An Interpretable Modelling Pipeline for Observational, Multi-modal Biomedical Data

Vineet Raghu[1], Xiaoyu Ge[1], Panos K. Chrysanthis[1], and Panayiotis V. Benos[1,2]

[1]Department of Computer Science, University of Pittsburgh
[2]Department of Computational and Systems Biology, University of Pittsburgh

Abstract

The abundance of measurable, personalized data has the potential to transform clinical decision making. Next generation sequencing technologies allow researchers to measure genes, proteins, and metabolites at the cellular level in individuals. Wearable and imaging data enable monitoring of clinical phenotypes and environmental factors. Since interventions can rarely be performed on human subjects, much of these data sources are observational. For modeling systems to have impact, causal knowledge must be inferred from these multi-scale observational data sources. In this paper, we propose an interpretable, causal, graphical modelling paradigm for multi-modal, observational data. Our paradigm builds upon existing domain knowledge to improve accuracy and interpretability, enabling knowledge discovery.

1 Introduction

Chronic disease are complex phenomena usually caused by the interaction of molecular and physiological factors with epidemiological and environmental factors. Technology has now advanced to a point where researchers can measure many of these factors. Next generation sequencing has enabled granular measurements of molecular data from individuals [24]. Advances in biomedical imaging technologies have provided a visual view of cellular phenotypes. Connected, wearable technology (e.g. smartphones, Fitbit, etc.) have enabled continuous monitoring of environmental and physiological signals [7]. Together, these data types give researchers the potential to 1) understand the fundamental causes of disease, 2) prioritize promising hypotheses, and 3) personalize medical treatments to individuals. The major challenge in this objective is the lack of effective multi-modal modelling techniques to understand complex interactions between these signals [8]. Machine learning (ML) is a popular tool for prediction in biomedical settings. However, ML is not suitable for these goals because ML models are not constrained to be interpretable by humans [21]. Further, ML models cannot infer causality from observational data [22]. Utilizing observational data alone to make causal predictions has been gaining popularity [16, 9, 1]; however, these models still have computational limitations and lack demonstrated successes [23, 15, 14].

Significance The scientific method is an iterative process to create knowledge consisting of: 1) Hypothesis formation from established knowledge, 2) Hypothesis testing through experimental intervention, and 3) Hypothesis evaluation from experimental data. Our vision is to automate hypothesis generation without targeted experiments. We aim to address the question: *Can simply observing a system lead to useful causal predictions?* To this end, we are developing a pipeline to integrate multi-modal observational data with domain knowledge to output causal predictions. With these, experiments can be prioritized and performed. This pipeline would be general enough to apply to any system, and understanding the causes of an outcome leads to actionable policy.

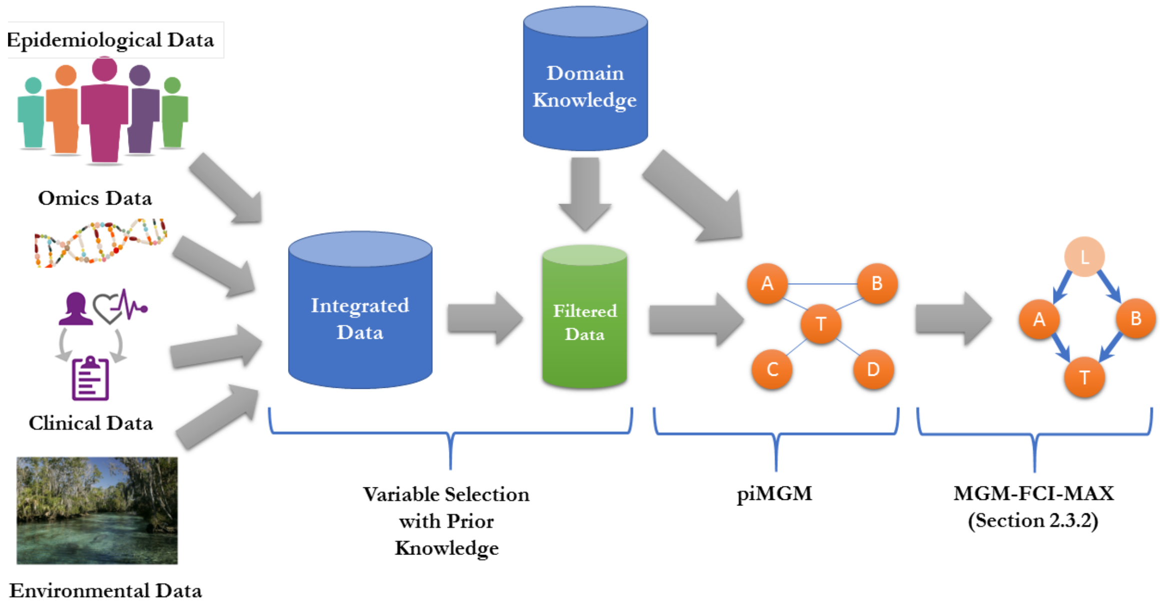


Figure 1: Our complete modeling pipeline. The pipeline takes multi-modal data as input (currently -omics and clinical data) and filters this data to fewer features that can be modeled. This data is merged with domain knowledge to learn an undirected graphical model (piMGM [12]). Finally, the undirected model is converted to a causal graph that identifies latent confounding (MGM-FCI-MAX Section 2.2.2)

Realizing Our Vision The first step in our vision is an efficient modeling pipeline to build upon domain knowledge with new observational data. The output is a network model of associations and experimental predictions. Our current pipeline (Figure 1) is built upon a class of models called Probabilistic Graphical Models (PGM's) [10]. PGM's are an unsupervised learning technique that model the joint distribution of variables as a graph where nodes are variables, and edges are conditional dependencies between variables. Directed edges correspond to causal relationships. We chose PGM's due to their inherent interpretability. Graphical representations of biological processes are common, which gives domain experts an intuitive understanding of their data. In addition, it is easy to query the graph to find causal relationships of interest.

Our complete pipeline consists of three components: 1) A feature selection algorithm to select relevant and diverse features while constructing meaningful aggregations of related features, 2) An algorithm to incorporate and evaluate known relationships to learn an undirected graphical model, and 3) A causal discovery algorithm to identify causal directions while accounting for latent confounding. Our pipeline is being implemented in an interactive web tool called Causal-MGM. This tool allows users to deploy our causal discovery methods on their observational data.

Contributions In this work, we present our causal modeling pipeline. Specifically,

- We discuss MGM-FCI-MAX [15]; a causal discovery algorithm to learn causal relationships in data with latent confounding (Section 2)
- We present an application of MGM-FCI-MAX to the early detection of lung cancer from low-dose CT scans and smoking factors [16] (Section 3)

2 Our Modeling Pipeline

Our modeling pipeline is based off of *mixed graphical models* [11, 20], which we discuss first and then discuss our method for learning causal models from multi-modal data with latent variables.

2.1 Mixed Graphical Models (MGM)

A PGM is a model that represents the joint distribution of variables as a graph which can be factored into local conditional distributions [10]. Originally, these models were only suitable for homogeneous data. Recently, MGM's have been proposed which model both categorical and continuous data in a single graph [11, 6, 3]. Based upon superior empirical performance, we chose to utilize the MGM learning algorithm from Lee and Hastie. Next, we briefly summarize the model they propose.

They parameterize the joint distribution of p continuous and q categorical variables (Equation 1). Here, β_{st} represents the linear interaction between continuous variables s and t. ρ_{sj} is a vector of parameters relating categorical variable j to continuous variable s, with one parameter for each category of j. Finally, ϕ_{rj} is a matrix representing the interactions between the categories of categorical variables r and j. This model generalizes two popular graphical models: the multivariate Gaussian for continuous data, and the pairwise Markov Random Field for categorical data.

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

Lee and Hastie optimize the parameters by minimizing the negative log-pseudolikelihood $(\widetilde{l}(\Theta))$[2]. In this model, the conditional distributions are 1) multivariate Gaussian for continuous variables with a mean given by a linear regression on the other variables and 2) Multinomial distribution for discrete variables with probabilities given by a multi-class logistic regression on the other variables. To prevent overfitting, they include sparsity penalties (Equation 2). In our work, we use separate sparsity parameters (λ) for each edge type (CC = Continuous-Continuous, CD = Continuous-Discrete, and DD = Discrete-Discrete), because this has shown better performance [20].

$$\underset{\Theta}{\text{minimize}}\, l_\lambda(\Theta) = \widetilde{l}(\Theta) + \lambda_{CC}\sum_{s=1}^{p}\sum_{t=1}^{s-1}|\beta_{st}| + \lambda_{CD}\sum_{s=1}^{p}\sum_{j=1}^{q}||\rho_{sj}||_2 + \lambda_{DD}\sum_{j=1}^{q}\sum_{r=1}^{j-1}||\phi_{rj}||_F \qquad (2)$$

2.2 Latent Variable Causal Discovery

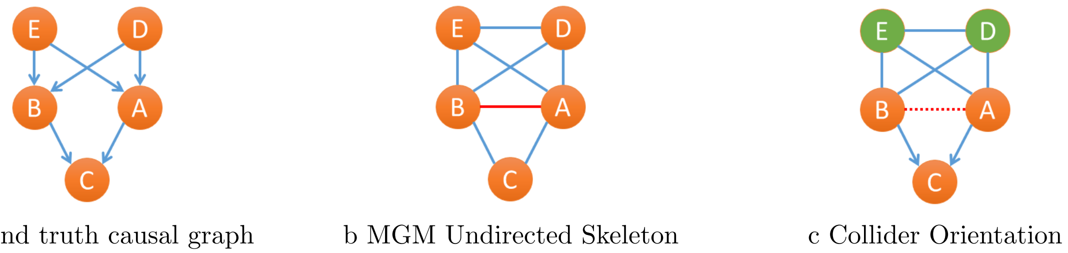
This section assumes knowledge of the causal discovery literature. Here we give a brief review, but for a thorough understanding of these concepts we refer the reader to [22, 14, 23].

2.2.1 Prior Work

The state of the art algorithm for learning causal relations from observational data with latent confounding had been the Fast-Causal Inference (FCI) algorithm. Several modifications of the algorithm have been proposed to improve FCI [18, 4, 5]; however, none are efficient and accurate enough for most applications. Integrated biomedical data poses another problem of mixed data. Causal discovery from mixed data requires a suitable independence test which have been explored [19], but not when data has latent confounding.

2.2.2 MGM-FCI-MAX

A crucial step of FCI is the ability to determine a separating set for each pair of variables because this determines orientations. In the original FCI algorithm, the separating set used is the smallest



a Ground truth causal graph b MGM Undirected Skeleton c Collider Orientation

Figure 2: Illustration of collider orientation process. The ground truth causal graph is given in (a). MGM would learn the graph in (b). In order to delete the edge, $A - B$ (orange edge), an appropriate separating set $\mathbf{S}$ must be found such that A is independent of B given $\mathbf{S}$. The MAX strategy would test all subsets of $\{C, D, E\}$ and find that $\{D, E\}$ (green nodes) is the best choice, thereby orienting the collider $B - C - A$

one, because this is first identified by the algorithm. Though this is the most efficient search strategy, in practice, FCI's causal orientations tend to be inaccurate.

To mitigate this problem, we proposed the MAX search (Figure 2) [15]. To test whether a triple $A - C - B$ should be oriented as a collider, the algorithm performs conditional independence tests using all subsets of the neighbors of A and C as conditioning sets. The MAX strategy chooses the subset S_{MAX} as the true separating set according to Equation 3, where $pval(A, B, S)$ is the p-value of the conditional independence test of A and B given $\mathbf{S}$.

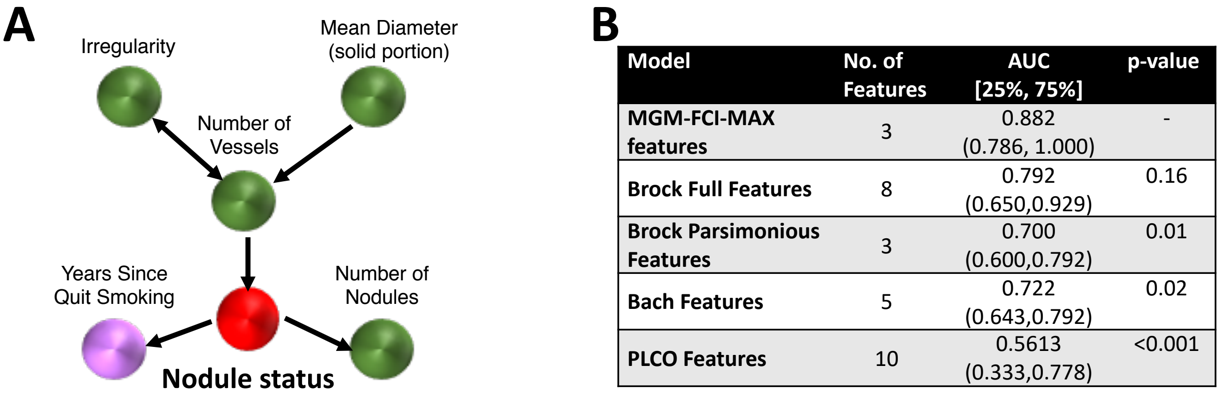
$$\mathbf{S}_{MAX} = \underset{\mathbf{S} \subset Adj(A) \cup Adj(C)}{\operatorname{argmax}} \quad pval(A, C, \mathbf{S}) \tag{3}$$

The intuition for this strategy stems from the fact that when A is dependent on B given $\mathbf{S}$, we expect the p-value distribution to be skewed towards 0. On the contrary, when A is independent of B given $\mathbf{S}$, we expect p-values to be uniformly distributed (null hypothesis). Thus, the maximum p-value is more likely to come from a true separating set where A and B are conditionally independent. After choosing the separating set, $\mathbf{S}_{MAX}$, we orient $A - C - B$ as a collider if $C \notin \mathbf{S}_{MAX}$, otherwise it is a collider. We refer to FCI with the MAX search technique as FCI-MAX. The limitation is that MAX requires the search to perform more conditional independence tests. For scalability, we first utilize the MGM algorithm to quickly learn the undirected graph. Then we use FCI-MAX to identify the causal relationships. Altogether we refer to this process as MGM-FCI-MAX.

3 Application: Early Detection of Lung Cancer

Low-dose computed tomography (LDCT) has become the standard method to screen for lung cancer [25]. Although, LDCT identifies nodules in 24% of the high-risk population, 96% of these nodules are benign. These result in extra costs through follow-up scans, invasive biopsies, etc. Many lung cancer prediction models have been developed; some of which use clinical characteristics to identify patients for screening [26]. More recently, models use CT scan features to predict probability of cancer [13]. In this section, we discuss the effectiveness of our causal model algorithm: MGM-FCI-MAX to identify a causal model of lung cancer and accurately predict lung cancer probability.

Data Source Our data is from a community-based research cohort called the Pittsburgh Lung Screening Study (PLuSS) [27]. This cohort consists of 3,642 current and former smokers who received a baseline and one year follow-up LDCT. A subset of PLuSS participants received biennial LDCT scans, yearly spirometry measurements and blood draws for 10 years. We used a randomly



Model	No. of Features	AUC [25%, 75%]	p-value
MGM-FCI-MAX features	3	0.882 (0.786, 1.000)	-
Brock Full Features	8	0.792 (0.650,0.929)	0.16
Brock Parsimonious Features	3	0.700 (0.600,0.792)	0.01
Bach Features	5	0.722 (0.643,0.792)	0.02
PLCO Features	10	0.5613 (0.333,0.778)	<0.001

Figure 3: a) Partial causal model learned by MGM-FCI-MAX. Three features were causally linked to cancer: number of nodules, blood vessels around the nodule, and years since the patient quit smoking. Directed arrows are causal relations and bidirected arrows are latent confounders. b) Cross-validation comparison against top lung cancer prediction models. MGM-FCI-MAX had the highest AUC in classifying indeterminate lung nodules. This figure has been adapted from [16]

selected cohort of 50 subjects with cancer and 50 subjects with benign nodules as a training set. The data consisted of 33 features from the CT scan, smoking history, and demographics.

Results Figure 3a presents part of our causal model. The model identifies three variables causally connected to cancer: blood vessels near the nodule is a *cause* of cancer (positively correlated), number of nodules, and years since the patient quit smoking are *results* of cancer (both negatively correlated). To validate, we used 10-fold cross validation. The entire model building process was run on each fold, and these three features were causes and effects of cancer in eight of ten folds.

Figure 3b presents the results of our cross-validation experiment. Our model performs the best in terms of area under the ROC curve. These results are statistically significant against all models except for the Bach model and the full Brock model which use eight and five features, respectively. For further validation, we applied our model to an independent cohort of 132 patients. We found that our model had higher prediction accuracy, but these results were not statistically significant against the two Brock models. The major use case of our model is preventing unnecessary follow-up tests for patients with a low probability of cancer. On the independent validation cohort, we found that a probability cutoff of 30% will accurately screen 28% of benign patients without missing a cancer case. This could drastically reduce health care costs without any risk to patients.

4 Discussion

The vision we proposed was an automated pipeline for hypothesis generation by only observing a system. Here, we presented a first step via our modeling pipeline. The pipeline takes integrated, observational data from several sources and outputs a model of the causal interactions in the data. The pipeline has three parts: 1) Feature selection and aggregation using domain knowledge, 2) Learning undirected graphical model structure with domain knowledge, 3) Learning causal associations from the undirected structure. We are testing our pipeline on applications to 1) breast cancer patient stratification and 2) understanding response to a prophylactic cancer vaccine, and have successfully applied the computational methods to diverse research problems [17, 16, 12]. Our completed pipeline is being implemented in a web server called CausalMGM (http://causalmgm.org/).

For future work, we aim to apply our completed pipeline to diverse data. Here we have focused upon the integration of transcriptomic and clinical variables; however, it is prudent to see if our methodology can handle environmental and epidemiological factors. Integrating these sources effectively will present interesting modeling problems. First, time-series data requires a causal framework to ensure causality moves forward in time. Next, context-specificity must be modeled, as causal relationships may only occur in certain situations (contexts). Finally, the interpretability of graphical models for knowledge discovery requires features (nodes) to be interpretable. A critical challenge is constructing human understandable features or aggregations of features from diverse sources (images, text, wearable sensor readings, etc.) in an automated or semi-automated way.

Acknowledgments: All software was implemented using Tetrad V from Carnegie Mellon University. We would like to thank our collaborators: David Wilson, Jiantao Pu, Dimitris Manatakis, Olivera Finn, Joseph Ramsey, Peter Spirtes, and Clark Glymour for their support. This work was funded in part by NIH grants: T32CA082084, R01LM012087, U01HL137159, U54HG008540. The content does not necessarily represent the official views of the National Institutes of Health.

References

[1] Irina Abecassis et al. Parp1 rs1805407 increases sensitivity to parp1 inhibitors in cancer cells suggesting an improved therapeutic strategy. *Nature Scientific Reports*, Forthcoming.

[2] Julian Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society, Series D*, pp 179–195, 1975.

[3] Jie Cheng, Tianxi Li, Elizaveta Levina, and Ji Zhu. High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 2016.

[4] Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[5] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, 40(1):294–321, 2012.

[6] Bernd Fellinghauer et al. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, 64:132–152, 2013.

[7] Mostafa Haghi, Kerstin Thurow, and Regina Stoll. Wearable devices in medical internet of things: scientific research and commercially available devices. *Healthcare informatics research*, 23(1):4–15, 2017.

[8] Konrad J Karczewski and Michael P Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299, 2018.

[9] Georgios D Kitsios, Adam Fitch, Dimitris V Manatakis, Sarah Rapport, Kelvin Li, Shulin Qin, Joseph Huwe, Yingze Zhang, John Evankovich, William Bain, et al. Respiratory microbiome profiling for etiologic diagnosis of pneumonia in mechanically ventilated patients. *Frontiers in microbiology*, 9:1413, 2018.

[10] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

[11] Jason D Lee and Trevor J Hastie. Learning the Structure of Mixed Graphical Models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.

[12] Dimitris V Manatakis, Vineet K Raghu, and Panayiotis V Benos. piMGM: Incorporating multi-source priors in mixed graphical models for learning disease networks. *Bioinformatics*, 34(17):i848–i856, 2018.

[13] Annette McWilliams et al. Probability of cancer in pulmonary nodules detected on first screening ct. *New England Journal of Medicine*, 369(10):910–919, 2013.

[14] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(0):96–146, 2009.

[15] Vineet K Raghu et al. Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *International Journal of Data Science and Analytics*, pp 1–13, 2018.

[16] Vineet K Raghu et al. Feasibility of lung cancer prediction from low-dose ct and smoking factors using causal models. *Thorax*, 2019.

[17] Vineet K Raghu, Xiaoyu Ge, Panos K Chrysanthis, and Panayiotis V Benos. Integrated theory-and data-driven feature selection in gene expression data analysis. In *Data Engineering (ICDE), 2017 IEEE 33rd International Conference on*, pp 1525–1532. IEEE, 2017.

[18] Joseph Ramsey, Jiji Zhang, and Peter L. Spirtes. Adjacency-Faithfulness and Conservative Causal Inference. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp 401–408, 2006.

[19] A.J. Sedgewick. *Graphical Models for De Novo and Pathway-Based Network Prediction Over Multi-Modal High-Throughput Biological Data.* dissertation, University of Pittsburgh, 2016.

[20] Andrew J. Sedgewick, Ivy Shi, Rory M. Donovan, and Panayiotis V. Benos. Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinformatics*, 17(S5):175, 2016.

[21] Galit Shmueli et al. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

[22] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search.* MIT press, 2000.

[23] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pp 3. SpringerOpen, 2016.

[24] Fuchou Tang et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377, 2009.

[25] National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[26] Kevin ten Haaf et al. Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. *PLoS medicine*, 14(4), 2017.

[27] David O Wilson et al. The pittsburgh lung screening study (pluss) outcomes within 3 years of a first computed tomography scan. *American journal of respiratory and critical care medicine*, 178(9):956–961, 2008.