Topic Model-based Mass Spectrometric Data Analysis in Cancer Biomarker Discovery Studies

Minkun Wang

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering

Guoqiang Yu, Chair
Yue Wang
Lamine Mili
Kenneth H. Wong
Habtom W. Ressom

April 28, 2017
Arlington, Virginia

Keywords: topic model, computational purification, Bayesian inference, biomarker discovery, liquid/gas chromatography-mass spectrometry (LC/GC-MS)

Topic Model-based Mass Spectrometric Data Analysis in Cancer Biomarker Discovery Studies

Minkun Wang

ABSTRACT

Identification of disease-related alterations in molecular and cellular mechanisms may reveal useful biomarkers for human diseases including cancers. High-throughput omic technologies for identifying and quantifying multi-level biological molecules (e.g., glycans, proteins, and metabolites) have facilitated the advances in biological research in recent years. Liquid (or gas) chromatography coupled with mass spectrometry (LC/GC-MS) has become an essential tool in such large-scale omic studies. Appropriate LC/GC-MS data preprocessing pipelines are needed to detect true differences between biological groups. Challenges exist in several aspects of MS data analysis. Specifically for biomarker discovery, one fundamental challenge in quantitation of biomolecules is owing to the heterogeneous nature of human biospecimens. Although this issue has been a subject of discussion in cancer genomic studies, it has not yet been rigorously investigated in mass spectrometry based omic studies. Purification of mass spectometric data is highly desired prior to subsequent differential analysis.

In this research dissertation, we majorly target at addressing the purification problem through probabilistic modeling. We propose an intensity-level purification model (IPM) to computationally purify LC/GC-MS based cancerous data in biomarker discovery studies. We further extend IPM to scan-level purification model (SPM) by considering information from extracted ion chromatogram (EIC, scan-level feature). Both IPM and SPM belong to the category of topic modeling approach, which aims to identify the underlying "topics"(sources) and their mixture proportions in composing the heterogeneous data. Additionally, denoise deconvolution model (DMM) is proposed to capture the noise signals in samples based on purified profiles. Variational expectation-maximization (VEM) and Markov chain Monte Carlo (MCMC) methods are used to draw inference on the latent variables and estimate the model parameters. Before we come to purification, other research topics in related to mass spectrometric data analysis for cancer biomarker discovery are also investigated in this dissertation.

Chapter 3 discusses the developed methods in the differential analysis of LC/GC-MS based omic data, specifically for the preprocessing in data of LC-MS profiled glycans. Chapter 4 presents the assumptions and inference details of IPM, SPM, and DDM. A latent Dirichlet allocation (LDA) core is used to model the heterogeneous cancerous data as mixtures of topics consisting of sample-specific pure cancerous source and non-cancerous contaminants. We evaluated the capability of the proposed models in capturing mixture proportions of contaminants and cancer profiles on LC-MS based serum and tissue proteomic and GC-MS based tissue metabolomic datasets acquired from patients with hepatocellular carcinoma (HCC) and liver cirrhosis. Chapter 5 elaborates these applications in cancer biomarker discovery, where typical single omic and integrative analysis of multi-omic studies are included.

Topic Model-based Mass Spectrometric Data Analysis in Cancer Biomarker Discovery Studies

Minkun Wang

GENERAL AUDIENCE ABSTRACT

This dissertation documents the methodology and outputs for computational deconvolution of heterogeneous omics data generated from biospecimens of interest. These omics data convey qualitative and quantitative information of biomolecules (e.g., glycans, proteins, metabolites, etc.) which are profiled by instruments named liquid (or gas) chromatography and mass spectrometer (LC/GC-MS). In the scenarios of biomarker discovery, we aim to find out the significant difference on intensities of biomolecules with respect to two specific phenotype groups so that the biomarkers can be used as clinical indicators for early stage diagnose. However, the purity of collected samples constitutes the fundamental challenge to the process of differential analysis. Instead of experimental methods that are costly and time-consuming, we treat the purification task as one of the topic modeling procedures, where we assume each observed biomolecular profile is a mixture of hidden pure source together with unwanted contaminants.

The developed models output the estimated mixture proportion as well as the underlying "topics". With different level's purification applied, improved discrimination power of candidate biomarkers and more biologically meaningful pathways were discovered in LC/GC-MS based multi-omic studies for liver cancer. This research work originates from a broader scope of probabilistic generative modeling, where rational assumptions are made to characterize the generation process of the observations. Therefore, the developed models in this dissertation have great potential in applications other than heterogeneous data purification discussed in this dissertation. A good example is to uncover the relationship of human gut microbiome with the host's phenotypes of interest (e.g., disease like type-II diabetes). Similar challenges exist in how to infer the underlying intestinal flora distribution and estimate their mixture proportions.

This dissertation also covers topics of related data preprocessing and integration, but with a consistent goal in improving the performance of biomarker discovery. In summary, the research help address sample heterogeneity issue observed in LC/GC-MS based cancer biomarker discovery studies and shed light on computational deconvolution of the mixtures, which can be generalized to other domains of interest.

Acknowledgments

First of all, I thank my advisor, Professor Guoqiang Yu for his support and guidance on my study, research and life. I am very grateful for his contributions of time, knowledge and expertise to my research during my Ph.D. study.

I would like to express my sincere gratitude to Professor Habtom W. Ressom, for all the help, support and insights he has provided to me throughout these years in Ressom Lab, where much of my research was conducted. In addition, his kindness, frankness, enthusiasm and positive energy have made it a great pleasure to work with him.

I want to thank the other members of my advisory committee: Professors Yue Wang, Lamine Mili and Kenneth H. Wong, for their invaluable feedback and constructive suggestions to improve the present work. I have been very fortunate to have several scholars who helped and supported me throughout these years. In particular, I would like to thank Dr. Tsung-Heng Tsai, for providing valuable suggestions and feedback on my work. Discussions with him have been a constant source of inspiration. His innovative thinking and high standard on research work have had a significant influence on how I approach a research problem. I thank Dr. Rong Zhang, and Dr. Nenghai Yu, my former supervisors who brought me into the area of informatics six years ago. Over the years, they have extended their support and continued to provide me very helpful suggestions. I enjoyed working with my colleagues in CBIL and Ressom Lab. Especially, I thank Cristina Di Poto, Yi Zhao, and Alessia Ferrarini, for their generous help and selfless contributions in our collaborative projects. Life in Washington D.C. is fantastic. I would love to give special thanks to Na Zhang, Boyu Lu, Frank Cheng, and Xianmei Wu, for sharing the good moments as well as helping me go through the tough days in my Ph.D. life.

Thanks sincerely to the founders of iCarbonX, Dr. Jun Wang and Yingrui Li. The platform they created, together with their tremendous support and encouragement enable me to consistently conduct my research of interest in this field. One year's internship at iCarbonX gives me a opportunity to join my research with applications of promise. Most importantly, I would like to thank my parents, Huaxin Chen and Jinsheng Wang, my uncle, Dr. Shangsheng Chen, my grandparents, Yi Chen and Yuanxiang Zhang. They have always motivated, encouraged, and supported me. This dissertation would not have been possible without their unconditional love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The field of molecular biology has seen many exciting technological developments over the past decades that facilitate the characterization of the average molecular state in a biological sample with unprecedented details. The development and application of high-throughput omic technologies have established a basis for many profound scientific investigations. For instance, one promising application is cancer biomarker discovery study, which associates changes in the levels of multiple biomolecules (e.g., proteins, glycans, glycoproteins, and metabolites) with the onset of cancer to identify clinically relevant diagnostic biomarkers [1–4]. The increased capacity of high-throughput omic approaches, at the same time, has necessitated a growing reliance on computational techniques to extract knowledge from the vast amount of biological data.

Liquid or gas chromatography coupled with mass spectrometry (LC/GC-MS) has been widely used for profiling expression levels of biomolecules in a variety of omic studies. This dissertation covers topics from LC/GC-MS based glycomic data preprocessing to integrative analysis of multi-omic data, and is focused on computational purification, which is a crucial step in the differential analysis of mass spectrum based omic data. These topics are linked through a consistent aim, that is to provide reliable techniques in differential analysis of LC/GC-MS based omic data, specifically for biomarker discovery. We investigate these problems from methodology development to practical application in this dissertation.

Each LC/GC-MS run generates data consisting of thousands of ion intensities characterized by their specific retention time (RT) and mass-to-charge ratio (m/z) values, thus enabling comprehensive profiling of a variety of biomolecules. This high-throughput technique is widely applied to identify candidate markers whose expression levels change between groups of distinct biological conditions [5–7]. In order to ensure an unbiased comparison of the ion intensities, several preprocessing steps including peak detection, retention time alignment, peak matching, normalization, and charge state deconvolution need to be appropriately handled [7]. Typically, these preprocessing steps generate a list of detected peaks with their RT, m/z values and intensities, which are subsequently analyzed using statistical tests to identify

significant differences between groups. Still there exist technical challenges in both the preprocessing pipeline and differential analysis. In LC-MS based glycomics, an individual glycan can generate a set of ion species with several charge states and different adducted forms. Thus, one of the data preprocessing problem is to compare different ion species, including isotopologue ions and multiply charged ions to recover the underneath unique compound information. Automated clustering of ions from the same analyte and annotating them contribute not only to obtaining a summarized feature list but also to achieving more accurate estimation of ion abundances, thereby, reducing the complexity for following statistical analysis and compound identification. In terms of biomarker discovery, we currently detected proteins, N-glycans, and metabolites significantly altered between groups using univariate statistical methods [8–10]. Additionally, multivariate statistical methods are desired to improve the ability to discriminate the cases from controls by taking advantage of the mutual information within the molecules detected by a single omic study as well as the combination of molecules from LC/GC-MS based multiple omic studies. It is challenging but of our interest to investigate if the synergy of multi-omic studies leads to improved performance in distinguishing cases from controls compared to the single omic study.

While the capability of high-throughput technology to yield comprehensive profiling and quantification offers a unique advantage in biomedical research, the heterogeneous nature of the biological samples poses a fundamental challenge in data analysis and interpretation. Specimens, such as tumor tissues and human blood, are typically mixtures of cells with distinct states and types, and usually only part of the constituent cell populations is relevant to the biological question of interest [11, 12]. In some cancer studies, heterogeneity is also observed within the malignant cell population, where multiple cancerous subtypes co-exist [13]. Ideally in a biomarker discovery study, one would perform between-group (cancer versus related disease, cancer versus healthy samples) differential expression analysis for type-specific constituents in samples [14]. However, biospecimens collected from patients usually exhibit some degree of heterogeneity. Moreover, the proportion of cancerous, other disease-related, and healthy components varies across individual samples pre-selected using pathological estimates if available. Therefore, the biomolecular measurements in expression profiles are attributed to multiple sites of origins with various mixture proportions. The cancerous profiles of interest are typically contaminated by other components, leading to unreliable results in downstream differential analyses. Purification of samples is hence highly desired to remove the effects of heterogeneity. Experimental methods for cleaning samples and isolating type-specific constituents are costly and time-consuming. Computational purification methods offer an attractive alternative that is inexpensive and efficient to implement, and can be applied to data already generated without any modifications on experimental procedures. Multiple approaches have been developed to deconvolute gene expression profiles in the past years, varying from linear regression based models [15, 16] to generative probabilistic models [17,18]. However, these methods are not directly applicable to mass spectrometric based omic studies involving quantitative analysis of proteins or metabolites, that is no such purification approaches have been designed to deal with the sample heterogeneity issue in this field. With the increasing volume of these data generated by LC/GC-MS, it is now necessary

to implement the purification of data prior to downstream differential analyses. The topic model-based methods I discuss in this thesis advance the state of the art models that are able to uncover source signals from heterogeneous profiles generated by LC/GC-MS.

Chapter 2

Background

The purpose of this chapter is to provide the background needed in molecular biology, LC/GC-MS data, and probabilistic models to understand the work presented in this thesis.

2.1 Molecular biology

In the field of molecular biology, we have seen many exciting technological developments over the past decade that enables the qualitative and quantitative analysis of the average molecular states of cells comprising a biological sample. Characterizing the association of biomolecules such as proteins, glycoproteins, glycans, and metabolites with various diseases including cancer has proven to be a promising strategy to discover candidate biomarkers [5–10].

Proteomics is the comprehensive analysis of all proteins in a biological system. Emerging technologies (e.g., LC-MS) facilitate delineation of changes on protein levels in a high-throughput fashion. [19] Glycosylation is one of the most common post-translational modifications of proteins. Altered patterns of glycosylation have been associated with various diseases and many currently used cancer biomarkers. Characterizing glycan modifications of proteins in complex proteomes is challenging as glycosylation can occur on multiple sites of peptides involving the attachment of different glycans to each site. Glycomics, as an effective alternative, is to analyze glycans released from proteins and associate the changes with pathological conditions of interest. N-glycans are of particular interest as their involvement in major biological processes, including cell-cell interactions and intracellular signaling, has important implications in disease progression. [8] Also, several enzymes that allow efficient release of this type of glycans have been made available. Through appropriate analytical methods that yield broad coverage of the glycome, characterizing glycomic patterns in serum/plasma of patients with cancer has proven a promising strategy to discover biomarkers for early diagnosis of cancer. Metabolites are molecular fingerprints of what cells do at a

particular point in time. Profiled by GC-MS in metabolomics, these fingerprints can reveal early signs of cancers when the chances for cure are highest.

Each of the individual omic studies can be carried to monitor the states on corresponding biomolecular level. Since these biomolecules are members of strongly intertwined biological pathways and are highly interactive with each other, integrative analysis on the same cohort of biological samples also offers an opportunity to help interpret such interactions and to identify reliable biomarkers.

2.2 LC/GC-MS data

With recent advances of mass spectrometry and separation methods, LC-MS and GC-MS have become essential analytical tools in biomedical research. There has been enormous progress in systems biology and biomarker discovery using LC/GC-MS-based omics [19–21]. Basic principles of LC/GC-MS and preprocessing pipelines for LC/GC-MS data analysis are introduced in this section.

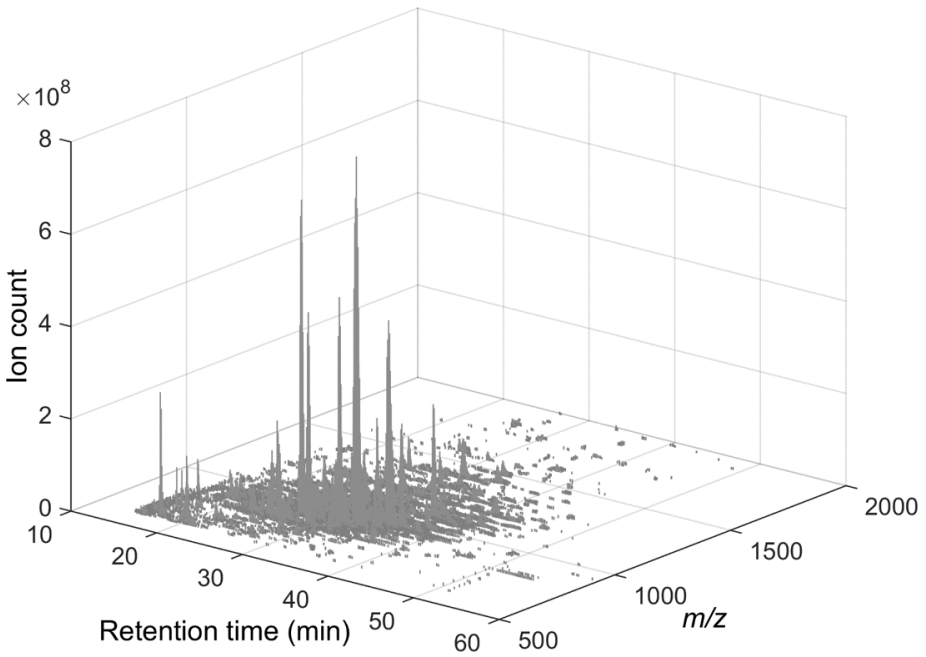


Figure 2.1: An LC-MS run contains RT information in chromatogram, mass-over-charge ratio (m/z) in MS spectrum, and relative ion abundance for each particular ion.

.

2.2.1 Liquid chromatography

Liquid chromatography (LC) is a chromatographic technique used to separate a mixture of compounds that are dissolved in a solvent. Reversed phase high-performance liquid chromatography (RP-HPLC) is the most commonly used method in LC-MS applications. In RP-HPLC, the mixture is dissolved in a mobile phase, composed of water and organic solvents. With a high-pressure pump, the mixture solution is directed into a RP-HPLC column (the stationary phase), using a solvent gradient with increasing organic concentration. The stationary phase is typically hydrophobic or non-polar, while the mobile phase is moderately polar. The choices of column material, type of solvent, and the solvent gradient all play a role in chromatographic separation. Different compounds in the mixture pass through the column at different rates due to the differences in their hydrophobicity and polarity. In RP-HPLC, hydrophilic compounds elute from the column earlier than hydrophobic compounds, and the time where a compound elutes from the column is called elution time or retention time (RT). [22]

2.2.2 Gas chromatography

Gas chromatography (GC) is in principle similar to LC, both for separating mixture of compounds, but has several notable differences. First, the process of separating the compounds in a mixture is carried out between a liquid stationary phase and a gas mobile phase, whereas in LC the stationary phase is a solid and the mobile phase is a liquid. Second, the column through which the gas phase passes is located in an oven where the temperature of the gas can be controlled. Finally, the concentration of a compound in the gas phase is solely a function of the vapor pressure of the gas. Typically, GC is used in seperation of molecules with comparably small weights, e.g., metabolites and lipids.

In most LC/GC-MS applications, a liquid or gas chromatography is coupled on-line to a mass spectrometer. Alternatively, compounds eluting from the column can be collected in aliquots and analyzed by the mass spectrometer afterwards.

2.2.3 Mass spectrometry

Mass spectrometry (MS) is an analytical technique that measures the mass-to-charge ratio (m/z) of charged molecules. The mass spectrometric analysis generates a mass spectrum summarizing abundance of detected ions distinguished in different m/z values. A mass spectrometer consists of three basic components: 1) an ion source that converts sample molecules into charged ions, 2) a mass analyzer that distinguishes the charged ions on the basis of their m/z values, and 3) a detector that counts the number of ions at each m/z value. Depending on the implementation of these components, there are different types of MS instruments. Major instrument configurations include: 1) electrospray ionization (ESI) and

matrix-assisted laser desorption/ionization (MALDI) for the ion source; 2) quadrupole (Q), ion-trap, time-of-flight (TOF), Fourier transform ion cyclotron resonance and Orbitrap for the mass analyzer; and 3) electron multiplier for the detector. The compatibility of different analyzers with different ionization methods varies. For example, while all the analyzers listed above can be used in conjunction with ESI ion source, MALDI is most commonly coupled to a TOF analyzer (MALDI-TOF). In addition, different configurations can be combined to achieve better performance or specific goals, e.g., quadrupole-time-of-flight (Q-TOF) and triple quadrupole (QqQ) mass spectrometry. [23–25]

2.2.4 LC/GC-MS

Mass spectrometers are often coupled with separation methods such as gas chromatography or liquid chromatography, to reduce the chance to analyze coincident molecules and increase the overall dynamic range of detection. Through LC/GC-MS, fewer ions are analyzed simultaneously by the mass spectrometer (compared to the whole sample injected at once). This reduces the ion suppression effect [26]. In addition, molecules with the same molecular weight but different hydrophobicity, e.g., isomers, or different affinity to the column may elute from the column and enter the mass spectrometer at different times, thus reducing ambiguity in differentiating these molecules. An LC/GC-MS run produces a set of MS spectra, acquired at multiple scans of different retention times.

2.2.5 LC/GC-MS data analysis

An LC/GC-MS run contains retention time information in a chromatogram, m/z value in MS spectrum, and relative ion abundance for each particular ion. MS signals of all ions throughout the chromatographic separation are formatted in a three-dimensional map that defines the LC/GC-MS data. An example LC-MS raw data is shown in Figure 2.1. A specified range of retention time and mass value will locate an individual feature (Figure 2.2) , corresponding to a biomolecule. LC/GC-MS can profile thousands of biomolecules in a single run, which necessities an automatic and reliable preprocessing pipeline to extract meaningful features. In order to ensure an unbiased comparison of the ion intensities, several preprocessing steps including noise filtering, deisotoping, peak detection, retention time alignment, peak matching and normalization need to be appropriately handled. Detailed preprocessing steps are discussed in Chapter 3 using LC-MS profiled glycomic data as example. Typically, these preprocessing steps generate a list of detected peaks characterized by their retention times, m/z values and ion intensities. Subsequent statistical analysis is used to identify significant differences in ion intensities across distinct groups. Variance of biology samples can be discovered and associated with the phenotype of interest. Whereas in biomarker discovery studies, we have to rely on the measured biomolecules to uncover the true variance as well as the mixture of other sources underneath the heterogeneous samples.

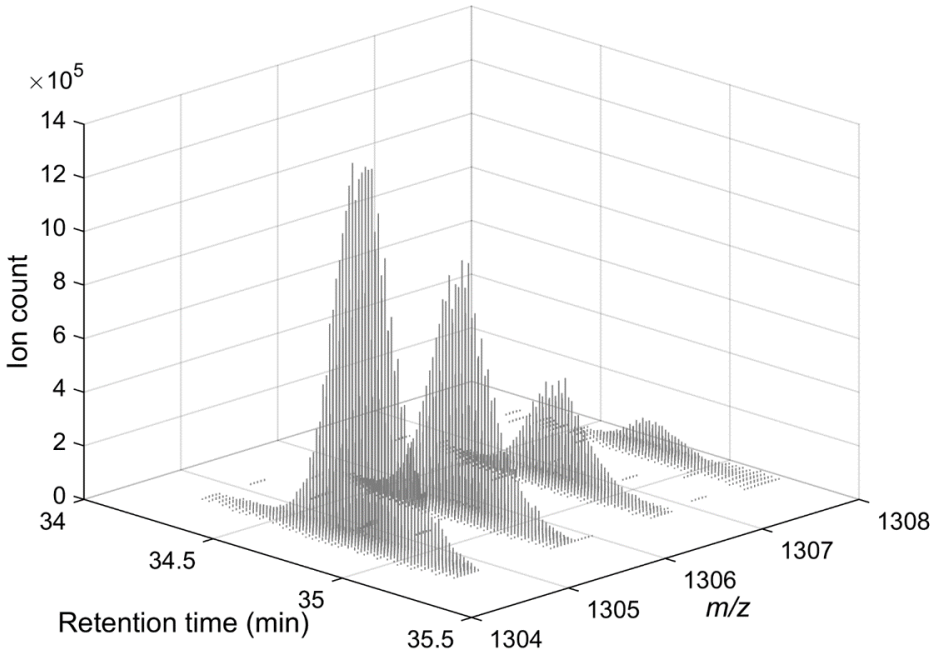


Figure 2.2: A typical feature in LC/GC-MS data corresponding to an individual compound. This compound has a monoisotopic m/z at 1304.8 Th and starts to elute at 34.3 min, lasting for 60 seconds. Multiples peaks are observed due to its isotope distribution.

2.3 Probabilistic generative models, inference, and learning

With reasonable hypotheses, we want to infer the sequence of underlying factors that may have given rise to observations, which could be the outcomes or measurements of a wide range of experiments in molecular biology. Compared to the observed signal itself, the *generative process* by which some unknown (or latent) causal events led to the outcomes should be characterized in a more detailed way [27]. In the cases where outcome or measurement demonstrates variability over multiple samples, the source of this variability with respect to the generative process should be identified as well. For instance, the aforementioned data heterogeneity can be treated as generative outcome of underneath meaningful or contaminant sources.

2.3.1 Probabilistic generative models

To build up a probabilistic generative model (PGM), we can formalize our assumptions of the underlying generative process using observed variables, latent variables, and model parameters. Our observations can be typically represented as a series of D data points each of dimensionality V , denoted as $\mathbf{x}_1, \cdots, \mathbf{x}_D \in \mathbb{R}^V$, and let $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_D\}$. Underlying causal factors are considered by incorporating the latent variables collectively as Z, for whom inference of their values is one of our goals. The relationship or joint probability distribution, of X and Z, is captured by a set of model parameters Θ. A probabilistic generative model is defined by its *complete likelihood function* $\mathcal{L}_C(\Theta|X, Z)$:

$$\mathcal{L}_C(\Theta|X, Z) = p(X, Z|\Theta) = p(Z|\Theta)p(X|Z, \Theta) \tag{2.1}$$

The above likelihood in Equation 2.1 describes how we assume the observations are generated in the creator's perspective. Latent variables Z are first sampled from the distribution $p(Z|\Theta)$, then the observed variables X are sampled from the conditional distribution $p(X|Z, \Theta)$. The objective function $\mathcal{L}_C(\Theta|X, Z)$ is used to evaluate the goodness of fit of the model parameters Θ to both the observed and latent variables. Fitting the parameters of model to the data, i.e., *learning* the model parameters that maximize a likelihood function, is one of the important objectives of PGM. Intuitively, we cannot maximize the complete likelihood function of Equation 2.1 since the latent variables are not observed. We can instead deal with marginal likelihood of the data which integrates out influence from various states of the latent variables:

$$\mathcal{L}_M(\Theta|X) = p(X|\Theta) = \int p(X, Z|\Theta)\partial Z \tag{2.2}$$

Then the parameters can be learned through maximum likelihood estimation (MLE) or maximum a posteriori (MAP) if an appropriate prior distribution over Θ is available in later case. Priors over parameter space help shrink the estimate towards pre-defined values, which prevent PGM from overfitting. The other objective of probabilistic generative modeling is to *infer* the posterior distribution over the latent variables Z given the observed data X:

$$p(Z|X, \Theta) = \frac{P(X|Z, \Theta)P(Z|\Theta)}{\int_Z P(X|Z, \Theta)P(Z|\Theta)\partial Z} \tag{2.3}$$

Using *Bayes Theorem*, the inference uncovers the status of latent variables Z and usually indicates the causal factors with meaningful explanation on the observations.

2.3.2 Expectation maximization

To obtain the MLEs of model parameters in probabilistic models with latent variables, we consider the Expectation Maximization (EM) algorithm, a method that is useful when

it is difficult to acquire the marginal likelihood by integrating out Z (Equation 2.2) but easy to compute the complete likelihood $p(X, Z|\Theta)$. Utilizing Bayes rule and an auxiliary distribution over latent variables Z, we can rewrite the marginal likelihood $P(X|\Theta)$ in the following form:

$$\begin{aligned} p(X|\Theta) &= \frac{p(X, Z|\Theta)}{p(Z|X, \Theta)} \\ &\Downarrow \\ \int q(Z) \cdot \ln p(X|\Theta) \cdot \partial Z &= \int q(Z) \cdot \ln \frac{p(X, Z|\Theta)}{q(Z)} \partial Z - \int q(Z) \cdot \ln \frac{p(Z|X, \Theta)}{q(Z)} \partial Z \\ &\Downarrow \\ \ln p(X|\Theta) &= \mathbb{E}_{q(Z)}[\ln \frac{p(X, Z|\Theta)}{q(Z)}] - \int q(Z) \cdot \ln \frac{p(Z|X, \Theta)}{q(Z)} \partial Z \\ &\triangleq \mathcal{F}(q, \Theta) + \mathrm{KL}(q||p) \end{aligned} \tag{2.4}$$

According to Jensen's Inequality [28], we have the Kullback−Leibler divergence $\mathrm{KL}(q||p) \geq 0$ and the equality holds iff $q(Z) = p(Z|X, \Theta)$. This property leads to the following two-step optimization method which iteratively estimates the parameters as well as infers the posterior $p(Z|X, \Theta)$. At iteration t:

1. **Expectation-step** (E-step): Set $q^{(t)}(Z) = p(Z|X, \Theta^{(t-1)})$, calculate the expected value of the log complete likelihood function w.r.t. $q^{(t)}(Z)$. Since $\mathrm{KL}(q||p) = 0$, $\mathcal{F}(q^{(t)}, \Theta) = \ln p(X|\Theta^{(t-1)}) \propto \mathbb{E}_{q^{(t)}(Z)}[\ln p(X, Z|\Theta)]$. This step changes the auxiliary distribution but updates no parameters.

2. **Maximization-step** (M-step): Maximize the expectation w.r.t. parameters Θ, i.e., find the new $\Theta^{(t)}$ that maximize $\mathcal{F}(q^{(t)}, \Theta)$ which is equivalent to maximize $\mathbb{E}_{q^{(t)}(Z)}[\ln p(X, Z|\Theta)]$. This step optimizes the estimates in parameter space and results in an increased marginal likelihood.

This type of coordinate descent algorithm allows us to alternate between optimizing Θ and inferring $q(Z)$ until convergence at iteration t^*. At convergence, $\mathcal{F}(q^{(t^*)}, \Theta^{(t^*)}) = \ln p(X|\Theta^{(t^*)})$ and hence optimizing $\mathcal{F}$ is equal to optimize the marginal likelihood $\ln p(X|\Theta)$.

2.3.3 Variational expectation maximization

In the scenario where we have difficulty in optimizing $\mathbb{E}_{q(Z)}[\ln p(X, Z|\Theta)]$ w.r.t Θ due to the fact that the posterior $q(Z) = p(Z|X, \Theta^{(t-1)})$ does not have an analytical form, the variational EM (VEM) algorithm can be used [29]. The EM algorithm performs unconstrained minimization of the term of KL divergence to its global minimum by setting $q^{(t)}(Z) = p(Z|X, \Theta^{(t-1)})$.

VEM aims to select q of properties in demand as an approximation of p. These auxiliary distribution q usually come from a specific family of distributions, enabling to calculate $\mathbb{E}_{q(Z)}[\ln p(X, Z|\Theta)]$ in an analytical way. The factorized family of distributions makes the following assumption:

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i), \quad q_{/j}(Z) = \prod_{i \neq j}^{M} q_i(Z_i) \tag{2.5}$$

Then the optimal solution for $q_j(Z_j)$ in this case is:

$$q_j(Z_j) = \frac{\exp(\mathbb{E}_{q_{/j}(Z)}[\ln p(X, Z|\Theta)])}{\int \exp(\mathbb{E}_{q_{/j}(Z)}[\ln p(X, Z|\Theta)])\partial Z_j} \tag{2.6}$$

Through iteratively update all factorized component in $q^{(t)}(Z)$ until convergence, we then update $\mathcal{F}(q^{(t)}, \Theta)$ in the same way as EM. Since we approximate the posterior with factorizable auxiliary distribution, the $\mathrm{KL}(q||p)$ is not guaranteed to be zero after E-step and therefore an increased marginal likelihood should not always be expected after each iteration.

2.3.4 Optimization of constrained variables

There may be no analytical form to maximize $\mathcal{F}(q^{(t)}, \Theta)$, or equivalently, $\mathbb{E}_{q^{(t)}(Z)}[\ln p(X, Z|\Theta)]$ in many cases. Numerical methods, such as first or second-order derivative-based algorithms [30] are typically used to optimize the model parameters. If the parameters are imposed with additional constraints, we can employ auxiliary variables to enable unconstrained parameter optimization. Specifically, for for a vector variable $\boldsymbol{\xi}$ with *threshold constraint* of C on each of its K attributes (e.g., $\xi_k \geq C, \forall k$), we can define an unconstrained variable $\boldsymbol{\omega}$ with the same dimensionality with $\boldsymbol{\xi}$ such that:

$$\xi_k = \exp(\omega_k) + C, \quad \forall k \tag{2.7}$$

The relationship of their first-order derivatives:

$$\frac{\partial \mathcal{F}}{\partial \omega_k} = \sum_{i=1}^{K} \frac{\partial \mathcal{F}}{\partial \xi_i} \cdot \frac{\partial \xi_i}{\partial \omega_k} = \frac{\partial \mathcal{F}}{\partial \xi_k} \cdot \exp(\omega_k), \quad \forall k \tag{2.8}$$

Another common type of constrained variables called *multinomial constraint* has non-negative attributes with their sum remaining a constant. For instance, considering a multinomial constraint on a vector variable $\boldsymbol{\theta}$ of length M, where $\theta_m \geq 0, \ \forall m$ and $\sum_{m=1}^{M} \theta_m = 1$, we define an unconstrained variable $\boldsymbol{\eta}$ with the same dimensionality with $\boldsymbol{\theta}$ such that:

$$\theta_m = \frac{\exp(\eta_m)}{\sum_{\mu=1}^{M} \exp(\eta_\mu)}, \quad \forall m. \tag{2.9}$$

We have,

$$\frac{\partial \theta_\mu}{\partial \eta_m|_{m\neq\mu}} = -\frac{\exp(\eta_m)\exp(\eta_\mu)}{\left(\sum_{j=1}^{M}\exp(\eta_j)\right)^2} = -\theta_\mu \cdot \theta_m \tag{2.10}$$

$$\frac{\partial \theta_m}{\partial \eta_m} = \frac{\exp(\eta_m)}{\sum_{j=1}^{M}\exp(\eta_j)} - \left(\frac{\exp(\eta_m)}{\sum_{j=1}^{M}\exp(\eta_j)}\right)^2 = \theta_m - \theta_m^2 \tag{2.11}$$

Therefore,

$$\frac{\partial \mathcal{F}}{\partial \eta_m} = \sum_{\mu=1}^{M}\frac{\partial \mathcal{F}}{\partial \theta_\mu}\cdot\frac{\partial \theta_\mu}{\partial \eta_m} = -\theta_m \sum_{\mu=1}^{M}\left(\frac{\partial \mathcal{F}}{\partial \theta_\mu}\cdot\theta_\mu\right) + \theta_m \cdot \frac{\partial \mathcal{F}}{\partial \theta_m}, \quad \forall m.$$

In the implementation, we use the PolakRibire conjugate gradient method [30] for the numerical optimization of model parameters.

2.3.5 Markov chain Monte Carlo

An alternative approach to estimate the posterior distribution of parameters is to use Markov chain Monte Carlo (MCMC) methods. Both variational inference (e.g., VEM) and MCMC algorithms have their own advantages and the appropriate approximation methods should be aptly selected according to the factors such as the model complexity and time cost of convergence. In summary, variational inference methods are deterministic and hence work usually faster for small to medium problems. MCMC sampling methods, on the other hand, are often easier to implement and hence applicable to a broader range of complex models. [31, 32]

A large number of samples generated from a Markov chain are employed to give a Monte Carlo estimate. The Markov chain has a unique property that the state of current sample only depends on its preceding sample, i.e., independent from states of other samples in the sequence. The state transition probability can be defined by $T(\boldsymbol{\Theta'}|\boldsymbol{\Theta})$ which indicates the probability of a state switch from $\boldsymbol{\Theta}$ to $\boldsymbol{\Theta'}$. If an infinite number of samples are generated from a target distribution, the Monte Carlo estimates would asymptotically converge to the target according to the law of large numbers. This requires, in MCMC, the Markov chain is able to converge to the target distribution $p(\boldsymbol{\Theta})$, e.g., the posterior distribution of model parameters $\boldsymbol{\Theta}$. Three fundamental properties are required to achieve the convergence and hence ensure the validity of the Monte Carlo estimates:

1. **Homogeneity**: a homogeneous Markov chain has the same transition probabilities for all pairs of adjacent variables in the chain.

2. **Ergodicity**: the sampling can pass from any of the states to anther, irrespective of the choice of initial distribution $p(\boldsymbol{\Theta}^{(0)})$.

3. **Invariant distribution**: the transition probability $T(\boldsymbol{\Theta'}|\boldsymbol{\Theta})$ of the chain leaves the distribution $p(\boldsymbol{\theta})$ invariant in all steps:

$$p(\boldsymbol{\Theta'}) = \sum_{\boldsymbol{\Theta}} T(\boldsymbol{\Theta'}|\boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}), \forall \boldsymbol{\Theta'} \tag{2.12}$$

A sufficient condition for ensuring the hold of the last property is to choose the transition probability satisfying *detailed balance*:

$$T(\boldsymbol{\Theta'}|\boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}) = T(\boldsymbol{\Theta}|\boldsymbol{\Theta'}) \cdot p(\boldsymbol{\Theta'}), \forall \boldsymbol{\Theta'} \text{ and } \boldsymbol{\Theta} \tag{2.13}$$

This property ensures the distribution is left invariant:

$$p(\boldsymbol{\Theta'}) = p(\boldsymbol{\Theta'}) \cdot \sum_{\boldsymbol{\Theta}} T(\boldsymbol{\Theta'}|\boldsymbol{\Theta}) = \sum_{\boldsymbol{\Theta}} T(\boldsymbol{\Theta}|\boldsymbol{\Theta'}) \cdot p(\boldsymbol{\Theta'}) = \sum_{\boldsymbol{\Theta}} T(\boldsymbol{\Theta'}|\boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}) \tag{2.14}$$

Two basic MCMC methods that satisfy detailed balance are Metropolis-Hastings algorithm [31, 33, 34] and Gibbs sampling [35, 36]. As design tools, they are often used to construct a transition of Markov chain for a target distribution, which is typically too complex to sample directly from it.

Metropolis-Hastings algorithm, proposed in 1970 by Hastings [31] was generalized from Metropolis algorithm first published by Metropolis *et al.* in 1953. [33] At particular step t of the algorithm, in which the current state is $\Theta^{(t)}$, a new instance $\Theta^{\star}$ is sampled from a proposal distribution $q(\Theta|\Theta^(t))$ and its acceptance probability $A(\Theta^{\star}|\Theta^(t))$ is determined by the ratio of the target distribution $p(\Theta^{\star})/p(\Theta^{(t)})$ and the transition ratio $q(\Theta^{\star}|\Theta^(t))/q(\Theta^{(t)}|\Theta^{\star})$:

$$A(\Theta^{\star}|\Theta^{(t)}) = \min\left(1, \frac{p(\Theta^{\star})q(\Theta^{(t)}|\Theta^{\star})}{p(\Theta^{(t)})q(\Theta^{\star}|\Theta^{(t)})}\right) \tag{2.15}$$

and,

$$T(\Theta'|\Theta) = A(\Theta'|\Theta) \cdot q(\Theta'|\Theta) \tag{2.16}$$

The property of detailed balance still holds for the transitions defined in Metropolis-Hastings algorithm:

$$\begin{aligned} T(\Theta'|\Theta)p(\Theta) &= \min\left(1, \frac{p(\Theta')q(\Theta|\Theta')}{p(\Theta)q(\Theta'|\Theta)}\right) \cdot q(\Theta'|\Theta)p(\Theta) \\ &= \min\left(p(\Theta)q(\Theta'|\Theta), p(\Theta')q(\Theta|\Theta')\right) \\ &= \min\left(\frac{p(\Theta)q(\Theta'|\Theta)}{p(\Theta')q(\Theta|\Theta')}, 1\right) \cdot q(\Theta|\Theta')p(\Theta') \\ &= T(\Theta|\Theta')p(\Theta') \end{aligned} \tag{2.17}$$

The above procedure shows a transfer of ways in sampling from targeted complex distributions to simple feasible proposal distribution. If the proposal distribution $q(\boldsymbol{\Theta'}|\boldsymbol{\Theta})$ is symmetric, the overall updating scheme is degraded to Metropolis algorithm. In Metropolis-Hastings algorithm, the acceptance probability depends on the ratio of target distributions, which requires no calculation in exact value of the target distribution itself.

Gibbs sampling was first introduced by Geman brothers with an application of image restoration in 1984 [35] and can be treated as a special case of the Metropolis-Hastings algorithm, considering the following proposal distribution:

$$q(\boldsymbol{\Theta'}|\boldsymbol{\Theta}) = p(\Theta'_i|\boldsymbol{\Theta}_{\setminus i})\mathbb{I}[\boldsymbol{\Theta}'_{\setminus i} = \boldsymbol{\Theta}_{\setminus i}] \tag{2.18}$$

The algorithm repeatedly updates each component of the state Θ_i by sampling from its full condition $p(\Theta'_i|\boldsymbol{\Theta}_{\setminus i})$, i.e., conditional distribution of the to-be-updated component given other components fixed (indicated by $\mathbb{I}[\boldsymbol{\Theta}'_{\setminus i} = \boldsymbol{\Theta}_{\setminus i}]$). Updated components can be utilized immediately in the sampling of next component. This design resulted in a merit that its acceptance probability is one:

$$\begin{aligned} A(\Theta'|\Theta) &= \min\left(1, \frac{p(\boldsymbol{\Theta}')q(\boldsymbol{\Theta}|\boldsymbol{\Theta}')}{p(\boldsymbol{\Theta})q(\boldsymbol{\Theta}'|\boldsymbol{\Theta})}\right) \\ &= \min\left(1, \frac{p(\boldsymbol{\Theta})p(\Theta_i|\boldsymbol{\Theta}'_{\setminus i})\mathbb{I}[\boldsymbol{\Theta}_{\setminus i} = \boldsymbol{\Theta}'_{\setminus i}]}{p(\boldsymbol{\Theta}')p(\Theta'_i|\boldsymbol{\Theta}_{\setminus i})\mathbb{I}[\boldsymbol{\Theta}'_{\setminus i} = \boldsymbol{\Theta}_{\setminus i}]}\right) \\ &= \min\left(1, \frac{p(\Theta_i)p(\Theta_i|\boldsymbol{\Theta}_{\setminus i})p(\Theta_i|\boldsymbol{\Theta}'_{\setminus i})\mathbb{I}[\boldsymbol{\Theta}_{\setminus i} = \boldsymbol{\Theta}'_{\setminus i}]}{p(\Theta'_i)p(\theta'_i|\boldsymbol{\Theta}'_{\setminus i})p(\Theta'_i|\boldsymbol{\Theta}_{\setminus i})\mathbb{I}[\boldsymbol{\Theta}'_{\setminus i} = \boldsymbol{\Theta}_{\setminus i}]}\right) \\ &= 1 \end{aligned} \tag{2.19}$$

Gibbs sampling is particularly useful when the full conditional distribution is tractable and can be sampled efficiently by avoiding the meticulous design of proposal distribution and selections of samples.

2.4 Finite mixture models

Finite mixture model, one popular example of probabilistic generative models, is formulated as a convex combination of limited number of components, belonging to the same parametric family of distributions. The property of each combined individual probability density function plays a synergic role in the mixed distribution, which leaves the mixture model a powerful and flexible tool for modeling complex data. [37]

2.4.1 General mixture models

For a given data $\mathbf{X}$ with N observations, each generated from one component distribution parameterized by $\boldsymbol{\Theta}_k$, indicated by latent variable z_n, the marginal likelihood of a general mixture model assuming the observed data are independently distributed is given by:

$$\mathcal{L}_M(\boldsymbol{\Theta}|\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\Theta}) = \prod_{n=1}^{N}\sum_{k=1}^{K} p(z_n = k|\boldsymbol{\Theta}) \cdot p(x_n|z_n, \boldsymbol{\Theta}) \tag{2.20}$$

$\mathbf{Z} = \{z_1, \cdots, z_N\}$ consists of N corresponding random latent variables specifying the identity of the mixture component of each observation, each distributed according to a K-dimensional categorical distribution $p(z_n = k|\boldsymbol{\Theta})$. Probability density function (PDF) of each component denoted by $p(x_n|z_n, \boldsymbol{\Theta})$ determines the probability distribution of an observation. The two most common choices of component PDF are *Multinomial* (for discrete observations) and *Gaussian* a.k.a.*'normal'* (for continuous-value observations) distributions. [38] Their plate notations are given in Figure 2.3.

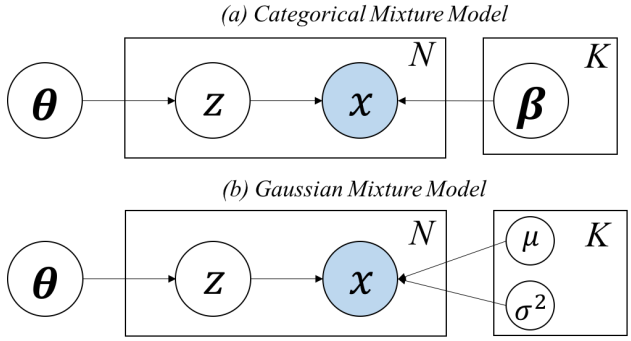


Figure 2.3: Probabilistic representation of two finite mixture models: (a) Categorical Mixture Model (b) Gaussian Mixture Model

Specifically, a typical non-Bayesian mixture model with categorical observations can be characterized using categorical distributions of K components $\{\boldsymbol{\beta}_k, k = 1, \cdots, K\}$, $\boldsymbol{\beta}_k \in \mathbb{R}^V$, mixture weight $\boldsymbol{\theta} \in \mathbb{R}^K$, and component indicator $z_n = \{1, \cdots, K\}$. The complete likelihood (in logarithm) is given by:

$$\mathcal{L}_{\text{CMM}} = \sum_{n=1}^{N} \ln p(x_n, z_n|\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{n=1}^{N} [\ln p(z_n|\boldsymbol{\theta}) + \ln p(x_n|\boldsymbol{\beta}, z_n)] \tag{2.21}$$

where,

$$p(z_n|\boldsymbol{\theta}) = \text{Multinomial}(z_n|\boldsymbol{\theta}) \tag{2.22}$$

$$p(x_n|\boldsymbol{\beta}, z_n) = \text{Multinomial}(x_n|\boldsymbol{\beta}_{z_n}) \tag{2.23}$$

Gaussian mixture model (Figure 2.3(b)) can be characterized in the same format except that the PDF of K components are parameterized by the mean $\{\mu_k, k = 1, \cdots, K\}$ and variance $\{\sigma_k^2, k = 1, \cdots, K\}$ of each component:

$$\mathcal{L}_{\text{CMM}} = \sum_{n=1}^{N} \ln p(x_n, z_n|\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{n=1}^{N} \left[\ln p(z_n|\boldsymbol{\theta}) + \ln p(x_n|\boldsymbol{\mu}, \boldsymbol{\sigma^2}, z_n)\right] \tag{2.24}$$

where,

$$p(z_n|\boldsymbol{\theta}) = \text{Multinomial}(z_n|\boldsymbol{\theta}) \tag{2.25}$$

$$x_n|\boldsymbol{\mu}, \boldsymbol{\sigma^2}, z_n \sim \mathcal{N}(\boldsymbol{\mu}_{z_n}, \boldsymbol{\sigma}^{\mathbf{2}}_{z_n}) \tag{2.26}$$

The parameters in above mixture models can be efficiently estimated via EM algorithms. [39]

2.4.2 Latent Dirichlet allocation

The proposed purification models elaborated in Chapter 4 are based on latent Dirichlet allocation (LDA) [40], a probabilistic topic model originated from natural language processing field. A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. The topic corresponds to the concept of component or underlying factor we early introduced. Therefore, we can treat LDA as a complicated instance of finite mixture models.

Basically, LDA regards words in each document as generated from a set of underlying topics, which are different probability distributions over words. Individual document in the collection has various mixture proportion of the topics. We will use the notation $t_{d,n}$ to represent individual n^{th} word seen in document d, mixing from topics $\{\boldsymbol{\beta}_k, k = 1, \cdots, K\}$. LDA assumes the following generative process:

1. Sample a set of mixture proportion $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$ that describe the composition of each document d $(d = 1, \cdots, D)$ with respect to K latent topics.
2. For each document d, sample N_d topic indicator variables $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$ for $n = 1, \cdots, N_d$, where the indicator $z_{d,n} = \{1, \cdots, K\}$.
3. For each topic indicator $z_{d,n}$ chosen, sample a word $t_{d,n} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{d.n}})$

We describe the complete likelihood of LDA as follow:

$$\mathcal{L}_{\text{LDA}}(\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{d=1}^{D} \ln p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \ln p(t_{d,n}, z_{d,n}|\boldsymbol{\theta}_d, \boldsymbol{\beta}) \tag{2.27}$$

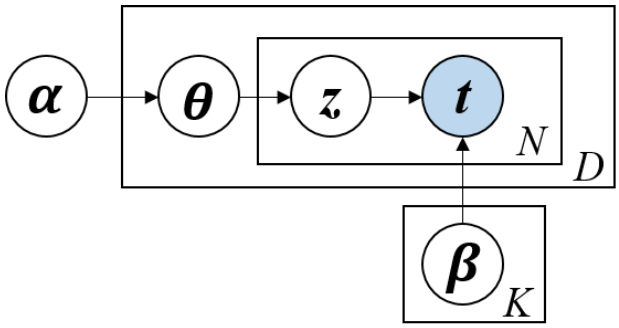


Figure 2.4: Probabilistic generative model of LDA.

where,

$$p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\theta}_d|\boldsymbol{\alpha}, 1) \tag{2.28}$$

$$p(t_{d,n}, z_{d,n}|\boldsymbol{\theta}_d, \boldsymbol{\beta}) = p(z_{d,n}|\boldsymbol{\theta}_d)p(t_{d,n}|z_{d,n}, \boldsymbol{\beta}) \tag{2.29}$$

$$p(z_{d,n}|\boldsymbol{\theta}_d) = \text{Multinomial}(z_{d,n}|\boldsymbol{\theta}_d) \tag{2.30}$$

$$p(t_{d,n}|z_{d,n}, \boldsymbol{\beta}) = \text{Multinomial}(t_{d,n}|\boldsymbol{\beta}_{z_{d,n}}) \tag{2.31}$$

The plate notation of LDA is shown in Figure 2.4. The model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are estimated using variational expectation maximization because for each document d, the posterior distribution over the relevant hidden variables $(\boldsymbol{\theta}_d, \boldsymbol{z}_d|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{t}_d)$ involves calculating the marginal likelihood, which has no analytical form:

$$p(\boldsymbol{t}_d|\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \int \left(\prod_{k-1}^{K} \theta_{d,k}^{\alpha_k - 1} \right) \left(\prod_{n=1}^{N_d} \sum_{k=1}^{K} \theta_{d,k} \beta_{k,t_{d,n}} \right) \partial\theta_d \tag{2.32}$$

The various approximation distribution q can be factorized into:

$$q(\boldsymbol{\theta}_d, \boldsymbol{z}_d|\boldsymbol{\eta}_d, \boldsymbol{\phi}_d) = q(\boldsymbol{\theta}_d|\boldsymbol{\eta}_d) \prod_{n=1}^{N_d} q(z_{d,n}|\phi_{d,n}) \tag{2.33}$$

.

VEM has the following updates in E-step [40]:

$$\eta_{d,k} = \alpha_k + \sum_{n=1}^{N_d} \phi_{d,n,k} \tag{2.34}$$

$$\phi_{d,n,k} \propto \beta_{k,t_{d,n}} \exp(\mathbb{E}_q[\log(\theta_{d,k})|\boldsymbol{\eta}_d]) \tag{2.35}$$

$$\mathbb{E}_q[\log(\theta_{d,k})|\boldsymbol{\eta}_d] = \boldsymbol{\Psi}(\eta_{d,k}) - \boldsymbol{\Psi}\left(\sum_{k'=1}^{K} \eta_{d,k'} \right) \tag{2.36}$$

Then variational M-step estimates the model parameters:

$$\beta_{k,\tau} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} [t_{d,n} = \tau] \phi_{d,n,k} \tag{2.37}$$

$$\frac{\partial \mathbb{E}_q[p(\boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{z} | \boldsymbol{\alpha}, \boldsymbol{\beta})]}{\partial \alpha_k} = D \cdot \left(\boldsymbol{\Psi} \left(\sum_{k'=1}^{K} \alpha_{k'} \right) - \boldsymbol{\Psi}(\alpha_k) \right) + \sum_{d=1}^{D} \left(\boldsymbol{\Psi}(\eta_{d,k}) - \boldsymbol{\Psi} \left(\sum_{k'=1}^{K} \eta_{d,k'} \right) \right) \tag{2.38}$$

The above calculations give the estimation of model parameters as well as the inference of latent variables.

2.5 Research Topics

In this dissertation, we present multiple research topics with respect to mass spectrometric data analysis in biomarker discovery studies. Three aspects of LC/GC-MS data analysis issues are discussed in subsequent chapters.

I. **LC-MS based glycomic data preprocessing**: The complexity of LC-MS data from glycan profiling forms a major bottleneck in data interpretation. One glycan compound is typically ionized with various charge and adduct states. The glycan profile annotation (GPA) work flow is proposed to detect peaks from LC-MS glycomic datasets and annotate them using charge states and adduct information. The GPA utilizes EIC shape information to give a highly confident annotation result. Features are annotated against mass difference table and user-provided known glycan list. Graph-based algorithm is implemented to rapidly separate coeluted compounds. GPA takes deconvoluted ions from the output of DeconTools program and aims to provide a list of unique neutral masses corresponding to putative glycan compounds.

II. **Computational purification**: In addressing sample heterogeneity issue in biomaker discovery, we propose topic models of IPM and SPM to computationally purify mass spectrometric data considering integrated peak intensities and scan-level features, i.e., extracted ion chromatograms (EICs), respectively. The model is further extended to DDM in deconvoluting mixture components including noise. Probabilistic generative models enable flexible representation in data structure and infer sample-specific pure resources. Scan-level modeling helps alleviate information loss during data preprocessing. We evaluated the capability of the proposed models in capturing mixture proportions of contaminants and cancer profiles on LC-MS based serum proteomic and GC-MS based tissue metabolomic datasets as well as synthetic data we generated based on the serum proteomic data in a liver cancer biomarker discovery study.

III. **Multi-omic data integration**: The benefit of an integrative analysis of LC-MS based proteomic, glycomic, and GC-MS based metabolomic datasets in improving our ability to distinguish disease groups is investigated. Complementary to univariate statistical methods, the integrative analysis utilizes mutual information among features to select a panel of features with improved ability to discriminate biologically distinct groups. We hypothesize that integration of multi-omic data by multivariate statistical or machine learning methods, combined with pathway-centric and network-based approaches, will help not only in identifying a panel of biomarkers that leads to improved diagnosis but also in gaining insight into the molecular mechanisms of cancer.

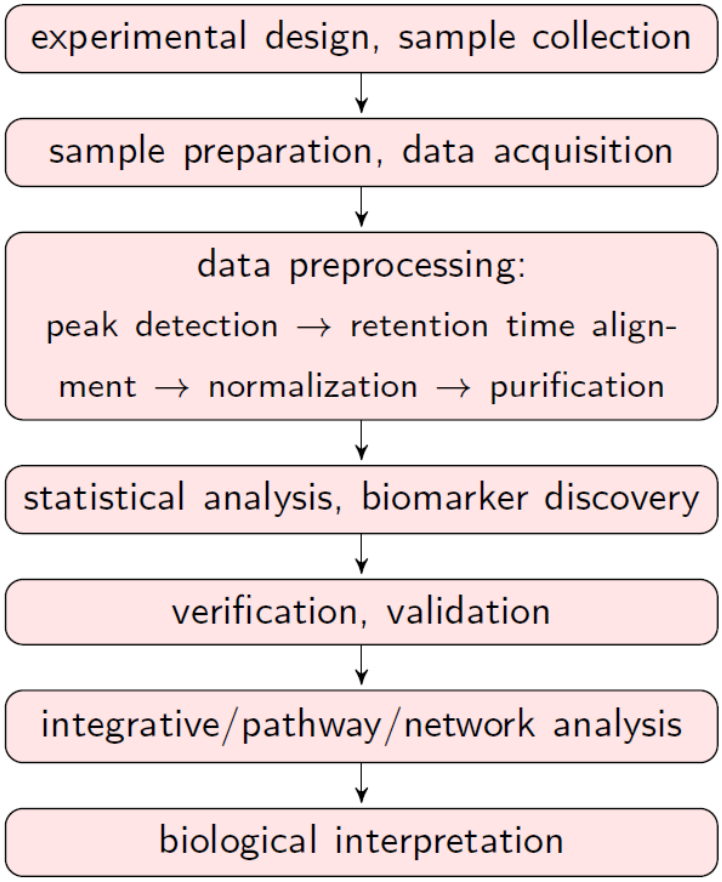


Figure 2.5: A general workflow of LC/GC-MS based omics: starting from experimental design, collected samples are prepared and injected into LC/GC-MS instruments; the acquired raw data are preprocessed to obtain corresponding identified and quantitated compound lists where statistical analysis can be applied to discover candidate biomarkers; further verification and downstream analysis, e.g., integrative, pathway, or network analyses, can be applied to give possible biological interpretation on selected markers.

These topics are tightly linked with consistent goals towards accurate biomarker discovery in

LC/GC-MS based omic studies, which typically follow the workflow depicted in Figure 2.5. The focus of this dissertation is to discuss topic model based purification methods (research topic II).

2.6 List of relevant publications

LC/GC-MS data preprocessing

1. **Wang M.**, Yu G., Mechref Y., Ressom H.W. (2013). GPA: an algorithm for LC-MS based glycan profile annotation. *In the proceedings of the 2013 IEEE International Conference on Bioinformatics and Biomedicine Workshop (BIBMW)*, pp. 16-22.

2. Tsai T.-H., **Wang M.**, Ressom H.W. (2016). Preprocessing and Analysis of LC-MS-Based Proteomic Data. *Statistical Analysis in Proteomics (Methods in Molecular Biology)*, pp. 63-76.

Integrative analysis

3. **Wang M.**, Yu G., Ressom H.W. (2015). Integrative analysis of LC-MS based glycomic and proteomic data. *In the proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 8185-8188.

4. **Wang M.**, Yu G., Ressom H.W. (2016). Integrative analysis of proteomic, glycomic, and metabolomic data for biomarker discovery. *IEEE J Biomed Health Inform*, 20(5), 1225-1231.

5. Ressom H.W., Di Poto C., Ferrarini A., Hu Y., Nezami Ranjbar M.R., Song E., Varghese R.S., **Wang M.**, Zhou S., Zhu R., Zuo Y., Tadesse M.G., Mechref Y. (2016). Multi-omic approaches for characterization of hepatocellular carcinoma. *IEEE Engineering in Medicine and Biology Society.*

Computational purification

6. **Wang M.**, Tsai T.-H., Yu G., Ressom H.W. (2015). Purification of LC/GC-MS based biomolecular expression profiles using a topic model. *In the proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 228-233.

7. **Wang M.**, Tsai T.-H., Di Poto C., Ferrarini A., Yu G., Ressom H.W. (2016). Topic model-based mass spectrometric data analysis in cancer biomarker discovery studies. *BMC Genomics*, 17(Suppl 4):545.

8. **Wang M**, Di Poto C, Ferrarini A, Yu G, Ressom HW (2016). Metabolomic data deconvolution using probabilistic purication models. *In the Proceedings of the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp.204-209. **Awarded the Best Student Paper(1/361)**

Cancer biomarker discovery

9. Tsai T.-H., **Wang M.**, Di Poto C., Hu Y., Zhou S., Zhao Y., Varghese R.S., Luo Y., Tadesse M.G., Ziada D.H., Desai C.S., Shetty K., Mechref Y., Ressom H.W. (2014). LC-MS profiling of N-Glycans derived from human serum samples for biomarker discovery in hepatocellular carcinoma. *J Proteome Res.* 13(11), pp. 4859-4868.

10. Tsai T.-H.*, Song E.*, Zhu R.*, Di Poto C.*, **Wang M.***, Luo Y., Varghese R.S., Tadesse M.G., Ziada D.H., Desai C.S., Shetty K., Mechref Y., Ressom H.W. (2015). LC-MS/MS based serum proteomics for identification of candidate biomarkers for hepatocellular carcinoma. *Proteomics*. 15(13), pp. 2369-2381. (* first authors)

11. Di Poto C., Ferrarini A., Zhao Y., Varghese R.S., Tu C., Zuo Y., **Wang M.**, et al (2016). Metabolomic characterization of hepatocellular carcinoma in patients with liver cirrhosis for biomarker discovery. *Cancer Epidemiol Biomarkers Prev* , DOI: 10.1158/1055-9965.EPI-16-0366.

2.7 Outline of the dissertation

The remainder of this dissertation is organized as follows. Chapter 3 introduces detailed steps in LC/GC-MS data analysis, e.g., the LC-MS profiled glycomic data preprocessing (research topic I). Chapter 4 presents the proposed intensity-level purification model (IPM), scan-level purification model (SPM), and denoise deconvolution model (DDM) in dealing with data heterogeneity issue existing in LC/GC-MS based omics (research topic II). Probabilistic models and corresponding inference methods are discussed. A set of single omic and integrative analysis (research topic III) based mutli-omic studies for biomarker discovery are introduced in Chapter 5. Finally, Chapter 6 concludes this dissertation with a summary and lists possible extension in future work.

Chapter 3

LC/GC-MS data preprocessing

In this chapter, details of mass spectrometric data preprocessing will be described. Specifically, an LC-MS based glycomic study is instanced to discuss the inherent challenges and proposed solutions.

3.1 LC-MS profiled glycans

As a common post-translational modification of proteins, glycosylation is considerably important because its alteration is biologically relevant with various processes in human diseases. Typically, technologies developed for analysis of released glycans aim at indicating the detailed glycan structure, which is crucial for downstream biomarker study and medicinal development. Mass spectrometry (MS) provides an approach for profiling compounds with high throughput and sensitivity, enabling us to deduce putative glycan compositions and abundances. MS records compounds' measured mass-to-charge values (m/z) and their ion counts. LC-MS is used to profile glycans derived from a complex biological sample. The additional chromatography dimension helps separate glycans suppressed in spectrum dimension and hence improves the capability to record more comprehensive information of the glycans present in sample.

Analyte mixture is firstly separated into individual compounds by the column through which the mobile phase analytes are passing. Upon ionization, an individual compound can generate a set of ion species with several charge states and different adducted forms. Thus, one of the data preprocessing steps is to compare different ion species, including isotopologue ions and multiply charged ions to recover the unique compound information underlying them. LC-MS data preprocessing steps include peak detection, deisotoping, deconvolution and feature clustering. A final glycan list with annotated compounds is desirable. However, it takes weeks to manually analyze a LC-MS data. Automated clustering of ions from the same analyte and annotating them contribute not only to obtaining a summarized feature

list but also to achieving more accurate estimation of ion abundances, thereby, reducing the complexity for following statistical analysis and compound identification.

During the past years, several software tools have been developed specifically to annotate glycan profiling data. *GlycoWorkbench* [41] contains functions of drawing glycan structures and matching putative glycan compositions based on matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) spectrum. *Cartoonist* [42] annotates MALDI mass spectral features from a candidate list of N-glycans using biosynthetic rules. Neither of these two tools is amenable to automatical analysis of 2-D LC-MS datasets (because MALDI mass spectrum is of 1 dimension). *GlycReSoft* [43] is designed to detect and score glycomic features from LC-MS datasets. Only ammonium adduct is considered in *GlycReSoft* for annotation. Yet in practice, more than one adduct state are observed in the detected ions. *MultiGlycan* [44] is a recently published tool for automatical annotation of ions in both MALDI and LC-MS datasets. It takes into consideration multiple adduct states (e.g., ammonium, sodium and potassium). However, both *MultiGlycan* and *GlycReSoft* match and group these ions only based on feature's mass information. This could lead to undesirable false annotation for two coeluted ions with close mass but belonging to different compounds.

In metabolomic study, *CAMERA* [45] following *XCMS* [46] automates metabolite profile detection and annotation, using similarity across chromatographic peak shapes in addition to the m/z difference to cluster related features. Glycans have larger mass (e.g., for N-glycan, the mass value is over 1164.6252 Da) than those analysed by *XCMS* and *CAMERA*, where input compounds are no more than 1000 Da. Furthermore, metabolite is ionized into single charge state while glycan can be multiply charged. Therefore, we can not directly apply these tools to glycomic study.

In this research, we proposed a new algorithm for glycan profile annotation (GPA). GPA integrates multiple methods for detecting peaks, clustering related features and annotating ion species based on a list of deconvoluted masses and abundances produced by DeconTools [47]. Compared with existing tools, GPA takes full advantage of extracted ion chromatographic information to give a better quantification and more confident clustering result. GPA workflow is tested using LC-MS data from a biomarker discovery study by profiling permethylated N-glycans derived from human serum.

3.2 Peak detection and quantification

Figure 3.1 shows the GPA workflow, which consists of two major parts: peak detection and feature clustering. Details in each part are explained in the following sections. In general, GPA converts raw LC-MS data into a peak list based on quantitative information derived from Decon2LS/DeconTools. LC-MS data contain thousands of features and ubiquitous noise. An original feature can be modeled as a product of its isotope pattern in the m/z dimension and elution profile in the retention time dimension. Features may interleave in

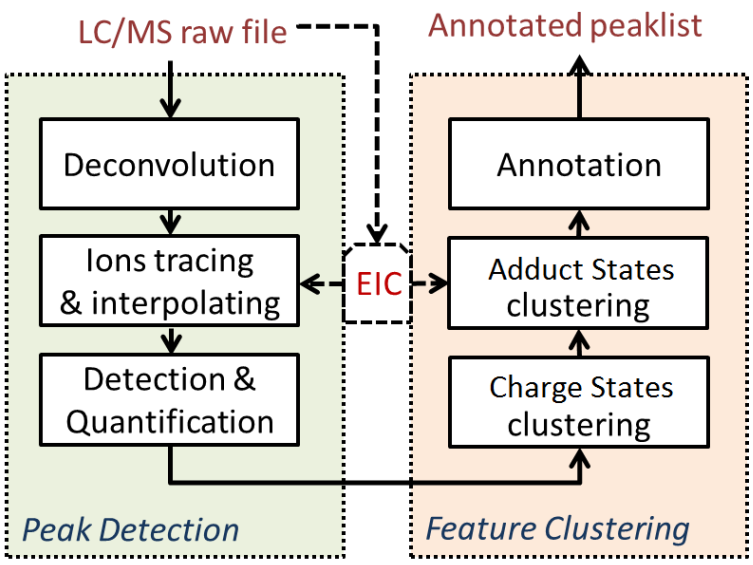


Figure 3.1: GPA workflow. The proposed workflow consists of two parts: peak detection part and feature clustering part. GPA transfers the input LC-MS dataset into annotated peaklist.

both dimensions. Hence feature deconvolution steps are necessary before quantification.

3.2.1 Deconvolution

Raw LC-MS data are analysed using the DeconTools, an open-source software package, to generate a deisotoped ions map. DeconTools employs the THRASH algorithm [48] to de-isotope mass spectra. By setting appropriate parameters and averagine formula, we can employ DeconTools to fit the isotope patterns and therefore deduce the charge states for candidate features. Based on m/z values and charge states, the putative mass values are easily recovered. Mass information is helpful for identification. Ions originating from the same analyte but ionized with different charge states (hence with different m/z values) would have the same mass value. An ion is represented by its molecular weight (MW), charge state, retention time and intensity information.

3.2.2 MW grouping and interpolating

Based on the deconvoluted map, ions are grouped along the mass dimension (MW grouping) across all the retention time points into traces with selected masses and fixed tolerance range. Each round of grouping starts from a seed ion selected from the ungrouped ions list with their abundances in descending order. Seed's mass value, MW_0, is set as the center of

molecular weight window $[MW_0 - \epsilon_G, MW_0 + \epsilon_G]$. The seed's scan (RT location) is set as the center of the final trace. From this scan on, forward tracing and backward tracing are carried out successively across scans. Within predetermined tracing steps, all the ions on those scans with mass falling into the MW window are grouped into this trace. After all the ions find their traces, the MW grouping is done.

We observed that some of these traces are not perfectly consecutive. Missing scans or intervals within a trace may be due to tight tolerance or false negatives mis-detected in DeconTools. A first-phase filter is set to screen out those traces with low quality, e.g. low scan density, low abundance or short elution time. Ions that may be mis-detected in DeconTools lead to flaws in traces. Evidences can be found by checking the corresponding extracted ion chromatograms (EICs) from raw data. A valid EIC profile has natural continuity without sharp flaws. We introduce interpolation, based on EIC information, to repair these flaws and give a more precise quantification.

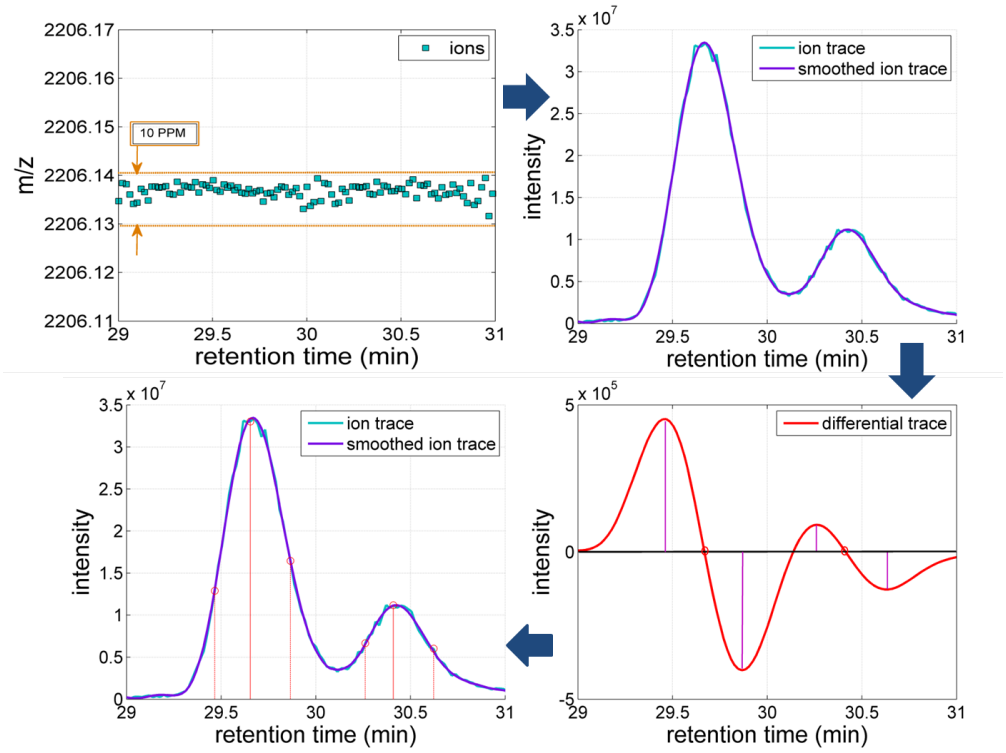


Figure 3.2: Peak detection diagram. Left top: tracing ions across scans; right top: smoothing trace; right bottom: taking first derivative; left bottom: latching peak locations.

3.2.3 Peak detection

To increase signal-to-noise ratio, a high order Savitzky-Golay filter is designed to smooth the traces. Detecting peaks in smoothed feature is more reliable. By convoluting smoothed trace with first derivative Guassian kernel, we can latch the position of a peak and its right and left boundary. The intensity of the peak is the sum of the ion counts between the boundaries. Because isomers exist, one trace may generate several peaks. Figure 3.2 illustrates through an example trace how the position of a peak and its boundaries are determined.

3.3 Feature clustering and annotation

The peak list consists of redundant information, because the same analyte is possibly ionized into different ion species due to various forms of adduct, which results in multiple peaks for LC-MS data. Without further preprocessing, the complexity of both downstream statistical analysis and subsequent compound identification is unduly increased, especially for untargeted experiments. We cluster features generated from the same analyte to reduce redundancy and improve quantification accuracy.

3.3.1 Clustering charge states

We cluster together ions that have the same mass value and retention time, but different charge states. GPA first splits the ungrouped peak list into several blocks by collecting the peaks with close mass value. Within each mass block, peaks with charge difference and close retention time are designated as a candidate cluster. From each cluster, we select the peak with the highest intensity as a representative and replace its intensity with the aggregate one from all peaks in this cluster. Since the same mass and RT information are shared by all peaks within a cluster, the only difference among the peaks is charge states.

3.3.2 Clustering adduct states

An analyte may attach different adducts during ionization. The adduct states include ammonium, sodium and potassium, resulting in multiple ions with various m/z sites eluted simultaneously. Hence the recovered mass values will be different. Software tools are available to cluster adducts including GlycReSoft and MultiGlycan. However, GlycReSoft only considers ammonium cases. MultiGlycan clusters peaks by matching retention time to user-defined molecular weight difference of possible adducts. This might lead to false positive cluster linking different analytes together, because two compounds may also closely coelute and their mass differences happen to match those among adduct pairs. A more reliable piece

of information is the profiling shape of peak. Similarity between a pair of peaks is employed to group or separate them.

Starting from the most intense feature, GPA first creates a time window/block with this feature's retention time as its center, assigning features closely coeluting with this feature to the same block. The peak list is, again, split into several blocks. Features within a block have close retention time (tolerance is at ±2s). A reasonable assumption is, among those features, those whose profiling shapes are similar with each other should come from the same analyte with higher level of confidence. Their extracted ion chromatograms (EICs), which contain the shape information, are truncated with same boundaries (length at n) for the sake of comparability. GPA calculates a pointwise pearson correlation coefficient (PCC) of the intensities between the chromatographic peak boundaries for all pairs of features in one block. PCC helps evaluate the similarity between each pair of features,

$$\mathrm{PCC}(\vec{X},\vec{Y}) = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}. \tag{3.1}$$

In a graph created for each block, the features, that may represent more than one underlying compounds, are set as vertices. The PCC values between each pair of vertices are set as the weights of edges connecting them. A predetermined threshold of PCC is empirically used to exclude weak connections in the graph, which could result in an unconnected graph. The *Floyd-Warshall algorithm* [49] is utilized to check the connectivity of a graph. If there exist two or more parts that are not connected within the graph, they are considered less likely to originate from the same glycan.

The proposed method for clustering adducts is modeled into a graph clustering problem, aiming to dig out the underlying subgraphs/clusters. In the field of graph theory, quite a few clustering algorithms are proposed to address this problem. From the perspective of clustering scale, the approaches can be categorized into *global clustering* and *local clustering*. It also depends on the application and the input graph whether it makes sense to utilize a *hierarchical clustering* or a *flat clustering* [50]. Even within hierarchy clustering, the methods could be quite different due to various aspects, e.g. *maximum flow* and *spectral method* . Considering implementation cost and the input weighted undirected graph, GPA employs *Highly-Connected-Subgraphs (HCS)* algorithm [51] based on the *minimum cut* method, that works top-down, recursively dividing the graph into clusters.

A good cluster should be the subgraph with high density. Typically, the density of graph G, with m edges (size = m) and n vertices (order = n), can be defined as the ratio of the number of edges present to the maximum possible,

$$\delta(G) = \frac{m}{\binom{n}{2}}. \tag{3.2}$$

HCS uses a density-based stopping condition. Density of G here is determined by graph's edge-connectivity K(G) , which indicates the minimum number of edges that would need to be removed from G in order to disconnect the graph. Generally, a cluster is achieved if the vertices involved are highly connected with each other, whereas the paths connecting them to vertices outside the cluster should have much smaller weights. In HCS, the highly connected graph refers to the one whose edge-connectivity is above half of the order n,

$$\mathrm{K}(G) \geq \frac{n}{2}. \tag{3.3}$$

We use cut to break down connections so as to obtain subgraphs or clusters. A cut of graph is a collection of edges whose removal disconnects the graph. The *minimum cut* is a cut with minimum number of edges (i.e. the size of minimum cut equals K(G)). In practice, the minimum cut in a given weighted graph can be found efficiently with a maximum-flow algorithm. HCS and minimum cut algorithm are described as follow.

Algorithm 1 Algorithm of Highly Connected Subgraphs

$C = \texttt{HCS}(G)\{$
%Input: A weighted connected graph G;
%Output: A collection of clusters C;

$\{(G_1, G_2), \mathrm{cut}\} = \texttt{MINCUT}(G)$

if $\mathrm{K}(G) \geq \frac{n}{2}$ **then**
  return G;
else
  $C_1 = \texttt{HCS}(G_1)$;
  $C_2 = \texttt{HCS}(G_2)$;
  return $[C_1, C_2]$;
end if
}

HCS groups features of each block into several clusters. At this stage, we are confident that features in each cluster may originate from the same compound.

Algorithm 2 Algorithm of Minimum Cut

$\{(G_1, G_2), \text{cut}\} = \texttt{MINCUT}(G)\{$
%Input: A weighted connected graph G,
% the weight of edge $\{v_i, v_j\}$ is $w_{i,j}$.
%Output: Minimum cut and separated parts G_1, G_2.

Arbitrarily select a vertex of G as v_1
$n = |V(G)|$
$S = \{v_1\}$

for $i = 2 : n$ **do**
  let v_i the vertex of $V \setminus S$
  if $v_i = \texttt{MAX}(\sum_{i \in S} w_{i,r})$ over all $v_r \in V \setminus S$ **then**
    $S = S \cup \{v_i\}$
  end if
end for
if $n = 2$ **then**
  return the cut $\delta(\{v_n\})$;
else
  Merge vertices in S to obtain G'
  cut = $\texttt{MINCUT}(G')$
  return $\texttt{MIN}(\text{cut}, \delta(\{v_n\}))$
end if
}

3.3.3 Annotation

An annotation carried out on each cluster will reduce the amount of features for subsequent analysis. The ion's mass value recovered by DeconTools is neutral molecular weight. DeconTools assumes all the analytes are ionized through protons rather than other cations. Hence the observed mass value (M_o) should be different from the real one (M) if its adduct formation varies. For example, a compound with neutral molecular weight at M is ionized into 2 features: $\text{Re}[M + z \cdot \text{H}]^{z+}$ and $\text{Re}[M + a \cdot \text{NH}_4 + b \cdot \text{H}]^{z+}, z = a + b$. DeconTools will recover M correctly for the first form, and report a shifted mass value M_o for the second

form:

$$
\begin{aligned}
(m/z)_1 &= \frac{M + z \times 1.0078}{z}, \\
&\Downarrow \\
M_o &= (m/z)_1 \times z - 1.0078 \times z = M; \\
\\
(m/z)_2 &= \frac{M + a \times 18.0344 + b \times 1.0078}{a + b} \\
&= \frac{M + a \times 17.0266 + (a + b) \times 1.0078}{a + b}, \\
&\Downarrow \\
M_o &= (m/z)_2 \times z - 1.0078 \times z = M + a \times 17.0266.
\end{aligned}
\tag{3.4}
$$

Based on these calculations, we learn that fixed mass difference (e.g. $a \times 17.0266$) exists between the observed mass pair of one specific adduct formation (e.g. $\mathrm{Re}[M + a \cdot \mathrm{NH_4} + b \cdot \mathrm{H}]^{z+}, z = a + b$) and simple proton formation. GPA then matches the pairwise mass differences within a cluster against the generated rule Table 3.1 to achieve a. The proton amount b can be deduced from charge state z. Hence given mass difference and charge state, the adduct form is confirmed.

Table 3.1: Δmass table for various adduct formations

Adduct\a	1	2	3	4
Ammonium	17.0266	34.0532	51.0798	68.1064
Sodium	21.9819	43.9638	65.9457	87.9276
Potasium	37.9559	75.9118	113.8677	151.8236

3.4 Experimental results

Experimental setup

GPA workflow is applied to detect and annotate permethylated N-linked glycans released from human serum samples. Figure 3.3 shows the sample preparation steps prior to analysis of serum by LC-MS.

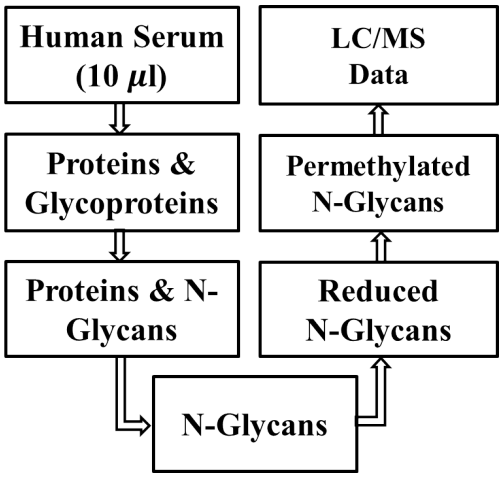


Figure 3.3: LC-MS data acquisition. Proteins and glycoproteins are extracted from human serum. N-glycans are then released from glycoproteins. Permethylation faciliates the detection of glycans in LC-MS

We employed the LTQ Orbitrap Velos MS (Thermo Scientific, Pittsburgh, PA, USA) for analysis of permethylated n-glycans. The mass spectrometry was operated at positive mode with m/z range of 600-2,000. FT mass analyzer was set at 15,000 resolution, 5 MS/MS scans were conducted after each MS scan in data dependant acquisition mode. A Dionex Ultimate 3000 UHPLC (Thermo Scientific, Pittsburgh, PA, USA) system was utilized for the LC separation, with the Acclaim C18 nano column (Thermo Scientific, Pittsburgh, PA, USA). The flow rate was set to 350 nL/min. The Ultimate 3000 LC system has a column oven to control column temperature. We use 55°C to get a better elution chromatography. The averagine formula is set to $C_{10}H_{18}N_{0.43}O_5S_0P_0$ in the deconvolution step where the open-source software DeconTools is employed.

3.4.1 Evaluation on LC-MS datasets

We tested GPA through LC-MS data acquired by analysis of N-glycans derived from a human serum sample. First, we evaluated the clustering and annotation components of GPA. We compared the performance of our proposed method with manual annotation and existing tool. Then, we demonstrated GPA's superiority in a case where other tools are helpless.

This evaluation focuses on the clustering and annotation part of GPA. An ungrouped peak list generated by the first part of GPA is reorganized into multiple blocks by grouping peaks with close retention time together. For a better illustration, we test the subsequent performance by looking into one block. We create the ground truth by manually separating and annotating the peaks in this block. In ungrouped peak list, the block consists of 24 peaks, 7 pairs of which are multiply charged. We list the peaks within the block in Table 3.2 after

Table 3.2: Block information

ID	Mass	Charge	RT(s)	Intensity
1	2237.167	2,3	1758.73	1.81E+09
2	2771.411	3,2	1761.01	1.37E+09
3	2254.194	2,3	1758.73	1.16E+09
4	2019.053	2	1758.73	1.12E+09
5	2788.440	3,2	1761.01	7.73E+08
6	2271.220	2,3	1758.73	5.31E+08
7	2149.113	2,3	1755.16	4.34E+08
8	2805.465	3,2	1762.13	3.78E+08
9	2166.142	2	1755.16	2.61E+08
10	1870.975	2	1759.82	1.97E+08
11	2276.171	2	1757.53	1.69E+08
12	2793.393	3	1761.01	1.42E+08
13	2183.167	2	1755.16	1.37E+08
14	2259.152	2	1757.53	1.34E+08
15	2809.397	3	1761.01	1.27E+08
16	2822.491	3	1761.01	7.18E+07
17	2188.120	2	1756.32	5.96E+07

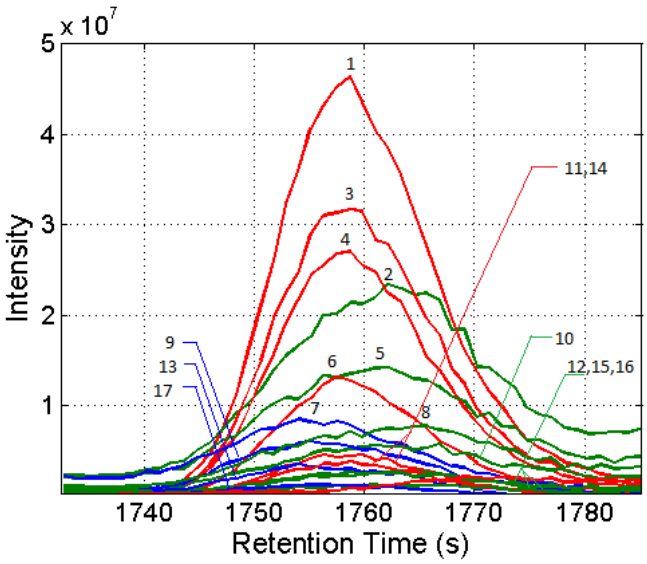


Figure 3.4: EICs of 17 closely coeluted peaks. Extracted ion chromatograms provide shape information for GPA to separate different compounds eluted together.

clustering different charge states. Adducts clustering and annotation are then performed on such a block.

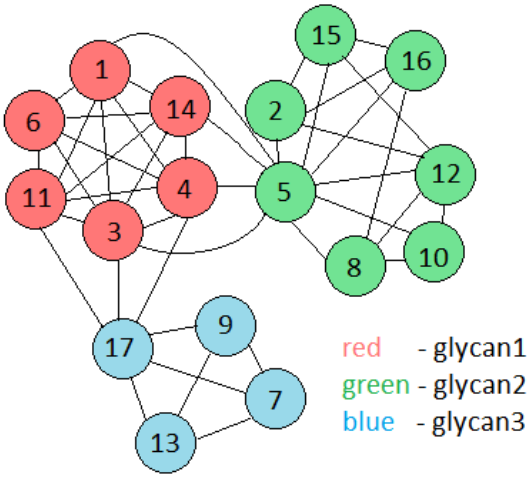


Figure 3.5: Constructed graph. GPA uses PCC values as edge weights. Vertices are named by peaks' IDs and edges indicate above-threshold PCC values. For this graph, HCS returns three clusters marked in red, green and blue.

Figure 3.4 depicts the EICs of these peaks extracted from the raw file. GPA employs pairwise PPC to construct a weighted graph (Figure 3.5,). Highly connected subgraphs are returned as clusters. For this example, GPA achieves three clusters marked in *red*, *green* and *blue* respectively. Three different underlying glycans are deduced by GPA.

Table 3.3 lists the result of annotation done within each cluster. Among 24 peaks, 22 have been assigned to one of the three glycans and annotated by different charge states and adduct forms. For each cluster, the peak adducted by proton is selected as the representative to match against user-defined putative structure list. For example, peak with mass of 2237.167 (ID = 1) represents the first glycan (G1) and is matched to composition $HexNac_4Hex_5deHex_1$.

Table 3.3: Annotation result

Glycan	ID	Mass	Charge	Annotation
G1	1	2237.167	2	$Re[M_1 + 2H]^{2+}$
G1			3	$Re[M_1 + 3H]^{3+}$
G1	3	2254.194	2	$Re[M_1 + H + NH_4]^{2+}$
G1			3	$Re[M_1 + 2H + NH_4]^{3+}$
G1	4	2019.053	2	unknown
G1	6	2271.220	2	$Re[M_1 + 2NH_4]^{2+}$
G1			3	$Re[M_1 + H + 2NH_4]^{3+}$
G1	11	2276.171	2	$Re[M_1 + H + K]^{2+}$
G1	14	2259.152	2	$Re[M_1 + H + Na]^{2+}$
G2	2	2771.411	3	$Re[M_2 + 3H]^{3+}$
G2			2	$Re[M_2 + 2H]^{2+}$
G2	5	2788.440	3	$Re[M_2 + 2H + NH_4]^{3+}$
G2			2	$Re[M_2 + H + NH_4]^{2+}$
G2	8	2805.465	3	$Re[M_2 + H + 2NH_4]^{3+}$
G2			2	$Re[M_2 + 2NH_4]^{2+}$
G2	10	1870.975	2	unknown
G2	12	2793.393	3	$Re[M_2 + 2H + Na]^{3+}$
G2	15	2809.397	3	$Re[M_2 + 2H + K]^{3+}$
G2	16	2822.491	3	$Re[M_2 + 3NH_4]^{3+}$
G3	7	2149.113	2	$Re[M_3 + 2H]^{2+}$
G3			3	$Re[M_3 + 3H]^{3+}$
G3	9	2166.142	2	$Re[M_3 + H + NH_4]^{2+}$
G3	13	2183.167	2	$Re[M_3 + 2NH_4]^{2+}$
G3	17	2188.120	2	$Re[M_3 + H + K]^{2+}$

For the complete LC-MS dataset, GPA initially generates 2566 ungrouped peaks. After grouping them using the charge state information, the number of peaks was reduced to 2376. These peaks were grouped into 307 clusters for the following annotation using the adduct information. Compared with GlycReSoft, GPA is able to consider more adducts. In addition, taking sodium and potassium into consideration could help assign glycans to correct structures. For example, in our experiment, GlycReSoft assigns the peak with mass at 2237.167 to two different structures $HexNac_4Hex_5deHex_1$ ($Re[M + 2H]^{2+}$) and $HexNac_4Hex_4NeuAc_1$ ($Re[M + H + NH_4]^{2+}$). However in GPA, the absence of a peak at 2220.140 ($HexNac_4Hex_4NeuAc_1$) and the presence of a peak at 2259.152 ($Re[M + H + Na]^{2+}$) result in a more confident assignment to structure $HexNac_4Hex_5deHex_1$.

3.4.2 Simulation on ambiguous case

GlycReSoft and MultiGlycan cluster peaks only using mass information. In some ambiguous cases, where mass information is not enough to distinguish two glycans because of their

close mass values, these tools will lead to false assignments. As comparison, GPA utilizes additional EIC shapes and uses graph clustering method to separate ambiguous compounds.

Table 3.4: Peaks to be clustered

ID	Mass	Charge	RT(s)	Intensity
A1	1787.942	2	1470.90	1.47E+08
A2	1804.964	2	1470.86	8.29E+07
B1	1770.926	2	1471.98	1.07E+08
B2	1787.952	2	1470.86	4.46E+07
B3	1792.913	2	1470.86	1.44E+07

It is very tedious to find sets of peaks from different compounds which coelute, due to the large size of our peaklist. Hence, we simulate the ambiguous case by artificially forcing peaks from two different compounds to coelute together. GPA was asked to recognize these peaks. While GlycReSoft and MultiGlycan were unable to distinguish two peaks (A1 and B2), because their mass values are too close.

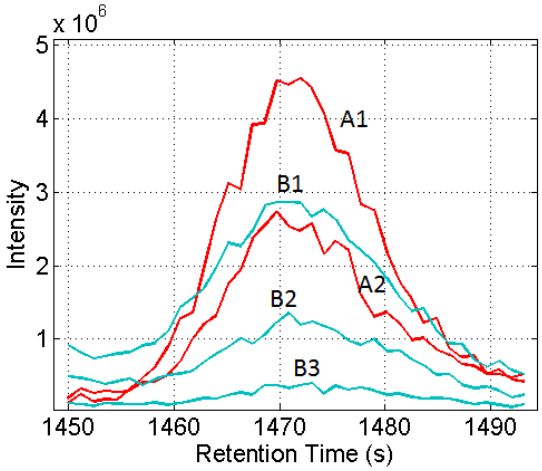


Figure 3.6: EICs of five peaks that coeluted closely

Table 3.5: Pairwise Pearson correlation coefficient

PCC	A1	A2	B1	B2	B3
A2	0.9873	-	-	-	-
B1	0.8576	0.8134	-	-	-
B2	0.8043	0.7911	0.9729	-	-
B3	0.7425	0.7604	0.9230	0.9386	-

Table 3.4 presents the five peaks derived from the two compounds. The EICs of these peaks are shown in Figure 3.6. Two of them belong to glycan A and the rest belong to glycan

Table 3.6: Annotation result

glycan	ID	mass	charge	Annotation
A	A1	1787.942	2	$Re[M_A + 2H]^{2+}$
A	A2	1804.964	2	$Re[M_A + H + NH_4]^{2+}$
B	B1	1770.926	2	$Re[M_B + 2H]^{2+}$
B	B2	1787.952	2	$Re[M_B + H + NH_4]^{2+}$
B	B3	1792.913	2	$Re[M_B + H + Na]^{2+}$

B. A1 and B2 have very close mass value. GlycReSoft and MultiGlycan reported these as one compound. GPA calculate pairwise PCC (Table 3.5). As shown in the table, peaks that originate from the same compound have stronger connections. Annotation result in Table 3.6 shows that GPA correctly recovers glycan A ($HexNac_3Hex_4deHex_1$) and glycan B ($HexNac_3Hex_3NeuAc_1$).

Chapter 4

Probabilistic purification models

In Chapter 3, we have elaborated the details of LC/GC-MS data preprocessing steps, which ascertain the rationality of compound identification and quantitation. This helps in monitoring disease-related alterations in molecular and cellular mechanisms that may reveal useful disease biomarkers. Discovery of clinically relevant biomarkers has potentially far reaching implications for disease management and patient treatment [8–10,52]. Similarly as glycomics, high-throughput omic technologies have facilitated the search for changes in multiple levels of various biomolecules (e.g., proteins, glycoproteins, metabolites, lipids, etc.) in biological samples [19, 53]. As aforementioned, biomolecules are separated, fragmented, ionized and detected in LC/GC-MS instruments. Abundances of ions with various retention time and mass values are recorded for downstream data processing. Valid signals are retrieved during the preprocessing. However, these signals may still lead to biased or even incorrect discovery of biomarker candidate if the injected samples are not homogeneous as we assumed. In this chapter, an in-depth discussion of this issue is given.

4.1 Sample heterogeneity in biomarker discovery

While the capability of high-throughput technology to yield comprehensive profiling and quantification offers a unique advantage in biomedical research, the heterogeneous nature of the biological samples poses a fundamental challenge in data analysis and interpretation. Specimens, such as tumor tissues and human blood, are typically mixtures of cells with distinct states and types, and usually only part of the constituent cell populations is relevant to the biological question of interest [11, 12]. In some cancer studies, heterogeneity is also observed within the malignant cell population, where multiple cancerous subtypes co-exist [13]. Ideally in a biomarker discovery study, one would perform between-group (cancer versus related disease, cancer versus healthy samples) differential expression analysis for type-specific constituents in samples [14]. However, biospecimens collected from patients

usually exhibit some degree of heterogeneity. Moreover, the proportion of cancerous, other disease-related, and healthy components varies across individual samples pre-selected using pathological estimates. Therefore, the biomolecular measurements in expression profiles are attributed to multiple sites of origins with various mixture proportions. The cancerous profiles of interest are typically contaminated by other components, leading to unreliable results in differential analyses. Purification of samples is hence highly desired to remove the effects of heterogeneity.

Experimental methods for cleaning samples and isolating type-specific constituents are costly and time-consuming. Computational purification methods offer an attractive alternative that is inexpensive and efficient to implement, and can be applied to data already generated without any modifications on experimental procedures. Multiple approaches have been developed to deconvolute gene expression profiles in the past years, including linear regression based models and generative probabilistic models [15–18]. Among these approaches, topic model based methods, e.g., ISOpure [12] and ISOLATE [54], showed promising performance in estimating the proportion of mixtures and inferring sample-specific purified profiles in heterogeneous genomic data. However, to the best of our knowledge, in omic studies involving quantitative analysis of proteins or metabolites, no such purification approaches have been applied to deal with the sample heterogeneity issue. With the increasing volume of these data generated by LC/GC-MS, it is necessary to implement the purification of data before downstream differential analyses. In this research, we first apply a topic model based purification approach to both synthetic and experimental data acquired from human sera and liver tissues by LC-MS and GC-MS, respectively. The purpose of this investigation is to test if sample heterogeneity issue in various biomolecular expression profiles can be addressed by adjusting ion intensities through topic models as in genomic studies. Also, we investigate the use of scan-level features, i.e. extracted ion chromatograms (EICs) instead of integrated peak intensities, to alleviate the information loss during the LC/GC-MS data preprocessing.

In the following sections, we introduce several topic model-based purification methods, including intensity-level and scan-level purification, and denoise deconvolution models, in chronological order of development. Assumptions and strategies within each topic model are elaborated respectively.

4.2 Intensity-level purification model

4.2.1 Mathematical modeling of ion counts

The LC/GC-MS instruments provide ion intensity values by counting the ions detected at specific m/z and retention time points. Due to the existence of heterogeneity, multiple constituents in the sample contribute to the observed expression profile. Therefore, we can model the expression profile of a heterogeneous sample $\boldsymbol{t}$ as a weighted mixture of expression

profiles of multiple sources, including a cancerous origin $\boldsymbol{\gamma}$ and non-cancerous contaminants $\boldsymbol{\beta}$. The expression distribution for every biomolecule in each of the sources plays a role as a "topic" contributing to the mixed expression profile. Basically, each ion in the observed profile is associated with a specific topic, i.e. a multinomial distribution of ion counts over biomolecules, determined by the corresponding source profile. In this model, expression profiles refer to integrated peak intensities.

4.2.2 Derivation of LDA and basic assumptions

The purification procedure can be realized through a set of topic models, which are generative probabilistic models typically applied to text corpora mining. Specifically, each expression profile is characterized by a probability distribution across topics. Topics are probability distributions across biomolecules. These distributions can be inferred based on the analysis of a collection of expression profiles through topic models. These hierarchical Bayesian models are variants of latent Dirichlet allocation (LDA) [40], a topic model that can 1) infer the posterior probability of topics given observed profiles, and 2) estimate the parameters that generate the latent mixture proportion and topic panel. These topic models have been adapted and applied to gene expression profiles in genomic studies [12, 54].

We use a modified topic model to purify the molecular expression profiles in cancer. Basically, three assumptions are made in developing the *intensity-level purification model* (IPM). First, the source contaminants in each expression profile $\{\boldsymbol{t}_d\}_{d=1,\cdots,D}$ are coming from the control group $\{\boldsymbol{\beta_m}\}_{m=1,\cdots,M}$ (i.e., healthy, non-cancerous profiles, etc.). It is commonly observed that the cancerous tissues are surrounded by adjacent non-cancerous tissues, which are typically used as controls in differential expression analysis. Second, the corresponding cancerous origins $\{\boldsymbol{\gamma}_d\}_{d=1,\cdots,D}$ share an average cancer profile $\boldsymbol{\gamma'}$. Individual cancerous profile can be treated as a noisy version of the average cancer profile. Third, the average cancer profile $\boldsymbol{\gamma}'$ has similar patterns as non-cancerous profiles $\{\boldsymbol{\beta}\}$, except for some sites (biomolecules) which are differentially expressed between case and control groups. Thus, the cancerous profile can be treated as a similar non-cancerous profile with several sites altered.

4.2.3 Probabilistic generative representation of IPM

The complete likelihood function in (1) describes how the profiles $\{\boldsymbol{t}_d\}_{(d=1,,D)}$ are generated. Specifically, we have two observable variables indicating D expression profiles in case group: $\{\boldsymbol{t}_d\}_{d=1,\cdots,D},\ \boldsymbol{t}_d \in \mathbb{R}^N$, and M non-cancerous profiles in control group: $\{\boldsymbol{\beta_m}\}_{m=1,\cdots,M},\ \boldsymbol{\beta_m} \in \mathbb{R}^L$. In our analysis, we normalize all profiles to have identical total ion counts of N and consider L biomolecules that are consistently detected in all the samples. For convenience, we represent the normalized profiles in two ways. Each heterogeneous cancer profile $\boldsymbol{t}_d$ is represented via N ions, with $t_{d,n} = \{1, 2, \cdots, L\}$ denoting the biomolecule corresponding to the n^{th} ion. Each non-cancerous profile $\boldsymbol{\beta}_m$ is represented via L biomolecules, with $\beta_{m,l}$

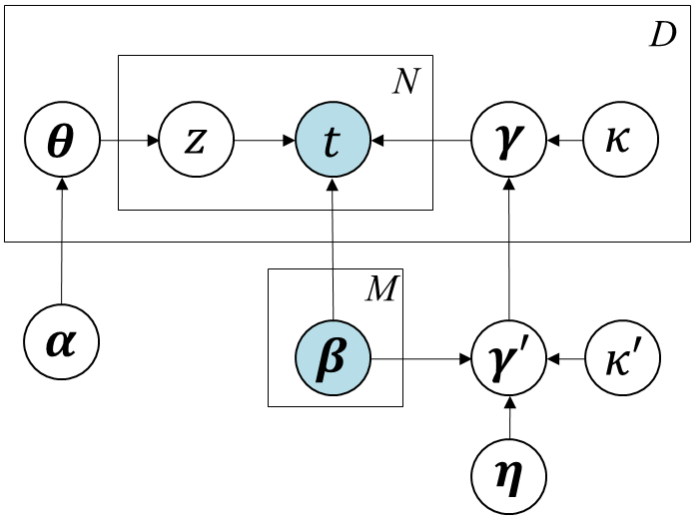


Figure 4.1: Graphical representation of the generative probabilistic model. Hyperparameters $\boldsymbol{\eta}$, κ' together with sources of contaminants $\{\boldsymbol{\beta}_m\}$ determine an average cancer profile $\boldsymbol{\gamma'}$. Each of the D profiles is associated with a mixture proportion $\boldsymbol{\theta}_d$ (regularized by hyperparameter α) and a topic panel consisting of $\{\boldsymbol{\beta_m}\}$ and γ' (generated from the average cancer profile). Each of the N ions in a profile $t_{n,d}$ is sampled from a topic indicated by $z_{n,d}$.

denoting the ion counts of the l^{th} biomolecule, and $\sum_{l=1}^{L} \beta_{m,l} = N$. The second expression can be further normalized by N to give a representation of multinomial distribution as a topic.

$$\begin{aligned}\mathcal{L}(\boldsymbol{t}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\gamma'} | \alpha, \boldsymbol{\beta}, \boldsymbol{\eta}, \kappa, \kappa') \\ = p(\boldsymbol{\gamma'}|\boldsymbol{\beta}, \boldsymbol{\eta}, \kappa') \cdot \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha) \cdot p(\boldsymbol{\gamma}_d|\boldsymbol{\gamma'}, \kappa_d) \\ \cdot \prod_{n=1}^{N} [p(z_{d,n}|\boldsymbol{\theta}_d) \cdot p(t_{d,n}|z_{d,n}, \boldsymbol{\theta}_d, \boldsymbol{\beta}, \boldsymbol{\gamma}_d)]\end{aligned} \tag{4.1}$$

The model also incorporates the following latent variables: the average cancer profile $\boldsymbol{\gamma'} \in \mathbb{R}^L$, sample-specific pure cancer profiles $\{\boldsymbol{\gamma}_d\}_{d=1,\cdots,D}$, $\boldsymbol{\gamma}_d \in \mathbb{R}^L$, sample-specific mixture proportions of topics $\{\boldsymbol{\theta}_d\}_{d=1,\cdots,D}$, $\boldsymbol{\theta}_d \in \mathbb{R}^{M+1}$, and sample-specific topic indicators $\{\boldsymbol{z}_d\}_{d=1,\cdots,D}$, $\boldsymbol{z}_d \in \mathbb{R}^N$, $z_{d,n} = \{1, \cdots, M, M+1\}$. Their relationships with observations and parameters are given as below.

$$p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\theta}_d|\boldsymbol{\alpha}, 1) \tag{4.2}$$

$$p(\boldsymbol{\gamma'}|\boldsymbol{\beta}, \eta, \kappa') = \text{Dirichlet}(\boldsymbol{\gamma'}|\boldsymbol{\eta}^T\boldsymbol{\beta}, \kappa') \tag{4.3}$$

$$p(\boldsymbol{\gamma}_d|\boldsymbol{\gamma'}, \kappa_d) = \text{Dirichlet}(\boldsymbol{\gamma}_d|\boldsymbol{\gamma'}, \kappa_d) \tag{4.4}$$

$$p(z_{d,n}|\boldsymbol{\theta}_d) = \text{Multinomial}(z_{d,n}|\boldsymbol{\theta}_d) \tag{4.5}$$

$$p(t_{d,n}|z_{d,n} \leq M, \boldsymbol{\theta}_d, \boldsymbol{\beta}, \boldsymbol{\gamma}_d) = \text{Multinomial}(t_{d,n}|\boldsymbol{\beta}_{z_{d,n}}) \tag{4.6}$$

$$p(t_{d,n}|z_{d,n} = M+1, \boldsymbol{\theta}_d, \boldsymbol{\beta}, \boldsymbol{\gamma}_d) = \text{Multinomial}(t_{d,n}|\boldsymbol{\gamma}_d) \tag{4.7}$$

The average cancer profile $\boldsymbol{\gamma}'$ is sampled from a Dirichlet distribution parameterized by a weighted mixture of non-cancerous profiles. Each pure cancer profile $\boldsymbol{\gamma}_d$ together with M contaminants $\{\boldsymbol{\beta}_m\}$ forms a sample-specific topic panel. The mixture proportion of topics determines $z_{d,n}$, indicating which source (i.e., γ_d or $\{\boldsymbol{\beta}_m\}$) each ion originates from. We infer the latent variables $\boldsymbol{\gamma}'$, $\{\boldsymbol{\gamma}_d\}_{d=1,\cdots,D}$, $\{\boldsymbol{\theta}_d\}_{d=1,\cdots,D}$, and estimate the parameters using the two-step learning approach developed based on variational EM algorithm (Appendix A). The graphical model representing the above topic model is shown in Figure 4.1. This three-level model allows a single profile to be associated with multiple topics (i.e., cancerous and non-cancerous origins). Such property of the LDA-framed models enable more flexible representation in data structure than that by other unigram models or mixture of unigrams [40]. Also in contrast to linear regression models, these methods use a multinomial noise model that is a better fit to noise measurement in biomolecular expression data. [17]

4.3 Scan-level purification model

4.3.1 Utilization of scan-level features

Here, we further extend the framework to scan-level purification model (SPM) by utilizing the scan-level measurements instead of the integrated peak intensities in IPM. During LC/GC-MS data preprocessing, ion intensity is obtained by integrating the scan-level measurements of a detected chromatographic peak within a specified retention time (RT) interval. This integration or truncation, however, inevitably brings in variances which interfere with original sample heterogeneity. Therefore, we propose to investigate LC/GC-MS data purification with scan-level measurements based on extracted ion chromatogram (EIC), which preserves scan-level peak shape information. We hypothesize that purification at the scan level leads to more accurate results and offers the opportunity to extend the model to characterize both ion abundance and peak shape.

After ion tracing and missing value interpolation, we can obtain a list of EICs for each sample. EIC is characterized by its retention time (corresponding to multiple scans), mass value, and ion abundance. In this scenario, the observed data $\{\boldsymbol{t}_d\}$ (same for $\{\boldsymbol{\beta}_m\}$) consists of multiple EIC peaks. Each peak is represented by ion abundances across S scans with a certain elution profile shape $\mathcal{F}(\cdot)$ as shown in Figure 4.2. Using these scan-level features, we model each EIC peak as shown in Eq. (4.8):

$$t_{d,n}(s) = x_{d,n} \cdot \delta_{d,n}(s) \cdot \mathcal{F}(s, \boldsymbol{\phi}_{d,n}) + e_{d,n}(s), \;\; s = 1, \cdots, S \tag{4.8}$$

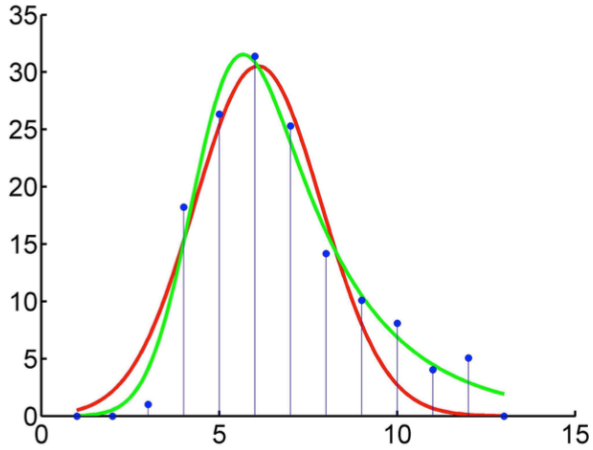


Figure 4.2: Extracted ion chromatography and peak shape function. Example of Gaussian (red) and exponentially modified Gaussian (green) peak shapes fitted to an experimental EIC involving 13 scans (blue).

where, $x_{d,n}$ is the ion abundance for n^{th} compound of d^{th} sample; $\delta_{d,n}(s)$ is a latent indicator to model the missing scans; the chromatographic peak shape is characterized by the exponentially modified Gaussian (EMG) function [55] parameterized by $\boldsymbol{\phi}$, as described in Eq. (4.9), and $e_{d,n}(s)$ is the random noise.

$$\mathcal{F}(s,\boldsymbol{\phi}) = \frac{1}{2}\zeta \exp\left(\frac{1}{2}\zeta(2\mu+\zeta\sigma^2-2s)\right)\cdot(1-\text{erf}(\frac{\mu+\zeta\sigma^2-s}{\sqrt{2}\sigma}),\ \boldsymbol{\phi}\doteq\{\mu,\zeta,\sigma\} \tag{4.9}$$

We hypothesize that the data heterogeneity in $t_{d,n}$ corresponds to the shape of the EIC (characterized by $\boldsymbol{\phi}$) as well as ion abundance $x_{d,n}$.

4.3.2 Probabilistic generative representation of SPM

We extend the purification model we used for integrated peaks by adding a lower layer to characterize the scan-level information as illustrated in Figure 4.3. The three assumptions are maintained in this model in terms of the dependancy of ion abundance variables. That is, Eq. (4.2-4.7) still hold for ion abundances $\boldsymbol{x}_t$, $\boldsymbol{x}_\beta$, $\boldsymbol{x}'_\gamma$, and $\boldsymbol{x}_\gamma$. We assume error terms in intensity measurements in Eq. (4.8) are independent random variables generated by a normal distribution with conjugate prior following an inverse-Gamma distribution:

$$e_{d,n}(s)|\sigma^2_{e_d} \sim \mathcal{N}(0,\sigma^2_{e_d}), \qquad \sigma^2_{e_d} \sim \mathcal{IG}(a_e,b_e). \tag{4.10}$$

The missing scan indicator variable $\delta_{d,n}(s)$ follows a Bernoulli distribution, parameterized by q_d with a prior of Beta distribution:

$$p(\delta_{d,n}(s)|q_d) = \text{Bernoulli}(\delta_{d,n}(s)|q_d), \qquad p(q_d|a_q,b_q) = \text{Beta}(q_d|a_q,b_q). \tag{4.11}$$

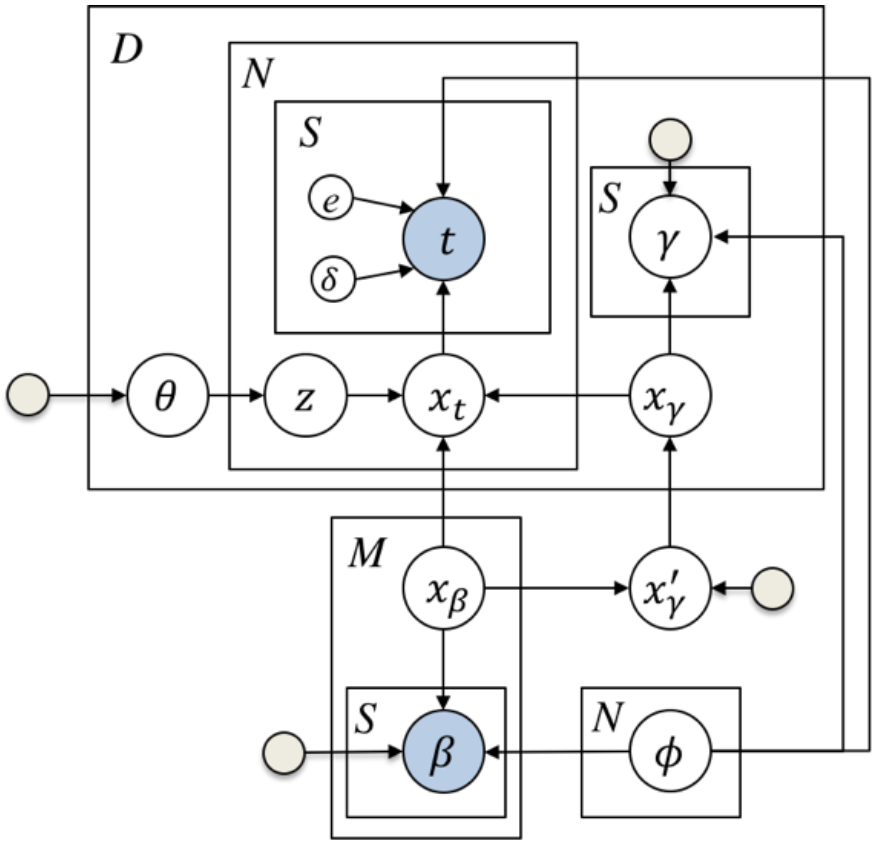


Figure 4.3: Graphical representation of the scan-level topic model. A lower layer to characterize the scan-level information is added. Ion abundances $\boldsymbol{x}_t$, $\boldsymbol{x}_\beta$, $\boldsymbol{x}'_\gamma$, and $\boldsymbol{x}_\gamma$ together with peak shape (parameterized in $\boldsymbol{\phi}$) determined the observed feature list $\boldsymbol{t}$, $\boldsymbol{\beta}$.

The observed data point therefore follows the distribution:

$$t_{d,n}(s)|x_{td,n}, q_d, \phi_{d,n}, \sigma^2_{e_d} \sim q_d \mathcal{N}(x_{td,n}\mathcal{F}(s, \phi_{d,n}), \sigma^2_{e_d}) + (1 - q_d)\mathcal{N}(0, \sigma^2_{e_d}). \tag{4.12}$$

The peak shape parameters $\boldsymbol{\phi}$ are considered to have a normal distribution and its detailed priors are described in [55]. The extended model contains variables that are mutually coupled, providing no analytical form for the posterior distribution in calculation. As a variational approximation, we can split the model into two components: 1) mixture model of underlying ion abundances, and 2) scan-level feature fitting. We adopt a two-phase approach to iteratively update the latent variables and estimate the parameters between the two parts. Specifically, we use a Markov chain Monte Carlo (MCMC) sampling method [55] (Appendix B) to infer the peak shape model parameters of the second part (i.e.,ion abundance $\boldsymbol{x}_t$, $\boldsymbol{x}_\beta$, and shape function parameters ϕ). We then treat $\boldsymbol{x}_t$, $\boldsymbol{x}_\beta$ as observed variables to implement the inference on the first part using the same algorithm [12] employed in the intensity-level purification. Once converged, the model outputs the sample-specific mixture

proportion $\boldsymbol{\theta}$, pure ion abundance $\boldsymbol{x}_\gamma$, shape function parameters $\boldsymbol{\phi}$ and related parameters. After purification is performed, ion intensity may be calculated by applying peak detection algorithms [56, 57] to the pure EIC peaks $\{\boldsymbol{\gamma}_{d,n}\}$ recovered using Eq. (4.8).

4.4 Denoise deconvolution model

4.4.1 Different derivation direction

The purification model is semi-supervised and specifically designed for studies of comparative analysis, with clearly labelled groups. Upon uncovering pure source profiles, we are interested in deconvoluting the observed heterogeneous measurements, by removing the noise contaminants, and retrieving the proportions of source components. By adjusting assumptions in a different direction, we modify the purification model into a denoise deconvolution model (DDM) which helps identify subcomponents among samples with respect to a specific disease in the same cohort. Figure 4.4 shows the graphical structure of the adjusted model, where we treat the S purified source profiles as the known topics $\boldsymbol{\beta}'$ and an additional noise prior $\boldsymbol{\delta}$ in generating sample-specific noise topics $\boldsymbol{\epsilon}$ in each of the measurements t.

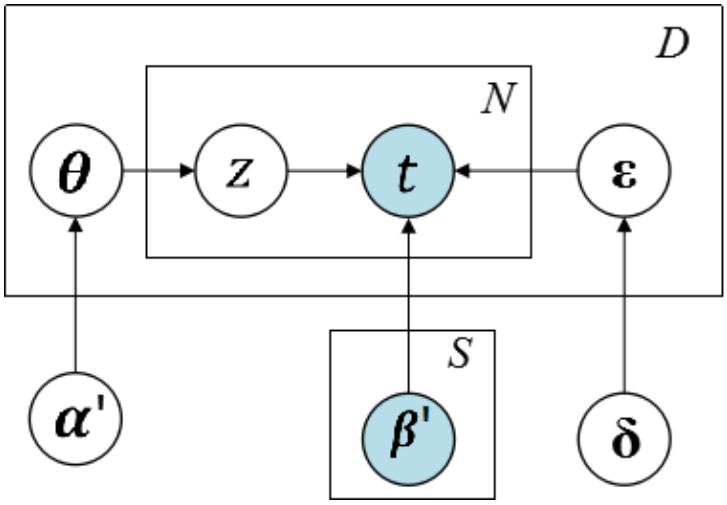


Figure 4.4: Graphical representation of the denoise deconvolution model: Each of the D profiles is associated with a mixture proportion $\boldsymbol{\theta}_d$ (regularized by hyperparameter α) and a topic panel consisting of S pure sources $\{\boldsymbol{\beta}'_{\boldsymbol{s}}\}$ and ϵ_d (independent from pure sources) with Dirichlet prior δ. Each of the N ions in a profile $t_{n,d}$ is sampled from a topic indicated by $z_{n,d}$.

4.4.2 Probabilistic generative representation of DDM

The complete likelihood function is modified to Eq. (4.13). Specifically, we have two observable variables indicating D expression profiles from all groups: $\{\boldsymbol{t}_d\}_{d=1,\cdots,D},\ \boldsymbol{t}_d \in \mathbb{R}^N$, and S pure source profiles (e.g., cancerous and noncancerous sources) recognized by purification model: $\{\boldsymbol{\beta}'_{\boldsymbol{s}}\}_{s=1,\cdots,S},\ \boldsymbol{\beta}_{\boldsymbol{s}} \in \mathbb{R}^L$. Additionally, we include a Dirichlet prior $\boldsymbol{\delta}$ to generate sample-specific noise topic $\{\boldsymbol{\varepsilon}_d\}_{d=1,\cdots,D},\ \boldsymbol{\varepsilon}_d \in \mathbb{R}^L$. In our analysis, we normalize all profiles to have identical total ion counts of N and consider L biomolecules that are consistently detected in all the samples. For convenience, we represented the normalized profiles in two ways. Each heterogeneous observed profile $\boldsymbol{t}_d$ is represented via N ions, with $t_{d,n} = \{1, 2, \cdots, L\}$ denoting the origin biomolecule of n^{th} ion. Each topic profile $\boldsymbol{\beta}_s$ is represented via L biomolecules, with $\beta_{s,l}$ denoting the ion counts of l^{th} biomolecule, and $\sum_{l=1}^{L} \beta_{s,l} = N$ (same for ε_d). The second expression can be further normalized by N to give a representation of multinomial distribution as a topic.

$$\begin{aligned}&\mathcal{L}(\boldsymbol{t}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\varepsilon} | \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\delta}) \\ &= \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha) \cdot p(\boldsymbol{\varepsilon}_d|\boldsymbol{\delta}) \cdot \prod_{n=1}^{N} [p(z_{d,n}|\boldsymbol{\theta}_d) \cdot p(t_{d,n}|z_{d,n}, \boldsymbol{\theta}_d, \boldsymbol{\beta}', \boldsymbol{\varepsilon}_d)]\end{aligned} \tag{4.13}$$

Where,

$$p(\boldsymbol{\varepsilon}_d|\boldsymbol{\delta}) = \text{Dirichlet}(\boldsymbol{\varepsilon}_d|\boldsymbol{\delta}) \tag{4.14}$$

$$p(t_{d,n}|z_{d,n} \leq S, \boldsymbol{\theta}_d, \boldsymbol{\beta}', \boldsymbol{\varepsilon}_d) = \text{Multinomial}(t_{d,n}|\boldsymbol{\beta}'_{z_{d,n}}) \tag{4.15}$$

$$p(t_{d,n}|z_{d,n} = S+1, \boldsymbol{\theta}_d, \boldsymbol{\beta}', \boldsymbol{\varepsilon}_d) = \text{Multinomial}(t_{d,n}|\boldsymbol{\varepsilon}_d) \tag{4.16}$$

Each noise profile $\boldsymbol{\varepsilon}_d$ together with S pure sources $\{\boldsymbol{\beta}'_s\}$ forms a sample-specific topic panel. The mixture proportion of topics determines $z_{d,n}$, indicating which source (i.e., ε_d or $\{\boldsymbol{\beta}'_s\}$) each ion originates from. We inferred the latent variables $\boldsymbol{\delta}, \{\boldsymbol{\varepsilon}_d\}_{d=1,\cdots,D}, \{\boldsymbol{\theta}_d\}_{d=1,\cdots,D}$, and estimated the parameters using the two-step learning approach developed based on variational EM algorithm same as LDA [12, 40]. Compared to the original purification model, the assumption that the latent topic originates from the known topics is eliminated because this latent variable $\boldsymbol{\delta}$ represents noise profile which is independent from other sources. When initializing the mixture proportion prior α', we assign a lower value to its attribute associated with $\boldsymbol{\varepsilon}_d$. Whereas in previous purification model, we expected a higher proportion on the underlying pure source $\boldsymbol{\gamma}_d$.

4.5 Mass spectrometric datasets and evaluation

The experimental data were acquired by analyses of tissue and blood samples from patients with hepatocellular carcinoma (i.e., HCC, case group) and liver cirrhosis (control group)

[8–10, 52]. HCC is a highly heterogeneous disease both at the molecular and clinical levels [58]. Whereas all patients in this study were diagnosed with liver cirrhosis, about half of them were also diagnosed with HCC. Contamination occurs due to the influence from cirrhotic constituents in HCC samples. In this study, we used GC-MS data acquired by analysis of metabolites in 105 tissues and LC-MS data acquired by analysis of proteins in sera from 116 subjects.

4.5.1 GC-MS based metabolomic dataset

As a pilot project, fifteen liver tissues were collected from 10 participants recruited at MedStar Georgetown University Hospital. As shown in Figure 4.5, the tissues were collected from 5 HCC cases (5 tumor and 5 adjacent cirrhotic tissues) and 5 patients with liver cirrhosis. Samples were profiled through Agilent 7890A gas chromatography coupled with LECO's time-of-flight mass spectrometer to characterize the metabolome alterations associated with HCC development in cirrhotic patients. We identified 559 metabolites after preprocessing the GC-MS raw data by ChromaTOF GC software with True Signal Deconvolution package (Leco Corporation). Two types of purification are investigated on the data. One is to purify HCC profiles by removing contaminants from cirrhotic profiles. The other is to purify adjacent cirrhotic profiles by reducing the impact of the profiles attributed to HCC.

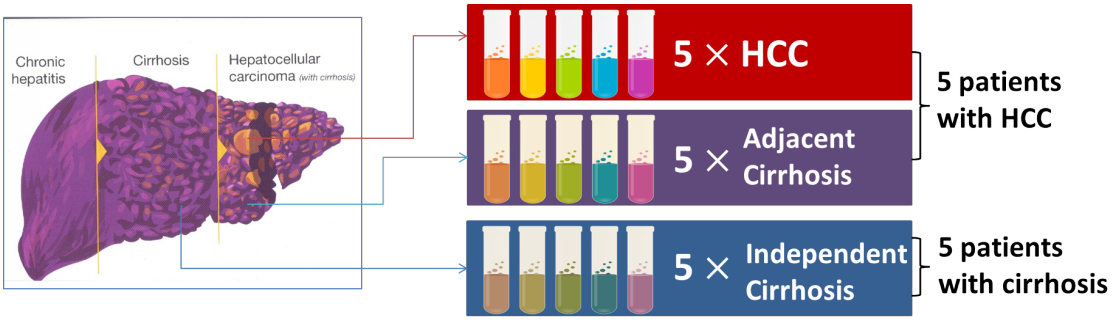


Figure 4.5: Fifteen tissue samples collected from 10 subjects (5 HCC cases and 5 cirrhotic controls). Five tumor and five adjacent cirrhotic tissues were obtained from the 5 HCC cases. Additional 5 cirrhotic tissues were obtained from the 5 independent subjects with liver cirrhosis.

4.5.2 LC-MS based proteomic dataset

We acquired 116 proteomic data by analysis of sera from 57 HCC cases and 59 patients with liver cirrhosis recruited from the hepatology clinics at MedStar Georgetown University Hospital. Following depletion and digestion, proteins extracted from sera were injected into a 3000 Ultimate nano-LC system interfaced to LTQ Orbitrap Velos and TSQ Vantage mass

spectrometers in untargeted and targeted analyses, respectively. Proteins were identified and quantified by MaxQuant [59] and Skyline [60] in preprocessing untargeted and targeted LC-MS data, respectively. Finally, 101 proteins that were consistently identified across 116 samples were selected as intensity-level features in expression profiles (i.e., $L = 101$). All profiles were normalized to the mean total-ion-counts at $N = 1.68 \times 10^8$. It is still not clear how the development of tumor in liver directly affect the alterations in blood. We hypothesize that there are some impacts from cirrhotic constituents contributing to the HCC profile in serum. The contamination may occur in an indirect way through, for example, secreted biomolecules instead of adjacent tissue cells. We apply the purification to remove the influence from cirrhotic contaminants.

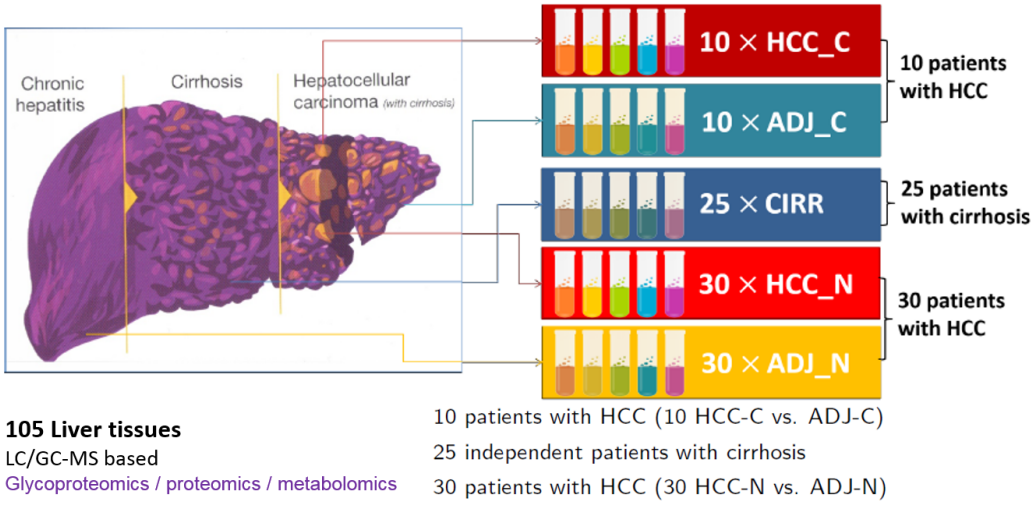


Figure 4.6: 105 liver tissues collected from 65 patients. 40 (10 tumor and 10 adjacent cirrhotic tissues from 10 patients, 30 tumor and 30 adjacent normal tissue from other 30 patients) were developed with HCC and 25 were only cirrhosis.

4.5.3 Multi-group metabolomic datasets

We later extend to use 105 liver tissues collected from 65 participants recruited at MedStar Georgetown University Hospital. As shown in Figure 4.6, the participants consist of 40 HCC cases (10 tumor and 10 adjacent cirrhotic tissues from 10 patients, 30 tumor and 30 adjacent normal tissue from other 30 patients) and 25 patients with liver cirrhosis. To characterize the metabolome alterations associated with HCC development in cirrhotic and normal tissues in patients. We performed metabolomic analysis of the tissues using two platforms (GC-TOF-MS and LC-QTOF-MS) The tissues were homogenized and metabolite extraction was performed simultaneously for both GC-MS and LC-MS analyses. Briefly, 10 mg of liver tissue was homogenized on ice with 1 mL of pre-chilled Methanol:Water (1:1) in order to extract the metabolites and precipitate the proteins. Samples were then centrifuged and the resulting supernatant was divided into two aliquots (one for GC-MS and

one for LC-MS analysis). GC-MS data were acquired by analysis of the extracted metabolites using Agilent 7890A GC coupled to LECO Pegasus HT, equipped with an electron ionisation source and TOF analyzer using splitless injection. We used ChromaTOF with True Signal Deconvolution package for data pre-processing, including baseline calculation, peak finding, deconvolution and identification. LECOs Statistical Compare software tool was used for alignment of the GC-MS data. LC-MS data were acquired by analysis of the extracted metabolites using Waters ACQUITY UPLC coupled to Synapt G2-Si QTOF-MS, operating in positive and negative polarity. LC-QTOF-MS data were first converted into Network Common Data Format (NetCDF) using DataBridge Program from the MassLynx software (Waters). Peak detection, alignment, and ion annotation were performed using XCMS and CAMERA. We detected 726, 2286, and 593 analytes in GC-MS, LC-MS positive and negative modes respectively.

4.5.4 Synthetic datasets

Before applying the models to experimental data, we generated synthetic datasets by artificially mixing real LC-MS data on both intensity and scan levels, and evaluated the model based on its performance of deconvolving the mixed data. We generated synthetic data based on the 116 LC-MS profiled serum proteomic dataset. We assume here that human sera are homogeneous specimens. Hence we can mix them to simulate heterogeneous cancer profiles. Figure 4.7 shows the generative process of 30 synthetic cancer profiles with contamination, following the steps below:

(i) Average the profiles of HCC group, $\{\boldsymbol{\lambda}_s\}_{s=1,\cdots,57}$, to obtain an average cancer profile $\boldsymbol{\gamma}'$, which is close to the real cancerous profile for HCC.

(ii) Sample 30 individual pure cancer profiles $\{\boldsymbol{\gamma}_d\}_{d=1,\cdots,30}$ from a Dirichlet distribution parameterized by $\boldsymbol{\gamma}'$ and $\kappa_d = \frac{1}{\min_l(\gamma'_l)}$.

(iii) Randomly select a subset of cirrhotic profiles $\{\boldsymbol{\beta}_m\}_{m=1,\cdots,M}$ ($M = 9$ in this simulation) as sources of contamination. Normalize them into form of multinomial distribution.

(iv) Combine M cirrhotic profiles with each of the pure cancer profiles to create 30 topic panels, each consisting of $M + 1 = 10$ profiles.

(v) Sample 30 mixture proportions $\{\boldsymbol{\theta}_d\}_{d=1,\cdots,30}$ from a Dirichlet distribution, as in Eq. (4.2), parameterized by $\boldsymbol{\alpha} = [1, \cdots, 1, 5]$, which is uniform for the first nine constituents (contaminants) and with a larger value assigned to last constituent (cancer origin). This ensures a larger proportion of cancerous component in final cancer profile.

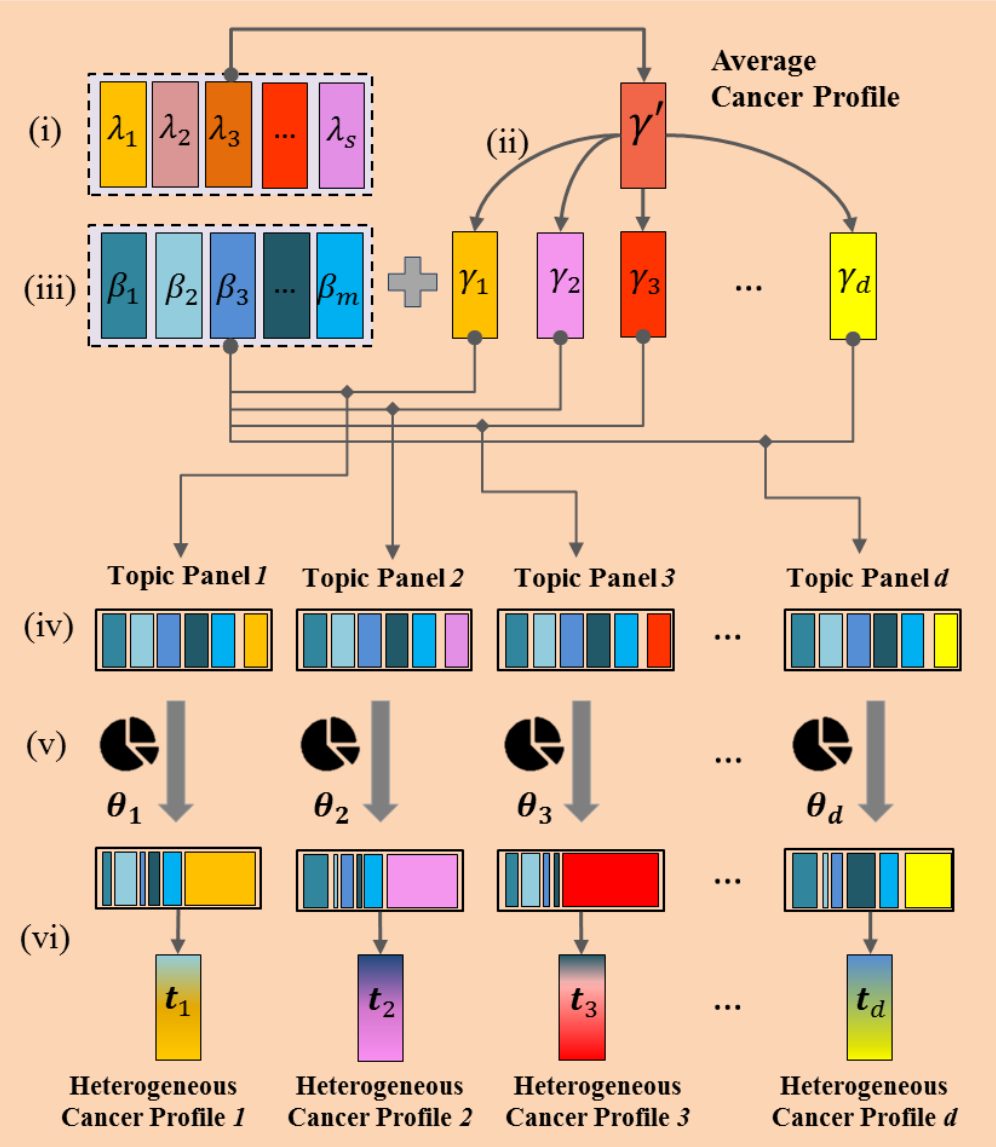


Figure 4.7: Generative process of heterogeneous cancer profiles. (i) average cancer profiles in case group; (ii) generate sample-specific pure cancer profile; (iii) select sources of contaminants in control group; (iv) form topic panels; (v) generate sample-specific mixture proportions; (vi) generate synthetic cancer profiles.

(vi) Sample a topic indicator $z_{d,n}$ from $\boldsymbol{\theta}_d$, and sample a $t_{d,n}$ from $\boldsymbol{\beta}_z$ if $z \leq M$ or $\boldsymbol{\gamma}'_d$ otherwise. Repeat the sampling for $N = 1.68 \times 10^8$ times to generate a synthetic cancer profile $\boldsymbol{t}_d$.

Each of these 30 heterogeneous cancer profiles is a mixture of a pure cancer profile and multiple contaminants. The intensity-level purification procedure will help retrieve the pure

cancer profile and estimate the sample purity as well as proportions of contaminants. Similar to intensity-level simulation, we generated heterogeneous dataset using scan-level features, i.e. EICs, exported from Skyline [60]. Corresponding to 101 proteins, 187 peptides with 561 scan-level features were extracted in each of the 116 samples. Each feature contains 60 scans representing a chromatographic peak as illustrated in Figure 4.8. We followed the same steps (i-vi) except that we average and blend EIC peaks instead of protein intensities. Finally, 30 heterogeneous cancerous samples, each characterized by a list of 561 EICs, are generated.

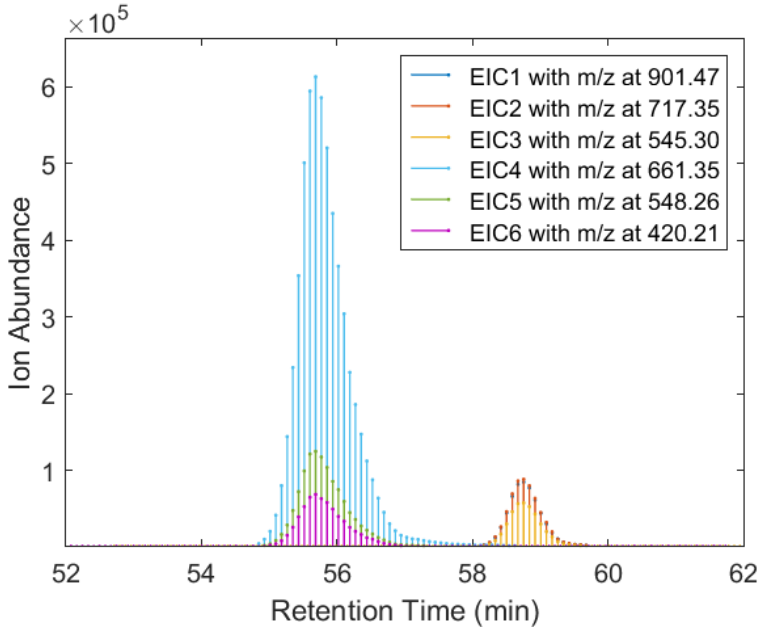


Figure 4.8: Extracted ion chromatograms from LC-MS based serum proteomic data Extracted ion chromatogram is characterized by m/z, retention time, and ion abundance.

4.5.5 Evaluation methods

We evaluated the performances of our proposed models on both synthetic and real experimental LC/GC-MS datasets in consideration of the following three goals: 1) to test on intensity level if the model can reasonably estimate the proportion of mixtures in each of the synthetic profiles and recover the pure cancer profiles underneath; 2) to demonstrate if the scan-level purification model gives more accurate estimation on synthetic data; 3) to investigate the benefits of using these models to purify samples from cancer patients collected in our previous differential analysis studies.

Outputs of intensity-level model include the sample-specific mixture proportions $\{\boldsymbol{\theta}_d^*\}$, pure cancer profiles $\{\boldsymbol{\gamma}_d^*\}$, and the estimated average cancer profile $\boldsymbol{\gamma}'^*$. Whereas, we expect outputs of sample-specific mixture proportion $\{\boldsymbol{\theta}_d^*\}$, pure ion abundance $\{\boldsymbol{x}_\gamma^*\}$, peak shape

function parameters $\boldsymbol{\phi}^*$ from extended model. For synthetic datasets, we compare the estimated proportions of mixtures $\{\boldsymbol{\theta}_d^*\}$ with the true ones ($\{\boldsymbol{\theta}_d\}$) used to generate the synthetic data. Estimation error ratio for a single sample is defined in Eq. (4.17).

$$\xi_d(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \frac{||\boldsymbol{\theta}_d^* - \boldsymbol{\theta}_d||_1}{||\boldsymbol{\theta}_d||_1} \times 100\%,\ d = 1, \cdots, 30 \tag{4.17}$$

Different from point-wise intensities, the scan-level estimation error ratio for a single sample is defined in Eg. (4.18)

$$\xi_d(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}) = \frac{||\sum_{s=1}^{S}[\boldsymbol{\gamma}_d^*(s) - \boldsymbol{\gamma}_d(s)]||_1}{||\sum_{s=1}^{S}\boldsymbol{\gamma}_d(s)||_1} \times 100\%,\ d = 1, \cdots, 30 \tag{4.18}$$

For experimental datasets, we evaluated the performances in multiple aspects including statistical significance of the candidate biomarkers, ROC curves in distinguishing the biological groups, and pathway analysis results.

4.6 Results and discussions

4.6.1 Synthetic datasets

We applied current model and the extended model to the synthetic intensity-level and scan-level LC-MS datasets, respectively. By incorporating peak detection algorithms, we can further compare the purification performances between the two topic models.

Intensity-level purification

We obtained an average error ratio of mixture proportion $\bar{\xi}_d(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ at 2.33%, indicating a good characterization of original proportions. The comparison of proportion parameters for the first six profiles is depicted in Figure 4.9 using radar charts and scatter plots. As shown in the figure, the estimation in each profile has captured consistent patterns as the ground truth in each of the 10 components. We achieved an average correlation coefficient between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_d^*$ at 0.975. The model accurately recognized those non-cancerous constituents contributed as small as 5% in each sample. The proportion of cancerous origin is overestimated in some samples due to the smaller contributions from the contaminants. The differences between $\boldsymbol{\theta}_d$ and $\boldsymbol{\theta}_d^*$ are also related to the recovered pure cancer profiles $\{\boldsymbol{\gamma}_d^*\}$. Similarly, we have the average estimation error ratio for sample-specific pure cancer profiles $\bar{\xi}_d(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}) = 6.51\%$, which is smaller than $\bar{\xi}_d(\boldsymbol{t}, \boldsymbol{\gamma}) = 16.57\%$, i.e., the error ratio between unpurified cancer profile and true cancer profile. Figure 4.10 shows scatter plots of 101 proteins in unpurified cancer profile $\{\boldsymbol{t_d}\}_{d=1,\cdots,6}$ versus true cancer profile (blue) and in purified cancer profile versus true cancer profile (orange). The average correlation coefficient increases from 0.986 to 0.999

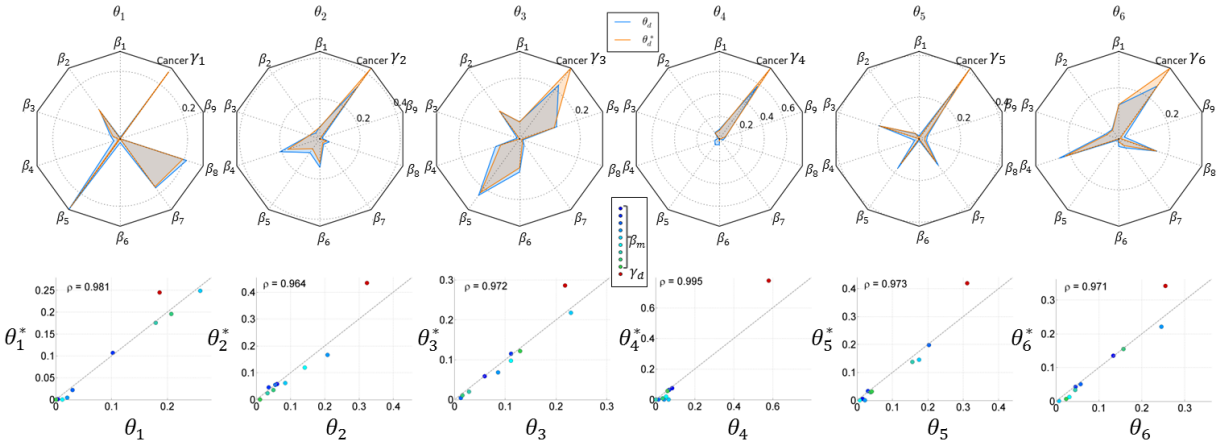


Figure 4.9: Similarity evaluation on θ. Comparison between estimated $\boldsymbol{\theta}^*$ and true mixture proportions $\boldsymbol{\theta}$ for the first six profiles. Top: radar charts with 10 spokes, each representing a source in topic panel. The proportion of each source is depicted by the length of lines with color (orange for estimation $\boldsymbol{\theta}^*$ and blue for ground truth $\boldsymbol{\theta}$). Bottom: scatter plots of corresponding proportions in ground truth $\boldsymbol{\theta}$ and estimation $\boldsymbol{\theta}^*$. The correlation coefficients ρ are given on the left-top.

Table 4.1: Estimation error ratio $\xi_d(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ means (standard deviations) based on 100 realizations.

SNR	LDA	DeMix	IPM
∞	30.87(8.95)	4.395(1.014)	**2.331**(0.541)
50	38.56(10.12)	6.753(2.455)	**4.198**(1.705)
25	51.45(12.21)	**12.74**(3.258)	13.71(4.302)
10	76.18(14.25)	35.78(9.854)	**32.25**(10.98)

after purification. The effects of purification are illustrated in Figure 4.11 by projecting the high-dimensional (dim = 101) profiles onto their top three principal components. We observe that the purified cancer profiles were more distant from non-cancerous profiles, and regularized towards an average cancer profile.

The above results were based on a synthetic dataset without considering noise. We repeated the realization by adding various energy of noises (with SNR at $50, 25, 10$ dB) to the generated heterogeneous data and compared the performances with *no purificaiton*, purification with *LDA*, and *DeMix* (version 1.0.1) [61], which is a currently available tool from genomic field. The results are summarized in the Table 4.1, 4.2, and 4.3.

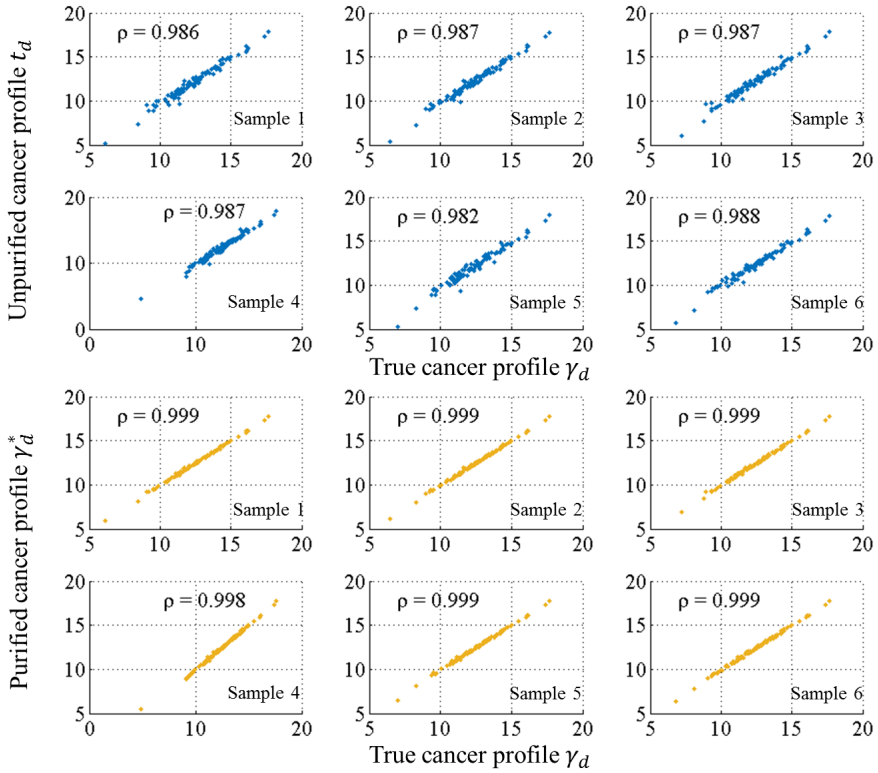


Figure 4.10: Similarity evaluation on γ. The first six out of 30 scatter plots of unpurified cancer profiles versus true cancer profiles (blue) and corresponding scatter plots of purified cancer profiles versus true cancer profiles (orange). The correlation coefficients ρ between each pair of profiles are given on the left-top.

Table 4.2: Estimation error ratio $\xi_d(\boldsymbol{\gamma}^*, \boldsymbol{\gamma})$: means (standard deviations) based on 100 realizations.

SNR	No purification⋆	LDA	DeMix	IPM
∞	16.57(3.432)	12.87(2.043)	7.294(1.821)	**6.510**(1.015)
50	24.11(5.217)	24.05(5.885)	16.33(3.753)	**10.20**(2.781)
25	30.66(7.514)	31.25(7.356)	**19.34**(4.255)	20.16(4.041)
10	39.78(8.021)	36.75(7.953)	25.64(4.863)	**21.53**(3.872)

Scan-level purification

We first evaluated the purification power in the case of scan-level features. The average estimation error ratio of mixture proportions is 3.57% by Eq. (4.17). In terms of recovering the underneath pure feature list, we achieved the average estimation error ratio for sample-specific pure cancerous feature list $\bar{\xi}_d(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}) = 3.12\%$, which is smaller than $\bar{\xi}_d(\boldsymbol{t}, \boldsymbol{\gamma}) = 9.61\%$, i.e., the error ratio between unpurified cancerous feature list and ground truth. The purifi-

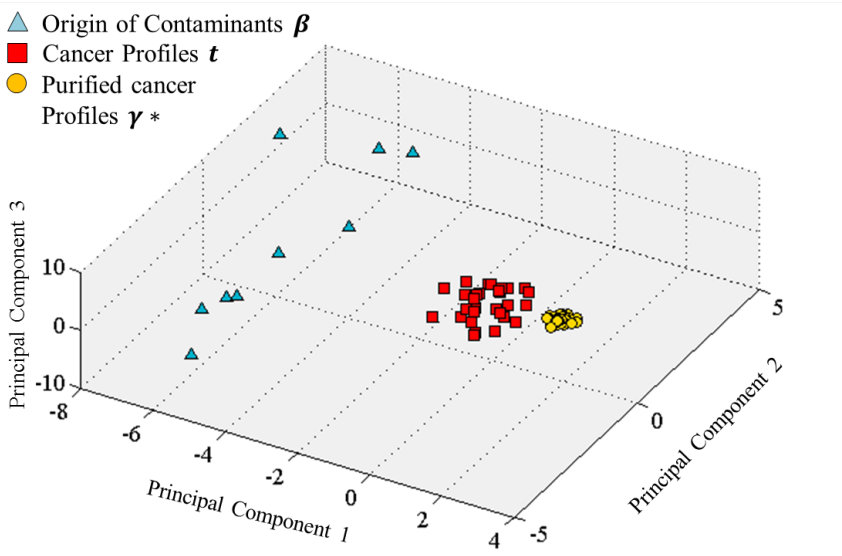


Figure 4.11: PCA analysis on simulated dataset. 30 cancer profiles $\{\boldsymbol{t}_d\}$ (red square), 30 purified cancer profiles $\{\boldsymbol{\gamma}_d^*\}$ (yellow circle), and 9 sources of cirrhotic contaminants $\{\boldsymbol{\beta}_m\}$(blue triangle).

Table 4.3: Correlation coefficients $\rho\langle\boldsymbol{\gamma}^*, \boldsymbol{\gamma}\rangle$: means (standard deviations) based on 100 realizations.

SNR	No purification⋆	LDA	DeMix	IPM
∞	0.985(0.002)	0.988(0.003)	0.998(2.455)	**0.999**(0.001)
50	0.947(0.005)	0.955(0.005)	0.988(0.002)	**0.995**(0.002)
25	0.875(0.025)	0.895(0.015)	0.926(0.014)	**0.950**(0.005)
10	0.755(0.035)	0.795(0.022)	0.890(0.015)	**0.940**(0.012)

cation with scan-level features works to some extent but it is also interesting to prove the extended model works in a more accurate way than intensity-level topic model. To allow intensity-level purification model to handle scan-level synthetic dataset, we employed peak detection algorithms (i.e., through successive convolution with a 4^{th} order Savitzky-Golay smoothing filter and a first-order derivative of a Gaussian kernel with window width of 25 scans, standard deviation of 3) to transfer EIC peaks into intensities using area under curve. The same algorithm is applied for transferring purified peak list resulted from the extended model. We obtained a greater distance of mixture proportion with $\bar{\xi}_d^{\mathcal{I}}(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ at 7.23% if using intensity-level purification model, compared to half ($\bar{\xi}_d^{\mathcal{S}}(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = 3.57\%$) achieved by extended scan-level purification model. Similarly, we compared the performances with consideration of noises, summarized in Table 4.4, 4.5, and 4.6. We observed that the SPM tend to be robust to the noise.

Table 4.4: Estimation error ratio $\xi_d(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ means (standard deviations) based on 100 realizations.

SNR	LDA	DeMix	IPM	**SPM**
∞	42.71(11.59)	10.69(3.01)	7.231(2.526)	**3.568**(1.422)
50	48.33(10.12)	13.75(2.55)	13.85(2.15)	**3.922**(1.305)
25	57.29(8.99)	22.57(4.58)	20.41(5.20)	**4.392**(1.823)
10	83.48(12.52)	27.62(5.84)	25.16(6.52)	**9.573**(2.117)

Table 4.5: Estimation error ratio $\xi_d(\boldsymbol{\gamma}^*, \boldsymbol{\gamma})$: means (standard deviations) based on 100 realizations.

SNR	No purification$\star$	LDA	DeMix	IPM	**SPM**
∞	9.61(1.432)	8.72(2.043)	4.239(1.821)	4.231(1.206)	**3.120**(0.085)
50	15.14(2.71)	14.47(2.15)	14.33(3.73)	13.85(2.15)	**4.201**(0.091)
25	29.32(4.54)	23.95(4.36)	19.13(3.27)	18.41(3.20)	**6.571**(0.523)
10	35.22(8.14)	34.52(7.51)	20.46(5.63)	23.16(5.85)	**10.454**(0.946)

Table 4.6: Correlation coefficients $\rho\langle\boldsymbol{\gamma}^*, \boldsymbol{\gamma}\rangle$: means (standard deviations) based on 100 realizations.

SNR	No purification$\star$	LDA	DeMix	IPM	**SPM**
∞	0.985(0.005)	0.988(0.003)	0.998(0.002)	**0.999**(0.001)	**0.999**(0.001)
50	0.935(0.005)	0.955(0.015)	0.988(0.002)	0.975(0.005)	**0.998**(0.002)
25	0.885(0.025)	0.895(0.015)	0.945(0.014)	0.955(0.005)	**0.990**(0.005)
10	0.785(0.035)	0.815(0.025)	0.920(0.015)	0.925(0.015)	**0.980**(0.015)

4.6.2 LC-MS based proteomic dataset

We treated all 59 cirrhotic profiles as origins of contaminants to purify 57 HCC profiles. We plotted these profiles using their first three principal components in Figure 4.12.

Similar to the simulation result, we observed a clearer distinction between HCC and cirrhotic profiles after purification. To further understand the improvements, we carried out the following analyses on both purified and unpurified profiles.

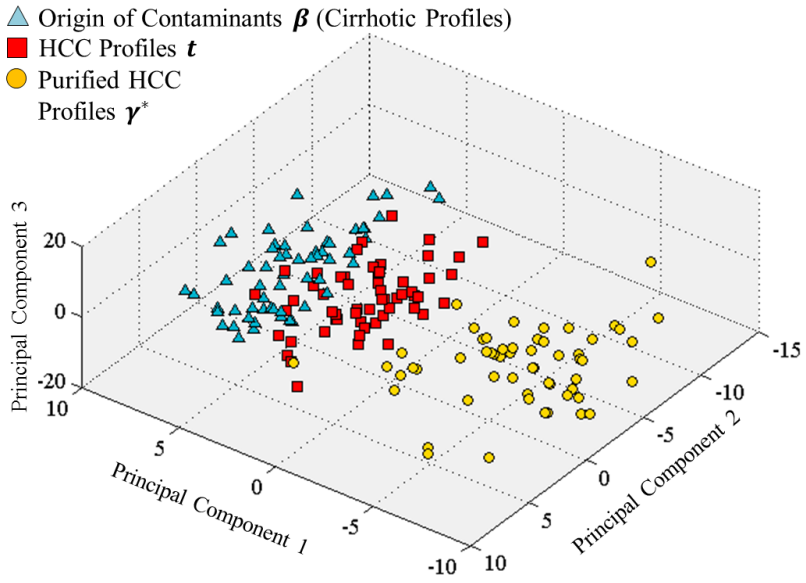


Figure 4.12: PCA analysis on proteomic dataset. 57 HCC profiles $\{\boldsymbol{t}_d\}$ (red square), 57 purified HCC profiles $\{\boldsymbol{\gamma}_d^*\}$ (yellow circle), and 59 sources of cirrhotic contaminants $\{\boldsymbol{\beta}_m\}$(blue triangle).

Firstly, in statistical analysis, the most relevant proteins with differential intensities between HCC cases and cirrhotic controls were selected using t-test, and the associated p-values were adjusted based on multiple testing correction (FDR ≤ 0.05). We found 43 proteins with significant change in expression between the two groups. The number of reported significant proteins under the same testing method increased from 43 to 75 after purification. The majority of the proteins identified in original profiles (40 out of 43) remained significant after purification. If purified based on scan-level features, the number of significant proteins also increased to 69, among which 38 and 61 are overlapped with unpurification and intensity-level purification results, respectively.

Figure 4.13a, 4.13b, and 4.13c show ROC curves for each of the 43, 75, and 69 significant proteins, respectively. A bootstrap method (1000 bootstrap replicates) was used to compute the 95% confidence interval (CI) of the area under each ROC curve. After intensity-level and scan-level purification we respectively achieved an average AUC of 0.793 (with 95% CI

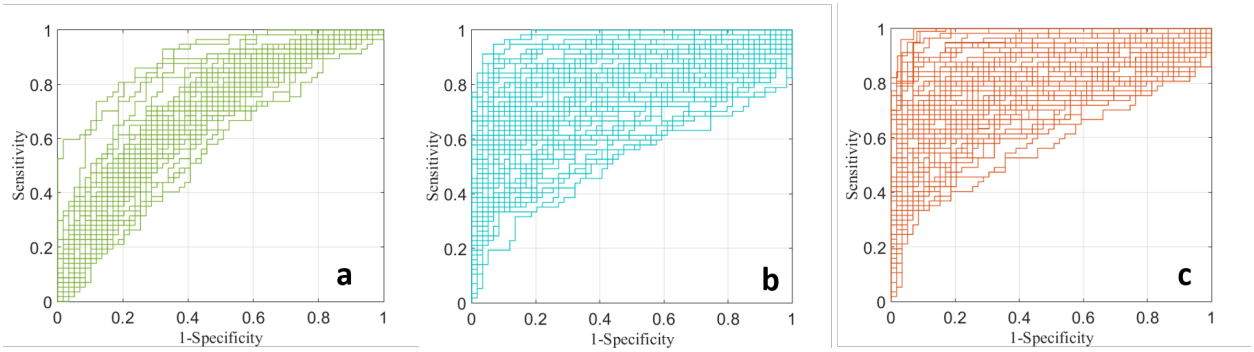


Figure 4.13: ROC curves of significant proteins. **a**: ROC curves for each of 43 significant proteins before purification ($\overline{\text{AUC}} = 0.706, 95\%\text{CI } [0.606, 0.795]$). **b**: ROC curves for each of 75 significant proteins after intensity-level purification ($\overline{\text{AUC}} = 0.793, 95\%\text{CI } [0.700, 0.863]$).**c**: ROC curves for each of 69 significant proteins after scan-level purification ($\overline{\text{AUC}} = 0.811, 95\%\text{CI } [0.719, 0.890]$).

Table 4.7: Signaling pathways (number of significant proteins involved in the pathway)

No Purification	**IPM**	**SPM**
Complement and coagulation cascades (13)	Complement and coagulation cascades (18)	Complement and coagulation cascades (19)
Systemic lupus erythematosus (5)	Systemic lupus erythematosus (6)	Systemic lupus erythematosus (4)
Prion diseases(4)	Prion diseases (4)	Prion diseases (4)
-	⋆ **PPAR signaling pathway** (5)	⋆**PPAR signaling pathway** (6)

at [0.700, 0.863]) and 0.811(with 95% CI at [0.719, 0.890]), both higher than 0.706 (with 95% CI at [0.606, 0.795]) for original biomarkers. More powerful biomarkers were selected after scan-level purification.

Finally, we used DAVID [62] (version 6.7) to identify significant signaling pathways, where the UniProt IDs of the significant proteins were mapped to the KEGG [63] database. As shown in Table 4.7, three pathways were reported from the original list of significant proteins. Following intensity-level and scan-level purifications, we found peroxisome proliferator-activated receptor (PPAR) signaling pathway with five and six significant proteins involved in addition to the three pathways (complement and coagulation casades, systemic lupus erythematosus, and prion disease) identified without purification. This is interesting in light of previous reports linking cancer and PPARs expressed in human liver [64].

4.6.3 GC-MS based metabolomic dataset

Heterogeneity issue is more intuitive in tissue samples, where the contaminations originate from the neighboring non-homogeneous cells. We first purified the HCC profiles $\{\boldsymbol{t_d}\}_{d=1,\cdots,5}$ using independent cirrhotic profiles $\{\boldsymbol{\beta_m}\}_{m=1,\cdots,5}$ as the sources of contamination. Without purification, none of the 559 metabolites passed the statistical test as significant (FDR adjusted p-value ≤ 0.05). However, seven metabolites were identified as significant after the profiles were purified. For the adjacent cirrhotic profiles $\{\boldsymbol{\psi_d}\}_{d=1,\cdots,5}$, we applied the model to remove contaminations from any neighboring cancerous cells. We expected to observe that the purified adjacent cirrhotic profiles became close to independent cirrhotic profiles. The dissimilarity, defined in Eq. (4.17), between independent and adjacent cirrhotic profiles is $\bar{\xi}(\boldsymbol{\psi}, \boldsymbol{\beta}) = 28.3\%$, and goes down to $\bar{\xi}(\boldsymbol{\psi^*}, \boldsymbol{\beta}) = 24.9\%$ after purification. The improvements are less substantial compared to the previous datasets, presumably due to the limited sample size and potential overfitting issue.

4.6.4 Multi-group metabolomic datasets

As for multi-group metabolomic datasets, four types of purification and a denoise deconvolution method will be investigated on the data matrices.

P1: Purify HCC profiles (HCCc) by removing contaminants from adjacent cirrhotic tissues (CIRRh).

P2: Purify cirrhotic profiles (CIRRh) by removing contaminants from adjacent cancerous tissues (HCCc).

P3: Purify normal profiles (NORMh) by removing contaminants from adjacent cancerous tissues (HCCn)

P4: Purify HCC profiles (HCCn) by removing contaminants from adjacent normal tissues (NORMh).

D: Deconvolute original profiles using uncovered pure sources (pHCCc, pHCCn, pCIRRh, CIRR, pNORMh).

As illustrated in Figure 4.14, we first purify each group using profiles generated from the adjacent tissues. Then we randomly select five purified profiles in each group and combine the 25 sources (10 cirrhosis, 10 HCC, and 5 normal) as a panel for further deconvolution.

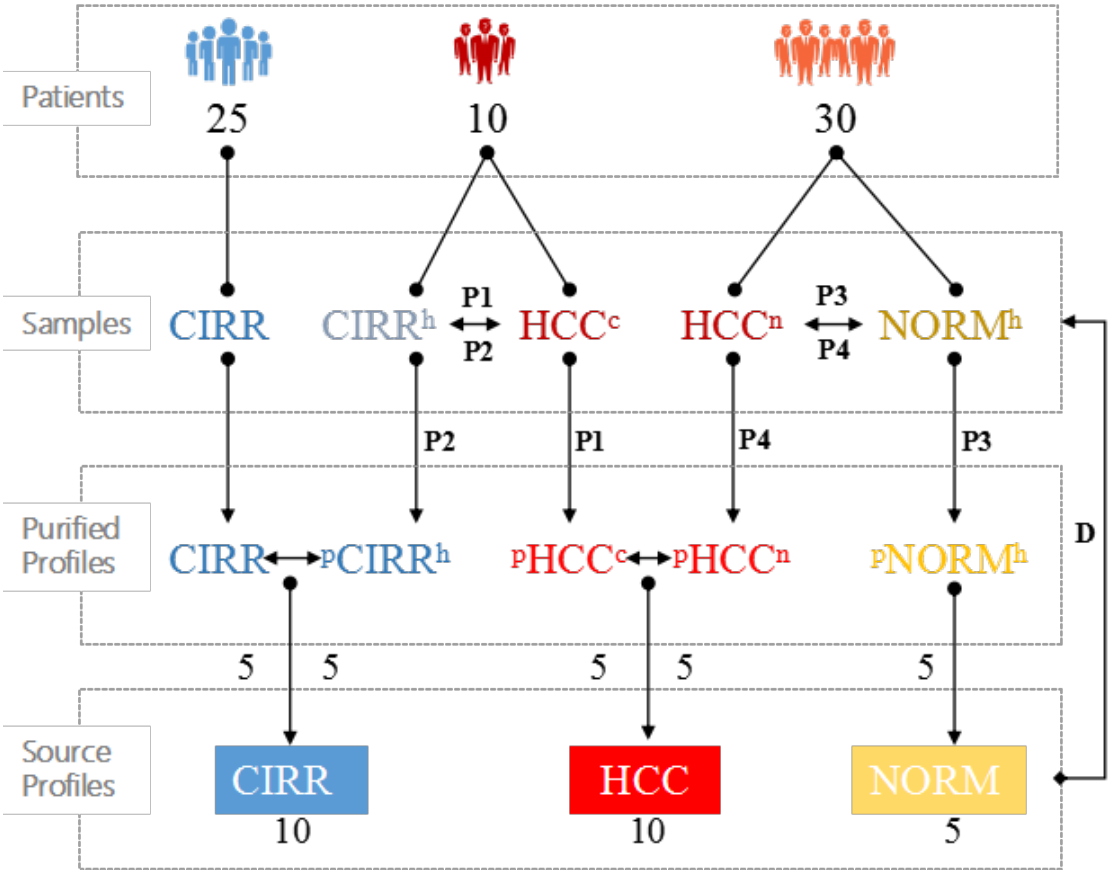


Figure 4.14: The workflow design for applications of purification and deconvolution models. We first purify heterogeneous samples between HCC samples and their adjacent normal/cirrhotic tissues (P1-P4). Then a deconvolution (D) of original profiles is conducted based on uncovered pure sources.

We evaluate the performances of the purification and denoise deconvolution methods on both GC-MS and LC-MS (positive and negative modes) datasets in consideration of the following three standards: 1) increased discrimination is observed in each pair of groups after purification; 2) samples could be clustered into sub-groups according to their labels; 3) deconvolution model can predict the group labels of measured samples.

Pairwise purification

Figure 4.15 shows the principal component analysis results on GC-MS dataset for each purification procedure. Profiles before and after purification as well as the reference contaminants are scattered onto three principal components. Increased distances are observed in all pairs of groups after purification. Pure profiles tend to cluster to a centroid and the intra-group variances are reduced. Similar performances are observed on two LC-MS datasets.

Table 4.8 presents the performances before and after purification in terms of 1) No. Sig:

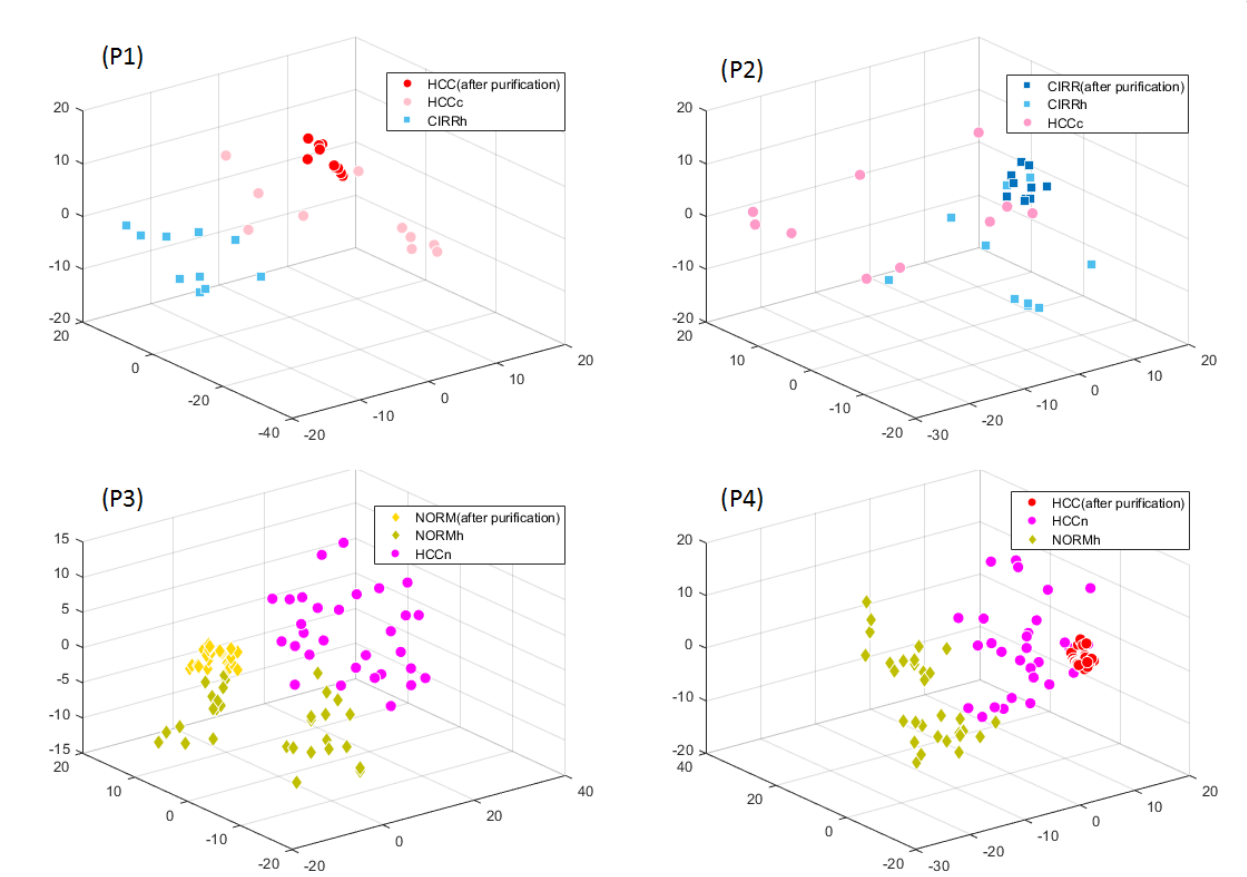


Figure 4.15: Samples expressed by the first three principal components after PCA in each of the four purification procedures based on GC-MS data. Compared samples are contaminants, purified groups and original groups.

the number of significant metabolites using the Student's t-test (adjusted $p-$value ≤ 0.05); 2) AUC: the average area under curve of ROC analysis for each significant metabolite and 95% confidence interval based on bootstrap methods (1000 times); 3) ACC: the classification accuracy using SVM-RFE with 10-fold cross validation. We consistently observed increase in the number of significant metabolites and improved classification accuracy after purification. We concatenate four purified profiles together with independent cirrhotic profiles and apply PCA. As shown in Figure 4.16 based on GC-MS the samples from the same groups tend to have less separation than samples from different groups. However, the two types of adjacent HCC are not clustered together as we expected. Two distinct contaminant references (i.e., CIRRh, NORMh) may lead to purification of adjacent HCC in different directions. Also, the HCC and adjacent cirrhotic tissues have relatively small sample size, which affected the performance of our methods for these groups. In the deconvolution step, we keep these HCC profiles (pHCCc, pHCCn) as two classes of pure sources.

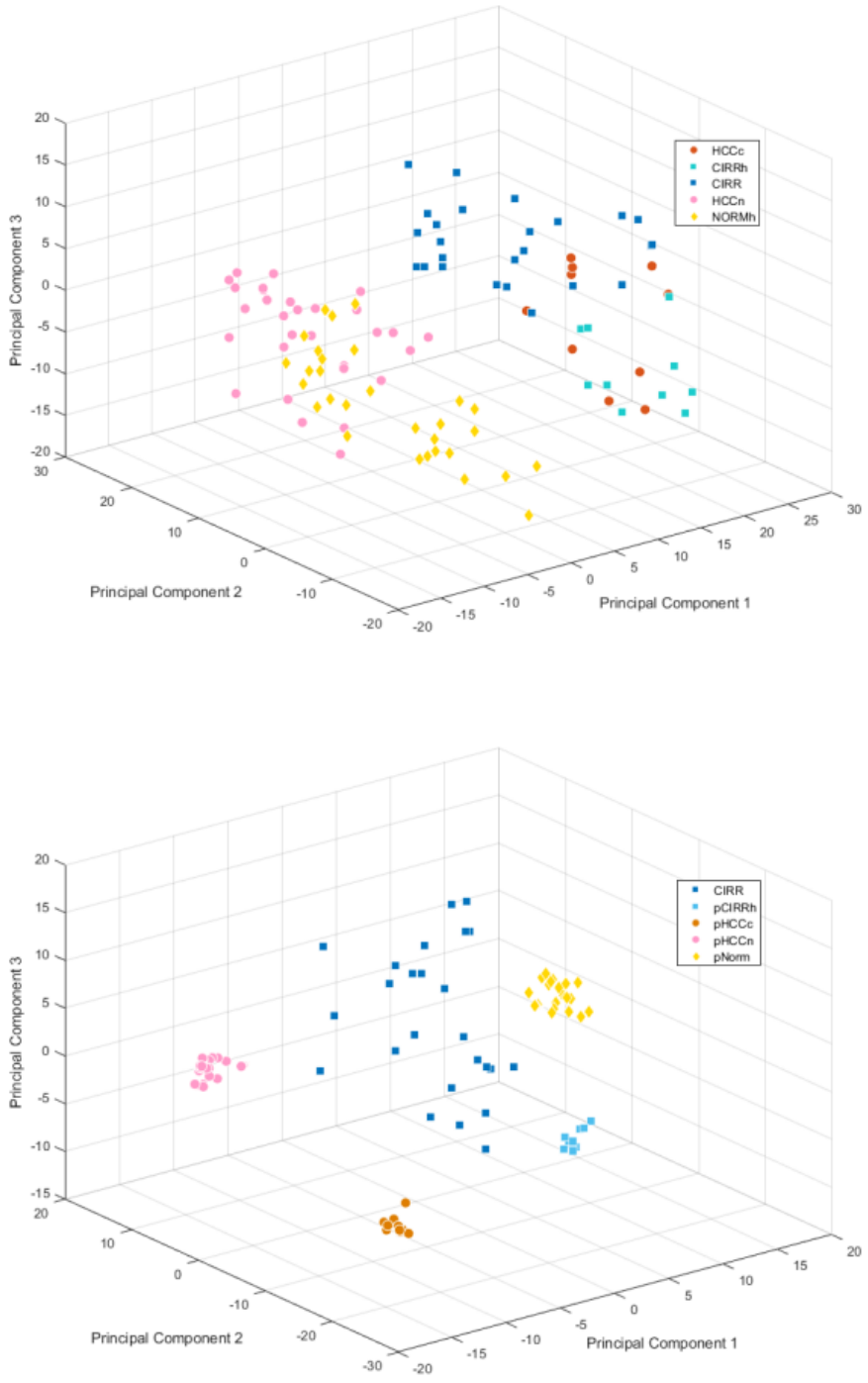


Figure 4.16: All samples (analyzed by GC-MS) expressed by the first three principal components based on PCA before and after purification (lower panel).

Table 4.8: Performance comparison before and after purification

		Before Purification			After Purification		
		No. Sig/all	AUC [95%CI]	ACC	No. Sig/all	AUC [95%CI]	ACC
GC-MS	P1	24/726	0.74[0.61, 0.85]	0.73	39/726	0.79[0.66, 0.92]	0.82
	P2	28/726	0.75[0.63, 0.84]	0.68	42/726	0.81[0.67, 0.93]	0.79
	P3	37/726	0.81[0.65, 0.90]	0.84	52/726	0.86[0.73, 0.95]	0.89
	P4	33/726	0.82[0.66, 0.91]	0.83	50/726	0.88[0.77, 0.97]	0.88
LC-MS (pos)	P1	104/2286	0.82[0.71, 0.90]	0.80	135/2286	0.84[0.74, 0.93]	0.86
	P2	97/2286	0.81[0.69, 0.89]	0.80	128/2286	0.85[0.76, 0.96]	0.86
	P3	127/2286	0.85[0.74, 0.92]	0.84	159/2286	0.89[0.79, 0.98]	0.90
	P4	132/2286	0.86[0.75, 0.93]	0.86	161/2286	0.89[0.80, 0.98]	0.92
LC-MS (neg)	P1	21/593	0.71[0.59, 0.82]	0.75	41/593	0.81[0.72, 0.92]	0.82
	P2	18/593	0.72[0.61, 0.83]	0.75	39/593	0.80[0.71, 0.92]	0.82
	P3	19/593	0.78[0.67, 0.88]	0.80	36/593	0.88[0.79, 0.99]	0.86
	P4	25/593	0.82[0.73, 0.92]	0.86	47/593	0.90[0.81, 0.99]	0.92

Deconvolution using purified profiles

Since the intra-group variances are small, we randomly select five profiles in each group as pure sources to deconvolute original measurements. The denoise deconvolution model estimated the mixture proportion of components from the 25 pure sources and sample-specific noise.

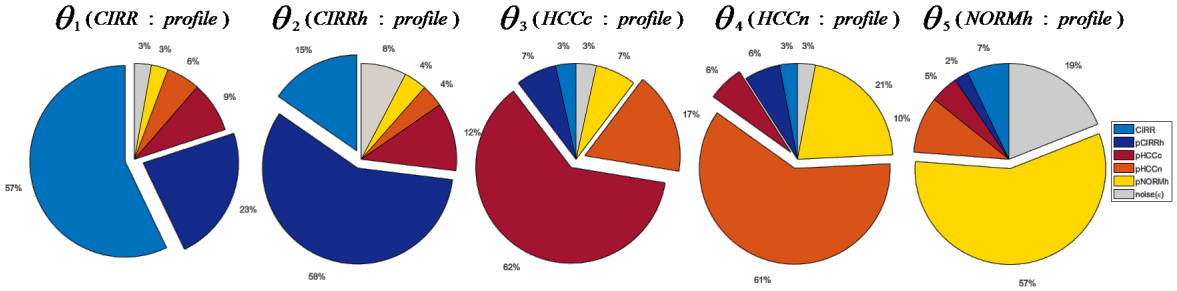


Figure 4.17: Mixture proportions estimated by denoise deconvolution model for multi-group samples (based on GC-MS). The predominant source correspond to the group labels.

Figure 4.17 shows, as examples, the mixture proportion θ of topics in multi-group samples.

The domain proportion helps allocate the group information correctly. We used this deconvolution method to assign the labels to all 105 samples. We obtained 95.2%, 97.1%, and 92.4% classification accuracy based on the GC-MS, LC-MS (positive) and LC-MS (negative) data, respectively.

In summary, we introduced a probabilistic purification model-based inference method to computationally address heterogeneity issue in liver tissues analyzed by LC-MS and GC-MS. The topic model gives a probabilistic explanation on the corpus of multi-group metabolomic profiles acquired by LC-MS or GC-MS. The purity of the samples is critical for disease characterization and biomarker discovery. We observed increased discrimination power between case and control groups after purification. By purifying the heterogeneous profiles, we obtained the underlying pure sources. These source profiles in return help deconvolute the original measured data, remove the unwanted noise, and successfully allocate the label information. We demonstrated that the improvements we observed in small sample size experiments can be pronounced when applied to larger sample size. However, we observed pitfalls such as the divergence of cancerous profiles and overfitting issue in deconvolution. Additional clinical information retrieved from the pathology report, cross validation of the deconvolution model, evaluation of candidate biomarkers by literature survey, will be performed in the future to further investigate the findings from this study.

Chapter 5

Applications to cancer biomarker discovery

The ultimate goal of developing the proposed methods of preprocessing and purification is to clear the obstacles in discovering candidate biomarkers from LC/GC-MS profiled biomolecules. In this chapter, we shift our focus towards the overall omic based applications in cancer biomarker discovery, including the disease of interest, biological hypotheses, experimental design, data acquisition and analysis, and consequentially the biomarker candidates we found.

5.1 Background

Hepatocellular carcinoma (HCC) is the third leading cause of cancer mortality worldwide with five-year relative survival rates less than 15%. [65–67] Most of the risk factors for HCC, including chronic infection with hepatitis B virus (HBV) or hepatitis C virus (HCV), lead to the development of liver cirrhosis, which is present in 80-90% of patients with HCC. [68] The malignant conversion of cirrhosis to HCC is often fatal in part because adequate biomarkers are not available for diagnosis during the progression stages of HCC. Survival rates of patients with HCC can be significantly improved if the diagnosis is made at earlier stages, when treatment is more effective. [67, 69] Alpha-Fetoprotein (AFP), the serologic biomarker for HCC in current use, is not effective for early diagnosis due to its low sensitivity. [70, 71] Therefore, more potent biomarkers for early stage HCC are needed.

5.2 Individual omic study for biomarker discovery

5.2.1 Glycomics

Glycosylation is one of the most common post-translational modifications of proteins. Altered patterns of glycosylation have been associated with various diseases, and many currently used cancer biomarkers, including AFP, are glycoproteins. [72, 73] The analysis of glycosylation is particularly relevant to liver pathology because of the major influence of this organ on the homeostasis of blood glycoproteins. Mass spectrometry is an essential tool for the analysis of glycosylation. As protein glycosylation can occur on multiple sites involving the attachment of different glycans to each site, analysis of glycoproteins requires site-specific elucidation of glycan heterogeneity. [74] This is further complicated by the different chemical properties between glycans and peptides, and analysis by mass spectrometry typically involves enrichment of glycoproteins or glycopeptides. [75] An effective alternative is to analyze glycans released from proteins and associate the glycomic changes with pathological conditions of interest. N-Glycans are of particular interest as their involvement in major biological processes, including cellcell interactions and intracellular signaling, has important implications in disease progression.8 Also, several enzymes that allow efficient release of this type of glycans have been made available. [76] Through appropriate analytical methods that yield broad coverage of the glycome, characterizing glycomic patterns in serum/plasma of patients with cancer has proven a promising strategy to discover biomarkers for early diagnosis of cancer. [5, 76, 77] In particular, mass spectrometry is an enabling technology for analysis of glycans in cancer biomarker discovery. [76] The use of matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) to identify N-glycan biomarkers for HCC has been widely applied and discussed. [78–81] With recent advances in mass spectrometry and separation methods, liquid chromatographymass spectrometry (LCMS) is capable of profiling hundreds of glycans including isomeric glycoforms. [82, 83] Mass spectrometry using electrospray ionization (ESI-MS) is especially well-suited to the LCMS based glycomic analysis. Higher sensitivity of LCESI-MS over MALDI-MS and LCMALDI-MS in detecting permethylated N-glycans derived from serum has been demonstrated in a recent study. [82] However, to date glycomic profiling using LCESI-MS has not been fully exploited for large-scale biomarker discovery studies, and there is still a lack of appropriate computational tools. [77]

The present study applies LCESI-MS based serum glycomics for HCC biomarker discovery in patients with liver cirrhosis. Workflow of the proposed glycomic analysis is shown in Figure 5.1. Sera were collected from patients recruited in Egypt and the U.S. We utilized two complementary platforms to perform global profiling and targeted quantitation of N-glycans and identified candidate biomarkers that distinguish HCC cases from cirrhotic controls. Global profiling was performed using a high-resolution mass spectrometer (LTQ Orbitrap Velos), while targeted quantitation was performed using a triple quadrupole (QqQ) mass spectrometer in multiple reaction monitoring (MRM) mode. [84] The integrative workflow

consisting of global profiling and targeted quantitation is widely applied in LCMS based proteomic studies but to our best knowledge has not yet been exploited in glycomics. This study revealed 26 N-glycans with statistically significant differences between HCC cases and cirrhotic controls through global profiling and 15 through targeted quantitation. Eleven of these candidate N-glycan biomarkers were identified by both quantitation approaches and match closely with the implications of important glycosyltransferases in cancer progression and metastasis. The results of this study illustrate the power of the integrative approach combining LCESI-MS based global profiling and targeted quantitation for a comprehensive serum glycomic analysis to investigate changes in N-glycan levels between HCC cases and patients with liver cirrhosis.

Preprocessing of LC-ESI-MS data

The data were analyzed using a preprocessing pipeline consisting of in-house-developed algorithms and open-source software tools. The preprocessing steps include deisotoping of mass spectra, peak detection, peak alignment, and normalization. We performed the deisotoping of mass spectra using DeconTools (v1.0.4672, October 16, 2012), [47] where the monoisotopic mass and charge state were deduced. DeconTools allows us to specify an appropriate average residue composition for the calculation of isotopic distribution. The average composition for the monosaccharides ($C_{10}H_{18}N_{0.43}O_5S_0$) was determined on the basis of the permethylated N-glycans commonly found in our previous studies. After the deisotoping step, peak detection was performed using an in-house-developed algorithm. Briefly, deisotoped ions with the same molecular weight (with 10 ppm tolerance) were linked along scans to generate a chromatographic trace. Low-quality traces were screened out according to user-defined criterions (i.e., minimum scans of 20 to define a peak, minimum summed intensity of 100,000, minimum density of 0.3 for valid scans in a trace, and allowable missing values of 35 between adjacent scans). Missing values in the remaining traces were interpolated using corresponding extracted ion chromatograms from raw data. The interpolated trace was further processed through successive convolution with a Savitzky-Golay smoothing filter (order of 5 and half of trace length as the window width) and a first-order derivative of a Gaussian kernel (window width of 30 scans, standard deviation of 3) to identify the position and boundary of the chromatographic peak at zero-crossing and enclosing local extrema, respectively. A detected peak was characterized by the following properties: monoisotopic mass, charge state, intensity (area under curve within boundary), and retention time. A normalization step was then applied to ensure that the summed intensity of detected peaks was identical in all of the LCESI-MS runs from the same batch. Peaks detected in multiple runs were aligned and matched using the simultaneous multiple alignment (SIMA, version of 2010) model [85] with the following parameters: -R 50 -M 0.1. The resulting peak list of the LCESI-MS runs was further refined such that only the peaks detected in over half of the runs in either case or control group were retained. Finally, missing values owing to either peak detection or alignment were interpolated using their corresponding extracted ion chromatograms. The

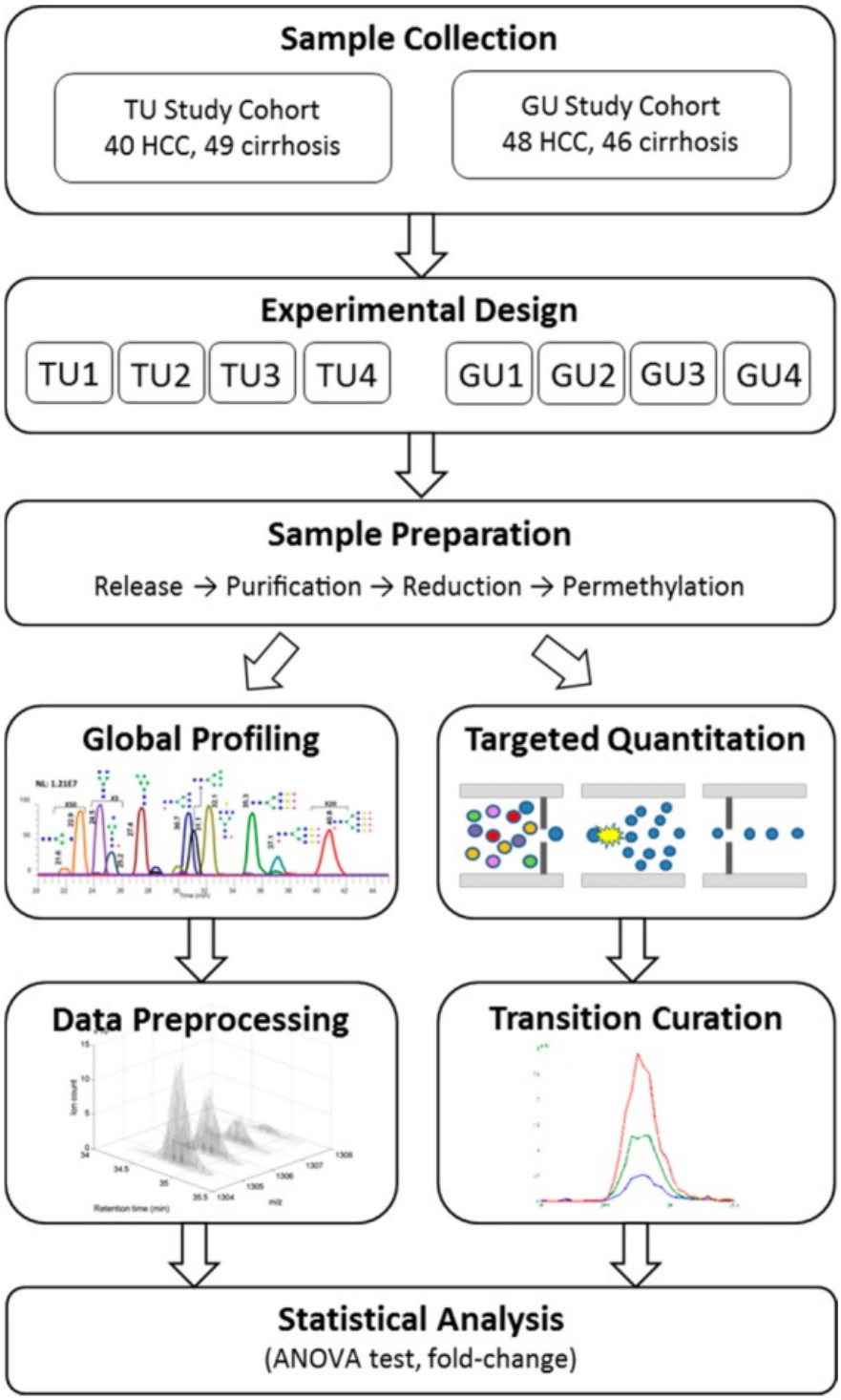


Figure 5.1: Workflow for the LC-ESI-MS analysis of N-glycans in sera from patients in two study cohorts (TU and GU).

preprocessing pipeline resulted in a consensus peak list of the LC-ESI-MS runs for subsequent analysis.

Statistical analysis

Following data preprocessing, the most relevant peaks with differential abundance between HCC cases and cirrhotic controls were selected using a two-way analysis of variance (ANOVA) model. Peaks from the four batches were matched upfront (m/z tolerance of 10 ppm and RT tolerance of 50 s), and peak intensity was modeled in terms of group effect (HCC versus cirrhosis), batch effect, interaction between group and batch, and random error associated with each sample. Specifically, the intensity for peak k from group i in batch j is modeled by $y_{ijk} = \mu + G_i + B_j + (G \times B)_{ij} + \epsilon_{ijk}$, where μ is the overall mean of the samples; G_is are the group effects ($\sum_i G_i = 0, i = 1, 2$); B_js are the batch effects ($\sum_j B_j = 0, j = 1, 2, 3, 4$); $(G \times B)_{ij}$s are the interaction between group and batch ($\sum_i (G \times B)_{ij} = \sum_j (G \times B)_{ij} = 0$); and ϵ_{ijk}s are the random errors from a zero-mean normal distribution. We calculated p-values with the null hypothesis that the group means within each batch are the same. Peaks with a p-value < 0.05 and having a consistent direction of fold change (FC) between groups in all four batches were selected as statistically significant.

Results and discussion

Global Profiling: LC-ESI-MS data were preprocessed in batch and peaks out of the expected RT range of glycans (1550 min) were excluded from subsequent analysis. In the TU cohort, 2609, 3130, 2903, and 2808 peaks were detected in batches TU1, TU2, TU3, and TU4, respectively. In the GU cohort, 2559, 2519, 2556, and 2922 peaks were detected in batches GU1, GU2, GU3, and GU4, respectively. Peak lists from the four batches in each cohort were then matched. This yielded 1628 and 1500 common peaks in the TU and GU cohorts, respectively, of which 262 and 254 peaks are associated with glycans. Prior to the statistical analysis, a logarithmic transformation was applied to ensure validity of the normal distribution assumption in the ANOVA model. We evaluated the quantitation variability based on coefficient of variation (CV) of peak intensities across samples. The ranges of the CVs in the TU cohort and GU cohort are $2\% \sim 40\%$ (with median at 8%) and $3\% \sim 58\%$ (with median at 10%), respectively. Using the two-way ANOVA model, we found 78 peaks in the TU cohort and 91 peaks in the GU cohort that are statistically significant (p-value < 0.05) and have consistent fold change across the four batches within each cohort. Putative glycan structures were assigned to the selected peaks by matching experimentally measured mass values with theoretical values of human serum N-glycans (with tolerance of 2 ppm) that were previously characterized in consideration of different charge states and adduct forms. Matched glycans are represented by the number of five monosaccharides: N-acetylglucosamine (GlcNAc), mannose, galactose, fucose, and N-acetylneuraminic acid (NeuNAc). This resulted in 18 significant N-glycans (11 up-regulated in HCC versus cirrhosis and 7 down-regulated) in

the TU cohort and 11 significant N-glycans (6 up-regulated and 5 down-regulated) in the GU cohort (Supplemental Table C.1-C.2). Three glycans were found significant in both cohorts. While [4-3-2-1-0] (up-regulated) and [5-3-0-0-0] (down-regulated) have the same fold change direction in both cohorts, [4-3-0-1-0] is up-regulated in the GU cohort and down-regulated in the TU cohort.

Targeted Quantitation: Manual curation was performed to eliminate channels with unfavorable chromatographic profiles or significant noise and to determine appropriate RT windows for quantitation. Owing to the unit resolution in Q1 and Q3, interference may appear across channels with close m/z values in their transitions. The observed elution order of N-glycans on the Orbitrap system was used to elucidate some ambiguous cases in the MRM analysis. Among the 213 transition channels, 65 channels representing 82 potential isomeric peaks of 52 N-glycans were detected consistently and quantitated for subsequent analysis. As in the global profiling analysis, peak intensities were log-transformed prior to the statistical analysis. A normalization step was also applied to ensure that the mean of the log-transformed peak intensities is identical in all the LC-ESI-MRM-MS runs from the same batch. The ranges of the CVs are 116% for both the TU and GU cohorts with median at 3% and 4%, respectively. Using the two-way ANOVA model, we selected significant N-glycans (p-value < 0.05) and those with consistent fold changes across the four batches in each cohort. We identified 11 significant glycans (7 up-regulated and 4 down-regulated) in the TU cohort and 5 significant glycans (4 up-regulated and 1 down-regulated) in the GU cohort. Consistent alteration (decreased level) was observed for the glycan [5-3-1-1-1] in both cohorts. Most of the significant glycans were also identified by the global profiling analysis, i.e., [4-3-1-0-0], [4-3-1-1-0], and [4-3-2-0-0] in the GU cohort and [5-3-0-0-0], [5-3-1-0-0], [5-3-1-0-1], [5-3-3-0-2], [5-3-3-0-3], [6-3-4-0-2], [6-3-4-0-3], and [6-3-4-0-4] in the TU cohort. Their MRM quantitation results are shown in Figures 5.2.

In summary, we analyzed over 1500 peaks, of which over 250 are associated with glycans in global profiling. In the targeted quantitation, we monitored 82 putative isomeric peaks of 52 glycans by MRM with improved sensitivity and accuracy compared to the global profiling. Smaller CVs for quantitated glycans in MRM analysis (median CV at 4%) compared to global profiling (median CV at 10%) demonstrate the improved accuracy.

Through statistical analysis, we identified 26 and 15 significant N-glycans by global profiling and targeted quantitation, respectively. These represent 30 unique glycans, because 11 glycans overlapped between the two approaches, while the remaining 19 glycans were selected by only one of the two approaches. The latter is in part due to the stringent criterion we selected. For example, manual curation was performed before targeted quantitation to eliminate channels with unfavorable chromatographic profiles or significant noise and to keep only the reliable information. Also, only those that displayed consistent fold change direction (either up- or down-regulation) across all batches were considered.

Most of the significant N-glycans discovered in this study are cohort-specific. This might be owing to the difference in etiologic factors between the two cohorts. For example, the

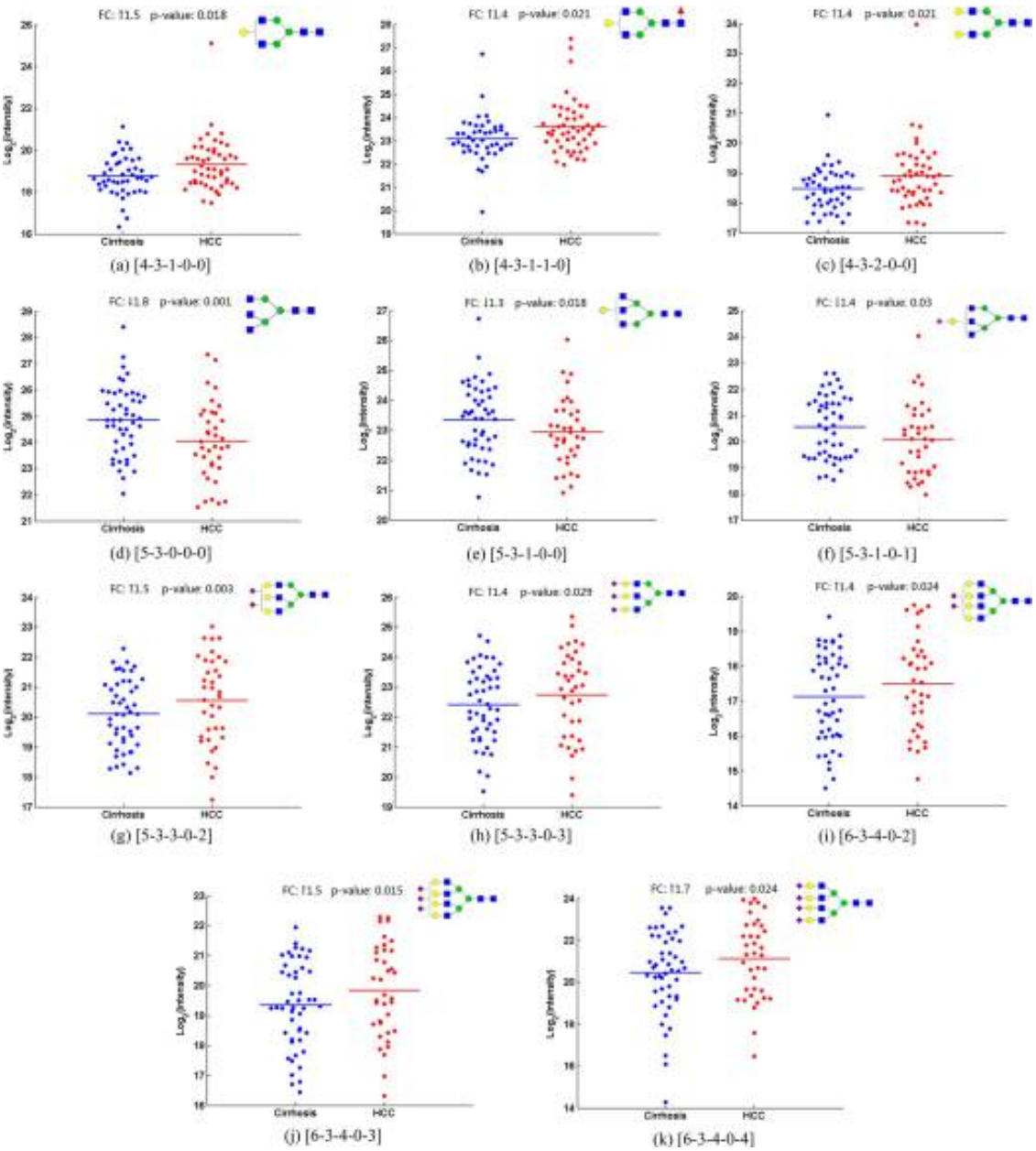


Figure 5.2: Quantitation results of 11 candidate N-glycan biomarkers in sera of HCC cases and cirrhotic controls by the MRM analysis. (a-c) Up-regulated biantennary glycans in the GU cohort. (d-f) Down-regulated β-1,6-GlcNAc branching glycans in the TU cohort. (g, h) Up-regulated β-1,6-GlcNAc branching glycans in the TU cohort. (i-k) Up-regulated tetra-antennary glycans in the TU cohort. FC=fold change. Blue square=GlcNAc, green circle=mannose, yellow circle=galactose, red triangle=fucose, purple diamond =NeuNAc.

Egyptian participants are all HCV positive and nearly all are HBV negative, whereas about half of the U.S. participants are HCV positive and about a third are HBV positive. Also, while the Egyptian participants are homogeneous Middle Eastern, the U.S. participants are approximately 56% Caucasian, 30% African American, 10% Asian, and 5% Hispanic. Moreover, about three-quarters of the Egyptian HCC cases are Stage I HCC, while Stage I HCC accounts for about half of the U.S. HCC cases. However, it should be noted that the sample size in this study is not large enough to draw solid conclusion in terms of etiology factors.

5.2.2 Proteomics

Similarly, we performed untargeted proteomic analysis to identify proteins showing statistically significant differences in sera from HCC cases and patients with liver cirrhosis. These proteins were further analyzed through targeted quantitation by MRM, which yielded more sensitive and accurate quantitation results. A high-resolution mass spectrometer (LTQ Orbitrap Velos) was used for untargeted proteomic analysis, while targeted quantitation was performed by MRM on a triple quadrupole (QqQ) mass spectrometer. We confirmed 21 candidates that showed significant changes in protein expression between HCC cases and cirrhotic controls in both cohorts. The results of this study demonstrate the power of combining untargeted and targeted quantitation methods for a comprehensive serum proteomic analysis, to investigate changes in protein levels between HCC cases and patients with liver cirrhosis. Figure 5.3 summarizes the workflow of this biomarker discovery study.

Depletion, digestion, and data acquisition

Sera were subjected to depletion using Agilent Plasma 7 Multiple Affinity Removal Spin Cartridge from Agilent Technologies (Santa Clara, CA). This cartridge depletes the seven most abundant human serum proteins, namely albumin, IgG, antitrypsin, IgA, transferrins, haptoglobin and fibrinogen. A 15-μl aliquot of serum was depleted as stated in the protocol provided by the manufacturer. The buffer of the depleted sample was exchanged into 50 mM ammonium bicarbonate (pH 8.0) using 3 kDa MWCO Amicon Ultra 0.5 mL centrifugal filters from Merck Millipore (Tullagreen, Carrigtwohill, Co. Cork). This buffer was used for tryptic digestion.

Prior to trypsin digestion, the protein concentration of depleted serum was determined by micro BCA protein assay following the protocols recommended by the vendor (Thermo Scientific/Pierce, Rockford, IL). A 20-μg aliquot of depleted serum proteins that corresponds to 0.4 μl of original serum was transferred to an Eppendorf tube, to which 100-μl of 50 mM ammonium bicarbonate was then added. Thermal denaturation was performed at 65 °C for 10 min. DTT and IAA solutions were prepared in 50 mM ammonium bicarbonate. Sample was reduced by adding a 1.25-μl aliquot of 200 mM DTT solution and incubated at 60 °C

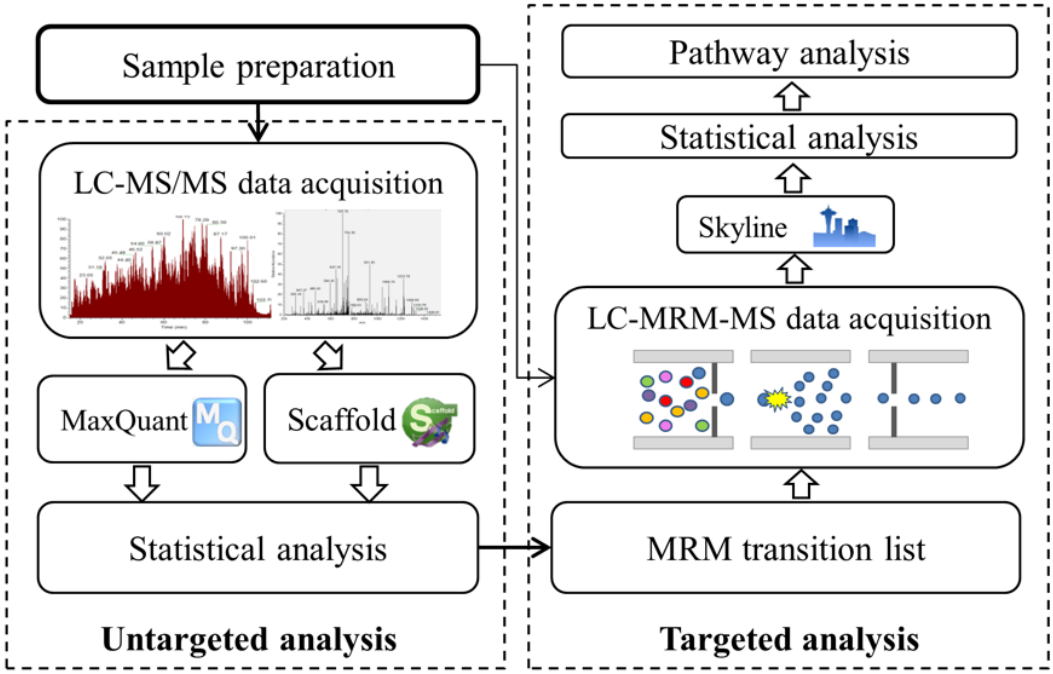


Figure 5.3: Workflow of the proposed biomarker discovery study involving untargeted and targeted analysis of sera.

for 45 min. The reduced proteins were then alkylated by adding of a 5-μl aliquot of 200 mM of IAA and incubated at 37.5 °C for 45 min. A second 1.25-μl aliquot of 200 mM DTT was added and followed by incubation at 37.5 °C for 30 min to consume excess IAA. A 0.8-μg aliquot of trypsin was added to the sample (enzyme/substrate ratio of 1:25 w/w), and then incubated at 37.5 °C overnight. This was followed by microwave-assisted digestion at 45 °C for 30 min at the power of 50 W. The enzymatic digestion was quenched by adding 0.5-μl neat FA to the samples. Then, the samples were speed-vacuum dried and re-suspended in 0.1% FA prior to LC-MS/MS and LC-MRM-MS analyses.

Analysis of sera was performed on a Dionex 3000 Ultimate nano-LC system (Dionex, Sunnyvale, CA) interfaced to an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, San Jose, CA), which is equipped with a nano-ESI source. LC-MS/MS analysis was performed on a tryptic digest corresponding to 1 μg of proteins which was derived from 0.2 μl of original serum considering the whole depletion and digestion process. The samples were online-purified using Acclaim PepMap100 C18 cartridge (3 μm, 100$\mathring{A}$, Dionex). The purified samples were then separated using Acclaim PepMap100 C18 capillary column (75 μm id $\times$ 150 mm, 2 μm, 100 $\mathring{A}$, Dionex). The separation of the digests was achieved at 350 nl/min flow rate, using the following gradient: 0-10 min sustaining 5% solvent B (98% ACN with 0.1% FA), 10-65 min ramping solvent B 5-20%, 65-90 min ramping solvent B 20-30%, 90-105 min ramping solvent B 30-50%, 105-106 min ramping solvent B 50-80%, 106-110 min maintaining solvent B at 80%, 110-111 min decreasing solvent B 80-5%, and 111-120 min maintaining solvent B at 5%. Solvent A was a 2% ACN aqueous solution containing 0.1% FA. The separation and scan time were set to 120 min.

The LTQ Orbitrap Velos mass spectrometer was operated with two scan events. The first scan event was a full FTMS scan of 380-2000 m/z with a mass resolution of 15,000 at m/z of 400. The second scan event was collision induced dissociation (CID) MS/MS of parent ions selected from the first scan event with an isolation width of 3.0 m/z. Normalized collision energy was set to 35% with an activation Q value of 0.250 and an activation time of 10 ms. The CID MS/MS was performed on the five most intense ions observed from the first MS scan event.

In the untargeted LC-MS/MS analysis, candidate protein biomarkers were identified using MaxQuant [59] and Scaffold [86]. We merged the results by both approaches and evaluated these proteins through targeted quantitation by MRM. For each targeted protein, one or two associated peptides were selected using the following rules [87]: (1) identified in the untargeted analysis with a Scaffold probability greater than 95%, (2) completely digested by trypsin, (3) 7-25 amino acid residues, (4) excluding the first 25 amino acids at the N-terminus of proteins, (5) excluding peptides with M, RP, KP and glycosylation site (NXS/T), (6) excluding peptides with ragged ends (tryptic peptides cleaved between R/K, K/R, R/R and K/K), and (7) fixed carbamidomethylation of Cysteine. Next, five transitions of selected peptides were determined using the following rules: (1) precursor ions with charge states of two or three, (2) y series of fragment ions greater than y3 with a charge state of one, (3) the five most intense fragment ions in the MS/MS spectra from untargeted analysis, and (4)

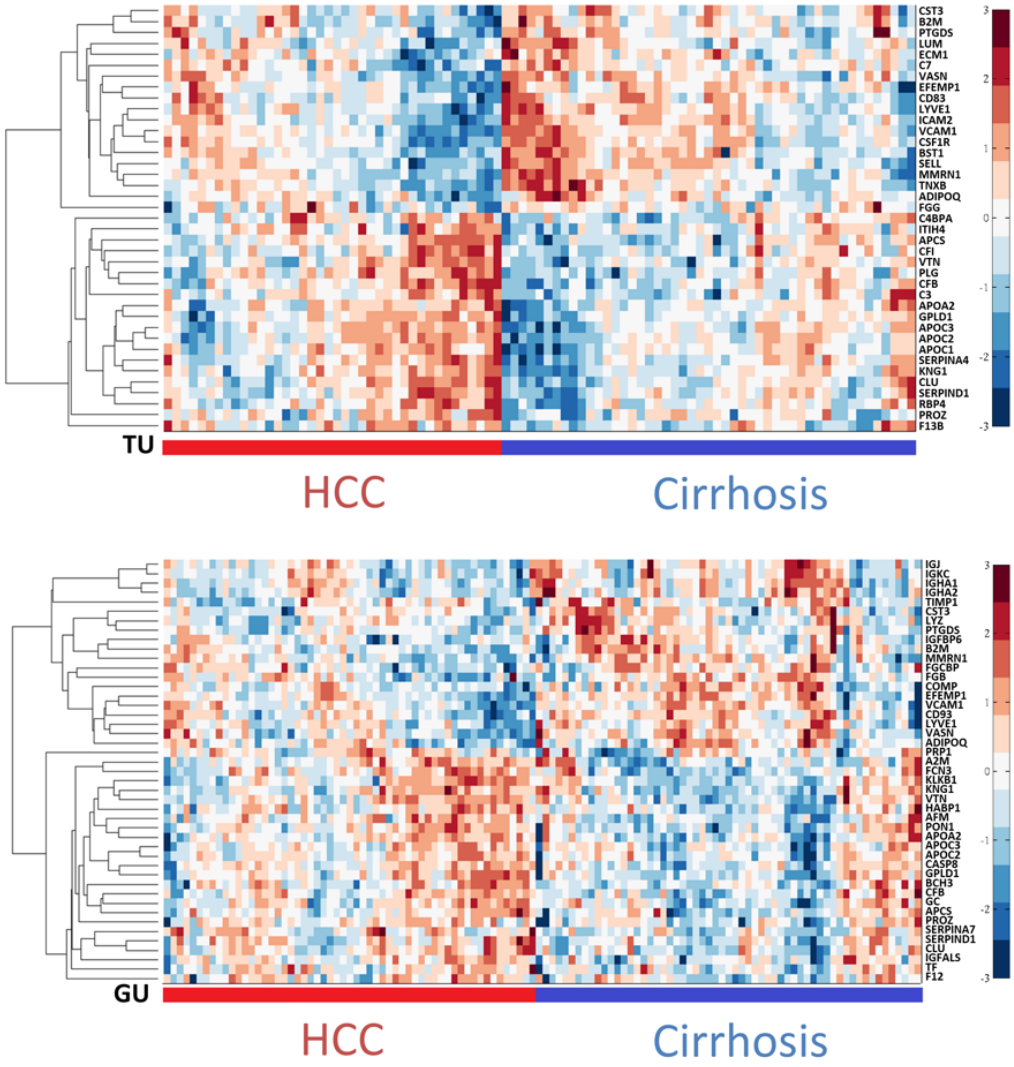


Figure 5.4: Heatmaps for significant proteins measured by MRM in the TU (top panel) and GU (bottom panel) cohorts.

m/z values of precursor and transition ions between 300 and 1500.

An RT segment was set to 12 min for each targeted peptide with its expected RT in the center based on the pooled sample analysis. The three most intense transition ions of each peptide were selected as the final transitions. Peptides and transitions were removed from transition list if any of them was not detected in the pooled sample analysis. In total, 101 targeted proteins with 187 peptides and 561 transitions were scheduled and subjected to the LC-MRM-MS experiments. With the abovementioned 12-min RT segment, a minimum 30 ms dwell time was assigned to each transition.

The LC-MRM-MS data were analyzed using Skyline [60] (version 2.5.0.6079). Peptide search results from Andromeda, i.e., msms.txt and mqpar.xml, were used to recognize the monitored transitions from LC-MRM-MS data. The Skyline determined the RT location and integration boundaries for each peptide in each run independently. By comparing the same peptide across runs, we adjusted the RT location and integration boundaries to exclude interfering regions. We selected the peak closest to the RT center of segment if multiple peaks were detected. Each proteins intensity was quantitated using the summation of intensities from its corresponding transitions. The difference between total area and background was assigned to quantify a transition. [87] Prior to the statistical analysis, the quantitated protein intensities were log-transformed and normalized by the summed intensity. The most relevant proteins with differential abundance between HCC cases and cirrhotic controls were selected using t-test, and the associated p-values were adjusted based on multiple testing correction (FDR < 0.05).

Results and discussion

Through targeted quantitation of the 101 candidate proteins by MRM, we found 61 proteins that are statistically significant (adjusted p-value < 0.05). These represent 39 and 43 significant proteins in the TU and GU cohorts, respectively, with 21 overlapping in both cohorts (Table 5.1). Heatmaps of hierarchical clustering results based on the significant proteins in each cohort are presented in Figure 5.4. Among the 21 proteins found significant in both cohorts, 11 are up-regulated in HCC versus cirrhosis, while 10 are down-regulated in HCC. While the reported AUC for each single biomarker is moderate, a panel selected by the SVM-RFE algorithm [88] from 21 proteins has led to a significant improvement over individual biomarkers including AFP. Specifically, the algorithm selected six proteins, namely, clusterin (CLU, P10909), vascular cell adhesion protein 1 (VCAM1, P19320), prostaglandin-H2 D-isomerase (PTGDS, P41222), phosphatidylinositol-glycan-specific phospholipase D (GPLD1, P80108), vasorin (VASN, Q6EMK4) and lymphatic vessel endothelial hyaluronic acid receptor 1 (LYVE1, Q9Y5Y7). Figure 5.5 depicts the ROC curves for AFP, a panel of six proteins, and the six proteins combined with AFP. We used a bootstrap method (1000 bootstrap replicates) to compute the 95% confidence interval (CI) of the area under each ROC curve. The six proteins in a panel achieved a higher AUC (95% CI [0.72, 0.88], with mean of 0.80) than

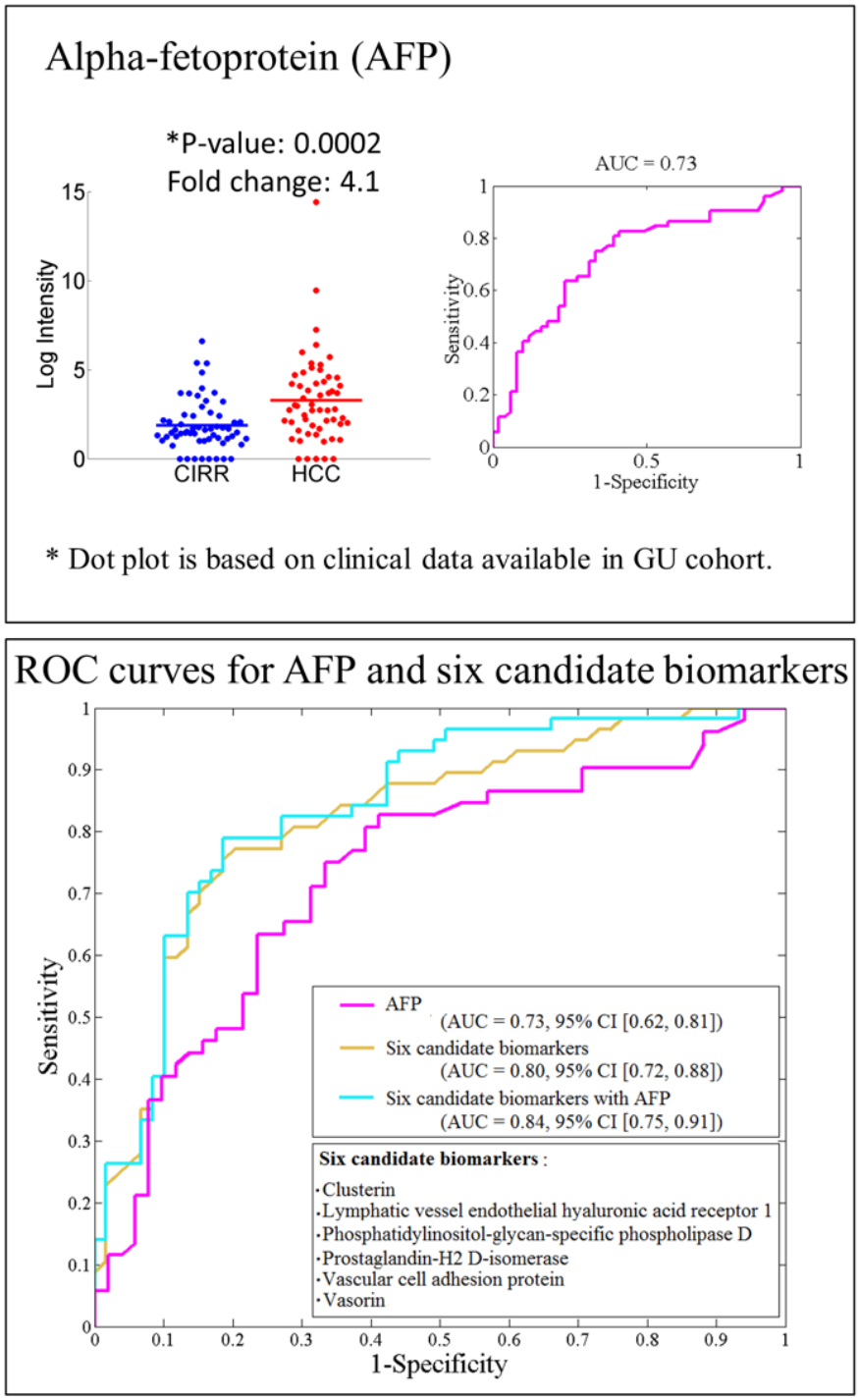


Figure 5.5: Top panel: Dot plot and ROC curve for AFP. Bottom panel: ROC curves for AFP, a panel of six proteins, and a panel of six proteins combined with AFP (mean AUC and 95% confidence interval).

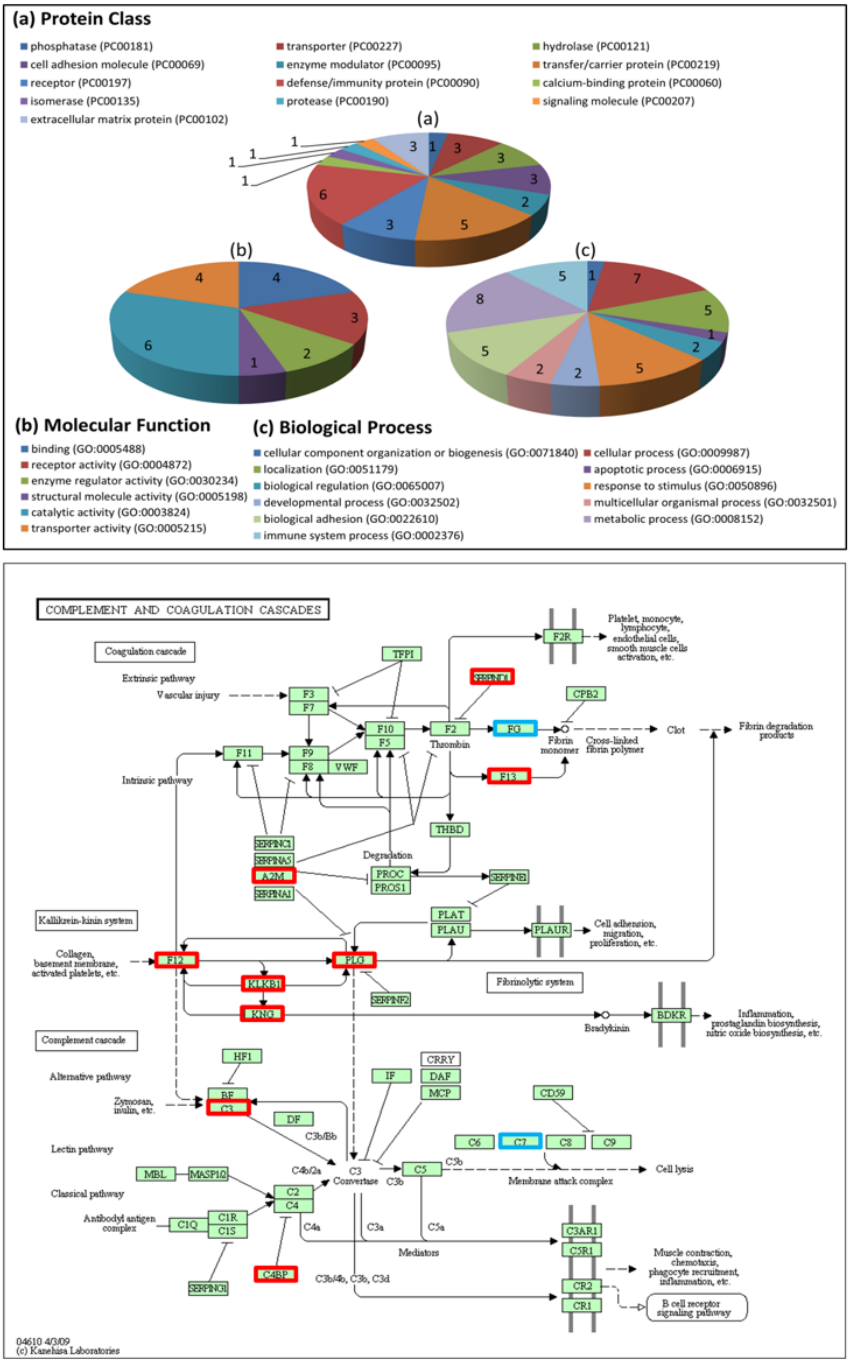


Figure 5.6: Top panel: Gene ontology analysis by PANTHER (Protein ANalysis THrough Evolutionary Relationships). Bottom panel: Complement and coagulation cascades pathway involving both up-regulated (red) and down regulated biomarkers (blue) in KEGG database.

Table 5.1: Protein candidate biomarkers identified by untargeted analysis and confirmed by targeted quantitation in both TU and GU cohorts. Fold change is the ratio of the mean intensity measured by MRM in the HCC group to the mean intensity in the cirrhotic group.

Protein Name	UniProt ID	TU cohort		GU cohort	
		Adjusted p-value	Fold Change	Adjusted p-value	Fold Change
Apolipoprotein A-II	P02652	0.025	↑1.4	0.006	↑1.3
Apolipoprotein C-II	P02655	0.028	↑1.9	0.017	↑1.6
Apolipoprotein C-III	P02656	0.032	↑1.8	0.045	↑1.6
Clusterin	P10909	0.014	↑1.3	0.002	↑1.4
Complement factor B	P00751	0.008	↑1.4	0.024	↑1.1
Heparin cofactor 2	P05546	0.025	↑1.5	0.011	↑1.4
Kininogen-1	P01042	0.014	↑1.3	0.023	↑1.2
Phosphatidylinositol-glycan-specific phospholipase D	P80108	0.019	↑1.6	0.004	↑1.5
Serum amyloid P-component	P02743	0.005	↑1.6	0.034	↑1.6
Vitamin K-dependent protein Z	P22891	0.016	↑1.4	0.012	↑1.3
Vitronectin	P04004	0.016	↑1.4	0.006	↑1.2
Adiponectin	Q15848	0.021	↓1.7	0.035	↓1.5
Beta-2-microglobulin	P61769	0.025	↓1.3	0.042	↓1.7
Complement component C1q receptor	Q9NPY3	0.040	↓1.4	0.011	↓1.3
Cystatin-C	P01034	0.037	↓1.2	0.023	↓1.3
EGF-containing fibulin-like extracellular matrix protein 1	Q12805	0.019	↓1.3	0.008	↓1.4
Lymphatic vessel endothelial hyaluronic acid receptor 1	Q9Y5Y7	0.019	↓1.5	0.048	↓1.4
Multimerin-1	Q13201	0.019	↓1.5	0.017	↓1.5
Prostaglandin-H2 D-isomerase	P41222	0.027	↓1.3	0.003	↓1.9
Vascular cell adhesion protein 1	P19320	0.006	↓1.5	0.048	↓1.3
Vasorin	Q6EMK4	0.031	↓1.2	0.003	↓1.2

AFP alone (95% CI [0.62, 0.81], with mean of 0.73). While the mean AUC was increased to 0.84 when the six proteins were combined with AFP in a panel, the addition of AFP does not improve the performance since the 95% CI [0.75, 0.91] overlaps substantially with the ROC based on the six proteins alone. An SVM classifier trained to minimize the misclassification rate by combining these markers and evaluated through cross-validation yielded higher sensitivity and specificity (0.75 and 0.77) compared with the performance of AFP alone (0.7 and 0.62). This comparison was performed using the GU study cohort, because the clinical measurement of AFP for the cirrhotic controls in the TU cohort was not available.

We used PathwayLinker [89] and DAVID [90] (version 6.7) to further identify significant signaling pathways, where the UniProt IDs of the significant proteins confirmed by MRM and their interacting neighbors were considered. We used PathwayLinker to obtain the first neighbor interactors of the 286 proteins detected in this study, where three interaction databases, BioGrid [91], STRING [92], and HPRD [93], were considered. Using the detected

proteins and their interacting neighbors as a reference, we analyzed the proteins found significant in this study to determine relevant signaling pathways in KEGG [63] through the DAVID functional annotation tool. By mapping 34 up-regulated proteins from both cohorts against the 286 detected proteins and their interacting neighbors, we found complement and coagulation cascades as the most significantly enriched signaling pathway, as shown in Figure 5.6. This pathway involves 11 proteins that were up-regulated in this study (A2M, F12, F13B, SERPIND1, CFI, KLKB1, KNG1, PLG, CFB, C3, and C4BPA). When we analyzed the 27 down-regulated proteins from the two cohorts, we found antigen processing and presentation, as the most significant pathway. Among the 27 down-regulated proteins, three (FGB, FGG, and C7) are involved in complement and coagulation cascades pathway.

5.3 Integrative analysis: multi-omics study

5.3.1 Introduction

So far, characterizing the association single level of biomolecules such as glycans or proteins with cancer has proven to be a promising strategy to discover candidate biomarkers. [8, 9] Because these biomolecules are members of strongly intertwined biological pathways and are highly interactive with each other, integrative analysis offers a great opportunity to help interpret such interactions and to identify reliable biomarkers. In addition to glycomics and proteomics, we also used gas chromatography coupled with mass spectrometry (GC-MS) to analyze metabolites in blood [10]. We detected proteins, N-glycans, and metabolites significantly altered in hepatocellular carcinoma (HCC) cases compared to patients with liver cirrhosis using univariate statistical methods. However, multivariate statistical or machine learning methods are desirable to improve the ability to discriminate the cases from controls by taking advantage of the mutual information within the molecules detected by a single omic study as well as the combination of molecules from multiple omic studies. The integrative analysis will allow us to investigate if the synergy of the three omic studies leads to improved performance in distinguishing cases from controls compared to the a single omic study. As a pilot study, we obtained improvement in discriminating HCC cases from cirrhotic controls using a panel of proteins and N-glycans selected by integrating proteomic and glycomic datasets [94].

5.3.2 Multi-omic data preprocessing and integration

In this research, we consider three datasets we previously generated by proteomic, glycomic, and metabolomic analysis of blood samples from HCC cases and patients with liver cirrhosis to identify proteins, N-glycans, and metabolites that are significantly altered in HCC versus cirrhosis. The goal of this investigation is to evaluate the improvement in dis-

ease classification achieved by integrating the data from the three studies. Support vector machine-recursive feature elimination (SVM-RFE) is used to select an optimal set of features that leads to highly discriminant classifier. [88] This not only helps recognize relevant patterns in the feature space, but also reduces dimensionality to overcome the risk of overfitting. Through a 10-fold cross validation, we evaluated the classification performances of the features selected from each omic studies as well as the combined features. We observed that improved performances can be achieved through the integrative analysis compared to a single omic study.

Experimental design

The proposed integrative analysis is performed on LC-MS-based proteomic and glycomic datasets and GC-MS-based metabolomic dataset we acquired by analysis of blood samples from HCC cases and patients with liver cirrhosis recruited in Egypt and the U.S. [8–10]. The participants in Egypt and the U.S. were recruited through protocols approved by the Ethics Committee at Tanta University Hospital and the Institutional Review Board at Georgetown University, respectively. Specifically, adult patients were recruited from the outpatient clinics and inpatient wards of the Tanta University Hospital (TU cohort) in Tanta, Egypt and from the hepatology clinics at MedStar Georgetown University Hospital (GU cohort) in Washington, DC, USA. The TU cohort consists of a total of 89 subjects (40 HCC cases and 49 patients with liver cirrhosis), and the GU cohort comprises of 116 subjects (57 HCC cases and 59 patients with liver cirrhosis).

Figure 5.7 depicts the overall workflow of our experimental design. Briefly, targeted quantitative analysis of selected proteins and N-glycans in blood samples was performed by multiple reaction monitoring (MRM) using a Dionex 3000 Ultimate nano-LC system (Dionex Sunnyvale, CA) interfaced to TSQ Vantage mass spectrometer (Thermo Scientific, San Jose CA). The targets were selected from our previous LC-MS-based untargeted proteomic and glycomic analyses and by text mining. Also, metabolites selected from a previous untargeted study were subjected for a targeted analysis in blood samples by selected ion monitoring (SIM) using an Agilent 7890A GC interfaced to a single quadrupole Agilent 5975C MSD (Agilent Technologies, Santa Clara, CA). The datasets from these omic studies were analyzed using Skyline [60], GPA [56], and SIMAT [95], respectively. Results from univariate statistical analysis have been previously reported in [8–10]. In the following, we introduce how we integrate the three datasets for feature selection that lead to improved performance on disease classification.

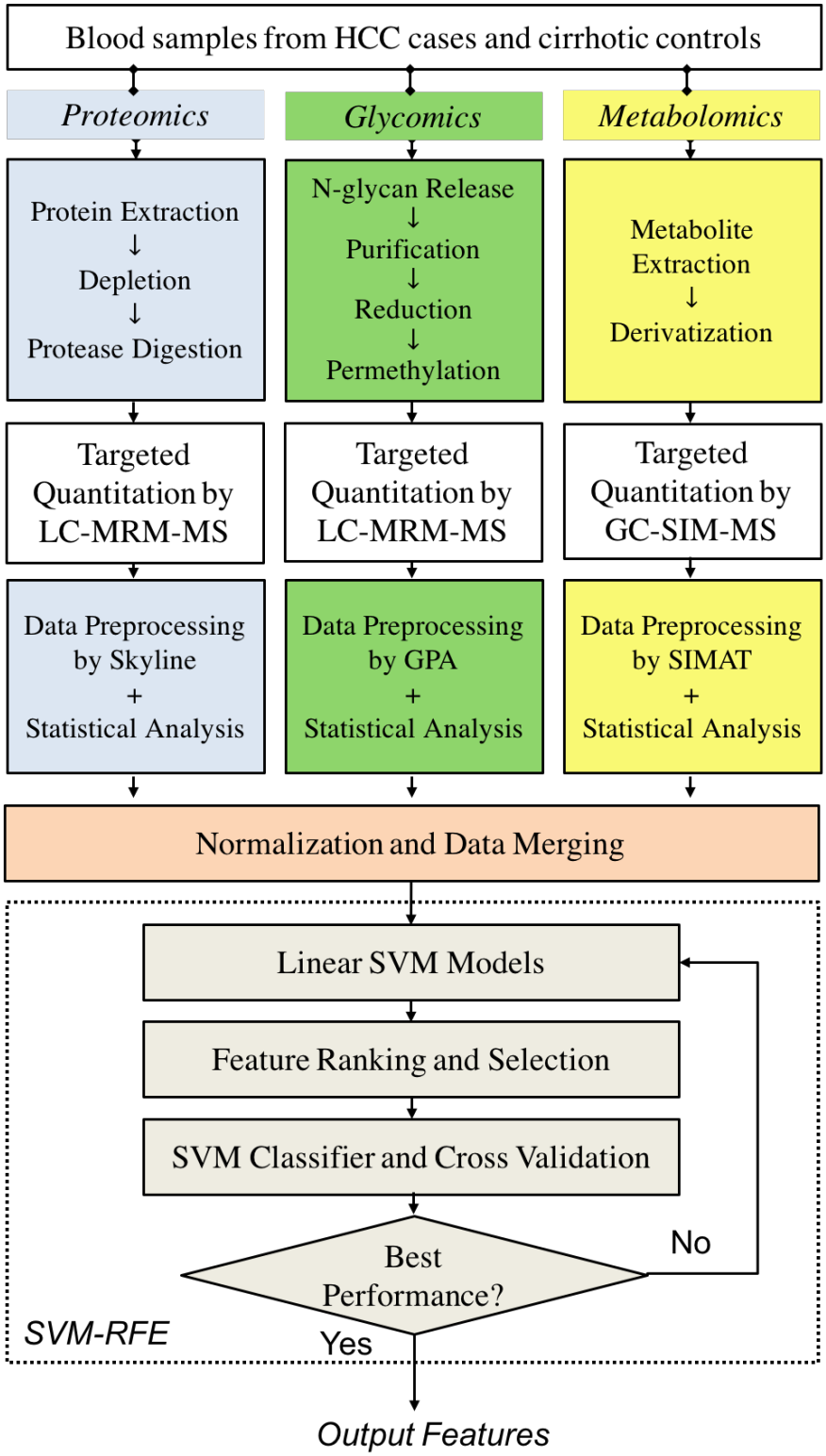


Figure 5.7: Workflow of integrative analysis of multi-omic data.

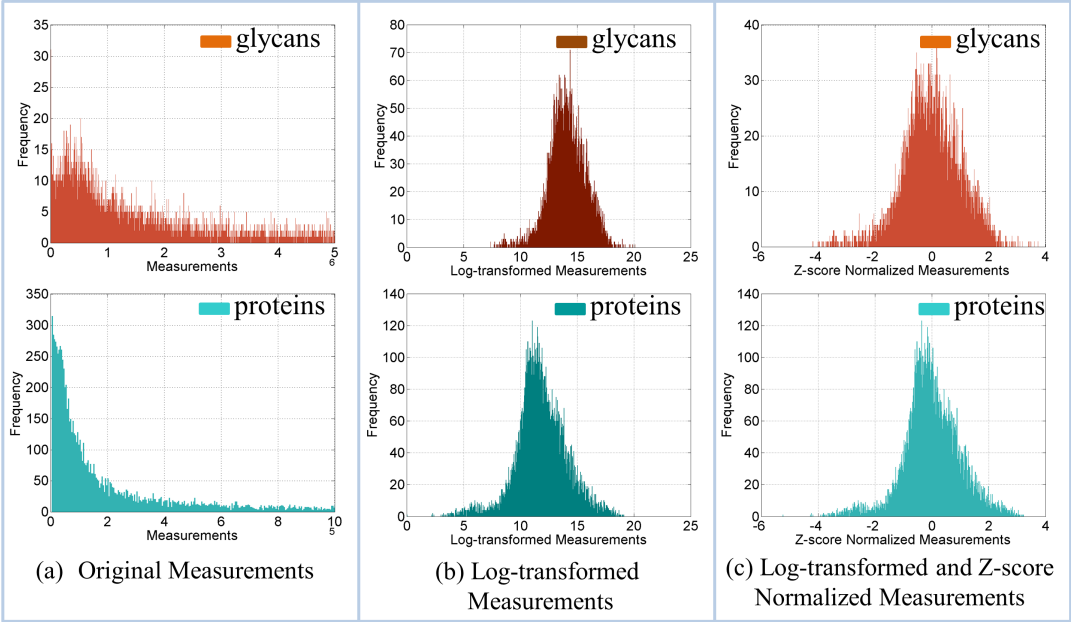


Figure 5.8: The distributions of raw glycomic (orange) and proteomic (cyan) datasets (a); log-transformed data (b); data after log-transformation and Z-score normalization.

Feature selection and classification

Linear SVMs were trained to classify samples in case and control groups using features from each of the three omic studies (proteomics, glycomics, and metabolomics) separately and by combining features from the three. Equation ?? presents the decision function in SVM model for an input sample x_t.

$$D(\mathbf{x}_t) = \mathbf{w} \cdot \mathbf{x}_t + b, \quad \text{where } \mathbf{w} = \sum_k \alpha_k y_k \mathbf{x}_k \quad \text{and } b = \langle y_k - \mathbf{w} \cdot \mathbf{x}_k \rangle. \tag{5.1}$$

The feature weight vector w determined by support vectors is used as feature ranking criterion by the recursive feature elimination (RFE) algorithm [6]. SVM-RFE eliminates redundant features iteratively and yields better and more compact feature subsets. The major steps include 1) training the SVM classifier; 2) ranking the features according to weight vector w of the learned SVM; 3) eliminating features with the smallest ranking criterion; 4) retraining SVM model with the remaining features; 5) estimating the performance of the model using cross-validation to check if the optimal subset is obtained. In this research, we applied SVM-

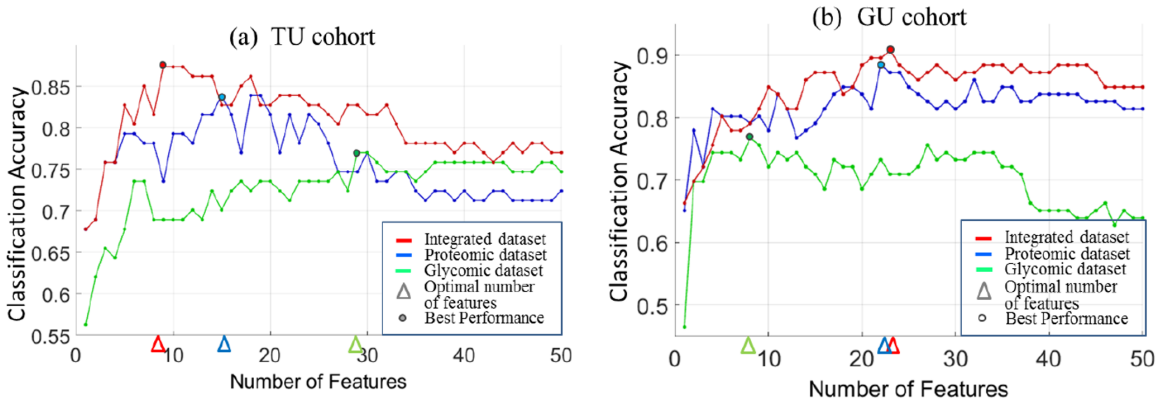


Figure 5.9: Classification accuracy at each iteration step for the top 50 features from glycomic (green), proteomic (blue), and integrated datasets (red) in the TU and GU cohorts. The optimal numbers of features (indicated by triangles) correspond to the best classification accuracy (indicated by circles).

RFE to select highly discriminative sets of proteins, N-glycans, and metabolites as well as features selected from an integrated set consisting of proteins, N-glycans, and metabolites.

5.3.3 Results and discussion

Integratibe analysis of proteins and N-glycans

Separate SVM-RFE models were trained for each of the three datasets. We started from the whole feature list in each dataset, and eliminated one feature in each iteration step till feature set was empty. 10-fold cross validation was employed to evaluate the average classification performance (i.e., accuracy, sensitivity, and specificity) at each step.

Figures. 5.9a and 5.9b depict the classification accuracy achieved at each iteration step for the top 50 features selected from the three datasets in the TU and GU cohorts, respectively. Also, the figures show the optimal number of features that leads to the best classification accuracy. We observed that, in most iteration steps, features selected from the integrated dataset yield higher accuracies compared to the same number of features selected from either the glycomic or proteomic dataset.

Receiver operating characteristics (ROC) curves were estimated by varying the SVM threshold parameter ($y_r = \hat{w} \cdot x - \hat{b}$). The 95% confidence intervals of area under the ROC (AUC) were calculated using bootstrap method with 1000 resampled replicates. Table 5.2 shows the disease classification performance with optimal subset of features in each dataset of the TU cohort. As shown in the table, SVM-RFE selected 29 out of 82 N-glycans and 15 out of 101

Table 5.2: Performance comparison based on the optimal number of features selected in the TU cohort

TU Cohort	***Glycomic***	***Proteomic***	***Integrated (P & G)***
Accuracy	0.77	0.84	**0.87**
Sensitivity	0.82	0.83	**0.90**
Specificity	0.75	0.84	**0.85**
AUC	0.87	**0.93**	0.92
(95% CI)	(0.78, 0.93)	**(0.83, 0.97)**	(0.81, 0.97)
Optimal Number of Features	29/82	15/101	**9**/183
Selected Features[b]	[25000] [43000][a]	P01024[a]	
	[53111][a] [34100][a]	P02743	
	[63402][a] [43202][a,c]	P02750	
	[53313] [53000][a]	P02753[a]	**P02743**
	[53323] [33101]	P02763	**P02763**
	[34110] [63403][a]	P03952	**P05160**
	[53311][c] [53311][c]	P04004[a]	**P06727**
	[43110][a] [53010]	P05160	**P0C0L4**
	[63413][a] [53302][a]	P06727	**P22891**[a]
	[53411] [34101]	P0C0L4	**P35858**
	[63423] [63404][a]	P13598[a]	**[43000]**[a]
	[53312] [29000]	P13796	**[26000]**
	[53101][a] [73514]	P22891[a]	
	[53201] [43202][a,c]	P27918	
	[2 10 000]	P35858	

[a] Significant (p value ≤ 0.05) in univariate statistical analysis
[b] N-glycans are characterized by GlcNAc, mannose, galactose, fucose, and NeuNAc, and proteins are indicated by Uniprot IDs
[c] Isomers with different retention times.

proteins as the optimal number of features. Among these, 13 glycans and 5 proteins were also selected as significantly altered in cases versus controls through univariate statistical test [8, 9]. Out of 183 integrated features, 7 proteins and 2 N-glycans in a panel were selected by SVM-RFE. The panel includes 2 that were also found significant in the univariate statistical analysis. The integrative analysis led to a significantly smaller number of features with a slight improvement on the disease classification accuracy compared to those selected by analysis of individual datasets. This phenomenon is observed consistently across the entire iteration steps, as illustrated in Figure 5.9a.

Similar results are obtained in the GU cohort (Table 5.3), in which SVM-RFE selected 18 proteins and 5 N-glycans in a panel yielded better performance than 22 proteins or the 8 glycans selected by analysis of individual datasets. Among the 23 features selected by the integrative analysis, four N-glycans and 10 proteins were also reported as significant by univariate statistical analysis. As shown in Figure 5.9b, the integrative analysis yielded improved performance compared to the analysis based on the individual datasets in the majority of the iteration steps. In both cohorts, we captured features with synergic contributions to the discrimination, which provide complementary information to univariate analysis. Although

Table 5.3: Performance comparison based on the optimal number of features selected in the GU cohort

GU Cohort	***Glycomic***	***Proteomic***	***Integrated (P & G)***
Accuracy	0.77	0.88	**0.91**
Sensitivity	0.79	0.86	**0.89**
Specificity	0.75	0.91	**0.93**
AUC (95% CI)	0.83 (0.71, 0.91)	0.95 (0.89, 0.98)	**0.96 (0.89, 0.99)**
Optimal Number of Features	8/82	22/101	**23**/183
Selected Features[b]	[43100][a] [53313] [53000][a] [43212] [53411] [53312] [53200] [63434]	O75015 O75636[a] P00748[a] P01023[a] P01877[a] P02741 P02766 P02771[a] P02790 P04278 P05155 P05452 P06727 P13796 P27169[a] P41222[a] P49747[a] P61626[a] P61769[a] Q15848[a] Q96KN2 Q9Y6R7[a]	**O75015 O75636[a] P01023[a] P01034[a] P01877[a] P02771[a] P04278 P05155 P05452[a] P08294 P13796 P41222[a] P61626[a] Q13201[a] Q15848[a] Q96KN2 [43100][a] [53313] [53000][a] [43200][a] [53411] [53200] [53111][a]**

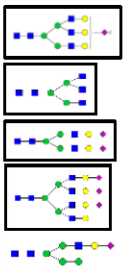
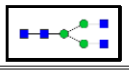
[a] Significant (p value ≤ 0.05) in univariate statistical analysis
[b] N-glycans are characterized by GlcNAc, mannose, galactose, fucose, and NeuNAc, and proteins are indicated by Uniprot IDs

we did not observe overlapping features between the optimal sets of features in the two cohorts, we were able to achieve AUCs greater than 0.73 when we trained SVMs based on the data the integrated panel learned from TU cohort and tested it on the GU cohort, and vice versa. In addition, we investigated the performance for each dataset by setting the feature size to five. We compared the performances of the best five features selected by SVM-RFE from each of the three datasets. While the integrative analysis outperformed the analysis based on individual dataset in TU cohort (Table 5.4), both the integrated features and the protein features led to similar performances in the GU cohort (Table 5.5).

Integratibe analysis of proteins, N-glycans, and metabolites

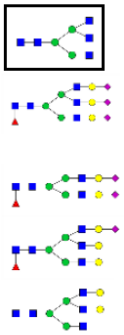
We present here the improvement in disease classification by including a dataset from a targeted analysis of 50 metabolites in blood samples. Thus, a total of 233 features (101 proteins, 82 N-glycans, and 50 metabolites) were considered for integrative analysis. The same normalization method was applied when merging features from the new dataset. Table 5.6 presents the performance of features selected by SVM-RFE from the metabolites only and

Table 5.4: Performance comparison based on the top ranking five features selected in the TU cohort

TU cohort	***Glycomic***	***Proteomic***	***Integrated (P & G))***
Accuracy	0.68	0.79	**0.83**
Sensitivity	0.71	0.79	**0.82**
Specificity	0.67	0.79	**0.85**
AUC (95% CI)	0.77 (0.65, 0.59)	0.88 (0.77, 0.94)	**0.89 (0.80, 0.95)**
Number of Selected Features	**5**/82	**5**/101	**5**/183
Selected Features		• Complement C3[a] • Serum amyloid P-component • Alpha-1-acid glycoprotein 1 • Coagulation factor XIII B chain • Apolipoprotein A-IV	• Serum amyloid P-component • Alpha-1-acid glycoprotein 1 • Coagulation factor XIII B chain • Apolipoprotein A-IV 

[a] Significant (p value ≤ 0.05) proteins in univariate statistical analysis. N-glycans that found significant (p value ≤ 0.05) in univariate statistical analysis are shown in boxes.

Table 5.5: Performance comparison based on the top ranking five features selected in the GU cohort

GU cohort	***Glycomic***	***Proteomic***	***Integrated (P & G)***
Accuracy	0.74	**0.80**	**0.80**
Sensitivity	0.74	**0.82**	**0.82**
Specificity	0.75	**0.79**	**0.79**
AUC (95% CI)	0.82 (0.70, 0.89)	0.85 (0.74, 0.92)	**0.87 (0.77, 0.93)**
Number of Selected Features	**5**/82	**5**/101	**5**/183
Selected Features		• Ficolin-3[a] • Alpha-2-macroglobulin[a] • Plasma protease C1 inhibitor • Tetranectin[a] • Prostaglandin-H 2 D-isomerase[a]	• Ficolin-3[a] • Alpha-2-macroglobulin[a] • Plasma protease C1 inhibitor • Sex hormone binding globulin 
Annotation	N-acetylglucosamine, mannose, galactose, fucose, NeuNAc		

[a] Significant (p value ≤ 0.05) proteins in univariate statistical analysis. N-glycans that found significant (p value ≤ 0.05) in univariate statistical analysis are shown in boxes.

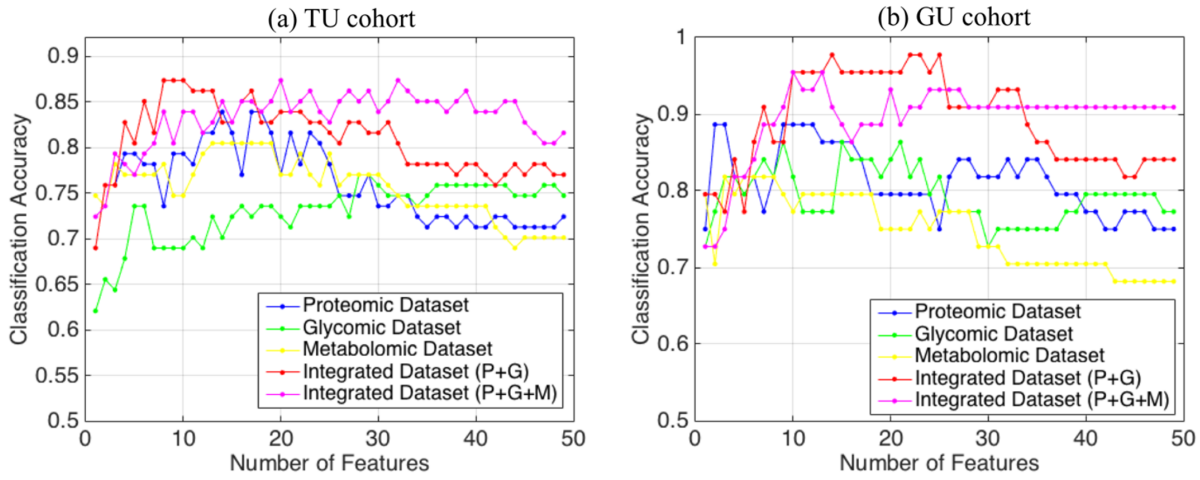


Figure 5.10: Classification accuracy at each iteration step for the top 50 features from glycomic (green), proteomic (blue), and integrated datasets (red) in the TU and GU cohorts. The optimal numbers of features (indicated by triangles) correspond to the best classification accuracy (indicated by circles).

the improvement achieved by combining the metabolites with proteins and glycans in the TU cohort. From the 50 metabolites, SVM-RFE selected 14 that showed better performance than those selected from the protein and N-glycan list presented in Table 5.2 on the same TU cohort representing 89 participants. A panel consisting of 10 proteins, 5 glycans, and 6 metabolites selected from the integrated dataset outperformed all other panels selected by SVM-RFE from single omic dataset or by combining proteomic and glycomic datasets. Figure 5.10a shows the classification accuracy at each iteration step for the top 50 features from three single datasets and two integrated datasets. We observe that the two integrated datasets (colored in red and magenta) have overall higher classification accuracies than any of the single omic datasets. Although the addition of metabolites to proteins and N-glycans did not improve the classification accuracy when relatively smaller number of features are selected, a more stable and discriminative performance is achieved as the feature size increases.

We performed integrative analysis of proteomic, glycomic, and metabolomics datasets acquired by analysis of blood samples from 44 subjects in the GU cohort. Since the number of overlapping samples in the three omic datasets is different from the number of overlapping proteomic and glycomic datasets reported in Tables 5.3 and 5.5, we repeated all multivariate analyses for appropriate comparison. Table 5.7 presents the performances of features selected from each of the three datasets as well as two integrated datasets. A panel of 10 features consisting of 4 proteins, 3 N-glycans, and 3 metabolites led to the best performance. Seven of these 10 features were also reported previously to have shown statistically significant changes in HCC vs. cirrhosis. [8–10] As illustrated in Figure 5.10b, features selected

Table 5.6: Performance comparison based on the optimal number of features selected in the TU cohort

TU Cohort	***Metabolomics***	***Integrated (P + G + M)***
Accuracy	0.86	**0.90**
Sensitivity	**0.91**	**0.91**
Specificity	0.84	**0.89**
AUC (95% CI)	0.93 (0.84, 0.97)	**0.99 (0.95, 0.99)**
Optimal # of Features	14/50	**21**/233
Selected Features[b]	• L-glutamic acid[a] • L-valine[a] • L-(+) lactic acid[a] • N-acetyl-5-hydroxyt ryptamine • L-threonine • Diglycerol • Urea • Arachidic acid • Trans-aconitic acid • L-proline • N, N-dimethyl-1 4-phenylenediamine • D-glucose L-serine • L-cystine	• P01024[a] • P01591 • P02743[a] • P02763 • P05160[a] • P06727 • P13591 • P13598[a] • P22891[a] • P35858 • [43000] • [53000][a] • [63423] • [28000] • [66012][a] • L-glutamic acid[a] • L-valine[a] • L-(+) lactic acid[a] • L-threonine • Urea • L-cystine

[a] Significant (p value ≤ 0.05) in univariate statistical analysis
[b] N-glycans are characterized by GlcNAc, mannose, galactose, fucose, and NeuNAc, and proteins are indicated by Uniprot IDs

from integrated datasets tend to have the best classification accuracy in most iterations. Integration of metabolites with proteins and N-glycans improves the classification accuracy as the number of features increases.

5.3.4 Summary

In this study, we investigated the benefit of an integrative analysis of proteomic, glycomic, and metabolomic datasets in improving our ability to distinguish HCC cases from patients with liver cirrhosis. Through SVM-RFE, a panel of features was selected from 101 proteins, 82 N-glycans, and 50 metabolites acquired by targeted analysis of blood samples using LC-MS and GC-MS. Complementary to univariate statistical methods, the integrative analysis utilizes mutual information among features to select a panel of features with improved ability

Table 5.7: Performance comparison based on the optimal number of features selected in the GU cohort (44 samples)

TU Cohort	***Proteomics***	***Glycomics***	***Metabolomics***	***Integrated (P + G)***	***Integrated (P+G+M)***
Accuracy	0.89	0.91	0.84	**0.98**	**0.98**
Sensitivity	0.94	0.87	0.85	0.95	**0.96**
Specificity	0.85	0.95	0.83	**0.99**	**0.99**
AUC (95% CI)	0.87 (0.72, 0.96)	0.97 (0.89, 0.99)	0.91 (0.77, 0.97)	**0.99 (0.95, 0.99)**	**0.99 (0.95, 0.99)**
Optimal Number of Features	3/101	10/82	4/50	15/183	10/233
Selected Features[b]	• O75636[a] • P00736 • P00751[a]	[43100][a] [53313] [53000][a] [34110] [53100] [53302] [53411][a] [43201] [63402] [53111][a]	• Ethanolamine • L-(+) lactic acid[a] • Oxalic acid • Putrescine	• O75636[a] • P01023[a] • P02774[a] • P04278 • P16070 • P41222[a] • P80108[a] • [53111][c] • [53313] • [34110] • [43110][a] • [43200][a] • [43201] • [73514] • [53111][a,c]	• O75636[a] • P01876[a] • P14151 • P41222[a] • [43100][a] • [53101][a] • [53111][a] • Malonic acid • Putrescine • Sorbose[a]

[a] Significant (p value ≤ 0.05) in univariate statistical analysis
[b] N-glycans are characterized by GlcNAc, mannose, galactose, fucose, and NeuNAc, and proteins are indicated by Uniprot IDs
[c] Isomers with different retention times.

to discriminate biologically distinct groups. In this study, we observe that features selected by merging the proteomic, glycomic, and metabolomic datasets lead to better disease classification accuracy compared to those selected from one or two of the three datasets. It should be emphasized that the improvement achieved by the integrative analysis was observed not only in using SVM-RFE, but also through other methods such as a sequential feature selection coupled with quadratic discriminant analysis. We believe that integration of multi-omic data by multivariate statistical or machine learning methods, combined with pathway-centric and network-based approaches, will help not only in identifying a panel of biomarkers that leads to improved diagnosis but also in gaining insight into the molecular mechanisms of cancer.

Chapter 6

Conclusion

While LC/GC-MS has provided us a high-throughput fashion in profiling multi-level biomolecules, appropriate LC/GC-MS data preprocessing pipelines are needed to ensure the detection of true differences between biological groups. This dissertation aims to address issues from several aspects of LC/GC-MS data analysis with consistent goal towards more reliable output of cancer biomarker discovery. As a conclusion, this chapter summarizes my original contributions of the research work and discusses potential future directions.

6.1 Summary of original contributions

In this thesis, we focus on LC/GC-MS based omics, aiming to address data processing, multi-omic integration, and sample heterogeneity issues for differential analysis of biomolecular data. Our final research goal is to develop computational models to facilitate cancer biomarker discovery studies. We have achieved improved performance of proposed methods in each of the research topics.

I. **LC-MS based glycomic data preprocessing**: In Chapter 3, we propose a workflow GPA that is designed to detect peaks from LC/MS glycomic datasets and annotate them using charge states and adduct information. We evaluate the performance of GPA using LC/MS datasets by analysis of permethylated N-glycan. A simulation study is carried out to demonstrate GPAs ability to handle ambiguous cases. This research enables the downstream statistical and integrative analysis in research topics II and III.

II. **Computational purification**: In Chapter 4 , we propose topic model-based purification methods IPM and SPM as the major research topic of this thesis. The results we obtained by analysis of the synthetic data demonstrated that both intensity-level and

scan-level purification models can accurately infer the mixture proportions and the underlying true cancerous sources with small average error ratios between estimation and ground truth. By applying the topic model-based purification to mass spectrometric data, we found more proteins and metabolites with significant changes between HCC cases and cirrhotic controls. Candidate biomarkers selected after purification yielded biologically meaningful pathway analysis results and improved disease discrimination power in terms of the area under ROC curve compared to the results found prior to purification. By purifying the heterogeneous profiles, we obtained the underlying pure sources. These source profiles in return help deconvolute the original measured data, remove the unwanted noise, and successfully allocate the label information. The performances of our models in estimating mixture proportion and retrieving underlying true cancer profile are evaluated through well-designed synthetic data. We observed that incorporation of scan-level features gives more accurate purification results by alleviating the loss in information caused as a result of integrating peak intensity values. Through GC-MS metabolomic and LC-MS proteomic datasets we acquired from tissues and blood samples, respectively, we showed the benefit of applying topic-model based purification of the data prior to statistical and pathway analyses.

III. **Multi-omic data integration**: Chapter 5.2 introduces an integrative analysis for biomarker discovery study. Through SVM-RFE, a panel of features was selected from 101 proteins, 82 N-glycans, and 50 metabolites acquired by targeted analysis of blood samples using LC-MS and GC-MS. We observed that features selected by merging the proteomic, glycomic, and metabolomic datasets lead to better disease classification accuracy compared to those selected from one or two of the three datasets. We would like to emphasize that the improvement achieved by the integrative analysis was observed not only in using SVM-RFE, but also through other methods such as a sequential feature selection coupled with quadratic discriminant analysis.

6.2 Future directions

There are several remaining topics that can be further explored. Future research include extension of current models and investigation of more potential applications. In terms of purification model, we demonstrated that the improvements we observed in small sample size experiments can be pronounced when applied to larger size of samples. However, we observed pitfalls such as the divergence of cancerous profiles, lack of biological evidence, and overfitting issue in deconvolution. Additional clinical information retrieved from the pathologist evaluation, cross validation of the deconvolution model and systems biological association of candidate biomarkers, will be prioritize to further investigate the findings from this research. Although this dissertation is focused on LC/GC-MS based omics data, some of the developed methodologies can be adjusted and used beyond the field of LC/GC-MS data analysis and cancer biomarker discovery. There is a broad interest in uncovering the rela-

tionship of human gut microbiome with the host's phenotypes, such as type-II diabetes [96], rheumatoid arthritis [97], obesity [98], and many other diseases. The challenges exist in how to infer the underlying intestinal flora distribution and estimate their mixture proportions. With additional efforts to translate and further characterize the flora deconvolution problem, the idea of probabilistic purification models developed in this dissertation may be useful in searching for the corresponding gut microorganisms.

6.3 Conclusion

Appropriate LC/GC-MS data preprocessing steps are in demand to identify the true and accurate differences between biological groups in LC/GC-MS based omic studies. Ion adduct clustering, heterogeneous profiles purification, and multi-omic data integration are listed among the most important yet challenging steps for biomarker discovery studies. In this dissertation, we investigated these problems from methodology development to practical applications through three research topics: 1) the development of glycan profile annotation algorithm; 2) the development of probabilistic purification models with utilization of different assumptions and features; 3) the applications in both individual and multiple omic based liver cancer biomarker discovery. Specifically, the major research work focuses on the proposed intensity-level, scan-level purification, and denoise deconvolution models, which have been evaluated and compared with its competitive models respectively using synthetic and real-world LC/GC-MS data sets from different omic studies. Experimental results show improved performance by the proposed models in identifying more discriminative and powerful marker candidates. This research greatly relieves the cost in dealing with sample heterogeneity issue, which is crucial yet not well investigated in LC/GC-MS based omic studies. Finally, several related research topics are proposed for future work, and some are already under way.

Bibliography

[1] Matthias Mann and Ole N Jensen. Proteomic analysis of post-translational modifications. *Nature biotechnology*, 21(3):255–261, 2003.

[2] Arun Sreekumar, Laila M Poisson, Thekkelnaycke M Rajendiran, Amjad P Khan, Qi Cao, Jindan Yu, Bharathi Laxman, Rohit Mehra, Robert J Lonigro, Yong Li, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457(7231):910–914, 2009.

[3] Emanuel F Petricoin, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, et al. Use of proteomic patterns in serum to identify ovarian cancer. *The lancet*, 359(9306):572–577, 2002.

[4] Gerald W Hart and Ronald J Copeland. Glycomics hits the big time. *Cell*, 143(5):672–676, 2010.

[5] Hyun Joo An, Scott R Kronewitter, Maria Lorna A de Leoz, and Carlito B Lebrilla. Glycomics and disease markers. *Current opinion in chemical biology*, 13(5):601–607, 2009.

[6] Adam M Hawkridge and David C Muddiman. Mass spectrometry–based biomarker discovery: toward a global proteome index of individuality. *Annual review of analytical chemistry (Palo Alto, Calif.)*, 2:265, 2009.

[7] Rasmus Madsen, Torbjörn Lundstedt, and Johan Trygg. Chemometrics in metabolomicsa review in human disease diagnosis. *Analytica Chimica Acta*, 659(1):23–33, 2010.

[8] Tsung-Heng Tsai, Minkun Wang, Cristina Di Poto, Yunli Hu, Shiyue Zhou, Yi Zhao, Rency S Varghese, Yue Luo, Mahlet G Tadesse, Dina Hazem Ziada, et al. LC–MS profiling of N-glycans derived from human serum samples for biomarker discovery in hepatocellular carcinoma. *Journal of Proteome Research*, 13(11):4859–4868, 2014.

[9] Tsung-Heng Tsai, Ehwang Song, Rui Zhu, Cristina Di Poto, Minkun Wang, Yue Luo, Rency S Varghese, Mahlet G Tadesse, Dina Hazem Ziada, Chirag S Desai, et al. LC-MS/MS-based serum proteomics for identification of candidate biomarkers for hepatocellular carcinoma. *Proteomics*, 2015.

[10] Mohammad R Nezami Ranjbar, Yue Luo, Cristina Di Poto, Rency S Varghese, Alessia Ferrarini, Chi Zhang, Naglaa I Sarhan, Hanan Soliman, Mahlet G Tadesse, Dina H Ziada, et al. GC-MS based plasma metabolomics for identification of candidate biomarkers for hepatocellular carcinoma in Egyptian cohort. *PloS ONE*, 10(6):e0127299, 2015.

[11] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10):883–892, 2012.

[12] Gerald Quon, Syed Haider, Amit G Deshwar, Ang Cui, Paul C Boutros, and Quaid Morris. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med*, 5(3):29, 2013.

[13] Montserrat Garcia-Closas, Per Hall, Heli Nevanlinna, Karen Pooley, Jonathan Morrison, Douglas A Richesson, Stig E Bojesen, Børge G Nordestgaard, Christen K Axelsson, Jose I Arias, et al. Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet*, 4(4):e1000054, 2008.

[14] Shai S Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L Bodian, Frank Staedtler, Nicholas M Perry, Trevor Hastie, Minnie M Sarwal, Mark M Davis, and Atul J Butte. Cell type–specific gene expression differences in complex tissues. *Nature Methods*, 7(4):287–289, 2010.

[15] Niya Wang, Ting Gong, Robert Clarke, Lulu Chen, Ie-Ming Shih, Zhen Zhang, Douglas A Levine, Jianhua Xuan, and Yue Wang. Undo: a bioconductor r package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics*, 31(1):137–139, 2015.

[16] Charles L Lawson and Richard J Hanson. *Solving least squares problems*, volume 161. SIAM, 1974.

[17] Wenlian Qiao, Gerald Quon, Elizabeth Csaszar, Mei Yu, Quaid Morris, and Peter W Zandstra. Pert: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput Biol*, 8(12):e1002838, 2012.

[18] Alexandra Posekany, Klaus Felsenstein, and Peter Sykacek. Biological assessment of robust noise models in microarray data analysis. *Bioinformatics*, 27(6):807–814, 2011.

[19] Eleftherios P Diamandis. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool opportunities and potential limitations. *Molecular & Cellular Proteomics*, 3(4):367–378, 2004.

[20] Christian H Ahrens, Erich Brunner, Ermir Qeli, Konrad Basler, and Ruedi Aebersold. Generating and navigating proteome maps using mass spectrometry. *Nature reviews Molecular cell biology*, 11(11):789–801, 2010.

[21] Matthias Gstaiger and Ruedi Aebersold. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nature Reviews Genetics*, 10(9):617–627, 2009.

[22] Tsung-Heng Tsai. Bayesian alignment model for analysis of lc-ms-based omic data. 2014.

[23] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[24] Joshua E Elias, Wilhelm Haas, Brendan K Faherty, and Steven P Gygi. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature methods*, 2(9):667–675, 2005.

[25] Bruno Domon and Ruedi Aebersold. Mass spectrometry and protein analysis. *science*, 312(5771):212–217, 2006.

[26] Thomas M Annesley. Ion suppression in mass spectrometry. *Clinical chemistry*, 49(7):1041–1044, 2003.

[27] Gerald Quon. *Probabilistic Models for the Analysis of Gene Expression Profiles*. PhD thesis, University of Toronto, 2012.

[28] Marek Kuczma. *An introduction to the theory of functional equations and inequalities: Cauchy's equation and Jensen's inequality*. Springer Science & Business Media, 2009.

[29] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[30] Stephen Wright and Jorge Nocedal. Numerical optimization. *Springer Science*, 35:67–68, 1999.

[31] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[32] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[33] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[34] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

[35] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[36] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[37] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[38] Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker, 1988.

[39] Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

[40] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022, 2003.

[41] Alessio Ceroni, Kai Maass, Hildegard Geyer, Rudolf Geyer, Anne Dell, and Stuart M Haslam. Glycoworkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *Journal of proteome research*, 7(4):1650–1659, 2008.

[42] David Goldberg, Mark Sutton-Smith, James Paulson, and Anne Dell. Automatic annotation of matrix-assisted laser desorption/ionization n-glycan spectra. *Proteomics*, 5(4):865–875, 2005.

[43] Evan Maxwell, Yan Tan, Yuxiang Tan, Han Hu, Gary Benson, Konstantin Aizikov, Shannon Conley, Gregory O Staples, Gordon W Slysz, Richard D Smith, et al. Glycresoft: a software package for automated recognition of glycans from lc/ms data. *PloS one*, 7(9):e45474, 2012.

[44] Chuan-Yih Yu, Anoop Mayampurath, Yunli Hu, Shiyue Zhou, Yehia Mechref, and Haixu Tang. Automated annotation and quantification of glycans using liquid chromatography–mass spectrometry. *Bioinformatics*, 29(13):1706–1707, 2013.

[45] Carsten Kuhl, Ralf Tautenhahn, Christoph Bottcher, Tony R Larson, and Steffen Neumann. Camera: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical chemistry*, 84(1):283–289, 2011.

[46] Colin A Smith, Elizabeth J Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006.

[47] Navdeep Jaitly, Anoop Mayampurath, Kyle Littlefield, Joshua N Adkins, Gordon A Anderson, and Richard D Smith. Decon2ls: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC bioinformatics*, 10(1):1, 2009.

[48] David M Horn, Roman A Zubarev, and Fred W McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 11(4):320–332, 2000.

[49] Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345, 1962.

[50] Stijn Marinus Van Dongen. Graph clustering by flow simulation. 2001.

[51] Erez Hartuv and Ron Shamir. A clustering algorithm based on graph connectivity. *Information processing letters*, 76(4):175–181, 2000.

[52] Jun Feng Xiao, Rency S Varghese, Bin Zhou, Mohammad R Nezami Ranjbar, Yi Zhao, Tsung-Heng Tsai, Cristina Di Poto, Jinlian Wang, David Goerlitz, Yue Luo, et al. LC–MS based serum metabolomics for identification of hepatocellular carcinoma biomarkers in Egyptian cohort. *Journal of Proteome Research*, 11(12):5914–5923, 2012.

[53] Michael L Metzker. Sequencing technologiesthe next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[54] Gerald Quon and Quaid Morris. ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, 25(21):2882–2889, 2009.

[55] Mohammad R Nezami Ranjbar, Mahlet G Tadesse, Yue Wang, and Habtom W Ressom. Bayesian normalization model for label-free quantitative analysis by lc-ms. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 12(4):914–927, 2015.

[56] Minkun Wang, Guoqiang Yu, Yehia Mechref, and Habtom W Ressom. Gpa: An algorithm for lc/ms based glycan profile annotation. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, pages 16–22. IEEE, 2013.

[57] Matthew E Monroe, Jason L Shaw, Don S Daly, Joshua N Adkins, and Richard D Smith. Masic: A software program for fast quantitation and flexible visualization of chromatographic profiles from detected lc–ms (/ms) features. *Computational biology and chemistry*, 32(3):215–217, 2008.

[58] Jean-Charles Nault and Augusto Villanueva. Intratumor molecular and phenotypic diversity in hepatocellular carcinoma. *Clinical Cancer Research*, 21(8):1786–1788, 2015.

[59] Jürgen Cox and Matthias Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, 2008.

[60] Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, and Michael J MacCoss. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, 2010.

[61] Jaeil Ahn, Ying Yuan, Giovanni Parmigiani, Milind B Suraokar, Lixia Diao, Ignacio I Wistuba, and Wenyi Wang. Demix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, 29(15):1865–1871, 2013.

[62] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2008.

[63] Minoru Kanehisa and Susumu Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[64] Keisuke Tachibana, Daisuke Yamasaki, Kenji Ishimoto, and Takefumi Doi. The role of PPARs in cancer. *PPAR Research*, 2008, 2008.

[65] Jacques Ferlay, Hai-Rim Shin, Freddie Bray, David Forman, Colin Mathers, and Donald Maxwell Parkin. Estimates of worldwide burden of cancer in 2008: Globocan 2008. *International journal of cancer*, 127(12):2893–2917, 2010.

[66] Alla Arzumanyan, Helena MGPV Reis, and Mark A Feitelson. Pathogenic mechanisms in hbv-and hcv-associated hepatocellular carcinoma. *Nature Reviews Cancer*, 13(2):123–135, 2013.

[67] Rebecca Siegel, Elizabeth Ward, Otis Brawley, and Ahmedin Jemal. Cancer statistics, 2011. *CA: a cancer journal for clinicians*, 61(4):212–236, 2011.

[68] Hashem B El-Serag and K Lenhard Rudolph. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology*, 132(7):2557–2576, 2007.

[69] Eldad S Bialecki and Adrian M Di Bisceglie. Diagnosis of hepatocellular carcinoma. *Hpb*, 7(1):26–34, 2005.

[70] Franco Trevisani, Paola Emanuela D'Intino, Antonio Maria Morselli-Labate, Giuseppe Mazzella, Esterita Accogli, Paolo Caraceni, Marco Domenicali, Stefania De Notariis, Enrico Roda, and Mauro Bernardi. Serum α-fetoprotein for diagnosis of hepatocellular carcinoma in patients with chronic liver disease: influence of hbsag and anti-hcv status. *Journal of hepatology*, 34(4):570–575, 2001.

[71] Samir Gupta, Stephen Bent, and Jeffrey Kohlwes. Test characteristics of α-fetoprotein for detecting hepatocellular carcinoma in patients with hepatitis ca systematic review and critical analysis. *Annals of internal medicine*, 139(1):46–50, 2003.

[72] Mark M Fuster and Jeffrey D Esko. The sweet and sour of cancer: glycans as novel therapeutic targets. *Nature Reviews Cancer*, 5(7):526–542, 2005.

[73] Bram Blomme, Christophe Van Steenkiste, Nico Callewaert, and Hans Van Vlierberghe. Alteration of protein glycosylation in liver diseases. *Journal of hepatology*, 50(3):592–603, 2009.

[74] Hyun Joo An, John W Froehlich, and Carlito B Lebrilla. Determination of glycosylation sites and site-specific heterogeneity in glycoproteins. *Current opinion in chemical biology*, 13(4):421–426, 2009.

[75] Joseph Zaia. Mass spectrometry and the emerging field of glycomics. *Chemistry & biology*, 15(9):881–892, 2008.

[76] Yehia Mechref, Yunli Hu, Aldo Garcia, and Ahmed Hussein. Identifying cancer biomarkers by mass spectrometry-based glycomics. *Electrophoresis*, 33(12):1755–1767, 2012.

[77] L Renee Ruhaak, Suzanne Miyamoto, and Carlito B Lebrilla. Developments in the identification of glycan biomarkers for the detection of cancer. *Molecular & Cellular Proteomics*, 12(4):846–855, 2013.

[78] Habtom W Ressom, Rency S Varghese, Lenka Goldman, Christopher A Loffredo, Mohamed Abdel-Hamid, Zuzana Kyselova, Yehia Mechref, MILOS NOVOTNY, and Radoslav Goldman. Analysis of maldi-tof mass spectrometry data for detection of glycan biomarkers. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 216. NIH Public Access, 2008.

[79] Zhiqun Tang, Rency S Varghese, Slavka Bekesova, Christopher A Loffredo, Mohamed Abdul Hamid, Zuzana Kyselova, Yehia Mechref, Milos V Novotny, Radoslav Goldman, and Habtom W Ressom. Identification of n-glycan serum markers associated with hepatocellular carcinoma from mass spectrometry data. *Journal of proteome research*, 9(1):104, 2010.

[80] Radoslav Goldman, Habtom W Ressom, Rency S Varghese, Lenka Goldman, Gregory Bascug, Christopher A Loffredo, Mohamed Abdel-Hamid, Iman Gouda, Sameera Ezzat,

Zuzana Kyselova, et al. Detection of hepatocellular carcinoma using glycomic analysis. *Clinical Cancer Research*, 15(5):1808–1813, 2009.

[81] Toshiya Kamiyama, Hideki Yokoo, Jun-Ichi Furukawa, Masaki Kurogochi, Tomoaki Togashi, Nobuaki Miura, Kazuaki Nakanishi, Hirofumi Kamachi, Tatsuhiko Kakisaka, Yosuke Tsuruga, et al. Identification of novel serum biomarkers of hepatocellular carcinoma using glycomic analysis. *Hepatology*, 57(6):2314–2325, 2013.

[82] Yunli Hu and Yehia Mechref. 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, 2012.

[83] Janie L Desantos-Garcia, Sarah I Khalil, Ahmed Hussein, Yunli Hu, and Yehia Mechref. Enhanced sensitivity of lc-ms analysis of permethylated n-glycans through online purification. *Electrophoresis*, 32(24):3516–3525, 2011.

[84] Paola Picotti, Oliver Rinner, Robert Stallmach, Franziska Dautel, Terry Farrah, Bruno Domon, Holger Wenschuh, and Ruedi Aebersold. High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. *Nature methods*, 7(1):43–46, 2010.

[85] Björn Voss, Michael Hanselmann, Bernhard Y Renard, Martin S Lindner, Ullrich Köthe, Marc Kirchner, and Fred A Hamprecht. Sima: simultaneous multiple alignment of lc/ms peak lists. *Bioinformatics*, 27(7):987–993, 2011.

[86] Brian C Searle. Scaffold: a bioinformatic tool for validating ms/ms-based proteomic studies. *Proteomics*, 10(6):1265–1269, 2010.

[87] Vinzenz Lange, Paola Picotti, Bruno Domon, and Ruedi Aebersold. Selected reaction monitoring for quantitative proteomics: a tutorial. *Molecular systems biology*, 4(1):222, 2008.

[88] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[89] Illés J Farkas, Ádám Szántó-Várnagy, and Tamás Korcsmáros. Linking proteins to signaling pathways for experiment design and evaluation. *PloS one*, 7(4):e36202, 2012.

[90] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009.

[91] Chris Stark, Bobby-Joe Breitkreutz, Andrew Chatr-Aryamontri, Lorrie Boucher, Rose Oughtred, Michael S Livstone, Julie Nixon, Kimberly Van Auken, Xiaodong Wang, Xiaoqi Shi, et al. The biogrid interaction database: 2011 update. *Nucleic acids research*, 39(suppl 1):D698–D704, 2011.

[92] Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, et al. String 8a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, 37(suppl 1):D412–D416, 2009.

[93] TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database2009 update. *Nucleic acids research*, 37(suppl 1):D767–D772, 2009.

[94] Minkun Wang, Guoqiang Yu, and Habtom W Ressom. Integrative analysis of lc-ms based glycomic and proteomic data. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 8185–8188. IEEE, 2015.

[95] Mohammad R Nezami Ranjbar, Cristina D Poto, Yue Wang, and Habtom W Ressom. Simat: Gc-sim-ms data analysis tool. *BMC bioinformatics*, 16(1):259, 2015.

[96] Junjie Qin, Yingrui Li, Zhiming Cai, Shenghui Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, Wenwei Zhang, Yuanlin Guan, Dongqian Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[97] Jose U Scher and Steven B Abramson. The microbiome and rheumatoid arthritis. *Nature Reviews Rheumatology*, 7(10):569–578, 2011.

[98] Husen Zhang, John K DiBaise, Andrea Zuccolo, Dave Kudrna, Michele Braidotti, Yeisoo Yu, Prathap Parameswaran, Michael D Crowell, Rod Wing, Bruce E Rittmann, et al. Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences*, 106(7):2365–2370, 2009.

Appendix A

Variational EM in IPM

We adopt a two-phase updating rules:

1. Treat γ' as consistent cancer origin for all profiles. Each profile is mixed from topic panel $\{\beta_1, ..., \beta_M, \gamma'\}$. Use the same variational EM framework as LDA to estimate α, κ', and infer $\{\theta_d\}, \gamma'$.

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = D \cdot \left[\Psi\left(\sum_{k=1}^{M+1} \alpha_k \right) - \Psi(\alpha_k) \right] + \sum_{d=1}^{D} \log(\theta_{d,k})$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{d,k}} = \frac{\alpha_k - 1}{\theta_{d,k}} + \sum_{n=1}^{N} \frac{t_{d,n}\beta_{n,k}}{\sum_{k'=1}^{M+1} \beta_{n,k'}\theta_{d,k'}}$$

$$\frac{\partial \mathcal{L}}{\partial \gamma'_n} = \frac{\kappa_n - 1}{\gamma'_n} + \sum_{d=1}^{D} \frac{t_{d,n}\theta_{n,M+1}}{\sum_{k'=1}^{M+1} \beta_{n,k'}\theta_{d,k'}}$$

2. Fix the cancer mixing proportion $\theta_{d,M+1}$, and average cancer origin γ' as prior to infer sample-specific pure cancer profile γ_d and contaminant mixing proportion $\{\theta_{d,k}\}, k = 1, ..., M$. Iteratively maximize the complete log likelihood function through conjugate gradient descent till convergence.

$$\frac{\partial \mathcal{L}}{\partial \theta_{d,k}} = \frac{\alpha_k - 1}{\theta_{d,k}} + \sum_{n=1}^{N} \frac{t_{d,n}\beta^{(d)}_{n,k}}{\sum_{k'=1}^{M+1} \beta^{(d)}_{n,k'}\theta_{d,k'}}, k \leq M$$

$$\frac{\partial \mathcal{L}}{\partial \gamma_{d,n}} = \frac{\kappa_d \gamma'_n - 1}{\gamma_{d,n}} + \sum_{d=1}^{D} \frac{t_{d,n}\theta_{n,M+1}}{\sum_{k'=1}^{M+1} \beta^{(d)}_{n,k'}\theta_{d,k'}}$$

$$\frac{\partial \mathcal{L}}{\partial \kappa_d} = \left[\left(\sum_{n=1}^{N} \gamma'_n \right) \Psi\left(\sum_{n=1}^{N} \kappa_d \gamma'_n \right) - \left(\sum_{n=1}^{N} \gamma'_n \Psi(\kappa_d \gamma'_n) \right) \right] + \sum_{n=1}^{N} \gamma'_d \log(\gamma_{d,n})$$

Appendix B

MCMC in SPM

In this research, we use a variety of Markov chain Monte Carlo (MCMC) methods in order to obtain the posterior distribution of parameters of SPM. Bayes rule was used to find the posterior of each variable $\boldsymbol{\Theta}_\iota$ given all other variables $\boldsymbol{\Theta}_{\setminus\iota}$ and observed data, $\{\mathbf{t}\}$:

$$\begin{aligned} P(\boldsymbol{\Theta}_\iota|\mathbf{t},\boldsymbol{\Theta}_{\setminus\iota}) &\propto P(\mathbf{t}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}_\iota|\boldsymbol{\Theta}_{\setminus\iota}) \\ &\propto P(\mathbf{t}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}_{\setminus\iota}|\boldsymbol{\Theta}_\iota)P(\boldsymbol{\Theta}_\iota) \end{aligned} \tag{B.1}$$

Here, we include the general closed form of the full conditionals based on the hierarchical model. We begin with the full conditionals for the first layer, starting with ion abundances:

$$\begin{aligned} P(\mathbf{x}|\mathbf{t},\boldsymbol{\Theta}_{\setminus x}) &\propto P(\mathbf{Y}|\boldsymbol{\Phi},\mathbf{x},\boldsymbol{\Delta},\boldsymbol{\Sigma}_e)P(\mathbf{x}|\tilde{\mathbf{x}},\boldsymbol{\Sigma}_x) \\ &\propto \prod_{i=1}^{m}\prod_{j=1}^{n}\Bigg(P(x_{i,j}|\tilde{x}_i,\sigma^2_{\eta_i}) \\ &\quad \times \prod_{t=1}^{T_{ij}} P\Big(\mathbf{t}_{i,j}(t)\Big|\boldsymbol{\phi}_{i,j},x_{i,j},\delta_{i,j}(t),\sigma^2_{e_i}\Big)\Bigg) \end{aligned} \tag{B.2}$$

Mean abundances, $\tilde{\mathbf{x}}$:

$$\begin{aligned} P(\tilde{\mathbf{x}}|\mathbf{t},\boldsymbol{\Theta}_{\setminus\tilde{x}}) &\propto P(\mathbf{t}|\boldsymbol{\Phi},\mathbf{x},\boldsymbol{\Delta},\boldsymbol{\Sigma}_e)P(\mathbf{x}|\tilde{\mathbf{x}},\boldsymbol{\Sigma}_x)P(\tilde{\mathbf{x}}|x_0,\sigma^2_{x_0}) \\ &\propto \prod_{i=1}^{m}\left(P(\tilde{x}_i|x_0,\sigma^2_{x_0})\prod_{j=1}^{n}P(x_{i,j}|\tilde{x}_i,\sigma^2_{x_i})\right) \end{aligned} \tag{B.3}$$

and the variance of the abundances, $\boldsymbol{\Sigma}_z$:

$$\begin{aligned} P(\boldsymbol{\Sigma}_x|\mathbf{t},\boldsymbol{\Theta}_{\setminus\sigma^2_x}) &\propto P(\mathbf{t}|\boldsymbol{\Phi},\mathbf{x},\boldsymbol{\Delta},\boldsymbol{\Sigma}_e)P(\mathbf{x}|\tilde{\mathbf{x}},\boldsymbol{\Sigma}_x)P(\boldsymbol{\Sigma}_x|a_x,b_x) \\ &\propto \prod_{i=1}^{m}\left(P(\sigma^2_{x_i}|a_x,b_x)\prod_{j=1}^{n}P(x_{i,j}|\tilde{x}_i,\sigma^2_{x_i})\right) \end{aligned} \tag{B.4}$$

Missing scan indicator:

$$P(\boldsymbol{\Delta}|\mathbf{t},\boldsymbol{\Theta}_{\setminus\Delta}) \propto P(\mathbf{t}|\boldsymbol{\Phi},\mathbf{x},\boldsymbol{\Delta},\boldsymbol{\Sigma}_e)P(\boldsymbol{\Delta}|\boldsymbol{\lambda}) \tag{B.5}$$
$$\propto \prod_{i=1}^{m}\prod_{j=1}^{n}\prod_{t=1}^{T_{i,j}} P\Big(y_{i,j}(t)\Big|\boldsymbol{\phi}_{i,j},x_{i,j},\delta_{i,j}(t),\sigma^2_{e_i}\Big)P\Big(\delta_{i,j}(t)\Big|\lambda_i\Big)$$

Missing rate of the scans:

$$\begin{aligned} P(\boldsymbol{\lambda}|\mathbf{t},\boldsymbol{\Theta}_{\setminus\lambda}) &\propto P(\mathbf{t}|\boldsymbol{\Phi},\mathbf{x},\boldsymbol{\Delta},\boldsymbol{\Sigma}_e)P(\boldsymbol{\lambda}|\boldsymbol{\Delta}) \\ &\propto P(\boldsymbol{\Delta}|\boldsymbol{\lambda})P(\boldsymbol{\lambda}|a_\lambda,b_\lambda) \\ &\propto \prod_{i=1}^{m}\left(P\Big(\lambda_i\Big|a_\lambda,b_\lambda\Big)\prod_{j=1}^{n}\prod_{t=1}^{T_{i,j}}P\Big(\delta_{i,j}(t)\Big|\lambda_i\Big)\right) \end{aligned} \tag{B.6}$$

Intensity error terms have a full conditional in the form of:

$$P(\boldsymbol{\Sigma}_e|\mathbf{t},\boldsymbol{\Theta}_{\setminus\sigma^2_e}) \propto P(\mathbf{t}|\boldsymbol{\Phi},\mathbf{x},\boldsymbol{\Delta},\boldsymbol{\Sigma}_e)P(\boldsymbol{\Sigma}_e|a_e,b_e) \tag{B.7}$$
$$\prod_{i=1}^{m}\left(P\Big(\sigma^2_{e_i}\Big|a_e,b_e\Big)\prod_{j=1}^{n}\prod_{t=1}^{T_{ij}}P\Big(y_{ij}(t)\Big|\boldsymbol{\phi}_{ij},x_{ij},\delta_{ij}(t),\sigma^2_{e_i}\Big)\right)$$

Peak shape function parameters, i.e. the lower layer:

$$\begin{aligned} P(\boldsymbol{\Phi}|\mathbf{t},&\boldsymbol{\Theta}_{\setminus\Phi}) \propto P(\mathbf{t}|\boldsymbol{\Theta})P(\boldsymbol{\Phi}|\boldsymbol{\Theta}_{\setminus\Phi}) \\ &\propto \prod_{i=1}^{m}\prod_{j=1}^{n}\Bigg(P\left(\boldsymbol{\phi}_{i,j}|\boldsymbol{\mu}_i,\mathbf{B},\mathbf{A}_i,\boldsymbol{\Sigma}_\epsilon\right) \\ &\quad\times\prod_{t=1}^{T_{i,j}}P\Big(y_{i,j}(t)\Big|\boldsymbol{\phi}_{i,j},x_{i,j},\delta_{i,j}(t),\sigma^2_{e_i}\Big)\Bigg) \end{aligned} \tag{B.8}$$

Error terms:

$$P(\boldsymbol{\Sigma}_\epsilon|\mathbf{t},\boldsymbol{\Theta}_{\setminus\Sigma_\epsilon}) \propto P(\boldsymbol{\Phi}|\boldsymbol{\mu},\mathbf{B},\mathcal{A},\boldsymbol{\Sigma}_\epsilon)P(\boldsymbol{\Sigma}_\epsilon|a_\epsilon,b_\epsilon) \tag{B.9}$$
$$\propto \prod_{k=1}^{r}\left(P(\sigma^2_{\epsilon_k}|a_\epsilon,b_\epsilon)\prod_{i=1}^{m}\prod_{j=1}^{n}P(\phi_{i,j,k}|\mu_{i,k},\boldsymbol{\beta}_k,\boldsymbol{\alpha}_{i,k},\sigma^2_{\epsilon_k})\right)$$

We assume the mean variable are independent for each ion and each peak shape parameter:

$$\begin{aligned} P(\boldsymbol{\mu}|\mathbf{t},\boldsymbol{\Theta}_{\setminus\mu}) &\propto P(\boldsymbol{\Phi}|\boldsymbol{\mu},\mathbf{B},\mathcal{A},\boldsymbol{\Sigma}_\epsilon)P(\boldsymbol{\mu}|\boldsymbol{\phi}_0,\boldsymbol{\Sigma}_\mu) \\ &\propto \prod_{i=1}^{m}\left(P(\boldsymbol{\mu}_i|\boldsymbol{\phi}_0,\boldsymbol{\Sigma}_\mu)\prod_{j=1}^{n}P(\boldsymbol{\phi}_{i,j}|\boldsymbol{\mu}_i,\mathbf{B},\mathbf{A}_i,\boldsymbol{\Sigma}_\epsilon)\right) \end{aligned} \tag{B.10}$$

Fixed effects coefficients were also assumed to be i.i.d. normally distributed:

$$\begin{aligned} P(\mathbf{B}|\mathbf{t}, \boldsymbol{\Theta}_{\setminus B}) &\propto P(\boldsymbol{\Phi}|\boldsymbol{\mu}, \mathbf{B}, \mathcal{A}, \boldsymbol{\Sigma}_\epsilon) P(\mathbf{B}|\boldsymbol{\Sigma}_\beta) \\ &\propto \prod_{i=1}^{m} \prod_{j=1}^{n} P(\boldsymbol{\phi}_{i,j}|\boldsymbol{\mu}_i, \mathbf{B}, \mathbf{A}_i, \boldsymbol{\Sigma}_\epsilon) \prod_{k=1}^{r} P(\boldsymbol{\beta}_k|\boldsymbol{\Sigma}_\beta) \end{aligned} \tag{B.11}$$

and their covariance matrices:

$$\begin{aligned} P(\boldsymbol{\Sigma}_\beta|\mathbf{t}, \boldsymbol{\Theta}_{\setminus \Sigma_\beta}) &\propto P(\mathbf{B}|\boldsymbol{\Sigma}_\beta) P(\boldsymbol{\Sigma}_\beta|a_\beta, b_\beta) \\ &\propto \prod_{k=1}^{r} P(\boldsymbol{\beta}_k|\boldsymbol{\Sigma}_\beta) P(\boldsymbol{\Sigma}_\beta|a_\beta, b_\beta) \end{aligned} \tag{B.12}$$

Random effects coefficients were considered to be independent for each peak shape parameter, but correlated across different ions:

$$\begin{aligned} P(\mathcal{A}|\mathbf{t}, \boldsymbol{\Theta}_{\setminus \mathcal{A}}) &\propto P(\boldsymbol{\Phi}|\boldsymbol{\mu}, \mathbf{B}, \mathcal{A}, \boldsymbol{\Sigma}_\epsilon) P(\mathcal{A}|\boldsymbol{\Sigma}_\alpha) \\ &\propto \prod_{i=1}^{m} \left(\prod_{j=1}^{n} P(\boldsymbol{\phi}_{i,j}|\boldsymbol{\mu}_i, \mathbf{B}, \mathbf{A}_i, \boldsymbol{\Sigma}_\epsilon) \prod_{k=1}^{r} P(\boldsymbol{\alpha}_{i,k}|\boldsymbol{\Sigma}_{\alpha_k}) \right) \end{aligned} \tag{B.13}$$

so for related covariance matrices we have:

$$\begin{aligned} P(\boldsymbol{\Sigma}_\alpha|\mathbf{t}, \boldsymbol{\Theta}_{\setminus \Sigma_\alpha}) &\propto P(\mathcal{A}|\boldsymbol{\Sigma}_\alpha) P(\boldsymbol{\Sigma}_\alpha|\boldsymbol{\Psi}_\alpha, \boldsymbol{\nu}) \\ &\propto \prod_{k=1}^{r} \left(P(\boldsymbol{\Sigma}_{\alpha_k}|\nu_k, \boldsymbol{\Psi}_{\alpha_k}) \prod_{i=1}^{m} P(\boldsymbol{\alpha}_{i,k}|\boldsymbol{\Sigma}_{\alpha_k}) \right) \end{aligned} \tag{B.14}$$

We split SPM into two components:

P1. Mixture model of underlying ion abundances (same as IPM).

P2. Scan-level feature modeling and inference.

In this point of view, we can adopt a two-step updating rule:

S1. MCMC sampling → peak shape model parameters in P2 (i.e.,ion abundance $\boldsymbol{x}_t$, $\boldsymbol{x}_\beta$, and shape function parameters ϕ). Gibbs sampling for variables (indicated by $\boldsymbol{\Theta}_g$) with known posterior density function

$$\boldsymbol{\Theta}_1^{(\wp+1)} \sim P(\boldsymbol{\Theta}_1|\mathbf{Y}, \boldsymbol{\Theta}_2^{(\wp)}, \boldsymbol{\Theta}_3^{(\wp)}, \ldots, \boldsymbol{\Theta}_{N_G}^{(\wp)}, \boldsymbol{\Upsilon}^{(\wp)})$$

$$\vdots$$

$$\boldsymbol{\Theta}_{N_G}^{(\wp+1)} \sim P(\boldsymbol{\Theta}_{N_G}|\mathbf{Y}, \boldsymbol{\Theta}_1^{(\wp+1)}, \ldots, \boldsymbol{\Theta}_{N_G-1}^{(\wp+1)}, \boldsymbol{\Upsilon}^{(\wp)})$$

Metropolis–Hastings for the rest $\boldsymbol{\Upsilon}_{mh}$ with proposal distribution $Q_\jmath()$, multivariate Gaussian.

$$\boldsymbol{\Upsilon}_\jmath^\star \sim Q_\jmath(\boldsymbol{\Upsilon}_\jmath|\mathbf{Y}, \boldsymbol{\Theta}_1^{(\wp)}, \ldots, \boldsymbol{\Theta}_{N_G}^{(\wp)}, \boldsymbol{\Upsilon}_{\setminus\jmath}^{(\wp)})$$

$$r_\jmath = \min\left(1, \frac{P(\boldsymbol{\Upsilon}_\jmath^\star) Q_\jmath(\boldsymbol{\Upsilon}_\jmath^{(\wp)}; \boldsymbol{\Upsilon}_\jmath^\star)}{P(\boldsymbol{\Upsilon}_\jmath^{(\wp)}) Q_\jmath(\boldsymbol{\Upsilon}_\jmath^\star; \boldsymbol{\Upsilon}_\jmath^{(\wp)})}\right)$$

$R \sim \mathcal{U}(0,1)$

$$\boldsymbol{\Upsilon}_\jmath^{(\wp+1)} = \begin{cases} \boldsymbol{\Upsilon}_\jmath^{(\wp+1)} & R \leq r_\jmath \\ \boldsymbol{\Upsilon}_\jmath^{(\wp)} & R > r_\jmath \end{cases}$$

S2. Treat $\boldsymbol{x}_t$, $\boldsymbol{x}_\beta$ as observed variables to implement the inference of P1 using the same VEM algorithm employed in IPM.

Appendix C

Supplemental Table

Glycans are characterized by the number of five monosaccharides: GlcNAc, mannose, galactose, fucose, and NeuNAc. The monosaccharide compositions were assigned through accurate mass matching (< 2 ppm). Retention times (RT, in min) in the first batch are reported. Adduct form is presented by charge state and number of protons replaced by ammonium: z(No. $[H]^+ \rightarrow [NH_4]^+$). Fold change is based on the comparison of HCC versus cirrhosis, where ↑ and ↓ denote up-regulation down-regulation, respectively.

Table C.1: Candidate N-glycan biomarkers identified in glycomic study

monosaccharide composition	cohort	quantitation approach	RT (min)	adduct: z (no. $[H]^+ \rightarrow [NH4]^+$)	p-value	fold change
[4-3-1-0-0]	GU	LC–ESI-MS	23.3	2 (0)	0.042	↑1.7
		MRM	26.0	2 (0)	0.018	↑1.5
[4-3-1-1-0]	GU	LC–ESI-MS	25.5	2 (0), 2 (1), 2 (2), 3 (0)	0.021, 0.001, 0.0002	↑1.4–1.7
		MRM	28.5	2 (0)	0.021	↑1.4
[4-3-2-0-0]	GU	LC–ESI-MS	24.8	2 (0)	0.009	↑1.6
		MRM	28.0	2 (0)	0.021	↑1.4
[5-3-3-0-2]	TU	LC–ESI-MS	33.0	3 (0), 4 (0)	0.005, 0.048	↑1.6
			34.6	3 (0), 4 (0)	0.022, 0.004	↑1.3–1.8
		MRM	34.3	3 (0)	0.003	↑1.5
[5-3-3-0-3]	TU	LC–ESI-MS	34.8	4 (0)	0.033	↑1.4
			36.7	3 (0)	0.007	↑1.6
		MRM	36.5	3 (0), 4 (0)	0.010, 0.029	↑1.4
[6-3-4-0-2]	TU	LC–ESI-MS	35.7	3 (0), 4 (0)	0.042, 0.030	↑1.4–1.6
		MRM	38.3	3 (0), 4 (0)	0.032, 0.024	↑1.3–1.4
[6-3-4-0-3]	TU	LC–ESI-MS	37.7	4 (0)	0.023	↑1.7
		MRM	39.5	4 (0)	0.015	↑1.5

Table C.2: Candidate N-glycan biomarkers identified in glycomic study (cont.)

monosaccharide composition	cohort	quantitation approach	RT (min)	adduct: z (no. $[H]^+ \rightarrow [NH4]^+$)	p-value	fold change
[6-3-4-0-4]	TU	LC–ESI-MS	38.8	3 (0), 4 (0)	0.014, 0.032	↑1.5–1.6
			39.3	3 (0), 4 (0)	0.021, 0.048	↑1.5–1.7
			39.9	4 (0)	0.001	↑1.9
		MRM	36.5	4 (0)	0.018	↑1.4
			40.8	4 (0)	0.024	↑1.7
[5-3-0-0-0]	TU	LC–ESI-MS	27.6	2 (0), 2 (1), 3 (0)	0.002, 0.003, 0.011	↓1.7
		MRM	29.5	2 (0)	0.0009	↓1.8
	GU	LC–ESI-MS	25.5	2 (0), 3 (0)	0.002, 0.014	↓1.1–1.2
[5-3-1-0-0]	TU	LC–ESI-MS	29.0	2 (0), 2 (1), 3 (0)	0.008, 0.009, 0.007	↓1.3
		MRM	30.3	2 (0), 3 (0)	0.018, 0.020	↓1.3
[5-3-1-0-1]	TU	LC–ESI-MS	31.8	2 (0), 2(1)	0.026, 0.008	↓1.5–1.8
		MRM	33.8	2 (0)	0.003	↓1.4
[3-4-1-0-0]	GU	MRM	29.5	2 (0)	0.041	↑1.3
[5-3-3-2-1]	TU	MRM	37.0	3 (0), 3 (0), 3 (0)	0.025, 0.027, 0.019	↑1.3–1.4
[6-6-0-1-2]	TU	MRM	39.5	3 (1)	0.041	↑1.3
[5-3-1-1-1]	TU	MRM	36.0	2 (0)	0.027	↓1.4
	GU	MRM	35.8	2 (0)	0.002	↓1.5
[3-3-0-0-1]	TU	LC–ESI-MS	24.5	2 (1)	0.020	↑1.4
[3-3-1-1-0]	GU	LC–ESI-MS	24.3	2 (0)	0.020	↑1.7
[4-3-2-1-0]	TU	LC–ESI-MS	30.1	2 (0)	0.006	↑4.8
	GU	LC–ESI-MS	27.1	2 (0), 2 (1)	0.012, 0.005	↑1.4–1.5
			28.0	2 (0)	0.041	↑1.4
[4-3-2-0-2]	TU	LC–ESI-MS	40.2	3 (0)	0.038	↑1.4
[5-3-3-0-1]	TU	LC–ESI-MS	30.9	3 (0)	0.022	↑1.3
			32.5	3 (0)	0.006	↑1.3
[6-3-4-0-0]	TU	LC–ESI-MS	33.1	3 (0)	0.017	↑1.8
[6-3-4-0-1]	TU	LC–ESI-MS	33.7	3 (0)	0.013	↑1.3
[4-3-0-0-0]	TU	LC–ESI-MS	23.6	2 (1)	0.008	↓1.9
[4-3-0-1-0]	TU	LC–ESI-MS	25.8	2 (1)	0.009	↓1.4
	GU	LC–ESI-MS	23.3	2 (1)	0.001	↑2.3
			24.8	2 (1)	0.002	↑1.52
			36.0	3 (0)	0.040	↓1.2
[4-3-2-0-0]	TU	LC–ESI-MS	32.3	2 (0)	0.008	↓1.4
			33.3	2 (0)	0.007	↓1.5
[5-3-2-1-0]	TU	LC–ESI-MS	39.5	2 (0), 3 (0)	0.039, 0.022	↓1.2
[4-3-1-0-1]	GU	LC–ESI-MS	26.2	2 (0)	0.013	↓1.2
[4-4-2-0-2]	GU	LC–ESI-MS	33.5	3 (0)	0.017	↓1.2
[6-3-4-1-2]	GU	LC–ESI-MS	35.1	4 (0)	0.031	↓1.1
[6-3-4-1-3]	GU	LC–ESI-MS	37.0	3 (0)	0.022	↓1.1