2016

Biological network models for inferring mechanism of action, characterizing cellular phenotypes, and predicting drug response

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

BIOLOGICAL NETWORK MODELS FOR INFERRING MECHANISM OF ACTION, CHARACTERIZING CELLULAR PHENOTYPES, AND PREDICTING DRUG RESPONSE

by

PAULA J. GRIFFIN

B.A., Boston University, 2010
M.A., Boston University, 2010
M.A., Boston University, 2013

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2016

Approved by

First Reader

Eric D. Kolaczyk, PhD
Professor of Mathematics & Statistics
Boston University

Second Reader

Josée Dupuis, PhD
Professor of Biostatistics
Boston University

Third Reader

W. Evan Johnson, PhD
Associate Professor of Medicine
Boston University

To Alex

Acknowledgments

I owe a debt of gratitude to many people for their help and support in completing this dissertation.

First, thank you to my advisor, Prof. Eric Kolaczyk. I am extremely grateful for the time, energy, and kindness you have spent in teaching me, especially in light of some difficult geographic constraints. Whenever I have found myself stuck or out of ideas, you have provided perspective and another approach. I've learned so much more than statistics from working with you.

Thank you also to my committee members, Profs. Josée Dupuis, Evan Johnson, Edo Airoldi, and Jacqueline Milton. Thank you all for sharing your expertise, guidance, and thoughtful suggestions. Further thanks to the other members of the BU Biostatistics and Statistics faculty, who have given me incredible advice throughout this process.

Thank you to the coauthors of the papers on which this dissertation is based: Eric Kolaczyk (Chapters 2, 4), Evan Johnson (2, 4), Edo Airoldi (3), Tatsunori Hashimoto (3), Mumtahena Rahman (3), and Shelley MacNeil (3). Thanks also to Mladen Kolar for sharing code for estimation of multi-omics networks (2), and to Daniel Lancour and Alexander Blocker for their help in processing data from TCGA (2).

Thank you also to the funding agencies that have supported me during my PhD. In particular, I am grateful for the support of training grants from the National Institute of General Medicine (T32 GM74905) and the National Heart, Lung, and Blood Institute (5T32 HL007501). I also thank the Boston University Research Computing Services and the FAS Division of Science Research Computing Group at Harvard University for providing the computational resources for the analyses contained in this dissertation.

To my colleagues at Quora, I am deeply appreciative of your confidence in me and the encouragement to see my thesis through. To my friends and family, thank you for your support and your tolerance for my admittedly unexciting discussions about this dissertation. I consider myself extraordinarily lucky to have had so many people cheering me on.

Lastly, thank you to my husband, Alexander Blocker. Thank you for your support, advice, proofreading, sanity-checking, and code reviews. Most of all, thank you for your patience and unwavering confidence.

BIOLOGICAL NETWORK MODELS FOR INFERRING MECHANISM OF ACTION, CHARACTERIZING CELLULAR PHENOTYPES, AND PREDICTING DRUG RESPONSE

(Order No.)

PAULA J. GRIFFIN

Boston University, Graduate School of Arts and Sciences, 2016

Major Professor: Eric D. Kolaczyk, Professor of Mathematics & Statistics

ABSTRACT

A primary challenge in the analysis of high-throughput biological data is the abundance of correlated variables. A small change to a gene's expression or a protein's binding availability can cause significant downstream effects. The existence of such chain reactions presents challenges in numerous areas of analysis. By leveraging knowledge of the network interactions that underlie this type of data, we can often enable better understanding of biological phenomena. This dissertation will examine network-based statistical approaches to the problems of mechanism-of-action inference, characterization of gene expression changes, and prediction of drug response.

First, we develop a method for multi-target perturbation detection in multi-omics biological data. We estimate a joint Gaussian graphical model across multiple data types using penalized regression, and filter for network effects. Next, we apply a set of likelihood ratio tests to identify the most likely site of the original perturbation. We also present a conditional testing procedure to allow for detection of secondary perturbations.

Second, we address the problem of characterization of cellular phenotypes via Bayesian regression in the Gene Ontology (GO). In our model, we use the structure of the GO to assign changes in gene expression to functional groups, and to model the covariance between these groups. In addition to describing changes in expression, we use these functional activity estimates to predict the expression of unobserved genes. We further determine when

such predictions are likely to be inaccurate by identifying GO terms with poor agreement to gene-level estimates. In a case study, we identify GO terms relevant to changes in the growth rate of *S. cerevisiae*.

Lastly, we consider the prediction of drug sensitivity in cancer cell lines based on pathway-level activity estimates from ASSIGN, a Bayesian factor analysis model. We use penalized regression to predict response to various cancer treatments based on cancer subtype, pathway activity, and 2-way interactions thereof. We also present network representations of these interaction models and examine common patterns in their structure across treatments.

Contents

List of Tables

List of Figures

List of Abbreviations

ASSIGN	Adaptive Signature Selection and InteGratioN (Shen et al., 2015)
AUC	Area under [receiver operator characteristic] curve
BIC	Bayesian information criterion
CNV	Copy number variation
EBIC	Extended Bayesian information criterion (Chen and Chen, 2008)
ENCODE	Encyclopedia of DNA Elements (ENCODE Project Consortium, 2004)
GI_{50}	Dose required to cause 50% growth inhibition
GO	Gene Ontology (Ashburner et al., 2000)
GSEA	Gene Set Enrichment Analysis (Subramanian et al., 2005)
GWAS	Genome-wide association study
HDP	Hierarchical Dirichlet Process (Teh et al., 2006)
HMEC	Human mammary epithelial cell
ICBP	Integrative Cancer Biology Program (Daemen et al., 2013)
KEGG	Kyoto encyclopedia of genes and genomes (Kanehisa and Goto, 2000)
MSE	Mean squared error
MRCA	Most recent common ancestor
NF	Network filtering
PCA	Principal components analysis
ROC	Receiver-operator characteristic
SNP	Single nucleotide polymorphism

SSEM	Sparse simultaneous equation model
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins (Szklarczyk et al., 2011)
TCGA	The Cancer Genome Atlas (Cancer Genome Atlas Network, 2012)

Chapter 1

Introduction

This dissertation examines several applications of network models to the biological problems of drug response prediction and mechanism-of-action inference. In Chapter 2, we develop a method for perturbation detection in multi-omics biological data, using a conditional Gaussian graphical model and a series of likelihood ratio tests in order to determine the most likely perturbation site. Chapter 3 deals with characterization of cellular phenotypes via Bayesian regression in the Gene Ontology (GO; Ashburner et al., 2000). We use the GO to map genes to interpretable functional groups, and the relationships in the ontology to inform the covariance between these groups. Lastly, Chapter 4 describes prediction of drug sensitivity in cancer cell lines based on pathway-level activity. We use a penalized regression model informed by biological plausibility to predict response to cancer treatments, and construct networks to show relationships between these drugs.

Chapter 2 deals with the problem of mechanism-of-action inference. Small perturbations to a cell may propagate throughout a gene or protein interaction network, and can have wide-ranging downstream effects. Given a snapshot of cellular activity, it can be difficult to tell where a disturbance originated. As a further challenge, scientists often collect multiple forms of data in order to study a phenomenon from all angles, or to get the most information out of a limited number of samples. While additional data can provide richer detail, statistical methods and models available to cope with multiple data types are less well-developed.

We approach the subproblem of perturbation detection by extending the notion of network filtering (Cosgrove et al., 2008) to multi-attribute data. We first We construct a joint

network combining all data types. For a single node in this network (all measurements associated with one gene), we might have gene expression, methylation, and protein abundance. We estimate a joint Gaussian graphical model across multiple data types using block-penalized regression in control data (Kolar et al., 2014) under the assumption that no external perturbations are present. We then use the corresponding estimated covariance matrix to filter for network effects in the case data. To determine the most likely perturbation site, we conduct a series of likelihood ratio tests, conditioning on the existence of a single perturbation. In addition, we present a conditional testing procedure to allow for detection of multiple perturbations. We demonstrate the efficacy of this method through simulation studies, and an analysis of data from The Cancer Genome Atlas (TCGA; Cancer Genome Atlas Network, 2012).

Chapter 3 uses a somewhat different type of network to characterize changes in gene expression at a functional level. Instead of looking for changes at the gene level, where individual measurements may be noisy or relevance poorly understood, we can aggregate genes that have been identified as belonging to functional groups. In addition, we can leverage the relationships between these functional groups through use of a network-informed prior distribution for more accurate inference. Several models have been proposed in this vein, but usually require harsh tradeoffs between predictive capabilities and interpretability of resulting gene groups (Eisen et al., 1998; Troyanskaya et al., 2003; Friedman, 2004).

We propose a biological-function oriented regression framework based on the Gene Ontology (GO; Ashburner et al., 2000), with the goal of characterizing biologically relevant gene groups to cellular phenotypes. The GO provides information regarding the functional role of these genes, both in terms of their particular function and the wider role that they play in cellular regulation. We model the expression of sets of genes involved in biological functions based on the the experimental setting (such as a gene knockouts or limited nutrient access) and phenotypes, within a Bayesian hierarchical formulation. Linear responses are combined according to membership in GO terms, the covariance of which depend upon their relationship in the ontology. We apply this model to analyze cellular

phenotypes in a structured experiment to determine drivers of cellular growth in yeast. Agilent Yeast V2 microarray measurements were taken on 36 CEN.PK derived *S. cerevisiae* chemostat cultures, grown in limiting quantities of glucose, nitrogen, phosphate, sulfur, leucine, and uracil. Brauer et al. (2008) provide experimental details.

We consider two benchmarks for comparison: linear pooling and a hierarchical Dirichlet process. In linear pooling, gene-level coefficients are estimated via simple linear regression and averaged to obtain GO term-level inferences. By contrast, the hierarchical Dirichlet process (Teh et al., 2006) clusters genes into groups based on the data. Similar models have been proposed for this variety of study (Airoldi et al., 2009; Wang and Wang, 2013), and comparison demonstrates the potential gains of a structured Bayesian regression over nonparametric and simplified parametric models. The Bayesian regression framework we have proposed also enables the prediction of expression for unobserved genes with GO annotation. Furthermore, we find that the discrepancy between gene-level and expected gene-level predictions (according to GO term coefficients) correlates well with the ability to predict out-of-sample genes. As such, not only can we make predictions about unobserved genes, but we can determine when these predictions are likely to be reliable.

Finally, we consider methods for predicting drug sensitivity in cancer cell lines in Chapter 4. Several personalized medicine models have been proposed that utilize principal components analysis or factor analysis to find gene expression signatures that correlate well with drug response (see Saeys et al., 2007, for an overview). In practice, these models often overfit the observed data, and may generate gene groups that are uninterpretable from a biological standpoint.

We use results generated by the Adaptive Signature Selection and InteGratioN method (ASSIGN; Shen et al., 2015) to predict drug response. ASSIGN seeks to improve on the factor analysis model by including controlled experimental data. In these experiments, cultured cells are transfected with adenoviruses that cause overexpression of particular genes and begin a cascade of interactions. The gene expression pattern that results may be considered a pathway activation signature. A series of these experiments is performed with

4 genes (*AKT*, *HER2*, *IGF1R*, and *BAD*) to generate 4 signatures. Differential expression from these experiments relative to control is used to inform a prior distribution for the gene signatures.

For the data of interest (such as patient samples), we perform a Bayesian factor analysis using the ASSIGN model to estimate posterior activation probabilities for each sample-pathway combination. We then fit a logistic classifier for response status based on subtype, pathway activations, and 2-way interactions thereof for 82 different drugs, using the penalized regression of Bien et al. (2013). Our analyses suggest that using multiple pathways, and modeling interactions between these pathways in the logistic model to predict drug sensitivity offers better accuracy than single-pathway estimation, in particular for general chemotherapy or DNA drugs. In addition, we construct network representations of the drugs and cell lines under study.

The studies in this dissertation demonstrate the potential benefit of network models in biological applications. For inferring mechanism of action, networks help us to clear away the ripple effects to find the source of an initial perturbation. In the case of ontology-based regression, we look to a different type of network to provide structure in our regression and provide biologically meaningful results. Lastly, looking at pathway-level interactions in cancer cell lines enables better predictions of drug sensitivity. By building models that include network representations of complex biological phenomena, we can make better inferences about mechanism of action, drivers of cellular phenotypes, and drug response.

Chapter 2

Detection of multiple perturbations in multi-omics biological networks

2.1 Introduction

Activity within a cell is governed by a complex set of molecular interactions. In such an intricate system, the introduction of a perturbation to a single element in the network can have widespread effects throughout the system. For mechanism-of-action inference or intervention targeting, it is a critical and difficult task to distinguish the site of the original perturbation from the downstream ripple effects. For example, testing genes one-by-one in an isolated manner, as in differential expression analyses, may be able to identify changes between two states, but the site of the largest change is not necessarily the site of an original disturbance. Our goal is to invert the process by which the effect propagates throughout the network, and identify the site of the initial perturbation to the system.

Previous work demonstrates the importance of considering network effects in analysis of gene expression data. di Bernardo et al. (2005) proposed mode-of-action by network identification (MNI), which used a large microarray compendium to construct a gene interaction network, then "filtered" expression profiles to identify the direct gene targets of each perturbation. Later, Cosgrove et al. (2008) provided a more statistically principled approach, SSEM-Lasso (sparse simultaneous equations model via lasso). This latter method consists of network estimation using lasso estimation, followed by filtering for network effects using the estimated regression parameters. Subsequently, genes are ranked as likely perturba-

tion sites according to the magnitude of their residuals. The theoretical properties of this method are explored in Yang and Kolaczyk (2010). Both of these methods were shown to be capable of providing improved detection of perturbation sites over methods that did not incorporate network structure, such as differential expression analysis. Other researchers consider this problem at the level of pathways rather than individual genes. Pham et al. (2011) build a pathway-level network based on differential expression and KEGG (Kanehisa and Goto, 2000) pathway membership in order to identify pathways of interest. Ma and Zhao (2012) pursue joint modeling in a different way, using drug sensitivity data and gene expression measurements in a Bayesian factor analysis to identify drug targets.

In addition to the difficulty of isolating the primary mover from the vast chain of trailing interactions, the recent trend of data integration introduces further modelling complexity. Researchers often collect measurements of multiple types on a single subject or sample, quantifying phenomena like gene expression, methylation status, and protein abundance. Recent efforts have established that examining a biological phenomenon from multiple 'angles' using multiple types of data can provide important additional mechanistic insight (Bordbar et al., 2012; Zhang et al., 2012; MacNeil et al., 2015). For human studies, multiple types of measurements may be taken in order to get the most information out of a limited pool of subjects.

Though multiple measures are often collected now, the analytic techniques to cope simultaneously with multiple data types are still developing. In many studies, each data type is analyzed separately and then subjected to some joint postprocessing, such as a check for correlation, or annotation for proximity between sets of results (for example, Fournier et al. 2010; Lee et al. 2011; Varambally et al. 2005; Tsavachidou-Fenner et al. 2010). Alternatively, one data type may be used as a discovery data set, while a second is reserved for validation. Analyses of this variety assume that there should be some mirroring of effects between data types, but typically ignore the inherent dependency between biological elements. For instance, the quantity of mRNA transcript is not independent of the abundance of its protein product, nor of its own methylation status. Various methods

exist for inference of potential drug targets (for an overview, see Lecca and Priami 2013 and Csermely et al. 2013), but to our knowledge none have addressed the question of how to jointly model multi-type data while explicitly filtering out effects due to network-based propogation.

In this chapter, we present a strategy for identifying gene-level perturbation sites in multi-type biological data. We construct a joint Gaussian graphical model incorporating all data types. Next, we estimate network structure using a graphical lasso, informed by prior data regarding gene-gene interactions. After then filtering for network effects, we develop a ranking of likely primary perturbation sites based on a series of likelihood ratio tests. We also offer an extension for inference of secondary sites. We demonstrate the efficacy of this methodology in a simulation study, and in an application to joint methylation and gene expression data from The Cancer Genome Atlas (TCGA; Cancer Genome Atlas Network 2012).

2.2 Joint Gaussian graphical model

In defining a framework to model cellular activity, we adopt a gene-centric perspective. Specifically, we match attributes of K different types to form a joint gene-level "node." We then form a graph $G = \{V, E\}$ of gene-wise interactions across these joint nodes. For example, a node may be constructed with a gene's $K = 3$ attributes of expression, methylation status, and protein abundance. Since we expect biologically that cross-gene interactions are relatively rare compared to interactions across measurement types, this joint-node simplification facilitates estimation, reducing the number of potential edges in G from $\frac{pK(pK-1)}{2}$ to $\frac{p(p-1)}{2}$, for p genes.

In more detail, for a single node $i \in \{1, \ldots, p\}$, we have K measurements $Y_i = [Y_i^{(1)}, \ldots, Y_i^{(K)}]^T$. These nodes are are combined into a "stacked" vector Y by node, writing

$Y = [Y_1^{(1)}, Y_1^{(2)}, \ldots, Y_1^{(K)} \ldots, Y_p^{(1)}, Y_p^{(2)}, \ldots, Y_p^{(K)}]^T$. We then specify a conditional Gaus-

sian graphical model, in which each element may be expressed as a linear combination of its neighbors, plus some perturbation μ and error ϵ:

$$y_i^{(k)}|y_{(-i)}, y_i^{(-k)} = \mu_i^{(k)} + \sum_{l \neq k} b_{ii}^{(k,l)} y_i^{(l)} + \sum_{l=1}^{K} \sum_{i \sim j} b_{ij}^{(k,l)} y_j^{(l)} + \epsilon_i^{(k)} \ , \tag{2.1}$$

with $\epsilon_i^{(k)} \sim N(0, \sigma^2)$. The additional term $\mu_i^{(k)}$ represents an external perturbation to $Y_i^{(k)}$ that results in a mean-shift, and is distinct from the effects of i's neighbors. Taking all nodes jointly, we can rewrite the model of Equation (2.1) as

$$Y \sim N((I-B)^{-1}\mu, (I-B)^{-1}\sigma^2) \tag{2.2}$$

$$Y \sim N(\Sigma\mu, \Sigma\sigma^2) \ . \tag{2.3}$$

Derivation of this formulation follows as in Cressie (1993). The matrix B is constructed from coefficients in the conditional formulation, and so an entry $b_{ij}^{(k,l)} = 0$ indicates $y_j^{(l)}$ does not directly influence $y_i^{(k)}$, and results in a zero in the precision matrix $\Omega = \Sigma^{-1}$. The vector of external perturbations μ is believed to be sparse, and our goal will be to identify likely nonzero entries in μ, corresponding to perturbation sites.

In practice, we do not know Σ, and must estimate it from our data. If there are no external perturbations to the network ($\mu = 0$), then we have $Y \sim N(0, \Sigma)$, which allows estimation of Σ. We define a perturbation as occurring relative to a control in case/treated data. We assert $\mu = 0$ holds in the control data, and estimate Σ with control samples only. We will then use $\hat{\Sigma}$ to make inferences about μ in case/treated samples.

As the number of entries in Σ far exceeds the available sample size, we apply a variant on the regularization of Kolar et al. (2014) in estimation of $\hat{\Sigma}$. For precision matrix Ω, we

build a block matrix according to node membership.

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} & \cdots & \Omega_{1p} \\ \Omega_{21} & \Omega_{22} & \cdots & \Omega_{2p} \\ \vdots & & \ddots & \vdots \\ \Omega_{p1} & \Omega_{p2} & \cdots & \Omega_{pp} \end{bmatrix} \tag{2.4}$$

In estimation of $\hat{\Omega}$, we apply a penalty to the Frobenius norms of these submatrices, and optimize according to

$$\hat{\Omega} = \text{argmin}_{\Omega \succ 0} \left(\text{tr}(S\Omega) - \log|\Omega| + \lambda \sum_{a,b} w_{ab}^{-1} \|\Omega_{ab}\|_F \right) \tag{2.5}$$

Penalizing on the level of these submatrices encourages entire $(K \times K)$ blocks in $\hat{\Omega}$ to zero. As previously noted, if submatrix $\Omega_{ab} = 0_{K \times K}$, then nodes a and b are conditionally independent. This type of variable selection procedure is a variant of covariance selection (Dempster, 1972). Further, a zero entry in the covariance matrix $\Sigma = \Omega^{-1}$ further indicates a lack of indirect influence, meaning the nodes are in separate components of the graph G. Building our network this way offers an attractive compromise between allowing interactions across data types and limiting the number of edges that must be estimated. Optimization based on Equation (2.5) proceeds according to approximate block-gradient descent, with details in Kolar et al. (2014). We recommend selection of the tuning parameter λ based on minimum extended Bayesian information criterion with $\gamma = 0.5$ (EBIC; Chen and Chen, 2008), which we have found offers better network recovery than the Bayesian information criterion (BIC) for small sample sizes.

In addition to the block structure, we allow an optional weight to increase the penalty on biologically unlikely edges. In Equation (2.5), w_{ab} represents a plausibility score for between-node interactions. This offers biologically reasonable interactions a lower barrier to entry in the model. Such scores can be constructed using a database such as STRING

(Szklarczyk et al., 2011), as we do in Section 4, or ENCODE (ENCODE Project Consortium, 2004). The weights may also be left at a constant value if insufficient prior information exists for the scenario at hand. This can facilitate estimation of larger networks with relatively few samples.

2.3 Perturbation site identification

2.3.1 Multi-attribute testing procedure

Given an estimate $\hat{\Omega}$, we now proceed to our main problem of interest, i.e., inference on perturbation site in case data, through inference on μ. Cosgrove et al. (2008) introduce the method of using an estimate of the covariance matrix to invert the propagation of network effects, which they called "network filtering." We can extend this concept to multi-type data by using a joint covariance matrix, obtained by the previously outlined method. In order to ascertain which node has been perturbed, we propose the use of node-wise likelihood ratio tests. Note that, as the material that follows in this section and the next do not depend directly on the particular choice of estimator $\hat{\Omega}$ adopted in Section 2, we present our proposed methodology in terms of known Ω (or Σ), and then address the question of how estimation of Ω impacts the overall procedure through a general analysis.

For a given node i, we test the hypothesis that only the entries in μ corresponding to node i (that is, $\mu_i = [\mu_i^{(1)}, \ldots, \mu_i^{(K)}]^T$) are nonzero ($\mu_i \neq 0$, $\mu_{(-i)} = 0$), against the null hypothesis of an entirely zero mean-shift vector ($\mu = 0$). This may be interpreted as a test of whether a particular gene has been perturbed, conditional on it being the only perturbation.

Without loss of generality, we consider a test at the first node, i.e., a test that $\mu_1 \neq 0$. We invert the network propagation and filter the data to obtain $Z = \Omega Y \sim N(\mu, \Omega)$. That is, through 'network filtering' we produce an alternative representation of the data with mean μ, rather than $\Sigma\mu$. In this parametrization, we obtain the maximum likelihood

estimator for μ_1 under the alternative hypothesis as

$$\hat{\mu}_1 \quad = \bar{z}_1 + \Sigma_{11}^{-1}\Sigma_{1\cdot}\bar{z}_{\cdot} \tag{2.6}$$

where $\bar{z}_{\cdot}$ indicates the mean of the filtered data not being presently tested (i.e., $\bar{z}_{(-1)}$), $\Sigma_{\cdot\cdot}$ indicates the corresponding submatrix in Σ, and so on. The resulting likelihood ratio test may be written

$$T_1 \quad = n\left(\bar{z}^T\Sigma\bar{z} - \bar{z}_{\cdot}^T(\Sigma_{\cdot\cdot} - \Sigma_{\cdot 1}\Sigma_{11}^{-1}\Sigma_{1\cdot})\bar{z}_{\cdot}\right) \quad . \tag{2.7}$$

Note that the precision of the filtered data is the covariance of the data on the original scale, Σ. The formula for the conditional precision $Z_{\cdot}$ given Z_1 is $\mathrm{Prec}(Z_{\cdot}|Z_1) = \Sigma_{\cdot\cdot} - \Sigma_{\cdot 1}\Sigma_{11}^{-1}\Sigma_{1\cdot}$. As such, the form of this test statistic is reminiscent of Hotelling's T^2 statistic on the filtered data ($\bar{z}^T\mathrm{Prec}(Z)\bar{z}$), less its portion deriving from the portion of μ that has been assumed-zero ($\bar{z}_{\cdot}^T\mathrm{Prec}(Z_{\cdot}|Z_1)\bar{z}_{\cdot}$). We perform this test for each node in turn, and then rank their likelihood of being the true perturbation site by test statistics $T_1, T_2, \ldots, T_p$.

Under the null hypothesis of $\mu = 0$, $T_j \sim \chi^2_K(0)$ for all j. Under the alternative hypothesis of $\mu \neq 0$, each test statistic T_j has a noncentral chisquare distribution. For example, for $j = 1$, this takes the general form

$$T_1 \quad \sim \quad \chi^2_K\left(\mu^T\begin{pmatrix}\Sigma_{11} & \Sigma_{1\cdot} \\ \Sigma_{\cdot 1} & \Sigma_{\cdot 1}\Sigma_{11}^{-1}\Sigma_{1\cdot}\end{pmatrix}\mu\right) \quad . \tag{2.8}$$

Suppose that the true perturbation is located at the first gene, i.e., that $\mu_1 \neq 0$ and $\mu_{\cdot} = 0$. Comparing T_1 with a test at another node $j \neq 1$, we obtain

$$T_1 \quad \sim \quad \chi^2_K(\mu_1^T\Sigma_{11}\mu_1) \tag{2.9}$$

$$T_j \quad \sim \quad \chi^2_K(\mu_1^T\Sigma_{1j}\Sigma_{jj}^{-1}\Sigma_{j1}\mu_1) \quad . \tag{2.10}$$

Since $\Sigma_{11} - \Sigma_{1j}\Sigma_{jj}^{-1}\Sigma_{j1}$ is positive-definite, $(\mu_1^T\Sigma_{11}\mu_1) > (\mu_1^T\Sigma_{1j}\Sigma_{jj}^{-1}\Sigma_{j1}\mu_1)$, and T_1 stochastically dominates T_j for any node j not containing a true perturbation.

While these derivations are shown here as a node-wise test, this test can be applied to any predefined sets of nodes, of arbitrary size and overlap. In principle, testing could be based on individual elements of μ, or on entire pathways. The test statistics T may not be directly compared if groups of varying sizes are tested, but p-values may be calculated on the basis of the chisquare distribution, with degrees of freedom equal to the total number of nodes in the group being tested.

2.3.2 Sequential multi-target testing

We have so far considered the occurrence of a single perturbation, but this is not always realistic. A treatment may have off-target effects, resulting in multiple interaction sites (Afzal et al., 2014), or a disease may be caused by perturbations to more than one gene. In such a case, interpretation of the previously described results becomes less straightforward. Since each of our previously described tests assumes that all other nodes have zero mean, we automatically perceive nodes *near* the truly perturbed node to be likely sites, so a near-target effect may be confused with a distinct, off-target effect. Once we have identified a primary perturbation site, we may wish to consider the most likely site for a secondary perturbation, in a manner that accounts for the location of the first.

Nested likelihood ratio tests provide a natural framework for a sequential ranking. At step $s+1$, we denote the sites already identified in steps $1, \ldots, s$ as a set S. Having already determined that that the subvector μ_S of μ contains nonzero entries, we can conduct a likelihood ratio test on the remaining nodes to search for additional perturbations. Thus, at step $s+1$, for node i, we test the hypothesis that an additional perturbation is located at node i ($\mu_i \neq 0, \mu_S \neq 0, \mu_{-(S,i)} = 0$) against the null that no perturbations outside of S exist ($\mu_S \neq 0, \mu_{-(S)} = 0$). We perform this calculation for all nodes i not determined to be perturbation sites in steps $1, \ldots, s$.

The resulting test statistic $T_i^{[s+1]}$ may be written as a difference of unadjusted likelihood

ratio test statistics:

$$T_i^{[s+1]} = T_{(i,S)} - T_S \ , \tag{2.11}$$

where T_S corresponds to testing $\mu_S \neq 0$, $\mu_{-(S)} = 0$ against $\mu = 0$, and $T_{(i,S)}$ corresponds to testing $\mu_i \neq 0$, $\mu_S \neq 0$, $\mu_{-(i,S)} = 0$ against $\mu = 0$, Inference can proceed on the conditional sequence, or p-values can be calculated and adjusted to maintain an appropriate false discovery rate across s using the method of Benjamini and Yekutieli (2001).

The magnitude and direction of the difference between this value and the original test statistic depends upon the correlation between the node currently being tested and the nodes already "found" by the sequential procedure. Theorems 2.1 and 2.2 establish some properties relevant to the relative ranking of the adjusted test statistics.

Theorem 2.1 *Given a set of nodes already found to have nonzero mean in steps* $1, \ldots s$*, consider testing for a perturbation at an additional node* i *in step* $s+1$*. Denote the indices in* $Z = \Omega Y$ *corresponding to the nodes found in steps* $1, \ldots, s$ *as* S*.*

We can write the expected difference between the original test statistic and the test statistic adjusted for perturbations in S *as*

$$E(T_i - T_i^{[s+1]}) = \mu_i^T \left(\Sigma_{i,S}\Sigma_{S,i}\right) \mu_i + 2\mu_i^T \left(\Sigma_{i,S}\right) \mu_S + \mu_S^T \left(\Sigma_{S,i}\Sigma_{i,S}\right) \mu_S \ .$$

In the special case that $\mu_i = 0$*,*

$$E(T_i - T_i^{[s+1]} | \mu_i = 0) \geq 0 \ .$$

As such, if no perturbation is truly present at node i, we expect its adjusted test statistic to be no larger than the unadjusted statistic.

Theorem 2.2 *Under the same conditions outlined in the general case of Theorem 2.1, if*

$\Sigma_{S,i} = 0$, *then*

$$T_i^{[s+1]} = T_i \ .$$

The proofs of Theorems 2.1 and 2.2 are given in Appendix A.2. Taken together, these facts give us insight into the way that secondary targets are identified. Suppose we test for secondary perturbations at nodes i and j after finding an initial set of nodes S. When i and S are not connected in our graph, the sequential test statistic for i is the same as the unadjusted statistic. Simultaneously, a correlation between measurements on j and S removes the near-target effects due to proximity to S, resulting in an expected decrease in $T_j^{[s+1]}$ compared to T_j by $\mu_S^T \Sigma_{S,j} \Sigma_{j,S} \mu_S$. Since at any step s we are concerned with relative ranking of test statistics, the decreased $T_j^{[s+1]}$ relative to $T_i^{[s+1}$ makes i a better candidate for an additional perturbation than it was previously. Accordingly, this procedure has the largest potential benefit when the two perturbations are completely separated in the graph.

For an illustration, see Figure 2.1. This simple network of $n = 100$ samples has only $p = 3$ nodes, each with $K = 2$ attributes, and a single edge between nodes 1 and 2. In Ω, we set the within-node partial correlation ρ_{in}, to 0.8 and the between-node partial correlation ρ_{out} to 0.2. In Figure 2.1(a), only a single perturbation is present, at node 1, with signal-to-noise ratio (the value of the perturbation size of μ relative to a diagonal element of Ω) SNR = 1. Node 1 is ranked as the most likely perturbation site, followed by node 2.

This is desirable behavior in 2.1(a) – if we know that only one perturbation exists, then node 2 is the next-best choice. In 2.1(b,c), we add a second perturbation at node 3 with a weaker signal (SNR = 0.25). According to the initial multi-attribute network filtering (NF) ranking shown in 2.1(b), node 2 is the runner-up due to its proximity to node 1. However, if we condition on the presence of a perturbation at node 1 as in 2.1(c), then node 3 is considered a more likely site for a *second* perturbation than node 2.

Performance of the sequential procedure is discussed in Section 2.4.2.

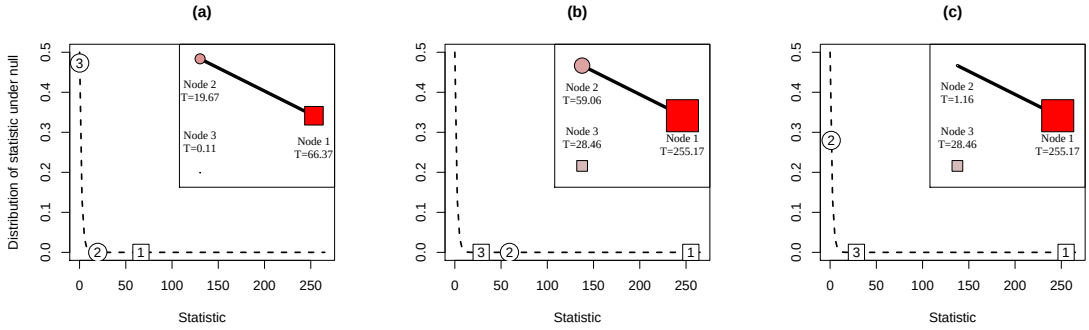


Figure 2.1: A toy example illustrating the properties of the multi-attribute NF in a 3-node network. Perturbed nodes are shown as squares, and node area is representative of test statistic size. Nodes 1 and 2 are neighbors. *(a)* Node 1 is perturbed. As a neighbor to the perturbed node, 2 is identified as the second most likely site for a perturbation if only one exists. *(b)* Nodes 1 and 3 are perturbed, and multi-attribute network filtering (NF) is applied. Node 2 is identified as the second most likely perturbation site because of the shared edge with node 1. *(c)* As in (b), nodes 1 and 3 are perturbed, but the sequential NF procedure is applied. After conditioning on node 1, node 3 is identified as the most likely site for a second perturbation.

2.3.3 Accuracy

We have described our proposed procedure for detecting multiple perturbation sites in multi-omics data as if the precision Ω (or covariance Σ) were known. In practice, of course, to expect exact knowledge of Ω is unrealistic. Firstly, error in estimation may occur. In addition, we take the network estimated in the control data to be representative of the network in the case/treated data, but if the network itself is dysregulated, this may not be an appropriate assumption. While a detailed practical examination of these various sources of errors and their impact on our procedure is beyond the scope of this chapter, we provide here a general characterization result.

Without loss of generality, let $\sigma^2 = 1$ and consider the case of T_j for $j = 1$. Let $\tilde{\Omega} = \Omega + \Delta$ be an erroneous version of the true Ω, and denote by $\tilde{T}_1$ the corresponding version of T_1 resulting from using $\tilde{\Omega}$ in place of Ω. Our interest will be on the distribution

of the discrepancy $T_1 - \tilde{T}_1$. Towards that end, we define the $K \times K$ matrix

$$D = \Omega_{11} - \Omega_{1\cdot}\Omega_{\cdot\cdot}^{-1}\Omega_{\cdot 1} - \left(\tilde{\Omega}_{11} - \tilde{\Omega}_{1\cdot}\tilde{\Omega}_{\cdot\cdot}^{-1}\tilde{\Omega}_{\cdot 1}\right) \ .$$

Assume Σ_{11} is positive definite. For the product $D\Sigma_{11}$, express its spectral decomposition as

$$D\Sigma_{11} = \sum_{k=1}^{s} a_k E_k \ ,$$

such that $\text{rank}(E_k) = r_k$ (corresponding to the multiplicity of the eigenvalue a_k) and $\sum_{k=1}^{s} r_k = K$.

We then have the following result.

Theorem 2.3 *Under the conditions above, the discrepancy $T_1 - \tilde{T}_1$ is equal in distribution to a linear combination of mutually independent, noncentral chisquare random variables,*

$$\sum_{k=1}^{s} a_k \chi^2_{r_k}(\delta_k) \ , \tag{2.12}$$

where

$$\delta_k = (n/2)\mu^T \Sigma_{\cdot 1} E_k \Sigma_{11}^{-1} \Sigma_{1\cdot} \mu \ .$$

Accordingly,

$$E\left[T_1 - \tilde{T}_1\right] = tr\left(D\Sigma_{11}\right) + \frac{n}{2}\mu^T \Sigma_{\cdot 1} D \Sigma_{1\cdot} \mu \tag{2.13}$$

and

$$Var\left(T_1 - \tilde{T}_1\right) = 2tr\left((D\Sigma_{11})^2\right) + 2n\mu^T \Sigma_{\cdot 1} D \Sigma_{11} D \Sigma_{1\cdot} \mu \ . \tag{2.14}$$

The proof of this theorem is given in Section A.3. The distributional result follows from application of Baldessari (1967) Theorem 1, while the moment results follow from definition of first second and moments of noncentral chisquare random variables. In the case that Σ_{11} is not positive definite, more general results in Tan (1977) may be used, at the cost of additional notation and conditions.

Note that D in our results above, as a function of $\Delta = \tilde{\Omega} - \Omega$, plays the key role of capturing the impact of the discrepancy between Ω and $\tilde{\Omega}$. A more relaxed – but arguably more informative – statement of our moment results is the following, wherein the role of Δ is made explicit.

Corollary 2.1 *Let* $||\cdot||_2$ *denote the spectral norm. Then*

$$E\left[T_1 - \tilde{T}_1\right] = O\left(||\Delta||_2\right) \quad \textit{and} \quad \textit{Var}\left(T_1 - \tilde{T}_1\right) = O\left(||\Delta||_2^2\right) \ .$$

Hence, we see that for a given discrepancy Δ between the true Ω and the value $\tilde{\Omega}$, the expected level of discrepancy between the corresponding statistics T_1 and $\tilde{T}_1$, as well as the standard deviation, are both of magnitude on the order of the spectral norm of Δ. Proof of the corollary may also be found in Section A.3.

2.4 Simulation

2.4.1 Single-target simulations

We want to consider two aspects of potential performance gains: (1) conducting a network-aware analysis method, and (2) using multiple data sources. To our knowledge, no other method has yet been proposed for joint modeling and detection of perturbations in this multi-attribute setting. As such, we conduct comparisons in simulation against established methods for single-type data, and a naïve extension of these methods to accommodate multi-type data. To assess gains from network analysis, we compare our method with simple differential expression (t-tests for single-attribute data, and Hotelling's T^2 for multi-attribute). To examine the benefit from considering multiple data sources, we consider the improvement obtained from using $K = 2$ sources, versus a single data type. We also perform SSEM-Lasso (Cosgrove et al., 2008) for the single-attribute case.

We simulate data across a range of network conditions, varying the strength of associations between data types and nodes. We construct a network of $p = 20$ nodes according to a

stochastic block model (Holland et al., 1983), with $n = 50$ cases and controls. The network is divided into two groups of nodes, where cross-block connections are more likely to occur within a block (probability $\theta_{within} = 0.4$) than between blocks (probability $\theta_{across} = 0.2$). Network links are assigned $-\rho_{out}$ in the precision matrix.

For each node with $K = 2$ attributes, we first assign all within-node correlations the value $-\rho_{in}$ in the precision matrix, creating a block-structure along the diagonal. A small value is added to the diagonal of Ω until the minimum eigenvalue is at least 0.5 to ensure invertibility, then the precision matrix is scaled to have diagonal 1. For each network constructed, for node i to be perturbed means that a mean-shift μ_i is applied to its elements. We simulate null data from $N(0, \Sigma)$ and perturbed data with one nonzero node in μ from $N(\Sigma\mu, \Sigma)$, and perform the aforementioned estimation and testing procedure.

From the likelihood ratio tests, we obtain a ranked list of nodes, with our truly perturbed node sitting at rank r. For each of 100 simulated networks, we perturb each of the $p = 20$ nodes in turn and observe their rank according to the multi-attribute network filtering (NF) procedure. We average over the proportion of sites occurring in our ranked list and construct receiver-operator characteristic (ROC) curves. These curves can be directly related to an empirical CDF, with positions along the x-axis indicating the proportion of total sites in a top k list. The y-axis, then, indicates the probability that the true perturbation site was included in that list of k sites. Results for single-perturbation simulations are shown in Figure 2.2. In addition, the probability that the top-ranked site correctly identifies the perturbation is shown in Table 2.1.

Across a range of partial correlations, multi-attribute network filtering (NF) has most successful recovery of the perturbed site with respect to AUC and the probability of selecting the true perturbation as the top-ranked site (an "ideal detection"). Multi-attribute NF is followed by its single-attribute counterpart and SSEM-lasso. Hotelling's T^2 follows, narrowly but consistently outperforming standard differential expression on a single attribute. Under all correlation settings considered here, the multi-attribute modeling strategy identifies the site correctly more than half of the time. On average, such ideal detections are

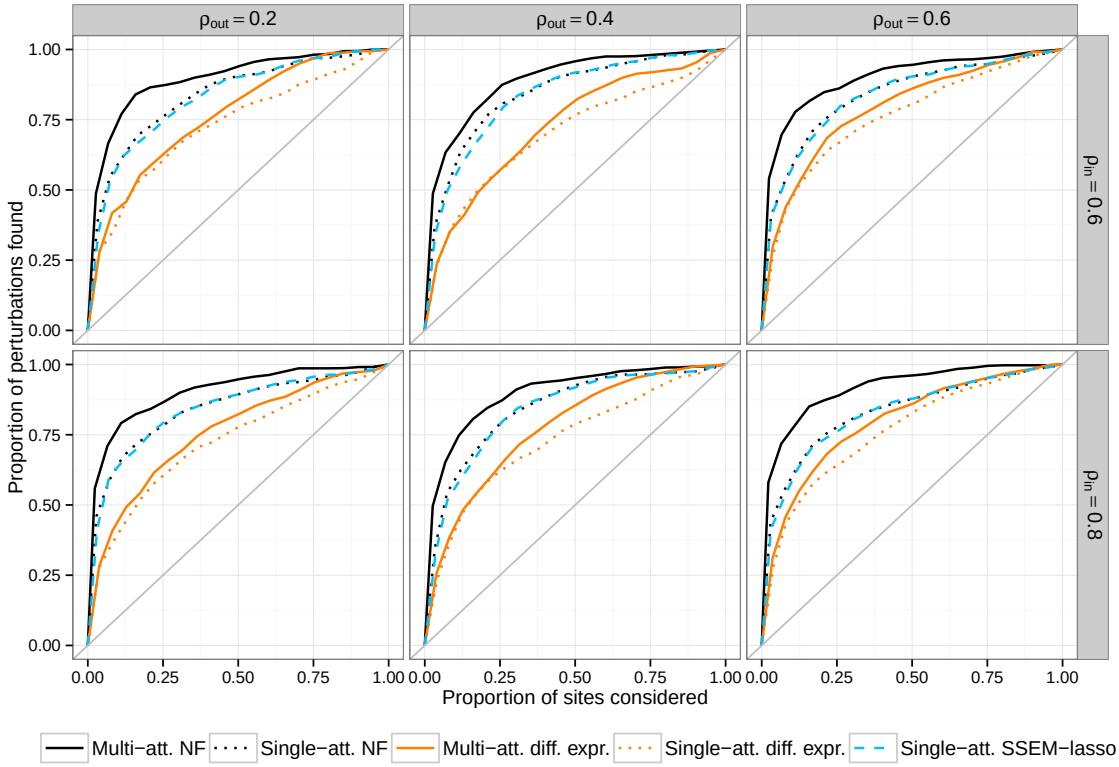


Figure 2.2: Single-site recovery from a stochastic block model simulation with $p = 20$ nodes, $n = 50$ cases and controls, and SNR $= 0.20$. Along the x-axis, we consider the proportion of all sites in a top k list, and along the y-axis, the probability that the truly perturbed site is contained within that top k list. In each plot, the jump at the leftmost edge of the graph corresponds to the probability of identifying the true perturbation as the highest-ranked site (values in Table 2.1).

.

made 54.0% of the time for multi-attribute NF, 42.8% for its single-attribute counterpart. By contrast, differential expression ranks the truly perturbed site first only 27.0% of the time using either method. SSEM-lasso with a single attribute identifies the true perturbation first 39.3% of the time, despite a comparable AUC to the single-attribute NF method, as shown in Table 2.1.

Table 2.1: Probability that the top-ranked site is the true perturbation site and (AUC) for simulations shown in Figure 2.2. ρ_{in} indicates the strength of within-node partial correlation, and ρ_{out} of cross-node partial correlations.

		NF methods		Differential expression		SSEM-lasso
ρ_{in}	ρ_{out}	Multi-att.	Single-att.	Multi-att.	Single-att.	Single-att.
0.8	0.2	0.56 (0.90)	0.46 (0.84)	0.28 (0.76)	0.28 (0.72)	0.41 (0.84)
	0.4	0.50 (0.90)	0.35 (0.84)	0.26 (0.77)	0.23 (0.73)	0.32 (0.84)
	0.6	0.58 (0.92)	0.44 (0.83)	0.31 (0.80)	0.28 (0.76)	0.41 (0.83)
0.6	0.2	0.49 (0.89)	0.39 (0.84)	0.28 (0.76)	0.28 (0.73)	0.34 (0.83)
	0.4	0.49 (0.89)	0.37 (0.84)	0.24 (0.73)	0.23 (0.70)	0.34 (0.84)
	0.6	0.54 (0.90)	0.41 (0.84)	0.30 (0.79)	0.27 (0.76)	0.40 (0.83)

2.4.2 Multi-target simulations

We also wish to evaluate the performance of the sequential procedure when multiple perturbations are present. As previously noted, any advantage over simply taking the initial rankings will depend upon the network structure and the distance between perturbations. If two perturbations occur adjacent to one another, the near-target and off-target effects will be aligned, and the ranking will not be substantively changed. However, if the perturbations are far apart in the graph, this procedure may substantially improve the chances of detecting both effects.

We extend our previous simulations study to include a second perturbation. In the context of a stochastic block model, we simulate two perturbations: a nonzero node in the first block with SNR = 0.20 as before, and a second, weaker perturbation in the second block with SNR = 0.10. We then vary the probability of a cross-block edge (θ_{across}) relative to the probability of an edge within each block (θ_{within}) to demonstrate the role of distance on the graph in the efficacy of the sequential procedure. We consider $\theta_{across}/\theta_{within} = 0.25$ (slight separation), 0.125 (moderate separation), and 0 (complete separation). Table 2.2 shows the probability of ranking both true perturbations in the top two sites, and Figure 2.3 shows the ROC curves for identifying both perturbations. The sequential procedure outperforms the initial ranking on both counts for cases shown, with gains increasing according

to separation between the perturbations for probability of ideal identification.

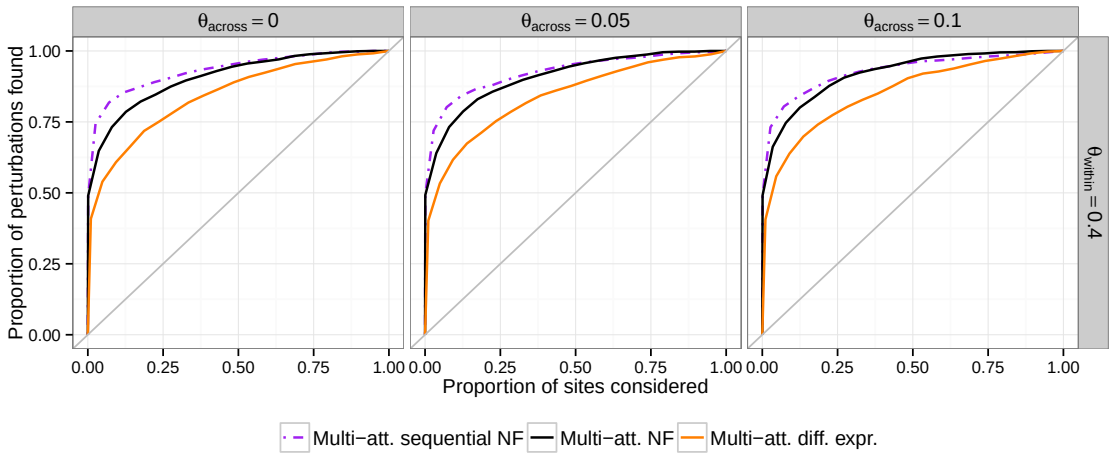


Figure 2.3: Simulations showing improvement of the sequentially restricted NF procedure versus the standard multi-atrribute NF and Hotelling's T^2 ranking when two perturbations are present, located in different blocks in a stochastic block model. The expected distance between these two perturbations are on the graph is determined by $\theta_{across} = (0, 0.05, 0.1)$, corresponding to complete, moderate, and slight separation between the two blocks, relative to the within-block edge probability of 0.4. Benefits from the sequential procedure are largest when the two perturbations are not connected in the graph (left).

Table 2.2: Probability of identifying the both truly perturbed sites in the first two ranked positions and (AUC), considering only multi-attribute methods. Corresponding plots are shown in Figure 2.3.

$\theta_{across}/\theta_{within}$	Sequential multi-att. NF	Multi-att. NF	Multi-att. diff. expr.
0.250	0.74 (0.93)	0.67 (0.92)	0.57 (0.85)
0.125	0.73 (0.93)	0.65 (0.91)	0.54 (0.84)
0.000	0.76 (0.93)	0.66 (0.91)	0.55 (0.84)

In certain circumstances, the sequential procedure may produce suboptimal results. For example, suppose that the first identification is a false positive due to proximity to a true perturbation. The truly perturbed site will have a lower ranking after conditioning for the false positive site, as this procedure would adjust away some of that node's own signal. This is particularly likely to occur when signal-to-noise ratio is low, or when multiple perturbations have common neighbors. As such, we recommend the use of this procedure when an

unambiguous initial identification has been made, and suspected secondary perturbations are not in close proximity to the initial site.

2.4.3 Comparison to post-analysis aggregation

While the multi-attribute NF method provides improved perturbation site detection over single-attribute methods and multivariate differential expression, we wish to consider how much is gained by considering cross-attribute relationships, as opposed to some comparatively simpler 'aggregation' of single-attribute results. This benchmark is of particular interest given the popularity of network recovery methods by Guo et al. (2011) and Danaher et al. (2013) for simultaneous inference of multiple, related networks across data types, but without cross-type interactions. Following the same simulation strategy as described in Section 2.4.1, we consider the performance of a "separated" ranking procedure, in which we estimate and filter for separate networks for each data type, then combine results into a block-precision matrix to rank individual biological attributes, setting cross-type entries to zero. This amounts to asserting independence between each data type. Results are shown in Figure 2.4, and Table 2.3. Note that for the separated procedure, we look for the probability that both attributes of the perturbed node are ranked highly.

Table 2.3: Probability that the top-ranked sites are the truly perturbed gene and (AUC) for simulations shown in Figure 2.4. These simulations feature a single perturbation.

		NF methods		
ρ_{in}	ρ_{out}	Multi-att.	Separated	Single-att.
0.8	0.2	0.55 (0.90)	0.44 (0.86)	0.41 (0.84)
	0.4	0.56 (0.90)	0.40 (0.84)	0.41 (0.84)
	0.6	0.46 (0.92)	0.37 (0.85)	0.42 (0.83)
0.6	0.2	0.57 (0.89)	0.44 (0.85)	0.48 (0.84)
	0.4	0.55 (0.89)	0.41 (0.83)	0.44 (0.84)
	0.6	0.55 (0.90)	0.38 (0.84)	0.41 (0.84)

The multi-attribute NF performs best in terms of AUC and the probability of ideal identification. Separated and single-attribute methods perform comparably to each other

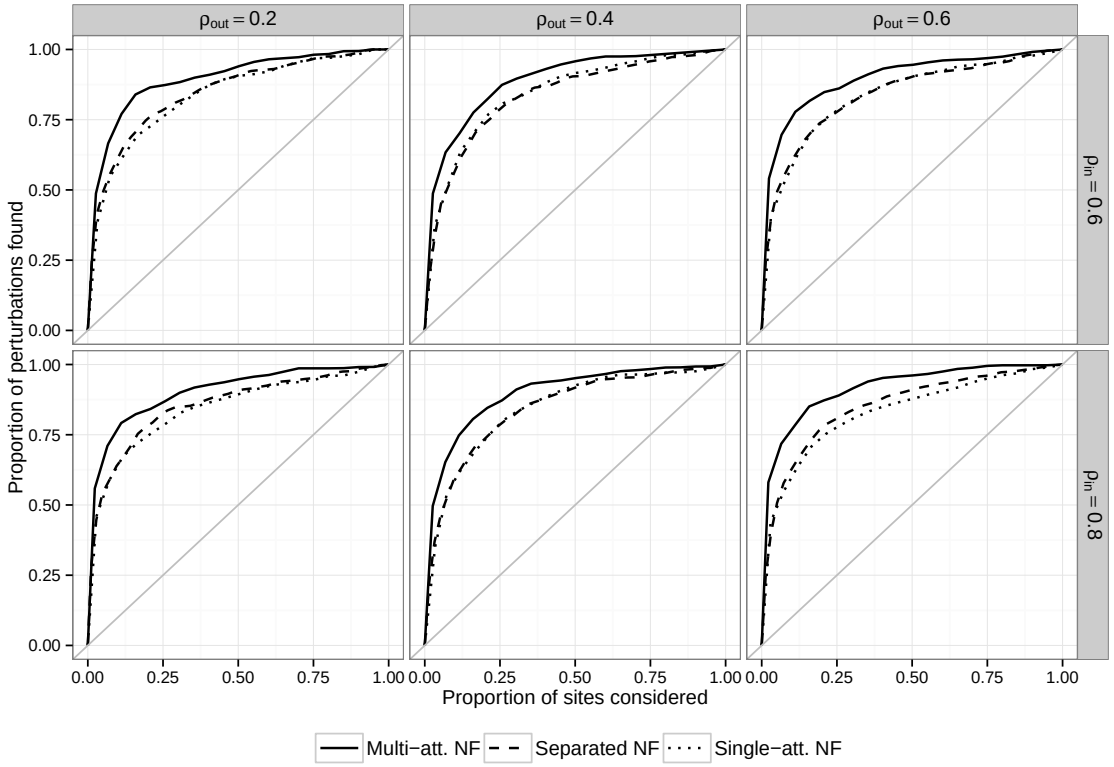


Figure 2.4: Comparison of network filtering methods in a single-perturbation setting. ROC curves show perturbation site recovery from a stochastic block model simulation scheme with $p = 20$ nodes, $n = 50$ cases and controls, and SNR $= 0.20$. "Separated NF" indicates that the network estimation and filtering procedures were performed in isolation on each data type and then combined for ranking.

by both of these metrics. This also holds if we rank according to the first appearance of a gene's measurements, rather than requiring top ranks for both. Given that a slightly higher burden is imposed on the separated method than the single-attribute (two attributes must be ranked highly rather than one), this is a slight advantage to the separated method over analysis of a single attribute. Nevertheless, our results indicate that most benefits attained from this type of data integration emerge from consideration of interaction *between* attributes when such interactions are present in the underlying data. The design of our model specifically exploits the existence of cross-type interactions, and is able to better discover perturbation sites as a result.

2.5 Analysis of TCGA breast cancer data

We apply this methodology in an analysis of breast cancer data from The Cancer Genome Atlas (TCGA). We have gene expression and methylation data obtained from tumor samples of 60 patients with metastatic cancer and and 569 with nonmetastatic cancer. Both the expression and methylation data were downloaded as Level 3 normalized data, and then processed to achieve approximately Gaussian distributions. RNA-seq data was preprocessed by TCGA using RSEM (RNASeq by Expectation Maximization; Li and Dewey, 2011) and MapSplice (Wang et al., 2010). Transcripts per million (TPM) were then transformed via quantile normalization on $\log_2(\text{TPM}+1)$. The 450k methylation array data was preprocessed by TCGA using the ratio of the intensity of methylated probes to the total probe intensity to produce β values (Du et al., 2010). We then transformed these values according to $\log_2\left(\frac{\beta}{1-\beta}\right)$. For our analysis, we extracted measurements from 133 genes belonging to the *WNT* signaling pathway in KEGG (Kanehisa and Goto, 2000) from samples with both transcript and methylation data. If more than one measurement was present per gene attribute (multiple methylation sites or transcript segments), a 90% trimmed mean was taken. Subjects were considered to have metastatic cancer if classified as such at baseline or at any subsequent follow-up. Details of the data processing may be found in Appendix A.

We first estimate the block-precision matrix of the network using the $n = 569$ tumor samples from nonmetastatic cancers. Using our estimated precision matrix $\hat{\Omega}$, we filter for network effects in the data from $n = 60$ metastatic cases, and perform gene-wise likelihood ratio tests in order to ascertain which gene is the most perturbation candidate.

The top-ranked sites are shown in Table 2.4. The highest-ranked site is *PP3CC* ($T = 14.35$), which has previously been implicated in prostate cancer (Hornstein et al., 2008), though it does not achieve group-wise significance (raw $p = 0.00076$). A drop-off in the test statistic is visible after the 4th position (for *WNT11*, $T = 8.69$ while the next gene *WNT10A* has $T = 7.61$). This difference is visible in the top panel of Figure 2.5. As such,

we consider the top 4 genes in Table 2.4 to be the most plausible primary perturbation sites.

For additional verification of our results, we perform cross-validation to assess the predictive accuracy of the mean vector implied by each gene ranking. We divide metastatic case data into 10 groups of approximately equal size. For each fold f, we use 90% of the data to estimate $\hat{\mu}_{1,f}, \ldots, \hat{\mu}_{p,f}$ according to Equation 2.6. We then predict the mean of Y^{test} for each gene j by taking $\hat{\Sigma}[0, \hat{\mu}_{j,f}^{train}, 0]^T$. Through this method we obtain mean-squared-error

$$\text{MSE}_{j,f} = \frac{1}{\sum_i I(i \in f)} \sum_{i \in f} (Y_i^{test} - \hat{\Sigma}[0, \hat{\mu}_{j,f}^{train}, 0]^T)^2 \tag{2.15}$$

for each gene under each fold. We take the average of these errors to obtain a ranking of predictive ability by cross-validation, with smallest MSE indicating the best accuracy. Rankings obtained by this cross-validation procedure show agreement with rankings from multi-attribute NF for the top-ranked site (Figure 2.5, bottom panel).

Table 2.4: Top-ranked genes from multi-attribute NF analysis of TCGA methylation and gene expression data. The top 5 genes for each method are included.

	Multi-att NF		Sequential NF		Diff. expr.		Cross-val.	
Gene	Statistic	Rank	Statistic	Rank	Statistic	Rank	MSE	Rank
PPP3CC	14.36	1	14.36	1	8.01	3	264.99	1
WNT7B	10.52	2	9.32	4	4.45	8	272.43	132
PRKACB	9.29	3	9.28	5	3.34	21	266.98	40
WNT11	8.69	4	10.47	3	4.43	9	275.74	133
WNT10A	7.62	5	7.77	7	2.14	34	267.69	121
NFATC2	7.59	6	12.34	2	7.85	4	270.39	128
SERPINF1	4.40	22	4.31	26	1.30	62	265.21	2
INVS	3.63	32	3.66	32	0.39	105	265.43	4
LRP5	3.33	36	3.24	39	0.77	90	265.40	3
FZD9	2.08	51	3.29	37	7.45	5	271.59	131
FBXW11	1.93	52	2.15	52	0.46	101	265.53	5
WNT10B	0.03	85	0.03	86	8.25	2	266.79	32
TCF7L2	0.03	87	0.03	88	12.31	1	266.90	34

We also show results from a joint differential expression analysis using Hotelling's T^2

test in Figure 2.5 and Table 2.4. While *PPP3CC* and the other top 6 multi-attribute NF results are also ranked highly in differential expression results, some genes such as *TCF7L2* show strong differential expression without strong network-filtered evidence. It is plausible that differentially expressed genes are identified because of an accumulation of network effects, rather than an external pressure applied to the system. Similarly, perturbation sites may not necessarily exhibit differential expression; network effects may compensate to restore the perturbed gene to normal levels.

Considering the possibility of multiple perturbations, we also performed the sequential multi-attribute NF procedure as described in Section 2.3.2. At each step, the node with the largest test statistic in the previous step is conditioned on as a nonzero portion of the mean vector, and testing is performed to ascertain whether additional nodes are nonzero. In the second panel of Figure 2.5, we see that after adjusting for the first perturbation at *PPP3CC*, the most plausible site for a secondary perturbation is at *NFATC2*, which was ranked the sixth most plausble primary perturbation site. *NFATC2* has been implicated in breast cancer cell invasion in a previous study by Yiu and Toker (2006).

2.6 Remarks

The multi-attribute network filtering methodology does suffer from some limitations. It relies upon the assumption that the network structure encoded in Ω does not vary between the control data and the case data. As such, this method is likely best suited to experimental settings in which it may be plausible to believe under investigator-limited perturbations that the underlying network relationships are fairly similar between case and control settings.

The framework here also depends upon multivariate Gaussian distributions for all data types. An extension of this network filtering procedure to non-Gaussian distributions would enable inclusion of additional phenotypes, such as SNP and CNV data. This extension has not been undertaken even in the univariate case thus far, but semi-parametric copula

methods (such as those by Liu et al., 2012) show promise for the network estimation portion of this problem.

As is always a concern with large network models, computational costs in estimation of Ω may be prohibitive. This is particularly the case in recovery of large, densely connected networks. As noted by Kolar et al. (2014), the block gradient descent algorithm employed here performs most efficiently when the graph can be separated into smaller connected components (as a rough guide, we recommend use of this algorithm when the largest connected component has fewer than 200 joint nodes). If estimation of the block-precision matrix is infeasible, use of a separated estimation procedure with network filtering, such as the joint graphical lasso (Guo et al., 2011; Danaher et al., 2013), may still be employed. This is expected to yield a large performance improvement over differential expression procedures, and potentially a smaller additional improvement over an analysis of a single attribute.

Our work shows that if cross-attribute interactions are present in the data, benefits from data integration are strongest when these interactions are explicitly modeled. Across all tested network settings, the multi-attribute NF procedure provides better detection of perturbation sites than any single-attribute method, or multi-attribute method that ignored the network structure. In addition, we found that there were substantial gains to be had from a network-filtering based ranking on a single attribute alone compared with differential expression– it easily outperformed Hotelling's T^2 statistic, and provided a greater chance of an ideal identification than SSEM-lasso. The results in this chapter underscore the need to take network effects into account when working with bioinformatic data, and offers a statistically principled method for a truly integrative analysis of multi-attribute data for better understanding cellular mechanism-of-action.

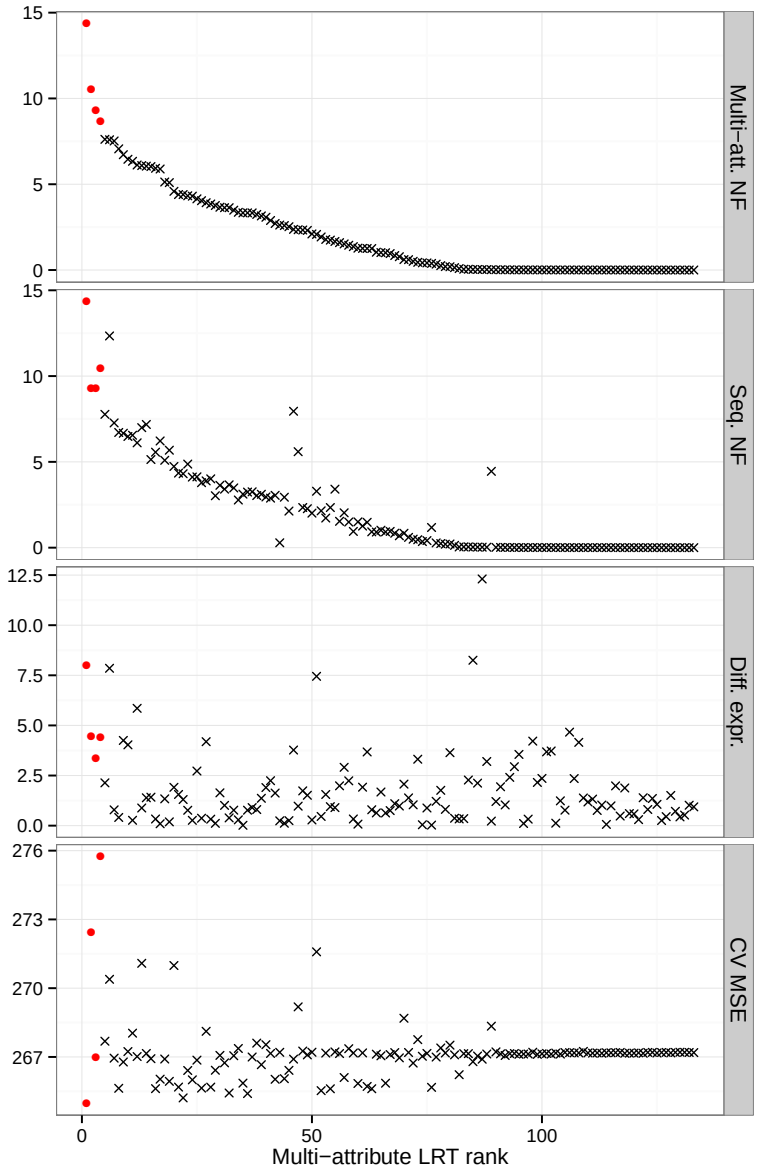


Figure 2.5: Results from an analysis of data from TCGA. Rank according to the non-sequential multi-attribute NF ranking is shown along the x-axis for all plots. Panels show NF statistic, differential expression statistic, and cross-validation MSE. The top 4 results shown in Table 2.4 are highlighted in red.

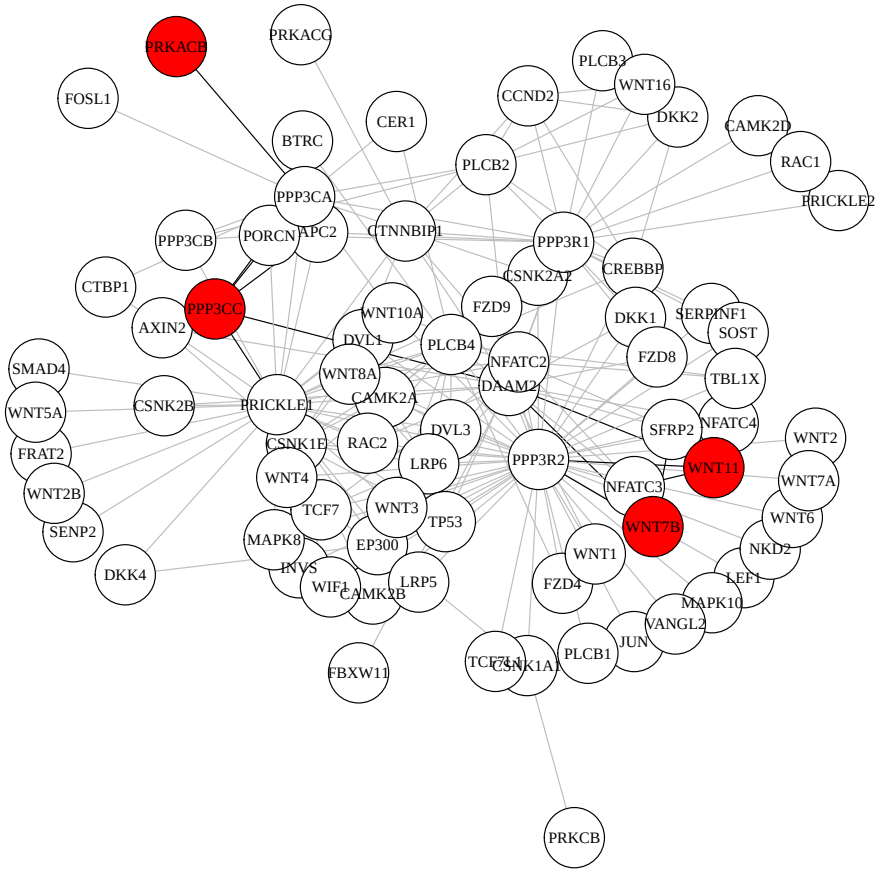


Figure 2.6: A graph showing the connected subgraph of TCGA genes. The top 4 genes shown in Table 2.4 are highlighted in red.

Chapter 3

Characterizing cellular phenotypes via Bayesian regression in the Gene Ontology

3.1 Introduction

A fundamental task in many bioinformatic studies is the explanation of phenotypic variance by differences in gene expression. Many approaches exist for this purpose, among them regression and differential expression approaches, clustering methods, and database-informed analyses. These vary in complexity from sophisticated machine learning models to simple sets of gene-by-gene linear regressions. A principal challenge in analyses of this type is finding an appropriate balance between interpretability and and predictive power.

Gene-by-gene analyses often fail to produce results that are either generalizable or biologically interpretable. A typical differential expression analysis may result in a list of genes that are significantly different between case and control, but it may be unclear what underlying mechanism at work. Some methods, such as GSEA (Subramanian et al., 2005) conduct group-based "enrichment" analyses. These strategies group genes according to various databases, and then analyze whether the significant genes in those group are jointly over- or under-expressed. These methods offer improved interpretability over single-gene and cluster-based analyses, and may offer power improvements in certain circumstances. Still, they typically do not capture the relationships that exist between these gene groups, aside from genes which share membership (and sometimes not even then).

On the other end of the spectrum, clustering-based methods allow for data-driven

groupings of genes. In recent years, methods as those described by Eisen et al. (1998), Shamir and Sharan (2001), and Jiang et al. (2004) have gained in popularity. Cluster-based models typically achieve good model fit according to residuals, but result in gene sets that are not interpretable without significant biological investigation and expertise. In addition, these models are completely unable to make predictions about genes not observed in the training set, as they offer no information as to the cluster assignment of new genes.

In this paper, we propose ontological regression, which aims to strike a balance between the interpretability of database-organized methods and the flexibility of clustering models. We perform a Bayesian hierarchical regression informed by the annotations and structure of the Gene Ontology (GO). The GO aims to be "a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism" (Ashburner et al., 2000). The biological processes domain organizes cellular functions into a directed acyclic graph, where the root of the tree is the general heading of "biological processes" and subsequent branches make finer and finer distinctions between groups of processes. For example, beginning at the root node of biological process (GO:0008150), we can take the branch of cellular process (GO:0009987), then metabolic process (GO:0008152), then biosynthetic processes (GO:0009058), then cellular biosynthetic process (GO:0044249), then membrane lipid biosynthetic process (GO:0046467), and after four more steps, we finally arrive at the leaf node of lipid A biosynthetic process (GO:0009245), which has no child nodes.

Each of these nodes in the ontology is referred to as a "GO term". Genes are annotated with one or more GO terms by experimental or computational evidence, or by curatorial judgement. We use both the gene annotations and the tree structure of the ontology itself in our regression, which models the expression of genes as mixtures of activity realized from distributions at the GO term level. We demonstrate the abilities of this model in a case study using data from a series of experiments studying gene expression in *Saccharomyces cerevisiae* under different nutrient limitations (see Brauer et al., 2008; Airoldi et al., 2009, for experimental details).

3.2 Model

Brauer et al. (2008) describe a series of experiments on *S. cerevisiae* in a chemostat. Yeast cells were restricted in of one of six nutrients (glucose, nitrogen, phosphate, sulfur, leucine, or uracil) to control their growth rate to specific values. Each observation of the data set consists of the restricted nutrient, the growth rate of the cells, and the resulting gene expression. We construct a Bayesian regression to model gene expression by growth rate, incorporating the natural structure of biological processes and the imposed structure of the experiment itself.

As it has previously been found that "expression of more than one quarter of all yeast genes is linearly correlated with growth rate, independent of the limiting nutrient" (Brauer et al., 2008), we begin by assuming a linear relationship between the growth rate X and each gene's observed expression Y per restriction environment f. Next, we aggregate condition-specific gene effects by their GO term annotations through a mapping matrix γ, and model the activity of GO terms. Finally, we model a common prior distribution for the activity of GO terms in each of these restriction environments. A diagram describing our regression model is provided in Figure 3.1.

Structural information from the GO is encoded in two different places in this model: the assignment matrix of genes to GO terms (γ) and in the prior for covariance between GO terms (Σ_{MRCA}). The assignment matrix is a row-normalized mapping of n_g genes by n_t GO terms. That is, $\gamma_{ij} = I(\text{gene } i \text{ in term } j)/\sum_k I(\text{gene } i \text{ in term } k)$. A gene that is annotated with many GO terms will have its influence divided evenly across all such terms, in contrast to an enrichment-style analysis.

Relationships between GO terms are also incorporated into our ontological regression model. In enrichment analyses, all GO terms with gene assignments are typically considered to be candidates for enrichment, and no adjustments are made in consideration of the relationships between GO terms. We encode these relationships as the scale matrix in the prior distribution for the covariance between GO term activity. In the GO space, we

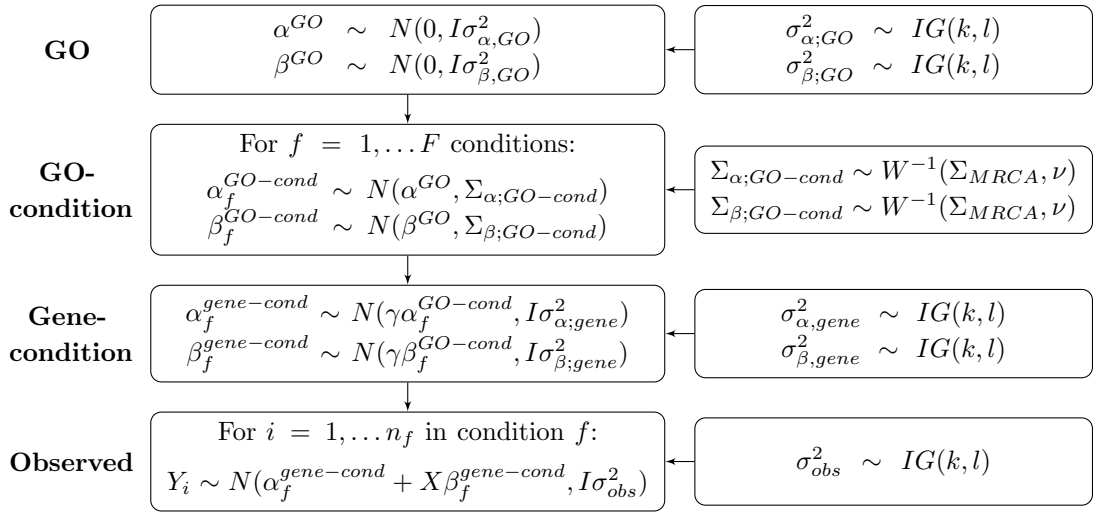


Figure 3.1: Generative model diagram for ontological regression. The model has four levels: global gene ontology terms (GO), gene ontology terms by sample condition (GO-condition), genes by sample environment (Gene-condition), and observed data (Observed). Note that additional covariates beyond X may be included in a parallel manner. The notation $N(\mu, \Sigma)$ indicates a multivariate normal with mean μ and covariance matrix Σ, $IG(k, l)$ indicates an inverse gamma distribution with shape parameter k and scale parameter l, and $W^{-1}(\Sigma, \nu)$ indicates an inverse Wishart distribution with scale matrix Σ and degrees of freedom ν.

posit that the root node has activity level distributed according to $Z_0 \sim N(0, \sigma_0^2)$. At each step down the tree, children j of node k have activity levels distributed $Z_j \sim N(Z_k, \sigma_s^2)$. Given this structure, we can approximate covariance between the activity of GO terms as a function of the distances to their most recent common ancestor (abbreviated MRCA). Denote $A(i, j)$ to be the most recent common ancestor of i and j in the ontology, and $D(l, m)$ the distance between l and m. We can write the first-order approximation as

$$\Sigma_{MRCA(i,j)} = \frac{\sigma_0^2}{\sqrt{D(A(i,j),i)\sigma_s^2 + \sigma_0^2}\sqrt{D(A(i,j),j)\sigma_s^2 + \sigma_0^2}} \ . \tag{3.1}$$

The details of posterior distributions are given in Appendix B.2. All priors are conjugate, and inference is done via Gibbs sampling (for software details, see the link in Appendix B.1).

Since gene-by-condition coefficients have expectations that are a function of their GO-by-condition coefficients and γ, we can calculate expected gene-level coefficients (and thus

predict expression) for genes outside of our training set. This is in contrast to clustering-based methods, which provide no guidance as to group membership for unobserved genes, and list-based analyses, which at best suggest the sign of the coefficient for left-out genes. By the same token, we can propagate errors at the gene level up to GO terms to analyze their functional significance. This means that if we identify which genes we have little ability to predict, and believe that those inadequacies stem from problems at the GO term level, we can identify the GO terms responsible for these bad predictions.

We note that when we have more GO terms than genes ($n_t \geq n_g$), an informative Σ_{MRCA} is required to ensure a positive-definite covariance matrix. The hyperparameters σ_0^2 and σ_s^2 may be manipulated to construct a near-diagonal scale matrix if a weaker prior is preferred.

3.3 Results

3.3.1 Analysis of cell growth experiment

From a single run of the Gibbs sampler, we obtain Monte Carlo estimates of posterior means for each gene-condition pairing, each GO term-condition pairing, and GO terms overall. We compare results against two benchmarks: linear pooling and a hierarchical Dirichlet process (HDP). For linear pooling, we merely average the coefficients from a series of linear regressions by GO term in order to obtain a functional summary of activity. By contrast, the HDP is a nonparametric model that clusters genes according to the data. Again, gene-level coefficients are averaged in order to obtain interpretable biological groups.

The top five GO terms by each method are shown in Table 3.1, ranked by p-value or Bayesian analogue. Note that directions of effect are consistent across all three methods for these GO terms. The GO terms identified by ontological regression largely agree with those identified by HDP, and in some cases by linear pooling. As this method effectively compromises between these two extremes, it is encouraging to see this mixing borne out in the top results. In addition, we note in particular that GO:0050896 (response to stimulus)

and GO:0006950 (response to stress) are identified more readily by ontological regression than the other two. These are two terms that we would certainly expect to appear for the experiments of Brauer et al. (2008).

		Ontoreg		Linear Pooling		HDP	
GOID	Term	Rank	Mean	Rank	Mean	Rank	Mean
GO:0050896	response to stimulus	1	-78.16	6	-93.85	2	-4.47
GO:0006950	response to stress	2	-61.35	9	-77.87	7	-3.30
GO:0006412	translation	3	50.80	2	126.29	3	3.90
GO:0044249	cellular biosynthetic process	4	50.78	21	56.48	1	4.85
GO:0044260	cellular macromolecule metabolic process	5	46.98	63	41.89	4	3.59
GO:0009059	macromolecule biosynthetic process	7	44.65	35	50.63	5	3.58
GO:0042254	ribosome biogenesis	11	35.05	1	129.44	12	2.26
GO:0006396	RNA processing	13	29.38	5	97.30	14	1.91
GO:0006364	rRNA processing	15	23.45	3	112.21	18	1.50
GO:0016072	rRNA metabolic process	16	23.40	4	108.12	17	1.52

Table 3.1: Top GO terms obtained by ontological regression, linear pooling, and HDP. GO terms are ranked according to p-values or Bayesian analogue.

Figure 3.2 shows the densities of gene-condition slopes and 10-fold cross-validation residuals to observed gene expression. Ontological regression shows a fatter-tailed distribution of regression coefficients, a feature expected in such a Bayesian hierarchical regression. Despite the wider spread in coefficients, it shows cross-validated residuals that more closely resemble those from simple linear regression (SLR) than HDP.

3.3.2 Predicting out-of-sample genes

As previously noted, the existence of GO-level activity estimates and the mapping matrix γ allows us to predict the expression of genes that are not in our training set. Recall $E(\beta^{gene-cond}) = \gamma\beta^{GO-cond}$. To predict the expression of an unobserved gene, we merely add a row to γ containing the assignments of the new gene and obtain γ_+. To verify the accuracy of these predictions, we compare the gene-condition coefficients from a full run of the ontological regression (all genes included) against GO-derived predictions obtained from a run with a fixed percentage of the genes removed.

We perform this procedure leaving out 10%, 20%, 30%, 40%, and 50% of all genes, with 10 replicates per proportion left-out. Results are shown in Figure 3.3. With only 10% of genes removed, out-of-sample predictions are correlated with sampled estimates at

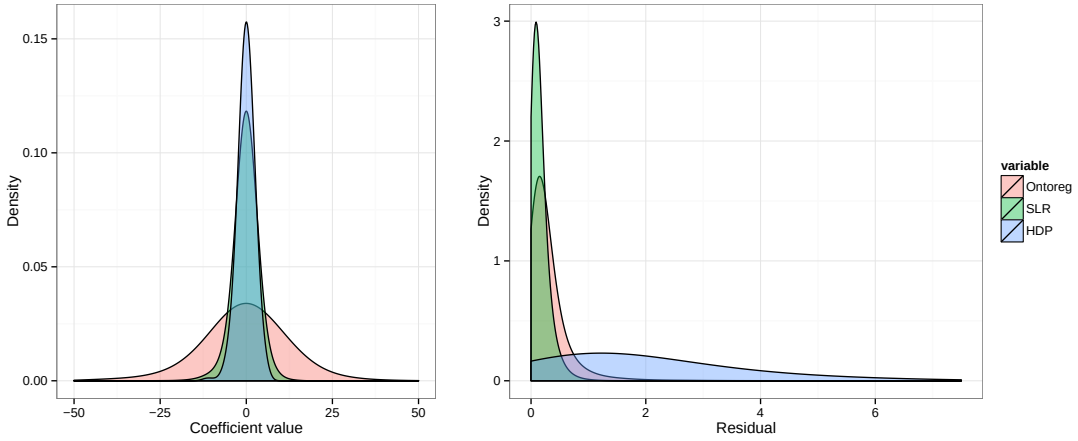


Figure 3.2: Densities of gene-level coefficients and 10-fold cross-validation residuals by method. HDP results in a spiky distribution of slopes (irregular features are more visible in Supplementary Figure B.1), while linear pooling and ontological regression generate smoother densities. The density of ontological regression coefficients is fatter-tailed, a feature expected for a hierarchical model. We note that although ontological regression results in a wider spread of coefficients, cross-validation residuals are nearly as small as simple linear regression and significantly better than those achieved by HDP.

$r = 0.78$. This correlation decreases as higher percentages of genes are removed, with 50% removal resulting in $r = 0.61$. The fact that such a strong correlation is maintained even when only half of the genes are used to estimate coefficients speaks to the ability of the Gene Ontology and this model to provide structured estimates of gene activity.

To confirm that these results are not due simply to the projection of gene-wise results into a higher-dimensional space, we obtain a bootstrapped null distribution by resampling the assignment of genes to GO terms and performing the same procedure. Results are shown in Figure 3.4. Bands represent the 25% and 75% quantiles of correlation based on replicates in each condition. A null mapping between genes and GO terms results in effectively zero correlation between sampled coefficients and out-of-sample predicted coefficients, indicating that the biological relevance of GO terms drives the correlation demonstrated in Figure 3.3.

Figure 3.3: Correlation between out-of-sample gene-factor slope predictions based on GO-condition coefficients and sampled gene-factor slopes from a run with all data.

3.3.3 Predicting model failure

Since we have demonstrated that ontological regression can predict the expression of out-of-sample genes with reasonable accuracy, the next natural question would be how to know when these predictions are reliable. We note that our model also contains an internal measure of validity: the agreement (or lack thereof) between gene-condition estimates and the expectation of those gene-condition coefficients implied by the GO-condition mean. We quantify this as the predicted gene expression according to gene-condition estimates, less the predicted gene expression according to GO-condition estimates. We compare the absolute value of this in-sample residual against out-of-sample gene expression residuals (actual gene expression less gene expression predicted by GO-condition estimates).

This measure of internal validity correlates with the error in residuals obtained from the estimated gene-level coefficients, as shown in Figure 3.5. Again, we consider out-of-sample predictions with 10%, 20%, 30%, 40%, and 50% of genes left out of the sampler. The differene between out-of-sample predictions and fully sampled values is compared to the average difference of in-sample prediction and sampled values. Absolute residuals are again correlated most strongly when relatively few genes are omitted ($r = 0.65$ for the 10% left out set, versus $r = 0.41$ for the 50% set). That is, if a large training set is available, it is easier to determine whether out-of-sample predictions will be accurate.

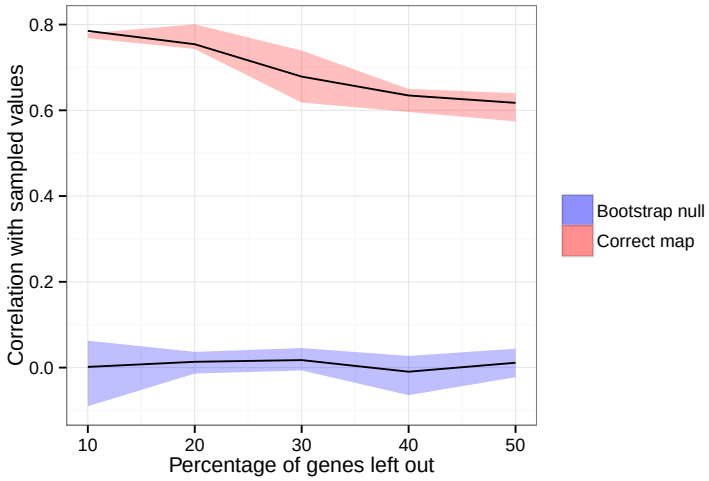


Figure 3.4: Correlation between sampled and out-of-sample predicted gene-condition slopes from using the true mapping versus a bootstrapped null mapping between genes and GO terms. An irrelevant mapping does not yield any substantial correlation between gene-condition slopes and those predicted from GO-condition slopes.

In a scenario with truly unobserved genes, we would be unable to compare the in-sample prediction residuals to calculate this measure directly, as we would not have any gene-level coefficients to compare against. However, we can perform this in-sample residual calculation for all genes in our training set, and determine which GO terms are are most strongly associated with poor out-of-sample predictions. In the Brauer et al. (2008) data set, the GO terms contributing most to model failure are shown in Table 3.2. Several metabolic terms are contained in this list, including broad categories such as GO:0006521 (regulation of cellular amino acid metabolic process) and GO:0045763 (negative regulation of cellular amino acid metabolic process). For out-of-sample genes for which we want to predict expression, membership in one of these GO terms would indicate an increased likelihood of inaccurate predictions.

3.4 Remarks

In this paper we have presented a Bayesian hierarchical model that leverages the Gene Ontology for characterization of changes in gene expression based on functional groups.

Figure 3.5: Correlation between absolute out-of-sample gene expression residuals (actual gene expression less gene expression predicted by GO-condition estimates) and absolute in-sample gene prediction disagreement (predicted gene expression according to gene-condition estimates less predicted gene expression according to GO-condition estimates).

By using the GO to link genes to functional terms and to approximate covariance between GO terms, we obtain biologically interpetable groups, and enable clearer attribution of effects between GO terms (relative to list-based methods). Beyond simple descriptions of relevant functional groups, we can predict the expression of genes not present in the set of observed genes. Further, we can identify which GO terms are likely to lead to accurate out-of-sample predictions by evaluating the reliability of gene-level estimates overall.

While this model does offer clear benefits over either a simple averaging-based model or a nonparametric model such as HDP, it comes at a cost. The Gibbs sampler – which at each iteration retains thousands of mean estimates and covariance matrix entries – requires computational resources that may make this method infeasible for more complex organisms with larger numbers of genes and GO terms. There are also many assumptions made in the course of this model construction, and each provides an opportunity for error. That said, we have limited informativeness of priors where feasible, and feel comfortable that the model is providing reasonable results in our case study.

This work could be extended to non-Gaussian phenotypes at the GO level using a generalized linear model framework for a fixed Σ_{MRCA}. Additionally, it would be desirable to allow uncertainty in assignment of genes to GO terms directly into the model, instead

	GOID	Term	Total residual
1	GO:0006521	regulation of cellular amino acid metabolic process	8058.87
2	GO:0045763	negative regulation of cellular amino acid metabolic process	8054.17
3	GO:0009896	positive regulation of catabolic process	7834.36
4	GO:0045913	positive regulation of carbohydrate metabolic process	7541.55
5	GO:0035065	regulation of histone acetylation	6847.11
6	GO:0042816	vitamin B6 metabolic process	6283.57
7	GO:0016239	positive regulation of macroautophagy	5644.11
8	GO:0031056	regulation of histone modification	5453.54
9	GO:0045732	positive regulation of protein catabolic process	4936.82
10	GO:0015936	coenzyme A metabolic process	4760.00

Table 3.2: GO terms most likely to yield poor out-of-sample predictions. The total residual column indicates the error summed over all member genes (normalized by the number of GO terms to which they belong).

of fixing it to a constant mapping.

Chapter 4

Prediction of drug sensitivity by gene signature activation patterns

4.1 Introduction

Since the arrival of high-throughput gene expression measurement technology, many techniques have been developed to construct gene expression signatures that may be used to classify disease states and predict drug response. This approach is particularly prevalent in cancer research, where disease drivers may vary between patients, between tumors, or even within a single tumor (Gerlinger et al., 2012). As such, a great deal of methodological research has been performed in search of methods that can accurately predict efficacy for individual subjects (Golub et al., 1999; Saeys et al., 2007; Van De Vijver et al., 2002).

The construction of these expression signatures is usually performed via one of three methods: (i) grouping genes according to database annotations, (ii) experimental generation of gene expression signatures (measured after some perturbation which activates or deactivates a given pathway), or (iii) factor analysis methods. Database and experimentally generated pathways have the advantage of providing biologically interpretable results, but may be of limited relevance to the phenomenon currently under study. They may fail to generalize across tissue types, or may be dependent upon unmeasured quantities, such as methylation status. Factor analysis models typically offer better fit, but generate groups that are not easily interpretable from a biological standpoint. In addition, these models may overfit, and generated signatures often fail to replicate in subsequent analyses.

Shen et al. (2015) recently published Adaptive Signature Selection and InteGratioN (ASSIGN), a method which attempts a compromise between the experimental and factor analysis strategies. ASSIGN incorporates information from lab experiments with a flexible factor analysis that can adapt to multiple tissue types or other background conditions. Experimental signatures are input as prior information into a Bayesian factor analysis model, which allows for some deviation between the experimental signature and the final pathway. The model also includes a term for background adjustment that can vary by context (tissue, cell line, etc). This has the advantage of allowing for adaptation when the patient samples differ in disease state or tissue context from the cells in which the experimental signature was generated. In addition, though experimental signatures are generated in isolation, ASSIGN permits simultaneous inference of activations for the samples of interest. Shen et al. (2015) examined the ability of these pathway activations to differentiate between different cancer subtypes, and found that they individually offered better discrimination than Bayesian Factor Regression Models (BFRM; Bernardo et al., 2003) and FacPad (Ma and Zhao, 2012). As one of the stated advantages of ASSIGN over models like GSEA (Subramanian et al., 2005) or BFRM is that it can infer multiple pathway activations simultaneously, we consider the ability these pathways to jointly predict drug response. This paper proposes a simple method for predicting drug response based on ASSIGN pathway activity estimates, and presents a case study in data from the Integrative Cancer Biology Program (ICBP; Daemen et al., 2013).

4.2 Methods

Broadly speaking, we use ASSIGN to estimate pathway activations, and then predict drug sensitivity via a penalized logistic regression model. The pathways and interactions to be incorporated into the final prediction model will be determined through use of the lasso on a 2-way interaction model by Bien et al. (2013).

4.2.1 Pathway signatures

As previously noted, ASSIGN requires data from two different sources: experimental data in which genes have been over- or under-expressed, and the patient/cell line samples for which we want to know pathway activations.

We present a brief overview of the model here; details may be found in Shen et al. (2015). The first stage of ASSIGN is the construction of a set of gene expression signatures. Microarray or RNA-Seq measurements are taken on a set of cells both before and after perturbations to particular genes/pathways, through methods such as gene knockdown experiments or adenovirus transfection. The differential expression between these states is then used to construct an informative prior on which genes should be included in a signature of activity for that pathway. Next, the patient or cell line samples of interest are assessed for activity in these pathways in a Bayesian factor analysis model.

For N patient/cell line samples on G genes and k experimentally perturbed pathways, the ASSIGN factor model may be written

$$Y_{G\times N} = B_{G\times 1}1'_{N\times 1} + S_{G\times k}A_{k\times N} + E_{G\times N} \ , \tag{4.1}$$

where Y denotes the expression of genes in the patient or cell line samples, B is the background expression level, S denotes the matrix containing the pathway signatures, and A contains sample-specific activations of those signatures. Sample j's expression follows $Y_{,j} \sim N(B + SA_{,j}, \Sigma)$, where $\Sigma = \text{diag}(\tau_1^{-1}, \ldots, \tau_G^{-1})$. Precision for the error terms τ_g are distributed Gamma(u, v), usually chosen to be non-informative. The background vector is distributed $B \sim N(\mu_B, S_B)$, in which μ_B and S_B are determined by the experimental data.

The matrix S is constructed according to a spike-and-slab prior on the experimental data, where, $S_{g,k}|\delta_{g,k} \sim (1 - \delta_{g,k})N(0, \omega_0^2) + \delta_{g,k}N(0, \omega_1^2)$. In this equation, $\delta_{g,k} \sim$ Bernoulli$(\pi_{g,k})$, with $\pi_{g,k}$ determined by the probability of differential expression between control and perturbed samples in the experimental data. ASSIGN takes $\omega_1^2 = 1$ and

$\omega_1^2 = 0.1$. That is, when $\delta_{g,k} = 1$, a diffuse prior on $S_{g,k}$ is used, and gene g contributes non-negligibly to the pathway k signature.

The matrix A is of primary interest for our application – for each sample, it contains the activation scores of each pathway for each signal. We take the posterior means of entries in A as predictors of drug sensitivity in a series of regression models. In the standard ASSIGN model, the entries of A follow a modified spike-and-slab distribution to encourage sparsity within columns (that is, an individual sample will exhibit only a limited number of pathway activations):

$$A_{k,j} \sim (1 - \gamma_{k,j})N(0, \omega_0^2) + \gamma_{k,j} \frac{\frac{1}{\omega_1} N(0,1)}{\Phi(\frac{1}{\omega_1}) - \Phi(0)} \ , \tag{4.2}$$

where Φ is the cumulative distribution function of the standard normal. Typically $\gamma_{k,j} \sim$ Bernoulli($\lambda_{k,j}$), with $0 < \lambda_{k,j} < 1$. Since our response prediction procedure will incorporate a selection procedure on the pathway activations, we do not wish to induce sparsity in A, and instead effectively set $\lambda_{k,j} = 1$ for all j, k. This may be accomplished by setting the parameter `mixture_beta=FALSE` in the ASSIGN Bioconductor package from Shen et al., 2013.

4.2.2 Drug response prediction

ASSIGN provides us with sample-specific estimates of pathway activation. We model sensitivity according to a logistic regression model, allowing for main effects due to cancer subtype and pathway activation. We also permit two-way interactions between subtype and pathway activations, and between pathways. As subtypes are mutually exclusive, subtype-subtype interactions are not considered.

Responsiveness to treatment is determined according to thresholding in GI_{50}, the dose required to result in a 50% inhibition in growth. These thresholds vary by drug; the cutoffs for each of the 82 treatments we will consider later can be found in the supplementary materials to Daemen et al. (2013).

Denote R_i the (unobserved) probability that sample i has a GI_{50} indicative of sensitivity to the drug under consideration, and $S(i)$ the transcriptional subtype of sample i. Denote the subtype indicator function

$$\mathbf{1}_s(S(i)) = \begin{cases} 0 & S(i) \neq s \\ 1 & S(i) = s \end{cases} . \tag{4.3}$$

Our model may be written

$$\text{logit}(R_i) = \sum_s \alpha_s \mathbf{1}_s(S(i)) + \sum_k \beta_k A_{k,i} + \frac{1}{2}\sum_s \sum_k \gamma_{s,k} A_{k,i} \mathbf{1}_s(S(i)) + \frac{1}{2}\sum_{j \neq k} \theta_{k,j} A_{k,i} A_{j,i} \ . \tag{4.4}$$

We encourage sparsity in this model using an ℓ_1-norm penalty according to the model of Bien et al. (2013), as implemented in the `hierNet` package in `R` (Bien and Tibshirani, 2014). Write our coefficients $\zeta = (\alpha, \beta)$, and $\Psi = \left(\begin{smallmatrix} 0 & \Gamma \\ \Gamma^T & \Theta \end{smallmatrix}\right)$. We impose a weak hierarchy on our model, such that an interaction between two variables requires at least one of them to have a nonzero main effect in the regression model. Given logit loss function $q(\zeta_0, \zeta^+ - \zeta^-, \Psi)$, we can rewrite the form of Equation 4.4 as

$$\text{logit}(R_i) = \hat{\zeta}_0 + (S(i)^T, A_{,i}^T)(\zeta^+ - \zeta^-) + (S(i)^T, A_{,i}^T)\Psi(S(i), A_{,i}) \tag{4.5}$$

and optimize according to

$$\min_{\zeta_0 \in \mathbb{R}, \zeta \in \mathbb{R}^P, \Psi \in \mathbb{R}^{p \times p}} q(\zeta_0, \zeta^+ - \zeta^-, \Psi) + \lambda 1^T(\zeta^+ + \zeta^-) + \frac{\lambda}{2}||\Psi||_1 \tag{4.6}$$

$$\left.\begin{array}{ll} \text{s.t.} & ||\Psi_j||_1 \leq \zeta^+ + \zeta^- \\ & \zeta_j^+ \geq 0, \zeta^- \geq 0 \end{array}\right\} \text{ for } j = 1, \ldots, p \ . \tag{4.7}$$

An additinal l_2 penalty is optionally applied for stability. Details of this optimization and additional model properties may be found in Bien et al. (2013).

4.3 Results

4.3.1 Prediction of drug response

We apply ASSIGN to RNA-Seq from the Integrative Cancer Biology Program (ICBP; Daemen et al., 2013). We use data from a set of adenovirus transfection experiments to obtain activation scores for 4 pathways: *AKT*, *HER2*, *IGF1R*, and *BAD*. First, cultured human mammary epithelial cells are infected with adenoviruses that overexpress a particular gene. After sufficient time is allowed for the cell to reach its new steady state, gene expression is measured RNA-seq. This is effectively the inverse of the more familiar knockdown experiment. The resulting gene signature prior is incorporated into the main ASSIGN factor analysis. From this, we use the resulting activation matrix A and the strategy outlined in Sections 4.2.1-4.2.2 to predict response to 82 drugs.

To evaluate performance of our logistic models of the form of Equation 4.4, for a single treatment, we first construct a set of interaction model with all samples $i = 1, \ldots, n$ and select the parameters (λ_1, λ_2) which minimize BIC. As recommended by Bien et al. (2013), we take the elastic net parameter $\lambda_2 = 10^{-8}\lambda_1$, which drastically reduces the size of the search space, while still providing some stability in variable selection. We then fit the model with $(\lambda_1^*, \lambda_2^*)$ n times, each time omitting one sample from the data used to estimate α, β, γ, and θ. We use the estimated model to predict the probability of response in the left-out sample according to the fitted model. Using the predicted probabilities from the leave-one-out procedure and the true response status, we calculate the AUC of each of these models by comparing to the binary drug response data. This procedure is performed for all 82 treatments for which we have drug response data.

We compare performance of the interaction models with regression models that feature only main effects. These models fall into three classes: (i) subtype effects only, (ii) subtype and a single pathway activation, and (iii) subtype and all pathway activations. No variable selection procedures are included in fitting these models. It should be noted that previous papers (Daemen et al., 2013; Shen et al., 2015) effectively consider only model classes (i)

and (ii). We evaluate all of these on the basis of AUC generated from the same leave-one-out procedure, where a model is fit n times, each with one observation excluded, whose probability of response is predicted according to the fitted logistic model.

Model	Mean AUC	Mean improvement	Count best model
Subtype only	0.70		0
Subtype and *AKT*	0.75	0.09	0
Subtype and *BAD*	0.73	0.07	2
Subtype and *HER2*	0.75	0.08	2
Subtype and *IGF1R*	0.74	0.08	0
Subtype and main effects	0.81	0.04	27
Full interaction lasso	0.85	0.10	51

Table 4.1: Summary of model class performance over all treatments. AUC is calculated according to the leave-one-out procedure described in Section 4.2.2. Mean improvement is relative to the next-simplest class of model (subtype and a single pathway show average gains over subtype only, subtype and main effects show improvement over subtype and the single best pathway, and the interaction model shows improvement over subtype and all main effects), and only calculated for treatments in which AUC increases. The times that the model has the highest AUC is also shown (out of all 82 treatments).

Figure 4.1 shows relative model performance according to AUC. Out of the 82 treatments considered, 51 had improved performance from use of an interaction model over the next-best method. 27 treatments showed the highest AUC in a model with subtype and all main effects, while only 4 treatments performed best with subtype and a single pathway. Of samples where the interaction model increased AUC over the next-best model, the interaction model offered an improvement of 0.10 on average. In 6 cases, the interaction model led to perfect discrimination (leave-one-out AUC= 1.00) among our samples. Tables 4.1 and C.1 provide additional details.

While the interaction model clearly offers the best improvement in the majority of treatments, it is also worth noting that the model including subtypes and all main effects also offers a non-negligible improvement over single-pathway or subtype-only models. This class of model achieves an average AUC of 0.81, and contributing a mean 0.04 improvement over a single-pathway model. Out of the 31 cases in which the interaction model was not the best performer, the main effects model prevailed in 27. This indicates that the current

strategy of fitting all pathway activations simultaneously but only performing drug response predictions one-at-a-time sells short the advantage of joint modeling.

4.3.2 Details of interaction models

When we examine the results of the interaction models, patterns of interactions can be observed. Figure 4.2 shows a heatmap of these coefficients by treatment from models fit on all observations. For example, an interaction of *AKT* and *HER2* generally results in a positive coefficient, corresponding to increased odds of response to these treatments when both are active. By contrast, the interaction term between *BAD* and *IGF1R* is typically negative, indicating decreased odds of response when both are active.

We can also examine the interactions between subtypes and pathways. The luminal subtype interaction with *HER2* is typically positive, meaning that luminal cancers are more likely to respond to these treatments when *HER2* is also active. The converse is true of the basal subtype and *HER2*, which typically results in a negative interaction term.

These interactions may also be represented in a network diagram, as in Figure 4.3. Nodes indicate the main effects present in the final interaction model, and edges the interactions between these nodes. In this diagram, nodes and edges are color-coded according to the direction of model coefficients (red denoting a negative value, and blue positive). The intensity of the color indicates the magnitude of the coefficient. Figure 4.3a shows such a diagram for treatment Glycyl H1152. Cell lines with luminal subtype are more likely to respond to this drug, while the presence of *AKT*, *BAD*, or *IGF1R* activation negatively affects the odds of response. Luminal subtype with *IGF1R* activation leads to lower odds of response than would be expected from either the subtype or *IGF1R* activation individually. Diagrams for all treatments showing improvement from the interaction model are shown in Appendix C.3.

4.4 Remarks

Shen et al.'s ASSIGN algorithm offers an exciting step forward in terms of allowing the flexibility of machine learning models and the interpretability of experimentally derived pathways. We demonstrate here that the power of these signature activations to predict drug response is actually greater than previously published. By using all available signature information to model sensitivity, we obtain more significantly improved predictions. Because ASSIGN pathways have clearly defined origins, we can interpret the interactions between subtypes and pathways in order to gain deeper insight into mechanisms at play. Furthermore, we can examine the model coefficients for all treatments and look for common patterns of response.

The overall method outlined in this paper bears with it the limitations of each of its two main components: ASSIGN and lasso models. Both may produce misleading results in the presence of collinearity between signatures, or when interactions between pathways are strongly nonlinear. As such, we recommend this procedure only for situations in which no more than 12 correlated pathways are included (as per Shen et al., 2015). Additional diagnostics may be performed to ensure that this methodology is suitable, such as examining trace plots and Gelman-Rubin convergence statistics for signs that the ASSIGN factor analysis model has successfully converged (Gelman and Rubin, 1992).

The use of this style of interaction modeling represents only a first step in the use of pathway activation to predict drug response. In particular, the penalty term in the lasso model may be modified to accomodate prior information regarding the plausibililty of biological interactions with respect to drug rsponse. Lu et al. (2013) used pathways from the Kyoto encyclopedia of genes and genomes (KEGG; Kanehisa and Goto, 2000) in this fashion to evaluate gene-gene interactions in GWAS of continuous traits. As has been demonstrated here however, the ability to predict response based on pathway-level activity can be greatly improved by even relatively unsophisticated joint modeling procedures.

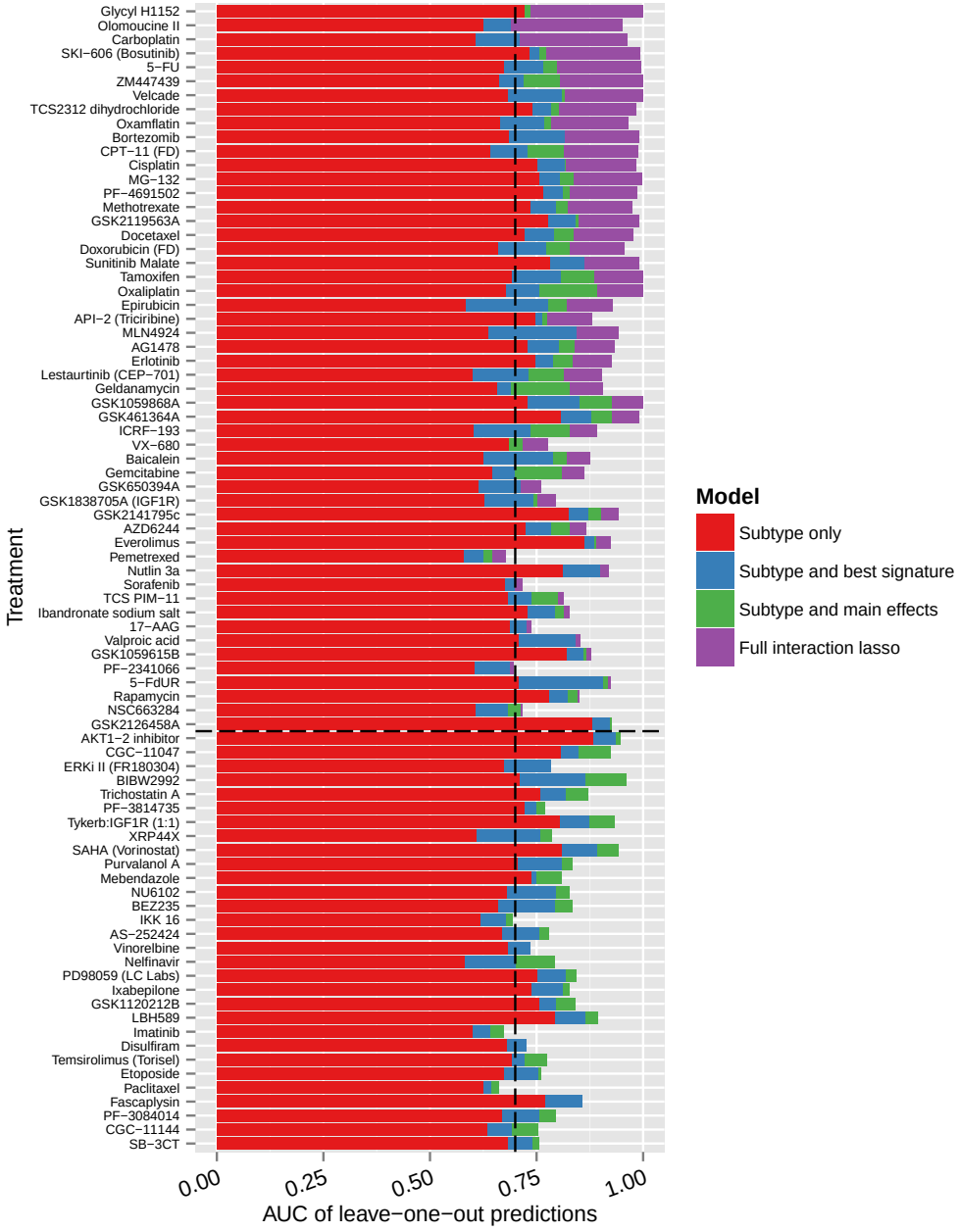


Figure 4.1: AUC based on leave-one-out models for subtype-only models, subtype a single pathway (best of *AKT*, *BAD*, *HER2*, and *IGF1R* shown), subtype and all pathways, and the full interaction lasso. Treatments are ordered according to the gain in leave-one-out AUC from the interaction model over the next-best performer. The vertical dashed line indicates AUC=0.70, and the horizontal dashed line indicates the treatment for which no improvement is obtained from the interaction model.

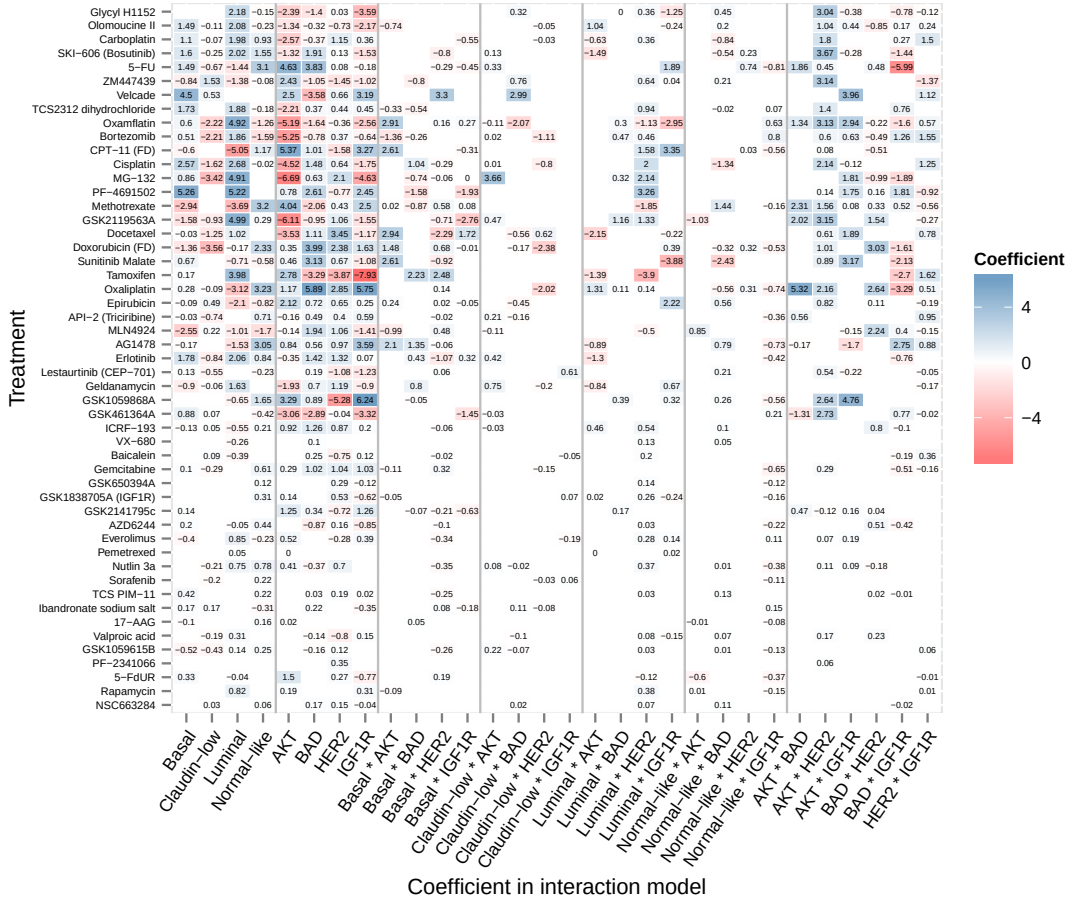
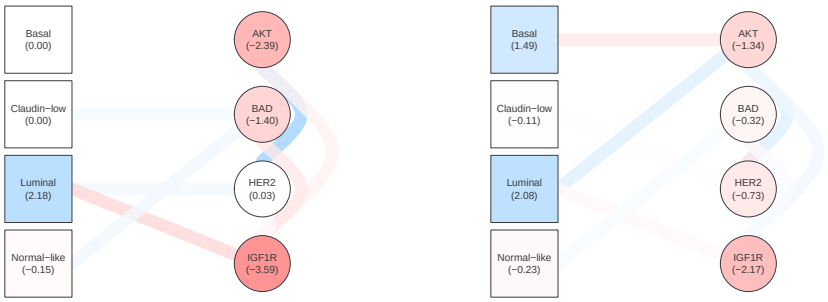


Figure 4.2: Coefficients from the interaction logistic model for treatments in which the interaction model offers the best performance by leave-one-out AUC. Each coefficient x may be converted to a multiplier on the odds of response by taking $\exp(x)$.



(a) Glycyl H1152 (AUC=1.00) (b) Olomoucine II (AUC=0.95)

Figure 4.3: Network representations of the lasso interaction models for Glycyl H1152 and Olomoucine II.

Chapter 5

Conclusion

In this dissertation, we have presented several network-based approaches to biological data, each using a different notion of a biological network.

Chapter 2 contained a method to detect perturbations in multi-omics biological data based on a conditional Gaussian Graphical model. Likelihood ratio tests provide a formal statistical framework for determining the most likely site for a perturbation for a given network structure. If multiple perturbations are suspected, a procedure for sequential inference is provided as well. The efficacy of this model compared to other inference strategies is demonstrated in simulation studies and in an analysis of breast cancer data from The Cancer Genome Atlas.

Chapter 3 used the network of the Gene Ontology to map genes to relevant functional groups and to approximate covariance between these groups. Using a Bayesian hierarchical regression, we obtain both gene- and GO term-level regression coefficients, which enable us This method was demonstrated via a case study of gene expression in controlled experiments of *S. cerevisiae* (Brauer et al., 2008). It produced biologically reasonble results in terms of easily interpretable groups.

Chapter 4 considers a networks constructed of pathway-pathway and pathway-subtype interactions as determined by a penalized regression model to predict drug response. Pathway activations are first estimated by application of the ASSIGN algorithm (Shen et al., 2015). Next, these activations are used with subtype in a penalized logistic model with main effects and 2-way interactions to predict response status for 82 different drugs. The resulting analysis showed a dramatic increase for most drugs in accuracy of response predic-

tion based on interaction models. Visualization of these models through network diagrams allows for ready interpretation, and can aid investigators in identifying relevant patterns.

Each of the methods discussed here has further room to grow as well. The perturbation detection models of Chapter 2 could be extended to non-Gaussian phenotypes to allow for the inclusion of a wider variety of data types. In addition, the decision to pursue gene-level inferences is a simple first approach, but further work could be done to determine optimal granularity and groupings for such tests. The Bayesian regression framework of Chapter 3 may benefit from models that incorporate uncertainty in assignment of genes to ontology terms. Chapter 4's regressions may be modified to incorporate a interaction-specific penalties based on prior biological data, similar to the gene-gene interaction models developed by Lu et al. (2013).

Though these models vary greatly in the way that networks are used, they all show the benefits of modeling complicated biological data within the context of the interactions that define it. When identifying perturbations, the largest change is not always the site of an initial disturbance. In describing genome-wide expression changes, better inference is permitted by considering membership in multiple functional groups and the relationships between those groups. Consideration of interactions between pathways and between pathways and subtypes can go beyond improving drug response prediction accuracy, but also allows for better understanding of the models that achieve it. By allowing our models a small measure of the interaction complexity inherent in high-throughput biological data, we can achieve better understanding of the phenomena under study.

Appendix A

Supplementary materials: "Detection of multiple perturbations in multi-omics biological networks"

A.1 Software

Supplementary files, including the simulation pipeline and TCGA scripts/processed data, may be found at `https://github.com/paulajgriffin/mapggm_supplemental`.

An R package **mapggm** containing methods for multi-attribute network estimation and perturbation detection is available at `https://github.com/paulajgriffin/mapggm`. To use this package, install the **devtools** package from CRAN and run:

```
library(devtools)
install_github('paulajgriffin/mapggm')
```

A.2 Properties of sequential tests

We prove the following theorems, presented in Section 2.3.2.

Theorem A.1 *Given a set of nodes already found to have nonzero mean in steps* $1, \dots s$*, consider testing for a perturbation at an additional node* i *in step* $s+1$*. Denote the indices in* $Z = \Omega Y$ *corresponding to the nodes found in steps* $1, \dots, s$ *as* S*.*

We can write the expected difference between the original test statistic and the test statistic adjusted for perturbations in S *as*

$$E(T_i - T_i^{[s+1]}) = \mu_i^T \left(\Sigma_{i,s}\Sigma_{s,i}\right) \mu_i + 2\mu_i^T \left(\Sigma_{i,s}\right) \mu_S + \mu_S^T \left(\Sigma_{s,i}\Sigma_{i,s}\right) \mu_S \ .$$

In the special case that $\mu_i = 0$,

$$E(T_i - T_i^{[s+1]} | \mu_i = 0) \geq 0 \ .$$

Proof of Theorem A.1. First, recall the form of T_j, the unadjusted test statistic for testing the alternative hypothesis that $\mu_j \neq 0$ and $\mu_{(-j)} = 0$ against a null of $\mu = 0$. We can rewrite the test statistic in terms of the unfiltered data y and obtain

$$\begin{aligned}
T_j &= n\left(\bar{z}^T\Sigma\bar{z} - \bar{z}_{(-j)}^T\left(\Sigma_{(-j),(-j)} - \Sigma_{(-j),j}\Sigma_{j,j}^{-1}\Sigma_{j,(-j)}\right)\bar{z}_{(-j)}\right) \\
&= n\left(\bar{z}^T\Sigma\bar{z} - \bar{z}^T\begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{(-j),(-j)} - \Sigma_{(-j),j}\Sigma_{j,j}^{-1}\Sigma_{j,(-j)} \end{pmatrix}\bar{z}\right) \\
&= n\left(\bar{y}^T\Omega\Sigma\Omega\bar{z} - \bar{y}^T\Omega\begin{pmatrix} 0 & 0 \\ 0 & \Omega_{(-j),(-j)}^{-1} \end{pmatrix}\Omega\bar{y}\right) \\
&= n\left(\bar{y}^T\Omega\bar{z} - \bar{y}^T\begin{pmatrix} \Omega_{j,(-j)}\Omega_{(-j),(-j)}^{-1}\Omega_{(-j),j} & \Omega_{j,(-j)} \\ \Omega_{(-j),j} & \Omega_{(-j),(-j)} \end{pmatrix}\bar{y}\right) \\
&= n\left(\bar{y}^T\begin{pmatrix} \Omega_{j,j} - \Omega_{j,(-j)}\Omega_{(-j),(-j)}^{-1}\Omega_{(-j),j} & 0 \\ 0 & 0 \end{pmatrix}\bar{y}\right) \\
&= n\left(\bar{y}^T\begin{pmatrix} \Sigma_{j,j}^{-1} & 0 \\ 0 & 0 \end{pmatrix}\bar{y}\right) \ .
\end{aligned}$$

The mean of the unfiltered data has distribution $\bar{y} \sim N(\Sigma\mu, \Sigma/n)$. Taking the expectation of our test statistic T_j, we obtain

$$E(T_j) = nE(\bar{y})^T\begin{pmatrix} \Sigma_{j,j}^{-1} & 0 \\ 0 & 0 \end{pmatrix}E(\bar{y}) + \mathrm{Tr}\left(\left(\begin{pmatrix} \Sigma_{j,j}^{-1} & 0 \\ 0 & 0 \end{pmatrix}\Sigma\right)\right.$$

$$
\begin{aligned}
=& n \begin{pmatrix} \Sigma_{j,j}\mu_j + \Sigma_{j,(-j)}\mu_{(-j)} \\ \Sigma_{(-j),j}\mu_j + \Sigma_{(-j),(-j)}\mu_{(-j)} \end{pmatrix}^T \begin{pmatrix} \Sigma_{j,j}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Sigma_{j,j}\mu_j + \Sigma_{j,(-j)}\mu_{(-j)} \\ \Sigma_{(-j),j}\mu_j + \Sigma_{(-j),(-j)}\mu_{(-j)} \end{pmatrix} \\
&+ \mathrm{Tr}\left(\begin{pmatrix} I & \Sigma_{j,j}^{-1}\Sigma_{j,(-j)} \\ 0 & 0 \end{pmatrix} \right) \\
=& n\left(\mu_j^T \Sigma_{j,j}\mu_j + 2\mu_j^T \Sigma_{j,(-j)}\mu_{(-j)} + \mu_{(-j)}^T \Sigma_{(-j),j}\Sigma_{j,(-j)}\mu_{(-j)} \right) + k_j \ ,
\end{aligned}
$$

where k_j indicates the number of attributes for node j.

Denote the indices in $Z = \Omega Y$ corresponding to the nodes found in steps $1, \ldots, s$ as S, and the indices corresponding to the node currently under consideration as i. Denote all other indices X.

In the sequential testing procedure, we test the alternative hypothesis of $\mu_i \neq 0$, $\mu_S \neq 0$, $\mu_X = 0$ against the null that $\mu_S \neq 0$, $\mu_i = 0$, $\mu_X = 0$. We can write the adjusted test statistic $T_i^{[s+1]}$ as the difference of two unadjusted test statistics

$$
T_i^{[s+1]} = T_{(i,S)} - T_S \ . \tag{A.1}
$$

We are interested in $E(T_i - T_i^{[s+1]})$, the expected difference between the original and adjusted test statistic.

$$
\begin{aligned}
E(T_i - T_i^{[s+1]}) =& \ E(T_i) + E(T_S) - E(T_{(i,S)}) \\
=& \ n\left(\mu_i^T \Sigma_{ii}\mu_i + 2\mu_i^T \Sigma_{i,S}\mu_s + 2\mu_i^T \Sigma_{i,X}\mu_X + \right. \\
& \left. \mu_S^T \Sigma_{S,i}\Sigma_{i,S}\mu_S + \mu_X^T \Sigma_{X,i}\Sigma_{i,X}\mu_X \right) + k_i \\
& + n\left(\mu_S^T \Sigma_{S,S}\mu_S + 2\mu_S^T \Sigma_{S,i}\mu_i + 2\mu_S^T \Sigma_{S,X}\mu_X + \mu_i^T \Sigma_{i,S}\Sigma_{S,i}\mu_i + \right. \\
& \left. \mu_X^T \Sigma_{X,S}\Sigma_{S,X}\mu_X \right) + \sum_{j \in S} k_j \\
& - n\left(\mu_i^T \Sigma_{i,i}\mu_i + 2\mu_i^T \Sigma_{i,S}\mu_S + \mu_S^T \Sigma_{S,S}\mu_S + 2\mu_i^T \Sigma_{i,X}\mu_X + 2\mu_S^T \Sigma_{S,X}\mu_X \right.
\end{aligned}
$$

$$+ \mu_X^T \Sigma_{X,i} \Sigma_{i,X} \mu_X + \mu_X^T \Sigma_{X,S} \Sigma_{S,X} \mu_x) + \left(k_i + \sum_{j \in S} k_j \right)$$

By gathering common terms, we obtain

$$E(T_i - T_i^{[s+1]}) = \mu_i^T \left(\Sigma_{i,S} \Sigma_{S,i} \right) \mu_i + 2\mu_i^T \left(\Sigma_{i,S} \right) \mu_S + \mu_S^T \left(\Sigma_{S,i} \Sigma_{i,S} \right) \mu_S \ .$$

In the special case that $\mu_i = 0$ (no perturbation exists at the node under consideration),

$$E(T_i - T_i^{[s+1]} | \mu_i = 0) = \mu_S^T \left(\Sigma_{S,i} \Sigma_{i,S} \right) \mu_S \geq 0 \ ,$$

since $(\Sigma_{S,i}\Sigma_{i,S})$ is by definition positive semi-definite.

Theorem A.2 *Under the same conditions outlined in the general case of Theorem A.1, if* $\Sigma_{S,i} = 0$*, then*

$$T_i^{[s+1]} = T_i \ .$$

Proof of Theorem A.2. Denote the indices in $Z = \Omega Y$ corresponding to the nodes found in steps $1, \ldots, s$ as S, and the indices corresponding to the node currently under consideration as i. Denote all other indices X.

For any $\Sigma_{S,i}$ we can write the test statistic T_i for the unconditional test as

$$\begin{aligned} T_i &= n(\bar{z} - \hat{\mu}_A)^T \Sigma (\bar{z} - \hat{\mu}_A) - n(\bar{z} - \hat{\mu}_0)^T \Sigma (\bar{z} - \hat{\mu}_0) \\ &= n \left((\bar{z} - \hat{\mu}_A) - (\bar{z} - \hat{\mu}_0) \right)^T \Sigma \left((\bar{z} - \hat{\mu}_A) + (\bar{z} - \hat{\mu}_0) \right) \ , \end{aligned}$$

where $\hat{\mu}_0$ and $\hat{\mu}_A$ denote the maximum likelihood estimators for μ under the null and alternative hypothesis, respectively. Without loss of generality, we reorder the filtered data so that $Z = (Z_i', Z_S', Z_X')$

Following formula (7) in the main paper, for the unconditional test, we have

$$\hat{\mu}_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad ,$$

and

$$\hat{\mu}_A = \begin{pmatrix} \bar{z}_i - \Sigma_{i,i}^{-1}\Sigma_{i,(SX)}\bar{z}_{(SX)} \\ 0 \\ 0 \end{pmatrix} \quad .$$

Similarly, a nested likelihood ratio test that conditions on the presence of nonzero mean values for indices S has the form

$$\begin{aligned} T_i^{[s+1]} &= n(\bar{z} - \hat{\mu}_A^{[s+1]})^T \Sigma (\bar{z} - \hat{\mu}_A^{[s+1]}) - n(\bar{z} - \hat{\mu}_0^{[s+1]})^T \Sigma (\bar{z} - \hat{\mu}_0^{[s+1]}) \\ &= n\left((\bar{z} - \hat{\mu}_A^{[s+1]}) - (\bar{z} - \hat{\mu}_0^{[s+1]})\right)^T \Sigma \left((\bar{z} - \hat{\mu}_A^{[s+1]}) + (\bar{z} - \hat{\mu}_0^{[s+1]})\right) \quad , \end{aligned}$$

with restricted MLEs

$$\hat{\mu}_0^{[s+1]} = \begin{pmatrix} 0 \\ \bar{z}_S + \Sigma_{S,S}^{-1}\Sigma_{S,(iX)}\bar{z}_{(iX)} \\ 0 \end{pmatrix} \quad , \text{ and}$$

$$\hat{\mu}_A^{[s+1]} = \begin{pmatrix} \bar{z}_{(iS)} + \Sigma_{(iS),(iS)}^{-1}\Sigma_{(iS),X}\bar{z}_X \\ 0 \end{pmatrix} \quad .$$

Recall $\Sigma_{i,S} = 0$ by assumption. We may rewrite $\hat{\mu}_A$, $\hat{\mu}_0^{[s+1]}$, and $\hat{\mu}_A^{[s+1]}$ as

$$\begin{aligned}
\hat{\mu}_A &= \begin{pmatrix} \bar{z}_i + \Sigma_{i,i}^{-1}(\Sigma_{i,S}\bar{z}_S + \Sigma_{i,X}\bar{z}_X) \\ 0 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} \bar{z}_i + \Sigma_{i,i}^{-1}\Sigma_{i,X}\bar{z}_X \\ 0 \\ 0 \end{pmatrix} \\
\hat{\mu}_0^{[s+1]} &= \begin{pmatrix} 0 \\ \bar{z}_S + \Sigma_{S,S}^{-1}(\Sigma_{S,i}\bar{z}_i + \Sigma_{S,X}\bar{z}_X) \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ \bar{z}_S + \Sigma_{S,S}^{-1}\Sigma_{SX}\bar{z}_X \\ 0 \end{pmatrix} \\
\hat{\mu}_A^{[s+1]} &= \begin{pmatrix} \bar{z}_{(iS)} + \Sigma_{(iS),(iS)}^{-1}\Sigma_{(iS),X}\bar{z}_X \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} \bar{z}_i + \Sigma_{i,i}^{-1}\Sigma_{i,X}\bar{z}_X \\ \bar{z}_S + \Sigma_{S,S}^{-1}\Sigma_{S,X}\bar{z}_X \\ 0 \end{pmatrix} .
\end{aligned}$$

Our unadjusted test yields

$$T_i = n \left(\begin{pmatrix} -\Sigma_{i,i}^{-1}\Sigma_{i,X}\bar{z}_X \\ \bar{z}_S \\ \bar{z}_X \end{pmatrix} - \begin{pmatrix} \bar{z}_i \\ \bar{z}_S \\ \bar{z}_X \end{pmatrix} \right)^T \Sigma \left(\begin{pmatrix} -\Sigma_{i,i}^{-1}\Sigma_{i,X}\bar{z}_X \\ \bar{z}_S \\ \bar{z}_X \end{pmatrix} + \begin{pmatrix} \bar{z}_i \\ \bar{z}_S \\ \bar{z}_X \end{pmatrix} \right)$$

$$= n \begin{pmatrix} -\bar{z}_i - \Sigma_{i,i}^{-1}\Sigma_{i,X}\bar{z}_X \\ 0 \\ 0 \end{pmatrix}^T \Sigma \begin{pmatrix} \bar{z}_i - \Sigma_{i,i}^{-1}\Sigma_{i,X}\bar{z}_X \\ 2\bar{z}_S \\ 2\bar{z}_X \end{pmatrix}$$

By a similar process, the adjusted test statistic is

$$T_i^{[s+1]} = n \begin{pmatrix} -\bar{z}_i - \Sigma_{i,i}^{-1}\Sigma_{i,X}\bar{z}_X \\ 0 \\ 0 \end{pmatrix}^T \Sigma \begin{pmatrix} \bar{z}_i - \Sigma_{i,i}^{-1}\Sigma_{i,X}\bar{z}_X \\ 2\bar{z}_S - 2\Sigma_{S,S}^{-1}\Sigma_{SX}\bar{z}_X \\ 2\bar{z}_X \end{pmatrix} \quad .$$

Note that both of these statistics have the form

$$\begin{aligned} T &= nd'\Sigma a \\ &= n \begin{pmatrix} d_i \\ d_S \\ d_X \end{pmatrix}^T \begin{pmatrix} \Sigma_{ii} & 0 & \Sigma_{iX} \\ 0 & \Sigma_{SS} & \Sigma_{SX} \\ \Sigma_{Xi} & \Sigma_{XS} & \Sigma_{XX} \end{pmatrix} \begin{pmatrix} a_i \\ a_S \\ a_X \end{pmatrix} \\ &= d_i^T \left(\Sigma_{ii}a_i + \Sigma_{iX}a_X\right) + d_S^T \left(\Sigma_{SS}a_S + \Sigma_{SX}a_X\right) + d_X^T \left(\Sigma_{Xi}a_i + \Sigma_{XS}a_S + \Sigma_{XX}a_X\right) \\ &= d_i^T \left(\Sigma_{ii}a_i + \Sigma_{iX}a_X\right) \quad . \end{aligned}$$

In both T_i and $T_i^{[s+1]}$, we have $d_i = -\bar{z}_i - \Sigma_{i,i}^{-1}\Sigma_{i,X}\bar{z}_X$, $a_i = \bar{z}_i - \Sigma_{i,i}^{-1}\Sigma_{i,X}\bar{z}_X$, and $a_X = 2\bar{z}_x$. Therefore, $T_i^{[s+1]} = T_i$.

A.3 Bounds on error in test statistic

We prove the following theorem, presented in Section 2.3.3.

Theorem A.3 *Under the conditions above, the discrepancy $T_1 - \tilde{T}_1$ is equal in distribution to a linear combination of mutually independent, noncentral chisquare random variables,*

$$\sum_{k=1}^{s} a_k \chi_{r_k}^2 \left(\delta_k\right) \quad , \tag{A.2}$$

where

$$\delta_k = (n/2)\mu^T \Sigma_{\cdot 1} E_k \Sigma_{11}^{-1} \Sigma_{1\cdot} \mu \ .$$

Accordingly,

$$E\left[T_1 - \tilde{T}_1\right] = tr\left(D\Sigma_{11}\right) + \frac{1}{2} n\mu^T \Sigma_{\cdot 1} D \Sigma_{1\cdot} \mu \tag{A.3}$$

and

$$Var\left(T_1 - \tilde{T}_1\right) = 2tr\left((D\Sigma_{11})^2\right) + 2n\mu^T \Sigma_{\cdot 1} D\Sigma_{11} D\Sigma_{1\cdot} \mu \ . \tag{A.4}$$

Proof of Theorem A.3. Begin by noting that $T_1 - \tilde{T}_1 = X^T D X$, where

$$D = \Omega_{11} - \Omega_{1\cdot}\Omega_{\cdot\cdot}^{-1}\Omega_{\cdot 1} - \left(\tilde{\Omega}_{11} - \tilde{\Omega}_{1\cdot}\tilde{\Omega}_{\cdot\cdot}^{-1}\tilde{\Omega}_{\cdot 1}\right) \ ,$$

as defined in the paper, and X is a multivariate normal random variable with mean $n^{1/2}\Sigma_{1\cdot}\mu$ and covariance Σ_{11}. Since D is symmetric and Σ_{11} is symmetric and positive definite (the latter by assumption), it follows from Lemma 1 of Baldessari (1967) that $D\Sigma_{11}$ has spectral decomposition

$$D\Sigma_{11} = \sum_{k=1}^{s} a_k E_k \ ,$$

such that $\text{rank}(E_k) = r_k$ (corresponding to the multiplicity of the eigenvalue a_k) and $\sum_{k=1}^{s} r_k = K$. By direct application of Theorem 1 of Baldessari (1967), the expression in (A.2) then follows.

As for the mean and variance expressions in (A.3) and (A.4), we see that

$$\begin{aligned} E\left[T_1 - \tilde{T}_1\right] &= E\left[\sum_{k=1}^{s} a_k \chi^2_{r_k}(\delta_k)\right] \\ &= \sum_{k=1}^{s} a_k (r_k + \delta_k) \\ &= \sum_{k=1}^{s} a_k r_k + \sum_{k=1}^{s} a_k \delta_k \\ &= tr(D\Sigma_{11}) + \frac{n}{2}\mu^T \Sigma_{\cdot 1} D \Sigma_{1\cdot} \mu \ , \end{aligned}$$

and similarly,

$$\begin{aligned}
\text{Var}\left(T_1 - \tilde{T}_1\right) &= \text{Var}\left(\sum_{k=1}^{s} a_k \chi^2_{r_k}(\delta_k)\right) \\
&= \sum_{k=1}^{s} a_k^2 (2r_k + 4\delta_k) \\
&= 2\sum_{k=1}^{s} a_k^2 r_k + 4\sum_{k=1}^{s} a_k^2 \delta_k \\
&= 2tr\left((D\Sigma_{11})^2\right) + 2n\mu^T \Sigma_{\cdot 1} D \Sigma_{11} D \Sigma_{1\cdot} \mu \ ,
\end{aligned}$$

where we have exploited independence among the chisquare random variables in both cases.

The following corollary was also provided in Section 3.3 of the paper.

Corollary A.1 *Let* $||\cdot||_2$ *denote the spectral norm. Then*

$$E\left[T_1 - \tilde{T}_1\right] = O\left(||\Delta||_2\right) \quad \textit{and} \quad Var\left(T_1 - \tilde{T}_1\right) = O\left(||\Delta||_2^2\right) \ .$$

Proof of Corollary A.1. The statements in this corollary follow through application of bounds on the trace of matrix products and repeated application of Cauchy-Schwartz, coupled with an appeal to the Lipschitz smoothness of the mapping between Ω and the expression $\Omega_{11} - \Omega_{1\cdot}\Omega_{\cdot\cdot}^{-1}\Omega_{\cdot 1}$. The latter follows from a straightforward Taylor series argument and the continuity of matrix inversion.

In Wang et al. (1986) it is established that for two matrices M and N, with N symmetric and positive semidefinite, that $|tr(MN)| \leq ||M||_2\, tr(N)$. For the mean, therefore, we have that $tr(D\Sigma_{11}) \leq ||D||_2\, tr(\Sigma_{11})$. At the same time,

$$\left|\frac{n}{2}\mu^T \Sigma_{\cdot 1} D \Sigma_{1\cdot}\mu\right| \leq \frac{n}{2}||\Sigma_{1\cdot}\mu||_2^2 ||D||_2 \ .$$

As a result, we find that $E[T_1 - \tilde{T}_1] = O(||D||_2)$.

Similarly, for the variance

$$2tr\left((D\Sigma_{11})^2\right) \leq 2tr\left(D^2\Sigma_{11}^2\right) \leq 2||D^2||_2\, tr(\Sigma_{11}^2) \leq 2||D||_2^2\, tr(\Sigma_{11}^2) \ ,$$

where the first inequality follows from Theorem 1 of Chang (1999). Additionally,

$$|2n\mu^T\Sigma_{\cdot 1}D\Sigma_{11}D\Sigma_{1\cdot}\mu| \leq 2n||\Sigma_{11}\mu||_2^2||\Sigma_{11}||_2||D||_2^2 \ .$$

Hence, $\mathrm{Var}(T_1 - \tilde{T}_1) = O(||D||_2^2)$.

Recall that the quantity $||D||_2$ depends upon our choice of $j = 1$. In order to have a general result, applicable to $T_j - \tilde{T}_j$ for all j, we prefer a bound in terms of the overall error $\Delta = \tilde{\Omega} - \Omega$. Without loss of generality, define a function $f(\Omega) = \Omega_{11} - \Omega_{1\cdot}\Omega_{\cdot\cdot}^{-1}\Omega_{\cdot 1}$. That this function is Lipschitz smooth is straightforward to show, as mentioned previously. As a result,

$$||D||_2 = ||f(\Omega) - f(\tilde{\Omega})||_2 \leq K||\Omega - \tilde{\Omega}||_2 = K||\Delta||_2 \ .$$

The results of the corollary then follow.

A.4 Additional simulations

Additional simulations are provided to demonstrate predictive ability at lower signal-to-noise (SNR) thresholds. Comparisons across values of ρ_{in} and ρ_{out} in the main paper were shown with SNR = 0.20; these additional simulations show SNR = 0.10 and SNR = 0.05.

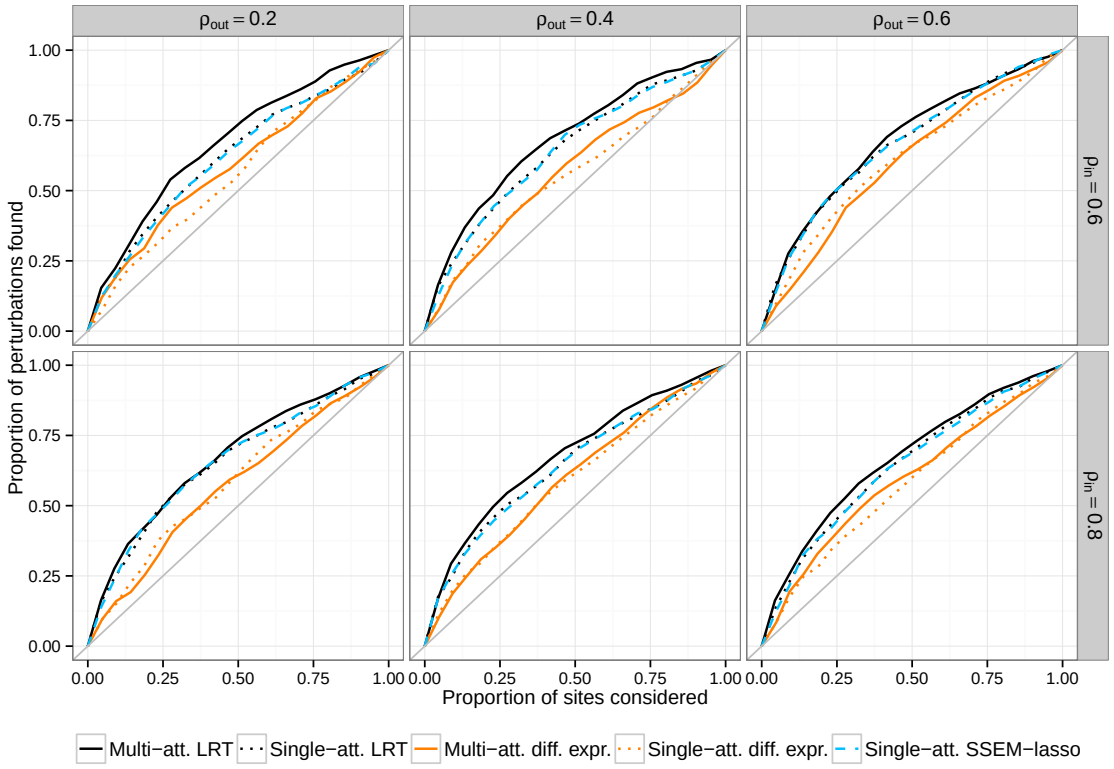


Figure A.1: Single-site recovery from a stochastic block model simulation with $p = 20$ nodes, $n = 50$ cases and controls, and SNR $= 0.05$.

Table A.1: Probability that the top-ranked site is the true perturbation site and (AUC) for simulations shown in Figure A.1. ($SNR = 0.10$)

		LRT methods		Differential expression		SSEM-lasso
ρ_{in}	ρ_{out}	Multi-att.	Single-att.	Multi-att.	Single-att.	Single-att.
0.8	0.2	0.14 (0.61)	0.12 (0.61)	0.10 (0.58)	0.11 (0.55)	0.11 (0.60)
	0.4	0.21 (0.65)	0.14 (0.63)	0.12 (0.59)	0.12 (0.57)	0.12 (0.62)
	0.6	0.16 (0.68)	0.14 (0.63)	0.09 (0.58)	0.08 (0.55)	0.14 (0.62)
0.6	0.2	0.21 (0.70)	0.15 (0.65)	0.14 (0.64)	0.14 (0.58)	0.17 (0.65)
	0.4	0.09 (0.64)	0.11 (0.64)	0.09 (0.58)	0.10 (0.57)	0.11 (0.63)
	0.6	0.15 (0.67)	0.11 (0.64)	0.07 (0.55)	0.09 (0.55)	0.12 (0.63)

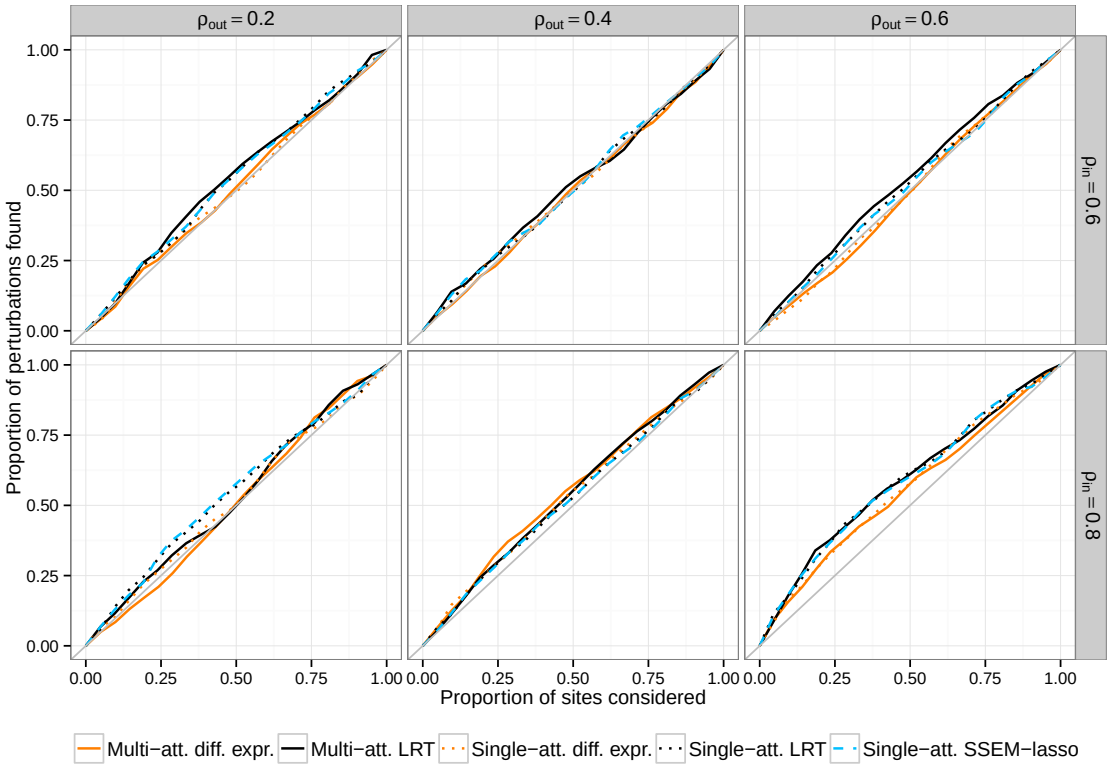


Figure A.2: Single-site recovery from a stochastic block model simulation with $p = 20$ nodes, $n = 50$ cases and controls, and SNR $= 0.05$.

.

Table A.2: Probability that the top-ranked site is the true perturbation site and (AUC) for simulations shown in Figure A.2. ($SNR = 0.10$)

		LRT methods		Differential expression		SSEM-lasso
ρ_{in}	ρ_{out}	Multi-att.	Single-att.	Multi-att.	Single-att.	Single-att.
0.8	0.2	0.09 (0.52)	0.09 (0.51)	0.05 (0.52)	0.06 (0.52)	0.07 (0.51)
	0.4	0.09 (0.55)	0.08 (0.55)	0.07 (0.54)	0.05 (0.55)	0.07 (0.54)
	0.6	0.07 (0.55)	0.08 (0.54)	0.08 (0.48)	0.05 (0.48)	0.07 (0.54)
0.6	0.2	0.10 (0.56)	0.08 (0.57)	0.07 (0.52)	0.06 (0.52)	0.09 (0.56)
	0.4	0.06 (0.55)	0.09 (0.52)	0.08 (0.54)	0.07 (0.56)	0.07 (0.52)
	0.6	0.09 (0.56)	0.10 (0.53)	0.07 (0.52)	0.06 (0.52)	0.10 (0.54)

Appendix B

Supplementary materials: "Characterizing cellular phenotypes via Bayesian regression in the Gene Ontology"

B.1 Software

The `ontoreg` package implementing this model in `R` is available at is available at `https://github.com/paulajgriffin/ontoreg`.

B.2 Posterior distributions

Conjugate priors have been used throughout, so all conditional posteriors have closed-form solutions. As several of these updates follow the same form, we provide a template and the substitutions necessary for each individual update.

B.2.1 Variance (scalar)

Prior distribution:

$$\sigma^2 \sim \text{Inv. Gamma}(k, l) \tag{B.1}$$

$$p(\sigma^2) = \frac{l^k}{\Gamma(k)}(\sigma^2)^{-k-1} \exp\left(\frac{-l}{\sigma^2}\right) \tag{B.2}$$

Data distribution:

$$y_1, \ldots, y_n | \sigma^2 \sim N(\mu, \sigma^2) \tag{B.3}$$

$$p(y_1, \ldots, y_n | \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(\frac{-\sum_i (y_i - \mu)^2}{2\sigma^2} \right) \tag{B.4}$$

Posterior:

$$p(\sigma^2 | y_1, \ldots, y_n) \propto (\sigma^2)^{-k-1-n/2} \exp\left(\frac{-l}{\sigma^2} + \frac{-\sum_i (y_i - \mu)^2}{2\sigma^2} \right) \tag{B.5}$$

$$\implies \sigma^2 | y_1, \ldots, y_n \sim \text{Inv. Gamma}\left(k + \frac{n}{2}, l + \frac{\sum_i (y_i - \mu)^2}{2} \right) \tag{B.6}$$

Instances in which this substitution is used:

1. Observation-level error (σ^2_{obs})
   - $k = k_{obs}$
   - $l = l_{obs}$
   - $n = G \times F$
   - $\sum_i (y_i - \mu)^2 = \sum_{g,f} (Y_{g,f} - (\alpha^{gene-cond}_{g,f} + \beta^{gene-cond}_{g,f} X_{g,f}))^2$
2. Gene-condition level error (intercept error $\sigma^2_{\alpha;gene-cond}$)
   - $k = k_{gene}$
   - $l = l_{gene}$
   - $n = G \times F$
   - $\sum_i (y_i - \mu)^2 = \sum_f (\alpha^{gene-cond}_f - \gamma \alpha^{GO-cond}_f)^2$
3. Gene-condition level error (coefficient error $\sigma^2_{\beta;gene-cond}$)
   - $k = k_{gene}$
   - $l = l_{gene}$
   - $n = G \times F$
   - $\sum_i (y_i - \mu)^2 = \sum_f (\beta^{gene-cond}_f - \gamma \beta^{GO-cond}_f)^2$
4. GO overall error (intercept error $\sigma^2_{\alpha;GO}$)

- $k = k_{GO}$
- $l = l_{GO}$
- $n = L$
- $\sum_i (y_i - \mu)^2 = \sum_L (\alpha_L^{GO})^2$

5. GO overall error (coefficient error $\sigma^2_{\beta;GO}$)

- $k = k_{obs}$
- $l = l_{obs}$
- $n = L$
- $\sum_i (y_i - \mu)^2 = \sum_L (\beta_L^{GO})^2$

B.2.2 Covariance (matrix)

Prior distribution:

$$\Sigma \sim \text{Inv. Wishart}(\nu, \Psi) \tag{B.7}$$

$$p(\Sigma) = \frac{|\Psi|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\nu/2)} |\Sigma|^{(-\nu+p+1)/2} \exp\left(-\frac{1}{2}\text{tr}(\Psi\Sigma^{-1})\right) \tag{B.8}$$

Data distribution:

$$\mathbf{y}_1, \ldots, \mathbf{y}_n | \Sigma \sim N(\mu, \Sigma) \tag{B.9}$$

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_n | \Sigma) = \left(\frac{1}{\sqrt{(2\pi)^n |\Sigma|}}\right)^n \exp\left(-\frac{1}{2}\sum_i (\mathbf{y}_i - \mu)^T \Sigma^{-1} (\mathbf{y}_i - \mu)\right) \tag{B.10}$$

Posterior:

$$p(\Sigma | \mathbf{y}_1, \ldots, \mathbf{y}_n) \propto |\Sigma|^{(-\nu+p+1+n)/2} \exp\left(-\frac{1}{2}\text{tr}\left(\Psi\Sigma^{-1} + \sum_i (\mathbf{y}_i - \mu)(\mathbf{y}_i - \mu)^T \Sigma^{-1}\right)\right) \tag{B.11}$$

$$\implies \Sigma | \mathbf{y}_1, \ldots, \mathbf{y}_n \sim \text{Inv. Wishart}\left(\nu + n, \Psi + \sum_i (\mathbf{y}_i - \mu)(\mathbf{y}_i - \mu)^T\right) \tag{B.12}$$

Instances in which this substitution is used:

1. GO-condition covariance (intercept $\Sigma_{\alpha;GO-cond}$)

   - $\nu = \nu_{GO}$
   - $\Psi = \Sigma_{MRCA}$
   - $n = F$
   - $\sum_i (\mathbf{y}_i - \mu)(\mathbf{y}_i - \mu)^T = \sum_f (\alpha_f^{GO-cond} - \alpha^{GO})(\alpha_f^{GO-cond} - \alpha^{GO})^T$

2. GO-condition covariance (coefficient $\Sigma_{\beta;GO-cond}$)

   - $\nu = \nu_{GO}$
   - $\Psi = \Sigma_{MRCA}$
   - $n = F$
   - $\sum_i (\mathbf{y}_i - \mu)(\mathbf{y}_i - \mu)^T = \sum_f (\beta_f^{GO-cond} - \beta^{GO})(\beta_f^{GO-cond} - \beta^{GO})^T$

B.2.3 Mean (vector)

Prior distribution:

$$\theta \sim N(\mu, \Sigma_\theta) \tag{B.13}$$

$$p(\theta) = \left(\frac{1}{\sqrt{(2\pi)|\Sigma_\theta|}} \right) \exp\left(-\frac{1}{2}(\theta - \mu)^T \Sigma_\theta^{-1} (\theta - \mu) \right) \tag{B.14}$$

Data distribution (A is a mapping matrix, constant):

$$x_1, \ldots, x_n | \theta \sim N(A\theta, \Sigma_x) \tag{B.15}$$

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n | \theta) = \left(\frac{1}{\sqrt{(2\pi)|\Sigma_x|}} \right) \exp\left(-\frac{1}{2} \sum_i (x_i - A\theta)^T \Sigma_x^{-1} (x_i - A\theta) \right) \tag{B.16}$$

Posterior:

$$p(\theta | \mathbf{x}_1, \ldots, \mathbf{x}_n) \propto \exp\left(-\frac{1}{2} \left((\theta - \mu)^T \Sigma_\theta^{-1} (\theta - \mu) + \sum_i (x_i - A\theta)^T \Sigma_x^{-1} (x_i - A\theta) \right) \right) \tag{B.17}$$

$$\implies \theta | x_1, \ldots, x_n \sim N\left(\left(\Sigma_\theta^{-1} + nA^T \Sigma_X^{-1} A \right)^{-1} \left(\Sigma_\theta^{-1} \mu + nA^T \Sigma_X^{-1} \bar{x} \right), \left(\Sigma_\theta^{-1} + nA^T \Sigma_X^{-1} A \right)^{-1} \right) \tag{B.18}$$

Instances in which this substitution is used:

1. Gene-condition means (combined vector; $\theta = [(\alpha_f^{gene-cond})^T, (\beta_f^{gene-cond})^T]$)
   - $\mu = \left[(\gamma\alpha_f^{GO-cond})^T, (\gamma\beta_f^{GO-cond})^T\right]$
   - $\Sigma_\theta = \begin{pmatrix} I\sigma^2_{\alpha;gene-cond} & 0 \\ 0 & I\sigma^2_{\beta;gene-cond} \end{pmatrix}$
   - $n = F \times R$
   - $A = [DX]$, where D is a design matrix that indicates the genes to which Y corresponds
   - $\Sigma_x = \text{diag}(\sigma^2_{obs})$
   - $x_i = y_i$, the observed expression data
2. GO-condition means (intercept $\alpha_f^{GO-cond}$)
   - $\mu = \alpha^{GO}$
   - $\Sigma_\theta = \Sigma^{\alpha;GO-cond}$
   - $n = 1$
   - $A = \gamma$
   - $\Sigma_x = I\sigma^2_{\alpha;gene-cond}$
   - $x_i = \alpha_f^{gene-cond}$
3. GO-condition means (coefficient $\beta_f^{GO-cond}$)
   - $\mu = \beta^{GO}$
   - $\Sigma_\theta = \Sigma^{\beta;GO-cond}$
   - $n = 1$
   - $A = \gamma$
   - $\Sigma_x = I\sigma^2_{\beta;gene-cond}$
   - $x_i = \beta_f^{gene-cond}$
4. GO overall mean (intercept α^{GO})
   - $\mu = 0$
   - $\Sigma_\theta = I\sigma^2_{\alpha,GO}$
   - $n = F$
   - $A = I$

- $\Sigma_x = \Sigma_{\alpha;GO-cond}$
- $x_i = \alpha_f^{GO-cond}$

5. GO overall mean (coefficient β^{GO})

- $\mu = 0$
- $\Sigma_\theta = I\sigma^2_{\beta_GO}$
- $n = F$
- $A = I$
- $\Sigma_x = \Sigma_{\beta;GO-cond}$
- $x_i = \beta_f^{GO-cond}$

B.3 Additional model details & extended results

This section presents extended results for the ontological regression, linear pooling, and HDP models described in the main paper.

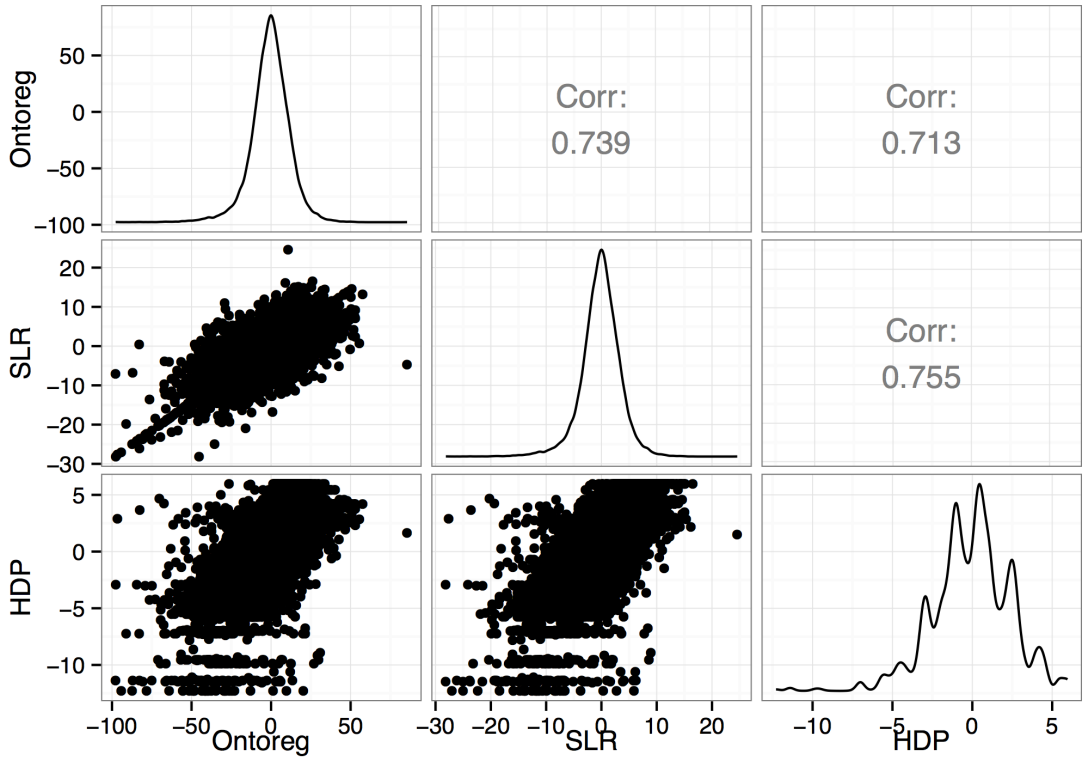


Figure B.1: Distribution of gene-by-factor slopes according to each of the three models, and correlations between them. Correlation between methods indicates some consistency between methods. Ontological regression results in a fatter-tailed distribution than linear pooling, and both ontological regression and linear pooling result in smoother distributions than HDP. An overlaid plot is given in Figure 3.2

B.3.1 Ontological regression

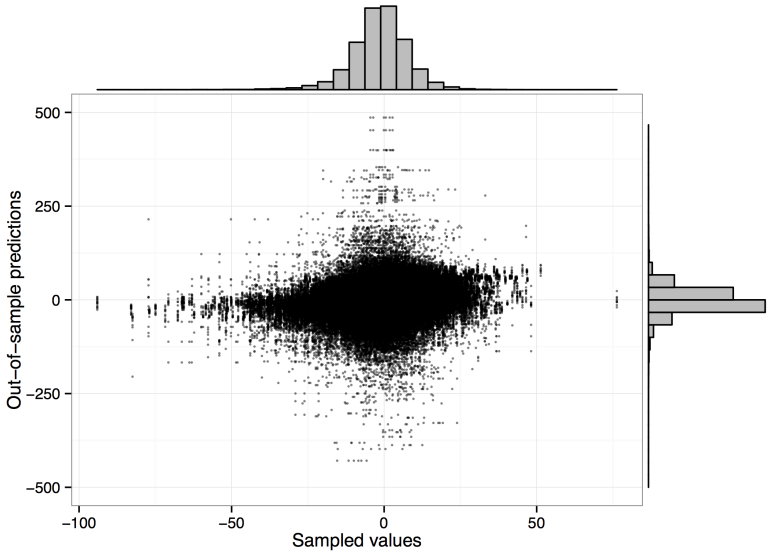


Figure B.2: Out-of-sample predicted coefficients vs sampled coefficients (average across all leave-out proportions; $r = 0.22$).

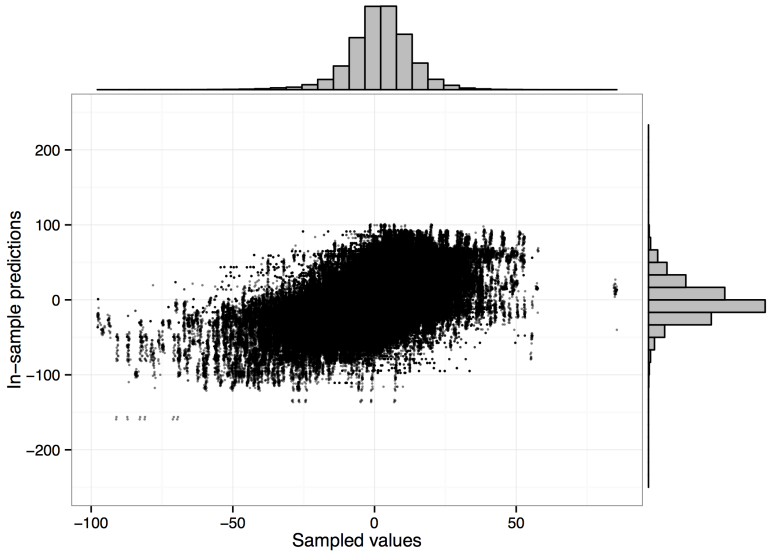


Figure B.3: In-sample predicted coefficients vs sampled coefficients (average across all leave-out proportions; $r = .56$).

	GOID	Term	Mean coefficient
1	GO:0050896	response to stimulus	-78.16
2	GO:0006950	response to stress	-61.35
3	GO:0006412	translation	50.80
4	GO:0044249	cellular biosynthetic process	50.78
5	GO:0044260	cellular macromolecule metabolic process	46.98
6	GO:0044267	cellular protein metabolic process	46.32
7	GO:0009059	macromolecule biosynthetic process	44.65
8	GO:0043170	macromolecule metabolic process	40.29
9	GO:0019538	protein metabolic process	39.97
10	GO:0009056	catabolic process	-37.66
11	GO:0042254	ribosome biogenesis	35.05
12	GO:0009987	cellular process	-34.42
13	GO:0006396	RNA processing	29.38
14	GO:0044248	cellular catabolic process	-27.59
15	GO:0006364	rRNA processing	23.45
16	GO:0016072	rRNA metabolic process	23.40
17	GO:0050794	regulation of cellular process	-22.65
18	GO:0016070	RNA metabolic process	21.88
19	GO:0007154	cell communication	-21.18
20	GO:0005975	carbohydrate metabolic process	-18.97
21	GO:0042221	response to chemical	-17.73
22	GO:0044262	cellular carbohydrate metabolic process	-17.12
23	GO:0006810	transport	-16.29
24	GO:0006508	proteolysis	-15.67
25	GO:0006091	generation of precursor metabolites and energy	-14.95

Table B.1: Top 25 GO terms by ontological regression.

B.3.2 Linear pooling

We use the term linear pooling to describe a process for summarizing the results of a series of simple linear regressions.

1. Fit a series of regressions for $f = 1, \ldots, F$ conditions and $g = 1, \ldots, G$ genes, to obtain $(\hat{\alpha}_{f,g}, \hat{\beta}_{f,g})$, the ordinary least-squares estimators for the regression $Y_i = \alpha_{f,g} + \beta_{f,g} x_i$.
2. For each term $l = 1, \ldots, L$ of the GO terms to which the genes have been mapped, calculate the average θ_l of all $\hat{\beta}_{f,g}$ across all conditions f and genes mapped to l.
3. Calculate p-values testing the hypothesis that $\theta_l \neq 0$ against $\theta_l = 0$

	GOID	Term	Mean coefficient
1	GO:0042254	ribosome biogenesis	135.23
2	GO:0006412	translation	156.20
3	GO:0006364	rRNA processing	96.53
4	GO:0016072	rRNA metabolic process	95.88
5	GO:0006396	RNA processing	117.40
6	GO:0050896	response to stimulus	-174.21
7	GO:0042274	ribosomal small subunit biogenesis	55.44
8	GO:0030490	maturation of SSU-rRNA	49.28
9	GO:0006950	response to stress	-127.13
10	GO:0009991	response to extracellular stimulus	-48.15
11	GO:0009605	response to external stimulus	-47.77
12	GO:0007154	cell communication	-73.97
13	GO:0044262	cellular carbohydrate metabolic process	-57.48
14	GO:0009056	catabolic process	-125.09
15	GO:0042273	ribosomal large subunit biogenesis	36.74
16	GO:0006399	tRNA metabolic process	39.21
17	GO:0009451	RNA modification	26.55
18	GO:0044248	cellular catabolic process	-99.01
19	GO:0006357	reg. of transcr. from RNA polymerase II promoter	-61.65
20	GO:0044267	cellular protein metabolic process	124.54
21	GO:0044249	cellular biosynthetic process	138.23
22	GO:0042594	response to starvation	-30.23
23	GO:0006091	generation of precursor metabolites and energy	-44.70
24	GO:0015980	energy derivation by oxidation of organic compounds	-42.05
25	GO:0042255	ribosome assembly	22.82

Table B.2: Top 25 GO terms by linear pooling analysis.

B.3.3 Hierarchical Dirichlet Process

The hierarchical Dirichlet process (HDP) offers a nonparametric approach to modeling this data. The basic intuition behind this is that genes are grouped into clusters, from which coefficients are drawn. The set of clusters is determined by a Dirichlet process.

The top 25 GO terms (summarized according to the same methodology as the linear pooling results are shown in Table B.3. The BUGS model that demonstrates the structure of this model is provided in Figure B.4

	GOID	Term	Mean coefficient
1	GO:0044249	cellular biosynthetic process	4.85
2	GO:0050896	response to stimulus	-4.47
3	GO:0006412	translation	3.90
4	GO:0044260	cellular macromolecule metabolic process	3.59
5	GO:0009059	macromolecule biosynthetic process	3.58
6	GO:0044267	cellular protein metabolic process	3.52
7	GO:0006950	response to stress	-3.30
8	GO:0043170	macromolecule metabolic process	3.18
9	GO:0019538	protein metabolic process	3.17
10	GO:0044237	cellular metabolic process	2.99
11	GO:0007005	mitochondrion organization	2.38
12	GO:0042254	ribosome biogenesis	2.26
13	GO:0009056	catabolic process	-2.06
14	GO:0006396	RNA processing	1.91
15	GO:0044248	cellular catabolic process	-1.71
16	GO:0016070	RNA metabolic process	1.55
17	GO:0016072	rRNA metabolic process	1.52
18	GO:0006364	rRNA processing	1.50
19	GO:0007154	cell communication	-1.47
20	GO:0050794	regulation of cellular process	-1.41
21	GO:0009987	cellular process	1.29
22	GO:0006807	nitrogen compound metabolic process	1.26
23	GO:0042221	response to chemical	-1.23
24	GO:0006996	organelle organization	1.15
25	GO:0006508	proteolysis	-1.12

Table B.3: Top 25 GO terms by HDP analysis.

```
model{
  #Data - Y, X
  # Gene level Dirichlet Process prior (C is cutoff parameter)
  # Precision Parameter
  alpha ~ dexp(0.1)

  # Constructive DPP (via stick breaking)
  p[1] <- r[1]
  for (j in 2 : C) {
      p[j] <- r[j] * (1 - r[j - 1]) * p[j -1 ] / r[j - 1]
  }
  p.sum <- sum(p[])
  for (j in 1:C){
    taupri[j] ~ dexp(0.1)
    for(k in 1:M){
      theta[j,k,1] ~ dnorm(0,taupri[j])
      theta[j,k,2] ~ dnorm(0,taupri[j])
    }
    r[j] ~ dbeta(1, alpha)

    # Scaling to ensure sum to 1
    pi[j] <- p[j] / p.sum
  }

  # Gene level
  for( i in 1 : N ) {
    # Draw from DPP.
    S2[i] ~ dcat(pi[])
    for(j in 1:M){
      # Nutrient level
      tau[i,j] ~ dgamma(5,0.1)
      mu[i,j,1] <- theta[S2[i],j,1]
      mu[i,j,2] <- theta[S2[i],j,2]

      # Data
      # mu[i,j,2] intercept
      # mu[i,j,1] coefficient
      for(k in 1:L){
        pmean[i,j,k] <- X[i,j,k]*mu[i,j,1] + mu[i,j,2]
        Y[i,j,k] ~ dnorm(pmean[i,j,k],tau[i,j])
      }
    }
  }
}
```

Figure B.4: BUGS model code for the HDP.

B.4 Sensitivity analysis

We perform additional analyses on the Brauer dataset to determine the sensitivity of our results to the assignment of genes to GO terms. We randomly select 1%, 5%, and 10% of gene-GO assignments and move the link to a different GO term. The scatterplots below show the correlation between of the overall GO-level estimates α^{GO} and β^{GO} estimated from the full data with the true γ, against those obtained from an altered assignment matrix $\tilde{\gamma}$.

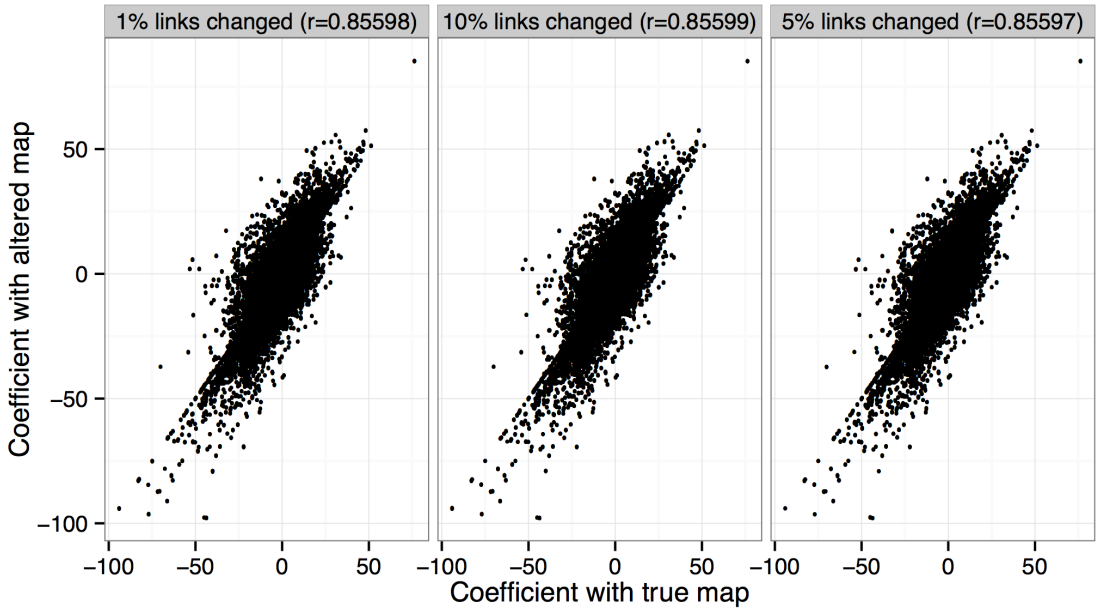


Figure B.5: Scatterplots showing average gene-condition coefficients from a correct map according to GO against maps in which 1%, 5%, or 10% of connections have been altered to have a different GO term endpoint. Correlation with the full model is fairly consistent across percentages of edges altered, which suggests that this method is somewhat robust to deviations in the GO map at these levels.

Appendix C

Supplementary materials: "Prediction of drug sensitivity by gene signature activation patterns"

C.1 Software

Software to reproduce analyses and figures in this paper is available at is available at `https://github.com/paulajgriffin/drug_response_pathways`.

C.2 Detailed AUC results

This section includes additional details of model performance.

Model \Comparison	Subtype	*AKT*	*BAD*	*HER2*	*IGF1R*	Main	Interact
Subtype only		13	15	16	21	1	8
Subtype and *AKT*	69		46	36	43	3	14
Subtype and *BAD*	65	35		30	37	3	14
Subtype and *HER2*	64	44	51		46	4	13
Subtype and *IGF1R*	61	36	41	34		4	11
Subtype and main effects	81	76	79	75	78		28
Full interaction lasso	74	67	67	69	71	52	

Table C.1: Relative performance all models considered. Each entry in this table is the number of times that the model on the row outperforms the model in the column. For example, the subtype and *AKT* model outperforms a subtype-only model in 69 cases out of a possible 82. Column names are shortened for brevity.

Figure C.1: AUC based on leave-one-out models for subtype-only models, subtype a single pathway (best of *AKT*, *BAD*, *HER2*, and *IGF1R* shown by color), subtype and all pathways, and the full interaction lasso. This is an extended version of Figure 4.1.

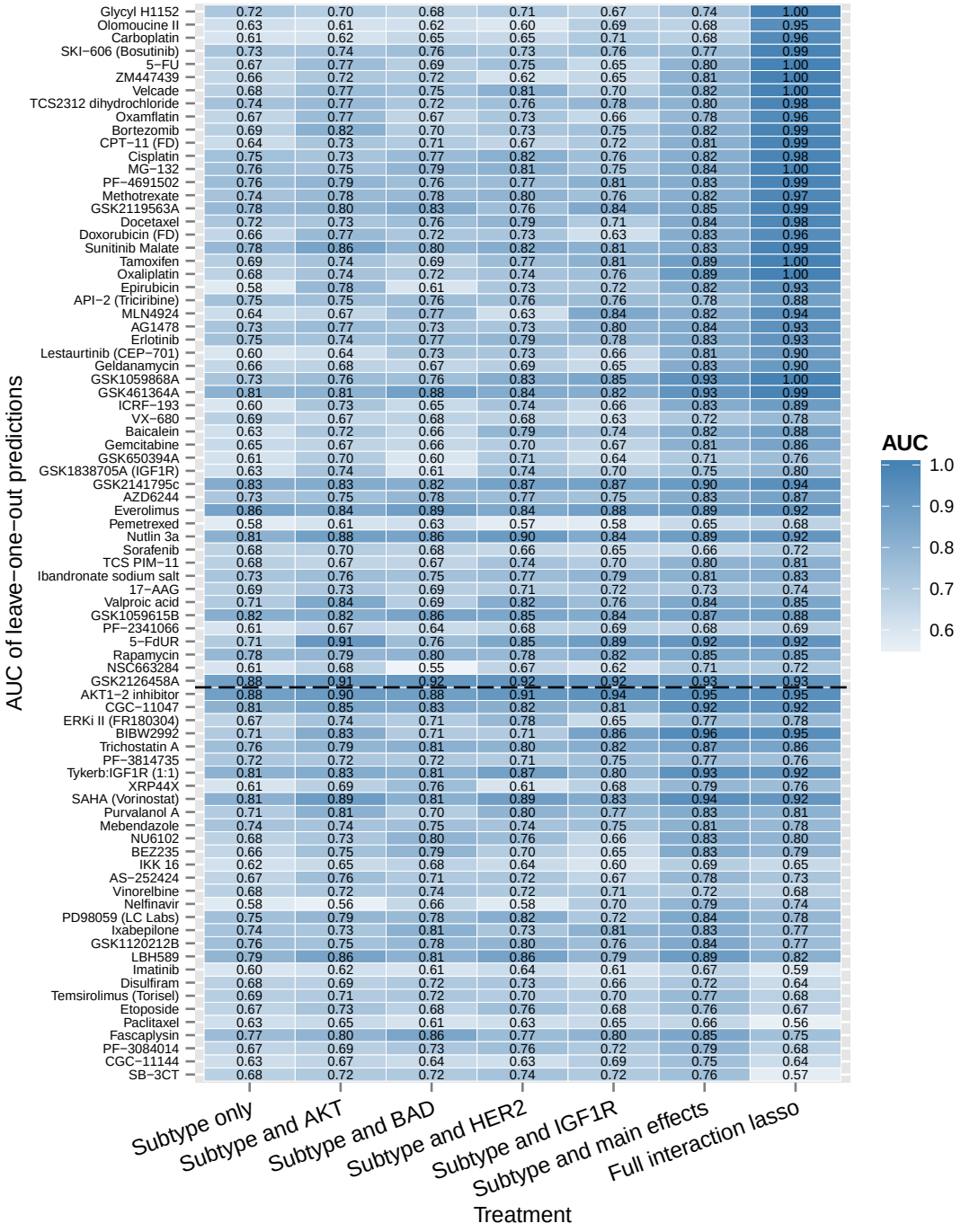


Figure C.2: AUC based on leave-one-out models for subtype-only models, subtype a single pathway (best of *AKT*, *BAD*, *HER2*, and *IGF1R* shown by color), subtype and all pathways, and the full interaction lasso. The horizontal line indicates the treatment for which AUC is no longer improved by interaction modeling.

C.3 Interaction model details

This section contains details of the interaction models described Chapter 4.

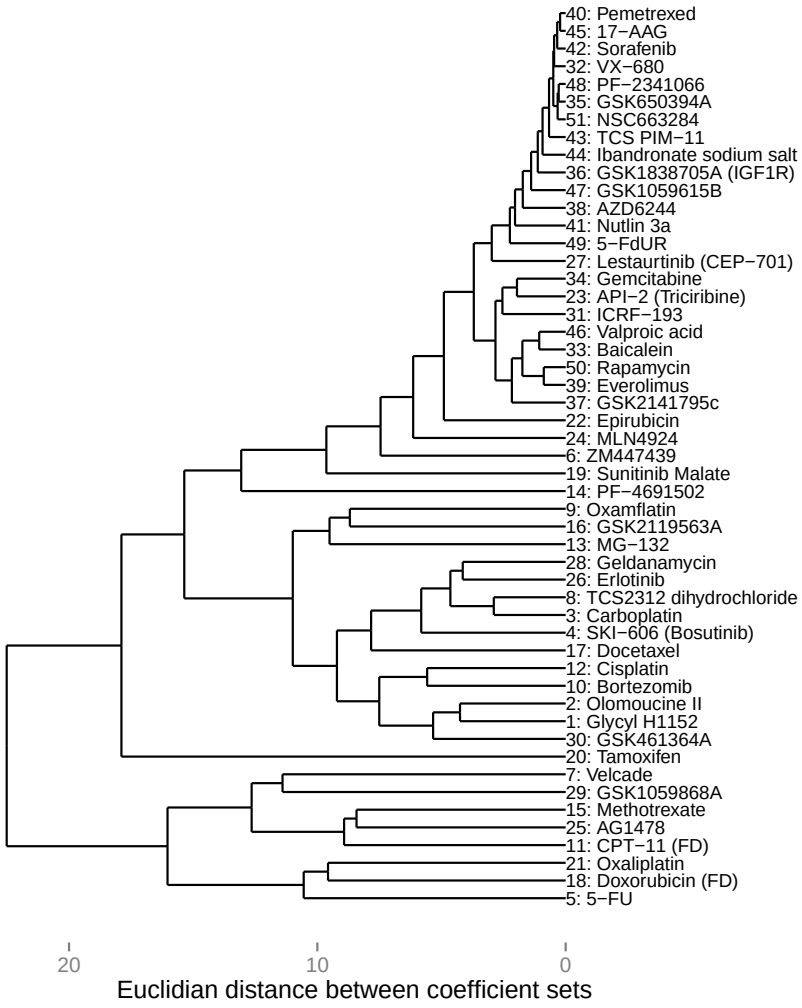


Figure C.3: A tree produced by hierarchical clustering of interaction models for which the interaction model provided superior performance. The number preceding each of the treatment names indicates the rank of improvement obtained by the interaction model (row number in Figure 4.1).

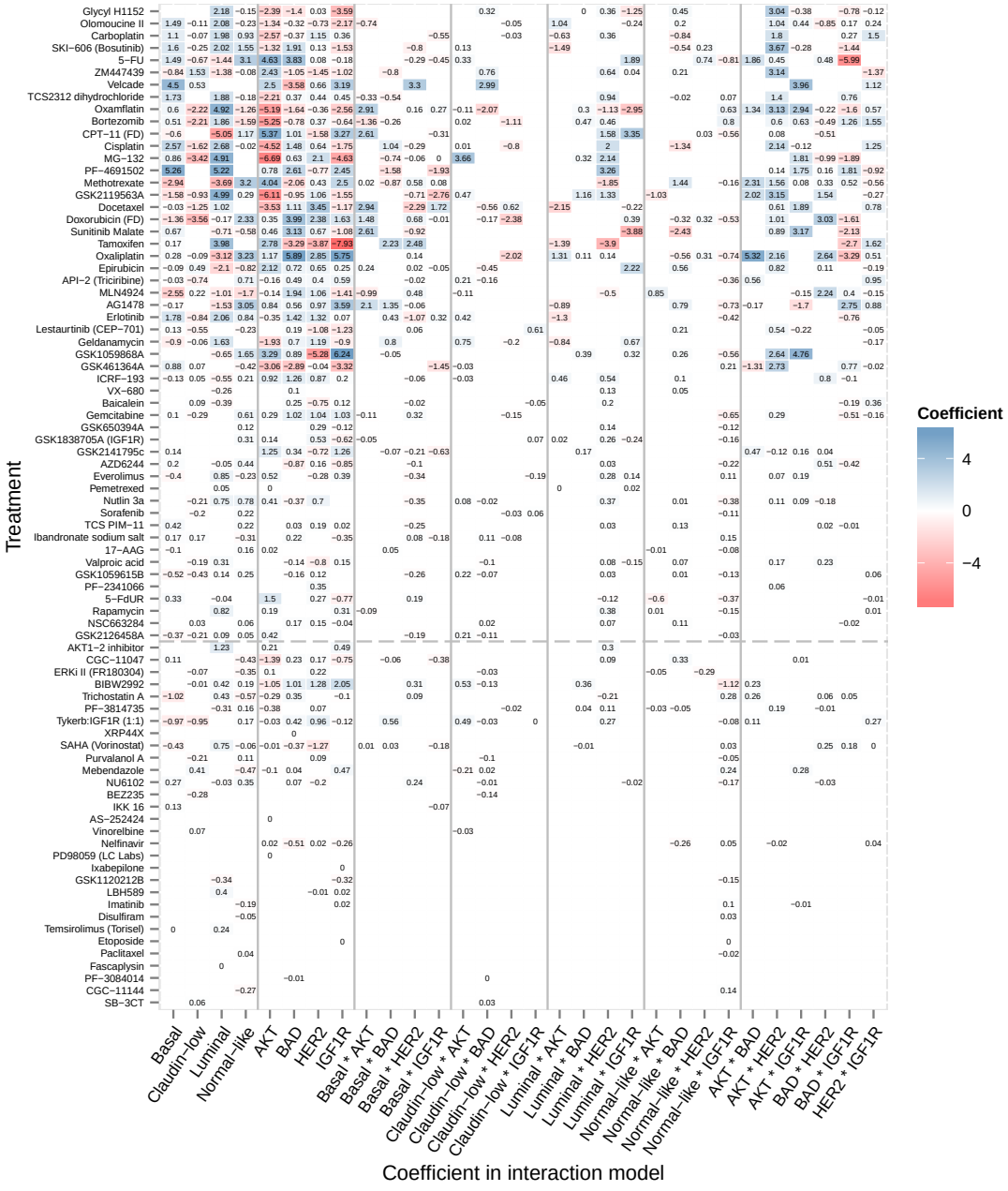
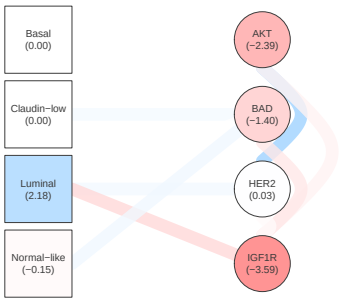
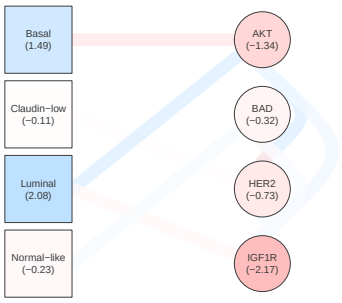


Figure C.4: Coefficients from the interaction lasso model, showing all treatments. Treatments below the dashed gray line performed worse under the interaction model than at least one of the other model types (subtype only, subtype and one pathway, or subtype and all pathways). This is an extended version of Figure 4.2.

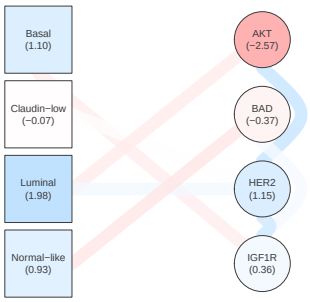
.



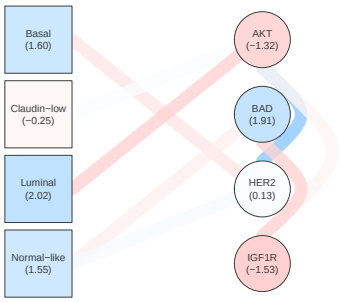
(a) Glycyl H1152 (AUC=1.00)



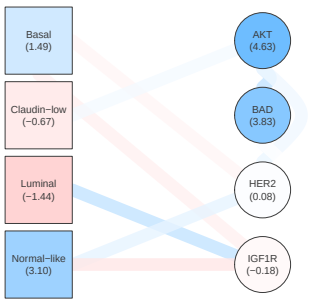
(b) Olomoucine II (AUC=0.95)



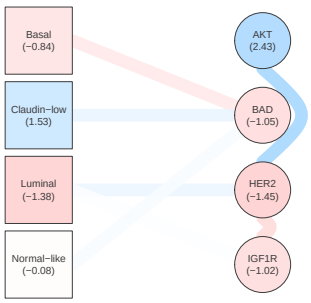
(c) Carboplatin (AUC=0.96)



(d) SKI-606 (Bosutinib) (AUC=0.99)

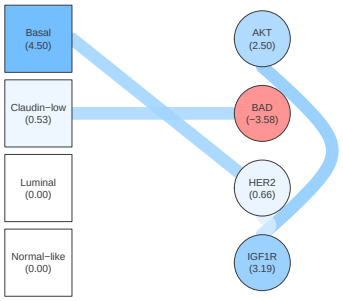


(e) 5-FU (AUC=1.00)

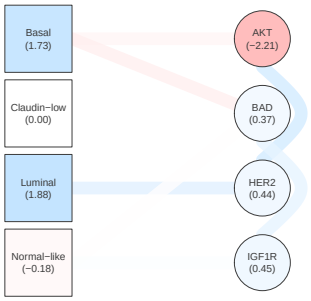


(f) ZM447439 (AUC=1.00)

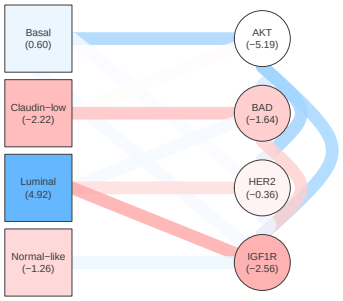
Figure C.5: Network representations of interaction models for response to treatments Glycyl H1152, Olomoucine II, Carboplatin, SKI-606 (Bosutinib), 5-FU, and ZM447439.



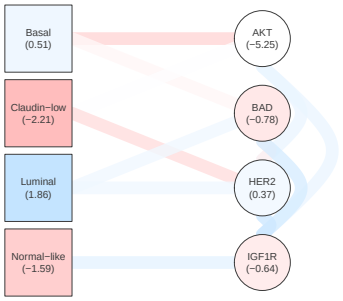
(a) Velcade (AUC=1.00)



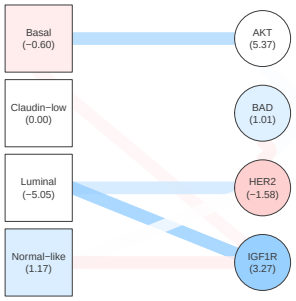
(b) TCS2312 dihydrochloride (AUC=0.98)



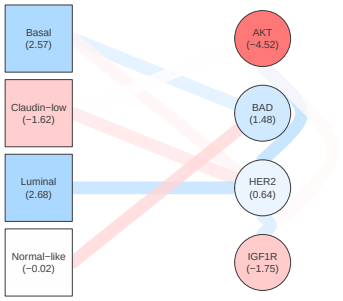
(c) Oxamflatin (AUC=0.96)



(d) Bortezomib (AUC=0.99)

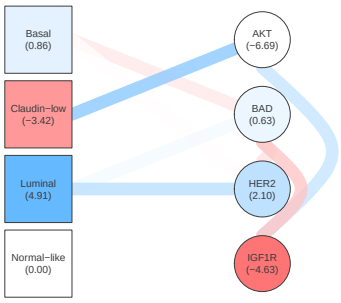


(e) CPT-11 (FD) (AUC=0.99)

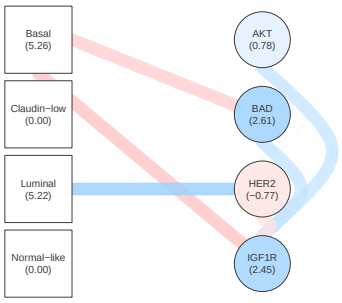


(f) Cisplatin (AUC=0.98)

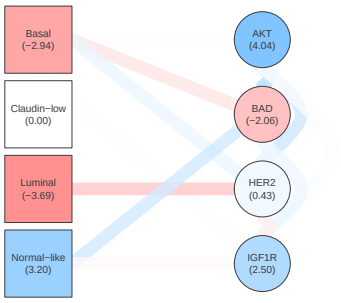
Figure C.6: Network representations of interaction models for response to treatments Velcade, TCS2312 dihydrochloride, Oxamflatin, Bortezomib, CPT-11 (FD), and Cisplatin.



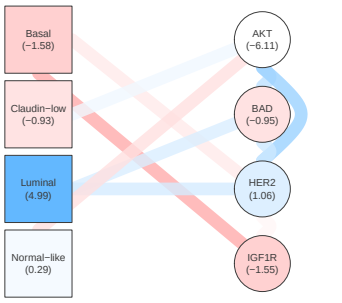
(a) MG-132 (AUC=1.00)



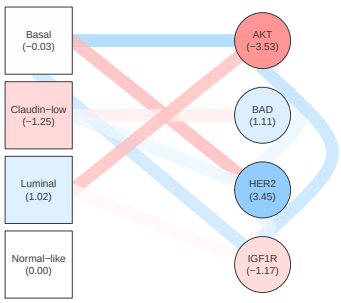
(b) PF-4691502 (AUC=0.99)



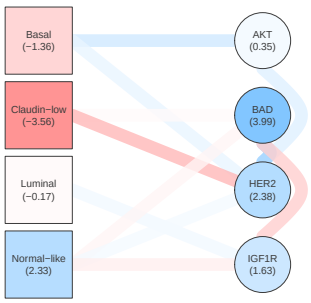
(c) Methotrexate (AUC=0.97)



(d) GSK2119563A (AUC=0.99)

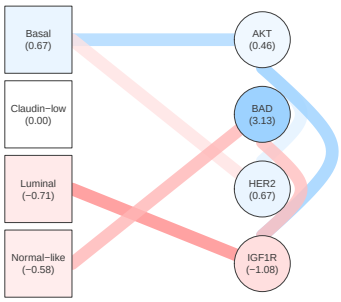


(e) Docetaxel (AUC=0.98)

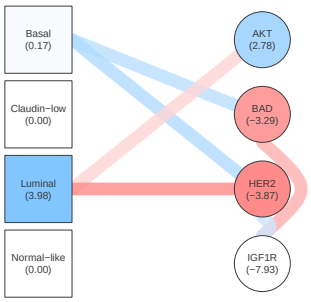


(f) Doxorubicin (FD) (AUC=0.96)

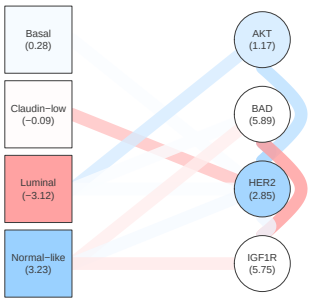
Figure C.7: Network representations of interaction models for response to treatments MG-132, PF-4691502, Methotrexate, GSK2119563A, Docetaxel, and Doxorubicin (FD).



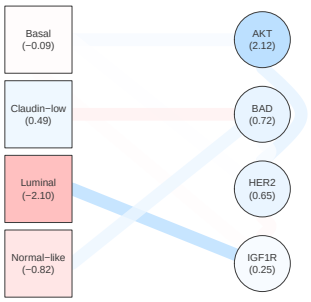
(a) Sunitinib Malate (AUC=0.99)



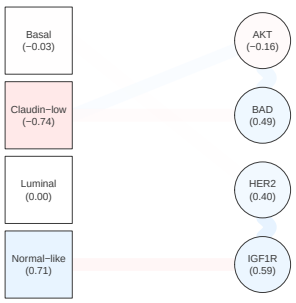
(b) Tamoxifen (AUC=1.00)



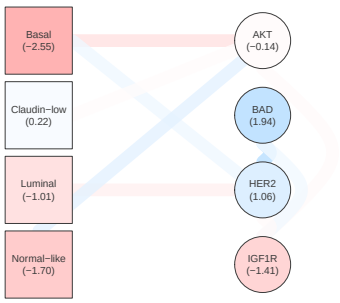
(c) Oxaliplatin (AUC=1.00)



(d) Epirubicin (AUC=0.93)

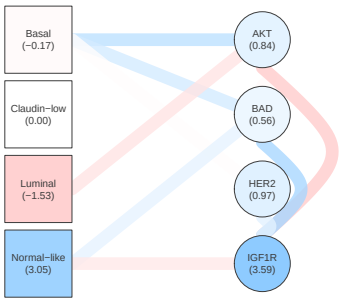


(e) API-2 (Triciribine) (AUC=0.88)

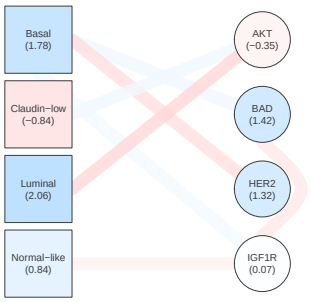


(f) MLN4924 (AUC=0.94)

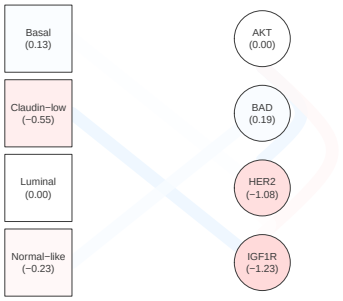
Figure C.8: Network representations of interaction models for response to treatments Sunitinib Malate, Tamoxifen, Oxaliplatin, Epirubicin, API-2 (Triciribine), and MLN4924.



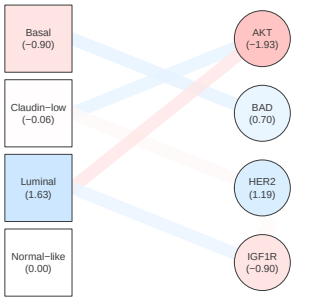
(a) AG1478 (AUC=0.93)



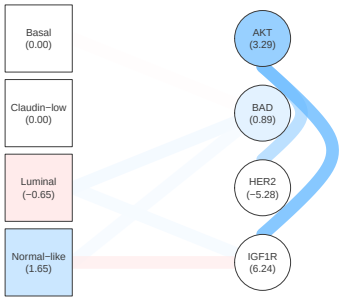
(b) Erlotinib (AUC=0.93)



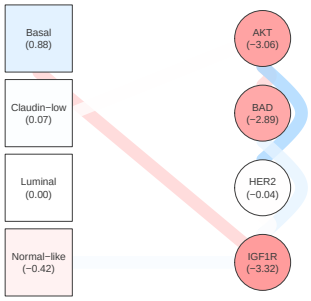
(c) Lestaurtinib (CEP-701) (AUC=0.90)



(d) Geldanamycin (AUC=0.90)

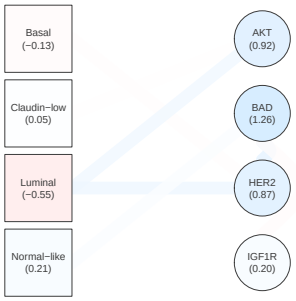


(e) GSK1059868A (AUC=1.00)

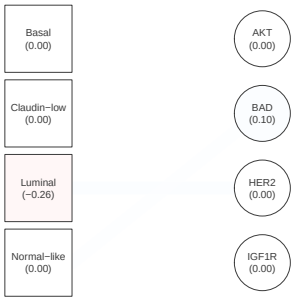


(f) GSK461364A (AUC=0.99)

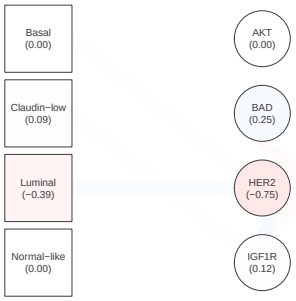
Figure C.9: Network representations of interaction models for response to treatments AG1478, Erlotinib, Lestaurtinib (CEP-701), Geldanamycin, GSK1059868A, and GSK461364A.



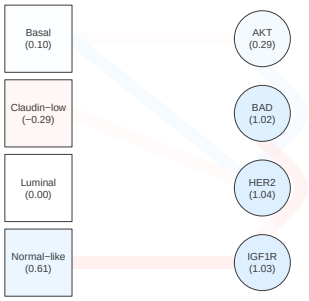
(a) ICRF-193 (AUC=0.89)



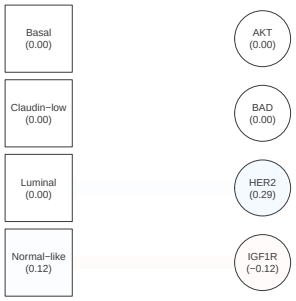
(b) VX-680 (AUC=0.78)



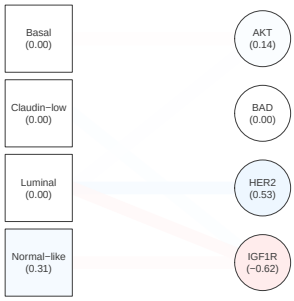
(c) Baicalein (AUC=0.88)



(d) Gemcitabine (AUC=0.86)

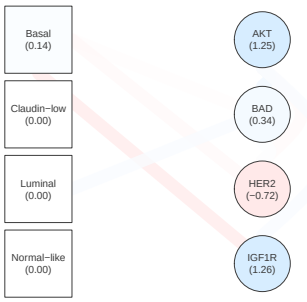


(e) GSK650394A (AUC=0.76)

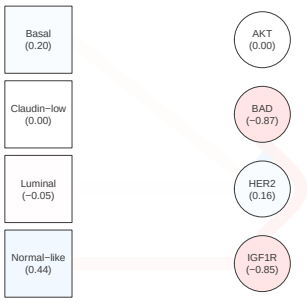


(f) GSK1838705A (IGF1R) (AUC=0.80)

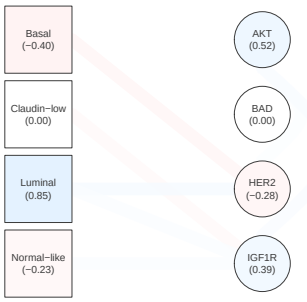
Figure C.10: Network representations of interaction models for response to treatments ICRF-193, VX-680, Baicalein, Gemcitabine, GSK650394A, and GSK1838705A (IGF1R).



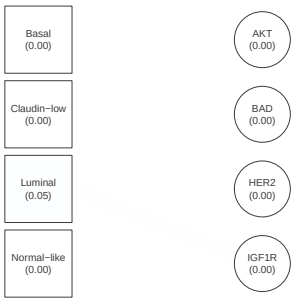
(a) GSK2141795c (AUC=0.94)



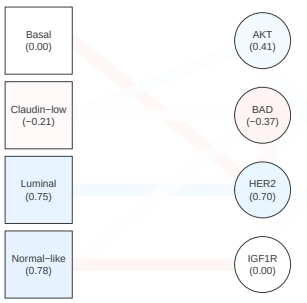
(b) AZD6244 (AUC=0.87)



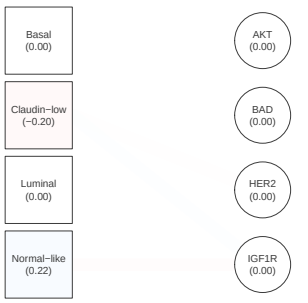
(c) Everolimus (AUC=0.92)



(d) Pemetrexed (AUC=0.68)

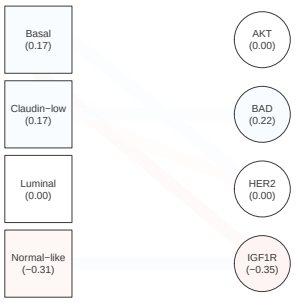


(e) Nutlin 3a (AUC=0.92)

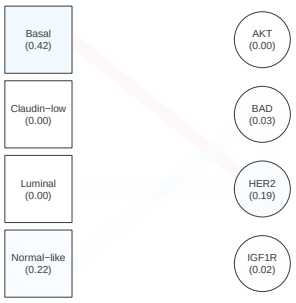


(f) Sorafenib (AUC=0.72)

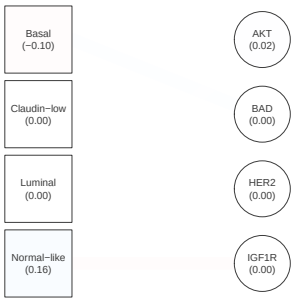
Figure C.11: Network representations of interaction models for response to treatments GSK2141795c, AZD6244, Everolimus, Pemetrexed, Nutlin 3a, and Sorafenib.



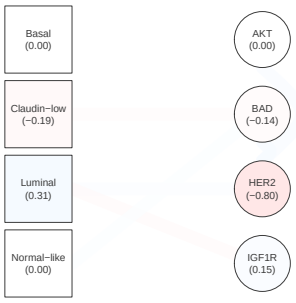
(a) Ibandronate sodium salt (AUC=0.83)



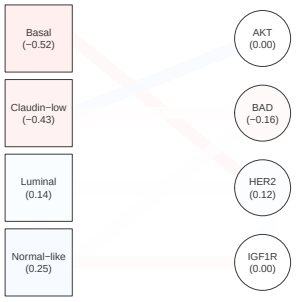
(b) TCS PIM-11 (AUC=0.81)



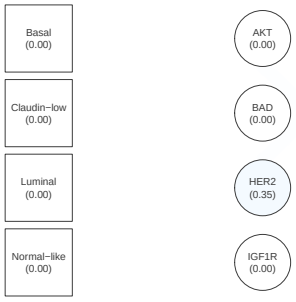
(c) 17-AAG (AUC=0.74)



(d) Valproic acid (AUC=0.85)

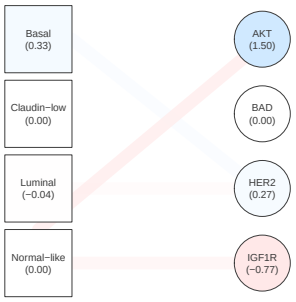


(e) GSK1059615B (AUC=0.88)

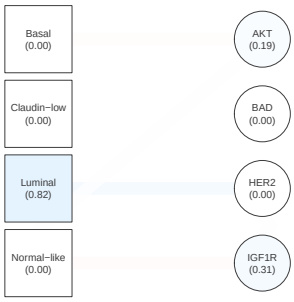


(f) PF-2341066 (AUC=0.69)

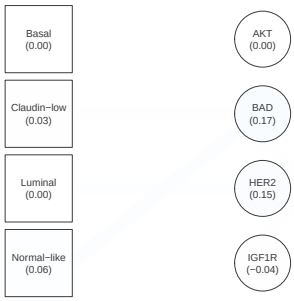
Figure C.12: Network representations of interaction models for response to treatments Ibandronate sodium salt, TCS PIM-11, 17-AAG, Valproic acid, GSK1059615B, and PF-2341066.



(a) 5-FdUR (AUC=0.92)



(b) Rapamycin (AUC=0.85)



(c) NSC663284 (AUC=0.72)

Figure C.13: Network representations of interaction models for response to treatments 5-FdUR, Rapamycin, and NSC663284.

Bibliography

Afzal, A. M., Mussa, H. Y., Turner, R. E., Bender, A., and Glen, R. C. (2014). Target fishing: A single-label or multi-label problem? *arXiv preprint arXiv:1411.6285* .

Airoldi, E. M., Huttenhower, C., Gresham, D., Lu, C., Caudy, A. A., Dunham, M. J., Broach, J. R., Botstein, D., and Troyanskaya, O. G. (2009). Predicting cellular growth from gene expression signatures. *PLoS Computational Biology* **5,** e1000257.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics* **25,** 25–29.

Baldessari, B. (1967). The distribution of a quadratic form of normal random variables. *The Annals of Mathematical Statistics* pages 1700–1704.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* pages 1165–1188.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). Bayesian factor regression models in the large p, small n paradigm. *Bayesian Statistics* **7,** 733–742.

Bien, J., Taylor, J., Tibshirani, R., et al. (2013). A lasso for hierarchical interactions. *The Annals of Statistics* **41,** 1111–1141.

Bien, J. and Tibshirani, R. (2014). *hierNet: A Lasso for Hierarchical Interactions.* R package version 1.6.

Bordbar, A., Mo, M. L., Nakayasu, E. S., Schrimpe-Rutledge, A. C., Kim, Y.-M., Metz, T. O., Jones, M. B., Frank, B. C., Smith, R. D., and Peterson, S. N. (2012). Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation. *Molecular Systems Biology* **8,** 558.

Brauer, M. J., Huttenhower, C., Airoldi, E. M., Rosenstein, R., Matese, J. C., Gresham, D., Boer, V. M., Troyanskaya, O. G., and Botstein, D. (2008). Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Molecular Biology of the Cell* **19,** 352–367.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70.

Chang, D.-W. (1999). A matrix trace inequality for products of Hermitian matrices. *Journal of Mathematical Analysis and Applications* **237,** 721–725.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95,** 759–771.

Cosgrove, E. J., Zhou, Y., Gardner, T. S., and Kolaczyk, E. D. (2008). Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. *Bioinformatics* **24,** 2482–2490.

Cressie, N. (1993). *Statistics for Spatial Data: Wiley Series in Probability and Statistics*. Wiley-Interscience New York.

Csermely, P., Korcsmáros, T., Kiss, H. J., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology & Therapeutics* **138,** 333–408.

Daemen, A., Griffith, O. L., Heiser, L. M., Wang, N. J., Enache, O. M., Sanborn, Z., Pepin, F., Durinck, S., Korkola, J. E., Griffith, M., et al. (2013). Modeling precision treatment of breast cancer. *Genome Biology* **14,** R110.

Danaher, P., Wang, P., and Witten, D. M. (2013). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

Dempster, A. P. (1972). Covariance selection. *Biometrics* pages 157–175.

di Bernardo, D., Thompson, M. J., Gardner, T. S., Chobot, S. E., Eastwood, E. L., Wojtovich, A. P., Elliott, S. J., Schaus, S. E., and Collins, J. J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature Biotechnology* **23,** 377–383.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11,** 587.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95,** 14863–14868.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640.

Fournier, M. L., Paulson, A., Pavelka, N., Mosley, A. L., Gaudenz, K., Bradford, W. D., Glynn, E., Li, H., Sardiu, M. E., Fleharty, B., et al. (2010). Delayed correlation of mRNA and protein expression in rapamycin-treated cells and a role for GGC1 in cellular sensitivity to rapamycin. *Molecular & Cellular Proteomics* **9,** 271–284.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **303,** 799–805.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* pages 457–472.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine* **366,** 883–892.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98,** 1–15.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5,** 109–137.

Hornstein, M., Hoffmann, M. J., Alexa, A., Yamanaka, M., Müller, M., Jung, V., Rahnenführer, J., and Schulz, W. A. (2008). Protein phosphatase and TRAIL receptor genes as new candidate tumor genes on chromosome 8p in prostate cancer. *Cancer Genomics-Proteomics* **5,** 123–136.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering* **16,** 1370–1386.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28,** 27–30.

Kolar, M., Liu, H., and Xing, E. P. (2014). Graph estimation from multi-attribute data. *The Journal of Machine Learning Research* **15,** 1713–1750.

Lecca, P. and Priami, C. (2013). Biological network inference for drug discovery. *Drug Discovery Today* **18,** 256–264.

Lee, M., Topper, S. E., Hubler, S. L., Hose, J., Wenger, C. D., Coon, J. J., and Gasch, A. P. (2011). A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. *Molecular Systems Biology* **7,**.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC bioinformatics* **12,** 323.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics* **40,** 2293–2326.

Lu, C., Latourelle, J., OConnor, G. T., Dupuis, J., and Kolaczyk, E. D. (2013). Network-guided sparse regression modeling for detection of gene-by-gene interactions. *Bioinformatics* **29,** 1241–1249.

Ma, H. and Zhao, H. (2012). iFad: an integrative factor analysis model for drug-pathway association inference. *Bioinformatics* **28,** 1911–1918.

MacNeil, S. M., Johnson, W. E., Li, D. Y., Piccolo, S. R., and Bild, A. H. (2015). Inferring pathway dysregulation in cancers from multiple types of omic data. *Genome Medicine* **7,** 1–12.

Pham, L., Christadore, L., Schaus, S., and Kolaczyk, E. D. (2011). Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis. *Proceedings of the National Academy of Sciences* **108,** 13347–13352.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* **23,** 2507–2517.

Shamir, R. and Sharan, R. (2001). Algorithmic approaches to clustering gene expression data. In *Current Topics in Computational Biology*. Citeseer.

Shen, Y., Bild, A. H., and Johnson, W. E. (2013). *ASSIGN: Adaptive Signature Selection and InteGratioN (ASSIGN)*. R package version 1.2.0.

Shen, Y., Rahman, M., Piccolo, S. R., Gusenleitner, D., EI-Chaar, N. N., Cheng, L., Monti, S., Bild, A. H., and Johnson, W. E. (2015). ASSIGN: Context-specific genomic profiling of multiple heterogeneous biological pathways. *Bioinformatics* page btv031.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545–15550.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* **39,** D561–D568.

Tan, W. (1977). On the distribution of quadratic forms in normal random variables. *Canadian Journal of Statistics* **5,** 241–250.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101,**.

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., and Botstein, D. (2003). A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences* **100,** 8348–8353.

Tsavachidou-Fenner, D., Tannir, N., Tamboli, P., Liu, W., Petillo, D., Teh, B., Mills, G., and Jonasch, E. (2010). Gene and protein expression markers of response to combined antiangiogenic and epidermal growth factor targeted therapy in renal cell carcinoma. *Annals of Oncology* **21,** 1599–1606.

Van De Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **347,** 1999–2009.

Varambally, S., Yu, J., Laxman, B., Rhodes, D. R., Mehra, R., Tomlins, S. A., Shah, R. B., Chandran, U., Monzon, F. A., Becich, M. J., et al. (2005). Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell* **8,** 393–406.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research* **38,** e178–e178.

Wang, L. and Wang, X. (2013). Hierarchical dirichlet process model for gene expression clustering. *EURASIP Journal on Bioinformatics and Systems Biology* **2013,** 1–14.

Wang, S.-d., Kuo, T.-S., and Hsu, C.-F. (1986). Trace bounds on the solution of the algebraic matrix Riccati and Lyapunov equation. *IEEE Transactions on Automatic Control* **31,** 654–656.

Yang, S. and Kolaczyk, E. D. (2010). Target detection via network filtering. *IEEE Transactions on Information Theory* **56,** 2502–2515.

Yiu, G. K. and Toker, A. (2006). NFAT induces breast cancer cell invasion by promoting the induction of cyclooxygenase-2. *Journal of Biological Chemistry* **281,** 12210–12217.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research* page gks725.

Curriculum Vitae