

University of Louisville

ThinkIR: The University of Louisville's Institutional Repository


Electronic Theses and Dissertations

8-2018

Bayesian analytical approaches for metabolomics : a novel method for molecular structure-informed metabolite interaction modeling, a novel diagnostic model for differentiating myocardial infarction type, and approaches for compound identification given mass spectrometry data.

Patrick J. Trainor
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

 Part of the [Analytical Chemistry Commons](#), [Applied Statistics Commons](#), [Biochemical Phenomena, Metabolism, and Nutrition Commons](#), [Biochemistry Commons](#), [Bioinformatics Commons](#), [Biostatistics Commons](#), [Cardiology Commons](#), [Cardiovascular Diseases Commons](#), [Computational Biology Commons](#), [Diagnosis Commons](#), [Medical Biomathematics and Biometrics Commons](#), [Medical Molecular Biology Commons](#), [Molecular Biology Commons](#), [Programming Languages and Compilers Commons](#), [Statistical Methodology Commons](#), [Systems Biology Commons](#), and the [Theory and Algorithms Commons](#)

Recommended Citation

Trainor, Patrick J., "Bayesian analytical approaches for metabolomics : a novel method for molecular structure-informed metabolite interaction modeling, a novel diagnostic model for differentiating myocardial infarction type, and approaches for compound identification given mass spectrometry data." (2018). *Electronic Theses and Dissertations*. Paper 3023.
<https://doi.org/10.18297/etd/3023>

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

BAYESIAN ANALYTICAL APPROACHES FOR METABOLOMICS: A
NOVEL METHOD FOR MOLECULAR STRUCTURE-INFORMED
METABOLITE INTERACTION MODELING, A NOVEL DIAGNOSTIC
MODEL FOR DIFFERENTIATING MYOCARDIAL INFARCTION TYPE,
AND APPROACHES FOR COMPOUND IDENTIFICATION GIVEN
MASS SPECTROMETRY DATA

By
Patrick J. Trainor
M.S. in Biostatistics, 2014
M.A. in Mathematics, 2014
B.S. in Mathematics, 2012

A Dissertation
Submitted to the Faculty of the
School of Interdisciplinary and Graduate Studies of the University of
Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Interdisciplinary Studies: Bioinformatics

School of Interdisciplinary and Graduate Studies
University of Louisville
Louisville, Kentucky

August 2018

Copyright 2018 by Patrick J. Trainor

All rights reserved

BAYESIAN ANALYTICAL APPROACHES FOR METABOLOMICS: A
NOVEL METHOD FOR MOLECULAR STRUCTURE-INFORMED
METABOLITE INTERACTION MODELING, A NOVEL DIAGNOSTIC
MODEL FOR DIFFERENTIATING MYOCARDIAL INFARCTION TYPE,
AND APPROACHES FOR COMPOUND IDENTIFICATION GIVEN
MASS SPECTROMETRY DATA

By

Patrick J. Trainor
M.S. in Biostatistics, 2014
M.A. in Mathematics, 2014
B.S. in Mathematics, 2012

Dissertation approved on

July 19, 2018

by the following dissertation Committee:

Shesh N. Rai, Ph.D.

Andrew P. DeFilippis, M.D., M.Sc.

Eric C. Rouchka, D.Sc.

Juw W. Park, Ph.D.

DEDICATION

This dissertation is dedicated to my family: Colleen, Jack, Katie, Winn, Aaron, Matt, Abby, Cassie, Sam, Betty, Jim, Jane, and Edmond; to the ESL students at the BLC; and to the 3rd-5th grade students at the BLC.

ACKNOWLEDGMENTS

Acknowledgments are due to the past and present members of the Atherosclerosis / Atherothrombosis Research Laboratory (AARL), especially: Dr. Andrew DeFilippis, Dr. Alok Amraotkar, Dr. Mahrokh Nokhbehzaeim, Amanda Coulter, Allison Smith, Sharon Vincent, Ayesha Singh, Mallory Hatfield, and Dr. Yong Siow. In addition to the laboratory, special thanks are due to the human participants who have graciously agreed to participate in the clinical studies conducted by the AARL to further research into the leading cause of global mortality: heart disease.

Acknowledgments are due to collaborators who have assisted in this work including: Samantha Carlisle, Dr. Hunter Moseley, and Joshua Mitchell.

Acknowledgments are due to Dr. Shesh Rai, Dr. Aruni Bhatnagar, and Dr. Andrew DeFilippis for providing mentorship and support.

Finally, acknowledgments are due to the following organizations for providing financial support: the American Heart Association (11CRP7300003), the National Institutes of Health (1P20 GM103492-06 & Common Fund metabolomics pilot and feasibility award), and the Alpha Phi Foundation (2016 Heart-to-Heart Grant).

ABSTRACT

BAYESIAN ANALYTICAL APPROACHES FOR METABOLOMICS: A
NOVEL METHOD FOR MOLECULAR STRUCTURE-INFORMED
METABOLITE INTERACTION MODELING, A NOVEL DIAGNOSTIC
MODEL FOR DIFFERENTIATING MYOCARDIAL INFARCTION TYPE,
AND APPROACHES FOR COMPOUND IDENTIFICATION GIVEN
MASS SPECTROMETRY DATA

Patrick J. Trainor

July 19, 2018

Metabolomics, the study of small molecules in biological systems, has enjoyed great success in enabling researchers to examine disease-associated metabolic dysregulation and has been utilized for the discovery biomarkers of disease and phenotypic states. In spite of recent technological advances in the analytical platforms utilized in metabolomics and the proliferation of tools for the analysis of metabolomics data, significant challenges in metabolomics data analyses remain. In this dissertation, we present three of these challenges and Bayesian methodological solutions for each. In the first part we develop a new methodology to serve a basis for making higher order inferences in metabolomics, which we define as the testing of hypotheses that are more complex than single metabolite hypothesis tests. This methodology utilizes informative priors that are generated via the analysis of molecular structure similarity to enable the estimation of metabolite “interactomes” (or probabilistic models) which are organism-, sample media-, and condition-specific as well as comprehensive; and that can serve as reference models for studying perturbations in metabolic

systems. After discussing the development of our methodology, we present an evaluation of its performance conducted using simulation studies, and we use the methodology for estimating a plasma metabolite interactome for stable heart disease. This interactome may serve as a reference model for evaluating systems-level changes that occur with acute disease events such as myocardial infarction (MI) or unstable angina. In the second part of this work, we present the challenge of developing diagnostic classification models which utilize metabolite abundances and that do not “overfit” relatively small sample sizes, especially given the high dimensionality of metabolite data acquired using platforms such as liquid chromatography-mass spectrometry. We use a Bayesian methodology for estimating a multinomial logistic regression classifier for the detection and discrimination of the subtype of acute myocardial infarction utilizing metabolite abundance data quantified from blood plasma. As heart disease is the leading cause of global mortality, a blood-based and non-invasive diagnostic test that could differentiate between MI types at the time of the event would have great utility. In the final part of this dissertation we review Bayesian approaches for compound identification in metabolomics experiments that utilize liquid chromatography-mass spectrometry which remains a challenging problem.

TABLE OF CONTENTS

Dedication	iii
Acknowledgments	iv
Abstract	v
List of Tables	x
List of Figures	xi
INTRODUCTION	1
Challenges in metabolomic data analyses	1
Bayesian methodology for metabolomics	4
Layout of the dissertation	4
PRELIMINARIES: METABOLOMICS	6
Metabolism	6
Complexity of a single metabolic process	7
Dysregulation of metabolism and disease	9
From metabolism to metabolomics	13
Mass spectrometry-based metabolomics	16
PRELIMINARIES: BAYESIAN STATISTICS	19
Bayes rule	20
Bayesian inference regarding a parameter	20
Example Bayesian inference regarding a parameter	21
Specification of prior distribution	24
MCMC methods for inference	25
PRELIMINARIES: GAUSSIAN GRAPHICAL MODELING	30
Introduction	30
Gaussian graphical models	30
Bayesian inference given G -Wishart priors	37
HIGHER-ORDER INFERENCE IN METABOLOMICS	38
Statistical tests for pathway enrichment	39
Exact tests of pathway enrichment	39
Metabolite set enrichment analysis	41
Chemical Similarity Enrichment Analysis approach	42
The need for interactomes in metabolomics	43

A specific biomedical science use case	44
The challenge of inferring high-dimensional interactomes	45
INFERRING METABOLITE INTERACTOMES VIA BAYESIAN GRAPHICAL MODEL SELECTION UTILIZING MOLECULAR STRUCTURE INFORMATIVE PRIORS	
Desired inference and sketch of evaluation	47
Molecular structure priors and the BGL	48
R package development	51
Efficacy analysis via simulation studies	52
Results of the simulation studies	54
A PLASMA INTERACTOME FOR STABLE HEART DISEASE	
Cohort	59
Plasma metabolomics	59
Generation of chemical structure informed priors	60
Gibbs sampling	63
Discussion	66
A BAYESIAN DIAGNOSTIC MODEL FOR DIFFERENTIATING MI TYPE	
Acute Myocardial Infarction	71
Clinical cohort and samples	72
Feature selection	72
Multinomial logistic regression model	75
MCMC sampling of the posterior distribution	75
Results	77
Discussion	87
BAYESIAN APPROACHES FOR COMPOUND IDENTIFICATION GIVEN LC-MS DATA	
The challenge of identifying compounds from mass spectrometry data	90
Utilizing feasible metabolic reactions to generate conditional distributions	92
Utilizing isotope patterns in conditional distributions	94
“ProbMetab”: Formalization and integration of pathway databases	94
Infinite Gaussian mixture models	96
Bayesian clustering of mass features	97
CONCLUSION, DISCUSSION, AND FUTURE DIRECTIONS	
Summary	99
Future directions: Bayesian interactome models	101
Future directions: Diagnostic modeling	101
REFERENCES	104
APPENDIX A: ACRONYMS UTILIZED	126
APPENDIX B: SOFTWARE DEVELOPED FOR THE BAYESIAN GRAPHICAL LASSO	128

High-level R code	128
Preamble and internal C++ functions	134
Regular Bayesian Graphical Lasso C++ code	135
Adaptive Bayesian Graphical Lasso C++ code	137
CURRICULUM VITAE	142

LIST OF TABLES

1	Example 2×2 table for two random variables X and Y	39
2	Example Fisher’s exact test for determining if a pathway is over-represented in a list of significant metabolites	40
3	Table of results from the AR(1) simulation studies	57
4	Bayesian multinomial logistic regression model goodness of fit showing the model, the Widely Applicable Information Criteria (WAIC), and the Standard Error (SE)	77
5	Leave-one-out Cross-validation (LOOCV) estimated confusion matrix for the multinomial logistic regression model without clinical troponin values fit by Maximum Likelihood Estimation	86
6	Leave-one-out Cross-validation (LOOCV) estimated confusion matrix for Bayesian multinomial logistic regression model without clinical troponin values	86
7	Leave-one-out Cross-validation (LOOCV) estimated confusion matrix for Bayesian multinomial logistic regression model with clinical troponin values	86
8	Leave-one-out cross-validation (LOOCV) estimated group probabilities from the Bayesian multinomial logistic regression models for each of human subject in the cohort.	87

LIST OF FIGURES

1	Diagram of the tricarboxylic acid (TCA) cycle in <i>Homo sapiens</i> . . .	7
2	Diagram of the cholesterol biosynthesis pathway in <i>Homo sapiens</i> . . .	8
3	Example Bayesian estimation of a Poisson rate parameter	23
4	Graphical representation of the source of bias in discrete pathway enrichment analyses	41
5	Theoretical relationship between the structural similarity of two metabolites and the expected value of the shrinkage parameter λ_{ij} .	50
6	Graphical models from a randomly selected AR(1) simulation study	55
7	Histogram showing the results of the AR(1) simulation studies . . .	58
8	Heatmap showing the molecular structure similarity between the metabolites that were detected and quantified from the analysis of blood plasma from thrombotic MI, non-thrombotic MI, and stable CAD subjects	62
9	Time series plots for the MCMC sampler for the shrinkage param- eter λ_{ij} and for the concentration matrix entry ω_{ij} for the following metabolite pairs: (cholate, tyrosine), (cholate, cortisone), (cholate, glycochenodeoxycholate)	63
10	Graphical representations of the plasma metabolite interactome es- timated by the chemical structure adaptive Bayesian Graphical Lasso (BGL) for stable heart disease with subfigures showing the global model as well as the model focused on cholate and its neighbors	65
11	Correlation plot showing the metabolites that were considered in Bayesian multinomial logistic regression models	74
12	Time series plot of the MCMC chain for 3-hydroxypyridine sulfate multinomial logistic regression coefficient and resulting histogram showing the posterior distribution	78
13	Bayesian credible intervals for multinomial logit model parameters without troponin	79
14	Bayesian credible intervals for multinomial logit model parameters with troponin	80
15	3-dimensional scatterplot of MCMC samples for both linoleoylglyc- erols (18:2) with the acyl side chain in different positions and the log-posterior probability of each sampled model	82

16	Scatterplot of MCMC samples for both linoleoylglycerols (18:2) with the acyl side chain in different positions colored by log-posterior probability	83
17	A segment of the first MCMC chain for simulating the posterior distribution of group membership probability estimates for a specific human subject.	84
18	Histograms showing the simulated posterior distribution of group membership probability estimates for a specific human subject . . .	84

CHAPTER I

INTRODUCTION

1 Challenges in metabolomic data analyses

Metabolomics, or the study of small molecules in biological systems, has enjoyed great success in enabling researchers to make inferences regarding disease associated metabolic dysregulation as well as in furnishing biomarkers of disease and phenotypic states [Dunn and Hankemeier, 2013, Carlisle et al., 2016, Johnson et al., 2016, Newgard, 2017]. In spite of recent technological advances in the analytical platforms utilized in metabolomics [Gowda and Djukovic, 2014, Nagana Gowda and Raftery, 2016], advances in strategies for studying disease associated changes in metabolism [Bruntz et al., 2017, Krycer et al., 2017], and advances in the democratization (such as by web-based tools) of metabolomics data analyses [Warth et al., 2017, Guijas et al., 2018, Chong et al., 2018], significant challenges in metabolomics data analyses remain. The first challenge that we confront in the current work is making “higher order inference” in metabolomics (Chapters V,VI, and VII). Prior to arriving at our discussion of higher order inference in metabolomics, we introduce metabolism and metabolomics more generally in Chapter II. We define higher order inference as the testing of hypotheses that are more complex than single metabolite hypothesis tests. In studying metabolism, research questions are often over systems of metabolites or metabolic processes. For example, a biomedical scientist might wish to know how a statin impacts cholesterol metabolism. Likewise a cancer researcher may wish to know how the knockdown or knockout of a specific gene alters cellular metabolism. Formulating such research questions as hypothesis tests is not straightforward. The first step

in developing such statistical tests is developing a reference model for the system of interest. In the example of a biomedical scientist studying a statin, a reference model of how the metabolites of cholesterol metabolism interact (probabilistically) is required in order to determine how the system has changed following the treatment of a human (or model organism) with a statin. As we argue in Chapter V, the current paradigm of pathway-based or *a priori* knowledge-based enrichment analyses may suffer from extreme bias, especially in the case of analyzing biofluids such as blood plasma, serum, or urine. Unbiased, or data-dependent approaches such as the construction of correlation networks also suffer from significant methodological issues, as we discuss (Chapter V). Consequently, we propose a new approach that allows for the construction of metabolite “interactomes” or probabilistic models which are organism and sample media specific that can be used as reference models for studying perturbations in metabolic systems. Given the high-dimensionality of metabolite abundance data from untargeted metabolomics experiments, we propose utilizing informative priors that are generated via the analysis of molecular structure similarity in the estimation of such models. After discussing the development of our methodology, and presenting an evaluation of its performance through simulation studies, we use the methodology to build a plasma metabolite interactome for stable heart disease. This interactome may serve as a reference model for evaluating systems-level changes that occur with acute disease events such as myocardial infarction (MI) or unstable angina.

In the second part of this work, we confront the challenge of developing diagnostic classification models utilizing metabolite abundances that do not “overfit” relatively small sample sizes. Often poor generalization error results from unrestricted optimization of likelihoods in models with a high ratio of model parameters to sample size of a training dataset [Hastie et al., 2009]. This is often referred to as the $p \gg n$ problem (where, in this context, p is the number of metabolites and n is the number of samples). Many metabolomics datasets are characterized by a large p (e.g. $> 30,000$ mass features can be observed in high resolution mass

spectrometry experiments), while the substantial costs associated with the acquisition of metabolomics data implies that small datasets are a persistent feature of the field. As more resources are devoted to personalized medicine [Hamburg and Collins, 2010, Wishart, 2016] and deep phenotyping [Delude, 2015], diagnostic classifiers that are robust given $p \gg n$ and minimize the likelihood of overfitting are essential.

In Chapter VIII we discuss such a classification methodology, applied to a critical clinical problem: the detection and discrimination of the subtype of acute myocardial infarction (colloquially: heart attack) utilizing metabolite abundance data quantified from blood plasma. Heart disease is the leading cause of global mortality [Benjamin et al., 2017]. Myocardial Infarction (MI), an acute manifestation of heart disease, is characterized by heterogeneous etiology [Thygesen et al., 2012]. Consequently, a blood-based and non-invasive diagnostic test that could be administered upon presentation to an emergency department and could differentiate between thrombotic MI and non-thrombotic MI would be of great utility. Our laboratory (the Atherosclerosis / Atherothrombosis Research Laboratory [AARL] at the University of Louisville) has recruited two clinical cohorts that are designed for the development and validation of a diagnostic test capable of differentiating thrombotic MI, non-thrombotic MI, and stable coronary artery disease; the first cohort has been described previously [DeFilippis et al., 2015, DeFilippis et al., 2017, Trainor et al., 2017]. As a critical step towards the development of such a diagnostic, in Chapter VIII we present multinomial logistic regression models utilizing the metabolite abundances for the metabolites that were selected as part of a separate feature selection work [Trainor et al., 2018]. We compare the performance of the Bayesian approach we have employed for estimating model coefficients with maximum likelihood estimation and discuss the merits of the Bayesian approach.

In the final part of this dissertation, we discuss one of the most challenging aspects of mass spectrometry-based metabolomics [Dunn et al., 2012], that is, compound identification. In this part, we review Bayesian approaches for compound

identification in metabolomics experiments that utilize liquid chromatography-mass spectrometry (LC-MS).

2 Bayesian methodology for metabolomics

While this dissertation presents three different analytical problems that practitioners in the field of metabolomics currently face, the common aspect of the solutions that we describe is Bayesian reasoning. Bayesian methodologies for the three types of data analysis problems discussed in the current work are seldom utilized and enjoy scant popularity in the field of metabolomics. As we hope to demonstrate in the current work, Bayesian approaches allow practitioners to incorporate prior scientific knowledge into the data analysis process. In the first part, (the generation of metabolite interactomes) the prior belief that is postulated is that the enzyme catalyzed biochemical reactions that convert one intermediate into another generate both dependence in molecular structure similarity and statistical dependence in the distribution of metabolite abundances. In the second part (generation of a statistical classifier for discriminating MI type), the prior belief that is postulated is that the magnitude of regression coefficients should be small given the relatively small sample size (and high ratio of p to n). In the final part (Bayesian methods for LC-MS compound identification), various prior beliefs are incorporated such as that the presence of one compound in a dataset implies that the probability closely related (by chemical modifications) compounds is more likely.

3 Layout of the dissertation

The layout of the dissertation is as follows: Chapter II provides preliminary theory regarding metabolism and metabolomics, Chapter III provides a cursory introduction to Bayesian statistics, Chapter IV provides an introduction to the class of model we will utilize for generating metabolite interactomes, Chapter V presents the current paradigms for making higher order inferences in metabolomics and sets the stage for Chapter VI in which we present our novel methodology for using

molecular structure similarity to generate informative priors for the estimation of metabolite interactomes. Chapter VII presents such an interactome for stable heart disease. We then transition to the development of a Bayesian diagnostic model for the detection and differentiation of MI type in Chapter VIII. In Chapter IX we transition again, and provide a review of Bayesian approaches utilized to identify compounds in metabolomics experiments using LC-MS. Finally, we conclude this dissertation in Chapter X by presenting conclusions we have drawn from the complete work and a discussion of the methodologies herein, as well as future directions that we are pursuing as a result of this work.

CHAPTER II

PRELIMINARIES: METABOLOMICS

1 Metabolism

Metabolism is broadly defined as the chemical reactions that enable life [Voet et al., 2013]. The metabolic reactions that sustain life can be categorized as catabolic reactions, or the reactions that entail breaking down and oxidizing molecules to provide energy or substrates for other reactions, and anabolic reactions in which complex molecules are synthesized. Series of coupled reactions are referred to as pathways. Anabolic pathways are characterized by endergonic processes as the synthesis of macromolecules from smaller molecules requires energy. Conversely catabolic pathways entail the break down of larger molecules into smaller molecules for use either as inputs in anabolic reactions or for the release of free energy via oxidation and reduction reactions.

An example catabolic process is the tricarboxylic acid (TCA) cycle which is illustrated in Fig. 1. In the TCA cycle an input molecule of Acetyl-CoA (from the catabolism of glucose, fats, or proteins) enters into a sequence of reactions which produces three molecules of NADH, one molecule of FADH₂, and one molecule of GTP. The molecules of NADH and FADH₂ produced during TCA cycle also serve as substrates for the electron transport chain which generates additional molecules of ATP. An example anabolic pathway is the cholesterol biosynthesis pathway shown in Fig. 1. In this pathway, acetyl-CoA molecules as a substrate are chemically transformed through a sequence of reactions with intermediates including HMG-CoA, mevalonate, isopentenyl phosphate, squalene, lanosterol, and

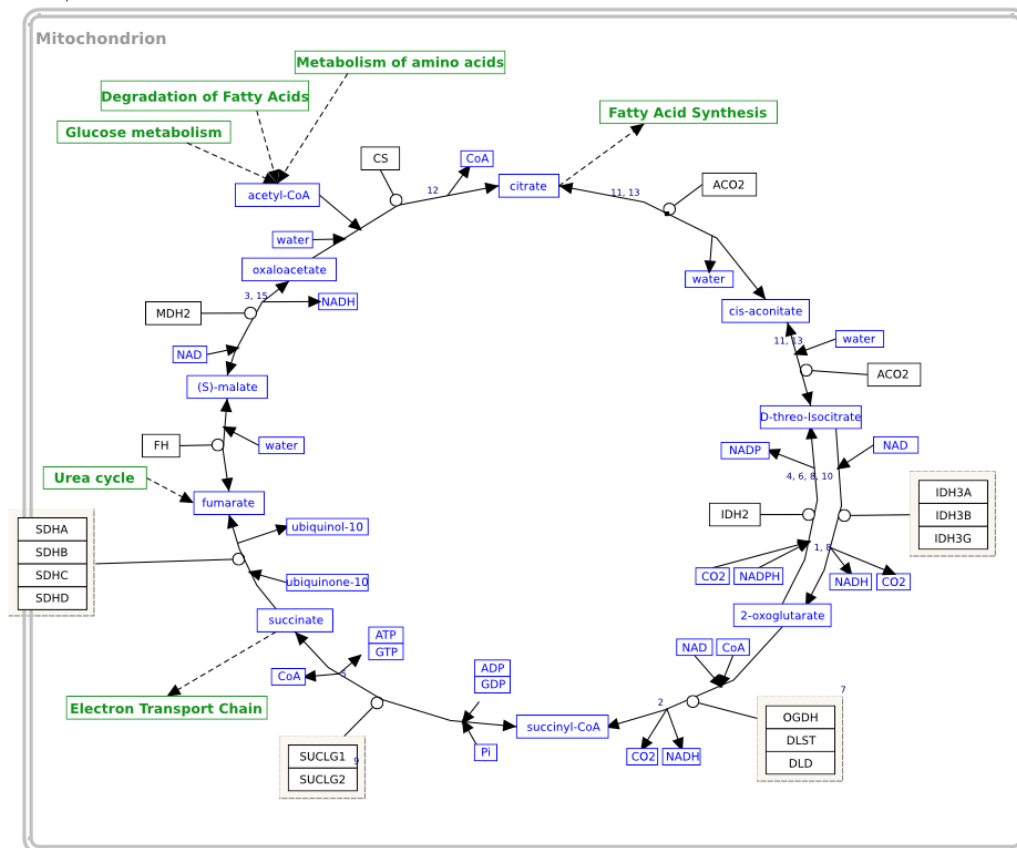


Figure 1. Diagram of the tricarboxylic acid (TCA) cycle in *Homo sapiens*.
 Wikipathways: WP78 Revision 90661.

finally cholesterol. The cholesterol biosynthesis pathway entails over 20 steps, and requires ATP as an energetic input and the cofactors NADH and NADPH.

2 Complexity of a single metabolic process

While many of the biochemical processes of metabolism, including glucose catabolism, glycogen metabolism, gluconeogenesis, lipid metabolism, TCA cycle, electron transport chain, nucleotide metabolism, and protein synthesis take place in the cell, many of the signaling processes that control metabolism as well as many systems of transport are extracellular [Voet et al., 2013]. As an example, consider lipid metabolism. While the catabolic process of breaking down fatty acids to yield energetic substrates, β -oxidation, takes place in the mitochondrial matrix of a cell, many of the processes prior to the point of β -oxidation are extracellular or take

Title: Cholesterol Biosynthesis
Last modified: 2/22/2013
Organism: Homo sapiens

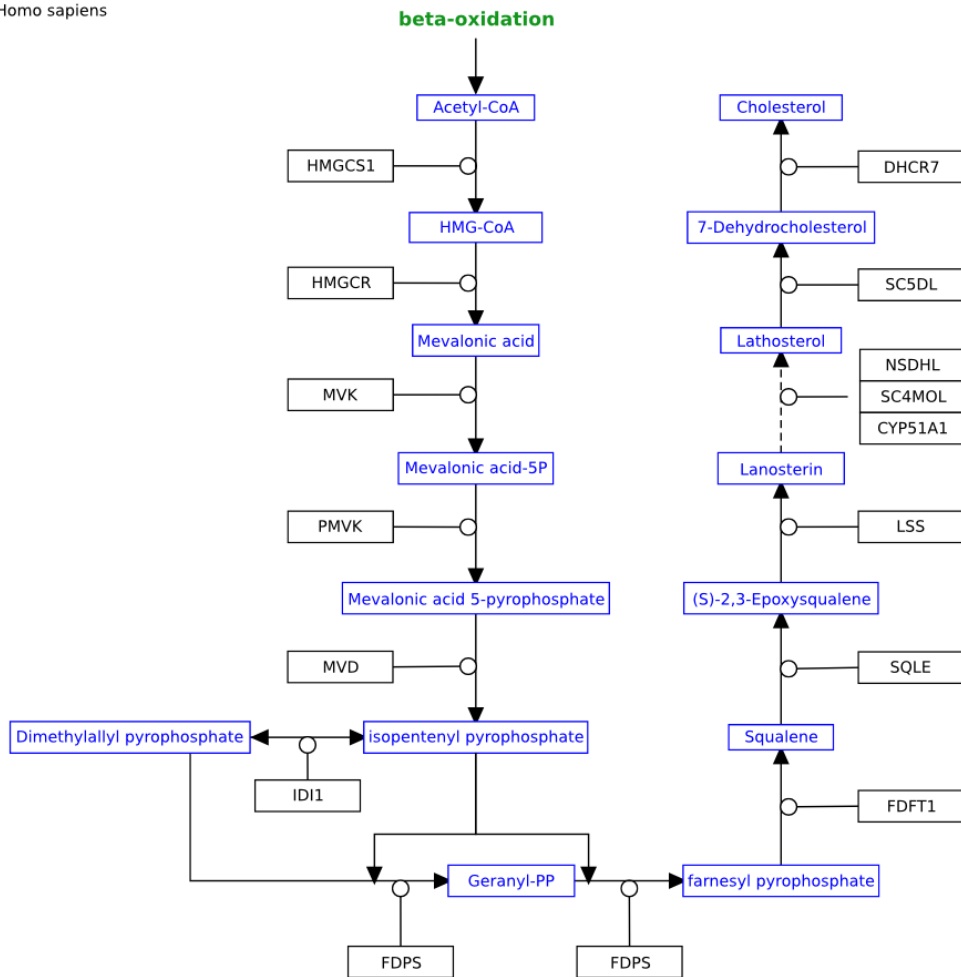


Figure 2. Diagram of the cholesterol biosynthesis pathway in *Homo sapiens*.

Wikipathways: WP197 Revision 91292

place in distinct cells from the cell in which a specific fatty acid molecule is being broken down. Within adipocytes (specialized cells for the storage of fatty acids), fatty acids are stored as triacylglycerols. When activated by hormone sensitive lipase, triacylglycerols are hydrolyzed yielding free fatty acids which are released into the bloodstream. These fatty acid molecules may be transported in blood in complex with the protein albumin, and may be taken up by other cells (especially by hepatocytes in the liver) for fatty acid oxidation. In addition to the oxidation of fatty acids to yield citrate and acetyl-CoA, cells (especially hepatocytes) also synthesize fatty acids for storage to meet future energetic needs. Newly synthesized fatty acid molecules may be transported in the bloodstream in the form of

triacylglycerols within chylomicrons as well as in VLDL. These molecules can be utilized to transport fatty acids for storage in adipocytes.

Fatty acid metabolism is largely controlled by hormonal signaling [Voet et al., 2013]. The activity of the enzyme hormone sensitive lipase is modulated by phosphorylation and dephosphorylation by the enzyme PKA. The activation of PKA is in turn regulated by cAMP concentrations which are influenced by multiple hormonal signals including epinephrine, norepinephrine, and glucagon. In addition to modulating the rate of triacylglycerol lipase in liver and muscle cells, PKA also is an important regulator of fatty acid synthesis. In this way the control of fatty acid breakdown and synthesis is controlled in a coupled manner (when the rate of one process increases, the rate of the other process decreases).

In summary then, a single metabolic process may involve biochemical reactions that take place in multiple cells (of different types), across multiple tissues, with metabolic control influenced by paracrine signaling, endocrine signaling, cellular metabolite concentrations, and extracellular metabolite concentrations.

3 Dysregulation of metabolism and disease

Dysregulation of metabolic processes is a prominent feature of human diseases. As a disease, cancer provides a salient example of the relationship between metabolism and disease. While cancer represents a broad class of diseases characterized by abnormal cell growth affecting multiple tissue and organ types, the reprogramming of cellular metabolism is a common feature of all cancers [Pavlova and Thompson, 2016]. For cancer cells to continue to grow and divide, new sources of nutrients are required and the more efficient use of nutrients in the tumor environment is required. These requirements lead to metabolic changes associated with cancer that have been described [Pavlova and Thompson, 2016] as six features: (1) changes in the glucose and amino acid uptake by cancer cells and tumor tissue, (2) “opportunistic” manners of nutrient utilization, (3) increased utilization of glycolysis and the TCA cycle intermediates for synthesizing intermediates as well as for increas-

ing the production of cofactors, (4) increased requirement of nitrogen, (5) changes in gene expression that are modulated by metabolite concentrations, and (6) abnormal metabolic interactions between cells and the surrounding environment.

In addition to cancer, metabolic syndromes are marked by dysregulation of metabolism. Changes in the metabolism of branch-chain amino acids have emerged as hallmarks of metabolic syndromes, the development of type 2 diabetes, and insulin resistance in general [Newgard, 2017]. Increases in the concentrations of branch chain amino acids (BCAAs) have been observed in human subjects that progress to type 2 diabetes versus those who do not in the Framingham study [Wang et al., 2011]. Similarly, 2-Aminoadipic acid has been shown to be associated with development of type 2 diabetes, as well as with insulin resistance and beta cell function [Wang et al., 2013]. The model posited for which these findings are consistent is that increased concentrations of BCAAs interfere with lipid oxidation in skeletal muscle which may increase insulin resistance [Newgard, 2017]. While much of the metabolomics studies using human cohorts to study metabolic syndromes have focused on biomarkers of disease risk and progression, other studies have focused on the mechanistic roles of specific metabolites. For example, while previous studies had shown the association between 3-carboxy-4-methyl-5-propyl-2-furanpropanoic acid (CMPF) in type 2 and gestational diabetes development, CMPF was shown to directly induce β cell dysfunction [Prentice et al., 2014] and to induce metabolic remodeling [Liu et al., 2016].

Coronary heart disease, a consequence of atherosclerosis, is the leading cause of death throughout the world [Benjamin et al., 2017]. Atherosclerosis is a disease process that is defined by the accumulation of lipids and other elements in the sub-endothelial space of arteries [Tabas et al., 2015, Falk, 2006, Lusis, 2000]. Multiple different cell types play a role in atherosclerosis including leukocytes, endothelial cells, and smooth muscle cells. Atherosclerosis is also characterized by inflammatory activation in both endothelial and smooth muscle cells. The stages of atherosclerotic plaque development have been described previously [Stoll and

Bendszus, 2006], and include endothelial dysfunction, the accumulation of lipoproteins in the vessel intima, recruitment of leukocytes, and the formation of foam cells. In the later stages of plaque progression, an atherosclerotic plaque evolves as fibrosis and calcification take place. While the progression of atherosclerotic plaques can be a complication on their own as the vessel lumen narrows, plaque rupture or erosion can have a severe consequence—acute coronary events [Arbab-Zadeh et al., 2012]. Plaque rupture occurs when a defect in the fibrous cap of an atherosclerotic cap exposes the core of the cap to blood, which can precipitate the formation of a thrombus (or blood clot) [Bentzon et al., 2014]. Plaque erosion, in contrast, entails the erosion of the endothelium and subsequent formation of a thrombus overlaying the plaque.

A thrombus overlaying a disrupted plaque in a coronary artery that disrupts the flow of blood can lead to acute coronary syndromes such as myocardial infarction [Arbab-Zadeh et al., 2012]. Myocardial infarction (MI) is defined as myocardial cell death that occurs due to ischemia, or inadequate blood supply [Thygesen et al., 2012]. MI that follows the rupture or erosion of an atherosclerotic plaque is known as Type 1 MI. In addition to this type of myocardial infarction, MI may occur due to other etiological causes such as coronary artery vasospasm or a fixed supply-demand imbalance that follows the narrowing of an artery due to atherosclerosis [Thygesen et al., 2012]. Other classifications of MI include MI associated with cardiac death and MI associated with revascularization procedures.

Metabolic processes are central to the etiology of both the constitutive disease processes that define heart disease (e.g. atherosclerosis) and acute disease events such as myocardial infarction. The role of lipid metabolism in atherosclerosis has been documented for over sixty years (see, for example, the seminal article “Lipid metabolism and atherosclerosis” by Gould, 1951 [Gould, 1951]). For atherosclerotic plaques to form, excess LDL and triglyceride-rich lipoproteins must be present in circulation [Nordestgaard, 2016]. Levels of both remnant cholesterol and triglyceride-rich lipoproteins are both a function of the rate of flux through

metabolic processes including cholesterol synthesis, reverse cholesterol transport, and the utilization of cholesterol for anabolic processes [Baynes and Dominiczak, 2014]. A ubiquitous therapeutic intervention that has shown benefits in the treatment of coronary heart disease is statins, a class of drugs whose mechanism of action is to inhibit the enzyme 3-hydroxy-3-methylglutaryl-coenzyme A-reductase, an enzyme involved in the synthesis of cholesterol [Chait and Eckel, 2016]. As endothelial cell dysfunction is a component of the formation of atherosclerotic lesions, metabolic changes in endothelial cells impact the process of lesion formation [Gimbrone and Garca-Cardea, 2016]. While endothelial cell dysfunction and activation leads to a generally maladaptive inflammatory response that contributes to the development and progression of atherosclerotic lesions, the metabolic state of macrophages is also a factor [Bories and Leitinger, 2017]. In addition to total number of macrophages present, the ratio of pro-inflammatory M1 macrophages versus anti-inflammatory M2 macrophages, which have distinct bioenergetic and metabolic phenotypes, may play a role in the progression of atherosclerosis [Bories and Leitinger, 2017].

Myocardial infarction (MI) as an acute disease event, may occur as a function of the disposition or fate of a disrupted atherosclerotic plaque in a coronary artery [Arbab-Zadeh and Fuster, 2015]. First, the formation and progression of atherosclerotic plaques is required, which is largely a product of disordered lipid metabolism and immunometabolic processes, and second a pathological response to plaque disruption is required. A pathological response to plaque rupture may be conceptualized as an imbalance between pro-thrombotic factors and pro-resolving factors [Arbab-Zadeh and Fuster, 2015]. Thrombosis is the process of blood coagulation in a localized area. Blood coagulation is a metabolic process that involves multiple serine proteases that circulate as proenzymes. A very brief and general description of the process of thrombosis is as follows [Baynes and Dominiczak, 2014, Gale, 2010, Hoffman and Monroe, 2007]. First, activated platelets adhere to and aggregate at the site of vessel injury and form an initial platelet plug [Gale,

2010]. The “intrinsic pathway” is then activated by platelets (the “extrinsic pathway” may simultaneously be activated due to tissue damage) which consists of a sequence of reactions that culminates in the final “common pathway” [Hoffman and Monroe, 2007]. The common pathway includes a prothrombin activator complex that forms to activate prothrombin to thrombin, and the activation of fibrinogen to fibrin by thrombin. This final step of the common pathway allows for the formation of a fibrin mesh in which platelets and erythrocytes are also trapped. Fibrinolysis, which is also a metabolic process defined by serine proteases that circulate as proenzymes, works to resolve blood clots formed by cross-linked fibrin [Gale, 2010]. Briefly, this process involves the activation of plasminogen into plasmin by either urokinase-type or tissue-type plasminogen activators. Once plasmin is generated from plasminogen, it can cleave fibrin allowing for the breakdown of a clot. While the presentation of thrombosis and fibrinolysis presented herein are fairly simplified, it can be observed that the formation and resolution of blood clots which could occlude a coronary artery and could lead to myocardial infarction are defined by the dysregulation of metabolic processes.

4 From metabolism to metabolomics

While multiple definitions of metabolomics have been proposed, the fundamental conceptual definition is that metabolomics is the study of small molecules in a biological sample [Nicholson and Lindon, 2008, Johnson et al., 2016, Newgard, 2017]. The term metabonomics has previously been used almost synonymously with metabolomics, although the focus of this discipline as posited previously study of the changes in metabolic processes following experimental manipulation [Nicholson and Lindon, 2008]. Given these definitions, it can be noted that metabonomics requires metabolomics, that is in order to study changes in metabolic processes that follow an experimental manipulation, one must measure small molecules from biological samples. In other words, “metabonomics” represents the goal of many “metabolomics” studies. For the purposes of the current work we use the term

“metabolomics” to refer to both the analysis of small molecules from a sample, as well as the study of changes in metabolic processes that follow an experimental manipulation.

As an -omics discipline, several aspects of metabolomics are unique. The first such aspect is that the background set of which metabolites could be present in humans is unknown and remains elusive. Specifically 18,557 unique metabolites have been detected and quantified in humans (as recorded in the HMDB), while 82,274 “expected” metabolites have been determined [Wishart et al., 2018]. Of these, 5,498 metabolite-disease associations have been noted in the Human Metabolome Database (HMDB). “Expected” metabolites are metabolites that have not been detected in human samples, yet the molecular structure is known and it is hypothesized that due to exposure, these metabolites are expected to be present in humans. This stands in sharp contrast to genomics, in which a reference genome was produced following the completion of the human genome project in 2003 [Collins, 2003, O’Leary et al., 2016]. While functional annotation of the genome remains ongoing [O’Leary et al., 2016], there is substantially less uncertainty as to what the “reference” genome should be than what a “reference” metabolome should be. Second, the determination of tissue specificity is substantially less advanced in metabolomics than in transcriptomics and proteomics, as well in evaluating the effect of gene variants on tissue specific gene expression. For example, the Tissue-specific Gene Expression and Regulation (TiGER) database began in 2007 with an expressed sequence tag (EST) approach to cataloging tissue-specific gene expression profiles [Liu et al., 2008]. The study of tissue-specific gene expression has continued with the utilization of RNA-seq data as has been done with the Genotype-Tissue Expression (GTEx) project [Lonsdale et al., 2013]. Additional efforts have used genomic sequencing in conjunction with gene expression profiling to evaluate whether expression quantitative trait loci (eQTLs) are active in specific tissues [Aguet et al., 2017, Brown et al., 2017]. To generate a map of tissue-specific protein expression Uhlen, et al. [Uhlen et al., 2015] used both

RNA-seq data and antibodies corresponding to 16,975 protein-encoding genes to quantify tissue-specificity, which is included in the Human Protein Atlas [Uhlen et al., 2010]. In contrast to genomics, transcriptomics, and proteomics, a comprehensive mapping of metabolite tissue-specificity does not yet exist. To summarize the first two nuances we have described, it is often not feasible to have *a priori* knowledge of what metabolites could or should be present in a human sample irrespective of the type of sample (tissue, urine, plasma, serum, cell culture).

The third aspect that distinguishes mass spectrometry-based metabolomics from other -omics disciplines is that often a substantial proportion of the mass features that are detected in a sample (using un-targeted mass spectrometry approaches) are not able to be identified (often greater than 2/3 of the features) [Newgard, 2017]. At this point it is informative to differentiate the analytical strategies employed in mass spectrometry-based metabolomics studies into three different types: targeted, un-targeted / non-targeted, and stable isotope resolved. Stable isotope resolved metabolomics can be conducted using either a targeted or non-targeted approaches. Targeted metabolomics corresponds to the measurement of metabolites specified in advance, for which an analytical techniques such as selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) assays have been developed [Roberts et al., 2012, Zamboni et al., 2015]. Multiple reaction monitoring makes use of mass spectrometers in which parent ions are first selected, fragmented into smaller ions, with these ions being selected and detected [Roberts et al., 2012]. The use internal standards allows for such methods to be quantitative or semi-quantitative. Targeted metabolomics experiments require substantial upfront methods development in order to determine the specific transitions for each metabolite as well as retentions times and other experimental parameters prior to analysis of samples. In contrast to this approach, un-targeted or non-targeted metabolomics does not stipulate the prior specification of compounds to be quantified in the analysis [Zamboni et al., 2015, Putri et al., 2013, Roberts et al., 2012]. The principal challenge of un-targeted metabolomics is the identi-

fication and resolution of the chemical identity of the thousands of features that are detected from any one biological sample [Zamboni et al., 2015]. To summarize metabolomics is distinct as an -omics discipline as there is a lack of knowledge about what compounds (of endogenous or exogenous origin) should be found in a human, a lack of knowledge about where such compounds should be localized to, and a relatively low identification rate in determining the chemical identity of ions in mass spectrometry-based metabolomics.

Stable Isotope Resolved Metabolomics, while a subset of metabolomics experiments, represents a distinct approach to studying metabolism using metabolomics. SIRM entails the treatment of a system with nutrients that are enriched with isotopes such as ^{13}C or ^{15}N [Bruntz et al., 2017, Newgard, 2017, Higashi et al., 2014, Fan et al., 2012]. A specific amount of time is allowed to elapse (specific to the design of experiments) to allow the system to incorporate the stable isotope labeled intermediates and for them to propagate across metabolic pathways. Targeted or un-targeted quantification of metabolites by either mass spectrometry or nuclear magnetic resonance may then be utilized. By examining the fractional enrichment of isotopologues, inference can be made regarding the flux through specific metabolic pathways [Higashi et al., 2014, Fan et al., 2012]. Critically, this approach can be utilized to study metabolic reprogramming, as is associated with the initiation and progression of cancer [Bruntz et al., 2017].

5 Mass spectrometry-based metabolomics

Mass spectrometry has been a central analytical tool in facilitating the rapid growth of metabolomics [Dunn and Hankemeier, 2013]. The fundamental principal of mass spectrometry is the sorting along and determination of the mass-to-charge ratio of molecules following ionization. The principle advantage of utilizing mass spectrometry in conjunction with or in lieu of nuclear magnetic resonance (NMR), is the relatively high sensitivity of mass spectrometry [Lei et al., 2011, Gowda and Djukovic, 2014, Gross, 2017]. Mass spectrometry allows for the determina-

tion of accurate mass, isotope distribution, and often the spectra of fragment ions may be used for compound identification. Multiple experimental designs (in terms of instruments and parameters) have been utilized in mass spectrometry-based metabolomics. Critical design facets include: the type of (or absence of) chromatographic separation prior to injection into the mass spectrometer, the ionization source (and relevant parameters), and type of mass analyzer. In this section we briefly discuss some of these design facets. Various techniques are employed in common practice to ionize molecules in a sample [Lei et al., 2011, Gowda and Djukovic, 2014, Gross, 2017]. Electron ionization (EI) involves bombarding molecules to be analyzed with electrons to form ions. EI is considered to be hard ionization technique as the energy impacted by the electron collisions can cause the fragmentation of molecules. In contrast, electrospray ionization (ESI) is considered to be a soft ionization technique as it does not lead to fragmentation. ESI is most common ionization technique with liquid chromatography as it efficiently ionizes molecules in the liquid phase. Other ionization methods utilized in metabolomics include matrix assisted laser desorption/ionization (MALDI) and atmospheric pressure chemical ionization (APCI) [Lei et al., 2011].

Once molecules in a sample have been ionized, resulting in positively or negatively charged molecules, they can be detected by a mass analyzer. High resolution mass analyzers include Fourier transform ion cyclon resonance (FT-ICR) mass analyzer and Orbitrap mass analyzers [Lei et al., 2011, Gowda and Djukovic, 2014]. These types of mass analyzers make it possible for many molecules to enumerate molecular formula from accurate mass alone. Time of flight (TOF) mass analyzers, quadrupole mass analyzers and quadrupole ion traps mass analyzers typically have less resolution than FT-ICR but may possess other advantages. For example, quadrupole mass analyzers allow for developing targeted assays via studying the transitions that occur with collision-induced dissociation.

While typically mass spectrometers are coupled with chromatography for separation, direct injection (DI) MS analyses are also technically feasible [Lei et al.,

2011], especially given high resolution mass detectors. A complication of DI-MS analyses is that they are often susceptible to ion suppression. Ion suppression is a phenomenon in which molecules with low affinity for electrons or protons are less likely to ionize when they co-elute with molecular ions with higher electron or proton affinity [Gowda and Djukovic, 2014]. The most common forms of chromatographic separation are gas chromatography (GC), liquid chromatography (LC), and capillary electrophoresis (CE). Gas chromatography may achieve a high degree of separation between molecules especially when conducted as two-dimensional GCxGC-MS. A drawback of GC is that it requires compounds to be volatile (and thermally stable), which is often not the case for many biomolecules [Lei et al., 2011]. To ensure compounds are in a volatile state, a derivatization procedure such as alkylation or silylation may be applied. LC is a commonly employed with ESI as the efficiency of ionization is high [Lei et al., 2011]. Often multiple types of LC are combined to provide greater profiling of the metabolites in a sample—for example hydrophilic interaction liquid chromatography (HILIC) coupled with reverse-phase liquid chromatography. This coverage can be increased further when both positive and negative mode ESI are applied.

The use of NMR in metabolomics has a few advantages including a relatively high degree of reproducibility, it is relatively quantitative, the identity of metabolites can often be resolved unambiguously, and it is non-destructive [Nagana Gowda and Raftery, 2016]. However, as the metabolomics data analyzed within this work was generated using mass spectrometry we do not discuss NMR-based metabolomics in the present work.

CHAPTER III

PRELIMINARIES: BAYESIAN STATISTICS

Bayesian statistics represents both a sub-discipline of statistics as well as a philosophical paradigm. A philosophical discussion of the relative merits of frequentist versus Bayesian statistics is beyond the scope of this text. However, we do note the following general distinctions between frequentist and Bayesian statistics. First, frequentist approaches assume that a population parameter is fixed, while samples are drawn (a stochastic or random process) from a population to estimate this population parameter. Contrastingly, Bayesian approaches assume that the sample data is fixed, but the population parameter is a random variate for which a probability distribution can be specified. Given a fixed population parameter, frequentist approaches are based on the long run behavior of drawing repeated samples from a population. For example, a 95% confidence interval (CI) for a population parameter means that if 100 samples were randomly drawn a CI constructed in an identical manner 95 would contain the population value on average. This implies that the frequentist approaches consider the likelihood of both observed and unobserved data. In contrast, Bayesian approaches evaluate the probability distribution of a population parameter the observed sample. This probability distribution can be determined by considering the integrating the likelihood of the observed sample given each possible parameter values times the probability of these parameter values, divided by the likelihood of the sample. Consequently, Bayesian approaches require the prior stipulation of the probability distribution of the population parameter.

1 Bayes rule

The central foundation of Bayesian statistics is Bayes rule [Gelman et al., 2004].

Theorem 1.1. *If A and B are events with $P(B) > 0$, then:*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

.

Further, Bayes rule can be extended to any arbitrary partitioning of a sample space [Casella and Berger, 2002].

Theorem 1.2. *Let $\mathcal{P} = \{A_i : i = 1, 2, \dots, N\}$ represents an arbitrary partitioning of a sample space S , then for a specific A_i :*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{A_j \in \mathcal{P}} P(B|A_j)P(A_j)}. \quad (2)$$

2 Bayesian inference regarding a parameter

Let θ (or $\boldsymbol{\theta}$ in the multivariate case) represent a parameter of a probability mass function (pmf) or probability distribution function (pdf) $f(x|\theta)$ of a random variable X . Without loss of generality, we discuss the univariate case of one random variable X and one population parameter θ , although multivariate generalization is straightforward. The objective of statistical inference regarding θ is to utilize a random sample X_1, X_2, \dots, X_n from a population with pmf/pdf $f(x|\theta)$ to determine likely values of θ [Casella and Berger, 2002]. Denoting a fixed sample as $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, we note that \mathbf{x} is a realization from the sample space \mathcal{X} of all possible samples from the population. In Bayesian inference, the population parameter θ is regarded as unknown [Hoff, 2009]. It is assumed that the uncertainty regarding the true population value of θ , can be represented by the prior probability distribution $p(\theta)$. Since the true population value is unknown, a set of possible values for θ , that is the parameter space Θ comprises the support of $p(\theta)$. In Bayesian analysis, a sampling model describing the probability of observing

realization x of the random variable X conditioned on a fixed value of the parameter θ , that is $p(x|\theta)$ must also be specified. When considering a fixed sample, this term is also referred to as the likelihood [Casella and Berger, 2002].

Theorem 2.1. *If X_1, X_2, \dots, X_n are independent and identically distributed (iid) random variables, then $f(x_1, x_2, \dots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$.*

More specifically, we refer to the joint distribution function of a sample x , conditional on the value of the parameter θ as the likelihood of theta given the sample, or $\mathcal{L}(\theta|\mathbf{x})$. Given a prior distribution for $\theta \in \Theta$, a random sample from the population \mathbf{x} , and a sampling model or likelihood, the posterior distribution for θ can then be determined utilizing Bayes rule, as below [Hoff, 2009]:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int_{\tilde{\theta} \in \Theta} p(\mathbf{x}|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}. \quad (3)$$

The denominator term is referred to as the marginal likelihood. As the marginal likelihood involves only a fixed sample, the denominator is a constant, which justifies the following relation involving the posterior distribution of θ :

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)p(\theta). \quad (4)$$

From this relation, it can be noted that the posterior distribution for θ , after observing a random sample \mathbf{x} is proportional to the likelihood times the prior.

3 Example Bayesian inference regarding a parameter

Consider a mass spectrometer with a counting detector that records the counts per second over a fixed m/z window. The number of counts recorded by the detector over a fixed time interval may be a Poisson process. Let X be the random variable representing the number of counts recorded by the detector over a one second time interval, with an average number of counts per second (population parameter) of λ . We then say that $X \sim Pois(\lambda)$, and note that the probability mass function for X is:

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \mathbb{N}, \quad \lambda \in [0, \infty) \quad (5)$$

Assume a sample has been observed with the following counts: {107, 103, 110, 99, 108, 108, 105} and that we wish to determine plausible values of the rate parameter λ . In order to determine plausible values, we will conduct Bayesian inference over λ . Assume that we have collected ten prior count observations and that the sum of counts from these observations was 1,000. We may then choose a prior distribution for λ of $Gamma(\alpha, \beta)$ with parameters $\alpha = 1000$ and $\beta = 10$. We note the following form of the Gamma distribution pdf:

$$f(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda \in [0, \infty), \quad \alpha, \beta > 0. \quad (6)$$

We note that the likelihood for the Poisson distribution factorizes as in the following:

$$\mathcal{L}(\lambda|\mathbf{x}) = p(\mathbf{x}|\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \left(\prod_{i=1}^n x_i! \right)^{-1} \times \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}. \quad (7)$$

From Eq. 7 we can note that the likelihood term that involves λ and are is not fixed by the sample \mathbf{x} is $\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}$. Thus the posterior distribution for λ has the following form:

$$p(\lambda|\mathbf{x}) \propto p(\lambda)p(\mathbf{x}|\lambda) \propto p(\lambda)\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda} \quad (8)$$

Considering the prior distribution that we have specified:

$$p(\lambda|\mathbf{x}) \propto \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda} \quad (9)$$

$$\propto \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha+\sum_{i=1}^n x_i-1} e^{-(\beta+n)\lambda}, \quad (10)$$

we can note that:

$$p(\lambda|\mathbf{x}) = c(\mathbf{x}, \alpha, \beta)^{-1} \lambda^{\alpha+\sum_{i=1}^n x_i-1} e^{-(\beta+n)\lambda}, \quad (11)$$

where $c(\mathbf{x}, \alpha, \beta)$ is a normalizing constant that depends only on the prior distribution parameters and the observed sample. Thus it is evident that the posterior distribution for λ is also a Gamma distribution, specifically $p(\lambda|\mathbf{x}) \sim Gamma(\alpha + \sum_{i=1}^n x_i, \beta + n)$. Fig. 3 shows the pdf of both the prior and posterior probability distributions for the Poisson rate parameter λ .

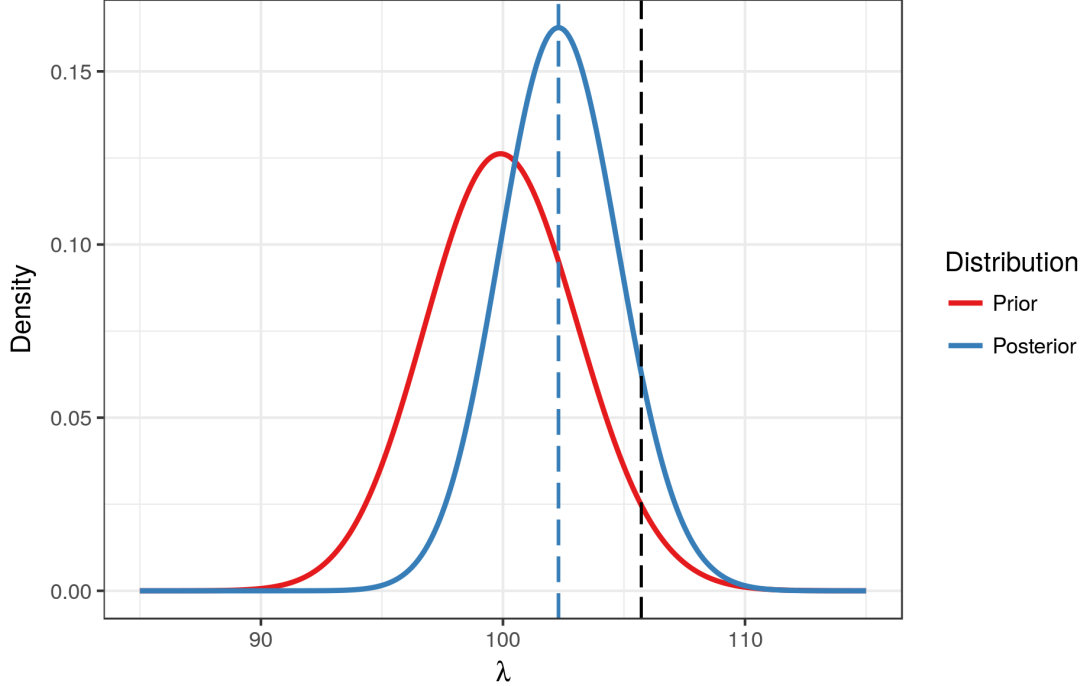


Figure 3. Example Bayesian estimation of a Poisson rate parameter. The plot shows the probability density function of the prior on posterior distributions for the Poisson rate parameter λ . Dashed blue vertical line shows the mode of the posterior distribution. Also shown as a dashed black vertical line is the maximum likelihood estimator determined from the sample \mathbf{x} .

As can be observed, the center (mode or mean) of the posterior probability distribution for λ lies between the maximum likelihood estimator (MLE) for λ and the prior distribution. In fact, while the MLE for λ is $\hat{\lambda}_{MLE} = 105.7143$, the mode of the posterior distribution of λ is:

$$\tilde{\lambda} = \frac{\alpha + \sum_{i=1}^n x_i - 1}{\beta + n} = \frac{1000 + 105.7143}{10 + n} = 102.2941.$$

In addition to point-wise summaries (such as the mode of the posterior distribution), interval summaries can be provided utilizing the posterior distribution of a parameter. In this case, a Bayesian credible interval [Gelman et al., 2004] can be analytically computed using a quantile function as the analytical form of the posterior distribution is known. Specifically, a two-sided 95% Bayesian credible interval (CI) can be determined by finding the unique solution to (l, u) , such that $F(l|\alpha + \sum_{i=1}^n x_i - 1, \beta + n) = 0.025$ and $F(u|\alpha + \sum_{i=1}^n x_i - 1, \beta + n) = .975$ where

$F(x)$ is the cumulative distribution function for the Gamma distribution. In the current example, this yields a 95% CI of (97.600, 107.218).

4 Specification of prior distribution

Given that the posterior distribution of parameters depends on both the data likelihood and the prior distribution, the specification of prior distribution is of critical interest in Bayesian inference. While a general distinction is made between “informative” priors and “non-informative” priors [Gelman, 2006], “weakly informative” priors represent an in-between. Non-informative priors are prior distributions that are vague, flat, diffuse [Gelman et al., 2004], and do not incorporate external information into posterior probability distributions. Weakly informative priors generally incorporate some other goal, such as encouraging regularization without introducing specific *a priori* knowledge about the prior likelihood of parameter values. In contrast, informative priors do incorporate external information (outside of the empirical data) [Gelman et al., 2004]. Informative priors may be generated from previous experimentation, often by the use of conjugate distributions [Hoff, 2009].

Definition 4.1. *Conjugate probability distribution* Let \mathcal{F} be a class of probability distributions from which x may be sampled with parameter θ , that is $f(x|\theta)$ and let \mathcal{P} be a class of prior distributions for θ . \mathcal{P} is said to be conjugate to \mathcal{F} if $p(\theta|x) \in \mathcal{P} \forall f(\cdot|\theta) \in \mathcal{F}$ and $p(\cdot) \in \mathcal{P}$.

Following these definitions, we return to our discussion of the counting mass detector example. This example utilized an informative conjugate prior distribution for the Poisson rate parameter. Given the form of the data-augmented posterior distribution, $p(\lambda|\mathbf{x}) \sim \text{Gamma}(\alpha + \sum_{i=1}^n x_i, \beta + n)$, the posterior Gamma distribution parameters can be noted as $\alpha' = \alpha + \sum_{i=1}^n x_i$ and $\beta' = \beta + n$, where α and β are the prior parameters. This provides a natural way to encapsulate prior empirical evidence. Noting that prior expectation for λ is $E(\lambda) = \frac{\alpha}{\beta}$, while the

posterior expectation conditional expectation is:

$$E(\lambda|\mathbf{x}) = \frac{\alpha + \sum_{i=1}^n x_i}{\beta + n} = \frac{\alpha}{\beta + n} + \frac{\sum_{i=1}^n x_i}{\beta + n}, \quad (12)$$

it can be seen that if the sum of counts observed in β previous experiments (where β is a natural number) was α , then the conditional expectation of λ is a weighted average of the prior observed counts and the current observed counts (weighted by the respective number of observations).

5 MCMC methods for inference

Many posterior probability distributions do not have an expression in analytical form, or it may be challenging to determine such a form. In such cases, Markov chain Monte Carlo (MCMC) methods may be utilized for simulating the posterior distribution. The fundamental principal of Monte Carlo methods is that if an integral $I = \int_{\Omega} f(\mathbf{x})d\mathbf{x}$ cannot be obtained analytically, then the law of large numbers justifies generating a random sample $\{\mathbf{X}_i\}_{i=1}^n$ from a distribution $p(x)$ over Ω and using it for approximation [Gelman et al., 2004], that is:

$$I \approx \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i), \quad (13)$$

where $g(x) = f(x)/p(x)$. Two commonly utilized MCMC approaches are Metropolis-Hastings type algorithms which rely on formulating a proposal density for accepting or rejecting random moves according to the target distribution and Gibbs sampling which draws from conditional distributions.

Gibbs Sampling

In addition to not having an analytical form of the joint posterior distribution of model parameters (which justifies the use of posterior simulation) it may be the case that it is straightforward to sample from conditional distributions of parameters but not for sampling from the joint distribution. In such cases, Gibbs sampling may be employed [Gelman et al., 2004, Hoff, 2009]. The idea of Gibbs sampling is

straightforward. First, parameters estimates are initialized to some initial values. Next, one parameter is fixed as the target parameter. A sample is then drawn from the conditional distribution of this target parameter (conditioned on both the non-target parameter estimates and the observed sample). Gibbs sampling can be extended to using target sets of parameters without loss of generality. A more precise definition of a Gibbs sampler is as follows and pseudocode is presented as Alg. 1. Let $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_p\}$ be a vector of parameters and let $\boldsymbol{\theta}_j$ be a vector defined by a subset of the parameters with the subsets indexed by $j = 1, 2, \dots, J$. Let $\tilde{\boldsymbol{\theta}}_j^{(t)}$ denote the estimate of the subset of parameters at time t .

Algorithm 1 Gibbs sampler

```

1: Set  $\boldsymbol{\theta}^{(0)}$  to some initial values
2: while  $t \leq t_{max}$  or convergence criteria not met do
3:   Set  $\tilde{\boldsymbol{\theta}} \leftarrow \tilde{\boldsymbol{\theta}}^{(t-1)}$ 
4:   for  $j = 1$  to  $J$  do
5:     Sample  $\tilde{\boldsymbol{\theta}}_j^* \sim p(\boldsymbol{\theta}_j | \tilde{\boldsymbol{\theta}}_{-j}, \mathbf{x})$ 
6:     Update  $\tilde{\boldsymbol{\theta}}_j \leftarrow \tilde{\boldsymbol{\theta}}_j^*$ 
7:   end for
8:   Set  $\tilde{\boldsymbol{\theta}}^{(t)} \leftarrow \tilde{\boldsymbol{\theta}}$  and return  $\tilde{\boldsymbol{\theta}}^{(t)}$ 
9:    $t \leftarrow t + 1$ 
10: end while

```

Example application of Gibb’s sampling

Continuing with the example of a mass spectrometer with a counting detector, we illustrate the use of Gibb’s sampling for simulating the posterior distribution of a set of model parameters. In the previous example, we assumed that there was only one population rate parameter. In the current example, we consider a case in which, over a fixed m/z window, counts may be observed for three types of ions within the same m/z window as opposed to one. To the observer, it is not clear which ions the counts are originating from. We would expect, however that the population rate parameters are different for each ion. This can be formulated using the following model [Diebolt and Robert, 1994, Everitt, 2004]. First we note

the pmf for individual observations x_i :

$$f(x_i|\boldsymbol{\lambda}, \mathbf{p}) = \sum_{j=1}^3 p_j \frac{e^{-\lambda_j} \lambda_j^{x_i}}{x_i!} \quad (14)$$

where j is the indicator for the source ion, each x_i is a random variable representing the counts, λ_j represents the population rate parameter for each j source ion, and p_j represents the probability that a count observation belongs to the j source ion.

For each observation i we then introduce a vector of latent variable for indicating which source ion the observation originated from: $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})$ where:

$$z_{ij} = \begin{cases} 1 & \text{if the counts } x_i \text{ originate from source ion } j \\ 0 & \text{else} \end{cases} . \quad (15)$$

The individual z_{ij} are then Bernoulli distributed with probability p_j and the individual vectors \mathbf{z}_i are distributed as $\mathbf{z}_i \sim \text{Multi}(1, p_1, p_2, p_3)$ with $\sum_{j=1}^3 p_j = 1$. Given the observations are exchangeable, we then have the indicator vector likelihood :

$$f(\mathbf{z}|\mathbf{p}) = \prod_{i=1}^n f(\mathbf{z}_i|\mathbf{p}) = \prod_{i=1}^n \prod_{j=1}^3 p_j^{z_{ij}}. \quad (16)$$

We then can specify a (conjugate) prior distribution for \mathbf{p} , that is $f(\mathbf{p}) \sim \text{Dirichlet}(\boldsymbol{\alpha})$.

We choose the non-informative prior $\alpha_j = 1 \ \forall j$. We place a Gamma distribution with shape and scale priors r and s over the rate parameters, that is: $\lambda_j \sim \text{Gamma}(r, s)$. and we assume that the rate parameters are independent (of each other prior parameter). By application of Bayes' rule and the chain rule, we can determine the posterior distribution of the parameters:

$$\begin{aligned} f(\mathbf{z}, \mathbf{p}, \boldsymbol{\lambda}|\mathbf{x}) &\propto f(\mathbf{x}|\mathbf{z}, \mathbf{p}, \boldsymbol{\lambda})f(\mathbf{z}, \mathbf{p}, \boldsymbol{\lambda}) \\ &\propto f(\mathbf{x}|\mathbf{z}, \mathbf{p}, \boldsymbol{\lambda})f(\mathbf{z}, \mathbf{p})f(\boldsymbol{\lambda}) \\ &\propto f(\mathbf{x}|\mathbf{z}, \mathbf{p}, \boldsymbol{\lambda})f(\mathbf{z}|\mathbf{p})f(\mathbf{p})f(\boldsymbol{\lambda}). \end{aligned} \quad (17)$$

Denoting the gamma pdf's for each λ_j as $f_{\lambda_j}(\lambda_j)$, and likewise the pdf for the Dirichlet prior as $f_{\mathbf{p}}(\mathbf{p})$, we can then enumerate Eq. 17 as:

$$f(\mathbf{z}, \mathbf{p}, \boldsymbol{\lambda}|\mathbf{x}) \propto \prod_{i=1}^n \prod_{j=1}^3 \left(\frac{e^{-\lambda_j} \lambda_j^{x_i}}{x_i!} \right)^{z_{ij}} \times \prod_{i=1}^n \prod_{j=1}^3 (p_j)^{z_{ij}} \times \prod_{j=1}^3 f_{\lambda_j}(\lambda_j) \times f_{\mathbf{p}}(\mathbf{p}) \quad (18)$$

From this, conditional distributions for each parameter are apparent. For the vector of Bernoulli parameters:

$$\begin{aligned}
f(\mathbf{p}|\mathbf{z}, \boldsymbol{\lambda}, \mathbf{x}) &\propto \prod_{i=1}^n \prod_{j=1}^3 (p_j)^{z_{ij}} f_{\mathbf{p}}(\mathbf{p}) \\
&\propto \prod_{i=1}^n \prod_{j=1}^3 (p_j)^{z_{ij}} \\
&\propto \prod_{j=1}^3 p_j^{\sum_{i=1}^n z_{ij}} \\
&\propto \text{Dirichlet}(\boldsymbol{\alpha}'),
\end{aligned} \tag{19}$$

where $\boldsymbol{\alpha}' = (\alpha'_1, \alpha'_2, \dots, \alpha'_k)$ and $\alpha_j = 1 + \sum_{i=1}^n z_{ij}$.

For the rate parameters conditional on the other parameters (using the first as an example), we have:

$$\begin{aligned}
f(\lambda_1|\boldsymbol{\lambda}^{(-1)}, \mathbf{p}, \mathbf{z}, \mathbf{x}) &\propto \prod_{i=1}^n \prod_{j=1}^3 \left(\frac{e^{-\lambda_j} \lambda_j^{x_i}}{x_i!} \right)^{z_{ij}} f_{\lambda_1}(\lambda_1) \\
&\propto \prod_{i=1}^n \left(\frac{e^{-\lambda_j} \lambda_j^{x_i}}{x_i!} \right)^{z_{ij}} \frac{s_1^{r_1}}{\Gamma(r_1)} \lambda_1^{r_1-1} e^{-s_1 \lambda_1} \\
&\propto \prod_{i=1}^n (e^{-\lambda_j} \lambda_j^{x_i})^{z_{ij}} (\lambda_1^{r_1-1} e^{-s_1 \lambda_1}) \\
&= e^{-\lambda_1(\sum_{i=1}^n z_{i1} + s_1)} + \lambda_1^{\sum_{i=1}^n x_i z_{i1} + r_1 - 1}.
\end{aligned} \tag{20}$$

It can thus be seen that the conditional distribution for λ_1 (conditioned on the remaining parameters and the data) is conjugate and is proportional to the following distribution:

$$f(\lambda_1|\boldsymbol{\lambda}^{(-1)}, \mathbf{p}, \mathbf{z}, \mathbf{x}) \propto \text{Gamma} \left(r_1 + \sum_{i=1}^n x_i z_{i1}, s_1 + \sum_{i=1}^n z_{i1} \right). \tag{21}$$

Finally, for the latent indicator variables:

$$\begin{aligned}
f(z_{ij} = 1|x_i, \boldsymbol{\lambda}, \mathbf{p}) &= \frac{f(x_i|\boldsymbol{\lambda}, \mathbf{p}, z_{ij} = 1)f(z_{ij} = 1|\boldsymbol{\lambda}, \mathbf{p}, x_i)}{\sum_{j=1}^3 f(x_i|\boldsymbol{\lambda}, \mathbf{p}, z_{ij} = 1)f(z_{ij} = 1|\boldsymbol{\lambda}, \mathbf{p}, x_i)} \\
&= \frac{f(x_i|\boldsymbol{\lambda}, z_{ij} = 1)f(z_{ij} = 1|\boldsymbol{\lambda}, \mathbf{p})}{\sum_{j=1}^3 f(x_i|\boldsymbol{\lambda}, z_{ij} = 1)f(z_{ij} = 1|\boldsymbol{\lambda}, \mathbf{p})} \\
&= \frac{f(x_i|\lambda_j)p_j}{\sum_{j=1}^3 f(x_i|\lambda_j)p_j} \\
&= \frac{f(x_i|\lambda_j)p_j}{f(x_i)}.
\end{aligned} \tag{22}$$

Thus it can be seen that the conditional distribution for each $z_i^{(-i)}$ is:

$$z_i^{(-i)} | (\boldsymbol{\theta}, \mathbf{p}, \mathbf{z}, \mathbf{x}) \sim \text{Multi} \left(1, \frac{f(x_i | \lambda_1) p_1}{f(x_i)}, \dots, \frac{f(x_i | \lambda_3) p_3}{f(x_i)} \right) \quad (23)$$

From these conditional distributions, a Gibbs sampler can be formulated.

CHAPTER IV

PRELIMINARIES: GAUSSIAN GRAPHICAL MODELING

1 Introduction

Across multiple domains including finance, economics, molecular biology, and machine vision, stochastic systems are observed via the realization of multiple random variables and the determination of the relationship between the random variables is an essential inferential task. For example, researchers in the field of metabolomics often sample from the repertoire of small molecules contained in a cell or biofluid to make inference regarding metabolic responses to the environment or differences across phenotypes [Johnson et al., 2016]. Central to making inferences regarding metabolic processes is determining the structure of probabilistic interactions between sets metabolites. In macroeconomics, to quantify financial risk that could propagate across borders, international bank settlements may be sampled and interrogation of the dependence structure of these random variables reveals cross border flows that may result in cross-border contagion during a financial crisis [Giudici and Spelta, 2016].

2 Gaussian graphical models

An undirected probabilistic graphical model also known as a Markov Random Fields (MRFs) is a graph $G = (V, E)$ in which random variables $X_i \in V$, $i \in \{1, 2, \dots, p\}$ are represented by vertices and edges in the edge set $E \subseteq V \times V$ represent probabilistic interactions [Koller and Friedman, 2009]. If the joint distribution of the random variables is assumed to be a multivariate normal distribution,

that is $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1})$ where $\boldsymbol{\Omega}$ is the concentration matrix and the inverse of the covariance matrix, then the MRF is a Gaussian Graphical Model (GGM), in which each vertex X_i has a marginal normal distribution, and normal conditional distributions $X_i|X_j$. We immediately note the likelihood given a sample $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_1, \dots, \mathbf{x}_n)^T$:

$$L(\boldsymbol{\Omega}|\mathbf{X}) = (2\pi)^{-np/2} |\boldsymbol{\Omega}|^{n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Omega} (\mathbf{x}_i - \boldsymbol{\mu})\right) \quad (24)$$

$$= (2\pi)^{-np/2} |\boldsymbol{\Omega}|^{n/2} \exp\left(-\frac{1}{2} \langle \mathbf{K}, \boldsymbol{\Omega} \rangle\right) \quad (25)$$

where $\sum_{i=1}^n$

In order to determine a Gaussian Graphical model, the graph topology and parameters must be estimated separately [Meinshausen and Bhlmann, 2006] or jointly [Friedman et al., 2007, Yuan and Lin, 2007, Banerjee et al., 2008]. Given a mean centered data matrix \mathbf{X} with $\dim(\mathbf{X}) = n \times p$, estimation of the concentration matrix (inverse of the joint distribution covariance) $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ fully determines the graph topology as well as the multivariate Gaussian distribution parameters. The entries of $\boldsymbol{\Omega}$ are of particular importance; ω_{ij} is the partial correlation between X_i and X_j . Consequently, $\omega_{ij} = 0$ implies X_i and X_j are conditionally independent.

The Graphical Lasso

To find the maximum likelihood estimator of $\boldsymbol{\Omega}$ the log likelihood of the concentration matrix is noted:

$$l(\boldsymbol{\Omega}) \propto \log(\det \boldsymbol{\Omega}) - \text{tr}(\mathbf{S}\boldsymbol{\Omega}), \quad (26)$$

where $\mathbf{S} = \frac{1}{n} \mathbf{X}^T \mathbf{X}$ is the empirical covariance matrix. In the case that $p > n$ maximization of the log likelihood function is not guaranteed to be convex. To overcome this problem, several approaches have been proposed for maximizing the L_1 norm penalized log-likelihood [Friedman et al., 2007, Yuan and Lin, 2007, Banerjee et al., 2008]:

$$l(\boldsymbol{\Omega}) \propto \log(\det \boldsymbol{\Omega}) - \text{tr}(\mathbf{S}\boldsymbol{\Omega}) - \rho \|\boldsymbol{\Omega}\|_1 \quad (27)$$

where ρ is the penalty parameter and the optimization is over the space of positive definite matrices of the same dimension as $\mathbf{\Omega}$. The solution proposed by [Friedman et al., 2007], known as the graphical Lasso employs a block coordinate descent for maximizing the penalized likelihood. To formulate a block-wise procedure the [Friedman et al., 2007] first note the following partitioning along the last row and last column of \mathbf{S} and $\mathbf{W} = \hat{\Sigma}$:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{12}^T & w_{22} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{S}_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{12}^T & s_{22} \end{bmatrix} \quad (28)$$

Given this partition, the maximization of the log-likelihood (Eq. 27) is equivalent to the following constraint problem [Banerjee et al., 2008]:

$$\mathbf{w}_{12} = \operatorname{argmin}_{\mathbf{y}} \{ \mathbf{y}^T \mathbf{W}_{11}^{-1} \mathbf{y} : \|\mathbf{y} - \mathbf{s}_{12}\|_{\infty} \leq \rho \}. \quad (29)$$

[Banerjee et al., 2008] show that by convex duality, if $\hat{\beta}$ minimizes:

$$\hat{\beta} = \operatorname{argmin}_{\beta} \left\{ \frac{1}{2} \|\mathbf{W}_{11}^{1/2} \beta - b\|^2 + \rho \|\beta\|_1 \right\} \quad (30)$$

where $b = \mathbf{W}_{11}^{-1/2} \mathbf{s}_{12}$, then $\mathbf{w}_{12} = \mathbf{W}_{11} \hat{\beta}$.

The SCAD penalty and the Adaptive Graphical Lasso

It has been shown that the linear increase penalization relative to the norm incurred with L_1 regularization of parameters introduces bias, especially in the case of parameters that have large magnitude [Fan et al., 2009, Lam and Fan, 2009]. Two alternative penalized likelihoods have been proposed as a solution to this known problem and have been shown to satisfy the oracle property: the smoothly clipped absolute deviation (SCAD) penalization and the adaptive lasso. It has been shown that the oracle property is not universally satisfied by the lasso [Zou, 2006]. The lasso, SCAD penalized estimation, and adaptive lasso share similar developmental histories, being developed first for variable selection and linear/generalized linear model parameter estimation with later extension to GGM parameter estimation. The smoothly clipped absolute deviation is a non-concave

penalty function defined as [Fan and Li, 2001]:

$$f_{\lambda,a}(\theta) = \begin{cases} \lambda|\theta| & \text{if } |\theta| \leq \lambda \\ -\left(\frac{|\theta|^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)}\right) & \text{if } |\theta| \in (\lambda, a\lambda] \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\theta| > a\lambda \end{cases} \quad (31)$$

The behavior of this penalty with respect to coefficient magnitude can be examined given the continuous derivative derivative of the SCAD penalty function:

$$f'_{\lambda,a}(\theta) = \lambda \left[I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right] \quad (32)$$

for $\theta > 0$ and $a > 2$. In general, the advantage of this penalization over L_1 norm penalization is that large values of θ are not excessively penalized. The two parameters of the penalty function can be optimized by cross-validation or via minimization of Bayes risk as was done in [Fan and Li, 2001]. To generalize the SCAD penalty for regularized estimation of GGMs, the SCAD penalized log-likelihood is noted:

$$l(\mathbf{\Omega}) \propto \log(\det \mathbf{\Omega}) - \text{tr} \left(\frac{\mathbf{S}}{n} \mathbf{\Omega} \right) - \sum_{i=1}^p \sum_{j=1}^p f_{\lambda,a}(|\omega_{ij}|)$$

After demonstrating the conditions in which lasso variable selection is not guaranteed to be consistent and hence not satisfying the oracle property, [Zou, 2006] modified the L_1 lasso penalty $\lambda \|\boldsymbol{\theta}\|_1$ to incorporate parameter specific weights \mathbf{w} yielding penalty $\lambda \|\mathbf{w}\boldsymbol{\theta}\|_1$. The penalized likelihood for the adaptive graphical lasso is then:

$$\log(\det \mathbf{\Omega}) - \text{tr} \left(\frac{\mathbf{S}}{n} \mathbf{\Omega} \right) - \lambda \sum_{1 \leq i \leq p} \sum_{1 \leq j \leq p} w_{ij} |\omega_{ij}|. \quad (33)$$

In this likelihood, the weights that contribute to the penalization are $w_{ij} = |\hat{\omega}_{ij}|^\alpha$ for a fixed $\alpha > 0$ and a consistent estimate of the concentration matrix with entries $\hat{\omega}_{ij}$.

The Bayesian Graphical Lasso

A Bayesian interpretation of the regular graphical lasso [Friedman et al., 2007] has been shown previously [Wang, 2012]. Given the following hierarchical model:

$$p(\mathbf{x}_i|\mathbf{\Omega}) = \mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{-1}) \quad \text{for } i = 1, 2, \dots, n \quad (34)$$

$$p(\mathbf{\Omega}|\lambda) = \frac{1}{C} \prod_{i<j} \text{DE}(\omega_{ij}|\lambda) \prod_{i=1}^p \text{EXP}(\omega_{ii}|\lambda/2) \cdot \mathbf{1}_{\mathbf{\Omega} \in M^+}, \quad (35)$$

it has been demonstrated the mode of the posterior distribution of $\mathbf{\Omega}$ is the graphical lasso estimate given penalty parameter $\rho = \lambda/n$. In this model, the prior distribution of the off-diagonal entries of the concentration matrix follow a double exponential distribution centered at zero with scale parameter λ , while the prior distribution of the diagonal entries of the concentration matrix is exponential with scale parameter $\lambda/2$. [Wang, 2012] noted that the hierarchical model in (35) can be represented as a scale mixture of normal distributions [Andrews and Mallows, 1974, West, 1987] leading to the following prior distribution:

$$p(\boldsymbol{\omega}|\boldsymbol{\tau}, \lambda) = \frac{1}{C_{\boldsymbol{\tau}}} \prod_{i<j} \left[\frac{1}{\sqrt{2\pi\tau_{ij}}} \exp\left(-\frac{\omega_{ij}^2}{2\tau_{ij}}\right) \right] \prod_{i=1}^p \left[\frac{\lambda}{2} \exp\left(-\frac{\lambda}{2}\omega_{ii}\right) \right] \cdot \mathbf{1}_{\mathbf{\Omega} \in M^+} \quad (36)$$

[Wang, 2012] exploited this representation to develop a block Gibbs sampler for simulating the posterior distribution. This sampler is predicated on identical partitioning of the concentration matrix $\mathbf{\Omega}$, products matrix \mathbf{S} , and latent scale parameters matrix $\boldsymbol{\mathcal{T}}$:

$$\begin{bmatrix} \mathbf{\Omega}_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{12}^T & \omega_{22} \end{bmatrix}, \begin{bmatrix} \mathbf{S}_{11} & \mathbf{s}_{12} \\ \mathbf{s}_{12}^T & s_{22} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\mathcal{T}}_{11} & \boldsymbol{\tau}_{12} \\ \boldsymbol{\tau}_{12}^T & \tau_{22} \end{bmatrix}. \quad (37)$$

The Gibbs sampler then samples from the conditional distribution of the last column, $(\boldsymbol{\omega}_{12}, \omega_{22})^T$ of $\mathbf{\Omega}$:

$$p(\boldsymbol{\omega}_{12}, \omega_{22}|\mathbf{\Omega}_{11}, \boldsymbol{\mathcal{T}}, \mathbf{X}, \lambda) \propto (\omega_{22} - \boldsymbol{\omega}_{12}^T \mathbf{\Omega}_{11}^{-1} \boldsymbol{\omega}_{12})^{n/2} \exp\left(-\frac{1}{2} [\boldsymbol{\omega}_{12}^T \mathbf{D}_{\boldsymbol{\tau}} \boldsymbol{\omega}_{12} + 2\mathbf{s}_{12}^T \boldsymbol{\omega}_{12} + (s_{22} + \lambda)\omega_{22}]\right) \quad (38)$$

In the same work, [Wang, 2012] proposed a Bayesian approach to the adaptive graphical lasso developed by [Fan et al., 2009]. The hierarchical model proposed

by for the adaptive lasso is:

$$p(\mathbf{x}_i|\mathbf{\Omega}) = \mathcal{N}(\mathbf{0}, \mathbf{\Omega}^{-1}) \quad \text{for } i = 1, 2, \dots, n \quad (39)$$

$$p(\mathbf{\Omega}|\{\lambda_{ij}\}_{i \leq j}) = C^{-1} \prod_{i < j} \text{DE}(\omega_{ij}|\lambda_{ij}) \prod_{i=1}^p \text{EXP}(\omega_{ii}|\lambda_{ii}/2) \cdot \mathbf{1}_{\mathbf{\Omega} \in M^+} \quad (40)$$

$$p(\{\lambda_{ij}\}_{i < j} | \{\lambda_{ii}\}_{i=1}^p) \propto C_{\{\lambda_{ij}\}_{i \leq j}} \prod_{i < j} \text{GA}(r, s) \quad (41)$$

As with the frequentist adaptive graphical lasso, the Bayesian adaptive graphical lasso proposed by [Wang, 2012] incorporates differential shrinkage of entries ω_{ij} according to an estimate $\hat{\omega}_{ij}$. However, as opposed to a fixed relationship between the norm of the current estimate $\hat{\omega}_{ij}$ and the size of the shrinkage parameter λ uncertainty about λ is incorporated by assuming a gamma distribution for λ_{ij} . Conditional on the concentration matrix, λ_{ij} are distributed as:

$$\lambda_{ij}|\mathbf{\Omega} \sim \text{GA}(1 + r, |\omega_{ij}| + s) \quad (42)$$

Given that the Bayesian adaptive graphical lasso allows for differential shrinkage for each ω_{ij} via shrinkage parameter λ_{ij} drawn from a non-informative distribution, it is natural to suppose that adaptive penalization would allow for the incorporation of apriori knowledge about the conditional relationship between X_i and X_j . [Peterson et al., 2013] propose that rather than fixing a non-informative prior gamma distribution scale s parameter, this parameter could capture prior belief regarding the conditional relationship between X_i and X_j . They assume that if an a priori defined unweighted graph describes the biological relationship between random variables X_i , then the pairwise distance between vertices X_i and X_j , that is $d(X_i, X_j)$, could be used as a hyperparameter for s_{ij} . In their specific case they set:

$$s_{ij} = \begin{cases} d_{ij}^{-1} \cdot 10^{-6+c} & \text{if } d_{ij} < \infty \\ 10^{-6} & \text{if } d_{ij} = \infty \end{cases}, \quad (43)$$

where $d_{ij} = d(X_i, X_j)$ was defined as the minimum path distance determined via breadth-first search, and $c \in (0, 6)$ is a positive constant. This specification of the

gamma scale parameter has the effect of shifting the mean of the prior distribution for the penalty parameter λ_{ij} towards zero when vertices X_i and X_j are close with respect to the a priori graph.

An important contrast between the frequentist and Bayesian graphical lasso is that given a continuous prior distribution, the Bayesian lasso does not possess an innate edge selection ability. Specifically, as the Bayesian graphical lasso given a continuous prior places positive probability on the posterior entries of the concentration matrix, ω_{ij} , a graph selected from the posterior distribution of $\mathbf{\Omega}$ will be fully connected, that is an edge will exist between each pair of vertices of the graph G . Heuristics for conducting edge selection are discussed in both [Wang, 2012] and [Peterson et al., 2013]. The selection operator discussed in [Wang, 2012] was based on a heuristic appeal to the argument made in [Carvalho et al., 2010]. [Carvalho et al., 2010] discusses a discrete mixture model:

$$\theta_i \sim (1 - p)\delta_0 + pg(\theta_i) \quad (44)$$

where δ_0 is the Dirac distribution and p is the prior mixing probability. Under this model the posterior mean of θ_i is then:

$$E(\theta_i|\mathbf{X}) = P(\theta_i \neq 0|\mathbf{X}) E_g(\theta_i|\mathbf{X}, \theta_i \neq 0). \quad (45)$$

[Wang, 2012] then claims that using the Wishart conjugate prior for the concentration matrix as the distribution g , the same factorization can be used to substantiate the claim that $\omega \neq 0$ if and only if:

$$\tilde{\pi}_{ij} = \frac{\tilde{\rho}_{ij}}{E_g(\rho_{ij}|\mathbf{X})} > 0.5 \quad (46)$$

where $\tilde{\rho}_{ij}$ is the posterior mean estimator of ρ_{ij} and $\tilde{\pi}_{ij}$ is the amount of shrinkage enforced by the graphical lasso prior. In contrast, [Peterson et al., 2013] proposes that a rejection region approach could be employed using posterior credible intervals. Specifically, they suppose edge selection should include an edge between X_i and X_j if and only if the 95% posterior credible interval for ω_{ij} does not include 0. The authors then use a thresholding approach in the application discussed, omitting edges between vertices if $|\omega_{ij}| \leq 0.1$.

3 Bayesian inference given G -Wishart priors

An alternative Bayesian treatment focuses on [Dawid and Lauritzen, 1993, Roverato, 2002, Atay-Kayis and Massam, 2005, Dobra and Lenkoski, 2011, Dobra et al., 2011, Wang, 2012, Cheng and Lenkoski, 2012] the estimation of GGMs via conjugate inference for $\mathbf{\Omega}$ given that $\mathbf{\Omega}$ is faithful to a fixed graph G , as opposed to estimation with shrinkage.

Early work [Dawid and Lauritzen, 1993] focused on the case of decomposable GGMs. A decomposition of a graph G is a pair subsets (R, S) , $R \subseteq V$, $S \subseteq V$ such that $V = R \cup S$, the graph defined by $R \cap S$ is complete, and $R \cap S$ is the separator of R and S —that is any path between R and S must go through $R \cap S$. A graph G is a decomposable graph if it complete or there exists a proper decomposition (R, S) of G . Equivalently, a graph is decomposable if each prime component of the graph is complete [Roverato, 2002]. [Dawid and Lauritzen, 1993] first describe the Hyper Inverse Wishart distribution for cliques $C \in \mathcal{C}$ where $\{C_1, C_2, \dots, C_k\}$ is a perfectly ordered set of cliques in the decomposable graph G and S .

$$f_G(\mathbf{\Sigma}^\mathcal{V} | \delta, \mathbf{D}^\mathcal{V}) = \frac{\prod_{j=1}^k f_{C_j}(\mathbf{\Sigma}_{C_j C_j} | \delta, \mathbf{D}_{C_j C_j})}{\prod_{j=2}^k f_{S_j}(\mathbf{\Sigma}_{S_j S_j} | \delta, \mathbf{D}_{S_j S_j})} \quad (47)$$

[Roverato, 2002] advanced the theory substantially by proposing methodology for inference given arbitrary graphs (including non-decomposable graphs). [Roverato, 2002] shows that if $G = (V, E)$ is an arbitrary graph then $\mathbf{\Sigma} \sim HIW_G$

This G -Wishart distribution has the following form:

$$p(\mathbf{\Omega} | G) = I_G(b, D)^{-1} |\mathbf{\Omega}|^{(b-2)/2} \exp \left[-\frac{1}{2} \text{tr}(D\mathbf{\Omega}) \right] 1_{\{\mathbf{\Omega} \in M^+(G)\}} \quad (48)$$

where $I_G(b, D)$ is a normalization constant that ensures that $p(\mathbf{\Omega} | G)$ is a proper distribution:

$$I_G(b, D) = \int |\mathbf{\Omega}|^{(b-2)/2} \exp \left[-\frac{1}{2} \text{tr}(D\mathbf{\Omega}) \right] 1_{\{\mathbf{\Omega} \in M^+(G)\}} d\mathbf{\Omega} \quad (49)$$

CHAPTER V

HIGHER-ORDER INFERENCE IN METABOLOMICS

Higher-order inference in metabolomics corresponds to the testing of hypotheses that are more complex than single metabolite hypothesis tests. For example, an experimenter might want to determine if a specific biological process (e.g. cholesterol biosynthesis) differs between two or more sample phenotypes. The central question of higher order inference is how to formally test a hypothesis such as:

- H_0 : cholesterol biosynthesis does not differ between two (or more) phenotypes
- H_A : cholesterol biosynthesis does differ between two (or more) phenotypes

Or formulated as an equivalence test:

- H_0 : cholesterol biosynthesis is the same between two (or more) phenotypes
- H_A : cholesterol biosynthesis is not the same between two (or more) phenotypes

Other forms of higher-order inference are not explicitly hypothesis driven. Many metabolomics experiments are motivated to address open-ended questions such as “What are the metabolic consequences of a specific disease or disease state?” or “What is the impact on cellular metabolism of the up or down-regulation of a specific gene”. An example of the first type can be found in Trainor et al. [Trainor et al., 2017], who sought to determine the metabolic consequence of thrombotic MI using blood plasma across a disease state transition. An

example of the second type can be found in Carlisle et al. [Carlisle et al., 2016]. Previous methods to make such higher order inferences generally fit one of the following categories: (1) statistical tests for pathway enrichment, (2) metabolite set enrichment analyses, or (3) correlation based analyses.

1 Statistical tests for pathway enrichment

Pathway enrichment analyses seek to determine if specific pathways are represented more than expected by chance alone in a list of differentially abundant metabolites (or other biomolecules such as mRNA transcripts). These analyses take the results of univariate statistical tests that first identify sets of metabolite features for which significant evidence of differences between experimental conditions or phenotypes are observed. After identifying interesting metabolite features, these sets can be tested for enrichment of specific metabolic pathways or biological processes greater than that expected by chance [Goeman and Buhlmann, 2007, Xia and Wishart, 2010a].

2 Exact tests of pathway enrichment

Table 1. Example 2×2 table for two random variables X and Y .

	X	
Y	$n_{x_1y_1}$	$n_{x_2y_1}$
	$n_{x_1y_2}$	$n_{x_2y_2}$

To discuss the use of Fisher’s exact for determining pathway over-representation we first introduce the computation of the probabilities of 2×2 contingency tables. The probability of a specific 2×2 contingency table (as shown in Table 1) can be computed using the hypergeometric distribution as follows [Agresti, 2013]:

$$p = \frac{\binom{n_{x_1y_1} + n_{x_2y_1}}{n_{x_1y_1}} \binom{n_{x_1y_2} + n_{x_2y_2}}{n_{x_1y_2}}}{\binom{n}{n_{x_1y_1} + n_{x_1y_2}}}, \tag{50}$$

where $n = n_{x_1y_1} + n_{x_2y_1} + n_{x_1y_2} + n_{x_2y_2}$. Fisher’s exact test is a test of the independence of rows and columns. In other words the test evaluates whether the

random variables X and Y are independent. In order to determine the statistical significance of the test, the probability of the observed table can be compared to the probability of tables with the same marginal totals. In the case of evaluating whether a specific metabolic pathway is over-represented within a list of differentially abundant metabolites, we can represent pathway membership as the first random variable and differential abundances indicator as the second random variable.

Table 2. Example Fisher’s exact test for determining if a pathway is over-represented in a list of significant metabolites.

	Differentially abundant (DA)	Not DA	Total
Pathway A	10	20	30
Not in Pathway A	40	350	390
Total	50	370	420

A hypothetical exact test for determining if a pathway is over-represented in a list of significant metabolites is presented in Table 2. In this example, it is assumed that a total of 420 metabolites were observed. Of these metabolites, 30 were in Pathway A, while 390 were not in Pathway A. Of the metabolites in Pathway A, 10 were observed to be differentially abundant, while of the metabolites not in Pathway A, 40 were observed to be differentially abundant. A two-sided Fisher’s test given the observed contingency table yields a p-value of 0.001, indicating that there is evidence of an association between the differential abundance indicator and the pathway membership indicator.

While metabolic pathway over-representation analyses are commonplace in metabolomics, there are significant sources of bias that dramatically reduce the utility of such approaches. A graphical representation of this issue is presented in Figure 4.

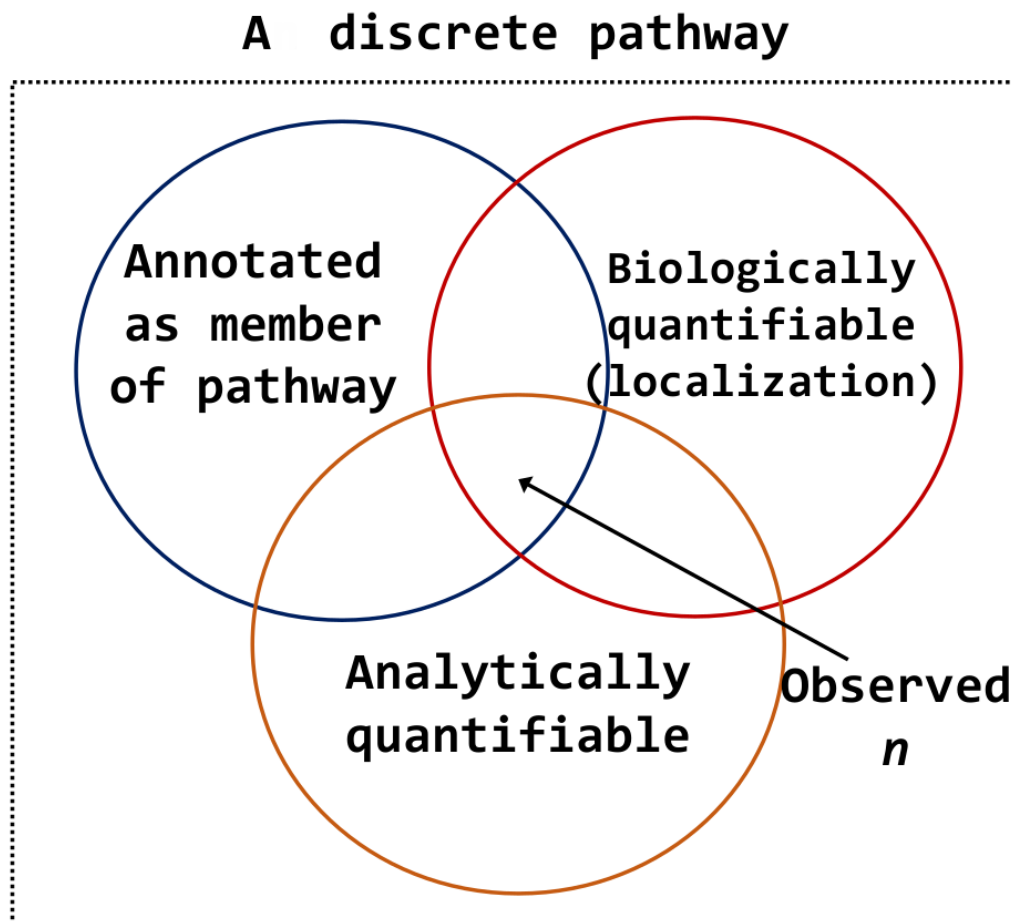


Figure 4. Graphical representation of the source of bias in discrete pathway enrichment analyses. A single metabolic pathway is shown as a bordered rectangle. The blue circle represents cases in which a metabolite that is in a metabolic pathway is annotated as being a member of that pathway in pathway databases. The red circle represents cases in which a metabolite that is a member of a metabolic pathway is quantifiable given the sample medium utilized in the analysis. For example, a metabolite localized to the mitochondria may not be detectable in plasma samples. The orange circle represents cases in which a metabolite is detectable from a sample given the analytical technology and platform utilized. For example, a metabolite may be not detectable given the sensitivity of NMR.

3 Metabolite set enrichment analysis

Metabolite set enrichment analysis (MSEA)[Xia and Wishart, 2010b] is a generalization of the concept of Gene Set Enrichment Analysis (GSEA) [Mootha et al.,

2003, Subramanian et al., 2005] to metabolomics data. The GSEA methodology was developed to address the arbitrary nature of using p -value thresholds in discrete pathway enrichment tests in transcriptomics [Subramanian et al., 2005]. As opposed to focusing on binary “significant”/“non-significant” cutoffs for conducting discrete statistical tests, GSEA takes a ranked list of genes as an input. This ranking may be based on statistical significance of differential expression, or correlation with a specific phenotypic attribute of interest. To determine the degree of enrichment of a specific pathway, a score is computed by descending the ranked list of genes and incrementing the score up or down based on an indicator of pathway membership and by a weight that is a monotonic function of the ordinal position in the rank list. Under the null hypothesis that a specific pathway is not overrepresented early or late in the ranked list, the running score will exhibit a random walk behavior about zero.

In a similar vein a global score test was developed for evaluating the strength of the association between groups of genes and a nominal, discrete, or continuous variable of interest [Goeman et al., 2003]. The statistic that underlies this test is a multiple of the squared covariance between gene expression and the variable of interest and is evaluated within pathways. This “set enrichment” test was utilized by Xia and Wishart [Xia and Wishart, 2010b] in the development of MSEA. In developing MSEA as a tool, extensions to other pathway / ontological databases were made, especially to the Small Molecule Pathway Database (SMPDB) [Frolkis et al., 2010]. However, in spite of these extensions, MSEA suffers from the same sources of biases that have been described previously, and which are depicted in Figure 4.

4 Chemical Similarity Enrichment Analysis approach

A promising alternative to pathway analyses discussed in [Barupal and Fiehn, 2017] is to use structural similarity and chemical ontology as *a priori* knowledge to generate study-specific metabolite sets for contextualizing empirical results.

Barupal, et al. frame the need for the method they developed with statistics on the coverage of detected metabolites in metabolism databases. For example in a study of metabolic dysregulation in diabetic mice, they observed that only 135 of 385 metabolites that were detected by mass spectrometry and for which the chemical structure could be elucidated were members of the KEGG database [Kanehisa et al., 2016]. The extraordinarily low coverage rate suggested to the authors that they should utilize chemical ontology for conducting enrichment analyses as opposed to metabolic pathways. Another justification they cite for this is that chemical annotations are unique, while metabolites may be members of multiple pathways. Briefly, Chemical Similarity Enrichment Analysis (ChemRICH) uses Medical Subject Headings (MeSH) chemical ontologies. The first step is to map the multiple possible metabolite identifiers to MeSH chemical ontologies. If a metabolite can not be mapped to an ontological class, the Tanimoto coefficient for structural similarity using PubChem 881 bit substructure fingerprints is utilized to join metabolites to existing ontological classes (based on similarity) or to make new ones. Once these classes have been defined (existing MeSH chemical classes or new ones based in structural similarity clustering), enrichment analysis is conducted using a similar approach to GSEA.

5 The need for interactomes in metabolomics

Untargeted profiling of the metabolome of an organism provides a view into the small molecule determinants of phenotype. While the genome of an organism may be conceptualized as a blueprint for the composition and organization of an organism that is largely immutable (barring epigenetic modifications, DNA damage, and genetic mutations) [Gao et al., 2015, Keating and El-Osta, 2015, Martincorena and Campbell, 2015], the metabolome of an organism is dynamic and variable [Dallmann et al., 2012, Krycer et al., 2017]. Sources of variation within the metabolome of a single organism include tissue-, cell-, and organelle-specific localization of metabolic processes [Shlomi et al., 2008, Voet et al., 2013]; envi-

ronmental exposures [Southam et al., 2014]; and host-microbe interactions and metabolite exchange [Moriya et al., 2017]. While the generation of reference human genomes has facilitated the interrogation of gene-phenome associations (including human disease associations), the intrinsic variability and dynamic nature of the metabolome of an organism likely precludes the generation of such a reference model. While a single reference model of the metabolome of an organism may not be sensible, significant efforts such as the HUSERMET project [Dunn et al., 2014] have been undertaken to quantify the repertoire of metabolites in specific biofluids for examining metabolite-metabolite and metabolite-phenotype associations. In order to make systems-level comparisons of the differences in the metabolome across phenotypes, models of the conditional relationships between metabolites are necessary. By the determination of sample media and analytical platform-specific probabilistic interaction models, henceforth called “interactomes”, systems-level comparisons of phenotypes that can be made. A specific use case for such an interactome model is the generation of a plasma interactome for stable heart disease.

6 A specific biomedical science use case

Heart disease is the most prevalent cause of death globally [Benjamin et al., 2017]. As a disease, heart disease does not represent a uniform condition, but rather a collection of diseases of varying etiologies [Kasper, 2015]. Of particular interest in the study of coronary artery disease (CAD) is the elucidation of the precipitants of acute disease events such as myocardial infarction [Arbab-Zadeh and Fuster, 2015] or unstable angina, metabolic pathways associated with disease phenotypes [Fan et al., 2016], and determining the metabolic consequences of acute events [Trainor et al., 2017]. To date, an interactome describing the conditional relationships between blood plasma metabolites in humans with heart disease does not exist. If such a reference model were determined, it would facilitate making systems-level inferences regarding metabolic perturbations that accompany acute disease events

such as unstable angina or acute myocardial infarction (MI).

7 The challenge of inferring high-dimensional interactomes

A significant challenge in evaluating the relationships between metabolites in an untargeted metabolomics experiment is that the dimension of metabolites may be greater than the number of samples. Even given a relatively high ratio of samples to metabolites detected, in the evaluation of pairwise conditional relationships between metabolites, the number of parameters to be estimated can be prohibitive. For example, if $p = 500$ metabolites are detected, an evaluation of all pairwise conditional relationships would require the simultaneous estimation of 124,750 parameters. The use of regularization is a well-established approach for guaranteeing the existence of Gaussian Graphical Model parameters, amenable to the case that the sample size n is less than p [Banerjee et al., 2008, Fan et al., 2009, Friedman et al., 2007, Meinshausen and Bhlmann, 2006, Yuan and Lin, 2007].

Penalized estimation of GGM parameters provides a natural mechanism for integrating a priori knowledge regarding the molecular structure of metabolites with experimental metabolomics data. The integration of empirical data and scientific knowledge regarding metabolism is common in metabolomics studies. The current work is of a similar paradigm and predicated on the assumption that the individual biochemical reactions that result in statistical dependence between metabolic intermediates also generate statistical dependence in structural similarity between the same intermediates. In our application, structural similarity is determined by an approach that considers overlap in shared local structure between metabolites. Rather than considering fixed sets of metabolites such as pathways, sets, or modules and subsequently quantifying enrichment of these sets in empirical results, we consider a priori knowledge of the relationships between metabolites as probabilistic statements about the relatedness of compounds. Thus, the a priori scientific knowledge is used to generate prior probability distributions that influence GGM

model selection, so that posterior inference probabilistically combines empirical data and prior scientific knowledge to yield an updated model of the probabilistic interactions between metabolites. In the present work, we introduce a methodology for using molecular structure similarity to generate prior distributions that control the degree of penalization in parameter estimation for learning a GGM metabolite interactome from metabolomics data.

CHAPTER VI

INFERRING METABOLITE INTERACTOMES VIA BAYESIAN GRAPHICAL MODEL SELECTION UTILIZING MOLECULAR STRUCTURE INFORMATIVE PRIORS

1 Desired inference and sketch of evaluation

As stated previously, a reference model describing the probabilistic interactions between metabolites that is organism-specific, sample-media specific, and specific to the metabolites that have been detected is warranted for making higher order inference regarding systems of metabolites that change between phenotypes. We propose constructing such interactomes utilizing a Bayesian approach for estimating Gaussian Graphical Models (GGMs) that utilizes prior information about the molecular structure similarity between pairs of metabolites. In the first part of the chapter, we discuss our proposed methodology. We then evaluate the effectiveness of the methodology and in the following chapter utilize the methodology to construct a stable heart disease plasma metabolite interactome. We evaluated the methodology using simulation studies that follow two different schema. Under the first schema, autoregressive processes were simulated for representing linear biological processes in which the correlation between simulated metabolites decreased in tandem with decreasing structural similarity. Under the second schema, random covariance matrices corresponding to structural similarity were simulated using the C-vines method [Lewandowski et al., 2009]. In this case, a hierarchical model was used, in which structural similarity was simulated first, followed by abundance distributions in which the correlation between abundances increased with structural simi-

larity. Given both schema, we evaluated the ability of the proposed method to recover the true pairwise conditional correlations structures that were specified in advance. We then evaluate our methodology for generating graphical models for representing the relationships between metabolites detected and quantified from human plasma, and specifically for the development of a reference model for stable coronary artery disease.

2 Molecular structure priors and the BGL

We propose that to incorporate prior knowledge regarding the relatedness of compounds, the scale hyperparameter can be linked to structural similarity, that is by specifying the prior distribution $\lambda_{ij} \sim \text{Gamma}(r, s_{ij})$ where s_{ij} is a measure of structural similarity between compound i and compound j . The conditional expected value of each λ_{ij} is then: $E(\lambda_{ij}|\Omega) = (1 + r)/(|\omega_{ij}| + s_{ij})$.

To generate informative shrinkage priors for the adaptive Bayesian graphical Lasso, we utilized a local structure similarity metric. This metric was adapted from the previously described Chemically Aware Substructure Search (CASS) algorithm [Mitchell et al., 2014]. In this adaptation, the structural similarity between any two chemical structures (A and B) was estimated using strings representing local chemical structure (referred to as the atoms color) centered at every atom in the two structures. The color of every atom was constructed as follows. First, for every bonded atom, its element type and the order of the bond connecting it to the center atom are joined to form a component string that is added to a list of components. For example, if the center atom has a double bonded oxygen, this would contribute a "O2" component to the component list. Every component represents a portion of the local bonded structure at the center atom. Second, the components strings are then sorted alphanumerically and concatenated to produce a description of the bonded structure one bond away from the center atom. Finally, to the front of this string, the element type of the center atom is then added to yield the

atoms color. Each color uniquely maps to a single locally bonded structure (e.g. the “CC1O1O2” coloring represents a carbon of a carboxylate). Since the component list was first sorted alphanumerically, this color is consistent for all identical local structures regardless of how they are ordered in their representation. Each chemical structure can be represented as the list of its constituent atoms colors and these lists of colors can be compared to determine structural similarity. To determine structural similarity between compounds using the color string representations, the Tanimoto coefficient between pairs of compounds was computed, which is defined as [Chen and Reynolds, 2002]:

$$s(A, B) = \frac{\sum_{i=1}^m \min(n_i(A), n_i(B))}{\sum_{i=1}^m n_i(A) + \sum_{i=1}^m n_i(B) - \sum_{i=1}^m \min(n_i(A), n_i(B))} \quad (51)$$

where $n_i(A)$ represents the count of unique atom pairs indexed by $i = 1, 2, \dots, m$ for molecule A . The Tanimoto dissimilarity is then $d(A, B) = 1 - s(A, B)$.

After determining the Tanimoto dissimilarity between each pair of metabolites, the gamma hyperprior distribution for the shrinkage parameter λ can be determined by linking the gamma distribution shape to the dissimilarity, that is by setting $s_{ij} = f(1 - d(i, j))$ where i, j index metabolites and $f(x)$ is a monotonic function. The conditional distribution of the shrinkage parameter is then $\lambda_{ij} | \Omega \sim \text{Gamma}(1 + r, |\omega_{ij}| + s_{ij})$. A plot of the relationship between the structural similarity of two hypothetical metabolites and the expected value of the shrinkage parameter is shown in Fig. 5.

To determine a graphical model (or a set of models of high probability) given structural priors samples may be drawn from the posterior distribution of $p(\Omega | X)$, using a Gibbs sampler as discussed in an earlier chapter and similar to that introduced by Wang [Wang, 2012]. Using a scale mixture of normals representation [West, 1987] and introducing the latent scale parameter matrix

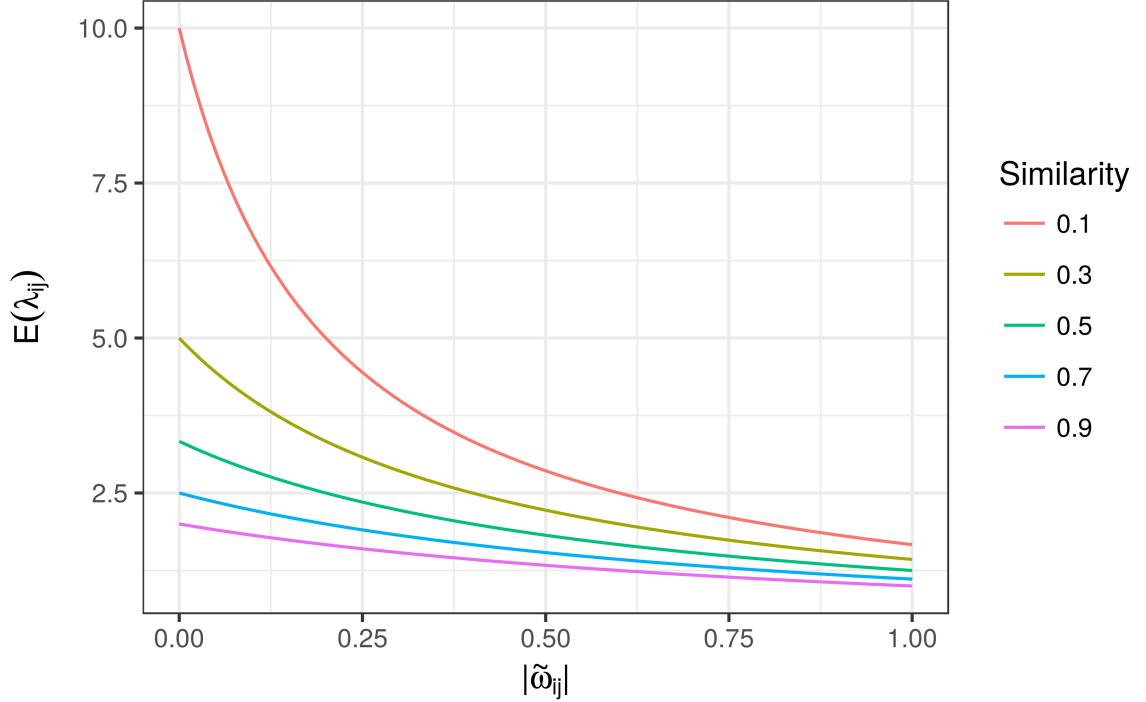


Figure 5. Theoretical relationship between the structural similarity of two metabolites and the expected value of the shrinkage parameter λ_{ij} . On the vertical axis the expected value of the shrinkage parameter is shown. On the horizontal axis the magnitude of the current estimate of a concentration matrix parameter, $|\tilde{\omega}_{ij}|$, is shown. Finally, lines are colored by theoretical molecular structure similarity between two metabolites.

τ , the unnormalized posterior distribution can be written as [Wang, 2012]:

$$p(\mathbf{\Omega}, \boldsymbol{\tau} | \mathbf{X}, \mathbf{\Lambda}) \propto |\mathbf{\Omega}|^{n/2} \exp\left(-\text{tr}\left(\frac{1}{2}\mathbf{S}\mathbf{\Omega}\right)\right) \prod_{i < j} \left(\tau_{ij}^{-1/2} \exp(-\omega_{ij}^2/(2\tau_{ij})) \exp\left(-\frac{1}{2}\lambda_{ij}^2\tau_{ij}\right)\right) \prod_{i=1}^p \exp(-\frac{1}{2}\lambda_{ij}\omega_{ij}) \mathbf{1}_{\mathbf{\Omega} \in M^+} \quad (52)$$

The block Gibbs sampler cycles through column-wise partitions of $\mathbf{\Omega}$, drawing from the conditional distribution of a single column of the matrix $\mathbf{\Omega}$, conditioned on the current values of the remaining columns.

3 R package development

We developed an R package, *BayesianGLasso*, for implementing this and other samplers for the Bayesian Graphical Lasso. The underlying sampler was written in C++ using *Rcpp* [Eddelbuettel and François, 2011, Eddelbuettel, 2013] and *RcppArmadillo* [Eddelbuettel and Sanderson, 2014] to make use of the *Armadillo* [Sanderson and Curtin, 2016] linear algebra library. The sampler was written in C++ as loops often execute faster in C++ than in R. In addition to providing Gibbs sampling methods the R package developed by our group includes classes for storing the Markov chains generated by the sampler along with relevant parameters and hyperparameters, and methods for conducting statistical inference over the simulated posterior distributions.

Rcpp [Eddelbuettel and François, 2011, Eddelbuettel, 2013] facilitates integration of C++ through an approachable API that also has many “syntactic sugar” functions. In order to implement functions utilizing the C++ language, a method for mapping R data structures to C data structures is required (although the underlying R data structures at the low level are C data structures). *Rcpp* provides functions for mapping R data structures to C++ types and C++ templates for returning the data structures utilizing in C++ to R [Eddelbuettel and François, 2011]. Many of the R to C++ type conversions are also done implicitly when passing data between R and C++. “Syntactic sugar” functions are C++ templated functions that can be used within C++ code, but are similar to R functions [Eddelbuettel, 2013]. These functions include logical operators (including logical operators for vector objects), many mathematical functions, as well as random number generators. The random number generation C++ templated functions are particularly useful in the context of statistical programming and for generating methods for sampling for posterior distributions.

The purpose of the *Armadillo* library is to provide linear algebra routines for usage in C++ [Sanderson and Curtin, 2016]. Syntactically, matrix algebra

and linear algebra functions and operations using *Armadillo* are similar to the Matlab language. The *Armadillo* library utilizes routines from BLAS (Basic Linear Algebra Subprograms) and LAPACK (Linear Algebra Package) libraries as building blocks. Consequently, the efficiency of Armadillo routines depends on the version of BLAS and LAPACK that are compiled against. Using the Armadillo library within C++ code written so that the compiled code can be sourced in R can be done utilizing *RcppArmadillo* [Eddelbuettel and Sanderson, 2014]. This R package provides bindings for the Armadillo and also provided the ability to map R data structures to Armadillo types.

In the process of building the R package *BayesianGLasso*, many of the functions of the package *devtools* [Wickham et al., 2018] were utilized. For example, functions from this package were used for creating the package skeleton and directories, for compiling and testing source code, and linking and testing the installation and loading of the package on Unix and Windows platforms. The package *roxygen2* [Wickham et al., 2017] was utilized to facilitate “online” package documentation. “Online” package documentation allows the developer to use tags within R source code to generate R markdown documentation files.

4 Efficacy analysis via simulation studies

To evaluate the efficacy of the proposed method, we employed simulation studies. We sought to evaluate the relative performance of the adaptive Bayesian Graphical Lasso (BGL) using informative priors versus (1) the adaptive Bayesian Graphical Lasso using non-informative priors, and (2) the Bayesian Graphical Lasso (non-adaptive). For the informative prior case, we further manipulated the degree to which the priors were accurate relative to the partial correlation structure utilized to generate the data. We evaluated the methods by simulating both simple partial correlation structures as well as more complex structures utilizing two simulation schemas. Under the first schema, a simple autoregressive (AR) process of order 1 was simulated for representing a

linear biological process with decreasing structural similarity with increasing process distance. Simulated structural similarity was taken to be deterministically known, that is a structural similarity matrix was defined as: $\Sigma = [\sigma_{ij}]$ where $\sigma_{ij} = \rho^{|i-j|}$. To simulate metabolite abundances, a random matrix was sampled from the multivariate normal distribution $N(\mathbf{0}, \Sigma)$. In the accurate informative prior case, the shrinkage hyperprior s_{ij} was defined as $s_{ij} = \omega_{ij}^{-1}$, where ω_{ij}^{-1} are the elements of $\Omega = \Sigma^{-1}$. After generating simulated datasets, the adaptive BGL (with informative and non-informative priors) as well as the non-adaptive BGL were utilized for estimating the concentration matrix and corresponding graph topology. Given the simple dependence structure in the AR(1) case, a measure of ground truth was available as the existence of edges between simulated metabolites was known a priori. We evaluated the sensitivity, specificity, and F_1 measure of each method for detecting the presence of edges by utilizing the magnitude of the estimated concentration matrix entries: $|\omega_{ij}|$. In addition, we report the area under the receiver operating characteristic curve for assessing each technique, which considers the range of possible fixed cutoff values of $|\omega_{ij}|$ for estimating the presence or absence of edges. While each technique draws shrinkage parameters from a Gamma distribution, the shape and scale of each distribution depends both on empirical data and hyperparameters. We conducted shape and scale hyperparameter optimization separately for each technique via a grid search over simulated datasets prior to the evaluation of performance.

The simulation studies were conducted using Amazon Elastic Compute Cloud (EC2) Linux instances. The EC2 instances were comprised of multiple Intel Xeon 3.0 GHz processors running Ubuntu 16.04 (Xenial) 64 bit server edition. The Intel Math Kernel library (MKL) was installed on each instance providing BLAS and LAPACK libraries that were optimized for the processor architecture. Significant performance improvement was observed when the C++ files were compiled against the BLAS and LAPACK libraries provided

by Intel MKL relative to multi-threaded openBLAS and the standard BLAS library that was included by default with the Ubuntu distribution.

5 Results of the simulation studies

Results from the simulation studies given an autoregressive covariance structure are shown in Table 1 and Figures 6-7. In Figure 6 a graphical model representation of the underlying covariance structure is shown along with the graphical model representations of the sample covariance matrix and the concentration matrices estimated by the multiple techniques evaluated in this study. These figures were generated from a randomly sampled simulation study.

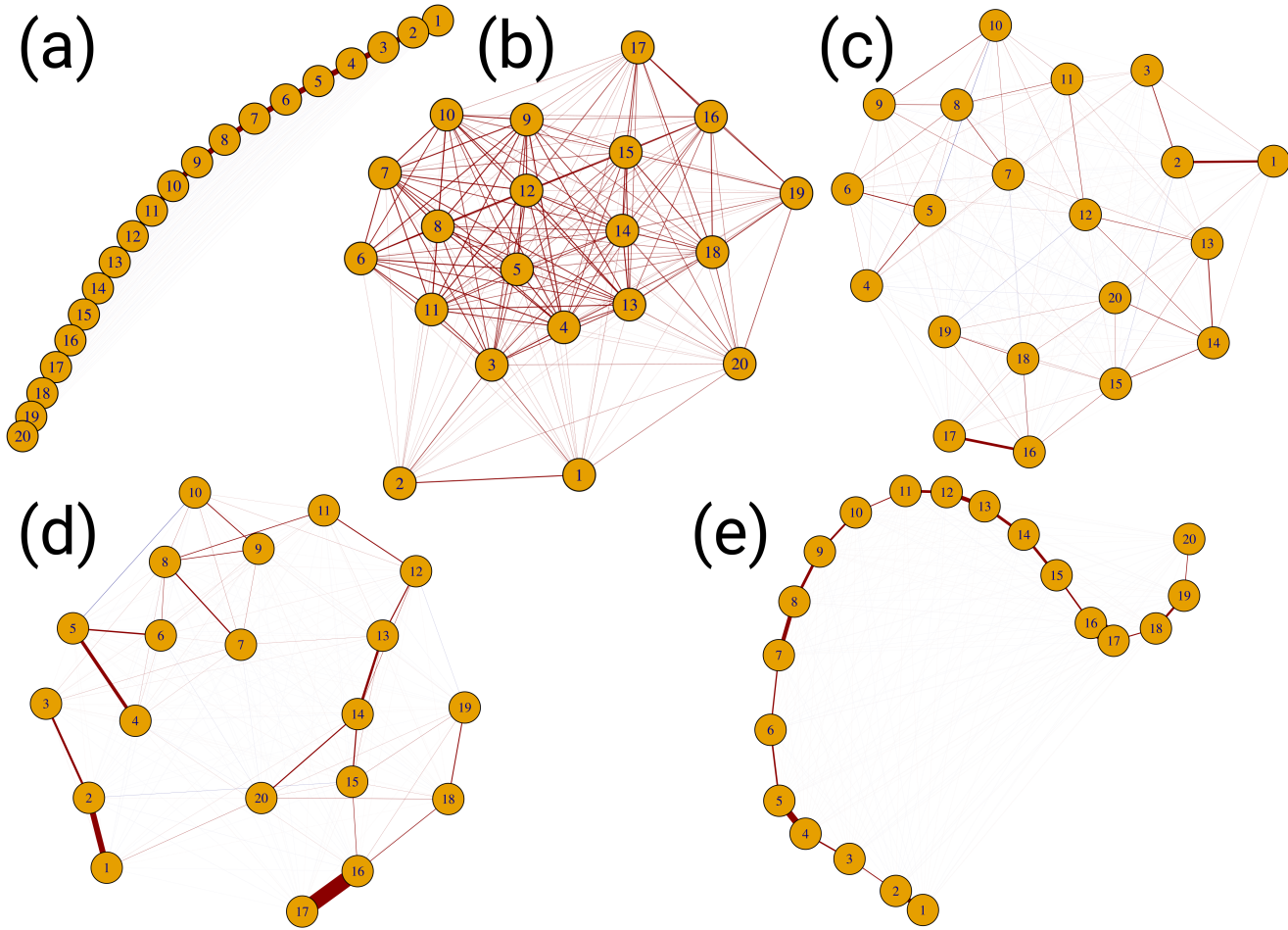


Figure 6. See next page for caption

Figure 6. True and estimated concentration matrix graphs for a randomly selected AR(1) simulation study with $p = 20$ simulated random variates (metabolites) and a simulated sample size of $n = 10$. Graphs represent the: (a) true concentration structure given an AR(1) covariance structure with $\rho = 0.95$, (b) sample covariance matrix, (c) concentration matrix estimated by the non-adaptive BGL, (d) concentration matrix estimated by the adaptive BGL with non-informative priors, (e) concentration matrix estimated by the adaptive BGL with chemical structure informative priors.

Results of the simulation studies for the autoregressive case are shown in Figure 7 with numerical summaries provided in Table 3. GGM estimation by the Bayesian Graphical Lasso (BGL) and Adaptive BGL exhibited similar performance characteristics with respect to sensitivity, specificity, AUC, and F_1 measure. The performance of the chemical structure informative adaptive BGL varied significantly based on the suitability of the informative prior distribution for shrinkage parameters. In the good prior case in which it is assumed that the structural similarity and data generating process were deterministically linked, the structure adaptive BGL demonstrated significantly higher sensitivity, specificity, AUC, and F_1 measure than the other techniques. Conversely, in the poor prior case in which the relationship between the simulated structural similarity and the data generating process was masked by gaussian noise, average AUC and F_1 measure were significantly lower for the structure adaptive BGL than the remaining techniques.

Table 3. Results of the AR(1) simulation studies. For each of the compared techniques, sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and F1 measure are reported. Reported values represent the sample mean and standard deviation over the simulation study replicates

Technique	Sensitivity	Specificity	AUC	F₁ Measure
Bayesian Graphical Lasso (BGL)	0.6792 ± 0.073	0.9896 ± 0.007	0.9740 ± 0.014	0.7786 ± 0.054
Adaptive BGL	0.6702 ± 0.079	0.9936 ± 0.006	0.9793 ± 0.014	0.7827 ± 0.062
Chemical Structure Adaptive BGL (Good prior)	0.9894 ± 0.019	0.9997 ± 0.001	1.000 ± 0.000	0.9938 ± 0.011
Chemical Structure Adaptive BGL (Poor prior)	0.8175 ± 0.068	0.8763 ± 0.033	0.9148 ± 0.036	0.6448 ± 0.066

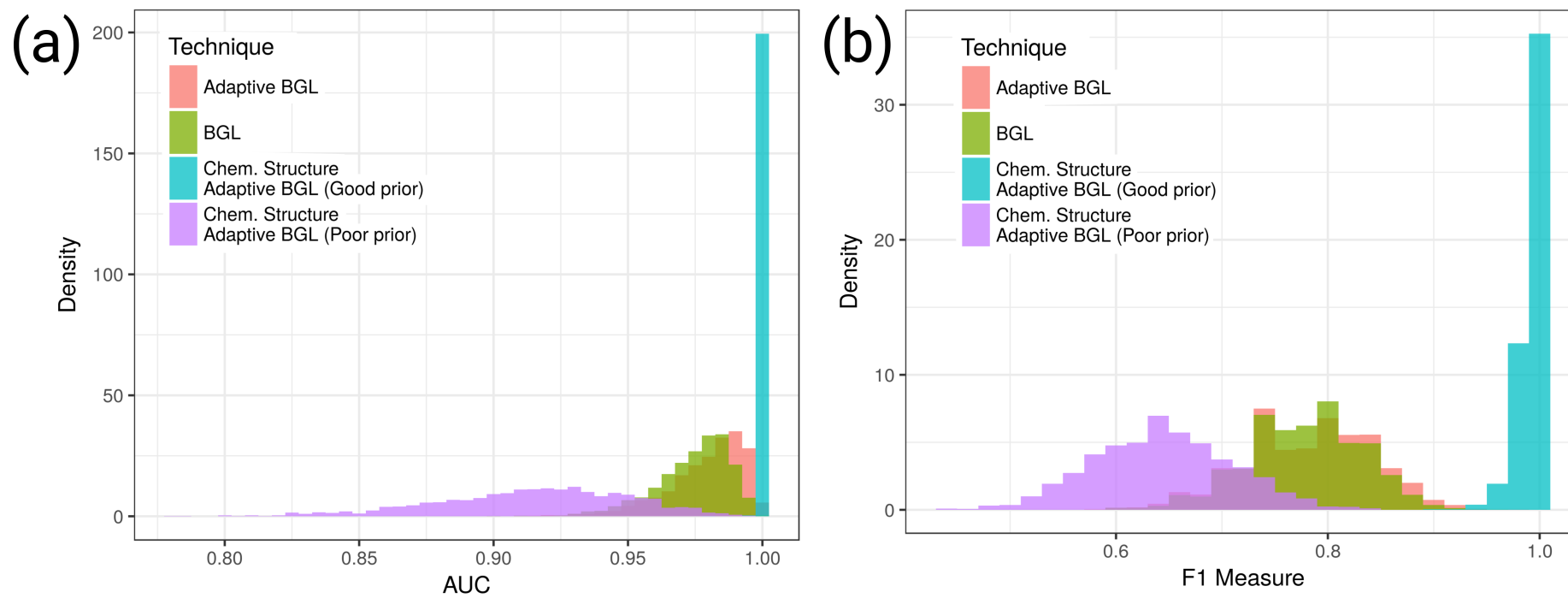


Figure 7. Results of the AR(1) simulation studies. The comparative performance of the techniques (as the ability to detect an edge, given an edge is truly present) is presented. Subfigure (a) shows a histogram of the observed area under the receiver operating characteristic curve (AUC) values, while (b) shows the F1 measure.

CHAPTER VII

A PLASMA INTERACTOME FOR STABLE HEART DISEASE

1 Cohort

In order to determine changes in the plasma metabolome associated with myocardial infarction (MI) characterized by thrombotic etiology versus non-thrombotic etiology, DeFilippis and colleagues assembled a human cohort as previously described [DeFilippis et al., 2015, DeFilippis et al., 2017, Trainor et al., 2017]. Briefly, 80 human subjects presenting with suspected acute MI or stable coronary artery disease (CAD) were enrolled. Utilizing a stringent criteria based on clinical presentation, angiographic evidence, and histological evidence, MI subjects were adjudicated as thrombotic MI or non-thrombotic MI. Blood samples were collected at the time of acute presentation (presentation to the coronary artery catheterization lab prior to procedures) and at a follow-up evaluation approximately three months later. To estimate the structure of a stable heart disease plasma interactome, we used the follow-up evaluations from all available MI subjects as well as the evaluations from stable CAD subjects. The analytical sample thus consisted of 47 whole blood samples from human subjects with definitive heart disease who were not experiencing an acute event at the time of sampling.

2 Plasma metabolomics

Details of the metabolite quantification have been described previously [Trainor et al., 2017], but a brief overview is provided as follows. Plasma samples were

prepared from whole blood and a recovery standard was added. Vigorous shaking was applied utilizing a GenoGrinder 2000 (Glen Mills, Metuchen, NJ) and methanol was added and to precipitate proteins. The extract containing small molecules was divided into five aliquots, four of which were analyzed using different platforms while the remaining aliquot was reserved. Two aliquots were analyzed by ultra-performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS) with negative and positive ion mode electrospray ionization (ESI). A third aliquot was also analyzed by UPLC-MS/MS with negative ion mode ESI and a method optimized for polar metabolite detection. The fourth aliquot was analyzed by gas chromatography-mass spectrometry (GC-MS). 1,032 chemical features were detected utilizing the multiple platforms in the analysis of the plasma samples. Of these, 590 compounds were identified by matching to authentic standards based on retention index, mass to charge ratio, and MS2 data; 73 were identified based on experimental data matched to curated databases; and 369 could not be confidently identified. As the original data dependent acquisition was conducted utilizing both acute event samples and stable heart disease samples, metabolites not detected in the stable heart disease samples were removed. Metabolites missing from greater than 70% of the samples or without compound identification were also removed, resulting in a final dataset with 522 metabolites across 47 samples. Minimum values were then imputed for the remaining metabolite relative abundances with missing data. As many of the metabolites exhibited approximately log-normal relative abundance distributions, metabolite abundances were log-transformed. Finally, the data was mean centered so that each metabolites relative abundance distribution was centered about zero.

3 Generation of chemical structure informed priors

A heatmap representation of the structural similarity between metabolites is shown in Figure 8. This heatmap was constructed using agglomerative hi-

erarchical clustering using Wards method and squared distances (with distance computed as $d_{ij} = 1 - s_{ij}$, where s_{ij} is the structural similarity between compounds). For illustrative purposes, the cluster containing cholate was retrieved from the root dendrogram by extracting the branch with height X, as the structural-adaptive BGL subnetwork generated by cholate is considered later. Considering clusters generated by branches with low merge heights (high structural similarity), cholate was a member of a cluster with other closely related compounds such as deoxycholate, 3 β -hydroxy-5-cholenoic acid, and glycocholate. Considering the more inclusive cluster generated by the branch at join height x, other members included many intermediates in progestagen, androgen, glucocorticoid, and mineralocorticoid steroid metabolic pathways. These steroid hormone metabolites were all members of a cluster with similar within-cluster distances. Finally, at the same branch height that joined steroid hormone and cholate metabolites, a branch consisting of tocopherols cluster and squalene cluster was also joined.

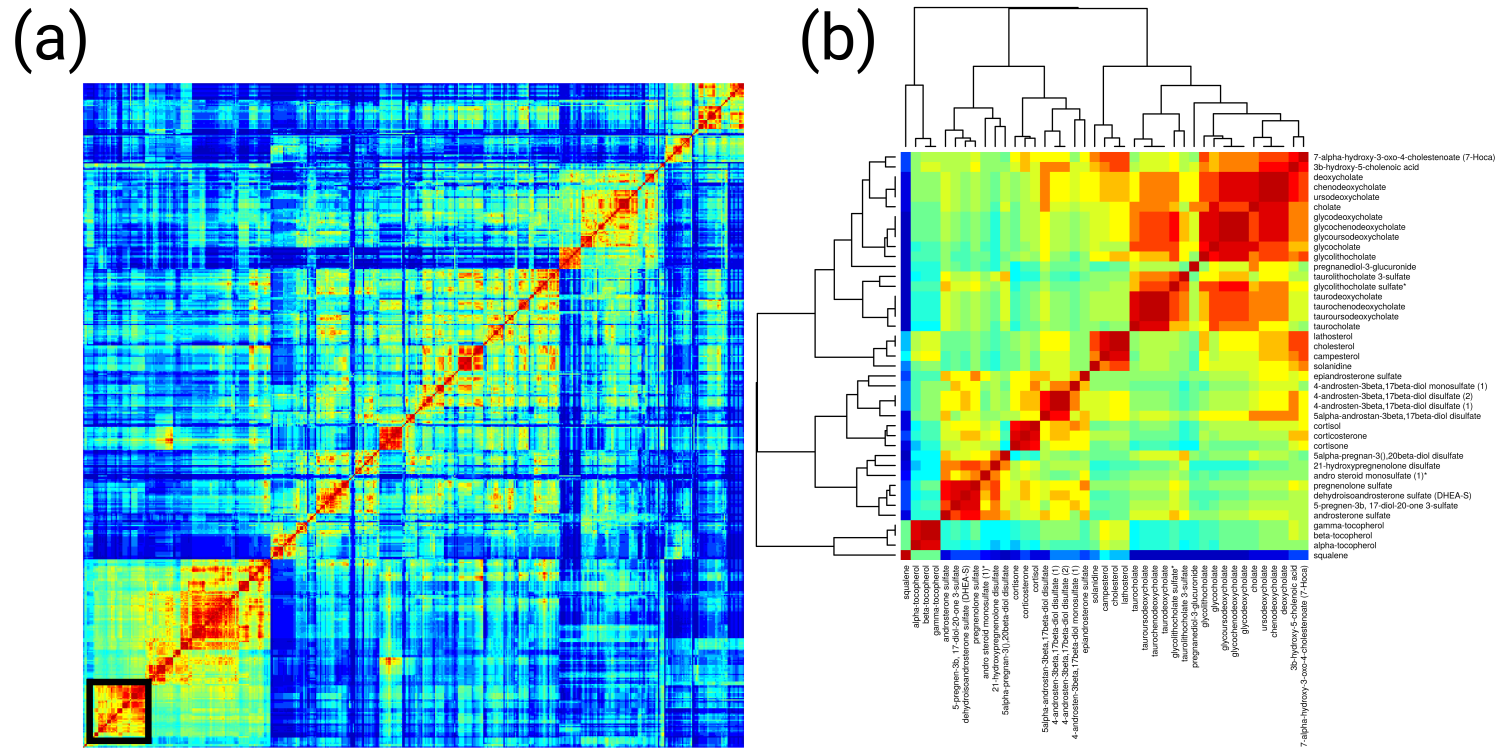


Figure 8. Heatmap showing the molecular structure similarity between the metabolites that were detected and quantified from the analysis of blood plasma from thrombotic MI, non-thrombotic MI, and stable CAD subjects. Subfigure (a) shows all metabolites, subfigure (b) highlights the dendrogram branch that contains cholate.

4 Gibbs sampling

To approximate the posterior distribution of $p(\boldsymbol{\Omega}|\mathbf{X})$, a Markov Chain was generated of length 1,000 with a 250 iteration burn-in period. From each sample from the posterior distribution $p(\boldsymbol{\Omega}|\mathbf{X})$, a partial correlation coefficient matrix was computed, yielding a simulated posterior distribution for the matrix of partial correlation coefficients.

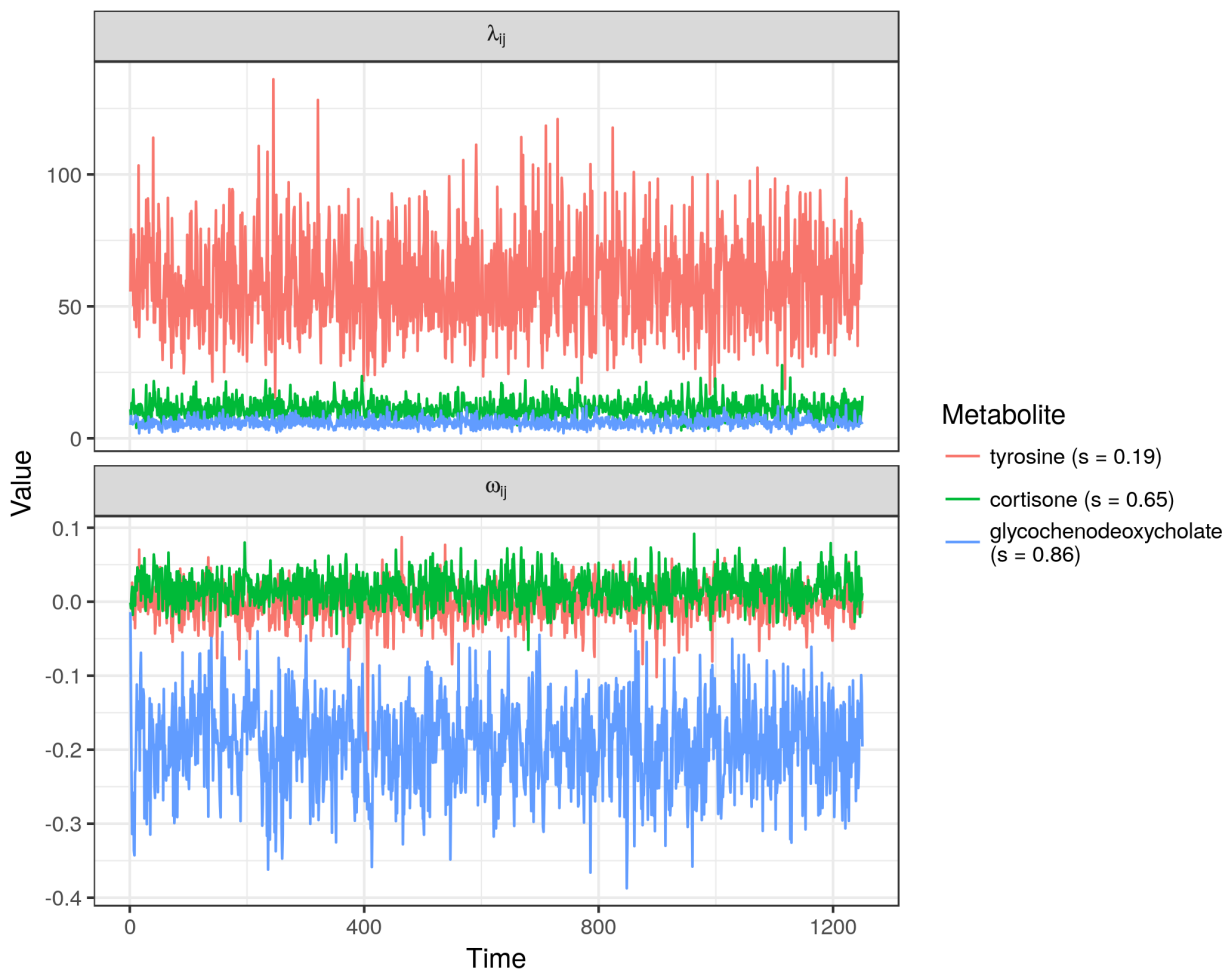


Figure 9. Time series plots for the MCMC sampler for the shrinkage parameter λ_{ij} and for the concentration matrix entry ω_{ij} for the following metabolite pairs: (cholate, tyrosine), (cholate, cortisone), (cholate, glycochenodeoxycholate).

Elements of the MCMC sampling iterations are presented in Figure 9. Continuing with the working example of the metabolite cholate, Markov chains

are presented for the estimation of the concentration parameters for the pairs (cholate, tyrosine), (cholate, cortisone), and (cholate, glychochenodeoxycholate). These metabolites are highlighted as exemplars of metabolites with relatively low, medium, and relatively high chemical structure similarity with cholate. The time series of the MCMC sampling for the shrinkage parameter, λ_{ij} , demonstrated differences between the three pairs. Averaged across iterations, more shrinkage was applied with decreasing chemical similarity between the metabolite pairs. While the shrinkage parameter sample values for (cholate, cortisone) tended to be significantly smaller than the sample values for (cholate, tyrosine), substantial overlap was observed in the posterior distribution of the concentration parameters for the same pairs.

From the simulated posterior distribution of $\mathbf{\Omega}$, the posterior mean $E(\mathbf{\Omega}|\mathbf{X})$ was estimated after discarding burn in iterations. The posterior mean of the distribution of partial correlation coefficients was also computed. The resulting plasma metabolite interactome inferred by the structure-adaptive Bayesian Graphical Lasso is presented in Figure 10. This figure presents both the entire graph representing the posterior mean partial correlations as well as the sub-graph generated by considering the neighbors of cholate. For ease of viewing, only edges for which $|\rho_{ij}| > 0.05$ are plotted in the presentation of the full graph.

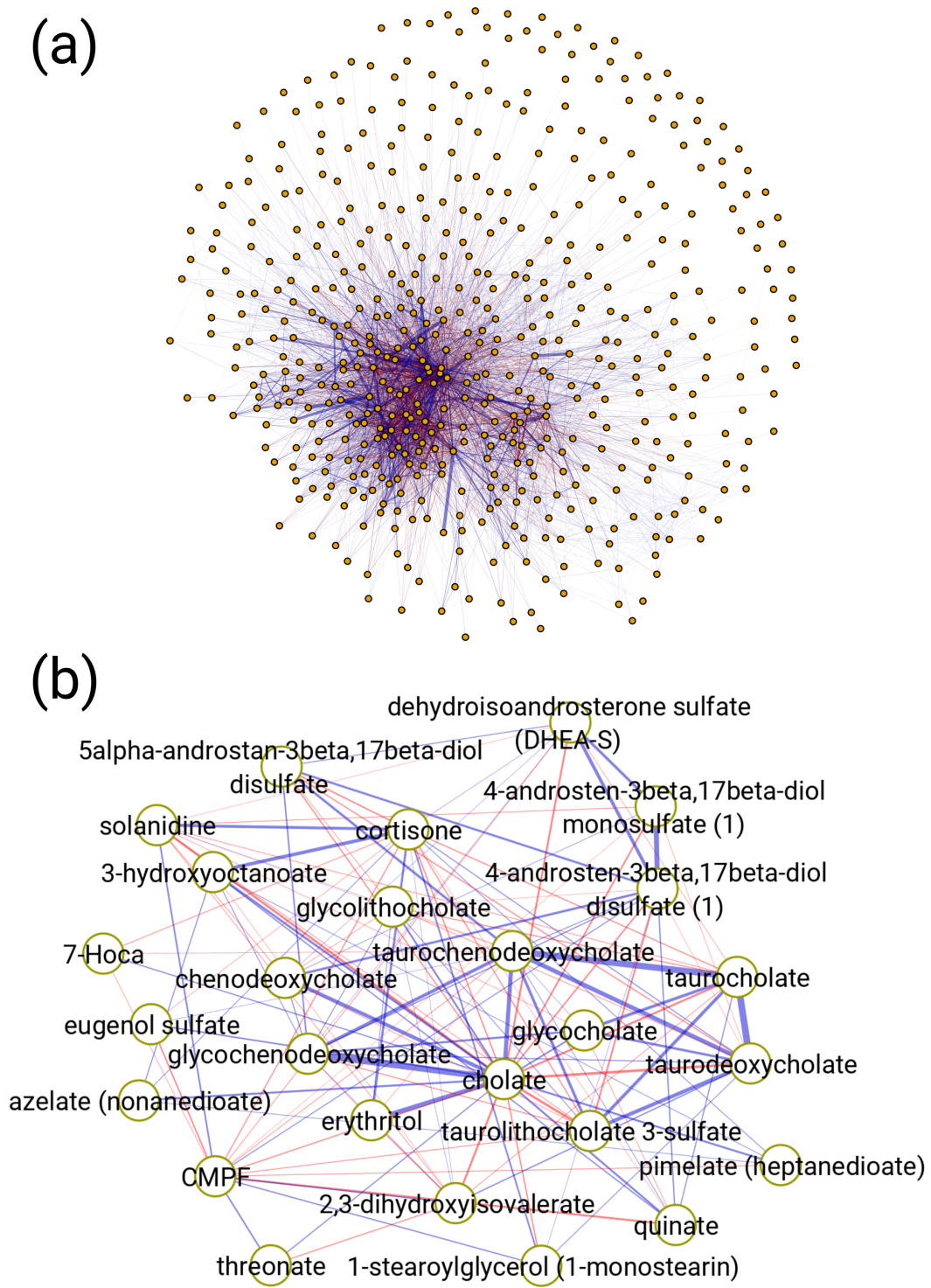


Figure 10. Graphical representation of the plasma metabolite interactome estimated by the chemical structure adaptive Bayesian Graphical Lasso (BGL) for stable heart disease. A simulated posterior distribution for the matrix of partial correlation coefficients was determined from the simulated posterior distribution of the concentration matrix Ω . Mean values of each partial correlation coefficient were then determined and are represented as colored edges (negative values represented in red, positive values in blue). Subfigure (A) shows all metabolites in the interactome along with edges for which $|\rho| > 0.05$.

Positive partial correlation coefficients were observed between cholic acid the following other primary bile acids: glycocholic acid, chenodeoxycholic acid, and glycochenodeoxycholic acid. Negative partial correlation coefficients were observed between cholate and the following: taurocholic acid, taurodeoxycholic acid, and taurochenodeoxycholic acid. In addition, the bile acid 7-Hoca and the bile acid conjugate tauroolithocholate 3-sulfate were first neighbors of cholic acid. Multiple conjugated androsteroes were observed to be first neighbors of cholic acid as was the glucocorticoid cortisone and the steroidal alkaloid solanidine. Other metabolites that were first neighbors of cholic acid included: 3-Carboxy-4-methyl-5-propyl-2-furanpropionic acid (CMPF), eugenol sulfate, erythritol, 2,3-dihydroxyisovalerate, threonate, quinate, pimelate, and azelate.

5 Discussion

Making inferences regarding how metabolic processes differ between phenotypes is the ultimate goal of most metabolomics and systems biology studies. Yet, unlike comparing the concentration or abundance of one metabolite across two or more phenotypes, for which simple statistical tests such as t-tests, Wilcoxon Rank-Sum tests, or multi-group analogues are readily available, a statistical framework for determining if and how metabolic processes differ between phenotypes remains elusive. Both strictly empirical methods (e.g. correlation analyses) and *a priori* knowledge based approaches (e.g. pathway enrichment analyses) suffer from substantial flaws. In terms of empirical methods, the analysis of correlations (such as by the Pearson, Spearman or biweight midcorrelation coefficient) reveals the marginal associations between metabolites; however, these methods do not uncover the relationship between a pair of metabolites conditional on the abundances of the remaining metabolites. Gaussian graphical models have been proposed previously in the context of metabolomics [Krumsiek et al., 2011] as an alternative, as GGMs can be utilized to determine the partial or conditional relationship between metabo-

lites. In their work Krumsiek et al. [Krumsiek et al., 2011] show that GGM edges (or concentration matrix entries) estimated from the analysis of blood serum samples from a large human cohort correspond to known metabolic pathway interactions. Consistent with this, our simulation studies illustrate the advantage of analyzing metabolite-metabolite interactions using the partial correlation coefficients from a GGM as opposed to correlation networks. In the case of an autoregressive correlation structure as might be observed given a linear metabolic pathway, correlation networks exhibit extremely high connectivity, and consequently could not be utilized to elucidate the order of reactions. In contrast, we observe high sensitivity and specificity in detecting the true edges using a Bayesian Graphical Lasso estimated GGM. While the approach utilized by Krumsiek et al. [Krumsiek et al., 2011] was appropriate for the analysis of their data, it would not be possible to apply this approach in studies in which the sample size is smaller than the number of metabolites, as is common in many metabolomics studies. Frequentist regularization methods represent a class of solutions for ensuring that the concentration matrix, or equivalent GGM topology is estimable. In an implicit manner, frequentist regularization methods for estimating GGMs place a higher *a priori* probability on models with concentration matrix entries of smaller magnitude [Wang, 2012]. However, this implicit prior cannot incorporate *a priori* knowledge as to whether some metabolites are more likely to be related than others.

In contrast to empirical methods, *a priori* knowledge based approaches such as pathway enrichment analyses consider the relationships between metabolites to be deterministically known which are then used to contextualize empirical results. Previous work [Barupal et al., 2012, Barupal and Fiehn, 2017] has highlighted that the coverage of metabolites detected in metabolomics studies in commonly utilized metabolic pathway and reaction databases may be extremely low. For example, Barupal et al. [Barupal and Fiehn, 2017] observe that given 385 metabolites identified from the plasma of non-obese diabetic

(NOD) mice, only 135 metabolites (or 35.1%) could be mapped to KEGG pathways [Kanehisa and Goto, 2000, Kanehisa et al., 2016]. To address this problem, Barupal et al. [Barupal and Fiehn, 2017] propose an alternative approach that utilizes both existing chemical ontological terms and chemical similarity between metabolites to develop coherent categories of metabolites for enrichment analyses. In the current work, we have sought a framework for balancing the benefits of empiricism with the benefits of a priori knowledge based approaches, while seeking to minimize the risks associated with both approaches. As opposed to considering metabolites as deterministically assigned to fixed pathways, our approach assumes that metabolites that are linked by biochemical reactions will exhibit overlap in local substructures. From this, our approach generates prior distributions for shrinkage parameters for the estimation of Gaussian graphical models. The posterior distribution of GGM parameters is thus proportional to the likelihood of the concentration matrix parameters (or the partial correlations between metabolites) times the prior probability of the concentration matrix parameters (which are linked to the structural similarity between metabolites). Similar to the non-informative BGL approach, this approach ensures that the concentration matrix is estimable via the Bayesian analog of regularization, however the regularization is applied given the prior belief that stronger associations are a priori more likely given structurally related compounds than unrelated compounds. While we find better justification for using structural similarity to generate prior probability distributions for shrinkage parameters in estimating a GGM, this approach would generalize to the use of priors from metabolic pathway maps. A previous work sought to estimate a GGM using 17 compounds quantified by NMR from 24 microglia cell culture samples using priors determined from KEGG [Peterson et al., 2013].

In addition to evaluation via simulation studies, we have applied the chemical structure adaptive BGL to generate a media-specific (blood plasma) metabo-

lite interactome for stable heart disease. This model may serve as a reference model for comparing how the probabilistic interactions between metabolites in circulation change during acute disease events such as myocardial infarction or unstable angina. From this model, we have observed probabilistic interactions that are consistent with previous research in metabolism, as can be observed by focusing on the metabolite cholate. Bile acids are the major catabolic intermediate of cholesterol [Russell, 2003]. Within mammals, the bile acid pool consists of primary bile acids such as cholic acid and chenodeoxycholic acid which are synthesized from cholesterol by enzymes expressed in hepatocytes, as well as secondary bile acids that are synthesized from primary bile acids by bacteria in the gut [Garca-Caaveras et al., 2012, Hofmann et al., 2010, Russell, 2003]. In addition to bile acids aiding in the digestion of nutrients in the gut, bile acids also act as signaling molecules that have been shown to regulate glucose and lipid metabolism [Ferrebee and Dawson, 2015, Khurana et al., 2011]. Given the substantial proportion of cholesterol that is converted to bile acids leading to elimination, bile acid metabolism is linked to atherosclerosis [Meissner et al., 2013]. In addition bile acids as signaling molecules affect cardiac [Desai et al., 2017, Rainer et al., 2013] and circulatory physiology [Khurana et al., 2011] via direct effects such as taurodeoxycholic acid mediated vasodilation [Khurana et al., 2005]. With respect to the current work, we observed relatively strong partial correlation between cholic acid and other primary bile acids. Additionally, partial correlations were observed between cholic acid and steroid hormones that share cholesterol as a common precursor. Given the importance of bile acids in cholesterol metabolism, atherosclerosis, cardiac physiology and circulatory physiology, a reference model of the probabilistic interactions of bile acids in circulation can help elucidate how acute disease events impact bile acid metabolism.

As with any Bayesian approach, the choice of prior probability distribution has a direct influence on the posterior distribution of model parameters

[Gelman et al., 2004]. In the current work, we have utilized informative priors that are linked to chemical structure similarity. This represents a potential limitation of the current work. Over the course of the simulation studies, we observed, unsurprisingly, that by introducing random noise into the simulated structural similarity the performance of the chemical structure adaptive BGL deteriorated. Further, the performance of the technique given poor prior information was, on average, worse than the performance of techniques such as the non-adaptive BGL that rely on non-informative priors. One element of the chemical structure adaptive BGL is worth noting in this context. In our proposed formulation, other monotonic functions for relating structural similarity to the Gamma scale parameter may be employed, as well as different shape and scale hyperparameters can be utilized. In this manner, the experimenter can diminish or strengthen the degree to which structural similarity impacts shrinkage. A second limitation of the current work is the choice of a multivariate Gaussian distribution for representing the joint distribution of metabolite abundances. While transformations in the stable heart disease data were applied over each metabolite to reduce the degree of departure from normality, the underlying intensity data is not normally distributed. Further, there are many cases in which approximate normality is not an achievable aim. A metabolite that is only present in some samples (e.g. acetaminophen metabolites that are present in some human subjects who have taken this medication, but not others) is one such case. Following a missing value imputation procedure, such a metabolite would exhibit a bimodal distribution that would not be well described by a Gaussian model.

CHAPTER VIII

A BAYESIAN DIAGNOSTIC MODEL FOR DIFFERENTIATING MI TYPE

1 Acute Myocardial Infarction

Acute Myocardial Infarction (AMI), which is an acute manifestation of coronary heart disease, is defined by myocardial ischemia (exposure of cardiac myocytes to oxygen deprivation) and necrosis (a form of cell death) [Trainor et al., 2018]. AMI may occur following atherosclerotic plaque disruption or other conditions which cause demand ischemia [Thygesen et al., 2012, Arbab-Zadeh et al., 2012]. Irrespective of underlying cause, ischemia and necrosis are the common pathological characteristics of all AMI. Thrombotic MI (MI results from spontaneous atherosclerotic plaque disruption with the formation of an occluding coronary thrombus) versus non-thrombotic MI represents an important etiological distinction [DeFilippis et al., 2017], as both types necessitate different treatment approaches [Thygesen et al., 2012]. A diagnosis of AMI can be substantiated by blood-based diagnostic tests that measure isoforms of the protein troponin which is released into the circulation following necrosis [Newby et al., 2012]. To date a blood-based diagnostic test capable of discriminating between thrombotic and non-thrombotic MI has not been developed, although it has been previously shown that a metabolic signature may differentiate between the types [DeFilippis et al., 2017, Trainor et al., 2018]. In the present work, we set out to develop a Bayesian model for differentiating thrombotic MI from both non-thrombotic MI and stable coronary artery disease (CAD) using metabolites detected in blood plasma. We regard plasma as a promising media for developing a non-invasive test as

plasma contains hormones, enzymes, lipoproteins, and other metabolic intermediates found in circulation. As metabolite concentrations are a product of genetic factors, environmental exposures, and the interaction between the two, sampling metabolites may provide a more robust characterization of the state of an organism than other approaches such as genomics.

2 Clinical cohort and samples

Towards the effort of developing a Bayesian diagnostic model, we utilized previously collected samples from a patient cohort that was recruited specifically for contrasting thrombotic MI from multiple control phenotypes [DeFilippis et al., 2015, DeFilippis et al., 2017]. This cohort was comprised of three phenotypic groups of human subjects: thrombotic MI, non-thrombotic MI, and stable CAD. In reference to the thrombotic MI group, both control groups (non-thrombotic MI and stable CAD) served as procedural controls as all groups underwent a cardiac catheterization procedure. The stable CAD group provides a stable disease control as both thrombotic MI subjects and stable CAD subjects have underlying coronary artery disease. Non-thrombotic MI subjects presented with myocardial necrosis and thus the non-thrombotic MI group serves as an acute disease event control.

Whole blood was collected shortly before cardiac catheterization from all study subjects. Details of the sample handling, sample processing, separation by liquid or gas chromatography, and mass spectrometry analysis for quantifying metabolites are provided in Chapter VII Section 2.

3 Feature selection

A significant analytical challenge in developing a blood-based diagnostic test for differentiating MI types is to determine a small set of metabolites that should be included in the statistical model from a limited number of training samples. Specifically, 1,032 chemical features (identified compounds or unknown compounds) were

detected from 11 thrombotic MI, 12 non-thrombotic MI, and 15 stable CAD subjects. To ensure that the statistical model is estimable from a small sample size, and as building a targeted MRM assay or multiplexed ELISA assay can only accommodate a small number of compounds, feature selection is a critical task. While other dimension reduction techniques such as latent variable approaches (e.g. Partial Least Squares models) that create new variables which are linear combinations of metabolites would be amenable for reducing the number of coefficients to be estimated in a classification model, these approaches would not reduce the number of metabolites needing to be quantified by future targeted assays. Consequently, we prioritized reducing the number of metabolites considered. To determine a statistical classifier with five metabolites, 9.7×10^{12} combinations are possible. In order to search the space of possible models we employed a feature selection technique that utilizes an evolutionary algorithm and seeks a consensus solution over bootstrapped datasets as described in Trainor et al. [Trainor et al., 2018]. In this work, small sets of metabolites were included in a multinomial logit classifier. Each model represented an individual in a population. Genetic fitness was determined as the likelihood of each individual model. These populations of models were allowed to evolve given evolutionarily inspired processes such as birth, recombination, and death. Populations were grown over bootstrapped datasets to increase diversity and reduce the correlation between models in the populations. Finally, the frequency that an individual metabolite was included in models within the final population after epochs of evolution was determined yielding a variable importance score for each metabolite. A correlation plot illustrating the metabolites with greatest variable importance score using the described technique is shown in Figure 11.

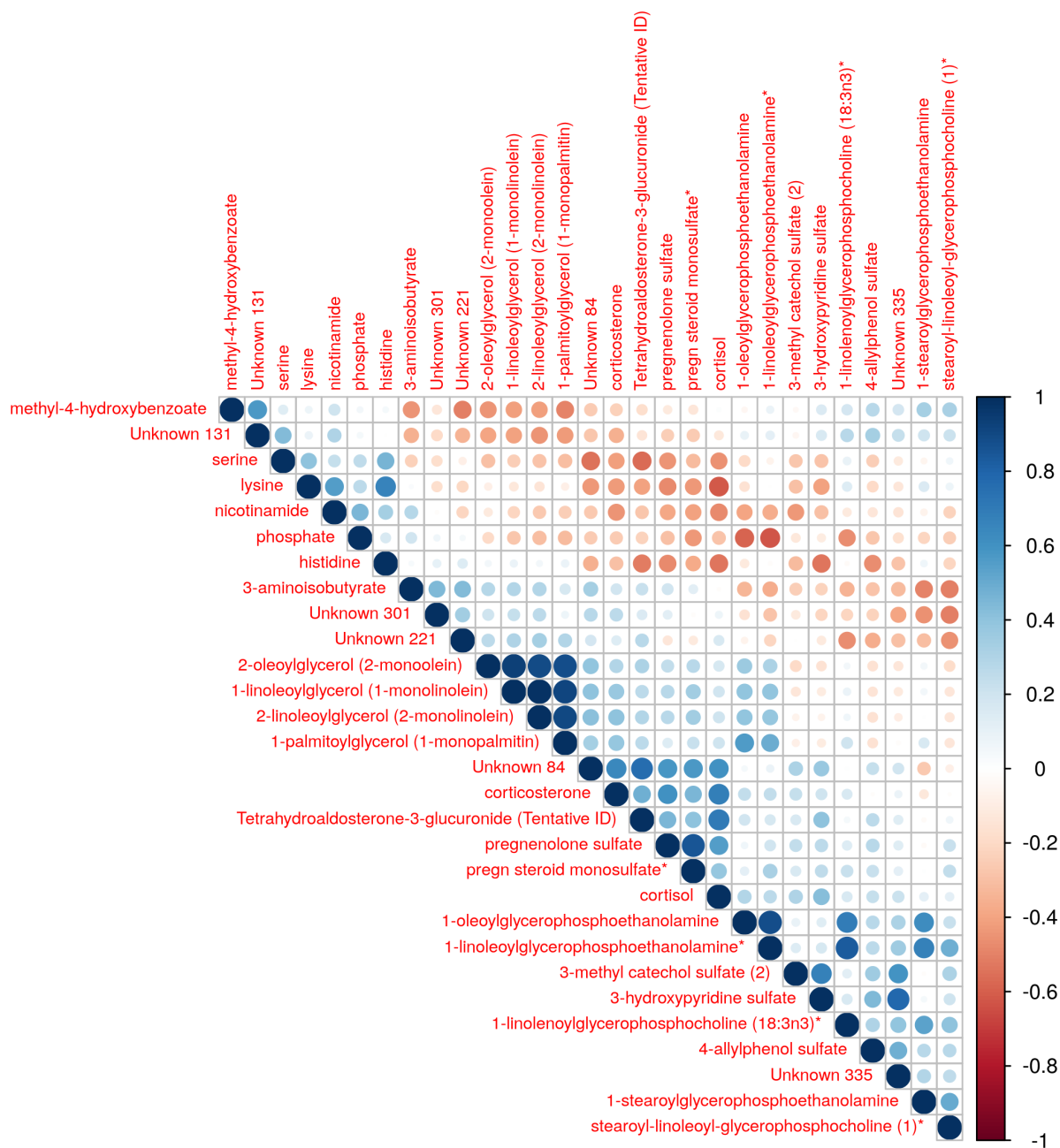


Figure 11. Metabolites with the highest variable importance score given the feature selection method we employed in a previous work [Trainor et al., 2018]. The Pearson correlation coefficients between metabolite transformed and scaled abundances are shown.

4 Multinomial logistic regression model

A multinomial logistic regression model was assumed for determining the probability a sample from a clinical subject was member of the thrombotic MI, non-thrombotic MI, or a stable CAD study group. The multinomial logistic regression model is a generalized linear model and has the following link function and model form:

$$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + \mathbf{x}_i^T \boldsymbol{\beta}_j, \quad (53)$$

where i indexes individual samples, j is the index of study groups, $\boldsymbol{\beta}_j$ is a vector of regression coefficients (with separate coefficients for each group), α_j is a group specific intercept term, and J represents the reference group. Given a specific value for the link function, the probability a sample belongs to specific study group can be computed as:

$$\hat{\pi}_{ij} = \frac{\exp \hat{\eta}_{ij}}{\sum_{k=1}^J \exp \hat{\eta}_{ik}}. \quad (54)$$

The model can be stated as a Bayesian model with the following priors and deterministic component:

$$\begin{aligned} \alpha_j &\sim N(0, 4) \\ \boldsymbol{\beta}_j &\sim N(0, 1) \\ \log \frac{\pi_{ij}}{\pi_{iJ}} &= \alpha_j + \mathbf{x}_i^T \boldsymbol{\beta}_j \\ Y &\sim \text{Multinom}(\boldsymbol{\pi}) \end{aligned} \quad (55)$$

5 MCMC sampling of the posterior distribution

A MCMC sampler known as the ‘‘No-U-Turn’’ sampler was utilized to simulate the posterior distribution of model parameters [Hoffman and Gelman, 2014]. This algorithm is an extension of the Hamiltonian Monte Carlo algorithm. The Hamiltonian Monte Carlo algorithm is designed to model a target probability distribution as a Hamiltonian system [Betancourt, 2017]. In such a system, a parameter vector

θ is viewed as a particle in a D -dimensional space [Neal, 2011] by defining a vector field that is aligned with the typical set (the region of a parameter space with both significant volume and density) [Betancourt, 2017]. Defining such a vector field is a complex task. While a vector field could be defined using the gradient of the target probability distribution, this vector field would pull a particle towards the mode of the distribution. By adding a momentum term, the vector field can be defined so as to restrict a particle to maintaining the Hamiltonian at a fixed value. The Hamiltonian Monte Carlo algorithm utilizes Metropolis-Hastings proposals for updating both the momentum and position variables [Neal, 2011, Betancourt, 2017]. A critical aspect of Hamiltonian Monte Carlo sampling is determining the optimal integration time for a particle to travel along a Hamiltonian path. One approach to dynamically determining this parameter is the “No-U-Turn” termination criteria [Hoffman and Gelman, 2014, Betancourt, 2017]. Conceptually, this criteria ensures that expansion of a trajectory continues to visit previously unexplored neighborhoods while terminating the trajectory when it returns to previously explored neighborhoods.

The “No-U-Turn” Hamiltonian MCMC sampler was implemented by others in *Stan*, a probabilistic programming language [Carpenter et al., 2017]. *Stan* is written in C++ and provides the “No-U-Turn” sampler, a Hamiltonian Monte Carlo Sampler, a variational inference algorithm, and multiple optimizers. *Stan* provides grammar and syntax for succinctly specifying Bayesian models; can design a sampler for the posterior distribution of model parameters; and using a C++ compiler, *Stan* compiles the sampler to byte code. Additionally, *Stan* provides data structures for storing MCMC chains and model output. An R to *Stan* interface, *rstan* [Stan Development Team, 2018], has been developed allowing an end user to pass datasets between *Stan* and to return MCMC chains and model results.

6 Results

In order to simulate the posterior distribution of model parameters, four MCMC chains were generated utilizing the No-U-Turn algorithm. An example MCMC chain is illustrated in Figure 13(a). In this figure 2,000 of the 10,000 iterations of the regression coefficient parameter for 3-hydroxypyridine sulfate from one of the MCMC chains (chain #1) is shown. Additionally, this figure shows how the posterior distribution is generated from the MCMC chain. In Figure 13(b), by-group parameter histograms are shown for the multinomial logit model parameters corresponding to 3-hydroxypyridine sulfate. For this metabolite, the probability mass of posterior coefficient estimates is centered slightly below zero for non-thrombotic MI, while the center of the posterior distribution is above zero for thrombotic MI. From the MCMC simulated posterior distribution, Bayesian credible intervals were determined for each model parameter as shown in Figure 13.

As the objective of the present analysis was to develop a model capable of discriminating between the three phenotypes utilizing only the information available at the presentation of patients and in a non-invasive manner, we also considered introducing clinical troponin values. Bayesian credible intervals are presented from the posterior distribution given the introduction of troponin in Figure 14. Substantial qualitative changes in the credible intervals were not observed following the introduction of troponin. To compare the models, the “Widely applicable information criterion” was utilized (see Table 4). A χ^2 test on one degree of freedom reveals that troponin did not lead to a significant improvement overall ($p = 0.36$).

Table 4. Evaluation of goodness of model fit with or without clinical troponin values. Inclusion of troponin did not result in a statistically meaningful improvement in goodness of fit. WAIC: Widely applicable information criteria. SE: Standard error.

Model	WAIC	SE
Without troponin	15.43	4.05
With troponin	14.60	3.72
Difference	0.83	0.87

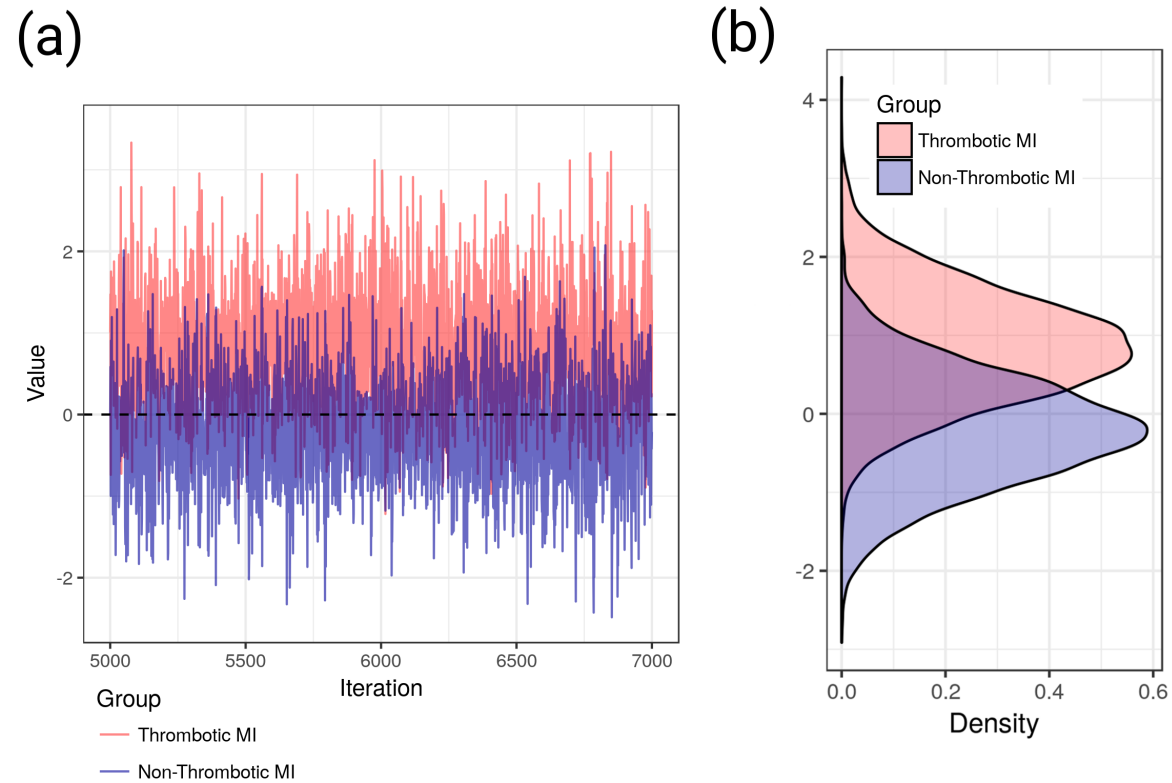


Figure 12. (a) MCMC chains for the by-group multinomial logit model parameters corresponding to 3-hydroxypyridine sulfate. (b) Histogram showing the simulated posterior probability distribution for the same parameters.

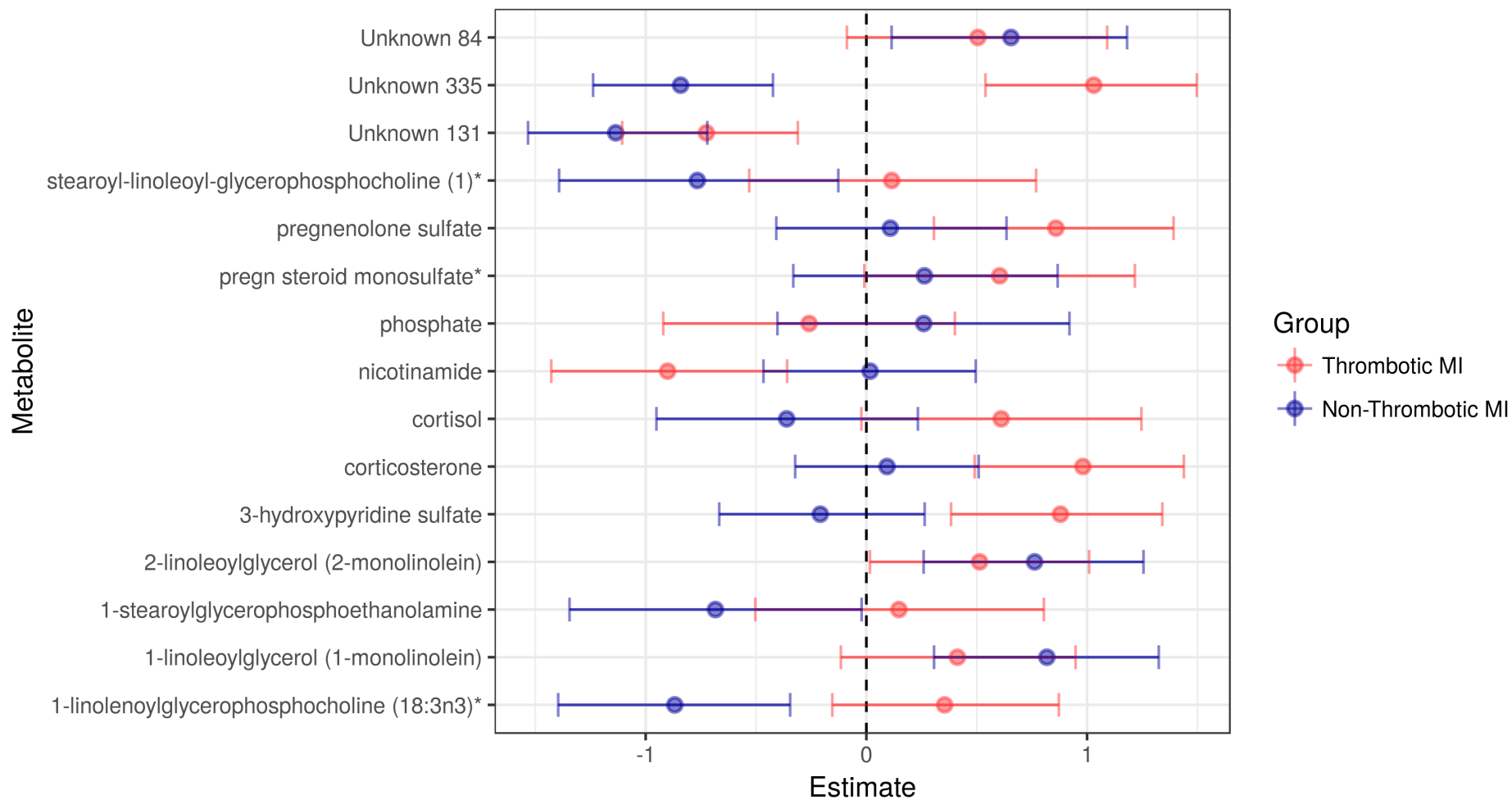


Figure 13. 50% Bayesian credible intervals for the by-group multinomial logistic regression coefficients. The multinomial models did not include clinical troponin assay values.

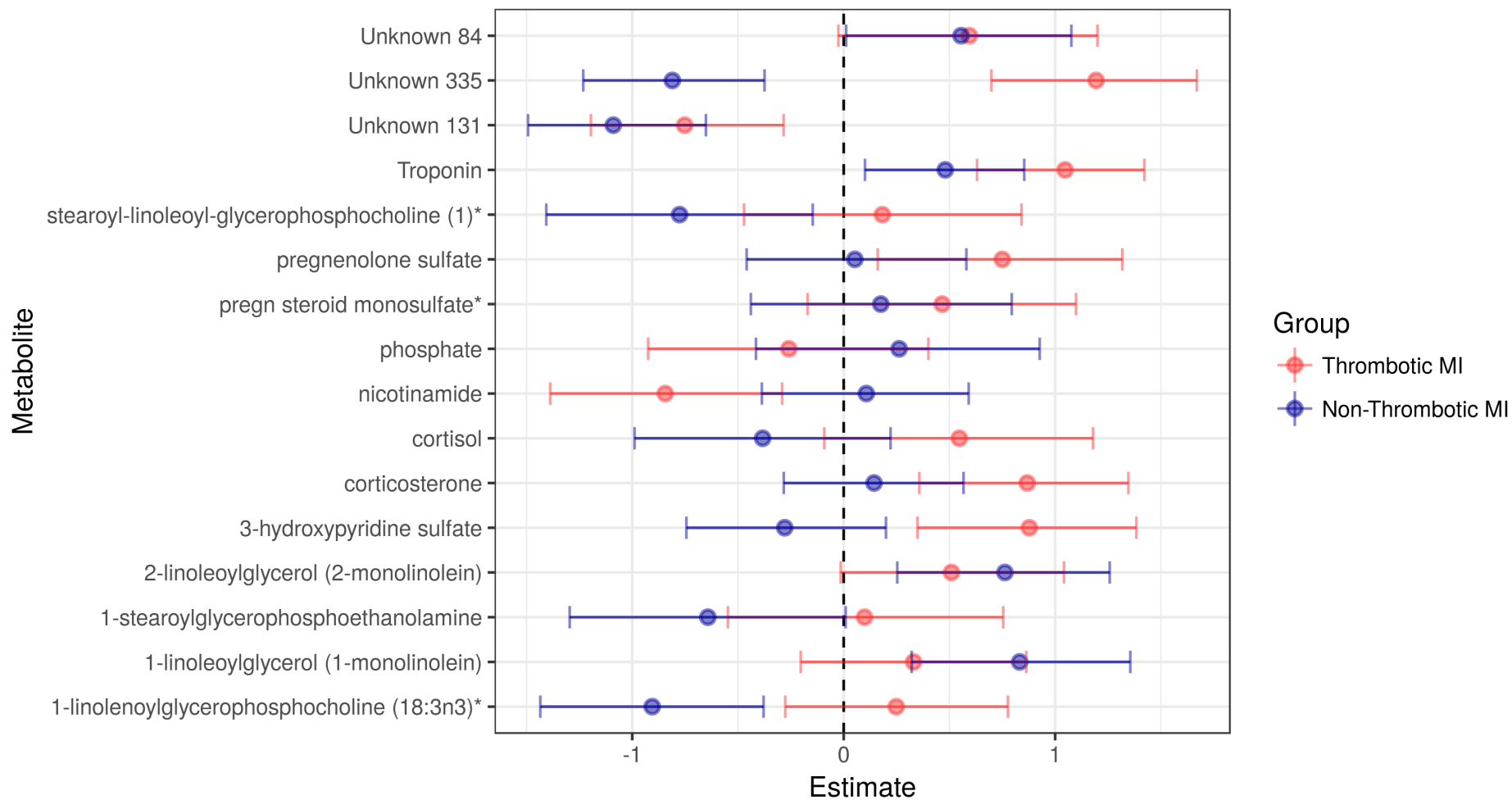


Figure 14. 50% Bayesian credible intervals for the by-group multinomial logistic regression coefficients. The multinomial models did include clinical troponin assay values.

Given that the MCMC chains provide the joint posterior distribution of model parameters, the correlation between parameter estimates can be evaluated. Two parameter estimates were significantly correlated, both of which were for monoacylglycerols [1-linoleoylglycerol (18:2)] and [2-linoleoylglycerol (18:2)]. This relationship was further explored. Figure 15 shows a 3-dimensional scatterplot of samples from the simulated posterior distribution. In the first plane, the parameter estimates for both monoacylglycerols are shown, while the remaining axis shows the log-posterior probability of the model with these parameterizations. Similarly, Figure 16 shows the same slice of the posterior distribution with the log-posterior probability axis removed and represented as a color instead. The relatively strong negative correlation between the parameter estimates, without a systemic change in the log-posterior probability suggests that these two monoacylglycerols modulate the probability of the phenotypes in a similar way.

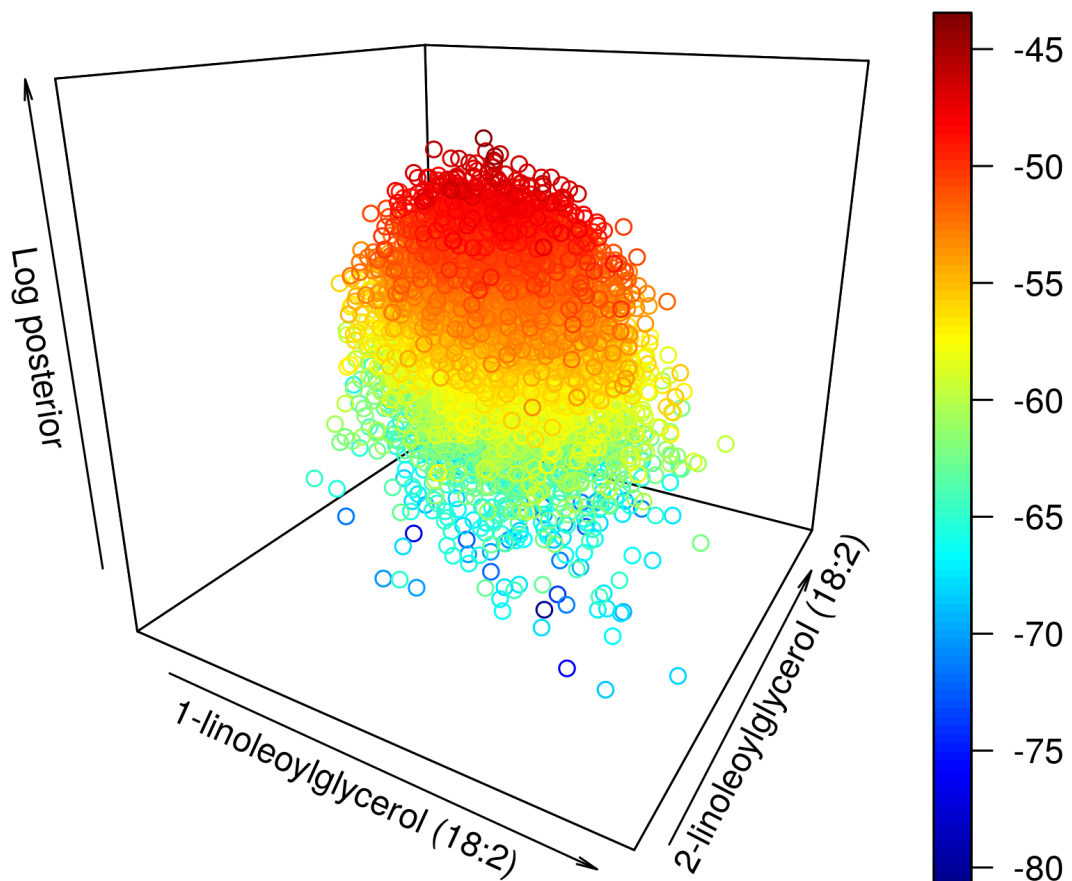


Figure 15. 3-dimensional scatterplot of MCMC samples for both linoleoylglycerols (18:2) with the acyl side chain in different positions (in the same plane), as well as the log-posterior probability of each sampled model.

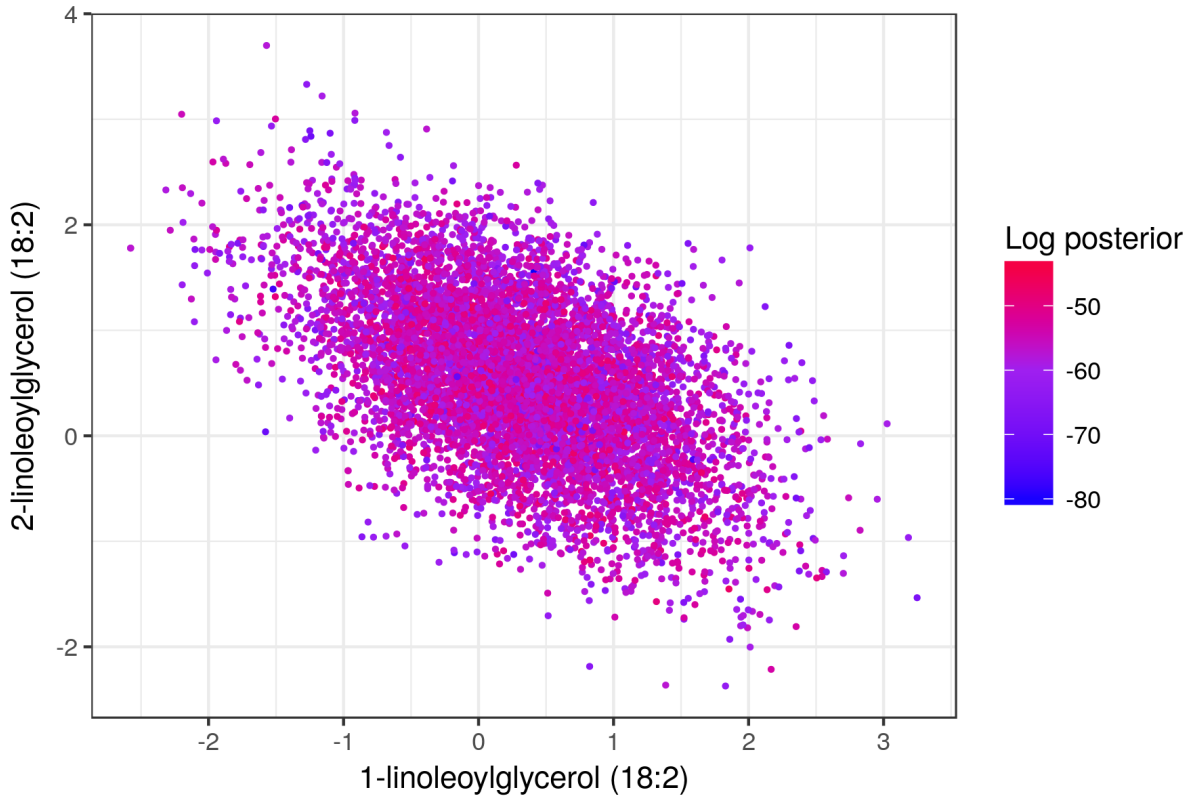


Figure 16. Scatterplot of MCMC samples for both linoleoylglycerols (18:2) with the acyl side chain in different positions colored by log-posterior probability. MCMC sampled models with the greatest log-posterior probability are shown in red, while models with lesser log-posterior probability are shown in blue.

From the simulated posterior distribution of models, a predictive phenotype probability distribution can be generated for each human subject. The MCMC chain for the phenotype probabilities of a selected clinical subject (#14) is presented in Figure 17. The maximum *a posteriori*) group label from both models evaluated properly classified this subject as a stable CAD subject, however in the leave-one-out cross-validation estimation, this subject misclassified as a Thrombotic MI subject. In the models (not-LOOCV estimates), a substantial probability mass was also placed on the posterior probability that this subject was a Non-Thrombotic MI.

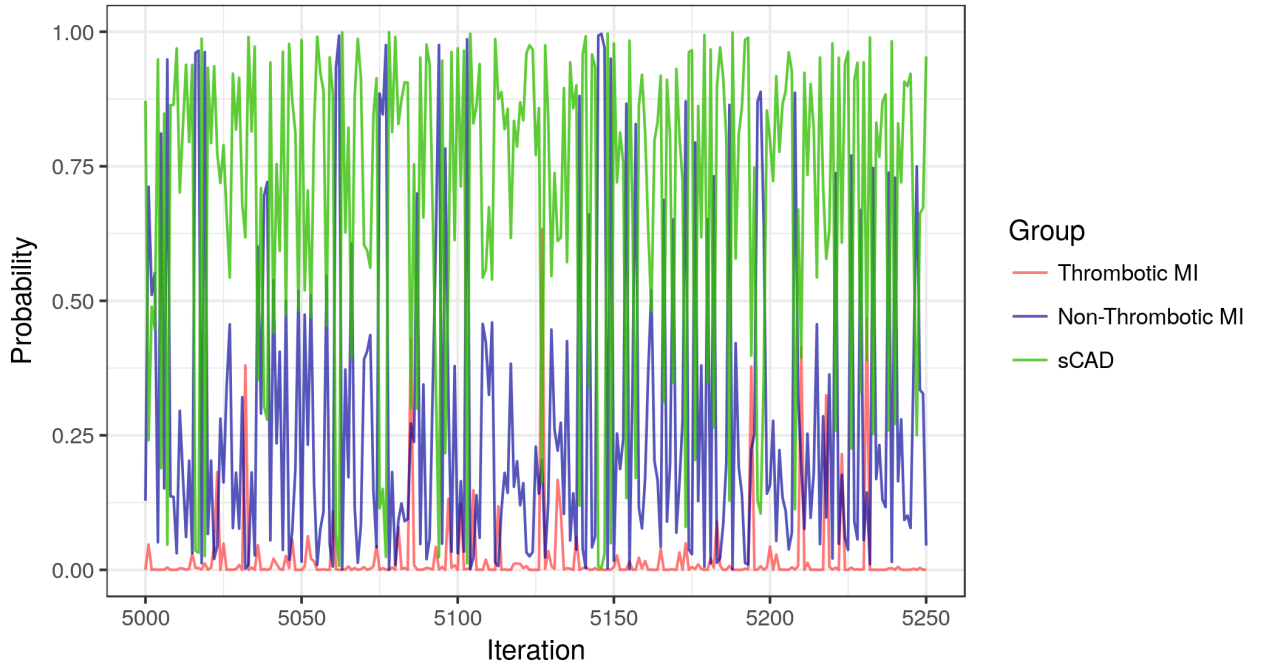


Figure 17. A segment of the first MCMC chain for simulating the posterior distribution of group membership probability estimates for a specific human subject.

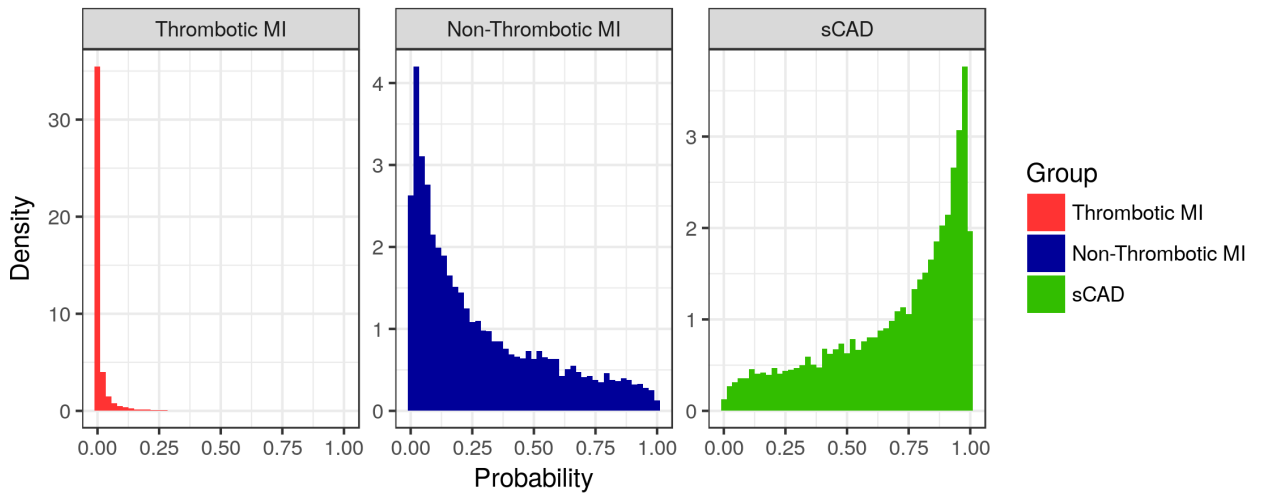


Figure 18. Histograms showing the simulated posterior distribution of group membership probability estimates for a specific human subject.

The LOOCV-estimated confusion matrix is shown in Table 5 for the model fit by maximum likelihood estimation (MLE), Table 6 for the Bayesian model without clinical troponin values, and Table 7 for the Bayesian model with clinical

troponin values. For the Bayesian models, group membership was determined by maximum *a posteriori* estimates. Finally, the raw probability estimates for each human subject from the Bayesian models are shown in Table 8. The LOOCV-estimated sensitivity for detecting and discriminating thrombotic MI was 90.9% for the Bayesian models with and without clinical troponin values. The sensitivity for the model fit by MLE was 81.8%. Without clinical troponin values, the estimated specificity for the Bayesian model was 96.3%, while with clinical troponin values included the estimated specificity improved to 100.0%. The model fit by MLE without clinical troponin values had a specificity of 92.6%. In terms of the detection and discrimination of non-thrombotic MI, the LOOCV-estimated sensitivity was 83.3% for both Bayesian models, and the estimated specificity was 88.5% for both. In terms of the misclassification rate the model estimated by MLE had a rate of 21.1%, while both Bayesian models had a rate of 13.2%. Two stable CAD subjects were classified by both Bayesian models as non-thrombotic MI subjects. Both of these subjects had relatively small estimated probabilities for thrombotic MI. Two non-thrombotic MI subjects were misclassified by both models. The first (subject #26) was classified as stable CAD by both models. The second subject (#27) was first classified as thrombotic MI (model without clinical troponin values), and then classified as stable CAD (model with clinical troponin values). This subject appeared to have nearly equal ($\approx 30\%$) probability mass placed over all three study groups. Finally, one thrombotic MI subject was classified as non-thrombotic MI by both models, however the difference between probability mass for thrombotic MI and non-thrombotic MI was slight (46.3% vs 39.0% in the model without clinical troponin values; 44.8% vs 36.0% for the model with clinical troponin values).

Table 5. Leave-one-out Cross-validation (LOOCV) estimated confusion matrix for the multinomial logistic regression model without clinical troponin values fit by Maximum Likelihood Estimation (MLE). The leftmost column shows the true group membership of the human subjects, while the remaining columns show the predicted probabilities (maximum *a posteriori*).

Group	Predicted		
	sCAD	Thrombotic MI	Non-Thrombotic MI
sCAD	13	0	2
Thrombotic MI	1	9	1
Non-Thrombotic MI	2	2	8

Table 6. Leave-one-out Cross-validation (LOOCV) estimated confusion matrix for Bayesian multinomial logistic regression model without clinical troponin values. The leftmost column shows the true group membership of the human subjects, while the remaining columns show the predicted probabilities (maximum *a posteriori*).

Group	Predicted		
	sCAD	Thrombotic MI	Non-Thrombotic MI
sCAD	13	0	2
Thrombotic MI	0	10	1
Non-Thrombotic MI	1	1	10

Table 7. Leave-one-out Cross-validation (LOOCV) estimated confusion matrix for Bayesian multinomial logistic regression model with clinical troponin values. The leftmost column shows the true group membership of the human subjects, while the remaining columns show the predicted probabilities (maximum *a posteriori*).

Group	Predicted		
	sCAD	Thrombotic MI	Non-Thrombotic MI
sCAD	13	0	2
Thrombotic MI	0	10	1
Non-Thrombotic MI	2	0	10

Table 8. Leave-one-out cross-validation (LOOCV) estimated group probabilities from the Bayesian multinomial logistic regression models for each of human subject in the cohort. M1: Model without clinical troponin values. M2: Model with clinical troponin values. Blue-to-red continuous scale ranks estimated probabilities from low-to-high.

ID	True Group	sCAD		Non-thrombotic MI		Thrombotic MI		Predicted Group	
		M1	M2	M1	M2	M1	M2	M1	M2
1	sCAD	0.998	0.998	0.001	0.002	0.001	0.001	sCAD	sCAD
2	sCAD	0.997	0.995	0.003	0.005	0.000	0.000	sCAD	sCAD
3	sCAD	0.990	0.993	0.010	0.008	0.000	0.000	sCAD	sCAD
4	sCAD	0.983	0.978	0.003	0.004	0.014	0.018	sCAD	sCAD
5	sCAD	0.983	0.981	0.014	0.016	0.004	0.003	sCAD	sCAD
6	sCAD	0.975	0.979	0.020	0.019	0.004	0.002	sCAD	sCAD
7	sCAD	0.975	0.979	0.001	0.001	0.025	0.020	sCAD	sCAD
8	sCAD	0.939	0.950	0.002	0.002	0.059	0.048	sCAD	sCAD
9	sCAD	0.931	0.938	0.068	0.062	0.001	0.001	sCAD	sCAD
10	sCAD	0.875	0.916	0.002	0.001	0.124	0.083	sCAD	sCAD
11	sCAD	0.813	0.881	0.041	0.035	0.146	0.084	sCAD	sCAD
12	sCAD	0.802	0.805	0.037	0.024	0.161	0.171	sCAD	sCAD
13	sCAD	0.786	0.767	0.214	0.233	0.000	0.000	sCAD	sCAD
15	sCAD	0.202	0.262	0.717	0.690	0.081	0.049	Non-Thromb. MI	Non-Thromb. MI
14	sCAD	0.146	0.178	0.832	0.811	0.022	0.011	Non-Thromb. MI	Non-Thromb. MI
16	Non-Thromb. MI	0.007	0.007	0.992	0.993	0.001	0.001	Non-Thromb. MI	Non-Thromb. MI
17	Non-Thromb. MI	0.007	0.004	0.992	0.990	0.001	0.006	Non-Thromb. MI	Non-Thromb. MI
18	Non-Thromb. MI	0.043	0.047	0.956	0.952	0.001	0.001	Non-Thromb. MI	Non-Thromb. MI
19	Non-Thromb. MI	0.023	0.012	0.914	0.721	0.062	0.268	Non-Thromb. MI	Non-Thromb. MI
20	Non-Thromb. MI	0.101	0.105	0.878	0.881	0.021	0.014	Non-Thromb. MI	Non-Thromb. MI
21	Non-Thromb. MI	0.134	0.206	0.838	0.773	0.029	0.022	Non-Thromb. MI	Non-Thromb. MI
22	Non-Thromb. MI	0.002	0.003	0.825	0.844	0.173	0.153	Non-Thromb. MI	Non-Thromb. MI
23	Non-Thromb. MI	0.132	0.162	0.791	0.799	0.078	0.040	Non-Thromb. MI	Non-Thromb. MI
24	Non-Thromb. MI	0.084	0.101	0.739	0.748	0.177	0.151	Non-Thromb. MI	Non-Thromb. MI
25	Non-Thromb. MI	0.327	0.258	0.672	0.739	0.001	0.003	Non-Thromb. MI	Non-Thromb. MI
26	Non-Thromb. MI	0.402	0.438	0.371	0.380	0.228	0.182	sCAD	sCAD
27	Non-Thromb. MI	0.345	0.422	0.257	0.248	0.398	0.330	Thrombotic MI	sCAD
28	Thrombotic MI	0.001	0.008	0.000	0.001	0.999	0.990	Thrombotic MI	Thrombotic MI
29	Thrombotic MI	0.000	0.000	0.001	0.002	0.999	0.998	Thrombotic MI	Thrombotic MI
30	Thrombotic MI	0.000	0.000	0.003	0.006	0.997	0.994	Thrombotic MI	Thrombotic MI
31	Thrombotic MI	0.014	0.042	0.000	0.001	0.986	0.957	Thrombotic MI	Thrombotic MI
32	Thrombotic MI	0.026	0.125	0.000	0.002	0.974	0.873	Thrombotic MI	Thrombotic MI
33	Thrombotic MI	0.038	0.008	0.085	0.150	0.878	0.841	Thrombotic MI	Thrombotic MI
34	Thrombotic MI	0.110	0.004	0.029	0.008	0.861	0.988	Thrombotic MI	Thrombotic MI
35	Thrombotic MI	0.010	0.006	0.208	0.189	0.783	0.806	Thrombotic MI	Thrombotic MI
36	Thrombotic MI	0.013	0.051	0.219	0.403	0.769	0.546	Thrombotic MI	Thrombotic MI
37	Thrombotic MI	0.299	0.089	0.054	0.045	0.647	0.866	Thrombotic MI	Thrombotic MI
38	Thrombotic MI	0.147	0.193	0.463	0.448	0.390	0.360	Non-Thromb. MI	Non-Thromb. MI

7 Discussion

The most notable conclusion from the construction of a Bayesian multinomial logistic regression model is that there is strong evidence that thrombotic MI, non-thrombotic MI, and stable CAD can be discriminated by a non-invasive method using a small set of circulating metabolites. High estimated sensitivity and speci-

ficity was observed for detecting and discriminating thrombotic MI from both non-thrombotic MI and stable CAD. The estimated sensitivity and specificity for detecting and discriminating non-thrombotic MI was lower for non-thrombotic MI. Given the heterogeneity of this etiology of acute MI, this observation is not surprising.

There are many advantages from utilizing a Bayesian model estimation as in this context. First, given a relatively small sample size, maximum likelihood estimation for the multinomial logistic model may not allow for convex optimization. The specification of prior distributions with significant probability mass near zero for regression coefficients allows for the encouragement of sparsity. In addition, complete or quasi-complete separation of classes can occur utilizing maximum likelihood estimation, leading to unstable estimates for regression coefficients. As before, by specifying prior distributions for regression parameters that place significant probability mass near zero, the problem of complete or quasi-separation can be avoided. In general, a Bayesian model with appropriate priors will ensure that the resulting model has not overfit the training data with poor generalizability. A benefit of the Bayesian regression paradigm is the treatment of regression coefficients as random variables. In the Bayesian paradigm, variables that may contribute to the separation of groups, but for which there is much uncertainty can be retained in the model without adversely impacting model performance. Variables with correlated parameter estimates can be determined via the interrogation of samples from the posterior distribution of model coefficients. In our case, the regression coefficients for 1-linoleoylglycerol (18:2) and 2-linoleoylglycerol (18:2) were observed to be highly correlated in MCMC samples. This is not surprising given that these metabolites have the same glycerol moiety, and the same acyl side chain (although it is in different positions). Not surprisingly, the high degree of structural similarity is accompanied by a high degree of correlation in the abundance of each metabolite, and in the estimates of multinomial logistic regression parameters for each.

The frequentist model estimated by MLE exhibited inferior performance than either Bayesian model with respect to misclassification rate / accuracy, sensitivity, and specificity.

CHAPTER IX

BAYESIAN APPROACHES FOR COMPOUND IDENTIFICATION GIVEN LC-MS DATA

1 The challenge of identifying compounds from mass spectrometry data

The identification of compounds from chromatography-coupled mass spectrometry experiments remains a great challenge of untargeted metabolomics [Domingo-Almenara et al., 2017, Uppal et al., 2017, Uppal et al., 2016, Dunn et al., 2012]. Many pieces of information may be available to assist in determining the chemical identity of features detected in such experiments including [Dunn et al., 2012]: the mass-to-charge ratio (m/z) of the ions that have been observed, the fragmentation pattern of either parent or fragment ions, isotopic distribution (e.g. the relative abundance of isotopes can be evaluated by comparing the ratio of ^{13}C to ^{12}C), the observed elution times from separation such as liquid or gas chromatography, and peak shape. Which of these classes of information should be used to identify compounds is largely analytical platform dependent. A further complication of assigning compound identity to the features observed in an experiment is that each compound that elutes from a column and is ionized (such as by electrospray ionization) generates multiple types of ions and adduct ions. In GC-MS analyses the derivatization process is not a uniform process [Halket and Zaikin, 2003]. When electrospray ionization (ESI) is utilized to ionize injected molecules, multiple types of adducted ions may be formed (e.g. sodium adducts) in protonated, non-protonated, or de-protonated form dependent on ESI mode [Dunn et al., 2012]. Additionally, other isotopologue ions may be present [Kuhl et al., 2011].

A system for representing the level of confidence in a compound annotation has been proposed previously as part of the Metabolomics Standards Initiative (MSI) [Sumner et al., 2007]. The first level (Level 1) represents the most confident type of identification and requires matching two “orthogonal” properties of an observed feature with the properties observed from the analysis of an authentic standard under identical analytical conditions. An example of two orthogonal properties for an LC-MS analysis would be observed m/z and retention time or retention index. In an LC-MS/MS experiment observed m/z and MS/MS fragmentation pattern compared to an authentic standard analyzed under identical analytical condition would constitute a Level 1 identification. A Level 2 identification also requires matching on “orthogonal” properties, but in comparison to previously archived data (e.g. databases such as the NIST libraries) as opposed to authentic standards analyzed under identical conditions). A Level 3 identification is more general and requires association of the properties of an observed feature to those of classes of biochemicals as opposed to specific compounds. This framework asserts in a general way that metabolite annotation is a probabilistic process, in which the likelihood of a match from mass feature to compound label depends on both *a priori* knowledge and empirical data. Consequently, a Bayesian approach to compound identification is sensible. In this chapter we describe previous Bayesian approaches for compound identification utilizing LC-MS data. Currently, the annotation of mass features in LC-MS(n) data is predominantly conducted using either matching experimental features to authentic standards analyzed under similar conditions, or by using retention time, m/z , and MS² (or MS[n]) spectra [Daly et al., 2014]. However, there is great utility in a probabilistic method that does not rely on authentic standards (as it is costly to acquire a comprehensive library of standards) and spectral libraries are somewhat incomplete—although there is rapid progress in their growth with such databases as Metlin [Smith et al., 2005, Tautenhahn et al., 2012, Guijas et al., 2018]. In this section, we do not detail Bayesian approaches for compound ID for MS(n) data, as we are unaware of such efforts to

date. Given LC-MS data we describe the assignment of compound labels to m/z features as peak annotation.

2 Utilizing feasible metabolic reactions to generate conditional distributions

Rogers, et al. [Rogers et al., 2009] introduce the use of prior information for peak annotation in terms of conditional probabilities (although they do not explicitly label it as such). They note that an intense monoisotopic peak for a specific compound should be accompanied by isotopic peaks at specific m/z shifts and with specific intensity ratios. For example the natural abundance ratios of ^{12}C and ^{13}C may be known in advance. Consequently, the annotation of a compound label to a m/z peak is more likely if the isotopic peaks are also present and at predictable abundance ratios, than a compound label assignment to a m/z peak that would be “missing” relevant isotopic peaks. In addition, Rogers et al. invoke previous work [Breitling et al., 2006] that argues that there is only a small set of possible chemical reactions with a specific compound as an intermediate; consequently the presence of a specific intermediate in a sample should alter the prior (conditional) belief of the likelihood of a limited number of products that could be generated with that intermediate as a substrate being present in the sample.

Rogers et al. introduce an mass measurement likelihood using the following formulation:

$$p(x_m | z_{cm} = 1, y_c, \gamma) = N\left(\frac{x_m}{y_c} | 1, \gamma^{-1}\right). \quad (56)$$

In this model, the authors use x_m to represent the observed mass of the m^{th} m/z peak, z_{cm} is a random variable for the assignment of the c^{th} molecular formula to the m^{th} peak, y_c to represent a theoretical mass of the c^{th} molecular formula, and γ for the precision of the mass spectrometer. We streamline the notation by

introducing the following notation:

m_i := a mass feature (peak) indexed by integer i , $i \in \{1, 2, \dots, n\}$

x_{m_i} := measured m/z for mass feature m_i

t_{m_i} := retention time for mass feature m_i

\mathbf{r}_{m_i} := vector of isotope ratios associated with a mass feature m_i

ℓ_j := a possible compound label for a mass feature $j \in \{1, 2, \dots, J\}$

x_{ℓ_j} := monoisotopic mass of a compound with label ℓ_j

A_k := a possible adduct $k \in \{1, 2, \dots, K\}$

\mathcal{I} := Indicator matrix with mass features as rows and labels as columns

The mass error model described by Rogers et al. is then:

$$\frac{x_{m_i}}{x_{\ell_j}} \Big| \{I(m_i = \ell_j) = 1\} \sim N(1, \gamma^{-1}). \quad (57)$$

Rogers et al. argue that the conditional probabilities of formula assignments should be predicated on current assignments, and how pairs of molecules can be related by chemical reactions. They present the following form for the conditional probability (rewritten using our notation):

$$p(I(m_i = \ell_j) = 1 | \mathcal{I}, \delta) = \frac{\beta_{ji} + \delta}{N\delta + \sum_{j'i} \beta_{j'i}} \quad (58)$$

where:

$$\beta_{ji} = \mathbf{W}_j \mathcal{I} \mathbf{1} - \mathbf{W}_j \mathcal{I}_i, \quad (59)$$

with \mathbf{W} as an indicator matrix that represents possible biochemical reactions, with $\dim(\mathbf{W}) = N \times N$, and δ is a hyperparameter. Given this formulation β_{ji} represents a count of the number of biochemical transformations from compound labels currently assigned to result in a new annotation. Rogers et al. then state the discrete conditional distribution of possible compound labels for a mass feature as:

$$p\left(\frac{x_{m_i}}{x_{\ell_j}}\right) \propto N(1, \gamma^{-1}) \frac{\beta_{ji} + \delta}{N\delta + \sum_{j'i} \beta_{j'i}}. \quad (60)$$

Rogers et al. in their supplemental material specify a Dirichlet-categorical prior for each set of compounds that could be generated by chemical reactions:

$$p(\boldsymbol{\beta}|\boldsymbol{\delta}) = \frac{\Gamma(\sum_j \delta_j) \prod_j \Gamma(\beta_{ji} + \delta_j)}{\Gamma(\sum_j \beta_{ji} + \delta_j) \prod_j \Gamma(\delta_j)}. \quad (61)$$

Although not explicitly stated, the posterior probability distribution of assignment in the form $\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$ described in Rogers et al. is:

$$\begin{aligned} p(\mathcal{I}|\mathbf{X}) &\propto p(\mathbf{X}|\mathcal{I})p(\mathcal{I}) \\ p(\mathcal{I}|\mathbf{X}) &\propto p(\mathbf{X}|\mathcal{I})p(\boldsymbol{\beta}|\boldsymbol{\delta}) \end{aligned} \quad (62)$$

From this a Gibbs sampler was formulated for simulating the posterior distribution of compound labels for each mass feature.

3 Utilizing isotope patterns in conditional distributions

Rogers et al. [Rogers et al., 2009] additionally discuss a method for incorporating prior information regarding monoisotopic-isotopic peak presence and intensity information. The additional likelihood term described is presented here (in our notation), although this formulation does not provide sufficient detail for understanding the process of incorporating this likelihood term:

$$p(I(m_i = \ell_j) = 1|\omega_{ji}, \delta) = \frac{\omega_{j'i} + \delta}{N\delta + \sum_{j'} \omega_{j'i}}. \quad (63)$$

In this equation, $\omega_{j'i}$ is said to represent the number of possible mass features (already assigned a compound label) that generate an isotopic peak m_i , or possible isotopic peaks that could be associated with a monoisotopic peak. Roger’s et al. describe using heuristic rules such as a relative tolerance for matching a hypothetical isotope-to-monoisotope intensity ratio versus theoretical expectation.

4 “ProbMetab”: Formalization and integration of pathway databases

Silva et al. [Silva et al., 2014] built on the ideas elaborated in [Rogers et al., 2009]. First, Silva et al. restate the conditional probability of compound label to mass feature assignment in a more general form. Before we introduce this conditional

probability, we present the full posterior distribution (which is not stated in their work) in our notation (with \mathbf{Y} representing the vector of all empirically observed attributes such as m/z , retention time, and isotope ratios):

$$\begin{aligned}
 p(\mathcal{I}|\mathbf{Y}) &\propto p(\mathbf{Y}|\mathcal{I})p(\mathcal{I}) \\
 p(\mathcal{I}|\mathbf{Y}) &\propto p_N(\mathbf{Y}|\mathcal{I}) \cdot p_{rt}(\mathbf{Y}|\mathcal{I}) \cdot p_{iso}(\mathbf{Y}|\mathcal{I}) \cdot p(\mathcal{I}),
 \end{aligned}
 \tag{64}$$

where $p_N(\cdot)$ is an error model for the measurement of m/z , $p_{rt}(\cdot)$ is a retention time / index error model, $p_{iso}(\cdot)$ is an isotopic distribution error model, and $p(\mathcal{I})$ is the prior probability of compound label to mass feature assignments. The probability distribution of compound labels for a specific mass feature, conditioned on other assignments, is then:

$$p(I(m_i = \ell_j)|\mathcal{I}^{(-m_i)}) \propto p_N(x_{m_i}|I(m_i = \ell_j)) \cdot p_{rt}(t_{m_i}|\mathcal{I}^{(-m_i)}) \cdot p_{iso}(\mathbf{r}_{m_i}|\mathcal{I}^{(-m_i)}) \cdot p(\mathcal{I}^{(-m_i)}).
 \tag{65}$$

While Rogers et al. [Rogers et al., 2009] had principally focused on the likelihood contribution from m/z measurement error, Silva et al. [Silva et al., 2014] demonstrate how other experimental attributes can be incorporated. While these likelihood terms are mentioned, it is not clear from the authors’ description how these models should be formulated. The authors do describe a reformulation of the m/z measurement error model. Specifically they state the following (in our notation):

$$p_N(x_{m_i}|I(m_i = \ell_j), w) \propto I(x_{\ell_j} \in N_w(x_{m_i})) \left(1 - \Phi \left(\frac{|x_{m_i}/x_{\ell_j} - 1| - \mu_w}{\sigma_w} \right) \right),
 \tag{66}$$

where $N_\varepsilon(x)$ is a neighborhood about the point x with radius ε . The inclusion of the indicator function with radius parameter w was designed to incorporate the belief that beyond a fixed m/z error tolerance, a compound label to mass feature assignment should not have positive probability. As with the Rogers et al. [Rogers et al., 2009] work, Silva et al. [Silva et al., 2014] use the conditional probability distribution to develop a Gibbs sampler. The author’s implement their sampler in the R package *ProbMetab*, which is available from the authors but not via CRAN.

5 Infinite Gaussian mixture models

Before introducing an advancement in Bayesian approaches for LC-MS feature annotation we discuss the Infinite Gaussian Mixture model. We begin with a description of the Dirichlet process mixture model [Ferguson, 1973]:

$$p(y|\mu_1, \mu_2, \dots, \mu_k, s_1, s_2, \dots, s_k, \pi_1, \pi_2, \dots, \pi_k) = \sum_{j=1}^k \pi_j N(\mu_j, s_j^{-1}). \quad (67)$$

In this model, mixture components are indexed by $j = \{1, 2, \dots, k\}$, with component means, μ_j and precisions s_j . Finally the mixing proportions are π_j . The priors for the component means can be specified as $p(\mu_j|\lambda) \sim N(\lambda, r^{-1})$, where in the classically described Dirichlet process mixture model, the priors for the component means are relatively diffuse, with $p(\lambda) \sim N(\mu_y, \sigma_y^2)$ and $p(r) \sim \text{Gamma}(1, \sigma_y^{-2})$. As for the precision parameters, Gamma priors are also specified, that is: $p(s_j|\beta, w) \sim \text{Gamma}(\beta, w^{-1})$, with hyperpriors $p(\beta^{-1}) \sim \text{Gamma}(1, 1)$ and $p(w) \sim \text{Gamma}(1, \sigma_y^2)$. While the previous parameters are involved in modeling cluster attributes (means, precisions), priors are also required for the mixing proportions and ‘‘occupation numbers’’ (number of observations in each cluster). The mixing proportions are assumed to follow a Dirichlet distribution, that is $p(\pi_1, \pi_2, \dots, \pi_k|\alpha) \sim \text{Dirichlet}(\alpha/k, \alpha/k, \dots, \alpha/k)$ where α/k represents a concentration parameter. Through integration and algebraic manipulation, [Rasmussen, 1999] showed that the prior probability of cluster indicators, conditioned on the concentration parameters is:

$$p(c_1, c_2, \dots, c_k|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{j=1}^k \frac{\Gamma(n_j + \alpha/k)}{\Gamma(\alpha/k)}. \quad (68)$$

From this, a conditional probability for a fixed cluster indicator is:

$$p(c_i = j|\mathbf{c}_{(-i)}, \alpha) = \frac{n_{(-i),j} + \alpha/k}{n - 1 + \alpha}. \quad (69)$$

By taking the limit as the number of classes goes to infinity, that is $k \rightarrow \infty$, [Rasmussen, 1999] derived the posterior probabilities for indicators, conditioned on the remaining indicators:

$$p(c_i = j|\mathbf{c}_{(-i)}, \alpha)p(y_i|\mu_j, s_j, \mathbf{c}_{(-i)}) \propto \frac{n_{(-i),j}}{n - 1 + \alpha} \sqrt{s_j} \exp(-s_j(y_i - \mu_j)^2/2). \quad (70)$$

6 Bayesian clustering of mass features

The work of Daly [Daly et al., 2014], et al. from the same group as [Rogers et al., 2009], represents a significant paradigm shift in Bayesian approaches for LC-MS mass feature annotation. The novel idea described in that work is to consider both monoisotopic peaks, isotopic peaks, and adduct peaks as being members of a cluster. Daly et al. [Daly et al., 2014], use the notation $\mathbf{d}_n = (x_n, w_n, r_n)$ to refer to the experimental attributes of a single mass feature, where x_n corresponds to the m/z value, w_n corresponds to the intensity of the peak at that m/z value, and r_n refers to the retention time of the feature. They then index the clusters by k and define an indicator variable, v_{nkai} for cluster membership. The subscripts of this indicator variable represent the index of the mass feature, n ; the cluster, k ; the adduct, a ; and the isotope i . The number of peaks that are in a cluster is then c_k where:

$$c_{nk} = \sum_{a=1}^A \sum_{i=1}^I v_{nkai} \quad (71)$$

although in the Daly article [Daly et al., 2014], it appears this quantity is labeled differently. In terms of the data likelihood terms, a mass term, an intensity term, and a retention time term are stated.

The MetAssign algorithm proceeds by first initializing random clustering. This clustering assumes an infinite Gaussian mixture model, as discussed above. For each cluster, mass features are assigned labels and posterior probabilities for current cluster membership and data likelihoods are computed. At this point the probability of new proposals is then computed. If a proposal is accepted then metabolite, adduct, and isotope data likelihoods are re-evaluated. This procedure can proceed until labels of mass feature to cluster membership and compound labels stabilize. The likelihoods utilized are described below.

The mass term supposes the following likelihood:

$$p(x_n | v_{nkai} = 1) = N(\log(x_n) | \log(y_{\phi_k ai}), \zeta^{-1}). \quad (72)$$

This assumes that on a log-scale the difference between the observed m/z and the expected given the compound, adduct, and isotope assignment to the peak is normally distributed with an instrument-specific precision parameter ζ . The intensity likelihood is much more involved and is stated as:

$$p(w_n | v_{nkai} = 1) = N(\beta_{\phi_k ai} \lambda_*, \kappa^{-1} + \beta_{\phi_k ai}^2 \kappa_*^{-1}). \quad (73)$$

According to the authors [Daly et al., 2014], this arises from considering individual peak intensities given assignments as $w_k \sim N(\beta_{\phi_k ai} \lambda_{ka}, \kappa^{-1})$, where $\beta_{\phi_k ai}$ represents the expected proportion of the total intensity considering isotope distributions and κ is a precision parameter. They then place a prior distribution for the total intensity, that is: $\lambda_{ka} \sim N(\lambda_0, \kappa_0^{-1})$. For the retention time likelihood it is assumed that:

$$p(r_n | v_{nkai} = 1) = N(r_n | \mu_*, \delta_*^{-1} + \gamma^{-1}). \quad (74)$$

where μ_* is drawn from a normal distribution with mean set to the mean of the retention times observed and γ^{-1} is a hyper parameter specified by the authors.

The authors [Daly et al., 2014] evaluate their method using data from a mixture of chemical standards. By utilizing a database of decoy compounds, the authors evaluate the precision (positive predictive value) and recall (sensitivity) of their method especially in comparison to mzMatch and CAMERA [Kuhl et al., 2011], and observe substantial improvement over these methods.

CHAPTER X

CONCLUSION, DISCUSSION, AND FUTURE DIRECTIONS

1 Summary

Collectively the current work covers three subtopics in the field of Metabolomics: (1) making higher order inferences in metabolomics using untargeted mass spectrometry data, (2) developing diagnostic models for discriminating disease phenotypes, and (3) determining the identity of compounds in LC-MS data. For the first two subtopics we have introduced novel methodology, while we have only provided a review of a class of methods in the third. What unites these three disparate aims into a cohesive work is the use of Bayesian statistical methodology for integrating extra-experimental scientific knowledge for solving high-dimensional challenges in metabolomics.

To introduce our novel methodology for making higher order inferences in metabolomics (Chapter VI) and to apply this methodology to characterizing the probabilistic interaction of metabolites given stable heart disease (Chapter VII) we provided some preliminary background. In Chapter II we provided a cursory overview of metabolism, the link between metabolism and multiple diseases, and the use of the techniques of metabolomics for studying metabolism and human diseases. In Chapter III we provided an introduction to Bayesian reasoning and in Chapter IV we introduced Gaussian Graphical Models (GGMs) and both frequentist and Bayesian approaches for the estimation of such models. These preliminaries were necessary foundations for developing our novel methodology for using molecular structure to generate informative priors for estimating GGMs for making higher order inferences in metabolomics. This section culminated in a plasma

metabolite interactome for stable heart disease (Chapter VII). In this section we highlighted how higher order inferences can be made by focusing on the primary bile acid cholate. We focused on cholate as dysregulated bile acid metabolism is a component of atherosclerosis, and thus heart disease. Interestingly, we observed that our methodology balanced both the molecular structure prior information as well as the empirical data. Relatively low penalization of edges between cholate and steroids hormones, and cholate and secondary bile acids was observed, however the posterior distribution showed higher posterior probability of strong edges between cholate and secondary bile acids.

In the second part of the current work, we focused on developing a statistical classifier capable of discriminating between three phenotypes: stable coronary artery disease (CAD), non-thrombotic myocardial infarction (MI), and thrombotic MI (Chapter VIII). We constructed two models, the first based on metabolite abundances only, and the second with the inclusion of clinical troponin values as these values would be available at presentation as well. We observed that a Bayesian multinomial logistic regression model demonstrated positive performance characteristics in discriminating between the phenotypes. Both Bayesian models performed better than the frequentist model with the same variables fit by maximum likelihood estimation. We hypothesize that the superior performance of the Bayesian models is due to a reduced likelihood of overfitting.

In the final section (Chapter IX) we discuss Bayesian approaches for compound identification given LC-MS data. The first two methods that we discussed utilized a Dirichlet-categorical prior over possible compound label to mass feature assignments. The third method relied instead on Gaussian mixture modeling over mass features. While only three Bayesian approaches for compound identification have been developed to date (to our knowledge), these methods have increased in sophistication over time. The first method described primarily included a likelihood term based in the measurement error of mass-to-charge ratio (although retention time error and isotope profile error were discussed). The prior probability dis-

tribution utilized possible transformations that could account for the dependence between compounds observed. The next method was largely similar, with a more complex measurement error model and the ability to use metabolite databases (e.g. KEGG) for generating priors. Finally, the third method recast the identification problem as a clustering followed by identification problem, which is sensible given that the mass features occur in clusters comprised of one compound in multiple adduct / ionic forms.

2 Future directions: Bayesian interactome models

Through the development and utilization of our novel methodology for estimating metabolite interactomes future directions have been suggested which will be pursued. First, the prior distributions of the concentration matrix parameters are continuous with all real numbers as the support. As a result, the posterior probability density for each concentration matrix parameter has the same support and the probability of a concentration matrix parameter being equal to zero is zero. Consequently, edge selection, or the process of setting some concentration matrix parameters equal to zero must happen after model selection. This suggests that a hierarchical model or a model that allows probability mass at zero would be beneficial. In the first case, a hierarchical model could include a Bernoulli component that models whether an edge between metabolites is present or absent, with a continuous prior distribution given that an edge is present. In the second case, a “slab and spike” prior could be utilized, which has a point mass at zero and more conventional continuous tails.

3 Future directions: Diagnostic modeling

The classification model described in the current work is not yet applicable for use in clinical practice. Ignoring, for the sake of discussion, the regulatory approval process, multiple steps remain in furthering this methodology. Firstly, the metabolite abundances described in the current work are relative abundances from

a data dependent acquisition. Consequently, the measurement scale from the current relative abundances would not be preserved in future measurements of these same metabolites. Further, the relationship between relative abundances and concentrations are non-linear and depend greatly on the analytical platform. As a result, we applied for, and received a grant to develop, in collaboration with Dr. Tong Shen and Dr. Oliver Fiehn of the University of California at Davis a targeted analysis of the metabolites that were utilized in this analysis (as well other metabolites observed to be important. The development of this multiple reaction monitoring assay with stable isotope dilution for targeted absolute quantification has been completed and currently the same set of human samples as discussed in the current work are being analyzed. Additionally, a data-dependent LC-MS proteomics analysis has been conducted by Dr. Qin Fu and Dr. Jennifer Van Eyk at Cedars-Sinai and Dr. Scott Peterman at Thermo Fisher Scientific; we are currently analyzing the results of this data acquisition. The acquisition of proteomics data from the same sample will allow for the development of a classifier model utilizing both proteomic and metabolite abundances.

In terms of statistical methodology, the set of metabolites that were included in the multinomial logistic regression models were drawn from previous work. An important future direction that we are currently pursuing is conducting the feature selection and model fitting process in conjunction (or at least using the same general methodological approach). Specifically, we are evaluating utilizing different prior probability distributions such as the Horseshoe prior and the slab and spike prior for conducting feature selection.

An important next step in furthering a classifier capable of detecting and discriminating myocardial infarction types will be the validation of this classifier in independent cohorts. To date, 168 human subjects presenting with acute coronary syndrome and 27 subjects with stable coronary artery disease have been enrolled in a new cohort by the Atherosclerosis / Atherothrombosis Research Laboratory (AARL) at the University of Louisville towards this and other aims. An evaluation

of the performance characteristics of the classifier developed utilizing the cohort described in the current work will take place in this new cohort. Finally, while this validation is an important step, the classifier should also be evaluated in a cohort of “all comers” presenting to the emergency department to evaluate the sensitivity and specificity for detecting and discriminating MI type in a more general setting. For example, while the cohort described in the current work has subjects with stable coronary artery disease and MI, an “all comers” cohort would also have patients presenting with symptoms such as chest pain that may be consistent with MI, but who are not experiencing an MI. Finally, once a diagnostic model (or models) have been fully developed and validated in independent cohorts, the development of a single test capable of being conducted within a clinical laboratory can occur.

REFERENCES

- [Agresti, 2013] Agresti, A. (2013). *Categorical data analysis*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 3rd edition.
- [Aguet et al., 2017] Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., Segr, A. V., Strober, B. J., Zappala, Z., Cummings, B. B., Gelfand, E. T., Hadley, K., Huang, K. H., Lek, M., Li, X., Nedzel, J. L., Nguyen, D. Y., Noble, M. S., Sullivan, T. J., Tukiainen, T., MacArthur, D. G., Getz, G., Addington, A., Guan, P., Koester, S., Little, A. R., Lockhart, N. C., Moore, H. M., Rao, A., Struewing, J. P., Volpi, S., Brigham, L. E., Hasz, R., Hunter, M., Johns, C., Johnson, M., Kopen, G., Leinweber, W. F., Lonsdale, J. T., McDonald, A., Mestichelli, B., Myer, K., Roe, B., Salvatore, M., Shad, S., Thomas, J. A., Walters, G., Washington, M., Wheeler, J., Bridge, J., Foster, B. A., Gillard, B. M., Karasik, E., Kumar, R., Miklos, M., Moser, M. T., Jewell, S. D., Montroy, R. G., Rohrer, D. C., Valley, D., Mash, D. C., Davis, D. A., Sobin, L., Barcus, M. E., Branton, P. A., Abell, N. S., Balliu, B., Delaneau, O., Frsard, L., Gamazon, E. R., Garrido-Martn, D., Gewirtz, A. D. H., Gliner, G., Gloudemans, M. J., Han, B., He, A. Z., Hormozdiari, F., Li, X., Liu, B., Kang, E. Y., McDowell, I. C., Ongen, H., Palowitch, J. J., Peterson, C. B., Quon, G., Ripke, S., Saha, A., Shabalina, A. A., Shimko, T. C., Sul, J. H., Teran, N. A., Tsang, E. K., Zhang, H., Zhou, Y.-H., Bustamante, C. D., Cox, N. J., Guig, R., et al. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213.

- [Andrews and Mallows, 1974] Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102.
- [Arbab-Zadeh and Fuster, 2015] Arbab-Zadeh, A. and Fuster, V. (2015). The myth of the "vulnerable plaque": transitioning from a focus on individual lesions to atherosclerotic disease burden for coronary artery disease risk assessment. *J Am Coll Cardiol*, 65(8):846–55.
- [Arbab-Zadeh et al., 2012] Arbab-Zadeh, A., Nakano, M., Virmani, R., and Fuster, V. (2012). Acute coronary events. *Circulation*, 125(9):1147–1156.
- [Atay-Kayis and Massam, 2005] Atay-Kayis, A. and Massam, H. (2005). A monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models. *Biometrika*, 92(2):317–335.
- [Banerjee et al., 2008] Banerjee, O., Ghaoui, L. E., and d’Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516.
- [Barupal and Fiehn, 2017] Barupal, D. K. and Fiehn, O. (2017). Chemical similarity enrichment analysis (chemrich) as alternative to biochemical pathway mapping for metabolomic datasets. *Scientific Reports*, 7(1).
- [Barupal et al., 2012] Barupal, D. K., Haldiya, P. K., Wohlgemuth, G., Kind, T., Kothari, S. L., Pinkerton, K. E., and Fiehn, O. (2012). Metamapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics*, 13(1):99.
- [Baynes and Dominiczak, 2014] Baynes, J. W. and Dominiczak, M. H. (2014). *Medical Biochemistry*. Elsevier Health Sciences.
- [Benjamin et al., 2017] Benjamin, E. J., Blaha, M. J., Chiuve, S. E., Cushman, M., Das, S. R., Deo, R., de Ferranti, S. D., Floyd, J., Fornage, M., Gillespie, C.,

- Isasi, C. R., Jimnez, M. C., Jordan, L. C., Judd, S. E., Lackland, D., Lichtman, J. H., Lisabeth, L., Liu, S., Longenecker, C. T., Mackey, R. H., Matsushita, K., Mozaffarian, D., Mussolino, M. E., Nasir, K., Neumar, R. W., Palaniappan, L., Pandey, D. K., Thiagarajan, R. R., Reeves, M. J., Ritchey, M., Rodriguez, C. J., Roth, G. A., Rosamond, W. D., Sasson, C., Towfighi, A., Tsao, C. W., Turner, M. B., Virani, S. S., Voeks, J. H., Willey, J. Z., Wilkins, J. T., Wu, J. H. Y., Alger, H. M., Wong, S. S., and Muntner, P. (2017). Heart disease and stroke statistics–2017 update: A report from the american heart association. *Circulation*, 135(10):e146–e603.
- [Bentzon et al., 2014] Bentzon, J. F., Otsuka, F., Virmani, R., and Falk, E. (2014). Mechanisms of plaque formation and rupture. *Circulation Research*, 114(12):1852–1866.
- [Betancourt, 2017] Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *ArXiv*.
- [Bories and Leitinger, 2017] Bories, G. F. P. and Leitinger, N. (2017). Macrophage metabolism in atherosclerosis. *FEBS Letters*, 591(19):3042–3060.
- [Breitling et al., 2006] Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M. L., and Barrett, M. P. (2006). Ab initio prediction of metabolic networks using fourier transform mass spectrometry data. *Metabolomics*, 2(3):155–164.
- [Brown et al., 2017] Brown, A. A., Viuela, A., Delaneau, O., Spector, T. D., Small, K. S., and Dermitzakis, E. T. (2017). Predicting causal variants affecting expression by using whole-genome sequencing and rna-seq from multiple human tissues. *Nature Genetics*, 49(12):1747–1751.
- [Bruntz et al., 2017] Bruntz, R. C., Lane, A. N., Higashi, R. M., and Fan, T. W. M. (2017). Exploring cancer metabolism using stable isotope-resolved metabolomics (sirm). *Journal of Biological Chemistry*, 292(28):11601–11609.

- [Carlisle et al., 2016] Carlisle, S. M., Trainor, P. J., Yin, X., Doll, M. A., Stepp, M. W., States, J. C., Zhang, X., and Hein, D. W. (2016). Untargeted polar metabolomics of transformed mda-mb-231 breast cancer cells expressing varying levels of human arylamine n-acetyltransferase 1. *Metabolomics*, 12(7).
- [Carpenter et al., 2017] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).
- [Carvalho et al., 2010] Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.
- [Casella and Berger, 2002] Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition.
- [Chait and Eckel, 2016] Chait, A. and Eckel, R. H. (2016). Lipids, lipoproteins, and cardiovascular disease: Clinical pharmacology now and in the future. *The Journal of Clinical Endocrinology and Metabolism*, 101(3):804–814.
- [Chen and Reynolds, 2002] Chen, X. and Reynolds, C. H. (2002). Performance of similarity measures in 2d fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *Journal of Chemical Information and Computer Sciences*, 42(6):1407–1414.
- [Cheng and Lenkoski, 2012] Cheng, Y. and Lenkoski, A. (2012). Hierarchical gaussian graphical models: Beyond reversible jump. *Electronic Journal of Statistics*, 6(0):2309–2331.
- [Chong et al., 2018] Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D. S., and Xia, J. (2018). Metaboanalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research*, 46(W1):W486–W494.

- [Collins, 2003] Collins, F. S. (2003). The human genome project: Lessons from large-scale biology. *Science*, 300(5617):286–290.
- [Dallmann et al., 2012] Dallmann, R., Viola, A. U., Tarokh, L., Cajochen, C., and Brown, S. A. (2012). The human circadian metabolome. *Proceedings of the National Academy of Sciences*, 109(7):2625–2629.
- [Daly et al., 2014] Daly, R., Rogers, S., Wandy, J., Jankevics, A., Burgess, K. E. V., and Breitling, R. (2014). Metassign: probabilistic annotation of metabolites from lcms data using a bayesian clustering approach. *Bioinformatics*, 30(19):2764–2771.
- [Dawid and Lauritzen, 1993] Dawid, A. P. and Lauritzen, S. L. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317.
- [DeFilippis et al., 2015] DeFilippis, A. P., Chernyavskiy, I., Amraotkar, A. R., Trainor, P. J., Kothari, S., Ismail, I., Hargis, C. W., Korley, F. K., Leibundgut, G., Tsimikas, S., Rai, S. N., and Bhatnagar, A. (2015). Circulating levels of plasminogen and oxidized phospholipids bound to plasminogen distinguish between atherothrombotic and non-atherothrombotic myocardial infarction. *J Thromb Thrombolysis*.
- [DeFilippis et al., 2017] DeFilippis, A. P., Trainor, P. J., Hill, B. G., Amraotkar, A. R., Rai, S. N., Hirsch, G. A., Rouchka, E. C., and Bhatnagar, A. (2017). Identification of a plasma metabolomic signature of thrombotic myocardial infarction that is distinct from non-thrombotic myocardial infarction and stable coronary artery disease. *Plos One*, 12(4):e0175591.
- [Delude, 2015] Delude, C. M. (2015). Deep phenotyping: The details of disease. *Nature*, 527(7576):S14–S15.
- [Desai et al., 2017] Desai, M. S., Mathur, B., Eblimit, Z., Vasquez, H., Taegtmeier, H., Karpen, S. J., Penny, D. J., Moore, D. D., and Anakk, S. (2017). Bile

- acid excess induces cardiomyopathy and metabolic dysfunctions in the heart. *Hepatology*, 65(1):189–201.
- [Diebolt and Robert, 1994] Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375.
- [Dobra and Lenkoski, 2011] Dobra, A. and Lenkoski, A. (2011). Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993.
- [Dobra et al., 2011] Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433.
- [Domingo-Almenara et al., 2017] Domingo-Almenara, X., Montenegro-Burke, J. R., Benton, H. P., and Siuzdak, G. (2017). Annotation: A computational solution for streamlining metabolomics analysis. *Analytical Chemistry*, 90(1):480–489.
- [Dunn et al., 2012] Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., and Viant, M. R. (2012). Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*, 9(S1):44–66.
- [Dunn and Hankemeier, 2013] Dunn, W. B. and Hankemeier, T. (2013). Mass spectrometry and metabolomics: past, present and future. *Metabolomics*, 9(S1):1–3.
- [Dunn et al., 2014] Dunn, W. B., Lin, W., Broadhurst, D., Begley, P., Brown, M., Zelena, E., Vaughan, A. A., Halsall, A., Harding, N., Knowles, J. D., Francis-McIntyre, S., Tseng, A., Ellis, D. I., OHagan, S., Aarons, G., Benjamin, B.,

- Chew-Graham, S., Moseley, C., Potter, P., Winder, C. L., Potts, C., Thornton, P., McWhirter, C., Zubair, M., Pan, M., Burns, A., Cruickshank, J. K., Jayson, G. C., Purandare, N., Wu, F. C. W., Finn, J. D., Haselden, J. N., Nicholls, A. W., Wilson, I. D., Goodacre, R., and Kell, D. B. (2014). Molecular phenotyping of a uk population: defining the human serum metabolome. *Metabolomics*, 11(1):9–26.
- [Eddelbuettel, 2013] Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York. ISBN 978-1-4614-6867-7.
- [Eddelbuettel and François, 2011] Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.
- [Eddelbuettel and Sanderson, 2014] Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063.
- [Everitt, 2004] Everitt, B. S. (2004). Mixture distributions. *Encyclopedia of statistical sciences*, 7.
- [Falk, 2006] Falk, E. (2006). Pathogenesis of atherosclerosis. *Journal of the American College of Cardiology*, 47(8):C7–C12.
- [Fan et al., 2009] Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, 3(2):521–541.
- [Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.
- [Fan et al., 2012] Fan, T. W. M., Lorkiewicz, P. K., Sellers, K., Moseley, H. N. B., Higashi, R. M., and Lane, A. N. (2012). Stable isotope-resolved

- metabolomics and applications for drug development. *Pharmacology & Therapeutics*, 133(3):366–391.
- [Fan et al., 2016] Fan, Y., Li, Y., Chen, Y., Zhao, Y.-J., Liu, L.-W., Li, J., Wang, S.-L., Alolga, R. N., Yin, Y., Wang, X.-M., Zhao, D.-S., Shen, J.-H., Meng, F.-Q., Zhou, X., Xu, H., He, G.-P., Lai, M.-D., Li, P., Zhu, W., and Qi, L.-W. (2016). Comprehensive metabolomic characterization of coronary artery diseases. *Journal of the American College of Cardiology*, 68(12):1281–1293.
- [Ferguson, 1973] Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.
- [Ferrebee and Dawson, 2015] Ferrebee, C. B. and Dawson, P. A. (2015). Metabolic effects of intestinal absorption and enterohepatic cycling of bile acids. *Acta Pharmaceutica Sinica B*, 5(2):129–134.
- [Friedman et al., 2007] Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.
- [Frolkis et al., 2010] Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., Liu, P., Gautam, B., Ly, S., Guo, A. C., Xia, J., Liang, Y., Shrivastava, S., and Wishart, D. S. (2010). Smpdb: The small molecule pathway database. *Nucleic Acids Research*, 38(Supplement 1):D480–D487.
- [Gale, 2010] Gale, A. J. (2010). Continuing education course #2: Current understanding of hemostasis. *Toxicologic Pathology*, 39(1):273–280.
- [Gao et al., 2015] Gao, X., Jia, M., Zhang, Y., Breitling, L. P., and Brenner, H. (2015). Dna methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of dna methylation studies. *Clinical Epigenetics*, 7(1).
- [Garca-Caaveras et al., 2012] Garca-Caaveras, J. C., Donato, M. T., Castell, J. V., and Lahoz, A. (2012). Targeted profiling of circulating and hepatic bile acids

- in human, mouse, and rat using a uplc-mrm-ms-validated method. *Journal of Lipid Research*, 53(10):2231–2241.
- [Gelman, 2006] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534.
- [Gelman et al., 2004] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Texts in statistical science. Chapman & Hall / CRC, Boca Raton, Florida.
- [Gimbrone and Garca-Cardea, 2016] Gimbrone, M. A. and Garca-Cardea, G. (2016). Endothelial cell dysfunction and the pathobiology of atherosclerosis. *Circulation Research*, 118(4):620–636.
- [Giudici and Spelta, 2016] Giudici, P. and Spelta, A. (2016). Graphical network models for international financial flows. *Journal of Business & Economic Statistics*, 34(1):128–138.
- [Goeman and Buhlmann, 2007] Goeman, J. J. and Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.
- [Goeman et al., 2003] Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2003). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99.
- [Gould, 1951] Gould, R. G. (1951). Lipid metabolism and atherosclerosis. *The American Journal of Medicine*, 11(2):209–227.
- [Gowda and Djukovic, 2014] Gowda, G. A. N. and Djukovic, D. (2014). Overview of mass spectrometry-based metabolomics: Opportunities and challenges. *Mass Spectrometry in Metabolomics*, 1198:3–12.

- [Gross, 2017] Gross, J. H. (2017). *Mass spectrometry : a textbook*. Springer Berlin Heidelberg, New York, NY, 3rd edition. edition.
- [Guijas et al., 2018] Guijas, C., Montenegro-Burke, J. R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., Koellensperger, G., Huan, T., Uritboonthai, W., Aisporna, A. E., Wolan, D. W., Spilker, M. E., Benton, H. P., and Siuzdak, G. (2018). Metlin: A technology platform for identifying knowns and unknowns. *Analytical Chemistry*, 90(5):3156–3164.
- [Halket and Zaikin, 2003] Halket, J. M. and Zaikin, V. G. (2003). Derivatization in mass spectrometry1. silylation. *European Journal of Mass Spectrometry*, 9(1):1–21.
- [Hamburg and Collins, 2010] Hamburg, M. A. and Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304.
- [Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics,. Springer, New York, NY, 2nd edition.
- [Higashi et al., 2014] Higashi, R. M., Fan, T. W. M., Lorkiewicz, P. K., Moseley, H. N. B., and Lane, A. N. (2014). Stable isotope-labeled tracers for metabolic pathway elucidation by gc-ms and ft-ms. *Methods in molecular biology*, 1198:147–167.
- [Hoff, 2009] Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer, New York, NY.
- [Hoffman and Monroe, 2007] Hoffman, M. and Monroe, D. M. (2007). Coagulation 2006: A modern view of hemostasis. *Hematology/Oncology Clinics of North America*, 21(1):1–11.
- [Hoffman and Gelman, 2014] Hoffman, M. D. and Gelman, A. (2014). The no-urn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

- [Hofmann et al., 2010] Hofmann, A. F., Hagey, L. R., and Krasowski, M. D. (2010). Bile salts of vertebrates: structural variation and possible evolutionary significance. *Journal of Lipid Research*, 51(2):226–246.
- [Johnson et al., 2016] Johnson, C. H., Ivanisevic, J., and Siuzdak, G. (2016). Metabolomics: beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17(7):451–459.
- [Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30.
- [Kanehisa et al., 2016] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Res*, 44(D1):D457–62.
- [Kasper, 2015] Kasper, D. L. (2015). *Harrison’s principles of internal medicine*. McGraw Hill Education, New York, 19th edition edition.
- [Keating and El-Osta, 2015] Keating, S. T. and El-Osta, A. (2015). Epigenetics and metabolism. *Circulation Research*, 116(4):715–736.
- [Khurana et al., 2011] Khurana, S., Raufman, J.-P., and Pallone, T. L. (2011). Bile acids regulate cardiovascular function. *Clinical and Translational Science*, 4(3):210–218.
- [Khurana et al., 2005] Khurana, S., Yamada, M., Wess, J., Kennedy, R. H., and Raufman, J.-P. (2005). Deoxycholytaurine-induced vasodilation of rodent aorta is nitric oxide- and muscarinic m3 receptor-dependent. *European Journal of Pharmacology*, 517(1-2):103–110.
- [Koller and Friedman, 2009] Koller, D. and Friedman, N. (2009). *Probabilistic graphical models : principles and techniques*. Adaptive computation and machine learning. MIT Press, Cambridge, MA.

- [Krumstiek et al., 2011] Krumstiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, 5(1):21.
- [Krycer et al., 2017] Krycer, J. R., Yugi, K., Hirayama, A., Fazakerley, D. J., Quek, L.-E., Scalzo, R., Ohno, S., Hodson, M. P., Ikeda, S., Shoji, F., Suzuki, K., Domanova, W., Parker, B. L., Nelson, M. E., Humphrey, S. J., Turner, N., Hoehn, K. L., Cooney, G. J., Soga, T., Kuroda, S., and James, D. E. (2017). Dynamic metabolomics reveals that insulin primes the adipocyte for glucose metabolism. *Cell Reports*, 21(12):3536–3547.
- [Kuhl et al., 2011] Kuhl, C., Tautenhahn, R., Bittcher, C., Larson, T. R., and Neumann, S. (2011). Camera: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, 84(1):283–289.
- [Lam and Fan, 2009] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278.
- [Lei et al., 2011] Lei, Z., Huhman, D. V., and Sumner, L. W. (2011). Mass spectrometry strategies in metabolomics. *Journal of Biological Chemistry*, 286(29):25435–25442.
- [Lewandowski et al., 2009] Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.
- [Liu et al., 2008] Liu, X., Yu, X., Zack, D. J., Zhu, H., and Qian, J. (2008). Tiger: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, 9(1):271.
- [Liu et al., 2016] Liu, Y., Prentice, K., Eversley, J., Hu, C., Batchuluun, B., Leavey, K., Hansen, J., Wei, D., Cox, B., Dai, F., Jia, W., and Wheeler, M.

(2016). Rapid elevation in cmpf may act as a tipping point in diabetes development. *Cell Reports*, 14(12):2889–2900.

[Lonsdale et al., 2013] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N. J., Nicolae, D. L., Gamazon, E. R., Im, H. K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E. T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585.

[Lusis, 2000] Lusis, A. J. (2000). Atherosclerosis. *Nature*, 407(6801):233–241.

[Martincorena and Campbell, 2015] Martincorena, I. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489.

[Meinshausen and Bhlmann, 2006] Meinshausen, N. and Bhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

[Meissner et al., 2013] Meissner, M., Wolters, H., de Boer, R. A., Havinga, R., Boverhof, R., Bloks, V. W., Kuipers, F., and Groen, A. K. (2013). Bile acid

- sequestration normalizes plasma cholesterol and reduces atherosclerosis in hypercholesterolemic mice. no additional effect of physical activity. *Atherosclerosis*, 228(1):117–123.
- [Mitchell et al., 2014] Mitchell, J. M., Fan, T. W. M., Lane, A. N., and Moseley, H. N. B. (2014). Development and in silico evaluation of large-scale metabolite identification methods using functional group detection for metabolomics. *Frontiers in Genetics*, 5.
- [Mootha et al., 2003] Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–73.
- [Moriya et al., 2017] Moriya, T., Satomi, Y., Murata, S., Sawada, H., and Kobayashi, H. (2017). Effect of gut microbiota on host whole metabolome. *Metabolomics*, 13(9).
- [Nagana Gowda and Raftery, 2016] Nagana Gowda, G. A. and Raftery, D. (2016). Recent advances in nmr-based metabolomics. *Analytical Chemistry*, 89(1):490–510.
- [Neal, 2011] Neal, R. M. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11).
- [Newby et al., 2012] Newby, L. K., Jesse, R. L., Babb, J. D., Christenson, R. H., De Fer, T. M., Diamond, G. A., Fesmire, F. M., Geraci, S. A., Gersh, B. J., Larsen, G. C., Kaul, S., McKay, C. R., Philippides, G. J., Weintraub, W. S., Harrington, R. A., Bhatt, D. L., Anderson, J. L., Bates, E. R., Bridges, C. R., Eisenberg, M. J., Ferrari, V. A., Fisher, J. D., Garcia, M. J., Gardner, T. J.,

- Gentile, F., Gilson, M. F., Hernandez, A. F., Hlatky, M. A., Jacobs, A. K., Kaul, S., Linderbaum, J. A., Moliterno, D. J., Mukherjee, D., Rosenson, R. S., Stein, J. H., Weitz, H. H., and Wesley, D. J. (2012). Accf 2012 expert consensus document on practical clinical considerations in the interpretation of troponin elevations. *Journal of the American College of Cardiology*, 60(23):2427–2463.
- [Newgard, 2017] Newgard, C. B. (2017). Metabolomics and metabolic diseases: Where do we stand? *Cell Metabolism*, 25(1):43–56.
- [Nicholson and Lindon, 2008] Nicholson, J. K. and Lindon, J. C. (2008). Systems biology: Metabonomics. *Nature*, 455(7216):1054–1056.
- [Nordestgaard, 2016] Nordestgaard, B. G. (2016). Triglyceride-rich lipoproteins and atherosclerotic cardiovascular disease. *Circulation Research*, 118(4):547–563.
- [O’Leary et al., 2016] O’Leary, N. A., Wright, M. W., Brister, J. R., Ciuffo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O’Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745.
- [Pavlova and Thompson, 2016] Pavlova, N. and Thompson, C. (2016). The emerging hallmarks of cancer metabolism. *Cell Metabolism*, 23(1):27–47.

- [Peterson et al., 2013] Peterson, C., Vannucci, M., Karakas, C., Choi, W., Ma, L., and Maletic-Savatic, M. (2013). Inferring metabolic networks using the bayesian adaptive graphical lasso with informative priors. *Statistics and Its Interface*, 6(4):547–558.
- [Prentice et al., 2014] Prentice, K., Luu, L., Allister, E., Liu, Y., Jun, L., Sloop, K., Hardy, A., Wei, L., Jia, W., Fantus, I. ., Sweet, D., Sweeney, G., Retnakaran, R., Dai, F., and Wheeler, M. (2014). The furan fatty acid metabolite cmpf is elevated in diabetes and induces cell dysfunction. *Cell Metabolism*, 19(4):653–666.
- [Putri et al., 2013] Putri, S. P., Yamamoto, S., Tsugawa, H., and Fukusaki, E. (2013). Current metabolomics: Technological advances. *Journal of Bioscience and Bioengineering*, 116(1):9–16.
- [Rainer et al., 2013] Rainer, P. P., Primessnig, U., Harenkamp, S., Doleschal, B., Wallner, M., Fauler, G., Stojakovic, T., Wachter, R., Yates, A., Groschner, K., Trauner, M., Pieske, B. M., and von Lewinski, D. (2013). Bile acids induce arrhythmias in human atrial myocardiumimplications for altered serum bile acid composition in patients with atrial fibrillation. *Heart*, 99(22):1685–1692.
- [Rasmussen, 1999] Rasmussen, C. E. (1999). The infinite gaussian mixture model. *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 554–560.
- [Roberts et al., 2012] Roberts, L. D., Souza, A. L., Gerszten, R. E., and Clish, C. B. (2012). Targeted metabolomics. *Current Protocols in Molecular Biology*, 98(1):30.2.1–30.2.24.
- [Rogers et al., 2009] Rogers, S., Scheltema, R. A., Girolami, M., and Breitling, R. (2009). Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, 25(4):512–518.

- [Roverato, 2002] Roverato, A. (2002). Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411.
- [Russell, 2003] Russell, D. W. (2003). The enzymes, regulation, and genetics of bile acid synthesis. *Annual Review of Biochemistry*, 72(1):137–174.
- [Sanderson and Curtin, 2016] Sanderson, C. and Curtin, R. (2016). Armadillo: a template-based c++ library for linear algebra. *The Journal of Open Source Software*, 1(2):26.
- [Shlomi et al., 2008] Shlomi, T., Cabili, M. N., Herrgrd, M. J., Palsson, B., and Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nature Biotechnology*, 26(9):1003–1010.
- [Silva et al., 2014] Silva, R. R., Jourdan, F., Salvanha, D. M., Letisse, F., Jamin, E. L., Guidetti-Gonzalez, S., Labate, C. A., and Vencio, R. Z. N. (2014). Probmetab: an r package for bayesian probabilistic annotation of lc-ms-based metabolomics. *Bioinformatics*, 30(9):1336–1337.
- [Smith et al., 2005] Smith, C. A., Maille, G. O., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., and Siuzdak, G. (2005). Metlin. *Therapeutic Drug Monitoring*, 27(6):747–751.
- [Southam et al., 2014] Southam, A. D., Lange, A., Al-Salhi, R., Hill, E. M., Tyler, C. R., and Viant, M. R. (2014). Distinguishing between the metabolome and xenobiotic exposome in environmental field samples analysed by direct-infusion mass spectrometry based metabolomics and lipidomics. *Metabolomics*, 10(6):1050–1058.
- [Stan Development Team, 2018] Stan Development Team (2018). *RStan: the R interface to Stan*. R package version 2.17.3.

- [Stoll and Bendszus, 2006] Stoll, G. and Bendszus, M. (2006). Inflammation and atherosclerosis: Novel insights into plaque formation and destabilization. *Stroke*, 37(7):1923–1932.
- [Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50.
- [Sumner et al., 2007] Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W. M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., and Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3(3):211–221.
- [Tabas et al., 2015] Tabas, I., Garca-Cardea, G., and Owens, G. K. (2015). Recent insights into the cellular biology of atherosclerosis. *The Journal of Cell Biology*, 209(1):13–22.
- [Tautenhahn et al., 2012] Tautenhahn, R., Cho, K., Uritboonthai, W., Zhu, Z., Patti, G. J., and Siuzdak, G. (2012). An accelerated workflow for untargeted metabolomics using the metlin database. *Nature Biotechnology*, 30(9):826–828.
- [Thygesen et al., 2012] Thygesen, K., Alpert, J. S., Jaffe, A. S., Simoons, M. L., Chaitman, B. R., White, H. D., Joint, E. S. C. A. A. H. A. W. H. F. T. F. f. U. D. o. M. I., Authors/Task Force Members, C., Thygesen, K., Alpert, J. S., White, H. D., Biomarker, S., Jaffe, A. S., Katus, H. A., Apple, F. S., Lindahl, B., Morrow, D. A., Subcommittee, E. C. G., Chaitman, B. R., Clemmensen, P. M., Johanson, P., Hod, H., Imaging, S., Underwood, R., Bax, J. J., Bonow, J. J., Pinto, F., Gibbons, R. J., Classification, S., Fox, K. A., Atar, D., Newby,

L. K., Galvani, M., Hamm, C. W., Intervention, S., Uretsky, B. F., Steg, P. G., Wijns, W., Bassand, J. P., Menasche, P., Ravkilde, J., Trials, Registries, S., Ohman, E. M., Antman, E. M., Wallentin, L. C., Armstrong, P. W., Simoons, M. L., Trials, Registries, S., Januzzi, J. L., Nieminen, M. S., Gheorghide, M., Filippatos, G., Trials, Registries, S., Luepker, R. V., Fortmann, S. P., Rosamond, W. D., Levy, D., Wood, D., Trials, Registries, S., Smith, S. C., Hu, D., Lopez-Sendon, J. L., Robertson, R. M., Weaver, D., Tendera, M., Bove, A. A., Parkhomenko, A. N., Vasilieva, E. J., Mendis, S., Guidelines, E. S. C. C. f. P., Bax, J. J., Baumgartner, H., Ceconi, C., Dean, V., Deaton, C., Fagard, R., Funck-Brentano, C., Hasdai, D., Hoes, A., Kirchhof, P., Knuuti, J., Kolh, P., McDonagh, T., Moulin, C., Popescu, B. A., Reiner, Z., Sechtem, U., Sirnes, P. A., Tendera, M., Torbicki, A., Vahanian, A., Windecker, S., Document, R., Morais, J., Aguiar, C., Almahmeed, W., et al. (2012). Third universal definition of myocardial infarction. *J Am Coll Cardiol*, 60(16):1581–98.

[Trainor et al., 2017] Trainor, P. J., Hill, B. G., Carlisle, S. M., Rouchka, E. C., Rai, S. N., Bhatnagar, A., and DeFilippis, A. P. (2017). Systems characterization of differential plasma metabolome perturbations following thrombotic and non-thrombotic myocardial infarction. *Journal of Proteomics*, 160.

[Trainor et al., 2018] Trainor, P. J., Yampolskiy, R. V., and DeFilippis, A. P. (2018). Wisdom of artificial crowds feature selection in untargeted metabolomics: An application to the development of a blood-based diagnostic test for thrombotic myocardial infarction. *Journal of Biomedical Informatics*, 81:53–60.

[Uhlen et al., 2015] Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szgyarto, C. A. K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P. H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk,

- J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419.
- [Uhlen et al., 2010] Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Bjrling, L., and Ponten, F. (2010). Towards a knowledge-based human protein atlas. *Nature Biotechnology*, 28(12):1248–1250.
- [Uppal et al., 2017] Uppal, K., Walker, D. I., and Jones, D. P. (2017). xmsannotator: An r package for network-based annotation of high-resolution metabolomics data. *Analytical Chemistry*, 89(2):1063–1067.
- [Uppal et al., 2016] Uppal, K., Walker, D. I., Liu, K., Li, S., Go, Y.-M., and Jones, D. P. (2016). Computational metabolomics: A framework for the million metabolome. *Chemical Research in Toxicology*, 29(12):1956–1975.
- [Voet et al., 2013] Voet, D., Voet, J. G., and Pratt, C. W. (2013). *Fundamentals of biochemistry : life at the molecular level*. Wiley, Hoboken, NJ, 4th edition.
- [Wang, 2012] Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886.
- [Wang et al., 2011] Wang, T. J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., Lewis, G. D., Fox, C. S., Jacques, P. F., Fernandez, C., O’Donnell, C. J., Carr, S. A., Mootha, V. K., Florez, J. C., Souza, A., Melander, O., Clish, C. B., and Gerszten, R. E. (2011). Metabolite profiles and the risk of developing diabetes. *Nature Medicine*, 17(4):448–453.
- [Wang et al., 2013] Wang, T. J., Ngo, D., Psychogios, N., Dejam, A., Larson, M. G., Vasan, R. S., Ghorbani, A., OSullivan, J., Cheng, S., Rhee, E. P., Sinha, S., McCabe, E., Fox, C. S., ODonnell, C. J., Ho, J. E., Florez, J. C., Magnusson, M., Pierce, K. A., Souza, A. L., Yu, Y., Carter, C., Light, P. E., Melander, O.,

- Clish, C. B., and Gerszten, R. E. (2013). 2-aminoadipic acid is a biomarker for diabetes risk. *Journal of Clinical Investigation*, 123(10):4309–4317.
- [Warth et al., 2017] Warth, B., Levin, N., Rinehart, D., Teijaro, J., Benton, H. P., and Siuzdak, G. (2017). Metabolizing data in the cloud. *Trends in Biotechnology*, 35(6):481–483.
- [West, 1987] West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3):646.
- [Wickham et al., 2017] Wickham, H., Danenberg, P., and Eugster, M. (2017). *roxygen2: In-Line Documentation for R*. R package version 6.0.1.
- [Wickham et al., 2018] Wickham, H., Hester, J., and Chang, W. (2018). *devtools: Tools to Make Developing R Packages Easier*. R package version 1.13.5.
- [Wishart, 2016] Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery*, 15(7):473–484.
- [Wishart et al., 2018] Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vazquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., Serra-Cayuela, A., Liu, Y., Mandal, R., Neveu, V., Pon, A., Knox, C., Wilson, M., Manach, C., and Scalbert, A. (2018). Hmdb 4.0: the human metabolome database for 2018. *Nucleic Acids Research*, 46(D1):D608–D617.
- [Xia and Wishart, 2010a] Xia, J. and Wishart, D. S. (2010a). Metpa: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*, 26(18):2342–2344.
- [Xia and Wishart, 2010b] Xia, J. and Wishart, D. S. (2010b). Msea: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Research*, 38(Web Server):W71–W77.

- [Yuan and Lin, 2007] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.
- [Zamboni et al., 2015] Zamboni, N., Saghatelian, A., and Patti, G. (2015). Defining the metabolome: Size, flux, and regulation. *Molecular Cell*, 58(4):699–706.
- [Zou, 2006] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

APPENDIX A: ACRONYMS UTILIZED

AARL: Atherosclerosis / Atherothrombosis Research Laboratory
APCI: Atmospheric Pressure Chemical Ionization
ATP: Adenosine triphosphate
AUC: Area Under the receiving operating characteristic Curve
BCAA: Branch chain amino acid
BGL: Bayesian Graphical Lasso
BLAS: Basic Linear Algebra Subprograms
CAD: Coronary Artery Disease
cAMP: Cyclic adenosine monophosphate
CASS: Chemically Aware Substructure Search algorithm
CE: Capillary Electrophoresis
ChemRICH: Chemical Similarity Enrichment Analysis
CI: Confidence Interval / Credible Interval
CoA: Coenzyme A
DA: Differentially Abundant
DI: Direct Injection
EC2: Elastic Compute Cloud
EI: Electron Ionization
ELISA: Enzyme-Linked Immunosorbent Assay
eQTL: Expression Quantitative Trait Loci
ESI: Electrospray Ionization
EST: Expressed Sequence Tag
FADH: Flavin adenine dinucleotide
FT-ICR: Fourier Transform Ion Cyclon Resonance
GC: Gas Chromatography
GGM: Gaussian Graphical Model
GSEA: Gene Set Enrichment Analysis
GTP: Guanosine triphosphate
HILIC: Hydrophilic Interaction Liquid Chromatography
HMDB: Human Metabolome Database
HMG-CoA: 3-hydroxy-3-methylglutaryl-CoA
KEGG: Kyoto Encyclopedia of Genes and Genomes
LAPACK: Linear Algebra PACKage
LASSO: Least Absolute Shrinkage and Selection Operator
LC: Liquid Chromatography
LOOCV: Leave One Out - Cross Validation
MALDI: Matrix Assisted Laser Desorption / Ionization
MCMC: Markov Chain Monte Carlo
MeSH: Medical Subject Headings

MI: Myocardial Infarction
MLE: Maximum Likelihood Estimation MRM: Multiple Reaction Monitoring
mRNA: Messenger ribonucleic acid
MS: Mass Spectrometry
MSEA: Metabolite Set Enrichment Analysis
NAD: Nicotinamide adenine dinucleotide
NMR: Nuclear Magnetic Resonance
PKA: Protein kinase A
RNA: Ribonucleic acid
RP: Reversed Phase
SCAD: Smoothly Clipped Absolute Deviation
SMPDB: Small Molecule Pathway Database
SRM: Selected Reaction Monitoring
TCA: Tricarboxylic acid
TiGER: Tissue-specific Gene Expression and Regulation database
TOF: Time of Flight
UPLC: Ultra Performance Liquid Chromatography

APPENDIX B: SOFTWARE DEVELOPED FOR THE BAYESIAN GRAPHICAL LASSO

High-level R code

Main high-level Function for generating a BGL object

```
#' Block Gibbs sampler for Bayesian Graphical Lasso  
#'  
#' Blockwise sampling from the conditional distribution of a  
permuted column/row  
#' for simulating the posterior distribution for the  
concentration matrix specifying  
#' a Gaussian Graphical Model  
#' @param X Data matrix  
#' @param iterations Length of Markov chain after burn-in  
#' @param burnIn Number of burn-in iterations  
#' @param adaptive Logical; Adaptive graphical lasso (TRUE)  
or regular (FALSE). Default is FALSE.  
#' @param lambdaPriora Shrinkage parameter (lambda) gamma  
distribution shape hyperparameter  
#' (Ignored if adaptive=TRUE)  
#' @param lambdaPriorb Shrinkage parameter (lambda) gamma  
distribution scale hyperparameter  
#' (Ignored if adaptive=TRUE)  
#' @param adaptiveType Choose of adaptive type. Options are  
"norm" for norm of concentration  
#' matrix based adaptivity and "priorHyper" for informative  
adaptivity  
#' @param priorHyper Matrix of gamma scale hyper parameters  
(Ignored if adaptiveType="norm")  
#' @param gammaPriors lambda_ij gamma distribution shape  
prior (Ignored if adaptive=FALSE)  
#' @param gammaPriort lambda_ij gamma distribution rate  
prior (Ignored if adaptive=FALSE)  
#' @param lambda_ii lambda_ii hyperparameter (Ignored if  
adaptive=FALSE)  
#' @param keepLambdas Logical: Should lambda MCMC chain  
(vector / matrix) be kept?
```

```

#' @param illStart Method for generating a positive definite
  estimate of the sample covariance matrix if sample
  covariance matrix is not semi-positive definite
#' @param rho Regularization parameter for the graphical
  lasso estimate of the sample covariance matrix (if
  illStart="glasso")
#' @param verbose logical; if TRUE return MCMC progress
#' @details Implements the block Gibbs sampler for the
  posterior distribution of
#' a GGM concentration matrix estimated via the Bayesian
  Graphical Lasso
#' introduced by Wang (2012) or the Bayesian Adaptive
  Graphical Lasso. For the adaptive case,
#' the element-wise shrinkage parameter is drawn from a
  gamma distribution where the scale
#' parameter is adaptively modulated directly from the
  estimated concentration matrix
#' or a user can supply their own informative prior.
#'
#' @return
#' \item{Omega}{List of concentration matrices from the
  Markov chains}
#' \item{Lambda}{Vector of simulated lambda parameters}
#' @author Patrick Trainor (University of Louisville)
#' @author Hao Wang
#' @references Wang, H. (2012). Bayesian graphical lasso
  models and efficient
#' posterior computation. \emph{Bayesian Analysis, 7}(4).
  <doi:10.1214/12-BA729> .
#' @examples
#' \donttest{
  #' # Generate true covariance matrix:
  #' s<-.9**toeplitz(0:9)
  #' # Generate multivariate normal distribution:
  #' set.seed(5)
  #' x<-MASS::mvrnorm(n=100,mu=rep(0,10),Sigma=s)
  #' blockGLasso(X=x)
  #' }
#' # Same example with short MCMC chain:
#' s<-.9**toeplitz(0:9)
#' set.seed(6)
#' x<-MASS::mvrnorm(n=100,mu=rep(0,10),Sigma=s)
#' blockGLasso(X=x,iterations=100,burnIn=100)
#' @export
blockGLasso<-function(X,iterations=2000,burnIn=1000,adaptive=FALSE,
  lambdaPriora=1,lambdaPriorb=1/10,
  adaptiveType=c("norm","priorHyper"),priorHyper=NULL,
  gammaPriors=1,gammaPriort=1,

```



```

lambdaI=1,keepLambdas=TRUE,illStart=c("identity","glasso"),rho=.1,
verbose=TRUE,...)
{
  if(adaptive)
  {
    UseMethod("blockAdGLasso")
  }else{
    UseMethod("blockGLasso")
  }
}
}

```

Function for calling the non-adaptive block Gibbs sampler

```

#' @export
blockGLasso.default<-function(X,iterations=2000,burnIn=1000,
  lambdaPriora=1,lambdaPriorb=1/10,
  illStart=c("identity","glasso"), keepLambdas=TRUE,rho=.1,
  verbose=TRUE,...) {
  # Total iterations:
  totIter<-iterations+burnIn

  # Ill conditioned start:
  illStart<-match.arg(illStart)

  # Sum of product matrix, covariance matrix, n
  S<-t(X)%*%X
  n=nrow(X)
  Sigma=S/n
  p<-dim(Sigma)[1]

  # Concentration matrix and it's dimension:
  if(rcond(Sigma)<.Machine$double.eps){
    if(illStart=="identity"){
      Omega<-diag(nrow(Sigma))+1/(p**2)
    }
    else{
      Omega<-glasso::glasso(cov(X),rho=rho)$wi
      + 1/(p**2)
    }
  }
  else{
    Omega<-MASS::ginv(Sigma)
  }
  rownames(Omega)<-rownames(Sigma)
  colnames(Omega)<-colnames(Sigma)

  # Gamma distribution posterior parameter a:
  lambdaPosta<-(lambdaPriora+(p*(p+1)/2))

```

```

# Keep lambdas?
keepLambdas<-ifelse (keepLambdas ,1L,0L)

# Call sampler:
bglObj<-bgl (n=n, iters=totIter , lambdaPriorb=lambdaPriorb ,
  lambdaPosta=lambdaPosta ,
  S=S , Sigma=Sigma , Omega=Omega , keepLambdas=keepLambdas)

bglObj<-c (bglObj , burnIn=burnIn)
class (bglObj)<-"BayesianGLasso"
return (bglObj)
}

```

Function for calling the non-adaptive block Gibbs sampler

```

#' @export
blockAdGLasso.default<-function (X, iterations=2000, burnIn=1000,
  adaptiveType=c ("norm" ," priorHyper" ) ,
  priorHyper=NULL, gammaPriors=1, gammaPriorT=1,
  lambdaI=1, keepLambdas=TRUE, illStart=c (" identity" ," glasso" ) ,
  rho=.1, verbose=TRUE, ...) {
  # Total iterations:
  totIter<-iterations+burnIn

  # Ill conditioned start:
  illStart<-match.arg (illStart)

  # Sum of product matrix, covariance matrix, n
  S<-t (X)%*%X
  n=nrow (X)
  Sigma=S/n
  p<-dim (Sigma) [1]

  # Adaptive type:
  adaptiveType<-match.arg (adaptiveType)
  if (adaptiveType==" priorHyper" ) {
    if (is.null (priorHyper)) stop ("Must specify a
      matrix of prior hyperparameters")
    if (class (priorHyper)!=" matrix" ) stop (" Prior
      hyperparameters must be provided as a
      matrix")
    if (!all (dim (Sigma)==dim (priorHyper)))
      stop (" Dimesion of hyperparameters does not
      equal dimension
      ..... of the concentration matrix")
    priorLogical<-TRUE
  } else {
    priorLogical<-FALSE
    priorHyper<-matrix (1 , nrow=p , ncol=p)
  }
}

```

```

}

# Concentration matrix and it's dimension:
if (rcond(Sigma) < .Machine$double.eps) {
  if (illStart == "identity") {
    Omega <- diag(nrow(Sigma)) + 1/(p**2)
  }
  else {
    Omega <- glasso::glasso(cov(X), rho=rho)$wi + 1/(p**2)
  }
} else {
Omega <- MASS::ginv(Sigma)
}
rownames(Omega) <- rownames(Sigma)
colnames(Omega) <- colnames(Sigma)

# Keep lambdas?
keepLambdas <- ifelse(keepLambdas, 1L, 0L)

# Call sampler:
bglObj <- bAdgl(n=n, iters=totIter, gammaPriors=gammaPriors,
gammaPriort=gammaPriort, lambdaIi=lambdaIi, S=S, Sigma=Sigma,
Omega=Omega, priorLogical=priorLogical,
priorHyper=priorHyper,
keepLambdas=keepLambdas)

bglObj <- c(bglObj, burnIn=burnIn)
class(bglObj) <- "BayesianGLasso"
return(bglObj)
}

```

Posterior inference method for BGL object

```

#' Posterior inference
#'
#' Inferential statistics for the simulated posterior
distributions of the covariance
#' and concentration parameter matrices for Gaussian
Graphical Models estimated
#' via the adaptive or non-adaptive Bayesian Graphical Lasso
#'
#' @param x Bayesian Graphical Lasso object as returned by
the function blockGLasso
#' @param parameter Options are "Omega" for the posterior
distribution of the
#' concentration matrix or "Sigma" for the covariance matrix
#' @return
#' \item{posteriorMean}{Matrix of mean values of the
simulated posterior distribution}

```

```

#' \item{posteriorSd}{Matrix of standard deviation values of
  the simulated posterior distribution}
#' \item{posteriorMedian}{Matrix of median values of the
  simulated posterior distribution}
#' \item{lowerCI}{Lower limit of the Bayesian credible
  interval for each matrix parameter}
#' \item{upperCI}{Upper limit of the Bayesian credible
  interval for each matrix parameter}
#' @export
posteriorInference<-function(bglObj , parameter="Omega" ,
alpha=.05 ,...){
  UseMethod("posteriorInference")
}

#' @export
posteriorInference.BayesianGLasso<-function(bglObj , parameter="Omega" ,
alpha=.05){
  # Check parameters
  if(! parameter %in% c("Omega" ,"Sigma")){
    stop("Parameter argument must be Omega_
      (concentration matrix) or
      _____Sigma_(covariance matrix)")
  }

  stats<-c("mean" ,"sd" ,"length" ,"median")

  # Discard burn in:
  x2<-bglObj
  x2$Sigmas<-x2$Sigmas[(x2$burnIn+1):length(x2$Sigmas)]
  x2$Omegas<-x2$Omegas[(x2$burnIn+1):length(x2$Omegas)]
  if(! is.null(x2$lambdas))
    x2$lambdas<-x2$lambdas[(x2$burnIn+1):length(x2$lambdas)]

  for(stat in stats){
    statMatrix<-matrix(NA,nrow=nrow(x2$Omegas[[1]]),
      ncol=ncol(x2$Omegas[[1]]))
    for(i in 1:nrow(statMatrix)){
      for(j in 1:ncol(statMatrix)){
        statMatrix[i , j]<-eval(call(stat ,
          sapply(x2$Omegas ,FUN=function(y)
            y[i , j])))
      }
    }
    assign(paste0("stat" ,stat) ,statMatrix)
  }

  lq<-function(z) quantile(z , probs=alpha/2)
  uq<-function(z) quantile(z , probs=1-(alpha/2))

```

```

lqMatrix<-uqMatrix<-matrix(NA,nrow=nrow(x2$Omegas[[1]]),
  ncol=ncol(x2$Omegas[[1]]))
for(i in 1:nrow(lqMatrix)){
  for(j in 1:ncol(lqMatrix)){
    lqMatrix[i,j]<-eval(call("lq",sapply(x2$Omegas,
      FUN=function(y) y[i,j])))
    uqMatrix[i,j]<-eval(call("uq",sapply(x2$Omegas,
      FUN=function(y) y[i,j])))
  }
}

```

```

list(parameter=parameter, alpha=alpha, posteriorMean=statmean, posteriorS
posteriorMedian=statmedian, lowerCI=lqMatrix, upperCI=uqMatrix)
}

```

Preamble and internal C++ functions

```

#include <RcppArmadillo.h>
// #include <gperftools/profiler.h>

// [[Rcpp::depends(RcppArmadillo)]]

using namespace Rcpp;
// using namespace arma;

double rInvG(double mu, double lambda){
  double val = 0;
  double z,y,x,u;
  z = rnorm(1)[0];
  y = z*z;
  x = mu + 0.5*mu*mu*y/lambda - 0.5*(mu/lambda) *
    sqrt(4*mu*lambda*y + mu*mu*y*y);
  u = runif(1)[0];
  if(u <= mu/(mu+x)){
    val = x;
  } else {
    val = mu*mu/x;
  }
}
return(val);
}

```

```

arma::vec upperTri(arma::mat a){
  arma::mat b = trimatu(a);
  arma::mat b2 = trimatl(a);
  arma::uvec c = find(b);
  arma::uvec c2 = find(b2);
  NumericVector d = wrap(c);
  NumericVector d2 = wrap(c2);
  NumericVector e = setdiff(d,d2);
}

```

```

    e=e.sort();
    arma::uvec f = as<arma::uvec>(e);
    arma::vec g = a.elem(f);
    return g;
}

IntegerMatrix permFun(int p){
    IntegerVector permInt = seq_len(p) - 1;
    IntegerMatrix perms(p-1,p);
    IntegerVector permInt2(p);
    for(int i = 0; i < p; i++)
    {
        permInt2 = permInt[permInt != i];
        perms(_,i) = permInt2;
    }
    return perms;
}

```

Regular Bayesian Graphical Lasso C++ code

```

// [[Rcpp::export]]
List bgl(int n, int iters, double lambdaPriorb, double
lambdaPosta,
arma::mat S, arma::mat Sigma, arma::mat Omega, int
keepLambdas){
    // ProfilerStart("myprof.log");

    int p = Omega.n_cols;
    IntegerMatrix idk = permFun(p);
    List OmegaList(iters);

    NumericVector lambdaList;
    if(keepLambdas==1){
        lambdaList = NumericVector(iters);
    }
    else{
        lambdaList = NumericVector(1);
    }

    for(int iter=0; iter < iters; ++iter){

        // Sample lambda:
        arma::vec Omega2 = vectorise(Omega);
        double lambdaPostb = lambdaPriorb +
            sum(abs(Omega2)) / 2.0;
        double lambda = R::rgamma(lambdaPosta, 1.0 /
            lambdaPostb);

        if(keepLambdas==1){

```

```

        lambdaList(iter) = lambda;
    }

    // Mu prime:
    arma::vec OmegaTemp = abs(upperTri(Omega));
    // arma::vec OmegaTemp =
    //   abs(Omega.elem(find(trimatu(Omega,1))));
    arma::vec mup = lambda/OmegaTemp;
    /* double mupThresh = pow(10.0,12.0);
    arma::uvec mupReplace = find(mup >
        mupThresh);
    mup.elem(mupReplace).fill(mupThresh); */

    // Sample tau:
    arma::vec rIG(p*(p-1)/2);
    for(unsigned int i = 0; i < mup.n_elem; i++){
        rIG(i) = rInvG(mup[i], lambda*lambda);
    }
    arma::vec tauVec = 1.0 / rIG;
    arma::mat
        tau(Omega.n_rows, Omega.n_cols, arma::fill::zeros);
    int k = 0;
    for(unsigned int j = 0; j < tau.n_cols; j++){
        for(unsigned int i = 0; i < j; i++){
            tau(i, j) = tauVec(k);
            tau(j, i) = tauVec(k);
            k++;
        }
    }

    IntegerVector tauIPerms(p);
    arma::uvec tauIPerms2(p), tauII(1);
    arma::vec Sigma12(p-1), rnorm1(p-1),
        beta(p-1);
    arma::mat tauI(p-1,1), Sigma11(p-1,p-1),
        Omega11inv(p-1,p-1), Ci(p-1,p-1),
        CiChol(p-1,p-1), mui(p-1,1);
    arma::mat OmegaInvTemp(p-1,1);
    double gamm;
    arma::mat gamm2;
    for(int i = 0; i < p; i++){
        tauIPerms = idk(_, i);
        tauIPerms2 =
            as<arma::uvec>(tauIPerms);
        tauII(0) = i;
        tauI = tau.submat(tauIPerms2, tauII);
        Sigma11 = Sigma.submat(tauIPerms2,
            tauIPerms2);
    }

```

```

        Sigma12 = Sigma.submat(tauIPerms2 ,
            tauII);
        Omega11inv = Sigma11 - ((Sigma12 *
            Sigma12.t()) / Sigma(i, i));
        Ci = (S(i, i) + lambda) * Omega11inv
            + diagmat(1 / tauI);
        CiChol = arma::chol(Ci);
        mui = arma::solve(-Ci,
            S.submat(tauIPerms2 , tauII));
        rnorm1 = rnorm(p-1);
        beta = mui + arma::solve(CiChol,
            rnorm1);
        Omega.submat(tauIPerms2 , tauII) =
            beta;
        Omega.submat(tauII , tauIPerms2) =
            beta.t();
        gamm = rgamma(1, n/2+1, 2/(S(i, i) +
            lambda)) [0];
        gamm2 = gamm + beta.t() * Omega11inv
            * beta;
        Omega(i, i) = gamm2(0,0);
        OmegaInvTemp = Omega11inv * beta;
        Sigma.submat(tauIPerms2 , tauIPerms2)
            = Omega11inv + (OmegaInvTemp *
            OmegaInvTemp.t()) / gamm;
        Sigma.submat(tauIPerms2 , tauII) =
            -OmegaInvTemp / gamm;
        Sigma.submat(tauII , tauIPerms2) =
            -OmegaInvTemp.t() / gamm;
        Sigma(i, i) = 1 / gamm;
    }
    OmegaList(iter) = Omega;
    Rcpp::Rcout << "iter = " << iter + 1 <<
        std::endl;

}
// ProfilerStop();

return List::create(Named("lambdas") = lambdaList,
    Named("Omegas") = OmegaList);
}

```

Adaptive Bayesian Graphical Lasso C++ code

```

// [[Rcpp::export]]
List bAdgl(int n, int iters, double gammaPriors, double
    gammaPriort,
    double lambdaii, arma::mat S, arma::mat Sigma,
    arma::mat Omega, bool priorLogical, arma::mat priorHyper,

```



```

int keepLambdas){
    // ProfilerStart("myprof.log");

    int p = Omega.n_cols;
    IntegerMatrix idk = permFun(p);

    List OmegaList(iters);

    List LambdaList;
    if(keepLambdas==1){
        LambdaList = List(iters);
    }
    else{
        LambdaList = List(1);
    }

    for(int iter=0; iter < iters; ++iter){
        // Sample lambdas:
        arma::vec Omega2 = vectorise(Omega);
        arma::vec OmegaTemp = abs(upperTri(Omega));
        // arma::vec OmegaTemp =
            abs(Omega.elem(find(trimatu(Omega,1))));

        // Gamma distribution conditional parameters:
        arma::vec tt = OmegaTemp + gammaPrior;

        // If informative prior:
        if(priorLogical){
            arma::vec priorHyperVec =
                vectorise(priorHyper);
            priorHyperVec = upperTri(priorHyper);
            // priorHyperVec =
                priorHyperVec.elem(find(trimatu(priorHyper,1)));
            tt = tt + priorHyperVec;
        }

        arma::vec s(tt.n_elem);
        s.fill(gammaPriors + 1);

        // Sample lambda:
        arma::vec lambda(tt.n_elem);
        for(unsigned int ii = 0; ii < tt.n_elem; ++ii){
            lambda(ii) = R::rgamma(s(ii),
                1.0/tt(ii)); // Change to Rcpp
                sugar version
        }

        // Mu prime:

```

```

arma::vec mup = lambda/OmegaTemp;
double mupThresh = pow(10.0,12.0);
arma::uvec mupReplace = find(mup >
    mupThresh);
mup.elem(mupReplace).fill(mupThresh);

// Sample tau:
arma::vec rIG(p*(p-1)/2);
for(unsigned int i = 0; i < mup.n_elem; i++){
    rIG(i) =
        rInvG(mup(i),pow(lambda(i),2));
}
arma::vec tauVec = 1.0 / rIG;
arma::mat
    tau(Omega.n_rows, Omega.n_cols, arma::fill::zeros);
int k = 0;
for(unsigned int j = 0; j < tau.n_cols; j++){
    for(unsigned int i = 0; i < j; i++){
        tau(i,j) = tauVec(k);
        tau(j,i) = tauVec(k);
        k++;
    }
}

```

```

IntegerVector tauIPerms(p);
arma::uvec tauIPerms2(p), tauII(1);
arma::vec Sigma12(p-1), rnorm1(p-1),
    beta(p-1);
arma::mat tauI(p-1,1), Sigma11(p-1,p-1),
    Omega11inv(p-1,p-1), Ci(p-1,p-1),
    CiChol(p-1,p-1), mui(p-1,1);
arma::mat OmegaInvTemp(p-1,1);
double gamm;
arma::mat gamm2;
for(int i = 0; i < p; i++){
    tauIPerms = idk(_,i);
    tauIPerms2 =
        as<arma::uvec>(tauIPerms);
    tauII(0) = i;
    tauI = tau.submat(tauIPerms2, tauII);
    Sigma11 = Sigma.submat(tauIPerms2,
        tauIPerms2);
    Sigma12 = Sigma.submat(tauIPerms2,
        tauII);
    Omega11inv = Sigma11 - ((Sigma12 *
        Sigma12.t()) / Sigma(i,i));
    Ci = (S(i,i) + lambdaii) *
        Omega11inv + diagmat(1 / tauI);
}

```

```

        CiChol = arma::chol(Ci);
        mui = arma::solve(-Ci,
            S.submat(tauIPerms2, tauII));
        rnorm1 = rnorm(p-1);
        beta = mui + arma::solve(CiChol,
            rnorm1);
        Omega.submat(tauIPerms2, tauII) =
            beta;
        Omega.submat(tauII, tauIPerms2) =
            beta.t();
        gamm = rgamma(1, n/2+1, 2/(S(i, i) +
            lambda_ii)) [0];
        gamm2 = gamm + beta.t() * Omega11inv
            * beta;
        Omega(i, i) = gamm2(0,0);
        OmegaInvTemp = Omega11inv * beta;
        Sigma.submat(tauIPerms2, tauIPerms2)
            = Omega11inv + (OmegaInvTemp *
                OmegaInvTemp.t())/gamm;
        Sigma.submat(tauIPerms2, tauII) =
            -OmegaInvTemp/gamm;
        Sigma.submat(tauII, tauIPerms2) =
            -OmegaInvTemp.t()/gamm;
        Sigma(i, i) = 1/gamm;
    }
    OmegaList(iter) = Omega;

    arma::mat
        lambdaMat(Omega.n_rows, Omega.n_cols, arma::fill::zero);
    int kk = 0;
    for(unsigned int j = 0; j <
        lambdaMat.n_cols; j++){
        for(unsigned int i = 0; i < j; i++){
            lambdaMat(i, j) = lambda(kk);
            kk++;
        }
    }
    if(keepLambdas==1){
        LambdaList(iter) = lambdaMat;
    }

    double detOmega = arma::det(Omega);
    Rcpp::Rcout << "iter _=" << iter + 1 << " _
        det(Omega) _=" << detOmega << std::endl;
}

return List::create(Named("lambdas") = LambdaList,

```

```
} Named("Omegas") = OmegaList);
```

CURRICULUM VITAE

Patrick Trainor

Objective

My objective is to obtain a research position in computational metabolomics. My primary research interests are: (1) developing Bayesian methodology for integrating extra-experimental knowledge and data sources for integrated -omics analyses, (2) developing Bayesian methodologies for solving problems in analytical chemistry, and (3) developing computational tools for metabolomics data analyses.

Education

PhD Bioinformatics

University of Louisville, Louisville, KY, 2015-2018 (expected)

MS Biostatistics

University of Louisville, Louisville, KY, 2012-2014

MA Mathematics

University of Louisville, Louisville, KY, 2012-2014

BS Mathematics

Seattle University, Seattle, WA, 2012

Research Positions

Graduate Research Assistant 2012-2014, 2015-Current
Institute of Molecular Cardiology & J.G. Brown Cancer Center (formerly),
University of Louisville

Consultant & Intern 2012-2013
Theravance Inc. (Theravance Biopharma)

Teaching Positions

Graduate Teaching Assistant 2012-2014
University of Louisville

Undergraduate Research Assistant 2010-2012
Seattle University

Other Professional Experience

Clinical Analytics Consultant 2014-2015
Humana, Inc., Louisville, KY

Rifleman & Infantry Squad Leader 2007-2013
4th Anti-Terrorism Battalion
United States Marine Corps (R)

Publications

*Denotes equally contributing authorship

Trainor, P.J., Yampolskiy, R.V.*, & DeFilippis, A.P.* (2018). Wisdom of artificial crowds feature selection in untargeted metabolomics: An application to the development of a blood-based diagnostic test for thrombotic myocardial in-

farction. *Journal of Biomedical Informatics*, 81. doi: 10.1016/j.jbi.2018.03.007

DeFilippis, A.P., **Trainor, P.J.** (2018). When given a lemon, make lemonade: Revising cardiovascular risk prediction scores. *Annals of Internal Medicine*, online ahead of print. doi: 10.7326/M18-1175.

Carlisle, S.M., **Trainor, P.J.**, Doll, M.A., Stepp, M.W., Klinge, C.M., & Hein, D.W. (2018). Knockout of human arylamine N-acetyltransferase 1 (NAT1) in MDA-MB-231 breast cancer cells leads to increased reserve capacity, maximum mitochondrial capacity, and glycolytic reserve capacity. *Molecular Carcinogenesis*, online ahead of print. doi: 10.1002/mc.22869

Trainor, P.J., DeFilippis, A.P., & Rai, S.N. (2017). Classifier performance for multiclass phenotype discrimination in metabolomics. *Metabolites*, 7(2). doi: 10.3390/metabo7020030

Trainor, P.J., Hill, B.G., Carlisle, S.M., Rouchka, E.C., Bhatnagar, A., Rai, S.N., & DeFilippis, A.P. (2017). Systems characterization of differential plasma metabolome perturbations following thrombotic and non-thrombotic myocardial infarction *Journal of Proteomics*, 160. doi: 10.1016/j.jprot.2017.03.014.

DeFilippis, A.P., **Trainor, P.J.**, Hill, B.G., Amraotkar, A.R., Rai, S.N., Hirsch, G.A., Rouchka, E., & Bhatnagar, A. (2017). Identification of a plasma metabolomic signature of thrombotic myocardial infarction that is distinct from non-thrombotic myocardial infarction and stable coronary artery disease. *PLoS One*, 12(4). doi: 10.1371/journal.pone.0175591.

Gibb, A.A., Epstein, P.N., Uchida, S., McNally, L.A., Obal, D., Katragadda, K., **Trainor, P.J.**, Conklin, D.J., Brittan, K.R., Tseng, M.T., Wang J., Jones,

S.P., Bhatnagar, A., & Hill, B.G. (2017). Exercise-Induced Changes in Glucose Metabolism Promote Physiologic Cardiac Growth *Circulation*, *136*(21). doi: 10.1161/CIRCULATIONAHA.117.028274

Khanal, S. **Trainor, P.J.**, Zahin, M., Ghim, S.J., Joh, J., Rai, S.N., Jenson, A.B., & Shumway, B.S. (2017). Histologic variation in high grade oral epithelial dysplasia when associated with high-risk human papillomavirus. *Oral Surgery, Oral Medicine, Oral Pathology, and Oral Radiology*, *123*(5). doi: 10.1016/j.oooo.2017.01.008.

Sultan, A., Zheng, Y., **Trainor, P.J.**, Yong, S., Amraotkar, A.R. Hill, B.G., & DeFilippis, A.P. (2017). Circulating prolidase activity in patients with myocardial infarction. *Frontiers in Cardiovascular Medicine*. doi: 10.3389/fcvm.2017.00050.

Amraotkar, A.R., Ghafghazi, S., **Trainor, P.J.**, Hargis C.W., Irfan A.B., Rai S.N., Bhatnagar A., & DeFilippis A.P. (2017). Presence of multiple coronary angiographic characteristics for the diagnosis of acute coronary thrombus. *Cardiology Journal*, *24*(1). doi: 10.5603/CJ.a2017.0004.

Khosravi, F., **Trainor, P.J.**, Lambert, C., Kloecker, G., Wickstrom, E. Rai, S.N., & Panchapakesan, B. (2016). Static micro-array isolation, dynamic time series classification, capture and enumeration of spiked breast cancer cells in blood: the nanotube–CTC chip. *Nanotechnology*, *27*(44). doi: 10.1088/0957-4484/27/44/44LT03.

Carlisle, S.M., **Trainor, P.J.**, Yin, X., Doll, M.A., Stepp, M.W., States, J.C., Zhang, X., & Hein, D.W. (2016). Untargeted polar metabolomics of transformed MDA-MB-231 breast cancer cells expressing varying levels of human arylamine N-acetyltransferase 1. *Metabolomics*, *12*(7). doi: 10.1007/s11306-

016-1056-z.

Amraotkar, A.R., Song, D.D., **Trainor, P.J.**, Ismail, I., Kothari, K., Sing, A., Moore, J.B., Rai, S.N., Bhatnagar, A., & DeFilippis, A.P. (2016). Platelet count and mean platelet volume at the time of and after acute myocardial infarction. *Clinical and Applied Thrombosis/Haemostasis*, *23*(8). doi: 10.1177/1076029616683804.

DeFilippis, A.P., Chernyavskiy, I., Amraotkar, A.R., **Trainor, P.J.**, Kothari, S., Ismail, I., Hargis, C.W., Korley, F.K., Tsimikas, S., Rai, S.N., & Bhatnagar, A. (2016). Circulating levels of plasminogen and oxidized phospholipids bound to plasminogen distinguish between atherothrombotic and non-atherothrombotic myocardial infarction. *Journal of Thrombosis and Thrombolysis*, *42*(1). doi: 10.1007/s11239-015-1292-5.

Egan, J.M., Slughter, J.M., Cardoso, T., **Trainor, P.J.**, Wu, K., Safford, H., & Fournier, D. (2016). Multi-temporal ecological analysis of Jeffrey pine beetle outbreak dynamics within the Lake Tahoe Basin. *Population Ecology*, *58*(3). doi: 10.1007/s10144-016-0545-2.

Rai, S.N.*, **Trainor, P.J.***, Khosravi, F., & Panchapakesan, B. (2016). Classification of biosensor time series using dynamic time warping: applications in screening cancer cells with characteristic biomarkers. *Open Access Medical Statistics*, *6*. doi: 10.2147/OAMS.S104731.

Khosravi, F., **Trainor, P.J.**, Rai, S.N., Kloecker, G., Wickstrom, E. & Panchapakesan, B. (2016). Label-free capture of breast cancer cells spiked in buffy coats using carbon nanotube antibody micro-arrays. *Nanotechnology*, *27*(13). doi: 10.1088/0957-4484/27/13/13LT02.

Khanal, S., Cole, ET., Joh, J., Ghim, S.J., Jenson, A.B., Rai, S.N., **Trainor, P.J.**, & Shumway, B.S. (2015). Human papillomavirus detection in histologic samples of multifocal epithelial hyperplasia: a novel demographic presentation. *Oral Surgery, Oral Medicine, Oral Pathology, and Oral Radiology*, 120(6). doi: 10.1016/j.oooo.2015.07.035.

Published Software

Trainor, P.J. & Wang H. (2017). *BayesianGLasso: Bayesian Graphical Lasso*. Available from <https://CRAN.R-project.org/package=BayesianGLasso>

Accepted Publications

Khanal S., Shumway B.S., Zahin M., Redman R., **Trainor P.J.**, Rai S.N., Ghim S.J., Jenson A.B., & Joh J. (2018). Human Papillomavirus DNA integration, but not viral methylation, is associated with malignant progression of head and neck cancers. *Oncotarget*, in production.

Invited Talks

Trainor, P.J. (2017). Metabolomics of thrombotic myocardial infarction: systems characterization of plasma metabolome perturbations and the development of a diagnostic classifier. *Systems Biology Seminar Series*, University of Kentucky, Lexington, Kentucky.

Oral / Poster Presentations

Trainor, P.J. (2018). Inferring metabolite interactomes via Bayesian graphical model selection utilizing molecular structure informative priors. *14th Annual Conference of the Metabolomics Society*, Seattle, Washington.

Trainor, P.J. (2018). High Performance R: Parallel computing, seamless C++ integration, optimizing linear algebra routines, and deep learning in the R language. *Kentucky Biomedical Research Infrastructure Network Seminar Series*, University of Louisville, Louisville, Kentucky.

Trainor, P.J., Carlisle, S.M., Hill, B.G., Rouchka, E.C., Bhatnagar, A., Rai, S.N., & DeFilippis, A.P. (2017). Metabolomic analysis of thrombotic and non-thrombotic Myocardial Infarction: From systems biology to diagnostic classification. *Symposium in Molecular Biology “Metabolism: Disease Models and Model Organisms”*. Pennsylvania State University, State College, Pennsylvania.

Trainor, P.J., Carlisle, S.M., DeFilippis, A.P.*, & Rai, S.N.* (2017). Molecular fingerprinting for inferring a Gaussian Graphical Model representation of a stable coronary artery disease plasma interactome using adaptive graphical Lasso penalization. *UT-KBRIN Annual Bioinformatics Summit*, Burns, Tennessee.

Trainor, P.J., Bhatnagar, A., Rai, S.N., & DeFilippis, A.P. (2016). Dynamic changes in topologically related plasma metabolites following atherothrombotic and non-atherothrombotic myocardial infarction: A weighted network analysis. *Kentucky Academy of Science Annual Meeting*, University of Louisville, Louisville, Kentucky.

Trainor, P.J., & Rai, S.N. (2015). Comparison of partial least squares-discriminant analysis, random forests, and rotation forests for phenotype discrimination given metabolomic data. *Research!Louisville*, University of Louisville, Louisville, Kentucky.

Trainor, P.J. & Rai, S.N. (2013). Patient rule induction method for identifying subgroups in clinical studies. *Kentucky Academy of Science Annual Meeting*, Morehead State University, Morehead, Kentucky.

Trainor, P.J., & Barnes, C. (2013). Statistical considerations for dedicated cardiovascular safety studies. *Theravance, Inc. (now Theravance Biopharma) company seminar*, San Francisco, California.

Trainor, P.J., & Slougher, J.M. (2012). Generalized Linear Modeling in Population Ecology. *Northwest Undergraduate Mathematics Symposium*, Lewis & Clark College, Portland, Oregon.

Grants

Co-investigator. NIH West Coast Metabolomics Center Pilot and Feasibility Award: “Targeted plasma metabolomic profiling of thrombotic myocardial infarction”. PI: DeFilippis, A.P. (2017-2018). \$50,000 direct cost.

Co-Investigator. Alpha Phi Foundation Heart-to-Heart Grant (2016 award): “Circulating Levels of Oxidized Phospholipids, Sub-Clinical Atherosclerosis and Atherosclerotic Cardiovascular Disease”. PI: DeFilippis, A.P. (2016-2017). \$100,000 direct costs.

Computing Experience

Programming Languages: R, Python, C++, MATLAB/Octave

Machine Learning & Deep Learning: TensorFlow, Keras, WEKA

Statistical Software: R, SAS

Bioinformatics Software: Many packages distributed via Bioconductor

Computer Algebra Systems: SAGE, Mathematica

Operating Systems: UNIX, Linux (openSUSE is my favorite flavor), and Windows

AWARDS

Metabolomics Society Student Travel Award, 2018, Metabolomics Society

2nd Place Graduate Research Competition in Health Sciences, 2016, Kentucky Academy of Science

1st Place Graduate Research Competition in Mathematics, 2013, Kentucky Academy of Science

Janet E. Mills Award for Outstanding Undergraduate Research in Mathematics, Seattle University