GLYCANSTRUCTBENCH: BENCHMARKING STRUCTURE-FAITHFUL GLYCAN REPRESENTATIONS WITH LINKAGE-AWARE ROOTED PATH TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Glycans pose a distinctive challenge for representation learning because their functions arise from branched architectures, not residue composition alone. Identical residue counts can differ in linkage, stereochemistry, root orientation, and terminal organization, so composition does not determine structure. This raises a central question: do models capture structural information or exploit compositional shortcuts? We introduce GlycanStructBench, a normalized-key-disjoint benchmark for evaluating whether glycan representations preserve structural distinctions across fifteen structural, taxonomic, and biological tasks. Its identity unit is the normalized structure key, not chemical identity. The benchmark includes composition-disjoint and motif-profile/cluster-disjoint stress-test splits, which withhold residue-count groups and whole clusters, respectively, and controlled probes of topology, linkage, and root-dependent processing. We also evaluate GlycoTransportFormer as a representative rooted-tree encoder with linkage-conditioned common-root-frame transport, not as a claim that this operator is optimal. Across seven models and five seeds, the encoder leads the structural and taxonomy families on the normalized-key and composition-disjoint partitions, with paired gains of 1.28 and 8.71 points, and 0.91 and 4.15 points under composition shift; all five paired seed-level differences are positive. On the motif-profile/cluster-disjoint partition it leads on subtype, type, and all eight taxonomy ranks. We report family-level paired effects and cross-partition consistency rather than a single mean rank. A parameter-free retrieval baseline exceeds the encoder on the two sparse biological audit tasks. These results show why glycan benchmarks should control for composition when assessing structure-faithful representations.

1 INTRODUCTION

Glycans are branched chains of sugars attached to cell surfaces and secreted proteins. They help regulate adhesion, receptor activation, immunity, host–pathogen recognition, and tissue organization (Varki, 2017; Moremen et al., 2012; Reily et al., 2019; Zhou & Cobb, 2021). Unlike proteins, glycans are not copied from a single linear template. Enzyme expression, cellular compartments, precursor availability, and local context jointly shape the structures a cell produces (Moremen et al., 2012; Schjoldager et al., 2020; Reily et al., 2019). Small structural changes can therefore alter binding or disease associations.

Two glycans can have the same residue counts and still differ in linkage positions, stereochemistry, root orientation, or terminal organization (Figure 1a). The ambiguity is not artificial: the human milk oligosaccharides LNT and LNnT are regioisomers that share both a composition and a symbol sequence but differ at one linkage. They are catabolized through distinct enzymatic routes and affect microbial physiology in different ways (Özcan & Sela, 2018) (Figure 1b). Yet a random split and one downstream score do not show whether a model retained these distinctions; it may instead rely on residue counts or frequent fragments (Figure 1c). This matters when predictions are used to prioritize biomarkers, interpret glycomic measurements, or design glycan-binding and glycoengineered therapeutics (Cummings & Pierce, 2014; Reily et al., 2019; Esmail & Manolson, 2021).

Glycoinformatics databases connect glycan structures to protein, pathway, tissue, and disease evidence (Ceroni et al., 2011; Aoki-Kinoshita et al., 2016; Tiemeyer et al., 2017; York et al., 2020; Thomès et al., 2021), but the same glycan can appear with different serializations, incomplete annotations, or source identifiers, so fair evaluation requires canonicalization and structure-level leakage control.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

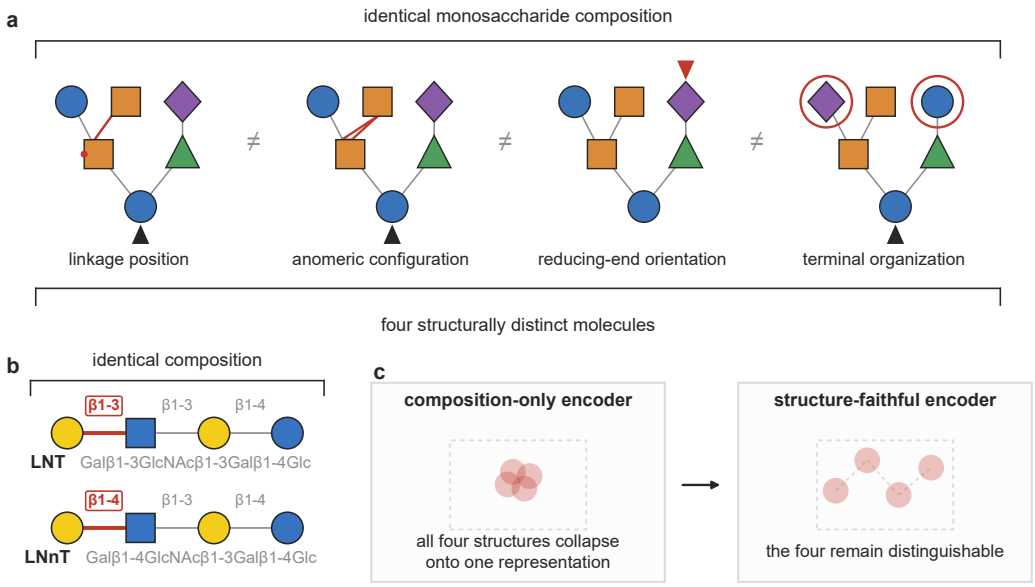


Figure 1: **Composition does not determine structure.** (a) Glycans with the same residue counts can differ in linkage, anomeric configuration, root orientation, or terminal organization. (b) LNT and LNnT share composition but differ at one linkage and follow distinct enzymatic routes (Özcan & Sela, 2018). (c) Composition-only representations collapse such cases, whereas a structure-sensitive evaluation tests whether the representation responds. Panels (a,c) are schematics; symbols follow the Symbol Nomenclature for Glycans (Neelamegham et al., 2019).

Existing task suites measure prediction, but a correct label does not reveal which structural evidence a model used; composition-disjoint evaluation withholds residue-count groups, and controlled probes test whether an implementation responds to specified edits.

We present GLYCANSTRUCTBENCH, which evaluates this question at four explicitly separated levels of evidence: normalized-key-disjoint in-distribution tasks, composition-disjoint splits, a motif-profile/cluster-disjoint stress test, and exact-isomorphism-disjoint partitions. It covers fifteen targets — structural attributes, eight taxonomic ranks, and biological annotations — under one canonicalization, feature schema, and validation-only selection protocol, with controlled edits and isomorphism-preserving controls measuring structural sensitivity separately from downstream accuracy.

The difference from *GlycanML* (Xu et al., 2025a), the closest prior task suite, is therefore what a reported number means rather than how many targets are listed. GlycanML fixes the task families and evaluates models on structure-level splits; we keep three of those families and rebuild everything a number depends on. The identity unit is declared rather than assumed — a persistent normalized structure key — and the four partitions above are applied to the same fifteen targets, each certified by an audit that reports zero leakage at its own level, so a score is attributed to key-disjoint, composition-disjoint, cluster-disjoint or isomorphism-disjoint generalization. On top of that we add a diagnostic layer no prediction suite provides: controlled linkage, anomer, branch and terminal edits with permutation controls through the frozen representation of *every* model, plus parameter-free retrieval and negative controls that bound what the sparse biological rows can support. The audits are model-independent and never involve GLYCOTF (Appendix E). The benchmark therefore has value independent of any new encoder: standard in-distribution scores cannot separate structural faithfulness from shortcut exploitation, because every model that reads composition can exploit it, and only a benchmark that withholds the shortcut makes the difference visible.

We also evaluate GLYCOTRANSPORTFORMER as a representative structure-aware model rather than as the benchmark's evidence base: a rooted-tree encoder in which bond descriptors parameterize orthogonal maps whose products place residue states in a shared reducing-end frame, and attention, bottom-up branch contraction, and terminal summaries combine global and hierarchical context. Same-source controls favor the orthogonal map over an unconstrained dense map on subtype, but additive and commutative variants remain competitive, so we describe the contribution as *linkage-conditioned common-root-frame transport*, a design we test rather than an established explanation for the model's overall performance.

Our contributions are:

- *A common evaluation layer over glycan task resources*: canonical structure keys, normalized-key-disjoint manifests, task-appropriate metrics, validation-only selection, and run-level provenance across fifteen targets.
- *A structural diagnostic suite* that separates sensitivity to topology and linkage from downstream accuracy, together with a bundled root-dependent stress test; it is defined on the benchmark, not on GLYCOTF.
- A representative structure-aware model, GLYCOTRANSPORTFORMER, evaluated under the benchmark: it makes linkage-conditioned common-root-frame transport explicit and combines it with branch and terminal summaries, while its advantage rests on paired composition-disjoint ablations and holds for the structural and taxonomy families rather than across all targets; the two sparse disease and tissue rows are audit cases rather than architecture evidence.

2 RELATED WORK

Glycan representations and benchmarks. Glycan learning has progressed from sequence-like and motif-based systems to graph convolutions, multi-relational encoders, higher-order message passing, and graph Transformers (Bojar et al., 2021; Burkholz et al., 2021; Carpenter et al., 2022; Lundström et al., 2022; Xu et al., 2025a;b; Joeres & Bojar, 2024; Kitani et al., 2026). Earlier tree kernels already identified rooted topology and bond type as useful signals (Yamanishi et al., 2007). Existing suites cover diverse prediction targets, including interaction and NMR tasks (Chen et al., 2023), but do not jointly test normalized-key-disjoint splits, composition shift, and direct structural probes. GlycanML (Xu et al., 2025a) is the closest prior suite and the reference our task families are measured against: it defines a shared prediction benchmark and reports accuracy on fixed structure-level splits. We reuse its families for the eight taxonomy ranks, type, and immunogenicity and claim no new labels there; what changes is the evaluation. The identity unit is declared and audited — the normalized structure key — and the same fifteen targets are re-partitioned under composition, motif-profile/cluster, and exact-isomorphism units, each certified leakage-free at its own level, so an accuracy difference is attributed to a stated form of generalization rather than to one split. Structural diagnostics are measured separately from accuracy, and curation and protocol are traceable, with the reused targets rebuilt from primary records rather than copied (Appendix E).

Glycoinformatics and structural identity. GlycoCT and WURCS encode residues, linkage positions, stereochemistry, and branching, while EUROCarbDB, GlyTouCan, GlyGen, and GlyConnect connect structures to biological context (Herget et al., 2008; Tanaka et al., 2014; Ceroni et al., 2011; Tiemeyer et al., 2017; York et al., 2020; Alocci et al., 2019). These standards motivate canonical structure keys and structure-grouped splits instead of raw string identity, and SNFG fixes the notation (Neelamegham et al., 2019).

Structural encodings and transport. Graph OOD work studies shifts in topology and context (Ying et al., 2021; Gui et al., 2022); grouping glycans by residue counts is such a shift, blocking the direct composition shortcut while retaining linkage and branching as predictive factors. Geometric networks use local frames and transport to compare features across positions (de Haan et al., 2021; Bronstein et al., 2021); our rooted-tree construction uses the reducing end as the common reference and requires no cycle consistency.

3 GLYCANSTRUCTBENCH

3.1 TASK SUITE

As summarized in Figure 2, the benchmark contains three structural-attribute tasks (composition, type, and subtype), eight single-label taxonomy tasks from domain to species, and four biological tasks (disease direction, immunogenicity, disease association, and tissue localization). Composition is a representation sanity check; type and subtype are multi-label structural annotations. Taxonomy uses macro-F1 because rare taxa matter, the large multi-label tasks use micro-F1, and disease association uses micro-average precision (micro-AP) because 170 sparse labels make thresholded F1 unstable. Three task families are reused rather than invented: the eight taxonomy ranks, glycan type, and immunogenicity follow the targets of GlycanML (Xu et al., 2025a), and we claim no new labels there. Everything a reported number depends on is rebuilt: subtype is a separately curated 37-label GlyGen annotation, composition is derived from the parsed structure, disease direction, disease association,

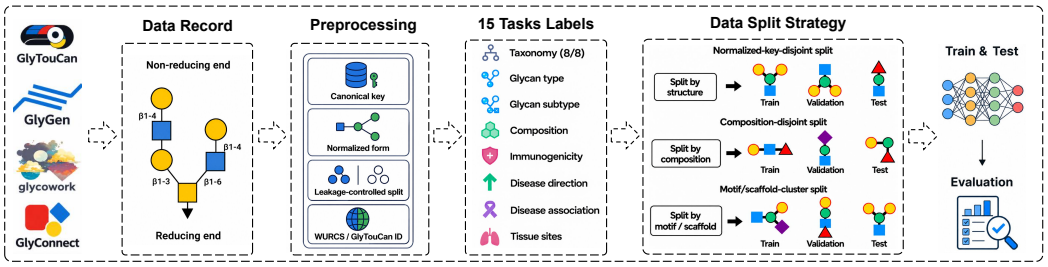


Figure 2: Benchmark construction and evaluation logic. Source records are canonicalized and assigned leakage-controlled keys. Normalized-key-disjoint evaluation prevents duplicate keys crossing splits; composition-disjoint evaluation withholds residue-count groups; the motif-profile/cluster-disjoint split holds out whole clusters; controlled edits test structural sensitivity. The composition and cluster splits test different forms of generalization and are not external-cohort validation.

and tissue localization are annotation-prediction tasks integrated here, and all splits and label maps are refit on the normalized key (Table 12 itemises this reuse against the rebuild; sizes are in Appendix B).

3.2 Normalized-key-disjoint ID evaluation

We normalize parsed IUPAC structures and assign a persistent structure key before splitting, so *structural identity is defined operationally as equality of that key* and we name the ordinary splits *normalized-key-disjoint*. Normalized-key disjointness is not exact isomorphism: isomorphic structures may carry different keys and can therefore cross a split. Grouping the 30,384-structure subtype universe by exact rooted-tree isomorphism gives 16,990 classes, of which 8,042 hold more than one key (70.5% of structures lie in such a class) and 1,805 classes straddle train and test (Appendix D). No key crosses a partition, but the ordinary splits are therefore *not* isomorphism-disjoint, so **the benchmark's identity unit is the normalized structure key, not chemical identity**, and every split, audit and reported number inherits that scope. Taxonomy is class-stratified and closed-set at all eight ranks, checked on the exported splits (Appendix C). The finer ranks are long-tailed, so macro-F1 over the full label set is a within-split diagnostic rather than a balanced estimate. The remaining tasks use a deterministic structure-key allocation, and label maps are fitted on training data.

The benchmark has four explicitly separated levels of evidence, a vocabulary we use throughout: *Level 1, normalized-key leakage control* — the ordinary in-distribution splits, in which no structure key crosses a partition, although isomorphic variants under different keys do; *Level 2, composition shortcut control* — the composition-disjoint splits, which withhold residue-count groups so that direct composition reuse cannot explain a score, and as a by-product keep every same-composition isomer pair inside one partition (Appendix D); *Level 3, motif-profile/cluster-disjoint stress test* — whole motif-profile clusters are held out from one task universe for all fifteen tasks, with the eight taxonomy ranks sharing one split. It is a cluster-disjoint stress test, not a scaffold-novelty guarantee: levels 1–3 each admit isomorphic variants under different keys on opposite sides of a partition, and only Level 4 closes that gap. The partition, its rule, and every task's train/validation/test sizes are recorded in Appendix E.1. *Level 4, exact-isomorphism disjointness* — the strictest unit: whole canonical AHU classes (a complete invariant for rooted labelled trees) are assigned to a partition, so no isomorphic pair and no structure key crosses a boundary (Section 6.4). Cluster-level out-of-distribution claims therefore rest on Level 3 alone, and isomorphism-invariant structural claims on Level 4. All 525 benchmark runs share one feature schema, optimization recipe, and validation-only selection protocol, so the comparison is architectural rather than a claim of a shared code snapshot; a scan of all eight task exports finds no repeated key within a split and zero key-identical overlap across partitions (Appendices R and B). All models receive the same 24-dimensional residue tensor, 20-dimensional bond tensor, bidirectional edge index, and labels; missing linkage positions and anomericity are explicit unknown entries rather than imputed, and unused bond dimensions are zero.

3.3 Positive-unlabelled disease association

Disease association has positive–unlabelled supervision: its 170 labels record curated evidence of association, while an unobserved pair is not a verified negative. The dense training matrix contains only 1.43% positives (test: 1.35%), so plain binary cross-entropy and thresholded F1 are dominated by the negative prior. The ID export also has 55 labels with no test positive and a train–test label-

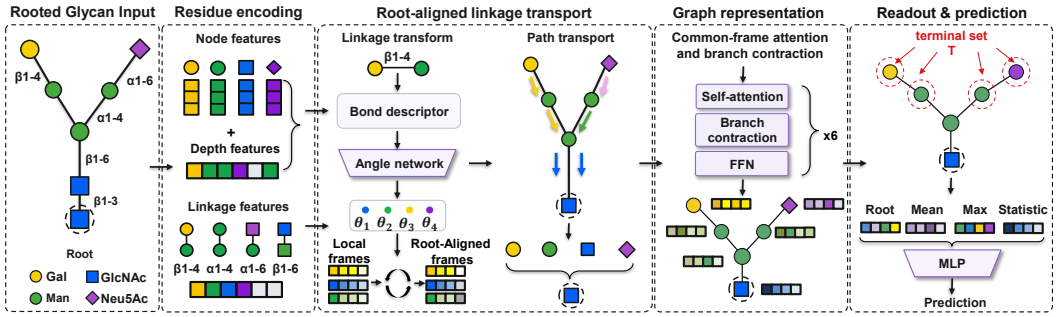


Figure 3: **GlycoTransportFormer.** Rooted linkage-labelled trees with the reducing end as root (1); residue states combine node, depth and incident-bond features (2); an angle network turns each bond descriptor into four bounded angles whose orthogonal Givens product carries residue frames into a shared root-aligned frame along the root-to-node path (3); six common-frame blocks alternate attention, gated branch contraction and a feed-forward network (4); a terminal-aware readout pools *Root* (z_r), *Mean* (μ_T) and *Max* (m_T) over the terminal set T, and *Statistic* $= \mathrm{vec}(C_T)$, the centered terminal cross-covariance of Eq. equation 11 (5).

frequency Jensen–Shannon divergence of 0.0964. We therefore use micro-AP as the primary disease metric and do not interpret it as accuracy, calibration, or recovery of verified negatives.

3.4 Composition-disjoint evaluation

To block direct reuse of observed residue-count vectors, structures are grouped by their canonical residue multiset (the sorted residue tokens of the IUPAC string) and groups, not individual structures, are assigned to train/validation/test, so a test composition is absent from training; the subtype split uses approximately 70/10/20 group allocation and selection remains validation-only. Five seeds are available for the primary subtype linkage and transport comparison; focused readout controls use three. This is a shortcut-controlled split, not an external-cohort test: a held-out composition group may still share motifs, branch patterns, and scaffolds with training structures, which is why the cluster split is reported separately (Table 3). On the subtype task the reported group assignment holds out 418 of the 2,112 composition groups with zero train/test group overlap within a partition. Alternative seeded assignments of the same universe hold out 415–456 groups, and any two of them share only 9–14% of their test groups (pairwise Jaccard 0.091–0.139), so the reported partition is one draw from a family of shortcut-controlled splits.

3.5 Controlled structural diagnostics

We edit anomer, linkage, branch, or terminal-identity fields in up to 224 base structures (220 with a structure-changing branch edit; two further branch edits are isomorphic to their base and act as a serialization control) and run the same probe set through all seven trained models, comparing every frozen representation distance against a node-permutation control. The probe separates them: three models are blind to anomericity, one to linkage, and one graph Transformer is not permutation-invariant (Appendix P). These probes establish feature sensitivity and invariance, not chemical validity or biological utility (Section V).

4 GlycoTransportFormer

4.1 Design objective and problem formulation

Let a glycan be a rooted, edge-labelled tree $G = (V, E, r)$ whose root r is the reducing end. Each residue $v \in V$ has a parsed feature vector $x_v \in \mathbb{R}^{24}$, and every glycosidic bond $(u, v) \in E$ has a descriptor $e_{uv} \in \mathbb{R}^{20}$ containing the available linkage and stereochemical fields. We direct the tree away from r and denote the unique root-to-node path by $\pi(v) = (r = v_0, v_1, \ldots, v_m = v)$. The objective is a graph representation $g(G)$ that is invariant to input node numbering and sibling serialization, yet can distinguish changes in edge labels, their order along a rooted path, and the joint organization of terminal residues. The model factorizes node states into frame channels, generates an orthogonal map for each bond, composes these maps along root-to-node paths, and reads out root and terminal statistics (Figure 3; Appendix R).

4.2 RESIDUE STATES AS FRAME CHANNELS

The residue encoder produces

$$H_v = \mathrm{reshape}_{c\times k}\big[\mathrm{LN}(\mathrm{GELU}(W_x x_v + b_x)) + d_{\delta(v)}\big], \tag{1}$$

where $d_{\delta(v)}$ is a learned depth embedding, $k = 8$, and $c = d/k = 32$ for $d = 256$; one small transport matrix is applied across all channels. Depth records distance from the root, whereas transport depends on the ordered bond sequence.

4.3 BOND-CONDITIONED ORTHOGONAL CONNECTION

For bond (u, v), a two-layer angle network produces four bounded angles,

$$\theta_{uv} = \pi \tanh(W_2\,\mathrm{SiLU}(W_1 e_{uv} + b_1) + b_2) \in [-\pi, \pi]^4. \tag{2}$$

Angle $\theta_{uv,j}$ parameterizes a Givens rotation $G_{a_j b_j}(\theta_{uv,j})$ on coordinate plane (a_j, b_j). We use overlapping planes $(0, 1), (1, 2), (2, 3), (3, 4)$ and define

$$R_{uv} = G_{a_4 b_4}(\theta_{uv,4}) \cdots G_{a_1 b_1}(\theta_{uv,1}) \in \mathrm{SO}(k). \tag{3}$$

Thus $R_{uv}^\top R_{uv} = I$. Overlapping rotations need not commute, which is what lets the construction represent reversed linkage order (Proposition 2). We separate what the experiments support from what the construction merely permits, and let the former carry the narrative: *linkage conditioning* (the bond descriptor determines the transform) and *transport into one shared rooted frame* are the parts our controls test, and we refer to them jointly as *linkage-conditioned common-root-frame transport*; orthogonality is a *parameterisation* that keeps channel norms bounded, and ordered non-commutativity is a *theoretical expressivity* property that our commutative control does not find necessary.

4.4 ROOT-ANCHORED PATH TRANSPORT

For the unique path $\pi(v)$, GLYCOTRANSPORTFORMER composes bond operators in path order,

$$P_r = I, \qquad P_v = P_u R_{uv} \quad \text{for } u = \mathrm{pa}(v), \qquad Z_v = H_v P_v. \tag{4}$$

This shallow-to-deep recursion expresses every Z_v in the reducing-end frame. Orthogonality gives the pre-encoder identity

$$\|Z_{v,j:}\|_2^2 = H_{v,j:} P_v P_v^\top H_{v,j:}^\top = \|H_{v,j:}\|_2^2 \tag{5}$$

for every channel j. This guarantee applies to coordinate alignment; later attention and feed-forward operations remain learned. A root change alters path products, while root-preserving reindexing only permutes $\{Z_v\}$ (Proposition 1).

4.5 COMMON-FRAME ATTENTION AND BRANCH CONTRACTION

After flattening $Z_v \in \mathbb{R}^{32\times 8}$ to $z_v^{(0)} \in \mathbb{R}^{256}$, six transport blocks mix context. For layer ℓ, global attention is

$$A^{(\ell)} = \mathrm{MHA}\Big(\mathrm{LN}(Z^{(\ell)}), \mathrm{LN}(Z^{(\ell)}), \mathrm{LN}(Z^{(\ell)})\Big), \tag{6}$$

$$\widetilde{z}_v^{(\ell)} = z_v^{(\ell)} + A_v^{(\ell)}. \tag{7}$$

Attention therefore mixes distant branches after root alignment (Vaswani et al., 2017). We next contract child states toward their parent. Let $s_v^{(\ell)} = \sum_{u:\mathrm{pa}(u)=v} \widetilde{z}_u^{(\ell)}$. Then

$$\gamma_v^{(\ell)} = \sigma\Big(W_g[\widetilde{z}_v^{(\ell)} \| s_v^{(\ell)}] + b_g\Big), \tag{8}$$

$$\widehat{z}_v^{(\ell)} = \widetilde{z}_v^{(\ell)} + \gamma_v^{(\ell)} \odot \frac{s_v^{(\ell)}}{1 + \delta(v)}, \tag{9}$$

$$z_v^{(\ell+1)} = \widehat{z}_v^{(\ell)} + \mathrm{FFN}(\mathrm{LN}(\widehat{z}_v^{(\ell)})). \tag{10}$$

The child sum is sibling-order invariant; its gate and depth normalization expose the rooted hierarchy without allowing repeated addition alone to dominate.

4.6 BOUNDARY STATISTICS AND PREDICTION

Let T be the terminal set, $\mu_T = |T|^{-1}\sum_{t\in T} z_t$, and $m_T = \max_{t\in T} z_t$. To capture interactions beyond these first-order summaries, bias-free projections $q_t = W_q z_t$ and $k_t = W_k z_t \in \mathbb{R}^{64}$ form

$$C_T = \frac{1}{|T|}\sum_{t\in T}(q_t - \bar{q})(k_t - \bar{k})^\top, \quad \bar{q} = \frac{1}{|T|}\sum_t q_t, \quad \bar{k} = \frac{1}{|T|}\sum_t k_t. \tag{11}$$

The graph descriptor is

$$g(G) = \mathrm{MLP}([z_r \| \mu_T \| m_T \| \mathrm{vec}(C_T)]) \,. \tag{12}$$

Here C_T is invariant to terminal enumeration and vanishes when $|T| = 1$. The parameter-matched control retains W_q, W_k but substitutes $C_T^{\mathrm{mean}} = (W_q\mu_T)(W_k\mu_T)^\top$, isolating centered per-terminal cross-statistics from a bilinear statistic of the mean.

4.7 Formal properties and scope

Appendix U proves reindexing invariance, transport norm preservation, and sensitivity to reversed linkage order, and records complexity and scope for rooted acyclic glycans.

5 Experimental protocol

Models. We compare SimpleGCN, SweetNet (Burkholz et al., 2021), a GlycanAA-inspired residue-level GCN (local baseline; not the atom-level, multi-scale-pretrained GlycanAA model) (Xu et al., 2025b), a multi-relational R-GCN corresponding to the GlycanML family (Xu et al., 2025a), a GlycanGT-style graph Transformer (Kitani et al., 2026), a rooted GlycanTreeFormer (GTF), and GlycoTF. All baselines are local adaptations trained from scratch under one protocol, not official pretrained reproductions of the published systems. Task-head parameter counts and a scoped RTX 5090 throughput and peak-memory microbenchmark for GlycanGT, GTF, and GlycoTF (excluding parsing and cache construction) appear in Appendix S and Table 36. GlycoTF has the lowest peak training memory but also the lowest throughput; because the headline ranks are descriptive rather than capacity-controlled, we flag that gap as a limitation.

Training and selection. Each model–task pair is trained with seeds 42–46, batch size 2048, Adam-style optimization at learning rate 2×10^{-4}, weight decay 10^{-4}, five warm-up epochs, automatic mixed precision, no class weights, and no label smoothing; the checkpoint maximizing the task's primary validation metric is evaluated once on test, and we report the mean and sample SD over five seeds. Because the eight taxonomy ranks share one structure split, the unweighted mean rank is descriptive rather than fifteen independent trials; we therefore also report family-balanced ranks. Two further partitions stress structural shift: a motif-profile/cluster-disjoint split of a $73{,}442$-structure universe that holds out whole clusters for all fifteen tasks, and an exact-isomorphism-disjoint split whose unit is the AHU canonical class, certified here for all fifteen tasks (Table 4). Both are reported under their own partition and do not expand the headline rank. For disease association, a frozen-export audit computes micro-AP, macro-AP, calibration error, and support-stratified AP without tuning thresholds or treating unobserved pairs as verified negatives; because the compact run JSONs hold aggregate metrics, these summaries are retained as audit artifacts.

Ablations and uncertainty. We retrain without linkage descriptors, without path transport, and with commutative and no-cumulant variants as additional structural controls. The subtype composition-disjoint linkage and transport ablations use five paired seeds, as does the completed tissue no-cumulant rerun. The disease mean-outer-product control uses three seeds with 2,000 paired bootstrap resamples over test structures per seed, so its intervals quantify structure-level uncertainty within a seed rather than across datasets. All deltas are variant minus full unless stated otherwise.

Root and terminal interventions. The wrong-root and frozen-terminal interventions are specified in Appendix K.

6 Results and structural diagnostics

6.1 A broad but non-uniform ID advantage

GlycoTF leads the structural and taxonomy families on this split: the paired family differences are $+1.28$ points (structural) and $+8.71$ (taxonomy), each with all five seed-level differences positive, while GTF keeps disease direction and immunogenicity. Aggregate rank summaries (family-balanced 1.25 against 2.00 for GTF, highest mean on 9 of the thirteen) are descriptive, per the family-level analysis predefined in Section 5. No individual task survives Holm correction at five seeds, so Table 21 reports effect sizes and seed-level consistency rather than significance. The ID disease and tissue gaps are not encoder-specific: label frequency, co-occurrence, gold taxonomy and a kNN control all match or exceed the learned rows, so both are audit cases (Appendices M, N) and capacity-paired subtype controls differ by less than the between-seed spread on both splits (Table 14).

Table 1: **In-distribution benchmark.** Five-seed mean ± sample SD in per cent under validation-only selection; bold marks the row maximum. Disease association and tissue localization are audit rows, not architecture evidence. Baselines are *local* adaptations of the published architectures, not official reproductions (Section 5).

Task	SimpleGCN	SweetNet	GlycanAA	R-GCN	GlycanGT	GTF	GLYCOTF
Composition	82.79±6.20	89.18±0.52	89.46±0.27	87.58±0.59	93.57±0.22	94.73±0.09	**95.23±0.05**
Subtype	33.24±0.01	51.75±4.35	59.64±5.94	57.85±7.44	70.57±11.62	93.07±0.64	**96.08±0.10**
Type	56.67±3.41	94.18±0.72	95.66±0.39	96.41±0.57	96.32±0.29	98.82±0.37	**99.15±0.08**
Domain	26.66±10.56	42.53±4.95	45.07±0.17	45.38±0.15	43.34±0.32	**66.09±8.49**	59.31±7.13
Kingdom	21.93±0.35	25.40±0.91	24.95±0.35	25.43±0.23	27.40±5.29	**47.70±4.00**	46.12±6.92
Phylum	6.77±0.73	9.44±0.55	10.19±0.36	11.40±0.62	17.82±0.38	29.79±2.54	**39.34±4.98**
Class	2.04±1.10	7.59±0.68	8.75±0.84	9.77±0.57	18.57±0.58	30.45±1.66	**38.87±2.63**
Order	0.47±0.69	4.35±0.49	5.98±0.44	6.97±0.51	19.94±0.87	28.72±0.41	**40.28±0.76**
Family	0.15±0.04	2.84±0.34	3.50±0.49	4.63±0.12	14.07±0.38	21.29±0.68	**36.56±1.97**
Genus	0.07±0.02	1.70±0.91	2.31±1.04	3.54±0.11	14.21±0.82	17.23±1.33	**34.75±0.51**
Species	0.05±0.01	1.07±0.44	1.51±0.41	1.99±0.21	11.01±0.62	12.44±0.92	**28.17±0.22**
Disease direction	50.19±10.82	63.85±7.93	55.77±12.82	52.93±10.62	48.51±7.06	**74.67±2.45**	64.60±3.81
Immunogenicity	86.58±3.73	93.89±2.99	97.47±1.20	93.82±0.46	93.54±1.22	**99.32±0.71**	98.14±1.10
Disease association (micro-AP)	1.27±0.09	1.52±0.13	1.52±0.15	1.41±0.20	1.59±0.41	1.49±0.15	**23.76±5.22**
Tissue localization	3.29±0.99	2.95±0.50	2.36±0.51	2.93±0.55	2.73±0.66	3.33±0.85	**34.68±0.76**

Table 2: **Composition-disjoint benchmark.** Same protocol, metrics, and formatting as Table 1; the two audit rows are not architecture evidence.

Task	SimpleGCN	SweetNet	GlycanAA	R-GCN	GlycanGT	GTF	GLYCOTF
Composition	82.12±5.39	90.22±0.22	89.73±0.21	87.98±0.36	94.59±0.19	**95.35±0.05**	95.21±0.13
Subtype	32.59±1.06	52.83±5.64	52.55±3.06	57.32±1.73	54.86±16.12	94.34±1.01	**96.72±0.13**
Type	57.95±1.46	94.02±0.77	94.97±0.40	95.04±0.94	96.55±0.25	98.60±0.42	**99.08±0.11**
Domain	36.30±9.89	43.01±0.33	43.64±0.62	43.80±0.54	42.51±0.53	57.93±2.31	**59.44±4.24**
Kingdom	24.46±0.25	26.86±0.25	27.93±0.60	28.18±0.55	39.05±5.18	47.05±0.72	**48.63±5.36**
Phylum	6.93±0.41	11.60±0.78	12.31±0.74	12.85±1.04	22.97±2.30	32.72±2.09	**35.48±1.42**
Class	2.17±0.99	8.59±0.45	10.59±0.72	12.02±0.71	22.63±1.78	32.82±1.26	**33.28±4.00**
Order	0.23±0.07	4.74±0.56	6.27±0.47	7.38±0.51	18.51±0.32	27.03±1.39	**31.10±2.37**
Family	0.15±0.05	3.11±0.23	4.25±0.40	5.10±0.28	16.67±1.12	20.31±1.15	**27.08±0.96**
Genus	0.07±0.02	1.67±0.84	2.61±0.42	2.43±1.31	13.30±0.63	15.68±1.34	**23.19±0.79**
Species	0.05±0.01	0.71±0.79	1.48±0.59	1.76±0.29	11.82±0.71	11.81±0.30	**20.35±0.37**
Disease direction	53.69±11.03	61.79±9.34	63.06±6.58	55.32±11.87	65.88±4.26	**70.38±0.97**	65.64±4.04
Immunogenicity	85.21±1.47	91.77±1.93	93.23±1.28	92.32±0.61	93.05±0.90	95.39±1.20	**96.33±0.62**
Disease association (micro-AP)	1.14±0.15	1.28±0.11	1.30±0.14	1.21±0.18	1.36±0.19	1.26±0.06	**33.11±1.21**
Tissue localization	2.87±0.73	2.95±0.87	2.37±0.40	2.80±0.60	2.37±0.67	3.76±1.20	**40.65±0.98**

6.2 LINKAGE-CONDITIONED TRANSPORT UNDER COMPOSITION-DISJOINT EVALUATION

Across the composition-disjoint matrix the family ordering repeats: the paired family differences are $+0.91$ (structural) and $+4.15$ (taxonomy) points, five of five seeds each, so the same two families lead under composition shift, which is the cross-partition claim; the aggregate counts and ranks — highest mean on $13/15$ rows, family-balanced 1.28 against 2.10 for GTF — remain descriptive. GTF is nominally ahead on composition and disease direction (Appendix I). Removing linkage descriptors or path transport lowers subtype micro-F1 by 2.97 and 2.90 points over five seeds, with all ten paired differences favoring the full model, which supports *linkage-conditioned rooted-path encoding* (Section 4.3, Appendix V). The bundled root intervention changes subtype micro-F1 by -0.21 points ID versus -2.50 under composition shift (Appendix K).

6.3 MOTIF-PROFILE/CLUSTER-DISJOINT STRESS TEST

On the motif-profile/cluster-disjoint partition (whole clusters held out, no cluster crossing) the ranking matches the other two partitions: GLYCOTF has the best mean on subtype and type (96.09, 99.07) while GTF is nominally ahead on composition (97.96 versus 97.75, inside the between-seed spread). The same ordering holds out of cluster across the eight taxonomy ranks, where GLYCOTF leads all eight with absolute macro-F1 far below in-distribution, while both sparse rows remain audit cases — GLYCOTF reaches 17.66 micro-F1 on tissue localization against a parameter-free rooted-linkage nearest-neighbour control at 28.17, and 6.95 micro-AP on disease association against 15.17 (Appendix N).

Task	SimpleGCN	SweetNet	GlycanAA	R-GCN	GlycanGT	GTF	GLYCOTF
Subtype	42.77±0.00	56.63±1.68	57.79±1.80	65.79±3.32	72.10±1.60	94.98±0.85	**96.09±0.66**
Type	75.90±0.00	97.86±0.26	97.89±0.12	97.98±1.17	96.59±0.92	98.51±0.40	**99.07±0.40**
Composition	86.85±6.34	95.59±0.78	95.55±0.23	93.71±0.87	97.39±0.11	**97.96±0.09**	97.75±0.17
Domain	32.13±0.07	32.10±0.00	33.70±5.10	32.09±0.01	42.93±1.04	50.33±3.51	**53.98±2.57**
Kingdom	8.50±0.83	12.07±3.60	14.35±0.22	14.84±0.57	16.64±5.52	24.62±0.43	**28.17±3.40**
Phylum	4.10±0.51	5.90±1.37	6.76±0.27	6.03±1.47	3.84±0.00	12.79±5.07	**15.59±2.48**
Class	2.17±0.26	2.33±0.45	2.45±0.68	2.25±0.23	2.02±0.00	11.94±0.95	**13.44±1.16**
Order	0.96±0.00	2.04±0.30	2.28±1.94	1.32±0.50	6.66±1.46	7.13±3.62	**10.34±1.30**
Family	0.84±0.06	1.16±0.40	1.29±0.26	1.69±1.60	6.46±0.71	6.71±0.71	**9.92±1.69**
Genus	0.65±0.00	0.67±0.03	0.73±0.16	0.66±0.01	3.26±2.46	4.69±2.27	**7.97±0.90**
Species	0.63±0.01	0.70±0.14	0.71±0.18	0.62±0.00	1.57±1.30	2.68±1.95	**6.56±1.62**
Disease direction	50.17±11.85	**65.48±3.04**	61.93±10.97	54.19±12.75	44.88±0.00	59.84±13.80	58.00±6.75
Immunogenicity	47.77±2.89	53.73±10.91	64.18±17.55	70.06±19.34	72.83±2.73	**88.46±5.49**	77.02±8.09
Disease association (micro-AP)	2.42±0.97	1.89±0.17	1.79±0.39	1.79±0.13	1.89±0.23	2.09±0.40	**6.95±2.35**
Tissue localization	3.19±1.23	3.39±0.58	3.13±0.76	3.27±1.35	3.52±2.04	3.34±0.96	**17.66±4.75**

Table 3: **Motif-profile/cluster-disjoint stress test** (whole clusters held out, no cluster crossing; cluster-disjoint, not a scaffold-novelty guarantee). Five-seed mean ± sample SD (%) under validation-only selection; bold marks the row maximum. Disease association uses micro-AP and the other rows the task's primary metric; its row and tissue localization are audit cases, and absolute taxonomy scores are not comparable to Table 1.

Task	SimpleGCN	SweetNet	GlycanAA	R-GCN	GlycanGT	GTF	GLYCOTF
Subtype	33.39±0.79	49.85±2.82	51.71±2.79	59.51±1.25	57.02±15.35	94.35±0.95	**97.29±0.14**
Type	55.49±0.53	93.24±0.57	94.10±0.89	94.67±0.97	96.35±0.74	98.70±0.53	**99.26±0.17**
Composition	82.90±6.01	90.68±0.42	89.98±0.14	88.56±0.18	94.50±0.09	95.20±0.05	**95.72±0.02**
Domain	36.10±9.82	42.83±0.44	43.56±0.54	44.24±0.20	47.15±7.23	**66.95±5.75**	64.85±7.51
Kingdom	21.86±0.37	25.27±0.57	26.25±0.29	26.82±0.20	41.95±0.84	49.62±6.33	**59.57±2.69**
Phylum	7.23±0.95	12.33±0.75	12.73±0.73	13.48±0.80	22.40±5.56	34.08±3.69	**39.44±1.68**
Class	1.85±1.10	8.54±0.74	12.14±0.70	13.76±1.05	27.77±1.43	38.99±3.15	**40.17±2.26**
Order	0.33±0.09	5.54±0.60	7.14±0.47	8.69±0.76	23.76±2.25	32.18±1.67	**36.15±1.02**
Family	0.17±0.06	4.24±0.51	4.84±0.43	5.98±0.52	21.16±1.88	24.80±1.00	**33.71±1.34**
Genus	0.09±0.01	2.58±0.56	4.12±0.34	4.14±0.21	18.35±0.35	19.68±1.26	**31.91±1.05**
Species	0.06±0.02	1.65±0.48	2.59±0.62	2.62±0.14	15.60±0.61	15.85±0.96	**28.51±0.79**
Disease direction	54.75±11.27	61.75±10.41	62.34±10.26	57.44±11.79	65.80±11.96	68.23±7.00	**69.89±2.94**
Immunogenicity	79.35±18.12	90.54±0.45	93.30±1.93	91.66±0.83	92.44±1.04	95.89±1.07	**96.44±0.52**
Disease association	2.57±0.41	3.20±0.92	2.70±0.50	2.75±0.24	2.92±0.63	3.34±0.83	**31.86±0.98**
Tissue localization	2.92±0.86	2.87±0.91	2.27±0.35	2.68±0.48	2.36±0.77	3.62±1.75	**44.76±1.00**

Table 4: **Exact-isomorphism-disjoint partition (Level 4).** Fifteen units, 525 runs: whole canonical AHU classes are the split unit, so no isomorphic pair and no structure key crosses a partition. Five-seed mean ± sample SD (%), validation-only selection, bold marks the row maximum; the disease-association and tissue-localization rows are audit cases.

6.4 EXACT-ISOMORPHISM-DISJOINT PARTITION

Here the split unit is the exact rooted-tree isomorphism class (canonical AHU form), so no isomorphic pair and no structure key crosses a partition. GLYCOTF keeps the best mean on all three structural tasks (subtype 97.29, type 99.26, composition 95.72; Table 4) and, across all fifteen units of the same matrix, on fourteen of them, the domain rank the single exception (64.85 against 66.95 for GTF); rank agreement with the other partitions stays high (Appendix F.3).

7 CONCLUSION

GLYCANSTRUCTBENCH reports every number at the level its split supports, under a declared identity unit rather than an assumed one. On the composition-disjoint partition GLYCOTF leads the structural and taxonomy families with all five paired seed differences positive, and that ordering survives both the cluster-disjoint stress test and the strictest unit, where it holds the best five-seed mean on fourteen of the fifteen tasks. The diagnostic layer separates structural learning from shortcut exploitation: controlled edits and permutation controls show which distinctions each model preserves, and the retrieval control shows that the two sparse biological rows cannot support an architecture claim. The advantage is family-specific, no single task survives Holm correction at five seeds, and external transfer, matched capacity and chemically validated targets remain open (Appendices I and V).

Reproducibility statement

All headline comparisons use validation-only checkpoint selection, a fixed benchmark registry (SHA-256 `e8e267bc39f2aa4fa4f8e2cb7b9f1bb900a9af8e49756bb7279e00079dc09a8a`), registered split manifests, and five seeded runs per cell, with one test evaluation after selection. The anonymous review repository is `https://anonymous.4open.science/r/GlycanStructBench-AB18`; it provides the processed manifests, model and feature schemas, run-level provenance, machine-readable metrics, integrity audits, and the scripts that regenerate every table and figure, together with the release index that maps each reported number to what produced it. Optimization settings, task sizes and splits, parameter counts, and the metric choice for positive–unlabelled disease association are in Appendices R, B, S and Section 3.3. The identity unit is the normalized structure key (Section 3.2), and the four levels of evidence are reported separately rather than pooled.

Ethics statement

This work uses publicly available glycan records and curated annotation tables; it does not collect data from human or animal subjects and does not perform clinical intervention. Disease-association labels are positive–unlabelled evidence rather than verified absence labels, so the reported disease scores should not be interpreted as clinical risk estimates or causal biological claims. The benchmark is intended for representation evaluation and method development; any downstream biomedical use requires independent validation, appropriate consent and governance, and domain-expert review.

AI use statement

Generative AI tools were used as research and engineering assistants for: *implementation* (data pipeline, training, evaluation, analysis and export code); *dataset cleaning* (deduplication, normalisation, split construction, and the integrity audits that check them); *hypothesis generation and method feedback* (baselines, ablations, controls, and critique of the design); *theory and mathematical claims* (drafting and checking the transport propositions and proofs); *result interpretation* (paired statistics and drafts of the outcomes); *conceptual framing* of the levels of evidence; and language editing, LaTeX, figure code, and consistency checking. No manuscript content was translated, and no reported number was produced by a generative model: every value is emitted by a deterministic export script from stored per-run payloads, and an automated claim audit recomputes each quoted value against those artifacts. The tools were not used to fabricate data or measurements, invent citations, or alter experimental results. The human authors directed the work, selected the methods and claims, and remain responsible for the manuscript. AI-assisted illustrations are conceptual diagrams and contain no measured data.

References

Davide Alocci, Julien Mariethoz, Alessandra Gastaldello, Elisabeth Gasteiger, Niclas G. Karlsson, Daniel Kolarich, Nicolle H. Packer, and Frédérique Lisacek. Glyconnect: Glycoproteomics goes visual, interactive, and analytical. *Journal of Proteome Research*, 18(2):664–677, 2019. doi: 10.1021/acs.jproteome.8b00766.

Kiyoko F. Aoki-Kinoshita, Sanjay Agravat, Nobuyuki P. Aoki, et al. Glytoucan 1.0—the international glycan structure repository. *Nucleic Acids Research*, 44(D1):D1237–D1242, 2016. doi: 10.1093/nar/gkv1041.

Daniel Bojar, Riley K. Powers, Diogo M. Camacho, and James J. Collins. Deep-learning resources for studying glycan-mediated host–microbe interactions. *Cell Host & Microbe*, 29(1):132–144.e3, 2021. doi: 10.1016/j.chom.2020.10.004.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Rebekka Burkholz, John Quackenbush, and Daniel Bojar. Using graph convolutional neural networks to learn a representation for glycans. *Cell Reports*, 35(11):109251, 2021. doi: 10.1016/j.celrep.2021.109251.

Eric J. Carpenter, Sandeep Seth, Nirajan Nepal, et al. Glynet: a multi-task neural network for predicting protein–glycan interactions. *Chemical Science*, 13:6871–6886, 2022. doi: 10.1039/D1SC05681F.

Alessio Ceroni, Anne Dell, and Stuart M. Haslam. EUROCarbDB: An open-access platform for glycoinformatics. *Glycobiology*, 21(4):493–502, 2011. doi: 10.1093/glycob/cwq188.

Zizhang Chen, Ryan Paul Badman, Lachele Foley, Robert J. Woods, and Pengyu Hong. Glyconmr: Dataset and benchmarks for nmr chemical shift prediction of carbohydrates with graph neural networks. *arXiv preprint arXiv:2311.17134*, 2023.

Richard D. Cummings and J. Michael Pierce. The challenge and promise of glycomics. *Chemistry & Biology*, 21(1):1–15, 2014. doi: 10.1016/j.chembiol.2014.01.013.

Pim de Haan, Maurice Weiler, Taco Cohen, and Max Welling. Gauge equivariant mesh cnns: Anisotropic convolutions on geometric graphs. In *International Conference on Learning Representations*, 2021.

Sally Esmail and Morris F. Manolson. Advances in understanding n-glycosylation structure, function, and regulation in health and disease. *European Journal of Cell Biology*, 100(7–8):151186, 2021. doi: 10.1016/j.ejcb.2021.151186.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Stefan Herget, René Ranzinger, Kai Maass, and Claus-Wilhelm von der Lieth. GlycoCT—a unifying sequence format for carbohydrates. *Carbohydrate Research*, 343(12):2162–2171, 2008. doi: 10.1016/j.carres.2008.03.011.

Roman Joeres and Daniel Bojar. Higher-order message passing for glycan representation learning. *arXiv preprint arXiv:2409.13467*, 2024.

Akihiro Kitani, Bingyuan Zhang, Koichi Himori, and Yusuke Matsui. Glycangt: A pretrained graph transformer framework for glycan graph representation and generative learning. *Bioinformatics*, 42(4), 2026. doi: 10.1093/bioinformatics/btag147.

Jon Lundström, Emma Korhonen, Frédérique Lisacek, and Daniel Bojar. Lectinoracle: A generalizable deep learning model for lectin–glycan binding prediction. *Advanced Science*, 9(1):2103807, 2022. doi: 10.1002/advs.202103807.

Julien Mariethoz, Khaled Khatib, Davide Alocci, Matthew P. Campbell, Niclas G. Karlsson, Nicolle H. Packer, Elaine H. Mullen, and Frederique Lisacek. Sugarbinddb, a resource of glycan-mediated host-pathogen interactions. *Nucleic Acids Research*, 44(D1):D1243–D1250, 2016. doi: 10.1093/nar/gkv1247.

Kelley W. Moremen, Michael Tiemeyer, and Alison V. Nairn. Vertebrate protein glycosylation: diversity, synthesis and function. *Nature Reviews Molecular Cell Biology*, 13(7):448–462, 2012. doi: 10.1038/nrm3383.

Sriram Neelamegham, Kiyoko Aoki-Kinoshita, Evan Bolton, et al. Updates to the symbol nomenclature for glycans guidelines. *Glycobiology*, 29(9):620–624, 2019. doi: 10.1093/glycob/cwz045.

Ezgi Özcan and David A. Sela. Inefficient metabolism of the human milk oligosaccharides lacto-n-tetraose and lacto-n-neotetraose shifts *Bifidobacterium longum* subsp. *infantis* physiology. *Frontiers in Nutrition*, 5:46, 2018. doi: 10.3389/fnut.2018.00046.

Colin Reily, Tyler J. Stewart, Matthew B. Renfrow, and Jan Novak. Glycosylation in health and disease. *Nature Reviews Nephrology*, 15(6):346–366, 2019. doi: 10.1038/s41581-019-0129-4.

Katrine T.-B. G. Schjoldager, Yoshiki Narimatsu, Hiren J. Joshi, and Henrik Clausen. Global view of human protein glycosylation pathways and functions. *Nature Reviews Molecular Cell Biology*, 21 (12):729–749, 2020. doi: 10.1038/s41580-020-00294-x.

Kenichi Tanaka, Kiyoko F. Aoki-Kinoshita, Masaaki Kotera, Hiromichi Sawaki, Shinichiro Tsuchiya, Noriaki Fujita, Toshihide Shikanai, Masaki Kato, Shin Kawano, Issaku Yamada, and Hisashi Narimatsu. Wurcs: The web3 unique representation of carbohydrate structures. *Journal of Chemical Information and Modeling*, 54(6):1558–1566, 2014. doi: 10.1021/ci400571e.

Luc Thomès, Rebekka Burkholz, and Daniel Bojar. Glycowork: A python package for glycan data science and machine learning. *Glycobiology*, 31(10):1240–1244, 2021. doi: 10.1093/glycob/cwab067.

Michael Tiemeyer, Kazuhiro Aoki, James Paulson, et al. GlyTouCan: an accessible glycan structure repository. *Glycobiology*, 27(10):915–919, 2017. doi: 10.1093/glycob/cwx066.

Ajit Varki. Biological roles of glycans. *Glycobiology*, 27(1):3–49, 2017. doi: 10.1093/glycob/cww086.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Minghao Xu, Yunteng Geng, Yihang Zhang, Ling Yang, Jian Tang, and Wentao Zhang. Glycanml: A multi-task and multi-structure benchmark for glycan machine learning. In *International Conference on Learning Representations*, 2025a.

Minghao Xu, Jiaze Song, Keming Wu, Xiangxin Zhou, Bin Cui, and Wentao Zhang. Modeling all-atom glycan structures via hierarchical message passing and multi-scale pre-training. *arXiv preprint arXiv:2506.01376*, 2025b.

Yoshihiro Yamanishi, Francis R. Bach, and Jean-Philippe Vert. Glycan classification with tree kernels. *Bioinformatics*, 23(10):1211–1216, 2007. doi: 10.1093/bioinformatics/btm090.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, 2021.

William S. York, Raja Mazumder, Rene Ranzinger, Nathan Edwards, Robel Kahsay, Kiyoko F. Aoki-Kinoshita, Matthew P. Campbell, Richard D. Cummings, Ten Feizi, Maria Martin, et al. Glygen: Computational and informatics resources for glycoscience. *Glycobiology*, 30(2):72–73, 2020. doi: 10.1093/glycob/cwz080.

Julie Y. Zhou and Brian A. Cobb. Glycans in immunologic health and disease. *Annual Review of Immunology*, 39:511–536, 2021. doi: 10.1146/annurev-immunol-101819-074237.

A Reading the result tables and units

How this appendix is organised. Sections B–E define the benchmark: task sizes and label counts for both protocol partitions, the split-integrity scan, the structural-identity audits, and the per-task provenance and release registration. Sections F–I hold result detail behind the main-text scorecards, and Sections J–Q hold the controls, ablations, and diagnostics, grouped by what they control for. Sections R–U give the protocol, cost, and formal statements, and Section V closes with the limitations. The reproducibility, ethics, and AI use statements required by the submission template are placed at the end of the main text, before the references, and they name the release index that maps every reported number to what produced it.

Unit convention for this appendix: numeric appendix tables report metric values as fractions of 1 (so a subtype micro-F1 of 0.9609 is 96.09%), whereas the main-text scorecard tables and every difference quoted in the main text are percentages and percentage points; a caption that uses another unit says so. The main-text Tables 1 and 2 carry the full seven-model matrix for both splits, so this appendix does not repeat it. Values are percentages and mean$\pm$sample SD over five seeds. The metric is micro-F1 for composition, subtype, type, and tissue localization, micro-AP for disease association, and macro-F1 for the taxonomy ranks and the two remaining biological rows. In every result table, "GlycanAA" and "GlycanGT" denote the local residue-level and graph-Transformer adaptations of Section 5; both are trained from scratch without official pretrained weights, and those caveats apply to all appendix scorecards.

Table 5: In-distribution (ID) task sizes in structures, as used by the 525 ID runs, with the label-space size after the train-only label map. Taxonomy contributes eight tasks that share one structure split and differ only in label space, so it appears once with the range over the eight ranks. The composition-disjoint counterpart of this table is Table 13.

Task	Train	Validation	Test	Labels
Taxonomy (8 ranks)	11,808–12,198	1,525–1,668	1,525–1,668	5–1,399
Subtype	24,258	3,036	3,090	37
Type	27,022	3,378	3,383	7
Composition	34,578	4,329	4,424	16
Disease direction	370	46	47	2
Immunogenicity	958	119	121	2
Disease association	913	110	141	170
Tissue localization	5,165	652	711	174

Table 6: Sizes for the audited composition-disjoint exports. The subtype and disease controls use distinct exported snapshots; the complete 15-task OOD scorecard in Table 2 is reconstructed from the corresponding per-run payloads.

Task	Train	Validation	Test
Subtype (audited replication export)	20,911	3,102	6,285
Disease association (audited replication export)	790	132	240

B DATASET STATISTICS

Disease direction is retained as an exploratory task; its small labeled support and high between-seed variance motivate the cautious interpretation in the main text. Disease association is additionally positive–unlabelled (Section 3.3): among the 141×170 test entries only 1.35% are positive, and 55 of the 170 labels have no test positive at all. Label counts for these audited exports are unchanged by the shift because the label space is fitted on training data and held fixed.

Table 7: Fresh integrity scan of the registered ID exports (one record per structure key per split; sizes in Table 5). Key-identical overlap must be zero for a normalized-key-disjoint protocol; composition overlap separates that protocol from the composition-disjoint one, and JSD is the Jensen–Shannon divergence between train and test label frequencies. The taxonomy row is a registry export, not the rank-specific split, so its JSD range describes that export only.

Task	Key-identical overlap	Composition overlap	Label JSD
Composition	0	2,121	0.0002
Type	0	1,387	< 0.0001
Subtype	0	1,265	0.0002
Taxonomy registry export	0	348	0.007–1.0000
Disease direction	0	27	< 0.0001
Immunogenicity	0	63	< 0.0001
Disease association	0	76	0.0964
Tissue localization	0	347	0.0107

The audit found no duplicate structure record within any split and no key-identical overlap for any train–validation, train–test, or validation–test pair. The nonzero composition counts above are expected for the ordinary ID protocol and motivate the separate composition-disjoint evaluation.
The taxonomy row is a *registry export* and is a different artifact from the rank-specific taxonomy tasks of Table 5. The export holds $13{,}169/951/1{,}128$ records, whereas the per-rank tasks hold $11{,}808$–$12{,}198$ / $1{,}525$–$1{,}668$ / $1{,}525$–$1{,}668$: a record enters a rank-specific task only when it carries a label at that rank, so ranks with more missing labels yield smaller tasks, while the export is a single unfiltered partition. The JSD range in Table 7 therefore describes the export rather than the training split. Within that export the species partition places train and test in *disjoint class-index supports*, and Jensen–Shannon divergence equals its upper bound of 1.0000 exactly when the two supports are disjoint; the value sits at 1.0000 rather than merely approaching it. This is an indexing property of the export, not evidence of genuine label shift, and we do not use it to characterize the taxonomy tasks, whose per-rank partitions contain no test-only classes. Taxonomy is consequently read as a long-tailed within-split diagnostic, and we report family-balanced summaries rather than treating the eight ranks as independent replicates.

C SPLIT AND STRUCTURAL AUDIT

The final read-only audit of the optimized subtype release found 24,258, 3,036, and 3,090 graphs in train, validation, and test, respectively. Each split contains unique structure keys, with zero train–validation, train–test, or validation–test key-identical intersections and no non-finite graph tensors. Coarse scaffold signatures nevertheless overlap across the partitions (1,940 train–validation, 2,011 train–test, and 777 validation–test signatures). This is why the composition-disjoint benchmark is described as shortcut-controlled rather than scaffold-disjoint; the motif-profile diagnostic is reported separately and is not folded into the headline rank.

The audit manifest, registry hash, split counts, structure-key hashes, and the unavailable-pretraining decision are retained with the machine-readable run artifacts. These checks constrain interpretation but do not certify chemical validity of every parsed edit.

A read-only inventory of the active result store parsed 318 finalized run payloads without malformed JSON or non-finite values. It also confirmed a single registry hash within the audited controls and one source hash within each five-seed orthogonal, dense, and additive transport group. This inventory checks the available working copy; it does not replace the registered 525-run benchmark matrix.

D STRUCTURAL IDENTITY: CANONICALIZATION AND ISOMER-PAIR AUDITS

What the structure key does and does not collapse. Table 8 probes the operational identity used by every split. Over $3{,}000$ structures the parser is deterministic ($3{,}000/3{,}000$ re-parses agree) and the graph signature is a genuine isomorphism invariant ($3{,}000/3{,}000$ invariant under node relabelling), with no parse failures. It is *not* a full canonical form: reversing the order of sibling branch groups in the IUPAC string reproduces the same key in only $541/1{,}469$ multi-branch cases (36.8%), so alternative serializations of one molecule can be entered as distinct structures. A WURCS/GlycoCT round-trip check would test this more directly but was not executable in the build environment, and we do not claim it.

Table 8: Canonicalization probes over $3{,}000$ corpus structures. The key is deterministic and node-relabelling invariant, but it is not a chemical canonical form.

Probe	Result	Reading
Re-parse determinism	$3{,}000/3{,}000$	stable
Node-relabelling invariance	$3{,}000/3{,}000$	invariant
Parse failures	$0/3{,}000$	clean
Sibling-branch-order collapse	$541/1{,}469$ (36.8%)	**incomplete**
WURCS/GlycoCT round-trip	not available	untested

Canonical grouping: how far the key is from chemical identity. Table 9 answers the sharper question directly. Because the benchmark's graphs are rooted, edge-labelled trees, isomorphism for them is the root-preserving labelled-tree isomorphism of Proposition 1, and its canonical form (AHU) is a *complete* invariant — so grouping by it gives exact isomorphism classes, not a Weisfeiler–Lehman heuristic. The split unit is coarser than the input unit: $8{,}042$ of $16{,}990$ classes contain more than one normalized key, and 70.5% of structures sit in such a class, meaning the same input graph reaches the model under several keys. The largest class holds 16 distinct keys, e.g. `?-Man(?1-?)?-Man(?1-` and `?-Man?(?1-?)?-Man?(?1-`; the common causes are sibling-branch order and dangling reducing-end linkage fragments. This is a statement about the benchmark's own featurization, which is the only object a model sees, and not a claim that the corpus holds duplicate molecules.

Table 9: Canonical grouping of the $30{,}384$-structure subtype universe by exact rooted-tree isomorphism class, per partition. The structure set is shared, so the class statistics are identical across partitions; only the crossing counts differ.

Partition	Classes	Classes with >1 key	Classes crossing train/test
Registered ID	$16{,}990$	$8{,}042$	$\mathbf{1{,}805}$ (5,207 structures)
Composition-disjoint	$16{,}990$	$8{,}042$	$\mathbf{0}$
v5 motif-profile	$16{,}990$	$8{,}042$	20 (90 structures)

Same-composition isomer pairs from real records. Because the key is not a canonical form, splits disjoint in the key are not necessarily disjoint in chemistry. We quantify that exposure with a pair benchmark built entirely from existing database records — no synthetic editing: two records form a *minimal isomer pair* when they share a composition (the multiset of residue tokens), carry

different structure keys, and their multisets of (anomer, position) linkage tokens differ in exactly one token. Pairs in which either member has an unknown linkage position are counted separately, since a multiset difference is then chemically uninterpretable. Table 10 reports the result.

Table 10: Same-composition minimal isomer pairs over the subtype universe (30,384 structures). "Known positions" restricts both members to fully specified linkage tokens. Nothing here is synthetic: every pair joins two existing database records.

Partition	Minimal pairs	Known positions	Straddle train/test
Registered ID	153,038	145,302	24,300
Composition-disjoint	153,038	145,302	**0**
v5 motif-profile	153,038	145,302	17,545

Two readings follow. First, under the ordinary in-distribution splits a model can be trained on one member of a real isomer pair and tested on its near-identical partner in 24,300 cases; a high score there therefore does not by itself demonstrate fine structural discrimination, which is exactly why we report the controlled structural probes of Section P alongside accuracy. Second, the composition-disjoint protocol keeps *all* pairs inside a single partition, because it groups by composition before splitting: withholding residue-count groups also removes this memorisation route. That property was not a design goal and is reported as an observed benefit, not as a claim that composition-disjoint evaluation tests structure. The motif-profile partition reduces the exposure to 17,545 pairs but does not remove it, which is a further reason to treat its result as one diagnostic rather than as broad structural OOD.

Table 11: Examples of fully specified minimal isomer pairs straddling the registered ID train/test boundary.

Train member	Test member
`Gal(b1-2)Fuc`	`Gal(b1-4)Fuc`
`GalNAc(a1-3)Gal`	`GalNAc(a1-6)Gal`
`Gal(b1-3)GlcNAc`	`Gal(b1-4)GlcNAc`

E TASK PROVENANCE AND BENCHMARK CONTRIBUTION

Table 12 is the itemised comparison with the prior task suites: per task it records what already existed and what this release rebuilt. "Prior" means that a closely corresponding prediction target appears in GlycanML or the source toolkit; it does not imply identical records, normalization, label maps, or splits. Read as a contribution statement, the table separates four groups. *(i)* For the eight taxonomy ranks, type, and immunogenicity the prediction targets are prior work, so no new label is claimed: what is new is the normalized-key identity unit, the four-level partition hierarchy built on it, and the audits that certify each partition. *(ii)* Subtype is a separately curated 37-label GlyGen structural annotation, so its target and its split are both new here. *(iii)* Composition is derived from the parsed structure, and disease direction, disease association, and tissue localization are annotation-prediction tasks integrated here; the last three would not otherwise appear in a glycan representation benchmark, and two of them — disease association and tissue localization — are the rows on which a parameter-free retrieval control beats the trained models, so they are reported as audit cases rather than as architecture evidence while disease direction is retained as exploratory. *(iv)* For all fifteen tasks the records are re-parsed, re-keyed, re-split, and re-labelled from the primary sources, so no number in this paper is a re-run of an existing split. The benchmark's principal innovation is therefore the identity-and-split design together with the structure-faithfulness diagnostics, with data integration as the enabling step rather than the claim: a user who already has the GlycanML targets gains the ability to state *which* generalization a number supports — key-disjoint, composition-disjoint, cluster-disjoint, or isomorphism-disjoint — and to test structural fidelity directly instead of inferring it from task accuracy.

GlyGen and GlyConnect distribute data under CC BY 4.0, while GlyTouCan requests attribution and some incorporated resources retain their own terms. The release preserves source identifiers and does not relicense upstream records. GlyConnect is cited to its database release (Alocci et al., 2019); SugarBindDB is cited to its database paper (Mariethoz et al., 2016); SugarBase denotes the versioned glycowork snapshot recorded in Table 12 (Thomès et al., 2021).

What "experimentally specified" means here, and what it does not. The structural distinctions the benchmark tests — linkage position and anomericity, branch order, root orientation, residue composition, and terminal organization — are read from the deposited source entries under the

normalized parsing described in Appendix C; they are not inferred, predicted, or scored by any model. They are therefore specified by the curated records rather than by our architecture. This is weaker than a validated chemical ground truth: the records are curated annotations, some entries are deposited from computational or literature curation rather than a fresh experimental determination, and the downstream taxonomic and biological label spaces are annotation targets whose quality we audit separately (Appendices M, N, and O) rather than assume. The claim the paper makes is about whether a representation preserves structural distinctions that the records specify; it is not a claim that those records are complete, unbiased, or experimentally confirmed in every entry.

Table 12: Task-level provenance: what existed before against what this release rebuilt, task by task. The eight taxonomy ranks share one entry. All tasks use the normalized structure key as split unit and a train-only label map. "Prior target" marks an existing prediction target; "Added here" records what was rebuilt, recurated, or newly defined. The four groups these rows fall into, and what each gives a user, are set out in the text.

Task(s)	Primary source	Prior target	Added here	Metric	Data terms
Taxonomy (8 ranks)	GlyGen/GlyTouCan	Yes	normalized closed-set split, per rank	macro-F1	CC BY / attribution
Subtype	GlyGen	Yes	deduplication, train-only labels	micro-F1	CC BY 4.0
Type	GlyGen	Yes	deduplication, train-only labels	micro-F1	CC BY 4.0
Composition	normalized IUPAC	No	derived residue-count labels	micro-F1	follows source
Disease direction	SugarBase evidence	No	direction ontology, filtering	macro-F1	snapshot terms
Immunogenicity	SweetOrigins/glycowork	Yes	IUPAC remap, structure split	macro-F1	snapshot terms
Disease association	SugarBase/GlyConnect/SugarBindDB	No	evidence merge, ID alignment	micro-AP	preserved per source
Tissue localization	GlyConnect/GlyGen	No	tissue harmonization	micro-F1	CC BY 4.0

Across all rows, the common protocol contribution is a persistent structure-key split, training-only vocabulary construction, validation-only model selection, and run-level provenance.

E.1 Benchmark release: registration and split sizes

Registration and entry points. The release is registered under benchmark version `glycobench-iclr2027-v1` with a registry hash that every run record carries, so a result can be traced back to the release that produced it; the hash itself is stated in the reproducibility statement. Splits are stored as JSONL records in the registered task directories, the tensor files the runner consumes are built from those records by one registered builder, and every task is evaluated through the same runner. Per-task sources, data terms, and metrics are in Table 12, the in-distribution sizes in Table 5, and Table 13 gives the sizes of the composition-disjoint partition used by the out-of-distribution scorecards.

Table 13: Composition-disjoint (motif-profile) partition: task sizes in structures, train / validation / test. This is the partition behind Table 3, separate from the in-distribution table (Table 5) and from the audited replication exports (Table 6); its rule is recorded in the release. The taxonomy ranks share one split, so they appear once; their scored test sets are smaller than the declared split where the train-only label map drops test records.

Task(s)	Train	Validation	Test
Taxonomy (8 ranks, one split)	6,099	591	1,806 (1,703–1,806 scored)
Subtype	19,523	3,043	7,818
Type	22,252	3,170	8,361
Composition	30,521	3,686	9,124
Disease direction	199	103	156
Immunogenicity	340	80	114
Disease association	520	146	360
Tissue localization	2,715	1,339	2,388

Which split backs Table 3. Those rows come from one partition and one partition only: a task-universe clustering whose universe is the union of the structures of the three structural tasks (73,442 structures). k=20 motif-profile clusters are formed on motif-composition features, whole clusters are held out, and the target fractions (20% test, 10% validation) were fixed before any model was trained on the partition. The rule reads structural coverage only, never a task label. Two earlier motif exports are used nowhere in this paper and must not be confused with this partition: in one, the held-out cluster contains a single subtype structure, and in the other the subtype and composition test partitions hold 80 and 5 structures. Table 3 is derived exclusively from the partition described here. Because disease association is positive–unlabelled, its primary metric is re-derived from each run's own per-example scores rather than read from a thresholded summary, so a result produced in another environment can be rescored for this task without our training stack; that recomputation first

reproduces the published in-distribution value (23.76) and only then reports this partition's numbers. The payloads of this partition record no micro-AP field, which is why the row is recomputed rather than read, exactly as for the other two partitions.

F SUPPLEMENTARY EXPERIMENT SCORECARD

Table 14 collects completed supplementary experiments. These analyses are reported separately from the fifteen-task headline scorecard; the detailed task-specific tables and seed-level controls follow below.

Table 14: Supplementary experiment scorecard. Values are test means ± sample SD over five seeds unless stated otherwise. All checkpoints use validation-only selection. Rows marked "diagnostic" report thresholded F1 from an export that carries no ranking scores; the primary metric for disease association is micro-AP (Section 3.3), which that baseline export does not provide.

Experiment	Split/task	Variant	Metric	Result	Seeds
Transport control	Composition-disjoint subtype	Orthogonal transport	micro-F1	0.9695 ± 0.0026	5
Transport control	Composition-disjoint subtype	Dense transport	micro-F1	0.9587 ± 0.0074	5
Transport control	Composition-disjoint subtype	Additive transport	micro-F1	0.9697 ± 0.0010	5
Transport control	Composition-disjoint subtype	Commutative transport	micro-F1	0.9689 ± 0.0036	5
Capacity control	Composition-disjoint subtype	Matched six-layer GTF	micro-F1	0.9462 ± 0.0068	5
Capacity control	ID subtype	Matched GTF-large	micro-F1	0.9677 ± 0.0010	5
Capacity control	ID subtype	Matched GlycoTF-small	micro-F1	0.9646 ± 0.0011	5
Capacity control	Composition-disjoint subtype	Matched GTF-large	micro-F1	0.9688 ± 0.0018	5
Capacity control	Composition-disjoint subtype	Matched GlycoTF-small	micro-F1	0.9712 ± 0.0017	5
Motif-profile/cluster-disjoint stress test	Motif-profile subtype and composition	GLYCOTF, GTF	micro-F1	see Table 3	5
Weak baseline	ID disease association	Fingerprint MLP	micro-F1 (diagnostic)	0.0309 ± 0.0075	5
Weak baseline	ID tissue localization	Fingerprint MLP	micro-F1	0.0342 ± 0.0134	5
Efficiency profile	Isolated RTX 5090 microbenchmark	GTF/GlycoTF/GlycanGT	time, memory	18 measurements	Complete

The per-example probabilities behind the probability and PU analyses are retained in dedicated frozen-export artefacts rather than embedded in every run payload; the compact run records keep the aggregate metrics only.

The local pretraining compatibility audit found no official pretrained checkpoint or canonical loader metadata for any of the seven compared models. Task-trained local checkpoints were therefore not relabelled as pretrained weights, and no pretrained-SOTA claim is made.

F.1 TASK-UNIVERSE MOTIF-PROFILE/CLUSTER SCORECARD

The three-panel task-universe scorecard is Table 3 in the main text: the three structural tasks, the eight taxonomy ranks, and the four biological rows on one partition. This appendix subsection retains what that table does not carry — the partition construction (whole motif-profile clusters held out, no cluster crossing) and the declared split sizes (Table 13).

F.2 CLASS-SAFE MOTIF-PROFILE/CLUSTER SCORECARD (V6)

The v5 partition above clusters individual structure keys, which leaves 20 isomorphism classes crossing train|test. Its successor makes the split unit the exact rooted-tree isomorphism class: one motif profile per class is clustered ($k = 20$, seed 42, rule fixed before training and reading no task label) and whole classes are assigned to a partition, so no isomorphic pair and no structure key can cross a boundary. On the 73,442-structure universe this gives 52,801 classes, of which 11,450 hold more than one structure key, and the builder records **zero** classes and zero keys crossing train|validation, train|test or validation|test. The test split holds 9,473 subtype, 10,027 type and 10,970 composition structures; all 105 runs used the registered recipe, and the scorecard in Table 15 is generated from the per-seed payloads. The ordering is the one the other partitions give — GLYCOTF leads subtype (93.37 against 82.59 for GTF, whose between-seed SD is 19.73) and type, while GlycanGT leads composition — so the Level-3 conclusion does not rest on how the v5 clusters were drawn.

Task	SimpleGCN	SweetNet	GlycanAA	R-GCN	GlycanGT	GTF	GLYCOTF
Subtype	40.33±5.21	51.25±5.13	45.21±9.31	58.97±3.77	65.43±8.12	82.59±19.73	**93.37±0.41**
Type	65.25±0.00	97.88±0.30	97.69±0.37	97.08±0.82	96.85±0.27	98.12±0.46	**98.58±0.85**
Composition	84.79±4.80	96.47±0.24	96.27±0.07	95.88±0.24	97.58±0.14	**98.27±0.09**	97.93±0.17

Table 15: Class-safe motif-profile/cluster-disjoint scorecard (v6): whole exact-isomorphism classes are assigned to a partition, so no isomorphic pair and no structure key crosses one. Five-seed mean±SD (%); best per row in bold.

F.3 EXACT-ISOMORPHISM-DISJOINT SCORECARD

The strictest partition in this work makes the split unit the *exact rooted-tree isomorphism class*. Candidate structures are grouped by their canonical AHU form — a complete invariant for rooted labelled trees, so two structures share a class exactly when they are isomorphic — and whole classes are assigned to a partition. No isomorphic pair and no structure key therefore crosses a boundary; the partition audit reports zero classes and zero keys crossing train|validation, train|test and validation|test. The test split holds $6{,}110$ subtype, $6{,}772$ type and $8{,}640$ composition structures. All 105 runs used the registered recipe, and the scorecard in Table 4 is generated from the per-seed payloads rather than transcribed. The partition generalises to the remaining twelve evaluation units under one shared rule: the class-to-split bucket depends only on the isomorphism class and the seed, so the extension needed only its own tensor files. Two checks make the two matrices commensurable: the registered label map of every task directory is byte-identical between the in-distribution root and the isomorphism root, so each unit is scored on the same class space; and the taxonomy partition is closed-set there as well, with $3{,}067$ of $3{,}069$ registered ids carrying a training example and no test record carrying an id absent from training. The two uncovered ids occur only in records that are not rooted trees, so they lie outside the isomorphism unit by construction rather than being silently scored.

The stricter unit costs the other baselines far more than the graph Transformers. Relative to the registered exact-structure-ID split, GLYCOTF changes by $+1.22$ points on subtype, $+0.12$ on type and $+0.49$ on composition, and GTF by $+1.28$, -0.12 and $+0.47$; several other baselines instead lose ground on subtype, most sharply GlycanGT (-13.55) and GlycanAA (-7.93), with SweetNet at -1.90 and the two weakest models roughly unchanged (SimpleGCN $+0.15$, R-GCN $+1.66$). On type every baseline except GlycanGT loses 0.9–1.7 points, GlycanGT is flat ($+0.03$), and GTF and GLYCOTF stay within 0.12. The pattern is consistent with the interpretation in the main text: the normalized-key protocol leaves duplicates of the same rooted tree to be split across partitions, and the models that profit from that redundancy are the ones whose scores drop once the unit is tightened, whereas a construction that encodes the tree explicitly loses nothing. Rank agreement with the other two partitions stays high (Spearman $+0.89$ to $+1.00$ over the seven models), so the ordering this paper reports is a property of the models rather than of one split. The full fifteen-unit matrix, including the twelve extension units, is Table 4 in the main text.

G INDEPENDENT OOD REPLICATION SLICE

The OOD scorecards above depend on one export each. As an independent check, we completed a separate composition-disjoint slice on a different remote host under the same registered benchmark version and source hash. Table 16 reports the disease association task for all three compared tree models over seeds 42–44, together with the corresponding GlycoTransportFormer subtype slice. The replication bundle records one source hash for all of its result files. This is a completed task slice rather than a complete 15-task OOD matrix and is kept separate from the headline aggregate. Because the bundle stores thresholded metrics rather than per-example probabilities, this slice cannot be re-scored in micro-AP; we therefore read the disease row only as a sign-and-ordering check against the primary micro-AP results reported for this task in the main text.

H SECONDARY METRIC MODEL COMPARISONS

The headline tables use one primary metric per task. To expose the precision–recall trade-offs behind those scores, Tables 17 and 18 report additional test metrics reconstructed from the same five-seed metadata. Values are percentages and are mean± sample SD; checkpoints remain selected using validation data only.

Table 16: Independent composition-disjoint replication slice from a separate remote host. The subtype column is test micro-F1 mean ± sample SD over seeds 42–44. The disease column is the *thresholded* micro-F1 of the same slice and is reported only as a secondary diagnostic: disease is a positive–unlabelled task whose primary measure is micro-AP throughout this paper, and the frozen payloads of this replication bundle retain thresholded metrics only, so its micro-AP cannot be reconstructed. The disease column is therefore not comparable with the micro-AP columns elsewhere.

Model	Disease association	Subtype
GLYCOTRANSPORTFORMER	0.2954 ± 0.0384	0.9658 ± 0.0038
GlycanTreeFormer	0.0290 ± 0.0048	–
GlycanGT	0.0281 ± 0.0060	–

Table 17: Secondary metrics for composition, subtype, type, and disease-direction tasks. μP and μR are micro-precision and micro-recall; M-F1 is macro-F1. The micro-F1 column is omitted here because it is the registered primary metric already reported in the main scorecards.

Task	Model	μP	μR	M-F1	Subset acc.
Composition	SimpleGCN	87.73±9.48	78.53±3.58	28.30±7.81	44.34±16.65
	SweetNet	97.51±0.82	82.16±0.68	41.03±2.12	62.16±0.97
	GlycanAA	97.46±0.72	82.68±0.37	43.11±1.14	62.67±0.81
	R-GCN	97.25±0.65	79.67±1.09	36.66±2.37	57.37±2.41
	GlycanGT	97.95±0.20	89.57±0.53	59.78±0.98	77.15±0.82
	GTF	98.19±0.09	91.51±0.13	65.10±0.56	81.54±0.30
	GLYCOTF	97.51±0.07	93.05±0.13	74.33±0.12	83.73±0.11
Subtype	SimpleGCN	32.74±0.02	33.76±0.00	2.65±0.00	0.00±0.00
	SweetNet	82.28±3.84	38.02±5.43	10.73±3.80	9.33±6.71
	GlycanAA	85.01±2.68	46.40±7.25	16.17±4.74	19.68±9.39
	R-GCN	80.61±10.01	45.13±5.99	14.08±5.09	17.29±8.66
	GlycanGT	77.34±8.16	65.20±13.38	28.60±11.83	38.77±17.75
	GTF	96.65±0.52	89.75±0.83	52.42±2.83	80.27±1.27
	GLYCOTF	97.42±0.15	94.76±0.12	79.80±0.39	88.15±0.23
Type	SimpleGCN	55.71±1.16	57.82±6.07	11.89±3.87	55.08±0.17
	SweetNet	93.38±0.75	94.99±0.92	61.71±1.25	91.71±1.22
	GlycanAA	94.90±0.45	96.43±0.51	64.31±0.94	93.82±0.42
	R-GCN	95.78±1.02	97.05±0.14	65.93±0.83	95.31±1.09
	GlycanGT	96.71±0.33	95.93±0.26	66.83±4.39	95.87±0.25
	GTF	98.91±0.32	98.72±0.42	88.17±11.17	98.71±0.44
	GLYCOTF	99.21±0.05	99.08±0.12	91.72±4.22	99.04±0.09
Disease direction	SimpleGCN	77.87±11.42	77.78±12.42	50.19±10.82	57.45±23.79
	SweetNet	63.52±8.29	84.81±9.11	63.85±7.93	35.32±17.97
	GlycanAA	74.55±10.10	79.26±13.06	55.77±12.82	54.04±22.75
	R-GCN	71.96±15.29	77.04±8.03	52.93±10.62	48.51±27.47
	GlycanGT	82.85±0.28	73.33±2.48	48.51±7.06	68.09±0.00
	GTF	74.86±4.66	90.00±2.48	74.67±2.45	61.70±9.63
	GLYCOTF	72.28±3.17	77.41±6.60	64.60±3.81	55.32±4.51

Table 18: Secondary metrics for disease association and tissue localization. For these highly imbalanced multi-label tasks, the thresholded F1 and subset accuracy are diagnostic; the ranking-based micro-AP in the main table and the PU audit remain the primary disease measures.

Task	Model	μP	μR	μF1	M-F1	Subset acc.
Disease association	SimpleGCN	1.35±0.17	37.46±15.12	2.55±0.22	1.09±0.42	0.00±0.00
	SweetNet	1.68±0.30	31.58±23.11	3.09±0.44	1.17±0.82	0.00±0.00
	GlycanAA	1.58±0.12	39.26±18.99	2.99±0.20	1.45±0.55	0.00±0.00
	R-GCN	1.49±0.44	18.95±16.18	2.59±0.74	0.75±0.61	0.00±0.00
	GlycanGT	1.57±0.25	23.65±6.88	2.94±0.50	0.65±0.21	0.00±0.00
	GTF	2.25±0.93	12.14±7.58	3.35±0.81	0.32±0.21	0.00±0.00
	GLYCOTF	38.12±3.04	17.15±7.77	22.46±8.14	11.53±6.05	18.30±6.13
Tissue localization	SimpleGCN	1.84±0.64	25.66±13.93	3.29±0.99	0.71±0.41	0.00±0.00
	SweetNet	1.78±0.65	24.04±17.84	2.95±0.50	1.25±0.97	0.00±0.00
	GlycanAA	1.33±0.13	34.72±19.52	2.36±0.51	1.24±0.68	0.00±0.00
	R-GCN	1.55±0.34	33.19±12.14	2.93±0.55	1.60±0.65	0.00±0.00
	GlycanGT	1.51±0.41	14.52±1.86	2.73±0.66	0.39±0.04	0.00±0.00
	GTF	2.01±0.87	16.26±7.28	3.33±0.85	0.54±0.26	0.00±0.00
	GLYCOTF	52.41±2.01	25.92±0.55	34.68±0.76	14.75±0.58	35.19±0.58

I PAIRED TESTS, EFFECT SIZES, AND MULTIPLICITY

Every reported comparison is paired by seed: the seven models share seeds 42–46, so the unit of analysis is the five per-seed differences on a task, not two independently estimated means. Table 19 and Table 20 give, for each task, the comparison against the best trained baseline on that task, the mean paired difference, a percentile bootstrap interval over the five pairs, the paired effect size d_z, the number of seeds with a positive difference, and the exact two-sided Wilcoxon p after Holm correction across the thirteen non-audit tasks. Table 21 aggregates to the task families.

Two properties of this design have to be stated rather than glossed over. First, with five pairs the exact two-sided Wilcoxon p cannot fall below $2/2^5 = 0.0625$, so no individual task survives Holm correction across thirteen comparisons, and the tables report *effect sizes and seed-level sign consistency* instead of significance stars. Reaching a Holm-adjusted $p < 0.05$ over thirteen tasks would require $p \leq 0.0038$ per task, i.e. at least nine paired seeds; that is a concrete requirement for a future revision and is not claimed here. Second, the intervals resample five paired differences and therefore inherit their granularity, and the family-level pattern they summarize is the one the main text reports; the two audit rows stay outside the correction family because a parameter-free kNN already exceeds the model on them (Section N).

Table 19: Paired comparisons against the best trained baseline per task, ID split. Every value is in percentage points of the task's *headline* metric – macro-F1 for the eight taxonomy ranks, micro-F1 for the structural and remaining biological tasks, and micro-AP for disease association (recomputed from the same per-example predictions as Tables 1 and 2, not the run's micro-F1 selection metric); every comparison is paired by seed, so Δ is the mean of the five per-seed differences. The interval is a percentile bootstrap over those five pairs, d_z the paired effect size, "seeds" the number of seeds with a positive difference, and p_{Holm} the exact two-sided Wilcoxon p after Holm correction over the 13 non-audit tasks. The last two rows are the audit tasks, which are excluded from the family summaries; † marks their uncorrected exact p, since they are not in the correction family.

Task	Best baseline	GLYCOTF	Baseline	Δ	95% CI	d_z	seeds	p_{Holm}
Composition	GTF	95.23	94.73	+0.50	$[+0.46, +0.56]$	7.34	5/5	0.812
Subtype	GTF	96.08	93.07	+3.00	$[+2.61, +3.60]$	4.58	5/5	0.812
Type	GTF	99.15	98.82	+0.33	$[+0.07, +0.59]$	0.96	4/5	0.812
Taxonomy domain	GTF	59.31	66.09	-6.78	$[-11.61, -1.95]$	-1.07	1/5	0.812
Taxonomy kingdom	GTF	46.12	47.70	-1.58	$[-8.51, +6.32]$	-0.17	2/5	0.812
Taxonomy phylum	GTF	39.34	29.79	+9.54	$[+4.15, +13.74]$	1.51	4/5	0.812
Taxonomy class	GTF	38.87	30.45	+8.42	$[+5.62, +11.08]$	2.37	5/5	0.812
Taxonomy order	GTF	40.28	28.72	+11.55	$[+11.03, +12.17]$	16.35	5/5	0.812
Taxonomy family	GTF	36.56	21.29	+15.27	$[+13.15, +16.40]$	6.45	5/5	0.812
Taxonomy genus	GTF	34.75	17.23	+17.52	$[+16.60, +18.44]$	14.66	5/5	0.812
Taxonomy species	GTF	28.17	12.44	+15.73	$[+15.11, +16.32]$	20.02	5/5	0.812
Disease direction	GTF	64.60	74.67	-10.07	$[-13.40, -6.74]$	-2.36	0/5	0.812
Immunogenicity	GTF	98.14	99.32	-1.19	$[-2.04, -0.51]$	-1.22	0/5	0.812
Disease association	GlycanGT	23.76	1.59	+22.17	$[+17.71, +24.95]$	4.39	5/5	0.062†
Tissue localization	GTF	34.68	3.33	+31.35	$[+30.67, +32.15]$	31.95	5/5	0.062†

Table 20: Paired comparisons against the best trained baseline per task, composition-disjoint split. Every value is in percentage points of the task's *headline* metric – macro-F1 for the eight taxonomy ranks, micro-F1 for the structural and remaining biological tasks, and micro-AP for disease association (recomputed from the same per-example predictions as Tables 1 and 2, not the run's micro-F1 selection metric); every comparison is paired by seed, so Δ is the mean of the five per-seed differences. The interval is a percentile bootstrap over those five pairs, d_z the paired effect size, "seeds" the number of seeds with a positive difference, and p_{Holm} the exact two-sided Wilcoxon p after Holm correction over the 13 non-audit tasks. The last two rows are the audit tasks, which are excluded from the family summaries; † marks their uncorrected exact p, since they are not in the correction family.

Task	Best baseline	GLYCOTF	Baseline	Δ	95% CI	d_z	seeds	p_{Holm}
Composition	GTF	95.21	95.35	-0.13	$[-0.23, -0.06]$	-1.23	0/5	0.812
Subtype	GTF	96.72	94.34	+2.38	$[+1.62, +3.19]$	2.36	5/5	0.812
Type	GTF	99.08	98.60	+0.47	$[+0.28, +0.78]$	1.42	5/5	0.812
Taxonomy domain	GTF	59.44	57.93	+1.51	$[-1.81, +4.83]$	0.35	3/5	1.000
Taxonomy kingdom	GTF	48.63	47.05	+1.57	$[-2.19, +6.46]$	0.28	3/5	1.000
Taxonomy phylum	GTF	35.48	32.72	+2.76	$[+0.41, +5.09]$	0.88	3/5	1.000
Taxonomy class	GTF	33.28	32.82	+0.47	$[-3.72, +4.27]$	0.09	3/5	1.000
Taxonomy order	GTF	31.10	27.03	+4.07	$[+1.90, +6.32]$	1.39	5/5	0.812
Taxonomy family	GTF	27.08	20.31	+6.77	$[+5.37, +8.17]$	3.78	5/5	0.812
Taxonomy genus	GTF	23.19	15.68	+7.52	$[+6.35, +8.58]$	5.33	5/5	0.812
Taxonomy species	GlycanGT	20.35	11.82	+8.53	$[+7.93, +9.20]$	9.96	5/5	0.812
Disease direction	GTF	65.64	70.38	-4.74	$[-7.83, -1.65]$	-1.19	1/5	0.812
Immunogenicity	GTF	96.33	95.39	+0.94	$[+0.01, +1.79]$	0.79	3/5	1.000
Disease association	GlycanGT	33.11	1.36	+31.75	$[+30.74, +32.77]$	24.45	5/5	0.062†
Tissue localization	GTF	40.65	3.76	+36.89	$[+35.50, +38.28]$	20.98	5/5	0.062†

Table 21: Family-level paired differences against the best trained baseline per task. Each value averages the per-seed differences over the family's tasks and then bootstraps the five resulting seed-level means. The audit family mixes micro-AP (disease association) with micro-F1 (tissue localization), so its average is not a comparable quantity and is reported only for completeness; the two audit tasks are interpreted separately in the text.

Split	Family	Δ	95% CI	d_z	seeds
ID	structural	+1.28	$[+1.13, +1.47]$	5.78	5/5
ID	taxonomy	+8.71	$[+7.05, +9.98]$	4.59	5/5
ID	biology (non-audit)	-5.63	$[-7.29, -3.96]$	-2.63	0/5
ID	audit rows	+26.76	$[+24.52, +28.28]$	10.60	5/5
composition-disjoint	structural	+0.91	$[+0.69, +1.12]$	3.32	5/5
composition-disjoint	taxonomy	+4.15	$[+3.16, +5.14]$	3.24	5/5
composition-disjoint	biology (non-audit)	-1.90	$[-3.74, -0.06]$	-0.79	1/5
composition-disjoint	audit rows	+34.32	$[+33.59, +35.32]$	30.87	5/5

Task (partition)	Variant	Metric (%)	Δ vs full	Seeds worse
Type (composition-disjoint)	full	99.09	–	–
Type (composition-disjoint)	no linkage	98.13	-0.96	5/5
Type (composition-disjoint)	no transport	98.03	-1.06	5/5
Type (composition-disjoint)	commutative	99.09	+0.01	3/5
Type (composition-disjoint)	no cumulant	99.11	+0.02	1/5
Taxonomy family rank (composition-disjoint)	full	28.21	–	–
Taxonomy family rank (composition-disjoint)	no linkage	19.53	-8.68	5/5
Taxonomy family rank (composition-disjoint)	no transport	19.00	-9.20	5/5
Taxonomy family rank (composition-disjoint)	commutative	26.37	-1.84	4/5
Taxonomy family rank (composition-disjoint)	no cumulant	30.63	+2.42	1/5
Taxonomy species rank (composition-disjoint)	full	20.28	–	–
Taxonomy species rank (composition-disjoint)	no linkage	13.82	-6.46	5/5
Taxonomy species rank (composition-disjoint)	no transport	13.91	-6.37	5/5
Taxonomy species rank (composition-disjoint)	commutative	20.97	+0.68	2/5
Taxonomy species rank (composition-disjoint)	no cumulant	23.12	+2.84	0/5

Table 22: Core GLYCOTRANSPORTFORMER ablations on the composition-disjoint partition, extended beyond subtype to type and to the family and species taxonomy ranks. Every variant is retrained from scratch on that task's own partition and compared with the same-source *full* reference over the same seeds; the metric column is the registered primary metric in per cent (micro-F1 for type, macro-F1 for the taxonomy ranks) and Δ is the paired mean per-seed difference in points, so a negative value means the variant loses.

J ADDITIONAL ABLATION DETAILS

The same-source composition-disjoint subtype deltas, paired by training seed against the orthogonal reference, are $(-0.0333, -0.0262, -0.0261, -0.0299, -0.0329)$ for the linkage arm and $(-0.0328, -0.0250, -0.0284, -0.0277, -0.0311)$ for the transport arm; all ten differences are negative.

Table 22 extends the same ablation from subtype to the rest of the composition-disjoint family: type, and the family and species ranks. Removing the linkage descriptors or the path transport costs 0.96 and 1.06 micro-F1 points on type and 8.68/9.20 and 6.46/6.37 macro-F1 points on the family and species ranks, and all thirty paired seed-level differences are negative. The two remaining variants move the metric far less: commutative stays within 1.9 points in either direction, and no-cumulant is never worse on average ($+0.02$, $+2.42$, $+2.84$) although it is worse in one seed of five on type and on the family rank. The variant ordering is therefore the one subtype already gives, and the size of the loss tracks the headroom of the task rather than the arm: type sits at 99.09 micro-F1 and loses about a point, while the two fine ranks, at 28.21 and 20.28 macro-F1, lose six to nine. Every arm was retrained from scratch on its own partition at batch 1024 with the same five seeds as its reference, and no queue mixes batch sizes.

The same two arms on the class-safe motif-profile/cluster-disjoint partition are much smaller (Table 23): the full reference sits at 94.31 micro-F1, and removing linkage descriptors or path transport costs 0.73 points each, reaching 93.58 with four of the five paired differences negative, against the 2.97 and 2.90 points the same arms cost on composition-disjoint subtype. The arms are therefore necessary on every partition we test, but their size is a property of the partition rather than of the arm: where whole isomorphism classes are held out, linkage and transport contribute a fraction of a point, so the conclusion holds with a much smaller margin.

Task	Variant	Metric (%)	Δ vs full	Seeds worse
Subtype	full	94.31	–	–
Subtype	no linkage	93.58	-0.73	4/5
Subtype	no transport	93.58	-0.73	4/5

Table 23: Removing linkage descriptors or path transport on the class-safe motif-profile/cluster-disjoint subtype task. Each variant is retrained from scratch on that partition and paired by seed against the same-source *full* reference; the metric is micro-F1 in per cent and Δ is the paired mean difference in points, so a negative value means the variant loses. The two arms are complete over five seeds and Δ pairs on the seeds the reference has finished (5 of 5 at this revision).

For disease association, mean-outer-product minus full micro-AP is $(-0.1323, -0.2422, -0.2535)$ ID and $(-0.3127, -0.3067, -0.3238)$ OOD. Each reported confidence interval resamples test structures within one trained seed. The three OOD values share the same held-out composition partition. The registered functional-integrity audit covers the headline subtype and disease-control rows. The completed remote reruns below are reported as secondary checkpoint controls with their seed-level summaries and are not used to expand the headline OOD scorecard.

K ROOT AND TERMINAL INTERVENTIONS

Table 24 separates retrained controls from a frozen-model diagnostic. The wrong-root control uses the same model size and training protocol but orients every eligible tree from a deterministic non-reducing leaf. The terminal intervention leaves the trained checkpoint fixed and sets only the $\mathrm{vec}(C_T)$ input coordinates to zero at inference.

Table 24: Paired micro-F1 changes (intervention minus full), mean±sample SD over seeds 42–44. Terminal-count rows use tissue localization.

Control	ID	Composition-disjoint
Wrong root, subtype	-0.0021 ± 0.0015	-0.0250 ± 0.0035
Zero frozen C_T, all tissue	-0.1821 ± 0.0131	-0.1808 ± 0.0045
one terminal	0.0000 ± 0.0000	0.0000 ± 0.0000
two terminals	-0.0588 ± 0.0124	$+0.0001 \pm 0.0141$
at least three terminals	-0.2978 ± 0.0169	-0.2924 ± 0.0096

The single-terminal row is an exact negative control because centering makes $C_T = 0$ when $|T| = 1$. Bootstrap intervals over structure keys exclude zero for the all-structure and three-or-more-terminal strata in every seed. The OOD two-terminal response varies around zero, reinforcing that the largest frozen dependence is specific to richer terminal organization.

L SAME-SOURCE TRANSPORT AND COMPONENT CONTROLS

The next five sections are the controls and diagnostics that the main text cites. Each records its protocol registry and source hash; cells that are not part of the current artifact are marked pending rather than imputed.

All arms share the composition-OOD subtype split and the common recipe (100 epochs, patience 10, tree-batch 1024, AMP, validation-only selection, test evaluated once) and differ in one component. Every arm records the same protocol registry (`e8e267bc39f2`) and a source hash for its own adapter files; paired differences and their confidence intervals are computed per training seed against the untreated orthogonal reference (`f65cd6f2a583`), which is bit-identical to the parent model.

Table 25: Same-source transport and component controls. Paired Δ against the orthogonal reference (0.9695 ± 0.0026, $n = 5$); p is Holm-adjusted over the four complete arms. Only arms with a full five-seed record are listed.

Arm	micro-F1	Δ	95% CI (paired t)	d_z / p
orthogonal (reference)	0.9695 ± 0.0026	—	—	—
commutative	0.9689	-0.0007	$[-0.0074, +0.0060]$	-0.12 / 1.0000
no_cumulant	0.9704	$+0.0008$	$[-0.0055, +0.0072]$	$+0.15$ / 1.0000
no_transport	0.9405	-0.0290	$[-0.0325, -0.0255]$	-9.49 / 0.0001
no_linkage	0.9399	-0.0297	$[-0.0337, -0.0257]$	-8.54 / 0.0001

The per-seed paired differences, in seed order $42, 43, 44, 45, 46$, are $(-0.0101, +0.0006, +0.0054, +0.0020, -0.0013)$ for commutative, $(-0.0082, +0.0040, +0.0058, +0.0032, -0.0006)$ for no-cumulant, $(-0.0328, -0.0250, -0.0284, -0.0277, -0.0311)$ for no-transport, and $(-0.0333,$

$-0.0262, -0.0261, -0.0299, -0.0329$) for no-linkage; every transport and linkage difference is negative. The transport and no-cumulant arms share one adapter source hash (`4f0188799f3b`) because they differ by a runner flag rather than by files, and the linkage arm carries `5facad9ac425`. An earlier realization of the same contrast, paired against a different reference run, agreed in sign and in the ordering of the two effects; we report the registered same-source family of Table 25.

Ordered non-commutative composition is not necessary: the commutative control keeps the angle network, dimensions, initialisation and parameter count while making the rotations commute, and is statistically indistinguishable from the full model ($\Delta = -0.0007$, $d_z = -0.12$, adjusted $p = 1.0$). The cumulant readout is likewise indistinguishable ($\Delta = +0.0008$). Transport and linkage conditioning are load-bearing: identity bonds and zeroed bond descriptors cost 0.0290 and 0.0297 micro-F1 on average, and all ten paired differences are negative with intervals excluding zero. What matters is that residue states are moved into a common rooted frame by linkage-conditioned operators, not that those operators fail to commute.

Earlier incomplete intervention artifacts were excluded after a provenance audit; Table 25 reports only complete five-seed runs whose intervention flags are recorded in the result configuration.

M Disease and tissue negative controls

We audit disease association and tissue localization with nine controls. Table 26 reports test micro-AP; positive density is 1.35% (disease) and 1.28% (tissue). Five controls use five seeds; frequency, co-occurrence, taxonomy, and nearest-neighbour rows are deterministic.

Table 26: Negative controls, test micro-AP (mean $\pm$ SD over five seeds where stochastic). "oracle" uses side information unavailable at deployment.

Control	Disease	Tissue	Notes
composition_only	0.0134 ± 0.0006	0.0160 ± 0.0032	
terminal_count_only	0.0142 ± 0.0016	0.0158 ± 0.0008	
terminal_mean_only	0.0128 ± 0.0011	0.0148 ± 0.0011	
random_train_labels	0.0143 ± 0.0008	0.0142 ± 0.0010	
shuffled_train_pairs	0.0143 ± 0.0008	0.0142 ± 0.0010	
label_frequency_only	0.0537	0.0699	deterministic
label_cooccurrence_only	0.1461	0.2177	oracle (label count)
taxonomy_only	0.1912	0.1106	oracle (gold taxonomy)
nearest_neighbour_similarity	0.3143	**0.4240**	rooted linkage-aware WL kNN

Both columns are test micro-AP, whereas the main table reports micro-F1 for tissue localization, so these metrics are not interchangeable. Simple composition, terminal, and label-permutation controls remain near the positive density, but label-frequency, label co-occurrence, and gold taxonomy provide nontrivial signals. A rooted, linkage-aware Weisfeiler–Lehman kNN obtains 0.3143 disease AP and 0.4240 tissue AP on the structure-key split. This shows that substantial task signal can be recovered without training a model; it does not identify whether this reflects biological regularity, annotation practice, or dataset construction.

N Retrieval under coarse-scaffold-signature masking and composition shift

To test whether retrieval only transfers labels between near-duplicates, every training neighbour sharing the query's coarse scaffold signature is masked to zero similarity before top-k selection; the coarse scaffold signature is defined as the sorted monosaccharide multiset plus the sorted undirected degree sequence, so linkage and anomericity are ignored. The A/B is repeated on the composition-disjoint split.

Table 27: Structural retrieval control, test micro-AP. "all" uses every training neighbour, "disjoint" masks all neighbours sharing the same coarse scaffold signature.

Split	Task	all	disjoint
structure-key	disease	0.3143	0.3833
composition-disjoint	disease	0.3885	0.3870
structure-key	tissue	0.4240	0.4501
composition-disjoint	tissue	0.4587	0.4636

The motif-profile rows give the same reading on the stricter partition (Table 28): a kNN that never trains on the task exceeds GlycoTF on both sparse biological tasks (15.17 against 6.95 micro-AP for

Table 28: Motif-profile audit rows under the same parameter-free retrieval control (task-universe v5 partition, the partition of Table 3). Each row uses the metric the main table uses for that task (micro-AP for disease association, micro-F1 for tissue localization), so the model and the control are comparable within a row. The control's other metric is 20.29/20.81 micro-F1 for disease association and 22.21/22.10 micro-AP for tissue localization (all/masked arms). "masked" zeroes the similarity of every training neighbour sharing the query's coarse scaffold signature.

Audit row	Metric	GLYCOTF	kNN, all	kNN, masked
Disease association	micro-AP	6.95	14.65	**15.17**
Tissue localization	micro-F1	17.66	28.20	**28.17**

disease association, 28.17 against 17.66 micro-F1 for tissue localization), and masking same-scaffold neighbours changes each arm by at most 0.53 points, so the margin is not near-duplicate retrieval. These are the numbers cited in Section 6.3.

All cells in Table 27 are micro-AP; the tissue column of the main table is micro-F1, so they are not directly comparable. Masking neighbours sharing the same coarse scaffold signature preserves or improves retrieval under both splits, showing that the signal is not restricted to examples with a shared coarse scaffold signature. The disease retrieval AP is higher than the corresponding benchmark headline AP on both splits, but this deterministic comparison does not identify biological mechanism. Validation selected the largest grid value $k = 50$ in every composition-disjoint arm, so the grid boundary was active and a larger neighbourhood may change performance. Frequency and co-occurrence controls are available on the structure-key audit split, but have not been evaluated under composition-disjointness; support-stratified AP is reported inTable 29 for the available control export.

O LABEL-SUPPORT STRATIFICATION

Table 29: Disease macro-AP by training-label support. Of 170 labels, 3 have no training support and 55 have no test positive.

Control	support 1–4	5–19	20+
nearest_neighbour_similarity	**0.5407**	**0.5183**	0.3684
taxonomy_only (oracle)	0.1483	0.1820	0.1873
label_cooccurrence_only (oracle)	0.1745	0.1464	0.1632
composition_only	0.0421	0.0360	0.0870
label_frequency_only	0.0111	0.0125	0.0399

Retrieval is strongest on the rarest labels and weakest on the most frequent, the opposite of the frequency intuition: highly supported disease labels are generic and noisy, whereas rarely supported ones tend to be specific structural associations a nearest neighbour retrieves directly. Only 6.1% of scored labels have AP below 0.05 under retrieval, against 93.9% under the frequency control, which quantifies which labels admit direct structural inference.

P FROZEN-ENCODER LINEAR PROBES

Table 30 runs one probe set through all seven trained models. The node-permutation control is an exact graph isomorphism, so a permutation-invariant encoder must place it at 0; six models do. The exception, GlycanGT, moves 4.6, on the same scale as its own linkage response: it is a TokenGT-style model whose token sequence follows input node order (its incidence and position encodings are built from that order), so the row measures order sensitivity rather than a defect in the probe or the training run, and it should be read as an order-sensitivity diagnostic rather than a like-for-like invariance comparison. The table also shows which structural distinctions survive training: SimpleGCN is blind to linkage and anomericity, R-GCN and GTF to anomericity, while branch and terminal edits move every model. GLYCOTF has the smallest linkage and anomer distances among the structure-sensitive models (3.9 and 3.2) and separates every edit from the permutation floor perfectly (AUROC 1.000), so its responses are small but exactly discriminative rather than absent. The branch-serialization control contains two eligible structures and is therefore a targeted invariance check, not a population-level estimate.

A single linear layer is fitted on frozen GlycoTF node embeddings, with splits assigned per structure key and the probe fitted on the training split only.

All four attributes are linearly decodable far above the majority baseline, supporting the claim that the representation retains them; as with the edit probes this establishes decodability, not chemical validity

Table 30: **Structural edit and invariance probes for every benchmark model.** Values are $100\times$ the mean L2 distance between L2-normalised penultimate representations of a base structure and its edited or permuted copy, over five seeds and up to 224 bases (220 with a structure-changing branch edit). "perm" is the node-permutation control, an exact isomorphism that a permutation-invariant encoder must place at 0; AUROC separates each edit from it (1.0 = perfect separation). Every model sees the same probe set.

Model	Linkage	Anomer	Branch	Terminal	perm	$\text{AUROC}_{\text{min}}$
SimpleGCN	0.00	0.00	1.22	3.68	0.00	0.53
SweetNet	19.31	18.70	34.04	39.49	0.00	1.00
GlycanAA	43.27	33.70	61.50	38.25	0.00	1.00
R-GCN	51.17	0.00	70.65	30.70	0.00	0.50
GlycanGT	16.35	11.53	22.74	28.08	4.57	0.93
GTF	11.33	0.00	58.94	72.08	0.00	0.40
GLYCOTF	3.92	3.17	38.91	37.25	0.00	1.00

Table 31: Frozen-encoder linear probes (4,000 graphs per split, $n_{\text{test}} \approx$ 34–38k nodes).

Target	Probe acc.	Probe macro-F1	Majority acc.	Majority macro-F1
linkage position	0.9057	0.7685	0.3972	0.0711
anomer	0.9573	0.9544	0.4111	0.1942
root direction	0.8645	0.8654	0.3726	0.1357
terminal organisation	0.9682	0.9639	0.6861	0.4069

or biological utility. A 300-graph smoke run gave linkage accuracy 0.373, below the majority baseline, so probe sample size materially changes the conclusion. This probe table covers the reference encoder only; the edit and invariance probes of Table 30 cover all seven models.

Q EXTERNAL TRANSFER AVAILABILITY

Table 32: External-transfer data gate.

Dataset	Verdict	Scale	Note
protein–glycan interaction	conditional pass	799,408 rows	43.5% map to corpus; 56.5% external
lectin binding	fail	0 bytes	placeholder; no records
GlyConnect protein–disease	conditional pass	23,126 rows	accessions only; no structure map
glycan interaction matrix	fail	931 cols	no documented label semantics

A valid protocol must restrict evaluation to the 1,626 glycans absent from the corpus, and no dataset passed unconditionally, so external transfer is retained as a limitation and no proxy results are created. This is consistent with the earlier protein–glycan experiments, where frozen-embedding Spearman is at noise level and fine-tuned Spearman stays at 0.11–0.17.

R OPTIMIZATION CONFIGURATION

Table 33 records the single configuration used for every model and task, and Table 34 the architecture widths that the shared factory instantiates. The publication runs pass `--training-recipe common`, which selects one optimizer, loss, and schedule for all seven architectures; the task- and model-adaptive branch that also exists in the runner was not used, so no per-task tuning separates the comparison.

S COMPARISON SCOPE: CAPACITY MATCHING AND COST

The audit instantiates every model through the same publication entry point (Table 35); all parameters are trainable. This is an architecture-level comparison rather than a capacity-matched contest. GTF and GLYCOTF are relatively close in size, whereas comparisons with the small GCN baselines may reflect both inductive bias and capacity. We therefore use component-matched interventions for mechanism claims and treat the seven-model rank as descriptive coverage. Parameter matching is closest on the subtype task, where GTF (5,677,758) and GlycanGT (4,071,973) bracket GLYCOTF (6,049,865); GlycoTF uses 6.6% more parameters than GTF while requiring fewer floating-point operations per graph (59.63M versus 69.29M in the efficiency microbenchmark of Appendix T). We do not report a strictly capacity- and compute-budget-matched edge-aware Transformer comparator, so the overall ranking is a comparison under one implementation and training recipe rather than a controlled capacity study.

Table 33: Optimization configuration. Every entry is identical for all seven models and all fifteen tasks; no per-task or per-model tuning is applied in the reported runs.

Component	Setting
Optimizer	AdamW, single parameter group
Learning rate	2×10^{-4}
Weight decay	1×10^{-4}
Schedule	linear warm-up over 5 epochs, then cosine decay to zero over the remaining 95
Epoch budget	100
Early stopping	patience 10 epochs, min_delta $= 10^{-4}$, on the validation primary metric
Gradient clipping	global norm 1.0
Batch size	2048 for every model (tree batch 2048, GlycanGT batch 2048)
Precision	automatic mixed precision (FP16) with loss scaling
Class weighting	none
Label smoothing	none
Loss, single-label	cross-entropy with logits, unweighted
Loss, multi-label	binary cross-entropy with logits
Model selection	highest validation primary metric; test evaluated once
Seeds	42–46
Feature schema	24-dim residue tensor, 20-dim bond tensor, bidirectional edge index

Table 34: Architecture configuration of the seven compared models, including GLYCOTF, as instantiated by the shared factory. Widths, depths, and dropout are identical across tasks; only the output head depends on the label count.

Model	Depth, width, heads	Dropout
SimpleGCN	graph convolution baseline	–
SweetNet	3 layers, width 96	–
GlycanAA	3 layers, width 96	–
R-GCN	3 layers, width 96, 20 relations	–
GlycanGT	6 layers, width 256, 8 heads	0.1
GTF	6 layers, width 256, 8 heads	0.1
GLYCOTF	6 layers, width 256, 8 heads, frame width 8	0.1

Table 35: Audited cost profile per model, for the shared 37-label head. Parameters are trainable counts after instantiation through the benchmark factory and vary slightly with output dimension, so they expose capacity rather than establish matching. Forward FLOPs are per graph at the most common node-size bucket (batch 64, random initialisation), measured with PyTorch's operation counter; they exclude parsing, data transfer, and cache creation, and are a contention-independent compute diagnostic rather than end-to-end throughput.

Model	Parameters	FLOPs/graph	FLOPs/parameter
SimpleGCN	43,141	0.29M	434
SweetNet	62,437	0.64M	657
GlycanAA	52,261	0.50M	617
R-GCN	594,949	7.02M	755
GlycanGT	4,071,973	144.92M	2,278
GTF	5,677,758	69.29M	781
GLYCOTF	6,049,865	59.63M	631

GlycoTF carries 6.6% more parameters than GTF but a lower forward cost per graph, so its ranking cannot be attributed to more compute per graph. Wall-clock step times were measured while other jobs shared the device and include the per-graph Python tree construction that the registered protocol amortises through the tree cache; they are therefore not comparable to the isolated efficiency microbenchmark and are excluded from the fairness argument.

T SCOPED EFFICIENCY MICROBENCHMARK

Table 36: Isolated RTX 5090 microbenchmark at batch sizes 64, 256, and 1024. The table reports the batch-1024 slice; throughput is graphs/s and memory is peak allocated GiB. Three warm-up and ten timed steps; parsing and cache construction are excluded. Parameter counts are in Table 35.

Model	Train/s	Infer/s	Train GiB	Infer GiB
GlycanGT	10,039	12,880	1.318	0.252
GTF	1,136	2,413	1.005	0.299
GLYCOTF	849	2,040	0.771	0.306

U FORMAL STATEMENTS AND PROOFS

We restate the propositions of Section 4.7 with proofs. All three statements concern the parameterization rather than the fitted models.

A *root-preserving labelled-tree isomorphism* is a bijection $\sigma\colon V \to V'$ with $\sigma(r) = r'$, $x'_{\sigma(v)} = x_v$, $e'_{\sigma(u)\sigma(v)} = e_{uv}$, and preserved directed adjacency. Write $g(G)$ for the descriptor of Equation 12 and $\{Z_v\}$ for the transported states of Equation 4.

Proposition 1 (Root-preserving permutation invariance)**.** *If σ is a root-preserving labelled-tree isomorphism, then $Z'_{\sigma(v)} = Z_v$ for every $v \in V$, and $g(G') = g(G)$.*

Proposition 2 (Ordered-path separation)**.** *For the overlapping planes $(0,1),(1,2),(2,3),(3,4)$ there exist bond descriptors e, f with $R(e)R(f) \neq R(f)R(e)$. Thus two equal-depth terminals with reversed linkage order can receive different transported states whenever $H(R(e)R(f) - R(f)R(e)) \neq 0$; commuting or additive path controls assign them the same state.*

Lemma 1 (Norm preservation)**.** *Under orthogonal per-bond operators, $\|Z_{v,j:}\|_2 = \|H_{v,j:}\|_2$ for every channel j (Equation 5).*

For n residues, $m = n - 1$ bonds, hidden width d, frame width k, L layers, and p boundary dimensions, path operators cost $O(mk^3 + ndk)$, global attention $O(Ln^2d)$, branch contraction $O(Lnd)$, and the boundary moment $O(|T|p^2)$. The formulation is limited to rooted acyclic glycans with parsed bond descriptions; it does not infer three-dimensional conformation or loop holonomy.

Notation. Let $G = (V, E, r)$ be a rooted, edge-labelled tree with node features x_v, bond descriptors e_{uv}, and rooted depths $\delta(v)$. For node v, let $\pi(v) = (v_0, \dots, v_m)$ be the root-to-node path with $v_0 = r$ and $v_m = v$. Write R_{uv} for the connection of Equation 3 and $R(\theta)$ for the composition evaluated at angle vector θ.

Proof of Proposition 1. Let σ be a root-preserving labelled-tree isomorphism. Rooted depth is determined by the rooted labelled tree, so $\delta(\sigma(v)) = \delta(v)$; therefore the residue encoder of Equation 1 satisfies $H'_{\sigma(v)} = H_v$ for every v, since it is applied pointwise and its only structural input is the depth embedding.

For the transport recursion, σ preserves directed adjacency and edge labels, so the image of $\pi(v) = (v_0, \dots, v_m)$ is $\pi(\sigma(v)) = (\sigma(v_0), \dots, \sigma(v_m))$ with $e'_{\sigma(v_i)\sigma(v_{i+1})} = e_{v_i v_{i+1}}$ for every i. The recursion of Equation 4 composes the operators in path order, and that order is indexed by depth rather than by node identifier. Hence $P'_{\sigma(v)} = R'_{\sigma(v_0)\sigma(v_1)} \cdots R'_{\sigma(v_{m-1})\sigma(v_m)} = R_{v_0 v_1} \cdots R_{v_{m-1} v_m} = P_v$, giving $Z'_{\sigma(v)} = H'_{\sigma(v)} P'_{\sigma(v)} = Z_v$. The collection $\{Z_v\}$ is therefore permuted, not altered.

Every remaining operation is a set operation or is indexed by structure rather than by node identity. Global attention (Equation 7) has no positional term other than the depth-conditioned inputs, so it is permutation equivariant. The child sum in Equation 10 is taken over the child set, and the normalization depends only on $\delta(v)$, so sibling order does not enter. The readout (Equation 12) uses the root state z_r, the terminal mean, the terminal max, and the centered terminal moment C_T; C_T is a sum over the terminal set of centered outer products and is therefore invariant to terminal enumeration, and the normalization by $|T|$ removes cardinality effects. Consequently $g(G') = g(G)$. □

Proof of Proposition 2. Both parts are established by explicit matrices in the frame dimensions actually used by the model ($k = 8$, planes $(0,1),(1,2),(2,3),(3,4)$), evaluated by the released non-commuting transport witness.

Non-commuting witness. Take $\theta = (\pi/2, \pi/2, 0, 0)$ and $\varphi = (0, 0, \pi/2, \pi/2)$. Direct evaluation gives $\max_{ij} |(R(\theta)R(\varphi) - R(\varphi)R(\theta))_{ij}| = 1$ and $\|R(\theta)R(\varphi) - R(\varphi)R(\theta)\|_F = \sqrt{6} \approx 2.4495$, whereas angle vectors whose support is on disjoint planes commute exactly ($\|R(\theta)R(\varphi) - R(\varphi)R(\theta)\|_F = 0$). The witness therefore depends on the overlapping-plane chain and not on Givens rotations as such. Two further random angle pairs also give nonzero Frobenius separation, so non-commuting pairs are not isolated.

Separation of ordered paths. Construct a tree with root r and two children a, b, each with one further child, so that t_1 (below a) and t_2 (below b) both have $\delta = 2$. Assign descriptors e to (r, a), f to (a, t_1), f to (r, b), and e to (b, t_2). Then $P_{t_1} = R(e)R(f)$ and $P_{t_2} = R(f)R(e)$, which differ by the witness. Give the two terminals identical residue features, so $H_{t_1} = H_{t_2} = H$ with $H \neq 0$. If $H(R(e)R(f) - R(f)R(e)) \neq 0$ then $Z_{t_1} \neq Z_{t_2}$: the two terminals share a rooted depth and carry the same descriptors in opposite order, yet receive different transported states. The set of H for which this fails is the kernel of a nonzero linear map, hence a proper closed subset.

Order-invariant operators cannot separate them. The additive control accumulates $P_v = I + \sum_{(u,w)\in\pi(v)}(R_{uw}-I)$. The two paths above receive $I+(R(e)-I)+(R(f)-I) = R(e)+R(f)-I$ and $I+(R(f)-I)+(R(e)-I)$, the same matrix; measured on the witness inputs, $\max_{ij}|P_{t_1}-P_{t_2}| = 0$. A commutative connection gives $R(e)R(f) = R(f)R(e)$ directly. Both therefore collapse the two terminals, which is what makes the additive family a test of this proposition rather than a tuning knob. □

Proof of Lemma 1. Let P_v be the ordered product of the bond operators on $\pi(v)$. Each factor satisfies $R_{uv}^\top R_{uv} = I$, so $P_v^\top P_v = R_{v_m}^\top \cdots R_{v_1}^\top R_{v_1}\cdots R_{v_m} = I$. For channel j, $Z_{v,j:} = H_{v,j:}P_v$ and hence $\|Z_{v,j:}\|_2^2 = H_{v,j:}P_vP_v^\top H_{v,j:}^\top = \|H_{v,j:}\|_2^2$, which is Equation 5. The identity depends only on orthogonality: replacing R_{uv} by an unconstrained operator removes the guarantee. On the witness inputs an unconstrained dense operator has a larger Frobenius norm against $\sqrt{8} = 2.8284$ for an orthogonal one, so channel norms can be amplified. The lemma concerns the pre-encoder alignment step only; attention, branch aggregation, and the feed-forward network remain unconstrained. □

V LIMITATIONS

Public glycan annotations are sparse and incomplete. Disease association is positive–unlabelled: its 170 labels record evidence of association, so the 98.6% of the target matrix that is zero mixes genuine negatives with pairs that were simply never annotated, and the binary cross-entropy used for every model treats the two identically. The task is therefore reported as a ranking problem, the micro-AP of the baselines sits at the positive density, and the ordering among unannotated pairs is not identified by this supervision. Disease direction and immunogenicity are small; their estimates should not be read as clinical or causal evidence. The ID disease split also exhibits measurable label-frequency shift (train–test Jensen–Shannon divergence 0.0964), and 55 labels have no positive test example. Random-label, shuffled-pair, composition-only, terminal-feature, label-frequency, label-co-occurrence, taxonomy-only, and rooted linkage-aware nearest-neighbour controls are now available. The simple feature and permutation controls remain near the positive-density floor, but structural retrieval reaches micro-AP 31.43 on the structure-key split and 38.85 on the composition-disjoint split; excluding neighbours sharing the same coarse scaffold signature does not remove this signal. The disease improvement therefore remains retrieval-explainable and is not evidence of an encoder-specific mechanism. Because unlabeled pairs are treated as negatives by the shared binary-cross-entropy recipe, ECE and Brier score are diagnostics against recorded labels only, not calibration to true disease-association probabilities. Normalized-key disjointness prevents duplicate-key leakage but does not remove scaffold overlap or prove graph-isomorphic chemical identity across alternative valid string notations: over $3{,}000$ structures the parser is deterministic ($3{,}000/3{,}000$ re-parses agree) and the graph signature is a genuine isomorphism invariant ($3{,}000/3{,}000$ invariant under node relabelling), yet reversing the order of sibling branch groups in the IUPAC string reproduces the same structure in only $541/1{,}469$ multi-branch cases (36.8%). Our identity control is therefore a *canonicalized-serialization-level* control rather than a chemical-isomorphism guarantee, although the exact-isomorphism split (Section 6.4) removes the isomorphic variants that either key would have admitted. WURCS/GlycoCT round-trip consistency was not tested, because no converter for those notations is available in this environment. The composition-disjoint evaluation uses five seeded group partitions over one structure universe; their held-out groups overlap, so the reported seed variation is not population-level uncertainty. Probability-level quantities such as ECE and Brier score require the separate frozen-export audit artifacts.

The benchmark evaluates curated annotation prediction, not direct molecular function, binding, or causal biological understanding. An external-data audit found a protein–glycan interaction file with 799,408 rows but only 43.5% corpus key coverage, an empty lectin-binding file, and a protein–disease table that still requires accession mapping; no external-transfer result is therefore reported. The main comparison uses one registry, benchmark release, feature schema, optimization recipe, and validation-only selection protocol, retaining a source snapshot per model implementation. The benchmark's audits are model-independent; only the comparative ranking depends on GLYCOTF. It does not include a strictly parameter- and compute-matched edge-aware Transformer comparator. Baseline capacity is not fully matched, and local from-scratch training does not reproduce every pretrained resource used in the source publications. The completed capacity-matched subtype control narrows this concern for that task and split but does not replace matched controls for every baseline and task. A validation-only, equal-budget hyperparameter search covers all seven architectures on the three structural tasks (subtype, type, composition: 20 configurations from the first grid plus 64

supplementary cells, one seed each, learning rate and weight decay only, read against the registered seed-42 value for the same task and model); it does not extend to the taxonomy or biological task families, and with one seed per configuration a delta smaller than the seed-to-seed spread is not interpretable. The isolated profile reports parameter count, throughput, and peak memory for GlycanGT, GTF, and GLYCOTF; a separate FLOP estimate uses a synthetic six-node batch and is a scoped operation-count diagnostic rather than end-to-end cost. Mean rank also weights eight correlated taxonomy levels equally. Statistical comparisons are paired by seed, but only five seeds exist per cell, so the exact two-sided Wilcoxon p cannot fall below 0.0625 and no individual task survives Holm correction over the thirteen non-audit comparisons; a Holm-adjusted $p < 0.05$ would require at least nine paired seeds even if every sign agreed. The reported evidence is therefore the paired effect size, the seed-level sign consistency, and reproducibility across the four partitions, not significance, and the two audit rows are excluded from the correction family rather than corrected within it (Appendix I).

Mechanistically, linkage removal and transport removal are supported on one composition-disjoint subtype split. Same-source orthogonal, dense, and additive controls show that the orthogonality constraint helps relative to an unconstrained dense map, but the additive control matches the ordered reference. These task-scoped results therefore do not establish that non-commutativity is necessary. The disease readout result compares a centered cross-moment with a bilinear mean outer product and therefore does not isolate polynomial order; its relation to a separately retrained no-cumulant model is also unstable across composition-disjoint seeds. The wrong-root control changes the anchor, path directions, and terminal set together, and therefore does not isolate which rooted operation produces the OOD decrease. The frozen terminal intervention measures reliance after training, while the no-cumulant experiment measures adaptation under retraining; neither alone establishes a causal biological role for terminal interactions. Synthetic edits test parsed-feature sensitivity, not chemical validity, binding energy, three-dimensional conformation, or biological causation. Finally, our rooted-tree formulation does not address cyclic or ambiguous structures that lack a unique directed tree.