GLYCOTwinning Scientific Network Meeting
&
2nd Summer School 2024

Structural Glycobioinformatics

Serge Perez

What is structural glycobioinfomatics?

Structural glyco-bioinformatics is a specialized field within bioinformatics focused on studying and analyzing glycans and their interactions with proteins, lipids, and other biological molecules.

This field combines principles of structural biology, bioinformatics, and glycobiology to understand the roles and structures of glycans in biological processes.

By leveraging bioinformatics techniques, structural glyco-bioinformatics aims to elucidate the complexities of glycan biology, contributing to advancements in areas such as drug discovery, disease diagnostics, and biotechnology.

What is structural glycobioinfomatics?

- **Glycan Structure Analysis**: Understanding the complex structures of glycans, which are often branched and diverse. This involves determining the monosaccharide composition, linkages, and 3D conformations of glycans.
- **Database Development**: Creating and maintaining databases that store information about glycan structures, glycosylation sites on proteins, and glycan-related biological functions.
- **Glycan-Protein Interactions**: Studying how glycans interact with proteins, is crucial for understanding processes like cell-cell recognition, signaling, and immune responses. (1) modeling, (2) predicting binding sites and affinities.
- **Glyco-proteo-lipidomics**: Integrating glycan data with proteomics and lipidomics to study glycoproteins and their roles in various biological processes. (1) identifying glycosylation sites, (2) characterizing the glycan moieties attached to proteins,….
- **Computational Tools and Algorithms:** Developing and utilizing computational tools and algorithms to analyze glycan structures, predict glycan functions, and model glycan interactions. These tools help visualize glycan structures and simulate their dynamics.
- **Structural Prediction and Modeling**: Using computational methods to predict the 3D structures of glycans and their complexes with other biomolecules helps understand the functional implications of glycan structures and modifications.
- **Functional Annotation:** Annotating glycan structures with functional information, such as their roles in disease, their involvement in biological pathways, and their evolutionary significance.
- **Data Integration:** Integrating glycan data with other biological data types, such as genomic, transcriptomic, and proteomic data, to provide a holistic view of glycan functions in various biological contexts.

What are the sources of data for structural glycobioinfomatics?

1. EXPERIMENTAL DATABASES

- **Protein Data Bank (PDB):** The primary repository for 3D structural data of proteins and nucleic acids obtained through X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy.
- **Electron Microscopy Data Bank (EMDB):** A repository for 3D electron microscopy maps and models of macromolecular complexes and organelles.
- **Biological Magnetic Resonance Data Bank (BMRB):** Stores NMR spectroscopy data of biomolecules,….
- **Small Angle Scattering Biological Data Bank (SASBDB):** A repository for small-angle scattering data from X-ray (SAXS) and neutron (SANS) experiments.

2. SEQUENCE DATABASES

- **UniProt:** Provides comprehensive information on protein sequences and functional information.
- **Pfam:** A database of protein families that includes their alignments and hidden Markov models (HMMs).
- **NCBI GenBank:** A nucleotide sequence database that provides data on the sequences of DNA, RNA, and proteins.
- **CAZY:** Families of structurally related catalytic and carbohydrate-binding modules of Carbohydrate-Active enzymes

3. FUNCTIONAL ANNOTATION DATABASES

- **The Gene Ontology (GO)** knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

ALPHA Fold Protein Structure Database

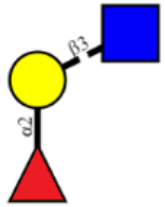
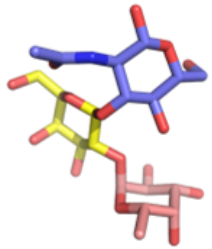
Symbol Nomenclature for Graphical Representation of Glycans

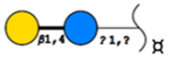
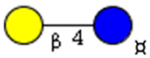
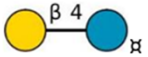
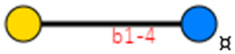
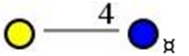
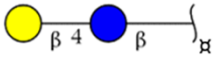
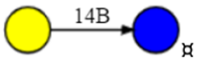
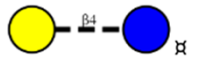
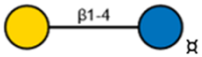
SHAPE	White (Generic)	Blue	Green	Yellow	Orange	Pink	Purple	Light Blue	Brown	Red
Filled Circle	Hexose	Glc	Man	Gal	Gul	Alt	All	Tal	Ido	
Filled Square	HexNAc	GlcNAc	ManNAc	GalNAc	GulNAc	AltNAc	AllNAc	TalNAc	IdoNAc	
Crossed Square	Hexosamine	GlcN	ManN	GalN	GulN	AltN	AllN	TalN	IdoN	
Divided Diamond	Hexuronate	GlcA	ManA	GalA	GulA	AltA	AllA	TalA	IdoA	
Filled Triangle	Deoxyhexose	Qui	Rha		6dGul	6dAlt		6dTal		Fuc
Divided Triangle	DeoxyhexNAc	QuiNAc	RhaNAc			6dAltNAc		6dTalNAc		FucNAc
Flat Rectangle	Di-deoxyhexose	Oli	Tyv		Abe	Par	Dig	Col		
Filled Star	Pentose		Ara	Lyx	Xyl	Rib				
Filled Diamond	3-deoxy-nonulosonic acids		Kdn				Neu5Ac	Neu5Gc	Neu	Sia
Flat Diamond	3,9-dideoxy-nonulosonic acids		Pse	Leg		Aci		4eLeg		
Flat Hexagon	Unknown	Bac	LDmanHep	Kdo	Dha	DDmanHep	MurNAc	MurNGc	Mur	
Pentagon	Assigned	Api	Fru	Tag	Sor	Psi				

Glycobiology,**(2015) 25,** 1323-1324

A. VARKI, R.D. CUMMINGS, M. AEBI, N.H. PARKER, P.H. SEEBERGER, J.D. ESKO, P. STANLEY, G. HART, A. DARVILL, T. KINOSHITA, J.J. PRESTEGARD, R.L. SCHNAAR, H.H. FREEZE, J.D. MARTH, C.R. BERTOZZI, M.E. ETZLER, M. FRANK, J.F.G. VLIEGENTHART, T. LUTTEKE, S. PEREZ, E. BOLTON, P. RUDD, J. PAULSON, M. KANEHISA, P. TOUKACH, K.F. AOKI-KINOSHITA, A. DELL, H. NARIMATSU, W. YORK, N. TANIGUCHI & S. KORNFELD,

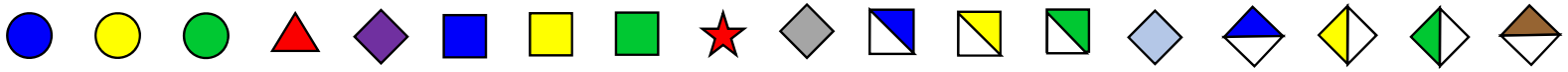
Terminology for Displaying Glycans Structures

Sketching (drawing)	• the most basic rendering • easy to make, understand • conveys a general idea • does not need to be correct	
Building	• advanced rendering • complex to put together • conveys precise data • correct, for scientific purposes	
Displaying	• molecular (3D) representation • complex to put together, render • conveys scientific data: angles,(x,y,z) coordinates • correct, closer to reality	
Virtual Reality	upon reaching a certain point, it may be interesting to "interact" with the generated structures	

Software¤	Image·(representation)¤
SugarSketcher¤	
Sugarbind·(Glycan·Builder)¤	
DrawGlycan-SNFG¤	
DrawRINGS¤	
LiGraph¤	
Glycam·(Carbohydrate·Builder)¤	
Polys·Glycan·Builder¤	
CHARMM-GUI·(Glycan·Reader·&·Modeler)¤	
Glycano¤	
GlycoGlyph¤	
Glyco.me·(SugarBuilder)¤	

Computational tools for drawing, building, and displaying carbohydrates: a visual guide. *Beilstein J. Org. Chem.* **2020,** *16,* 2448–2468.

Extending the Symbolic Representation of Monosaccharides



<anomeric prefix><prefix for absolute configuration><the monosaccharide code><suffix for ring configuration>_[<O-ester and O-ether substitutions>].

Residue Letter Name: Rib, Ara, Xyl, Lyx, All, Alt, Glc, Man, Gul, Ido, Gal, Tal,….

[O-ester and ethers]: (when present) are shown attached to the symbol with a number, e.g.
6Ac for 6-*O*-acetyl group, 3S for 3-*O*-sulfate group
6P for 6-*O*-phosphate group, 6Me for 6-*O*-Methyl group
36Anh for 3,6-anhydro, Pyr for pyruvate group

Absolute Configuration: D or L

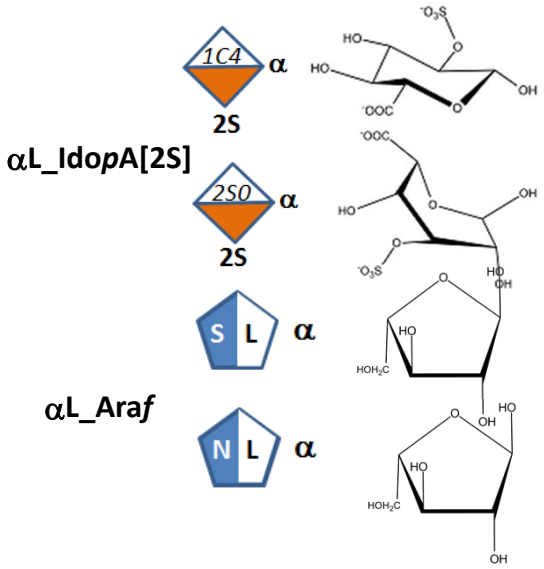
The D-configuration for monosaccharide and the L configuration for Fucose and Idose are implicit and does not appear in the symbol. Otherwise the L configuration, is indicated in the symbol, as in the case of Arabinose or L-Galactose.
For those occurring in the furanose form, a letter *N* or *S* is inserted in the symbol, indicating the northern *(N*) or Southern (*S*) conformation of the five membered ring.

Anomeric Configuration.

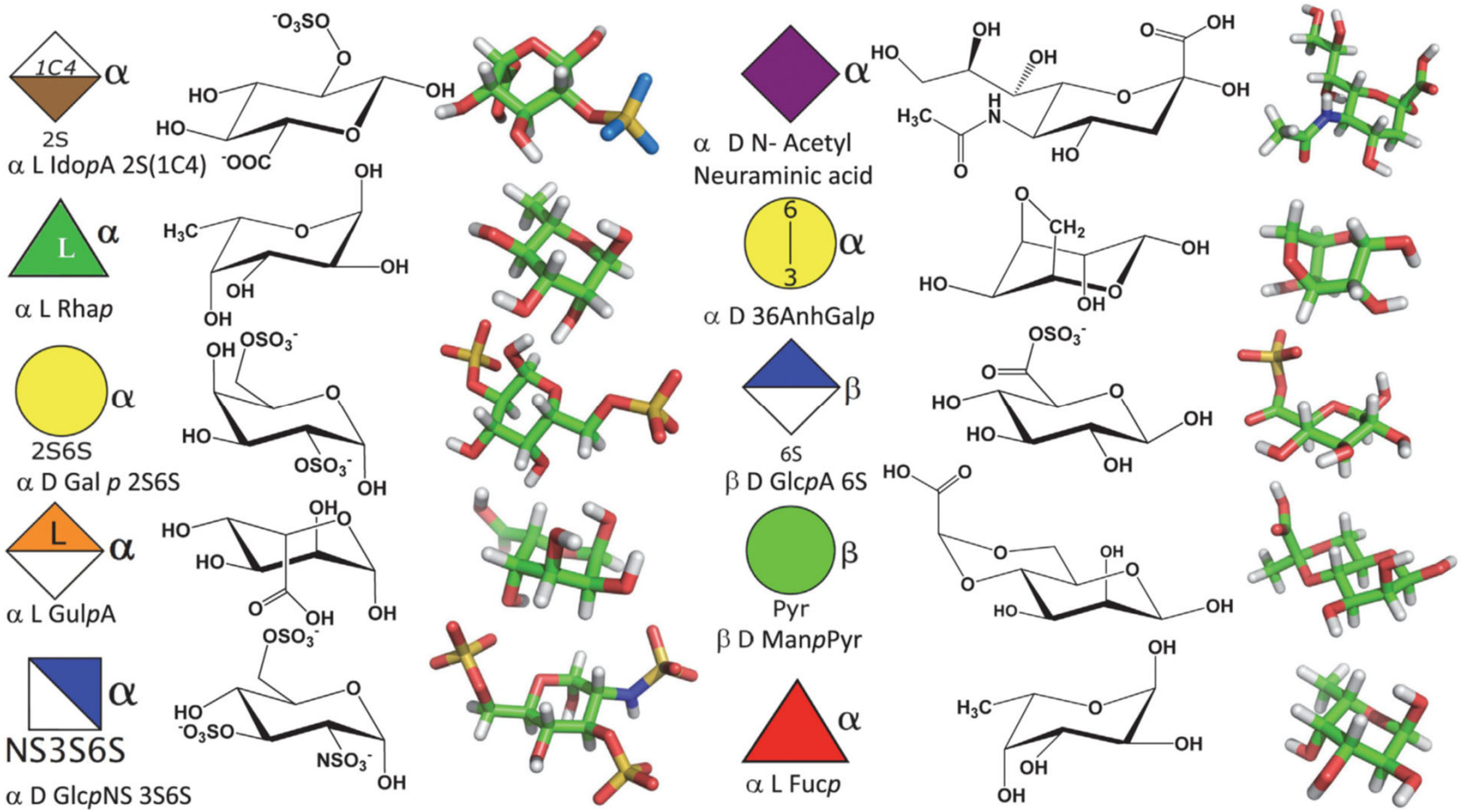
The nature of the glycosidic configuration (α or β) is explicitly set within the symbol.

Ring Conformation.

All pyranoses in the D-configuration are assumed to have $^4C_1$ chair conformation; those in the L configuration are assumed to have $^1C_4$ chair conformation. Otherwise, the symbol indicates the ring conformation as $^2S_0$ in the case of α-L-Idopyranose.
N or *S* indicates the conformation of the five membered rings on the conformational wheel.



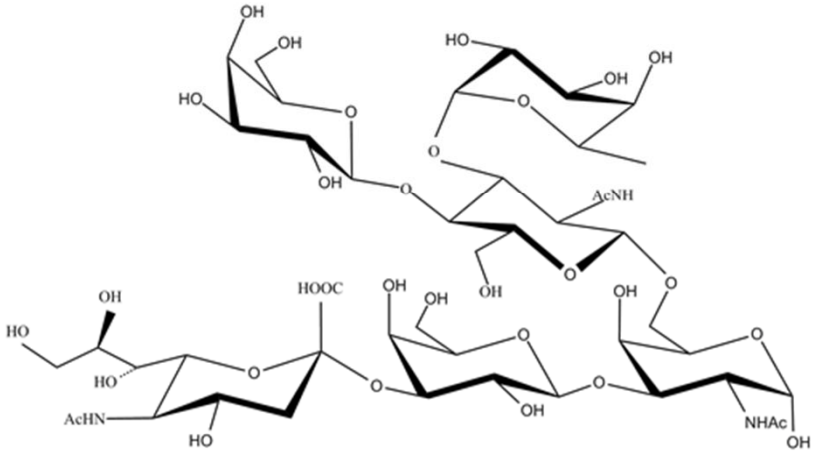
From Symbol Representation to 3D-Structures

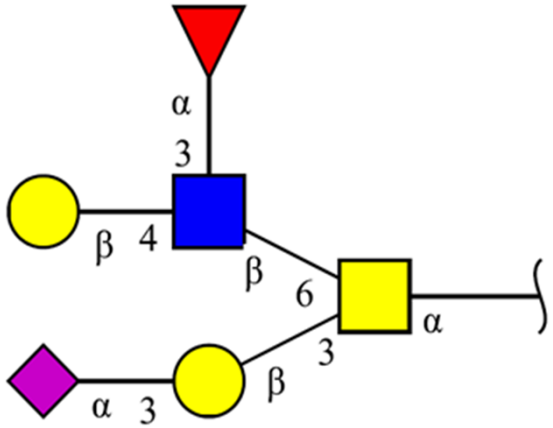


Encoding of Glycan Structures

Lewis X and Sialyl Acid on Core 2

Neu5Ac a2-3 Gal b1-3 (Gal b1-4 (Fuc a1-3) GlcNAc b1-6) GalNAc





RES
1b:a-dgal-HEX-1:5
2s:n-acetyl
3b:b-dgal-HEX-1:5
4b:a-dgro-dgal-NON-2:6|1:a|2:keto|3:d
5s:n-acetyl
6b:b-dglc-HEX-1:5
7s:n-acetyl
8b:a-lgal-HEX-1:5|6:d
9b:b-dgal-HEX-1:5
LIN
1:1d(2+1)2n
2:1o(3+3)3d
3:3o(3+2)4d
4:4d(5+1)5n
5:1o(6+1)6d
6:6d(2+1)7n
7:6o(3+1)8d
8:6o(4+1)9d

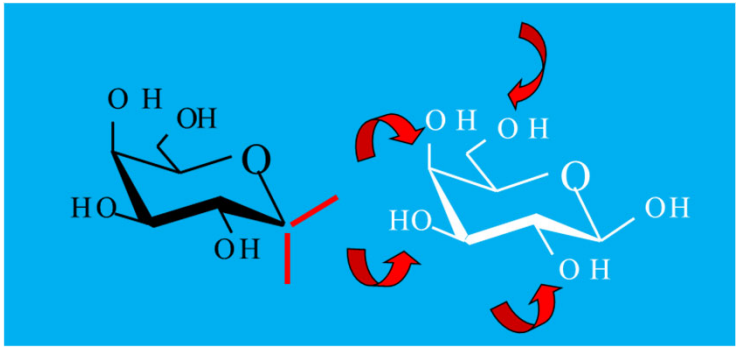
GlycoCT

Disaccharides & Higher Oligosaccharides

•Have a very high number of monomers (substitution…).

• Have many different ways of connecting monomers.

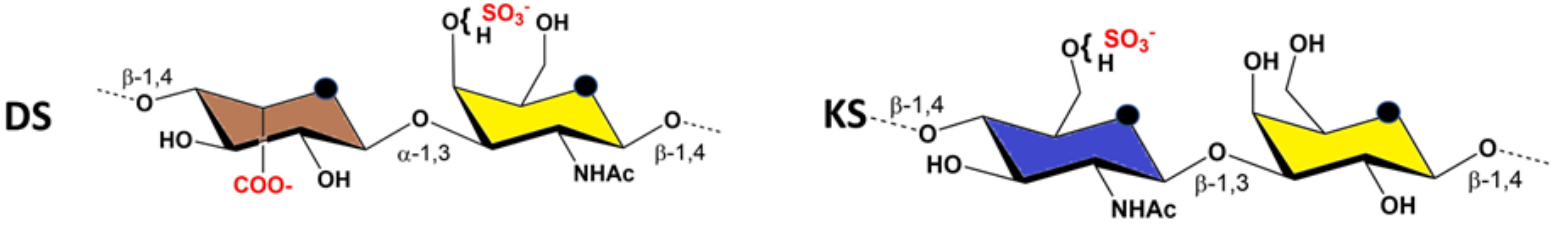
• Have branching points.



•Are difficult to synthesize and to crystallize.

• Are not the direct producs of a gene (≠ proteins).

• Cannot be amplified by PCR (≠ Nucleic acids).

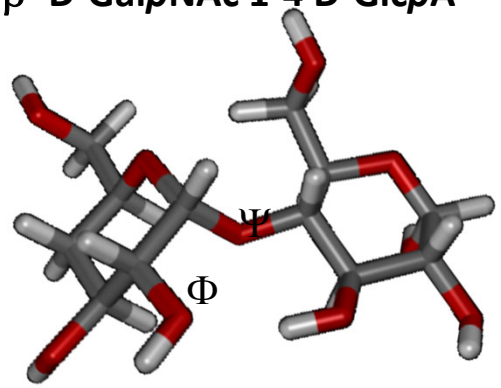


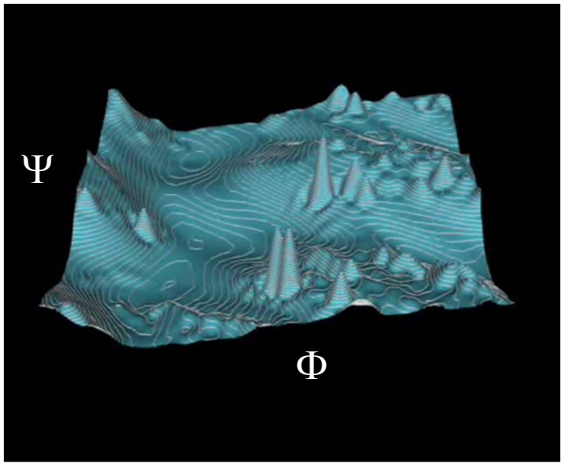
All chemical compounds are described with IUPAC, Simplified Molecular Input Line Entry Specification syntax (SMILES), and InChi encodings that are readable by most chemo-informatics tools.

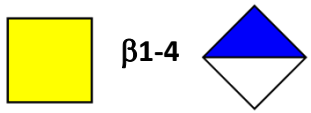
Glycans are encoded in GlycoCT, WURCS (Web3 Unique Representation of Carbohydrate Structures) LINUCS (LInear Notation for Unique Description of Carbohydrate Sequences).

Disaccharide: Structural Descriptors

β–D-Gal*p*NAc 1-4 D-Glc*p*A

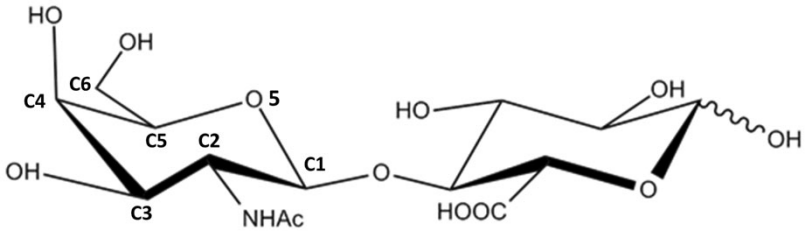


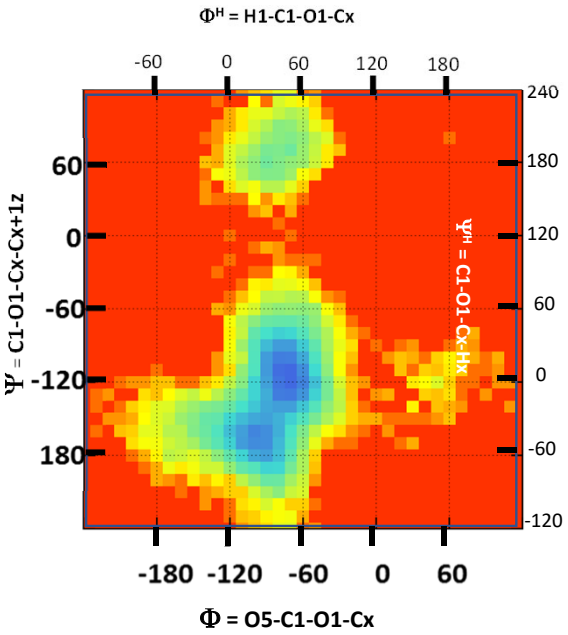


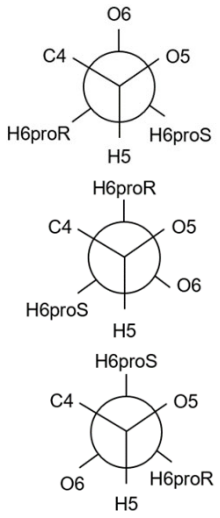


β1-4

Gal*p*NAc β1-4 Glc*p*A

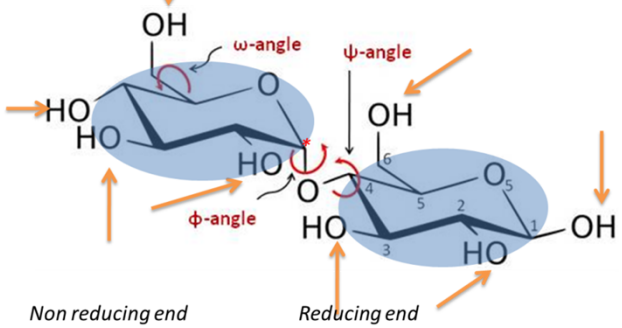


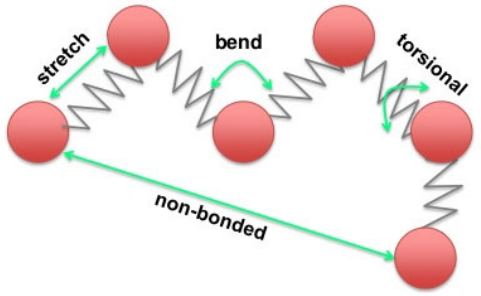


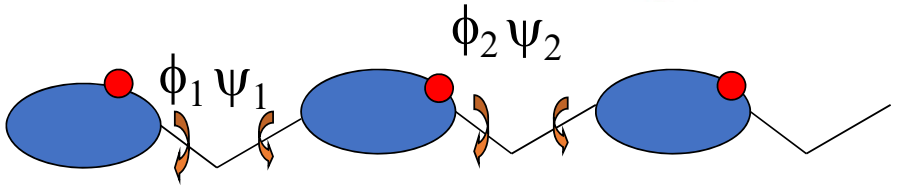


(Φ, Ψ) = -80; -120
(Φ, Ψ) = -100; -160
(Φ, Ψ) = -90; 70
(Φ, Ψ) = 60; -120

Conformational Space of Oligosaccharides







Combinatorial building

Assumption:

Because of the bulky and (almost) rigid nature of the monosaccharide unit, the conformation of each linkage is independent on the other

Methods :

Combine the lowest energy minima of each disaccharide map

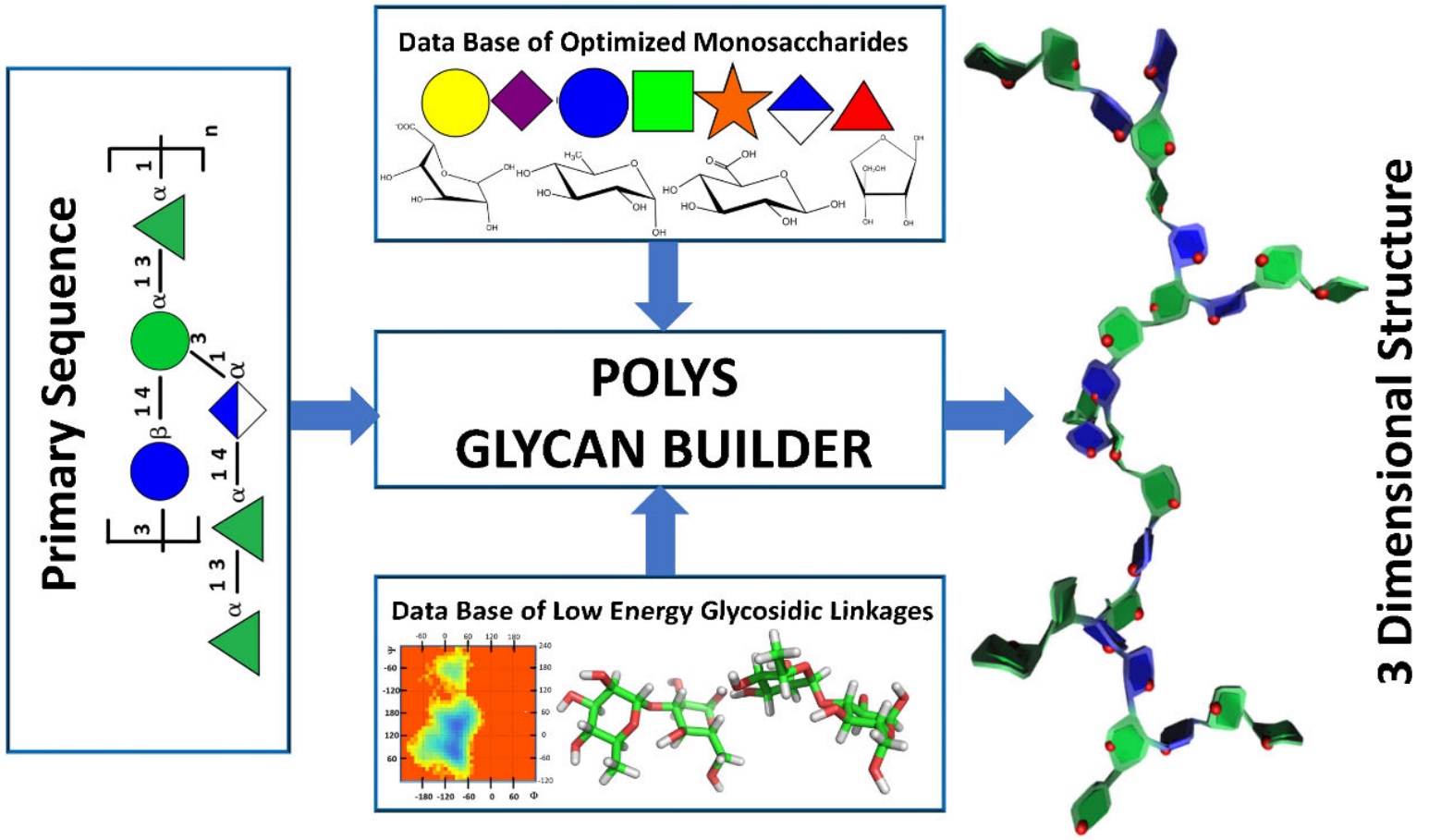
Not true for

- long range interactions
- branched structures

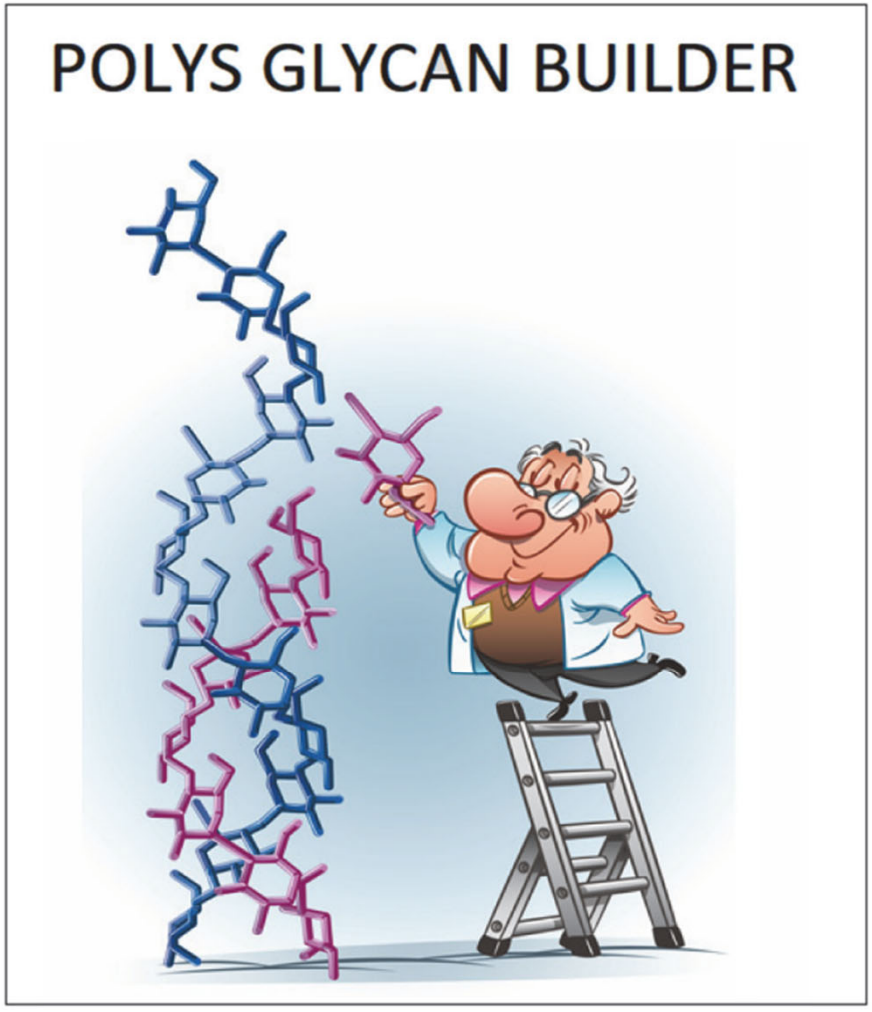
....

But very useful for building starting structures!

PolysGlycanBuilder



PolysGlycanBuilder



ALGEA BUILDER

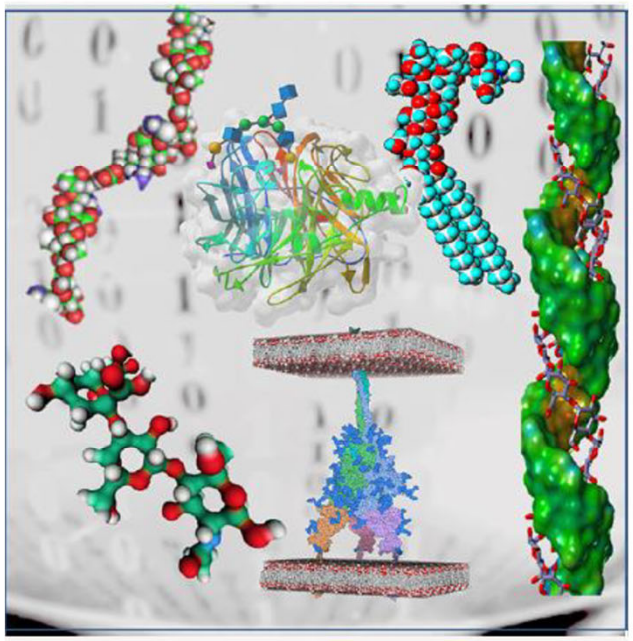
BACTERIA BUILDER

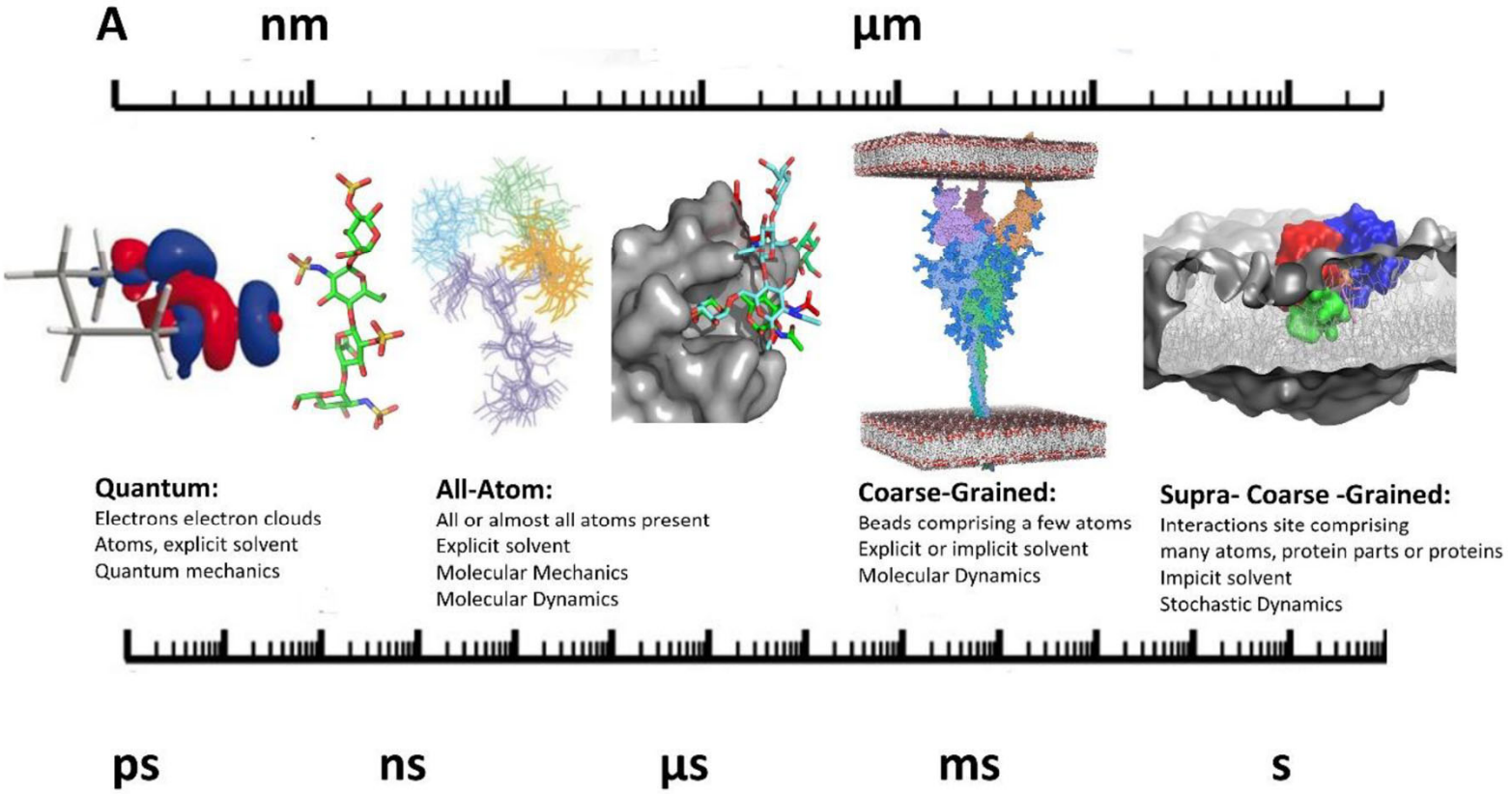
GAG BUILDER

N-O LINKED BUILDER

PLANT BUILDER

Multifaceted Computational Modeling in Glycoscience





Multifaceted Computational Modeling in Glycoscience, Chemical Reviews https://doi.org/10.1021/acs.chemrev.2c00060,

Molecular Mechanics / Dynamics

Initial positions given by the **PDB**
Initial velocities determined based on a Boltzmann distribution of velocities at the target temperature

$$\vec{F} = m\vec{a} = -\frac{dU}{dr}$$

New positions and velocities through integration

MD run → trajectory

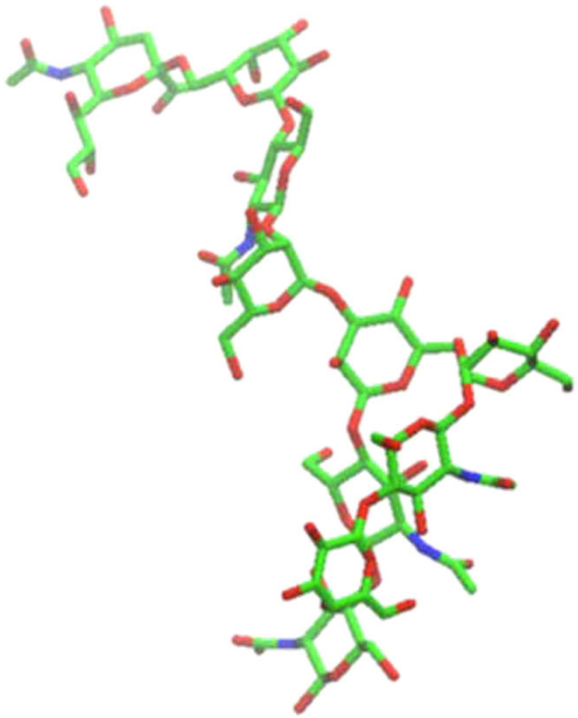
THE FORCE FIELD

$$v^{\text{Coulomb}}(r) = \frac{Q_1 Q_2}{4\pi\epsilon_0 r},$$

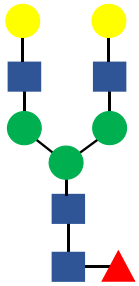
$$v^{\text{LJ}}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right].$$

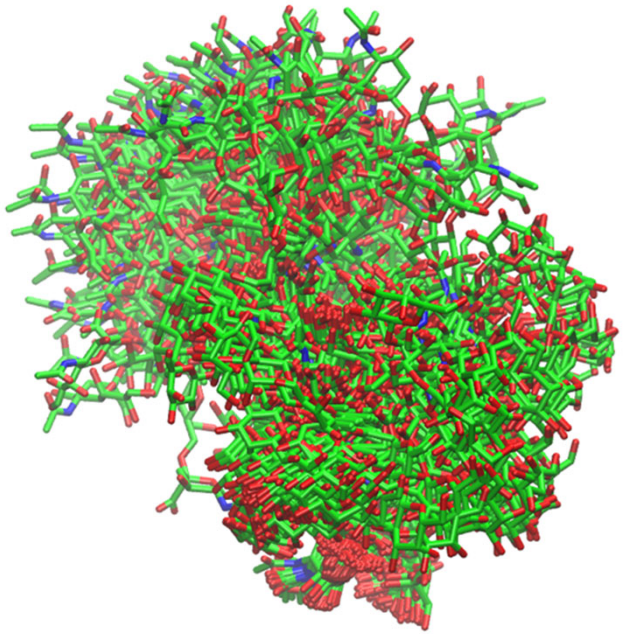
$$\begin{aligned} \mathcal{U}_{\text{intramolecular}} = & \frac{1}{2}\sum_{\text{bonds}} k_{ij}^{r}\left(r_{ij} - r_{\text{eq}}\right)^2 \\ & + \frac{1}{2}\sum_{\substack{\text{bend}\\ \text{angles}}} k_{ijk}^{\theta}\left(\theta_{ijk} - \theta_{\text{eq}}\right)^2 \\ & + \frac{1}{2}\sum_{\substack{\text{torsion}\\ \text{angles}}}\sum_{m} k_{ijkl}^{\phi,m}\left(1 + \cos(m\phi_{ijkl} - \gamma_m)\right) \end{aligned}$$

Glycans Can be Highly Flexible and Dynamic

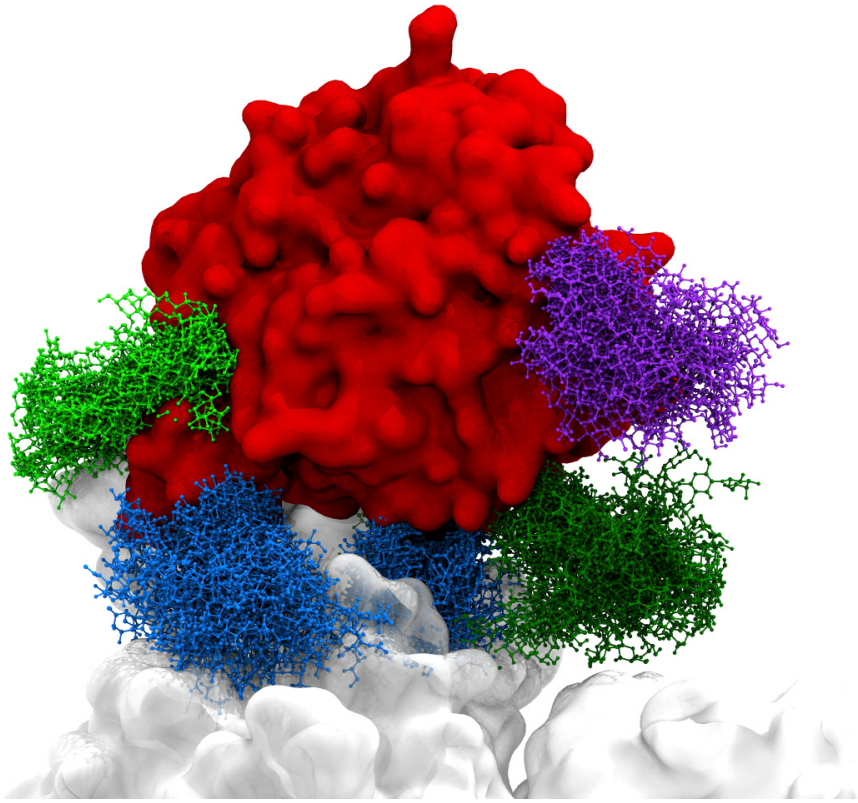


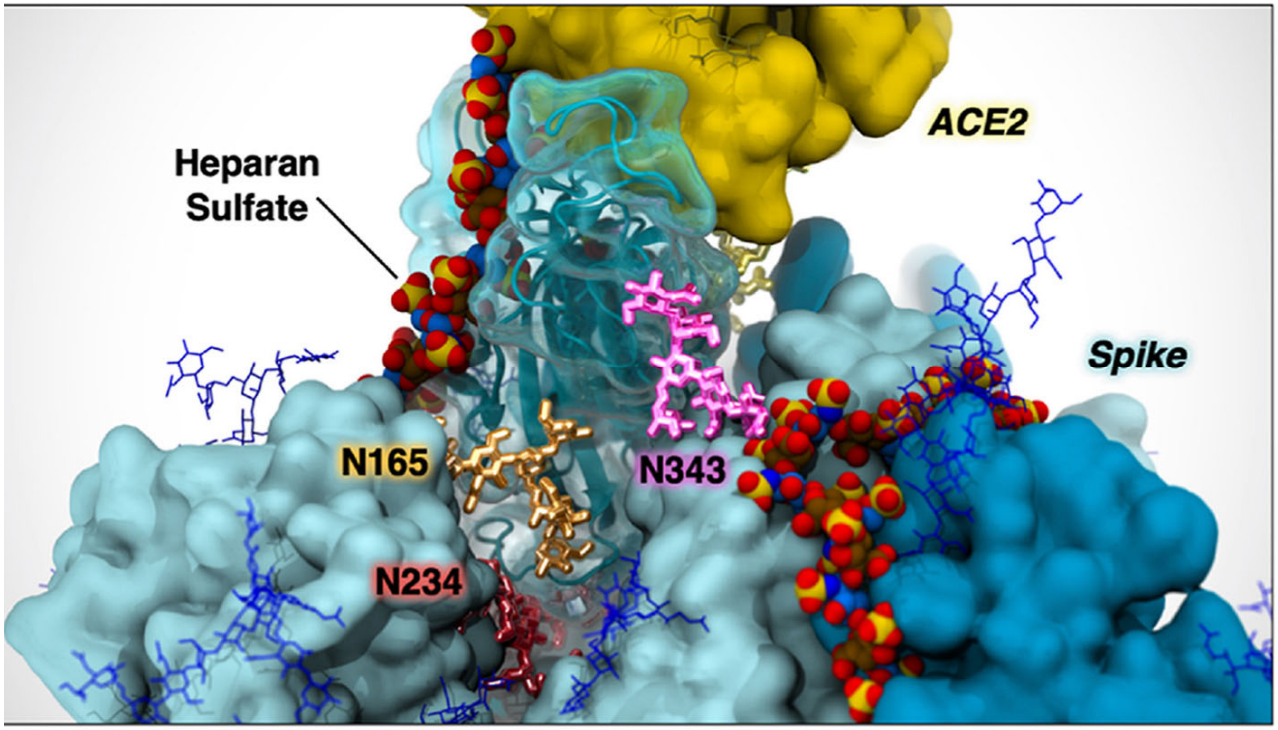
250 ns single trajectory





This ensemble of different structures is not necessarily ‘a mess’ not all possible conformations are allowed or equally populated and some of these conformations may actually be functionally important

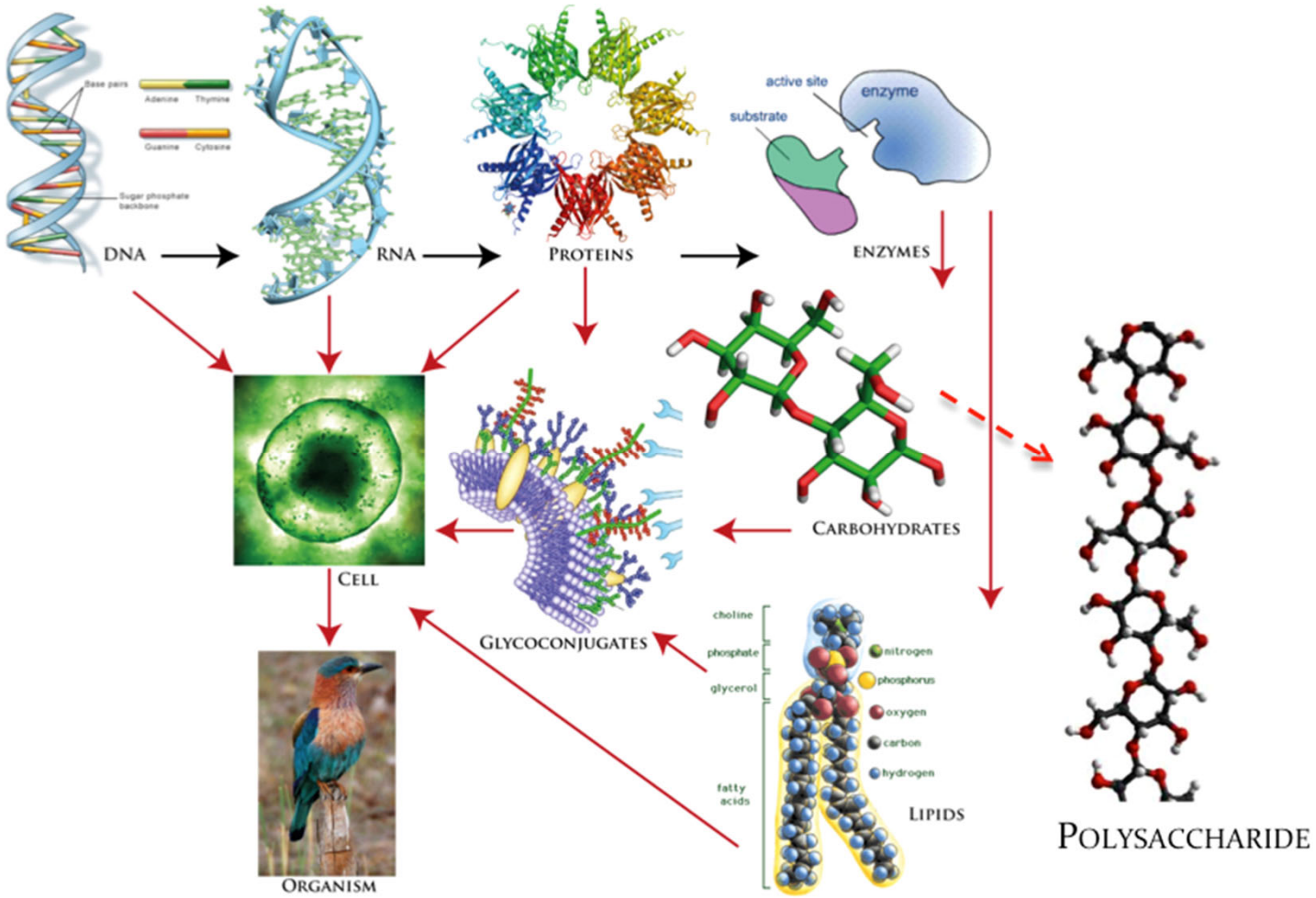




Interactions of glycosylated soluble human angiotensin-converting enzyme 2 (ACE2) and glycosylated SARS-CoV-2 S trimer. ACE2 is colored red, with ACE2 glycans interacting with the spike protein's upper part. (Cell Host Microbe 2020)

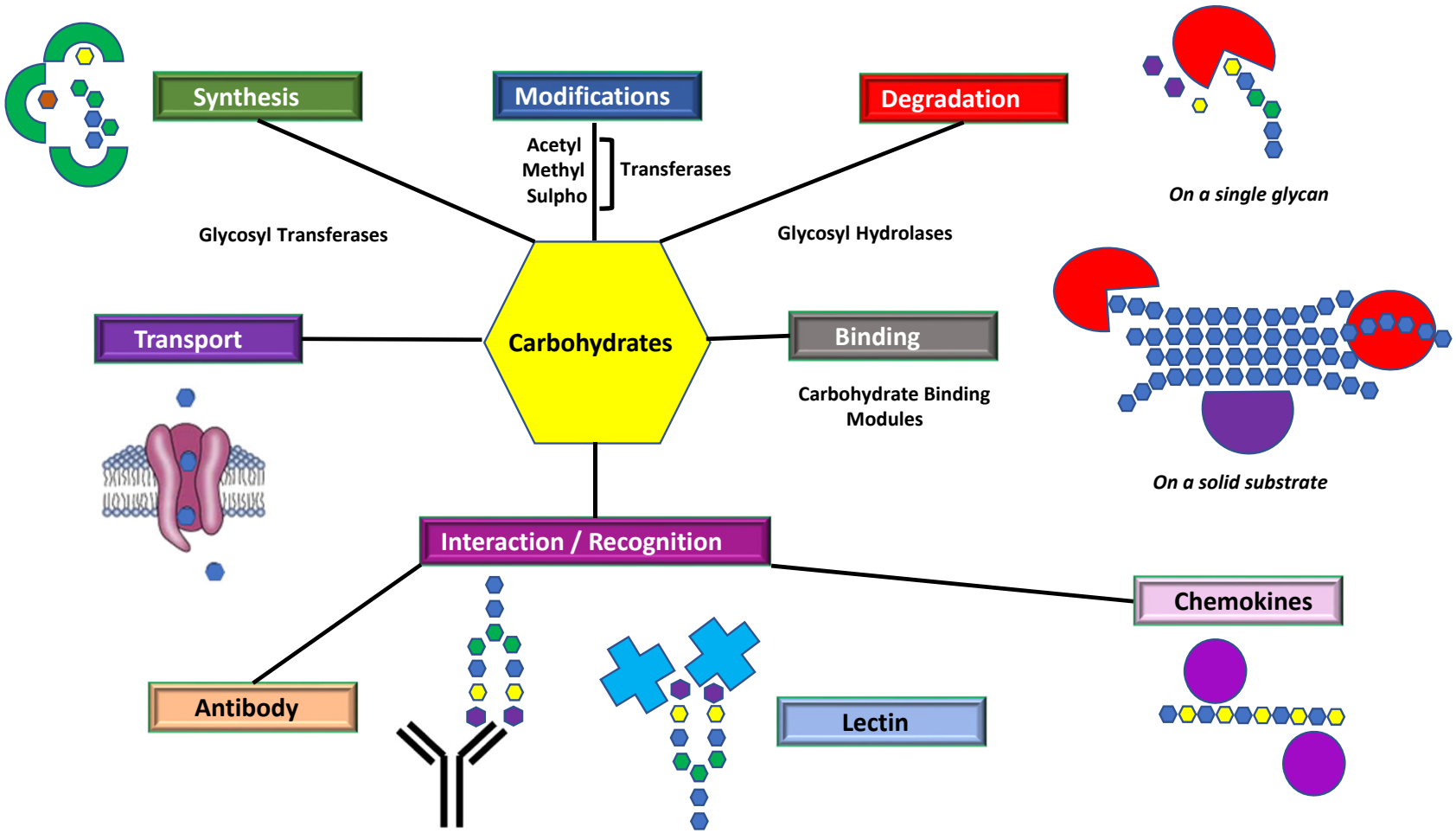
F.L. Kearns et al., Current Opinion in Structural Biology Volume 76, October 2022, 102439

Carbohydrates in the Scheme of the Central Dogma of Life



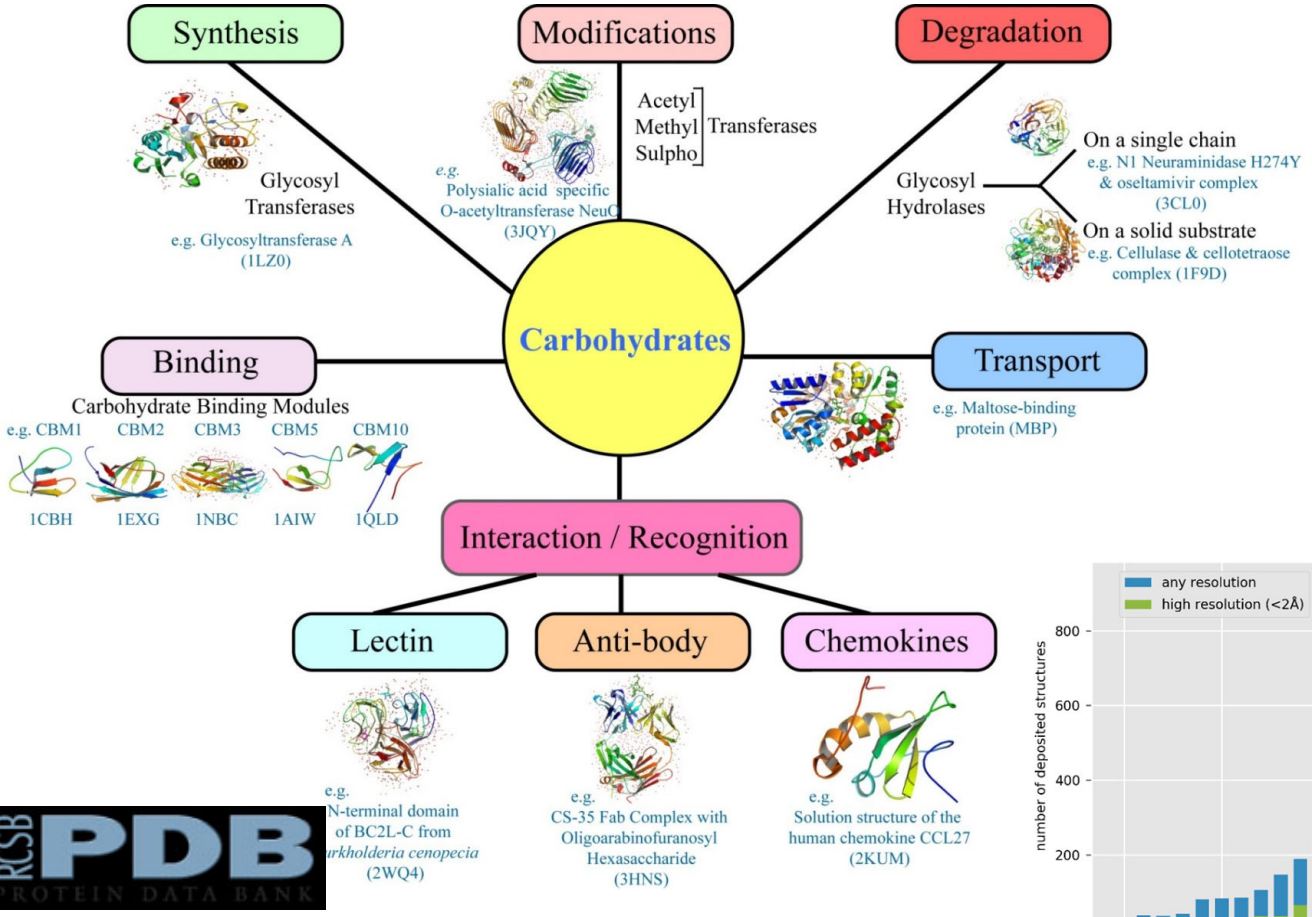
Inspired by: A. Varki. **Historical Background and Overview**. In: *Essentials of Glycobiology.* Edited by Varki A, Cummings RD, Esko JD, Freeze HH, Hart GW, Etzler ME, 2nd Edition Cold Spring Harbour (NY) (2008)

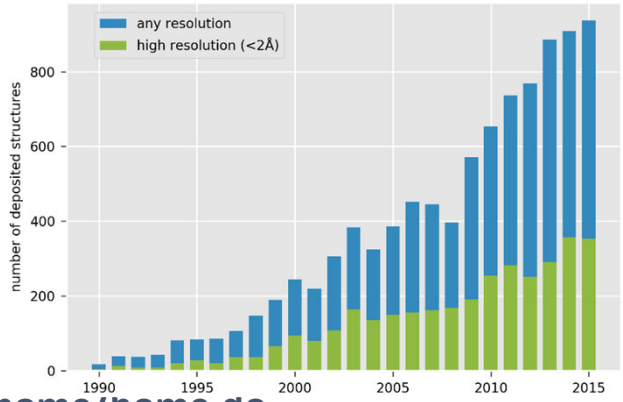
Glycan Active Proteins



Protein-Carbohydrate Crystal Structures

Protein-Carbohydrate Interactions

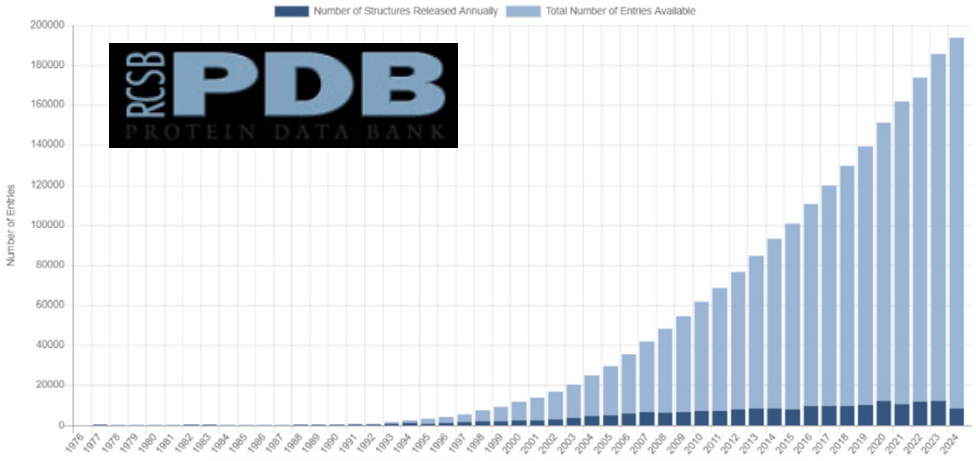




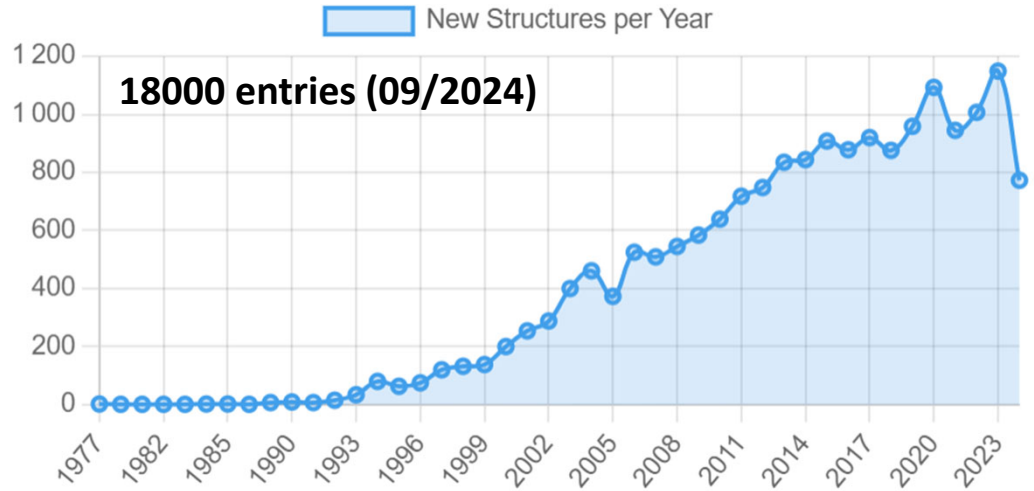
RCSB PDB PROTEIN DATA BANK

Protein Data Bank : http://www.rcsb.org/pdb/home/home.do

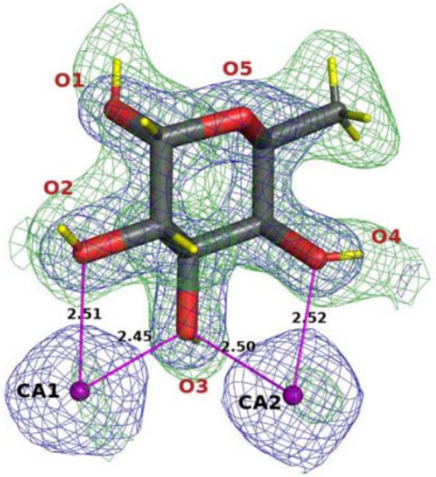
Protein-Carbohydrate Crystal Structures



Protein Data Bank : http://www.rcsb.org/pdb/



Carbohydrate-containing structures added yearly to the PDB

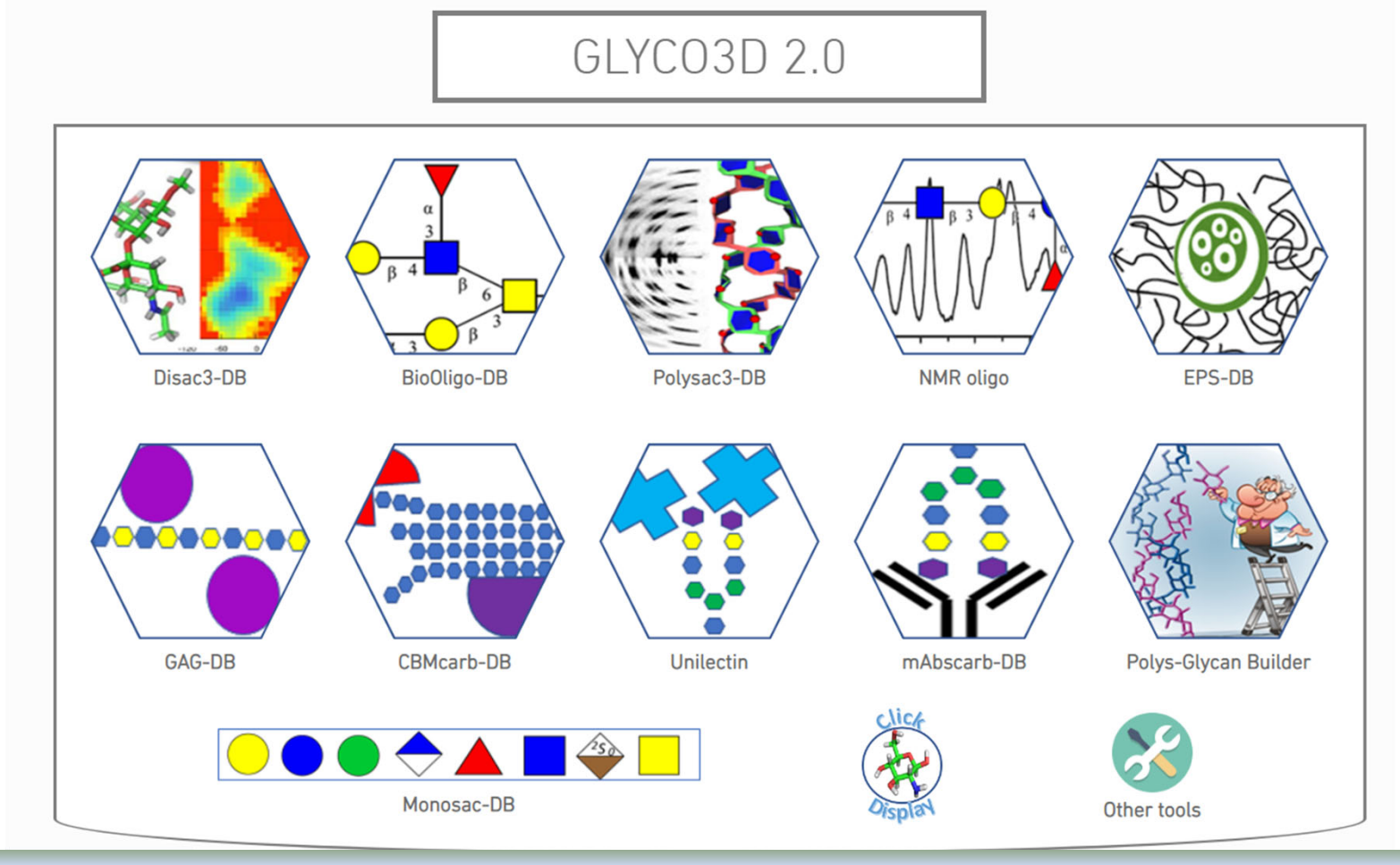


Neutron Protein crystallography
< 200 entries
4 Protein-Carbohydrate Complexes

How many entries are in the PDB /AlphaFold ?

- AlphaFold is an AI system developed by Google DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.
- Google DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) have partnered to create **AlphaFold DB** to make these predictions freely available to the scientific community.
- The latest database release contains over 200 million entries, providing broad coverage of UniProt (the standard repository of protein sequences and annotations).
- Individual downloads are provided for the human proteome and for the proteomes of 47 other key organisms important in research and global health.
- Download for the manually curated subset of UniProt (Swiss-Prot).

Structural DataBases



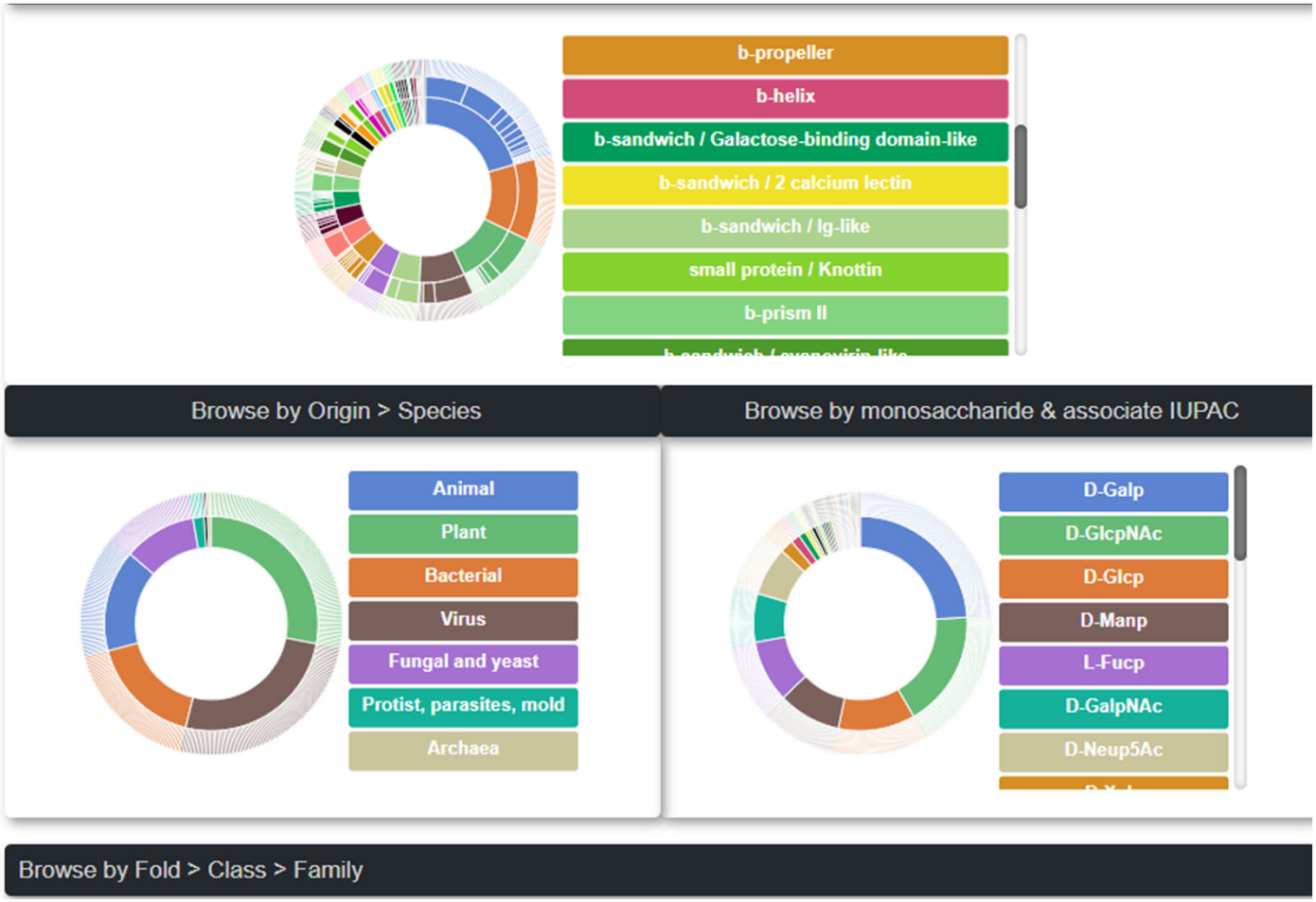
Glyco3D:https://glyco3d.cermav.cnrs.fr/home.php

Unilectin3D Curated Database

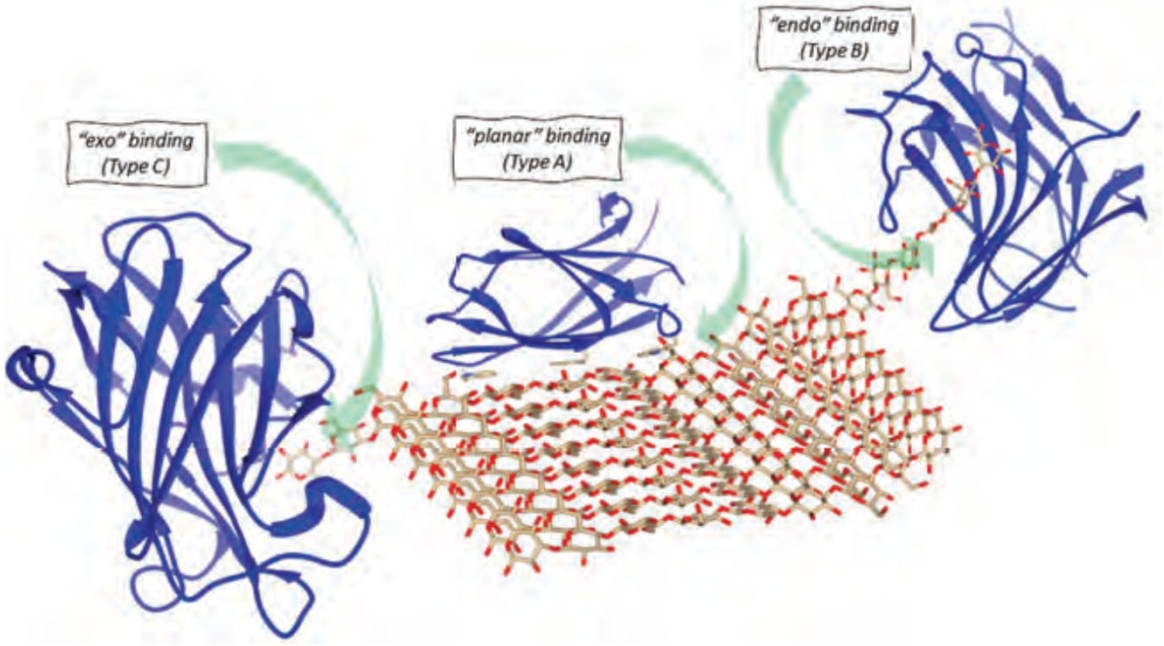
Unilectin3D provides curated information on 3D structures of lectins grouped into families based on the carbohydrate binding domains

How many lectins and structures?
2639 3D XRay structures
(1622 with interacting glycans)
686 distinct lectin sequences
247 distinct glycans
1162 articles

Invited to attend Anne's Imberty Presentation on Friday



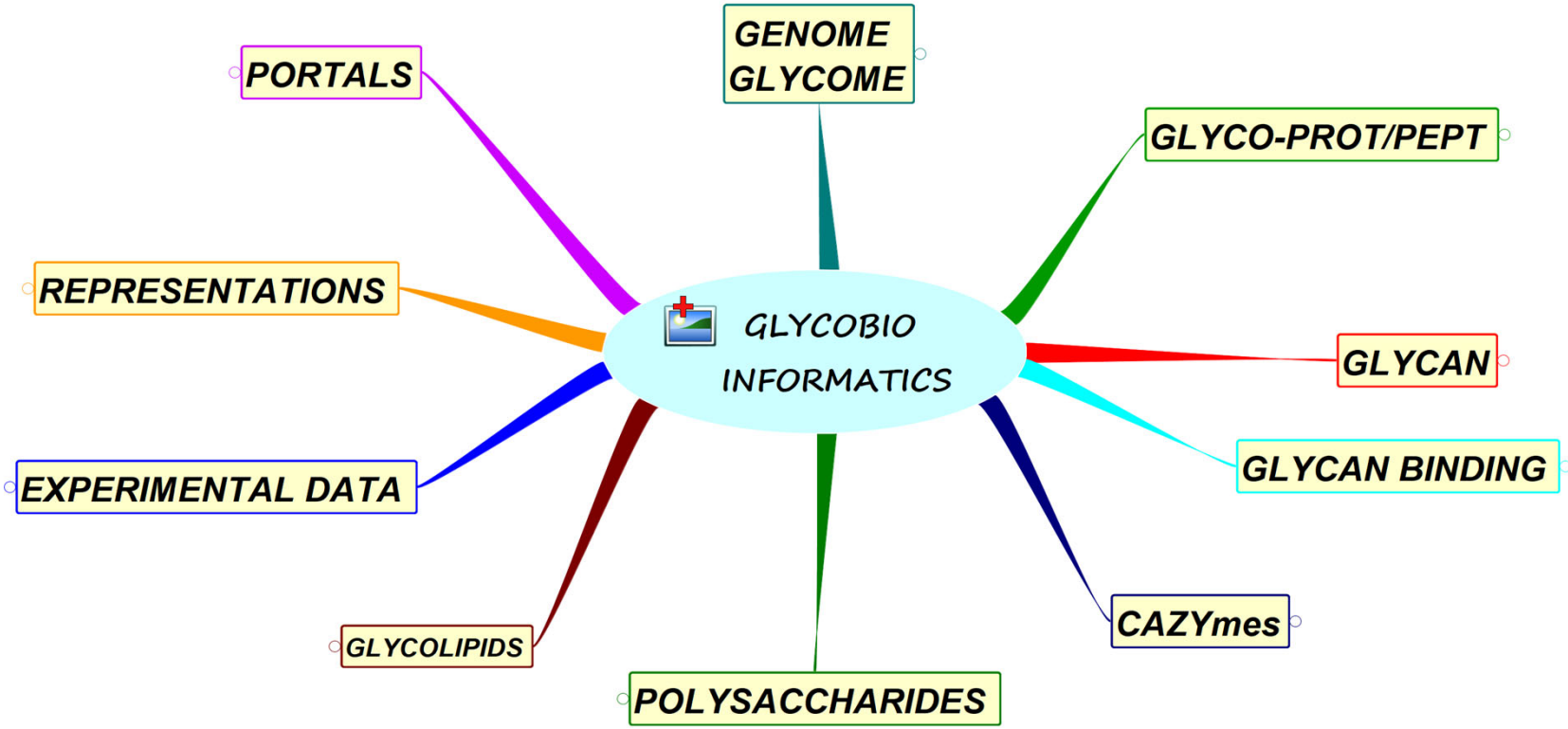
CBMcarb-DB: Interface of the 3D Landscape of Carbohydrate-Binding Modules



CBMs: a class of carbohydrate-binding proteins, defined as non-catalytic protein domains with amino acid sequences ranging from 30 to 200 amino acids

Ribeiro et al, 2024 https://cbmcarb.webhost.fct.unl.pt.

Tools and DataBases



Multifaceted Computational Modeling in Glycoscience; Serge Perez and Olga Makshakova
Chemical Reviews (2022), Cite This: https://doi.org/10.1021/acs.chemrev.2c00060

Biomolecular databases

RCSB PDB PROTEIN DATA BANK — 224,004 Structures from the PDB; 1,068,577 Computed Structure Models (CSM)

UniProt — BLAST Align Peptide search ID mapping SPARQL

Binding affiinities databses

BindingDB

Binding MOAD — Mother of All Databases

PDB bind

Protein-carbohydrate databases

CAZy — Carbohydrate-Active enZymes

UniLectin — Unified exploration platform for manually curated and predicted lectins

Protein domain classifications

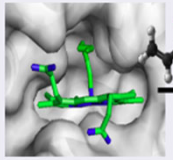
SCOPe

CATH

ECOD:

Evolutionary Classification of proteins Domains

Protein-ligand database



BioLiP2 for Ligand-protein binding database

Protein family classications

Pfam

Enzyme classifications

BRENDA

Ressources for Glycoscience

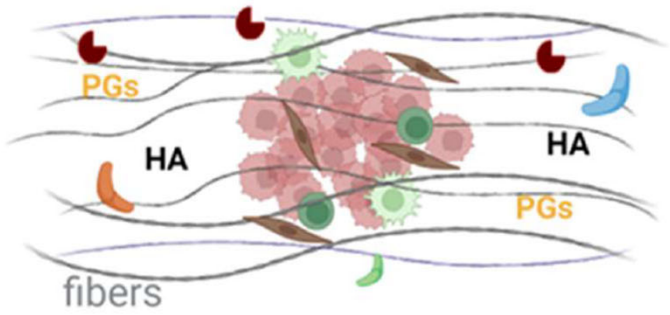
GlyGen

GlyConnect GLYCOEXPASY

GlyCosmos Portal

ProteoGlycan-mediated mechanosensing pathways in health and disease

ECM stiffening in cancer microenvironment

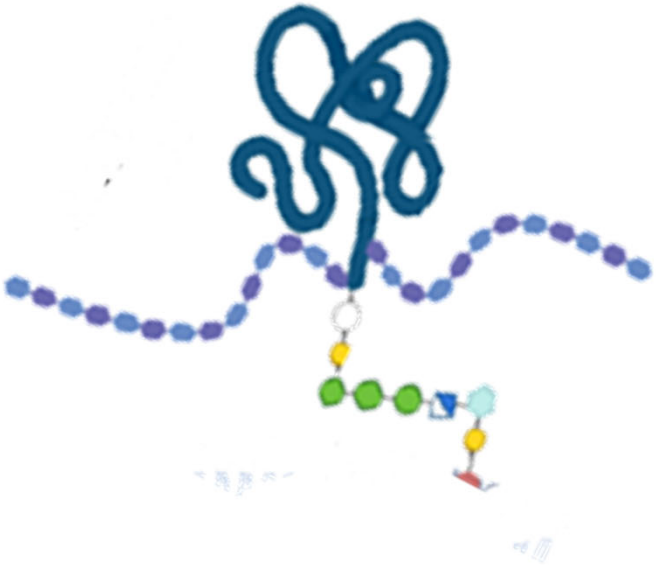


GF, cytokines, mitogens
Fiber alignment
Fiber crosslinking (elastin-collagen)
Enzymes (MMPs)
Desmoplasia

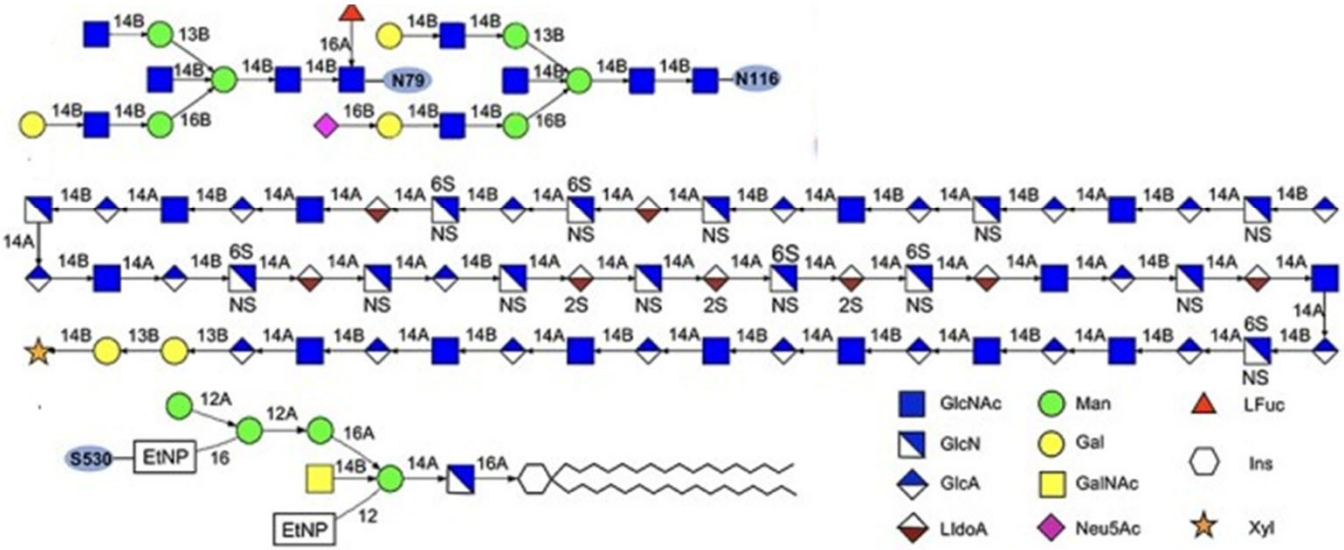
Modulations
Receptors: integrin R, hyaluronan R, RAGE, DRR 1+2, RTKs
Transmembrane PGs: syndecans
Post-translational modifications of ECM proteins
Signaling molecules: FAK/Scr, LOX, ROCK, PI3K
Cytoskeletal changes: F-actin, laminin
YAP nuclear translocation

Motility, survival, metastasis, chemotherapy resistance

Human Glypican-1



```
>sp|P35052|GPC1_HUMAN Glypican-1 OS=Homo sapiens OX=9606 GN=GPC1 PE=1 SV=2
MELRARGWWLLCAAAALVACARGDPASKSRSCGEVRQIYGAKGFSLSDVPQAEISGEHLR
ICPQGYTCCTSEMEENLANRSHAELETALRDSSRVLQAMLATQLRSFDDHFQHLLNDSER
TLQATFPGAFGELYTQNARAFRDLYSELRLYYRGANLHLEETLAEFWARLLERLFKQLHP
QLLLPDDYLDCLGKQAEALRPFGEAPRELRLRATRAFVAARSFVQGLGVASDVVRKVAQV
PLGPECSRAVMKLVYCAHCLGVPGARPCPDYCRNVLKGCLANQADLDAEWRNLLDSMVLI
TDKFWGTSGVESVIGSVHTWLAEAINALQDNRDTLTAKVIQGCGNPKVNPQGPGPEEKRR
RGKLAPRERPPSGTLEKLVSEAKAQLRDVQDFWISLPGTLCSEKMALSTASDDRCWNGMA
RGRYLPEVMGDGLANQINNPEVEVDITKPDMTIRQQIMQLKIMTNRLRSAYNGNDVDFQD
ASDDGSGSGSGDGCLDDLCSRKVSRKSSSSRTPLTHALPGLSEQEGQKTSAASCPQPPTF
LLPLLLFLALTVARPRWR
```



Be FAIR to Glycans…

Update on Standards: Glycan data management and exchange require consolidation and compliance to standards,: Minimum Information Data Required for Glycomics (***MIRAGE***)

FAIR Principles; ***Findability, Accessibility, Interoperability and Reusability.***
Many data are not fully characterized, the lack of information on the metadata (explaining and characterizing the measured or computed data), the ontologies relationships in metadata), and the workflow of different research groups are difficult to adjust. ***Most research data are neither, findable nor interoperable.***

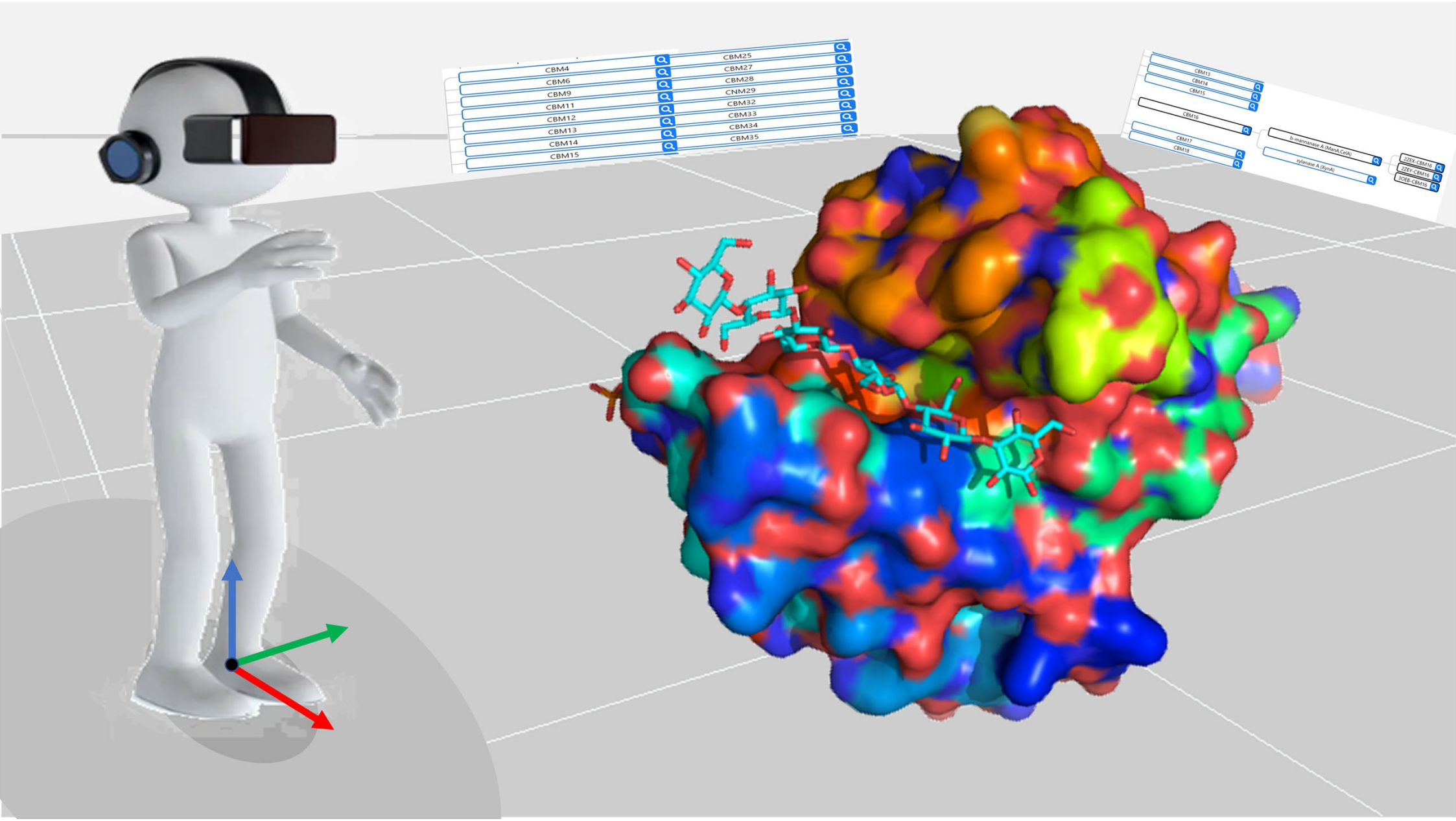
TRUST Principles: ***Transparency, Responsibility, User focus, Sustainability, Technology***

Cross-Referencing: Linking experimental, theoretical, and biological data using **common schemes** and **ontology** will generate a new level of Glycoscience

Data Modeling: Implementing multiscale data (spatial & temporal) faces heterogeneities: simulation steups, force fields, meaning and representation of the produced data
Need for selection and compressions strategies compatible with the type and amount of data

Big Data and AI Approach : *Standardized, structured & well annotated data required to Deep Learning methods*

!!! Support DataBase funding !!!


CBM4
CBM6
CBM9
CBM11
CBM12
CBM13
CBM14
CBM15
CBM25
CBM27
CBM28
CNM29
CBM32
CBM33
CBM34
CBM35
CBM13
CBM14
CBM15
CBM16
CBM17
CBM18
b-mannanase A (ManA, CelA)
xylanase A (XynA)
2JEX-CBM16
2JEV-CBM16
3OEB-CBM16