Glycan Reachability Analysis: A Bottleneck-Aware Framework for Inferring Tissue-Specific Glycan Biosynthetic Potential from Transcriptomics

Yusuke Matsui[1,2*]

[1]Biomedical and Health Informatics Unit, Department of Integrated Health Sciences, Nagoya University Graduate School of Medicine, Nagoya, Japan
[2]Institute for Glyco-core Research (iGCORE), Nagoya University, Nagoya, Aichi, Japan
[*]Corresponding author: Y. Matsui

Abstract

Glycan biosynthesis requires the coordinated expression of glycosyltransferases, modifying enzymes, and nucleotide sugar donor synthesis/transport machinery. Existing computational tools predict glycan structures from gene expression using binary thresholds, losing quantitative information about relative biosynthetic capacity across tissues. Here we present glycan biosynthetic reachability analysis—a framework that quantifies the capacity of a tissue to synthesize a given glycan structure based on the expression of the least-abundant pathway component—integrating expression-based Z-scores across biosynthetic pathway steps using AND/OR logic and a minimum-aggregation (bottleneck) principle to produce continuous, tissue-comparable scores of glycan biosynthetic capacity. Applied to 17,382 RNA-seq samples across 54 human tissues from GTEx v8, we demonstrate that reachability scores capture biologically meaningful tissue-specific variation in sialyl Lewis X (sLeX), ganglioside, and heparan sulfate biosynthetic capacity. Critically, we show that binary threshold approaches lose the quantitative resolution needed to distinguish tissues where all pathway enzymes are detectable yet expressed at uniformly low levels (e.g., pancreas for sLeX: 96% binary-positive, reachability $Z = -1.86$) from tissues with genuinely high biosynthetic capacity. Systematic comparison of aggregation functions confirms that min-aggregation (bottleneck principle) consistently outperforms naive mean expression in predicting downstream biological outcomes. We emphasize that reachability scores reflect transcriptomic potential — not proven enzymatic activity or glycan abundance — as post-transcriptional regulation, protein stability, and substrate competition are not captured. Our method requires only bulk RNA-seq data, no training data or kinetic parameters, and can in principle be applied to other species through ortholog mapping, though cross-species calibration and biological validity remain context-dependent.

1 Introduction

Glycosylation is among the most prevalent and complex post-translational modifications, with the human glycome estimated to encompass thousands of distinct structures synthesized by approximately 700 glycosylation-related genes [1, 2]. The tissue-specific expression of glycosyltransferases, glycosidases, and nucleotide sugar metabolism genes gives rise to organ-specific glycan repertoires that influence cell signaling, immune recognition, and pathogen interactions [3].

Recent advances in computational glycobiology have produced tools that leverage transcriptomic data to predict glycan structures. GlycoMaple [4, 5] maps RNA-seq expression onto 21 curated glycan metabolic pathways, predicting the presence or absence of glycan structures based

on whether pathway enzymes exceed a TPM threshold of 1.0. Glycologue [6] uses rule-based logic to predict products from a set of active enzymes. glycoPATH [7] employs machine learning regression to predict quantitative N-glycan profiles, but requires paired glycomic–transcriptomic training data and is limited to N-glycans. Glycopacity [8] and glycoCARTA [9] analyze glycogene expression patterns from single-cell transcriptomics but do not integrate pathway logic to assess biosynthetic potential.

A common limitation of these approaches is that they either (i) reduce expression to binary presence/absence, discarding quantitative information about relative biosynthetic potential, or (ii) require experimental glycomic data for training. None provides a framework for statistically comparing biosynthetic potential across tissues or conditions using expression data alone.

Here we address this gap with glycan reachability analysis. We define the reachability of a glycan structure as the minimum normalized expression across all required biosynthetic steps — including both enzymatic reactions and nucleotide sugar donor pathways. This continuous score, termed *biosynthetic reachability* (or simply reachability hereafter), operationalizes a simple heuristic: the least-expressed pathway component constrains the overall transcriptomic potential of the pathway. The underlying assumption is that, when enzymes operate in the linear (unsaturated) regime where reaction rate scales approximately with enzyme concentration, the weakest link in the chain sets an upper bound on pathway output. This assumption is strongest when substrate concentrations are below K_m for most enzymes, a condition that may not hold universally but provides a useful first-order approximation. We emphasize that reachability scores measure transcriptomic biosynthetic potential — not enzymatic activity or glycan abundance — as gene expression is an imperfect proxy for protein function. Unlike metabolic control analysis (MCA), which demonstrates that flux control is distributed among multiple enzymes with flux control coefficients summing to unity [30], our approach does not claim to predict metabolic flux or identify true rate-limiting steps. Rather, it provides a rank-ordering of tissues by their transcriptomic capacity for each biosynthetic pathway.

2 Methods

2.1 Reachability score definition

For a glycan biosynthetic pathway P consisting of enzymatic steps $E_1, E_2, \ldots, E_k$ and required donor substrates $D_1, D_2, \ldots, D_m$, the reachability score for sample s is defined as:

$$\mathrm{Reach}(P, s) = \min\big(E_1(s),\, E_2(s),\, \ldots,\, E_k(s),\, D_1(s),\, D_2(s),\, \ldots,\, D_m(s)\big) \tag{1}$$

Each enzymatic step E_i may be catalyzed by one or more isozymes. We aggregate isozyme contributions using the arithmetic mean:

$$E_i(s) = \mathrm{mean}\big(Z(e_{i1}, s),\, Z(e_{i2}, s),\, \ldots,\, Z(e_{in}, s)\big) \tag{2}$$

Sensitivity of results to isozyme grouping and OR (isozyme-family aggregation) choices is detailed in Supplementary Methods.

where $Z(g, s)$ is the per-gene Z-score computed as:

$$Z(g, s) = \frac{\log(1 + \mathrm{TPM}(g, s)) - \mu_g}{\sigma_g} \tag{3}$$

with μ_g and σ_g being the mean and standard deviation of $\log(1 + \mathrm{TPM})$ across all samples.

Each donor substrate D_j requires the coordinated expression of synthesis enzymes, activating enzymes, and Golgi transporters with AND logic:

$$D_j(s) = \min\big(\mathrm{synth}(s),\, \mathrm{activation}(s),\, \mathrm{transport}(s)\big) \tag{4}$$

For donor substrates synthesized by alternative pathways (e.g., GDP-Fuc via de novo or salvage), OR logic is applied between pathways:

$$\text{GDP-Fuc}(s) = \max\big(\text{de_novo}(s),\ \text{salvage}(s)\big) \tag{5}$$

2.2 Glycan pathways modeled

Five glycan families were modeled, yielding 23 reachability metrics (Tables 1–2; pathway schematics in Fig 1A–E; metric definitions in Fig 1F; machine-readable definitions in Supplementary File S1). Several biological simplifications are noted: MGAT4A/B (tri-antennary branching) and MGAT5 (tetra-antennary branching) are grouped as a single OR step despite catalyzing distinct reactions; the HS3ST family (HS3ST1–6) is modeled as one OR group despite distinct substrate specificities; and B4GALT1–6 are grouped for LacNAc synthesis despite differential glycoprotein/glycolipid preferences (see Supplementary Methods for sensitivity analyses). Within each pathway, enzymatic steps are connected by AND logic (min-aggregation; all components required). Where multiple isozymes catalyze the same reaction, OR logic (mean-aggregation) is applied. Enzyme complexes are modeled with appropriate subunit logic: EXT1/EXT2 (required heteromeric complex, AND), ALG13/ALG14 (required heterocomplex, AND), and the OST complex (catalytic STT3A/B as OR; structural subunits RPN1, RPN2, DAD1, DDOST, OSTC as AND; oxidoreductase TUSC3/MAGT1 as OR). All LLO assembly genes (DPAGT1 through ALG10) are modeled as sequential AND steps, reflecting that each is essential and non-redundant (deficiency causes Congenital Disorders of Glycosylation). Five nucleotide sugar donor supply chains, shared across multiple pathways, are summarized in Table 1. GALE appears in both the UDP-Gal and UDP-GalNAc donor pathways, as it catalyzes epimerization of both UDP-Glc/Gal and UDP-GlcNAc/GalNAc [27].

Table 1: Nucleotide sugar donor substrate supply chains. Each uses AND (min) logic across sequential biosynthetic, activation, and Golgi transport steps. GDP-Fuc uniquely uses OR_{alt} (max) between de novo and salvage branches. GALE appears in both UDP-Gal and UDP-GalNAc pathways as it catalyzes both UDP-Glc/Gal and UDP-GlcNAc/GalNAc epimerization.

Donor	Sequential steps (AND)	Used by
CMP-Sia	GNE → NANS → NANP → CMAS → SLC35A1	sLeX, Ganglio., N-gly., O-GalNAc
GDP-Fuc	max(GMDS → TSTA3, FUK → FPGT) → SLC35C1	sLeX, N-glycan
UDP-Gal	UGP2 → GALE → SLC35A2	sLeX, Ganglio., O-GalNAc
PAPS	PAPSS1/2 *(OR)* → SLC35B2/B3 *(OR)*	HS
UDP-GalNAc	UAP1 → GALE → SLC35D1/A2 *(OR)*	Ganglio., O-GalNAc

Table 2: Glycan biosynthetic pathway steps and reachability metrics. Steps within each pathway are connected by AND (min) logic; isozyme groups use isozyme-family aggregation (OR = mean). Required donors refer to Table 1. 22 glycan pathway metrics plus PAPS donor supply as a standalone metric (23 total; see S4 Table for a complete listing with pathway family, step count, gene count, and primary/composite classification). SHH-like differs from WNT-like by requiring 3-O sulfation. Display names (Metric column) are used throughout the main text; corresponding code variable names are listed in S4 Table.

Step	Enzymes (logic)	Metric
Sialyl Lewis X (sLeX)	donors: CMP-Sia, GDP-Fuc, UDP-Gal	
LacNAc formation	B4GALT1–6 (OR)	
α2,3-sialylation	ST3GAL3/4/6 (OR)	
α1,3/4-fucosylation	FUT3/5/6 (OR)	sLeX reachability
Gangliosides	donors: CMP-Sia, UDP-Gal, UDP-GalNAc	
LacCer synthesis	UGCG + B4GALT5/6 (OR)	
GM3	+ ST3GAL5	GM3 reach
GM2	+ B4GALNT1	GM2 reach
GM1	+ B3GALT4	GM1 reach
GD3	GM3 + ST8SIA1	GD3 reach
Heparan sulfate (HS)	donor: PAPS	
Chain polymerization	EXT1 + EXT2 (AND)*	HS polymerization
N-sulfation	NDST1–4 (OR)	HS N-sulfation
2-O-sulfation	HS2ST1	HS 2-O sulfation
6-O-sulfation	HS6ST1–3 (OR)	HS 6-O sulfation
3-O-sulfation	HS3ST1–6 (OR)	HS 3-O sulfation
FGF-like profile	Core + N + 2-O (AND)	HS FGF-like
WNT-like profile	Core + N + 6-O (AND)	HS WNT-like
SHH-like profile	Core + N + 6-O + 3-O (AND)	HS SHH-like
N-glycan processing	donors: GDP-Fuc (core fuc.), CMP-Sia (sia.)	
LLO initiation	DPAGT1	
LLO assembly	ALG13/14 (AND)‡, ALG1, 2, 11, 3, 9, 12, 6, 8, 10, 5 (AND)	
OST transfer	STT3A/B (OR) + structural** (AND) + TUSC3/MAGT1 (OR)	
Mannosidase I	MAN1A1/A2/B1/C1 (OR)	
GnT-I	MGAT1	
Mannosidase II	MAN2A1/A2 (OR)	
GnT-II	MGAT2	Ng complex
Branching	MGAT4A/B, MGAT5 (OR)	Ng branch
Core fucosylation	FUT8	Ng core Fuc
Bisecting GlcNAc	MGAT3	Ng bisect
Sialylation	ST6GAL1/2 (OR)	Ng sialylation
O-GalNAc (mucin-type)	donors: UDP-GalNAc, UDP-Gal, CMP-Sia	
Tn initiation	GALNT1–20 (OR)	OGN Tn

continued on next page

Table 2 continued

Step	Enzymes (logic)	Metric
Core 1	C1GALT1 + COSMC (AND)	OGN Core 1
Core 2	GCNT1	OGN Core 2
Sialylation	ST3GAL1/2 (OR)	OGN sialylation

*EXT1/EXT2 form a required heteromeric complex [26]; AND (min) logic. ‡ALG13/14 form a heterocomplex; all other ALG genes act sequentially (AND); each is essential (CDG). **OST structural subunits: RPN1, RPN2, DAD1, DDOST, OSTC (all required).

2.3 Data and preprocessing

GTEx v8 [10]: Gene-level TPM values for 17,382 RNA-seq samples across 54 tissue types from 948 post-mortem donors were obtained from the GTEx Portal (open access; sample sizes per tissue in S3 Table). Gene symbols from the GCT Description field were used; duplicate symbols (multiple Ensembl IDs mapping to the same symbol) were averaged.

2.4 Statistical analysis

Kruskal–Wallis tests assessed global tissue-level variation per metric. Pairwise Wilcoxon rank-sum tests with Benjamini–Hochberg (BH) FDR correction were performed for all tissue pairs (1,225 pairs per metric). Cliff's delta was computed as a non-parametric effect size. Effect size guidelines follow Romano et al. (2006): small ≥ 0.147, medium ≥ 0.33, large ≥ 0.474 [29]. Tissues with fewer than 20 samples were excluded from pairwise comparisons (retaining 50 tissues).

Tissues with fewer than 20 samples were excluded from analyses requiring tissue-level summaries, removing Bladder, Cervix Ectocervix, Cervix Endocervix, and Fallopian Tube (retaining 50 tissues). Throughout this study, we emphasize effect sizes (Cliff's delta) alongside p-values to ensure that statistically significant differences reflect biologically meaningful magnitudes.

All p-values were corrected using the Benjamini–Hochberg procedure; a detailed multiple testing correction strategy is provided in Supplementary Methods.

2.5 Comparison with GlycoMaple-like binary approach

To compare against the threshold-based approach used by GlycoMaple [4], we reimplemented binary pathway assessment using the same TPM ≥ 1 threshold and AND logic described in Huang et al. (2021). For each pathway step, a step was considered "active" if at least one enzyme in the OR group exceeded the threshold. A sample was classified as "reachable" (binary) if all steps were active. Tissue-level binary reachability was computed as the percentage of samples classified as reachable. Spearman correlation between binary percentage and median reachability Z-score was computed across tissues.

2.6 Aging analysis

Age-related changes in reachability were assessed using Spearman correlation between donor age (bracket midpoint) and per-sample reachability scores within each tissue. Tissues with fewer than 3 age brackets containing at least 10 samples each were excluded. Correlations were classified as significant decline or increase at FDR < 0.05 (BH correction within each tissue across all 23 metrics). Tissue-cascade coherence between age-reachability and age-downstream target correlations was assessed as described in Section 2.7, Panel C.

2.7 Functional validation via downstream target correlation

To assess whether reachability scores predict biologically meaningful downstream consequences, we tested three independent glycan-signaling cascades where the glycan product has a well-characterized effect on a receptor signaling pathway:

These three cascades were selected *a priori* based on two criteria: (i) the glycan product must have a well-established, experimentally validated effect on a specific receptor signaling pathway with published mechanistic evidence, and (ii) the downstream signaling targets must be canonical transcriptional readouts that are widely used as pathway activity indicators in the literature. Other glycan-signaling relationships exist (e.g., Notch/O-fucosylation, TGF-β/HS) but were excluded because their downstream transcriptional signatures are less specific or require cell-type-specific interpretation that bulk tissue RNA-seq cannot resolve.

(i) WNT cascade: HS N- and 6-O-sulfation (HS WNT-like reachability) promotes WNT ligand binding and co-receptor engagement. Downstream targets: AXIN2, LEF1, TCF7L2, MYC, CCND1, canonical WNT/β-catenin transcriptional readouts [11, 12].

(ii) EGFR cascade: GM3 ganglioside (GM3 reachability) inhibits EGFR dimerization and activation [13, 14]. Downstream targets: AREG, EREG, EGR1 (direct EGFR-induced transcription), DUSP6, SPRY2 (EGFR/MAPK negative feedback genes widely used as pathway activity indicators [15]). Because GM3 is an inhibitor of EGFR signaling, the downstream score was sign-inverted prior to partial correlation analysis to test the expected inverse relationship (see Panel B).

(iii) Selectin cascade: sLeX (sLeX reachability) is the minimal glycan determinant for selectin-mediated leukocyte rolling. Downstream targets: ICAM1, VCAM1, ITGB2 (endothelial adhesion molecules engaged after selectin-mediated rolling), SELPLG/PSGL-1 (the primary P-selectin glycoprotein ligand that presents sLeX on its surface; note that SELPLG is a downstream target presenting the glycan, not a biosynthetic enzyme), CCL2, CXCL8 (chemokines induced by selectin-mediated NF-κB activation [16]).

For each cascade, per-sample scores were computed as follows. **Reachability** was obtained from pre-computed per-sample reachability scores. Two naive mean expression baselines were defined: (i) **Mean (all pathway genes)** — the mean Z-score of all genes in the pathway (both catalytic enzymes and substrate supply genes), used as the comparator in Panel A; and (ii) **Mean (core catalytic only)** — the mean Z-score of core catalytic enzymes only (glycosyltransferases and modification enzymes), uniformly excluding substrate supply genes, used as the control variable in the Panel B partial correlation. This two-baseline design tests whether reachability outperforms a naive average of the same gene set (Panel A), and whether it carries unique predictive information beyond an enzymatic expression baseline that excludes substrate supply information (Panel B). The core catalytic gene sets were: WNT — EXT1/2, NDST1–4, HS6ST1–3 (9 core enzymes; excluding PAPSS1/2, SLC35B2/B3); EGFR — UGCG, B4GALT5/6, ST3GAL5 (4 core transferases; excluding GNE, NANS, NANP, CMAS, SLC35A1, UGP2, GALE, SLC35A2); Selectin — B4GALT1–6, ST3GAL3/4/6, FUT3/5/6 (12 core transferases; excluding 13 nucleotide sugar synthesis/transport genes).

Analyses were performed at the tissue level using median values across samples within each tissue ($n = 50$ tissues with at least 20 samples).

Panel A (tissue-level correlation): Spearman correlations between each predictor (reachability or naive mean expression) and downstream target score were computed independently for each cascade.

Panel B (partial correlation): The unique predictive contribution of reachability beyond naive mean expression was assessed using two complementary approaches. (1) Spearman partial correlation was computed using the ppcor R package [17], testing whether reachability correlates with downstream activity after controlling for mean expression. For the EGFR cascade, the downstream score was sign-inverted prior to partial correlation to reflect the expected inhibitory relationship between GM3 and EGFR signaling. (2) Nested linear model F-test: a base model

(downstream $\sim$ mean expression) was compared against a full model (downstream $\sim$ mean expression + reachability) using ANOVA F-test; the incremental variance explained (ΔR^2) was reported. These two tests are methodologically independent: the Spearman partial correlation assesses rank-based (potentially nonlinear) association, while the F-test assesses linear incremental predictive power.

Panel C (aging cascade coherence): For each tissue-cascade pair, Spearman correlations of age (bracket midpoint) with reachability and with downstream score were computed at the individual sample level. Tissue-cascade pairs with fewer than 3 age brackets containing at least 10 samples were excluded, yielding 144 evaluable pairs. Concordance between age-reachability and age-downstream correlations was assessed across tissue-cascade pairs.

Multiple testing correction: All p-values within each panel were adjusted using the Benjamini–Hochberg procedure. Panel A: 6 tests (3 cascades $\times$ 2 predictors). Panel B: 3 tests per test type (Spearman partial correlation and F-test corrected separately). Panel C: 144 tests per correlation type.

Additional robustness analyses (permutation tests, bootstrap confidence intervals, nonlinearity assessment, and sensitivity to pseudocount and normalization choices) are described in Supplementary Methods. Methods for the aggregation function comparison, GlycoMaple KO validation, and additional validation analyses are also provided in Supplementary Methods.

3 Results

3.1 Glycosylation gene expression varies dramatically across human tissues

To establish the transcriptomic landscape underlying tissue-specific glycan biosynthesis, we surveyed the expression of 97 glycosylation-related genes — spanning 28 functional categories including donor substrate synthesis/transport, glycosyltransferases, glycosidases, and sulfotransferases — across all 54 GTEx tissue types (Fig 1; S1 Fig; S2 Fig).

Hierarchical clustering of median Z-scores at the pathway-category level revealed pronounced tissue-specific expression programs (S1 Fig A). Brain regions formed a tight cluster characterized by elevated ganglioside synthase expression but reduced donor substrate and HS-related gene expression. Connective tissue-rich organs (nerve, fibroblasts, ovary) showed the highest HS-related expression. Gastrointestinal tissues clustered together with elevated sLeX pathway components, consistent with mucosal glycan biology. Whole blood, skeletal muscle, and heart consistently ranked lowest across most categories, reflecting the reduced glycosylation machinery in terminally differentiated cells.

This landscape provides the biological premise for our central question: can quantitative integration of these expression patterns infer tissue-specific glycan biosynthetic potential more accurately than binary threshold approaches?

3.2 Reachability analysis: A bottleneck-aware quantitative framework

We developed glycan reachability analysis, a method that integrates per-gene Z-scores across biosynthetic pathway steps using AND/OR logic and a minimum-aggregation (bottleneck) principle (Fig 1F). For each sample, the reachability score of a glycan pathway is defined as the minimum Z-score across all required enzymatic steps and donor substrate pathways (see Methods). This design captures the heuristic that the least-expressed pathway component constrains overall transcriptomic biosynthetic potential. While metabolic control analysis demonstrates that flux control is distributed rather than concentrated at a single step, the min-aggregation provides a useful first-order approximation that is empirically validated in Section 3.8.

We computed 23 reachability metrics spanning five glycan families for all 17,382 GTEx samples. All metrics showed highly significant tissue-level variation (Kruskal–Wallis $p < 10^{-300}$ for all; Table 3). For the 14 primary metrics (sLeX, gangliosides including GM2, and HS),

chi-squared statistics ranged from 9,458 (GD3) to 14,649 (HS polymerization). We note that reachability scores are not directly comparable across pathways of different lengths, because pathways with more steps produce systematically lower minimum Z-scores (expected minimum $\approx -\sqrt{2 \ln k}$ for k steps); all cross-tissue comparisons in this study are therefore performed within each metric.

Hierarchical clustering of tissue median reachability scores (S10 Fig) revealed biologically coherent groupings that differed from and complemented the raw expression landscape in S1 Fig A. While the expression landscape shows which genes are transcribed, the reachability heatmap reveals which biosynthetic pathways have expression-complete profiles. For example, brain regions express ganglioside synthases at high levels (S1 Fig A) but show only moderate ganglioside reachability (S10 Fig), because the bottleneck principle identifies donor substrate availability as the potential rate-limiting factor (see Section 3.4).

3.3 Reachability provides quantitative resolution complementary to binary approaches

Binary threshold methods such as GlycoMaple [4] address a fundamentally different question — "can this tissue produce glycan X?" — whereas reachability analysis asks "how does this tissue's transcriptomic potential compare to other tissues?" These approaches are complementary, not competitive, and we compare them here to illustrate the additional information provided by quantitative scoring.

We compared reachability scores against a GlycoMaple-like binary threshold approach (TPM ≥ 1 per enzyme) across 50 tissues with sufficient sample size ($n \geq 20$).

For sLeX, binary positive rates ranged from $< 1\%$ (whole blood, skeletal muscle) to 100% (minor salivary gland, esophagus mucosa), while reachability Z-scores ranged from -2.56 to -0.11. The two measures showed a weak positive correlation that did not reach significance (Spearman $\rho = 0.259$, $p = 0.070$; Fig 2A), indicating that binary detection and quantitative reachability capture partially overlapping but distinct information.

The most striking discordance was observed in pancreas: 96% of samples had all sLeX pathway enzymes above the 1 TPM threshold (binary: "capable"), yet the median reachability Z-score was -1.86 (among the lowest of all tissues). This occurs because each enzyme is individually detectable but expressed at uniformly low levels; the min-aggregation across all steps amplifies these individually modest deficits into a substantially reduced biosynthetic potential score.

The well-documented clinical use of sLeX-related antigens (CA19-9/sialyl Lewis A) as pancreatic cancer biomarkers, with their marked elevation in malignant versus normal tissue [33], is consistent with our prediction that normal pancreas has low basal sLeX biosynthetic capacity. However, this case illustrates a key limitation of bulk tissue analysis: pancreatic ductal cells, although a minority population ($< 5\%$ of pancreatic mass), are known to express sLeX at higher levels than acinar cells, and their signal is likely masked in bulk RNA-seq data dominated by acinar transcripts. The low reachability score thus reflects the dominant cell type's program rather than the tissue's full biosynthetic repertoire.

Conversely, nerve tibial tissue showed only $\sim$5% binary positive rate for sLeX (due to low FUT3/5/6 expression below threshold), yet its reachability score (-0.40) ranked among the highest, driven by strong expression of upstream pathway components and donor substrate machinery.

For GM3, the binary–continuous correlation was stronger ($\rho = 0.701$, $p = 1.5 \times 10^{-8}$; Fig 2B), reflecting a simpler pathway with fewer steps where binary and quantitative measures more naturally agree. However, six brain regions showed $> 97\%$ binary positivity yet reachability in the bottom quartile, identifying donor substrate availability as the hidden bottleneck.

3.4 Identification of pathway bottlenecks

A key advantage of reachability analysis is the ability to identify which pathway step limits overall biosynthetic potential — information that is entirely absent from binary classifications.

sLeX bottleneck analysis (Fig 2C): We compared the per-step Z-scores between minor salivary gland (highest sLeX reachability, $Z = -0.11$) and pancreas (near-lowest, $Z = -1.86$). Minor salivary gland showed elevated expression across nearly all pathway components, with particularly high fucosyltransferase (FUT) and GDP-Fuc synthesis scores. In contrast, pancreas showed uniformly depressed Z-scores across all 11 steps, with the deepest deficit in UDP-Gal synthesis ($Z = -1.84$) and CMP-Sia activation by CMAS ($Z = -1.23$). This pattern — all steps detectable but none strongly expressed — is precisely what binary approaches cannot detect: a "capable but barely" biosynthetic profile.

GM3 bottleneck analysis (Fig 2D): We compared brain amygdala (high GM3 binary rate of 98% but low reachability $Z = -1.56$) against nerve tibial (100% binary, reachability $Z = -0.26$). At the step level, while ST3GAL5 (the GM3-specific synthase) showed comparable expression in both tissues (brain $Z = 0.50$ vs. nerve $Z = 0.40$), the upstream steps — UGCG (brain $Z = -1.37$ vs. nerve $Z = 0.22$) and CMP-Sia transport via SLC35A1 (brain $Z = -0.77$ vs. nerve $Z = 1.14$) — were dramatically lower in brain. The reachability framework identifies UGCG as the transcriptomic bottleneck.

Failure mode — cell-type heterogeneity: This result must be interpreted with caution. The brain is the most ganglioside-rich organ in the body, with gangliosides constituting 10–12% of neuronal membrane lipid content [32]. The apparently low UGCG expression in bulk brain RNA-seq likely reflects the dilution of neuronal transcripts by glial cells (astrocytes, oligodendrocytes, microglia), which constitute the majority of brain cells and have lower ganglioside biosynthetic activity. This illustrates a fundamental limitation of bulk tissue analysis: cell-type heterogeneity can produce misleading bottleneck predictions when the biosynthetically active cell type is a minority population. Single-cell or spatial transcriptomic resolution would be required to accurately assess neuronal ganglioside biosynthetic capacity.

3.5 Statistical framework: Quantifying tissue differences

To provide a rigorous statistical foundation for tissue-level comparisons, we performed pairwise Wilcoxon rank-sum tests across all 1,225 tissue pairs for each reachability metric.

Signed effect size heatmaps (Fig 3A,B) display $-\log_{10}(\text{FDR})$ values signed by the direction of difference for sLeX reachability and HS FGF-like reachability, respectively. These reveal the global pattern of tissue differentiation: for sLeX (Fig 3A), gastrointestinal and salivary tissues form a high-reachability block sharply separated from brain, muscle, and blood. For HS FGF-like (Fig 3B), the pattern inverts, with connective tissue-rich organs showing the highest potential and blood/muscle remaining low.

Across the 14 primary metrics, 89–94% of tissue pairs showed significant differences at FDR < 0.05, with median absolute Cliff's delta ranging from 0.53 (GD3) to 0.91 (HS polymerization) (Fig 3C). HS-related metrics generally showed larger effect sizes than glycolipid metrics, suggesting greater tissue-specificity of proteoglycan biosynthetic programs. Effect size guidelines (Cliff's delta: small ≥ 0.147, medium ≥ 0.33, large ≥ 0.474) indicate that the majority of significant tissue pairs exhibit large effect sizes, confirming that reachability scores capture biologically substantial differences rather than statistically significant but trivial variation.

3.6 Age-related changes in glycan biosynthetic potential

To assess whether glycan biosynthetic potential changes with aging, we computed Spearman correlations between donor age and per-sample reachability scores within each tissue across all 23 metrics (Fig 4).

N-glycan processing potential showed the most consistent age-related decline: Ng complex, Ng branch, and Ng bisect each showed significant decline in 13 of 48 evaluable tissues (FDR < 0.05), with median ρ ranging from -0.070 to -0.078 (Fig 4A). Ng core Fuc similarly declined in 13 tissues (median $\rho = -0.067$). These results are consistent with the known age-related reduction in N-glycan processing complexity observed in glycoproteomic studies.

Among the primary 14 metrics, HS 2-O sulfation showed the most tissues with significant age-related decline (12/48), followed by HS 6-O sulfation (11/48) and sLeX reachability (9/48). In contrast, HS 3-O sulfation was the only metric with more tissues showing significant increase (12/48) than decline (4/48), suggesting a distinct age-related regulatory program for this rare sulfation modification.

Ganglioside metrics showed mixed patterns: GM3 reach increased in 9 tissues but declined in 5, while GM1 and GD3 showed similar modest trends. The tissue-specificity of these age-related changes — varying in both direction and magnitude across organs — underscores the importance of sample-level analysis rather than aggregate summaries.

Results were essentially unchanged when controlling for sex via partial Spearman correlation (median $\Delta\rho < 0.01$ across all tissue–metric pairs). RIN and ischemic time are not available as per-sample covariates in GTEx v8 open-access data.

We emphasize that the age-related effects reported here are small in magnitude (median $|\rho| < 0.10$ for most metrics) and should be considered exploratory. Age in GTEx is confounded with post-mortem interval, cause of death, and other donor-level variables that are incompletely controlled in the open-access data. These results generate testable hypotheses but do not establish causal age-related changes in glycan biosynthetic capacity.

3.7 Reachability correlates with downstream signaling beyond naive mean expression

To assess whether reachability scores capture biologically meaningful information beyond simple gene expression averages, we tested three independent glycan-signaling cascades at the tissue level ($n = 50$ tissues): WNT (HS sulfation $\rightarrow$ WNT/β-catenin targets), EGFR (GM3 ganglioside $\rightarrow$ EGFR activity, inverse relationship), and Selectin (sLeX $\rightarrow$ selectin-mediated adhesion/inflammation targets). We note that these tissue-level correlations do not establish causality, and that confounding by tissue identity remains a possibility (see Limitations). Reachability scores also correlated with cognate receptor/ligand expression across tissues (S3 Fig), providing additional support for biological coherence.

Reachability outperforms naive mean expression (Fig 5A). At the tissue level ($n = 50$), reachability scores showed consistently stronger Spearman correlations with downstream target expression than naive mean expression of all pathway genes: WNT $\rho = 0.83$ vs. 0.72, EGFR $|\rho| = 0.66$ vs. 0.47, Selectin $\rho = 0.53$ vs. 0.45 (all $p_{\mathrm{BH}} < 0.002$; Fig 5A). The improvement was largest for EGFR ($\Delta\rho = 0.18$), where the multi-step ganglioside pathway includes upstream ceramide synthesis and CMP-sialic acid supply — bottleneck information that is absent from a simple transferase average.

Size-matched permutation testing (10,000 random gene sets) yielded permutation p-values: WNT $p = 0.0003$, EGFR $p = 0.0012$, Selectin $p = 0.0041$. However, when using variance-matched permutation (gene sets drawn with probability weights proportional to cross-tissue expression variance similarity), only WNT remained significant ($p = 0.013$), while EGFR ($p = 0.49$) and Selectin ($p = 0.80$) did not. This indicates that the EGFR and Selectin correlations may partly reflect the general tendency of tissue-specific genes to co-vary across tissues, rather than glycan-specific biological signal. Importantly, the WNT cascade — which shows the strongest correlation and the largest improvement over mean expression — survived this stringent control, supporting a genuine biological relationship. The EGFR and Selectin cascades should therefore be interpreted as suggestive but not definitive.

Bootstrap 95% confidence intervals (2,000 iterations) for the reachability-downstream correlations were: WNT $\rho = 0.83$ [0.72, 0.90], EGFR $|\rho| = 0.66$ [0.46, 0.80], Selectin $\rho = 0.53$ [0.29, 0.71]. Donor-level block bootstrap (resampling donors rather than tissues to account for non-independence) yielded tighter intervals due to the large donor overlap ($\sim$139 shared donors per tissue pair): WNT [0.83, 0.85], EGFR [0.64, 0.68], Selectin [0.50, 0.55].

Within-tissue sample-level validation. To address the ecological fallacy concern, we computed within-tissue Spearman correlations between per-sample reachability and per-sample downstream target expression. The within-tissue correlations were weaker than between-tissue correlations (as expected given the narrower dynamic range) but consistently positive: WNT median $\rho = 0.54$ (48/50 tissues significant at BH < 0.05), EGFR median $\rho = -0.36$ (45/50), Selectin median $\rho = 0.33$ (41/50). These within-tissue correlations provide evidence that the associations are not solely driven by tissue identity confounding.

Tissue-subtype collapsing. To estimate the effective sample size, we collapsed 50 tissues to 30 organ-level representatives and recomputed correlations. Results remained significant: WNT $\rho = 0.84$ ($p = 5 \times 10^{-9}$), EGFR $|\rho| = 0.67$ ($p = 6 \times 10^{-5}$), Selectin $\rho = 0.42$ ($p = 0.022$).

Formal test of improvement. Steiger's test for dependent correlations confirmed that reachability significantly outperformed mean expression: WNT $\Delta\rho = 0.31$ ($p = 0.0002$), EGFR $\Delta|\rho| = 0.34$ ($p < 0.0001$), Selectin $\Delta\rho = 0.18$ ($p = 0.027$).

Restricted downstream gene sensitivity. Using only the most pathway-specific downstream targets (WNT: AXIN2 and LEF1; EGFR: AREG, EREG, EGR1; Selectin: SELPLG, ITGB2, ICAM1) yielded significant but attenuated correlations: WNT $\rho = 0.51$, EGFR $|\rho| = 0.49$, Selectin $\rho = 0.51$ (all $p < 0.001$), confirming that the results are not driven solely by broadly regulated genes such as MYC and CCND1.

Sensitivity to tissue clustering. Leave-one-tissue-group-out analysis confirmed that no single tissue group drives the validation results. Removing all brain tissues ($n = 13$), gastrointestinal tissues ($n = 7$), or reproductive tissues ($n = 5$) reduced sample sizes but correlations remained significant (WNT: $\rho = 0.78$–0.81; EGFR: $|\rho| = 0.58$–0.64; Selectin: $\rho = 0.44$–0.51; all $p < 0.01$). Block bootstrap confidence intervals (resampling tissue groups rather than individual tissues) were wider but remained significant: WNT reachability ρ 95% CI $[0.65, 0.89]$ (vs. $[0.73, 0.90]$ for standard bootstrap).

Reachability carries unique predictive information (Fig 5B). We next asked whether reachability predicts downstream targets after controlling for mean expression, using Spearman partial correlation and nested model F-tests. For WNT and EGFR, both tests confirmed significant unique contributions: WNT partial $\rho = 0.76$ ($p_{\mathrm{BH}} = 8.6 \times 10^{-10}$, $\Delta R^2 = 0.46$, F-test $p_{\mathrm{BH}} = 1.2 \times 10^{-11}$); EGFR partial $\rho = -0.69$ ($p_{\mathrm{BH}} = 4.9 \times 10^{-8}$, $\Delta R^2 = 0.52$, F-test $p_{\mathrm{BH}} = 1.8 \times 10^{-9}$). For EGFR, the downstream score was sign-inverted to test the expected GM3-mediated inhibition of EGFR dimerization [13]; the significant partial correlation in the expected direction indicates that GM3 biosynthetic potential carries unique information about EGFR pathway modulation beyond what core transferase expression alone provides.

For Selectin, the Spearman partial correlation was nominally significant (partial $\rho = 0.44$, $p_{\mathrm{BH}} = 1.5 \times 10^{-3}$), but the linear F-test showed negligible incremental variance explained ($\Delta R^2 = 0.03$, $p_{\mathrm{BH}} = 0.24$). The discrepancy between the rank-based and linear tests, combined with the small ΔR^2, indicates that the unique contribution of sLeX reachability beyond core enzyme expression is weak for the Selectin cascade. This likely reflects both biological and methodological factors: (i) sLeX-mediated selectin binding is a necessary but not sufficient condition for leukocyte adhesion, which also requires chemokine receptor engagement and integrin activation [16]; (ii) reachability and mean expression are highly collinear for this pathway ($R^2 = 0.67$; S6 Fig), making it difficult to separate their contributions with $n = 50$; and (iii) the cellular heterogeneity of bulk tissue data, where sLeX-producing cells and downstream target-expressing cells are distinct populations. We consider the Selectin cascade evidence as weak and exploratory, in contrast to the stronger WNT and EGFR results.

Nonlinear control. To confirm that the unique contribution of reachability is not an artifact of nonlinearity, we re-computed partial correlations controlling for a natural cubic spline (3 df) of mean expression. Results were similar: WNT partial $\rho = 0.72$ (vs. 0.76 with linear control), EGFR partial $|\rho| = 0.65$ (vs. 0.69), Selectin partial $\rho = 0.39$ (vs. 0.44), confirming that the unique contribution reflects genuinely independent biological information.

Age-related reachability changes predict downstream target changes (Fig 5C). If reachability captures genuine biosynthetic capacity, then age-related changes in reachability should predict concordant changes in downstream signaling. Across 144 tissue-cascade pairs, age-reachability correlations ($\rho_{\text{age,reach}}$) and age-downstream correlations ($\rho_{\text{age,down}}$) showed significant positive concordance, confirming that tissues where reachability declines with age also show declining downstream target expression. This coherence provides longitudinal support for the cross-sectional associations observed in Panels A and B.

3.8 Min-aggregation outperforms mean expression across all validation cascades

A central design choice in reachability analysis is the use of min-aggregation (bottleneck principle) rather than simple averaging across pathway genes. To evaluate whether this choice is justified, we systematically compared four aggregation methods across the three validation cascades (Fig 6).

Full correlations (Fig 6A). Min-aggregation (bottleneck) consistently produced the strongest correlations with downstream targets: WNT $\rho = 0.85$, |EGFR ρ| $= 0.72$, Selectin $\rho = 0.56$. The mean of all pathway genes was weaker in every case: WNT $\rho = 0.72$ ($\Delta = -0.14$), |EGFR ρ| $= 0.47$ ($\Delta = -0.25$), Selectin $\rho = 0.45$ ($\Delta = -0.12$). The mean of core catalytic enzymes only was weaker still: WNT $\rho = 0.52$, |EGFR ρ| $= 0.32$, Selectin $\rho = 0.35$. Interestingly, replacing the OR=mean aggregation with OR=max produced nearly identical results (WNT $\rho = 0.87$, |EGFR ρ| $= 0.72$, Selectin $\rho = 0.59$), indicating that the key driver of improved prediction is the AND=min (bottleneck) logic, not the specific choice of OR aggregation. (Note: the reach values here differ slightly from Section 3.7 because the aggregation comparison recomputes Z-scores internally across the analyzed samples for a fair comparison of methods, while Section 3.7 uses the canonical pre-computed per-sample reachability scores.)

Partial correlations controlling for core expression (Fig 6B). To assess unique information content, we computed Spearman partial correlations controlling for the core-only mean. Min-aggregation maintained strong unique contributions across all cascades: WNT partial $\rho = 0.79$ ($p_{\text{BH}} = 4.1 \times 10^{-11}$), |EGFR partial ρ| $= 0.74$ ($p_{\text{BH}} = 2.2 \times 10^{-9}$), Selectin partial $\rho = 0.51$ ($p_{\text{BH}} = 2.5 \times 10^{-4}$). In contrast, the mean of all genes showed substantially weaker unique information: WNT partial $\rho = 0.65$, |EGFR partial ρ| $= 0.40$, Selectin partial $\rho = 0.37$ (all $p_{\text{BH}} < 0.01$). These results confirm that min-aggregation captures bottleneck information from substrate supply pathways that is absent from any averaging approach.

Near-optimality of min-aggregation. Systematic variation of the AND-logic aggregation function from min (quantile 0) through median to max (quantile 1.0) revealed that performance is highest in the low-quantile range and degrades monotonically as the aggregation approaches the maximum (S7 Fig). Specifically, WNT performance peaked at quantile 0.20–0.25 ($\rho = 0.867$) rather than at the strict minimum ($\rho = 0.831$); EGFR peaked at quantile 0 ($|\rho| = 0.718$); Selectin peaked at quantile 0 ($\rho = 0.564$). The performance difference between quantile 0 and 0.25 was modest for all cascades ($|\Delta\rho| \leq 0.04$), and all cascades showed a steep decline beyond quantile 0.5 (median). These results support the use of min-aggregation as a good practical default rather than a universally optimal choice: for WNT, quantile 0.20–0.25 was marginally better ($\Delta\rho = 0.04$), suggesting that a slightly relaxed bottleneck may better capture pathways where multiple near-limiting steps contribute. We retain the strict minimum for simplicity and interpretability.

3.9 Pathway topology validated using published knockout data

Pathway topology was validated using published HEK293 knockout glycomics data [4] (S11 Fig; Supplementary Methods). Direction prediction accuracy was 76–100% for early-pathway KOs (MGAT1, MGAT2, SLC35C1, MAN1A1A2B1 triple), confirming that the pathway definitions correctly capture gene–glycan relationships. We note that these direction predictions rely primarily on the pathway topology rather than on the min-aggregation quantitative scoring per se; any method encoding the correct pathway logic would make similar qualitative predictions. The KO analysis therefore validates the pathway definitions, not the bottleneck aggregation principle.

3.10 Supplementary validation: DepMap essentiality and cross-species glycomics

Two additional validation approaches were attempted but yielded negative results, which we report transparently for completeness.

DepMap essentiality analysis (S4 Fig). We tested whether glycosylation genes predicted as pathway bottlenecks show greater essentiality (more negative Chronos scores) in cancer cell lines from the DepMap/Project Achilles dataset [23]. Across the three validation cascades, only 2 of 23 tested bottleneck genes showed significantly different essentiality between bottleneck and non-bottleneck cell lines after BH correction: HS6ST1 ($p_{\text{adj}} = 0.018$) and HS6ST2 ($p_{\text{adj}} = 0.012$) in the WNT/HS cascade. The overall median Chronos scores for bottleneck vs. non-bottleneck cell lines were not meaningfully different for any cascade. This negative result is expected: glycan biosynthesis is generally non-essential for cancer cell proliferation in standard 2D culture conditions, as essentiality screens primarily detect growth/survival phenotypes rather than glycan-dependent functions such as cell-cell adhesion, immune recognition, and signaling modulation.

Cross-species tissue glycomics comparison (S5 Fig). We compared predicted N-glycan reachability scores against published tissue-level glycomic profiles from mouse [37] ($n = 15$ matched tissue types). None of the four glycan features tested showed significant correlation with reachability predictions: complex N-glycans ($\rho = 0.16$, $p = 0.56$), branching ($\rho = -0.37$, $p = 0.17$), sialylation ($\rho = 0.15$, $p = 0.60$), and core fucosylation ($\rho = 0.21$, $p = 0.44$). Multiple factors likely contribute to this negative result: (i) species-specific differences in glycoenzyme expression and regulation between human and mouse; (ii) post-transcriptional regulation not captured by RNA-seq; (iii) limited sample size ($n = 15$) providing insufficient statistical power for detecting moderate correlations; and (iv) the min-aggregation produces identical reachability scores for multiple N-glycan metrics when a shared upstream step is rate-limiting, reducing the effective dynamic range for correlation analysis.

4 Discussion

4.1 Relationship to existing methods

Glycan reachability analysis occupies a distinct methodological niche compared to existing computational glycobiology tools (Table 2). Unlike GlycoMaple [4], which answers "can this tissue produce glycan X?" (binary), our method answers "how does this tissue's transcriptomic potential to produce glycan X compare to other tissues?" (quantitative, relative). Unlike glycoPATH [7], which predicts absolute glycan abundances via supervised machine learning requiring paired glycomic–transcriptomic training data, our method is unsupervised and requires only RNA-seq data. Unlike kinetic simulation approaches (GNAT, GlycoMME, GlyCompute), which require kinetic rate parameters rarely available for most biological systems, our method relies on the biologically motivated assumption that the least-expressed pathway component limits overall pathway potential.

The bottleneck (min-aggregation) principle is a heuristic motivated by the observation that, in linear-regime enzyme kinetics (substrate concentration well below K_m), reaction rate scales with enzyme concentration and the least-abundant enzyme constrains overall pathway throughput. We explicitly note that this is a simplification: metabolic control analysis (MCA) demonstrates that flux control is distributed among multiple enzymes, with flux control coefficients summing to unity [30, 31]. Our approach does not claim that a single step has complete flux control; rather, it uses the minimum expression as a proxy for the overall transcriptomic capacity of the pathway. The empirical validation (Fig 6, S7 Fig) supports this heuristic: min-aggregation outperforms mean-based alternatives for predicting downstream biological correlates, though the optimal quantile is near but not always exactly at the minimum (see below).

While conceptually related to constraint-based metabolic modeling approaches such as FBA [19], our method differs fundamentally in scope. FBA operates on genome-scale stoichiometric matrices with mass conservation constraints and linear programming. Expression-constrained approaches such as iMAT [20] and GIMME [21] share our goal of integrating transcriptomic data with pathway knowledge, but require genome-scale metabolic models. The trade-off is explicit: our method sacrifices mechanistic detail for simplicity, interpretability, and broad applicability to any pathway definable with AND/OR logic.

Our systematic comparison of aggregation functions (Fig 6) provides direct empirical support for the min-aggregation choice: across all three independent validation cascades, bottleneck-based scores outperformed naive mean expression of all pathway genes, which in turn outperformed the mean of core catalytic enzymes only. This result confirms that substrate supply chain information (captured by min-aggregation across donor pathway steps) contributes meaningful predictive power beyond simple enzymatic expression.

4.2 Biological interpretation of discordant cases

The case studies below illustrate two systematic failure modes of bulk tissue reachability analysis — cell-type heterogeneity masking minority populations, and cell-type composition differences mimicking biosynthetic differences — that define the interpretive boundaries of our approach.

The pancreas case (Fig 2C) exemplifies where quantitative assessment provides additional resolution. While a binary approach classifies pancreas as capable of sLeX synthesis (96% of samples have all enzymes detectable), the quantitative reachability score ($Z = -1.86$) indicates uniformly low expression across all pathway steps. The exocrine pancreas, which dominates the bulk tissue transcriptome, primarily produces digestive enzymes and maintains glycosylation machinery at basal levels. We note, however, that pancreatic ductal cells — a minority population — are known to express sLeX and related antigens at higher levels, and this cell-type heterogeneity limits the interpretation of bulk tissue predictions for the pancreas.

The brain ganglioside case (Fig 2D) illustrates both the utility and limitations of the reachability framework. The model identifies upstream steps (UGCG, CMP-Sia transport) as transcriptomic bottlenecks in brain tissue. However, as noted in Section 3.4, the brain is the most ganglioside-rich organ in the body [32], which means the low reachability score likely reflects an artifact of bulk tissue heterogeneity rather than a genuine biosynthetic limitation. Ganglioside synthesis is concentrated in neurons, which are a minority of total brain cells; the bulk tissue signal is dominated by glial transcripts. This case serves as a cautionary example: reachability analysis is most reliable when the biosynthetically active cell population is well-represented in the bulk tissue profile. Application to single-cell or spatially resolved transcriptomic data would address this limitation.

4.3 Negative validation results and limitations

We have reported two supplementary validation analyses that yielded negative results (Section 3.10). The DepMap analysis (S4 Fig) was based on the premise that bottleneck genes

should show greater essentiality, but this premise does not hold for glycosylation in standard cancer cell culture: glycan-dependent phenotypes (adhesion, immune evasion, signaling modulation) are not captured by proliferation-based essentiality screens. The cross-species glycomics comparison (S5 Fig) was limited by sample size ($n = 15$), species differences, and the structural constraint that min-aggregation can produce identical reachability scores for multiple N-glycan features when upstream processing is rate-limiting. These negative results define the boundary conditions of our approach: reachability analysis infers relative biosynthetic potential from transcriptomics, but does not predict absolute glycan abundance, cell-line essentiality, or cross-species glycomic profiles.

The primary limitation of our validation framework is transcriptomic circularity: both predictors and validation readouts are derived from RNA-seq, so our analyses cannot establish that glycan products mediate the observed associations. Direct glycomic measurements across matched tissues would be needed to confirm this causal chain. Furthermore, the tissue-level validation ($n = 50$) is susceptible to ecological fallacy — tissues differing in reachability also differ in many other transcriptomic programs. The within-tissue sample-level correlations (Section 3.7) partially address this concern, but single-cell validation would provide stronger support. The 17,382 GTEx samples derive from only 948 donors with substantial tissue overlap (mean ≈ 139 shared donors per tissue pair), raising the concern of donor-level pseudoreplication. Within each tissue, each donor contributes exactly one sample, so within-tissue comparisons are unaffected. For the tissue-level validation (Fig 5), where the same donors contribute to multiple tissue medians, we performed donor-level block bootstrap (2,000 iterations resampling donors with replacement and reconstructing tissue medians from each resampled donor set). The resulting 95% CIs remained tight and excluded zero for all three cascades: WNT ρ [0.83, 0.85], EGFR $|\rho|$ [0.64, 0.68], Selectin ρ [0.50, 0.55] (S2 Table). For the pairwise Wilcoxon tissue-difference tests, within each tissue each donor contributes exactly one sample, so individual test statistics are computed on non-overlapping sample sets. However, donors contributing to both tissue groups introduce a residual non-independence that is not fully captured by our within-tissue analysis. Donor-aggregated sensitivity analysis (computing donor medians per tissue) confirmed identical results across all 23 metrics ($> 89\%$ significant tissue pairs, negligible change in Cliff's delta), but we emphasize that the large effect sizes (median |Cliff's delta| = 0.54–0.90) provide stronger evidence of genuine tissue differences than p-values alone, which may be anti-conservative due to the shared-donor structure.

Gene expression is an imperfect proxy for enzymatic activity: post-translational regulation, protein stability, and subcellular localization are not captured. Reachability scores should therefore be interpreted as transcriptomic biosynthetic potential rather than as measures of enzymatic activity or glycan abundance. The model does not account for glycosidase-mediated trimming, though our glycosidase survey (S8 Fig) revealed strong positive correlations between biosynthetic and catabolic enzymes (e.g., NEU1 vs. GM3 reach $\rho = 0.78$; MAN2B1 vs. Ng complex $\rho = 0.82$), indicating coordinated glycan turnover rather than confounding. Golgi compartmentalization, post-mortem RNA degradation effects [22], and bulk tissue heterogeneity (which can mask cell-type-specific biosynthetic programs, as illustrated by the brain ganglioside case in Section 3.4) represent additional sources of noise.

The current implementation covers five glycan families; expansion requires additional curation. Within the HS pathway, the HS3ST isoforms (HS3ST1–6) are modeled as a single OR group despite having distinct substrate specificities [28], and the 6-O endosulfatases SULF1/SULF2, which modulate WNT ligand binding [25], are omitted. The model also omits inter-pathway competition for shared nucleotide sugar donors (CMP-Sia, UDP-Gal, GDP-Fuc), two donor pathways (UDP-GlcNAc, UDP-GlcA), the HS linker tetrasaccharide (XYLT1/2, B4GALT7, B3GALT6, B3GAT3), and acceptor protein availability (e.g., PSGL-1 for sLeX, syndecans/glypicans for HS). Finally, Z-scores are relative to the GTEx population, making scores dataset-dependent, and min-aggregation produces systematically negative scores for pathways with many steps, pre-

cluding direct cross-pathway comparison (see Section 3.2 for details).

4.4 Perspectives

Many of the predictions generated by reachability analysis are directly testable with existing experimental platforms. The most immediate validation would be multi-tissue glycomic profiling by mass spectrometry: our model predicts, for example, that pancreas has low sLeX abundance despite near-universal enzyme detectability (Fig 2C), and that brain ganglioside levels are not limited by ST3GAL5 but by upstream UGCG and CMP-Sia transport (Fig 2D). Tissue-matched glycomic–transcriptomic datasets, though currently scarce, would allow direct evaluation of whether reachability scores predict relative glycan abundance across tissues. Similarly, the predicted age-related decline in N-glycan processing capacity (Fig 4) generates a testable hypothesis that aged tissues show reduced N-glycan complexity, a trend for which preliminary glycoproteomic evidence exists but systematic multi-tissue data are lacking.

The bottleneck identification capability suggests a strategy for targeted perturbation experiments. If the model correctly identifies the expression-limiting step, then CRISPR-mediated upregulation of the predicted bottleneck gene should produce a larger increase in pathway output than upregulation of non-bottleneck components. The HEK293 knockout analysis (S11 Fig) provides initial support for the pathway topology, though it validates pathway definitions rather than the quantitative bottleneck scoring; activation experiments using CRISPRa in primary cells or organoids would offer a more direct test of the bottleneck principle.

Beyond validation, the framework enables several classes of biological inquiry. Applied to disease transcriptomes, reachability analysis could identify pathways whose biosynthetic potential shifts between healthy and pathological states — for instance, whether tumor-associated glycan changes (e.g., increased sLeX, altered branching) reflect transcriptomic reprogramming of specific bottleneck steps. Application to single-cell atlases would resolve the bulk tissue heterogeneity limitation and reveal cell-type-specific glycan biosynthetic programs, addressing cases such as the brain ganglioside paradox described above. Integration with spatial transcriptomics could further map glycan biosynthetic potential onto tissue architecture, linking reachability gradients to functional microenvironments. An R package implementing all methods described here is available at https://github.com/matsui-lab/glycoreach.

Tables

Table 3: Kruskal–Wallis tests for tissue-level variation in reachability scores (14 primary metrics). All p-values are below machine precision ($< 2.2 \times 10^{-308}$).

Metric	χ^2	df	p-value
sLeX reachability	10,128	53	$< 2.2 \times 10^{-308}$
GM3 reach	9,980	53	$< 2.2 \times 10^{-308}$
GM2 reach	10,502	53	$< 2.2 \times 10^{-308}$
GM1 reach	11,330	53	$< 2.2 \times 10^{-308}$
GD3 reach	9,458	53	$< 2.2 \times 10^{-308}$
HS polymerization*	14,649	53	$< 2.2 \times 10^{-308}$
PAPS	13,113	53	$< 2.2 \times 10^{-308}$
HS N-sulfation	10,991	53	$< 2.2 \times 10^{-308}$
HS 2-O sulfation	10,407	53	$< 2.2 \times 10^{-308}$
HS 6-O sulfation	11,561	53	$< 2.2 \times 10^{-308}$
HS 3-O sulfation	13,125	53	$< 2.2 \times 10^{-308}$
HS FGF-like	13,109	53	$< 2.2 \times 10^{-308}$
HS WNT-like	12,673	53	$< 2.2 \times 10^{-308}$
HS SHH-like**	12,398	53	$< 2.2 \times 10^{-308}$

*Uses AND (min) for EXT1/EXT2 heteromeric complex. **Includes 3-O sulfation requirement (differs from WNT-like).

Table 4: Comparison with existing computational glycobiology methods.

Method	Input	Output	Quantitative?	Pathway logic	Training data	Multi-pathway
GlycoMaple [4]	RNA-seq TPM	Binary structure prediction	No (TPM $\geq$ 1)	AND (sequential)	No	Yes (21)
Glycologue [6]	Active enzyme set	Predicted products	No	Rule-based	No	Yes
glycoPATH [7]	RNA-seq + glycomics	N-glycan abundances	Yes (regression)	ML-learned	Yes (paired)	No (N-glycan)
GNAT/GlycoMME	Kinetic params	Glycan distributions	Yes (simulation)	Kinetic model	Yes (rate constants)	Limited
Glycopacity [8]	scRNA-seq	Glycogene scores	Semi	No (gene-level)	No	No
glycoCARTA [9]	scRNA-seq	Glycogene modules	Semi	No (correlation)	No	Limited
Reachability	**RNA-seq TPM**	**Continuous potential score**	**Yes (Z-score)**	**AND/OR + min**	**No***	**Yes**

*No training data in the supervised/statistical learning sense. However, pathway definitions encode expert-curated biological knowledge (enzyme groups, AND/OR logic assignments, donor substrate chains) that represents the primary source of domain-specific input.

Table 5: Comprehensive validation statistics with bootstrap confidence intervals and permutation p-values.

Cascade	Metric	ρ	95% CI (tissue bootstrap)	95% CI (tissue-group block)	Perm. p (size-m
WNT	Full ρ (reach)	0.83	[0.73, 0.90]	[0.65, 0.89]	0.0003
WNT	Full ρ (mean)	0.72	[0.55, 0.83]	—	—
WNT	Partial ρ	0.76	[0.58, 0.87]	—	—
EGFR	Full $\lvert\rho\rvert$ (reach)	0.66	[0.47, 0.79]	[0.38, 0.81]	0.0012
EGFR	Full $\lvert\rho\rvert$ (mean)	0.47	[0.22, 0.66]	—	—
EGFR	Partial $\lvert\rho\rvert$	0.69	[0.48, 0.82]	—	—
Selectin	Full ρ (reach)	0.53	[0.31, 0.70]	[0.24, 0.72]	0.0041
Selectin	Full ρ (mean)	0.45	[0.21, 0.64]	—	—
Selectin	Partial ρ	0.44	[0.18, 0.63]	—	—

Supplementary Methods

Isozyme grouping and OR aggregation sensitivity

We note that grouping enzymes as "isozymes" (OR logic) is a simplification: many grouped enzymes have overlapping but non-identical substrate specificities. For example, in the sLeX pathway, B4GALT1 is the dominant glycoprotein LacNAc synthase, while B4GALT5/6 preferentially act on glycolipid substrates. We include B4GALT1–6 in the sLeX OR group for completeness, but a sensitivity analysis restricting this group to B4GALT1–4 (glycoprotein-relevant members) yielded nearly identical scores (Spearman $\rho = 0.999$ with original) and identical validation results. Even restricting to B4GALT1 alone (the dominant glycoprotein LacNAc synthase) yielded $\rho = 0.957$ with original scores and slightly improved downstream correlation ($\rho = 0.62$ vs. 0.56), confirming that the results are robust to isozyme grouping choices (S2 Table). Similarly, MGAT4A/B (tri-antennary branching) and MGAT5 (tetra-antennary branching) catalyze distinct reactions; separating them into individual metrics is biologically more accurate and is provided as an alternative output. These simplifications are made explicit to facilitate informed interpretation.

We chose mean rather than max aggregation for isozymes based on three considerations. First, many isozyme families (e.g., B4GALT1–6 for LacNAc synthesis, ST3GAL3/4/6 for α2,3-sialylation) comprise enzymes with overlapping but distinct substrate specificities; co-expression of multiple family members can increase the total enzymatic potential beyond that of the single highest-expressed member. Second, the mean is more robust to outlier expression of a single family member than the max, reducing the risk of overestimating potential when one isozyme is spuriously highly expressed. Third, our systematic comparison of OR=mean versus OR=max aggregation across three independent validation cascades (Fig 6) demonstrates that the two produce nearly identical results ($\Delta\rho < 0.02$ across all cascades), confirming that the AND=min (bottleneck) logic—not the OR aggregation function—drives the predictive improvement over averaging approaches. We note that this OR aggregation choice applies only to isozyme groups within a single reaction step; multi-subunit complexes (EXT1/EXT2, ALG13/14, OST structural subunits) use AND logic, as all subunits are required for function.

Multiple testing correction strategy

All p-values are corrected using the Benjamini–Hochberg (BH) procedure within each analysis (S2 Table provides a complete enumeration of all statistical tests performed, totaling approximately 18,700 tests). The vast majority (17,150 = 1,225 pairs $\times$ 14 primary metrics) are pairwise Wilcoxon tests corrected within each metric. Excluding these, approximately 1,575 independent tests were performed across all other analyses. We adopt a hierarchical correction strategy: BH correction within each analysis panel, with paper-level interpretation guided by a conservative Bonferroni threshold ($\alpha/1575 \approx 3.2 \times 10^{-5}$). Under this threshold, the WNT and EGFR validation results (full and partial correlations) and all Kruskal–Wallis tests survive correction. The Selectin partial correlation and the aging analyses are interpreted as exploratory.

Validation robustness analyses

Permutation test for validation specificity: To confirm that the observed correlations reflect genuine biological relationships rather than general tissue-level gene expression co-variation, we performed two complementary permutation tests. (1) *Size-matched permutation:* 10,000 random gene sets of the same size as the true downstream targets were drawn uniformly from all available genes (excluding pathway and downstream genes). (2) *Variance-matched permutation:* 1,000 random gene sets were drawn with probability weights proportional to the similarity of each gene's cross-tissue expression variance to that of the true downstream targets (Gaussian kernel

on tissue-median standard deviation). The variance-matched test controls for the possibility that tissue-specific genes naturally co-vary, providing a more conservative null distribution.

Bootstrap confidence intervals: 95% bootstrap confidence intervals (2,000 iterations, tissue-level resampling) were computed for all Spearman correlations and partial correlations reported in Panels A and B.

Nonlinearity assessment: For the Selectin cascade, where the Spearman partial correlation was significant but the linear F-test was not, we fitted a generalized additive model (GAM; `mgcv` R package) with a smooth term for reachability and compared against the linear model using an F-test.

Pseudocount sensitivity analysis. To assess sensitivity to the pseudocount in the log transformation (Eq. 3), we tested values of 0.1, 0.5, 1 (default), and 5. Validation correlations were robust across all choices ($\Delta\rho < 0.03$ for all cascade–pseudocount combinations; S9 Fig).

Normalization sensitivity analysis. We compared three normalization approaches: (1) $\log(1 + \text{TPM})$ Z-score (default), (2) quantile normalization followed by Z-score, and (3) rank-based inverse normal transformation (INT). Results were robust between the default Z-score and INT (WNT $\rho = 0.83$–0.86, EGFR $|\rho| = 0.72$, Selectin $\rho = 0.55$–0.56; S9 Fig). Quantile normalization substantially degraded performance (WNT $\rho = 0.46$, EGFR $|\rho| = 0.03$), likely because it removes the between-tissue expression rank structure that reachability analysis relies upon. We therefore recommend the default $\log(1 + \text{TPM})$ Z-score or rank-based INT.

Aggregation function comparison

To evaluate whether the bottleneck (min-aggregation) principle provides better predictions than simpler averaging approaches, we compared four aggregation methods for computing tissue-level glycan biosynthetic scores across the same three validation cascades:

1. **Min-aggregation (bottleneck):** Standard reachability scores using OR=mean, AND=min logic (the default method).
2. **Min-aggregation (OR=max):** AND=min with OR=max for isozyme groups, testing sensitivity to the OR aggregation function.
3. **Mean (all genes):** Flat mean of Z-scores across all pathway genes (both catalytic enzymes and substrate supply genes).
4. **Mean (core only):** Mean of Z-scores for core catalytic enzymes only (ctrl_genes).

For each method-cascade combination, Spearman correlation with downstream target expression was computed at the tissue level ($n = 50$). Additionally, Spearman partial correlations controlling for the core-only mean were computed for methods 1–3, assessing unique predictive information beyond simple enzymatic expression.

GlycoMaple KO validation with HEK293 glycomics

To test whether reachability scores can predict the directional effect of gene knockouts on glycan profiles, we used published N-glycan mass spectrometry data from wild-type and knockout HEK293 cells [4]. Six KO conditions were analyzed: MGAT1_KO (GnT-I), TKO_MAN1A1A2B1 (triple mannosidase I knockout), SLC35C1_KO (GDP-Fuc transporter), MGAT2_KO (GnT-II), MGAT4A/B_KO, and MGAT5_KO.

For each KO, reachability was computed using the WT HEK293 expression profile with the targeted gene(s) set to zero expression. For each detected glycan, the predicted direction of change (increase, decrease, or no change) was compared against the observed direction (log2 fold-change of KO vs. WT glycan abundance). Direction accuracy was computed as the fraction of glycans with correctly predicted direction. Statistical significance was assessed using a one-sided binomial test against the null hypothesis of 50% accuracy (chance level). KO conditions with fewer than 5 testable glycans were flagged as having insufficient statistical power.

Additional validation analyses

Leave-one-tissue-group-out analysis. To assess whether correlated tissue groups (e.g., multiple brain regions) drive the validation results, we systematically removed all tissues belonging to each of three groups—brain ($n = 13$), gastrointestinal ($n = 7$), and reproductive ($n = 5$)—and recomputed tissue-level Spearman correlations for all three cascades.

Block bootstrap. To account for non-independence among related tissues, we performed block bootstrap with 2,000 iterations, resampling tissue groups (brain, artery, skin, esophagus, adipose, heart, colon, kidney, and remaining tissues as individual blocks) with replacement, and computed 95% confidence intervals.

Glycosidase expression survey. To assess whether glycan-degrading enzymes confound reachability predictions, we surveyed the expression of 12 glycosidases—categorized by subcellular localization (lysosomal: NEU1, MAN2B1/2, FUCA1, HEXA/B, GBA, GALC; plasma membrane: NEU3, NEU4; cytoplasmic: NEU2; ER/Golgi: FUCA2)—and computed Spearman correlations with corresponding reachability metrics across 50 tissues (S8 Fig).

Donor substrate demand analysis. To assess the potential impact of inter-pathway competition for shared nucleotide sugar donors, we computed aggregate demand indices for CMP-Sia, UDP-Gal, and GDP-Fuc per tissue by summing the tissue-median Z-scores of all consuming enzymes across all five pathway families (S1 Table).

Generalized aggregation analysis. To empirically evaluate the optimality of min-aggregation, we systematically varied the AND-logic aggregation function using quantile-based interpolation from min (quantile 0) through median (quantile 0.5) to max (quantile 1.0), keeping OR-logic (mean) unchanged. For each quantile level, we recomputed pathway scores and tissue-level Spearman correlations with downstream targets (S7 Fig).

Donor-level block bootstrap. To assess the impact of donor-level non-independence in GTEx (948 donors contributing to multiple tissues), we performed donor-level block bootstrap: 2,000 iterations resampling donors (not tissues) with replacement, reconstructing tissue medians from each resampled donor set, and computing Spearman correlations (S2 Table).

Tissue-subtype collapsing. To estimate the effective sample size after accounting for non-independent tissue subtypes, we collapsed the 50 tissues to approximately 30 organ-level representatives (e.g., all 13 brain regions to a single "Brain", 3 arteries to "Artery") and recomputed tissue-level correlations at this reduced sample size (S2 Table).

Within-tissue sample-level correlations. To test whether the tissue-level associations hold at the individual sample level (addressing the ecological fallacy), we computed within-tissue Spearman correlations between per-sample reachability and per-sample downstream target expression for each of the 50 tissues and each cascade (S2 Table).

Steiger's test for correlation differences. To formally test whether reachability significantly outperforms mean expression, we applied the Steiger (1980) test for comparing two dependent correlations sharing a common variable (downstream targets) (S2 Table).

Restricted downstream gene sensitivity. To assess sensitivity to the choice of downstream target genes, we repeated the primary validation using restricted gene sets containing only the most pathway-specific targets (WNT: AXIN2, LEF1; EGFR: AREG, EREG, EGR1; Selectin: SELPLG, ITGB2, ICAM1), excluding broadly regulated proliferation genes such as MYC and CCND1 (S2 Table).

Data and Code Availability

All gene expression data used in this study are publicly available from the GTEx Portal (https://gtexportal.org/, GTEx v8 open access). HEK293 KO glycomics data were obtained from [4]. DepMap Chronos scores were obtained from the DepMap Portal (https://depmap.org/). All analysis scripts, pathway definitions, and computed reachability scores are available at https://

github.com/matsui-lab/glycoreach (commit: `0ae9769`; Zenodo DOI to be assigned upon acceptance). Pathway definitions are provided in machine-readable JSON format (`data/pathway_definitions.js` to facilitate reuse and extension. The core reachability computation requires only an R installation with the `data.table` package and a gene-level TPM matrix as input.

Acknowledgments

This work was supported by the Human Glycome Atlas Project (HGA) and JSPS KAKENHI Grant Number JP20H04282.

References

[1] Neelamegham S, Aoki-Kinoshita KF, Bolton E, Frank M, Lisacek F, Lütteke T, et al. Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology*. 2019;29(9):620–624.

[2] Groth T, Diehl AD, Gunawan R, Neelamegham S. GlycoEnzOnto: a GlycoEnzyme pathway and molecular function ontology. *Bioinformatics*. 2022;38(24):5413–5420.

[3] Varki A. Biological roles of glycans. *Glycobiology*. 2017;27(1):3–49.

[4] Huang YF, Aoki K, Akase S, Ishihara M, Liu YS, Yang G, et al. Global mapping of glycosylation pathways in human-derived cells. *Dev Cell*. 2021;56(8):1195–1209.

[5] Kong WZ, Fujita M. GlycoMaple: recent updates and applications in visualization and analysis of glycosylation pathways. *Anal Bioanal Chem*. 2025;417:885–894.

[6] McDonald AG, Tipton KF, Davey GP. A knowledge-based system for display and prediction of O-glycosylation network behaviour in response to enzyme activities. *PLoS Comput Biol*. 2016;12(4):e1004844.

[7] Alvarez MRS, Holmes XA, Oloumi A, Grijaldo-Alvarez SJ, Schindler R, Zhou Q, et al. Integration of RNAseq transcriptomics and N-glycomics reveal biosynthetic pathways and predict structure-specific N-glycan expression. *Chem Sci*. 2025;16:7155–7172.

[8] Dworkin LA, Clausen H, Joshi HJ. Applying transcriptomics to study glycosylation at the cell type level. *iScience*. 2022;25(6):104419.

[9] Chrysinas P, Venkatesan S, Ang I, Ghosh V, Chen C, Neelamegham S, et al. Cell- and tissue-specific glycosylation pathways informed by single-cell transcriptomics. *NAR Genomics Bioinf*. 2024;6(4):lqae169.

[10] GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318–1330.

[11] Janda CY, Waghray D, Levin AM, Thomas C, Garcia KC. Structural basis of Wnt recognition by Frizzled. *Science*. 2012;337(6090):59–64.

[12] Nusse R, Clevers H. Wnt/β-catenin signaling, disease, and emerging therapeutic modalities. *Cell*. 2017;169(6):985–999.

[13] Kawashima N, Yoon SJ, Itoh K, Nakayama KI. Tyrosine kinase activity of epidermal growth factor receptor is regulated by GM3 binding through carbohydrate to carbohydrate interactions. *J Biol Chem*. 2009;284(10):6147–6155.

[14] Yoon SJ, Nakayama KI, Hikita T, Handa K, Hakomori SI. Epidermal growth factor receptor tyrosine kinase is modulated by GM3 interaction with N-linked GlcNAc termini of the receptor. *Proc Natl Acad Sci USA*. 2006;103(49):18987–18991.

[15] Pratilas CA, Taylor BS, Ye Q, Viale A, Sander C, Solit DB, et al. V600EBRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. *Proc Natl Acad Sci USA*. 2009;106(11):4519–4524.

[16] Ley K, Laudanna C, Cybulsky MI, Nourshargh S. Getting to the site of inflammation: the leukocyte adhesion cascade updated. *Nat Rev Immunol*. 2007;7(9):678–689.

[17] Kim S. ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Commun Stat Simul Comput*. 2015;44(4):1062–1078.

[18] Gagiannis D, Gossrau R, Reutter W, Zimmermann-Kordmann M, Horstkorte R. Engineering the sialic acid in organs of mice using N-propanoylmannosamine. *Biochim Biophys Acta*. 2007;1770(2):297–306.

[19] Orth JD, Thiele I, Palsson BO. What is flux balance analysis? *Nat Biotechnol*. 2010;28(3):245–248.

[20] Shlomi T, Cabili MN, Herrgård MJ, Palsson BO, Ruppin E. Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol*. 2008;26(9):1003–1010.

[21] Becker SA, Palsson BO. Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol*. 2008;4(5):e1000082.

[22] Ferreira PG, Muñoz-Aguirre M, Reverter F, Sá Godinho CP, Sousa A, Amadoz A, et al. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat Commun*. 2018;9(1):490.

[23] Dempster JM, Boyle I, Vazquez F, Root DE, Boehm JS, Hahn WC, et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biol*. 2021;22(1):343.

[24] Rapraeger AC, Krufka A, Olwin BB. Requirement of heparan sulfate for bFGF-mediated fibroblast growth and myoblast differentiation. *Science*. 1991;252(5013):1705–1708.

[25] Ai X, Do AT, Lozynska O, Kusche-Gullberg M, Lindahl U, Emerson CP Jr. QSulf1 remodels the 6-O sulfation states of cell surface heparan sulfate proteoglycans to promote Wnt signaling. *J Cell Biol*. 2003;162(2):341–351.

[26] McCormick C, Duncan G, Goutsos KT, Tufaro F. The putative tumor suppressors EXT1 and EXT2 form a stable complex that accumulates in the Golgi apparatus and catalyzes the synthesis of heparan sulfate. *Proc Natl Acad Sci USA*. 2000;97(2):668–673.

[27] Openo KK, Schulz JM, Vargas CA, Orton CS, Epstein MP, Schnur RE, et al. Epimerase-deficiency galactosemia is not a binary condition. *Am J Hum Genet*. 2006;78(1):89–102.

[28] Shworak NW, Liu J, Petros LM, Zhang L, Kobayashi M, Copeland NG, et al. Multiple isoforms of heparan sulfate D-glucosaminyl 3-O-sulfotransferase. *J Biol Chem*. 1999;274(9):5170–5184.

[29] Romano J, Kromrey JD, Coraggio J, Skowronek J. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and similar surveys? In: *Annual Meeting of the Florida Association of Institutional Research*; 2006:1–33.

[30] Kacser H, Burns JA. The control of flux. *Symp Soc Exp Biol.* 1973;27:65–104.

[31] Heinrich R, Rapoport TA. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur J Biochem.* 1974;42(1):89–95.

[32] Schnaar RL. Gangliosides of the vertebrate nervous system. *J Mol Biol.* 2016;428(16):3325–3336.

[33] Ballehaninna UK, Chamberlain RS. The clinical utility of serum CA 19-9 in the diagnosis, prognosis and management of pancreatic adenocarcinoma: an evidence based appraisal. *J Gastrointest Oncol.* 2012;3(2):105–119.

[34] Cliff N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol Bull.* 1993;114(3):494–509.

[35] Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* New York: Chapman & Hall; 1993.

[36] Phipson B, Smyth GK. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol.* 2010;9(1):Article 39.

[37] Otaki M, Hirane N, Natsume-Kitatani Y, Nogami Itoh M, Shindo M, Kurebayashi Y, et al. Mouse tissue glycome atlas 2022 highlights inter-organ variation in major N-glycan profiles. *Sci Rep.* 2022;12(1):17804.

Figures

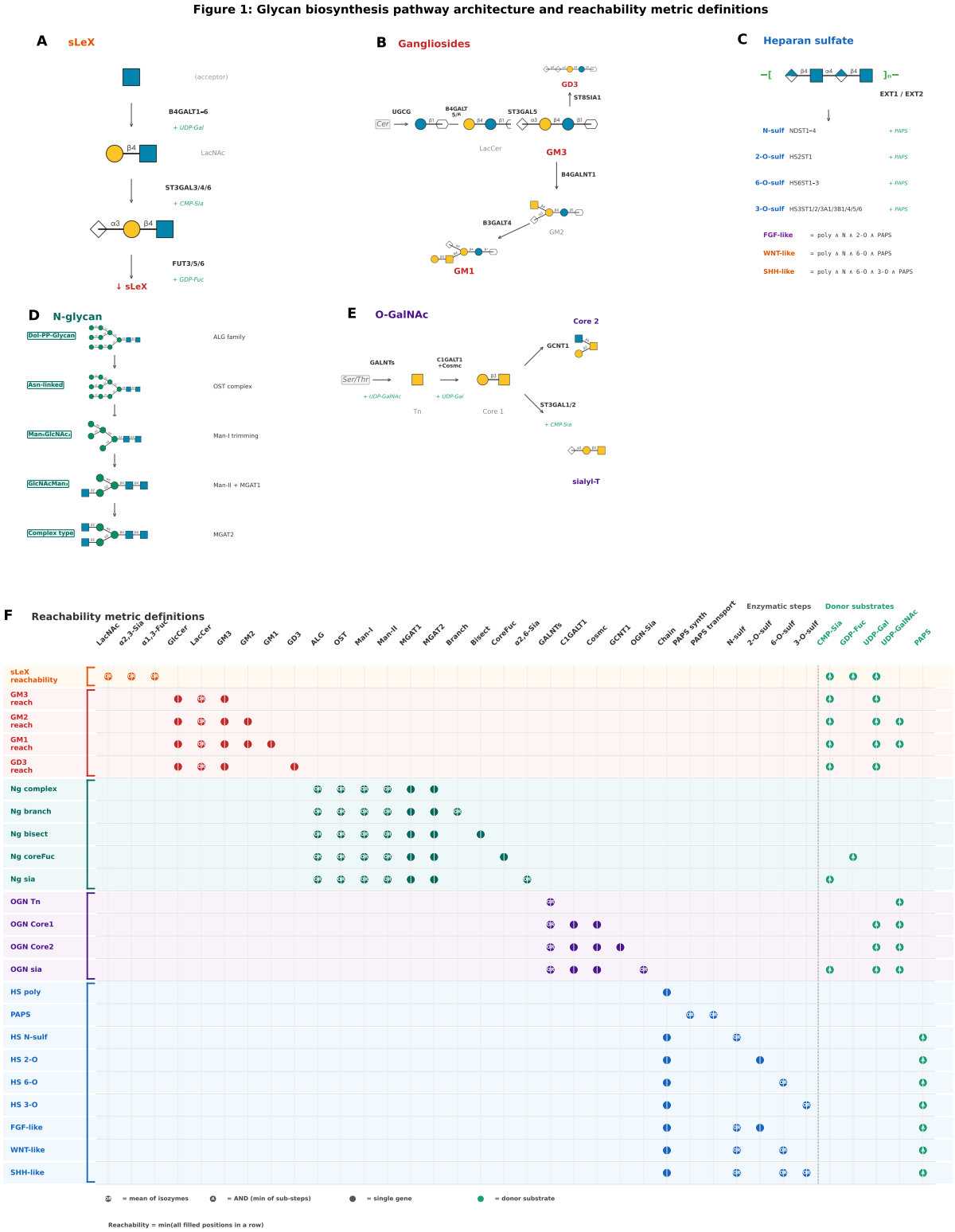

Fig. 1: Glycan biosynthesis pathway architecture and reachability metric definitions. (A–E) Schematic diagrams of five glycosylation pathway families modeled in this study: (A) sLeX, (B) ganglioside, (C) heparan sulfate, (D) N-glycan, and (E) O-GalNAc. Each diagram shows the sequential enzymatic steps (black) and required nucleotide sugar donors (green). **(F)** Reachability metric definition matrix for all 23 metrics. Each row defines one metric; columns represent enzymatic steps and shared donor substrates. OR nodes (isozymes) are aggregated by mean; AND nodes (sequential steps) by minimum. The reachability score equals the minimum across all required positions, identifying the expression-limiting bottleneck.

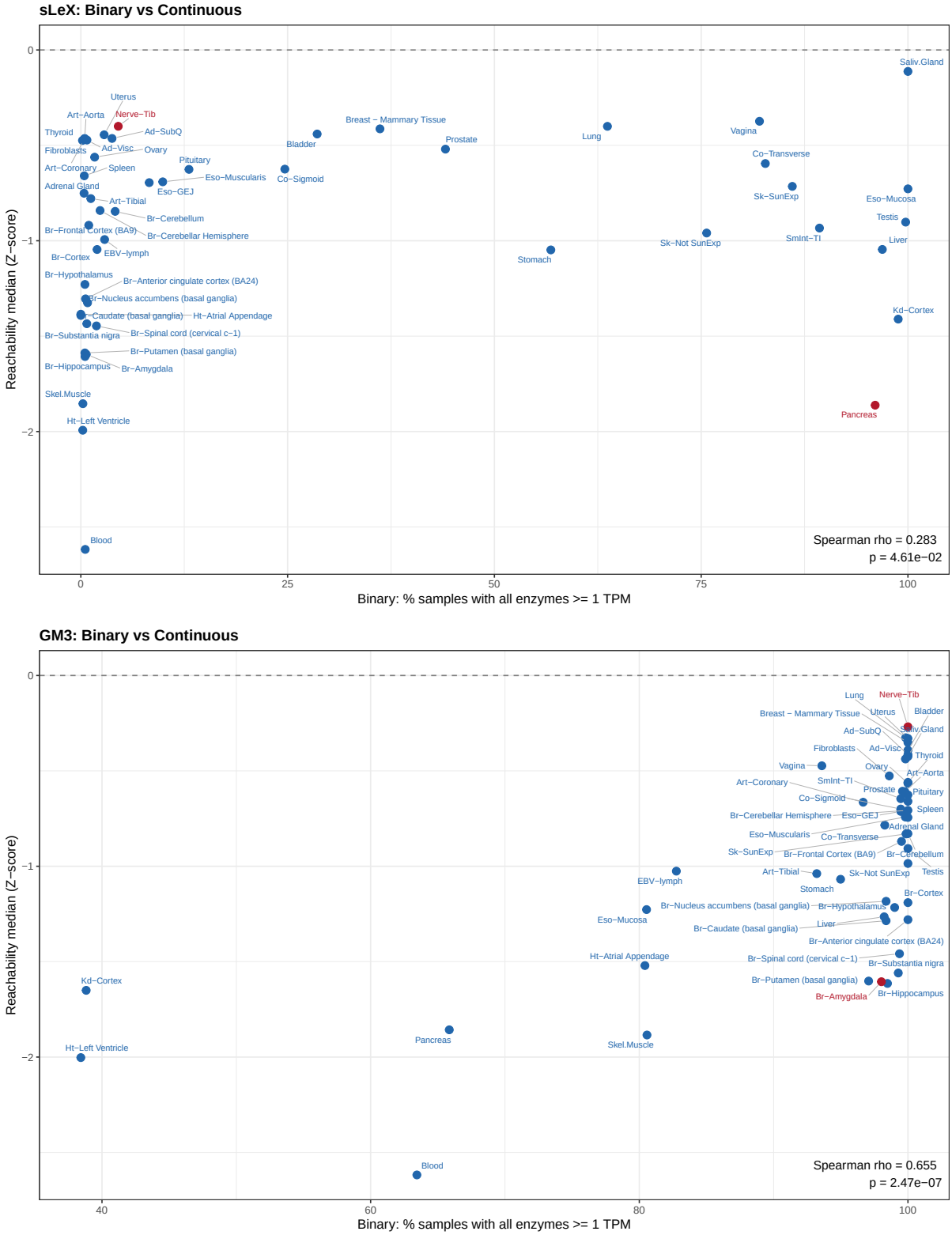

Fig. 2: Binary threshold approaches lose quantitative resolution in biosynthetic potential assessment. (A) Scatter plot of GlycoMaple-like binary positive rate (% samples with all sLeX pathway enzymes $\geq$ 1 TPM) versus median sLeX reachability Z-score across 50 tissues. Spearman $\rho = 0.259$, $p = 0.070$ (not significant). **(B)** Same as (A) for GM3. Spearman $\rho = 0.701$, $p = 1.5 \times 10^{-8}$.

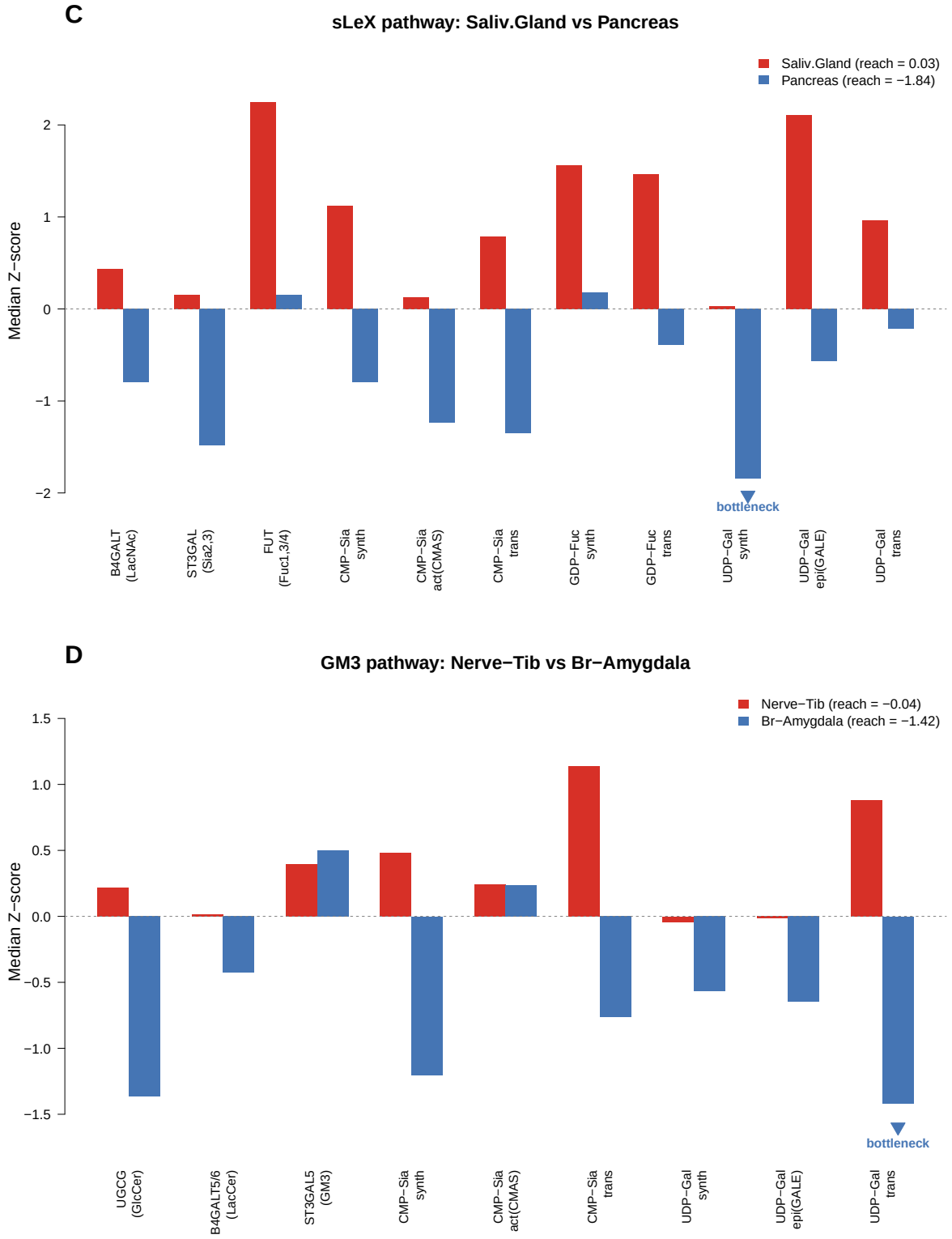

Fig. 2 (continued). **(C)** Per-step median Z-scores for the sLeX pathway comparing minor salivary gland (red) and pancreas (blue). Arrowhead marks the bottleneck in pancreas. **(D)** Per-step median Z-scores for the GM3 pathway comparing nerve-tibial (red) and brain-amygdala (blue).

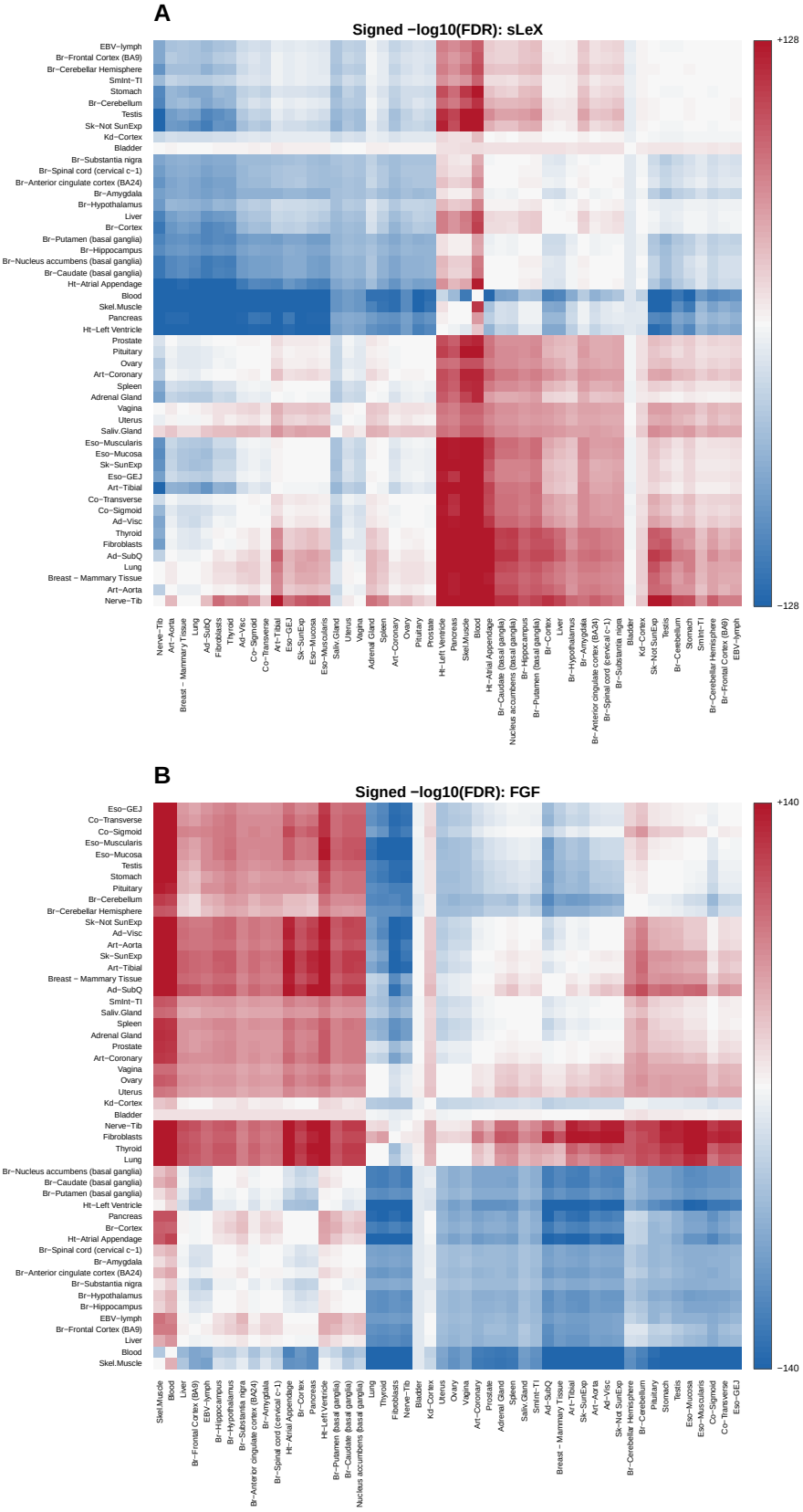

Fig. 3: Statistical framework for tissue-level comparison. **(A)** Signed $-\log_{10}(\text{FDR})$ heatmap for pairwise Wilcoxon tests of sLeX reachability across 50 tissues. **(B)** Same as (A) for HS FGF-like reachability, showing inverted tissue hierarchy compared to sLeX.

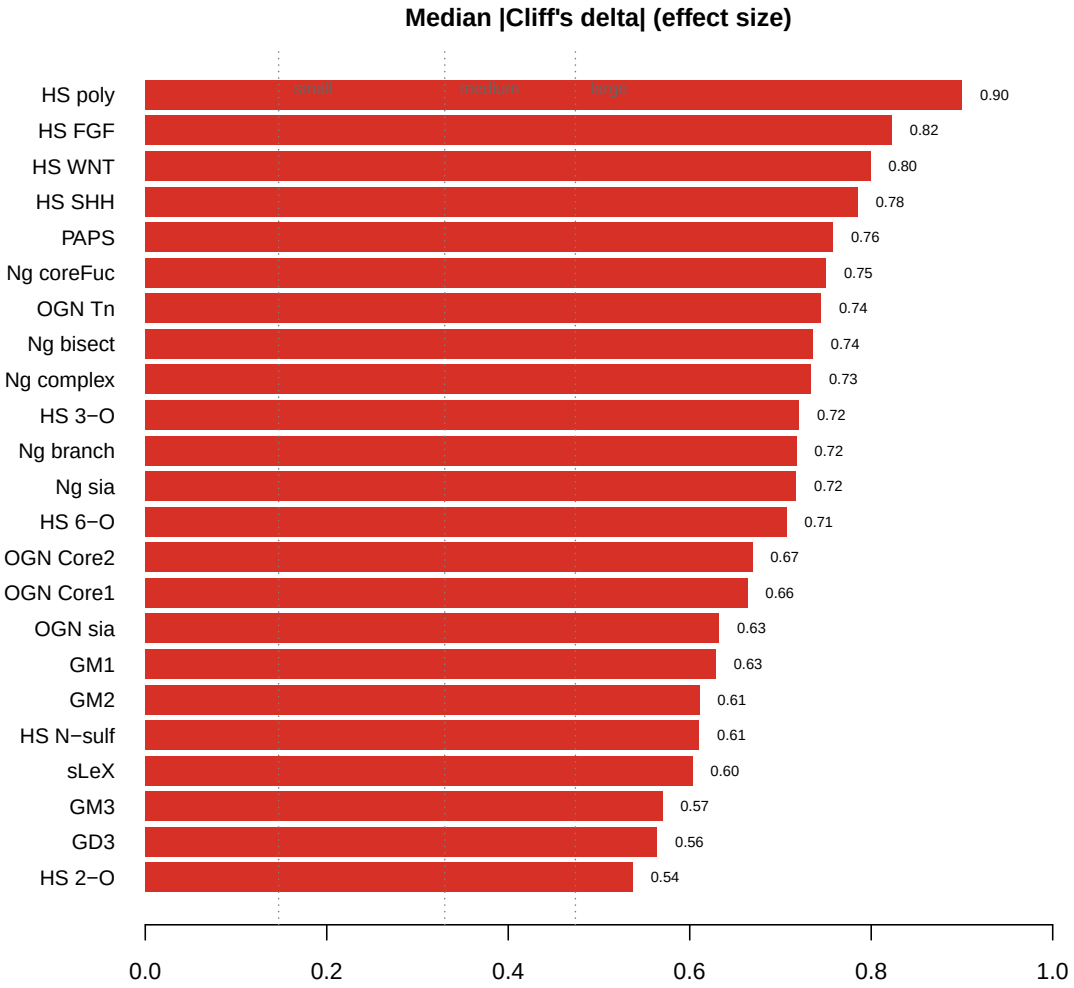

Fig. 3 (continued). (C) Median absolute Cliff's delta (effect size) across all 23 reachability metrics, sorted by discriminatory power. All metrics showed 89–94% of tissue pairs with significant differences (FDR < 0.05). Effect size guidelines: small ≥ 0.147, medium ≥ 0.33, large ≥ 0.474.

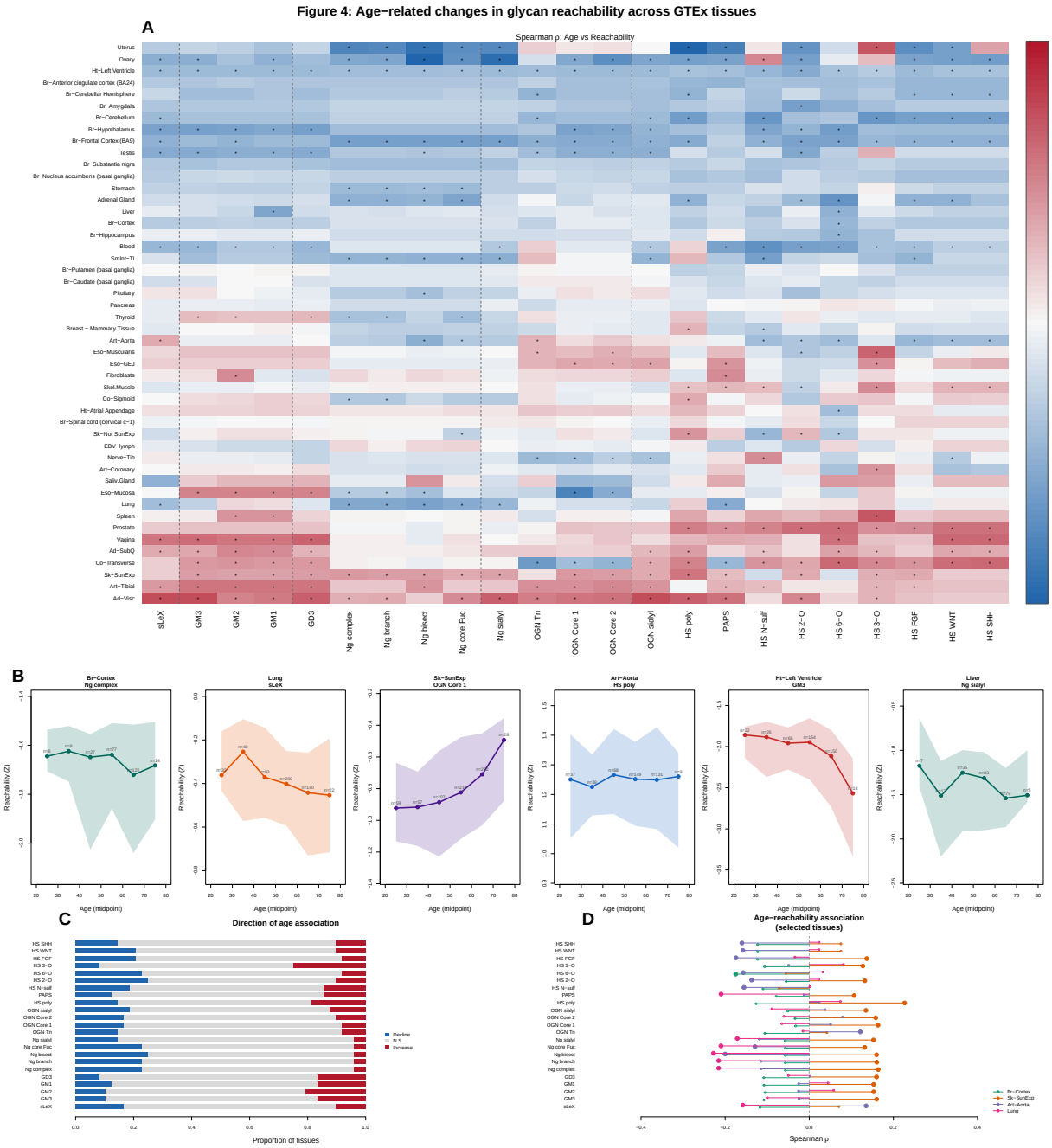

Fig. 4: Age-related changes in glycan biosynthetic potential. (A) Summary of age-reachability Spearman correlations across 23 metrics and 48 evaluable tissues. Bar heights show the number of tissues with significant decline (blue, below) or increase (red, above) at FDR < 0.05. **(B)** Representative tissue-level age-reachability trajectories for selected metrics and tissues, showing the diversity of age-related patterns across organs.

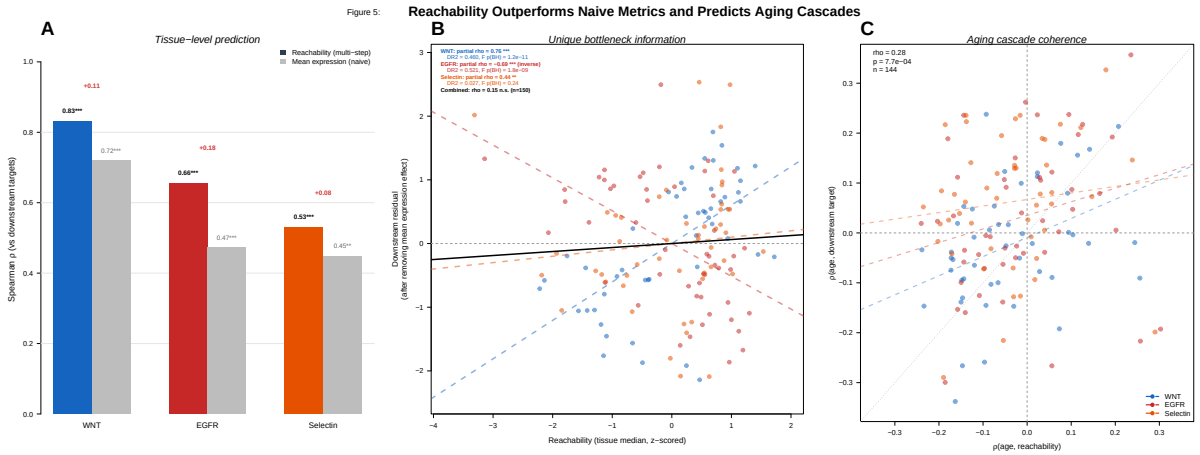

Fig. 5: Reachability predicts downstream signaling outcomes beyond naive mean expression. (A) Tissue-level Spearman correlation (ρ) between each predictor and downstream target expression for three independent cascades. Blue bars: reachability; grey bars: naive mean expression of all pathway genes. **(B)** Partial correlation analysis showing the unique contribution of reachability after controlling for mean expression. **(C)** Aging cascade coherence. Each point represents one tissue-cascade pair ($n = 144$). Positive concordance indicates that tissues where reachability declines (or increases) with age show parallel changes in downstream target expression.

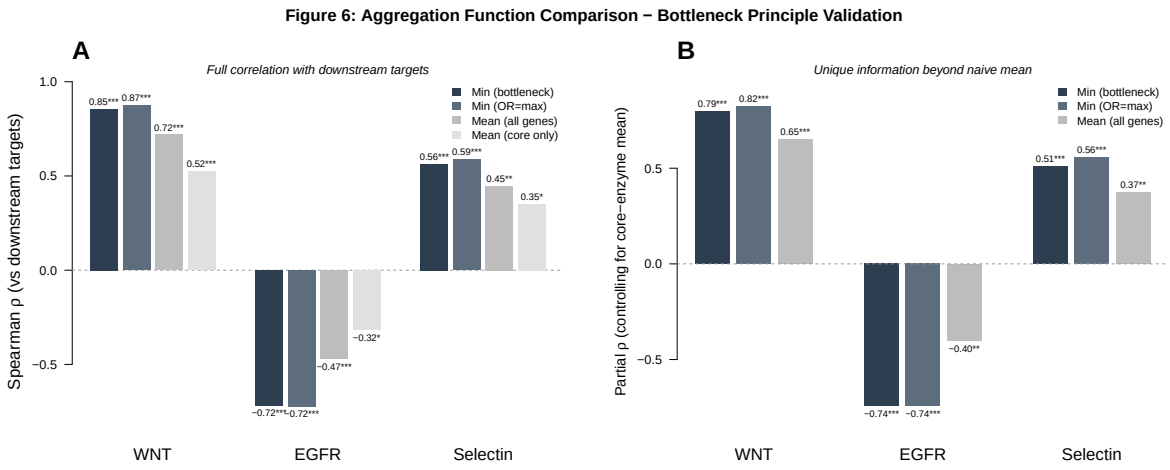

Fig. 6: Min-aggregation (bottleneck principle) outperforms averaging across all validation cascades. (A) Spearman correlation (ρ) between predicted biosynthetic score and downstream target expression for four aggregation methods across three cascades. **(B)** Spearman partial correlation (controlling for core-only mean) for the three methods that include substrate supply information. Min-aggregation maintains the strongest unique contribution across all cascades.