Fast and Accurate Deconvolution of Glycan Mixtures: Bringing Data Science to Glycomics

Rakesh Agrawal
Data Insights Laboratories
San Jose, California, USA

Odysseas Papapetrou
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

Thomas R. Rizzo
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

***Abstract*—We rock! :-)**

I. Introduction

Glycans are one of the four fundamental classes of macromolecules (along with nucleic acids, proteins, and lipids) that comprise living systems [26]. They are ubiquitous. Nearly every cell of every organism is coated with a layer of glycans. They play central role in almost every biological process and are indicated in every major disease, including cancer, diabetes, cardiovascular, congenital, immunological and infectious disorders [1]. Commercially, glycan-based drugs and therapeutics represent a greater than $20 billion market [2].

On the heels of rapid advances in the understanding of nucleic acids (genomics) and proteins (proteomics), glycomics, the study of glycan expression in biological systems, has emerged as the next frontier in the molecular biology revolution [28]. But, compared to other biomolecules, our understanding of glycans is still in infancy. Invention of new and effective analytical tools and techniques is of paramount importance for making further progress in this arena [1].

This paper studies the problem of deconvolving an unknown mixture of glycans. We want to discover what glycans are present in the mixture and in what proportion. Our work has been prompted by two recent technical developments. First, radically new instruments are becoming available for sequencing glycans at a very high rate. For example, the Molecular Physical Chemistry Laboratory at EPFL has recently built an instrument that can simultaneously take three measurements for a given glycan: i) its mass, ii) its collisional cross section that reflects its overall shape, and iii) its high-resolution, vibrational spectrum [19]. These three measurements together can serve as a signature for a glycan. Next, taking cue from how database-centered technologies radicalized genotyping [5], [32], there is consensus and accelerated efforts to build databases of standard glycans [24], [27].

It has been observed that the vibrational spectra are fundamental properties of the molecules and not strongly sensitive to the experimental conditions [9]. As long as measurements are taken at sufficiently low temperatures, the vibrational spectrum obtained will be the same every time, even under slightly different conditions [19]. Thus, having a database of glycans in which each glycan is represented by its signature, one might measure the mass, collisional cross section, and vibrational spectrum of the unknown mixture and use these measurements in conjunction with the database to infer the mixture composition. We present in this paper a highly effective technique for exactly accomplishing this goal.

One unique aspect of our approach is that rather than focussing on peaks as has been conventionally done, we make use of all the information present in a spectrum. We also directly tackle the combinatoric question as to what combination of glycans can best explain the mixture on hand. We view the mixture deconvolution problem as a mathematical optimization problem, and more particularly, a linear programming (LP) problem, which minimizes the difference between what is present in the mixture and what can be accounted by a given combination of glycans. We have not seen the approach we are advocating discussed in the literature [9], [15], [16].

The LP we formulate is large. It has as many pairs of inequalities as the number of data points in the spectrum. Each inequality can have as many terms as the number of glycans in the database. For example, the EPFL instrument measures spectrum values at wave numbers from 3200 cm^{-1} to 3700 cm^{-1} taken at 0.5 cm^{-1} intervals, which gives rise to 1000 pairs of inequalities (not counting the non-neagtivity constraints). Although they are currently not this large, the number of glycans in the database in the future can easily get to in hundreds of millions. But by making use of database indices, we reduce the number of terms present in the inequalities. We only have terms corresponding to those glycans that have mass and cross section values present in the mixture. Thus, we have a large LP, but one that can easily be solved by modern solvers on modern computing infrastructure.

We validate the effectiveness of our approach using real measurements from EPFL's Molecular Physical Chemistry Laboratory. This data contains mixtures of disaccharides, trisaccharides, hexasaccharides, and combinations thereof. In all cases, our technique is able to accurately identify the components and their proportions in the mixture.

In order to test the scalability of our technique, we design generators for creating synthetic datasets based on real glycan data as model. These datasets allow us to test our technique under various operating regions. They also enable us to test the sensitivity of our technique to potential noises in the measurements. Our experiments show that our technique scales well and is robust.

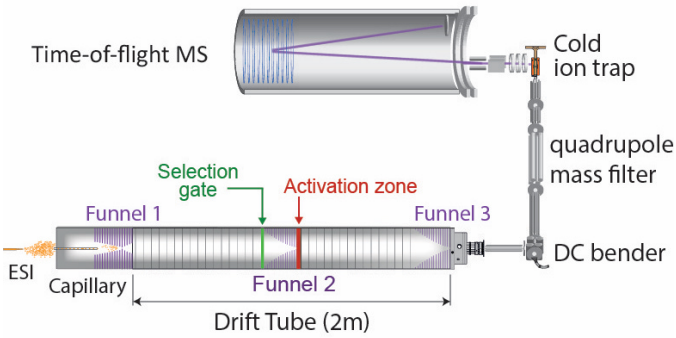

Figure 1: Schematic of the EPFL instrument that combines an electrospray ion source with a time-of-flight mass spectrometer, an ion-mobility-spectrometry drift tube, and a cryogenic ion trap for simultaneously measuring the glycan mass, its average cross section (*aka* drift time), and its infrared vibrational spectrum [19].

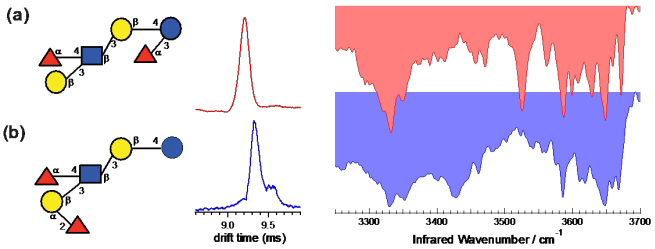

Figure 2: Broad spectra of two human milk hexasaccharides [14][Check with Tom the correct citation]. The zero of the Y-axis in the plots of the spectra in this figure is at the top and the positive values are downwards. The collisional cross section (drift time) has a distribution of values, but only the peak value is generally recorded in the database.

We have made the code and datasets pertaining to our paper publicly available at *Github URL* . [substitute with the real URL]

A. Paper Layout

The rest of the paper is organized as follows. Section II gives background information on instruments used in experimental molecular chemistry that is relevant for understanding the present paper. This section also discusses related work. Section III gives the model and formally defines the problems addressed in the paper. For ease of exposition, Problem 1 considers the special case when only the vibrational spectra are available in the mixture as well as in the database. In Problem 2, in addition to spectra, the mass and collision cross section values are also available. Section IV presents the LP formulation for solving Problem 1. This solution is then enhanced with database indexes to reduce the dimensionality of LP. Section V presents experimental results. We conclude with a summary and directions for future work in Section VI.

II. Backround and Related Work

We first discuss work related to the instrument aspect of our work and then the work related to the data science aspect.

A. Instruments

Several different tools have been used in the past for identifying and characterizing glycans, including various types of high performance liquid chromotography, capilary electrophoresis, nuclear magnetic resonance, and mass spectrometry. Unfortunately, when used alone, none of them can disambiguate all the isomeric glycans. Therefore, the instruments that couple various techniques started to gather attention [8], [13].

In this vein, several research groups reported success in resolving many of the glycan isomers by coupling ion mobility spectrometry (IMS) to mass spectrometry (MS). It was followed by attempts to aditionally bring in spectroscopic dimension. Indeed, IMS-MS coupled with infrared multiphoton dissociation (IRMPD) spectroscopy was successful in indentifying several small glycans [25]. But, the room temperature IRMPD spectra was found to be too broad to uniquely identify large glycans like isomeric disaccharides.

Recently, free-electron laser has been used to obtain spectroscopic fingerprints of oligosaccharides cooled in liquid helium droplets and it was shown that sufficiently resolved spectra could distinguish various types of isomerism [22]. However, the non-linear nature of the spectroscopic technique employed complicates the comparison of spectra across different platforms. In another recent development, EPFL's Molecular Chemistry Laboratory used message-tagging infrared spectroscopy in a cryogenic ion trap in combination with IMS-MS to successfully characterize isomeric glycans [14], [19]. Fig. 1 shows the schematic of the EPFL instrument, which is what we use for collecting data for our experiments.

B. Mixture Deconvolution in Analytical Chemistry

Spectroscopy-based methods for deciphering components of a solutions have been in use in analytical chemistry for a long time. See [9], [15], [16], [31] for comprehensive overviews. Of particular relevance to our work are the library-based methods that attempt to identify an unknown sample by comparing its spectrum with the spectra of the reference molecules kept in a library [11], [12]. Some of the tools used for mixture analysis in the past include linear regression, principal component analysis, pattern recognition procedures, and knowledge-based and expert systems. Experimental results reported in the literature are generally for small number of molecules in the library and mixture consisting of very small number of them [17], [18], [21].

In our literature search, a common attribute we found in prior work is what we consider over-reliance on the peaks in the spectra vs. using all the information present in a spectrum as we do. We make direct use of the combinatorics of what combination of glycans can best explain the mixture on hand. Thus, a simple mixture consisting of just two hexasaccharides shown in Fig. 2 presents tremendous challenge to existing techniques, but our technique easily deconvolves it. Fig. 3 shows the spectra of four hypothetical glycans and their equal proportion mixture. The spectrum of this mixture is completely flat. Thus, any peaks-based technique fails miserably in de-

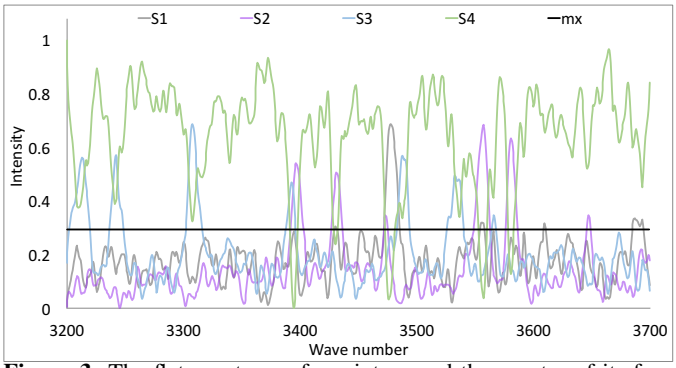

Figure 3: The flat spectrum of a mixture and the spectra of its four constituent glycans. Our technique deconvolves this mixture accurately (with precision=recall=1 and with RMSE$\leq 5 \times 10^{-7}$) whereas the techniques based on identifying peaks in the spectra fail to deconvolve it.

convolving this mixture, but our techniques again correctly convolves it.

C. Mixture Deconvolution in Other Application Domains

The problem of deconvolving mixtures arises naturally in many domains other than molecular chemistry. For example, in forensic science, deconvolution is used to ascertain whether an individual might have contributed to a DNA sample left at a scene of crime [6]. Linear mixture analysis [23] and least square deconvolution [29] are two numerical methods often employed in this deconvolution. Spectral mixture analysis has been widely used in remote sensing applications such as material discrimination where a pixel is generally mixed by a number of materials [10]. The goal of hyperspectral unmixing is to separate the pixel spectra from a hyperspectral image into a collection of constituent spectra and a set of fractional abundances. A good survey of techniques used for such unmixing is available in [3]. One learning from these prior works is that the specific technique that can find good deconvolution of a mixture is dependent on the rules governing how constituents components express in the mixture. The technique we propose for deconvolving glycan mixtures exploits the rules governing the composition of the signature of the mixture from the signature of the constituent glycans.

III. Model and Problems

A. Definitions and Notations

The signature Φ of a glycan consists of its mass $\mathcal{M}$, collision cross section (*aka* drift time) Ω, and infrared spectrum Γ, that is, $\Phi = (\mathcal{M}, \Omega, \Gamma)$. The mass $\mathcal{M} \in \mathbb{R}_{>0}$ is usually obtained using mass spectrometry, whereas the collision cross section $\Omega \in \mathbb{R}_{>0}$ is determined using ion mobility spectrometry [26]. The spectrum Γ is a vector of intensity values measured at different wave numbers in the diagnostic range of interest. We will re-base the lowest wave number of interest to 1 and represent spectrum as $\Gamma = \vec{u}$, where $\forall i \in [1, L]\ \vec{u}[i] \in \mathbb{R}_{\geq 0}$. The EPFL instrument uses cryogenic spectroscopy in an ion trap to obtain infrared spectrum of high resolution [19].

We assume that the signatures of all of the N glycans are available in the database $\mathcal{D}$. Our design point is $\mathcal{D}$ containing hundreds of million signatures ($|\mathcal{D}| = N$ = 1 Billion). Based on the most effective diagnostic range, we take $L = 1000$.

B. Signature of a Mixture of Glycans

We are presented with a mixture X of an unknown number of glycans. We do not know the type of glycans present in X. We also do not know the proportion of various types of glycans in X. However, we may use our instrument to obtain the signature of the mixture, Φ_X.

Although the composition of X is unknown, certain rules govern how its signature Φ_X would look like [9], [26]. It will consist of a set of triples. This set will have as many triples as the number of unique combinations of equal mass and collisional cross section values present in the mixture. Each triple is comprised of a mass, a collisional cross section, and a spectrum. The spectrum is a linear combination of the spectra of the glycans having equal values for mass and collisional cross section. The linear combination is weighted by the relative proportion of the corresponding glycans in the mixture. Thus, suppose there were k glycans present in X, each of which had mass $\mathcal{M}$, endcollisional cross section Ω, and their relative proportions were p_i ($i = 1 \ldots k, \sum_{i=1}^{k} p_i = 1$), then the corresponding triple will have the value $(\mathcal{M}, \Omega, \{\vec{u} \mid \vec{u}[j] = \sum_{i=1}^{k} p_i \vec{u}_i[j], \forall j \in [1, L]\})$.

Example 1: A mixture X consists of three glycans, G_1, G_2, and G_3. Their signatures are as follows: $\Phi_1 = (\mathcal{M}, \Omega, \{u_1^1, u_1^2, \ldots, u_1^L\})$, $\Phi_2 = (\mathcal{M}, \Omega, \{u_2^1, u_2^2, \ldots, u_2^L\})$, and $\Phi_3 = (\mathcal{M}', \Omega', \{u_3^1, u_3^2, \ldots, u_3^L\})$. The relative proportions of G_1 and G_2 are p_1 and p_2 respectively. Since X has two unique combinations of mass and collisional cross section values, namely $(\mathcal{M}, \Omega)$ and $(\mathcal{M}', \Omega')$, the signature of X will consist two triples, and we will have $\Phi_X = \{(\mathcal{M}, \Omega, \{p_1 u_1^1 + p_2 u_2^1, p_1 u_1^2 + p_2 u_2^2, \ldots, p_1 u_1^L + p_2 u_2^L\}), (\mathcal{M}', \Omega', \{u_3^1, u_3^2, \ldots, u_3^L\})\}$.

C. Problems

Our problem is to determine the composition of mixture X, given Φ_X and $\mathcal{D}$. That is, determine the number of each type of glycans present in X.

It has been hypothesized that the spectra obtained using the spectroscopic methods are unique for different glycans [19], [22]. Our first problem, therefore, considers the case where the signatures only have the spectrum component. That is, $\Phi^{\Gamma} = (-, -, \Gamma)$.

Problem 1 (Composition Discovery Using Spectrum-Only signatures): Given a mixture X of glycans and its spectrum-only signature Φ_X^{Γ}, determine the number of each type of glycans present in X, assuming the availability of the database $\mathcal{D}$ of the spectrum-only signatures of all glycans.

Our next problem considers the case when the full signatures are available. That is, $\Phi = (\mathcal{M}, \Omega, \Gamma)$. Our composition discovery algorithm makes use of the additional components of the signatures to gain execution efficiency.

Problem 2 (Composition Discovery Using Full signatures): Given a mixture X of glycans and its full signature Φ_X, determine the number of each type of glycans present in X, assuming availability of the database $\mathcal{D}$ of signatures of all glycans.

IV. Solutions

A. Solution for Problem 1

We have with us the spectrum-only signature of the mixture $\Phi_X^\Gamma = \vec{u}_X$. We need to find k glycans (k can be any value from 1 to N) as well as their proportion in the mixture such that the component-wise weighted sum of their spectra adds up to the spectrum of the mixture (see Section III-B).
We make the following observations:

- The intensity value at each wave number in the mixture is a linear combination of the intensity values of the constitute molecules at the same number.
- The intensity value at each wave number in the mixture can be considered independent of the values at other wave numbers.

These observations lead to the formulation of the problem as a convex optimization program stated below.
This formulation writes the objective as the sum of point-wise absolute differences in the observed spectrum and the spectrum obtained from taking the weighted sum of the spectra of the glycans in the database. The weights are the (unknown) proportions of the various glycans in the mixture.

Convex Program for Problem 1.

$$\begin{aligned} \min \quad & \sum_{\nu=1}^{L} \left| \vec{u}_X[\nu] - \sum_{g=1}^{N} p_g \vec{u}_g[\nu] \right| \\ \text{s.t.} \quad & \sum_{g=1}^{N} p_g = 1 \\ & p_g \geq 0, \quad g = 1 \ldots N \end{aligned}$$

This convex programming problem can be transformed into a linear programming problem, using well-known techniques [4]. The resultant linear program is as follows.

Linear Program for Problem 1.

$$\begin{aligned} & \min \quad \sum_{\nu=1}^{L} \epsilon_\nu \\ \text{s.t.} \quad & -\epsilon_\nu \leq \vec{u}_X[\nu] - \sum_{g=1}^{N} p_g \vec{u}_g[\nu] \leq \epsilon_\nu, \quad \nu = 1 \ldots L \\ & \sum_{g=1}^{N} p_g = 1 \\ & p_g \geq 0, \quad g = 1 \ldots N \end{aligned}$$

This linear program has N variables, $p_1, \ldots, p_N$, one each corresponding to every glycan in the database. The variable p_g gives the proportion of the gth glycan in the mixture. Since p_g values sum up to 1, p_g can also be viewed as the probability that the glycan g is present in the mixture. The larger the probability, the larger the fraction value of the corresponding glaycan in the mixture.
There are L pairs of inequalities, one pair for every wave number ν of the spectra in the diagnostic range. The νth pair corresponds to the intensity values $\vec{u}[\nu]$ at the wave number ν in the spectra of the various glycans. They specify that the weighted sum of these intensity values should add up to the corresponding intensity value of the mixture $\vec{u}_X[\nu]$ at this wave number, modulo some tolerance ϵ (set to 0.00001) to accommodate imprecision in measuring intensity values. The weighted sum is taken over all the glycans in the database, the N of them.
The solution to this linear program will find the optimum combination of glycans and their proportions that best explains the spectrum of the mixture.
Linear programming is well-known to have polynomial time complexity [20] and several fast solvers are readily available [7], [30].

B. Solution for Problem 2

Having the full signature of each of the glycans in the database dramatically cuts down the dimensionality of the linear program.
Given a mixture X, we need to consider the spectra of only those glycans for whom the combination of mass and cross section values is present in the signature of X. This observation follows from the rules for the signature of a mixture outlined in Section III-B. We can thus omit from each of the inequalities many terms in our linear program.
Specifically, we can omit from the inequality j the term $p_i \vec{u}_i[j]$ inside the summation for any glycan i for which

$$\mathcal{M}_i \notin \mathcal{M}_X \text{ OR } \Omega_i \notin \Omega_X.$$

This filtering can be done very efficiently by building indices on the mass and cross section fields of the signatures in $\mathcal{D}$.

V. Experiments

We next present the experiments we performed to study the performance characteristics of our proposed technique. We wanted to assess of our technique can correctly identify the composition of glycan mixtures even if the number of glycans in the mixture is large and there is noise in the measurements of data. We also wanted to examine the scalability of our technique as the number of glycans in the database increases. We start by describing the datasets, both real and synthetic, used in our study. We then describe the performance evaluation methodology, including the performance metrics we employed. Finally, we present the experimental results.
We focus on the performance of our technique on Problem 1. The solution of Problem 2 is predicated on the effectiveness

Mnemonic	Name	$\mathcal{M}$	Ω
hexaS1	Name can be large	fill	fill
hexaS2	fill	fill	fill
triS1	fill	fill	fill
triS2	fill	fill	fill
diS1	fill	fill	fill
diS2	fill	fill	fill
diS3	fill	fill	fill
diS4	fill	fill	fill
diS5	fill	fill	fill
diS6	fill	fill	fill
diS7	fill	fill	fill

Table I: Names, $\mathcal{M}$, and Ω of the real of glycans used in the experiments

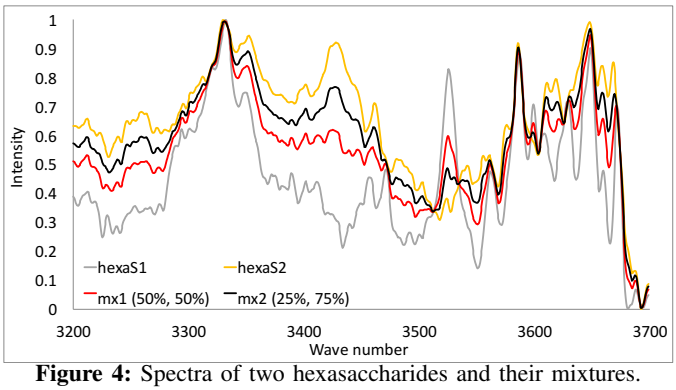

Figure 4: Spectra of two hexasaccharides and their mixtures.

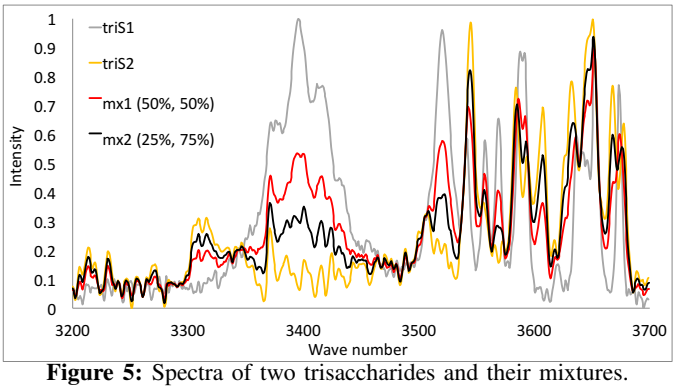

Figure 5: Spectra of two trisaccharides and their mixtures.

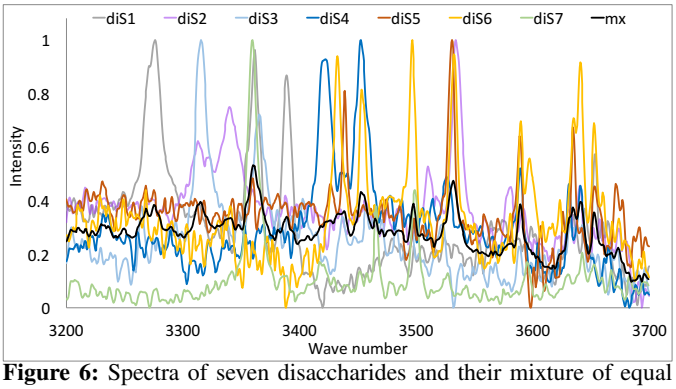

Figure 6: Spectra of seven disaccharides and their mixture of equal proportions.

of indexes on mass and collisional cross section to reduce the candidate set of glycans to those glycans that have the mass and collisional cross section values present in the mixture. The database indexing technology is quite mature by now and the effectiveness of database indices to cut down the size of consideration set of glycans is not questionable any more.

A. Datasets

1) Real Glycans Data: The Laboratory for Molecular Physical Chemistry at École Polytechnique Fédérale de Lausanne provided us the signature data for 11 glycans. Seven of the glycans in this dataset were hexasaccharides, two were trisaccharides, and seven were disaccharides. This dataset comes from the measurements taken on the hybrid instrument designed in the said laboratory following the protocol described in [19]. In this dataset, the saccharides in each category have identical values for mass $\mathcal{M}$ and collision cross section Ω, putting the deconvolution burden on the spectrum Γ alone. Table I shows the names, and $\mathcal{M}$ and Ω values for various glycans [Get table values from Tom]. Their spectra can be seen in Figs. 4-6.

2) Synthetic Glycan Data: We observed from the spectra measured on the EPFL instrument that each spectrum had some high peaks, lesser number of medium peaks, and very many small peaks. Each peak had varying around of spread of intensities around it. There was also varying amount of distance between peaks. Algorithm 1 shows the psuedocode for generating a synthetic spectrum. It samples from uniform distributions the number of high peaks and medium peaks. It then samples from another uniform distribution the wave numbers where the peaks should be placed. Then, for each peak, it samples the intensity value and places a Gaussian at the corresponding wave number. Finally, it places a small Gaussian on every wave number at which no peak has been

Algorithm 1 Glycan Spectrum Generator

Input: a) Mean number of high (medium) resonance peaks, $R_h(R_m)$. b) Parameter values: $L = 1000$, $R_h = 3$, $R_m = 5$, $\tau_h = 0.6$, $\tau_m = 0.1$, $\sigma^2 =$ computed value $(100u_\nu)$.

Sample r_h, the number of high resonance peaks, from $\mathcal{N}(R_h, R_h)$. Similalry, sample r_m, the number of medium peaks, from $\mathcal{N}(R_m, R_m)$.
$F_h \leftarrow$ Sample w/o replacement r_h wave numbers from $\mathcal{U}\{1, L\}$. Similarly F_m.
foreach ν in F_h **do**
    $u_\nu \leftarrow \tau_h$ + offset sampled from $\mathcal{U}(0, 1 - \tau_h)$
    Overlay at ν the Gaussian $u(x) = u_v \exp(-x^2/2\sigma^2)$
foreach ν in F_m **do**
    $u_\nu \leftarrow \tau_m$ + offset sampled from $\mathcal{U}(0, \tau_h - \tau_m)$
    Overlay at ν the Gaussian $u(x) = u_v \exp(-x^2/2\sigma^2)$
foreach $\nu \notin F_h \cup F_m$ **do**
    $u_\nu \leftarrow$ offset sampled from $\mathcal{U}(0, \tau_m)$
    Overlay at ν the Gaussian $u(x) = u_v \exp(-x^2/2\sigma^2)$
Add all the intensity values u_ν at each wave number ν.
Scale the intensity values to lie between 0 and 1.

Algorithm 2 Mixture Generator

1: **Input:** a) Desired number of glycans in the mixture, $|X|$, Minimum proportion of any glycan in the mixture, $p_{\min}$, such that $p_{\min} \times |X| \leq 1$. b) Parameter values: $|X| \in \{1, \ldots, 32\}$, $p_{\min} = 0.03$.

2: /* Pick the glycans that will be present in the mixture */
3: $\vec{x} \leftarrow |X|$ values sampled w/o replacement from $\mathcal{U}\{1, |\mathcal{D}|\}$
4: /* Now pick their corresponding proportions */
5: /* Each glycan should have proportion at least $p_{\min}$ */
6: **for** i in $1 \ldots |X|$ **do**
7: $\quad p_i \leftarrow p_{\min}$
8: /* Partition the remaining proportion to the glycans */
9: residue $\leftarrow 1 - |X| \times p_{\min}$
10: $\vec{s} \leftarrow (|X| - 1)$ values sampled from $\mathcal{U}(0, \text{residue})$
11: Sort values in $\vec{s}$ ascending, s.t. $s_i \leq s_{i+1}$
12: prev $\leftarrow 0$
13: **for** i in $1 \ldots |X| - 1$ **do**
14: $\quad p_i \leftarrow p_i + (s_i - \text{prev})$
15: $\quad \text{prev} \leftarrow s_i$
16: $p_N \leftarrow p_N + (1 - \text{prev})$
17: Output $< x_1, p_1 >, < x_2, p_2 >, \ldots, < x_{|X|}, p_{|X|} >$

placed so far. In the end, the intensity values at every wave number are summed and scaled with respect to the maximum intensity value so that all intensities lie between 0 and 1.

The database $\mathcal{D}$ is created by sampling as many $\mathcal{M}$ and Ω values as the desired number of glycans in the database and generating spectrum Γ for them using Algorithm 1. The glycans are numbered from 1 to $|\mathcal{D}|$.

3) Mixture Generator: We had real data for individual glycans, but did not have the mixtures and their signatures. So, we formulate synthetic mixtures. They are anyway needed for synthetic glycans. Synthetic mixtures also enable us to explore the performance of our algorithm under various operating regions. Algorithm 2 shows the pseudocode of the mixture generator. Having <glycon number, proportion value> pairs from line 17, generating the mixture X using the rules of composition from Section III-B is straightforward. Figs. 4-6 include spectra of the mixtures of some hexa-, tri-, and disaccharides.

4) Distorting Spectra: There can be two sources of errors in measuring a spectrum: i) error in measuring an intensity value, u, and ii) error in measuring a wave number, ν. Algorithm 3 shows the pseudocode for distorting a given spectrum due to these errors.

The EPFL instrument can measure wave numbers with five decimal place accuracy. However, the error in measuring the intensity values can be up to 5%. Our experiments, therefore, focus on sensitivity to error in intensity values in the measured spectrum of the mixtures.

The measurements of mass and collisional cross section values of a glaycan can have small spread around the peak values, but the deviation is tiny and only their peak values are recorded.

Algorithm 3 Distorting a spectrum

1: **Input:** a) spectrum Γ, assuming Γ was generated to have $L + \Delta\nu$ intensity values. b) Parameter values: $\Delta u \in 0.01 \ldots 0.05$, $\Delta\nu \in 1 \ldots 5$.

2: /* error in measuring intensity */
3: **for** i in $1 \ldots L$ **do**
4: $\quad \delta u \leftarrow$ value sampled from $\mathcal{U}(-\Delta u, \Delta u)$
5: $\quad \Gamma[i] \leftarrow \Gamma[i] + \delta u \times \Gamma[i]$
6: Scale Γ s.t. $\forall i\ \Gamma[i] \in 0 \ldots 1$
7: /* error in measuring wave number */
8: $\delta\nu \leftarrow$ value sampled from $\mathcal{U}\{1, \Delta\nu\}$
9: **for** i in $1 \ldots L$ **do**
10: $\quad \Gamma[i] \leftarrow \Gamma[i + \delta\nu]$

B. Performance Metrics

For measuring the quality of the results produced by our algorithm, we take a mixture of glycans of known composition, but hide this information from the algorithm. We run the algorithm on this mixture and quantify the performance with respect to the following metrics:

- *Recall*: the fraction of the glycans present in the mixture that are included in the result produced by the algorithm. That is, recall is defined as:

$$\frac{|\{\text{glycans in the mixture}\} \cap \{\text{glycans in the result}\}|}{|\{\text{glycans in the mixture}\}|}.$$

- *Precision*: the fraction of the glycans that are correct out of those included in the result produced by the algorithm. That is, precision is defined as:

$$\frac{|\{\text{glycans in the mixture}\} \cap \{\text{glycans in the result}\}|}{|\{\text{glycans in the result}\}|}.$$

We will sometime combine the recall and precision values into the one measure:

- *F-score*: the harmonic mean of recall and precision values. That is, F-score is defined as:

$$\text{F} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}.$$

In order to quantify the extent of error made by the algorithm in determining the proportions of various glycans present in a mixture, we use the following additional performance metric:

- *Root Mean square Error (RMSE)*: the square root of the arithmetic mean of the square of the differences between the actual proportion of various glycans in the mixture and their reported proportion in the result. If a glycan is not included in the result, its reported proportion is taken as zero. Suppose the actual proportions of various glycans present in a mixture is $p_1, \ldots, p_{|\mathcal{D}|}$, but the algorithm reports their proportions as $p'_1, \ldots, p'_{|\mathcal{D}|}$. Then, RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_i - p'_i)^2}.$$

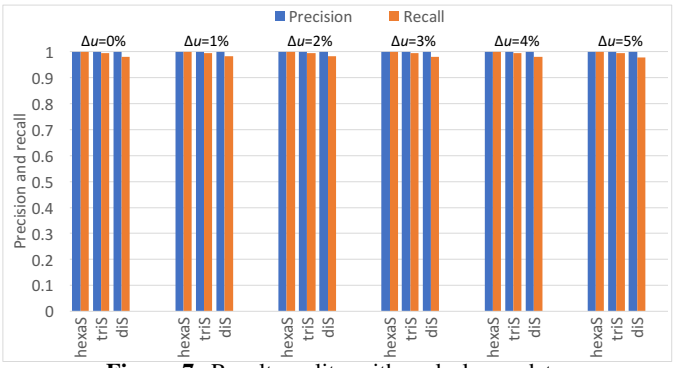

Figure 7: Result quality with real glycan data.

where $n \leq |\mathcal{D}|$ is the number of glycans for which either the reported proportion or the actual proporiton is non-zero.

The linear program provides in its result the proportion values for all the glycans. We process the output of the linear program in the following way to compute the above metrics. First, sort the proportion values in a decreasing order. Now, discard from the result all the proportion values (and their corresponding glycans) that are below a threshold, ρ. Thus, all the above metrics are parameterized with ρ. Increasing the value of ρ generally reduces recall but increases precision. Since more glycans now might have zero proportion values, increasing ρ also increases RMSE. We assume that the presence of less than 1% of a glycan in a mixture is not meaningful and set ρ to 0.01 throughout the paper. Hence, to avoid clutter, we will not write the ρ value when mentioning any of the metrics (unless necessary).

C. Experimental Setup

[State here the LP solver used as well as the important specs of the hardware/software environment.]

D. Results of Experiments using real glycan data

Fig. 7 shows the results our algorithm produces using the spectra of real glycans described in Section V-A1. For a given set of glycans, we generate 100 different mixtures. The proportion of various glycans in a mixture is randomly chosen using steps 8 through 16 of Algorithm 2 given in Section V-A3. Having thus generated the spectrum of a glycan, we distort it using the procedure described in Section V-A4. The results for a set of glycans are averaged over these 100 mixtures. [Increase 100 to 1000 (and verify the confidence interval)? Or, run as many times till the results lie in 95% confidence interval?]

In Fig. 7, we only show the precision and recall values. The root mean square error was negligible in all cases (its maximum value was 0.007 [Check]. There are six sets of histrograms in this figure, one each for different amount of noise in the mixture spectrum. Each set of histograms exhibit the results for three different types of glaycans: hexa-, tri-, and disaccherides.

We see that our algorithm is able to almost perfectly identify various glyans and their proportions in the mixtures, even in the presence of high noise.

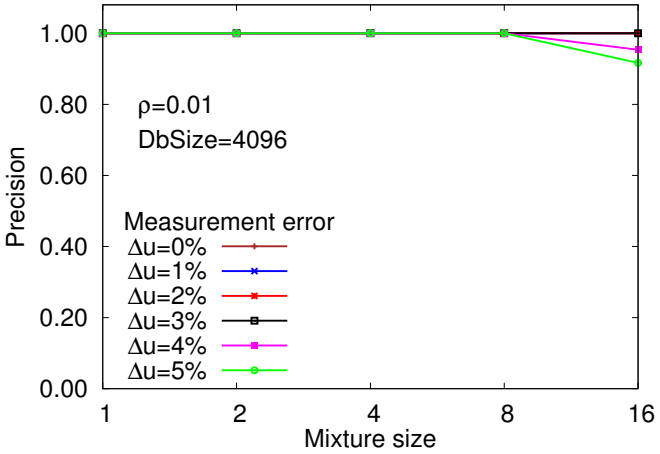

Figure 8: Result quality with synthetic glycan data for varying mixture sizes and levels of measurement errors: Precision.

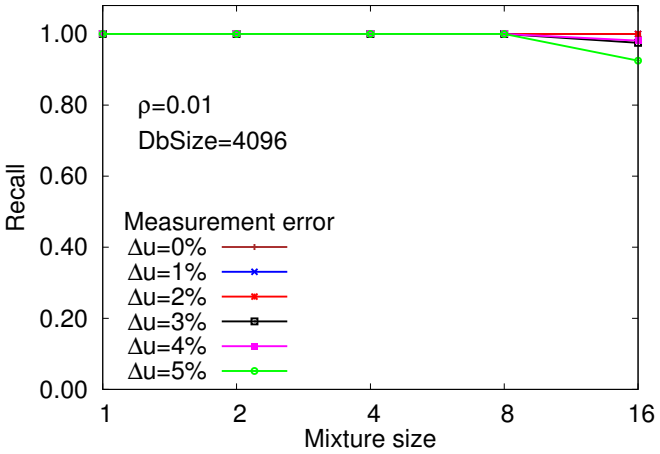

Figure 9: Result quality with synthetic glycan data for varying mixture sizes and levels of measurement errors: Recall.

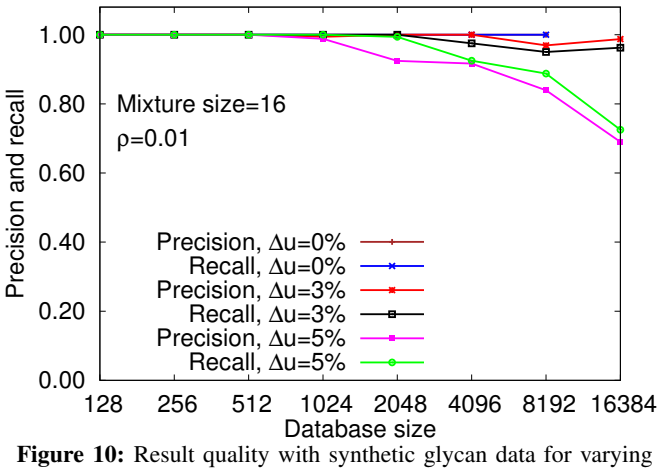

Figure 10: Result quality with synthetic glycan data for varying database sizes.

E. Results of Experiments using synthetic glycan data

In order to stress test our algorithm, we create synthetic glycans, using the procedure described in Section V-A2. We then generate mixtures of them using the procedure described

in Section V-A3 and distorting the mixture spectra using the procedure given in Section V-A4. For each experiment, we create 100 mixtures and average results over them. [Increase 100 to 1000 (and verify the confidence interval)? Or, run as many times till the results lie in 95% confidence interval?]

Figs. 9 and 10 show the results. We again only show the precision and recall values as the root mean square error was negligible in all cases.

In the first figure, we keep the size of the database fixed at 2048 glycans and increase the number of glycans in the mixture from 1 to 128 [check]. We show the plots for various levels of noise in the mixture. Increasing the mixture size or increasing the noise, makes the mixture harder to deconvolve. Note that the X-axis in this figure has logarithmic scale. We again see that our algorithm has excellent performance and this performance is maintained for large mixtures and noise levels.

In figure 10, we keep the mixture size fixed at 32 and vary the number of glycans in the database from 1 to 16384 [check]. We again show the plots for various levels of noise in the mixture and use logarithmic scale for X-axis. We see that our algorithm scales very well.

[Execution time numbers in support of "Fast" in the title of the paper?]

VI. Conclusion

Given the emergent importance of glycomics for future advances in the molecular biology revolution [28], we study the specific problem of discovering the composition of a glycan mixture. This problem arises naturally in many crucial applications, including glycan sequencing, disease diagnosis, and drugs and theropeutics design [2].

Our major technical contributions in this paper include:

- We formulated the problem of deconvolving an unknown mixture of glycans into its constituent components as a convex optimization problem, which we then transformed into a linear programming programming problem for efficiency reasons. The solution of the linear program yields the glycans present in the mixture as well as their proportions.
- We used real glycan data obtained from EPFL's Molecular Chemistry Laboratory to study the the quality of results produced by approach, which showed that our approach is able to correctly discover the mixture compositions.
- In the absence of the availability of large amount of glycan data, we devised synthetic data generators to study the scalability of our approach. These generators are able to create synthetic glycans and their mixtures and admit the modeling of measurement errors. These generators were based on data for real glycans and should be of indendent interest to the glycomics researchers.
- Using the synthetic data thus obtained, we carried out an extensive study of the performance characteristics of our appraoch. This study showed that our approach scales well with increase in database size and the mixture size and able to withstand large measurement erros.

It has been suggested that close collaboration between molecular biologists and chemists and information scientists will be critical for new discoveries in glycomics [1], [28]. The work presented here is the result of a collaborative effort between the Molecular Chemistry Laboratory and the Data-Intensive Applications and Systems Laboratory in EPFL. A byproduct of this work is that we have abstracted an important problem in glycomics and desribed it in a terminology familiar to researchers in data science, paving way for them to enhance and extend the techniques presented in this paper. At the same time, we have provided introduction to some tools and techniques from data science to the researchers in molecular biology and chemistry for them to potentially use them in their own work in the future.

Our approach makes the usual closed world assumption in that the data about all the individual glycans present in a mixture is available in the glycan database. In the future, we would like to investigate how this assumption can be weakened.

Acknowledgments RA's work was partially funded by EPFL - School of Computer and Communication Sciences, Data Intensive Applications and Systems Laboratory. The authors thank Anadiotis Angelos Christos for his insightful comments and suggestions.

References

[1] Agre, P., Bertozzi, C., Bissell, M., Campbell, K. P., Cummings, R. D., et al. Training the next generation of biomedical investigators in glycosciences. *The Journal of clinical investigation 126*, 2 (2016), 405.

[2] Bielik, A. M., and Zaia, J. Historical overview of glycoanalysis. *Functional Glycomics: Methods and Protocols* (2010), 9–30.

[3] Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., et al. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE journal of selected topics in applied earth observations and remote sensing 5*, 2 (2012), 354–379.

[4] Boyd, S., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

[5] Couzin, J. The hapmap gold rush: researchers mine a rich deposit. *Science 312*, 5777 (2006), 1131–1131.

[6] Cowell, R. G., Lauritzen, S. L., and Mortera, J. Identification and separation of dna mixtures using peak area information. *Forensic Science International 166*, 1 (2007), 28–34.

[7] Ferris, M. C., Mangasarian, O. L., and Wright, S. J. *Linear programming with MATLAB*. SIAM, 2007.

[8] Gaunitz, S., Nagy, G., Pohl, N. L., and Novotny, M. V. Recent advances in the analysis of complex glycoproteins. *Analytical chemistry 89*, 1 (2016), 389–413.

[9] Harvey, D. Analytical chemistry 2.0—an open-access digital textbook. *Analytical and bioanalytical chemistry 399*, 1 (2011), 149–152.

[10] Heinz, D. C., and Chang, C.-I. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE transactions on geoscience and remote sensing 39*, 3 (2001), 529–545.

[11] Ito, H., Kameyama, A., Sato, T., Kiyohara, K., Nakahara, Y., and Narimatsu, H. Molecular-weight-tagged glycopeptide library: Efficient construction and applications. *Angewandte Chemie 117*, 29 (2005), 4623–4625.

[12] Kameyama, A., Kikuchi, N., Nakaya, S., Ito, H., Sato, T., et al. A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. *Analytical chemistry 77*, 15 (2005), 4719–4725.

[13] Kanie, Y., and Kanie, O. Addressing the glycan complexity by using mass spectrometry: In the pursuit of decoding glycologic. *Biochemical Compounds 5*, 1 (2017), 3.

[14] KHANAL, N., MASELLIS, C., KAMRATH, M. Z., CLEMMER, D. E., AND RIZZO, T. R. Glycosaminoglycan analysis by cryogenic messenger-tagging ir spectroscopy combined with ims-ms. *Analytical Chemistry* (2017).

[15] KUMAR, N., BANSAL, A., SARMA, G., AND RAWAL, R. K. Chemometrics tools used in analytical chemistry: An overview. *Talanta 123* (2014), 186–199.

[16] LAVINE, B. K., AND WORKMAN JR, J. Chemometrics. *Analytical chemistry 85*, 2 (2012), 705–714.

[17] LO, S.-C., AND BROWN, C. W. Infrared spectral search for mixtures in large-size libraries. *Applied spectroscopy 45*, 10 (1991), 1628–1632.

[18] LO, S.-C., AND BROWN, C. W. Infrared spectral search for mixtures in medium-size libraries. *Applied spectroscopy 45*, 10 (1991), 1621–1627.

[19] MASELLIS, C., KHANAL, N., KAMRATH, M. Z., CLEMMER, D. E., AND RIZZO, T. R. Cryogenic vibrational spectroscopy provides unique fingerprints for glycan identification. *Journal of American Society of Mass Spectrometry* (June 2017).

[20] MEGIDDO, N. On the complexity of linear programming. In *Advances in Economic Theory*, T. F. Bewley, Ed. Cambridge University Press, 1987.

[21] MEYER, K., MEYER, M., HOBERT, H., AND WEBER, I. Qualitative and quantitative mixture analysis by library search: infrared analysis of mixtures of carbohydrates. *Analytica chimica acta 281*, 1 (1993), 161–171.

[22] MUCHA, E., GONZÁLEZ FLÓREZ, A. I., MARIANSKI, M., THOMAS, D. A., HOFFMANN, W., ET AL. Glycan fingerprinting using cold-ion infrared spectroscopy. *Angewandte Chemie International Edition* (2017).

[23] PERLIN, M. W., AND SZABADY, B. Linear mixture analysis: a mathematical approach to resolving mixed dna samples. *Journal of Forensic Science 46*, 6 (2001), 1372–1378.

[24] RANZINGER, R., MAASS, K., AND LÜTTEKE, T. Bioinformatics databases and applications available for glycobiology and glycomics. In *Functional and Structural Proteomics of Glycoproteins*, R. Owens and J. Nettleship, Eds. Springer, 2011, ch. 3, pp. 59–90.

[25] SCHINDLER, B., BARNES, L., GRAY, C., CHAMBERT, S., FLITSCH, S., ET AL. IRMPD spectroscopy sheds new (infrared) light on the sulfate pattern of carbohydrates. *The Journal of Physical Chemistry A 121*, 10 (2017), 2114–2120.

[26] VARKI, A., CUMMINGS, R. D., ESKO, J. D., FREEZE, H. H., STANLEY, P., ET AL. *Essentials of glycobiology*. Cold Spring Harbor laboratory Press, New York, 2009.

[27] WALSH, I., ZHAO, S., CAMPBELL, M., TARON, C. H., AND RUDD, P. M. Quantitative profiling of glycans and glycopeptides: an informatics' perspective. *Current opinion in structural biology 40* (2016), 70–80.

[28] WALT, D., AOKI-KINOSHIT, K. F., BENDIAK, B., BERTOZZI, C. R., BOONS, G.-J., ET AL. *Transforming glycoscience: a roadmap for the future*. National Academies Press, 2012.

[29] WANG, T., XUE, N., AND DOUGLAS BIRDWELL, J. Least-square deconvolution: A framework for interpreting short tandem repeat mixtures. *Journal of forensic sciences 51*, 6 (2006), 1284–1297.

[30] WOLFRAM, S. *The Mathematica Book, 5th Edition*. Wolfram Media, Inc, 2003.

[31] ZAIA, J. Mass spectrometry of oligosaccharides. *Mass spectrometry reviews 23*, 3 (2004), 161–227.

[32] ZOLDOŠ, V., HORVAT, T., AND LAUC, G. Glycomics meets genomics, epigenomics and other high throughput omics for system biology studies. *Current opinion in chemical biology 17*, 1 (2013), 34–40.