Structural Organization of the Glycosylation Machinery: A Computational Investigation of CAZyme Assemblies in the Golgi

Thesis director: Marc F. Lensink

Summary

Protein glycosylation is a central post-translational modification that profoundly affects protein stability, trafficking, and molecular recognition. Glycosylation represents one of the most complex and dynamically regulated pathways, one that is able to generate an extraordinary diversity of glycan structures that play essential roles in development, immunity, and disease. In contrast to templated macromolecules, glycan biosynthesis is not directly encoded in the genome but emerges from the coordinated action of multiple glycosyltransferases, glycosylhydrolase and transporters (CAZymes) localized along the secretory pathway. These enzymes act sequentially and competitively on shared substrates, yet the structural and organizational principles governing their cooperation remain largely unresolved.

This PhD project aims to elucidate the structural basis of functional organization within the glycosylation pathways by focusing on protein-protein interactions between consecutively acting Golgi-resident CAZymes. Building upon the conceptual and methodological framework established in the team (MassiveFold, CDP, GlycoPlay, add refs), the project will apply state-of-the-art structural bioinformatics and large-scale assembly modeling to identify, characterize, and rationalize putative multi-enzyme complexes, focusing on those CAZymes involved in O-glycan biosynthesis. Internal collaboration within the host institute (UGSF) will guide target selection and interpretation, but also provide experimental validation, ensuring a tight computational-experimental feedback loop.

By combining deep learning-based structure prediction, massive sampling of protein assemblies, and graph-based analysis of interaction interfaces, this project seeks to provide molecular-level evidence for the existence of glycosylation metabolons and to clarify how structural organization contributes to pathway specificity, robustness, and regulation.

State Of The Art

Recent advances in deep learning-based protein structure prediction, most notably AlphaFold and its derivatives, including MassiveFold developed in the host team [1], have fundamentally changed the landscape of structural biology. These methods have demonstrated an unprecedented ability to infer three-dimensional structures and assemblies from sequence information alone, capturing evolutionary constraints that are particularly informative for interacting protein systems [2–4]. Massive sampling strategies, such as those implemented in MassiveFold, have further extended these capabilities by enabling the exploration of large conformational spaces as well as alternative complex compositions and stoichiometries for protein assemblies [5]. These developments make it now feasible to address long-standing questions about the structural organization of glycosylation pathways at a molecular scale.

Experimental studies have demonstrated that Golgi glycosyltransferases are not randomly distributed but instead form homo- and heteromeric complexes whose assembly influences enzymatic activity, spatial organization within Golgi membranes, and pathway-level substrate processing [6–8]. Evidence for such interactions has been reported for several CAZymes involved in both N- and O-glycosylation, suggesting that transient and dynamic protein assemblies may be a general organizational principle of the Golgi glycosylation machinery [9]. However, due to the membrane-associated nature of these proteins and the weak, short-lived character of their interactions, high-resolution structural data for these complexes remain scarce. At present, genetics allows identifying the genes involved and glycomics allows identifying the glycans on cell surfaces, but experimental evidence as to the molecular mechanisms underlying glycosylation remain out of reach. Here lies the strength of present-day computational approaches as they allow us to sketch a reliable image of the interactions involved, including their dynamics.

O-glycosylation of proteins

Mucin-type O-glycosylation is initiated in the Golgi by a large family of polypeptide N-acetylgalactosaminyltransferases (GalNAc-Ts), which transfer GalNAc residues to serine and threonine residues on protein substrates. Subsequent elongation and modification steps involve multiple CAZyme families, including core-specific galactosyltransferases, GalNAc-Ts, sialyltransferases, fucosyltransferases, chaperones and transporters of *e.g.* of sugar-nucleotides and ions. Many of these are shared with or closely related to enzymes acting in other glycosylation pathways. The combinatorial nature of these reactions underlies the structural diversity of glycans including O-glycans but also raises fundamental questions regarding pathway control and fidelity [10]. Pathways that are built stepwise by multiple Golgi CAZymes are the most likely candidates for functional enzyme assemblies and substrate channeling, because multiple sequential catalytic steps occur in the same compartment and involve shared intermediates and donor/acceptor logistics. This makes the iterative, Golgi-localized biosynthetic cascade of mucin-type O-glycosylation particularly relevant and testable for the metabolon hypothesis.

Objectives

The central objective of this project is to identify and structurally characterize protein-protein interactions between CAZymes acting sequentially in the O-glycosylation pathway. Through systematic assembly modeling and interface analysis, the project aims to determine whether these enzymes form recurring, evolutionarily constrained interaction patterns consistent with the concept of a metabolon. A secondary objective is to relate the predicted interaction interfaces to enzyme specificity, regulation, and known pathological variants affecting O-glycosylation. Phylogenetic input will be used to prioritize conserved enzyme pairs and to interpret interaction models in an evolutionary context, while experimental validation will be carried out by partners within the host institute.

Methodology

The project will primarily rely on structural bioinformatics approaches developed in the host team. Target enzymes involved in O-glycan initiation and elongation will be selected based on biological relevance, availability of sequence data, and guidance from phylogenetic

analyses. For each enzyme, high-confidence monomeric models will be generated using AlphaFold-based approaches, with particular attention to luminal catalytic domains and flexible stem regions characteristic of type II Golgi membrane proteins.

Assembly modeling will then be performed using MassiveFold to generate large ensembles of putative homo- and heteromeric complexes between consecutively acting enzymes. This approach allows extensive sampling of relative orientations and interfaces on GPU-accelerated architectures, capturing alternative interaction modes that may correspond to transient or condition-dependent assemblies. Complementary docking methods will be employed where appropriate to assess methodological robustness and to incorporate experimental or bioinformatics restraints.

The resulting ensembles of structural models will be analyzed using graph-based consensus scoring and network centrality measures to identify recurrent interaction interfaces and key amino acid residues contributing to complex stability. These residues will be mapped onto sequence alignments and, where available, interpreted in light of co-evolutionary signals. The integration of structural and evolutionary information will enable the formulation of testable hypotheses regarding interaction specificity and functional relevance.

All computational results, including predicted structures, confidence metrics, and interface annotations, will be curated into a structured dataset. Experimental partners will use this information to design site-directed mutagenesis and biochemical assays aimed at validating the predicted interactions and assessing their impact on O-glycosylation outcomes. Although experimental work is not part of this PhD project, continuous interaction within the host institute will ensure iterative refinement of computational models based on empirical feedback.

Gantt Chart

The following Gantt chart is indicative for the development of the project:

Milestone	**Year 1**	**Year 2**	**Year 3**
1. Identification of relevant proteins in the pathway 2. Generation of interaction models			
3. Co-occurence analysis 4. Generation of interaction models			
5. Identification of interaction surface; mutagenesis design 6. Experimental assays			
7. Publication of results; thesis writing			

Expected Outcomes and Impact

This project is expected to deliver the first systematic, structure-based characterization of enzyme assemblies in the O-glycosylation pathway. By providing molecular-level models of CAZyme interactions, it will advance our understanding of how spatial organization within the

Golgi contributes to glycan diversity and pathway regulation. Beyond O-glycosylation, the methodological framework developed here will be broadly applicable to other metabolic pathways involving sequentially acting enzymes.

More generally, the project will contribute to establishing structural bioinformatics as a central tool for studying dynamic, weakly interacting protein systems that have traditionally remained inaccessible to high-resolution experimental techniques. In close synergy with our partners in the host institute and cross-disciplinary PIE project (Protein Interaction Evolution), this work will help substantiate the metabolon concept in glycosylation biology and open new perspectives on the evolution, regulation, and dysfunction of complex biosynthetic pathways.

Host Institute and Team

The research topics in the host institute UGSF evolve around the study of glycoconjugates, including their structural biology and modeling, biosynthesis and degradation, and function and pathologies due to deregulation. The institute adopts a multi-disciplinary and multi-scale approach to this combining experimental and computational approaches at all levels. The present project fits the scope of these research activities perfectly and aligns with two of its research axes: “Biomass polysaccharides” and “Glycosylations, regulation and pathologies”.

The host team, “Biomolecular interactions and dynamics”, led by Dr. Marc Lensink, plays a key role in the CAPRI project, where they are responsible for the organization of prediction rounds and the development of assessment criteria. Through their on-going collaboration with CASP, they have been closely observing the advent of AlphaFold and can be considered experts in its application. In France, they are involved in the installation and provisioning of AlphaFold, ColabFold, and other structural modeling tools based on deep learning approaches on national computing infrastructures. The team has a strong expertise in structural bioinformatics methods, including modeling-based approaches for the study of protein function and network approaches to study protein-protein interaction networks and pathways. They are a member of the French GDR BIM “Bioinformatics for molecular biology”.

Local Environment And Institutional Synergies

The project will benefit from a strong internal UGSF collaboration, notably dr. U. Cenci and co-workers (“Integrative biology of storage polysaccharides”) who will provide phylogenetic analyses, and dr. A. Harduin-Lepers and co-workers (“Molecular mechanisms of glycosylation in health and disease”) who will perform experimental validation. In addition, this PhD project is embedded in a highly integrated local research environment at the University of Lille, combining recognized expertise in glycobiology with advanced structural bioinformatics and evolutionary analysis. The host team is a coordinating member of the cross-disciplinary PIE project, which aims to understand how protein-protein interactions emerge, evolve, and constrain biological function. The present project directly contributes to these objectives by investigating evolutionarily conserved interaction patterns between consecutively acting CAZymes and by integrating phylogenetic signals into large-scale structural modeling of protein assemblies.

The project will also strongly benefit from recent investments in digital research infrastructure at the University of Lille. Through the cross-disciplinary PIE project and the involvement of the host team in the national MUDIS4LS initiative, coordinated by the *Institut Francais de Bioinformatique* and funded within the French *Programme d'Investissements d'Avenir*, 22 GPU's have been added to the local high-performance computing infrastructure. These resources are specifically adapted to deep learning-based protein structure prediction, large-scale docking, and massive sampling approaches such as MassiveFold. This infrastructure enables statistically robust exploration of CAZyme assemblies and ensures efficient integration with parallel experimental efforts, reinforcing the University of Lille's strategic positioning at the interface of glycoscience, evolutionary biology, and AI-driven structural biology.

References

1. Raouraoua, N. *et al.* MassiveFold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling. *Nat. Comput. Sci.* **4**, 824–828 (2024).
2. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
3. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
4. Lensink, M. F. *et al.* Impact of AlphaFold on structure prediction of protein complexes: The CASP15-CAPRI experiment. *Proteins* **91**, 1658–1683 (2023).
5. Raouraoua, N., Lensink, M. F. & Brysbaert, G. MassiveFold Data for CASP16-CAPRI: A Systematic Massive Sampling Experiment. *Proteins* https://doi.org/10.1002/prot.70040 (2025) doi:10.1002/prot.70040.
6. Harrus, D. *et al.* The dimeric structure of wild-type human glycosyltransferase B4GalT1. *PloS One* **13**, e0205571 (2018).
7. Khoder-Agha, F. *et al.* Assembly of B4GALT1/ST6GAL1 heteromers in the Golgi membranes involves lateral interactions via highly charged surface domains. *J. Biol. Chem.* **294**, 14383–14393 (2019).
8. Petit, D., Teppa, E., Cenci, U., Ball, S. & Harduin-Lepers, A. Reconstruction of the sialylation pathway in the ancestor of eukaryotes. *Sci. Rep.* **8**, 2946 (2018).
9. Kellokumpu, S., Hassinen, A. & Glumoff, T. Glycosyltransferase complexes in eukaryotes: long-known, prevalent but still unrecognized. *Cell. Mol. Life Sci. CMLS* **73**, 305–325 (2016).
10. Gagneux, P., Panin, V., Hennet, T., Aebi, M. & Varki, A. Evolution of Glycan Diversity. in *Essentials of Glycobiology* (eds Varki, A. et al.) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2022).