A COMPUTATIONAL PIPELINE FOR
GRAPHICAL MODELING OF INTEGRATED
BIOMEDICAL DATA

by

Vineet K. Raghu

BS Computer Science and Mathematical Biology, University of
Pittsburgh, 2014

Submitted to the Graduate Faculty of
the Department of Computer Science in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019
This dissertation was presented

by

Vineet K. Raghu

It was defended on

May 8th 2019

and approved by

Panos K. Chrysanthis, Department of Computer Science

Milos Hauskrecht, Department of Computer Science

Rebecca Hwa, Department of Computer Science

Olivera Finn, Department of Immunology

Peter Spirtes, Department of Philosophy, Carnegie Mellon

Dissertation Advisors: Panos K. Chrysanthis, Department of Computer Science, Panayiotis V. Benos, Department of Computational and Systems Biology
Clinical decision making and biomedical research have the potential to be revolutionized by the abundance of readily available, multi-modal data. One of the main drivers of this wealth of data is next-generation sequencing technologies such as RNA-Seq and Single-cell RNA-Seq. These methods enable high-throughput measurements of the genome at a granular level. However, to truly understand the causes of disease and the effect of medical interventions, this data must be integrated with phenotypic, environmental, and behavioral data from individuals. In addition, effective modeling methods that can infer causal relationships from this data are required. This presents a host of modeling challenges such as 1) high-dimensionality (low sample size and many variables), 2) redundancy among features, and 3) unmeasured variables that may confound the system under study. In addition, due to ethical concerns and cost, much of this data is observational. This means that no experimentally controlled perturbation was performed to measure the data. In this thesis, I present a pipeline to mine causal relationships from this integrated, observational biomedical data. The pipeline consists of three components: 1) Feature selection and clustering with prior knowledge, 2) Learning undirected graphical model structure to represent the joint distribution of mixed data, and 3) Learning causal graphical models with latent confounding. I demonstrate how this pipeline extracts useful knowledge via two cancer research applications: 1) Prediction of response to a prophylactic cancer vaccine and 2) Early detection of lung cancer from low-dose CT scans.
# TABLE OF CONTENTS

PREFACE .................................................................................. xiv

1.0 INTRODUCTION ...................................................................... 1
  1.1 Challenges ........................................................................... 2
  1.2 Our Approach ....................................................................... 2
  1.3 Research Contributions .......................................................... 4

2.0 PRELIMINARIES .................................................................... 5

3.0 EVALUATION AND INTEGRATION OF MULTI-SOURCE PRIOR
   INFORMATION ........................................................................... 8
  3.1 Background .......................................................................... 8
    3.1.1 Bayesian Approaches ......................................................... 9
    3.1.2 Regularization Approaches ............................................. 10
  3.2 Methods .............................................................................. 11
    3.2.1 Problem Definition .......................................................... 12
    3.2.2 Limiting Parameter Range .............................................. 13
    3.2.3 Determining Empirical Probabilities ............................... 15
    3.2.4 Evaluating Prior Knowledge Sources ............................. 16
    3.2.5 Parameter Selection ........................................................ 17
  3.3 Discussion ........................................................................... 20

4.0 VARIABLE SELECTION FOR GRAPHICAL MODELING .......... 22
  4.1 Background .......................................................................... 22
    4.1.1 Feature Selection in Genomics ...................................... 23
    4.1.2 Incorporating Prior Knowledge ....................................... 24
4.1.3 Graphical Models for Feature Selection ........................................... 25
4.1.4 Problem Statement ................................................................. 26
4.2 Methods ......................................................................................... 27
  4.2.1 Preferential Diversity (Pref-Div) .................................................. 27
  4.2.2 Prior Information Pref-Div (piPref-Div) ....................................... 29
  4.2.3 Simulated Datasets .................................................................... 29
  4.2.4 Real Data and Curated Prior Information ................................. 31
  4.2.5 Evaluation Metrics .................................................................. 33
4.3 Results .......................................................................................... 34
  4.3.1 Prior Knowledge Evaluation Results ......................................... 34
  4.3.2 Effect of Amount and Quality of Prior Knowledge ................... 34
  4.3.3 Sensitivity Analysis ................................................................. 37
  4.3.4 Breast Cancer Outcome Prediction .......................................... 40
  4.3.5 Stratification of Breast Cancer Sub-types ................................. 42
4.4 Discussion ...................................................................................... 44

5.0 UNDIRECTED GRAPHICAL MODELS FOR REPRESENTING JOINT
  DISTRIBUTIONS OF MIXED DATA ...................................................... 47
5.1 Background ..................................................................................... 47
  5.1.1 Markov Random Field ............................................................ 48
  5.1.2 Mixed Graphical Models .......................................................... 49
  5.1.3 Regularization and Prior Knowledge ......................................... 51
5.2 Methods ......................................................................................... 51
  5.2.1 Prior Information Mixed Graphical Models (piMGM) ................. 52
  5.2.2 Description of Data Sources ..................................................... 54
5.3 Results ............................................................................................ 56
  5.3.1 Prior Knowledge Evaluation Results ....................................... 57
  5.3.2 Structure Learning Results ..................................................... 60
5.4 Discussion ....................................................................................... 64

6.0 PREDICTION OF RESPONDERS TO A PROPHYLACTIC CANCER VACCINE ................. 66
6.1 Background ......................................................... 66
6.2 Methods ......................................................... 67
  6.2.1 Description of Data Sources ................................. 68
  6.2.2 Model Development and Evaluation ......................... 71
6.3 Results .......................................................... 71
  6.3.1 Evaluation of Prior Knowledge Sources .................... 72
  6.3.2 Evaluation of Week 2 Model Predictions .................. 72
  6.3.3 Biological Significance of Week 2 Models ................ 77
  6.3.4 Graphical Models from Pre-Vaccination Data .............. 81
6.4 Discussion ....................................................... 83

7.0 CAUSAL DISCOVERY FROM DATA WITH LATENT VARIABLES 86
7.1 Background ...................................................... 86
  7.1.1 Causal Structure Learning Algorithms ..................... 89
  7.1.2 Structure Learning from Data with Latent Confounders .... 91
7.2 Methods ......................................................... 95
  7.2.1 MGM-FCI-MAX ............................................... 95
  7.2.2 Parallelization and Optimization ........................... 95
  7.2.3 Independence Tests for Mixed Data ....................... 97
  7.2.4 Dataset Simulation and Evaluation Metrics ................. 98
7.3 Results .......................................................... 100
  7.3.1 Comparison of FCI and Variants ........................... 100
  7.3.2 Incorporating MGM for Fast Skeleton Discovery ........... 103
  7.3.3 Scaling to 500 Node Networks .............................. 106
7.4 Discussion ....................................................... 108

8.0 IDENTIFICATION OF CAUSAL FACTORS FOR EARLY DETECTION OF LUNG CANCER 110
8.1 Background ...................................................... 110
8.2 Methods ......................................................... 111
  8.2.1 Study population ............................................. 111
  8.2.2 Development and Evaluation of Prediction Model .......... 112
LIST OF TABLES

5.1 Differentially regulated pathways by receptor status (positive: Luminal A and B; negative: HER2, triple negative) in Breast Cancer (FDR < 0.1) Pathways in red found by piMGM but not NetGSA. .............................................. 59

6.1 Clinical characteristics of patients whose PBMC’s were sequenced at two weeks post-vaccination. ................................................................. 68

6.2 Clinical characteristics of non-responder Group 1 patients whose PBMC’s were sequenced pre-vaccination. .................................................. 69

6.3 Clinical characteristics of non-responder Group 2 patients whose PBMC’s were sequenced pre-vaccination. .................................................. 70

6.4 Evaluation of Prior Knowledge Sources used in the MUC1 Modeling Study . 73

6.5 Evaluation of STRINGdb component sources used in the MUC1 Modeling Study 74

6.6 Comparison of prediction models for antibody response to MUC1 vaccine built from RNA-Seq data two weeks post-vaccination. ................. 75

6.7 Top five pathways significantly different than random priors on RNA-Seq samples from non-responders. ......................................................... 79

6.8 Top five pathways significantly different than random priors on RNA-Seq samples from responders. ......................................................... 80

6.9 Evaluation of structure learning methods to discriminate responders from non-responders using the pre-vaccination RNA-Seq data. .................... 82

7.1 Experimental Orientation Score to compute orientation precision and recall for latent variable causal discovery algorithms. ......................... 100
8.1 Features selected by MGM-FCI-MAX for lung cancer prediction in each round of cross validation .................................................. 114
8.2 Description and performance of Lung nodule discrimination models on the training cohort. AUC is reported with 25\textsuperscript{th} and 75\textsuperscript{th} percentiles. .................. 116
LIST OF FIGURES

1.1 Our complete modeling pipeline. The pipeline takes multi-modal data as input (currently -omics and clinical data) and filters this data to fewer features that can be modeled. This data is merged with domain knowledge to learn an undirected graphical model. Finally, the undirected model is converted to a causal graph that identifies latent confounding. ................................. 3

3.1 Illustration of procedure to limit tested parameter range. ....................... 15

3.2 Subsampling procedure to determine empirical probabilities for each feature to be selected. $B$ is the base selection procedure given hyperparameter $\lambda$ ... 16

3.3 Merging of prior and empirical information for a feature. Panel (a) shows normal distributions for the amount of expected occurrences across subsamples for each prior and for their weighted mixture (black line). (b) The mixture distribution ($F_k^{KL}$) is approximated by a normal distribution. (c) This normal distribution is combined with the empirical distribution to give a posterior normal distribution. .............................................. 19

4.1 Cluster Simulation (left) and Pathway Simulation (right) for piPref-Div simulated datasets. Purple nodes are master regulators of a cluster, blue nodes are causal parents of the target variable, gray variables belong to multiple clusters, and the beige node is the target variable. .............................................. 30

4.2 Predicted Weight vs. Net Reliability for each prior knowledge source in simulated experiments for piPref-Div for (left) 50 samples and (right) 200 samples. 35
4.3 Accuracy of predicted clusters for varying amount and reliability of prior knowledge. Graphs were simulated according to the cluster simulation method. Sample size was set to 50 (left) and 200 (right). .......................... 36

4.4 Accuracy of predicted clusters for varying amount and reliability of prior knowledge. Graphs were simulated according to the pathway simulation method. Sample size was set to 50 (left) and 200 (right). .......................... 37

4.5 Accuracy of predicted clusters for varying amount and reliability of prior knowledge on large datasets. Graphs were simulated according to the cluster simulation method. Sample size was set to 50 (left) and 200 (right). .......................... 38

4.6 Impact of the WP Cutoff parameter on the ability of piPref-Div to select relevant clusters. .......................... 38

4.7 Sensitivity of piPref-Div to the number of features selected. Graphs were simulated according to the cluster simulation method with 50 clusters and 25 relevant (true) clusters. Each point is averaged over 15 simulations. Left graph shows the proportion of all clusters selected as a function of number of selected features. Right graph shows true clusters only. .......................... 39

4.8 AUC of predicting RFS using several feature selection methods on six independent breast cancer microarray datasets. .......................... 41

4.9 Stability of learned models for predicting RFS using several feature selection methods on six independent breast cancer microarray datasets. .......................... 43

4.10 Graphical model of breast cancer sub-type. Size of each edge represents the number of times a similar cluster was selected to be related to Subtype in each of the cross-validation folds. .......................... 45

5.1 Predicted deviance score $\tau$ vs. true reliability for simulated data of 100 variables, 1000 samples (left) and 200 samples (right) .......................... 58

5.2 Edge differences between receptor positive and negative breast tumor samples for a common sub-network of the T cell and B cell receptor signaling pathways (left) and the T cell receptor signaling pathway (right). Red (green) edges have higher (lower) empirical probability in receptor positive cancer according to piMGM .......................... 60
5.3 Accuracy of piMGM inferred networks vs. other approaches on simulated data with 100 variables, 500 samples and 5 prior sources. The datasets differ in the percent of edges with prior information (0, 10, 30, 60%) and the number of reliable priors (1, 3, 5 out of 5). The blue bars represent the $F_1$ score accuracy of the baseline networks (no priors). Priors gave information about only true edges (left) or all edges (right) chosen randomly.

5.4 Accuracy of piMGM inferred networks vs. other approaches on simulated data with 100 variables, 200 samples and 5 prior sources. Priors gave information about only true edges (left) or all edges (right) chosen randomly.

5.5 Results of applying piMGM to the TCGA Breast Cancer dataset with varying prior information sources. Classification accuracy (left), feature stability (center), and number of features (right) were computed.

6.1 Predicted vs. Actual Log IgG at 12 weeks post vaccination for three different prediction models: StEPS (Left), piMGM with String Prior (Center), and PD-PCA + StEPS (Right).

6.2 Higher levels of NFkB signaling via phosphorylation of p65 in High Responders at Week 2 (left). Larger change from baseline to Week 2 in NFkB signaling via phosphorylation of p65 in Responders (right). Figure courtesy of Dr. Mark Cameron at Case Western Reserve University.

6.3 Features selected by piMGM with STRING Pathway prior knowledge in any round of cross-validation. Node size indicates number of cross-validation rounds in which it was selected. Green color is a positive association with response. Clusters were selected separately using piPref-Div and StEPS. Transparent nodes were not selected from the full dataset.

6.4 Visualization of model predictions (left) and learned graphical model structure (right). Model was learned from the baseline RNA-Seq data pre-vaccination.

6.5 Lower frequency of B cells in Responders at Week 0 (left). Higher change in B-cell frequency in responders from baseline to Week 2 (right). Figure courtesy of Dr. Mark Cameron at Case Western Reserve University.
7.1 Precision-Recall Plots of FCI, FCI-MAX, CFCI and BCCD for 50 Node Scale-Free Networks with 1000 Samples and 5 latent variables. Plots are separated by adjacency and orientation recovery in each row, and by edge type in each column; (CC) refers to only edges between two Continuous variables, (CD) refers to edges between Continuous and Discrete variables, while (DD) refers to edges between two Discrete variables.

7.2 Precision-Recall Plots of FCI, FCI-MAX, CFCI and BCCD for 50 Node Scale-Free Networks with 200 Samples and 5 latent variables. Plots are separated by adjacency and orientation recovery in each row, and by edge type in each column; (CC) refers to only edges between two Continuous variables, (CD) refers to edges between Continuous and Discrete variables, while (DD) refers to edges between two Discrete variables.

7.3 Runtime Experiment with 50 Nodes, 1000 Samples, and 5 Latent Variables. All MGM algorithms were set to $\lambda = 0.15$.

7.4 Runtime (s) measured as a function of the number of processors for varying chunk size parameter $\beta$ on 50 Node Networks with 1000 Samples.

7.5 Precision-Recall Plots of CFCI, FCI-MAX, and MGM Variants for 50 Node Scale-Free Networks with 1000 Samples and 5 latent variables.

7.6 Precision-Recall Plots of FCI, MGM-FCI, and MGM-FCI-MAX for 500 Node Scale-Free Networks with 500 Samples and 20 latent variables.

7.7 Runtime Experiment with 500 Nodes, 500 Samples, and 20 Latent Variables.

8.1 Learned causal model using MGM-FCI-MAX on the high-risk training cohort for lung cancer detection.

8.2 ROC Curves of Lung Cancer Prediction on Cross-Validation Testing Sets.

8.3 Kernel density plot of predicted probability of lung cancer for the Brock Parsimonious Model (left) and for LCCM (right).

8.4 Sensitivity-Specificity plot of predictions made by LCCM.
PREFAE

Just like raising a child, it takes a village to complete a doctoral degree. Here, I would like to acknowledge my village; those individuals who gave me academic guidance, moral support, and emotional stability during my years in graduate school.

Thank you to my advisors Takis Benos and Panos Chyrsanthis for agreeing to take a chance on me as your student. The ability to create a critical and educational environment for students that is still nurturing is a unique gift. Thank you for sharing all of the science and laughs with me these past five years. Moreover, thank you for the freedom that you have entrusted to me to explore my research topic and mentor the next generation of young scientists. I hope to one day meet the standards that you have shown me when I am advising students of my own.

Thank you to the rest of my committee for all of your advice and support during my comprehensive examination, my proposal defense, and my dissertation defense. The ideas and results of this thesis are as much yours as they are my own.

Thank you to my family: Mom, Dad, Vikram, Shobhitha, and Vedyan for everything. A roller coaster metaphor does not come close to capturing the ups and downs of the research journey. The unconditional love and support that you provide has meant the world to me. To my parents, thank you for the financial support that made my academic journey significantly less stressful. You allowed me to carve my own path and learn from my own mistakes, and for that I am forever grateful.

Lastly, thank you to my colleagues and collaborators. Thank you to Olja Finn, Michelle Miller, and Kristen Critelli for their help on the MUC1 vaccine project and for teaching me basic Immunology, and thank you to Mark and Cheryl Cameron for their work on this project. Thank you to David Wilson and Jiantao Pu for their help to build the lung cancer
prediction model and for providing me with the CT scan data. Thank you to Clark Glymour, Joe Ramsey, and Peter Spirtes for all of the enjoyable and crazy meetings to get us up to speed on causal discovery and for helping me with the theoretical aspects of my projects. Thank you to Dimitrios Manatakis for starting me down the path of prior knowledge and for being a great collaborator. Finally, thank you to the administrators and staff at the Department of Computer Science and Computational and Systems Biology for being a joy to work with, and for doing all of the behind the scenes work that makes everything tick.

This thesis is dedicated to the loving memory of my grandmothers: Ranga Paati and Viji Paati.
1.0 INTRODUCTION

Biomedical research has long been impeded by the inability to measure factors necessary to understand disease onset, progression, and treatment effectiveness. However, technology has advanced to a point where researchers *can* measure many of these factors. Next generation sequencing has enabled granular measurements of molecular data from individuals [130]. Advances in biomedical imaging technologies have provided a visual view of cellular phenotypes. Connected, wearable technology (e.g. smartphones, watches, etc.) have enabled continuous monitoring of environmental and physiological signals [43]. Together, these multi-modal data types give researchers the potential to 1) understand the fundamental causes of disease, 2) prioritize promising hypotheses about human biology, and 3) personalize medical treatments to individuals. The major challenge in achieving these objectives is the lack of effective modelling techniques to understand interactions between these complex signals [61].

Machine learning (ML) is a popular tool for prediction and modeling in biomedical settings [31]. However, ML is not suitable for these goals because ML models are not constrained to be interpretable by humans [114]. Further, researchers often do not have a specific hypothesis in mind, and wish to generate hypotheses from just *observing* a system. Hypothesis generation requires causal knowledge which ML models cannot infer from observational data [122].

Probabilistic graphical models (PGM’s) are a tool that may be able to address the problem of hypothesis generation from observational data [65]. Graphical models are a data exploration technique which represent the variables in a dataset as a graph. In the graph, nodes correspond to variables in the data, and edges correspond to statistical dependence. With this representation, researchers can quickly query the graph to understand the direct relationships among molecular, environmental, and behavioral factors with health outcomes.
This graphical form is an intuitive one to understand biological phenomena, as a large portion of biological knowledge is understood as networks (pathways) [60, 35]. Under certain assumptions, these models can be interpreted as causal models [124, 94]. These models still have computational limitations on large data [98] and lack demonstrated successes on complex, multi-modal biomedical data [63, 64, 111].

1.1 CHALLENGES

Despite their potential, learning PGM structure from biomedical data presents a number of obstacles to overcome. Integrated biomedical datasets have mixtures of continuous and categorical data, which most structure learning algorithms are not well suited to handle [20, 25]. Further, most genomic datasets are high-dimensional (many variables and few samples) [152] and have large correlations among measured genes [139]. Algorithmic techniques to learn graphical models must be efficient to handle the high-dimensionality of the data, robust to learn a stable model despite this correlated nature, and accurate to prevent spurious hypotheses. I define an effective graphical model as one that meets these three conditions.

1.2 OUR APPROACH

A way to address the inefficiency of graphical structure learning is to use a multi-step approach where promising relationships in the data are prioritized quickly before fine-tuning the final model. A way to address the highly correlated nature of expression data is to use external domain knowledge to identify the most biologically plausible of the many nearly equivalent sets of variables. To this end, the central hypothesis of this thesis is that learning graphical model structure from integrated biomedical data will be more effective when using multi-step pipelines and when incorporating external domain knowledge. Since ground truth graphical model structure is not available for most biological applications, I use prediction of important target variables as a proxy for graph accuracy.
To address the hypothesis of this thesis, I have designed a full graphical modeling pipeline (Figure 1.1). The pipeline begins with an integrated biomedical dataset from several sources. In this thesis, we focus upon integrating clinical, demographic, and genomic data. First, this dataset is filtered using a variable selection scheme that incorporates prior knowledge to identify a set of variables that is relevant to outcomes of interest, yet diverse. Then, prior knowledge is combined with the filtered data to learn undirected graphical structure. Finally, cause and effect relationships are mined using a search procedure that can identify and handle latent variables in observational data. I test the developed methods on two biomedical research applications. First, I test the variable selection and undirected structure learning methods on an application to predict response to a prophylactic cancer vaccine. Then, I test the causal discovery method on an application to identify causal factors of lung cancer and to build a model to accurately detect lung cancer in a high-risk population.
1.3 RESEARCH CONTRIBUTIONS

Each chapter in this thesis is structured as follows. First, I give a brief introduction to the topic covered in the chapter. Then, I give necessary background information and a review of relevant literature. Next, I discuss the developed computational methods in detail and the evaluation protocol for each method. After, I present evaluation results of each computational method on both synthetic datasets and real datasets to ensure correctness and practicality. Lastly, I give a brief discussion about the method including the benefits, limitations, and potential future work. In this thesis, I occasionally use the word “we” to recognize work that was completed collaboratively.

Overall, this thesis presents a computational pipeline for knowledge discovery from integrated biomedical data. The specific contributions are as follows:

- I present a computational framework to evaluate and integrate potentially unreliable and incomplete prior information from multiple sources (Chapter 3).
- I apply the framework to develop a method for variable selection as a precursor to graphical modeling, and I evaluate the approach on synthetic data and real breast cancer transcriptomic data. (piPref-Div, Chapter 4).
- I again apply the framework to develop a method to learn undirected graphical model structure from categorical and continuous data, and evaluate this approach on synthetic and real breast cancer data transcriptomic and clinical data. (piMGM, Chapter 5).
- I apply both the developed methods to a biomedical research application: predicting response to a prophylactic cancer vaccine from clinical trial transcriptomic and clinical data (Chapter 6).
- I present an algorithm for causal discovery from data with latent variables and demonstrate its efficiency and accuracy on synthetic data with mixed variables. (MGM-FCI-MAX, Chapter 7).
- I apply the causal discovery method to build a model for the early detection of lung cancer from low-dose CT scan, smoking, and demographic factors. (Chapter 8).
2.0 PRELIMINARIES

In this chapter, I give important definitions about graphs that are useful to understand the remainder of the thesis. The definitions presented here are applicable to all chapters in the thesis, and further definitions important to specific chapters can be found in the background sections. Most of these concepts are from graph theory, so I refer the reader to [9, 65] for further reading.

**Definition 2.1.** A **Graph** \((G = (V, E))\) consists of a set of vertices \(V\), and a set of edges \(E\). An **edge** \((E = (V_i, V_j, EP_x, EP_y))\) is a four-tuple consisting of two vertices \(V_i\) and \(V_j\) and two endpoints \(EP_x\) and \(EP_y\).

Depending upon the type of graph, the set of possible endpoints may differ. For simplicity, in this thesis we represent the edge in short form. For example, the edge \(E = (X, Y, -, >)\) would be represented as \(X \rightarrow Y\). Adjacencies refer to edge existence, and orientations refer to edge endpoints.

**Definition 2.2.** An **Undirected Edge** is an edge \(E\) such that both endpoints have no orientation (e.g. \(E = (V_1, V_2, -, -)\))

**Definition 2.3.** A **Directed Edge** from \(X\) to \(V\) is an edge \(E = (X, V, - , >)\). This will be written as \(X \rightarrow V\) in this thesis.

**Definition 2.4.** A **Parent** of a variable \(V\) in graph \(G\) is a variable \(X\) such that the edge \(E = (X, V, - , >)\) \(\in G\). For a causal graph, \(X\) would be a direct cause of \(V\). The set of all parents of \(V\) in \(G\) is denoted \(Pa (V, G)\).

**Definition 2.5.** A **Child** of a variable \(V\) in graph \(G\) is a variable \(X\) such that \(V\) is a parent of \(X\) in \(G\). The children set of a variable \(X\) in graph \(G\) is denoted \(Ch (X, G)\)
Definition 2.6. A variable $V$ is adjacent to a variable $X$ in $G$ if and only if there exists an edge $E = (X, V, *, *) \in G$ where $*$ denotes any valid endpoint.

Definition 2.7. The Markov Blanket of a variable $V$ in a graph $G$ is the set of variables $S$ where $S$ is defined as:

$$ S = \text{Adj} (V, G) \bigcup_{X \in \text{Ch}(V, G)} \text{Pa} (X, G) / V \tag{2.1} $$

The Markov Blanket in a graph is given by the parents of the variable, children of the variable, and other parents of the children of the variable (spouses). This set of variables is often used for prediction, as it contains all of the information about a variable that the graph has to offer.

Definition 2.8. The Skeleton of a graph $G$ is a graph where all variables and edges are the same, except the endpoints of all edges are changed to ")-".

Definition 2.9. A Path between $X$ and $V$ in $G$ is a sequence of edges $S$ such that all edges in $S$ are in $G$ and:

$$ S = [(X, V_1, *, *), (V_1, V_2, *, *), \ldots, (V_k, V, *, *)] \tag{2.2} $$

Definition 2.10. A Directed Path from $X$ to $V$ in $G$ is a sequence of edges $S$ such that all edges in $S$ are in $G$ and:

$$ S = [(X, V_1, -, >), (V_1, V_2, -, >), \ldots, (V_k, V, -, >)] \tag{2.3} $$

An Almost or Semi-Directed Path replaces all ")-" endpoints in $S$ with "*--".

Definition 2.11. A Cycle in a Graph $G$ is a directed path that begins and ends at the same variable.

In the causal discovery algorithms discussed here, I assume that no cycles exist in the data generating causal process.

Definition 2.12. A Descendant of a variable $X$ in a graph $G$ is any variable $V$ such that there is a directed path from $X$ to $V$ in $G$. The set of all descendants of $X$ in $G$ is denoted $\text{De} (X, G)$.
Definition 2.13. A Clique $C = (V_1, \ldots, V_k)$ in a graph $G$ is a set of variables such that:

$$\forall i, j \ V_i \in C \ and \ V_j \in C \rightarrow V_i \in Adj(V_j, G)$$ (2.4)

In other words, this is a set of variables such that all variables are adjacent to one another in the graph.

Definition 2.14. A Maximal Clique is a clique $C$ in $G$ such that

$$\forall X \ s.t. \ X \notin C \ and \ X \in G, \exists V \in C \ s.t. \ X \notin Adj(V, G)$$ (2.5)

This is a clique where no more variables can be added to the set such that $C$ remains a clique.

Definition 2.15. A Collider in a graph $G$ is a triple of variables $C = (X, Y, Z)$ such that the following edges $E_1, E_2 \in G$. $E_1 = (X, Y, *, >)$ and $E_2 = (Z, Y, *, >)$. If $X \notin Adj(Z, G)$ then this is referred to as an Unshielded Collider, otherwise it is Shielded.

Definition 2.16. A Collider $C$ on a Path $P$ is a collider $C$ such that all edges in $C$ are contained in the path $P$.

Definition 2.17. A Unshielded Triple is an Unshielded Collider without the arrowhead orientations. Formally, a triple of variables $C = (X, Y, Z)$ is an unshielded triple in a graph $G$ if and only if $(X, Y)$ and $(Y, Z)$ are connected with an undirected edge in $G$, and $X \notin Adj(Z, G)$
This chapter discusses strategies for using prior knowledge to improve model selection. We present a general approach that evaluates and integrates prior information from multiple sources and uses this information for model selection. This approach is instantiated to particular modeling problems of interest in Chapters 4 and 5 for variable selection and undirected graphical modeling, respectively.

### 3.1 BACKGROUND

There are a wealth of external databases with rich knowledge about the relationships between genes, proteins, and phenotypic outcomes (diseases). This creates an opportunity for modeling methods to leverage this information to improve the accuracy and stability of learned models. This also creates challenges, since this information is largely incomplete and in some cases, inaccurate for the analysis task at hand. As an example, prior knowledge for genomic data consists of gene interaction maps or gene regulatory pathways. Though these regulatory interactions are correct in general, in a specific cell or disease context these interactions may be disrupted. This necessitates building methods to incorporate prior knowledge that better handle these situations.

Using prior knowledge to improve statistical inference and modeling is a well-studied topic. In this section, I give some background as to how prior knowledge has been used to improve models. Since this is a very broad topic, I narrow the focus to approaches that have been used for genomic datasets because these approaches take into account the modeling
challenges inherent to biomedical data. These approaches cluster into two main categories: those that use Bayesian strategies to infer a posterior distribution for a parameter of interest and those that use prior knowledge to group related variables when developing an objective function to optimize.

3.1.1 Bayesian Approaches

Bayesian approaches are a broad category of statistical models that utilize Bayes Theorem (Equation 3.1) to combine prior knowledge with empirical data. Here, \( p(M \mid D) \) is referred to as the posterior distribution over models, \( p(D \mid M) \) is referred to as the likelihood of the data given the model, and \( p(M) \) is the prior distribution over models. \( p(D) \) is a normalizing constant equal across all models that ensures the posterior distribution satisfies the rules of a probability distribution.

\[
p(\text{Model } M \mid \text{Data } D) = \frac{p(D \mid M) \ast p(M)}{p(D)} \tag{3.1}
\]

Once the posterior distribution over models is computed, it can be used to determine the most likely model. Alternatively, Bayesian Model Averaging can be used to extract weighted predictions from all models [52]. As empirical evidence is collected, the likelihood function is updated which results in an updated posterior distribution. In this way, the Bayesian approach allows for continuous updates to a model with increasing data. The challenge in Bayesian modeling is to specify the prior distribution over models so that it is grounded in some truth but is computationally feasible to evaluate for the large space of models [138, 22]. One method is to use an uninformative prior that results in an analytical solution for the posterior that avoids computational hardness. When used in this way, the Bayesian approach is beneficial to avoid overfitting with small samples [117, 71]. This however does not leverage informative prior knowledge to improve models.

Some Bayesian approaches have attempted to curate external data to form informative prior distributions. In [95], the authors develop a Bayesian framework to select informative genes for a target variable. They develop a prior distribution to quantify a degree of belief as to whether a gene is an important predictor of a target. This prior is proportional to
whether other genes in co-occurring pathways are predictive of the target. A similar use of prior knowledge for gene selection is given in [67]. Here, the authors select genes by learning an undirected graphical model. The prior knowledge consists of a degree of belief in whether genes should be connected to one another in the graph. The difficulty with these approaches is that the methods do not specify how to integrate multiple sources of knowledge coherently. An approach that directly addresses this concern to learn a Bayesian Network of genomic data is given in [82]. The authors require users to specify concordance functions which determine how well a graphical structure agrees with prior knowledge about graph topology or individual connections. A weighted combination of these concordance functions are then taken to arrive at a final prior distribution. Despite this solution, challenges still remain in determining the relative importance of priors versus likelihood and determining how to set hyper-parameters for these models. Further, these methods assume that the user has given only accurate sources of prior information which may not be the case in practice.

3.1.2 Regularization Approaches

A second class of methods use prior knowledge to group features based on their functional similarity or relatedness. A motivating example for this approach is addressing the problem of feature selection for predictive modeling. Here, the aim is to identify a small set of features that accurately predict a variable of interest. A full discussion of the feature selection problem is given in Chapter 4, so here we focus upon one particular method that has shown effectiveness in incorporating prior knowledge: Least Absolute Shrinkage and Selection Operator (LASSO) regression [133].

Assume that there is a continuous target variable of interest $Y$, then the original LASSO formulation is given in Equation 3.2. The first term is the same as least squares linear regression with target variable $Y$, data matrix $X$ ($X_i$ refers to the $i^{th}$ sample of $X$) with $N$ independent samples, and coefficient vector $\beta$.

$$\begin{align*}
\argmin_{\beta} = \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \sum_j | \beta_j |
\end{align*}$$

The second term of this expression represents the $L_1$ norm of the coefficient vector. This
norm induces sparsity on the coefficient vector, which means that many of the coefficients will be shrunk to zero. The nonzero coefficients then correspond to the ”selected” features with the hyper-parameter $\lambda$ controlling the number of selected features. Though this approach has been successfully utilized for a host of problems in biology, LASSO faces an issue when dealing with a set of correlated features. The algorithm tends to select just one of the correlated features and excludes the rest, which can lead to good accuracy but limited interpretability [148]. Prior knowledge can be used to address this concern via the group LASSO (Equation 3.3) [151].

$$\begin{align*}
\text{argmin}_\beta &= \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \sum_{l=1}^{L} \sqrt{df_l} \| \beta_l \|_2 \\
(3.3)
\end{align*}$$

Here, the first term is equivalent to the LASSO formulation; however, the penalty term is an $L_2$ norm that is summed across different groups of features. $L$ refers to a set number of these groups and $df_l$ is the degrees of freedom for group $l$ which allows for standardization across groups of different sizes. This penalty term induces sparsity across groups, meaning that groups of features are selected together. Prior knowledge for this type of method typically consists of pre-defined biological pathways or networks of genes [4, 5, 153]. Again, these methods require gene groups to be correctly specified, which is difficult since our current understanding of functional gene groups is limited [99]. A slew of methods have incorporated prior knowledge in a similar way for other classification models [140, 73] and for graphical structure learning [143, 13]. However, these methods still suffer from the same problems.

### 3.2 METHODS

In this section, we propose a general method for incorporating incomplete and potentially unreliable prior knowledge from several sources. The method aims to select reliable features based upon the data and the prior knowledge sources. In addition, the method evaluates each prior knowledge source to determine how well the prior knowledge is expressed in the data (the prior’s reliability). I begin by defining the problem, and then I discuss the details of the method to evaluate the prior knowledge sources and to select quality features.
3.2.1 Problem Definition

Assume that we have a dataset $S$ of size $n \times p$ where $n$ is the number of samples (rows) and $p$ is the number of variables (columns). Assume further that this data was generated according to some true structural equation model (SEM) (Equation 7.1). The desired output for this problem is some set of features that describe the graph $G$ of the SEM. Some specific examples that will be discussed in this dissertation are to select the edges in $G$ or to select the neighborhood of a variable of interest in $G$. Lastly, assume that we have some base selection procedure $B(\lambda, S)$, where $\lambda$ is a set of parameters for the selection procedure. We assume further that the amount of features selected by $B$ is inverse monotonically related to the value of $\lambda$ (larger $\lambda$ means fewer selected features). Lastly, we assume that the way $B$ uses the parameter value $\lambda$ can be applied separately to two sets of potential features: those where prior information exists ($wp$ "with prior") and those where prior information is not available ($np$ "no prior"). An example of a choice for $B$ for the feature selection problem is LASSO regression, where features that are in $wp$ could be penalized separately from those in $np$.

Next, assume we have $R$ sources of prior information $T = (t_1, t_2, \ldots, t_R)$. The prior information for a particular source $t_r$ is a vector of size $K$ where each element $m_{tk}^r$ describes the belief the expert has that the corresponding feature $k$ should appear in the graph structure. $K$ here is the total number of possible features (e.g. for the neighborhood selection problem $K = p - 1$ ). Note that the prior information source may not provide information for all possible features, in which case $m_{tk}^r$ is NULL if prior information for element $k$ from source $t_r$ is unknown.

The intended outputs are twofold. First, the algorithm should accurately output the features of interest in the graph structure. In particular, the algorithm should select features more accurately when prior information is more plentiful and more accurate. But, it should perform no worse than the base procedure when prior information is unreliable. Secondly, the algorithm should accurately evaluate each prior knowledge source and determine an appropriate set of weights $w$ such that $\sum_{i=1}^{R} w_i = 1$. These weights represent the relative confidence that the model has in each prior information source.
The main idea of the method is to use a Bayesian approach to identify a posterior distribution for the probability of each individual feature. Then by using all of these distributions, the method separately selects a set of parameters for features where prior information is available and where no prior information is available. This is done by comparing the output of the base selection procedure across a range of parameters to what the posterior distribution expects. Finally the base selection procedure is run with this set of selected parameters to get a final output.

Next, we give the details of the method. It proceeds in four main steps.

1. First, an appropriate parameter range is determined. This is done by identifying the range where the fewest features are selected yet changing the parameter values slightly results in a large change in the number of selected features.
2. Then, a subsampling approach is used to compute empirical probabilities of appearance for each feature using the base selection procedure on the data alone.
3. Next, the prior knowledge sources are evaluated against these empirical probabilities across all features.
4. Finally, the posterior distributions are computed and $\lambda^*$ is selected based upon its concordance to the posteriors.

### 3.2.2 Limiting Parameter Range

For biomedical research applications, a sparse model is preferable to an overly dense representation because the aim of these models is to suggest potential hypotheses to test. In addition, for graphical models in genomics, the expectation is that the underlying structure should be sparse [143]. Here, we discuss a procedure to limit the parameter selection range for $\lambda$ based on these principles. The idea is to choose a relatively large value for $\lambda$ to ensure a sparse representation, but test a range where a small change in $\lambda$ results in a large change in the number of selected features. This ensures that each value in the tested range is meaningfully different.

Figure 3.1 illustrates the method, and the procedure is formalized in Algorithm 3.1. Initially $J$ parameter values are tested across the full range of possibilities: $\Lambda = (\lambda_1, \ldots, \lambda_J)$. 
Algorithm 3.1 Parameter Range Selection Method

1: procedure RANGE-RESTRICT(Data $D$, Parameter Set $\Lambda$, Base Selector $B$)
2: 
3: \hspace{.2cm} count = 1
4: \hspace{.2cm} numFeatures = []
5: \hspace{.2cm} for $\lambda \in \Lambda$ do
6: \hspace{.4cm} numFeatures[count] = $| B(\lambda, D) |$
7: \hspace{.2cm} $KP_1 = \text{getKneePoint}(numFeatures)$
8: \hspace{.2cm} $KP_2 = \text{getKneePoint}(numFeatures[1 : KP_1])$
9: \hspace{.2cm} $KP_3 = \text{getKneePoint}(numFeatures[KP_1 : | \Lambda |])$
10: \hspace{.2cm} return $\text{GetRange}(KP_2, KP_3, \Lambda)$

11: procedure getKneePoint(Array of Feature Counts $numFeatures$)
12: \hspace{.2cm} minError = $\infty$
13: \hspace{.2cm} minIndex = -1
14: \hspace{.2cm} length = $| numFeatures |$
15: \hspace{.2cm} for $i \leftarrow 2$ to $length - 1$ do
16: \hspace{.4cm} error = LinearFit($numFeatures[1 : i]$)
17: \hspace{.4cm} error = error + LinearFit($numFeatures[i + 1 : length]$)
18: \hspace{.4cm} if $error < minError$ then
19: \hspace{.6cm} minIndex = i
20: \hspace{.6cm} minError = error
21: \hspace{.2cm} return minIndex

22: procedure GetRange(Knee Point Index i, Knee Point Index j, Parameter Set $\Lambda$)
23: \hspace{.2cm} length = $| \Lambda |$
24: \hspace{.2cm} $\Lambda^* \leftarrow$ Empty Array of Size length
25: \hspace{.2cm} $\Lambda^*[1] \leftarrow \Lambda[i], \Lambda^*[length] \leftarrow \Lambda[j]$
26: \hspace{.2cm} rangeFactor $\leftarrow \frac{\Delta[length]-\Delta[1]}{\Delta[j]-\Delta[i]}$
27: \hspace{.2cm} for $n \leftarrow 2$ to $length - 1$ do
28: \hspace{.4cm} $\Lambda^*[n] \leftarrow \Lambda^*[n-1] + \frac{\Delta[n]-\Delta[n-1]}{\text{rangeFactor}}$
29: \hspace{.2cm} return $\Lambda^*$
This produces a monotonically decreasing curve of the value of $\lambda$ vs. the number of selected features. First, a "knee point" is identified by examining each point at a time and determining which point best separates the curve into two lines. This procedure is then repeated for the left side of the "knee point" and the right side of the "knee point". The identified points from steps 2 and 3 give the final range to test. Finally, we use $J$ parameter values linearly spaced between the points identified in steps 2 and 3 as the final set $\Lambda$.

### 3.2.3 Determining Empirical Probabilities

Using this set of parameter values $\Lambda$, we next determine the empirical probability of each possible feature. For this, we randomly draw $q$ subsamples of size $b \leq n$ from the dataset $S$: $Q = (S_1, \ldots, S_q)$. Here, each subsample is drawn without replacement. Then, for each $\lambda_i \in \Lambda$ and each $S_j \in Q$, we compute $B(\lambda_i, S_j)$ using the base selection procedure $B$ (Figure 3.2).

This results in a set of selected features for each iteration $F = (F_1, \ldots F_{q*J})$. Then the

---

*Figure 3.1: Illustration of procedure to limit tested parameter range.*

![Figure 3.1: Illustration of procedure to limit tested parameter range.](image)
probability of presence $P_k$ for a particular feature $k$ is given in Equation 3.4. $P_k$ is an empirical frequency of appearance across all parameter values tested and all subsampled datasets. This quantity is next used to evaluate each prior knowledge source.

$$P_k = \frac{N_k}{q \times J}, \text{ where } N_k = \sum_{i=1}^{q \times J} \mathbb{1}_{k \in F_i} \quad (3.4)$$

### 3.2.4 Evaluating Prior Knowledge Sources

Next, we discuss how to evaluate prior knowledge sources with respect to the empirical probability of feature presence computed in the previous section. Recall that each prior information source $(t_r)$ is a vector of size $K$ where each element $m^{t_r}_k$ is the probability of feature $k$ to appear in the true graph structure. First, we compute the expected number of times for feature $k$ to appear across the subsamples according to prior $t_r (\phi^{t_r}_k)$ (Equation 3.5).

$$\phi^{t_r}_k = m^{t_r}_k \times q \quad (3.5)$$
\[ \mu_k = P_k * q \]  
(3.6)

The corresponding expected number of appearances from the subsampling procedure \( \mu_k \) can be computed using the probability of presence from the data \( P_k \) (Equation 3.6). Using these quantities, we then estimate the reliability of source \( t_r \) based on the average deviance between the expected and actual rates of appearance (Equation 3.7). Here, \( w_{p^{tr}} \) refers to the set of features where source \( t_r \) provided non-null prior information.

\[ \tau_{tr} = \frac{\sum_{k=1}^{\left| w_{p^{tr}} \right|} \left| \phi_k^{tr} - \mu_k \right|}{\left| w_{p^{tr}} \right|} \]  
(3.7)

\( \tau_{tr} \) is a measure of deviance for the prior \( t_r \). A high value of \( \tau \) implies low reliability. The final step to get weights for each prior source is to convert \( \tau \) into a standardized reciprocal form (Equation 3.8).

\[ w^{tr} = \frac{a^{tr}}{\sum_{i=1}^{R} a^{ti}} \cdot \frac{\tau_{ti}}{\tau_{tr}} \]  
(3.8)

Statistical significance of the deviance of the prior compared to a random prior of equivalent size is assessed using the following procedure. Given a prior knowledge source \( t_r \), an empirical null distribution is estimated by randomly permuting the labels of the information provided by \( t_r \) repeatedly. For each of these permutations, \( \tau \) is computed. Then, the \( p \)-value of \( t_r \) is the percent of \( \tau \) values in the null distribution greater than \( \tau_{tr} \). Thus, this \( p \)-value quantifies how well represented the source is by the system under study. Later in this dissertation, this quantity is used to evaluate pathway activity in gene expression data.

### 3.2.5 Parameter Selection

The final step in this procedure is to compute posterior distributions for each feature and choose \( \lambda^* \) for \( B \) based upon concordance to these posteriors. We first describe how to compute a posterior distribution and choose parameters for features where prior information
is unavailable across all sources. Then we describe how to do the same for features where prior information is available from at least one source.

**Choosing Parameters in the Absence of Priors** For features without any prior information, we use the data alone to compute an empirical distribution for the feature. We assume that the number of times the feature appeared across subsamples is binomially distributed $Bin(q, P_k)$. Using this, we compute the concordance ($\theta_{k}^{\lambda_i}$) of each $\lambda_i \in \Lambda$ for each feature $k$ without prior information (Equation 3.9). Here, $z_{k}^{\lambda_i}$ is the number of times feature $k$ was selected across all subsamples using parameter $\lambda_i$.

$$\theta_{k}^{\lambda_i} = \left( \frac{q}{z_{k}^{\lambda_i}} \right) \times (P_k)^{z_{k}^{\lambda_i}} \times (1 - P_k)^{q - z_{k}^{\lambda_i}}$$ \hspace{1cm} (3.9)

When the data has few samples, a desirable property of any model is stability or insensitivity to data variations. To incorporate stability into parameter selection, we compute stability for each feature across subsamples (Equation 3.10).

$$g_{k}^{\lambda_i} = 4 \times f_{k}^{\lambda_i} \times (1 - f_{k}^{\lambda_i}) \quad \text{where} \quad f_{k}^{\lambda_i} = \frac{z_{k}^{\lambda_i}}{q}$$ \hspace{1cm} (3.10)

$(0 \leq g_{k}^{\lambda_i} \leq 1)$ is a measure of instability of the feature $k$ for parameter $\lambda_i$. The ideal parameter then would have a large value of $\theta_{k}^{\lambda_i}$ and a small value for $g_{k}^{\lambda_i}$. This motivates the scoring function used for each potential parameter value (Equation 3.11). We choose the parameter value that maximizes this score for features without prior information.

$$score_{\lambda_i} = \sum_{k=1}^{np} \theta_{k}^{\lambda_i} \times (1 - g_{k}^{\lambda_i})$$ \hspace{1cm} (3.11)

**Choosing Parameters with Priors** Finally, we describe how to choose $\lambda$ for features in $wp$. Earlier, we represented the empirical distribution of the appearance of each feature as a binomial distribution. In order to make calculations tractable to compute a posterior, we approximate this binomial distribution with a normal distribution $N = (\mu_k, Var_k)$ (Equation
Figure 3.3: Merging of prior and empirical information for a feature. Panel (a) shows normal distributions for the amount of expected occurrences across subsamples for each prior and for their weighted mixture (black line). (b) The mixture distribution ($F^k_{KL}$) is approximated by a normal distribution. (c) This normal distribution is combined with the empirical distribution to give a posterior normal distribution.

3.12) \[89]\.

$$\mu_k = P_k \ast q$$
$$\text{Var}_k = P_k \ast (1 - P_k) \ast q$$

Figure 3.3 gives the main idea of the remainder of the approach. The first step is to represent each prior information source for each feature as a normal distribution and compute the weighted mixture of these normal distributions based on the reliability weights from the previous section (Equation 3.13). In this equation, $T_k$ represents the set of prior information sources that provided non-null information about feature $k$.

$$\Phi_k(x) = \sum_{i=1}^{\mid T_k \mid} w^t_{ir} \ast N \left( \phi^t_{kr}, \left( \tau^t_{kr} \right)^2 \right)$$

(3.13)

In order to synthesize this information to a normal distribution, we approximate the mixture using the normal distribution with the minimal KL divergence. This can be computed
analytically (Equation 3.14) [105]. Here, $\mu_k^{KL}$ and $\text{Var}_k^{KL}$ are the mixture mean and variance respectively.

$$\mu_k^{KL} = \sum_{i=1}^{|T_k|} w_{t_i} \phi_{t_i}^k$$

$$\text{Var}_k^{KL} = \sum_{i=1}^{|T_k|} w_{t_i} \left( (\tau_{t_i})^2 + (\mu_k^{KL} - \phi_{t_i}^k)^2 \right)$$

Now, the prior and likelihood (empirical) distributions are both normal distributions with previously specified parameters. Thus, a posterior distribution ($N = (\mu^*_k, \text{Var}^*_k)$) can now be computed analytically for each feature $k$ [11] (Equation 3.15).

$$\mu^*_k = \frac{\mu_k \text{Var}_k^{KL} + \mu_k^{KL} \text{Var}_k}{\text{Var}_k^{KL} + \text{Var}_k}$$

$$\text{Var}^*_k = \frac{\text{Var}_k^{KL} \text{Var}_k}{\text{Var}_k^{KL} + \text{Var}_k}$$

Now, a posterior distribution is set for every possible selected feature. The final step of the method is to choose $\lambda$ for those features where prior information is available. The score function for each value of $\lambda$ remains the same in this case (Equation 3.11). The two changes are that only features with prior information are considered, and $\theta^\lambda_k$ is computed according to Equation 3.16. Since a continuous probability distribution cannot be queried at a single point, we use a small range around the point to integrate the distribution. This gives an estimate of the probability that the feature appeared $z_k^\lambda$ times. Again, the chosen value of $\lambda_{WP}$ is the one that maximizes the score function.

$$\theta^\lambda_k = \int_{z_k^\lambda - \epsilon}^{z_k^\lambda + \epsilon} N(x \mid \mu^*_k, \text{Var}^*_k) \, dx$$

3.3 DISCUSSION

In this chapter, we have broadly discussed the integration of prior knowledge to improve modeling approaches for a variety of biomedical applications. We have presented an algo-
rithmic strategy to incorporate prior knowledge into model selection given a base selection procedure. This strategy addresses the fact that many prior knowledge sources may be incomplete and unreliable. The approach weights each prior knowledge source based on its concordance with the data. It then synthesizes their information to select informed parameters for the base selection procedure. The useful outputs of the strategy are threefold. First, a posterior distribution for each feature is given, which quantifies the uncertainty around whether the feature should be selected. Second, a weight for each prior knowledge source is computed along with a p-value that the source’s predictions are significantly more reflected in the data than a random source of equal size. Lastly, a hyperparameter is selected based upon model stability and concordance to the posterior distribution. This gives a principled way of selecting hyperparameters which could be used with or without prior information.

In the following chapters, I apply this general approach to two problems of interest in the overall exploratory pipeline: variable selection for graphical modeling (Chapter 4) and learning undirected graphical model structure (Chapter 5). In both cases, I show how to instantiate the approach with a particular base selection procedure, and I demonstrate how to adapt the methodology to address the problem of interest. Finally, I apply this methodology to an interesting biomedical application in the prediction of responders to a prophylactic cancer vaccine (Chapter 6).
4.0 VARIABLE SELECTION FOR GRAPHICAL MODELING

The integrated biomedical datasets that I aim to analyze with the proposed pipeline are too large to be modeled by graphical model structure learning algorithms. Thus, the first step of the pipeline is to pre-select variables to include in the graphical models. In this chapter, I investigate this problem. In the integrated biomedical datasets of interest in this thesis, continuous variables tend to outnumber the categorical variables by orders of magnitude. Further, the categorical variables tend to be important demographic and phenotypic information of interest. For this reason, the feature selection methodology I propose focuses upon selecting continuous variables.

First, I give some background about how feature selection has been applied to genomic and clinical data, and how prior knowledge has been proposed to improve these approaches. Then, I define a modified version of the feature selection problem, which I propose to be more appropriate in the context of graphical model structure learning. Next, I use the prior knowledge incorporation framework (Chapter 3) to both evaluate and incorporate external data. Lastly, I explore the performance of this approach on feature selection and prediction tasks for simulated and real biomedical data.

4.1 BACKGROUND

Graphical models are an effective way to understand the complex relationships in an integrated biomedical dataset, but to utilize them for this task requires addressing some of the difficulties inherent to biomedical data. High throughput sequencing assays (RNA-Seq, Microarray) often measure over 20,000 covariates (genes) in one experiment. Applying graphical
model learning algorithms to this data can be computationally intractable or costly, especially when integrating these assays with other data sources. Further, interpretation of such a large graphical model is near impossible unless a single variable of interest is being queried. Lastly, these datasets have an additional problem of high correlation among features. Graphical models have difficulty addressing high co-linearity as this can result in the formation of cliques disconnected from the rest of the graph [72].

One way to address these issues is by pre-selecting variables to model. The machine learning community refers to this problem as feature selection. There, the aim is to find the subset of features that best predict a target variable of interest. Though some of these approaches are applied to integrated biomedical data, they still fail to address the aforementioned challenges. High correlation among features results in unstable prediction models, and harms biological coherence.

In this section, I give background information on how the feature selection problem has been addressed with a focus on approaches tailored to genomic data. Then, I discuss how prior knowledge has been used to try to improve these approaches, and how graphical models can be used for feature selection. Finally, I present the specific problem statement to be addressed by this chapter.

### 4.1.1 Feature Selection in Genomics

Feature selection aims to identify a subset of features in a dataset that together best predict a target variable [42]. The main purpose of feature selection is to improve model training efficiency and to prevent overfitting the model to the data. In machine learning, feature selection approaches fall into three broad classes: filter methods, wrapper methods, and embedded methods [45]. Filter methods select features using univariate ranking scores such as a Wilcoxon test or a t-test between a co-variate and a target variable. Wrapper methods use a predictive model like the Support Vector Machine to select a set of features that gives high accuracy when used as input to the prediction model. Two popular wrapper methods are recursive feature elimination and greedy forward search, which select the best feature to eliminate (or include, respectively) in a step-wise fashion. Finally, embedded methods are
predictive models which select features automatically as part of their optimization procedure. The most popular example of this is the LASSO regression method, which uses an $L_1$ norm penalty to shrink many feature coefficients to zero in a linear regression setup.

Recently, a study was performed investigating the performance of these techniques on genomic datasets to predict breast cancer relapse [45]. Overall, none of these methods had significantly better accuracy than randomly choosing features and only filter based methods were more stable (insensitive to variations in the data). Since this was true even within a single dataset, it could only be attributed to the statistical issues inherent to genomic data. This suggests that tailored approaches are necessary to improve feature selection from this type of data.

4.1.2 Incorporating Prior Knowledge

One way to improve these approaches is to use domain knowledge about the relationships between genes and between their protein products. Three main sources of prior knowledge have been explored: gene ontology (GO) terms, protein-protein interaction networks (PPI’s), and biological pathways [50].

GO is an attempt to group genes based on known biological functions (e.g. cell cycle or angiogenesis). Several approaches have leveraged GO terms as prior information to construct gene clusters [88, 19, 17]. The main drawback of these methods is the nature of GO terms themselves. Not all genes belong to a functional group in the GO, and these methods chose to discard those genes. In addition, GO terms tend to define very broad functional classes which may have limited relevance to high throughput sequencing data.

PPI’s are networks that encode protein interactions known to occur in normal cellular activity. Many methods for gene selection have been built off of these networks, and they were reviewed and evaluated in [30]. Many of these approaches aim to either 1) group genes based on the edges in the network and penalize them together [4, 140, 116, 155] or 2) use the network information to determine gene importance [131, 59].

Pathway based approaches are similar in principle, but use biological pathways which represent a module of the network that carries out a specific function. These are normally
taken from a pathway database such as KEGG or I2D [60, 12]. Biological pathway-based feature selection (BPFS) is one such method. BPFS is a step-wise method that uses mutual information to the target variable as a scoring criterion. To avoid redundancy, the method does not choose multiple genes from the same area of a constructed graph consisting of the union of all the pathways [7]. In [41], the authors attempt to construct a single feature for each pathway by aggregating information across multiple genes. A similar method is taken in [3] except that the pathways are constructed using the data.

Multiple studies have examined these methods compared to methods that do not use prior knowledge and have found no significant benefit in prediction accuracy [30, 125]. However, these methods do appear to give more biologically interpretable signatures.

4.1.3 Graphical Models for Feature Selection

Though this chapter is focused upon using feature selection methods to choose the input to a graphical model, it is important to note that graphical models themselves can be used for feature selection. The typical procedure for this is as follows. First, the graphical model structure is learned using the full dataset. Then the neighbors or the Markov Blanket of a target variable is extracted, and these features are used in a downstream prediction model. I demonstrate applications of this procedure with the graphical modeling methods developed here in Chapters 6 and 8.

Alternatively, the Markov Blanket of the target variable could be directly constructed. A popular method for this is the Max-Min Parents and Children (MMPC) algorithm [135], which uses conditional independence tests to identify the parents and children of a target variable in the graphical structure without learning the full graph structure. More recently, other methods have been proposed for this which attempt to deal with the high correlations of genomic data by finding clusters of variables in the parents and children set [56].

However, if the full graph structure is desired, then the methods to learn this structure are too computationally costly to run on large omic and clinical datasets. Thus, feature selection must be used prior to learning graphical model structure. Unlike for some prediction tasks, an important consideration in selecting these features is redundancy. Highly redundant (or
correlated) features must not be selected together to prevent isolated sub-networks from appearing in the final graphical model. In addition, one of the redundant features must not be chosen at random, because this limits biological interpretability.

4.1.4 Problem Statement

To this end, we propose to address the following modified feature selection problem. This was first proposed in the database community to identify query results relevant to the user but also diverse to give a broad snapshot of the underlying data [39]. The problem was referred to as the Top-K relevant and diverse set problem, and is as follows.

Definition 4.1. Top-K Relevant and Diverse Set. Given \(0 \leq r \leq 1\) a radius of similarity, a set of variables \(V\), an output size \(k\), a similarity function \(Sim(V_i, V_j)\), and a relevance function \(Rel(V_i)\).

\[
\begin{align*}
\text{maximize} & \quad \sum_{X_i \in S} Rel(X_i) \\
\text{subject to} & \quad S \subset V \\
& \quad |S| = k \\
& \quad \forall i, j \ V_i \in S \text{ and } V_j \in S \rightarrow Sim(V_i, V_j) < r
\end{align*}
\]

Intuitively, this problem aims to find a set of variables that are relevant to the user with the constraint that no pair of chosen variables are similar to one another. The method we propose to solve this problem is similar in principle to two filter methods: Correlation-based feature selection [44] and maximum relevance minimum redundancy (mRMR) feature selection [33]. Both of these are greedy approaches which select the feature that optimizes an objective function that balances relevance and diversity. The main difference in our approach is that we require no redundancy, and that we quantify redundancy using prior knowledge. Lastly, to ensure stability of the downstream model, we report the selected features as clusters with redundant variables included in each cluster (instead of discarding them). This allows the user to understand the correlations present in the data.

Another popular approach that follows this principle is the Weighted Gene Correlation Network Analysis (WGCNA) [154]. Briefly, this method aims to learn a weighted undirected
correlation network by converting correlation to edge weight. With this network, the dissimilarity between nodes in the network is inferred, and network characteristics (e.g. hub nodes) are used to select important genes. This method differs in that it uses soft thresholding to define correlated variables. In addition, the inferred network is a correlation network, whereas the graphical models here aim to identify conditional dependence relationships.

4.2 METHODS

In this section, I describe an algorithm to solve the problem, built upon the prior information method from Chapter 3. I instantiate this problem with a specific relevance and similarity score, and I discuss how to set appropriate parameters for the problem.

4.2.1 Preferential Diversity (Pref-Div)

To solve the Top-K Relevant and Diverse feature problem, the original authors suggest the Preferential Diversity (Pref-Div) algorithm (Algorithm 4.2). Here, $\text{sim}_r(X,Y)$ is true if $X$ and $Y$ are within the radius of similarity $r$ of one another. $I(G)$ is the relevance value of variable $G$. This is an iterative procedure that in each iteration: 1) attempts to include the most relevant $N$ features remaining, 2) excludes those that are too similar to any variable in the current result set, and 3) excludes those that are too similar to another variable recently included in the result set. The algorithm is tuned by the "Accuracy" parameter $A$ which controls how much the user values relevance ($A = 1$) versus diversity ($A = 0$). When $A = 1$, the top-K variables are returned regardless of their redundancy, whereas when $A = 0$ the search continues to the most irrelevant variables until a set of variables is found such that no two variables are pairwise related.

For the purpose of identifying variables to include in a graphical model, we instantiate the Pref-Div algorithm with the following choices. First, $A$ is set to 0 to avoid including any pairwise redundant variables. The absolute value of the Pearson correlation between variables is used as a similarity metric, since both correlated and anti-correlated variables
Algorithm 4.2 Preferential Diversity algorithm

1: procedure PREFERENCE DIVISION (Set of variables $S$, size $k$, relevance parameter $A$, radius $r$)
2:   $T \leftarrow \emptyset$
3:   $turnCounter = 0$
4:   while there exists unmarked items in $S$ and $|R| < k$ do
5:     Increase $turnCounter$ by 1
6:     $T \leftarrow$ Pick $k$ items with highest intensity from $S$
7:     for variable $G_i \in R$ do
8:       for variable $G_j \in T$, s.t. $\text{sim}_r(G_i, G_j)$ do
9:         Mark $G_j$ as "Eliminated"
10:    while there exists unmarked items in $T$ do
11:       $R = R \cup G_i$, s.t. $G_i \in T$ is unmarked and $I(G_i) \geq I(G_j)$ : $\forall G_j \in T$
12:       for unmarked $G_u \in T$ do
13:         if $\text{sim}_r(G_i, G_u)$ then
14:           mark $G_u$ as "Eliminated"
15:    while number of unmarked items in $T < A \cdot k$ do
16:       $R = R \cup G_i$, s.t. $G_i \in S$ is unmarked and $I(G_i) \geq I(G_j)$ : $\forall G_j \in T$
17:       $A = A \cdot 0.5$
18:     if $turnCounter == 1$ then
19:       create new set $N \leftarrow \forall G_j \in T$, s.t. $G_j$ is marked
20:   $S = S - (S \cap T)$
21: if $|R| < k$ and $\forall G_j \in S$, s.t. $G_j$ are marked then
22:   while $|R| < k$ do
23:     $R = R \cup G_j$, s.t. $G_j \in N$ and $I(G_j) \geq I(G_i)$ : $\forall G_i \in N$
24:   Return $R$

are redundant in a linear model. Relevance is defined as the absolute value of the correlation between the variable and the target(s). The number of variables to select ($k$) is user-defined, and in the experiments here, cross validation is used to select this value. Finally, we make one
modification to the Pref-Div algorithm. Instead of discarding redundant variables, we include them as a cluster with the most relevant variable of the cluster being the representative. In this way, aggregation methods (i.e. PCA) can be used to combine the information of redundant variables, and no biological information is lost.

4.2.2 Prior Information Pref-Div (piPref-Div)

The prior knowledge framework (Chapter 3) is used to evaluate and utilize prior information sources to choose the hyperparameter: the radius of similarity \( r \) for Pref-Div. For the base selection procedure \( B \), a correlation graph \( G \) is constructed based on the value of \( r \). An undirected edge \( E = (V_i, V_j) \in G \leftrightarrow 1 - | \text{Cor}(V_i, V_j) | < r \). So, \( r \) acts as an absolute threshold to determine edge existence in the correlation graph. The variables to be selected by this method are the edges in the correlation graph. The reason that Pref-Div itself is not used as the base selection procedure is that this focuses on only the variables related to the target, but all relationships can be used to select a sensible hyperparameter value (since prior information is available for non-target relationships).

The output of the framework are two values for the hyperparameter of interest \( r_{WP} \) and \( r_{NP} \), which are radii of similarity for relationships where prior information exists and for those where it does not. Using these thresholds, we use the Pref-Div algorithm as usual except that determining whether \( V_1 \) and \( V_2 \) are redundant is based upon \( r_{WP} \) if prior information is available for the relationship between \( V_1 \) and \( V_2 \) and \( r_{NP} \) otherwise. Since correlation is a relatively stable measure, we slightly modify the variables included in the \( WP \) group. We exclude relationships where the weighted probability of occurrence according to the prior information sources is less than a cutoff. For all of our experiments, we set this cutoff value to 0.5. The impact of this cutoff value is tested in the simulated results (Figure 4.6).

4.2.3 Simulated Datasets

Simulated datasets were used to ensure algorithmic correctness and to understand the impact of prior information sources. For the simulated experiments in this chapter, data was generated from a linear Gaussian graphical model. Edge coefficients were drawn uniformly
Clustered Simulation
In the clustered simulation (Figure 4.1, left), each variable belonged to one of $C$ clusters. In these clusters, each pair of variables in the cluster were connected by an edge. $c < C$ clusters had one randomly chosen variable connected to the target variable (relevant clusters), while the remaining $C - c$ clusters were disconnected from the rest of the network. Each cluster consisted of an equal number of variables. To represent a master regulator and force correlated structure, each cluster had a single latent (unmeasured) variable that influenced the value of all variables in the cluster.

Pathway Simulation
The main difference between this pathway simulation and the clustered simulation is that variables may belong to multiple clusters and connections can exist between clusters. The specific procedure to generate the graph structure is as follows. First, the number of clusters each gene belongs to is pulled from an exponential distribution with mean 1, constrained so that each variable belongs to at least one cluster and no variable belongs to more than $C$ clusters. The clusters that each variable belong to are pulled at random from the set $[-1.5, -0.5] \cup [0.5, 1.5]$. Error terms for each variable were zero mean with variance randomly drawn from the set $[0.01, 2]$. Graphical structure was simulated using two types of simulations: clustered simulation and pathway simulation.
uniformly at random. Next, edges are placed between the variables in each cluster uniformly at random. Given that the cluster has $V_i$ variables, the number of edges for cluster $i$ is drawn from the set $\left[0, \frac{V_i \times (V_i - 1)}{2}\right]$. Lastly, for $c$ clusters, one variable is randomly chosen to be connected to the target variable (relevant clusters). The other $C - c$ clusters are irrelevant to the target of interest. Like before, a latent variable for each cluster is used as a master regulator.

Prior knowledge was simulated for two types of prior information sources: reliable priors and unreliable priors. All prior sources give information based on a beta distribution; however, the parameters of this distribution differ based on the type of prior and whether the variables in question belong to the same cluster. An unreliable prior gives information drawn from $Beta(4, 4)$ for both true and false relationships, whereas a reliable prior draws from $Beta(10, 2)$ for true relationships, and $Beta(2, 10)$ for false relationships. The amount of prior information varies based on the experiment. To determine whether prior information is available for each relationship, each relationship gets a value $b \sim U(0, 1)$, and each prior information source has a value $c \in [0, 1]$. The prior gives information about the relationship if $b < c$. In this way, the simulated data reflects the fact that some relationships are more well-studied than others.

4.2.4 Real Data and Curated Prior Information

To evaluate the performance of piPref-Div, we apply it to two real applications. The first is six publicly available breast cancer microarray datasets [107, 57, 142, 93, 120, 32]. These datasets have been used in several previous analyses and represent a baseline to evaluate prediction methods [30, 125, 45]. Each dataset consists of microarray expression data for between 159 and 286 patients, and the target variable of interest in this data was whether or not the patient had relapse free survival for 5 years.

The second dataset is RNA-Seq data from the Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) project. This data included gene expression measurements from 784 breast tumor samples and 13,994 genes. Breast cancer diagnosis and prognosis are commonly divided into five main subtypes: Luminal A, Luminal B, HER2+, Triple-Negative,
and Basal. Breast cancer sub-type information for each tumor sample was obtained from [58], which did not distinguish between Triple-Negative and Basal, so we focused on differentiating these four types. The main driving distinction for these subtypes is the presence or absence of hormone receptors on the tumor cell surface, which can lead to varying prognoses. In these experiments, we aim to identify clusters distinguishing the four sub-types from expression data. To determine stability of each of these clusters, a 10-fold cross validation was performed, and the stability of each cluster was the number of times a similar cluster (similarity $> 0.85$) was selected in each fold.

Prior knowledge consisted of five distinct sources of information. Physical gene distance quantified the base pair distance between two genes on the chromosome. If two genes were on separate chromosomes, then this value was set to zero. Otherwise given gene $G_i$ from base pairs $B_1^i$ to $B_2^i$ and gene $G_j$ from base pairs $B_1^j$ to $B_2^j$, and full chromosome length $C$, the physical distance prior is given by Equation 4.2. This represented the proportion of chromosome distance covered by the space between these two genes.

$$
\text{Phys} (G_i, G_j) = 1 - \frac{\text{max} (B_2^i, B_2^j) - \text{min} (B_1^i, B_1^j)}{C}
$$

(4.2)

Gene family information was curated from the Human Genome Organization (HUGO). Gene families are groups of genes related by sequence and/or function. A single gene can belong to multiple gene families. Thus, we represent each gene as a vector of families with one-hot encoding. To compute the similarity between these vectors, we use the Jaccard similarity metric which is the number of families in common divided by the total number of unique families either gene belongs to.

A similar approach is used for gene-disease mapping from the DisGeNet [96]. This database gives scores quantifying the level of knowledge that a change in a gene is related to a disease. We use the guilt by association principle to compute whether two genes are related. We represent a gene by a vector of scores to the diseases in the database, and we compute the cosine similarity between two gene vectors. Since all scores are positive, this metric is between 0 and 1, and can be used directly as a probability.

Finally, we use gene-gene similarity data from two sources: Harmonizome [104] and STRING [129]. Harmonizome similarity data was curated from the Molecular Signatures
Database [128]. This consisted of correlation between gene expression across several microarray experiments. STRINGdb curates gene-gene relationship scores based on several factors such as: co-expression, literature co-occurrence, experimental evidence, other databases, etc. STRINGdb scores were scaled from their (0,1000) range to (0,1).

### 4.2.5 Evaluation Metrics

The metric we use for evaluation of selected clusters on simulated data is called cluster accuracy. This metric compares the relevant clusters output by piPref-Div to the true relevant clusters in the data generating graph. A relevant cluster is a cluster with at least one parent of the target variable. To compute cluster accuracy, first, an optimal matching between the predicted and actual clusters is found using the Hungarian Algorithm. The cost of assigning a predicted cluster to an actual cluster is given by $1 -$ the Jaccard similarity between the clusters. If multiple predicted clusters are best assigned to the same actual cluster, then these clusters are combined. Finally, the average Jaccard similarity between the combined predicted clusters and their matched actual clusters are computed as the score.

For real datasets, the predictive models were judged using two evaluation metrics: accuracy and stability. For accuracy, area under the ROC curve was used to compare predicted probability of RFS for five years vs. actual outcome. For stability, the metric used is a cluster similarity score originally reported in [56]. The overall score is a pairwise similarity between the clusters selected in each fold. To compute the similarity between two sets of clusters, first an optimal matching is found between the clusters using the Hungarian Algorithm (with the score function being the Tanimoto Set Similarity). Next, the average Tanimoto similarity between the optimal matching is used as the final score. For single variable selection, this is equivalent to the Tanimoto set similarity between the feature sets.
4.3 RESULTS

In this section, I present an evaluation of piPref-Div on simulated and real datasets. I first explore the performance of piPref-Div in evaluating prior information sources and identifying relevant clusters on simulated datasets of varying size. I then test the sensitivity of piPref-Div to ad-hoc choices that must be made to run the algorithm. Finally, we evaluate the performance of piPref-Div on the breast cancer microarray datasets in predicting relapse free survival.

4.3.1 Prior Knowledge Evaluation Results

First, we tested the ability of piPref-Div to accurately evaluate prior knowledge sources on 15 simulated datasets of 500 variables with 50 clusters and 25 relevant clusters, 200 and 50 samples, and 5 prior knowledge sources (3 reliable) with a random amount of prior information. The results are presented in Figure 4.2. Here the ”Net Reliability” on the y-axis refers to the sum of the probabilities given to true relationships minus the sum of the probabilities given to false relationships for each prior. The predicted weight for each prior knowledge source given by piPref-Div shows a clear association to the reliability score. A benefit of this approach is that this weight does not appear to be dependent on the amount of prior information. Even with very little prior information (blue circles), piPref-Div assigns a relatively accurate weight to the knowledge sources.

4.3.2 Effect of Amount and Quality of Prior Knowledge

The next experiment investigated the impact of the amount and quality of prior knowledge sources on the ability for piPref-Div to identify relevant clusters of variables in the data. In these experiments, we test the method using the same experimental parameters as the prior knowledge evaluation section, and using a larger dataset with 3000 variables, 300 clusters, and 75 relevant clusters. For each experimental setting, 15 graphs/datasets were generated and the results are presented cumulatively over these simulations.

The results for the cluster simulation on the small datasets are given by Figure 4.3.
Sample size is the most significant factor in increasing the accuracy of the selected clusters. Prior information can result in a modest improvement in selected clusters, but this benefit is only seen when there is at least 50% prior information and at least 3 reliable sources out of 5. However, when all sources are unreliable, there is no decrease in accuracy unless there is a large amount of information present. The benefit of prior information is drastically reduced in cases with sufficient sample size (200 sample case). This is intuitive, as with more data, correlation becomes a very stable measure, and prior information can be ignored.

This experiment was repeated for the pathway simulation and the results are presented in Figure 4.4. These results are similar to the cluster simulation, except that all amounts of
Figure 4.3: Accuracy of predicted clusters for varying amount and reliability of prior knowledge. Graphs were simulated according to the cluster simulation method. Sample size was set to 50 (left) and 200 (right).

prior information perform substantially worse. This is to be expected because the pathway simulation prevents large intercorrelations among the variables in a single cluster. The influence of other clusters mitigates these correlations, making it difficult to detect independent clusters for the base selection procedure regardless of prior knowledge.

Lastly, the ability of piPref-Div to detect clusters from a larger graph (Figure 4.5) was examined. Here, the pattern is largely similar to the small simulations, except the impact of prior knowledge is increased. In particular, even 25% prior information results in a substantial increase in accuracy over having no prior information. This impact is again larger when the sample size is smaller, though it is present in both cases. Further, the impact of more samples is more pronounced in the larger dataset. An increase from 50 to 200 samples results in an increase in accuracy from 0.65 to over 0.8 for all prior information amounts.
Figure 4.4: Accuracy of predicted clusters for varying amount and reliability of prior knowledge. Graphs were simulated according to the pathway simulation method. Sample size was set to 50 (left) and 200 (right).

4.3.3 Sensitivity Analysis

Next, the sensitivity of the algorithms were examined. The choice of the number of features to select and the choice of the cutoff to include a relationship in the "with prior" WP group was explored. Figure 4.7 shows the efficiency with which piPref-Div selects unique clusters (left) and clusters related to the target (right). These plots demonstrate that the amount and quality of prior information does not have a significant impact on whether the algorithm selects unique clusters. This is expected because the radius of similarity has a minimal impact on which features are selected (since it does not change the relevance scores of each variable). Sample size appears to have a significant impact on how efficiently the algorithm identifies clusters related to the target. When the algorithm selects 20 features from a 200 sample dataset, it selects a feature from around 55% of true clusters and 35% of all clusters. With 50 samples, only 40% of true clusters are represented, suggesting that the algorithm selects variables from irrelevant clusters (the slope of the line on the left does not flatten).
Figure 4.5: Accuracy of predicted clusters for varying amount and reliability of prior knowledge on large datasets. Graphs were simulated according to the cluster simulation method. Sample size was set to 50 (left) and 200 (right).

Figure 4.6: Impact of the WP Cutoff parameter on the ability of piPref-Div to select relevant clusters.
Figure 4.7: Sensitivity of piPref-Div to the number of features selected. Graphs were simulated according to the cluster simulation method with 50 clusters and 25 relevant (true) clusters. Each point is averaged over 15 simulations. Left graph shows the proportion of all clusters selected as a function of number of selected features. Right graph shows true clusters only.

Figure 4.6 shows the effect of changing the cutoff to include a relationship in the WP (with prior) group. For a wide range of cutoff values, there is no significant impact on accuracy. The impact is most largely felt at the extreme choices of the cutoff value, with low sample size, and with lesser and unreliable prior information. It is important to note; however, that this is dependent upon the simulation parameters. In this simulation, the reliable priors generated edge probabilities by altering their beta distribution parameters, and 0.5 is a good discrimination line between the true and false edges. In real applications, this parameter could have a more significant impact. The most conservative choice would be to have no cutoff and include all edges with prior information in the WP set. The performance drop from this appears to be relatively small in our simulated experiments.
4.3.4 Breast Cancer Outcome Prediction

To determine the performance of piPref-Div on real datasets, we applied the algorithm to the aforementioned breast cancer microarray datasets. The task was to build a prediction model to predict five year relapse free survival (RFS) of cancer patients based on expression data. To determine whether these models would generalize to unseen data, a five-fold cross validation approach was used for each dataset. In particular, each dataset was split into five independent folds, and in each iteration one fold was treated as a testing set while the other four folds were treated as a training set. For each training set, features were selected using either a Pref-Div algorithm alone or a combination of Pref-Div and an undirected graphical structure learning algorithm [112]. For the undirected graphs, the variables directly connected to RFS were treated as the selected features. For all feature sets, the selected features were used as input to a logistic regression model to predict RFS.

The learned models were evaluated for accuracy and stability. Here, accuracy was computed using area under the ROC curve (AUC) comparing the predicted probability of 5 year RFS versus actual RFS. Stability was defined as the average Jaccard similarity between the features selected in each fold.

Six variations of piPref-Div were tested. piPref-Div alone with prior information (PD), piPref-Div with a graphical modeling algorithm (PD-Graph), piPref-Div with and without prior information (NP = No Prior) with clusters aggregated into summarized features using principal component analysis (PD-PCA, PDNP-PCA), and each of the PCA approaches with graphical modeling (PD-PCA-Graph, PDNP-PCA-Graph). For the Pref-Div approaches without graphical modeling, an inner 3-fold cross-validation loop was used to determine the number of selected features (1, 3, 5, and 10 features were tested). For the graphical modeling approaches 100 features were selected by Pref-Div to be input into the graphical model. Genes with less than 0.5 standard deviation across samples in the training set were removed from the dataset prior to feature selection. For comparison purposes, two methods that performed well in a previous study were included in the analysis: Hybrid-Huberized SVM (HH-SVM) and Recursive-Reweighted Feature Elimination (RRFE) [30].

Figure 4.8 presents the accuracy results across the six independent breast cancer datasets.
Figure 4.8: AUC of predicting RFS using several feature selection methods on six independent breast cancer microarray datasets.
Across the datasets, the consistent best performing methods are PD-NP-PCA and PD-PCA (sea-green and light blue, respectively). This suggests that cluster selection and representing individual features as clusters offers a substantial benefit to selecting single genes alone. However, this is dataset dependent, as Pref-Div alone (PD, yellow box) matches these methods on 2 of the datasets (Sotiriou and Desmedt) and performs better on 1 dataset (Wang). Overall, these results show no significant difference between using prior information (PD, PD-Graph, PD-PCA, and PD-PCA-Graph) and not using prior information. Using prior information with PCA clustering shows a slight improvement on the Ivshina dataset, but none of the others. In this experiment, including graphical modeling often performed worse than selecting features directly, which is to be expected since this is an unsupervised technique. In addition, in some experiments the graphical model suggested that no features were related to RFS, which is why some of the AUC values are close to 0.5 for all runs. It could be the case that though models were predictive in cross-validation, they may just be spurious correlations that will not generalize to unseen data.

Figure 4.9 presents the stability of the learned models across all six datasets. The results confirm previous works that identifying a stable model for breast cancer outcome prediction is a difficult problem [30]. In general, across all six datasets only the RRFE algorithm shows somewhat consistent stability; however, a contributing factor is that this algorithm uses on average 119 selected features, whereas HH-SVM averages around 6 and the PD approaches average around 1 feature (or cluster). Among, the Pref-Div based approaches, PD-PCA with and without prior information show the most consistent stability. On nearly all datasets they are near to, or on par with RRFE despite choosing significantly fewer features. One important consideration is that the graphical modeling based approaches often did not identify any variable related to the target. In these instances, stability was computed as 0, whereas in practice this is an indication that no reliable model can be identified.

### 4.3.5 Stratification of Breast Cancer Sub-types

Due to its consistent performance, PD-PCA was evaluated based on its ability to mine interesting and relevant clusters to breast cancer sub-type discrimination. PD-PCA selected
Figure 4.9: Stability of learned models for predicting RFS using several feature selection methods on six independent breast cancer microarray datasets.
features from the data which were passed to a graphical modeling algorithm [112]. An MGM model was learned on a dataset consisting of only the selected clusters and the Subtype variable. To summarize clusters into single names, the Ingenuity Pathway Analysis (IPA) regulator analysis was used, and the KEGG Pathway database was queried (corrected p-values < 0.05 were chosen as candidates). Following this step, specific pathways and regulators were chosen as the names of the clusters.

The learned graphical model is presented in Figure 4.10. Two clusters were unable to be mapped coherently to any biological function (single gene representatives were TMEM41A, and TSPAN15); however, these clusters were relatively unstable. The two most stable clusters were: Fanconi Anemia/ Hereditary Breast Cancer pathway, and a set of genes regulated by MYCN. Fanconi Anemia and the Hereditary Breast Cancer pathways are known to share common genes [2] and hereditary breast cancer tends to be associated with ER+ breast cancer [83]. MYC family pathways and the transcription factors themselves are known to be differentially expressed across sub-types, and the MYCN factor in particular has shown differences between triple-negative and other sub-types [53]. FOXA1 along with GATA3 and ESR1 are necessary for maintaining a luminal phenotype of breast cancer [15], and AGR2 is upregulated by FOXA1 but only in an estrogen receptor dependent manner [146]. This implies that the FOXA1-AGR2 loop will only be upregulated in ER+ breast cancer. Though it is unclear how KRT14 regulated genes distinguish sub-types of breast cancer, it is known that upregulation of KRT14 reduces the ability of breast tumor to metastasize and invade the extracellular matrix [144]. Overall, PD-PCA-Graph constructs and selects reasonable candidate clusters for scientists to experimentally probe.

4.4 DISCUSSION

In this chapter, I discussed the problem of selecting variables to include in the graphical model. The proposed method was an instantiation of the prior knowledge incorporation method from the previous chapter. The method aimed to identify a set of variables that
Figure 4.10: Graphical model of breast cancer sub-type. Size of each edge represents the number of times a similar cluster was selected to be related to Subtype in each of the cross-validation folds.

balanced relevance to target variable(s) of interest with limited redundancy between variables in the data. In addition, the method identified variables related to the selected variables and included these clusters in the model to prevent information loss.

On simulated data the method demonstrated resilience to unreliable prior information and limited improvement when accurate prior information was used. The improvement was maximal when the dataset had few samples and when the dataset had a large number of variables. This may be due to the fact that Pearson correlation is a relatively stable measure (insensitive to small variations in the data). It could be that with a more unstable measure such as partial correlation, that prior knowledge will have a greater impact. In addition, the results showed that the method was insensitive to changes in the hyperparameters except at extreme values.
On real biological data, the variable selection method gives models that are as accurate or better than state of the art gene selection methods. Using PCA to summarize related genes into clusters appears to produce more accurate and stable models. Incorporating prior information into these variable selection methods does not appear to improve predictive accuracy. In discriminating breast cancer subtype, we find that using our variable selection strategy with graphical modeling produces a useful gene signature that is supported by recent studies. In addition, the model gives novel candidates for future studies.

Limitations of the approach include the fact that variable relationships are considered regardless of directionality. This could present issues when interpreting a cluster as a coherent module, since some of the features may be positively related to the target(s) and some features may be negatively related. Using a direction-aware aggregation scheme may alleviate this issue [4]. In addition, the improvement from using prior knowledge was limited. This is somewhat expected since the prior knowledge was used only to choose hyperparameters for the model. A potential workflow for this method would be to first evaluate the prior knowledge sources and determine their reliability. As a next step, reliable sources could be used as absolute or soft constraints on the feature selection method.

Another alternative to this approach would be to change the base feature selection procedure. An algorithm like the graphical LASSO [38] method could be used to learn the partial correlation network (instead of the correlation network). This could then be used to select features related to the target or network connectivity features could be used to select important genes for downstream graphical modeling. The downside of this approach is the limited scalability of graphical LASSO compared to computing a correlation matrix and that graphical LASSO is affected by high correlations in the data.

In the next chapter, I examine a second instantiation of the prior knowledge incorporation scheme: learning the structure of undirected graphical models for mixed continuous and categorical datasets.
5.0 UNDIRECTED GRAPHICAL MODELS FOR REPRESENTING JOINT DISTRIBUTIONS OF MIXED DATA

The previous chapter discussed methods for feature selection using prior knowledge. The method I propose for this problem is an instantiation of the prior knowledge evaluation and incorporation framework (Chapter 3). In the context of the pipeline of this thesis, the variable selection method is used to determine which features should be included in the downstream graphical models. The remainder of this thesis discusses how to learn these graphical model structures from data.

In this chapter, I focus upon learning undirected graphical model structure from mixed continuous and categorical data. I begin by giving some background on graphical models for mixed datasets and the necessary hyperparameters for model selection methods. I then present a method for utilizing prior information to learn undirected structure from mixed data. Finally, I evaluate this approach on simulated, mixed datasets as well as on a real application to breast cancer transcriptomic and clinical data from the Cancer Genome Atlas (TCGA).

5.1 BACKGROUND

Graphical models represent the joint distribution of a set of variables as a graph where nodes correspond to variables and edges correspond to conditional dependence relationships among connected variables. These models have natural interpretability in biological applications where pathway graphs are a popular way of structuring human knowledge [60, 35].

There are two broad classes of graphical models: directed graphical models and undi-
rected graphical models. In both cases the joint probability distribution factors according to the graph, meaning that the joint probability distribution can be represented as a product of conditional distributions based on the graphical structure. In the directed case, each factor in this product is the conditional distribution of each variable given its directed parents in the graph. The most popular of these is the Bayesian Network [65]. On the other hand, several undirected models exist which can be factored in different ways.

Undirected graphical models give the user a ”snapshot” picture of the relationships present in their data. This is beneficial to determine parsimonious prediction models for a target variable of interest, or to perform statistical inference on a query concerning multiple variables (e.g. How likely is it for A, B, and C to have high values?). For the purposes of this thesis, we focus upon a specific undirected model, the Markov Random Field

5.1.1 Markov Random Field

A Markov Random Field (MRF) is defined by a Graph $G = (V, E)$ and a joint probability distribution over the variables in the graph $P$. Together these must satisfy three properties (Definitions 5.1 to 5.3). These three Markov properties are equivalent for any strictly positive joint density [65].

**Definition 5.1.** The Pairwise Markov Property states that $\forall X_i, X_j \in V, \ Adj (X_i, X_j, G) \rightarrow X_i \Perp\Perp X_j | V/X_i, X_j$

**Definition 5.2.** The Local Markov Property states that $\forall X_i, X_j \in V$ and $X_j \notin Adj (X_i, V) X_i \Perp\Perp X_j | Adj (X_i, V)$

**Definition 5.3.** The Global Markov Property states that $\forall X_i, X_j, X_S \in V X_i \Perp\Perp X_j | X_S$ if every path from $X_i$ to $X_j$ crosses a node in $X_S$

To avoid parameterization of the full joint distribution, MRF’s can be parametrized using local sub-graphs of the full graph. One example of this is clique factorization of an MRF (Equation 5.1), where $C (G)$ represents the maximal cliques in the graph $G$, $Y$ is a possible value assignment for the variables in the graph, and $\phi$ is the positive potential function. This function defines how ”preferable” a particular configuration of the variables in the clique $C$ is. Since the values of this potential function are unconstrained positive
values, $Z$ is a normalizing constant to ensure that the probability distribution sums to 1 (or integrates to 1 for continuous distributions). One difficulty in estimating a MRF from data is the computation of the partition function, since this involves a combinatorially large summation (or integration).

$$p(X_1, \ldots, X_N = X) = \prod_{c \in C(G)} \frac{1}{Z} \phi(X_c)$$

$$Z = \sum_Y \prod_{c \in C(G)} \phi(X_c = Y)$$

(5.1)

Another way to parameterize these models is the pairwise Markov Random Field (Equation 5.2). Here, a potential function is defined for each edge in $G$.

$$p(X_1, \ldots, X_N = X) = \prod_{\text{Edge } e = (X_i, X_j) \in G} \frac{1}{Z} \phi(X_i, X_j)$$

(5.2)

These models can also be used with continuous data if a parametrization that satisfies the Markov properties can be formed. One continuous undirected model is the Gaussian Graphical Model (GGM) [28]. This model assumes that the joint distribution of the variables is a multivariate Gaussian distribution. A graph that satisfies the Markov conditions for this joint distribution can be constructed by placing edges between variables with nonzero entries in the inverse covariance matrix ($\Sigma^{-1}$) [38].

### 5.1.2 Mixed Graphical Models

Here, we focus upon models of mixed categorical and continuous data types because these models could help solve an outstanding problem in analyzing biomedical datasets: data integration. Currently, many experimental and observational datasets have information from various sources: genetics, epigenetics, demographics, phenotypes, and transcriptomics. Developing a single modelling approach capable of handling these data requires modelling both types of variables among others (ordinal, censored data, etc.).

Identifying a joint distribution that satisfies the Markov properties is more difficult in the case of data with continuous and categorical variables. The conditional Gaussian model
[68] parameterizes the joint distribution with a separate multivariate Gaussian for each configuration of the categorical variables. This model has the benefit of flexibility in capturing changing relationships in different contexts. However, in practice, the model is intractable to estimate because the number of parameters increases exponentially with the number of categorical variables in the data.

More recently, some computationally feasible approaches have been proposed. In [136], the authors propose qp-graphs which can be estimated from high dimensional data. The idea is to use a linear measure of association called limited-order correlations to remove edges from the model. These limited-order correlations can be estimated from small sample size data. However, this type of model assumes that there are no edges between categorical variables which is a limiting assumption for clinical data.

Several authors have proposed using regression based methods to estimate the conditional dependencies among pairs of variables to infer the edges in the graph. Fellinghauer et al. use a random forest regression approach to rank edges for inclusion into a graphical model among mixed variables [36]. Yang et al. assume that the conditional distributions of each type of variable come from the exponential family and use node-wise regression approaches to estimate the parameters of the model [150]. Other authors have proposed similar techniques to the aforementioned methods [18, 16, 38].

Another way to estimate a mixed graphical model is the pseudolikelihood approach [8]. This approach uses the product of conditional distributions of the variables as a consistent estimator of the true likelihood without computing the partition function. Then a gradient based optimization method can be used to find maximum pseudolikelihood estimates of the parameters. Lee and Hastie propose a mixed graphical model that naturally generalizes the GGM and the MRF to data with mixed continuous and categorical variables [70]. They demonstrate that using the pseudolikelihood approach shows better empirical performance than using separate regressions. We give more details about this method in the Methods section.
5.1.3 Regularization and Prior Knowledge

To avoid overfitting these models to a high-dimensional dataset, sparsity promoting penalties on edges are a popular technique. The Graphical LASSO (gLASSO) promotes sparsity in the Gaussian Graphical Model by penalizing nonzero entries in the inverse covariance matrix [38]. Similar penalty approaches have been taken for mixed datasets [70]. One difficulty with penalization of likelihoods is that a hyperparameter ($\lambda$) is used to control the severity of the penalty for including edges in the model. Standard techniques to choose this hyperparameter are data-driven techniques like Akaike Information Criterion (AIC) [1], Bayesian Information Criterion (BIC) [109], and Cross-Validation. The main concern with these data-driven procedures is that they can increase bias towards the training dataset and do not leverage prior information.

An alternate approach is to use prior information to select these hyperparameters [143, 156, 74]. A popular method for this is to differentially penalize edges based upon whether or not they are suggested by the prior information sources. In this way, even with nominal improvements in model fit to the data, known relationships will be included in the model. However, unlike treating prior knowledge as absolute constraints, edges that do not improve model fit will still be excluded due to the penalty. The limitation of this approach is the necessity to curate accurate prior knowledge to achieve model accuracy. For a particular dataset, some sources of prior knowledge may not be reflected in the data (e.g. a dysregulated genomic pathway), and these should be addressed by the model learning procedure.

5.2 METHODS

The method we propose to learn undirected structure merges the model proposed by Lee and Hastie [70] with our prior information framework (Chapter 3). We choose the Lee and Hastie model due to its superior empirical performance and its ability to learn interactions among categorical variables. We first describe this model in detail, and then we describe how it is used as the base learning procedure in the prior information evaluation procedure.
Lastly, we describe how we evaluate the method on simulated data, and how we curate prior knowledge sources for genomic applications.

5.2.1 Prior Information Mixed Graphical Models (piMGM)

A Mixed Graphical Model (MGM) is an undirected graphical model proposed by [70] to characterize the joint distribution (Equation 5.3) over a dataset with both continuous and discrete variables.

\[
p(x, y; \theta) \propto \exp \left( \sum_{s=1}^{p} \sum_{t=1}^{p} -\frac{1}{2} \beta_{st} x_s x_t + \sum_{s=1}^{p} \alpha_s x_s + \sum_{s=1}^{p} \sum_{j=1}^{q} \rho_{sj} (y_j) x_s + \sum_{j=1}^{q} \sum_{r=1}^{q} \phi_{rj} (y_r, y_j) \right) \tag{5.3}
\]

Here, \(\theta\) represents the full set of parameters, \(x_s\) represents the \(s^{th}\) of \(p\) continuous variables and \(y_j\) represents the \(j^{th}\) of \(q\) discrete variables. \(\beta_{st}\) represents the edge potential between continuous variables \(s\) and \(t\), \(\alpha_s\) represents the continuous node potential for \(s\), \(\rho_{sj}\) represents the edge potential between continuous variable \(s\) and discrete variable \(j\), and finally \(\phi_{rj}\) represents the edge potential between discrete variables \(r\) and \(j\). This model has the favorable property that its conditional distributions are given by Gaussian linear regression and Multiclass Logistic Regression for continuous and discrete variables respectively (Equations 5.4 and 5.5).

\[
p (x_s \mid x_{/s}, y, \Theta) = \frac{\sqrt{\beta_{ss}}}{2\pi} \exp \left( -\frac{\beta_{ss}}{2} \left( \frac{\alpha_s + \sum_j \rho_{sj} (y_j) - \sum_{t \neq s} \beta_{st} x_t}{\beta_{ss}} - x_s \right)^2 \right) \tag{5.4}
\]

\[
p (y_r \mid y_{/r}, x, \Theta) = \frac{\exp \left( \sum_s \rho_{sr} (y_r) x_s + \phi_{rr} (y_r, y_r) + \sum_{j \neq r} \phi_{rj} (y_r, y_j) \right)}{\sum_{i=1}^{L_r} \exp \left( \sum_s \rho_{sr} (i) x_s + \phi_{rr} (i, i) + \sum_{j \neq r} \phi_{rj} (i, y_j) \right)} \tag{5.5}
\]

Learning this model over high dimensional datasets directly is computationally infeasible due to the computation of the partition function, so to avoid this, a proximal gradient method is used to optimize the penalized negative log pseudolikelihood. This negative log
pseudolikelihood is given in Equation 5.6. To prevent overfitting, nonzero parameters are penalized with $\lambda$ selected using the method described in [112] (StEPS, Equation 5.7). Here, $\lambda_{CC}$ is a penalty parameter only for edges between continuous variables (CC = Continuous-Continuous), $\lambda_{CD}$ and $\lambda_{DD}$ are for mixed edges and edges only using discrete variables, respectively. $\| \cdot \|_F$ refers to the Frobenius norm of a matrix. For evaluation experiments in this chapter and Chapter 6, StEPS is used as a baseline approach without prior information.

$$\tilde{l}(\Theta|x,y) = -\sum_{s=1}^{p} \log p(x_s|x/s, y; \Theta) - \sum_{r=1}^{q} \log p(y_r|x, y/r; \Theta) \quad (5.6)$$

$$\minimize_{\Theta} l_{\lambda}(\Theta) = \tilde{l}(\Theta) + \lambda_{CC} \sum_{s=1}^{p} \sum_{t=1}^{s-1} |\beta_{st}| + \lambda_{CD} \sum_{s=1}^{p} \sum_{j=1}^{q} ||\rho_{sj}||_2 + \lambda_{DD} \sum_{j=1}^{q} \sum_{r=1}^{j-1} ||\phi_{rj}||_F \quad (5.7)$$

To optimize this objective function the proximal gradient optimization method is used. We use the algorithm as specified in [112] with the modification for efficiency that instead of waiting for the edge parameters themselves to converge, we terminate the algorithm when the number of different edges in the graph produced by the algorithm remains unchanged for three consecutive iterations. The parameters themselves are not critical to this chapter as only the graphical structure is studied (non-zero parameters).

To develop a structure learning algorithm that incorporates prior knowledge, we apply the prior information incorporation and evaluation method (Chapter 3) to this problem. Here, the variables to be selected are the edges in the graphical model. The mixed graphical model learning procedure is the base selection procedure, and the $\lambda$ parameters are the hyperparameters to be selected. We refer to the full procedure as Prior Information Mixed Graphical Models (piMGM).

Like before, each parameter is split into a parameter for variables where prior information exists ($WP$) and where prior information is not available ($NP$) (Equation 5.8). Note that when performing the subsampling procedure (Figure 3.2), each $\lambda$ value is treated independently, instead of ranging over all possible combinations of the three parameters. Though it is expected that these parameters will interact, it is computationally infeasible to perform
a large grid search over these parameters. Similarly, computing the score (Equation 3.11), for each parameter is done independently and only for the subset of variables referred to by the parameter (e.g. for $\lambda_{CC}^{NP}$ only edges between continuous variables are used in Equation 3.11).

$$\min_{\Theta} l_\lambda(\Theta) = \tilde{l}(\Theta) + \lambda_{CC}^{NP} \sum_s \sum_{t<s} |\beta_{st}| \mathbb{1}_{(s,t) \notin WP} + \lambda_{CC}^{WP} \sum_s \sum_{t<s} |\beta_{st}| \mathbb{1}_{(s,t) \in WP} + \lambda_{CD}^{NP} \sum_j \sum_{r>j} ||\rho_{sj}||_F \mathbb{1}_{(s,j) \notin WP} + \lambda_{CD}^{WP} \sum_j \sum_{r>j} ||\rho_{sj}||_F \mathbb{1}_{(s,j) \in WP} + \lambda_{DD}^{NP} \sum_j \sum_{r>j} ||\phi_{rj}||_F \mathbb{1}_{(r,j) \notin WP} + \lambda_{DD}^{WP} \sum_j \sum_{r>j} ||\phi_{rj}||_F \mathbb{1}_{(r,j) \in WP}$$

(5.8)

5.2.2 Description of Data Sources

To determine whether piMGM can accurately learn undirected graphical model structure from mixed data, we used both simulated and real expression data from the Cancer Genome Atlas (https://www.cancer.gov/tcga). For the expression data, two main research questions were addressed: 1) scoring biological pathway activity and 2) network structure learning to understand the drivers of disease sub-type.

**Simulated Datasets** Simulated datasets of varying sizes were generated using the Lee and Hastie simulation method from Tetrad VI (http://www.phil.cmu.edu/tetrad/). First, a Directed Acyclic Graph (DAG) was generated uniformly at random with number of edges equal to the number of nodes. Each node was randomly assigned to be a continuous or categorical variable with equal probability. This DAG was parameterized with random edge weights in the range: $[-1.5, -0.5], [0.5, 1.5]$. Independent samples were generated from the DAG to produce a final dataset. To compare the estimated graph learned by piMGM, the DAG was converted to its equivalent ‘moralized’ undirected version, which maintains the independence relationships present in the original DAG [27].

Prior sources were generated based on the ground truth DAG using two different methods for: 1) scoring prior sources or 2) evaluating network structure inference. To evaluate scoring prior sources, a source $i$ was a random selection of $E_i$ edges, where $10 \leq E_i \leq 2N$ and $N$ is
the number of edges in the data generating moralized graph. \( T_i \) edges were randomly selected from the ground truth graph to include in the prior, where \( T_i \) is draw uniformly at random from the range: \([1, \min(E_i, N)]\). Thus, the reliability of prior \( i \) is \( \frac{T_i}{E_i} \), which measures what percent of a prior’s edges are present in the data generating graph. Finally, we randomly selected \( E_i - T_i \) ‘false’ edges from the set of all edges not in the data generating graph. These prior sources were given as ‘hard priors’ which means that the prior information given by a source only contained the values null and 1 corresponding to the absence or presence of an edge in the prior, respectively. This was to ensure that these simulated prior sources would resemble real prior information in the form of biological pathways.

For the full network inference experiments, it was assumed that all prior sources provide information about the same edges, but with different reliability, to test the ability of piMGM to successfully synthesize prior information from multiple sources. We tested cases where the sources only provide information for the true edges in the ground truth graph and for all edges uniformly at random. The number of edges provided, and the number of experts were experimentally controlled. The edges for which the experts give prior information were determined randomly, and the information itself was a ‘soft prior’ with a real numbered value ranging from \((0.6, 1)\) for a reliable source giving information about a true edge. All other cases the soft prior was drawn uniformly at random from \((0, 1)\). This is in contrast to the ‘hard priors’ from the previous ‘pathway’ experiment.

**TCGA Breast Cancer Data** To evaluate piMGM on real data, the TCGA-BRCA RNA-Seq expression dataset was used. This data included gene expression measurements from 800 breast tumor samples and 95 matched normal samples. Prior information consisted of 33 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [60], which were selected by the same criteria as a related pathway enrichment method [80], but excluding those with fewer than 10 expressed genes or fewer than 20 gene–gene interactions. The gene–gene interactions encoded in each pathway were used as a ‘hard prior’ with a value of 1 if the edge existed in the pathway, and no information otherwise. 645 genes were included in the final dataset, consisting of the union of the genes in the 33 pathways, excluding 2735 consistently lowly expressed genes or those without a variance of at least 0.5.

To evaluate the usefulness of full network inference for classification, breast cancer sub-
type information for each tumor sample was obtained from [58]. This determination was used as a ground truth for classification experiments, and the fifty genes commonly used to compute this classification on microarray data, Prediction Analysis of Microarray 50 (PAM50) was used as a prior information source. In this case, the edges between each gene in the PAM50 list and the categorical variable denoting sub-type was given a prior probability of 1, while all other edges had no prior information. 'Coexpression' priors consisted of gene-gene similarity data from two popular databases: STRINGdb [129] and Harmonizome [104]. 'Random' priors were shuffled versions of the PAM50 gene sets, where 50 randomly selected genes were chosen to have a 100% probability of being connected to the target variable. This was to test the ability of piMGM to handle incorrect prior knowledge on real data.

Breast cancer diagnosis and prognosis are commonly divided into four main subtypes: Luminal A, Luminal B, HER2+, and Triple-Negative breast cancer. The main driving distinction for these subtypes is the presence or absence of hormone receptors on the tumor cell surface. In the pathway analysis experiments, we aim to identify biological pathways distinguishing between hormone receptor positive (Luminal A and Luminal B) vs negative (HER2+, Triple-Negative) sub-types. In the prediction experiments, we aim to build a model to distinguish between all four groups based on gene expression and clinical data.

5.3 RESULTS

Next, we demonstrate the effectiveness of piMGM on simulated and real datasets. First, we discuss the ability of piMGM to evaluate prior knowledge sources on simulated datasets. Subsequently, we show how this evaluation method can be used to identify differentially regulated pathways between two sub-types of breast cancer. Next, we discuss the ability of piMGM to learn more accurate network structure and how the quality and amount of prior information affects its accuracy. Lastly, we use the network learned by piMGM from the TCGA expression data to predict breast cancer sub-types.
5.3.1 Prior Knowledge Evaluation Results

We first evaluate how well piMGM can assess the reliability of a prior information source relative to the dataset being analyzed.

Figure 5.1 presents the results of applying piMGM on 25 simulated datasets with 100 variables each, 15 hard prior information sources, with 200 and 1000 samples. The figure demonstrates the strong inverse correlation between the predicted deviance ($\tau$: Equation 3.7) of each simulated prior from the ground truth graph and its aforementioned reliability score. The major outliers from the trendline are those pathways that provide information about relatively few edges. This is because if a prior has little information, it may or may not match spurious correlations in the dataset by chance. Thus, it is prudent to consider both the reliability score of the prior as well as its P-value to determine its reliability. On a graph by graph basis, for the 200 sample case, the mean correlation between $\tau$ and the reliability score across all 25 datasets was 0.992 ($\pm$0.006). This further confirms the accuracy of the piMGM model for each individual dataset. When increasing the sample size of the data to 1000 samples, there is no significant difference in terms of the correlation between the predicted and actual reliability. The major difference here is that the trendline moves toward a -1 slope. This indicates that with more samples, small changes in the true reliability do not have a disparate effect on the estimated value of $\tau$. As a result of this analysis, when applying the prior evaluation scheme to piMGM, we normalize the unreliability score $\tau$ by dividing by the mean of the null distribution. This prevents disparate effects due to the amount of prior information each source gives.

As a first application of piMGM on biological data, we used it to identify differentially regulated pathways in breast cancer patients from TCGA. In this application, the pathways are prior knowledge sources and we scored their reliability on tumor and normal samples. ‘Reliable’ pathways are then those that are active in the given sample dataset (low p-value), and ‘unreliable’ are those that are not. 33 pathways from KEGG were used for this experiment, as described in the Methods section.

piMGM found eight of the 33 pathways to be differentially regulated (FDR-corrected P-value $< 0.1$) between receptor positive (Luminal A and B sub-types) and receptor negative
(HER2 and Triple-Negative) sub-types (Table 5.1). Five of the eight pathways were also found to be differentially expressed by another pathway identification method, NetGSA [80], so we further examined the three remaining pathways to understand what piMGM found mechanistically. Here, we focus on T and B Cell Receptor Signaling.

Both the T and B cell receptor signaling pathways shared a common sub-network that was driving their identification by piMGM as significant. This network included genes AKT3, PIK3CA, PIK3CD and PIK3R1 (Figure 5.2, left), all of which are genes of significance in cancer. Several studies have found changes in the regulation of AKT3 in receptor negative breast cancer ([21]; [85]) and one recent study has found changes in expression of PI3K/Akt across receptor sub-types [23], consistent with our identified sub-network. An interesting finding from this pathway is that the direct connection between PIK3CA and AKT3 is more present in receptor negative tumors. PIK3CA is an oncogene whose aberrant activation results in AKT3 activation which can lead to uncontrolled cell proliferation and tumorigenesis [48]. It is possible that this sub-network elucidates distinct mechanisms of AKT3 over-
Table 5.1: Differentially regulated pathways by receptor status (positive: Luminal A and B; negative: HER2, triple negative) in Breast Cancer (FDR < 0.1) Pathways in red found by piMGM but not NetGSA.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>P-value (+)</th>
<th>P-value (-)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glutathione metabolism</td>
<td>0.507</td>
<td>0.091</td>
<td>[76]</td>
</tr>
<tr>
<td>Glycolysis</td>
<td>&lt; 10^{-5}</td>
<td>0.129</td>
<td>[108]</td>
</tr>
<tr>
<td>Neurotrophin signaling</td>
<td>0.702</td>
<td>0.074</td>
<td>[92]</td>
</tr>
<tr>
<td>Notch signaling</td>
<td>&lt; 10^{-5}</td>
<td>0.223</td>
<td>[54]</td>
</tr>
<tr>
<td>Pentose phosphate</td>
<td>0.025</td>
<td>0.239</td>
<td>[14]</td>
</tr>
<tr>
<td>B Cell Receptor signaling</td>
<td>0.141</td>
<td>0.004</td>
<td>[49]</td>
</tr>
<tr>
<td>Insulin signaling</td>
<td>0.098</td>
<td>0.384</td>
<td>(see text)</td>
</tr>
<tr>
<td>T cell receptor signaling</td>
<td>0.507</td>
<td>0.058</td>
<td>(see text)</td>
</tr>
</tbody>
</table>

activation in different breast cancer sub-types. piMGM also identified an NFATC1 sub-network of the T cell receptor pathway as critical for breast cancer development (Figure 5.2). NFAT1 is a nuclear factor that alters transcription in T-cells in response to T-cell receptor stimulation, but NFAT1 mRNA is also found in breast tissue [106]. A recent study found that the NFAT1 pathway was active in triple negative breast cancer but not in other sub-types, and that the pathway was critical for metastasis and tumorigenesis [97]. Though KEGG labels these pathways as lymphocyte related pathways (T and B cells), the reason these pathways were found to be differentially regulated was due to the sub-components critical to breast cancer progression in different sub-types.

All eight differentially regulated pathways appear to have an established relationship to breast tumor sub-type. piMGM not only identified them as significant, but it also identified which parts of them are the most significant in breast cancer sub-typing (Figure 5.2), thus pointing to the mechanisms that influence this sub-typing. Follow up studies can further investigate the specific mechanisms implied by piMGM networks. piMGM tends to identify
Figure 5.2: Edge differences between receptor positive and negative breast tumor samples for a common sub-network of the T cell and B cell receptor signaling pathways (left) and the T cell receptor signaling pathway (right). Red (green) edges have higher (lower) empirical probability in receptor positive cancer according to piMGM.

fewer differentially regulated pathways than conventional analysis techniques, largely because piMGM uses stronger conditions than typical pathway enrichment methods. piMGM uses independence changes between genes while conventional enrichment methods ignore network connectivity and focus on absolute expression changes. Individual gene expression changes may be related to established pathways, but independence relationships query the precise network information stored in KEGG pathways to determine differential regulation.

5.3.2 Structure Learning Results

Next, we present results on simulated and biological data for inference of the full network structure on datasets with mixed continuous and categorical variables. For simulated datasets, we compare the learned network structure using piMGM to the ground truth moralized graph. For real datasets we use a prediction task for a target variable of interest as a proxy for correct network structure, because the true gene regulatory network is unknown.

To evaluate the ability of piMGM to incorporate prior information, piMGM was compared to several baseline approaches on the simulated data described in Methods. Four
Figure 5.3: Accuracy of piMGM inferred networks vs. other approaches on simulated data with 100 variables, 500 samples and 5 prior sources. The datasets differ in the percent of edges with prior information (0, 10, 30, 60%) and the number of reliable priors (1, 3, 5 out of 5). The blue bars represent the $F_1$ score accuracy of the baseline networks (no priors). Priors gave information about only true edges (left) or all edges (right) chosen randomly.

Baseline methods were used for comparison (Figure 5.3, left columns, 0% prior). STARS [77] is a network stability approach that tunes a single regularization parameter ($\lambda$) for all edges in MGM without using prior information. StEPS [112] is an extension of this approach that uses three regularization parameters, one for each edge type (Continuous-Continuous, Continuous-Discrete, Discrete-Discrete). The Oracle graph is MGM run with the set of three regularization parameters that maximize accuracy, and Oracle One $\lambda$ is an equivalent approach using only one regularization parameter for the whole network.

We compared the baseline results with piMGM given five prior information sources, where both the percent of edges with prior information (10%, 30%, 60%) and the number of reliable experts among those 5 (reliable experts: 1, 3 or all 5) were varied. Figure 5.3 displays the amount of prior information (x axis) vs. the $F_1$ score of the learned graph compared to the ground truth (moralized) graph. As expected, an increase in the ratio
Figure 5.4: Accuracy of piMGM inferred networks vs. other approaches on simulated data with 100 variables, 200 samples and 5 prior sources. Priors gave information about only true edges (left) or all edges (right) chosen randomly.

of reliable to unreliable priors increases overall accuracy ($F_1$ score) for all percentages of edges with prior information. We also see that in most cases the $F_1$ score is not affected by unreliable priors even when 4 out of 5 experts are unreliable (Figure 5.3, yellow bars; compared to STARS and STEPS methods with no prior). Interestingly, if prior information is available for 10% of edges, piMGM is equivalent to approaches with no priors, regardless of the reliability of the information sources. However, if 60% of edges have prior information, piMGM outperforms even the Oracle graph given that the prior information is somewhat reliable ($p < 0.01$). If the prior information is highly unreliable, with 30% prior, piMGM does not have degrading performance, and it maintains at least the quality that it would have had without any prior ($p > 0.05$). This is desirable in cases where prior information may or may not be well represented in the system under study. When ‘experts’ provide priors for edges that are not in the data generating graph the results are nearly identical (Figure 5.3).

We evaluated piMGM further on datasets generated from graphs with 200 samples, and
these results follow a similar pattern (Figure 5.4). The first observation here is that there is a much larger gap between the baseline approach and the optimal hyperparameter setting with fewer samples. piMGM smoothly fills this gap, as even with 10% prior information, regardless of how many reliable priors are available piMGM performs better than the baseline approach. When more prior information is available, piMGM can reach near optimal performance with 60% prior information. These results hold regardless of whether the reliable sources provide information only about true edges (left), or both true and false edges (right).

Lastly, we evaluate the ability of piMGM to recover network structure from biomedical mixed data. Since ground truth for the whole network is impossible to obtain in biological systems due to incomplete understanding, we evaluate piMGM by using the learned network for classification of a single target variable. The Markov Blanket of the target variable was used to predict breast cancer sub-type.

As mentioned above, breast cancer consists of four sub-types with varying prognoses: triple negative, Luminal A, Luminal B and HER2+. We used a 10-fold nested cross validation setup to test the ability of piMGM learned models to classify these sub-types. For each training set, we ran piMGM with each of these sources as prior information and inferred a full network from data consisting of the genes described in the Methods section along with clinical variables of: Gender, Age, Vital Status, and Sub-type. Then, we used the genes connected to the sub-type variable as features in a multi-class logistic regression model to evaluate prediction accuracy on the testing set. We compared classification accuracy on networks learned using five sets of priors: No prior information, five random sets of fifty genes each connected to the Sub-type variable (Random Prior), the two gene-gene similarity matrices from STRING and Harmonizome (Coexpression Prior), the PAM50 set of genes along with the coexpression matrices (All Relevant Priors) and all of these prior information sources together (All Priors). We found that there is no significant difference in classification accuracy between each of the five models (Figure 5.5, left columns), though the "No Prior", "Co-expression Prior", and "All Priors" are trending higher. This means that piMGM is not significantly affected by random (unreliable) priors even when they are the only source of information. On the other hand, incorporating appropriate prior information results into models that require significantly fewer features to achieve the same accuracy (Figure 5.5, right
Figure 5.5: Results of applying piMGM to the TCGA Breast Cancer dataset with varying prior information sources. Classification accuracy (left), feature stability (center), and number of features (right) were computed.

In particular, using no prior information resulted in a very dense model with no improvement in accuracy. This could be suggestive of overfitting. The best balance between prediction accuracy, stability, and parsimony was achieved by using the co-expression priors with piMGM.

5.4 DISCUSSION

In this chapter, we have discussed learning an undirected probabilistic graphical model from data by evaluating and incorporating prior information. The method we propose (piMGM) is an extension of the MGM method propose by Lee and Hastie that models continuous and categorical data using linear and logistic regressions, respectively. piMGM is an instantiation of the prior information evaluation and incorporation scheme described in Chapter 3.

After evaluation on simulated datasets, we find that piMGM is able to successfully evaluate prior information sources based on how well their information is reflected in the data.
This appears to render piMGM resilient to unreliable prior information, as even when a majority of priors are unreliable, the algorithms performs no worse than having no prior information at all. Meanwhile, with reliable prior information the algorithm can offer a significant performance improvement.

On real expression data, we find that the prior information evaluation scheme can be used successfully to identify active pathways within breast cancer sub-types. The corresponding p-values can be used to understand differences between sub-types. Though, evaluating the full learned graphical model is difficult, we found that when evaluating the model on a specific prediction task, prior information about gene co-expression gives a more stable model than having no prior information at all. Including random prior information did not significantly alter the prediction model learned by piMGM, reinforcing the findings from simulated datasets.

Overall, piMGM is similar to piPref-Div in that it appears to be a conservative way to incorporate prior information. Even with very accurate prior information, the potential performance benefit is somewhat modest due to the fact that prior information is only used to choose hyperparameters for the model instead of as absolute constraints. For the same reason, very inaccurate prior information does not significantly adversely affect the learned model. Future studies for the methodology can focus on how to allow for a wider performance improvement while maintaining the risk avoidance. One strategy could be to include edges that are widely agreed upon by prior information sources deemed to be reliable, regardless of what the data says. Another future direction is to explore promising sources of prior information. Some potential sources that were not explored here are: literature mining to identify genes causally related to clinical variables (Gender, Outcomes, etc.), correlation matrices from past expression analyses using the Gene Expression Omnibus (GEO), guilt by association from genes that are targeted by similar drugs, and similar gene base pair sequences themselves. As new potential sources of information are developed, the prior information evaluation framework can be used to quickly evaluate these sources to benefit future analyses.
6.0 PREDICTION OF RESPONDERS TO A PROPHYLACTIC CANCER VACCINE

In the previous chapters I have introduced a framework to evaluate and incorporate external domain knowledge. I applied this framework to two distinct problems in variable selection and in learning the structure of undirected graphical models. In this chapter, I apply these methods to a biomedical application: prediction of response to a prophylactic cancer vaccine. First, I give some background about cancer and the vaccine. Then, I discuss the methods and prior knowledge used to evaluate the learned prediction models. Finally, I discuss the accuracy of the learned models and explain their potential biological interpretation.

6.1 BACKGROUND

Despite advances in cancer treatment such as immunotherapy, cancer remains a costly disease to patients both emotionally and fiscally. An alternative to immunotherapy is immunoprevention that becomes possible with the advent of preventative cancer vaccines. Cancer vaccines for cancers of viral etiology can be prevented by vaccines that prevent the initial infection. The majority of human tumors, however, are of non-viral origin and vaccines for the prevention of non-viral cancers typically target a tumor-associated antigen (TAA). These are mutated or non-mutated antigens that are presented differently on tumor cells compared to normal tissues. This allows the immune system to discriminate between these tissue types [84]. Previously, vaccines based on TAAs have been administered as treatments to patients with cancer; however, these have had limited therapeutic benefit [79, 40]. This is now known to be due to the immunosuppressive nature of the environment surrounding the tumor [134].
A more promising approach could be to prophylactically administer vaccines to patients at high-risk for cancer [37]. One vaccine that has shown promise in this setting is a vaccine targeting the TAA Mucin 1 (MUC1). MUC1 is a protein expressed on the cell surface of normal epithelial tissue, but in tumor tissue it is over-expressed in a hypoglycosylated form [62]. Abnormally expressed MUC1 has been shown to induce immune responses, and it has been found in the premalignant setting as well [51]. These properties make it an appropriate target for a prophylactic vaccine.

Recently, a feasibility clinical trial for a vaccine targeting MUC1 was carried out in individuals with a history of advanced colonic adenomas (precursors to colon cancer). These individuals are at increased risk for recurrence of adenomas and progression to colon cancer. The end points of the trial were safety and immunogenicity of the vaccine. Immunogenicity was measured by the vaccine’s ability to elicit anti-MUC1 Immunoglobulin G (IgG) [62]. The vaccine was found to be safe but immunogenic in fewer than 50% of the 43 vaccinated individuals. These results have now been repeated in a much larger placebo controlled clinical trial of 55 vaccinated participants and 55 placebo controls. The main question for both trials was whether there was a genomic basis that predisposed some individuals to respond to the vaccine, and secondly, whether response to the vaccine could be predicted. To address this question, we applied the variable selection and graphical modeling methods to transcriptomic data obtained on peripheral blood mononuclear cells (PBMC) obtained pre-vaccination and 2 weeks after the first vaccine, as well as the placebo controls. We aimed to identify genomic markers predictive of response and to build a model that accurately predicts response using pre-vaccination measurements.

6.2 METHODS

In this section, I briefly describe the data sources used for the study and the methodology used to build and evaluate the predictive models.
Table 6.1: Clinical characteristics of patients whose PBMC’s were sequenced at two weeks post-vaccination.

<table>
<thead>
<tr>
<th></th>
<th>Responders (N = 26)</th>
<th>Non-Responders (N = 20)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort (% Mayo)</td>
<td>57.69</td>
<td>0.00</td>
<td>0.0028</td>
</tr>
<tr>
<td>Gender (% Female)</td>
<td>69.23</td>
<td>40.00</td>
<td>0.093</td>
</tr>
<tr>
<td>Age</td>
<td>55.5 (6.97)</td>
<td>58.8 (8.21)</td>
<td>0.158</td>
</tr>
<tr>
<td>IgG Week 0</td>
<td>0.240 (0.190)</td>
<td>0.148 (0.179)</td>
<td>0.101</td>
</tr>
<tr>
<td>IgG Week 12</td>
<td>1.826 (1.239)</td>
<td>0.15 (0.187)</td>
<td>&lt; 10^{-4}</td>
</tr>
<tr>
<td>BMI (%Overweight)</td>
<td>53.85</td>
<td>35.0</td>
<td>0.3782</td>
</tr>
<tr>
<td>BMI (%Obese)</td>
<td>26.92</td>
<td>45.0</td>
<td></td>
</tr>
</tbody>
</table>

6.2.1 Description of Data Sources

The dataset for this study was RNA-Seq data from peripheral blood mononuclear cells (PBMC’s). Each patient had an advanced adenoma (premalignant growth) of the colon, and was administered the MUC1 vaccine. The RNA sequencing was collected both pre-vaccination and two weeks post vaccination. Both datasets were analyzed in this study. The goal was to determine if graphical modeling methods can develop effective prediction models at both time points to discriminate responders from non-responders.

The Week 2 dataset consisted of 59 samples (patients) from two cohorts (one from the feasibility trial - Finn and the other from the placebo controlled trial - Mayo; both trials were identical in design and execution) with 26,424 measured genes. After removing genes consistently not expressed (sum of log counts per million reads were less than 0.25 on average) and removing genes with less than 0.5 variance, 7,968 genes remained. The log counts per million read (log CPM) of the genes were used to address the strong skew in the data. Some patients in the dataset had high anti-MUC1 IgG levels pre-vaccination. These patients were referred to as disease positive responders, having spontaneously generated antibodies to the
Table 6.2: Clinical characteristics of non-responder Group 1 patients whose PBMC’s were sequenced pre-vaccination.

<table>
<thead>
<tr>
<th></th>
<th>Responders (N = 17)</th>
<th>Non-Responders (N = 29)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort (% Mayo)</td>
<td>47.06</td>
<td>82.76</td>
<td>0.027</td>
</tr>
<tr>
<td>Gender (% Female)</td>
<td>58.82</td>
<td>27.59</td>
<td>0.075</td>
</tr>
<tr>
<td>Age</td>
<td>55.12 (7.59)</td>
<td>60.97 (7.99)</td>
<td>0.018</td>
</tr>
<tr>
<td>IgG Week 0</td>
<td>0.2382 (0.1983)</td>
<td>0.0634 (0.2186)</td>
<td>0.315</td>
</tr>
<tr>
<td>IgG Week 12</td>
<td>1.115 (1.127)</td>
<td>0.3034 (0.2394)</td>
<td>0.0094</td>
</tr>
<tr>
<td>BMI (%Overweight)</td>
<td>58.82</td>
<td>34.48</td>
<td>0.266</td>
</tr>
<tr>
<td>BMI (%Obese)</td>
<td>23.53</td>
<td>41.38</td>
<td></td>
</tr>
</tbody>
</table>

MUC1 expressed on the adenoma prior to its removal. These samples were removed from the dataset since their biology is expected to be different from non-responders and responders. After this, 46 patients remained in the data. When graphical model learning was done alone, for feasibility, only 750 genes with the highest variance were used. When variable selection was performed, the entire set of 7,968 genes were included. The clinical characteristics of the cohort are given in Table 6.1. None of these characteristics are significantly different in responders vs. non-responders except for IgG at Week 12 (the outcome we are predicting), and cohort because one cohort only provided responders for this study.

The pre-vaccination dataset had 62 samples (patients) from the same cohorts with 30,017 measured genes. After the same pre-processing steps, 7,267 genes and 46 patients remained. Since the trial is placebo-controlled, patients that received the placebo are distinct from the true non-responders (that were administered the vaccine). Though placebo status is still confidential for this trial, we were given two groups of non-responders, one of which were true non-responders. Thus, the pre-vaccination analysis was done on each group separately, with the initial aim of identifying the true non-responders.

The clinical characteristics of the cohort are given in Tables 6.2 and 6.3. Overall, we find
Table 6.3: Clinical characteristics of non-responder Group 2 patients whose PBMC’s were sequenced pre-vaccination.

<table>
<thead>
<tr>
<th></th>
<th>Responders (N = 17)</th>
<th>Non-Responders (N = 17)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort (% Mayo)</td>
<td>47.06</td>
<td>70.59</td>
<td>0.2958</td>
</tr>
<tr>
<td>Gender (% Female)</td>
<td>58.82</td>
<td>47.06</td>
<td>0.7312</td>
</tr>
<tr>
<td>Age</td>
<td>55.12 (7.59)</td>
<td>58.65 (7.850)</td>
<td>0.1923</td>
</tr>
<tr>
<td>IgG Week 0</td>
<td>0.2382 (0.1983)</td>
<td>0.1953 (0.1214)</td>
<td>0.453</td>
</tr>
<tr>
<td>IgG Week 12</td>
<td>1.115 (1.127)</td>
<td>0.2035 (0.1206)</td>
<td>0.0043</td>
</tr>
<tr>
<td>BMI (%Overweight)</td>
<td>58.82</td>
<td>23.53</td>
<td>0.111</td>
</tr>
<tr>
<td>BMI (%Obese)</td>
<td>23.53</td>
<td>41.18</td>
<td></td>
</tr>
</tbody>
</table>

that the non-responder group 1 is more dissimilar in terms of clinical characteristics with significant differences in Age and a trending difference in Gender. In Group 2 there appears to be no clear clinical differences between responders and non-responders, though BMI has a slight trend.

Prior information was curated from the same sources as described in Chapter 4. The databases used were DisGeNet, MSigDB, HUGO Gene Families, Physical gene distance, and STRING db. Since STRING db is a compendium of several channels of information (text mining, databases, coexpression, etc.), in this chapter we do a further analysis by evaluating each source of information separately. In addition, significant pathways were evaluated as in Chapter 5. For these evaluations the Hallmark gene lists from MSigDB were used [75], consisting of KEGG, Reactome, and Biocarta pathways. Only those connections that were transcription factor to target relationships as defined by the RegNetwork database [78] were included, since protein-protein interactions do not necessarily imply co-expression.
6.2.2 Model Development and Evaluation

In order to develop and evaluate model predictions, a leave-one-out cross validation approach was used. Iteratively, each individual sample is used as a testing set with the remaining samples used as a training set. On the training set, feature selection was performed, predictive features were extracted, and a prediction model was learned to predict week 12 antibody levels (IgG Week 12). This model was then applied to the single left out sample and the predictions were saved. After employing this process for all samples, the predictive accuracy of the model was estimated using root-mean-square error from the true antibody levels (both raw and log-transformed). In addition, the predicted antibody level was used as a score and AUC of response vs. non-response was computed, with response defined as twice baseline levels of IgG.

Prediction models were developed by using undirected graphical structure learning to find variables adjacent to the IgG Week 12 variable. These variables were then fed as input to a linear regression model. Feature stability is defined as the average intersection over union (Tanimoto set similarity) of the features selected in each fold.

6.3 RESULTS

In this section, we evaluate the developed models of response using piMGM and piPref-Div. Since we expect discrimination between responders and non-responders to be easier post-vaccination, we begin by evaluating our pipeline on the Week 2 data. Then, we apply the best performing methods to the pre-vaccination dataset to determine the most likely true non-responder group and to build a predictive model of response prior to vaccine administration.

First, we evaluate the prior knowledge sources by examining the “trust” our methods placed in the prior knowledge sources. Then, we perform leave-one-out cross validation to build graphical models. We evaluate these graphical models for different prior information sources and with/without our variable selection method. Finally, we discuss the biological significance of the models.
6.3.1 Evaluation of Prior Knowledge Sources

Two separate analyses were performed to evaluate the sources of prior knowledge. The first analysis aimed to determine how reliable external databases were to the RNA-Seq data under study. For this analysis, both piMGM and piPref-Div were used so that their results could be compared. The results are reported in Table 6.4. This analysis shows that all databases are significantly more represented in the dataset than random datasets of equal size (all p-values < $10^{-3}$) for both piMGM and piPref-Div. The most noticeable differences between the piMGM and the piPref-Div evaluations are for the co-expression data and the physical distance data which are both much better expressed according to piPref-Div. The co-expression information is expected as this is exactly what piPref-Div is querying: correlations between genes as opposed to conditional dependence between genes. Physical distance is known to be related to co-expression, so this may be a noisy version of a similar phenomena [81].

To better understand the STRING database, which is a composite score from multiple distinct sources, we performed the same analysis on each of its individual components (Table 6.5). When examining the prior sources at this level, the difference between piMGM and piPref-Div is more substantial. Three components: fusion, co-occurrence, and neighborhood are all substantially different, and the first two are not statistically significant according to piMGM. Fusion refers to genes that are one continuous gene in other organisms, and this has shown to be related to co-expression [149]. Co-occurrence and neighborhood refer to genes that occur in organisms and genes that appear near each other in other genomes respectively. Due to the low weight given by piMGM, and the fact that these sources have little information, they were excluded from the rest of the analysis. In addition, in lieu of the STRING composite score, we choose to use the remaining STRING component sources, so they can be evaluated accordingly.

6.3.2 Evaluation of Week 2 Model Predictions

Next, we used these sources to build a prediction model of IgG Week 12. First, we tested the ability of piMGM to build an accurate prediction model with several sources of prior
<table>
<thead>
<tr>
<th>Database Sources</th>
<th>$\tau$</th>
<th>p-Value (Pref-Div)</th>
<th>p-Value (piMGM)</th>
<th>Percent Prior (Pref-Div)</th>
<th>Percent Prior (piMGM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DisGeNet Gene-Disease Mapping</td>
<td>0.985</td>
<td>0.9922</td>
<td>$&lt; 10^{-3}$</td>
<td>1.87</td>
<td>2.59</td>
</tr>
<tr>
<td>MSigDB Coexpression</td>
<td>0.8762</td>
<td>0.9461</td>
<td>$&lt; 10^{-3}$</td>
<td>0.492</td>
<td>0.403</td>
</tr>
<tr>
<td>STRINGdb Composite</td>
<td>0.9479</td>
<td>0.973</td>
<td>$&lt; 10^{-3}$</td>
<td>2.46</td>
<td>3.21</td>
</tr>
<tr>
<td>Physical Gene Distance</td>
<td>0.9468</td>
<td>0.9836</td>
<td>$&lt; 10^{-3}$</td>
<td>4.09</td>
<td>4.04</td>
</tr>
<tr>
<td>HUGO Gene Families</td>
<td>0.9427</td>
<td>0.9451</td>
<td>$&lt; 10^{-3}$</td>
<td>0.499</td>
<td>0.693</td>
</tr>
</tbody>
</table>
Table 6.5: Evaluation of STRINGdb component sources used in the MUC1 Modeling Study

<table>
<thead>
<tr>
<th>Prior Source</th>
<th>$\tau$ (Pref-Div)</th>
<th>$\tau$ (piMGM)</th>
<th>p-Value (Pref-Div)</th>
<th>p-Value (piMGM)</th>
<th>Percent Prior (Pref-Div)</th>
<th>Percent Prior (piMGM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>0.866</td>
<td>0.9298</td>
<td>$&lt; 10^{-3}$</td>
<td>$&lt; 10^{-3}$</td>
<td>0.198</td>
<td>0.29</td>
</tr>
<tr>
<td>Textmining</td>
<td>0.9617</td>
<td>0.9803</td>
<td>$&lt; 10^{-3}$</td>
<td>$&lt; 10^{-3}$</td>
<td>1.92</td>
<td>3.45</td>
</tr>
<tr>
<td>Fusion</td>
<td>0.8623</td>
<td>1.021</td>
<td>$&lt; 10^{-3}$</td>
<td>0.1979</td>
<td>$1.87E^{-3}$</td>
<td>$7.04E^{-4}$</td>
</tr>
<tr>
<td>Coexpression</td>
<td>0.8919</td>
<td>0.937</td>
<td>$&lt; 10^{-3}$</td>
<td>$&lt; 10^{-3}$</td>
<td>0.568</td>
<td>1.26</td>
</tr>
<tr>
<td>Co-occurrence</td>
<td>0.8894</td>
<td>0.9661</td>
<td>$&lt; 10^{-3}$</td>
<td>0.1113</td>
<td>0.014</td>
<td>$7.40E^{-3}$</td>
</tr>
<tr>
<td>Neighborhood</td>
<td>0.9054</td>
<td>0.9734</td>
<td>$&lt; 10^{-3}$</td>
<td>0.0011</td>
<td>0.115</td>
<td>0.085</td>
</tr>
<tr>
<td>Experimental</td>
<td>0.9389</td>
<td>0.9615</td>
<td>$&lt; 10^{-3}$</td>
<td>$&lt; 10^{-3}$</td>
<td>0.824</td>
<td>0.774</td>
</tr>
</tbody>
</table>

information. Full Databases refers to the all of the sources evaluated in the previous section except for the three excluded STRING components. Pathway Prior refers to KEGG and MSigDB gene lists with connections given by transcription factor - target relationships from RegNetwork. STRING Pathway Prior are the same relationships, except instead of using probability 1.0 for all edges, the value that the composite STRING database has for the edge is used.

The prediction results are given in Table 6.6. In terms of prediction accuracy, the best performing method is StEPS, which does not use prior information to learn an undirected model. However, we note that this method uses on average 25 features to predict antibody level. In a dataset consisting of only 46 samples, this is suggestive of overfitting. Using prior information for graphical modeling (piMGM) appears to improve the parsimony of the model; however, this comes at a slight loss in accuracy and a larger loss of stability. Incorporating Pref-Div into the pipeline appears to stabilize these models without any significant loss of accuracy. Using prior information as part of the full pipeline appears to improve accuracy, but only when the prior information is used only by Pref-Div (PD-PCA + StEPS). When using prior information at both stages of the pipeline, the results appear to be substantially
Table 6.6: Comparison of prediction models for antibody response to MUC1 vaccine built from RNA-Seq data two weeks post-vaccination.

<table>
<thead>
<tr>
<th>Algorithm (Prior Source)</th>
<th>Feature Stability</th>
<th>Mean Features</th>
<th>Mean Cluster Size</th>
<th>RMSE</th>
<th>Log RMSE</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>StEPS (None)</td>
<td>0.762 (0.125)</td>
<td>25.43 (2.28)</td>
<td>-</td>
<td>1.656</td>
<td>1.28</td>
<td>0.819</td>
</tr>
<tr>
<td>piMGM (Pathway)</td>
<td>0.69 (0.166)</td>
<td>5.07 (1.31)</td>
<td>-</td>
<td>1.576</td>
<td>1.416</td>
<td>0.735</td>
</tr>
<tr>
<td>piMGM (STRING Pathway)</td>
<td>0.687 (0.179)</td>
<td>5.17 (1.43)</td>
<td>-</td>
<td>1.549</td>
<td>1.39</td>
<td>0.735</td>
</tr>
<tr>
<td>piMGM (Full Database)</td>
<td>0.593 (0.196)</td>
<td>3.78 (1.53)</td>
<td>-</td>
<td>1.93</td>
<td>1.47</td>
<td>0.662</td>
</tr>
<tr>
<td>PD-PCA + StEPS (No Prior)</td>
<td><strong>0.885 (0.128)</strong></td>
<td><strong>2.15 (0.42)</strong></td>
<td><strong>2.39 (1.76)</strong></td>
<td>1.97</td>
<td>1.77</td>
<td>0.748</td>
</tr>
<tr>
<td>PD-PCA + StEPS (Full Database)</td>
<td>0.837 (0.149)</td>
<td>2.43 (0.54)</td>
<td>3.84 (3.09)</td>
<td>1.602</td>
<td>1.79</td>
<td>0.744</td>
</tr>
<tr>
<td>PD-PCA + piMGM (Full Database)</td>
<td>0.847 (0.144)</td>
<td>2.41 (0.54)</td>
<td>3.90 (3.09)</td>
<td>2.08</td>
<td>1.73</td>
<td>0.704</td>
</tr>
</tbody>
</table>
worse in terms of prediction accuracy. This could be due to the choice of prior information used. Though using the full databases appeared strongly favored by piPref-Div, they were only somewhat favored by piMGM (Tables 6.4 and 6.5). One potential solution could be to use separate prior sources for each stage of the pipeline, based upon the information needed (correlation for piPref-Div vs. conditional independence for piMGM).

In order to evaluate the models further, we next examined the predictions made by StEPS (High accuracy), PD-PCA + StEPS (Full Database) (Good stability and accuracy), and piMGM (STRING Pathway) (Good accuracy and parsimonious) (Figure 6.1). From these plots, it is clear that the linear assumption appears to hold for this dataset, as the predicted and actual IgG values appear to have a linear relationship. Further, we note that both StEPS and piMGM allow for a threshold to be drawn to distinguish most responders from non-responders, though the predictions of all three models are not substantially different from one another.
6.3.3 Biological Significance of Week 2 Models

To understand whether the modeling methods proposed are interpretable, we evaluated the learned models based on the mechanistic knowledge they provide. Since piMGM with String and Pathway prior performed relatively well in prediction, we used the prior reliability score calculated by this model to see which pathways were well reflected in the data. In particular, we looked for pathways that were significantly expressed in the responder and non-responder groups (measured by low $\tau$ value by piMGM).

The results are presented in Table 6.7 for non-responders and Table 6.8 for responders. The first observation from these results is that the non-responder group appears to have a more standard biological phenotype, as most of the pathways are statistically different than random according to piMGM. This is not the case for the pathways in the responder group. From the "driving connections" column, we observe that the main factors distinguishing the responder and non-responder group are NFkB and STAT1. This is expected as these genes have a significant role in regulating the immune response [115, 46]. Since NFkB was independently implicated by differential expression analyses, our collaborators at Case Western performed phosphorylation flow-cytometry experiments to confirm differential activity of NFkB in B-Cells of Responders vs. Non-responders. These results confirmed that high-responders had higher levels of phosphorylated (active) NFkB at Week 2 and a higher change in activity from Baseline to Week 2 (Figure 6.2). Also, we note that nearly all of the pathways identified by the algorithm are directly involved in the immune response. Though, this does not give novel information, it confirms the relevance of prior evaluation.

Next, we examined the individual genes and gene clusters selected by piMGM and piPref-Div respectively, and the results are presented in Figure 6.3. Here, the size of the variable indicates the stability of the feature across cross-validation folds, and the color of the feature indicates direction of association with response (green = positive association). Four genes appear to be the prominent causes of response. CD79A is a protein involved in the B-Cell antigen receptor complex, which is the functional unit on a B-cell that is necessary for recognition of antigens and activation. Since, B-cells are the cell which produce antibody this is a logical relationship. Similarly HLA-DQA2 is involved in the MHC II complex which
help present antigen (MUC1) to CD4+ T-cells to stimulate the immune response. PGM5 is slightly overexpressed in CD33+ Myeloid cell populations which is a known inhibitor of the humoral immune response [62]. It is not immediately clear what functional relevance MYO16 contribute to response, but this could be a topic for further investigation.

Next, we examined the clusters frequently predicted by piPref-Div to be related to response. Two major clusters of genes were identified. The first cluster (JPX, KDM5C, SCARNA9L, ACSL4, TXLNG, XIST, ERCC6L) consisted of seven genes. This exact cluster was identified in 11 of 46 folds. All of these genes are X-linked genes overexpressed in females (corrected $p < 0.05$), except for ACSL4. This association of gender and response has been suggested by the previous study [62]. The second stable cluster (CEP55,RP11-384K6-6,KIF11,DLGAP5,CCNA2) appeared in 15 of the 46 folds. These genes are all overexpressed in B lymphoblasts (a progenitor of a B-Cell) [147, 127]. This cluster is underexpressed in responders, which may indicate that post-vaccination, there should not be a high level of circulating B-progenitors but functional B cells instead.
Table 6.7: Top five pathways significantly different than random priors on RNA-Seq samples from non-responders.

<table>
<thead>
<tr>
<th>Name</th>
<th>P-Value Corr (NR)</th>
<th>P-Value Corr (R)</th>
<th>Driving Connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>PID ATF2 PATHWAY</td>
<td>0.014</td>
<td>0.001</td>
<td>JUN - PLA, FOS - IL8</td>
</tr>
<tr>
<td>REACTOME INTERFERON SIGNALING</td>
<td>0.014</td>
<td>0.001</td>
<td>STAT1 - IFIT3</td>
</tr>
<tr>
<td>KEGG PATHWAYS IN CANCER</td>
<td>0.020</td>
<td>0.002</td>
<td>SPI1 - RXRB, SPI1 - CKS1B,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NFKB2 - TGFA, NFKB2 - RARA,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NFKB1 - LAMB3,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>FOS - LAMB3, FOS - DVL2</td>
</tr>
<tr>
<td>REACTOME CYTOKINE SIGNALING IN IMMUNE SYSTEM</td>
<td>0.020</td>
<td>0.002</td>
<td>STAT1 - FCGR1B, STAT1 - CISH</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NFKB2 - NFKB1B,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>EGR1 - REL, REL - NFKB2,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>JUN - IL6, TCEB2 - SOCS3</td>
</tr>
</tbody>
</table>
Table 6.8: Top five pathways significantly different than random priors on RNA-Seq samples from responders.

<table>
<thead>
<tr>
<th>Name</th>
<th>P-Value Corr (R)</th>
<th>P-Value (R)</th>
<th>P-Value Corr (NR)</th>
<th>Driving Connections</th>
</tr>
</thead>
<tbody>
<tr>
<td>REACTOME CYTOKINE SIGNALING IN IMMUNE SYSTEM</td>
<td>0.058</td>
<td>0.003</td>
<td>0.020</td>
<td>STAT1 - YES1, NFKB2 - IL6,</td>
</tr>
<tr>
<td>REACTOME TRANSCRIPTIONAL REGULATION OF WHITE ADIPOCYTE DIFFERENTIATION</td>
<td>0.058</td>
<td>0.003</td>
<td>0.074</td>
<td>IRF3 - PIN1, STAT1 - LCK</td>
</tr>
<tr>
<td>KEGG TOLL LIKE RECEPTOR SIGNALING PATHWAY</td>
<td>0.192</td>
<td>0.016</td>
<td>0.068</td>
<td>CEBPB - CEBPD</td>
</tr>
<tr>
<td>PID RB 1PATHWAY</td>
<td>0.257</td>
<td>0.028</td>
<td>0.020</td>
<td>CEBPA - CEBPD</td>
</tr>
<tr>
<td>REACTOME INTERFERON SIGNALING</td>
<td>0.400</td>
<td>0.055</td>
<td>0.014</td>
<td>STAT1 - SOCS3</td>
</tr>
</tbody>
</table>
6.3.4 Graphical Models from Pre-Vaccination Data

Based on the previous analysis, two methods were evaluated on the pre-vaccination data: piMGM (String Pathway) and piPref-Div + StEPS. The leave-one-out cross validation results are presented in Table 6.9 with separate rows for Group 1 non-responders and Group 2 non-responders. The graphical models can distinguish between Group 2 non-responders and true responders, but this is not the case for Group 1 (AUC is close to 0.5). This implies that Group 2 are the true non-responders who did not receive the placebo. Based on this result, we focus upon the piMGM model for Group 2, since piMGM had good discrimination accuracy (AUC) and good prediction of antibody level (low RMSE).

A visualization of the predictions made for each patient (left) and the learned graphical model structure (right) are displayed in Figure 6.4. The main observation from the predic-
Table 6.9: Evaluation of structure learning methods to discriminate responders from non-responders using the pre-vaccination RNA-Seq data.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Group</th>
<th>Stability</th>
<th>Number of Features</th>
<th>AUC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>piMGM</td>
<td>Group 1</td>
<td>0.769 (0.195)</td>
<td>4.30 (1.28)</td>
<td>0.586</td>
<td>0.857</td>
</tr>
<tr>
<td>piMGM</td>
<td>Group 2</td>
<td>0.630 (0.154)</td>
<td>7.59 (1.67)</td>
<td>0.637</td>
<td>0.863</td>
</tr>
<tr>
<td>piPref-Div and STEPS</td>
<td>Group 1</td>
<td>0.787 (0.249)</td>
<td>2.57 (1.24)</td>
<td>0.562</td>
<td>0.971</td>
</tr>
<tr>
<td>piPref-Div and STEPS</td>
<td>Group 2</td>
<td>0.964 (0.105)</td>
<td>2.41 (0.86)</td>
<td>0.664</td>
<td>1.29</td>
</tr>
</tbody>
</table>

The antibody titer predictions are relatively accurate, but the discrimination between responders and non-responders is not as accurate as Week 2. This could suggest that the transcriptomic signal pre-vaccination is similar to two weeks, but reduced. Examination of the stably selected genes gives a similar interpretation as the Week 2 model. DHRS9 and LPL are two genes that appear to be present in CD33+ Myeloid cell populations and underexpressed in responders (like PGM5 in the post-vaccination model). SIGLEC14 is a gene that is important to innate immunity and is expressed on monocytes, and the SIGLEC family is known to be a regulator of both innate and adaptive immunity [29]. It is unclear exactly how SIGLEC14 may relate to an antibody response. BMI was an interesting association, as overweight individuals were better responders than both normal and obese individuals. Lastly, MYOM2 is expressed primarily on Natural Killer cells (CD56+). But up-regulation of MYOM2 in the blood was previously implicated in a general immune disorder called Common Variable Immunodeficiency which decreases antibody production [91]. This is consistent with our results, since MYOM2 is down-regulated in responders.

Next, we compared the Week 2 and Baseline signatures. Overall, we find that all of the genes in the Week 2 signature are related to response the same way on the pre-vaccination data (up vs. down regulated), except for CD79A. We found CD79A to be overexpressed
CD79A is a B-cell marker so the expression of this gene could be indicative of the relative frequency of B-cells in the blood. Experiments performed by our collaborators confirmed that responders have a lower frequency of B-cells pre-vaccination, but a higher change in B-cell frequency than non-responders (Figure 6.5). Only one gene (HLA-DQA2) from the Week 2 signature appears in the cross-validation experiment for the pre-vaccination samples. None of the genes from the pre-vaccination signature appear in the Week 2 cross-validation experiment. This could indicate that there are better markers to distinguish the mechanism of response, but the pre-vaccination signature is what "primes" an individual to be a good responder. An interesting next step is to follow the expression of both signatures throughout the trial.

### 6.4 DISCUSSION

In this chapter, I presented an application of the first part of the pipeline to study response to a prophylactic cancer vaccine from transcriptomic and clinical data. Here, both piPref-Div
and piMGM were applied to this problem to 1) build a model for response to the vaccine and 2) uncover biological knowledge as to why patients are primed to respond (Baseline data) and also the mechanism of response (Week 2 Data).

Overall, the Week 2 prediction model successfully distinguishes responders and non-responders as measured by AUC from a leave-one-out cross validation experiment. The results show that the most accurate model was identified by using graphical model learning with no prior information; however, this model was relatively unstable and used a large number of features. Using piMGM, the number of features used was drastically reduced without a substantial change in accuracy. Though piPref-Div did not have a substantial change in accuracy, it added a significant amount of stability to the model and further reduced the model complexity. Finally, the best performing models were evaluated in terms of their biological information and uncovered reasonable relationships between gene expression and vaccine response from both piMGM and piPref-Div.

The best performing methods were applied to the pre-vaccination data to determine which individuals were true non-responders and to find a transcriptomic signature to predict response. Based on these results, we hypothesize that the second group of non-responders
were administered the vaccine. We then built a predictive model to discriminate responders from these non-responders. Though this signature is not as accurate as the Week 2 model, we find that there is a predisposition for some individuals to respond to the vaccine. In addition, we find consistent biological meaning from both the pre-vaccination and the Week 2 gene signatures. Some of these analyses were confirmed via ex vivo experiments performed by our collaborators.

This use case demonstrated the ability of piPref-Div and piMGM to learn a reasonable model for a real biomedical research problem. However, in this application, prior knowledge did not have a significant impact on the results. This could be due to the specific prior knowledge databases used. It is future work to explore the biomedical databases further and identify promising choices for use with both piPref-Div and piMGM. Also, to truly validate the prognostic ability of the graphical models, they need to be evaluated on an independent validation cohort. Lastly, in this situation, these graphical models were used to predict a specific target variable instead of as a data exploration tool to suggest experimental interventions. We leave to future work investigating these two important extensions.
Though undirected graphical models can factor joint distributions into conditional dependence relationships in a dataset, they do not give any information about causal direction between variables. An undirected edge in a graphical model could be the result of a true causal relationship, a latent confounder, or selection bias in the dataset. Causal discovery algorithms aim to distinguish these cases and identify the causal relationships between all variables in a dataset. In this chapter, we explore search algorithms for causal discovery from mixed, observational datasets. We use an extensive empirical evaluation to demonstrate the best performing search algorithms for data with latent variables.

7.1 BACKGROUND

In this section, we provide some necessary background information about the causal discovery problem. Intuitively, the causal discovery problem is to use a tabular dataset as input in order to output a directed graphical model where directions indicate causal relationships. The difference between these models and the models discussed thus far are that undirected graphical models are probability models. These models parameterize the joint probability density of the variables in the data in terms of local conditional probability densities.

Causality cannot be captured by parameterizing the joint probability density alone, because causality relies upon understanding the probability density of variables following interventions. \( p(X|Y) \) gives a prediction for the variable \( X \) when observing the value of \( Y \); however, it does not give the distribution of \( X \) when \( Y \) is experimentally manipulated. I follow the work of Pearl and others to define the intervention density of \( X \) when \( Y \) is
manipulated as \(p(X|do(Y))\) [94]. If there is no causal relationship between \(X\) and \(Y\), then 
\[ p(X|do(Y)) = p(X|Y) \] However, if \(Y\) is a true cause of \(X\), then in general these densities are not equal. To understand the causal relations among a set of variables, it is no longer sufficient to identify a single joint probability density. Instead, a set of densities corresponding to all possible manipulations of observed variables must be identified. This is the main focus of the causal discovery problem.

Representing this combinatorially large set of probability densities can be simplified by using \textbf{Structural Equation Models} or SEM’s. A SEM \((S = (G,V,e,F))\) is a four-tuple where \(G\) is a graph, \(V\) is a set of variables of interest, \(e\) is a set of error terms, and \(F\) is a set of functional relationships. In a SEM, each variable is functionally determined by its parents. Each variable \(X \in V\) is given by Equation 7.1, where \(Pa(X)\) refers to the parents of variable \(X\) in the graph, and \(e_X\) is the error term for variable \(X\).

\[
X = f(Pa(X), e_X) \tag{7.1}
\]

As a concrete example, if the SEM has linear functional relationships and Normally distributed errors, then \(X\) is given by Equation 7.2, where \(w_i\) is the linear parameter relating \(X\) to its \(i\)'th parent.

\[
X = \sum_{i=1}^{|Pa(X)|} w_i \ast Pa(X)_i + N(\mu_x, \sigma_x) \tag{7.2}
\]

The joint probability density can be determined from these functional equations because the joint factors according to the graph. This means that the full joint density can be represented by a product of local conditional densities [124, 65]. In addition, each of these equations is interpreted to be causal. This means that intervening on a parent of a variable \(X\) redefines \(X\) according to the intervention performed and the functional definition of \(X\). Thus, a SEM gives post-intervention densities for all variables in the graph.

The causal discovery problem in general aims to identify the SEM that generated a sample of a set of variables in the graph (with the assumption that such a SEM exists). The set of variables that are in the sample will be referred to as the \textbf{observed variables}. Since our main goal is to suggest hypotheses to be experimentally validated, we constrain the problem to just identifying the correct graphical structure of the SEM instead of both
the graph and the functional relationships. The graphical structure allows a researcher to answer the questions: 1) What variables can be intervened upon to control the value of $T$? and 2) Will intervening on a variable $X$ result in a change in the variable $Y$?

**Assumptions** Before discussing established algorithms for causal discovery, we first give the necessary assumptions for the methods discussed here. The Causal Markov Condition (Definition 7.1) states that each variable is independent of its non-descendants given its parents. Here, $De(V, G)$ refers to any descendant of $V$ in the graph of the SEM. This assumption implies that every association in the dataset is the result of a causal relationship. Selection bias and redundant variable definitions are two potential cases where this assumption is violated. For a thorough discussion, we refer the reader to [124, 137]. The Causal Faithfulness Assumption (Definition 7.2) links statistical independence measures from the data to inference of graphical structure. Most of the algorithms mentioned here make this assumption; however, recently some methods have tried to relax this [100, 72].

**Definition 7.1. Causal Markov Assumption.** Let $G$ be the graph of the SEM, then

$$\forall V \in G, \forall X \notin De(V, G), (V \perp \!\!\!\!\!\perp X | Pa(V, G))$$

**Definition 7.2. Causal Faithfulness Assumption** A probability density $p$ is **faithful** to a causal graph $G$ iff every conditional independence in $p$ is entailed by $G$.

Lastly, it is assumed that samples are statistically independent and all come from identical distributions (i.i.d.). These three assumptions are generally made in most causal discovery works. In this thesis, we make two additional assumptions. We assume acyclicity, meaning that there are no directed cycles in the SEM graph which generated the data. In addition, we assume that there is no selection bias in the data. This implies that the selection criteria for inclusion in the sample is not based upon a variable with more than one parent in the causal graph.
7.1.1 Causal Structure Learning Algorithms

In this section, we begin to explore algorithms to solve the causal discovery problem. First, we discuss algorithms that make an additional assumption called \textbf{Causal Sufficiency} (Definition 7.3). This definition just says that there are no unobserved variables that are causes of multiple observed variables. This is because these unobserved variables could induce correlations in the data.

\textbf{Definition 7.3. Causal Sufficiency} A set of variables $O$ is \textit{causally sufficient} for a set of variables $V$ and a causal graph $G = (V, E)$ iff $\forall X \in V, |\text{Children}(X, G)| > 1 \rightarrow X \in O$

Without additional assumptions on the functional relationships in the data or on the causal graph [55, 113], one cannot find the unique causal graph in general. Conditional independence alone can only suggest causal relations up to \textbf{Markov Equivalence} (Definition 7.4)

\textbf{Definition 7.4. Markov Equivalent Graphs} A set of graphs $G = (G_1 = (V_1, E_1), ..., G_N = (V_N, E_N))$ are Markov Equivalent iff $\forall i, j \text{Skeleton}(G_i) = \text{Skeleton}(G_j)$ and $G_i$ and $G_j$ share the same unshielded colliders.

With these assumptions, there are two sets of algorithms for learning a causal graph from a data sample: constraint-based algorithms and score-based algorithms. Constraint-based algorithms identify conditional independencies in the data in order to constrain the space of causal graphs consistent with the data. Score-based algorithms attempt to maximize model fit while penalizing complexity of the model using a statistical score like Bayesian Information Criterion. The most popular of these approaches is the FGES algorithm, which has demonstrated great accuracy and scalability on continuous datasets [101, 20]. Constraint-based approaches have shown better performance on mixed datasets with latent variables, and so we choose to focus our discussion on constraint-based approaches. The classic constraint-based algorithm for identifying a Markov Equivalent set of causal graphs is the PC algorithm [123] (Algorithm 7.3). Here, $O(X, Y, S)$ outputs whether $X$ and $Y$ are independent given $S$ in the true causal graph. With a data sample, this is equivalent to performing a conditional independence test, though in this case errors may be made.
Algorithm 7.3 PC Algorithm

1: procedure PC Adjacency(CI Oracle O, Variables V)
2: \( G \leftarrow \) Fully Connected Undirected Graph(V)
3: \( n = -1 \) \hspace{1cm} \( \triangleright \) Conditioning Set Size
4: Sepsets = Empty List
5: while \( \exists V_i, V_j \in V \text{s.t.} |\text{Adj}(V_i, G) \setminus V_j| > n \) do
6: \( n = n + 1 \)
7: for Edge \( E \in G \) do \hspace{1cm} \( \triangleright E = (V_1, V_2) \)
8: if \( |\text{Adj}(V_1, G) \setminus V_2| \geq n \) then
9: for \( S \in \text{Adj}(V_1, G) \text{s.t.} |S| = n \) do
10: if \( O(V_1, V_2, S) \) then \( \triangleright \) Oracle says \( V_1 \) and \( V_2 \) are independent given \( S \)
11: Delete Edge \( E = (V_1, V_2) \) from \( G \)
12: Sepsets = Sepsets + (\( E, S \)) \hspace{1cm} \( \triangleright \) Which set deleted \( E \)?
13: return \( G, \text{Sepsets} \)

14: procedure PC Orientation(Graph G, Sepsets)
15: \( G' = \text{null} \)
16: while \( G' \neq G \) do
17: \( G' = G \)
18: for Unshielded Triple \( (A, B, C) \in G \) do
19: if \( B \notin \text{Sepsets}(A, C) \) then
20: Orient \( A \rightarrow B \) and \( C \rightarrow B \) in \( G \) \hspace{1cm} \( \triangleright \) Orient Colliders
21: for \( (A, B, C) \in G \text{s.t.} A \rightarrow B \text{ and } B \rightarrow C \text{ and } C \notin \text{Adj}(A, G) \) do
22: Orient \( B \rightarrow C \) in \( G \) \hspace{1cm} \( \triangleright \) Orient away from non-colliders
23: for \( (A, B) \in G \text{s.t.} B \in \text{De}(A, G) \text{ and } A \rightarrow B \) do
24: Orient \( A \rightarrow B \) in \( G \) \hspace{1cm} \( \triangleright \) Prevent Directed Cycles
25: return \( G \)
The PC Algorithm begins with a fully connected, undirected causal graph. It uses conditional independence tests with increasing conditioning set size in order to eliminate edges. Finally it orients edges with causal direction using three rules: 1. Orient Colliders 2. Orient away from definite non-colliders, 3. Orient away from cycles. This algorithm is provably sound and complete for identifying the Markov Equivalence Class (MEC) given an oracle for conditional independence tests (large sample limit).

The algorithm outputs the MEC via a **Partially Directed Acyclic Graph** (PDAG). A PDAG represents the MEC with directed edges and undirected edges. A directed edge \( A \to B \) means that all members of the MEC contain this edge direction. An undirected edge means that all members of the MEC contain this edge; however, they differ about the orientation of the edge.

Several modifications have been proposed for the PC Algorithm to improve its accuracy and efficiency on small sample data. PC-Stable is a popular modification that deletes edges connecting independent variables in parallel to avoid the order-dependence of the original PC algorithm [25]. Many strategies have been proposed to handle potential mistakes in identifying the correct separating set. Conservative-PC (CPC) [100] tests all possible separating set to ensure that they all agree about collider orientations. Majority-rule PC uses a majority vote of these potential separating sets.

### 7.1.2 Structure Learning from Data with Latent Confounders

When the set of observed variables is not causally sufficient for the set of variables in the full causal graph, the PC algorithm is no longer accurate [123]. The modified version of the PC algorithm to handle this case is the Fast Causal Inference (FCI) algorithm. To understand this algorithm, a few more definitions are necessary.

**Definition 7.5. A Maximal Ancestral Graph** or MAG is a graph with three edge types: Directed \( \to \), Undirected \( - \) and Bi-directed \( \leftrightarrow \). A MAG has cycles or almost directed cycles (ones which follow directed and bi-directed edges). Lastly, the graph is maximal if for any non-adjacent vertices, there is a set that m-separates them.
Definition 7.6. Variables $A$ and $B$ are m-separated by $S$ in a graph $G$ iff there is no path $p$ between $A$ and $B$ in $G$ such that

- Every non-collider in $p$ is not a member of $S$
- Every collider in $p$ has a descendant in $S$

Definition 7.7. A Discriminating Path in a MAG $G$ for variable $V$ is a path $p = (X, ..., Y, Z, V)$ such that:

- $p$ contains at least three edges
- $V \neq X$ and $V \neq Y$ and $V \in \text{Adj}(Y, G)$
- $X \notin \text{Adj}(Y, G)$, and every vertex between $X$ and $V$ is a collider on $p$ and a parent of $V$

Definition 7.8. A Partial Ancestral Graph or PAG represents an equivalence class of MAG’s. All MAG’s in a PAG have the following properties:

- They have the same adjacencies
- They have the same unshielded colliders
- If any MAG contains a discriminating path $p$ for variable $V$. Then if $V$ is a collider on $p$ in one MAG, it must be a collider on $p$ for all MAG’s.

The PAG (Definition 7.8) represents the set of MAG’s (Definition 7.5) using three edge endpoints. Tails (−) suggest a causal relationship. Arrowheads (>) suggest a lack of causality, and circles (o) are edges that vary for different MAG’s in the equivalence class. With these definitions, we give the FCI algorithm (Algorithm 7.4). The PCAdjacency method performs the adjacency search portion of the PC algorithm, OrientCircles orients all edges in the graph as o−o and ColliderOrient performs collider orientations based on the separating sets identified by PCAdjacency.

The major difference between PC and FCI is that FCI must use conditioning sets outside of just the adjacencies in order to remove edges. Instead of using all possible conditioning sets, this computation can be reduced by the Possible D-Sep characterization [123]. This constrains the necessary conditioning sets that must be tested for each adjacency. For
Algorithm 7.4 FCI Algorithm

1: procedure FCI Adjacency(CI Oracle O, Variables V)  
2: \[ G, \text{Sepsets} \leftarrow \text{PCAdjacency}(O, V) \]  
3: \[ G \leftarrow \text{OrientCircles}(G) \] \quad \triangleright \text{Orients all endpoints as circles}  
4: \[ G \leftarrow \text{ColliderOrient}(G, \text{Sepsets}) \]  
5: \[ G, \text{Sepsets} \leftarrow \text{PossibleD-Sep}(G) \]  
6: return \( G \)  

7: procedure FCI Orientation(Graph G, Sepsets S)  
8: \[ G \leftarrow \text{ColliderOrient}(G, S) \]  
9: \( G' = \text{null} \)  
10: while \( G' \neq G \) do  
11: \[ G' = G \]  
12: \quad for Edge \( E = (A, B) \in G \) do  
13: \quad \quad if \( \exists C \text{ s.t. } A \rightarrow \nabla C \text{ and } B \rightarrow \nabla C \text{ and } A \notin \text{Adj}(C,G) \) then  
14: \quad \quad \quad Orient \( B \rightarrow C \) in \( G \) \quad \triangleright \text{Rule 1: Non-Colliders}  
15: \quad \quad \quad \text{if } \exists C \text{ s.t. } (A \rightarrow B \text{ and } B \rightarrow \nabla C) \text{ OR } (A \rightarrow \nabla C \text{ and } B \rightarrow C) \text{ then}  
16: \quad \quad \quad \quad \text{if } A \rightarrow \nabla C \text{ then}  
17: \quad \quad \quad \quad \quad \text{Orient } A \rightarrow \nabla C \text{ as } A \rightarrow \nabla C \text{ in } G \quad \triangleright \text{Rule 2: Cycles}  
18: \quad \quad \quad \quad \text{if } \exists C \text{ s.t. } A \rightarrow \nabla B \text{ and } C \rightarrow \nabla B \text{ and } A \notin \text{Adj}(C,G) \text{ then}  
19: \quad \quad \quad \quad \quad \text{if } \exists D \text{ s.t. } A \rightarrow \nabla D \text{ and } D \rightarrow \nabla C \text{ then}  
20: \quad \quad \quad \quad \quad \quad \text{Orient } D \rightarrow \nabla B \text{ as } D \rightarrow \nabla B \text{ in } G \quad \triangleright \text{Rule 3: Cycle-Collider Conflict}  
21: \quad \quad \quad \text{if } \exists X_1, \ldots, X_N \text{ s.t. } X_N, \ldots, X_1, B, A \text{ is a discriminating path for } A \text{ then}  
22: \quad \quad \quad \text{if } B \rightarrow \nabla A \text{ and } B \in \text{Sepset}(X_N, A) \text{ then}  
23: \quad \quad \quad \quad \text{Orient } B \rightarrow \nabla A \text{ as } B \rightarrow A \text{ in } G \quad \triangleright \text{Rule 4: Discriminating Path}  
24: \quad \quad \text{else}  
25: \quad \quad \quad \text{Orient } X_1 \leftrightarrow \nabla B \text{ and } B \leftrightarrow A \text{ in } G
orientations, FCI uses the same rules as PC, except it begins with $o-o$ edges and it can use discriminating paths to achieve further orientations.

There have been a number of proposed extensions to improve the speed and output accuracy of the algorithm. Really Fast Causal Inference (RFCI) is a modification of FCI that uses fewer independence tests than FCI to achieve a significant speedup at the cost of reduced information in the final output graph [26]. Anytime FCI is a modification of FCI that allows the algorithm to terminate FCI’s original conditional independence search early with the guarantee that no incorrect information will be presented, though the output may again be less informative than a full run of FCI [121].

Several other algorithms have been proposed to handle the presence of latent variables in causal discovery. The FindOneFactorClusters (FOFC) algorithm imposes restrictions on the graphical structure of the true causal model, thereby rendering it unsuitable to the general causal discovery problem [66]. Overcomplete ICA was used to learn a causal ordering amongst the variables which allows an experimenter to learn experimental predictions, however this approach is limited in scale and requires few latent variables relative to sample size [55]. In addition, none of these approaches are intended to deal with both continuous and categorical data simultaneously.

GFCI [86] is a new procedure that uses a parallelization of the Greedy Equivalence Search algorithm (GES)[101] as an initial step before running FCI; however, to apply this algorithm to mixed variables requires a hybrid scoring function which are still under development for large datasets [6]. Sokolova et al. [119] have developed a scoring function for mixed data that has unsuitable assumptions for categorical data. Their original algorithm, the Bayesian Constraint-Based Causal Discovery algorithm (BCCD) [24] uses a hybrid constraint and score based approach to perform causal search in the presence of latent variables. The extension of this work to handle both discrete and continuous variables uses a modification of the Bayesian Information Criterion (BIC) score. In our evaluation section, we compare our approach to BCCD; however, this mixed score assumes monotonic relationships between the variables, and is only computationally feasible for relatively small datasets.
7.2 METHODS

Here, we discuss the algorithmic changes made to the FCI algorithm to better handle mixed continuous and categorical data. In addition, we give optimizations to the method to improve time complexity and allow it to scale to several hundred variables.

7.2.1 MGM-FCI-MAX

The method we propose is called MGM-FCI-MAX (Algorithm 7.5). It makes two main algorithmic changes to FCI. The primary change is in the collider orientation process. Instead of using the first separating set identified as the true separating set for each disconnected pair of variables, the algorithm uses the MAX search technique to identify this set. It tests all potential separating sets (according to Possible D-Sep) and utilizes the separating set with the largest p-value in its conditional independence test. Then the collider orientation phase is carried out as usual with these selected sepsets.

Due to the MAX search technique, the algorithm has greater time complexity. In order to mitigate this, we also change the adjacency search. Instead of initializing the PC adjacency search with a fully connected graph, it uses the undirected graph learned by the MGM method. Since this is a moralized version of the causal graph, it is guaranteed to be a super-set of the adjacencies in the true causal graph.

7.2.2 Parallelization and Optimization

In this section, we discuss parallelization and optimizations for FCI-MAX to achieve better runtime on high-dimensional data. First we implemented the PC-Adjacency search in parallel using the technique given by [69] without modification. Unfortunately, parallelizing this portion alone is not enough to even equal the runtime of the FCI algorithm due to the extra conditional independence tests performed by MAX. For FCI-MAX, we parallelize the collider orientation phase as well, by giving each thread a job requiring a similar amount of independence tests, computed based upon the number of edges adjacent to the nodes involved in the current unshielded collider.
Algorithm 7.5 MGM-FCI-MAX Algorithm

1: \textbf{procedure} MGM-FCI-MAX \textsc{Adjacency}(Dataset \( D \), Variables \( V \))

2: \hspace{1em} \( G \leftarrow \text{MGMLearn}(D) \) \hspace{1em} \( \triangleright \) Learn Moralized Graph

3: \hspace{1em} \( G \leftarrow \text{PCAdjacency}(D,G, V) \) \hspace{1em} \( \triangleright \) Remove Extra Edges

4: \hspace{1em} \( G \leftarrow \text{OrientCircles}(G) \) \hspace{1em} \( \triangleright \) Orient all edges as circles

5: \hspace{1em} \( G \leftarrow \text{ColliderOrientMAX}(G) \) \hspace{1em} \( \triangleright \) Orient Colliders with MAX strategy

6: \hspace{1em} \( G, \text{Sepsets} \leftarrow \text{Possible D-Sep MAX}(G) \) \hspace{1em} \( \triangleright \) Possible D-Sep with MAX for sepsets

7: \hspace{1em} \text{return} \( G, \text{Sepsets} \)

8: \textbf{procedure} MGM-FCI-MAX \textsc{Orientation}(Graph \( G \),\text{Sepsets} \( S \))

9: \hspace{1em} \( G \leftarrow \text{ColliderOrientMAX}(G) \)

10: \hspace{1em} \( G \leftarrow \text{FCIOrientation}(G) \)

11: \textbf{procedure} ColliderOrientMAX(Graph \( G \))

12: \hspace{1em} \textbf{for} \text{Unshielded Collider} \( C = (V_2,V_1,V_3) \in G \) \textbf{do}

13: \hspace{1em} \hspace{1em} \( \text{max} = -1, \text{set} = \emptyset \)

14: \hspace{1em} \hspace{1em} \textbf{for} \( S \subseteq \text{Adj}(V_2,G) \text{or} S \subseteq \text{Adj}(V_3,G) \) \textbf{do}

15: \hspace{1em} \hspace{1em} \hspace{1em} \( p = \text{CITest}(V_2,V_3,S) \)

16: \hspace{1em} \hspace{1em} \hspace{1em} \textbf{if} \( p > \text{max} \) \textbf{then}

17: \hspace{1em} \hspace{1em} \hspace{1em} \hspace{1em} \( \text{max} = p, \text{set} = S \)

18: \hspace{1em} \hspace{1em} \hspace{1em} \textbf{if} \( V_1 \notin \text{set} \) \textbf{then}

19: \hspace{1em} \hspace{1em} \hspace{1em} \hspace{1em} \hspace{1em} \hspace{1em} \text{Orient} \( V_3* \rightarrow V_1 \) \text{ and} \( V_2* \rightarrow V_1 \) \text{ in} \( G \)
In particular, the task to orient all colliders was subdivided recursively and split into two jobs if the total number of adjacent edges among all of the involved colliders in the current job was greater than the chunk size. The chunk size was set to the total number of edges in the graph \( \|E\| \) multiplied by \( \beta \) (a user-specified parameter) divided by the number of available processors \( \|\text{Cores}\| \). Given \( n \) unshielded colliders \( (X_i - Y_i - Z_i) \), and Graph \( G = (V, E) \) the following rule defines when recursive subdivisions occur:

\[
\text{If } \left( \sum_i |\text{Adj}(X_i, G)| + |\text{Adj}(Z_i, G)| \right) > \frac{|E|}{\|\text{Cores}\|} \text{ then Subdivide} \tag{7.3}
\]

The Subdivide operation splits a job (which would normally be assigned to a single thread) into two jobs which can be executed in parallel by multiple threads. If these jobs are still too large by the above rule, then they can be further subdivided, and so on. In addition, an optimization to the FCI-MAX algorithm was included involving the Possible D-Sep phase. Often times, this phase results in no change to many adjacencies of the graph. Taking advantage of this, we retain colliders whose adjacency sets were unchanged by the Possible D-Sep phase instead of retesting them to determine orientations. This change results in no loss of accuracy and can lead to a substantial reduction in runtime by eliminating redundant independence tests. Note that some independence tests will still be repeated in our current scheme since caching all independence test results is exponential in the number of nodes in the graph. It is future work to explore this time-memory trade-off to determine if caching more independence tests is beneficial for FCI-MAX.

### 7.2.3 Independence Tests for Mixed Data

For using MGM-FCI-MAX to learn a causal graph from a set of samples with mixed continuous and categorical variables, a suitable independence test is necessary. We propose a conditional independence test based upon linear and logistic regressions \([110]\). First, all discrete variables are transformed into several binary indicator variables, one for each category of the original discrete variable. Then, to test the independence of two continuous variables \( X \) and \( Y \), given a conditioning set \( S \), we perform a linear regression of \( X \) onto \( Y \) and \( S \), and we perform a t-test on the coefficient of \( Y \) in this regression. If the p-value of this test
is less than a specified threshold $\alpha$, we reject the null hypothesis of zero partial correlation between $X$ and $Y$ and determine that $X$ and $Y$ are conditionally dependent given $S$. Note that in this situation $S$ may contain both continuous and categorical variables and for the categorical variables, the aforementioned binary indicator variables are used as the predictors. Alternatively, if $X$ or $Y$ is a categorical variable, then we utilize a likelihood ratio test between logistic regression models including the conditioning set $S$ as predictors (null model) and without the conditioning set as predictors. For all causal discovery methods in this thesis, we use this independence test.

### 7.2.4 Dataset Simulation and Evaluation Metrics

Latent variable methods were tested on random graphs of 50 and 500 variables consisting of 50% continuous and 50% three-category discrete variables. The 50 node networks had edge amounts normally distributed with a mean of 100 edges and a standard deviation of 30, while the 500 node networks had edge amounts normally distributed with a mean of 750 edges and a standard deviation of 200.

In order to generate a network with latent confounders, each network had 5 or 20 variables selected to act as latent variables for the 50 and 500 node networks respectively. These nodes were removed from the dataset used for the algorithmic search procedures. To ensure that these variables created confounding effects, only variables with at least two observed children were selected to become latent. To convert this original DAG with latent variables to a ground truth Partial Ancestral Graph (PAG), an oracle for conditional independence relations from the underlying DAG along with the set of nodes from the DAG (excluding the latent variables) were given as input to the FCI algorithm [102].

Datasets were generated from the underlying network using the Lee and Hastie model [70]. Parents contribute to the values of their children in different ways depending upon the type of parent and the type of child. Parents of continuous variables contribute linearly to the means of their children. In particular, continuous parents are multiplied by a single edge parameter, whereas discrete parents have a distinct edge parameter associated with each category of the discrete variable. Parents of discrete variables contribute log-linearly to
the probabilities of each category of their children, with separate edge parameters for each
category of the child. Thus, when discrete variables have 3 categories, edges connecting two
continuous variables consist of a single edge parameter, edges connecting a continuous and
discrete variable consist of 3 parameters, and edges connecting two discrete variables consist
of 9 parameters.

Edge weights were drawn uniformly at random from the union of the regions $[-1.5, .5]$ and $[.5, 1.5]$. For continuous-continuous (CC) edges, the edge parameter is equal to the drawn weight. For continuous-discrete (CD) edges, a vector of 3 values are drawn uniformly from $[0, 1]$, and these values are scaled such that they sum to 0 and shifted so that the largest edge parameter is equal to the original drawn weight. For discrete-discrete (DD) edges, we use the same method as for CD edges, except that we use three permutations of the original vector to fill all 3 rows of the matrix. To ensure the discrete variables behaved as categorical variables, the CD and DD parameters were permuted such that no edge relationship between two variables was monotonic. In this manner, treating all relationships as linear would not suffice in producing correct output, as would be the case with categorical variables in practice.

With these parameters, data was generated from the network. For discrete variables, we first make a random draw uniformly from the interval $[0, 1]$ to be used as an error term that determines the actual value of the variable given the probabilities of each category. For continuous variables, Gaussian error terms with mean equal to zero, and standard deviation uniformly drawn from $[1, 2]$ are set for each variable. Since the dataset is based off of a DAG structure, convergence of the mean values for the variables occurs by starting at the root nodes of the graph, and propagating the values of children. To simulate latent variables in our dataset, we take our original simulated dataset and remove variables corresponding to those set as latent in the true PAG.

Latent variable causal discovery algorithms were evaluated using two scores: adjacency
precision and recall and an experimental orientation score (Table 7.1). Adjacency precision
and recall use the usual definitions of precision and recall where true positives (TP) correspond to correctly identified edges (regardless of orientation), false positives (FP) are edges predicted to be present but absent in the ground truth, and false negatives (FN) are edges predicted to be absent, but present in the ground truth. The experimental orientation
Table 7.1: Experimental Orientation Score to compute orientation precision and recall for latent variable causal discovery algorithms.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>True Edge: A *-o B</td>
<td>1 TP</td>
<td>If (¬Ancestor (B,A))</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 TP Else, 1 FN</td>
</tr>
<tr>
<td>True Edge: A*→ B</td>
<td>½ FP</td>
<td>1 TP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 TP</td>
</tr>
<tr>
<td>True Edge: A*-B</td>
<td>½ FN</td>
<td>1 FN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 TP</td>
</tr>
</tbody>
</table>

score defines TP, FP, and FN according to ancestral predictions, where a correct prediction of ancestry or non-ancestry is a true positive, and false positives and negatives are defined accordingly.

7.3 RESULTS

Here, I present an evaluation of causal discovery methods on synthetic datasets. I focus upon algorithms that do not assume causal sufficiency, and I focus on datasets with latent variables. To discover the contributions of the various strategies to improve the FCI algorithm, I first compare FCI with its variants (FCI-MAX and CFCI) and BCCD. Then I explore how using MGM affects the best performing methods.

7.3.1 Comparison of FCI and Variants

The first experiment aimed to determine which algorithms were most accurate and efficient on 50 Node datasets. Figure 7.1 shows accuracy results for these algorithms across both Adjacency Recovery and Orientation Recovery for each edge type: Continuous-Continuous (CC), Continuous-Discrete (CD), and Discrete-Discrete (DD). In these plots, a single data
Figure 7.1: Precision-Recall Plots of FCI, FCI-MAX, CFCI and BCCD for 50 Node Scale-Free Networks with 1000 Samples and 5 latent variables. Plots are separated by adjacency and orientation recovery in each row, and by edge type in each column; (CC) refers to only edges between two Continuous variables, (CD) refers to edges between Continuous and Discrete variables, while (DD) refers to edges between two Discrete variables.

point refers to an average precision and recall measurement for a particular setting of parameters across all graphs. For each trial $\alpha$ was selected from the set $[0.001, 0.01, 0.05, 0.1]$ to determine the impact of the parameter across a range of values. For BCCD, all combinations of edge probability cutoffs and orientation probability cutoffs were chosen from the set $[0.25, 0.5, 0.75]$, and so there are nine data points for this algorithm. The three columns of the figure correspond to edges between different types of variables (CC, CD, DD).

All FCI modifications produced similar adjacencies but the adjacency recovery for all these algorithms is significantly better than BCCD for edges involving categorical variables (CD, DD) and slightly worse for CC edges. This is expected since BCCD can handle monotonic relationships well (which applies to the CC edges only). In addition, among the correctly identified adjacencies, FCI-MAX tends to orient these with the best balance between
precision and recall. Conservative FCI (CFCI) has high recall in all cases, and FCI has high precision in all cases. This is due to the fact that CFCI refrains from orienting many edges with arrowheads (thereby avoiding false negatives), and FCI orients nearly all edges as arrowheads (thereby avoiding false positives). This may be because FCI uses the first separating set it finds as the true separating set, which results in inaccurate orientations on high dimensional data. Both of these scenarios are not ideal, as it becomes difficult to discern true causal relationships for both CFCI and FCI. In addition, we note that FCI-MAX tends to outperform BCCD in all cases either by having similar precision and better recall or similar recall and better precision.

This experiment was repeated for a more realistic sample size of the data and similar results were obtained (Figure 7.2). The major difference that we note in this experiment is
Figure 7.3: Runtime Experiment with 50 Nodes, 1000 Samples, and 5 Latent Variables. All MGM algorithms were set to $\lambda = 0.15$.

the larger spread of the data points across different parameter settings, thereby increasing the importance of selecting a suitable value for the $\alpha$ parameter. A nice feature of these algorithms for adjacencies is that the $\alpha$ value tends to balance between precision and recall (a high value for $\alpha$ leads to good recall and poorer precision, and a low value leads to the opposite).

### 7.3.2 Incorporating MGM for Fast Skeleton Discovery

Next, the runtime of these algorithms was explored for the same 50 Node, 1000 Sample synthetic datasets on a four core machine. Runtime is displayed both in terms of wall clock time (Figure 7.3b) and in terms of the number of independence tests performed (Figure 7.3a) because wall clock time can be influenced by algorithm independent factors (CPU Specs, Server load, etc.). BCCD was not included in this experiment as the available implementation was in MATLAB, which would result in a biased comparison against the other algorithms written in Java. In addition, since BCCD is a hybrid constraint and score based algorithm, measuring the performance by number of independence tests was not a suitable substitute.
These figures demonstrate that CFCI performs significantly more independence tests which results in large runtime. Similarly, FCI-MAX is significantly slower than FCI. Since FCI is known to not scale well to high dimensional datasets, FCI-MAX and CFCI will also need optimizations in order to scale.

However, the same figure demonstrates the runtime improvement from using the MGM approach for all FCI modifications. In particular, MGM-FCI and MGM-FCI-MAX improve runtime to faster levels than the FCI baseline approach itself. Clearly, the number of independence tests performed when using the MGM approach is significantly less than the standard approaches, and thus the overhead for computing the MGM graph is the only drawback. Note that these results are presented for a low value of $\lambda = .15$ which actually require the largest runtime. This is because low values of $\lambda$ tend to produce the densest graphs and thereby require the most independence tests. The 500 node networks demonstrate that increasing the value of $\lambda$ can lead to even more significant runtime savings (Figure 7.7b.

Next, the parallelization of MGM-FCI-MAX was evaluated. Figure 7.4 shows the effect on runtime of changing the chunk size and the number of cores available for processing. The chunk size determines how small the recursive subdivisions in the orienting colliders phase
should be before processing begins. Thus, a larger chunk size means that processing will begin earlier on larger jobs. The figure shows that parallelization has a significant impact on performance, but that the impact has diminishing returns for a larger number of cores. The chunk size factor $\beta$ does not have a straightforward relationship to runtime as the optimal chunk size depends on the number of cores used. However, $\beta$ equal to or less than 1 is efficient across all scenarios as long as some parallelism is employed (greater than 1 core).

Figure 7.5 displays accuracy results for adding the MGM approach to the FCI modifications with $\lambda$ chosen from the set [0.1 0.15 0.25]. The best performing FCI modifications without MGM were also included in these graphs for comparison purposes (FCI-MAX and CFCI). Note that MGM does not appear to have a significant impact on the learned causal graphs in most cases. MGM-CFCI marginally improves over the original CFCI algorithm in causal orientations, and MGM-FCI changes the FCI orientations to favor recall instead of precision. MGM-FCI-MAX and FCI-MAX are nearly indistinguishable across most parame-
ter settings for CC and CD edges. Some parameter settings for MGM tend to hurt adjacency recall on DD edges. This may be because equal $\lambda$ parameter values were used for all types of edges, though it is known that MGM tends to perform better when a smaller penalty is applied to DD edges. Using a stability based procedure to determine these parameters could enable more accurate predictions in practice [112].

7.3.3 Scaling to 500 Node Networks

To determine the scalability of the approaches, next the algorithms were tested on 500 Node Networks. FCI, MGM-FCI, and MGM-FCI-MAX were evaluated in this section. The CFIC algorithms were not tested as they did not complete a search on the 500 Node Networks (out of memory), and FCI-MAX was not tested as it had nearly identical performance to MGM-FCI-MAX but less efficient runtime. BCCD was unable to complete a search on a 500 Node network within a half day so it was also excluded from this section.

Figure 7.6 displays the accuracy results on these networks, and a similar pattern to the smaller networks appears. The algorithms all produce similar adjacencies though MGM allows the algorithms to maintain high precision across all parameter values. In orientation accuracy, MGM-FCI-MAX produces the most accurate orientations, achieving high precision and recall with almost no effect from changing the parameters of the algorithm. FCI demonstrates a slight edge in recall of DD edges; however, the parameter setting of FCI which achieves the same precision as the MGM approaches tends to have the same recall. This implies that this is a simply a precision-recall trade-off.

Figure 7.7b and Figure 7.7a display the results of measuring the runtime and number of independence tests performed for these algorithms on 8 core machines. Larger values of the $\alpha$ independence test threshold FCI has significantly larger runtime than the MGM based approaches. For small values of $\alpha$ there is no significant difference between FCI and the MGM based approaches in run time. In practice, $\alpha = .05$ tends to be a commonly accepted parameter setting. At this threshold, MGM improves runtime significantly. In all cases, FCI performs significantly more independence tests, and thus will certainly not be scalable if a more computationally costly independence test is used such as a Nonlinear regression
Figure 7.6: Precision-Recall Plots of FCI, MGM-FCI, and MGM-FCI-MAX for 500 Node Scale-Free Networks with 500 Samples and 20 latent variables.

Figure 7.7: Runtime Experiment with 500 Nodes, 500 Samples, and 20 Latent Variables.
test. Overall, we find that MGM-FCI-MAX tends to balance accuracy and efficiency most effectively in our simulated experiments.

7.4 DISCUSSION

In this chapter, I have presented and analyzed several approaches for causal discovery from mixed datasets. I first presented and empirically analyzed strategies for causal discovery from data with latent variables. In addition, I presented a strategy called MGM-FCI-MAX which builds off of the FCI algorithm to improve accuracy and efficiency on large datasets. The results show that using the MAX search strategy to orient edges helps balance precision and recall among all edge types (CC, CD, DD). In addition, using MGM for skeleton identification overcomes the increased runtime of MAX and allows for scalability beyond FCI. Using MGM also maintains the improved orientation predictions of the MAX strategy. Though determining the specific parameters to use for the algorithm is important, in most cases the accuracy tends to be similar across a wide range of parameter values. This demonstrates the usability of the algorithms without an automated parameter selection procedure. One caveat of this analysis is the orientation score used compares how accurate the learned graph is to the best possible causal graph that could be identified using conditional independence. Though this is useful for research, practitioners may be more interested in whether the causal predictions themselves are plentiful and accurate. Developing a score that meets this criteria and evaluating causal discovery algorithms in this way is left to future work.

Here I evaluated the ability of latent variable causal discovery algorithms to learn a causal graph from data. One interesting direction is to evaluate whether using latent variable causal discovery algorithms provides a practical benefit over classical methods like PC. In addition, I assume throughout the chapter that linear and logistic regressions were accurate models of the interactions between continuous and discrete variables. However, it is not clear how well these assumptions will hold in biomedical data due to non-linearity. Some future directions of this chapter include generalizing our approaches to conditions where linearity is not a suitable assumption for the continuous variables. Finally, there is a pressing need to
broaden the empirical analyses done in this chapter to larger datasets and more complicated
distributional assumptions. Another interesting direction would be to analyze real datasets
in several domains with known interventions to determine the accuracy of causal directions.
8.0 IDENTIFICATION OF CAUSAL FACTORS FOR EARLY DETECTION OF LUNG CANCER

In this chapter, I study the usefulness of MGM-FCI-MAX on a real application: early detection of lung cancer. I identify a causal model of lung cancer from low-dose CT scan, smoking history, and demographic data. I show that a prediction model based off of this causal model can predict lung cancer status better than the state of the art on an independent validation cohort of patients. I further show how the model can eliminate 28% of benign cases without missing a single case of lung cancer.

8.1 BACKGROUND

Lung cancer is the leading cause of cancer-related death worldwide, and the second most common cancer diagnosis in the United States [10]. For many years the standard procedure to detect lung cancer was via a chest x-ray, but the National Lung Screening Trial (NLST) demonstrated that low-dose CT scans were a preferable alternative [87]. Though those screened with low-dose CT (LDCT) showed a 20% reduction in mortality, 96% of positive screens turned out to be false positives. These false positives can lead to unnecessary follow-up procedures and emotional and monetary cost to patients [90, 132].

Here, we used demographic data, smoking history, comorbidities and LDCT scan features of lung nodules from the Pittsburgh Lung Screening Study (PLuSS) cohort [145]. PLuSS is a community-based research cohort that recruited 3642 smokers (current or former) from 2002 to 2006. All PLuSS participants received a baseline LDCT scan, and 3423 participants received a follow-up LDCT scan 1 year later. In addition, each PLuSS participant completed
a questionnaire on smoking history, underwent spirometry for pulmonary function testing, and provided a blood sample. A subset of 970 PLuSS participants received biennial LDCT scans from 2006 to 2016, and yearly spirometry and blood draws.

8.2 METHODS

In this section, I describe the data used to build the causal model, how we converted the PAG learned by MGM-FCI-MAX into a prediction model, and our methodology for evaluating the prediction model.

8.2.1 Study population

Training cohort The training cohort included 50 subjects with cancer detected on their baseline LDCT scan and 50 subjects with screen-detected nodules (benign) from a previously evaluated subset of PLuSS participants drawn from the 970 subjects mentioned above. The benign status of the nodules was further confirmed through prolonged follow-up (2 to 15 years). Eight control subjects were excluded because of missing information about CT scan variables (seven) and number of nodules (one). Thus, the final training cohort had 50 cases and 42 controls \((N = 92)\). Based on the inclusion criteria (age 55 to 77, pack-years > 30 and quit smoking < 15 years), the PLuSS cohort has a very homogeneous population and the selected subjects were at very high risk of lung cancer. As a result, age, sex and smoking history were similar in subjects with malignant and benign nodules.

Validation cohort The validation cohort consisted of 126 subjects (44 cases and 82 controls) from the newer PLuSS XX cohort. The data from these subjects were collected independently of the training cohort. Age, sex and smoking history were similar in subjects with cancer and benign nodules in the validation cohort as well. For nodules < 3 cm the validation cohort consisted of 39 cases and 63 control subjects.

Processed Dataset Altogether, the final dataset consisted of 33 features (9 categorical and 24 continuous). This included 18 features manually extracted from the LDCT, 5 features
related to smoking history and intensity, 6 clinical history features, and 4 demographic features. To extract the CT features, an experienced thoracic radiologist identified and characterized the nodules according to size, presence, type (solid, non-solid or part-solid), and where in the lung the nodule appeared. The same methods were used to extract the radiographic features in cancer cases and controls in both cohorts. Each patient could have multiple nodules and some features were specific to individual nodules. Thus, we only included the malignant tumor in cancer patients and the largest benign nodule in control cases. On the validation set, we evaluate our model’s predictions on a per-nodule basis as well.

8.2.2 Development and Evaluation of Prediction Model

Model Building The first step in applying MGM-FCI-MAX to this data was to select the hyper-parameters of the algorithm. The three sparsity parameters $\lambda_{CC}, \lambda_{CD}, \lambda_{DD}$ for the MGM step were selected automatically using StEPS [112]. The value of $\alpha$ for the independence test threshold for MGM-FCI-MAX was set to 0.05. After learning the full network using MGM-FCI-MAX, the direct children and parents of the lung cancer variable were used as predictors in a logistic regression model of lung cancer probability.

Cross-Validation To ensure that the learned causal model and prediction model were insensitive to small variations in the data, we used a 10-fold cross validation procedure. For each training set, we repeated the entire process of learning the causal model, extracting the parents and children, learning a logistic regression model, and evaluating the model on the test set. The stability of each feature was the percent of folds in which it was selected to be a cause or effect of lung cancer. To compare multiple prediction models, we compared the test-set predictions using a bootstrap-based test (pROC package [103]).

For alternate models, we only extract their feature sets and retrained the coefficient weights to learn a logistic regression model. This was done in order to test the quality of the features and not the cohorts used to train these models. On the validation dataset, the model coefficients from the original publications were used as well to determine the applicability of the MGM-FCI-MAX model to the clinical setting.
8.3 RESULTS

In this section, I explore the performance of MGM-FCI-MAX in identifying a reasonable causal model, and as a first step to developing an accurate prediction model.

8.3.1 Learned Causal Model

The causal model learned using MGM-FCI-MAX on the full training cohort is given in Figure 8.1. We note several interesting characteristics of the causal model. First, there is limited interaction between the extracted features from the CT scan and the clinical information. This may be because the imaging of the nodule is at a more granular level than the clinical information which describe long-term behavior and characteristics of the patients. Despite this, we note that many of the interactions found in the model are logical and supported by other works. Sex and education may interact to influence pack-years according to the model, which is supported by sociological studies [34]. Emphysema and bronchitis were associated to one another through a hidden confounder. These two conditions are clinically difficult to distinguish [118].

The three variables directly connected to lung cancer status in the model are the number of vessels surrounding the nodule, the years since the patient quit smoking, and the total number of nodules on the scan. The model suggests that the blood vessels around the nodule are a cause of lung cancer, which is supported by recent publications [141, 47]. The model suggests that the years since the patient quit smoking and the number of nodules are effects of lung cancer. In our training cohort, only 19% of former smokers with benign nodules quit smoking within 2 years of their CT scan. This number jumps to 44% in subjects with lung cancer. This could imply that the onset of lung cancer symptoms influenced the decision to quit smoking; however, this would need to be validated in a behavioral study. Lastly, the number of nodules in a scan could be influenced by whether or not mutations leading to cancer have occurred. A large number of nodules in a scan could be an effect of a non-cancerous state such as histoplasmosis [126].

To ensure that these variables were stable, we examined the variables selected variables
Table 8.1: Features selected by MGM-FCI-MAX for lung cancer prediction in each round of cross validation

<table>
<thead>
<tr>
<th>CV Round</th>
<th>Vessel Number</th>
<th>Nodule Count</th>
<th>Years quit</th>
<th>Nodule Location</th>
<th>Nodule Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 8.1: Learned causal model using MGM-FCI-MAX on the high-risk training cohort for lung cancer detection.

In each round of cross-validation (Table 8.1). We find that the number of nodules, and the number of vessels appear in all ten rounds of cross-validation. The years since the patient quit smoking is still relatively stable, appearing as a causal factor in eight out of ten folds.

8.3.2 Training Cohort Results

To determine the predictive utility of MGM-FCI-MAX, we compare the performance of logistic regression using the features selected by MGM-FCI-MAX versus the performance of features from other lung nodule discrimination models. The models compared in this analysis are given in Table 8.2. The ROC curves for each of the models on the aggregated set of testing folds are given in Figure 8.2. The logistic regression model using MGM-FCI-MAX’s selected features had the highest AUC in cross-validation, though this was not statistically significant against the Brock Full model ($p = 0.16$) and the Bach model ($p = 0.06$). However, there is a large difference in AUC, and these models use 8 and 5 features, respectively. This suggests that MGM-FCI-MAX has selected a predictive yet more parsimonious set of features.
Table 8.2: Description and performance of Lung nodule discrimination models on the training cohort. AUC is reported with $25^{th}$ and $75^{th}$ percentiles.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>p-val</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGM-FCI-MAX</td>
<td>0.882</td>
<td>-</td>
<td>Smoking: Years Quit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Radiographic: Nodule Count, Vessel Number</td>
</tr>
<tr>
<td></td>
<td>0.7857,1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brock Full</td>
<td>0.792</td>
<td>0.16</td>
<td>Demographics: Age, Sex, Family History Cancer</td>
</tr>
<tr>
<td></td>
<td>0.650,0.929</td>
<td></td>
<td>Comorbidities: Emphysema</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Radiographic: Nodule Size, Type, Location, Count</td>
</tr>
<tr>
<td>Brock Parsimonious</td>
<td>0.700</td>
<td>0.01</td>
<td>Demographics: Sex</td>
</tr>
<tr>
<td></td>
<td>0.600,0.792</td>
<td></td>
<td>Radiographic:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Nodule Location, Size</td>
</tr>
<tr>
<td>Bach Features</td>
<td>0.722</td>
<td>0.06</td>
<td>Demographics: Age, Sex</td>
</tr>
<tr>
<td></td>
<td>0.643,0.792</td>
<td></td>
<td>Smoking: Cigs Per Day, Duration, Years Quit</td>
</tr>
<tr>
<td>PLCO Features</td>
<td>0.5613</td>
<td>&lt;0.01</td>
<td>Demographics: BMI, Education,</td>
</tr>
<tr>
<td></td>
<td>0.333,0.778</td>
<td></td>
<td>Family History Ca, Race</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Comorbidities: Ca History, COPD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Smoking: Duration, Status, Intensity, Years Quit</td>
</tr>
</tbody>
</table>
8.3.3 Validation Cohort Results

Next, we compared the model with comparable parameters (Brock Parsimonious) to the logistic regression model trained on the full training set using the features selected by MGM-FCI-MAX (referred to as the Lung Cancer Causal Model or LCCM). For this analysis, we used both the Brock model with the coefficients from the original paper along with the Brock model with retrained coefficients from our training cohort. On the validation cohort, LCCM was significantly more accurate than both Brock models. Specifically, the AUC for LCCM was 0.903 (±0.061), against 0.812 (±0.077) and 0.757 (±0.086) for the original and retrained Brock parsimonious models respectively. These differences were statistically significant at $p < 0.02$.

The previous results were generated using only the largest benign nodule for the control patients. So, we also examine predictions on all benign nodules. We evaluate these predictions on a patient by patient basis, by treating the probability for a patient to have cancer as the largest probability for any of the nodules according to the model. These results are similar to the previous results. LCCM ($AUC = 0.888$) is significantly better than the pretrained Brock model ($AUC = 0.678$, $p < 0.01$) and not statistically significantly better than the original Brock model ($AUC = 0.843$, $p = 0.225$).
Benign Nodule Screening An important use case of this model is to prevent unnecessary, invasive follow-up procedure for those with only benign nodules. Figure 8.4 displays
an important property of our model. A threshold can be drawn at 60.9% which allows a clinician to eliminate 28.3% of patients with benign nodules without misclassifying a single subject with cancer. The kernel density plot (Figure 8.3 shows why this is the case. The probability score for subjects with cancer (red density) is ≥ 0.9 for 82% of cancer subjects. These results demonstrate the potential of this model to improve clinical cost-effectiveness.

8.4 DISCUSSION

In this chapter, I discussed an application of MGM-FCI-MAX to the problem of estimating the probability of lung cancer from LDCT imaging features, smoking history, and clinical factors. The algorithm identified a reasonable causal model for lung cancer from the training cohort with three key factors as causes and effects of cancer: Years since the patient quit smoking, number of nodules, and vessels around the nodule. These three factors are logical based upon prior work and are highly stable in cross-validation experiments.

A logistic regression predictive model based on the learned causal graphs predicted cancer well in cross-validation experiments on the training cohort and on an application to an independent validation cohort. The model had the highest AUC in both cases, and was statistically significantly better than the state of the art model with a comparable number of features (Brock Parsimonious). The model shows potential use in clinical applications as it can screen nearly 29% of benign cases without missing a cancer case. In addition, the model is interpretable, as the three risk factors all have causal links to lung cancer. Before the model can be directly applied to patients, it must be validated in a larger, prospective cohort study.

Though MGM-FCI-MAX was successful in building an accurate prediction model of lung cancer, this use case is only one example. The algorithm still must be validated on higher dimensional datasets. In addition, the cause and effect associations were validated only using other published work as it is impossible to perform interventions on these variables. An important future direction is to apply MGM-FCI-MAX to data where interventions could be performed to impact clinical care such as medicating patients or hospital process analyses.
9.0 CONCLUSIONS

In this chapter, I discuss the results of this dissertation. I outline the specific contributions made by this work. I discuss several open challenges related to the problem addressed by this thesis, and I discuss the potential impact of the computational methods and applications presented here.

9.1 SUMMARY OF CONTRIBUTIONS

The availability of new measurement technology requires novel modeling pipelines uniquely equipped to handle the challenges of integrated biomedical data: high-dimensionality (low sample size, many variables), mixed continuous and categorical data, and large correlations among measured variables. To this end, I proposed an interpretable modeling pipeline to produce causal knowledge from observational data. The causal modeling pipeline gives models that are interpretable and can be used to 1) understand the fundamental causes of disease, 2) prioritize promising hypotheses about human biology, and 3) personalize medical treatments to individuals. The central hypothesis was that learning graphical model structure from integrated biomedical data will be more effective when using multi-step pipelines and when incorporating external domain knowledge. The specific contributions of this thesis are:

- A framework to evaluate and integrate prior information from multiple sources (Chapter 3).
- An application of this framework to address two computational modeling problems: variable selection and learning undirected graphical model structure. I demonstrated that
these methods improve upon baseline approaches with reliable prior information, and are resilient to unreliable prior information. I found that these methods have improved accuracy and stability on real applications to breast cancer outcome and subtype prediction (Chapters 4 and 5).

• I applied these methods to develop a model that accurately predicted response to a prophylactic cancer vaccine from clinical trial transcriptomic and clinical data. I demonstrated that the best models were constructed when the pipeline was used and prior knowledge was incorporated. (Chapter 6).

• I presented an algorithm for causal discovery from data with latent variables and demonstrated its efficiency and accuracy on synthetic data with mixed variables. (MGM-FCI-MAX, Chapter 7).

• I discussed an application of MGM-FCI-MAX to low-dose CT, smoking and demographic data to build a model for the early detection of lung cancer. I evaluated this model and showed that it detects lung cancer better than state of the art approaches in a high-risk cohort (Chapter 8).

9.2 OPEN PROBLEMS

There are several open challenges that motivate potential future work from this thesis. First, I assumed simple linear and logistic regressions were suitable models for the continuous and categorical variables, respectively. Though this may hold for some datasets, it would be a useful contribution to be able to automatically identify where nonlinear interactions could improve models and use nonlinear functions accordingly. In addition, it would be a useful contribution to characterize the expected functional relationships between classes of biological variables (genes, proteins, phenotypes, SNP’s, etc.) based on prior knowledge and to use these accordingly.

Though the prior knowledge methods demonstrated effectiveness on simulated datasets, the results on real datasets were not substantially different from using no prior information for some evaluation criteria. One potential reason could be the curation of prior knowledge.
The sources used in this work were mostly third party sources, as opposed to past genomics experiments. As more graphical modeling or related network analyses are performed, a useful extension of this work would be to curate these sources into a database and use them as prior knowledge directly. Past models may serve as a more effective source of prior knowledge. In addition, the prior knowledge incorporation scheme could be altered to rely more heavily on the reliable sources of prior information, instead of using them solely to choose hyperparameters.

The causal modeling algorithms discussed here made some assumptions known to be unrealistic in biomedical data. Cycles or feedback loops are known to be important components of biological pathways, and should be modeled accordingly. Here, I focused upon continuous and categorical measurements, but an important future direction is to incorporate ordinal and censored data into these causal models. Lastly, some aspects of causality are not captured by these models such as mediating effects and context-specific interactions. Extending the causal models developed here to handle these types of causal relationships would be a useful contribution for biomedical applications.

9.3 BROADER IMPACT

Overall, this dissertation has demonstrated that multi-step pipelines and prior knowledge can be effective strategies to improve graphical modeling of integrated biomedical data. Experiments on simulated datasets demonstrate that these methods show consistent improvements over state of the art approaches, especially in cases where sample size is limited. On a real application to prediction of responders to a prophylactic cancer vaccine, prior knowledge helps identify a parsimonious prediction model, and using multi-step pipelines greatly improves model stability. Further, a multi-step pipeline for causal modeling (MGM-FCI-MAX) identifies a causal model of lung cancer that is more predictive than state of the art models and is interpretable to domain experts. I believe that these experiments and applications are sufficient evidence to support use of this pipeline in particular, and graphical modeling approaches in general, for biomedical research applications. Finally, I note that many of the
concepts and approaches applied here are not specific to biomedical applications. I believe that any high-dimensional dataset could be effectively modeled by these approaches, and would be improved by appropriate domain knowledge.
BIBLIOGRAPHY


