Application of Network Smoothing to Glycan LC-MS Profiling

Joshua Klein, Luis Carvalho, Joseph Zaia

April 27, 2018

Contents

1 Experimental Samples Used

We demonstrate our algorithm on several samples from a variety of instruments and conditions described in Table S 1. We present two samples in the main text, QTOF analysis of Native, Formate adducted N-glycans from Influenza strain Phil-BS virions *20141103-02-Phil-BS*, and Orbitrap analysis of Permethylated and Reduced Ammonium adducted N-glycans from human serum *Perm-BS-070111-04-Serum*.

Sample Name	Instrument	Derivatization	Adduction	Source	Taxon
20150930-06-AGP	QTOF	Native	Formate (1)	Khatri *et al.* (2016a)	Human
20141031-07-Phil-82	QTOF	Native	Formate (3)	Khatri *et al.* (2016a)	Human Virus in Avian Tissue
20141103-02-Phil-BS	QTOF	Native	Formate (3)	Khatri *et al.* (2016a)	Human Virus in Avian Tissue
20151002-02-IGG	QTOF	Native	Formate (2)	Khatri *et al.* (2016b)	Human
20141128-11-Phil-82[1]	QTOF	Deutero-reduced, Permethylated	Ammonium (3)	Khatri *et al.* (2016a)	Human Virus in Avian Tissue
AGP-DR-Perm-glycans-1[1]	Orbitrap	Deutero-reduced, Permethylated	Ammonium (3)	Khatri *et al.* (2016a)	Human
AGP-permethylated-2ul-inj-55-SLens[1]	Orbitrap	Reduced, Permethylated	Ammonium (3)	Khatri *et al.* (2016a)	Human
Perm-BS-070111-04-Serum[1]	Orbitrap	Reduced, Permethylated	Ammonium (3)	Yu *et al.* (2013); Hu and Mechref (2012)	Human

[1] Included MS^n Scans

Table 1: Samples Used

As Table 1 describes, we analyze data from several different combinations of configurations of instrument, derivatization, and reduction.

2 Database Comparison

The three databases we used were overlapping but distinct. The size of these overlaps is shown in Figure S 1.

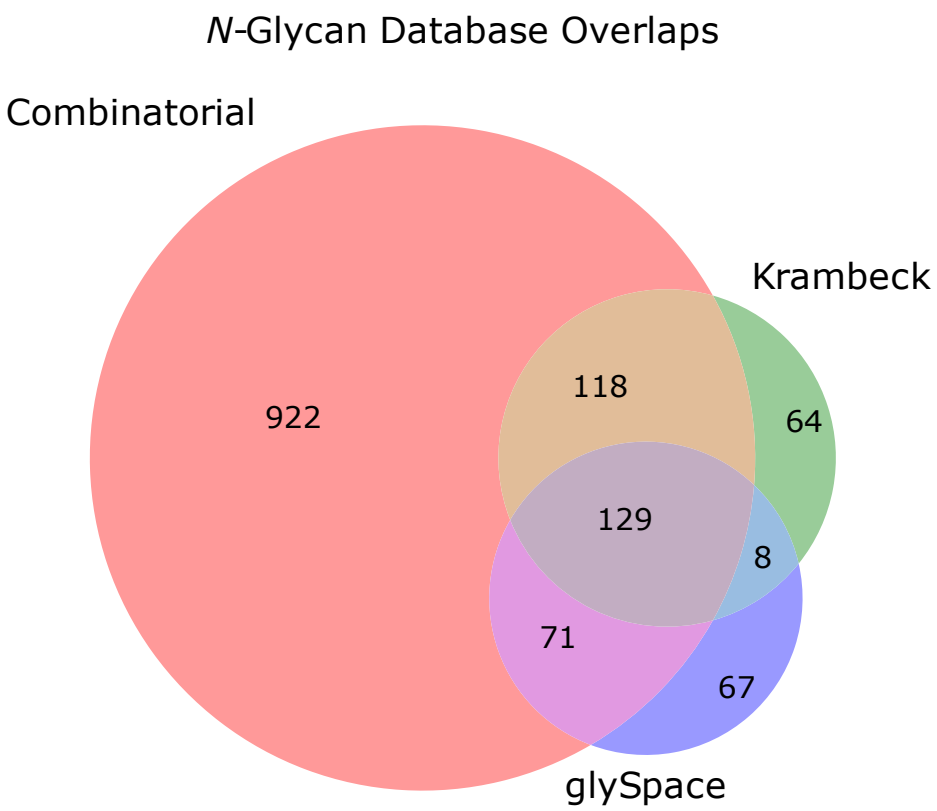


Figure 1: The overlap of the source databases used. As expected, the combinatorial database contains an enormous number of compositions not found in either other database, many of which are not biosynthetically feasible for humans. Those found in the Krambeck database but not the combinatorial or glySpace database are derived from lactosamine extensions run to the limit of the biosynthetic process covered in the original simulation (Krambeck and Betenbaugh, 2005). The glySpace database contained composition units not found in the other two databases, such as Xylose, Sulfate, and Phosphate.

3 Chromatographic Feature Evaluation

For each candidate feature, we computed several metrics to estimate how distinguishable the observed signal was from random noise. We use the quantities described in Table S 2 from each LC-MS feature.

All metrics are penalized by an $\epsilon = 1e-6$ to prevent scores from actually achieving a value of 1.0 which would make the logit value infinite. If a metric's value would be less than $0+\epsilon$, it is given a value of ϵ instead to prevent the logit value from being undefined.

3.1 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

Table 2: 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
T_j	The time of the jth scan of the experiment
$\mathbf{env}_{i,j,k}$	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

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 ordinary least squares regression linear 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. $\mathscr{L}_i$ is bounded in $(-\infty, 1]$, where 1 corresponds to a perfect fit, and 0 would correspond to the peak shape fit being no better than the OLS straight line fit. This metric is thresholded at 0.15, with any chromatogram scoring below 0.15 being discarded as having insufficient peak shape evidence to interpret.

3.2 Composition Dependent Charge State Distribution

As the number of monosaccharides composing a glycan increases, the number of possible sites for charge localization increases. This relationship is visualized in Figure S 2. Under normal conditions, we would expect to observe the same molecule in multiple charge states (Maxwell *et al.* (2012)). 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

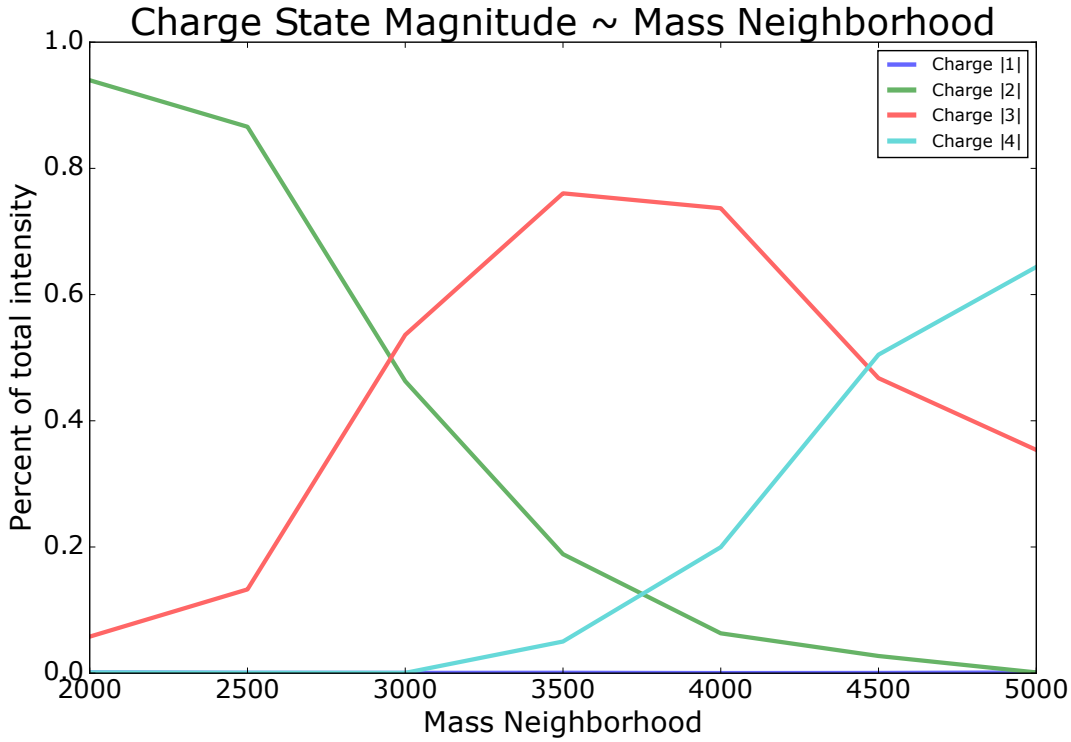


Figure 2: The trend of charge state relative abundance for acidic glycans

permethylated glycans, charge is carried by protons or metallic cation adducts like sodium, the relationship between acidic monosaccharides and charge state propensities is weaker.

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

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. If the model is complete, then this metric is bounded within $[0, 1]$ where 0 corresponds to having no observed charge states and 1 corresponds to all expected charge states being observed. In practice, the model is not complete, where an existing mass range may be missing a charge state in which case $P(c, m)$ is the average over all known values of c in m. When a mass range is required but missing from the model, the model will fall back to a naive model where $P(c, m) = 0.4 \; \forall \; c$ and as such this metric must be clamped to not exceed 1.0. This metric has an exceptional threshold of 0.05 instead of 0.15.

3.3 Adduction Frequency

For the datasets *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 ammonium adduction is not expected to be composition dependent.

$$
\begin{aligned}
m_i &= (\lfloor(\mathcal{M}_i/w)/10\rfloor + 1) * 10 \\
\mathcal{H}_{i,j} &= \frac{\boldsymbol{\mathcal{I}}_{i,a=j}}{\boldsymbol{\mathcal{I}}_i} \\
P(a, m) &= \frac{\sum_{m_i \in m} \mathcal{H}_{i,j}}{\sum_j \sum_{m_i \in m} \mathcal{H}_{i,j}} \\
\mathscr{A}_i &= \sum_{a_{i,j} \in \mathbf{a}_i} P(a_{i,j}, m_i)
\end{aligned} \tag{3}
$$

We fit an ammonium adduction model on *AGP-permethylated-2ul-inj-55-SLens* in order to make our comparison to third-party data less biased given limited sample data. This metric is bounded within $[0, 1]$ where 0 corresponds to having no observed adduction states within the model and 1 corresponds to all observing all adduction states in the model. This metric follows the same behavior as the charge state distribution metric w.r.t. missing information within the model, but will reject chromatograms when this metric score is below 0.15.

We fit a sialylation-aware formate adduction model on a collection of sialylated and unsialylated native N-glycan samples from replicates of the *20150930-06-AGP*, *20151002-02-IGG*, and *20141031-07-Phil-82* datasets. This model was used for *20150930-06-AGP*, *20151002-02-IGG*, *20141031-07-Phil-82* and *20141103-02-Phil-BS*. This model had its upper limit set to 0.7, so it could not contribute a large positive number to the score of a match after logit transformation. This is desirable because we want to be able to eliminate matches which are made with improbable formate adducts when no reasonable adduction state is present.

3.4 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 $\hat{g}_i = \emptyset$and any mass shifting molecular adduct or neutral loss for the matched peak. This computes a per-peak intensity weighted mean G-test comparing the goodness of fit between the experimental envelope and the theoretical isotopic pattern. This metric is bounded within $(-\infty, \infty)$ as the G-test achieves its optimal value at 0, and can take on extreme values towards either signed ∞, however because of the previous deconvolution process, in practice it cannot take on such extreme values and is bounded within $(-\infty, 1]$. This metric is thresholded at 0.15, with any chromatogram scoring below 0.15 being discarded as having insufficient isotopic consistency to interpret.

3.5 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.

$$\begin{aligned} d_i &= \frac{1}{T_i - T_j} \\ \mathscr{T}_i &= 1 - 2\boldsymbol{\mathcal{I}}_i^{-t}\mathbf{I}_i \sum_{j=1}^{J} \boldsymbol{\mathcal{I}}_{i,j} f\left(d_i\left(\mathbf{t}_{i,j} - \mathbf{t}_{i,j-1}\right)\right) \end{aligned} \tag{5}$$

As this metric depends heavily on the speed of the mass spectrometer, a scaling function f must be estimated from the total ion chromatogram to reduce the penalty on slower instruments. When $\frac{1}{J}\sum_j^J T_j - T_{j-1} > 0.2$, $f(x) = x/\left(\frac{1}{J}\sum_j^J T_j - T_{j-1} * 15\right)$, otherwise $f(x) = x$. This metric is bounded within $(-\infty, 1]$ as $(\mathbf{t}_{i,j} - \mathbf{t}_{i,j-1})$ is always positive. This metric is thresholded at 0.15, with any chromatogram scoring below 0.15 being discarded as having insufficient detection consistency to interpret.

3.6 Summarization Score

Each scoring metric $\in [\mathscr{L}_i, \mathscr{C}_i, \mathscr{I}_i, \mathscr{T}_i, \mathscr{A}_i]$ 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 scores and is unexpected to be real, while $s_i > 15$ is consistent with model expectations.

Table 3: Score Thresholds

Chromatographic Peak Shape	0.15
Charge State Distribution	0.05
Adduction Frequency	0.15
Isotopic Pattern Consistency	0.15
Observation Spacing	0.15

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

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

$$\begin{aligned}
0 &= \frac{\partial \ell}{\partial \phi_m}\left((\mathbf{s}-\phi_\mathbf{o})^t(\mathbf{s}-\phi_\mathbf{o}) + \lambda\begin{bmatrix}\phi_o-\tau_o, & \phi_m-\tau_m\end{bmatrix}\begin{bmatrix}\mathbf{L_{oo}} & \mathbf{L_{om}}\\ \mathbf{L_{mo}} & \mathbf{L_{mm}}\end{bmatrix}\begin{bmatrix}\phi_o-\tau_o\\ \phi_m-\tau_m\end{bmatrix}\right) \quad (7)\\
&= \lambda(\phi_o-\tau_o)^t\mathbf{L_{om}} + \lambda\mathbf{L_{mo}}(\phi_o-\tau_o) + \lambda(\phi_m-\tau_m)^t(\mathbf{L_{mm}}^t + \mathbf{L_{mm}})\\
&= 2\lambda\mathbf{L_{mo}}(\phi_o-\tau_o) + 2\lambda\mathbf{L_{mm}}(\phi_m-\tau_m)\\
-\mathbf{L_{mm}}(\phi_m-\tau_m) &= \mathbf{L_{mo}}(\phi_o-\tau_o)\\
(\phi_m-\tau_m) &= -\mathbf{L_{mm}}^{-1}\mathbf{L_{mo}}(\phi_o-\tau_o)\\
\hat{\phi}_m &= -\mathbf{L_{mm}}^{-1}\mathbf{L_{mo}}(\phi_o-\tau_o)+\tau_m \quad (8)
\end{aligned}$$

and w.r.t. ϕ_o

$$\begin{aligned}
0 &= \frac{\partial \ell}{\partial \phi_o}\left((\mathbf{s}-\phi_\mathbf{o})^t(\mathbf{s}-\phi_\mathbf{o}) + \lambda\begin{bmatrix}\phi_o-\tau_o, & \phi_m-\tau_m\end{bmatrix}\begin{bmatrix}\mathbf{L_{oo}} & \mathbf{L_{om}}\\ \mathbf{L_{mo}} & \mathbf{L_{mm}}\end{bmatrix}\begin{bmatrix}\phi_o-\tau_o\\ \phi_m-\tau_m\end{bmatrix}\right) \quad (9)\\
&= -2\mathbf{s} + 2\phi_o + \lambda\left(\mathbf{L_{oo}} + \mathbf{L_{oo}}^t\right)(\phi_o-\tau_o) + \lambda\mathbf{L_{om}}(\phi_m-\tau_m) + \lambda\mathbf{L_{mo}}^t(\phi_m-\tau_m)\\
&= -2\mathbf{s} + 2\phi_o + 2\lambda\mathbf{L_{oo}}(\phi_o-\tau_o) + 2\lambda\mathbf{L_{om}}(\phi_m-\tau_m)\\
\mathbf{s} &= \phi_o + \lambda\left(\mathbf{L_{oo}}(\phi_o-\tau_o) + \mathbf{L_{om}}(\phi_m-\tau_m)\right)\\
&= \phi_o + \lambda\left(\mathbf{L_{oo}}(\phi_o-\tau_o) + \mathbf{L_{om}}(-\mathbf{L_{mm}}^{-1}\mathbf{L_{mo}}(\phi_o-\tau_o)+\tau_m-\tau_m)\right)\\
&= \phi_o + \lambda\left(\mathbf{L_{oo}}(\phi_o-\tau_o) - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}(\phi_o-\tau_o)\right)\\
\mathbf{s}-\tau_o &= \phi_o-\tau_o + \lambda\left(\mathbf{L_{oo}}(\phi_o-\tau_o) - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}(\phi_o-\tau_o)\right)\\
&= \mathbf{I}(\phi_o-\tau_o) + \lambda\left(\mathbf{L_{oo}}(\phi_o-\tau_o) - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}(\phi_o-\tau_o)\right)\\
&= \left[\mathbf{I} + \lambda\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)\right](\phi_o-\tau_o)\\
(\phi_o-\tau_o) &= \left[\mathbf{I} + \lambda\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)\right]^{-1}(\mathbf{s}-\tau_o)\\
\hat{\phi}_o &= \left[\mathbf{I} + \lambda\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)\right]^{-1}(\mathbf{s}-\tau_o)+\tau_o \quad (10)
\end{aligned}$$

5 Estimation of Laplacian Regularization Parameters

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

$$(\mathbf{s}|\phi_\mathbf{o},\tau) \sim \mathcal{N}(\phi_\mathbf{o},\Sigma) \quad (11)$$

$$\Sigma = \rho\mathbf{I} \quad (12)$$

$$\left(\begin{bmatrix}\phi_\mathbf{o}\\ \phi_\mathbf{m}\end{bmatrix}\middle|\tau\right) \sim \mathcal{N}(\mathbf{A}\tau, \lambda^{-1}\mathbf{L}^-) \quad (13)$$

$$(\phi_\mathbf{o}|\tau) \sim \mathcal{N}\left(\mathbf{A_o}\tau, \Sigma_{\phi_o}\right) \quad (14)$$

$$\Sigma_{\phi_o} = \lambda^{-1}\left(\mathbf{L_{oo}} - \mathbf{L_{om}}\mathbf{L_{mm}^{-1}}\mathbf{L_{mo}}\right)^{-1} \quad (15)$$

$$\tau \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right) \quad (16)$$

Fully expanded, this becomes

$$\begin{bmatrix}\mathbf{s}\\ \phi_\mathbf{o}\\ \tau\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\Sigma+\Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \sigma^2\mathbf{A_o}\\ \Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \Sigma_{\phi_o}+\sigma^2\mathbf{A_o}\mathbf{A_o}^t & \sigma^2\mathbf{A_o}\\ \sigma^2\mathbf{A_o}^t & \sigma^2\mathbf{A_o}^t & \sigma^2\mathbf{I}\end{bmatrix}\right) \quad (17)$$

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

$$\mu_{\tau|\mathbf{s}} = 0 + (\sigma^2 \mathbf{A_o}^t)\left(\Sigma + \Sigma_{\phi_o} + \sigma^2 \mathbf{A_o}\mathbf{A_o^t}\right)^{-1}\mathbf{s} \tag{18}$$

$$= \mathbf{A_o}^t\left(\tilde{\rho}\mathbf{I} + \frac{1}{\tilde{\lambda}}\mathbf{L_{oo}^-} + \mathbf{A_o}\mathbf{A_o^t}\right)^{-1}\mathbf{s} \tag{19}$$

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.

$$\mathbf{e} = \mathbf{s} - \hat{\phi}_\mathbf{o} \tag{20}$$

$$\mathbf{H} = \left(\mathbf{I} + \tilde{\lambda}\mathbf{L}\right)^{-1} \tag{21}$$

$$\arg\min_{\tilde{\lambda}} \sum_i^n \left(\frac{e_i}{1 - h_{i,i}}\right)^2 \tag{22}$$

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:

$$\bar{\lambda}_j = \arg\min_{\lambda_i} r_{\lambda_i,\gamma_j,PRESS} \tag{23}$$

$$\tau_{\gamma_j} = |r_{\bar{\lambda}_j,\gamma_j,\tau}| * \left(\frac{\gamma_j}{b} + (1 - \frac{1}{b})\right) \tag{24}$$

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 $\bar{\gamma} = \left[\gamma_j \leftarrow \tau_{\gamma_j} \geq \bar{\tau}_\gamma * 0.95\right]$. 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:

$$\bar{\tau}_\gamma = \max \tau_\gamma \tag{25}$$

$$\bar{\gamma} = \left\{\gamma_j \leftarrow \tau_{\gamma_j} \geq \bar{\tau}_\gamma * 0.9\right\} \tag{26}$$

$$\bar{\lambda} = \left\{\bar{\lambda}_j \leftarrow \gamma_j \in \bar{\gamma}\right\} \tag{27}$$

$$\mathbf{s}_{\gamma_\mathbf{j}} = \left\{s_i \leftarrow s_i > \gamma_j\right\} \tag{28}$$

$$\bar{\tau}_\mathbf{j} = \mu_{\tau|\mathbf{s}_{\gamma_j},\bar{\lambda}_j} \tag{29}$$

$$\hat{\lambda} = \frac{1}{|\bar{\lambda}|}\sum_j \bar{\lambda}_j \tag{30}$$

$$\hat{\tau} = \frac{1}{|\bar{\tau}|}\sum_j \bar{\tau}_\mathbf{j} \tag{31}$$

$$\hat{\gamma} = \frac{1}{|\bar{\gamma}|}\sum_j \bar{\gamma}_j \tag{32}$$

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 10, labeled **??** in the main text.

6 MS^n Signature Ion Criterion

This feature was not used in the main article in order to make the comparison between our results and previously published work more straight forward.

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":

$$I = max(intensity(p)) \tag{33}$$

$$t = I * 0.01 \tag{34}$$

$$p_{oxonium} = \{p_i \leftarrow |ppmerror(mass(p_j), mass(f_g))| < e, f_g \in oxonium(g), f_g \neq \text{Fucose}, intensity(p_i) > t\} \tag{35}$$

$$p_{edges} = \{(p_i, p_j) \leftarrow |ppmerror(mass(p_j) - mass(p_i), mass(f_g))| < e, \tag{36}$$
$$oxonium(f_g) \in g, intensity(p_i) > t, intensity(p_j) > t\}$$

$$s_{oxonium} = \frac{1}{|p_{oxonium}|} \sum_{p_i}^{p_{oxonium}} \left(\frac{intensity(p_i)}{I} \right) * min(log_4|p_{oxonium}|, 1) \tag{37}$$

$$s_{edges} = \frac{1}{|p_{edges}|} \sum_{p_i, p_j}^{p_{edges}} \left(\frac{intensity(p_i) + intensity(p_j)}{I} \right) * min(log_4|p_{edges}|, 1) \tag{38}$$

$$s_g = max(s_{oxonium}, s_{edges}) \tag{39}$$

$$\tag{40}$$

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.

7 Algorithmic Performance on All Datasets

For more details on each sample, please see Table 1.

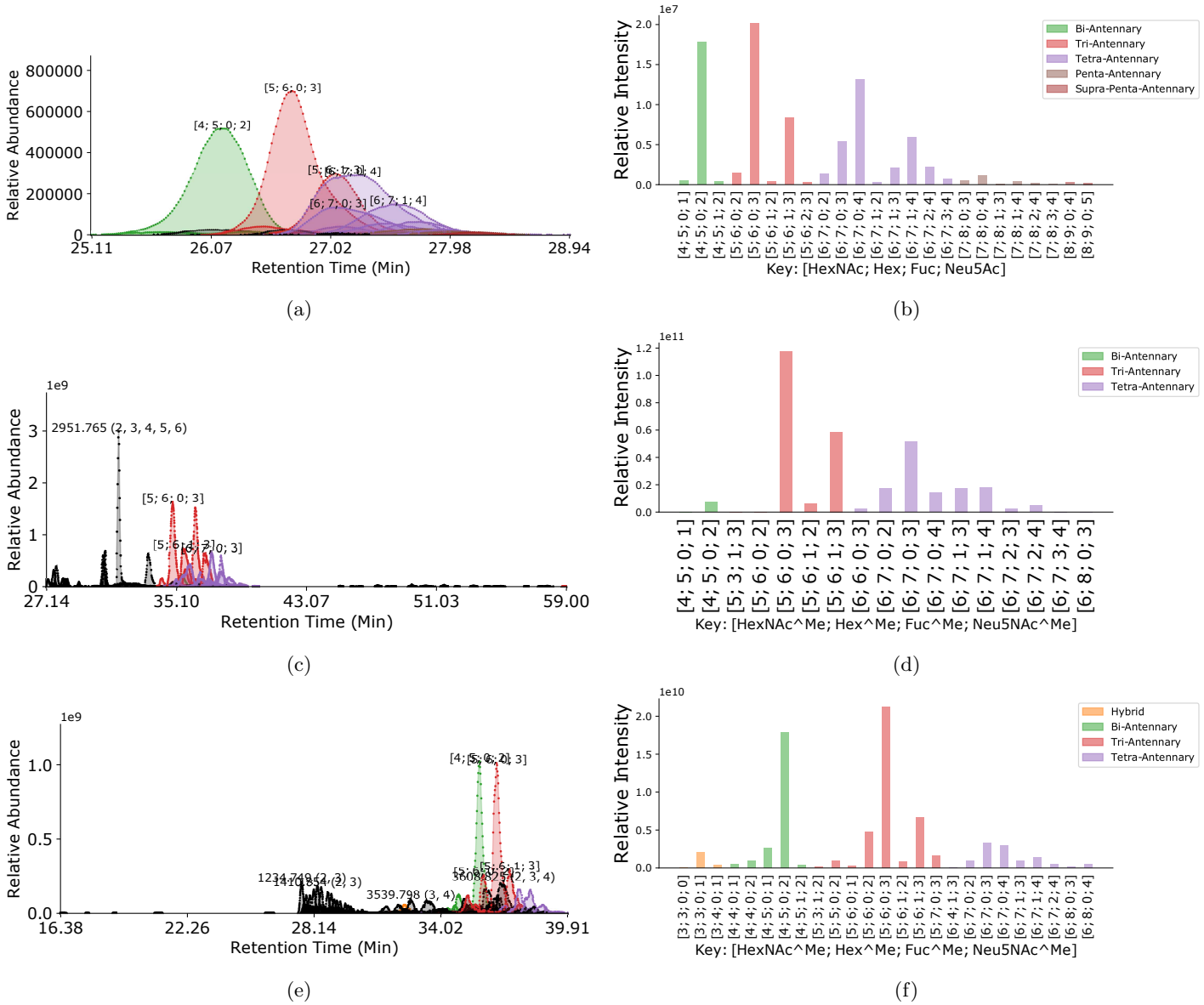
7.1 Results for AGP

We analyzed three different sample workups of N-glycans released from Alpha 1 Acid Glycoprotein. See Table 4 for a comparison of estimated τ values for each sample. For *AGP-DR-Perm-glycans-1* and *AGP-permethylated-2ul-inj-55-SLens*, we used an MS^n Signature Ion Criterion threshold of 0.17 to filter out large contaminants that may be introduced by permethylation reagents.

The estimate of γ for *20150930-06-AGP* was larger than the score for the larger penta-antennary

τ_i	*20150930-06-AGP*	*AGP-DR-Perm-glycans-1*	*AGP-permethylated-2ul-inj-55-SLens*
high-mannose	0.000	0.000	0.000
hybrid	11.520	7.240	21.092
bi-antennary	15.691	12.859	20.627
asialo-bi-antennary	0.000	0.000	13.253
tri-antennary	21.752	21.693	21.550
asialo-tri-antennary	0.000	0.000	6.792
tetra-antennary	15.993	15.276	17.452
asialo-tetra-antennary	0.000	0.000	0.000
penta-antennary	11.446	10.127	7.282
asialo-penta-antennary	0.000	0.000	0.000
hexa-antennary	2.211	0.000	0.000
asialo-hexa-antennary	0.000	0.000	0.000
hepta-antennary	0.000	0.000	0.000
asialo-hepta-antennary	0.000	0.000	0.000
$\hat{\lambda}$	0.99	0.99	0.99
$\hat{\gamma}$	15.74	16.22	17.64

Table 4: Estimated values of smoothing parameters τ, λ, and γ for each AGP-based dataset and using a combinatorial database



(a) (b) (c) (d) (e) (f)

Figure 3: Chromatogram Assignments for *20150930-06-AGP* (a, b), *AGP-DR-Perm-glycans-1* (c, d) and *AGP-permethylated-2ul-inj-55-SLens* (e, f)

7.2 Results for Phil-82

We analyzed native and deutero-reduced and permethylated N-glycans released from virions of Influenza-A Virus strain Phillipines 1982, both samples acquired on a QTOF mass spectrometer. See Table 5 for a comparison of estimated τ values for each sample. In the case of *20141128-11-Phil-82*, MS^nscans were acquired, resulting in lower resolution chromatographic peaks. We observed little ammonium adduction in *20141128-11-Phil-82*. As expected, we observed abundant formate adduction in *20141031-07-Phil-82*, particularly on the high mannose glycans. *20141128-11-Phil-82* also displays considerable in-source fragmentation of the high mannose series, defined by the multimodal chromatographic peaks of smaller high mannose glycans appearing in lower abundance directly under larger peaks for high mannose glycans. This fragmentation, combined with permethylation altering the ionization efficiency of these analytes, makes a direct comparison of glycan composition abundance between *20141031-07-Phil-82* and *20141128-11-Phil-82* inadvisable. We observe markedly different peak shapes between *20141031-07-Phil-82* and *20141128-11-Phil-82* but the relative order of elution is preserved, with the largest high mannose glycans eluting later than the largest observed complex type.

τ_i	*20141031-07-Phil-82*	*20141128-11-Phil-82*
high-mannose	17.070	19.395
hybrid	14.039	17.147
bi-antennary	0.000	0.000
asialo-bi-antennary	16.287	17.689
tri-antennary	0.000	0.000
asialo-tri-antennary	15.220	18.865
tetra-antennary	0.000	0.000
asialo-tetra-antennary	7.103	7.660
penta-antennary	0.000	0.000
asialo-penta-antennary	0.000	3.365
hexa-antennary	0.000	0.000
asialo-hexa-antennary	0.000	0.000
hepta-antennary	0.000	0.000
asialo-hepta-antennary	0.000	0.000
$\hat{\lambda}$	0.99	0.99
$\hat{\gamma}$	16.51	15.50

Table 5: Estimated values of smoothing parameters τ, λ, and γ for each Phil-82-based dataset and using a combinatorial database

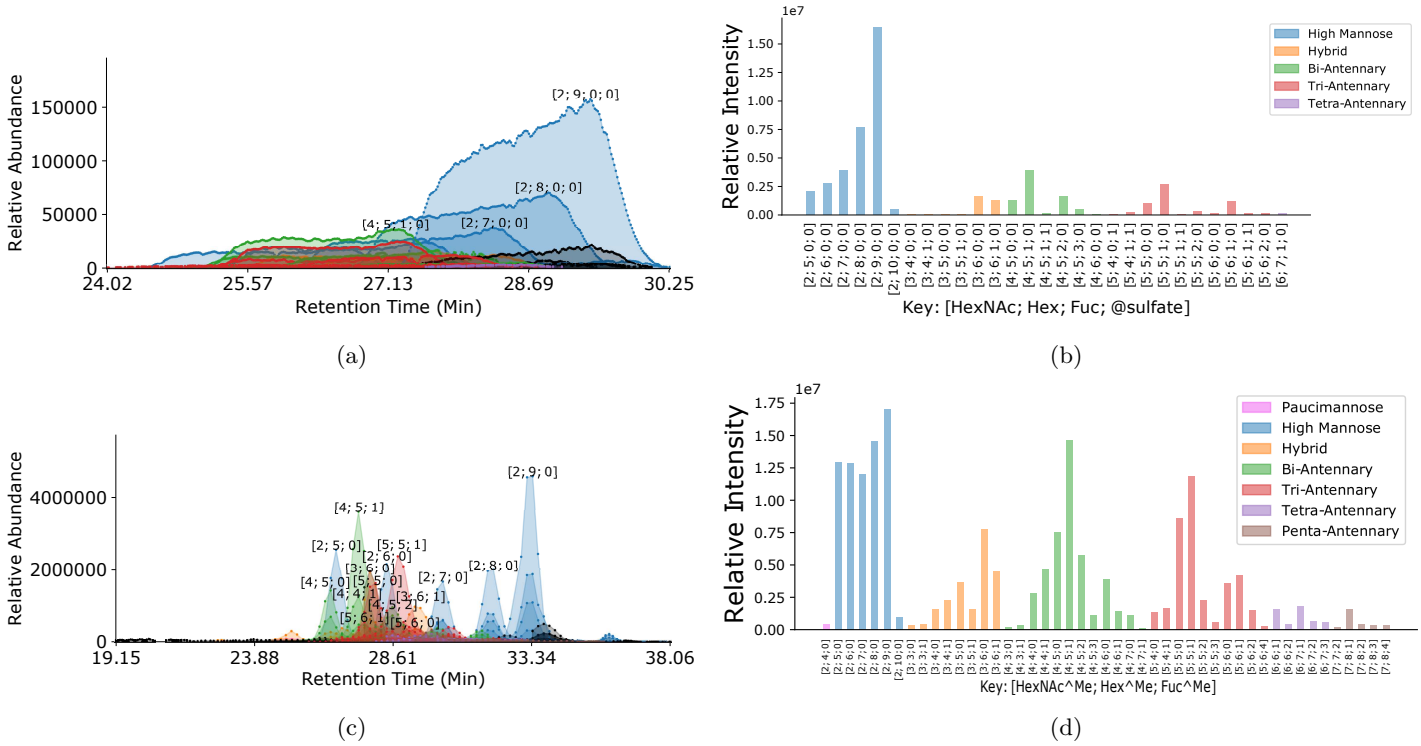


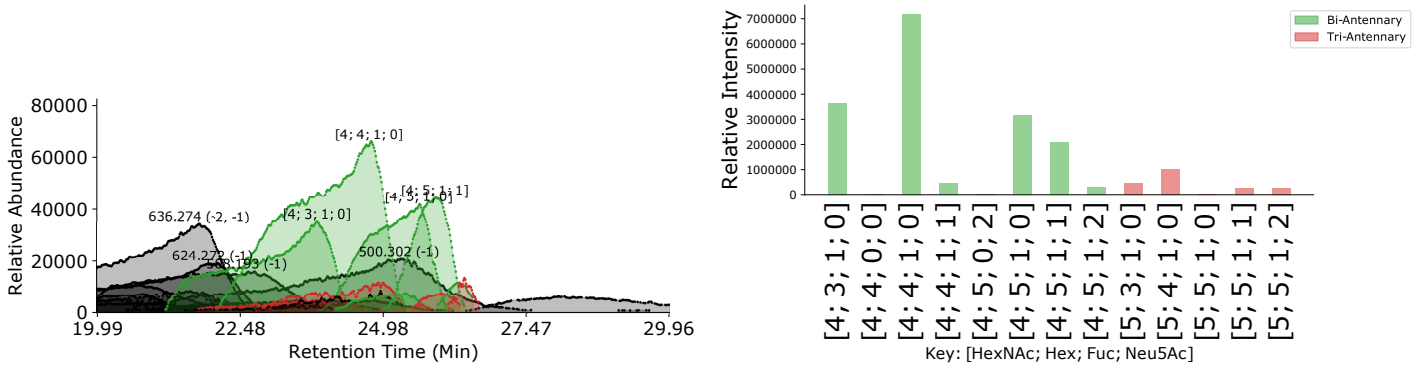
Figure 4: Chromatogram Assignments for *20141031-07-Phil-82* (a, b) and *20141128-11-Phil-82* (c, d)

7.3 Results for IGG

We analyzed native *N*-glycans released from IgG. The estimated τ values shown in Table 6are consistent with the expectation that IgG glycans will be either hybrid or small complex-type structures. These findings are consistent with the results from Peltoniemi *et al.*, 2013, though their study used different sample preparation and instrumentation, and their data were not available for side-by-side comparison. The EICs and integrated abundances for this sample are shown in Figure 5.

τ_i	*20151002-02-IGG*
high-mannose	0.000
hybrid	15.737
bi-antennary	12.594
asialo-bi-antennary	13.614
tri-antennary	7.657
asialo-tri-antennary	15.724
tetra-antennary	4.252
asialo-tetra-antennary	0.000
penta-antennary	0.000
asialo-penta-antennary	0.000
hexa-antennary	0.000
asialo-hexa-antennary	0.000
hepta-antennary	0.000
asialo-hepta-antennary	0.000
$\hat{\lambda}$	0.99
$\hat{\gamma}$	14.12

Table 6: Estimated values of smoothing parameters τ, λ, and γ for IGG using a combinatorial database



(a) (b)

Figure 5: Chromatogram Assignments for *20151002-02-IGG*

8 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.

9 glySpace Integration and Upload

We extracted *N*-glycan structures from GlyTouCan Query Endpoint (http://ts.glytoucan.org/sparql) using the SPARQL query

```
PREFIX glycan: <http://purl.jp/bio/12/glyco/glycan#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX glycoinfo: <http://rdf.glycoinfo.org/glycan/>

SELECT DISTINCT ?saccharide ?glycoct ?motif WHERE {
    ?saccharide a glycan:saccharide .
    ?saccharide glycan:has_glycosequence ?sequence .
    ?saccharide skos:exactMatch ?gdb .
    ?gdb glycan:has_reference ?ref .
    ?ref glycan:is_from_source ?source .
    ?source glycan:has_taxon ?taxon
    FILTER CONTAINS(str(?sequence), "glycoct") .
    ?sequence glycan:has_sequence ?glycoct .
    ?saccharide glycan:has_motif ?motif .
    FILTER(?motif in (glycoinfo:G00026MO))
}
```

and converted each structure into a glycan composition, followed by substituent separation for sulfated and phosphorylated monosaccharides, and filtering out compositions containing units not in [**Hex, HexNAc, Fuc, Neu5Ac, sulfate**]. This procedure is implemented in Python in the included "glyspace_extract_nglycans.py" script.

Note that lines 4-7 restricts the query to only compositions which were in Glycome-DB which came from externally curated sources with taxonomic information, though it is not limited to human *N*-glycans specifically. If these lines are omitted, the query will return over 800 compositions, compared to the expected 275, but the additional compositions will not have been curated. The precise number of compositions returned by this modified query is not fixed as GlyTouCan is a living database, accepting new submissions.

We converted our *N*-glycan compositions into partially determined topologies assuming that the chitobios core was present to ensure that they were classified as *N*-glycans.

From *Perm-BS-070111-04-Serum*

```
{Fuc:1; Hex:5; HexNAc:3; Neu5Ac:1}
{Fuc:2; Hex:5; HexNAc:4; Neu5Ac:2}
{Fuc:2; Hex:6; HexNAc:5; Neu5Ac:3}
{Fuc:2; Hex:7; HexNAc:6; Neu5Ac:3}
{Fuc:2; Hex:7; HexNAc:6; Neu5Ac:4}
{Hex:7; HexNAc:6; Neu5Ac:2}
{Hex:7; HexNAc:6; Neu5Ac:3}
{Hex:8; HexNAc:7; Neu5Ac:3}
{Hex:8; HexNAc:7; Neu5Ac:4}
{Hex:9; HexNAc:8; Neu5Ac:2}
```

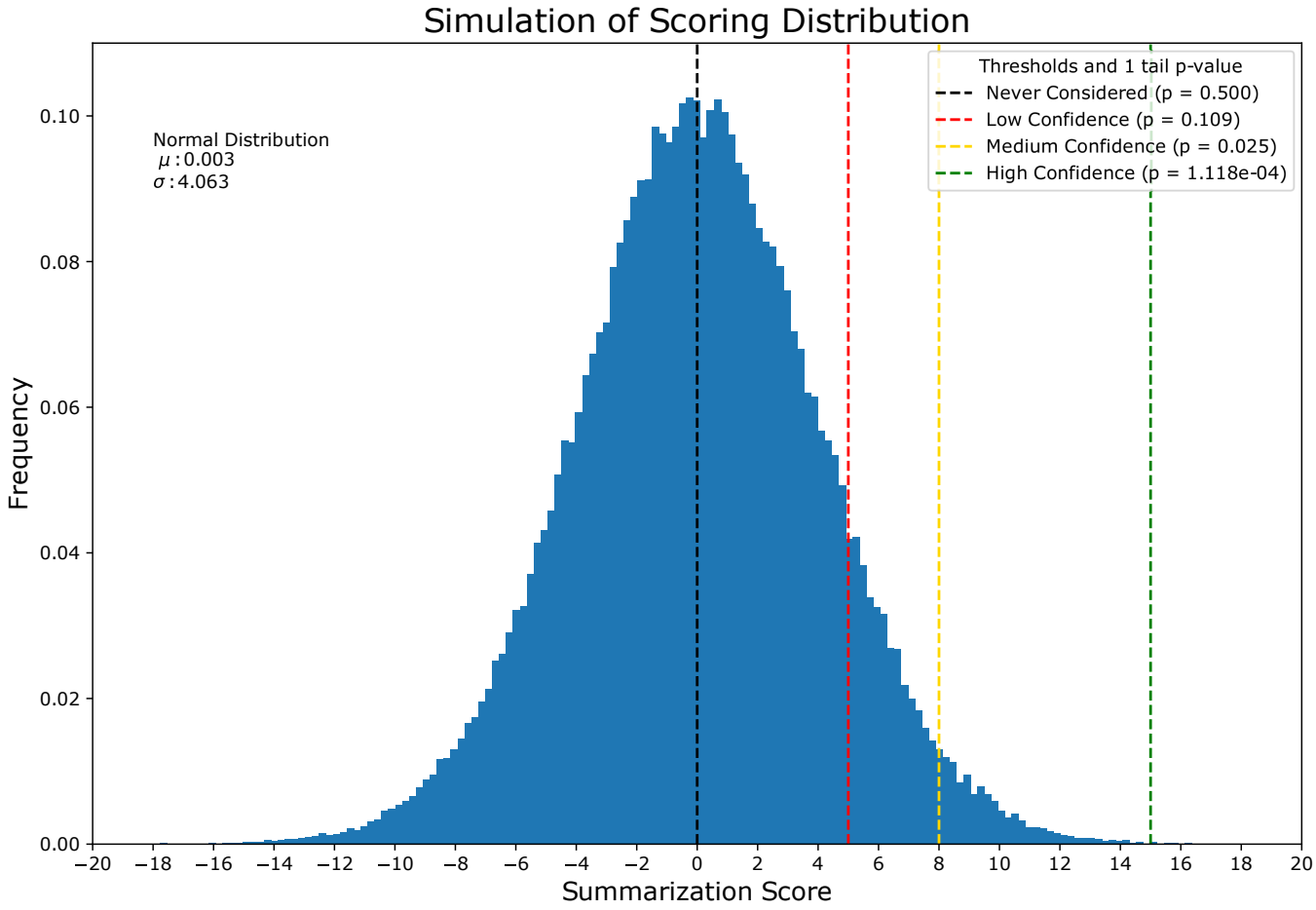
From *20141103-02-Phil-BS*

```
{@sulfate:1; Fuc:1; Hex:4; HexNAc:5}
{@sulfate:1; Fuc:1; Hex:5; HexNAc:4}
{@sulfate:1; Fuc:1; Hex:5; HexNAc:5}
{@sulfate:1; Fuc:2; Hex:4; HexNAc:5}
{@sulfate:1; Fuc:2; Hex:6; HexNAc:5}
{@sulfate:1; Fuc:2; Hex:9; HexNAc:8}
{@sulfate:1; Fuc:3; Hex:4; HexNAc:5}
{@sulfate:1; Fuc:3; Hex:6; HexNAc:5}
{@sulfate:1; Fuc:3; Hex:9; HexNAc:8}
{@sulfate:1; Fuc:4; Hex:6; HexNAc:5}
{@sulfate:1; Fuc:4; Hex:8; HexNAc:7}
{@sulfate:1; Fuc:4; Hex:9; HexNAc:8}
{@sulfate:1; Hex:10; HexNAc:9}
{@sulfate:1; Hex:4; HexNAc:5}
{@sulfate:1; Hex:5; HexNAc:4}
{Fuc:2; Hex:8; HexNAc:7}
{Fuc:3; Hex:7; HexNAc:6}
{Fuc:3; Hex:8; HexNAc:7}
{Fuc:4; Hex:8; HexNAc:7}
{Hex:10; HexNAc:9}
```

10 Simulation of Summarization Score

To simulate the summarization score, we assume that each component scoring feature is drawn from an independent uniform distribution. We sample these five distributions 100,000 times, and for each set of five feature scores compute $\sum_j \text{logit}(f_{i,j})$. According to the central limit theorem, the distribution of the summarization score should be normal, with a mean at approximately 0. We noted two score thresholds, 8 and 15 for lower confidence and high confidence matches. We connect these score thresholds to p values from one-sided tests for significance from thes simulated distribution. The threshold of 8 has a p value of ~ 0.025, while 15 has a p value of $\sim 1.1 \times 10^{-4}$. This distribution is visualized in Figure S 6.

Figure 6: Simulation of Summarization Score



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.

Khatri, K., Klein, J. A., White, M. R., Grant, O. C., Lemarie, N., Woods, R. J., Hartshorn, K. L., Zaia, J., Leymarie, N., Woods, R. J., Hartshorn, K. L., and Zaia, J. (2016a). Integrated omics and computational glycobiology reveal structural basis for Influenza A virus glycan microheterogeneity and host interactions. *Molecular & cellular proteomics : MCP*, **13975**(615).

Khatri, K., Klein, J. A., and Zaia, J. (2016b). Use of an informed search space maximizes confidence of site-specific assignment of glycoprotein glycosylation. *Analytical and Bioanalytical Chemistry*.

Krambeck, F. J. and Betenbaugh, M. J. (2005). A mathematical model of N-linked glycosylation. *Biotechnology and Bioengineering*, **92**(6), 711–728.

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.

Peltoniemi, H., Natunen, S., Ritamo, I., Valmu, L., and Räbinä, J. (2013). Novel data analysis tool for semiquantitative LC-MS-MS2 profiling of N-glycans. *Glycoconjugate journal*, **30**(2), 159–70.

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.