

GlycanGT: a pretrained graph transformer framework for glycan graph representation and generative learning

Akihiro Kitani^{1,2, }, Bingyuan Zhang^{2, }, Koichi Himori^{2, }, Yusuke Matsui^{*,1,2, }

¹Biomedical and Health Informatics Unit, Department of Integrated Health Sciences, Nagoya University Graduate School of Medicine, 1-1-20 Daiko-minami, Higashi-ku, Nagoya, Aichi, 461-8673, Japan

²Systems Biology Division, Institute for Glyco-core Research (iGCORE), Nagoya University, 1-7 Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-0814, Japan

*Corresponding author. Yusuke Matsui, Biomedical and Health Informatics Unit, Department of Integrated Health Sciences, Nagoya University Graduate School of Medicine, Nagoya, Japan; Systems Biology Division, Institute for Glyco-core Research (iGCORE), Nagoya University, 1-7 Furo-cho, Chikusa-ku, Nagoya, Aichi 464-0814, Japan. E-mail: matsui.yusuke.d4@f.mail.nagoya-u.ac.jp

Associate Editor: Arne Elofsson

Abstract

Motivation: Glycans are highly diverse biological sequences, but their functional understanding has lagged behind proteins and nucleic acids. Many glycans remain ambiguously annotated, limiting computational analyses. Existing computational approaches are primarily graph-based, capturing local structural features but struggling to model global patterns and incomplete sequences.

Results: We present GlycanGT, a graph-transformer-based pretrained model for glycans. Glycans were represented as graphs of monosaccharides and glycosidic bonds, and the model was pretrained using a masked language modeling objective. GlycanGT demonstrated higher performance than existing methods across 8 benchmark tasks (e.g., 0.844 AUPRC for immunogenicity classification), and its embeddings formed biologically relevant clusters that recovered known N- and O-glycan categories. Moreover, GlycanGT accurately proposed candidates for ambiguous sequences, maintaining >80% top-5 accuracy for both monosaccharide and glycosidic bond predictions under high masking levels.

Availability and implementation: The source code used in this study is available at <https://github.com/matsui-lab/GlycanGT> and archived on Zenodo (DOI: 10.5281/zenodo.18636040); pretrained model weights are provided via Hugging Face (<https://huggingface.co/Akikitani295/GlycanGT>).

Contact: matsui.yusuke.d4@f.mail.nagoya-u.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Glycans are structurally diverse biomolecules that play essential roles in protein folding, cell–cell communication, and immune regulation (Wells *et al.* 2001, Bektas and Rubenstein, 2011, Hart *et al.* 2011, Gu *et al.* 2012, Varki, 2017, Pinho *et al.* 2023, He *et al.* 2024, Hao *et al.* 2025). Aberrant glycosylation is implicated in many diseases such as cancers (Pinho and Reis, 2015), neurological disorders (Pradeep *et al.* 2023, Zhang *et al.* 2024), and infectious diseases (Zhang and Qu, 2021), yet our understanding of glycans remains far behind that of proteins and nucleic acids due to their nonlinear branching, compositional heterogeneity, and stochastic biosynthesis (Pothukuchi *et al.* 2019). Public repositories such as GlyTouCan (Tiemeyer *et al.* 2017, Fujita *et al.* 2021) contain over 250,000 entries, but many sequences remain incomplete or ambiguously annotated, underscoring the need

for computational methods to capture glycan complexity. In this study, we focus on structurally defined carbohydrate chains represented in GlyTouCan, rather than intact glycoconjugates or large polysaccharide assemblies, which constitute the broader biochemical definition of glycans.

Early computational approaches relied on motif-based features (Porter *et al.* 2010, Coff *et al.* 2020, Bao *et al.* 2021, Li *et al.* 2022), while more recent machine learning methods have leveraged sequence models and graph neural networks (GNNs). RNN-based language models (e.g., SweetTalk (Bojar *et al.* 2021)) treated glycans as “glycowords,” and GNN-based methods (e.g., SweetNet (Burkholz *et al.* 2021), GIFFLAR (Joeres and Bojar, 2024), and GlycanAA (Xu *et al.* 2025)) represented glycans as graphs, achieving improved performance in multiple prediction tasks (e.g., taxonomy, glycosylation, and immunogenicity). However, GNNs suffer from over-smoothing—where node

features become indistinguishable as layers deepen—and over-squashing—where long-range dependencies are compressed into limited representations—both of which limit their ability to capture global structural patterns (Oono and Suzuki, 2019). Moreover, there remains no large-scale pretrained model for glycans that can jointly learn whole-structure representations and support contextual completion of incompletely annotated glycan sequences.

Here, we present GlycanGT (Glycan Graph Transformer), a pretrained model based on a graph transformer (Yun *et al.* 2019, Ying *et al.* 2021, Kim *et al.* 2022, Rampásek *et al.* 2022) architecture pretrained with a masked language modeling objective. The acronym “GT” denotes Graph Transformer. Unlike GNNs, graph transformers apply attention mechanisms (Vaswani *et al.* 2017) across entire graphs, enabling learning of both local and global structural relationships. GlycanGT achieves state-of-the-art performance across multiple prediction tasks, uncovers biologically meaningful structural clusters, and accurately proposes candidates for incomplete sequences. Collectively, these findings position GlycanGT as a pretrained graph transformer for glycans, capable of supporting a broad spectrum of downstream tasks—including classification, functional prediction, and sequence completion—as well as contributing to glycan database expansion and translational applications.

2 Materials and methods

2.1 Data sources

We obtained pretraining glycans from GlyCosmos (Yamada *et al.* 2020), which provides access to GlyTouCan (Tiemeyer *et al.* 2017, Fujita *et al.* 2021) entries, as of 17 June 2025 ($n = 244,842$). Glycans containing ambiguous symbols (“?”) were excluded. For entries with uncertain linkage notations such as “ $\alpha 1-3/5$,” we expanded them into all admissible alternatives (e.g., “ $\alpha 1-3$ ” and “ $\alpha 1-5$ ”), treating each alternative as an independent glycan entry. Downstream benchmark datasets (taxonomy, glycosylation, immunogenicity) were taken from GlycanML (Xu *et al.* 2024). In GlycanML, motif frequencies were precomputed and used to perform k-means clustering ($k = 10$), with eight clusters designated for training and the remainder split into validation and test sets; we adopted these official splits. To prevent data leakage, any glycans overlapping downstream tasks were removed from pretraining, yielding 83,739 glycans. The full list of glycans used for pretraining, the entries removed due to ambiguous symbols, and the expanded linkage variants are provided in [Supplementary Table S1](#), available as [supplementary data](#) at *Bioinformatics* online.

2.2 Graph construction

Glycans in IUPAC-condensed notation were converted to graphs using `glycowork glycan_to_nxGraph` function and extracted head–relation–tail (node–edge–node) triples. Monosaccharides were treated as node tokens and glycosidic bonds as edge tokens. Dictionaries for monosaccharides and glycosidic bonds were constructed based on the GlyTouCan v3.0 notation, after removing ambiguous symbols (“?”, “/”). Stereochemistry

and modifications (e.g., L/D forms; “GalNAc3,4,6Ac3,” “GalNAc3,4,6Me3”) were encoded as distinct tokens. Linkages included canonical forms (e.g., “ $\alpha 1-1$ ”) and allowed variants (“1-1,” “1-P-1,” “ $\alpha 1-1\alpha$ ”). Terminal linkages were excluded from prediction targets because they lack sufficient contextual information for reliable modeling. The graph-construction scripts are publicly available at <https://github.com/matsui-lab/GlycanGT>.

2.3 GlycanGT architecture

Here we adopt a TokenGT (Kim *et al.* 2022) -style pure graph transformer (Yun *et al.* 2019, Ying *et al.* 2021, Rampásek *et al.* 2022) tailored to glycans: every monosaccharide (node) and glycosidic bond (edge) is treated as an independent token, with no graph-specific message passing or locality constraints hard-coded. Building upon TokenGT, we further introduced a masked language modeling (MLM) pretraining objective over both node and edge tokens, enabling self-supervised representation learning tailored to glycans. As shown by Kim *et al.* (2022), each token embedding is the sum of (i) a linear projection of its content features, (ii) orthogonal random features (ORFs) used as node identifiers—a node v receives $[P_v, P_v]$ and an edge (u, v) receives $[P_u, P_v]$ —where the mutual orthogonality of P_v allows the model to infer which nodes and edges are connected through attention inner products, and (iii) a trainable type embedding that distinguishes nodes from edges. A trainable [Graph] token is prepended, and all tokens are projected to model width d . The sequence is processed by post-norm Transformer encoder blocks, where each sub-layer (multi-head self-attention or MLP) is followed by LayerNorm and a residual connection. Attention weights (normalized dot products) (Vaswani *et al.* 2017) aggregate information across all tokens so the model can capture both local branch motifs and long-range cross-branch relations; multi-head attention encourages diverse structural dependencies. The [Graph] token serves as the glycan representation for downstream classifiers. We instantiated four scales (ss/small/medium/large) by varying depth, width, and heads ([Supplementary Table S2](#), available as [supplementary data](#) at *Bioinformatics* online). Kim *et al.* (2022) showed theoretically that TokenGT with node/type identifiers is at least as expressive as a 2-IGN (and thus 2-WL), exceeding message-passing GNNs in expressiveness; empirically, TokenGT outperformed GNN baselines and effectively captured global dependencies via self-attention. Because it avoids message passing, it does not inherit typical over-smoothing issues. Full attention is $O((n+m)^2)$ with the number of tokens per glycan, where n denotes the number of monosaccharide residues (nodes) and m the number of glycosidic bonds (edges); this cost remains practical for typical glycan sizes. For identifier construction, we use fixed, seeded orthogonal random features (ORFs) to ensure reproducibility while avoiding Laplacian-based eigendecomposition. Implementation details and hyperparameters are available in [Supplementary Table S2](#), available as [supplementary data](#) at *Bioinformatics* online and our public repository.

2.4 Pretraining objective

We used masked language modeling (MLM) over both node (monosaccharide) and edge (linkage) tokens. Masking ratios of

5%, 15%, 25%, 35%, 45%, and 55% were evaluated. The masking ratio for the final model was determined based on the Macro-F1 scores for taxonomy and glycosylation tasks, and AUPRC for the immunogenicity prediction tasks (Supplementary Figure S3, available as [supplementary data](#) at *Bioinformatics* online). The loss was cross-entropy with an edge loss weight of 0.5. Optimization used AdamW (learning rate 1×10^{-6} , weight decay 0.01). A 90/10 train/validation split with early stopping (no mean validation-loss improvement for 15 epochs for ss/small or 30 epochs for medium/large) determined the number of epochs, after which we retrained on the full set.

2.4.1. Ablation

To evaluate the contribution of each token type to model learning, we compared a balanced 35% node–35% edge masking strategy with two single-modality variants: (i) 70% node masking with 0% edge masking, and (ii) 70% edge masking with 0% node masking, ensuring a comparable total masking ratio ($\sim 35\%$) across settings. These experiments were designed to assess whether jointly masking both token types improves contextual learning. Final-layer attention weights were averaged across heads and visualized as token–token attention matrices to examine modality-specific focus patterns (Supplementary Figure S6d–f, available as [supplementary data](#) at *Bioinformatics* online).

2.5 Downstream task evaluation

For each downstream task, glycans were encoded by the pre-trained model and the [Graph] token embedding was used as the feature vector. We adopted three benchmark tasks—taxonomy, glycosylation, and immunogenicity—from the GlycanML dataset (Xu *et al.* 2024), which provides standardized task definitions and dataset splits. These tasks were chosen because they represent biologically meaningful and well-established prediction problems in glycoinformatics. The taxonomy task involves predicting glycan origin at eight hierarchical levels (domain, kingdom, phylum, class, order, family, genus, and species), comprising 13,209 glycans from SugarBase with 4–1,737 classes per level. The glycosylation task aims to classify glycans as N-linked, O-linked, or free, using 1,683 glycans curated from GlyConnect. The immunogenicity task involves binary classification of 1,320 glycans annotated for immune activity based on literature evidence in SugarBase. All three tasks employ motif-based dataset splits (8:1:1 for training, validation, and test) to evaluate generalization across structurally distinct glycans. Tasks and metrics were: taxonomy (domain to species; Macro-F1, accuracy), glycosylation (N-linked, O-linked, free; Macro-F1, accuracy), and immunogenicity (binary; AUPRC, accuracy). We trained support vector machines (SVM) (Boser *et al.* 1992) and LightGBM (Ke *et al.* 2017) classifiers. These classifiers were chosen as lightweight yet strong baseline models that are well-established for tabular and structured data. Our primary goal was to evaluate the quality of GlycanGT representations rather than to optimize classifier architectures; thus, we used SVM and LightGBM for their simplicity, robustness, and reproducible performance across diverse datasets. For SVM, we performed randomized hyperparameter search over 10 iterations, testing radial basis function (RBF) and linear kernels with regularization parameter

$C \sim \log U (10^{-3}, 10^2)$ and $\gamma \sim \log U (10^{-4}, 10^{-1})$ for RBF. For LightGBM, we ran 15 randomized search iterations with the following ranges: `n_estimators` $\in [100, 1000]$, `learning_rate` $\sim \log U (0.01, 0.3)$, `num_leaves` $\in [20, 150]$, `max_depth` $\in [5, 20]$, `reg_alpha`, `reg_lambda` $\sim \log U (10^{-2}, 10^1)$, and `colsample_bytree`, `subsample` $\sim U (0.6, 1.0)$. All classifiers used class-balanced weighting and were trained with consistent random seeds for comparability. We followed the official fixed train/validation/test splits of GlycanML. Within the combined training set (`train` $\cup$ `validation`), hyperparameters were selected by RandomizedSearch with 3-fold cross-validation. The best model was then evaluated once on the held-out test split. To account for stochasticity, we repeated the entire procedure with three different random seeds and reported mean $\pm$ s.d. over the three runs, following the evaluation protocol of GlycanML.

2.6 Baseline methods

We compared GlycanGT against four representative graph-based models: SweetNet (Burkholz *et al.* 2021), GlycanAA (Xu *et al.* 2025) Graphormer (Ying *et al.* 2021), and an RGCN implementation provided by GlycanML (Xu *et al.* 2024). SweetNet applies GNNs to glycan graphs, focusing on local branching motifs. GlycanAA represents glycans as heterogeneous graphs with atom and monosaccharide nodes and hierarchical message passing between these levels using relational graph convolution (RGConv) (Schlichtkrull *et al.* 2017), enabling multi-scale feature learning. Graphormer is a type of Graph Transformer that incorporates three structural encodings—Centrality, Spatial, and Edge—into the attention mechanism as additive attention biases, thereby directly integrating pairwise structural relationships between nodes into the self-attention computation. For comparison with GlycanGT, Graphormer was trained using monomer- and linkage-level graph representations instead of atom-level inputs. A large configuration (12 layers, 768-dim, 32 heads) with AdamW (lr $1e-5$, wd 0.01), batch size 32, dropout 0.1, and early stopping was adopted. RGCN models multi-relational edges to capture diverse linkage types; we used the official GlycanML implementation (accessed in July 2025) with default hyperparameters as described in Xu *et al.* (2024). All baselines were trained and evaluated on the same datasets and data splits described in Section 2.5 to ensure fair comparison.

2.7 Structural similarity analysis using graph kernels

To evaluate whether GlycanGT captures global structural patterns beyond local substructure similarities, we computed pairwise graph similarities using the Weisfeiler–Lehman (WL) subtree kernel (Nguyen *et al.* 2021) implemented in the GraKeL library. Glycans were first converted from IUPAC-condensed notation into NetworkX graphs using the `glycan_to_nxGraph` function of glycowork. Each node was labeled by its monosaccharide identity, and edges represented glycosidic linkages. The WL subtree kernel iteratively compares relabeled node neighborhoods to quantify structural similarity across branching patterns. A normalized similarity matrix K_{global} was computed for all pairwise glycan combinations. The resulting similarity matrices were visualized as hierarchical cluster heatmaps.

2.8 Visualization and motif enrichment analysis

To visualize the structural organization of glycan representations learned by GlycanGT, we projected the pretrained [Graph] token embeddings into two dimensions using t-distributed stochastic neighbor embedding (t-SNE) implemented in scikit-learn. Each glycan was represented by the [Graph] embedding obtained from the pretrained large model without additional fine-tuning. The t-SNE algorithm was initialized by PCA and optimized for 1,000 iterations with a perplexity of 30. Subsequently, to examine functional relationships among embedding clusters, we conducted motif enrichment analysis. GlycanGT embeddings for the glycosylation dataset were reduced to 50 dimensions by PCA and clustered by k-means ($k = 6$). Cluster-wise motif enrichment was assessed using glycowork's `get_pvals_motifs` function (Thomès *et al.* 2021). For each cluster, IUPAC-condensed glycan names were labeled as target (1) or non-target 0, and motif occurrence ratios were compared using Welch's t-test with FDR correction and Cohen's d effect sizes. Motifs with $p < 0.05$ after FDR correction were considered significantly enriched.

2.9 Ambiguous glycan analysis

Because masked language modeling (MLM) enables masked token prediction, we evaluated GlycanGT's ability to reconstruct incomplete or uncertain glycan sequences resembling real-world ambiguity patterns observed in GlyTouCan (Tiemeyer *et al.* 2017, Fujita *et al.* 2021). We considered two representative scenarios derived from empirical distributions of ambiguous annotations (Supplementary Figure S7b, available as [supplementary data](#) at *Bioinformatics* online): (i) 0% masking of linkages with 10–90% masking of monosaccharides, and (ii) 100% masking of linkages with 0–90% masking of monosaccharides. To ensure strict separation from pretraining, we used 5678 glycans excluded from the pretraining set due to compositional overlap with downstream datasets. Predictions were ranked by softmax probabilities over the masked token vocabulary, and performance was measured by hit@K ($K = \{1, 2, 3, 5, 10, 20, 30\}$), defined as the fraction of cases where the correct token (monosaccharide or linkage) appeared within the top-K predictions. This experiment was designed to assess the model's capability for automated reconstruction of incompletely annotated glycans.

2.10 Statistics

Unless noted, error bars denote mean $\pm$ s.d. across replicates. Model metrics (Macro-F1, AUPRC, accuracy) were averaged over three runs. Multiple testing was corrected by Benjamini–Hochberg.

2.11 Reproducibility and computing environment

Experiments ran on NVIDIA L40 GPUs (46 GB memory). Software: Python 3.12, PyTorch 2.5.1 (CUDA 12.1). Codes are provided at <https://github.com/matsui-lab/GlycanGT> and archived on Zenodo (DOI: 10.5281/zenodo.18636040). Pretrained model

weights are available on Hugging Face: <https://huggingface.co/Akikitani295/GlycanGT>.

3 Results

3.1 Pretraining of GlycanGT using masked language modeling

For model pretraining, glycans with complete structural information were curated from GlyCosmos (Yamada *et al.* 2020, Fujita *et al.* 2021). Glycans containing ambiguous symbols (e.g., “?”) as well as those included in downstream analyses were excluded to avoid data leakage, resulting in a pretraining dataset of 83,739 glycans. Each glycan was converted into a graph representation, with monosaccharides treated as nodes and glycosidic bonds as edges, and provided as input to TokenGT (Kim *et al.* 2022), a variant of the graph transformer architecture (Yun *et al.* 2019, Ying *et al.* 2021, Rampásek *et al.* 2022) (Fig. 1 and Supplementary Figure S1, available as [supplementary data](#) at *Bioinformatics* online, see Materials and Methods for details).

Pretraining was performed using a masked language modeling (MLM) objective, originally developed for natural language processing (Devlin *et al.* 2019), in which monosaccharides and glycosidic bonds were randomly masked at specified probabilities and predicted from the surrounding context. Six masking ratios (5%, 15%, 25%, 35%, 45%, and 55%) were examined (Supplementary Figure S2, Supplementary Tables S2 and S3, available as [supplementary data](#) at *Bioinformatics* online). Four model scales (ss, small, medium, and large) were trained, among which the large model consistently achieved the lowest loss and was therefore used for further exploration of optimal masking ratios.

Embeddings obtained from the pretrained models were applied as input features for downstream prediction tasks, with support vector machines (SVM) (Boser *et al.* 1992) and LightGBM (Ke *et al.* 2017) used for classification. Three types of classification tasks were evaluated: taxonomy, glycosylation, and immunogenicity, using previously published benchmark datasets (Xu *et al.* 2024). Across tasks, masking ratios of 25–45% generally yielded the best performance, with a masking ratio of 35% providing consistently strong results (Supplementary Figure S3, available as [supplementary data](#) at *Bioinformatics* online). Therefore, we adopted the large model pretrained with a 35% masking ratio for subsequent evaluations.

3.2 Benchmarking GlycanGT against existing models

We next evaluated the performance of GlycanGT against previously reported methods (Xu *et al.* 2024). As baselines, we considered SweetNet (Burkholz *et al.* 2021), GlycanAA (Xu *et al.* 2025), Graphormer (Ying *et al.* 2021), and the relational graph convolutional network (RGCN) implemented in GlycanML (Xu *et al.* 2024). For taxonomy classification spanning domain to species, GlycanGT achieved the best Macro-F1 performance at all levels except the class level (Fig. 2a). However, when evaluated by accuracy (Supplementary Figure S4, available as [supplementary data](#) at *Bioinformatics* online), GlycanAA performed best at five

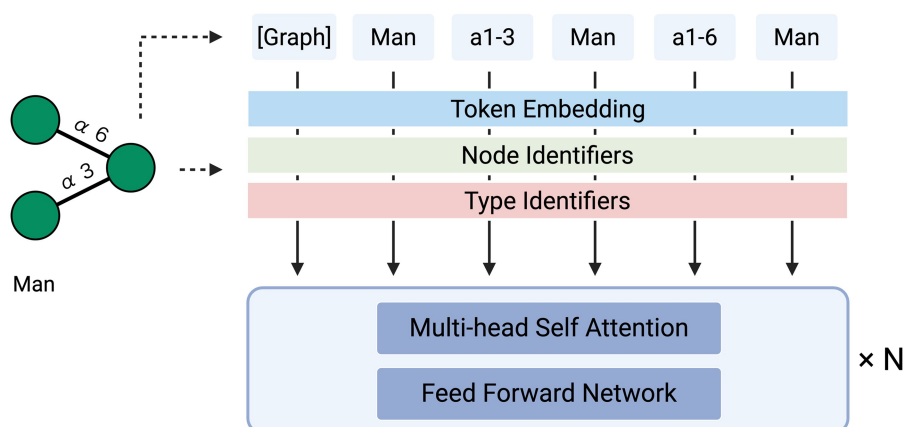


Figure 1 Overview of GlycanGT architecture. Glycans are represented as graphs, where monosaccharides correspond to node tokens and glycosidic bonds to edge tokens. These tokens are processed by a Tokenized Graph Transformer (TokenGT). Each token embedding is constructed as the sum of three components: (1) Token embeddings, derived from token IDs; (2) Node identifiers, orthonormal vectors assigned to nodes to encode graph connectivity; and (3) Type identifiers, trainable vectors indicating whether a token corresponds to a node or an edge. The resulting augmented embeddings are input to the Transformer encoder layers to learn glycan graph representations.

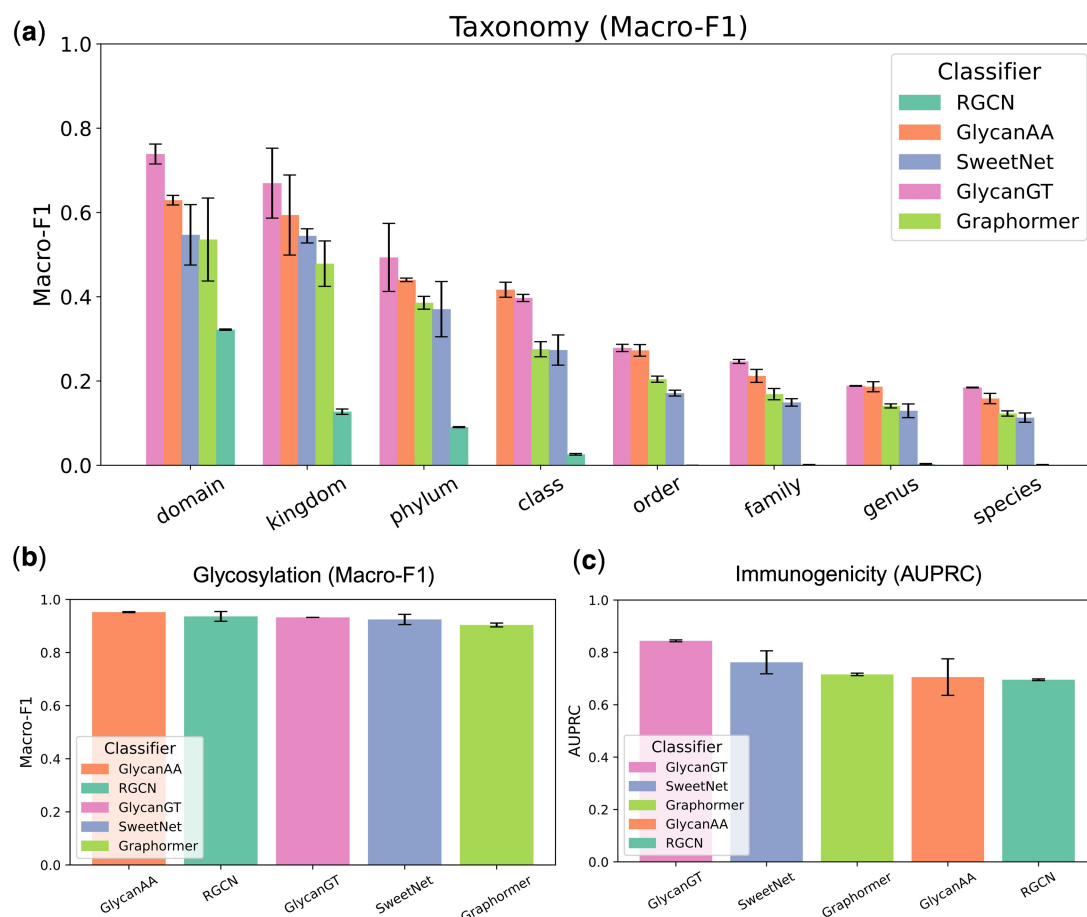


Figure 2 Performance comparison of GlycanGT and baseline models on prediction tasks. (a) Taxonomy classification evaluated by Macro-F1 across eight hierarchical levels (domain to species). (b) Glycosylation prediction evaluated by Macro-F1. (c) Immunogenicity prediction evaluated by AUPRC. Error bars represent mean $\pm$ s.d. across replicates.

taxonomic levels, whereas GlycanGT was best at two levels and Graphormer at one level (see [Supplementary Table S4](#) for details, available as [supplementary data](#) at *Bioinformatics*

online). For glycosylation prediction, GlycanGT yielded a Macro-F1 of 0.932, which was lower than that of GlycanAA (0.952) and RGCN (0.936) (Fig. 2b). For immunogenicity prediction, GlycanGT

achieved the best performance, with an AUPRC of 0.844, compared with 0.705 for GlycanAA and 0.695 for RGCN. Similar trends were also observed when evaluated using accuracy (Supplementary Figure S4 and Supplementary Table S4, available as [supplementary data](#) at *Bioinformatics* online).

3.3 Global structural patterns captured by GlycanGT

GlycanGT showed high performance in taxonomy classification. We hypothesized that this improvement stems from GlycanGT's ability to capture the global structural patterns of glycans more effectively than existing GNN-based models. To test this hypothesis, we applied a graph kernel approach to compute pairwise similarities between glycans and examined structural clustering using hierarchical clustering. Specifically, we employed the Weisfeiler–Lehman (WL) subtree kernel (Nguyen *et al.* 2021), which quantitatively compares branching patterns of glycans by evaluating the similarity of their substructures. Using this approach, structural clustering did not recapitulate distinctions based on taxonomy (Supplementary Figure S5a, available as [supplementary data](#) at *Bioinformatics* online). In contrast, glycans were clustered by glycosylation type (N-linked, O-linked, or free glycans) and by immunogenicity status (immunogenic vs. non-immunogenic) (Supplementary Figure S5b and c, available as [supplementary data](#) at *Bioinformatics* online). Taxonomy classification may depend on distributed structural cues across branches rather than localized subtree similarity, making global-context modeling advantageous. Taken together, these results suggest that taxonomy classification is not strongly dependent on substructural similarities, but rather benefits from the ability of the graph transformer pretraining to capture global structural patterns of glycans.

3.4 Biological interpretation via motif enrichment analysis

To assess the effect of pretraining, glycan representations were projected into two dimensions using t-distributed stochastic neighbor embedding (t-SNE) and plotted for each dataset—taxonomy (Fig. 3a), glycosylation (Fig. 3b), and immunogenicity (Fig. 3c). The embeddings effectively separated classes in taxonomy, glycosylation, and immunogenicity tasks, with additional subclusters observed within each class. To further examine their biological relevance, we clustered the glycosylation embeddings and performed motif enrichment analysis, which identified six clusters characterized by distinct glycan motifs (Fig. 3d, Supplementary Table S5, available as [supplementary data](#) at *Bioinformatics* online). Notably, these clusters corresponded to known biological categories. For example, Cluster 0 was enriched for sialylated N-glycan motifs such as NeuNAc(α 2-3/6/?Gal, which are known to promote tumor cell survival, invasion, and immune evasion through enhanced α 2-3/6-linked sialylation and interaction with Siglec receptors in the tumor microenvironment (van Vliet and van Kooyk, 2025). In contrast, Cluster 2 contained O-glycan cores initiated by GalNAc such as GlcNAc(b1-6)GalNAc, Gal(b1-3)GalNAc—structures typical of mucin-type glycoproteins involved in epithelial protection and cell–cell

adhesion (Bektas and Rubenstein, 2011). Other clusters also captured biologically relevant motifs such as high-mannose, and Plant- and insect-type xylosylated N-glycans. These observations indicate that GlycanGT embeddings preserve functional hierarchies of glycan biosynthesis and organization, capturing biologically meaningful structural diversity that aligns with known N- and O-glycan classes.

3.5 Ablation analysis of masking strategies

To assess the necessity of masking both token types, we compared the balanced 35%–35% masking strategy with two alternatives: 70% monosaccharide-only masking and 70% glycosidic bond-only masking. The balanced strategy consistently yielded superior downstream performance (Supplementary Figure S6a–c, available as [supplementary data](#) at *Bioinformatics* online). Attention weight visualizations further showed that the balanced model distributed attention across both token types, whereas the single-modality models placed disproportionate weight on the unmasked tokens (Supplementary Figure S6d–f, available as [supplementary data](#) at *Bioinformatics* online).

3.6 Prediction of ambiguous glycan sequences

Many glycans in GlyTouCan contain ambiguous sequence information (e.g., “?” or “/”). As of June 2025, more than half of the registered glycans lacked a defined IUPAC-condensed name, and among those with names, ~60% contained ambiguous symbols (Supplementary Figure S7a, available as [supplementary data](#) at *Bioinformatics* online). The distribution of such ambiguities followed two common patterns: (i) unmasked glycosidic bonds with 10–90% of monosaccharides masked, and (ii) fully masked glycosidic bonds with 0–90% of monosaccharides masked (Supplementary Figure S7b, available as [supplementary data](#) at *Bioinformatics* online).

To evaluate GlycanGT under these conditions, we masked complete glycan sequences excluded from pretraining and assessed prediction accuracy. In the first scenario, top-1 accuracy for monosaccharide prediction remained above 80% across most masking ratios (Fig. 4a). In the second, accuracy declined at high masking levels (dropping to ~50% at 90% masking), but the correct monosaccharide or glycosidic bond appeared within the top-5 predictions in >80% of cases (Fig. 4b and c; Supplementary Table S6, available as [supplementary data](#) at *Bioinformatics* online). These results indicate that GlycanGT can propose contextually consistent candidates for glycans with masked or incomplete structural information. This capability suggests potential utility in supporting the curation of glycan databases by prioritizing plausible structural elements, as illustrated by our model-based predictions for GlyTouCan entries with ambiguous annotations (Supplementary Table S7, available as [supplementary data](#) at *Bioinformatics* online).

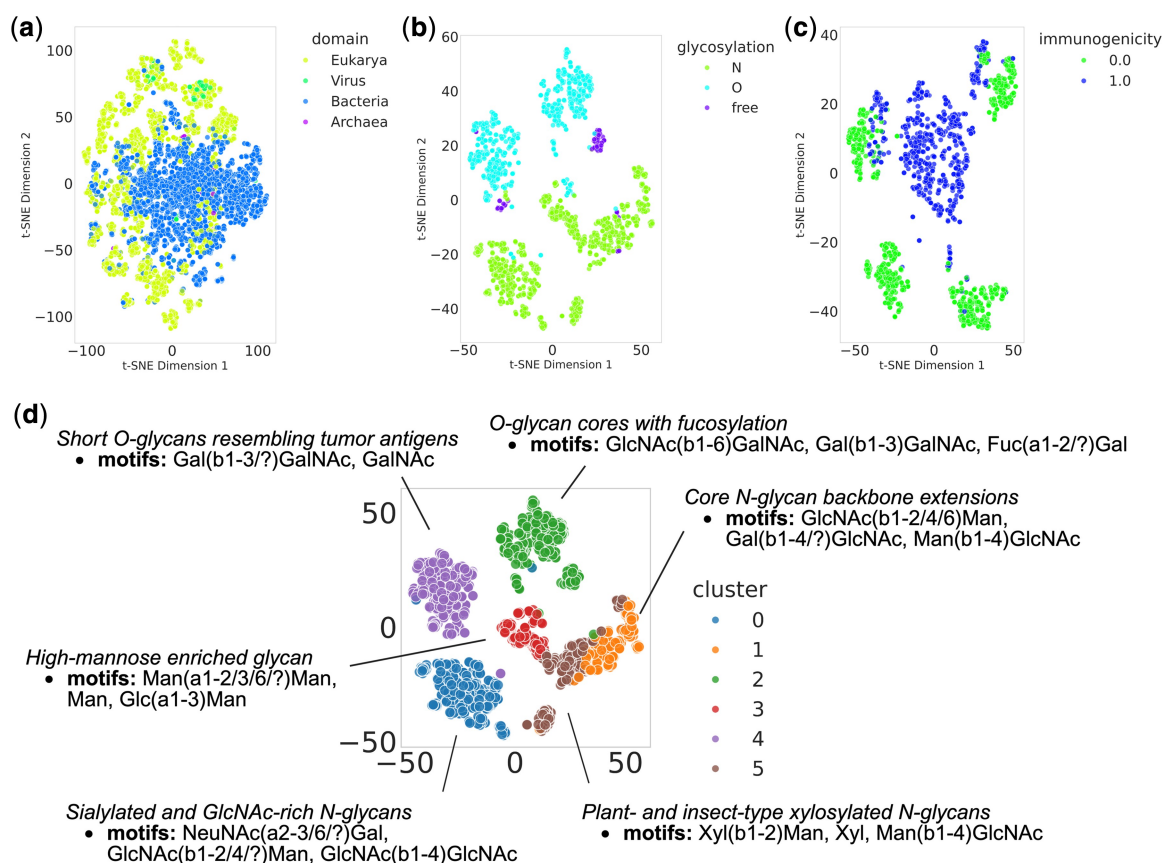


Figure 3 Visualization of glycan representations learned by GlycanGT. Two-dimensional plots of glycan embeddings projected by t-distributed stochastic neighbor embedding (t-SNE). (a) Taxonomy classification at the domain level (eukarya, virus, bacteria, archaea). (b) Glycosylation types (N-linked, O-linked, and free glycans). (c) Immunogenicity (0.0: non-immunogenic, 1.0: immunogenic). (d) Motif enrichment analysis of clusters derived from glycosylation embeddings.

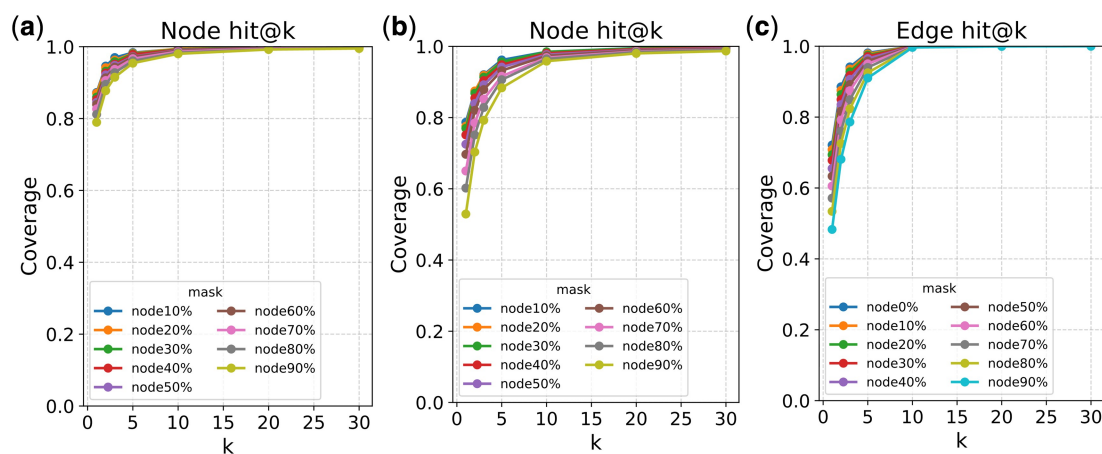


Figure 4 Prediction performance of GlycanGT for ambiguous glycans. (a) Monosaccharide prediction accuracy (hit@K) when glycosidic bonds were not masked (0% masking) and monosaccharide masking ratios varied from 0% to 90%. (b, c) Prediction accuracy when glycosidic bonds were fully masked (100%). (b) Monosaccharide prediction with varying monosaccharide masking ratios (0-90%). (c) Glycosidic bonds prediction with varying monosaccharide masking ratios (0-90%).

4 Discussion

In this study, we developed GlycanGT, a pretrained graph transformer model for glycans based on the graph transformer (Yun *et al.* 2019, Ying *et al.* 2021, Kim *et al.* 2022, Rampásek *et al.* 2022) architecture. By employing masked language modeling for pretraining, GlycanGT outperformed previous approaches such as SweetNet (Burkholz *et al.* 2021), GlycanAA (Xu *et al.* 2025), and RGCN across multiple benchmark tasks. Although GlycanAA achieved slightly better performance in certain taxonomy and glycosylation classification tasks, this may reflect its atom- and residue-level feature representations that emphasize local structural motifs. In contrast, GlycanGT focuses on capturing long-range contextual dependencies across entire glycan structures, which likely accounts for its superior performance in taxonomy and immunogenicity prediction and ambiguous sequence completion. Notably, GlycanGT effectively mitigated limitations of conventional GNN-based models, such as over-smoothing and over-squashing (Oono and Suzuki, 2019), by leveraging full self-attention across all node–edge tokens. This architecture removes the locality constraints inherent to message-passing networks and enables efficient propagation of information across distant branches. Similar graph transformer architectures have been theoretically shown to overcome these limitations (Kim *et al.* 2022, Rampásek *et al.* 2022), and our benchmarking and structural analyses (Sections 3.2 and 3.3) empirically support this improvement, demonstrating enhanced capture of global structural dependencies and superior predictive performance compared with GNN-based models such as SweetNet and GlycanAA. Furthermore, because it was pretrained with a masked language modeling (MLM) objective, GlycanGT can infer missing structural elements by predicting masked monosaccharides and linkages from their surrounding context. This capability enables automated reconstruction of incompletely annotated glycans, as demonstrated by its >80% Top-1 accuracy and >80% Top-5 hit rates in ambiguous sequence prediction (Section 3.6). Such context-aware completion could facilitate the curation of public databases GlyTouCan (Tiemeyer *et al.* 2017, Fujita *et al.* 2021) by suggesting plausible candidates for entries with uncertain or partially defined structures. Although Graphormer also uses global self-attention, our Graphormer baseline was trained from scratch under the same supervised downstream settings and was not pretrained with an MLM objective. Therefore, the observed gap may reflect both architectural differences and the benefit of large-scale pretraining in GlycanGT (which learns connectivity via ORF-based identifiers rather than predefined attention biases).

Despite these strengths, our study has certain limitations. First, the number of fully defined glycan structures remains limited, and the pretraining dataset used in this work comprised only ~80,000 glycans, which may restrict the model's generalization to rare or synthetic structures. In transformer-based models, performance is known to follow scaling laws with respect to model size, data quantity, and computational budget (Kaplan *et al.* 2020). Therefore, we expect that pretraining on larger and more diverse glycan datasets will further improve GlycanGT's performance and generalizability. Second, inconsistencies in glycan notation required stereochemistry and modifications (e.g., L/D forms, acetylation, methylation) as well as linkage

variants (e.g., “ α 1-1,” “1-1,” “1-P-1”) to be treated as distinct tokens. This redundancy can lead to sparse or overlapping representations, potentially affecting embedding stability, even though these variants may represent nearly identical entities. In addition, the quadratic complexity of full self-attention poses computational challenges for scaling to extremely large glycan graphs. Lastly, the masked completion experiment mirrors the MLM pretraining objective. Although high accuracy indicates strong contextual recovery, it does not fully represent real-world structural ambiguity. These results therefore demonstrate contextual completion rather than definitive resolution of uncertain glycans in a biological context. These limitations are expected to be mitigated as larger and more standardized glycan datasets become available, and future work will explore scalable attention mechanisms and cross-modal pretraining that integrates glycomics and glycoproteomics data to further enhance model robustness and generalizability.

Overall, GlycanGT demonstrated strong performance across classification tasks and ambiguous sequence completion tasks, achieving up to 0.844 AUPRC in immunogenicity prediction and maintaining >80% Top-1 accuracy for masked monosaccharide prediction (Sections 3.2 and 3.6). The model further provided biologically interpretable embeddings that facilitated motif discovery and immunogenicity analysis. As both pretrained weights and code are publicly available, GlycanGT can be readily fine-tuned or adapted to additional glycan-related datasets and downstream tasks. We envision GlycanGT serving as a reusable pretrained model for glycans—bridging structural, functional, and omics-level analyses—and contributing to the mechanistic understanding of disease and the development of glycan-based therapeutic strategies.

Author contributions

Akihiro Kitani(Conceptualization [Lead], Data curation [Lead], Formal analysis [Lead], Investigation [Lead], Methodology [Lead], Software [Lead], Validation [Lead], Visualization [Lead], Writing—original draft [Lead], Writing—review & editing [Lead]), Zhang Bingyuan(Conceptualization [Supporting], Formal analysis [Supporting], Investigation [Supporting], Methodology [Supporting], Writing—review & editing [Supporting]), and Koichi Himori(Conceptualization [Supporting], Formal analysis [Supporting], Investigation [Supporting], Methodology [Supporting], Writing—review & editing [Supporting]), Yusuke Matsui(Funding acquisition [Lead], Project administration [Lead], Supervision [Lead], Writing—review & editing [Lead])

Supplementary data

[Document S1. Supplementary Figures S1–S8—pdf file](#)

[Supplementary Table S1](#) Pretraining glycan dataset construction: original IDs, ambiguity removal, and linkage expansion—xlsx file

[Supplementary Table S2](#) Model architectures and number of trainable parameters—xlsx file

[Supplementary Table S3](#) Best validation loss and epochs for GlycanGT models at different masking ratios—xlsx file

[Supplementary Table S4](#) Comparison of GlycanGT with baseline methods—xlsx file

[Supplementary Table S5](#) Motif enrichment analysis of clusters derived from glycosylation embeddings—xlsx file

[Supplementary Table S6](#) Evaluation of GlycanGT prediction performance for ambiguous glycans—xlsx file

[Supplementary Table S7](#) Prediction of ambiguous glycan sequences in GlyTouCan using GlycanGT—xlsx file

Conflict of interest

None declared.

Funding

This work was supported by the Human Glycome Atlas Project (HGA).

Figures were created using [Bio-Render.com](#).

Data availability

All data used in this study were obtained from publicly available sources. The pretrained GlycanGT model weights and usage scripts are available on Hugging Face: <https://huggingface.co/Akikitani295/GlycanGT>. Additional source code for constructing GlycanGT and reproducing the analyses in this study is provided at: <https://github.com/matsui-lab/GlycanGT>, and archived on Zenodo (DOI: 10.5281/zenodo.18636040).

References

- Bao B, Kellman BP, Chiang AWT *et al.* Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis. *Nat Commun* 2021;**12**:4988.
- Bektas M, Rubenstein DS. The role of intracellular protein O-glycosylation in cell adhesion and disease. *J Biomed Res* 2011;**25**:227–36. [https://doi.org/10.1016/S1674-8301\(11\)60031-6](https://doi.org/10.1016/S1674-8301(11)60031-6)
- Bojar D, Powers RK, Camacho DM *et al.* Deep-learning resources for studying glycan-mediated host-microbe interactions. *Cell Host Microbe* 2021;**29**:132–44.e3. <https://doi.org/10.1016/j.chom.2020.10.004>
- Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational learning theory*. New York: Association for Computing Machinery, 1992, 144–52.
- Burkholz R, Quackenbush J, Bojar D. Using graph convolutional neural networks to learn a representation for glycans. *Cell Rep* 2021;**35**:109251. <https://doi.org/10.1016/j.celrep.2021.109251>
- Coff L, Chan J, Ramsland PA *et al.* Identifying glycan motifs using a novel subtree mining approach. *BMC Bioinformatics* 2020;**21**:42.
- Devlin J, Chang MW, Lee K *et al.* 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT, 2019*, pp. 4171–86.
- Fujita A, Aoki NP, Shinmachi D *et al.* The international glycan repository glytoucan version 3.0. *Nucleic Acids Res* 2021;**49**:D1529–33. <https://doi.org/10.1093/nar/gkaa947>
- Gu J, Isaji T, Xu Q *et al.* Potential roles of N-glycosylation in cell adhesion. *Glycoconj J* 2012;**29**:599–607. <https://doi.org/10.1007/s10719-012-9386-1>
- Hao C, Zou Q, Bai X *et al.* Effect of glycosylation on protein folding: From biological roles to chemical protein synthesis. *iScience* 2025;**28**:112605. <https://doi.org/10.1016/j.isci.2025.112605>
- Hart GW, Slawson C, Ramirez-Correa G *et al.* Cross talk between O-glcNacetylation and phosphorylation: roles in signaling, transcription, and chronic disease. *Annu Rev Biochem* 2011;**80**:825–58. <https://doi.org/10.1146/annurev-biochem-060608-102511>
- He M, Zhou X, Wang X *et al.* Glycosylation: mechanisms, biological functions and clinical implications. *Signal Transduct Target Ther* 2024;**9**:194.
- Joeres R, Bojar D. 2024. Higher-order message passing for glycan representation learning. arXiv [cs.LG], preprint: not peer reviewed.
- Kaplan J, McCandlish S, Henighan T *et al.* 2020. Scaling laws for neural language models. arXiv [cs.LG], preprint: not peer reviewed.
- Ke G, Meng Q, Finley T *et al.* LightGBM: a highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems (NeurIPS 2017)*, pp. 3146–3154.
- Kim J, Nguyen D, Min S *et al.* Pure Transformers are Powerful Graph Learners. In: *Advances in Neural Information Processing Systems (NeurIPS 2022)*, pp. 14582–14595.
- Li H, Chiang AWT, Lewis NE. Artificial intelligence in the analysis of glycosylation data. *Biotechnol Adv* 2022;**60**:108008. <https://doi.org/10.1016/j.biotechadv.2022.108008>
- Nguyen DH, Nguyen CH, Mamitsuka H *et al.* Learning subtree pattern importance for Weisfeiler-Lehman based graph kernels. *Mach Learn* 2021;**110**:1585–607.
- Oono K, Suzuki T. Graph Neural Networks exponentially lose expressive power for node classification. In: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
- Pinho SS, Alves I, Gaifem J *et al.* Immune regulatory networks coordinated by glycans and glycan-binding proteins in autoimmunity and infection. *Cell Mol Immunol* 2023;**20**:1101–13. <https://doi.org/10.1038/s41423-023-01074-1>
- Pinho SS, Reis CA. Glycosylation in cancer: mechanisms and clinical implications. *Nat Rev Cancer* 2015;**15**:540–55. <https://doi.org/10.1038/nrc3982>
- Porter A, Yue T, Heeringa L *et al.* A motif-based analysis of glycan array data to determine the specificities of glycan-binding proteins. *Glycobiology* 2010;**20**:369–80. <https://doi.org/10.1093/glycob/cwp187>
- Pothukuchi P, Agliarulo I, Russo D *et al.* Translation of genome to glycome: role of the golgi apparatus. *FEBS Lett* 2019;**593**:2390–411. <https://doi.org/10.1002/1873-3468.13541>
- Pradeep P, Kang H, Lee B *et al.* Glycosylation and behavioral symptoms in neurological disorders. *Transl Psychiatry* 2023;**13**:154.
- Rampášek L, Galkin M, Dwivedi AP *et al.* Recipe for a general, powerful, scalable graph Transformer. In: *Advances in Neural*

- Information Processing Systems (NeurIPS) 2022, pp. 14501–515.
- Schlichtkrull M, Kipf TN, Bloem P *et al.* Modeling relational data with graph convolutional networks. In: Proceedings of the European Semantic Web Conference (ESWC 2018), pp. 593–607.
- Thomès L, Burkholz R, Bojar D. Glycwork: A python package for glycan data science and machine learning. *Glycobiology* 2021; **31**:1240–4. <https://doi.org/10.1093/glycob/cwab067>.
- Tiemeyer M, Aoki K, Paulson J *et al.* GlyTouCan: An accessible glycan structure repository. *Glycobiology* 2017; **27**:915–9. <https://doi.org/10.1093/glycob/cwx066>.
- Van Vliet SJ, Van Kooyk Y. Sialic acids in cancer biology and immunity-recent advancements. *J Biol Chem* 2025; **301**:110641 <https://doi.org/10.1016/j.jbc.2025.110641>. PMC: 40885395
- Varki A. Biological roles of glycans. *Glycobiology* 2017; **27**:3–49. [10.1093/glycob/cww086](https://doi.org/10.1093/glycob/cww086).
- Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS 2017)*, pp. 5998–6008.
- Wells L, Vosseller K, Hart GW. Glycosylation of nucleocytoplasmic proteins: signal transduction and O-glcnaC. *Science* 2001; **291**:2376–8. <https://doi.org/10.1126/science.1058714>.
- Xu M, Geng Y, Zhang Y *et al.* GlycanML: A Multi-Task and Multi-Structure Benchmark for Glycan Machine Learning. In: *Proceedings of the International Conference on Learning Representations (ICLR 2025)*.
- Xu M, Zong J, Wu K *et al.* Modeling all-atom glycan structures via hierarchical message passing and multi-scale pre-training. In: *Proceedings of the International Conference on Machine Learning (ICML 2025)*.
- Yamada I, Shiota M, Shinmachi D *et al.* The glycosmos portal: A unified and comprehensive web resource for the glycosciences. *Nat Methods* 2020; **17**:649–50. <https://doi.org/10.1038/s41592-020-0879-8>.
- Ying C, Bourgeois D, You J, Zitnik M, Leskovec J. Do Transformers really perform bad for graph representation? In: *Advances in Neural Information Processing Systems (NeurIPS 2021)*, pp. 28877–28888.
- Yun S, Jeong M, Kim R, Kang J, Kim HJ. Graph Transformer Networks. In: *Advances in Neural Information Processing Systems (NeurIPS 2019)*, pp. 11960–11970.
- Zhang W, Chen T, Zhao H *et al.* Glycosylation in aging and neurodegenerative diseases. *Acta Biochim Biophys Sin (Shanghai)* 2024; **56**:1208–20. <https://doi.org/10.3724/abbs.2024136>.
- Zhang X-L, Qu H. The role of glycosylation in infectious diseases. *Adv Exp Med Biol* 2021; **1325**:219–37. https://doi.org/10.1007/978-3-030-70115-4_11.