PREDICTING CANCER DRUG RESPONSE

USING A RECOMMENDER SYSTEM

CHAYAPORN SUPHAVILAI

NATIONAL UNIVERSITY OF SINGAPORE

2019

PREDICTING CANCER DRUG RESPONSE

USING A RECOMMENDER SYSTEM

CHAYAPORN SUPHAVILAI

(B.S. KU; M.S., IUPUI)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2019

Supervisors:

Professor Wong Lim Soon, Main Supervisor

Associate Professor Niranjan Nagarajan, Co-Supervisor

Examiners:

Dr Anders Jacobson Skanderup

Dr Ng See Kiong

Professor Michael Gromiha M, Indian Institute of Technology Madras

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

ชยพร ศุภวิไล

Chayaporn Suphavilai

7th May 2019

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Associate Professor Niranjan Nagarajan, for his encouragement, suggestions, critical comments for my study progress, as well as his advice on essential skills for my future research. Thanks to Professor Wong Lim Soon for his supervision. Also thanks to my thesis committees for their extensive guidance.

I would like to thank all of the current and previous team members in Nagarajan's Lab at Genome Institute of Singapore for their support and their great suggestions, especially Denis Bertrand for his guidance at the beginning of my cancer research. Also, thanks to the members of DasGupta's Lab for providing important data to fulfill this study.

Special thanks to my best badminton friends for their tremendous moral support, and thanks to the team members and coach from NUS Badminton Team for keeping me up. Enjoying sports with lovely friends always helps me to get through stressful days.

Last but definitely not the least, I would like to express my deepest gratitude to my parents, my sisters, my cousins, and my best companion for their consistent support and love, which supply me the strengths to move forward.

Abstract

As we move towards an era of precision medicine, the ability to predict patient-specific drug responses in cancer based on molecular information such as gene expression data represents both an opportunity and a challenge. In particular, methods are needed that can accommodate the high-dimensionality of data to learn interpretable models with the goal of providing the right drug for the right patient at the right time.

We propose a method based on ideas from recommender systems (CaDRReS) that predicts cancer drug responses for unseen cell lines/patients based on learning projections for drugs and cell-lines into a latent pharmacogenomic space. Comparisons with other proposed approaches for this problem based on large public datasets (CCLE, GDSC) show that CaDRReS provides consistently good models and robust predictions even across unseen patient-derived cell line datasets. Also, analysis of the pharmacogenomic spaces inferred by CaDRReS can be used to understand drug mechanisms, identify cellular subtypes, and characterize drug-pathway associations.

Furthermore, we propose a modified version of CaDRReS for single-cell RNA-seq data to investigate intra-patient drug response heterogeneity, using head and neck cancer as a case study. We showed that systematically combining cell-type specific drug response predictions provided better concordance with *in vitro* drug response when comparing to prediction based on bulk gene expression. Finally, to transfer our *in silico* prediction to a clinic, we incorporate clinical drug response information to predict an upfront patient-specific drug combination that could inhibit multiple cell types identified within a patient, resolving intra-patient heterogeneity.

Contents

List of Figures

Chapter 1

Introduction

Advances in DNA sequencing technologies allow us to now generate large amounts of data to understand cancer, a genetic disease that causes millions of deaths worldwide every year. A key question in cancer genetics and treatment is to identify the genetic basis of uncontrolled growth in cancer cells, and to identify vulnerabilities that can be the target of anti-cancer drugs. Another critical challenge is to understand the heterogeneity of response to anti-cancer drugs across patients. In the last few years, several international consortiums have generated datasets based on systematic screening of anti-cancer drugs against established cancer cell lines, along with their genetic and transcriptional molecular profiles. Consequently, these datasets allow us to identify patterns in these molecular profiles that can explain the varying levels of drug sensitivity across patients, an opportunity that forms the central focus of this thesis.

In the rest of this chapter, we will review the motivation for this thesis including a more detailed introduction to cancer genomics and the need for patient-specific cancer drug response prediction. This will be followed by the list of publications as well as my contributions to each work in the areas of cancer driver and drug response prediction that represent a major part of this thesis.

Subsequent chapters in this thesis will expand on these topics. In **Chapter 2**, we will discuss the challenge of identifying genetic drivers of cancer and the development of a new "consensus-based" cancer driver prediction method. **Chapter 3** introduces the concept of "precision medicine" as it relates to cancer, and details different existing models for cancer drug response prediction, together with their strengths and limitations. In **Chapter 4**, we propose a new cancer drug response prediction model (CaDRReS,), and compare its performance against state-of-the-art methods. **Chapter 5** expands on this theme and discusses the utility of the "pharmacogenomic space" model that is learnt by CaDRReS. **Chapter 6**, discusses challenges related to tumor heterogeneity and how methods such as CaDRReS can be extended and applied in this context. Finally, **Chapter 7** summarizes the major conclusions from this thesis and discusses some important directions for future work.

1.1 Background: Cancer genomics

Cancer causes several million deaths worldwide, and the number of new cases is rising[1]. It is well-known as a genetic disease caused by changes in DNA sequence, i.e., mutations. Mutations can introduce abnormal behaviors in healthy cells through a variety of mechanisms, such as, altering gene expression and protein function. Breakthroughs in sequencing technologies allow us now to rapidly and accurately detect mutations and measure gene expression. Consequently, international efforts such as The Cancer Genome Atlas (TCGA)[2] and International Cancer Genome Consortium (ICGC)[3] have been conducted to collect *omics* information for several thousand tumors, across multiple cancer types, in an effort to understand the underlying mechanisms.

One of the major challenges in cancer genomics is to search for 'driver' mutations that lead to tumor formation, typically by granting a growth advantage to cancer cells. A driver gene harboring driver mutations can be identified by detecting signals of positive selection, cancer-specific signatures in the mutated DNA sequence, as well as from the effects of mutations on transcriptomic profiles and protein functions [4]. The ability to discover driver genes and their related biological functions/pathways has a significant impact on cancer treatment as cancer driver genes can be candidate drug targets [5]. Moreover, besides genomic data, others types of omic profiles including transcriptomes, epigenomes, metabolomes and proteomes also provide complementary information, enabling us to gain insights into the complex mechanisms underlying cancer [6].

Genomic and transcriptomic data are the two most commonly available data types that are used for studying cancer biology. The ability to identify key mutated genes, i.e., biomarkers, that can be used for determining drug response in cancer cells can help us to better target treatments. However, variations in drug response across different cancer cell lines still exists within a group that harbors the same mutational biomarkers [7,8]. Also, it has been shown that these different drug responses in cancer cell lines can be explained by differences in transcriptomic profiles[9]. The heterogeneity of drug response across patients leads us to the challenge of predicting patient-specific drug response based on their multi-omic profiles.

1.2 Motivation: the need of patient-specific drug response prediction

In precision medicine as it pertains to cancer, it is crucial to understand heterogeneity across cancer types as well as across different patients. While the

ultimate goal is to provide the right drug to the right patient at the right time to maximize treatment effectiveness, one of the main challenges is the ability to predict drug response based on the unique molecular profiles of patients. Besides accurate drug response prediction, interpretability is also an essential property of any drug recommender system, allowing us to understand drug response mechanisms that play a role.

Many existing methods focus on constructing a model to predict drug responses of cancer cell lines, independently for each drug [7,9,10]. Although this strategy allows the model to learn specific mechanisms for a given drug, performance and robustness are limited by the small number of cell types tested for each drug. Subsequently, models based on multitask learning, in which parameters are shared across drugs, have been proposed to increase the number of training samples [11]. To capture relationships between drugs and cell lines, a few models based on collaborative filtering techniques have been proposed [12]. These models simultaneously learn hidden properties of cell lines and drugs to predict sample-specific drug responses.

To exploit a larger number of samples to construct a more generalized model and overcome limitations of existing models – such as an inability to predict drug response for unseen samples and information losing in data normalization steps, we have developed Cancer DRug Response prediction using Recommender System (CaDRReS). CaDRReS learns a latent pharmacogenomic space can predict patient-specific drug response based on transcriptomic profiles. Comparisons with other existing methods based on large public datasets shows that CaDRReS provides consistently good models and robust predictions even across unseen patient-derived cell-line datasets. Moreover, the pharmacogenomic space captures drug-drug, cell line-cell line, and drug cell line relationships. Our

extended downstream analyses demonstrate that the pharmacogenomics space learned can be used for understanding drug response mechanisms. Furthermore, we show that CaDRReS can be extended to the analysis of heterogenous patient tumors based on single-cell sequencing approaches to accurately predict cancer drug response and even recommend suitable drug combinations.

1.3 List of publications and contributions

1) Predicting Cancer Drug Response using a Recommender System

Chayaporn Suphavilai, Denis Bertrand, Niranjan Nagarajan. "Predicting Cancer Drug Response using a Recommender System." Bioinformatics.

This work describes the CaDRReS system and the pharmacogenomic space that it learns and is the major contribution of this thesis, including **Chapter 3** which introduces the drug response prediction problem and reviews existing methods, **Chapter 4** which details the development and benchmarking of CaDRReS, and **Chapter 5** which describes downstream analysis based on the pharmacogenomic space.

2) ConsensusDriver Improves upon Individual Algorithms for Predicting Driver Alterations in Different Cancer Types and Individual Patients

Bertrand, Denis, Sibyl Drissler, Burton K. Chia, Jia Yu Koh, Chenhao Li, Chayaporn Suphavilai, Iain Beehuat Tan, and Niranjan Nagarajan. "ConsensusDriver Improves upon Individual Algorithms for Predicting Driver Alterations in Different Cancer Types and Individual Patients." Cancer Research (2017).

This work details results from a systematic comparison of 18 diverse cancer driver prediction methods to understand their strengths and weaknesses. Some of the major results from this work and how this impacts precision oncology is highlighted in **Chapter 2**. Based on the orthogonality of predictions, we report a new consensus method that significantly improves over all existing methods.

In this work, my contributions are identifying relationships between cancer drivers and biological pathways via pathway analysis and studying the roles of cancer drivers as drug targets (actionable drivers). Our observations suggest that the actionable drivers could be used as biomarker features in a drug response prediction model.

3) Comprehensive characterization of cancer driver genes and mutations

Bailey, Matthew H., Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico *et al.* "Comprehensive characterization of cancer driver genes and mutations." Cell 173, no. 2 (2018): 371-385.

We joined The Cancer Genome Atlas Research Network to predict cancer drivers for a newly generated dataset consisting of over 11,000 tumors from 33 cancer types. In this paper, our contribution is to apply multiple tools to identify cancer driver and perform analyses to identify the effects of predicted cancer driver mutations on gene expression levels and pathways. The key results from our analysis are highlighted in **Chapter 2**.

4) Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics

Ding, Li, Matthew H. Bailey, Eduard Porta-Pardo, Vesteinn Thorsson, Antonio Colaprico, Denis Bertrand, David L. Gibbs *et al.* "Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics." Cell 173, no. 2 (2018): 305-320.

This follow-up paper provides an overview of different types of multi-omic analyses based on the newly generated pan-cancer driver gene list. Our contributions are summarized in Figure 5 (Relationships between Oncogenic Processes and Driver Genes) in the paper, including identifying the impacts of cancer drivers on various oncogenic processes, studying mutual exclusive patterns of cancer drivers, as well as uncovering the associations between cancer

types and different cancer-related biological processes, as detailed further in **Chapter 2**.

5) Predicting cancer drug response *in silico* for targeting tumor heterogeneity

Chayaporn Suphavilai, Ankur Sharma, Lorna Tu, Shumei Chia, Ramanuj DasGupta, and Niranjan Nagarajan (manuscript in preparation)

Tumor heterogeneity is well recognized as an important factor in defining treatment response and clinical outcomes in diverse cancer types. The increased availability of single-cell omics approaches opens up the possibility that tumor heterogeneity can be accounted for in computational models for drug response. In this manuscript, we applied CaDRReS-Sc, a new version of CaDRReS for single-cell data (**Chapter 6**), to explore heterogeneity in drug response in head and neck cancer. The ability to predict intra-patient drug response heterogeneity has important applications for combating drug resistance and metastasis, with ongoing efforts focused on confirming CaDRReS-Sc's utility for identifying complementary drug combinations.

Chapter 2

Genomic elements driving cancer

Cancers are believed to arise typically from the accumulation of mutations that eventually transform healthy cells into cancer cells[13]. The development of tumors resembles Darwinian evolution of species where individual cells continuously gain heritable genetic variations that increase their ability to compete, survive, and reproduce. Typically, genes that *drive* cancer can be classified into two types including *oncogenes*, which provide selective growth advantages, and *tumor suppressor genes*, which in their native form prevent uncontrolled cell growth[14].

One of the major challenges in cancer genomics is to search for 'driver' mutations in the sea of 'passenger' mutations that are not related to cancer. Driver mutations disrupt critical biological processes, giving rise to genomic instability, unlimited cell division, sustained proliferative signaling, evasion of growth suppression, altered cellular energetics and/or resistance to apoptosis[4]. Although the majority of cancer driver studies have focused only on genomic information, multi-omic datasets provide complementary information and could allow us to discover novel cancer drivers (**Figure 2.1A**). Several international efforts have generated cancer multi-omic profiles for the community to investigate and

develop tools for predicting cancer drivers. Some examples of large datasets include those from The Cancer Genome Atlas (TCGA)[15], International Cancer Genome Consortium (ICGC)[3] and Catalogue of Somatic Mutations in Cancer (COSMIC)[16].

In this chapter, we provide an overview of cancer driver prediction methods. We then present results from the three cancer driver papers listed in Section 1.3, including our work on consensus methods for predicting drivers and pan-cancer analysis with such tools that highlight the challenges of using mutation information for understanding cancer biology. Analysis of the relationships between actionable cancer drivers —identified based on different underlying hypotheses of how cancer drivers cause cancers— and drug responses suggest the roles of drivers as drug response biomarkers, which in turn could serve as features for predicting drug response in the future.

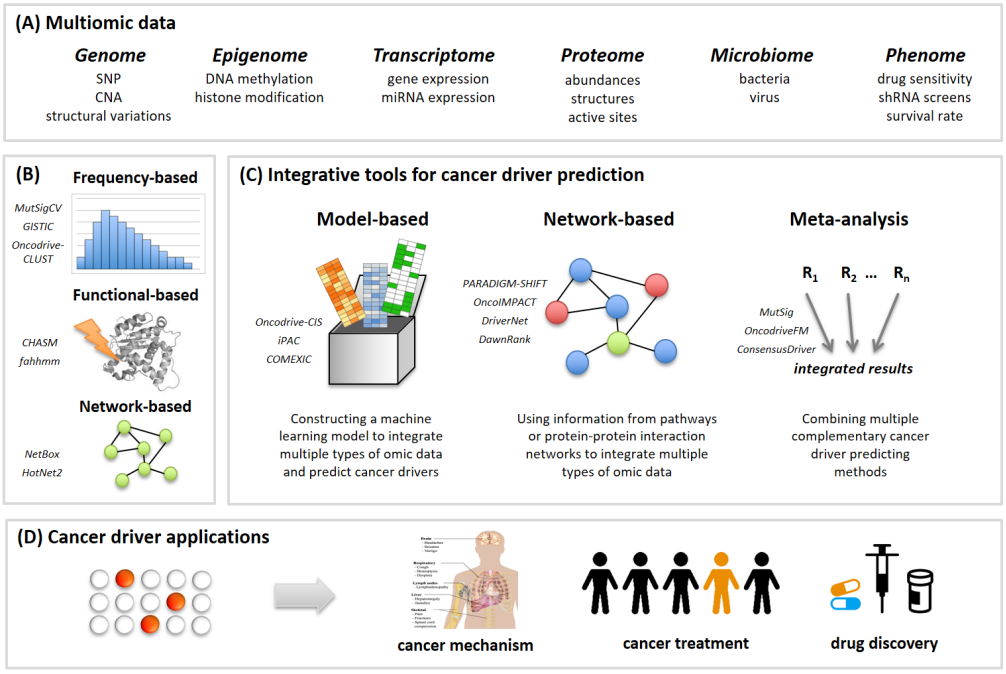

Figure 2.1 An overview of cancer driver prediction and applications. (A) Types of omic data related to cancer driver discovery. (B) Three types of methods for identifying cancer drivers based on genomic profiles. (C) An overview of integrative cancer driver prediction tools based on multi-omic profiles. (D) Applications of cancer drivers.

2.1 Cancer driver prediction

As shown in **Figure 2.1B-C**, despite the diverse approaches used, existing cancer driver prediction methods naturally fall into two classes, (i) those that are primarily based on genomic information and corresponding signals, and (ii) those that try to integrate across multiple omic profiles and thus improve their sensitivity. As we discuss later, their complementary strengths also enable the design of consensus approaches that are simultaneously more sensitive and specific.

Identifying cancer driver genes based on sample genomic profiles

We can categorize existing methods in this class according to their approach for detecting signals of cancer drivers — using background mutation rate to identify genes with high-frequency of mutations in cancer samples (frequency-based), calculating the impact of mutations on protein function (functional impact-based), and incorporating biological networks to identify coherent sets of driver genes (network-based) (**Figure 2.1B**).

Frequency-based tools rely on the hypothesis that frequently mutated genes across tumors likely have a signature of selective advantage. The more frequently a certain gene is mutated in a sample cohort, the more likely that the gene is a driver. This approach typically relies on statistical power based on access to data from a large number of samples, and the assumption that background mutation rates can be appropriately estimated across cancer types, mutation types, and genomic regions[17]. MutSigCV is one of the most widely used tools in this category and identifies frequently mutated genes with respect to the background mutation rate estimated for each gene by considering heterogeneities within and across samples[18]. GISTIC2.0 (Genomic Identification of Significant

Targets in Cancer) similarly identifies recurrent somatic copy number alterations (SCNA) by estimating overall background rates of formation, scoring the SCNAs according to their likelihood of occurrence to identify candidate driver genes[19]. OncodriveCLUST identifies genes containing a cluster of mutations with the underlying hypothesis that gain-of-function mutations tend to cluster in specific protein regions[20].

Functional impact-based tools evaluate the impact of point mutations on protein structure and functions based on information from evolutionary conservation, protein structures, and biochemical properties of mutated residues[21]. SIFT, PolyPhen2, and MutationAssessor[22–24] are some of the widely used tools in this category, but were not specifically designed for identifying cancer drivers. Methods such as CHASM and fathmm[25,26] improve on this basic paradigm by incorporating properties of known cancer genes. For example, CHASM (Cancer-specific High-throughput Annotation of Somatic Mutations) identifies missense mutations that most likely enhance tumor cell proliferation by using a random forest classifier trained on a list of known cancer genes. Their random forest model takes into account the average nucleotide-level conservation, SNP density, and frequency of different types of missense changes. Similarly, fathmm (functional analysis through hidden markov models) uses a machine-learning model that captures the properties of drivers from a list of known cancer genes, based on sequence conservation, disease association information, and functionally neutral amino acid substitutions as a control set. As CHASM and fathmm are based on limited training sets, their ability to generalize across different families of proteins may be limited.

Network-based tools incorporate biological networks such as pathways and protein-protein interaction networks that allow them to infer biologically

coherent groups of mutated genes that may contribute to cancer. For example, NetBox[27] searches for the shortest path between every pair of mutated genes to define a set of strongly interconnected mutated genes, while HotNet2[28] uses a heat diffusion process to model interactions among genes and identify sub-networks that are mutated more than expected by chance. Other types of biological networks are also studied; for example, MAXDRIVER constructs a combined network based on gene similarities, disease phenotype similarities, and gene-disease associations to identify driver genes targeted by CNAs[29]. Another strategy is to search for mutual exclusivity of mutated genes in biological networks based on the hypothesis that different patients may have different sets of driver genes that perturb the same pathways. While Miller *et al*[30] and CoMEt[31] directly identify a set of mutually exclusive mutated genes, Mutex identifies driver genes based on mutual exclusivity of mutated genes that have common downstream genes in a signaling pathway[32].

There are several limitations for methods that are solely based on genomic data. Firstly, estimating background mutations rate is a challenging task because mutation frequencies can vary across cancer types, tissue types, samples, and genomic regions. Secondly, recurrently copy number altered regions can contain many genes, of which only a few are expected to be drivers. Also, some frequency-based and network-based tools require a large number of samples to have sufficient statistical power for distinguishing drivers or detecting a set of mutually exclusive genes above background noise. For functional impact-based tools, although they directly assess the impact of mutations on protein functions and do not require a large number of samples, a high confidence list of cancer driver mutations or genes is still needed for model training. Additionally, methods

that only rely on evolutionary conservation tend to have a higher false positive rate as not every mutation in conserved regions is oncogenic[21].

Integrative methods for cancer driver prediction

Multiple types of omics data, i.e., different types of experimental data from various fields of biological study (-omics) can be integrated to identify cancer driver genes. Combining genomic data with other types of omics data, such as transcriptomic data, can allow us to discover rare and novel cancer driver mutations [33–35]. Integrative methods can be classified based on the strategies employed into three categories: model-based, integrative network-based, and meta-analysis tools.

Model-based tools construct a computational model to predict cancer drivers based on sample multi-omic profiles. For instance, Oncodrive-CIS measures cis effects, i.e., the impact of an alteration on the expression of the gene harboring it, and predicts genes that bias toward deregulation caused by copy number alterations as candidate drivers[36]. iPAC (in-trans Process Associated and Cis-correlated genes) uses statistical tests to assess both cis and trans effects, i.e., impacts of an alteration in a gene on other genes or biological processes as well, to identify candidate driver genes[37]. CONEXIC (COpy Number and EXpression In Cancer) uses a Bayesian network to discover associations between a candidate driver and a set of differentially expressed genes and searches for combinations of candidate drivers that most likely explain changes in expression across samples[38]. The method CNAmet integrates genomic, transcriptomic, and epigenomic data through a model that uses a signal-to-noise ratio statistic based on gene expression values across samples to calculate methylation and copy number weights for detecting driver genes[39]. Helios identifies significantly amplified regions and uses information on point mutations, gene expression data,

and shRNA screening data to prioritize driver genes inside copy number altered regions[40].

Integrative network-based tools incorporate biological networks to infer relationships among genes and link different types of omics data for identifying cancer drivers, especially driver genes that rarely harbor mutations. For example, PARADIGM incorporates a signed and directed biological network to construct a factor graph based on gene expression and mutations and calculates an activity value representing the probability that a given gene contribute to cancer[41]. PARADIGM-SHIFT predicts whether a particular mutation is neutral, gain-of-function, or loss-of-function for each sample by using activity scores calculated by PARADIGM. The model evaluates the degree to which a mutation disturbs the influence of upstream genes on downstream genes and then predicts genes having large perturbations as drivers[42]. Other integrative network-based tools require a simpler biological network, which does not provide directionality and interaction types but contains a larger set of genes. DriverNet calculates the impact of mutated genes on expression levels of the neighboring genes in the network by applying a greedy algorithm to identify the smallest set of mutated genes that covers the largest number of deregulated genes as driver genes[34]. OncoIMPACT evaluates the impact of mutated genes on changes in gene expression by identifying a gene module that consists of mutated genes and associated differentially expressed genes in each sample and then predicts sample-specific driver genes that comprehensively explain the deregulated genes in the sample [33]. DawnRank uses a modified PageRank algorithm to iteratively rank mutated genes according to their impact on differentially expressed genes and predicts genes with higher ranks as cancer drivers[35].

Meta-analysis based methods combine results from existing tools to predict a more comprehensive and robust list of driver genes. Cancer genes can have different signals of positive selection due to the functional diversity of their products. For example, the gene RB1 is identified by detecting significantly high mutation frequency but could not be detected by assessing a regional clustering of mutations, while the opposite is true for the gene HRAS[43]. These examples suggest that combining tools detecting different signals can increase the sensitivity of driver prediction as well as reduce the number of false positives. For instance, MutSig combines the p-values of MutSigCV and MutSigCL that identify alterations having a higher degree of positional clustering than expected, with MutSigFN that detects the accumulation of alterations at locations of higher conservation relative to other sites in the gene[44]. OncodriveFM computes functional impact scores from SIFT, PolyPhen and Mutation Assessor, and then combines the scores using Fisher's exact test[45]. Another way to combine results from different types of methods is to use a quasi-majority vote approach to prioritize the predicted genes. For example, the Integrative Onco Genomics (IntOGen) database allows users to visualize and search for cancer drivers predicted by a specific number of tools[46]. DriverDB is another database that provides cancer drivers predicted by different types of methods including frequency-based, network-based and functional impact-based methods[47]. Moreover, several machine-learning algorithms, including a random forest classifier[48], support vector machine[49] and multi-kernel learning[50], can be used to combine scores and p-values obtained from individual tools. For instance, Liu *et al.* combined the p-values and scores from multiple tools using an ensemble classifier [51].

2.2 ConsensusDriver

In this second part, we investigate a consensus approach to combine the strengths of different types of cancer driver prediction methods discussed in the first part. Motivated by the fact that driver prediction models are based on various assumptions, we have developed ConsensusDriver that allows us to combine the orthogonal strengths from 18 driver prediction methods on more than 3,400 tumor samples. ConsensusDriver uses a rank aggregation-based approach to systematically select a subset of methods from different classes to obtain a list of high-quality driver genes. We used a cancer gene gold-standard list compiled from several sources for evaluation. ConsensusDriver outperformed those individual methods and other meta-analysis tools for both cohort and patient-specific levels.

Besides the gold-standard gene list, we constructed a list of actionable driver genes that are targets of anticancer drugs and could be decision support in precision medicine. We analyzed the top 5 driver genes predicted from different methods and observed significant variability to predict actionable driver genes, highlighting differences in underlying models of various methods. We also found that methods that incorporate information of gene expression dysregulation predict actionable driver genes in a considerably higher number of patients. Still, low percentage of the patients with actionable genes was observed, suggesting that predicting actionable cancer drivers is a challenging problem.

2.3 Characterization of cancer drivers and oncogenic processes

The Cancer Genome Atlas Research Network (TCGA Research Network) has conducted the Multi-Center Mutation-Calling Multi-tumor Completion (MC3) network consisting of over 11,000 tumors from 33 cancer types. To study characteristic of cancer driver genes and understand impact of the driver genes

on oncogenic process, we jointly worked with The TCGA Research Network on two papers: 1) *"Comprehensive characterization of cancer driver genes and mutations,"* for which we applied our driver prediction pipeline to predict cancer driver genes for a subset of computational tools [5] and 2) "*Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics*" for which we examined impacts of tumor genome on transcriptome that can reveal disrupted biological processes [6].

The TCGA Research Network gathered multiple research groups across different countries to predict cancer driver genes, which could serve as new drug targets and help us to understand the mechanisms of each cancer type. Different classes of driver prediction tools output different sets of driver genes, highlighting that the problem is challenging and driver genes can cause cancers in various ways. Next, the outputs of different categories of cancer driver prediction tools were combined to obtain a final driver consensus list consisting of 299 genes, which consists of computationally predicted genes and driver genes obtained from manual curation of the literature. The 299 cancer genes, as well as mutations positioning within the genes, have been extensively analyzed in several aspects. For example, missense driver mutations occur more frequent in oncogene than tumor suppressor genes, while truncations or frameshifts occur more often in tumor suppressor genes. Also, therapeutic implications of the predicted driver genes were assessed based on databases of known drug biomarkers and more than half of the samples harbored at least one actionable driver genes. However, the gene list is limited by consist of mutations and small indels without other types of aberrations. Many important issues still need to be solved such as moving beyond the effect of a single gene and integrative analysis that takes into account multiple types of omic-profiles.

Subsequently, the TCGA Research Network summarize and expand the findings of the TCGA PanCancer Atlas Projects in a manuscript "*Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics*"[6]. The paper focuses on three aspects of oncogenic processes: 1) somatic and germline variants and their impacts in the tumors; 2) the impact of the tumor genome and epigenome on transcriptome and proteome; and 3) the relationships between tumor and microenvironments, and. Our contribution for this paper is for the second aspect, where we investigated the impacts of driver mutations on transcriptomes and discovered relationships between oncogenic processes and cancer types through the driver consensus list.

For the first aspect, analyzing the interaction between somatic and germline drivers revealed that germline variants are usually from genes involved in maintaining genomic stability, while somatic alterations are involved in cell cycle, metabolism, signaling, and transcriptional/translational regulation.

The second aspect studies cis-effects of driver mutations and mutation types on transcriptome indicated clear upregulation of cancer driver genes affected by missense mutations and downregulation affected by nonsense or frameshift mutations. Additionally, we used OncoIMPACT to investigate the effect of driver mutations on transcriptome by integrating protein interaction, transcriptomic, and mutation information. We found that driver mutations often affect the change of gene expression levels of interacting genes and genes in the same biological pathways that are general tumorigenic processes and are frequently deregulated across cancer types. We also observed associations between oncogenic processes and cancer types, as well as known pairs of mutually exclusive mutated genes, suggesting that multiple drivers are functionally interchangeable in particular contexts.

To understand tumor microenvironment, relationships between tumor cells and immune cells were studied. The analysis revealed relationships between lymphocytic infiltrate degrees measured by gene expression and feature extracted from imaging data. Also, it has been found that mutational diver genes may affect the transcriptional regulation that guides immune response, which can affect response to immunotherapy treatment.

In this chapter, we explored different types of cancer driver prediction methods and combined their orthogonal strengths to obtain a consensus list of cancer driver genes. Next, we studied the characteristics of cancer drivers including both oncogenes and tumor suppressor genes. We then investigated the impacts of cancer driver mutations on transcriptomic information of oncogenic processes, associations among driver genes, and relationships between oncogenic processes and cancer types. We learned that driver genes could serve as a drug target, which is useful for drug discovery, and gene expression could capture dysregulation of cancer-related biological processes, mutation effects, and tumor microenvironments. However, there are still remaining challenges in cancer treatment and one of the key challenges is the variability of drug response levels which could be different across patients. Cancer driver genes might serve as drug targets or biomarkers, while transcriptomic profiles of oncogenic pathways could be used for predicting patient-specific drug response. In **Chapter 3**, we discuss the challenge of drug response prediction and different types of computational models for predicting patient-specific drug responses.

Chapter 3

Drug response prediction

3.1 Cancer precision medicine

Precision medicine is about prevention and treatment that take into account individual variability [52]. In cancer precision medicine, the goal is to give *the right treatment for the right patient at the right time* based on multi-omic profiles to improve effectiveness and reduce the side effects of the treatment. The genomic and transcriptomic profiles can help us to understand drug response heterogeneity across patients [7,8] highlighting the need of computational models to learn patterns of the profiles to understand inter-patient drug response heterogeneity. To study relationships between genomic/transcriptomic profiles and drug response behaviors, international efforts such as CCLE[7], CTRP[8], GDSC[9], and NCI60[53] have generated datasets of drug-screening on several hundreds of cancer cell lines. A drug-screening experiment measures viabilities of a cell line for each drug at multiple concentrations and then a dose-response curve is fitted. Based on the dose-response curve, summary scores such as IC_{50} (the concentration of a drug where the cell viability is reduced by 50%), EC_{50} (the concentration of a drug that gives a half-maximal response), and AUC (the area under the dose-response curve) are calculated. The availability of both drug

response data and multi-omic profiles allow us to develop a computational model to predict patient-specific drug responses.

3.2 Computational models for drug response prediction

Several computational models have been proposed for predicting drug responses using transcriptomic, genomic, or other omic-features[11,54–56]. We categorized the existing methods into two groups, drug-specific models, and models based on collaborative filtering techniques. We discuss the underlying assumption of the models, as well as their strengths and weakness.

A drug-specific model for predicting drug response

A common strategy to predict drug response is to construct a model for each drug independently. With this strategy, the model learns patterns of transcriptomic/genomic that explain drug response heterogeneity of a given drug across cell lines and identify genes explaining drug responses. Many drug-specific models are based on gene expression, while few models are based on genomic information. For example, both CCLE and GDSC papers presented drug response prediction based on gene expression using an ElasticNet model. Additionally, for the GDSC, a logic model was trained based on genomic information to predict a combination of mutated genes that can explain drug response behaviors. Other drug-specific models include linear regression models based on gene expression[7,9,10] or based on a combination of gene expression and other genomic information such as copy number alterations and DNA methylation[57,58]. Nonlinear models include neural networks, random forests, support vector machines, kernel regression based on multiple types of genomic information[59–61], and a neural network model that also incorporates drug property information[62].

Although the drug-specific models showed good predictive performances, their generalizability is limited by a small number of cell lines tested for each drug. To increase the number of data points and obtain a more robust model, a Bayesian multitask multiple kernel learning (BMTMKL) approach has been proposed[11]. The BMTMKL jointly learned across multiple drugs and achieved the best performance in the DREAM challenge for drug response prediction. This promising result highlighted the usefulness of sharing information across drugs in improving the accuracy of drug response prediction. However, one limitation of multitask learning is that it learns parameters that are shared across all drugs, preventing us from learning parameters specific for each drug or different classes of drugs. Also, a multitask learning requires normalization of drug response values, causing a loss of drug ranking information for a given cell line. Consequently, the predicted drug response might not be suitable to use in cancer precision medicine, where we need a decision support system to suggest drugs for a given patient based on their molecular profiles.

Collaborative filtering based models

To predict the drug response of a given pair of drug and cell line, multitask learning assigns all drugs equal importance. However, it could be more meaningful to prioritize information from similar drugs, as is possible using collaborative filtering techniques. In other words, collaborative filtering techniques have an ability to predict drug response based on information from a subset of training cell lines (drugs) that have similar properties as a predicting cell line (drug). In the area of recommender systems, collaborative filtering is a framework to analyze relationships between users (cell-lines/patients) and dependencies among items (drugs) to identify new user-item associations (patient-specific drug response)[63]. The two major classes of collaborative filtering

techniques are (i) neighborhood methods, which predict the user-item association based on predefined user-user and item-item similarities, and (ii) latent factor models, which use matrix factorization to identify a latent space that captures user-item associations. Matrix factorization techniques, in particular, have shown promising results in the Netflix Prize, a competition for collaborative filtering methods to predict user ratings for movies based on a rating history[64].

The collaborative filtering techniques have been used for predicting patient-specific drug responses. Based on a neighborhood method, Sheng *et al.*[65] define drug-specific cell line similarity and drug structural similarity, then predict unobserved drug responses by calculating a weighted average of observed drug responses according to both drug and cell line similarity. However, the model is solely based on the assumption that the predefined similarities can explain drug responses and it does not take into account the observed drug response information to define drug similarity. For the matrix factorization approach, Khan *et al.*[12] construct component-wise kernelized Bayesian matrix factorization (cwKBMF) model to predict drug responses based on multiple cell line kernels. The model can identify drug-pathway associations and outperformed BMTMKL. However, both cwKBMF and BMTMKL require per-drug normalization of drug response values and this preprocessing step leads to a loss of information on relative ranking of drugs within each cell line. Also, we note that Wang *et al.* have proposed a matrix factorization model based on cell line and drug similarities (SRMF) that could outperform cwKBMF, but the model does not tailor for predicting drug response of unseen samples[66].

Overall, the small number of cell lines tested with each drug lead to a challenge of learning robust models that provide meaningful predictions in a new dataset. Some models do not have an ability to predict drug response for an

unseen patient/cell line or are not suitable for predicting drug response based on transcriptomic profiles measured from different technologies such as microarray and RNA sequencing (RNA-seq). Additionally, the interpretability of models and their usage to obtain biological insights has not been extensively explored in the field. In the next chapter, we proposed CaDDReS, a model based on matrix factorization technique to predict patient-specific drug response, address the limitations of existing models, as well as provide biological insights into drug response mechanisms. Although we mainly used cancer cell line datasets to develop and test our models, for the sake of completeness, we discuss other data sources that provide tumor and normal tissues information and existing studies that analyze these data sources.

3.3 Incorporating multi-omic profiles of tumors

Drug response prediction models are typically developed and tested on only cancer cell line datasets containing a thousand cell lines, while there are several datasets such as TCAG and ICGC that provide several thousands of tumors and normal tissues (**Chapter 2**). Although drug response information is not available for those samples, it could be useful to combine both tumor and cell line datasets. For example, Allen *et al.* constructed a database of drug-related tumor alterations for genomics-driven therapy (TARGET)[67]. They also proposed an algorithm, precision heuristics for interpreting the alterations landscape (PHIAL), to rank alterations observed in whole-exome sequencing (WES) studies based on TARGET database, known cancer genes, pathway information, mutation types, and gene expression. PHIAL can identify and prioritize clinically relevant alterations including both in standard-of-care and unexpected clinically relevant alterations. To assign clinical relevance to WES data, Ghazani *et al.* proposed a framework based on PHIAL for assessing somatic and germ-line variants detected

in patients, as well as a computational framework for annotating clinical relevant that has superior scalability comparing to a tumor expert board[68].

Besides methods for ranking actionable alterations based on the curated database, Lorio *et al.* integratively analyzed both human tumor, consisting of >6,800 samples, and cancer cell line datasets [9]. Based on the tumor dataset, they identified Cancer functional events (CFEs) comprising of: 1) predicted cancer driver genes based on multiple cancer driver prediction tools, 2) recurrently aberrant copy number segments, and 3) hypermethylated sites. They showed that a large panel of cell lines could cover CFEs identified in patients and then applied different computational models to determine the power of CFEs for drug sensitivity prediction on the cell line dataset. They constructed ANOVA analysis and LOBICO (a logic model) to identify drug response-related single CFEs and combination of CFEs, respectively. Finally, they measured the contributions of different types of omics data and observed that for pan-cancer analysis gene expression performed the best, while in the cancer-specific analysis the best performing models are based on genomic features.

Geeleher *et al.* recently proposed an imputed drug-wise association study (IDWAS), a framework for discovering pharmacogenomic biomarkers in the human tumor using a model learned from transcriptomic profiles of cancer cell lines[69]. With the ten times larger number of samples in tumor dataset compared to the cell line dataset, the framework could improve the detection of clinically actionable somatic alterations and identify new biomarkers, highlighting the usefulness of analyzing both cancer cell lines and human tumor datasets.

In this work, we mainly used cancer cell line datasets to construct and evaluate our models because they provide both drug response and molecular information of the samples (**Chapter 4-5**). However, we also analyzed TCGA

dataset that contains information about tumors, survivorship, and clinical drug response of the corresponding patients to study *intra-patient* heterogeneities (**Chapter 6**).

3.4 Toward cancer precision medicine

Several drug response prediction models have been proposed and showed reasonable performances. In most studies, the evaluations were usually done per drug, i.e., calculating a correlation between predicted and observed drug response values for each drug. However, in precision medicine, we need a decision support system that can prioritize drugs for a patient or a patient-derived cell line based on the multi-omic profile. Therefore, a predictive performance of each cell line should also be evaluated, especially the performance of unseen samples.

Interpretability is another aspect that has not been extensively explored. For the drug-specific models, roles of each gene in predicting drug response were not typically discussed. Also, in **Chapter 4**, we showed that the sets of genes that were selected to predict drug response were not robust due to a small number of cell lines tested for a drug, preventing us from identifying genes to study different drug response behaviors across patients. For matrix factorization-based methods, they learn a latent space that captures interactions among cell lines and drugs to predict drug response, but the latent spaces were not yet fully explored.

Normalization of predicting drug response values in many studies could lead to loss of information about drug ranking within a cell line. Ideally, we need a model that can predict an accurate patient-specific dosage, so normalizing the drug response values might prevent us from transferring knowledge learned from a model to a clinic. Besides the predicting drug response value, sample features such as gene expression might also need to be normalized or transformed.

Directly using gene expression values might not appropriate for applying a model learned on gene expression measured from microarray (CCLE and GDSC datasets) to predict samples with gene expression measured from RNA-seq. A transformed cell line feature should allow us to apply a model to different samples with molecular profiles measured from various technologies.

Finally, predicting drug response for a patient based on bulk gene expression of tumor might not accurately represent the real drug response behavior due to intra-patient heterogeneity. A tumor sample can consist of various cell types that respond to the same drug differently. Recently, single-cell RNA sequencing technologies have been introduced, and they allow us to measure gene expression of individual cells within a tumor. Therefore, an ability to predict drug response based on the gene expression of unseen samples – including cancer cell lines, tumors, and cells – is essential to study intratumoral drug response heterogeneity.

Chapter 4

Cancer DRug Response prediction using Recommender System (CaDRReS)

4.1 What is CaDRReS

We proposed Cancer Drug Response prediction using a Recommender System (CaDRReS) to predict drug response in cancer cell lines based on transcriptomic profiles[70]. The term 'recommender system' refers to a method for identifying relationships among users (patients) and items (drugs) for predicting patient-specific drug response. Inside CaDRReS, a matrix factorization technique is used for learning a latent *pharmacogenomic space*, which consists of drug and cell line vectors, across multiple drugs and cell types (**Figure 4.1A** and **4.1B** center). The *pharmacogenomic space* captures interactions between drugs and the genomic background of cell lines such that the dot product between a cell line vector and a drug vector ($\boldsymbol{p} \cdot \boldsymbol{q}$) represents the interaction between the drug and the cell line. As shown in **Figure 4.1B** (center), cell line u is sensitive to drug i and

drug j while not being sensitive to drug k. Similarly, cell line v and cell line u respond to drugs i and j differently.

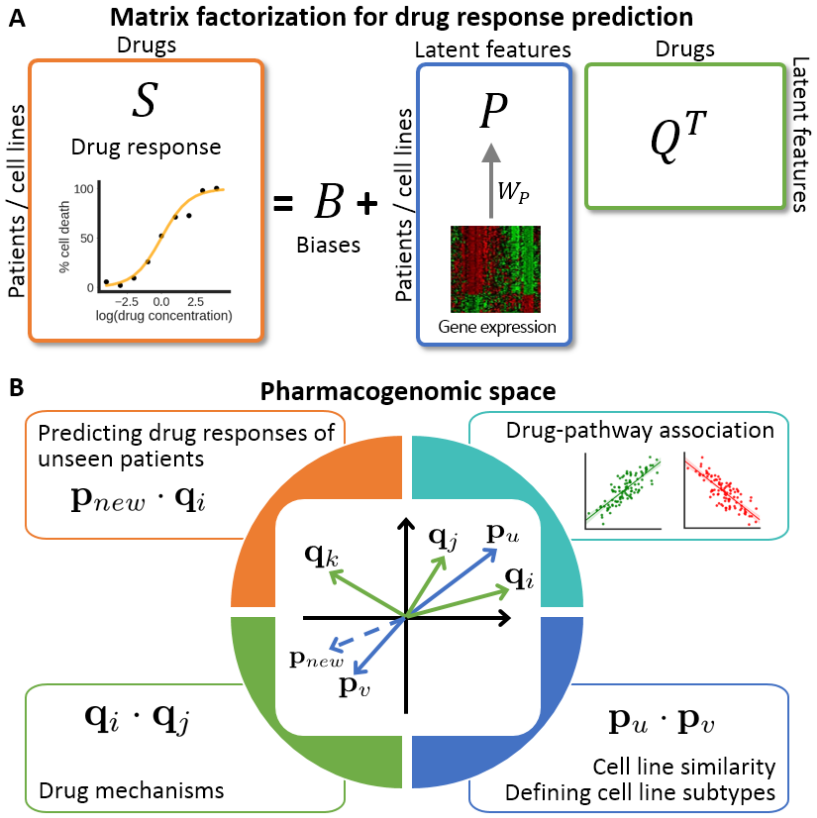

Figure 4.1 An overview of CaDRReS framework. (A) Schematic depicting the relationship between the drug response matrix S, the bias terms and factorized matrices for cell lines and drugs. A transformation matrix (W_P) is used for projecting cell lines onto the latent space. (B) The pharmacogenomic latent space captures interactions between drugs and cell lines and thus enables the study of drug-pathway associations, drug mechanism similarity, and cell line sub-types as discussed in later sections.

CaDRReS learns a matrix that transforms cell line transcriptomic features into cell line vectors on the pharmacogenomic space. The transformation matrix allows us to project unseen samples such as patient-derived cell lines, onto the pharmacogenomic space to predict drug response. Additionally, CaDRReS does not require normalization of drug response values and the objective function is to minimize the difference between predicted and observed drug response values, allowing us to predict exact drug response values that might be used in the following experimental validation. Moreover, to enable CaDRReS to work across

different datasets generated based on different transcriptomic measurement technologies, we construct a kernel feature that captures similarity between cell lines instead of directly using of gene expression values for drug response prediction.

As a result, the overall predictive performance of CaDRReS is among the best when we benchmark it with state-of-the-art methods (**Section 4.5**). Besides predicting drug response, the representation of cell lines and drugs on the pharmacogenomic space has many applications including (i) predicting drug responses of unseen samples (cell lines or patients), (ii) revealing drug mechanisms and (iii) subtypes of cell lines, and (iv) identifying drug-pathway associations (**Figure 4.1B** and **Chapter 5**).

4.2 Matrix factorization for drug response prediction

The process of matrix factorization for drug response prediction can be depicted as drug response matrix ($\boldsymbol{S}$) being factorized into biases ($\boldsymbol{B}$) and matrices of cell lines ($\boldsymbol{P}$) and drugs ($\boldsymbol{Q}$). Rows of the cell line matrix ($\boldsymbol{P}$) and the drug matrix ($\boldsymbol{Q}$) are vectors of cell lines and drugs in a latent space, respectively (**Figure 4.1A**). There are few other studies using matrix factorization techniques, but their models still have limitations as reviewed in **Section 3.2**. In particular, cwKBMF requires per-drug normalization, which leads to loss of relative ranking of drugs within a cell line, while SRMF does not provide a transformation matrix to facilitate prediction of unseen samples. To address both limitations, we proposed CaDRReS based on the following objective function:

Equation 1.

$$\begin{aligned}\hat{s}_{ui} &= \mu + b_i^{Q} + b_u^{P} + \boldsymbol{q}_i \cdot \boldsymbol{p}_u \\ &= \mu + b_i^{Q} + b_u^{P} + \boldsymbol{q}_i(\boldsymbol{x}_u \boldsymbol{W}_P)^T\end{aligned}$$

A drug sensitivity score used for training the model is equal to $-log(IC_{50})$, where the higher of the value, the more a cell line sensitive to a drug. In the objective function, $\hat{s}_{ui}$ is the predicted sensitivity score of cell line u to drug i, μ is the overall drug response, b_i^{Q} and b_u^{P} are bias terms for drug i and cell line u, respectively, $\boldsymbol{q}_i, \boldsymbol{p}_u \in \mathbb{R}^f$ are vectors for drug i and cell line u in the f-dimensional latent space and $\boldsymbol{W}_P \in \mathbb{R}^{d \times f}$ is a transformation matrix that projects cell line features $\boldsymbol{x}_u \in \mathbb{R}^d$ onto the latent space. We note that introducing drug and cell line bias terms allows us to train a model without per-drug normalization step and learning $\boldsymbol{W}_P$ allows us to project unseen samples onto the pharmacogenomic space. The value of f was set at 10 for both CCLE and GDSC datasets based on cross-validation performance.

4.3 Datasets and preprocessing

Drug-screening data for cancer cell lines were obtained from two large-scale studies, CCLE and GDSC, and all cell lines with baseline gene expression data were retained. Firstly, a Bayesian sigmoid curve fitting approach was applied to raw intensity data at different drug dosages to estimate IC_{50} (minimal concentration that induces 50% cell death) values that were more comparable across datasets. Each cell line in CCLE and GDSC was tested with each drug at 8 and 9 different concentrations, respectively. At each concentration, an activity value was calculated as follow:

$$Activity\ value = \frac{-(test - positive)}{(negative - positive)}$$

where $test$, $positive$ and $negative$ are intensity values (cell counts) measured from a well with the tested drug, positive control, and negative control, respectively.

Next, we applied Bayesian sigmoid curve fitting using JAGS [71] as described by Kruschke[72] to fit a dose-response curve for each pair of drug and cell line. Briefly, defining i as the cell line index, j as the drug index, x_{ij} as the log-scale dosage, and y_{ij} as the response for a given pair of drug and cell line, modelled as a Student's t distribution with the parameter β_{0_j} representing the center of the dose-response curve (i.e. IC_{50}) and β_{1_j} capturing the steepness of the curve, we get:

$$y_{ij} = \tau\left(\frac{1}{1+2^{(\beta_{0j}-x_{ij})\beta_{1j}}}, \frac{1}{\sigma^2}, \nu\right)$$

The two parameters of the drug-response curve β_{0_j} and $\beta_{1_j}$are assumed to follow Normal and Gamma distributions, respectively.

$$\beta_{0_j} \sim \mathcal{N}\left(\mu_{\beta_0}, \frac{1}{\sigma^2_{\beta_0}}\right)$$
$$\beta_{1_j} \sim \mathcal{G}\left(\lambda_{\beta_1}, r_{\beta_1}\right)$$

For the remaining parameters, non-informative priors were used as described by Kruschke[72]. Finally, we calculated log (IC_{50}) based on the fitted sigmoid curves, where IC_{50} is a concentration that inhibits 50% of the cells [73]. We observed that the newly calculated IC_{50} values from raw dose-response data show higher Spearman correlation (for each drug) between CCLE and GDSC datasets than the provided IC_{50} values (paired t-test; *p-value* < 0.01, **Figure 4.2A-B**). The re-estimated IC_{50} values were used for all methods and analyses in this thesis. We note that in the training step of CaDRReS to obtain a pharmacogenomic space we defined a drug sensitivity score as – $log(IC_{50})$, so the higher score the more drug

sensitivity, corresponding to higher values of the dot product between cell line and drug vectors.

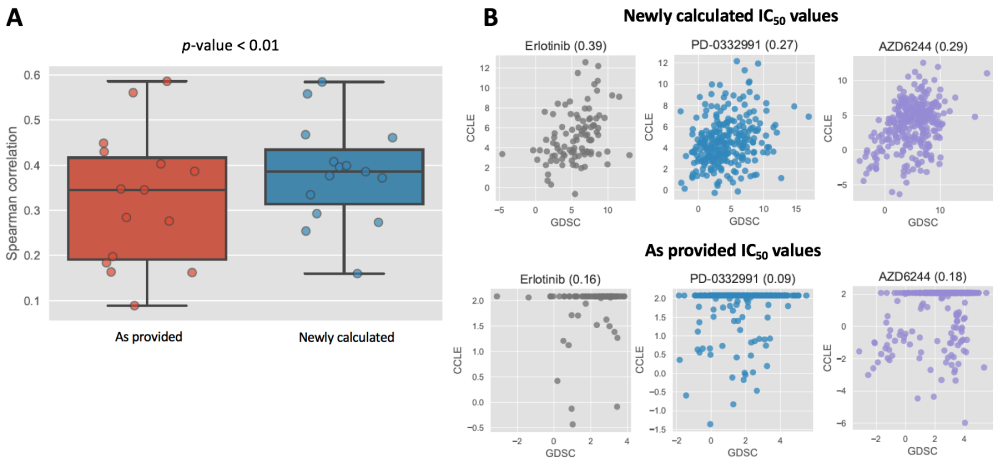

Figure 4.2 Estimating IC_{50} values. (A) Newly calculated IC_{50} values from raw dose-response data show higher Spearman correlation (for each drug) between CCLE and GDSC datasets than the provided IC_{50} values (paired t-test; *p-value* < 0.01). (B) Examples of Spearman correlation between CCLE and GDSC datasets for each drug.

Drugs with median IC_{50} less than 1 μM tend to be cytotoxic drugs with consistently high toxicity across cell lines (**Figure 4.3A-B**). Correspondingly, they make the drug response prediction problem easier, and so we excluded them to focus our efforts on predicting response for targeted cancer drugs. Our final dataset contained 491 cell lines, 19 drugs, and 9,096 experiments from CCLE, and 983 cell lines, 223 drugs, and 179,633 experiments from GDSC, providing a large dataset for training and validation of our models.

We also obtained an in-house dataset based on screening of 276 drugs (65 of which overlap with GDSC) on 8 head and neck cancer (HNC) patient-derived cell lines from 5 subjects[74]. Two of the cell lines were found to be not sensitive to any of the overlapping drugs (inhibition score <50 at $1\mu M$), while one was found to be sensitive to more than 25% of the overlapping drugs. These three cell lines were excluded as the single dosage they were tested on does not seem to allow discrimination across drugs and thus appropriate evaluation of drug response

models, leaving us with 325 data points from 5 cell lines to be used as an independent dataset to evaluate predictions from different models. Additionally, transcriptomic profiles of these patient-derived cell lines were measured by using RNA-seq, while CaDRReS was trained on microarray data (CCLE and GDSC). This allowed us to evaluate the performance of the model for unseen samples across different gene expression platforms.

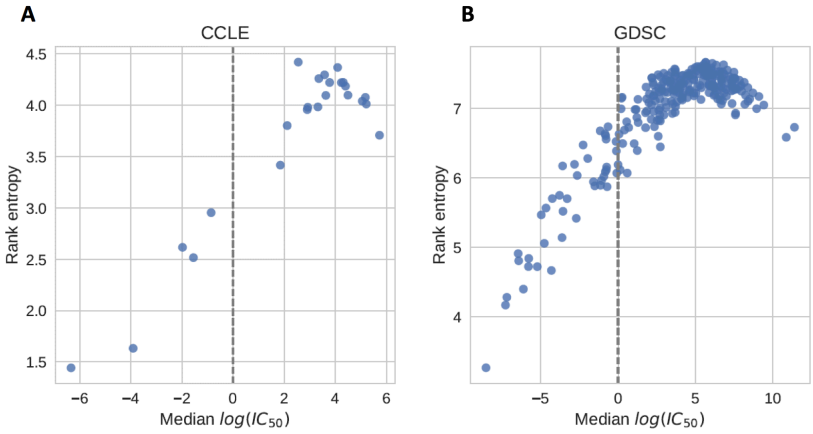

Figure 4.3 Comparison between median IC_{50} and rank entropy. (A) CCLE (B) GDSC. Drugs with median IC_{50} less than 1 μM were excluded. Drugs having low IC_{50} tend to have higher ranks across all the cell lines (low rank entropy) suggesting that they may lack of specificity and be cytotoxic drugs.

4.4 Model training

The first step in CaDRReS is to calculate cell line kernel features based on gene expression information. With the kernel feature, we can apply the model trained based on microarray gene expression on the samples' gene expression measured by RNA-seq. To do this, we normalized baseline gene expression values for each gene by computing fold-changes compared to the median value across cell lines. For the next step, since the drug response experiments in GDSC and CCLE aim to measure cell death, 1,856 essential genes identified based on large-scale CRISPR experiments[75] were selected to condense the expression information for each cell line. Pearson's correlation for every pair of cell lines was

calculated using the expression fold-changes of these essential genes. Thus, in total, we had 491, and 983 cell line features for CCLE and GDSC, respectively.

For training the model, a drug sensitivity score $s = -\log(IC_{50})$ was defined where the higher the score, the more sensitive the cell line is to the drug. Models were trained and tested independently for CCLE and GDSC to avoid biases towards either of the datasets[76,77]. According to **Equation 1**, the model was trained by optimizing the following 'sum of squared error' loss function:

Equation 2.

$$L(\theta) = \frac{1}{2|\kappa|}\sum_{u}\sum_{i} e_{ui}^2$$

$$\mathrm{e_{ui}} = \mathrm{s_{ui}} - \hat{\mathrm{s}}_{\mathrm{ui}}$$

where s_{ui} and $\hat{s}_{ui}$ are observed and predicted sensitivity scores for cell line u using drug i, respectively, $\boldsymbol{\theta} = \{b_i, b_u, \boldsymbol{W}_P, \boldsymbol{q}_i\}$, and $|\kappa|$ is the number of drug response experiments in the training dataset. Finally, we applied gradient descent to optimize this loss function and obtain all parameters in $\boldsymbol{\theta}$ based solely on the assayed drug-response values. Gradient functions for the parameters $b_i^Q, b_u^P, \boldsymbol{W}_P, \boldsymbol{W}_Q$ were calculated as follows:

$$\frac{\partial L}{\partial b_u^P} = \frac{-\sum_i e_{ui}}{|\kappa|}$$

$$\frac{\partial L}{\partial b_i^Q} = \frac{-\sum_u e_{ui}}{|\kappa|}$$

$$\frac{\partial L}{\partial \mathbf{W}_P} = \left(-\sum_i \sum_u e_{ui}\mathbf{q}_i^T\mathbf{x}_u\right)^T/|\kappa|$$

$$\frac{\partial L}{\partial \mathbf{W}_Q} = \left(-\sum_u \sum_i e_{ui}\mathbf{p}_u^T\mathbf{y}_i\right)^T/|\kappa|$$

We initialized b_i^Q, b_u^P to be zeros and $\boldsymbol{W}_P, \boldsymbol{W}_Q$ using small uniformly random numbers in the range of [-0.05, 0.05], and performed a batch gradient descent using the above-calculated gradients to optimize $b_i^Q, b_u^P, \boldsymbol{W}_P$ and $\boldsymbol{W}_Q$. Additionally, we tested CaDRReS' robustness by constructing ten different models using different random starting points for the gradient descent optimization. We observed that the models show similar performance (**Figure 4.4A**). We also compared the latent space of the models that were trained based on ten different random initializations using the CCLE dataset. Because the latent pharmacogenomic space captures the drug-drug, sample-sample, drug-sample relationships through the dot product, we calculated cosine similarity of all pairs of drug and cell line vectors, and then compared the cosine similarity values of the pharamacogenomic spaces learned based on different random initializations. High correlations of the cosine similarity were observed (**Figure 4.4B**; Pearson's correlation = [0.92, 0.97] for all 45 pairs of pharmacogenemic spaces), indicating that the pharamacogenomic spaces capture similar relationships. Taken together, these results suggest that CaDRReS is robust against random starting points.

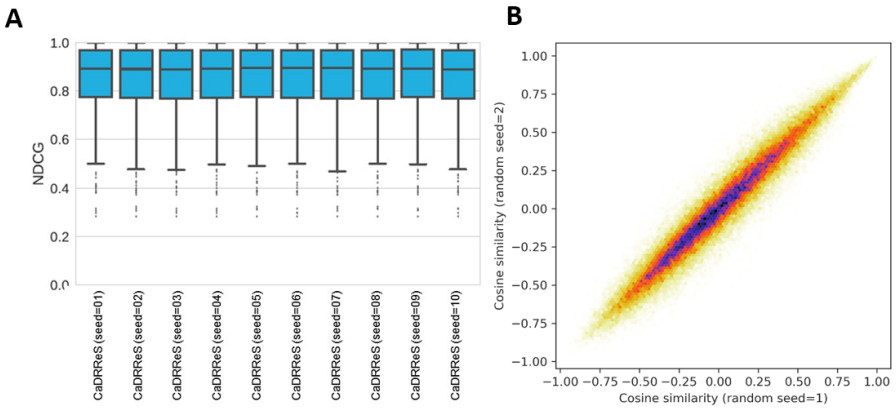

Figure 4.4. **Performance comparison of CaDRReS based on different initial values.** (A) Ten CaDRReS models trained based on different random starting points produced a similar performance per cell line. (B) A comparison of cosine similarity values of the pharamacogenomic spaces learned using different random initializations.

Finally, to prediction the drug response of unseen cell lines, we calculated the cell line kernel features, which are Pearson's correlation of the essential genes between the input cell lines and the cell lines used for training the model. Then we used the transformation matrix to project the input cell lines onto the pharmacogenomic space for drug response prediction.

4.5 Evaluations

We compared the predictive performance and robustness of CaDRReS against other existing methods including a method based on the elastic net regression model (**ElasticNet**[7,9], **cwKBMF**[12], the method from **Sheng *et al.***[65], **SRMF**[66], as well as a control method based on random permutations of the drug sensitivity scores for each cell line (**Control**). For ElasticNet, the model was trained for each drug as described previously[7,9] using the Elastic Net library from Scikit-learn[78] ($l1$-ratio = 0.5), where the model automatically selects the genes. For the method proposed by Sheng *et al.*[65], we re-implemented it as described in the paper, normalized drug response data, calculated drug similarity and drug-specific cell line similarity scores and set the parameters r_d (number of similar drugs) = 3 and r_c (number of similar cell lines) = 9 as used in the paper. For cwKBMF, drug response data were normalized for each drug as described in the paper and the provided MATLAB source code was used to train a model. For SRMF, cell line similarities were calculated as described in the paper and we set λ_d to zero because it has been shown that SRMF performed the best when drug similarity is ignored. We also set the number of dimensions to 10 as used in both cwKBMF and CaDRReS.

To benchmark the performance of drug response prediction methods, we evaluated the prediction for both drug and cell line perspectives. The drug aspect

measures an ability of the model to capture transcriptomic patterns that explain the different drug response levels for a particular drug, while cell line (sample) aspect evaluates potentiality of the model to suggest drugs for a given patient in precision oncology.

Firstly, for each drug, we calculated Spearman correlation (r_s) and reported the average correlation across drugs. The higher correlation suggests that the model can capture the mechanism of the drug. Next, to evaluate models for each cell line, the normalized discounted cumulative gain (NDCG), a widely used score for evaluating ranking recommendations, was calculated as follows:

$$NDCG(\hat{\boldsymbol{r}}, \boldsymbol{s}) = \frac{DCG(\hat{\boldsymbol{r}}, \boldsymbol{s})}{DCG(\boldsymbol{r}, \boldsymbol{s})}$$

$$DCG(\hat{\boldsymbol{r}}, \boldsymbol{s}) = \sum_i \frac{2^{s_i} - 1}{\log_2 \hat{r}_i + 1}$$

where $\hat{\boldsymbol{r}}$ is the predicted rank of drugs tested on a cell line, $\boldsymbol{s}$ is a list of observed drug sensitivity scores and $\boldsymbol{r}$ is the known ranking of drugs calculated based on the measured drug response values. NDCG ranges from 0 to 1, where 1 indicates that the model correctly predicts the ranking of drugs. The numerator in DCG is designed to give greater weight to a drug with higher sensitivity score, while the denominator gives preference to drugs predicted to have higher ranks. The higher NDCG suggests that the model can correctly suggest most effective drugs for each unseen cell line, highlighting the usefulness of the model in precision medicine.

We performed 5-fold cross-validation to evaluate the predictive performance of the models both drug and cell line perspectives. Moreover, we assessed the stability of the five models learned based on different sets of cell lines from cross-validation data. For ElasticNet, for each drug, we identified the selected genes (genes with non-zero coefficient) and counted the number of

overlapping genes across the five different models. For CaDRReS, we compared drug cosine similarity (and cell line cosine similarity) calculated on the pharmacogenomic space for the five models learned from the different set of cell lines (and drugs).

4.6 Performance and robustness

A common way to evaluate drug response prediction methods is to assess their correlation by comparing the predicted responses to known responses for each drug (across cell lines) in a cross-validation framework [7,9]. We performed ten sets of 5-fold cross-validation to measure the predictive performance of the models. Using the matrix-factorization based approaches, SRMF, CaDRReS, and cwKBMF showed significantly better performance than ElasticNet, Sheng *et al.*, as well as the Control method (*p-value* $<10^{-30}$) in both the CCLE and GDSC datasets (**Figure 4.5A** and **4.6A**).

While the ability to predict cell line responses for a given drug is useful for understanding drug efficacy and characterizing drug mechanisms, ranking drugs for a given unseen cell-line/patient may be more relevant for precision oncology applications. Based on a weighted scoring of rankings (NDCG), we noted that CaDRReS and ElasticNet exhibited similar performance and improved notably over cwKBMF, SRMF, Sheng *et al.*, and the Control method (*p-value* $<10^{-20}$; **Figure 4.5B** and **4.6B**). Taken together, these results suggest that CaDRReS improves over existing approaches in providing models that are useful for both drug response prediction across cell-lines and within a cell line.

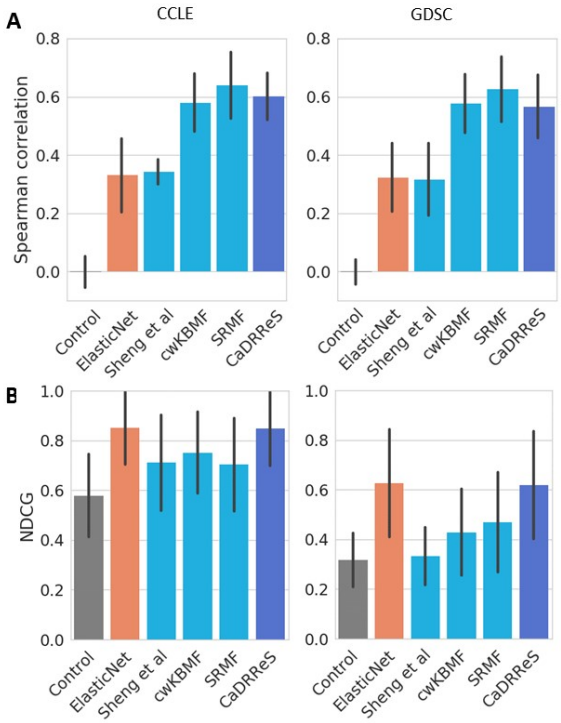

Figure 4.5 Performance and robustness of the CaDRReS model. (A) Average performance (Spearman correlation) across drugs based on 5-fold cross-validation (error bars represent 1 standard deviation). (B) Average NDCG scores across unseen cell-lines based on 5-fold cross-validation.

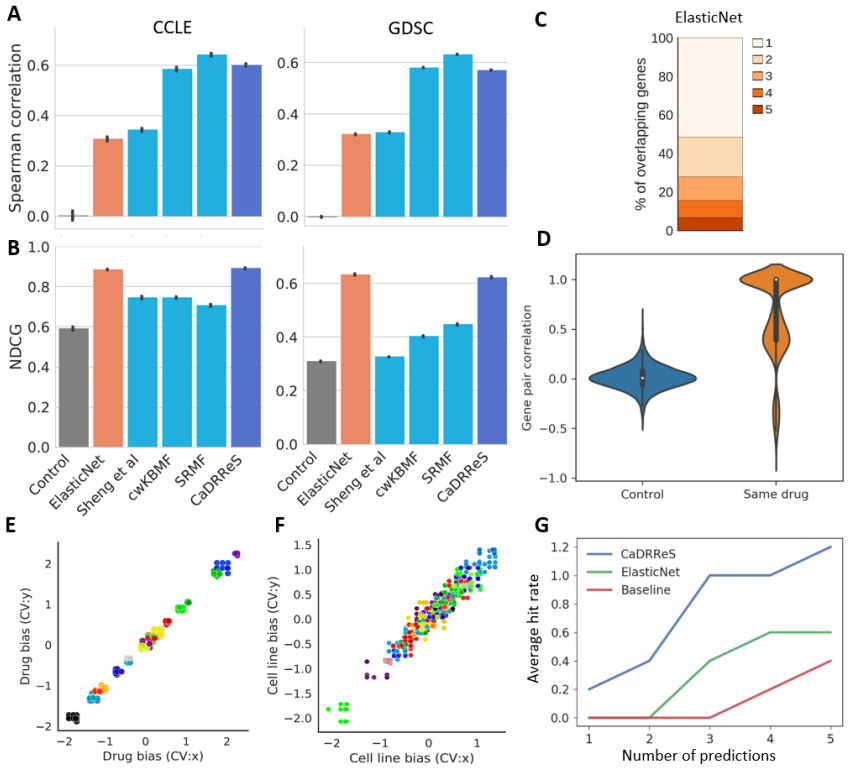

Figure 4.6 Performance and robustness of the CaDRReS model. (A) Average spearman correlation across 10 runs of 5-fold cross-validation (error bars represent 1 standard deviation). (B) Average NDCG scores across 10 runs of 5-fold cross-validation. (C) Average percentage of overlapping genes in ElasticNet across different CCLE cross-validation datasets. (D) A violin plot presents the distribution of gene expression correlations between genes identified based on every pair of ElasticNet models of the same drug, and the control is a distribution of gene expression correlations of random gene pairs. (E) Concordance between drug-specific bias terms as inferred by CaDRReS for every pair of models from the 5-fold cross-validation analysis. Each color represents a drug in the CCLE dataset. (F) Concordance between cell line bias terms as inferred by CaDRReS for every pair of models from the 5-fold cross-validation analysis. Each color represents a cell line in the CCLE dataset (first 50 cell lines). (G) Average hit rate (number of sensitive drugs identified) in the top five predictions of each method. Baseline refers to an approach that sorts drugs by their average sensitivity across cell lines.

For drug response prediction within a cell-line, although ElasticNet models were trained independently for each drug, their NDCG scores were surprisingly high. However, we suspected that there could be high variance among the models trained based on different sets of cell lines due to a limited number of cell lines for each drug. To assess this we evaluated the robustness of ElasticNet models learned across cross-validation runs and found that <10% of the selected genes were shared across folds and half of the genes were selected in only one-fold (**Figure 4.6C**). Although the number of overlapping genes were small, the expression levels of non-overlapping genes could be correlated. To investigate this, for each model pair of the same drug, we identified the best-matched gene, i.e., highest absolute correlation, for each gene in the smaller gene set (**Figure 4.6D**). We observed that approximately half of the genes were not overlapped and the expressions of the non-overlapping genes were not highly correlated. The ElasticNet models that were trained on different sets of cell lines selected different sets of genes for the same drug, limiting us from obtained a consistent interpretation from the model.

In contrast, CaDRReS showed consistently high correlation for drug biases (0.99; **Figure 4.6E**) and cosine similarity of inferred drug vectors (0.96) across cross-validation runs, as well as high correlation for cell line biases (0.96; **Figure**

4.6F) and cosine similarity of the inferred cell line vectors (0.88), highlighting the robustness of its models.

To further evaluate their performance, CaDRReS and ElasticNet models were trained on the GDSC dataset and tested on an independent dataset from patient-derived HNC cell-lines. Sheng *et al.* and cwKBMF were not included here because they require per-drug normalization of drug response values, which leads to a loss of drug ranking information within a cell line, while SRMF was excluded because it is not tailored for predicting drug response for unseen samples.

Despite having similar performance on the GDSC dataset, CaDRReS outperformed ElasticNet on this independent dataset (**Figure 4.6G**), emphasizing its ability to provide more robust and generalizable models. In particular, CaDRReS was able to identify on average at least one drug that elicited a strong response for each cell-line among its top three predictions, while a baseline method based on average response across cell lines identified none. The results also highlighted the usefulness of the cell line kernel features that allow the model to perform across different gene expression measurement platforms (microarray for GDSC and RNA-seq for the patient-derived cell lines).

Overall, we explained how to construct CaDRReS, as well as presented the benchmarking results of CaDRReS and other state-of-the-art models. We also showed in the head and neck case study that CaDRReS has an ability to predict drug response for unseen samples across different gene expression platforms. In the next chapter, we focused on the interpretability of CaDRReS by discussing applications of the pharmacogenomic space, including studying drug mechanisms and subtypes of cell lines, as well as identifying drug-pathway associations.

4.7 Discussion

We proposed CaDRReS, a model for predicting drug response based on transcriptomic profiles by using a matrix factorization technique that simultaneously learns across multiple drugs and samples to obtain a pharmacogenomic space. To train the model, we minimized the objective function, i.e., the overall error between observed and predicted IC_{50} values. However, the calculation of IC_{50} does not take into account cell growth rate, so the IC_{50} tends to be low for the cell lines with fast growth rate because the inhibition can be earlier observed[79]. Due to the lack of cell growth rate information of the cell line in the GDSC and CCLE datasets, we could not incorporate the growth rate information in our analysis. Nevertheless, with an availability of growth rate information in the future, growth rate index (GR_{50}) is a better drug response value because it corrects for the cell growth bias.

For the ElasticNet as well as other drug-specific models, the smaller number of cell lines tested for each drug might lead to overfitting problem, i.e. few hundreds cell lines (N) and ten thousand parameters for genes (G), as we observed when we applied the model to the head and neck patient-derived cell line dataset. In contrast, for CaDRReS, the number of parameters of the projection matrix $\boldsymbol{W}$ is 10N, where N the number of training cell lines and the dimension of the latent pharmacogenomic space is 10. The 10N parameters and N+D biases are trained based on ND data points, where D is the number of drugs. The larger number of data points with respect to the number of training parameters highlights the benefit of the ability to learn across multiple drugs to obtain a more robust model.

The inconsistency of CCLE and GDSC datasets have been discussed in several studies[76,80], including different types of negative and positive controls,

different post-treatment durations, and the dose-response curves were estimated by different computational methods. In this study, we obtained a raw intensity data and re-estimated the dose-response curve (Section 4.3) to unify the calculation of IC_{50} values, but disagreement of the response values was still observed. Therefore, we construct a model separately for each dataset to avoid a possible ambiguity in the performance evaluation step and the interpretation of the latent pharmacogenomic space. Nonetheless, we calculated a kernel feature to avoid direct using gene expression values, allowing the model to make a prediction based on gene expression values measured from different platforms.

To evaluate the drug response predictions, most studies focused on the predictive performance or accuracy for each drug, while the performance should also be assessed for each sample to determine the usefulness of a model in precision medicine. Therefore, in this study, we aim to measure the ability of the model for capturing transcriptomic patterns (that explain the different drug response levels for a particular drug) and suggesting drugs for a given patient in precision oncology. Therefore, we evaluated the predictions for both drug and cell line perspectives by using Spearman correlation and NDCG, respectively (described in Section 4.5). For each drug, the Spearman correlation suggests how good the model can rank cell lines by the drug sensitivity levels predicted based on transcriptomic profiles. For each cell line, NDCG suggests an ability of the model to predict top few drugs with strong observed sensitivity, while drugs at the bottom of the predicted list have lesser contributions to the score.

Additionally, in a cross-validation framework, the drug-response values are typically split into multiple folds regardless of samples, preventing us from evaluating the actual predictive performance for unseen samples. Hence, to we measure the predictive performance in different scenarios, we applied two cross-

validation schemes: 1) seen samples, where random cells in the drug response matrix (row=sample and column=drug) are held out; 2) unseen samples, where random rows in the drug response matrix are held out. The seen case measures the performance of the models for predicting the response of cell lines to a specific drug with prior knowledge of responses to other drugs, which could be useful in the case that the sample has already been tested for a subset of drugs. The unseen case measures the predictive performance for unseen samples, which correspond to the case that no prior drug response information.

Correlation between *in silico* predicted and *in vitro* observed drug responses were calculated to measure the correlation in our study as well as most of the other studies. However, the correlation does not fully capture the dosage error as the value can be high while the range of predicted and observed values are largely different, preventing us from using the predictions to determine dosage to be used in experimental validation or in a clinic.

Lastly, the concept of CaDRReS is not limited to predicting cancer drug response IC_{50} values. We could apply CaDRReS to predict other types of drug response measurements such as area under the dose-response curve, GR_{50}, and IC_{90}. Additionally, an ability to predict multiple values of the dose-response curve could be more useful than predicting a single value of drug response. The concept can also be applied to other domains such as studying epigenomic features that affect response to immunotherapy, predicting gene expression after treatment to investigate drug-resistant mechanism, and identifying anti-bacterial drugs.

Chapter 5

A pharmacogenomic space

A pharmacogenomic space allows us to study drug response mechanisms, highlighting interpretability of CaDRReS. This chapter presents applications of the pharmacogenomic space including explaining drug response mechanisms, classifying the cell lines, identifying drug similarity, and detecting drug-pathway associations. Firstly, we trained CaDRReS models on the full datasets to obtain drug and cell-line biases, as well as the pharmacogenomic spaces capturing drug-drug, cell line-cell line, and drug-cell line associations for both CCLE and GDSC (**Figure 5.1**).

To study drug mechanisms, we took vectors defined for each drug in the pharmacogenomic space, computed cosine similarities between every pair, and compared these to a commonly used drug structural similarity score (Tanimoto coefficient of SMILES calculated using the SMSD toolkit[81]). Drug cosine similarities were significantly higher for drug pairs having high structural similarities (Tanimoto coefficient > 0.3; Wilcoxon test *p-value* <0.04 for CCLE and <0.001 for GDSC), suggesting that in general, similarly structured drug pairs tend to have higher cosine similarity on the pharmacogenomic space and thus elicit similar responses (**Figure 5.2**).

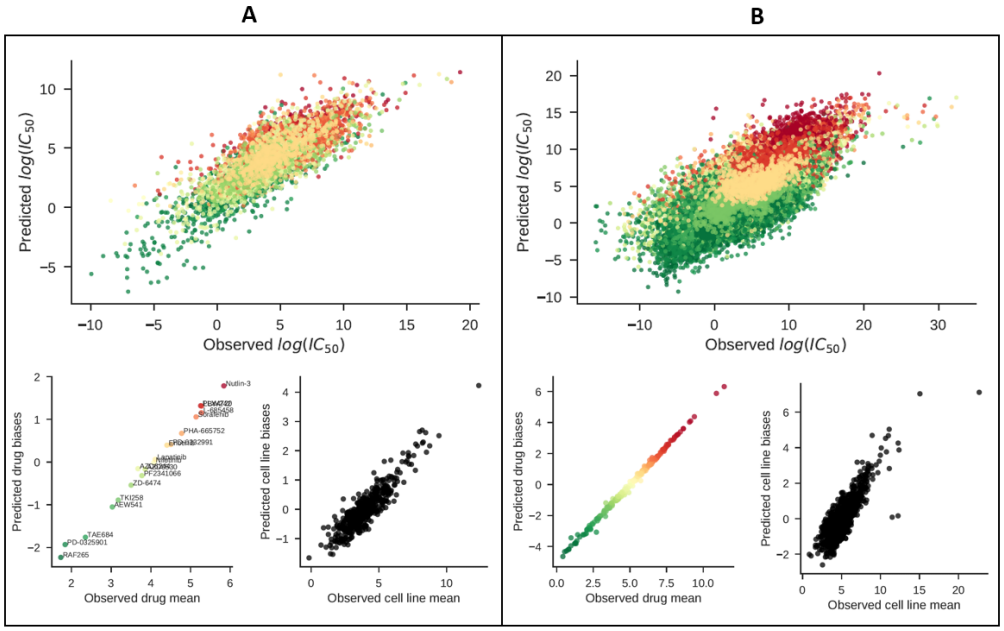

Figure 5.1 Comparison between observed and predicted IC_{50} for the full datasets. (A) CCLE (B) GDSC. Colors represent different drugs. The scatter plots show that the pharmacogenomic space correctly captured the observed drug responses, drug biases, and cell line biases for both datasets.

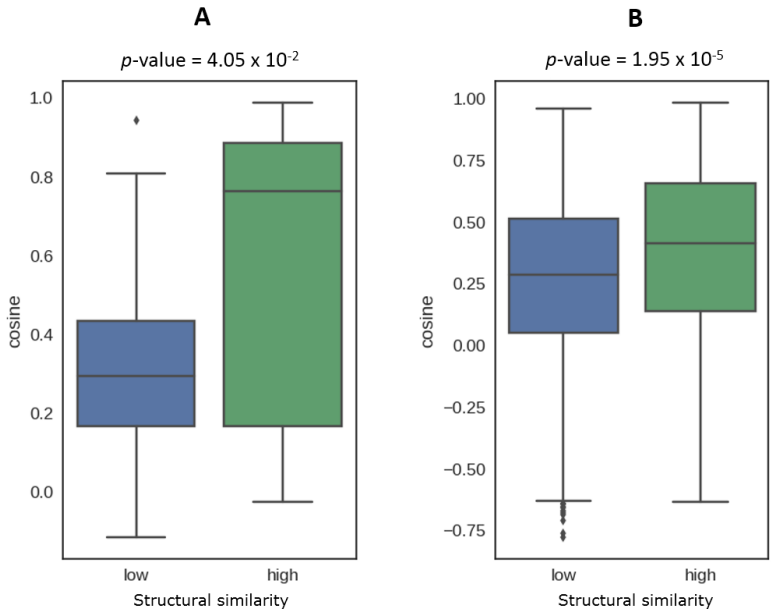

Figure 5.2 Comparison between structural similarity and cosine similarity between drugs. (A) CCLE (B) GDSC. A box-plot shows that drugs pairs with high structural similarities have significantly higher cosine similarity on the pharmacogenomic space, and thus have similar responses. x-axis represents high (>0.3) and low structural similarity and y-axis represents the cosine similarity. *p*-values were calculated based on the Wilcoxon test.

However, there are indeed exceptions to this rule where drugs that elicit a similar response profile have significantly different chemical structures. For instance, PD-0332991 and PHA-665752 have a relatively low structural similarity

(Tanimoto coefficient = 0.07), but high correlation for the observed drug responses (0.51 with *p-value* < 10^{-29}). This is likely due to the fact that PD-0332991 is a CDK4/6 inhibitor that can reduce RB phosphorylation[82], while PHA-665752 can inhibit c-MET and thus result in reduced phosphorylation of RB downstream [83]. Therefore drug similarity in the pharmacogenomic space has the potential to capture deeper similarities in drug response mechanisms beyond those observed purely based on drug structural similarity.

5.1 A pharmacogenomic space capturing drug response mechanisms

In the pharmacogenomic space, we observed that clusters of drugs frequently represent groups that target the same gene or pathway (**Figure 5.3A**). For example, EGFR inhibitors (Lapatinib, ZD-6474, AZD0530, Erlotinib), RAF inhibitors (RAF265, PLX4720) and MEK inhibitors (PD-0325901, AZD6244) in CCLE formed separate clusters based on cosine similarity. In addition, cosine similarities among the five MEK1 inhibitors in GDSC (CI-1040, PD-0325901, RDEA119, Trametinib, and selumetinib) were significantly higher than between MEK1 inhibitors and other drugs (*p-value* $<10^{-15}$). A similar trend was also observed for the four BRAF inhibitors, AZ628, Dabrafenib, PLX4720, and SB590885 (*p-value* $<10^{-7}$; **Figure 5.3B**). These observations are interesting given that CaDRReS was trained based solely on drug response data, without any other information on drug properties.

By examining dimensions of the pharmacogenomic space, we observed that each dimension captured different aspects of sensitivity to various drug classes (**Figure 5.3C**). For example, EGFR inhibitors dominated in the 5th and 9th dimensions and thus cell lines that were projected close to the positive sides of

these dimensions have higher EGFR inhibitor sensitivity. Additionally, we observed that MEK inhibitors lie on the negative side of the 8th dimension and the values of cell line vectors in this dimension were most positively correlated with activity scores for the EIF2 pathway (0.217), indicating that cell-lines with inactivated EIF2 pathway may be more sensitive to MEK inhibitors. This observation is in agreement with prior work showing that MEK inhibitors work by inducing activation of eIF-2B, which results in a shutdown of cellular protein synthesis and leads to apoptosis[84,85]. These results highlight the utility of the pharmacogenomic space learned by CaDRReS for capturing interpretable information related to drug mechanisms and pathways.

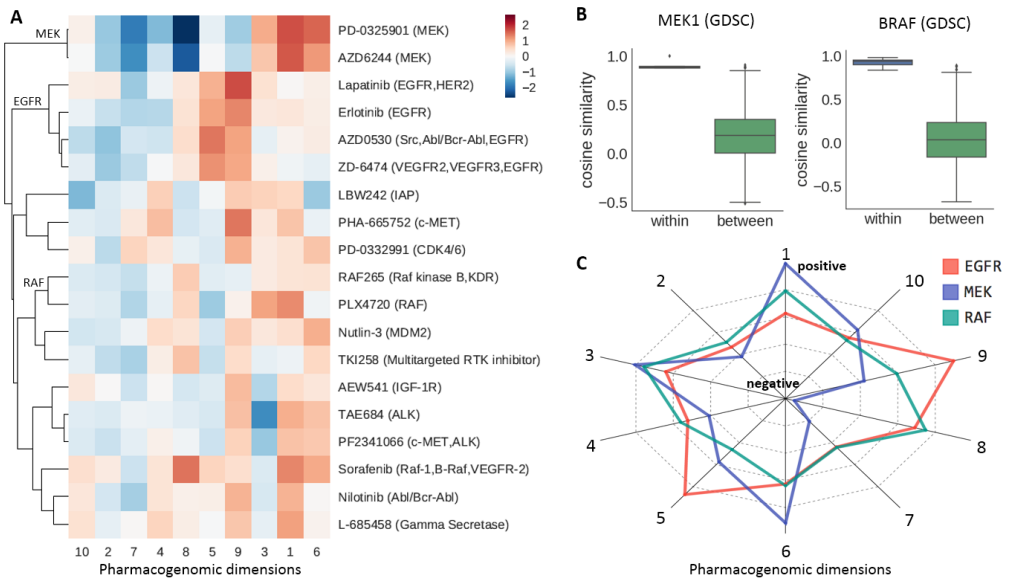

Figure 5.3 Clustering of drugs on the pharmacogenomic space and its relation to mechanism-of-action. (A) Heatmap presenting average linkage hierarchical clustering of drugs based on cosine similarity on the pharmacogenomic space (CCLE). (B) Distribution of within- and between-group cosine similarities of drugs targeting MEK1 (GDSC) and BRAF (GDSC). (C) Representation of dimensions of the pharmacogenomic space capturing different drug mechanisms. For each target, the average vector of the corresponding drugs was calculated for EGFR, RAF, and MEK inhibitors (CCLE).

5.2 Cell line subtypes in the pharmacogenomic space

Clusters of cell-lines in the pharmacogenomic space should in-principle be tuned to capture drug response similarities. However, not surprisingly we found

that they also capture tissue type signatures, with cell-lines from the same tissue type showing significantly higher cosine similarity than cell-lines from different tissue types (**Figure 5.4A**, **5.5A**), and also being visually distinct in t-SNE[86] 2D space (**Figure 5.4B**, **5.5B**). Further segregation into histological subtypes was not always as clear, though most small cell lung carcinoma (SCLC) cell-lines were distinct from non-small cell lung carcinoma (NSCLC) cell lines (except for NSCLC carcinoid cell-lines; **Figure 5.4C**). The placement of NSCLC carcinoid cell-lines with SCLC cell-lines is clearly reflected in their drug-response profiles: e.g., while NSCLC cell-lines were typically sensitive to PD-0325901 (MEK inhibitor), carcinoid cell-lines were not (**Figure 5.6**). Also, we found that cell lines with KRAS mutations had significantly higher predicted PD-0325901 sensitivity (adjusted *p-value* $<1.4 \times 10^{-8}$), and that KRAS mutations were common in NSCLC cell lines (~30%) but not seen often in SCLC or carcinoid cell-lines (~3%), in agreement with prior work on KRAS mutations being activation biomarkers for MEK inhibitors[87].

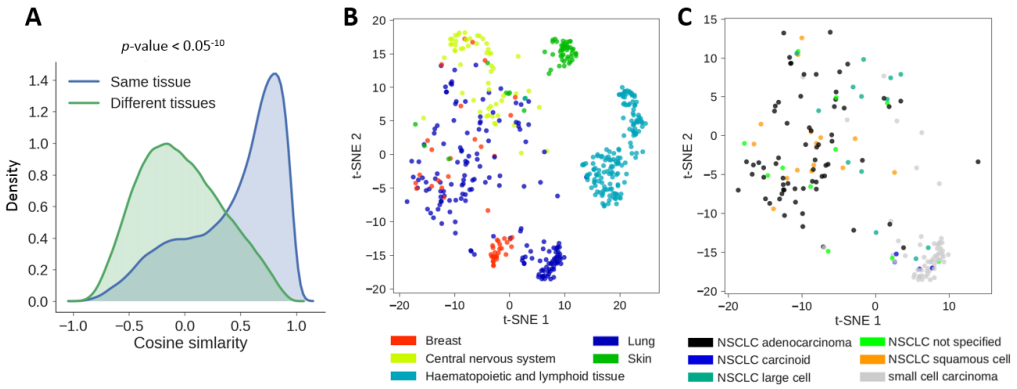

Figure 5.4 Subtypes of cell-lines on the pharmacogenomic space. (A) Kernel density plot showing distributions of cosine similarities between cell-lines of the same tissue type and of different tissue types (GDSC). (B) Visualization of GDSC cell-lines from top 5 most frequent tissue types using t-SNE. (C) Visualization of different subtypes of GDSC lung cancer cell lines using t-SNE.

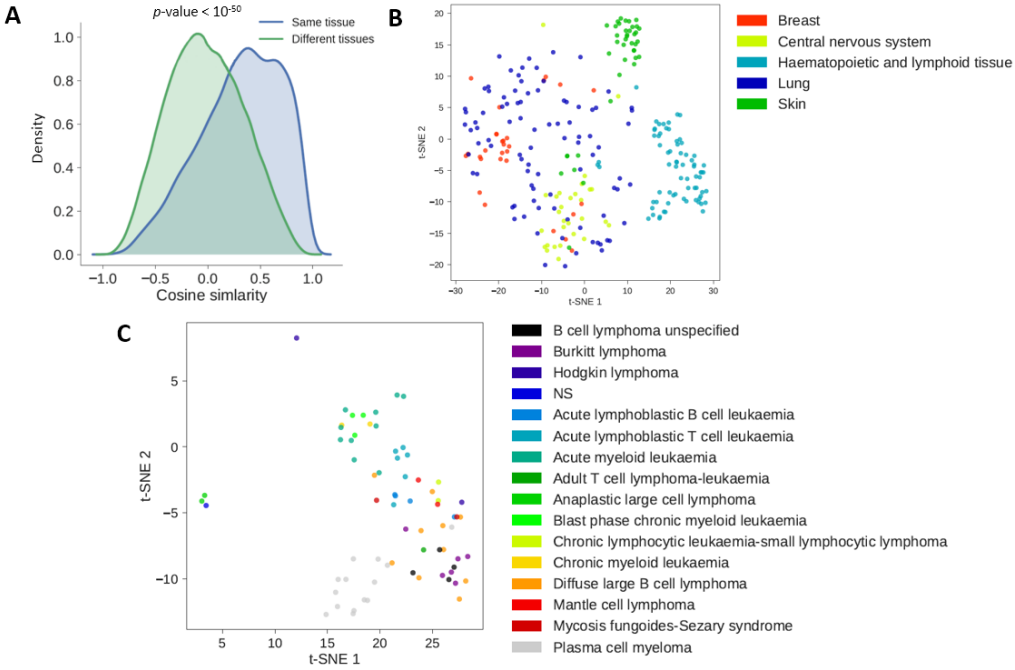

Figure 5.5 Subtypes of cell-lines on the pharmacogenomic space (CCLE). (A) Kernel density estimation plot showing cosine similarity within tissue type was significantly higher than between different tissue types. (B) t-SNE plot of top 5 tissue types. (C) t-SNE plot for subtypes of hematopoietic and lymphoid tissue cell lines.

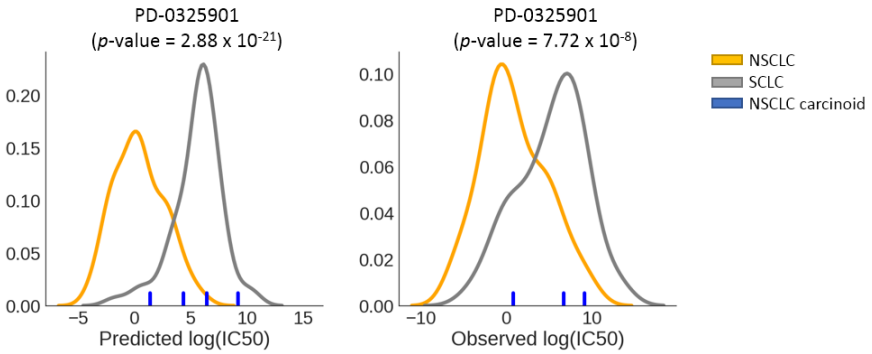

Figure 5.6 Comparison of drug response values between different cancer subtypes. (A) predicted drug response (B) observed drug response. Kernel density plot showing that NSCLC cell lines were more sensitive to PD-0325901 (inhibitor of MEK1 and MEK2). The NSCLC carcinoid cell lines seem to follow the distribution of SCLC rather than NSCLC cell lines.

By leveraging pathway information, we observed that activity scores for the ERK pathway in NSCLC cell-lines (mean=1.52) were significantly higher than for SCLC cell-lines (mean=-3.24; *p-value* $<1.3 \times 10^{-9}$), and the activation of ERK pathway due to KRAS mutation could play a role in the increased sensitivity to MEK inhibitors (RAF-MEK-ERK pathway; Stinchcombe and Johnson, 2014). In contrast, cell-lines with RB1 mutations had a significantly lower PD-0325901

sensitivity (adjusted *p-value* $< 7 \times 10^{-8}$), and correspondingly RB1 mutations were more common in SCLC cell-lines (67%) than in NSCLC cell-lines (10%). These observations corroborate earlier work suggesting that mutations in the RB1 pathway can inhibit the RAF-MEK-ERK pathway and thus induce resistance to MEK inhibitors [88]. Cell-line clusters determined by CaDRReS thus correlated well with mutation and pathway activation in explaining drug responses, and could serve to construct new testable hypotheses when such information is not known.

5.3 Association between drugs and pathways

Associations between cancer drugs and key pathways can be identified in the pharmacogenomic space based on pathway activity scores, cell line vectors, and drug vectors, as follows. Firstly, using 217 Biocarta pathway gene sets from MSigDB[85], pathway activity scores were calculated for each cell line by summing up gene expression fold-changes of genes in each pathway. To identify drug-pathway associations, we then calculated the Pearson correlation between pathway activity scores and predicted drug responses ($\mathbf{log}\,(\boldsymbol{IC_{50}})$; lower values indicate greater response), where a negative correlation suggests that a pathway is essential for drug effectiveness, while a positive correlation suggests that it plays a role in drug resistance.

As expected, we observed that drugs targeting the same gene were frequently associated with the same set of pathways (**Figure 5.7A**, **Figure 5.8**). For instance, four EGFR inhibitors had IC_{50} values that were negatively correlated with activation scores for the EGFR SMRTE pathway (assistant association), consistent with a study showing that amplification of the EGFR gene is correlated with high response to anti-EGFR agents[89]. Similarly, two RAF inhibitors showed

assistant associations with the VEGF-Hypoxia-Angiogenesis pathway (VEGF), in agreement with previous studies showing that VEGF expression induced by Raf promotes angiogenesis, while RAF inhibitors can block the RAF/MEK/ERK pathway and inhibit tumor angiogenesis [90,91].

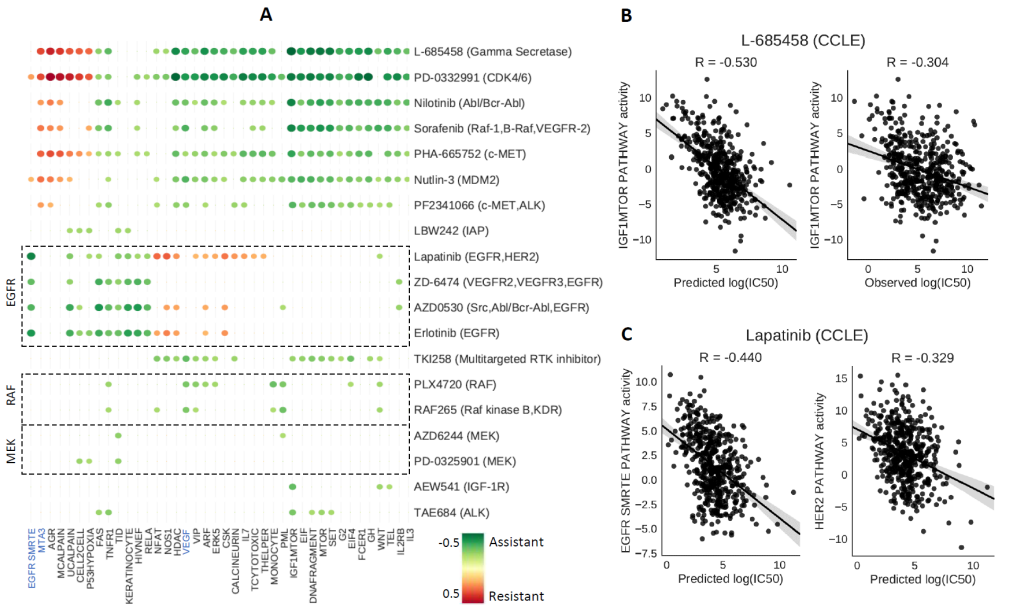

Figure 5.7 Drug-pathway associations identified on the pharmacogenomic space. (A) Drug-pathway associations based on CCLE data. For visualization, the top 40 pathways having the highest associations across drugs (average absolute correlation) were selected. Negative and positive correlations between pathway activity and drug sensitivity scores are denoted as being "assistant" and "resistant" associations, respectively. (B) Assistant associations between L-685458 (gamma-secretase inhibitor) and IGF-1 MTOR pathway. (C) Assistant associations between Lapatinib (EGFR inhibitor) and EGFR SMRTE and HER2 pathways.

We also observed resistant associations between the MTA3 pathway (MTA3) and multiple drugs such as L-685458 (gamma-secretase inhibitor) and PD-0332991 (CDK4/6 inhibitor), suggesting that the cell lines with inactivated MTA3 pathway tend to be sensitive to these drugs. In addition, the study of Fujita *et al.* showed that the absence of MTA3 leads to invasive growth in breast cancer[92]. Taken together, these observations suggest that drugs having a resistant association with MTA3 pathway might be effective when tumor growth is caused by the downregulation of the MTA3 pathway, although further work is needed to confirm this hypothesis.

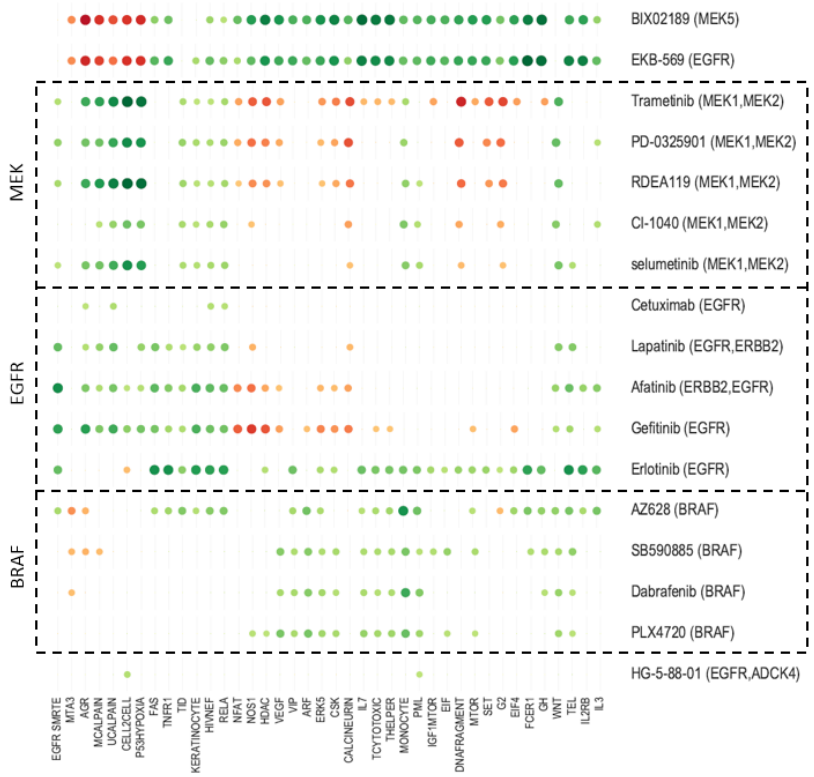

Figure 5. 8 Drug-pathway associations for GDSC drugs targeting EGFR, MEK, and BRAF.

In terms of drug-pathway associations, we noted that the strongest assistant association was observed between the drug L-685458 (gamma-secretase inhibitor) and the IGF-1 MTOR pathway (**Figure 5.7B**). This observation is also borne out in studies reporting that gamma-secretase inhibitors can inactivate MTOR signaling pathway and consequently induce apoptosis[93]. Interestingly, we observed a stronger association signal for predicted drug responses than observed drug responses, suggesting that CaDRReS may have the ability to reduce the noise observed in experimental drug response data. Stronger signals based on predicted drug responses were also observed for other known assistant associations, such as the one between Lapatinib (an EGFR inhibitor) and the EGFR SMRTE pathway (R=-0.440 vs -0.329; **Figure 5.7C**) as well as the HER2 pathway (R=-0.288 vs -0.242) (Harari, 2004; Medina and Goodin,

2008). These results highlight the utility of predictions from CaDDReS for discovering pathway biomarkers for drug sensitivity.

5.4 CaDRReS for cancer precision medicine

We proposed CaDRReS to predict patient-specific drug response based on transcriptomic profile, as well as addressed several issues in other existing models. We then showed that based on various evaluation criteria, CaDRReS' performance was among the best compared to other state-of-the-art methods (**Chapter 4**). Additionally, we investigated several applications of the pharmacogenomic space including drug response prediction for unseen patients/cell lines (**Chapter 4**), studying drug response mechanisms, classifying the cell lines according to their drug response profiles and tissue types, identifying a group of drugs having similar effects and targeting the same gene, and discovering drug-pathway associations.

However, we have only addressed inter-patient drug response heterogeneity by applying the model to predict patient-specific drug responses. Within a tumor or a cell line, there could be several clones of various cell types that respond to the same drug differently — *intra-patient* (or *intratumoral*) drug response heterogeneity. Therefore, predicting drug response based on the gene expression of bulk tumor might not accurately capture the right drug response behavior in a patient. In the next chapter, we investigated intra-patient drug response heterogeneity by analyzing tumor information obtained from TCGA dataset, as well as developed a new version of CaDRReS that account for several challenges in predicting drug response of single-cell data, which allow us to study different cell types within a tumor.

Chapter 6

Predicting cancer drug response in the presence of tumor heterogeneity

Each tumor is different as it progressively evolves into a complex system and interacts with its microenvironment[96,97]. As the disease progresses, clonal expansion, genetic diversification, and clonal selection iteratively occurs within tissue ecosystems[98]. Advances in genomic technologies have allowed us to observe an even greater than anticipated genetic, phenotypic and functional heterogeneity in cancer. These technologies have enabled us to study heterogeneity in cancer across different patients, different tumors within a patient, and different cell types within a tumor in order to understand the initiation, progression, and metastasis of cancer. Analyzing diversities of cell types (or cell states) within a tumor is one way to decipher intra-tumor heterogeneity, allowing us to understand the biological complexity of tumors and cell compositions, which in turn have been shown to affect patient survival rates and drug response[99].

While anti-cancer treatments can inhibit cancer clones, they can also induce selective pressure for the expansion of resistant variants, which could lead to therapeutic failure[98]. In previous chapters, we addressed inter-patient drug

response heterogeneity by using the information in tumor transcriptomic profiles. However, transcriptomic profiles measured from bulk tumors only represent an average gene expression across different cell types within a tumor, and using bulk gene expression could prevent us from identifying resistant cells that lead to therapeutic complications. Emerging single-cell technologies such as single-cell RNA sequencing (scRNA-seq) can enable us to measure the transcriptomic profile of thousands of individual cells within a tumor. The increasing availability of scRNA-seq data presents both opportunities and challenges to understand intra-tumor drug response heterogeneity for supporting precision oncology.

In this chapter, we study the impact of tumor heterogeneity on clinical outcomes, as well as investigate applications of CaDRReS to predict drug response in the presence of intra-tumor heterogeneity. Firstly, we performed a large scale tumor deconvolution analysis of more than ten thousand tumors from multiple cancer types. From the analysis, we observed relationships between tumor heterogeneity and clinical features such as patient survival rate and clinical drug responses, re-establishing the relationships found in smaller scale analyses[99,100].

Next, we propose CaDRReS-SC, a new version of CaDRReS that is more suitable for predicting cell type-specific drug response based on single-cell data, to study intra-tumor drug response heterogeneity. Patient-derived cell lines are cancer cells that are derived from a patient and grown in a laboratory for studying cancer biology and testing the response to cancer treatments. In this analysis, we obtained scRNA-seq data from cancer cell lines derived from head and neck cancer patients as a case study, as they allowed us to test drug response for a panel of drugs while being patient proximal and retaining a higher degree of cellular heterogeneity[74]. We then applied CaDRReS-SC to predict drug response for each

cell type observed in the patient derived cell lines based on the transcriptomic profile.

In addition, we introduce a Newton-like method for systematically aggregating cell type-specific drug response predictions to capture the overall response of a cell mixture. The aggregation method iteratively estimates an expected drug response of a cell line by taking into account the proportions, the sigmoid shape of dose-response curve, and the predicted drug responses of multiple cell types identified within the cell line. Compared to the predictions based on bulk gene expression, the aggregation of cell type-specific predictions based on single-cell gene expression provided better concordance with *in vitro* drug response measurements for heterogeneous cell lines.

Finally, because a single-drug treatment might not be able to inhibit some existing resistant cell types in a tumor that could lead to treatment complications, it is useful to identify a combination of drugs that together can inhibit various cell types. In the last section, we investigate the application of CaDRReS-SC to predict drug responses for different cell types in each patient, and then we incorporate clinically-feasible drug dosage information to identify patient-specific drug combinations to inhibit those cell types.

6.1 Methods

Identifying relationships between heterogeneity and clinical features

We obtained a dataset of 10,956 tumors from 32 cancer types (one of the most extensively characterized, publicly available tumor datasets) from The Cancer Genome Atlas Research Network (TCGA Research Network)[5]. The TCGA dataset provides multi-omic profiles of the tumor as well as clinical outcomes including survival days and clinical drug responses. For gene expression used in

this study, we obtained the TCGA pancan gene expression matrix[101] which contains RSEM-normalized RNA-Seq values, i.e., TPM (Transcripts per Million).

Next, we identified cell types in each tumor by using CIBERSORT[100], a tool for tumor deconvolution based on a transcriptomic profile. CIBERSORT requires a cell signature matrix that contains transcriptomic profiles of different cell types. Two cell type panels were used for our analysis including a default panel of immune cells (LM22) and a panel of cancer cells from different histology subtypes of GDSC (GDSC). For each input tumor, CIBERSORT outputs percentages for different cell types present in the tumor.

To measure the degree of heterogeneity in each tumor, we defined a heterogeneity score (H) as information entropy, $H = -\sum_i P_i log P_i$, where P_i is a percentage of cell type i identified in a tumor. Cell types present at a relatively small percentage (<5%) were excluded to reduce the impact of classification noise and make the score more robust. For each cancer type, we classified tumors into three categories based on their heterogeneity scores: low (<Q1, i.e. scores in the first quartile of the distribution), medium (Q1 to Q3), and high (>Q3). Finally, we performed survival analysis for each cancer type to identify differences in survival rates between different heterogeneity classes. The p-values were calculated based on pairwise the logrank test, i.e., three pairs for each cancer type, and corrected by the Bonferroni correction.

Comparing survivorship of tumors clustered by heterogeneity and transcriptomic profile

Besides the three groups of tumors categorized according to their heterogeneity scores, we clustered the tumors based on their transcriptomic profiles by using the non-negative matrix factorization (NMF) method[102]. NMF

clustering decomposes a gene expression matrix into two matrices (tumor and gene), and the latent dimension is equal to the number of clusters. To determine the appropriate number of clusters, we calculate a silhouette score that captures how similar a tumor is to its cluster compared to other clusters[103]. For each cancer type, the number of clusters corresponding to maximum silhouette score was used.

Single-cell data for head and neck cancer patient-derived cell lines and cell clustering

We used a previously published scRNA-seq dataset consisting of 12 cancer cell lines derived from both primary and metastatic tumors for six head and neck cancer patients[74]. The dataset contains 1,241 cells in total, and expression levels for 26,968 genes was measured by scRNA-seq. In addition, each patient-derived cell line was tested with anti-cancer drugs, and inhibition scores (the proportion of cancer cells inhibited by 1 uM of a drug) were calculated. In total, there are 90 drugs that are predictable by CaDRReS model that was trained on the GDSC dataset. Finally, we obtained cluster information for the 1,241 cells, clustered based on transcriptomic profiles using Seurat, a package for scRNA-seq data analysis[104].

CaDRReS for single-cell data (CaDRReS-SC)

CaDRReS-SC is a framework for predicting cell type-specific drug response based on scRNA-seq data. The framework consists of a new objective function for CaDRReS that is better suited for analyzing cell types that are not part of the training set, a combination of CaDRReS models trained for different drug classes, and a preprocessing step for scRNA-seq data that allows us to apply CaDRReS to predict cell type-specific drug response.

Firstly, we observed that the performance of CaDRReS for unseen samples was relatively lower than for seen samples, likely due to challenges in estimating bias terms for unseen samples. Therefore, to improve the performance of CaDRReS for predicting unseen samples, we propose a new objective function for training the model (**Equation 3**). Specifically, we removed the cell line bias term from the original objective function (**Equation 1**) to allow the pharmacogenomic space to directly capture the effect of cell line bias. Compared to the original objective function, the predictive performance across unseen cell lines improved (37% based on average per-drug Spearman correlation of 5-fold cross-validation on the GDSC dataset, *p-value* < $9.71e^{-37}$).

Equation 3.

$$\begin{aligned}\hat{s}_{ui} &= \mu + b_i^{Q} + \boldsymbol{q}_i \cdot \boldsymbol{p}_u \\ &= \mu + b_i^{Q} + \boldsymbol{q}_i(\boldsymbol{x}_u \boldsymbol{W}_P)^T\end{aligned}$$

Additionally, we trained two separate CaDRReS models for cytotoxic and targeted drugs, allowing the models to learn pharmacogenomic spaces that are more specific to these drug classes.

Due to the low sensitivity of scRNA-seq data, some expressed genes might not be detected in a subset of the cells[105]. Thus predictions based on applying CaDRReS to individual gene expression values for each cell might not be accurate and robust. Therefore, we aggregated transcriptomic profiles of cells in each cell cluster by calculating the 95th-percentile of expression for each gene. We observed that the 95th-percentile of each gene across all individual cells in a cell line showed higher concordance when compared with the bulk RNA-seq data of the cell line (Pearson correlation: 0.78 vs 0.70). For the following analysis, we refer to the prediction based on the aggregated gene expression of each cluster as cell type-

specific prediction, as each cluster consists of cells that have similar gene expression signature.

Cell type-specific predictions allow us to study drug response heterogeneity across different cell types within a cell line or tumor. We can aggregate these predictions to obtain an overall response of each a specific tumor, and the aggregated prediction can guild us to select drugs to inhibit the tumor. In addition, due to the lack of single-cell drug response information, we can compare the aggregated prediction to validate the cell type-specific predictions.

CaDRReS-SC can predict IC_{50} values, which define positions of the dose-response curves of different cell types, which occupy different proportions in a cell line or tumor (**Figure 6.1A**). Each dose-response sigmoid curve is defined by the following equation:

Equation 4.

$$y = \frac{1}{1 + 2^{(a-x)b}}$$

where y represents cell viability ranging from 0 to 1, x is drug concentration in the log scale, a and b define position and slope of the curve. The IC_{50} value is equal to a because it correspond to $y = 0.5$. Next, we define the aggregated dose-response curve as follow:

Equation 5.

$$y = \sum_i p_i \frac{1}{1 + 2^{(a_i-x)b_i}}$$

where p_i is a percentage of cell type or cluster i in a given cell line and a_i and b_i are position and slope of the dose-response curve of cell type i.

To obtain the overall drug response IC_{50} of a cell line or tumor, we have to find x that $y = 0.5$ (**Equation 5**). Therefore, we apply the Newton method to estimate IC_{50} of the aggregated dose-response sigmoid curve (**Figure 6.1B**). Firstly, we initiate x, calculate y based on Equation 5, and calculate a slope of the tangent line at (x, y). We then calculate a new x based on the slope and the new x corresponds to the value of y that is closer to 0.5. Finally, we stop the iterative process when we identify x that corresponds to y that is almost equal to 0.5. The algorithm for aggregating dose-response curves is summarized below.

A Newton-like algorithm for estimating IC_{50} of the combined dose-response curve

Initiation:	$x = \sum_i a_i p_i$	Initiate the dosage value.
Iterative estimation:	$m = \sum_i \frac{p_i 2^{b_i(a_i - x)}}{(1 + 2^{b_i(a_i - x)})^2}$	Calculate a slope of the tangent line.
	$y = p_i \sum_i \frac{1}{1 + 2^{(a_i - x)b_i}}$	Calculate drug response based on the combined dose-response curve, which is a weighted-average of the dose-response curves of all cell types.
	$x = x + \frac{0.5 - y}{m}$	Update x to the new dosage which is closer to 50% cell viability.
Stop criterion:	$\lvert y - 0.5 \rvert < \varepsilon$	Stop and return x when y is almost equal to 50% cell viability.

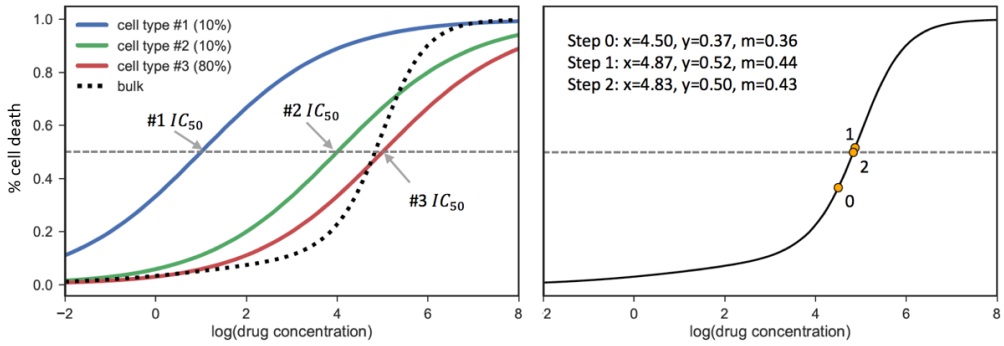

Figure 6.1 Aggregating cell type-specific drug response predictions. (A) An example of dose-response curves. Cell types response to the same drug differently as illustrated in three dose-response curves, while a dashed line represents IC_{50} of the bulk. (B) A Newton-like method to iteratively calculate IC_{50} of the combined curve by taking into account slope and position of each cell type's curve and percentage of the cell type in a cell line.

Comparing CaDRReS-SC prediction against observed drug responses in patient-derived cell lins

Several cancer drugs were tested on the head and neck and patient cell lines, and the inhibition score for each pair of drug and cell line was calculated. However, to evaluate the predictive performance of the models, the observed inhibition score should not be directly compared against the predicted IC_{50} because they are different units. Specifically, the inhibition score is a percentage of cells that are inhibited at 1uM of a drug, while IC_{50} is a dosage that the drug can inhibit 50% of the cells. Therefore, we applied a cutoff at 50% inhibition to classify *in vitro* drug response into two groups, sensitive and resistant. Finally, for each head and neck patient-derived cell line, we counted the number of sensitive drugs that were predicted among the top-5 predictions from each method.

Predicting patient-specific drug combination

We combined single-cell data obtained from both primary and metastatic cell lines derived from each head and neck cancer patient to study drug response heterogeneity at a patient-level, as a cancer treatment can affect both primary and metastatic tumors. We then applied CaDRReS-SC to predict cell type-specific drug

responses and identified a combination of drugs that can inhibit all cell types detected (at >10% frequency) in a patient.

Besides, the ranges of feasible concentrations are different across drugs, so predicting drug combination solely based on the predicted IC_{50} is not suitable. For example, we might tend to select a combination of drugs that have low IC_{50}, while the dosages used in a clinic are yet lower. To address this issue, we incorporated a clinically relevant value, i.e., the maximum plasma concentration (Cmax), which is the drug concentration observed in plasma when using maximum dosage indicated on the drug label[106]. Based on Cmax value, we can identify a combination of drugs that can inhibit different types of cells, as well as have appropriate dosages to be used in a clinic.

For a given pair of drug and cell type, we classified the predicted drug response into four groups according to drug-specific Cmax values. Firstly, we used Cmax as a drug-specific cutoff value to identify responders (predicted IC_{50} < Cmax) and non-responders (predicted IC_{50} > Cmax) cell types. We then classified the responders into three groups (low dosage responder, medium dosage responder, and high dosage responder) based on half Cmax and a quarter of the Cmax value. Lower dosages of the drug could reduce potential side effects, so we prefer drugs that have predicted IC_{50} values significantly lower than the Cmax. Next, we defined a weight for each group: non-responder = 0, high dosage responder = 1, medium dosage responder = 2, and low dosage responder = 3. Finally, we solved a weighted maximum cover problem to identify the appropriate patient-specific drug or drug combination. In the case that multiple combinations have equal maximum weight, we broke ties by prioritizing drugs with higher selectivity, i.e., a drug that has a smaller total weight (summing up across cell types).

6.2 Results

Identifying relationships between tumor heterogeneity and clinical outcomes

As cancers progress, clonal expansion, genetic diversification, and clonal selection lead to intra-tumor heterogeneity, i.e., multiple cell types with different genetic and transcriptomic profiles exist in a tumor. Traditionally, multi-omic profiles of a bulk tumor were measured to study the disease, but they do not directly capture information on proportions of various cell types within a tumor. For example, different proportions of immune cells have been shown to have significant associations with patient clinical features and cancer genetic alterations[99]. Also, tumors with high levels of intra-tumor heterogeneity can have inferior clinical outcomes due to the expansion of pre-existing subclonal populations or from the evolution of drug-tolerant cells[107].

In this section, we further evaluated the relationship between intra-tumor heterogeneity and treatment outcomes based on large cancer genomics datasets. The Cancer Genome Atlas (TCGA) provides transcriptomic and clinical outcome such as survivorship and clinical drug response data for several thousand cancer patients. The relationships between intra-tumor heterogeneity and treatment outcomes has not been investigated in such large datasets and across different cancer types. Here, we applied a tumor deconvolution approach to compute the proportion of different cell types in tumors to estimate their heterogeneity and association with clinical features (see **Methods**).

Firstly, we run the deconvolution method based on the immune cell panel (LM22), a default cell type panel. It has been shown that immune cell composition are associated with patient clinical features and response to immunotherapy in

some cancer types[99]. In our analysis, we observed that same cancer type patients could have different immune infiltration profiles, as the tumor samples did not cluster based on cancer types (**Figure 6.2A**). In addition, to study relationship between tumor heterogeneity and clinical features including survivorship and drug response provided by the TCGA dataset, we decomposed tumors based on the GDSC histology subtype panel to identify different cancer cell types within a tumor (**Figure 6.2B**). As expected, we observed clusters of tumors from the same cancer type, suggesting that different cancer types have unique cancer cell type compositions. As we did not consider immunotherapy treatment in our study, we focused on the deconvolution results based on the GDSC histology subtype panel, in line with the observation in Chapter 5 that histological subtypes can define drug response.

Next, we investigated the relationship between heterogeneity and clinical outcomes (see **Method**). Using survival analysis, we observed significantly different survivorships between low and high heterogeneity groups for some cancer types such as Low Grade Glioma (LGG, p-value < 1.24e-4) and Sarcoma (SARC, p-value < 2.88e-4). These results suggest that intra-tumor heterogeneity is associated with survivorship of the patients and the effects of heterogeneity degrees could vary across cancer types (GDSC cancer subtypes panel, **Figure 6.2A**).

Figure 6.2 Tumor deconvolution results. (A) LM22 immune cell type panel. (B) GDSC histological subtype panel. We clustered the tumors based on cell type compositions and observed that for (B) the tumors were clustered based on cancer type. For example, breast (histological subtype) was enriched for Breast invasive carcinoma (BRCA) tumors.

We hypothesized that the heterogeneity score might be superior in explaining patient survivorship compared to transcriptomic profiles. To examine this, we performed non-negative matrix factorization (NMF) clustering for all cancer types (see **Methods**). We applied NMF clustering on LGG and SARC samples based on transcriptomic profiles, and the numbers of clusters were set to 2, corresponding with the highest overall average silhouette score. Based on survival analysis, we observed a larger survivorship variation between different heterogeneity groups comparing to transcriptomic clusters, suggesting that heterogeneity could be a better indicator of survivorship (LGG: p-value < $3.72e^{-4}$ vs p-value < $1.94e^{-2}$, SARC: p-value < $8.64e^{-4}$ vs p-value < $5.98e^{-2}$, **Figure 6.3A-B**). Additionally, we investigated the difference between two cluster types by

counting the number of overlapping tumors between different clusters. For SARC, we observed overlaps between the medium/high heterogeneity groups and cluster 2 (Jaccard index = 0.38, 0.39), suggesting that heterogeneity information could further differentiate the tumor within cluster 2 into two groups that have different survivorship (**Figure 6.3C**).

Finally, we examined the relationship between intra-tumor heterogeneity and clinical drug response categories (see **Methods**). For Doxorubicin (a chemotherapy drug for blocking an enzyme that manages DNA tangles), we observed that heterogeneity scores were significantly different between subjects falling into the categories of "Complete Response" and "Clinical Progressive Disease" (p-value < 2.20e-4, **Figure 6.3D**), highlighting that intra-tumor heterogeneity plays a role in clinical drug response.

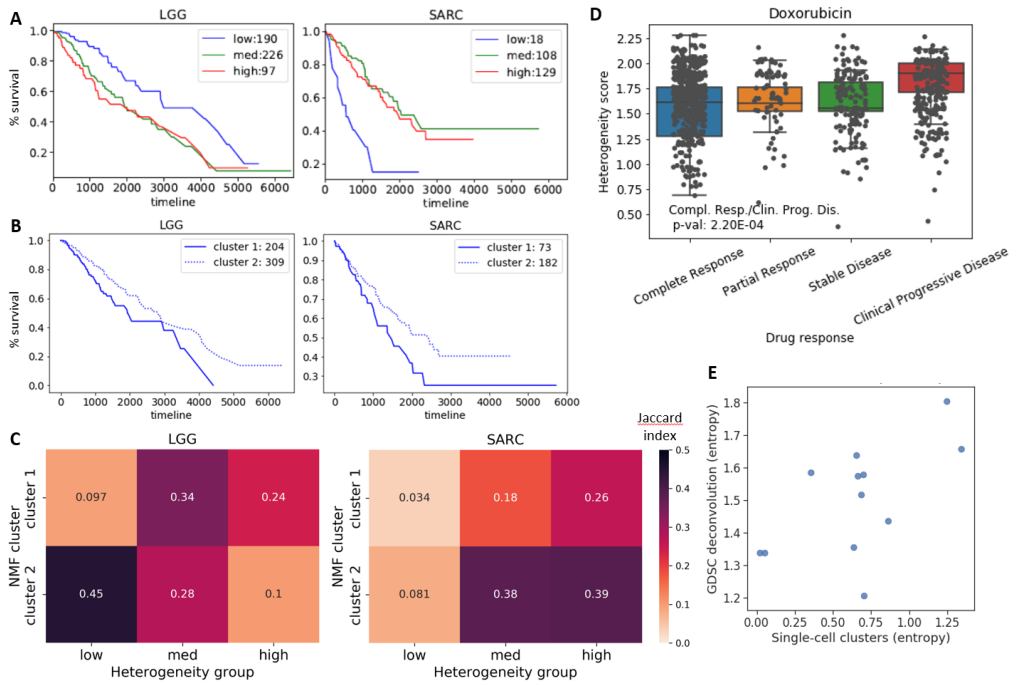

Figure 6.3 Intra-tumor heterogeneity and clinical patient features. (A) Survival analysis comparing the different degree of heterogeneity. (B) Survival analysis comparing different groups of patients based on NMF clustering using gene expression. (C) The overlap between heterogeneity and gene expression clusters. (D) Comparison of heterogeneity scores for patients from four Doxorubicin response groups. (E) Comparison of heterogeneity scores and entropy values calculated from cell cluster percentages based on single-cell RNA-seq data of 12 head and neck patient-derived cell lines.

We acknowledged that calculating intra-tumor heterogeneity scores based on GDSC histological subtypes is not the best strategy, and accuracy of heterogeneity scores needs to be evaluated. To address this issue, we compared entropy based on CIBERSORT's deconvolution results of the aggregated transcriptomic profiles (GDSC histological subtype panel, **Figure 6.2B**) and entropy based on cell cluster percentages (**Figure 6.4A**) across 12 head and neck patient-derived cell lines. Here we only considered the clusters or the histological subtypes with at least 5% presence to avoid the noise. Although the number of cell lines is small, we observed a significant correlation between the two types of entropy values (**Figure 6.3E**, Pearson correlation = 0.60, p-value < $4.08e^{-2}$), suggesting that the heterogeneity scores could capture the gene expression heterogeneity within a sample.

Using scRNA-seq data to predict drug response in the presence of cellular heterogeneity

Using scRNA-seq data, we can apply CaDRReS to predict drug response of each cell type identified within a tumor or patient. Firstly, we obtained scRNA-seq of head and neck patient-derived cell lines, as well as information of 22 cell clusters (**Figure 6.4A-B**). Next, we proposed a modified version of CaDRReS, *CaDRReS-SC,* to further improve the performance on an unseen sample, alleviate a challenge of scRNA-seq data which typically have low sensitivity in detecting genes, and predict overall drug response of a tumor by aggregating cell type-specific predictions (see **Method**).

Next, we compared CaDRReS-SC's drug response predictions (IC_{50}) with the *in vitro* observed drug response (inhibition score). Both CaDRReS (predictions based on bulk gene expression) and CaDRReS-SC (the combined cell type-specific drug predictions) showed better performance than the baseline prediction.

Although the predictive performance of CaDRReS and CaDRReS-SC were not significantly different, we observed that the predictions based on CaDRReS-SC were better than CaDRReS for the cell lines with a higher degree of heterogeneity; HN120M and HN160M that consist of three and four cell types, respectively (Figure 6.4A and 6.5). These results suggested that intra-tumor drug response heterogeneity should be considered in predicting drug response.

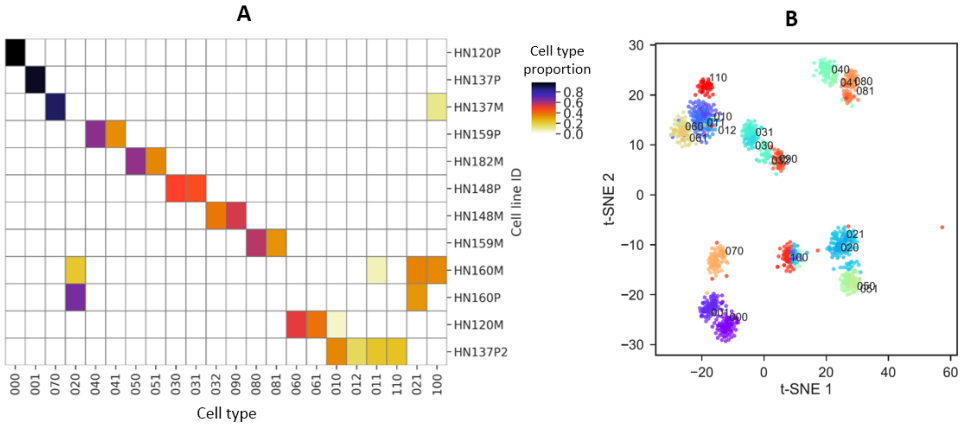

Figure 6.4 (A) Different cell types were identified in each patient-derived cell line. (B) The t-SNE plot shows clusters of single cells.

Predicting patient-specific drug combinations to treat heterogeneous tumors

In the presence of intra-patient heterogeneity, single drug treatments may not be sufficient to inhibit all subpopulations of cancer cells in a patient[108], and resistance to treatment can lead to the expansion of pre-existing subclonal populations or the evolution of drug-tolerant cells[107]. Therefore, we studied the application of CaDRReS-SC to predict drug combinations that can inhibit multiple cell types identified within a patient. Treatment with such a drug cocktail could potentially prevent the emergence of aggressive clones due to drug-tolerant cells. Additionally, a drug combination based treatment can allow for smaller dosages of individual drugs and thus lead to lower overall toxicity.

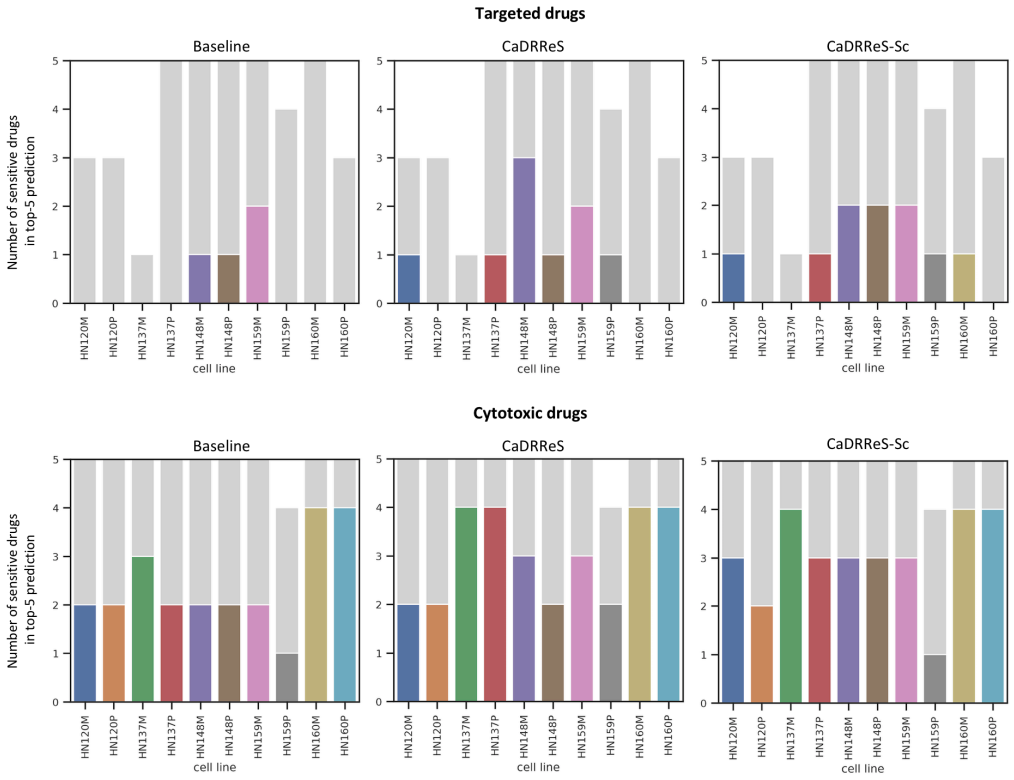

Figure 6.5 Comparison of in silico predicted drug response (IC_{50}) and the in vitro observed drug response (inhibition score) for 71 targeted drugs (top) and 19 cytotoxic drugs (bottom). Numbers of sensitive drugs detected in the top-5 predictions based on three different methods: Baseline prediction (left), CaDRReS prediction based on bulk gene expression (center), and CaDRReS-SC based on scRNA-seq data (right).

After combining single-cell data obtained from both primary and metastatic cell lines derived from each patient, we observed different heterogeneity degrees across patients (**Figure 6.6A**). Next, we applied CaDRReS-SC (without the aggregation of cell type-specific predictions) to predict cell type-specific responses for five standard-of-care drugs of head and neck cancer including Docetaxel, Paclitaxel, Cisplatin, Lapatinib, and Gefitinib. The drugs and cell types projected on the 10-D pharmocogenomic spaces of two drug groups, drugs with median $IC_{50} < 1$uM (cytotoxic) and those with $IC_{50} \geq 1$uM (targeted). We observed that drugs targeting EGFR (Lapatinib and Gefitinib) were clustered together, while Cisplatin exhibited distinct responses (**Figure 6.6B**).

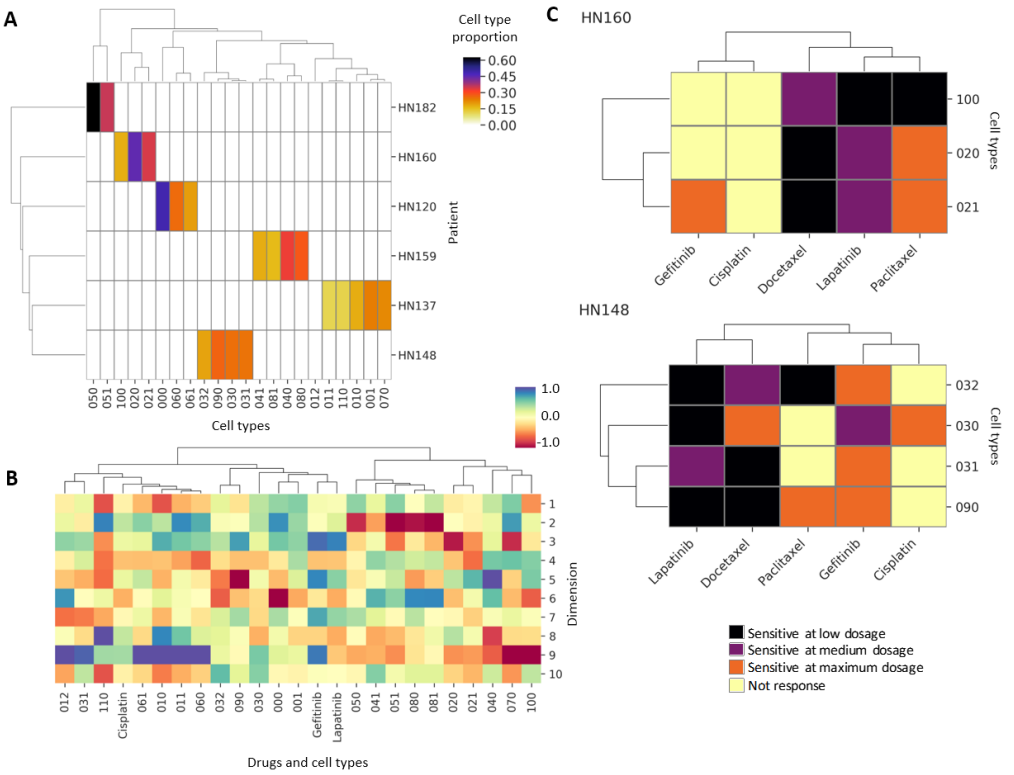

Figure 6.6 Predicting patient-specific drug combination. (A) The proportion of cell types in each head and neck cancer patient. (B) A heatmap visualizes the 10-D pharmacogenomic space of drugs with IC_{50} < 1uM and 22 cell types. Drugs and cell types were clustered based on cosine similarity. (C) Identifying a combination of Docetaxel and Paclitaxel for HN160 and a combination of Lapatinib and Docetaxel for HN148.

We obtained drug-specific Cmax values to classify the predicted IC_{50} into four response classes and solved a weighted maximum cover problem to identify a combination of drugs that can inhibit different cell types within a patient (See **Methods**). Based on the cell type-specific predictions, we identified drug combinations (or a single drug) as being suitable for the six patients. Among these, we identified a combination of Docetaxel and Paclitaxel for HN160 and a combination of Lapatinib and Docetaxel for HN148 (**Figure 6.6C**). Docetaxel and Paclitaxel are commonly used in breast cancer treatment[109] and were predicted for HN160 to inhibit three cell types identified within the patient. Both Docetaxel-Paclitaxel and Docetaxel-Lapatinib were able to inhibit the tree cell types at low dosage, but Paclitaxel was more selective to cell cluster 100 and have a lesser

effect on cluster 020 and 021. A combination of Lapatinib and Docetaxel has been extensively studied for advanced cancer treatment[110] and was predicted to inhibit four cell types identified within HN148. For this patient, Lapatinib could be the most suitable drug, but cell type 031 (25% of the tumors) might be resistant to the drug at low dosage. Combining with a low dosage of Docetaxel, it is possible to apply lower drug dosages, which could lower toxicity, to inhibit most cell types in the patient.

We also identified for HN120 a combination of Lapatinib and Gefitinib, which are *EGFR* inhibitors. Although further experimental validation is needed, a combination of drugs that target the same pathway could be more effectively inhibit the tumor due to a synergistic effect [111]. Taken together, these results suggested that the ability to predict drug response for multiple cell types within a patient is essential for identifying drug combinations to inhibit the tumors, as well as provides a promising step toward developing a decision-support system for precision oncology.

6.3 Discussion

We investigated intra-tumor drug response heterogeneity by analyzing a large tumor dataset (TCGA), as well as modifying and applying CaDRReS to predict cell type-specific prediction. Firstly, we applied a deconvolution algorithm to a large tumor dataset (TCGA) to evaluate the relationships between intra-tumor heterogeneity and survivorship across cancer types in the TCGA dataset. We also observed a pancan association between the heterogeneity degree and drug response, although it could be more informative to study the association between the heterogeneity and drug response for a specific cancer type. Due to the limited number of samples that have clinical drug response information, applying a model

for predicting drug response on the TCGA dataset might allow us to study roles of intra-tumor heterogeneity in drug response in different cancer types.

We note that there are studies which have analyzed the relationships between infiltrating immune cells and drug response behaviors[100,112]. However, those studies focus on infiltrating immune cells, while we aim to identify different cancer cell types in order to calculate a degree of heterogeneity for each tumor. Therefore, we use GDSC histological subtypes as proxies of different cancer cell types with the assumption that the public cell line panel can represent the diversity of cancer cells.

To study intra-tumor heterogeneity, we obtained single-cell RNA-seq data of head and neck patient-derived cell lines constructed from both primary and metastatic sites. Based on the assumption that the public cell line panel can capture the diversity of cancer cell types, we applied CaDRReS to predict drug response specifically for different clusters of single cells. Although it is not clear if the cell clusters represent different cell types or different cell states of the same cell type, each cell cluster has a unique transcriptomic profile that could lead to different drug response behaviors. A patient-derived cell line serves us a snapshot of a tumor, and understanding of drug response behavior of this snapshot can be useful for future studies of intra-tumor heterogeneity as well as drug-resistant mechanism.

In the single-cell analysis, the dropout issue could lead to missing information of some genes[113], so we alleviated this problem by clustering the cells based on gene expression and carefully combined the gene expression of cells in each cluster to obtain the profile that represents each cell type. For applying CaDRReS to single-cell data, we proposed a new objective function of CaDRReS that suits better for predicting drug response of unseen samples. However, we

acknowledged that improving the accuracy of sample bias prediction could be a better solution to improve the performance of the model for unseen samples.

Besides the new objective function, we trained two separated CaDRReS models for targeted and cytotoxic drugs as we found that the performance was improved comparing to a single model for all drugs, suggesting that training a model for each drug class could improve the predictive performance while the number of samples is not too small to train a robust models. Systematic pre-clustering of drugs could allow us to further improve the performance and obtain pharmacogenomic spaces that capture specific behaviors of each drug type. Moreover, drug properties information can be incorporated into the model, i.e., instead of directly calculating a vector $\boldsymbol{q}_i$ to represent drug i in the pharmacogenomic space (**Equation 1**), we calculate $\boldsymbol{q}_i = \boldsymbol{y}_i \boldsymbol{W}_Q$, where $\boldsymbol{y}_i$ represents drug features and $\boldsymbol{W}_Q$ is a matrix to project the drug features onto the pharmacogenomic space. This modification will allow the models to predict the dose responses for an unseen drug as well as learn new parameters that are specific for different drug classes or chemical groups.

We combined single-cell data of the cell lines derived from both metastatic and primary tumors to predict drug response at a patient level. Using single drug treatment might not be able to kill some existing cancer cell types, which could transform into aggressive resistant cells and activate metastasis. We proposed a method to predict an upfront combination of drugs to inhibit multiple detected cell types, and several known drug combinations were predicted. However, due to the lake of data, in this analysis, we did not consider drug-drug interactions such as antagonistic and synergistic interactions. Also, interactions between drugs could depend on drug dosages, and we still need to examine this aspect to improve patient-specific drug combination predictions.

Using the patient-derived cell line data allowed us to validate our predictions with the existing drug-screening data. The patient-derived cell lines are also more proximal to the patients than the standard cancer cell lines, suiting for future experimental validation and follow up. To compare our predictions with the in vitro drug response, we systematically aggregated cell-type specific predictions. We observed that a simple weighted average of the cell type-specific IC_{50}s was not accurate due to the sigmoid shape of the dose-response curves, so we proposed a Newton-like method to estimate the IC_{50} of the aggregated dose-response curve for heterogeneous cell lines. We found that the systematic aggregation of the predicted cell type-specific IC_{50}s were more accurate when compared to predictions based on the bulk gene expression. These results suggested that, ideally, we need a recommender system that can predict patient-specific drug response behavior need to consider both intra- and inter-patient heterogeneities for serving as a decision-support system in cancer treatment.

Chapter 7

Conclusion

The field of machine learning for drug response prediction is challenging and various types of models have been proposed. We discussed several existing drug response prediction methods as well as their strengths and limitations (Chapter 3). With the goals of improving predictive performance and enhancing interpretability of the existing models, we developed CaDRReS, a recommender system for patient-specific drug response prediction based on transcriptomic information (Chapter 4). CaDRReS is based on a matrix factorization technique that simultaneously learns across multiple drugs and samples (patients, cancer cell lines, and cancer cells) to obtain a hidden space that captures relationships among drugs and samples. This technique can increase the number of samples to learn a more robust model. We also introduced bias terms to remove the normalization step of drug response values, which lead to the loss of drug ranking information within a sample. Moreover, we created the kernel feature (based on transcriptomic information of essential genes) that is more robust than gene expression values and enables us to apply CaDRReS to datasets obtained from different transcriptomic platforms.

We note that predicting the response of a given drug in a patient via responses observed in cancer cell lines is performed based on the following assumptions. Firstly, the model learns from standard cancer cell lines which might not maintain intra-tumor heterogeneity. Secondly, we assume that each cell line represents a cancer cell type and that the cell line panel can represent the diversity of cancer cell types. Also, *in vitro* cell line models have their limitations: they could not capture effects of tumor microenvironments; they might only represent a specific tumor sector in a patient; clonal selection can occur in the model generating process[114]. However, it has been observed that patient-derived cell lines can capture drug response phenotypes and biomarkers that present in the corresponding patients[74]. Using *in vitro* model allows high-throughput screen for hundreds of drugs across multiple cell lines, which is currently not possible *in vivo* due to cost limitation. Therefore, as the first step toward applying the model for a patient, we use the standard cell lines as proxies of cancer cell types and develop a model to predict drug response based on their transcriptomic profiles.

Besides the predictive performance, interpretability of the models is another aspect that has been ignored in many studies. The pharmacogenomic spaces learned in CaDRReS captured and visualized drug-sample relationships, enabling us to study drug response mechanisms (Chapter 5). We showed that they could be used for identifying groups of drugs having similar mechanisms, subtypes of cell lines based on drug-response behaviors, and known drug-pathway associations. However, the biological interpretability of the pharmacogenomic spaces has not yet been fully explored. For example, the span of the space and biological interpretation of each dimension can be further analyzed. In addition, with a larger amount of drug response data (or with an ability to combine drug response data from various sources), a more complicated

pharmacogenomic space can be learned. For examples, a deep collaborative filtering model can be constructed to learn complicated patterns of gene expression to explain inter-patient or intratumoral drug response heterogeneities [115]. A multi-layer network might also help us to learn interactions among genes and embedding representations of samples through multi-omics data, as well as protein interaction networks can be integrated into the model through convolutional or manifold techniques [116].

Last but not least, we investigated intratumoral drug response heterogeneity, the other important aspect needed for identifying resistant cell types and addressing drug-resistant complication and metastasis in precision oncology (Chapter 6). We evaluated the relationships between intratumoral heterogeneity and clinical outcomes across cancer types by applying a deconvolution algorithm to a large tumor dataset (TCGA). Although the algorithm can decompose tumor bulk, the accuracy relies on the input cell type panel, and the tools might not be able to differentiate microenvironment effects. Recently, the availability of scRNA-seq data has allowed us to measure the gene expression of individual cells in tumor bulk, enabling us to study the complexity within tumors. Taken together with the ability of CaDRReS to predict drug response for unseen samples, we used single-cell data from head and neck cancer patients as a case study for CaDRReS to study intratumoral drug response heterogeneity.

We proposed CaDRReS-SC, a framework to apply a modified version of CaDRReS to scRNA-seq dataset: (1) to increase the sensitivity in detecting genes, we systematically aggregated transcriptomic profiles of the cells within the same cluster; (2) we modified the objective function of CaDRReS to improve the performance of predicting drug response for unseen samples; (3) we trained two separated CaDRReS models for cytotoxic and targeted drugs, allowing the models

to learn more specific drug response mechanisms of each drug class while retaining a large number of samples for robustness; and (4) we proposed the algorithm for aggregating cell type-specific predictions to predict overall response of a given tumor or patient.

Single drug treatment might not be able to inhibit some existing cancer cell types that could transform into aggressive resistant cells and lead to metastasis. Therefore, we combined single-cell data of the cell lines derived from both metastatic and primary tumors to predict patient-specific drug combinations, i.e., a combination of drugs that inhibits multiple cell types identified within a patient. Using scRNA-seq of cell lines derived from head and neck patients, we showed that the pharmacogenomic space could capture the interactions between the standard of care drugs and different cell types. We incorporated clinical dosages information to identify a combination of drugs that could inhibit multiple cell types in a given patient and that the dosages are in clinical ranges. It could be more useful to add clinical dosages information into the model training step, allowing the pharmacogenomic space to capture the clinical aspect.

Several challenges still exist in predicting patient-specific drug response. Firstly, an adverse effect of drugs has been ignored in most of the study due to the lack of drug response data of a healthy cell line. Nevertheless, by using a computational model that accurately predicts drug response in unseen samples, multi-omics profiles of adjacent normal tissues might allow studying drug adverse effect in patients. Secondly, tissue information could be useful, although constructing a model separately for each tissue type could dramatically reduce the number of samples. It has been shown that cancer type-level analysis showed a stronger predictive power of mutation information [9], and regressing out the

effect of tissue-specific genes might allow the model to focus on genes that explain drug response behavior.

Integrating other types of omics data, such as genomic and epigenomic data, in a meaningful manner can enrich information to explain drug response heterogeneity. Although CaDRReS was among the top performing models, it still considered only expression of essential genes. We previously tested whether integrating mutational status of genes could improve the performance, but we found that the performance was not significantly improved. This is probably because of the lower predictive performance of genomic data due to the low occurrence of mutations. Beside individual mutated genes, a combination of mutations and relationships between genomic and transcriptomic information could also explain drug response mechanisms. However, the 299 cancer genes or the comprehensive list of genes predicted by ConsensusDriver could help us to identify actionable driver genes (targets of anticancer drugs) in the patients, and these driver genes could be added into a model to further improve the performance and interpretability of drug response prediction.

In this thesis, we proposed CaDRReS, a recommender system that can predict patient-specific drug response based on the transcriptomic profile. The pharmacogenomic space of CaDRReS allows us to interpret and visualize the results and understand drug response mechanism. A modified version, CaDRReS-SC, can be applied to single-cell data to capture intratumoral drug response heterogeneity and to suggest a combination of drugs to inhibit multiple cell types identified within a tumor.

Constructing computational models to directly predict the best drug for a specific patient might not be able to accomplish in the near future due to multiple complex factors that alter drug response behaviors. Still, such the models can be

served as a decision support system for suggesting a small number of drugs to be extensively studied using *in vitro* cell lines or *in vivo* models. Also, the interpretability of the model can help us to understand drug response mechanisms and could enable drug discovery and repositioning.

Bibliography

1. Cancer Statistics - National Cancer Institute. Available at: https://www.cancer.gov/about-cancer/understanding/statistics.
2. The Cancer Genome Atlas (TCGA). *National Cancer Institute and National Human Genome Research Institute* Available at: http://cancergenome.nih.gov/.
3. International Cancer Genomc Consortium (ICGC). *International Cancer Genomc Consortium*
4. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
5. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).
6. Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173**, 305-320.e10 (2018).
7. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–7 (2012).
8. Broad Institute. Cancer Therapeutics Response Portal. Available at: https://portals.broadinstitute.org/ctrp.v2.1/.
9. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).
10. Geeleher, P., Cox, N. J. & Huang, R. S. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.* **15**, R47 (2014).
11. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1–103 (2014).
12. Khan, S., Malani, D., Murumägi, A. & Kallioniemi, O. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. (2016).
13. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–24 (2009).
14. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–58 (2013).
15. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–20 (2013).
16. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic

mutations in human cancer. *Nucleic Acids Res.* **43**, D805-811 (2014).

17. Marx, V. Cancer genomes: discerning drivers from passengers. *Nat. Methods* **11**, 375–379 (2014).

18. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–8 (2013).

19. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

20. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–44 (2013).

21. Pon, J. R. & Marra, M. A. Driver and Passenger Mutations in Cancer. *Annu. Rev. Pathol. Mech. Dis.* **10**, 25–50 (2015).

22. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).

23. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–9 (2010).

24. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).

25. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* **69**, 6660–7 (2009).

26. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).

27. Cerami, E., Demir, E., Schultz, N., Taylor, B. S. & Sander, C. Automated network analysis identifies core pathways in glioblastoma. *PLoS One* **5**, e8918 (2010).

28. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2014).

29. Chen, Y. *et al.* Identifying potential cancer driver genes by genomic data integration. *Sci. Rep.* **3**, 3538 (2013).

30. Miller, C. A., Settle, S. H., Sulman, E. P., Aldape, K. D. & Milosavljevic, A. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med. Genomics* **4**, 34 (2011).

31. Leiserson, M. D., Wu, H.-T., Vandin, F. & Raphael, B. J. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* **16**, 160 (2015).

32. Babur, Ö. *et al.* Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biol.* **16**, 45 (2015).

33. Bertrand, D. *et al.* Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res.* **43**, e44 (2015).

34. Bashashati, A. *et al.* DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* **13**, R124 (2012).

35. Hou, J. P. & Ma, J. DawnRank: discovering personalized driver genes in cancer. *Genome Med.* **6**, 56 (2014).

36. Tamborero, D., Lopez-Bigas, N. & Gonzalez-Perez, A. Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. *PLoS One* **8**, e55489 (2013).

37. Aure, M. R. *et al.* Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. *PLoS One* **8**, e53014 (2013).

38. Akavia, U. D. *et al.* An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–17 (2010).

39. Louhimo, R. & Hautaniemi, S. CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics* **27**, 887–8 (2011).

40. Sanchez-Garcia, F. *et al.* Integration of Genomic Data Enables Selective Discovery of Breast Cancer Drivers. *Cell* **159**, 1461–75 (2014).

41. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237-45 (2010).

42. Ng, S. *et al.* PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* **28**, i640–i646 (2012).

43. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).

44. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).

45. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169 (2012).

46. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–2 (2013).

47. Cheng, W.-C. *et al.* DriverDB: an exome sequencing database for cancer driver gene identification. *Nucleic Acids Res.* **42**, D1048-54 (2014).

48. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9 (2014).

49. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–5 (2014).

50. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–43 (2015).

51. Liu, Y., Tian, F., Hu, Z. & DeLisi, C. Evaluation and integration of cancer gene classifiers: identification and ranking of plausible drivers. *Sci. Rep.* **5**, 10204 (2015).

52. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J.*

Med. **372**, 793–795 (2015).

53. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6**, 813–823 (2006).

54. Azuaje, F. Computational models for predicting drug responses in cancer research. *Brief. Bioinform.* (2016).

55. McLeod, H. L. Cancer Pharmacogenomics: Early Promise, But Concerted Effort Needed. *Science (80-.).* **339**, (2013).

56. Wheeler, H. E., Maitland, M. L., Dolan, M. E., Cox, N. J. & Ratain, M. J. Cancer pharmacogenomics: strategies and challenges. *Nat. Rev. Genet.* **14**, 23–34 (2012).

57. Ding, Z., Zu, S. & Gu, J. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics* (2016).

58. Chen, T. & Sun, W. Prediction of cancer drug sensitivity using high-dimensional omic features. *Biostatistics* (2016).

59. Cortés-Ciriano, I. *et al.* Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics* **32**, 85–95 (2015).

60. Dong, Z., Zhang, N., Li, C., Wang, H. & Fang, Y. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC* (2015).

61. Gupta, S. *et al.* Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine. *Sci. Rep.* **6**, 23857 (2016).

62. Menden, M. P. M. *et al.* Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS One* **8**, (2013).

63. Koren, Y., Bell, R. & Volinsky, C. Matrix factorization techniques for recommender systems. *Computer (Long. Beach. Calif).* (2009).

64. Bennett, J. & Lanning, S. The netflix prize. *Proc. KDD cup Work.* (2007).

65. Sheng, J., Li, F. & Wong, S. Optimal drug prediction from personal genomics profiles. *IEEE J. Biomed.* (2015).

66. Wang, L., Li, X., Zhang, L. & Gao, Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* **17**, 513 (2017).

67. Van Allen, E. M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat. Med.* **20**, 628–688 (2014).

68. Ghazani, A. A. *et al.* Assigning clinical meaning to somatic and germ-line whole-exome sequencing data in a prospective cancer precision medicine study. *Genet. Med.* **19**, 787–795 (2017).

69. Geeleher, P. *et al.* Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Res.* **27**, 1743–1751 (2017).

70. Suphavilai, C., Bertrand, D., Nagarajan, N. & Wren, J. Predicting Cancer Drug Response using a Recommender System. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty452

71. Plummer, M. JAGS: A program for analysis of Bayesian graphical models

using Gibbs sampling. *Proc. 3rd Int. Work.* (2003).

72. Kruschke, J. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. (2014).

73. Sebaugh, J. L. Guidelines for accurate EC50/IC50 estimation. *Pharm. Stat.* **10**, 128–134 (2011).

74. Chia, S. *et al.* Phenotype-driven precision oncology as a guide for clinical decisions one patient at a time. *Nat. Commun.* **8**, 435 (2017).

75. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–101 (2015).

76. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–93 (2013).

77. Haverty, P. *et al.* Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* (2016).

78. Pedregosa, F., Varoquaux, G. & Gramfort, A. scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2825–2830 (2011). Available at: http://scikit-learn.org/.

79. Hafner, M., Niepel, M., Chung, M. & Sorger, P. K. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat. Methods* **13**, 521–527 (2016).

80. Safikhani, Z. *et al.* Revisiting inconsistency in large pharmacogenomic studies. *F1000Research* **5**, 2333 (2016).

81. Rahman, S. *et al.* Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminform.* **1**, 12 (2009).

82. Fry, D. W. *et al.* Specific inhibition of cyclin-dependent kinase 4/6 by PD 0332991 and associated antitumor activity in human tumor xenografts. *Mol. Cancer Ther.* **3**, 1427–38 (2004).

83. Ma, P. C. *et al.* Downstream signalling and specific inhibition of c-MET/HGF pathway in small cell lung cancer: implications for tumour invasion. *Br. J. Cancer* **97**, 368–377 (2007).

84. Quevedo, C., Alcázar, A. & Salinas, M. Two different signal transduction pathways are implicated in the regulation of initiation factor 2B activity in insulin-like growth factor-1-stimulated neuronal cells. *J. Biol. Chem.* **275**, 19192–7 (2000).

85. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

86. Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* (2008).

87. Stinchcombe, T. E. & Johnson, G. L. MEK inhibition in non-small cell lung cancer. *Lung Cancer* **86**, 121–5 (2014).

88. El-Naggar, S., Liu, Y. & Dean, D. C. Mutation of the Rb1 pathway leads to overexpression of mTor, constitutive phosphorylation of Akt on serine 473, resistance to anoikis, and a block in c-Raf activation. *Mol. Cell. Biol.* **29**, 5710–7 (2009).

89. Normanno, N. *et al.* Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene* **366**, 2–16 (2006).

90. McCubrey, J. A. *et al.* Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance. *Biochim. Biophys. Acta - Mol.*

Cell Res. **1773**, 1263–1284 (2007).

91. Liu, L. *et al.* Sorafenib blocks the RAF/MEK/ERK pathway, inhibits tumor angiogenesis, and induces tumor cell apoptosis in hepatocellular carcinoma model PLC/PRF/5. *Cancer Res.* (2006).

92. Fujita, N. *et al.* MTA3, a Mi-2/NuRD Complex Subunit, Regulates an Invasive Growth Pathway in Breast Cancer. *Cell* **113**, 207–219 (2003).

93. Shih, I.-M. & Wang, T.-L. Notch Signaling, gamma-secretase Inhibitors, and Cancer Therapy. *Cancer Res.* **67**, 1879–1882 (2007).

94. Harari, P. M. Epidermal growth factor receptor inhibition strategies in oncology. *Endocr. Relat. Cancer* **11**, 689–708 (2004).

95. Medina, P. & Goodin, S. Lapatinib: A dual inhibitor of human epidermal growth factor receptor tyrosine kinases. *Clin. Ther.* **30**, 1426–1447 (2008).

96. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–337 (2013).

97. Wang, M. *et al.* Role of tumor microenvironment in tumorigenesis. *J. Cancer* **8**, 761–773 (2017).

98. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).

99. Li, B. *et al.* Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17**, 174 (2016).

100. Newman, A. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

101. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).

102. Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. & Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.* **52**, 155–173 (2007).

103. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

104. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

105. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* **14**, 535–571 (2017).

106. Liston, D. R. & Davis, M. Clinically relevant concentrations of anticancer drugs: a guide for nonclinical studies. *Clin. Cancer Res.* **23**, 3489–3498 (2017).

107. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2017).

108. Bedard, P. L., Hansen, A. R., Ratain, M. J. & Siu, L. L. Tumour heterogeneity in the clinic. *Nature* **501**, 355–364 (2013).

109. Crown, J., O'Leary, M. & Ooi, W.-S. Docetaxel and paclitaxel in the treatment of breast cancer: a review of clinical experience. *Oncologist* **9**, 24–32 (2004).

110. LoRusso, P. M. *et al.* Phase I and Pharmacokinetic Study of Lapatinib and Docetaxel in Patients With Advanced Cancer. *J. Clin. Oncol.* **26**, 3051–3056 (2008).

111. Jia, J. *et al.* Mechanisms of drug combinations: interaction and network perspectives. *Nat. Rev. Drug Discov.* **8**, 111–28 (2009).

112. Gentles, A. J. *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21**, 938–945 (2015).

113. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).

114. Choi, S. Y. C. *et al.* Lessons from patient-derived xenografts for better in vitro modeling of human cancer. *Adv. Drug Deliv. Rev.* **79–80**, 222–237 (2014).

115. Yang, X., Guo, Y., Liu, Y. & Steck, H. A survey of collaborative filtering based social recommender systems. *Comput. Commun.* (2014).

116. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**, 20170387 (2018).