Bayesian Models for Multi-omic Multi-system Integration for Precision Oncology

by

Rupam Bhattacharyya

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2023

Doctoral Committee:

Professor Veerabhadran Baladandayuthapani, Co-Chair
Assistant Professor Nicholas Henderson, Co-Chair
Professor Jian Kang
Professor Alexey Nesvizhskii
Professor Xiang Zhou

Rupam Bhattacharyya

rupamb@umich.edu

ORCID iD: 0000-0003-4292-2372

To
Late Ajeet Kumar Bhattacharya,
Late Uma Bhattacharya,
Late Jyotiprasad Chakraborty,
Chandana Chakraborty.

ACKNOWLEDGEMENTS

Acknowledgements are futile. Neither because they are usually the part the typical reader skips the fastest, nor due to the sheer lack of uniqueness that they bear - although those are perfectly valid reasons to highlight their futility. But it's none of that. Acknowledgements are, rather, limited by the carriage that drives them - the alphabet. The good old collection of letters which, when arranged in certain familiar orders, can definitely guide expression of a plethora of emotions, is still nothing but a tiny, flickering lamp before the storms of human psyche. And yet, we do write them. All of us. The bestselling author penning their 50th book, the impressionable eighth grader writing their first project report, me shaping my last formal academic activity - we all try. We attempt to express our gratitude, in words, to people who deserve much more than that. People who deserve a tight hug, a holding of hands that seems to be eternal, or an evening spent in silence, just sitting side by side. But where's the time for that? This meeting today, that presentation tomorrow, yet another deadline day after - pushing everything away in time until you never even have the chance. So this will have to do. For now, at least. If you have made it this far, what's coming in the next paragraphs reaches nowhere close to expressing what these people mean to me. And yet, this will have to do, for now.

How do I arrange this? My OCD tells me I should fix some order at this point - so that's what I am going to do now. A habit I had during my childhood was to read comics from the last page. Weird, for sure - but I did like it. And it has been years since I have done that. So let's do that, shall we? I start from 2023, and I rewind back in time. And if I am doing so - then this must begin with the people who drove this dissertation towards the shape you chanced upon today. Prof. Veera Baladandayuthapani - the man who started his journey in the department during the same month as me, long back in September 2018 - has been the friend I needed through these five years. I am restless and impatient by nature - and his calm assessment of every situation - academic or non-academic - may not be something I am able to imitate in the near future, but is what kept me in line for the majority of the recent past. I do not need to tell you about the superlative science he does, a simple google search can yield that information for you. But I wish from the core of my heart that all of you have a VB in your lives - for the moments when everything seems to fall apart will come, and it is during those moments that you will need a guiding light like him. Prof. Nick Henderson - naturally reticent and possessor of a brilliant mind - acted not just as my co-advisor, but as an encyclopedia of useful directions when needed. He would probably speak three times during a one-hour meeting, and yet, those words would usually solve a challenge I spent three days thinking about. Most often, I would not even know if he was there while walking by his office - but I do know how much of this dissertation was molded by his careful and meticulous efforts. A few words do no justice to that.

This is where the academic environment in UM (and UM Biostat, in particular) reaches a level few other places in the world can offer. Most of us would be grateful to have one or two such people alongside, but UM offers many more. My committee members, for example, extended their helping hands whenever I needed them. Prof. Xiang Zhou spent his valuable time behind reading my works and returning with feedbacks that only raised their levels. He provided detailed and strong recommendations when I was looking for a postdoctoral positions, and the opportunity that awaits me after this is possible in large parts thanks to him. Prof. Jian Kang - the computational genius, as Veera likes to describe him - suggested tricks and tips that made my life exponentially easier. When I was looking for options to handle the flexible Gaussian process modeling in a varying coefficient context for Chapter V, it was his basis expansion technique that saved the day. He provided both relevant literature and computing guidance which made the project reach fruition. Prof. Alexey Nesvizhskii made sure I was aware of the scientific implications my works had. He pointed me towards emerging areas of precision oncology, including novel and enriched datasets that open up a window of possibilities for further research in the directions I pursue. And finally, all of them, amidst their busy schedules, managed to find time to provide thoughtful suggestions and allow me to schedule the proposal and defense meetings flexibly. I could not have asked for anything better.

Beyond my dissertation, UM has been a place of multidimensional research and learning for me. The biggest piece in that jigsaw comes from Prof. Bhramar Mukherjee, who made it a point to put her scientific and leadership skills into action during the COVID-19 pandemic. In the process, she inspired a large group of researchers from different domains and generations to conflate towards producing some of the most insightful knowledge about the disease - both in terms of prediction and prevention. Being a part of that effort contributed immensely to my growth as a collaborative scientist - seeking productive interactions both within and outside the department. Between working with her, with Veera, and collaborations outside the department, I met people who I each learnt something from - be it a style of writing, or a good practice for code management, or curiosity in general. Maxwell Salvatore, Prof. Debashree Ray, Dr. Lili Wang, Prof. Min Jin Ha, Prof. Sayantan Banerjee, Dr. Rajmohan Panda, Indrajit Banerjee - working closely with these people and many others I am missing here taught me the value of interactions and exchanging opinions in science. The BayesRx lab, that I have been a part of during the entirety of my stay here, brought me in contact with some of the most brilliant and versatile minds. Prof. Shariq Mohammed, (soon-to-be Dr.) Tsung-Hung Yao, Qingzhi Liu, Liying Chen, Nathaniel Osher, Chinmay Raut, Aditya Jalin - if you are reading this, it has been a pleasure being the part of a team with you all!

My UM story is incomplete without mentioning the excellent teaching I have received here - Prof. Rod Little's course on the history of statistics, Prof. Zhenke Wu's course on repeated measures data, Prof. Shawn Lee's human genetics course, Prof. Hui Jiang's big data computing class, Prof. Peisong Han's course on missing data, and Prof. Marisa Eisenberg's class on mathematical modeling in epidemiology are only a few of the several classes I have substantially benefitted from. A special mention should go to Prof. Jeremy Taylor, Prof. Tom Braun, and Prof. Phil Boonstra - who taught me possibly the most significant biostat course I have ever taken - namely, Biostat 699. I came to the department as a student with the necessary quantitative and mathematical skills, but little idea of how to make my work presentable and readable. I was a different student after the conclusion of the course - all thanks to the painstaking efforts of correcting the minutest details in my reports that these three people undertook. The department staff always were there to make every technicality simpler. Nicole Fenech, David Kubacki, Mandi Larson, Fatma Nedjari, Kyle Terwillegar, Wendy Mashburn, Sabrina Olsson, Lauren Detzler - they all guided me through things as simple as submitting a form or as complex as scheduling events and booking locations for them. Dan Barker and Mike Kleinsasser provided computing reinforcements whenever I needed them for my research, sometimes even responding during off-hours or weekends to cater to emergency needs. None of the shiny dashboards or cluster computations mentioned in this dissertation would materialized without them.

I absolutely hate the idea of living in a metro city. Hence, Ann Arbor was the perfect place for me to spend some of the key years of my life. Quaintly flavored with modern offerings, state-of-the-art campus yet comprising of historical bits, an old bookshop right beside a polished restaurant - Ann Arbor offered me the exact combination of twentyfirst century life and old-school sentiments I craved. The garnish on that curry was added by the friends I have had here through these years. Dr. Diptavo Dutta, Dr. Anwesha Bhattacharyya, Dr. Aritra Guha, Dr. Avijit Shee, Dr. Debarghya Mukherjee (none of whom I address by those honorifics) introduced me to the yearly fun of barbeques or the weekly fun of goat biryani at Paradise. Soumik, Debraj, Sunrit, Saptarshi, Pramit, Arkajit, Ritoban, Swaraj, Soham B, Soham D, and Bishvanwesha joined the crew over the years, and together we made memories that I will cherish forever. I wish I could insert an emotional "will miss you all" here, but to your grave misfortune, I am continuing at Ann Arbor for a couple more years now. This list would remain incomplete without the four grand old men - Dr. Shariq Mohammed, Dr. Satwik Acharyya, Dr. Somnath Mahapatra, and Dr. Sagnik Bhaduri - who have gone through several bombardments of curiosities on life in general in past and recent years from my end, and yet have never failed to be there for me. But when it comes to being there, how can I forget the family I have had here away from family? Prof. Mousumi Banerjee's efforts to make us feel at home on every cultural occassion brought me fresh air whenever I needed it, (along with plenty of food and songs to go with, of course). She even made me perform Rabindrasangeet - a feat that I felt would better be restricted to empty rooms than in front of an audience. Prof. Ananda Sen, Mala Chakraborty, and Prof. Abhijit Biswas made me a part of MILITS, opening up a new world of cultural exploration. Ashish Kaku, Jayanti Kakima, and Prof. Upali Nanda always made sure I was doing fine in every phase of my presence here. They are huge reasons behind why I decided to stay here for a while more.

The me that came to Ann Arbor in August 2018 lacked a lot of what I have now, except for the zeal towards exploring new horizons that I still carry. A substantial portion of that was formulated at the Indian Statistical Institute, Kolkata, where I spent some of the best years of my life. Being part of my ISI cohort has brought me many experiences, most of which I will remember fondly forever. The football discussions with Shounak and Anamitra, the evenings of gossip in Soumya's and Sumit's rooms, the absolutely off-key singing of Bohemian Rhapsody and We are the Champions at the hostel wings with Imon, the hours of mischief in the computer lab with Subrata, Santanu and Souvik, scanning pages of solved problems from Sayan's notebooks, Sarthak and Nilashis cheering for us from the back during hostel quizzes, trying to stop Sukanya from fleeing from the genetics class - it was an unending kaleidoscope of moments that I hold very close to my heart. Every single quiz I have been to with Arnab and Dipanjan (the latter sadly has stated in recent times he is unlikely to tread on those paths again) was an enthralling encounter beyond the points scored and the trophies won. Last but not the least, ISI made me familiar with Subha's cooking skills, which I have had the luxury of enjoying more than any of my batchmates, thanks to having him as my housemate at Ann Arbor. As a collective, we faced some of the darkest scenarios of our lives in ISI, including the sudden loss of a friend, but we all had each other's backs. I am proud to have been a part of this cohort.

The other thing ISI offered me was a support system like no other. Beyond teaching and providing recommendation letters, Prof. Ayanendranath Basu, Prof. Tapas Samanta, Prof. Arijit Chakrabarti, Prof. Sumitra Purkayastha, Prof. Rita Saha Ray, Prof. Atanu Biswas and many others extended their guiding arms in ways I could not expect. Prof. Parthanil Roy, Prof. Rajat Subhra Hazra, and Prof. Raghunath Chatterjee taught me to always follow my instincts and stand firmly with my own decisions, even if the circumstances were challenging. It is due largely to the inspiration and belief provided by them that I pursued biostatistics when none of my cohort-mates were willing to do so. But this point about joining a biostat PhD brings me to probably the hardest set of sentences I have to zot down today. I have to write about the man who always would tell me to chase my dreams. He was the person that encouraged my homesick mind to go out and explore the world. From recommendations about airlines to stories from conferences, he painted in front of me the life I am living now long before I gave it any serious thought. He assured me that it was okay to have a quirky and sometimes politically incorrect sense of humor in a world that is desparately trying to hide its flaws behind a fake sense of vanity. Most significantly, he taught me the importance of keeping in touch, rekindling lost connections, and being genuine to people. Prof. Saurabh Ghosh, I remember every bit of advice you ever gave me, along with every inappropriate joke that we shared. The flame you ignited is still here with all its strength.

This luck of having supportive mentors has followed me since the very beginning. In Santipur, I was taught by Mr. Jayanta Acharya, who I fondly call Mukulmama. He has been with me from the beginning of my academic journey - back when I had only joined primary school. He inculcated in me the seeds of being ambititious, which is something he does successfully even to this date. Alongside, I have come into contact with people like Nayan Kumar Sarkar, Prakash Chandra Dey, Jayanta Banerjee, Chandan Debnath, Sujoy Pramanik, Atanu Saha, and many others, who have always seen my ups and downs as their own. Some of them taught me within the school premises - some outside - but the affection I enjoy from them till this date has remained the same even after these many years. If not for Rahul Basu, who taught me math during my 11th and 12th standards, I would probably have skipped sitting for the ISI entrance exam thinking it was too tough for me to crack, thus missing out on the entirety of this journey I have been a part of since. I am poor at saying thanks - therefore, the gratitude I have for these people will probably never be expressed as I want them to be. But then, something is better than nothing, I guess.

I did already mention quizzing in this write-up. And quizzing landed me a friend like Ranit Paul, eight years back from now. What started as a decision to team up for a single contest turned out to be one of the best decisions of my life. The last time we competed in a quiz together was five years ago - but when has that stopped us from staying in touch? Ranit has been an inspiration for me, in both his striving for excellence and his humane ways of looking at the world. Quiz also connected me to the online group we call Starcrazy, which was created during the pandemic. Discussing trivias, obscure and well-known alike, over zoom, this group was one of the things that kept me sane during the lockdown. Numerous seniors, juniors, and batchmates connected weekends after weekends, which would otherwise have been extremely boring for me. Likewise, the other group that we started during the pandemic was Football Kyajra. Intended to be a no-holds-barred space for trolls, memes, and digs related to football, it has now expanded into a multiverse of frenzy and laughter that cleanses my mind at the end of every single day. I can't stress enough how fortunate I am to have got in touch with these folks. I have two loving families now - one each from mine and my wife's side, and I have people who I know I can rely upon. Pishimoni (Ranjana Chakraborty), Pishemoshai (Milan Kumar Chakraborty), Babu Dada (Shuvendu Chakraborty) have never failed to check in on me even when I forgot to call them, pressurized by work. Samir Basu and Lily Basu - my parents-in-law, who I still call Kaku and Kakima to this date because old habits die hard, celebrate my every joy like theirs. I could keep on listing people here who have showered me with unconditional love, but then that would be the dissertation itself.

It is only fit that I talk about the four pillars of my life at the end of this - the four people who enrich me every single day in ways they probably aren't aware themselves. Let's start with talking about Dr. Tuhin Majumder. People say it is sheer luck to find a friend who shares the same madness as you, and by that parameter, I have n times that luck. Little did I have any idea for the first 18 years of my life that a boy with the same screws loose as me did exist somewhere in the same state as I resided in. Yet, for the last 10 years, I have never felt for a single moment that I did not know him before. I do not have clear memories of a day when we did not share at least one meme between one another or did not mention each other on some stupid post on Facebook. His support towards my efforts and ambitions are sometimes only comparable to blind faith - but I can safely guarantee that that feeling is mutual. Tuhin, I can't tell you how glad I am that you have confirmed your alliance towards biostatistics. I wish nothing but success for you in all fronts - except for East Bengal winning a trophy that actually matters.

My wife, Dishari, who I know for the last 12 years, but have only paid attention to for the past four, is best described as a curious specimen. I have never seen another person who has the same level of belief in others, yet keeps underrating themselves like hell. Dishari, as I tell you always, you are destined for great things. If you are not sure, believe in my words - for I have believed in yours every single time when I have felt stuck and down. This last year, in particular, has been magical thanks to your presence beside me every single moment. In all turmoils and trepidations, I have had a home to come back to in you. The world for sure is getting worse in many ways - but I still dare to dream for a better tomorrow because you are there with me. I wish I can inspire you like you have inspired me at every stage of my progress, towards every word I have written in this document. I want to be there for you on every big occasion of yours like you have been here for me. Besides, we still have many more countries to explore, don't we?

I will finish this section with a story about two persons. They got married more than 28 years back from now. Their lives got uprooted several times afterwards, and they had to sail through many a hurricanes for over a decade. They made sacrifice their go-to option when it came to their own comfort, and focused on their son's upbringing as their singular priority. During the time when it was becoming a norm in India for parents to push their children toward their own choices, I was fortunate to have them as my parents, always keeping faith upon me. To this date, I know how it pains them to meet me only for a single month or a bit more every year, and yet they keep their emotions in chains so that I don't become weak mentally. Instead, they engage themselves in learning new things everyday. As I finish my dissertation, one of them is researching better choices of soil for gardening, and the other is preparing to take the exam for her driving license. I have close to 25 years to spend before being in the same age group as they are now, and if I retain half the excitement they have in exploring new avenues by then, I will consider myself blessed.

If you, the reader, have made it through to this sentence, I have only two things to offer to you - congratulations and condolences, for being able to go through this manuscript-length compendium of ramblings. What awaits in the rest of the document is nowhere as boring, I assure, and logically far more sound. I hope you venture into those waters!

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

The molecular heterogeneity of cancer makes it challenging to delineate the underlying mechanisms and optimize therapeutic avenues. Large-scale cancer datasets across multiple dimensions (such as omics and clinical data types, or cancer systems including patients and cell lines) offer assistance towards mitigating these challenges via a granular yet holistic view of the disease. While integrative approaches have the potential to both unmask novel functional mechanisms and prioritize therapeutic targets, their development and implementation is challenging due to data variety and the underlying dependence within/between such datasets. In this dissertation, I focus on developing Bayesian statistical procedures that can take advantage of the diversity offered by such databases while taking into account the associated biological and statistical challenges.

In Chapter II, I develop TransPRECISE, a multiscale Bayesian network modeling framework, to analyze the pan-cancer patient and cell line interactome. I assess pan-cancer pathway activities of patients from 31 tumor types and cell lines from 16 lineages, along with the cell lines' response to 481 drugs. TransPRECISE captures differential and conserved proteomic pathway circuitries between multiple patient and cell line lineages. Tumor stratification using these learned networks uncovers distinct clinical subtypes of patient cancers characterized by different cell line avatars. High predictive accuracy is observed for cell line drug sensitivities using Bayesian additive regression tree models with TransPRECISE pathway scores as predictors.

In Chapter III, I propose fiBAG, an integrative hierarchical Bayesian framework for modeling the fundamental biological relationships underlying cross-platform molecular features of cancer. Using Gaussian process models, fiBAG identifies upstream functional evidence for proteogenomic biomarkers. By mapping said evidence to prior inclusion probabilities, a calibrated Bayesian variable selection (cBVS) model is built to identify biomarkers associated with an outcome of interest. Simulation studies show that cBVS has higher power to detect disease-related markers than non-integrative approaches. Via an integrative proteogenomic analysis of 14 cancer datasets, several known and novel genes/proteins associated with cancer stemness and patient survival are identified.

While multi-omic patient databases have sparse drug response, cancer model systems databases provide extensive pharmacogenomic profiles, albeit with lower sample sizes, resulting in reduced statistical power. For this reason, in Chapter IV, I propose BaySyn - a hierarchical Bayesian evidence synthesis framework that detects functionally relevant driver genes based on their associations with upstream regulators and uses this evidence to calibrate Bayesian variable selection models in the (drug) outcome layer. I use BaySyn to analyze multi-omic patient and cell line datasets across pan-gynecological cancers. BaySyn mechanistic models implicate several known functional genes in GO and KEGG gene sets of interest in the cancers assessed. Further, the BaySyn outcome model makes more discoveries than its uncalibrated counterparts under equal Type I error control.

In Chapter V, I focus on incorporating tumor heterogeneity in clinicogenomic models. To this end, I propose GPVIBES, a Gaussian process-based varying coefficient model using Bayesian variable selection, to model the association between a biomarker and an outcome as a function of a hierarchical covariate equipped with horseshoe prior-based shrinkage. Simulation studies with one or more hierarchical covariates show that at the same signal-to-noise and sample-size-to-dimensionality ratios, GPVIBES yields improved selection performance alongside accurate estimates of the coefficient function, compared

to other varying-coefficient-based models. A pan-cancer integrative analysis of 16 cancers identified modulation of proteomic associations via several known signatures.

CHAPTER I

Introduction

Background Across several applied corridors of research, complex and high-dimensional datasets are becoming increasingly common, focusing on a wide range of natural and interdisciplinary sciences. Thus, data integration methods, equipped to handle complex, heterogeneous datasets in order to synthesize information from them in a systematic manner have become increasingly common in the recent times (Boehm et al., 2022). Across many scientific domains, data integration approaches have relied on several common techniques to achieve a few broad goals – (i) to aggregate and synthesize information in a synchronized manner, and (ii) to learn complementary information from sources that by themselves may not be complete enough to provide a clear answer to the scientific question at hand. Thus, the primary challenges of creating and implementing such procedures are also common across different sciences – the size, complexity, heterogeneity, and possible interdependence of the data sources can all contribute to the said overall challenge (Mirza et al., 2019)). Biological datasets, in particular, are of no exception – it has become increasingly customary to use multiple omics (i.e., data on several components contributing to cellular functions, such as genomics, proteomics, metabolomics, epigenomics, and transcriptomics) and phenotypes (potentially both genetic marker-based or related to survival or health indicators) simultaneously to uncover the mechanisms underlying a disease or optimize therapeutic solutions governing them (Vahabi and Michailidis, 2022). These methods cover an extensive spectrum of methodological research and may belong to categories such as learning protein interaction and functions (e.g., Wang et al. (2013a); Ma et al. (2014); Alonso-López et al. (2019); Vitrinel et al. (2019)), identifying biomarker association and prioritization of such markers (e.g., Hwang et al. (2012); Bromberg (2013); Leclercq et al. (2019)), optimization and discovery of therapy and/or drug regimens (e.g., Katsila and Matsoukas (2018); Hudson (2021); Pak et al. (2023)), and personalized medicine (e.g., Harris et al. (2014); Ozer et al. (2020)). In particular, data integration methods have been particularly useful in precision oncology (Nicora et al., 2020), which is the primary application area of this thesis, as I discuss next.

Precision oncology: data availability and potential for integration Cancer is a complex disease characterized by accumulation of small changes at a cellular level across several molecular layers (Byrne et al., 2006; Gentles and Gallahan, 2011; La Porta and Zapperi, 2017). Thus, to build a holistic view of a cancer of interest it is essential to consider multiple features of the disease via several platforms that can provide complementary and high-resolution snapshots. Pan-omic-clinical integrative approaches, hence, are of supreme importance in cancer (Boehm et al., 2022). Innovation towards such procedures has particularly been fueled by the increasing availability of data via databases covering not only multiple sequencing platforms, clinical results, and phenotypes, but also different cancer microenvironments such as patient tumors and model systems. Such databases may focus on patients (e.g., International Cancer Genome Consortium (ICGC, Zhang et al. (2019a)), The Cancer Genome Atlas (TCGA, Weinstein et al. (2013)), Pan-Cancer Analysis of Whole Genomes (PCAWG, Hoadley et al. (2018)), The Cancer Proteome Atlas (TCPA, Li et al. (2013, 2017a))), model systems (e.g., Genomics of Drug Sensitivity in Cancer (GDSC, Yang et al. (2012)), the Cancer Cell Line Encyclopedia (CCLE, Barretina et al. (2012)), the MD Anderson Cell Lines Project (MCLP, Li et al. (2017b))), along with drug profiling (e.g., NCI60 (Grever et al., 1992), the Library of Integrated Network-Based Cellular Signatures (LINCS, Keenan et al. (2018)), Broad Institute Connectivity Map (CMAP, Lamb et al. (2006); Lamb (2007); Subramanian et al. (2017)), The Cancer Dependency Map (DepMap, Tsherniak et al. (2017))), and so on. This dissertation is focused on formulating integrative procedures that can accommodate different combinations of such data and guide the detection of cellular cancer mechanisms as well as potential avenues for treatment, as is described next.

Outline and progression of the dissertation projects In essence, there are at least two distinct dimensions along which integrative methods in precision oncology must successfully be implemented – (i) the variability and heterogeneity in multiplatform data must be acknowledged and utilized, (ii) the difference in cancer microenvironments in the patient's tumor and a model system must be incorporated in the integrative procedure. This dissertation focuses on a sequence of research projects aiming to assess these dimensions both individually and simultaneously, as summarized in Figure 1.1. The individual chapters are organized as follows. Chapter II focuses on TransPRECISE, a Bayesian network-based integration procedure to combine and compare multi-system proteomic pathways. TransPRECISE analyzes the pan-cancer patient and cell line interactome – to both globally assess cell lines as representative models for patients, and develop drug sensitivity prediction models. TransPRECISE captures differential and conserved proteomic pathway circuitries between multiple patient and cell line lineages and uncovers distinct clinical subtypes based on tumor stratification using the learned networks. Chapter III describes fiBAG, a hierarchical Bayesian framework to integrate multi-omics and clinical data from patients. fiBAG identifies upstream functional evidence for proteogenomic biomarkers and mapping said evidence to prior inclusion probabilities, builds a calibrated Bayesian variable selection (cBVS) model to identify the biomarkers associated with an outcome of interest. cBVS shows higher power to detect disease-related markers than non-integrative approaches. Chapter IV covers BaySyn, a multi-stage Bayesian pipeline to integrate multi-platform data across both patient tumors and cancer models. While multi-omic patient databases have sparse drug response, cancer model systems databases, despite covering a wide range of pharmacogenomic platforms, provide lower sample sizes, resulting in reduced statistical power. To address this, BaySyn detects functionally relevant driver genes based on their associations with upstream regulators and combines evidence from multiple systems to calibrate Bayesian variable selection models in the (drug) outcome layer. The calibrated BaySyn outcome model makes more discoveries than its uncalibrated counterparts under equal Type I error control. Chapter V is focused on development of GPVIBES, a Gaussian process-based varying coefficient model using Bayesian variable selection. GPVIBES models the association between a biomarker and an outcome/phenotype using a Gaussian process specification as a function of the hierarchical covariate equipped with horseshoe prior-based shrinkage. Using simulation studies based on synthetic datasets with one or more hierarchical covariates, the performance of GPVIBES is compared with existing varying coefficient-based frequentist and Bayesian procedures. At the same signal-to-noise and sample size to dimensionality ratios, GPVIBES yields improved selection performance in terms of the AUC and Matthew's correlation coefficient, alongside accurate estimates of the regression coefficient function. A pan-cancer integrative analysis is also performed using GPVIBES, based on data from 16 TCGA cancers, utilizing overall survival as the outcome, more than 200 proteomic expressions as the biomarkers, and a total of 68 immune signatures as the hierarchical covariates. The pan-cancer analysis identifies several known key signatures, such as the modulation of the association of EGFR and YAP protein expressions with survival by CD8 T lymphocyte proportion in the tumor microenvironment for BRCA. Each chapter contains method-specific review of relevant literature along with presentation of the developed methodology and the analysis results. We conclude with a discussion on future research directions in Chapter VI.

Key scientific and statistical themes The progressive chapters of this dissertation are inherently connected via two thematic axes. First, the chapters and the methods presented in them are linked via the scientific challenges that motivate them from the context of cancer data integration. As presented in Figure 1.1, the methods utilize increasingly multi-platform and multi-system data to decipher cancer mechanisms and improve detection of biomarker associations and identification of potential therapeutic targets. In each project, the primary target is to summarize subsets of the cellular oncological mechanism. In Chapter II, this is achieved by modeling the interactions between the proteins in a pathway of interest. In Chapter III and Chapter IV, this is performed by modeling the association between proteogenomic biomarkers and their corresponding upstream DNA-level information. Finally, in Chapter IV, the tumor microenvironment is accommodated in the model via capturing the modulation of proteomic expression and its association with outcomes/phenotypes by component-specific summaries of the tumor microenvironment. Further, in all the chapters, a key focus is on estimating the association of cellular expression-level summaries with outcomes of interest – such as the the association of proteomic pathway scores with drug response in Chapter II, the association of proteogenomic expression with cancer stemness index and overall survival in Chapter III, the association of mRNA expression with drug response in Chapter IV, and finally the association of proteomic expression with overall survival in Chapter V. The second axis of connection between the chapters stems from

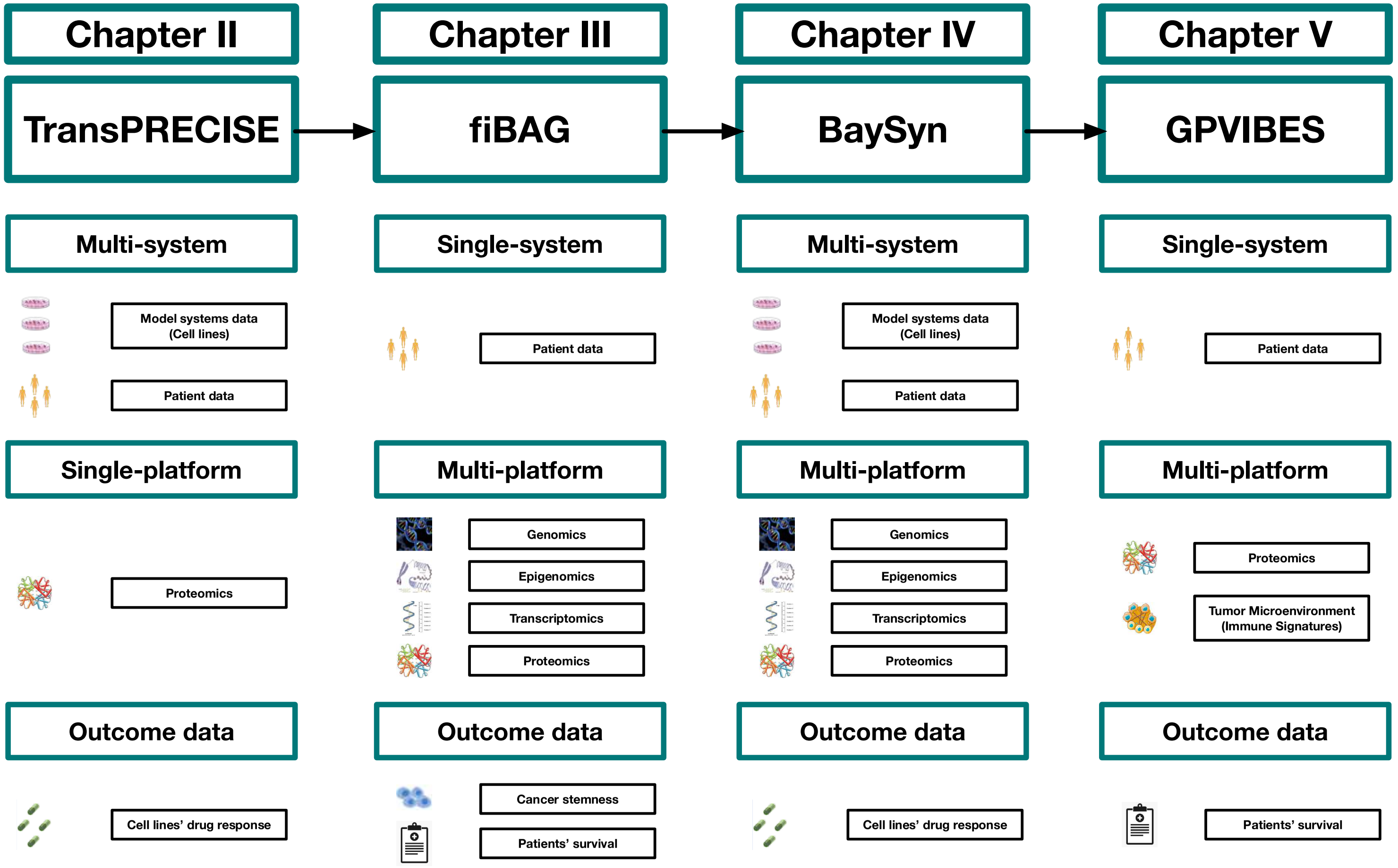


Figure 1.1: Overview of dissertation chapters. Chapter II focuses on TransPRECISE, a Bayesian network-based integration procedure to combine and compare multi-system proteomic pathways. Chapter III describes fiBAG, a hierarchical Bayesian framework to integrate multi-omics and clinical data from patients. Chapter IV covers BaySyn, a multi-stage Bayesian pipeline to integrate multi-platform data across both patient tumors and cancer models. Chapter V describes GPVIBES, a Gaussian process-based varying coefficient modeling procedure using Bayesian variable selection to integrate tumor microenvironment summaries in clinicogenomic models.

several methodological techniques employed. In all chapters, we use Bayesian machine learning procedures to incorporate information from the data at hand and any potential prior knowledge – either from previous, existing studies, or from multi-level models in the framework itself. In particular, I use a Bayesian network regression model in Chapter II for modeling the protein-protein interaction in the pathways. A Gaussian process speficiation is used to capture potential nonlinear effects or interactions in both the mechanistic models in Chapter III and Chapter IV, and for the varying coefficients in Chapter V. Bayesian variable selection procedures are employed via the Bayesian additive regression tree-based drug response models in Chapter II, the calibrated spike-and-slab prior-based outcome models in Chapter III and Chapter IV, and the beta-binomial prior-based selection procedure in Chapter V. Further, in Chapter V, we induce sparsity in the varying coefficients using a horseshoe prior specification on the coefficient parameters corresponding to the basis expansion of the modified squared exponential kernel. Last but not the least, since each project focuses on outcome association models that contain multiple covariates, corrections for multiple comparisons become essential. Since each outcome model allows us to compute a posterior probability of inclusion for each covariate, this is typically performed via false discovery rate control procedures employed on these probabilities.

Scientific end-user resources of the chapters Each chapter in this dissertation is aimed at eliciting interpretable outputs from the integrative models that can guide future oncological investigation and decision-making, as is discussed below. To ensure improved accessibility in part of users from all scientific domains, I built interactive R shiny-based dashboards for each project summarizing the key overview of the computational framework, some details on the sample size and the cancer types in the data used for the integrative analyses, and the end-user outputs. Each dashboard also offers the processed datasets and source codes for download, in order to maintain reproducibility of the results presented in this dissertation.

- In Chapter II, the key end-user outputs are the estimated cancer- and sample-specific proteomic pathway networks, and the rankings of the pathways specific to their association with a given drug/treatment regime for a specific cancer tissue. The patient and cell line level networks provide a summary of the conserved and differential pathway circuitry for each cancer type, and the pathway rankings provide a prioritization of potential drug targets for future investigations for specific cancers. The shiny dashboard for this chapter is available at `https://bayesrx.shinyapps.io/TransPRECISE/`.

- In Chapter III and Chapter IV, the mechanistic evidence quantities (log- or pseudo-Bayes factors) quantify the functional relevance of a gene/protein for a given cancer, and also at a pan-cancer level for cancer groups of interest. The posterior inclusion probabilities for each gene/protein from the calibrated outcome models help prioritizing the proteogenomic biomarkers in terms of their association with the outcome. The specific magnitude and direction of this association, adjusted for the presence of other biomarkers, is provided by the estimated coefficient parameter from the outcome model. The shiny dashboard for Chapter III is available at `https://bayesrx.shinyapps.io/Functional_iBAG/`, and the shiny dashboard for Chapter IV is available at `https://bayesrx.shinyapps.io/BaySyn/`.

- In Chapter V, the estimated posterior inclusion probabilities quantify the selection of a proteomic covariate in terms of its association with the outcome (survival in the real-data analyses). The association itself is now no more quantified as a point estimate, but is estimated as a function of the value of the hierarchical covariate (some immune signature score for the real-data analyses). Thus, this offers a quantification of the modulation of the association between the outcome and the proteomic covariate due to these tumor microenvironment summaries. The shiny dashboard is available at `https://bayesrx.shinyapps.io/GPVIBES/`.

CHAPTER II

Personalized Network Modeling of the Pan-Cancer Patient and Cell Line Interactome

2.1 Introduction

Precision medicine aims to improve clinical outcomes by optimizing treatment to each individual patient. The rapid accumulation of large-scale pan-omic molecular data across multiple cancers on patients (ICGC (Zhang et al., 2019a), TCGA (Weinstein et al., 2013), PCAWG (Hoadley et al., 2018), TCPA (Li et al., 2013, 2017a)) and model systems (GDSC (Yang et al., 2012), CCLE (Barretina et al., 2012), MCLP (Li et al., 2017b)), along with extensive drug profiling data (NCI60 (Grever et al., 1992), LINCS (Keenan et al., 2018), CMAP (Lamb et al., 2006; Lamb, 2007; Subramanian et al., 2017), DepMap (Tsherniak et al., 2017)) have generated information-rich and diverse community resources with major implications for translational research in oncology (Goodspeed et al., 2016). However, a major challenge remains: to bridge anticancer pharmacological data to large-scale omics in the paradigm wherein patient heterogeneity is leveraged and inferred through rigorous and integrative data-analytic approaches across patients and model systems.

Complex diseases, such as cancer, are often characterized by small effects in multiple genes and proteins that are interacting with each other by perturbing downstream cellular signaling pathways (Boyle et al., 2017; Creixell et al., 2015; Yao et al., 2018). It is well established that complex molecular networks and systems are formed by a large number of interactions of genes and their products operating in response to different cellular conditions and cell environment, i.e., model systems (Bandyopadhyay et al., 2010). To date, most, if not all approaches for mechanism and drug discovery have been constrained by the biological system (patients or cell-lines) (Geeleher et al., 2014; Kim et al., 2018), specific cancer lineage (Sinha et al., 2017; Sun and Liu, 2015), or by prior knowledge of specific genomic alterations (Domcke et al., 2013; Jiang et al., 2016). Hence, there is a critical need for robust computational methods that integrate molecular profiles across large cohorts of patients and model systems from multiple lineages in an unbiased data-driven manner to delineate specific regulatory mechanisms, uncover drug targets and pathways, and develop individualized predictive models in cancer.

Recently, a network-based framework called PRECISE (personalized cancer-specific integrated network estimation model) has been developed to estimate cancer-specific networks, infer patient-specific networks, and elicit interpretable pathway-level signatures (Ha et al., 2018). Using a large cohort of patients from TCGA across 30+ tumor types, PRECISE identifies pan-cancer commonalities and differences in proteomic network biology within and across tumors, allows robust tumor stratification that is both biologically and clinically informative, and has superior prognostic power compared to multiple existing approaches (Ha et al., 2018). In this chapter, I present translational PRECISE (TransPRECISE, in short), a generalization of the PRECISE framework, to establish the translational relevance of these pathway signatures. Briefly, TransPRECISE uses a multiscale Bayesian modeling strategy that infers de novo differential and conserved networks of intra-pathway circuitry between the two biological systems (patients and cell lines) for multiple cancers. Further, it identifies cell line “avatars” for patients based on pathway activities, and also develops machine learning based predictive models for drug sensitivity in both cell lines and patients to potentially guide pathway-based individualized medical decision-making. I also have developed an online, publicly available, comprehensive interactive database and visualization tool of our findings along with software code, hosted at `https://bayesrx.shinyapps.io/TransPRECISE/`.

2.2 Datasets and Methods

2.2.1 Cancer Patients' Proteomic Data

I used a dataset of 7,714 patient samples across 31 different cancer types available from the Cancer Proteome Atlas (TCPA) (Li et al., 2013, 2017a), as summarized in Table S2.1. TCPA offers reverse-phase protein array (RPPA)-based proteomics datasets, profiled using extensively validated antibodies to nearly 200 proteins and phosphoproteins. The functional space of the antibodies covers major functional and signaling pathways relevant to human cancers. For this work, I used a total of 12 pathways, including DNA damage response, EMT, hormone signaling, apoptosis, TSC/mTOR, and RAS/MAPK (Table S2.2).

2.2.2 Cancer Cell Lines' Proteomic and Drug Sensitivity Data

I used RPPA-based protein expression data for cell lines available via the MD Anderson Cell Lines Project (MCLP) (Li et al., 2017b). In set of 640 cancer cell lines spanning across 16 lineages, each cell line has RPPA expression data based on the same set of proteins as in the patient tumors (Table S2.3). Additionally, I used drug sensitivity data from the Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al., 2012) database, with the sensitivity of 481 drugs assessed on a subset of 254 cell lines (Table S2.4). For the entirety of this chapter, I denote cell line samples in lowercase and patient samples in uppercase letters.

Imputing missing cell line expressions Unlike the patient tumor expression data, the cell lines expression data has some amount (approximately 6%) of missing values. I use the function `impute.knn` from the Bioconductor package `impute` (Hastie et al., 2011) for imputation. For the k-nearest neighbor imputation implementation to run, each sample must have $<$ 50% of missing data across all the variables, and each variable must have $<$ 80% of missing data across all the samples. The original RPPA data set collected from MCLP has expression data on 651 cell lines from 194 genes that are common with our patient tumor RPPA dataset. Using the missing data upper bounds, I end up with 648 cell lines. I further removed eight more cell lines from some lineages (prostate (3), cervix (1), thyroid (1) and missing lineage (3)) that had too small sample sizes to be useful in fitting a stable Bayesian graphical regression model. I also removed one gene by the missing data criterion. Thus, I obtained our final set of 640 cell lines from 16 different cancer lineages with data on 193 proteins. I executed the imputation on this global profile consisting of all proteins at hand, instead of only the subset of proteins in the 12 functional pathways of interest, since the imputation would be more informative this way and would not reflect any undue bias towards possible interactions within and between the pathways of interest. After completing imputation, I simply use the subset of the imputed data set with the proteins in the pathways of interest.

2.2.3 TransPRECISE Framework

The TransPRECISE implementation can broadly be classified into three modules (Figure 2.1). The first module takes as input the combined proteomics data from patients and cell lines (as described above). The second module implements a Bayesian network modeling framework, providing the cancer-specific pathway networks and sample-specific pathway scores as outputs. The final module predicts patient drug responses based on models trained on the cell lines. The network module uses a Bayesian regression model, in which each protein is regressed on all the other proteins in the same pathway. The selected set of interacting protein-protein pairs then constitute the population-level cancer-specific pathway network. Given this population-level cancer-type-specific network, I deconvolve it to the sample-specific networks and scores, using the status of each protein (node) in the networks as neutral, suppressed, or activated. I then use these cancer-specific networks for pan-cancer and across model systems identification of conserved and differential pathway activities. In the final module, I use these sample-specific scores for identifying matching avatar cell lines for patient samples and predicting drug sensitivity. In the rest of this section, I describe the computational steps and details behind the TransPRECISE framework.

Step 1: Bayesian estimation of cancer-specific pathway networks I aim to estimate cancer-specific pathway networks using Bayesian regression methods on each of the proteins. I begin with fixing one cancer type in one model system and one of the 12 pathways of interest. In this pathway of interest, for an interactive relationship among proteins, let $w_{ij} = w_{ji}$ be the weight connecting protein i and protein j, with the number of proteins in the pathway of choice denoted by p. I set $w_{ij} = w_{ji} = 0.5 \; \forall i \neq j$ and $w_{ii} = 0 \; \forall i$. Suppose $\boldsymbol{y}_i$ is the $n \times 1$ vector containing expression values of protein i for n samples from the fixed cancer lineage. For protein i, the $n \times 1$ expression vector $\boldsymbol{y}_i$ (centered with its mean) is modeled as the following.

$$\boldsymbol{y}_i = \sum_{j \neq i} \beta_{ij} \boldsymbol{y}_j + \boldsymbol{\varepsilon}_i = \boldsymbol{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \tag{2.1}$$

where $\boldsymbol{\varepsilon}_i \sim \boldsymbol{N}_n(\boldsymbol{0}_n, \sigma_i^2 \boldsymbol{I}_n)$. $\boldsymbol{\beta}_i$ is the vector of all the regression coefficients, and $\boldsymbol{Z}_i$ is the design matrix consisting of expression data for all the other proteins. I employ *Zellner's g-prior* (Agliari and Parisetti, 1988) on $\boldsymbol{\beta}_i$ as the following.

$$\boldsymbol{\beta}_i \mid g = \boldsymbol{N}_{p-1}(\boldsymbol{0}_{p-1}, \sigma_i^2 \left(\boldsymbol{Z}_i^T \boldsymbol{Z}_i / g\right)^{-1}). \tag{2.2}$$

The hyper-parameter g reflects the prior on $\boldsymbol{\beta}_i = 0$. A higher value of g implies more probable deviation from $\boldsymbol{\beta}_i = 0$, and I assigned $g = n$ by default, which yields

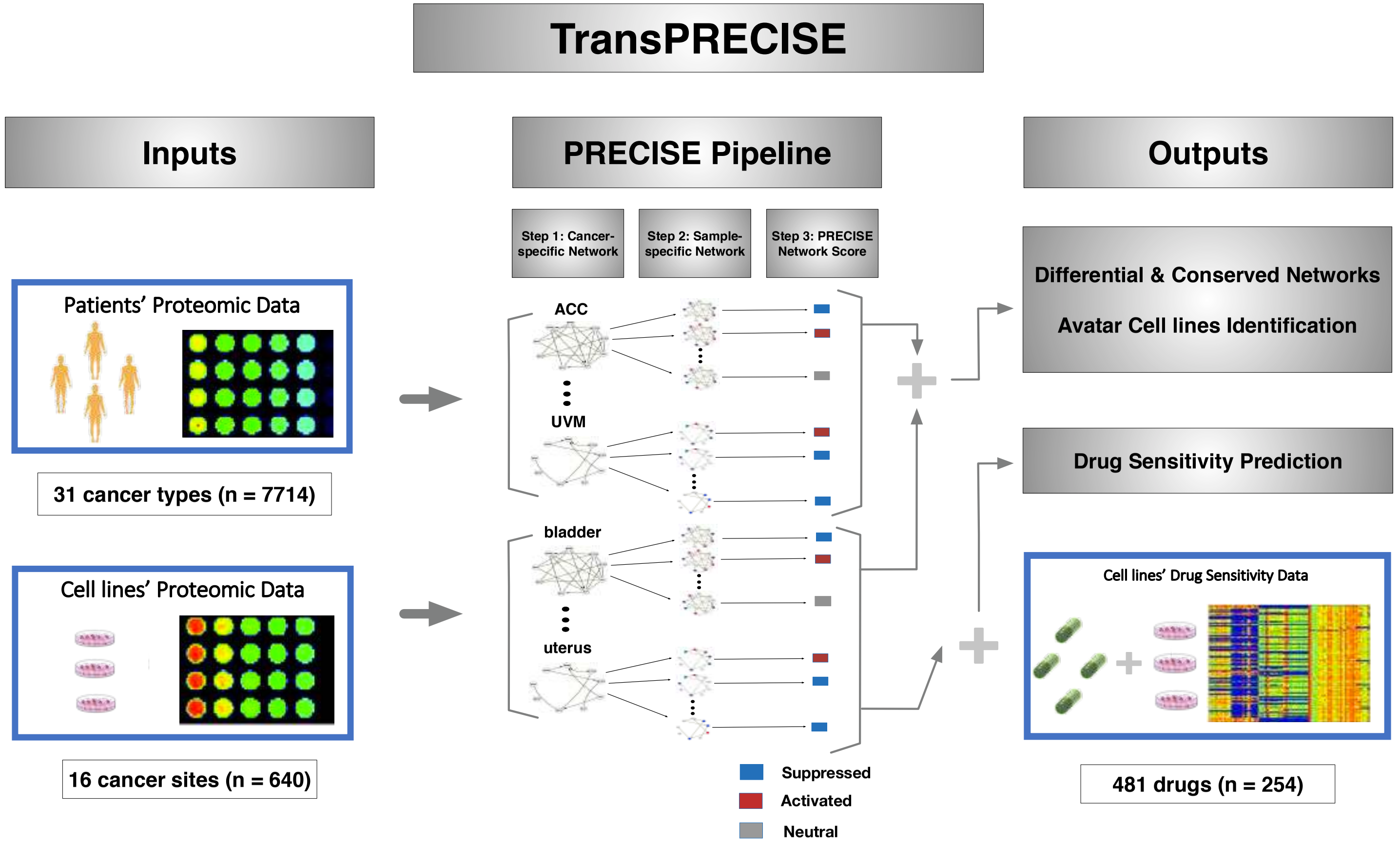

TransPRECISE
Inputs
PRECISE Pipeline
Outputs
Patients' Proteomic Data
31 cancer types (n = 7714)
Cell lines' Proteomic Data
16 cancer sites (n = 640)
Step 1: Cancer-specific Network
Step 2: Sample-specific Network
Step 3: PRECISE Network Score
ACC
UVM
bladder
uterus
Suppressed
Activated
Neutral
Differential & Conserved Networks
Avatar Cell lines Identification
Drug Sensitivity Prediction
Cell lines' Drug Sensitivity Data
481 drugs (n = 254)

Figure 2.1: Overview of the TransPRECISE framework. The first step involves running the PRECISE pipeline on two sets of RPPA protein expression data—namely, cancer patients and cancer cell lines. For each cancer-pathway combination, three consecutive steps are followed: (step 1) fitting cancer-specific protein networks using Bayesian graphical regression, (step 2) deconvolving these cancer-specific networks to fit sample-specific pathways networks, and (step 3) aggregating the sample-specific networks to obtain calibrated TransPRECISE scores and pathway activity status. The cancer-specific networks from step 1 are compared across patients and cell lines for each pathway towards a pan-cancer identification of differential and conserved pathway activities. The TransPRECISE scores from steps 2 and 3 are used to identify potential avatar cell lines and the lineages for patient tumors and to construct prediction models for drug sensitivity trained in in vivo drug sensitivity and used for in silico drug sensitivity prediction of patients' drug response.

the unit information prior in this case. I also set the prior for σ_i to be proportional to σ_i^{-1}. Then I performed a full enumeration of the joint model across all proteins as specified by the conditional distributions specified by Equation (2.1) using the Markov chain Monte Carlo (MCMC) algorithm. After performing all node-wise regressions following the above model, I select the median probability model to infer the posterior pathway network. Specifically, for each pair of proteins i and j, a cancer-specific *edge* between them is established if both β_{ij} and β_{ji} correspond to posterior inclusion probabilities (PIPs, defined as the sum of the posterior model probabilities for models where the covariate protein was included) > 0.5. Since a regression for protein i on protein j adjusts for all the other proteins present in the pathway, the connection is present only if the partial correlation between the expression profiles of the two proteins for samples from that cancer is high. Thus, biologically, an edge between two proteins in a pathway for a specific cancer simply means that the data has enough evidence for the two proteins to be associated based on their partial correlation after adjusting for the other proteins in the pathway.

Thus, learning the network structure put on the pathway based on the edges defined as above (with edge-specific weights determined by the posterior inclusion probabilities) allows us to learn the *wiring* of the pathway for the specific cancer type. I define *rewiring* of a pathway network as the changes in these sets of edges and weights between the same set of nodes i.e. proteins for a pathway from cancer to cancer and model system to model system. Biologically, rewiring of a pathway across different patient and cell line cancers essentially captures the changes in the network topology based on various stress conditions that are either static or dynamic.

Step 2: Construction of patient-specific networks via deconvolution I further obtain a cancer-specific network with sample-specific labels on the nodes (proteins). Specifically, the activation statuses of the nodes are evaluated by estimating the posterior predictive density for each protein for each sample. To determine the activation status of a protein i for a sample j (y_{ij}), I compute the posterior probabilities of the protein expression to lie in the δ-interval around 0 (p^0_{ij}), to be greater than δ (p^+_{ij}), or less than $-\delta$ (p^-_{ij}). Then, I decide whether a protein is *neutral, activated, or suppressed*, depending on the maximum of these three posterior probabilities. Thus, samples from the same cancer lineage may have different node labels as suppressed, neutral, or activated, while the structure of the networks stay same across all such samples. Using $\delta = 0.5$, I calculate TransPRECISE networks across all samples for each of the 31 patient cancers and 16 cell line cancers, and each of the 12 pathways.

Step 3: Calibrating patient-specific pathway scores and status To compute an aggregated pathway activity score for each sample, I derive summary measures from the TransPRECISE networks obtained from previous step, indicating the entire pathway as neutral, activated, or suppressed. Under the TransPRECISE networks, the number of nodes that are connected to protein i ($|\{j : i \leftrightarrow j\}|$) is denoted by C_i. For a given pathway with p genes, the pathway activity scores for a sample j are given by the following equation, where $\bullet$ is one of 0, +, or $-$.

$$\kappa^\bullet_j = \sum_{i=1}^{p} p^\bullet_{ij}(C_i + 1)/p. \tag{2.3}$$

Note that these sample-specific pathway scores are weighted averages of the posterior probabilities for the corresponding statuses of the proteins with the number of connected proteins informing the weights. Therefore, *hub proteins* in the pathway that exercise more control over the network through higher number of interacting proteins get *higher weights* towards determining the cumulative network score. For a given pathway and sample j, the TransPRECISE pathway status – which indicates whether the pathway is suppressed, neutral, or activated for the sample – is then decided by the maximum of the three TransPRECISE pathway scores ($\kappa_j^{\bullet}$s).

2.2.4 Post-processing of Cancer- and Sample-specific TransPRECISE Networks

Comparing network activity levels across model systems For a given cancer lineage with respect to a pathway, based on the fitted network wiring, I want to quantify the *signaling* in the pathway for the specific cancer. From a biological perspective, signaling is the indication that there are sufficient non-random interactions between the proteins in that pathway among the samples from the cancer of interest. I also define the term *cross-signaling* as the evidence that two cancers (among patients and cell lines i.e. across model systems) exhibit similar levels of signaling in the same pathway. For this purpose, I define a quantity termed the connectivity score.

Briefly, the connectivity score (CS) is the ratio of the observed number of edges in the cancer-specific network to the total number of possible edges in the pathway ($p(p-1)/2$, if the pathway has p proteins). The suggestion of significant interaction exhibited within a pathway for a cancer type is then quantified by finding a randomCS proportion based on the CS, a low value of which indicates higher nonrandom interactions in the pathway for that cancer type. Briefly, for each cancer type and pathway, I randomly select the same number of proteins as in that given pathway from the pool of all proteins across the 12 pathways. After constructing TransPRECISE networks from a total of 1000 such random permutations of the proteins, a 1000 iterations of the CS (called randomCS, coming from the underlying null distribution if there were no significant interaction within the pathway, in addition to the possible interactions already present in the global pool of proteins) for that cancer type and pathway are obtained. The randomCS proportion corresponding to this cancer type and pathway pair is then defined by the following equation.

$$p_{CS_{\text{Obs}}} = \sum_{s=1}^{1000} I(CS_s > CS_{\text{Obs}})/1000. \tag{2.4}$$

Here $I(\bullet)$ is an indicator function, CS_s is the randomCS value obtained from the s^{th} permutation and CS_{Obs} is the actual CS value obtained from the data. Now, suppose I have a patient cancer type P and a cell line cancer type C. For a pathway G fixed, I say that the triplet P-G-C is connected in terms of similar network activity, if I have both $p_{CS_{\text{P, G}}}$ and $p_{CS_{\text{C, G}}} < \epsilon$ for some small chosen value of ϵ. A higher value of ϵ would result in a higher number of connections and decreasing the cutoff would lead to more refined sets of edges that have strong suggestion towards conserved network activity.

Correlating patient and cell line tumors based on PRECISE scores For a given pathway, the *network aberration score* of a sample j is defined as $\kappa_j^+ + \kappa_j^-$, where the two terms are the respective activated and suppressed TransPRECISE scores, as defined in the previous subsection. Using the resulting TransPRECISE network aberration score matrix across the 12 pathways as the input data, I obtain a robust pan-cancer and pan-model systems stratification. The data matrix has 8354 rows, the first 7714 of them corresponding to one patient each and the next 640 corresponding to one cell line each, and 12 columns, each corresponding to a pathway. Based on the Euclidean distance of the score matrix, I apply hierarchical clustering using Ward's method (Ward Jr, 1963). To determine the number of clusters, I use the gap statistic (Tibshirani et al., 2001).

2.2.5 Drug Response Prediction using TransPRECISE Scores

Conversion of continuous drug sensitivity data to binary The drug sensitivity data (continuous, IC_{50}) were collected from the GDSC and the conversion to binary sensitive/resistant responses were executed following the methods used in the CCLE, as described below (Yang et al., 2012; Barretina et al., 2012; Haibe-Kains et al., 2013).

1. Extract the drug sensitivity measurements, either IC_{50} or AUC (I use IC_{50} for illustration throughout).

2. Sort increasing $\log(\text{IC}_{50})$ values of the cell lines to generate a waterfall distribution.

3. If the waterfall distribution is nonlinear (Pearson correlation coefficient to the linear fit ≤ 0.95), estimate the major inflection point of the $\log(\text{IC}_{50})$ curve as the point on the curve with the maximal distance to a line drawn between the start and end points of the distribution.

4. If the waterfall distribution appears linear (Pearson correlation coefficient to the linear fit > 0.95), then use the median IC_{50} instead.

5. Cell lines with lower IC_{50} values than this cut-off are defined as sensitive, and those with IC_{50} values higher than this are called resistant.

I additionally require at least five sensitive and resistant cell lines each after applying these criteria, to ensure the stability of the trained models.

Training models for cell lines' drug response Out of the 254 cell lines that have drug sensitivity information, eight cancer lineages can be obtained with at least 10 samples (Table S2.4). For each lineage-drug combination with at least 10 non-missing responses, I fit a Bayesian additive regression tree (BART) (Chipman et al., 2010) model using the package `bartMachine` (Kapelner and Bleich, 2013) with the predictors being the 12 pathway network aberration scores and the response variable being the binary (sensitive/resistant) drug response. I compute the area under the receiver operating characteristics curve (AUC) for each model with a five-fold cross-validation, and only keep those models for further inferences that had test-set AUC > 0.85. I obtain the ranking of predictors in each model using the variable inclusion proportions, defined as the number of times a covariate is selected in a model within the 1000 iterations after a 250-iteration burn-in.

Predicting patient drug responses (binary) For each of the eight cell line lineages with at least 10 samples in the drug sensitivity data, I use the corresponding models for predicting the drug sensitivity in patient tumors with the same tissue type. The predictions are in a continuous scale of $0-1$, and I mark a sample as "sensitive" if the predicted response is > 0.5. I define the *response rate* of a patient cancer with respect to a drug as the proportion of samples labeled sensitive within each for that cancer-drug combination and rank the drugs within each cancer type in decreasing order of these response rates.

Evaluating prediction performances I further evaluated the predictive performance of our TransPRECISE algorithm using the drug exposure data for TCGA patients obtained from the Gene-Drug Interactions for Survival in Cancer (GDISC) study (Spainhour and Qiu, 2016; Spainhour et al., 2017). GDISC integrates gene copy number data, drug exposure data, and patient survival data to infer gene-drug interactions impacting survival. In addition to the collection of all analyzed gene-drug interactions in 32 cancer types organized and presented in a searchable web-portal, GDISC provides the standardized drug exposure data, which is the resource used by us to quantify predictive capabilities of our pipeline (Spainhour and Qiu, 2016; Spainhour et al., 2017). I computed the proportion of correct predictions for 10 TCGA cancer types and 51 drugs that are common between our cell lines drug sensitivity data obtained from the GDSC and the GDISC drug exposure data. I used Bayesian additive regression tree (BART) models as before for this purpose. Briefly, for a specific cancer type and a drug with n patient samples, let us denote the true drug exposure vector from the GDISC as Y (binary), and the predicted response as described in the previous subsection as $\hat{Y}$ (continuous in [0, 1]). Then, the proportion of correct prediction is computed as $\sum_{i=1}^{n} I(Y_i = I(\hat{Y}_i > 0.5))/n$.

2.3 Pan-cancer Multi-system Proteomic Analyses

2.3.1 Differential and Conserved Rewiring and Circuitry of Cancer-specific Networks

Using the de-novo cancer-specific population-level networks (from Step 1 of TransPRECISE), I evaluated intra-pathway edge rewiring (Section 2.2.3) across lineages of the two model systems to identify highly conserved and differential edges, and to link patient and cell line tumor types by measuring intra-pathway circuitry.

Network rewiring across model systems I determined the extent to which protein-protein edges in each of the pathways were shared across tumor sites in the patients and the cell lines. I found highly conserved edges across lineages for both cell lines and patients (Figures 2.2, S2.1-S2.10). All of the 12 pathways had at least one link that was shared across more than 20 lineages among the patient cancer types, and 11 pathways (with the exception of hormone signaling) had at least one link that was shared across more than eight lineages among the cell line lineages. The conserved edges were further classified into three categories: (a) patients-cell lines, (b) patients only, and (c) cell-lines only. For category (a), I identified significant correlation of CCNE2-FOXM1 (10 cell line lineages, 17 patient cancer types) in a cell cycle, CTNNB1-SERPINE1 (eight cell line lineages, 17 patient cancer types) in EMT, and RB1-RPS6 (eight cell line lineages, 20 patient cancer types) in TSC/mTOR pathways, respectively.

Linking tumor types between model systems based on network circuitry I investigated the shared cross-signaling between cell line and patient tumor types. As a measure of the level of cross-signaling (Section 2.2.3) of a specific pathway network, I defined the connectivity score (CS) as the ratio of the observed number of edges in a given network to the total number of possible edges in the pathway, as more edges imply a higher level of cross-signaling within a pathway (Table S2.5). In addition, I quantified the level of significance

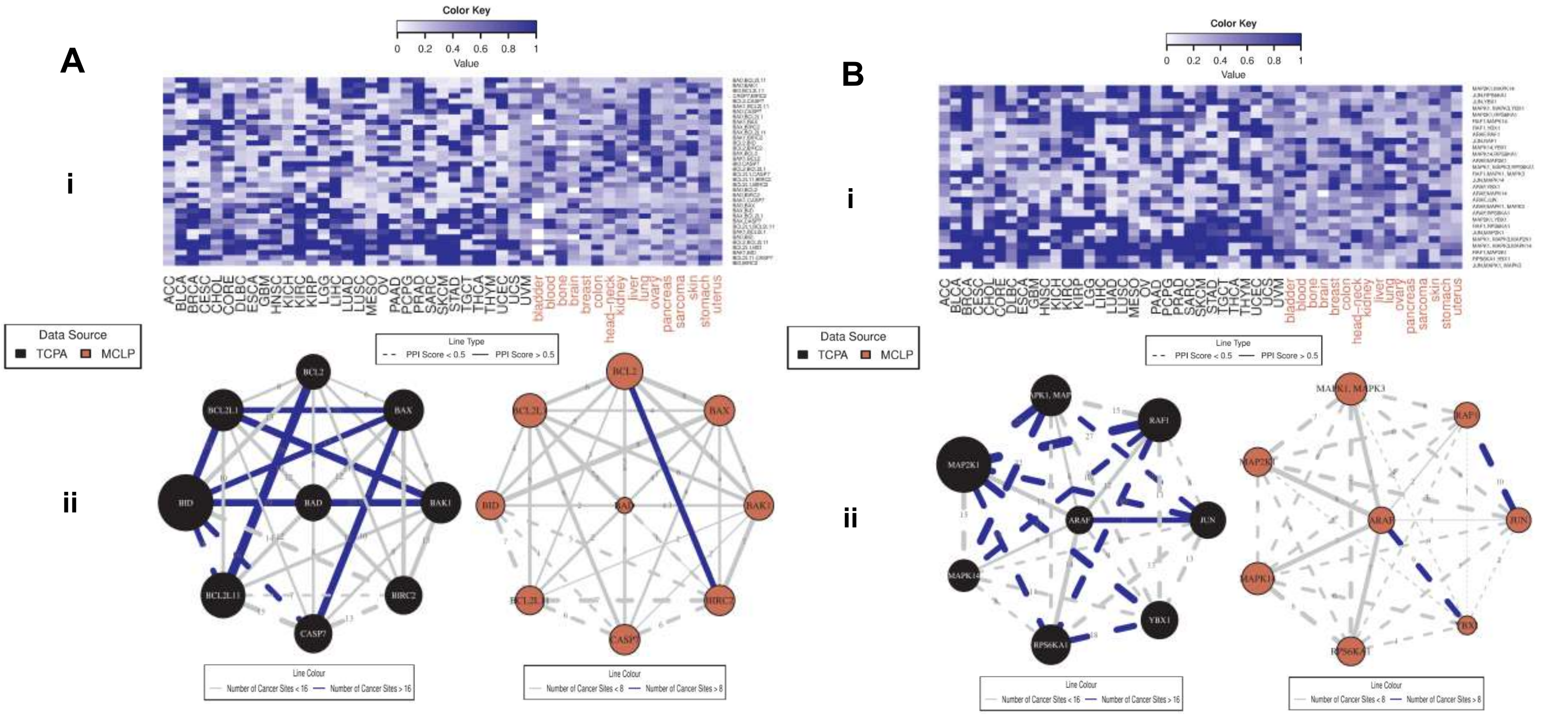


Figure 2.2: Pan-cancer summary of protein networks for (A) apoptosis and (B) RAS/MAPK pathways. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Networks depicting pan-cancer commonalities and differences in cancer-specific network structures: edges are weighted by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge with a posterior probability > 0.5, and labeled by solid lines if the edges are confirmed by the interaction scores from STRING database. The left and right panels are networks for patients and cell lines, respectively.

for the observed CS value by comparing it with CS values obtained from random permutation of the network, called randomCS; lower values of randomCS provide evidence against the observed CS value being obtained under random chance (Section 2.2.4). Based on the randomCS, I evaluated the similarity between cell line and patient tumor types in terms of network cross-signaling. Specifically, I declared two lineages were similar for a pathway if both of them showed high levels of cross-signaling (i.e., low randomCS proportions). Some key triplets of cell line-pathways-patient are summarized in Figure 2.3.

2.3.2 Pan-cancer Stratification across Model Systems Based on TransPRECISE Scores

I deconvolved the global population-level networks to obtain sample-specific pathway-level functional summaries of the proteomic crosstalk within a pathway—in other words, for a given pathway, each sample has three different scores for activated, neutral, and suppressed statuses of the pathway. For the tumor stratification, I used the *network aberration score* defined as the sum of the activated and suppressed TransPRECISE scores for each sample.

For linking cell lines and patients, I computed the Pearson's correlation for aberration score vectors (across twelve pathways) from each cell line-patient pair. Majority of the cell line-patient pairs for sarcoma-SARC (green), kidney-KIRC (light green), breast-BRCA (orange), and brain-LGG and GBM (light green and yellow) (edge colors in Figure 2.4 parenthesized) showed absolute correlations > 0.9. Interestingly, the pancreatic and brain cancers were highly correlated across model systems: 93% of GBM-pancreas pairs (besides 99% of pancreas-HNSC pairs and 92% of the PAAD-head&neck pairs) had absolute correlations > 0.9, and most of these connections appear to be driven by high aberration scores in the DNA damage response pathway (Figure 2.5).

To find robust pan-cancer stratification across model systems, I applied hierarchical clustering using the complete linkage method (Sorensen, 1948) on the correlations of the

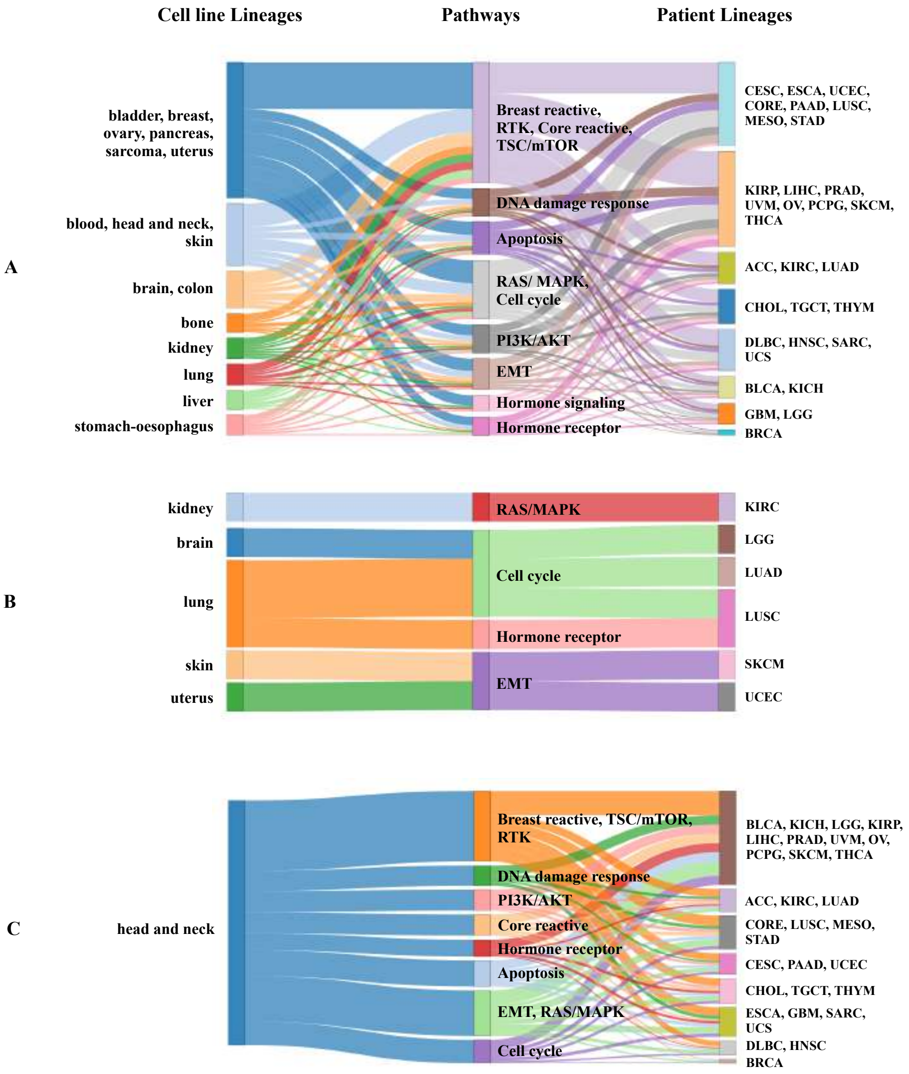


Figure 2.3: Sankey diagrams for patient and cell line cancers with conserved pathway-specific connectivity. A. The columns contain cell line cancers, pathways, and patient cancers from left to right, respectively. A cell line cancer tissue is connected to a pathway if the connectivity score (CS) for that cancer type-pathway pair (defined as the proportion of edges out of all possible undirected edges in the pathway that are held by that cancer type) is more than 900 out of 1000 randomCS values computed for that cancer type, with repeated random selection of the same number of proteins as in the pathway from the pool of all proteins across the 12 pathways. The connection between a patient cancer type to a pathway is also determined by the same rule. The length of the middle (pathway) column pieces indicate the participation of that pathway in driving the conservation across the two model systems. As seen in panel A, Ovary and uterus cell lines were connected via the hormone signaling (breast) pathway with BRCA; lung, kidney, and stomach-oesophagus cell lines were linked together with two clusters of patient cancers (KICH, KIRP, PRAD, LGG and LUSC, UCEC, STAD) via the RTK pathway. B. The sankey diagram contains only the subset of cell line cancer (i.e., patient cancer pairs that have same tissue-specific lineage), and the cutoff for CS values is higher than 800 of the 1000 randomCSs obtained using the random selection of proteins. Panel B presents clear confirmations of conservation of activities across model systems within cancer tissues, some specific examples being bladder-core reactive-BLCA, kidney-RTK-KICH & KIRP, kidney-hormone receptor-KIRC, ovary-hormone signaling-OV, and stomach-hormone receptor-ESCA & STAD. C. The sankey diagram contains only the subset of the edges that are originating from the head and neck cancer cell line type, and the cutoff for CS values is higher than 800 of the 1000 randomCSs obtained using the random selection of proteins.

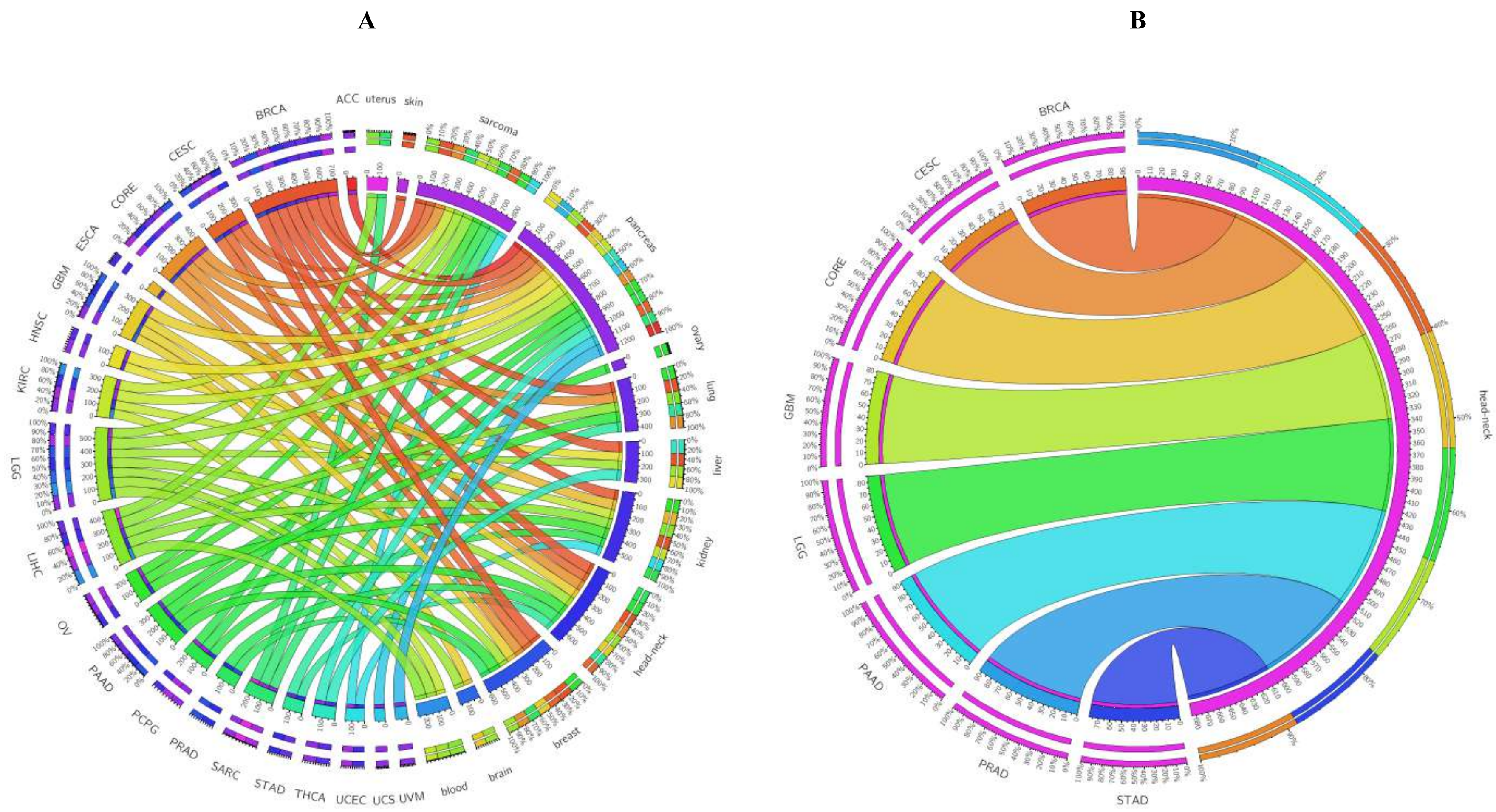


Figure 2.4: Circos plots summarizing high correlations of network aberration scores between patient and cell line cancers. A. An edge exists between a patient cancer type and a cell line cancer lineage if more than 75% of all possible patient-cell line pairs for that pair of cancers have a Pearson correlation of magnitude 0.9 or higher between their sets of the 12 pathway network aberration scores (sum of TransPRECISE sample-specific pathway activation and suppression scores). The edge strengths are determined by these percentages, as well. The edge colors indicate the patient cancers from which the edge originates, and the lengths of the innermost node pieces indicate the neighborhood size of the corresponding node. The two circular axes in the exterior indicate relative strengths of the edges originating from the same node, and the sections are colored by the opposite node to which that edge is connected, with the edges now arranged according to decreasing order of strength. B. This panel contains the subset of the plot in Panel A with only the connections originating from the head and neck cell line type visible.

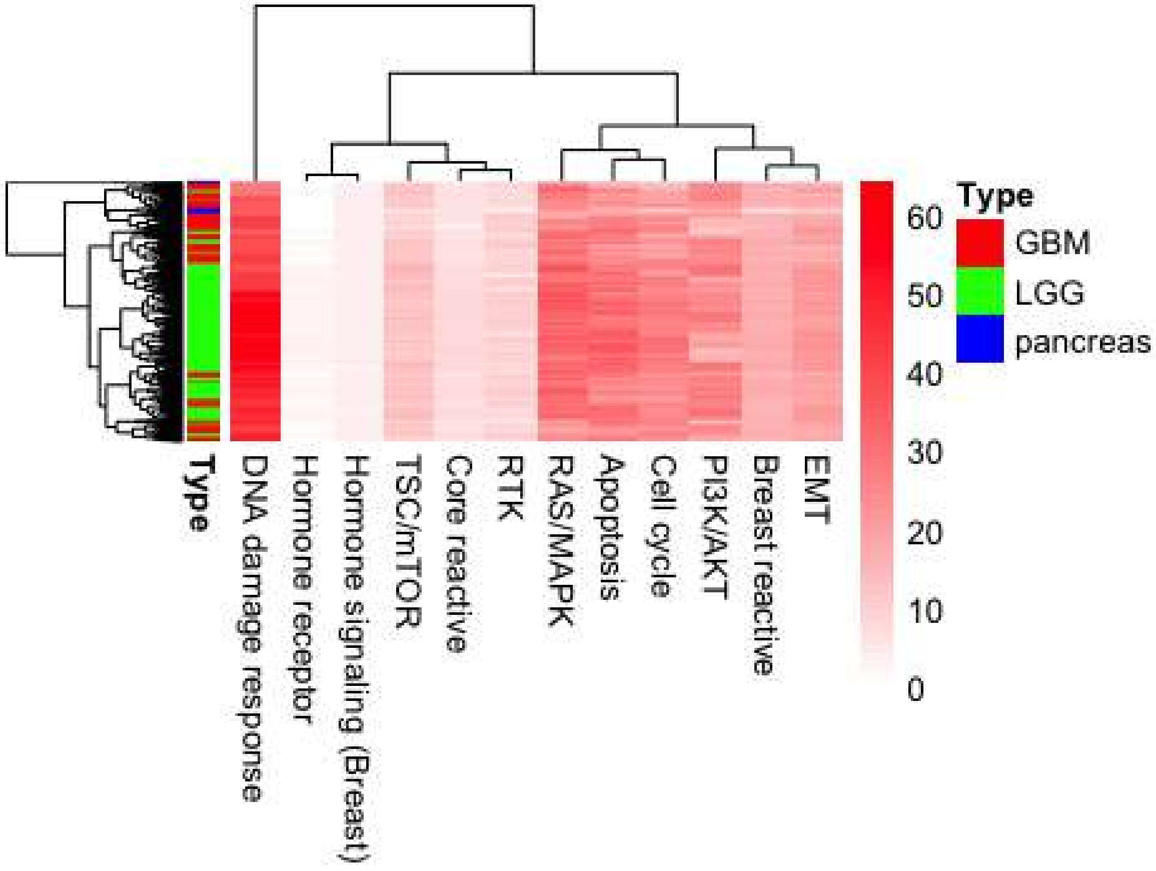


Figure 2.5: Heatmap depicting network aberration scores (combined activation and suppression TransPRECISE pathway scores) for the GBM and LGG cancer patients and pancreas cell lines across 12 proteomic signaling pathways. The leftmost annotation bar indicates sample types.

aberration scores. Among the 29 optimal clusters across patients and cell lines (Figure 2.6 and Table S2.6), most of the cell lines have mixed membership with patient tumors in eight clusters (C2, C3, C4, C9, C13, C14, C19, and C23), while cluster C29 includes only cell lines (48 out of 640 in total, 7.5%). The cluster C4 showed a high level of fidelity in lineages between cell line and patient tumor types; it includes 81% of ovary cell lines and 11% of OV patients, 72% of head&neck cell lines and 38% of HNSC patients, and 20% of pancreas cell lines (another 70% of them being located in C2 with notable aberration of RAS/MAPK pathway), and 80% of PAAD patients exhibiting high aberration in apoptosis and DNA damage response pathways (Table S2.7). Within cluster C4, I observed significant correlations between the patient-cell line samples from ovary-PAAD, OV, BLCA, skin-PAAD, and head&neck-BLCA, HNSC (Figure 2.7). More specifically, the HNSC samples were almost exclusively divided into the two clusters, C4 (n = 78, 38%) and C15 (n = 122, 60%), that include 38 (73%) head&neck cell lines and 5 (100%) oesophagus cell lines, respectively (Table S2.6). The co-occurrence of squamous cell carcinoma of the head&neck and esophageal cancer is not uncommon (McGuirt et al., 1982; Jain et al., 2013).

2.3.3 Characterization of head&neck Cancer Cell Lines and Patients

I focused on a case study using only the head&neck cell lines in conjunction with all the patient samples from TCGA. As presented in Figure 2.3C, I observed connections from the head&neck cell lines to the patient cancers across the pathways at a threshold of randomCS proportion < 0.2. One significant observation is that the head&neck cell lines are connected to the HNSC samples via several pathways including RTK, apoptosis, cell cycle and EMT. Notably, the set of patient cancers, for which at least 75% of the sample-sample pairs with the head&neck cell lines have highly correlated network aberration scores across all pathways, includes the BRCA, CORE, LGG and GBM samples but does not include the

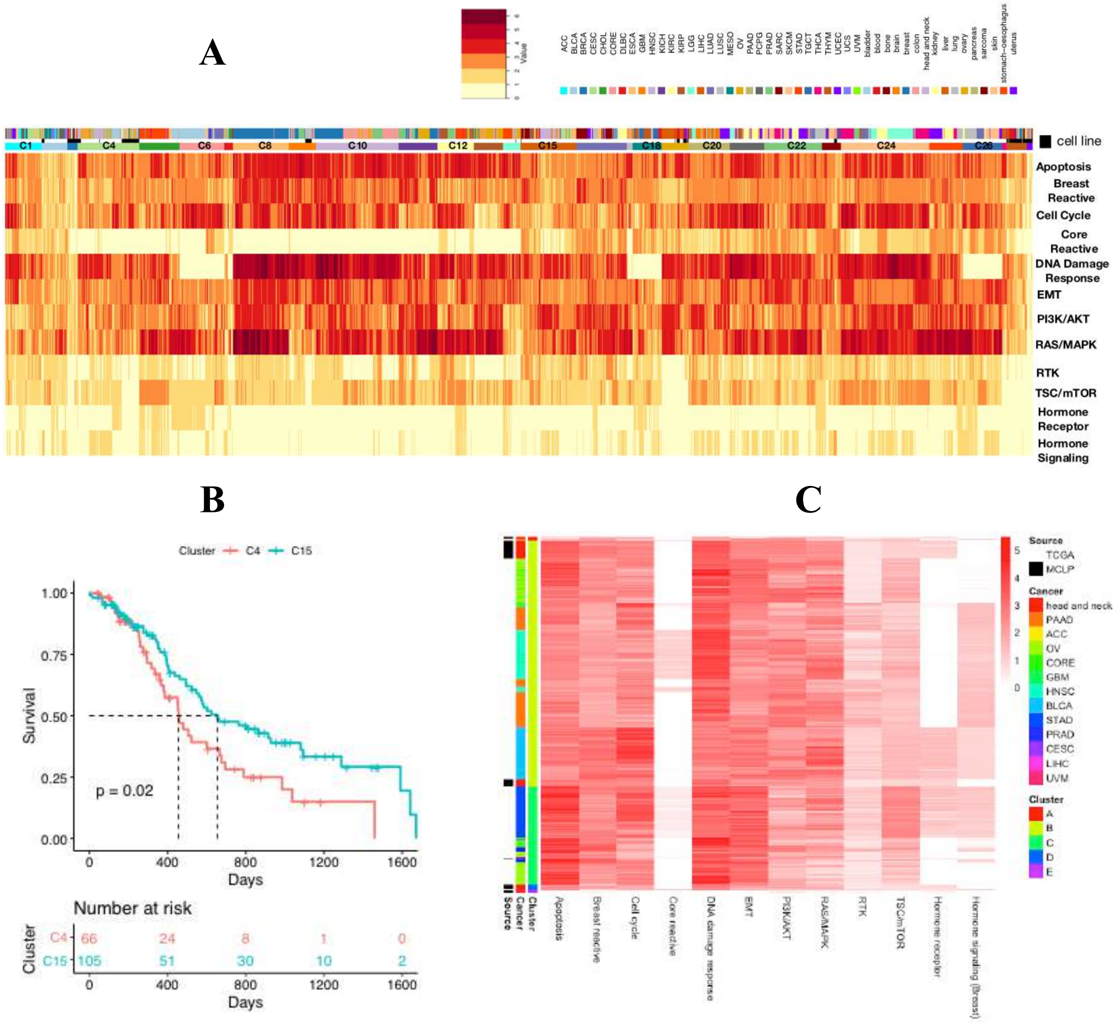


Figure 2.6: Avatar cell lines identification and selection of driving pathways using network aberration scores. A. Heatmap depicting network aberration scores (combined activation and suppression TransPRECISE pathway scores) after running unsupervised hierarchical clustering of the score matrix consisting of 8354 samples (7714 patients across 31 cancer lineages and 640 cell lines across 16 cancer types) and 12 proteomic signaling pathways. 29 clusters are identified by gap statistic. Out of the three annotation bars, the topmost one indicates tumor types, the middle one indicates whether the sample is a patient or a cell line, and the bottom one indicates cluster participation according to which the samples are grouped. B. Kaplan-Meier curves depicting difference between survival times of HNSC patients that are clustered in clusters C4 and C15 using the hierarchical clustering method on TransPRECISE network aberration scores. C. Heatmap depicting network aberration scores (combined activation and suppression TransPRECISE pathway scores) after running unsupervised hierarchical clustering of the score matrix consisting of all patient samples and only the head and neck cell line samples across the 12 pathways. Out of the three annotation bars, the leftmost one indicates whether the sample is a patient or a cell line, the middle one indicates the cancer type, and the rightmost one indicates cluster participation according to which the samples are grouped.

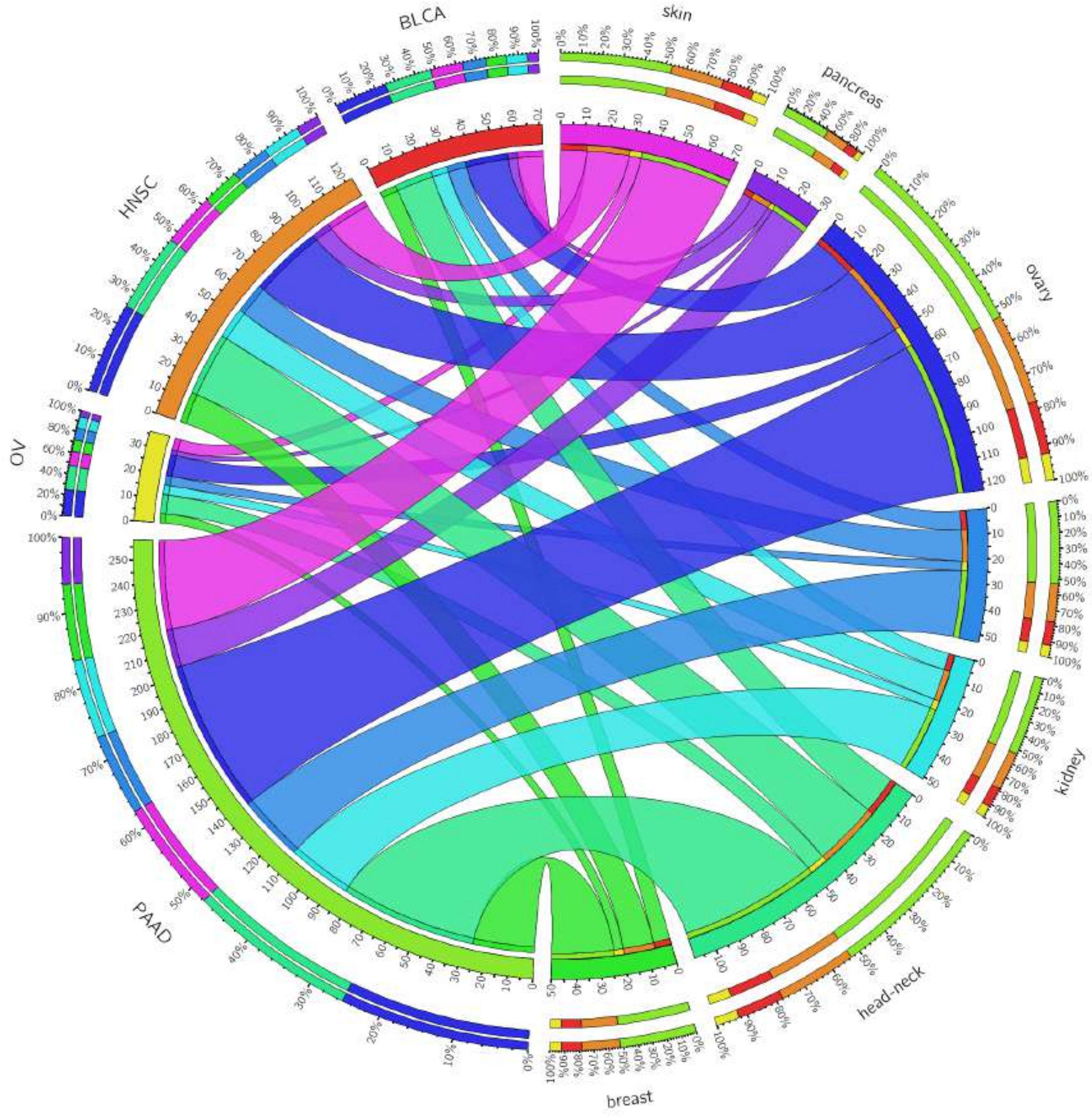


Figure 2.7: Circos plot summarizing joint membership of patient and cell line samples in cluster C4. An edge exists between a patient cancer type and a cell line cancer lineage if more than 10% of total number of samples for both of the lineages are located in cluster C4. The edge strengths are determined by the product of the two percentages for the two nodes (lineages) scaled by 100. The edge colors pertain to the cell line cancers that the edge originates from, and the lengths of the innermost node pieces indicate the neighborhood size of the corresponding node. The two circular axes in the exterior indicate relative strengths of the edges originating from the same node, and the pieces here are colored by the opposite node to which that edge is connected, with the edges now arranged according to decreasing order of strength.

HNSC samples, which is in line with the findings from Figure 3C since those connections were stronger than the connection with HNSC (Figure 2.4B). In hierarchical clustering of the head&neck cell lines and all the patient samples, a subset of the head&neck cell lines cluster with a subset of the HNSC patients with high aberration in the DNA damage response pathway. In the hierarchical clustering based on all patients and cell lines, I found a significant difference in the survival outcome between HNSC patients in C4 and C15: the median survival was 456 days and 654 days for C4 and C15, respectively, with a p-value of 0.02 (Figure 2.6B). The patients in C15 that were represented by esophagus cell lines showed better survival than those in C4, which includes head&neck cell lines – this indicates that our TransPRECISE scores captured distinct prognostic information in HNSC patients. Moreover, the patterns of pathway activity and status were significantly different between the two clusters. The HNSC patients in both C4 and C15 had high aberration scores in apoptosis, PI3K/AKT and DNA damage response pathways. Specifically, for the DNA damage response pathway, the two clusters exhibited significantly distinct TransPRECISE statuses; 72% patients in C4 showed suppression and 65% patients in C15 showed activation (Chi-squared test p-value < 0.0001).

2.3.4 Drug Response Prediction using TransPRECISE Scores

Training drug response prediction models in cell lines For the subset of cell lines where drug sensitivity data are available (Table S2.4), I used Bayesian additive regression trees (BART) (Chipman et al., 2010), a machine learning method, to build predictive models from the network aberration scores for the 12 pathways. For each cancer, I fit BART, with drug response (sensitive or resistant) as a binary outcome and TransPRECISE scores as predictors, for the drugs having profiles of ≥ 10 cell lines for that cancer type.

I found that TransPRECISE scores conferred high predictive power, translating to high median test-set areas under the receiver operating characteristic curves (AUCs) across the lineages; all lineages had median AUCs > 0.8, with lung, breast, and colon being the top three, having median AUCs > 0.9 (Figure 2.8). From the radar plot summarizing the top pathway predictors across all drugs for each lineage (Figure 2.9A), I observed some notable evidence of predictive affinity for certain pathways to specific lineages: hormone receptor in breast, core reactive, RTK and TSC/mTOR in colon, RAS/MAPK in liver, DNA damage response and PI3K/AKT in lung, apoptosis, cell cycle and EMT in ovary, and DNA damage response and TSC/mTOR in pancreas cell lines. Further, I investigated pathway interaction in predicting drug sensitivity (Figure 2.9B, S2.11). The breast cancer related pathways, breast reactive, and hormone receptor pathways were highly synergistic in predicting the responses of five drugs including ML311 in breast cancer cell lines (Bashari et al., 2016).

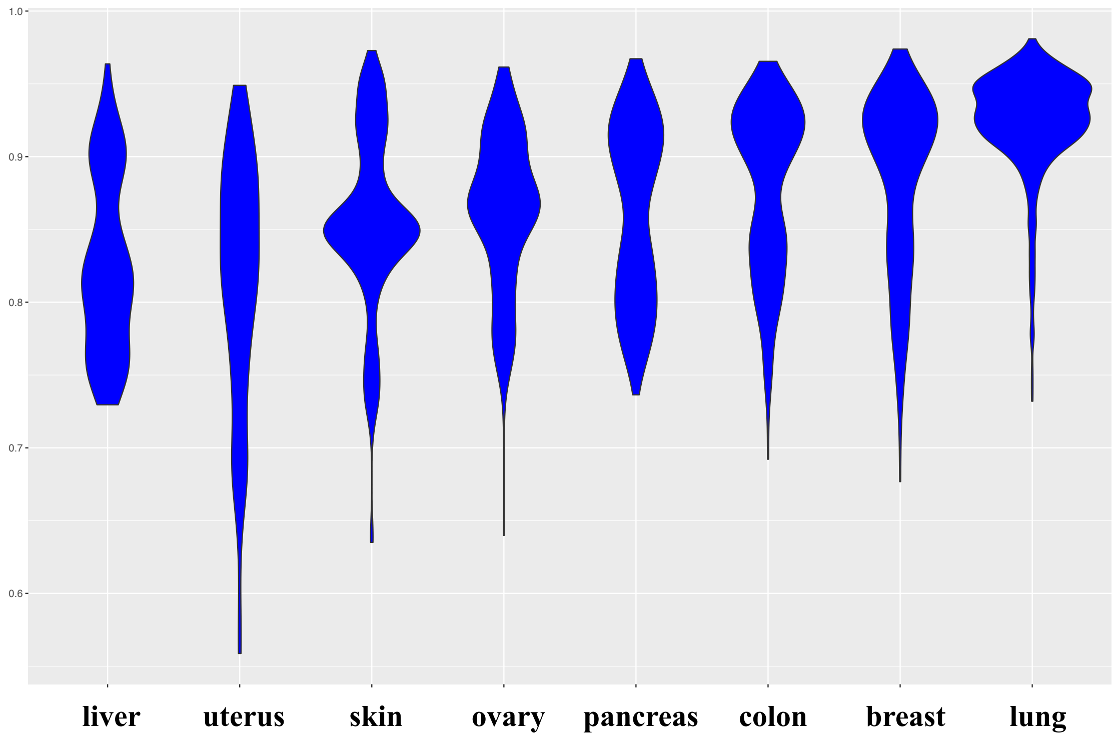


Figure 2.8: Performance of drug sensitivity prediction models across cell line lineages. Each column exhibits a violin plot for AUCs from all the fitted BART models for the corresponding cancer type for each drug with at least 10 responses available for the samples in that lineage.

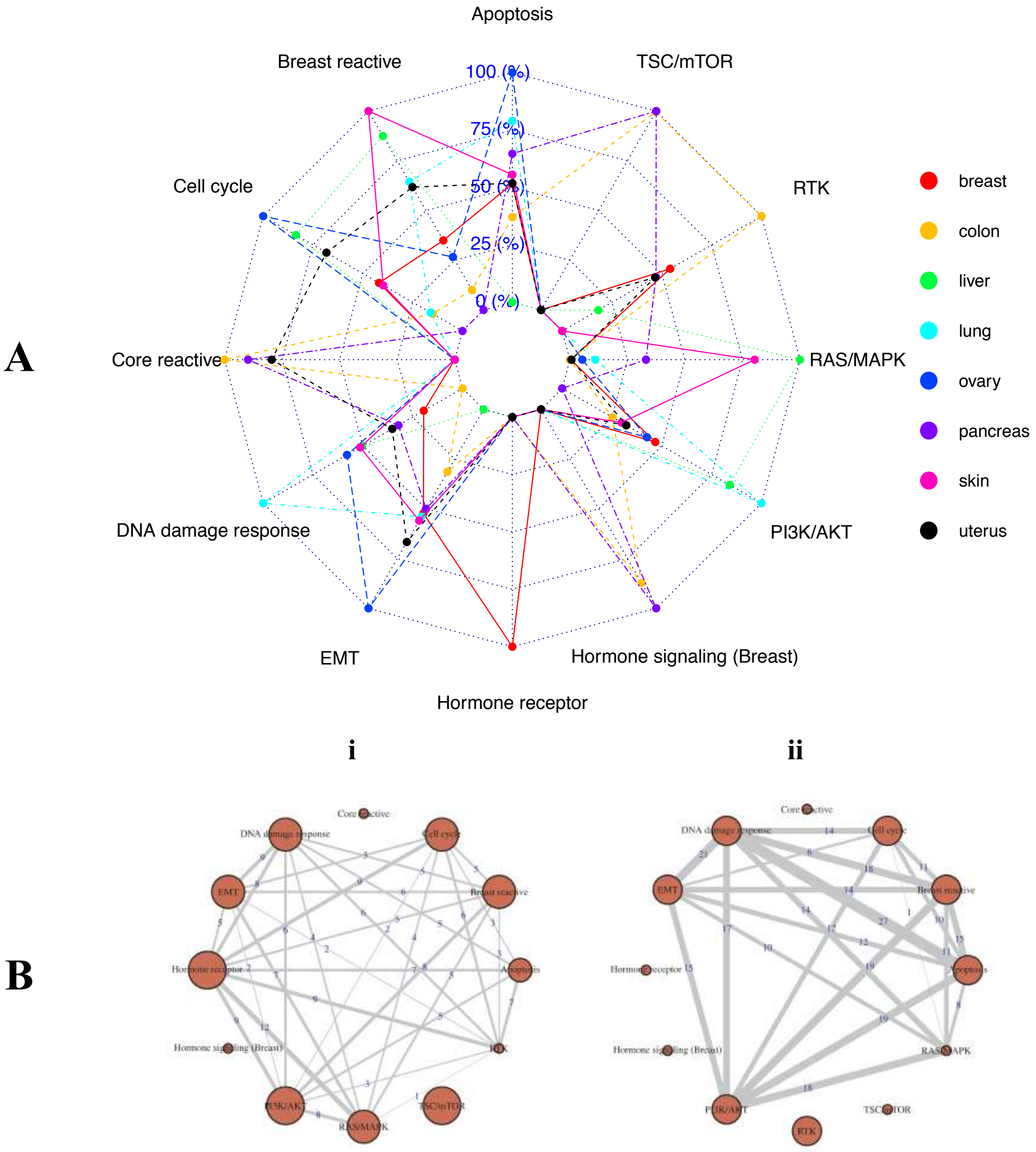


Figure 2.9: Performance of pathways in drug response prediction for cell lines across cancer lineages, based on test-set AUC values evaluated from five-fold cross validation. A. For a tissue type, I only look at the subset of drugs for which I have at least 10 response profiles from cell lines in that lineage and at least 0.85 test-set AUC using a five-fold cross-validation in the BART models. Then, for each pathway I compute the proportion of times it is the top predictor in models for such drugs. The radar plot shows these proportions in a ln(1 + •)-transformed scale. The significance and ranking of each of the twelve pathways in a model are quantified by posterior probabilities of inclusion in such a final predictive model for drugs. B. Networks showing the number of times (within models satisfying the criteria in panel A) a pair of pathways are the top two predictive pathways in a BART model. Panel i is for the breast cancer cell lines and panel ii is for the lung cancer cell lines.

Predicting drug sensitivity in patient tumors For each of the cell line cancer lineage, for which the training models were fitted with the TransPRECISE pathway scores (as above), I predict drug sensitivity in patient tumors within matched tissue type (total 10 lineages). I found drugs that had 100% response rate especially in BRCA, CORE, LIHC, PAAD and SKCM; some of which are under clinical investigations in their respective cancers (Table S2.8). For example, all BRCA patients were predicted to be responsive to Ibrutinib that targets Bruton tyrosine kinase (BTK) with RAS/MAPK, PI3K/AKT, EMT as the top predictive pathways (Table S2.8). Using patient drug exposure data from GDISC, I evaluated the models' predictive performances, following procedures described in Section 2.2.5. For all the CORE patients our model trained on the colon cell lines for the drug lapatinib predicts the true exposure correctly (note the same drug-cancer drug combination was also predicted to have a 100% response; Table S2.8). Further, for $>$ 90% of the OV patients our model fitted on the ovary cell lines managed to correctly predict the response to the drug paclitaxel, which, by very current standards, remains an integral part of the chemotherapeutic treatment of ovarian cancer (Boyd and Muggia, 2018; Kampan et al., 2015; Kumar et al., 2010).

2.4 Discussion and Future Work

Overview The investigation of patient tumors and cell-line interactome offers insights into the translational potential of preclinical model systems. This requires development of analytical models that capture the molecular heterogeneity of a cancer type in an unbiased manner and accurate calibration of aberrant biological pathways. I propose TransPRECISE, a multi-scale Bayesian network modeling framework, whose overarching goals are three-fold: identify differential and conserved intra-pathway activities between two different model systems (patient tumors and cell lines) across multiple cancers; globally assess cell lines as representative in vitro models for patients based on their inferred pathway circuitry; and build drug sensitivity prediction models for both cell lines and patients to aid pathway-based personalized medical decision-making. To the best of our knowledge, TransPRECISE is the first computational approach that provides a conflation of these goals.

Application to multi-system pharmacoproteomic datasets As a proof-of-concept study, I illustrate the utility of TransPRECISE using RPPA-based proteomic expression profiles from patients and cell lines across several functional pathways, and cell lines' drug response. The protein interactions that were present in both model systems offer valuable insights into the shared pathway circuitry across model systems, which has potential translational utility to study the role of tumor microenvironment. For example, the robust link CCNE2-FOXM1 within cell cycle pathway has been identified to have important implications in modulations of several cancers, such as breast (Zanin et al., 2019), prostate cancer subtype 1 (Ketola et al., 2017), hepatocellular carcinoma (Zhang et al., 2019b), and osteosarcoma (Grant et al., 2013). The aberration of the highly shared edge CTNNB1-SERPINE1 in EMT pathway has been found to effect the growth of malignant cell masses in several cancers, including cancers of the gastric system (Tanabe et al., 2016; Xu et al., 2019), pancreatic cancer (Wu et al., 2019), and breast cancer (Asiedu et al., 2011). I also found high degree of fidelity to their histological sites between model systems based on the level of network cross-signaling, e.g., RTK pathway in kidney cancers (Patel et al., 2006), and the hormone signaling pathway in ovarian cancers (Hao et al., 2019; Zhang et al., 2017). As further validation, TransPRECISE implicated cross-signaling in EMT pathway in SKCM and UCEC, which are expected since the SKCM cohort contains many metastatic samples (Akbani et al., 2015) and UCEC includes epithelial-like endometrioid samples as well as mesenchymal-like serous samples (Levine, 2013). TransPRECISE implicated the hormone receptor pathway in lung cancer, which is another known observation that is being studied for its translational potential (Chen et al., 2017b). Our sample-specific inference of pathway activity provided robust tumor stratification across model systems that includes distinct prognostic information (Figure 2.6). These robust edges and cross-signaling of pathways across model systems and cancer sites will potentially provide complementary information in terms of disease characterization and therapeutic targets. Our Bayesian prediction models using the pathway scores on cell line's drug sensitivity provided high prediction accuracies (median test-set AUC $>$ 0.8 across all drugs and all cancers) and selected cancer-specific pathway signatures in predicting drug response, such as hormone receptor-breast (Lumachi et al., 2013), and TSC/mTOR-pancreas (Ayuk and Abrahamse, 2019; Iriana et al., 2016). Our training models using cell lines were used to predict patients' drug response and validated with their known sensitivities. For example, ibrutinib, which had high predicted sensitivity for all the BRCA samples, has been investigated for its impact on HER2-amplified breast cancers (Chen et al., 2016). Similarly, lapatinib, in combination with trastuzumab, has recently been tested clinically for HER2-amplified metastatic colorectal cancer (Sartore-Bianchi et al., 2016).

General applicability of TransPRECISE to state-of-the-art proteomic datasets One of the key strengths of the TransPRECISE algorithm is its generalizability, as it can be applied to any disease system that has matched genomic or molecular data on model and primary samples. For example, the transition from RPPA to other advanced high-throughput platforms and development of databases, such as CPTAC (Ellis et al., 2013), opens up the opportunity to include more proteins (thus, more pathways) in the network analyses: leading to a more global coverage of the proteomic crosstalk between model systems. Further, the PRECISE (Ha et al., 2018) pipeline, which lies at the core of the TransPRECISE analyses, allows integration of upstream regulatory information and multi-omics layers such as mutations, copy number, methylation and mRNA expression. These modalities can be leveraged for better and holistic rewiring of pathway circuitry. Finally, our framework can be, in principle, applied to emerging model systems, such as patient-derived xenografts (Lai et al., 2017; Siolas and Hannon, 2013), and organoids (Drost and Clevers, 2018), that allow better recapitulation of the human tumor microenvironment. In summary, TransPRECISE offers the potential to bridge the gap between human and pre-clinical models to delineate actionable cancer-pathway-drug interactions to assist personalized systems biomedicine approaches in the clinic.

Data and material availability I have created an online, publicly available R shiny app (available at `https://bayesrx.shinyapps.io/TransPRECISE/`) that is a comprehensive database and visualization repository of our findings. All codes used in generating our results are available, along with the documentation, on `https://github.com/bayesrx/TransPRECISE`.

2.5 Supplementary Figures

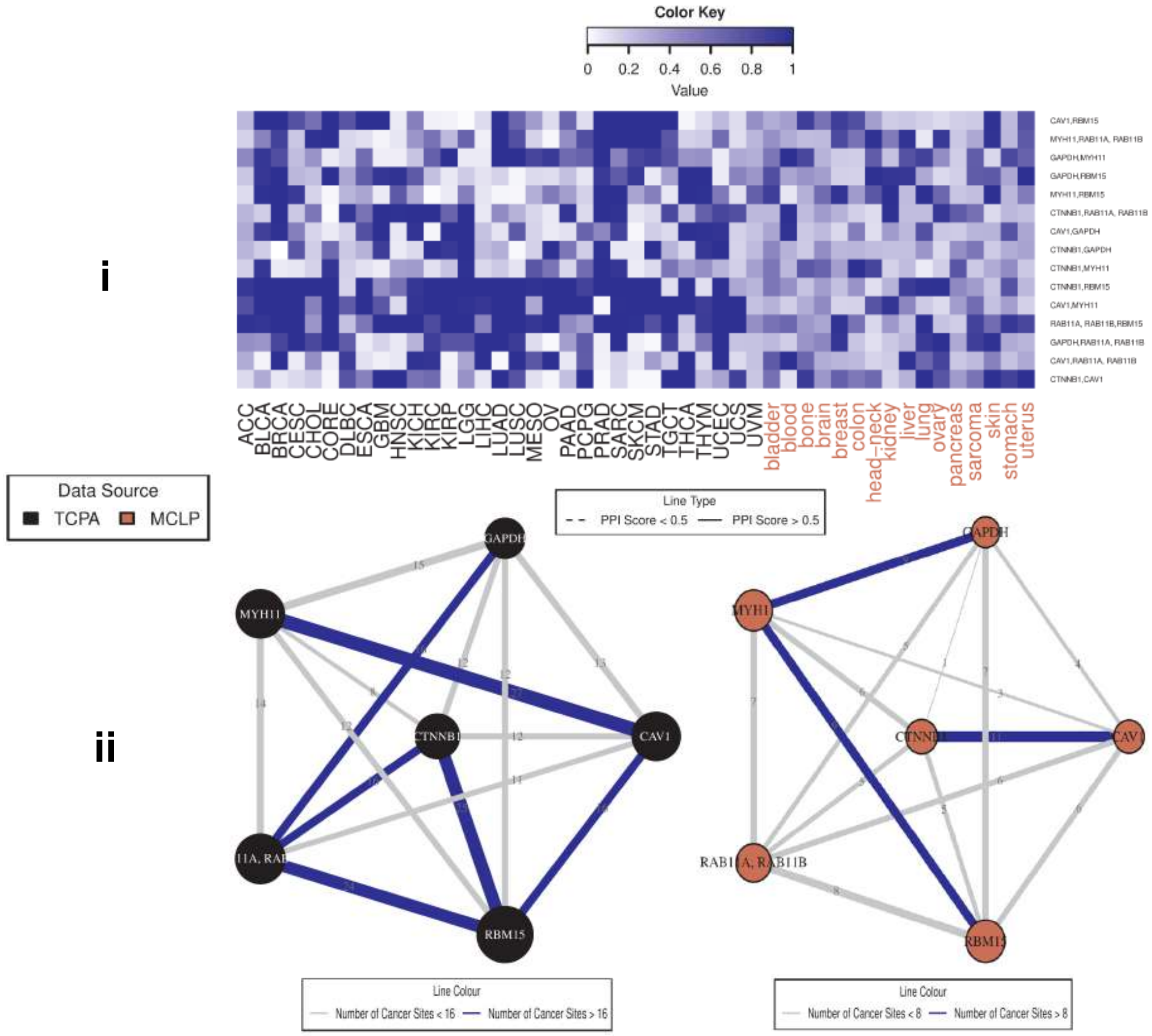


Figure S2.1: Pan-cancer summary of protein networks for breast reactive pathway. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Left panel exhibits a network with its edges weighted and labeled by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge, also presenting the a priori known strength of the edge using the protein-protein interaction score from the STRING database. The right panel is the corresponding network across cell line cancers.

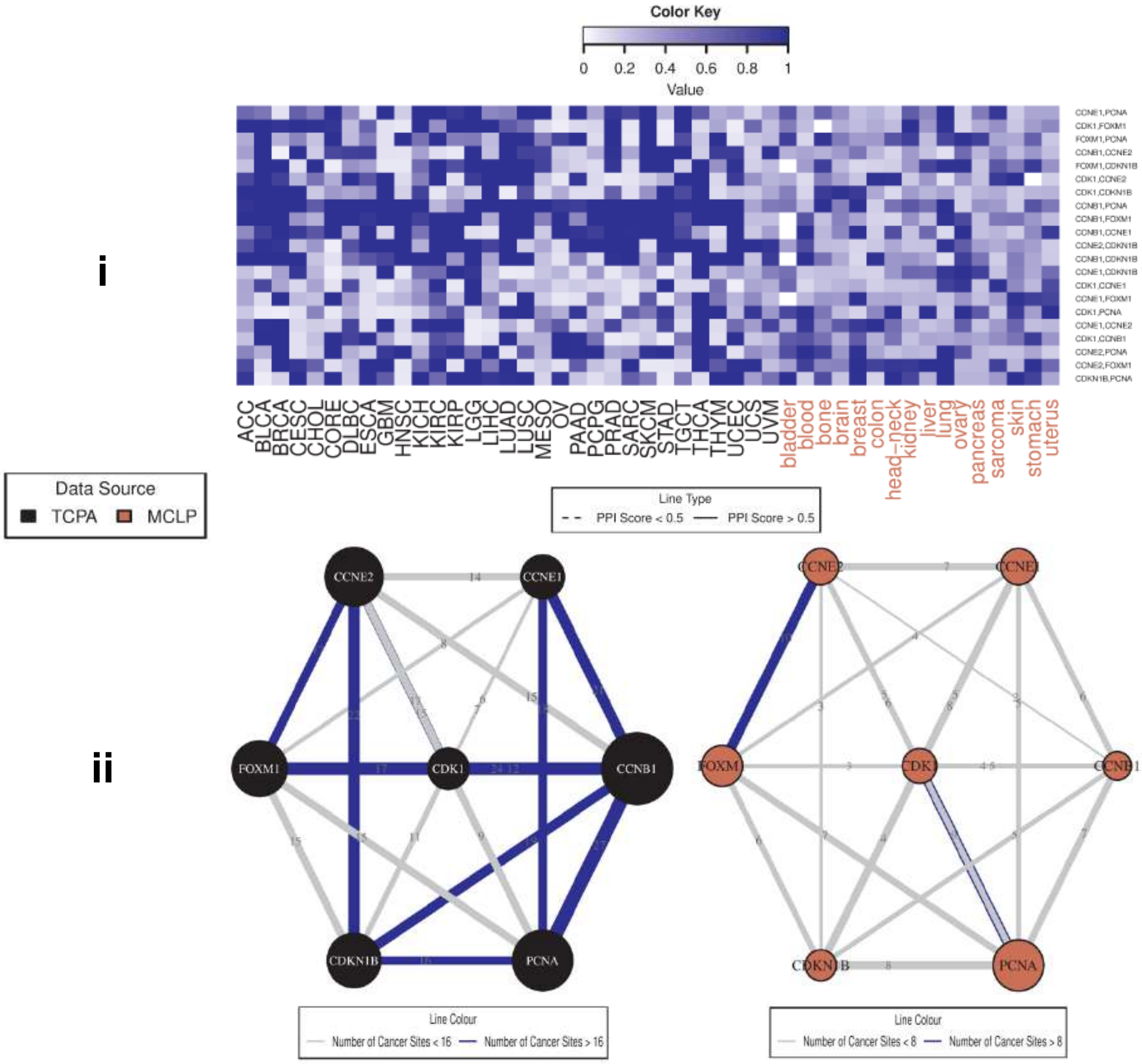


Figure S2.2: Pan-cancer summary of protein networks for cell cycle pathway. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Left panel exhibits a network with its edges weighted and labeled by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge, also presenting the a priori known strength of the edge using the protein-protein interaction score from the STRING database. The right panel is the corresponding network across cell line cancers.

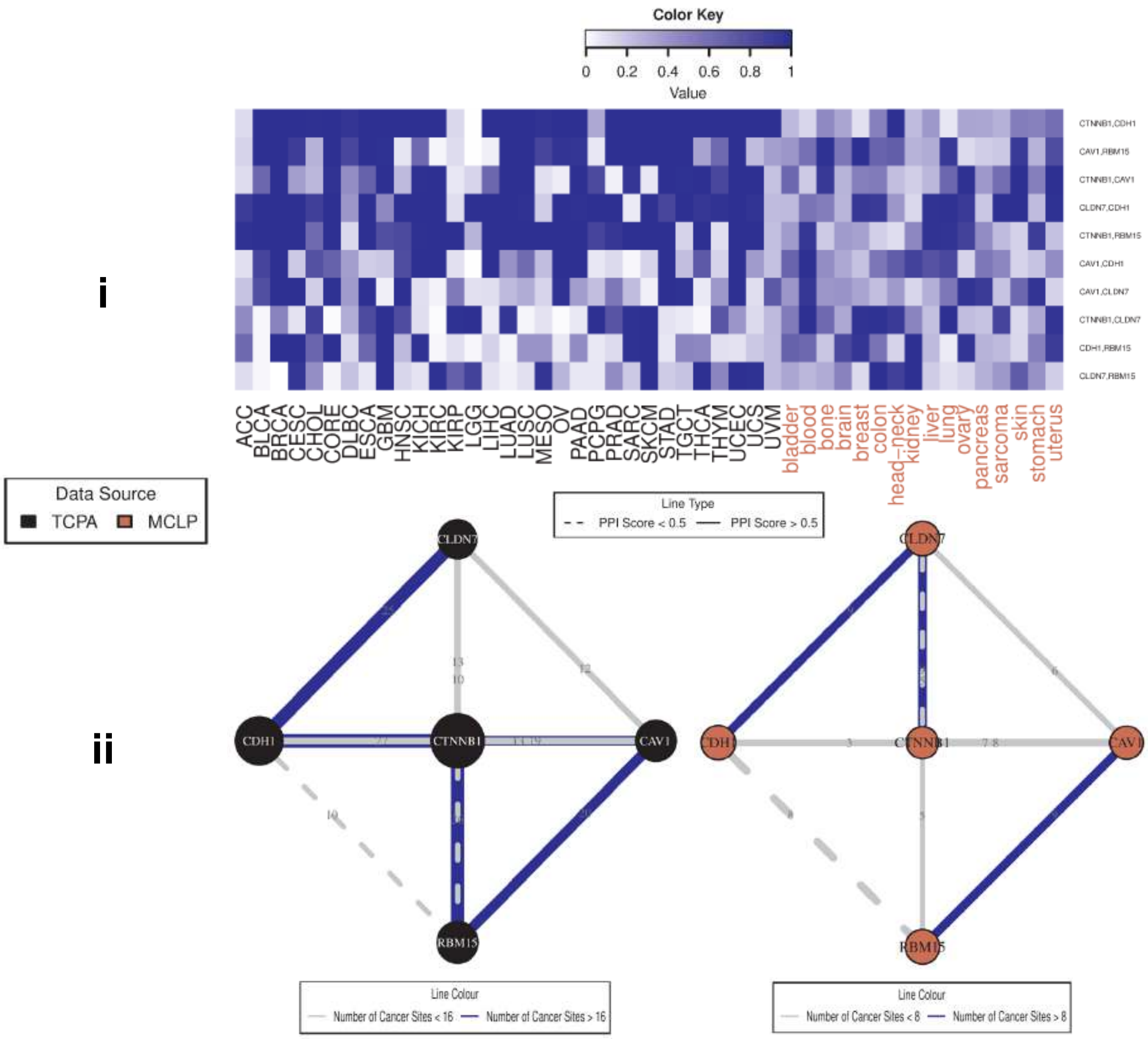


Figure S2.3: Pan-cancer summary of protein networks for core reactive pathway. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Left panel exhibits a network with its edges weighted and labeled by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge, also presenting the a priori known strength of the edge using the protein-protein interaction score from the STRING database. The right panel is the corresponding network across cell line cancers.

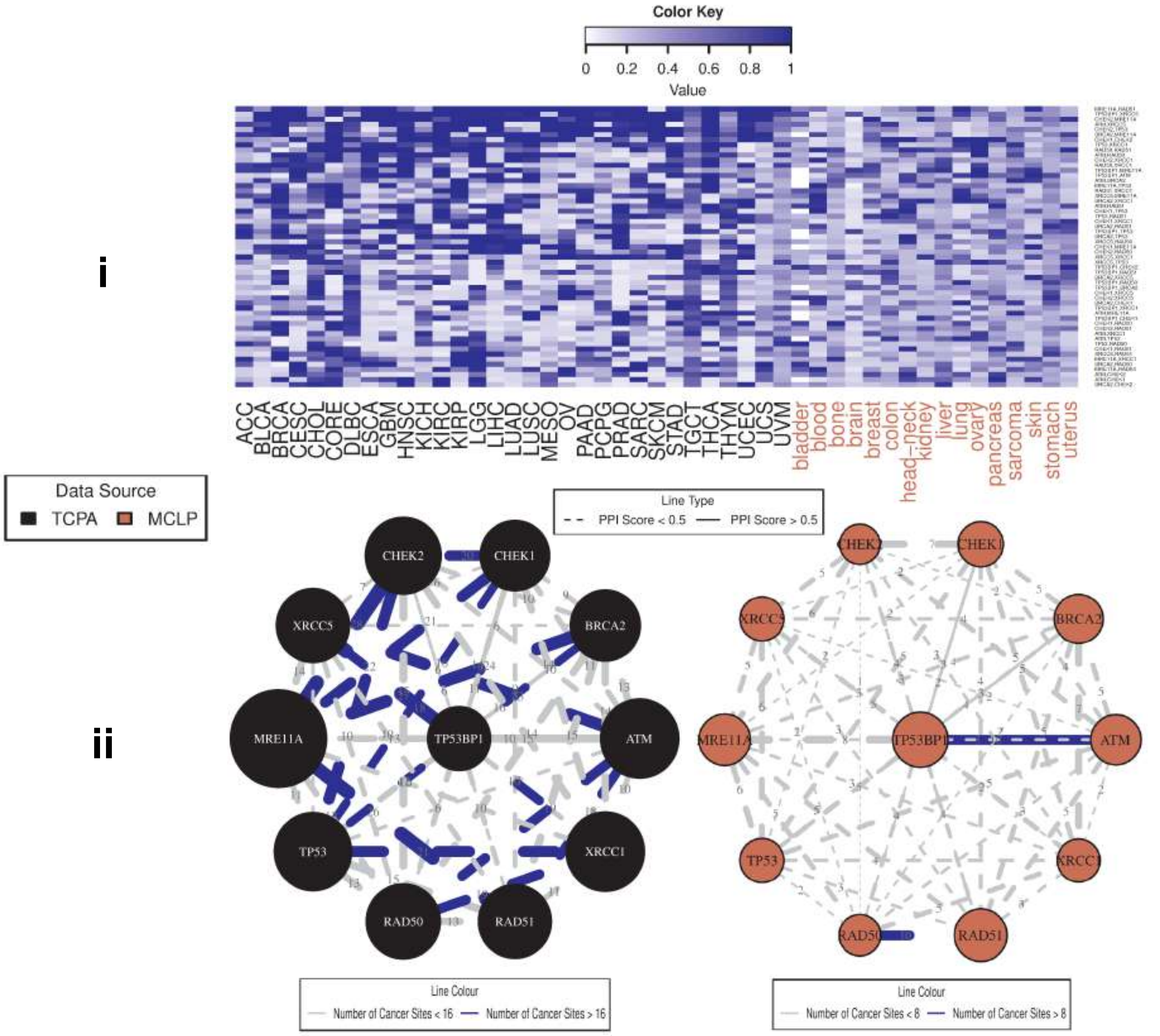


Figure S2.4: Pan-cancer summary of protein networks for DNA damage response. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Left panel exhibits a network with its edges weighted and labeled by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge, also presenting the a priori known strength of the edge using the protein-protein interaction score from the STRING database. The right panel is the corresponding network across cell line cancers.

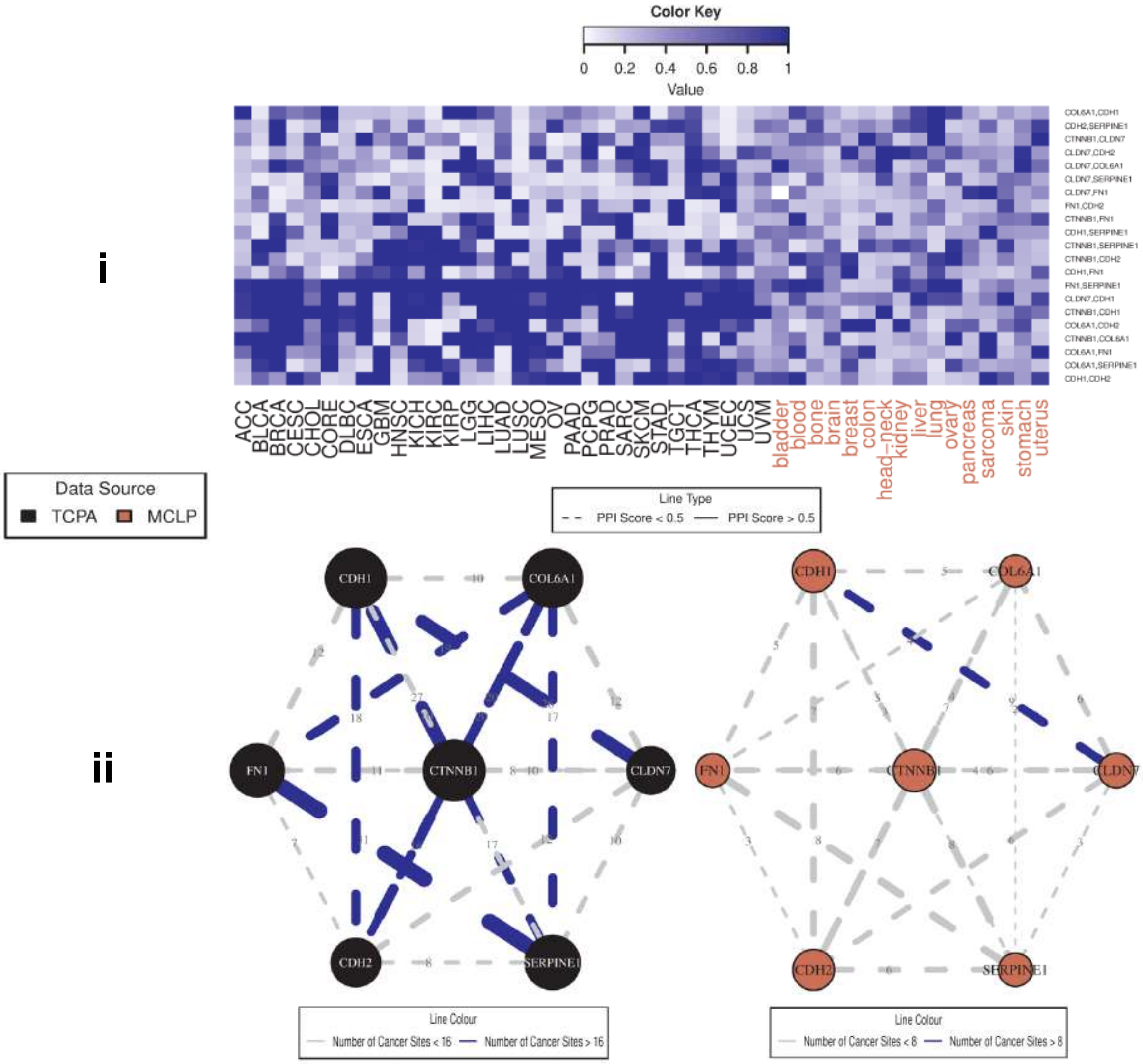


Figure S2.5: Pan-cancer summary of protein networks for EMT pathway. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Left panel exhibits a network with its edges weighted and labeled by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge, also presenting the a priori known strength of the edge using the protein-protein interaction score from the STRING database. The right panel is the corresponding network across cell line cancers.

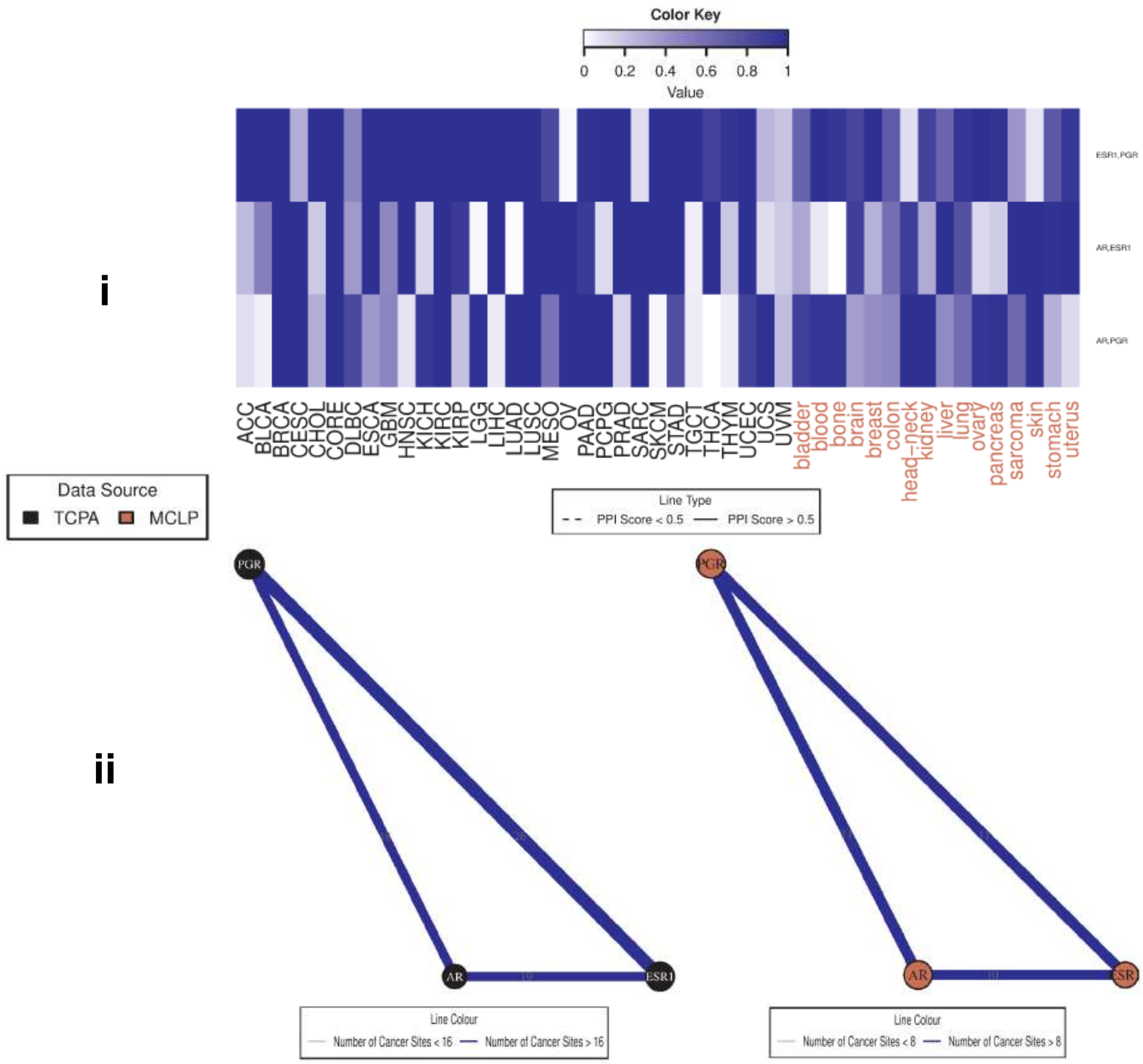


Figure S2.6: Pan-cancer summary of protein networks for hormone receptor pathway. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Left panel exhibits a network with its edges weighted and labeled by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge, also presenting the a priori known strength of the edge using the protein-protein interaction score from the STRING database. The right panel is the corresponding network across cell line cancers.

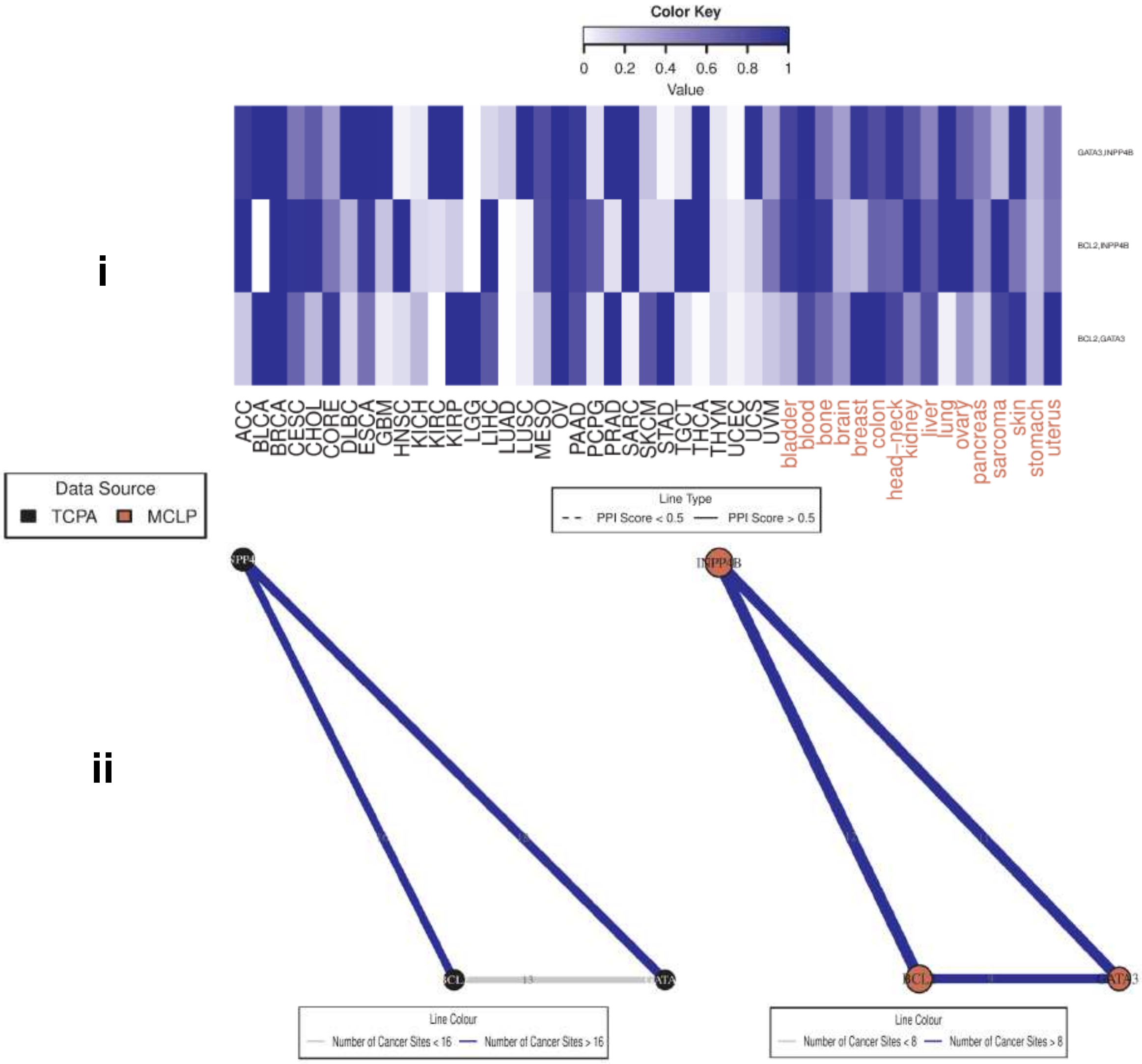


Figure S2.7: Pan-cancer summary of protein networks for hormone signaling (breast) pathway. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Left panel exhibits a network with its edges weighted and labeled by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge, also presenting the a priori known strength of the edge using the protein-protein interaction score from the STRING database. The right panel is the corresponding network across cell line cancers.

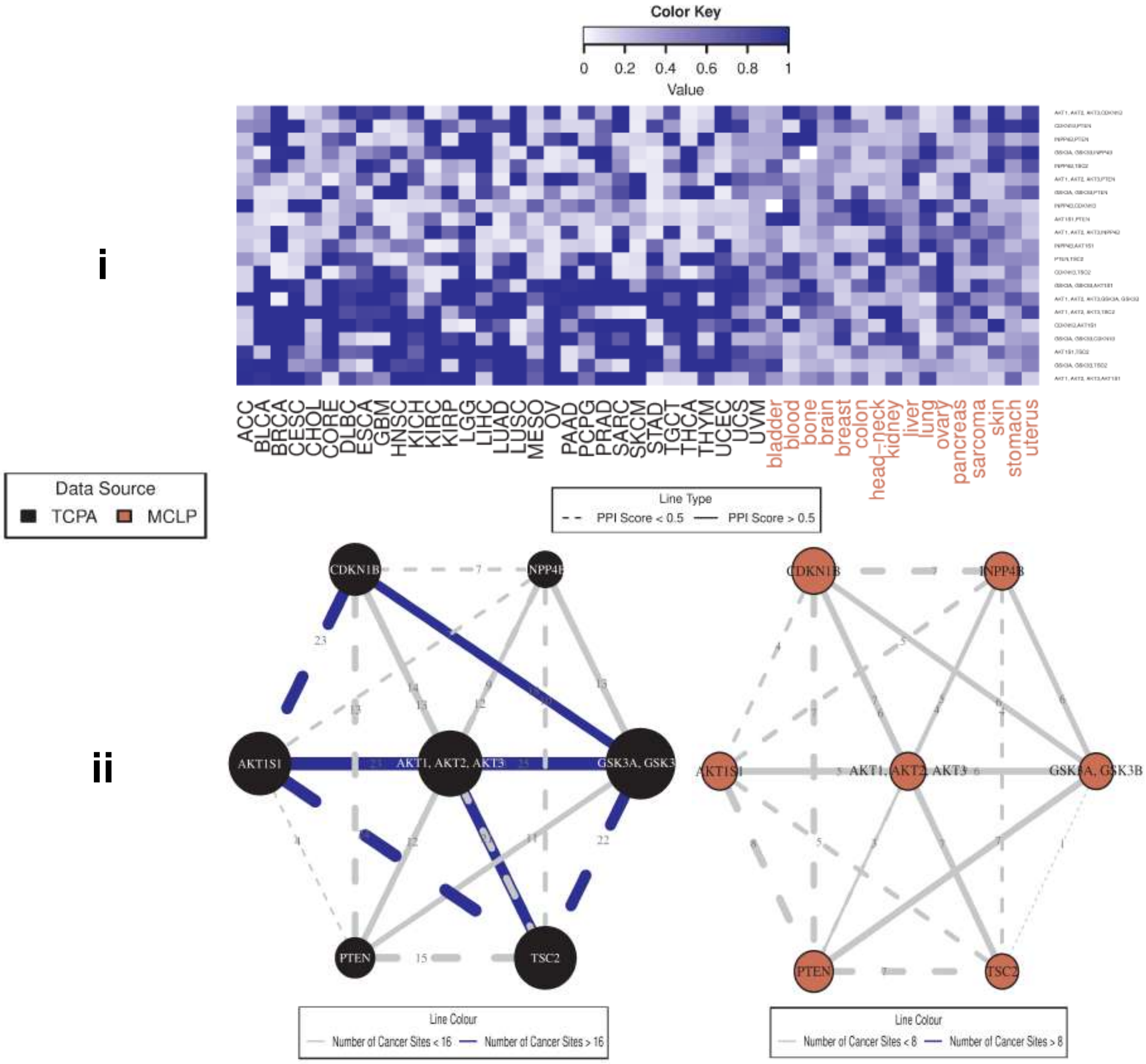


Figure S2.8: Pan-cancer summary of protein networks for PI3K/AKT pathway. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Left panel exhibits a network with its edges weighted and labeled by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge, also presenting the a priori known strength of the edge using the protein-protein interaction score from the STRING database. The right panel is the corresponding network across cell line cancers.

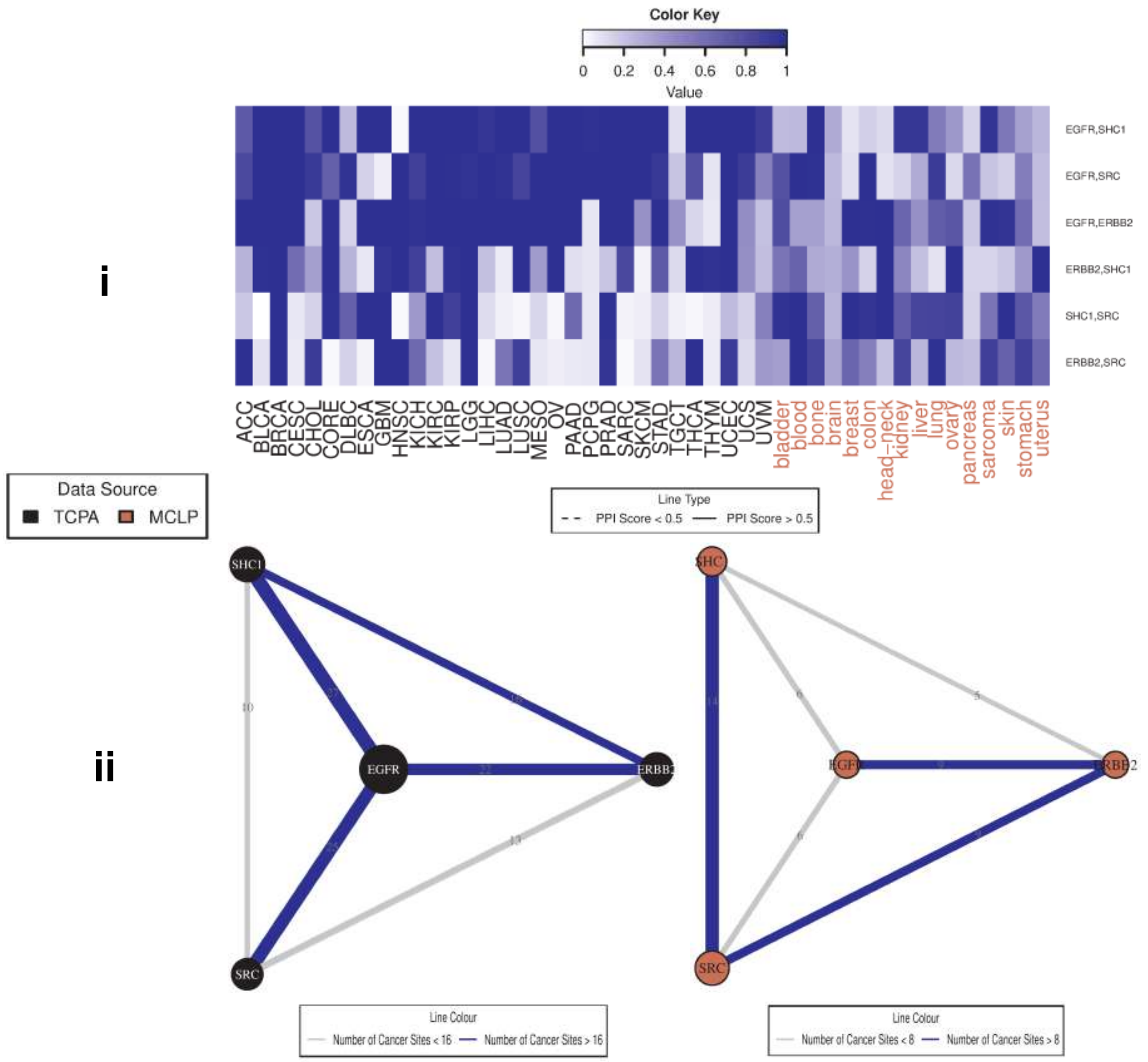


Figure S2.9: Pan-cancer summary of protein networks for RTK pathway. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Left panel exhibits a network with its edges weighted and labeled by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge, also presenting the a priori known strength of the edge using the protein-protein interaction score from the STRING database. The right panel is the corresponding network across cell line cancers.

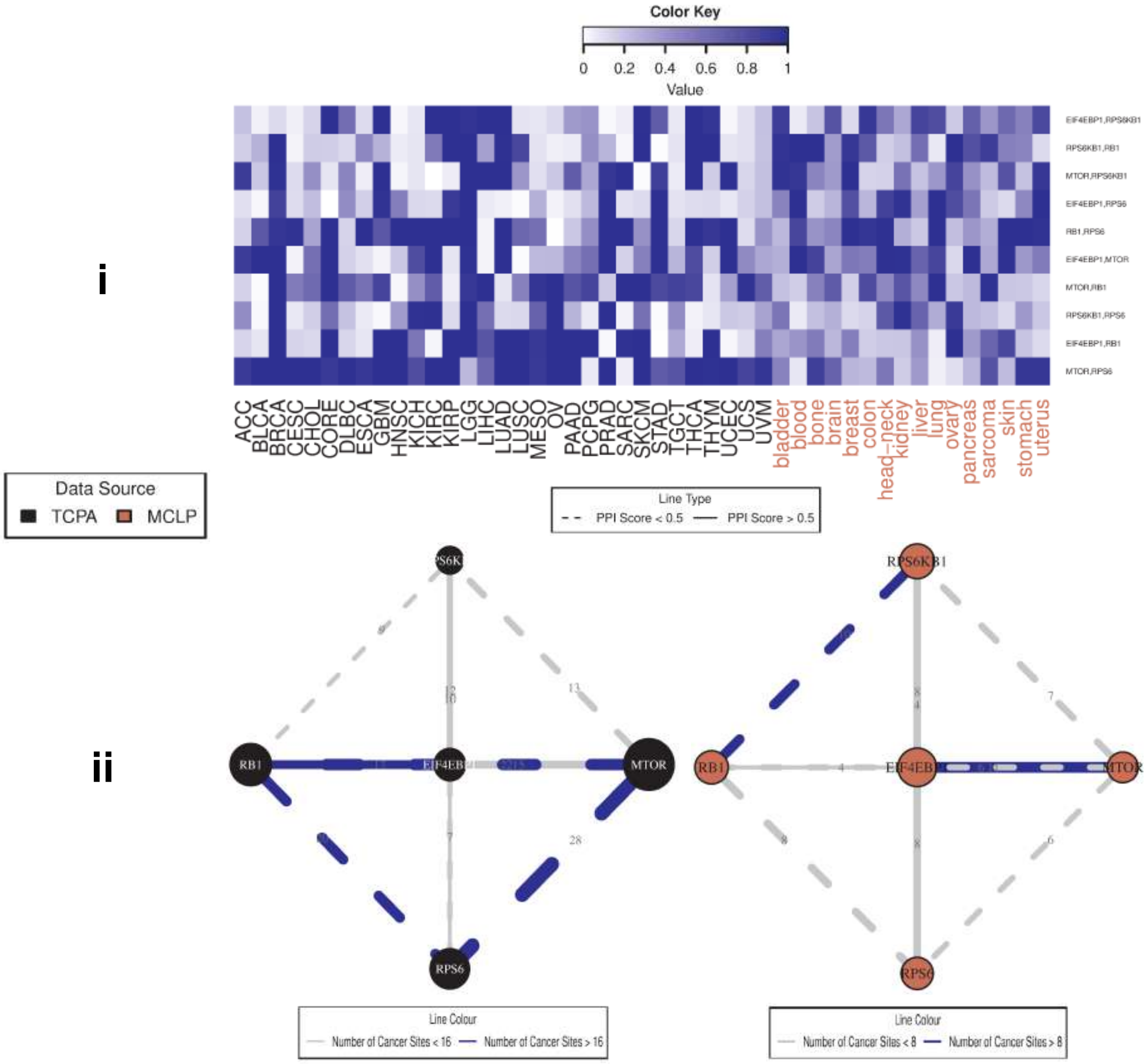


Figure S2.10: Pan-cancer summary of protein networks for TSC/mTOR pathway. i. Heatmap depicting strengths of all possible protein-protein edges within the pathway, across all 47 patient and cell line tumor lineages, quantified by the posterior inclusion probabilities of the edges based on the fitted Bayesian graphical regression model. ii. Left panel exhibits a network with its edges weighted and labeled by the edge consistencies (ECs), which are quantified by the number of patient tumor types holding that particular edge, also presenting the a priori known strength of the edge using the protein-protein interaction score from the STRING database. The right panel is the corresponding network across cell line cancers.

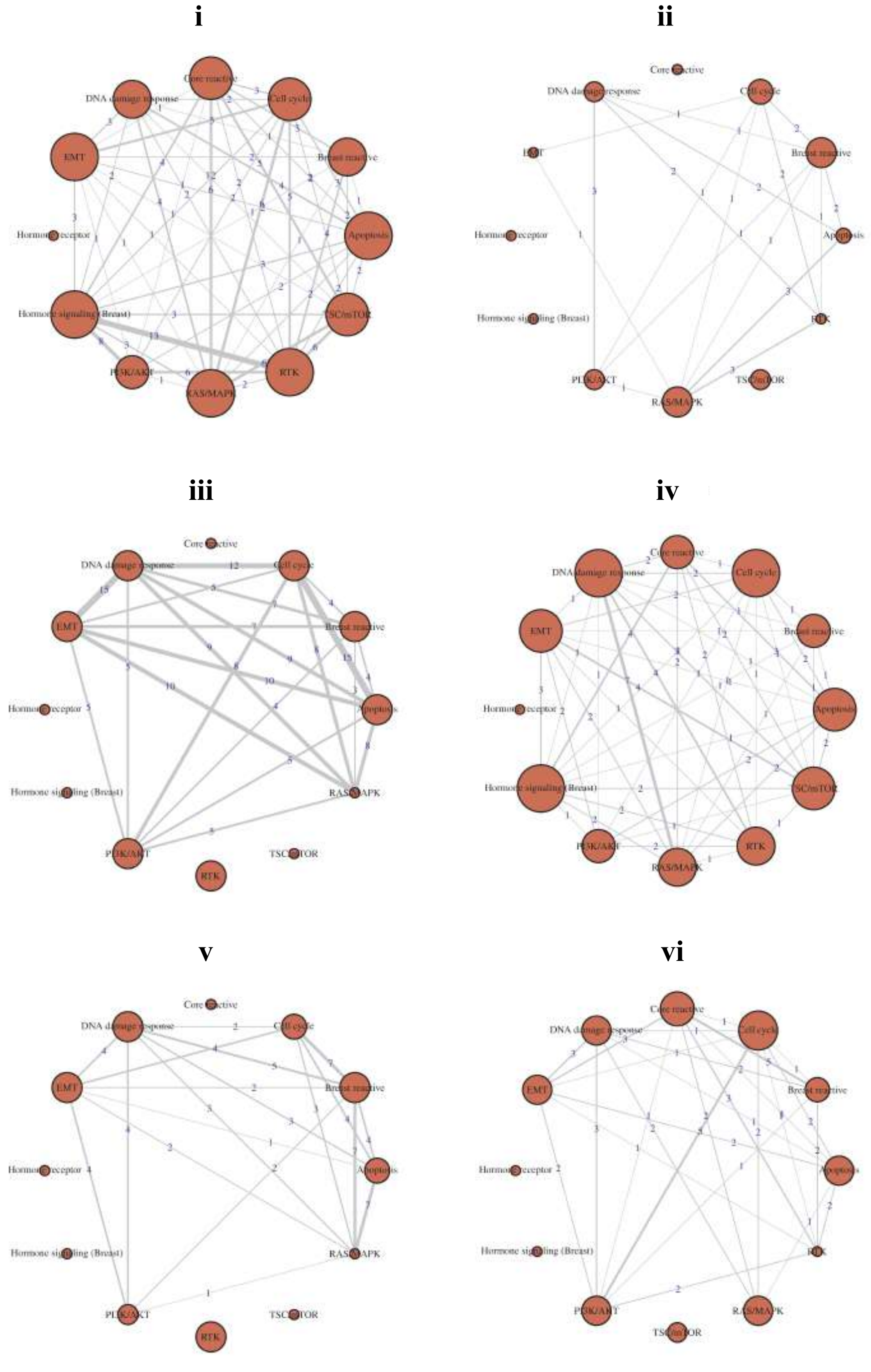


Figure S2.11: Shared presence of pathways in cell line drug sensitivity prediction models. For each lineage, I only look at drugs with at least 10 response profiles available within cell lines from that lineage and the corresponding BART model having a >0.85 test-set AUC based on five-fold cross-validations. Edge weights indicate the number of times the two nodes (pathways) are the top two predictors within all such BART models for a lineage. Each panel corresponds to a different lineage, as follows: i. colon, ii. liver, iii. ovary, iv. pancreas, v. skin, vi. uterus.

2.6 Supplementary Tables

Table S2.1: Summary of patient tumor sample sizes according to lineages.

Cancer Lineage	Study Abbreviation	Number of Patients
Adrenocortical carcinoma	ACC	46
Bladder Urothelial Carcinoma	BLCA	343
Breast invasive carcinoma	BRCA	878
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	171
Cholangiocarcinoma	CHOL	30
Colon/Rectum adenocarcinoma	CORE	491
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	33
Esophageal carcinoma	ESCA	126
Glioblastoma multiforme	GBM	232
Head and Neck squamous cell carcinoma	HNSC	203
Kidney Chromophobe	KICH	63
Kidney renal clear cell carcinoma	KIRC	469
Kidney renal papillary cell carcinoma	KIRP	217
Brain Lower Grade Glioma	LGG	432
Liver hepatocellular carcinoma	LIHC	184

Table S2.1: Summary of patient tumor sample sizes according to lineages.

Cancer Lineage	Study Abbreviation	Number of Patients
Lung adenocarcinoma	LUAD	362
Lung squamous cell carcinoma	LUSC	325
Mesothelioma	MESO	61
Ovarian serous cystadenocarcinoma	OV	431
Pancreatic adenocarcinoma	PAAD	122
Pheochromocytoma and Paraganglioma	PCPG	82
Prostate adenocarcinoma	PRAD	351
Sarcoma	SARC	224
Skin Cutaneous Melanoma	SKCM	355
Stomach adenocarcinoma	STAD	392
Testicular Germ Cell Tumors	TGCT	122
Thyroid carcinoma	THCA	380
Thymoma	THYM	90
Uterine Corpus Endometrial Carcinoma	UCEC	439
Uterine Carcinosarcoma	UCS	48
Uveal Melanoma	UVM	12

Table S2.2: Summary of the genes that the 12 pathways consist of.

Pathway	Genes / Proteins
Apoptosis	BAD, BAK1, BAX, BCL2, BCL2L1, BID, BCL2L11, CASP7, BIRC2
Breast reactive	CTNNB1, CAV1, GAPDH, MYH11, RAB11A, RAB11B, RBM15
Cell cycle	CDK1, CCNB1, CCNE1, CCNE2, FOXM1, CDKN1B, PCNA
Core reactive	CTNNB1, CAV1, CLDN7, CDH1, RBM15
DNA damage response	TP53BP1, ATM, BRCA2, CHEK1, CHEK2, XRCC5, MRE11A, TP53, RAD50, RAD51, XRCC1
EMT	CTNNB1, CLDN7, COL6A1, CDH1, FN1, CDH2, SERPINE1
Hormone receptor	AR, ESR1, PGR
Hormone signaling (Breast)	BCL2, GATA3, INPP4B
PI3K/AKT	AKT1, AKT2, AKT3, GSK3A, GSK3B, INPP4B, CDKN1B, AKT1S1, PTEN, TSC2
RAS/MAPK	ARAF, JUN, RAF1, MAPK8, MAPK1, MAPK3, MAP2K1, MAPK14, RPS6KA1, YBX1
RTK	EGFR, ERBB2, ERBB3, SHC1, SRC
TSC/mTOR	EIF4EBP1, MTOR, RPS6KB1, RB1, RPS6

Table S2.3: Summary of cell line expression sample sizes according to lineages.

Cancer Lineage	Number of Cell lines
bladder	11
blood	101
bone	20
brain	6
breast	57
colon	35
head and neck	53
kidney	29
liver	17
lung	124
ovary	47
pancreas	20
sarcoma	29
skin	46
stomach-oesophagus	13
uterus	32

Table S2.4: Summary of drug sensitivity data availability for cell lines. DR: drug response.

Cancer Lineage	**Cell lines with DR Data**	**Average DR Available / Cell line**
bladder	3	73
blood	4	132
bone	7	85
brain	1	88
breast	35	105
colon	23	96
head and neck	3	84
kidney	6	93
liver	13	75
lung	72	87
ovary	25	78
pancreas	17	82
sarcoma	4	64
skin	15	105
stomach-oesophagus	9	81
uterus	17	78

Table S2.5: Connectivity scores (randomCS proportions) of the integrated cancer-specific networks. The connectivity scores with randomCS proportions < 0.15 are **bold**.

Cancer Lineage	**Apoptosis**	**Breast reactive**	**Cell cycle**	**Core reactive**	**DNA damage response**	**EMT**	**Hormone receptor**	**Hormone signaling (Breast)**	**PI3K/AKT**	**RAS/MAPK**	**RTK**	**TSC/mTOR**
Patient Cancers												
ACC	0.28 (0.96)	0.4 (0.689)	**0.67 (0.007)**	0.4 (0.626)	0.45 (0.164)	0.5 (0.35)	0.33 (0.872)	0.67 (0.292)	0.33 (0.813)	**0.57 (0.046)**	0.67 (0.186)	0.4 (0.594)
BLCA	0.47 (0.837)	**0.77 (0.07)**	**0.71 (0.115)**	0.7 (0.298)	0.44 (0.864)	0.57 (0.467)	0.67 (0.757)	0.67 (0.594)	0.55 (0.589)	0.52 (0.693)	0.67 (0.567)	0.35 (0.967)
BRCA	0.58 (0.776)	**0.8 (0.099)**	0.71 (0.372)	**0.9 (0.067)**	**0.69 (0.073)**	**0.79 (0.103)**	**1 (0)**	**1 (0)**	0.67 (0.507)	0.71 (0.163)	**1 (0)**	**0.9 (0.099)**
CESC	0.51 (0.23)	0.5 (0.491)	0.52 (0.325)	**0.75 (0.042)**	**0.55 (0.017)**	0.52 (0.343)	**1 (0)**	**1 (0)**	0.52 (0.31)	0.45 (0.613)	0.67 (0.341)	0.3 (0.935)
CHOL	0.53 (0.269)	0.5 (0.55)	0.45 (0.696)	**0.7 (0.148)**	**0.74 (0.001)**	0.6 (0.238)	0.5 (0.615)	0.67 (0.455)	0.33 (0.938)	0.57 (0.187)	0.67 (0.317)	0.45 (0.653)
CORE	0.58 (0.53)	0.6 (0.592)	0.52 (0.84)	0.8 (0.178)	**0.63 (0.143)**	0.69 (0.194)	**1 (0)**	**1 (0)**	0.62 (0.483)	0.59 (0.535)	0.83 (0.296)	0.7 (0.353)
DLBC	0.4 (0.593)	0.43 (0.488)	0.52 (0.15)	0.5 (0.356)	**0.69 (0)**	0.33 (0.805)	**1 (0)**	0.5 (0.344)	**0.64 (0.022)**	0.45 (0.353)	0.42 (0.499)	0.5 (0.335)
ESCA	0.44 (0.494)	0.57 (0.244)	**0.6 (0.117)**	**0.8 (0.034)**	0.36 (0.875)	0.5 (0.445)	**1 (0)**	**1 (0)**	**0.71 (0.006)**	**0.57 (0.113)**	0.5 (0.637)	0.4 (0.785)
GBM	0.51 (0.521)	0.6 (0.404)	0.6 (0.361)	**0.8 (0.079)**	0.54 (0.242)	0.5 (0.634)	0.67 (0.634)	0.33 (0.944)	**0.67 (0.1)**	0.5 (0.589)	0.67 (0.447)	**0.75 (0.091)**
HNSC	0.44 (0.596)	0.47 (0.63)	0.33 (0.936)	0.6 (0.359)	**0.52 (0.094)**	0.55 (0.261)	0.67 (0.515)	0.33 (0.89)	**0.67 (0.026)**	0.48 (0.456)	0.67 (0.281)	0.5 (0.579)
KICH	0.4 (0.897)	**0.7 (0.097)**	0.62 (0.199)	0.7 (0.255)	0.48 (0.314)	0.62 (0.175)	0.67 (0.668)	0 (1)	0.55 (0.411)	0.39 (0.934)	**1 (0)**	0.55 (0.578)
KIRC	0.54 (0.613)	0.67 (0.303)	**0.81 (0.015)**	0.6 (0.621)	0.54 (0.563)	0.5 (0.833)	**1 (0)**	0.33 (0.959)	0.64 (0.295)	**0.71 (0.04)**	0.67 (0.577)	0.7 (0.393)
KIRP	0.56 (0.439)	0.53 (0.745)	0.64 (0.273)	0.4 (0.955)	**0.58 (0.146)**	0.55 (0.682)	0.67 (0.738)	0.67 (0.628)	0.55 (0.657)	0.62 (0.225)	0.83 (0.268)	0.6 (0.584)
LGG	0.53 (0.902)	0.73 (0.335)	0.76 (0.162)	0.3 (0.995)	0.61 (0.476)	0.55 (0.853)	0.67 (0.782)	0.33 (0.97)	0.69 (0.35)	0.62 (0.557)	0.75 (0.39)	**0.9 (0.071)**
LIHC	0.6 (0.172)	0.33 (0.992)	0.55 (0.62)	0.5 (0.804)	0.49 (0.577)	0.62 (0.357)	0.67 (0.782)	0.67 (0.698)	0.62 (0.332)	0.5 (0.691)	0.5 (0.886)	0.6 (0.61)
LUAD	0.58 (0.296)	0.63 (0.348)	**0.74 (0.046)**	0.65 (0.366)	0.52 (0.421)	0.64 (0.244)	0.67 (0.681)	0 (1)	0.55 (0.523)	0.45 (0.847)	0.67 (0.529)	0.7 (0.322)
LUSC	0.46 (0.768)	0.6 (0.411)	**0.67 (0.145)**	0.7 (0.278)	0.38 (0.954)	0.57 (0.439)	**1 (0)**	0.33 (0.946)	0.64 (0.174)	**0.64 (0.111)**	0.83 (0.205)	0.5 (0.768)
MESO	0.31 (0.913)	0.5 (0.26)	0.38 (0.648)	0.5 (0.41)	0.35 (0.689)	0.48 (0.27)	**1 (0)**	0.67 (0.308)	0.38 (0.568)	0.41 (0.421)	0.58 (0.209)	0.55 (0.216)
OV	0.53 (0.217)	0.5 (0.569)	0.36 (0.937)	0.6 (0.378)	0.5 (0.238)	**0.62 (0.143)**	0.67 (0.544)	**1 (0)**	0.57 (0.246)	**0.7 (0.006)**	0.67 (0.391)	0.4 (0.878)
PAAD	0.4 (0.795)	0.57 (0.336)	0.55 (0.303)	**0.8 (0.028)**	0.45 (0.461)	0.45 (0.663)	**1 (0)**	**1 (0)**	0.57 (0.202)	0.45 (0.634)	0.67 (0.3)	0.6 (0.342)
PCPG	0.4 (0.693)	0.57 (0.229)	0.38 (0.832)	0.5 (0.557)	0.41 (0.517)	0.55 (0.192)	0.67 (0.487)	0.33 (0.885)	0.52 (0.243)	**0.64 (0.02)**	0.33 (0.885)	0.5 (0.568)
PRAD	0.67 (0.192)	0.73 (0.266)	0.6 (0.721)	0.65 (0.563)	0.52 (0.816)	0.57 (0.784)	0.67 (0.756)	0.67 (0.713)	0.6 (0.664)	0.59 (0.644)	0.83 (0.352)	0.6 (0.772)
SARC	0.5 (0.367)	**0.77 (0.02)**	0.48 (0.676)	0.6 (0.395)	0.43 (0.639)	0.5 (0.534)	0.67 (0.601)	0.67 (0.489)	**0.69 (0.022)**	0.52 (0.387)	0.67 (0.396)	0.4 (0.863)
SKCM	0.39 (0.961)	0.5 (0.731)	0.62 (0.232)	0.7 (0.244)	0.51 (0.339)	**0.71 (0.062)**	0.67 (0.682)	0.33 (0.932)	0.43 (0.884)	**0.66 (0.063)**	0.5 (0.827)	**0.75 (0.135)**
STAD	0.54 (0.692)	0.53 (0.792)	0.57 (0.673)	0.7 (0.378)	0.46 (0.934)	0.69 (0.208)	**1 (0)**	0.33 (0.959)	0.43 (0.96)	0.57 (0.53)	0.83 (0.285)	0.7 (0.373)
TGCT	0.46 (0.596)	0.6 (0.272)	0.6 (0.218)	0.55 (0.527)	**0.78 (0)**	0.48 (0.602)	0.33 (0.92)	0.33 (0.901)	0.55 (0.265)	0.5 (0.442)	0.17 (0.999)	0.35 (0.928)
THCA	0.53 (0.761)	0.6 (0.606)	**0.74 (0.117)**	0.75 (0.206)	0.55 (0.611)	**0.79 (0.037)**	0.67 (0.729)	0.67 (0.657)	0.57 (0.636)	0.68 (0.15)	0.67 (0.57)	0.7 (0.383)
THYM	0.54 (0.153)	**0.63 (0.146)**	0.43 (0.762)	0.65 (0.174)	**0.67 (0)**	0.55 (0.349)	0.33 (0.932)	0 (1)	0.48 (0.542)	0.46 (0.577)	0.33 (0.928)	0.5 (0.577)
UCEC	0.44 (0.923)	0.67 (0.256)	0.64 (0.254)	**0.9 (0.023)**	0.47 (0.808)	0.67 (0.183)	**1 (0)**	0 (1)	0.62 (0.286)	**0.7 (0.053)**	0.83 (0.181)	0.45 (0.83)
UCS	0.32 (0.797)	0.4 (0.571)	0.38 (0.576)	0.6 (0.155)	0.39 (0.337)	**0.52 (0.121)**	0.33 (0.829)	0.33 (0.803)	**0.62 (0.017)**	0.48 (0.175)	0.5 (0.465)	0.3 (0.845)
UVM	0.58 (0.596)	0.37 (0.852)	0.55 (0.506)	0.3 (0.922)	0.58 (0.906)	0.52 (0.57)	0 (1)	0.67 (0.405)	0.55 (0.507)	**0.77 (0.114)**	0.67 (0.265)	0.6 (0.319)
Cell line Cancers												
bladder	0.39 (0.767)	0.63 (0.336)	0.38 (0.867)	0.8 (0.185)	0.3 (0.897)	0.57 (0.425)	0.83 (0.786)	0.83 (0.746)	0.67 (0.242)	**0.71 (0.115)**	0.75 (0.483)	0.7 (0.321)
blood	0.44 (0.71)	0.5 (0.911)	0.5 (0.774)	0.7 (0.43)	0.42 (0.495)	0.6 (0.236)	0.67 (0.999)	**1 (0)**	0.55 (0.497)	0.43 (0.927)	0.83 (0.425)	0.55 (0.926)
bone	**0.6 (0.038)**	0.57 (0.427)	0.4 (0.884)	0.6 (0.507)	0.35 (0.787)	0.5 (0.546)	0.67 (0.998)	**1 (0)**	0.36 (0.969)	0.38 (0.903)	0.83 (0.202)	0.6 (0.543)
brain	0.33 (0.168)	0.7 (0.2)	0.55 (0.19)	0.65 (0.753)	0.13 (0.584)	0.4 (0.701)	**1 (0)**	**1 (0)**	0.52 (0.233)	0.41 (0.2)	0.5 (0.956)	0.65 (0.723)
breast	0.47 (0.398)	0.43 (0.973)	0.52 (0.497)	0.6 (0.678)	0.45 (0.348)	0.4 (0.961)	0.67 (0.991)	0.67 (0.982)	0.45 (0.844)	0.46 (0.648)	0.58 (0.85)	0.6 (0.649)
colon	0.49 (0.367)	0.5 (0.546)	0.31 (0.976)	**0.8 (0.04)**	0.48 (0.27)	0.43 (0.747)	**1 (0)**	**1 (0)**	0.38 (0.848)	**0.59 (0.101)**	0.67 (0.327)	0.5 (0.544)
head and neck	0.42 (0.675)	0.53 (0.532)	0.5 (0.492)	0.65 (0.323)	0.4 (0.501)	0.4 (0.917)	0.67 (0.971)	**1 (0)**	0.38 (0.97)	0.38 (0.908)	0.5 (0.976)	0.65 (0.381)
kidney	0.51 (0.222)	0.4 (0.978)	0.57 (0.224)	0.45 (0.968)	0.35 (0.881)	0.38 (0.971)	0.83 (0.521)	0.83 (0.525)	0.55 (0.354)	**0.61 (0.063)**	0.83 (0.208)	0.55 (0.794)
liver	0.38 (0.904)	0.33 (0.995)	0.43 (0.885)	0.5 (0.835)	0.52 (0.239)	**0.71 (0.045)**	0.83 (0.569)	**1 (0)**	**0.64 (0.13)**	0.52 (0.355)	0.58 (0.761)	0.75 (0.164)
lung	0.51 (0.335)	0.57 (0.576)	0.62 (0.15)	0.65 (0.474)	0.46 (0.44)	0.55 (0.501)	**1 (0)**	0.67 (0.969)	0.48 (0.871)	0.55 (0.297)	0.83 (0.281)	0.65 (0.449)
ovary	0.47 (0.633)	0.53 (0.661)	0.62 (0.153)	0.55 (0.79)	0.46 (0.532)	0.52 (0.597)	0.67 (0.985)	0.83 (0.389)	0.55 (0.487)	0.34 (0.998)	0.58 (0.815)	0.5 (0.926)
pancreas	0.35 (0.894)	0.4 (0.746)	0.36 (0.834)	0.6 (0.232)	**0.55 (0.141)**	0.36 (0.831)	0.67 (0.521)	0.67 (0.396)	0.55 (0.191)	**0.55 (0.149)**	0.17 (0.992)	0.6 (0.206)
sarcoma	0.49 (0.477)	0.63 (0.216)	0.55 (0.375)	0.5 (0.895)	0.44 (0.653)	0.4 (0.921)	0.83 (0.439)	0.67 (0.962)	0.52 (0.5)	0.59 (0.154)	0.67 (0.67)	0.6 (0.577)
skin	0.4 (0.826)	0.5 (0.806)	0.52 (0.475)	0.6 (0.713)	0.35 (0.904)	**0.62 (0.123)**	0.67 (0.995)	**1 (0)**	0.57 (0.253)	0.52 (0.303)	0.67 (0.777)	0.6 (0.668)
stomach-oesophagus	0.53 (0.856)	0.57 (0.755)	0.62 (0.585)	0.65 (0.59)	0.63 (0.579)	0.62 (0.609)	0.83 (0.447)	0 (1)	0.64 (0.53)	0.46 (0.948)	**1 (0)**	0.55 (0.81)
uterus	0.5 (0.338)	0.57 (0.402)	0.45 (0.741)	0.65 (0.274)	0.42 (0.695)	0.62 (0.146)	0.67 (0.953)	0.83 (0.346)	0.6 (0.163)	0.41 (0.823)	0.5 (0.954)	0.6 (0.491)

Table S2.6: Membership of samples by cancer lineages in the 29 clusters obtained from hierarchical clustering on the network aberration scores matrix. The color codes are as follows – patient cancers, cell line cancers, clusters with notable membership from both the model systems.

Cancer Lineages	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
ACC	30	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BLCA	0	0	0	76	50	209	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRCA	0	0	0	0	0	0	0	455	57	230	114	5	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CESC	87	0	0	0	0	0	0	0	0	0	0	0	0	1	82	1	0	0	0	0	0	0	0	0	0	0	0	0	0
CHOL	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CORE	0	0	0	17	0	0	0	1	47	298	27	1	54	14	0	0	32	0	0	0	0	0	0	0	0	0	0	0	0
DLBC	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0
ESCA	24	1	0	0	21	0	0	0	0	0	0	0	0	0	55	21	0	0	4	0	0	0	0	0	0	0	0	0	0
GBM	12	0	0	9	0	0	1	0	1	0	0	4	0	0	45	2	0	0	0	154	4	0	0	0	0	0	0	0	0
HNSC	0	0	0	78	0	0	0	0	0	0	0	0	0	0	122	0	0	0	0	3	0	0	0	0	0	0	0	0	0
KICH	3	0	0	0	0	0	0	0	0	0	0	0	0	0	5	2	0	5	0	0	0	5	43	0	0	0	0	0	0
KIRC	1	0	0	0	0	0	0	0	6	0	0	0	0	0	3	78	0	0	0	2	125	56	0	87	87	24	0	0	0
KIRP	7	0	0	0	0	0	0	0	0	0	0	82	0	0	41	12	0	0	0	15	35	1	0	20	4	0	0	0	0
LGG	2	0	0	0	4	0	0	0	1	0	0	5	0	0	2	7	0	0	0	44	61	23	0	283	0	0	0	0	0
LIHC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0	0	0	0	0	0	33	1	0
LUAD	47	0	0	0	0	75	0	0	1	0	0	28	0	0	13	33	0	2	0	27	0	31	0	54	0	51	0	0	0
LUSC	0	0	0	0	16	19	0	0	0	0	0	5	0	0	0	39	0	13	0	0	0	53	29	0	50	101	0	0	0
MESO	7	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OV	0	0	0	49	0	0	0	0	2	61	87	19	155	40	2	0	16	0	0	0	0	0	0	0	0	0	0	0	0
PAAD	0	0	1	97	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PCPG	26	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRAD	0	0	0	0	1	0	0	0	53	84	90	55	10	13	0	0	0	0	0	1	0	0	0	44	0	0	0	0	0
SARC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	67	157	0	0	0	0	0	0	0	0	0	0	0	0	0
SKCM	25	0	0	0	0	29	0	0	0	0	0	0	0	0	5	6	0	12	0	0	5	125	42	0	0	106	0	0	0
STAD	0	0	0	0	224	0	0	0	0	0	0	87	0	75	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0
TGCT	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	87	0	0	0	0	0	0	0	7	0	0	0
THCA	0	0	0	0	9	0	0	0	0	0	0	3	0	0	6	1	0	0	0	3	31	72	0	230	25	0	0	0	0
THYM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90	0	0	0	0	0	0	0	0	0	0	0
UCEC	26	0	0	0	0	6	0	0	0	0	0	0	0	0	2	49	0	0	0	78	20	105	15	0	108	30	0	0	0
UCS	0	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
UVM	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
bladder	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
blood	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	48	0
bone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	5
brain	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
breast	0	0	7	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	6
colon	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	1
head and neck	0	0	4	38	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	6	4
kidney	0	0	4	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	4
liver	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	2
lung	0	0	7	44	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	9
ovary	0	0	2	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	4
pancreas	0	14	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
sarcoma	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	7	0
skin	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	6
stomach-oesophagus	0	0	4	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	3	1
uterus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	7	5

Table S2.7: Shared memberships of cell lines and cancer patients in clusters obtained from hierarchical clustering on the network aberration scores, along with the pathways driving the shared presence.

Cluster	Cell Line Lineages	Patient Tumors	Driving Pathways
C2	pancreas (70%) colon (26%)	Mesothelioma (89%) Uveal Melanoma (83%) Pheochromocytoma and Paraganglioma (68%) Adrenocortical carcinoma (33%)	RAS/MAPK
C4	ovary (81%) head and neck (72%) skin (48%) lung (35%) kidney (34%) breast (33%) pancreas (20%)	Pancreatic adenocarcinoma (80%) Head and Neck squamous cell carcinoma (38%) Bladder Urothelial Carcinoma (22%) Ovarian serous cystadenocarcinoma (11%)	Apoptosis DNA Damage Response
C9	lung (40%)	Prostate adenocarcinoma (15%)	Apoptosis Breast Reactive Core Reactive DNA Damage Response
C15	stomach-oesophagus (38%)	Head and Neck squamous cell carcinoma (60%) Cervical squamous cell carcinoma and endocervical adenocarcinoma (48%) Esophageal carcinoma (40%) Sarcoma (30%) Kidney renal papillary cell carcinoma (19%) Glioblastoma multiforme (19%)	DNA Damage Response PI3K/AKT
C19	uterus (62%) blood (36%) sarcoma (34%)	Liver hepatocellular carcinoma (82%)	DNA Damage Response PI3K/AKT RAS/MAPK
C23	colon (71%)	Kidney Chromophobe (68%)	Apoptosis Cell Cycle EMT

Table S2.8: Summary of Bayesian additive regression tree models for each cancer type and each drug. Models are fit using cell lines' drug sensitivity and network aberration scores. Then the patient drug responses for a cancer type are predicted using the model for that drug and the cell line cancer lineage for the same tissue. The response rate is defined as the proportion of samples for that cancer type with a > 0.5 predicted response with respect to that drug. Drugs marked with an asterisk ($*$) have been or are being investigated through clinical trials on cancers of the corresponding tissues (see Discussion section of the paper for references).

Patient Cancer	**Drugs/Compounds (Response Rates)**	**Top 3 Predictive Pathways in Cell Line Training Models**	**Targets**
BRCA	Ibrutinib* (100%)	RAS/MAPK, PI3K/AKT, EMT	BTK
CORE	Lapatinib* (100%)	RAS/MAPK, Cell cycle, EMT	ERBB2, EGFR
	KU.0060648 (98.2%)	Apoptosis, Hormone signaling, RAS/MAPK	PI3K, DNA-PK
	PD318088 (96.7%)	RTK, PI3K/AKT, Core reactive	MEK1/2
	Canertinib (94.9%)	Core reactive, RAS/MAPK, Hormone signaling	EGFR, HER-2, ErbB-4
	Neratinib (93.5%)	Cell cycle, RTK, EMT	Her2, EGFR
	Afatinib (93.5%)	DNA damage response, Hormone signaling, Breast reactive	ERBB2, EGFR
	Navitoclax (68.4%)	EMT, Breast reactive, Cell cycle	BCL2, BCL-XL, BCL-W
LIHC	PD318088 (100%)	PI3K/AKT, Cell cycle, EMT	MEK1/2
	UNC0638 (100%)	Apoptosis, PI3K/AKT, DNA damage response	G9a, GLP
	Linifanib (100%)	PI3K/AKT, EMT, RAS/MAPK	VEGFR1-3, CSF1R, FLT3, KIT
	Pandacostat (99.5%)	Breast reactive, RTK, RAS/MAPK	HDAC1-9
PAAD	NPC.26 (100%)	Cell cycle, Breast reactive, TSC/mTOR	mPTP
	Trametinib (72.1%)	Apoptosis, TSC/mTOR, Cell cycle	MEK1/2
	BRD9876 (62.3%)	RAS/MAPK, TSC/mTOR, PI3K/AKT	Eg5
	UNC0638 (58.2%)	EMT, Breast reactive, Cell cycle	G9a and GLP methyltransferases
	Methotrexate (50.8%)	Apoptosis, Cell Cycle, Breast reactive	Antimetabolite
SKCM	PLX.4032 (100%)	EMT, DNA damage response, PI3K/AKT	B-Raf
	GDC.0879 (100%)	PI3K/AKT, EMT, Cell cycle	B-Raf
	PLX.4720 (81.4%)	RAS/MAPK, Apoptosis, PI3K/AKT	B-Raf
	VAF.347 (52.7%)	PI3K/AKT, DNA damage response, RAS/MAPK	AhR
UCEC	BRD.K97651142 (74.9%)	Breast reactive, RAS/MAPK, RTK	EGFR, ERBB2

CHAPTER III

Functional Integrative Bayesian Analysis of High-dimensional Multiplatform Genomic Data

3.1 Introduction

Rapid advancements in collection, processing, and dissemination of multi-platform molecular patient data has resulted in enormous opportunities to aggregate such data in order to understand, prevent, and treat diseases. This has catalyzed development of integrative methods that can collectively mine multiple types and scales of multi-omics data, in order to provide a more holistic view of the human disease evolution and progression (Subramanian et al., 2020). Specifically, in context of cancer, a disease driven predominantly by agglomerations of several molecular changes (Sun et al., 2021), the importance of synthesizing information from multi-platform omics and clinical sources to understand the cellular basis and behavior of the disease is even further underscored. Cellular oncological mechanisms, triggered at different molecular levels of the DNA → RNA → Protein path, can confer profound phenotypic advantages (or disadvantages). While significant improvements have been made in multi-omics data integration methods to unveil such mechanisms, focused on both prognosis (Duan et al., 2021) and treatment (Finotello et al., 2020), the precise functions governing these mechanisms needs detailed and data-driven de-novo evaluations. Our work, in the same vein, aims at two different but inter-related scientific axes: (i) selection of biomarkers associated with cancer prognosis and clinical outcomes, (ii) learning the mechanism of these biomarkers' effects upon such outcomes via integrating upstream molecular information - we provide some additional scientific context below.

Classes of integrative omics models First, I briefly discuss existing integrative omics approaches in order to contextualize the need for our framework. Broadly, most of the existing integrative statistical methods can be classified into two categories - horizontal (meta-analysis type) and vertical (multi-omics) integration procedures (Tseng et al., 2015). Horizontal meta-analysis methods focus on integrating data on similar omics features from different sources such as laboratories, cohorts, sites, etc; examples include works by Tu et al. (2015) and Angel et al. (2020). Vertical integrative methods, on the other hand, are focused on integrating data sets on the same cohort of samples obtained from different omics experiments, wherein the data sets can be vertically aligned; examples include works by Cheng et al. (2015) and Kaplan and Lock (2017). (See Richardson et al. (2016) and Morris and Baladandayuthapani (2017) for a comprehensive review of integrative methods.) Most, if not all such studies perform the integration in an agnostic manner – they neither take into account known biological structures nor utilize data illustrating functional roles of the markers of interest. Incorporating such structures and molecular regulatory information into integrative models can improve both the power to detect true biomarkers of a disease and the understanding of their cellular roles in the progression of it, as I discuss next.

Importance of functional information Broadly, by *functional information*, I mean the knowledge of the molecular functions of the cellular genomic, epigenomic, and transcriptomic elements, leading to disease outcomes. Incorporation of such information in biomarker association models is important due to several reasons. First, different omics components of the molecular configuration of a disease, while interconnected and hierarchical, can provide complementary information. A recent review by Buccitelli and Selbach (2020) indicates how in the DNA → RNA → Protein path, termed the ‘gene expression pathway’, lower or higher correlations of the expressions of proteins and their coding genes may be observed due to changes in the functional regulatory elements. Second, specifically for cancer, recent literature indicates that recurrent regulatory structures drive tumor progression through aberrations at different omics levels via common *master regulatory mechanisms* (Califano and Alvarez, 2017). Finally, experimental validation and characterization of functional information on a biomarker-by-biomarker basis is resource-intensive - especially with many plausible candidates. Thus, computational models that can identify and incorporate functional information about genes/proteins (referred to as *proteogenomic* data henceforth) into the models rather than post-hoc analyses in a natural, inherent way can facilitate this understanding and can lead to *de-novo data-driven prioritization* of the relevant biomarkers, especially for translational and clinical utility. To this end, I propose a framework called Functional Integrative Bayesian Analysis of High-dimensional Multiplatform Genomic Data (fiBAG, in short), that allows simultaneous identification of upstream functional evidence of proteogenomic biomarkers and the incorporation of such knowledge in Bayesian (biomarker) selection models to improve signal detection.

Goals and utility of fiBAG Our scientific goals are multifold. I focus on integrating high-dimensional multi-omics data and clinical responses in an approach similar to that of Wang et al. (2013b), deciphering and delineating functional roles of proteogenomic markers using mechanism-driven (*mechanistic*) models, and incorporating this functional information into *outcome* models, thus providing functional relevance to the findings. In particular, I want to contextualize the available functional information such as the type of proteogenomic activity following existing literature such as Gevaert et al. (2013) and Song et al. (2019). Using evidence from these mechanistic models, I then guide selection and hence the degree of penalization/prioritization of covariates in the final outcome models. Figure 3.1 provides a broad-scale summary of how the fiBAG procedure achieves these goals. The first column describes the upstream data utilized to infer functional information. For the purpose of this work, I only use copy number and DNA methylation; other data platforms having potential functional relevance (such as microRNA) may also be utilized. The middle column describes the three axes of mechanistic information that I intend to infer on using gene and protein expression data, namely, *driver gene* (dashed line), *driver protein* (dashed-dotted line), and *cascading protein* (dotted line). I describe the construction of these mechanistic models in more detail in the following Section(s). The third and final column describes the “calibrated" outcome model, where summary information from the mechanistic models are incorporated alongside outcome data to improve selection of genes and proteins. Briefly, our study offers both methodological novelty via proposing a calibrated Bayesian variable selection procedure for the outcome model, and scientific innovation via performing integration of both patient outcomes and tumor features with omics data.

Methodological novelty Using a mapping function to calibrate numerical evidences of significance obtained from the mechanistic models to a prior inclusion probability scale, I inform the outcome model of prior functional evidence in favor of specific proteogenomic candidates. Our method is flexible – the calibration function can be adapted according to the choice of the mechanistic model and the resulting quantification of significance. This hierarchical evidence sharing procedure allows our method to integrate data across any number of genomic, epigenomic, and other relevant platforms of choice. Using Gaussian processes to identify the mechanistic evidence, our model is better equipped to pick up nonlinear cellular associations than a standard linear model, as used in Wang et al.

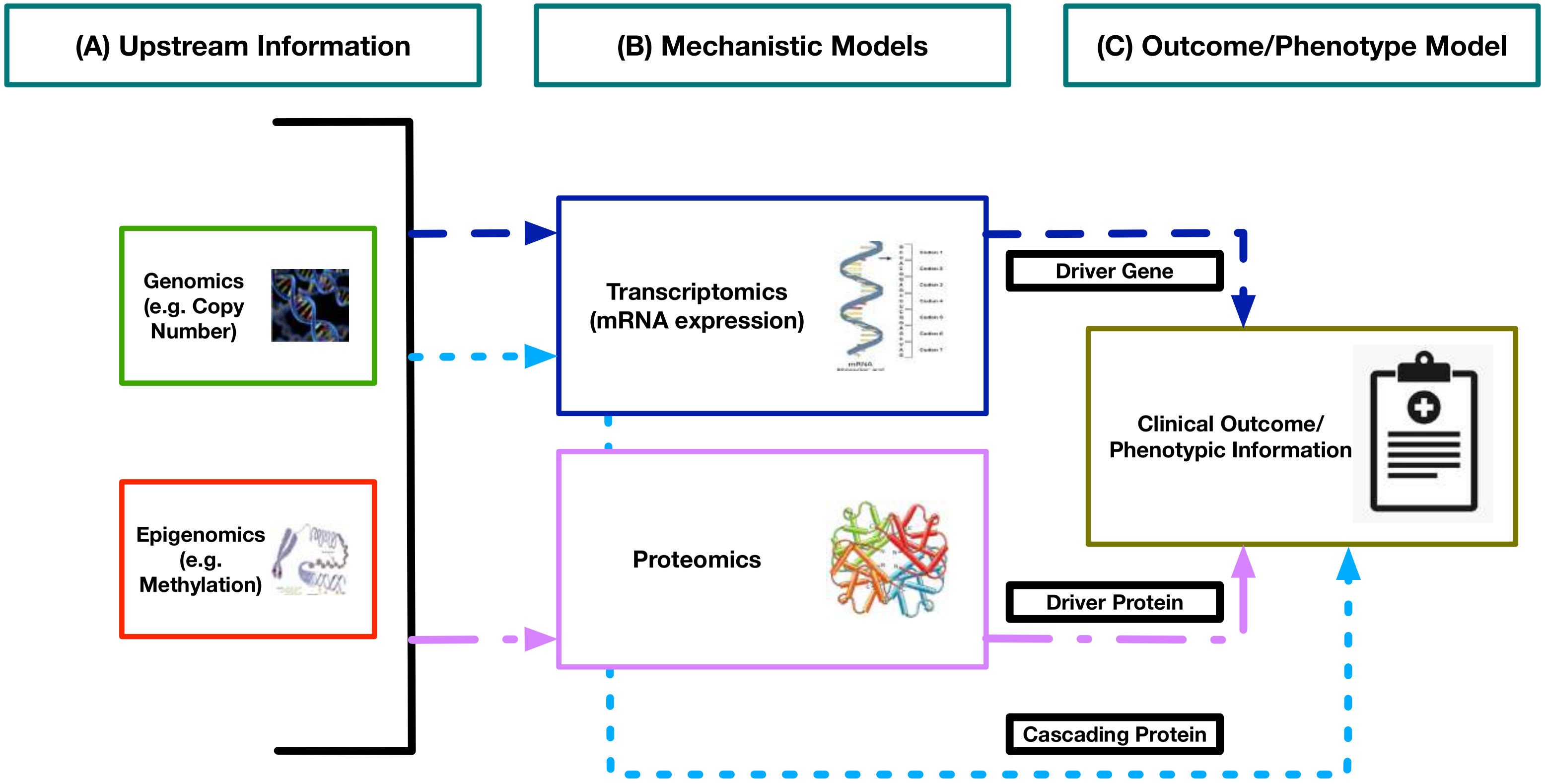

(A) Upstream Information
(B) Mechanistic Models
(C) Outcome/Phenotype Model
Genomics (e.g. Copy Number)
Epigenomics (e.g. Methylation)
Transcriptomics (mRNA expression)
Proteomics
Driver Gene
Driver Protein
Cascading Protein
Clinical Outcome/ Phenotypic Information

Figure 3.1: Conceptual schematic summarizing the fiBAG procedure. Panels (A), (B), and (C) respectively describe the upstream platforms, the mechanistic models, and the outcome model. The dashed, dotted-dashed, and dotted arrows indicate the regulatory paths for the driver gene, driver protein and cascading protein functional axes, respectively.

(2013b), and is computationally simpler, eliminating the needs of choosing the number of knots and incorporating penalization in a spline-type setting, such as the approach taken by Jennings et al. (2013). I calibrate the evidence summarized from these models to the Bayesian variable selection setting by proposing a generalized version of the spike-and-slab prior originally proposed by George and McCulloch (1997), termed the calibrated spike-and-slab prior. This calibrated prior structure improves the selection of the covariates by borrowing strength across multi-platform data, choosing to continuously up-weight prior inclusion probabilities of biomarkers in a data-driven manner. While I take the spike-and-slab route to build an adaptive and flexible mechanism to incorporate external knowledge in this work, other existing methods perform the incorporation of such *a priori* information in an outcome model via adaptive shrinkage, penalization, or some different prior structure. Using simulation studies under multiple synthetic and real data-based scenarios, I compare both the selection and estimation performances of our method against standard penalized regression (Tibshirani, 1996), grouped penalized regression (Boulesteix et al., 2017), and prior-informed selection (Velten and Huber, 2021; Zeng et al., 2021a) methods. Our method exhibits comparable selection and estimation performances with state-of-the-art methods for higher sample size to number of covariates ratios, and exhibits substantial improvement in performance over them for low-sample high-dimensional settings. I also offer computational flexibility using both Markov chain Monte Carlo (MCMC) implementation and a computationally efficient expectation-maximization based variable selection procedure (Ročková and George, 2014). I further perform pan-cancer integrative analyses of proteogenomic data with disease features previously unexplored in such settings.

Scientific innovation Multiple works from recent biostatistical literature have focused on incorporating existing evidence or external information into final models of interest via various approaches, such as data-adaptive shrinkage (Boonstra et al., 2015), adaptive Bayesian updates (Boonstra and Barbaro, 2020), or calibrated maximum-likelihood type procedures (Chatterjee et al., 2016). Our method is different from such approaches in the sense that it offers a framework to both learn de novo evidence within the pipeline and incorporate the said evidence into the final outcome models. As discussed before, omics elements at different hierarchical levels of the *gene expression pathway* may provide partly independent and complementary information (Buccitelli and Selbach, 2020). Our integrative approach allows learning such information across interconnected axes of functional acitivity such as DNA, RNA, and protein level quantifications. Additionally, our integrative analysis of pan-cancer proteogenomic data from the Cancer Genome Atlas utilizes both traditional *prognostic outcomes* (survival data) and recently developed *cellular descriptors* of cancer growth (stemness indices) to identify the proteogenomic signature driving such features, along with the molecular basis of these signatures. Cancer stem-like cells lead to sustained proliferation via resisting apoptosis, evading growth suppression, and exhibiting increased invasive and metastatic potential (Fulda, 2013; Adorno-Cruz et al., 2015). I look at the challenge of identifying the cellular molecular basis of the behavior of such stem-like cells, by using the mRNA-based stemness index (SI) proposed by Malta et al. (2018) as an outcome variable. From the survival and stemness outcome analyses across four common groups of cancers: Pan-gynecological, Pan-gastrointestinal, Pan-squamous and Pan-kidney, I identify both known and novel markers associated with the outcomes alongside insights on their functional roles. In particular, the genes RPS6KA1 (protein p90RSK) and YAP1 (proteins YAP and YAPPS127) are identified as top driver genes in the pan-gyne cancers, along with significant associations with the stemness outcome across multiple cancers in the group; both have been known to be crucial agents impacting gynecological cancers. Similarly, our analyses found the gene ERBB2 (protein HER2) to be positively associated with stemness for gastrointestinal cancers, supported by existing literature.

The rest of the chapter is organized as follows. Section 3.2 describe fiBAG both conceptually and with the mathematical details of the mechanistic and outcome models and the evidence calibration function, along with the computational steps behind the fitting, estimation, and selection procedures for the two models. Section 3.3 summarizes simulation studies in both synthetic and real data-based settings comparing our method to existing benchmarks. Section 3.4 summarizes results from our pan-cancer integrative proteogenomic analyses. I conclude the chapter with a discussion on the methodological and biological aspects of our work, along with some potential future directions in Section 3.5. All our results are available, alongside all the codes for our method and generating the plots, in an interactive R-Shiny dashboard, hosted at `https://bayesrx.shinyapps.io/Functional_iBAG/`.

3.2 fiBAG Model

3.2.1 Conceptual Factorization of the Multi-omics Model

I begin with the data structure and some notations. Let n denote the number of samples, q_g be the number of genes, and q_p be the number of proteins in the dataset of interest. The gene (mRNA) expression data matrix $\mathbf{G}$ and the protein expression matrix $\mathbf{P}$ respectively have dimensions $n \times q_g$ and $n \times q_p$. Further, let the data matrices corresponding to the upstream covariates copy number alteration and DNA methylation be denoted as $\mathbf{C}$ and $\mathbf{M}$, respectively, each having n rows. Let $\mathbf{Y}$ denote the $n \times 1$ outcome data vector. Thus, the proteogenomic, upstream, and outcome data available for a cohort of samples can be aligned vertically (i.e. matched across samples). I write the joint model of the outcome and the proteogenomic data conditional on the upstream data as

$$\mathcal{P}[\mathbf{Y}, \mathbf{G}, \mathbf{P} | \mathbf{C}, \mathbf{M}, \theta] = \underbrace{\mathcal{P}[\mathbf{Y} | \mathbf{G}, \mathbf{P}, \theta_Y]}_{Outcome} \underbrace{\mathcal{P}[\mathbf{G}, \mathbf{P} | \mathbf{C}, \mathbf{M}, \theta_F]}_{Mechanistic} . \tag{3.1}$$

Here, $\boldsymbol{\theta} = (\boldsymbol{\theta}_Y, \boldsymbol{\theta}_F)$ denotes a conceptual parameter (possibly multi-dimensional) that connects the two layers of models. The omics-only part ($\mathcal{P}[\mathbf{G}, \mathbf{P}|\mathbf{C}, \mathbf{M}, \boldsymbol{\theta}_F]$) represents the *mechanistic model*, which concerns the functional mechanisms of proteogenomic expression as driven by DNA-level cellular activities. Via the parameter $\boldsymbol{\theta}_F$, this mechanistic information is then learned and incorporated into the *outcome model* ($\mathcal{P}[\mathbf{Y}|\mathbf{G}, \mathbf{P}, \boldsymbol{\theta}_Y]$). The parameter vector $\boldsymbol{\theta}$ enables the mechanistic layer to inform the outcome layer, in line with our scientific aims of integrating functional information.

Biological rationale for the factorization The connection between the two layers drives potentially improved identification of proteogenomic features in the final outcome model. This conceptual framework aligns with the idea that different tumors, although potentially driven by changes in different agents or genomic locations, are controlled by a recurrent regulatory architecture where genomic alterations cluster upstream of functional proteins (master regulators) (Califano and Alvarez, 2017). The interconnections between these regulators form the tumor checkpoints that can potentially be useful as biomarkers and therapeutic targets. Further, epigenetic and genetic changes determining a cancer cell state can be intertwined, and the mutual dependencies between such traits can contribute to tumor progression via sequential layers of cellular activity (Alizadeh et al., 2015). Thus, it makes sense to model multi-platform omics data in a way where functional contributions to the variability in expressions of genes and proteins can be inferred separately and can be utilized to calibrate the identification of the roles of those genes and proteins in tumor progression. Over the next few subsections, I describe the specific approaches undertaken in this work to formulate these two model layers to utilize this biological framework.

3.2.2 Mechanistic Models

I are interested in *three axes of mechanistic information* (one for each gene and two for each protein) based on the available upstream data (Figure 3.1). I first describe the mathematical settings and the interpretations of the three axes. Let j denote the index for the specific proteogenomic biomarker of interest in a mechanistic model, with the understanding that it is a gene if $j \in \{1, \dots, q_g\}$, and a protein if $j \in \{q_g + 1, \dots, q_g + q_p\}$. For biomarker j and sample i, let the corresponding sub-vectors of $\mathbf{C}$ and $\mathbf{M}$ be denoted by $\mathbf{C}_{ij}$ and $\mathbf{M}_{ij}$, respectively. I now describe the general forms of the models corresponding to the three mechanistic axes.

$$\begin{aligned}
\textbf{Driver gene model:} \quad G_{ij} &= f_{1j}\big((\mathbf{C}_{ij}^T, \mathbf{M}_{ij}^T)^T\big) + e_{1ij}, \\
\textbf{Driver protein model:} \quad P_{ij} &= f_{2j}\big((\mathbf{C}_{ij}^T, \mathbf{M}_{ij}^T)^T\big) + e_{2ij}, \\
\textbf{Cascading Protein model:} \quad P_{ij} &= f_{3j}\big((G_{ij}, \mathbf{C}_{ij}^T, \mathbf{M}_{ij}^T)^T\big) + e_{3ij},
\end{aligned}$$

Here the $e_{\bullet}$s denote the error components. The upstream to gene correspondences are defined by the physical location of the coding segment of the gene in the genome. All copy repetitions within the gene frame and methylation sites across a window of ±500 kb are taken into account. The gene to protein correspondences are defined using which gene codes for which protein. The upstream to protein correspondences are defined by composing the previous two. The three models can be biologically interpreted in the following way.

1. **Driver gene:** Whether the regulations corresponding to the gene are unique at transcript levels, and whether there is a significant relationship between the genomic/epigenomic events and the resulting gene expression which, in turn, drives cancer progression and outcomes (Gevaert et al., 2013).

2. **Driver protein:** Whether the regulations corresponding to the protein are unique at transcript levels, and whether there is a significant relationship between the genomic/epigenomic events and the resulting protein expression which, in turn, drives cancer progression and outcomes.

3. **Cascading protein:** Whether the protein-specific regulations transit through multiple prior omics levels, i.e., whether there is a cascade of effects via the DNA → RNA → protein path (Song et al., 2019).

The middle column of Figure 3.1 describes the data structures as presented in the above equations. For each model, I are interested in a null hypothesis of the type $f_\bullet$ = constant. Compelling evidence against such hypotheses would provide evidence for the corresponding mechanistic activity being present for the gene/protein in question. For example, a strong level of evidence for the driver gene model would mean that the upstream DNA-level events impact the expression of the gene of interest significantly; the other models can be interpreted similarly. I now describe the characterization of these relationships via suitable modeling choices for the $f_\bullet$s, and the hypothesis testing setting to quantify the strength of evidence in the data.

Characterization of f via Gaussian processes The flexibility of our mechanistic model setting lies in the freedom in choosing the specific form of f. The simplest choice would be a linear function - which would lead to multiple linear regression models. These models, explored by Wang et al. (2013b), are easy to handle computationally and easy to interpret, since the regression parameters are explicitly available for inference (estimation/testing). However, they could miss potentially nonlinear associations prevalent across the multi-omics levels of cellular activity (Solvang et al., 2011; Litovkin et al., 2014). A more sophisticated choice of f, allowing such nonlinear associations would be to use a set of basis functions to describe f (such as using a spline model, as explored by Jennings et al. (2013)). Such models, however, require specifications of the knots based on a priori knowledge and demand additional penalization to obtain a stable fit, rendering the procedure to be more computationally intensive and less interpretable. To allow computationally tractable and interpretable identification of nonlinear associations while avoiding the need to specify knot locations over a multivariate domain, I use Gaussian process (GP) models. GPs have been utilized in the context of genomic data in past literature including modeling gene expression dynamics (Rattray et al., 2019), transcriptional regulations (Lawrence et al., 2007), and pathway analyses (Liu et al., 2007). Our simulation studies indicate that in scenarios with a high degree of non-linearity among the covariates in the generating model, GPs are better equipped in capturing significant associations than linear models (Section 3.3).

I now describe the GP specifics for the driver gene mechanistic model here; the other models can be expressed similarly. To recall, the j^{th} driver gene model is written as $G_{ij} = f_{1j}((\mathbf{C}_{ij}^T, \mathbf{M}_{ij}^T)^T) + e_{1ij}$, where I assume $e_{1ij} \overset{\text{iid}}{\sim} \text{N}(0, \tau_{1j}^2)$. Let us also denote $f_{1j}^{(i)} = f_{1j}((\mathbf{C}_{ij}^T, \mathbf{M}_{ij}^T)^T)$ for all i. Then, the GP prior on f_{1j} is placed as follows:

$$\begin{aligned} \textbf{GP prior:} \quad & (f_{1j}^{(1)}, \dots, f_{1j}^{(n)})^T \sim \mathbf{N}(\mathbf{0}, \mathbf{K}_{1j}), \\ \textbf{Covariance matrix:} \quad & \mathbf{K}_{1j(i,k)} = K_{1j}((\mathbf{C}_{ij}^T, \mathbf{M}_{ij}^T)^T, (\mathbf{C}_{kj}^T, \mathbf{M}_{kj}^T)^T), \\ \textbf{Kernel function:} \quad & K_{1j}(\mathbf{u}, \mathbf{v}) = g\tau_{1j}^2 \exp\big(- \frac{||\mathbf{u} - \mathbf{v}||^2}{\lambda_{1j}^2} \big). \end{aligned}$$

The hyperpriors specify $\tau_{1j}^2 \sim$ Inverse-Gamma$(\frac{\nu_{01j}}{2}, \frac{\nu_{01j}\tau_{01j}^2}{2})$ and $\lambda_{1j} \sim \exp(\lambda_{01j})$. For all our analyses, I set $g = n$. Although I use the standard squared exponential kernel, a common default choice (Micchelli et al., 2006), other kernels can be adopted as well.

Hypothesis tests for drivers and cascades via Bayes factors I describe our Bayesian hypothesis testing procedure for the mechanistic models using our driver gene models; the other models can be tested similarly. For the j^{th} driver gene model, I test $H_{0j} : f_{1j} =$ constant vs $H_{1j} : H_{0j}$ is false (equivalent to $H_{0j} : K_{1j}(\bullet, \bullet) \equiv 0$). I compute a log of the Bayes factor (lBF) corresponding to the comparison of the full model vs the null (mean) model to perform the test. Bayes factors (BFs) are particularly useful in this setup from both a statistical and a scientific point of view. From a methodological perspective, elegant almost sure convergence results for Bayes factors are available for rather general settings under standard assumptions (Chatterjee et al., 2020). Further, Bayes factors have been successfully utilized to quantify significance and compare model performances in omics models in past literature (Stephens and Balding, 2009). For our case, the final expression of the lBF for the driver gene model j, ignoring the constants a_n, b_n, c_n, and a (dependent on data dimensions and hyperparameters), is given below (other models can be derived similarly). The integral in Equation (3.2) is computed numerically. For any matrix $\mathbf{A}$, $\mathbf{A}_{\bullet j}$ denotes the j^{th} column of $\mathbf{A}$, and $\mathbf{A}_{i\bullet}$ denotes the i^{th} row of $\mathbf{A}$.

$$\begin{aligned} \mathrm{lBF}_{1j} = \Big[a_n + b_n \ln\Big(a + \sum_{i=1}^{n} G_{ij}^2 - \frac{(\sum_{i=1}^{n} G_{ij})^2}{c_n}\Big) \\ + \ln \int_0^\infty \exp(-\lambda_{01j}\lambda_{1j}) \frac{2^{b_n} |\mathbf{K}_{1j}/\tau_{1j}^2 + I|^{-\frac{1}{2}}}{\big\{\mathbf{G}_{\bullet j}^T(\mathbf{K}_{1j}/\tau_{1j}^2 + I)\mathbf{G}_{\bullet j} + a)\big\}^{b_n}} d\lambda_{1j}\Big] / \ln(10). \end{aligned} \tag{3.2}$$

I decide the strength of the evidence posed by the lBF from a mechanistic model using the following standard significance ranges: < 0.5 (no evidence), $0.5 - 1$ (substantial), $1 - 2$ (strong), and > 2 (decisive) (Kass and Raftery, 1995). For each gene, I have one lBF from the driver gene model, whereas for each protein, I have a maximum of two such lBFs from the driver and cascading protein models. I now describe our approach to calibrate these evidence quantities into the variable selection models for the outcomes of interest.

3.2.3 Calibrated Bayesian Variable Selection

Functional information is captured by the lBFs from the mechanistic models - each lBF summarizes the strength of evidence for the functional role of the corresponding gene/protein.

The lBF metric is particularly useful for two reasons - first, it provides us with an interpretable, unidirectional and continuous scale of evidence strength, and second, across a large proteogenomic panel, it provides a highly parallelizable procedure to gather scalar evidence of functional relevance which are useful numerical quantities in their own terms and pertinent prior information for future models, such as our outcome model.

Thus, following Equation (3.1), I want to incorporate this mechanistic information into our final outcome model. I pose this problem in context of a general regression framework where some quantitative summaries of covariate importance are available beforehand, and such information is to be combined with the mechanics of a typical variable selection setting. Generally, I denote such prior information as $\mathcal{E}$, with $\mathcal{E}_j$ denoting the possibly multidimensional evidence summary for covariate j (i.e., lBF for gene/protein j in our case). I intend to inform the final outcome model using these evidences in terms of selection/non-selection of the predictors. Specifically, if there is sufficiently *strong evidence* in favor of a covariate, I want to *up-weight its probability of inclusion*. Otherwise, I want to put a uniform probability on selection/non-selection for that particular covariate. To achieve this, I utilize a hierarchical Bayesian framework with spike-and-slab priors for each effect, with the spike probabilities calibrated using the evidence available. The rest of this subsection describes the components of our fiBAG outcome model, called calibrated Bayesian variable selection, or cBVS in short.

Notations Following notations introduced at the beginning of this section, let Y_i denote the outcome for individual i. For the purpose of exposition, I assume that the outcome of interest is continuous, and each Y_i is assumed to be observed. Generalizations to censored or categorical outcomes are straightforward. Let us denote the design matrix corresponding to any additional covariates (other than genes and proteins, such as clinical information or demographics) by $\mathbf{B}_{n\times q_b}$. The combined design matrix will have a dimension $p = 1 + q_b + q_g + q_p$. I then propose a hierarchical Bayesian outcome model, as described below.

$$\begin{aligned} Y_i &= \beta_0 + \mathbf{B}_{i\bullet}^T \boldsymbol{\beta}_B + \mathbf{G}_{i\bullet}^T \boldsymbol{\beta}_G + \mathbf{P}_{i\bullet}^T \boldsymbol{\beta}_P + \epsilon_i, \quad \forall i \in \{1, ..., n\}, \\ \epsilon_i &\overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2), \quad \forall i \in \{1, ..., n\}. \end{aligned}$$

Let $\boldsymbol{\beta}_{p\times 1}$ denote the complete regression coefficient vector. I set up a hierarchical calibrated spike-and-slab prior on its last $q_g + q_p$ components, as described below. A standard conjugate prior is put on the residual variance parameter as $\sigma^2 \sim \text{Inverse-Gamma}(\frac{\nu}{2}, \frac{\nu\lambda}{2})$.

$$\begin{aligned} (\beta_0, \boldsymbol{\beta}_B^T, \boldsymbol{\beta}_G^T, \boldsymbol{\beta}_P^T)^T = \boldsymbol{\beta} | \boldsymbol{\gamma}_{q_g+q_p\times 1}, \sigma &\sim \text{Normal}(\mathbf{0}, \mathbf{D}_{\gamma,\sigma}), \\ \gamma_j | \omega_j &\sim \text{Bernoulli}(\omega_j), \quad \forall j \in \{1, ..., q_g + q_p\}, \\ \omega_j &\sim \text{Beta}\Big(\mathcal{F}(\boldsymbol{\mathcal{E}}_j), \frac{1}{\mathcal{F}(\boldsymbol{\mathcal{E}}_j)}\Big), \quad \forall j \in \{1, ..., q_g + q_p\}, \end{aligned}$$

where $\mathbf{D}_{\gamma,\sigma} = \sigma^2 \mathbf{A}_\gamma, \mathbf{A}_{\gamma\ p\times p} = \text{diag}\{v_1 \mathbf{1}_{1+q_b}, \gamma_1 v_1 + (1-\gamma_1)v_0, \dots, \gamma_{q_g+q_p} v_1 + (1-\gamma_{q_g+q_p})v_0\}$ where $v_1 \geq v_0 > 0$ are respectively the slab and spike variances, and where $\mathcal{F}$ is a calibration function mapping the evidence $\boldsymbol{\mathcal{E}}_j$ to the prior inclusion probability ω_j.

Calibration functions for external information The function $\mathcal{F}$ in the outcome model maps the functional evidence from the mechanistic models to a parameter scale to inform the Beta hyperprior for selection of each covariate. The specific mathematical properties and form of this function depends on the user-specific context of what type, range, and scale of functional evidence is being used. Further, it is possible to garner multiple quantities of evidence from different sources for each covariate of interest, and depending on the scenario, the calibration function may take inputs in $\mathbb{R}^K$ for some $K > 1$.

For the suitability of exposition, I describe a specific form of the mapping function used for our analyses where it is assumed that the evidence measures $\mathcal{E}_j$ are in the form of scalar summary statistics s_j. As explained in the previous subsection, I have a single evidence quantity (the lBF from the driver gene model) for each gene and a maximum of two evidence quantities (lBFs from the driver and cascading protein models) for each protein. For genes (i.e., $j \in \{1, \dots, q_g\}$), I set $s_j = \text{lBF}_{1j}$, and for proteins (i.e., $j \in \{q_g + 1, \dots, q_g + q_p\}$), I set $s_j = \max(\text{lBF}_{2j}, \text{lBF}_{3j})$. The choice of taking the maximum of the two available lBFs for a protein is justified because the incorporation of this information in the outcome model is intended to improve selection of covariates that are important at a cellular functional level and the specific source of such functional activity is immaterial in the final model. Our calibration function then needs to be a map from $\mathbb{R} \rightarrow \mathbb{R}$ that can reflect the changes in the strength of evidence in the lBFs in the desired ranges.

For lBFs, the strength of evidence is unidirectional (i.e., higher evidence quantity implies higher strength of the evidence). In such a scenario, a mapping function is required to be non-decreasing. Such a mapping function can either be discrete or continuous, and the jumps of the discrete curve or the slope of the continuous curve can be tuned depending on the importance to be put on different scalar values of the evidence. In particular, I follow the standard lBF significance ranges as described before to a parameter describing a Beta prior distribution for the covariate-inclusion probabilities ω_j. The shape of the lBF to Beta prior mean curve and the representative densities for some lBF values of interest in each of the standard lBF significance ranges (< 0.5 (no evidence), $0.5 - 1$ (substantial), $1 - 2$ (strong), and > 2 (decisive)) for the calibration function used in this work are summarized in Figure 3.2. Effectively, I build a function having the following properties.

- If the mechanistic models *do not provide* us with *enough evidence* regarding the functional utility of a gene/protein, I put a *uniform prior* (mean prior probability $= 0.5$) for the corresponding selection probability parameter in the outcome model. This is illustrated by the leftmost point on the X axis of Figure 3.2A.
- If the mechanistic models provide us with *strong evidence*, on the other hand, I put a

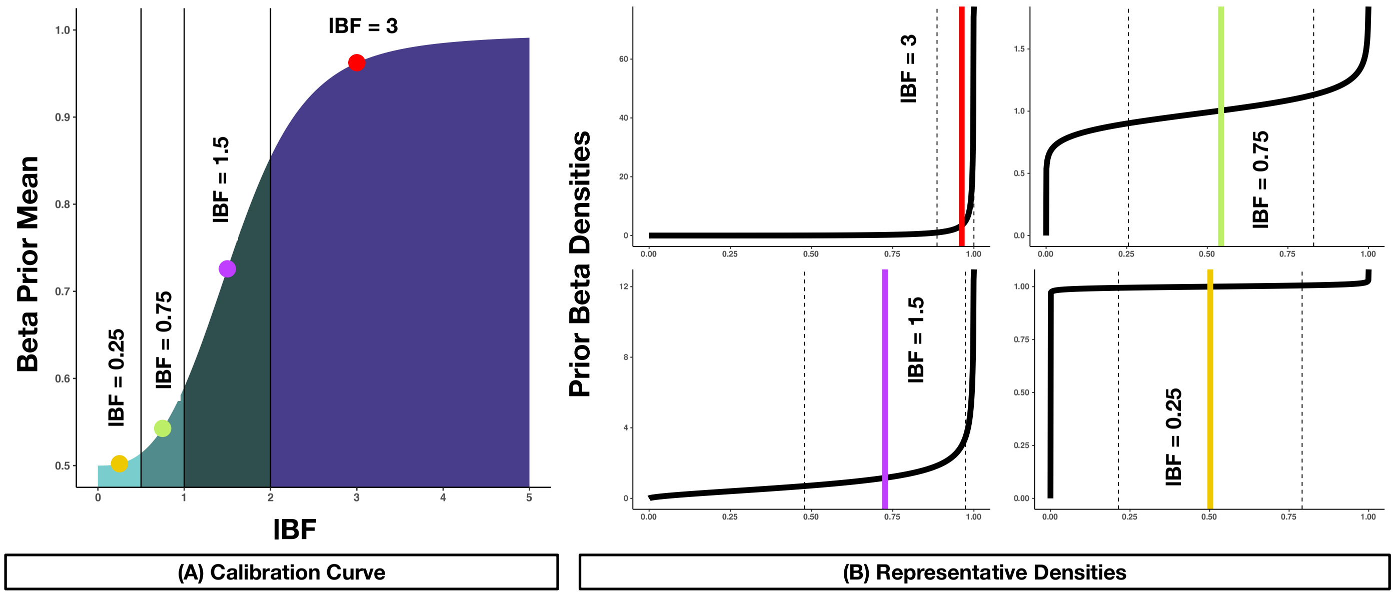


Figure 3.2: Calibration function for the outcome model. Panel (A) plots the calibrated prior Beta means against the lBFs. Panel (B) presents the densities (means indicated by solid vertical lines, broken vertical lines indicate the ±1 sd ranges around the means) for four representative values from the four ranges of interest in the x-axis of panel (A).

strong prior with a large prior mean on the selection probability parameter. This is illustrated by the increment in the Y axis of Figure 3.2A.

This construction enables the functional evidence to inform selection/estimation in the outcome model, while retaining the flexibility to still allow the model to ignore prior evidence, if necessary, for the proteogenomic markers without any observed association with the outcome of interest, as are illustrated by our pan-cancer applications presented in Section 3.4.

I perform the following steps to obtain the shape parameters of the final Beta prior as $\mathcal{F}(\mathcal{E}_j) = \mathcal{F}(s_j)$ and $1/\mathcal{F}(\mathcal{E}_j) = 1/\mathcal{F}(s_j)$. Note that thus the prior mean for ω_j takes the form $\mathcal{F}(s_j)\big\{\mathcal{F}(s_j)+1/\mathcal{F}(s_j)\big\}^{-1}$, and the corresponding variance is $\big[\big\{\mathcal{F}(s_j)+1/\mathcal{F}(s_j)\big\}^2\big\{\mathcal{F}(s_j)+1/\mathcal{F}(s_j)+1\big\}\big]^{-1}$. I first ensure that all small positive and non-positive lBFs from the mechanistic model are effectively truncated to zero (no evidence, to be mapped to a uniform prior), by setting $s_j^* = \max(s_j, 10^{-6})$. Then, our mapping function evaluates $\mathcal{G}(s_j) = \frac{1}{2}\big[\big\{1+(s_j^*/3)^{-2.75}\big\}^{-1}+1\big]$. Finally, I compute $\mathcal{F}(s_j) = (2\mathcal{G}(s_j))^4 = 16\mathcal{G}^4(s_j)$ to ensure sharp increase in prior mean probability of inclusion when lBF increases from 1 to 2. Effectively, $\mathcal{F}(\bullet)$ is a composition of three functions.

- A four-parameter logistic ($s_j^* \rightarrow \dfrac{1}{1+(s_j^*/3)^{-2.75}}$) allowing us to control the shape and scale of the prior probability calibration curve along with its asymptotic behavior and lower bound.

- A linear map ($x \rightarrow \frac{1}{2}(x-1)$) that takes [0, 1] to [0.5, 1] enabling us to make the prior noninformative when the lBF is small or negative.

- A polynomial map ($c \rightarrow (2c)^4$) allowing additional control over the steepness of the resulting prior probability calibrated means.

The tuning parameters with values 3 and 2.75 are chosen via computational checks across ranges of $[2, 4]$ and $[2, 3]$, respectively. The calibration functions for the other parameter values are presented in Figure S3.1-S3.14. For this specification, the prior mean probability of inclusion for a covariate ranges from 0.5 (when the lBF is 0 or negative) to approximately 1 (when the lBF is large, say > 5). The calibrated prior mean probabilities for some representative points from these intervals are as follows, with respect to Figure 3.2B: 0.502 (lBF = 0.25, bottom-right), 0.543 (lBF = 0.75, top-right), 0.726 (lBF = 1.5, bottom-left), 0.962 (lBF = 3, top-left).

Benefits and utilities of calibration The benefits of this calibrated model formulation are multifold. First, by modifying the calibration function $\mathcal{F}$, our framework allows the user to incorporate any form of model summary information (such as z-scores, p-values, etc.) into the outcome model. Unlike existing grouped shrinkage-based procedures, this eliminates the restriction of only using external information through categorical covariates. Second, unlike the shrinkage-based or Bayesian approaches where the external covariates directly inform the final model, our model only relies on the supply of the summary statistics s_j - rendering the computations less challenging and allowing the use of any upstream dataset, however large, in learning the functional information from multiple categorical/continuous sources. In the same vein, any such calibration function can easily be adapted to a scenario where multiple sources of evidence are available for a single covariate rather than just a single summary statistic s_j. For example, if there are K lBFs ($\mathcal{E}_j = (\mathrm{lBF}_{1j}, \ldots, \mathrm{lBF}_{Kj})$) available from as many mechanistic models for the jth biomarker, then I can think of the calibration function as a composition: $\mathcal{F}(\mathcal{E}_j) = \mathcal{F}_1(\mathcal{F}_2(\mathcal{E}_j))$, where $\mathcal{F}_1$ can be similar to the $\mathcal{F}$ I use, and $\mathcal{F}_2 : \mathbb{R}^E \to \mathbb{R}$ aggregates the multiple lines of evidence to a single scalar value s_j. One possible choice for this is a linear map, as $s_j = \mathcal{F}_2(\mathcal{E}_j) = \sum_{k=1}^{K} \alpha_{kj} \mathrm{lBF}_{kj}$. Here α_{kj}s are convex weights specific to biomarker j, interpreted as quantifications of the importance of each source of evidence. Several choices of the α_{kj}s are possible, as described below.

1. **Average evidence:** $\alpha_{kj} = 1/K$ (takes a simple average of all available evidences).

2. **Maximal evidence:** $\alpha_{kj} = I\big(\mathrm{lBF}_{kj} = \max_{k' \in \{1,\ldots,K\}} \mathrm{lBF}_{k'j}\big)$ (only takes into account the strongest evidence available from any source). This is akin to the approach I used for the proteomic markers with two mechanistic models.

3. **Precision-weighted evidence:** $\alpha_{kj} = \rho_{kj} / \sum_{k=1}^{K} \rho_{kj}$ (weights the evidences by some metric of reliability of the evidences, such as $\rho_{kj} = \hat{\tau}_{kj}^{-2}$ where $\hat{\tau}_{kj}^{2}$ is the estimated noise variance for the source mechanistic model of lBF_{kj}).

This provides an additional layer of flexibility that allows the user to choose how many sources of evidence to use and how best to combine them.

Parameters of interest and model dependence structure Having described the mechanistic and outcome layers of the modeling framework and established the connection between them via the evidence calibration function, I now revisit Equation (3.1) to interpret the conceptual parameters in context of our modeling scheme. For the mechanistic models, the quantities $\boldsymbol{\theta}_F$ represent the parameter vector for the Gaussian process models, namely, $\{(\tau_{lj}, \lambda_{lj}) : l$ varies across each mechanistic model available for each gene/protein $j\}$. The quantification of evidence is not performed via direct estimation of these parameters but via computing and calibrating the Bayes factors corresponding to each model. Therefore, one component of $\boldsymbol{\theta}_Y$ is driven by $\boldsymbol{\theta}_F$ via the evidence calibration. The rest of $\boldsymbol{\theta}_Y$ is specified by $(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\omega}^T, \sigma)^T$, i.e. the parameters of interest from the outcome model. The inferential task then is to fit the outcome model, estimate the parameters of interest $(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\omega}^T, \sigma)^T$, and perform covariate selection based on the posterior distribution of $(\boldsymbol{\gamma}^T, \boldsymbol{\omega}^T)^T$. The overall dependence structures in the mechanistic and outcome model settings are summarized in Figure 3.3. In the next section, I describe our model-fitting procedures.

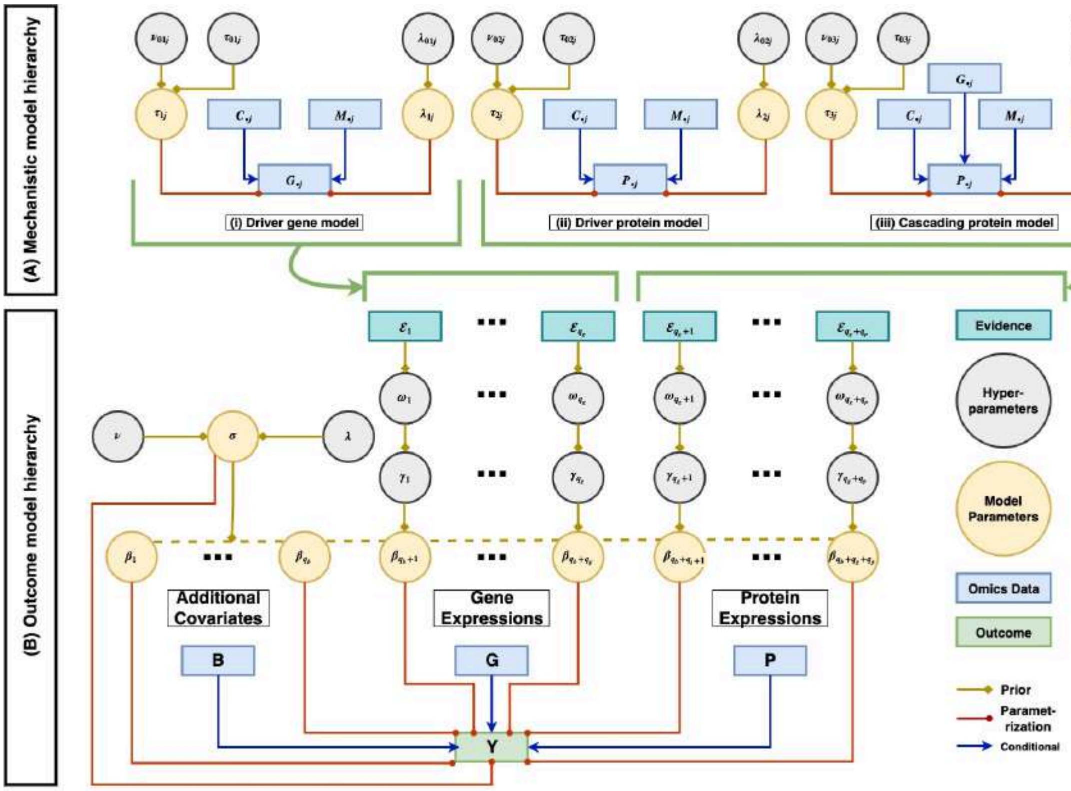


Figure 3.3: Plate diagram summarizing the dependence structure across the variables, parameters, and hyperparameters in fiBAG. Panels (A) and (B) respectively summarize the mechanistic and outcome models. **C**, **M**, **B**, **G**, and **P** each have n rows and are matched across samples, along with the outcome vector **Y**. **B**, **G**, and **P** respectively have q_b, q_g and q_p columns. **C** and **M** respectively represent copy number and methylation data. All the parameters and data structures are described in full detail in Section 3.2.

Generalizations to non-Gaussian outcomes Such generalizations typically involve minimal changes to the outcome model. As an example, I discuss an extension to survival outcomes briefly. Let us denote $Y_i = \log T_i$ as the possibly unobserved log-survival time and C_i as the possibly unobserved censoring time for individual i. The observed response for individual i is (Z_i, δ_i), where $Z_i = \min\{\log T_i, \log C_i\}$ and $\delta_i = I(T_i < C_i)$. Under the assumptions that the censoring distribution is free of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and that the censoring times are independent of the true outcomes conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, changing the outcome model from Gaussian to an accelerated failure time model with log-Normal outcomes only introduces truncations of the censored (unobserved) outcomes in the log-posterior. Therefore, the only change in the model fitting procedure occurs in the expression of the full log-posterior.

3.2.4 Model Fitting and Parameter Estimation

Mechanistic model fitting As described in Section 3.2.2, I are interested in quantifying the functional evidence for each gene/protein via the GP-based mechanistic model using the lBF - therefore, I do not require a full Bayesian exploration of the model. To ensure computational efficiency, I directly compute the lBF following expressions as in Equation (3.2). This requires the evaluation of an integral, which I perform numerically.

Outcome model fitting As described in the previous subsection, the complete set of parameters to be estimated by cBVS is $(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\omega}^T, \sigma)^T$. $\boldsymbol{\beta}^T$ provides estimates of the effect sizes of the proteogenomic covariates on the outcome, and the $(\boldsymbol{\gamma}^T, \boldsymbol{\omega}^T)^T$ guides their selection. The parameter estimation is focused on the posterior of $(\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\omega}^T, \sigma)^T$. For large proteogenomic panels, this posterior will be computationally resource-intensive to directly sample from. The question, therefore, is of a trade-off between computational simplicity and estimation accuracy. Due to this reason, I offer three implementations of cBVS in increasing order of computational efficiency – a standard Markov chain Monte Carlo (MCMC) using

Gibbs sampler to sample from the complete posterior, a selection-only MCMC to sample from the marginal posterior of $\boldsymbol{\gamma}$, and an expectation-maximization based variable selection (EMVS) procedure to approximate the posterior modes of the parameters. Briefly, the selection-only MCMC focuses on estimating $\boldsymbol{\gamma}$ first and then estimates $\boldsymbol{\beta}$ using a Bayesian model averaging-type procedure (Hinne et al., 2020), and the EMVS sacrifices the full posterior along with error estimates to achieve fast point estimation (Ročková and George, 2014). The exact details of each are as follows.

1. **Full MCMC:** The simplest approach is to perform a complete Markov chain Monte Carlo (MCMC) procedure to sample from the joint posterior – I utilize a Gibbs sampler for this. For both continuous and survival outcomes, I use the `rjags` (Plummer et al., 2016) package in R. The rjags model descriptions are available in our shiny app hosted at `https://bayesrx.shinyapps.io/Functional_iBAG/`.

2. **Selection-only MCMC:** One way to mitigate the time- and resource-intensiveness of the full MCMC is to focus on $\boldsymbol{\gamma}$ only - by integrating $(\boldsymbol{\beta}^T, \boldsymbol{\omega}^T, \sigma)^T$ out of the joint posterior. This results in $\Pi(\boldsymbol{\gamma}|\text{Data, Hyperparameters})$, to be then approximated via MCMC. The final estimates for $\boldsymbol{\omega}$ are computed by taking the average of the traversed MCMC path for $\boldsymbol{\gamma}$ post burn-in. The components of $\boldsymbol{\beta}$ are estimated using a Bayesian model averaging-type procedure (Hinne et al., 2020), taking a convex combination of draws from their conditional posterior at every step, weighted proportionally to the negative log posterior evaluated at that step. Since this MCMC only performs a search across a lattice space of size 2^p, it results in significant improvements in computation times. The overview of the selection-only MCMC procedure is presented in Algorithm 3.1. The codes are made available via our shiny app hosted at `https://bayesrx.shinyapps.io/Functional_iBAG/`.

3. **E-M based variable selection (EMVS):** The fastest alternative to a full MCMC is to sacrifice approximating the full posterior (along with error margins) and to only focus on point estimates of the parameters of interest. For this purpose, I adapt the EMVS procedure by Ročková and George (2014) for our continuous and survival settings. The EMVS procedure estimates the posterior mode instead of approximating the full posterior, resulting in faster iterations. The model codes for the implementation of EMVS in our setting are available in our shiny app hosted at `https://bayesrx.shinyapps.io/Functional_iBAG/`.

Algorithm 3.1. Selection-only MCMC algorithm

0. Initiate the MCMC with $\boldsymbol{\gamma} = \boldsymbol{\gamma}_{\text{old}}$, where $\gamma_{\text{old},j} \sim Ber(\frac{a(s_j)}{a(s_j) + \frac{1}{a(s_j)}}), \forall j$.

1. With probability $1/3$ each, select one of the following steps and perform the corresponding task.

   (a) **Add:** Randomly pick one of the $\boldsymbol{\gamma}_{\text{old}}$ indices with value 0, change it to 1, and call the resulting vector $\boldsymbol{\gamma}_{\text{new}}$. If there is no index j where $\gamma_{\text{old},j} = 0$, perform 1.2.

   (b) **Delete:** Randomly pick one of the $\boldsymbol{\gamma}_{\text{old}}$ indices with value 1, change it to 0, and call the resulting vector $\boldsymbol{\gamma}_{\text{new}}$. If there is no index j where $\gamma_{\text{old},j} = 1$, perform 1.1.

   (c) **Swap:** Randomly pick one of the $\boldsymbol{\gamma}_{\text{old}}$ indices with value 0, and independently, randomly pick one of the $\boldsymbol{\gamma}_{\text{old}}$ indices with value 1. Swap the values of these two indices, and call the resulting vector $\boldsymbol{\gamma}_{\text{new}}$. If there is no index j where $\gamma_{\text{old},j} = 1$, perform 1.1. If there is no index j where $\gamma_{\text{old},j} = 0$, perform 1.2.

2. Compute $\log(p^*) := \min\{0, \log(\Pi(\boldsymbol{\gamma}_{\text{new}}|.)) - \log(\Pi(\boldsymbol{\gamma}_{\text{old}}|.))\}$. The proposed $\boldsymbol{\gamma}_{\text{new}}$ is then accepted with probability p^*.

3. Iterate between 1-2 until convergence.

Model summaries The MCMC/EMVS procedures provide us with posterior inclusion probabilities (PIPs) $\hat{\omega}_j$ and regression coefficient estimates $\hat{\beta}_j$ for each covariate in the model. A cut-off on the PIPs is computed using a false discovery rate adjustment procedure at a specified level of significance, treating the $1 - \hat{\omega}_j$ as p-value type quantities. Suppose the estimated posterior probabilities of inclusion are denoted by $\{\hat{\omega}_j, j \in \{1, \dots, q_g + q_p\}\}$. Then, I define p-value type quantities $p_j = 1 - \hat{\omega}_j, \forall j \in \{1, \dots,_g +q_p\}$, and sort them in the increasing order of magnitude as $\{p_j^*, j \in \{1, \dots,_g +q_p\}\}$, with the understanding that for each j, $p_j^* = p_{k_j}$ for some $k_j \in \{1, \dots,_g +q_p\}$. Let the cumulative sums of these ordered quantities be denoted by $r_j = \sum_{l=1}^{j} p_j^*$ for each j. Let $j^* := \min\{j : r_j \geq \alpha\}$, where α is a pre-specified level for the false discovery rate control. Then I infer that the covariates with indices $k_1, \dots, k_{j^*}$ are selected in the outcome model.

I now have all the computational tools to implement the fiBAG framework. In the next two sections, I respectively describe our simulation studies and our pan-cancer proteogenomic analyses using fiBAG.

3.3 Simulation Studies

To illustrate the utilities of cBVS, I performed two simulation studies. *Simulation 1* deals with continuous outcomes in synthetic datasets, comparing metrics of selection and estimation from the cBVS procedure with existing benchmarks across a class of variable selection methods. I include both penalized/grouped penalized selection procedures and Bayesian prior-based selection procedures as benchmarks. Simulation 1 is expected to assess cBVS for both low and high sample size to dimension ratios and quantify the improvement and/or preservation of performance along this spectrum compared to the benchmarks. In *Simulation 2*, the data generation procedure is informed by the patient datasets for breast invasive carcinoma from the Cancer Genome Atlas. The next two subsections describe the design, settings, and results from these two simulation studies.

3.3.1 Simulation 1: Synthetic Data-based Simulations

Data generation and choices of true effects To compare performances across a grid of varying sample size/number of covariates (n/p) ratios, I fix the number of proteogenomic biomarkers generated in the simulated datasets at $p = 200$ and vary the sample size across $n = 50, 100, 200, 400, 800$, covering a range of $n/p = 1/4$ to 4. For simplicity, I assume that there is only one upstream covariate for each biomarker. 100 replicates are generated for each n. I follow the steps described below to generate one such replicate.

- Upstream covariates $\boldsymbol{U}_{n\times p}$ (comparable to copy number/methylation data in the biological setting) are generated such that each $U_{ij} \overset{\text{iid}}{\sim} \text{N}(0, 1)$.

- Let $\boldsymbol{X}_{n\times p}$ denote the design matrix of the proteogenomic expression data for the outcome model. Then the generating model for the j^{th} expression is $X_{ij} \overset{\text{ind}}{\sim} \text{N}(\xi_j U_{ij}, 1)$, $\forall i \in \{1, \ldots, n\}$. The mechanistic effect size ξ_j controls the level of evidence reflected by an lBF for the mechanistic model of $\boldsymbol{X}_{\bullet j}$ on $\boldsymbol{U}_{\bullet j}$. The correspondence between values of ξ_j and the four levels of lBF: no evidence, substantial, strong, and decisive, is established numerically.

  Among the 200 $\boldsymbol{X}_{\bullet j}$s, the first 60 are distributed into four groups of 15 each, in the order of no evidence, substantial, strong, and decisive, followed by two groups of five at the levels strong and decisive, respectively. Those first 60 are distributed further into varying outcome effect sizes in the later steps to cover the complete range of combinations of prior evidence and effect size. The latter two groups will be assigned no true effects to include an in-built checkpoint for false positive evidences.

- Among each group of 15 for the four levels of prior functional evidence, I put three groups of five $\boldsymbol{X}_{\bullet j}$s with effect sizes β_js generated respectively at the low ($U(0, 0.2)$), medium ($U(0.4, 0.6)$), and high ($U(0.9, 1.1)$) levels. All the other β_js are set $= 0$.

  This results in 12 groups of covariates of size five each, because there are 12 possible combinations from the three levels of effect sizes and the four levels of evidences. Additionally, I have two groups of size five each with strong/decisive level of evidence but no true effect. All other 130 covariates have no evidence and no true effect.

- Finally, I generate the continuous outcome data as $\mathbf{Y}_i \overset{\text{ind}}{\sim} \mathcal{N}_p(\boldsymbol{\beta}^T \mathbf{X}_{i\bullet}, 1)$.

Brief overview of benchmark methods I compare the performance of the cBVS model against three classes of methods that perform variable selection based on different approaches. I use one Bayesian variable selection method without any external evidence, namely, the expectation-maximization based variable selection (EMVS) (Ročková and George, 2014) - as an uncalibrated counterpart to cBVS. I also include two penalized selection procedures - additional to LASSO (Tibshirani, 1996), I include a grouped penalized selection procedure termed Integrative LASSO with Penalty Factors (IPF-LASSO) (Boulesteix et al., 2017). The reason behind including a grouped penalized regression procedure is that the prior evidence levels provide a natural grouping for the covariates to be used in a variable selection setting. I choose IPF-LASSO in particular since it has previously shown promise in integrative omics-based personalized medicine context. Finally, I implement two recently proposed Bayesian variable selection methods incorporating external information, namely, graper (Velten and Huber, 2021), which incorporates categorical external covariates only, and xtune (Zeng et al., 2021a), which can handle continuous covariates as well. Our simulation scheme (i) allows a comprehensive comparison of calibrated vs non-calibrated methods and (ii) provided a benchmark for calibration to evaluate cBVS.

Summary of metrics used Each method provides a coefficient estimate $\hat{\boldsymbol{\beta}}$. For LASSO, I get a $\hat{\boldsymbol{\gamma}}$ corresponding to the best-fit λ. For each of the other methods, I compute $\hat{\boldsymbol{\omega}}$, on which I use the FDR-based adjustment method as described in Section 3.2.4 to infer selection. I compute several standard metrics of selection performance, namely, area under the receiver operating characteristic curve (AUC), scaled AUC between 0.8 to 1 specificity (AUC20), true positive rate (TPR), false positive rate (FPR), and Matthew's correlation coefficient

(MCC). The use of MCC is particularly useful since at a specified selection threshold it aggregates performance summary from both the TPR and FPR metrics. Using AUC and AUC20 allows threshold-free evaluation of selection performance.

Results and discussion The results from Simulation 1 are summarized in Figure 3.4A. In terms of AUC, our calibrated outcome regression model performs the best for the smallest n/p ratio $= 1/4$, and is only second to graper by a very small margin for the n/p ratios $= 2, 4$. In terms of MCC, our calibrated outcome regression model is only second to xtune for n/p ratios $= 1/4, 1/2$, and is only behind graper and xtune for the n/p ratios $= 1, 2$, and 4. In terms of R^2, our calibrated outcome regression model performs the best in both ends of the n/p ratio spectrum, i.e., $n/p = 1/4, 1/2, 4$, and is marginally behind graper and xtune for the n/p ratios $= 1, 2$. These summaries indicate several utilities of the calibrated outcome regression procedure. First of all, based on the metrics summarized, calibration based on prior evidence seems to have an evident benefit, as all the methods utilizing prior evidence (cBVS, graper, xtune) outperform those not incorporating any prior evidence (EMVS, LASSO, IPF-LASSO) across all the n/p ratios used and across all metrics. Further, as evident from the AUC summaries, cBVS offers an improved preservation of selection performance for n/p ratios < 1 over the methods incorporating external covariates, while maintaining a comparable and satisfactory level of performance in the more favorable scenarios, i.e., $n/p > 1$, as well. This is particularly reassuring, since unlike MCC or other standard metrics of selection, AUC is a threshold-free summary which enables us to sum up the performance of the methods across the whole spectrum of lenient to stringent selection criteria.

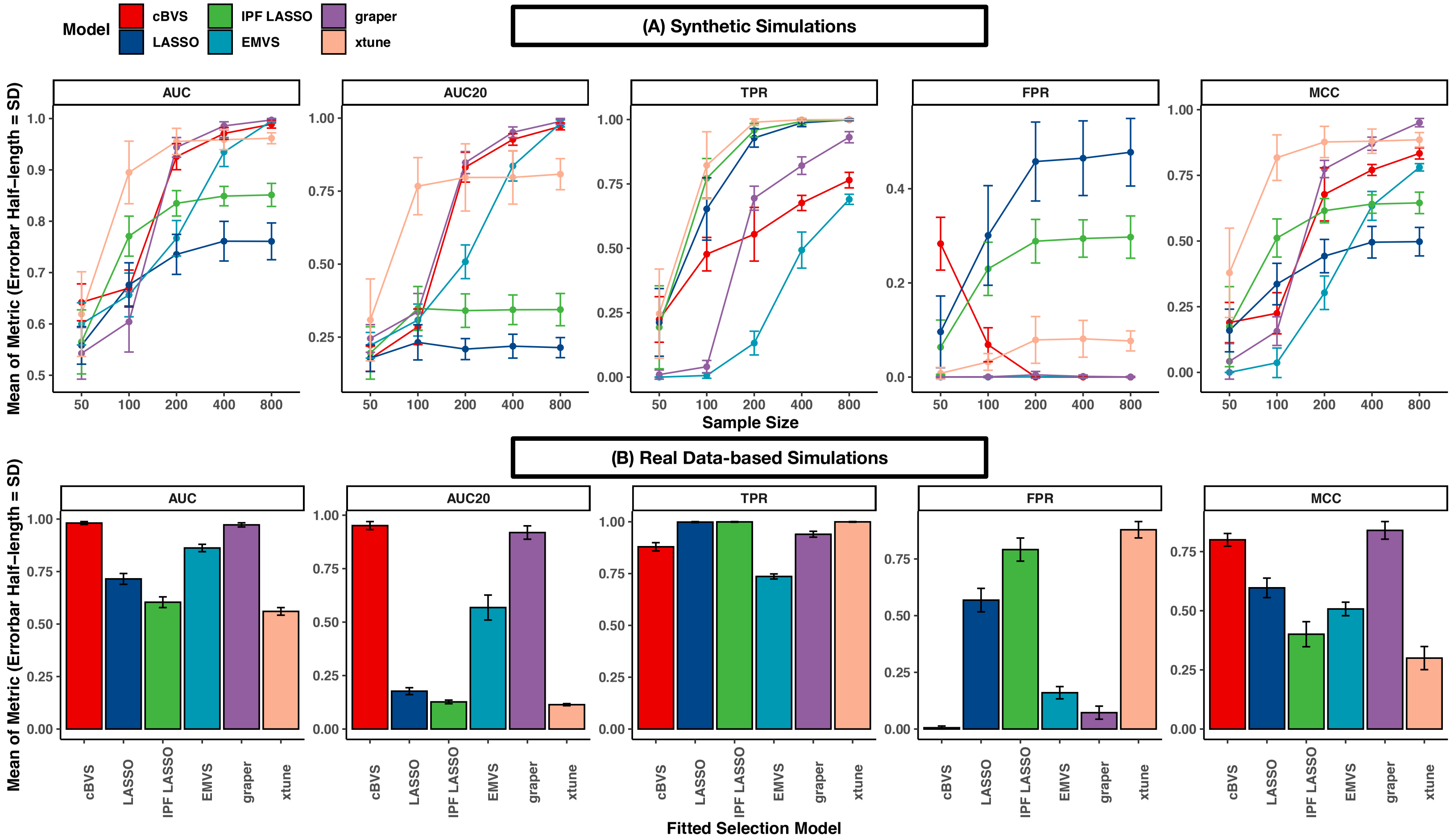


Figure 3.4: Summary of results from (A) synthetic and (B) real data-based simulations with continuous response.For (A), 100 replicates were generated for each sample size. The number of covariates was 200, with 60 of them having non-zero true effect sizes, mixed across a grid of low, medium, and high values, each with equal proportions of lBFs at the levels no evidence, substantial, strong, and decisive. For (B), 100 replicates were generated as well. Covariate matrix, effect sizes, and calibrated evidence were picked from a calibrated model for BRCA (breast invasive carcinoma).

3.3.2 Simulation 2: Real Data-based Simulations

Data generation and choices of true effects Before the pan-cancer multi-platform proteogenomic analysis using data from the Cancer Genome Atlas, I perform a modified version of Simulation 1, based on breast invasive carcinoma (BRCA) patient data. I aggregate and annotate DNA methylation, copy number alteration, proteogenomic expression, and censored survival data for BRCA following Section 3.4. Since the rest of the simulation scheme is mostly similar to Simulation 1, I only summarize the changes briefly below.

I denote the upstream covariate data and the proteogenomic expression data for BRCA respectively by $\mathbf{U}_{n\times q}$ and $\mathbf{X}_{n\times p}$. Note that since a single proteogenomic biomarker may have multiple corresponding methylation sites and copy number changes, generally, $q \geq p$. In particular, for the BRCA data I use, $n = 790$ and $p = 365$. For each $\mathbf{X}_{\bullet j}$, I fit mechanistic model(s) on the corresponding $\mathbf{U}_{\bullet j}$s, and compute lBFs, as described in Section 3.2.2. I then group these lBFs as before (< 0.5 (no evidence), $0.5 - 1$ (substantial), $1 - 2$ (strong), > 2 (decisive)), and assign the group-specific means of the lBFs to the $\mathbf{X}_{\bullet j}$s belonging to the group as their level of prior evidence in the simulation. I then build a cBVS model using survival outcome data to estimate $\hat{\boldsymbol{\beta}}_{p\times 1}$. I cluster $\hat{\boldsymbol{\beta}}$ into four groups based on the three quartiles of their absolute values, and assign the group means to each group to arrive at the final $\boldsymbol{\beta}_{p\times 1}$. This leads us to a total of $4 \times 4 = 16$ combinations of prior evidence × outcome effect size. The outcome data generation procedure, benchmark methods, summary metrics, and number of simulations remain the same as before.

Results and discussion The resulting metrics are summarized in Figure 3.4B. Notably, cBVS performs the best among all the methods compared in terms of AUC, AUC20 and FPR, and is at the second place, only behind graper, for MCC. Again, two calibrated selection methods (cBVS, graper) outperform uncalibrated selection methods (EMVS, LASSO, IPF-

LASSO) in terms of all metrics but TPR. Thus, cBVS is well-equipped for the real data scenarios to be encountered in the pan-cancer setting.

3.3.3 Additional Simulation Studies

Comparing linear models and Gaussian processes in capturing non-linear evidence I perform a simple simulation study to compare the performances of Gaussian process models and linear models in capturing significance based on data generated from varying orders of non-linear associations. Hypothetically, with increasing proportion of non-linearity in the generating model, the Bayes factors from the Gaussian process models should be better able to quantify significance than those from the linear models. To perform this, our generating models are defined as the following.

- For a given sample size $n = 100$ and number of covariates $p = 5$, I generate $\mathbf{X}_{n \times p}$ where each $X_{ij} \overset{iid}{\sim} U(0, 1)$.
- The fully nonlinear model is defined as the following.

  $Y_i \sim N(10\cos(X_{i1}) - 15X_{i2}^2 + 10\exp(-X_{i3})X_{i2} - 8\sin(X_{i3})\cos(X_{i4}) + 20X_{i1}X_{i5}, 1)$, i.e., $Y_i \sim N(\sum_{j=1}^{p} \beta_j f_{ij}, 1)$. For $l \in \{0, 1, \dots, 5\}$, a $20l\%$ nonlinear model is then generated from the following distribution.

  $Y_i \sim N(\sum_{j=1}^{l} \beta_j f_{ij} + \sum_{j=l+1}^{p} \beta_j X_{ij}, 1)$.
- For each non-linearity scenario, 100 independent replicates are generated. For each replicate, the Bayes factors corresponding to the Gaussian process model (as described in Section 3.2.2) and the linear model are computed.

As seen in Figure 3.5, for 0% and 20% levels of non-linearity, the median lBF from the linear models is larger than that from the Gaussian processes, whereas this pattern is reversed for all the higher levels of non-linearity. Noticeably, the difference between the

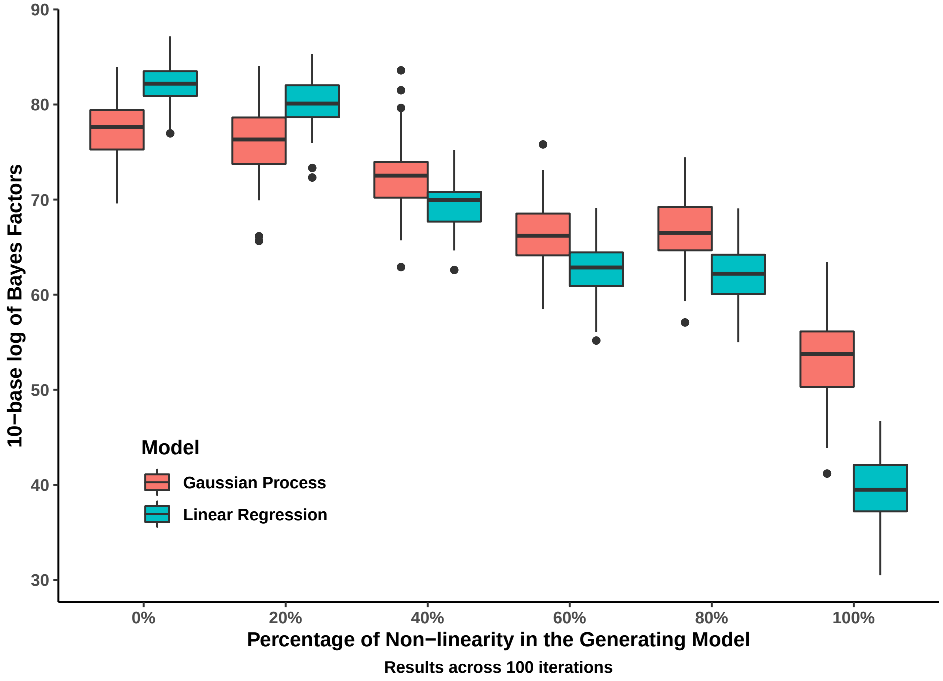


Figure 3.5: Results from Simulation comparing Gaussian processes and linear models. The x-axis is in increasing order of non-linearity (in terms of the mean function of the outcome based on the covariates) in the generating model. The y-axis presentes box plots across 100 replicates of the lBFs from Gaussian processes and linear models.

two sets of lBFs increase steadily with increasing non-linearity. This supports our claim that the *Gaussian process models are better equipped in identifying evidence* based on data originating from nonlinear generating models.

3.4 Integrative Pan-cancer Proteogenomic Analyses

3.4.1 Data Description, Cleaning, and Analysis

Pan-cancer multi-omics data I analyze pan-cancer multi-platform proteogenomic data from the Cancer Genome Atlas (TCGA). I include 14 cancers across four cancer groups, classified by commonalities in the sites/tissues of occurrence. For each cancer, DNA methylation, copy number alteration, and proteogenomic expression data are obtained. The list of cancers for each group are presented below. The sample size summaries across different platforms for each cancer are available in our shiny app hosted at `https://bayesrx.shinyapps.io/Functional_iBAG/`.

- **pan-gyne:** breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), uterine carcinosarcoma (UCS).
- **Pan-kidney:** kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP).
- **Pan-squamous:** esophageal carcinoma (ESCA, squamous), head and neck squamous cell carcinoma (HNSC), lung squamous cell carcinoma (LUSC).
- **Pan-gastrointestinal (pan-GI):** colon and rectum adenocarcinoma (CORE), esophageal carcinoma (ESCA, adeno), stomach adenocarcinoma (STAD).

Outcome data I investigate two different outcomes. I use the censored survival data available from TCGA and implement the cBVS model to identify cancer-specific proteogenomic biomarkers associated with survival. However, survival outcome alone does not provide biologically interpretable insights into the molecular progression of the disease. To this end, I use another outcome called mRNA-based stemness index (SI, in short) - a metric of cancer growth in terms of its cellular features. Briefly, SIs for TCGA cancers are computed using a one-class logistic regression model trained on pluripotent stem cell samples from the Progenitor Cell Biology Consortium dataset (Daily et al., 2017; Malta et al., 2018). The SIs quantify the stem-cell-like behavior of the tumor of interest. I build cBVS models selecting proteogenomic biomarkers associated with SIs.

Modeling For each cancer, the mechanistic model analyses and hypothesis tests for each gene and protein are performed using GPs as described in Section 3.2.2, and lBFs are computed. Two cBVS models - one using stemness and one using survival data - are built for each cancer, using the selection-only MCMC procedure as described in Section 3.2.4

to estimate $(\boldsymbol{\beta}^T, \boldsymbol{\omega}^T, \sigma)^T$ in each. For each gene and protein, I thus obtain five quantities of interest: lBF, and $(\hat{\beta}_j, \hat{\omega}_j)^T$ - one each from each outcome model. An overview of the analysis pipeline is presented in Figure 3.6.

Scientific Questions Our analyses are driven by three broad scientific aims. First, I intend to *identify cancer-specific and pan-cancer functional drivers and cascades* from the proteogenomic candidates using the functional evidence learned via the mechanistic models. Second, using the cBVS models, I *select cancer-specific biomarkers associated with changes in the survival and stemness outcomes*. Finally, I assess the mechanistic and outcome model results in conjunction to *evaluate the utility of calibrating outcome models using mechanistic evidence*. I present the numerical results in the following two subsections, followed by the biological interpretations and discussions of the results in Section 3.4.5.

3.4.2 Mechanistic Model Results for TCGA Cancer Groups

I summarize the mechanistic model outputs in a pan-cancer fashion for the pan-gyne and pan-GI cancer groups in Figure 3.7; the rest of the groups are presented in Figure S3.15-S3.20. Figure 3.7A, C exhibit heatmaps summarizing lBF classes for the gene-protein pairs which have some evidence across at least three-fourth of the cancers in at least two out of the three mechanistic model types, along with corresponding upset plots in Figure 3.7B, D describing the number of genes/proteins that are at the decisive level of significance for the three mechanistic models across the intersections of the different cancers.

For the pan-gyne cancers, 26 gene-protein pairs are at strong/decisive level of evidence across at least four cancers in at least two out of the three mechanistic model types, including genes such as RPS6KA1 (protein p90RSK), YAP1 (proteins YAP and YAPPS127), and DIABLO (protein SMAC) (Figure 3.7A). The largest sharing of decisive driver gene signatures is observed across BRCA, CESC, OV, and UCEC, and that for cascading proteins

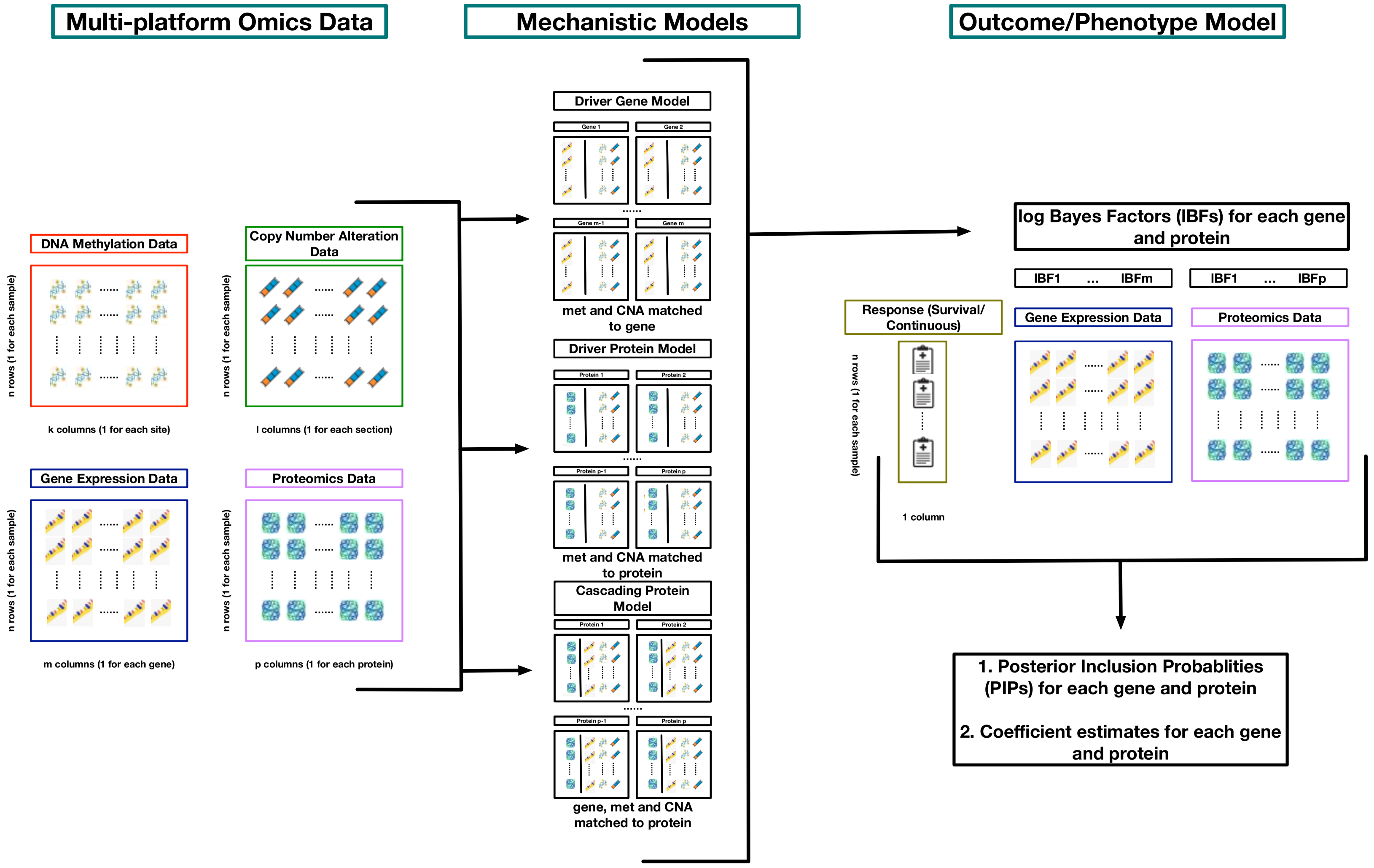

Multi-platform Omics Data
Mechanistic Models
Outcome/Phenotype Model
DNA Methylation Data
n rows (1 for each sample)
k columns (1 for each site)
Copy Number Alteration Data
n rows (1 for each sample)
l columns (1 for each section)
Gene Expression Data
n rows (1 for each sample)
m columns (1 for each gene)
Proteomics Data
n rows (1 for each sample)
p columns (1 for each protein)
Driver Gene Model
Gene 1
Gene 2
Gene m-1
Gene m
met and CNA matched to gene
Driver Protein Model
Protein 1
Protein 2
Protein p-1
Protein p
met and CNA matched to protein
Cascading Protein Model
Protein 1
Protein 2
Protein p-1
Protein p
gene, met and CNA matched to protein
log Bayes Factors (lBFs) for each gene and protein
lBF1 ... lBFm
lBF1 ... lBFp
Response (Survival/ Continuous)
n rows (1 for each sample)
1 column
Gene Expression Data
Proteomics Data
1. Posterior Inclusion Probablities (PIPs) for each gene and protein
2. Coefficient estimates for each gene and protein

Figure 3.6: Integrative analysis using the fiBAG pipeline. In each mechanistic model presented in the middle panel, upstream covariates are matched to each proteogenomic biomarker. The lBFs computed from each model are used in the cBVS outcome models to perform calibrated variable selection, as outlined in the right panel. The outcome models provide posterior inclusion probabilities for each biomarker along with estimated coefficient for association with the outcome of choice. The entire pipeline is independently applied on each cancer. I build one survival and one stemness outcome model for each cancer.

is observed across BRCA, OV, and UCEC (Figure 3.7B).

For the pan-GI cancers, eight gene-protein pairs are at strong/decisive level of evidence across all three cancers in at least two out of the three mechanistic model types, including genes such as ERBB2 (proteins HER2 and HER2PY1248), CCNE1 (protein CYCLINE1), and MAPK9 (protein JNK2) (Figure 3.7C). The largest number of decisive driver gene, driver protein, and cascading protein signatures is observed from CORE, and the largest numbers of pan-cancer signatures for the driver gene and the cascading protein models are observed between CORE and ESCA (Figure 3.7D).

3.4.3 Stemness cBVS Results for TCGA Cancers

I summarize the stemness outcome model outputs for each TCGA cancer using plots showing $-log_{10}(1 - \hat{\omega})$ on the y-axis and $\hat{\beta}$ on the x-axis for the nine cancers with the largest sample sizes in Figure 3.8; the rest are available in our shiny app at `https://bayesrx.shinyapps.io/Functional_iBAG/`. The shiny app also contains bar diagrams of the lBFs for the selected genes/proteins and histograms with density plots of lBFs for the predictors not selected, for each cancer. Here, I discuss the results from the BRCA and CORE analyses since they are the cancers with the largest sample sizes in the pan-gyne and pan-GI groups, respectively.

For BRCA, no protein is selected using the 10% FDR cut-off on $\hat{\boldsymbol{\omega}}$ from the stemness cBVS model. Several genes are selected, such as YAP1, DIABLO, and RAPTOR (Figure 3.8A). The genes selected cover a large span of evidence range from the mechanistic models, with genes like JUN and COL6A1 having no functional evidence and DVL3, MYH11, EIF4G1 at the decisive level of evidence, indicating that cBVS is capable of identifying associations even in the absence of prior evidence, as has been noted in the simulation studies before too (Section 3.3). The histogram indicates that a vast majority of the non-selected genes and proteins have little to no evidence from the mechanistic models,

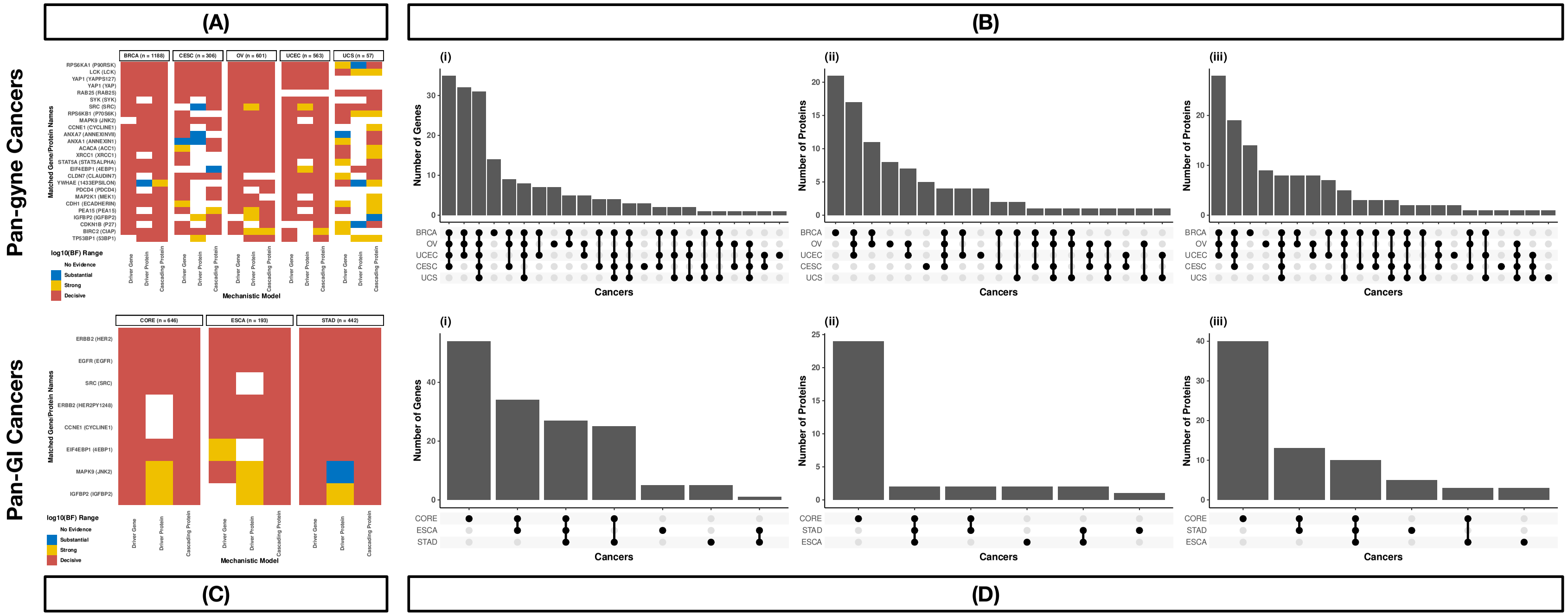


Figure 3.7: Summary of mechanistic model results for the pan-gyne group cancers: BRCA (breast invasive carcinoma), CESC (cervical squamous cell carcinoma and endocervical adenocarcinoma), OV (ovarian serous cystadenocarcinoma), UCEC (uterine corpus endometrial carcinoma), UCS (uterine carcinosarcoma) (panels A and B), and the pan-GI group cancers: CORE (colon and rectum adenocarcinoma), ESCA (esophageal carcinoma), STAD (stomach adenocarcinoma) (panels C and D). (A/C) The lBF ranges are defined as: < 0.5 (no evidence), $0.5 - 1$ (substantial), $1 - 2$ (strong), > 2 (decisive). Only the gene-protein pairs with some evidence across at least 75% of the cancers in at least two out of the three mechanistic model types are shown here. (B/D) Upset plots exhibiting the number of genes/proteins at the decisive level of significance for the (i) driver gene, (ii) driver protein, and (iii) cascading protein mechanistic models.

but 194 of them are at the strong and decisive levels of evidence and yet not selected by cBVS due to the absence of sufficient association with the outcome data.

For CORE, two genes - PEA15 and KDR are selected using the 10% FDR cut-off on $\hat{\boldsymbol{\omega}}$ from the stemness cBVS model, both negatively associated with the stemness index as can be seen from the sign of the estimated regression coefficients (Figure 3.8I). Both the genes selected are at the decisive level of functional evidence. The histogram again indicates that a vast majority of the non-selected genes and proteins have little to no evidence from the mechanistic models - however, 179 many are at the strong/decisive levels of evidence and are still not selected by cBVS.

3.4.4 Survival cBVS Results for TCGA Cancers

I summarize the survival cBVS outputs for each TCGA cancer using plots showing $-log_{10}(1 - \hat{\omega})$ on the y-axis and $\hat{\beta}$ on the x-axis for each cancer. Further, I present bar diagrams exhibiting the lBFs for the selected genes and proteins and histograms with density plots of lBFs for the predictors not selected. To ensure a cleaner presentation, I include these results only in the shiny app hosted at `https://bayesrx.shinyapps.io/Functional_iBAG/`. Here, similar to the previous subsection, I discuss the results from the BRCA and CORE analyses.

For BRCA, no gene is selected using the 10% FDR cut-off. The only protein selected is Collagen VI, which has decisive functional evidence. Again, a vast majority of the non-selected predictors had little to no functional evidence. For CORE, no gene/protein is selected at the 10% FDR threshold.

3.4.5 Biological Findings and Implications

Majority of the proteogenomic biomarkers identified in our analyses have supporting evidence from past literature in terms of their roles in cancer mechanism. In this subsection, I discuss some of these results in the light of past evidence regarding their molecular

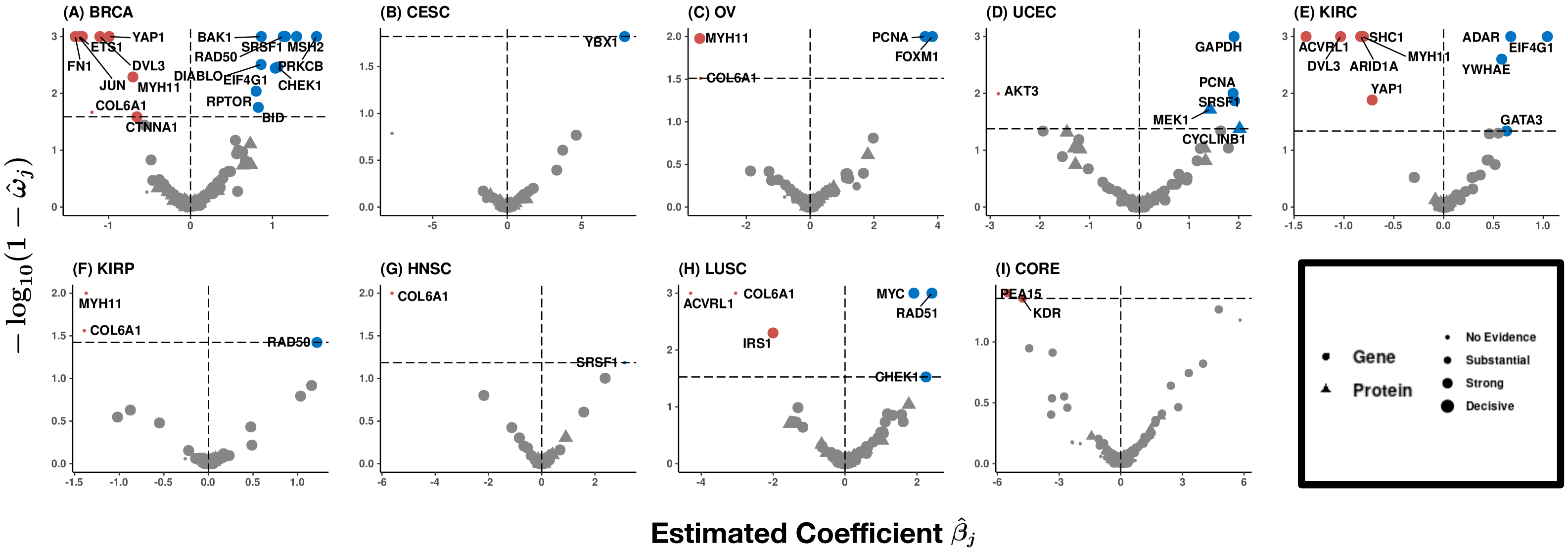


Figure 3.8: Plots summarizing outcome model results based on stemness indices (SI) for TCGA cancers. Proteins are represented by triangles and genes by circles. The shapes are colored red if the estimated $\hat{\beta}_j$ from fiBAG is negative, and green if positive. The x-axis shows the $\hat{\beta}_j$s, and the y-axis shows the $-\log_{10}(1-\hat{\omega}_j)$s. An FDR check to adjust for multiple comparisons is performed treating $1-\hat{\omega}_j$ as a p-value type quantity at the 10% FDR level. Only the selected biomarkers are marked in non-gray colors and labeled. The sizes of the points are in the increasing order of evidence from the mechanistic models: lBF ranges are defined as: < 0.5 (no evidence), $0.5-1$ (substantial), $1-2$ (strong), > 2 (decisive).

mechanisms in cancer progression and patient survival. All associations discussed below are significant at a 10% level of FDR control.

RPS6KA1 gene and p90RSK kinases in gynecological cancer progression The gene RPS6KA1 (protein p90RSK) has decisive evidence for all but two mechanistic models in the pan-gyne cancers (Figure 3.7A). The gene has been known to be differentially expressed in endometrial cancer tissue as opposed to benign endometrial tissue (Mamoor, 2021). Specifically, it is known to have a favourable prognostic effect on clinical outcomes in endometrial cancers (Bradfield et al., 2020). The p90RSK protein is selected in the survival cBVS model for OV. p90RSK has been known to impact metastatic seeding of ovarian cancer cells, effecting the invasiveness of the cancer via activating a self-reinforcing cell autonomous circuit (Torchiaro et al., 2016). The RPS6KA1 gene is also significantly associated with the increased risk of breast cancer (Shareefi et al., 2020).

YAP1 gene and YAP proteins in gynecological malignancies The gene YAP1 (proteins YAP, YAPPS127) has decisive evidence in all mechanistic models for all gynecological cancers except UCS (smallest sample size in the group, $n = 57$) (Figure 3.7A). YAP is a crucial agent impacting gynecological cancers. As a transcriptional co-activator within the Hippo pathway, over-activation of YAP leads to uncontrolled cell growth and malignant transformation in gynecological malignancies, including cervical, ovarian, and endometrial cancers (Wang et al., 2020). Further, YAP expression is associated with a poor prognosis for gynecological cancers - activation of YAP induces cancer cell proliferation and migration (breast: Guo et al. (2019), cervical: He et al. (2015), endometrial: Tsujiura et al. (2014)). This aligns with our identification of YAP1 as negatively associated with stemness for BRCA (Figure 3.8A) and with survival for OV and CESC.

DIABLO gene as a marker of gynecological tumors Another interesting candidate emerging from the mechanistic models is the gene DIABLO (protein Smac), which has been proposed as a biomarker for gynecological tumors, so far with little knowledge about its cellular mechanism. A recent study shows some evidence in favor of a positive association of Smac/DIABLO expression levels with estrogen receptor positivity in breast cancer (Espinosa et al., 2021). Our cBVS analyses identify higher DIABLO expression to be associated with higher tumor stemness for two gynecological cancers (BRCA and UCS, Figure 3.8A), which is in line with prior knowledge (Arellano-Llamas et al., 2006; Arbiser, 2018). On the other hand, DIABLO is associated with higher survival for OV, supported by earlier evidence - higher expression of DIABLO is a good prognostic sign for ovarian and endometrial cancers (Dobrzycka et al., 2010, 2015). However, neither DIABLO nor its corresponding protein Smac as top candidates in the mechanistic analyses.

HER2 as a therapeutic target in gastrointestinal cancers The gene ERBB2 (protein HER2) has decisive evidence for all mechanistic models in the pan-GI group, emerging as the top candidate (Figure 3.7C). HER2 overexpression has been known to be a frequent molecular abnormality in gastric cancers via gene amplification (Gravalos and Jimeno, 2008). HER2 has also been considered as a molecular therapeutic target for patients with advanced gastric cancer (Abrahao-Machado and Scapulatempo-Neto, 2016). In our cBVS analyses, both ERBB2/HER2 are positively associated with stemness for ESCA and STAD, which aligns with past knowledge - increased expression of HER2 leads to quicker growth and poorer progression of gastric and esophageal cancers (Malaguti et al., 2015; Lee et al., 2019).

EGFR exhibiting contradictory associations with gastrointestinal cancers The EGFR gene/protein emerge jointly on top in the pan-GI mechanistic models, with consistent decisive evidence (Figure 3.7C). Among gastric cancer patients, 2–35% are reported to have EGFR protein overexpression and/or gene amplification (Adashek et al., 2020). However, the utility of EGFR as a therapeutic target has been questionable at best so far, with no general consensus on its prognostic value in gastric cancer (Arienti et al., 2019). While some studies have indicated that high EGFR gene amplification is associated with poor outcome (Kandel et al., 2014), others have suggested the opposite (Aydin et al., 2014). In line with this, cBVS found contradictory associations - EGFR gene expression is negatively associated with both stemness and survival for ESCA but positively associated with both for STAD.

3.5 Discussion and Future Work

Overview I propose fiBAG, a hierarchical Bayesian framework to perform integration of multi-omics data and outcomes. Using GPs, I quantify the functional roles of proteogenomic biomarkers along three axes of interest (driver gene, driver protein, cascading protein). Then, using a Bayesian variable selection procedure with a calibrated spike-and-slab prior on the regression coefficients, I incorporate this functional evidence in the outcome model to improve covariate selection and effect size estimation. The framework offers novelty and utility in multiple directions. First, it offers the user liberty in terms of multi-platform integration, in the sense that depending on the data available, the mechanistic and outcome layers can be appended or modified along with additions or alterations of the complete set of parameters of interest. For example, I use DNA methylation and copy number alterations as upstream covariates in our mechanistic models, but other information such as miRNA expression could easily be incorporated. Second, the calibrated spike-and-slab prior addresses the more general statistical question of incorporating external information in a variable selection setting - by updating the mapping function used to calibrate evidences to a prior probability scale, the procedure can be adapted to other settings where the numerical evidences are in a different scale and/or sourced from different models. Our calibrated regression framework, i.e., cBVS is compared with multiple benchmarks via simulation studies in synthetic and real data-based settings. The benchmarks include uncalibrated Bayesian variable selection (EMVS), standard and grouped penalized regression (LASSO and IPF-LASSO), and Bayesian variable selection incorporating external information (graper and xtune). I use selection metrics like AUC and MCC and estimation metrics such as MAD and MSE to evaluate performances across a broad spectrum of n/p ratios. cBVS outperforms all the uncalibrated methods across all n/p settings, and performs comparably with the calibrated methods in the high n/p settings. For low n/p, cBVS offers improvement in selection compared to other calibrated methods (Figure 3.4).

Pan-cancer applications Our framework is cancer-specific (each mechanistic and outcome model is built for each cancer separately and then assessed in a pan-cancer fashion) for several reasons. First of all, non-omics covariates (**B** - e.g., demographic information such as gender or age, and clinical information such as cancer stage) would potentially be entirely different across cancers, and may have widely different scales of measure. Further, even the outcomes may not always be normalized across cancers - while a scaled outcome like the mRNAsi does not pose this problem, survival, for example, may differ considerably across different cancers. Finally, the mechanistic models need to be run separately for each gene and protein, and depending on the cancer of interest, the functional roles of these genes and proteins are potentially different, effectively rendering the measures of evidence to be cancer-specific as well. Having established the improved utility of our method in an evidence-based setting, I analyze pan-cancer multiomics data from TCGA, across a total of four cancer groups (pan-gyne, pan-kidney, pan-squamous and pan-GI) and 14 cancers. Our real data analyses identify both known and novel associations at cancer-specific and pan-cancer levels, such as the identification of the roles of RPS6KA1 gene and p90RSK kinases in progression of gynecological cancers, and the potential utility of the EGFR gene and protein as therapeutic targets for ESCA where their expressions are negatively associated with survival outcomes.

Sensitivity analysis in real data applications A key point of interest in validating such detections is to ensure that the implemented procedure is not overly sensitive to the hyperparameter specifications and that the results are not driven majorly by these choices. To this end, I performed the stemness outcome model analysis for the five pan-gynecological cancers using not only the specific set of tuning parameters for the calibration function in the calibrated spike-and-slab prior as described in Section 3.2.3, but also additional parameter values. To recall, the core part of the four-parameter logistic function contains the transformation $(s_j^*/3)^{-2.75}$. For the tuning parameter in the exponent, a grid between $[2.5, 3]$ with a spacing of 0.125 (five values) is used, and for the tuning parameters in the denominator, a grid between $[2.5, 3.5]$ with a spacing of 0.25 (five values) is used. I tabulate the proteogenomic biomarkers selected across each of the 25 combinations of the tuning parameters for each of the cancers. The 18 proteogenomic biomarkers selected for BRCA stemness outcome models as exhibited in Figure 3.8 are consistently selected in 21 out of the 25 tuning parameter combinations, with the exceptions all being cases where the denominator is 3.5. As can be noted from the mathematical form of the calibration function, increase in this parameter implies a decrease in the prior inclusion probability for the same value of mechanistic evidence, which leads to non-selection of some signals with strong or decisive evidence. As can be observed in Figure S3.10-S3.14, an even higher value of four for the denominator parameter leads to further lower prior inclusion probabilities for even lBFs higher than two (i.e. decisive). Overall, the results indicate that as long as the calibration is performed based on parameters that allow the prior mechanism to incorporate the evidence up to a reasonable strength of belief, the exact selection results would not vary substantially.

Flexibility and scientific novelty of the analysis scheme The flexible covariate-specific mechanistic modelling approach offers the feasibility of utilizing platforms with larger expression pools such as the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). Further, more cancers or cancer groups can be included in the analysis pipeline as well, since the cBVS model fitting is cancer-specific and I offer three options in decreasing order of computational complexity (full MCMC, selection-only MCMC, and EMVS). All these features make the whole procedure highly parallelizable (covariate level for mechanistic models and cancer level for outcome models) - the computation times for the simulation and real data settings described in this chapter reassure the utility of our pipeline in this direction. An important novelty of our integrative analyses is the use of mRNA-based cancer stemness indices as outcomes. Indeed, tumor stemness is a pertinent determinant of cancer growth and prognostic outcomes, and only recently there have been efforts to quantify stemness based on cellular signatures (Malta et al., 2018). To the best of our knowledge, our study is the first to look at potential associations of stemness with a pool of proteogenomic biomarkers in an oncological context. A relevant question, however, would be whether the associations identified are meaningful or artifacts of the construction of the index. To answer this, I looked at the machine learning procedure employed by Malta et al. (2018), specifically at the ranking of the biomarkers according to decreasing absolute value of weight in the final trained model. Reassuringly, none of our identified proteogenomic biomarkers across all the cancers belonged to the top 50 of the list, and very few belonged to top 100, 500, and 1000, with the highest magnitude of the weight among these selected covariates being at the order of 10^{-3}.

Future directions and reproducibility This work indicates a number of potential avenues for future statistical research. First, the flexibility in choosing the calibration function suggests that the function can even potentially be data-driven (such as choosing the tuning parameters via some pre-hoc analysis of the data). The cBVS framework can also be adapted to settings slightly different from ours - such as the link function being non-linear, or using bivariate outcomes and so on. To offer seamless interactive visualization of our integrative analysis results, I have built an R Shiny app, hosted at `https://bayesrx.shinyapps.io/Functional_iBAG/`. All our model and figure codes have been made available publicly via the app. I strongly believe that fiBAG will offer significant prognostic/therapeutic utility in oncological treatment and research.

3.6 Supplementary Figures

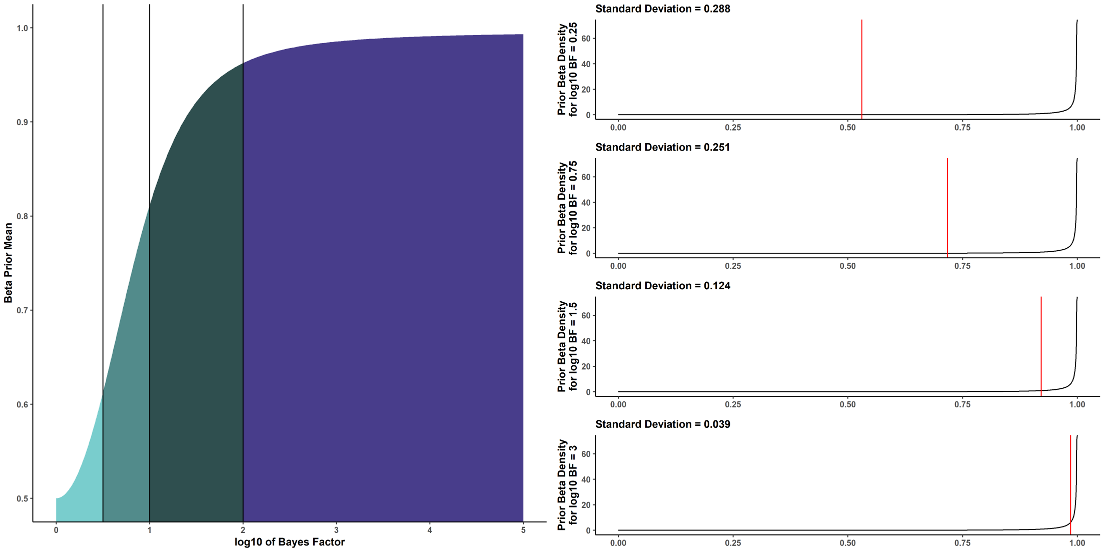


Figure S3.1: Evidence calibration function for tuning parameters (2, 2). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

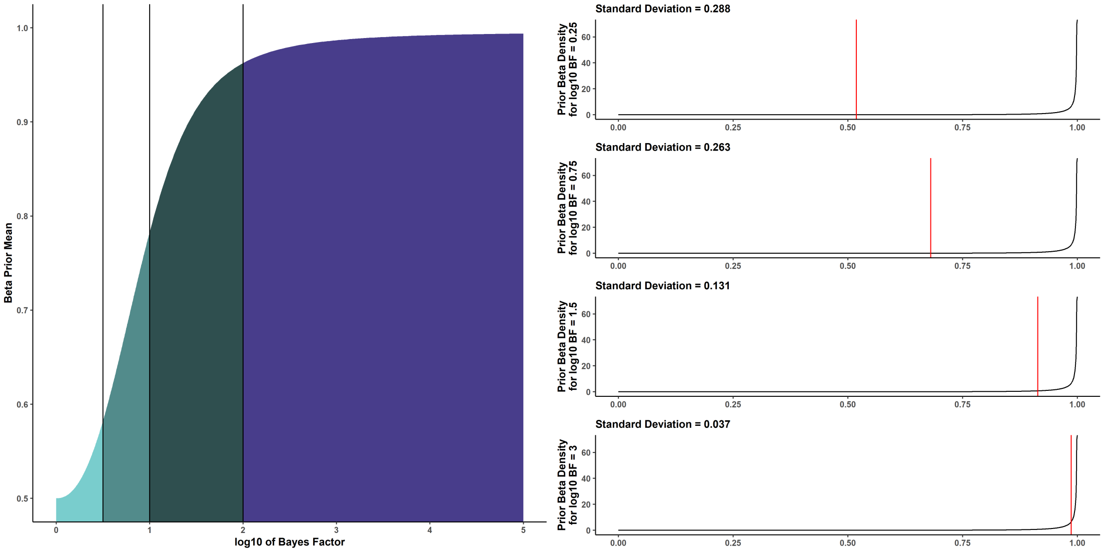


Figure S3.2: Evidence calibration function for tuning parameters (2, 2.25). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

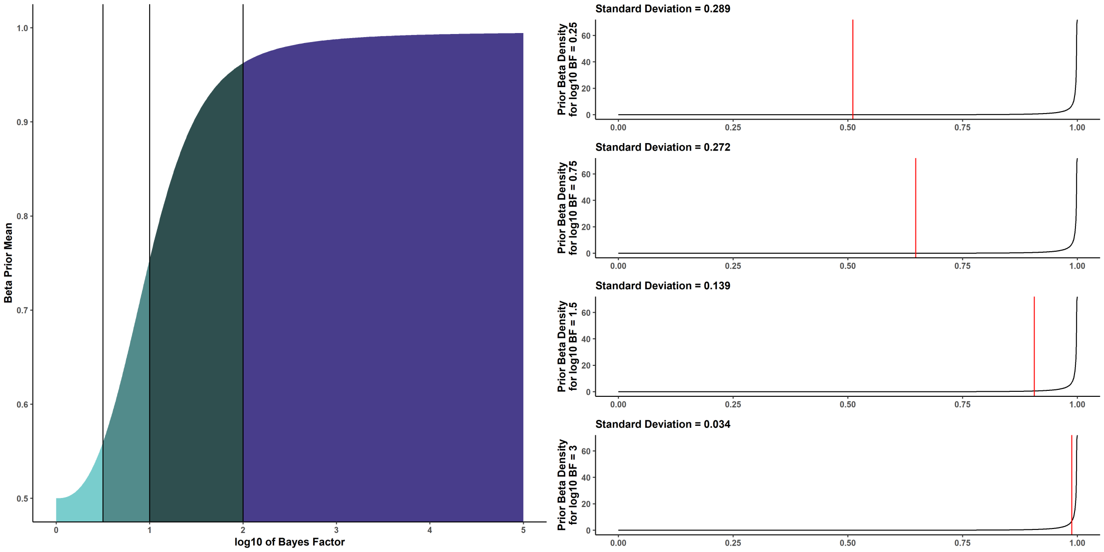


Figure S3.3: Evidence calibration function for tuning parameters (2, 2.5). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

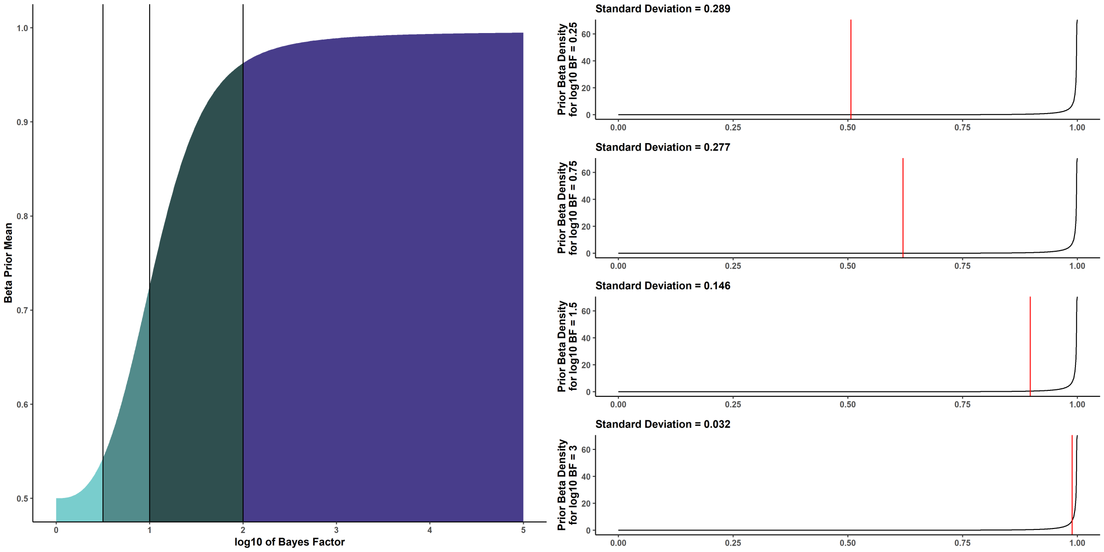


Figure S3.4: Evidence calibration function for tuning parameters (2, 2.75). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

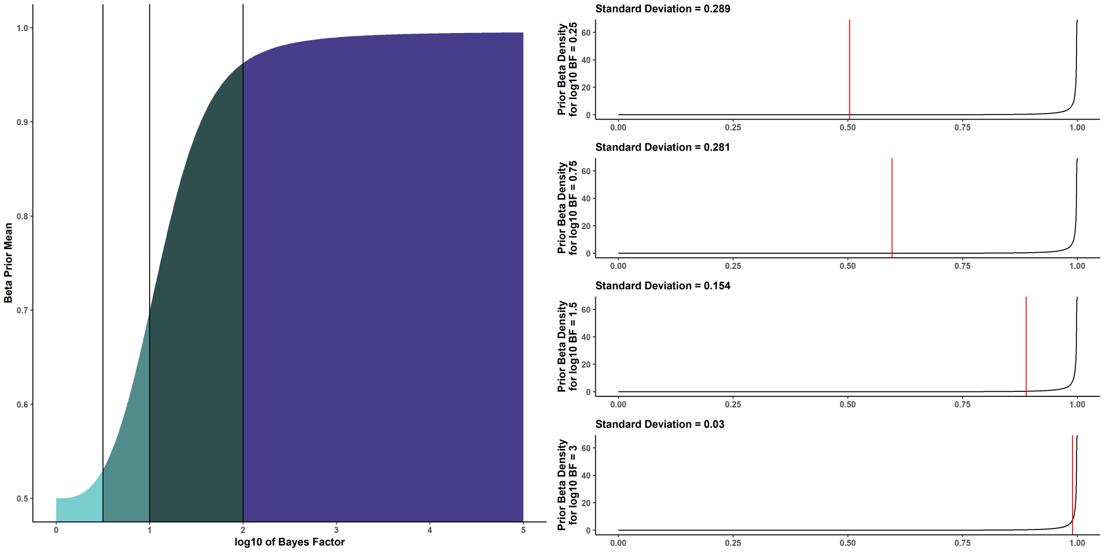


Figure S3.5: Evidence calibration function for tuning parameters (2, 3). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

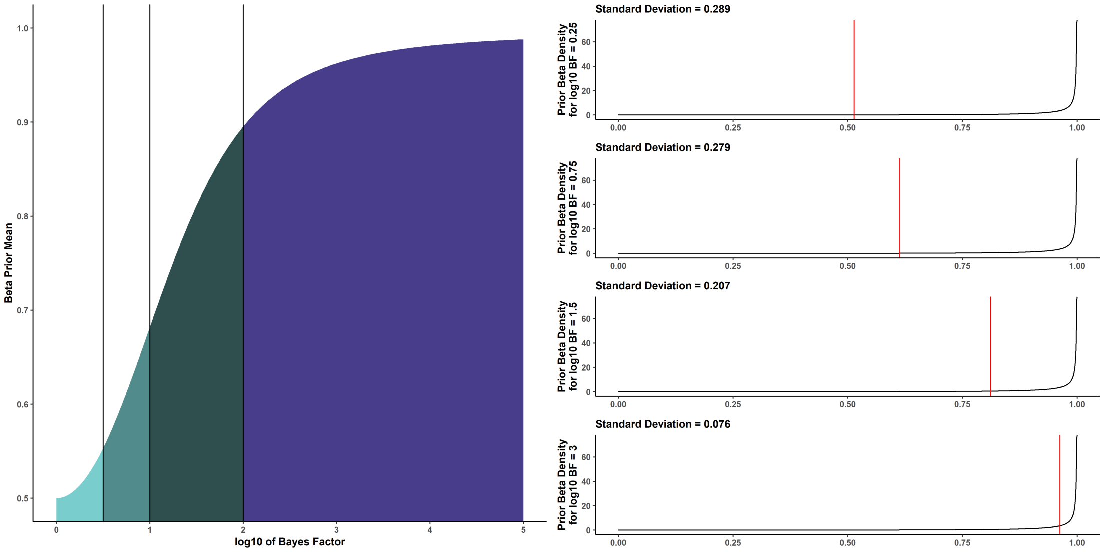


Figure S3.6: Evidence calibration function for tuning parameters (3, 2). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

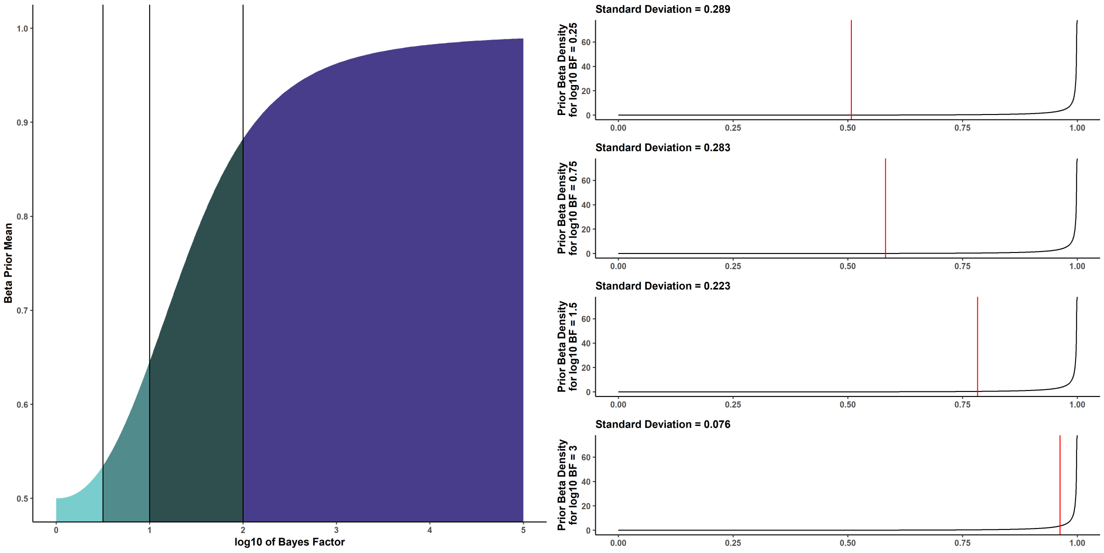


Figure S3.7: Evidence calibration function for tuning parameters (3, 2.25). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

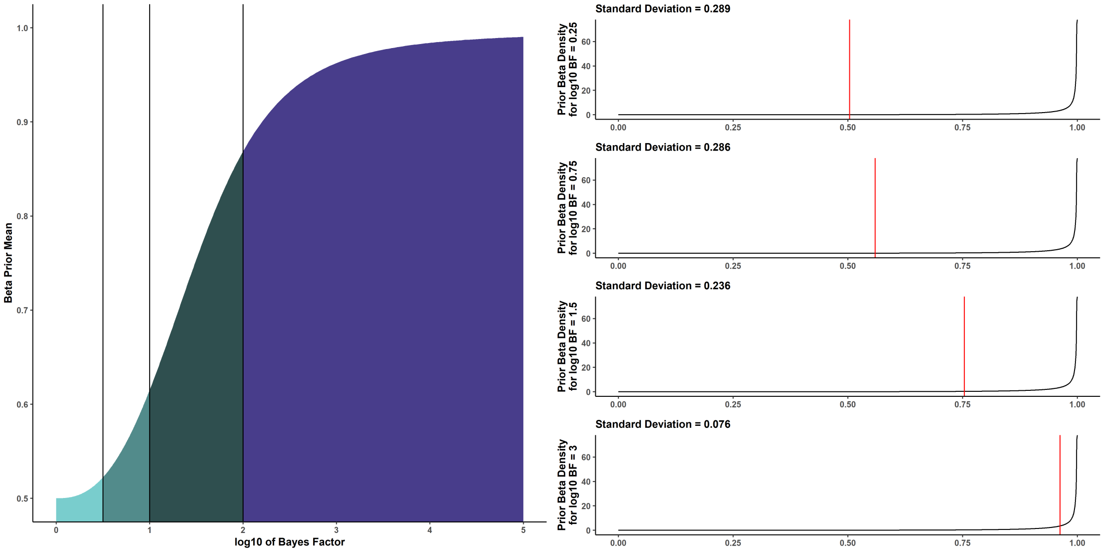


Figure S3.8: Evidence calibration function for tuning parameters (3, 2.5). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

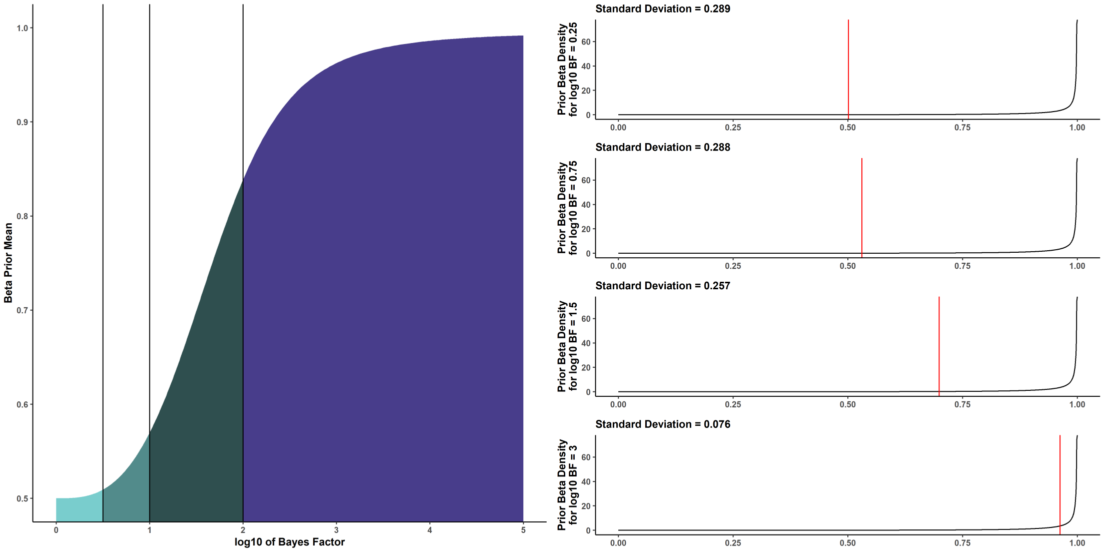


Figure S3.9: Evidence calibration function for tuning parameters (3, 3). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

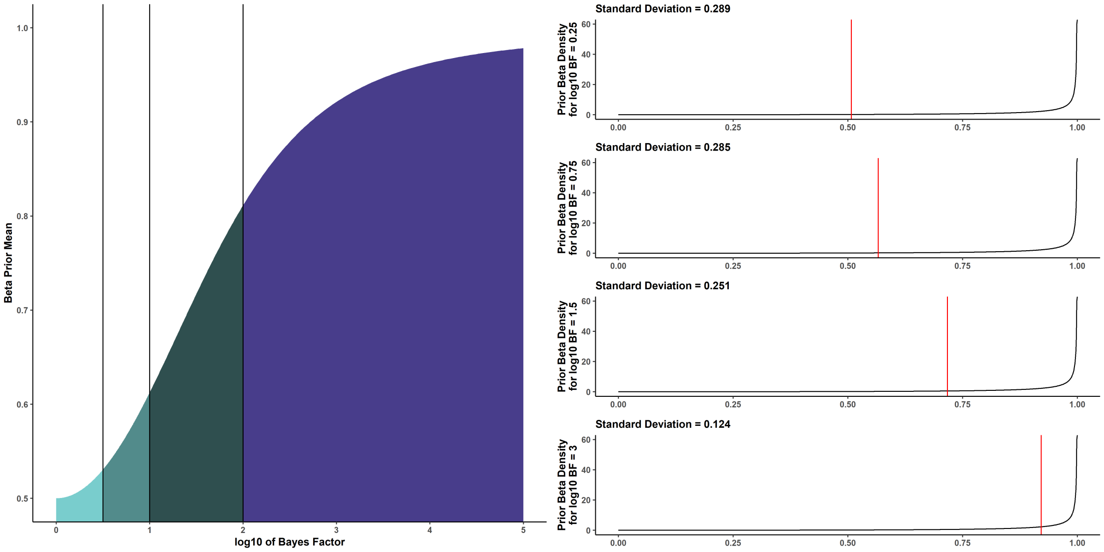


Figure S3.10: Evidence calibration function for tuning parameters (4, 2). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

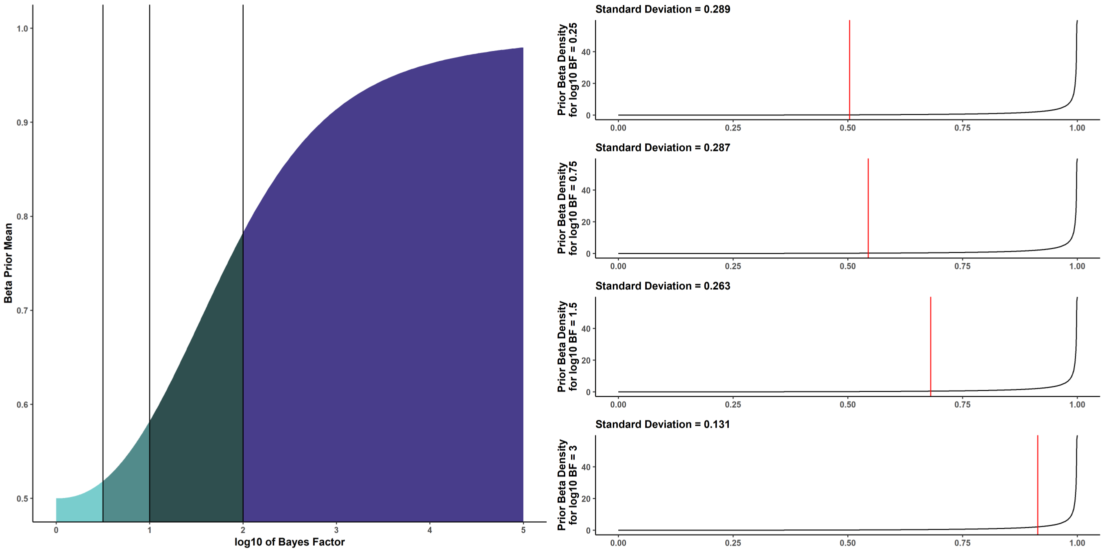


Figure S3.11: Evidence calibration function for tuning parameters (4, 2.25). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

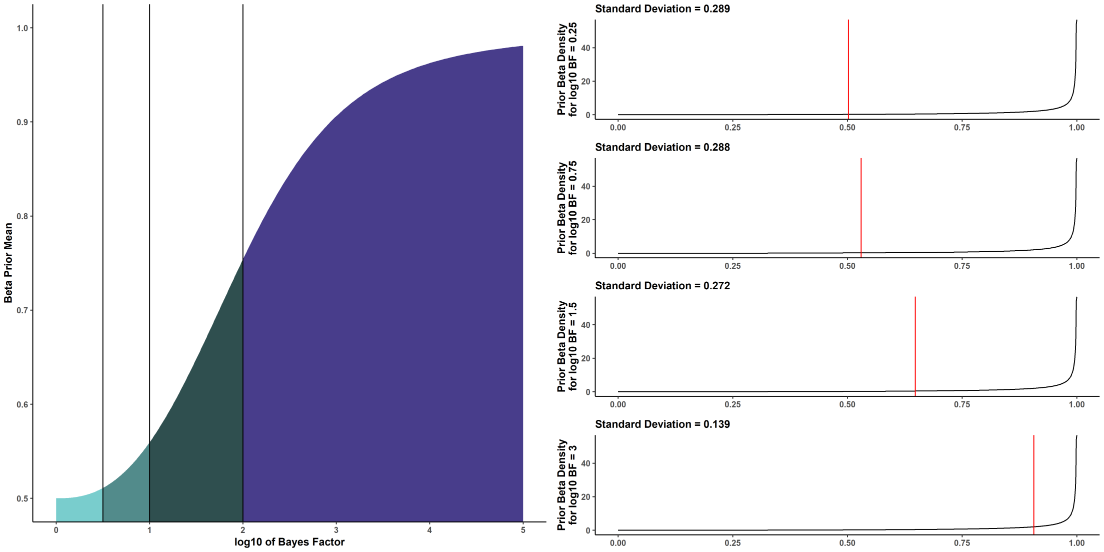


Figure S3.12: Evidence calibration function for tuning parameters (4, 2.5). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

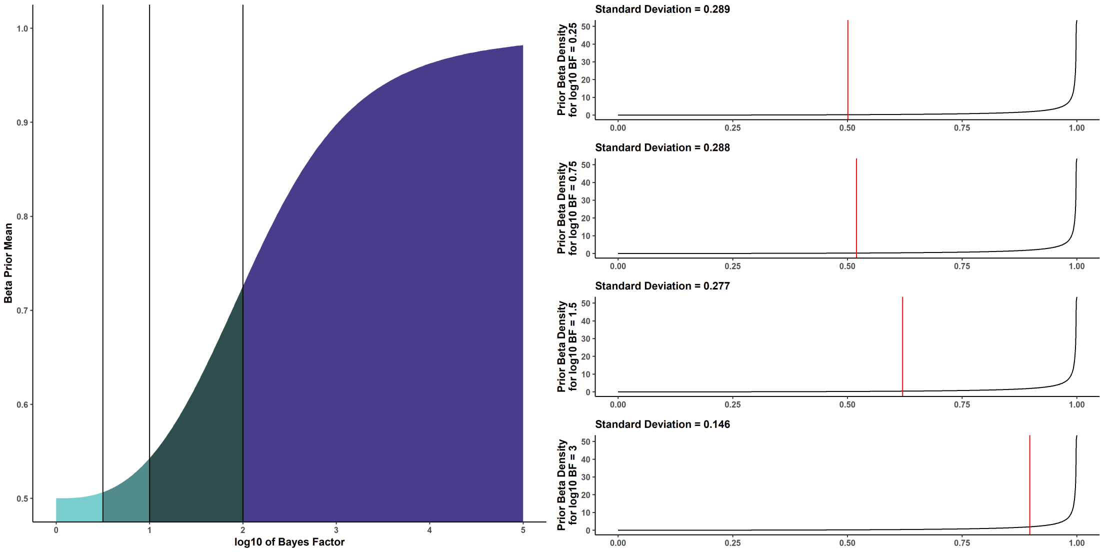


Figure S3.13: Evidence calibration function for tuning parameters (4, 2.75). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

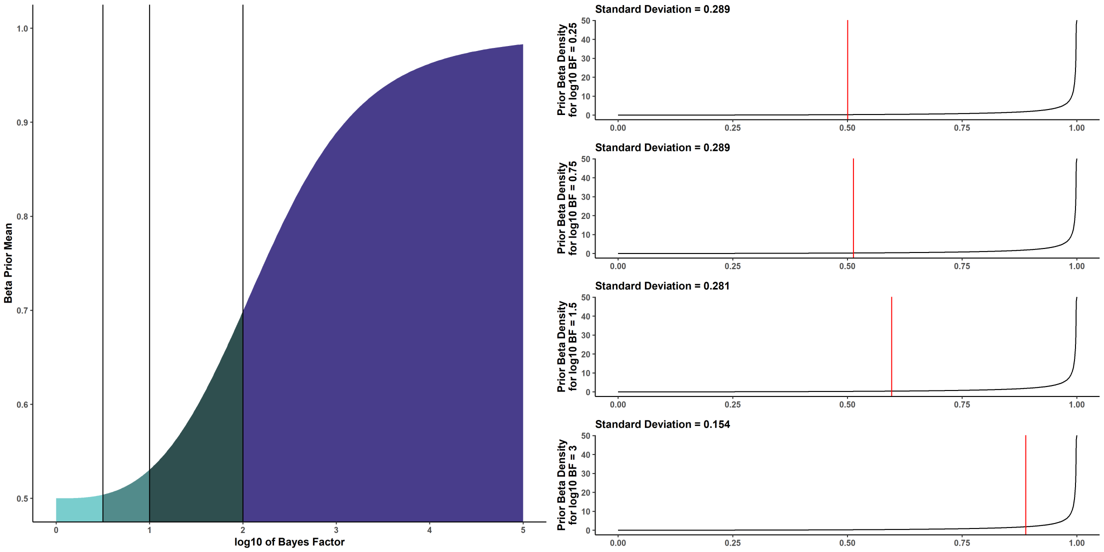


Figure S3.14: Evidence calibration function for tuning parameters (4, 3). The left panel summarizes the relationship between the beta prior mean and the lBF from the mechanistic model. The right panel presents the corresponding beta densities at lBF = 0.25, 0.75, 1.5, and 3.

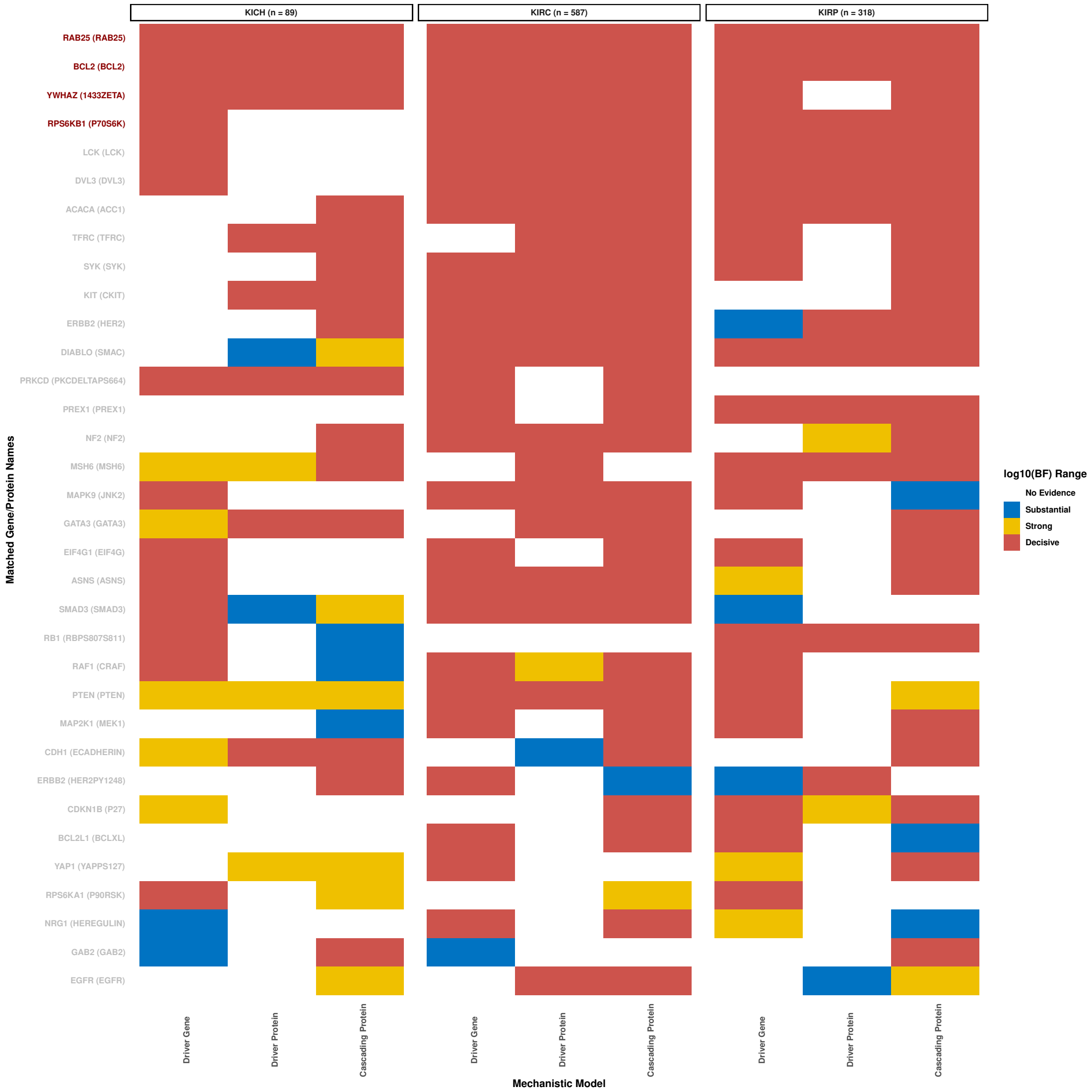


Figure S3.15: Mechanistic model heatmaps for the pan-kidney cancers. Each cancer column consists of three sub-columns, one each for the three mechanistic models (driver gene, driver protein and cascading protein). The lBF ranges are defined as: < 0.5 (no evidence), $0.5 - 1$ (substantial), $1 - 2$ (strong), > 2 (decisive). Only the gene-protein pairs which are at the decisive level of significance across all four cancers in at least two out of the three mechanistic model types are shown here.

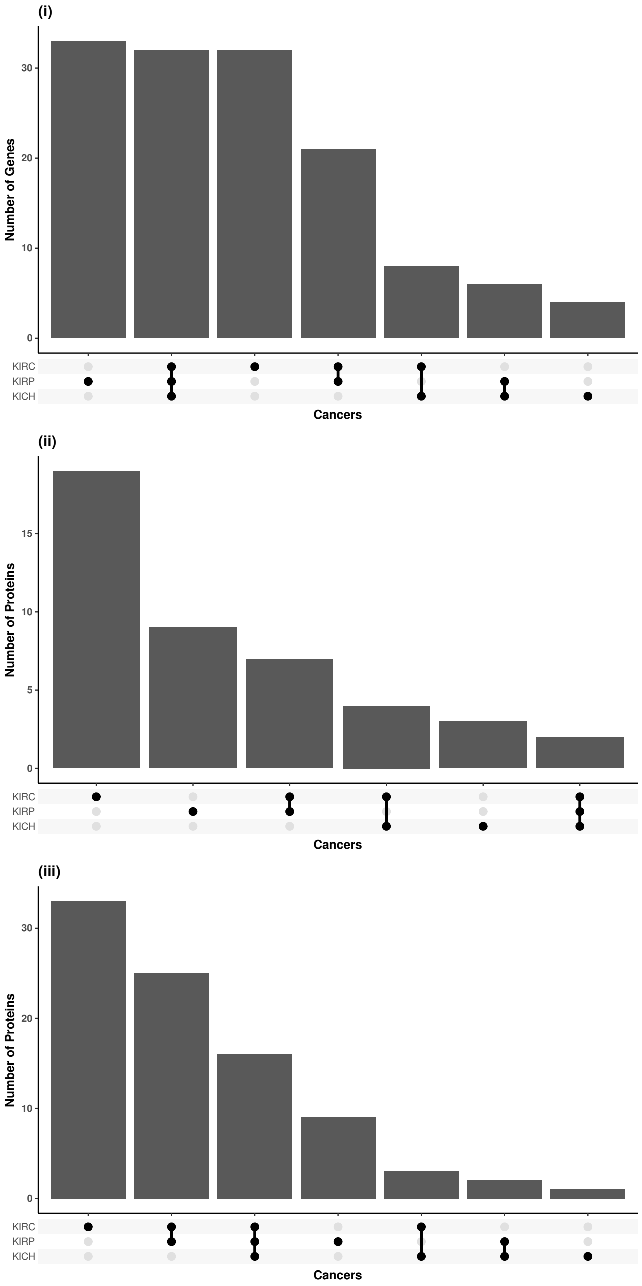


Figure S3.16: Mechanistic model upset plots for the pan-kidney cancers. Upset plots exhibit the number of genes (panel A) or proteins (panels B-C) that are at the decisive level of significance (lBF $>$ 2) for the (A) driver gene, (B) driver protein, and (C) cascading protein mechanistic models respectively, stratified by intersections across cancers.

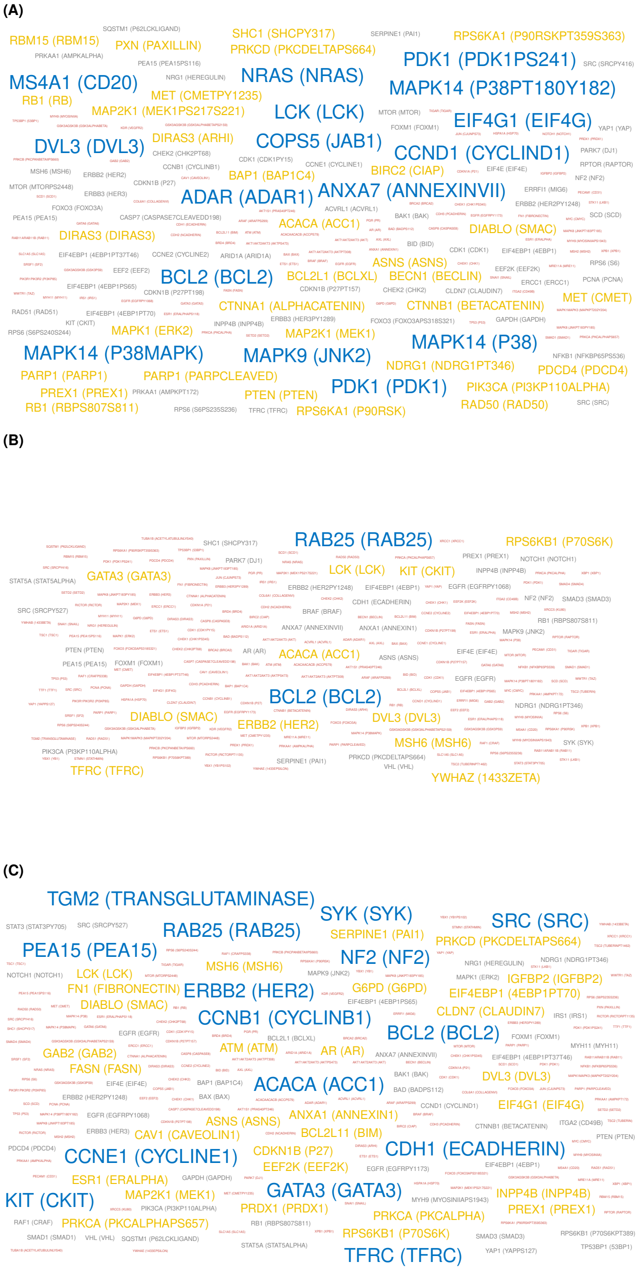


Figure S3.17: Mechanistic model word clouds for the pan-kidney cancers. Word clouds summarize pan-cancer mechanistic model results for genes (panel A) or proteins (panels B-C) for the (A) driver gene, (B) driver protein, and (C) cascading protein mechanistic models. The size of the gene/protein names are proportional to (no. of cancers where the gene/protein is at the decisive level of significance)^3. Here, decisive is defined as lBF > 2.

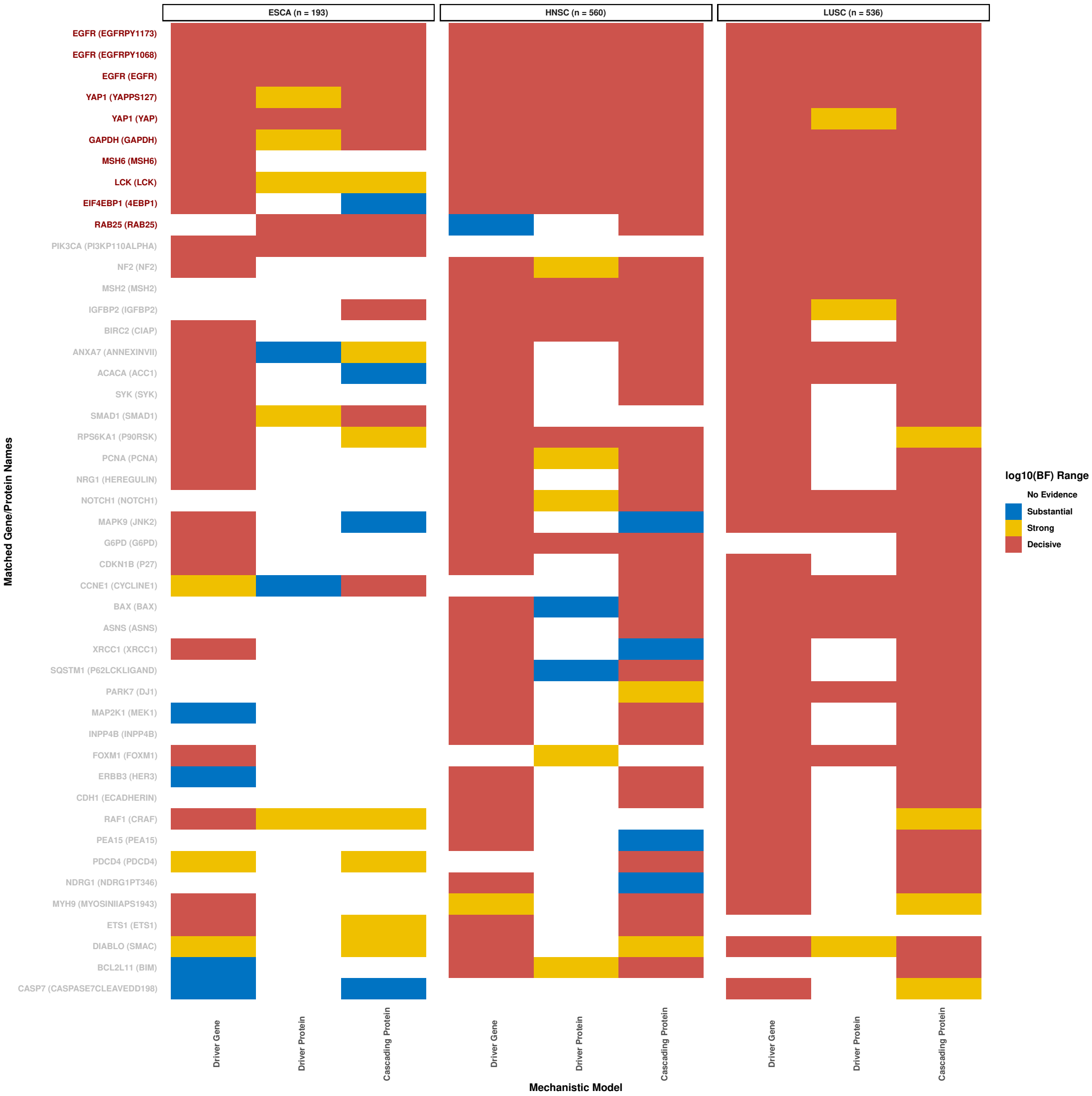


Figure S3.18: Mechanistic model heatmaps for the pan-squamous cancers. Each cancer column consists of three sub-columns, one each for the three mechanistic models (driver gene, driver protein and cascading protein). The lBF ranges are defined as: < 0.5 (no evidence), $0.5 - 1$ (substantial), $1 - 2$ (strong), > 2 (decisive). Only the gene-protein pairs which are at the decisive level of significance across all four cancers in at least two out of the three mechanistic model types are shown here.

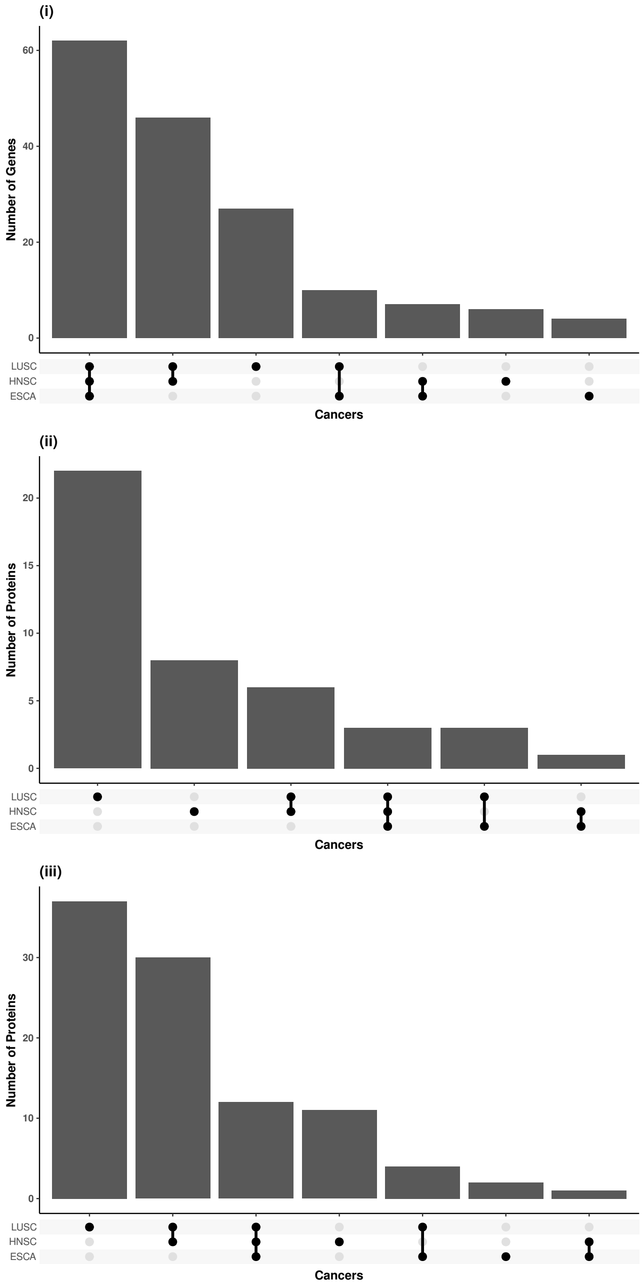


Figure S3.19: Mechanistic model upset plots for the pan-squamous cancers. Upset plots exhibit the number of genes (panel A) or proteins (panels B-C) that are at the decisive level of significance (lBF $>$ 2) for the (A) driver gene, (B) driver protein, and (C) cascading protein mechanistic models respectively, stratified by intersections across cancers.

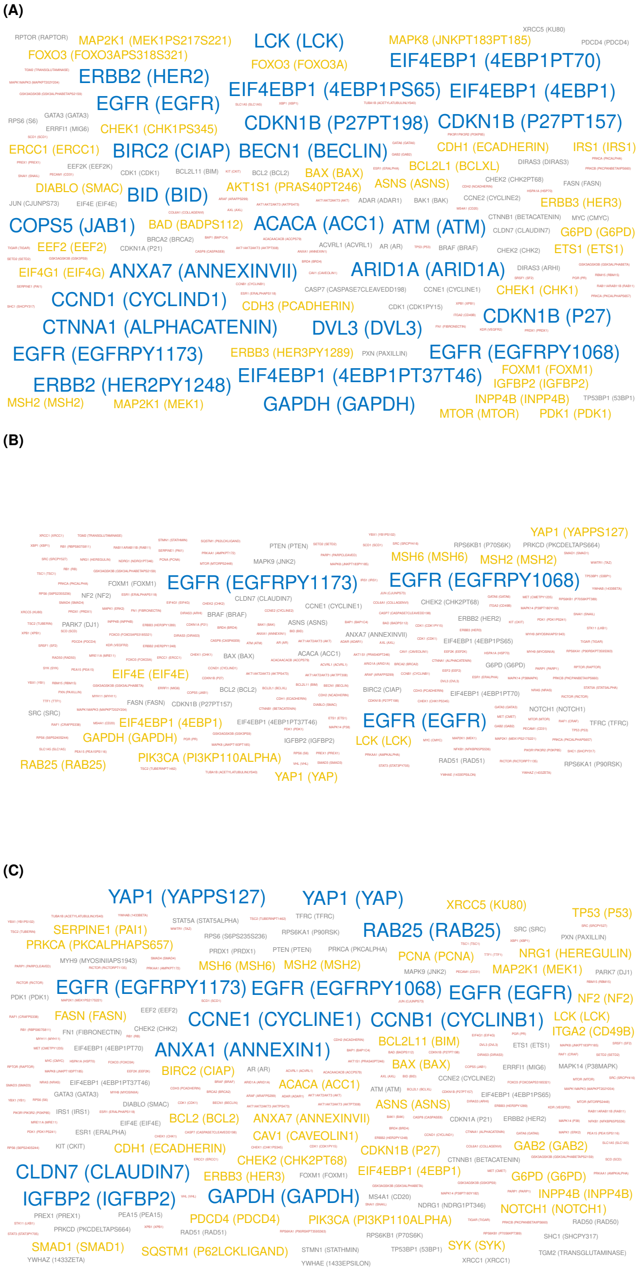


Figure S3.20: Mechanistic model word clouds for the pan-kidney cancers. Word clouds summarize pan-cancer mechanistic model results for genes (panel A) or proteins (panels B-C) for the (A) driver gene, (B) driver protein, and (C) cascading protein mechanistic models. The size of the gene/protein names are proportional to (no. of cancers where the gene/protein is at the decisive level of significance)^3. Here, decisive is defined as lBF > 2.

CHAPTER IV

Bayesian Evidence Synthesis for Multi-system Multiomic Integration

4.1 Introduction

With the advent of sophisticated techniques and platforms, large-scale datasets covering multiple layers of cellular omics are becoming increasingly available (Subramanian et al., 2020; Conesa and Beck, 2019). Consistent advancements have been made in the last few years towards adding more dimensions to these high-throughput datasets, namely (1) additional to patient-level disease databases, model systems such as cell lines, patient-derived xenografts and organoids are being studied extensively in context of cancer and other diseases (Ruggeri et al., 2014; Kim et al., 2020); (2) assessing clinical information and therapeutic response with omics data to make pharmacogenomic discoveries is becoming increasingly common (Relling and Evans, 2015; Roden et al., 2006). Multiple challenges arise during investigations of such datasets, including but not limited to computational inefficiency, complex nature of associations among the omic variables considered, and the biological interpretability and clinical implications of the results (Tarazona et al., 2021). Specifically in context of cancer, the necessity to not only detect biomarker associations with drug/treatment regimens but also to assess the functional relevance and mechanism of such associations is paramount, potentially guiding future therapeutic advances. Thus, novel algorithms that integrate multi-omics patient and model systems profiles can potentially reveal novel biomarkers, drug targets and predictive models in cancer.

Multi-dimensional data integration in cancer To address the wide range of complexity and variability in both detection and management of cancer, a number of multi-omics approaches have been able to uncover intricate molecular mechanisms and discover prognostic candidates (Chakraborty et al., 2018). Data integration approaches have proven particularly useful - both vertical (multiple experiments on a common cohort of samples) (Kaplan and Lock, 2017; Cheng et al., 2015) and horizontal (meta-analysis of different cohorts) (Angel et al., 2020; Tu et al., 2015) integration methods have been developed (Tseng et al., 2015). To simultaneously identify pharmacogenomic associations and corresponding functional mechanisms, singular usage of either of these dimensions is insufficient due to the richness of the currently available omics databases. Multi-omics patient databases of cancer such as The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013), while rich in transcriptomic, proteomic and other levels of omics profiles, do not typically provide comprehensive and systematic drug response on the same cohort of patients, restricting utilization of these profiles directly in pharmacogenomic contexts. Model systems databases such as the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) and Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al., 2012) provide both molecular profiles and drug sensitivity information on the same set of models, but the cancer- or lineage-specific sample sizes of such databases are lower than their patient counterparts and association models built solely on them may suffer from the lack of sufficient statistical power to detect all the true signals. In this work, I propose a solution to this, based on a multi-stage hierarchical Bayesian framework that synthesizes information from both patient and model system databases across multiomic levels to *improve the identification of novel cancer driver genes and association with drug responses*.

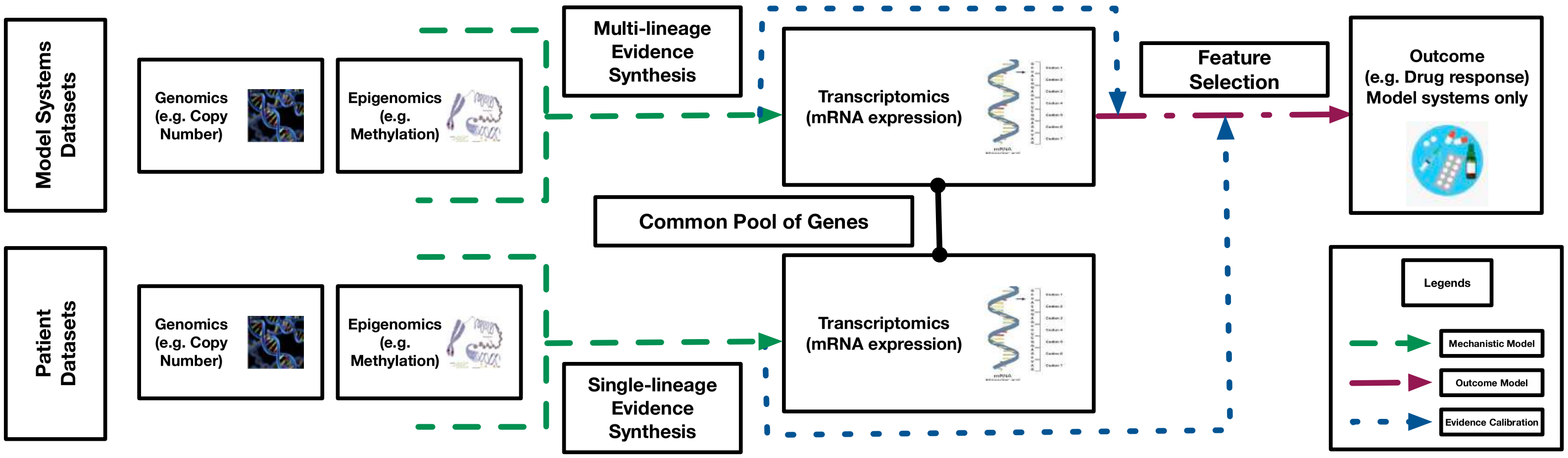

Model Systems Datasets
Patient Datasets
Genomics (e.g. Copy Number)
Epigenomics (e.g. Methylation)
Genomics (e.g. Copy Number)
Epigenomics (e.g. Methylation)
Multi-lineage Evidence Synthesis
Single-lineage Evidence Synthesis
Transcriptomics (mRNA expression)
Transcriptomics (mRNA expression)
Common Pool of Genes
Feature Selection
Outcome (e.g. Drug response) Model systems only
Legends
Mechanistic Model
Outcome Model
Evidence Calibration

Figure 4.1: Conceptual schematic of the BaySyn framework. Multi-lineage evidence synthesis is performed for the model systems datasets with lower sample sizes. Single-lineage evidence synthesis is performed for the patient datasets with larger sample sizes. The two sets of learned evidence are combined on a gene-by-gene basis and an outcome model for performing selection of genomic biomarkers is calibrated using the combined evidence.

A Bayesian evidence synthesis procedure Our integrative framework is called BaySyn: a multi-stage hierarchical Bayesian evidence synthesis pipeline for analysis of multi-system multiomic data. The first stage identifies *cancer driver genes* by detecting transcriptomic associations with upstream changes, which are then utilized to *inform biomarker association models* in the second stage to improve selection. Specifically, the first stage uses additive Gaussian process regression models to detect potential nonlinear associations of gene expression data with corresponding copy number and methylation profiles for both cell line cancer lineages and patient cancer types. To tackle the issue of lower sample size in cell line data, I propose multi-lineage versions of these mechanistic models that can deconvolve lineage and upstream main effects as well as any potential interactions, in addition to single-lineage versions of the same. Evidence synthesized across a common pool of genes from the two sources is then used in a calibrated Bayesian variable selection procedure in the second stage to identify genes having high association with an outcome variable of interest, such as drug response data. Specifically, the evidence quantifications from the mechanistic models are used in these outcome models to upweight the prior probability of selection of different biomarkers in a spike-and-slab prior setting. A conceptual schematic of the procedure is presented in Figure 4.1, providing a high-level summary of the multi-model system evidence synthesis through the mechanistic models and calibrated biomarker selection via the outcome models. I apply our framework to multiomic CCLE and TCGA datasets from pan-gynecological cancers (breast, ovary, and uterus lineages). Our mechanistic models provide cancer-specific and cross-lineage evidence that implicate several relevant functional genes such as PTPN6 and ERBB2 in the KEGG adherens junction gene set. Furthermore, our outcome model is able to make higher number of discoveries in drug response models than its uncalibrated counterparts under the same thresholds of type I error control, including detection of known lineage-specific biomarker associations such as

BCL11A in breast and FGFRL1 in ovarian cancers.

The rest of the paper is organized as follows. Section 4.2 summarizes the multi-stage data integration framework. Section 4.3 describes the CCLE and TCGA data processing and analysis procedures, along with summarization of interesting results. I finish with a brief discussion of our proposed procedure and findings in Section 4.4. All the processed datasets, R codes for the pipeline, and the complete set of real data results are available for access via an interactive R Shiny dashboard at `https://bayesrx.shinyapps.io/BaySyn`.

4.2 The BaySyn Framework

Multi-stage integration pipeline I propose BaySyn, a multi-stage hierarchical Bayesian evidence synthesis pipeline for multi-omics and multi-systems data, as outlined in Figure 4.1. For a given set of samples (patients/model systems), I build gene-specific mechanistic models to infer functional relevance of the genes in the samples of interest based on the association of the gene's expression pattern with its upstream covariates such as copy number changes or DNA methylation. Particularly, in case of model systems, certain cancer lineages may *contain a low number of samples and the mechanistic models may suffer from a lack of sufficient statistical power* to identify true associations with upstream factors. Therefore, I build two versions of the mechanistic models depending on the sample size scenarios - a multi-lineage model that can borrow strength across samples from different lineages (used in this work for modeling the cell line samples; Section 4.2.1), and a single-lineage version that can be applied to a set of samples from a single cancer lineage/type (used in this work in context of the patient samples; Section 4.2.2). Based on statistical summaries of significance of the upstream factors for each gene from these models, I then build the outcome-specific Bayesian hierarchical variable selection models (outcome models, in short; Section 4.2.3) that can incorporate such prior information and borrow strength to improve selection of genes. The specifics of each type of model are described in full detail in the rest of this section.

4.2.1 Multi-lineage Mechanistic Models

Mechanistic models For the mechanistic models, I investigate a gene of interest specifically in relation with its upstream factors to detect whether it is a functional driver, and repeat the procedure across the complete pool of genes included in the analyses. This approach offers a highly parallelizable framework, and the efficiency only depends on the *computational resources* used by each individual model. Further, the class of genomic associations with upstream factors that I are interested in may be highly nonlinear, as has been indicated in past cancer literature (Solvang et al., 2011; Litovkin et al., 2014). Therefore, I intend to equip our models with sufficiently *flexible specifications that can identify a broad range of association patterns*. Keeping these useful features in mind, I describe the mathematical details of the multi-lineage mechanistic models in this subsection and single-lineage mechanistic models in the next subsection.

Notations I begin with setting up some notations. Let M denote the number of lineages across which I intend to borrow strength in a single mechanistic model, and let $\{n_1, \dots, n_M\}$ denote the lineage-specific sample sizes, with $n = \sum_{c=1}^{M} n_c$ being the total sample size. Across a total of $j \in \{1, \dots, q\}$ genes, let G_{ij} denote the (continuous) normalized expression data for the j^{th} gene in the i^{th} sample. Let L_i denote the lineage (tissue/cancer type) of the i^{th} sample, and let $\mathbf{U}_{ij} = (U_{ij1}, \dots, U_{ijp_j})^T$ denote the $p_j \times 1$ vector of upstream information from sample i matched to gene j. Our mechanistic models are gene-specific, allowing different sample sizes for each gene. However, for simplicity of notations, I describe the models assuming a fixed n.

Model structure For the j^{th} gene, I build an additive multi-lineage mechanistic model containing separable components for the main effects of lineage and each upstream covariate, along with any possible interactions of lineage with the upstream factors. Assuming the G_{ij}s to be mean-centered, the general mathematical form of such a model is presented in the following equation.

$$G_{ij} = \underbrace{f_{1j}(L_i)}_{\text{Lineage main effect}} + \underbrace{\sum_{v=1}^{p_j} f_{2jv}(U_{ijv})}_{\text{Upstream main effects}} + \underbrace{\sum_{v=1}^{p_j} f_{3jv}(L_i, U_{ijv})}_{\text{Interaction effects}} + \underbrace{\varepsilon_{ij}}_{\text{Error}}, \tag{4.1}$$

$\forall i \in \{1, \ldots n\}$. The simplest choice is to specify each component $f_{\bullet}$ as a linear model. Such models have been explored in context of cancer omics (Wang et al., 2013b). Although they are computationally simple, they may not be fully able to capture the general range of cellular association patterns. An obvious nonlinear extension is to use splines to construct piece-wise linear mean profiles. Such approaches have also been explored in this context (McGuffey, 2015). However, there are multifold challenges – including specifying the number of knots (hence the degree of adaptable nonlinearity) and increasing computational intensity with increasing number of covariates. To build a general class of additive association models while maintaining a reasonable extent of computational efficiency, I use Gaussian process (GP) models.

To build an additive GP model with interaction effects, I adapt an existing approach proposed in context of longitudinal data (Timonen et al., 2021). In a repeated measures setting, this approach provides a way to incorporate sample-level baseline effects and treatment effects in a nonlinear fashion. I extend this idea to our scenario to include *lineage-level baseline effects* (treating the experiments on cell lines from the same lineage akin to a repeated experiment setting) and *changes in the effects of upstream covariates across different lineages*. While samples belonging to cancers sharing some larger group-specific commonalities (e.g. all gynecological cancers) may share patterns of mechanistic impacts of upstream platforms on gene expressions, there may still be cancer-specific differences in the exact effects. Briefly, I use a GP equipped with a zero-sum (zs) kernel for the main effect of the categorical lineage variable, one with an exponentiated quadratic (eq) kernel for the main effects of the continuous upstream variables, and a product of the zs and eq kernels for their interactions, following existing approaches (Kaufman and Sain, 2010; Timonen et al., 2021). I now discuss the specifics of the GP model along with the prior choices.

GP specification and priors I build a Gaussian likelihood by first assuming $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_j^2)$, and I then build the $f_\bullet$ components using GP priors for each component. For each component $f_\bullet$, let us define $f^{(\bullet)} := [f_\bullet(\mathbf{x}_1), \ldots, f_\bullet(\mathbf{x}_n)]^T$ where $\mathbf{x}_i$ generally denotes the vector of all possible covariates for sample i. I assume that this has a multivariate normal prior as $f^{(\bullet)} \sim \mathbf{N}_n(\mathbf{0}, \mathbf{K}^{(\bullet)})$ where the $n \times n$ covariance matrix has entries $\mathbf{K}_{ih}^{(\bullet)} = \alpha_\bullet^2 \mathscr{k}_\bullet(\mathbf{x}_i, \mathbf{x}_h)$. There are a few things worth noting about this model. First, the components $f_\bullet$ are assumed independent *a priori*, hence their sum is also a zero-mean GP with kernel $\mathscr{k}(\mathbf{x}, \mathbf{x}') = \sum_\bullet \alpha_\bullet^2 \mathscr{k}_\bullet(\mathbf{x}, \mathbf{x}')$ (Williams and Rasmussen, 2006). Second, each $\alpha_\bullet^2$ controls the marginal variance of the corresponding component, while the base kernel function $\mathscr{k}_\bullet$ controls the component's shape and the induced covariance structure. Third, in our applications, kernels $\mathscr{k}_{1j}$ and $\mathscr{k}_{2jv}$s are functions of the lineage and upstream covariates only, respectively, while the kernels $\mathscr{k}_{3jv}$s are functions of all available covariates. The final step in the model building is to choose the specific kernel functions for each component according to the types and scales of the covariates they take as inputs, as described below.

1. For component 1 (one categorical covariate), I use the zero-sum (zs) kernel. The marginal variance is denoted by α_{1j}^2. Assuming that the model includes samples from

a total of M lineages, the kernel is defined as $\mathscr{k}_{1j}(L_i, L_h) = \begin{cases} 1 & \text{if } L_i = L_h, \\ \dfrac{1}{1-M} & \text{else.} \end{cases}$ Note that this choice of kernel function is equivalent to assuming the lineage effects follow a standard random effects model with a zero-sum constraint on the random effects. Namely, this is equivalent to assuming that $f_{1j}(L) \sim N(0, \alpha^2_{1j})$ for $L = 1, \ldots, M$ independently, with the constraint that $\sum_{L=1}^{M} f_{1j}(L) = 0$.

2. For component 2 (only continuous upstream covariates), I use the exponentiated quadratic (eq) kernel on each covariate. The kernel for the v^{th} upstream covariate corresponding to gene j is defined as $\mathscr{k}_{2jv}(U_{ijv}, U_{hjv}) = \exp(-\dfrac{(U_{ijv} - U_{hjv})^2}{2l^2_{2jv}})$. The j^{th} marginal variance is denoted by α^2_{2jv}.

3. For component 3 (interactions between categorical lineage information and continuous upstream covariates), I use the product of zs and eq kernels on each interaction. For the v^{th} interaction, the kernel is defined as $\mathscr{k}_{3jv}((L_i, U_{ijv})^T, (L_h, U_{hjv})^T) = \mathscr{k}_{1j}(L_i, L_h)\exp(-\dfrac{(U_{ijv} - U_{hjv})^2}{2l^2_{3jv}})$ following existing approaches (Kaufman and Sain, 2010). The j^{th} marginal variance is denoted by α^2_{3jv}.

Each marginal standard deviation $\alpha_{\bullet}$ is given a Student-t^+_{20} prior, and each length-scale parameter $l_{\bullet}$ is given a Log-Normal(0, 1) prior, independently. The residual variance parameters σ^2_j are assigned an Inverse-Gamma(2, 1) prior.

Model fitting The interest is in building mechanistic models that would allow us to test for different main and interaction effects of interest. Due to the nature of the zs kernel, the interaction components will also have the zero-sum property (Timonen et al., 2021), which makes it simple to extract and interpret the interaction effects separately. I use a dynamic Hamiltonian Monte Carlo (HMC) sampler as implemented in the R package `lgpr` to obtain draws from the posterior distributions of the parameters, and arrive at the posterior of the functional components analytically for Gaussian likelihood (Timonen et al., 2021). While selection of the specific components and hence covariates is possible based on ranking Bayesian variable relevance statistics or following a minimal subset selection-type approach using such statistics, I are more interested in quantifying the significance of the main and interaction effects as separate collectives, and follow the approach described below.

Model comparison and hypothesis testing Since I are interested in evaluating the roles of lineage, upstream factors, and any possible interactions in explaining the variability in gene expressions, I are interested in testing the following hypotheses for the j^{th} gene.

1. **Lineage main effect:** $H_{0Lj} : f_{1j} = \text{constant}$.

2. **Upstream main effects:** $H_{0Uj} : f_{2jv} = \text{constant}, \forall v \in \{1, \dots, p_j\}$.

3. **All upstream effects:** $H_{0UIj} : f_{2jv}, f_{3jv} = \text{constant}, \forall v \in \{1, \dots, p_j\}$.

To perform these tests, I need to be able to construct models that contain the additive components of interest and compare them against submodels without those components. I use log-posteriors of the parameters in a model to perform the model comparisons, computing HMC-based pseudo-Bayes factors (pBF$_{\bullet j}$s) as scalar summaries of component significance. First, I describe the models I construct and the log-posterior (LP) quantities for each below. Here $\mathbf{G}_{.j} = (G_{1j}, \dots, G_{nj})^T$.

(M1) **Lineage-only model:** Components $f_{2\bullet}$ and $f_{3\bullet}$ in Equation (4.1) are not included. The expression for the log-posterior of this model is given below. Here $A = -\frac{n}{2}\ln(2\pi)$ and $B = \ln \frac{2\Gamma(10.5)}{\sqrt{20\pi}\Gamma(10)}$ are constants free of the model parameters and data. $\boldsymbol{\Sigma}_{1j} = \mathbf{V}_{0j} + \mathbf{V}_{1j}$ where $\mathbf{V}_{0j} = \sigma_j^2 \mathbf{I}_n$ and $\mathbf{V}_{1jih} = \alpha_{1j}^2 ꝁ_{1j}(L_i, L_h)$.

$$\begin{aligned}
LP_{1j} &= \ln[\mathscr{p}(\mathbf{G}_{\bullet j}|\alpha_{1j}, \sigma_j^2).\mathscr{p}(\sigma_j^2).\mathscr{p}(\alpha_{1j})] \\
&= \ln[(2\pi)^{-\frac{n}{2}}|\mathbf{\Sigma}_{1j}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mathbf{G}_{\bullet j}^T\mathbf{\Sigma}_{1j}^{-1}\mathbf{G}_{\bullet j}\}.\Gamma(2)^{-1}\sigma_j^{-6}\exp(-\sigma_j^{-2}) \\
&\quad .\frac{2\Gamma(10.5)}{\sqrt{20\pi}\Gamma(10)}(1+\frac{\alpha_{1j}^2}{20})^{-10.5}I(\alpha_{1j}>0)] \\
&= A + B - \frac{\ln|\mathbf{\Sigma}_{1j}| + \mathbf{G}_{\bullet j}^T\mathbf{\Sigma}_{1j}^{-1}\mathbf{G}_{\bullet j}}{2} - 6\ln\sigma_j - \sigma_j^{-2} - 10.5\ln(1+\frac{\alpha_{1j}^2}{20}) + \ln I(\alpha_{1j}>0).
\end{aligned}$$

(M2) **Upstream-only model:** Components f_1 and $f_{3\bullet}$ in Equation (4.1) are not included. The expression for the log-posterior of this model is given below. Here $\mathbf{\Sigma}_{2j} = \mathbf{V}_{0j} + \mathbf{V}_{2j}$ where $\mathbf{V}_{2jih} = \sum_{v=1}^{p_j} \alpha_{2jv}^2 \mathscr{k}_{2jv}(U_{ijv}, U_{hjv})$.

$$\begin{aligned}
LP_{2j} &= \ln[\mathscr{p}(\mathbf{G}_{\bullet j}|\alpha_{2j1},\ldots,\alpha_{2jp_j}, l_{2j1},\ldots,l_{2jp_j},\sigma_j^2).\mathscr{p}(\sigma_j^2).\prod_{v=1}^{p_j}\{\mathscr{p}(\alpha_{2jv})\mathscr{p}(l_{2jv})\}] \\
&= \ln[(2\pi)^{-\frac{n}{2}}|\mathbf{\Sigma}_{2j}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mathbf{G}_{\bullet j}^T\mathbf{\Sigma}_{2j}^{-1}\mathbf{G}_{\bullet j}\}.\Gamma(2)^{-1}\sigma_j^{-6}\exp(-\sigma_j^{-2}) \\
&\quad .\prod_{v=1}^{p_j}\{\frac{2\Gamma(10.5)}{\sqrt{20\pi}\Gamma(10)}(1+\frac{\alpha_{2jv}^2}{20})^{-10.5}I(\alpha_{2jv}>0)\} \\
&\quad .\prod_{v=1}^{p_j}\{\frac{1}{\sqrt{2\pi}l_{2jv}}\exp(-\frac{(\ln(l_{2jv}))^2}{2})\}] \\
&= \frac{n+p_j}{n}A + p_jB - \frac{\ln|\mathbf{\Sigma}_{2j}| + \mathbf{G}_{\bullet j}^T\mathbf{\Sigma}_{2j}^{-1}\mathbf{G}_{\bullet j}}{2} - 6\ln\sigma_j - \sigma_j^{-2} \\
&\quad - 10.5\sum_{v=1}^{p_j}\ln(1+\frac{\alpha_{2jv}^2}{20}) + \sum_{v=1}^{p_j}\ln I(\alpha_{2jv}>0) \\
&\quad - \sum_{v=1}^{p_j}\ln(l_{2jv}) - \frac{1}{2}\sum_{v=1}^{p_j}(\ln(l_{2jv}))^2.
\end{aligned}$$

(M3) **All main effects model:** Components $f_{3\bullet}$ in Equation (4.1) are not included. The expression for the log-posterior of this model is given below. Here $\mathbf{\Sigma}_{3j} = \mathbf{V}_{0j} + \mathbf{V}_{1j} + \mathbf{V}_{2j}$.

$$LP_{3j} = \frac{n+p_j}{n}A + (p_j+1)B - \frac{\ln|\mathbf{\Sigma}_{3j}| + \mathbf{G}_{\bullet j}^T\mathbf{\Sigma}_{3j}^{-1}\mathbf{G}_{\bullet j}}{2} - 6\ln\sigma_j - \sigma_j^{-2}$$

$$
\begin{aligned}
&- 10.5 \ln(1 + \frac{\alpha_{1j}^2}{20}) + \ln I(\alpha_{1j} > 0) \\
&- 10.5 \sum_{v=1}^{p_j} \ln(1 + \frac{\alpha_{2jv}^2}{20}) + \sum_{v=1}^{p_j} \ln I(\alpha_{2jv} > 0) - \sum_{v=1}^{p_j} \ln(l_{2jv}) - \frac{1}{2} \sum_{v=1}^{p_j} (\ln(l_{2jv}))^2.
\end{aligned}
$$

(M4) **Interactions model:** All components in Equation (4.1) are included. The expression for the log-posterior of this model is given below. Here $\mathbf{\Sigma}_{4j} = \mathbf{V}_{0j} + \mathbf{V}_{1j} + \mathbf{V}_{2j} + \mathbf{V}_{3j}$ where $\mathbf{V}_{3jih} = \sum_{v=1}^{p_j} \alpha_{3jv}^2 \mathscr{k}_{3jv}((L_i, U_{ijv})^T, (L_h, U_{hjv})^T)$.

$$
\begin{aligned}
LP_{4j} = &\frac{n + 2p_j}{n} A + (2p_j + 1)B - \frac{\ln |\mathbf{\Sigma}_{4j}| + \mathbf{G}_{\bullet j}^T \mathbf{\Sigma}_{4j}^{-1} \mathbf{G}_{\bullet j}}{2} - 6 \ln \sigma_j - \sigma_j^{-2} \\
&- 10.5 \ln(1 + \frac{\alpha_{1j}^2}{20}) + \ln I(\alpha_{1j} > 0) - 10.5 \sum_{v=1}^{p_j} \ln(1 + \frac{\alpha_{2jv}^2}{20}) + \sum_{v=1}^{p_j} \ln I(\alpha_{2jv} > 0) \\
&- 10.5 \sum_{v=1}^{p_j} \ln(1 + \frac{\alpha_{3jv}^2}{20}) + \sum_{v=1}^{p_j} \ln I(\alpha_{3jv} > 0) - \sum_{v=1}^{p_j} \ln(l_{2jv}) - \frac{1}{2} \sum_{v=1}^{p_j} (\ln(l_{2jv}))^2 \\
&- \sum_{v=1}^{p_j} \ln(l_{3jv}) - \frac{1}{2} \sum_{v=1}^{p_j} (\ln(l_{3jv}))^2.
\end{aligned}
$$

Sequential evidence detection using pBFs Based on these quantities I now perform the tests of hypotheses as follows. I focus first on model M3 to test whether the lineage component has any effect at all, and move on to M4 (including interactions) only if the answer to the previous question is yes. For model M3, let S denote the number of draws from the HMC sampler, and let $\boldsymbol{\phi}_j^{(s)} = (\alpha_{1j}^{(s)}, \alpha_{2j1}^{(s)}, \dots, \alpha_{2jp_j}^{(s)}, l_{2j1}^{(s)}, \dots, l_{2jp_j}^{(s)}, \sigma_j^{(s)})^T$ denote the vector of sampled parameter values at the s^{th} iteration, $s \in \{1, \dots, S\}$. Let $LP_{3j}^{(s)}$ and $LP_{2j}^{(s)}$ denote the values of LP_{3j} and LP_{2j} respectively, evaluated at $\boldsymbol{\phi}_j^{(s)}$.

Let $\text{pBF}_{Lj} = \frac{1}{S} \sum_{s=1}^{S} (LP_{3j}^{(s)} - LP_{2j}^{(s)})$ be defined as the pseudo-Bayes factor for testing $H_{0Lj} : f_{1j} =$ constant (lineage main effect). Note that this quantity is an approximation for the log-Bayes factor (lBF) for comparing models M3 and M2 under equal model priors. To compute the lBF, one has to compute the expected posteriors for each models, followed by taking a ratio of the two quantities, followed by a log. Here, I are computing an empirical average of the difference of log-posteriors of the model parameters based on the HMC samples. I use standard cut-offs for significance used for lBFs at a $\log_{10}(\bullet)$-scale: < 0.5 (no evidence), $0.5 - 1$ (substantial), $1 - 2$ (strong), and > 2 (decisive) (Kass and Raftery, 1995). From now on, by pBF I always mean a quantity already in this scale. If this test indicates non-significance (neither strong nor decisive evidence), I test $H_{0Uj} : f_{2jv} = \text{constant}, \forall v \in \{1, \dots, p_j\}$ (upstream main effects), by comparing M3 with M1 via computing $\text{pBF}_{Uj} = \frac{1}{S}\sum_{s=1}^{S}(LP_{3j}^{(s)} - LP_{1j}^{(s)})/\ln(10)$ following similar notations as before. The mechanistic evidence $\mathcal{E}_{j1}$ is then set equal to pBF_{Uj}.

If pBF_{Lj} falls in the strong or decisive evidence range, I test $H_{0UIj} : f_{2jv}, f_{3jv} = \text{constant}, \forall v \in \{1, \dots, p_j\}$ (all upstream effects), comparing M4 with M1. For model M4, following our previous notations, let S again denote the number of draws from the HMC sampler, and let $\boldsymbol{\psi}_j^{(s)} = (\alpha_{1j}^{(s)}, \alpha_{2j1}^{(s)}, \dots, \alpha_{2jp_j}^{(s)}, \alpha_{3j1}^{(s)}, \dots, \alpha_{3jp_j}^{(s)}, l_{2j1}^{(s)}, \dots, l_{2jp_j}^{(s)}, l_{3j1}^{(s)}, \dots, l_{3jp_j}^{(s)}, \sigma_j^{(s)})^T$ denote the vector of sampled parameter values at the s^{th} iteration, $s \in \{1, \dots, S\}$. Let $LP_{4j}^{(s)}$ and $LP_{1j}^{(s)}$ now denote the values of LP_{4j} and LP_{1j} respectively, evaluated at $\boldsymbol{\psi}_j^{(s)}$. Let $\text{pBF}_{UIj} = \frac{1}{S}\sum_{s=1}^{S}(LP_{4j}^{(s)} - LP_{1j}^{(s)})/\ln(10)$ be defined as the pseudo-Bayes factor for all (main + interaction) upstream effects. Based on pBF_{UIj}, I can then: (1) assign a level of significance using the standard cut-offs as before, and (2) assign the mechanistic evidence for gene j as $\mathcal{E}_{j1} = \text{pBF}_{UIj}$. The entire testing procedure is then performed independently for each of the q genes, as described in Figure 4.2.

4.2.2 Single-lineage Mechanistic Models

These models do not include any lineage main or interaction effects. Thus, from Equation (4.1), the full models reduce to the following for the j^{th} gene, using same notations as before.

$$G_{ij} = \underbrace{\sum_{v=1}^{p_j} f_{jv}(U_{ijv})}_{\text{Upstream main effects}} + \underbrace{\varepsilon_{ij}}_{\text{Error}}, \forall i \in \{1, \dots n\}. \tag{4.2}$$

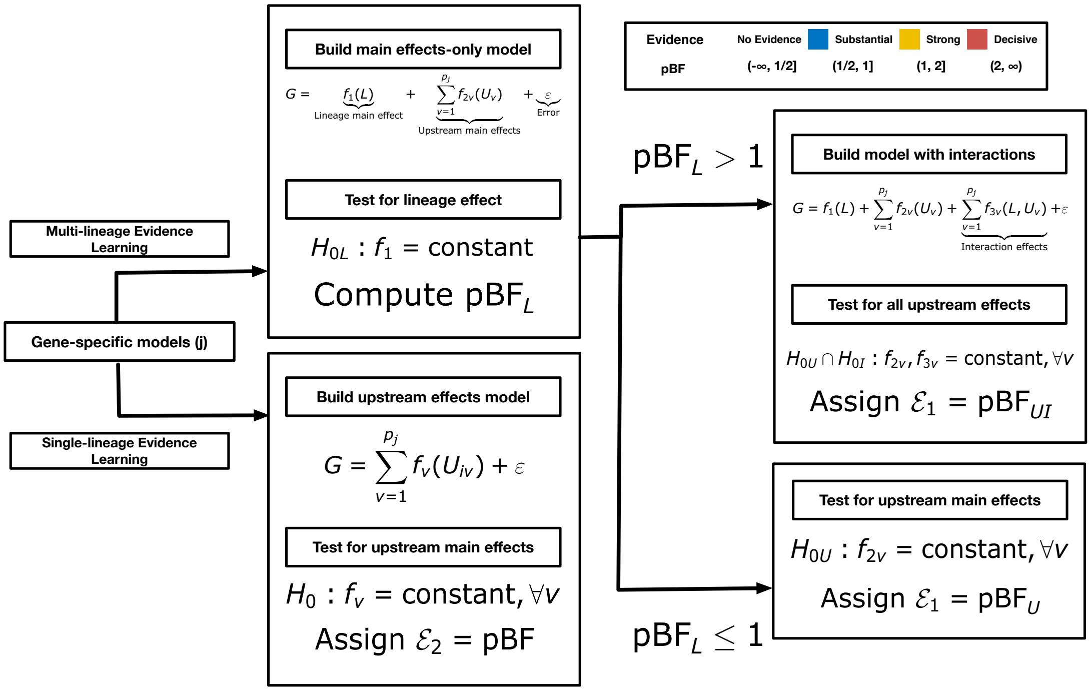

Evidence
No Evidence
Substantial
Strong
Decisive
pBF
(-∞, 1/2]
(1/2, 1]
(1, 2]
(2, ∞)
Multi-lineage Evidence Learning
Gene-specific models (j)
Single-lineage Evidence Learning
Build main effects-only model
$G = \underbrace{f_1(L)}_{\text{Lineage main effect}} + \underbrace{\sum_{v=1}^{p_j} f_{2v}(U_v)}_{\text{Upstream main effects}} + \underbrace{\varepsilon}_{\text{Error}}$
Test for lineage effect
$H_{0L} : f_1 = \text{constant}$
Compute pBF_L
Build upstream effects model
$G = \sum_{v=1}^{p_j} f_v(U_{iv}) + \varepsilon$
Test for upstream main effects
$H_0 : f_v = \text{constant}, \forall v$
Assign $\mathcal{E}_2 = \text{pBF}$
$\text{pBF}_L > 1$
$\text{pBF}_L \leq 1$
Build model with interactions
$G = f_1(L) + \sum_{v=1}^{p_j} f_{2v}(U_v) + \underbrace{\sum_{v=1}^{p_j} f_{3v}(L, U_v)}_{\text{Interaction effects}} + \varepsilon$
Test for all upstream effects
$H_{0U} \cap H_{0I} : f_{2v}, f_{3v} = \text{constant}, \forall v$
Assign $\mathcal{E}_1 = \text{pBF}_{UI}$
Test for upstream main effects
$H_{0U} : f_{2v} = \text{constant}, \forall v$
Assign $\mathcal{E}_1 = \text{pBF}_U$

Figure 4.2: The sequential evidence detection procedure for identifying driver genes within the mechanistic layer of the BaySyn framework. All pseudo-Bayes factors (pBFs) are assumed to be at the $\log_{10}(\bullet)$-scale.

I use the same eq kernel parametrization for the GP priors on each $f_\bullet$ as I used for the $f_{2\bullet}$ components in the multi-lineage models. I now test $H_{0j} : f_{jv} = \text{constant}, \forall v \in \{1, \ldots, p_j\}$ for each gene. I compare the full model in Equation (4.2) with a noise-only null model, as described below.

(M5) **Null model:** Components $f_\bullet$ in Equation (4.2) are not included. The expression for the log-posterior of this model is given below. Here A, $\mathbf{G}_{\bullet j}$, and $\mathbf{V}_{0j}$ are as before.

$$\begin{aligned} LP_{5j} &= \ln[\wp(\mathbf{G}_{\bullet j}|\sigma_j^2).\wp(\sigma_j^2)] \\ &= \ln[(2\pi)^{-\frac{n}{2}}|\mathbf{V}_{0j}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mathbf{G}_{\bullet j}^T\mathbf{V}_{0j}^{-1}\mathbf{G}_{\bullet j}\}.\Gamma(2)^{-1}\sigma_j^{-6}\exp(-\sigma_j^{-2}) \\ &= A - \frac{\ln|\mathbf{V}_{0j}| + \mathbf{G}_{\bullet j}^T\mathbf{V}_{0j}^{-1}\mathbf{G}_{\bullet j}}{2} - 6\ln\sigma_j - \sigma_j^{-2}. \end{aligned}$$

(M6) **Full model:** Components $f_\bullet$ in Equation (4.2) are included. The expression for the log-posterior of this model is given below. Here $\mathbf{\Sigma}_{5j} = \mathbf{V}_{0j} + \mathbf{V}_{5j}$ where $\mathbf{V}_{5jih} =$

$\sum_{v=1}^{p_j} \alpha_{jv}^2 \mathscr{k}_{jv}(U_{ijv}, U_{hjv}) = \sum_{v=1}^{p_j} \alpha_{jv}^2 \exp(-\frac{(U_{ijv} - U_{hjv})^2}{2l_{jv}^2})$, and A, B are as before.

$$
\begin{aligned}
LP_{6j} &= \ln[\mathscr{p}(\mathbf{G}_{\bullet j}|\alpha_{j1}, \dots, \alpha_{jp_j}, l_{j1}, \dots, l_{jp_j}, \sigma_j^2).\mathscr{p}(\sigma_j^2). \prod_{v=1}^{p_j}\{\mathscr{p}(\alpha_{jv})\mathscr{p}(l_{jv})\}] \\
&= \ln[(2\pi)^{-\frac{n}{2}}|\mathbf{\Sigma}_{5j}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}\mathbf{G}_{\bullet j}^T\mathbf{\Sigma}_{5j}^{-1}\mathbf{G}_{\bullet j}\}.\Gamma(2)^{-1}\sigma_j^{-6}\exp(-\sigma_j^{-2}) \\
&\quad . \prod_{v=1}^{p_j}\{\frac{2\Gamma(10.5)}{\sqrt{20\pi}\Gamma(10)}(1 + \frac{\alpha_{jv}^2}{20})^{-10.5} I(\alpha_{jv} > 0)\}. \prod_{v=1}^{p_j}\{\frac{1}{\sqrt{2\pi}l_{jv}} \exp(-\frac{(\ln(l_{jv}))^2}{2})\}] \\
&= \frac{n + p_j}{n}A + p_jB - \frac{\ln|\mathbf{\Sigma}_{5j}| + \mathbf{G}_{\bullet j}^T\mathbf{\Sigma}_{5j}^{-1}\mathbf{G}_{\bullet j}}{2} - 6\ln\sigma_j - \sigma_j^{-2} \\
&\quad - 10.5\sum_{v=1}^{p_j}\ln(1 + \frac{\alpha_{jv}^2}{20}) + \sum_{v=1}^{p_j}\ln I(\alpha_{jv} > 0) - \sum_{v=1}^{p_j}\ln(l_{jv}) - \frac{1}{2}\sum_{v=1}^{p_j}(\ln(l_{jv}))^2.
\end{aligned}
$$

For testing $H_{0j} : f_{jv} = \text{constant}, \forall v \in \{1, \dots, p_j\}$, I compare M6 with M5 via computing $\text{pBF}_j = \frac{1}{S}\sum_{s=1}^{S}(LP_{6j}^{(s)} - LP_{5j}^{(s)})/\ln(10)$, where all the quantities are defined similar to the previous subsection. I assign the mechanistic evidence $\mathcal{E}_{j2} = \text{pBF}_j$ for the gene of interest. As before, I use the following standard significance ranges for these quantities to categorize levels of evidence: < 0.5 (no evidence), $0.5 - 1$ (substantial), $1 - 2$ (strong), and > 2 (decisive) (Kass and Raftery, 1995). This procedure is then performed independently for each of the q genes, as described in Supplementary Figure 4.2.

4.2.3 Outcome Model

For a given pool of genes, it is possible to compute multiple lines of evidence ($\boldsymbol{\mathcal{E}}_j = (\mathcal{E}_{j1}, \dots, \mathcal{E}_{jE})^T$ for gene j). For example, for a given gene j, I may compute *one pBF from a multi-lineage model* built on cell line samples, and *another pBF from a single-lineage model* built on patient samples ($E = 2$). With interest in some disease- or therapy-related phenotype/outcome Y and the selection of biomarkers associated with it, the goal is to *inform the outcome model* about any level of evidence captured in these $\mathcal{E}_{je}$s in a covariate-specific way to possibly improve selection.

1. Sufficiently strong evidence in favor of a covariate $\Longrightarrow$ higher prior probability of inclusion.

2. Otherwise, a uniform prior is placed on selection/non-selection for that particular covariate.

As in Section 3.2.3, I utilize a hierarchical Bayesian setting with calibrated spike-and-slab priors, described below. Let Y_i be the mean-centered continuous outcome for the i^{th} sample. As described in Section 3.2.3, simple extensions to categorical/censored outcomes are possible, but in this work I only focus on continuous outcomes. The mathematical form of the updated calibrated Bayesian variable selection (cBVS) model is then the following.

$$Y_i = \sum_{j=1}^{q} \underbrace{\beta_j}_{\text{Gene expression coefficients}} G_{ij} + \underbrace{\eta_i}_{\text{Error}}, i \in \{1, \dots, n\}. \tag{4.3}$$

Model and prior specifications The errors η_i are iid $\text{N}(0, \tau^2), \forall i \in \{1, ..., n\}$. A standard conjugate prior is used for $\tau^2 \sim$ Inverse-Gamma$(\frac{\nu}{2}, \frac{\nu\lambda}{2})$. Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^T$ denote the q-dimensional vector of regression coefficients. I place a calibrated hierarchical spike-and-slab prior on $\boldsymbol{\beta}$, similar to Section 3.2.3.

$$\begin{aligned}
\boldsymbol{\beta} | \boldsymbol{\delta}, \tau &\sim \mathbf{N}_q(\mathbf{0}, \mathbf{D}_{\delta,\tau}), \\
\delta_j | \theta_j &\sim \text{Bernoulli}(\theta_j), \quad \forall j \in \{1, ..., q\}, \\
\theta_j &\sim \text{Beta}\Big(\mathcal{F}(\boldsymbol{\mathcal{E}}_j), \frac{1}{\mathcal{F}(\boldsymbol{\mathcal{E}}_j)}\Big), \quad \forall j \in \{1, ..., q\}.
\end{aligned}$$

Here $\mathbf{D}_{\delta,\tau} = \tau^2 \mathbf{A}_\delta$, where $\mathbf{A}_\delta$ is the $q \times q$ diagonal matrix $\mathbf{A}_\delta = \text{diag}\{\delta_1 v_1 + (1 - \delta_1) v_0, \dots, \delta_q v_1 + (1 - \delta_q) v_0\}$ and $v_1 \geq v_0 > 0$ are respectively the slab and spike variances. The binary latent variables δ_j are variable inclusion indicators with $\delta_j = 1$ meaning that the j^{th} variable is included in the model. $\mathcal{F}$ is a calibration function mapping the evidence vector $\boldsymbol{\mathcal{E}}_j = (\mathcal{E}_{j1}, \dots, \mathcal{E}_{jE})^T$ to the prior covariate inclusion probability θ_j. The advantages of the hierarchical formulation coupled with the evidence calibration function $\mathcal{F}$ are multifold. First, by adapting $\mathcal{F}$, our framework allows the user to incorporate other significance quantities (such as p-values) into the final outcome model. Any external upstream information, including categorical and continuous covariates, can be used in the mechanistic layer to compute such summary statistics. Finally, by tuning $\mathcal{F}$ appropriately, our framework allows the user to control the impact of the prior information on selection, as I show below.

Choice of evidence calibration function I use a calibration function $\mathcal{F}$ on $\mathbb{R}^E \to [0,1]$ to aggregate multi-dimensional prior evidence into a scalar prior probability. This calibration function has to serve two primary purposes. First, it should be able to aggregate multi-dimensional prior evidence into a scalar prior probability, which means it needs to be a function on $\mathbb{R}^E \to [0,1]$. Second, since the evidence quantities in our case belong to a continuous and nondecreasing spectrum of evidence strength, the function needs to preserve this unidirectional nature – increment in one source of evidence while keeping the others fixed should result in equal or higher calibrated prior probability. To serve these two purposes, I decompose the function as $\mathcal{F}(\boldsymbol{\mathcal{E}}_j) = \mathcal{F}_1(\mathcal{F}_2(\boldsymbol{\mathcal{E}}_j))$. $\mathcal{F}_2 : \mathbb{R}^E \to \mathbb{R}$ aggregates the multiple lines of evidence to a single scalar value $\bar{\mathcal{E}}_j$. I explore a linear map for this, as $\bar{\mathcal{E}}_j = \mathcal{F}_2(\boldsymbol{\mathcal{E}}_j) = \sum_{e=1}^{E} \omega_{je}\mathcal{E}_{je}$. Here ω_{je}s are convex weights specific to gene j, interpreted as quantifications of the importance of each source of evidence for that gene. Several choices of the ω_{je}s are possible, as described below.

1. **Average evidence:** $\omega_{je} = 1/E$ (takes a simple average of all available evidences).

2. **Maximal evidence:** $\omega_{je} = I\left(\mathcal{E}_{je} = \max_{e' \in \{1,\dots,E\}} \mathcal{E}_{je'}\right)$ (only takes into account the strongest evidence available from any source).

3. **Precision-weighted evidence:** $\omega_{je} = \rho_{je} / \sum_{e=1}^{E} \rho_{je}$ (weights the evidences by some metric of reliability of the evidences, such as $\rho_{je} = \hat{\sigma}_{je}^{-2}$ where $\hat{\sigma}_{je}^2$ is the estimated

noise variance for the source model of $\mathcal{E}_{je}$).

$\mathcal{F}_1 : \mathbb{R} \to [0, 1]$ maps the scalar evidence summary $\bar{\mathcal{E}}_j$ to the beta parameter. In our setting, this function is required to have the following features: for small positive or nonpositive $\bar{\mathcal{E}}_j$ (indicating small to no evidence for gene j) the beta parameter should be close to one, resulting in a prior distribution close to U(0, 1) for θ_j; for larger values of $\bar{\mathcal{E}}_j$ the prior distribution should put increasing mass towards one. The rate of increase is guided by the cut-off ranges for the $\bar{\mathcal{E}}_j$s as described before (Kass and Raftery, 1995). Since these requirements are similar to that of the calibration function used in fiBAG, I use the calibration function from Section 3.2.3 as $\mathcal{F}_1$ here. – namely, $\mathcal{F}_1(\bar{\mathcal{E}}_j) = [[1 + \{\max(\bar{\mathcal{E}}_j, 10^{-6})/3\}^{-2.75}]^{-1} + 1]^4$. As illustrated previously in Figure 3.2, the prior distribution of θ_j shifts from an uniform prior to one concentrated close to one with increase in prior evidence strength.

Variable selection Inference is centered around the posterior $p(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\theta}, \tau | \boldsymbol{Y}, \boldsymbol{G}, \boldsymbol{\mathcal{E}}, \nu, \lambda, v_0, v_1)$, where $\boldsymbol{\beta}, \boldsymbol{\delta}$, and $\boldsymbol{\theta}$ are the $q \times 1$ vectors of all β_js, δ_js, and θ_js respectively, $\boldsymbol{Y}_{n\times 1}$ is the outcome vector, $\boldsymbol{G}_{n\times q}$ is the design matrix, and $\boldsymbol{\mathcal{E}}_{q\times E}$ is the matrix of the $\mathcal{E}_{je}$s. I approximate this using a Gibbs sampler implemented via the `rjags` R package (Plummer et al., 2016). I obtain posterior estimates of the parameters (i.e., $\hat{\beta}_j$s, $\hat{\theta}_j$s, and $\hat{\tau}$) as their corresponding empirical posterior means. Model selection is performed using the collection of $1 - \hat{\theta}_j$ as p-value type quantities and applying a false discovery rate (FDR) control procedure (Baladandayuthapani et al., 2010), as described previously in Section 3.2.4.

The overall BaySyn procedure is then summarized in Algorithm 4.1.

Algorithm 4.1 The BaySyn Procedure

```
procedure MECHANISTIC MODEL
    procedure MULTI-LINEAGE EVIDENCE SYNTHESIS                ▷ For model system data
        for $j$ in $1 \rightarrow p$ do                       ▷ $p$ = number of genes
            Build multi-lineage mechanistic model
            Compute evidence $\mathcal{E}_{j1}$
        end for
    end procedure
    procedure SINGLE-LINEAGE EVIDENCE SYNTHESIS               ▷ For patient data
        for $j$ in $1 \rightarrow p$ do
            Build single-lineage mechanistic model
            Compute evidence $\mathcal{E}_{j2}$
        end for
    end procedure
end procedure
procedure OUTCOME MODEL                                       ▷ For model system data
    procedure EVIDENCE CALIBRATION
        for $j$ in $1 \rightarrow p$ do
            Compute hyperparameter $\mathcal{F}(\boldsymbol{\mathcal{E}}_j)$    ▷ $\mathcal{F}$: calibration function, $\boldsymbol{\mathcal{E}}_j = (\mathcal{E}_{j1}, \mathcal{E}_{j2})^T$
        end for
    end procedure
    procedure CALIBRATED BAYESIAN VARIABLE SELECTION
        Build cBVS model with one outcome and $p$ covariates
        Estimate parameters $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\tau}$
        Apply FDR control on $\hat{\boldsymbol{\theta}}$
    end procedure
end procedure
```

4.3 Multi-system Multi-platform Integrative Analyses of Pan-Gynecological Cancers

I perform an integrative analysis of cancer cell lines data from CCLE and patient samples from TCGA (Barretina et al., 2012; Weinstein et al., 2013). Using multi-lineage mechanistic models for cell line samples and single-lineage mechanistic models for patient samples, I quantify gene-specific associations of expression with corresponding copy number and methylation data. I then use the pBFs from these two sources to inform and build cBVS models of *drug response* on gene expression based on the cell line samples. Specifically, our multi-lineage mechanistic models on the cell line samples borrow strength by combining data across three gynecological lineages - *breast, ovary, and uterus*. The single-lineage mechanistic models on the patient samples are built separately for each of the three corresponding TCGA cancer types by tissue - *breast invasive carcinoma (BRCA), ovarian serous cystadenocarcinoma (OV), and uterine carcinosarcoma (UCS)*. The outcome models on the cell line samples are built in a lineage-specific way for a collection of drugs of interest in gynecological cancers. Our investigations are aimed broadly at answering two sets of questions.

1. I assess within-system and between-system patterns of functional evidence garnered by the mechanistic models (i.e., a gene may have strong mechanistic evidence of association with the upstream factors for the cell lines only, the patients only, both, or none).

2. I identify panels of genes whose expressions are associated with responses to specific drugs in the cell line samples, potentially offering novel introspection into treatment selection and the cellular mechanisms/targets of such drugs.

4.3.1 Data Processing and Analysis Pipeline

Multi-omics cell line and patient data Gene expression, copy number, and DNA methylation data on cancer cell lines from CCLE, drug response data from GDSC, along with annotation information to match genes to upstream information, are downloaded from the depmap portal (Tsherniak et al., 2017). Gene expression, copy number, and DNA methylation data on TCGA patient samples, along with annotation information matching genes to upstream covariates, are downloaded from the Xena browser (Goldman et al., 2020). Only the genes satisfying the following set of requirements in the cell lines data from CCLE are included in all analyses.

1. Minimum sample size of 100 across breast, ovary, and uterus lineages.
2. At least two matched upstream covariate (copy number or methylation) available in the dataset
3. The coefficient of variation (CV, percentage scale) across the merged multi-lineage gene expression data is at least 25 (genes with too low CV have low variability in the samples of interest and are less informative for the second-stage cBVS model).

These cleaning steps result in a panel of 5,792 genes that pass all the tests. All these genes are included in the mechanistic and outcome model building procedures. Expression data for each gene is mean-centered before the analyses. Among the drugs available in the CCLE dataset, only those with at least 20 samples in all three lineages were included in the outcome model analyses, resulting in a total of 65 drugs/treatments. IC50 values (log-scale) are used as outcome variables in the drug response models, after mean centering for each drug × lineage combination. Summary information on each dataset are available in Figure S4.1-S4.6 and Table S4.1.

BaySyn analysis of gynecological cancers For each gene, a multi-lineage mechanistic model with $M = 3$ (breast, ovary, uterus) is constructed (termed the CL model hereafter) and hypothesis tests are performed as described in Figure 4.2. Further, for each gene, three single-lineage mechanistic models (one for each cancer type – BRCA, OV, UCS) are built on the patient samples and upstream effects are quantified following Figure 4.2. As a post-model fitting investigation, I perform gene set enrichment analyses (GSEA) (Subramanian et al., 2005) using these four sets of evidence (CL, BRCA, UCS, OV) for the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) and gene ontology (GO) gene sets (Ashburner et al., 2000; Consortium, 2021). For our analyses, I use the gene set enrichment (GAGE) procedure implemented in the `gage` R package due to the reason that our pBFs are on a different scale than typical expression levels or fold-change summaries (Luo et al., 2009). The gene set-specific hypothesis that I test is whether the set in question exhibits significantly higher level of activity as summarized by the evidence statistics compared to the genes outside the gene set, due to the unidirectional nature of the pBFs. For each lineage, drug-specific response association models are built using the cBVS procedure, and variable selection is performed using a 10% FDR control threshold.

4.3.2 Results

Utility of borrowing strength to detect mechanistic evidence Figure 4.3a summarizes the number of genes inferred to be at the decisive level of evidence (in favor of associations with corresponding upstream covariates) across the three single-lineage models for each TCGA patient cancer type and the multi-lineage model for the cell lines data. The connected dots at the bottom indicate the intersection of the mechanistic models for which the number of genes summarized by the bar height are decisive. The top three combinations of models in terms of detecting decisive evidence all belong to some combination of the TCGA data sets (BRCA only, BRCA and OV, BRCA and UCS - in decreasing order). However, except for the BRCA dataset which utilizes > 750 samples for all genes to build the mechanistic models, the cell lines mechanistic models borrowing strength across three lineages detect more unique signals (4th in the ranking) than the other TCGA datasets. This further validates the utility of building joint nonlinear association models with main and interaction components that can identify shared patterns of association across smaller datasets which would potentially be missed in dataset-specific models. The list of genes uniquely identified by the cell lines mechanistic model is available in Table S4.2.

KEGG gene set enrichment analyses illustrate utility of mechanistic evidences To assess the utility of the mechanistic evidence quantities and to validate their use in future detection of novel functional drivers, I perform GSEA using the four evidence sources and the KEGG and GO gene sets. I only discuss the KEGG results here; the GO results are presented in the shiny app at `https://bayesrx.shinyapps.io/BaySyn`. Several KEGG gene sets have been implicated to have significant roles generally in cancer (Chen et al., 2017a; Yuan et al., 2018) and specifically in gynecological cancers (Campos-Parra et al., 2016; Yang et al., 2018; Zhang et al., 2018; Chen et al., 2020). The results from our KEGG GSEA are summarized in Figure 4.3b, exhibiting the seven gene sets with FDR-controlled q-value < 0.2. The gene set-specific mechanistic evidences are summarized in Figure 4.3c-d for the top two KEGG gene sets; the rest are presented in Figure S4.7-S4.11. The top gene set identified in the KEGG analyses is the herpes simplex infection pathway (p-value $= 3.88\times10^{-16}$) (Figure 4.3b). This gene set contains a large cluster of genes exhibiting decisive evidence across majority of the mechanistic models, as can be seen in Figure 4.3c. Following these genes are two major clusters - one containing genes at the decisive level for the BRCA, OV, and CL mechanistic models, and one containing genes at the decisive level for all three TCGA cancers. The consistent nature of functional evidence across this gene set is

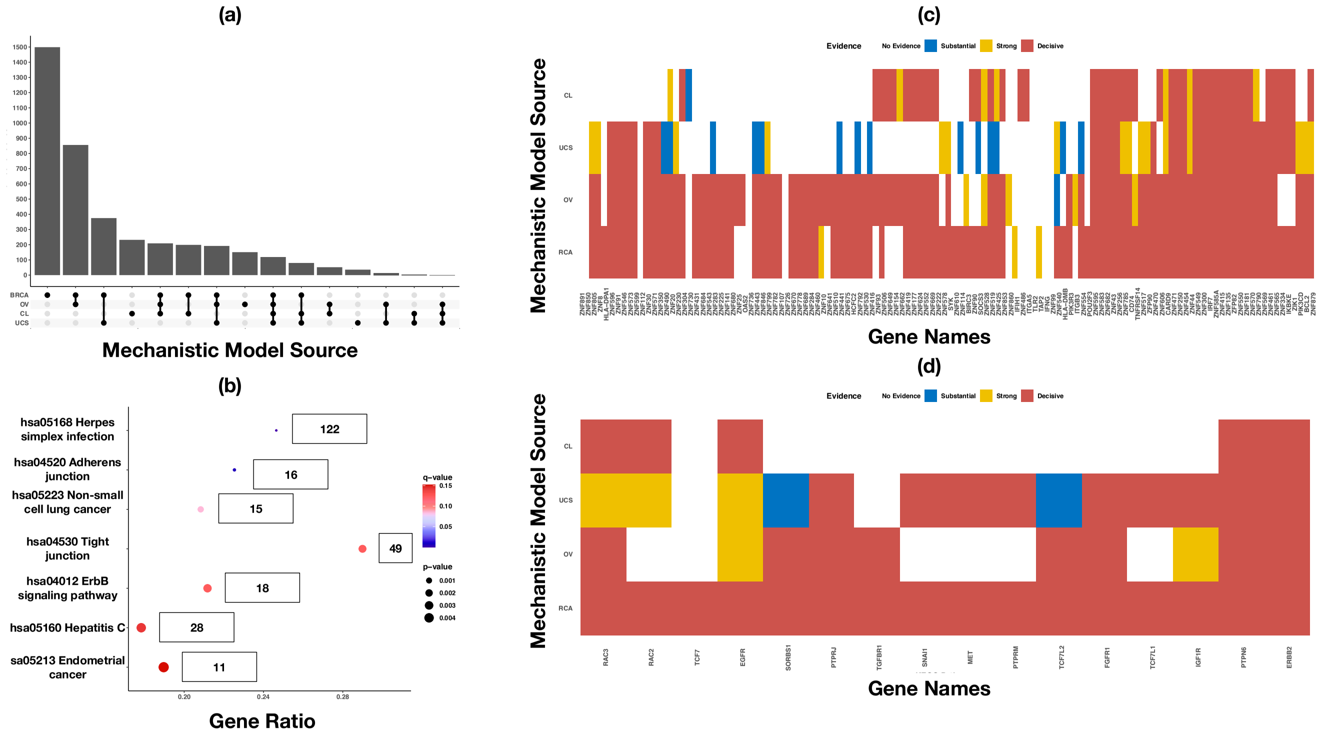


Figure 4.3: Mechanistic evidence summary and gene set enrichment results. Panel (a) presents an upset plot of the number of genes at the decisive level of evidence based on the mechanistic models for different intersections of the patient and cell line datasets. Panel (b) presents a dotplot summarizing significance levels for KEGG gene sets. The gene sets are ordered from top to bottom in decreasing order of q-values (≤ 0.2 included). The labels beside the dots indicate set sizes in our analyses. Panels (c) and (d) present heatmaps summarizing levels of mechanistic evidence for the genes in KEGG herpes simplex infection and adherens junction gene sets respectively. Genes in the rows are ordered based on clusters resulting from the evidence statistics.

in agreement with findings from past investigations - multiple studies have indicated the prognostic value of members of this pathway in gynecological cancers - including breast (Ghouse et al., 2020), ovarian (Nakamori et al., 2003), and endometrial (Zhou et al., 2022) cancer. The second-highest gene set in the KEGG analyses is the adherens junction gene set (p-value = 5.52×10^{-5}) (Figure 4.3b). The genes PTPN6 and ERBB2 exhibit decisive levels of mechanistic evidence in all four models (Figure 4.3d). Different upstream mechanisms of the ERBB2 gene have been implicated in different gynecological cancers, such as copy number changes in ovarian tumors (Dimova et al., 2006) and somatic mutations in breast cancer (Hou et al., 2020). The EGFR gene has also shown promise as a potential therapeutic target in multiple gynecological cancers (Reyes et al., 2014; Kim et al., 2015), which is in alignment with our findings of some signal in all the TCGA and cell line models (Figure 4.3d).

Calibrated drug response models identify high-association lineage-specific biomarkers I build calibrated hierarchical Bayesian variable selection-based drug response models for each lineage × drug combination across all 65 drugs and all three cell line lineages. Figure 4.4a presents a wordcloud where each gene is weighted by the total number of times it is selected in a drug response model at the 10% FDR-controlled cutoff. The genes BAHCC1, ALOX12P2, and SYCP2 emerge as the top candidates, with selection in 14, 12, and 12 models respectively. While this summary allows us to identify general candidates for future pharmacogenomic investigations, it does not indicate any potential lineage-specific utility of these genes. To this end, Figure 4.4b summarizes the number of times the top genes across all drug response models are selected in each lineage. For breast, genes BAHCC1, BCL11A, and SYCP2 are at the top, with respectively eight, eight, and six detected drug associations. *The role of BCL11A in triple-negative breast cancer (TNBC) stemness is well* *known, and it is considered to be one of the first utilizable targets for treatment of TNBCs* (Errico, 2015). A similar confirmation can be obtained for SYCP2, which has recently been identified as a prognostic biomarker in breast cancer (Wu and Tuo, 2019). However, to the best of our knowledge, BAHCC1 has not so far been identified to have breast cancer-specific functional roles, which renders it as a novel detection that deserves deeper investigations. Top genes in the two other lineages also include both novel and known functional drivers - such as ALOX12P2 (nine selections, novel) and FGFRL1 (eight selections, known) (Tai et al., 2018) for ovary and FBXO17 (seven selections, novel) for uterus.

Calibration improves statistical power to detect gene-drug associations To assess the discoveries for specific lineage × drug combinations, I focus on two drugs with known use in specific cancer lineages - docetaxel for breast and cisplatin for ovary. The number of discoveries across different FDR thresholds for these are presented in Figure 4.4c-d and the corresponding discoveries are summarized in Table S4.3-S4.4. Similar plots and tables for all other models are available in our R Shiny dashboard at `https://bayesrx.shinyapps.io/BaySyn`. Evidently, compared to an uncalibrated Bayesian variable selection procedure implemented via the `BMS` R package, cBVS models make more discoveries at the same level of error control, allowing a continuum of assessment for top candidates emerging across increasing control thresholds. This indicates the utility of synthesizing mechanistic evidence and calibrating the outcome models with such evidences. Several examples of cell lines-based discoveries guided by evidences discovered in patient data emerge. For example, the model for docetaxel response in breast cell lines identify an association with the gene GRK5 at 10% FDR control. Cell lines overexpressing GRK5 have previously been observed to demonstrate an increase in resistance to docetaxel in male gynecological cancers (Black et al., 2018), and our finding suggests that it deserves further investiga-

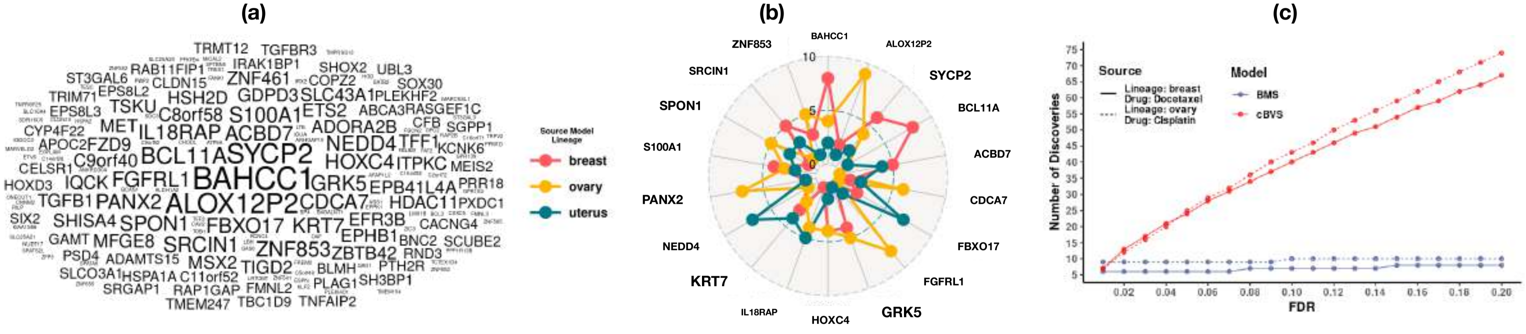


Figure 4.4: Drug response model summaries from BaySyn multi-system multi-omic analyses. Panel (a) presents a wordcloud of top genes across all the drug response models (three lineages × 65 drugs). The sizes of the words are proportional to the total number of times across all models that a gene is selected based on a 10% FDR-controlled threshold. Panel (b) presents a radar chart of the top 18 genes (selected in at least nine drug response models) according to the three lineages. Panel (c) presents a discovery plot across increasing FDR control thresholds for the drug docetaxel in lineage breast and the drug cisplatin in lineage ovary. BMS refers to an uncalibrated Bayesian variable selection model based on the Bayesian model averaging procedure.

tions in female gynecological cancers as well. Another top discovery at the same FDR threshold is the gene CD83, expression of which is known to be enhanced by docetaxel in metastatic breast cancers (Buoncervello et al., 2012). For the response model of cisplatin in the ovarian lineage, multiple solute-carrier family (SLC) genes are selected at the 10% FDR threshold. These genes are known potential biomarkers of ovarian cancer and are under investigation for prognostic utility (Chen et al., 2021). Another interesting discovery is that of the CDCA7 gene from the cell division cycle pathway, silencing of which has recently been shown to downregulate cisplatin resistance in lung cancer subtypes, making it a potential therapeutic target (Zeng et al., 2021b). Our finding seems to indicate similar scope in ovarian cancer, demanding further investigation. Notably, *all four of these discussed findings had no cell lines-based mechanistic evidence, but had decisive evidence from at least one TCGA source* – which further underscores the importance of synthesizing evidence across model systems.

4.4 Discussion and Future Work

Overview I propose BaySyn, a hierarchical multi-stage Bayesian evidence synthesis procedure for multi-system multiomic integration. BaySyn detects functionally relevant driver genes based on their associations with upstream regulators and uses this information to guide variable selection in outcome association models. I apply our framework to multiomic cancer cell line and patient datasets for pan-gynecological cancers. pBFs from the mechanistic layer of BaySyn exhibit high enrichment in previously known KEGG gene sets and detect driver genes known to have functional roles in the cancers studied. Calibrated outcome models for drug responses identify several confirmatory and novel lineage-drug-gene combinations providing further evidence on the profitability of our approach towards future precision oncology endeavors.

General applicability of BaySyn to enriched multi-system multiomic datasets Several features of our framework makes it readily adaptable to more general settings and richer datasets. The calibrated spike-and-slab prior can be generalized to include any number (more upstream platforms such as miRNA or mutation) and form (other evidence metrics such as p-values) of prior information by tuning the calibration function accordingly. The outcome model can easily be extended to include other biomarkers such as proteomics. While I use cell lines data to illustrate the integrative approach across model systems, it is straightforward to apply our pipeline to datasets from cancer model systems with higher fidelity to human tumors (Goodspeed et al., 2016) - such as organoids (Drost and Clevers, 2018) or patient-derived xenografts (Invrea et al., 2020) - as such databases become increasingly comprehensive and available. Further, both the stages of our framework are highly parallelizable and individual runs are quite efficient - a single gene-specific multi-lineage mechanistic model with interactions takes approximately 20 minutes on average to complete, while a single lineage-drug specific outcome model takes approximately 12 minutes on average (both based on runs on a single core of a 2015 Macbook Air with 8 GB memory and Intel i5 processor). Thus, extending our analyses to include larger gene-drug panels with similar sample sizes is straightforward with existing parallel computing resources.

Methodological and scientific limitations Certain improvements are of interest given the biological context of our work. First, although I assess mechanistic relevance at a gene-by-gene basis, at a molecular level, genes interact in functional pathways to result in downstream modifications. This motivates joint models for driver genes in a multivariable setting accounting for underlying gene-gene interactions. Second, the relatively low lineage-specific sample sizes in cell lines data make fully Bayesian exploration of the posteriors feasible in the outcome models. Higher data dimensions would result in increased computation times; where-in approximate Bayesian computation schemes such as the E-M based variable selection (Ročková and George, 2014) or variational Bayes (Fox and Roberts, 2012) would need to be employed. Third, while our framework allows integration of covariate-specific prior information in a variable selection framework, more granular information (both sample- and covariate-specific) may be available, allowing improved learning of the molecular functions driving the changes in an outcome of interest. For example, sample-specific data on tumor heterogeneity may be available, and such data may need to be incorporated in the outcome models driving changes in the covariate effects. Finally, as outlined in Section 4.2, in the presence of multiple lines of evidence, how best to aggregate them depends heavily on the context - while multiple possible approaches exist, a case-specific decision must be made to ensure best utilization of the evidences. A data-driven procedure of choosing evidence weights would eliminate this requirement. I leave these tasks for future exploration.

Reproducibility To ensure easier access to all our integrative analyses results and analysis codes in part of the readers, I have made these resources publicly available at an interactive R Shiny dashboard hosted at `https://bayesrx.shinyapps.io/BaySyn`. I believe that BaySyn, in its current form, will prove to be a useful resource in context of precision oncology and guide future pharmacogenomic investigations.

4.5 Supplementary Figures

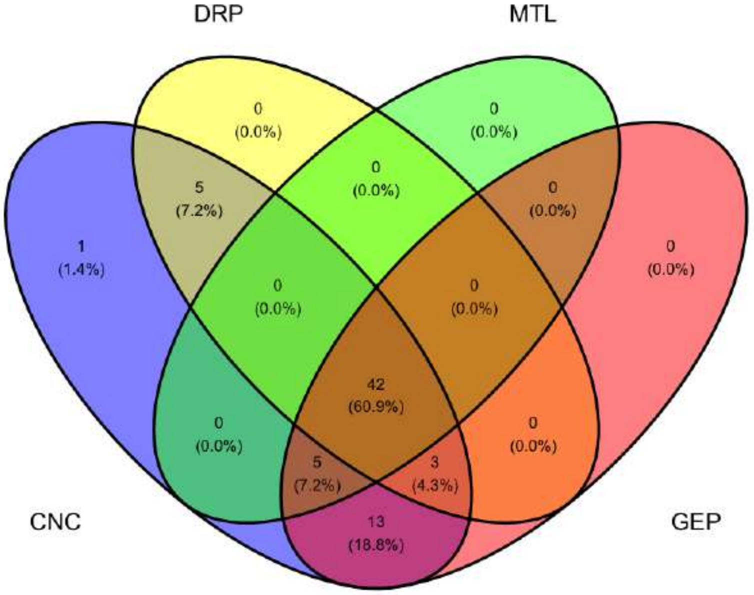


Figure S4.1: Sample size summaries across the different platforms for breast cell lines from CCLE. The acronyms for the different data platforms are as the following - CNC: copy number, GEP: gene expression, MTL: DNA methylation, DRP: drug response.

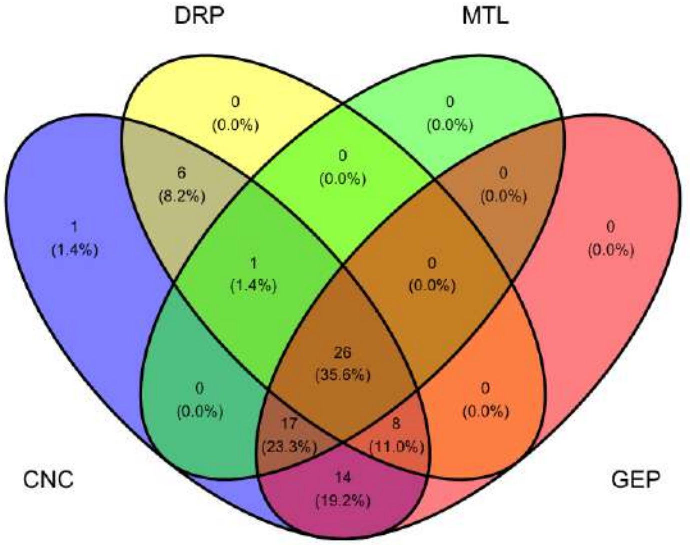


Figure S4.2: Sample size summaries across the different platforms for ovary cell lines from CCLE. The acronyms for the different data platforms are as the following - CNC: copy number, GEP: gene expression, MTL: DNA methylation, DRP: drug response.

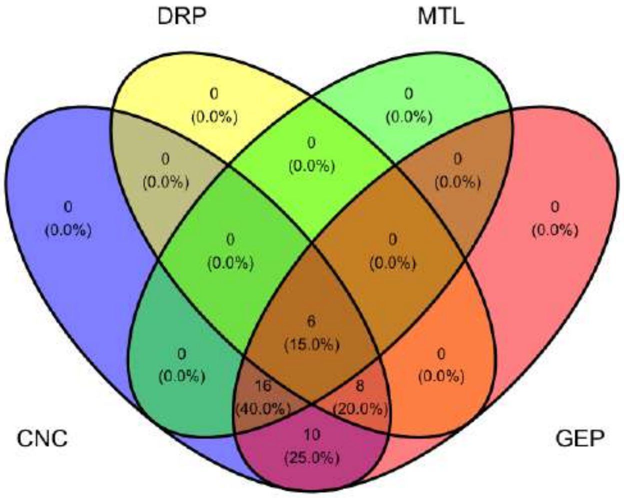


Figure S4.3: Sample size summaries across the different platforms for uterus cell lines from CCLE. The acronyms for the different data platforms are as the following - CNC: copy number, GEP: gene expression, MTL: DNA methylation, DRP: drug response.

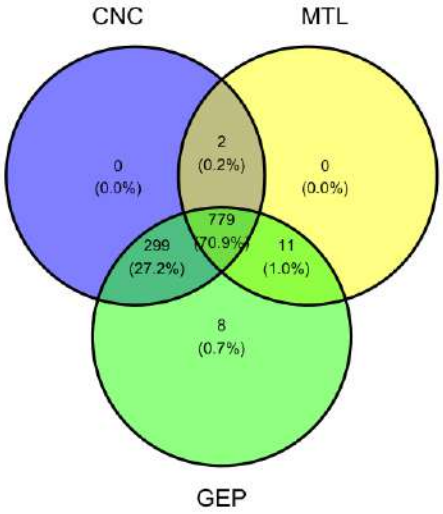


Figure S4.4: Sample size summaries across the different platforms for BRCA patients from TCGA. The acronyms for the different data platforms are as the following - CNC: copy number, GEP: gene expression, MTL: DNA methylation, DRP: drug response.

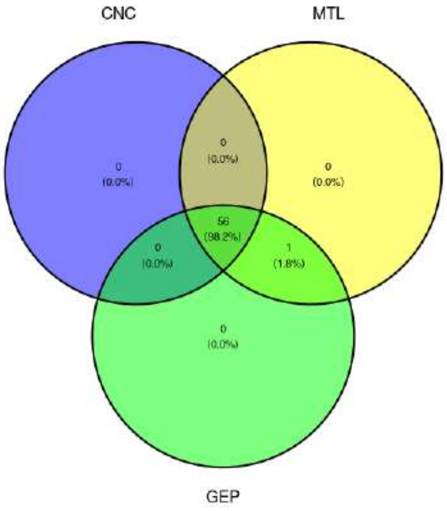


Figure S4.5: Sample size summaries across the different platforms for UCS patients from TCGA. The acronyms for the different data platforms are as the following - CNC: copy number, GEP: gene expression, MTL: DNA methylation, DRP: drug response.

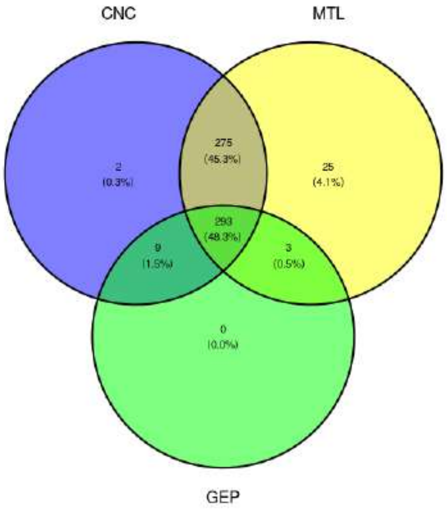


Figure S4.6: Sample size summaries across the different platforms for OV patients from TCGA. The acronyms for the different data platforms are as the following - CNC: copy number, GEP: gene expression, MTL: DNA methylation, DRP: drug response.

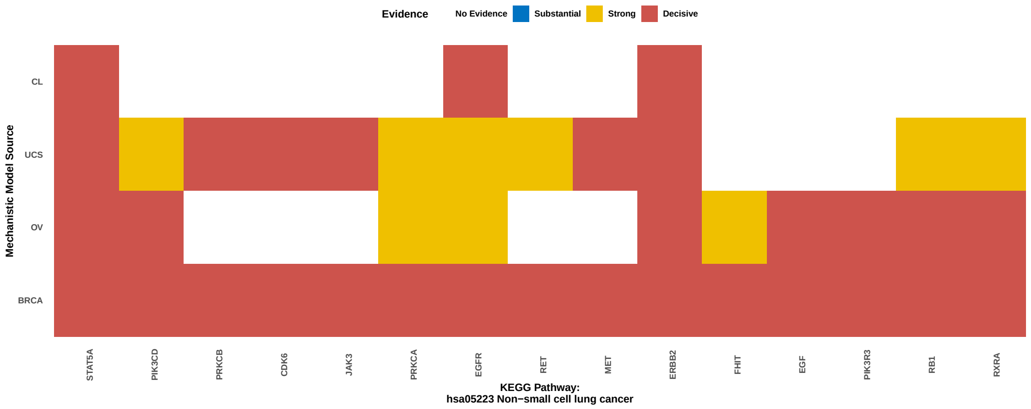


Figure S4.7: Heatmap summarizing levels of mechanistic evidence for the genes in KEGG non-small cell lung cancer gene set. Genes in the rows are ordered based on clusters resulting from the evidence statistics.

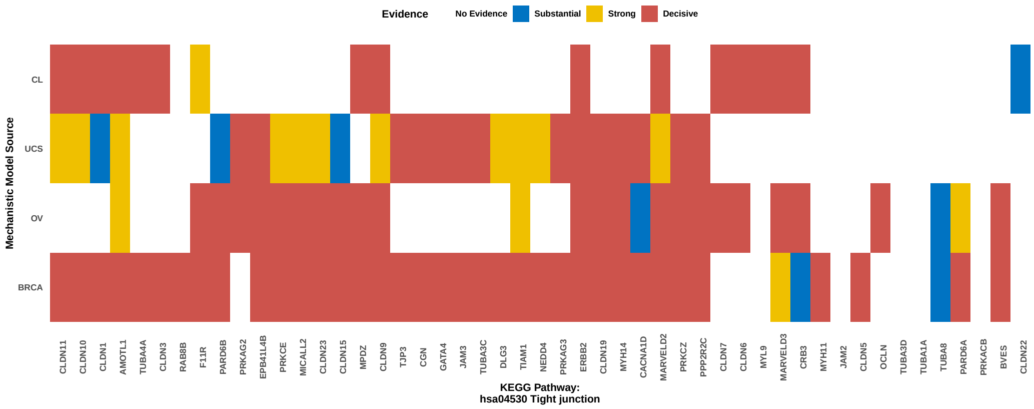


Figure S4.8: Heatmap summarizing levels of mechanistic evidence for the genes in KEGG tight junction gene set. Genes in the rows are ordered based on clusters resulting from the evidence statistics.

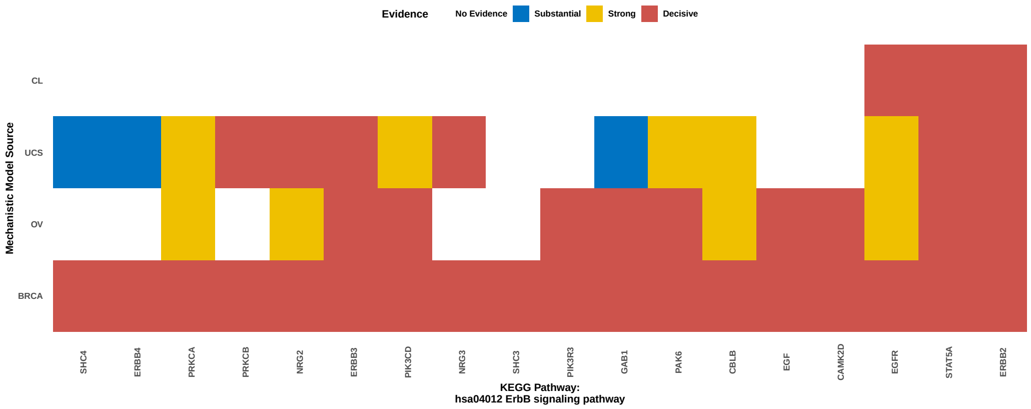


Figure S4.9: Heatmap summarizing levels of mechanistic evidence for the genes in KEGG ERBB signaling gene set. Genes in the rows are ordered based on clusters resulting from the evidence statistics.

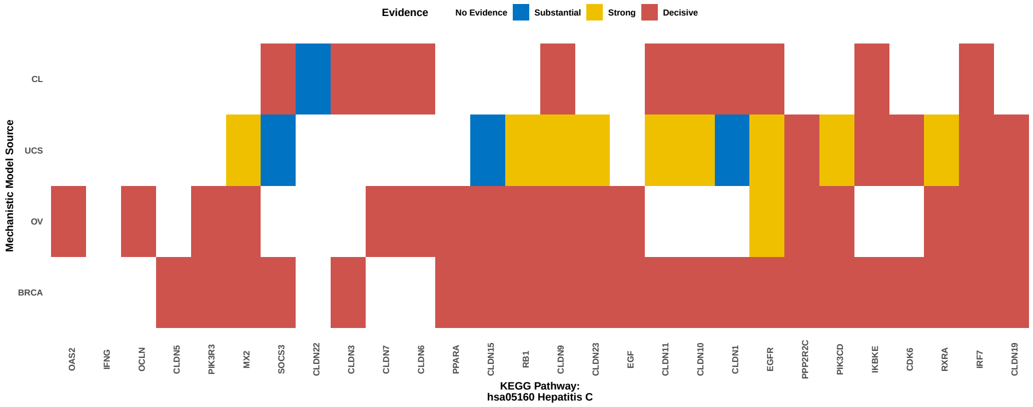


Figure S4.10: Heatmap summarizing levels of mechanistic evidence for the genes in KEGG hepatitis C gene set. Genes in the rows are ordered based on clusters resulting from the evidence statistics.

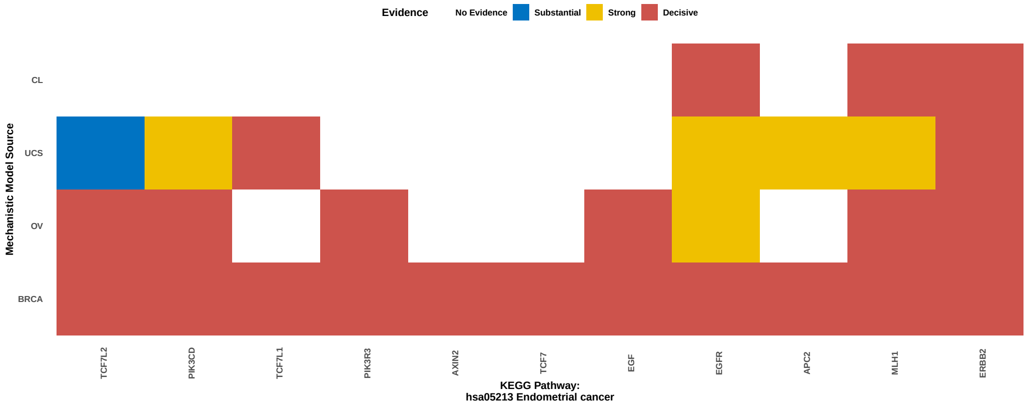


Figure S4.11: Heatmap summarizing levels of mechanistic evidence for the genes in KEGG endometrial cancer. Genes in the rows are ordered based on clusters resulting from the evidence statistics.

4.6 Supplementary Tables

Table S4.1: Summary information on sources and platforms for multiomic cancer cell line and patient data. Only 10 samples are available for the Methylation450 platform for OV, which is why Methylation27 was used instead.

Cell Lines Data (CCLE and GDSC)		
Data	**Platform**	**Download Link**
Gene expression	RNAseq TPM RSEM (log2 transformed using a pseudo-count of 1)	Link
Copy Number	Gene-level (log2 transformed with a pseudo-count of 1). Inferred from WGS, WES or SNP array depending on availability. Calculated by mapping genes onto segment level calls and computing a weighted average.	Link
Methylation	Reduced representation bisulfite sequencing (promoter CpG clusters).	Link
Drug Response	Multiple dose-response parameters available. IC50s used.	Link
Metadata	—	Link
Patient Data (TCGA)		
Data	**Platform**	**Download Link**
Gene Expression	Illumina HiSeq 2000 (log2 transformed RSEM normalized count).	BRCA UCS OV
Copy Number	Gene-level copy number variation (CNV) estimated using GISTIC2.	BRCA UCS OV
Methylation	Illumina Infinium HumanMethylation450 platform (BRCA, UCS) Illumina Infinium HumanMethylation27 platform (OV)	BRCA UCS OV

Table S4.2: Mechanistic evidence summary for genes with decisive evidence only from the cell lines multi-lineage model.

Patient Model			**Genes**
BRCA	**UCS**	**OV**	
No Evidence	No Evidence	No Evidence	See shiny app
Substantial	No Evidence	No Evidence	ADAMTS4, PCDHGA3, RTP3, SMC1B
No Evidence	Strong	No Evidence	AGBL1, CNPY1, PLA2G12B
No Evidence	No Evidence	Strong	APOC2, C16orf86, UPK3B, ZNF626
Substantial	Substantial	No Evidence	AQP12B
No Evidence	Substantial	No Evidence	CASR, MORC1
Strong	No Evidence	No Evidence	CCT8L2, HES5, MAB21L1, MFAP2
No Evidence	No Evidence	Substantial	ELF3, PCDHGA8, PLBD1
Strong	No Evidence	Strong	IER3
Strong	No Evidence	Substantial	RIN1
No Evidence	Strong	Strong	ZNF880

Table S4.3: Summary for genes selected in the docetaxel response model for breast cell lines. PIP denotes the posterior inclusion probability in the calibrated Bayesian variable selection model. The last four columns indicate the level of mechanistic evidence determined by the pBFs from the corresponding models for that gene.

Gene	PIP	CL	BRCA	OV	UCS
LIPH	0.9995	No Evidence	Decisive	Decisive	No Evidence
ZNF728	0.9992	Decisive	No Evidence	No Evidence	No Evidence
TNFRSF25	0.9992	No Evidence	Decisive	Decisive	No Evidence
BTG3	0.9986	Decisive	Decisive	Decisive	Strong
ANXA9	0.9984	Decisive	Decisive	Decisive	Strong
ZIC4	0.9984	No Evidence	Decisive	No Evidence	No Evidence
CD83	0.9983	No Evidence	Decisive	No Evidence	No Evidence
BLMH	0.9983	No Evidence	Decisive	Decisive	No Evidence
SLIT3	0.9982	No Evidence	Decisive	No Evidence	Strong
ITPR1	0.9982	No Evidence	Decisive	No Evidence	Substantial
CTSC	0.9982	Strong	Decisive	Strong	Decisive
KIAA1614	0.9981	No Evidence	Decisive	No Evidence	Strong
GRK5	0.9979	No Evidence	Decisive	No Evidence	Strong
ADAM19	0.9978	Decisive	Decisive	No Evidence	Strong
ZBTB42	0.9977	No Evidence	Decisive	Decisive	Strong
HDAC11	0.9977	No Evidence	Decisive	Decisive	Substantial
MICAL1	0.9975	No Evidence	Decisive	No Evidence	Strong
CNFN	0.9975	No Evidence	Decisive	Decisive	Strong
S1PR2	0.9974	No Evidence	Decisive	Decisive	No Evidence
FSCN1	0.9974	No Evidence	Decisive	Strong	Decisive

Table S4.3: Summary for genes selected in the docetaxel response model for breast cell lines. PIP denotes the posterior inclusion probability in the calibrated Bayesian variable selection model. The last four columns indicate the level of mechanistic evidence determined by the pBFs from the corresponding models for that gene.

Gene	PIP	CL	BRCA	OV	UCS
RHOB	0.9974	No Evidence	Decisive	Substantial	Substantial
MFGE8	0.9973	No Evidence	Decisive	Substantial	Substantial
PRR15	0.9973	Decisive	Decisive	No Evidence	No Evidence
TAF4B	0.9972	No Evidence	Decisive	Decisive	No Evidence
PKNOX2	0.9971	No Evidence	Decisive	No Evidence	Decisive
LRRCC1	0.9971	No Evidence	Decisive	Decisive	No Evidence
AEBP1	0.9971	Decisive	Decisive	No Evidence	No Evidence
SCML2	0.997	No Evidence	Decisive	No Evidence	Strong
LZTS1	0.997	No Evidence	Decisive	No Evidence	Substantial
FAT2	0.997	No Evidence	Decisive	No Evidence	No Evidence
PTK6	0.997	Decisive	Decisive	Decisive	Strong
HAL	0.9968	No Evidence	Decisive	Decisive	No Evidence
ACVRL1	0.9968	No Evidence	Decisive	No Evidence	Strong
TMEM51	0.9968	No Evidence	Decisive	Decisive	Substantial
SLC6A5	0.9967	Decisive	No Evidence	Decisive	No Evidence
ZNF670	0.9967	No Evidence	Decisive	Decisive	No Evidence
NEBL	0.9966	No Evidence	Decisive	Strong	Substantial
SHISA2	0.9966	No Evidence	Decisive	No Evidence	Decisive
CNGA3	0.9966	No Evidence	Decisive	No Evidence	Substantial
RAB39B	0.9966	No Evidence	Decisive	No Evidence	No Evidence

Table S4.3: Summary for genes selected in the docetaxel response model for breast cell lines. PIP denotes the posterior inclusion probability in the calibrated Bayesian variable selection model. The last four columns indicate the level of mechanistic evidence determined by the pBFs from the corresponding models for that gene.

Gene	PIP	CL	BRCA	OV	UCS
WBP2NL	0.9966	No Evidence	Decisive	No Evidence	Strong
BAHCC1	0.9965	No Evidence	Decisive	No Evidence	Decisive
CXXC5	0.9965	No Evidence	Decisive	Decisive	Decisive
PRSS41	0.9965	Strong	Decisive	No Evidence	Strong
SUSD3	0.9965	Decisive	Decisive	Decisive	Substantial
GATA3	0.9965	Decisive	Decisive	No Evidence	Decisive
SDR16C5	0.9964	No Evidence	Decisive	No Evidence	Strong
METRN	0.9964	Decisive	Decisive	No Evidence	No Evidence
ST8SIA6	0.9964	No Evidence	Decisive	No Evidence	No Evidence
MYRIP	0.9963	No Evidence	Decisive	Substantial	Strong
BTN1A1	0.9963	No Evidence	Decisive	No Evidence	No Evidence
FUT11	0.9963	No Evidence	Decisive	Decisive	No Evidence
SLC23A1	0.9963	No Evidence	Decisive	Substantial	Substantial
ZNF283	0.9962	No Evidence	Decisive	Decisive	Substantial
TMEM65	0.9962	Decisive	Decisive	Decisive	Strong
KIF21B	0.9962	No Evidence	Decisive	No Evidence	Substantial
DNAJC5B	0.9962	No Evidence	Decisive	No Evidence	No Evidence
APBB2	0.9962	No Evidence	Decisive	Strong	No Evidence
VAMP1	0.9962	No Evidence	Decisive	Decisive	No Evidence
LINC00226	0.9962	Decisive	No Evidence	No Evidence	No Evidence

Table S4.3: Summary for genes selected in the docetaxel response model for breast cell lines. PIP denotes the posterior inclusion probability in the calibrated Bayesian variable selection model. The last four columns indicate the level of mechanistic evidence determined by the pBFs from the corresponding models for that gene.

Gene	**PIP**	**CL**	**BRCA**	**OV**	**UCS**
LRP12	0.9962	No Evidence	Decisive	Decisive	Substantial
MIPOL1	0.9962	Decisive	Decisive	Strong	Substantial
GPC2	0.9962	No Evidence	Decisive	No Evidence	No Evidence
PRKCA	0.9961	No Evidence	Decisive	Strong	Strong
ISYNA1	0.9961	Decisive	Decisive	Decisive	Strong
RGS20	0.9961	No Evidence	Decisive	Strong	Substantial
ANGPT1	0.9961	No Evidence	Decisive	No Evidence	No Evidence

Table S4.4: Summary for genes selected in the cisplatin response model for ovary cell lines. PIP denotes the posterior inclusion probability in the calibrated Bayesian variable selection model. The last four columns indicate the level of mechanistic evidence determined by the pBFs from the corresponding models for that gene.

Gene	PIP	CL	BRCA	OV	UCS
CDCA7	0.9992	Strong	Decisive	No Evidence	No Evidence
SLC24A3	0.9986	No Evidence	Decisive	No Evidence	No Evidence
SLC27A5	0.9986	No Evidence	Decisive	Decisive	Strong
PXDC1	0.9986	Decisive	No Evidence	No Evidence	No Evidence
PLVAP	0.9985	No Evidence	Decisive	No Evidence	No Evidence
MUC4	0.9985	No Evidence	Decisive	Strong	No Evidence
TNFAIP2	0.9983	Decisive	Decisive	Decisive	Substantial
CYBA	0.9982	Decisive	Decisive	Decisive	Strong
TBL1X	0.9982	No Evidence	Decisive	No Evidence	Decisive
TRIM21	0.9979	No Evidence	Decisive	Decisive	Substantial
EPPK1	0.9979	Decisive	Decisive	Decisive	Strong
CROT	0.9978	Decisive	No Evidence	Decisive	Substantial
PTH2R	0.9978	Decisive	Decisive	No Evidence	Decisive
BANK1	0.9978	No Evidence	Decisive	Decisive	Strong
PRR18	0.9978	No Evidence	Decisive	No Evidence	Strong
KLF2	0.9977	No Evidence	Decisive	No Evidence	No Evidence
MGST2	0.9977	Decisive	Decisive	Decisive	Strong
CADM1	0.9977	No Evidence	Decisive	Decisive	Decisive
ZNF652	0.9977	Decisive	Decisive	Decisive	Strong
ZNF506	0.9977	Decisive	Decisive	Decisive	No Evidence

Table S4.4: Summary for genes selected in the cisplatin response model for ovary cell lines. PIP denotes the posterior inclusion probability in the calibrated Bayesian variable selection model. The last four columns indicate the level of mechanistic evidence determined by the pBFs from the corresponding models for that gene.

Gene	PIP	CL	BRCA	OV	UCS
PCDHGB4	0.9976	No Evidence	No Evidence	Decisive	No Evidence
ESPN	0.9976	No Evidence	Decisive	Decisive	Decisive
ZNF572	0.9976	Decisive	Decisive	Decisive	Decisive
CST6	0.9976	Decisive	Decisive	No Evidence	No Evidence
LTB	0.9976	No Evidence	Decisive	No Evidence	No Evidence
ITGA3	0.9976	No Evidence	Decisive	Decisive	Strong
PODN	0.9975	No Evidence	Decisive	No Evidence	No Evidence
RGS2	0.9975	No Evidence	Decisive	No Evidence	No Evidence
SRR	0.9975	No Evidence	Decisive	Decisive	No Evidence
FZD9	0.9975	No Evidence	Decisive	Decisive	No Evidence
ATP2B2	0.9975	No Evidence	Decisive	Decisive	Decisive
MICAL2	0.9974	No Evidence	Decisive	Decisive	No Evidence
TNFAIP3	0.9974	No Evidence	Decisive	No Evidence	Strong
TRPV2	0.9973	Decisive	Decisive	No Evidence	No Evidence
PHLDB2	0.9973	No Evidence	Decisive	No Evidence	Strong
ASS1	0.9973	Decisive	Decisive	Decisive	Decisive
HOXB7	0.9972	No Evidence	Decisive	No Evidence	Substantial
C5orf38	0.9972	Decisive	Decisive	No Evidence	No Evidence
EPHA2	0.9972	Decisive	Decisive	Decisive	No Evidence
LY6K	0.9972	Decisive	Decisive	Decisive	Decisive

Table S4.4: Summary for genes selected in the cisplatin response model for ovary cell lines. PIP denotes the posterior inclusion probability in the calibrated Bayesian variable selection model. The last four columns indicate the level of mechanistic evidence determined by the pBFs from the corresponding models for that gene.

Gene	PIP	CL	BRCA	OV	UCS
BLMH	0.9971	No Evidence	Decisive	Decisive	No Evidence
EPHB1	0.9971	No Evidence	Decisive	No Evidence	Decisive
BCL2L11	0.997	No Evidence	Decisive	Decisive	No Evidence
SMIM22	0.997	Decisive	No Evidence	No Evidence	No Evidence
F11R	0.997	Strong	Decisive	Decisive	No Evidence
CMTM3	0.997	Decisive	Decisive	Decisive	No Evidence
IL15	0.997	No Evidence	Decisive	Decisive	No Evidence
SLC16A7	0.997	No Evidence	Decisive	Decisive	No Evidence
DLX1	0.997	No Evidence	Decisive	Decisive	No Evidence
LHFPL2	0.997	No Evidence	Decisive	Substantial	Substantial
PAK6	0.997	No Evidence	Decisive	Decisive	Strong
PTPRZ1	0.9969	No Evidence	Decisive	No Evidence	Decisive
SEMA6B	0.9969	Strong	Decisive	Decisive	Decisive
PPM1M	0.9969	No Evidence	Decisive	Substantial	No Evidence
ZNF425	0.9969	Strong	Decisive	Decisive	Substantial
TET1	0.9969	No Evidence	Decisive	No Evidence	No Evidence
EPB41L4A	0.9969	No Evidence	Decisive	Decisive	Strong
KRT8	0.9968	No Evidence	Decisive	Decisive	Decisive
DENND1B	0.9967	No Evidence	Decisive	Decisive	Substantial
OR2T11	0.9967	Decisive	No Evidence	No Evidence	No Evidence

Table S4.4: Summary for genes selected in the cisplatin response model for ovary cell lines. PIP denotes the posterior inclusion probability in the calibrated Bayesian variable selection model. The last four columns indicate the level of mechanistic evidence determined by the pBFs from the corresponding models for that gene.

Gene	**PIP**	**CL**	**BRCA**	**OV**	**UCS**
EVPLL	0.9967	No Evidence	Decisive	No Evidence	No Evidence
SLCO6A1	0.9966	No Evidence	Decisive	No Evidence	No Evidence
IFI27	0.9966	Decisive	Decisive	No Evidence	Decisive
GDPD3	0.9966	No Evidence	Decisive	Strong	No Evidence
AGBL1	0.9966	Decisive	No Evidence	No Evidence	Strong
HOXC4	0.9966	No Evidence	Decisive	Decisive	Decisive
SDR42E1	0.9966	Decisive	Decisive	No Evidence	Decisive
SUSD2	0.9966	No Evidence	Decisive	Decisive	No Evidence
NUPR1	0.9966	Decisive	Decisive	No Evidence	Strong
MCTS2P	0.9966	Decisive	No Evidence	No Evidence	No Evidence
B4GALT6	0.9965	No Evidence	Decisive	Decisive	No Evidence
ST14	0.9965	Decisive	Decisive	Decisive	Decisive
NPEPL1	0.9965	No Evidence	Decisive	Strong	No Evidence
TMC4	0.9965	No Evidence	Decisive	Decisive	No Evidence

CHAPTER V

Bayesian Gaussian Process-based Varying Coefficient Models for Incorporating Tumor Heterogeneity in Clinicogenomic Studies

5.1 Introduction

One of the key characteristics of cancer that needs to be addressed in order to accurately prevent, diagnose, and treat the disease is tumor heterogeneity. Traditionally, tumor heterogeneity was defined and understood in genetic terms, focusing on the intra- and inter-tumor divergence within the same tissue propagated by cellular genetic contributions (Marusyk and Polyak, 2010). More recently, the scientific literature on tumor heterogeneity has expanded its focus to admit epigenetic and potentially non-genetic sources of such variation (Pe'er et al., 2021). In general, diversity within and between tumor cell clusters can be exhibited through modulation of the cellular oncological mechanism via different sources - for example, evolutionary mechanisms may contribute to differential expression of the disease across populations (Heng et al., 2011), non-genetic components to intra-tumor heterogeneity are observed commonly via the tumor epigenome and immune microenvironments (Black and McGranahan, 2021), and even ethnic or demographic features may interact with the disease progression resulting in distinct patterns of incidence and/or mortality (Roshandel et al., 2014). In general, clinicogenomic studies attempt to study clinical outcomes or phenotypes in conjunction with genomic and other omics datasets to in order to identify potential therapeutic targets and assess treatment response (Veltman and Lupski, 2015). While population-level inferences based on integrative approaches implemented in the clinicogenomic context are exciting prospects towards improving therapeutic paths and controlling disease progression, the individuality of a single patient tumor should be taken into account while implementing such procedures to arrive at personalized detection of aberrant cellular function and the treatment of the same. In this regard, the next paragraph provides an overview of the manifestation of tumor heterogeneity within and across cancers and its relevance in integrative quantitative procedures.

The existence and relevance of tumor heterogeneity Several studies from the past few decades have yielded substantial evidence in favor of the existence and role of tumor heterogeneity in multiple cancers. Some examples include breast cancer (Martelotto et al., 2014), colorectal cancer (Zlatian et al., 2015; Testa et al., 2018), prostate cancer (Brady et al., 2021), and lung cancer (Lim and Ma, 2019). Recent precision oncology literature has majorly tended to signify tumor heterogeneity as a systemic barrier against successful treatment – for example, El-Sayes et al. (2021) point out that in the case of immunotherapy, the immune system's response against specific tumor antigens may induce selective bias towards antigen-negative cells, which, in turn, is a common cause of relapse. Thus, quantifying such heterogeneity at a tumor-specific (or even further granular) level and incorporating such quantifications in integrative statistical models carry utmost importance. It is crucial, at this stage of the discussion, to distinguish between inter- and intra-tumoral heterogeneity. The differential behavior between multiple tumors can be inferred from multi-platform molecular profiles of such tumors, obtained via collaborative efforts such as TCGA. On the other hand, even within a single tumor, compositions of cell types may interact with the cancer mechanism and treatment efforts. Quantifying the latter is substantially more challenging at the level of bulk-sequencing data; single-cell and other such deep sequencing procedures offer great potential in this regard (Levitin et al., 2018). This is where the concept of tumor microenvironment becomes crucial – in addition to malignant cells, the population of cells contributing to the heterogeneity in the tumor microenvironment diverse immune cells (lymphocytes, myeloid cells, dendritic cells, etc.), cell types involved in the tumor's blood supply, and other stromal populations. Quantitative summaries that rely on this information, then, can be helpful to guide personalization of statistical integrative models, as is discussed next.

Scientific and clinical importance of the tumor microenvironment Broadly, tumor microenvironment is defined as the ecosystem housing a tumor inside the body, including immune cells, the extracellular matrix, and blood vessels. A tumor and its microenvironment constantly interact with and influence each other, and the result of such interactions may be beneficial or detrimental towards the disease progression. The non-malignant cells and other components have unique immunological capabilities determining the potential of the tumor to progress and survive, and quantifying such influences is necessary to accurately decipher the cancer mechanism (see Arneth (2019) for a comprehensive review). Several quantitative pipelines have been proposed in the recent decades in order to perform this exact task: some of these methods rely on machine-learning based combination of bulk sequencing data with purified immune cell samples (e.g., CIBERSORT by Chen et al. (2018)), some utilize normalization and deconvolution techniques followed by constrained regression procedures (e.g., quanTIseq by Plattner et al. (2020)), and some others perform clustering-driven identification of immune signature sets from a large class of expression signatures (e.g. immune signature scores constructed by Thorsson et al. (2018)). The human tumor immune microenvironment cell-type composition database provides an extensive online compendium of quantifications obtained via more than ten such methods across more than 500 datasets including all TCGA datasets (Wang et al., 2023). Such resources offer an excellent promise towards our overarching goal – developing integrative procedures that can accommodate these quantifications in addition to the multi-system multi-platform datasets explored in the previous chapters will allow individual-specific assessment of tumor characteristics. For example, the outcome models in Chapter III and Chapter IV estimate biomarker associations globally for a system-specific dataset – however, there may be distinct sub-types of the samples driven by different proteogenomic markers, as evidenced in previous cancer studies (breast: Skibinski and Kuperwasser (2015), bladder: da Costa et al. (2018), pancreas: Cros et al. (2018)). Therefore, the associations of the same biomarker with the same outcome may be different depending on which subset of samples we are looking at. To identify such distinctions in an integrated fashion, it then becomes necessary to allow association models to estimate individually varying coefficients. We now discuss the existing statistical procedures allowing such integrative analyses, and specific gaps addressed by our approach.

Association models incorporating individual characteristics The interest in assessing whether the association between two variables is different depending on the value of one or more other variables has generally been explored in previous statistical literature, both within and outside the context of precision oncology. In parametric or semiparametric regression settings, interaction terms between covariates of interest and other variables have been widely used to answer questions similar to that of us. Examples of these include interactions among risk factors in logistic regression models of case-control status (Qiu et al., 2008), gene-environment interactions in logistic and Cox regression models (Wu et al., 2011), and interactions of disease predictors with miRNA functionality in penalized regression models (Qabaja et al., 2013). While such models provide simple and interpretable estimates of the associations of interest, they assume a linear parametric form for the mean of the outcome of interest (modulo the use of a link function), thus restricting any possible shifts due to the interactions to also be linear. As previously discussed in Chapter III and Chapter IV, cellular oncological mechanisms can departure substantially from linear patterns of association. Hence, modeling approaches that capture interactions beyond linear parametric forms are necessary.

A particular class of statistical models that eliminates the requirement of linear interactions relies on using varying coefficients, wherein the regression coefficients themselves are modeled as functions of hierarchical covariates that can impact the association inferred from these coefficients (Hastie and Tibshirani, 1993). Several recent works have explored this class of models utilizing different specifications of the varying coefficient functions to perform efficient selection and estimation. Such approaches include Bayesian hierarchical varying-sparsity regression by Ni et al. (2019), nonparametric varying coefficient spike-and-slab lasso for Bayesian estimation and variable selection by Bai et al. (2019), and varying coefficient models using Bayesian additive regression trees by Deshpande et al. (2020). All these methods address the nonlinearity in the coefficient function estimation using different formulations – Ni et al. (2019) and Bai et al. (2019) utilize a spline-based expansion of the varying coefficients, and Deshpande et al. (2020) use additive regression trees to formulate the same coefficients. In similar spirit, I propose a Gaussian Process-based Varying coeffIcient model using Bayesian variablE Selection (GPVIBES), as described below.

Statistical and scientific novelty GPVIBES relies upon a Gaussian process (GP) regression procedure to model the varying coefficients as explicit functions of the hierarchical covariates. The procedure is developed with an incorporation of Bayesian variable selection priors associated with the varying coefficients themselves. Therefore, GPVIBES offers two distinct axes of statistical advantage: flexibility in modeling a general class of patterns of the modulation of covariate effects (allowed by the GP modeling of the coefficient parameters), and sparsity in terms of the number of nonzero coefficients (controlled by the Bayesian variable selection mechanics). This conflation of estimation and selection, as is shown via synthetic simulation studies, leads to a computationally efficient inference procedure inheriting the merits of the previous approaches while improving scalability of the procedure for high-dimensional datasets. The simulation studies also provide evidence in favor of GPVIBES yielding improved estimation and selection metrics compared to the other varying coefficient-based procedures under similar sample size to number of covariates ratios and at the same level of type I error control. To illustrate the utility of GPVIBES in real clinicogenomic studies, I perform an integrative analysis using data from 16 TCGA cancers on more than 200 proteomic (reverse-phase protein array) expressions, 68 immune signatures summarizing the tumor microenvironment, and overall survival as the outcome of interest. The pan-cancer integrative study identifies several known key signatures, such as the modulation of the association of EGFR and YAP protein expressions with survival by CD8 T lymphocyte proportion in the tumor microenvironment for BRCA.

The rest of the chapter is organized as follows. Section 5.2 describes the GPVIBES procedure including the methodological details regarding the Gaussian process and variable selection specifications, along with details on the computational algorithm. Section 5.3 summarizes the settings and the key results from the simulation studies comparing the performance of GPVIBES with other procedures. Section 5.4 presents the integrative pan-cancer clinicogenomic analysis in conjunction with immune signatures and highlights the key confirmatory and novel results. The chapter is concluded with a discussion on the methodological and scientific aspects of the work and possible future directions in Section 5.5. All the real data results, along with the processed datasets,

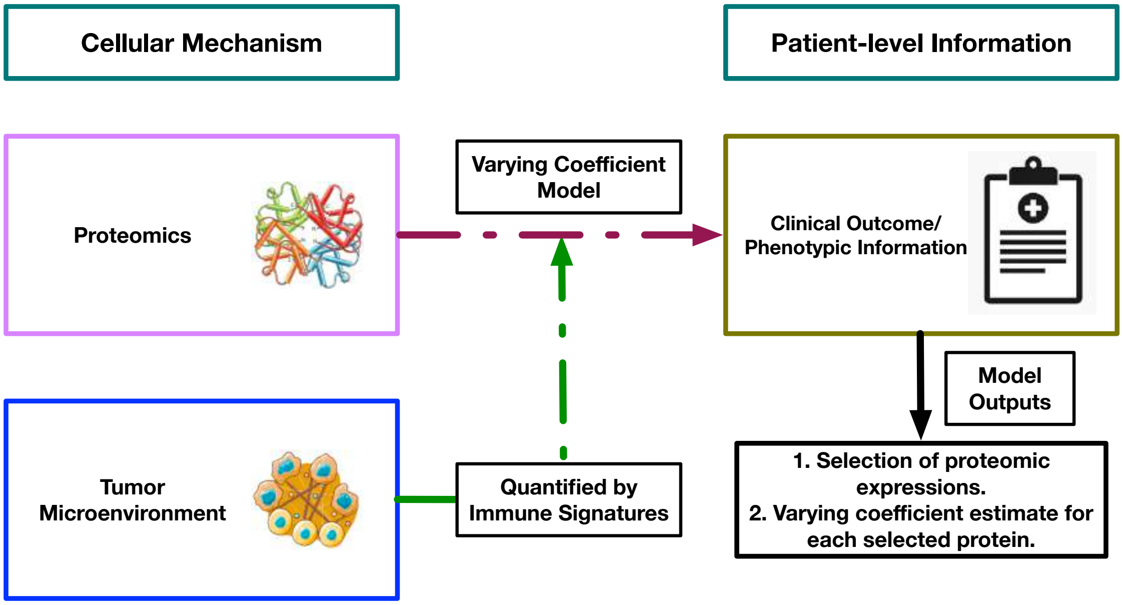

Cellular Mechanism
Patient-level Information
Proteomics
Varying Coefficient Model
Clinical Outcome/ Phenotypic Information
Model Outputs
Tumor Microenvironment
Quantified by Immune Signatures
1. Selection of proteomic expressions.
2. Varying coefficient estimate for each selected protein.

Figure 5.1: Overview of the GPVIBES model in the clinicogenomic context. The Bayesian variable selection model is built with a patient-level outcome modeled on proteomic expressions. The regression coefficients are specified as functions of quantitative summaries of the heterogeneity in the tumor microenvironment.

and the computational codes are available in an interactive R shiny dashboard hosted at `https://bayesrx.shinyapps.io/GPVIBES/`.

5.2 The GPVIBES Model

5.2.1 Notations

Let the dataset of interest consist of n samples (patient tumors in the context of the motivating biological problem), and let i be the index denoting the sample of interest (thus $i \in \{1, \ldots, n\}$). Let it also be assumed that the data is annotated in such a way so that the information from different clinicogenomic platforms can be aligned horizontally, i.e., matched at a sample level. Let the outcome of interest for the i^{th} sample be denoted by Y_i. The variable Y is assumed to be continuous and mean-centered across samples for ease of exposition. Generalizations to other classes of outcomes (such as survival) are straightforward and are discussed in Section 5.2.6. Let p be the total number of covariates of interest, and let X_{ji} denote the value of the covariate j for the sample i. In the scientific context of this project, this denotes the expression of the candidate biomarker j for patient i. For exposition, let us assume a single hierarchical covariate for now, denoted by Z_i for the i^{th} sample. In context of the scientific motivation for this chapter, this will indicate the value of an immune signature of interest for the i^{th} sample. Generalizations to more than one such covariate is also straightforward, and is described in Section 5.2.6.

5.2.2 The Clinicogenomic Biomarker Selection Model with Varying Coefficients

Following the notations introduced above, the fundamental clinicogenomic model with varying coefficients can be generally written as follows.

$$Y_i = \sum_{j=1}^{p} \beta_j(Z_i) X_{ji} + \varepsilon_i. \tag{5.1}$$

The errors ε_i are assumed to be iid with distribution $N(0, \sigma^2)$. $\beta_j(\bullet)$ denotes the varying coefficient function for the j^{th} covariate. Depending on how we mathematically specify the form of the $\beta_j(\bullet)$s, there can be two distinct approaches to building this model. The key difference between the two approaches lies in the fact that one approach attempts to perform the selection of the global effects of the candidate biomarkers and the estimation of the hierarchical covariate effects within them simultaneously via a joint prior structure, and the other approach parametrizes $\beta_j(\bullet)$ in a way such that these two procedures are controlled by separate mechanisms. We discuss these two approaches below.

Simultaneous selection and estimation In this case, we directly specify each $\beta_j(\bullet)$ as a Gaussian process (GP), writing the following.

$$\beta_j(Z_i) = f_j(Z_i). \tag{5.2}$$

The GP specification then follows from putting the following prior structure on the $f_j(\bullet)$s. Let us denote $f_j^{(i)} = f_j(Z_i)$, and $\mathbf{f}_j = (f_j^{(1)}, \dots, f_j^{(n)})^T$. Then, the GP prior is specified as:

$$\mathbf{f}_j \overset{\text{ind}}{\sim} \mathbf{N}_n(\mathbf{0}_n, \mathbf{K}_j). \tag{5.3}$$

Here the $(i, k)^{\text{th}}$ element of the covariance matrix $\mathbf{K}_j$ is given by $\sigma_j^2 K_j(Z_i, Z_k)$, where $K_j(\bullet)$ is a suitable kernel function. We discuss the choice of this function and how it affects the computations in Section 5.2.3. Assuming such a choice is fixed, the selection and estimation procedure then relies on the prior specifications for the σ_js. The general form of a prior that can accommodate our needs is as follows.

$$\sigma_j \sim (1 - \gamma_j)\delta_{\sigma_j = 0} + \gamma_j D_j. \tag{5.4}$$

Here γ_j is a selection indicator for the j^{th} biomarker effect. It is possible to specify prior distributions on these parameters using information from previous studies or other models, such as the calibrated priors discussed in Chapter III. Otherwise, it is possible to place standard beta-binomial type variable selection priors on these parameters. We discuss such choices in Section 5.2.4. D_j, on the other hand, is a prior distribution with support $(0, \infty)$, covering the case when $\sigma_j > 0$, i.e., the j^{th} covariate is deemed to have a non-zero association with Y. This prior may or may not be indexed by j, depending on whether we are interested in placing independent or shrinkage priors on them. Such choices are discussed in Section 5.2.4.

Deconvolved selection and estimation Following Kuo and Mallick (1998), it is possible to rewrite Equation (5.2) as the following.

$$\beta_j(Z_i) = \gamma_j f_j(Z_i). \tag{5.5}$$

All the notations are the same as the previous paragraph. Essentially, this formulation separately specifies priors on each of the components concerning selection and estimation.

For the selection indicators γ_j, we can follow the same prior structures described in Section 5.2.4 depending on whether prioritization of the biomarkers based on prior evidence is necessary. For the functional components due to the hierarchical covariates, the GP specifications can remain the same as described previously. The key difference now is in the specification of the prior for σ_j. Since the γ_js already take care of the biomarker selection separately, we can directly write $\sigma_j \sim D_j$. D_j can then be specified using the same priors described in Section 5.2.4. For all the computations described in the rest of the chapter, the deconvolved selection and estimation specification of the model as described in Equation (5.5) is used.

5.2.3 Kernel Function for Gaussian Process Specification of Varying Coefficients

A common default choice for the kernel functions $K_j(\bullet)$s is the squared exponential (SE) kernel, specified as $K_j(Z_i, Z_k) = \exp(-b_j(Z_i - Z_k)^2)$. A typical parametrization is performed using $b_j = 1/2\lambda_j^2$, where λ_j is interpreted as a length-scale parameter (Ulapane et al., 2020). It is possible to fix the value of this parameter in advance, following approaches such as that of Sun et al. (2020). In this particular approach, several values of the $\hat{\lambda}_j^2$ are chosen on a grid specified by the observed values of $(Z_i - Z_k)^2/2$ over $i, k \in \{1, \ldots, n\}$. However, while this reduces the number of parameters to be dealt with, this does not get rid of the large covariance matrices in the GP prior. Thus, any computational procedure would involve the calculation, addition, and inversions of these matrices, potentially affecting efficiency. Hence, we follow an alternative approach using a modified SE (MSE) kernel, which reduces the variance structures using basis expansions. We follow the same parametrization as used by Wu et al. (2022). The MSE kernel is written as follows.

$$K_j(Z_i, Z_k) = \exp(-a_j(Z_i^2 + Z_k^2) - b_j(Z_i - Z_k)^2). \tag{5.6}$$

As is obvious from the expression, $a_j = 0$ reduces the MSE kernel to the SE kernel. The advantage of using the MSE kernel is that the eigendecomposition of the MSE kernel has a closed form expression. For a general d-dimensional case, if the decomposition includes a maximal polynomial degree of M, then the decomposition involves $L = \binom{M+d}{d}$ basis vectors. Since this formulation involves a GP prior on a single hierarchical covariate, the interest lies in individual effects of the hierarchical covariates, meaning that $d = 1$ in this case. This implies that $L = M + 1$. For the covariance kernel $K_j(\bullet)$, let the eigenvalues be denoted by η_{jl} and the corresponding basis functions are $\psi_{jl}(\bullet)$, $l \in \{1, \ldots, L\}$. Then using this decomposition, our biomarker selection model from Equation (5.1) with the specification described in Equation (5.5) can be written as the following.

$$Y_i = \sum_{j=1}^{p} \gamma_j \sum_{l=1}^{L} \theta_{jl} U_{jli} + \varepsilon_i. \tag{5.7}$$

Here U_{jli}s are our new covariates of interest in the selection model, defined as $U_{jli} = \sqrt{\eta_{jl}}\psi_{jl}(Z_i)X_{ji}$. Note that these are fully known and this reduces our model to a linear form. The θ_{jl}s are now our regression coefficients of interest, with $\theta_{jl} \sim N(0, \sigma_j^2)$, retaining the variance parameters from the GP prior. All the prior choices discussed in the previous subsection can be used as before.

5.2.4 Prior Choices for Selection and Estimation Parameters

For the overall noise variance σ^2, a standard prior of the form $\mathrm{IG}(\alpha_0, \beta_0)$ is used. The other parameters that require the specification of priors for completing the description of the model are the p selection indicator parameters γ_j and the p kernel variance parameters σ_j^2. The rest of this subsection describes the potential prior choices for these.

Priors for selection indicators Both the simultaneous and deconvolved selection and estimation procedures involve the selection indicators γ_js, respectively as a spike-and-slab-type indicator quantity in the priors for the kernel variances, and as an explicit multiplicative component of the varying coefficient. One option is to therefore utilize these parameters to prioritize candidate biomarkers X_js with available prior evidence $\mathcal{E}_j$ (possibly multi-dimensional). As specified in Section 3.2.3, let us write the prior on γ_j as $\gamma_j \sim \text{Ber}(\omega_j)$. Then, it is possible to specify a hierarchical calibrated prior as $\omega_j \sim \text{Beta}(\mathcal{F}(\mathcal{E}_j), 1/\mathcal{F}(\mathcal{E}_j))$. $\mathcal{F}(\bullet)$ can be chosen depending on the type and scale of the evidence quantities and the importance to be placed on the said evidence. For the computations that follow in the rest of this chapter, such a calibrated prior structure is not utilized. Instead, a standard Beta-Bernoulli specification is used, by writing $\gamma_j \overset{\text{iid}}{\sim} \text{Ber}(\rho)$, and then $\rho \sim \text{Beta}(a, b)$. Here ρ represents the prior proportion of selected covariates, thus directly quantifying sparsity. The hyperparameters a, b can either be chosen based on information available about the relevance of the covariates in the model or in a noninformative manner.

Priors for variance parameters Recall that based on notations introduced previously, the marginal prior distribution on the kernel standard deviation parameters σ_j is denoted as D_j. The simplest possible choice of such a prior setting is to assume that the σ_js are independent of each other, and all the D_js are the same. A natural option would be to assume $\sigma_j^2 \sim \text{IG}(\alpha_0, \beta_0)$ under D_j, similar to the prior placed on the overall noise variance. While this leads to a computationally simple specification and straightforward approximations of the posterior, since the σ_j^2s represent the overall variability in GP components due to the same hierarchical covariate, it is prudent to utilize a prior specification that can exploit the possible interdependence between these parameters to ensure lesser false signals. In these lines, a shrinkage-type prior D on $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_p)^T$ can be employed. Some potential choices are discussed next.

- Following Zhang et al. (2014), a possible choice for D is to specify the prior as shown below.

$$\log(\boldsymbol{\sigma}) \sim \mathbf{N}_p\left(\mathbf{0}_p, \tau^2 \begin{bmatrix} \tau_1^2 & & \\ & \ddots & \\ & & \tau_p^2 \end{bmatrix}\right). \tag{5.8}$$

It is then possible to utilize a shrinkage prior as $\tau_j^2 \overset{iid}{\sim} \mathrm{Exp}(\xi^2/2)$, and then some non-informative prior on each of τ and ξ.

- Following Bitto and Frühwirth-Schnatter (2019), it is possible to build a prior on the variance parameters that does not convert them to a log scale, thereby eliminating the potential issue about the interpretation of a zero variance in the log scale. The prior looks as follows.

$$\sigma_j^2 \sim \mathrm{Gamma}(1/2, 1/2\tau_j^2). \tag{5.9}$$

$$\tau_j^2 \sim \mathrm{Gamma}(\alpha, \alpha\xi^2/2). \tag{5.10}$$

  As in the previous case, non-informative priors can then be used for α and ξ.

- An interesting choice of prior is inspired by the idea of the horseshoe prior proposed by Carvalho et al. (2009). Following them, it is possible to write $\sigma_j^2 = \sigma^2\tau^2\tau_j^2$. Then, the prior specification on the local shrinkage parameters is as follows.

$$\tau_j \sim C^+(0, 1). \tag{5.11}$$

  Here $C^+(0, 1)$ is a half-Cauchy distribution, which is also used as the prior distribution for τ.

The horseshoe is particularly interesting in the context of this chapter since it provides a way to perform aggressive shrinkage on the varying coefficients that are constant at 0, thus reducing false discovery further, in conjunction with the selection. The heavy tails induced by the prior still allows the non-zero varying coefficients to move away from zero and be approximated flexibly via the GP. This behavior is confirmed via the simulation studies described in Section 5.3, where the horseshoe-GP combination is observed to perform empirically well in terms of making true discoveries while controling the false positives. Hence, for the remainder of this chapter, all computations are based on this specification.

5.2.5 Conditional Posteriors and their Approximation

Summary of key modeling choices To sum up the discussions and modeling decisions made in the previous subsections, I begin here with a brief overview of how the final model is formulated. The linearized form of the model after the basis expansion of the MSE kernel is given by Equation (5.7). The errors ϵ_i are assumed to be iid with distribution $N(0, \sigma^2)$. The kernel variance parameters are assumed to have the form $\sigma_j^2 = \sigma^2\tau^2\tau_j^2$. The priors on each parameter of interest are as follows.

1. $\sigma^2 \sim \mathrm{IG}(\alpha_0, \beta_0)$.
2. $\theta_{jl} \overset{\text{ind}}{\sim} N(0, \sigma^2\tau^2\tau_j^2)$.
3. $\tau \sim C^+(0, 1)$.
4. $\tau_j \overset{\text{iid}}{\sim} C^+(0, 1)$.
5. $\gamma_j \overset{\text{iid}}{\sim} \mathrm{Ber}(\rho)$.
6. $\rho \sim \mathrm{Beta}(a, b)$.

With this specification, the Bayesian computational task is then to draw samples from the joint posterior of the parameters of interest, namely, $\pi\Big((\sigma, \tau, \boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \rho)^T | \mathbf{Y}, \mathbf{U}, \mathbf{Z}, \alpha_0, \beta_0, a, b\Big)$. Here $\boldsymbol{\tau}$ is the $p \times 1$ vector of the τ_js, $\boldsymbol{\theta}$ is the $pL \times 1$ vector of the θ_{jl}s, $\boldsymbol{\gamma}$ is the $p \times 1$ vector of the γ_js, $\mathbf{Y}$ is the $n \times 1$ vector of the Y_is, $\mathbf{U}$ is the $n \times pL$ matrix of the U_{jli}s, and $\mathbf{Z}$ is the $n \times 1$ vector of the Z_is. I now discuss how to perform this task based on the model formulation.

Conditional posteriors and sampling In all the following expressions, $\pi(\bullet|\mathcal{P}^{(-)})$ indicates the posterior distribution for a parameter given all the parameters but itself, all the hyperparameters, and all data. For some of the parameters, an explicit closed-form conditional posterior can be obtained, and a straightforward sampling procedure can be implemented to update them. For the others, it is necessary to use some existing alternative approaches to perform the update efficiently. I summarize all these steps below.

1. $\pi(\sigma^2|\mathcal{P}^{(-)}) = \mathrm{IG}((n + pL)/2 + \alpha_0, (\mathrm{SSE} + \mathrm{SS}\theta)/2 + \beta_0)$. Here $\mathrm{SSE} = \sum_{i=1}^{n} e_i^2$, where $e_i = Y_i - \sum_{j=1}^{p} \gamma_j \sum_{l=1}^{L} \theta_{jl} U_{jli}$ and $\mathrm{SS}\theta = \sum_{j=1}^{p} \sum_{l=1}^{L} \theta_{jl}^2/\tau^2\tau_j^2$.

2. The global scale parameter τ and the local scale parameters τ_js do not yield a closed-form solution for the conditional posterior. However, once τ is fixed, the τ_js are independent in the posterior under the assumed setting. Hence, following Polson et al. (2014), τ can be updated first, and then the τ_js can be updated in a single block. All these updates are performed via slice sampling, an implementation of which is available in the R package `horseshoe`.

3. $\pi(\theta|\mathcal{P}^{(-)}) \sim \mathbf{N}_{pL}\Big(\mathbf{G}^{-1}\mathbf{U}_\gamma^T\mathbf{Y}, \sigma^2\mathbf{G}^{-1}\Big)$. Here $\mathbf{G} = \mathbf{U}_\gamma^T\mathbf{U}_\gamma + \mathbf{H}^{-1}$, $\mathbf{U}_\gamma$ is the $n \times pL$ matrix of the $\gamma_j U_{jli}$s, and $\mathbf{H} = \tau^2 \mathrm{diag}(\tau_1^2\mathbf{1}_L, \ldots, \tau_p^2\mathbf{1}_L)$. For the high-dimensional settings typical to the motivating biological scenarios in this chapter, it is expensive to perform this stage of the sampling. Hence, I utilize a data augmentation-based alternative approach proposed by Bhattacharya et al. (2016), implemented in the R package `horseshoe`. This algorithm scales linearly with the data dimension.

4. For each j, $\pi(\gamma_j|\mathcal{P}^{(-)}) \sim \mathrm{Ber}\Big(\psi_j/(1 + \psi_j)\Big)$. Here $\psi_j = \dfrac{\rho}{1-\rho}\exp\Bigg(\Big(\sum_{i=1}^{n}\sum_{l=1}^{L}\theta_{jl}U_{jli}\cdot (2Y_i - \theta_{jl}U_{jli})\Big)/2\sigma^2\Bigg)$. Note that these updates can be made independently, and the ψ_js can be computed parallelly after each previous set of updates.

5. $\pi(\rho|\mathcal{P}^{(-)}) \sim \mathrm{Beta}(a + \sum_{j=1}^{p}\gamma_j, b + p - \sum_{j=1}^{p}\gamma_j)$.

Unless otherwise specified, for the rest of this chapter, the sampling procedure as above is implemented with the following choices.

1. $a_j = 0.01, b_j = 100$, and $M = 30$ in the MSE kernel specifications.

2. The hyperparameter values are chosen to be $a = b = 1$, $\alpha_0 = \beta_0 = 0.5$.

3. Two parallel MCMC chains, each with 10,000 burn-ins and 20,000 draws, with a thinning factor of 2.

5.2.6 Generalizations of Interest

Generalization to censored survival outcome It is straightforward to generalize the GPVIBES model as discussed in this section to a censored survival outcome using a log-normal accelerated failure time (AFT) specification. For a censored survival outcome setting where T represents the true survival time variable and C represents the censoring time variable, the observed outcome variables in the data are assumed to be $Y = \log(\min(T, C))$, and event indicator $\delta = I[T < C]$. Then, prior to each sampling step for the parameters as described above, it is necessary to update the censored survival times (i.e. for which $\delta = 0$) by sampling from their conditional distributions given the rest of the data, and all the parameter values at the latest update, and the hyperparameters. This update can be performed using the same truncated normal distribution used in Chapter III for the generalization described in Section 3.2.3.

Generalization to multiple hierarchical covariates In specific settings, it may be of interest to incorporate multiple hierarchical covariates Z_rs in the GPVIBES model. For example, in the context of the biological motivation for this chapter, there may be multiple immune signatures or cell type summaries describing different but potentially interacting components of the tumor microenvironment, and assessing the modulation of the biomarker associations in the collective presence of these signatures may be interesting. In such a scenario, it is straightforward to generalize the GPVIBES model and the sampling algorithm to accommodate multiple hierarchical covariates.

To outline this briefly, let the updated notations be as follows. Let Z_{ri} denote the value of hierarchical covariate r for sample i, $r \in \{1, \dots, R\}$. Then, the deconvolved form of the varying coefficients can be written as the following, generalizing the Gaussian process specifications to additive Gaussian processes (AGP).

$$\beta_j(\mathbf{Z}_i) = \gamma_j \sum_{r=1}^{R} f_{jr}(Z_{ri}). \tag{5.12}$$

In this updated setting, then, let $f_{jr}^{(i)} = f_{jr}(Z_{ri})$, and $\mathbf{f}_{jr} = (f_{jr}^{(1)}, \dots, f_{jr}^{(n)})^T$. The AGP prior is specified as:

$$\mathbf{f}_{jr} \sim \mathbf{N}_n(\mathbf{0}_n, \mathbf{K}_{jr}). \tag{5.13}$$

Here the (i, k)th element of the covariance matrix $\mathbf{K}_{jr}$ is given by $\sigma_{jr}^2 K_{jr}(Z_{ri}, Z_{rk})$, where $K_{jr}(\bullet)$ is the MSE kernel function, expressed in the following way.

$$K_{jr}(Z_{ri}, Z_{rk}) = \exp(-a_{jr}(Z_{ri}^2 + Z_{rk}^2) - b_{jr}(Z_{ri} - Z_{rk})^2). \tag{5.14}$$

The same basis expansion technique as described in Section 5.2.3 can be applied to each kernel separately. For the covariance kernel $K_{jr}(\bullet)$, let the eigenvalues be η_{jrl} and the corresponding basis functions be $\psi_{jrl}(\bullet)$, $l \in \{1, \dots, L\}$. In this case, the aggregated GPVIBES model can be written as the following.

$$Y_i = \sum_{j=1}^{p} \gamma_j \sum_{r=1}^{R} \sum_{l=1}^{L} \theta_{jrl} U_{jrli} + \varepsilon_i. \tag{5.15}$$

Here U_{jrli}s are the new covariates of interest in the selection model, defined as $U_{jrli} = \sqrt{\eta_{jrl}} \psi_{jrl}(Z_{ri}) X_{ji}$. Note that these are again fully known and this reduces the model to a linear form. The θ_{jrl}s are now the regression coefficients of interest, with $\theta_{jrl} \sim N(0, \sigma_{jr}^2)$, retaining the variance parameters from the AGP prior. Then, the horseshoe prior setting can be implemented by decomposing $\sigma^2_{jr} = \sigma^2\tau^2\tau^2_{jr}$. The sampling procedure can accordingly be updated, and all the computational choices regarding the sampling and the hyperparameters can remain the same.

5.3 Simulation Studies

To assess the selection and estimation performance of GPVIBES, I perform two simulation studies where the outputs of GPVIBES are compared to outputs from other varying coefficient-type models. In Simulation 1, the synthetic data is generated from a mechanism with only one hierarchical covariate Z that modulates the effects of the model covariates X on the outcome of interest Y. In Simulation 2, the data is generated from a setting with two hierarchical covariates that modulate the covariate effects additively. In both scenarios, the outcome variable is assumed to be continuous, a total of $p = 100$ covariates are assessed, each setting is replicated 100 times, and standard selection and estimation metrics are summarized, as described in the rest of this section.

5.3.1 Simulation 1: Single Hierarchical Covariate

Data generation Let the sample size in a given simulation setting be denoted by n and the preferred signal-to-noise ratio be denoted by SNR. For a particular simulation replicate, the following steps are performed to generate the hierarchical covariate, covariate, and outcome data.

1. $\mathbf{Z} = (Z_1, \dots, Z_n)^T$ is chosen to be the equidistant grid of size n between the range $[0, 1]$.

2. Among the $p = 100$ signals, the last 90 are assumed to be zero. This means that the varying coefficient function $\beta_j(\bullet) = 0$ for all $j > 10$.

3. For the first five varying coefficients, each is assumed to take a form $\beta_j(z) = c_j f_j(z)$, where c_j is chosen such that the integral L^2-norm of the function on the region $[0, 1]$

is exactly 1. This means that for each $j \in \{1, \dots, 5\}$, $c_j = \left(\sqrt{\int_0^1 f_j^2(z)dz}\right)^{-1}$. The functions chosen are as follows: $f_1(z) = 1$, $f_2(z) = z$, $f_3(z) = z^5$, $f_4(z) = (z - 0.5)^2$, $f_5(z) = \sin(9z)$. The intention behind choosing these functions is to ensure that the performance of the modeling procedures being assessed can be evaluated for all types of functions including constant, linear, polynomial, and more nonlinear than polynomials (such as trigonometric).

4. For each $j \in \{6, \dots, 10\}$, $\beta_j(z) = -\beta_{j-5}(z)$. This is done to ensure that the assessment is performed for both positive and negative functions of the same absolute signal level.

5. Let X_{ji} denote the value of the j^{th} covariate for the i^{th} sample. Then for all j and i, $X_{ji} \overset{\text{iid}}{\sim} \text{N}(0, 1)$.

6. Then, the outcome is generated as $Y_i \overset{\text{ind}}{\sim} \text{N}(\sum_{j=1}^{p} \beta_j(Z_i)X_{ji}, 1/\text{SNR}^2)$ for each i.

The sample size n is varied over $200, 500, 1000, 2000$, and the SNR is varied over $0.5, 1, 1.5, 2$, resulting in a total of 16 simulation scenarios.

Competing methods and model outputs GPVIBES is implemented as described in Section 5.2.5. GPVIBES yields two quantities of interest: an estimated posterior inclusion probability $\hat{\gamma}_j$ for each covariate, and corresponding estimates $\hat{\beta}_j(Z_i)$ for each varying coefficient at each observed value of the hierarchical covariate. I use one competing varying coefficient model with similar flavor, namely, VCBART, which models the varying coefficients using a Bayesian additive regression tree formulation (Deshpande et al., 2020). The number of chains and sampling iterations, along with all other relevant details, are kept exactly the same as that of GPVIBES. The model outputs from VCBART are also exactly same as that of GPVIBES - however, the posterior inclusion probabilities are no more estimates of an explicit selection indicator γ_j; rather, they are posterior average inclusions of a tree corresponding to the hierarchical covariate in the model. Further, two nonparametric varying coefficient modeling procedures are used, following Bai et al. (2019), as implemented in the R package `NVCSSL`. The first of these, denoted by VCFREQ hereon, is a frequentist version of the model using basis expansions with LASSO-based regularization on groups of basis coefficients. For the implementation in this simulation study, the optimal regularization parameter λ is chosen from a grid of size 10,000 between $[10^{-3}, 10^{3}]$, equidistant in the $\log_{10}(\bullet)$ scale, based on minimum AIC of the regularized model. This model provides a binary inclusion decision for each covariate and corresponding estimates of the varying coefficient similar to the previous two procedures. The second version, denoted by VCSSLL henceforth, is a Bayesian version of the model with spike-and-slab lasso specifications. This model is implemented with a fixed slab hyperparameter $\lambda_1 = 1$, and the spike hyperparameter λ_0 chosen from a grid of size 10,000 between $[1, 10^{3}]$, equidistant in the $\log_{10}(\bullet)$ scale, based on minimum AIC of the regularized model. The outputs of this model are the same as that of VCFREQ.

Evaluation metrics for selection and estimation Two classes of metrics are used to evaluate the performance of the competing methods. First, based on the continuous posterior inclusion probabilities or the binary selection indicators, we compute several standard metrics for selection performance based on the true generating mechanism. For the two procedures that yield a continuous posterior inclusion probability for each covariate, variable selection is performed using FDR control on these probabilities at a 10% level, following the same procedure described in Section 3.2.4. Based on these, true positive rate (TPR), false positive rate (FPR), Matthew's correlation coefficient (MCC), and true positive rates specifically for the linear and nonlinear functions (TPRL and TPRNL) are computed. Using a grid of selection cut-offs for GPVIBES and VCBART, and the grid of regularization parameters for VCFREQ and VCSSLL, area under the receiver operating characteristic curve (AUC) and scaled AUC between specificity of 0.8 to 1 (AUC20) are also computed. These metrics are threshold-free and provide an overall evaluation of the selection performances. For estimation, we compare the $n \times p$ varying coefficient matrix $\boldsymbol{\beta}$ and its estimate provided by the procedures, via the mean squared error (MSE). An overall MSE is computed, along with versions of it for specific subsets of the effects: MSETrue (for the first 10 true signals), MSEFalse (for the last 90 signals which are zero), MSELinear (for the four linear effects, $j \in \{1, 2, 6, 7\}$), and MSENonLinear (for the other six nonzero nonlinear effects). These metrics are helpful in assessing the overall estimation performance as well as variations in the performance specific to particular classes of effects.

Key results The first set of interesting results from Simulation 1 are provided by the simulation metrics. These are summarized in Figure 5.1 and Table S5.1-S5.7. Across all signal-to-noise ratios and all sample sizes, GPVIBES is the best in terms of MCC, indicating excellent selection performance at the 10% FDR threshold (Figure 5.1C). In particular, this result is driven by the fact that GPVIBES yields the lowest number of false positives compared to all the other competitors (Figure 5.1B). This indicates that the Bayesian variable selection deconvolved via the γ_js combined with the sparsity control using the horseshoe prior performs satisfactorily in terms of filtering the false signals out. GPVIBES also yields AUCs which are either the best or second best among all the competitors across all the simulation scenarios, as can be seen in Figure 5.1D. This indicates a commendable selection performance across thresholds between 0 and 1, since AUC is a threshold-free summary of selection performance over this range.

The estimation metrics summarized in Figure 5.2 and Table S5.8-S5.12 make the case in favor of GPVIBES even further compelling. As can be observed in Figure 5.2A-E,

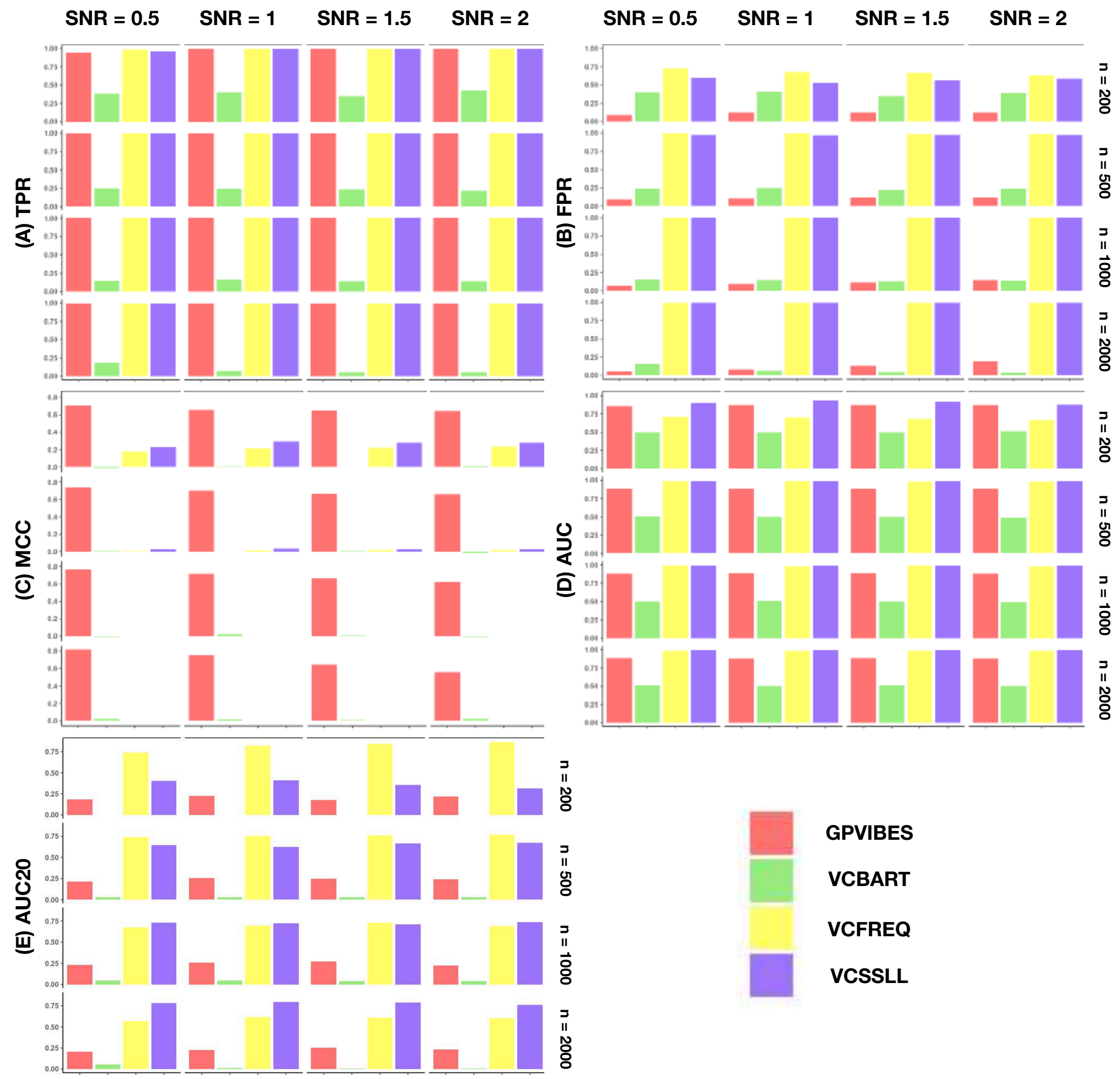


Figure 5.1: Simulation 1 results for the selection metrics. The height of each bar is the average value of the metric across 100 replicates. The signal-to-noise ratio varies across the columns, as labeled at the top. The sample size varies across the rows, as labeled at the right. The abbreviations are as follows: TPR - true positive rate, FPR - false positive rate, MCC - Matthew's correlation coefficient, AUC - area under the receiver operating characteristic curve, AUC20 - scaled AUC in the range of specificity between 0.8 and 1.

GPVIBES produces the least replicate-averaged mean squared error across all simulation scenarios and for all kinds of signals (overall, true, false, linear, and nonlinear). To further illustrate the significance of this performance, the true and estimated varying coefficient functions for one linear and nonlinear signals are visualized in Figure 5.3-5.4. It is easy to observe that for low sample size ($n = 200, 500$) and low signal-to-noise (SNR = 0.5, 1) combinations, GPVIBES (red lines) approximates the true functions better than its competitors, reinforcing the reason behind the excellent performance in terms of the MSE even for these scenarios. For the higher sample size ($n = 1000, 2000$) and low signal-to-noise (SNR = 1.5, 2) combinations, all procedures produce comparable results.

5.3.2 Simulation 2: Multiple Hierarchical Covariates

For Simulation 2, most of the key details remain the same as Simulation 1. Hence, in the rest of this section, the major differences from Simulation 1 are pointed out.

Data generation The choices for the signal-to-noise ratios, and the number of simulation replicates remain the same as those in Simulation 1. The sample size n is chosen to be $100, 400, 900, 1600$. For each n, the two hierarchical covariates $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are taken to be grids of length $\sqrt{n}$, similar to $\mathbf{Z}$ in Simulation 1. All possible pairings across the two grids then result in a total of n observations. A total of $p = 60$ covariates are generated, with the last 54 of them corresponding to signals (varying coefficients) fixed at zero. The first three varying coefficients are each assumed to follow the additive form $\beta_j(z_1, z_2) = c_j(f_{j1}(z_1) + f_{j2}(z_2))$. Again, c_j is chosen such that the integral L^2-norm of the function on the region $[0, 1]^2$ is exactly 1. The functions are chosen to be the following: $f_{11}(z) = f_{32}(z) = z$, $f_{12}(z) = f_{21}(z) = (z - 0.5)^2$, $f_{22}(z) = f_{31}(z) = \sin(9z)$. These functions are chosen so that we can use all possible combinations of linear, polynomial, and trigonometric functions to assess estimation performances. As before, the next three non-zero varying coefficients

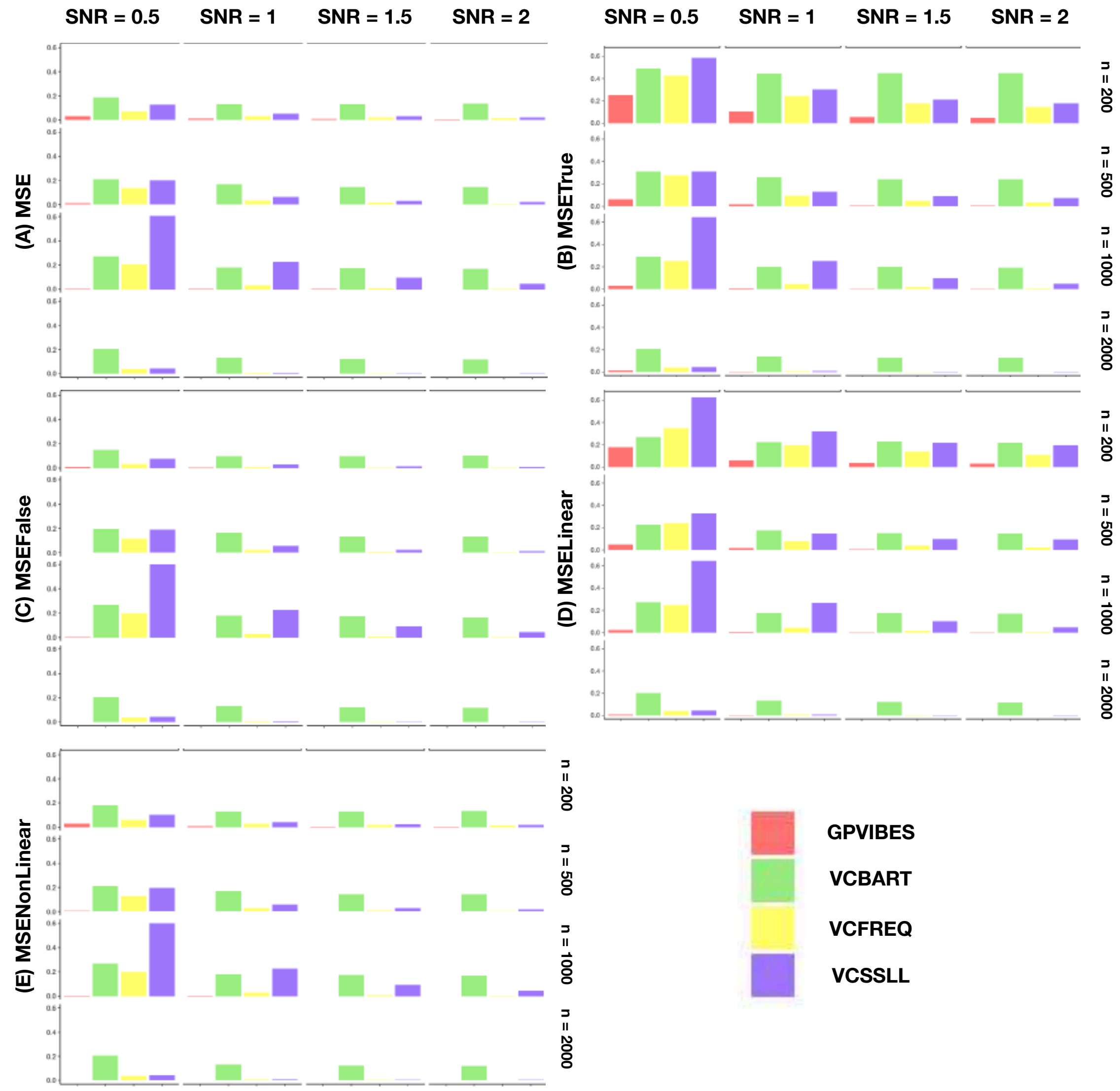


Figure 5.2: Simulation 1 results for the estimation metrics. The height of each bar is the average value of the metric across 100 replicates. The signal-to-noise ratio varies across the columns, as labeled at the top. The sample size varies across the rows, as labeled at the right. MSE means the overall mean squared error across all 100 varying coefficients. For any label W, MSEW means the MSE computed only across the varying coefficients for which the generating mechanism is W. W varies across true (nonzero), false (constant at zero), linear, and nonlinear.

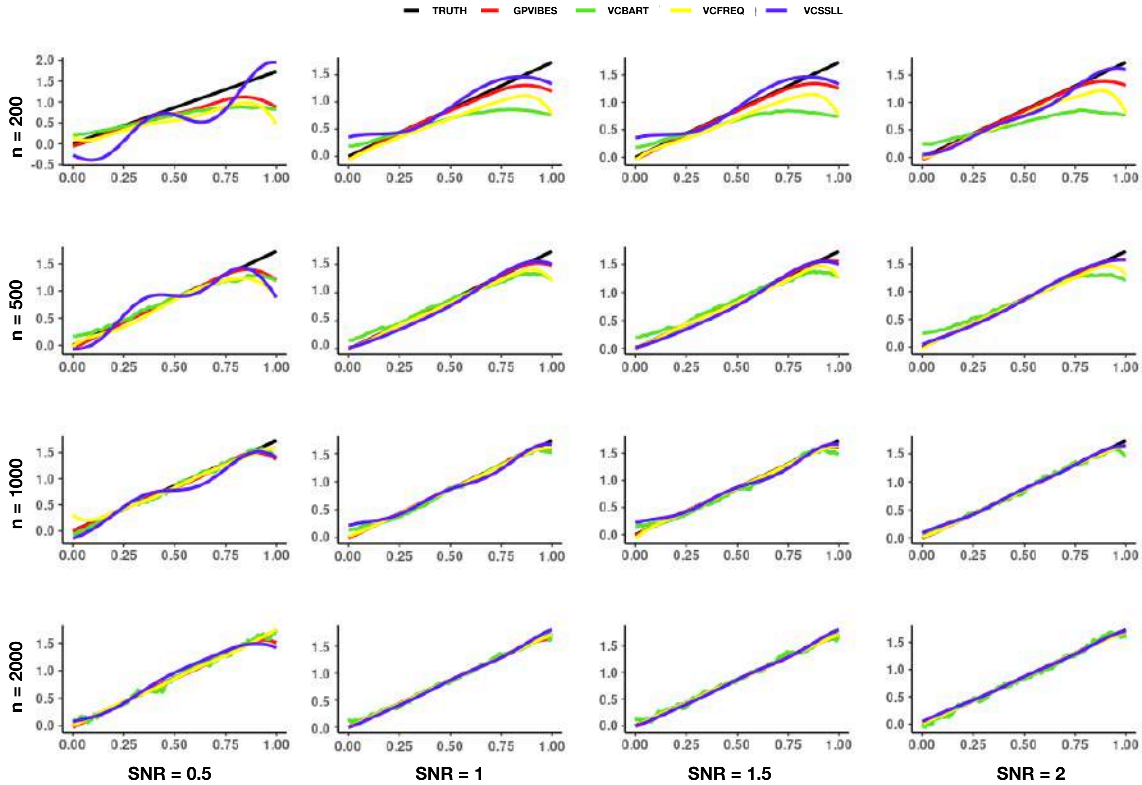


Figure 5.3: Simulation 1 estimation results for a linear function. The signal-to-noise ratio varies across the columns, as labeled at the bottom. The sample size varies across the rows, as labeled at the left. Each curve for the estimation procedures is an average over the 100 simulation replicates. The true function is $\beta(z) = \sqrt{3}z$.

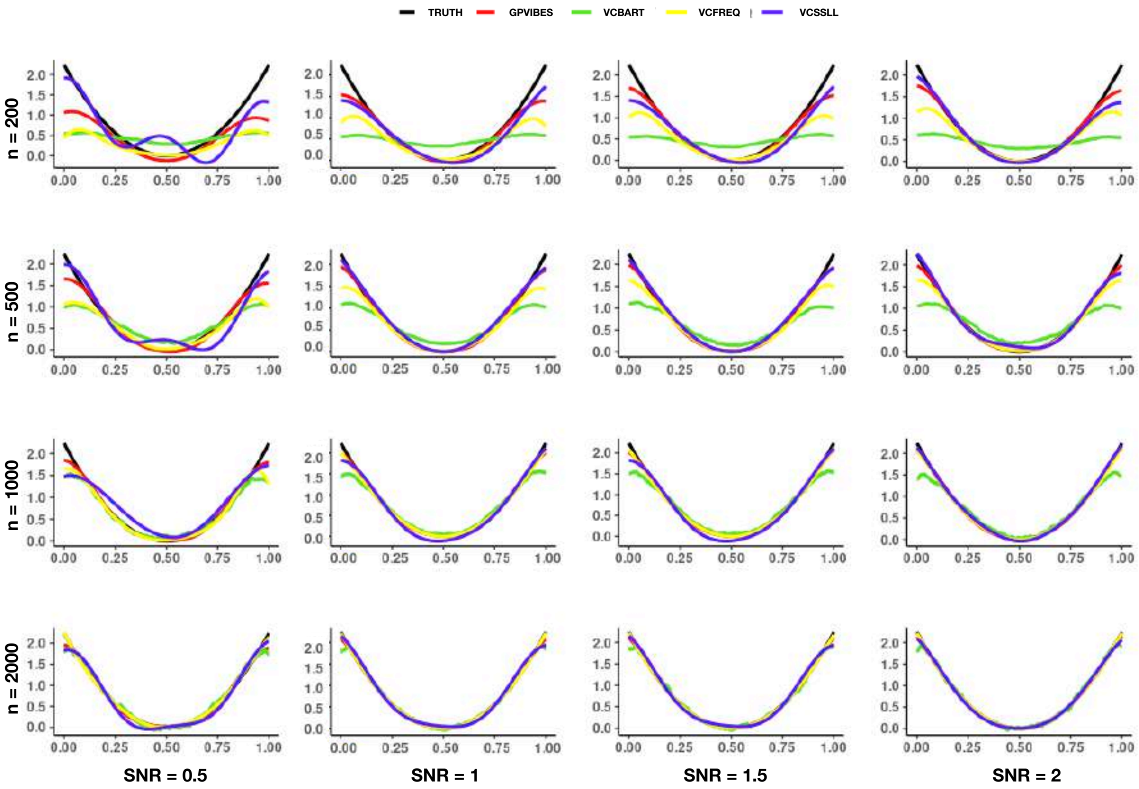


Figure 5.4: Simulation 1 estimation results for a nonlinear function. The signal-to-noise ratio varies across the columns, as labeled at the bottom. The sample size varies across the rows, as labeled at the left. Each curve for the estimation procedures is an average over the 100 simulation replicates. The true function is $\beta(z) = 8.944(z - 0.5)^2$.

are assumed to be the negatives of these first three coefficients. The generating mechanism for the covariates and the outcome remain the same.

Competing methods and evaluation metrics The nonparametric varying coefficient models VCFREQ and VCSSLL as proposed by Bai et al. (2019) and implemented in the R package `NVCSSL` can only accommodate a single hierarchical covariate Z, interpreted as time variable for repeated measurements in their setting. Hence, these procedures are omitted from the comparative evaluation scheme for this particular simulation. The details regarding the implementation of VCBART and GPVIBES remain the same as those in Simulation 1, along with the selection and estimation metrics used there. Note that in this case it is possible to separately estimate the functional components individually from the overall additive varying coefficients.

Key results The simulation and estimation metrics are summarized via bar diagrams in Figure 5.5. As in Simulation 1, GPVIBES yields better metrics than VCBART across all simulation scenarios. In particular, GPVIBES maintains near-perfect true positive rate while still maintaining lower false positive rate than VCBART except for the scenarios where $n = 1600$ and SNR ≥ 1. The estimation performance is also excellent, as is evidenced by the MSEs across different categories of the signals. In all of these categories, GPVIBES yields substantially lower replicate-averaged MSE than VCBART. This indicates that even for models with more than one hierarchical covariate of interest, GPVIBES performs accurate estimation of the true signals via the varying coefficients specified as Gaussian processes.

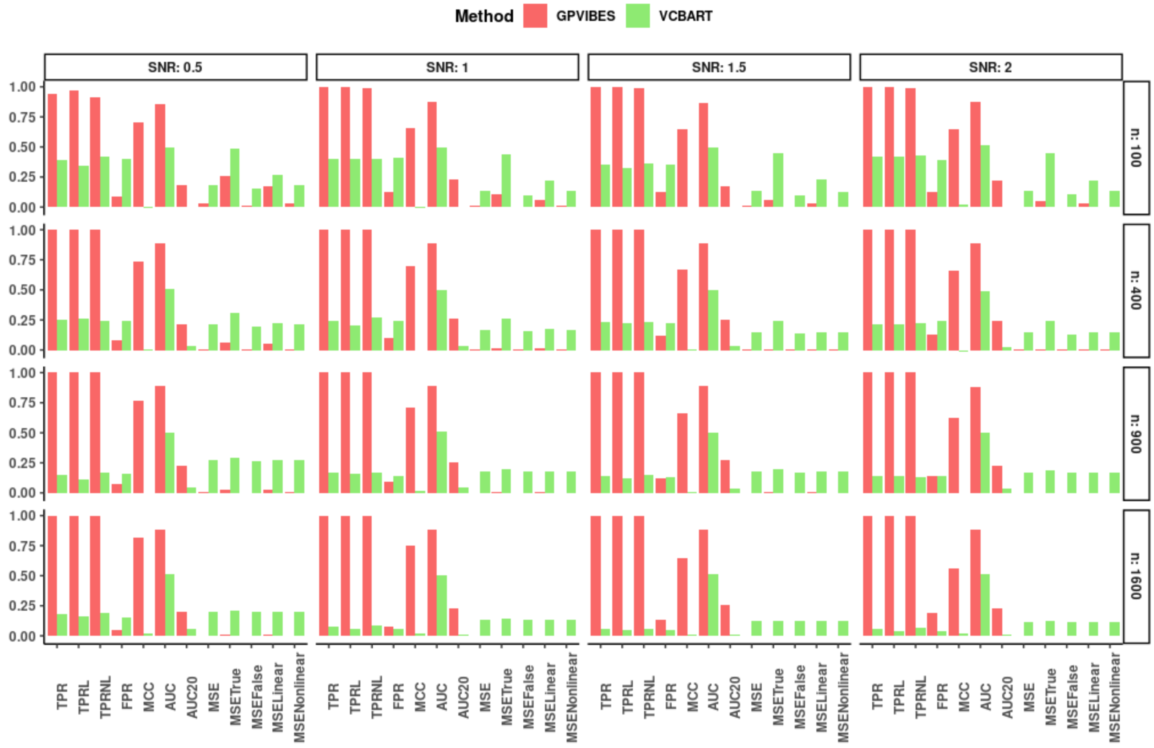


Figure 5.5: Simulation 2 results for the selection and estimation metrics. The height of each bar is the average value of the metric across 100 replicates. The signal-to-noise ratio varies across the columns, as labeled at the top. The sample size varies across the rows, as labeled at the right. The abbreviations are as follows: TPR - true positive rate, FPR - false positive rate, MCC - Matthew's correlation coefficient, AUC - area under the receiver operating characteristic curve, AUC20 - scaled AUC in the range of specificity between 0.8 and 1, MSE - overall mean squared error. For any label W, MSEW means the MSE computed only across the varying coefficients for which the generating mechanism is W. W varies across true (nonzero), false (constant at zero), linear, and nonlinear.

5.4 Integrative Clinicogenomic Pan-cancer Analysis incorporating Immune Signatures

In order to illustrate the utility of GPVIBES in the scientific context, I perform an integrative analysis of pan-cancer data from The Cancer Genome Atlas (TCGA) using proteomic expressions as the covariates of interest and overall survival as the outcome. The role of the hierarchical variables summarizing the heterogeneity in the tumor microenvironment is played by cellular composition-based signatures computed from the genomic profile of the tumor samples. First, I describe the source of and some key overview regarding the data.

5.4.1 Data Description

Data on all 32 of the TCGA cancers are obtained from the Xena browser (Goldman et al., 2020). The overall survival times along with the censoring information are obtained from the combined phenotype dataset, the proteomic expressions are obtained from the data covering the reverse-phase protein array platform, and the immune signatures constructed by Thorsson et al. (2018) are obtained from the cellular signature data release. I begin with annotating all the three platforms using the TCGA patient tumor barcodes, and retain only the samples available across all platforms. Only the 16 cancers that retain at least 200 samples after this step are utilized for all the analyses hereafter. The smallest sample sizes are available for KIRP and OV (201 each), while the highest sample size of 860 is available for BRCA. A summary of the sample sizes is presented in Figure S5.1. The proteomic data contains a total of 210 proteins, and a total of 68 immune signatures (including key cellular components such as B cell, T cell, CD8-T cells, Chemokines, and Interleukins) are available in the data.

5.4.2 Implementation of GPVIBES and Scientific Questions

For each cancer-signature combination, a single GPVIBES model is built, using the single signature as the hierarchical covariate Z and all the 210 proteins as the X covariates. The details regarding the Bayesian implementation (i.e., number of chains, number of burn-ins, post burn-in iterations, and thinning proportion) remain the same as those used in Simulations 1 and 2. As before, GPVIBES yields, for each of the $16 \times 68 = 1,088$ models, two sets of estimates for each protein j: the posterior probability of inclusion $\hat{\gamma}_j$, and the estimated varying coefficients $\hat{\beta}_j(Z_i)$ at the observed signature values. Based on these outputs, the interest lies in answering two primary scientific questions. First, at a pan-signature level, it is interesting to assess whether there are cancers that exhibit higher modulation of the proteomic associations with survival via the tumor immunogenic heterogeneity than the rest of the cancers. For these cancers, it is also interesting to note whether there are specific proteins that are selected in the GPVIBES models more often, indicating strong evidence for pan-cancer and/or cancer-specific associations with survival. Second, for a particular signature, there may be key proteins that are modulated via this signature for a specific set of cancers. The exact form of these modulations can be quantified using the estimated varying coefficients.

5.4.3 Key Results and Biological Interpretations

Cross-cancer variability of immunogenic modulation Figure 5.6A summarizes the total proportion of varying coefficients found to be significant at a 10% FDR control cut-off across all signature-specific models for each cancer. While the sample size availability for each cancer (hence the statistical power offered by the data) directly contributes to this quantitative summary, there are additional evidences of immunogenic modulation of proteomic associations with survival. For example, KIRC, SARC, THCA, and SKCM - four cancers which all have lower number of samples than BRCA in our dataset, are found to capture more significant modulations than BRCA at the same level of type I error control. The majority of these cancers (or certain subgroups of them) are known to be immunologically hot, meaning that they trigger strong immune responses (KIRC: Bi et al. (2021), THCA: Du et al. (2021), SKCM: Li et al. (2023)). While SARC is known to be immunologically cold, recent efforts have explored the use of immunotherapeutic agents in the management of sarcomas (Rytlewski et al., 2021). To further investigate these modulations, the proportion of signatures for which a protein was deemed to be significant is summarized at a cancer-specific level in Figure 5.6B. Here, only the proteins that were significant for at least half of the signatures in at least six out of the 16 cancers are presented. Several known associations are identified here, such as the protein EGFR which is found to be significant in 100% of the signature-specific models for both of the gynecological cancers BRCA and OV. Aberrant EGFR expression is known to be correlated with disease progression, resistance to radiation and chemotherapy, and poor clinical prognosis in BRCA (Kumaraswamy et al., 2015). It is also known as a prognostic biomarker and therapeutic target in ovarian cancer (Mehner et al., 2017), thus aligning the findings from GPVIBE regarding it with existing biological knowledge.

Differential and conserved immunogenic modulation of proteomic associations across cancers Figure 5.7 summarizes the cancer-specific posterior inclusion probabilities of the top proteins for three tumor microenvironment signatures: (A) B cell, (B) Interleukin-12, and (C) PDL1. Certain cancers and proteins exhibit strong levels of evidence in favor of immunogenic modulation - for example, the cancer KIRC and the protein PEA15 are on top across all three signatures. PEA15 is known to be prognostic in multiple stages of renal cell carcinomas (Han et al., 2017). Another conserved association is observed for the cancer BRCA

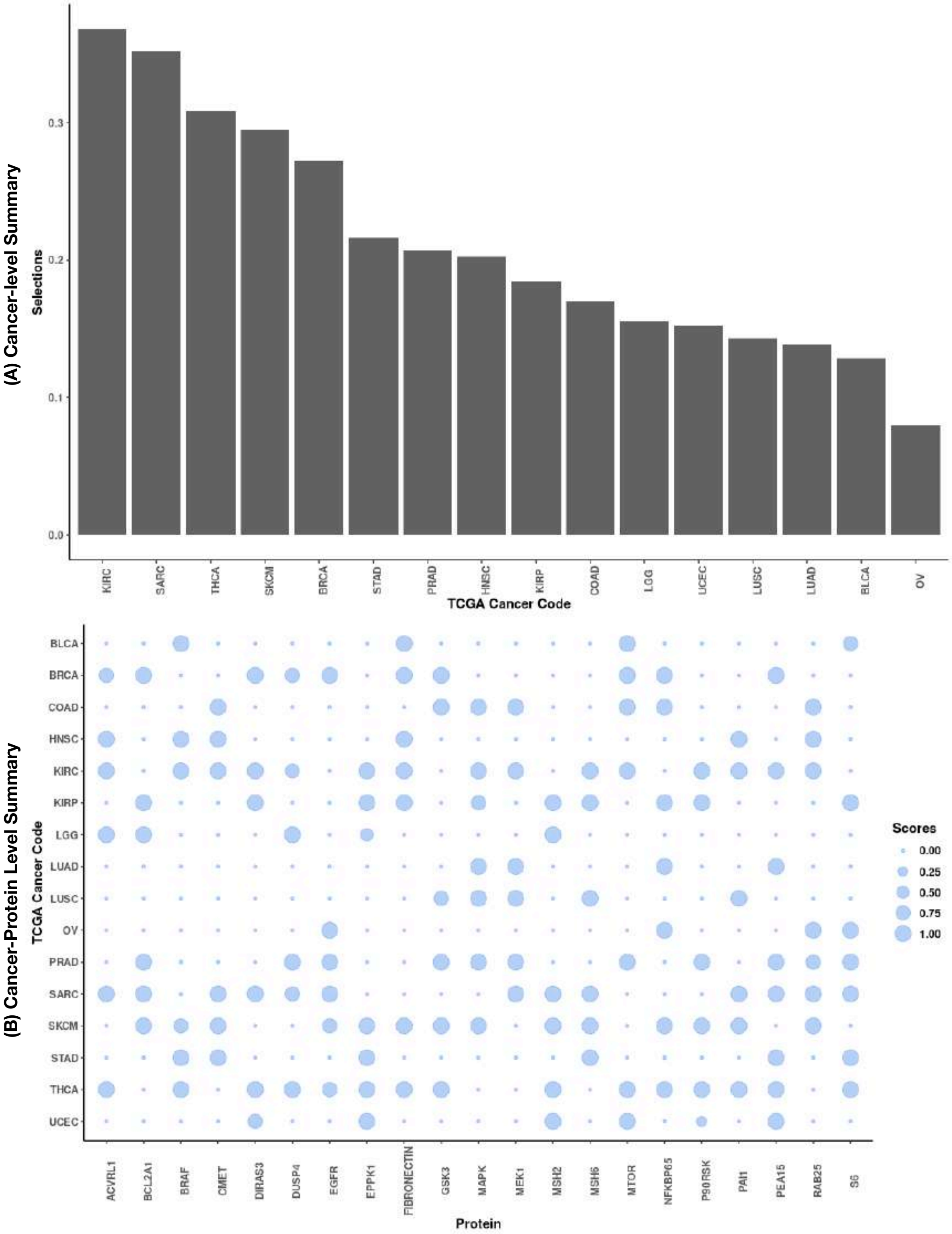


Figure 5.6: Pan-cancer summary of variable selection in the integrative analysis. In panel (A), the height of each bar represents the proportion of varying coefficients selected across all 68 models and 210 proteins for the corresponding cancer. In panel (B), the Y axis presents only the proteins which are significant in more than 50% of the immune signature-specific models across at least six of the 16 cancers. For each cancer-protein combination, the size of the bubble is proportional to the proportion of immune signature-specific models in which the varying coefficient corresponding to the protein is statistically significant at a 10% level of FDR control.

and the protein fibronectin. Previously, a dynamic relationship between tumor and stromal cells within the tumor microenvironment has been established in breast cancer, in which the levels and fibrillarization of fibronectin in the extracellular matrix are modulated during different stages of disease progression (Libring et al., 2020). On the other hand, a key example of differential signal is presented by the association of ribosomal protein S6 expression with overall survival in BLCA, which is statistically significant in the B cell and PDL1 models, but not in the Interleukin model. The therapeutic potential of S6 has been explored in recent oncological research (Yi et al., 2021). However, Interleukin-12 is already well established as an option for intravesical immunotherapy in bladder cancers (Nguyen et al., 2021). The fact that GPVIBES identifies the S6 association for other signatures but not for Interleukin-12 may suggest a mechanistic modulation of the S6 expression in presence of high Interleukin-12 activity in the tumor microenvironment, potentially guiding future investigations.

Contradictory associations of PEA15 with overall survival in KIRC To illustrate the utility of the flexible Gaussian process-based modeling of the varying coefficients via GPVIBES, the association of PEA15 with overall survival in KIRC is further discussed as a case study here. The estimated varying coefficients for the same signatures considered in Figure 5.7 are visualized in Figure 5.8. While all three estimates lie on both sides of the constant zero signal depending on the value of the signature score, the significant regions based on the 95% posterior credible intervals differ. The significant portion of the varying coefficient is positive for the B cell model, while the same is negative for the other two models. As discussed in the previous paragraph, PEA15 is a known prognostic agent in renal carcinomas. However, several studies have indicated that changes in the tumor microenvironment such as phosphorylation of agents in the ERK signaling cascade can modify PEA15 from

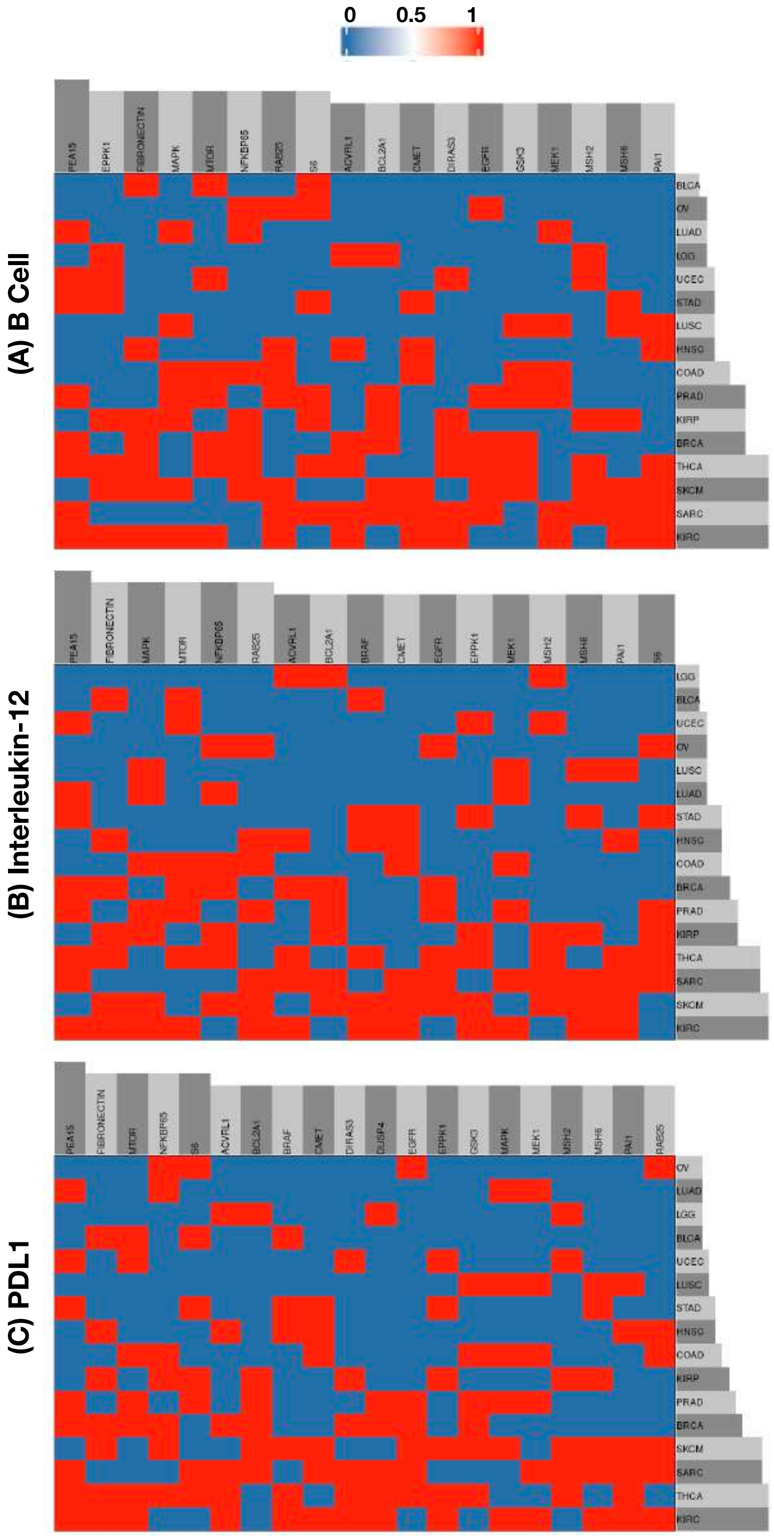
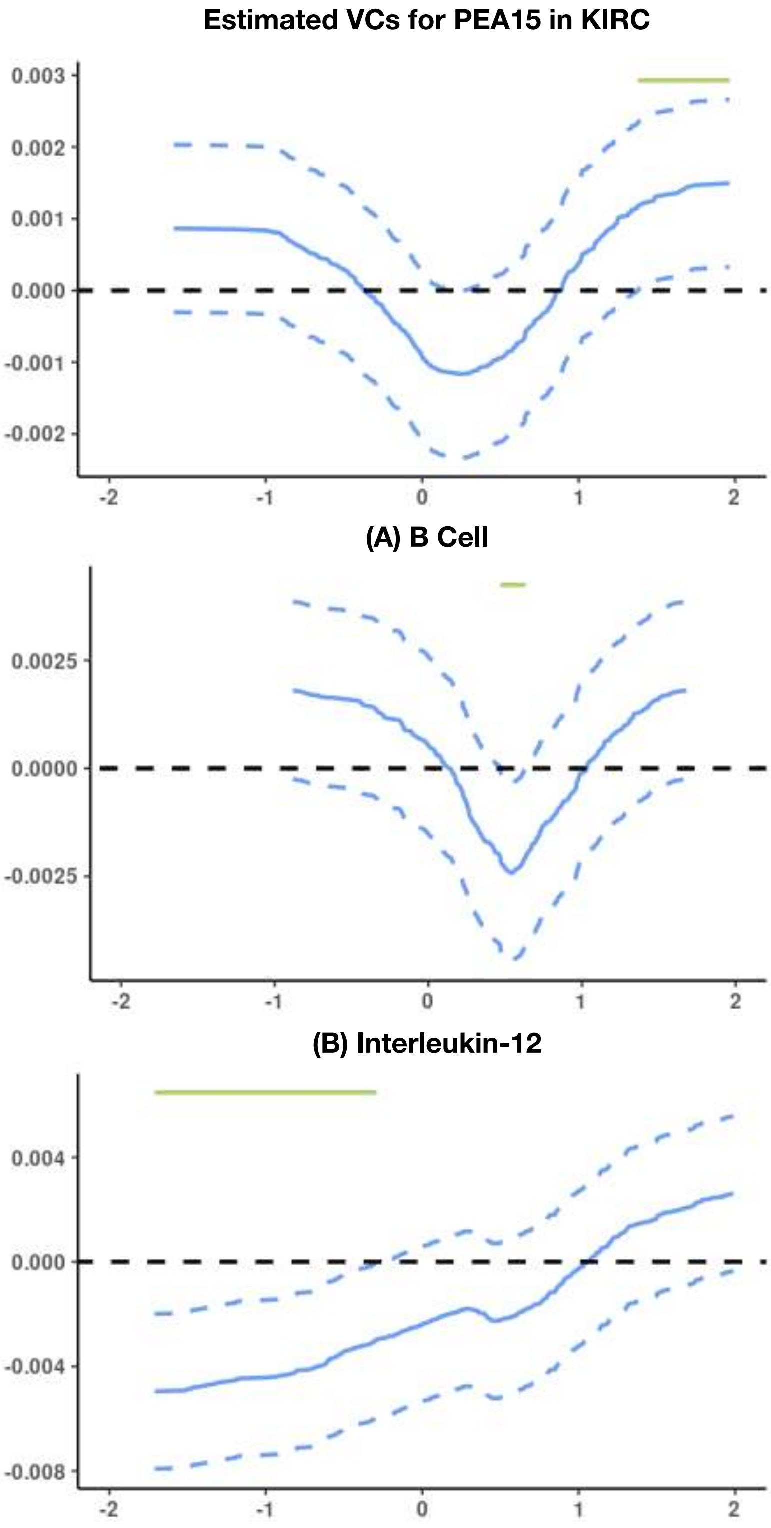


Figure 5.7: Signature-specific pan-cancer summary of selection. In each panel, for a specific immune signature, a heatmap is presented across all cancers in the rows and the top proteins in the columns. The color coding in the heatmap cells corresponds to the posterior inclusion probability $\hat{\gamma}$ for the particular gene in the column in the signature-specific model for the cancer in the row. The rows and columns are ordered according to the average posterior inclusion probability across each dimension.

a tumor suppressor to a tumor promoter. In the GPVIBES estimates, PEA15 expression is positively associated with overall survival in KIRC for high positive B cell score, but is negatively associated with the same for small positive Interleukin-12 score and negative PDL1 score. These results illustrate the potential that GPVIBES offers to identify tumor microenvironment elements that play crucial roles in modulating roles of genomics in patient survival.



(A) B Cell

(B) Interleukin-12

(C) PDL1

Figure 5.8: Estimated varying coefficients from signature-specific models for the protein pea15 and cancer KIRC. In each panel, the dark dashed horizontal line is at zero, the blue solid line indicates the estimated varying coefficient, and the blue dashed lines indicate the upper and lower 95% pointwise posterior credible intervals for the same. The green solid line on top of each panel indicates the regions where the intervals do not contain the zero lines.

5.5 Discussion and Future Work

Overview In this chapter, I propose GPVIBES, a Gaussian process-based varying coefficient model framework using Bayesian variable selection. GPVIBES offers flexible modeling options for varying coefficients as functions of hierarchical covariates via the Gaussian process specification while enforcing sparsity in the number of selected associations via an amalgamation of Bayesian variable selection and shrinkage techniques. Simulation studies using synthetic datasets with one or more hierarchical covariates across a broad spectrum of sample size to number of covariates ratios and signal to noise ratios exhibit substantial improvements in the part of GPVIBES in terms of both selection and estimation compared to other varying coefficient-based parametric and nonparametric models. To illustrate the utility of GPVIBES in integrative clinicogenomic studies in context of precision oncology, I perform a pan-cancer analysis of overall survival and proteomics data from TCGA using immune signatures as the hierarchical covariates. Our analysis uncovers several interesting and interpretable results, such as the differential behavior of PEA15 expression in context of survival of KIRC patients across different signatures. Additional to the flexible yet sparse modeling schema that GPVIBES offers, a key improvement is due to the computational advantages that GPVIBES offers against the other varying coefficient models, as is discussed next.

Sensitivity to kernel hyperparameters A key point in the process of building a GPVIBES model is the choice of hyperparameters driving the MSE kernel-based expansion technique. As described in Section 5.2.5, I use fixed values of the kernel parameters a and b across all simulation and real data scenarios, along with a fixed degree M for the maximal polynomial included in the expansion. Initial investigations into the approximation of known nonlinear functions indicate that a small value for a in the order of 10^{-2} or 10^{-3} is sufficient, since the role of a is only to ensure that the expansion is possible, which is the case as long as $a > 0$. Similarly, for b, a value in the order of 10^k for some $k \geq 2$ is usually sufficient, It is possible to tune these parameters to a specific hierarchical covariate. One such option is provided by the procedure followed in Sun et al. (2020). For a given Z, they compute quantile-based summaries of $(Z_i - Z_j)^2$, and choose multiple values of the lengthscale parameter $\lambda = 1/\sqrt{2b}$ across a grid. Then, an optimal value is chosen based on cross-validating model performance. A similar approach can be followed for choosing a and b in the setting described in this chapter. For the class of nonlinear associations modeled in this chapter, approximations with polynomials of a maximum degree $M = 30$ suffice. Initial simulations indicate that depending on the data input, even approximations of maximum degree $M = 10$ or 15 may be sufficient. Since majority of our models use a single hierarchical covariate, the approximation even with $M = 30$ is superfast and does not pose any substantial computational burden, which is the reason why it is used as a fixed, singular choice.

Computational efficiency of GPVIBES An attractive feature of GPVIBES is its computational efficiency. Due to the fast data augmentation-based update of the regression coefficients in the horseshoe setting, as well as the deconvolution of the selection and estimation mechanics in the Gaussian process specification, GPVIBES can handle large datasets at ease. This is illustrated in Figure S5.2, which summarizes the average time in seconds per unit operation in Simulation 1 for each method. As can be observed, despite being a fully Bayesian procedure, GPVIBES is the fastest among the methods compared. To illustrate the utility of GPVIBES in context of integrative clinicogenomic studies, the runtimes of GPVIBES across the signature-specific models are summarized via individual boxplots for each cancer in Figure S5.3. As can be observed, all runtimes for all the cancers are under one hour, with the third quartile of runtimes falling under 10 minutes for all cancers except KIRC, which is the cancer with the second largest sample size. This makes GPVIBES an extremely realistic option for clinicogenomic studies with large omics panels and several hierarchical signatures of interest.

Future directions From a scientific perspective, a key direction of interest to pursue as a follow-up to the research performed in this chapter is to incorporate higher quality genomic data in the integrative analysis. The proteomic data used in this case comes from a reverse-phase protein array, which only includes a targeted set of proteins which are of oncological interest. However, deeper proteomics panels covering additional dimensions of cellular oncological activity such as phosphorylation are becoming increasingly available. An example of this, as mentioned in previous chapters, is given by the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (Ellis et al., 2013). Emerging data dissemination efforts such as LinkedOmics have made such data available for multiple cancers, aligned with other omics platforms including mRNA expression and copy number variation, along with sample-specific phenotypes and clinical variables (Vasaikar et al., 2018). Thus, there is promising potential towards integrative GPVIBES analysis with more enriched datasets. From the methodological angle, two particular developments are of interest. First, in this chapter, the GPVIBES model with more than one hierarchical variables was only explored in Simulation 2. It would be interesting to assess the computational evolution of such a model with increasing number of hierarchical variables, and any potential challenge for substantially higher dimensions. Additionally, in the scientific context, employing such a model may be of interest since many tumor microenvironment components can potentially interact among themselves. Including such signatures in a single model may be beneficial to uncover underlying tumor growth mechanisms. The other methodological pursuit of interest is to combine the idea of biomarker prioritization via calibrated priors as discussed in Chapter III and Chapter IV. Such a component can be implemented via changing the prior structure on the γ_js following Section 5.2.4. It would be interesting to investigate whether such a combined modeling scheme is feasible computationally, and if yes, whether it results in more interpretable estimates and improved selection.

Reproducibility All the analysis codes developed for the purpose of this chapter, along with the processed datasets are available on the R shiny dashboard hosted at `https://bayesrx.shinyapps.io/GPVIBES/`. The dashboard also serves as an interactive domain for the visualization and presentation of all the results from our pan-cancer integrative clinicogenomic analyses.

5.6 Supplementary Figures

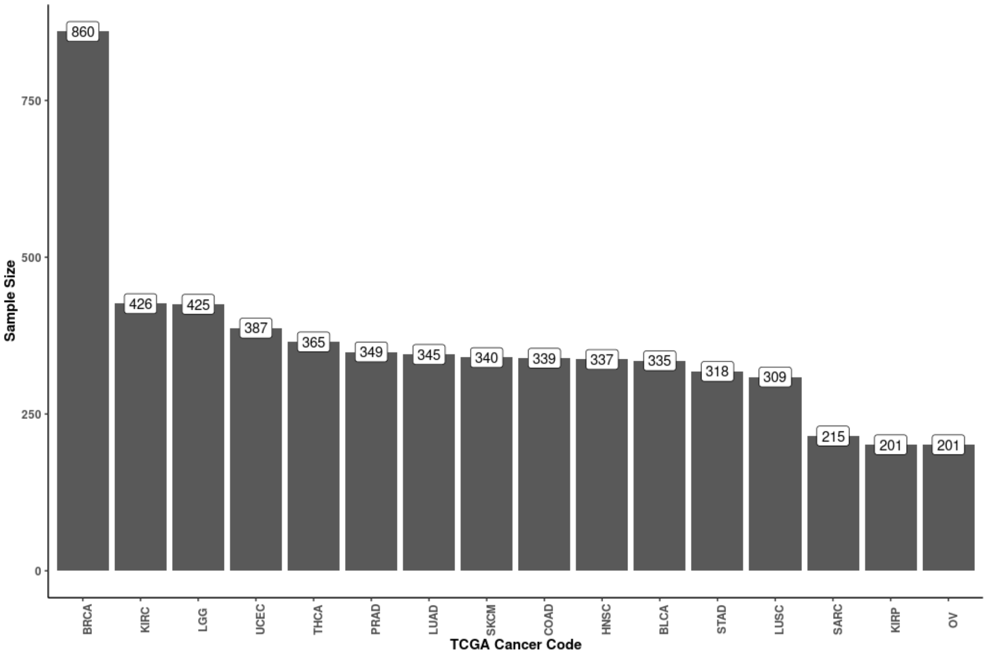


Figure S5.1: Sample sizes of TCGA cancers used in the integrative clinicogenomic analysis. In each case, the sample size indicates the total number of samples for which data from all three sources - overall survival, immune signatures, and proteomic expressions - are available.

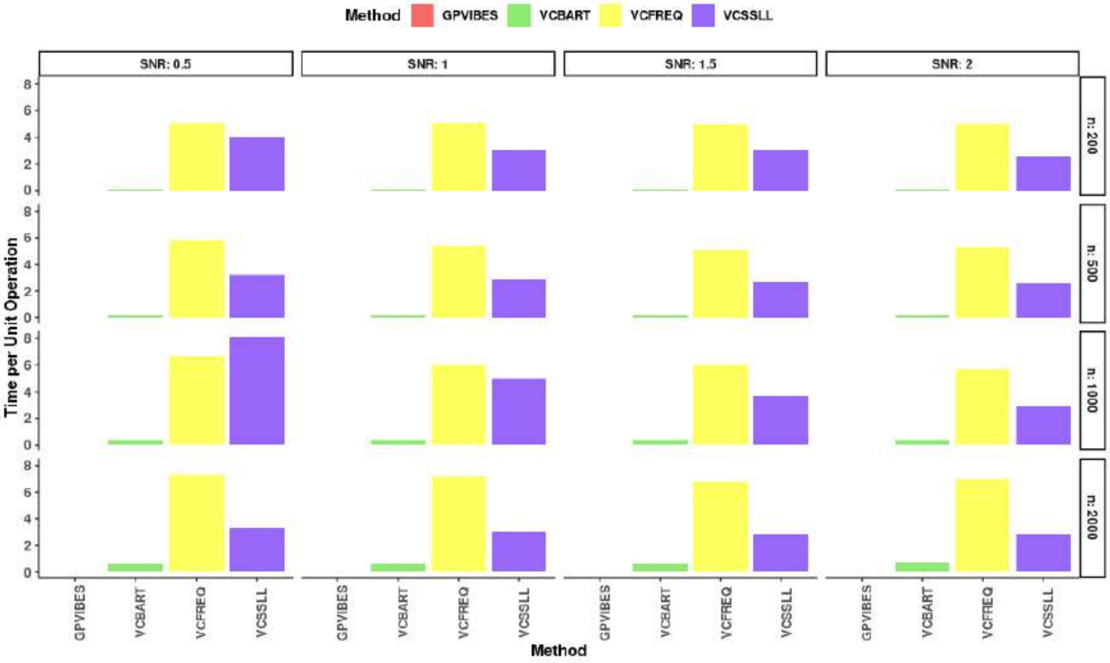


Figure S5.2: Average runtimes per unit operation for Simulation 1. The height of each bar is the average runtime per operation across 100 replicates. For GPVIBES and VCBART, unit operation means a single update in the MCMC procedure. For VCFREQ, unit operation means fitting the model for a single regularization parameter. For VCSSLL, unit operation means one iteration of the posterior approximation procedure for a single regularization parameter. The signal-to-noise ratio varies across the columns, as labeled at the top. The sample size varies across the rows, as labeled at the right. MSE means the overall mean squared error across all 100 varying coefficients.

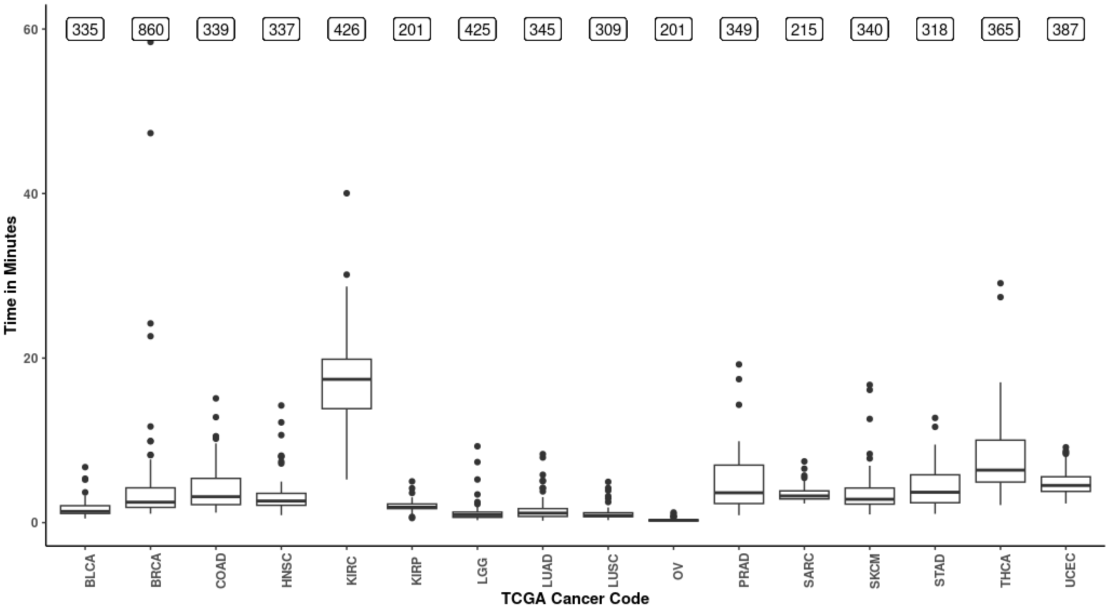


Figure S5.3: Runtimes across signature-specific models for TCGA cancers. The boxplots are based on a total of 68 values, one for each signature-specific model for a given cancer. The labels on top indicate the sample sizes for the TCGA cancers.

5.7 Supplementary Tables

Table S5.1: True positive rate results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.937 (0.13)	0.390 (0.20)	0.992 (0.03)	0.959 (0.08)
	1	0.995 (0.02)	0.402 (0.16)	1.000 (0.00)	0.998 (0.01)
	1.5	0.996 (0.02)	0.350 (0.21)	1.000 (0.00)	1.000 (0.00)
	2	0.994 (0.02)	0.424 (0.17)	1.000 (0.00)	1.000 (0.00)
500	0.5	1.000 (0.00)	0.248 (0.14)	1.000 (0.00)	1.000 (0.00)
	1	1.000 (0.00)	0.242 (0.15)	1.000 (0.00)	1.000 (0.00)
	1.5	1.000 (0.00)	0.228 (0.13)	1.000 (0.00)	1.000 (0.00)
	2	1.000 (0.00)	0.218 (0.14)	1.000 (0.00)	1.000 (0.00)
1000	0.5	1.000 (0.00)	0.147 (0.11)	1.000 (0.00)	1.000 (0.00)
	1	1.000 (0.00)	0.164 (0.13)	1.000 (0.00)	1.000 (0.00)
	1.5	1.000 (0.00)	0.140 (0.10)	1.000 (0.00)	1.000 (0.00)
	2	1.000 (0.00)	0.136 (0.10)	1.000 (0.00)	1.000 (0.00)
2000	0.5	1.000 (0.00)	0.180 (0.12)	1.000 (0.00)	1.000 (0.00)
	1	1.000 (0.00)	0.074 (0.09)	1.000 (0.00)	1.000 (0.00)
	1.5	1.000 (0.00)	0.055 (0.07)	1.000 (0.00)	1.000 (0.00)
	2	1.000 (0.00)	0.054 (0.07)	1.000 (0.00)	1.000 (0.00)

Table S5.2: True positive rate for the linear functions results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.968 (0.13)	0.345 (0.25)	0.998 (0.03)	0.983 (0.07)
	1	1.000 (0.00)	0.405 (0.22)	1.000 (0.00)	1.000 (0.00)
	1.5	1.000 (0.00)	0.328 (0.27)	1.000 (0.00)	1.000 (0.00)
	2	0.998 (0.03)	0.420 (0.27)	1.000 (0.00)	1.000 (0.00)
500	0.5	1.000 (0.00)	0.260 (0.21)	1.000 (0.00)	1.000 (0.00)
	1	1.000 (0.00)	0.200 (0.19)	1.000 (0.00)	1.000 (0.00)
	1.5	1.000 (0.00)	0.222 (0.21)	1.000 (0.00)	1.000 (0.00)
	2	1.000 (0.00)	0.212 (0.21)	1.000 (0.00)	1.000 (0.00)
1000	0.5	1.000 (0.00)	0.112 (0.16)	1.000 (0.00)	1.000 (0.00)
	1	1.000 (0.00)	0.160 (0.21)	1.000 (0.00)	1.000 (0.00)
	1.5	1.000 (0.00)	0.120 (0.15)	1.000 (0.00)	1.000 (0.00)
	2	1.000 (0.00)	0.138 (0.16)	1.000 (0.00)	1.000 (0.00)
2000	0.5	1.000 (0.00)	0.162 (0.20)	1.000 (0.00)	1.000 (0.00)
	1	1.000 (0.00)	0.060 (0.12)	1.000 (0.00)	1.000 (0.00)
	1.5	1.000 (0.00)	0.050 (0.11)	1.000 (0.00)	1.000 (0.00)
	2	1.000 (0.00)	0.040 (0.09)	1.000 (0.00)	1.000 (0.00)

Table S5.3: True positive rate for the nonlinear functions results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.917 (0.15)	0.420 (0.24)	0.988 (0.04)	0.943 (0.10)
	1	0.992 (0.04)	0.400 (0.20)	1.000 (0.00)	0.997 (0.02)
	1.5	0.993 (0.03)	0.365 (0.26)	1.000 (0.00)	1.000 (0.00)
	2	0.992 (0.04)	0.427 (0.20)	1.000 (0.00)	1.000 (0.00)
500	0.5	1.000 (0.00)	0.240 (0.20)	1.000 (0.00)	1.000 (0.00)
	1	1.000 (0.00)	0.270 (0.20)	1.000 (0.00)	1.000 (0.00)
	1.5	1.000 (0.00)	0.232 (0.16)	1.000 (0.00)	1.000 (0.00)
	2	1.000 (0.00)	0.222 (0.17)	1.000 (0.00)	1.000 (0.00)
1000	0.5	1.000 (0.00)	0.170 (0.16)	1.000 (0.00)	1.000 (0.00)
	1	1.000 (0.00)	0.167 (0.16)	1.000 (0.00)	1.000 (0.00)
	1.5	1.000 (0.00)	0.153 (0.15)	1.000 (0.00)	1.000 (0.00)
	2	1.000 (0.00)	0.135 (0.13)	1.000 (0.00)	1.000 (0.00)
2000	0.5	1.000 (0.00)	0.192 (0.17)	1.000 (0.00)	1.000 (0.00)
	1	1.000 (0.00)	0.083 (0.12)	1.000 (0.00)	1.000 (0.00)
	1.5	1.000 (0.00)	0.058 (0.10)	1.000 (0.00)	1.000 (0.00)
	2	1.000 (0.00)	0.063 (0.10)	1.000 (0.00)	1.000 (0.00)

Table S5.4: False positive rate results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.085 (0.06)	0.402 (0.10)	0.730 (0.05)	0.600 (0.24)
	1	0.122 (0.06)	0.409 (0.07)	0.683 (0.05)	0.527 (0.20)
	1.5	0.128 (0.06)	0.350 (0.15)	0.663 (0.05)	0.558 (0.21)
	2	0.126 (0.05)	0.396 (0.09)	0.636 (0.07)	0.587 (0.27)
500	0.5	0.085 (0.04)	0.239 (0.06)	0.998 (0.00)	0.972 (0.05)
	1	0.104 (0.06)	0.244 (0.06)	0.997 (0.01)	0.965 (0.06)
	1.5	0.118 (0.05)	0.227 (0.05)	0.994 (0.01)	0.972 (0.05)
	2	0.123 (0.05)	0.243 (0.07)	0.993 (0.01)	0.978 (0.04)
1000	0.5	0.071 (0.04)	0.156 (0.05)	1.000 (0.00)	1.000 (0.00)
	1	0.095 (0.05)	0.144 (0.05)	1.000 (0.00)	1.000 (0.00)
	1.5	0.119 (0.05)	0.134 (0.05)	1.000 (0.00)	1.000 (0.00)
	2	0.143 (0.06)	0.143 (0.05)	1.000 (0.00)	1.000 (0.00)
2000	0.5	0.052 (0.04)	0.156 (0.05)	1.000 (0.00)	1.000 (0.00)
	1	0.077 (0.04)	0.060 (0.04)	1.000 (0.00)	1.000 (0.00)
	1.5	0.131 (0.05)	0.049 (0.03)	1.000 (0.00)	1.000 (0.00)
	2	0.192 (0.07)	0.041 (0.03)	1.000 (0.00)	1.000 (0.00)

Table S5.5: Matthews correlation coefficient results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.706 (0.12)	-0.008 (0.12)	0.183 (0.03)	0.227 (0.12)
	1	0.659 (0.10)	-0.004 (0.09)	0.211 (0.03)	0.293 (0.12)
	1.5	0.651 (0.11)	-0.001 (0.10)	0.220 (0.02)	0.278 (0.12)
	2	0.647 (0.08)	0.017 (0.10)	0.234 (0.03)	0.280 (0.21)
500	0.5	0.733 (0.10)	0.007 (0.10)	0.006 (0.01)	0.028 (0.05)
	1	0.698 (0.11)	-0.001 (0.11)	0.009 (0.02)	0.038 (0.05)
	1.5	0.668 (0.10)	0.001 (0.10)	0.016 (0.02)	0.033 (0.05)
	2	0.658 (0.09)	-0.017 (0.10)	0.018 (0.02)	0.027 (0.04)
1000	0.5	0.769 (0.10)	-0.007 (0.10)	0.000 (0.00)	0.000 (0.00)
	1	0.713 (0.10)	0.017 (0.12)	0.000 (0.00)	0.000 (0.00)
	1.5	0.663 (0.09)	0.005 (0.09)	0.000 (0.00)	0.000 (0.00)
	2	0.627 (0.10)	-0.005 (0.09)	0.000 (0.00)	0.000 (0.00)
2000	0.5	0.821 (0.11)	0.020 (0.10)	0.000 (0.00)	0.000 (0.00)
	1	0.750 (0.09)	0.018 (0.11)	0.000 (0.00)	0.000 (0.00)
	1.5	0.643 (0.08)	0.007 (0.10)	0.000 (0.00)	0.000 (0.00)
	2	0.557 (0.09)	0.024 (0.11)	0.000 (0.00)	0.000 (0.00)

Table S5.6: AUC results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.855 (0.05)	0.496 (0.10)	0.714 (0.05)	0.899 (0.07)
	1	0.872 (0.04)	0.497 (0.08)	0.702 (0.04)	0.929 (0.08)
	1.5	0.870 (0.04)	0.496 (0.09)	0.683 (0.04)	0.921 (0.08)
	2	0.873 (0.04)	0.513 (0.08)	0.663 (0.04)	0.880 (0.08)
500	0.5	0.886 (0.03)	0.507 (0.07)	0.990 (0.01)	0.994 (0.00)
	1	0.886 (0.04)	0.497 (0.08)	0.990 (0.01)	0.990 (0.01)
	1.5	0.889 (0.03)	0.502 (0.07)	0.986 (0.01)	0.991 (0.01)
	2	0.887 (0.04)	0.487 (0.08)	0.987 (0.01)	0.992 (0.01)
1000	0.5	0.885 (0.04)	0.499 (0.06)	0.991 (0.01)	0.994 (0.00)
	1	0.887 (0.03)	0.511 (0.08)	0.989 (0.01)	0.991 (0.01)
	1.5	0.888 (0.04)	0.504 (0.06)	0.991 (0.01)	0.991 (0.01)
	2	0.881 (0.04)	0.497 (0.06)	0.989 (0.01)	0.991 (0.01)
2000	0.5	0.887 (0.03)	0.516 (0.06)	0.988 (0.01)	1.000 (0.00)
	1	0.881 (0.04)	0.503 (0.06)	0.990 (0.01)	1.000 (0.00)
	1.5	0.885 (0.03)	0.517 (0.05)	0.987 (0.01)	1.000 (0.00)
	2	0.881 (0.03)	0.509 (0.05)	0.986 (0.01)	1.000 (0.00)

Table S5.7: AUC20 results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.186 (0.20)	0.001 (0.01)	0.739 (0.14)	0.404 (0.29)
	1	0.229 (0.22)	0.000 (0.00)	0.821 (0.10)	0.411 (0.34)
	1.5	0.176 (0.19)	0.000 (0.00)	0.844 (0.09)	0.357 (0.34)
	2	0.220 (0.21)	0.001 (0.01)	0.867 (0.08)	0.313 (0.37)
500	0.5	0.217 (0.21)	0.029 (0.06)	0.745 (0.15)	0.648 (0.21)
	1	0.259 (0.22)	0.029 (0.05)	0.754 (0.14)	0.624 (0.18)
	1.5	0.252 (0.21)	0.031 (0.06)	0.760 (0.14)	0.669 (0.18)
	2	0.242 (0.23)	0.026 (0.05)	0.769 (0.14)	0.675 (0.17)
1000	0.5	0.228 (0.21)	0.045 (0.05)	0.678 (0.17)	0.734 (0.16)
	1	0.255 (0.20)	0.046 (0.04)	0.696 (0.19)	0.724 (0.16)
	1.5	0.272 (0.21)	0.039 (0.04)	0.733 (0.19)	0.713 (0.16)
	2	0.221 (0.21)	0.038 (0.04)	0.687 (0.18)	0.739 (0.15)
2000	0.5	0.204 (0.22)	0.054 (0.05)	0.566 (0.20)	0.785 (0.15)
	1	0.226 (0.20)	0.013 (0.02)	0.616 (0.19)	0.794 (0.13)
	1.5	0.252 (0.22)	0.008 (0.01)	0.611 (0.19)	0.792 (0.14)
	2	0.233 (0.21)	0.007 (0.01)	0.601 (0.19)	0.761 (0.17)

Table S5.8: Overall MSE results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.032 (0.01)	0.185 (0.06)	0.071 (0.01)	0.127 (0.05)
	1	0.013 (0.01)	0.134 (0.04)	0.033 (0.01)	0.054 (0.02)
	1.5	0.007 (0.01)	0.132 (0.06)	0.023 (0.01)	0.033 (0.01)
	2	0.005 (0.01)	0.138 (0.08)	0.017 (0.01)	0.025 (0.02)
500	0.5	0.009 (0.00)	0.209 (0.08)	0.133 (0.02)	0.203 (0.06)
	1	0.002 (0.00)	0.170 (0.08)	0.033 (0.00)	0.063 (0.02)
	1.5	0.001 (0.00)	0.146 (0.07)	0.013 (0.00)	0.032 (0.01)
	2	0.001 (0.00)	0.143 (0.06)	0.007 (0.00)	0.022 (0.01)
1000	0.5	0.004 (0.00)	0.269 (0.07)	0.202 (0.03)	0.605 (0.14)
	1	0.001 (0.00)	0.181 (0.06)	0.030 (0.00)	0.229 (0.07)
	1.5	0.001 (0.00)	0.175 (0.06)	0.010 (0.00)	0.094 (0.02)
	2	0.000 (0.00)	0.168 (0.05)	0.004 (0.00)	0.047 (0.01)
2000	0.5	0.002 (0.00)	0.204 (0.02)	0.037 (0.00)	0.043 (0.00)
	1	0.001 (0.00)	0.132 (0.01)	0.007 (0.00)	0.011 (0.00)
	1.5	0.000 (0.00)	0.122 (0.01)	0.003 (0.00)	0.005 (0.00)
	2	0.000 (0.00)	0.118 (0.01)	0.001 (0.00)	0.003 (0.00)

Table S5.9: Nonzero functions MSE results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.254 (0.11)	0.487 (0.09)	0.425 (0.07)	0.585 (0.20)
	1	0.106 (0.05)	0.441 (0.07)	0.244 (0.05)	0.304 (0.10)
	1.5	0.057 (0.05)	0.444 (0.08)	0.179 (0.04)	0.212 (0.09)
	2	0.046 (0.05)	0.444 (0.11)	0.145 (0.05)	0.179 (0.12)
500	0.5	0.061 (0.02)	0.309 (0.11)	0.277 (0.05)	0.312 (0.12)
	1	0.017 (0.00)	0.258 (0.09)	0.094 (0.02)	0.130 (0.04)
	1.5	0.008 (0.00)	0.240 (0.08)	0.045 (0.01)	0.088 (0.03)
	2	0.005 (0.00)	0.238 (0.07)	0.029 (0.01)	0.074 (0.03)
1000	0.5	0.026 (0.00)	0.288 (0.07)	0.250 (0.06)	0.641 (0.21)
	1	0.008 (0.00)	0.201 (0.06)	0.046 (0.01)	0.252 (0.11)
	1.5	0.004 (0.00)	0.199 (0.06)	0.016 (0.00)	0.100 (0.03)
	2	0.002 (0.00)	0.189 (0.05)	0.008 (0.00)	0.050 (0.02)
2000	0.5	0.013 (0.00)	0.210 (0.03)	0.040 (0.01)	0.044 (0.01)
	1	0.004 (0.00)	0.139 (0.02)	0.009 (0.00)	0.012 (0.00)
	1.5	0.002 (0.00)	0.127 (0.02)	0.004 (0.00)	0.005 (0.00)
	2	0.001 (0.00)	0.124 (0.02)	0.002 (0.00)	0.003 (0.00)

Table S5.10: Zero functions MSE results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.008 (0.00)	0.151 (0.06)	0.032 (0.01)	0.076 (0.04)
	1	0.003 (0.00)	0.100 (0.04)	0.010 (0.00)	0.026 (0.01)
	1.5	0.001 (0.00)	0.097 (0.06)	0.005 (0.00)	0.013 (0.01)
	2	0.001 (0.00)	0.104 (0.08)	0.003 (0.00)	0.008 (0.01)
500	0.5	0.003 (0.00)	0.198 (0.07)	0.117 (0.02)	0.191 (0.06)
	1	0.001 (0.00)	0.160 (0.08)	0.026 (0.00)	0.056 (0.02)
	1.5	0.000 (0.00)	0.135 (0.07)	0.010 (0.00)	0.026 (0.01)
	2	0.000 (0.00)	0.133 (0.06)	0.005 (0.00)	0.016 (0.00)
1000	0.5	0.002 (0.00)	0.267 (0.07)	0.197 (0.03)	0.601 (0.14)
	1	0.000 (0.00)	0.179 (0.06)	0.028 (0.00)	0.227 (0.07)
	1.5	0.000 (0.00)	0.172 (0.06)	0.009 (0.00)	0.093 (0.02)
	2	0.000 (0.00)	0.166 (0.05)	0.004 (0.00)	0.046 (0.01)
2000	0.5	0.001 (0.00)	0.204 (0.02)	0.037 (0.00)	0.043 (0.00)
	1	0.000 (0.00)	0.132 (0.01)	0.007 (0.00)	0.011 (0.00)
	1.5	0.000 (0.00)	0.122 (0.01)	0.003 (0.00)	0.005 (0.00)
	2	0.000 (0.00)	0.118 (0.01)	0.001 (0.00)	0.003 (0.00)

Table S5.11: Linear functions MSE results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.175 (0.12)	0.266 (0.11)	0.352 (0.09)	0.623 (0.28)
	1	0.062 (0.02)	0.221 (0.09)	0.193 (0.05)	0.321 (0.14)
	1.5	0.034 (0.02)	0.229 (0.12)	0.138 (0.04)	0.220 (0.12)
	2	0.028 (0.03)	0.219 (0.10)	0.113 (0.05)	0.192 (0.16)
500	0.5	0.049 (0.02)	0.224 (0.12)	0.241 (0.06)	0.326 (0.14)
	1	0.014 (0.00)	0.175 (0.10)	0.077 (0.02)	0.143 (0.05)
	1.5	0.007 (0.00)	0.150 (0.08)	0.037 (0.01)	0.100 (0.05)
	2	0.004 (0.00)	0.144 (0.07)	0.024 (0.01)	0.092 (0.05)
1000	0.5	0.023 (0.01)	0.270 (0.08)	0.248 (0.09)	0.642 (0.30)
	1	0.007 (0.00)	0.176 (0.07)	0.045 (0.02)	0.268 (0.14)
	1.5	0.003 (0.00)	0.178 (0.06)	0.016 (0.01)	0.104 (0.05)
	2	0.002 (0.00)	0.170 (0.06)	0.008 (0.00)	0.050 (0.03)
2000	0.5	0.011 (0.00)	0.204 (0.04)	0.040 (0.01)	0.045 (0.02)
	1	0.003 (0.00)	0.133 (0.03)	0.009 (0.00)	0.012 (0.00)
	1.5	0.001 (0.00)	0.121 (0.02)	0.004 (0.00)	0.005 (0.00)
	2	0.001 (0.00)	0.119 (0.02)	0.002 (0.00)	0.003 (0.00)

Table S5.12: Nonlinear functions MSE results from Simulation 1. n denotes the sample size, SNR denotes the signal-to-noise ratio. Each cell contains the average value of the metric across 100 replicates (with the standard deviation across the 100 replicates in parentheses).

n	SNR	GPVIBES	VCBART	VCFREQ	VCSSLL
200	0.5	0.026 (0.01)	0.181 (0.06)	0.059 (0.01)	0.107 (0.05)
	1	0.011 (0.01)	0.130 (0.04)	0.027 (0.01)	0.043 (0.02)
	1.5	0.006 (0.00)	0.128 (0.06)	0.018 (0.00)	0.025 (0.01)
	2	0.005 (0.00)	0.135 (0.08)	0.013 (0.00)	0.018 (0.01)
500	0.5	0.007 (0.00)	0.209 (0.08)	0.129 (0.02)	0.198 (0.06)
	1	0.002 (0.00)	0.170 (0.08)	0.031 (0.00)	0.060 (0.02)
	1.5	0.001 (0.00)	0.145 (0.07)	0.012 (0.00)	0.029 (0.01)
	2	0.001 (0.00)	0.143 (0.06)	0.007 (0.00)	0.019 (0.01)
1000	0.5	0.003 (0.00)	0.269 (0.07)	0.200 (0.03)	0.603 (0.14)
	1	0.001 (0.00)	0.181 (0.05)	0.029 (0.00)	0.228 (0.07)
	1.5	0.000 (0.00)	0.175 (0.06)	0.009 (0.00)	0.094 (0.02)
	2	0.000 (0.00)	0.168 (0.05)	0.004 (0.00)	0.047 (0.01)
2000	0.5	0.002 (0.00)	0.204 (0.02)	0.037 (0.00)	0.043 (0.00)
	1	0.000 (0.00)	0.132 (0.01)	0.007 (0.00)	0.011 (0.00)
	1.5	0.000 (0.00)	0.122 (0.01)	0.003 (0.00)	0.005 (0.00)
	2	0.000 (0.00)	0.118 (0.01)	0.001 (0.00)	0.003 (0.00)

CHAPTER VI

Conclusion

Overview In this dissertation, I developed several Bayesian integrative statistical procedures motivated by problems in context of precision oncology. In Chapter II, I developed TransPRECISE, a Bayesian network-based integration procedure to combine and compare multi-system proteomic pathways. In Chapter III, I worked on fiBAG, a hierarchical Bayesian framework to integrate multi-omics and clinical data from patients. In Chapter IV, I formulated BaySyn, a multi-stage Bayesian pipeline to integrate multi-platform data across both patient tumors and cancer models. Finally, in Chapter V, I proposed GPVIBES, a Gaussian process-based varying coefficient modeling procedure using Bayesian variable selection to integrate tumor microenvironment summaries in clinicogenomic models.

Methodological and Scientific Contributions From a methodological perspective, in each chapter, I have investigated the mechanics of the proposed procedure in great detail. In Chapters III and V, I performed extensive simulation studies to compare the methods developed by me against state-of-the-art methods developed to tackle similar problems. In all chapters, I have tested the utility of the methods and the interpretability of the results via performing large-scale pan-cancer integrative analyses of patient and model system data from biological databases of interest. I assessed these results in context of existing scientific knowledge by investigating the existing clinical oncological literature extensively and aligning such knowledge with outputs from my models. In each chapter, I discuss potential future directions both through the scientific and methodological axes that can further boost the growing knowledge base of integrative biostatistical research. To ensure reproducibility of the model results and accessibility of the scientific outputs in part of readers from all knowledge domains, I have developed R shiny-based interactive dashboards for each chapter, as listed in Chapter I and described in each individual chapter. These dashboards provide software codes and processed datasets to reproduce the results reported in this dissertation, as well as an interactive collection of such results for the users to probe into. I sincerely believe that these collections will contribute significantly both to the utility and merit of this dissertation as well as to the advancement of the greater scientific cause.

Evolution of increasingly enriched multi-omic and clinical datasets The rapid advancement of genomic sequencing methods in terms of both computational efficiency and cellular granularity has multiplied the opportunities in the integrative and personalized inference paradigms by manifolds. While evolution of techniques such as next-generation sequencing has made aggregation of high-sample-size patient data repositories possible, single cell and spatial omics methods have begun to offer increasingly microscopic and structured observation of the tumor microenvironment (Kumar et al., 2019; Kashima et al., 2020). However, publicly available repositories aggregating such novel datasets are typically in progress and yet to achieve completion, with new samples and cancer types still being recruited and processed. Hence, the data applications illustrated in each chapter of this dissertation relies on omics databases reliant on bulk sequencing techniques, both for patients (TCGA) and cancer model systems (CCLE). Such databases provide complete, processed, and normalized multi-omics datasets that have repeatedly been validated in recent studies and utilized to map pan-cancer and cancer-specific features of the genome. However, all the statistical procedures developed in this dissertation can be generalized for the purpose of applications to databases of higher depth. For the purpose of exposition, I discuss one such deep omics dataset, namely, the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC, Ellis et al., 2013). Across three stages of data collection completed so far, CPTAC provides information on more than 1,500 samples across more than 10 cancer types. The data types available include DNA-level information based on whole genome sequencing, mRNA expression quantifications based on whole exome sequencing, harmonized deep- and phospho-proteomic expression data, along with clinical and biospecimen information. Such high-resolution multi-platform data increases the potential to uncover cellular mechanisms of cancer along with therapeutic opportunities in a data-driven manner. At the same time, the deeper sequencing techniques lead to higher complexity and variability in the datasets, leading to a necessity to adapt the methods proposed in this dissertation to such real datasets in a suitable manner. I conclude with a discussion of how this can be executed in the case of CPTAC, as described below.

Applicability of the developed methods to CPTAC data The proteomic data used in this dissertation is based on reverse-phase protein array (RPPA), which only focuses on a specific handpicked set of proteins (around 200 in number) that are of functional and therapeutic importance in the cancers assessed. However, CPTAC provides a global proteomic profile of the samples, with both standard and phosphorylated proteins included, making it possible to glean the functional changes in the cellular cycle in a more comprehensive manner compared to a targeted panel such as RPPA (Deb et al., 2020; Liu et al., 2022). Typically, a cancer-specific CPTAC proteomics panel contains anywhere between 7,500-12,000 proteins. This does not demand any major changes in the pathway-specific Bayesian neighborhood selection model in Chapter II or the protein-specific Gaussian process-based mechanistic models in Chapter III and Chapter IV, since these models can be parallelized respectively across the pathways and the proteins. Rather, CPTAC datasets contain the potential to assess the functional relevance of more proteins and pathways than those assessed previously. The calibrated Bayesian variable selection (cBVS)-based outcome models in Chapter III and Chapter IV and the varying coefficient-based Bayesian variable selection model in Chapter V, on the other hand, utilize all the proteomic expressions together. Hence, to generalize their applications to the CPTAC panel, these variable selection models need to be able to accommodate $p = 1,000$ or more variables at once. The simulation studies in Chapter III indicate that for n/p ratios $\geq 1/4$, the cBVS model performs reasonably. In Chapter V, the simulation studies illustrate that the varying coefficient-based variable selection model performs convincingly for n/p ratios ≥ 1. For cases where the sample size is not sufficient to accommodate all the protein expressions from CPTAC, the proteomic panel is required to be filtered to a smaller dimension. Compared to RPPA datasets which typically contain $5 - 10\%$ of missing data, CPTAC proteomic datasets contain higher proportions of missing data, sometimes ranging to more than 50%. Hence, cut-offs based on data missingness and quality can be employed to select only the proteins with the highest sample sizes for incorporation into the models. Further, at a cellular level, proteins typically act in pathways in an interconnected manner. Such biological knowledge may be leveraged to include representatives from such pathways to bring the overall number of proteomic candidates further down.

Bibliography

Abrahao-Machado, L. F. and Scapulatempo-Neto, C. (2016). Her2 testing in gastric cancer: An update. *World Journal of gastroenterology*, 22(19):4619.

Adashek, J., Arroyo-Martinez, Y. M., Menta, A. K., Kurzrock, R., and Kato, S. (2020). Therapeutic implications of epidermal growth factor receptor (egfr) in the treatment of metastatic gastric/gej cancer. *Frontiers in Oncology*, 10:1312.

Adorno-Cruz, V., Kibria, G., Liu, X., Doherty, M., Junk, D. J., Guan, D., Hubert, C., Venere, M., Mulkearns-Hubert, E., Sinyuk, M., et al. (2015). Cancer stem cells: targeting the roots of cancer, seeds of metastasis, and sources of therapy resistance. *Cancer research*, 75(6):924–929.

Agliari, A. and Parisetti, C. C. (1988). A-g reference informative prior: A note on zellner's g-prior. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 37(3):271–275.

Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., Arachchi, H., Arora, A., Auman, J. T., Ayala, B., et al. (2015). Genomic classification of cutaneous melanoma. *Cell*, 161(7):1681–1696.

Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., et al. (2015). Toward understanding and exploiting tumor heterogeneity. *Nature medicine*, 21(8):846–853.

Alonso-López, D., Campos-Laborie, F. J., Gutiérrez, M. A., Lambourne, L., Calderwood, M. A., Vidal, M., and De Las Rivas, J. (2019). Apid database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database*, 2019.

Angel, P. W., Rajab, N., Deng, Y., Pacheco, C. M., Chen, T., Lê Cao, K.-A., Choi, J., and Wells, C. A. (2020). A simple, scalable approach to building a cross-platform transcriptome atlas. *PLoS computational biology*, 16(9):e1008219.

Arbiser, J. L. (2018). Diablo: A double-edged sword in cancer? *Molecular Therapy*, 26(3):678–679.

Arellano-Llamas, A., Garcia, F. J., Perez, D., Cantu, D., Espinosa, M., De la Garza, J. G., Maldonado, V., and Melendez-Zajgla, J. (2006). High smac/diablo expression is associated with early local recurrence of cervical cancer. *BMC cancer*, 6(1):1–10.

Arienti, C., Pignatta, S., and Tesei, A. (2019). Epidermal growth factor receptor family and its role in gastric cancer. *Frontiers in oncology*, 9:1308.

Arneth, B. (2019). Tumor microenvironment. *Medicina*, 56(1):15.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Asiedu, M. K., Ingle, J. N., Behrens, M. D., Radisky, D. C., and Knutson, K. L. (2011). Tgfβ/tnfα-mediated epithelial–mesenchymal transition generates breast cancer stem cells with a claudin-low phenotypetgfβ/tnfα and bcscs. *Cancer research*, 71(13):4707–4719.

Aydin, K., Okutur, S. K., Bozkurt, M., Turkmen, I., Namal, E., Pilanci, K., Ozturk, A., Akcali, Z., Dogusoy, G., and Demir, O. G. (2014). Effect of epidermal growth factor receptor status on the outcomes of patients with metastatic gastric cancer: A pilot study. *Oncology letters*, 7(1):255–259.

Ayuk, S. M. and Abrahamse, H. (2019). mtor signaling pathway in cancer targets photodynamic therapy in vitro. *Cells*, 8(5):431.

Bai, R., Boland, M. R., and Chen, Y. (2019). Fast algorithms and theory for high-dimensional bayesian varying coefficient models. *arXiv preprint arXiv:1907.06477*.

Baladandayuthapani, V., Ji, Y., Talluri, R., Nieto-Barajas, L. E., and Morris, J. S. (2010). Bayesian random segmentation models to identify shared copy number aberrations for array cgh data. *Journal of the american statistical association*, 105(492):1358–1375.

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.-K., Chuang, R., Jaehnig, E. J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., et al. (2010). Rewiring of genetic networks in response to dna damage. *Science*, 330(6009):1385–1389.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607.

Bashari, M. H., Fan, F., Vallet, S., Sattler, M., Arn, M., Luckner-Minden, C., Schulze-Bergkamen, H., Zörnig, I., Marme, F., Schneeweiss, A., et al. (2016). Mcl-1 confers protection of her2-positive breast cancer cells to hypoxia: therapeutic implications. *Breast Cancer Research*, 18(1):1–15.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, page asw042.

Bi, X., Lu, Y., Bi, X., and Luo, Z. (2021). Hot clear cell renal cell carcinoma immune status: Illusion or reality? *Journal of Clinical Oncology*, 39(15_suppl):e16575–e16575.

Bitto, A. and Frühwirth-Schnatter, S. (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, 210(1):75–97.

Black, J. B., Purayil, H. T., Mahmud, I., Liao, D., and Daaka, Y. (2018). Abstract lb-312: Grk5 activity mediates in vivo and in vitro prostate cancer progression and chemoresistance. *Cancer Research*, 78(13_Supplement):LB–312.

Black, J. R. and McGranahan, N. (2021). Genetic and non-genetic clonal diversity in cancer evolution. *Nature Reviews Cancer*, 21(6):379–392.

Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J., and Shah, S. P. (2022). Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, 22(2):114–126.

Boonstra, P. S. and Barbaro, R. P. (2020). Incorporating historical models with adaptive bayesian updates. *Biostatistics*, 21(2):e47–e64.

Boonstra, P. S., Taylor, J. M., and Mukherjee, B. (2015). Data-adaptive shrinkage via the hyperpenalized EM algorithm. *Statistics in biosciences*, 7(2):417–431.

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017). Ipf-lasso: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and mathematical methods in medicine*, 2017.

Boyd, L. R. and Muggia, F. M. (2018). Carboplatin/paclitaxel induction in ovarian cancer: the finer points. *Oncology (08909091)*, 32(8).

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186.

Bradfield, A., Button, L., Drury, J., Green, D. C., Hill, C. J., and Hapangama, D. K. (2020). Investigating the role of telomere and telomerase associated genes and proteins in endometrial cancer. *Methods and protocols*, 3(3):63.

Brady, L., Kriner, M., Coleman, I., Morrissey, C., Roudier, M., True, L. D., Gulati, R., Plymate, S. R., Zhou, Z., Birditt, B., et al. (2021). Inter-and intra-tumor heterogeneity of metastatic prostate cancer determined by digital spatial gene expression profiling. *Nature communications*, 12(1):1426.

Bromberg, Y. (2013). Chapter 15: disease gene prioritization. *PLoS computational biology*, 9(4):e1002902.

Buccitelli, C. and Selbach, M. (2020). mrnas, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10):630–644.

Buoncervello, M., Borghi, P., Romagnoli, G., Spadaro, F., Belardelli, F., Toschi, E., and Gabriele, L. (2012). Apicidin and docetaxel combination treatment drives ctcfl expression and hmgb1 release acting as potential antitumor immune response inducers in metastatic breast cancer cells. *Neoplasia*, 14(9):855–IN19.

Byrne, H., Alarcon, T., Owen, M., Webb, S., and Maini, P. (2006). Modelling aspects of cancer dynamics: a review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 364(1843):1563–1578.

Califano, A. and Alvarez, M. J. (2017). The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nature reviews Cancer*, 17(2):116–130.

Campos-Parra, A. D., Padua-Bracho, A., Pedroza-Torres, A., Figueroa-González, G., Fernández-Retana, J., Millan-Catalan, O., Peralta-Zaragoza, O., de León, D. C., Herrera, L. A., and Pérez-Plasencia, C. (2016). Comprehensive transcriptome analysis identifies pathways with therapeutic potential in locally advanced cervical cancer. *Gynecologic oncology*, 143(2):406–413.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial intelligence and statistics*, pages 73–80. PMLR.

Chakraborty, S., Hosen, M., Ahmed, M., Shekhar, H. U., et al. (2018). Onco-multi-omics approach: a new frontier in cancer research. *BioMed research international*, 2018.

Chatterjee, D., Maitra, T., and Bhattacharya, S. (2020). A short note on almost sure convergence of bayes factors in the general set-up. *The American Statistician*, 74(1):17–20.

Chatterjee, N., Chen, Y.-H., Maas, P., and Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117.

Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M., and Alizadeh, A. A. (2018). Profiling tumor infiltrating immune cells with cibersort. *Cancer Systems Biology: Methods and Protocols*, pages 243–259.

Chen, H., Han, T., Zhao, Y., and Feng, L. (2021). Identification of solute-carrier family 27a molecules (scl27as) as a potential biomarker of ovarian cancer based on bioinformatics and experiments. *Annals of Translational Medicine*, 9(15).

Chen, J., Kinoshita, T., Sukbuntherng, J., Chang, B. Y., and Elias, L. (2016). Ibrutinib inhibits erbb receptor tyrosine kinases and her2-amplified breast cancer cell growth. *Molecular cancer therapeutics*, 15(12):2835–2844.

Chen, J., Liu, C., Cen, J., Liang, T., Xue, J., Zeng, H., Zhang, Z., Xu, G., Yu, C., Lu, Z., et al. (2020). Kegg-expressed genes and pathways in triple negative breast cancer: Protocol for a systematic review and data mining. *Medicine*, 99(18).

Chen, L., Zhang, Y.-H., Lu, G., Huang, T., and Cai, Y.-D. (2017a). Analysis of cancer-related lncrnas using gene ontology and kegg pathways. *Artificial Intelligence in Medicine*, 76:27–36.

Chen, X.-Q., Zheng, L.-X., Li, Z.-Y., and Lin, T.-Y. (2017b). Clinicopathological significance of oestrogen receptor expression in non-small cell lung cancer. *Journal of International Medical Research*, 45(1):51–58.

Cheng, C., Tseng, G., Ghosh, D., and Zhou, X. J. (2015). From transcription factor binding and histone modification to gene expression: Integrative quantitative models. *Integrating Omics Data*, page 380.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Conesa, A. and Beck, S. (2019). Making multi-omics data accessible to researchers. *Scientific data*, 6(1):1–4.

Consortium, G. O. (2021). The gene ontology resource: enriching a gold mine. *Nucleic acids research*, 49(D1):D325–D334.

Creixell, P., Schoof, E. M., Simpson, C. D., Longden, J., Miller, C. J., Lou, H. J., Perryman, L., Cox, T. R., Zivanovic, N., Palmeri, A., et al. (2015). Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell*, 163(1):202–217.

Cros, J., Raffenne, J., Couvelard, A., and Poté, N. (2018). Tumor heterogeneity in pancreatic adenocarcinoma. *Pathobiology*, 85(1-2):64–71.

da Costa, J. B., Gibb, E. A., Nykopp, T. K., Mannas, M., Wyatt, A. W., and Black, P. C. (2018). Molecular tumor heterogeneity in muscle invasive bladder cancer: biomarkers, subtypes, and implications for therapy. In *Urologic Oncology: Seminars and Original Investigations*. Elsevier.

Daily, K., Sui, S. J. H., Schriml, L. M., Dexheimer, P. J., Salomonis, N., Schroll, R., Bush, S., Keddache, M., Mayhew, C., Lotia, S., et al. (2017). Molecular, phenotypic, and sample-associated data to describe pluripotent stem cell lines and derivatives. *Scientific data*, 4(1):1–10.

Deb, B., Sengupta, P., Sambath, J., and Kumar, P. (2020). Bioinformatics analysis of global proteomic and phosphoproteomic data sets revealed activation of nek2 and aurka in cancers. *Biomolecules*, 10(2):237.

Deshpande, S. K., Bai, R., Balocchi, C., Starling, J. E., and Weiss, J. (2020). Vcbart: Bayesian trees for varying coefficients. *arXiv preprint arXiv:2003.06416*.

Dimova, I., Zaharieva, B., Raitcheva, S., Dimitrov, R., Doganov, N., and Toncheva, D. (2006). Tissue microarray analysis of egfr and erbb2 copy number changes in ovarian tumors. *International Journal of Gynecologic Cancer*, 16(1).

Dobrzycka, B., Mackowiak-Matejczyk, B., Terlikowska, K. M., Kulesza-Bronczyk, B., Kinalski, M., and Terlikowski, S. J. (2015). Prognostic significance of pretreatment vegf, survivin, and smac/diablo serum levels in patients with serous ovarian carcinoma. *Tumor Biology*, 36(6):4157–4165.

Dobrzycka, B., Terlikowski, S. J., Bernaczyk, P. S., Garbowicz, M., Niklinski, J., Chyczewski, L., and Kulikowski, M. (2010). Prognostic significance of smac/diablo in endometrioid endometrial cancer. *Folia histochemica et cytobiologica*, 48(4):678–681.

Domcke, S., Sinha, R., Levine, D. A., Sander, C., and Schultz, N. (2013). Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature communications*, 4(1):1–10.

Drost, J. and Clevers, H. (2018). Organoids in cancer research. *Nature Reviews Cancer*, 18(7):407–418.

Du, J.-w., Li, G.-q., Li, Y.-s., and Qiu, X.-g. (2021). Identification of prognostic biomarkers related to the tumor microenvironment in thyroid carcinoma. *Scientific Reports*, 11(1):1–15.

Duan, R., Gao, L., Gao, Y., Hu, Y., Xu, H., Huang, M., Song, K., Wang, H., Dong, Y., Jiang, C., et al. (2021). Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLoS computational biology*, 17(8):e1009224.

El-Sayes, N., Vito, A., and Mossman, K. (2021). Tumor heterogeneity: a great barrier in the age of cancer immunotherapy. *Cancers*, 13(4):806.

Ellis, M. J., Gillette, M., Carr, S. A., Paulovich, A. G., Smith, R. D., Rodland, K. K., Townsend, R. R., Kinsinger, C., Mesri, M., Rodriguez, H., et al. (2013). Connecting genomic alterations to cancer biology with proteomics: the nci clinical proteomic tumor analysis consortium. *Cancer discovery*, 3(10):1108–1112.

Errico, A. (2015). Bcl11a—targeting triple-negative breast cancer? *Nature Reviews Clinical Oncology*, 12(3):127–127.

Espinosa, M., Lizárraga, F., Vázquez-Santillán, K., Hidalgo-Miranda, A., Piña-Sánchez, P., Torres, J., García-Ramírez, R. A., Maldonado, V., Melendez-Zajgla, J., and Ceballos-Cancino, G. (2021). Coexpression of smac/diablo and estrogen receptor in breast cancer. *Cancer Biomarkers*, pages 1–18.

Finotello, F., Calura, E., Risso, D., Hautaniemi, S., and Romualdi, C. (2020). Multi-omic data integration in oncology. *Frontiers in oncology*, 10:1768.

Fox, C. W. and Roberts, S. J. (2012). A tutorial on variational bayesian inference. *Artificial intelligence review*, 38(2):85–95.

Fulda, S. (2013). Regulation of apoptosis pathways in cancer stem cells. *Cancer letters*, 338(1):168–173.

Geeleher, P., Cox, N. J., and Huang, R. S. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitrodrug sensitivity in cell lines. *Genome biology*, 15(3):1–12.

Gentles, A. J. and Gallahan, D. (2011). Systems biology: Confronting the complexity of cancerconfronting the complexity of cancer. *Cancer research*, 71(18):5961–5964.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373.

Gevaert, O., Villalobos, V., Sikic, B. I., and Plevritis, S. K. (2013). Identification of ovarian cancer driver genes by using module network integration of multi-omics data. *Interface focus*, 3(4):20130013.

Ghouse, S. M., Nguyen, H.-M., Bommareddy, P. K., Guz-Montgomery, K., and Saha, D. (2020). Oncolytic herpes simplex virus encoding il12 controls triple-negative breast cancer growth and metastasis. *Frontiers in oncology*, 10:384.

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., et al. (2020). Visualizing and interpreting cancer genomics data via the xena platform. *Nature biotechnology*, 38(6):675–678.

Goodspeed, A., Heiser, L. M., Gray, J. W., and Costello, J. C. (2016). Tumor-derived cell lines as molecular models of cancer pharmacogenomicscancer cell lines as pharmacogenomic models. *Molecular Cancer Research*, 14(1):3–13.

Grant, G. D., Brooks 3rd, L., Zhang, X., Mahoney, J. M., Martyanov, V., Wood, T. A., Sherlock, G., Cheng, C., and Whitfield, M. L. (2013). Identification of cell cycle–regulated genes periodically expressed in u2os cells and their regulation by foxm1 and e2f transcription factors. *Molecular biology of the cell*, 24(23):3634–3650.

Gravalos, C. and Jimeno, A. (2008). Her2 in gastric cancer: a new prognostic factor and a novel therapeutic target. *Annals of oncology*, 19(9):1523–1529.

Grever, M. R., Schepartz, S. A., and Chabner, B. A. (1992). The national cancer institute: cancer drug discovery and development program. In *Seminars in oncology*, volume 19, pages 622–638.

Guo, L., Chen, Y., Luo, J., Zheng, J., and Shao, G. (2019). Yap 1 overexpression is associated with poor prognosis of breast cancer patients and induces breast cancer cell growth by inhibiting pten. *FEBS Open Bio*, 9(3):437–445.

Ha, M. J., Banerjee, S., Akbani, R., Liang, H., Mills, G. B., Do, K.-A., and Baladandayuthapani, V. (2018). Personalized integrated network modeling of the cancer proteome atlas. *Scientific reports*, 8(1):1–14.

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J., and Quackenbush, J. (2013). Inconsistency in large pharmacogenomic studies. *Nature*, 504(7480):389–393.

Han, G., Zhao, W., Song, X., Kwok-Shing Ng, P., Karam, J. A., Jonasch, E., Mills, G. B., Zhao, Z., Ding, Z., and Jia, P. (2017). Unique protein expression signatures of survival time in kidney renal clear cell carcinoma through a pan-cancer screening. *BMC genomics*, 18(6):79–93.

Hao, D., Li, J., Wang, J., Meng, Y., Zhao, Z., Zhang, C., Miao, K., Deng, C., Tsang, B. K., Wang, L., et al. (2019). Non-classical estrogen signaling in ovarian cancer improves chemo-sensitivity and patients outcome. *Theranostics*, 9(13):3952.

Harris, M., Bhuvaneshwar, K., Natarajan, T., Sheahan, L., Wang, D., Tadesse, M. G., Shoulson, I., Filice, R., Steadman, K., Pishvaian, M. J., et al. (2014). Pharmacogenomic characterization of gemcitabine response–a framework for data integration to enable personalized medicine. *Pharmacogenetics and genomics*, 24(2):81–93.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

Hastie, T., Tibshirani, R., Narasimhan, B., Chu, G., and Narasimhan, M. B. (2011). Package 'impute'. *biocViews Bioinformatics, Microarray*.

He, C., Mao, D., Hua, G., Lv, X., Chen, X., Angeletti, P. C., Dong, J., Remmenga, S. W., Rodabaugh, K. J., Zhou, J., et al. (2015). The hippo/yap pathway interacts with egfr signaling and hpv oncoproteins to regulate cervical cancer progression. *EMBO molecular medicine*, 7(11):1426–1449.

Heng, H. H., Stevens, J. B., Bremer, S. W., Liu, G., Abdallah, B. Y., and Christine, J. Y. (2011). Evolutionary mechanisms and diversity in cancer. *Advances in cancer research*, 112:217–253.

Hinne, M., Gronau, Q. F., van den Bergh, D., and Wagenmakers, E.-J. (2020). A conceptual introduction to bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3(2):200–215.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.

Hou, J. Y., Wright, J. D., and Achariyapota, V. (2020). An application of erbb2 receptor inhibitors in a rare case of s310f somatic erbb2 mutation of primary signet-ring cell adenocarcinoma of vagina: A case report and review literature of s310f somatic erbb2 mutation in breast and gynecologic cancers. *Gynecologic Oncology Reports*, 32.

Hudson, I. L. (2021). Data integration using advances in machine learning in drug discovery and molecular biology. *Artificial Neural Networks*, pages 167–184.

Hwang, T., Atluri, G., Xie, M., Dey, S., Hong, C., Kumar, V., and Kuang, R. (2012). Co-clustering phenome–genome for phenotype classification and disease gene discovery. *Nucleic acids research*, 40(19):e146–e146.

Invrea, F., Rovito, R., Torchiaro, E., Petti, C., Isella, C., and Medico, E. (2020). Patient-derived xenografts (pdxs) as model systems for human cancer. *Current opinion in biotechnology*, 63:151–156.

Iriana, S., Ahmed, S., Gong, J., Annamalai, A. A., Tuli, R., and Hendifar, A. E. (2016). Targeting mtor in pancreatic ductal adenocarcinoma. *Frontiers in oncology*, 6:99.

Jain, K. S., Sikora, A. G., Baxi, S. S., and Morris, L. G. (2013). Synchronous cancers in patients with head and neck cancer: risks in the era of human papillomavirus-associated oropharyngeal cancer. *Cancer*, 119(10):1832–1837.

Jennings, E. M., Morris, J. S., Carroll, R. J., Manyam, G. C., and Baladandayuthapani, V. (2013). Bayesian methods for expression-based integration of various types of genomics data. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013(1):1–11.

Jiang, G., Zhang, S., Yazdanparast, A., Li, M., Pawar, A. V., Liu, Y., Inavolu, S. M., and Cheng, L. (2016). Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. *BMC genomics*, 17(7):281–301.

Kampan, N. C., Madondo, M. T., McNally, O. M., Quinn, M., and Plebanski, M. (2015). Paclitaxel and its evolving role in the management of ovarian cancer. *BioMed research international*, 2015.

Kandel, C., Leclair, F., Bou-Hanna, C., Laboisse, C. L., and Mosnier, J.-F. (2014). Association of her1 amplification with poor prognosis in well differentiated gastric carcinomas. *Journal of clinical pathology*, 67(4):307–312.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30.

Kapelner, A. and Bleich, J. (2013). bartmachine: Machine learning with bayesian additive regression trees. *arXiv preprint arXiv:1312.2171*.

Kaplan, A. and Lock, E. F. (2017). Prediction with dimension reduction of multiple molecular data sources for patient survival. *Cancer informatics*, 16:1176935117718517.

Kashima, Y., Sakamoto, Y., Kaneko, K., Seki, M., Suzuki, Y., and Suzuki, A. (2020). Single-cell sequencing techniques from individual to multiomics analyses. *Experimental & Molecular Medicine*, 52(9):1419–1427.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Katsila, T. and Matsoukas, M.-T. (2018). How far have we come with contextual data integration in drug discovery? *Expert Opinion on Drug Discovery*, 13(9):791–794.

Kaufman, C. G. and Sain, S. R. (2010). Bayesian functional {ANOVA} modeling using gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–149.

Keenan, A. B., Jenkins, S. L., Jagodnik, K. M., Koplev, S., He, E., Torre, D., Wang, Z., Dohlman, A. B., Silverstein, M. C., Lachmann, A., et al. (2018). The library of integrated network-based cellular signatures nih program: system-level cataloging of human cells response to perturbations. *Cell systems*, 6(1):13–24.

Ketola, K., Munuganti, R. S., Davies, A., Nip, K. M., Bishop, J. L., and Zoubeidi, A. (2017). Targeting prostate cancer subtype 1 by forkhead box m1 pathway inhibitiontargeting pcs1 by foxm1 pathway inhibition. *Clinical Cancer Research*, 23(22):6923–6933.

Kim, J., Koo, B.-K., and Knoblich, J. A. (2020). Human organoids: model systems for human biology and medicine. *Nature Reviews Molecular Cell Biology*, 21(10):571–584.

Kim, K. K., Han, A., Yano, N., Ribeiro, J. R., Lokich, E., Singh, R. K., and Moore, R. G. (2015). Tetrathiomolybdate mediates cisplatin-induced p38 signaling and egfr degradation and enhances response to cisplatin therapy in gynecologic cancers. *Scientific reports*, 5(1):1–11.

Kim, Y., Dillon, P. M., Park, T., and Lee, J. K. (2018). Concord biomarker prediction for novel drug introduction to different cancer types. *Oncotarget*, 9(1):1091.

Kumar, K. R., Cowley, M. J., and Davis, R. L. (2019). Next-generation sequencing and emerging technologies. In *Seminars in thrombosis and hemostasis*, volume 45, pages 661–673. Thieme Medical Publishers.

Kumar, S., Mahdi, H., Bryant, C., Shah, J. P., Garg, G., and Munkarah, A. (2010). Clinical trials and progress with paclitaxel in ovarian cancer. *International journal of women's health*, 2:411.

Kumaraswamy, E., Wendt, K. L., Augustine, L. A., Stecklein, S. R., Sibala, E. C., Li, D., Gunewardena, S., and Jensen, R. A. (2015). Brca1 regulation of epidermal growth factor receptor (egfr) expression in human breast cancer cells involves microrna-146a and is critical for its tumor suppressor function. *Oncogene*, 34(33):4333–4346.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

La Porta, C. A. and Zapperi, S. (2017). Complexity in cancer stem cells and tumor evolution: Toward precision medicine. In *Seminars in cancer biology*, volume 44, pages 3–9. Elsevier.

Lai, Y., Wei, X., Lin, S., Qin, L., Cheng, L., and Li, P. (2017). Current status and perspectives of patient-derived xenograft models in cancer research. *Journal of hematology & oncology*, 10(1):1–14.

Lamb, J. (2007). The connectivity map: a new tool for biomedical research. *Nature reviews cancer*, 7(1):54–60.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K. N., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935.

Lawrence, N. D., Sanguinetti, G., and Rattray, M. (2007). Modelling transcriptional regulation using gaussian processes. *Advances in Neural Information Processing Systems*, 19:785.

Leclercq, M., Vittrant, B., Martin-Magniette, M. L., Scott Boyer, M. P., Perin, O., Bergeron, A., Fradet, Y., and Droit, A. (2019). Large-scale automatic feature selection for biomarker discovery in high-dimensional omics data. *Frontiers in genetics*, 10:452.

Lee, J., Franovic, A., Shiotsu, Y., Kim, S. T., Kim, K.-M., Banks, K. C., Raymond, V. M., and Lanman, R. B. (2019). Detection of erbb2 (her2) gene amplification events in cell-free dna and response to anti-her2 agents in a large asian cancer patient cohort. *Frontiers in oncology*, 9:212.

Levine, D. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73.

Levitin, H. M., Yuan, J., and Sims, P. A. (2018). Single-cell transcriptomic analysis of tumor heterogeneity. *Trends in cancer*, 4(4):264–268.

Li, H., Sun, X., Zhao, Y., Zhang, C., Jiang, K., Ren, J., Xing, L., and He, M. (2023). Pan-cancer analysis of tasl: a novel immune infiltration-related biomarker for tumor prognosis and immunotherapy response prediction. *BMC cancer*, 23(1):1–22.

Li, J., Akbani, R., Zhao, W., Lu, Y., Weinstein, J. N., Mills, G. B., and Liang, H. (2017a). Explore, visualize, and analyze functional cancer proteomic data using the cancer proteome atlas. *Cancer research*, 77(21):e51–e54.

Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P. L., Liu, W., Yang, J.-Y., Broom, B. M., Verhaak, R. G., Kane, D. W., et al. (2013). Tcpa: a resource for cancer functional proteomics data. *Nature methods*, 10(11):1046–1047.

Li, J., Zhao, W., Akbani, R., Liu, W., Ju, Z., Ling, S., Vellano, C. P., Roebuck, P., Yu, Q., Eterovic, A. K., et al. (2017b). Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer cell*, 31(2):225–239.

Libring, S., Shinde, A., Chanda, M. K., Nuru, M., George, H., Saleh, A. M., Abdullah, A., Kinzer-Ursem, T. L., Calve, S., Wendt, M. K., et al. (2020). The dynamic relationship of breast cancer cells and fibroblasts in fibronectin accumulation at primary and metastatic tumor sites. *Cancers*, 12(5):1270.

Lim, Z.-F. and Ma, P. C. (2019). Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. *Journal of hematology & oncology*, 12(1):1–18.

Litovkin, K., Joniau, S., Lerut, E., Laenen, A., Gevaert, O., Spahn, M., Kneitz, B., Isebaert, S., Haustermans, K., Beullens, M., et al. (2014). Methylation of pitx2, hoxd3, rassf1 and tdrd1 predicts biochemical recurrence in high-risk prostate cancer. *Journal of cancer research and clinical oncology*, 140(11):1849–1861.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088.

Liu, W., Cui, Y., Liu, W., Liu, Z., Xu, L., and Li, E. (2022). Deep proteome profiling promotes whole proteome characterization and drug discovery for esophageal squamous cell carcinoma. *Cancer Biology & Medicine*, 19(3):273.

Lumachi, F., Brunello, A., Maruzzo, M., Basso, U., and Mm Basso, S. (2013). Treatment of estrogen receptor-positive breast cancer. *Current medicinal chemistry*, 20(5):596–604.

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). Gage: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, 10(1):1–17.

Ma, X., Chen, T., and Sun, F. (2014). Integrative approaches for predicting protein function and prioritizing genes for complex phenotypes using protein interaction networks. *Briefings in Bioinformatics*, 15(5):685–698.

Malaguti, P., D'Aloia, M. M., and Alimandi, M. (2015). The erbb2 receptor in gastric cancer: the quick-change artist. *Translational Gastrointestinal Cancer*, 4(4):282–293.

Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L., Weinstein, J. N., Kamińska, B., Huelsken, J., Omberg, L., Gevaert, O., et al. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell*, 173(2):338–354.

Mamoor, S. (2021). Over-expression of ribosomal protein s6 kinase a1 in human endometrial cancer. *OSF Preprints*.

Martelotto, L. G., Ng, C. K., Piscuoglio, S., Weigelt, B., and Reis-Filho, J. S. (2014). Breast cancer intra-tumor heterogeneity. *Breast Cancer Research*, 16:1–11.

Marusyk, A. and Polyak, K. (2010). Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1805(1):105–117.

McGuffey, E. J. (2015). *Statistical methods for integrating genomics data*. Texas A&M University.

McGuirt, W. F., Matthews, B., and Koufman, J. A. (1982). Multiple simultaneous tumors in patients with head and neck cancer. a prospective, sequential panendoscopic study. *Cancer*, 50(6):1195–1199.

Mehner, C., Oberg, A. L., Goergen, K. M., Kalli, K. R., Maurer, M. J., Nassar, A., Goode, E. L., Keeney, G. L., Jatoi, A., Radisky, D. C., et al. (2017). Egfr as a prognostic biomarker and therapeutic target in ovarian cancer: evaluation of patient cohort and literature review. *Genes & cancer*, 8(5-6):589.

Micchelli, C. A., Xu, Y., and Zhang, H. (2006). Universal kernels. *Journal of Machine Learning Research*, 7(12).

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Genes*, 10(2):87.

Morris, J. S. and Baladandayuthapani, V. (2017). Statistical contributions to bioinformatics: Design, modelling, structure learning and integration. *Statistical modelling*, 17(4-5):245–289.

Nakamori, M., Fu, X., Meng, F., Jin, A., Tao, L., Bast Jr, R. C., and Zhang, X. (2003). Effective therapy of metastatic ovarian cancer with an oncolytic herpes simplex virus incorporating two membrane fusion mechanisms. *Clinical cancer research*, 9(7):2727–2733.

Nguyen, K. G., Wagner, E. S., Vrabel, M. R., Mantooth, S. M., Meritet, D. M., and Zaharoff, D. A. (2021). Safety and pharmacokinetics of intravesical chitosan/interleukin-12 immunotherapy in murine bladders. *Bladder Cancer*, 7(4):427–437.

Ni, Y., Stingo, F. C., Ha, M. J., Akbani, R., and Baladandayuthapani, V. (2019). Bayesian hierarchical varying-sparsity regression models with application to cancer proteogenomics. *Journal of the American Statistical Association*, 114(525):48–60.

Nicora, G., Vitali, F., Dagliati, A., Geifman, N., and Bellazzi, R. (2020). Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. *Frontiers in oncology*, 10:1030.

Ozer, M. E., Sarica, P. O., and Arga, K. Y. (2020). New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines. *Omics: a journal of integrative biology*, 24(5):241–246.

Pak, M., Lee, S., Sung, I., Koo, B., and Kim, S. (2023). Improved drug response prediction by drug target data integration via network-based profiling. *Briefings in Bioinformatics*, 24(2):bbad034.

Patel, P., Chaganti, R., and Motzer, R. (2006). Targeted therapy for metastatic renal cell carcinoma. *British journal of cancer*, 94(5):614–619.

Pe'er, D., Ogawa, S., Elhanani, O., Keren, L., Oliver, T. G., and Wedge, D. (2021). Tumor heterogeneity. *Cancer cell*, 39(8):1015–1017.

Plattner, C., Finotello, F., and Rieder, D. (2020). Deconvoluting tumor-infiltrating immune cells from rna-seq data using quantiseq. In *Methods in enzymology*, volume 636, pages 261–285. Elsevier.

Plummer, M., Stukalov, A., Denwood, M., and Plummer, M. M. (2016). Package 'rjags'. *Vienna, Austria*.

Polson, N. G., Scott, J. G., and Windle, J. (2014). The bayesian bridge. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 713–733.

Qabaja, A., Alshalalfa, M., Bismar, T. A., and Alhajj, R. (2013). Protein network-based lasso regression model for the construction of disease-mirna functional interactions. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013:1–11.

Qiu, H., Yu, I., Wang, X.-R., Fu, Z.-M., and Tse, S. (2008). Study on the interaction under logistic regression modeling. *Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi*, 29(9):934–937.

Rattray, M., Yang, J., Ahmed, S., and Boukouvalas, A. (2019). Modelling gene expression dynamics with gaussian process inference. *Handbook of Statistical Genomics: Two Volume Set*, pages 879–20.

Relling, M. V. and Evans, W. E. (2015). Pharmacogenomics in the clinic. *Nature*, 526(7573):343–350.

Reyes, H. D., Thiel, K. W., Carlson, M. J., Meng, X., Yang, S., Stephan, J.-M., and Leslie, K. K. (2014). Comprehensive profiling of egfr/her receptors for personalized treatment of gynecologic cancers. *Molecular diagnosis & therapy*, 18(2):137–151.

Richardson, S., Tseng, G. C., and Sun, W. (2016). Statistical methods in integrative genomics. *Annual review of statistics and its application*, 3:181–209.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Roden, D. M., Altman, R. B., Benowitz, N. L., Flockhart, D. A., Giacomini, K. M., Johnson, J. A., Krauss, R. M., McLeod, H. L., Ratain, M. J., Relling, M. V., et al. (2006). Pharmacogenomics: challenges and opportunities. *Annals of internal medicine*, 145(10):749–757.

Roshandel, G., Boreiri, M., Sadjadi, A., and Malekzadeh, R. (2014). A diversity of cancer incidence and mortality in west asian populations. *Annals of global health*, 80(5):346–357.

Ruggeri, B. A., Camp, F., and Miknyoczki, S. (2014). Animal models of disease: preclinical animal models of cancer and their applications and utility in drug discovery. *Biochemical pharmacology*, 87(1):150–161.

Rytlewski, J., Milhem, M. M., and Monga, V. (2021). Turning ‘cold’tumors ‘hot’: immunotherapies in sarcoma. *Annals of Translational Medicine*, 9(12).

Sartore-Bianchi, A., Trusolino, L., Martino, C., Bencardino, K., Lonardi, S., Bergamo, F., Zagonel, V., Leone, F., Depetris, I., Martinelli, E., et al. (2016). Dual-targeted therapy with trastuzumab and lapatinib in treatment-refractory, kras codon 12/13 wild-type, her2-positive metastatic colorectal cancer (heracles): a proof-of-concept, multicentre, open-label, phase 2 trial. *The Lancet Oncology*, 17(6):738–746.

Shareefi, G., Turkistani, A. N., Alsayyah, A., Kussaibi, H., Had, M. A., and Alkharsah, K. R. (2020). Pathway-affecting single nucleotide polymorphisms (snps) in rps6ka1 and mbip genes are associated with breast cancer risk. *Asian Pacific Journal of Cancer Prevention: APJCP*, 21(7):2163.

Sinha, R., Winer, A. G., Chevinsky, M., Jakubowski, C., Chen, Y.-B., Dong, Y., Tickoo, S. K., Reuter, V. E., Russo, P., Coleman, J. A., et al. (2017). Analysis of renal cancer cell lines from two major resources enables genomics-guided cell line selection. *Nature communications*, 8(1):1–10.

Siolas, D. and Hannon, G. J. (2013). Patient-derived tumor xenografts: transforming clinical samples into mouse models. *Cancer research*, 73(17):5315–5319.

Skibinski, A. and Kuperwasser, C. (2015). The origin of breast tumor heterogeneity. *Oncogene*, 34(42):5309–5316.

Solvang, H. K., Lingjærde, O. C., Frigessi, A., Børresen-Dale, A.-L., and Kristensen, V. N. (2011). Linear and non-linear dependencies between copy number aberrations and mrna expression reveal distinct molecular pathways in breast cancer. *BMC bioinformatics*, 12(1):1–12.

Song, X., Ji, J., Gleason, K. J., Yang, F., Martignetti, J. A., Chen, L. S., and Wang, P. (2019). Insights into impact of dna copy number alteration and methylation on the proteogenomic landscape of human ovarian cancer via a multi-omics integrative analysis. *Molecular & Cellular Proteomics*, 18(8):S52–S65.

Sorensen, T. A. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, 5:1–34.

Spainhour, J. C. G., Lim, J., and Qiu, P. (2017). Gdisc: a web portal for integrative analysis of gene–drug interaction for survival in cancer. *Bioinformatics*, 33(9):1426–1428.

Spainhour, J. C. G. and Qiu, P. (2016). Identification of gene-drug interactions that impact patient survival in tcga. *BMC bioinformatics*, 17(1):1–8.

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051.

Sun, G., Rong, D., Li, Z., Sun, G., Wu, F., Li, X., Cao, H., Cheng, Y., Tang, W., and Sun, Y. (2021). Role of small molecule targeted compounds in cancer: Progress, opportunities, and challenges. *Frontiers in Cell and Developmental Biology*, page 2043.

Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature methods*, 17(2):193–200.

Sun, Y. and Liu, Q. (2015). Deciphering the correlation between breast tumor samples and cell lines by integrating copy number changes and gene expression profiles. *BioMed research international*, 2015.

Tai, H., Wu, Z., Sun, S., Zhang, Z., and Xu, C. (2018). Fgfrl1 promotes ovarian cancer progression by crosstalk with hedgehog signaling. *Journal of immunology research*, 2018.

Tanabe, S., Kawabata, T., Aoyagi, K., Yokozaki, H., and Sasaki, H. (2016). Gene expression and pathway analysis of ctnnb1 in cancer and stem cells. *World journal of stem cells*, 8(11):384.

Tarazona, S., Arzalluz-Luque, A., and Conesa, A. (2021). Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science*, 1(6):395–402.

Testa, U., Pelosi, E., and Castelli, G. (2018). Colorectal cancer: genetic abnormalities, tumor progression, tumor heterogeneity, clonal evolution and tumor-initiating cells. *Medical Sciences*, 6(2):31.

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Yang, T.-H. O., Porta-Pardo, E., Gao, G. F., Plaisier, C. L., Eddy, J. A., et al. (2018). The immune landscape of cancer. *Immunity*, 48(4):812–830.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Timonen, J., Mannerström, H., Vehtari, A., and Lähdesmäki, H. (2021). lgpr: an interpretable non-parametric method for inferring covariate effects from longitudinal data. *Bioinformatics*, 37(13):1860–1867.

Torchiaro, E., Lorenzato, A., Olivero, M., Valdembri, D., Gagliardi, P. A., Gai, M., Erriquez, J., Serini, G., and Di Renzo, M. F. (2016). Peritoneal and hematogenous metastases of ovarian cancer cells are both controlled by the p90rsk through a self-reinforcing cell autonomous mechanism. *Oncotarget*, 7(1):712.

Tseng, G., Ghosh, D., and Zhou, X. J. (2015). *Integrating omics data*. Cambridge University Press.

Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., et al. (2017). Defining a cancer dependency map. *Cell*, 170(3):564–576.

Tsujiura, M., Mazack, V., Sudol, M., Kaspar, H. G., Nash, J., Carey, D. J., and Gogoi, R. (2014). Yes-associated protein (yap) modulates oncogenic features and radiation sensitivity in endometrial cancer. *PloS one*, 9(6):e100974.

Tu, Z., Zhang, B., and Zhu, J. (2015). Network integration of genetically regulated gene expression to study complex diseases. *Integrating Omics Data*, 88:88–109.

Ulapane, N., Thiyagarajan, K., and Kodagoda, S. (2020). Hyper-parameter initialization for squared exponential kernel-based gaussian process regression. In *2020 15th IEEE Conference on Industrial Electronics and applications (ICIEA)*, pages 1154–1159. IEEE.

Vahabi, N. and Michailidis, G. (2022). Unsupervised multi-omics data integration methods: a comprehensive review. *Frontiers in genetics*, 13.

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). Linkedomics: analyzing multi-omics data within and across 32 cancer types. *Nucleic acids research*, 46(D1):D956–D963.

Velten, B. and Huber, W. (2021). Adaptive penalization in high-dimensional regression and classification with external covariates using variational bayes. *Biostatistics*, 22(2):348–364.

Veltman, J. A. and Lupski, J. R. (2015). From genes to genomes in the clinic.

Vitrinel, B., Koh, H. W., Kar, F. M., Maity, S., Rendleman, J., Choi, H., and Vogel, C. (2019). Exploiting interdata relationships in next-generation proteomics analysis. *Molecular & Cellular Proteomics*, 18(8):S5–S14.

Wang, D., He, J., Dong, J., Meyer, T. F., and Xu, T. (2020). The hippo pathway in gynecological malignancies. *American journal of cancer research*, 10(2):610.

Wang, H., Huang, H., Ding, C., and Nie, F. (2013a). Predicting protein–protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. *Journal of Computational Biology*, 20(4):344–358.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2013b). ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159.

Wang, X., Chen, L., Liu, W., Zhang, Y., Liu, D., Zhou, C., Shi, S., Dong, J., Lai, Z., Zhao, B., et al. (2023). Timedb: tumor immune micro-environment cell composition database with automatic analysis and interactive visualization. *Nucleic Acids Research*, 51(D1):D1417–D1424.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Wu, B., Guo, Y., and Kang, J. (2022). Bayesian spatial blind source separation via the thresholded gaussian process. *Journal of the American Statistical Association*, pages 1–12.

Wu, C. and Tuo, Y. (2019). Sycp2 expression is a novel prognostic biomarker in luminal a/b breast cancer. *Future Oncology*, 15(8):817–826.

Wu, J., Li, H., Shi, M., Zhu, Y., Ma, Y., Zhong, Y., Xiong, C., Chen, H., and Peng, C. (2019). Tet1-mediated dna hydroxymethylation activates inhibitors of the wnt/β-catenin signaling pathway to suppress emt in pancreatic tumor cells. *Journal of Experimental & Clinical Cancer Research*, 38(1):1–17.

Wu, Y., Zhang, L., Liu, L., Zhang, Y., Zhao, Z., Liu, X., and Yi, D. (2011). A multifactor dimensionality reduction-logistic regression model of gene polymorphisms and an environmental interaction analysis in cancer research. *Asian Pac J Cancer Prev*, 12(11):2887–2892.

Xu, B., Bai, Z., Yin, J., and Zhang, Z. (2019). Global transcriptomic analysis identifies serpine1 as a prognostic biomarker associated with epithelial-to-mesenchymal transition in gastric cancer. *PeerJ*, 7:e7091.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., et al. (2012). Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961.

Yang, X., Zhu, S., Li, L., Zhang, L., Xian, S., Wang, Y., and Cheng, Y. (2018). Identification of differentially expressed genes and signaling pathways in ovarian cancer by integrated bioinformatics analysis. *OncoTargets and therapy*, 11:1457.

Yao, V., Wong, A. K., and Troyanskaya, O. G. (2018). Enabling precision medicine through integrative network models. *Journal of molecular biology*, 430(18):2913–2923.

Yi, Y. W., You, K. S., Park, J.-S., Lee, S.-G., and Seong, Y.-S. (2021). Ribosomal protein s6: a potential therapeutic target against cancer? *International Journal of Molecular Sciences*, 23(1):48.

Yuan, F., Lu, L., Zhang, Y., Wang, S., and Cai, Y.-D. (2018). Data mining of the cancer-related lncrnas go terms and kegg pathways by using mrmr method. *Mathematical Biosciences*, 304:1–8.

Zanin, R., Pegoraro, S., Ros, G., Ciani, Y., Piazza, S., Bossi, F., Bulla, R., Zennaro, C., Tonon, F., Lazarevic, D., et al. (2019). Hmga1 promotes breast cancer angiogenesis supporting the stability, nuclear localization and transcriptional activity of foxm1. *Journal of Experimental & Clinical Cancer Research*, 38(1):1–23.

Zeng, C., Thomas, D. C., and Lewinger, J. P. (2021a). Incorporating prior knowledge into regularized regression. *Bioinformatics*, 37(4):514–521.

Zeng, W., Tan, H., Jamal, M., Xie, T., Li, J., Song, H., Jang, W., Huang, J., Zhang, Q., Xie, S., et al. (2021b). Expression of cdca7 is a prognostic biomarker and potential therapeutic target in non-small cell lung cancer. *Research Square*.

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L. D., and Ferretti, V. (2019a). The international cancer genome consortium data portal. *Nature biotechnology*, 37(4):367–369.

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., and Do, K.-A. (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(4):595–620.

Zhang, Q., Madden, N. E., Wong, A. S. T., Chow, B. K. C., and Lee, L. T. O. (2017). The role of endocrine g protein-coupled receptors in ovarian cancer progression. *Frontiers in endocrinology*, 8:66.

Zhang, T., Guo, J., Gu, J., Chen, K., Wang, Z., Li, H., Wang, G., and Wang, J. (2019b). Kiaa0101 is a novel transcriptional target of foxm1 and is involved in the regulation of hepatocellular carcinoma microvascular invasion by regulating epithelial-mesenchymal transition. *Journal of Cancer*, 10(15):3501.

Zhang, T., Xu, J., Deng, S., Zhou, F., Li, J., Zhang, L., Li, L., Wang, Q.-E., and Li, F. (2018). Core signaling pathways in ovarian cancer stem cell revealed by integrative analysis of multi-marker genomics data. *PLoS One*, 13(5):e0196351.

Zhou, X.-Y., Dai, H.-Y., Zhang, H., Zhu, J.-L., and Hu, H. (2022). Signal transducer and activator of transcription family is a prognostic marker associated with immune infiltration in endometrial cancer. *Journal of Clinical Laboratory Analysis*, 36(4):e24315.

Zlatian, O. M., Comanescu, M. V., Rosu, A. F., Rosu, L., Cruce, M., Gaman, A. E., Calina, C. D., and Sfredel, V. (2015). Histochemical and immunohistochemical evidence of tumor heterogeneity in colorectal cancer. *Rom J Morphol Embryol*, 56(1):175–81.