Methods for Utilizing Co-expression Networks for Biological Insight

by

Teal Guidici

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2018

Doctoral Committee:

Professor Charles Burant, Co-Chair
Professor George Michailidis, Co-Chair
Assistant Professor Johann Gagnon-Barsch
Professor Elizaveta Levina
Professor Kerby Shedden

Teal Guidici

tealg@umich.edu

ORCID iD: 0000-0003-0783-1598

for everyone who gave me grace and hope

ACKNOWLEDGEMENTS

As I reflect on the experiences that have led to this dissertation, I am struck by how much grace and kindness I have received along the way.

Throughout this journey, I have been fed, clothed and housed by friends and acquaintances. I have been driven to the grocery store, the airport, and the hospital. I have been comforted, given hope and good counsel. Important people with full schedules have freely given their time when I needed it.

I am thankful for all the grace I have received here. These are the memories I hope to carry with me, long after the rest has faded.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

DA Differential Abundance

FFA free fatty acids

GSA Gene Set Analysis

GSEA Gene Set Enrichment Analysis

LEVCD Leading Eigenvector Community Detection

PCST Prize Collecting Steiner tree

PUFA polyunsaturated fatty acid

ABSTRACT

The explosion of high-throughput Omics assays in past 15 years has led to a revolution in the quantity of data and the number of data types which are available to biological researchers. This has necessitated a second revolution in the development of analytical tools to handle this wealth and variety of data. No longer is it practical for a researcher to simply examine a list of differentially expressed compounds and draw meaningful insight about the biological processes at hand; these differentially expressed compounds must be put into context with each other, and integrated with existing biological knowledge. Co-expression techniques, where the simultaneous expression of two or more compounds is analyzed, have become a powerful tool for biological insight in high-throughput Omics settings.

The primary goal of this dissertation is to develop techniques for identifying and characterizing patterns of co-expression. In our first project, we develop a Differentially Weighted Factor Model for estimating covariance matrices related through structured experimental design. Our factor model allows us to estimate common structural elements using all available data, and to estimate unique structural elements in a condition specific manner. We develop a method for visualizing the resulting estimates, and implement the method in an R package, `DWFM`. The second project presents a method using the Prize Collecting Steiner Tree algorithm to integrate and identify modules in lipid and untargeted metabolomic assays in a data-driven manner. These assays are integrated over a co-expression network specific to the applied setting in question, allowing us to capture modules unique to this setting. Our final project presents a second technique for identifying modules

of co-expressed biomolecules. This technique addresses a major limitation of PCST based approaches, namely that one is required to choose a cutoff to obtain a list of differentially expressed compounds used as input into the algorithm. Additionally, this second method utilizes a meta-analytic inspired approach to identify patterns of co-expression across multiple data sets, thus reducing the impact of a single noisy assay.

CHAPTER I

Introduction

1.1 Background and motivation

1.1.1 Co-expression networks

Over the past decade, co-expression networks have proven themselves to be a powerful tool for analyzing high-dimensional Omics data. In such settings, the nodes correspond to biomolecules (genes, proteins, metabolites, lipids), while the edges capture statistical associations between them. These networks can be drawn from well annotated biological or pathway databases; it has been noted, however, that networks comprising of genome-wide (proteome-, metabolome-) interactions derived from experimental or observational data may contain novel interaction information not covered annotated in existing databases [2]. Studying such associations has enhanced our understanding of a number of biological phenomena, including dynamics of human disease [3], transcriptional changes associated with aging [4], and condition-specific alterations to metabolic pathways [5].

A number of techniques are available in the literature to estimate networks from data, including correlation based methods [6, 7, 8] and partial correlation ones [9, 10, 11]. The former are straight forward to calculate, but focus on highly connected compounds which may not be particularly informative, being potentially driven by

artifacts [12], while also not differentiating between direct and indirect interactions between compounds. Partial correlations have been used extensively in Omics settings, but require large sample sizes to calculate (see discussion in [10].)

Coexpression networks on a single dataset have been used to identify hub genes and pathway associations in retinoblastoma [13], to functionally annotate long noncoding RNAs (lncRNAs) and identify their potential cancer associations [14], and to identify potential treatment targets of peripheral arterial disease [15].

Differential analysis of co-expression networks, where the networks themselves are tested for differences between conditions, have been used to identify novel glioblastoma linked gene sets [16] and capture changes in functional interactions resulting from genetic/epigenetic changes that affect co-expression, but not expression, in the molecular pathogenesis of multiple sclerosis [17].

Co-expression networks aggregated over multiple related data sets have been used to gain insight into the ways in which cancer affects perturbs co-expression relationships [18], illuminate the role of indirect connections in gene networks [19] and identify functional modules of genes in a meta-analytic fashion [20].

1.1.2 Metabolomics

Metabolites are small molecules which are chemically transformed during metabolism; the metabolome is the collection of these metabolites in an organism. The metabolome is closely linked to phenotype [21], more so than the genome or proteome, as metabolites are direct signatures of biochemical activity. We are particularly interested in a subset of metabolites known as lipids - a group of organic compounds which are crucial for understanding cellular physiology and pathology. Both targeted and untargeted mass spectrometry assays are used in metabolomics and lipidomics studies; targeted assays seek to detect and measure a single metabolite or a select group of metabolites, while untargeted assays seek to measure *any* metabolite feature present

in the data.

Advancements in mass spectrometry, coupled with the development of more sophisticated data processing tools and comprehensive spectral libraries, have enabled researchers to probe deeper into the metabolome and lipodome; thousands of metabolite features can be measured simultaneously, necessitating the development of new techniques for functional interpretation of such data. In some ways, the field of metabolomics stands at a similar point to where the field of genomics stood 15 years ago, when gene set enrichment and pathway analysis became the tool of choice for gaining insight into the underlying biology of differentially expressed genes [22].

Just as Gene Set Enrichment Analysis (GSEA) [23] and Gene Set Analysis [24] approaches were developed to analyze high throughput genomics data, so various metabolite set enrichment analysis methods have been developed to understand high throughput metabolomics data [25, 26, 27]. Many of these methods, for any Omics data types, rely on canonical pathways or other knowledge drawn from existing biological knowledge bases, such as KEGG [28] or Gene Ontology [29].

This reliance on prior biological knowledge for gaining insight poses a particular problem in analyzing the lipidome, which is essentially unannotated to canonical features. Additionally, there are certain lipid classes which are tightly linked, as lipids in one class serve as precursors for those in another class; for example, there is the association between the lysophospholipid, diacylglycerol, and triglyceride classes that are involved in the triglyceride synthesis pathway. Such biochemical constraints give rise to more structured co-expression patterns which, if appropriately leveraged, can provide insight into disease processes.

1.2 Dissertation Overview

The primary subject of this thesis is identifying and characterizing patterns of co-expression; each remaining chapter presents a variation on this theme. We focus

on Omics applications where prior biological knowledge is lacking (e.g. lipidomics and untargeted metabolomics), and in each chapter the results of the method are used to generate hypothesis about the broader biological phenomenon at play.

In Chapter 2 we introduce a differentially weighted factor model capable of jointly estimating the structure of multiple related covariance matrices. Studying co-expression networks from high dimensional Omics data has been used to enhance our understanding of a wide range of biological phenomena. Often the data from such studies can be partitioned intro groups corresponding to experimental conditions or disease states. Traditional approaches analyze these groups separately, and thus do not make full use of all sample information. Our method uses all available data for the identification of common structural elements and group/condition specific data to estimate differential weights across conditions. We also introduce a method of visualization to aid in summarizing and interpreting the results of our method. The method's utility is demonstrated on lipidomics data from breast cancer patients.

The proposed method and visualizations have been implemented in the R-package `DWFM` available at https://github.com/tguidici/DWFM.

Chapter 3 presents a data-driven method utilizing the Prize Collecting Steiner Tree (PCST) algorithm to integrate and identify modules of differentially abundant lipids and untargeted metabolites, while also also incorporating relevant, non-differential compounds. We apply our method to data from a controlled feeding study and use condition specific co-expression networks to identify dietary linked modules of lipids and small polar molecules. We then confirmed that the identified modules were altered in an animal model of differential metabolite utilization.

The method we present in Chapter 4 addresses one of the limitations of the method in Chapter 3. As PCST methods rely on choosing a significance cutoff, biologically meaningful features which miss this cutoff could be missed. Identifying modules based only on patterns of co-expression between variables helps overcome this limitation.

We present a method for identifying modules based on co-expression patterns across multiple datasets. The method is applied to the data set used in Chapter 3, and the discovered modules are used to illuminate systems level changes in a second, more complex data set.

CHAPTER II

A Differentially Weighted Factor Model for Estimating Multiple Related Covariance Matrices with Applications to Lipidomics

2.1 Introduction

There has been much work in recent years on estimating co-expression networks from high-dimensional Omics data. In such networks, the nodes correspond to biomolecules (genes, proteins, metabolites, lipids), while the edges capture *statistical associations* between them. Associations can correspond to Pearson correlations or more robust variants such as Spearman's ρ or Kendall's τ, or partial correlations [30]. As noted in [31], networks comprising of genome-wide (proteome-, metabolome-) interactions derived from experimental or observational data, may contain novel interaction information not covered by the standard pathways. Studying such associations has enhanced our understanding of a number of biological phenomena, including dynamics of human disease [3], transcriptional changes associated with aging [4], and condition-specific alterations to metabolic pathways [5]. Patterns of co-expression are also important in studies involving metabolomics and/or lipidomics data, since changes across experimental conditions or disease states can provide insights into the flow of metabolites through latent metabolic processes. Further, since most lipid

species are not currently mapped to canonical pathways, studying their co-expression can provide useful information of identifying sets of lipids, possibly from different classes (see Appendix Section A.4 for a very brief primer on lipid classes), that can be the focus of downstream analysis. Note that in the case of lipids, there are certain lipid classes which are tightly linked, as lipids in one class serve as precursors for those in another class; for example, there is the association between the lysophospholipid, diacylglycerol, and triglyceride classes that are involved in the triglyceride synthesis pathway. Thus, such biochemical constraints give rise to more structured co-expression patterns which, if appropriately leveraged, can provide insight into disease processes. Such a setting arises due to the dietary and enzymatic influences that affect lipid metabolism [32, 33].

The data from many biomedical studies involving Omics data can be naturally partitioned into groups, corresponding to either different experimental conditions or disease states. The standard way to conduct analysis of co-expression patterns in this setting would be to estimate from the Omics measurements a *separate* co-expression network for each group and subsequently examine them for common patterns. Under the assumption that the groups exhibit *similarities* in their co-expression due to relationships between the experimental conditions or disease states, a better strategy would be to develop a statistical model that would enable *joint estimation* of all co-expression networks, thus utilizing the sample information across all groups efficiently, but at the same time allow for differences across groups to manifest themselves.

The motivation for such a modeling framework comes from examining co-expression lipidomics data for a set of women with early stage breast cancer being treated with aromatase inhibitor (AI) therapy. The subject's lipidomes are assayed before beginning the therapy, and 3 months into treatment. Half of the women were unable to tolerate the AI therapy for more than 6 months due to significant side effects (more details in Section 2.3.3 and Appendix Section A.4). A visual exploration of the co-

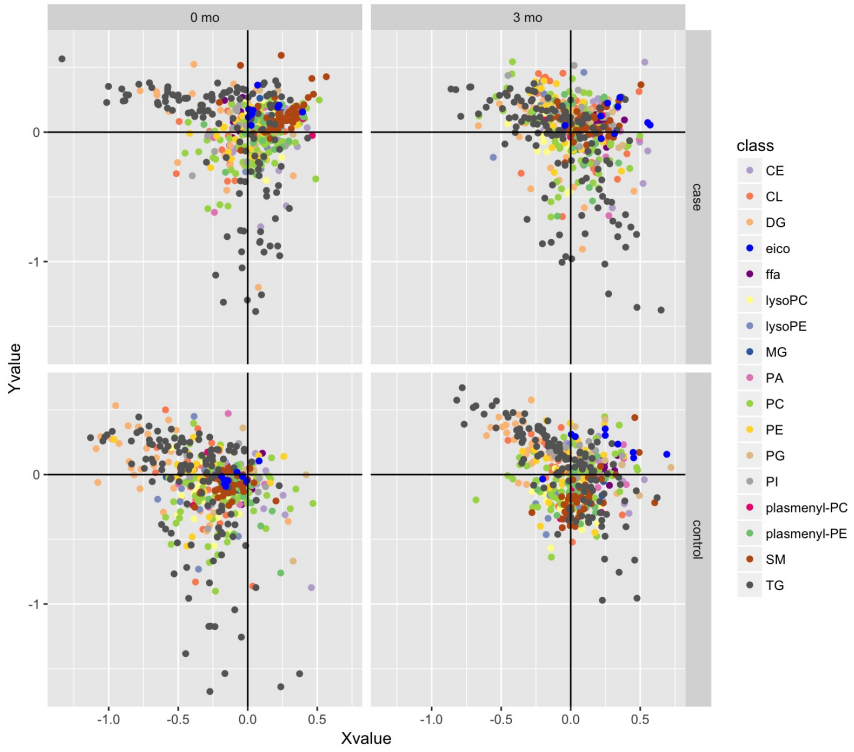


Figure 2.1: *Scaled first and second eigenvectors from each condition in AI dataset.* Eigenvectors taken from the eigendecomposition of the sample covariance matrices. Metabolites are colored by class. The TGs, colored in grey, provide a good illustration of the overall similarities and condition specific differences between the four conditions. They are relatively similarly placed in the four plots, but different subgroups of the TGs expand or contract from condition to condition.

variance matrices for the four groups shows a number of striking similarities across all of them, but also group specific patterns within each individual data set. These are depicted in Figure 2.1, where the first two principal components of each data set's covariance matrix are plotted, with the points colored by lipid class. It can be seen that a common structure is shared amongst the data sets, but the patterns are expanded or contracted in each data set; i.e., the group specific co-expression patterns are roughly *proportional* to some common structure shared across the data sets.

To formalize this finding, we propose a factor analysis model, wherein the factor loadings are decomposed onto a common set of loadings, shared across all groups, but which we allow to be differentially weighted within each group. Thus, when esti-

mating the common loadings, information from all samples in the study is utilized, a particularly attractive feature in the analysis of data sets with a small number of samples in each group (a common enough setting for Omics data), while the differential weights are calculated based on the sample information from each group.

This modeling framework can inform biomedical researchers of the set of variables (biomolecules) which are acting in a concordant manner and characterize the underlying biological processes in the data, while their differential weighting sheds light on how these common patterns of association are modulated or disrupted in a specific group. Such information could then suggest avenues for further investigation into potential mechanisms underlying the observed differences between groups.

This problem has received some attention in the bioinformatics and biostatistics literature. A number of approaches focuses on data sets representing different Omics modalities (genomics, proteomics, etc.) but acquired on the *same* set of samples and the goal is either to remove idiosyncratic variability [34], or find common signatures across Omics modalities for subsets of samples [35, 36]. These approaches use matrix decompositions (e.g. principal components analysis or non-negative matrix factorization) appropriately constraining their parameters across data sets; for a review of related approaches see also [37]. When the statistical associations in the co-expression networks correspond to partial correlations, one has also to contend with the issue that there are more variables present than samples, and therefore such networks can only be estimated from Omics data only if additional structural assumptions are imposed; the most common and popular assumption being that of sparsity - namely, that most interactions are non-present. There is an expanding body of literature on this topic -see e.g. [38, 39, 40] and references therein. Finally, [41] focuses on the problem of testing for differences between two covariance matrices defined on the same set of variables, obtained from two different data sets, without assuming any structural similarities between them.

In this work, the focus is on identifying common, but "differentially weighted" co-expression patterns, which shares some conceptual commonalities with the work on partial correlation networks mentioned above. However, the technical framework and developments are different. Finally, it should be noted that the proposed methodology is capable of identifying data driven strongly interacting modules of biomolecules, that can be of interest to test for enrichment, as discussed in [31]. This feature is particularly useful in studies involving lipidomics data, where as previously mentioned, canonical pathways are not well delineated. Hence, their enrichment analysis has the potential to provide a systems perspective which can lead to deeper biological insights.

2.2 Methods

2.2.1 Differential Weighted Factor Model

We start the presentation by providing some basic background on the factor model for a *single* data set, before generalizing to the K-group setting. Let X denote a data set of size $n \times p$, containing Omics measurements on p variables (biomolecules) collected from n samples. Since the focus is on understanding co-expression/co-variation patterns amongst variables, a factor model represents a standard tool for this task. Formally, let x be a p-dimensional vector of random variables with distribution function H representing the p Omics measurements. Hence, the n samples are independent and identically realizations from the distribution H. The classical factor model [30] posits that the i-th variable can be expressed as a linear combination of m latent common factors $\{f_j\}_{j=1}^m$ and an idiosyncratic error; namely $x_i = \sum_{j=1}^m \lambda_{ij} f_j + \epsilon_i$, where λ_{ij} is a weight and ϵ_i an error term. Since both the factors and the errors are unobservable, for identifiability purposes it is assumed that $F = [f_1, \cdots, f_m]'$ and $E = [\epsilon_1, \cdots, \epsilon_p]'$ satisfy $\mathbb{E}(f) = 0, \mathbb{E}(E) = 0$,

$\mathbb{E}(FF') = I, \mathbb{E}(EE') = \Psi^k, \mathbb{E}(F'E) = 0$, where Ψ is a *diagonal matrix* containing the variances of the error term. Therefore, the common factors and the idiosyncratic components are uncorrelated and also uncorrelated between them. Then, some standard algebra gives that $\text{Cov}(x) \equiv \Sigma = (\Lambda)(\Lambda)' + \Psi$, where Λ is a $p \times m$ matrix of factor loadings with elements $\{\lambda_{ij}\}_{i=1,j=1}^{p,m}$. The latter equation can be used to estimate from the data set X, the corresponding model parameters (Λ, Ψ) by applying for example an eigenvalue decomposition on the sample covariance matrix $S = (X'X)$ [30]. Note that Λ is identified up to rotations, since for any orthonormal matrix Φ, we have that $\Sigma = \tilde{\Lambda}\tilde{\Lambda}' + \Psi$, where $\tilde{\Lambda} = \Lambda\Phi$ and $\tilde{F} = \Phi F$.

Our objective is to model *jointly* the co-expression/co-variation of the K groups. To that end, we posit the following model for the p random vector x^k that generates the observed data in set X^k comprising of n_k samples:

$$X^k = \Lambda^k F^k + \epsilon^k = B^k Q F + \epsilon^k, \quad k = 1, \cdots, K, \tag{2.1}$$

where $\Lambda = B^k Q$, with Q being a $p \times m$ matrix of *common factor loadings* and B^k, a $p \times p$ diagonal matrix, being a set of *differential weights*. Further, the assumptions on the factors F^k and idiosyncratic components E^k are as in the single model case, and further we assume that $\mathbb{E}(F^k(F^\ell)') = 0, \mathbb{E}(E^k, (E^\ell)') = 0, \mathbb{E}(F^k(E^\ell)') = 0, \quad \forall\ k \neq \ell = 1, \cdots, K$. Similar calculations as above yield that each group specific covariance matrix $\Sigma^k = B^k(QQ')B^k + \Psi^k$. Note that the model reduces significantly the number of loading related model parameters to $(p \times m) + Kp$ from $k \times (p \times m)$, in the unconstrained version above.

To complete the model formulation, we need to impose a further identifiability constraint on the differential weight matrices B^k; otherwise, we can inflate all of them by a factor c, deflate the elements of Q by the same factor c and the model would not change. To that end, we propose the following two identifiability constraints.

The first is $\sum_{k=1}^{K} B^k = I_p$ (ID1), where I_p denotes the p-dimensional identity matrix. According to it, the weights for each variable i must sum to 1 across the K conditions. The second one is motivated by applied settings where the K groups can be naturally arranged according to the levels of a two-way factorial design. For example, suppose we have $K = 4$ groups of disease and normal, female and male patients. Hence, our two grouping factors are disease status with $K_1 = 2$ levels and sex with $K_2 = 2$ levels. The second identifiability constraint is then given be $\sum_{k=1}^{K_1} B^k = I_p$ (ID2) for those groups k, where one of the grouping factors is fixed at a certain level (e.g. normal) and we are normalizing over the levels of the other grouping factor.

Remark II.1. The model formulation assumes that the factors F^k for each group span different (orthogonal) subspaces, and their coupling comes from the structure imposed on the factor loading matrices Λ^k. An alternative would have been to allow the factor $F = [F^1, \cdots, F^K]'$ to have correlated elements between the various F^k's, but not within. That is $\mathbb{E}(F^k(F^k)') = 0$ for all $k = 1, \cdots, K$, but $\mathbb{E}(F^k(F^\ell))' \neq 0$. In this case, to identify the model parameters we would require that the factor loading matrix Λ containing the loadings for all K factors $F^k, k = 1, \cdots, K$ be block diagonal. The most restrictive model formulation assumes that $F^k = \tilde{F},\ \forall\ k = 1, \cdots, K$, that is there is a single latent space explaining the co-variation structure of all the K data sets, but this proves excessively stringent, since this would suggest to put all K data sets together and analyze them as a *single* one.

2.2.2 Estimation Procedure

Since we assume that the data X^k span different subspaces, we can estimate the model parameters by the following two stage procedure:

1. Estimate Λ^k through the eigenvalue decomposition of S^k the empirical covariance matrix of Σ^k.

2. Estimate the model parameters by solving $\Lambda^k = B^k Q$ subject to the identifiability constraints (ID1) or (ID2).

Specifically, since we require a rank-m solution, an application of the Eckart-Young theorem gives that the best rank-m approximation (in squared Frobenius norm) of $\tilde{S}^k = U^k_{[1:m]} D^k_{[1:m]} (U^k_{[1:m]})'$, i.e. the eigenvectors corresponding to the largest m eigenvalues of S^k. Therefore, a rank(m) estimate for Λ^k is $\widehat{\Lambda}^k = U^k_{[1:m]} \sqrt{D^k_{[1:m]}}$. Then, from step (2) above, we get using (ID1) that

$$\widehat{Q} = \sum_{k=1}^{K} \widehat{\Lambda}^k \tag{2.2}$$

or under the (ID2) constraint, the estimate is given by $\widehat{Q} = \frac{1}{K_1} \sum_{k=1}^{K} \widehat{\Lambda}^k$ or replacing K_1 by K_2.

Then, the estimates for the differential weights are obtained under both normalization constraints by

$$\widehat{B}^k_{ii} = \frac{1}{m} \sum_{j=1}^{m} \frac{\widehat{\Lambda} ij}{\widehat{Q}_{ij}}. \tag{2.3}$$

Next, we briefly discuss two implementation issues. The first has to do with selecting the number of factors; one can use the standard strategies employed in factor and principal components analysis for each Σ^k, namely employ the scree plot and spot where the knee in the eigenvalues occurs, or require that the total variance explained by the first m factors exceeds a certain percentage (e.g. 60% or 70%). The final selection for m would be that value that is most compatible across all K groups. The second issue deals with the rotational invariance of the factor model discussed above, which is also a feature of the eigenvalue decomposition [30] used to obtain $\widehat{\Lambda}^k$. Since the $\widehat{Q}$ is an average of K $\widehat{\Lambda}^k$'s, we need to ensure that all of them have the same "orientation". The latter is addressed through a Procrustes rotation (see [42].)

Extension to Sparse Factor Models: For many data sets, the factor structure may be sparse, wherein different subsets of variables load on different factors. The estimation of sparse Λ^k's becomes more involved. There are a number of proposals in the literature that address this issue, including (i) sparse principal component analysis (SPCA) [43, 44] that employs an elastic net (a combination of a lasso and a ridge) penalty to obtain sparse principal components, while relaxing their orthogonality constraint commonly assumed in classical PCA, (ii) the standard eigenvalue decomposition used in Step 1 of the proposed estimation procedure, where the eigenvectors are truncated by magnitude [45] (EDTM) and (iii), a novel method presented here, similar to (ii) but the eigenvectors are truncated by the cardinality of their support (EDTC).

Note that approaches (i) and (ii) require careful selection of tuning parameters; specifically, for SPCA the regularization parameters the control the lasso and ridge penalties, while for EDTM selecting a universal threshold value. Without a priori knowledge of the sparse patterns in Q, tuning these methods directly proves challenging, as manifested in our simulation studies. Instead, we developed an alternative approach that provides good estimates of the number of factors m together with their support using a community detection algorithm. Each covariance matrix is regarded as a graph, with edge weights given by the absolute value of the covariance matrix. The leading eigenvector community detection algorithm (LEVCD) [46] is used to find groups of highly connected nodes (correlated variables). These communities correspond to the columns of Q and the size of each community roughly corresponds to the number of non-zero components (the cardinality of the support) for a given column of Q. This last fact is key - we could have estimated the number of factors from a scree plot, but it would not have provided the support size for each factor. We set the number of factors, m, equal to the smallest number of communities within a data set, across all K data sets. The cardinality of non-zero support in each factor is set to the

cardinality of these communities $s = (s_1, \cdots s_m)$, where $s_i > s_{i+1}$. These estimates for m and s, can then be used to estimate $\widehat{\Lambda}^k$ via either SPCA (Appendix Algorithm 4) or EDTC (Appendix Algorithm 3).

All algorithms are detailed in full in Appendix A.1, along with a performance review on selecting the tuning parameters, and a comparison of all three methods in Appendix A.2.

2.2.3 Illustration of the Method through a Toy Example

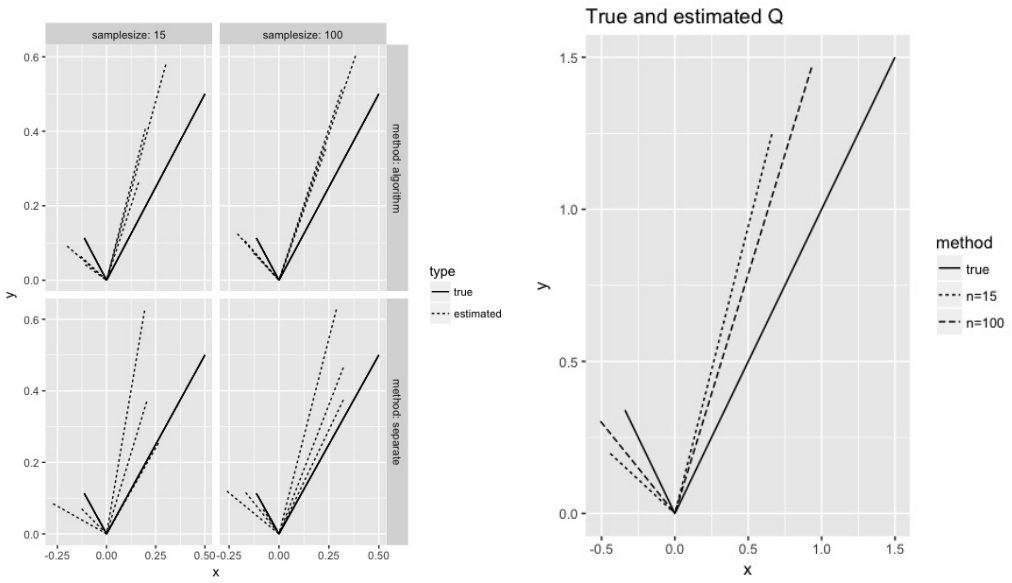
An intuitive understanding of how the method works can be gained via the following small example. Suppose we have data from an experiment with 3×1 experimental design. The data have 2 common latent factors, 2 variables are observed across data sets. The B^k_{ii} are equal across all data sets ($B^k_{ii} = 0.33$), but the noise in the observed data (E^k) increases with each condition. The error terms are generated so that $cor(X^1_1, X^1_2) \approx 0.9$, $cor(X^2_1, X^2_2) \approx 0.75$ and $cor(X^3_1, X^3_2) \approx 0.6$.

We can compare the results of our method with the standard approach of separate eigen-decompositions by computing the reconstruction loss in Equation A.5. We expect our method to perform well when the sample size is large; we also expect gains over the standard approach when the sample size is small. In panel (a) of Figure 2.2 we can see that this is indeed the case.

We can also see in panel (b) of Figure 2.2 and Table 2.1 that our estimation of B^k and Q moves closer to the truth as sample size increases.

2.2.4 Visualization

Note that the output of the proposed model comprises of the common factor loadings $\widehat{Q}$ and the differential weights $\{B^k\}_{k=1}^K$. The common factor loadings can be represented as in any standard factor or principal components analysis through a scatterplot of its elements. Similarly, the reconstructed covariance matrices $\widehat{\Sigma}^k = \widehat{B^k}\widehat{Q}\widehat{Q}'\widehat{B^k}$



(a) $\widehat{\Lambda}^k$ estimated with different sample sizes (b) true Q and estimated $\widehat{Q}$

Figure 2.2: *Toy Example:* (a) Both our method and separate eigen-decompositions improve with increasing sample size. Our algorithm (top row) does a better job of estimating the $\widehat{\Lambda}^k$ similarly, while the estimates from separate eigen-decompositions are spread further apart, reflecting the influence of increasing noise in the data. Left to right by row, the reconstruction loss values are: 0.153 (method, n=15), 0.108 (method, n=100), 0.161 (eigen-decomp, n=15), 0.109 (eigen-decomp, n=100). (b) As the sample size increases, $\widehat{Q}$ moves closer to the true Q.

	B^1	B^2	B^3
true B^k_{11}, B^k_{22}	0.33	0.33	0.33
$n = 100, \widehat{B}^k_{11}$	0.25	0.34	0.41
$n = 100, \widehat{B}^k_{22}$	0.24	0.35	0.41
$n = 15, \widehat{B}^k_{11}$	0.25	0.30	0.46
$n = 15, \widehat{B}^k_{22}$	0.21	0.33	0.46

Table 2.1: *Toy Example Performance.* $\widehat{B}^k$ for $n = 100$ and $n = 15$. True $B^k_{11} = B^k_{22} = 0.33, \forall k$.

can be visualized through heatmaps. For the differential weights, we propose to compute a quantity similar in spirit to the traditional log-fold-change. For comparing groups k and ℓ, we define $r^{k\ell}_i = \log(\widehat{B}^k_{ii}/\widehat{B}^\ell_{ii}), i = 1, \cdots, p$. This ratio provide information about how the loadings $\widehat{Q}_{\cdot i}$ for variable i are modulated between groups k and ℓ; for example, if $r^{k\ell}_i$ is close to 0, then the patterns of association involving variable i are similar in both groups. As $r^{k\ell}_i$ moves away from 0 and becomes more positive, the associations involving i become, on average, stronger in group k relative to ℓ, while if it becomes increasingly negative, the associations become weaker in group k relative to ℓ. The ratios $r^{k\ell}$ are visualized as bar charts, with bars extending to either side of the central axis (representing 0). A visualization of synthetic data corresponding to 4 groups organized according to two experimental design factors with two levels each are depicted in Figure 2.4.

Figure 2.3 illustrates our approach to visualization in a setting with 3×1 experimental design structure.

2.3 Results

Next, we provide an in depth performance evaluation of the proposed methodology based on synthetic data. We start with a comparison over estimating K separate models, and then focus on the performance of the method under different settings

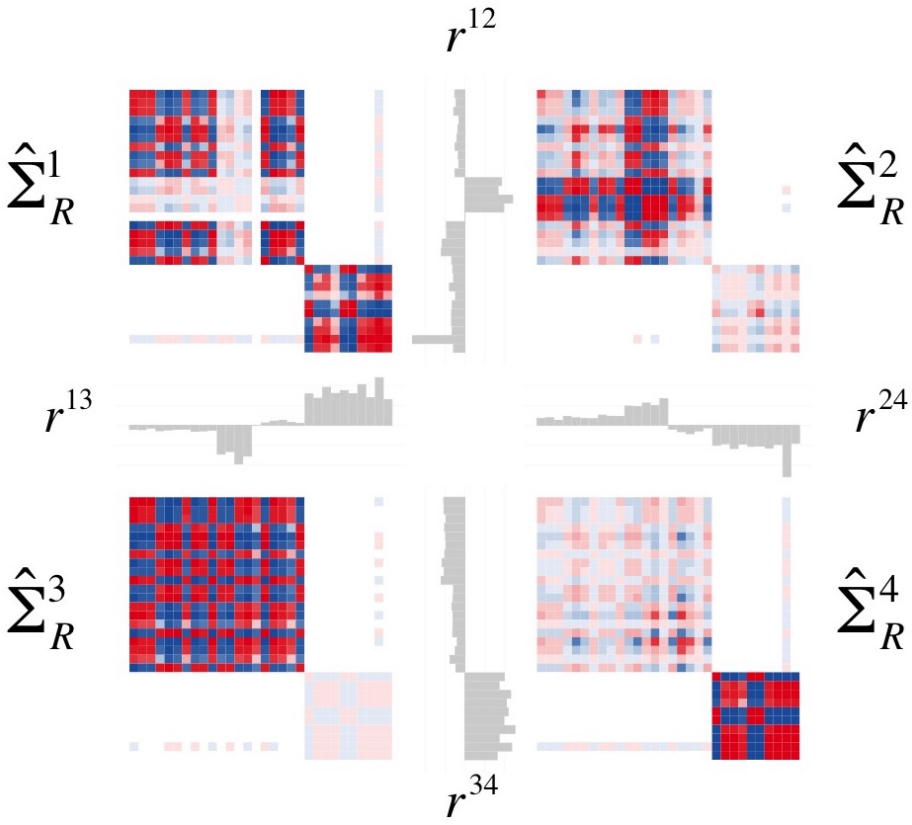


Figure 2.3: *Visualization for* 2×2 *experimental design.* Visualization for results from a simulated 2×2 setting with the following selection of model parameters; $m = 2, p = 30, n = 100$. Latent factors account for 75% of observed variance, while Q_1 has 20 non-sparse entries and Q_2 has 10. Reconstruction loss is 0.1486. The reconstructed covariance matrix heatmaps are interpreted in the usual manner. The bar charts are oriented so that longer bars extending in the direction of a specific reconstructed covariance matrix k indicate that those B^k values are larger than the same values in the condition the bars are pointing away from. We can see that the visualization allows us to see the common block structure, as well as specific differences between conditions (compare, for example, the lower right block and the relevant sections of the bar charts between all 4 heatmaps.) This figure also allows us to observe the impact of the experimental design on the normalization - $\widehat{B}^1$ and $\widehat{B}^2$ are normalized together, as are $\widehat{B}^3$ and $\widehat{B}^4$. This can be seen in the figure, where $\widehat{\Sigma}_R^1$ and $\widehat{\Sigma}_R^2$ have patterns of light and dark cross-hatching which are complimentary to each other, and very different from the patterns present in $\widehat{\Sigma}_R^3$ and $\widehat{\Sigma}_R^4$

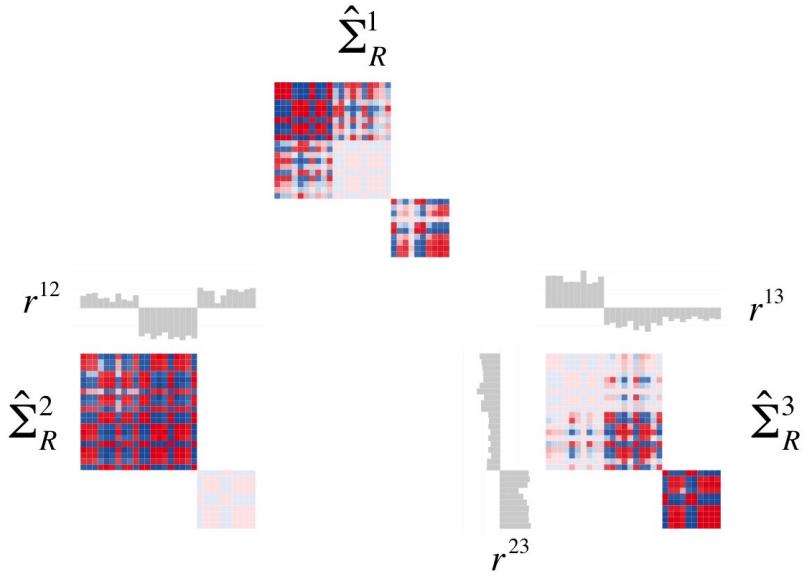


Figure 2.4: *Visualization for* 3×1 *experimental design.* We observe that the visualization allows us to see the common structure present in all of the covariance matrices - they all share a block diagonal structure, with a large block in the upper left (representing the first factor), and a small block in the lower right (the second factor). Differences between the conditions can be seen in both the color and intensity of the individual cells in the heatmaps, and in the bar charts. We can see this concretely by regarding the lower right block in $\widehat{\Sigma}_R^1$. This represents the variables which load onto Q_2. Overall this square is darker than the equivalent square in $\widehat{\Sigma}_R^2$ and lighter than the one in $\widehat{\Sigma}_R^3$. This relationship between $\widehat{\Sigma}_R^1$ and the other two datasets is also reflected in the bar charts - in r^{12} the bars corresponding to the second factor point towards $\widehat{\Sigma}_R^1$, but in r^{13} these same bars point away from $\widehat{\Sigma}_R^1$ and towards $\widehat{\Sigma}_R^3$.

involving the key model parameters p, n, m, K and the noise level.

2.3.1 Comparison to performing separate analyses

Using synthetic data across a range of simulation settings: 3×1, 2×2 and 5×1 experimental design settings, with 2 latent factors, 50 or 100 variables, and a range of sample sizes. Further, the latent factors account for 50% of the observed variation. We compare the reconstruction error from using K separate scaled eigen-decompositions

or using the proposed method to estimate Λ^k. We compute

$$\frac{1}{K}\sum_{k=1}^{K}\frac{||\tilde{\Lambda}^k - \Lambda^k||}{||\Lambda^K||} \tag{2.4}$$

where $\tilde{\Lambda}^k$ is either $\widehat{\Lambda}^k$, or $\widehat{B}^k\widehat{Q}$. The ratio of this reconstruction error for the two approaches is given in Table 2.2. In general, the proposed method outperforms undertaking K separate analyses. The greatest gains are observed when the experimental design necessitates constraint ID2 (the 2×2 experimental design setting), or when noisy entries of a modest size (cutoff = 0.05) are removed. In settings where ID1 is used (3×1 and 5×1), and the cutoff is too large (0.2), our method slightly under-performs. This is primarily due to the fact that the selected threshold 0.2 sets some true non-zero entries to zero (a common practice in classical factor analysis to aid interpretation), and the additional steps included in the algorithm to deal with noisy entries slightly decrease the accuracy in these scenarios.

	3×1	2×2	5×1
cutoff = 0	0.9516	0.8903	0.9221
cutoff = 0.05	0.9498	0.8860	0.9264
cutoff = 0.1	0.9772	0.8925	0.9987
cutoff = 0.2	1.0223	0.9416	1.0092

Table 2.2: *Average loss ratio:* (method based reconstruction loss)/(separate eigen-decomposition reconstruction loss). Results are averaged over 500 error realizations, all n and all p. Entries with magnitude less than the cutoff are set to 0.

2.3.2 Performance evaluation based on Synthetic Data

Next, we consider a large set of scenarios to test the performance of the model and the corresponding estimation strategies for both sparse and non-sparse Q, with varying experimental design structures. For each simulation setting, we measured the algorithm's performance on estimating a range of model elements. Here, we present the algorithm's performance in computing $\widehat{B}^k$, for a subset of simulation settings that

give an overall sense for the algorithm's behavior.

This performance can be measured by *B-loss*, calculated as

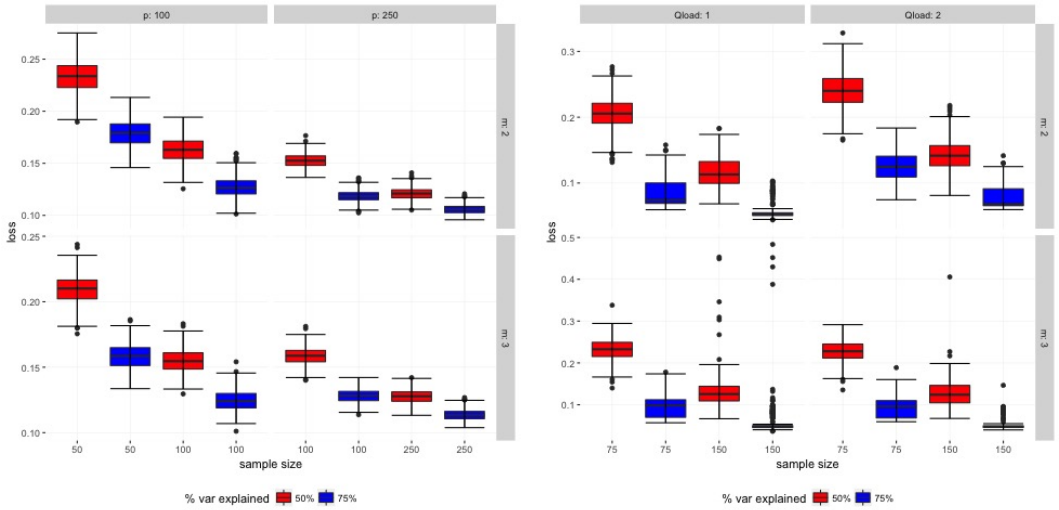
$$B\text{-}loss\text{:}\ \frac{||B-\widehat{B}||_F}{||B||_F} \tag{2.5}$$

where B is a $p \times K$ matrix whose k^{th} column is the diagonal entries in B^k. Figure 2.5 shows this loss for 3×1 experimental design structure (a single design factor with 3 levels), with non-sparse Q settings in panel (a) and sparse Q settings in panel (b). See figure legend for complete simulation details.

It can be seen that the algorithm behaves in an expected manner; specifically, loss decreases as the percentages of variance explained by the latent factors increases, or as sample or variable size increase. We also observe that the standard deviations of the loss values are quite small - indicating that the algorithm consistently exhibits good performance.

When Q is sparse, these trends still apply, but the loss values are generally higher and the range of loss values is broader. This can be attributed to the challenge of identifying the sparse support of Q with a very high degree of accuracy. In the non-sparse case, mean *B-loss* is always less than 0.233, with half of the instances even below 0.11. For the sparse case, the mean *B-loss* is higher, ranging from 0.042 to 0.24, but in the majority of the settings the loss values remain below 0.15.

In Appendix A.3results on the model's performance in estimating $\widehat{Q}$ and reconstructing $\widehat{\Lambda}^k$, across additional simulation settings (2×2 experimental design structures are also tested) are provided. Further, a discussion of a scaling factor, s_{ij}, included to adjust for noisy entries and complete technical details on the simulation procedure are also given.



(a) *B-loss* with non-sparse Q

(b) *B-loss* with sparse Q

Figure 2.5: *B-loss for* 3×1 *experimental design.* We consider $m = 2, 3$, and Ψ such that Λ^k explains 50% or 75% of the variance present in the simulated data. For the non-sparse Q (each column in Q is fully populated with non-zero entries), we consider $p = 100, 250$ with $n = 50, 100, 250$, depending on p. When Q is sparse, we test $p = 100$ and $n = 75, 150$. For both 2 and 3 factors in the sparse scenario, we also vary the amount of sparsity in each column of Q. The first column of Q always has the greatest number of non-zero entries, followed by the second column and so on. Every row in Q has a single non-zero entry, giving the columns of Q distinct, non-overlapping support. For $m = 2$ we look at Q having $(0.55p, 0.45p)$ and $(0.7p, 0.3p)$ non-sparse elements, while for $m = 3$ we test Q with $(0.39p, 0.33p, 0.28p)$ and $(0.5p, 0.3p, 0.2p)$ non-sparse elements. These patterns of loading are referred to as "Qload: 1" and "Qload: 2". All scenarios are run for 500 error realizations with a single realization of Q and B^k, $\forall\, k$. Our preferred method for estimating sparse $\widehat{\Lambda}^k$ is the EDTC method, as it yielded the best results (see Appendix Algorithm 3 and Appendix Section A.2 for more details.)

2.3.3 Application to lipidomics data

We further illustrate the usefulness of the proposed model by analyzing the following lipidomics data set. Aromatase inhibitor (AI) adjuvant therapy is effective in reducing the recurrence early stage hormone receptor-positive of breast cancer by inhibiting the conversion of androgens to estrogen, interrupting estrogen-dependent cancer cell growth [47]. AI use is associated with adverse events, including muscle and joint pain that occurs in up to 50% of treated patients [48]. These side effect can affect quality of life and lead to discontinuation of the drug in a significant proportion of symptomatic patients [49] .

Dietary fatty acid intake, especially ω3- and ω6-fatty acids, can affect overall inflammation by altering the production of pro- and anti-inflammatory cytokines [50]. An earlier study showed that women treated with daily ω3-fatty acids showed no difference in incidence of AI side effects when compared to soybean/corn oil supplementation [51]. However, it was noted that both groups showed a significantly higher improvement during the intervention phase than had been seen in other intervention trials to modulate side effects following AI inhibition. One potential explanation for this finding is that the "placebo" soybean and corn oil have over 50% content of ω6-fatty acids, which also can have anti-inflammatory effects [52].

Given the above, we employed the proposed model to identify potential differences in patterns of co-variation of lipids in the serum lipidome of women who developed symptomatic arthralgias following treatment with AI (Cases $n = 24$) compared to women who remained asymptomatic (Controls $n = 25$). Samples were derived from a prospective clinical trial (more details in Appendix A.4.) Cases were defined as women who were unable to continue treatment for more than 6 months due to the development of musculoskeletal pain, whereas controls remained symptom free (defined as pain $\leq 2/10$) for at least 24 months. Clinical characteristics of the women were not different between cases and controls (Appendix Table A.1) nor were there a

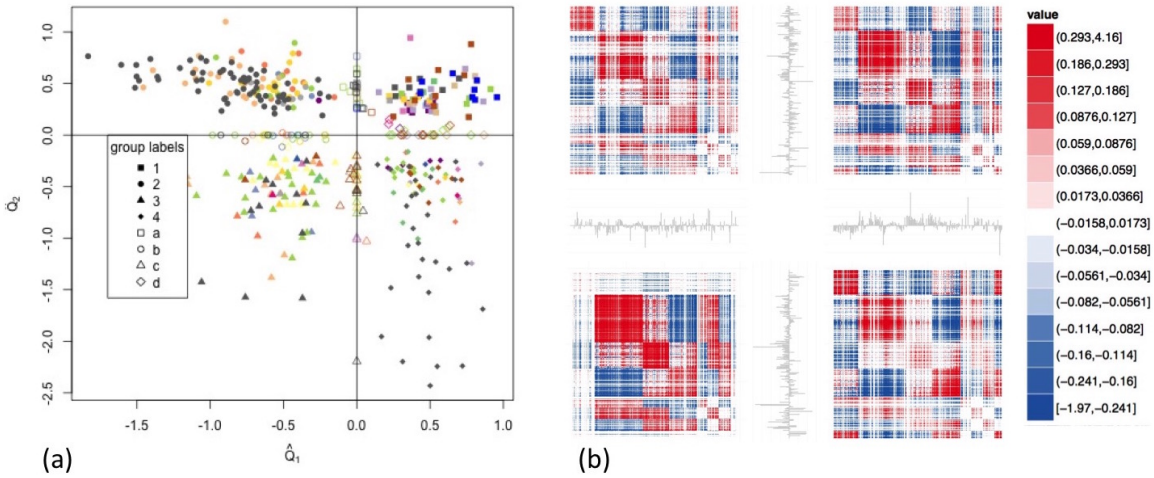


Figure 2.6: *Visualization of results for AI data* (a) $\widehat{Q}$ for AI data. Metabolites are colored as in Figure 2.1, and are grouped according to their loadings onto $\widehat{Q}$. (b) $\widehat{\Sigma}_R^k$ and $r^{kk'}$ for AI data. Variables in each heatmap are sorted by group assignment from (a). Clockwise from upper left, the reconstructed heatmaps are: Cases at baseline, Cases at 3 months, Controls at 3 months, Controls at baseline. In the same order, the bar charts are: Case at baseline vs Cases at 3 months, Cases at 3 months vs Controls at 3 months, Controls at baseline vs Controls at 3 months, Controls at baseline vs Cases at baseline.

difference in the type of AI used.

Subjects' lipidomes were assayed at baseline, and 3 months following initiation of AI treatment; a total of 442 lipids in 15 classes were identified and their relative levels determined by LC/MS. We also measured a small set of eicosanoids which are potential modulators of inflammation [53], and free fatty acids (FFA).

A description of the normalization procedure used, as well as pre- and post-processing steps for the method and wet lab lipidomics methods can be found in the Appendix A.4.. In short, after normalizing and selecting samples which were present in all three analytical platforms, there were 24 subjects who could not continue receiving the drug (cases) and 25 subjects who could continue being on the drug (controls). For these 49 subjects, data were available on 467 compounds (lipid species, FFA and eicosanoids).

In preliminary analysis of the data, paired and unpaired t-tests were used for identifying differentially expressed compounds related to the following comparisons: (i) cases at baseline vs cases at 3 months, (ii) controls at baseline vs controls at 3 months, (iii) cases vs controls at baseline and (iii) cases vs controls at 3 months. After correcting the resulting p-values for multiple comparisons using the Benjamini-Hochberg False Discovery Rate adjustment procedure [1], there were only 3 metabolites (CE 22:6, TG 62:13, TG 62:14) with an adjusted p-value less than a 0.1 threshold, in the controls at baseline vs controls at 3 month comparison. No statistically significant differential abundance was found for any other comparison.

We then turned our attention to co-variation patterns exhibited across the four groups, using the proposed factor model. Based on scree plots, we set the rank to be $m = 2$. The differential weights were normalized based on the (ID2) constraint applied to the samples in the controls and the cases, respectively. To enhance interpretation, we thresholded small values of $\widehat{Q}$ (removing values between -0.2 and 0.2) and the results are depicted in Figure 2.6, where the coloring scheme of the compounds corresponds to their class. Based on their location on this plot, each compounds can be assigned to a group; this assignment is reflected by the plotting symbols used in Figure 2.6.

The common factor loadings reflect the association of each variable with the latent factors, with a positive value indicating that the specific compound is positively associated with the corresponding factor. This fact provides a data driven strategy for grouping variables (compounds) for further analysis, in settings where canonical pathways are not defined, or well studied.

With the compounds divided into groups, we tested those groups for over-representation of each lipid class, and saturation level. Similar to the approach often taken in the analysis of gene expression data organized into canonical pathways, we also tested each group for differential abundance using the GSA technique [24]. More specific

methodological details and full results for these tests can be found in Appendix Section A.5, and Appendix Tables A.2 and A.3.

We found that group 1 enriched for saturated compounds, group 2 for monounsaturated, and both groups 3 and 4 strongly enriched for polyunsaturated lipids. Group 1 included many FFA and eicosanoids, having over half of the FFA and all of the eicosanoids. Group 2 mainly contained the DG and TG lipid classes. Group 3 was enriched for the lysoPC and PC lipid classes, while Group 4 was enriched for the plasmenyl-PE and TG lipid classes.

Using GSA we tested all 2-way comparisons of interest (described above), and found that only group 4 showed any differential abundance, with the controls exhibiting lower abundance of the respective compounds at month 3, vs the controls at baseline. As previously mentioned, group 4 contains mostly polyunsaturated lipids, which have polyunsaturated fatty acid (PUFA) tails. Since longer chain PUFA species with multiple double bonds are primarily derived from the elongation and desaturation of dietary essential fatty acids [54], the significant decrease in Group 4 lipids following AI treatment in Controls could be due to differences in PUFA intake or metabolism.

Women have higher levels of the ω3-fatty acid, docosahexaenoic acid (DHA) than men due to estrogen, but these levels fall in the absence of estrogen due to decreased conversion of dietary ω6-fatty acids to DHA [55]. The fall in the levels of the very long chain fatty acids in women who remained asymptomatic could be due to a decrease in the intake of ω3-fatty acids with an increase in the relative intake of ω6-fatty acids, including α-linoleic acid. Accumulating evidence suggests that dietary α-linoleic acid may have potent anti-inflammatory effects and reduces cardiovascular risk factors. Given the greater than expected response from the 'placebo' arm which contained high levels of α-linoleic in studies testing the efficacy of fish oil to alleviate arthralgias in AI treatment may due to salutary biological effects of lipids in soybean/corn oil in

the placebo.

2.4 Discussion

There is increased interest in studying co-variation patterns across multiple data sets, as in routinely done with analysis of variance for examining mean changes. The proposed method fills in this gap. As extensive numerical work shows, the method performs well across a range of simulation settings in identifying common co-variation patterns in complex experimental settings and can also accommodate sprase factor structures.

As the AI case study shows, the resulting estimates of the common factors and their differential weights prove useful for gaining deeper biological insights. To do so effective visualization strategies are needed and this work provides plots that successfully summarize the bulk of the information produced by the algorithm.

At present, the model serves as an exploratory analysis tool. If one is interested in a more confirmatory approach, it would be possible to do hypothesis testing on the elements of B^k. Bootstrapping could be used to generate confidence intervals for B^k. Fox example, if one's experimental design allowed for constraint (ID1), and one had K total data sets, a tight confidence interval around $\frac{1}{K}$ for B_i^k would imply that there is no (or little) condition-specific modulation for the i^{th} metabolite across the K conditions. Further, permutation tests could be used to test $||B^k - B^{k'}||$ for two conditions k, k', where $||\cdot||$ is an appropriate norm that reflects well the magnitude of the differences in the two conditions. If $||B^k - B^{k'}||$ was not significantly different from zero, then one could consider treating k and k' as the same condition - allowing the researcher to pool the data and increase sample size for that condition. This would, of course, change the calculations for the normalization constraints, and could even change which normalization constraint was applied. This pooling would need to be done with care, as would consideration of the label swapping in the permutation

test. Additionally, one could test $||B_i^k - B_i^{k'}||$, over a subset of metabolites, such as those with high loadings on a particular column of $\widehat{Q}$.

At a technical level, the model can be used with more complicated experimental designs (such as a 5×4 design). Visualization in these scenarios is less straightforward and more advanced tools such as those provided in [56] and Shiny [57] need to be employed to enable researchers to fully explore the results and understand co-variation patterns in their data.

Finally, the model assumes that all patterns of co-variation observed in the data are up or down modulations of common motifs when, in truth, there could be additive differences in these patterns as well. Additionally, in a biological context, it is much more likely that the columns of Q would have sparse, overlapping support, rather than the non-overlapping sparse scenarios investigated here. In the case where the conditions do not all have the same number of factors, the model will still be suitable, but one would need to be careful about the interpretation of the results. For example, an experimental condition having 4 factors, while the remaining ones only have have 2 could indicate that that dataset is more variable than the others and should be treated with care.

CHAPTER III

Integrative data-driven module discovery of metabolic perturbations induced by diet

3.1 Introduction

Advancements in mass spectrometry, coupled with the development of more sophisticated data processing tools and comprehensive spectral libraries, have enabled researchers to probe deeper into the metabolome and lipodome. This has allowed new insights into biological mechanisms and disease progression, such as creating measurements of internal body time [58], integrating metabolomics and genomics to identify features associated with poor prognosis in neuroendocrine cancers [59], or illuminating metabolic profiles predictive of future diabetes [60].

While primary and secondary metabolites are well mapped to canonical pathways (such as KEGG [28]), and this prior knowledge can be leveraged for biological insights, the lipidome remains essentially unannotated in this way. Metabolites in the blood have notable interconnectedness which can result in substantial co-expression. This holds true both for metabolites sequentially formed and consumed in metabolic pathways, and metabolites that are metabolized by the same enzymes. The high correlation of various lipids within and across lipid classes is due to both step-wise metabolic anabolism and multiple lipid species competing as substrate. Because

metabolomics analysis of small polar compounds and lipidomics are often done on separate platforms, integration of these data sets is important to gain broader insights into how a physiological or disease state affects the interconnected domains of intermediary metabolism. The presence of unannotated features in the "untargeted" metabolomics profiles also limits the utilization of all information derived from mass spectrometry analyses [12]. Recent studies have demonstrated that networks beyond those defined by canonical pathways can provide novel biological insights [2, 61, 62]. These factors make network based methods particularly attractive in studying Omics data in general, and metabolomics data in particular. Recent applications of network based methods include detecting novel candidate drivers in melanoma [63], illuminating a link between pre-existing cellular phenotype and certain genetic alterations in glioblastoma [64] and identifying relevant unannotated compounds in a longitudinal study of women's aging [10].

A number of techniques are available in the literature to estimate networks from data, including correlation based methods [6, 7, 8] and partial correlation ones [9, 10, 11]. The former are straight forward to calculate, but focus on highly connected compounds which may not be particularly informative, being potentially driven by artifacts [12], while also not differentiating between direct and indirect interactions between compounds. Partial correlations have been used extensively in Omics settings, but require large sample sizes to calculate (see discussion in [10].) A common analytic pipeline leveraging such data driven networks is to first calculate (partial) correlation networks from metabolomics (or other Omics) profiles and subsequently extract strongly connected components from them. These components are then considered as sets and examined for enrichment using one of the numerous methods available in the literature [65, 66, 25].

In this paper, we employ the PCST algorithm that *integrates* the above two-step process and *simultaneously* considers differential expression/abundance and interac-

tions between compounds. This algorithm requires as input a network (nodes and their corresponding edge set) together with a cost associated with each edge and a prize associated with a subset of the nodes. Its output comprises of a tree that maximizes profit (total prizes - total cost). It has been successfully applied to identify relevant modules (subnetworks) across a range of Omics data types. For example, Bailly-Bechet *et al.* [67] used it to uncover the role of a previously uncharacterized protein, leveraging a yeast protein interaction network (edge costs) combined with p-values obtained from testing for differential expression across experimental conditions related to response to yeast (prizes). A human protein interaction network was also employed in Balbin *et al.* [68] together with differential abundance (prizes) between non-small cell lung cancer phenotypes in transcript, protein and phosphoprotein datasets to help discover a protein in the KRAS pathway that could serve as a drug target. More recently Pirhaji *et al.* [69] used a network of protein-protein and protein-metabolite interactions, constructed from integrating multiple databases, while the prizes reflected the significance of a metabolomic feature's disregulation between conditions in a Huntington's disease cell-line model.

This brief literature overview shows that the PCST is used in conjunction with a network of physical interactions (protein-protein, protein-DNA, protein-metabolite) obtained from curated biological databases. However, as argued in Creixell *et al.* [2] such interactions may not reflect the current physiological state, since different physiological or pathological conditions may directly or indirectly alter the interactions, limiting the technique's potential. Indeed, in earlier studies we found that the relative correlations of metabolite can be affected in individuals prone to type 2 diabetes [10].

Metabolites in the blood have notable interconnectedness which can result in substantial co-expression patterns [70]. This arises from the metabolites sequentially formed and consumed in metabolic pathways, and metabolites that are metabolized by the same enzymes. A high correlation of various lipids species, within and across

lipid classes, is due to both to step wise metabolic anabolism and multiple lipid species competing as substrate. For example, glycerol-3-phosphate acyltransferase (GPAT) enzymes catalyze the conversion of multiple diacylglycerol species to triacylglycerol [71] resulting in high co-expression of diacylglycerol with triacylglycerols in plasma. Increased dietary carbohydrates (CHO) intake result in increase in de novo lipogenesis of saturated and monounsaturated fatty acids [72] while polyunsaturated fatty acids (PUFA) from the diet correlate with fat intake [73]. These correlations extend to a variety of other metabolites such as amino acids and acylcarnitines [10].

In the present study, we performed a controlled feeding study to assess the alterations in the metabolome as a first step in developing an objective assessment of dietary intake. Our goal is to identify modules of co-varying lipids and small polar molecules by analyzing the results of LC-MS based lipidomic and untargeted metabolomics profiles, integrating these platforms to gain insights into both the metabolic pathways that are affected by the diets and the metabolite signature of each diet.

In the application, we exploit the high degree of interconnectedness amongst molecular entities in metabolomics datasets, especially those derived from lipidomics profiling and take an agnostic approach using networks with data driven edge weights, instead of those based on physical interactions obtained from curated biological databases. We combine these networks with node prizes based on differential abundance and apply the PCST to identify biologically relevant modules in the plasma metabolome from individuals fed two different diets, one high on polyunsaturated fats (PUFA) and another high on carbohydrates (CHO). The proposed approach yields insights into both the metabolic pathways that are affected by the diets and the metabolite signature of each diet. We further show that identified modules can be coordinately altered in the plasma of an animal model of altered fatty acid oxidation, suggesting that the intimate association of metabolites may provide a statistically

tractable way of identifying variable interactions amongst metabolites that provides both biomarkers for dietary intake, but also may provide insights into alterations in underlying physiological processes in physiological or disease states.

3.2 Integration Methodology

The Prize Collecting Steiner Tree algorithm [74] is employed to integrate, in a condition specific manner, lipidomic and untargeted metabolomic data sets, with a focus on lipids and metabolites that are significantly changed across dietary conditions. As mentioned in the introductory Section, the PCST algorithm requires as input an undirected network $G(V, E, c(e), p(v))$, with node set V, and edge set E. The function $p(v)$ assigns a prize, $p(v) \geq 0$ to each node $v \in V$, and nodes with $p(v) > 0$ are referred to as *terminal* nodes. The function $c(e) > 0$ assigns a cost to each $e \in E$. The aim is to find a tree $T(V_T, E_T)$ that maximizes the objective function:

$$profit(T) = \sum\nolimits_{v \in V_t} \beta p(v) - \sum\nolimits_{e \in E_t} c(e) \tag{3.1}$$

Nodes with $p(v) = 0$ which are returned in V_T as part of the solution will be referred to as *Steiner* nodes.

Effectively the PCST algorithm is used to sparsify a condition-specific co-expression network, where the latter is sparsified *around* nodes of interest (often those which show differences between experimental conditions or groups). Further, in a setting where the co-expression network itself is substantially different between conditions, such as the feeding study, our method enables researchers to utilize this condition-specific connectivity information to generate hypothesis about systems level differences in behavior.

To ensure robust results, the algorithm is run multiple times while applying a

small amount of noise to the cost function. The results are then used to create consensus graphs in which the approach identifies modules of metabolites that are strongly linked by their patterns of co-variation. The resulting stable modules can then by characterized by testing for enrichment or depletion in relevant characteristics and compared across conditions in order to obtain a ranking of their contribution. This workflow is outlined graphically in Figure 3.1.

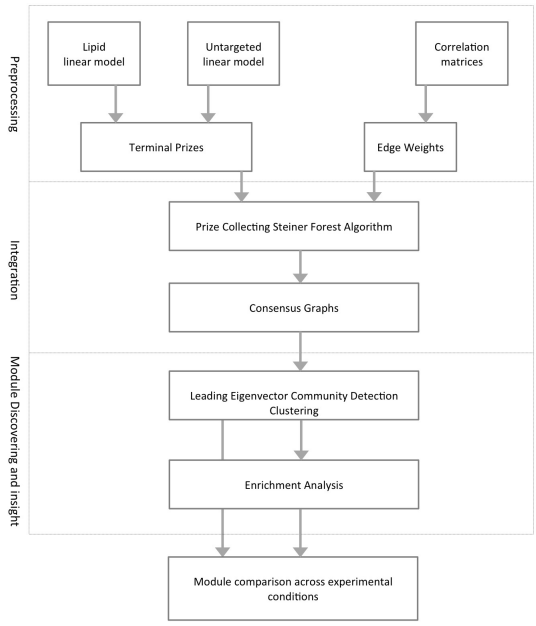


Figure 3.1: *Method Schematic:* Differentially abundant lipids and untargeted metabolites are designated as terminal nodes and assigned prizes. Condition specific correlation matrices are used to calculate edge weights. The Prize Collecting Steiner Tree Algorithm is run with noise over an undirected network as described in main text. Consensus graphs are created from PCST output; modules are identified in these graphs. Modules are tested for enrichment in relevant characteristics and compared across experimental conditions.

In the discovery phase of this work, we analyze data from a controlled feeding study on humans. Twelve healthy adults (6M/6F) were fed a diet high in polyunsaturated fatty acids (PUFA) for 3 weeks, immediately followed by a diet high in carbohydrates (CHO) for 3 weeks. Each subject's plasma lipidome was assayed at days 0, 2, 7, 21,

23, 28, and 42, while their plasma metabolome was assayed at days 0, 21 and 42. For additional details see Methods.

3.2.1 Prize Collecting Steiner Tree algorithm

The PCST algorithm employed is implemented in the OmicsIntegrator software [75] (local copy, downloaded on 2/2/2017) which solves a modified version of the Prize Collecting Steiner Forest (PCSF) problem. Formulated similarly to the PCST problem, the PCSF allows the solution set to be comprised of disjoint trees. OmicsIntegrator further incorporates additional tuning parameters into the objective function, most of which we set to their default values. For complete details of the formulation, see Tuncbag *et al.* [75].

We supplied the node and edge sets, edge weights, and terminal prizes (detailed below), using *forest.py* with the *forest-only* and *doRandom* options. We modified the original *forest.py* file so that betweenness was not calculated when merging the results from noisy PCST runs. Forest parameters were as follows: $w = 4$ (controls the number of trees), $b = 20$ (controls the trade off between including more terminals and using less reliable edges), $D = 10$ (controls maximum depth of trees), processes= 1 (number of processors to spawn when doing randomization runs), threads= 2 (number of threads to use in optimization algorithm), noise= 0.005 (the standard deviation of the Gaussian noise added to edge costs). All other parameters were set to their default values. For robustness, the algorithm was run 50 times over the same set of terminal nodes and prizes for each network, with a small amount of noise added to the edge weights each time.

The four networks considered - labeled as

$$G_{21u}(V, E_{21}, c(e), p_{21u}(v)), \quad G_{42u}(V, E_{42}, c(e), p_{42u}(v))$$
$$G_{21d}(V, E_{21}, c(e), p_{21d}(v)), \quad G_{42d}(V, E_{42}, c(e), p_{42d}(v))$$

- are described fully below. Each network has the same node sets and cost function, while the edge weights and prize functions differed.

3.2.2 Terminal prizes

To determine terminal nodes and the corresponding prizes, we used a linear model to test for differences in abundance levels between any two time-points assayed. We modeled the abundance of the k^{th} lipid (or metabolite) L_k as

$$L_k = d + u \tag{3.2}$$

where d is a factor, with levels, d_i, representing each day assayed in the study ($d_i, i \in \{0, 2, 7, 21, 23, 28, 42\}$ for lipids or $i \in \{0, 21, 42\}$ for untargeted metabolites.) The random effects u for the n^{th} subject on day k of diet j are specified as: $u_n + u_{nj} + u_{nk}$; note that d_0, d_2, d_7, d_{21} are classified as diet 1, PUFA; the remainder of the days are classified as diet 2, CHO.) TThis specification allows us to account for overall subject level variation, as well as differences in the way each subject processes each dietary intervention across time.

Differential abundance between any two levels i and j of d were then tested using contrast vectors based on the R-language implementation in the `lme4` and `pbkrtest` packages. Formally, the hypothesis of interest is given by

$$\begin{aligned} H_0 &: d_i - d_j = 0 \\ H_\alpha &: d_i - d_j \neq 0 \end{aligned} \tag{3.3}$$

While we included all of the data in our linear model, our analysis focused on the dynamics between days 0, 21 and 42 as these time points captured the most meaningful differences between dietary interventions.

We identified three different terminal node sets of interest: $21d$, the lipids and

metabolites for which there is a significant difference between d_0 and d_{21}; $42d$, the lipids and metabolites for which there is a significant difference between d_{21} and d_{42}; and $21u = 42u$, the union of those two sets. Significance was defined for lipids as a False Discovery Rate corrected (via the Benjamini-Hochberg procedure) p-value of $p < 0.1$; for untargeted metabolites a p-value of $p < 0.2$. Let p_{0v21} denote the FDR-corrected p-value from testing $d_0 - d_{21} = 0$, and p_{21v42} be the corresponding p-value for testing $d_{21} - d_{42}$. Terminal prizes for $21u$ and $42u$ were set to $p_{21u}(v) = p_{42u}(v) = -log(\min(p_{0v21}, p_{21v42}))$. For $21d$, $p_{21d}(v) = -log(p_{0v21})$ and for $42d$, $p_{42d}(v) = -log(p_{21v42})$.

3.2.3 Network Edge Weights

Edge weights for the diet specific networks were generated by applying Fisher's z-transformation [76] to the 603×603 matrix of correlations between all lipids and metabolites at a single time point. We employ Fisher's z-transformation to test whether correlation amongst lipids and/or metabolites is zero or not. Specifically, letting r denote the sample correlation between two lipids, then $z(r) = \frac{1}{2}\log(\frac{1+r}{1-r}) = \text{arctanh}(r)$. Further, if the measurements of the two lipids are assumed to be independent and identically distributed from a bivariate normal distribution with correlation ρ, then $z(r) \sim N(\frac{1}{2}\log(\frac{1+\rho}{1-\rho}), \frac{1}{\sqrt{N-3}})$, where N denotes the corresponding sample size. This allows us to test for correlations that are significantly different from 0.

In our implementation of the OmicsIntegrator, we set edge weights to $w(e) = 1 - p_e$, where p_e is the FDR-corrected p-value from testing the null hypothesis $H_0 : z_e = 0$ for the pair of lipids defined by edge e, versus the alternative of being different than 0. Hence, the corresponding edge cost is set to $c(e) = p_e$, so that edges with more significant p-values are more likely to be included in the solution set. Further, the sample correlation matrix at day 21 was used to create E_{21}, while E_{42} was created from the sample correlation matrix at day 42; these correlation coefficients are then

converted to z values using the above transformation and tested for significance, yielding p-values for the edge costs.

3.2.4 Consensus graphs, module discovery, enrichment analysis

Inspired by the consensus clustering approach [77], we created consensus graphs from the PCST results on each graph by selecting nodes which appeared in $> 80\%$ of all noisy runs. Any edge chosen by the algorithm to connect these nodes, regardless of how frequently that edge appeared in the solution set, was included in the consensus graph. We we refer to the consensus graph from the results on G_i as CG_i.

Modules in each consensus graph were identified via the leading eigenvector community detection algorithm [46] (LEVCD), which identifies highly connected subgroups/modules within a larger network. For this clustering step, an edge between nodes i and j was given weight equal to $|r_{ij}|$, where r_{ij} is the sample correlation between the nodes at the relevant time point. The k^{th} module from consensus graph CG_i is labeled $M_{i:k}$. These modules were then tested for enrichment in relevant characteristics using the hypergeometric test. The hypergeometric test indicates whether, in a given module, there are more lipids/metabolites with a certain characteristic than one would expect by chance (in which case the module is enriched in said characteristic), or if there are fewer than one would expect (in which case the module is depleted.) (see Cao and Zhang [78] for a good review on the hypergeometric test).

Finally, percent density for the consensus graphs is calculated as:

$$D = 100\frac{2|E|}{|V|(|V|-1)}$$

3.2.5 Validation in Rat data

The discovered modules based on the data from the human feeding study were subsequently assessed for concordance and biological relevance in metabolomic and

lipid profiles obtained from the rat model (see details below). For the comparison, we first identified modules in which at least 70% of the module's lipids/metabolites are also present in the rat data. These modules were then tested for differential abundance between pairs of conditions using the Gene Set Analysis (GSA) approach [24] implemented in the R package `GSA`. We used the maxmean method, with $s0 = 0$, and without restandardization. The GSA method [24] is a more powerful and robust version of the popular Gene Set Enrichment Analysis (GSEA) procedure [23]. The maxmean statistic used in GSA can detect subtle, concordant changes in a group of biomolecules across a wide range of settings.

3.3 Results

3.3.1 Method Results

To create our list of terminal nodes, we used a linear model to test for DA across the course of each dietary intervention. Lipids and untargeted metabolites were modeled and tested separately, as the untargeted metabolomics assay was only run on days 0, 21 and 42.

Lipids showed high levels of activity across all classes - 180 lipids exhibited statistically significant changes from day 0 to day 21, and 222 from day 21 to day 42. Out of these, 130 lipids exhibited statistically significant differences under both diets (FDR adjusted p-value < 0.1.) Untargeted metabolites showed much lower levels of activity overall - out of 147 metabolites, only 9 showed statistically significant differences under PUFA, and 10 changed under CHO (FDR adjusted p-value < 0.2). Finally, no metabolites were differentially expressed under both diets. These cutoffs gave G_{21u} and G_{42u} 291 terminal nodes, while G_{21d} and G_{42d} had 189 and 232 ones, respectively.

As previously mentioned, the PCSF algorithm was applied on the following four

Graph	Nodes in CG_i	Edges in CG_i	# of modules	# of modules with >8 members
G_{21d}	224	619	21	8
G_{42d}	291	1475	16	6
G_{21u}	323	967	20	7
G_{42u}	346	1668	16	9

Table 3.1: Summary of results from PCST and module discovery steps

networks: G_{21u}, G_{42u}, G_{21d}, and G_{42d}. In each case, it always returned the terminal nodes and a fairly stable subset of Steiner nodes (Appendix Table B.1). We chose to include Steiner nodes which occurred in $> 80\%$ of the solutions in the consensus graphs. The edge sets were much more variable - relatively few edges were chosen every time, with many more "moderately often" chosen edges (Appendix Table B.1.) This variability led us to include in the consensus graphs all edges chosen by the algorithm which connected our terminal and selected Steiner nodes.

The resulting consensus graphs are overall quite sparse - only 1.86% dense, 2.79% dense, 2.52% dense and 3.54% dense for CG_{21u}, CG_{42u}, CG_{21d}, and CG_{42d} respectively. The Leading Eigenvector Community Detection (LEVCD) algorithm identifies between 16 and 21 modules in each consensus graph (Table 3.1); the j^{th} module from CG_i is referred to as $M_{i:j}$ (i.e. $M_{21u:2}$ refers to the second identified module from CG_{21u}.) Between 6 and 9 modules in each consensus graph had more than 8 members. These larger modules were tested for enrichment in the observed lipid classes, saturation levels, and other relevant characteristics. The enrichment and depletion results for select modules are summarized in Table 3.2, with complete results for all modules in CG_{21u} and CG_{42u} available in Supplemental Table A4.

In general, the untargeted metabolites are fairly well integrated with lipids in the identified submodules. Although many modules do not contain any metabolites, those that do, also contain lipids from a range of lipid classes. Hence, we conclude that the approach successfully integrates lipids and metabolites and does not segregate the two classes of biomolecules into their own modules ((see Appendix C)

Module	CE	DG	lysoPC	lysoPE	MG	PC	PE	plasmenyl-PC	plasmenyl-PE	SM	TG	SFA	MUFA	PUFA
$M_{21u:2}$	0.795	0.106	0.921	0.728	0.595	0.338	0.886	0.523	0.928	0.859	0.000	0.426	0.008	0.366
$M_{21u:3}$	0.648	0.000	0.921	0.627	0.595	0.558	0.886	0.523	0.928	0.404	0.137	0.977	0.922	0.004
$M_{21u:4}$	0.441	0.707	0.000	0.361	0.057	0.251	0.886	0.523	0.860	0.429	0.990	0.009	0.017	0.720
$M_{21u:6}$	0.441	0.735	0.921	0.465	0.282	0.595	0.886	0.523	0.860	0.810	0.000	0.856	0.922	0.001
$M_{21u:20}$	0.211	0.953	0.921	0.465	0.208	0.131	0.593	0.612	0.000	0.404	0.990	0.856	0.253	0.002
$M_{42u:2}$	0.485	0.000	0.920	0.602	0.567	0.923	0.950	0.790	0.816	0.965	0.000	0.860	0.192	0.009
$M_{42u:5}$	0.485	0.225	0.755	0.602	0.192	0.923	0.950	0.320	0.685	0.965	0.000	0.906	0.472	0.000
$M_{42u:14}$	0.383	0.958	0.000	0.602	0.192	0.923	0.950	0.394	0.224	0.965	0.097	0.823	0.609	0.001
$M_{42u:15}$	0.485	0.958	0.055	0.000	0.192	0.644	0.950	0.175	0.816	0.965	0.998	0.469	0.472	0.960
$M_{42u:16}$	0.712	0.958	0.755	0.602	0.567	0.605	0.950	0.790	0.816	0.965	0.000	0.457	0.245	0.029

Table 3.2: *Enrichment analysis of identified modules.* Selected modules were tested for enrichment in notable classes and for saturated fatty acids (SFA), monounsaturated fatty acids (MUFA) and polyunsaturated fatty acids (PUFA). *P*-values are corrected column-wise for multiple testing using the Benjamini-Hochberg False Discovery Rate adjustment procedure [1]. Results for all modules in CG_{21u} and CG_{42u} available in Supplemental Table A4.

Comparison tested	$M_{21u:2}$	$M_{21u:3}$	$M_{21u:4}$	$M_{21u:6}$	$M_{21u:20}$	$M_{42u:2}$	$M_{42u:5}$	$M_{42u:14}$	$M_{42u:15}$	$M_{42u:16}$
$d_{21} < d_0$	0.000	0.000	0.000	0.997	0.000	0.000	0.032	0.000	0.000	0.000
$d_{42} < d_0$	0.993	0.993	0.013	0.993	0.013	0.993	0.993	0.061	0.027	0.993
$d_{42} < d_{21}$	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
$d_{21} > d_0$	1.000	1.000	1.000	0.053	1.000	1.000	1.000	1.000	1.000	1.000
$d_{42} > d_0$	0.107	0.018	0.993	0.200	0.993	0.018	0.064	0.993	0.993	0.190
$d_{42} > d_{21}$	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 3.3: *Differential abundance in modules from CG_{21u} and CG_{42u}* tested via GSA. End of PUFA (d_{21}) and end of CHO (d_{42}) tested against baseline (d_0), and each other. Values presented are fdr adjusted *p*-values for a subset of modules. Results from all modules in CG_{21u} and CG_{42u} available in Supplemental Table A2.

We primarily focus our analysis on a subset of modules from CG_{21u} and CG_{42u}.

3.3.2 Module metabolite participants and dynamics under different dietary conditions

We examined the properties of the metabolites that were identified as covarying under the two dietary conditions. Following consumption of a PUFA diet for 21 days, several modules were identified that were significantly enriched in some metabolite classes (Table 3.2). There was a large overlap between $M_{21d:3}$, $M_{21u:2}$, $M_{42u:16}$ (Appendix C), indicating that these metabolites are dynamically changed under each feeding condition. We focus first on $M_{21u:2}$, which was overall decreased by PUFA

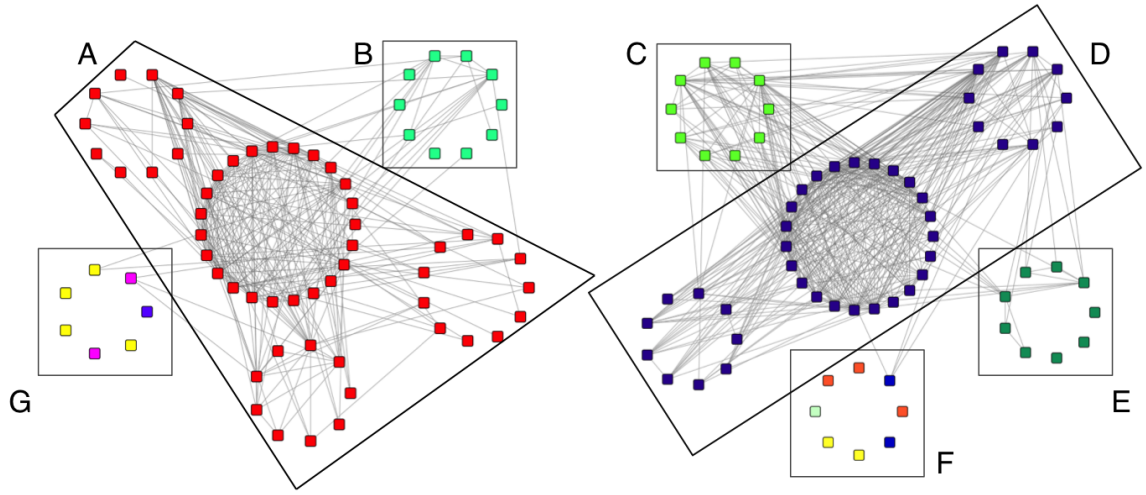


Figure 3.2: *All nodes present in either* $M_{21u:2}$ *or* $M_{42u:16}$. Nodes in left and right halves of image arranged identically. Edges on left taken from $CG_{21u:1}$, edges on right taken from $CG_{42u:16}$. Only edges connecting nodes within the figure are shown. Nodes colored by module label from $CG_{21:u}$ (left) and $CG_{42:16}$ (right). Clockwise from upper left, A: $M_{21u:2}$, B: Nodes in $M_{21:3}$ that are brought in as a group when transitioning to CHO, C: Subset of nodes shed as a group into $M_{42u:2}$ when transitioning to CHO D: $M_{42u:16}$, E: Subset of nodes shed as a group into $M_{42u:4}$ when transitioning to CHO, F: Nodes which are shed into many different modules when transitioning to CHO, G: Nodes with varying module memberships under PUFA, incorporated into $M_{42u:16}$

feeding and increased following CHO diet (select results in Table 3.3, complete results in Supplemental Table A2). The module was enriched in PCs and triglycerides (TGs), which we observe to be primarily shorter chain species, with relatively few saturated fatty acids (Table 3.2), likely reflecting the increase in de novo lipogenesis seen after consumption of a carbohydrate diet. This module contained 2 untargeted metabolites, carnitine and 3-hydroxy-3-methylbutryric acid. The latter is generated during leucine degradation and is found elevated in individuals with insulin resistance and diabetes [79]. Leucine was not associated with a module under either dietary condition. However, it has been previously observed that carbohydrate can reduced the oxidation of leucine [80] and dietary carbohydrates also reduce leucine catabolism [81], suggesting that the higher levels of its intermediate breakdown product is due to

Comparison tested	$M_{21u:3}$	$M_{21u:4}$	$M_{21u:6}$	$M_{21u:20}$	$M_{42u:2}$	$M_{42u:5}$	$M_{42u:14}$	$M_{42u:15}$	$M_{42u:16}$
HCR-AL < LCR-AL	0.011	0.025	0.108	0.000	0.090	0.089	0.040	0.011	0.170
HCR-CR < LCR-CR	0.381	0.278	0.278	0.493	0.278	0.278	0.278	0.278	0.493
LCR-CR < LCR-AL	0.000	0.000	0.033	0.013	0.000	0.017	0.000	0.063	0.033
HCR-CR < HCR-AL	0.213	0.213	0.213	0.870	0.213	0.213	0.200	0.748	0.213
HCR-AL < LCR-CR	0.853	0.511	0.853	0.483	0.853	0.783	0.853	0.133	0.853

Table 3.4: *Module dynamics in animal model.* Modules from CG_{21u} and CG_{42u} which had at least 70% overlap with animal data were tested for differential abundance in animal data using GSA. None of the tests in the opposite direction (HCR-AL > LCR-AL, HCR-CR > LCR-CR, etc) were significant. Values presented are fdr adjusted p-values for a subset of modules. Results from all modules in CG_{21u} and CG_{42u} available in Supplemental Table A3.

inhibition of the later part of leucine catabolism (see Discussion). Interestingly, the single metabolite whose levels covaried inversely to the others in $M_{21u:2}$ was carnitine. A previous study observed that carnitine levels rose significantly on a high fat diet compared to a high carbohydrate diet [82], suggesting that the increased oxidation of carbohydrates results in greater consumption of carnitine.

The majority of TG species in $M_{21u:2}$ overlap substantially with $M_{42u:16}$; the latter also being highly enriched in TGs (Figure 3.2). The 23 overlapping lipids are almost exclusively shorter chain, relatively saturated TGs (Figure 3.2, box A). The retention of a highly connected subset of TGs suggests that these lipids are generated and consumed in a highly parallel manner, allowing them to be maintained in synchronous plasma levels despite varying in concentration.

As module $M_{21u:2}$ transitions to module $M_{42u:16}$ following CHO feeding, longer polyunsaturated DGs and TGs, as well as some saturated PCs and TGs are shed (removed from the module) (Figure 3.2, boxes C and E). These lipids 'join' two large modules $M_{42u:2}$ and $M_{42u:4}$, the former enriched in DGs and TGs and the latter enriched in PCs (Table 3.2). These lipids were replaced by some shorter, mostly low saturated chain PCs and interestingly, some short, unsaturated PCs and TGs (PC 38:8, PC 40:4, TG 51:4, TG 52:7) (Figure 3.2, box B). Also shed were the two untargeted metabolites. The addition of the new lipid species in the module after

CHO diet suggests that these are newly formed lipids which have incorporate de novo generated fatty acids into a polyunsaturated lipid. A variety of other nodes have varying memberships after leaving $M_{21u:2}$ (Figure 3.2, box F) or before joining $M_{42u:16}$(Figure 3.2, box G).

Modules $M_{21u:4}$ and $M_{42u:14}$ were of interest as they were both enriched only in LPCs in which nearly half showed reciprocal changes under the different dietary conditions (Appendix C),. Module $M_{42d:7}$ was also enriched in these LPCs. Paradoxically, PUFA-diet associated Module $M_{21u:4}$ was enriched in saturated and monounsaturated fatty acids while the CHO-associated $M_{42u:14}$ was statistically enriched in polyunsaturated fatty acids. Plasma LPC is thought to be derived from phosphatidylcholine in lipoproteins acyltransferases and phospholipases [83]. LPC as well as its metabolite, lysophosphatidic Acid (LPA, generated by the removal of the choline headgroup), can differentially signal through cell surface receptors by depending on chain length and degree of saturation [84]. No specific pattern of increase and decrease in the levels of other LPCs in the data set were found; we do note that PCs populated Module $M_{42u:14}$, though it was not significantly enriched in this class. Module $M_{21u:4}$ also contained 4 notable untargeted metabolites, guanosine and threonine as well as palmitoylcarnitine decreased between baseline and d_{21}, while cholesterol increased over the same time. As with leucine, higher levels of threonine are found in insulin resistance and are reduced by weight loss [85, 86]. The reduction in palmitoylcarnitine is likely due to reduced entry of palmitic acid into mitochondria during PUFA feeding. These metabolites were shed in $M_{42u:14}$. Further exploration of these modules will be necessary to understand its unique behavior, but we note that the modules were also perturbed in the rat model (see discussion below).

We also recognized $M_{42u:15}$ as being unique. This 33 member module, visualized in Figure 3.3, was enriched in LPE and was devoid of triglycerides (Appendix C). The module contains long-chain acylcarnitines, derivatives of saturated and mononunsatu-

rated fatty acids, which decrease in PUFA diet as well as medium chain acylcarnitines, which increase in PUFA diet. The lower abundance of long-chain acylcarnitines suggests reduced entry of saturated and monounsaturated fatty acids into the mitochondria [87], potentially through competition by polyunsaturated fatty acids. This is supported by the observed increase in levels of octenoylcarnine, which is derived from polyunsaturated omega-6 fatty acids. In addition, $M_{42u:15}$ contained a number of acetylated amino acids. There is a dearth of literature on n-acetylated amino acids outside of N-acetylaspartate (NAA), which an abundant metabolite in the brain and is synthesized enzymatically from aspartate and acetyl-CoA in neurons [88]. We note that NAA was only found in $M_{42d:3}$, which also increases following the CHO diet. Non-enzymatic acetylation of lysine residues in the mitochondrial is well described [89] and it is tempting to suggest that under higher carbohydrate flux, excess acetyl-CoA generated in the mitochondria may modify amino acids. Finally, kynurenine, a product of mitochondrial tryptophan metabolism, is also associated with $M_{42u:15}$. Like many of the other metabolites found associated with modules elevated in the plasma following CHO diet, kynurenine (KYN) and its direct precursor tryptophan (TRP) are elevated in the plasma of insulin resistant individuals [90, 91]. However, unlike the other metabolites, KYN terminal metabolic fate is not oxidation, rather is used for niacin biosynthesis. The reason for an increase in kynurenine in insulin resistance has been attributed to increased expression of indoleamine 2,3-dioxygenase 1 [IDO1] by chronic inflammation [90, 92], increasing the conversion of TRP to KYN. The finding that KYN is associated with other presumptively mitochondrially generated metabolites could suggest that alteration in KYN mitochondrial metabolism may underlie its association with insulin resistance.

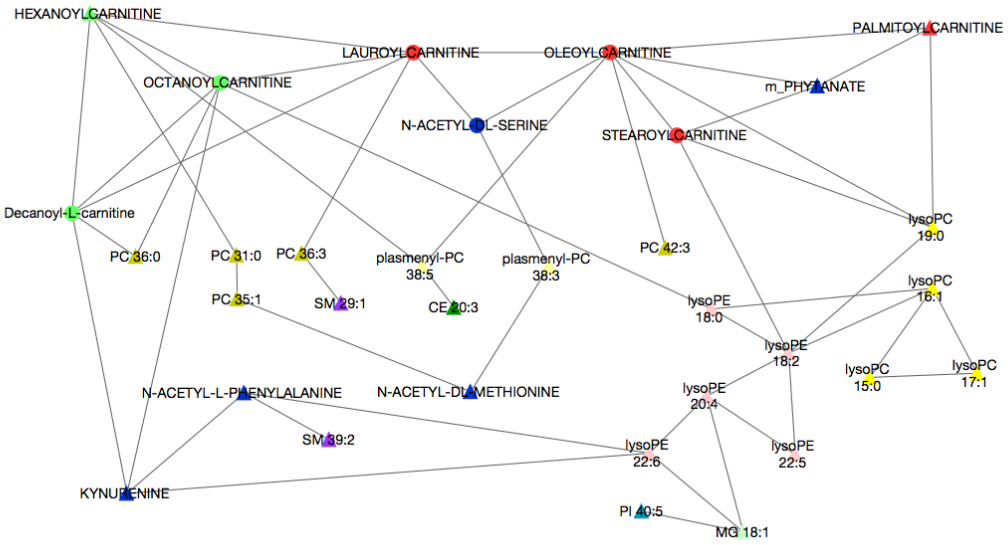


Figure 3.3: *Visualization of* $M_{42u:15}$ Nodes are colored according to their classes - most notably, the medium chain acylcarnitines in bright green and the long chain acylcarnitines in red. Terminal nodes are represented as triangles while steiner nodes (those which are not differentially abundant, but are brought into the consensus graph via their associations with differentially abundant nodes) are represented by circles. Note that only two acylcarnitines, Hexanoylcarnitine and Palmitoylcarnitine are terminal nodes; the other acylcarnitines in this module are brought in by their associations with these two nodes. Out of 11 acylcarnitines in the entire untargeted metabolomics dataset, 7 are incorporated into this module, with the remainder not being incorporated into any module from any consensus graph.

3.3.3 Module dynamics in animal model of differential metabolite utilization

The above results suggest that modules of plasma metabolites and lipids can be identified that change in a concordant manner following dietary changes. To answer the question as to whether these metabolite/lipid modules related under divergent dietary conditions, remain associated in different biological contexts, we examined the identified modules in a rat model which demonstrates differences in the utilization of fatty acids and carbohydrates at rest and during acute exercise [93]. Two lines of rats,

high capacity runners (HCR) and low capacity runners (LCR), were selectively bred for high and low intrinsic exercise capacity based on untrained aerobic capacity to create a model for studying aerobic exercise and its relationship to metabolic health [93]. We have shown that the HCR, compared to the LCR rats, have a higher fatty acid utilization, both at rest and during exercise. Specifically, the enhanced fatty acid utilization appears to underlie the enhanced running capacity of the HCR. An apparent inefficiency in the oxidation of fatty acids likely plays a role in the reduced weight gain observed in the HCR line [93], and may also play a role in the increase in lifespan in the HCR compared to the LCR [94]. Thus, to test whether the plasma metabolome would respond to alterations in fuel utilization (fatty acids v. CHO), as opposed to alterations in fuel input through dietary consumption, we identified PCST-identified human modules which had more than 70% of their members observed in the rat model. These modules were tested for differential abundance in the rat plasma using the GSA pathway enrichment methodology [24]. In addition, we also assess enrichment in HCR and LCR rat plasma following a 12-month period of caloric restriction (CR) to assess whether the lines had a similar metabolomic response.

Few plasma metabolites show differential levels between the HCR and LCR in either the ad lib fed or CR state. Only 49 lipids/metabolites were differentially abundant between LCR-AL and LCR-CR, having an fdr-corrected p-value ¡ 0.1. Between LCR-AL and HCR-AL, 26 lipids/metabolites were differentially abundant with the same cutoff. No lipids/metabolites were differentially abundant between any other 2-way comparison of interest. Each module found in the rats, however, was lower in the HCR (Table 3.4) and among the comparisons, the most statistically significant changes were seen in LCR-CR compared to the LCR-AL followed by HCR-AL compared to HCR-AL. The modules identified in the rat comparison are reduced in the plasma of PUFA diet and increased in the CHO diet. This may suggests that the increased utilization of fatty acids as fuel (or reduction of CHO utilization), is what

is impacting the levels of the modules.

Of considerable interest is the finding that none of the modules were changed between the HCR-CR compared to HCR-AL or the LCR-CR, suggesting that the perturbation in metabolism induced by caloric restriction "preexists" in the HCR such that there is not further change following CR. Importantly, as is found in caloric restriction in rodents, HCR rats have an extended lifespan in the ad libitum fed state [94], suggesting biological mimicry in the enhanced utilization of fatty acids in both high oxidative capacity and caloric restriction, both associated with improved health and longevity humans [95].

3.4 Discussion

In this paper we present a data-driven method for integrative analysis of lipids and untargeted metabolites. We focused on named untargeted metabolites as a first step, but the method can readily be extended to unnamed untargeted metabolites, or applied to other Omics settings where one wishes to identify modules outside of canonical pathways. Our method identifies modules of biologically relevant biomolecules, centered around those which are differentially abundant between conditions, while incorporating related biomolecules which did not reach a significance cutoff. Many potentially interesting modules were identified beyond the most biologically interesting ones highlighted here. While competing methods based only on the correlation matrices uncover some of the same features, our method identifies additional subtle features that other methods do not.

In our controlled human dietary intervention, each of the dietary feeding periods caused a significant change in the levels of multiple lipid species. Many were expected due to the anticipated influx of polyunsaturated fatty acids into lipids during the PUFA diet and a reduction in saturated and monounsaturated fatty acids; the converse happening during CHO diet. Additional Steiner nodes were identified

in each condition through the addition of lipids and untargeted metabolites due to their consistent interactions with the terminal nodes in each module. Though few of the metabolites detected in the untargeted platform were significantly different in transitioning from baseline to PUFA to CHO, the metabolites added to the modules using the data-driven PCST algorithm were both biologically consistent with previous observations and provided some potentially novel insights into alterations in whole body metabolism in people under fat and carbohydrate feeding.

The changes in the untargeted metabolites in modules identified via CG_{21u} and CG_{42u} were consistent with changes in handling of metabolites in the mitochondria (Figure 3.4). Our findings are consistent with increasing carbohydrate utilization following the change from the high fat PUFA diet to CHO. It is known that a relative increase in flux of glucose-derived pyruvate into the mitochondria increases the levels of acetyl-CoA which leads to increased malonyl-CoA [96] production and reduction of fatty acid uptake through CPT-1 [96, 97], reducing the levels of long- and short-chain acyl-CoAs. Increased acetyl-CoA production will also reduce oxidation of branched chain amino acids [80, 98], reflected in the elevated levels of 3-hydroxy-3-methyglutarate, an intermediate in leucine metabolism (Figure 3.4). Though speculative, the fall in carnitine levels during CHO feeding may be due to utilization in formation of medium chain acylcarnitines in the cell, accumulating in the plasma.

Excess mitochondrial acetyl-CoA is associated with an increase in mitochondrial protein acetylation [98]. Our finding relative increased N-acetylated amino acids in CHO vs. PUFA diet suggests that excess acyl-CoA formed during CHO diet can lead to accumulation of n-acetylation amino acids, potentially making N-acetylated amino acids markers of high carbohydrate intake. Whether this occurs in the mitochondria or in the cytoplasm or whether they are formed enzymatically is unclear. Finding the TRP metabolite KYN in a module with the medium chain acylcarnitines and n-acetyl amino acids is more puzzling. The degradation of TRP is complex and regulated in

large part by Indoleamine 2,3-dioxygenases (IDO) and as mostly been studied in relationship to inflammation-related metabolism [99]. One speculative reason for the finding of increased TRP and KYN in CHO is a an increase in mitochondrial redox, lowering FAD/FADH and reducing the FAD-dependent conversion of KYN to 3-hydroxykynurenine by IDO1. Our proposed model outlined in Figure 3.4 assumes that diet-induced alteration in mitochondrial metabolism, produces a signature in plasma. Studies of plasma following inherited disruption of mitochondrial enzymes support that this is a reasonable assumption [100].

The changes identified in the module metabolites in the HCR/LCR rat model under two different feeding conditions show that even across species, the metabolites identified in humans covary under different physiological states. We note that the NIH31 diet used in the HCR/LCR studies has a very high CHO and low fat content (72.2% and 7.4% of calories, respectively), which is similar to the CHO diet provided in the human studies. Despite this, we observe changes in metabolite modules in the HCR and during calorically restricted feeding that parallel the high PUFA diet. This intriguing result suggests that alterations in the modules may be more related to an increase in fatty acid vs. glucose utilization. Many of the alterations in plasma metabolites that we observe in individuals under a high carbohydrate diet have been seen in insulin resistance. Data supports the idea that a principal defect in people with insulin resistance is an impaired capacity to upregulate muscle lipid oxidation in the face of high fatty acid supply, principally in skeletal muscle [101, 102]. We suggest that the insulin resistance signature may be due to alterations in fuel selection that may be modified by changes in oxidative capacity observed in the HCR/LCR rat model.

We were limited by the relatively small sample size in defining covariant modules, but the human studies were aided by the sequential feeding in the same individuals, reducing inter-individual variations. We also incorporated only annotated metabolites

in the untargeted metabolomics data set in our analysis. Further studies demonstrating association of unannotated metabolites in "modules" in different metabolic states could potentially help in identifying these unknown metabolites.

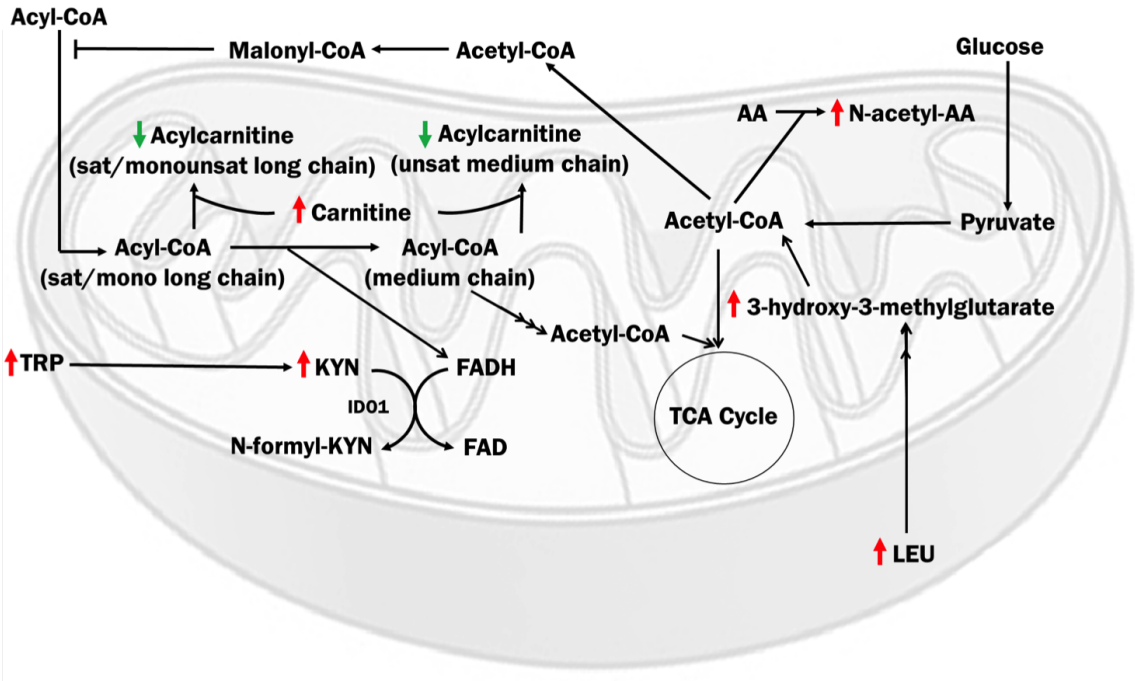


Figure 3.4: *Schematic of mitochondrial metabolism of untargeted metabolites identified by PCST.* Arrows preceding metabolites indicate the direct of change from PUFA to CHO diet. TRP = tryptophan; KYN = kynurenine; LEU = leucine; FAD = flavin adenine dinucleotide; CoA = Coenzyme A; AA = amino acid

3.5 Methods

3.5.1 Dietary intervention

All studies were approved by the Institutional Review Board of the University of Michigan and all participants provided written, informed consent. All methods and procedures were performed in accordance with the relevant guidelines and regulations. Twelve healthy participants (6 women, 6 men, Table 3.5) were provided two sequential isocaloric diets for 21 days each. The first diet was enriched in polyunsaturated fatty

	Females ($n = 6$)	Males ($n = 6$)	p-value
Age (yr)	29.3 ± 7.4	27.17 ± 4.1	0.55
Body mass index (kg/m^2)	21.9 ± 2.6	27.08 ± 4.0	0.02
Cholesterol (mg/dl)	163.3 ± 24.6	157.00 ± 28.9	0.69
Triglyceride (mg/dl)	68.2 ± 66.0	81.50 ± 57.9	0.72
High Density Lipoprotein Cholesterol (mg/dl)	63.7 ± 9.8	54.0 ± 10.9	0.13
Low Density Lipoprotein Cholesterol (mg/d)	92.3 ± 27.8	86.5 ± 30.2	0.74
Glucose (mg/dl)	86. 3 ± 9.8	86.7 ± 5.0	0.94
Homeostatic Measure of Insulin Resistance (HOMA)	2.5 ± 1.0	3.1 ± 1.4	0.40

Table 3.5: *Subject characteristics* Values given are mean ± standard deviation.

acids (PUFA) consisting of $12 - 15\%$ protein, $35 - 50\%$ carbohydrate, $40 - 50\%$ fat ($25 - 30\%$ polyunsaturated fatty acids, $5 - 10\%$ monounsaturated and $< 10\%$ saturated followed immediately by a diet enriched in carbohydrate (CHO) with $10 - 15\%$ protein, $70 - 80\%$ carbohydrate and $10 - 15\%$ fat ($< 10\%$ fat as SFA with 2% polyunsaturated fatty acids). All meals and snacks were provided by the Nutrition Assessment Laboratory (NAL) of Nutrition Obesity Research Center at the University of Michigan. Participants came to the facility at least twice per week during this 6-week controlled feeding trial for food pick-ups and weigh-ins. The diet was adjusted to maintain initial body weight. This resulted in a mean relative weight standard deviation of 0.8% throughout the study period. Fasting blood was drawn in the morning at baseline (Day 0) and at days 2, 7, 21 (PUFA), and 23, 28 and 42 days (CHO) and EDTA plasma was collected and aliquoted for analysis.

3.5.2 HCR/LCR rat model

These studies were approved by the University of Michigan Institutional Animal Care and Use Committee. All methods and procedures were performed in accordance with the relevant guidelines and regulations. In the present study, generation 20 male HCR and 20 LCR were randomly selected to receive a fortified NIH31 diet (Taconic, Rensselaer, NY) ad libitum (ADLIB) or a calorically restricted diet (CR) for 12 months. Plasma was obtained from cardiac puncture between 5 and 6 pm and frozen

	LCR-ADLIB ($n = 11$)	LCR-CR ($n = 12$)	HCR-ADLIB ($n = 10$)	HCR-CR ($n = 11$)
Start weight (gm)	93.8 ± 5.1	93.6 ± 4.9	78.6 ± 4.2	78.8 ± 4.2
End weight (gm)	517.3 ± 15.9	331.3 ± 10.7	388.5 ± 15.2	286.5 ± 8.6
Body weight gain (gm)	440.4 ± 15.4	245.9 ± 9.3	330.7 ± 14.7	208.8 ± 7.4

Table 3.6: *Animal model subject characteristics.* Values given are mean ± standard deviation.

at -80°C until analysis. Additional details on animals subjects in Table 3.6 and further details of the entire study are the subject of another publication.

3.5.3 Lipodomic profiling

Lipids were extracted from 50 μl of plasma using a modified Bligh-Dyer Method. The extraction was performed using water/methanol/dichloromethane (2:2:2 v/v/v) at room temperature after spiking internal standards. The organic layer was then collected and dried under a stream of nitrogen before being re-suspended in 100μL of Buffer B [acetonitrile/water/isopropanol (10:5:85 v/v/v) containing 10mM ammonium acetate]. The lipid extract was injected onto a 1.8μm particle 50x2.1mm internal diameter Waters Acquity HSS T3 column (Waters, Milford, MA) that was heated to 55°C. Four injections were performed with either 60% or 2% of a solution of acetonitrile/water/isopropanol (10:5:85 v/v/v) for each sample. This produced a total run-time of 20 minutes. Data were acquired in positive and negative mode using data-depended MS/MS with dynamic mass exclusion. Pooled human plasma sample and pooled experimental sample (prepared by combining small aliquots of all experimental samples) were used to control for the quality of sample preparation and analysis. Furthermore, a randomization scheme was used to distribute pooled samples within the set. A mixture of pure authentic standards was used to monitor instrument performance on a regular basis. Lipids were identified using the LIPIDBLAST computer-generated tandem MS library [103]. This database contains 212,516 spectra covering 119,200 compounds representing 26 lipid classes, including phospholipids, glycerolipids, bacterial lipoglycan, and plant glycolipids. Quantification of lipids was

completed using AB-SCIEX MultiQuant software. The nomenclature used for individual lipids begins with the abbreviation of the lipid class followed by the number of carbon atoms in the molecule, and, finally, by the number of double bonds.

3.5.4 Untargeted metabolite profiling

Fasting plasma (50 μl) was extracted by adding 280 μl of extraction solvent (1:1:1 methanol: acetonitrile: acetone) containing internal standards; vortexing for 10 sec, allowing to rest on ice for 5 min, and then centrifuging at 4°C for 10 min. The supernatant was dried by vacuum centrifuge at 45°C and resuspended in 200 μl of 8:2 methanol:water. Metabolites were analyzed by LC-MS using an Agilent 1260 infinity LC connected to an Agilent 6520 quadrupole time-of-flight MS. MS parameters were as follows: full-scan negative ion mode (m/z 50 to 1,200), acquisition rate 1 spectrum/sec, capillary voltage 3500 V, gas temperature 350°C, drying gas 10 l/min, nebulizer pressure 20 psig, and reference mass correction enabled. RPLC was performed using a Waters Acquity HSS T3 column, 1.8 μm particle size, 2.1 x 100 mm i.d. (Milford, MA), with a flow rate of 0.25 ml/min. The gradient consisted of a 7-min linear ramp from 0 to 99% B, 3 min at 99% B, and 5 min of re-equilibration at 0% B. Mobile phase A was 0.1% of formic acid in water and mobile phase B was 0.1% of formic acid in 8:2 of isopropanol:acetonitrile.

Untargeted feature peak areas were initially quantified using Profinder version B.08.00 (Agilent Technologies, Santa Clara, CA) and re-quantified using Agilent Masshunter Quantitative Analysis software for quadrupole time-of-flight MS version B.07.00. Peaks were re-quantified by peak area using the "Agile2" or "spectrum summation" peak integrator. Untargeted metabolite identification was performed using accurate mass and retention time from authentic standards by MS or using MS/MS fragmentation pattern referenced from
www.lipidmaps.org/resources/lipidmapspresentations/EB2009/BrownEB2009.pdf. Un-

targeted metabolite annotation was performed using Human Metabolome Database (www.hmdb.ca) and LIPID MAPS Lipidomics Gateway (www.lipidmaps.org).

3.5.5 Data normalization

Lipidomics data were normalized to remove batch and run order effects. Each lipid was normalized individually, without the use of internal standards. Positive and Negative modes treated separately, until the final step of removing duplicate lipids. Pooled samples are the pooled samples from the test data. Lipids which were missing excessive data from either the pooled samples or the subject samples were removed. Robust regression on the pooled data was used to calculate an adjustment ratio between batches; this ratio was then used to remove batch effects. For each lipid i, we calculate a batch-adjustment factor β_i. If there are two batches, this is essentially the slope from the robust regression of one batch on the other, without an intercept. Let b_i^1 be the measurements for lipid i in batch 1 and b_i^2 be the measurements for lipid i in batch 2. We want to calculate $b_i^2 = \beta_i b_i^1$.

If there are more than two batches, then one batch is picked as the reference, and all other batches are regressed against the reference batch, one at a time. We use the *lmrob* function from the R package *robustbase* for calculating the adjustment ratio between batches. Once the adjustment factors have been calculated, missing data are imputed using the `knn` function from the R `pamr` package. Imputation takes into account the batch number, run order and sample label. Then, the adjustment factor is used to remove batch effects by updating b_i^2 to be $\frac{1}{\beta} b_i^2$.

Next, loess smoothing is used to remove the remaining effects of run order. Loess tuning parameters are calculated on the pooled samples, and then used to smooth the original samples. Once all batch and run order effects have been adjusted for, positive and negative modes are combined and repeated lipids are removed. If a lipid is present in only one mode, but with multiple ions, we keep the ion with lowest

variability as measured by relative standard deviation (RSD), where RSD of the i^{th} lipid, l_i, is equal to 100stdev(l_i)/mean(l_i). If a lipid is present in both modes, we pick the mode that has the most lipids of that lipid's class, and keep the ion w/ the lowest RSD within that mode. If a lipid is present in both modes, and there are the same number of ions/lipids in both modes, we keep the ion with the lowest RSD across both modes After normalization and the elimination of duplicates, there were 458 lipids in the controlled feeding experiment. All data was then $\log_2$-transformed.

The untargeted metabolomics were normalized similarly. We started with 1588 untargeted metabolites, run in one batch. We removed metabolites that had fewer than 5 pooled samples, had an RSD over 30, or were missing more than 25% of their samples across all timepoints. Where there were multiple instances of a single metabolite, we retained the instance with the lowest RSD and discarded the others. Additionally, we chose to analyze only named, non-lipid untargeted metabolites. Selected exogenous compounds (acetaminophen and caffeine) were removed. The resulting set of untargeted metabolites had 147 members. Missing data were imputed, positive and negative modes were combined, the data was median centered by subject and finally log2 transformed.

Animal data (lipids and untargeted metabolomics) were normalized with the lipidomics normalization workflow. After normalization, the data had 478 lipids and 188 untargeted metabolites, out of which, 304 and 75 overlapped with the human data, respectively.

CHAPTER IV

Consensus Correlation Modules for Discovery and Insight

4.1 Introduction

In this chapter, we present a variation on the method presented in Chapter III. Previously, we showed that condition specific modules could be discovered by using the Prize Collecting Steiner Tree to integrate differentially abundant lipids and untargeted metabolites via a data driven network, while also incorporating a small number of related biomolecules that were not differentially expressed. This new method creates a robust, consensus co-expression network by combining information from multiple correlation matrices. Modules are then identified in this new network, and treated similarly to the modules in Chapter III.

The method presented here has two main differences compared to the method of Chapter III. The first being that instead of anchoring the modules with differentially abundant biomolecules, we examine the entire interactome for modules, considering *only* the biomolecules' patterns of co-expression. Focusing on differentially abundant variables requires one to impose a significance cutoff, which could leave important and relevant features out of a module simply because they were not differentially abundant. Analysis techniques that consider only biomolecules having the largest

differences between conditions have several major limitations. The most pertinent to this chapter is that after correcting for multiple testing relatively few features may be differentially abundant, especially if the biological differences are small relatively to the noise in the assay. A range of techniques have been developed to counter this problem, such as Gene Set Enrichment Analysis (GSEA) [23] and GSA [24]. These techniques also address a related problem - many biological processes involve a set of biomolecules working in concert. In such a setting, a small increase in the abundance of many members of a biological pathway could result in a more meaningful biological difference than a very significant increase in a single member of that pathway. Our method is similarly motivated, though we start with the aim of summarizing patterns of co-expression, not differential expression.

The second main difference is that instead of creating a co-expression network from a single dataset, we use a variation on the Hedges-Olkin method [104] for combining estimates of correlation coefficients to create a consensus correlation network on which modules are identified. Combining correlation coefficients in this fashion allows one to find common patterns across datasets while decreasing the influence of noise in a single dataset. Examples in the literature include Lee et al [20], who perform a large-scale analysis of 60 human mRNA microarray datasets, combining the co-expression profiles to create a high-confidence network of genes which contains functionally coherent modules of genes. Choi et al [18] use gene expression datasets from cancers of 13 different tissues to construct 2 distinct co-expression networks (tumor and normal).They compare these networks to elucidate the ways in which cancer affects many co-expression relationships leading to functional changes in energy metabolism, promotion of cell growth and immune activity. Gillis and Pavlidis [19] present another variant of this concept, summing individual co-expression matrices from microarray data to illuminate the role of indirect connections in gene networks in predicting function.

	Protein	Carbohydrate	Fat	SFA	MUFA	PUFA
PUFA, n=35	15.42 (2.38)	40.54 (2.02)	44.02 (2.02)	5.55 (0.86)	11.38 (2.04)	24.24 (2.13)
CHO, n=36	15.37 (2.15)	71.66 (3.33)	13.06 (2.67)	4.17 (1.31)	4.2 (0.94)	3.22 (0.83)
B, n=72	14.71 (0.83)	49.48 (0.89)	35.74 (0.96)	14.26 (0.93)	11.19 (0.86)	7.08 (0.75)
HF, n=231	15.21 (1.04)	24.72 (0.84)	60.03 (0.97)	27.87 (1.87)	18.16 (1.85)	8.89 (2.12)
HC, n=273	14.89 (0.65)	74.72 (0.62)	10.38 (0.61)	3.56 (0.52)	3.11 (0.32)	2.35 (0.39)
US diet	15.8 (0.1)	48.5 (0.2)	33.7 (0.1)	10.8 (0.1)		

Table 4.1: *Summary statistics for percent calorie intake from select macronutrients.* US diet data taken from dietary recall interviews for 2011-2014, reported in [106]. Remainder of data taken from NDSR output. Values are mean(sd) or mean(se) for the US diet data.

We are interested in identifying modules in the PUFA/CHO dietary intervention, described in Chapter III, based *only* on the common patterns of co-expression between lipids across all days of each dietary intervention. We use this new method to identify consensus modules in PUFA and CHO, which are then used to identify differences in the lipidomes of subjects on a second dietary intervention. This second intervention has some notable differences from the first - 24 healthy adults are fed for 3 days on a standardized diet (represented by B), meant to reflect the median standard American diet [105]. Subjects were then randomized to either a high fat (HF, n=11) or a high carbohydrate (HC, n=13) diet. While the HF/HC diets are more extreme in the percentage of total calories from fat or carbohydrates (compared to the PUFA/CHO diets), the more notable difference comes from the relative proportions of saturated (SFA), monounsaturated (MUFA) and polyunsaturated (PUFA) fatty acids.

As shown in Table 4.1, the PUFA diet is high in polyunsaturated fatty acids, while the HF diet has an even greater percentage of calories coming from saturated fatty acids. While humans cannot make polyunsaturated fatty acids, saturated and monounsaturated fatty acids can come either directly from dietary intake, or indirectly from dietary intake through de novo lipogenesis when excess carbohydrates are consumed [72]. As a result, it is anticipated the HF/HC dietary interventions will present a more complex and potentially subtle signal (relative to each other) than the PUFA/CHO dietary interventions.

4.2 Consensus Correlations Networks

4.2.1 Theory background

Let (X, Y) be a pair of random variables with a bivariate normal distribution and true correlation ρ,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\mu, \Sigma)$$

where

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Given a set of N sample pairs $(X_i, Y_i), i \in \{1, \cdots N\}$, let r be the sample correlation coefficient,

$$r = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

Fisher's z-transformation [76, 107] of r, $f_z(r)$ and the inverse transformation are defined as

$$f_z(r) = z := \frac{1}{2} ln \left(\frac{1+r}{1-r} \right) = arctanh(r) \tag{4.1}$$

$$r = tanh(z) \tag{4.2}$$

With (X, Y) as described above, and X_i, Y_i being independent and identically

distributed, then z is approximately normally distributed $z \sim N(\mu_z, \sigma_z^2)$,

$$\mu_z = \frac{1}{2} ln \left(\frac{1+\rho}{1-\rho} \right) \text{ and } \sigma_z = \frac{1}{\sqrt{N-3}}$$

where N is the number of samples and ρ the true (population) correlation coefficient.

Also of use is the fact that given $z_i, i \in \{1, \cdots, K\}$ where $z_i \sim N(0,1)$, the weighted sum, Z, has a standard normal distribution -

$$Z := \frac{\sum_{i=1}^{K} w_i z_i}{\sqrt{\sum_{i=1}^{K} w_i^2}} \sim N(0,1) \tag{4.3}$$

4.2.2 Consensus Correlation Network

Suppose that one had K related data sets, $X_k, k \in \{1, \cdots K\}$, each with N_k observations and sample correlation matrix S_k. These could be multiple datasets from the same experiment, or datasets from separate experiments investigating similar phenomena (as in the case of meta-analysis).

Let r_{kij} be the sample correlation between variables i and j in dataset k. Fisher's z-transformation is used to transform S_k into Z_k by applying Equation 4.1 to each r_{kij} in S_k. For simplicity, the elements of Z_k are standardized.

$$f_z(r_{kij}) = z_{kij} = \frac{arctanh(r_{kij}) - \frac{1}{2} ln \left(\frac{1+\rho}{1-\rho} \right)}{\sqrt{N_k - 3}} \sim N(0,1) \tag{4.4}$$

We then create a consensus Z matrix, $Z_{\mathcal{K}}$, from the weighted sum of the Z_k matrices.

$$Z_{\mathcal{K}} = \frac{\sum_{k=1}^{K} w_k Z_k}{\sqrt{\sum_{k=1}^{K} w_k^2}}$$

The weights w_k allow flexibility in how the Z_k matrices are combined- all datasets

could count equally, or emphasis could be placed on certain data sets which were more important or had more reliable data.

For each $z_{\mathcal{K}_{ij}}$, we can test whether this z variable is significantly different from 0. Formally, the hypothesis of interest is given by:

$$H_0 : \rho = 0 \tag{4.5}$$

$$H_\alpha : \rho \neq 0 \tag{4.6}$$

which is tested by calculating

$$p(z_{\mathcal{K}_{ij}}) = \mathbb{P}(z_{\mathcal{K}_{ij}} \neq 0 | z_{\mathcal{K}_{ij}} \sim N(0,1)) \tag{4.7}$$

Testing this hypothesis for each element of $Z_\mathcal{K}$ gives us $p(Z_\mathcal{K}) = P'_\mathcal{K}$. This matrix of probability values can then be corrected for multiple testing. Let $P_\mathcal{K}$ be the version of this matrix with p-values corrected for multiple testing using the Benjamini-Hochberg procedure [1].

The inverse Fisher's z-transformation can be applied to $Z_\mathcal{K}$ to give us a consensus correlation matrix $S_\mathcal{K} := tanh(Z_\mathcal{K})$. $S_\mathcal{K}$ can be sparsified by setting to 0 any $S_{\mathcal{K}_{ij}}$ which has corresponding $P_{\mathcal{K}_{ij}} > c$ for some significance cutoff c. This sparsified matrix can be thought of as a network, $G_\mathcal{K}$, with nodes corresponding to the original variables observed in the data. Nodes i and j are connected by an edge with non-zero weight $e_{ij} = |S_{\mathcal{K}_{ij}}|$ if $P_{\mathcal{K}_{ij}} < c$, and are unconnected otherwise. This network is then used for all future analyses.

4.2.3 Module discovery and analysis

Relevant modules can be identified on $G_\mathcal{K}$, the sparse consensus correlation network, by using a community detection algorithm, e.g. leading eigenvector community detection (LEVCD) algorithm [46]. By using the entire interactome/graph, the

method identifies relevant modules based solely on their co-expression patterns. As there is no pre-selection step (as in Chapter III, where the modules are centered around a pre-selected subset of biomolecules), the method remains agnostic to any features of our variables beyond their co-expression patterns.

These modules can then be tested for enrichment in various characteristics using the hypergeometric test. This test indicates whether, in a given module, there are more lipids with a certain characteristic than one would expect by chance (in which case the module is enriched in said characteristic), or if there are fewer than one would expect (in which case the module is depleted.) See Cao and Zhang [78] for a good review on the hypergeometric test.

The modules can also be tested for differential abundance DA between pairs of conditions; our preferred method for this is a custom implementation of the GSA method [24]. Inspired by the Gene Set Enrichment Analysis (GSEA) procedure of [23], Efron and Tibshirani developed the maxmean statistic as a more robust means of detecting group level differences in a wider range of settings. This statistic is computed in the following way:

Given some test statistic z, define

$$s(z) = (s^{(+)}(z), s^{(-)}(z)), \quad \left\{ \begin{array}{rl} s^{(+)}(z) = & \max(z, 0) \\ s^{(-)}(z) = & -\min(z, 0) \end{array} \right\}$$

For some set of biomolecules $\mathcal{S}$, containing m biomolecules g, each having a test statistic z, the maxmean statistic, $\mathcal{S}_{max}$ is defined as

$$\mathcal{S}_{max} := \max \left\{ \frac{\sum_{g \in \mathcal{S}} s^{(+)}(z)}{m}, \frac{\sum_{g \in \mathcal{S}} s^{(-)}(z)}{m} \right\} \tag{4.8}$$

As $\mathcal{S}_{max}$ is divided by the total number of molecules m, many small biomolecule scores will contribute more than a single large score. The statistic is robust by design,

and does not allow a few large biomolecule scores (positive or negative) to dominate.

4.3 Results

4.3.1 Module discovery in PUFA/CHO

Initial exploratory analysis of the PUFA/CHO data indicates that many lipids change quite quickly within the first two days of either dietary intervention, while others take the full 21 days to reach a significant change (see Section 4.3.2 and Figure 4.1 for complete details). The patterns of co-expression also change over time - some quickly, and others more gradually. By creating a consensus correlation matrix from each diet, the significance of co-expression patterns which are always strong is increased, and influence of noise inherent in a single data set is decreased.

Letting $Z_{d_i} = f_z(S_{d_i})$ for any $i \in \{0, 2, 7, 21, 23, 48\}$, consensus correlation matrices for the PUFA and CHO diets are created from an equally weighted sum of the days for each dietary intervention.

$$Z_{PUFA} = \frac{Z_{d_2} + Z_{d_7} + Z_{d_{21}}}{\sqrt{3}}$$

$$S_{PUFA} = tanh(Z_{PUFA})$$

$$Z_{CHO} = \frac{Z_{d_{23}} + Z_{d_{28}} + Z_{d_{42}}}{\sqrt{3}}$$

$$S_{CHO} = tanh(Z_{CHO})$$

The consensus correlation matrices, S_{PUFA} and S_{CHO}, are sparsified using a significance cutoff of 0.05 (controlling FDR at the 5% level) giving $G_{\mathcal{PUFA}}$ and $G_{\mathcal{CHO}}$. LEVCD identifies 4 larger modules in both PUFA and CHO; as can be seen in the module membership contingency table (Table 4.2) most of the lipids fall into one of these 4 groups, with a handful of remaining lipids clustered into modules with one or

	C_1	C_{13}	C_{14}	C_{15}	singletons/dyads	total
P_1	90	22	27	8	6	153
P_{16}	26	74	9	30	1	140
P_{17}	20	8	68	8	0	104
P_{18}	10	11	21	1	2	45
singletons/dyads	6	3	5	0	2	16
total	152	118	130	47	11	458

Table 4.2: *Contingency table of module membership.* PUFA modules are labeled as P_i and CHO modules as C_j. In each diet, 4 modules with more than 8 members are identified. No module is identical across either diet. Groups of lipids move together under either diet, joining together with different other lipid groups to form the larger modules.

two members (singleton/dyad modules).

Table 4.2 also shows us that none of the modules are identical across diets. Rather, groups of lipids (some quite large, others rather small) move together. These groups come together with other groups under PUFA to form PUFA modules, then detaching and associating with other groups of lipids to form CHO modules. This phenomenon is similar to the one illustrated in Figure 3.2.

Each of the identified modules is tested for enrichment (E) or depletion (D) in certain classes, saturation levels and behavioral characteristics (Table 4.3). A module is classified as enriched if it contains more lipids of a particular class of lipid (or saturation level) than one would expect by chance, and depleted if it contains less of the same than one would expect by chance. For convenience, the p-values, adjusted for multiple comparisons using the Benjamini Hochberg procedure [1], are discretized in the following manner:

$$\text{D/E} : 0.05 < p < 0.1$$

$$\text{DD/EE} : 0.001 < p < 0.05$$

$$\text{DDD/EEE} : 0.0001 < p < 0.001$$

$$\text{DDDD/EEEE} : p < 0.0001$$

We can see that P_1 is depleted in DGs and TGs, while P_{16} is highly enriched in these two classes, as is C_{13}. The PEs and PCs tend to group together in P_{17} and C_{14}, but under neither dietary condition do they group with the lysoPCs and lysoPEs.

dd	CE	DG	lysoPC	lysoPE	PC	PE	PI	plasmenyl-PC	plasmenyl-PE	SM	TG	SFA	MUFA	PUFA
P_1		D	EEEE	EEE		D		E	EEE	EEE	DDDD	EE	EE	DDD
P_{16}		EEEE	DD		DDDD	DDDD	D			DD	EEEE		DD	EE
P_{17}		DDD	DD		EEEE	EEEE			D		DDDD	DD	DD	EEE
P_{18}			D				EEEE			DD			EE	DD
C_1			EEEE	E		DDD			EEEE	EEEE	DDDD			DD
C_{13}	EE	EEEE			DDD	DDD			DD	DDDD	EEEE			
C_{14}		DDD	DD		EEEE	EEEE	EEEE		DD	DD	DDDD			EE
C_{15}			D			DD					EEEE	DD		EEE

Table 4.3: *Module characteristics.* Modules are tested separately for enrichment (E) or depletion (D) in each of the classes and saturation levels shown. Classes with fewer than 9 members observed are not shown (CerP, CL, MG, PA, PG). P-values, after adjusting for multiple comparisons, are discretized as follows: D/E:$0.05 < p < 0.1$, DD/EE: $0.001 < p < 0.05$, DDD/EEE: $0.0001 < p < 0.001$, DDDD/EEEE: $p < 0.0001$

4.3.2 Differential Abundance Testing

Using the linear model and hypothesis tests presented in Section 3.2.2, we can test for differential abundance between any two days measured in the PUFA/CHO experiment. These tests, summarized in Table 4.4, showed us that many lipids changed quickly, with 28% of the lipids changing significantly within the first 2 days of either dietary intervention. Most of these lipids decrease over the PUFA dietary intervention (35% of the lipids decrease significantly between d_{21} and d_0). In contrast, 35%

of the lipids increase over the CHO dietary intervention ($d_{42} - d_{21}$). We also see that relatively few lipids change between d_{42} and d_0 (baseline), suggesting that the individuals were consuming a relatively high carbohydrate diet prior to entry into the study. Overall, the model in Equation 3.2 does a good job of explaining the variation seen in the data, with 317/458 models having an adjusted p-value < 0.1.

	$d_2 - d_0$	$d_7 - d_2$	$d_{21} - d_7$	$d_{23} - d_{21}$	$d_{28} - d_{23}$	$d_{42} - d_{28}$	$d_{21} - d_0$	$d_{42} - d_{21}$	$d_{42} - d_0$
proportion negative	0.58	0.64	0.68	0.30	0.27	0.83	0.74	0.36	0.54
proportion signif ($p < 0.1$)	0.28	0.08	0.01	0.35	0.10	0.06	0.39	0.48	0.16
proportion signif & nega	0.21	0.02	0.01	0.04	0.00	0.06	0.35	0.13	0.13

Table 4.4: *Summary of time course dynamics in PUFA/CHO.* Proportions are calculated as (# with characteristic)/458.

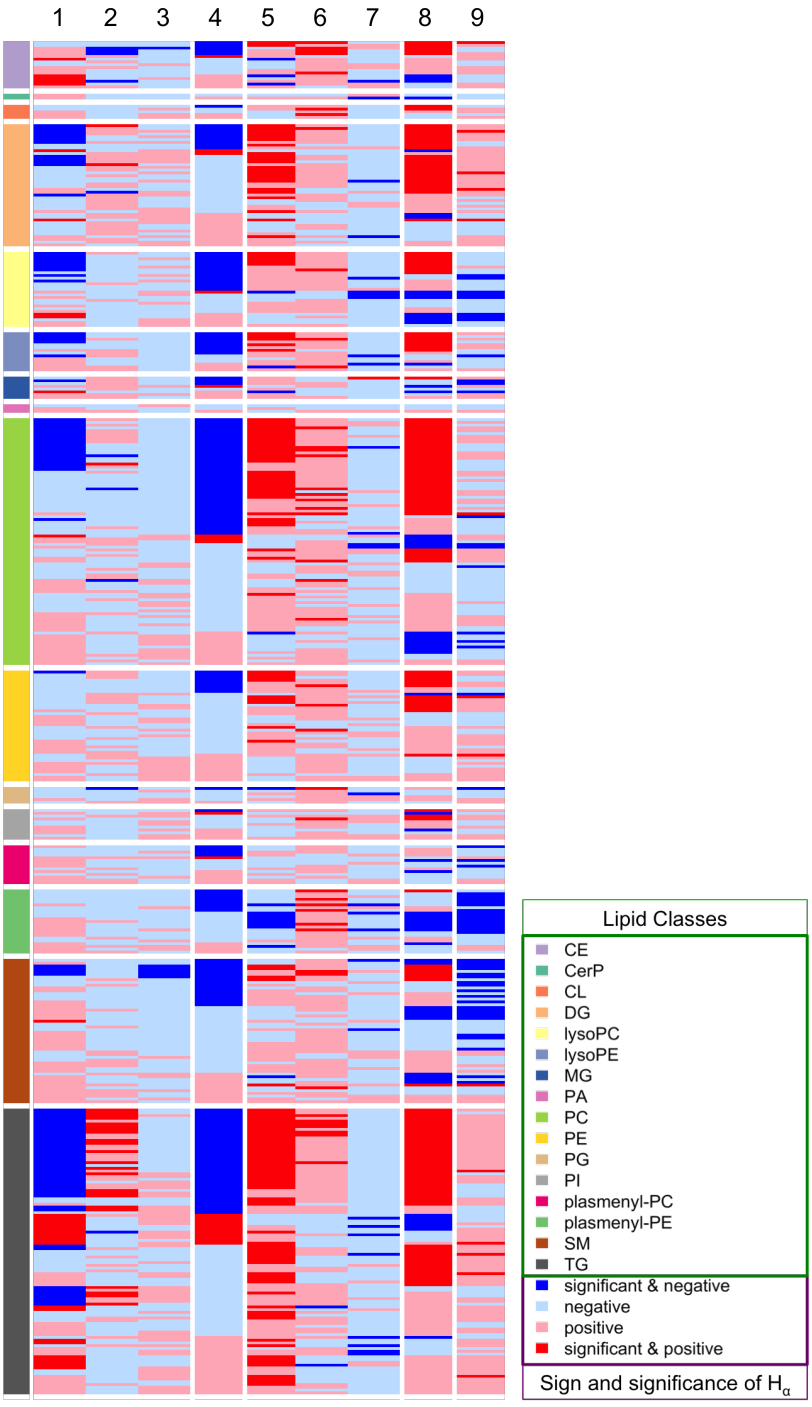
We can classify the lipids that show a significant change across either PUFA or CHO as fast or slow, based on how quickly they reach a significant change. Fast lipids, under either diet, change significantly in the first 2 days, and also change significantly between the beginning and end of the diet. Slow lipids are classified as those which do not change significantly within the first 2 days, but reach a significant change between the beginning and the end of the diet. These labels ($\mathcal{P}_F, \mathcal{P}_S, \mathcal{C}_F, \mathcal{C}_S$) have a significant association with each other, as can be seen in Table 4.5. We will use $\mathcal{C}_N$ and $\mathcal{P}_N$ to denote the lipids which do not change significantly over the course of either dietary intervention.

These dynamic changes can be summarized visually in Figure 4.1.

The consensus modules can be tested for enrichment or depletion in these dietary labels. Table 4.6 shows that modules which are enriched for $\mathcal{P}_F$ tend to be enriched for $\mathcal{C}_F$, and that some of the modules are enriched for lipids which show no significant

	$\mathcal{C}_F$	$\mathcal{C}_s$	$\mathcal{C}_N$
$\mathcal{P}_F$	62	20	11
$\mathcal{P}_S$	25	23	39
$\mathcal{P}_N$	39	53	186

Table 4.5: *Distribution of DA labels for each diet.* PUFA labels have significant association with CHO labels (Pearson's Chi-squared $p < 2.2e^{-16}$)



(a) (b)

Figure 4.1: *Visual representation of dynamic changes in PUFA/CHO.* Tests in each column are: 1: $d_2 - d_0$; 2: $d_7 - d_2$; 3: $d_{21} - d_7$; 4: $d_{21} - d_0$; 5: $d_{23} - d_{21}$; 6: $d_{28} - d_{23}$; 7: $d_{42} - d_{28}$; 8: $d_{42} - d_{21}$; 9: $d_{42} - d_0$

changes over either diet.

	$\mathcal{P}_F$	$\mathcal{P}_S$	$\mathcal{P}_N$	$\mathcal{C}_F$	$\mathcal{C}_S$	$\mathcal{C}_N$
P_1	DD	E		DD		EE
P_{16}	EEEE	DDD		EEEE	D	DD
P_{17}	D	E				
P_{18}	E					
C_1	DDDD		EE	DDD		EE
C_{13}	EEEE	DD	DD	EEEE		DD
C_{14}						
C_{15}						

Table 4.6: *Module enrichment in dietary labels* Modules are tested separately for enrichment and depletion in the diet labels shown. P-values are summarized in Table 4.3, after adjusting for multiple comparisons.

4.3.3 Module dynamics

For each consensus module the maxmean statistic, $\mathcal{S}_{max}$, is computed. P-values are obtained in the usual fashion by permuting sample labels (see Section 4.6.1 for additional details). We are particularly interested in whether a given module increases *or* decreases between d_0 and d_{21}, d_0 and d_{42}, and d_{21} and d_{42} (results in Table 4.7). As expected from the linear model results, all modules significantly decrease between d_0 and d_{21}. Likewise, all modules but one, C_1, increase between d_{21} and d_{42}. Module C_1, enriched in lipids which do not change under either diet and several classes of lipids, decreases significantly across this diet.

In Table 4.4, we saw that relatively few ($\sim 16\%$) of the lipids show a significant difference between baseline (d_0) and the end of the carbohydrate diet (d_{42}). These modules, however, help us to identify more subtle changes - about half of the identified modules decrease across the entire course of the experiment, while the other half increase. Further analysis is necessary to identify which lipids are responsible for these changes, and what characteristics they might have in common.

	P_1	P_{16}	P_{17}	P_{18}	C_1	C_{13}	C_{14}	C_{15}
$d_{42} > d_0$	1	0	1	0.128	1	0	1	0
$d_{21} > d_0$	1	1	1	1	1	1	1	1
$d_{42} > d_{21}$	0	0	0	0	1	0	0	0
$d_{42} < d_0$	0	1	0.04	1	0	1	0	1
$d_{21} < d_0$	0	0	0	0	0	0	0	0
$d_{42} < d_{21}$	1	1	1	1	0	1	1	1

Table 4.7: *GSA analysis in PUFA/CHO data.* Modules from PUFA/CHO are tested for DA using a custom implementation of GSA. End of PUFA (d_{21}) and end of CHO d_{42} are tested against baseline (d_0) and each other. Values presented are p-values adjusted for multiple comparisons.

4.4 HF/HC validation

The previous sections illustrate that our method can be used to identify biologically meaningful modules in a controlled feeding experiment where the dietary interventions are relatively simple and starkly different. In HF/HC experiment, the dietary interventions are more complex - all subjects are put on standardized diet for three days, before being randomized to HF or HC conditions. While the overall mean percent of calories from fat is more extreme in the HF/HC diets than in the PUFA/CHO diets, the more meaningful difference comes in the percent of calories from SFA and MUFA. In particular, the HF diet is substantially different from the PUFA diet in terms of what percentage of calories is coming from SFA vs MUFA vs PUFA. Given that humans make their own saturated and monounsaturated fatty acids through de novo lipogenesis when the body takes in more carbohydrates than necessary, the signal from the HF/HC dietary interventions is more complex and potentially more difficult to analyze.

4.4.1 Linear Modeling of HF/HC

The dynamics of these dietary interventions can be captured by a linear model similar to the one used for PUFA/CHO. The abundance of the k^{th} lipid is modeled

as:

$$L_k \sim \mu + r \tag{4.9}$$

where μ is a factor with 10 levels:

$$\mu \text{ levels: } \mu_{-3}, \mu_0, \mu_{F2}, \mu_{F7}, \mu_{F14}, \mu_{F21}, \mu_{C2}, \mu_{C7}, \mu_{C14}, \mu_{C21}$$

All subjects have measurements for levels μ_{-3} and μ_0, which represent the standardized diet on days -3 and 0. Levels $\mu_{F2}, \mu_{F7}, \mu_{F14}, \mu_{F21}$ represent the 3 weeks of HF diet, and the remaining levels correspond to the 3 weeks of HC diet.

The random effects, represented by r, for subject i on diet j are: $r_i + r_{ij}$. Each subject is on two different diets: B (corresponding to the standardized diet on day -3 and day 0) and HF or HC, corresponding to the remainder of the days assayed.

Out of 562 models, 281 were significant at the model level (adjusted p-value < 0.1). Linear model results are summarized in Table 4.8, where it can be see immediately that the time course dynamics are more complex than in PUFA/CHO. The majority of lipids which change significantly across either of the diets do decrease (comparing μ_{F21} and μ_{C21} to either μ_0 or μ_{-3}), but when μ_{C21} is compared with μ_{F21} a greater proportion of lipids are significantly decreasing than what was found in the equivalent comparison in PUFA/CHO. It is interesting to note that more lipids change significantly when transitioning from the standardized diet to the HC diet ($\mu_{C2} - \mu_0$) than when subjects transition to the HF diet ($\mu_{F2} - \mu_0$). The total percentage of macronutrients from fat is reduced by over 66%, in the former case, while almost being doubled in the latter case. Further investigations into this subset of lipids may yield further insight into the short term response of the metabolome to dietary perturbations.

	$\mu_0 - \mu_{-3}$	$\mu_{C2} - \mu_0$	$\mu_{C7} - \mu_{C2}$	$\mu_{C14} - \mu_{C7}$	$\mu_{C21} - \mu_{C14}$	$\mu_{C21} - \mu_0$	$\mu_{C21} - \mu_{-3}$	$\mu_{C21} - \mu_{F21}$
proportion negative	0.75	0.60	0.69	0.43	0.68	0.71	0.57	0.44
proportion signif ($p < 0.1$)	0.02	0.12	0.03	0.06	0.00	0.23	0.13	0.22
proportion signif & negative	0.02	0.08	0.03	0.01	0.00	0.21	0.10	0.12
		$\mu_{F2} - \mu_0$	$\mu_{F7} - \mu_{F2}$	$\mu_{F14} - \mu_{F7}$	$\mu_{F21} - \mu_{F14}$	$\mu_{F21} - \mu_0$	$\mu_{F21} - \mu_{-3}$	
proportion negative		0.65	0.68	0.55	0.40	0.77	0.60	
proportion signif ($p < 0.1$)		0.00	0.00	0.00	0.00	0.19	0.10	
proportion signif & negative		0.00	0.00	0.00	0.00	0.17	0.08	

Table 4.8: *Summary of time course dynamics in HF/HC.* Proportions are calculated as (# with characteristic)/562.

	C_1	C_{13}	C_{14}	C_{15}	singletons/dyads	total
P_1	30	11	14	3	3	61
P_{16}	11	60	5	22	0	98
P_{17}	10	3	48	5	0	66
P_{18}	3	8	13	0	1	25
singletons/dyads	3	0	1	0	0	4
total	57	82	81	30	4	254

Table 4.9: *Contingency table of module membership in HF/HC data set.* Module membership for PUFA and CHO modules, with *only* lipids measured in HF/HC

4.4.2 Module dynamics under more extreme dietary perturbations

Out of the 562 lipids measured in the HF/HC dataset, only 254 overlapped with the PUFA/CHO dataset. Fortunately, the modules are relatively well preserved, and the missingness is distributed relatively evenly across modules (Table 4.9.)

The modules defined in Section 4.3.1 can be tested for DA as in Section 4.3.3 to see if there are more substantial changes in HF/HC data at the module level than at the individual lipid level. As with the PUFA/CHO data, the three comparisons which are of most interest are: μ_{C21} vs μ_0, μ_{F21} vs μ_0, and μ_{C21} vs μ_{F21}. These results (Table 4.10) complement and magnify the linear model results, while also providing a more complex picture than what was originally observed in the PUFA/CHO experiment.

We see that when day 21 is compared against day 0, for HF or HC, the modules which change significantly decrease, with the exception of P_{16} and C_{13} which increase between across the HC diet. These modules, along with C_{15} increase between d_0 and d_{42} in the original CHO diet. P_{18} and C_{15} decrease over the HC diet, while remaining

	P_1	P_{16}	P_{17}	P_{18}	C_1	C_{13}	C_{14}	C_{15}
$\mu_{C21} > \mu_0$	1	0.016	1	1	1	0.016	1	1
$\mu_{F21} > \mu_0$	1	1	1	1	1	1	1	1
$\mu_{C21} > \mu_{F21}$	1	0	0	0.896	1	0	0.1184	0
$\mu_{C21} < \mu_0$:	0	0.998	0	0.0576	0	0.998	0	0.069
$\mu_{F21} < \mu_0$	0	0	0	0	0	0	0	0
$\mu_{C21} < \mu_{F21}$	0	1	1	0.875	0	1	1	1

Table 4.10: *GSA analysis in HF/HC data.* Modules from PUFA/CHO are tested for DA in HF/HC data set using a custom implementation of GSA. End of HC (μ_{C21}) and end of HF μ_{F21} are tested against μ_0 and each other. Values presented are p-values adjusted for multiple comparisons.

unchanged (P_{18}) or increasing (C_{15}) in the CHO diet.

It is particularly interesting that while P_1 increases from d_{21} to d_{42}, it decreases over the equivalent comparison in the HF/HC study. These subtle differences in module level activity may provide insight into more complex systems level dynamics regarding the metabolism of SFA, MUFA and PUFA.

4.5 Discussion

In this chapter we present an alternative method for identifying modules of lipids, one which does not rely on an arbitrary DA cutoff, and which combines data from multiple data sets to create more robust co-expression networks. This method prioritizes differential edges, instead of differential nodes (as in Chapter III). Ideally this method is suited to a scenario where one has weak mean differentials between conditions, and where the patterns of co-expression themselves are of primary interest.

We use the method to identify diet-linked modules, in the PUFA/CHO controlled feeding study, which are then are used to illuminate more complex dynamic behavior in a second controlled feeding study (HF/HC).

4.6 Methods

4.6.1 Gene Set Analysis Implementation

Let $\hat{\beta}_i$ be the estimate of the fixed effects from our linear model and $\hat{\Sigma}_i$ the associated covariance matrix from solving the appropriate regression equation for the i^{th} lipid of the relevant dataset. Given a contrast vector c, we can calculate the appropriate statistic

$$D_i = \frac{c'\hat{\beta}}{\sqrt{c'\hat{\Sigma}c}}$$

The test statistics D_i are used to calculate $\mathcal{S}_{max}$ (Equation 4.8) for a set $\mathcal{S}$, comprised of all the lipids in a consensus module (P_1, C_{16}, etc.) To calculate a p-value for each $\mathcal{S}_{max}$, sample labels are permuted n_{perm} times (we used $n_{perm} = 500$.) For paired data (as in PUFA/CHO, or certain comparisons in HF/HC), the labels are shuffled in pairs. For unpaired tests, all sample labels are permuted together. Sample labels which are not directly involved in a test remained fixed (ie: if testing d_{21} - d_0, the sample labels for the observations from d_2, d_7, d_{23}, d_{28} and d_{42} remain unchanged.)

$\mathcal{S}_{max}$ is recomputed on each permuted dataset, giving permuted values $\mathcal{S}^{*1}_{max}$,$\mathcal{S}^{*2}_{max}$,$\cdots$, $\mathcal{S}^{*n_{perm}}_{max}$. For the test $d_i - d_j$, two p-values are computed:

$$p - \text{value for } d_i > d_j = \frac{\sum_{k=1}^{n_{perm}} \mathbb{I}(\mathcal{S}^{*k}_{max} > \mathcal{S}_{max})}{n_{perm}} \tag{4.10}$$

$$p - \text{value for } d_i < d_j = \frac{\sum_{k=1}^{n_{perm}} \mathbb{I}(\mathcal{S}^{*k}_{max} < \mathcal{S}_{max})}{n_{perm}} \tag{4.11}$$

P-values are then adjusted for multiple comparisons across all modules tested using the Benjamini-Hochberg False Discovery Rate adjustment procedure [1].

4.6.2 Study methods

The diet intervention study was approved the Institutional Review Board of the University of Michigan (HUM00110543). Participants were recruited through umclinicalstudies.org website and with posters placed throughout the University of Michigan community. Subjects were required to have a body mass index (BMI) of $18.5 - 27 kg/m^2$ and between the age of 18-45 years. They were required to be weight stable for 6 months ($\pm$5 pounds), have no known food allergies and be willing to eat provided meals. Exclusions included active cigarette use within the previous 6 months, active cardiovascular disease, diabetes mellitus or other metabolic diseases, use of any medication known to alter metabolism, such as metformin. After obtaining informed consent, subjects were randomized within sex to either the a high carbohydrate (HC) or high fat (HF) diets (6 HC women, 6 HF women, 7 HC men, 5 HF men). Subjects were assigned a unique identification number for tracking. All collected data was de-identified prior to analysis using this unique number.

Plasma sampling. Plasma was collected from all subjects was obtained in three phases: Baseline, Standard Diet, and. was sampled at baseline, after 3 days of a Standard Diet and at days 2, 7, 14 and 21 of the HC or HF experimental diet. **Dietary intervention. Standard diets** reflected the 50^{th} percentile macronutrient intake ($\pm$2%) for the US population, as available in the 2015 Dietary Guidelines Advisory Committee Report (Committee, 2015) with the target intake of 15% protein, 35% fat, and 50% carbohydrate. Details of the interventions are detailed in past MCRU-NAL protocols [108, 109, 110].

Experimental Diets. The HF diet target was 60% fat, 25% carbohydrate and 15% protein while the HC diet target was 10% fat, 75% carbohydrate and 15% protein. Total calories provided during the Standard diet and Experimental diet phases was calculated to meet macronutrient requirements of the eucaloric diet for each individual. Weight varied less than 1% during the dietary interventions.

Dietary Analysis. For PUFA/CHO MCRU Nutrition Assessment Laboratory dietitians analyzed recorded the content of the diet over a random sample of 3 days of each study diet using the dietary analysis program Nutrition Data for Research (NDSR). Each subject had 3 days analyzed for each dietary intervention, except for subject 17, which only had 2 days for PUFA due to a processing error. For HF/HC, dieticians analyzed recorded content for each of the 24 subjects for each of the 3 standard diet days (B), as well as for each of the 21 HF or HC diet days.

4.6.3 Lipidomics methods

Lipidomics methods for PUFA/CHO are described in Chapter III.

HF/HC data were processed similar, with additional classes of lipids added to the insilico library.

4.6.4 Lipid Normalization

Lipids in both studies were normalized as in Chapter III. In HF/HC data, one subject sample (male, high carb group at day 0) was removed because of an injection error. After normalization, PUFA/CHO data had 458 lipids in 16 classes; HF/HC data had 562 lipids in 23 classes.

APPENDICES

APPENDIX A

Supplementary material for Chapter II

A.1 Algorithm Details

In the algorithm for estimating $\widehat{Q}$ and $\widehat{B}^k$, (Appendix Algorithm 1), we include two additional steps, one of which is optional, to help deal with noisy data. The identification restriction on the B^k_{ii}'s require them to be positive; hence, $\widehat{B}^k_{ii} < 0$ implies excessively noisy data for variable i, which in turn can result in the corresponding entries in the $\widehat{\Lambda}^k$ having differing signs. It is possible to have different signs in $\widehat{\Lambda}^k$ without having a $\widehat{B}^k_{ii} < 0$, if the signal is strong enough. If $\widehat{B}^k_{ii} < 0$, we pick the "correct" sign based on a majority rule, and remove from estimation the entries with the opposing sign.

If Q_{ij} is nonzero, then any corresponding entry $\widehat{\Lambda}^k_{ij}$ which is set to 0 in the course of the algorithm is done so because of noise. This means that in the calculation of $\widehat{Q}_{ij}$, this element is not getting the full contribution from each $\widehat{\Lambda}^k_{ij}$, and will be artificially low. This can be partially addressed by an optional scaling factor, s_{ij}, which was found to improve performance in some simulation settings.

$$s_{ij} = \frac{K}{\sum_{k=1}^{K} \mathbb{I}(\widehat{\Lambda}_{ij}^{k} \neq 0)} \tag{A.1}$$

$\widehat{Q}_{ij}$ is then updated to be $s_{ij}\widehat{Q}_{ij}$.

If all $\widehat{\Lambda}_{ij}^{k}$ are nonzero, then the scaling factor is 1. Otherwise, $s_{ij} > 1$, and will slightly inflate Q_{ij}. This adjustment would be included whenever $\widehat{Q}$ was calculated (immediately after line 1 in Appendix Algorithm 1). A brief discussion of the results of including this step is included in Appendix Section A.3.4.

Recall that certain constraints are placed on B^k, depending on the overall design structure, and how one wants to compare data sets. If comparisons between any two data sets in a 2×2 experimental design are desirable, one would normalize B^k across all levels of both design factors. If the first design factor has levels $\{1, \cdots K_1\}$ and the second $\{1, \cdots K_2\}$, then

$$\text{ID0} : \sum_{k=1}^{K_1} \sum_{j=1}^{K_2} B^{kj} = I \tag{A.2}$$

Given a single design factor with unordered levels, this constraint would reduce to

$$\text{ID1} : \sum_{k=1}^{K} B^{k} = I \tag{A.3}$$

If one were primarily interested in comparing across levels of the one design factor for a single, fixed, level of the second design factor, one might use the following constraint:

$$\text{ID2} : \sum_{k=1}^{K_2} B^{ik} = I \text{ (for each fixed } i \in \{1, \cdots, K_1\}.) \tag{A.4}$$

Algorithm 1: Estimating $\widehat{Q}$, $\widehat{B}^k$

Data: $\widehat{\Lambda}^k$, $k \in \{1, \cdots, K\}; m; d$
Result: $\widehat{B}^k, \widehat{Q}$
begin

1 $\quad \widehat{Q} \leftarrow \frac{1}{d} \sum_k \widehat{\Lambda}^k$

1.1 $\quad$ **if** *including* s_{ij} **then**

1.2 $\quad\quad \widehat{Q_{ij}} \leftarrow s_{ij}(\widehat{Q}_{ij})$

2 $\quad$ **for** $k \in \{1, \cdots, K\}$ **do**

3 $\quad\quad \widehat{B}^k_{ii} = \frac{1}{m} \sum_{j=1}^{m} \widehat{\Lambda}^k_{ij} / \widehat{Q}_{ij}$

4 $\quad\quad$ **if** $\widehat{Q}_{ij} = 0$ **then**

5 $\quad\quad\quad \widehat{B}^k_{ij} = 0$

6 $\quad$ Normalize $\widehat{B}^k$

7 $\quad$ **if** $\widehat{B}^k_{ii} < 0$ **then**

8 $\quad\quad$ **for** $j \in \{1, \cdots, m\}$ **do**

9 $\quad\quad\quad n_p = \sum_{k=1}^{K} \mathbb{I}(\widehat{\Lambda}^k_{ij} > 0)$, $n_n = \sum_{k=1}^{K} \mathbb{I}(\widehat{\Lambda}^k_{ij} < 0)$

10 $\quad\quad\quad$ **if** $n_p = n_n$ **then**

12 $\quad\quad\quad\quad \widehat{\Lambda^k_{ij}} = 0 \; \forall \, k$

13 $\quad\quad\quad$ **if** $n_p > n_n$ **and** $\widehat{\Lambda}^k_{ij} < 0$ **then**

14 $\quad\quad\quad\quad \widehat{\Lambda}^k_{ij} \leftarrow 0$

15 $\quad\quad\quad$ **if** $n_p < n_n$ **and** $\widehat{\Lambda}^k_{ij} > 0$ **then**

16 $\quad\quad\quad\quad \widehat{\Lambda}^k_{ij} \leftarrow 0$

16-21 $\quad$ repeat lines 1-6

where one is interested in comparing outcomes for each of the K_1 levels of D_1 across all K_2 levels of D_2.

In Algorithm 1, normalizing of $\widehat{B}^k$ is according to design structure and these constraints. d is equal to 1 if using the constraints in equations A.2 and A.3, and is equal to K_1 if using the constraint in equation A.4.

For the the algorithms calculating $\widehat{\Lambda}^k$, we need a few additional functions. Let $\phi_{\widehat{\Lambda}^k}(\widehat{\Lambda^{k'}})$ be the rotation of $\widehat{\Lambda^{k'}}$ so that it has maximum similarity with $\widehat{\Lambda}^k$ (the Procrustes rotation). The reference condition k can be chosen at random, or with

some consideration of the experimental design.

Let $\psi_c(\widehat{\Lambda}^k)$ be the thresholding of $\widehat{\Lambda}^k$ where all entries between $-c$ and c are set to 0.

Finally, let $\eta_s(\widehat{\Lambda}^k)$ be the function with thresholds $\widehat{\Lambda}^k$ by cardinality. Specifically, let $s = (s_1, s_2, \cdots, s_m)$ represent the target cardinality (the target number of non-zero entries) for each column of $\widehat{\Lambda}^k$. If $\widehat{\Lambda^k}$ has p rows, then, for each column j, the $p - s_j$ smallest entries are set to 0.

The estimation of non-sparse $\widehat{\Lambda}^k$ via a scaled eigendecomposition is presented in Algorithm 2. The addition of lines 3.1 and 3.2 are all that is required for a fast, decent approximation of a sparse $\widehat{\Lambda}^k$ via method EDTM. Method EDTC is presented in Algorithm 3, and the SPCA method for sparse $\widehat{\Lambda}^k$ is presented in Algorithm 4.

Algorithm 2: Estimation of non-sparse $\widehat{\Lambda}^k$ OR sparse $\widehat{\Lambda}^k$ via method EDTM

Data: $U^k D^k U^{k'}$, the eigen-decomposition of Σ^k, $k \in \{1, \cdots, K\}; m; c$
Result: $\widehat{\Lambda}^k$
begin
1 **for** $k \in \{1, \cdots, K\}$ **do**
2 $\widehat{\Lambda}^k \leftarrow U^k_{1:m}\sqrt{D^k_{1:m}}$
3 Pick $k \in \{1, \cdots, K\}$
3.1 **if** Q *is sparse* **then**
3.2 $\widehat{\Lambda}^k \leftarrow \psi_c(\widehat{\Lambda^k})$
4 **for** $k' \in \{1, \cdots, K\}, k' \neq k$ **do**
5 $\widehat{\Lambda^{k'}} \leftarrow \phi_{\widehat{\Lambda}^k}(\widehat{\Lambda^{k'}})$
6 **for** $k \in \{1, \cdots, K\}$ **do**
7 $\widehat{\Lambda}^k \leftarrow \psi_c(\widehat{\Lambda^k})$

In Algorithm 4, Σ_{spca} is the diagonal matrix having entries equal to the variance of each sparse principal component. In this specific case where we are using the *spca* function from *elasticnet*, the entries are equal to the total variance of the decomposition multiplied by the percent explained variance of each component. We use the *varnum* option of *spca*, in which the user supplies the number of non-zero components

Algorithm 3: Estimation of sparse $\widehat{\Lambda}^k$, via EDTC

Data: $U^k D^k U^{k'}$, the eigen-decomposition of Σ^k, $k \in \{1, \cdots, K\}; m; s$
Result: $\widehat{\Lambda}^k$
begin
1-3 As in Algorithm 2
4 $\widehat{\Lambda}^k \leftarrow \eta_s(\widehat{\Lambda}^k)$
5 **for** $k' \in \{1, \cdots, K\}, k' \neq k$ **do**
6 $\widehat{\Lambda^{k'}} \leftarrow \phi_k(\widehat{\Lambda^{k'}})$
7 $\widehat{\Lambda^{k'}} \leftarrow \psi_c(\widehat{\Lambda^{k'}})$

desired for each sparse principal component to be estimated.

Algorithm 4: Estimation of sparse $\widehat{\Lambda}^k$, via SPCA

Data: Σ^k, $k \in \{1, \cdots, K\}$, $m, (s_1, \cdots s_m)$
Result: $\widehat{\Lambda}^k$
begin
1 **for** $k \in \{1, \cdots, K\}$ **do**
2 $\widehat{\Lambda}^k \leftarrow$ spca(Σ^k) with m sparse principal components, having $(s_1, \cdots s_m)$ non-zero elements in each column.
3 $\widehat{\Lambda}^k \leftarrow \widehat{\Lambda}^k \sqrt{\Sigma_{spca}}$
4 $\widehat{\Lambda}^k \leftarrow \psi_c(\widehat{\Lambda^k})$
5 Pick $k \in \{1, \cdots, K\}$
6 **for** $k' \in \{1, \cdots, K\}, k' \neq k$ **do**
7 $\widehat{\Lambda^{k'}} \leftarrow \phi_{\widehat{\Lambda}^k}(\widehat{\Lambda^{k'}})$
8 **for** $k \in \{1, \cdots, K\}$ **do**
9 $\widehat{\Lambda}^k \leftarrow \eta_s(\widehat{\Lambda^k})$

A.2 Performance review of sparse PCA methods

A.2.1 selecting tuning parameters

For both SPCA and EDTC, we used the `leading.eigenvector.community` detection algorithm (LEVCD) in the `igraph` package in R to estimate the number of latent factors, m, and their cardinalities, $(s_1, \cdots s_m)$. The steps to estimate these

tuning parameters are presented in Algorithm 5.

Algorithm 5: Number and cardinality of latent factors

Data: $\Sigma^k \ \forall k$
Result: $m, (s_1, \cdots s_m)$
begin
  for $k \in \{1, \cdots, K\}$ **do**
    Convert Σ^k to an undirected graph with weighted edges, G_k.
    Calculate the number, c_k and size, $s_1^k, \cdots s_{c^k}^k$, $(s_i^k > s_{i+1}^k)$, of communities present in G_k.
  $m = \min_{k \in \{1, \cdots K\}} c^k$
  $(s_1, \cdots s_m) = (s_1^m, \cdots s_{c^m}^m)$

A.2.2 Performance Review

In the case where we believe that the underlying Q is sparse, our first challenge is in calculating an accurate sparse eigen-decomposition. We began by investigating SPCA ([43] [44]), implemented in the `spca` function of the R package `elasticnet`. In principal this method should work well when tuned appropriately. The default setting uses two different λ penalty parameters;in our investigations we were unable to come up with a good heuristic for guiding the tuning of these parameters.

The `spca` function in R has a second option, one which allows the user to put in the number of principal components to estimate, and the number of non-zero elements in each. When given the true values for these parameters, the method worked quite well. This raises another question though - how to estimate the values for those parameters? It turns out that LEVCD can be used quiet reliably to estimate those parameters.

The first quantity LEVCD needs to estimate is the number of communities (which we take as equivalent to the number of latent factors). Over all iterations of all the 2-factor simulation settings that we ran, LEVCD returned the correct number of factors 99.94% of the time. In the few cases where it mis-estimated the number of factors, it

added an additional factor.

For the 3-factor simulation settings, the performance was also quite good, returning the correct number of factors 99.6% of the time. Here too the method usually over-estimated the number of latent factors, most commonly when the factors only explained 50% of the observed variance and the communities all had similar sizes.

In both the 2 and 3 factor simulation settings, when LEVCD over estimated the number of factors/communities, the additional community was estimated to be very small. For the 2-factor settings, this third community was never estimated to have more than 2 members; for the 3-factor settings, the fourth community was never estimated to have more than 7 members (the majority of the time it was estimated as having only one member.)

Across all 2 factor simulation settings, LEVCD estimates the first and second community within ± 1 of their true size 98.3% of the time. In the 3 factor simulation settings, LEVCD estimates the first, second and third community sizes within ± 1 of their true size over 99.2% of the time.

While SPCA yielded good performance, the algorithm often took some time to run. Given that we had already estimated the number and cardinality of each of the factors, we wondered if there was a simpler method for creating the sparse eigendcomposition which could use these parameters, but offered improved performance or speed over SPCA. For this purpose, we developed an alternative eigen-decomposition method, where the eigenvectors were truncated by their cardinality (EDTC method, Algorithm 3), which also yields good results.

After close examination of the results from EDTC method, it was observed that many of the "incorrect" values were quite small, and would have been truncated if a simple cutoff had been used. This led us to test a simple eigen-decomposition, truncated by magnitude in the usual way (method EDTM). To make a fairer comparison, we used the number of communities estimated via LEVCD (so there would

be the same number of factors as the other methods), but truncated with a simple 0.1 threshold, the same as we used for the non-sparse eigen-decomposition.

Intuitively, whichever method has the best reconstruction of $\widehat{\Lambda}^k$, is also going to give us the best results for $\widehat{B}^k$ and $\widehat{Q}$. We used a modified version of the reconstruction loss (Equation A.5) as a proxy for how well each method would perform in the algorithm over all.

$$\textit{Reconstruction- loss v2:} \frac{1}{K}\sum_{k=1}^{K}\frac{||\,|\Lambda^k| - |\widehat{\Lambda}^k|\,||_F}{||\Lambda^k||_F} \tag{A.5}$$

We can compare this reconstruction loss for the three methods in Appendix Figures A.7, A.8 and A.9. We note that for the EDTM, we did use the number of communities generated by the LEVCD method, though this number could also be obtained from a scree plot. It can be seen that both of the eigen-decomposition methods outperform the spca method. Closer inspection shows that EDTC slightly outperforms EDTM, based on this reconstruction loss, but the performance differences is quite small. In summary, simple eigen-decomposition, truncated by cardinality, yields results just as good as the SPCA, with the advantage of being much faster to compute. Furthermore, if one did not want to go through the additional work of estimating parameters via LEVCD, one could get a decent sparse principal component approximation simply by truncation by magnitude.

A.3 Simulation Results

A.3.1 additional simulation details

For each $B^k \in \{1, \cdots, K\}$ we generate a random $b^k \sim Unif(0.15, 0.85)$. The set of B^k are then normalized according to the design structure, as discussed in Section A.1.

The factors F are generated from a $N(0,1)$ distribution, while the non-zero entries in Q are generated uniformly from $(-1,-0.5)\cup(0.5,1)$.

Once we have generated our B^k, F and Q, we generate our distribution for E. For each realization of E we sample from a $N(0,\Psi)$ distribution, where Ψ is designed so that Λ^k explains, on average, 50% or 75% of the variance in the observed data.

In the context of our motivating example, lipids in a certain class are more likely to have variances which are more similar to each other than to lipids from another class, due to technical reasons. With an eye towards this application, we also design Ψ as a block-diagonal matrix. In all of our simulations, Ψ has three blocks, each representing approximately 1/3 of the total variables. If the target variance (calculated as described below) is σ_t^2, then we set the variance within the blocks to be $((\sigma_t-0.05)^2,\sigma_t^2,(\sigma_t+0.06)^2)$, respectively.

The proportion of variance in our dataset, having p variables, explained by our factors can be written as:

$$\frac{\sum_{i=1}^{p}\left(\Lambda^k\Lambda^{k'}\right)_{ii}}{\sum_{i=1}^{p}\left(\Lambda^k\Lambda^{k'}\right)_{ii}+\Psi_{ii}} \tag{A.6}$$

If we want our factors to account for 50% of the total variance in the dataset, then

$$\sigma_t^2=\frac{1}{p}\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{p}\left(\Lambda^k\Lambda^{k'}\right)_{ii}$$

If we instead want our factors to account for 75% of the total variance in the dataset, then

$$\sigma_t^2=\frac{1}{3}\frac{1}{p}\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{p}\left(\Lambda^k\Lambda^{k'}\right)_{ii}$$

Each application of the algorithm includes an additional step in which we use

a procrustes rotation to align the $\widehat{\Lambda}^k$ with Λ^k, so that we accurately measure the performance of the algorithm, *not* the accuracy of the eigen-decomposition or `spca` in picking the correct orientation. We used the `procrustes` function in `vegan` package in R, with the `scale=false` option, which ensures that the transformation is only a rotation, and therefore orthogonal.

The performance of the algorithm is evaluated via the following equations:

$$\textit{B-loss: } \frac{||B - \widehat{B}||_F}{||B||_F} \tag{A.7}$$

$$\textit{Q-loss: } \frac{||\,|\widehat{Q}| - |Q|\,||_F}{||Q||_F} \tag{A.8}$$

$$\textit{Reconstruction-loss: } \frac{1}{K}\sum_{k=1}^{K} \frac{||\,|\widehat{B}^k\widehat{Q}| - |B^kQ|\,||_F}{||B^kQ||_F} \tag{A.9}$$

where B is a $p \times K$ matrix whose k^{th} column is the diagonal entries in B^k.

In the non-sparse simulation settings we used a threshold of $c = 0.1$ for the final truncation step in estimating $\widehat{\Lambda}^k$ (step 7 in Algorithm 2). We ran the algorithm with and without the s_{ij} scaling factor on the same datasets; in general, the adjustment step decreased the error in estimating B, while slightly increasing the error in estimating Q. These two effects combined to have a minuscule decrease in the error of $\widehat{\Lambda}$ in some settings. Overall, this step was shown to yield no increase in performance in the non-sparse settings (likely due to the strong signal), while yielding some improvements in the sparse scenario (possibly due to the difficulty in correctly estimating the sparse structure.) While the step increases the error in $\widehat{Q}$ slightly, it also serves to decrease the number of entries in $\widehat{Q}$ which have small loadings. This may be an advantageous in an application setting, as it was in our application on the AI data. Depending on which component was of greater interest in an application, B^k or Q, a researcher may wish to include the adjustment step or not.

A.3.2 Results for non-sparse Q, without s_{ij}

Overall, we have quite good results across all simulation settings, as can be seen in Figures A.1, A.2, and A.3. We see that in many ways that algorithm behaves as our intuition tells us it should - loss goes down as the error decrease and as the percentages of variance explained by the latent factors increases. Loss generally decreases with increasing sample or variable size, but increases (particularly in Q) with an increasing number of latent factors. Performance across all metrics tends to increase as the number of datasets increases, regardless of the experimental design structure. We also observe that the standard deviations of the loss values are quite small - indicating that the algorithm consistently has good performance.

The mean *B-loss* is always < 0.233, with half of the instances even below 0.11. Mean *Q-loss* values are also quite good- always below 0.23, with the majority of the settings having loss values < 0.17, and a quarter below 0.12. Mean *Reconstruction-loss* values are slightly higher, but still below 0.29, with half of the simulation settings having loss values between 0.083 and 0.174.

A.3.3 Results for sparse Q, without s_{ij}

Our preferred method for estimating sparse $\widehat{\Lambda}^k$ is the EDTC method, as it yielded the best results, shown in Figures A.4, A.5, and A.6. "Qload: 1" and "Qload: 2" refer to Q having $(0.55p, 0.45p)$ and $(0.7p, 0.3p)$ non-sparse elements respectively, when $m = 2$, or Q with $(0.39p, 0.33p, 0.28p)$ and $(0.5p, 0.3p, 0.2p)$ non-sparse elements, respectively, for $m = 3$.

Again we have quite good results across all simulation settings. With the addition of sparsity, the overall picture is somewhat more complex than previously. Overall, we see similar trends to the non-sparse case - loss decreases as the percentage of variance explained by our factors increases, or as sample size increases. Now however, we see that as the number of factors increases, the loss also tends to increase very slightly.

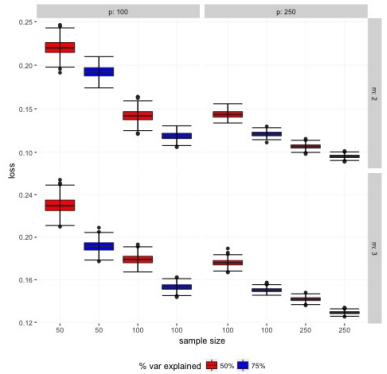
For two factors and either experimental design structure, the algorithm performs slightly better in the Qload: 1 settings than in the Qload: 2 settings for all of the performance metrics. This is also true for most of the three factor settings, but the performance increase is so slight as to be negligible.

Across all settings, the mean *Q-loss* ranges between 0.067 and 0.158, with half of the settings having loss values below 0.1. The mean *B-loss* is higher here, ranging from 0.042 to 0.24, but the majority of the settings have loss values below 0.15. Mean *Reconstruction-loss* values are similar in range to the *B-loss* values, 0.766 to 0.269, with most < 0.17.

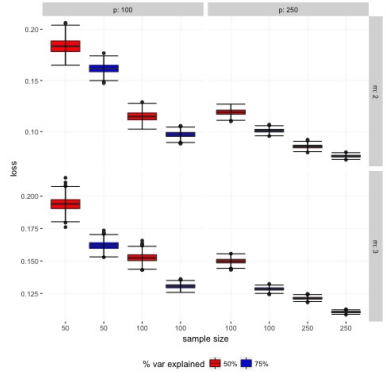
A.3.4 Results for sparse Q with s_{ij}

The results for the sparse scenarios with the additional adjustment step have the same general trends as without that adjustment step, as can be seen in Figures A.10, A.11 and A.12.

There are some slight differences - mean *Q-loss* ranges between 0.068 and 0.193, with most of the entries having loss < 0.139. Mean *B-loss* ranges from 0.042 to 0.183, with about half of the loss values being below 0.1. Mean *Reconstruction-loss* values are still fairly good, between 0.076 and 0.27, with 75% of those being < 0.171.

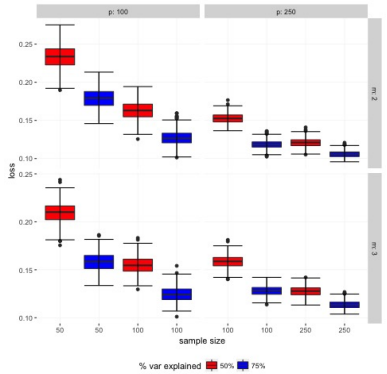


(a) 3×1 design

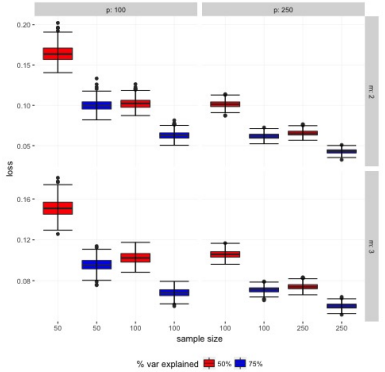


(b) 2×2 design

Figure A.1: *Q-loss*, non-sparse Q

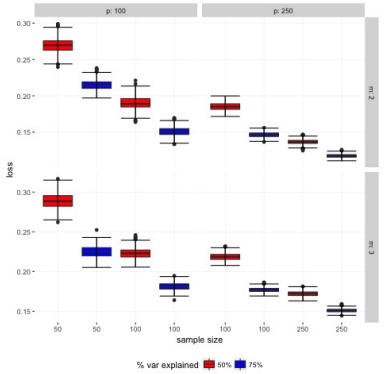


(a) 3×1 design

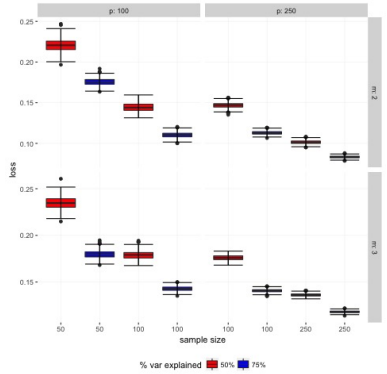


(b) 2×2 design

Figure A.2: *B-loss*, non-sparse Q

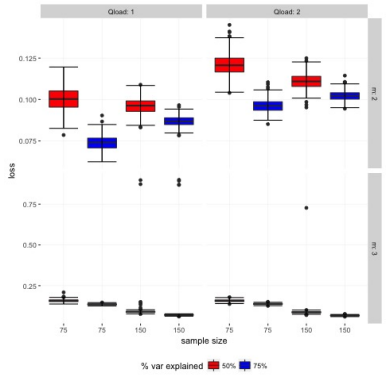


(a) 3×1 design

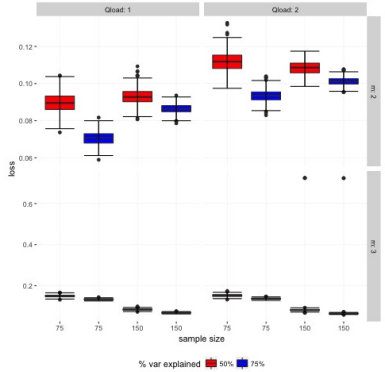


(b) 2×2 design

Figure A.3: *Reconstruction-loss*, non-sparse Q

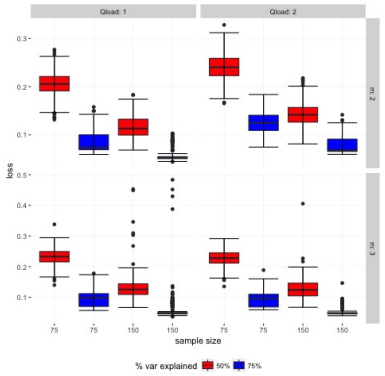


(a) 3×1 design

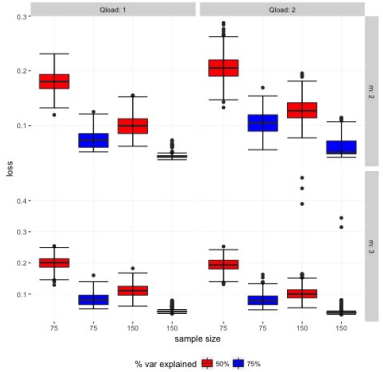


(b) 2×2 design

Figure A.4: *Q-loss* with sparse Q

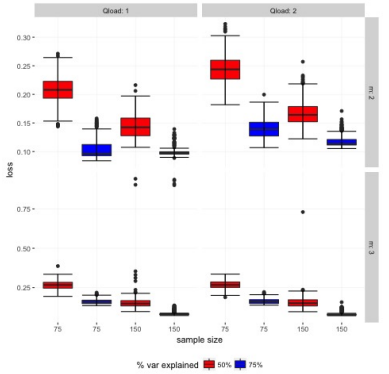


(a) 3×1 design

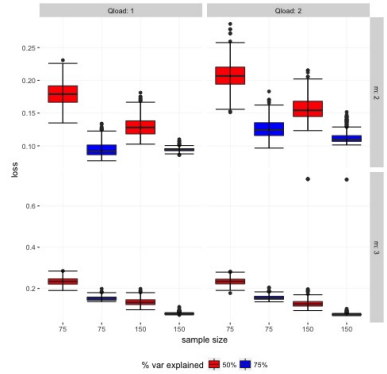


(b) 2×2 design

Figure A.5: *B-loss* with sparse Q

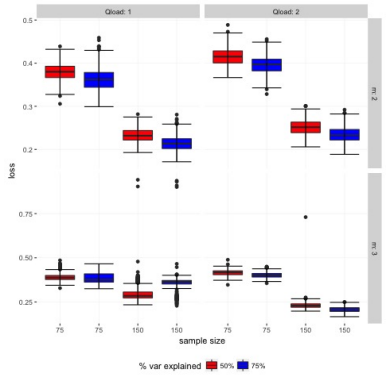


(a) 3×1 design

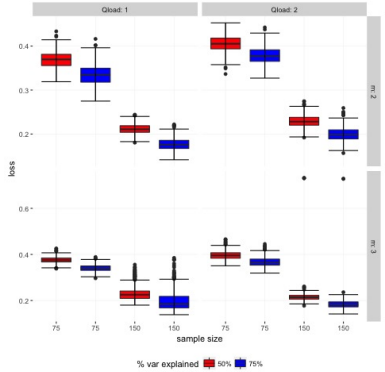


(b) 2×2 design

Figure A.6: *Reconstruction-loss* with sparse Q

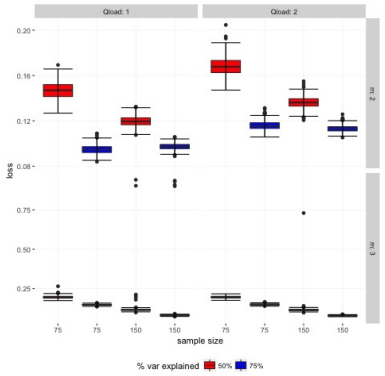


(a) 3×1 design

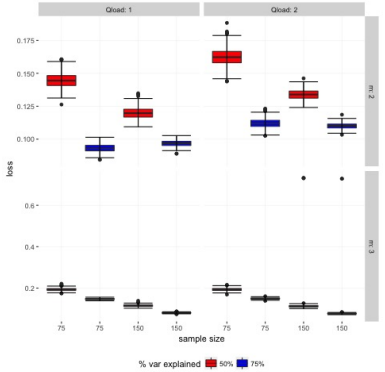


(b) 2×2 design

Figure A.7: *Reconstruction-loss v2*, SPCA

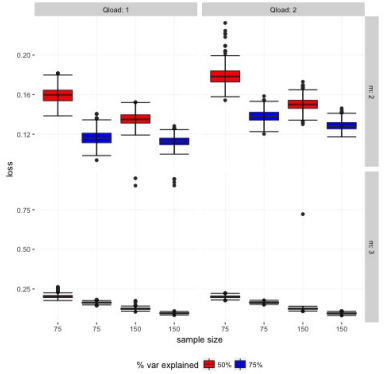


(a) 3×1 design

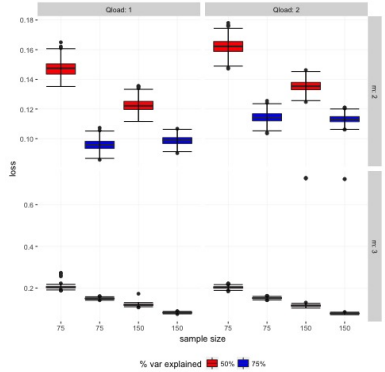


(b) 2×2 design

Figure A.8: *Reconstruction-loss v2*, EDTC

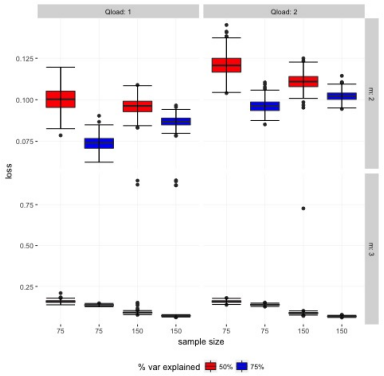


(a) 3×1 design

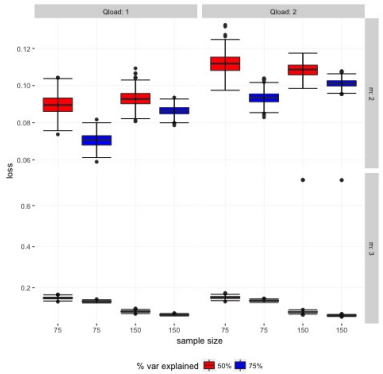


(b) 2×2 design

Figure A.9: *Reconstruction-loss v2*, EDTM

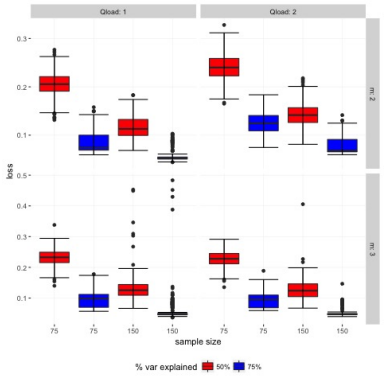


(a) 3×1 design

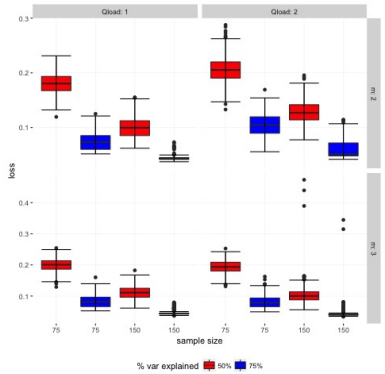


(b) 2×2 design

Figure A.10: *Q-loss* with sparse Q and s_{ij}

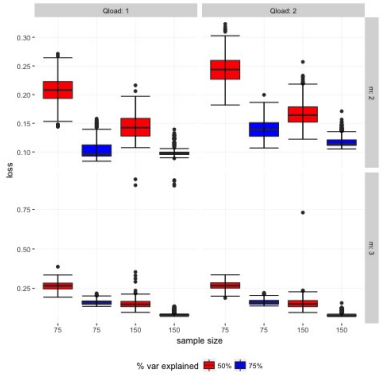


(a) 3×1 design

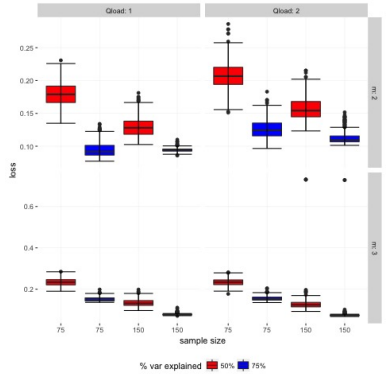


(b) 2×2 design

Figure A.11: *B-loss* with sparse Q and s_{ij}



(a) 3×1 design



(b) 2×2 design

Figure A.12: *Reconstruction-loss* with sparse Q and s_{ij}

A.4 Materials and Methods

A.4.1 Lipid background

Lipids are small biomolecules comprised of a head group attached to a fatty acid tail. The structure of the head group separates lipids into classes (such as diglycerides (DGs), triglycerides (TGs), and sphingomylins (SMs), among others). The fatty acid tail is comprised of chains of carbon atoms, connected by single or double bonds. Lipids can be classified by the length of these tails as short, medium or long chain lipids. They can also be classified as saturated (0 double bonds), monounsaturated (1 double bond) or polyunsaturated (2 or more double bonds).

A.4.2 Metabolomics methods

Plasma samples were spiked with internal standards lipids (LPC 17:0/0:0, PC 17:0/17:0, PE 17:0/17:0, SM 17:0/17:0, Ce 17:0/17:0, PG 17:0/17:0, PS 17:0/17:0, PA 17:0/17:0, PI 17:0/20:4, d5-DAG, TG 17:0/17:0/17:0 and D31-TAG, Avanti Polar Lipids (Alabaster, AL)) and lipids extracted using a modified Bligh-Dyer method [111] using 2:2:2 volume ratio of water/methanol/dichloromethane at room temperature as described previously [112]. The organic phase was collected and dried under nitrogen and reconstituted in $100\mu L$ of a buffer (10:85:5 ACN/IPA/H2O) containing $10mM$ ammonium acetate and analyzed using LC-MS based lipidomics. The data acquisition was performed in both positive and negative ionization modes, using a TripleTOF 5600 equipped with a DuoSpray ion source (AB Sciex, Concord, Canada). The lipids were identified using Lipid Blast [113] [114] ,software by matching MSMS spectra to different library and the data files were processed using MultiQuant 1.1.0.26 [115] (ABsciex, Concord, Canada).

A.4.3 Normalization procedure

Data was normalized to remove batch and run order effects, and log2 transformed. One case sample had lipid data, but no FFA or eicosanoid data, and so was removed (giving us 24 cases and 25 controls). There were 442 lipids, 9 eicosanoids and 16 FFA in the dataset.

Lipidomics data was normalized to remove batch and run order effects. Each lipid was normalized individually, without the use of internal standards. Positive and Negative modes treated separately, until the final step of removing duplicate lipids. Pooled samples are the pooled samples from the test data.

Each batch had 7 pooled samples. Lipids that were missing more than 2 pooled samples across both batches were removed (35 lipids in the positive mode, 5 lipids in the negative mode).

Robust regression on the pooled data was used to calculate an adjustment ratio between batches; this ratio was then used to remove batch effects.

For each lipid i, we calculate a batch-adjustment factor β_i. If there are two batches, this is essentially the slope from the robust regression of one batch on the other, without an intercept. Let b_i^1 be the measurements for lipid i in batch 1 and b_i^2 be the measurements for lipid i in batch 2. We want to calculate

$$b_i^2 = \beta_i b_i^1$$

If there are more than two batches, then one batch is picked as the reference, and all other batches are regressed against the reference batch, one at a time. We use the *lmrob* function from the R package *robustbase* for calculating the adjustment ratio between batches.

Once the adjustment factors have been calculated, and missing data imputed (using the *knn* function from the *pamr* package. Imputation was done taking into

account the batch number, run order and sample label (cases at baseline, controls at 3 months, etc).) we can then use the adjustment factor to remove batch effects by updating b_i^2 to be $\frac{1}{\beta} b_i^2$.

Next, loess smoothing is used to remove the remaining effects of run order. Loess tuning parameters are calculated on the pooled samples, and then used to smooth the original samples.

Once all batch and run order effects have been adjusted for, we combine the positive and negative modes and remove any duplicate lipids which appear more than once.

If a lipid is present in only one mode, but with multiple ions, we keep the ion with lowest variability as measured by relative standard deviation (RSD), where RSD of the i^{th} lipid, l_i, is equal to $100\text{stdev}(l_i)/\text{mean}(l_i)$.

If a lipid is present in both modes, we pick the mode that has the most lipids of that lipid's class, and keep the ion w/ the lowest RSD within that mode.

If a lipid is present in both modes, and there are the same number of ions/lipids in both modes, we keep the ion with the lowest RSD across both modes.

The FFA data was normalized in the same way as the lipidomics data. The eicosanoid data was run in a single batch and was median centered after being log2 transformed.

A.4.4 Case Study Details

A.4.4.1 Patient Characteristics

Samples were derived from a prospective clinical trial - all postmenopausal women, had completed surgery, radiation, and chemotherapy as indicated, participating in a randomized clinical trial (Exemestane and Letrozole Pharmacogenomics (ELPh)) comparing two aromatase inhibitors (letrozole and exemestane). Patients were followed prospectively during treatment to assess tolerance of medication, completed

		Cases (n=25)	Controls (n=25)
Median age (range)		60 (48-79)	60 (44-77)
Race	White Black	22 (88%) 3 (12%)	23 (92%) 2 (8%)
Body mass index (kg/m2), mean (SD)		30.5 (5.6)	29.3 (5.1)
Aromatase inhibitor (AI)	Exemestane Letrozole	16 (64%) 9 (36%)	15 (60%) 10 (40%)
Time on AI, months, median (range)		5.5 (2.9-6.0)	24.2 (23.6-25.1)
Prior chemotherapy		12 (48%)	11 (44%)
Prior tamoxifen		13 (52%)	12 (48%)

Table A.1: *Baseline Demographic and Medical Characteristics.* Data presented as $n(\%)$, $n(max, min)$, or mean (standard deviation)

the HAQ/VAS questionnaire at baseline, 1, 3, 6, 12, and 24 months. Those patients who discontinued therapy by 6 months because of increased pain were "cases" and those who didn't report a pain level $> 2/10$ during the 24 month follow-up on AI therapy were the "controls" [48]. A summary of subject characteristics can be seen in Appendix Table A.1. Serum was collected at baseline and after 3 months of therapy on subjects enrolled on this clinical trial.

A.4.4.2 Application of method

Based on scree plots of the covariance matrices for each of our 4 data sets, we decided on a 2-factor model with a non-sparse underlying Q. We chose the controls at time 0 as the reference condition for the Procrustes rotation. The rotated $\widehat{\Lambda}^k$, were then truncated with a threshold of ± 0.1 to remove small loadings.

We used constraint ID2 (equation A.4) to normalize the $\widehat{B}^k$ values, normalizing the $\widehat{B}^k$ for the controls together, and the $\widehat{B}^k$ for the cases together.

We chose to include the s_{ij} adjustment step for several reasons. It moves the metabolites further away from the axis of $\widehat{Q}$ and therefore helps make the grouping more clear. Additionally, the step makes it it so fewer metabolites are truncated because of low loadings.

After applying the method, 6 lipids were removed for having $\widehat{Q}_{i1} = \widehat{Q}_{i2} = 0$, and

additional 15 lipids and one FFA removed for having $|\widehat{Q}_{i1}| < 0.2$ and $|\widehat{Q}_{i2}| < 0.2$, leaving us with 445 metabolites.

A.5 Enrichment and differential abundance analysis

For the over representation/enrichment analysis, we tested whether a particular class of metabolite or saturation level was over represented in a given group - (ie: if there were more saturated lipids in group 1 than we would expect by chance, given the total number of saturated lipids in the set of all metabolites, and the size of group 1) by using the hypergeometric distribution (see [78] for a good review on the hypergeometric test in enrichment analysis). These results are summarized in Table A.2.

To test whether a higher expression in a given group was correlated with one of the two conditions in our two-way contrasts of interests (the same two-way contrasts as in the t-tests), we used the `GSA` function from the `GSA` package in R. We used the "maxmean" method, with $s_0 = 0$ and no restandardization. P-values for these tests can be seen in Table A.3.

Group	1	2	3	4	a	b	c	d
Saturated	0.0022	0.9277	0.7693	0.9872	0.9277	0.3887	0.7693	0.3887
Monounsaturate	0.9913	0.0196	0.9913	0.9913	0.9913	0.0038	0.0244	0.2723
Polyunsaturated	0.9984	0.9984	0.015	1e-04	0.1471	0.9984	0.9984	0.9984
ffa	4e-04	0.9573	0.9573	0.342	0.4839	0.89	0.89	0.4102
CE	4e-04	0.9941	0.9941	0.342	0.8384	0.4496	0.4496	0.4102
CL	0.9878	0.0575	0.5414	0.8454	0.3212	0.5414	0.8454	0.8454
DG	0.9913	2e-04	0.0919	0.9913	0.9913	0.8252	0.9913	0.9913
lysoPC	0.9741	0.9741	1e-04	0.9741	0.9741	0.0462	0.1548	0.9741
lysoPE	0.8906	0.8906	0.0894	0.8906	0.0894	0.0045	0.8906	0.8886
MG	0.2613	0.7406	0.2613	0.2613	0.2876	0.194	0.2876	0.2822
PA	0.0362	0.953	0.5367	0.5367	0.2786	0.5367	0.0774	0.5367
PC	0.9997	0.9997	0.0291	0.5607	0.1042	0.1193	0.4632	0.0291
PE	0.7149	0.0174	0.9558	0.9576	0.9558	0.3951	0.7149	0.9558
PG	0.0014	0.7695	0.7695	0.7695	0.7063	0.7063	0.7063	0.0175
PI	0.3613	0.3613	0.7622	0.9248	0.0652	0.7622	0.4065	0.3613
plasmenyl-PC	0.5366	0.7406	0.3355	0.3355	0.3355	0.3355	0.3355	0.0028
plasmenyl-PE	0.5582	0.9941	0.5582	0.0079	0.5359	0.5359	0.3145	0.596
SM	0.0648	1	1	0.2449	0.9608	0.9608	4e-04	0.0026
TG	0.9995	0	0.9995	0.0382	0.9995	0.9995	0.9995	0.9995
eico	0	0.953	0.9241	0.9241	0.4179	0.7697	0.7697	0.7697
LA	0.0053	0.0408	0.9982	0.9982	0.984	0.984	0.984	0.984
ALA	0.9889	0.11	0.0436	0.3093	0.9889	0.9889	0.9889	0.9889

Table A.2: *Enrichment analysis for AI data set.* Variables are partitioned into 8 groups $(1, 2, 3, 4, a, b, c, d)$, based on their loadings onto $\widehat{Q}$. Each group is then tested for over representation in a class or saturation level. P-values presented in the table are adjusted row-wise for multiple comparisons, using the Benjamini-Hochberg procedure [1]

Group	1	2	3	4	a	b	c	d
higher exp corr w/ controls at t 3 vs t 0	0.99	0.2	0.99	0.99	0.99	0.99	0.99	0.99
higher exp corr w/ cases at t 3 vs t 0	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
higher exp corr w/ controls vs cases at t 0	0.8571	0.925	0.8571	0.8571	0.8571	0.8571	0.8571	0.16
higher exp corr w/ controls vs cases at t 3	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
lower exp corr w/ controls at t 3 vs t 0	0.27	0.975	0.27	0.06	0.3543	0.336	0.3543	0.06
lower exp corr w/ cases at t 3 vs t 0	0.72	0.4333	0.32	0.32	0.4333	0.32	0.32	0.72
lower exp corr w/ controls vs cases at t 0	0.8457	0.6	0.68	0.8457	0.68	0.68	0.792	0.98
lower exp corr w/ controls vs cases at t 3	0.68	0.5543	0.5543	0.5543	0.5543	0.5543	0.5543	0.5543

Table A.3: *GSA analysis for AI data set.* P-values for using GSA to test for group level differences of abundance in AI data. Variables are partitioned as in Table A.2. P-values are adjusted row-wise using the Benjamini-Hochberg procedure.

APPENDIX B

Supplementary material for Chapter III

B.1 Module nomenclature

The j^{th} module from CG_i is referred to as $M_{i:j}$ (i.e. $M_{21u:2}$ refers to the second identified module from CG_{21u}.)

B.2 PCST node and edge frequencies

PCST solution set	0	(0.0,0.05]	(0.05,0.1]	(0.1,0.15]	(0.15,0.2]	(0.2,0.25]	(0.25,0.3]	(0.3,0.35]	(0.35,0.4]	(0.4,0.45]	(0.45,0.5]	(0.5,0.55]	(0.55,0.6]	(0.6,0.65]	(0.65,0.7]	(0.7,0.75]	(0.75,0.8]	(0.8,0.85]	(0.85,0.9]	(0.9,0.95]	(0.95,1]
G_{42d} edges	77128	836	347	135	130	50	59	42	37	15	22	20	31	9	15	15	16	9	16	15	56
G_{21d} edges	59640	445	145	60	64	37	46	27	32	15	30	13	12	13	12	8	18	11	10	20	68
G_{42u} edges	96321	929	340	120	139	60	66	30	45	33	26	16	21	12	24	10	23	9	22	25	75
G_{21d} edges	89209	496	185	68	79	36	57	28	36	21	36	19	23	14	21	18	24	26	18	27	84
G_{42d} nodes		27	11	6	6	6	6	3	5	0	9	4	7	2	3	6	6	2	6	11	272
G_{21d} nodes		26	14	10	8	4	7	9	5	5	5	5	4	6	7	5	5	5	5	7	207
G_{42u} nodes		36	7	4	3	4	8	2	3	2	4	3	4	2	10	1	5	4	6	8	328
G_{21d} nodes		26	17	4	7	5	4	6	9	2	3	2	3	3	4	5	3	6	7	6	304

Table B.1: *Prize Collecting Steiner Tree output summary.* Prize Collecting Steiner Tree node and edge frequencies over all 50 noisy runs on each graph. Table contains the total number of nodes/edges contained in $> 50X$ and $\leq 50Y$ solutions for column $(X, Y]$. In each scenario, all terminal nodes are chose in 100% of the runs.

B.3 Module differential abundance

B.4 Full enrichment/depletion tables

Comparison tested	$M_{21u:1}$	$M_{21u:2}$	$M_{21u:3}$	$M_{21u:4}$	$M_{21u:6}$	$M_{21u:7}$	$M_{21u:20}$	$M_{42u:2}$	$M_{42u:3}$	$M_{42u:4}$	$M_{42u:5}$	$M_{42u:6}$	$M_{42u:7}$	$M_{42u:14}$	$M_{42u:15}$	$M_{42u:16}$
$d_{21} < d_0$	0.000	0.000	0.000	0.000	0.997	0.000	0.000	0.000	0.000	0.000	0.032	0.000	0.000	0.000	0.000	0.000
$d_{42} < d_0$	0.993	0.993	0.993	0.013	0.993	0.013	0.013	0.993	0.993	0.021	0.993	0.013	0.100	0.061	0.027	0.993
$d_{42} < d_{21}$	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.027	1.000	1.000	1.000	1.000
$d_{21} > d_0$	1.000	1.000	1.000	1.000	0.053	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$d_{42} > d_0$	0.018	0.107	0.018	0.993	0.200	0.993	0.993	0.018	0.064	0.993	0.064	0.993	0.993	0.993	0.993	0.190
$d_{42} > d_{21}$	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000

Table B.2: *Complete module DA results.* Modules from CG_{21u} and CG_{42u} were tested for differential abundance using GSA. End of PUFA (d_{21}) and end of CHO (d_{42}) tested against baseline (d_0), and each other. Values presented are row-wise fdr adjusted p-values for all of modules from CG_{21u} and CG_{42u}.

Comparison Tested	$M_{21u:2}$	$M_{21u:3}$	$M_{21u:4}$	$M_{21u:6}$	$M_{21u:20}$	$M_{42u:2}$	$M_{42u:5}$	$M_{42u:14}$	$M_{42u:15}$	$M_{42u:16}$
HCR-AL < LCR-AL	0.17	0.011	0.025	0.108	0.000	0.090	0.089	0.040	0.011	0.170
HCR-CR < LCR-CR	0.4933	0.381	0.278	0.278	0.493	0.278	0.278	0.278	0.278	0.493
LCR-CR < LCR-AL	0.0286	0.000	0.000	0.033	0.013	0.000	0.017	0.000	0.063	0.033
HCR-CR < HCR-AL	0.2125	0.213	0.213	0.213	0.870	0.213	0.213	0.200	0.748	0.213
HCR-AL < LCR-CR	0.8533	0.853	0.511	0.853	0.483	0.853	0.783	0.853	0.133	0.853

Table B.3: *Complete module dynamics in animal model.* Modules from CG_{21u} and CG_{42u} which had at least 70% overlap with animal data were tested for differential abundance in animal data using GSA. None of the tests in the opposite direction (HCR-AL > LCR-AL, HCR-CR > LCR-CR, etc) were significant. Values presented are row-wise fdr adjusted p-values for all modules with requisite coverage.

Module	CE	DG	lysoPC	lysoPE	MG	PC	PE	PI	plasmenyl-PC	plasmenyl-PE	SM	TG	untarg	SFA	MUFA	PUFA
$M_{21u:1}$	0.211	0.707	0.921	0.714	0.583	0.338	0.001	0.733	0.008	0.632	0.615	0.990	1.000	0.887	0.922	0.001
$M_{21u:2}$	0.795	0.106	0.921	0.728	0.595	0.338	0.886	0.585	0.523	0.928	0.859	0.000	1.000	0.426	0.008	0.366
$M_{21u:3}$	0.648	0.000	0.921	0.627	0.595	0.558	0.886	0.746	0.523	0.928	0.404	0.137	1.000	0.977	0.922	0.004
$M_{21u:4}$	0.441	0.707	0.000	0.361	0.057	0.251	0.886	0.746	0.523	0.860	0.429	0.990	1.000	0.009	0.017	0.720
$M_{21u:6}$	0.441	0.735	0.921	0.465	0.282	0.595	0.886	0.585	0.523	0.860	0.810	0.000	1.000	0.856	0.922	0.001
$M_{21u:7}$	0.183	0.735	0.921	0.361	0.129	0.297	0.593	0.000	0.523	0.860	0.404	0.990	1.000	0.856	0.008	0.720
$M_{21u:20}$	0.211	0.953	0.921	0.465	0.208	0.131	0.593	0.733	0.612	0.000	0.404	0.990	1.000	0.856	0.253	0.002
$M_{42u:2}$	0.485	0.000	0.920	0.602	0.567	0.923	0.950	0.551	0.790	0.816	0.965	0.000	1.000	0.860	0.192	0.009
$M_{42u:3}$	0.614	0.909	0.936	0.602	0.589	0.036	0.000	0.352	0.790	0.816	0.965	0.998	1.000	0.860	0.472	0.000
$M_{42u:4}$	0.614	0.958	0.755	0.602	0.589	0.000	0.950	0.352	0.790	0.000	0.486	0.998	1.000	0.397	0.557	0.039
$M_{42u:5}$	0.485	0.225	0.755	0.602	0.192	0.923	0.950	0.551	0.320	0.685	0.965	0.000	1.000	0.906	0.472	0.000
$M_{42u:6}$	0.123	0.958	0.755	0.602	0.192	0.923	0.950	0.538	0.205	0.482	0.000	0.998	1.000	0.397	0.192	0.960
$M_{42u:7}$	0.499	0.225	0.755	0.602	0.310	0.036	0.950	0.516	0.175	0.762	0.002	0.998	1.000	0.469	0.245	0.106
$M_{42u:14}$	0.383	0.958	0.000	0.602	0.192	0.923	0.950	0.551	0.394	0.224	0.965	0.097	1.000	0.823	0.609	0.001
$M_{42u:15}$	0.485	0.958	0.055	0.000	0.192	0.644	0.950	0.352	0.175	0.816	0.965	0.998	0.290	0.469	0.472	0.960
$M_{42u:16}$	0.712	0.958	0.755	0.602	0.567	0.605	0.950	0.551	0.790	0.816	0.965	0.000	1.000	0.457	0.245	0.029

Table B.4: *Complete enrichment analysis of identified modules.* Modules from CG_{21u} and CG_{42u} were tested for enrichment in the classes and saturation levels listed (SFA: saturated fatty acids, MUFA: monounsaturated fatty acids, PUFA: polyunsaturated fatty acids). P-values are adjusted for multiple comparisons column-wise.

APPENDIX C

Supplementary table for Chapter III

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
(-)-salsolinol	untarg		-1	-1									0
(s)-3-methyl-2-oxopentanoate	untarg		-1	1									0
1-aminocyclopropane-1-carboxylate	untarg		1	-1									1
1,7-dimethyl uric acid	untarg		-1	-1			S				10		0
11-deoxycortisol	untarg		-1	1									1
2-acetylpyrrolidine	untarg		1	-1									0
2-deoxy-d-glucose	untarg		1	1									0
2-hydroxy-3-methylbutyric acid	untarg		-1	1	T		T	T	4		4	2	1
2-hydroxybutyrate	untarg		1	-1									1
2-piperidinone	untarg		-1	1		T	T	T		10	6	3	0
3-(4-hydroxyphenyl)lactate	untarg		-1	1									1
3-dehydroxycarnitine	untarg		1	-1									0
3-hydroxy-3-methylglutarate	untarg		-1	1				S				3	1
3-methoxytyrosine	untarg		1	-1				S				7	0
3,4-dihydroxybenzoate	untarg		1	-1									0
3beta-hydroxyandrost-5-en-17-one	untarg		1	-1									0
3beta-hydroxyandrost-5-en-17-one 3-sulfate	untarg		1	1									0
4-acetamidobutanoate	untarg		-1	1				S				3	1
4-methyl-2-oxovaleric acid	untarg		-1	1									0
4-nitrophenol	untarg		-1	1									1
4-pyridoxate	untarg		1	-1									1
5-hydroxytryptophan	untarg		-1	1									0
5-oxoproline	untarg		-1	1	S				6				1
5-tetradecenoylcarnitine (myristoyl)	untarg		-1	-1									0
5-valerolactone	untarg		1	1									1
5'-methylthioadenosine	untarg		1	-1									1
allose	untarg		-1	1									0
alpha-tocopherol	untarg		1	-1									0
ascorbate	untarg		1	1									1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
betaine	untarg		1	-1									0
biliverdin	untarg		-1	1									1
butyrylcarnitine	untarg		-1	-1									0
c17 sphinganine	untarg		1	-1									0
carnitine	untarg		1	-1		T	T	T		3	4	2	1
CE 16:0	CE	SFA	1	-1									1
CE 16:1	CE	MUFA	-1	1	T	T	T	T	1	1	1	1	0
CE 16:2	CE	PUFA	-1	1									0
CE 17:1	CE	MUFA	-1	1	T	T	T	T	4	6	8	4	1
CE 18:0	CE	SFA	-1	1	S		S		6		6		1
CE 18:1	CE	MUFA	-1	1			S				5		1
CE 18:2	CE	PUFA	1	-1	T		T	T	6		14	20	1
CE 18:3	CE	PUFA	-1	1	T	T	T	T	4	21	4	20	1
CE 20:1	CE	MUFA	-1	1									0
CE 20:2	CE	PUFA	1	-1	T		T	T	4		6	7	1
CE 20:3	CE	PUFA	-1	1	T	T	T	T	3	9	15	4	1
CE 20:4	CE	PUFA	1	-1	T		T	T	16		2	3	1
CE 20:5	CE	PUFA	-1	1	T	T	T	T	3	21	3	1	0
CE 22:2	CE	PUFA	1	-1		T	T	T		7	14	7	1
CE 22:4	CE	PUFA	-1	1									1
CE 22:5	CE	PUFA	-1	1									1
CE 22:6	CE	PUFA	1	-1									0
CerP 32:1	CerP	MUFA	1	-1	T		T	T	4		4	7	0
CerP 34:1	CerP	MUFA	-1	-1									1
cholate	untarg		1	-1									1
cholesterol	untarg		1	-1	T		T	T			6	4	0
choline	untarg		-1	-1									1
cis-7,10,13,16-docosatetraenoic acid	untarg		-1	1									1
citramalate	untarg		1	1									1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
citrate	untarg		1	-1									0
CL 70:5	CL	PUFA	-1	1	T		T	T	3		3	1	0
CL 72:7	CL	PUFA	-1	1	T	T	T	T	3	9	3	1	1
CL 74:3	CL	PUFA	1	-1	S		S		3		3		0
CL 74:7	CL	PUFA	1	-1	S		S		3		3		0
CL 78:3	CL	PUFA	-1	1									1
corticosterone	untarg		1	-1									0
cortisol	untarg		1	1				S				3	0
cortisone	untarg		-1	1									1
cycloheptanecarboxylic acid	untarg		1	1									1
decanoate	untarg		-1	1		S		S		21		12	0
decanoyl-l-carnitine	untarg		1	-1	S		S		6		15		0
deoxyadenosine	untarg		-1	1									1
deoxycholic acid	untarg		-1	1									1
deoxyuridine	untarg		-1	1									1
DG 30:0	DG	SFA	-1	1	T	T	T	T	16	21	2	1	0
DG 30:1	DG	MUFA	-1	1									1
DG 32:0	DG	SFA	1	1									1
DG 32:1	DG	MUFA	-1	1	T	T	T	T	16	3	2	2	1
DG 32:2	DG	PUFA	-1	1	T		T	T	16		2	3	0
DG 33:0	DG	SFA	-1	1	T		T	T	3		3	7	0
DG 33:1	DG	MUFA	-1	1	T		T	T	16		2	2	1
DG 33:2	DG	PUFA	-1	1	T	S	T	T	4	4	4	3	1
DG 34:0	DG	SFA	1	-1									1
DG 34:2	DG	PUFA	-1	1	T		T	T	16		2	3	1
DG 34:3	DG	PUFA	-1	1	T		T	T	16		2	2	0
DG 34:4	DG	PUFA	-1	1	T		T	T	2		16	4	0
DG 35:0	DG	SFA	1	-1									0
DG 35:1	DG	MUFA	-1	1	T		T	T	16		2	3	1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
DG 35:2	DG	PUFA	-1	1	T	T	T	T	16	3	2	2	1
DG 35:3	DG	PUFA	-1	1	T		T	T	16		2	2	1
DG 36:0	DG	SFA	1	-1									1
DG 36:1	DG	MUFA	-1	1									1
DG 36:2	DG	PUFA	-1	1	T	T	T	T	16	4	2	3	1
DG 36:3	DG	PUFA	1	1	S	T	T	T	16	4	2	3	1
DG 36:4	DG	PUFA	1	-1	T	T	T	T	3	4	2	3	0
DG 36:5	DG	PUFA	1	-1			S				2		0
DG 36:6	DG	PUFA	-1	1									0
DG 37:0	DG	SFA	1	-1									0
DG 37:5	DG	PUFA	-1	1									0
DG 38:0	DG	SFA	1	-1									1
DG 38:1	DG	MUFA	-1	1	T	T	T	T	6	21	5	4	0
DG 38:2	DG	PUFA	-1	1	T	T	T	T	3	17	3	3	1
DG 38:3	DG	PUFA	-1	1	T		T	T	16		2	3	1
DG 38:4	DG	PUFA	-1	1	T	T	T	T	3	4	3	3	1
DG 38:5	DG	PUFA	-1	1	T		T	T	16		2	3	1
DG 38:6	DG	PUFA	1	1	S		S		6		5		0
DG 38:7	DG	PUFA	1	1	T		T	T	6		5	1	0
DG 39:0	DG	SFA	1	-1				S				4	0
DG 40:0	DG	SFA	1	-1	T		T	T	8		7	4	0
DG 40:1	DG	MUFA	-1	1		S		S		7		4	0
DG 40:2	DG	PUFA	-1	1			S				3		0
DG 40:5	DG	PUFA	-1	1	T		T	T	16		2	2	0
DG 40:6	DG	PUFA	-1	1	T	T	T	T	8	3	14	2	1
DG 40:7	DG	PUFA	-1	1	T		T	T	7		5	3	1
DG 40:8	DG	PUFA	-1	1	T		T	T	7		2	3	0
DG 41:0	DG	SFA	1	-1	T		T	T	8		7	1	0
DG 42:0	DG	SFA	-1	1									0

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
DG 42:10	DG	PUFA	-1	1	T	T	T	T	2	7	16	3	0
dodecenoylcarnitine	untarg		1	-1									1
gamma-butyrolactone	untarg		-1	1		S		S		3		4	0
gamma-l-glutamyl-l-cysteine	untarg		1	1									0
glu-ile/leu / l-gamma-glutamyl-l-isoleucine	untarg		1	-1									0
gluconic acid	untarg		1	-1									1
glucose	untarg		-1	1									1
glutamate	untarg		1	1									0
glutamine	untarg		1	1									0
glutamyl-phenylalanine	untarg		-1	1									1
glutarate	untarg		-1	1		S		S		10		7	0
glyceraldehyde	untarg		-1	1									0
glycochenodeoxycholate	untarg		1	1									1
glycocholate	untarg		1	1									0
guanosine	untarg		-1	1	T		T	T	4		4	4	1
heptadecanoate	untarg		-1	1									0
hexadecasphinganine	untarg		1	1									0
hexanoylcarnitine	untarg		1	-1	T		T	T	5		15	1	1
hippurate	untarg		1	-1									0
histidinyl-tryptophan	untarg		-1	-1									0
hypaphorine	untarg		-1	-1		T	T	T		7	4	7	0
hypoxanthine	untarg		-1	1		S		S		7		1	0
ile-ile	untarg		-1	1									1
indole-3-acetate	untarg		1	1									0
inosine	untarg		-1	1	T		T	T	4		4	3	1
isoleucine	untarg		1	-1									1
isovalerylcarnitine	untarg		1	-1									1
krynurenic acid	untarg		-1	-1	S				4				1
kynurenine	untarg		-1	1	T		T	T	3		15	3	0

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
l-carnitine	untarg		1	-1									1
l-histidine	untarg		1	-1									0
l-rhamnose	untarg		1	1									1
lauroylcarnitine	untarg		-1	-1	S		S		6		15		1
leucine	untarg		-1	1									1
lignoceric acid	untarg		1	-1		T	T	T		2	6	3	1
lysine	untarg		1	1									1
lysoPC 14:0	lysoPC	SFA	-1	1	T	T	T	T	2	10	16	3	1
lysoPC 15:0	lysoPC	SFA	-1	1	T	T	T	T	7	9	15	4	1
lysoPC 16:0	lysoPC	SFA	-1	1		T	T	T		9	14	4	1
lysoPC 16:1	lysoPC	MUFA	-1	1	T	T	T	T	7	13	15	9	1
lysoPC 17:1	lysoPC	MUFA	-1	1	T	T	T	T	7	8	15	4	1
lysoPC 18:0	lysoPC	SFA	-1	1		T	T	T		4	14	3	1
lysoPC 18:1	lysoPC	MUFA	-1	-1	S	T	T	T	7	9	14	4	1
lysoPC 18:2	lysoPC	PUFA	1	-1	T		T	T	7		14	4	1
lysoPC 18:3	lysoPC	PUFA	-1	1	T	T	T	T	7	9	14	4	1
lysoPC 19:0	lysoPC	SFA	-1	-1		T	T	T		9	15	4	1
lysoPC 20:0	lysoPC	SFA	1	-1	T	T	T	T	7	9	14	4	1
lysoPC 20:1	lysoPC	MUFA	-1	-1	T		T	T	7		5	4	1
lysoPC 20:2	lysoPC	PUFA	1	-1	T		T	T	7		14	4	1
lysoPC 20:3	lysoPC	PUFA	-1	1	T	T	T	T	7	9	14	4	1
lysoPC 20:4	lysoPC	PUFA	-1	1		T	T	T		9	14	4	1
lysoPC 20:5	lysoPC	PUFA	1	-1	T		T	T	7		14	4	1
lysoPC 22:0	lysoPC	SFA	1	-1	T		T	T	4		4	4	1
lysoPC 22:1	lysoPC	MUFA	-1	1	S	T	T	T	7	9	14	4	1
lysoPC 22:4	lysoPC	PUFA	-1	1	T	T	T	T	7	9	14	4	1
lysoPC 22:5	lysoPC	PUFA	-1	1	T	T	T	T	7	9	14	4	1
lysoPC 22:6	lysoPC	PUFA	-1	1				S				4	1
lysoPC 23:0	lysoPC	SFA	-1	-1									1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
lysoPC 24:0	lysoPC	SFA	-1	-1	T		T	T	4		4	20	1
lysoPC 24:1	lysoPC	MUFA	-1	-1				S				4	0
lysoPC 26:0	lysoPC	SFA	-1	1									1
lysoPC 26:1	lysoPC	MUFA	-1	-1									0
lysoPC 26:2	lysoPC	PUFA	1	-1	S		S		5		4		1
lysoPE 16:0	lysoPE	SFA	-1	1	T	T	T	T	4	9	4	16	0
lysoPE 16:1	lysoPE	MUFA	1	1									1
lysoPE 18:0	lysoPE	SFA	-1	1		T	T	T		19	15	15	1
lysoPE 18:1	lysoPE	MUFA	-1	-1				S				4	1
lysoPE 18:2	lysoPE	PUFA	1	-1	T		T	T	7		15	4	0
lysoPE 18:3	lysoPE	PUFA	-1	1	T	T	T	T	5	7	3	9	1
lysoPE 20:3	lysoPE	PUFA	-1	1	T	T	T	T	2	10	16	3	1
lysoPE 20:4	lysoPE	PUFA	-1	1	T	T	T	T	7	10	15	7	0
lysoPE 20:5	lysoPE	PUFA	-1	-1									0
lysoPE 22:4	lysoPE	PUFA	-1	1	T	T	T	T	3	9	3	4	0
lysoPE 22:5	lysoPE	PUFA	-1	1	T	T	T	T	7	8	15	5	1
lysoPE 22:6	lysoPE	PUFA	-1	1	T	T	T	T	7	15	15	20	0
lysoPE 24:1	lysoPE	MUFA	-1	1									0
lysoPE 26:0	lysoPE	SFA	1	-1									1
malate	untarg		-1	-1									1
mandelic acid	untarg		1	1									1
methionine	untarg		1	1									1
methyl beta-d-galactoside	untarg		1	1		T	T	T		7	4	1	0
methyl indole-3-acetate	untarg		-1	1									0
MG 16:0	MG	SFA	-1	1	T	T	T	T	6	20	6	7	1
MG 17:0	MG	SFA	-1	-1									1
MG 18:0	MG	SFA	-1	-1		T	T	T		9	5	4	0
MG 18:1	MG	MUFA	-1	-1		T	T	T		9	15	4	0
MG 18:2	MG	PUFA	1	-1	T	T	T	T	6	16	14	4	0

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
MG 18:4	MG	PUFA	1	-1	T		T	T	10		9	20	0
MG 19:0	MG	SFA	1	-1									0
MG 24:0	MG	SFA	1	1									0
myo-inositol	untarg		-1	1									0
n-acetyl-d-tryptophan	untarg		1	1									1
n-acetyl-dl-methionine	untarg		-1	1	T		T	T	6		15	3	0
n-acetyl-dl-serine	untarg		-1	1	S		S		6		15		1
n-acetyl-l-alanine (-h+na)	untarg		-1	1									0
n-acetyl-l-aspartic acid	untarg		-1	1	S				3				0
n-acetyl-l-leucine	untarg		-1	1									1
n-acetyl-l-phenylalanine	untarg		-1	1	T		T	T	7		15	1	1
n-acetylglycine	untarg		-1	1									1
n-alpha-acetyl-l-lysine	untarg		1	-1									0
n-cyclohexylformamide	untarg		1	-1									0
n-gamma-l-glutamyl-l-methionine	untarg		-1	1									0
n-methyl-l-glutamate	untarg		1	1									1
n2_n2-dimethylguanosine	untarg		-1	1									0
n6,n6,n6-trimethyl-l-lysine	untarg		1	-1									1
nicotinamide	untarg		-1	1									1
o-acetylcarnitine	untarg		1	-1									0
octanoylcarnitine	untarg		1	-1	S		S		6		15		1
octenoylcarnitine	untarg		1	-1									0
oleoylcarnitine	untarg		-1	-1			S				15		1
PA 34:0	PA	SFA	-1	-1									0
PA 34:2	PA	PUFA	-1	-1									0
PA 40:1	PA	MUFA	1	1									0
palmitoylcarnitine	untarg		-1	1		T	T	T		9	15	4	1
pantothenate	untarg		-1	1									1
paraxanthine	untarg		-1	1	T		T	T	7		14	15	0

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
PC 19:2	PC	PUFA	-1	-1				S				4	1
PC 21:0	PC	SFA	-1	-1									0
PC 26:0	PC	SFA	-1	1	T	T	T	T	4	3	4	2	0
PC 28:0	PC	SFA	-1	1	T	T	T	T	5	7	4	2	0
PC 29:0	PC	SFA	-1	1	T	T	T	T	5	3	4	2	0
PC 30:0	PC	SFA	-1	1	T	T	T	T	1	8	4	18	1
PC 30:1	PC	MUFA	-1	1	T	T	T	T	4	3	4	2	1
PC 30:2	PC	PUFA	-1	1	T	T	T	T	4	10	4	3	1
PC 30:3	PC	PUFA	-1	1	T	T	T	T	4	7	4	4	0
PC 31:0	PC	SFA	-1	1		T	T	T		7	15	3	1
PC 31:1	PC	MUFA	1	-1	T		T	T	3		3	7	0
PC 32:0	PC	SFA	-1	1	T	T	T	T	3	7	3	6	1
PC 32:1	PC	MUFA	-1	1		S		S		3		2	1
PC 32:2	PC	PUFA	-1	1	S	T	T	T	2	4	16	3	1
PC 32:3	PC	PUFA	-1	1	T	T	T	T	3	9	4	4	1
PC 32:4	PC	PUFA	-1	1	T	T	T	T	4	3	4	2	0
PC 33:0	PC	SFA	1	1									1
PC 33:1	PC	MUFA	-1	1	T	T	T	T	3	13	3	20	1
PC 33:2	PC	PUFA	-1	-1		T	T	T		7	4	3	1
PC 33:3	PC	PUFA	-1	1	T	T	T	T	2	7	16	4	1
PC 34:0	PC	SFA	1	-1									0
PC 34:1	PC	MUFA	-1	1	T	T	T	T	3	7	3	7	1
PC 34:2	PC	PUFA	-1	1	T	T	T	T	2	8	5	20	1
PC 34:3	PC	PUFA	-1	-1									1
PC 34:4	PC	PUFA	-1	1	T	T	T	T	5	9	4	1	1
PC 34:5	PC	PUFA	-1	1	T	T	T	T	2	4	4	3	1
PC 35:0	PC	SFA	-1	1	T	S	T	T	4	6	4	10	1
PC 35:1	PC	MUFA	-1	1	T	T	T	T	6	3	15	2	1
PC 35:2	PC	PUFA	-1	-1									1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
PC 35:3	PC	PUFA	1	-1	T		T	T	4		4	1	1
PC 35:4	PC	PUFA	-1	1	T	T	T	T	11	9	3	3	1
PC 35:5	PC	PUFA	-1	1									0
PC 35:6	PC	PUFA	-1	1	T		T	T	3		3	1	0
PC 36:0	PC	SFA	-1	1	T	T	T	T	6	21	15	4	1
PC 36:1	PC	MUFA	-1	1	T	T	T	T	2	21	16	20	1
PC 36:2	PC	PUFA	-1	-1	T		T	T	2		16	4	1
PC 36:3	PC	PUFA	1	-1	T		T	T	7		15	1	1
PC 36:4	PC	PUFA	-1	1	S		S		6		7		1
PC 36:5	PC	PUFA	-1	1	T	T	T	T	2	3	5	2	1
PC 36:6	PC	PUFA	-1	1									1
PC 36:7	PC	PUFA	-1	1	S	T	T	T	2	9	4	3	0
PC 37:1	PC	MUFA	1	-1	T	S	T	T	6	9	6	4	1
PC 37:2	PC	PUFA	1	1									1
PC 37:3	PC	PUFA	-1	1									1
PC 37:4	PC	PUFA	-1	-1	S		S		8		14		1
PC 37:5	PC	PUFA	-1	1	T	T	T	T	4		4	4	1
PC 37:6	PC	PUFA	-1	1		T	T	T		3	4	14	1
PC 37:7	PC	PUFA	-1	1	T		T	T	3		3	1	1
PC 38:0	PC	SFA	1	-1	T	T	T	T	3	9	3	4	0
PC 38:1	PC	MUFA	-1	1									1
PC 38:2	PC	PUFA	-1	-1	S		S		3		3		1
PC 38:3	PC	PUFA	-1	1	T	T	T	T	2	4	16	2	1
PC 38:4	PC	PUFA	-1	1	T	T	T	T	8	7	7	1	1
PC 38:5	PC	PUFA	-1	1	T	T	T	T	6	15	7	20	1
PC 38:6	PC	PUFA	-1	1	T		T	T	3		14	14	1
PC 38:7	PC	PUFA	-1	1		T	T	T		21	4	3	1
PC 38:8	PC	PUFA	-1	1	T	T	T	T	2	9	16	3	1
PC 38:9	PC	PUFA	-1	1	T	T	T	T	4	14	4	13	0

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
PC 39:2	PC	PUFA	1	-1									0
PC 39:3	PC	PUFA	-1	1									1
PC 39:4	PC	PUFA	-1	1	T	T	T	T	6	7	4	1	1
PC 39:5	PC	PUFA	-1	1	T	T	T	T	3	18	3	19	1
PC 39:6	PC	PUFA	-1	1	T	T	T	T	7	9	8	4	1
PC 39:7	PC	PUFA	-1	1									0
PC 39:8	PC	PUFA	-1	1									0
PC 40:1	PC	MUFA	1	-1	T		T	T	4		7	4	0
PC 40:10	PC	PUFA	-1	1	S		S		4		4		1
PC 40:2	PC	PUFA	1	-1	T		T	T	5		13	4	1
PC 40:3	PC	PUFA	-1	1		T	T	T		8	14	20	1
PC 40:4	PC	PUFA	-1	1	T	T	T	T	2	7	16	7	1
PC 40:5	PC	PUFA	-1	1	T	T	T	T	16	6	2	20	1
PC 40:6	PC	PUFA	-1	1	T	T	T	T	8	3	7	20	1
PC 40:7	PC	PUFA	-1	-1									1
PC 40:8	PC	PUFA	-1	-1	T		T	T	14		11	4	1
PC 41:6	PC	PUFA	-1	1									1
PC 42:1	PC	MUFA	-1	1									1
PC 42:10	PC	PUFA	-1	1	T	T	T	T	3	10	3	4	1
PC 42:11	PC	PUFA	-1	-1		S				10			1
PC 42:2	PC	PUFA	1	-1	T		T	T	3		3	20	1
PC 42:3	PC	PUFA	1	-1	T	T	T	T		6	15	20	1
PC 42:4	PC	PUFA	-1	-1		S		S		6		20	0
PC 42:5	PC	PUFA	-1	1	T	T	T	T	4	7	3	7	1
PC 42:6	PC	PUFA	-1	1	T		T	T	12		2	11	0
PC 42:7	PC	PUFA	-1	-1	S		S		4		4		1
PC 42:8	PC	PUFA	1	-1	T		T	T	3		3	7	1
PC 42:9	PC	PUFA	1	-1	T	T	T	T	8	9	14	16	0
PC 44:4	PC	PUFA	-1	-1									1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
PC 44:5	PC	PUFA	-1	1									1
PC 46:5	PC	PUFA	-1	1									0
PE 32:1	PE	MUFA	1	1	T		T	T	3		3	20	0
PE 32:2	PE	PUFA	1	1									0
PE 33:0	PE	SFA	1	-1									0
PE 33:1	PE	MUFA	-1	1	T		T	T	3		3	8	0
PE 33:2	PE	PUFA	-1	1	S		S		3		3		0
PE 34:0	PE	SFA	1	-1									1
PE 34:1	PE	PUFA	-1	-1	T	S	T	T	7	6	5	20	0
PE 34:2	PE	PUFA	-1	-1									0
PE 34:3	PE	PUFA	-1	1	T	T	T	T	4	21	4	3	1
PE 35:0	PE	MUFA	1	1									0
PE 35:1	PE	PUFA	-1	1									0
PE 35:2	PE	PUFA	-1	1									1
PE 35:3	PE	PUFA	-1	1	S		S		4		4		0
PE 35:4	PE	SFA	-1	1	S		S		3		3		0
PE 36:0	PE	MUFA	1	1									1
PE 36:1	PE	PUFA	-1	-1		S		S		8		20	0
PE 36:2	PE	PUFA	-1	-1									1
PE 36:3	PE	PUFA	1	-1	S		S		3		3		1
PE 36:4	PE	PUFA	-1	1	T		T	T	3		3	1	0
PE 36:5	PE	PUFA	-1	1	T	T	T	T	3	14	3	2	1
PE 37:2	PE	PUFA	-1	-1	S				4				0
PE 37:3	PE	PUFA	-1	1	T		T	T	3		3	7	0
PE 37:4	PE	PUFA	-1	1	T	T	T	T	3	9	3	1	0
PE 37:6	PE	PUFA	-1	1	S	S	S	S	4	7	4	4	0
PE 38:1	PE	MUFA	-1	1		T	T	T		7	10	7	0
PE 38:2	PE	PUFA	-1	1									1
PE 38:3	PE	PUFA	-1	-1									0

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
PE 38:4	PE	PUFA	-1	1	T	T	T	T	3	5	3	1	1
PE 38:5	PE	PUFA	-1	1	T		T	T	3		3	1	0
PE 38:6	PE	PUFA	-1	1	T		T	T	3		3	1	0
PE 38:7	PE	PUFA	-1	1									0
PE 39:6	PE	PUFA	-1	1	T	T	T	T	3	7	3	3	0
PE 40:3	PE	PUFA	-1	1		T	T	T		21	3	1	0
PE 40:4	PE	PUFA	-1	1	S	S	S	S	3	4	3	1	1
PE 40:5	PE	PUFA	-1	1	T	T	T	T	3	4	3	1	1
PE 40:6	PE	PUFA	-1	1	T		T	T	3		3	1	0
PE 40:7	PE	PUFA	1	-1	S		S		3		3		0
PE 40:8	PE	PUFA	1	-1	S		S		3		3		1
PE 42:8	PE	PUFA	-1	1		S	S	S		9	4	4	0
PE 42:9	PE	PUFA	1	-1									0
PG 33:0	PG	SFA	-1	-1									1
PG 34:1	PG	MUFA	-1	1									0
PG 34:2	PG	PUFA	1	-1									0
PG 36:0	PG	SFA	-1	1		T	T	T		8	4	20	0
PG 36:3	PG	PUFA	-1	1									0
PG 38:4	PG	PUFA	-1	-1			S				7		0
phenyl acetate	untarg		-1	1									1
phenylalanine	untarg		1	1									1
phenylephrine	untarg		1	-1		S		S		9		3	0
phytanate	untarg		-1	1		T	T	T		7	15	3	0
PI 34:1	PI	MUFA	-1	1	T	T	T	T	4	7	4	7	0
PI 34:2	PI	PUFA	1	1									1
PI 36:1	PI	MUFA	1	-1									0
PI 36:2	PI	PUFA	1	-1	T	S	T	T	3	7	3	7	1
PI 36:3	PI	PUFA	1	-1	T	T	T	T	4	7	4	7	0
PI 36:4	PI	PUFA	-1	1	T		T	T	2		3	7	1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
PI 38:3	PI	PUFA	-1	1									1
PI 38:4	PI	PUFA	-1	1									1
PI 38:5	PI	PUFA	-1	1									1
PI 40:5	PI	PUFA	-1	1	T		T	T	6		15	2	1
PI 40:6	PI	PUFA	1	1									0
pipecolate	untarg		1	1									1
piperine	untarg		-1	1		T	T	T		8	8	1	0
plasmenyl-PC 18:0	plasmenyl-PC	SFA	-1	1		T	T	T		9	14	4	1
plasmenyl-PC 20:0	plasmenyl-PC	SFA	1	-1	S		S		10		9		1
plasmenyl-PC 29:0	plasmenyl-PC	SFA	-1	1		T	T	T		12	8	1	0
plasmenyl-PC 34:1	plasmenyl-PC	MUFA	-1	-1									0
plasmenyl-PC 34:2	plasmenyl-PC	PUFA	-1	1		T	T	T		9	7	4	0
plasmenyl-PC 36:3	plasmenyl-PC	PUFA	-1	-1		S		S		3		2	0
plasmenyl-PC 36:4	plasmenyl-PC	PUFA	1	-1									0
plasmenyl-PC 38:1	plasmenyl-PC	MUFA	-1	-1		T	T	T		9	5	3	0
plasmenyl-PC 38:3	plasmenyl-PC	PUFA	1	-1	T		T	T	6		15	1	0
plasmenyl-PC 38:5	plasmenyl-PC	PUFA	-1	-1	T		T	T	3		15	1	0
plasmenyl-PC 40:5	plasmenyl-PC	PUFA	1	-1									0
plasmenyl-PC 40:6	plasmenyl-PC	PUFA	1	-1			S				8		0
plasmenyl-PC 42:5	plasmenyl-PC	PUFA	1	-1	S	T	T	T	6	5	6	1	0
plasmenyl-PC 44:4	plasmenyl-PC	PUFA	-1	1									0
plasmenyl-PE 32:1	plasmenyl-PE	MUFA	-1	1									0
plasmenyl-PE 34:0	plasmenyl-PE	SFA	1	-1									0
plasmenyl-PE 34:1	plasmenyl-PE	MUFA	-1	-1		T	T	T			14	20	1
plasmenyl-PE 34:2	plasmenyl-PE	PUFA	-1	-1	T		T	T	4		4	20	1
plasmenyl-PE 34:3	plasmenyl-PE	PUFA	-1	-1	T		T	T	6		6	1	0
plasmenyl-PE 36:0	plasmenyl-PE	SFA	-1	1	S	T	T	T	3	16	3	4	0
plasmenyl-PE 36:1	plasmenyl-PE	MUFA	-1	-1		T	T	T		8	4	20	1
plasmenyl-PE 36:2	plasmenyl-PE	PUFA	-1	-1	T		T	T	4		14	20	0

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
plasmenyl-PE 36:3	plasmenyl-PE	PUFA	-1	-1	T		T	T	4		4	20	1
plasmenyl-PE 36:4	plasmenyl-PE	PUFA	-1	-1	T		T	T	4		4	20	1
plasmenyl-PE 36:5	plasmenyl-PE	PUFA	-1	1		T	T	T		9	4	1	0
plasmenyl-PE 38:1	plasmenyl-PE	MUFA	-1	1									1
plasmenyl-PE 38:2	plasmenyl-PE	PUFA	1	-1									1
plasmenyl-PE 38:3	plasmenyl-PE	PUFA	-1	1	T	T	T	T	4	8	4	20	1
plasmenyl-PE 38:4	plasmenyl-PE	PUFA	-1	-1		T	T	T		8	4	20	1
plasmenyl-PE 38:5	plasmenyl-PE	PUFA	-1	-1	S	T	T	T	4	8	4	20	1
plasmenyl-PE 38:6	plasmenyl-PE	PUFA	-1	-1	T		T	T	4		4	20	1
plasmenyl-PE 40:3	plasmenyl-PE	PUFA	1	-1	S		S		4		4		0
plasmenyl-PE 40:4	plasmenyl-PE	PUFA	-1	-1									1
plasmenyl-PE 40:5	plasmenyl-PE	PUFA	-1	-1		T	T	T		8	14	20	1
plasmenyl-PE 40:6	plasmenyl-PE	PUFA	-1	-1	T		T	T	7		4	20	1
plasmenyl-PE 42:5	plasmenyl-PE	PUFA	-1	-1									1
plasmenyl-PE 42:6	plasmenyl-PE	PUFA	1	-1	T		T	T	7		5	4	0
possible peptide-262.1341-6.151	untarg		-1	1									0
possible peptide-310.1151-2.804	untarg		-1	1									0
possible peptide-414.2046-18.389	untarg		1	-1									0
proline	untarg		-1	1		S				8			1
propionylcarnitine	untarg		-1	1									1
pyridoxamine	untarg		1	1	S		S		3		3		0
raffinose	untarg		-1	1									0
s-allyl-l-cysteine	untarg		-1	1									0
sarcosine	untarg		1	-1									1
serine	untarg		1	1									1
serinyl-leucine	untarg		-1	1		S	S			10	6		0
SM 21:0	SM	SFA	-1	-1									0
SM 21:1	SM	MUFA	-1	-1	T	S	T	T	7	9	14	4	1
SM 29:1	SM	MUFA	-1	1		T	T	T		14	15	17	0

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
SM 30:0	SM	SFA	-1	1	T	T	T	T	14	3	4	2	0
SM 30:1	SM	MUFA	-1	1	T	T	T	T	4	7	4	3	0
SM 31:1	SM	MUFA	-1	1	T	T	T	T	4	7	4	7	1
SM 32:0	SM	SFA	-1	1	S	T	T	T	4	6	4	20	1
SM 32:1	SM	MUFA	-1	1	T	T	T	T	4	7	4	7	1
SM 32:2	SM	PUFA	-1	1	T	T	T	T	4	7	4	3	1
SM 33:0	SM	SFA	-1	-1		S		S		8		20	0
SM 33:1	SM	MUFA	-1	1		T	T	T		7	4	20	1
SM 33:2	SM	PUFA	-1	-1									0
SM 34:1	SM	MUFA	1	1									1
SM 34:2	SM	PUFA	-1	-1									1
SM 35:0	SM	SFA	-1	1									0
SM 35:1	SM	MUFA	1	-1	T	S	T	T	3	5	3	3	1
SM 35:2	SM	PUFA	-1	1		T	T	T		7	7	1	1
SM 36:0	SM	SFA	-1	1									1
SM 36:1	SM	MUFA	-1	1	T	T	T	T	9	7	7	1	1
SM 36:2	SM	PUFA	-1	-1	S		S	S	9		16	3	1
SM 37:0	SM	SFA	-1	1									0
SM 37:1	SM	MUFA	-1	-1		T	T	T		7	7	7	1
SM 37:2	SM	PUFA	-1	1		T	T	T		8	7	20	1
SM 38:0	SM	SFA	-1	-1									1
SM 38:1	SM	MUFA	1	-1									1
SM 38:2	SM	PUFA	-1	-1	T		T	T	9		8	3	1
SM 38:4	SM	PUFA	1	-1	T		T	T	8		7	1	0
SM 39:0	SM	SFA	-1	-1									0
SM 39:1	SM	MUFA	-1	-1	T	T	T	T	6	7	6	3	1
SM 39:2	SM	PUFA	-1	-1		T	T	T		6	15	10	1
SM 39:4	SM	PUFA	-1	-1									0
SM 39:5	SM	PUFA	-1	1									0

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
SM 40:0	SM	SFA	1	-1	T		T	T	6		6	4	0
SM 40:1	SM	MUFA	-1	-1	T		T	T	6		6	3	1
SM 40:2	SM	PUFA	-1	-1	T	T	T	T	3	3	3	2	1
SM 40:4	SM	PUFA	1	1									1
SM 40:5	SM	PUFA	-1	1									0
SM 41:1	SM	MUFA	-1	-1	T		T	T	6		3	4	1
SM 41:2	SM	PUFA	-1	-1			S				6		1
SM 41:4	SM	PUFA	-1	-1		T	T	T		7	6	4	1
SM 41:5	SM	PUFA	1	-1									0
SM 42:0	SM	SFA	1	-1	T		T	T	6		6	4	0
SM 42:1	SM	MUFA	-1	-1	T		T	T	6		6	4	1
SM 42:2	SM	PUFA	1	-1									1
SM 42:4	SM	PUFA	1	1	T		T	T	6		6	4	1
SM 42:5	SM	PUFA	1	1									1
SM 43:1	SM	MUFA	-1	-1		T	T	T		7	3	20	1
SM 43:2	SM	PUFA	-1	1									1
SM 43:4	SM	PUFA	-1	-1									0
SM 43:5	SM	PUFA	-1	1									0
SM 44:1	SM	MUFA	-1	1									1
SM 44:2	SM	PUFA	-1	-1									1
stearoylcarnitine	untarg		-1	-1			S				15		1
succinate	untarg		-1	1									1
sucrose	untarg		-1	1									1
TG 40:0	TG	SFA	-1	-1									1
TG 42:0	TG	SFA	-1	1		T	T	T		3	16	2	1
TG 42:1	TG	MUFA	-1	1	S	T	T	T	16	3	2	2	1
TG 42:2	TG	PUFA	-1	1	S		S		16		2		1
TG 44:1	TG	MUFA	-1	1	T	T	T	T	2	11	16	2	1
TG 44:2	TG	PUFA	-1	1	T	T	T	T	16	4	2	2	1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
TG 46:0	TG	SFA	-1	1	T	T	T	T	2	3	16	2	1
TG 46:1	TG	MUFA	-1	1	T	T	T	T	16	3	16	2	1
TG 46:2	TG	PUFA	-1	1	T	T	T	T	16	3	16	2	1
TG 46:3	TG	PUFA	-1	1	T	T	T	T	16	4	16	2	1
TG 48:0	TG	SFA	-1	1	T	T	T	T	2	3	16	2	1
TG 48:1	TG	MUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 48:2	TG	PUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 48:3	TG	PUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 49:0	TG	SFA	-1	1	T	T	T	T	2	3	16	2	1
TG 49:1	TG	MUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 49:2	TG	PUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 49:3	TG	PUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 50:0	TG	SFA	-1	1	T	T	T	T	16	3	16	2	1
TG 50:1	TG	MUFA	-1	1	T	T	T	T	16	3	2	2	1
TG 50:2	TG	PUFA	-1	1	T	T	T	T	5	3	4	2	1
TG 50:3	TG	PUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 50:4	TG	PUFA	-1	1	T	S	T	T	16	4	16	3	1
TG 50:5	TG	PUFA	-1	1	T	S	T	T	2	4	16	3	1
TG 51:1	TG	MUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 51:2	TG	PUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 51:3	TG	PUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 51:4	TG	PUFA	-1	1	S	S	S	S	2	4	16	3	1
TG 51:5	TG	PUFA	-1	1									1
TG 52:0	TG	SFA	-1	1	T	T	T	T	16	10	2	4	1
TG 52:1	TG	MUFA	-1	1	T	T	T	T	16	3	2	2	1
TG 52:2	TG	PUFA	-1	1	T	T	T	T	3	3	3	2	1
TG 52:3	TG	PUFA	-1	1	T	S	T	T	15	4	1	3	1
TG 52:4	TG	PUFA	1	-1		T	T	T		4	3	3	1
TG 52:5	TG	PUFA	-1	1	T		T	T	13		5	2	1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
TG 52:6	TG	PUFA	1	-1	S		S		2		5		1
TG 52:7	TG	PUFA	1	1	S		S	S	2		16	3	1
TG 53:0	TG	SFA	-1	1	T		T	T	2		16	4	1
TG 53:1	TG	MUFA	-1	1	T	T	T	T	2	3	16	2	1
TG 53:2	TG	PUFA	-1	1	T	T	T	T	16	3	16	2	1
TG 53:3	TG	PUFA	-1	1	T	S	T	T	2	4	16	2	1
TG 53:4	TG	PUFA	1	-1	T	T	T	T	6	4	5	3	1
TG 53:5	TG	PUFA	1	1									1
TG 53:6	TG	PUFA	-1	1	T		T	T	6		5	3	1
TG 53:7	TG	PUFA	-1	1									1
TG 54:1	TG	MUFA	-1	1	T	T	T	T	2	4	2	3	1
TG 54:2	TG	PUFA	-1	1	T	T	T	T	16	4	2	3	1
TG 54:3	TG	PUFA	1	-1	S		S		2		16		1
TG 54:4	TG	PUFA	1	-1	T	T	T	T	16	6	16	6	1
TG 54:5	TG	PUFA	1	-1	T	T	T	T	6	6	5	6	1
TG 54:6	TG	PUFA	1	-1	T	T	T	T	6	6	5	6	1
TG 54:7	TG	PUFA	1	-1	T	T	T	T	6	6	5	6	1
TG 54:8	TG	PUFA	1	1	S		S	S	2		5	6	1
TG 54:9	TG	PUFA	-1	1	T		T	T	6		5	3	0
TG 55:0	TG	SFA	-1	1									1
TG 55:1	TG	MUFA	-1	1	T	T	T	T	6	13	6	9	1
TG 55:2	TG	PUFA	-1	1	T	T	T	T	16	10	2	20	1
TG 55:3	TG	PUFA	-1	1									1
TG 55:4	TG	PUFA	1	-1									1
TG 55:5	TG	PUFA	1	-1									0
TG 55:6	TG	PUFA	-1	1	T	T	T	T	6	2	5	15	1
TG 55:7	TG	PUFA	1	1			S				14		1
TG 55:8	TG	PUFA	-1	1	T	T	T	T	6	2	5	1	1
TG 56:0	TG	SFA	1	1									1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
TG 56:1	TG	MUFA	-1	1	T	T	T	T	16	10	2	4	1
TG 56:10	TG	PUFA	1	1	S		S		6		5		1
TG 56:2	TG	PUFA	-1	1	T	T	T	T	16	4	2	3	1
TG 56:3	TG	PUFA	-1	1	S		S		3		3		1
TG 56:4	TG	PUFA	1	1	S		S		16		2		1
TG 56:5	TG	PUFA	1	1	S	S	S		7	6	5		1
TG 56:6	TG	PUFA	1	-1	S	S	S	S	7	6	14	6	1
TG 56:7	TG	PUFA	1	1									1
TG 56:8	TG	PUFA	1	-1		T	T	T		6	5	6	1
TG 56:9	TG	PUFA	1	-1	S	T	T	T	6	6	5	6	1
TG 57:1	TG	MUFA	-1	1									1
TG 57:2	TG	PUFA	-1	1	T		T	T	16		12	4	1
TG 57:3	TG	PUFA	1	1									1
TG 57:6	TG	PUFA	-1	1		T	T	T		20	4	15	0
TG 57:8	TG	PUFA	-1	1									0
TG 58:0	TG	SFA	-1	1	T		T	T	16		2	4	1
TG 58:1	TG	MUFA	-1	1									1
TG 58:10	TG	PUFA	1	-1	T	T	T	T	4	6	14	6	1
TG 58:11	TG	PUFA	1	1	S	T	T	T	6	7	5	6	1
TG 58:12	TG	PUFA	-1	1	T		T	T	6		5	4	0
TG 58:2	TG	PUFA	-1	-1	S	S	S	S	16	4	2	3	1
TG 58:3	TG	PUFA	1	-1	T		T	T	16		2	3	1
TG 58:4	TG	PUFA	1	1	S		S		3		3		1
TG 58:5	TG	PUFA	-1	1									0
TG 58:6	TG	PUFA	-1	1									1
TG 58:7	TG	PUFA	-1	1	T		T	T	7		14	6	1
TG 58:8	TG	PUFA	1	1	S		S		7		14		1
TG 58:9	TG	PUFA	1	-1	S	T	T	T	7	6	14	6	1
TG 60:10	TG	PUFA	-1	1	S		S		7		14		1

Table C.1: *Summary details for PCST results.* Saturation levels (sat) are abbreviated as SFA (saturated fatty acid), MUFA (monounsaturated fatty acid) and PUFA (polyunsaturated fatty acid). Columns $d_{21} - d_0$ and $d_{42} - d_{21}$ contain the sign of this difference. Consensus graph nodes are terminal (T), or steiner (S). Columns labeled M_{21u} and similar contain module numbers from associated consensus graph. A 1 in the column HCR/LCR indicates the compound was present in HCR/LCR data, 0 indicates otherwise.

Compound Name	class	sat	$d_{21} - d_0$	$d_{42} - d_{21}$	CG_{42d} node	CG_{21d} node	CG_{42u} node	CG_{21u} node	M_{42d}	M_{21d}	M_{42u}	M_{21u}	HCR/LCR
TG 60:11	TG	PUFA	-1	1	T	T	T	T	7	8	14	20	1
TG 60:12	TG	PUFA	1	1									1
TG 60:13	TG	PUFA	-1	1				S				2	1
TG 60:3	TG	PUFA	1	-1	S		S		16		2		1
TG 60:8	TG	PUFA	-1	1	T	S	T	T	7	8	14	20	0
TG 60:9	TG	PUFA	-1	1	T		T	T	7		14	3	1
TG 62:1	TG	MUFA	-1	1			S				16		1
TG 62:12	TG	PUFA	-1	1									0
TG 62:13	TG	PUFA	-1	1	T		T	T	4		4	1	1
TG 62:14	TG	PUFA	1	1									1
theobromine	untarg		-1	1	T		T	T	7		14	7	0
theophylline	untarg		-1	1									0
threonine	untarg		-1	-1		S		S		7		4	1
thymine-d4(methyl-d3,6-d1) [istd]	untarg		-1	1									0
thyroxine	untarg		-1	1		S		S		4		3	1
tryptophan	untarg		1	1									1
tyrosine	untarg		-1	1									1
urate	untarg		1	1									1
uridine	untarg		-1	-1									1
ursodiol	untarg		-1	1		T	T	T		4	3	3	1
valine	untarg		1	1									1
xanthine	untarg		-1	1									0
xylose	untarg		1	-1									0

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," 1995.

[2] t. M. C. Consortium, P. A. working group of the International Cancer Genome, P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander, B. J. Raphael, D. S. Marks, B. F. F. Ouellette, A. Valencia, G. D. Bader, P. C. Boutros, J. M. Stuart, R. Linding, N. Lopez-Bigas, and L. D. Stein, "Pathway and network analysis of cancer genomes," *Nature Methods*, vol. 12, pp. 615–621, jul 2015.

[3] J. W. K. Ho, M. Stefani, C. G. dos Remedios, and M. A. Charleston, "Differential variability analysis of gene expression and its application to human diseases.," *Bioinformatics (Oxford, England)*, vol. 24, pp. i390–8, jul 2008.

[4] J. Gillis and P. Pavlidis, "A methodology for the analysis of differential coexpression across the human lifespan.," *BMC bioinformatics*, vol. 10, p. 306, jan 2009.

[5] W. Weckwerth, M. E. Loureiro, K. Wenzel, and O. Fiehn, "Differential metabolic networks unravel the effects of silent plant phenotypes.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 7809–14, may 2004.

[6] R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth, "Observing and interpreting correlations in metabolomic networks.," *Bioinformatics (Oxford, England)*, vol. 19, pp. 1019–26, may 2003.

[7] L. Yetukuri, M. Katajamaa, G. Medina-Gomez, T. Seppänen-Laakso, A. Vidal-Puig, and M. Oresic, "Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis.," *BMC systems biology*, vol. 1, p. 12, feb 2007.

[8] E. Petsalaki, A. O. Helbig, A. Gopal, A. Pasculescu, F. P. Roth, and T. Pawson, "SELPHI: correlation-based identification of kinase-associated networks from global phospho-proteomics data sets.," *Nucleic acids research*, vol. 43, pp. W276–82, jul 2015.

[9] N. Krämer, J. Schäfer, and A.-L. Boulesteix, "Regularized estimation of large-scale gene association networks using graphical Gaussian models.," *BMC bioinformatics*, vol. 10, p. 384, nov 2009.

[10] S. Basu, W. Duren, C. R. Evans, C. F. Burant, G. Michailidis, and A. Karnovsky, "Sparse network modeling and Metscape-based visualization methods for the analysis of large-scale metabolomics data," *Bioinformatics*, vol. 33, p. btx012, jan 2017.

[11] J. Ma, A. Shojaie, and G. Michailidis, "Network-based pathway enrichment analysis with incomplete network information," *Bioinformatics*, vol. 32, pp. 3165–3174, oct 2016.

[12] G. J. Patti, O. Yanes, L. P. Shriver, J. P. Courade, R. Tautenhahn, M. Manchester, and G. Siuzdak, "Metabolomics implicates altered sphingolipids in chronic pain of neuropathic origin," *Nat Chem Biol*, vol. 8, no. 3, pp. 232–234, 2012.

[13] Q. Wang, X. Chen, M. Zhang, Q. Shen, and Z. Qin, "Identification of hub genes and pathways associated with retinoblastoma based on co-expression network analysis," *Genetics and Molecular Research*, vol. 14, pp. 16151–16161, dec 2015.

[14] S. B. Cogill and L. Wang, "Co-expression Network Analysis of Human lncRNAs and Cancer Genes," *Cancer Informatics*, vol. 13s5, p. CIN.S14070, jan 2014.

[15] D.-X. YIN, H.-M. ZHAO, D.-J. SUN, J. YAO, and D.-Y. DING, "Identification of candidate target genes for human peripheral arterial disease using weighted gene co-expression network analysis," *Molecular Medicine Reports*, vol. 12, pp. 8107–8112, dec 2015.

[16] A. Voigt, K. Nowick, and E. Almaas, "A composite network of conserved and tissue specific gene interactions reveals possible genetic interactions in glioma.," *PLoS computational biology*, vol. 13, p. e1005739, sep 2017.

[17] T. M. Creanza, M. Liguori, S. Liuni, N. Nuzziello, and N. Ancona, "Meta-Analysis of Differential Connectivity in Gene Co-Expression Networks in Multiple Sclerosis.," *International journal of molecular sciences*, vol. 17, jun 2016.

[18] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim, "Differential coexpression analysis using microarray data and its application to human cancer.," *Bioinformatics (Oxford, England)*, vol. 21, pp. 4348–55, dec 2005.

[19] J. Gillis and P. Pavlidis, "The role of indirect connections in gene networks in predicting function.," *Bioinformatics (Oxford, England)*, vol. 27, pp. 1860–6, jul 2011.

[20] H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis, "Coexpression analysis of human genes across many microarray data sets.," *Genome research*, vol. 14, pp. 1085–94, jun 2004.

[21] G. J. Patti, O. Yanes, and G. Siuzdak, "Metabolomics: the apogee of the omics trilogy," *Nature Reviews Molecular Cell Biology*, vol. 13, pp. 263–269, apr 2012.

[22] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: Current approaches and outstanding challenges," *PLoS Computational Biology*, vol. 8, no. 2, 2012.

[23] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 15545–50, oct 2005.

[24] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *The Annals of Applied Statistics*, vol. 1, pp. 107–129, jun 2007.

[25] J. Xia and D. S. Wishart, "MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data.," *Nucleic acids research*, vol. 38, pp. W71–7, jul 2010.

[26] M. Chagoyen and F. Pazos, "MBRole: enrichment analysis of metabolomic data," *Bioinformatics*, vol. 27, pp. 730–731, mar 2011.

[27] J. Xia and D. S. Wishart, "Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis," in *Current Protocols in Bioinformatics*, vol. 55, pp. 14.10.1–14.10.91, Hoboken, NJ, USA: John Wiley & Sons, Inc., sep 2016.

[28] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes.," *Nucleic acids research*, vol. 28, pp. 27–30, jan 2000.

[29] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill,

L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, may 2000.

[30] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, vol. 355. John Wiley & Sons, 2003.

[31] P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, A. Gonzalez-Perez, J. Pearson, C. Sander, and Others, "Pathway and network analysis of cancer genomes," *Nature methods*, vol. 12, no. 7, p. 615, 2015.

[32] L. D. Parnell, J. M. Ordovas, and C.-Q. Lai, "Environmental and epigenetic regulation of postprandial lipemia.," *Current opinion in lipidology*, vol. 29, pp. 30–35, feb 2018.

[33] C. B. Cole, M. Nikpay, and R. McPherson, "Gene-environment interaction in dyslipidemia.," *Current opinion in lipidology*, vol. 26, pp. 133–8, apr 2015.

[34] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types," *The annals of applied statistics*, vol. 7, no. 1, p. 523, 2013.

[35] K. Van Deun, T. F. Wilderjans, R. A. den Berg, A. Antoniadis, and I. Van Mechelen, "A flexible framework for sparse simultaneous component based data integration," *BMC bioinformatics*, vol. 12, no. 1, p. 448, 2011.

[36] Z. Yang and G. Michailidis, "A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data," *Bioinformatics*, vol. 32, no. 1, pp. 1–8, 2015.

[37] G. Zhou, Q. Zhao, Y. Zhang, T. Adal\i, S. Xie, and A. Cichocki, "Linked component analysis from matrices to high-order tensors: Applications to biomedical data," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 310–331, 2016.

[38] J. Guo, E. Levina, G. Michailidis, and J. Zhu, "Joint estimation of multiple graphical models," *Biometrika*, vol. 98, no. 1, pp. 1–15, 2011.

[39] C. Peterson, F. C. Stingo, and M. Vannucci, "Bayesian Inference of Multiple Gaussian Graphical Models," *Journal of the American Statistical Association*, vol. 110, pp. 159–174, apr 2015.

[40] J. Ma and G. Michailidis, "Joint structural estimation of multiple graphical models," *Journal of Machine Learning Research*, vol. 17, no. 166, pp. 1–48, 2016.

[41] T. Cai, W. Liu, and Y. Xia, "Two-Sample Covariance Matrix Testing and Support Recovery in High-Dimensional and Sparse Settings," *Journal of the American Statistical Association*, vol. 108, pp. 265–277, mar 2013.

[42] J. C. Gower and G. B. Dijksterhuis, *Procrustes problems*. Oxford; New York: Oxford University Press, 1 ed., 2004.

[43] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[44] H. Zou, T. Hastie, and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[45] R. A. R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*. Prentice Hall, 1992.

[46] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 74, no. 3, 2006.

[47] C. J. Fabian, "The what, why and how of aromatase inhibitors: hormonal agents for treatment and prevention of breast cancer.," *International journal of clinical practice*, vol. 61, pp. 2051–63, dec 2007.

[48] N. L. Henry, F. Azzouz, Z. Desta, L. Li, A. T. Nguyen, S. Lemler, J. Hayden, K. Tarpinian, E. Yakim, D. A. Flockhart, V. Stearns, D. F. Hayes, and A. M. Storniolo, "Predictors of aromatase inhibitor discontinuation as a result of treatment-emergent symptoms in early-stage breast cancer.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 30, pp. 936–42, mar 2012.

[49] D. L. Hershman, L. H. Kushi, T. Shao, D. Buono, A. Kershenbaum, W.-Y. Tsai, L. Fehrenbacher, S. L. Gomez, S. Miles, and A. I. Neugut, "Early discontinuation and nonadherence to adjuvant hormonal therapy in a cohort of 8,769 early-stage breast cancer patients.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 28, pp. 4120–8, sep 2010.

[50] R. Zárate, N. El Jaber-Vazdekis, N. Tejera, J. A. Pérez, and C. Rodríguez, "Significance of long chain polyunsaturated fatty acids in human health.," *Clinical and translational medicine*, vol. 6, p. 25, dec 2017.

[51] D. L. Hershman, J. M. Unger, K. D. Crew, D. Awad, S. R. Dakhil, J. Gralow, H. Greenlee, D. L. Lew, L. M. Minasian, C. Till, J. L. Wade, F. L. Meyskens, and C. M. Moinpour, "Randomized Multicenter Placebo-Controlled Trial of Omega-3 Fatty Acids for the Control of Aromatase Inhibitor-Induced Musculoskeletal

Pain: SWOG S0927.," *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 33, pp. 1910–7, jun 2015.

[52] E. Tortosa-Caparrós, D. Navas-Carrillo, F. Marín, and E. Orenes-Piñero, "Anti-inflammatory effects of omega 3 and omega 6 polyunsaturated fatty acids in cardiovascular disease and metabolic syndrome.," *Critical reviews in food science and nutrition*, vol. 57, pp. 3421–3429, nov 2017.

[53] E. Tortosa-Caparrós, D. Navas-Carrillo, F. Marín, and E. Orenes-Piñero, "Anti-inflammatory effects of omega 3 and omega 6 polyunsaturated fatty acids in cardiovascular disease and metabolic syndrome.," *Critical reviews in food science and nutrition*, vol. 57, pp. 3421–3429, nov 2017.

[54] A. Jakobsson, R. Westerberg, and A. Jacobsson, "Fatty acid elongases in mammals: their regulation and roles in metabolism.," *Progress in lipid research*, vol. 45, pp. 237–49, may 2006.

[55] G. Burdge, "Alpha-linolenic acid metabolism in men and women: nutritional and biological implications.," *Current opinion in clinical nutrition and metabolic care*, vol. 7, pp. 137–44, mar 2004.

[56] Plotly, "Visualize Data, Together," 2017.

[57] RStudio, "Shiny," 2017.

[58] Y. Minami, T. Kasukawa, Y. Kakazu, M. Iigo, M. Sugimoto, S. Ikeda, A. Yasui, G. T. J. van der Horst, T. Soga, and H. R. Ueda, "Measurement of internal body time by blood metabolomics.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 9890–5, jun 2009.

[59] J. E. Ippolito, J. Xu, S. Jain, K. Moulder, S. Mennerick, J. R. Crowley, R. R. Townsend, and J. I. Gordon, "An integrated functional genomics and

metabolomics approach for defining poor prognosis in human neuroendocrine cancers.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 9901–6, jul 2005.

[60] T. J. Wang, M. G. Larson, R. S. Vasan, S. Cheng, E. P. Rhee, E. McCabe, G. D. Lewis, C. S. Fox, P. F. Jacques, C. Fernandez, C. J. O'Donnell, S. A. Carr, V. K. Mootha, J. C. Florez, A. Souza, O. Melander, C. B. Clish, and R. E. Gerszten, "Metabolite profiles and the risk of developing diabetes," *Nature Medicine*, vol. 17, pp. 448–453, apr 2011.

[61] G. Wu and L. Stein, "A network module-based method for identifying cancer prognostic signatures.," *Genome biology*, vol. 13, p. R112, dec 2012.

[62] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis.," *Molecular systems biology*, vol. 3, p. 140, 2007.

[63] U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway, and D. Pe'er, "An integrated approach to uncover drivers of cancer.," *Cell*, vol. 143, pp. 1005–17, dec 2010.

[64] A. M. Sonabend, M. Bansal, P. Guarnieri, L. Lei, B. Amendolara, C. Soderquist, R. Leung, J. Yun, B. Kennedy, J. Sisti, S. Bruce, R. Bruce, R. Shakya, T. Ludwig, S. Rosenfeld, P. A. Sims, J. N. Bruce, A. Califano, and P. Canoll, "The transcriptional regulatory network of proneural glioma determines the genetic alterations selected during tumor progression.," *Cancer research*, vol. 74, pp. 1440–1451, mar 2014.

[65] S. C. Booth, A. M. Weljie, and R. J. Turner, "COMPUTATIONAL TOOLS FOR THE SECONDARY ANALYSIS OF METABOLOMICS EXPERI-

MENTS," *Computational and Structural Biotechnology Journal*, vol. 4, jan 2013.

[66] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.," *Nucleic acids research*, vol. 37, pp. 1–13, jan 2009.

[67] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. François, and R. Zecchina, "Finding undetected protein associations in cell signaling by belief propagation.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, pp. 882–7, jan 2011.

[68] O. A. Balbin, J. R. Prensner, A. Sahu, A. Yocum, S. Shankar, R. Malik, D. Fermin, S. M. Dhanasekaran, B. Chandler, D. Thomas, D. G. Beer, X. Cao, A. I. Nesvizhskii, and A. M. Chinnaiyan, "Reconstructing targetable pathways in lung cancer by integrating diverse omics data.," *Nature communications*, vol. 4, p. 2617, jan 2013.

[69] L. Pirhaji, P. Milani, M. Leidl, T. Curran, J. Avila-Pacheco, C. B. Clish, F. M. White, A. Saghatelian, and E. Fraenkel, "Revealing disease-associated pathways by network integration of untargeted metabolomics.," *Nature methods*, vol. 13, no. 9, pp. 770–6, 2016.

[70] G. Astarita, J. H. McKenzie, B. Wang, K. Strassburg, A. Doneanu, J. Johnson, A. Baker, T. Hankemeier, J. Murphy, R. J. Vreeken, J. Langridge, and J. X. Kang, "A Protective Lipidomic Biosignature Associated with a Balanced Omega-6/Omega-3 Ratio in fat-1 Transgenic Mice," *PLoS ONE*, vol. 9, p. e96221, apr 2014.

[71] H. Wang, M. V. Airola, and K. Reue, "How lipid droplets TAG along: Glycerolipid synthetic enzymes and lipid storage," *Biochimica et Biophysica Acta*

(BBA) - Molecular and Cell Biology of Lipids, vol. 1862, pp. 1131–1145, oct 2017.

[72] J. M. Ntambi, "Regulation of stearoyl-CoA desaturase by polyunsaturated fatty acids and cholesterol.," *Journal of lipid research*, vol. 40, pp. 1549–58, sep 1999.

[73] E. Toledo, D. D. Wang, M. Ruiz-Canela, C. B. Clish, C. Razquin, Y. Zheng, M. Guasch-Ferré, A. Hruby, D. Corella, E. Gómez-Gracia, M. Fiol, R. Estruch, E. Ros, J. Lapetra, M. Fito, F. Aros, L. Serra-Majem, L. Liang, J. Salas-Salvadó, F. B. Hu, and M. A. Martínez-González, "Plasma lipidomic profiles and cardiovascular events in a randomized intervention trial with the Mediterranean diet.," *The American journal of clinical nutrition*, vol. 106, pp. 973–983, oct 2017.

[74] I. Ljubić, R. Weiskircher, U. Pferschy, G. W. Klau, P. Mutzel, and M. Fischetti, "An Algorithmic Framework for the Exact Solution of the Prize-Collecting Steiner Tree Problem," *Mathematical Programming*, vol. 105, pp. 427–449, feb 2006.

[75] N. Tuncbag, S. J. C. Gosline, A. Kedaigle, A. R. Soltis, A. Gitter, and E. Fraenkel, "Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package," *PLOS Computational Biology*, vol. 12, p. e1004879, apr 2016.

[76] R. A. Fisher, "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika*, vol. 10, p. 507, may 1915.

[77] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Ex-

pression Microarray Data," *Machine Learning*, vol. 52, no. 1/2, pp. 91–118, 2003.

[78] J. Cao and S. Zhang, "A Bayesian extension of the hypergeometric test for functional enrichment analysis.," *Biometrics*, vol. 70, pp. 84–94, mar 2014.

[79] S. Park, K. C. Sadanala, and E.-K. Kim, "A Metabolomic Approach to Understanding the Metabolic Link between Obesity and Diabetes," *Molecules and Cells*, vol. 38, pp. 587–596, jul 2015.

[80] M. G. Buse, S. Jursinic, and S. S. Reid, "Regulation of branched-chain amino acid oxidation in isolated muscles, nerves and aortas of rats.," *The Biochemical journal*, vol. 148, pp. 363–74, jun 1975.

[81] J. A. Vazquez, E. L. Morse, and S. A. Adibi, "Effect of dietary fat, carbohydrate, and protein on branched-chain amino acid catabolism during caloric restriction.," *Journal of Clinical Investigation*, vol. 76, pp. 737–743, aug 1985.

[82] G. Cederblad, "Effect of diet on plasma carnitine levels and urinary carnitine excretion in humans," *The American Journal of Clinical Nutrition*, vol. 45, pp. 725–729, apr 1987.

[83] Y. Zhao and V. Natarajan, "Lysophosphatidic acid signaling in airway epithelium: Role in airway inflammation and remodeling," *Cellular Signalling*, vol. 21, pp. 367–377, mar 2009.

[84] K. Hayashi, M. Takahashi, W. Nishida, K. Yoshida, Y. Ohkawa, A. Kitabatake, J. Aoki, H. Arai, and K. Sobue, "Phenotypic modulation of vascular smooth muscle cells induced by unsaturated lysophosphatidic acids.," *Circulation research*, vol. 89, pp. 251–8, aug 2001.

[85] T. Esko, J. N. Hirschhorn, H. A. Feldman, Y.-H. H. Hsu, A. A. Deik, C. B. Clish, C. B. Ebbeling, and D. S. Ludwig, "Metabolomic profiles as reliable biomarkers of dietary composition," *The American Journal of Clinical Nutrition*, vol. 105, pp. 547–554, mar 2017.

[86] M. A. Lips, J. B. Van Klinken, V. van Harmelen, H. K. Dharuri, P. A. 't Hoen, J. F. Laros, G.-J. van Ommen, I. M. Janssen, B. Van Ramshorst, B. A. Van Wagensveld, D. J. Swank, F. Van Dielen, A. Dane, A. Harms, R. Vreeken, T. Hankemeier, J. W. Smit, H. Pijl, and K. Willems van Dijk, "Roux-en-Y Gastric Bypass Surgery, but Not Calorie Restriction, Reduces Plasma Branched-Chain Amino Acids in Obese Women Independent of Weight Loss or the Presence of Type 2 Diabetes," *Diabetes Care*, vol. 37, pp. 3150–3156, dec 2014.

[87] M. G. Schooneman, A. Napolitano, S. M. Houten, G. K. Ambler, P. R. Murgatroyd, S. R. Miller, C. E. Hollak, C. Y. Tan, S. Virtue, A. Vidal-Puig, D. J. Nunez, and M. R. Soeters, "Assessment of plasma acylcarnitines before and after weight loss in obese subjects," *Archives of Biochemistry and Biophysics*, vol. 606, pp. 73–80, sep 2016.

[88] D. J. Rigotti, M. Inglese, J. S. Babb, M. Rovaris, B. Benedetti, M. Filippi, R. I. Grossman, and O. Gonen, "Serial whole-brain N-acetylaspartate concentration in healthy young adults.," *AJNR. American journal of neuroradiology*, vol. 28, pp. 1650–1, oct 2007.

[89] J. Baeza, M. J. Smallegan, and J. M. Denu, "Mechanisms and Dynamics of Protein Acetylation in Mitochondria," *Trends in Biochemical Sciences*, vol. 41, pp. 231–244, mar 2016.

[90] M. Favennec, B. Hennart, R. Caiazzo, A. Leloire, L. Yengo, M. Verbanck, A. Arredouani, M. Marre, M. Pigeyre, A. Bessede, G. J. Guillemin, G. Chinetti,

B. Staels, F. Pattou, B. Balkau, D. Allorge, P. Froguel, and O. Poulain-Godefroy, "The kynurenine pathway is activated in human obesity and shifted toward kynurenine monooxygenase activation," *Obesity*, vol. 23, pp. 2066–2074, oct 2015.

[91] B. Strasser, K. Berger, and D. Fuchs, "Effects of a caloric restriction weight loss diet on tryptophan metabolism and inflammatory biomarkers in overweight adults," *European Journal of Nutrition*, vol. 54, pp. 101–107, feb 2015.

[92] J. E. Ho, M. G. Larson, A. Ghorbani, S. Cheng, M.-H. Chen, M. Keyes, E. P. Rhee, C. B. Clish, R. S. Vasan, R. E. Gerszten, and T. J. Wang, "Metabolomic Profiles of Body Mass Index in the Framingham Heart Study Reveal Distinct Cardiometabolic Phenotypes," *PLOS ONE*, vol. 11, p. e0148361, feb 2016.

[93] K. Overmyer, C. Evans, N. Qi, C. Minogue, J. Carson, C. Chermside-Scabbo, L. Koch, S. Britton, D. Pagliarini, J. Coon, and C. Burant, "Maximal Oxidative Capacity during Exercise Is Associated with Skeletal Muscle Fuel Selection and Dynamic Changes in Mitochondrial Protein Acetylation," *Cell Metabolism*, vol. 21, pp. 468–478, mar 2015.

[94] L. G. Koch, O. J. Kemi, N. Qi, S. X. Leng, P. Bijma, L. J. Gilligan, J. E. Wilkinson, H. Wisloff, M. A. Hoydal, N. Rolim, P. M. Abadir, E. M. van Grevenhof, G. L. Smith, C. F. Burant, O. Ellingsen, S. L. Britton, and U. Wisloff, "Intrinsic Aerobic Capacity Sets a Divide for Aging and Longevity," *Circulation Research*, vol. 109, pp. 1162–1172, oct 2011.

[95] S. N. Blair, J. B. Kampert, H. W. Kohl, C. E. Barlow, C. A. Macera, R. S. Paffenbarger, and L. W. Gibbons, "Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women.," *JAMA*, vol. 276, pp. 205–10, jul 1996.

[96] D. W. Foster, "Malonyl-CoA: the regulator of fatty acid synthesis and oxidation.," *The Journal of clinical investigation*, vol. 122, pp. 1958–9, jun 2012.

[97] A. M. Giudetti, E. Stanca, L. Siculella, G. V. Gnoni, and F. Damiano, "Nutritional and Hormonal Regulation of Citrate and Carnitine/Acylcarnitine Transporters: Two Mitochondrial Carriers Involved in Fatty Acid Metabolism.," *International journal of molecular sciences*, vol. 17, may 2016.

[98] F. Pietrocola, L. Galluzzi, J. M. Bravo-San Pedro, F. Madeo, and G. Kroemer, "Acetyl Coenzyme A: A Central Metabolite and Second Messenger," *Cell Metabolism*, vol. 21, pp. 805–821, jun 2015.

[99] M. H. E. Christensen, D. J. Fadnes, T. H. Røst, E. R. Pedersen, J. R. Andersen, V. Våge, A. Ulvik, Ø. Midttun, P. M. Ueland, O. K. Nygård, and G. Mellgren, "Inflammatory markers, the tryptophan-kynurenine pathway, and vitamin B status after bariatric surgery," *PLOS ONE*, vol. 13, p. e0192169, feb 2018.

[100] E. S. Goetzman, Z. Gong, M. Schiff, Y. Wang, and R. H. Muzumdar, "Metabolic pathways at the crossroads of diabetes and inborn errors," *Journal of Inherited Metabolic Disease*, vol. 41, pp. 5–17, jan 2018.

[101] J. E. Galgani, C. Moro, and E. Ravussin, "Metabolic flexibility and insulin resistance.," *American journal of physiology. Endocrinology and metabolism*, vol. 295, pp. E1009–17, nov 2008.

[102] S. Schenk and J. F. Horowitz, "Acute exercise increases triglyceride synthesis in skeletal muscle and prevents fatty acid-induced insulin resistance.," *The Journal of clinical investigation*, vol. 117, pp. 1690–8, jun 2007.

[103] T. Kind, K. H. Liu, D. Y. Lee, B. DeFelice, J. K. Meissen, and O. Fiehn, "LipidBlast in silico tandem mass spectrometry database for lipid identification," *Nat Methods*, vol. 10, no. 8, pp. 755–758, 2013.

[104] L. V. Hedges, I. Olkin, L. V. Hedges, and I. Olkin, "Combining Estimates of Correlation Coefficients," in *Statistical Methods for Meta-Analysis*, pp. 223–246, Elsevier, 1985.

[105] E. S. Ford and W. H. Dietz, "Modeling dietary patterns to assess sodium recommendations for nutrient adequacy," *The American Journal of Clinical Nutrition*, vol. 97, pp. 848–853, apr 2013.

[106] N. C. f. H. S. (US), *Health, United States, 2016*. National Center for Health Statistics (US), 2017.

[107] R. A. Fisher, "On the probable error of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, pp. 3–32, 1921.

[108] A. S. Cornford, A. L. Barkan, and J. F. Horowitz, "Rapid suppression of growth hormone concentration by overeating: potential mediation by hyperinsulinemia.," *The Journal of clinical endocrinology and metabolism*, vol. 96, pp. 824–30, mar 2011.

[109] S. L. Hummel, E. M. Seymour, R. D. Brook, T. J. Kolias, S. S. Sheth, H. R. Rosenblum, J. M. Wells, and A. B. Weder, "Low-sodium dietary approaches to stop hypertension diet reduces blood pressure, arterial stiffness, and oxidative stress in hypertensive heart failure with preserved ejection fraction.," *Hypertension (Dallas, Tex. : 1979)*, vol. 60, pp. 1200–6, nov 2012.

[110] A. V. Mathew, E. M. Seymour, J. Byun, S. Pennathur, and S. L. Hummel, "Altered Metabolic Profile With Sodium-Restricted Dietary Approaches to Stop Hypertension Diet in Hypertensive Heart Failure With Preserved Ejection Fraction.," *Journal of cardiac failure*, vol. 21, pp. 963–7, dec 2015.

[111] E. G. Bligh and W. J. Dyer, "A RAPID METHOD OF TOTAL LIPID EX-

TRACTION AND PURIFICATION," *Canadian Journal of Biochemistry and Physiology*, vol. 37, pp. 911–917, aug 1959.

[112] F. Afshinnia, T. M. Rajendiran, A. Karnovsky, T. Soni, X. Wang, D. Xie, W. Yang, T. Shafi, M. R. Weir, J. He, C. S. Brecklin, E. P. Rhee, J. R. Schelling, A. Ojo, H. Feldman, G. Michailidis, S. Pennathur, L. J. Appel, A. S. Go, J. Kusek, J. P. Lash, and R. R. Townsend, "Lipidomic Signature of Progression of Chronic Kidney Disease in the Chronic Renal Insufficiency Cohort.," *Kidney international reports*, vol. 1, pp. 256–268, nov 2016.

[113] T. Kind, J. K. Meissen, D. Yang, F. Nocito, A. Vaniya, Y.-S. Cheng, J. S. VanderGheynst, and O. Fiehn, "Qualitative analysis of algal secretions with multiple mass spectrometric platforms," *Journal of Chromatography A*, vol. 1244, pp. 139–147, jun 2012.

[114] J. K. Meissen, B. T. K. Yuen, T. Kind, J. W. Riggs, D. K. Barupal, P. S. Knoepfler, and O. Fiehn, "Induced pluripotent stem cells show metabolomic differences to embryonic stem cells in polyunsaturated phosphatidylcholines and primary metabolism.," *PloS one*, vol. 7, p. e46770, oct 2012.

[115] C. S. Ejsing, E. Duchoslav, J. Sampaio, K. Simons, R. Bonner, C. Thiele, K. Ekroos, and A. Shevchenko, "Automated identification and quantification of glycerophospholipid molecular species by multiple precursor ion scanning.," *Analytical chemistry*, vol. 78, pp. 6202–14, sep 2006.