

*Figure X – Application of Glycompare substructure analysis to multiple datatypes. (A-B) Biclustered heatmaps of compositional (A) site-specific N-glycan data from mouse brain {Riley,2019}. (B) The same compositional data was substructure-decomposed to calculate substructure abundances presented in another biclustered heatmap. (C-D) The Pearson Correlation Coefficient was calculated for the compositional and composition substructure abundance for each glycosylation site across proteins. Biclustered heatmaps of the resulting Correlation coefficients are present as biclustered heatmaps. Biclustering used a complete agglomerative approach. (E) A re-analysis of mucin-type O-glycans {Jin,2019} from normal, tumor-proximal and gastrointestinal cancer biopsies, transformed to epitope abundance showed multiple core 2 epitopes depleted in gastrointestinal cancer. (F-G) A re-analysis of ganglioside glycolipid substructure abundance from a lacose root clarified distinct glycosylation in the retina {Sibelle,2016}*

**Results: Substructure and epitope abundance provides additional insight across glycan datatypes**

To explore the broad applicability of glycan substructure (from the root) and epitope (from the leaves) abundance, we next reanalyzed data from multiple studies with unique datatypes to explore how findings of each study could be enriched (**Fig X**).

Examining site-specific N-glycan compositional data from rat brain {Riley,2019}, we found that the decomposition of composition abundance into composition substructure abundance reveals additional potential signal. As previously shown, the sparsity of the abundance matrix decreases, and the comparability of profiles is improved when glycan data is aggregated over substructures (**Fig XA-B**). Futher, the correlation structure of substructure aggregated abundance (**Fig XD**) appears more robust than its compositional counterpart (**Fig XC**); there are more clusters with clearer boarders, multiple clear off-diagonal clusters and the median R2 is approximately doubled. While it is possible that the higher correlation is indicative of an increased background, that is unlikely considering the increase in visible correlation is structured, not randomly distributed through the background.

We next explored mucin-type O-glycans from a recent study in gastrointestinal cancer{Jin,2019}. Rather than building substructures from a biosynthetic core, and aggregating from larger to subsumed structures—this core-based analysis having the biosynthetic implications previously discussed, here we aggregate starting from monosaccharides to calculate epitope abundance. We found a substantial depletion in the tumor samples of five core 2 structures. These structures included three fucosylated and two with an I branches. The largest structures were over 30-fold depleted in tumors (FDR<0.03, **Fig XE**). We examined all glycan and epitope abundance changes relative to the tumor/normal comparison. **These five epitopes were significantly (after multiple test correction, FDR) different between tumor and normal samples while there were zero glycan structures that distinguished tumor from normal samples.** While these structures were observed in the original publication, the continuity across similar epitopes was not clear in the original work; that continuity is recovered using substructure aggregation.

Finally, we examined lactose core-based ganglioside abundance from a recent study of gangliosides in the eye, brain and blood {Sibille,2016}. We examined measured abundance from two gangliosides (GD3 and GM2) and their corresponding lactose-based substructure abundance. When ganglioside and substructure abundance was pooled by ceramide types, we found the GD3 substructure enriched in retina relative to brain and plasma, while the GD3 ganglioside abundance showed no coherent effect. Similarly, the GM2 substructure was enriched across several ceramide types in the Ciliary-body relative to the retina, while the GM2 ganglioside showed no coherent effect. By aggregating over subtypes we can account for confounding biosynthetic complexity thereby simplifying analyses and making crucial insights more accessible.

**Methods:** Core, Epitope and Compositional Substructure Abundance Calculations

In this manuscript we discuss three types of substructure abundance: core-based substructures, epitope substructures and compositional substructures. Each is a slight modification on the core algorithm with distinct implications.

Core-based substructures are discussed in the context of EPO N-glycosylation (N-glycan core), HMO abundance (lactose core), and gangliosides (lactose core). These analyses start with a glycan core (e.g. lacose) and then aggregate from larger structures towards the core. For example, any HMO abundance is added to the core lactose abundance because every HMO contains lactose. Similarly, DSLNT abundance is added to LNT abundance between DSLNT contains LNT. Core-base substructure aggregation accounts for the “disappearance” of glycan abundance as smaller structures are consumed to make larger structures; this aggregation approach accounts for biosynthetic variation in glycan abundance and should facilitate biosynthetic insights.

Epitope substructures are discussed in the context of mucin-type O-glycans. The epitope aggregation approach starts from every terminal monosaccharide and aggregates over larger structures. For examples, VIM substructure abundance (Neu5Ac-Gal-GlcNAc) would be added to sialic acid abundance. This type of aggregation focuses on substructures originating at the terminal monosaccharides and therefore can be used to identify relevant epitopes.

Finally, compositional substructures are discussed in the context of site-specific N-glycosylation data. The compositional substructure strategy sums over larger and subsuming structures in a compositional network. Consider the compositional abundance of a structure: HexNac(I)Hex(J)Fuc(K). Instead of abundance of HexNAc=I, Hex=J, and Fuc=K, we examine the compositional abundance for all measurements where HexNAc>=I, Hex>=J, and Fuc>=K. The nextwork structure can be constrained to provide additional insight (e.g. Glyconnect Compozitor). But currently, the aggregation criteria remains simple.