# The Glycan Structure Dictionary - a dictionary describing commonly used glycan structure terms.

Jeet Vora[1*], Rahi Navelkar[1], K.Vijay-Shanker[2], Nathan Edwards[3], Xiying Ding[1], Tianyi Wang[1], Peng Su[2], Karen Ross[3], Frederique Lisacek[4], Catherine Hayes[4], Karina Martinez[1], Robel Kahsay[1], Rene Ranzinger[5], Michael Tiemeyer[5], Raja Mazumder[1]

[1]Department of Biochemistry & Molecular Medicine, The George Washington School of Medicine and Health Sciences, Washington, DC, 20052, USA. [2]Department of Computer and Information Science, University of Delaware, Newark, Delaware, 19716 USA. [3]Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, 20007, Washington, DC, USA. [4]University of Geneva and Swiss Institute of Bioinformatics, CUI - 7, route de Drize, Geneva, 1211, Switzerland. [5]Complex Carbohydrate Research Center, The University of Georgia, Athens, GA, 30602, USA.

*To whom correspondence should be addressed.

## Abstract

Glycosylation is one of the most frequent and important post-translational modifications wherein glycans affect several physicochemical properties of a protein, including protein folding, stabilization, and protein interactions, mediate host-pathogen interactions and are widely used to develop therapeutics and small-molecule mimetics. Recent technological advancements in the field of glycobiology have resulted in a large influx of data; however, authors still use different terms to report the glycan structural features in the publications. To address this shortcoming, the Glycan Structure Dictionary has been developed as a reference dictionary to provide a standardized list of widely used glycan terms that will help in the curation and mapping of glycan structures described in the publications.

Glycans mediate important biological functions, serve as biomarkers for impactful diseases, regulate host-pathogen interactions, and contribute to ongoing efforts to develop novel biotherapeutics, bioenergy sources, and biomaterials. Glycan-related research, databases, and bioinformatics tools have advanced in the last twenty years, providing ever-increasing volumes of glycoproteomics and glycomics data. However, significant differences exist in the terms that various authors use to describe and report glycan structural features in the literature. The lack of a reference dictionary, that can be leveraged by researchers to report glycan structures in a standardized manner, complicates efforts to integrate glycan information from the literature into bioinformatics databases. To bridge this gap, we have developed a Glycan Structure Dictionary that encompasses a comprehensive reference list of widely used glycan terms and their definitions.

The Glycan Structure Dictionary is designed to generate a reference list for meta-analysis and text-mining to help researchers and informaticians extract, transfer and search glycan information efficiently. To achieve this goal, a term list was derived from automated text mining. Just like a term in a language dictionary, each term collected in the dictionary is annotated with additional information such as a definition, literature evidence (PMID), a sentence from a paper where the term is found, cross-links from other open-access databases, synonyms, functions, disease associations, Wikipedia pages, and relevant chapter number(s) in Essentials of Glycobiology. The primary identifier for the terms is a GlyTouCan accession (Fujita et al. 2020), where available. The Glycan Structure Dictionary also captures cross-references from GlycoMotif (https://glycomotif.glyomics.org/), GlycoEpitope (https://www.glycoepitope.jp/), PubChem (Kim et al. 2021), ChEBI (Hastings et al. 2015), and other databases. Currently, the dictionary has over 180 glycan structure terms and is updated regularly. The current terms can be accessed via GlyGen Wiki (https://wiki.glygen.org/index.php/Glycan_structure_dictionary) or as a downloadable dataset via https://data.glygen.org/GLY_000557. The links to the Glycan Structure Dictionary terms are also available from the List of Motifs details and Glycan details page (https://www.glygen.org/list-of-motifs/) on the GlyGen portal (York et al. 2019) (https://www.glygen.org/) for the terms containing cross-references to GlyTouCan and GlycoMotif.

The Glycan Structure Dictionary will become increasingly useful as it is populated by terms frequently used by the community. Therefore, collaborations and contributions from a broad range of glycobiologists and other scientists are welcomed. Users, researchers, bioinformaticists, and other interested colleagues can submit a single term using the online form at https://data.glygen.org/gsd/, or they can submit multiple terms through a file upload mechanism (with a sample template) provided on the Glycan Structure Dictionary Wiki Page (GlyGen Consortium 2021). The term will be accepted if it 1) describes a motif, type, subtype, branching, or terminal structure of a glycan, 2) has at least two associated publications (PMIDs/DOIs) where the term is represented exactly or in its synonym form, and 3) is not already included in the dictionary. Users can also submit updates such as annotations (PMIDs, function, disease association, etc.) to existing terms.

The broad adoption of the Symbol Nomenclature for Glycans (SNFG) (Neelamegham et al. 2019) provided a blueprint for the visual representation of monosaccharides and glycan structures in a standardized way. Adaptation of this nomenclature by researchers, authors, bioinformatics databases, and peer-reviewed journals has led to faster, more effective curation. Similarly, we believe that the glycan dictionary can help map glycan structures described in publications to databases such as GlyGen, GlyConnect (Alocci et al. 2018), GlyCosmos (Yamada et al. 2020), etc., thus positively impacting the curation process and enhancing access to knowledge in the glycoscience domain.

## Funding

This work was supported by the National Institutes of Health [1U01GM125267 - 01]

## References

Alocci D, Mariethoz J, Gastaldello A, Gasteiger E, Karlsson NG, Kolarich D, Packer NH, Lisacek F. 2018. GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. Journal of Proteome Research. 18(2):664–677. doi:10.1021/acs.jproteome.8b00766.

Fujita A, Aoki NP, Shinmachi D, Matsubara M, Tsuchiya S, Shiota M, Ono T, Yamada I, Aoki-Kinoshita K. 2020. The international glycan repository GlyTouCan version 3.0. Nucleic Acids Research. 49(D1):D1529–D1533. doi:10.1093/nar/gkaa947.

GlyGen Consortium. 2021. Glycan Structure Dictionary - Submit New Terms. GlyGen Wiki. https://wiki.glygen.org/index.php/Glycan_structure_dictionary#Submit_new_terms.

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. 2015. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research. 44(D1):D1214–D1219. doi:10.1093/nar/gkv1031.

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, et al. 2021. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Research. 49(D1):D1388–D1395. doi:10.1093/nar/gkaa971. [accessed 2021 Feb 5]. https://academic.oup.com/nar/article/49/D1/D1388/5957164.

Neelamegham S, Aoki-Kinoshita K, Bolton E, Frank M, Lisacek F, Lütteke T, O'Boyle N, Packer N, Stanley P, Toukach P, et al. 2019. Updates to the Symbol Nomenclature for Glycans guidelines. Glycobiology. 29(9):620–624. doi:10.1093/glycob/cwz045.

Yamada I, Shiota M, Shinmachi D, Ono T, Tsuchiya S, Hosoda M, Fujita A, Aoki NP, Watanabe Y, Fujita N, et al. 2020. The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. Nature Methods. 17(7):649–650. doi:10.1038/s41592-020-0879-8. https://pubmed.ncbi.nlm.nih.gov/32572234/.

York WS, Mazumder R, Ranzinger R, Edwards N, Kahsay R, Aoki-Kinoshita KF, Campbell MP, Cummings RD, Feizi T, Martin M, et al. 2019. GlyGen: Computational and Informatics Resources for Glycoscience. Glycobiology. 30(2):72–73. doi:10.1093/glycob/cwz080.