Table 1: Chromatogram Feature Definitions

$\mathcal{M}_i$	The neutral mass of the ith chromatogram
$\boldsymbol{\mathcal{I}}_i$	The total intensity array assigned to the ith chromatogram
$\boldsymbol{\mathcal{I}}_{i,j}$	The sum of all peak intensities for peaks observed in the jth scan for the ith chromatogram
$\mathcal{I}_{i,j,k}$	The intensity assigned to the kth peak at the jth scan for the ith chromatogram
$\mathbf{c}_i$	The set of charge states observed for the ith chromatogram
$\boldsymbol{\mathcal{I}}_{i,c=j}$	The total intensity assigned to the ith chromatogram with charge state j
$\mathbf{t}_{i,j}$	The time of the jth scan of the ith chromatogram
$\mathbf{env}_{i,j,k}$	The normalized experimental isotopic envelope composing the kth peak of the jth scan of the ith chromatogram, whose members sum to 1
$\mathbf{a}_i$	The set of adduction states observed for the ith chromatogram
$\boldsymbol{\mathcal{I}}_{i,a=j}$	The total intensity assigned to the ith chromatogram with adduct j
$\hat{g}_i$	The glycan composition assigned to the ith chromatogram, or $\emptyset$ if there was no matched glycan composition

1 Chromatographic Feature Evaluation

For each candidate feature, we computed several statistics to estimate how distinguishable the observed signal was from random noise. We use the following quantities from each LC-MS feature:

1.1 Chromatographic Peak Shape

An LC-MS elution profile should be composed of one or more peak-like components, each following a bi-Gaussian peak shape model (Yu and Peng (2010)) or in less ideal chromatographic circumstances, a skewed Gaussian peak shape model. We fit these models using non-linear least squares (NLS). As measures of goodness of fit are not generally available for NLS, we use the following criterion:

$$\begin{aligned}
\hat{y}_i &= NLS(\boldsymbol{\mathcal{I}}_i, \mathbf{t}_i) \\
e_{i,NLS} &= \boldsymbol{\mathcal{I}}_i - \hat{y}_i \\
\bar{y}_i &= \mathbf{t}_i \left(\left(\mathbf{t}_i^t \mathbf{t}_i \right)^{-1} \mathbf{t}_i \boldsymbol{\mathcal{I}}_i \right) \\
e_{i,null} &= \boldsymbol{\mathcal{I}}_i - \bar{y}_i \\
\mathscr{L}_i &= 1 - \frac{\sum e_{i,NLS}^2}{\sum e_{i,null}^2}
\end{aligned} \tag{1}$$

where line score describes how much the peak shape fit improves on a straight line fit null model.

We apply two competitive peak fitting strategies to address distorted, overlapping, or multimodal elution profiles. The first works iteratively by finding a best-matching peak shape using non-linear least squares, subtracting the fitted signal and checks if there is another peak with at least half as tall as the removed peak, if so repeating the process until no peak can be found, saving each peak model so constructed. The second approach starts by locating local minima between putative peaks, and partitioning the chromatogram into sub-groups which would are fit independently. This method generates a candidate list of minima, and selects the case which has the greatest difference between the minimum and its pair of maxima to split the feature at. The strategy which produces the maximum $\mathscr{L}_i$ is chosen.

1.2 Composition Dependent Charge State Distribution

As the number of monosaccharides composing a glycan increases, the number of possible sites for charge localization increases. Under normal conditions, we would expect to observe the same molecule in multiple charge states (Maxwell *et al.* (2012)).

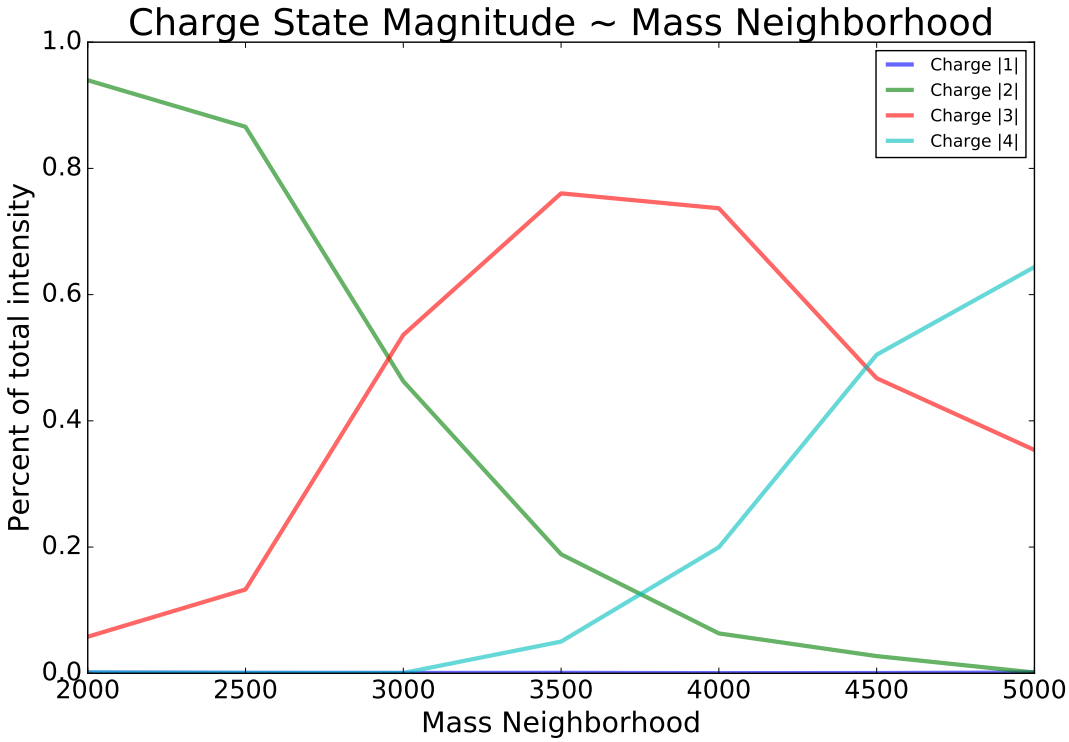

Figure 1: The trend of charge state relative abundance for acidic glycans

Which charge states are expected would depend upon the size of the molecule and it's constituent units' electronegativity. In it's native state, **NeuAc**'s acidic group causes glycans with one or more **NeuAc** to have a propensity for higher negative charge states (Varki and Schauer (2009)). To capture this relationship, we modeled the probability of observing a glycan composition for sialylated and unsialylated compositions separately. For permethylated glycans, charge is carried by protons or metallic cation adducts like sodium, the relationship between acidic monosaccharides and charge state propensities is weaker.

$$
\begin{aligned}
m_i &= (\lfloor(\mathcal{M}_i/w)/10\rfloor + 1) * 10 \\
\mathcal{H}_{i,j} &= \frac{\boldsymbol{\mathcal{I}}_{i,c=j}}{\boldsymbol{\mathcal{I}}_i} \\
P(c, m) &= |m| \sum_{m_i \in m} \mathcal{H}_{i,j} \\
\mathscr{C}_i &= \sum_{c_{i,j} \in \mathbf{c}_i} P(c_{i,j}, m_i)
\end{aligned} \tag{2}
$$

where w is the width of the mass bin divided by 10 and $P(c, m)$ is defined as part of the model estimation procedure.

1.3 Adduction Rate

For the samples *AGP-permethylated-2ul-inj-55-SLens* and *Perm-BS-070111-04-Human-Serum* we also include an Adduction Frequency model score $\mathscr{A}_i$, following the same pattern as the charge state distribution, with the same extension of justification from Maxwell *et al.* (2012). We use one mass scaling model for all glycan compositions as ammmonium adduction is not expected to be composition dependent.

$$
\begin{aligned}
\mathcal{H}_{i,j} &= \frac{\boldsymbol{\mathcal{I}}_{i,a=j}}{\boldsymbol{\mathcal{I}}_i} \\
P(a, m) &= |m| \sum_{m_i \in m} \mathcal{H}_{i,j} \\
\mathscr{A}_i &= \sum_{a_{i,j} \in \mathbf{a}_i} P(a_{i,j}, m_i)
\end{aligned} \tag{3}
$$

We fit the adduction rate model on *AGP-permethylated-2ul-inj-55-SLens* in order to make our comparison to third-party data less biased given limited sample data.

1.4 Isotopic Pattern Consistency

Our ahead-of-time deconvolution procedure uses an averagine isotopic model and does not capture the consistency of the isotopic pattern that was fit with the isotopic pattern of the glycan composition that matched that peak. The criterion

$$\mathscr{I}_i = 1 - 2\boldsymbol{\mathcal{I}}_i^{-t}\mathbf{I}_i \sum_{j}^{J}\sum_{k}^{K} \mathcal{I}_{i,j,k}\mathbf{env}_{i,j,k}^{t}\left(\ln \mathbf{env}_{i,j,k} - \ln \mathbf{tid}_i\right) \tag{4}$$

where **tid** is the theoretical isotopic pattern derived from either $\hat{g}_i$ or an averagine interpolated for $\mathcal{M}_i$ if $\hat{g}_i = \emptyset$. This computes a per-peak intensity weighted mean G-test comparing the goodness of fit between the experimental envelope and the theoretical isotopic pattern.

1.5 Observation Spacing Score

The less time between observations of a glycan composition the less likely the chromatogram is to contain peaks missing or caused by isotopic pattern interference or missing information.

$$\mathscr{T}_i = 1 - 2\boldsymbol{\mathcal{I}}_i^{-t}\mathbf{I}_i \sum_{j=1}^{J} \boldsymbol{\mathcal{I}}_{i,j}(\mathbf{t}_{i,j} - \mathbf{t}_{i,j-1}) \tag{5}$$

1.6 Summarization Score

Each scoring feature $\in [\mathscr{L}_i, \mathscr{C}_i, \mathscr{I}_i, \mathscr{T}_i]$ is penalized by $\epsilon = 1\text{e}{-6}$ bounded in the range $[0, 1)$, with values below 0 set to ϵ.

$$s_i = \sum_{f_{i,j} \in \text{features}_i} \ln \frac{f_{i,j}}{1 - f_{i,j}} \tag{6}$$

producing a value between $(-\infty, \infty)$. $s_i < 8$ reflects multiple poor feature scores and is unexpected to be real, while $s_i > 15$ is consistent with model expectations.

2 MS^n Signature Ion Criterion

When MS^nscans are present, it may be useful to consider only those MS^1 features which are associated with MS^nscans that contain glycan-like signature ions. We include an algorithm for classifying an MS^nscan as being "glycan-like":

$$I = max(intensity(p)) \tag{7}$$

$$t = I * 0.01 \tag{8}$$

$$p_{oxonium} = \{p_i \leftarrow |ppmerror(mass(p_j), mass(f_g))| < e, f_g \in oxonium(g), f_g \neq \text{Fucose}, intensity(p_i) > t\} \tag{9}$$

$$\begin{aligned} p_{edges} = \{&(p_i, p_j) \leftarrow |ppmerror(mass(p_j) - mass(p_i), mass(f_g))| < e, \\ &oxonium(f_g) \in g, intensity(p_i) > t, intensity(p_j) > t\} \end{aligned} \tag{10}$$

$$s_{oxonium} = \frac{1}{|p_{oxonium}|} \sum_{p_i}^{p_{oxonium}} \left(\frac{intensity(p_i)}{I}\right) * min(log_4|p_{oxonium}|, 1) \tag{11}$$

$$s_{edges} = \frac{1}{|p_{edges}|} \sum_{p_i,p_j}^{p_{edges}} \left(\frac{intensity(p_i) + intensity(p_j)}{I}\right) * min(log_4|p_{edges}|, 1) \tag{12}$$

$$s_g = max(s_{oxonium}, s_{edges}) \tag{13}$$

$$\tag{14}$$

Where p is the set of peaks in the scan, g is the glycan compostion, e the required parts-per-million mass accuracy. $oxonium()$ is a function that given a glycan composition g, produces fragments f_g of g composed of between one and three monosaccharides, commonly observed as oxonium ions alone, or as the mass difference between two peaks formed from consecutive fragmentation of a glycosidic bond. This method is not intended to identify a glycan structure, just detect patterns in the signal peaks of the MS^nscan that could indicate the fragmentation of a glycan.

3 A more complete derivation of $\hat{\phi}$

To obtain the optimal ϕ, we take the partial derivative of ℓ w.r.t ϕ_m

$$
\begin{aligned}
0 &= \frac{\partial \ell}{\partial \phi_m}\left((\mathbf{s}-\phi_{\mathbf{o}})^t(\mathbf{s}-\phi_{\mathbf{o}}) + \lambda\begin{bmatrix}\phi_o-\tau_o, & \phi_m-\tau_m\end{bmatrix}\begin{bmatrix}\mathbf{L_{oo}} & \mathbf{L_{om}}\\ \mathbf{L_{mo}} & \mathbf{L_{mm}}\end{bmatrix}\begin{bmatrix}\phi_o-\tau_o\\ \phi_m-\tau_m\end{bmatrix}\right) \qquad (15)\\
&= \lambda(\phi_o-\tau_o)^t\mathbf{L_{om}} + \lambda\mathbf{L_{mo}}(\phi_o-\tau_o) + \lambda(\phi_m-\tau_m)^t(\mathbf{L_{mm}}^t + \mathbf{L_{mm}})\\
&= 2\lambda\mathbf{L_{mo}}(\phi_o-\tau_o) + 2\lambda\mathbf{L_{mm}}(\phi_m-\tau_m)\\
-\mathbf{L_{mm}}(\phi_m-\tau_m) &= \mathbf{L_{mo}}(\phi_o-\tau_o)\\
(\phi_m-\tau_m) &= -\mathbf{L_{mm}}^{-1}\mathbf{L_{mo}}(\phi_o-\tau_o)\\
\hat{\phi}_m &= -\mathbf{L_{mm}}^{-1}\mathbf{L_{mo}}(\phi_o-\tau_o)+\tau_m \qquad (16)
\end{aligned}
$$

and w.r.t. ϕ_o

$$
\begin{aligned}
0 &= \frac{\partial \ell}{\partial \phi_o}\left((\mathbf{s}-\phi_{\mathbf{o}})^t(\mathbf{s}-\phi_{\mathbf{o}}) + \lambda\begin{bmatrix}\phi_o-\tau_o, & \phi_m-\tau_m\end{bmatrix}\begin{bmatrix}\mathbf{L_{oo}} & \mathbf{L_{om}}\\ \mathbf{L_{mo}} & \mathbf{L_{mm}}\end{bmatrix}\begin{bmatrix}\phi_o-\tau_o\\ \phi_m-\tau_m\end{bmatrix}\right) \qquad (17)\\
&= -2\mathbf{s} + 2\phi_o + \lambda\left(\mathbf{L_{oo}}+\mathbf{L_{oo}}^t\right)(\phi_o-\tau_o) + \lambda\mathbf{L_{om}}(\phi_m-\tau_m) + \lambda\mathbf{L_{mo}}^t(\phi_m-\tau_m)\\
&= -2\mathbf{s} + 2\phi_o + 2\lambda\mathbf{L_{oo}}(\phi_o-\tau_o) + 2\lambda\mathbf{L_{om}}(\phi_m-\tau_m)\\
\mathbf{s} &= \phi_o + \lambda\left(\mathbf{L_{oo}}(\phi_o-\tau_o) + \mathbf{L_{om}}(\phi_m-\tau_m)\right)\\
&= \phi_o + \lambda\left(\mathbf{L_{oo}}(\phi_o-\tau_o) + \mathbf{L_{om}}(-\mathbf{L_{mm}}^{-1}\mathbf{L_{mo}}(\phi_o-\tau_o)+\tau_m-\tau_m)\right)\\
&= \phi_o + \lambda\left(\mathbf{L_{oo}}(\phi_o-\tau_o) - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}(\phi_o-\tau_o)\right)\\
\mathbf{s}-\tau_o &= \phi_o-\tau_o + \lambda\left(\mathbf{L_{oo}}(\phi_o-\tau_o) - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}(\phi_o-\tau_o)\right)\\
&= \mathbf{I}(\phi_o-\tau_o) + \lambda\left(\mathbf{L_{oo}}(\phi_o-\tau_o) - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}(\phi_o-\tau_o)\right)\\
&= \left[\mathbf{I} + \lambda\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)\right](\phi_o-\tau_o)\\
(\phi_o-\tau_o) &= \left[\mathbf{I} + \lambda\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)\right]^{-1}(\mathbf{s}-\tau_o)\\
\hat{\phi}_o &= \left[\mathbf{I} + \lambda\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)\right]^{-1}(\mathbf{s}-\tau_o)+\tau_o \qquad (18)
\end{aligned}
$$

4 Estimation of Laplacian Regularization Parameters

We model the relationship between $\mathbf{s}$, $\phi_{\mathbf{o}}$, and τ as a set of gaussian distribution.

$$(\mathbf{s}|\phi_{\mathbf{o}},\tau) \sim \mathcal{N}(\phi_{\mathbf{o}},\Sigma) \qquad (19)$$

$$\Sigma = \rho\mathbf{I} \qquad (20)$$

$$\left(\begin{bmatrix}\phi_{\mathbf{o}}\\ \phi_{\mathbf{m}}\end{bmatrix}\middle|\tau\right) \sim \mathcal{N}(\mathbf{A}\tau, \lambda^{-1}\mathbf{L}^-) \qquad (21)$$

$$(\phi_{\mathbf{o}}|\tau) \sim \mathcal{N}\left(\mathbf{A_o}\tau, \Sigma_{\phi_o}\right) \qquad (22)$$

$$\Sigma_{\phi_o} = \lambda^{-1}\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)^{-1} \qquad (23)$$

$$\tau \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right) \qquad (24)$$

Fully expanded, this becomes

$$\begin{bmatrix}\mathbf{s}\\ \phi_{\mathbf{o}}\\ \tau\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\Sigma+\Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \sigma^2\mathbf{A_o}\\ \Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \sigma^2\mathbf{A_o}\\ \sigma^2\mathbf{A_o}^t & \sigma^2\mathbf{A_o}^t & \sigma^2\mathbf{I}\end{bmatrix}\right) \qquad (25)$$

We can form the conditional distribution $\tau|\mathbf{s}$ which has a mean

$$
\begin{aligned}
\mu_{\tau|\mathbf{s}} &= 0 + (\sigma^2\mathbf{A_o}^t)\left(\Sigma+\Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o^t}\right)^{-1}\mathbf{s} \qquad (26)\\
&= \mathbf{A_o}^t\left(\tilde{\rho}\mathbf{I} + \frac{1}{\tilde{\lambda}}\mathbf{L_{oo}^-} + \mathbf{A_o}\mathbf{A_o^t}\right)^{-1}\mathbf{s} \qquad (27)
\end{aligned}
$$

We assume that $\sigma^2 \gg 1$, and treat λ and ρ as relative to σ^2, as $\tilde{\rho}$ and $\tilde{\lambda}$. This model gives us an estimate for τ given a value for ρ and λ. As ρ has no direct role in the central tendency of ϕ or $\mathbf{s}$, we choose to fix the value of $\tilde{\rho} = 0.1$, which leaves only $\tilde{\lambda}$. We estimate the optimal $\tilde{\lambda}$ by grid search, minimizing the predicted residual error sum of squares (PRESS) statistic.

$$\arg\min_{\tilde{\lambda}} \frac{\mathbf{s} - \hat{\phi}_{\mathbf{o}}}{\left(1 - \left(\mathbf{I} + \tilde{\lambda}\mathbf{L}\right)^{-1}\right)^2} \tag{28}$$

This formulation depends upon the value of $\mathbf{s}$ and is sensitive to low scoring matches, which can lead to incorrect estimates of τ and PRESS. We therefore perform a grid search over both $\tilde{\lambda}$ and a minimum threshold for $\mathbf{s}$, γ.

As we increase γ we remodel the graph $\mathcal{G}$, removing nodes whose score is below γ. For each pair of neighbors of removed node g_m, (g_u, g_v), if $L_1(g_u, g_v) > L_1(g_u, g_m) + L_1(g_m, g_v)$, we add an edge from g_u to g_v with weight $\frac{1}{L_1(g_u,g_m)+L_1(g_m,g_v)}$, up to a limit of $L_1(g_k, g_m) < 5$. We give the result of this grid search the name $\mathbf{r}$. At each point, on the grid, we save the value of τ in $r_{\lambda_i,\gamma_j,\tau}$ and the PRESS in $r_{\lambda_i,\gamma_j,PRESS}$. To select the optimal parameters, we traverse the grid along γ, computing τ_γ:

$$\bar{\lambda}_j = \arg\min_{\lambda_i} r_{\lambda_i,\gamma_j,PRESS} \tag{29}$$

$$\tau_{\gamma_j} = |r_{\bar{\lambda}_j,\gamma_j,\tau}| * \left(\frac{\gamma_j}{b} + (1 - \frac{1}{b})\right) \tag{30}$$

where b is a bias factor defining how much weight to give to higher values of γ which correspond to networks made up of higher confidence assignments. We chose $b = 4$. We define $\bar{\tau}_\gamma = \max \tau_\gamma$ and define the vector $\bar{\gamma} = \left[\gamma_j \leftarrow \tau_{\gamma_j} \geq \bar{\tau}_\gamma * 0.9\right]$. This favors values of γ where large values of τ are selected, meaning that the neighborhoods are well populated, while also giving an estimate for $\tilde{\lambda}$ that is non-zero. We term the values of γ in $\bar{\gamma}$ the *target thresholds* of $\mathbf{s}$.

To estimate $\tilde{\lambda}$ and τ from these results, we select the columns of the grid $\mathbf{r}$ at each $\gamma_j \in \bar{\gamma}$ and applied the following procedure:

$$\bar{\tau}_\gamma = \max \tau_\gamma \tag{31}$$

$$\bar{\gamma} = \left\{\gamma_j \leftarrow \tau_{\gamma_j} \geq \bar{\tau}_\gamma * 0.9\right\} \tag{32}$$

$$\bar{\lambda} = \left\{\bar{\lambda}_j \leftarrow \gamma_j \in \bar{\gamma}\right\} \tag{33}$$

$$\mathbf{s}_{\gamma_\mathbf{j}} = \left\{s_i \leftarrow s_i > \gamma_j\right\} \tag{34}$$

$$\bar{\tau}_\mathbf{j} = \mu_{\tau|\mathbf{s}_{\gamma_j},\bar{\lambda}_j} \tag{35}$$

$$\hat{\lambda} = \frac{1}{|\bar{\lambda}|}\sum_j \bar{\lambda}_j \tag{36}$$

$$\hat{\tau} = \frac{1}{|\bar{\tau}|}\sum_j \bar{\tau}_\mathbf{j} \tag{37}$$

$$\hat{\gamma} = \frac{1}{|\bar{\gamma}|}\sum_j \bar{\gamma}_j \tag{38}$$

where $\mathbf{s}_{\gamma_j}$ is the set of observed scores which are greater than γ_j, but where the estimation of is carried out with the complete Laplacian $\mathbf{L}$, not the reduced network used to compute $\mathbf{r}$. This set of averaged estimates of $\hat{\lambda}$ and $\hat{\tau}$ are then used to estimate $\hat{\phi}_o$ by 18, labeled **??** in the main text.

5 Algorithmic Performance on All Datasets

TODO

6 Differences in Assigned Glycans for *Perm-BS-070111-04-Serum*

Of the compositions assigned by our algorithm that were not mentioned in Yu *et al.* (2013) but were annotated in the original publication of this dataset in Hu and Mechref (2012) include **HexNAc3 Hex4**, **HexNAc3 Hex4 NeuAc1**, and **HexNAc5 Hex3**. Because our database was constructed based on combinatorial rules that did not take into account all biosynthetic constraints, we include infeasible compositions in our search space, such as **HexNAc2 Hex10 Fuc1** and **HexNAc5 Hex3**

Fuc1 NeuAc2. Future work could be done to restrict the database to only biosynthetically feasible glycan compositions. This would also have benefits for the construction of the composition network where only those compositions which have an enzymatic reaction to from one to the other would have an edge connecting them, such that **HexNAc5 Hex6 NeuAc2** would not have an edge to **HexNAc5 Hex7 NeuAc2** as in our current model.

References

Hu, Y. and Mechref, Y. (2012). Comparing MALDI-MS, RP-LC-MALDI-MS and RP-LC-ESI-MS glycomic profiles of permethylated N-glycans derived from model glycoproteins and human blood serum. *Electrophoresis*, **33**(12), 1768–1777.

Maxwell, E., Tan, Y., Tan, Y., Hu, H., Benson, G., Aizikov, K., Conley, S., Staples, G. O., Slysz, G. W., Smith, R. D., and Zaia, J. (2012). GlycReSoft: a software package for automated recognition of glycans from LC/MS data. *PloS one*, **7**(9), e45474.

Varki, A. and Schauer, R. (2009). *Sialic Acids*. Cold Spring Harbor Laboratory Press.

Yu, C.-Y. C.-Y., Mayampurath, A., Hu, Y., Zhou, S., Mechref, Y., and Tang, H. (2013). Automated annotation and quantification of glycans using liquid chromatography-mass spectrometry. *Bioinformatics*, **29**(13), 1706–1707.

Yu, T. and Peng, H. (2010). Quantification and deconvolution of asymmetric LC-MS peaks using the bi-Gaussian mixture model and statistical model selection. *BMC bioinformatics*, **11**(1), 559.