia pestis, Plague

The publication contains the following compound(s):

- *Compound ID:* 4209

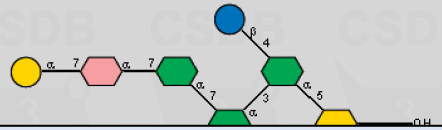


1 = a-Kop
2 = Et N

Show legend
Show as text

Yersinia pestis KM218
CSDB ID 10076 (all data & tools)

- *Compound ID:* 4210



Search for bibliography

Authors: "Knirel YA" OR Toukach Author index start with: tou

Help on author/keyword query syntax ä ö ü á é í ó č š

Title: pestis OR plague* search also in abstract

(content of title) Help on title/abstract query syntax

Keywords: structure? OR composition? search also in title

(content of keyword section) Help on author/keyword query syntax

Journal:
Carbohydrate Letters
Carbohydrate Polymers
Carbohydrate Research
Cell
Cell Chemical Biology
Cell and Tissue Research

Year: = > < 1983 1984 1985 1986 1987 1988

Vol: *

Page: *

Search scope:

- Search the whole database
- Search in the result of the previous query (logical AND)
- Combine with the result of the previous query (logical OR)
- Negate search (find results NOT matching current query)

Previous results: 3 publications: 2953,201,1525

Publications with structure elucidation only

Restrict taxonomical domain: All domains

Go! & display 30 records per page.

PubMed XML Home Help

Author index:

Toubetto K Toussaint A
Toukach FV

The listed author names start with 'Tou'.
Click an author name to copy it to the author field in the caller form.

Close this window

Conformation search

Search for disaccharide conformation maps

Use the following criteria alone or in any combination to search for conformation maps.

Conformation ID: *Type CSDB conformation ID or range, e.g. 1-5,10,12*
ANY
Search!

Model: *Use selectors or wizard*
α | L-Rhap | →4 | β | D-GlcpNAc
(only those components are listed for which conformation maps are stored)

or type dimeric fragment in CSDB encoding
aLRhap(1-4)[Ac(1-2)]bDGlcpN

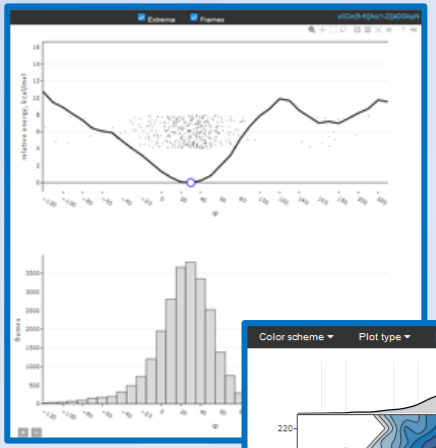
Force field: *Filter by MD method*
MM3-2000

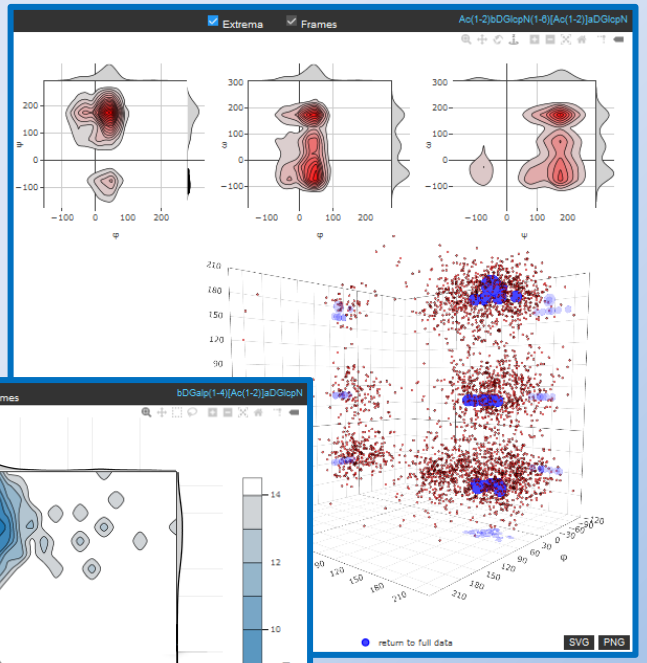
Temperature: *Filter by MD temperature*
1000

Solvent model: *Filter by solvent model*
any

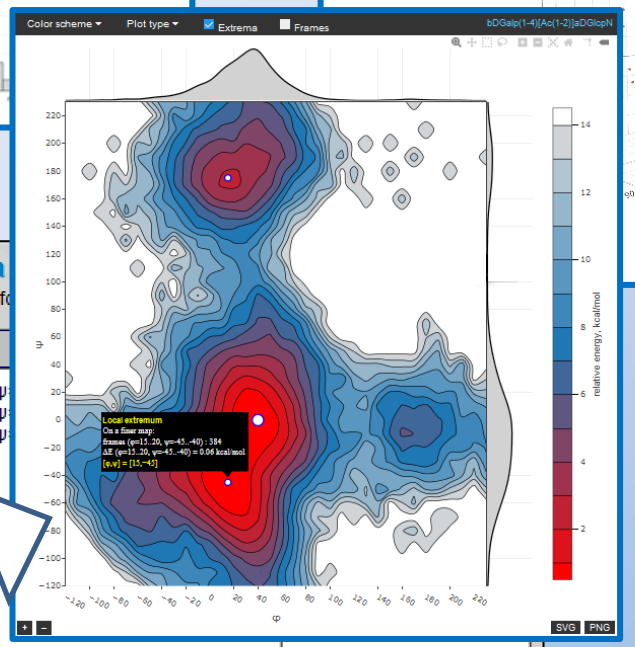
Home Help

1D





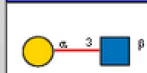
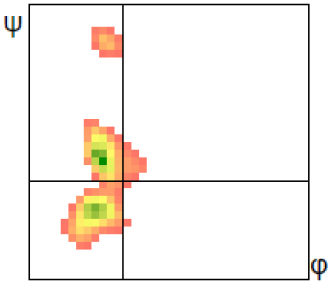
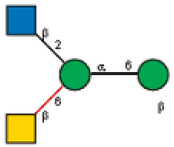
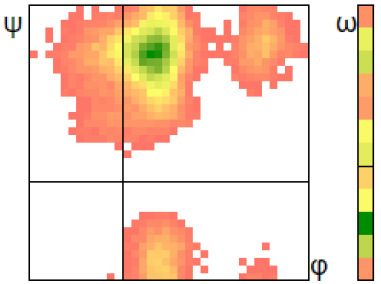
3D



2D

CSDB conformation data

12 conformation maps have been f

Model structure	Conformation map		
	 map live view ID: 1610	φ=-30, ψ= φ=-35, ψ= φ=-25, ψ=	
 (bond: )	 map live view ID: 907	φ=25, ψ=165, ω=-75 ΔE=0.00 Kcal/mol φ=45, ψ=170, ω=-60 ΔE=0.00 Kcal/mol φ=45, ψ=195, ω=-75 ΔE=0.38 Kcal/mol φ=50, ψ=185, ω=-60 ΔE=0.72 Kcal/mol φ=25, ψ=165, ω=180 ΔE=0.72 Kcal/mol φ=30, ψ=155, ω=-60 ΔE=0.72 Kcal/mol φ=35, ψ=180, ω=180 ΔE=0.85 Kcal/mol φ=45, ψ=215, ω=-60 ΔE=0.85 Kcal/mol φ=40, ψ=170, ω=165 ΔE=0.99 Kcal/mol φ=20, ψ=185, ω=165 ΔE=0.99 Kcal/mol φ=55, ψ=195, ω=165 ΔE=0.99 Kcal/mol φ=55, ψ=155, ω=-60 ΔE=1.13 Kcal/mol φ=45, ψ=185, ω=165 ΔE=1.13 Kcal/mol φ=30, ψ=145, ω=-60 ΔE=1.13 Kcal/mol	Force field: MM3-1996 Solvent model: None MD temperature: 1000 MD duration: 30 ns Frames: 30K MD summary file: download

Glycosyltransferase search

CSDB glycosyltransferase search

Use the following conditions alone or in any combination to search for glycosyltransferazes.
Any field may be left blank for no restrictions.

GT names and IDs: Enzyme name — *Type enzyme name, e.g. "Orf10". Wildcards (* and ?) are supported.* ANY

Organism: *Select species* Escherichia coli — *Type strain/serogroup*

Molecule role: *Filter by target structure* O-antigen

Synthesized bond: *Type dimeric fragment in CSDB encoding or use tools* ANY(1-?)?DGlcpN Use Wizard

Donor & acceptor: *Type donor CSDB encoding or use tools* ANY Use Wizard

Type acceptor CSDB encoding or use tools ANY Use Wizard

Treat donor/acceptor as fragments

Confirmation status: *Filter results to those* evidenced or predicted

Search!

Home Help HELP !!!

CSDB glycosyltransfer

42 glycosyltransferase activities ha
Please note that GTR database covers only two species: *A*

Enzyme	Gene	Activity		
Name: WbbD UniProt ID: Q03084*	?	Synthesized dimer: bDGalp(1-3)aDGlcpN Donor (ID 19342): DGalp(1-P-P-5)nucU Acceptor (ID 19715): Ph(1-11)[Ac(1-2)aDGlcpN(1-P-1)]Subst // Subst = undecan-1,11-diol Status: evidence *in vitro* Confirmation methods: *in vitro* (crude extract) ID: 2053	Organism (ID 8 Full structure (I 1 = u Analog	
Name: WbbG UniProt ID: Q0H8C8*		Synthesized dimer: aDGlcp(1-3)aDGlcpN Status: indirect evidence *in vivo* Confirmation methods: mutation (knockout) Notes: Repeating unit of the O148 antigen. ID: 2151	Organism (ID 8 Full structure (I CSDB ID(s): 11572, 21578, 23062, 26257, 27289 Molecule role: O-antigen	
Name: WbaD UniProt ID: Q1L815*	Name: wbaD GenBank ID: 7156002*	Synthesized dimer: bDManp(1-3)aDGlcpN Donor (ID 19855): DManp(1-P-P-5)nucG	Organism (ID 1863): *Escherichia coli* O77 Full structure (ID 4600):	**Zhou et al. 2016** DOI: 10.1016/j.carres.2016.02.007 Wang et al. 2007 DOI: 10.1099/mic.0.2007

Taxon clusterization

Scope settings

Limit taxonomical scope to: phylum

Display groups: ☑ bacteria ☑ archaea ☑ protista ☐ algae ☐ fungi ☐ plants ☐ animals

Phylum: (select multiple with CTRL key)

(unspecified bacteria)
(unspecified protista)
Actinobacteria
Bacteroidetes
Chlamydiae
Chloroflexi
Crenarchaeota
Cyanobacteria

General settings

species — **Rank** of taxons to compare (should be lower than selected scope). Specify exact species *(all)*

☐ 50 — **Taxon population threshold.** Minimal number of structures* assigned to a taxon or its subtaxons, to include this taxon in calculation (affects selection of taxons). Check to use this filter.

☑ 15 ‰ — **Normalized taxon population threshold.** Minimal part of structures* assigned to a taxon or its subtaxons, to include this taxon in calculation (affects selection of taxons). Normalized by the total number of structures* in the database. Check to use this filter.

50 — **Structure abundance threshold.** Minimal number of structures* in which a fragment should be contained to be qualified as 'present in biota' (affects selection of fragments)

60 — **Fragment abundance threshold.** Minimal number of instances* in which a fragment should be present to be qualified as 'present in biota' (affects selection of fragments)

2 — **Fragment presence threshold.** Minimal number of instances* in which a fragment should be present in organisms of a taxon to be qualified as present in this taxon (affects occurence codes and thus, taxon dissimilarity)

two residues — **Type** of fragments to analyze (dimeric or monomeric)

only polymers — **Type of structures** to analyze. Only structures of this type are considered in fragment analysis and where marked by (*). 'Optimized' = only polymers from bacteria, archaea and fungi, and only mono/oligomers from plants.

R-project — **Format** of the dissimilarity matrix

Fragment pool generation settings

☐ **Combine anomeric forms.** All sugar residues will be treated as 'any anomer'

☑ **Exclude underdetermined residues.** Residues with unknown anomeric, absolute or ringsize configuration will be omitted from analysis.

☐ **Exclude monovalent residues.** Residues like Me, Ac, etc. will be omitted from analysis. Please note, that Ac in N-acetylated amnosugars is a separate residue.

☑ **Exclude superclasses.** Fragments with residues represented by aliases and superclasses will be omitted from analysis.

☐ **Differentiate aliases.** Residue aliases (used for atypical residues) will be differentiated by actual residue names, otherwise they are combined under an alias name.

☑ **Sugars only.** Fragments with non-sugar residues (including monovalent residues, like N-acetyls) will be omitted from analysis.

☑ **Exclude aglycons.** Fragments with atypical residues at non-reducing ends will be omitted from analysis.

☐ **Differentiate location.** The same fragments at different locations (inline, terminal, reducing) will be treated as different.

☑ **Strict comparizon** of fragments. Unknown configurations and ringsizes are always unequal to those known (otherwise a fuzzy comparizon is performed.

Distance matrix based on fragment presence

The analysis was performed over all cellular organisms

```
Prepared 20 monomers
Prepared 32 genera
Generated occurence bit-codes. Show
Generated dissimilarity matrix. Show
```

Calculation parameters:

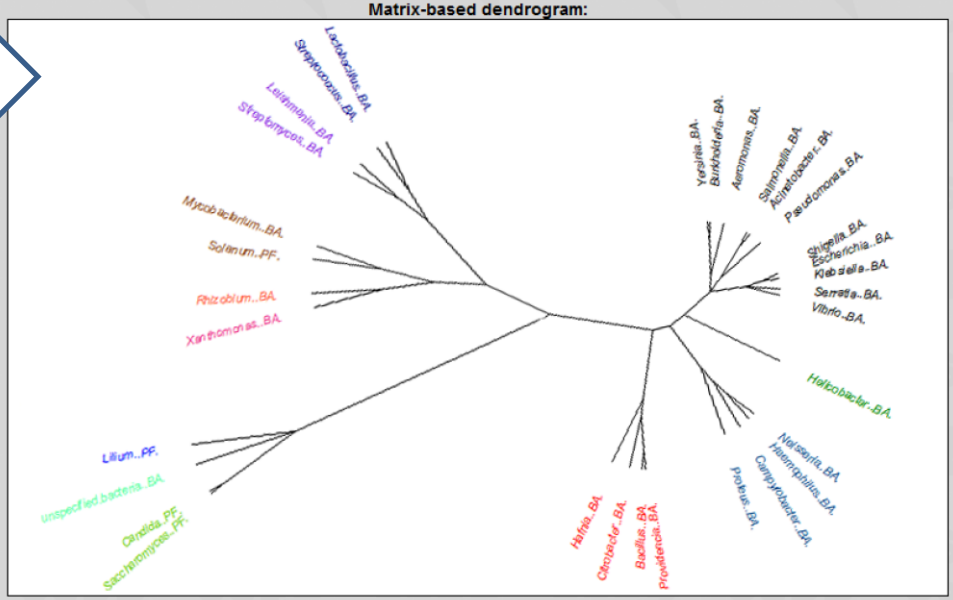
```
Hamming mode: YES
Fragment size: monomer
Fragment abundance filter: instance threshold: 550
Fragment abundance filter: structures threshold: 500
Fragment presence threshold: 2
                                  (atypical residues): NO
                structures of this type: any
Filter: differentiate monomer positions (inline/terminal/reducing) in structures: NO
Matrix data format: R
```

Coverage data on used taxons:

(taxons, number of organisms in a taxon, number of structures assigned to these organisms)

```
Acinetobacter (BA)      68      140
Aeromonas (BA)          64      122
Bacillus (BA)           95      234
Burkholderia (BA)
                        36      219
Solanum (PF)            46      127
```

Matrix-based dendrogram:



Your job name is *dsmatrix_2014Nov09_21-34-09*.

Use these persistent links to download all job data or the distance matrix alone in R format

Build a new unrooted tree and colorize 12 cluster(s): **Rebuild dendrogram** and Export Newick tree

Structure prediction

Structure generation constraints:

*The structure contains **6** residue(s):* Add residue

	α / β	D / L	Residue	Ring form
1.	?	D	galact-2N-uronic acid	pyranose
2.			acetic acid	
3.		D	show all residues	
4.			phosphoric acid	
5.	α	?	any octose	pyranose
6.		L	alanine	

Allowed linkages:

	C1	C2	C3	C4	C5	C6	C7+
D-Gal*p*NA			✓	✓		✓	None
Ac	✓						
D-Rib-ol	✓	✓	✓	✓		None	
P							
α-Oct*p*		✓	✓	✓	✓	✓	None
L-Ala	✓	✓		None			

Advanced options: Hide

	Min in	Max in	Location	Ac at N	Acceptors	Remove
D-Gal*p*NA	1	2	any	demanded	any	☒
Ac					1	
D-Rib-ol	?	?	any		any	☒
P			terminal		3,5	☒
α-Oct*p*	?	?	reducing			☒
L-Ala	?	?	any	forbidden	1	☒

Search depth	Scope	Advanced scope
Widespread structures only	✓ oligomers ☐ polymers⚠	β-anomers: = 1	CH_2 carbons: ?	☐ no furanoses

Find best matching structures:

Experimental ^{13}C NMR spectrum in water (24 signals of 24 expected):

```
17.4 22.9 34.7 50.5 52.4 63.9 64.9 66.2 68.3 70.6 72.3 72.4 72.7 73.3 73.6 76.5
78.6 78.8 99.2 102.6 171.2 175.2 176.0 176.5
```

± 2 signals

✓ Find 15 best-fitting structures

✓ Save generated structures

Go!

E-mail for results: (why?) user@gmail.com

Top 15 matches:

#Rank **Mean deviation** Linear correlation RMS deviation Trustworthiness	Structure Experimental spectrum Simulated spectrum Comments
#1. **Δ ~ 0.94 ppm** Corr = 1.000 RMS dev = 1.46 ppm Trust = 46%	Sim assignment Structure as text 1 = b-D-GalpNA 2 = L-Ala Expt: Sim:
#2. **Δ ~ 0.95 ppm**	
#15. **Δ ~ 1.42 ppm** Corr = 0.999 RMS dev = 1.99 ppm Trust = 49%	Sim assignment Structure as text 1 = b-D-GalpNA 2 = L-Ala