UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Towards Precision Glycotherapeutics Through Genome Editing and Next-Generation Glycoanlytics

Permalink

https://escholarship.org/uc/item/1s79t0vc

Author

Li, Haining

Publication Date

2025

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Towards Precision Glycotherapeutics Through Genome Editing and Next-Generation Glycoanlytics

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioengineering

by

Haining Li

Committee in charge:

Professor Nathan Lewis, Chair
Professor Xiaohua Huang, Co-Chair
Professor Kamil Godula
Professor Reem Khojah

2025

Copyright

Haining Li, 2025

All rights reserved.

The Dissertation of Haining Li is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2025

EPIGRAPH

THE MORE I STUDY SCIENCE, THE MORE I BELIEVE IN GOD.

- Albert Einstein

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE iii

EPIGRAPH iv

TABLE OF CONTENTS v

LIST OF FIGURES viii

LIST OF TABLES ix

LIST OF ABBREVIATIONS x

ACKNOWLEDGEMENTS xi

VITA xiii

PUBLICATIONS xiii

FIELD OF STUDY xiv

ABSTRACT OF THE DISSERTATION xv

CHAPTER 1 INTRODUCTION 1

1.1 Glycobiology: a gateway to novel biotherapeutics 1

1.2 Advancing glycomics through artificial intelligence 1

1.3 Development of AI-based methods for overcoming the curse of glycomics data complexity 4

1.4 Multi-omics based approaches 7

1.5 Thesis overview 8

CHAPTER 2 LeGenD: high-throughput N-glycan screening using explainable AI and lectin profiling 10

2.1 Introduction 10

2.2 Results 13

2.2.1 Lectin profiling cannot unambiguously predict glycoprofiles 13

2.2.3 Simulated and experimental lectin-binding profiles are similar 15

2.2.4 LeGenD accurately predicts N-glycosylation in A1AT 19

2.2.5 Crucial Lectins can be identified using SHAP 20

2.3 Discussion 24

2.4 Methods 29

2.4.1 Lectin Selection 29

2.4.2 N-glycan release and preparation from glycoproteins for LC-based glycan characterization 29

2.4.3 IgG and Fetuin B *N*-glycan purification and procainamide labeling 29

2.4.4 UPLC analysis of *N*-glycans from IgG from human sera and fetal bovine Fetuin B. 30

2.4.5 CHO-derived recombinant protein purification and HILIC-UPLC *N*-glycan analysis 31

2.4.6 Analysis of *O*-glycan analysis by MALDI-TOF Mass Spectrometry 32

2.4.7 Enzyme-linked lectin assay 32

2.4.8 Relative and absolute quantities of *N*-glycans 33

2.4.9 Analysis of UPLC chromatograms 33

2.4.10 Analysis of ELLA binding profiles 34

2.4.11 Lectin profile simulation 34

2.4.12 Model tuning and training 35

2.4.13 Output Collection and Normalization 37

2.4.14 SHapley Additive explanations analysis 38

2.4.15 Statistical Analyses 38

2.5 Conclusion 38

2.6 Supplementary figures and tables 39

2.7 Acknowledgements 49

CHAPTER 3 Delineating the Transcriptional Regulation and Phenotypic Impact from Biotherapeutic Glycoengineering 50

3.1 Introduction 50

3.2 Results 53
3.2.1 A panel of 166 clones with 54 different glycoengineering designs 53
3.2.2 Dominant isozymes and off-target effects are apparent from glycoprofiles 54
3.2.3 Changes in glycosylation co-varies with compensatory glycosyltransferase gene expression 56
3.2.4 Stratification of molecular phenotype induced by glycoengineering 59
3.2.5 Metabolic reprogramming underlying cellular response induced by glycoengineering 63
3.3 Discussion 67
3.4 Methods 73
3.4.1 Glycoprofile analysis 73
3.4.2 RNA sample preparation and sequencing 74
3.4.3 RNA-Seq quantification and differential gene expression 74
3.4.4 Cluster comparison analysis 75
3.4.5 Bioprofile FLEX Analyzer and data processing 76
3.4.6 Similarity Network Fusion Analysis 77
3.4.7 Gene Set Variation Analysis 77
3.4.8 Ordinal regression analysis 78
3.4.9 CHO-S derived recombinant protein purification and data collection 78
3.4.10 Statistical Analyses 79
3.5 Conclusion 80
3.6 Supplementary Figures and Tables 80
3.7 Acknowledgements 84
REFERENCES 86

LIST OF FIGURES

Figure 1.1 Involvement of AI models in glycobiology for basic science and medicine. 3
Figure 2.1 Lectin profiles cannot be connected to a single glycoprofile. 15
Figure 2.2 Predicting relative N-glycan abundance using machine learning. 16
Figure 2.3 Simulating lectin profiles for model training data. 18
Figure 2.4 Predicting the A1AT N-glycoprofiles from measured lectin binding profiles. 20
Figure 2.5 The SHAP values explain the contribution to prediction from each lectin. 23
Figure 3.1 Experimental workflow for characterizing glycoengineered CHO cell lines. 54
Figure 3.2 Glycoprofiles of EPO produced in the geCHO cells clustered mutants into similar genotypes. 56
Figure 3.3 Non-targeted glycosylation machinery can be impacted by glycoengineering. 58
Figure 3.4 Bioprofile FLEX phenotype clustering. 61
Figure 3.5 Pathways deregulations in different glycoengineered mutants clustered into three groups based on Similarity Network Fusion. 63
Figure 3.6 Ordinal regression results of GSVA of pathways in the St3gal-family glycoengineered strains. 67

Supplementary Figure 2.1 40
Supplementary Figure 2.2 44
Supplementary Figure 3.1 81
Supplementary Figure 3.2 82
Supplementary Figure 3.3 83
Supplementary Figure 3.4 84

LIST OF TABLES

Table 1.1 AI-based and multi-omics integration methods for analyzing glycomics data from 2017-2021. 5

Table 2.1 UPLC solvent mixtures for IgG and Fetuin B profiling 31

Supplementary Table 2.1 Selected lectins for *N*-glycan lectin profiling 45

Supplementary Table 2.2 RMSE values of the model when tuning the hyperparameters. 48

LIST OF ABBREVIATIONS

CHO	Chinese hamster overy
geCHO	glycoengineered Chinese hamster ovary
HPLC	high-performance liquid chromatography
UPLC	ultra-performance liquid chromatography
MS	mass spectrometry
NGS	next-generation sequencing
AI	artificial intelligence
SHAP	shapley additive explanations
ANN	artificial neural network
ELLA	enzyme-linked lectin assays
rhA1AT	recombinant human Alpha1-antitrypsin
pdA1AT	plasma-derived Alpha1-antitrypsin
EPO	erythropoietin
RMSE	root mean square error
GNN	graph neural network
ZFNs	zinc-finger nucleases
GSVA	gene set variation analysis

ACKNOWLEDGEMENTS

First and foremost, I extend my deepest gratitude to my advisor, Dr. Nathan Lewis, whose unwavering belief in my potential transformed my academic trajectory. When I first joined the lab, Dr. Lewis welcomed me into his research group with trust and enthusiasm, granting me the freedom to explore glycobiology while providing the mentorship I needed to grow as a scholar. Over the years, he has been a constant source of wisdom, patience, and inspiration, guiding me through challenges and celebrating my successes as if they were his.

I am also expressing my deepest gratitude to God, whose unwavering grace and guidance have been the cornerstone of this journey. Through every challenge and triumph, His steadfast presence has provided clarity, strength, and purpose. I am equally thankful for my cherished church family, whose fervent prayers, encouraging words, and steadfast support have been a constant source of renewal. From fellowship meals to midweek prayer circles, your collective faith and kindness created a nurturing environment that sustained me through difficult moments. Special thanks to Pastor Wen for spiritual counsel, Clara Hu and Michelle Ma for practical assistance, and every member who uplifted me in prayer – your ministry of presence made this milestone possible.

Chapter 2, in full, is currently being prepared for submission for publication of the material in a manuscript titled "LeGenD: determining glycoprofiles using an explainable AI-leveraged model with lectin profiling", Haining Li, Angelo G. Peralta, Sanne Schoffelen, Anders Holmgaard Hansen, Johnny Arnsdorf, Song-Min Schinn, Jonathan Skidmore, Biswa Choudhury, Mousumi Paulchakrabarti, Bjorn G. Voldborg, Austin W.T. Chiang, Nathan E. Lewis. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the material in a manuscript titled "Delineating the Transcriptional Regulation and Phenotypic Impact from

Biotherapeutic Glycoengineering", Haining Li, Austin W.T. Chiang, Vahid H. Gazestani, Bokan Bao, Shangzhong Li, Patrice Ménard, Johnny Arnsdorf, Zulfiya Sukhova, Sara Petersen Bjørn, Karen Kathrine Brøndum, Anders Holmgaard Hansen, Bjørn G. Voldborg and Nathan E. Lewis. The dissertation author was the primary researcher and author of this paper.

VITA

2020 Bachelor of Science in Biomedical Engineering, University of Connecticut

2025 Doctor of Philosophy in Bioengineering, University of California San Diego

PUBLICATIONS

Li H., Peralta A G., Schoffelen S., Hansen A H., Arnsdorf J., Schinn S., Skidmore J., Choudhury B., Paulchakrabarti M., Voldborg B G., Chiang A.W.T., Lewis N. E. LeGenD: determining glycoprofiles using an explainable AI-leveraged model with lectin profiling. bioRxiv, Mar 2025, revised and submitted.

Li H., Chiang A.W.T., Gazestani V. H., Bao B., Li S., Ménard P., Arnsdorf J., Sukhova Z., Bjørn S. P., Brøndum K. K., Hansen A. H., Voldborg B. G. and Lewis N.E. Delineating the Transcriptional Regulation and Phenotypic Impact from Biotherapeutic Glycoengineering. Mar 2025, in preparation.

Kulakova L., **Li K.H.**, Chiang A.W.T., Schwoerer M.P., Schoffelen S., Elkholy K., Chao K.L., Shahid S., Kumar B., Murray N.B., Archer-Hartmann S., Azadi P., Voldborg B.G., Marin A., Mariuzza R.A., Andrianov A., Ploss A., Lewis N.E., Toth E.A., and Fuerst T.R. Glycoengineering of the hepatitis C virus E2 glycoprotein leads to improved biochemical properties and enhanced immunogenicity. (2025). NPJ Vaccines, submitted.

Tsai C.M., Hajam I.A., Caldera J.R., Chiang A.W.T., Gonzalez C., Du X., Choudhruy B., **Li H.**, Suzuki E., Askarian F., Clark T., Lin B., Wierzbicki I.H., Riestra A.M., Conrad D.J., Gonzalez D.J., Nizet V., Lewis N.E., Liu G.Y. Pathobiont-driven antibody sialylation through IL-10 undermines vaccination. (2024). JCI, 10.1172/JCI179563.

Jumabay M., Abud E.M., Okamoto K., Dutta P., Chiang A.W.T., **Li H.**, Manresa M., Zhu Y.P., Frederick D., Kurten R., Croker B., Lewis N.E., Kennedy J.L., Dohil R., Croft M., Ay F., Wechsler J.B., Aceves S.S. Eosinophilic Esophagitis Drives Tissue Fibroblast Regenerative Programs Towards Pathologic Dysfunction. Journal of Allergy and Clinical Immunology, 10.1016/j.jaci.2024.11.028.

Li H., Chiang A.W.T., Lewis N.E. Artificial Intelligence in the Analysis of Glycosylation Data. (2022). Biotechnology Advances, 10.1016/j.biotechadv.2022.108008.

Tsai C.M., Caldera J.R., Hajam I.A., Chiang A.W.T., Tsai C.H., **Li H.**, Lázaro-Díez M., Gonzalez C., Trieu D., Martins G.A., Underhill D.M., Arditi M., Lewis N.E., Liu G. (2022) Non-protective

immune imprint underlies failure of S. aureus IsdB vaccine. Cell Host & Microbe, 10.1016/j.chom.2022.06.006.

Zhu T., Liang G., Zhu C., **Li H.**, & Bi, J. An Efficient Algorithm for Deep Stochastic Contextual Bandits. (2021). Association for the Advancement of Artificial Intelligence-21, 10.48550/arXiv.2104.05613.

Zhu C., Zhu T., **Li H.**, Bi J., & Song M. Accelerating Large-Scale Molecular Similarity Search through Exploiting High-Performance Computing. (2019). IEEE bioinformatics and biomedical conference 2019, 10.1109/BIBM47256.2019.8982950.

FIELD OF STUDY

Major Field: Bioengineering

Studies in Bioinformatics, Glycobiology, and Machine Learning
Professor Nathan E. Lewis

ABSTRACT OF THE DISSERTATION

Towards Precision Glycotherapeutics Through Genome Editing and Next-Generation Glycoanlytics

by

Haining Li

Doctor of Philosophy in Bioengineering

University of California San Diego, 2025

Professor Nathan E. Lewis, Chair
Professor Xiaohua Huang, Co-Chair

Glycosylation is crucial to many biological functions. Therefore, its control and monitoring are essential in many fields, from fundamental science to biotechnology and clinical diagnostics applications. For example, there is a pressing need for effective and affordable glycoprotein biopharmaceuticals to combat diseases that affect millions worldwide. For these, it is essential to

control their glycan structures precisely, which can considerably impact their safety and efficacy. However, glycoengineering faces challenges due to the deficient understanding of glycosyltransferases—the enzymes that orchestrate glycan synthesis—and the subsequent effects of their manipulation on the host cell. Moreover, there is a need for high-throughput, cost-effective glycoprofiling techniques that can accurately characterize the resulting glycoforms of valuable biopharmaceuticals. In this dissertation, a comprehensive study will be conducted on the effects of glycoengineering on the cells used for biomanufacturing. This will be done by integrating glycoprofiles, bioprocessing phenotypes, and RNA-Seq data of glycoengineered CHO cells. Accelerating the characterization of their glycoprotein products will also be followed by an AI-based approach to assay N-glycan structures and their relative abundance using lectin-binding patterns. Finally, through the development of further computational modeling lectin binding kinetics and genomic methods, this method will be translated to a Next-generation sequencing platform. Thus, this dissertation will create a powerful new toolkit that will help accelerate cell line development for biomanufacturing

CHAPTER 1 INTRODUCTION

1.1 Glycobiology: a gateway to novel biotherapeutics

Glycosylation is the decoration of biomolecules with carbohydrates(*Essentials of Glycobiology*, 2015; Schjoldager et al., 2020). Through a series of enzymatic steps, glycans can mature and play essential roles in diverse biological functions(Schjoldager et al., 2020). Aberrant glycosylation of proteins can promote disease, including cancer, autoimmune diseases, and inflammatory diseases(M. Wang et al., 2019). Detection of altered glycosylation can provide invaluable information for basic biomedical discoveries and clinical practice(Aizpurua-Olaizola et al., 2018). Indeed, changes in glycan structure are a common feature of oncogenesis and can be used to identify critical glycans for diagnosis or monitoring treatment effectiveness(Silva, 2019). Awareness of glycobiology is increasingly important for biomarker discovery and drug discovery. However, there is increasing need for data analysis tools to enable the discovery of how dysregulated glycosylation results in disease or altered physiology. It is important to understand the machinery for glycan synthesis and degradation as these underly the dysregulated glycosylation.

1.2 Advancing glycomics through artificial intelligence

Glycan biosynthesis is a non-templated enzymatic process(Lopez Aguilar et al., 2017). Specifically, expression levels and kinetics of a variety of glycosyltransferases (GTs) and glycosidases affect the quantity and structures of glycans produced in a specific cell type under different environmental circumstances(Moremen et al., 2012; Nairn et al., 2008a; Spahn et al., 2016; Taujale et al., 2021). The large diversity of enzymatic reactions involved in glycan biosynthesis results in a wide range of complex glycan structures, causing glycomic data analysis

to be more challenging and to progress more slowly when compared with genomics and proteomics(Gebrehiwot et al., 2018). However, advances in glycoanalytics have improved the resolution of glycan measurements, including using mass-spectrometry (MS)(Klein et al., 2018; Ruhaak et al., 2018), high-performance liquid chromatography (HPLC)(Jansen et al., 2018), or glycan-binding molecules (e.g., lectins, antibodies). The burgeoning glycomics datasets(Ranzinger et al., 2011) have been generated with these advanced technologies. These data, both new datasets and those catalogued in databases, provide opportunities for the use of Artificial intelligence (AI) for analyzing experimental glycan data(Chocholova et al., 2018). However, development of AI-based methods in this field are just beginning to be explored, and many opportunities exist to help it achieve its full potential in providing insights into the mechanisms of glycan synthesis and analysis of glycomics data. Five examples of major questions that could be addressed with AI models in glycobiology are summarized in **Figure 1.1**.

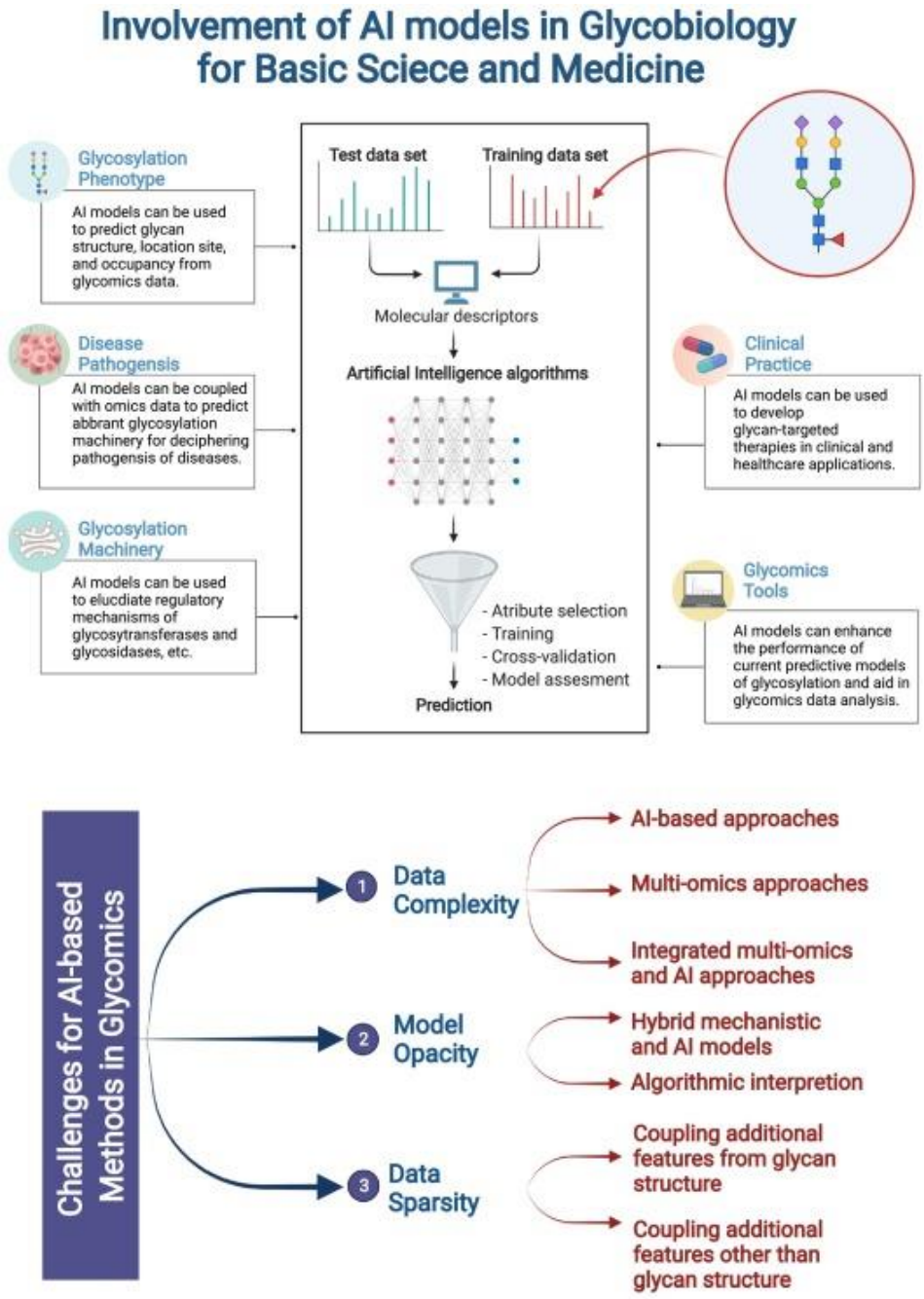


Figure 1.1 Involvement of AI models in glycobiology for basic science and medicine.
(Top) We highlight here five major applications of AI models in glycomics data analysis: 1) to understand glycosylation phenotypes (e.g., glycan structure, location site, and occupancy), 2) to understand glycan-linked disease pathogenesis, 3) to decipher complex glycosylation machinery, 4) to develop therapies to offset effects of aberrant glycosylation, and 5) to enhance current glycomics tools. **(Bottom)** Current challenges and potential solutions in analyzing glycomics data using AI-based methods. There are three major challenges for developing AI-based methods in analyzing glycomics data: 1) data complexity, 2) model opacity, and 3) data sparsity. However, for each challenge, there are methods that could be used for addressing this challenge (right).

1.3 Development of AI-based methods for overcoming the curse of glycomics data complexity

AI, including machine learning and deep learning, has been invaluable for many omics data analyses, such as in genomics, proteomics, and transcriptomics(Mirza et al., 2019), and a few specific applications in glycomics (e.g., to predict glycosylation sites(Akmal et al., 2017) and glycan profiles(Burkholz et al., 2021)). AI-based methods are distinguished by their exceptional ability in finding patterns buried deep in large amounts of data that are impossible for the human brain to process. Without much knowledge required for the underlying biological mechanisms, computational tools can find associations between input samples and labels to make accurate predictions using multivariate statistical models(Theodoratou et al., 2016). However, there has been less AI-based research in glycomics.

Table 1.1 AI-based and multi-omics integration methods for analyzing glycomics data from 2017-2021.

Category	Omic(s)	AI Algorithm	Biological Questions	Reference
AI-based Method	Glycomic (glycosylation reaction)	Random Forest	Predictions of Stereoselectivities	Sooyeon et al. (2020)
	Glycomic	Artificial Neural Network	Predictions of glycoform	Antonakoudis et al. (2021)
	Glycomic (glycan structure)	Deep Learning	Glycan analysis and representation	Thomes et al. (2021)
	Glycomic (UniProt glycosylation dataset)	Artificial Neural Network	Identification of N-linked glycosylation sites	Akmal et al. (2017)
	Proteomic (N-GlycositeAtlas dataset)	Deep Learning	Predictions of structural features and evolutionary information	Pakhrin et al. (2021)
	Manganese, galactose and fucose supplementation	Artificial Neural Network	Predictions of glycoform	Kotidis et al. (2020)
	Glycomic (UniProt glycosylation dataset)	Positive Unlabeled Learning	Predictions of glycosylation site	Li et al. (2019)
	Glycomic (UniProt, dbPTM, GlycoProtDB, etc)	Deep Learning and Support Vector Machine	Predictions of glycosylation site	Taherzadeh et al. (2019)
	Glycomic (UniProt, dbPTM, and O-GlycBase)	XGBoost	Predictions of glycosylation site	Chien et al. (2020)
	Glycomic	Deep Learning	Predictions of lectin–glycan interactions	Lundstrøm et al. (2021)
Multi-omics integration Method	Transcriptomic, proteomic, and glycomic	Non-AI (Statistical method)	Correlations between genes, proteins and N-glycans	Lakshmanan et al. (2019)
	Genomics, proteomics, and O-glycoproteomics	Non-AI (Statistical method)	Characterization of O-glycosylation in tumor-stroma crosstalk	Bouchard et al. (2021)
	Transcriptomic and glycomic	Non-AI (Mechanistic model)	Predictions of enzymes that synthesize human milk oligosaccharides	Kellman et al. (2020)
	Glycomic, metabolomic, and transcriptomic	Non-AI (Mechanistic model)	Predictions of glycoform	Bezjak et al. (2021)
	Glycomic, metabolomic, and transcriptomic	Non-AI (Statistical method)	Predictions of glycoform	Sumit et al. (2019)
	Glycomic and metabolomic	Non-AI (Statistical method)	Predictions of (dynamic) glycoform	Zürcher et al. (2020)

Table 1.1 summarizes several example studies in this field. Due to space limitations, we will present here only the computational methods that employed AI or multi-omic integration methods for analyzing glycomics data within 5 years (2017-2021). Sooyeon et al.(Moon et al., 2020) recently introduced a regression-based random forest algorithm to predict stereoselectivity of glycosylation. This study used descriptors that precisely describe steric and electronic parameters of every reaction participant in related chemical reactions when constructing random forest trees, thus having an extraordinary accuracy when predicting the subtle nuances of stereoselectivity. By illustrating the influence on glycosylation stereoselectivity from factors like different nucleophiles, electrophiles, catalysts, and solvents, this model provides researchers new insights into the glycan biosynthesis reactions. Another method was developed by Antonakoudis et al.(Antonakoudis et al., 2021), which employed a stoichiometric model, combined with artificial neural networks (ANN) to predict glycoforms. The stoichiometric metabolic model portion describes the glycosylation-related metabolic reactions, including 101 metabolites and 144 reactions. Meanwhile the ANN took nucleotide sugar donor fluxes from the stoichiometric model to calculate glycan abundances on the protein of interest. This hybrid model is free of parameterization, making accurate glycoprofile predictions with concentrations solely from key cell culture metabolites. Besides the glycosylation prediction, Thomes et al.(Thomès et al., 2021) developed Glycowork, which focuses on the glycan data analysis. Glycowork is designed with powerful glycan-focused machine learning functions to predict which organism a studied glycan stems from and to identify the most informative motif for this prediction.

These AI-based glycosylation methods make accurate predictions based on large amounts of training data, and they therefore do not require biological parameters such as enzyme kinetics.

However, despite their great prediction performance, these models do not provide details into the mechanisms underlying the model predictions.

1.4 Multi-omics based approaches

Although glycan synthesis is not template-driven, the glycosyltransferases that catalyze the reactions for attaching or detaching monosaccharides onto the glycans are DNA-encoded(Chang et al., 2019). Therefore, omics data enable us to gain valuable information about the relationships between glycans and enzymes in the glycosylation pathways. Multi-omics integration (e.g., using transcriptomics, proteomics, metabolomics, etc.) has been developed for other biological systems and brings benefits to fields such as precision medicine(Huang et al., 2017). Multi-omics integration approaches that include glycomics data for investigating glycosylation mechanisms, however, are appearing more frequently now, and we highlight a few such studies from the past 5 years in Table 1.1.

One recent study combined transcriptomic, proteomic, and glycomic data to investigate the heterogeneity of CHO cells caused by genetic diversity(Lakshmanan et al., 2019). Specifically, the transcriptomic and proteomic data were deployed to train a CHO metabolic network to assess the enriched metabolic pathways. Then, the glycomic data were further analyzed to detect unique glycan species of each CHO parental cell line. The authors concluded that CHO cell heterogeneity causes differences in growth phenotype and glycosylation profiles, which highlighted the interplay between metabolism and glycosylation. Another study conducted by Bouchard et al.(Bouchard et al., 2021) also integrated genomics, proteomics, and O-glycoproteomics to understand how different components cooperate with each other. They found that gene and protein expression can be used to determine enriched pathways that are regulated by O-glycoproteins, which provides mechanistic insights into tumor-stroma crosstalk. In another study, Zachary et al.(Klamer et al.,

2019) integrated on-chip glycan modification data with glycan-array data to help interpret mass spectrometry-based glycomic data and gained valuable information of glycosyltransferase isomers.

The multi-omics integration approaches demonstrate their strengths in connecting the different data types together to study complex processes associated with glycosylation. However, we only see attempts to integrate fewer than three types of omics data in glycomics studies. One major obstacle is the complexity of study design and efforts required for data analysis exponentially increasing as the dimension of omics data types increases.

1.5 Thesis overview

In Chapter 2, I will introduce a method called LeGenD (Lectin to Glycoprofile ENhanced with Data-driven methods) glycoscreening, which uses an artificial neural network (ANN) model to quantify major N-glycan structures using enzyme-linked lectin assays (ELLA). I will train the LeGenD model and validate it using measured recombinant human Alpha1-antitrypsin (rhA1AT) and plasma-derived Alpha1-antitrypsin (pdA1AT) lectin profiles. The glycoprofiles predicted by LeGenD will be compared with those derived from UPLC. Finally, I will use SHAP to identify the dominant lectins responsible for reconstructing the glycoprofile. Taken together, LeGenD opens the door to an era of AI-enabled restoration of glycan structures through high-throughput and cost-effective lectin analysis, thereby demonstrating the untapped potential of lectins for glycan analysis.

In Chapter 3, I will assess the impact of glycoengineering on CHO cells using an expansive library of 166 cell lines, representing 54 different combinations of glycosyltransferase knockouts and knockins. Through integrated multi-omics analysis of glycomics, transcriptomics, and phenotypic data across these cell lines, I will illuminate the complex interplay between glycoengineering and cellular homeostasis, revealing previously unknown pathways by which

glycosylation changes drive phenotypic alterations. These findings will provide critical insights into the broader cellular impact of glycoengineering and underscore the fundamental importance of glycosylation in mammalian cell biology. These insights will have important implications for optimizing biopharmaceutical development and manufacturing processes.

CHAPTER 2 LeGenD: high-throughput N-glycan screening using explainable AI and lectin profiling

2.1 Introduction

Glycosylation affects protein structure, function, and interactions by ensuring protein stability, proper folding, and solubility. This post-translational modification further influences protein function in diverse processes including pathogen binding, cell adhesion, signal transduction, and molecular trafficking(Kukuruzinska and Lennon, 1998; Moremen et al., 2012). Thus, changes in glycan structure can influence many biological and physiological processes(Rocamora et al., 2023; Varki, 2017). For example, changes in glycosylation can modulate inflammatory responses, facilitate viral immune escape, aid metastasis, orchestrate apoptosis, and participate in the pathophysiology of various genetic and infectious diseases(Ohtsubo and Marth, 2006; Reily et al., 2019). Comprehensive profiling of protein glycosylation is critical for biomedical research, including the study of how single sugar residues can considerably alter protein function and activity(Gupta et al., 2023; Shade et al., 2020, 2015), and the identification of prognostic and diagnostic biomarkers for diverse diseases and genetic defects(Drake et al., 2010; Hwang et al., 2010; Kailemia et al., 2017).

Several methods can be used to measure glycan features, such as high-performance liquid chromatography (HPLC)(Dreyfuss et al., 2011), capillary electrophoresis(Mechref et al., 2013; Molnarova et al., 2022), mass spectrometry (MS)(Wuhrer, 2013; Zaia, 2010), nuclear magnetic resonance(Battistel et al., 2014), lectin arrays(Hu and Wong, 2009; Smith and Cummings, 2013), and enzyme-linked immunosorbent assays(Pattathil et al., 2015). While each continues to provide invaluable information addressing specific biological questions, glycomics continues to trail behind the strides achieved by other omics, especially in next-generation sequencing (NGS) of nucleic acids(Hart and Copeland, 2010; Kunej, 2019; Lageveen-Kammeijer and Kuster, 2022;

Raman et al., 2005; Syed et al., 2016). Considering the growing needs for glycobiology research, alternative techniques for assaying and quantifying glycan structures could be beneficial.

Lectins are carbohydrate-binding proteins that recognize mono- and oligosaccharide structures and can be applied in diverse analytical formats(Haab, 2012; Hirabayashi, 2014; Rüdiger and Gabius, 2001) to measure glycan features(Chen et al., 2021; Hu and Wong, 2009). Despite their value for quantifying glycosylation patterns across diverse samples and conditions, lectin arrays and assays remain underutilized, and have yet to be used to measure entire glycan structures(Hu and Wong, 2009). Moreover, samples with considerably different glycan compositions can produce similar lectin profiles, making these differences difficult to observe(Kuzmanov et al., 2013; Lee et al., 2010). Lastly, although successful work has elucidated many lectin-binding motifs, lectin binding affinities can remain ambiguous if they have only been tested on single glycan structures or on a single type of array, nor tested on variable external factors(Klamer et al., 2022; Lehmann et al., 2006; Yom et al., 2024). Therefore, it is challenging to leverage lectin-based approaches to comprehensively elucidate intact glycoform structures in a given glycoprotein, and there is a clear need for more data-driven innovations.

In recent years, artificial intelligence (AI) models have been applied to many biological challenges to analyze diverse data fragments(Bhardwaj et al., 2022), and complex omics data for diverse applications, such as the reconstruction of metabolic pathways(Shah et al., 2021), drug discovery, and biomarkers(Dara et al., 2022; Martorell-Marugán et al., n.d.). Glycomics undoubtedly involve the deciphering of hidden connections, especially given the ambiguity found in glycan structures (from their branching nature and inclusion of repeating structural units) and the various molecules that bind to glycans. Several studies have applied AI tools(Carpenter et al., 2022; Li et al., 2022; Lundstrøm et al., 2022) to glycomics research. Essentially, these studies

successfully extrapolated glycan-binding properties without performing costly and labor-intensive experiments by harnessing neural-network models. Additionally, AI-based approaches(Burkholz et al., 2021; Dai et al., 2021) have been adopted to illustrate the structure and functional relationships of glycans. To facilitate the interpretation of such models applied in glycobiology, explainable AI (XAI) could be used with the black-box AI models. With such XAI methods, explainability is generally defined as the assignment of meaning between model input and output using model-agnostic methods. One such method, SHAP (Shapley Additive Explanations)(Lundberg and Lee, 2017; Lundberg et al., 2020, 2018) is an approach, based on game theory, that dissects the individual predictions of AI models, allowing the computation of Shapley values for the sets of essential input features that guide local and global model result explainability. For black-box AI models in general, explainability as a concept is essential for both i) understanding the real-world significance of the relationship observed between input features and model results, and ii) guiding the refinement of said models to improve upon the predictive accuracy for the objective of interest. As such, the explainability of AI-based methods in glycobiology is necessary to both resolve mechanisms of glycan synthesis and evaluate glycomic data.

Here, we present a method called LeGenD (*Le*ctin to *G*lycoprofile *EN*hanced with *D*ata-driven methods), which uses an artificial neural network (ANN) model to quantify major *N*-glycan structures using enzyme-linked lectin assays (ELLA). Specifically, we generated a large dataset of diverse ultra-performance liquid chromatography (UPLC) glycan profiling experiments using a genetically engineered CHO cell panel. By harnessing established lectin specificities(Bojar et al., 2022), we simulated the lectin-binding profiles of the 309 collected glycoprofiles. We fine-tuned the data using paired lecting profiles and glycan profiles of additional proteins and used this to

train the LeGenD model. The model was then validated using independent lectin profiling data from recombinant human alpha1-antitrypsin (rhA1AT) and plasma-derived alpha1-antitrypsin (pdA1AT). The glycoprofiles predicted by ANN were generally consistent with those derived from UPLC, although there were minor differences in the accuracy of certain glycan species. Finally, we used SHAP to identify the dominant lectins responsible for reconstructing the glycoprofile. Taken together, LeGenD opens the door to an era of AI-enabled restoration of glycan structures through high-throughput and cost-effective lectin analysis, thereby demonstrating the untapped potential of lectins for glycan analysis.

2.2 Results

2.2.1 Lectin profiling cannot unambiguously predict glycoprofiles

Lectins are valuable reagents for measuring glycan features; however, the diversity of glycan features that can be oriented in glycan structures(Hirabayashi, 2004; Nilsson, 2003; Rüdiger and Gabius, 2001) has limited the ability to use lectins to unambiguously determine glycan structures. Furthermore, many lectins have secondary specificities, thereby increasing the potential glycan structures that they bind to(Kletter et al., 2013). To estimate the scale of this problem, we aimed to quantify the diversity of glycoprofiles that could be obtained from a single lectin profile. We generated 2912 random glycoprofiles from a set of 68 known glycan structures that were measured from a panel of 10 glycoproteins expressed in 30 glycoengineered Chinese hamster ovary (geCHO) cell lines (**Methods 2.4**). From these glycoprofiles, we simulated their lectin profiles (details in following sections), which effectively incorporated affinity factors between each glycan feature and lectin (**Methods 2.4**).

From 2912 randomly generated glycoprofiles, we found 100 highly similar simulated lectin profiles ($R > 0.95$), clearly demonstrating that it is nearly impossible to unambiguously reconstruct glycoprofiles from lectin-based glycan features. To gain further insight into the diversity of

glycoprofiles that can be obtained from a single lectin profile, we performed hierarchical clustering using the Voorhees algorithm, which revealed at least seven groups of potential glycoprofiles (depicted as black squares) highly correlated with the simulated mean lectin profiles (**Figure 2.1**). However, each cell constrains the glycoprofiles they produce, for example, based on the species, their enzyme repertoires, monosaccharide availability, and structure of the receiving glycoprotein; thus, the range of possible glycoprofiles would be greatly reduced(Medzihradszky et al., 2015; Raju et al., 2000). Because the quantitative influence of many of these constraints is not well understood, we hypothesized that this can be feasibly overcome by training an ANN on highly diverse glycoprofiles, which accounts for the inherent constraints on glycan patterns imposed by its training data.

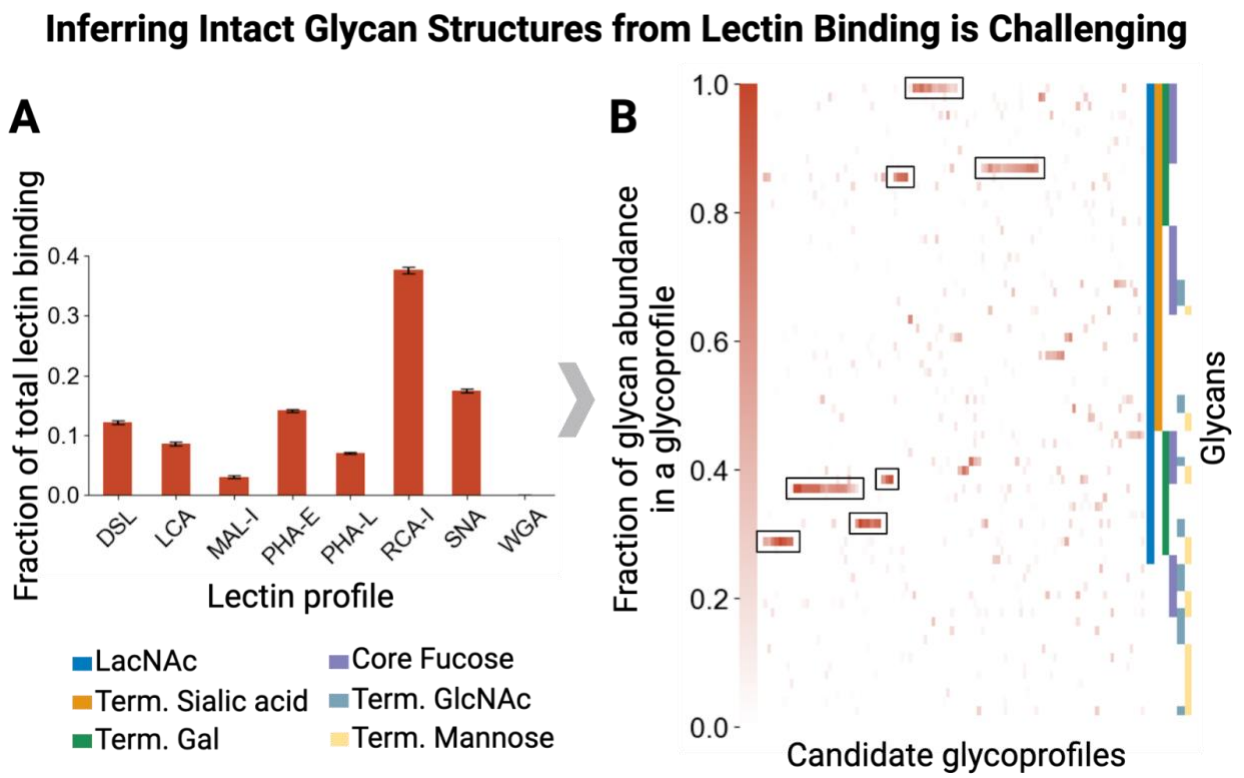


Figure 2.1 Lectin profiles cannot be connected to a single glycoprofile.
From 2912 randomly generated glycoprofiles, we simulated their lectin profiles and then plotted the glycoprofiles corresponding to the 100 most similar simulated lectin profiles (Pearson R correlation > 0.95). **A** Bar plot of mean lectin profiles representing their fraction of total lectin binding, with standard error bars showing the variability for each lectin. **B** The heatmap shows the similar candidate glycoprofiles (black squares) derived from lectin profiles after clustering using the Voorhees algorithm. The glycan features are annotated on the right.

2.2.3 Simulated and experimental lectin-binding profiles are similar

LeGenD uses an ANN to connect patterns of lectin binding to glycoprofiles (**Figure 2.2A**). However, there is a paucity of available datasets containing detailed lectin-binding patterns and their corresponding glycan structural information. To address this challenge, we used 30 geCHO cell lines comprising different combinations of gene knockouts and knockins of 17 glycosyltransferases (**Figure 2.2B**). These cell lines were used to produce a range of different glycoforms of ten recombinant glycoproteins, including monoclonal antibodies, fusion proteins, cytokines, and enzymes (manuscript in preparation). We performed N-glycan analysis using

HILIC-UPLC profiling, which yielded 309 glycoprofiles. We then simulated the lectin profiles from this large glycoprofile dataset (**Figure 2.3**).

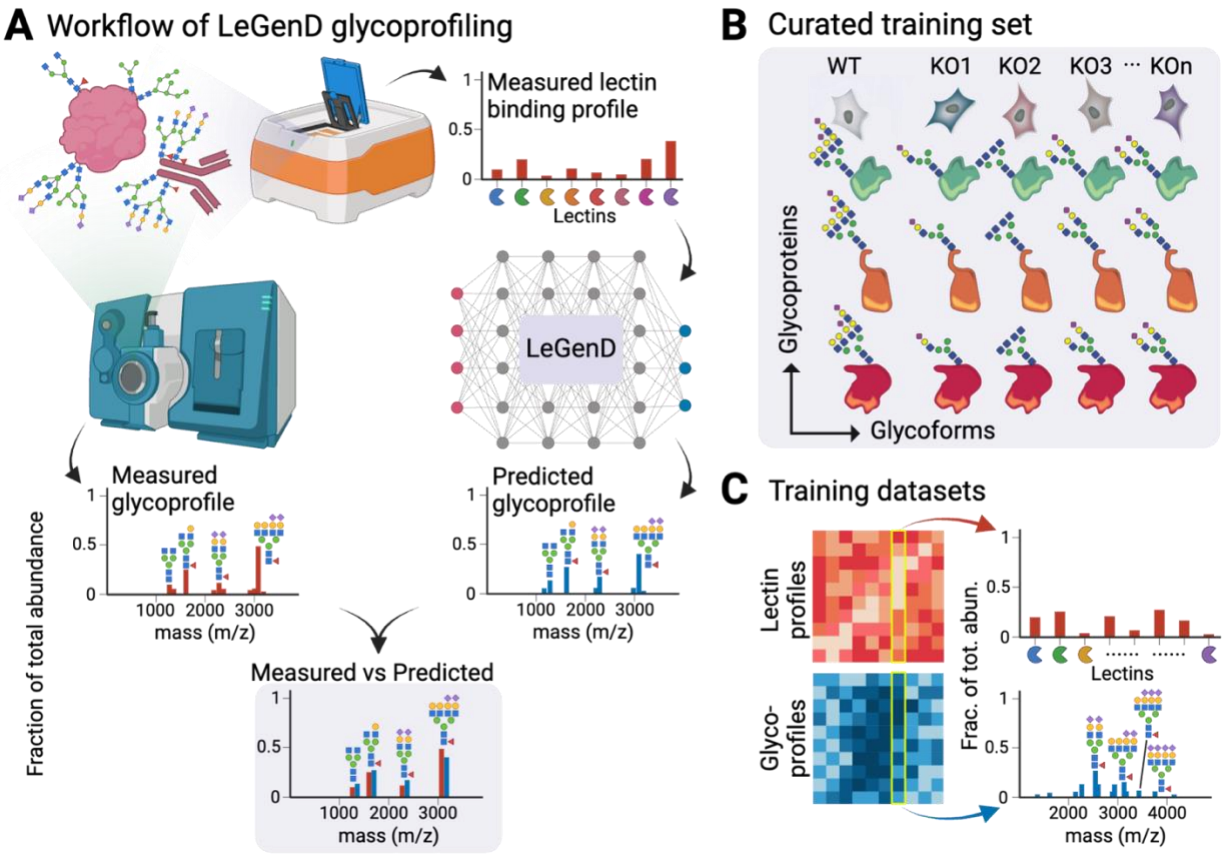


Figure 2.2 Predicting relative N-glycan abundance using machine learning.
A Workflow of LeGenD utilizing neural networks to predict N-glycoprofiles based on lectin-binding profiles. **B** 309 N-glycoprofiles were obtained from 10 glycoproteins expressed in a large panel of 30 glycoengineered Chinese hamster ovary cells (geCHO), yielding a comprehensive view of N-glycans for training LeGenD. **C** Matrix representations of training datasets consisting of lectin profiles and glycoprofiles.

The simulation of the lectin binding profiles was performed by multiplying a glycan-feature matrix (where each glycan is decomposed into counts of expressed features with each glycan feature annotated as per Linearcode(Banin, 2002)) by a binding-rules matrix (**Figure 2.3A**). **Figure 2.3B** demonstrates examples of simulated lectin profiles from measured EPO glycoprofiles. We then aligned some simulated lectin profiles with their experimentally observed counterparts to enhance the credibility of our simulated lectin profiles (**Figure 2.3C**).

While our simulated lectin-binding patterns are based on published binding factors(Bojar et al., 2022), it is imperative to test whether the values are consistent with our experimental data. This is especially crucial to validate our assumption, that the binding intensity is proportional to the amount of glycan features to which the lectin binds. Thus, we validated this using IgG from human sera and bovine Fetuin B as representative glycoproteins, which contain different *N*-glycan structures(Carr et al., 1993; Windwarder and Altmann, 2014; Wuhrer et al., 2007). The *N*-glycans of both glycoproteins were characterized separately using Ultra-Performance Liquid Chromatography (UPLC)(Hanić et al., 2019; Pučić-Baković, 2017), and the relative amounts of respective *N*-glycans were estimated from the peak regions of the fluorescence chromatograms (FLC) (**Figure 2.3C**). Adhering to our simulation rules, simulated lectin profiles were constructed for each glycoprotein. Concurrently, we used ELLA to quantify N-linked glycan characteristics by selecting lectins with glycan epitopes consistently present in both glycoproteins, serving as a benchmark for analytical development (**Methods 2.4**) (**Supplementary Table 2.1**). The selected lectins were also evaluated for cross-reactivity with O-linked glycans by analyzing lectin specificities from the CFG microarray data (**Supplementary Figure 2.1**) and O-glycan modifications in Fetuin B using MALDI-TOF (**Supplementary Figure 2.1**). Although the lectin microarray suggested potential cross-reactivities for certain lectins, O-glycan profiling of Fetuin B revealed the absence of their primary binding epitopes. Thereafter, experimentally derived lectin profiles were generated (**Figure 2.3D**). Linear regression analysis indicated strong correspondence between the simulated and experimental lectin profiles. Moreover, the analysis provided insights into the relationship between the experimental and simulated lectin profiles, allowing us to align the simulation data with the observed data for increased accuracy. Following the application of a linear regression model, our results demonstrate a significant resemblance of simulated lectin

profiles derived from FLC data to the actual lectin profiles for both IgG (RMSE=2.63e-03) and Fetuin B (RMSE=3.85e-03) (**Figure 2.3D**). Hence, the simulated lectin profiles were expected to be suitable training data for LeGenD in subsequent experiments.

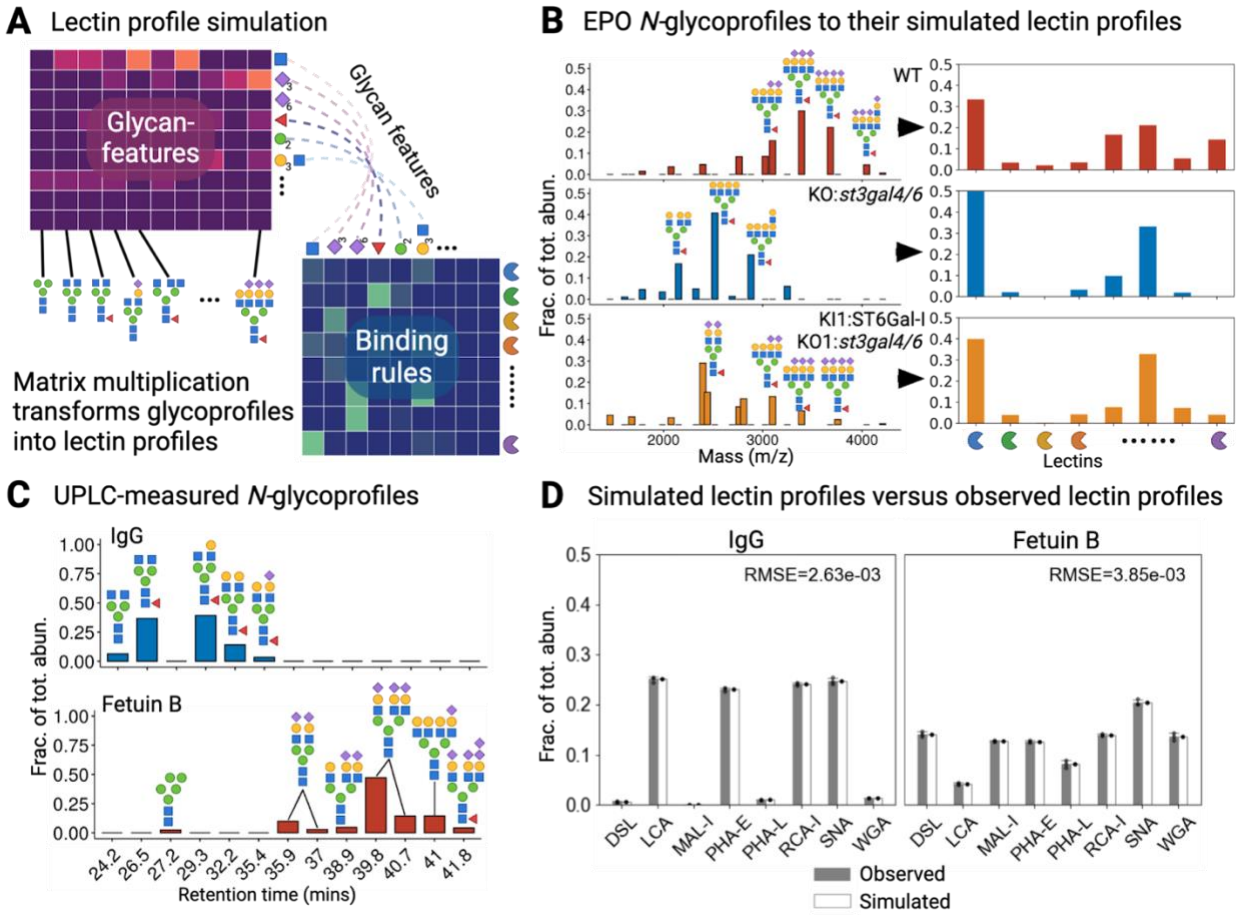


Figure 2.3 Simulating lectin profiles for model training data.
A Schematic approach for simulating lectin profiles from glycoprofiles through matrix multiplication using lectin-binding rules. Glycans are mapped to lectins by decomposing each glycan into its constituent glycan features recognized by each lectin. **B** Simulation of glycoprofiles with their respective lectin profiles. **C** UPLC-derived N-glycoprofiles for IgG from human sera and bovine Fetuin B, Validation of simulated lectin profiling by comparing simulated lectin profiles of IgG and Fetuin B, generated from their UPLC-derived N-glycoprofiles. **D** Validation of lectin enzyme-linked lectin assay (ELLA) data for Bovine Fetuin B samples and IgG for human sera, generated from UPLC analysis, against actual ELLA experiments with standard error bars. The comparison revealed low root mean square error (RMSE) values, confirming the accuracy of our simulation.

2.2.4 LeGenD accurately predicts N-glycosylation in A1AT

We built a fully connected neural-network model (**Figure 2.4A**), connecting eight lectin measurements with 74 distinct *N*-glycan structures stemming from 309 geCHO glycoprofiles. To optimize model performance, we performed hyperparameter tuning and confirmed a model architecture consisting of four hidden layers and 20 nodes in each layer performed the best (RMSE of -3.80E-02) (**Methods 2.4**). The detailed hyperparameter sets tested are listed in **Supplementary Table 2**.

This model was evaluated for its ability to computationally predict N-glycans from the ELLA-derived lectin profiles. We first provided rhA1AT and pdA1AT lectin profiles to the ANN to predict the identity and abundance of the rhA1AT and pdA1AT N-glycans(Rocamora et al., 2024) (**Figure 2.4B and 2.4C**). To denoise, we employed a threshold score of 0.02 and then normalized the predictions to have a summed abundance of 1. The predictive accuracy of the glycoprofiles was evaluated by comparing observed and predicted values, as illustrated in Figure 4. For the rhA1AT glycoprofile, the model achieved a root mean square error (RMSE) of 6.8 and an error of 7.5, indicating moderate agreement between predictions and experimental data. In contrast, the pdA1AT glycoprofile demonstrated markedly higher precision, with an RMSE of 0.056 and an error of 0.0739, reflecting strong alignment between predicted and observed trends. These results suggest that the model performs robustly for specific glycoprofile components, particularly in scenarios where lower error margins are critical. The observed variability in accuracy highlights the potential influence of structural or experimental complexity on prediction efficacy, warranting further optimization for challenging cases. Overall, the findings underscore the model's capability to capture key features of glycoprofiles while identifying areas for refinement.

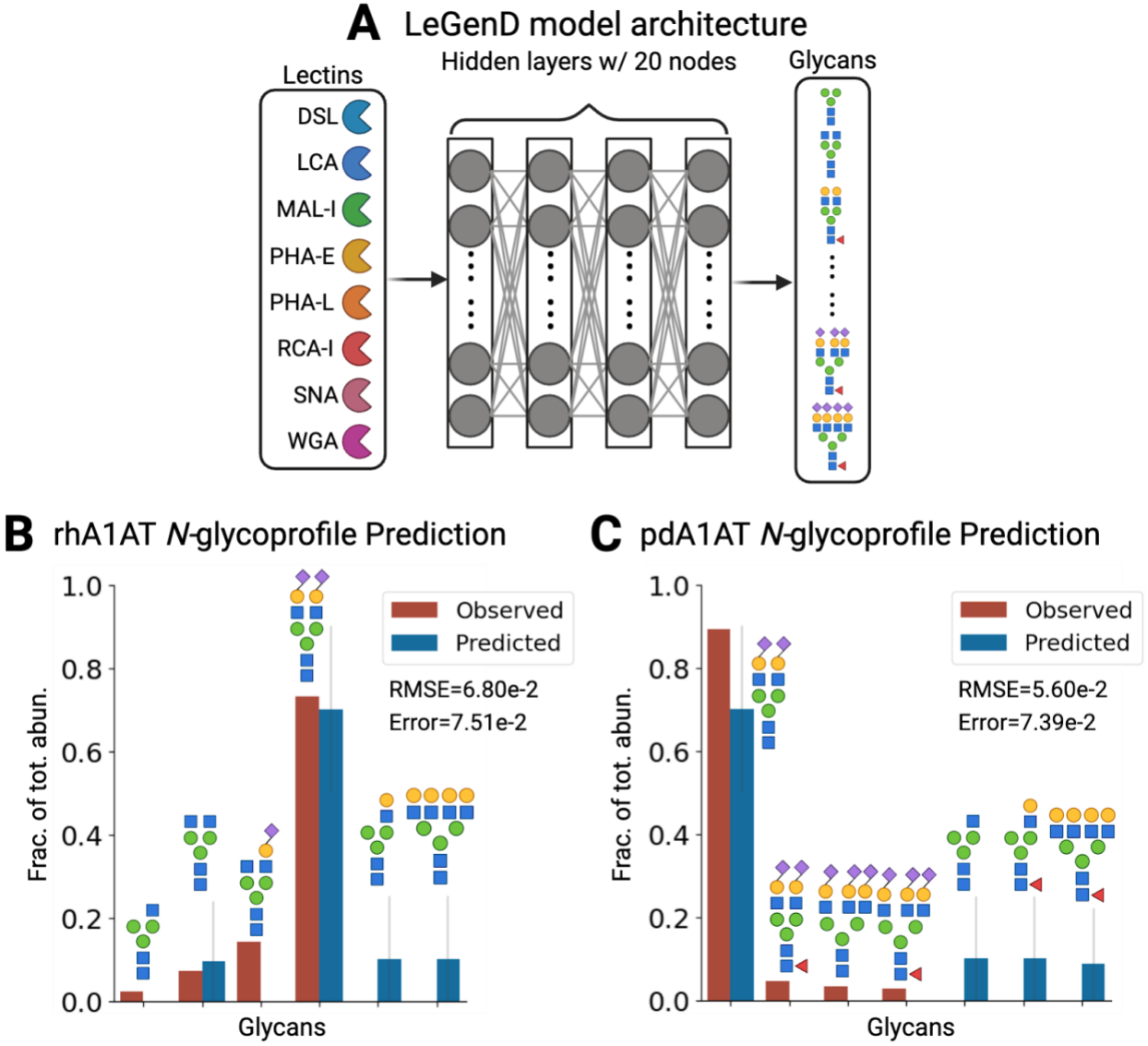


Figure 2.4 Predicting the A1AT N-glycoprofiles from measured lectin binding profiles. **A** The architecture of the LeGenD neural network model, consisted of four fully connected hidden layers with 20 nodes each. **B** Comparison between the experimentally observed rhA1AT-glycoprofile and the LeGenD-predicted N-glycoprofile. The red bar represents the glycoprofile obtained using UPLC, whereas the blue bar represents the predicted glycoprofiles obtained by LeGenD using observed lectin assays as inputs. The fraction of total abundance for each glycan structure was determined. Error bars indicate standard error. The symbol for each glycan follows the SNFG standard. **C** Comparison between the experimentally observed pdA1AT-glycoprofile and the LeGenD-predicted N-glycoprofile.

2.2.5 Crucial Lectins can be identified using SHAP

To interpret the "black box" model in LeGenD, we employed a model-agnostic SHapley Additive exPlanations (SHAP) analysis(Lundberg and Lee, 2017; Lundberg et al., 2020, 2018).

This approach identified key lectins and their associated glycan epitopes contributing to glycoprofile predictions derived from ELLA lectin-binding profiles. For rhA1AT (**Figure 2.5A**), SHAP analysis highlighted SNA, RCA-I, and LCA as the most influential lectins, with absolute SHAP importance values of ~0.5 and ~0.25 above the background (averaged values of glycan abundance in training dataset), respectively. In pdA1AT (**Figure 2.5B**), SNA and LCA retained prominence, while PHA-L and WGA were among the least predictive lectins in both proteins.

SNA's positive SHAP values align with its specificity for alpha2,6-linked sialic acid(Dalziel et al., 1999; Gagneux, 2003), a well-documented modification on A1AT(Castilho, 2014; Rocamora et al., 2024). This reinforces the model's ability to link SNA's binding profile to sialylation levels. LCA, which binds fucosylated glycans, also displayed positive SHAP values. Here, diminished LCA signals—reflecting reduced fucosylation—correlated directly with the model's prediction of abundant afucosylated glycans. This relationship underscores the model's capacity to interpret lectin-binding patterns in a biologically meaningful way, where low LCA activity serves as a predictor of afucosylation.

MAL-I, targeting Siaα2-3Galβ1-4GlcNAc epitopes(Bojar et al., 2022; Geisler and Jarvis, 2011; Knibbs et al., 1991), contributed positively despite the absence of its canonical glycan in A1AT. The model likely associates MAL-I's binding behavior with structural features co-occurring in the dataset, enabling accurate predictions even in the absence of its primary epitope. These results highlight the framework's ability to integrate diverse lectin signals, translating them into coherent glycomic insights without relying on indirect interpretations.

By anchoring SHAP values in A1AT-specific glycan biology, the analysis demonstrates how lectin-binding profiles inform predictions. For example, the bipartite plot for pdA1AT LCA (Fig. 5B) illustrates that low LCA signals (indicative of afucosylation) strongly align with high

predicted afucosylated glycan levels—a direct reflection of established biochemical principles. This approach ensures clarity and relevance, emphasizing the model's fidelity to empirical data.

To interpret the "black box" model in LeGenD, we employed a model agnostic SHapley Additive exPlanations (SHAP) analysis(Lundberg and Lee, 2017; Lundberg et al., 2020, 2018). This elucidated specific lectins and their associated glycan epitopes, which contributed to the prediction of glycoprofiles using lectin profiles acquired through ELLA. The Shapley importance values for a given feature can be understood as the average incremental contribution that the feature offers to the model considering all potential groupings, in our case, the set of lectins chosen. For rhA1AT (**Figure 2.5A**), SHAP revealed that SNA, RCA-I and LCA were the most important lectins, with absolute SHAP importance values of ~0.5 and ~0.25 above the background (averaged training data), respectively. In contrast, MAL-I and PHA-E, while still important, contributed less than 0.1 above the background. In pdA1AT (**Figure 2.5B**), a similar ranking was observed for SNA and LCA based on the aggregate sum of SHAP values. In addition, SHAP analysis indicated PHA-L and WGA as the least important lectins in predicting glycan structures in both rhA1AT and pdA1AT.

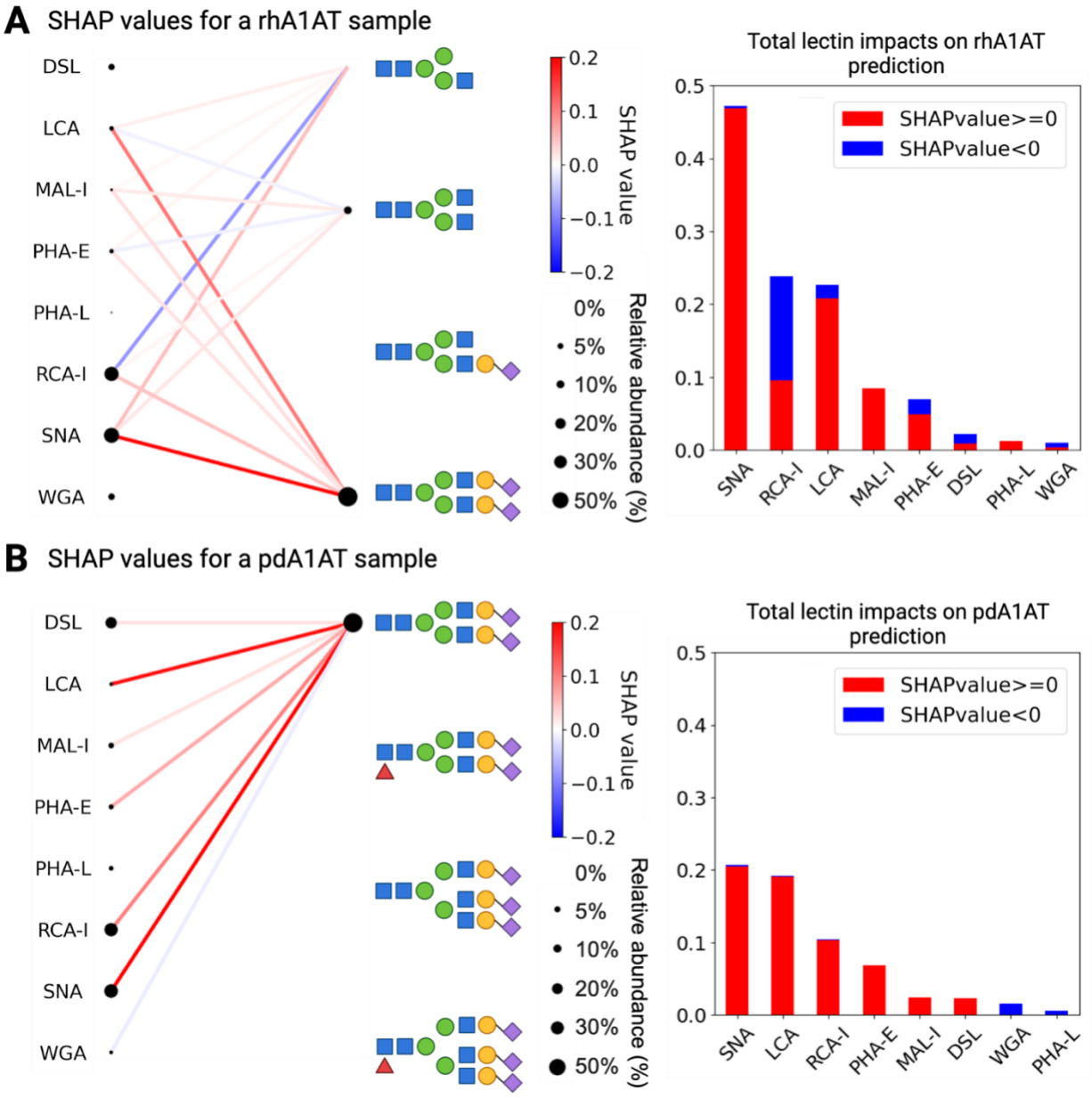


Figure 2.5 The SHAP values explain the contribution to prediction from each lectin.
In the bipartite network on the left, SHapley Additive exPlanations (SHAP) values of lectins and their glycan structures in rhA1AT are displayed. The color scale represents SHAP values, whereas the dot size indicates the relative abundance of a lectin or glycan. To the right, a bar graph breaks down the sum of the SHAP values, illustrating the contribution of each of the eight lectins. Purple diamonds oriented towards the right and left represent α-2,6-linked sialic acids, respectively. **A** a rhA1AT sample. **B** a pdA1AT sample.

2.3 Discussion

In this paper, we introduced LeGenD, a method for deciphering lectin-binding patterns and predicting the identities and quantities of the dominant N-glycans in a given sample. Our AI model, trained with diverse glycoproteins, captures inherent relationships between glycans as determined by the cell. It provides sufficient resolution and sensitivity for quantifying highly expressed glycans and is compatible with different analytical platforms, enabling miniaturization and high-throughput glycan characterization. LeGenD can complement current analytical methods that can encounter challenges in deciphering glycan structural features due to isobaric monosaccharides, stereochemistry(Gray et al., 2019; Wooke et al., 2019), and the need for comprehensive preliminary structural assignments and subsequent validation using orthogonal technologies implemented by experts in glycoanalytics(Chen et al., 2021; Varki et al., 2022). Although several studies have attempted to profile glycans holistically(Rosenfeld et al., 2007) or to predict single glycans from lectin binding patterns(Yom et al., 2024), mapping these patterns to an actual glycoprofile is a highly under-constrained problem, wherein nigh infinite glycoprofiles could have the same lectin profile (**Figure 2.1**). However, LeGenD overcomes this challenge by using AI with a large glycomics training set to successfully compute an N-glycan profile based on a lectin-binding profile. In conjunction with explanatory AI approaches, such as SHAP analysis, we also identified the most predictive lectins.

In the present study, we demonstrated that ANNs can be used to identify an accurate glycoprofile from a lectin-binding profile, despite the immense diversity of glycans that could exist based solely on known enzyme functions. Glycans are highly complex, highly branched, nonlinear biopolymers that are synthesized through seemingly stochastic interactions between glycosyl hydrolases and glycosyltransferases(Lairson et al., 2008; Nairn et al., 2008b; Spahn et al., 2016).

However, it is anticipated that the physicochemical properties of the cell and the glycosylated proteins will limit the range of possible glycan structures and possibly which ones are more likely to co-exist in any given sample (Collins et al., 2017; Hamouda et al., 2014; Reily et al., 2019). Our study shows that ANNs capture these constraints on glycan synthesis, thus enabling the accurate determination of the correct glycoprofiles despite the expected theoretical complexity.

While conventional neural networks have been successful here and in other studies for studying glycosylation patterns and processes (Bao et al., 2021; Bojar and Lisacek, 2022; Coff et al., 2020; Senger and Karim, 2008), further developments in A.I. have presented new model topologies that could further refine technologies, such as ours, for more accessible technologies in glycobiology. For example, glycans have proven to be well-suited for graph neural network (GNN) applications(Bojar et al., 2020; Burkholz et al., 2021; Carpenter et al., 2022; Mohapatra et al., 2022). Indeed, GNNs leverage the full molecular structure or even the geometry for inputs, and the GNN itself learns informative molecular representations to predict the given target properties(Reiser et al., 2022; Wu et al., 2021; Zhang et al., 2021). In the context of glycans, they can comprehensively account for monosaccharides and their linkages, represented as nodes and edges, respectively, to form “glycowords”(Bojar et al., 2020; Daponte et al., 2021). This strategic representation facilitates the capture of unique structural features and contextual information pertaining to glycans including branching substructures, anomericity, and various types of linkages(Daponte et al., 2021). GNNs have been successfully employed for motif exploration within glycan substructures, enabling categorization by *O*-/*N*-linkage, immunogenic properties, evolutionary origin, and viral protein recognition(Bojar et al., 2020; Burkholz et al., 2021). Therefore, we anticipate that integrating alternative model architectures beyond standard neural

networks will further enhance applications for studying glycan structural patterns. These results are highly compatible with our workflow, resulting in an improved prediction performance.

SHAP analysis provided further insights into glycan epitopes that require better reagents for quantification. For example, we found that WGA and PHA-L lectins, which target complex epitopes such as β-1,4-GlcNAc-linked and sialic acid-containing residues(Nabi et al., 2018), tri-, and tetraantennary structures(Bojar et al., 2022), did not significantly improve the predictive performance of LeGenD for rhA1AT and pdA1AT. We hypothesized that, while these lectins can recognize various structures, such as internal complex *N*-glycan structures, there is a need for more specific carbohydrate-binding proteins to enhance the model's overall performance. Therefore, future studies are needed to identify and screen proteins with better binding specificities to the internal structure of glycans. Another important goal for development is to increase the available repertoires of lectins that can be utilized. Not all lectins perform optimally during *in vitro* analyses (Bojar et al., 2022; Haab, 2012; Xie et al., 2020a). However, recent advances have largely demonstrated alternatives to overcome this limitation (Collins et al., 2017; Li et al., 2024; Sterner et al., 2016). Therefore, we anticipate the development of new recognition molecules will aid in accelerating the emergence of high throughput glycomics.

While SHAP effectively reveals the individual contribution of each lectin to glycoprofile prediction, its primary utility is that of an audit of model behavior. This audit facilitates the biological interpretation of model prediction in relation to supplied features. Specifically, when visualized as bipartite plots (**Figure 2.5A and 2.5B**), the local SHAP scores (Shapley values) provide accessible explanations of how lectin abundances influence the model's predictions for rhA1AT and pdA1AT glycoprofiles independently. These explanations reflect the average contribution of a feature value to the prediction across different coalitions, rather than indicating

how the prediction would change if the feature were removed from the model. Consequently, SHAP provides a quantitative rationale for the model's prediction based on the supplied features, similar to explainable models such as linear regression models(Lundberg and Lee, 2017). Therefore, the results should be interpreted as correlations rather than causal relationships(Lee et al., 2023). In summary, SHAP enhances our understanding of the data beyond just the weights from a neural network. Lastly, SHAP will also be valuable in future studies for evaluating different models and improving their explainability, ultimately leading to more accurate predictions and insights from LeGenD.

In addressing the limits of traditional methods for assessing glycan structures and abundances, such as throughput constraints and sensitivity concerns (Thompson et al., 2011; Vliegenthart, 2017), particular examples such as microtiter plate-based techniques (ELLA/ELISA) highlight the need for improvement. These methods, which detect only one target within a single reaction well(Engvall and Perlmann, 1971; Ling et al., 2007) and struggle to achieve high sensitivity comparable to PCR due to their inability to utilize direct amplification techniques(Y. Wang et al., 2019), underscore the importance of finding solutions to enhance the development of in vitro glycan research tools. In response, the integration of glycan binding proteins (GBPs) like lectins with DNA barcoding is promising. This approach enables the identification of lectin-glycan interactions via PCR, facilitating highly-sensitive quantitative glycan analysis (Minoshima et al., 2021; Odaka et al., 2022). Furthermore, within the expanding omics paradigm, multiplexing methodologies are becoming increasingly vital for evaluating glycan complexity and diversity at high throughput scales.

LeGenD is poised to emerge as a rapid, streamlined technology for discerning highly abundant glycoforms across biotherapeutic batches, enabling high-throughput glycan mapping and

monitoring. By combining the quantitative precision of structural composition analysis (e.g., chromatography and mass spectrometry) with the linkage-specific insights of lectin arrays, our method addresses the unmet need for platforms that track predominant glycosylation changes across diverse contexts, offering potential to establish clinically relevant N-glycan signatures. While orthogonal separation-based methods can validate LeGenD's predictions for critical applications, the framework's versatility allows extension to other glycosylation types (e.g., O-glycans) and formats such as lectin arrays(Pilobello et al., 2007; Shibata et al., 2020; Zou et al., 2017), microfluidics(Shang et al., 2016), or Next Generation Sequencing(Kearney et al., 2021; Minoshima et al., 2021; Oinam and Tateno, 2022), each requiring tailored training datasets. To broaden its utility, future iterations would benefit from expanded training data generated via multivariate probability distributions (e.g., Dirichlet distributions), encompassing understudied features like bisecting GlcNAc, non-core fucose, modified sialic acids, and polysialic acids. Though this study focused on N-glycans from standardized glycoproteins (rhA1AT and pdA1AT) to mirror training data constraints, the modular design readily accommodates O-glycans and novel glycoproteins, positioning LeGenD as a scalable tool for advancing glycobiological discovery.

In summary, as awareness of the importance of glycosylation increases, new technologies are needed to increase the accessibility and throughput of glycan research tools. LeGenD provides a complementary concept that with further development could aid in the proliferation of glycoanalytics in routine biological research. Indeed, leveraging AI in glycoanalytics can help manage the complexity of this field, thereby enabling a much larger user base for glycomics and allowing researchers to link changes in glycosylation with phenotypic changes.

2.4 Methods

2.4.1 Lectin Selection

We selected lectins capable of providing a wide coverage across a broad spectrum of N-glycoform epitopes found in our glycoprofiling training dataset. Thirteen preliminary lectins were chosen (data not shown). However, by removing weak binding lectins or are unavailable in biotinylated form, preventing the use of multimerization to increase in vitro binding strength, as well as the inability to perform linear regression during lectin profile simulation resulted in refining our selection to an array eight lectins that still enables a comprehensive characterization of N-glycosylation features(Bojar et al., 2022; Klamer et al., 2019). Specifically, these lectins can distinguish glycan structures such as branches with up to four branches (GlcNAc-β1,2/4/6), LacNAc elongation (GlcNAc-β1,3), epitope monosaccharides (e.g., fucose), and high mannose structures (see **Supplementary Table 1**).

2.4.2 N-glycan release and preparation from glycoproteins for LC-based glycan characterization

150 µg IgG (from human serum; ≥95% SDS-PAGE, Sigma-Aldrich, St Louis, MO) and 300 µg fetal bovine Fetuin (purity ≥90%, Sigma-Aldrich) were each treated with a mixture of PNGase F (purity ≥95%, based on SDS-PAGE and ESI-MS, NEB, Boston, MA, USA) following the manufacturer's protocol.

2.4.3 IgG and Fetuin B *N*-glycan purification and procainamide labeling

The *N*-glycans obtained by PNGase F treatment of Fetuin and IgG were purified using C18 Sep-Pak® 1cc (Waters, Milford, MA, USA) columns. To eliminate salts and hydrophilic species, the purified samples were subjected to an additional cleanup step using graphite cartridges. The C18 Sep-Pak column was initially conditioned with 1 ml of 10% aq MeOH + 1% HOAc three times, followed by sequential addition of 50% aq MeOH, 100% MeOH, and Chloroform. Subsequently, the column was washed twice with 1 ml of 100% MeOH, 50% aq. MeOH, and 1 ml of H_2O five

times. Similarly, HyperSep Hypercarb 25mg Porous graphitic carbon (PGC) columns (Thermo Fisher Scientific) were conditioned with 1 ml of 1% trifluoroacetic acid (TFA), 50% aq acetonitrile (ACN), and 100% ACN three times, followed by 1 ml 50% aq ACN twice, and with 1 ml H_2O three times. The Sep-Pak and PGC columns were further washed with 1 and 3 ml of H_2O, respectively. Finally, *N*-glycans were eluted using 3 ml of 30% aqueous ACN containing 0.1% TFA, and the elution fraction was collected and lyophilized for further analysis.

The reconstituted *N*-glycans were labeled with procainamide via reductive amination, as described by Xie et al(Xie et al., 2021). Briefly, a procainamide solution was prepared by dissolving procainamide in sodium cyanoborohydride in a DMSO:acetic acid:H_2O (280:120:100 v/v/v) solution and mixed thoroughly. Labeling solution was added to each sample and incubated at 65°C for 1 h. ACN was mixed with each sample, and SPE cleanup was performed using GlycoClean S-Cartridges (ProZyme, Hayward, CA, USA). After cleanup, the samples were washed thrice with ACN. Finally, the procainamide-labeled *N*-glycans were eluted with H_2O.

2.4.4 UPLC analysis of *N*-glycans from IgG from human sera and fetal bovine Fetuin B

N-glycan analysis was performed using a Waters Acquity UPLC system attached to a fluorescence detector (Besson Scientific, San Diego, CA, USA). Specified quantities of *N*-glycan were tagged with the fluorophore procainamide and profiled on a Glycan BEH amide column (150 mm x 2.1 mm x 1.7um) (Waters, Milford, MA, USA). A gradient mixture of solvents containing A:100mM ammonium formate (pH-4.5) and B: ACN was used to profile glycans. The solvent gradient conditions used are listed in Table 1. Detection was performed using a fluorescent detector set at excitation and emission wavelengths of 310 and 370 nm, respectively. The column temperature was set at 60°C and the autosampler temperature was maintained at 10°C. The data were processed using Chromeleon™ software (Thermo Fisher Scientific)

Table 2.1 UPLC solvent mixtures for IgG and Fetuin B profiling

Time (min)	Flow (ml/min)	%A	%B
0	0.5	22	78
10	0.5	22	78
52	0.5	47	53
55	0.25	100	0
57	0.25	100	0
60	0.5	22	78
70	0.5	22	78

2.4.5 CHO-derived recombinant protein purification and HILIC-UPLC *N*-glycan analysis

Herceptin, Rituximab and Enbrel were purified by protein A affinity chromatography. For each protein glycoform, a 1-mL MAbSelect Extra column (Cytiva) was equilibrated with 5 column volumes (CV) of 20 mM sodium phosphate, 0.15 M NaCl, pH 7.2. Next, 30-mL supernatant was loaded, the column was washed with 20 CV of 20 mM sodium phosphate, 0.15 M NaCl, pH 7.2, and the protein was eluted using 0.1 M citrate, pH 3.0. The elution fractions (0.5 mL) were collected in deep-well plates containing 100 µL of 1 M Tris at pH 9 per well.

Alpha-1-antitrypsin, butyrylcholinesterase, protein Z, serpin A5, serpin A10, serpin C1, and EPO, all C-terminally tagged with the HPC4 tag (amino acids EDQVDPRLIDGK), were purified over a 1-mL column of anti-protein C affinity matrix according to the manufacturer's protocol (Roche, cat. no. 11815024001). 1 mM CaCl2 was added to the supernatant, equilibration buffer, and wash buffer, respectively. The proteins were eluted in 0.5 mL fractions using 5 mM EDTA in the elution buffer.

For all proteins, elution fractions containing the highest protein concentration were concentrated using Amicon Ultra centrifugal filter units (MWCO 10 kDa). 12 µL of concentrated

protein solutions (concentrations varying between 0.1 and 1 mg/mL) were subjected to *N*-glycan labeling using the GlycoWorks RapiFluor-MS *N*-Glycan Kit (Waters, Milford, MA, USA). Labeled glycans were analyzed by HILIC-FLR using an ACQUITY UPLC Glycan BEH Amide column (2.1 x 150 mm, 1.7 µm) (Waters, Milford, MA, USA) mounted on an Ultimate 3000 UPLC system and a Fusion Orbitrap mass spectrometer (Thermo Fisher Scientific). Acetonitrile (100%) and ammonium formate (50 mM, pH 4.4) were used as mobile phases. The data were processed using Thermo Xcalibur software v4.3 (Thermo Fisher Scientific).

2.4.6 Analysis of *O*-glycan analysis by MALDI-TOF Mass Spectrometry

O-glycan analysis was performed similarly according to Zhong et al (Zhong et al., 2020). Briefly, a known amount of Fetuin B was treated with 50 mM NaOH and 1 M NaBH4 for 16 hours at 45°C. The reaction mixture was slowly neutralized with ice-cold 30% acetic acid, purified using Dowex 50-X cation exchange resin, and lyophilized. Excess boric acid generated was removed by co-evaporation with acidified methanol, followed by methanol. The *O*-glycan was then purified using a C18 cartridge, methylated, and analyzed after permethylation. Mass spectral data were acquired using a Bruker AutoFlex mass spectrometer in positive reflectron mode and analyzed and annotated using GlycoWork Bench.

2.4.7 Enzyme-linked lectin assay

96-well high-bind microplates (Corning, Corning, NY, USA) were coated overnight at 4°C with either 100 µl of 10 µg/mL IgG or fetal bovine Fetuin B using carbonate coating buffer (100mM $NaHCO_3$, 30 mM $NaCO_3$, pH 9.5). The plate was then blocked with a blocking buffer (1X PBS with 2% w/v polyvinylpyrrolidone) for 1 h at 37°C. Biotinylated lectins (Vector Labs, Burlingame, CA, USA) were diluted to 10 µg/ml in binding buffer solution (10mM HEPES, 0.15M NaCl, 0.1mM $CaCl_2$, pH 7.5) and added to the plate for 30 min at room temperature. The avidin-

biotinylated HRP complex (ABC reagents; Vector Lab, Burlingame, CA, USA) was subsequently added according to the manufacturer's instructions. Each step was washed thrice with PBST (Sigma-Aldrich, St Louis, MO, USA). HRP activity was determined by incubation with TMB (Sigma-Aldrich, St Louis, MO, USA) and H_2O_2 at room temperature for 1–10 min and quenching with 1 M H_3PO_4. The color was detected using a Biotek Synergy Mx plate reader (Agilent, Santa Clara, CA, USA) at 450 nm.

2.4.8 Relative and absolute quantities of *N*-glycans

The relative quantity (%) of each *N*-glycan was calculated from the sum of the individual UPLC peak areas in the chromatogram(Kim et al., 2023). Each chromatogram area was generated using LC-ESI-HCD-MS/MS, and the quantity (%) of each *N*-glycan (>0.1 %) was determined relative to the total amount of *N*-glycans (100 %)(Kim et al., 2023). The absolute quantity of non-overlapping glycan peaks was estimated from the fluorescence intensity of the glycan peak in the UPLC chromatogram, using a linear calibration curve (r2 = 0.99) generated from various concentrations of AB or ProA. The quantities of the remaining glycans were calculated using the relative quantities of fetuin and IgG from the absolute quantity represented by the glycan peak(Kim et al., 2023). The glycan quantities of rhA1AT and pdA1AT was obtained from this work(Rocamora et al., 2024).

2.4.9 Analysis of UPLC chromatograms

The main glycan peaks were identified and annotated for each profile. The total abundance fraction of each *N*-glycan was calculated from the fraction of individual UPLC peak areas relative to the total area. Owing to some unannotated peaks, most samples did not reach a summed abundance of 1. Some poorly annotated samples identified as outliers by the Matplotlib boxplot(Hunter, 2007) were excluded. We normalized the rest of the glycoprofile dataset for each

sample to obtain a total abundance of 1. The raw UPLC peak images provided only the overall glycan composition. Therefore, detailed structures and glycosidic linkages were derived based on references from the literature(Szabo et al., 2018; Varki et al., n.d.). Finally, we converted the glycan annotation into linear code(Banin et al., 2002) style for our choice of nomenclature.

2.4.10 Analysis of ELLA binding profiles

The assay signal for each lectin was calculated as the average background-subtracted intensity for each sample. Subsequently, the data for each sample were normalized to a total abundance of 1.

2.4.11 Lectin profile simulation

Based on the literature, we selected eight lectins that mutually recognize ten distinctive common *N*-linked glycan features (Supplementary Table 1). These eight lectins were selected based on two critical factors. First, the selected set of lectins should be able to recognize most of the *N*-linked glycans found in the glycoprofile dataset. Second, these lectins exhibit high specificity and affinity towards their intended glycan epitopes.

For any given glycoprofile (GP), lectin profile (LP) can be generated by using Equations 1 and 2 below.

$$LPg_{i,j} = GF_{i,k} * W_{k,j}, \quad \text{(Equation 2.1)}$$

$LPg_{i,j}$ is the lectin binding profiles for given glycans, where each row represents a glycan and each column represents a lectin; $GF_{i,k}$ indicates the expressed glycan feature k on glycan i; and $W_{k,j}$ is lectin binding rules that tells the frequency of glycan feature k recognized by lectin j. In this study, we assume that glycan features are recognized by their occurrence counts. Binding rules were

adapted from literature(Bojar et al., 2022), where they provided p-values for significant epitopes bound by each lectin. We converted the p values to z-score and used it as our binding rule matrix. The strength of the binding rules is reflected through the z-scores: high z-scores indicate strong binding affinities, as they correspond to low p-values, indicating significant and reliable binding interactions; lower z-scores on the other hand, correspond to higher p-values, suggesting less significant binding interactions.

$$LPg_{n,j} = GPg_{n,i} * LPg_{i,j}, \quad \text{(Equation 2.2)}$$

$LPg_{n,j}$ is the lectin binding profiles for given glycoprofiles, where each row represents a specific glycoprofile, and each column represents a lectin; $GPg_{n,i}$ is the signal intensity (relative MS/HPLC/UPLC relative area under peak) of glycan i in the given glycoprofile k.

To calibrate the simulation, we performed regression fitting using the sklearn.linear_model.LinearRegression function. Regression analysis was used to calculate the estimated coefficients and intercepts for each lectin using the measured and simulated lectin profiles from our IgG and Fetuin B samples. Finally, we applied this linear fitting to $LPg_{n,j}$ to correct the lectin profile simulation.

To demonstrate the problem of mapping a single lectin binding profile to a specific glycoprofile, we generated random glycoprofiles and simulated lectin profiles until we obtained 100 highly correlated profiles (with a Pearson R-correlation above 0.95) from the same glycoprofile. This requires the generation of 2912 different glycoprofiles.

2.4.12 Model tuning and training

We developed a neural network model using lectin profile data as input, where lectins were represented as features in the rows. The model produced multiple continuous values corresponding

to the relative abundance fraction of each glycan. Model training was constructed using the TensorFlow framework, with RMSprop as the optimizer and softmax activation for the output layer. We trained the model on 80% of our training dataset, using a batch size of 4 for 200 epochs, and explored all possible combinations within predefined hyperparameter ranges (Supplementary Table 2) to identify the optimal set for the best model performance. To ensure robust evaluation and mitigate overfitting, we employed 3-fold cross-validation during hyperparameter tuning. The dataset was partitioned into three subsets, with the model trained on two subsets and validated on the remaining subset in a rotating manner. This procedure was repeated three times, ensuring each subset served as the validation set once. The average RMSE across the three folds was used to determine the optimal hyperparameter combination. This entire process was repeated ten times to evaluate the average performance.

A threshold was established to determine meaningful improvements in the model's loss value. When the loss value decrease was less than 0.002, based on the statistical significance within the 95% confidence interval, the model was not considered to have a significant improvement. This threshold ensured that only statistically significant improvements were acknowledged. The model configuration with four hidden layers and 20 nodes yielded the best performance (loss value bolded in **Supplementary Table 2.2**). The relationship between model performance and computational cost is visualized in **Supplementary Figure 2.2.** As shown, computation time consistently increases with the number of layers and nodes in the model. Notably, the steepness of the trend lines indicates that the 4-layer models strike a good balance between complexity and performance, making them particularly effective for the task at hand. This observation further supports our decision to select the model with 4 layers and 20 nodes.

Utilizing the identified optimal hyperparameters, we constructed ten replicate neural network models to analyze IgG and Fetuin B samples. Each model was trained using the full training dataset. The architecture of our models consisting of 4 layers with 20 neurons each, and the rectified linear unit (ReLU) was employed as the activation functions. The learning rate was set at the default value of 0.001 and Mean Squared Error (MSE) was used as the loss function. To ensure robustness and generalizability, we conducted 500 training epochs for each replicate model. For reproducibility, we initialized each model with a randomly generated seed, which also governed the data shuffling process.

2.4.13 Output Collection and Normalization

The outputs generated by each model were systematically collected and normalized using a cutoff threshold of 0.02. To identify the optimal threshold, we evaluated a range of cutoff thresholds (0, 0.01, 0.02, 0.03, 0.05, 0.1) against two key performance metrics: Root Mean Square Error (RMSE) and accuracy. RMSE measures the average magnitude of error between predicted and actual values, providing insight into overall prediction error, While the accuracy metric assesses the binary classification performance of the models. By converting predictions into binary 0/1 outputs, we calculated the percentage of correctly predicted presences or absences of the glycans, irrespective of quantity.

The model performance at different thresholds is detailed in **Supplementary Table 3**. Our analysis determined that a threshold of 0.02 provided the best balance between RMSE and accuracy. Consequently, the threshold was selected for normalizing model outputs in subsequent analyses.

2.4.14 SHapley Additive explanations analysis

We performed SHapley Additive explanations (SHAP) analysis using the KernelExplainer function from SHAP library (version 0.41.0)(Lundberg and Lee, 2017). SHAP provides local interpretation, so we randomly selected one sample from the rhA1AT and pdA1AT data for analysis. SHAP values of all the eight lectins were determined. As we trained and tested ten replicate models, the SHAP values are averages derived from all ten models. These SHAP values, either positive or negative, delineate the extent of their impact within the model. This relationship between the relative abundance of certain lectin-glycan interactions in a given lectin profile and their correlation with a given *N*-glycan structure is effectively illustrated by the corresponding bipartite network diagrams (**Figure 2.5A and 2.5B**).

2.4.15 Statistical Analyses

All statistical calculations were implemented in Python (v3.9; https://www.python.org). For comparing LeGenD predictive performance with UPLC-derived glycoprofiles, we used RMSE. To determine the variability of the predictions across the replicate models, we used standard error. Pearson correlation analysis was employed to assess the similarities between the simulated lectin profiles. Pearson correlation coefficients were reported to quantify the magnitude and the direction of their linear correlations. KDE plots were generated using the Seaborn (v0.13.1; https://seaborn.pydata.org/) library. The glycan structures were drawn using GlycoGlyph(Mehta and Cummings, 2020) (https://glycotoolkit.com/Tools/GlycoGlyph/)

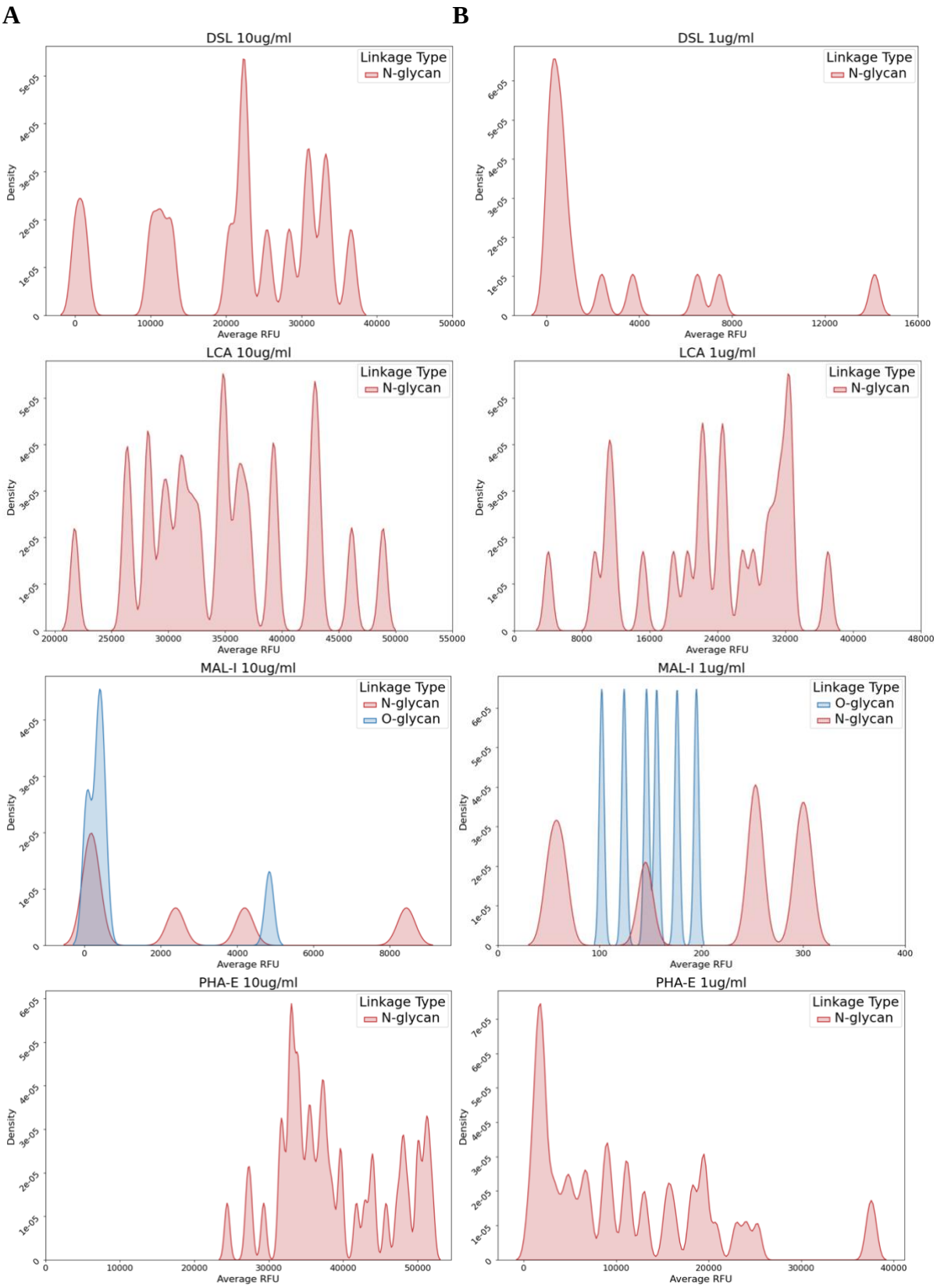
2.5 Conclusion

In summary, as awareness of the importance of glycosylation increases, new technologies are needed to increase the accessibility and throughput of glycan research tools. LeGenD provides a complementary concept that with further development could aid in the proliferation of

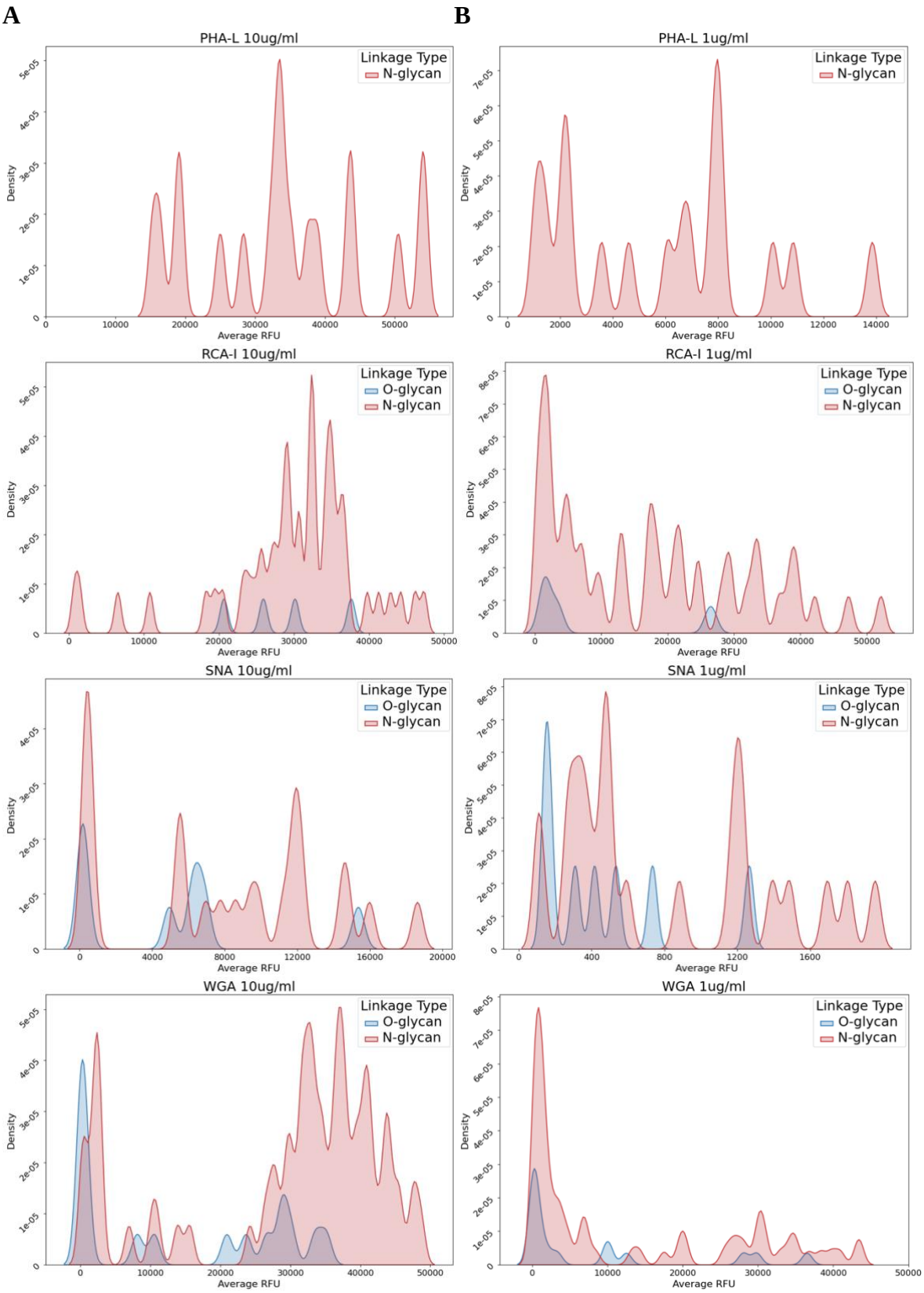
glycoanalytics in routine biological research. Indeed, leveraging AI in glycoanalytics can help manage the complexity of this field, thereby enabling a much larger user base for glycomics and allowing researchers to link changes in glycosylation with phenotypic changes.

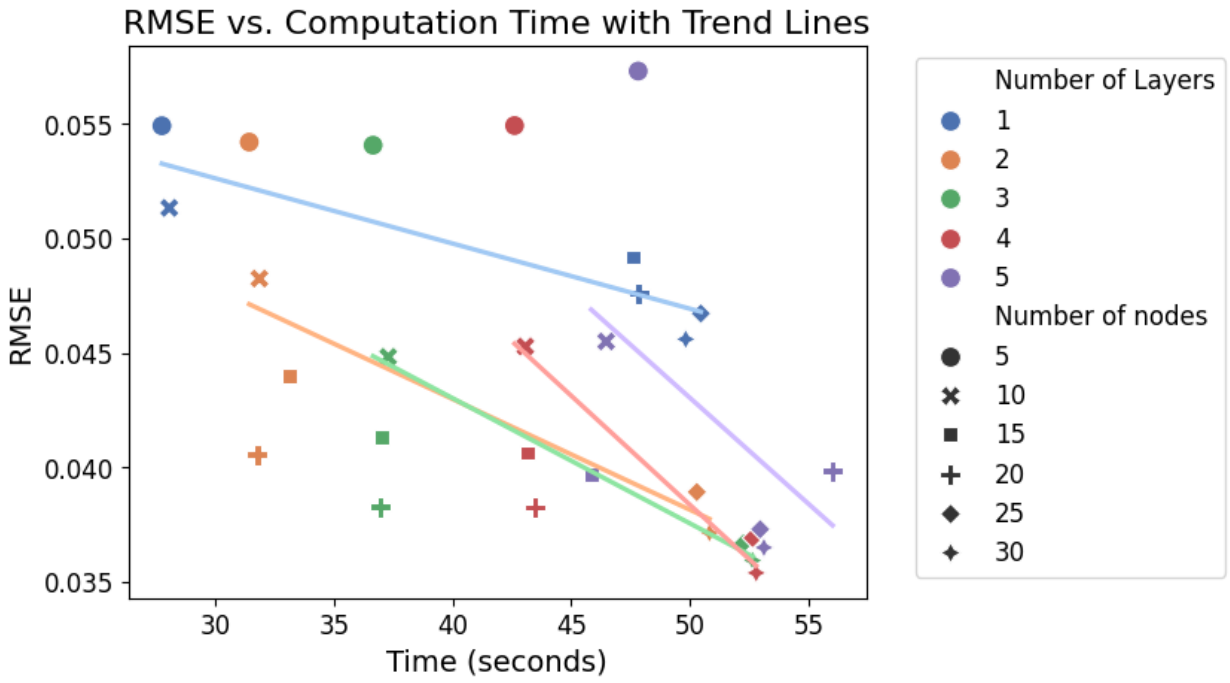
2.6 Supplementary figures and tables

Supplementary Figure 2.1
Kernel density estimate (KDE) plots illustrating the specificity of 8 lectins towards N-glycan and O-glycan epitopes as employed in LeGenD. CFG microarray data from biotinylated lectins at concentrations of A 10µg/ml and B 1µg/ml were analyzed to determine the distribution of N-glycan and O-glycan epitopes targeted by the lectins used in LeGenD (rows). The Average Relative Fluorescence Units (RFU) of lectin signals were plotted, with peaks indicating binding signals for either *N*-glycan (red) or *O*-glycan (blue) epitopes. Higher peak density corresponds to a greater concentration of epitopes at a given Average RFU level.


A
B
DSL 10ug/ml
DSL 1ug/ml
LCA 10ug/ml
LCA 1ug/ml
MAL-I 10ug/ml
MAL-I 1ug/ml
PHA-E 10ug/ml
PHA-E 1ug/ml
Linkage Type
N-glycan
O-glycan
Density
Average RFU

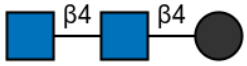
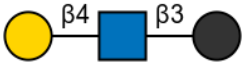
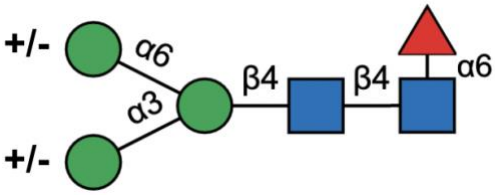
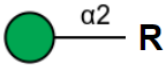
Supplementary Figure 2.2, Continued
Kernel density estimate (KDE) plots illustrating the specificity of 8 lectins towards N-glycan and O-glycan epitopes as employed in LeGenD. CFG microarray data from biotinylated lectins at concentrations of A 10µg/ml and B 1µg/ml were analyzed to determine the distribution of N-glycan and O-glycan epitopes targeted by the lectins used in LeGenD (rows). The Average Relative Fluorescence Units (RFU) of lectin signals were plotted, with peaks indicating binding signals for either *N*-glycan (red) or *O*-glycan (blue) epitopes. Higher peak density corresponds to a greater concentration of epitopes at a given Average RFU level.


A
B
PHA-L 10ug/ml
PHA-L 1ug/ml
RCA-I 10ug/ml
RCA-I 1ug/ml
SNA 10ug/ml
SNA 1ug/ml
WGA 10ug/ml
WGA 1ug/ml
Linkage Type
O-glycan
N-glycan
Density
Average RFU

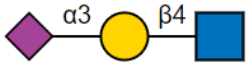
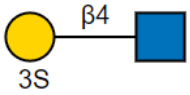
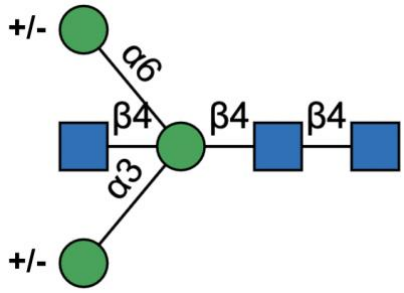
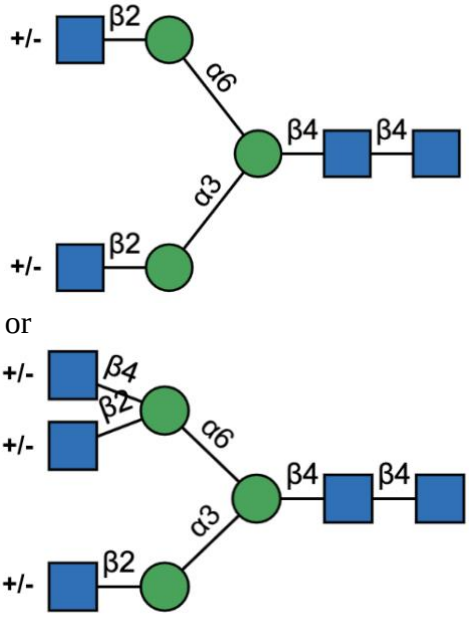


Supplementary Figure 2.3
Scatter plot illustrating how computation time increases as the model gets more layers/nodes. The time reported is for training each model with 80% of the samples over 200 epochs. The different markers represent models with varying numbers of nodes, while the colors correspond to different numbers of layers. Trend lines are generated using linear regression to visualize the relationship between computation time and the number of nodes. The steepness suggests that the 4-layer models balance complexity and performance well.

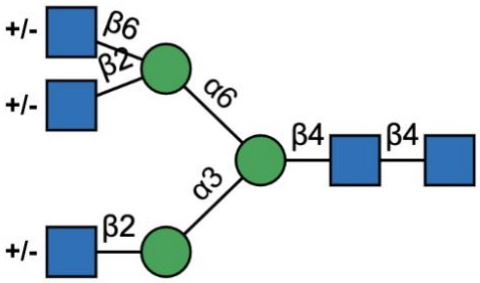
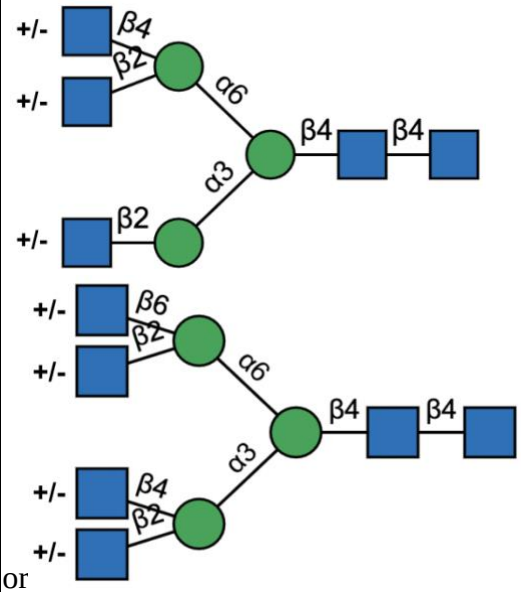
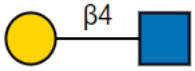
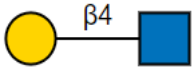
Supplementary Table 2.1 Selected lectins for *N*-glycan lectin profiling

Lectins	**Full Name**	**Preferred *N*-Glycan Epitopes**	**SNFG Icon**
DSL	Datura Stramonium Lectin	Chitin(Crowley et al., 1984; Hořejší and Kocourek, 1978)	 *n
		Type II poly LacNAc(Cummings and Kornfeld, 1984; Itakura et al., 2017)	 *n
LCA	Lens Culinaris Agglutinin	Core fucose(Tateno et al., 2009)	
		Terminal Man α1,2(Tateno et al., 2009)	

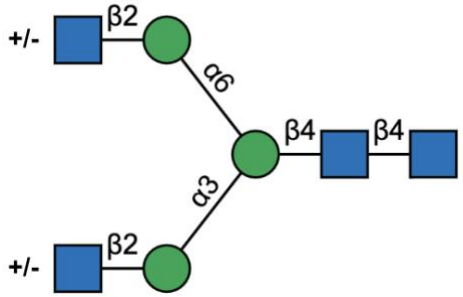
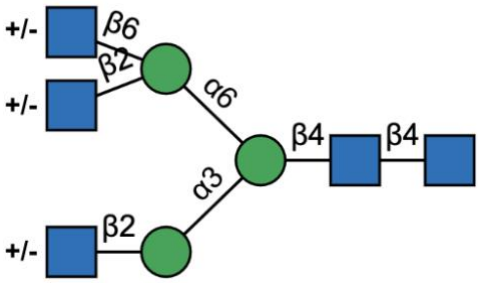
Supplementary Table 2.2 Selected lectins for *N*-glycan lectin profiling, Continued

MAL-I	Maackia Amurensis-I	α2,3 sialylated LacNAc(Geisler and Jarvis, 2011)	
		Terminal 3-O sulfated Gal on LacNAc(Geisler and Jarvis, 2011)	
PHA-E	Phaseolus Vulgaris Erythroagglutinin	Bisecting GlcNAc(Kaneda et al., 2002)	
		Bi/Tri-antennary(Cummings and Kornfeld, 1982; Nagae et al., 2014; Xie et al., 2020a)	
		Galactosylated(Cummings and Kornfeld, 1982; Nagae et al., 2014; Xie et al., 2020a)	

Supplementary Table 2.3 Selected lectins for *N*-glycan lectin profiling, Continued

PHA-L	Phaseolus Vulgaris Leucoagglutinin	β1,6-branched(Cummings et al., 1982; Kaneda et al., 2002)	
		Galactosylated(Bojar et al., 2022; Xie et al., 2020a)	 or
		Galactosylated(Xie et al., 2020b)	
RCA-I	Ricinus Communis Agglutinin I	Terminal type 2 LacNAc(Wu et al., 2006)	
		Galβ1,4(Wang et al., 2011; Wu et al., 2006)	
SNA	Sambucus Nigra Agglutinin	α2,6-sialic acid(Shibuya et al., 1987) (Neu5Ac α2,6(Fischer and Brossmer, 1995))	

Supplementary Table 2.4 Selected lectins for *N*-glycan lectin profiling, Continued

WGA	Wheat Germ Agglutinin	Terminal GlcNAc(Gao et al., 2019)	
		Sialic acid(Ryva et al., 2019)	
		Biantennary(Gao et al., 2019)	
		2,2,6-form triantennary(Gao et al., 2019)	

Supplementary Table 2.5 RMSE values of the model when tuning the hyperparameters.

		Neurons per layer				
		5	10	15	20	25
Number of Layers	1	0.055	0.051	0.049	0.048	0.047
	2	0.054	0.048	0.044	0.041	0.039
	3	0.054	0.045	0.041	0.038	0.037
	4	0.055	0.045	0.041	**0.038**	0.037
	5	0.057	0.046	0.040	0.040	0.037

2.7 Acknowledgements

Chapter 2, in full, is currently being prepared for submission for publication of the material in a manuscript titled “LeGenD: determining glycoprofiles using an explainable AI-leveraged model with lectin profiling”, Haining Li, Angelo G. Peralta, Sanne Schoffelen, Anders Holmgaard Hansen, Johnny Arnsdorf, Song-Min Schinn, Jonathan Skidmore, Biswa Choudhury, Mousumi Paulchakrabarti, Bjorn G. Voldborg, Austin W.T. Chiang, Nathan E. Lewis. The dissertation author was the primary investigator and author of this paper.

CHAPTER 3 Delineating the Transcriptional Regulation and Phenotypic Impact from Biotherapeutic Glycoengineering

3.1 Introduction

There are many therapeutic glycoproteins used for diverse diseases and disorders(Dörner and Kay, 2015; Gils et al., 2017; Halpern and Hutto, 2013; Tomlinson, 2004). The biopharmaceutical landscape has seen remarkable growth, with 197 new biopharmaceutical approvals in the United States and European Union between 2018-2022, including 50 oncology products, 15 for inflammatory disorders, 12 for neutropenia, and 10 for diabetes(Walsh and Walsh, 2022). However, there remain considerable challenges in developing more effective, safer, and more affordable glycoprotein drugs to treat millions of patients across diverse disease indications.

Glycoengineering has emerged as a valuable approach to enhance many therapeutic glycoproteins by modulating their activity, half-life, and immunogenicity(Chiang et al., 2016; Meuris et al., 2014; O'Flaherty et al., 2020; Rocamora et al., 2023; Sha et al., 2016; Solá and Griebenow, 2010). While glycosylation can be controlled in host cells through various strategies (e.g., media optimization, chemical treatments, or genetic modifications(Chang et al., 2019; Makrydaki et al., 2024; Romann et al., 2024)), such interventions are often implemented with a limited understanding of their cellular impacts. This knowledge gap is particularly concerning since glycans play critical roles in cellular processes including protein folding, trafficking, cell signaling, and cell-cell interactions. Studies in animal models and human diseases have shown that changes in glycosylation machinery can have profound physiological effects, with glycosyltransferase deficiencies leading to various pathological conditions(Reily et al., 2019), and many genetic variants likely impact proximal glycan structure(Kellman et al., 2024).

The significance of this knowledge gap is validated by regulatory requirements. The U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) require strict control of glycoprotein quality attributes due to glycans' crucial role in therapeutic efficacy and half-life(Elliott et al., 2003; Ma et al., 2020; Rocamora et al., 2023; "Specifications: Test procedures and acceptance criteria for biotechnological/biological products," 2016; Walsh and Jefferis, 2006). While studies have successfully engineered CHO cells to produce more homogeneous N-glycosylation(Kwak et al., 2017; Yang et al., 2015; Yin et al., 2017), the broader impacts of glycan modifications on host cell gene expression and phenotype remain poorly understood. Understanding these effects is crucial not only for optimizing production but also for ensuring the stability and reliability of production cell lines.

The connection of glycosylation to host cell physiology is complex and demands a comprehensive and integrative approach to understand the impacts of glycoengineering. While single-omic analyses can provide relevant information, the integration of multiple omics allows a much more comprehensive view of the underlying biological mechanisms(Fiocchi, 2023). Glycomics and phenomics provide essential data, reporting glycan structures(Adua et al., 2017) and phenotypic readouts(Bochner, 2003; Cox et al., 2020; Houle et al., 2010). However, integration of additional omics, e.g., transcriptomics, provides crucial information to bridge the gap between glycotype and phenotype(Gutierrez Reyes et al., 2024; Kunej, 2019). Transcriptomics can quantify the expression of genes associated with adding, removing, remodeling, and transporting sugar molecules to create the diverse glycans(Duchez et al., 2011; Nairn and Moremen, 2010), while also shedding light on other pathways and processes that are impacted(Strasser, 2023). This multi-omics integration is particularly useful as it can reveal molecular bases for phenotypic changes and identify compensatory mechanisms in the

glycosylation machinery that might be missed by analyzing any single ome in isolation(Fiocchi, 2023).

In this study, we assess the impact of glycoengineering on CHO cells using an expansive library of 166 cell lines, representing 54 different combinations of glycosyltransferase knockouts and knockins (**Supplementary Table 3.1**). Through integrated multi-omics analysis of glycomics, transcriptomics, and phenotypic data across these cell lines, we investigated three fundamental questions: 1) How do genetic modifications of specific glycosyltransferases affect the biosynthetic steps in glycosylation? 2) How does the knockout of different glycosyltransferase types impact cellular phenotypes? and 3) Which metabolic functions and pathways are altered when glycosyltransferases are knocked out or in? Our analysis revealed that mutants cluster into three distinct groups based on their glycan synthetic patterns compared to wild type, with each group showing unique phenotypic signatures. Notably, we discovered that different glycosyltransferase families (Mgat, B4galt, St3gal) exhibited distinct correlations with phenotypic and pathway alterations. For instance, perturbations in the St3gal family led to significant changes in metabolic pathways, ion transport, and DNA repair processes. To ensure robustness of our findings, we validated the major phenotypic perturbations through analysis of an independent set of CHO clones.

Through this systematic analysis, we illuminate the complex interplay between glycoengineering and cellular homeostasis, revealing previously unknown pathways by which glycosylation changes drive phenotypic alterations. Our findings provide critical insights into the broader cellular impact of glycoengineering and underscore the fundamental importance of glycosylation in mammalian cell biology. These insights have important implications for optimizing biopharmaceutical development and manufacturing processes.

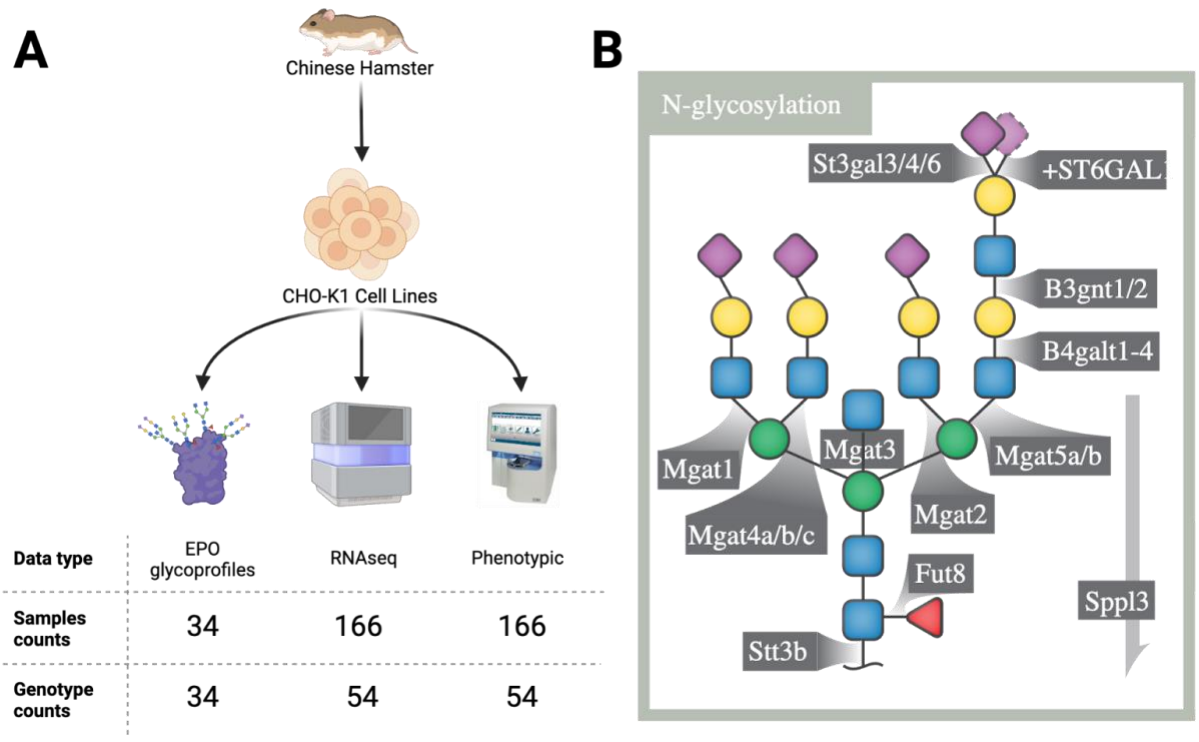
3.2 Results

3.2.1 A panel of 166 clones with 54 different glycoengineering designs

To investigate the cellular impacts of altered glycosylation, an extensive panel of glycoengineered CHO cell lines was developed and characterized through multiple stages. The study design consisted of five major phases (**Figure 3.1**). Initially, a panel of glycoengineered CHO-K1 cell lines using zinc-finger nucleases (ZFNs) to create targeted knockouts of glycosyltransferase genes. These engineered lines were then used to produce and analyze recombinant erythropoietin (EPO), with glycoprofile data from their distinct genotypes reported previously(Yang et al., 2015).

Building on this foundation, we expanded the characterization of these cell lines through comprehensive molecular and phenotypic profiling. From the original panel, we obtained 54 unique genotypes, each represented by approximately 3 biological replicates, yielding 166 total samples. These samples underwent RNA sequencing to profile the transcriptome and phenotypic assessment with the Bioprofile FLEX Analyzer (measures key cellular phenotypes including cell size, growth rate, viability, and concentrations of glucose, lactate, glutamine and glutamate). The multi-omics dataset was then integrated to identify key molecular signatures and pathway alterations associated with specific glycoengineering modifications.

To validate our findings, we generated an independent panel of 30 glycoengineered cell lines derived from the CHO-S cell line and studied these as they transiently expressed 10 different recombinant proteins. This validation cohort underwent both RNA sequencing and glycoprofiling of the recombinant proteins of interest to confirm the key molecular and glycoform changes observed in the primary CHO-K1 panel.



Data type	EPO glycoprofiles	RNAseq	Phenotypic
Samples counts	34	166	166
Genotype counts	34	54	54

Figure 3.1 Experimental workflow for characterizing glycoengineered CHO cell lines. **A)** The data were generated from CHO-K1 cells with combinations of glycosyltransferases knocked out or knocked in. We adapted EPO glycoprofiling from 34 genotypes from Yang's study. Then, we performed a multi-omics analysis of an expanded panel with 54 genotypes (166 total samples), including RNA-Seq and phenotypic characterization. **B)** Key glycosylation-related enzymes targeted in this study, including genes involved in N-glycan branching (Mgat1, Mgat3, Mgat4a/b/c, Mgat5a/b), sialylation (St3gal3/4/6, St6GAL1), galactosylation (B4galt1-4), fucosylation (Fut8), and other glycan-modifying enzymes (B3gnt1/2, Stt3b, Spp13). Knockouts (e.g., St3gal3/4/6) and knock-ins (e.g., +ST6GAL1) were designed to perturb specific glycan biosynthesis pathways, enabling systematic analysis of their roles in glycosylation efficiency and product quality.

3.2.2 Dominant isozymes and off-target effects are apparent from glycoprofiles

Glycosyltransferases are vital for eukaryotic physiology, with enzyme deficiencies potentially altering cell functions(Conroy et al., 2021; Schjoldager et al., 2020; Stanley, 2016). However, to understand how glycoengineering impacts the host cell, we first need to assess which genetic changes induce a meaningful change to glycosylation. EPO was previously expressed in 34 different glycoengineered cell line genotypes and subsequently glycoprofiled(Yang et al., 2015). We reanalyzed these data using Glycompare(Bao et al., 2021) to quantify all glycan intermediates

(glyco-motifs) for each of the 34 EPO glycoprofiles(Yang et al., 2015). This allowed us to identify glycoengineering designs that led to structurally similar glycans. After obtaining the glyco-motif abundance profiles, we subjected these to analysis using hierarchical clustering to identify mutants with similar glycan biosynthesis.

Upon analyzing the EPO glyco-motifs, we found the mutant cell lines clustered into a few groups with similar glycoprofiles (**Figure 3.2**). In particular, several single-glycosyltransferase knockouts clustered tightly with the wild-type parental CHO-K1 cell line. This "Wild type-like" group (green dendrogram), have mutations in glycosyltransferases that are either minimally functional, are supported by active and redundant isozymes. The next cluster, or "Moderate" group (blue dendrogram), included genotypes with 1 or 2 glycosyltransferase knockouts. As expected, with additional knockouts, we observed a clear clustering of samples with extensive changes in glycosylation, which we designate as the "Substantial" group (red dendrogram).

Each group included knockouts involved in similar glycan synthesis pathways. As expected, clones with multiple sialyltransferases targeted (e.g., St3gal4/6, St3gal3/4, and St3gal4) grouped together, separate from the 'Wild type-like' group. Similarly, samples from clones wherein Mgat5, Mgat4a/4b, and Mgat4b, were knocked out also clustered more tightly and showed a decrease in the tetra-antennary-related glyco-motif clusters (GM clusters: GM6 and GM7).

The clustering of glyco-motifs informs which glycosyltransferases cause minor modifications and which combinations disrupt the wild-type glycoform. Indeed, different isozymes within the same glycosyltransferase family could show different magnitudes of impact on the glycoprofiles. For example, we knocked out multiple St3gal family isozymes (i.e., St3gal3, St3gal4, and St3gal6), but only mutants with St3gal4 knocked out showed a decrease in sialylation; meanwhile, the St3gal3 mutant showed "wild type-like" glycoprofiles and the St3gal6 knockout

clustered as “Substantial” due this clone resulting in decreased lacNAc elongation compared to the other St3gal KOs. Similar observations were found with the Mgat family. Dominant isozymes in the moderate and substantial groups included Mgat4b/5 and B4galt1, St3gal4, and St3gal6.

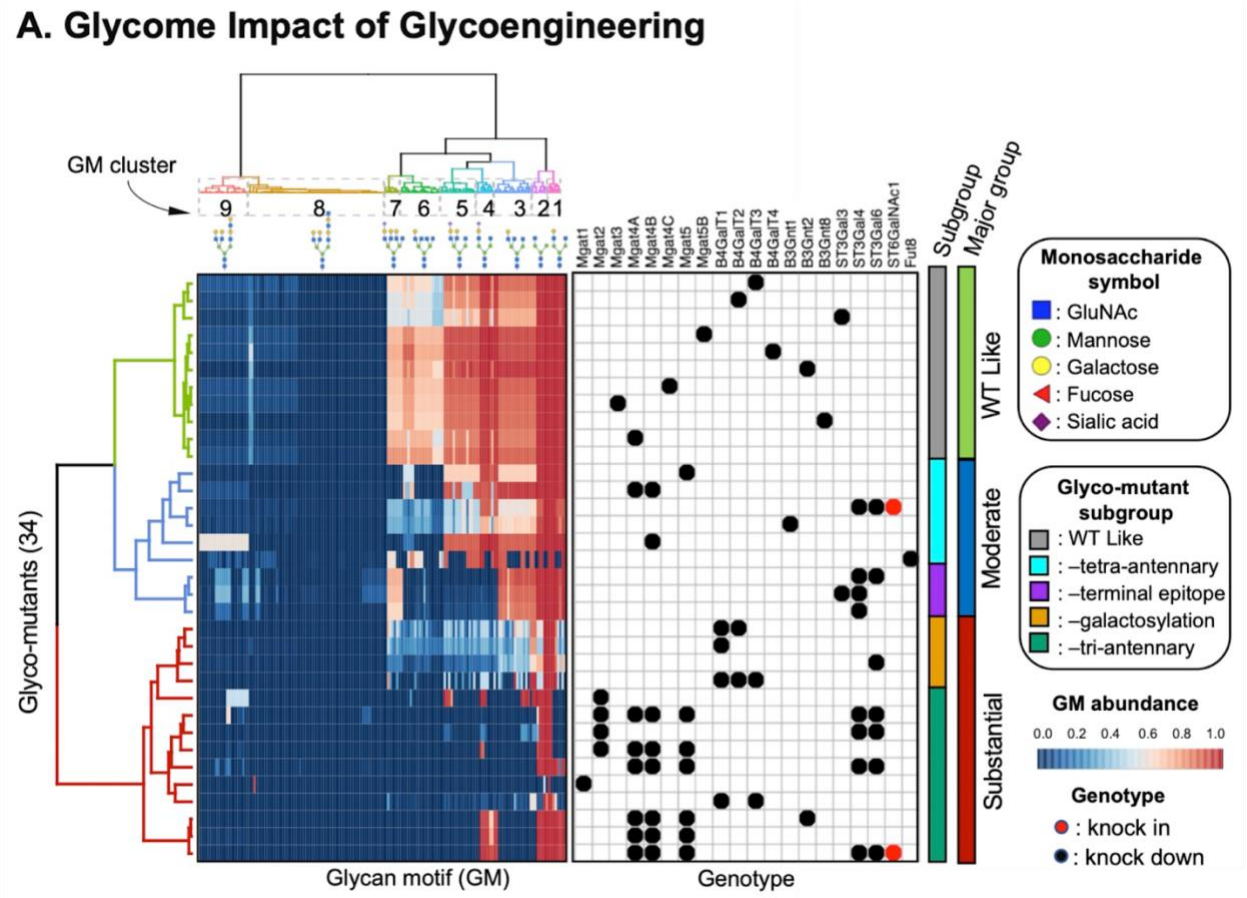


Figure 3.2 Glycoprofiles of EPO produced in the geCHO cells clustered mutants into similar genotypes.
Glycoprofiles were clustered using GlyCompare, and mutants with similar glycoprofiles were identified. Clustering was driven by nine glyco-motif clusters (GM 01-09), in which each cluster is annotated with the representative glycan topology (greater than 60% of the glyco-motif members within a cluster share the same substructure). There are three primary clusters of mutants on the 34 glycoengineered samples: WT-like (green), Moderate (blue), and Substantial (red).

3.2.3 Changes in glycosylation co-varies with compensatory glycosyltransferase gene expression

The clustering of glycoprofiles identifies mutants with similar glycosylation, but do the cell lines show compensatory gene expression? We next tested if the changes in glycosylation result in differential expression of the remaining glycosyltransferases. Here, we first tested if

unexpected changes in the glycome (i.e., beyond the expected changes from the known enzyme knockouts) could be explained by further transcriptomic changes. This survey identified a short list of genes with variations sufficient for delineating a transcriptomic signature specific for each glycoengineered mutant (**Figure 3.3**). We compared the clustering of the mutant glycoprofiles and the transcriptomic signature for each mutant and found that both data types clustered similarly (**Figure 3.3A**), as quantified using the Fowlkes-Mallows score (FMS) ($p = 2 \times 10^{-4}$) (**Figure 3.3B**). The clusters have an FM index of 0.64, indicating a high level of agreement between the glycosylation clusters and the gene expression clusters.

A subsequent differential gene expression analysis (**Figure 3.3C**) showed multiple glycosyltransferase genes were impacted in some mutants, beyond just the targeted genes. For example, in the mutant "B3gnt2/Mgat4a/4b/5", St3gal3 and St8sia4 were down-regulated, and B3galt6 was up-regulated (highlighted in **Figure 3.3C**). Taken together, we conclude that glycosylation changes lead to gene expression changes in the glycosylation machinery.

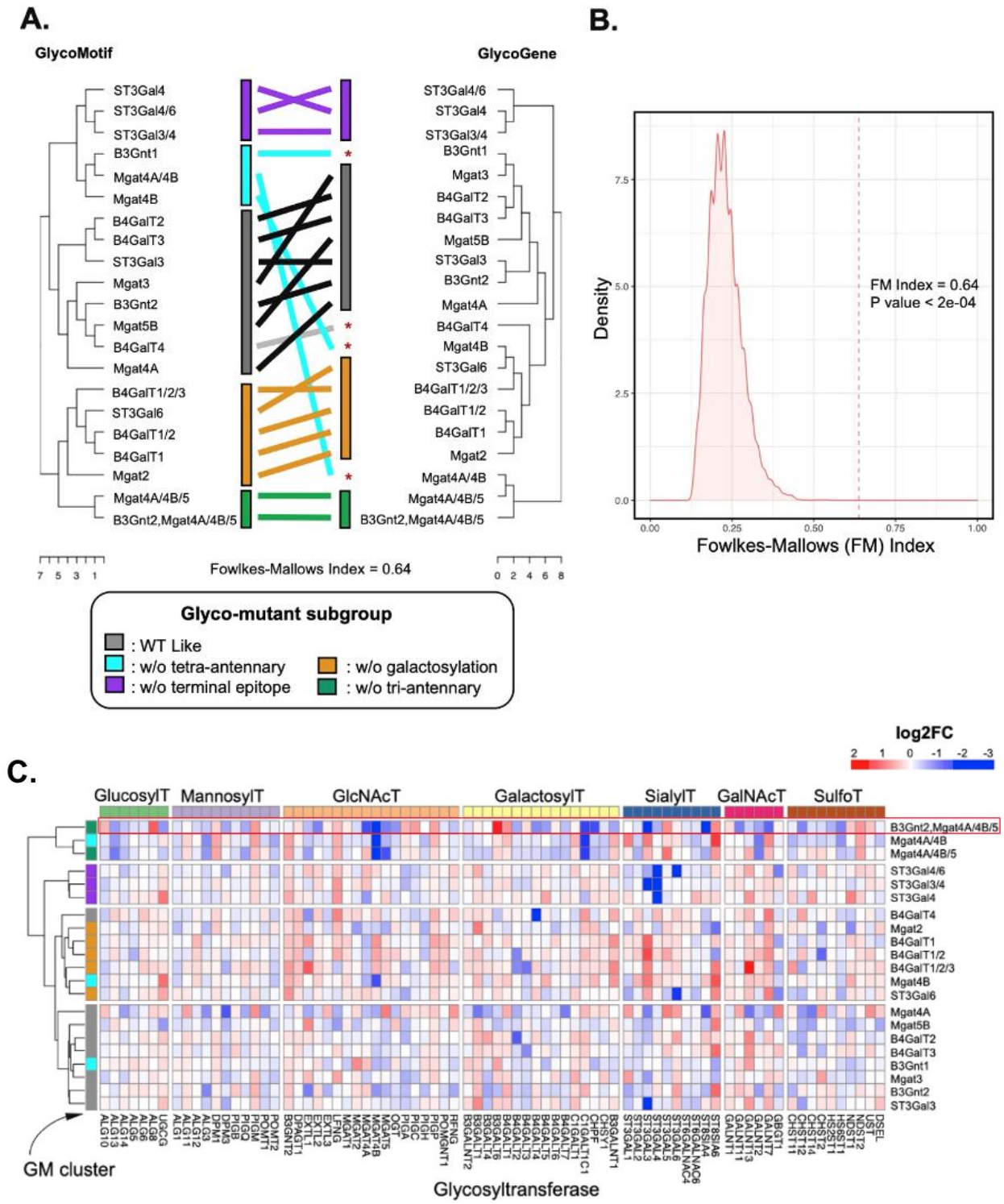


Figure 3.3 Non-targeted glycosylation machinery can be impacted by glycoengineering.
A) Glycoprofiles and gene expression of mutants clustered similarly. **B)** The FM index suggests a high similarity between the two clusterings. The distribution of randomized FMS across various clustering comparisons is visualized in a density plot, with the FM index plotted along the x-axis and the density values along the y-axis **C)** Differential gene expression analysis shows changes in the expression of additional glycosyltransferases beyond the target genes.

3.2.4 Stratification of molecular phenotype induced by glycoengineering

Another major concern in glycoengineering is whether the glycoengineered mutants have significantly altered phenotypes that might harm product production or threaten host cell survival. To explore the phenotypic relationships among these glycoengineered mutants, we analyzed the phenotypic profiles of 54 glycoengineered mutants using the Bioprofile FLEX analyzer. Hierarchical clustering showed that many of the sialyltransferase (e.g., St3gal3, St3gal4, St3gal6) knockout mutants clustered tightly as a group (**Figure 3.4A).** This 'B01' group (green dendrogram) showed the 'Growth Rate' and 'VCD' are significantly increased. The second group of mutants included galactosyltransferase (e.g., B4galt1, B4galt2, B4galt3, B4galt4) knockout mutants (termed the 'B02' group; red dendrogram); in these cell lines, the 'Cell size' significantly increased. The branching-related glycosyltransferases (e.g., Mgat4a, Mgat4b, and Mgat5) knockout mutants showed another distinct 'B03' group, in which the 'Viability' is significantly increased.

To corroborate the robustness of the clustering and to further explore the spatial distribution of these phenotypic variations, Principal Component Analysis (PCA) was performed (**Figure 3.4B**). The first PC, which accounts for 33.9% of the total variation, separates the glycoengineered mutants with phenotypes from the 'Cell Size' and 'Growth Rate.' The second PC, which accounts for 28.5% of the total variation, separates the glycoengineered mutants with phenotypes from the 'Viability' and other phenotypes. The dots were colored to correspond to their previous hierarchical cluster assignments. Notably, when we color the dots corresponding to the groups from the hierarchical clustering, the PCA visualization revealed three separate clusters, thereby underscoring the presence of distinct phenotypic patterns within the dataset. These results suggest that different glycosyltransferase mutations impact phenotypes differently. Building on this premise, we conducted targeted analyses to ascertain which glycosyltransferase mutations exert the most pronounced effects on specific phenotypic features (**Figure 3.4C**). We focused on the

knockout effects of St3galt, B4gal, and Mgat families on four distinct phenotypic features: cell size, VCD, viability, and growth rate. We employed t-tests to compare the phenotypes of knockout cell lines against all other cell lines. Among these comparisons, our results revealed notable changes in ‘Cell Size’ following the knockout of Mgat (p-value = 0.0008), ‘VCD’ and ‘Growth Rate’ following the knockout of St3gal (p-value = 0.004, 0.006), and ‘Viability’, ‘VCD’, and ‘Growth Rate’ following the knockout of B4galt (p-value = 0.045, 0.019, 0.03). These findings unraveled the intricate relationships between particular glycosyltransferase activities and their phenotypic manifestations, enabling glycoengineering to screen the knockouts better.

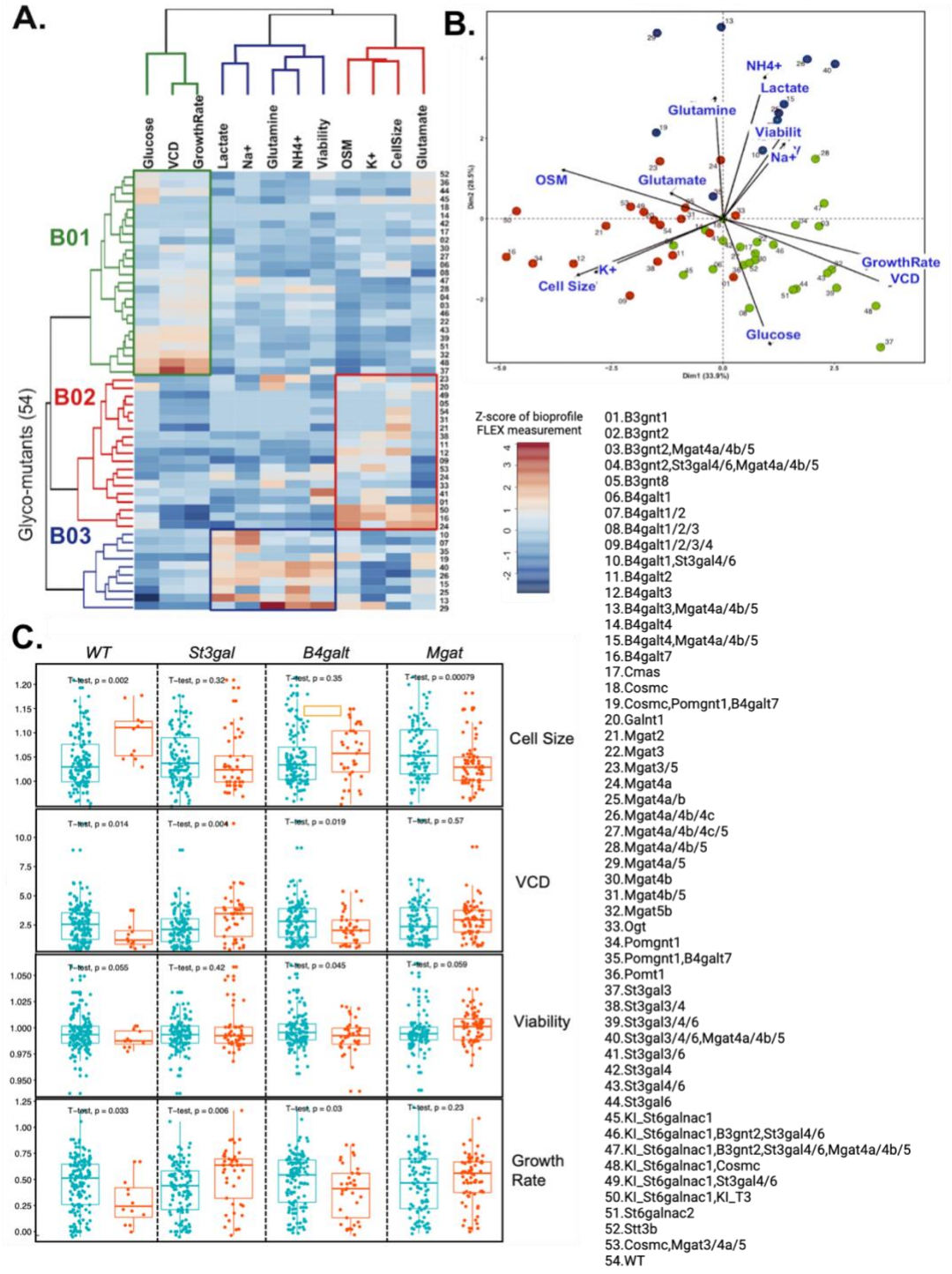


Figure 3.4 Bioprofile FLEX phenotype clustering.
(A) There are three dominant row clusters on the 54 glycoengineered samples: B01 (Growth rate and VCD; green), B02 (Cell size; red), and B03 (Viability; blue). **(B)** Principal component analysis on the 54 glycoengineered samples. Dots are colored to correspond to the cluster colors in (A). **(C)** Phenotypic alterations in different glycosyltransferase-family knockout.

We further integrated phenotypic and gene expression data using Similarity Network Fusion (SNF)(Wang et al., 2014) to reveal the molecular mechanisms underlying cellular response to glycoengineering. SNF allows one to combine distinct data types into a unified similarity network of samples (see Methods for details). In essence, the samples grouped represent a subset of the most similar cellular responses and gene expressions. SNF clustering on the integrated data revealed three latent molecular phenotype (MP) subgroups in the glycomutants (**Figure 3.5**), indicating differences in the cellular responses to different glycoengineering strategies, and each subgroup exhibits distinct transcriptional signatures. The identified subtypes are strongly associated with the PCA clusters (**Figure 3.4B**). Specifically, the 'MP01' mutants (red) exhibited significantly elevated 'Cell size' but reduced 'Growth rate' and 'Viability,' and the 'MP02' mutants (green) exhibited significantly increased 'Ammonia' and 'Lactate.' Interestingly, the 'MP03' mutants (blue) showed an opposite phenotype to the 'MP01' mutants, which significantly enhanced 'Growth Rate' and 'Glucose consumption rate.' To characterize the dysregulated pathways underpinning the variability of cellular responses exhibited by the identified MPs, we assessed the importance of all pathways in the Reactome collection concerning glycoengineering. Our results show that 'MP01' is associated with PC3 and PC4 pathways, 'MP02' is associated with PC1,2,3,5 and 7 pathways, and 'MP03' is associated with PC6 pathways. However, the question remains about how the transcriptional reorganization underpinning each MP-subgroup represents the cellular state of each glycoengineered cell line.

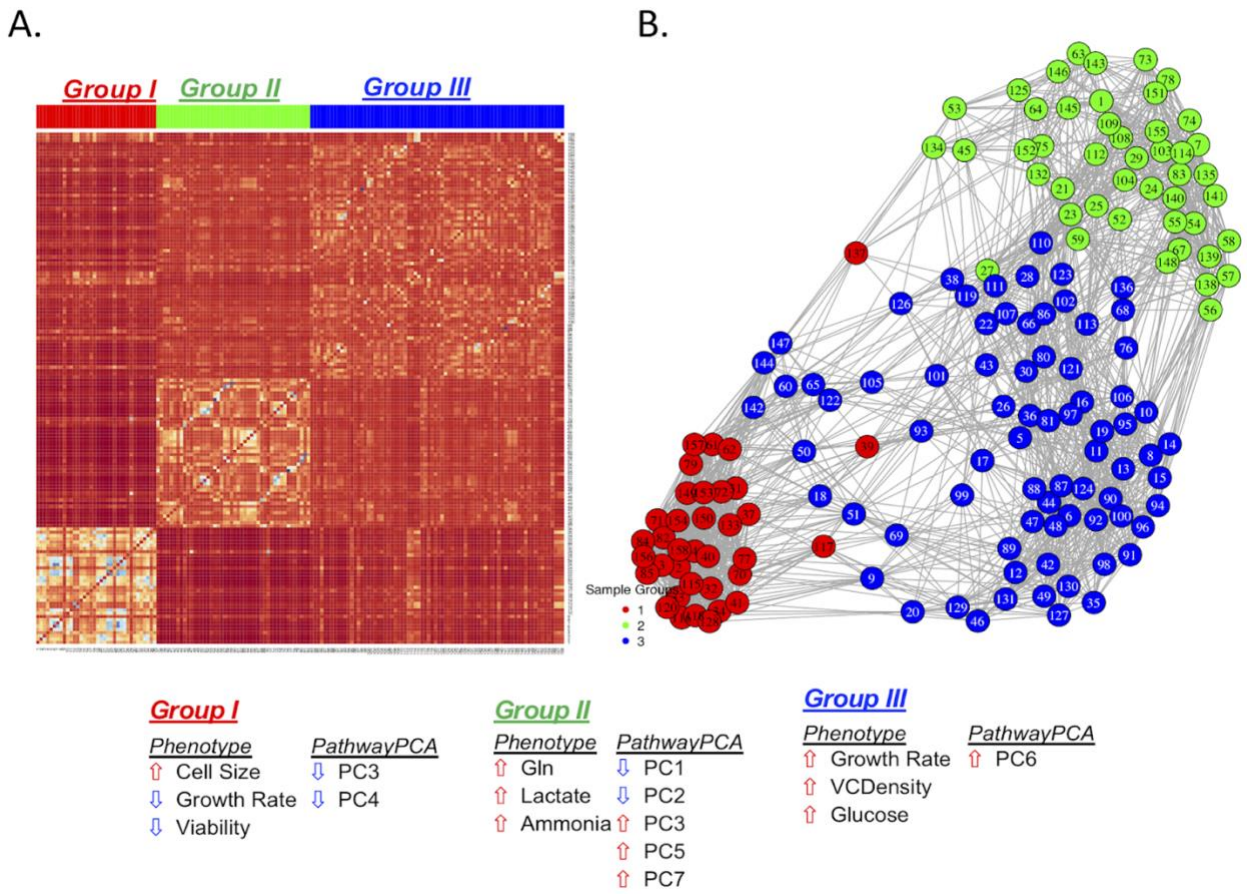


Figure 3.5 Pathways deregulations in different glycoengineered mutants clustered into three groups based on Similarity Network Fusion.
(A-B) SNF-combined similarity matrix and network for integrated phenotypic and gene expression data. Three groups are found, and each has a different cellular response and transcriptional signature.

3.2.5 Metabolic reprogramming underlying cellular response induced by glycoengineering

To determine if the glycoengineered cells exhibit significant metabolic changes contributing to the altered phenotypes, we performed a gene set variation analysis (GSVA) to identify significant alterations in phenotypic traits. Utilizing the GSVA results, we then conducted ordinal regression to identify biological pathways that were increasingly dysregulated with increasing numbers (0-3) of isozymes targeted in mutant cell lines. Below, we summarize the findings across different glycosyltransferase families.

The Mgat-family mutants exhibit several significant changes in many cellular pathways, and here we present the top 3 up- and 3 down-regulated pathways as more and more Mgat genes are knocked out. Notably, there were increased activities in pathways involving activated point mutants of FGFR2, regulation of the pyruvate dehydrogenase (PDH) complex, and regulation of Rheb GTPase activity by AMPK (**Supplementary Figure 3.2B, a-c**). The activation of FGFR2 suggests an enhancement in oncogenic signaling, potentially promoting cell proliferation and survival(Chang et al., 2018). The increased activity of PDH complex regulation likely reflects cellular adaptation to handle elevated pyruvate levels resulting from increased glycolysis(Harris et al., 2002). Furthermore, the upregulation of AMPK activity and subsequent inhibition of Rheb GTPase and mTORC1 signaling appears to be a compensatory response to metabolic stress, possibly through enhanced autophagy to maintain cellular homeostasis (MDPI)(Garcia and Shaw, 2017). These pathway activities, combined with our observation of increased lactate secretion, suggest a metabolic reprogramming toward the Warburg effect, with its high glycolytic activity even in the presence of oxygen(Vander Heiden et al., 2009). This metabolic phenotype is commonly observed in rapidly proliferating cells(Abdel-Haleem et al., 2017), wherein enhanced glycolysis supports both energy production and biosynthetic needs. Conversely, pathways were downregulated associated with p75NTR signaling complexes, acyl-chain remodeling of phosphatidylinositol (PI), and SHC1 events in ERBB4 signaling (**Supplementary Figure 3.2B, d-f**). The downregulation of p75NTR signaling may promote survival and inflammatory responses, potentially reducing apoptosis(Kraemer et al., 2014). Reduced acyl-chain remodeling of PI could affect membrane dynamics and signaling, potentially disrupting lipid metabolism and membrane-associated processes(Bozelli and Epand, 2019). Finally, the decreased activity in SHC1-mediated

ERBB4 signaling suggests a reduction in proliferative and survival signals, which could impact cellular responses to growth factors and overall cell growth(Mehlitz et al., 2010).

The B4galt-family mutants, on the other hand, exhibit enhanced stress responses and reduced critical signaling pathways as more and more B4galt genes are knocked out. We observed an upregulation in the pathway activities involving acyl-chain remodeling of phosphatidylglycerol (PG), the unfolded protein response (UPR), and the calnexin-calreticulin cycle (**Supplementary Figure 3.3B, a-c**). The upregulation of acyl-chain remodeling suggests enhanced lipid metabolism and membrane dynamics, likely as an adaptive response to maintain membrane integrity under stress conditions(Milacic et al., 2024; Patton-Vogt and de Kroon, 2020). The increased activity in the UPR indicates substantial ER stress as cells attempt to cope with the accumulation of misfolded proteins, a critical process for maintaining cellular homeostasis(Bartoszewska and Collawn, 2020; Lee and Ozcan, 2014). Furthermore, the upregulation of the calnexin-calreticulin cycle, part of the ER quality control system, underscores the need for efficient protein folding and processing in response to increased ER stress(Camacho et al., 2003; Kozlov and Gehring, 2020). Contrarily, there was a downregulation in pathways associated with Signal Regulatory Protein (SIRP)-family interactions, regulation of interferon-alpha (IFN-α) signaling, and regulation of KIT signaling (**Supplementary Figure 3.3B, d-f**). The downregulation of SIRP-family interactions indicates a potential impairment in immune cell signaling and cell-cell communication, which could affect phagocytosis inhibition(Sadria and Layton, 2021). Similarly, the reduction in IFN-α signaling pathways suggests compromised antiviral responses and immune regulation, potentially increasing susceptibility to viral infections(Hu et al., 2022; Sun et al., 2022). Additionally, the decreased activity in KIT signaling pathways, essential for cell proliferation and survival, implies potential disruptions in these processes(Huang et al., 2014; Lee et al., 2011).

The dominant pathway changes that correlated with the number of mutated St3gal genes included those involved in metabolism, ion transport, and DNA repair. When more St3gal isozymes were knocked out, the Endogenous Sterols and Reversible Hydration of Carbon Dioxide pathways were upregulated (**Figure 3.6B, a-c**). The Endogenous Sterols pathway involves the biosynthesis and regulation of sterols, which are essential components of cellular membranes and precursors for bioactive molecules such as hormones(Lakshmanan et al., 2019). Similarly, the Reversible Hydration of Carbon Dioxide pathway is crucial also for maintaining acid-base homeostasis through the reversible conversion of carbon dioxide and water to bicarbonate and protons(Milacic et al., 2024). In contrast, several pathways were downregulated, indicating Ligand Gated Ion Channel Transport, Abacavir Transport and Metabolism, Homologous Recombination Repair of Replication Independent Double Strand Breaks, and Double Strand Break Repair (**Figure 3.6B, d-f**). The downregulation of the Ion Channel Transport suggests a potential disruption in cellular communication and signal transduction, which could affect membrane polarization(Hey et al., 2023). Finally, the down-regulation of the double strand breaks pathways points to a compromised genomic stability, increasing the risk of mutations and chromosomal aberrations(Shimizu et al., 2014; Wozny et al., 2020).

To validate our findings, we performed similar analyses using an independent dataset of 363 glycoengineered CHO-S samples. In this dataset, there are 30 geCHO cell lines comprising different knockout/in combinations of 17 glycosyltransferase gene combinations to produce a range of different glycoforms of ten recombinant glycoproteins, including monoclonal antibodies, fusion proteins, cytokines, and enzymes. We performed RNA-Seq to assess the transcriptome of this dataset. We tested if DNA-damage-repair-related pathways exhibited significant changes in expression with increasing KOs of St3gal genes. Despite being an independently generated cell

line panel, implemented in a different CHO cell lineage, we found that double-strand break repair is suppressed. This pathway showed significantly different enrichment scores between 0 knockouts (mean = 0.215) and 3 knockouts (mean = -0.299) (Mann-Whitney U test, $p < 0.001$) (**Supplementary Figure 3.4**).

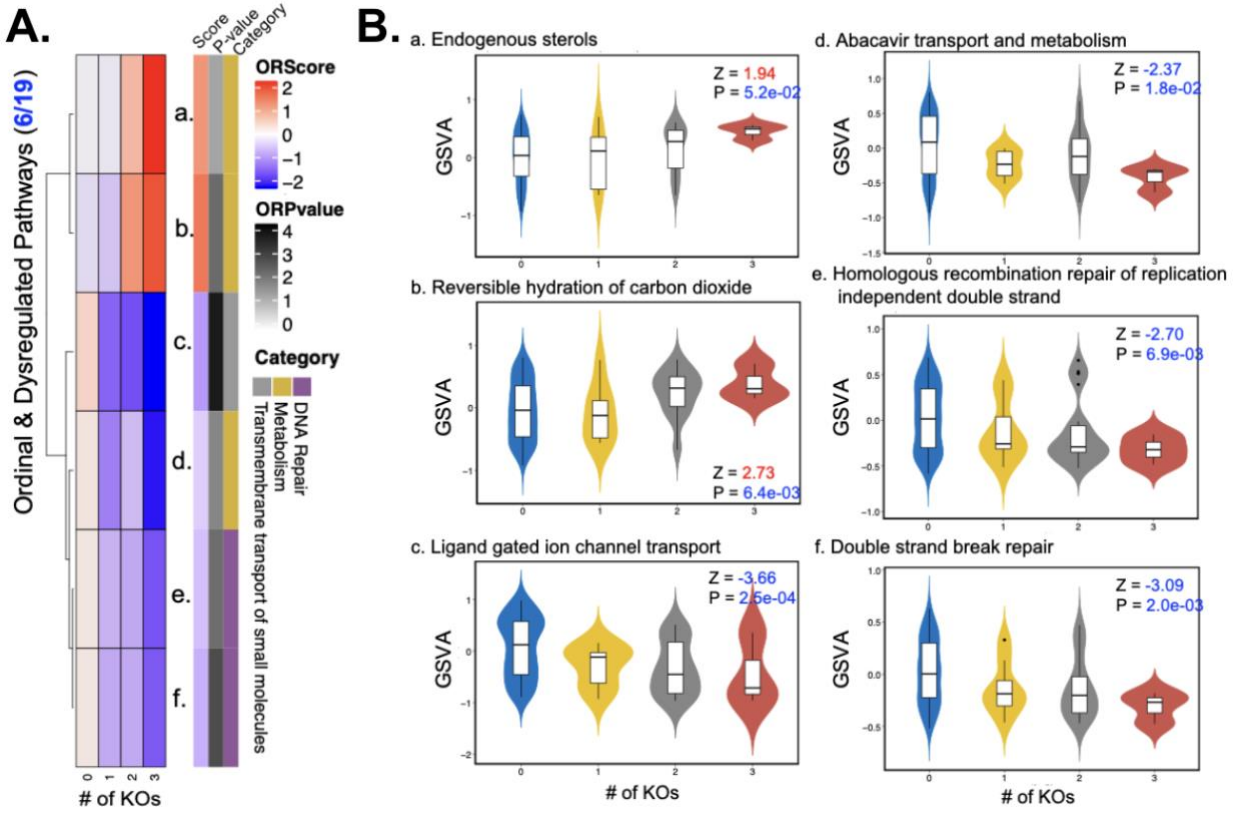


Figure 3.6 Ordinal regression results of GSVA of pathways in the St3gal-family glycoengineered strains.
(A) The heatmap of statistically significant dysregulated pathways with an ordinal trend of GT knockouts. **(B)** Statistically significant dysregulated pathways.

3.3 Discussion

Here, we uncovered fundamental insights into how glycoengineered CHO cells respond to alterations in their glycosylation machinery. First, we demonstrated that glycoengineered cell lines cluster into three distinct groups based on their glycoprofiles (wild-type-like, Moderate, and

Substantial), with these glycomic changes strongly correlating with specific transcriptional signatures. Second, we discovered that different glycosyltransferase families exhibit distinct cellular responses - St3gal modifications increased growth rate and cell density, B4galt affected cell size, and Mgat enhanced cell viability. Third, we revealed previously unknown metabolic reprogramming specific to each glycosyltransferase family, including alterations in energy metabolism, stress responses, and DNA repair mechanisms in St3gal mutants. These findings have important implications for glycoengineering strategies in biotherapeutic production. The observed compensatory transcriptional responses, particularly in non-targeted glycosyltransferases, suggest that cells actively maintain glycan homeostasis through complex regulatory networks. Moreover, the perturbation of critical cellular pathways, such as DNA repair in St3gal mutants, indicates that glycoengineering can have broader impacts on cell physiology than previously appreciated. Understanding these cellular responses is crucial for developing more effective and stable glycoengineered cell lines for biotherapeutic production.

While different glycoengineering approaches produced distinct glycoforms (**Figure 3.2**, left panels), our results further indicate the specific glycans that are either present or absent as a result of glycoengineering (**Figure 3.2**). For instance, bi-antennary and triantennary glycans (GM cluster 1-4; red bars) were significantly presented, whereas tetra-antennary glycans (GM cluster 6-7; blue bars) were significantly absent in the mgat4a and mgat4b double knockout glycoprofile (**Figure 3.2**). The most intriguing observation is that the glycoprofiles of the three major glycosyltransferase knockouts—Mgat (branching), B4galt (galactosylation), and St3gal (sialylation)—show a clear shift in N-glycan structure from tetra-antennary to bi-antennary forms (**Figure 3.2**). This effect could be explained by substrate competition between the knocked-out glycosyltransferase and the enzymes responsible for the biosynthesis of bi-, tri-, and tetra-

antennary structures. A study in CHO cells with human chorionic gonadotropin reported that B4galt4 knockout to 40% of its original expression level resulted in a significant shift in N-glycan structures towards a higher degree of branching, suggesting that B4galt4 plays a substantial role in N-glycan processing(Shauchuk et al., 2020).

Recent reports provide empirical evidence that compensatory mechanisms exist within the complex glycosylation system, demonstrating their roles in maintaining glycan epitope homeostasis. Homeostasis is crucial for robustness in living systems. Several studies have reported examples illustrating how the glycosylation machinery maintains glycan epitope homeostasis(Akama et al., 2006; Krambeck et al., 2017). Specifically, gene mutations in glycosyltransferases within the glycan biosynthesis pathway can result in the global compensation of missing glycan epitopes in other glycans. For example, Mgat4a and Mgat4b enable the synthesis of the GlcNAcβ1-4 branch on the Manα1-3 arm of the N-glycan core. This branch increases glycan epitope density in N-linked glycoproteins, enhancing overall lectin binding to N-linked glycans. Interestingly, Mgat4a/b double mutant mice exhibit significantly decreased N-glycan branches. However, the missing N-linked branch is observed as LacNAc epitopes on the remaining linear arms of the N-linked glycans(Takamatsu et al., 2010). These results demonstrate that the glycosylation machinery can effectively maintain glycan epitope density at homeostatic states. Although the underlying mechanisms by which the glycosylation machinery maintains glycan epitope homeostasis are mainly unknown, elucidating these mechanisms will undoubtedly reveal novel aspects of glycosylation control.

Our observations suggest that glycosylation alterations are influenced by several dominant isoenzymes (**Figure 3.2**). Specifically, the glycoform resulting from Mgat4b knockout exhibits a pronounced reduction in tetra-antennary glycans (GM cluster 6-7) compared to Mgat4A

knockout's glycoform. Moreover, B4galt1 knockout leads to a significant reduction in many hybrid glycans (GM cluster 1, 3-7) than knockouts of other isoenzymes: B4galt2, B4galt3, and B4galt4. Lastly, St3gal4 appears to be the dominant isoenzyme affecting the expression of tri- and tetra-antennary glycans (GM cluster 4-7), as also supported by a previous study(Chung et al., 2015). These findings provide insights into how glycan compensatory mechanisms could be elucidated by the biochemical network.

Based on the observed pathway activity changes in the Mgat-family glycomutants, several conclusions can be drawn that relate to smaller cell size and higher growth rate. The observed significant upregulation of pathways involving activated FGFR2, PDH complex regulation, and AMPK-mediated Rheb GTPase regulation suggests a metabolic shift toward the Warburg effect, which is characterized by high glycolytic activity even in the presence of oxygen. Supporting this metabolic shift, we observed increased lactate secretion as more genes were knocked out in Mgat mutants. This metabolic phenotype is commonly observed in rapidly proliferating cells and can support higher growth rates through enhanced glycolysis for both energy production and biosynthetic needs. Conversely, we observed downregulation in pathways associated with acyl-chain remodeling of phosphatidylinositol (PI) and SHC1 events in ERBB4 signaling. Reduced acyl-chain remodeling of PI could affect membrane composition and cellular signaling, potentially disrupting lipid metabolism and membrane-associated processes that are important for CHO cell growth. The decreased activity in SHC1-mediated ERBB4 signaling suggests a reduction in proliferative and survival signals in these production cells. These pathway alterations align with our Bioprofile FLEX analysis, which showed smaller cell size and higher growth rate in Mgat mutant cell lines, though the precise mechanistic connections between these pathway changes and phenotypes require further investigation.

Our data suggests that B4galt-family mutants decreased in viable cell density (VCD) and growth rate. The significant upregulation in pathways related to enhanced lipid metabolism, adaptive membrane dynamics, and substantial endoplasmic reticulum (ER) stress due to the accumulation of misfolded proteins indicates that these cells may be under increased metabolic strain. This strain can negatively impact cell growth and viability as cells prioritize survival and maintenance functions over proliferation. Conversely, the downregulated pathways indicate impairments in immune cell signaling, compromised antiviral responses, and disruptions in cell proliferation and survival. These disruptions further contribute to reduced cell viability and growth. Collectively, the enhanced stress responses and the downregulation of critical signaling pathways in B4galt-family glycomutants reflect the cellular efforts to adapt and survive under stress conditions while potentially compromising immune and proliferative functions. These findings align with our observed phenotypic changes of decreased VCD and growth rate.

The Reactome pathway enrichment analysis following St3gal knockout revealed significant metabolic and activity pathway alterations (**Figure 3.5**). Notably, the upregulated pathways likely represent compensatory mechanisms to maintain cellular homeostasis under altered glycosylation conditions, which can support increased cell membrane integrity and efficient pH regulation, thereby promoting cell growth and survival. Unfortunately, the downregulated pathways indicate potential impairments in DNA repair mechanisms. The upregulation of metabolic pathways might be an adaptive response to the stress induced by the knockouts, aiming to stabilize cellular functions. However, the concomitant downregulation of key DNA repair pathways suggests compromised genomic integrity, which could lead to reduced repair efficiency, shift in repair pathway choice, and impact on cell cycle(Wozny et al., 2020). We also found these findings align with the increased VCD and cell growth rate we found earlier. The

balance between enhanced metabolic activity and reduced DNA repair efficiency suggests that the cells are prioritizing growth and proliferation at the potential cost of long-term genomic stability. These findings highlight a complex interplay where metabolic adjustments are countered by vulnerabilities in DNA repair and cellular signaling, underscoring the crucial role of St3gal in maintaining cellular equilibrium and its potential impact on cell growth and viability.

A critical challenge in biopharmaceutical development is establishing mammalian cell lines that can efficiently and effectively perform glycoengineering of therapeutic proteins(Lewis et al., 2013; Noh et al., 2013; Reinhart et al., 2019; Tejwani et al., 2018; Thiele et al., 2018). While numerous cell lines have been developed, productivity loss remains a major concern in biomanufacturing. Research has demonstrated that this productivity instability often stems from genome instability(Bandyopadhyay et al., 2019; Hopkins et al., 2022), where chromosomal abnormalities lead to reduced transgene copy numbers and decreased protein titer, as genomic instability disrupts the integrity of integrated transgenes(Spahn et al., 2022). Recent studies have revealed that deficiencies in DNA damage repair pathways can be a root cause of this genome instability(Pradhan et al., 2022; Terabayashi and Hanada, 2018). Given these findings, investigating whether glycoengineered mutants exhibit significant alterations in DNA damage and repair mechanisms becomes crucial, as these changes could either benefit or harm product production and host cell survival. Understanding the underlying mechanisms of these cellular responses could provide valuable insights for rational optimization of expression systems, ultimately advancing the development of cell lines for sustainable glycoengineered therapeutic protein production.

3.4 Methods

3.4.1 Glycoprofile analysis

Glycoprofiles of EPO were already measured and published(Yang et al., 2015). EPO glycoprofiling was performed on 34 different glycoengineered genotypes using MALDI-TOF (matrix-assisted laser desorption/ionization–time of flight) analysis of released permethylated N-glycans. For each sample, approximately 25 μg of purified EPO was digested overnight at 37°C with 2U PNGase F in 50 mM ammonium bicarbonate to release the N-glycans. Released N-glycans were separated using C18 Stagetips and incubated in 100 mM ammonium acetate (pH 5.0) for 60 minutes at 22°C.

The glycan samples were permethylated and extracted with chloroform. The organic phase was washed repeatedly with deionized water for desalting, then evaporated under nitrogen gas. Permethylated N-glycans were dissolved in 20 μl of 50% (v/v) methanol and co-crystallized with 2,5-dihydroxybenzoic acid matrix (10 mg/ml in 70% acetonitrile, 0.1% trifluoroacetic acid, and 2.5 mM sodium acetate). MALDI-TOF analysis was performed in positive reflector mode, acquiring 2,000 shots per spot in the m/z 1,000-7,000 mass range.

The resulting glycoprofile data was reanalyzed using GlyCompare(Bao et al., 2021) to quantify all glycan intermediates (glyco-motifs) for each sample. Hierarchical clustering analysis was performed using the ward.D2 agglomeration method to identify mutants with similar glycan biosynthesis patterns. Prior to clustering, missing values in the glyco-motif abundance matrix were set to zero. The dendrogram was cut to identify three distinct clusters (k=3) representing different levels of glycosylation modification. For visualization, a symmetric color palette was generated using RColorBrewer's RdBu palette, and the clustering results were displayed using annHeatmap2 with both row and column dendrograms. Nine major glyco-motif clusters (GM 01-09) were identified through column clustering, with each cluster annotated based on representative glycan

topology shared by >60% of glyco-motif members within that cluster. The robustness of the clustering was assessed by comparing the branching patterns in the dendrograms.

3.4.2 RNA sample preparation and sequencing

WT Rit, 6xKO Rit, 11xKO Rit cells (5 × 106) sampled from the DASGIP bioreactor on days 4, 6, and 8 were resuspended in RLT buffer (Qiagen, Hilden, Germany) containing 40-mM DTT. After storage at −80 °C, RNA was extracted using the RNeasy kit (Qiagen) followed by on-column DNAse digestion. RNA was eluted in 40 µl nuclease-free water, concentration was measured by Qubit (Thermo Fisher scientific) and the purity was checked on a Fragment Analyzer (Advanced Analytical). Complementary DNA synthesis was obtained from 1 µg of RNA using the High capacity cDNA RT kit (Thermo Fisher scientific). Samples were diluted to 60 ng µL−1 in 50 µL and library preparation was performed with the TruSeq Stranded mRNA Library Prep Kit (Illumina, San Diego, CA, USA). Final RNA libraries were first quantified by Qubit and the size was checked on a Fragment Analyzer. Libraries were normalized to 10 nM and pooled and the final pool of libraries was run on the NextSeq platform with high output flow cell configuration (NextSeq® 500/550 High Output Kit v2, 300 cycles, Illumina). Raw data are deposited at the Gene Expression Omnibus and Short Read Archive with accession number.

3.4.3 RNA-Seq quantification and differential gene expression

RNA-Seq quality was assessed using FastQC. Adapter sequences and low quality bases were trimmed using Trimmomatic(Bolger et al., 2014). Sequence alignment was performed using STAR(Dobin et al., 2013) against the CHO genome (GCF_000419365.1_C_griseus_v1.0) with the default parameters. The expression of each gene was quantified using HTSeq(Anders et al., 2015).

Gene counts were normalized using variance stabilizing transformation. To facilitate cross-species comparison, CHO genes were mapped to their mouse orthologs using a curated orthology

mapping table. Gene symbols were obtained by mapping the mouse Entrez IDs to official gene symbols using the org.Hs.eg.db annotation package. For downstream analysis, we focused on glycosylation-related genes, including glycosyltransferases, glycan-binding proteins, and other glycosylation-related enzymes based on curated annotation lists.

Differential expression analysis was performed using DESeq2(Applied Research Press, 2015) to compare each glycoengineered cell line to the wild-type control. For each comparison, we generated log2 fold changes and corresponding p-values. A gene was considered differentially expressed if it showed an absolute log2 fold change > 1.5 and adjusted p-value < 0.05. The results were compiled into a comprehensive differential expression table containing both the log2 fold changes and p-values for all genes across different glycoengineering conditions. For visualization and interpretation, differentially expressed genes were grouped by glycosyltransferase families (Mgat, B4galt, St3gal, and B3gnt) and their biological functions. Expression changes were visualized using hierarchical clustering with Ward's method (ward.D2).

3.4.4 Cluster comparison analysis

To assess the relationship between glycomotif patterns and gene expression, we performed a quantitative comparison of the hierarchical clustering results using the dendextend R package. The dendrogram structures were first aligned by identifying overlapping samples between datasets using tree intersection methods. The dendrograms were then optimized through a two-step process: an initial random search algorithm with 2,000 iterations, followed by stepwise rotation optimization to minimize dendrogram entanglement.

The similarity between clustering patterns was evaluated using the Fowlkes-Mallows (FM) index, which quantifies the agreement between two hierarchical clusterings. Statistical significance was assessed through permutation testing with 100,000 iterations. For each iteration, cluster labels were randomly permuted while maintaining the dendrogram structure, and the FM

index was recalculated using the FM_index_R function. The empirical p-value was determined by comparing the observed FM index to the distribution of permuted values, with $p < 1e-5$ considered significant. The entanglement metric was also calculated to quantify the amount of crossing between dendrograms, with lower values indicating better agreement between the clustering patterns.

3.4.5 Bioprofile FLEX Analyzer and data processing

The concentrations of glucose, lactate, ammonium ($NH4^+$), and glutamine in spent media were measured using the BioProfile 400 (Nova Biomedical). Also measured key cellular phenotypes including cell size, growth rate, and viability.

Cell culture measurements were processed to calculate key metabolic and growth parameters. The exponential growth rate (r) was calculated using the equation N(t) = N(0)e^(rt), where N(0) and N(t) represent the viable cell density (VCD) at the initial and 48-hour timepoints, respectively. The cell population doubling time was derived as ln(2)/r.

Metabolic uptake and secretion rates were calculated for glutamine, glutamate, glucose, lactate, $NH4^+$, Na^+, K^+, and osmolality. The specific rates were determined using the equation (C2-C1)/(0.5(t2-t1)(VCD1+VCD2)), where C2 and C1 are the metabolite concentrations at 48 hours and 0 hours, t2-t1 is the time interval (48 hours), and VCD1 and VCD2 are the viable cell densities at the respective timepoints.

Fold changes between 0 and 48 hours were calculated for cell size, VCD, and viability as the ratio of final to initial values. For baseline media composition, average values of initial metabolite concentrations were used across all samples to establish consistent reference conditions.

Data was first visualized using hierarchical clustering with Ward's method (ward.D2). Then, for phenotype visualization, measurements (cell size, VCD, viability, and growth rate) were

normalized by subtracting wild-type values. The normalized data was visualized using boxplots with individual data points overlaid using ggplot2. Samples were grouped by glycosyltransferase family (WT, Mgat, B4gal, ST3gal), with other perturbations classified as "Other". For each phenotype, the comparison between glycosyltransferase families was visualized using position-adjusted boxplots.

3.4.6 Similarity Network Fusion Analysis

To integrate multiple-omics data types and identify patterns in cellular responses to glycoengineering, we employed Similarity Network Fusion (SNF)(Wang et al., 2014) to construct and analyze patient-similarity networks. SNF enables integration of phenotypic and gene expression data by constructing a network for each data type and then fusing them into one comprehensive similarity network that captures both shared and complementary information.

We first constructed a sample similarity network represented as a graph G = (V,E), where vertices V correspond to the cell line samples and edges E are weighted by pairwise sample similarities. Edge weights were computed using a scaled exponential similarity kernel. We employed normalized mutual information (NMI) to validate the compatibility of the integrated data sources. All SNF analyses were performed using the SNFtool R package with default parameters unless otherwise specified.

3.4.7 Gene Set Variation Analysis

Prior to regression, pathway activity scores were computed using Gene Set Variation Analysis (GSVA)(Hänzelmann et al., 2013). RNA-Seq count data were normalized via variance-stabilizing transformation and mapped to mouse orthologs. Reactome pathways(Milacic et al., 2024) were selected for their relevance to glycosylation, metabolism, and stress responses. GSVA enrichment scores quantified the relative activity of each pathway across samples, enabling systematic comparison of metabolic and functional shifts associated with glycoengineering.

3.4.8 Ordinal regression analysis

Ordinal regression was performed using cumulative link models (CLMs)(Christensen and Haubo, 2015) to analyze ordered categorical outcomes. The analysis was implemented in R using the ordinal package's clm function, which estimates threshold parameters and predictor effects simultaneously while maintaining the ordered nature of response categories. Pathway activity scores, derived from Gene Set Variation Analysis (GSVA), were modeled as continuous responses, with the number of targeted glycosyltransferase genes (e.g., St3gal, Mgat, or B4galt family) knockouts as an ordinal predictor.

The CLM framework assumes proportional odds, where the effect of increasing knockouts on pathway activity is consistent across thresholds. Coefficients (ORscore) were interpreted as log-odds ratios, where positive values indicated pathways upregulated with increasing knockouts, and negative values indicated downregulation. ORPvalue is the statistical significance of the ORscore, derived from the CLM. Significance thresholds were set at an absolute log2 fold change >1.5 and adjusted p-value <0.05 after Benjamini-Hochberg correction for multiple testing.

Pathways with significant ordinal trends ($p < 0.05$) were visualized as heatmaps (Figure 6a, S2a, S3a). Top 3 up- and down-regulated pathways were visualized as box plots (Figure 6b, S2b, S3b).

3.4.9 CHO-S derived recombinant protein purification and data collection

Herceptin, Rituximab and Enbrel were purified by protein A affinity chromatography. For each protein glycoform, a 1-mL MAbSelect Extra column (Cytiva) was equilibrated with 5 column volumes (CV) of 20 mM sodium phosphate, 0.15 M NaCl, pH 7.2. Next, 30-mL supernatant was loaded, the column was washed with 20 CV of 20 mM sodium phosphate, 0.15 M NaCl, pH 7.2, and the protein was eluted using 0.1 M citrate, pH 3.0. The elution fractions (0.5 mL) were collected in deep-well plates containing 100 µL of 1 M Tris at pH 9 per well.

Alpha-1-antitrypsin, butyrylcholinesterase, protein Z, serpin A5, serpin A10, serpin C1, and EPO, all C-terminally tagged with the HPC4 tag (amino acids EDQVDPRLIDGK), were purified over a 1-mL column of anti-protein C affinity matrix according to the manufacturer's protocol (Roche, cat. no. 11815024001). 1 mM $CaCl_2$ was added to the supernatant, equilibration buffer, and wash buffer, respectively. The proteins were eluted in 0.5 mL fractions using 5 mM EDTA in the elution buffer.

For all proteins, RNA-Seq quality was assessed using FastQC. For details, refer to the RNA-Seq quantification section above. Following RNA-seq processing, Gene Ontology (GO)(Boyle, 2004) term analysis was performed to annotate biological pathways. GSVA was then applied to quantify pathway activity, generating enrichment scores for each sample. Samples were stratified into two groups based on knockout status (0 vs. 3) to compare pathway dysregulation. Differential pathway activity between groups was assessed using the Wilcoxon rank-sum test, and pathways with statistically significant differences ($p < 0.05$) were retained. From these, DNA damage repair-related pathways were selectively identified based on functional relevance. To visualize the most pronounced changes, the top three pathways exhibiting the strongest upregulation and downregulation (ranked by effect size) were plotted as boxplots, illustrating the distribution of GSVA scores across knockout conditions.

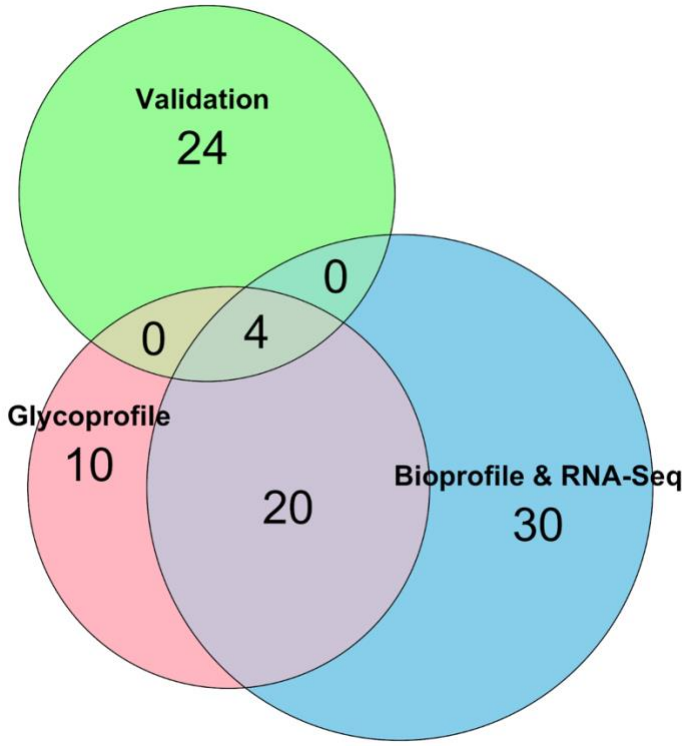
3.4.10 Statistical Analyses

Statistical analyses were performed using R (version 4.1.0), and data visualization was carried out using custom R scripts incorporating the ggplot2 and related packages. Principal Component Analysis (PCA) was performed using the PCA function from FactoMineR package.

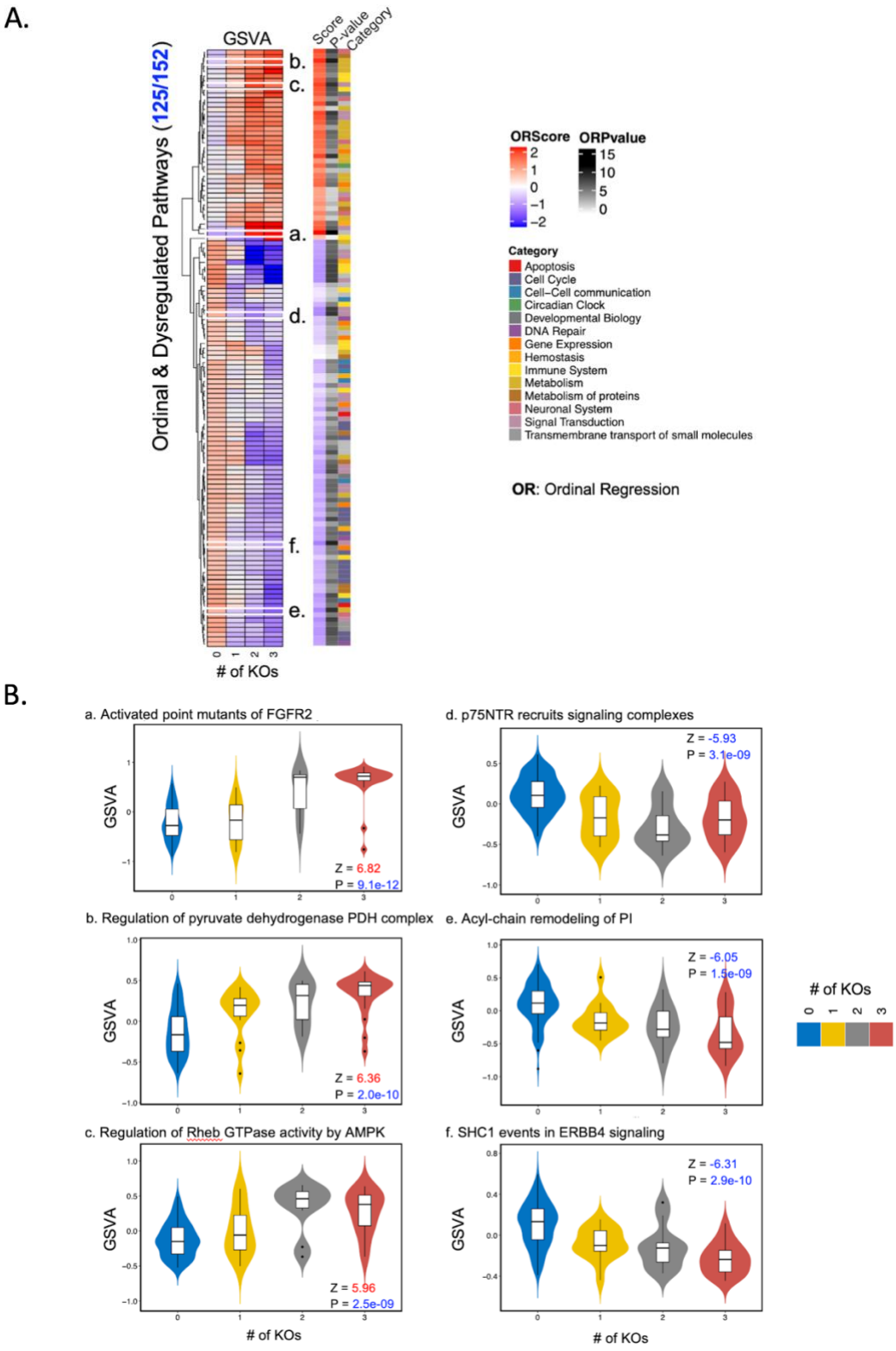
3.5 Conclusion

Here, we demonstrated that glycoengineering not only impacts glycosylation but that it yields phenotypic alterations in CHO cell lines. Furthermore, these molecular mechanisms are linked to the phenotypic alterations as glycosylation changes. Studies have shown that metabolism is associated with several CHO cell phenotypes(Hefzi et al., 2016; Lao and Toth, 1997; Yang and Butler, 2000). Our findings provide insightful knowledge on the effects of glycoengineering and detailed mechanisms about how glycosylation and pathway interplay can lead to phenotype changes following glycoengineering on the CHO cells. Thus, these insights provide further tools to enable the development of a robust host with desired phenotypes when performing glycoengineering on the CHO cells for biotherapeutic products.

3.6 Supplementary Figures and Tables

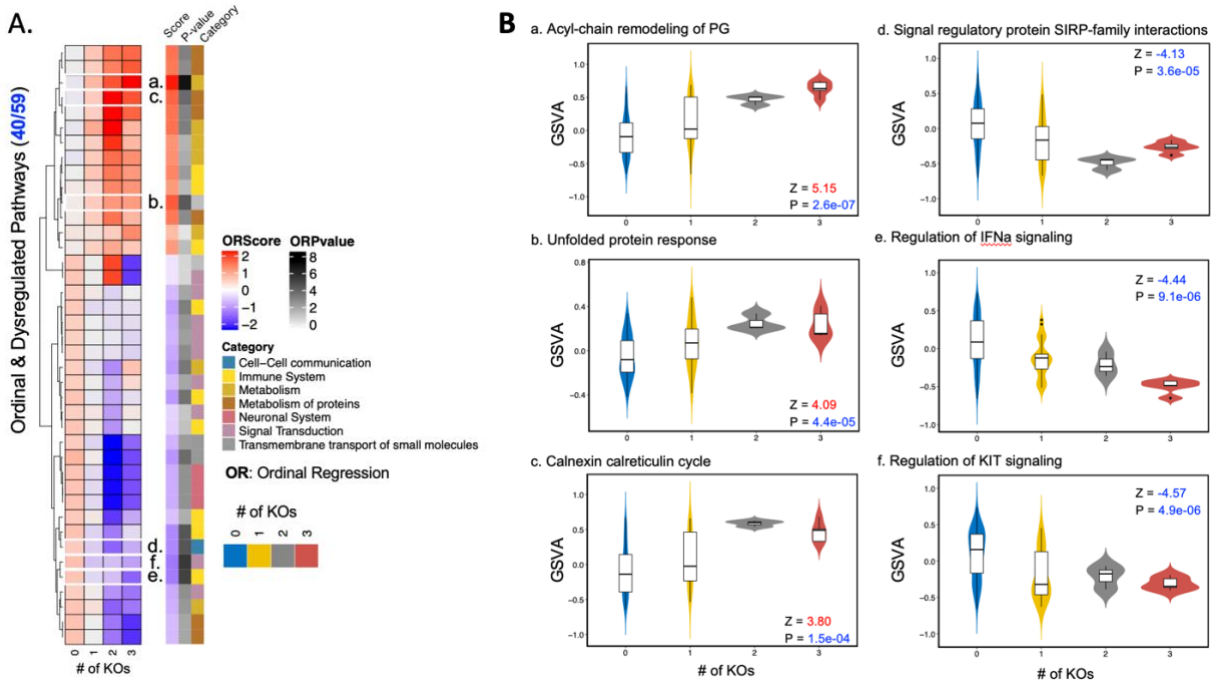


Supplementary Figure 3.1
Venn diagram illustrating overlap and uniqueness among Validation, Glycoprofile, and Bioprofile & RNA-Seq datasets. The Glycoprofile dataset comprises 34 genotypes, of which 24 overlap with the Bioprofile & RNA-Seq dataset. The Bioprofile & RNA-Seq dataset contains 30 unique genotypes not shared with other datasets. The Validation dataset (RNA-Seq) includes 24 unique genotypes.



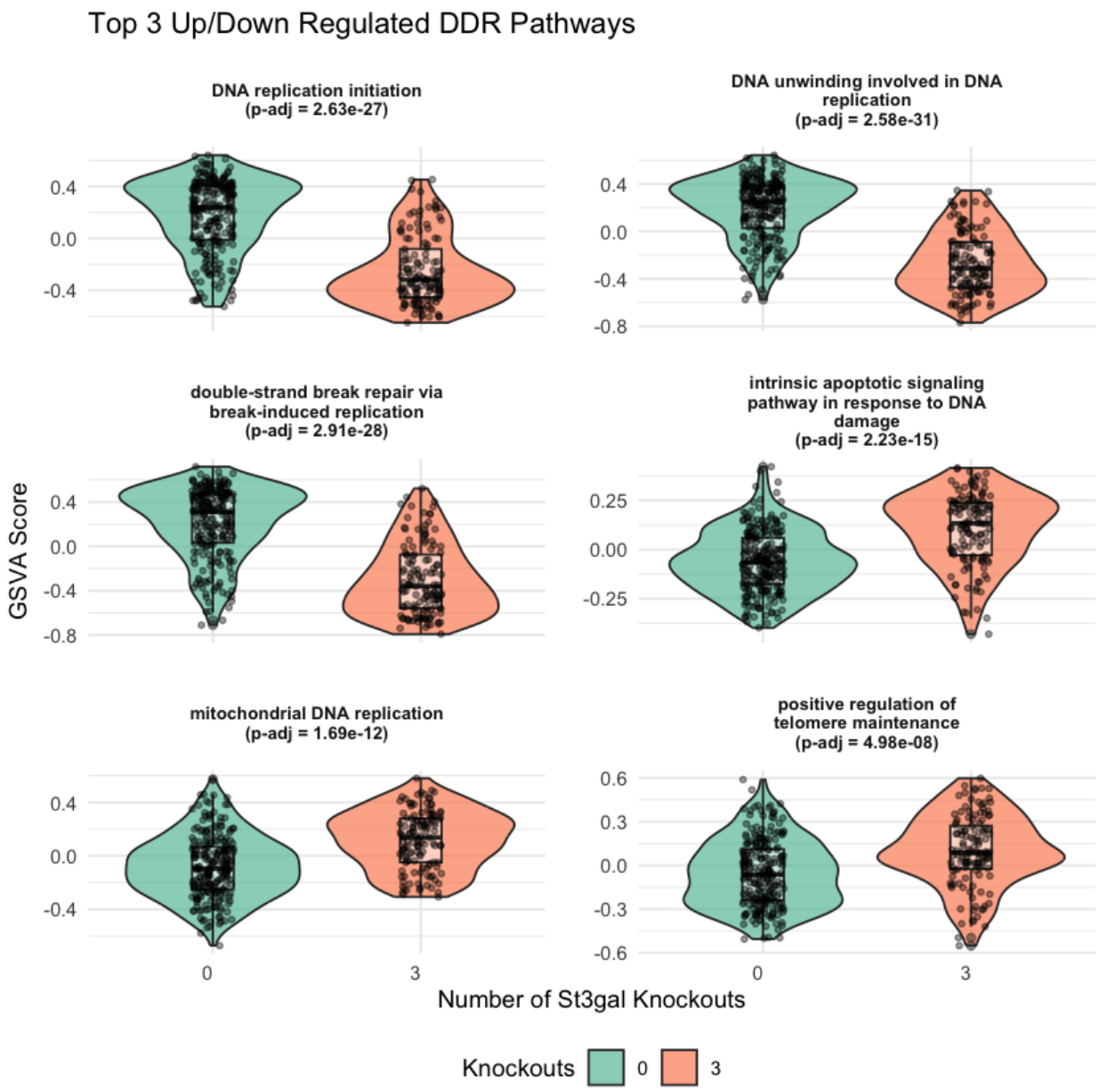
Supplementary Figure 3.2

Ordinal regression results of GSVA of pathways in the Mgat-family glycoengineered strains. (A) The heatmap of statistically significant dysregulated pathways with an ordinal trend of GT knockouts. (B) Statistically significant dysregulated pathways.



Supplementary Figure 3.3

Ordinal regression results of GSVA of pathways in the B4galt-family glycoengineered strains. **(A)** The heatmap of statistically significant dysregulated pathways with an ordinal trend of GT knockouts. **(B)** Statistically significant dysregulated pathways.



Supplementary Figure 3.4
Validation results of GSVA of pathways in the St3gal-family glycoengineered strains. Statistically significant dysregulated DNA-damage-repair related pathways.

3.7 Acknowledgements

Chapter 3, in full, is currently being prepared for submission for publication of the material in a manuscript titled “Delineating the Transcriptional Regulation and Phenotypic Impact from Biotherapeutic Glycoengineering”, Haining Li, Austin W.T. Chiang, Vahid H. Gazestani, Bokan Bao, Shangzhong Li, Patrice Ménard, Johnny Arnsdorf, Zulfiya Sukhova, Sara Petersen Bjørn,

Karen Kathrine Brøndum, Anders Holmgaard Hansen, Bjørn G. Voldborg and Nathan E. Lewis.

The dissertation author was the primary investigator and author of this paper.

REFERENCES

Abdel-Haleem, A.M., Lewis, N.E., Jamshidi, N., Mineta, K., Gao, X., Gojobori, T., 2017. The Emerging Facets of Non-Cancerous Warburg Effect. Front Endocrinol (Lausanne) 8, 279.

Adua, E., Russell, A., Roberts, P., Wang, Y., Song, M., Wang, W., 2017. Innovation Analysis on Postgenomic Biomarkers: Glycomics for Chronic Diseases. OMICS 21, 183–196.

Aizpurua-Olaizola, O., Sastre Toraño, J., Falcon-Perez, J.M., Williams, C., Reichardt, N., Boons, G.-J., 2018. Mass spectrometry for glycan biomarker discovery. Trends Analyt. Chem. 100, 7–14.

Akama, T.O., Nakagawa, H., Wong, N.K., Sutton-Smith, M., Dell, A., Morris, H.R., Nakayama, J., Nishimura, S.-I., Pai, A., Moremen, K.W., Marth, J.D., Fukuda, M.N., 2006. Essential and mutually compensatory roles of α-mannosidase II and α-mannosidase IIx in N-glycan processing in vivo in mice. Proceedings of the National Academy of Sciences 103, 8983–8988.

Akmal, M.A., Rasool, N., Khan, Y.D., 2017. Prediction of N-linked glycosylation sites using position relative features and statistical moments. PLoS One 12, e0181966.

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq-a Python framework to work with high-throughput sequencing data. bioinformatics 31, 166–169.

Antonakoudis, A., Strain, B., Barbosa, R., Jimenez del Val, I., Kontoravdi, C., 2021. Synergising stoichiometric modelling with artificial neural networks to predict antibody glycosylation patterns in Chinese hamster ovary cells. Comput. Chem. Eng. 154, 107471.

Applied Research Press, 2015. Differential expression analysis for sequence count data. Createspace Independent Publishing Platform, North Charleston, SC.

Bandyopadhyay, A.A., O'Brien, S.A., Zhao, L., Fu, H.-Y., Vishwanathan, N., Hu, W.-S., 2019. Recurring genomic structural variation leads to clonal instability and loss of productivity: BANDYOPADHYAY. Biotechnol. Bioeng. 116, 41–53.

Banin, E., 2002. A novel linear code® nomenclature for complex carbohydrates. Trends in Glycoscience and Glycotechnology 14, 127–137.

Banin, E., Neuberger, Y., Altshuler, Y., Halevi, A., Inbar, O., Nir, D., Dukler, A., author_in_Japanese, 2002. A Novel Linear Code® Nomenclature for Complex Carbohydrates. Trends Glycosci. Glycotechnol. 14, 127–137.

Bao, B., Kellman, B.P., Chiang, A.W.T., Zhang, Y., Sorrentino, J.T., York, A.K., Mohammad, M.A., Haymond, M.W., Bode, L., Lewis, N.E., 2021. Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis. Nat. Commun. 12, 4988.

Bartoszewska, S., Collawn, J.F., 2020. Unfolded protein response (UPR) integrated signaling networks determine cell fate during hypoxia. Cell. Mol. Biol. Lett. 25, 18.

Battistel, M.D., Azurmendi, H.F., Yu, B., Freedberg, D.I., 2014. NMR of glycans: shedding new light on old problems. Prog. Nucl. Magn. Reson. Spectrosc. 79, 48–68.

Bhardwaj, A., Kishore, S., Pandey, D.K., 2022. Artificial Intelligence in Biological Sciences. Life 12. https://doi.org/10.3390/life12091430

Bochner, B.R., 2003. New technologies to assess genotype-phenotype relationships. Nat. Rev. Genet. 4, 309–314.

Bojar, D., Camacho, D.M., Collins, J.J., 2020. Using Natural Language Processing to Learn the Grammar of Glycans. bioRxiv. https://doi.org/10.1101/2020.01.10.902114

Bojar, D., Lisacek, F., 2022. Glycoinformatics in the Artificial Intelligence Era. Chem. Rev. 122, 15971–15988.

Bojar, D., Meche, L., Meng, G., Eng, W., Smith, D.F., Cummings, R.D., Mahal, L.K., 2022. A Useful Guide to Lectin Binding: Machine-Learning Directed Annotation of 57 Unique Lectin Specificities. ACS Chem. Biol. 17, 2993–3012.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Bouchard, G., Marques, F.J.G., Karacosta, L.G., Zhang, W., Bermudez, A., Riley, N.M., Mehl, L.C., Benson, J.A., Shrager, J.B., Bertozzi, C.R., Pitteri, S., Giaccia, A.J., Plevritis, S.K., 2021. Multi-Omics Analysis of Fibroblasts from the Invasive Tumor Edge Reveals that Tumor-Stroma Crosstalk Induces O-glycosylation of the CDK4-pRB Axis. bioRxiv. https://doi.org/10.1101/2021.05.28.446229

Boyle, E.I., 2004. GO:: TermFinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20, 3710–3715.

Bozelli, J.C., Jr, Epand, R.M., 2019. Specificity of Acyl Chain Composition of Phosphatidylinositols. Proteomics 19, e1900138.

Burkholz, R., Quackenbush, J., Bojar, D., 2021. Using graph convolutional neural networks to learn a representation for glycans. Cell Rep. 35, 109251.

Camacho, P., John, L., Li, Y., Paredes, R.M., Roderick, H.L., 2003. Calnexin and Calreticulin, ER Associated Modulators of Calcium Transport in the ER, in: Eggleton, P., Michalak, M. (Eds.), Calreticulin: Second Edition. Springer US, Boston, MA, pp. 126–132.

Carpenter, E.J., Seth, S., Yue, N., Greiner, R., Derda, R., 2022. GlyNet: a multi-task neural network for predicting protein-glycan interactions. Chem. Sci. 13, 6669–6686.

Carr, S.A., Huddleston, M.J., Bean, M.F., 1993. Selective identification and differentiation of N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. Protein Sci. 2, 183–196.

Castilho, A., 2014. Proteolytic and N-glycan processing of human α 1-antitrypsin expressed in Nicotiana benthamiana. Plant physiology 166, 1839–1851.

Chang, M.M., Gaidukov, L., Jung, G., Tseng, W.A., Scarcelli, J.J., Cornell, R., Marshall, J.K., Lyles, J.L., Sakorafas, P., Chu, A.-H.A., Cote, K., Tzvetkova, B., Dolatshahi, S., Sumit, M., Mulukutla, B.C., Lauffenburger, D.A., Figueroa, B., Jr, Summers, N.M., Lu, T.K., Weiss, R., 2019. Small-molecule control of antibody N-glycosylation in engineered mammalian cells. Nat. Chem. Biol. 15, 730–736.

Chang, M.-M., Lai, M.-S., Hong, S.-Y., Pan, B.-S., Huang, H., Yang, S.-H., Wu, C.-C., Sun, H.S., Chuang, J.-I., Wang, C.-Y., Huang, B.-M., 2018. FGF9/FGFR2 increase cell proliferation by activating ERK1/2, Rb/E2F1, and cell cycle pathways in mouse Leydig tumor cells. Cancer Sci. 109, 3503–3518.

Chen, S., Qin, R., Mahal, L.K., 2021. Sweet systems: technologies for glycomic analysis and their integration into systems biology. Crit. Rev. Biochem. Mol. Biol. 56, 301–320.

Chiang, A.W.T., Li, S., Spahn, P.N., Richelle, A., Kuo, C.-C., Samoudi, M., Lewis, N.E., 2016. Modulating carbohydrate–protein interactions through glycoengineering of monoclonal antibodies to impact cancer physiology. Curr. Opin. Struct. Biol. 40, 104–111.

Chocholova, E., Bertok, T., Jane, E., Lorencova, L., Holazova, A., Belicka, L., Belicky, S., Mislovicova, D., Vikartovska, A., Imrich, R., Kasak, P., Tkac, J., 2018. Glycomics meets artificial intelligence - Potential of glycan analysis for identification of seropositive and seronegative rheumatoid arthritis patients revealed. Clin. Chim. Acta 481, 49–55.

Christensen, R., Haubo, B., 2015. Analysis of ordinal data with cumulative link models-estimation with the R-package ordinal. R-package version 28.

Chung, C.-Y., Yin, B., Wang, Q., Chuang, K.-Y., Chu, J.H., Betenbaugh, M.J., 2015. Assessment of the coordinated role of ST3GAL3, ST3GAL4 and ST3GAL6 on the α2,3 sialylation linkage of mammalian glycoproteins. Biochem. Biophys. Res. Commun. 463, 211–215.

Coff, L., Chan, J., Ramsland, P.A., Guy, A.J., 2020. Identifying glycan motifs using a novel subtree mining approach. BMC Bioinformatics 21, 42.

Collins, B.C., Gunn, R.J., McKitrick, T.R., Cummings, R.D., Cooper, M.D., Herrin, B.R., Wilson, I.A., 2017. Structural Insights into VLR Fine Specificity for Blood Group Carbohydrates. Structure 25, 1667-1678.e4.

Conroy, L.R., Hawkinson, T.R., Young, L.E.A., Gentry, M.S., Sun, R.C., 2021. Emerging roles of N-linked glycosylation in brain physiology and disorders. Trends Endocrinol. Metab. 32, 980–993.

Cox, M.J., Jaensch, S., Van de Waeter, J., Cougnaud, L., Seynaeve, D., Benalla, S., Koo, S.J., Van Den Wyngaert, I., Neefs, J.-M., Malkov, D., Bittremieux, M., Steemans, M., Peeters, P.J., Wegner, J.K., Ceulemans, H., Gustin, E., Chong, Y.T., Göhlmann, H.W.H., 2020. Tales of 1,008 small molecules: phenomic profiling through live-cell imaging in a panel of reporter cell lines. Sci. Rep. 10, 13262.

Crowley, J.F., Goldstein, I.J., Arnarp, J., Lönngren, J., 1984. Carbohydrate binding studies on the lectin from Datura stramonium seeds. Arch. Biochem. Biophys. 231, 524–533.

Cummings, R.D., Kornfeld, S., 1984. The distribution of repeating [Gal beta 1,4GlcNAc beta 1,3] sequences in asparagine-linked oligosaccharides of the mouse lymphoma cell lines BW5147 and PHAR 2.1. J. Biol. Chem. 259, 6253–6260.

Cummings, R.D., Kornfeld, S., 1982. Characterization of the structural determinants required for the high affinity interaction of asparagine-linked oligosaccharides with immobilized Phaseolus vulgaris leukoagglutinating and erythroagglutinating lectins. J. Biol. Chem. 257, 11230–11234.

Cummings, R.D., Trowbridge, I.S., Kornfeld, S., 1982. A mouse lymphoma cell line resistant to the leukoagglutinating lectin from Phaseolus vulgaris is deficient in UDP-GlcNAc: alpha-D-mannoside beta 1,6 N-acetylglucosaminyltransferase. J. Biol. Chem. 257, 13421–13427.

Dai, B., Mattox, D.E., Bailey-Kellogg, C., 2021. Attention please: modeling global and local context in glycan structure-function relationships. bioRxiv. https://doi.org/10.1101/2021.10.15.464532

Dalziel, M., McFarlane, I., Axford, J.S., 1999. Lectin analysis of human immunoglobulin G N-glycan sialylation. Glycoconj. J. 16, 801–807.

Daponte, V., Hayes, C., Mariethoz, J., Lisacek, F., 2021. Dealing with the Ambiguity of Glycan Substructure Search. Molecules 27. https://doi.org/10.3390/molecules27010065

Dara, S., Dhamercherla, S., Jadav, S.S., Babu, C.M., Ahsan, M.J., 2022. Machine Learning in Drug Discovery: A Review. Artif Intell Rev 55, 1947–1999.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Dörner, T., Kay, J., 2015. Biosimilars in rheumatology: current perspectives and lessons learnt. Nat. Rev. Rheumatol. 11, 713–724.

Drake, P.M., Cho, W., Li, B., Prakobphol, A., Johansen, E., Anderson, N.L., Regnier, F.E., Gibson, B.W., Fisher, S.J., 2010. Sweetening the pot: adding glycosylation to the biomarker discovery equation. Clin. Chem. 56, 223–236.

Dreyfuss, J.M., Jacobs, C., Gindin, Y., Benson, G., Staples, G.O., Zaia, J., 2011. Targeted analysis of glycomics liquid chromatography/mass spectrometry data. Anal. Bioanal. Chem. 399, 727–735.

Duchez, S., Pascal, V., Cogné, N., Jayat-Vignoles, C., Julien, R., Cogné, M., 2011. Glycotranscriptome study reveals an enzymatic switch modulating glycosaminoglycan synthesis during B-cell development and activation. Eur. J. Immunol. 41, 3632–3644.

Elliott, S., Lorenzini, T., Asher, S., Aoki, K., Brankow, D., Buck, L., Busse, L., Chang, D., Fuller, J., Grant, J., Hernday, N., Hokum, M., Hu, S., Knudten, A., Levin, N., Komorowski, R., Martin, F., Navarro, R., Osslund, T., Rogers, G., Rogers, N., Trail, G., Egrie, J., 2003. Enhancement of therapeutic protein in vivo activities through glycoengineering. Nat. Biotechnol. 21, 414–421.

Engvall, E., Perlmann, P., 1971. Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. Immunochemistry 8, 871–874.

Essentials of Glycobiology, 2015. . Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).

Fiocchi, C., 2023. Omics and multi-omics in IBD: No integration, no breakthroughs. Int. J. Mol. Sci. 24. https://doi.org/10.3390/ijms241914912

Fischer, E., Brossmer, R., 1995. Sialic acid-binding lectins: submolecular specificity and interaction with sialoglycoproteins and tumour cells. Glycoconj. J. 12, 707–713.

Gagneux, P., 2003. Human-specific regulation of α2-6-linked sialic acids. Journal of Biological Chemistry 278, 48245–48250.

Gao, C., Hanes, M.S., Byrd-Leotis, L.A., Wei, M., Jia, N., Kardish, R.J., McKitrick, T.R., Steinhauer, D.A., Cummings, R.D., 2019. Unique Binding Specificities of Proteins toward Isomeric Asparagine-Linked Glycans. Cell Chem Biol 26, 535-547.e4.

Garcia, D., Shaw, R.J., 2017. AMPK: Mechanisms of Cellular Energy Sensing and Restoration of Metabolic Balance. Mol. Cell 66, 789–800.

Gebrehiwot, A.G., Melka, D.S., Kassaye, Y.M., Rehan, I.F., Rangappa, S., Hinou, H., Kamiyama, T., Nishimura, S.-I., 2018. Healthy human serum N-glycan profiling reveals the influence of ethnic variation on the identified cancer-relevant glycan biomarkers. PLoS One 13, e0209515.

Geisler, C., Jarvis, D.L., 2011. Letter to the Glyco-Forum: Effective glycoanalysis with Maackia amurensis lectins requires a clear understanding of their binding specificities. Glycobiology 21, 988–993.

Gils, A., Bertolotto, A., Mulleman, D., Bejan-Angoulvant, T., Declerck, P.J., 2017. Biopharmaceuticals: Reference Products and Biosimilars to Treat Inflammatory Diseases. Ther. Drug Monit. 39, 308–315.

Gray, C.J., Migas, L.G., Barran, P.E., Pagel, K., Seeberger, P.H., Eyers, C.E., Boons, G.-J., Pohl, N.L.B., Compagnon, I., Widmalm, G., Flitsch, S.L., 2019. Advancing Solutions to the Carbohydrate Sequencing Challenge. J. Am. Chem. Soc. 141, 14463–14479.

Gupta, A., Kao, K.S., Yamin, R., Oren, D.A., Goldgur, Y., Du, J., Lollar, P., Sundberg, E.J., Ravetch, J.V., 2023. Mechanism of glycoform specificity and in vivo protection by an anti-afucosylated IgG nanobody. Nat. Commun. 14, 2853.

Gutierrez Reyes, C.D., Alejo-Jacuinde, G., Perez Sanchez, B., Chavez Reyes, J., Onigbinde, S., Mogut, D., Hernández-Jasso, I., Calderón-Vallejo, D., Quintanar, J.L., Mechref, Y., 2024. Multi Omics Applications in Biological Systems. Curr. Issues Mol. Biol. 46, 5777–5793.

Haab, B.B., 2012. Using lectins in biomarker research: addressing the limitations of sensitivity and availability. Proteomics Clin. Appl. 6, 346–350.

Halpern, W., Hutto, D., 2013. Biopharmaceuticals, in: Haschek and Rousseaux's Handbook of Toxicologic Pathology. Elsevier, pp. 751–782.

Hamouda, H., Kaup, M., Ullah, M., Berger, M., Sandig, V., Tauber, R., Blanchard, V., 2014. Rapid analysis of cell surface N-glycosylation from living cells using mass spectrometry. J. Proteome Res. 13, 6144–6151.

Hanić, M., Lauc, G., Trbojević-Akmačić, I., 2019. N-Glycan Analysis by Ultra-Performance Liquid Chromatography and Capillary Gel Electrophoresis with Fluorescent Labeling. Curr. Protoc. Protein Sci. 97, e95.

Hänzelmann, S., Castelo, R., Guinney, J., 2013. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 14, 7.

Harris, R.A., Bowker-Kinley, M.M., Huang, B., Wu, P., 2002. Regulation of the activity of the pyruvate dehydrogenase complex. Adv. Enzyme Regul. 42, 249–259.

Hart, G.W., Copeland, R.J., 2010. Glycomics hits the big time. Cell 143, 672–676.

Hefzi, H., Ang, K.S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C.A., Baycin-Hizal, D., Huang, Y., Ley, D., Martinez, V.S., Kyriakopoulos, S., Jiménez, N.E., Zielinski, D.C., Quek, L.-E., Wulff, T., Arnsdorf, J., Li, S., Lee, J.S., Paglia, G., Loira, N., Spahn, P.N., Pedersen, L.E., Gutierrez, J.M., King, Z.A., Lund, A.M., Nagarajan, H., Thomas, A., Abdel-Haleem, A.M., Zanghellini, J., Kildegaard, H.F., Voldborg, B.G., Gerdtzen, Z.P., Betenbaugh, M.J., Palsson, B.O., Andersen, M.R., Nielsen, L.K., Borth, N., Lee, D.-Y., Lewis, N.E., 2016. A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. Cell Syst 3, 434-443.e8.

Hey, G., Rao, R., Carter, A., Reddy, A., Valle, D., Patel, A., Patel, D., Lucke-Wold, B., Pomeranz Krummel, D., Sengupta, S., 2023. Ligand-Gated Ion Channels: Prognostic and Therapeutic Implications for Gliomas. J Pers Med 13. https://doi.org/10.3390/jpm13050853

Hirabayashi, J., 2014. Lectin-based glycomics: how and when was the technology born? Methods Mol. Biol. 1200, 225–242.

Hirabayashi, J., 2004. Lectin-based structural glycomics: glycoproteomics and glycan profiling. Glycoconj. J. 21, 35–40.

Hopkins, J.L., Lan, L., Zou, L., 2022. DNA repair defects in cancer and therapeutic opportunities. Genes Dev. 36, 278–293.

Hořejší, V., Kocourek, J., 1978. Studies on lectins: XXXVII. Isolation and characterization of the lectin from Jimson-weed seeds (Datura stramonium L.). Biochimica et Biophysica Acta (BBA) - Protein Structure 532, 92–97.

Houle, D., Govindaraju, D.R., Omholt, S., 2010. Phenomics: the next challenge. Nat. Rev. Genet. 11, 855–866.

Hu, B., Yu, M., Ma, X., Sun, J., Liu, C., Wang, C., Wu, S., Fu, P., Yang, Z., He, Y., Zhu, Y., Huang, C., Yang, X., Shi, Y., Qiu, S., Sun, H., Zhu, A.X., Zhou, J., Xu, Y., Zhu, D., Fan, J., 2022. IFNα Potentiates Anti–PD-1 Efficacy by Remodeling Glucose Metabolism in the Hepatocellular Carcinoma Microenvironment. Cancer Discov. 12, 1718–1741.

Hu, S., Wong, D.T., 2009. Lectin microarray. Proteomics Clin. Appl. 3, 148–154.

Huang, S., Chaudhary, K., Garmire, L.X., 2017. More is better: Recent progress in multi-omics data integration methods. Front. Genet. 8, 84.

Huang, Z., Ruan, H.-B., Xian, L., Chen, W., Jiang, S., Song, A., Wang, Q., Shi, P., Gu, X., Gao, X., 2014. The stem cell factor/Kit signalling pathway regulates mitochondrial function and energy expenditure. Nat. Commun. 5, 4282.

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95.

Hwang, H., Zhang, Jianpeng, Chung, K.A., Leverenz, J.B., Zabetian, C.P., Peskind, E.R., Jankovic, J., Su, Z., Hancock, A.M., Pan, C., Montine, T.J., Pan, S., Nutt, J., Albin, R., Gearing, M., Beyer, R.P., Shi, M., Zhang, Jing, 2010. Glycoproteomics in neurodegenerative diseases. Mass Spectrom. Rev. 29, 79–125.

Itakura, Y., Nakamura-Tsuruta, S., Kominami, J., Tateno, H., Hirabayashi, J., 2017. Sugar-Binding Profiles of Chitin-Binding Lectins from the Hevein Family: A Comprehensive Study. Int. J. Mol. Sci. 18. https://doi.org/10.3390/ijms18061160

Jansen, B.C., Hafkenscheid, L., Bondt, A., Gardner, R.A., Hendel, J.L., Wuhrer, M., Spencer, D.I.R., 2018. HappyTools: A software for high-throughput HPLC data processing and quantitation. PLoS One 13, e0200280.

Kailemia, M.J., Park, D., Lebrilla, C.B., 2017. Glycans and glycoproteins as specific biomarkers for cancer. Anal. Bioanal. Chem. 409, 395–410.

Kaneda, Y., Whittier, R.F., Yamanaka, H., Carredano, E., Gotoh, M., Sota, H., Hasegawa, Y., Shinohara, Y., 2002. The High Specificities of Phaseolus vulgaris Erythro- and Leukoagglutinating Lectins for Bisecting GlcNAc or β1–6-Linked Branch Structures, Respectively, Are Attributable to Loop B*. J. Biol. Chem. 277, 16928–16935.

Kearney, C.J., Vervoort, S.J., Ramsbottom, K.M., Todorovski, I., Lelliott, E.J., Zethoven, M., Pijpers, L., Martin, B.P., Semple, T., Martelotto, L., Trapani, J.A., Parish, I.A., Scott, N.E., Oliaro, J., Johnstone, R.W., 2021. SUGAR-seq enables simultaneous detection of glycans, epitopes, and the transcriptome in single cells. Sci Adv 7. https://doi.org/10.1126/sciadv.abe3610

Kellman, B.P., Sandoval, D., Zaytseva, O.O., Brock, K., Baboo, S., Nachmanson, D., Irvine, E.B., Armingol, E., Mih, N., Zhang, Y., Jeffris, M., Bartels, P., Nguyen, T., Tam, A., Gasman, S., Ilan, S., Shamie, I., Diedrich, J.K., Wang, X., van Woudenbergh, E., Altman, M., Aylward, A., Bao, B., Castro, A., Sorrentino, J., Chiang, A., Campbell, M., Bartsch, Y., Aguilar-Calvo, P., Sigurdson, C., Alter, G., Lauc, G., Yates, J.R., III, Marks, D., Lisacek, F., Lewis, N.E., 2024. Protein structure, a genetic encoding for glycosylation. bioRxiv. https://doi.org/10.1101/2024.05.15.594261

Kim, A., Kim, Jeongeun, Park, C.S., Jin, M., Kang, M., Moon, C., Kim, M., Kim, Jieun, Yang, S., Jang, L., Jang, J.Y., Kim, H.H., 2023. Peptide-N-glycosidase F or A treatment and procainamide-labeling for identification and quantification of N-glycans in two types of mammalian glycoproteins using UPLC and LC-MS/MS. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 1214, 123538.

Klamer, Z., Hsueh, P., Ayala-Talavera, D., Haab, B., 2019. Deciphering protein glycosylation by computational integration of on-chip profiling, glycan-array data, and mass spectrometry. Mol. Cell. Proteomics 18, 28–40.

Klamer, Z.L., Harris, C.M., Beirne, J.M., Kelly, J.E., Zhang, J., Haab, B.B., 2022. CarboGrove: a resource of glycan-binding specificities through analyzed glycan-array datasets from all platforms. Glycobiology 32, 679–690.

Klein, J., Carvalho, L., Zaia, J., 2018. Application of network smoothing to glycan LC-MS profiling. Bioinformatics 34, 3511–3518.

Kletter, D., Cao, Z., Bern, M., Haab, B., 2013. Determining lectin specificity from glycan array data using motif segregation and GlycoSearch software. Curr. Protoc. Chem. Biol. 5, 157–169.

Knibbs, R.N., Goldstein, I.J., Ratcliffe, R.M., Shibuya, N., 1991. Characterization of the carbohydrate binding specificity of the leukoagglutinating lectin from Maackia amurensis. Comparison with other sialic acid-specific lectins. J. Biol. Chem. 266, 83–88.

Kozlov, G., Gehring, K., 2020. Calnexin cycle - structural features of the ER chaperone system. FEBS J. 287, 4322–4340.

Kraemer, B.R., Yoon, S.O., Carter, B.D., 2014. The biological functions and signaling mechanisms of the p75 neurotrophin receptor. Handb. Exp. Pharmacol. 220, 121–164.

Krambeck, F.J., Bennun, S.V., Andersen, M.R., Betenbaugh, M.J., 2017. Model-based analysis of N-glycosylation in Chinese hamster ovary cells. PLoS One 12, e0175376.

Kukuruzinska, M.A., Lennon, K., 1998. Protein N-glycosylation: molecular genetics and functional significance. Crit. Rev. Oral Biol. Med. 9, 415–448.

Kunej, T., 2019. Rise of Systems Glycobiology and Personalized Glycomedicine: Why and How to Integrate Glycomics with Multiomics Science? OMICS 23, 615–622.

Kuzmanov, U., Kosanam, H., Diamandis, E.P., 2013. The sweet and sour of serological glycoprotein tumor biomarker quantification. BMC Med. 11, 31.

Kwak, C.-Y., Park, S.-Y., Lee, C.-G., Okino, N., Ito, M., Kim, J.H., 2017. Enhancing the sialylation of recombinant EPO produced in CHO cells via the inhibition of glycosphingolipid biosynthesis. Sci. Rep. 7, 13059.

Lageveen-Kammeijer, G.S.M., Kuster, B., 2022. High sensitivity glycomics in biomedicine. Mass Spectrom.

Lairson, L.L., Henrissat, B., Davies, G.J., Withers, S.G., 2008. Glycosyltransferases: structures, functions, and mechanisms. Annu. Rev. Biochem. 77, 521–555.

Lakshmanan, M., Kok, Y.J., Lee, A.P., Kyriakopoulos, S., Lim, H.L., Teo, G., Poh, S.L., Tang, W.Q., Hong, J., Tan, A.H.-M., Bi, X., Ho, Y.S., Zhang, P., Ng, S.K., Lee, D.-Y., 2019. Multi-omics profiling of CHO parental hosts reveals cell line-specific variations in bioprocessing traits. Biotechnol. Bioeng. 116, 2117–2129.

Lao, M.S., Toth, D., 1997. Effects of ammonium and lactate on growth and metabolism of a recombinant Chinese hamster ovary cell culture. Biotechnol. Prog. 13, 688–691.

Lee, A., Nakano, M., Hincapie, M., Kolarich, D., Baker, M.S., Hancock, W.S., Packer, N.H., 2010. The lectin riddle: glycoproteins fractionated from complex mixtures have similar glycomic profiles. OMICS 14, 487–499.

Lee, J., Ozcan, U., 2014. Unfolded protein response signaling and metabolic diseases. J. Biol. Chem. 289, 1203–1211.

Lee, Y.-G., Oh, J.-Y., Kim, D., Kim, G., 2023. SHAP Value-Based Feature Importance Analysis for Short-Term Load Forecasting. Journal of Electrical Engineering & Technology 18, 579–588.

Lee, Y.-N., Brandal, S., Noel, P., Wentzel, E., Mendell, J.T., McDevitt, M.A., Kapur, R., Carter, M., Metcalfe, D.D., Takemoto, C.M., 2011. KIT signaling regulates MITF expression through miRNAs in normal and malignant mast cell proliferation. Blood 117, 3629–3640.

Lehmann, F., Tiralongo, E., Tiralongo, J., 2006. Sialic acid-specific lectins: occurrence, specificity and function. Cell. Mol. Life Sci. 63, 1331–1354.

Lewis, N.E., Liu, X., Li, Y., Nagarajan, H., Yerganian, G., O'Brien, E., Bordbar, A., Roth, A.M., Rosenbloom, J., Bian, C., Xie, M., Chen, W., Li, N., Baycin-Hizal, D., Latif, H., Forster, J., Betenbaugh, M.J., Famili, I., Xu, X., Wang, J., Palsson, B.O., 2013. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. Nat. Biotechnol. 31, 759–765.

Li, H., Chiang, A.W.T., Lewis, N.E., 2022. Artificial intelligence in the analysis of glycosylation data. Biotechnol. Adv. 60, 108008.

Li, Y., Wang, H., Chen, Y., Ding, L., Ju, H., 2024. In Situ Glycan Analysis and Editing in Living Systems. JACS Au 4, 384–401.

Ling, M.M., Ricks, C., Lea, P., 2007. Multiplexing molecular diagnostics and immunoassays using emerging microarray technologies. Expert Rev. Mol. Diagn. 7, 87–98.

Lopez Aguilar, A., Briard, J.G., Yang, L., Ovryn, B., Macauley, M.S., Wu, P., 2017. Tools for studying glycans: Recent advances in chemoenzymatic glycan labeling. ACS Chem. Biol. 12, 611–621.

Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. arXiv [cs.AI].

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell 2, 56–67.

Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.-W., Newman, S.-F., Kim, J., Lee, S.-I., 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2, 749–760.

Lundstrøm, J., Korhonen, E., Lisacek, F., Bojar, D., 2022. LectinOracle: A generalizable deep learning model for lectin-glycan binding prediction. Adv. Sci. 9, e2103807.

Ma, B., Guan, X., Li, Y., Shang, S., Li, J., Tan, Z., 2020. Protein Glycoengineering: An Approach for Improving Protein Properties. Front Chem 8, 622.

Makrydaki, E., Donini, R., Krueger, A., Royle, K., Moya Ramirez, I., Kuntz, D.A., Rose, D.R., Haslam, S.M., Polizzi, K.M., Kontoravdi, C., 2024. Immobilized enzyme cascade for targeted glycosylation. Nat. Chem. Biol. 20, 732–741.

Martorell-Marugán, J., Tabik, S., Benhammou, Y., del Val, C., Zwir, I., Herrera, F., Carmona-Sáez, P., n.d. Deep Learning in Omics Data Analysis and Precision Medicine, in: Husi, H. (Ed.), Computational Biology. Codon Publications, Brisbane (AU).

Mechref, Y., Hu, Y., Desantos-Garcia, J.L., Hussein, A., Tang, H., 2013. Quantitative glycomics strategies. Mol. Cell. Proteomics 12, 874–884.

Medzihradszky, K.F., Kaasik, K., Chalkley, R.J., 2015. Tissue-Specific Glycosylation at the Glycopeptide Level. Mol. Cell. Proteomics 14, 2103–2110.

Mehlitz, A., Banhart, S., Mäurer, A.P., Kaushansky, A., Gordus, A.G., Zielecki, J., Macbeath, G., Meyer, T.F., 2010. Tarp regulates early Chlamydia-induced host cell survival through interactions with the human adaptor protein SHC1. J. Cell Biol. 190, 143–157.

Mehta, A.Y., Cummings, R.D., 2020. GlycoGlyph: a glycan visualizing, drawing and naming application. Bioinformatics 36, 3613–3614.

Meuris, L., Santens, F., Elson, G., Festjens, N., Boone, M., Dos Santos, A., Devos, S., Rousseau, F., Plets, E., Houthuys, E., Malinge, P., Magistrelli, G., Cons, L., Chatel, L., Devreese, B., Callewaert, N., 2014. GlycoDelete engineering of mammalian cells simplifies N-glycosylation of recombinant proteins. Nat. Biotechnol. 32, 485–489.

Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., Petryszak, R., Ragueneau, E., Rothfels, K., Sevilla, C., Shamovsky, V., Stephan, R., Tiwari, K., Varusai, T., Weiser, J., Wright, A., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P., 2024. The Reactome Pathway Knowledgebase 2024. Nucleic Acids Res. 52, D672–D678.

Minoshima, F., Ozaki, H., Odaka, H., Tateno, H., 2021. Integrated analysis of glycan and RNA in single cells. iScience 24, 102882.

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N.C., Ping, P., 2019. Machine learning and integrative analysis of biomedical big data. Genes (Basel) 10, 87.

Mohapatra, S., An, J., Gómez-Bombarelli, R., 2022. Chemistry-informed macromolecule graph representation for similarity computation, unsupervised and supervised learning. Mach. Learn.: Sci. Technol. 3, 015028.

Molnarova, K., Cokrtova, K., Tomnikova, A., Krizek, T., Kozlik, P., 2022. Liquid chromatography and capillary electrophoresis in glycomic and glycoproteomic analysis. Monatsh. Chem. 153, 659–686.

Moon, S.-Y., Chatterjee, S., Seeberger, P.H., Gilmore, K., 2020. Predicting glycosylation stereoselectivity using machine learning. ChemRxiv. https://doi.org/10.26434/chemrxiv.12357302.v1

Moremen, K.W., Tiemeyer, M., Nairn, A.V., 2012. Vertebrate protein glycosylation: diversity, synthesis and function. Nat. Rev. Mol. Cell Biol. 13, 448–462.

Nabi, B., Rehman, S., Khan, S., Baboota, S., Ali, J., 2018. Ligand conjugation: An emerging platform for enhanced brain drug delivery. Brain Res. Bull. 142, 384–393.

Nagae, M., Soga, K., Morita-Matsumoto, K., Hanashima, S., Ikeda, A., Yamamoto, K., Yamaguchi, Y., 2014. Phytohemagglutinin from Phaseolus vulgaris (PHA-E) displays a novel glycan recognition mode using a common legume lectin fold. Glycobiology 24, 368–378.

Nairn, A., Moremen, K., 2010. Chapter 5 - Glycotranscriptomics, in: Cummings, R.D., Pierce, J.M. (Eds.), Handbook of Glycomics. Academic Press, San Diego, pp. 95–135.

Nairn, A.V., York, W.S., Harris, K., Hall, E.M., Pierce, J.M., Moremen, K.W., 2008a. Regulation of glycan structures in animal tissues. J. Biol. Chem. 283, 17298–17313.

Nairn, A.V., York, W.S., Harris, K., Hall, E.M., Pierce, J.M., Moremen, K.W., 2008b. Regulation of glycan structures in animal tissues: transcript profiling of glycan-related genes. J. Biol. Chem. 283, 17298–17313.

Nilsson, C.L., 2003. Lectins: proteins that interpret the sugar code. Anal. Chem. 75, 348A-353A.

Noh, S.M., Sathyamurthy, M., Lee, G.M., 2013. Development of recombinant Chinese hamster ovary cell lines for therapeutic protein production. Curr. Opin. Chem. Eng. 2, 391–397.

Odaka, H., Ozaki, H., Tateno, H., 2022. scGR-seq: Integrated analysis of glycan and RNA in single cells. STAR Protoc 3, 101179.

O'Flaherty, R., Bergin, A., Flampouri, E., Mota, L.M., Obaidi, I., Quigley, A., Xie, Y., Butler, M., 2020. Mammalian cell culture for production of recombinant proteins: A review of the critical steps in their biomanufacturing. Biotechnol. Adv. 43, 107552.

Ohtsubo, K., Marth, J.D., 2006. Glycosylation in cellular mechanisms of health and disease. Cell 126, 855–867.

Oinam, L., Tateno, H., 2022. Glycan Profiling by Sequencing to Uncover Multicellular Communication: Launching Glycobiology in Single Cells and Microbiomes. Front Cell Dev Biol 10, 919168.

Pattathil, S., Avci, U., Zhang, T., Cardenas, C.L., Hahn, M.G., 2015. Immunological Approaches to Biomass Characterization and Utilization. Front Bioeng Biotechnol 3, 173.

Patton-Vogt, J., de Kroon, A.I.P.M., 2020. Phospholipid turnover and acyl chain remodeling in the yeast ER. Biochim. Biophys. Acta Mol. Cell Biol. Lipids 1865, 158462.

Pilobello, K.T., Slawek, D.E., Mahal, L.K., 2007. A ratiometric lectin microarray approach to analysis of the dynamic mammalian glycome. Proc. Natl. Acad. Sci. U. S. A. 104, 11534–11539.

Pradhan, S., Gao, R., Bush, K., Zhang, N., Wairkar, Y.P., Sarkar, P.S., 2022. Polyglutamine expansion in huntingtin and mechanism of DNA damage repair defects in Huntington's disease. Front. Cell. Neurosci. 16, 837576.

Pučić-Baković, M., 2017. High-Throughput Analysis of the IgG N-Glycome by UPLC-FLR. Methods Mol. Biol. 1503, 21–29.

Raju, T.S., Briggs, J.B., Borge, S.M., Jones, A.J., 2000. Species-specific variation in glycosylation of IgG: evidence for the species-specific sialylation and branch-specific galactosylation and importance for engineering recombinant glycoprotein therapeutics. Glycobiology 10, 477–486.

Raman, R., Raguram, S., Venkataraman, G., Paulson, J.C., Sasisekharan, R., 2005. Glycomics: an integrated systems approach to structure-function relationships of glycans. Nat. Methods 2, 817–824.

Ranzinger, R., Herget, S., von der Lieth, C.-W., Frank, M., 2011. GlycomeDB--a unified database for carbohydrate structures. Nucleic Acids Res. 39, D373-6.

Reily, C., Stewart, T.J., Renfrow, M.B., Novak, J., 2019. Glycosylation in health and disease. Nat. Rev. Nephrol. 15, 346–366.

Reinhart, D., Damjanovic, L., Kaisermayer, C., Sommeregger, W., Gili, A., Gasselhuber, B., Castan, A., Mayrhofer, P., Grünwald-Gruber, C., Kunert, R., 2019. Bioprocessing of Recombinant CHO-K1, CHO-DG44, and CHO-S: CHO Expression Hosts Favor Either mAb Production or Biomass Synthesis. Biotechnol. J. 14, 1700686.

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., van Hoesel, C., Schopmans, H., Sommer, T., Friederich, P., 2022. Graph neural networks for materials science and chemistry. Commun Mater 3, 93.

Rocamora, F., Peralta, A.G., Shin, S., Sorrentino, J., Wu, M.Y.M., Toth, E.A., Fuerst, T.R., Lewis, N.E., 2023. Glycosylation shapes the efficacy and safety of diverse protein, gene and cell therapies. Biotechnol. Adv. 67, 108206.

Rocamora, F., Schoffelen, S., Arnsdorf, J., Toth, E.A., Abdul, Y., Cleveland, T.E., 4th, Bjørn, S.P., Wu, M.Y.M., McElvaney, N.G., Voldborg, B.G.R., Fuerst, T.R., Lewis, N.E., 2024. Glycoengineered recombinant alpha1-antitrypsin results in comparable in vitro and in vivo activities to human plasma-derived protein. bioRxivorg. https://doi.org/10.1101/2024.03.27.587088

Romann, P., Vuillemin, T., Pavone, S., Jordan, M., Perilleux, A., Souquet, J., Bielser, J.-M., Herwig, C., Villiger, T.K., 2024. Maduramycin, a novel glycosylation modulator for mammalian fed-batch and steady-state perfusion processes. J. Biotechnol. 383, 73–85.

Rosenfeld, R., Bangio, H., Gerwig, G.J., Rosenberg, R., Aloni, R., Cohen, Y., Amor, Y., Plaschkes, I., Kamerling, J.P., Maya, R.B.-Y., 2007. A lectin array-based methodology for the analysis of protein glycosylation. J. Biochem. Biophys. Methods 70, 415–426.

Rüdiger, H., Gabius, H.J., 2001. Plant lectins: occurrence, biochemistry, functions and applications. Glycoconj. J. 18, 589–613.

Ruhaak, L.R., Xu, G., Li, Q., Goonatilleke, E., Lebrilla, C.B., 2018. Mass Spectrometry Approaches to Glycomic and Glycoproteomic Analyses. Chem. Rev. 118, 7886–7930.

Ryva, B., Zhang, K., Asthana, A., Wong, D., Vicioso, Y., Parameswaran, R., 2019. Wheat Germ Agglutinin as a Potential Therapeutic Agent for Leukemia. Front. Oncol. 9, 100.

Sadria, M., Layton, A.T., 2021. Interactions among mTORC, AMPK and SIRT: a computational model for cell energy balance and metabolism. Cell Commun. Signal. 19, 57.

Schjoldager, K.T., Narimatsu, Y., Joshi, H.J., Clausen, H., 2020. Global view of human protein glycosylation pathways and functions. Nat. Rev. Mol. Cell Biol. 21, 729–749.

Senger, R.S., Karim, M.N., 2008. Prediction of N-linked glycan branching patterns using artificial neural networks. Math. Biosci. 211, 89–104.

Sha, S., Agarabi, C., Brorson, K., Lee, D.-Y., Yoon, S., 2016. N-Glycosylation Design and Control of Therapeutic Monoclonal Antibodies. Trends Biotechnol. 34, 835–846.

Shade, K.-T.C., Conroy, M.E., Washburn, N., Kitaoka, M., Huynh, D.J., Laprise, E., Patil, S.U., Shreffler, W.G., Anthony, R.M., 2020. Sialylation of immunoglobulin E is a determinant of allergic pathogenicity. Nature 582, 265–270.

Shade, K.-T.C., Platzer, B., Washburn, N., Mani, V., Bartsch, Y.C., Conroy, M., Pagan, J.D., Bosques, C., Mempel, T.R., Fiebiger, E., Anthony, R.M., 2015. A single glycan on IgE is indispensable for initiation of anaphylaxis. J. Exp. Med. 212, 457–467.

Shah, H.A., Liu, J., Yang, Z., Feng, J., 2021. Review of Machine Learning Methods for the Prediction and Reconstruction of Metabolic Pathways. Front Mol Biosci 8, 634141.

Shang, Y., Zeng, Yun, Zeng, Yong, 2016. Integrated Microfluidic Lectin Barcode Platform for High-Performance Focused Glycomic Profiling. Sci. Rep. 6, 20297.

Shauchuk, A., Szulc, B., Maszczak-Seneczko, D., Wiertelak, W., Skurska, E., Olczak, M., 2020. N-glycosylation of the human β1,4-galactosyltransferase 4 is crucial for its activity and Golgi localization. Glycoconj. J. 37, 577–588.

Shibata, M., Okamura, K., Yura, K., Umezawa, A., 2020. High-precision multiclass cell classification by supervised machine learning on lectin microarray data. Regen Ther 15, 195–201.

Shibuya, N., Goldstein, I.J., Broekaert, W.F., Nsimba-Lubaki, M., Peeters, B., Peumans, W.J., 1987. The elderberry (Sambucus nigra L.) bark lectin recognizes the Neu5Ac(alpha 2-6)Gal/GalNAc sequence. J. Biol. Chem. 262, 1596–1601.

Shimizu, I., Yoshida, Y., Suda, M., Minamino, T., 2014. DNA damage response and metabolic disease. Cell Metab. 20, 967–977.

Silva, M.L.S., 2019. Lectin biosensors in cancer glycan biomarker detection. Adv. Clin. Chem. 93, 1–61.

Smith, D.F., Cummings, R.D., 2013. Application of microarrays for deciphering the structure and function of the human glycome. Mol. Cell. Proteomics 12, 902–912.

Solá, R.J., Griebenow, K., 2010. Glycosylation of therapeutic proteins: an effective strategy to optimize efficacy. BioDrugs 24, 9–21.

Spahn, P.N., Hansen, A.H., Hansen, H.G., Arnsdorf, J., Kildegaard, H.F., Lewis, N.E., 2016. A Markov chain model for N-linked protein glycosylation--towards a low-parameter tool for model-driven glycoengineering. Metab. Eng. 33, 52–66.

Spahn, P.N., Zhang, X., Hu, Q., Lu, H., Hamaker, N.K., Hefzi, H., Li, S., Kuo, C.-C., Huang, Y., Lee, J.C., Davis, A.J., Ly, P., Lee, K.H., Lewis, N.E., 2022. Restoration of DNA repair mitigates genome instability and increases productivity of Chinese hamster ovary cells. Biotechnol. Bioeng. 119, 963–982.

Specifications: Test procedures and acceptance criteria for biotechnological/biological products, 2016. , in: Handbook of Pharmaceutical Manufacturing Formulations. CRC Press, pp. 123–130.

Stanley, P., 2016. What Have We Learned from Glycosyltransferase Knockouts in Mice? J. Mol. Biol. 428, 3166–3182.

Sterner, E., Flanagan, N., Gildersleeve, J.C., 2016. Perspectives on Anti-Glycan Antibodies Gleaned from Development of a Community Resource Database. ACS Chem. Biol. 11, 1773–1783.

Strasser, R., 2023. Plant glycoengineering for designing next-generation vaccines and therapeutic proteins. Biotechnol. Adv. 67, 108197.

Sun, Y., Han, X., Shang, C., Wang, Y., Xu, B., Jiang, S., Mo, Y., Wang, D., Ke, Y., Zeng, X., 2022. The downregulation of type I IFN signaling in G-MDSCs under tumor conditions promotes their development towards an immunosuppressive phenotype. Cell Death Dis. 13, 36.

Syed, P., Gidwani, K., Kekki, H., Leivo, J., Pettersson, K., Lamminmäki, U., 2016. Role of lectin microarrays in cancer diagnosis. Proteomics 16, 1257–1265.

Szabo, Z., Thayer, J.R., Reusch, D., Agroskin, Y., Viner, R., Rohrer, J., Patil, S.P., Krawitzky, M., Huhmer, A., Avdalovic, N., Khan, S.H., Liu, Y., Pohl, C., 2018. High performance anion exchange and hydrophilic interaction liquid chromatography approaches for comprehensive mass spectrometry-based characterization of the N-glycome of a recombinant human erythropoietin. J. Proteome Res. 17, 1559–1574.

Takamatsu, S., Antonopoulos, A., Ohtsubo, K., Ditto, D., Chiba, Y., Le, D.T., Morris, H.R., Haslam, S.M., Dell, A., Marth, J.D., Taniguchi, N., 2010. Physiological and glycomic

characterization of N-acetylglucosaminyltransferase-IVa and -IVb double deficient mice. Glycobiology 20, 485–497.

Tateno, H., Nakamura-Tsuruta, S., Hirabayashi, J., 2009. Comparative analysis of core-fucose-binding lectins from Lens culinaris and Pisum sativum using frontal affinity chromatography. Glycobiology 19, 527–536.

Taujale, R., Zhou, Z., Yeung, W., Moremen, K.W., Li, S., Kannan, N., 2021. Mapping the glycosyltransferase fold landscape using interpretable deep learning. Nat. Commun. 12. https://doi.org/10.1038/s41467-021-25975-9

Tejwani, V., Andersen, M.R., Nam, J.H., Sharfstein, S.T., 2018. Glycoengineering in CHO Cells: Advances in Systems Biology. Biotechnol. J. 13, e1700234.

Terabayashi, T., Hanada, K., 2018. Genome instability syndromes caused by impaired DNA repair and aberrant DNA damage responses. Cell Biol. Toxicol. 34, 337–350.

Theodoratou, E., Thaçi, K., Agakov, F., Timofeeva, M.N., Štambuk, J., Pučić-Baković, M., Vučković, F., Orchard, P., Agakova, A., Din, F.V.N., Brown, E., Rudd, P.M., Farrington, S.M., Dunlop, M.G., Campbell, H., Lauc, G., 2016. Glycosylation of plasma IgG in colorectal cancer prognosis. Sci. Rep. 6, 28098.

Thiele, M., Arnold, N., Scheidig, A.J., Winkler, K., 2018. Generation of Highly Productive Chinese Hamster Ovary Cells by Application of Metabolic Pool Selection During Cell Line Development. Current Biotechnology 7, 355–367.

Thomès, L., Burkholz, R., Bojar, D., 2021. Glycowork: A Python package for glycan data science and machine learning. Glycobiology 31, 1240–1244.

Thompson, R., Creavin, A., O'Connell, M., O'Connor, B., Clarke, P., 2011. Optimization of the enzyme-linked lectin assay for enhanced glycoprotein and glycoconjugate analysis. Anal. Biochem. 413, 114–122.

Tomlinson, I.M., 2004. Next-generation protein drugs. Nat. Biotechnol. 22, 521–522.

Vander Heiden, M.G., Cantley, L.C., Thompson, C.B., 2009. Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science 324, 1029–1033.

Varki, A., 2017. Biological roles of glycans. Glycobiology 27, 3–49.

Varki, A., Cummings, R.D., Esko, J.D., Stanley, P., Hart, G.W., Aebi, M., Mohnen, D., Kinoshita, T., Packer, N.H., Prestegard, J.H., Schnaar, R.L., Seeberger, P.H. (Eds.), n.d. Essentials of Glycobiology. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).

Varki, A., Cummings, R.D., Esko, J.D., Stanley, P., Hart, G.W., Aebi, M., Mohnen, D., Kinoshita, T., Prestegard, J.H., Packer, N.H., Schnaar, R.L., Seeberger, P.H., 2022. Essentials of Glycobiology. Cold Spring Harbor Laboratory Press.

Vliegenthart, J.F.G., 2017. The impact of defining glycan structures. Perspect. Sci. 11, 3–10.

Walsh, G., Jefferis, R., 2006. Post-translational modifications in the context of therapeutic proteins. Nat. Biotechnol. 24, 1241–1252.

Walsh, G., Walsh, E., 2022. Biopharmaceutical benchmarks 2022. Nat. Biotechnol. 40, 1722–1760.

Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A., 2014. Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods 11, 333–337.

Wang, M., Zhu, J., Lubman, D.M., Gao, C., 2019. Aberrant glycosylation and cancer biomarker discovery: a promising and thorny journey. Clin. Chem. Lab. Med. 57, 407–416.

Wang, Y., Jin, M., Chen, G., Cui, X., Zhang, Y., Li, M., Liao, Y., Zhang, X., Qin, G., Yan, F., Abd El-Aty, A.M., Wang, J., 2019. Bio-barcode detection technology and its research applications: A review. J. Advert. Res. 20, 23–32.

Wang, Y., Yu, G., Han, Z., Yang, B., Hu, Y., Zhao, X., Wu, J., Lv, Y., Chai, W., 2011. Specificities of Ricinus communis agglutinin 120 interaction with sulfated galactose. FEBS Lett. 585, 3927–3934.

Windwarder, M., Altmann, F., 2014. Site-specific analysis of the O-glycosylation of bovine fetuin by electron-transfer dissociation mass spectrometry. J. Proteomics 108, 258–268.

Wooke, Z., Nagy, G., Barnes, L.F., Pohl, N.L.B., 2019. Development of a Post-Column Liquid Chromatographic Chiral Addition Method for the Separation and Resolution of Common Mammalian Monosaccharides. J. Am. Soc. Mass Spectrom. 30, 419–425.

Wozny, A.-S., Alphonse, G., Cassard, A., Malésys, C., Louati, S., Beuve, M., Lalle, P., Ardail, D., Nakajima, T., Rodriguez-Lafrasse, C., 2020. Impact of hypoxia on the double-strand break repair after photon and carbon ion irradiation of radioresistant HNSCC cells. Sci. Rep. 10, 21357.

Wu, A.M., Wu, J.H., Singh, T., Lai, L.-J., Yang, Z., Herp, A., 2006. Recognition factors of Ricinus communis agglutinin 1 (RCA(1)). Mol. Immunol. 43, 1700–1715.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S., 2021. A Comprehensive Survey on Graph Neural Networks. IEEE Trans Neural Netw Learn Syst 32, 4–24.

Wuhrer, M., 2013. Glycomics using mass spectrometry. Glycoconj. J. 30, 11–22.

Wuhrer, M., Stam, J.C., van de Geijn, F.E., Koeleman, C.A.M., Verrips, C.T., Dolhain, R.J.E.M., Hokke, C.H., Deelder, A.M., 2007. Glycosylation profiling of immunoglobulin G (IgG) subclasses from human serum. Proteomics 7, 4070–4081.

Xie, Y., Mota, L.M., Bergin, A., O'Flaherty, R., Jones, A., Morgan, B., Butler, M., 2021. High-throughput and high-sensitivity N-Glycan profiling: A platform for biopharmaceutical development and disease biomarker discovery. Anal. Biochem. 623, 114205.

Xie, Y., Sheng, Y., Li, Q., Ju, S., Reyes, J., Lebrilla, C.B., 2020a. Determination of the glycoprotein specificity of lectins on cell membranes through oxidative proteomics. Chem. Sci. 11, 9501–9512.

Xie, Y., Sheng, Y., Li, Q., Ju, S., Reyes, J., Lebrilla, C.B., 2020b. Determination of the glycoprotein specificity of lectins on cell membranes through oxidative proteomics. Chem. Sci. 11, 9501–9512.

Yang, M., Butler, M., 2000. Effects of ammonia on CHO cell growth, erythropoietin production, and glycosylation. Biotechnol. Bioeng. 68, 370–380.

Yang, Z., Wang, S., Halim, A., Schulz, M.A., Frodin, M., Rahman, S.H., Vester-Christensen, M.B., Behrens, C., Kristensen, C., Vakhrushev, S.Y., Bennett, E.P., Wandall, H.H., Clausen, H., 2015. Engineered CHO cells for production of diverse, homogeneous glycoproteins. Nat. Biotechnol. 33, 842–844.

Yin, B., Wang, Q., Chung, C.-Y., Bhattacharya, R., Ren, X., Tang, J., Yarema, K.J., Betenbaugh, M.J., 2017. A novel sugar analog enhances sialic acid production and biotherapeutic sialylation in CHO cells. Biotechnol. Bioeng. 114, 1899–1902.

Yom, A., Chiang, A., Lewis, N.E., 2024. Boltzmann Model Predicts Glycan Structures from Lectin Binding. Anal. Chem. https://doi.org/10.1021/acs.analchem.3c04992

Zaia, J., 2010. Mass spectrometry and glycomics. OMICS 14, 401–418.

Zhang, X.-M., Liang, L., Liu, L., Tang, M.-J., 2021. Graph Neural Networks and Their Current Applications in Bioinformatics. Front. Genet. 12, 690049.

Zhong, C., Li, P., Argade, S., Liu, L., Chilla', A., Liang, W., Xin, H., Eliceiri, B., Choudhury, B., Ferrara, N., 2020. Inhibition of protein glycosylation is a novel pro-angiogenic strategy that acts via activation of stress pathways. Nat. Commun. 11, 6330.

Zou, X., Yoshida, M., Nagai-Okatani, C., Iwaki, J., Matsuda, A., Tan, B., Hagiwara, K., Sato, T., Itakura, Y., Noro, E., Kaji, H., Toyoda, M., Zhang, Y., Narimatsu, H., Kuno, A., 2017. A standardized method for lectin microarray-based tissue glycome mapping. Sci. Rep. 7, 43560.