Sample-Specific Models for Precision Medicine

Benjamin Lengerich

CMU-CS-20-139

December 2020

School of Computer Science
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Eric P. Xing, Chair
Zico Kolter
Ziv Bar-Joseph
Manolis Kellis (MIT)
Rich Caruana (MSR)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

This research was sponsored by National Institutes of Health award R01 5R01GM114311-04 and the Center for Machine Learning and Healthcare Fellowship. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: Personalized Machine Learning, Sample-Specific Models, Precision Medicine

To my Rose.

Abstract

Modern applications of artificial intelligence are often characterized by training large machine learning (ML) models on large datasets. These datasets are composed of overlapping groups of samples, either explicitly (e.g. the large dataset is created by combining multiple datasets) or implicitly (e.g. the samples belong to latent sub-populations). Population models prefer weakly-predictive global patterns over highly-predictive localized effects, a problem because localized effects are critical to understanding complex processes such as in applications to computational biology (in which samples come from latent cell types) and precision medicine (in which patients come from latent disease subtypes).

In this thesis, we propose that: The performance of intelligent computer systems can be improved by treating different samples as different tasks. This is especially helpful in domains such as computational biology and precision medicine, in which we care about understanding the highly specific context of each sample.

We propose to solve this problem by estimating a collection of many small models. For large collections, each model is responsible for only a small number of samples, enabling simultaneous interpretability and accuracy. As we show in this thesis, this framework can be scaled to estimate different model parameters for every sample.

This thesis begins by studying the challenges of characterizing real-world data with population-level models. Next, we develop the methodology of Personalized Regression. Finally, we apply sample-specific inference to computational biology and precision medicine by: (1) Identifying Discriminative Subtypes of Cancers from Histopathology Images and (2) Cell-Specific Transcriptomic Regulatory Network Inference.

Acknowledgments

Throughout my life, many people have supported me even more than I deserve. I'd like to thank a few of them here. First, I want to thank my wife, for "sticking with me". This PhD has defined much of our long-distance relationship, and your sacrifice and support has been integral to it.

I'm extremely grateful to my advisor Eric Xing. He has been a constant support and inspiration throughout my PhD. The SAILING lab group has been a tremendous environment for me to learn and grow, and I'm greatly appreciative of his urging to think bigger and to imagine how our work can guide the field and impact the world. I'll take these perspectives throughout my career.

Thank you to the other members of my thesis committee: Zico Kolter, Ziv Bar-Joseph, Rich Caruana, and Manolis Kellis. Your feedback has been extremely insightful, and I look forward to every time our paths cross.

Thanks to all my SAILING labmates who have helped me to learn, answered my many questions, and worked on research projects together. You all made our group meetings simultaneously fun, inspirational, and humbling. Thank you to Maruan Al-Shedivat, Willie Neiswanger, Haohan Wang, Xun Zheng, Helen Zhou, Micol Marchetti-Bowick, Avinava Dubey, Mrinmaya Sachan, Zhiting Hu, Jinliang Wei, Pengtao Xie, Aurick Qiao, Yaoliang Yu, Qirong Ho, Lisa Lee, Helen Zhou, Xiaodan Liang, and Amy Protos.

Thank you to my mentors who have guided individual research projects and my personal growth. Thank you Rich Caruana for your time, creativity, and advice – I feel incredibly lucky to have had that extra intern spot open up, and I've changed both personally and professionally as a result of our work together. Thank you Bryon Aragam for your ongoing research guidance and personal advice. Thank you to Chris Potts and Andrew Maas for guiding me to write my first paper – that was a very enjoyable summer.

Thank you to my friends who have navigated the PhD jungles with me. Thanks to Michael Kleyman for the years of friendship which defined this time. Thanks to Amir Alavi for putting up with my late nights, being a role model to find balance, and being my virtual groomsman. Thanks to many other friends who've shared both good and bad times with me: Francisco Sanguinetti, Anthony Platanios, Otilia Stretcu, Jennifer Williams, Bentley Wingert, Natalie Sauerwald, Cong Ma, Ian Welland, and many others. Thank you to the younger collaborators whom I have had the honor of working with: Sami Labbaki, Zhihao Ouyang, Stephen Tsou.

Thank you to the administrators Deb Cavlovich, Diane Stidle, and Thom Gulish who make everyone's lives much easier.

Finally, I'd like to thank my family. You have loved me, fed me, and pushed me to do "good work". I appreciate all of our conversations where you challenged my ideas, and all of those in which I'm sure I bored you with my research but you graciously listened anyway. I've been given much more than I deserve to be raised and surrounded by such a family. Thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Modern machine learning (ML) applications often seek to represent complex phenomena by estimating large models from large datasets. However, these datasets are often composed of overlapping groups of samples, either explicitly (e.g. the large datasets is created by combining multiple datsets) or implicitly (e.g. the samples belong to latent sub-populations). As a result, these datasets contain some globally predictive effects, but also contain many highly predictive localized effects (applicable to a smaller number of samples). When faced with a localized effect, traditional population-level ML models can either ignore this effect or encode the localized effect as an interaction of many input variables. This is an important problem because localized effects are critical to understanding *heterogeneous* datasets in which the samples represent different underlying processes. Heterogeneity is especially important in applications to computational biology (in which samples belong to latent cell types) and precision medicine (in which patients belong to latent disease sub-types). In these settings, we need models which can simultaneously capture sample variability while retaining the interpretability of small models.

We propose to explore a solution to this challenge by expanding the toolbox of heterogeneous models. Instead of representing localized effects as high-order interaction effects, we propose to estimate collections of models from which a different model can be chosen to be applied to each localized region. In this way, we can explicitly reason about the structure of the collection and the effects contained in each model.

To develop rigorous methodology which is useful for any level of localized effects, we go to the extreme and study processes which vary continuously between all samples. For these extreme problems, we propose to estimate parameters which take on different values for every sample in the dataset. These models are able to accurately fit the data while also providing human intelligibility.

Towards Precision Medicine

One of the fundamental goals of precision medicine is to understand the patterns of differentiation between patients such that the appropriate care can be provided for each individual. However, traditional cohort-level models estimate the same effect sizes for all patients. Since the patients actually have different histories, environments, and disease sub-types, no cohort-level model can appropriately model the patient journeys.

This perspective of individualized risk profiles aligns with clinicians' practical thinking of individualized treatments. As Yang et al. found in clinical evaluation of a predictive model:

> "Some voiced strong concerns that using [a ML model] was the same as applying 'populational statistics' to individual patient decision making. They felt this was unethical. Others proposed that 'instead of having one model that we apply to the entire population, we would have a group of models. Those models predict for that group of patients." [198]

In this thesis, we seek methods to estimate such groups of models to satisfy the clinicians' desires. If we could instead estimate model parameters which vary smoothly between samples, we could make principled sample-specific inferences (e.g. "is protein i influential for deciding the treatment of patient j?"). Unfortunately, prior ML work on collections of models requires either (1) a small number of sub-groups relative to the number of samples (e.g. mixture models) [162], (2) known patterns of variation [94, 144, 176], or (3) significant domain knowledge to constrain the solutions [196, 197]. These requirements are inappropriate for applications in computational biology and precision medicine, in which each individual has a slightly different version of the disease and the domain is too large to encode patterns of differentiation. Thus, we are motivated to design methods which can automatically extract patient similarity and provide patient-specific risk scores. We will especially focus on precision oncology, for which sample-specific models can provide insight into the molecular subtype of the tumor and suggest therapeutic targets.

Towards Intelligible Artificial Intelligence

In many domains, the application of AI is limited due to a strict requirement for intelligible decisions which can be audited by humans. Current standards of ML focus on optimizing predictive accuracy by estimating large models, but these large population-level models are often difficult to interpret. To overcome this problem, some works have proposed post-hoc procedures to approximate the large model with locally-interpretable models [160]. In this thesis, we propose to approach the same endpoint more directly: we seek to estimate a collection of simple models from the data without ever optimizing a black-box model. In this way, we provide direct interpretability of models and potentially unlock intelligible AI to be used in critical scenarios.

Biological Context and Datatypes Used

The main application area considered in this thesis is computational genomics. To supplement the introductions of each individual application, we describe here some of the datatypes and application goals used repeatedly throughout this thesis.

Gene Expression

The "central dogma" of molecular biology [39] states that genetic information flows from deoxyribonucleic acid (DNA) into expression physical traits through the formation of proteins which catalyze molecular reactions. Proteins are polypeptide chains [40] which catalyze molecular reactions and effect phenotypic changes.

To construct these chains, DNA is transcribed into messenger ribonucleic acid (mRNA), which is in turn translated by ribosomes into polypeptide chains of protein. The bulky nature of proteins has made it more convenient for biologists to design assays to measure expression levels of mRNA than expression levels of proteins. While all cells within an individual contain the same DNA sequence, cells are differentiated based on gene expression – which genes are turned on in which cells and what concentration levels. Each mRNA transcript can be connected (sometimes imperfectly) to its precursor gene and many to their resulting proteins (some RNA transcripts do not code for proteins). Thus, measuring mRNA concentration levels informs about the gene expression, and turns the static gene sequence into a snapshot of cell state.

While mRNA is a precursor to protein synthesis, measuring the expression levels of mRNAs provides only a messy picture of the concentration of effective proteins in cells for many reasons. For example, post-translational modifications of proteins can change the effectiveness of a protein's ability to catalyze a reaction. In other cases, there can exist an inverse relationship between the level of expression of an mRNA transcript and the corresponding protein when the increase in mRNA expression indicates a bottleneck in the production of the downstream protein (e.g. due to scarcity of a peptide needed to construct the poly-peptide chain). For these and other reasons, mRNA expression levels provide only noisy views of protein expression levels and must be de-noised through computational means.

The most common assay to measure mRNA expression levels is RNA-Seq. Traditional RNA-seq assays blend a population of cells together, extract the expressed RNA, and provide a count of the number of reads recovered for each RNA transcript. These integer counts are normalized and mapped to the precursor gene to give a continuous-valued measure of the expression of each gene in the cell population. For more information on RNA-seq processes and data, please refer to [99].

More recently, the biological community has developed tools to measure single-cell RNA-seq (scRNA-seq), which does not the blend the cell population together and instead provides the expression levels of genes within individual cells [166]. This view provides more granularity to observe heterogeneous processes which result in different levels of gene expression in different cells, but comes at the cost of extra technical noise. Thus, ML techniques are critical to making sense of scRNA-seq datasets.

Epigenetics

For a gene to be transcribed to RNA, it must be accessible by RNA polymerase (an enzyme tasked with synthesizing RNA). Thus, the physical configuration of chromatin (DNA) and epigenetic modifications dictates which genes can be expressed. As a result, we are interested in tying the gene regulation encoded by DNA accessibility to the gene expression (measured by RNA-seq).

A common way to measure DNA accessibility is through the "transposase-accessible chromatin using sequencing assay" (ATAC-seq) [16]. ATAC-seq begins by lysing cells to extract chromatin. The chromatin is then fragmented and tagged to identify which regions of the chromatin are accessible to the transposase enzyme. Chromatin regions can be mapped to functional elements such as genes, providing a measure of the accessibility of each gene to the machinery of gene expression.

Similarly to the recent advances in scRNA-seq, single-cell ATAC-seq (scATAC-seq) assays have also been developed to profile gene accessibility within individual cells [17]. Again, these assays provide sharper granularity of heterogeneous biological processes but come with the caveat of increased technical noise. As a result, there is a strong opportunity for ML to improve the analysis of scATAC-seq datasets.

Multi-Omic

Finally, it should be considered that the RNA-seq and ATAC-seq assays are destructive; that is, each cell can only be profiled a single time. As a result, we have not had simultaneous views of gene expression and gene accessibility. Very recent advances in multi-omic sequencing assays allow us to profile both of these simultaneously in the same cells. These assays provide, for the first time, a view of concurrent gene expression and epigenetic markers in the same cell. As is the trend with many of these high-resolution assays, by choosing this assays type we trade off improved specificity against more technical noise. We explore some of the potential of using this new datatype in Chapter 6.

1.2 Outline

This thesis is divided into several chapters:

- In Chapter 2, we study the challenges of using population-level models to understand complex phenomena. Because datasets often contain heterogeneous samples which have different underlying effects, large population-level models often permit interaction effects of many variables. As a result, estimation of these models can have excess variance.
- In Chapter 3, we show an equivalence between these high-order interaction effects and heterogeneous models and unify several types of heterogeneous models into a single framework for future analysis. We push heterogeneous models to the extreme of estimating model parameters which can be different for every observed sample.
- In Chapter 4, we apply these methods to precision oncology. First, we estimate discriminative subtypes of lung cancers from histopathology images. These discriminative subtypes permit optimal predictions from transcriptomic assays, with transcriptomic model parameters generated according to the inferred discriminative subtype. This procedure allows us to label phenotypic patterns according to transcriptomic variables.
- In Chapter 5, we propose “Personalized Regression” which is appropriate when we do not believe a one-to-one relationship between covariates and regression parameters exists. We apply this method to identify sample-specific patterns of cancer transcriptomics.
- Finally, in Chapter 6, we propose Contextualized NOTEARS, a method to estimate sample-specific Bayesian Networks. We apply this method to estimate patient-specific transcriptomic regulatory networks for cancer patients and cell-specific transcriptomic regulatory networks which can be rewired according to each cell’s epigenetic markers.

Chapter 2

Hidden Dangers in Population Models

Modern ML systems often seek to train large models on large datasets. To capture the complexity inherent in these large datasets, practitioners often prefer large model classes with large representational capactiy and few contraints. These models, such as fully-connected neural networks, often have too many parameters for each to be intelligible, leading to a black-box nature. There are many methods to peer inside trained black-box models, but can we actually trust what these population-scale models have learned? Below we discuss two common problems which lead us to believe that more targeted models can be preferred over population-scale black-box models.

Portions of this chapter have been previously published or are in review as [105, 106, 107].

2.1 Heterogeneous Effects

When analyzing datasets by estimating a single population model, it's possible to get very deceiving results if the population model is mis-specified – that is, there does not exist a population model which accurately summarizes the processes underlying the observed data. This often happens in medical datasets, for which a medication's treatment effect avereaged over the entire population can be extremely limited (or even harmful), but the same drug may be helpful to particular types of patients. Such treatment effects which vary between patients are often referred to as *heterogeneous treatment effects* [101, 189].

Heterogeneous Treatment Effects of Glucocorticoids on Covid-19

An interesting recent example of heterogeneous treatment effects is in the treatment of Covid-19 pneumonia. RCTs have shown that Glucocorticoids (GCs) can improve outcomes of patients with severe cases of Covid-19 [79]. Here, we examine a large dataset of Covid-19 patients to show that GCs may have a hetereogeneous treatment effect and thus could be targeted to patients with high Neutrophil/Lymphocyte Ratio (NLR) at time of admission.

Dataset

We are interested in the outcome of in-hospital mortality for Covid-19 patients. Our dataset consists of hospitalized patients who have lab-confirmed cases of Covid-19. To filter out patients who were hospitalized for reasons other than Covid-19, we exclude patients who have indicators of (1) pregnancy: outpatient prenatal vitamins, in-patient oxytocics, folic acid preparations; or (2) scheduled surgery: urinary tract radiopaque diagnostics, laxatives, general anesthetics, antiemetic/antivertigo agents, or antiparasitics. We also require that the patients have recorded temperature, age, BMI, and Admission Day. Finally, we remove patients who died within six hours of admission.

Patient Features To correct for patient risk confounding, we observe pre-admission features including demographics, comorbidities, outpatient medications, initial in-patient vitals, and initial in-patient lab tests. We exclude any measurement taken within 24 hours of the patient mortality. Neutrophil-Lymphocyte Ratio is calculated by dividing Neutrophil % by Lymphocyte %. We exclude all patients who are missing either of these lab values. Because GCs can affect patient NLR, we use only initial lab tests which are taken before in-patient medications are administered.

The patient population has changed over time. The majority of patients were admitted during the first recognized wave of the pandemic, which is approximately a 3-week period in this dataset. However, the range of NLR observed in the days and months following the peak of the pandemic remains wide.

Treatment We define "GC treatment" as treatment with a GC within 24 hours of hospital admission. Our control condition is not receiving any GC in the first 24 hours. Our analysis includes 3108 patients hospitalized for Covid-19 with 193 receiving GCs. This cohort includes patients hospitalized from March to August with an average mortality rate of 18.1%. Mortality rate peaked over 25% and decreased to less than 5% in August.

To ensure a proper linking of lab values, treatments, and outcomes, we consider only GC treatments that are given within 24 hours of the initial lab test and at least 24 hours before mortality. In this set of patients, the number of doses by GC medication are: 627 Methylprednisolone, 291 Prednisone, 186 Hydrocortisone, 165 Dexamethasone. The proportion of Covid-19 patients being prescribed GCs has increased over time from a minimum of less than 5% to a more recent peak near 30%.

Glucocorticoid prescriptions are correlated with: Admission Day (R=0.16), Chronic Obstructive Pulmonary Disease (R=0.13), Outpatient Beta-Adrenergic Agents (R=0.11), Valve Replacement (R=0.10), and Increased Charlson Score (R=0.10). Glucocorticoid prescriptions are anti-correlated with: In-Patient Hydroxychloroquine (R=-0.06), Hematocrit (R=-0.05), and In-Patient Analgesic/Antipyretics (R=-0.05).

Methods

Treatment protocols changed over time with more GC prescriptions at later dates [80]. To correct for this and other confounding, we use a two-stage machine learning procedure [23, 70] to estimate GC effect after correcting for patient mortality risk at admission.

We use generalized additive models (GAMs) to model patient mortality risk at admission. GAMs are a version of logistic regression that are able to model non-linear effects [4]. While logistic regression summarizes the influence of each feature with a single coefficient, GAMs estimate the influence of a feature for every value the feature can take on as a graph. This means GAMs can accommodate non-linear effects, which improves both model accuracy and interpretation. Modeling non-linear effects is particularly important when features have multiple regions of high or low risk (e.g., both hyperthermia and hypothermia are associated with high risk).

In particular, we use tree-based GAMs [23] implemented in the Python `Interpret` package [142]. These GAMs are invariant to all monotonic feature transforms, so log-transforms of lab values are not necessary.

The risk model achieves an ROC of 0.912 ± 0.001 and an F1-score of 0.598 ± 0.002 on held-out patients. This significantly outperforms a logistic regression model which achieves an ROC of 0.859 ± 0.001 and F1-score of 0.455 ± 0.002 on the same data.

If NLR is included in this model, it jumps to becoming the most important feature. This agrees with clinical understanding of NLR as a marker of inflammation linked to severe cases of Covid-19 [102]. Figure 2.1 demonstrates the effect of NLR on mortality risk estimated in this model: while elevated NLR is a strong predictor of mortality, it is not the only risk factor. Indeed, there are many patients with low NLR levels who nevertheless have large probabilities of mortality, and many of these patients died.

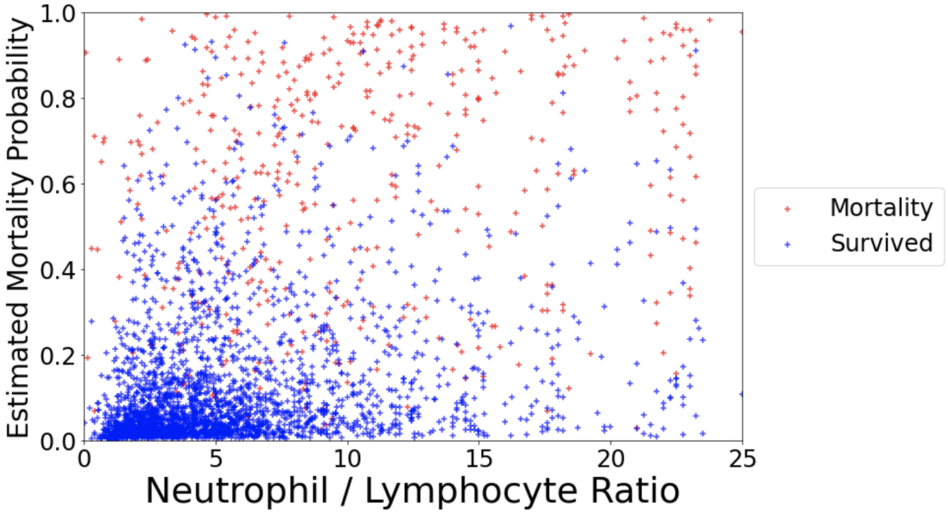


Figure 2.1: Estimated probability of mortality for each patient, as predicted by the risk model operating on patient features at admission. While there is a relation between elevated NLR and increased risk of mortality, there are a sizable number of patients who have a large probability of mortality without elevated NLR.

Results

Effect of NLR As a preliminary investigation, we explore marginalization. We see that elevated NLR is associated with elevated mortality risk (Figure 2.2). These marginalization plots indicate that the mortality rate of patients treated with GCs does not rise as rapidly with elevated

NLR as does the mortality rate of patients not treated with GCs. GCs are associated with improved outcomes for patients with NLR>6: for patients with NLR>6, 27.2% (31 of 114) treated with GCs died compared to 32.9% (393 of 1195) not treated with GCs. For patients with NLR below 6, 8.9% (7 of 79) treated with GCs die compared to 6.9% (120 of 1741) not treated with GCs.

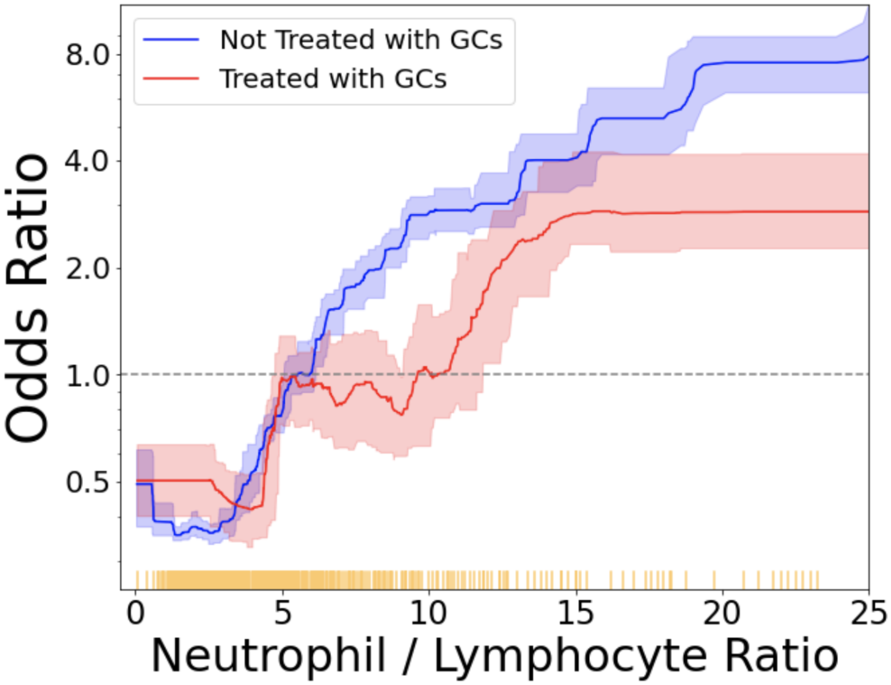


Figure 2.2: Marginal odds ratio of mortality by NLR in patients treated and not treated with GCs. Shaded regions are 95% CIs. For patients not treated with GCs (blue), the probability of mortality steadily increases with increased NLR; for patients treated with GCs (red), the probability of mortality does not rise rapidly through the region of NLR 6-10. Each yellow tick mark along the horizontal axis indicates 10 patients.

Lack of Homogeneous Treatment Effect of GCs We first seeek to estimate the effect of GCs as a homogeneous benefit to all Covid-19 patients. After correcting for patient risk factors, our analysis finds that GCs do not have a significant homogeneous effect, i.e., GCs do not benefit all patients equally. After correcting for confounding variables, we estimate an Odds Ratio (OR) of 0.96 with 95% confidence interval (CI) 0.86-1.09 across all patients (Table 2.1). Thus, we examine possible heterogeneous effects of GCs.

NLR-Mediated GC Effect We examine the heterogeneous effect of GCs with benefit modulated by NLR level. For each NLR value, the ratio $I(x) = T_1(x)/T_0(x)$ where T_1 is the OR for patients treated with GCs and T_0 is the OR for patients not treated with GCs, gives the benefit of treatment with GCs. Once again, despite the reduced sample size in the NLR bins, Figures 2.3

and 2.4 show that this benefit is best for patients with NLR between 6 and 10 and statistically most significant near NLR=8.5.

One major difference between this result and the marginalization result shown in Figure S5 is the estimated benefit to patients with extremely high NLR (above 10): while the marginalization showed lower mortality rate for patients in this group prescribed GCs, this analysis does not show any evidence of a beneficial effect of GCs to patients in this group. This suggests that the lower mortality rate for patients with NLR > 10 is explainable by other risk factors taken into account by the background risk model, e.g., recent patients who are more likely to be prescribed GCs may have fewer high-risk comorbidities.

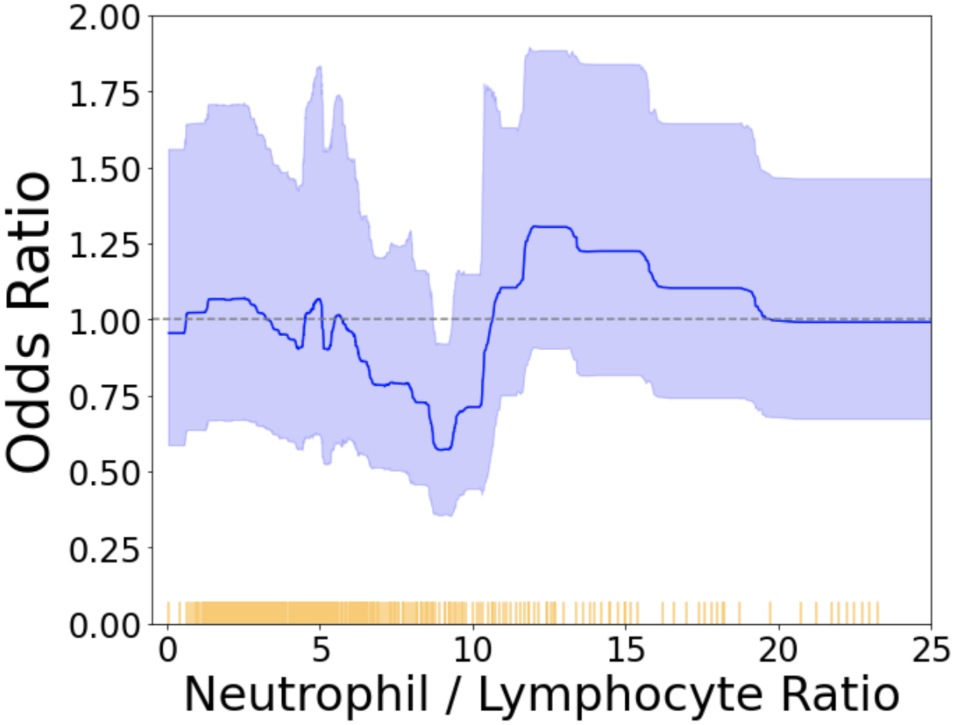


Figure 2.3: Estimated benefit I(x) of GC treatment for by NLR value, after correcting for patient risk. Shaded regions indicate 95% CIs. Despite the small sample size in each NLR bin, there is still statistical significance near NLR=8.5.

Estimating heterogeneous treatment effects for all NLR values greatly reduces statistical power and inflates the width of the CIs. To test for statistical significance of the GC benefit, we group NLR values into 3 ranges: NLR < 6, NLR 6-10, NLR > 10. Using these three ranges, we estimate ORs (and 95% CIs) for NLR 0-6: 0.97(0.61-1.51), NLR 6-10: 0.64(0.44-0.93) and NLR 10-25: 1.14(0.85-1.53) (corresponding sample sizees are given in Table 2.1). These values indicate that the benefit of GCs is statistically significant at p<0.05 for NLR 6-10, but not for either NLR < 6 or NLR > 10.

These results agree with current clinical understanding that GCs benefit patients with severe cases of Covid-19, and that elevated NLR is associated with mortality. However, there are also high-risk patients with NLR<6 (Figure 2.1), suggesting varying presentations of severe Covid-

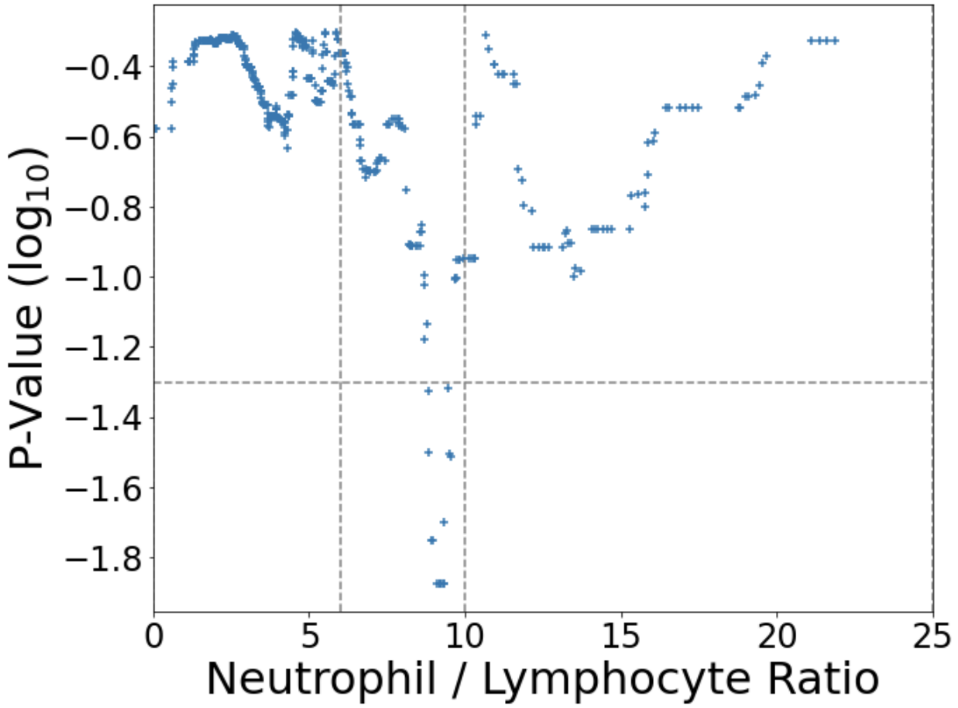


Figure 2.4: P-values of the estimated benefit of GC treatment by NLR value. Vertical gray lines demarcate the NLR bins 0-6, 6-10, and 10-25. The horizontal gray line indicates p=0.05.

19, some of which may not respond to GCs. Our study comes with all caveats of observational analyses, and we suggest randomized control trials to study heterogeneous treatment protocols.

Lessons

Our results suggest that GCs may have limited benefit to patients who are at high risk without having an elevated NLR. Using a single population model to estimate a homogeneous treatment effect would have estimated little to no benefit of GCs. Randomized control trials informed us that GCs are indeed likely to be an effective treatment, so wee searched for heterogeneous treatment effects of GCs and found a strong effect mediated by patient NLR. Automating this

NLR Values	N GC, N Control	Odds Ratio (OR)	95% CI	P-Value
0–6	79, 1741	0.97	0.61 – 1.51	0.896
6–10	45, 561	0.64	0.44 – 0.93	0.004
10–25	54, 518	1.14	0.85 – 1.53	0.419
0–25	178, 2800	0.96	0.86 – 1.09	0.495

Table 2.1: Odds Ratios of mortality for patients treated with GCs compared to patients not treated with GCs, calculated by patient NLR. ORs less than 1 indicate reduced mortality for patients treated with GCs, i.e., a beneficial effect.

process could help future investigations into designing more individualized treatment regimens for these complex diseases.

2.2 Interaction Effects

In the previous example, we saw that heterogeneous treatment effects can be critical to understanding patient outcomes and proper therapeutic choices. We discovered this heterogeneous effect by examining the interaction between NLR, GCs, and mortality. This choice of three features was chosen by domain knowledge and previous literature suggesting that NLR, GCs, and Covid-19 mortality relate to similar inflammatory processes. Instead of using domain knowledge to pick specific interaction effects to model, would it be possible to allow a large population model to simultaneously explore all interactions? To begin to answer this question, we need to define pure interaction effects.

Pure Interaction Effects

In the rest of this thesis, we will use the concept of *pure interaction effects* from [106]. According to this definition, a pure interaction effect is variance explained by a group of variables u that *cannot* be explained by any subset of u. This definition is equivalent to the functional ANOVA decomposition of the overall function F: Given a density $w(X)$ and $\mathcal{F}^u \subset \mathcal{L}^2(\mathbb{R}^u)$ the family of allowable functions for variable set u, the weighted functional ANOVA [42, 77, 78] is:

$$\{f_u(X_u) | u \subseteq [d]\} = \underset{\{g_u \in \mathcal{F}^u\}_{u \in [d]}}{\operatorname{argmin}} \int \Big(\sum_{u \subseteq [d]} g_u(X_u) - F(X) \Big)^2 w(X) dX,$$

where $[d]$ indicates the power set of d features, such that

$$\forall\ v \subseteq\ u, \quad \int f_u(X_u) g_v(X_v) w(X) dX = 0 \quad \forall\ g_v, \tag{2.1a}$$

i.e., each member f_u is orthogonal to the members which operate on a subset of the variables in u. Once this decomposition has been found, we have a set of functions f_u which all have zero-mean and can be analyzed independently. We say that an interaction effect f_u is of *order* k if $|u| = k$.

This definition of interaction effects demands a data distribution. As Lengerich et al. describe, the correct distribution to use is the data-generating distribution $p(x)$. In studies on real data, estimating $p(x)$ is one of the central challenges of machine learning; here, we use only simulation data for which we know $p(x)$ and can precisely study the effects of Dropout.

While multiplicative terms like $X_1 X_2$ are often used to encode "interaction effects", they are only *pure* interaction effects if X_1 and X_2 are completely uncorrelated and have mean zero. When the two variables are correlated, some portion of the variance in the outcome $X_1 X_2$ can be explained by main effects of each individual variable (e.g. if X_1 and X_2 are perfectly correlated, $X_1 X_2 = X_1^2$). Note, however, that in the general case correlation between two input variables does not imply an interaction effect on the outcome, and an interaction effect of two input variables on the outcome does not imply correlation between the variables.

Statistical (Un)Reliability of Interaction Effects

One reason why models which ignore high-order interaction effects can perform so well is the tremendous difficulty that higher-order interaction effects present to learning algorithms. When trying to learn high-order interaction effects, we are stuck between a rock and a hard place: the number of possible interaction effects grows exponentially (the number of k-order interaction effects possible from N input features is $\binom{N}{k}$), while the the variance of an interaction effect grows with the interaction order [110]. This quandary is intensified when the effect strength decreases with interaction order, which is reasonable for real data [57]. It is like searching for a needle in a haystack, but as we increase k, the haystack gets larger and the needle gets smaller. For large k, we are increasingly likely to select spurious effects rather than the true effect – at some point it is better to stop searching the haystack. Viewed this way, it is less surprising that in the absence of prior knowledge of which interaction effects, simple models are able to outperform large models.

Interaction Effects in Neural Networks

The function estimated by a neural network can be decomposed as:

$$\hat{F}(X) = \sum_{u \in [d]} \hat{f}_u(X_u) \tag{2.2}$$

by the functional ANOVA. We will use this decomposition to measure the interaction effects implicit in the estimated $\hat{F}$. To approximate this decomposition, we repeatedly apply model distillation [15, 76] using the XGBoost software package [28]. First, we train boosted stumps (XGBoost with max depth of 1) to approximate the output of the neural network using only main effects of individual variables. We successively increase the maximum depth of trees (corresponding to an increase in the maximum order of interaction effect permitted). By training on the residuals of the previous model, we ensure that the estimated effects are orthogonal. In the remainder of this paper, we will refer to $\text{Var}_X(\hat{f}_u(X))$ as the *effect size* of an estimated effect $\hat{f}_u$.

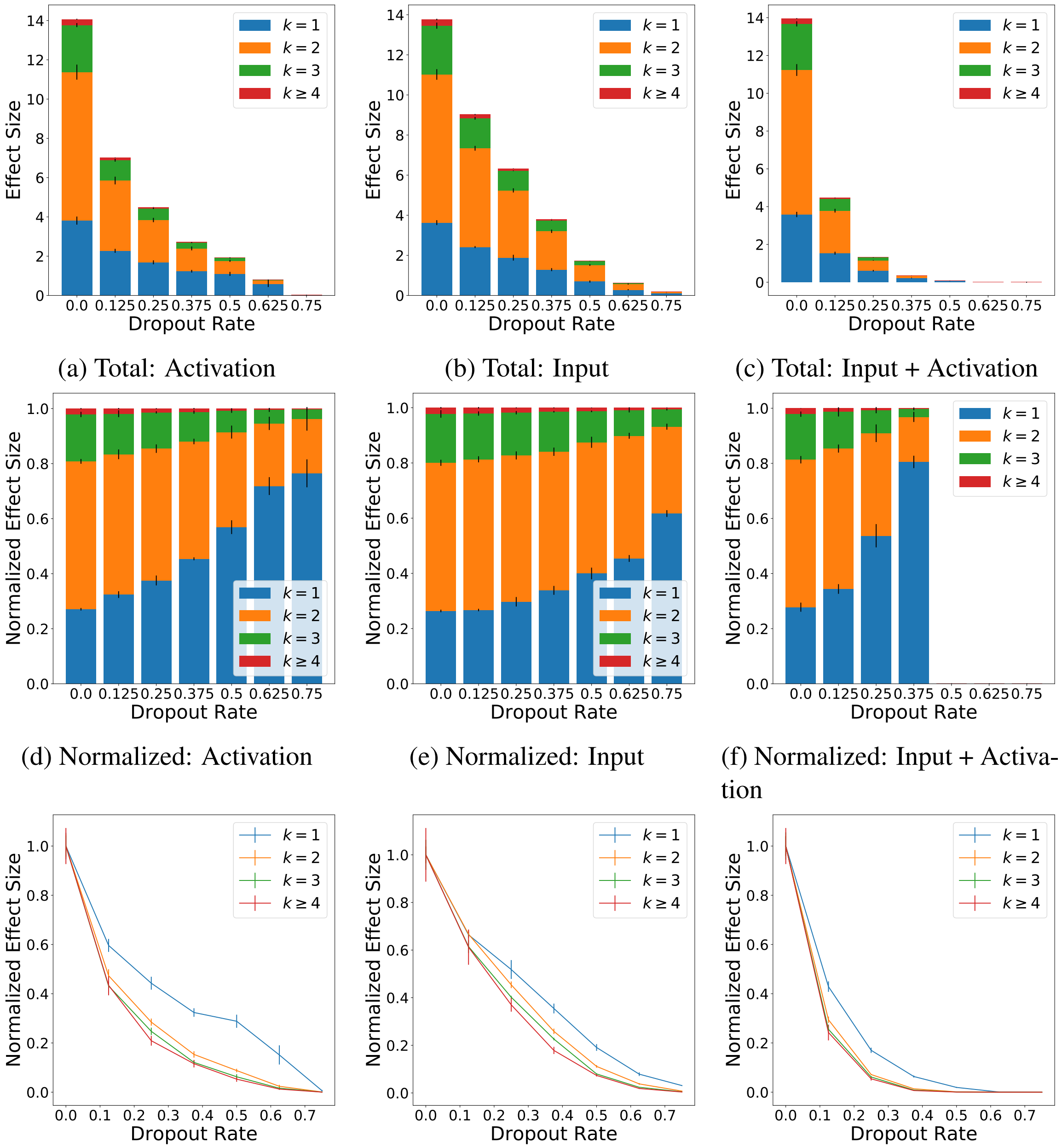
Dropout Regularizes Against Interaction Effects

Dropout operates by probabilistically setting values to zero. For clarity, we call this action "Input Dropout" if the perturbed values are input variables, and "Activation Dropout" if the perturbed values are activations of hidden nodes. Input Dropout, which targets the input variables, is equivalent to augmenting the training dataset with samples drawn from a perturbed distribution:

Theorem 1. *Let* $\mathbb{E}[Y|X] = \sum_{u \in [d]} f_u(X_u)$ *with* $\mathbb{E}[Y] = 0$. *Then sampling with Input Dropout at rate* p *has*

$$\mathbb{E}[Y|X \odot M] = \sum_{u \in [d]} (1-p)^{|u|} f_u(X) + \zeta \tag{2.3}$$

where M is the Dropout mask, $\odot$ is element-wise multiplication, and ζ is normally distributed with mean zero. Without changing the outcomes Y, Input Dropout drives the conditional expectation of $Y|X \odot M$ toward the marginal expectation of Y. Furthermore, it acts by preferentially



(a) Total: Activation

(b) Total: Input

(c) Total: Input + Activation

(d) Normalized: Activation

(e) Normalized: Input

(f) Normalized: Input + Activation

(g) Shrinkage: Activation

(h) Shrinkage: Input

(i) Shrinkage: Input + Activ.

Figure 2.5: Dropout regularizes neural networks by down-weighting higher-order interaction effects. In this experiment, we train fully-connected neural networks on a dataset of pure noise. Displayed values are the (mean ± std. over 10 initializations) of the trained model's variance explained by each order of interaction effect. The top row of graphs (a–c) shows the absolute variance of the models for different values of Dropout — as Dropout grows, overfitting is reduced and the variance of the predictions converges towards zero. The middle row (d–f) shows the relative effect sizes of interactions of degree 1, 2, 3, and 4 or greater. The bottom row (g–i) shows the effects normalized by their strength in the unregularized model.

targeting high-order interactions: the scaling factor grows exponentially with $|u|$. Because the distribution of training data is actually different for different levels of Input Dropout, we should expect that models will converge to different optima based on the level of Input Dropout (e.g., Input Dropout introduces bias). Finally, we note that Input Dropout acts on the data distribution, not the model, so it has the same effect on the learning process regardless of the downstream architecture.

To see this, we examine a set of neural networks trained to convergence with varying levels of Dropout. In this experiment, we use a simulation setting in which there is no signal (so any estimated effects are spurious). This gives us a testbench to easily see the regularization strength of different levels of Dropout. Specially, we generate 1500 samples of 25 input features where $X_i \sim Unif(-1, 1)$ and $Y \sim N(0, 1)$. We optimize neural networks with 3 hidden layers and ReLU nonlinearities. In Fig. 2.5, we see the results for neural networks with 32 units in each hidden layer. For this small network, both Activation and Input Dropout have strong regularizing effects on a neural net. Not only do they reduce the overall estimated effect size, both Activation and Input Dropout preferentially target higher-order interactions (e.g., the proportion of variance explained by low-order interactions monotonically increases as the Dropout Rate is increased for Figs. 2.5d,2.5e, and 2.5f).

Thus, Dropout does not simply introduce unbiased noise into learning — training with higher levels of Dropout produces models that are likely to learn weaker interaction effects. This suggests that even large black-box models such as deep neural networks suffer the pain of estimating interaction effects and benefit from constraints on interaction effects.

Lessons

We have seen that pure interaction effects are difficult for unconstrained models to learn from data. A major difficulty in learning these interaction effects is the number of interactions, which grows exponentially with the number of variables in the interaction. As a result, deep neural networks are improved by Dropout, which regularizes against interaction effects. Going forward, we should not expect larger population models to accurately capture high-order interaction effects solely from data.

Chapter 3

Samples as Tasks

In the previous chapter, we saw that estimating population-level models can hide sub-population heterogeneity. Moreover, this problem is not avoided by simply expanding the population model's representational capacity, because the interaction effects inherent in heterogeneous samples are difficult to find unless prior specification is provided. These challenges lead us to ask if we can instead use principled methods to share statistical power between samples while still permitting the estimated effects to vary between samples? This motivates us to examine the connection between sample heterogeneity and multitask learning.

3.1 Tasks as Interaction Effects

We often observe multiple related phenomena, or "tasks", which have distinct distributions $P_t(Y|X)$, where $t \in T$ indexes the task[1].Multitask learning [14, 22] seeks to improve the estimation of each $P_t(Y|X)$ by sharing power between distinct tasks t. In some cases, the task label is hidden and we are concerned with the single distribution shared by all tasks:

$$Y|X \sim \int_t P_t(Y|X)\lambda(t)dt, \tag{3.1}$$

where $\lambda(t)$ is the scalar weight of task t. In other cases, the task label is provided as an input and we are concerned with the task-specific distributions:

$$Y|X, t \sim P_t(Y|X). \tag{3.2}$$

We can view this latter distribution as an interaction effect of T and X on Y.

Theorem 2. *The task-specific distribution is the sum of the overall distribution and a task-specific pure interaction effect:*

$$Y|X, t = Y|X + \rho(Y|X, t) \tag{3.3}$$

where $\rho(Y|X, t)$ is a pure interaction effect.

[1] For this section, we consider only supervised learning of conditional distributions.

Proof. We show that $\rho(Y|X,t)$ satisfy the integral conditions of a pure interaction effect:

$$\rho(Y|X,t) = Y|X,t - Y|X \tag{3.4}$$

$$\int_t \rho(Y|X,t)p(X,t)dt = Y|X - \int_t Y|X,tp(X,t)dt = Y|X - Y|X = 0 \tag{3.5}$$

$$\int_{X_j} \rho(Y|X,t)p(X,t)dX_j = \int_{X_j} (Y|X_j, X_{\setminus j}, t - Y|X_j, X_{\setminus j})p(X,t)dX_j = 0 \tag{3.6}$$

□

This means that any estimator of $Y|X,t$ also gives us an estimator of $Y|X$ and $\rho(Y|X,t)$ because the interactions between X and T are exactly the differences between task distributions. Thus, if we regularize against interactions between X and T (as with Dropout on T), we can encourage similarity in the task-specific distributions. In addition, we can use the purification algorithm from [106] to recover the task-specific interaction and the main effects from $Y|X,t$.

3.2 Heterogeneous Models for Sample Sub-populations

Defining multitask learning as above shows us that the task of estimating different models for different samples is a form of multitask learning in which sample representations U are used as task representations t. Notice that as defined above, multitask learning does not require discrete task identifiers or specific algorithms for sharing models between tasks. Thus, in our transition to regarding sample representations as task identifiers, we are free to use whichever sample representation most accurately captures the patterns of variation underlying sample heterogeneity (if there is no underlying variation and only stochastic noise, the optimal would be sample representations which are all identical and thus produce a single population model). Our goal, then, in order to build a framework for learning heterogeneous models which vary between samples is two-fold: to identify meaningful sample covariates U which describe the processes underlying each sample, and to construct a framework to estimate similar $Y|X,U$ for similar U and dissimilar $Y|X,U$ for dissimilar U.

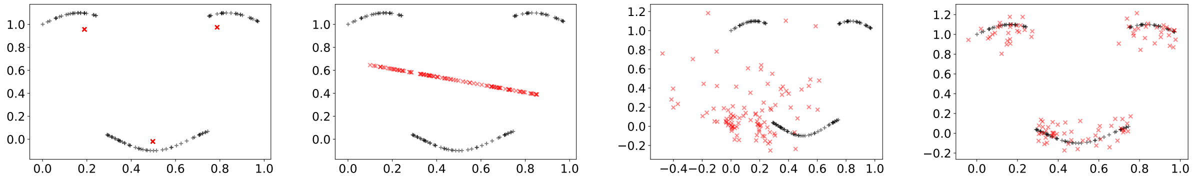
Heterogeneous Models and Interaction Effects

As a concrete example, let us consider the function

$$Y = \begin{cases} aX_1 + bX_2 & X_1X_2 < 0 \\ cX_1 + dX_2 & X_1X_2 \geq 0 \end{cases}$$

which generates outcomes for uncorrelated features X_1, X_2. This function can be captured in a simple, intelligible heterogeneous model $Y = \langle \theta(X_1, X_2), X_1 \rangle$ where $\theta(X_1, X_2) = [a + (c - a)\text{sign}(X_1X_2), b + (d - b)\text{sign}(X_1X_2)]$. In this form, we can inspect the value of θ for each sample and use the parameter values to readily identify the equivalence of models within each quadrant.

Alternatively, we could use a black-box model capable of modeling high-order interaction effects to fit this data. After fitting this model to the data, we could decompose it into main and



(a) Mixture model (b) Varying-Coefficient (c) Deep Neural Net (d) Personalized

Figure 3.1: Illustration of the benefits of personalized models. Each point represents the regression parameters for a sample. Black points indicate true effect sizes, while the red points are estimates. Mixture models (a) estimate a limited number of models. The varying-coefficients model (b) estimates sample-specific models but the non-linear structure of the true parameters violates the model assumptions, leading to a poor fit. The locally-linear models induced by a deep learning model (c) do not accurately recover the underlying effect sizes. In contrast, personalized regression (d) accurately recovers effect sizes.

interaction effects according to the algorithm in [106]. The resulting decomposition is proportional to

$$\begin{aligned}
F(X_1, X_2) &= f_1(X_1) + f_2(X_2) + f_{12}(X_1, X_2) \\
f_1(X_1) &= aX_1 \\
f_2(X_2) &= bX_2 \\
f_{12}(X_1, X_2) &= (a - c)X_1\mathcal{I}(X_1X_2 \geq 0) + (b - d)X_2\mathcal{I}(X_1X_2 \geq 0)
\end{aligned}$$

These forms represent equivalent functions, but the heterogeneous form is much easier to interpret. Intelligible parameters assist designers of regularization schemes, so the heterogenous model is preferable if we have prior knowledge to encode in the sample-specific. In much of the remainder of this thesis, we propose to use external data modalities to infer this prior knowledge and make adapative regularization schemes which dynamically tie samples together based on learned sample similarity.

3.3 The Extreme: Sample-Specific Models

Now we ask the question of what happens at the extreme: what if each sample is considered its own task? In other words, what if the sample representations U are unique (possibly due to continuous-valued attributes)? Despite the long history of statistical studies of heterogeneity, few methods have been designed to estimate sample-specific effects, and the ones which do typically require prior knowledge regarding the relation between samples (e.g. a network) [140]. At the same time, as datasets continue to increase in size and complexity, the possibility of inferring sample-specific phenomena by exploiting patterns in these large datasets has driven interest in important scientific problems such as precision medicine [18, 141]. The relevance and potential impact of sample-specific inference has also been widely acknowledged in applications including psychology [52], education [69], and finance [2].

Here, we explore a solution to this dilemma through the framework of "personalized" models. Personalized modeling seeks to estimate a *large* collection of *simple* models in which each model is tailored—or "personalized"—to a single sample. This is in contrast to models that seek to estimate a *single*, *complex* model. To make this more precise, suppose we have n samples $(X^{(i)}, Y^{(i)})$, where $Y^{(i)}$ denotes the response and $X^{(i)} \in \mathbb{R}^p$ are predictors. A traditional ML model would model the relationship between $Y^{(i)}$ and $X^{(i)}$ with a single function $f(X^{(i)}; \theta)$ parametrized by a complex parameter θ (e.g. a deep neural network). In a personalized model, we model each sample with its own function, allowing θ to be simple while varying with each sample. Thus, the model becomes $Y^{(i)} = f(X^{(i)}; \theta^{(i)})$. These models are estimated jointly with a single objective function, enabling statistical power to be shared between sub-populations.

The flexibility of using different parameter values for different samples enables us to use a simple model class (e.g. logistic regression) to produce models which are simultaneously interpretable and predictive for each individual sample. By treating each sample separately, it is also possible to capture heterogeneous effects *within* similar subgroups. Finally, the parameters learned through our framework accurately capture underlying effect sizes, giving users confidence that sample-specific interpretations correspond to real phenomena (Fig 3.1).

Motivating Example

Let us consider the problem of understanding election outcomes at the local level. For example, given data on a particular candidate's views and policy proposals, we wish to predict the probability that a particular locality (e.g. county, township, district, etc.) will vote for this candidate. In this example we focus on counties for concreteness. More importantly, in addition to making accurate predictions, we are interested in understanding and explaining how different counties react to different platforms. The latter information—in addition to simple predictive measures—is especially important to candidates and political consultants seeking advantages in major elections such as a presidential election. This information is also important to social and political scientists seeking to understand the characteristics of an electorate and how it is evolving. An application of this motivating example using personalized regression can be found in in Section 5.

One approach would be to build individual models for each county, using historical data from previous elections. Immediately we encounter several practical challenges: 1) By building independent models for each county, we fail to share information between related counties, resulting in a loss of statistical power, 2) Since elections are relatively infrequent, the amount of data on each county is limited, resulting in a further loss of power, and 3) To ensure that the models are able to explain the preferences of an electorate, we will be forced to use simple models (e.g. logistic regression or decision trees), which will likely have limited predictive power compared to more complex models. This simultaneous loss of power and predictive accuracy is characteristic of modeling large, heterogeneous datasets arising from aggregating multiple subpopulations. Crucially, in this example the *total* number of samples may be quite large (e.g. there are more than 3,000 US counties and there have been 58 US presidential elections), but the number of samples per subpopulaton is small. Furthermore, these challenges are in no way unique to this example: similar problems arise for examples in financial, biological, and marketing applications.

One way to alleviate these challenges is to model the ith county using a regression model $f(X;\theta^{(i)})$, where the $\theta^{(i)}$ are parameters that vary with each sample and are trained jointly using *all of the data*. This idea of *personalized modeling* allows us to train accurate models using only a single sample from each county—this is useful in settings where collecting more data may be expensive (e.g. biology and medicine) or impossible (e.g. elections and marketing). By allowing the parameter $\theta^{(i)}$ to be sample-specific, there is no longer any need for f to be complex, and simple linear and logistic regression models will suffice, providing useful and interpretable models for each sample.

Alternative approaches and related work. One natural approach to adapt to heterogeneity is to use mixture models, e.g. a mixture of regression [154] or mixture of experts model [61]. While mixture models present an intriguing way to increase power and borrow strength across the entire cohort, they are notoriously difficult to train and are best at capturing coarse-grained heterogeneity in data. Importantly, mixture models do not capture individual, sample-specific effects and thus cannot model heterogeneity *within* subgroups.

Furthermore, previous approaches to personalized inference [65, 116, 196, 197] assume that there is a *known* network or similarity matrix that encodes how samples in a cohort are related to each other. A crucial distinction between our approach and these approaches is that no such knowledge is assumed. Recent work has also focused on estimating sample-specific parameters for structured models [88, 89, 98, 111, 116, 188]; in these cases, prior knowledge of the graph structure or causal constraints enables efficient testing of sample-specific deviations.

More classical approaches include varying-coefficient (VC) models [49, 71, 171], where the parameter $\theta^{(i)} = \theta(U^{(i)})$ is allowed to depend on additional covariates U in some smooth way, and random effects models [90], where θ is modeled as a random variable. More recently, the spirit of the VC model has been adapted to use deep neural networks as encoders for complex covariates like images [4, 5] or domain adaptation [151, 163]. In contrast to our approach, which does not impose any regularity or structural assumptions on the model, these approaches typically require strong smoothness (in the case of VC) or distributional (in the case of random effects) assumptions.

Frameworks of Sample-Specific Model Parameters

Several frameworks have been proposed to estimate such models which use side information u to generate parameters for models which operate on x. In general, we can write the operation of these models with sample-specific parameters as $f(x^i;\theta^i)$ where

$$\theta^i = z^i Q, \tag{3.7}$$

$$z^i = g(u^i;\phi)^T. \tag{3.8}$$

This formulation encompasses several framework of sample-specific models:

- If g is a linear model and Q is an identity matrix, we have the varying-coefficients model [71].
- If g indexes the k-nearest neighbors in Q by a learned distance metric, we have low-rank personalized regression [108, 109].

- If g is a deep neural network and $||z^i|| = 1$, we have Contextual Explanation Networks (CEN) [4].
- If g is a few steps of gradient-descent applied to the instances in u^i beginning at the initialization ϕ, we have meta-learning.

Note here that we are not specifying the learning algorithm used to optimize the dictionary elements. For instance, CEN and the VC model optimize a loss defined on the sample-specific models, while personalized regression makes employs distance-matching regularization to fill in Q and a nearest neighbor criterion to select models from Q at test time.

Lessons

In this chapter, we have introduced the use of sample representations as task representation to estimate different models for different samples. We see that task-specific models are implicitly learning pure interactions between task representations and model parameters. As a result, heterogeneous models which change parameters are learn pure interactions between sample representations and model parameters. We suggest taking this approach to the extreme and estimating *sample-specific* model parameters which use continuous-valued sample representations to estimate model parameters which vary continuously between samples. Several existing methods of heterogeneous model estimation fit into this paradigm, and next seek to apply these methods to real data.

Chapter 4

Sample-Specific Models to Identify Discriminative Subtypes of Lung Cancer

In this chapter, we explore an application of sample-specific models to understand transcriptomic signatures of cancers. To design individualized treatment protocols for cancer patients, clinicians must synthesize information from multiple data modalities into a single, parsimonious description of the patient's condition. However, the most informative description of each patient's disease is fundamentally unknown, and thus tools for automatic, personalized subtyping of cancer have not been built. In this work, we propose to describe patient conditions with latent *discriminative subtypes*: sample representations which give the most informative context for a model to understand and make accurate predictions about the tumor. According to this definition, discriminative subtypes can be estimated from one data modality and used to improve predictions about the patient in another modality. We apply contextual deep learning to extract these discriminative subtypes from lung cancer histopathology images. Based on these subtypes, our framework produces sample-specific transcriptomic models which accurately classify samples as adenocarcinoma, squamous cell carcinoma, or healthy tissue (F1 score of 0.97, achieving a new state-of-the-art). Combining these data modalities via contextualization not only improves the predictive accuracy but also gives biological interpretations of subtypes and ties the morphological patterns present in histopathology images to transcriptomic processes.

Portions of this chapter are available as [104].

4.1 Motivation

In many biological settings, inter-sample heterogeneity is critical to understanding the complex biological processes under study. For example, in the analysis of cancer gene expression assays, each patient in a cohort may have a different somatic mutation. As these mutations adjust the baseline gene expression levels of each patient, the observed gene expression data must be interpreted with respect to patient-specific "contexts." These contexts can be described as *molecular subtypes* which are associated with patient outcomes [161] and can inform the decision-making process behind surgery and radiation therapy [38]. Our understanding of molecular subtypes is still advancing, and much recent work has focused on identifying fine-grained descriptions which

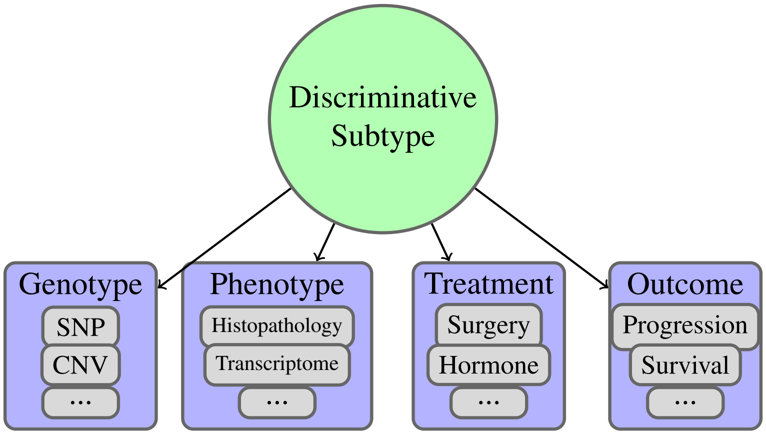


Figure 4.1: The most concise description of the many variables regarding oncology patients is an unobserved latent variable. In this work, we seek to estimate this latent variable as a *discriminative subtype* which improves the discriminative ability of downstream predictors.

comprise previously-characterized subgroups (e.g. in breast [177] or lung cancers [92]).

If the perfect contexts were annotated (e.g., if molecular subtypes were to correspond exactly to primary tissue sites), we could simply train a different probabilistic model for each context. However, meaningful contexts are often complex and possibly unknown. Simply increasing the capacity of the predictive model to handle multiple contexts is typically not feasible due to limited sample size and an emphasis on fitting interpretable models. Instead, we would like to use a model which captures the complexity of different contexts while retaining the domain knowledge and interpretability of structured probabilistic models.

To do so, we introduce the notion of a *discriminative subtype* (Figure 4.1). A discriminative subtype is a latent variable which captures the variation in many observable variables, and can be used for a variety of downstream tasks. Discriminative subtypes may correspond, but are not limited, to previously-characterized molecular subtypes.

In this work, we will use a deep neural network to estimate the discriminative subtype for each sample based on contextual data. We will use this subtype information to generate the parameters for an interpretable model for downstream tasks. This two-stage procedure allows us to analyze transcriptomic data in the context of complex context data such as diagnostic images. Specifically, we investigate the capacity of this architecture to predict disease subtypes from the histopathology images and transcriptomic data of patients diagnosed with two types of lung cancer: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC).

Lung Cancers

Lung cancer is one of the top ten most common cancers for males and females, and is estimated to account for roughly 13% of all new cancer cases[173]. For males, it is second only to prostate cancer, and for females it is second only to breast cancer. In terms of mortality, it is estimated to be the most lethal cancer for both sexes, accounting for 23% of cancer deaths in males and 22% of cancer deaths in females for the year 2020 [173]. It is estimated that in 2020 in the United

States, 228,820 new cases of lung cancer will be diagnosed and 135,720 people will die from lung cancer [173].

The most common category of lung cancer is Non-Small Cell Lung Cancer (NSCLC), accounting for 85% of cases. LUAD and LUSC are subtypes within this NSCLC category, and are the two most common types of lung cancers, accounting for 38.5% and 20% of all lung cancer cases respectively [41]. When lung cancer is diagnosed, it is categorized into one of four stages based on the TNM grading system (based on (T) the size of the tumor, (N) the spread of cancer to lymph nodes, and (M) metastasis) where a lower stage number means that there is less cancer in the patient. Early detection is critical in lung cancer: in the U.S., 61% of NSCLC patients diagnosed at the localized stage survived after five years compared to only 6% of those diagnosed at the distant stage.[1]

While LUAD and LUSC are both major subtypes of lung cancer, they have important distinctions in terms of their clinicopathology, tumor microenvironments, and molecular organization. For example, LUSC is associated with a history of smoking, while LUAD may occur in smokers but is the most common form of lung cancer in non-smokers [159]. These differences have implications for the best therapeutic approaches for each subtype. For example, epidermal growth factor receptor (EGFR) and anaplastic lymphoma kinase (ALK) are two commonly mutated genes in LUAD and are targets for current therapies, while they are infrequently mutated in LUSC [112]. In addition, LUAD and LUSC tumors each have different immune subtypes (tumor subtypes characterized by differences in infiltrating immune cell types and immunogenic expression), and these differences have been shown to have an effect on the efficacy of immune therapies (immune checkpoint blockade therapy) [169]. Given this variety of subtypes at different pathological levels, estimating a discriminative subtype that encapsulates these variations will be essential for machine learning tasks such as cancer type prediction.

Contributions and Generalizable Insights

As described above, effective machine learning systems for oncology require both the integration of data modalities (e.g. temporal, demographic, radiological, histopathological data) and model interpretability. The generic form of this problem is a significant open problem in machine learning. In this manuscript, we demonstrate that we can apply contextual deep learning to simultaneously achieve accuracy and interpretability for classification of lung cancers. This case study brings to light the following insights that apply in general to the problem of using machine learning techniques on multi-modal patient data:

- In applications with heterogeneous samples, understanding the context in which predictions are made is important to making accurate predictions. In this work, we show that sample-specific contexts can improve downstream predictions even when the contexts are latent and must be inferred from data (Section 4.3). Towards this end, we show that contextual deep learning is a promising tool for estimating these latent variables and linking different data modalities in a single pipeline.
- For transcriptomic analysis of lung cancer samples, we show that meaningful contexts can be inferred from histopathology samples. These meaningful contexts form *discriminative*

[1] `cancer.org`

subtypes (Section 4.3), which are not themselves correlated with the cancer type, provide meaningful context in which simple models can accurately label sample types.

- Finally, the learned transcriptomic models for different discriminative subtypes tend to give high attention to different biological processes (Section 4.3). This suggests that the sample-specific transcriptomic models learned from our discriminative subtypes are capturing the heterogenity of the biological processes underlying lung cancer. This highlights the promise of approaches which use contextual deep learning to learn biologically explainable/meaningful models for heterogenous data.

Related Work

Molecular profiling has increased understanding of the pathology of cancer [10, 34, 155, 183]. This understanding allows pathologists to use phenotypic molecular and morphological data to identify increasingly specific cancer subtypes [60, 84, 168]. Improved cancer subtype identification enables clinicians to devise more effective individualized treatment plans for patients with cancer. The machine learning community is increasingly working to assist pathologists further improve cancer subtype identification to better inform treatment based on molecular and morphological data [21, 58, 82, 96, 132, 138, 179, 186, 192, 193]. This common goal of better informing treatment has also led the machine learning community to work towards creating patient specific interpretable and explainable predictive models to provide actionable insights to clinicians about the prognosis of their patient's disease [132, 206].

Cancer Subtype Identification

Identifying cancer subtypes is crucial to devising individualized treatment protocols for patients with cancer, but subtype identification has been hindered by the lack of knowledge of the biological processes underlying tumor growth [44, 67, 73, 172]. In previous work, this was circumvented by using observed phenotypic and/or genotypic data to define cancer subtypes [21, 58, 82, 96, 132, 138, 179, 192, 193]. These papers defined specific cancer subtypes based on phenotypic and/or genotypic features which were meaningful for specific downstream tasks such as predicting healthy versus cancer. Recently, a new definition of subtype was introduced in which is based on mutations predicted from histopathology images within a clinically accepted cancer type [36]. This approach is promising, as it could assist pathologists detect gene mutations to inform treatment. While all of these definitions of a cancer subtype are useful, the features they identify only capture a subset of the aspects of the biological processes underlying tumor growth related to a specific task or data type. Our work is the first to our knowledge to associate discriminative models with cancer subtypes, which we call *discriminative subtypes* and infer from morphological data. This definition enables us to learn accurate patient-specific models for downstream prediction tasks.

Downstream Clinical Prediction

The key challenge to prediction of downstream clinical outcomes is to use multimodal data to generate predictions that are interpretable and explainable to clinicians. In previous work, interpretable predictive models have been based on only one type of data, such as histologic patterns which are interpretable to pathologists [36, 58, 120, 121, 192, 193, 201]. Multimodal models have leveraged different types of data to increase predictive capabilities, but their interpretability has been hindered by the complexity of the models [29, 68, 132, 152, 206]. The only way some of these complex models have been interpreted is with a retrospective analysis of the features selected in their neural networks, and it is unclear how to use this type of interpretation to inform clinical decisions [68, 132]. Our approach allows us to generate predictive models linking histopathology images and molecular data which are accurate, interpretable, and explainable to clinicians.

Sample-Specific Model Parameters

In this work, we identify discriminative subtypes by estimating a basis set of models which can be utilized to generate model parameters specific to each sample. Interest in such *sample-specific* model parameters has grown in recent years, as increasing evidence has pointed to fine-grained subtypes which do not form discrete clusters [18, 122]. Discovering these individualized molecular profiles could lead to substantial advances in the diagnosis and treatment of disease, in addition to refining our biological understanding of the mechanisms behind different diseases. Despite the recent surge of interest in personalized modeling [6, 97, 115, 119, 135, 187], basic statistical challenges remain unanswered and deserve further study. Patient-specific analysis of cancer patients has mainly focused on two areas: detection of personalized biomarkers [12, 108, 109, 137], and inference of personalized regulatory networks [20, 89]. In this work, we contextualize such personalized models with imaging data. We hope that this idea of contextualization can spur further development of such patient-specific modeling efforts that appear promising.

4.2 Methods

To combine multi-modal data in an interpretable pipeline, we use Contextual Explanation Networks (CENs) [4], illustrated in Figure 4.2. CEN architecture assumes that the data is represented by: (i) context variables, denoted C, (ii) semantically meaningful variables, denoted X, and (iii) target variables, denoted Y. Here, we will use histopathology images as contextual data and gene expression assays as semantically meaningful predictors of the cancer type. The model will represent conditional probability of the cancer type given the histopathology and transcriptomic inputs, $\mathbb{P}(Y \mid X, C)$, in the following form:

$$\mathbb{P}(Y \mid X, C) = \int \mathbb{P}(Y \mid X, \theta)\delta(\theta = \phi_w(C))d\theta = \mathbb{P}(Y \mid X, \phi_w(C)), \tag{4.1}$$

where $\mathbb{P}(Y \mid X, \theta)$ is a linear logistic model that predicts cancer types from gene expression information. Note that parameters (or weights) θ of the logistic model are a function of the contextual information, i.e., $\theta = \phi_w(C)$. In other words, CEN architecture produces a sample-specific parameterization of a linear probabilistic model that operates on transcriptomic data based on the sample's contextual data (i.e., based on histopathology).

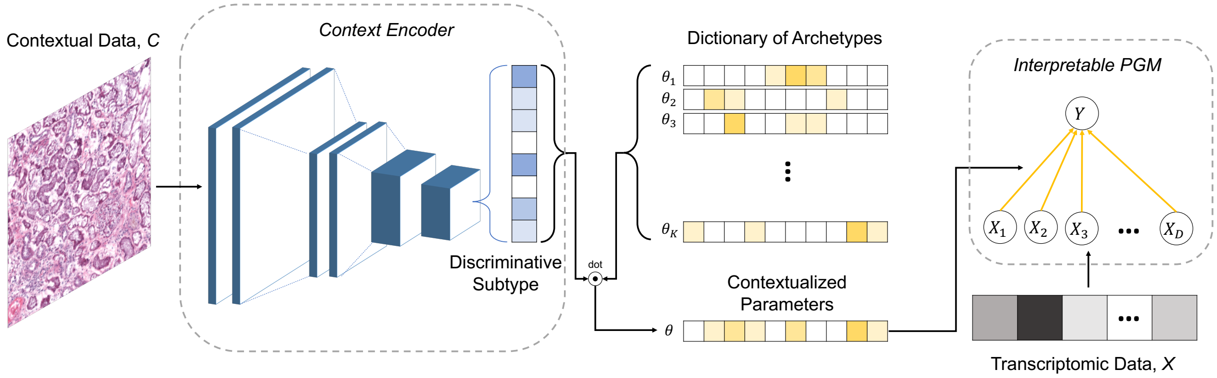


Figure 4.2: A CEN architecture combines "contextual" histopathology data (C) with gene expression data (X) through intermediate, sample-specific interpretable linear probabilistic models. We call the output of the context encoder the discriminitative subtype, as it is used to form sample-specific parameters θ by weighting the archetypal models stored in a dictionary. Finally, the formed sample-specific linear models operate on the transcriptomic data (X) to estimate the final target Y (in our experiments, cancer type).

Generation of parameters for sample-specific linear models is accomplished via a context encoder represented by a convolutional neural network (Figure 4.2). To reduce model complexity, parameters θ are further confined to be a linear combination of a small constant number (K) of "archetypes," denoted $\{\theta_1, \ldots, \theta_K\}$ or $\theta_{1:K}$. Specifically for a sample i with context c_i, weights θ_i are computed as follows:

$$\theta_i = \sum_{k=1}^{K} \alpha_w^k(c_i)\theta_k, \quad \text{where } \alpha_w^k(c_i) \geq 0,\ k = 1, \ldots, K \text{ and } \sum_{k=1}^{K} \alpha_w^k(c_i) = 1, \tag{4.2}$$

where $(\alpha^1, \ldots, \alpha^K)$ is a vector output of the context encoder, which we call *discriminative subtype*. This dictionary of parameter sets, $\theta_{1:K}$, is learned jointly with the context encoder, and the entire architecture is trained end-to-end via backpropagation. To summarize, the context encoder is a deep neural network that processes contextual data (histopathology imagery) and outputs a probability vector of length K that softly selects weights for a linear probabilistic model from a dictionary of archtetypes.

Architectural Components To encode histopathology imagery, following Coudray et al. [36], we used `InceptionV3` [180] architecture as our context encoder. Since this architecture naturally operated on images of size 299x299, to utilize it on the large histopathology images, we sliced the images into non-overlapping patches of size 299x299, filtering out patches that had more than 25% of background.

Further, for the interpretable probabilistic part of CEN, we used logistic regression (LR) that predicted cancer type from gene expression data. Predictions were computed both on the level of patches (in which case patches from the same slide were assigned the same slide-level labels and gene expression data) and on the level of slides (using majority vote over the corresponding patches).

Baselines

To benchmark the performance of the CEN model, we compare against both uni-modal and multi-modal baselines. Firstly, we compare against the two uni-modal parts of the CEN architecture: logistic regression on transcriptomic profiles and the convolutional neural network on histopathology images. In addition, we compare performance against two multi-modal baselines. Multi-modal baseline 1 ("Concatenated") is the same architecture used for the state-of-the-art subtype prediction in [132]. It predicts labels from a concatenation of the RNA-seq features and the output of the InceptionV3 network that encodes histopathology. Multi-modal baseline 2 ("Ensemble") is an ensemble (i.e., a weighted combination) of the two uni-modal models described above.

Cohort

For this investigation, we use lung cancer data available in the NCI Genomic Data Commons [63], which includes both TCGA and TCIA resources to provide multi-modal descriptions of a large number of cancer patients. As of April 2, 2020, this depository includes 585 LUAD cases and 504 LUSC cases, along with a variety of cases with cancer of other tissues. From this set, we selected all samples which have both eosin-stained histopathology whole-slide images and transcriptomic RNA-seq profiling. This reduces the dataset to a total of 992 patients. We selected these datatypes because inference of discriminative subtype is most clinically-useful if it can be done from phenotypic data. In addition, we look forward to the availability of transcriptomic data from multiple time points to make dynamic predictions which track tumor progression.

We split this dataset into training, validation, and testing partitions. The size and composition of each partition is shown in Table 4.1.

Partition	**Healthy**		**LUAD**		**LUSC**		**Total**	
	patches	slides	patches	slides	patches	slides	patches	slides
Training	59,530	89	658,498	562	722,396	518	1,440,424	1,169
Validation	10,478	17	132,101	120	156,698	116	299,277	253
Test	11,899	18	118,521	124	183,946	113	314,366	255

Table 4.1: Dataset sizes by counts of slides and patches. The slides are divided into patches.

Feature Choices

Image Subsampling. For the image analysis, we use the 20x magnification whole-slide images. To make these large images suitable for the pretrained Inception architecture, we split each whole-slide image into non-overlapping 299x299 image patches. The patches were controlled for quality by discarding any patches with more than 25% white pixels, following the procedure of Coudray et al. [36]. We treat each patch as a separate sample, with transcriptomic data duplicated over each patch.

Table 4.2: Performance of models on 3-way Normal/LUAD/LUSC classification. We report accuracy on both the "patch" level (where each prediction corresponds to a small patch in the image) and on the "sample" level (where a single prediction is made for the whole-slide image). Some models are designed to operate on transcriptomic data (T), while some operate on histopathology data (H), while others use both forms of data (H+T). CEN (our model) outperforms all existing models.

Level	Model	Data	Accuracy (%)	Macro F1
Patch	**CEN**	**H+T**	**96.18**	**96.97**
	Concatenated	H+T	95.32	93.65
	Ensemble	H+T	94.61	90.23
	Logistic Regression	T	94.05	91.40
	`InceptionV3`	H	69.14	65.85
Sample	**CEN**	**H+T**	**94.51**	**95.29**
	Concatenated	H+T	93.33	93.65
	Ensemble	H+T	92.94	90.23
	Logistic Regression	T	92.16	89.67
	`InceptionV3`	H	80.00	76.76

Transcriptomic Profiles. The transcriptomic profiling in TCGA captures the expression of over 60,000 distinct transcripts in each sample. To reduce this dimensionality, we select the 1000 transcripts with the highest variance in the non-lung cancer cases in TCGA. In addition, we augment this set with 695 transcripts corresponding to genes in the Catalogue of Somatic Mutations in Cancer (COSMIC, [11]). This leaves us with a final set of 1695 transcriptomic features.

4.3 Results

The goal of our study is not only to evaluate the predictive ability of the proposed multi-modal approach, but also to study the biological meaning and clinical utility of the inferred clusters. Towards that end, first, we measure predictive accuracy compare performance of our model against unimodal approaches: `InceptionV3` trained to predict cancer type from histopathology only [36] and logistic regression that predicts cancer type from transcriptomic data only. Then, we seek to assign biological meaning to the learned model archetypes, find descriptive histopathology images for each archetype, and analyze the discovered discriminative subtypes.

Contextualization Improves Prediction

First, we measure the ability of the model to discriminate between LUAD, LUSC, and healthy samples. As shown in Table 4.2, the multi-modal CEN achieves the best classification results, outperforming both the convolutional neural network (which operates solely on histopathology images, had `InceptionV3` architecture) and the regularized logistic regression (which operates

Table 4.3: Slide-level confusion matrices for baselines and CEN. Rows correspond to ground truth labels and columns to predictions made by the corresponding models.

	CEN			Logistic Regression			InceptionV3		
	Healthy	**LUAD**	**LUSC**	**Healthy**	**LUAD**	**LUSC**	**Healthy**	**LUAD**	**LUSC**
Healthy	18	0	0	15	1	2	12	5	1
LUAD	1	117	6	3	115	6	4	102	18
LUSC	0	7	106	0	8	105	1	22	90

	Concatenated			Ensemble		
Healthy	**LUAD**	**LUSC**	**Healthy**	**LUAD**	**LUSC**	
Healthy	17	1	0	15	1	2
LUAD	1	115	8	3	115	6
LUSC	0	7	106	0	6	107

solely on transcriptomic data). Our results for the `InceptionV3` model come very close to the performance reported by [36]. Based on these accuracy and macro F1 metrics, we can see that most of the signal to distinguish the cancer types is held in the transcriptomic data, but a single linear transcriptomic model is not flexible enough to achieve the best accuracy over the diverse cohort.

Archetypal Models Correspond to Biologically Meaningful Processes

Next, we turn to interpret the learned models. A natural question is to what degree the different model archetypes are redundant or look at different biological processes. Surprisingly, the model archetypes coefficient vectors are nearly all orthogonal (Fig. 4.3), indicating that the archetype emphasize distinct biological processes.

To identify these biological meanings, we turn to enrichment analysis. For each archetype, we have 3 vectors of coefficients defining the 3-way logistic regression model. For each vector, we sort the genes by the magnitude of the associated coefficient and we search the top 100 genes in this order for enriched biological terms against the background of the remaining 1595 genes using gProfiler[157]. We set a minimum intersection size of 3 genes (the number of genes selected from each term must be at least 3) and a maximum Bonferroni-corrected p-value of 0.05. Terms enriched in the archetypal models for prediction of the “normal” label (which corresponds to case/control prediction) are shown in Table 4.4. Here we see that a large number a of the 32 archetypal models are significantly enriched for a diverse set of processes which are markers for cancer.

As a vignette, below we discuss the significance terms in Table 4.4 which have a p-value of less than 0.01:

- **13: Factor: Smad4.** Intersection: CRNKL1, MYH11, ZRSR2.
  SMAD family member 4 (SMAD4) is a necessary component of the transforming growth factor beta (TGFβ) pathway and in this capacity regulates proliferation [199, 203, 205].

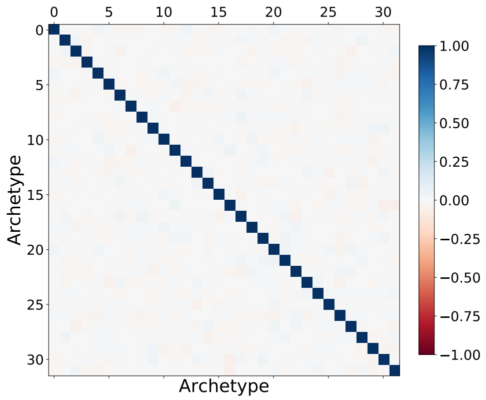


Figure 4.3: Kendall Tau similarity of ranks of transcripts selected by each entry in the dictionary. Nearly all off-diagonal comparisons have similarity less than 0.1, indicating that the models are orthogonal.

SMAD4 has been established as a tumor suppressor in pancreatic, colon and lung cancer [64, 124, 203]. Decreased expression of SMAD4 has been identified in NSCLC, and has been shown to be correlated with poor prognosis in LUAD [32, 64, 83, 124].

- **18: Factor: REST.** Intersection: FGFR1, NCOA1, CRTC1, CBLB, DGCR8, PRDM16, BAP1, OTOS, FOXL2, ETV5.
  RE1-silencing transcription factor (REST) is most widely known to repress neuronal genes in non-neuronal genes, but has also been found to be an oncogene or tumor suppressor gene in certain cancers as well [139]. In breast cancer, it has been shown to act as a tumor suppressor, and loss of this gene is associated with aggressive breast cancer [190]. In lung cancers, an isoform of REST has been indicated as a specific clinical marker for early detection of small cell lung cancers [37, 170].
- **26: Apoptosis - multiple species.** Intersection: BIRC3, BIRC6, CASP3.
  A hallmark of cancer cells is their ability to limit or avoid cell death induced by apoptosis [67]. In NSCLC multiple genes in the KEGG apoptosis pathway have been found to be correlated with tumorigenesis, chemoresistance, and poor prognosis [47, 50, 62, 91, 91, 114, 153, 181, 200]. As an example, altered expression and/or localization of caspases which are involved in the crucial last steps of the apoptosis pathway have been shown to make cells resistant to apoptosis enabling cancer to progress and making cells resistant to chemotherapy [50, 91, 145, 181, 200]. Both cancer progression and resistance to chemotherapy increase the likelihood of a poor prognosis.
- **28: Positive Regulation of Multicellular Organismal Process.** Intersection: PTPRC, CRTC1, CPEB3, COL1A1, IL6ST, PIM1, LEF1, BCL10, RUNX1, ATP1A1, STAT3, MYD88, SETD2, BCL9L, SRC, MIR126.
  This Gene Ontology term is broadly defined as those processes which up-regulate or activate the progress of organismal processes [9, 35]. In cancer, oncogenes activate tu-

Table 4.4: Terms enriched in case/control archetypal models ($p < 0.05$).

Archetype	Term ID	Term Name	P-Val
1	KEGG:04071 KEGG:04310	Sphingolipid signaling pathway Wnt signaling pathway	0.037 0.049
6	REAC:R-HSA-6802952	Signaling by BRAF and RAF fusions	0.039
8	TF:M06732	Factor: ZNF304	0.023
12	REAC:R-HSA-8939236 GO:0010629	RUNX1 regulates transcription of genes involved in differentiation of HSCs negative regulation of gene expression	0.022 0.039
13	TF:M09657_1	Factor: Smad4	0.004
15	GO:0071385	cellular response to glucocorticoid stimulus	0.013
17	REAC:R-HSA-400206	Regulation of lipid metabolism by PPARα	0.018
18	TF:M04726_1	Factor: REST	0.001
19	TF:M05327_1	Factor: WT1	0.025
21	TF:M01224_1	Factor: P50:RELA-P65	0.034
25	TF:M09611_0	Factor: ER81	0.003
26	KEGG:04215	Apoptosis - multiple species	0.006
28	GO:0051240	positive regulation of multicellular organismal process	0.006
30	REAC:R-HSA-4791275	Signaling by WNT in cancer	0.045

mor growth and proliferation processes by up-regulating other development processes, including angiogenesis, cell migration, and metastasis. As an example, aggressive cancers achieve metastasis by up regulating the genes involved in inflammation and migration of cells, which include genes that produce cell-to-cell or cell-to-extra-cellular-membrane adhesion molecules [66, 67].

Morphologic Patterns of Archetypal Models

One of the advantages of contextual deep learning is that it enables us to tie the morphology patterns recognized in histopathology images into the biological processes used to describe the transcriptomic models. Towards this end, we visualize representative patches which maximize the influence of archetypal models 12, 13, and 30 in Figure 4.4. We can see that the morphology is more similar within the clusters corresponding to archetypes than between the clusters, indicating that the transcriptomic patterns used for accurate downstream predictions have a correspondence with morphological changes on a larger scale.

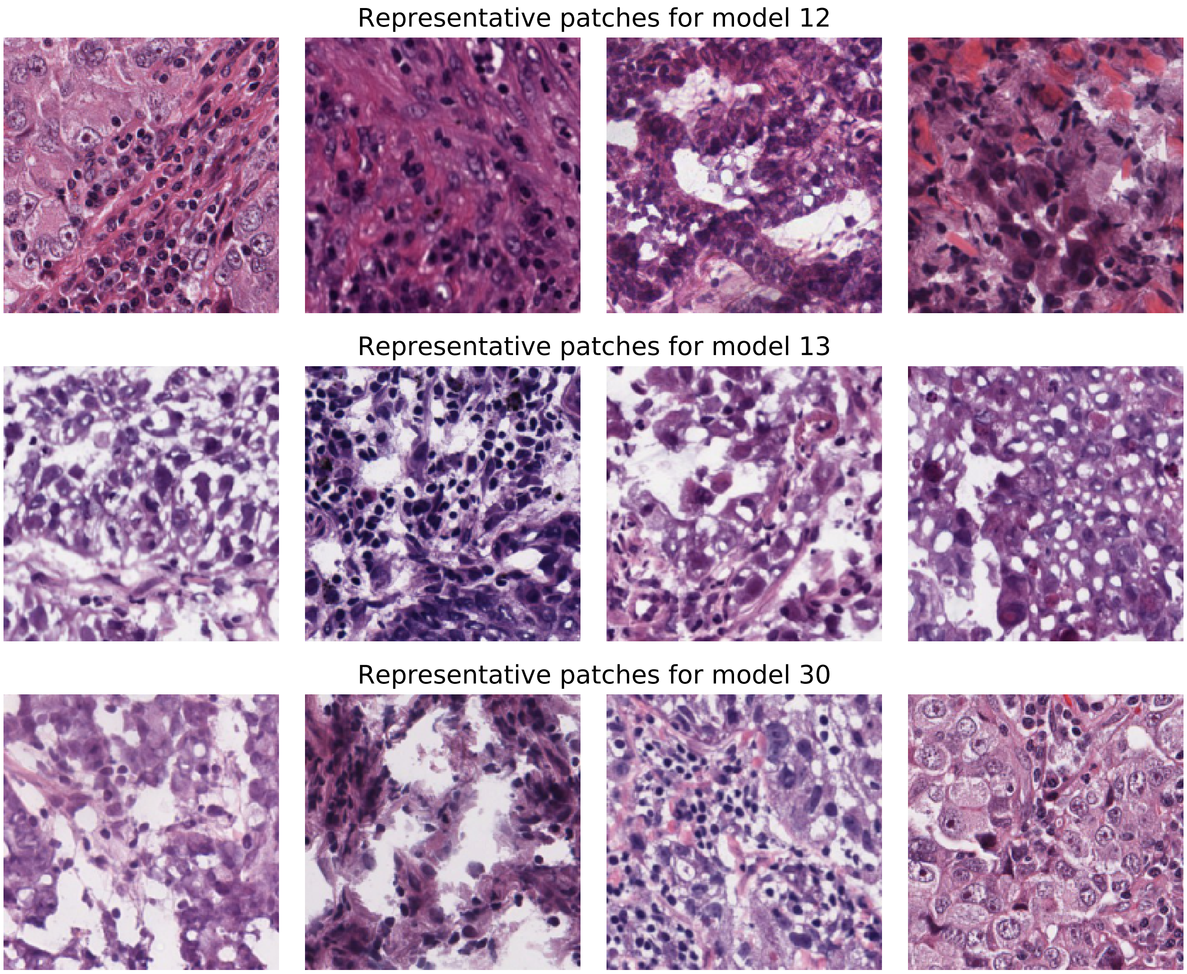

Representative patches for model 12
Representative patches for model 13
Representative patches for model 30

Figure 4.4: Patches that made CEN assign the highest weights to the corresponding models in the dictionary. Morphology of the patches is homogeneous within model-specific cluster and varies between the clusters.

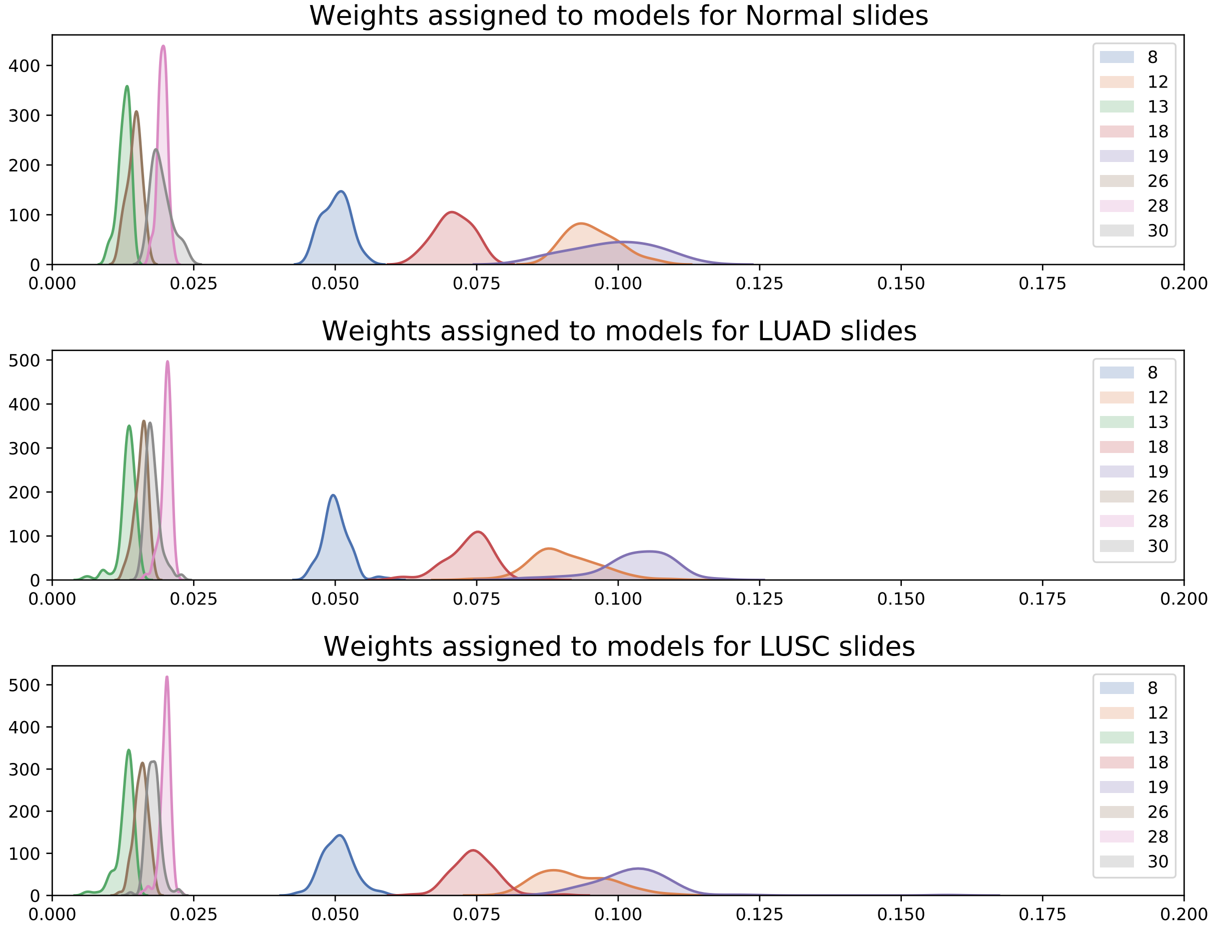


Figure 4.5: Class-conditional distributions over the weights assigned to different archetypes by CEN, visualized for healthy, LUAD, and LUSC slides. Observe only slight variation between distributions of the weights assigned to different archetypes, while the relative ordering between the archetypes being strictly preserved.

Discriminative Subtypes Display Intratumor Heterogeneity

The spatial diversity of cells within lung tumors has recently been shown to encode prognostic markers [1]; the CEN model enables us to ask questions regarding the localization of discriminative subtypes. By making predictions for each patch independently, the CEN model allows us to ask whether different locations in the image are indicative of distinct tumor subtypes and if morphological patterns of a particular subtype tend to be clustered in a single location. Firstly, we see that on held-out test data, in 46.3% of patches, the CEN assigns highest weight to a discriminative subtype which is not the discriminative subtype chosen for the full slide by a majority vote of the patches. This indicates that there may either be significant diversity or uncertainty of morphological patterns or tumor subtypes in each tumor. To examine the spatial locality, we visualize the predictions of discriminative subtypes. As shown in Figure 4.6, predictions of discriminative subtype exhibit spatial locality, with nearby patches tending to correspond to the same predicted subtype.

Analysis of the Discriminative Subtyping Hypothesis

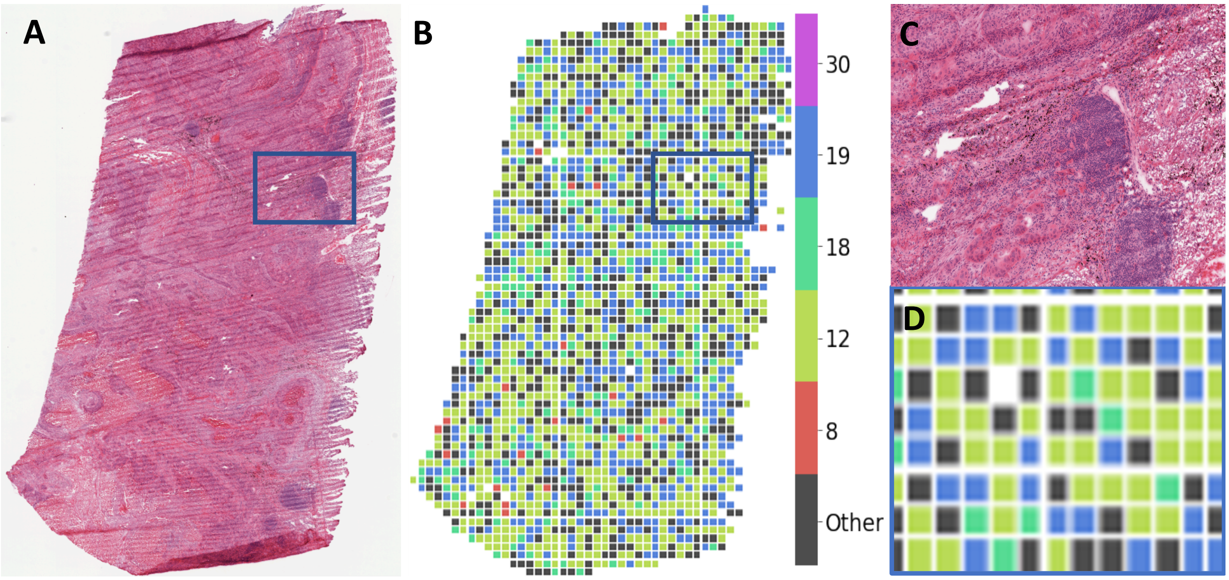
Our discussion started with a hypothesis that gene expression data must be interpreted with respect to patient-specific contexts that may correspond to different subtypes. The contextual learning approach we proposed to use in this work allowed us to discover latent subtypes, which we termed *discriminative* since each of the subtypes corresponded to an interpretable linear model that could accurately discriminate between different classes (cancer types in our case) based on transcriptomic information.

However, if samples of the same type always mapped to the same linear model, while samples of different types mapped to different models, this would have been troublesome—the biological interpretation of the archetypal coefficients (Section 4.3) would have been confounded by the assignment of different models to different classes. As shown in Figure 4.5, this is not the case: the distribution over the weights assigned to archetypes by the context encoder conditional on the target label are nearly identical across all classes. Thus, indeed linear models that correspond to each of the subtypes are ultimately discriminating between the classes.

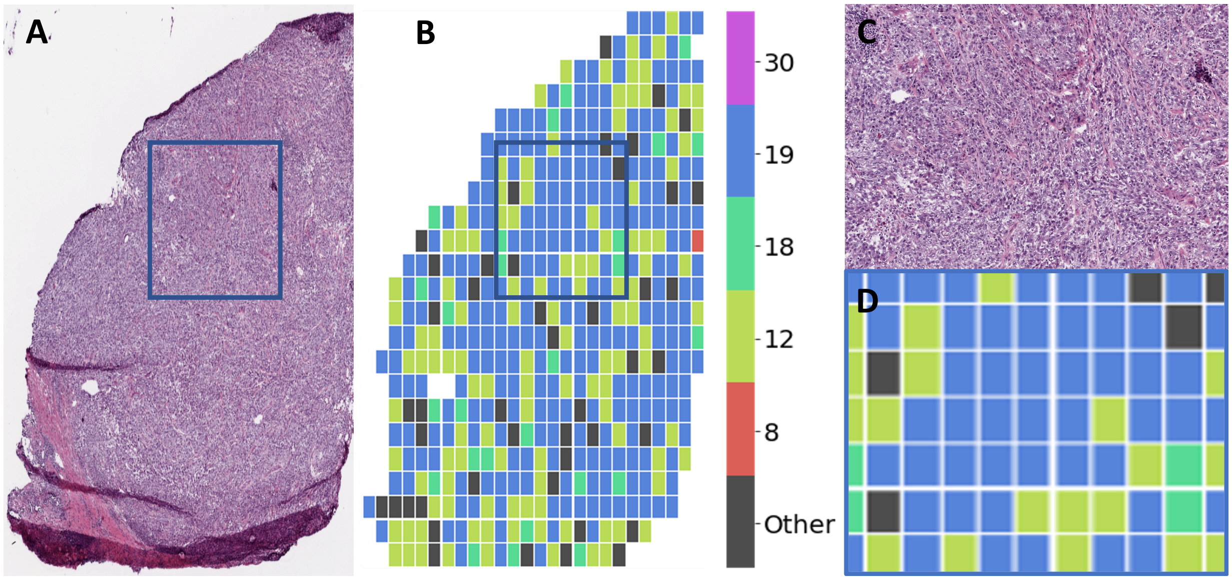
4.4 Discussion and Limitations

In this work, we have used contextual deep learning to estimate discriminative subtypes of lung cancers. These discriminative subtypes improve prediction of cancer type from transcriptomic data by allowing the sample-specific models to pay attention to different biological processes for different samples.

For machine learning practitioners, this has several implications. Firstly, multi-modal data analysis is critically important in healthcare applications (often because the technical noise in each measurement can only be reduced by having multiple views of each sample), but multi-modal pipelines are difficult to meld with interpretability. Contextual learning is a promising framework which can be used to design these interpretable multi-modal approaches. By allowing one data modality to select the model to be used on another modality, we can achieve both



(a) Representative results of patch-level subtype prediction for a LUSC sample held out from training.



(b) Representative results of patch-level subtype prediction for a LUAD sample held out from training.

Figure 4.6: Discriminative subtypes display significant heterogeneity throughout the sample, but cluster spatially. Pane A shows the tumor whole-slide image. Pane B displays the archetype which CEN assigned to each patch; for visual clarity, we display only the common archetypes 8, 12, 18, 19, and 30, and group all other archetypes under “other”. Pane C displays the inset square outlined in blue in panes A and B. Pane D displays the discriminative subtypes for the patches in the inset region. A similar result for a LUAD sample is shown in Figure S2, and matched control samples from the surrounding tissue are shown in Figure S3, S4.

accuracy and interpretability. This is extremely important for practitioners who want their machine learning models to be used in areas such as healthcare and public policy.

In addition, this work has clinical implications. Firstly, because the discriminative subtypes correspond to biologically-meaningful processes, we can surmise that these are distinctive signatures of cancers that may be used to improve personalized medicine beyond known histologic or molecular subtypes. Furthermore, in this study, a deep learning model trained on images was able to assign meaningful contexts for transcriptomic models, indicating that some patterns of transcriptomic aberrations are contained in histopathology data with only basic hematoxylin and eosin (H&E) staining. This concords with several recent works to predict genetic variants from H&E stained histopathology data [8, 56, 152], and we are excited to see future works improve on the connection between these datatypes to discover and analyze morphological patterns of cellular modifications. Future work in this area will enable the machine learning community to assist pathologists. Lastly, this approach of patient-specific modeling by contextualization has clinical relevance because it suggests that the inter-sample heterogeneity which hampers clinical predictive power may be overcome by more flexible modeling. Future work in this area may lead to machine learning practitioners being able to assist clinicians by providing patient-specific actionable insights about the prognosis of their patient's disease.

Limitations.

While our work advances our understanding of machine learning and healthcare, we note some important limitations. First, we only considered transcriptomic and histopathology data in our model. It would be beneficial to add known clinical confounders for lung cancer to our model such as smoking, age, and environment to add more relevant prognostic information to our model and enable prediction of more clinically-relevant outcomes such as treatment responses. Second, the labels for our histopathology data relied on the whole slide level labels provided by TCGA. While other works [36] have also used this labeling scheme, it adds noise to the labels because not every patch within a cancerous slide depicts cancerous tissue. Third, the distribution of healthy, LUAD, LUSC samples were imbalanced because we only used samples with both transcriptomic and histopathologic data and TCGA collected less transcriptomic data from healthy subjects. Future work should consider validating these results with an external balanced dataset. Lastly, we have focused on only one prediction task of classification. While this highlights the promise of contextual deep learning, we are excited for future work to use our approach to predict stage, survival, and optimal treatments.

Despite these limitations, our results suggest that with contextual deep models we can learn a context called a discriminative subtype from histopathological data and this discriminative subtype can produce sample-specific transcriptomic models which accurately classify LUAD, LUSC, and healthy tissue. Our work takes another step towards creating patient-specific interpretable predictive models of disease. We look forward to future work to expand on the use of contextual deep models to learn from multi-modal patient data.

Chapter 5

Personalized Regression: A Non-parametric Parameter Generator

In the previous chapter, we explored the use of histopathology images to contextualize transcriptomic models of cancer. This approach succeeded and indicated that histopathology images contain significant information about the genomic patterns in tumors. What can we do when the sample representations do not provide a complete picture of the sample-specific model parameters? Can we estimate sample-specific model parameters if there does not exist a one-to-one function mapping covariates to model paramters? In this chapter we propose a method for this case in which sample covariates only provide insight on sample similarity rather than directly leading to model parameters.

Portions of this chapter have been previously published as [108, 109].

Personalized Regression

Here, we propose a framework to estimate sample-specific models by learning patterns of differentiation between samples. Instead of learning a single model for an entire cohort, or learning a generating function for model parameters, our framework learns a distance metric between samples and uses this distance to encourage similarity between the model parameters of similar samples. The key is to leverage the fact that although each sample is expected to have a unique pattern of differentiation, these patterns are not independent of one another, and are expected to share substantial similarities. Leveraging this, we can "borrow strength" from the entire cohort to learn a useful model that is specific to a given sample. To do this, we propose a novel *distance matching* regularizer which estimates sample similarity and encourages parameter similarity for similar samples.

Instead of assuming a known parametric function that translates covariates into regression parameters, we aim to recover the personalized regression parameters by matching the pairwise distances implied by covariates U to the pairwise distances in regression parameters β. By focusing on matching the two measurements of sample *distance* rather than learning a function of the covariate *values*, our framework avoids assumptions of smoothness of the personalization effects. In addition, our framework requires only a set of K feature-wise distance metrics $d_{U_k} : \Omega_k^2 \to \mathbb{R}_{\geq 0}$ that each act on a single covariate feature. Note that here we do not require these distance metrics to be differentiable. This allows for a wide variety of distance metrics, such as the discrete metric $d_{U_k}(U_k^{(i)}, U_k^{(j)}) = \mathbb{I}_{\{U_k^{(i)} \neq U_k^{(j)}\}}$, and allows our framework to handle the realistic situation of categorical covariates without ordering, which previous approaches do not handle. By applying these feature-specific distance metrics, we produce K-dimensional distances between each pair i, j of covariate values $d_Z(U^{(i)}, U^{(j)}) = [d_{U_1}(U_1^{(i)}, U_1^{(j)}), ..., d_{U_K}(U_K^{(i)}, U_K^{(j)})]$.

Conveniently, the same analytic tools required to understand an individual sample in the context of similar samples will also enable us to look at each sample group in the context of other groups. In this way, our approach can lead to models of controllable granularity.

For clarity, we describe the main idea using a linear model for each personalized model; extension to arbitrary generalized linear models including logistic regression is straightforward. In Section 5, we include experiments using both linear and logistic regression. A traditional linear model would dictate $Y^{(i)} = \langle X^{(i)}, \theta \rangle + w^{(i)}$, where the $w^{(i)}$ are noise and the parameter $\theta \in \mathbb{R}^p$ is shared across different samples. We relax this model by allowing θ to vary with each sample, i.e.

$$Y^{(i)} = \langle X^{(i)}, \theta^{(i)} \rangle + w^{(i)}. \tag{5.1}$$

Clearly, without additional constraints, this model is overparametrized— there is a $(p-1)$-dimensional subspace of solutions to the equation $Y^{(i)} = \langle X^{(i)}, \theta^{(i)} \rangle$ in $\theta^{(i)}$ for each i. Thus, the key is to choose a solution $\theta^{(i)}$ that simultaneously leads to good generalization and accurate inferences about the ith sample. We propose two strategies for this: (a) a novel regularization scheme and (b) a low-rank latent representation of the parameters $\theta^{(i)}$.

Model

We are interested in learning which features $X \in \mathbb{R}^P$ are relevant for predicting a phenotype $Y \in \mathbb{R}$ such as disease status. At the same time, we assume we have access to clinical covariates $U \in \Omega_1 \times \cdots \times \Omega_K$ for each individual, which are allowed to be arbitrary—unordered or ordered, categorical or continuous, and even with missing values. Throughout, we let N denote the total number of patients in the cohort and use superscripts to identify samples. Thus, $Y^{(i)}$, $X^{(i)}$, and $U^{(i)}$ denote the data for the ith sample and $\beta^{(i)}$ denotes the personalized regression coefficients for the ith sample.

5.1 Distance-Matching Regularization

To recover personalized model parameters $\beta^{(i)}$ without *a priori* knowledge of how samples are related, we assume that there are *unknown* (pseudo)metrics d_β and d_U such that $d_\beta(\beta^{(i)}, \beta^{(j)}) \approx d_U(U^{(i)}, U^{(j)})$. That is, similarity in parameters is related to similarity in covariates, however, the nature of this similarity is unobserved, unknown, and may not correspond to usual notions of distance such as Euclidean distance. This is closely related to the notion of distance metric learning introduced by Xing et al. [195]. Existing work along these lines in the personalized estimation literature typically assumes that either (a) The metrics are Euclidean, or (b) The pairwise similarities are known [196, 197].

To learn these latent distance metrics, we model them as follows:

$$d_\beta(x,y) = \zeta\langle\phi_\beta, \Big[d_{\beta_1}(x_1,y_1),\dots,d_{\beta_P}(x_P,y_P)\Big]\rangle, \tag{5.2a}$$

$$d_U(x,y) = \langle\phi_U, \Big[d_{U_1}(x_1,y_1),\dots,d_{U_K}(x_K,y_K)\Big]\rangle, \tag{5.2b}$$

where $\langle\cdot,\cdot\rangle$ denotes the dot product of two vectors and d_{β_p} ($p = 1,\dots,P$) are user-specified metrics between scalars and d_{U_k} ($k = 1,\dots,K$) are user-specified metrics between covariates. Note that here we do not require these distance metrics to be differentiable. This allows for a wide variety of distance metrics, such as the discrete metric $d_{U_k}(x,y)$ that equals one if $x = y$ and is zero otherwise. This allows our framework to handle the realistic situation of categorical covariates without ordering. The parameters ϕ_β and ϕ_U represent unknown linear transformations of these "simple" distances into more useful latent distance metrics given by (5.2a) and (5.2b) with scale $\zeta > 0$.

Define pairwise distance vectors for each i, j by

$$\Delta_\beta^{(i,j)} = \Big[d_{\beta_1}(\beta_1^{(i)},\beta_1^{(j)}),\dots,d_{\beta_P}(\beta_P^{(i)},\beta_P^{(j)})\Big] \tag{5.3a}$$

$$\Delta_U^{(i,j)} = \Big[d_{U_1}(U_1^{(i)},U_1^{(j)}),\dots,d_{U_K}(U_K^{(i)},U_K^{(j)})\Big] \tag{5.3b}$$

Since the covariate values in U are fixed, $\Delta_U^{(i,j)}$ is also fixed, whereas $\Delta_\beta^{(i,j)}$ is not fixed since the values of $\beta^{(i)}$ and $\beta^{(j)}$ will change during training. For simplicity, we take $d_{\beta_p}(x,y) = |x-y|$ ($p = 1,\dots,P$) in the remainder, although this could be replaced with any distance metric that is differentiable for all $x \neq y$.

Now define the following *distance matching regularizer*:

$$\begin{aligned} \varrho_\gamma^{(i)}(d_\beta,d_U) &= \frac{\gamma}{2}\sum_{j\neq i}\Big(d_\beta(\beta^{(i)},\beta^{(j)}) - d_U(U^{(i)},U^{(j)})\Big)^2 \\ &= \frac{\gamma}{2}\sum_{j\neq i}\Big(\zeta\langle\phi_\beta,\Delta_\beta^{(i,j)}\rangle - \langle\phi_U,\Delta_U^{(i,j)}\rangle\Big)^2. \end{aligned} \tag{5.4}$$

This regularizer attempts to match the pairwise distances between covariate values to the pairwise distances in the learned regression parameters. Let f be a loss function, e.g. least squares for regression or logistic loss for classification. Define a sample-specific objective by

$$\mathcal{L}^{(i)}(\beta^{(i)}; d_\beta, d_U) \propto f(X^{(i)},Y^{(i)},\beta^{(i)}) + \rho_\lambda^\beta(\beta^{(i)}) + \varrho_\gamma^{(i)}(d_\beta,d_U).$$

where γ trades off sensitivity to prediction of the response variable against sensitivity to sample distances, $f(X^{(i)},Y^{(i)},\beta^{(i)})$ is the prediction loss for sample i, and $\rho_\lambda^\beta : \mathbb{R}^P \to \mathbb{R}_{\geq 0}$ regularizes $\beta^{(i)}$ with strength set by λ. Summing these, we obtain the complete objective function

$$\mathcal{L}(\boldsymbol{\beta},\phi_\beta,\phi_U,\zeta) \propto \sum_{i=1}^N \mathcal{L}^{(i)}(\beta,d_\beta,d_U) + \psi_\alpha^\beta(d_\beta) + \psi_\upsilon^U(d_U)$$

where ψ_α^β, and ψ_υ^U regularize the distance functions d_β, d_U with strengths set by α, υ, respectively.

Intuition

A visualization of this model is to imagine a set of marbles positioned at the top of a hill, with springs connecting each pair of marbles. When the spring constants are set to $+\infty$ and the springs have a resting length of 0, the marbles cannot move (as in a classical population estimator which constrains all estimates to be at the same point). When the spring constant is reduced and the resting lengths of the springs are increased (to create our personalized estimator), the marbles will be allowed to roll down the hill and will each follow their own path toward a minimum. In this visualization, the spring constant κ_{ij} between marbles i and j is given by $\kappa_{ij} = \frac{\gamma}{2}$ and the spring has a resting length of $\Delta_U^{(i,j)}\phi_U$. While the resting length of the spring may change over the course of the optimization (as ϕ_U is updated), the spring imparts the same force in opposing directions on marbles i and j, so it cannot move the midpoint of marbles i and j. Thus, the pairwise regularization cannot move the bericenter of the personalized solutions.

5.2 Low-Rank Personalized Regression

We constrain the matrix of personalized parameters $\Omega = [\theta^{(1)}|\cdots|\theta^{(n)}] \in \mathbb{R}^{p\times n}$ to be low-rank, i.e. $\theta^{(i)} = Q^T Z^{(i)}$ for some loadings $Z^{(i)} \in \mathbb{R}^q$ and some dictionary $Q \in \mathbb{R}^{q\times p}$. Letting $Z \in \mathbb{R}^{q\times n}$ denote the matrix of loadings, we have a low-rank representation of $\Omega = Q^T Z$. The choice of q is determined by the user's desired latent dimensionality; for $q \ll p$, using only $\Theta\big(q(n+p)\big)$ instead of the $\Theta(np)$ of a full-rank solution can greatly improve computational and statistical efficiency. In addition, the low-rank formulation enables us to use ℓ_2 distance in Z in Eq. (5.6) to restrict Euclidean distances between the $\theta^{(i)}$: After normalizing the columns of Q, we have

$$\|\theta^{(i)} - \theta^{(j)}\| \leq \sqrt{p}\|Z^{(i)} - Z^{(j)}\|. \tag{5.5}$$

This illustrates that closeness in the loadings $Z^{(i)}$ implies closeness in parameters $\theta^{(i)}$. This fact will be exploited to regularize $\theta^{(i)}$ (Section 5).

This use of a dictionary Q is common in multi-task learning [129] based on the assumption that tasks inherently use shared atomic representations. Here, we make the analogous assumption that samples arise from combinations of shared processes, so sample-specific models based on a shared dictionary efficiently characterize sample heterogeneity. Sparsity in θ can be realized by sparsity in Z, Q; for instance, effect sizes which are consistently zero across all samples can be created by zero vectors in the columns of Q. The low-rank formulation also implicitly constrains the number of personalized sparsity patterns; this can be adjusted by changing the latent dimensionality q.

Distance-matching of Low-Rank Personalized Models

By Eq. (5.5), in order for $\|\theta^{(i)} - \theta^{(j)}\| \approx \rho_\phi(U^{(i)}, U^{(j)})$, it suffices to require $\|Z^{(i)} - Z^{(j)}\| \approx \rho_\phi(U^{(i)}, U^{(j)})$. With this in mind, define the following *distance-matching regularizer*:

$$D_\gamma^{(i)}(Z,\phi) = \frac{\gamma}{2}\sum_{j\in B_r(i)}\Big(\rho_\phi(U^{(i)},U^{(j)}) - \|Z^{(i)} - Z^{(j)}\|^2\Big)^2, \tag{5.6}$$

where $B_r(i) = \{j : \|Z^{(i)} - Z^{(j)}\|^2 < r\}$. This regularizer promotes imitating the structure of covariate values in the regression parameters. By using Z instead of Ω in the regularizer, calculation of distances is much more efficient when $q \ll p$.

Missing values

When there is a missing value in the covariate data, we set the distance between this value and all others to zero. This underestimates the distance between samples, biasing the solution toward retaining a central population estimator rather than personalizing the models based on missing features.

Full Model Let $\ell(x, y, \theta)$ be a loss function, e.g. least-squares or logistic loss. For each sample i of the training data, define a regularized, sample-specific loss by

$$\mathcal{L}^{(i)}(Z, Q, \phi) = \ell(X^{(i)}, Y^{(i)}, Q^T Z^{(i)}) + \psi_\lambda(Q^T Z^{(i)}) + D_\gamma^{(i)}(Z, \phi), \tag{5.7}$$

where ψ_λ is a regularizer such as the ℓ_1 penalty and $D_\gamma^{(i)}$ is the distance-matching regularizer defined in Eq. (5.6). We learn Ω and ϕ by minimizing the following composite objective:

$$\mathcal{L}(Z, Q, \phi) = \sum_{i=1}^{n} \mathcal{L}^{(i)}(Z, Q, \phi) + \upsilon\|\phi - 1\|_2^2, \tag{5.8}$$

where the second term regularizes the distance function ρ_ϕ with strength set by υ, and we recall that $\Omega = Q^T Z$. The hyperparameter γ trades off sensitivity to prediction of the response variable against sensitivity to covariate structure.

Optimization We minimize the composite objective $\mathcal{L}(Z, Q, \phi)$ with subgradient descent combined with a specific initialization and learning rate schedule. An outline of the algorithm can be found in Alg. 1 below. In detail, we initialize Ω by setting $\theta^{(i)} \sim N(\widehat{\theta}^{\text{pop}}, \epsilon I)$ for a population model $\widehat{\theta}^{\text{pop}}$ such as the Lasso or elastic net and then initialize Z and Q by factorizing Ω with PCA. ϵ is a very small value used only to enable factorization by the PCA algorithm. Each personalized estimator is endowed with a personalized learning rate $\alpha_t^{(i)} = \alpha_t / \|\widehat{\theta}_t^{(i)} - \widehat{\theta}^{(\text{pop})}\|_\infty$, which scales the global learning rate α_t according to how far the estimator has traveled. In addition to working well in practice, this scheme guarantees that the center of mass of the personalized regression coefficients does not deviate too far from the intialization $\widehat{\theta}^{\text{pop}}$, even though the coefficients $\widehat{\theta}^{(i)}$ remain unconstrained. This property is discussed in more detail in Section 5.

Prediction Given a test point (X, U), we form a sample-specific model by averaging the model parameters of the k_n nearest training points, according to the learned distance metric ρ_ϕ:

$$\theta = \frac{1}{k_n} \sum_{j=1}^{k_n} \theta^{(\eta(\rho_\phi, U)[j])}, \qquad \eta(\rho_\phi, U) = \underset{1 \le i \le n}{\operatorname{argsort}}\, \rho_\phi(U, U^{(i)}), \tag{5.9a}$$

where $\operatorname{argsort}$ orders the indices $\{1, \ldots, n\}$ in descending order of covariate distance. Increasing k_n drives the test models toward the population model to control overfitting. In our experiments, we use $k_n = 3$.

Algorithm 1 Personalized Estimation

```
**Require:** $\widehat{\theta}^{pop}$, $\lambda$, $\gamma, \upsilon, \alpha, c$
$\theta^{(1)}, \dots, \theta^{(n)} \leftarrow \widehat{\theta}^{pop}$
$\Omega \leftarrow [\theta^{(1)} | \dots | \theta^{(n)}]$
$Z, Q \leftarrow \text{PCA}(\Omega)$

$\alpha \leftarrow \alpha_0$
**do**
    $\widetilde{Z}, \widetilde{Q}, \widetilde{\phi} \leftarrow Z, Q, \phi$
    $\phi \leftarrow \phi - \alpha \frac{\partial}{\partial \phi} \mathcal{L}(\widetilde{Z}, \widetilde{Q}, \widetilde{\phi}; \lambda, \gamma, \upsilon)$
    $Z^{(i)} \leftarrow Z^{(i)} - \frac{\alpha}{\|\theta^{(i)} - \widehat{\theta}^{\text{pop}}\|_\infty} \Big[ \frac{\partial}{\partial Z^{(i)}} \sum_{i=1}^n D_\gamma^{(i)}(\widetilde{Z}, \widetilde{\phi}) +$
        $\widetilde{Q} \big( \partial \ell(X^{(i)}, Y^{(i)}, \theta^{(i)}) + \partial \psi_\lambda(\theta^{(i)}) \big) \Big] \quad \forall\ i \in [1, \dots, n]$
    $Q \leftarrow Q - \alpha \Big[ \frac{\partial}{\partial Q} \sum_{i=1}^n D_\gamma^{(i)}(\widetilde{Z}, \widetilde{\phi}) + \sum_{i=1}^n \widetilde{Z}^{(i)} \big( \partial \ell(X^{(i)}, Y^{(i)}, \theta^{(i)})^T + \partial \psi_\lambda(\theta^{(i)})^T \big) \Big]$
    $\alpha \leftarrow \alpha c$
    $\theta^{(i)} \leftarrow Q^T Z^{(i)} \quad \forall\ i \in [1, \dots, n]$
    $\Omega \leftarrow [\theta^{(1)} | \dots | \theta^{(n)}]$
**while** not converged
**return** $\Omega, Z, Q, \phi$
```

We have intentionally avoided using X to select θ so that interpretation of θ is not confounded by X. In some cases, however, the sample predictors can provide additional insight to sample distances (e.g. [191]); we leave it to future work to examine how to augment estimations of sample distances by including distances between predictors.

Scalability Naïvely, the distance-matching regularizer has $O(n^2)$ pairwise distances to calculate, however this calculation can be made efficient as follows. First, the terms involving $d_\ell(U_\ell^{(i)}, U_\ell^{(j)})$ remain unchanged during optimization, so that their computation can be amortized. This allows the use of feature-wise distance metrics which are computationally intensive (e.g. the output of a deep learning model for image covariates). Furthermore, these values are never optimized, so the distance metrics d_ℓ need not be differentiable. This allows for a wide variety of distance metrics, such as the discrete metric for unordered categorical covariates. Second, we streamline the calculation of nearest neighbors in two ways: 1) Storing Z in a spatial data structure and 2) Shrinking the hyperparameter r used in (5.6). With these performance improvements, we are able to fit models to datasets with over 10,000 samples and 1000s of predictors on a Macbook Pro with 16GB RAM in under an hour.

Analysis

Initializing sample-specific models around a population estimate is convenient because the sample-specific estimates do not diverge from the population estimate unless they have strong reason to do so. Here, we analyze linear regression minimized by squared loss (e.g., $f(X^{(i)}, Y^{(i)}, \theta^{(i)}) =$

$(Y^{(i)} - X^{(i)}\theta^{(i)})^2)$, though the properties extend to any predictive loss function with a Lipschitz-continuous subgradient.

Theorem 3. *Let us consider personalized linear regression with* $\psi_\lambda(x) = \lambda\|x\|_1$ *(i.e.* ℓ_1 *regularization). Let* X *be normalized such that* $\max_i\|X^{(i)}\|_\infty \leq 1$, $\|X^{(i)}\|_1 = 1$. *Define* $\overline{\theta}_t := \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_t^{(i)}$, *where* $\widehat{\theta}_t^{(i)}$ *is the current value of* $\widehat{\theta}^{(i)}$ *after* t *iterations. Let the learning rate follow a multiplicative decay such that* $\alpha_t = \alpha_0 c^t$, *where* α_0 *is the initial learning rate and* c *is a constant decay factor. Then at iteration* τ,

$$\|\overline{\theta}_\tau - \widehat{\theta}^{\text{pop}}\|_\infty \in \mathcal{O}(\lambda). \tag{5.10}$$

That is, the center of mass of the personalized regression coefficients does not deviate too far from the initialization $\widehat{\theta}^{\text{pop}}$, even though the coefficients $\widehat{\theta}^{(i)}$ remain unconstrained. In addition, the distance-matching regularizer does not move the center of mass and the update to the center of mass does not grow with the number of samples. Proofs of these claims are included in [109].

5.3 Applications

We compare personalized regression (hereafter, PR) to four baselines: 1) Population linear or logistic regression, 2) A mixture regression (MR) model, 3) Varying coefficients (VC), 4) Deep neural networks (DNN). First, we evaluate each method's ability to recover the true parameters from simulated data. Then we present three real data case studies, each progressively more challenging than the previous: 1) Stock prediction using financial data, 2) Cancer diagnosis from mass spectrometry data, and 3) Electoral prediction using historical election data. The results are summarized in Table 5.3 for easy reference.

We believe the out-of-sample prediction results provide strong evidence that any harmful overfitting of PR is outweighed by the benefit of personalized estimation. This agrees with famous results such as [174], where it is showed that optimal ensembles of linear models consist of overfitted atoms; see especially Eq. 12 and Fig. 2 therein.

Baselines For each experiment, we use several baseline models to benchmark performance:

- *Population model.* First, we use elastic net regularization [207] as a generalizable population estimator.
- *Mixture of regressions.* To estimate a small collection of models, we use a standard mixture model optimized by expectation-maximization. Since this model does not share information between mixture components, the number of components must be much smaller than the number of samples.
- *Varying coefficient model.* To estimate sample-specific models, we use an ℓ_1-regularized linear varying-coefficients model [71].
- *Deep neural network.* Finally, to compare against models with large representational capacity, we include a neural network. This neural network contains 5 hidden layers, with layer sizes and nonlinearities treated as hyperparameters optimized for cross-validation loss by grid search. The final version contains 250 hidden nodes in each layer with sigmoid nonlinearities.

For the tasks with continuous outcomes, these are linear regression models; for classification tasks, these are logistic regression models.

Choice of hyperparameters While the personalized regression approach estimates a large number of parameters, there are relatively few hyperparameters. Hyperparameters to be selected are: λ the strength of the traditional regression regularizer, γ the strength of the distance-matching regularizer, r the diameter of the neighborhoods considered by the distance-matching regularizer, υ the strength of regularizer on ϕ, and q the latent dimensionality. λ should be set equivalent to the λ used in the population estimator. γ requires some tuning and should be set such that the distance-matching regularizer contributes the a same order of magnitude on the total loss as does the predictive loss. r should be set to reflect the user's desired neighborhood of personalization; larger r produces personalized estimates which reflect covariate distances even for very different samples, smaller r improves computation speed but decreases the size of the neighborhoods of personalization. Finally, υ regularizes ϕ and should be set to reflect the user's prior knowledge about the influence of each covariate on personalization.

For our experiments, we use the following hyperparameter values with PR:

- *Simulation.* $\lambda = 1\text{e}{-1}$, $\gamma = 1\text{e}5$, $\upsilon = 1\text{e}{-2}$, $q = 2$
- *Finance.* $\lambda = 1$, $\gamma = 1\text{e}8$, $\upsilon = 1\text{e}{-2}$, $q = 50$
- *Cancer.* $\lambda = 1$, $\gamma = 1\text{e}6$, $\upsilon = 1\text{e}{-2}$, $q = 50$
- *Election.* $\lambda = 1\text{e}{-2}$, $\gamma = 1\text{e}3$, $\upsilon = 1\text{e}{-2}$, $q = 2$

For all experiments, we dynamically set r such that each point has on average 10 neighbors, and use the learning rate schedule of $\alpha_0 = 1\text{e}{-4}$, $c = 1 - 1\text{e}{-4}$. For each baseline described in Section 5, hyperparameter values were selected by cross-validation.

Simulation Study We first investigate the capacity of personalized regression to recover true effect sizes in simulation studies. For these experiments, we generate data according to $X \sim \text{Unif}(-1, 1)^p$, $U \sim \text{Unif}(0, 1)^K$, $a \sim \text{Unif}(0, 1)^p$, $b \sim \text{Unif}(0, 1)^p$, $c \sim \text{Cat}(K)^p$, $\theta_j = \mathcal{I}_{\{U_{c_j} > a_j\}} + b_j \sin U_{c_j}$, $Y^{(i)} = X^{(i)}\theta^{(i)} + N(0, 0.01)$. These experiments all use $K = 5$ covariates. As shown in Fig. 3.1, this produces regression parameters with a discontinuous distribution.

The algorithms are given both X and U as input during training, and we use LIME [160] to generate local linear approximations to the DNN in order to estimate parameters $\theta^{(i)}$ for each sample. In this setting, there exists a discontinuous function which could output exactly the sample-specific regression models from the covariates. In this sense, the neural network is "correctly specified" for this dataset, testing how well locally-linear models approximate the true parameters. To estimate personalized models for the simulated dataset, we initialize the personalized estimations with a varying-coefficient model, and personalize according to the distance metric $d_1(x, y) = |x - y|$.

Results. In Table 5.1, we report results for varying n and in Table 5.2, we report results for varying p. The values reported are: (1) the recovery error of the true regression parameters in the training set, with (mean ± std) values calculated over 20 experiments with different values of

p	Model	$\|\|\hat{\Omega} - \Omega\|\|_2$	R^2	MSE
2	Pop.	9.97	0.87	0.13
	MR	9.86	0.88	0.12
	VC	14.55	0.76	0.22
	DNN	30.42	0.75	0.24
	PR	**7.82**	**0.89**	**0.09**
10	Pop.	15.19	0.79	0.73
	MR	14.81	0.80	0.70
	VC	23.86	0.69	1.09
	DNN	67.49	0.80	0.85
	PR	**14.52**	**0.82**	**0.65**
25	Pop.	25.86	0.85	1.26
	MR	25.75	0.86	1.20
	VC	38.77	0.66	3.05
	DNN	103.72	0.68	2.78
	PR	**24.53**	**0.87**	**1.10**

Table 5.1: Simulations with $n = 500$.

n	Model	$\|\|\hat{\Omega} - \Omega\|\|_2$	R^2	MSE
100	Pop.	6.36	0.90	0.23
	MR	6.48	0.90	0.23
	VC	10.75	0.78	0.50
	DNN	22.30	0.39	0.75
	PR	**6.03**	**0.91**	**0.21**
500	Pop.	11.83	0.84	0.29
	MR	11.78	0.84	0.30
	VC	19.06	0.74	0.49
	DNN	47.33	0.81	0.37
	PR	**10.30**	**0.86**	**0.26**
2500	Pop.	33.03	0.87	0.26
	MR	31.75	0.88	0.26
	VC	33.71	0.87	0.27
	DNN	102.88	0.88	0.29
	PR	**26.11**	**0.90**	**0.21**

Table 5.2: Simulations with $p = 5$.

X, U, w, (2) correlation coefficient of predictions on the test set, and (3) mean squared error of predictions on the test set.

As expected, the recovery error is much lower for PR, while the DNN shows competitive predictive error. The population estimator successfully recovers the mean effect sizes, but this central model is not accurate for any individual, resulting in poor performance both in recovering Ω and in prediction. Similarly, both MR and VC perform poorly. As expected, the deep learning model excels at predictive error, however, the local linear approximations do not accurately recover the sample-specific linear models. In contrast, PR exhibits both the flexibility and the structure to learn the true regression parameters while retaining predictive performance.

Financial Prediction A common task in financial trading is to predict the price of a security at some point in the future. This is a challenging task made more difficult by nonstationarity—the interpretation of an event changes over time, and different securities may respond to the same event differently.

Data. We built a dataset of security prices over a 30-year time frame by joining stock and ETF trading histories[1] to a database of global news headlines from Bloomberg [46] and Reddit[2]. We transform news headlines into continuous representations by tf-idf weighting averaging [7] of word embeddings under the GLoVE model [146] pre-trained on Wikipedia and Gigaword corpora[3]. After dimensionality reduction, this news dataset consists of a 50-dimensional vector

[1] https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs/version/3

[2] https://www.kaggle.com/aaron7sun/stocknews

[3] https://nlp.stanford.edu/projects/glove

Model	Financial		Cancer		Election	
	R^2	MSE	AUROC	Acc	R^2	MSE
Pop.	0.01	64144	0.794	0.962	0.00	0.019
MR	0.74	16146	0.876	0.939	-0.56	0.031
VC	0.06	60694	0.430	0.863	0.00	0.019
DNN	-0.02	63028	0.901	0.955	0.00	0.019
PR	$\mathbf{0.86}$	$\mathbf{4822}$	$\mathbf{0.923}$	$\mathbf{0.975}$	$\mathbf{0.45}$	$\mathbf{0.011}$

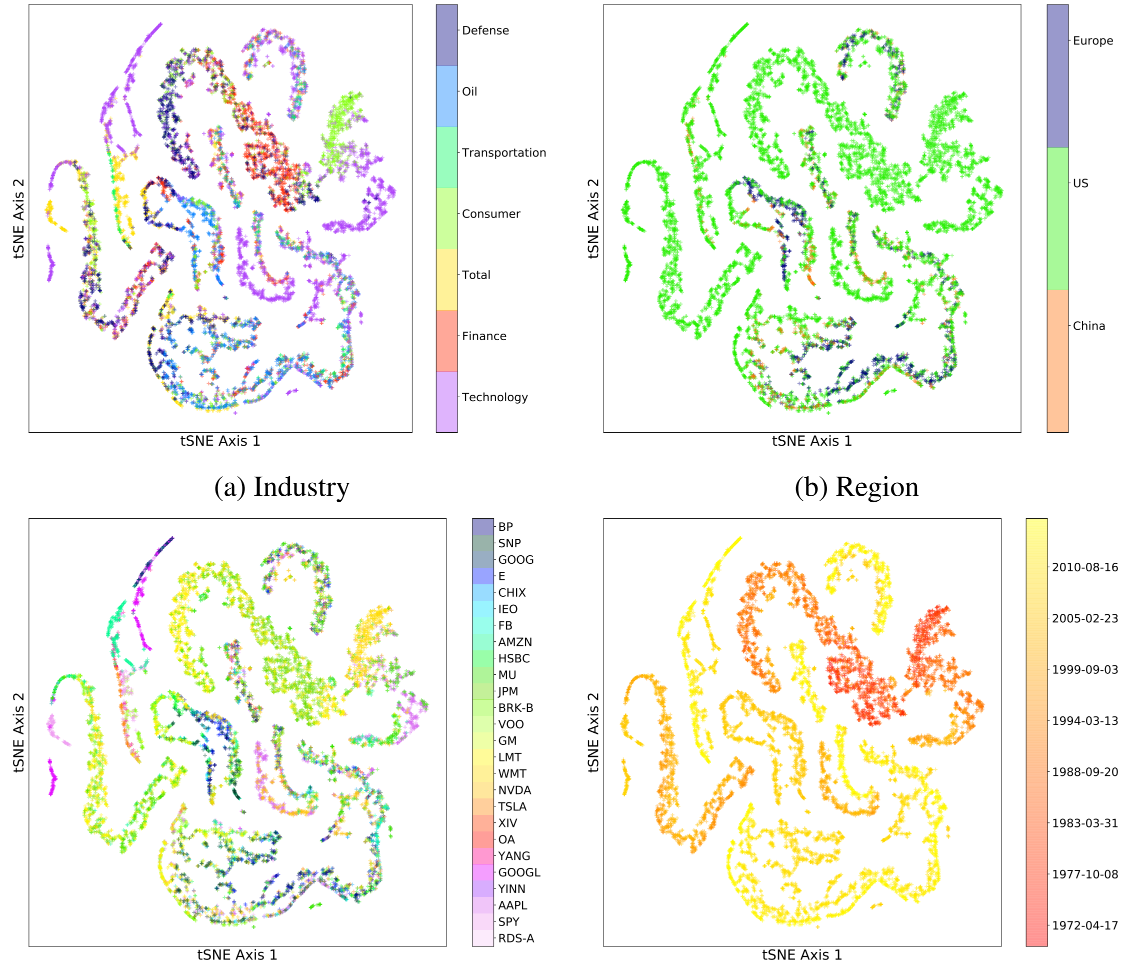
Table 5.3: Predictive performance on test sets of real data experiments. For continuous response variables, we report correlation coefficient (R^2) and mean squared error (MSE) of the predictions. For classification tasks, we report area under the receiver operating characteristic curve (AUROC) and the accuracy (ACC).

for each date. The predictors $X^{(i,t)}$ consist of the trading history of the 24 securities over the previous 2 weeks as well as global news headlines from the same time period. The covariates $U^{(i,t)}$ consist of the date and security characteristics (name, region, and industry). The target $Y^{(i,t)}$ is the price of this security 2 weeks after t. We split the dataset into training and test sets at the 80th percentile date, which is approximately the beginning of 2011. To estimate personalized models for the financial dataset, we initialize the personalized estimators with the population model and personalize according to the ℓ_1 distance for time and the discrete metric for the other covariates.

Results. PR significantly outperforms baseline methods to predict price movements (Table 5.3). In contrast to standard models which average effects over long time periods and/or securities, PR summarizes gradual shifts in attention. Shown in Fig. 5.1 are visualizations of the model parameters, colored by each of the covariates used for personalization. The strongest clustering behavior is due to time (Fig. 5.1d). For instance, models fit to samples in the era of U.S. "stagflation" (1973-1975) are overlaid on models for samples in the early 1990s U.S. recession. In both of these cases, real equity prices declined against the background of high inflation rates. In contrast, the recessions marked by structural problems such as the Great Financial Crisis of 2008 are separated from the others. Within each time period, we also see that industries (Fig. 5.1a), regions (Fig. 5.1b), and securities (Fig. 5.1c) are strongly clustered.

Cancer Analysis In cancer analysis, the challenges of sample heterogeneity are paramount and well-known. Increasing biomedical evidence suggests that patients do not fall into discrete clusters [18, 123], but rather each patient experiences a unique disease that should be approached from an individualized perspective [48]. Here, we investigate the capacity of PR to distinguish malignant from benign skin lesions using a dataset of desorption electrospray ionization mass spectrometry imaging (DESI-MSI) of a common skin cancer, basal cell carcinoma (BCC).

Data. This dataset, from [126], contains 17,053 total samples from 17 patients. Each sample consists of 2,734 spectra intensities and is labeled with a binary outcome (0=benign, 1=malignant). Data from 9 patients are used to fit models, while data from 8 patients are held-out for



(a) Industry

(b) Region

(c) Security

(d) Time

Figure 5.1: Personalized financial models using t-SNE [185] embedding. Each point represents a regression model for one security at a single date, colored according to covariate value.

evaluation. In this dataset, the only explicit covariate is the patient label. To produce covariates which are most useful for personalization, we augment the patient labels with 1500 of the predictive features compressed to 2 dimensions by t-SNE dimensionality reduction. These 1500 predictive features are excluded from the set of predictors for PR, while baseline methods use the entire set of features as predictors. We fit the personalized regression according to the distance function $d_1(x, y) = \mathcal{I}_{\{x \neq y\}}$, $d_2(x, y) = |x - y|$, $d_3(x, y) = |x - y|$, where the first function checks if the patients are the same and the final two calculate distance in the continuous covariates.

Results. As shown in Table 5.3, PR produces the best predictions of tumor status amongst the methods evaluated. The substantial improvement over competing methods is likely due to the long tail of the distribution of characteristic features. This may point to the "mosaic" view of tumors, under which single tumors are comprised of multiple cell lines [113]. This example underscores the benefits of treating sample heterogeneity as fundamental by designing algorithms to estimate sample-specific models.

Presidential Election Analysis Our last experiment illustrates a practical use case for the example of modeling election outcomes discussed in Section 1. The goals are twofold: 1) To predict county-level election results, and 2) To explore the use of distinct regression models as embeddings of samples in order to better understand voting preferences at the county (i.e. sample-specific) level.

Data. The election predictors are taken from the 2012 U.S. presidential election and consists of discrete representations of each candidate based on candidate positions compiled by ProCon.[4] Outcomes are the county-level vote proportions in the 2012 U.S. presidential election.[5] For the covariates U, we used county demographic information from the 2010 U.S. Census.[6] As the outcome varies across samples but the predictors remain constant, the personalized regression models must encode sample heterogeneity by estimating different regression parameters for different samples, thus creating county representations ("embeddings") which combine both voting and demographic data.

Results. The out-of-sample predictive error is significantly reduced by personalization (Table 5.3). Representations of the personalized models for Pennsylvania counties are shown in Fig. 5.2. Generating county embeddings based solely on demographics produce embeddings which do not strongly correspond to voting patterns (Fig. 5.2a), while voting outcomes are near a one-dimensional manifold (Fig. 5.2b). In contrast, the personalized models produce a structure which interpolates between the two types of data (Fig. 5.2e). These trends are not captured by the baseline methods, such as the varying-coefficients model (Fig. 5.2d). In addition, concatenating the demographic and voting outcomes does not recover the same structure (Fig. 5.2c).

An interesting case is that of the Lackawanna and Allegheny counties. While these counties had similar voting results in the 2012 election, their embeddings are far apart due to the difference in demographics between their major metropolitan areas. This indicates that the county populations may be voting for different reasons despite similar demographics, a finding that is not discovered by jointly inspecting the demographic and voting data (Fig. 5.2c). Thus, sample-specific models can be used to understand the complexities of election results.

Sample-Specific Pan-Cancer Analysis

A fundamental goal of personalized medicine is to understand the patterns of differentiation between individuals. With the advent of projects like The Cancer Genome Atlas[7] (TCGA) and the International Cancer Genome Consortium (ICGC)[8], genomic cancer data are generated at an unprecedented volume. We would like to use these data to understand patient-specific differences

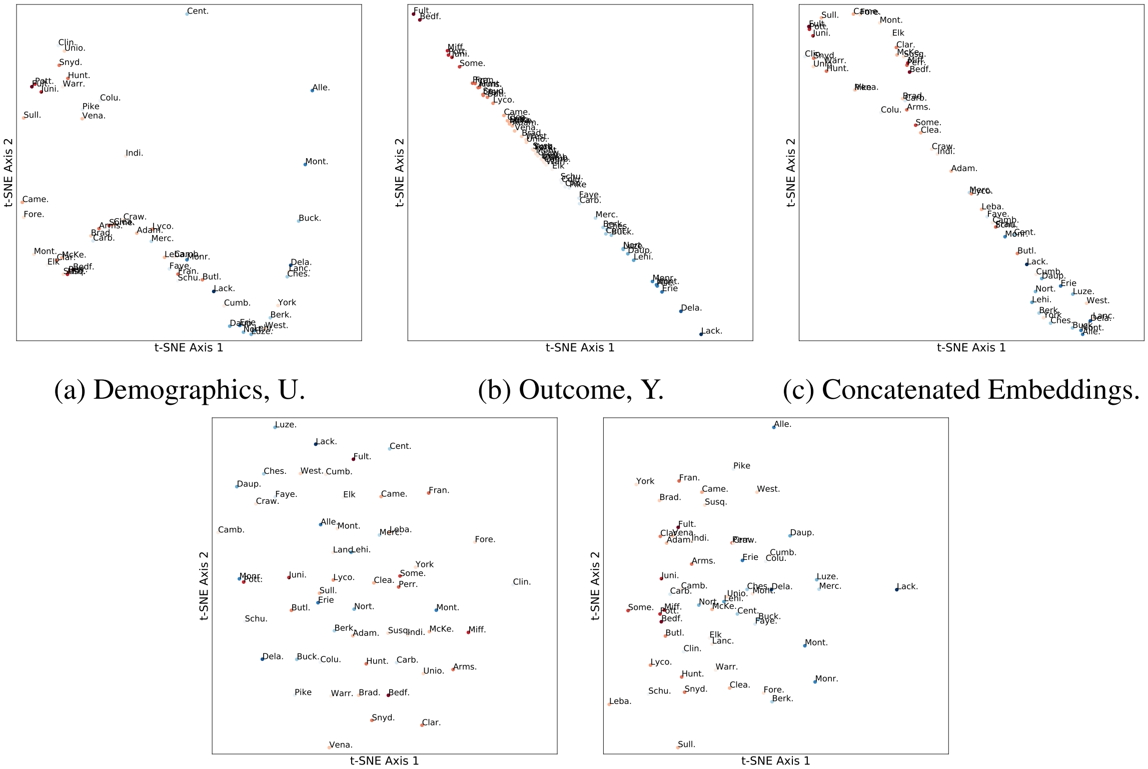
[4] https://2012election.procon.org/view.source-summary-chart.php

[5] https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/21919

[6] https://www.census.gov/data/datasets/2016/demo/popest/counties-detail.html

[7] cancergenome.nih.giv

[8] dcc.icgc.org



(a) Demographics, U.

(b) Outcome, Y.

(c) Concatenated Embeddings.

(d) VC Embeddings.

(e) Personalized Estimation, $\widehat{Z}$

Figure 5.2: Embeddings of Pennsylvania counties. Each point represents the t-SNE embedding of a representation of a county, with color gradient corresponding to the 2012 election result (red for Republican candidate, blue for Democratic candidate). (a) The county demographics (U) lie near a low-dimensional manifold that does not correspond to voter outcome. (b) The observed voting results lie near a one-dimensional manifold. (c) Personalized regression produces sample embeddings ($\widehat{Z}$) that trade off between demographic and voting information.

for personalized medicine, but many analysis pipelines discard sample heterogeneity in order to boost accuracy. Sample heterogeneity is particularly important for cancer, as cancer is increasingly appreciated as a complex disease in which many distinct underlying mutations may present with similar phenotypes [53]; even within a single patient, there is increasing evidence of tumor mosaics composed of distinct cell lines [127]. This difficulty with complex diseases like cancer motivates us to find new ways of analyzing data at increasingly small granularities.

Toward this aim, the bioinformatics community has developed increasingly specific assays [100]. From targeted microarrays to whole-genome RNA-Seq and single cell RNA-Seq, the granularity of data collected by genomic assays has continued to be refined, to the point that we now possess data points representing the state of an individual cell at a single time point, unlocking the potential to study inter-patient, inter-tissue, and inter-cell variability of complex diseases.

A classic approach to personalization is to assume that we have access to a large volume of multimodal data (e.g. clinical, genomic, proteomic, biometric, etc.) on each individual, which

Tissue	n	Tissue	n
Breast	1,092	Ovary	376
Lung	1,016	Liver	371
Kidney	885	Cervix	304
Brain	677	Soft Tissue	259
Colorectal	623	Adrenal Gland	258
Uterus	611	Pancreas	177
Thyroid	502	Esophagus	164
Head and Neck	501	Bone Marrow	151
Prostate	495	Eye	80
Skin	468	Lymph Nodes	48
Bladder	408	Bile Duct	36
Stomach	380		

Table 5.4: Number of samples by tissue in TCGA.

is used to build large predictive models. Given enough data per individual, clinical outcomes and decisions can be personalized [100, 150], and recent work along these lines has leveraged a dizzying array of complex models including Gaussian processes [6], neural networks [119], and tree-based models [135], just to name a few. Despite the successes of these methods, they are still limited to this 'one disease–one model' perspective, in which a single predictive model—often through model averaging—is built for a single cohort (e.g. corresponding to patients of a particular disease type). Furthermore, these complex models are often difficult to interpret and are not guaranteed to provide correct inference into the underlying biological drivers of disease.

Unfortunately, in many circumstances, we may only have access to a limited amount of measurements per individual (e.g. either for cost or privacy reasons). In this case, it is advantageous to leverage data from distinct but related cohorts in order to build personalized models for each individual. For example, in cancer applications we now have access to large datasets for commonly studied cancers such as breast and lung cancer through repositories such as TCGA. At the same time, less common cancers such as of the eye and lymph node, have much less data (Table 5.4). A true "pan-cancer" study would combine all of this data, exploiting the similarities between different types of cancer to improve the accuracy of models for eye and lymph node cancer. That such similarities exist is well-established in the literature [e.g. 194]. However, in the traditional 'one disease–one model' paradigm, data from other cancers play no role; while this makes sense for diseases which have a single root cause, the heterogeneity of complex diseases such as cancer renders these methods inadequate. Leveraging data from multiple cohorts while simultaneously obtaining distinct models for different diseases and different patients is a key challenge in personalized medicine.

Here, we investigate the potential of using personalized regression for personalized cancer analysis. We use gene expression (RNA-Seq) quantification data from The Cancer Genome Atlas (TCGA). This dataset compiles data from 37 projects spanning 36 disease types in 28 primary sites. After pruning for missing values, this dataset contains 9663 profiles for 8944 case and 719 matched control samples; we divide this set into 75% training data and 25% testing. While this full dataset is sizable, previous analyses have been hampered by the small number of samples for each particular cancer sub-type (e.g., there are only 36 cases present in the bile duct cancer dataset). Because our framework of personalized regression allows models to share information across diverse settings, we are able to jointly analyze the cancer subtypes while still recovering subtype-specific characteristics. The number of samples available from each dataset was shown in Table 5.4.

We subsample genes based on annotations in the COSMIC Catalogue of Somatic Mutations in Cancer [54], so that there is exactly one putatively causal gene for each 5 non-annotated genes. This resulting in $P = 4123$ features when an intercept term is added. We train each logistic regression model to predict the case/control status of each sample with ℓ_1 regularization to perform variable selection in order to study which genes are relevant for classification. Our baseline models include: ℓ_1-regularized logistic regression model trained on all pan-cancer data ("Population"), ℓ_1-regularized logistic regression model trained on each primary tissue type ("Tissue-Population"), ℓ_1-regularized mixture model with the number of clusters equal to the number of tissue types in the pan-cancer dataset ("Mixture"), a logistic regression model with parameters that follow a linear varying coefficients model ("VC"), and the mixed model recently proposed by Hayeck et al. 72.

In addition to the RNA-seq data, we used the following 14 covariates: disease type, primary tumor site, age of the patient at diagnosis, year of birth of the patient, the number of days to sample collection, gender of the patient, race of the patient, percent of neutrophil infiltration, percent monocyte infiltration, percent normal cells, percent tumor nuclei, percent lymphocyte infiltration, percent stromal cells, and percent tumor cells in the sample. These covariates span a range of different types, including both continuous and discrete values; for continuous-valued covariates, we use the ℓ_1 distance function, for discrete-valued covariates, we use the discrete distance metric. For the VC model, unordered discrete covariates such as primary tissue must be converted into one-hot vectors. This procedure increases the number of covariate features to 64, underscoring the benefit of our model's ability to directly use the 14 unordered, discrete covariates without modification.

To predict case/control status of each sample, we implemented the personalized logistic regression model with Lasso regularization. We selected λ in the population estimator by 10-fold cross-validation on the training set. This value of λ is held fixed between the population estimator and the personalized estimator. Next, we set γ so that the loss due to the distance matching regularizer is similar in magnitude to the prediction loss. Finally, we set υ and α so that the loss due to distance metric regularization is one order of magnitude smaller than the logistic classification loss. This heuristic represents our uncertainty in the form of personalization for cancer; we prefer to rely on the data than to set a rigid form of personalization. Empricially, we observe robustness in the solutions up to an order of magnitude change in these hyperparasecdepthmeters. By inspecting the variables (mRNA transcripts) selected by this method, we find that personalized regression identifies (1) individualized genetic aberrations, (2) interpretable patterns of differentiation, and (3) patient sub-typing that is more meaningful than clustering based on covariate data.

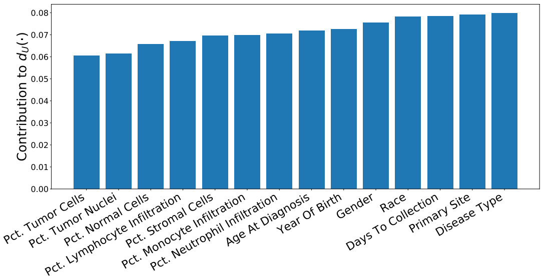


Figure 5.3: Contribution of each covariate to the learned personalization distance in the pan-cancer dataset. We see that, as expected, this method learns to upweight differences in disease type and primary site, along with other demographic features.

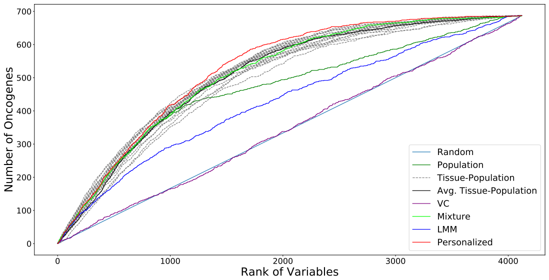


Figure 5.4: Overlap of selected variables with annotated oncogenes (best viewed in color). Results for each tissue-specific model are displayed in dashed gray lines, with the sample-weighted mean displayed in a solid black line. We see that the personalized models select oncogenes at higher ranks than do the baseline methods, especially for the long tail of low rank oncogenes.

Personalization Effects

We also examine the learned distance metrics for contributions to personalization by each covariate. The linear form of the distance msecdepthetric makes interpretation of ϕ_U straightforward by inspection of the loadings (Figure 5.3). As expected, the disease and primary tissue site of the sample have the heaviest influence on personalization, confirming our intuition that the variation between cell types is highest in cells of distinct differentiations. Next in importance to ϕ_U are demographic and clinical features, which may be interpreted as a coarse-grained view of the patient's SNPs. Important molecular markers of cancer subtype appear to be (a) percent of neutrophil infiltration, (b) percent monocyte infiltration, and (c) percent stromal cells, confirming clinical findings these phenotypic characteristics as indicative of molecular subtypes, especially in breast cancers [45, 85, 117].

Accurate Recovery of Personalized Parameters

Personalized regression selects variables on a sample-specific level. Such fine-grained analytic power, unobscured by cohort averaging, enables more accurate recovery of important features than is possible by population-scale models. As a result, the number of variables selected for each sample-specific model is much lower than the number of variables selected by the population estimator (Figure 5.5, top). In addition, the number of samples for which each variable is selected follow a long-tailed distribution in which a few genes are selected for many samples, but many genes are selected for a few samples (Figure 5.5, bottom). The set of common gene selections represents well-studied oncogenes that are common to many types of cancer while the infrequently selected genes may correspond to less common oncogenes.

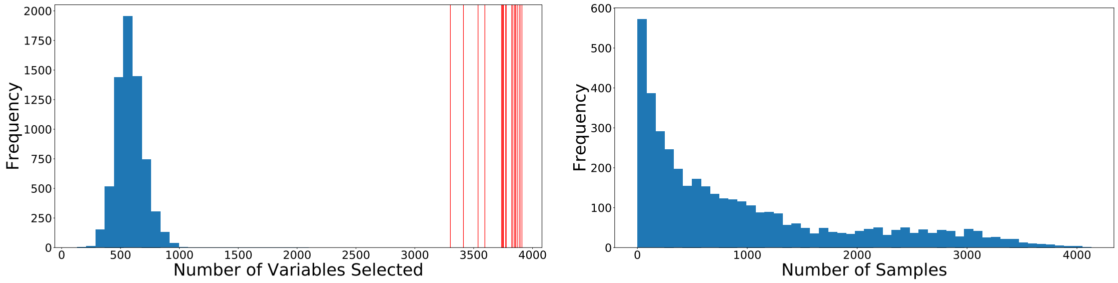


Figure 5.5: The sample-specific variable selection of personalized regression results in models with fewer selected variables than those selected by population-level models. (Top) Histogram of the number of variables selected for each patient by personalized regression. Vertical red lines indicate the number of variables selected by the Tissue-Population model trained on a single cancer type. Personalized models achieve similar or improved predictive performance with fewer selected genes. (Bottom) Histogram of the number of samples for which each gene is selected.

To investigate this possibility of many infrequently selected oncogenes, we further examine the oncogene distribution by rank of variable. Ranks are calculated by ordering the sums of the magnitudes of each coefficient along the sample axis (for population models, this is simply the magnitude of the coefficient associated with that variable). In this way, the rank captures both the number of samples for which the variable was selected and the magnitude of the implied effect size. As shown in Figure 5.4, the overlap between selected genetic markers and the annotations in COSMIC [54] is improved by the process of personalization. We see that the highly ranked oncogenes are efficiently selected by nearly all methods, but the performance of the baseline models lags as the rank diminishes. In particular, although the Tissue-Population models that are learned independently using only samples from a given tissue tend to select highly ranked genes that are also annotated in COSMIC, the performance in the long-tail of infrequently selected genes is less competitive compared to the personalized model. This confirms the intuition that personalization is the most useful in this latter regime.

To test whether this increase in oncogene selection is due to novel identification of genetic processes, we perform enrichment analysis of the ranked lists of genes. Reported in Table 5.5 are the most significant Gene Ontology (GO) terms from a ranked enrichment test using Panther 13.1 [131] on the Panther GO-SLIM Biological Process dataset [130] with a cutoff of $p < 0.05$ for the Bonferroni-corrected p-values. The genes selected by personalized models are enriched with similar GO terms compared to the baseline models, which is expected since the gene ontology is largely comprised of well-studied annotations from large cohorts as opposed to harder to detect personalized effects. This validates our hypothesis that the improved performance of variable selection is not due to identification of a single group of genes, but rather is due to the identification of many sample-specific effects.

Model	Biological Process	p-value
Population	mRNA Processing	2.06e-8
	DNA Metabolic Process	3.18e-6
	Organelle Organization	3.86e-2
Tissue-Population	mRNA Processing	3.09e-9
	Metabolic Process	3.26e-5
	Transcription, DNA-Dependent	9.61e-5
	DNA metabolic process	5.9e-3
Mixture	mRNA processing	1.45e-8
	DNA Metabolic process	1.96e-5
	transcription, DNA-dependent	2.62e-4
	organelle organization	7.32e-3
VC	None	NA
LMM	DNA metabolic process	2.02e-2
Personalized	mRNA processing	5.83e-6
	metabolic process	1.1e-3
	DNA metabolic process	3.15e-2

Table 5.5: Enrichment Analysis of Complete Variable Rankings

Discovery of Molecular Subtypes

The pattern of selection of genes is of particular interest for clinical application. As seen in Figure 5.6, there are a number of common oncogenes that are repeatedly selected throughout many cancer types, including FOXA1, HOXC13, and FCGR2B. This set combines with a sparse selection of a number of oncogenes specific to each cancer type. These cancer types span surface-level characteristics such as tissue type. Interestingly, we also see a small set of rarely selected oncogenes that are consistently selected for a cluster of about 300 patients (outlined in Figure 5.6). This set of oncogenes is highly over-represented for the GO biological process term "Modulation of Chemical Synaptic Transmission" (Bonferroni corrected p-values of 2.32e-2), which includes genes ATP1A2, SLC6A4, ASIC1, GRM3, and SLC8A3. These genes code for ion-transport processes, which have long been seen in vivo as an important system in thyroid cancer [51] and in vitro from leukemic cells [136], but only recently been appreciated as a functional marker across many different cancer types [167].

Figure 5.7 depicts a tSNE projection of the learned effect vector for each sample, colored by the primary tumor site. While the samples appear to form clusters, and the case samples are separated from the control samples by a large margin, again these clusters do not appear to correspond to any individual covariate. This complexity of personalization underscores the need for learned distant metrics to capture relationships corresponding to molecular characterization of tumors.

To identify molecular subtypes, we cluster the parameter embeddings using the HDBSCAN algorithm and perform an enrichment analysis of each cluster's variable selection. The top 3 over-enriched leaf terms from the GO biological process dataset are shown in Table 5.6. We see that the different clusters of models correspond to different biological processes. For instance, cluster 3 is enriched for several terms associated with extracellular interactions, while cluster 2 emphasizes terms associated with nucleotide modification via splicing and repair. These results suggest that the clusters discovered by personalized regression may correspond to clinically meaningful molecular subtypes.

Cluster	Biological Process	p-value
1	Symbiont Process	2.62e-3
	Regulation of Cellular Catabolic Process	1.96e-2
	Protein Modification Process	3.43e-2
2	DNA repair	3.21e-12
	RNA splicing, via Transesterification Reactions with Bulged Adenosine as Nucleophile	3.64e-7
	DNA Replication	1.00e-6
3	Symbiont Process	1.4e-3
	Antigen Processing and Presentation of Peptide Antigen	1.06e-2
	Antigen Processing and Presentation of Exogenous Antigen	1.08e-2
4	DNA Metabolic Process	3.83e-8
	DNA repair	1.68e-6
	Regulation of Cellular Macromolecule Biosynthetic Process	5.06e-6
5	Plasma Membrane Bounded Cell Projection Morphogenesis	1.45e-2
	Neuron Projection Development	3.02e-2
6	mRNA Catabolic Process	8.78e-4
	Gene Expression	6.02e-4
	Macromolecule Biosynthetic Process	3.32e-2
7	None	N/A
8	Generation of Precursor Metabolites and Energy	4.75e-5
	Oxidation-Reduction Process	4.52e-5
	Citrate Metabolic Process	9.84e-3
9	DNA Metabolic Process	3.96e-10
	Cellular Response to DNA Damage Stimulus	5.57e-9
	Protein Complex Subunit Organization	1.41e-4
10	DNA Metabolic Process	7.15e-8
	ncRNA Metabolic Process	1.33e-4
	Chromatin Organization	8.27e-4
11	Negative Regulation of Phosphorylation	3.74e-2
	Hematopoietic or Lymphoid Organ Development	4.46e-2

Table 5.6: Enrichment Analysis of Tumor Clusters

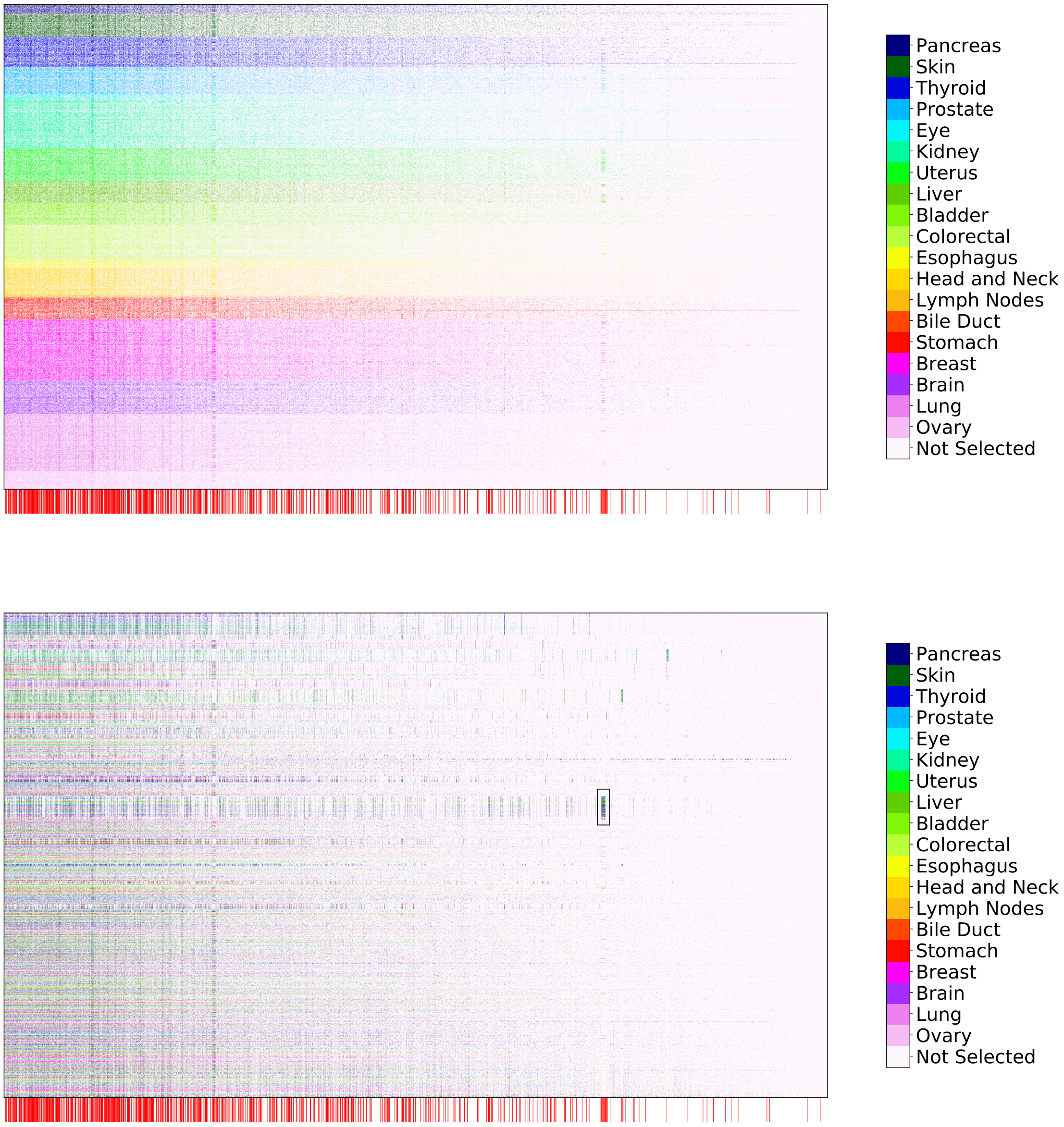


Figure 5.6: Selection of genetic markers as predictive of case/control status from a pan-cancer dataset. The horizontal axis denotes genes while the vertical axis indexes samples. Selected variables in each row are colored by the primary tumor site of the sample, with unselected variables colored white. We observe consistent selection of a number of common oncogenes throughout all cancer types along with the sparse selection of a small number of oncogenes specific to each cancer type. Genes annotated as oncogenes in the COSMIC census are marked by a red line along the horizontal axis (zoom in for more detail as these lines may be difficult to differentiate on some screens). (Top) Rows ordered by primary tissue site, (Bottom) Rows clustered according to personalized variable selection. The boxed region is analyzed in Section 5.

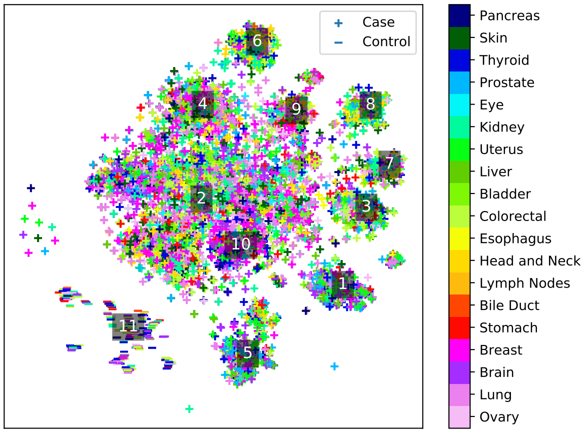


Figure 5.7: t-SNE projection of personalized regression parameters learned from a pan-cancer dataset. Each point represents a single sample with color indicating primary tumor site and marker type indicating case/control status of the patient. Labelled points indicate the centroids of clusters analyzed in Table 5.6.

Chapter 6

Sample-Specific Network Inference

6.1 Motivation

Estimating the structure of directed acyclic graphs (DAGs, i.e. Bayesian networks) is a classic problem in which we seek to summarize relationships between many observed variables through a graphical model. This approach has been frequently applied to many problems in biology [164], genetics [202], machine learning [95], and causal inference [178].

However, the relationships between variables are often non-stationary, e.g. interactions change over time. Time-varying networks [3, 144, 175] have enjoyed success in identifying changes in network connections over time by permitting coefficients to vary with time. However, when dealing with complex processes like cancer which vary according to many processes simultaneously, we often do not know *a priori* how the factors influence network development. Thus, we would like a method which can estimate Bayesian network with structures and coefficients that vary according to several, and possibly latent, continuous covariates. These algorithms will enable us to understand large heterogeneous datasets at new granularities.

Discovering New Molecular Profiles of Diseases

Complex diseases such as cancer[33, 53, 148] and Alzheimer's disease[134, 158, 182] are affected by many sources of variation which cause individuals to experience a unique patient journey. Traditional classification of diseases based on coarse-grained factors such as tissue morphology are increasingly outdated as fine-grained biological assays are revealing stunning heterogeneity at the granularity of individual cells[18, 128]. While these assays provide data at a finer resolution than previously possible, methods of analysis continue to rely on statistical methods which independently estimate cluster-level models. Increasing evidence points to molecular subtypes which do not form discrete clusters[122, 137]. Discovering sample-specific molecular profiles would refine our understanding of the mechanisms behind complex diseases. To this end, sample-specific inference aims to build statistical models that are capable of tailoring estimation for each sample.

Regulatory Network Inference in Alzheimer's Disease

Recent works have shown that neuronal cell types have strikingly distinct interactomes[133]; thus, estimating a single shared regulatory network will obscure the effects of the disease. With the wealth of single-cell transcriptomics data recently available[128], we can apply the methods of personalized network inference to estimate different regulatory networks for different cells. In this way, we can uncover the heterogeneity of regulatory networks on the cellular level, and investigate the effectiveness of pre-defined cell types to capture regulatory heterogeneity, or possibly define new cell types which have different regulatory modifications.

Sample-Specific Inferences for Multi-Modal Cancer Analysis

Sample heterogeneity is particularly important for cancer, as cancer is increasingly appreciated as a complex disease in which many distinct underlying mutations may present with similar phenotypes [53]; even within a single patient, there is increasing evidence of tumor mosaics composed of distinct cell lines [127]. Thus, we are interested to estimate regulatory networks for cell populations in individual cancer patients to examine which regulatory interactions are modulated in which patients.

6.2 Preliminaries and Related Work

DAG Learning

The basic DAG learning problem is formulated as: Let $X \in \mathbb{R}^{n \times p}$ be a data matrix consisting of n IID observations of the random vector $X = (X_1, \dots, X_p)$. Let $\mathbb{D}$ denote the space of DAGs $G = (V, E)$ of p nodes. Given X, we seek a DAG $G \in \mathbb{D}$ (i.e. Bayesian network) for the joint distribution $\mathbb{P}(\mathbb{X})$ where X is modeled via a structural equation model (SEM) defined by a weighted adjacency matrix $W \in \mathbb{R}^{p \times p}$ [95, 178].

Structural Equation Models For $W \in \mathbb{R}^{p \times p}$, we have a directed graph of p nodes with structure defined by the binarized adjacency matrix $A(W)$ where $A(W)_{ij} = 1 \iff w_{ij} \neq 0$. To translate this graph into a joint distribution on X, we use the linear SEM $X_j = w_j^T X + z_j$, where $X = (X_1, \dots, X_p)$ is a random vector and $z = (z_1, \dots, z_p)$ is a random noise vector. For the remainder, we assume that z is Gaussian with mean zero, but that is not strictly necessary. This form can be expanded to use any generalized linear model as a transition function where $\mathbb{E}(X_j | X_{pa(X_j)}) = f(w_j^T X)$ (e.g., for Boolean X_j, we can use a logistic link function).

For simplicity, we focus on this linear SEM with least-squares (LS) loss $\ell(W; X) = \frac{1}{2n} \|X - XW\|_F^2$, which has convenient properties such as consistency [19].

Network Inference via NOTEARS

Estimating DAG structure, even for a single population model, from data is NP-hard [30]. Much of this challenge is due to the acyclicity constraint: DAGs representing Bayesian network cannot contain cycles of any length (otherwise, the likelihood of a variable's value would be indirectly linked to itself). When expressed as a combinatorial optimization problem, this constraint is difficult to enforce efficiently while we simultaneous optimizing coefficients governing many relationships.

Recent work has translated this *combinatorial* optimization problem into a *continuous* program through the matrix exponential: $h(W) = tr(e^{W \cdot W}) - p = 0 \iff W \in R^{p \times p}$ is a DAG, where $\cdot$ is the Hadamard product and e^A is the matrix exponential of A. This expression of the DAG constraint is useful because $h(W)$ has a simple gradient: $\nabla h(W) = (e^{W \cdot W})^T \dot{2}W$. With this result, [204] developed the NOTEARS (Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning) algorithm for smooth optimization of DAGs:

$$\min_{W \in \mathbb{R}^{p \times p}} F(W) \quad \text{such that } h(W) = 0 \tag{6.1}$$

, which can be efficiently solved via the augmented Lagrangian to translate the equality constraint on $h(W)$ into a smooth optimization.

Sample-Specific DAG Learning

Here, we are interested in learning DAGs for which the structure varies between samples. That is, instead of estimating a single population Bayesian network representation by W, we seek sample-specific Bayesian networks each represented by $W^{(i)}$, where i is the sample index. For any sample, it is of course trivial to estimate a DAG which perfectly capitulates that sample's data. To ensure that the sample-specific DAGs vary according to underlying processes rather than overfitting individual samples, we will also observe sample representations $U = (U_1, \dots, U_m)$.

Related Work

In spirit, this problem of estimating sample-specific networks has its roots in context-specific independence (CSI) [13]. CSI permitted that the independence relations captured in Bayesian Network sparsity patterns may hold only for certain contexts. In general, context-specific independence transforms the traditional Bayesian Network DAG into a multi-graph which cannot be represented in the same data structure. Thus, a long line of work has gone into developing alternative representations and approaches for estimating CSIs [31, 55, 143, 147]. However, many of these approaches rely on the combinatorial view of Bayesian Network optimization and result in limited scalability. More recent approaches have used causality rules to perform instance-specific independence tests [86, 87, 89].

In addition, several approaches have sought to construct an estimator of sample-specific networks based on the divergence from a population network. In particular, LIONESS [98] estimates sample-specific linear networks as the weighted difference between the network estimated by all N samples and the network estimated by leaving out the ith sample. While this approach showcases the clever idea of the exchangability of these estimators and differences, it does not share statistical power between sample estimators (the statistical power is only shared for the population model). A similar approach was used to identify dysregulated pathways in a gene regulatory network by examining the neighborhood regression residual, although this approach was more robust than the above because it sought only to estimate dysregulation rather than a sample-specific network [20]. Other approaches to estimate sample-specific gene regulatory networks have typically eschewed the search for Bayesian Networks, and found other heuristics which can substitute for sample-specific regulatory networks. [43]

6.3 Contextualized NOTEARS

To estimate sample-specific Bayesian Networks, we propose *Contextualized NOTEARS*, which uses the smooth NOTEARS loss [204] to estimate sample-specific networks as the output of a function of contextual data.

Contextualized NOTEARS uses a version of Contextual Explanation Networks (CENs) [4], which takes data represented by context variables C and interpretable variables X for which we are interested in inferring the Bayesian Network W. The model represents conditional probability of the activations given contextual inputs and network weights, $\mathbb{P}(X \mid W, C)$, in the following form:

$$\mathbb{P}(X \mid C) = \int \mathbb{P}(X \mid W)\delta(W = \phi_w(C))dW = \mathbb{P}(X \mid \phi_\theta(C)), \tag{6.2}$$

where $\mathbb{P}(X|W)$ is the likelihood of X under the linear SEM for DAG W. Note that the DAG W is a function of the contextual information, i.e., $W = \phi_\theta(C)$. In other words, the CEN architecture produces a sample-specific parameterization of the Bayesian network for each sample.

Generation of parameters for sample-specific networks is accomplished via a context encoder. DAGs W are a linear combination of a small constant number (K) of "archetypes," denoted $\{W_1, \ldots, W_K\}$ or $W_{1:K}$. For a sample i with context C^i, DAG W^i is computed as:

$$W^i = \phi_\theta(C^i) = \sum_{k=1}^{K} \sigma\left(f_\theta(C^i)\right)_k W_k \tag{6.3a}$$

where $\sigma(\cdot)$ is the softmax function and $f_\theta(\cdot)$ is the context encoder with a K-dimensional output. The softmax constraint ensures that the weighting of each archetype is positive and sums to 1, permitting these values to be interpreted as the probability of cluster membership. The dictionary of DAGs, $W_{1:K}$, is estimated jointly with the context encoder, and the entire architecture is trained end-to-end via backpropagation after initialization of the archetypes to random vectors.

To ensure that each W^i is a DAG, we balance the estimation loss with the NOTEARS loss applied to the sample-specific model. Thus, Contextualized NOTEARS can be summarized as the following minimizer:

$$\underset{\theta, W_{1:k}}{\text{argmin}} \sum_{i=1}^{n} \frac{\alpha}{2}\left(X^i - X^i W^i\right)^2 + \beta \|W^i\|_1 + \gamma \text{tr}\left(e^{W^i \cdot W^i}\right) \tag{6.4}$$

where $W^i = \phi_\theta(C^i)$ and α, β, and γ, are hyperparameters that trade off predictive loss against DAG-ness and sparsity. Because the NOTEARS loss is smooth, this loss can be efficiently optimized through backpropagation with automated gradient solvers. A `Python` implementation is available at `github.com/blengerich/ContextualizedNOTEARS`.

Scalability: Low-Rank Contextualized NOTEARS

Contextualized NOTEARS estimates $\mathcal{O}(Kp^2 + |\theta|)$ parameters. As such, for large p, computationally permissive values of K may not have the representational capacity to accommodate our data. As such, we can use a straightforward adaptation to make the dictionary of W archetypes low-rank: $W_j = A_j B_j$ where $A_j \in \mathbb{R}^{p\times d}$, $B_j \in \mathbb{R}^{d\times p}$ for $j = 1, \dots, K$. This reduces the number of free parameters to $\mathcal{O}(2Kpd + |\theta|)$, allowing savings for $d < p$. Under this formulation, the N sample-specific DAGs are contained in a convex hull of K exterior points in a D-dimensional subspace of $\mathbb{R}^{p\times p}$.

Experiments

First, we test Contextualized NOTEARS and Low-Rank Contextualized NOTEARS on simulation data. We generate $N = 100$ Erdos-Renyi DAGs with $p = 16$ nodes in each. These DAGs are sparse, with a mean of only 8 edges in each. We observe 8 contextual features for each sample, which are PCA-compressed representations of the N DAGs obfuscated by a 1:1 signal-to-noise ratio.

As shown in Figure 6.1, both Contextualized NOTEARS and Low-Rank Contextualized NOTEARS significantly outperform population models and the LIONESS sample-specific model. For strong edges (with weight magnitudes above 0.2), edge recovery is quite strong, often above an F1-score of 0.7. We can also see that the performance of the Contextualized models do not decrease for increased K.

6.4 Gene Regulatory Network Inference

Next, we turn to estimate sample-specific gene regulatory networks. A common task in bioinformatics is to infer regulatory networks from gene expression data. That is, we seek a graphical model which we can use to understand which gene products regulate the production of other gene products. These networks enable systems biologists to probe the function of genes and proteins, and to understand causal mechanisms through knockout and perturbation analyses.

Historically, inference of these transcriptomic regulatory networks was performed at a population level. However, recent advances in both the experimental assays (with high-throughput single-cell sequencing) and the computational power have enabled new analyses to be performed at a more granular scale, motivating us to examine sample-specific networks estimated through Contextualized NOTEARS. Here, we briefly study two examples: personalized regulatory networks for cancer patients, and cell-specific networks for mouse brains.

Personalized Regulatory Networks for Analysis of Cancer

We first seek to estimate sample-specific regulatory network of cancer patients. Cancer is a highly individualized disease [137], for which distinct gene processes are important for different tumors, and tumors may vary continuously rather than falling into discrete clusters [184].

Thus, we would like to estimate patient-specific networks to identify dysregulated pathways and potential therapeutic targets.

Dataset

To estimate these patient-specific networks, we use bulk RNA-seq data from The Cancer Genome Atlas [1]. From the large number of RNA-seq transcripts available, we select 784 based on the standard deviation and annotation in the COSMIC gene census [54]. On the other axis, we have 9715 patients each with 5 covariates used for contextualization: Age in Days, Gender, Year of Birth, Race, and Sample Type. In addition, we hold out the following covariates for analysis of the estimated networks: Disease Type, Primary Site, Percent Stromal Cells.

Results

The estimated personalized networks are visualized in Figure 6.2, with color indicating covariate values. As expected, these networks cluster strongly with respect to the observed covariate of patient race. Interestingly, the networks capture some of the variance with respect to disease type and tissue. Finally, the networks do not cluster strongly with respect to the year of birth, indicating that networks of gene interaction in tumors may not vary significantly with patient age.

Cell-Specific Transcriptomic Regulatory Networks

Recent advances in high-throughput single-cell sequencing have enabled new analyses to be performed at a more granular scale. For example, inference of tissue-specific [125, 149, 165] and cell-type specific [25, 26, 74, 93, 133] regulatory networks have redefined biological understanding of many genetic functions.

Here, we propose to take these analyses to the next level of granularity: a different regulatory network for each cell. Successfully estimating different networks for every cell would enable the next stage of biological inquires to be performed at a new level of specificity, allowing biologists to understand the cellular diversity within and between defined cell types.

Dataset

To estimate single-cell regulatory networks, we require a covariate measured for each cell. Dual assay technology, which allows simultaneously profiling transcriptomic (RNA-seq) and epigenetic (ATAC-seq) markers have recently become available. In this example, we use the SNARE-seq dataset of 10,309 cells from adult mouse brain cells [27].

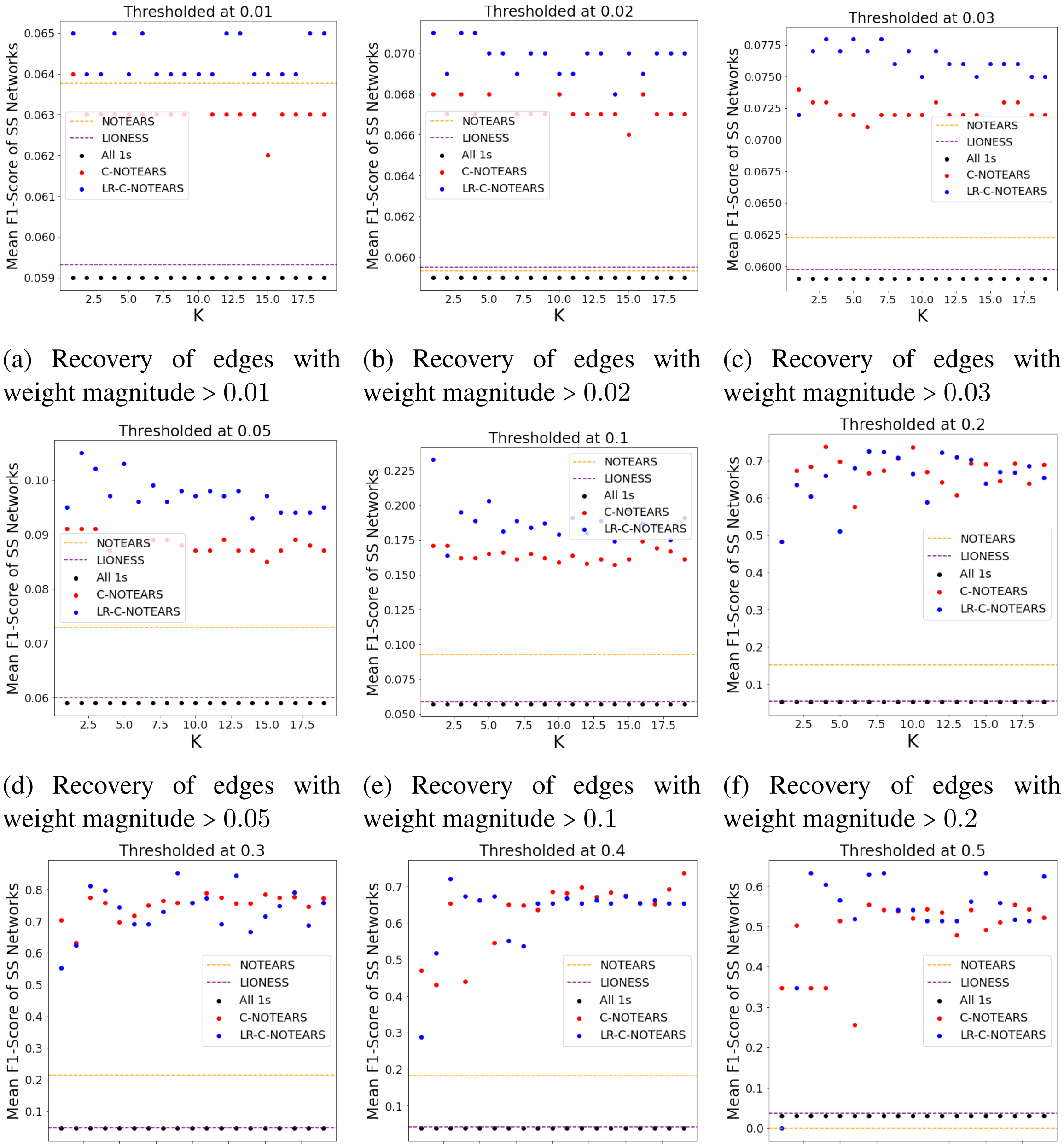
To preprocess the scRNA-seq in this dataset, we initial perform quality control with poor quality cell removal and doublet removal. Next, we remove empty droplets using `DropletUtils` and doublets using `scds` with a threshold of 1.0. Finally, we normalize reads with `linnorm`. Finally, we perform a PCA transform of the scATAC-seq data to compress to 5 covariates and

[1] `https://portal.gdc.cancer.gov/`

estimate cell-specific networks by fitting Low-Rank Contextualized NOTEARS with $K = 16$ and $d = 5$.

Results

Estimated cell-specific networks are shown in Figure 6.3. Even though the Contextualized NOTEARS model did not have access to the labelled cell types, the gene interaction networks tend to recapitulate these cell types. This indicates that the cell types, which are traditionally identified by expression of particular genes, also correspond to distinct epigenetic patterns and distinct gene–gene interaction networks. This demonstration of the Contextualized NOTEARS motivates future work to quantify these relationships between epigenetic regulators and gene–gene interaction networks.



(a) Recovery of edges with weight magnitude > 0.01

(b) Recovery of edges with weight magnitude > 0.02

(c) Recovery of edges with weight magnitude > 0.03

(d) Recovery of edges with weight magnitude > 0.05

(e) Recovery of edges with weight magnitude > 0.1

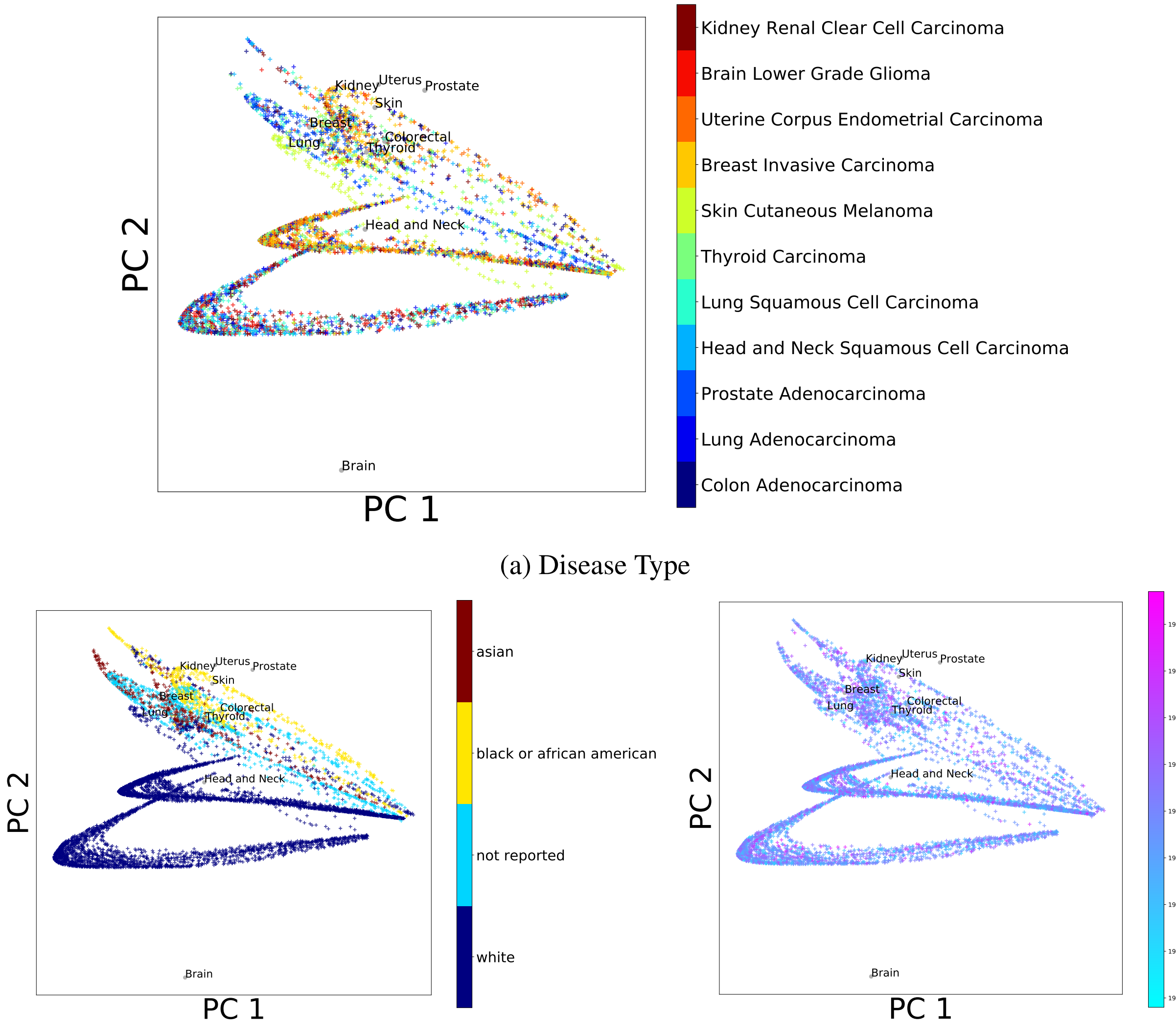
(f) Recovery of edges with weight magnitude > 0.2

(g) Recovery of edges with weight magnitude > 0.3

(h) Recovery of edges with weight magnitude > 0.4

(i) Recovery of edges with weight magnitude > 0.5

Figure 6.1: Recovery of Edges in Sample-Specific Graphs.



(a) Disease Type

(b) Patient Race

(c) Year of Birth

Figure 6.2: TCGA Personalized Networks with colored by covariates. For visualization, networks are compressed by PCA to 2 dimensions. Annotated gray points correspond to the population network estimated for samples of each particular tissue.

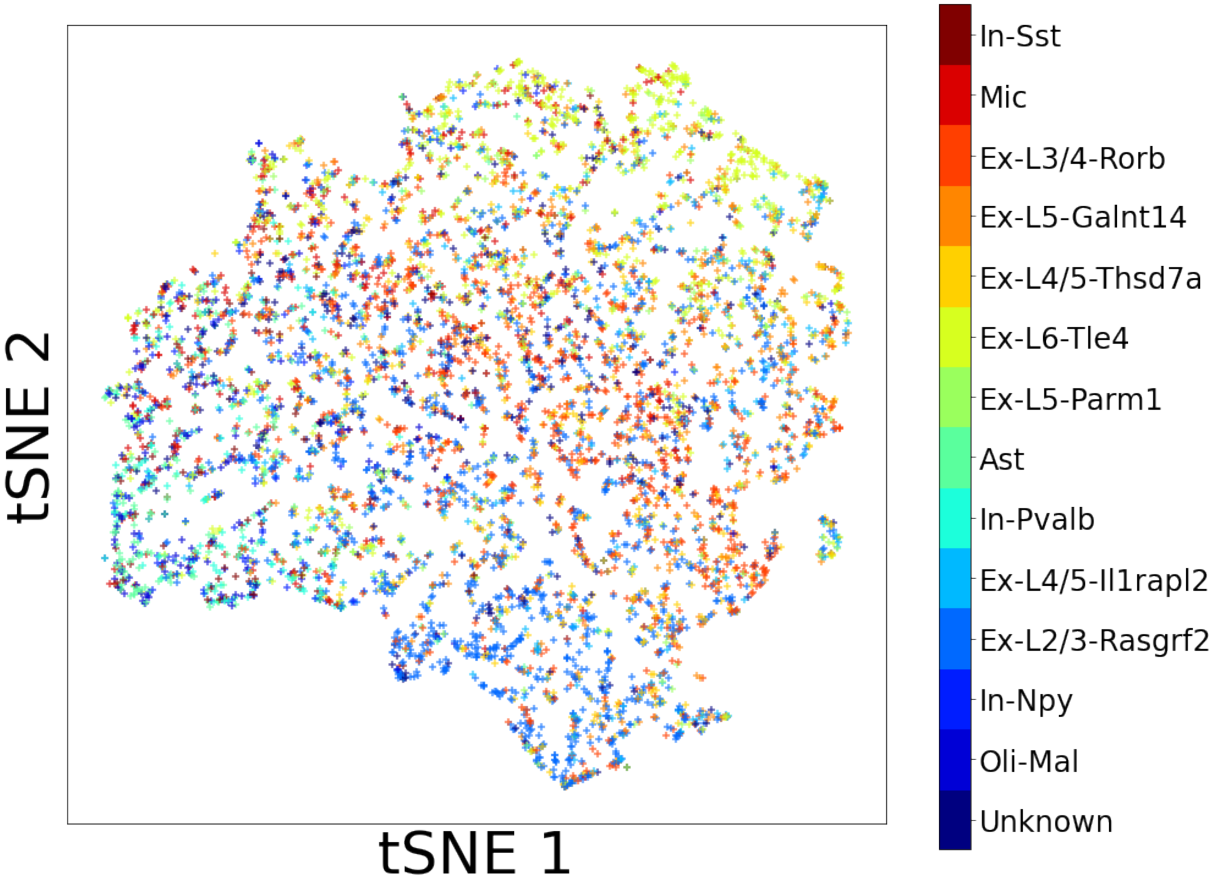


Figure 6.3: Cell-specific networks in mouse brains, colored according to cell type.

Chapter 7

Conclusions

7.1 Summary

In this thesis, we have explored the possibilities and challenges of *sample-specific models*. Since many scientific questions fundamentally ask about individual samples (e.g. “which drug should this patient receive?”), principled methods to estimate sample-specific inferences would be extremely useful. This is a very challenging problem because sample-specific inferences require flexibility between samples, which works against the law of large numbers. Thus, our challenge in performing sample-specific inference is to identify sample representations which inform us about sample membership in underlying sub-populations and use these representations to share power between similar samples.

We have shown this framework can be applied to histopathology images to estimate sample-specific transcriptomic models in cancer, and we have developed methods to estimate sample-specific model parameters even when a one-to-one function mapping sample representations to model parameters does not exist (such as for demographic features of cancer patients). Finally, we have extended these methods to estimate sample-specific network structure and begun to apply these techniques to estimate patient-specific gene networks for cancer patients and cell-specific gene networks for single-cell data. These experiments demonstrate that sample-specific models have potential to answer questions of sample-specific inference. Much work remains to improve methods for estimation of sample-specific models and post-hoc analysis of the estimated models.

7.2 Future Research Directions

This thesis has demonstrated the utility of sample-specific models and motivates further research into theory and applications of these flexible models.

Theoretical Understanding of Sample-Specific Models

Crystallizing Connection Between Interaction Effects, Dropout, and Multitask Learning As discussed in Chapter 3, sample-specific model estimation can be viewed as an extreme form of task-specific learning in which sample representations are used as task representations. Under this view, differences between task-specific models can be represented as pure interaction effects, and methods for sharing power between related tasks can be used to improve estimation of sample-specific models. In particular, I am interested in methods to use sample representations as input with a Dropout regularization to regularize the interaction between sample representations and model parameters (i.e. driving sample-specific models toward group-level models). In addition, these methods which rely on the accuracy of sample representations; extending methods to generate these sample representations from more diverse data sources (e.g. from distributional data and conceptual knowledge graphs [103]) could have a meaningful impact on the quality of the learned models. I believe these direction can give a new framework for estimating sample-specific models that is more scalable than the methods discussed in this thesis.

Network Estimation In Chapter 6, I proposed to estimate sample-specific networks via Contextualized NOTEARS. However, many questions regarding sample-specific networks remain open. For example, how much difficulty in estimating sample-specific networks is caused by the non-convexity of DAG sets? Are there conditions on the DAG archetype dictionary which could be enforced to guaranteed a convex set of archetypes, and if so, would those conditions assist or hamper estimation? Further, what is the consequences of choosing to limit matrix rank or choosing to limit the number of archetypes? These questions and others on the estimation of sample-specific networks remain open.

Analysis In addition to questions of estimation procedures, there are also open questions regarding the analysis of estimated sample-specific models. Once we have estimated sample-specific model parameters, what is the best way to summarize these new representations? Should we perform clustering on the estimated parameters, or is it best to present these models to users as sample-specific models?

Finally, when seeking to understand sample-specific models, we are interested in questions of identifiablity: how many sets of sample-specific models could equivalently recapitulate the observed data? For example, we know that both population and group-level linear models are identifiable in common conditions [75, 118, 156], but sample-specific models without covariates are not identifiable. What conditions on covariate structure must be met to retain identifiability of sample-specific models? Answering this question will help understanding which situations deserve analysis with sample-specific model estimation.

Discovering Molecular Profiles of Diseases

Complex diseases such as cancer [33, 53, 148] and Alzheimer's disease[134, 158, 182] are affected by many sources of variation which cause individuals to experience a unique patient journey. Traditional classification of diseases based on coarse-grained factors such as tissue morphology are increasingly outdated as fine-grained biological assays are revealing stunning heterogeneity at the granularity of individual cells [18, 128]. While these assays provide data at a finer resolution than previously possible, methods of analysis continue to rely on statistical methods which independently estimate cluster-level models. Increasing evidence points to molecular subtypes which do not form discrete clusters [122, 137]. Discovering sample-specific molecular profiles would refine our understanding of the mechanisms behind complex diseases, and may help to explain why different patients experience the same disease.

Regulatory Network Inference in Alzheimer's Disease In particular, methods of sample-specific network inference (discussed in Chapter 6) could be used to infer gene regulatory networks of complex diseases. Recent works have shown that neuronal cell types have strikingly distinct interactomes [133]; thus, estimating a single shared regulatory network will obscure the effects of neurodegenerative diseases. With the wealth of single-cell transcriptomics data recently available [128], we can apply the methods of personalized network inference to estimate different regulatory networks for different cells. In this way, we can uncover the heterogeneity of regulatory networks on the cellular level, and investigate the effectiveness of pre-defined cell types to capture regulatory heterogeneity, or possibly define new cell types which have different regulatory modifications.

Interpretable Models for Clinical Risk Assessment of Diverse Patient Cohorts In this thesis, we have developed some tools to identify risk factors in patient cohorts. These projects have included theoretical questions of model interpretability [24, 106] and applications to mortality risk in Covid-19. However, it is a well-known problem that machine learning models do not always generalize to diverse patient cohorts, including minority and underrepresented groups [59, 81]. In cohorts of diverse patients, the risk factors are not the same for every patient — thus, population-level models of risk are limited.

How can we extend clinical risk assessments to diverse patient cohorts? One approach is to develop statistical methods to overcome sample heterogeneity and estimate distinct risk models while also transferring statistical power for different groups of patients. Specifically, I aim to develop methods of automated cohort detection which can use the estimated patient-specific model coefficients to assist clinicians in understanding patient similarity. With these methods, we will be able to make risk assessments at the individual level and extend the benefits of the interpretable models to more diverse populations.

Multi-Cancer Discriminative Subtypes In addition to estimating discriminative subtypes of lung cancers, I am interested in estimating discriminative subtypes which may be shared be between cancer types. As in Chapter 5, applying machine learning techniques to estimate models shared among cancer types can increase statistical power and identify previously-unknown clusters of cancers which appear in multiple tissue types.

To perform this analysis, we could apply the procedures from Chapter 4 to the pan-cancer data in TCGA and TCIA. These methods naturally produce sample-specific embeddings which we can analyze (using the methods in Section 5) to understand whether discriminative subtypes are shared between or are unique to tissue types. Either result would have clinical implications; current clinical practice is to understand tumor through a top-down approach that first breaks tumors into groups by tissue, but from a molecular biology perspective it is not clear that this is the correct approach. Such analysis of multi-cancer discriminative subtypes could provide an indication of whether the most informative transcriptomic markers are shared between or unique to tissue types, giving insight on which direction the medical community should explore for more clinically-relevant tumor clusters.

Bibliography

[1] Khalid AbdulJabbar, Shan E Ahmed Raza, Rachel Rosenthal, Mariam Jamal-Hanjani, Selvaraju Veeriah, Ayse Akarca, Tom Lund, David A Moore, Roberto Salgado, Maise Al Bakir, et al. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nature Medicine*, pages 1–9, 2020. 4.3

[2] Ilieva I Ageenko, Kelley A Doherty, and Adrian Paul Van Cleave. Personalized lifetime financial planning tool, June 24 2010. US Patent App. 12/316,967. 3.3

[3] Amr Ahmed and Eric P Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29): 11878–11883, 2009. 6.1

[4] Maruan Al-Shedivat, Avinava Dubey, and Eric P Xing. Contextual explanation networks. *arXiv preprint arXiv:1705.10301*, 2017. 3.3, 3.3, 4.2, 6.3

[5] Maruan Al-Shedivat, Avinava Dubey, and Eric P Xing. Personalized survival prediction with contextual explanation networks. *arXiv preprint arXiv:1801.09810*, 2018. 3.3

[6] Ahmed M Alaa, Jinsung Yoon, Scott Hu, and Mihaela van der Schaar. Personalized risk scoring for critical care patients using mixtures of gaussian process experts. *ICML 2016 Workshop on Computational Frameworks for Personalization*, 2016. 4.1, 5

[7] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*, 2017. 5

[8] Jordan Ash, Gregory Darnell, Daniel Munro, and Barbara Engelhardt. Joint analysis of gene expression levels and histological images identifies genes associated with tissue morphology. *bioRxiv*, page 458711, 2018. 4.4

[9] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. 4.3

[10] D R Baldwin, B White, M Schmidt-Hansen, A R Champion, and A M Melder. Diagnosis and treatment of lung cancer: summary of updated NICE guidance. *BMJ*, 342, 2011. ISSN 0959-8138. doi: 10.1136/bmj.d2110. URL `https://www.bmj.com/content/342/bmj.d2110`. 4.1

[11] Sally Bamford, Emily Dawson, Simon Forbes, Jody Clements, Roger Pettett, Ahmet Do-

gan, A Flanagan, Jon Teague, P Andrew Futreal, Michael R Stratton, et al. The cosmic (catalogue of somatic mutations in cancer) database and website. *British journal of cancer*, 91(2):355–358, 2004. 4.2

[12] Adham Beykikhoshk, Thomas P Quinn, Samuel C Lee, Truyen Tran, and Svetha Venkatesh. Deeptriage: Interpretable and individualised biomarker scores using attention mechanism for the classification of breast cancer sub-types. *bioRxiv*, page 533406, 2019. 4.1

[13] Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-specific independence in bayesian networks. *arXiv preprint arXiv:1302.3562*, 2013. 6.2

[14] Leo Breiman and Jerome H Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):3–54, 1997. 3.1

[15] Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 2.2

[16] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12):1213, 2013. 1.1

[17] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, 2015. 1.1

[18] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155, 2015. 3.3, 4.1, 5, 6.1, 7.2

[19] Peter Bühlmann, Philipp Rütimann, Sara van de Geer, and Cun-Hui Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858, 2013. 6.2

[20] Kristina L Buschur, Maria Chikina, and Panayiotis V Benos. Causal network perturbations for instance-specific analysis of single cell and disease samples. *bioRxiv*, page 637710, 2019. 4.1, 6.2

[21] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0508-1. URL `https://doi.org/10.1038/s41591-019-0508-1`. 4.1, 4.1

[22] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 3.1

[23] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730. ACM, 2015. 2.1

[24] Chun-Hao Chang, Sarah Tan, Ben Lengerich, Anna Goldenberg, and Rich Caruana. How interpretable and trustworthy are gams? *arXiv preprint arXiv:2006.06466*, 2020. 7.2

[25] Deborah Chasman and Sushmita Roy. Inference of cell type specific regulatory networks on mammalian lineages. *Current opinion in systems biology*, 2:130–139, 2017. 6.4

[26] Chen Chen, Shihua Zhang, and Xiang-Sun Zhang. Discovery of cell-type specific regulatory elements in the human genome using differential chromatin modification analysis. *Nucleic acids research*, 41(20):9230–9242, 2013. 6.4

[27] Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457, 2019. 6.4

[28] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *ArXiv*, abs/1603.02754, 2016. 2.2

[29] Jun Cheng, Jie Zhang, Yatong Han, Xusheng Wang, Xiufen Ye, Yuebo Meng, Anil Parwani, Zhi Han, Qianjin Feng, and Kun Huang. Integrative Analysis of Histopathological Images and Genomic Data Predicts Clear Cell Renal Cell Carcinoma Prognosis. *Cancer Research*, 77(21):e91—-e100, 2017. ISSN 0008-5472. doi: 10.1158/0008-5472. CAN-17-0313. URL `https://cancerres.aacrjournals.org/content/77/21/e91`. 4.1

[30] David Maxwell Chickering. Learning bayesian networks is np-complete. In *Learning from data*, pages 121–130. Springer, 1996. 6.2

[31] David Maxwell Chickering and David Heckerman. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. *Machine learning*, 29 (2-3):181–212, 1997. 6.2

[32] Jean Chiou, Chia-Yi Su, Yi-Hua Jan, Chih-Jen Yang, Ming-Shyan Huang, Yung-Luen Yu, and Michael Hsiao. Decrease of FSTL1-BMP4-Smad signaling predicts poor prognosis in lung adenocarcinoma but not in squamous cell carcinoma. *Scientific reports*, 7(1):9830, aug 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-10366-2. URL `https://pubmed.ncbi.nlm.nih.gov/28852126https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5575295/`. 4.3

[33] Christine H Chung, Philip S Bernard, and Charles M Perou. Molecular portraits and the family tree of cancer. *Nature genetics*, 32(Supp):533, 2002. ISSN 1061-4036. 6.1, 7.2

[34] Eric A Collisson, Peter Bailey, David K Chang, and Andrew V Biankin. Molecular subtypes of pancreatic cancer. *Nature Reviews Gastroenterology & Hepatology*, 16(4): 207–220, 2019. ISSN 1759-5053. doi: 10.1038/s41575-019-0109-y. URL `https://doi.org/10.1038/s41575-019-0109-y`. 4.1

[35] Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019. 4.3

[36] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0177-5. URL `https://doi.org/10.1038/s41591-018-0177-5`. 4.1, 4.1, 4.2, 4.2, 4.3, 4.3, 4.4

[37] Judy M Coulson, Jodie L Edgson, Penella J Woll, and John P Quinn. A splice variant of the neuron-restrictive silencer factor repressor is expressed in small cell lung cancer: a potential role in derepression of neuroendocrine genes and a useful clinical marker. *Cancer research*, 60(7):1840–1844, 2000. 4.3

[38] Simon J Crabb, Maggie CU Cheang, Samuel Leung, Taina Immonen, Torsten O Nielsen, David D Huntsman, Chris D Bajdik, and Stephen K Chia. Basal breast cancer molecular subtype predicts for lower incidence of axillary lymph node metastases in primary breast cancer. *Clinical breast cancer*, 8(3):249–256, 2008. 4.1

[39] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970. 1.1

[40] Francis HC Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958. 1.1

[41] Charles S Dela Cruz, Lynn T Tanoue, and Richard A Matthay. Lung cancer: epidemiology, etiology, and prevention. *Clinics in chest medicine*, 32(4):605–644, 2011. 4.1

[42] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. An anova test for functional data. *Computational statistics & data analysis*, 47(1):111–122, 2004. 2.2

[43] Hao Dai, Lin Li, Tao Zeng, and Luonan Chen. Cell-specific network constructed by single-cell rna sequencing data. *Nucleic acids research*, 47(11):e62–e62, 2019. 6.2

[44] Mark A Dawson. The cancer epigenome: Concepts, challenges, and therapeutic opportunities. *Science*, 355(6330):1147–1152, 2017. ISSN 0036-8075. doi: 10.1126/science.aam7304. URL `https://science.sciencemag.org/content/355/6330/1147`. 4.1

[45] Jennifer B Dennison, Maria Shahmoradgoli, Wenbin Liu, Zhenlin Ju, Funda Meric-Bernstam, Charles M Perou, Aysegul A Sahin, Alana Welm, Steffi Oesterreich, Matthew J Sikora, et al. High intratumoral stromal content defines reactive breast cancer as a low-risk breast cancer subtype. *Clinical Cancer Research*, 22(20):5068–5078, 2016. 5

[46] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *IJCAI*, pages 2327–2333, 2015. 5

[47] Xin Dong, Dong Lin, Chris Low, Emily A. Vucic, John C. English, John Yee, Nevin Murray, Wan L. Lam, Victor Ling, Stephen Lam, Peter W. Gout, and Yuzhuo Wang. Elevated expression of birc6 protein in non-small-cell lung cancers is associated with cancer recurrence and chemoresistance. *Journal of Thoracic Oncology*, 8(2):161–170, 2013. ISSN

15561380. doi: 10.1097/JTO.0b013e31827d5237. 4.3

[48] Holly K Dressman, Andrew Berchuck, Gina Chan, Jun Zhai, Andrea Bild, Robyn Sayer, Janiel Cragun, Jennifer Clarke, Regina S Whitaker, LiHua Li, et al. An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *Journal of Clinical Oncology*, 25(5):517–525, 2007. 5

[49] Jianqing Fan and Wenyang Zhang. Statistical estimation in varying coefficient models. *Annals of Statistics*, pages 1491–1518, 1999. 3.3

[50] Dean A Fennell. Caspase Regulation in Non{\textendash}Small Cell Lung Cancer and its Potential for Therapeutic Exploitation. *Clinical Cancer Research*, 11(6):2097–2105, 2005. ISSN 1078-0432. doi: 10.1158/1078-0432.CCR-04-1482. URL `https://clincancerres.aacrjournals.org/content/11/6/2097`. 4.3

[51] Sebastiano Filetti, Jean-Michel Bidart, Franco Arturi, Bernard Caillou, Diego Russo, and Martin Schlumberger. Sodium/iodide symporter: a key transport system in thyroid cancer cell metabolism. *European Journal of Endocrinology*, 141(5):443–457, 1999. 5

[52] Aaron J. Fisher, John D. Medaglia, and Bertus F. Jeronimus. Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 115(27):E6106–E6115, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1711978115. URL `https://www.pnas.org/content/115/27/E6106`. 3.3

[53] R Fisher, L Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479, 2013. 5, 6.1, 6.1, 7.2

[54] Simon A Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, and Sari Ward. Cosmic: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*, 43(D1):D805–D811, 2014. 5, 5, 6.4

[55] Nir Friedman, Moises Goldszmidt, et al. Discretizing continuous attributes while learning bayesian networks. In *ICML*, pages 157–165, 1996. 6.2

[56] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vohringer, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *bioRxiv*, page 813543, 2019. 4.4

[57] Andrew Gelman. Statistical modeling, causal inference, and social science, Mar 2018. URL `https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/`. 2.2

[58] Arkadiusz Gertych, Zaneta Swiderska-Chadaj, Zhaoxuan Ma, Nathan Ing, Tomasz Markiewicz, Szczepan Cierniak, Hootan Salemi, Samuel Guzman, Ann E Walts, and Beatrice S Knudsen. Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. *Scientific Reports*, 9 (1):1483, 2019. ISSN 2045-2322. doi: 10.1038/s41598-018-37638-9. URL `https://doi.org/10.1038/s41598-018-37638-9`. 4.1, 4.1, 4.1

[59] Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, 2018. 7.2

[60] A Goldhirsch, E P Winer, A S Coates, R D Gelber, M Piccart-Gebhart, B Thürlimann, H-J Senn, and Panel Members. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Annals of oncology : official journal of the European Society for Medical Oncology*, 24(9):2206–2223, sep 2013. ISSN 1569-8041. doi: 10.1093/annonc/mdt303. URL `https://pubmed.ncbi.nlm.nih.gov/23917950https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755334/`. 4.1

[61] Isobel Claire Gormley, Thomas Brendan Murphy, et al. A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics*, 2(4): 1452–1477, 2008. 3.3

[62] A M Groeger, V Esposito, A De Luca, R Cassandro, G Tonini, V Ambrogi, F Baldi, R Goldfarb, T C Mineo, A Baldi, and E Wolner. Prognostic value of immunohistochemical expression of p53, bax, Bcl-2 and Bcl-xL in resected non-small-cell lung cancers. *Histopathology*, 44(1):54–63, 2004. doi: 10.1111/j.1365-2559.2004.01750.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2559.2004.01750.x`. 4.3

[63] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016. 4.2

[64] S M Haeger, J J Thompson, S Kalra, T G Cleaver, D Merrick, X-J Wang, and S P Malkoski. Smad4 loss promotes lung cancer formation but increases sensitivity to DNA topoisomerase inhibitors. *Oncogene*, 35(5):577–586, 2016. ISSN 1476-5594. doi: 10.1038/onc.2015.112. URL `https://doi.org/10.1038/onc.2015.112`. 4.3

[65] David Hallac, Jure Leskovec, and Stephen Boyd. Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 387–396. ACM, 2015. 3.3

[66] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000. 4.3

[67] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation, mar 2011. ISSN 00928674. 4.1, 4.3

[68] Jie Hao, Sai Chandra Kosaraju, Nelson Zange Tsaku, Dae Hyun Song, and Mingon Kang. PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data. In *Biocomputing 2020*, pages 355–366. WORLD SCIENTIFIC, nov 2019. ISBN 978-981-12-1562-9. doi: doi:10.1142/9789811215636_0032. URL `https://doi.org/10.1142/9789811215636{_}0032`. 4.1

[69] Sara A. Hart. Precision education initiative: Moving toward personalized education. *Mind,*

Brain, and Education, 10(4):209–211, 2016. doi: 10.1111/mbe.12109. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mbe.12109. 3.3

[70] Trevor Hastie and Rob Tibshirani. *Generalized Additive Models*. Chapman and Hall/CRC, 1990. 2.1

[71] Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993. 3.3, 3.3, 5

[72] Tristan J Hayeck, Noah A Zaitlen, Po-Ru Loh, Bjarni Vilhjalmsson, Samuela Pollack, Alexander Gusev, Jian Yang, Guo-Bo Chen, Michael E Goddard, Peter M Visscher, et al. Mixed model with correction for case-control ascertainment increases association power. *The American Journal of Human Genetics*, 96(5):720–730, 2015. 5

[73] Josie Hayes, Pier Paolo Peruzzi, and Sean Lawler. MicroRNAs in cancer: Biomarkers, functions and therapy, 2014. ISSN 1471499X. 4.1

[74] Yuan He, Surya B Chhetri, Marios Arvanitis, Kaushik Srinivasan, François Aguet, Kristin G Ardlie, Alvaro N Barbeira, Rodrigo Bonazzola, Hae Kyung Im, Christopher D Brown, et al. Mechanisms of tissue-specific genetic regulation revealed by latent factors across eqtls. *bioRxiv*, page 785584, 2019. 6.4

[75] Christian Hennig. Identifiablity of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296, 2000. 7.2

[76] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2.2

[77] Giles Hooker. *Diagnostics and Extrapolation in Machine Learning*. Stanford University, 2004. 2.2

[78] Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. 2.2

[79] Peter Horby, Wei Shen Lim, Jonathan R Emberson, Marion Mafham, Jennifer L Bell, Louise Linsell, Natalie Staplin, Christopher Brightling, Andrew Ustianowski, Einas Elmahi, et al. Dexamethasone in hospitalized patients with covid-19-preliminary report. *The New England journal of medicine*, 2020. 2.1

[80] Leora Horwitz, Simon A Jones, Robert J Cerfolio, Fritz Francois, Joseph Greco, Bret Rudy, and Christopher M Petrilli. Trends in covid-19 risk-adjusted mortality rates in a single health system. *medRxiv*, 2020. 2.1

[81] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019. 7.2

[82] Osamu Iizuka, Fahdi Kanavati, Kei Kato, Michael Rambeau, Koji Arihiro, and Masayuki Tsuneki. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Scientific Reports*, 10(1):1504, 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-58467-9. URL https://doi.org/10.1038/s41598-020-58467-9. 4.1, 4.1

[83] Marcin Imielinski, Alice H. Berger, Peter S. Hammerman, Bryan Hernandez, Trevor J. Pugh, Eran Hodis, Jeonghee Cho, James Suh, Marzia Capelletti, Andrey Sivachenko, Carrie Sougnez, Daniel Auclair, Michael S. Lawrence, Petar Stojanov, Kristian Cibulskis, Kyusam Choi, Luc De Waal, Tanaz Sharifnia, Angela Brooks, Heidi Greulich, Shantanu Banerji, Thomas Zander, Danila Seidel, Frauke Leenders, Sascha Ansén, Corinna Ludwig, Walburga Engel-Riedel, Erich Stoelben, Jürgen Wolf, Chandra Goparju, Kristin Thompson, Wendy Winckler, David Kwiatkowski, Bruce E. Johnson, Pasi A. Jänne, Vincent A. Miller, William Pao, William D. Travis, Harvey I. Pass, Stacey B. Gabriel, Eric S. Lander, Roman K. Thomas, Levi A. Garraway, Gad Getz, and Matthew Meyerson. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6): 1107–1120, sep 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.08.029. 4.3

[84] Kentaro Inamura. Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification. *Frontiers in Oncology*, 7:193, 2017. ISSN 2234-943X. doi: 10.3389/fonc.2017.00193. URL `https://www.frontiersin.org/article/10.3389/fonc.2017.00193`. 4.1

[85] Claudio Isella, Andrea Terrasi, Sara Erika Bellomo, Consalvo Petti, Giovanni Galatola, Andrea Muratore, Alfredo Mellano, Rebecca Senetta, Adele Cassenti, Cristina Sonetto, et al. Stromal contribution to the colorectal cancer transcriptome. *Nature genetics*, 47(4): 312, 2015. 5

[86] Fattaneh Jabbari and Gregory F Cooper. An instance-specific algorithm for learning the structure of causal bayesian networks containing latent variables. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 433–441. SIAM, 2020. 6.2

[87] Fattaneh Jabbari, Shyam Visweswaran, and Gregory F Cooper. Instance-specific bayesian network structure learning. *Proceedings of machine learning research*, 72:169, 2018. 6.2

[88] Fattaneh Jabbari, Shyam Visweswaran, and Gregory F. Cooper. Instance-specific bayesian network structure learning. In Václav Kratochvíl and Milan Studený, editors, *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*, volume 72 of *Proceedings of Machine Learning Research*, pages 169–180, Prague, Czech Republic, 11–14 Sep 2018. PMLR. URL `http://proceedings.mlr.press/v72/jabbari18a.html`. 3.3

[89] Fattaneh Jabbari, Shyam Visweswaran, and Gregory F Cooper. An empirical investigation of instance-specific causal bayesian network learning. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2582–2585. IEEE, 2019. 3.3, 4.1, 6.2

[90] Jiming Jiang. *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media, 2007. 3.3

[91] Bertrand Joseph, Jessica Ekedahl, Rolf Lewensohn, Philippe Marchetti, Pierre Formstecher, and Boris Zhivotovsky. Defective caspase-3 relocalization in non-small cell lung carcinoma. *Oncogene*, 20(23):2877–2888, 2001. ISSN 1476-5594. doi: 10.1038/sj.onc.1204402. URL `https://doi.org/10.1038/sj.onc.1204402`. 4.3

[92] Hyun Seok Kim, Saurabh Mendiratta, Jiyeon Kim, Chad Victor Pecot, Jill E Larsen, Iryna

Zubovych, Bo Yeun Seo, Jimi Kim, Banu Eskiocak, Hannah Chung, et al. Systematic identification of molecular subtype-selective vulnerabilities in non-small-cell lung cancer. *Cell*, 155(3):552–566, 2013. 4.1

[93] Sarah Kim-Hellmuth, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Valentin Wucher, Silva Kasela, Stephane E Castel, Andrew Hamel, Ana Viñuela, Amy L Roberts, et al. Cell type specific genetic regulation of gene expression across human tissues. *bioRxiv*, page 806117, 2019. 6.4

[94] Mladen Kolar, Le Song, and Eric P Xing. Sparsistent learning of varying-coefficient models with structural changes. In *Advances in neural information processing systems*, pages 1006–1014, 2009. 1.1

[95] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 6.1, 6.2

[96] Daisuke Komura and Shumpei Ishikawa. Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*, 16:34–42, 2018. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2018.01.001. URL `http://www.sciencedirect.com/science/article/pii/S2001037017300867`. 4.1, 4.1

[97] Marieke Lydia Kuijjer, Matthew Tung, GuoCheng Yuan, John Quackenbush, and Kimberly Glass. Estimating sample-specific regulatory networks. *arXiv preprint arXiv:1505.06440*, 2015. 4.1

[98] Marieke Lydia Kuijjer, Matthew George Tung, GuoCheng Yuan, John Quackenbush, and Kimberly Glass. Estimating sample-specific regulatory networks. *iScience*, 14:226–240, 2019. 3.3, 6.2

[99] Kimberly R Kukurba and Stephen B Montgomery. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11):pdb–top084970, 2015. 1.1

[100] Chandan Kumar-Sinha and Arul M Chinnaiyan. Precision oncology in the age of integrative genomics. *Nature biotechnology*, 36(1):46, 2018. 5

[101] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019. 2.1

[102] Francisco Alejandro Lagunas-Rangel. Neutrophil-to-lymphocyte ratio and lymphocyte-to-c-reactive protein ratio in patients with severe coronavirus disease 2019 (covid-19): A meta-analysis. *Journal of medical virology*, 2020. 2.1

[103] Ben Lengerich, Andrew Maas, and Christopher Potts. Retrofitting distributional embeddings to knowledge graphs with functional relations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2423–2436, 2018. 7.2

[104] Benjamin Lengerich, Maruan Al-Shedivat, Amir Alavi, Jennifer Williams, Sami Labbaki, and Eric P. Xing. Discriminative subtyping of lung cancers from histopathology images with contextual deep learning. *Under Review*, 2020. 4

[105] Benjamin Lengerich, Rich Caruana, and Yin Aphinyanaphongs. Targeting glucocorticoids

in covid-19 with neutrophil-lymphocyte ratio. 2020. 2

[106] Benjamin Lengerich, Sarah Tan, Chun-Hao Chang, Giles Hooker, and Rich Caruana. Purifying interaction effects with the functional anova: An efficient algorithm for recovering identifiable additive models. In *International Conference on Artificial Intelligence and Statistics*, pages 2402–2412, 2020. 2, 2.2, 2.2, 3.1, 3.2, 7.2

[107] Benjamin Lengerich, Eric P Xing, and Rich Caruana. On dropout, overfitting, and interaction effects in deep neural networks. *arXiv preprint arXiv:2007.00823*, 2020. 2

[108] Benjamin J Lengerich, Bryon Aragam, and Eric P Xing. Personalized regression enables sample-specific pan-cancer analysis. *Bioinformatics*, 34(13):i178–i186, 2018. 3.3, 4.1, 5

[109] Benjamin J. Lengerich, Bryon Aragam, and Eric P Xing. Learning sample-specific models with low-rank personalized regression. In *Advances in Neural Information Processing Systems (In Press)*, 2019. 3.3, 4.1, 5, 5

[110] Andrew C Leon and Moonseong Heo. Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Computational statistics & data analysis*, 53(3):603–608, 2009. 2.2

[111] Xiang Li, Shanghong Xie, Peter McColgan, Sarah J Tabrizi, Rachael I Scahill, Donglin Zeng, and Yuanjia Wang. Learning subject-specific directed acyclic graphs with mixed effects structural equation models from observational data. *Frontiers in genetics*, 9, 2018. 3.3

[112] Neal I Lindeman, Philip T Cagle, Mary Beth Beasley, Dhananjay Arun Chitale, Sanja Dacic, Giuseppe Giaccone, Robert Brian Jenkins, David J Kwiatkowski, Juan-Sebastian Saldivar, Jeremy Squire, Erik Thunnissen, and Marc Ladanyi. Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, 8 (7):823–859, jul 2013. ISSN 1556-1380. doi: 10.1097/JTO.0b013e318290868f. URL `https://pubmed.ncbi.nlm.nih.gov/23552377https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159960/`. 4.1

[113] Chong Liu, Jonathan C Sage, Michael R Miller, Roel GW Verhaak, Simon Hippenmeyer, Hannes Vogel, Oded Foreman, Roderick T Bronson, Akiko Nishiyama, and Liqun Luo. Mosaic analysis with double markers reveals tumor cell of origin in glioma. *Cell*, 146(2): 209–221, 2011. 5

[114] Guangbo Liu, Fen Pei, Fengqing Yang, Lingxiao Li, Amit Dipak Amin, Songnian Liu, J Ross Buchan, and William C Cho. Role of Autophagy and Apoptosis in Non-Small-Cell Lung Cancer. *International journal of molecular sciences*, 18(2):367, feb 2017. ISSN 1422-0067. doi: 10.3390/ijms18020367. URL `https://pubmed.ncbi.nlm.nih.gov/28208579https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5343902/`. 4.3

[115] Xiaoping Liu, Yuetong Wang, Hongbin Ji, Kazuyuki Aihara, and Luonan Chen. Personalized characterization of diseases using sample-specific networks. *Nucleic acids research*,

44(22):e164–e164, 2016. 4.1

[116] Xiaoping Liu, Yuetong Wang, Hongbin Ji, Kazuyuki Aihara, and Luonan Chen. Personalized characterization of diseases using sample-specific networks. *Nucleic acids research*, 44(22):e164–e164, 2016. 3.3

[117] Chad A Livasy, Gamze Karaca, Rita Nanda, Maria S Tretiakova, Olufunmilayo I Olopade, Dominic T Moore, and Charles M Perou. Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma. *Modern pathology*, 19(2):264, 2006. 5

[118] Lennart Ljung and Torkel Glad. On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2):265–276, 1994. 7.2

[119] Daniel Lopez-Martinez and Rosalind Picard. Multi-task neural networks for personalized pain recognition from physiological signals. *arXiv preprint arXiv:1708.08755*, 2017. 4.1, 5

[120] Marta Lucchetta, Isabelle da Piedade, Mohamed Mounir, Marina Vabistsevits, Thilde Terkelsen, and Elena Papaleo. Distinct signatures of lung cancer types: aberrant mucin O-glycosylation and compromised immune response. *BMC Cancer*, 19(1):824, 2019. ISSN 1471-2407. doi: 10.1186/s12885-019-5965-x. URL `https://doi.org/10.1186/s12885-019-5965-x`. 4.1

[121] Xin Luo, Xiao Zang, Lin Yang, Junzhou Huang, Faming Liang, Jaime Rodriguez-Canales, Ignacio I Wistuba, Adi Gazdar, Yang Xie, and Guanghua Xiao. Comprehensive Computational Pathological Image Analysis Predicts Lung Cancer Prognosis. *Journal of Thoracic Oncology*, 12(3):501–509, 2017. ISSN 1556-0864. doi: https://doi.org/10.1016/j.jtho.2016.10.017. URL `http://www.sciencedirect.com/science/article/pii/S1556086416312369`. 4.1

[122] Siyuan Ma, Shuji Ogino, Princy Parsana, Reiko Nishihara, Zhirong Qian, Jeanne Shen, Kosuke Mima, Yohei Masugi, Yin Cao, and Jonathan A Nowak. Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. *Genome biology*, 19(1):142, 2018. 4.1, 6.1, 7.2

[123] Siyuan Ma, Shuji Ogino, Princy Parsana, Reiko Nishihara, Zhirong Qian, Jeanne Shen, Kosuke Mima, Yohei Masugi, Yin Cao, Jonathan A. Nowak, Kaori Shima, Yujin Hoshida, Edward L. Giovannucci, Manish K. Gala, Andrew T. Chan, Charles S. Fuchs, Giovanni Parmigiani, Curtis Huttenhower, and Levi Waldron. Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. *Genome Biology*, 19(1):142, Sep 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1511-4. URL `https://doi.org/10.1186/s13059-018-1511-4`. 5

[124] Stephen P Malkoski and Xiao-Jing Wang. Two sides of the story? Smad4 loss in pancreatic cancer versus head-and-neck cancer. *FEBS letters*, 586(14):1984–1992, jul 2012. ISSN 1873-3468. doi: 10.1016/j.febslet.2012.01.054. URL `https://pubmed.ncbi.nlm.nih.gov/22321641https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3285395/`. 4.3

[125] Daniel Marbach, David Lamparter, Gerald Quon, Manolis Kellis, Zoltán Kutalik, and Sven Bergmann. Tissue-specific regulatory circuits reveal variable modular perturbations

across complex diseases. *Nature methods*, 13(4):366, 2016. 6.4

[126] Katherine Margulis, Albert S. Chiou, Sumaira Z. Aasi, Robert J. Tibshirani, Jean Y. Tang, and Richard N. Zare. Distinguishing malignant from benign microscopic skin lesions using desorption electrospray ionization mass spectrometry imaging. *Proceedings of the National Academy of Sciences*, 115(25):6347–6352, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1803733115. URL `https://www.pnas.org/content/115/25/6347`. 5

[127] Andriy Marusyk, Vanessa Almendro, and Kornelia Polyak. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 12(5):323, 2012. 5, 6.1

[128] Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, Jennie Z Young, Madhvi Menon, Liang He, Fatema Abdurrob, Xueqiao Jiang, et al. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337, 2019. 6.1, 6.1, 7.2, 7.2

[129] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *ICML (2)*, pages 343–351, 2013. 5

[130] Huaiyu Mi and Paul Thomas. *PANTHER Pathway: An Ontology-Based Pathway Database Coupled with Data Analysis Tools*, pages 123–140. Humana Press, Totowa, NJ, 2009. ISBN 978-1-60761-175-2. doi: 10.1007/978-1-60761-175-2_7. URL `https://doi.org/10.1007/978-1-60761-175-2_7`. 5

[131] Huaiyu Mi, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang, and Paul D. Thomas. Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1):D183–D189, 2017. doi: 10.1093/nar/gkw1138. URL +`http://dx.doi.org/10.1093/nar/gkw1138`. 5

[132] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee A D Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970—-E2979, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1717139115. URL `https://www.pnas.org/content/115/13/E2970`. 4.1, 4.1, 4.1, 4.2

[133] Shahin Mohammadi, Jose Davila-Velderrain, and Manolis Kellis. Reconstruction of cell-type-specific interactomes at single-cell resolution. *Cell Systems*, 9(6):559–568, 2019. 6.1, 6.4, 7.2

[134] Thomas J Montine and Kathleen S Montine. Precision medicine: Clarity for the clinical and biological complexity of alzheimer's and parkinson's diseases. *Journal of Experimental Medicine*, 212(5):601–605, 2015. 6.1, 7.2

[135] Hojin Moon, Hongshik Ahn, Ralph L Kodell, Songjoon Baek, Chien-Ju Lin, and James J Chen. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artificial intelligence in medicine*, 41(3):197–207, 2007. 4.1, 5

[136] Kevin Morgan, Gillian Spurlock, Richard C Brown, and M Afzal Mir. Release of a sodium transport inhibitor (inhibitin) from cultured human cancer cells. *Cancer research*, 46(12

Part 1):6095–6100, 1986. 5

[137] Thanos P Mourikis, Lorena Benedetti, Elizabeth Foxall, Julianne Perner, Matteo Cereda, Jesper Lagergren, Michael Howell, Christopher Yau, Rebecca Fitzgerald, Paola Scaffidi, et al. Patient-specific detection of cancer genes reveals recurrently perturbed processes in esophageal adenocarcinoma. *bioRxiv*, page 321612, 2018. 4.1, 6.1, 6.4, 7.2

[138] Hassan Muhammad, Carlie S. Sigel, Gabriele Campanella, Thomas Boerner, Linda M. Pak, Stefan Büttner, Jan N. M. IJzermans, Bas Groot Koerkamp, Michael Doukas, William R. Jarnagin, Amber Simpson, and Thomas J. Fuchs. Towards unsupervised cancer subtyping: Predicting prognosis using a histologic visual dictionary, 2019. 4.1, 4.1

[139] Sara Negrini, Ilaria Prada, Rosalba D'Alessandro, and Jacopo Meldolesi. Rest: an oncogene or a tumor suppressor? *Trends in cell biology*, 23(6):289–295, 2013. 4.3

[140] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418):61, 2012. 3.3

[141] Kenney Ng, Jimeng Sun, Jianying Hu, and Fei Wang. Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Summits on Translational Science Proceedings*, 2015:132, 2015. 3.3

[142] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019. 2.1

[143] Chris J Oates, Jim Q Smith, Sach Mukherjee, and James Cussens. Exact estimation of multiple directed acyclic graphs. *Statistics and Computing*, 26(4):797–811, 2016. 6.2

[144] Ankur P Parikh, Wei Wu, Ross E Curtis, and Eric P Xing. Treegl: reverse engineering tree-evolving gene networks underlying developing biological lineages. *Bioinformatics*, 27(13):i196–i204, 2011. 1.1, 6.1

[145] I Paul, A D Chacko, I Stasik, S Busacca, N Crawford, F McCoy, N McTavish, B Wilson, M Barr, K J O'Byrne, D B Longley, and D A Fennell. Acquired differential regulation of caspase-8 in cisplatin-resistant non-small-cell lung cancer. *Cell Death & Disease*, 3 (12):e449–e449, 2012. ISSN 2041-4889. doi: 10.1038/cddis.2012.186. URL `https://doi.org/10.1038/cddis.2012.186`. 4.3

[146] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 5

[147] Johan Pensar, Henrik Nyman, Timo Koski, and Jukka Corander. Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. *Data mining and knowledge discovery*, 29(2):503–533, 2015. 6.2

[148] P. Perez-Moreno, E. Brambilla, R. Thomas, and J. C. Soria. Squamous cell carcinoma of the lung: molecular subtypes and therapeutic opportunities. *Clin Cancer Res*, 18(9): 2443–51, 2012. ISSN 1078-0432 (Print) 1078-0432 (Linking). doi: 10.1158/1078-0432.CCR-11-2370. URL `https://www.ncbi.nlm.nih.gov/pubmed/22407829`. 6.1, 7.2

[149] Emma Pierson, Daphne Koller, Alexis Battle, Sara Mostafavi, GTEx Consortium, et al.

Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput Biol*, 11(5):e1004220, 2015. 6.4

[150] Jennifer Pittman, Erich Huang, Holly Dressman, Cheng-Fang Horng, Skye H Cheng, Mei-Hua Tsou, Chii-Ming Chen, Andrea Bild, Edwin S Iversen, and Andrew T Huang. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8431–8436, 2004. 5

[151] Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. Contextual parameter generation for universal neural machine translation. *arXiv preprint arXiv:1808.08493*, 2018. 3.3

[152] Vlad Popovici, Eva Budinska, Lenka Capkova, Daniel Schwarz, Ladislav Dusek, Josef Feit, and Rolf Jaggi. Joint analysis of histopathology image features and gene expression in breast cancer. *BMC Bioinformatics*, 17:1–9, 2016. 4.1, 4.4

[153] Milind M. Pore, T. J. Hiltermann, and Frank A. E. Kruyt. Targeting apoptosis pathways in lung cancer. *Cancer letters*, 332(2):359–368, May 28 2013. URL `http://pitt.idm.oclc.org/login?url=https://search-proquest-com.pitt.idm.oclc.org/docview/1668042138?accountid=14709`. Copyright - Copyright Elsevier Limited May 28, 2013; Last updated - 2015-04-10. 4.3

[154] Xavier Puig and Josep Ginebra. A bayesian cluster analysis of election results. *Journal of Applied Statistics*, 41(1):73–94, 2014. 3.3

[155] Martin Reck and Klaus F Rabe. Precision Diagnosis and Treatment for Advanced Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, 377(9):849–861, aug 2017. ISSN 0028-4793. doi: 10.1056/NEJMra1703413. URL `https://doi.org/10.1056/NEJMra1703413`. 4.1

[156] Olav Reiersøl. Identifiability of a linear relation between variables which are subject to error. *Econometrica: Journal of the Econometric Society*, pages 375–389, 1950. 7.2

[157] Jüri Reimand, Tambet Arak, Priit Adler, Liis Kolberg, Sulev Reisberg, Hedi Peterson, and Jaak Vilo. g: Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic acids research*, 44(W1):W83–W89, 2016. 4.3

[158] Christiane Reitz. Toward precision medicine in alzheimer's disease. *Annals of translational medicine*, 4(6), 2016. 6.1, 7.2

[159] Valeria Relli, Marco Trerotola, Emanuela Guerra, and Saverio Alberti. Abandoning the Notion of Non-Small Cell Lung Cancer. *Trends in Molecular Medicine*, 25 (7):585–594, 2019. ISSN 1471-4914. doi: https://doi.org/10.1016/j.molmed.2019.04.012. URL `http://www.sciencedirect.com/science/article/pii/S1471491419301017`. 4.1

[160] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016. 1.1, 5

[161] Sara Ricardo, André Filipe Vieira, Renê Gerhard, Dina Leitão, Regina Pinto, Jorge F Cameselle-Teijeiro, Fernanda Milanezi, Fernando Schmitt, and Joana Paredes. Breast cancer stem cell markers cd44, cd24 and aldh1: expression distribution within intrinsic molecular subtype. *Journal of clinical pathology*, 64(11):937–946, 2011. 4.1

[162] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396, 2014. 1.1

[163] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41 (4):801–814, 2018. 3.3

[164] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005. 6.1

[165] Ashis Saha, Yungil Kim, Ariel DH Gewirtz, Brian Jo, Chuan Gao, Ian C McDowell, Barbara E Engelhardt, Alexis Battle, François Aguet, Kristin G Ardlie, et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome research*, 27(11):1843–1858, 2017. 6.4

[166] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860, 2014. 1.1

[167] Claudio Scafoglio, Bruce A. Hirayama, Vladimir Kepe, Jie Liu, Chiara Ghezzi, Nagichettiar Satyamurthy, Neda A. Moatamed, Jiaoti Huang, Hermann Koepsell, Jorge R. Barrio, and Ernest M. Wright. Functional expression of sodium-glucose transporters in cancer. *Proceedings of the National Academy of Sciences*, 112(30):E4111–E4119, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1511698112. URL `http://www.pnas.org/content/112/30/E4111`. 5

[168] Stuart J Schnitt. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Modern Pathology*, 23(2):S60–S64, 2010. ISSN 1530-0285. doi: 10.1038/modpathol.2010.33. URL `https://doi.org/10.1038/modpathol.2010.33`. 4.1

[169] Jeong-Sun Seo, Ahreum Kim, Jong-Yeon Shin, and Young Tae Kim. Comprehensive analysis of the tumor immune micro-environment in non-small cell lung cancer for efficacy of checkpoint inhibitor. *Scientific Reports*, 8(1):14576, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-32855-8. URL `https://doi.org/10.1038/s41598-018-32855-8`. 4.1

[170] Masahito Shimojo, Yoshie Shudo, Masatoshi Ikeda, Tomoyo Kobashi, and Seiji Ito. The small cell lung cancer-specific isoform of re1-silencing transcription factor (rest) is regulated by neural-specific ser/arg repeat-related protein of 100 kda (nsr100). *Molecular Cancer Research*, 11(10):1258–1268, 2013. 4.3

[171] William M Shyu, Eric Grosse, and William S Cleveland. Local regression models. In

Statistical models in S, pages 309–376. Routledge, 1991. 3.3

[172] Wengong Si, Jiaying Shen, Huilin Zheng, and Weimin Fan. The role and mechanisms of action of microRNAs in cancer drug resistance. *Clinical Epigenetics*, 11(1):25, 2019. ISSN 1868-7083. doi: 10.1186/s13148-018-0587-8. URL `https://doi.org/10.1186/s13148-018-0587-8`. 4.1

[173] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1):7–30, 2020. 4.1

[174] Peter Sollich and Anders Krogh. Learning with ensembles: How overfitting can be useful. In *Advances in neural information processing systems*, pages 190–196, 1996. 5

[175] L. Song, M. Kolar, and E. P. Xing. Keller: estimating time-varying interactions between genes. *Bioinformatics*, 25(12):i128–36, 2009. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). 6.1

[176] Le Song, Mladen Kolar, and Eric P Xing. Time-varying dynamic bayesian networks. In *Advances in neural information processing systems*, pages 1732–1740, 2009. 1.1

[177] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning, and A. L. Borresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–74, 2001. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.191367098. URL `http://www.ncbi.nlm.nih.gov/pubmed/11553815`. 4.1

[178] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000. 6.1, 6.2

[179] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep neural network models for computational histopathology: A survey, 2019. 4.1, 4.1

[180] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 4.2

[181] Tetsuya Takata, Fumihiro Tanaka, Tomoko Yamada, Kazuhiro Yanagihara, Yosuke Otake, Yozo Kawano, Tatsuo Nakagawa, Ryo Miyahara, Hiroki Oyanagi, Kenji Inui, and Hiromi Wada. Clinical significance of caspase-3 expression in pathologic-stage I, nonsmall-cell lung cancer. *International Journal of Cancer*, 96(S1):54–60, jan 2001. ISSN 0020-7136. doi: 10.1002/ijc.10347. URL `https://doi.org/10.1002/ijc.10347`. 4.3

[182] Lin Tan, Teng Jiang, Lan Tan, and Jin-Tai Yu. Toward precision medicine in neurological diseases. *Annals of translational medicine*, 4(6), 2016. 6.1, 7.2

[183] William D Travis, Elisabeth Brambilla, Allen Burke, Alexander Marx, and Andrew G Nicholson. *WHO classification of tumours of the lung, pleura, thymus and heart*. International Agency for Research on Cancer, 2015. 4.1

[184] Oana Ursu, James T Neal, Emily Shea, Pratiksha I Thakore, Livnat Jerby-Arnon, Lan Nguyen, Danielle Dionne, Celeste Diaz, Julia Bauman, Mariam Mosaad, et al. Massively

parallel phenotyping of variant impact in cancer with perturb-seq reveals a shift in the spectrum of cell states induced by somatic mutations. *bioRxiv*, 2020. 6.4

[185] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014. (document), 5.1

[186] Michel E Vandenberghe, Marietta L J Scott, Paul W Scorer, Magnus Söderberg, Denis Balcerzak, and Craig Barker. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Scientific Reports*, 7(1):45938, 2017. ISSN 2045-2322. doi: 10.1038/srep45938. URL `https://doi.org/10.1038/srep45938`. 4.1

[187] Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. In *International Conference on Machine Learning*, pages 6448–6457, 2019. 4.1

[188] Shyam Visweswaran, Derek C Angus, Margaret Hsieh, Lisa Weissfeld, Donald Yealy, and Gregory F Cooper. Learning patient-specific predictive models from clinical data. *Journal of biomedical informatics*, 43(5):669–685, 2010. 3.3

[189] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018. 2.1

[190] Matthew P Wagoner, Kearney TW Gunsalus, Barry Schoenike, Andrea L Richardson, Andreas Friedl, and Avtar Roopra. The transcription factor rest is lost in aggressive breast cancer. *PLoS genetics*, 6(6), 2010. 4.3

[191] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333, 2014. 5

[192] Shidan Wang, Alyssa Chen, Lin Yang, Ling Cai, Yang Xie, Junya Fujimoto, Adi Gazdar, and Guanghua Xiao. Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Scientific Reports*, 8 (1):10393, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-27707-4. URL `https://doi.org/10.1038/s41598-018-27707-4`. 4.1, 4.1, 4.1

[193] Jason W Wei, Laura J Tafe, Yevgeniy A Linnik, Louis J Vaickus, Naofumi Tomita, and Saeed Hassanpour. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific Reports*, 9(1):3358, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-40041-7. URL `https://doi.org/10.1038/s41598-019-40041-7`. 4.1, 4.1, 4.1

[194] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013. 5

[195] Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003. 5

[196] Jianpeng Xu, Jiayu Zhou, and Pang-Ning Tan. Formula: Factorized multi-task learning for task discovery in personalized medical models. In *Proceedings of the 2015 International Conference on Data Mining*. SIAM, 2015. 1.1, 3.3, 5

[197] Makoto Yamada, Koh Takeuchi, Tomoharu Iwata, John Shawe-Taylor, and Samuel Kaski. Localized lasso for high-dimensional regression. *stat*, 1050:20, 2016. 1.1, 3.3, 5

[198] Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300468. URL `https://doi.org/10.1145/3290605.3300468`. 1.1

[199] J M Yingling, M B Datto, C Wong, J P Frederick, N T Liberati, and X F Wang. Tumor suppressor Smad4 is a transforming growth factor beta-inducible DNA binding protein. *Molecular and Cellular Biology*, 17(12):7019–7028, 1997. ISSN 0270-7306. doi: 10.1128/MCB.17.12.7019. URL `https://mcb.asm.org/content/17/12/7019`. 4.3

[200] Jin young Yoo, Chi Hong Kim, So Hyang Song, Byoung Yong Shim, Youn Ju Jeong, Meyung Im Ahn, Suji Kim, Deog Gon Cho, Min Seop Jo, Kyu Do Cho, Hong Joo Cho, Seok Jin Kang, and Hoon Kyo Kim. Expression of caspase-3 and c-myc in non-small cell lung cancer. *Cancer research and treatment : official journal of Korean Cancer Association*, 36(5):303–307, oct 2004. ISSN 2005-9256. doi: 10.4143/crt.2004.36.5.303. URL `https://pubmed.ncbi.nlm.nih.gov/20368820https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2843874/`. 4.3

[201] Kun-Hsing Yu, Ce Zhang, Gerald J Berry, Russ B Altman, Christopher Ré, Daniel L Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7(1):12474, 2016. ISSN 2041-1723. doi: 10.1038/ncomms12474. URL `https://doi.org/10.1038/ncomms12474`. 4.1

[202] Bin Zhang, Chris Gaiteri, Liviu-Gabriel Bodea, Zhi Wang, Joshua McElwee, Alexei A Podtelezhnikov, Chunsheng Zhang, Tao Xie, Linh Tran, Radu Dobrin, Eugene Fluder, Bruce Clurman, Stacey Melquist, Manikandan Narayanan, Christine Suver, Hardik Shah, Milind Mahajan, Tammy Gillis, Jayalakshmi Mysore, Marcy E MacDonald, John R Lamb, David A Bennett, Cliona Molony, David J Stone, Vilmundur Gudnason, Amanda J Myers, Eric E Schadt, Harald Neumann, Jun Zhu, and Valur Emilsson. Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, 153(3):707—720, April 2013. ISSN 0092-8674. doi: 10.1016/j.cell.2013.03.030. URL `http://europepmc.org/articles/PMC3677161`. 6.1

[203] Ming Zhao, Lopa Mishra, and Chu-Xia Deng. The role of TGF-β/SMAD4 signaling in cancer. *International journal of biological sciences*, 14(2):111–123, jan 2018. ISSN 1449-2288. doi: 10.7150/ijbs.23230. URL `https://pubmed.ncbi.nlm.nih.gov/29483830https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5821033/`. 4.3

[204] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483, 2018. 6.2, 6.3

[205] S Zhou, P Buckhaults, L Zawel, F Bunz, G Riggins, J L Dai, S E Kern, K W Kinzler, and B Vogelstein. Targeted deletion of Smad4 shows it is required for transforming growth factor beta and activin signaling in colorectal cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, 95(5):2412–2416, mar 1998. ISSN 0027-8424. doi: 10.1073/pnas.95.5.2412. URL `https://pubmed.ncbi.nlm.nih.gov/9482899https://www.ncbi.nlm.nih.gov/pmc/articles/PMC19358/`. 4.3

[206] X. Zhu, J. Yao, X. Luo, G. Xiao, Y. Xie, A. Gazdar, and J. Huang. Lung cancer survival prediction from pathological images and genetic data — an integration study. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1173–1176, 2016. 4.1, 4.1

[207] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. 5