Predicting Cancer Drug Effectiveness with Deep Learning Artificial Intelligence

by

Michael Qi Ding

Bachelor of Arts, Harvard University, 2012

Master of Science, University of Pittsburgh, 2017

Submitted to the Graduate Faculty of the

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Michael Qi Ding

It was defended on

April 7, 2020

and approved by

Dr. Gregory Cooper, Professor, Biomedical Informatics

Dr. Douglas Landsittel, Professor, Biomedical Informatics

Dr. Xiang-Qun Xie, Professor, Pharmacy

Dissertation Director: Dr. Xinghua Lu, Professor, Biomedical Informatics

Predicting Cancer Drug Effectiveness with Deep Learning Artificial Intelligence

Michael Qi Ding, PhD

University of Pittsburgh, 2020

Despite advances in molecular technologies, application of precision medicine in clinical oncology remains difficult. Although widely used, single-gene biomarkers are imperfect predictors for the effective administration of targeted therapy. Meanwhile, nonspecific cytotoxic medications lack established biomarkers to guide their usage, yet they remain first-line chemotherapy for many patients. As the formulary of precision medications and tissue-agnostic treatments expands, there is a pressing and growing need for sensitive and specific companion diagnostic testing. The effective application of powerful computational techniques holds great potential for assisting clinicians and patients in the navigation of ever-increasingly complex treatment decisions.

To address this challenge, we applied state of the art deep learning techniques and modern machine learning frameworks to develop a deep neural network autoencoder for learning latent representations of integrated omics data from cancer cell lines. We used these representations to build predictive models of drug sensitivity. We evaluated the effectiveness of these models using a variety of preclinical and clinical data to assess potential for translational impact.

This research is significant in three primary ways. First, we developed a novel data-driven approach to precision medicine. Second, we demonstrated the potential for this approach to improve clinical outcomes relative the current standard of care. Third, we demonstrated that this approach not only optimizes therapeutic efficacy in preclinical cancer models, but is also generalizable to real, clinical tumors.

Table of Contents

Preface xii

1.0 Introduction 1

1.1 Description of the Problem 1

1.2 Significance of this Research 2

2.0 Background 4

2.1 Brief History of Clinical Oncology 4

2.2 Overview of Drug Screening Experiments for Cancer Chemotherapy 10

2.3 Current Methods in Cancer Pharmacogenomics 19

2.4 Preliminary Work 27

2.4.1 Methods 27

2.4.1.1 Data Retrieval and Feature Engineering 27

2.4.1.2 Developing Latent Representations of Omics Data Using Deep Learning 29

2.4.1.3 Predicting Drug Sensitivity with Latent Representations 30

2.4.2 Results 31

2.4.3 External Validation 34

2.4.4 Conclusions 36

3.0 Research Design 37

3.1 Research Questions 37

3.2 Dissertation Overview 38

4.0 Aim 1: Deep Learning Optimization 40

4.1 Data Retrieval and Feature Engineering 40

4.1.1 Methods 41

4.1.1.1 Predictive Feature Data 41

4.1.1.2 Drug Sensitivity Data 42

4.2 TensorFlow Implementation 42

4.2.1 Methods 43

4.2.1.1 Developing Latent Representations of Omics Data Using Deep Learning 43

4.2.1.2 Predicting Drug Sensitivity with Latent Representations 44

4.2.2 Results 44

4.2.3 Discussion 45

4.2.4 Limitations 46

4.3 Deep Neural Network Architecture Optimization 46

4.3.1 Methods 47

4.3.1.1 Candidate Autoencoder Structure Construction 47

4.3.1.2 Candidate Autoencoder Structure Evaluation 49

4.3.1.3 Optimal Autoencoder Structure Design and Evaluation 50

4.3.2 Results 51

4.3.2.1 Candidate Autoencoder Structure Evaluation 51

4.3.2.2 Optimal Autoencoder Structure Design and Evaluation 53

4.3.3 Discussion 56

4.3.4 Limitations 57

4.4 Deep Neural Network Regularization 58

4.4.1 Methods 59

4.4.2 Results 60

4.4.3 Discussion 60

4.4.4 Limitations 61

5.0 Aim 2: Determine Potential for Clinical Impact 63

5.1 Cancer Cell Line Trial 63

5.1.1 Methods 64

5.1.1.1 Standard of Care Modeling 64

5.1.1.2 AI-Supported Decision Modeling 66

5.1.2 Results 68

5.1.3 Discussion 71

5.1.4 Limitations 72

6.0 Aim 3: Explore Generalizability to Organoids 73

6.1 Gene Expression Normalization 74

6.1.1 Methods 75

6.1.1.1 Data Retrieval and Feature Engineering 75

6.1.1.2 Clustering 76

6.1.2 Results 76

6.1.3 Discussion 78

6.1.4 Limitations 79

6.2 Nonparanormal Transformation 80

6.2.1 Methods 80

6.2.2 Results 81

6.2.3 Discussion **82**

6.2.4 Limitations **82**

6.3 Organoid Drug Sensitivity Prediction **83**

6.3.1 Methods **83**

6.3.1.1 Data Retrieval and Feature Engineering **83**

6.3.1.2 Developing Latent Representations with Deep Learning **84**

6.3.1.3 Predicting Drug Sensitivity with Latent Representations **85**

6.3.2 Results **86**

6.3.3 Discussion **87**

6.3.4 Limitations **87**

7.0 Aim 4: Incorporate Clinical Data with Transfer Learning **89**

7.1 Transfer Learning **90**

7.1.1 Methods **90**

7.1.1.1 Predictive Feature Data **90**

7.1.1.2 Developing Latent Representations with Deep Learning **91**

7.1.1.3 Predicting Drug Sensitivity with Latent Representations **92**

7.1.2 Results **93**

7.1.3 Discussion **94**

7.1.4 Limitations **94**

7.2 Clinical Patient Survival Prediction with Cell Line Based Models **95**

7.2.1 Methods **96**

7.2.2 Results **97**

7.2.3 Conclusions **98**

7.2.4 Limitations 99
7.3 Clinical Patient Survival Prediction with Clinical Tumor Based Models 99
7.3.1 Methods 100
7.3.1.1 Predictive Feature Data 100
7.3.1.2 Developing Latent Representations with Deep Learning 101
7.3.1.3 Predicting Drug Sensitivity with Latent Representations 102
7.3.2 Results 102
7.3.3 Conclusions 103
7.3.4 Limitations 104
8.0 Final Conclusions and Future Work 105
8.1 Deep Learning Autoencoders Enable Powerful Predictive Modeling 106
8.2 Accurate Predictive Models Exhibit Significant Potential Impact 107
8.3 Generalizable Predictive Models Function in Other Preclinical Settings 108
8.4 Transfer Learning Unlocks Translation to Clinical Setting 108
8.5 Future Work 109
Appendix A Clinical Tumor Response Prediction with Cell Line Based Models 111
Appendix A.1 Methods 111
Appendix A.2 Results 113
Appendix A.3 Conclusions 114
Appendix A.4 Limitations 115
Bibliography 116

List of Tables

Table 1: Large pharmacogenomics experiments. 16

Table 2: Autoencoder network architecture optimization. 54

Table 3: A simplified standard of care for NSCLC. 65

Table 4: A simplified standard of care for ADT cancers. 65

List of Figures

Figure 1: Learning cellular states using deep learning. 32

Figure 2: External validity of predictive models..... 35

Figure 3: Dissertation overview. 38

Figure 4: Learning cellular states with different autoencoder implementations..... 45

Figure 5: Autoencoder network architecture optimization. 52

Figure 6: Learning cellular states using an optimized autoencoder. 55

Figure 7: Learning cellular states using regularized autoencoders. 60

Figure 8: Evaluating the NSCLC standard of care with a cell line trial. 69

Figure 9: Evaluating the aero-digestive tract cancer standard of care with a cell line trial.70

Figure 10: K-means clustering of a combined GDSC and TCGA dataset. 77

Figure 11: Learning cellular states using transformed expression data..... 81

Figure 12: External validation on pancreatic cancer cell lines and bladder, colorectal, and liver cancer organoids..... 86

Figure 13: Learning cellular states using a TCGA autoencoder. 93

Figure 14: Predicting overall survival of TCGA lung cancer patients. 98

Figure 15: Predicting overall survival of TCGA colorectal cancer patients. 103

Figure 16: Predicting clinical FOLFIRI sensitivity from cell line based deep learning models for individual component medications..... 113

Preface

I am grateful to the National Library of Medicine and the Biomedical Informatics Training Program at the University of Pittsburgh for their financial support and the opportunity to earn two degrees; Dr. Wendy Mars and Dr. Rebecca Jacobson who first gave me the opportunity to transition from molecular biology to informatics; Dr. Gregory Cooper, Dr. Douglas Landsittel, Dr. Xiang-Qun Xie, Dr. Kayhan Batmanghelich, Dr. Harry Hochheiser, Dr. Gerald Douglas, and Dr. Roger Day for their guidance and tutelage through various projects; Dr. Vanathi Gopalakrishnan, Dr. David Boone, Dr. Songjian Lu, Dr. Richard Boyce, Dr. Tanja Bekhuis, Dr. Shyam Visweswaran, Dr. Katrina Romagnoli, and Dr. Michael Becich for their mentorship and support during my time in the department; and finally, Dr. Xinghua Lu for his unwavering enthusiasm, wisdom, and encouragement through my master's thesis and PhD dissertation.

Thank you to the members of the Lu laboratory, especially Dr. Jonathan Young, Dr. Chunhui Cai, Dr. Joyeeta Dutta-Muscato, and Samuel Ding for their contributions. Thank you to the staff who supported this work: Toni Porterfield, Bill Shirey, Victoria Khersonsky, Genine Bartolotta, Linda Mignogna, Cleat Szczepaniak, Barbara Karnbauer, Lucy Cafeo, and Rob Cecchetti. Thank you also to fellow students and friends in the department with whom I have worked in some capacity: Dr. Arielle Fisher, Dr. Andrew King, Dr. Amie Barda, Dr. Brian Liu, Dr. Jenna Schabdach, Lauren Rost, and Adriana Johnson.

Thank you to the long line of mentors who have guided my research career from the very beginning: Dr. Agnes Kane, Dr. David Lombard, Dr Bjoern Schwer, Dr. Frederick Alt, Dr. George Michalopoulos, and Dr. Wendy Mars. Lastly, special thanks to my wonderful family, without whose support this journey would not have been possible.

"Everything should be made as simple as possible, but not simpler."

-Albert Einstein

1.0 Introduction

Precision oncology, the practice of using data from analyses of biomarkers and next-generation sequencing to guide therapies, has the potential to revolutionize cancer medicine [1]. Over time, usage of the term has transitioned from restrictive descriptions of specific targeted therapies, such as tyrosine kinase inhibitors, to encompass the general prospect of directing therapy based on genomic profiles, independent of anatomically or histologically defined cancer types [2]. The success of this endeavor depends on two components: the identification of genomic alterations that drive individual tumors, and the development of effective methodologies for matching effective targeted therapies to these alterations once identified [3].

1.1 Description of the Problem

In the current practice of precision oncology, the prescription of molecularly targeted therapies is mainly based on the altered genomic status of a drug-target gene as a therapeutic indicator, but this approach benefits only a small percentage of patients. The cause is two-fold. First, the percentage of tumors with actionable targeted therapies using approved agents under current clinical guidelines is generally estimated to be in the single digits [4]. Second, once administered according to these guidelines, the effectiveness of a therapy is not guaranteed [5]. Most molecular targeted agents only partially inhibit the intended signaling pathway, and many suffer from poor selectivity, affecting up to as many as 17 different known targets. In contrast, nonspecific cytotoxic medications lack well-established biomarkers to guide their usage, but they

remain first-line chemotherapy for many patients [4]. The development of effective companion diagnostic tests for guiding treatment is critical for effectively utilizing the existing armamentarium of targeted and non-targeted medications, as well as for incorporating future advances in molecularly targeted therapy and immunotherapy.

Recent large-scale pharmacogenomics screening on cancer cell lines [6, 7] and patient-derived xenografts [8] has explored the effectiveness of both molecular targeted and nonspecific therapies in inhibiting the growth of in vitro and in vivo cancer models. These studies found that for the majority of molecularly targeted drugs, genomic markers are not accurate indicators. Not only did some cancer samples resist medications that genomic markers indicated should be effective, there existed other samples that were found to be sensitive to molecular targeted drugs even though genomic status of the corresponding target genes were wild type.

Translated into a clinical setting, these findings suggest that the presence of a single genetic indicator may not be sufficient for administering a molecular targeted therapy, and also that additional patients beyond those with aberrant genomic markers may potentially benefit from receiving such a targeted treatment. Accurately matching patients with effective therapies would both expand the application of existing anti-cancer drugs as well as reduce the rate of ineffective therapy. Achieving this goal requires progressing beyond the use of conventional genomic markers and utilizing modern advances in the collection and analysis of genome-scale omics data.

1.2 Significance of this Research

Clinical oncologists require companion diagnostic tests that provide sensitive and specific predictions about the potential effectiveness of a medication for a given patient [9]. Current clinical

guidelines regarding the administration of molecular targeted chemotherapies generally take into consideration some clinical variables and the genomic status of a small number of single nucleotide polymorphisms. These rule-based systems lack the capacity to accurately model the complex processes that give rise to cancer and have limited utility as predictors for drug effectiveness. The incorporation of a wider array of genetic and genomic information is key to building better suited models that can properly balance bias and variance. In this research, we apply a deep learning approach to utilize integrated omics data to significantly improve the practice of drug sensitivity prediction.

This dissertation recounts the journey of the creation and validation of a novel platform technology for the effective utilization of rapidly expanding molecular healthcare data in clinical care. Aside from the development of the technology itself, this research represents significant contributions to knowledge regarding the effective administration of chemotherapy and to theory of the optimization and behavior of deep neural networks. It is our deepest hope that this research will someday also contribute to improving the practice of precision medicine.

2.0 Background

2.1 Brief History of Clinical Oncology

The clinical practice of treating cancer dates back to at least 3000 BC, when the earliest known recorded cases of cancer were documented in Ancient Egypt [10]. The disease was not well-understood at the time, and although the Egyptians and contemporaneous civilizations attempted a range of remedies ranging from cauterization to arsenic and metallic pastes, there was no effective treatment. As knowledge and understanding of cancer grew due to advancements in technology and biological theory, the shape of clinical oncology evolved alongside it.

In Ancient Greece around 300 BC, Hippocrates and his students rejected superstitious explanations for cancer and argued that it was a natural disease. Maintaining consistency with the humoral theory, they posited that cancer was caused by an imbalance of black bile, yellow bile, phlegm, and blood. A distinction was made between surface lesions and deep cancers, the former being treated with cauterization and ointments, while the latter, if operable, were removed by surgery [11].

Shortly after the turn of the first millennium, the first description of what is now understood to be metastasis was recorded in Rome [12]. Unfortunately, the Romans did not agree with the Greeks in using surgery for cancer, and instead favored the use of purgatives to rebalance bodily humors, delaying progress for centuries. After the fall of the Roman empire, surgery as treatment for cancer spread and grew throughout Europe and the Middle East during the Dark Ages. This time period saw locally isolated innovations such as the removal of entire organs to prevent recurrence, and the initial development of wide excision techniques [13].

The invention of the printing press in 1450 signaled the end of the Dark Ages, and the resulting Renaissance movement removed many societal barriers that had hindered scientific progress. In medicine, increasing practice of autopsy and subsequent publication of findings in case reports led to rejection of the humoral theory [14]. Advances in anatomy and physiology coincided with the identification and characterization of cancers from various sites of origin, including the brain, lung, breast, colon, liver, cervix, and prostate, forming a very early foundation for modern understanding of tumor pathology [15]. These insights led to the refinement of earlier surgical techniques, and the development of procedures including radical mastectomy and lumpectomy [16].

The invention of the microscope in 1590 led to the examination of surgical samples by microscope and the growing adoption of microscopy in the diagnosis of cancer. The resulting descriptions of tumor physiology informed the creation of guidelines for operative surgery [17]. It was around this time that doctors began to propose environmental causes for cancer. The first two hypothesized carcinogens were tobacco [18] and chimney soot [19]. With the advent of cell theory in 1838 came the understanding that cancers are formations of cells in diseased organs with the ability to spread to other parts of the body through the vascular system [20]. Because cells can only arise from other cells, it was concluded that cancers must develop from normal tissue. It was quickly established that primary tumors can develop in every organ system, each with unique clinical and microscopic features. A medical movement towards subspecialization spurred the development of new tools and techniques to facilitate organ-specific studies and surgeries. Continued accumulation of knowledge in cancer microscopy led to the histological identification and classification of tumor subtypes.

The turn of the 19th and 20th centuries brought several major advances in the clinical treatment of cancer. Advances in anesthesia and sterile technique enabled the practice of more advanced surgical procedures [21]. In 1895, Wilhelm C. Rontgen discovered X-rays, creating the field of diagnostic radiology [22]. In the following years, X-rays were found to be both carcinogenic and highly destructive to living tissue. This discovery, combined with the isolation of radium by Pierre and Marie Curie in 1898, led to the development of the earliest forms of radiation therapy [23]. In 1909, Paul Ehrlich published the first book on chemotherapy, detailing the results of experiments in rodents [24]. Critically, he observed that neoplasms are composed of a combination of sensitive and resistant cells, causing nonuniform response to chemical treatment.

Over the following decades, scientists continued to search for causes of cancer [25], finding evidence for parasites [26], viruses [27], hormones [28], and pollutants [29], as potential triggers for the disease. From 1919 to 1940, James Ewing published four editions of his definitive textbook on neoplastic diseases [30]. Ewing believed cancers were caused by repeated damage of tissue by chemical and environmental agents. He also demonstrated that the malignancy of a tumor could be correlated to a number of histopathological features. These findings led to the eventual development of modern histologic grading systems.

The same years saw the development and adoption of smear techniques for collecting and visualizing cells from tumors at all body sites for diagnosis [31]. Combined with the microscopic examination of bodily fluids and secretions [32], this established the practice of cytology. Pioneering work by the likes of Theodor Kocher, William Halsted, and Harvey W. Cushing contributed to the continued specialization of surgery [33]. Large reductions in operative mortality were observed as a result, causing radical surgery to remain the treatment of choice for both primary and metastatic cancers. Although progress was being made in the nascent fields of

radiation therapy and chemotherapy, early successes were difficult to come by. Radiation therapy suffered from technical issues in instrumentation and manufacturing, as well from the unexplainable phenomenon of radioresistance [34]. The earliest chemotherapeutic agents, which included nitrogen mustard, were not favored due to extreme toxicity [35].

The elucidation of the structure of DNA by James Watson, Francis Crick, Rosalind Franklin, and Maurice Wilkins in 1953 opened the door to the modern understanding of the etiology of cancer. Prior to this discovery, many environmental causes for cancer had already been identified, typically through the observation of increased incidence of a specific type of cancer in a population with high exposure to a carcinogen. These included leukemia in radiologists [36], and lung carcinoma in smokers [37], asbestos workers [38], and pesticide factory employees [39]. However, it was only after the physical structure of the DNA molecule was determined that progress in genetics enabled the experiments with tumor viruses [40-42] and heredity studies [43, 44] that separately began to suggest the importance of the genome in oncogenesis.

Advances in cytology, including the clinical adoption of the Pap smear led to increased rates of and earlier detection of cancer in patients [45]. The resulting proliferation of clinical samples that required both proper identification and causal explanation created the field of surgical pathology. Over the next several decades, clinical tumor samples were extensively and systematically examined, diagnosed, and classified, creating the foundation of the modern organizational classification of cancers [46].

Continued tests of highly toxic chemotherapeutic agents such as urethane yielded only short remissions in patients until Sidney Farber demonstrated the efficacy of a folic acid antagonist in treating acute leukemia in 1948 [47]. This initial success in treating a systemic condition sparked a search for antimetabolites and antibiotics for the treatment of metastatic tumors, leading to

estrogen therapy for metastatic carcinoma of the prostate [48], and methotrexate for metastatic choriocarcinoma [49]. Advances in radiation therapy saw the replacement of radium with cobalt and proton beams [50]. During this time, radiation was determined to be particularly effective as a first-line treatment for lymphomas and started to become incorporated into adjuvant therapy in combination with surgery. Meanwhile, radical surgical procedures fell out of favor as continued technical progress made removing only the cancerous parts of affected tissue not only possible but also effective [51]. Early attempts at creating a vaccine against cancer were unsuccessful, but progress was made in the search for cancer-specific antigens, despite the relative lack of understanding of immunological processes at this time [52].

Despite an accumulation of evidence for environmental exposures including viruses, chemicals, and radiation as causes for cancer, the underlying molecular process of oncogenesis did not begin to become clear until 1976, when Harold E. Varmus and J. Michael Bishop determined that certain retroviruses can transform proto-oncogenes, specific genes in normal cells, into oncogenes responsible for the growth of cancer [53]. Over time, it was determined that many different genomic modifications, including translocation, amplification, and deletion could create oncogenes. Genes that control the process of cell division came to be known as tumor-suppressor genes, because they counteract the proliferative tendency of malignant tumors. Deactivation of such genes through genomic mutation could lead to cancer, making them oncogenes, as well [54]. One such gene and its associated product, tumor protein 53, was discovered in 1979. It has since been found to be the most frequently mutated gene in human cancer [55].

Angiogenesis was determined to be an important component of cancer growth [56]. With the discovery of tumor angiogenesis factors such as vascular endothelial growth factor and prostaglandin came the identification and therapeutic use of angiogenesis inhibitors such as

thalidomide [57]. This was the earliest practice of targeted therapy – the administration of a chemotherapeutic agent intended to act upon a specific gene or gene product. New chemotherapeutic agents, both cytotoxic and targeted, were developed and tested at an astonishing rate in the 1970s and 1980s [58]. Successful compounds ultimately required evaluation in humans for regulatory approval, leading to the formalization of a testing framework consisting of four phases of randomized clinical trials [59]. Effective primary chemotherapy regimens were incorporated into preoperative and postoperative adjuvant therapy alongside surgery and radiation [60]. Combination chemotherapy consisting of multiple compounds administered together were tested and found to be appropriate for certain cancers [61].

The 1980s and 1990s saw great advances in the practice of immunotherapy, made possible by lessons learned from progress in immunology. Experiments with monoclonal antibodies revealed that activated immune cells secreted cytokines as an antiproliferative response, leading to the therapeutic application of interleukins and interferons [62]. Further discoveries regarding the natural immune response to cancer led to the identification of additional targets for immunotherapy. Eventually, trastuzumab was the first therapeutic antibody approved for clinical use in 1998 [63].

In 2001, regulatory approval of the tyrosine kinase inhibitor imatinib for the treatment of chronic myeloid leukemia signaled the arrival of precision oncology as the latest development in clinical oncology [64]. Treatments tailored to a patient's specific tumor promised to be more effective and have fewer side effects than traditional systemic chemotherapy. Initially, efforts were focused on finding or developing small-molecule inhibitors for a small number of known oncogenic targets. Over time, as understanding of the molecular processes of cancer grew, the search expanded to cover more molecules and more targets [6, 7].

Most recently, precision oncology has expanded to include personalized immunotherapy. In 2017, chimeric antigen receptor T-cell therapy, in which a patient's own immune cells are harvested and modified outside the body to fight cancer, was approved to treat B-cell cancer [65]. In 2018, the antibody pembrolizumab was approved for the treatment of solid tumors with microsatellite instability or mismatch repair deficiency [66]. This is the first approval of a cancer treatment based solely on genetics, and its administration depends not on the tissue type or body site of the target tumor, but on the results of a companion diagnostic test [67]. This approval of a pioneering tissue-agnostic drug reflects an understanding that cancers of different tissues can share underlying activating mechanisms, potentially sensitizing them to the same treatments.

2.2 Overview of Drug Screening Experiments for Cancer Chemotherapy

When Paul Erlich published his work on small molecule screening, he detailed the search for a molecule effective against *Treponema pallidum*, the causative agent of syphilis [68]. Erlich's animal model of disease was a rabbit, but he performed initial screening of hundreds of different candidate treatments in mice infected with a different *Treponema* bacteria, for reasons including time and cost. The substitute worked; compound 606[1], which was effective in treating mice infected with trypanosomes, was subsequently found to also eradicate spirochete infections in chickens. Compound 606 went on to cure the rabbit, and ultimately succeeded in a clinical trial of 50 patients with late-stage syphilis. Erlich was extremely fortunate to be working at an institution with talented chemists and the equipment and expertise for high-throughput animal testing. Even

[1] Arsphenamine, also known as Salvarsan.

so, the screening experiment took five years to complete from its inception in 1904 to its publication in 1909. A more efficient system was clearly needed.

The earliest mouse model of cancer took the form of spontaneous tumors in inbred mouse lines that were selected for high susceptibility to cancer. Utilizing this model proved challenging, as each spontaneous tumor is distinct and thus must function as its own control [69]. In addition, the experimenter has no control over the creation of tumors, and it was often difficult to obtain spontaneous tumors in sufficient quantity to conduct a screening experiment. These shortcomings were addressed by the creation of transplantable tumor systems in rodents [70]. These systems enabled the transplantation of a single spontaneous tumor from one animal into many other animals. This enabled experimenters to transplant as many tumors as they needed at one time, with the added benefit that because the transplants were all from the same source, they would all be genetically identical, enabling comparability of drug responses between samples.

In 1935, Murray Shear used one such system, the murine S37 sarcoma model, to create an organized cancer drug screening program at the Office of Cancer Investigations of the United States Public Health Service, which would eventually merge with the National Institutes of Health Laboratory of Pharmacology to become the National Cancer Institute [71]. Over the course of the next 18 years, Shear's program would test over 3,000 compounds. Unfortunately, only two would make it into clinical testing, and they would be rejected due to extreme toxicity. Due to a lack of knowledge and understanding regarding the evaluation of toxic chemicals in humans, the program was brought to an end in 1953.

Those intervening years had seen the development of additional tumor transplantation systems. The S37 sarcoma model was replaced by the newer and more representative L1210 leukemia model [72] as the NCI's primary screening testbench. The L1210 contributed to progress

in therapy for acute lymphocytic leukemia and various lymphomas but did little to advance treatment for solid tumors [73]. It would remain in use until 1975, when it was supplanted by a panel of tumors utilizing new xenograft technology [74]. These xenografts better represented human tumors than murine transplantation models because they were derived from human cancer tissue, but screening remained slow and costly as they were still in vivo models. In 1985, the xenograft models were replaced by a panel of 60 primary tumor cell lines [75]. Advances in culture technique and cell line immortalization had finally made it possible to grow tumor models in vitro, greatly reducing the time and cost of performing drug screening experiments [76]. The NCI-60 screen, as it was called, continues to operate, screening 3000 small molecules per year, thus evaluating potential drug-cancer pairings at a rate over 1000 times that of Shear's S37 program.

The drug screening methods described thus far have been examples of phenotypic drug discovery (PDD). Under this modality, candidate molecules, often randomly selected, are evaluated in a battery of cells, tissues, or animals to determine whether or not they have the desired effect. The molecular mechanism of action of a successful molecule, or its target, is typically only determined later, after it has been established that the molecule works [77]. The advantage of PDD is that it does not require a strong understanding of the disease being treated, nor does it depend on extensive knowledge of the compounds being tested. In this way, it is a reflection of the field of cancer chemotherapy until relatively recently.

The alternative to PDD is target-based drug discovery (TDD). Under this alternate modality, candidate molecules are first selected based on their association with an entity that is known to be important to the disease. This association can be measured in many ways, ranging from simple enzymatic activity assays on one end to evaluation of complex protein-protein interactions on the other [78]. This selection step can typically be run in an extremely high

throughput manner, enabling the pre-screening of millions of candidate molecules in vitro before a select few are enhanced and evaluated in a more time and cost intensive manner in the disease model. This increased throughput is the primary advantage of TDD.

The discovery of the *BCR-ABL* fusion tyrosine kinase inhibitor imatinib is a major success story in TDD. Initially, a lead compound was identified in a screen for inhibitors of protein kinase C [79]. During optimization of the molecule through the addition and subtraction of organic functional groups, a variant was discovered that traded inhibition of protein kinase C for inhibition of tyrosine kinases. After further modification to improve delivery and bioavailability, the molecule now known as imatinib was selected for clinical development [80]. After success in cell line experiments, mouse models, and clinical trials, imatinib was approved for clinical use [81].

The success of imatinib, in the words of one author, "is proof of principle that rationally designed, molecularly targeted therapy works. Imatinib represents a paradigm shift in cancer drug development. It is hoped that this will pave the way for a new generation of specific, targeted therapies" [82]. Such optimism pervaded the field of drug discovery as PDD was abandoned for TDD virtually overnight. Unfortunately, the balms of Gilead failed to materialize. Instead of an explosion of designer targeted therapy, the ensuing years actually saw a decline in the number of candidate treatments progressing through clinical trials and eventually entering the market, despite increased investment in research and development [83]. Given the supposedly higher throughput of the new methodology, something was amiss.

TDD relies on a critical assumption that the target and assay chosen for the screening process are appropriate for identifying potential drug candidates. Although these decisions are made carefully and based on a current understanding of the disease pathology, the process is necessarily reductionistic. By performing the screen in vitro, selection of candidate compounds is

made on the basis of interaction between the drug and the target. However, in the administration of the drug to treat disease, the interaction between the drug and the organism is what really matters [84]. Sometimes, such as in the case of imatinib, the approximation works. However, the relationship between the *BCR-ABL* oncogene and chronic myeloid leukemia is a special case. More than 95% of people with the disease have the fusion protein targeted by imatinib [85]. In contrast, the underlying etiology of most other cancers is significantly more complicated. TDD was struggling to properly identify candidate molecules for these more complex disease states.

An analysis of first-in-class new molecular entities (NMEs) approved by the United States Food and Drug Administration (FDA) between 1999 and 2008 revealed that 28 approved medications were discovered using a PDD framework, while only 17 were found via TDD [86]. A subsequent study examining 48 NMEs approved between 2008 and 2013 found that the number of drugs found by PDD had declined to four, consistent with the framework's fall from favor. Meanwhile, TDD had contributed 29 medications. Upon closer inspection, 21 of those discoveries were tyrosine kinase inhibitors, which follow a relatively straightforward molecular mechanism of action (MMOA). If these were removed from consideration, then TDD had only discovered eight new compounds. Thus, the contributions of PDD and TDD during this time period were overshadowed by hybrid modes of drug discovery, which combined elements of both PDD and TDD to find 15 new therapies [77].

One such new framework, called mechanism-informed phenotypic drug discovery (MIPDD) involves the design of a drug-target interaction assay that functions in the context of a cell culture. In this manner, the screen is similar to PDD in that it takes place in vivo, but produces rationally designed candidates with known MMOAs, like TDD. The main drawback of MIPDD is that theoretically a compound that achieves the desired phenotype may not do so by acting through

the intended pathway. Despite this, MIPDD has proven useful especially for the development of second generation therapies [87]. Ultimately, PDD has proven to be the superior method for the identification of first-in-class therapies, while TDD/MIPDD is better suited for the development of next-generation best-in-class therapies based on known MMOAs[2].

With the development and refinement of microarrays, followed by next-generation sequencing techniques, it has become increasingly affordable to collect a large amount of genomic, and transcriptomic data to characterize a biological sample. These technologies have been incorporated into cell line-based drug screening experiments with the dual purpose of validating and confirming the identity of cell cultures maintained for decades in the laboratory, as well as identifying genetic targets for potential therapy. Microarray assessment of RNA expression in the NCI-60 cell lines was first done in the year 2000 [88], and DNA mutation profiles were completed shortly thereafter [89].

PDD screening experiments in which a large number of compounds are evaluated on a large number of cell lines while genomic and transcriptomic data are collected have come to be known as large pharmacogenomic studies, due to their potential for generating knowledge relevant to pharmacogenomics. Pharmacogenomics describes interactions between genes and medications, and is particularly significant in cancer chemotherapy, due to the genetic nature of the disease [90]. A summary of the landscape of large pharmacogenomic experiments that have been conducted for cancer is presented in Table 1.

---

[2] It is tempting to attribute this to the first or early-mover advantage enjoyed by PDD, but the fact remains that TDD/MIPDD has consistently produced more follow-on next-generation therapies than novel first-in-class drugs, while the opposite is true for PDD.

Table 1: Large pharmacogenomics experiments.

Study Name	Study Type	RNA Expression	DNA Mutation	DNA Copy Number	Cell Lines	Drug Compounds	Release Year
NCI-60	Large Pharmacogenomic	Various Microarrays, RNA-Seq	Various SNP Arrays, Whole Exome NGS	Various SNP Arrays	59	>88,000	2016
GlaxoSmithKline	Large Pharmacogenomic	Affymetrix U133 Plus 2.0 Microarray	Affymetrix 500k SNP Array	Affymetrix 500k SNP Array	311	19	2010
Cancer Cell Line Encyclopedia	Large Pharmacogenomic	Affymetrix U133 Plus 2.0 Microarray, Illumina RNA-Seq	OncoMap, Hybrid Capture NGS	Affymetrix SNP 6.0 Array	479	24	2012
Cancer Genome Project	Large Pharmacogenomic	Affymetrix HT-U133A Microarray	Sanger	Affymetrix SNP 6.0 Array	639	130	2012
Genentech Cell Line Screening Initiative	Large Pharmacogenomic	Illumina RNA-Seq	Sanger	Illumina 2.5M SNP Array	610	16	2015
Genomics of Drug Sensitivity in Cancer	Large Pharmacogenomic	Affymetrix U219 Microarray	Illumina HiSeq Whole Exome NGS	Affymetrix SNP 6.0 Array	1001	265	2016
Cancer Therapeutics Response Portal v2	Large Pharmacogenomic	Affymetrix U133 Plus 2.0 Microarray, Illumina RNA-Seq	OncoMap, Hybrid Capture NGS	Affymetrix SNP 6.0 Array	860	481	2018
Connectivity Map	Perturbational	L1000 Assay			76	19811	2017
Institute for Molecular Medicine Finland	Traditional Screening				106	308	2016

The earliest large pharmacogenomic screen was conducted and published by the pharmaceutical company GlaxoSmithKline, making available their cell line characterization and drug response data [91]. Two years later, the Cancer Cell Line Encyclopedia (CCLE) and the Cancer Genome Project (CGP) large pharmacogenomic studies were published concurrently in 2012, encouraging the development of computational models to predict drug sensitivity [6, 7]. A meta-analysis of overlapping data provided by the CCLE and CGP studies found generally strong correlation between the genomic data provided, but suggested inconsistencies in the drug response measurements when comparing calculated values for half maximal inhibitory concentrations (IC50) [92]. When the Genentech Cell Line Screening Initiative published the results of their study (cGSI), they confirmed agreement of genomic data between cGSI, CCLE, and CGP [93].

The putative assumption was that inconsistency between the two studies arose from a combination of technical and data processing discrepancies [94], and a follow-up study that processed the data through a unified IC50 estimation pipeline harmonized the results [95]. However, this was accomplished largely by placing an artificial upper limit on calculated IC50 values. The main culprit appeared to be incomplete and poorly overlapping dose response curves that made IC50 difficult to estimate and thus a poor representation of drug sensitivity [96]. Suggestions were made that the area under a normalized dose response curve be used instead, and that targeted precision compounds should be evaluated separately from nonspecific traditional chemotherapy medications [97]. The Institute for Molecular Medicine Finland created and published their own drug screening dataset (FIMM) of 308 drugs evaluated across 106 cancer cell lines sans genomic data, and argued that drug sensitivity results could be reconciled between FIMM, CGP, and CCLE through the use of a novel drug sensitivity score [98]. Additional

computational methods have been proposed to rectify the situation [99, 100]. The disagreement remains unresolved [101].

In the ensuing time, the research groups responsible for CGP and CCLE have continued their work, expanding the number of cell lines characterized as well as the number of drug compounds tested. The Genomics of Drug Sensitivity in Cancer (GDSC) is an expanded version of CGP and has replaced it for most purposes [102]. The Cancer Therapeutics Response Portal (CTRP) is an expansion of CCLE. Although the two experiments are maintained by the Broad Institute as separate studies, CTRP obtains most of its genomic data from CCLE [103].

Other experiments that are not technically large pharmacogenomic studies but are significant and worth mentioning include the Connectivity Map (CMap), a perturbational experiment that instead of focusing on phenotypic drug response, seeks to determine the genomic impact of exposure to specific compounds. It has quantified changes in gene expression caused by 28,000 perturbagens on 76 cell lines using the L1000 Assay [104]. In addition, there exist countless smaller pharmacogenomic datasets that are focused on characterizing cancer cell lines of a specific type, most notably for breast [105, 106].

A major criticism of cell line-based screening experiments is that the laboratory cell culture model may not be highly representative of the actual disease state. Although the lineage of a cancer cell line eventually traces back to a primary sample, the immortalization process necessarily changes the cell line, and its genome is vulnerable to drift over the course of many passages in the laboratory [107]. The alternative is to conduct screening experiments using patient-derived primary cells. While this can certainly be done [108, 109], it is not a perfect solution, as not all primary cell samples can be maintained in culture long enough to perform the experiment.

In an interesting twist, patient-derived xenograft (PDX) models have recaptured the attention of investigators as they seek a model that adheres to the actual disease state as closely as possible [110]. This process has drawbacks similar to those of the primary cell culture model, as not all tumor samples can successfully inoculate a PDX. However, it does achieve the goal of modeling a very realistic cancer, at the cost of decreased throughput that comes with performing an in vivo screen [8].

The latest development in drug screening experiments attempts to address this throughput problem via the use of patient-derived organoids [111]. Organoids are miniature models of organs. Thus, they are composed of a combination of many different cell types. This diversity, combined with local anatomy that is similar to real tissue, should cause them to be more realistic models than immortalized cell monocultures. Organoids are grown in culture, making them logistically easier to use than PDX models. Although the protocols for creating patient-derived organoids are still relatively new, tissue-specific pharmacogenomic organoid datasets are already being created [112-115]. If organoids prove to offer significant tangible advantages over cell line-based models, they will play an important role in the future of cancer drug screening.

2.3 Current Methods in Cancer Pharmacogenomics

The infancy of pharmacogenomics can be traced back as far as 510 BC, when Pythagoras recorded that consumption of fava beans induced a serious, sometimes fatal, reaction in some, but not all people [116]. It is now known that this reaction is caused by a deficiency in glucose-6-phosphate dehydrogenase. The digestion of fava beans creates highly reactive redox compounds that in turn produce reactive oxygen species in the blood. Ordinarily, this response is controlled

through cellular antioxidant mechanisms. However, these controls rely on glucose-6-phosphare dehydrogenase to function effectively. Individuals with a deficiency are less able to clear reactive oxygen species, and suffer severe oxidative damage as a result. This damage causes red blood cells to break down faster than they can be replaced, resulting in hemolytic anemia [117].

However, there was no way to capitalize on Pythagoras's insight[3], as there existed no formalized concept of genes until Gregor Mendel established his rules of heredity [118]. Even then, without modern molecular biology techniques for identifying, visualizing, and quantifying genetic changes, little progress could be made in exploring potential connections between genetic variability and drug metabolism. It was not until the 1950s that the maturation of genetic technologies such as gel electrophoresis would both enable and renew interest in the study of pharmacogenomics [119-121].

Although many types of genetic variation exist, pharmacogenomics exploration has largely centered on single nucleotide polymorphisms (SNPs) and their impact on individual drug response [122]. Early interest in the genetics behind pharmacokinetic variability yielded advances in the understanding of transport proteins and metabolizing enzymes, including cytochrome P450 [123]. Over time, focus shifted to emphasize variation in genes that encode drug targets [124]. Prominent among the successes is the identification of the role of *APOE* polymorphisms in Alzheimer's disease [125], which served as a proof of concept for both the selection of a subgroup of patients whose genetics predispose a better or worse response to treatment [126, 127], and the detection of a linkage disequilibrium locus [128], which would pave the way for genome-wide association studies (GWAS) after the completion of the Human Genome Project [129].

[3] Aside from prodigiously avoiding fava beans.

Despite these advances, there were few opportunities to integrate pharmacogenomics into clinical practice [130]. This difficulty arose from a combination of factors, including concerns about the usefulness of genetic testing in the primary care setting [131] and the prohibitive cost of early genotyping systems [132]. However, the creation and implementation of critical entities including genetics specialists [133], education programs [134], and computer decision support systems [135] slowly began to enable physicians to meet the growing public demand for genetic testing and counseling [136].

Although pharmacogenomics holds promise for an array of treatments across many diseases, the potential for impact is greatest in diseases that are inherently variable. These include infectious diseases caused by rapidly mutating bacteria and viruses, such as HIV and hepatitis, as well as chronic diseases with broadly diagnosable symptoms but numerous potential underlying molecular causes, like Alzheimer's and hypertension [124]. Cancer represents a combination of challenges from both groups. Like the former, it can mutate and develop resistance to previously effective therapy. Like the latter, it is highly heterogenous, and exhibits great variation in response to treatment as a result. Because of this, pharmacogenomics is uniquely positioned to provide powerful insights into the effective treatment of cancer [137].

The holy grail of cancer pharmacogenomics is the accurate prediction of drug response based on a patient's genetic profile in the form of an affordable, accessible genetic test performed on a readily available biological sample. The challenge lies in determining which genes to test and which polymorphisms to look for.

Early work in this direction focused on optimizing the administration of existing chemotherapies, a class of medications which initially consisted wholly of nonspecific, cytotoxic compounds. Guided by known mechanisms of action elucidated through traditional molecular

biology techniques, researchers conducted population studies to identify critical genes and gene products for a variety of chemotherapy medications. Many of these focused on predicting the ability of the body to properly metabolize the drug, controlling for toxicity risk and preventing adverse reactions. Results include thiopurine methyltransferase deficiency causing thiopurine toxicity [138, 139], and dihydropyrimidine dehydrogenase deficiency causing 5-Fluorouracil toxicity [140, 141]. Similar studies examining the genetic basis of resistance to platinum agents revealed the importance of various DNA excision and repair proteins [142, 143], as well as more general components of the cellular toxic response [144].

These findings, while interesting, were unfortunately generally not clinically actionable. In the cases of thiopurine and 5-Fluorouracil, the toxicity-inducing enzymatic deficiencies occur in the general population at rates of 0.3% and 0.1% respectively [137, 145]. This prevalence is so low that pre-screening for these conditions before treatment, while available, is not common practice. As for platinum therapy, the mechanistic significance of specific DNA repair SNPs was difficult to determine [146, 147], and the cellular processes for handling toxin export contain such redundancy that some very deleterious mutations inexplicably have no predictive power [148]. Overall, this suggests the existence of a much more complex series of interactions than can be accurately described by a small number of polymorphisms. The administration of these medications under current standard of care guidelines is based on tumor tissue type and staging, without genetic testing.

The advent of molecularly targeted medications signaled a potential change in this paradigm. When combined with powerful algorithms for identifying driving mutations in individual cancers, these precision medications raised the possibility of delivering effective therapy by matching a specific tumor's causal mutation with a drug that specifically targets that

alteration [149]. Unfortunately, the formulary of precision medications is limited, and prescription of these medications using the genomic status of a target gene as a therapeutic indicator benefits only an estimated 5-8% of patients [150, 151]. This is due in part to stringent clinical guidelines which often require a tumor to be from a specific tissue and contain a specific polymorphism in a single gene to approve the use of a precision therapy. Theoretically, a targeted medication could provide benefits over traditional chemotherapy for patients who do not meet the clinical guidelines for its use.

Proposals for repurposing agents across different disease and tumor types with the same target gene, across different alterations within the same gene, and based on strong but unintended binding kinetics could potentially increase the portion of patients benefiting from targeted medications to 40% [4]. The most recent work in this area suggests that combining a database of drug-target interactions with pathway knowledge enables repurposing of drugs across different targets in the same pathway, potentially bringing the portion of patients benefiting from existing precision therapies to over 90% [152]. Although drug repurposing systems can certainly suggest a targeted medication for a large proportion of cancer patients, the rate at which the suggested drugs will be effective in treating the specific tumors is likely to be much lower in reality. These systems typically operate under either a Mendelian or limited polygenic model for disease etiology. Under these models, the observable variability in disease phenotype, which includes clinically relevant traits such as treatment response, is attributable to genotypic variation in just a small number of genes. If accurate, then effective treatment decisions can be made based on knowledge of the mutation status of a few genomic locations.

However, in the case of cancer and other complex diseases, this assumption generally does not hold [153, 154]. Genome-wide association studies have been very good at identifying

important loci in complex diseases, but significant results from a typical GWAS, when combined, still only explain a fraction of observed variability [155]. The rest is attributable to a combination of a large number of common polymorphisms with effect sizes too small to be statistically identified in a genome-wide test, and very rare variants with large effect sizes that do not occur in the test sample [156, 157]. In contrast to Mendelian diseases which are typically caused by protein coding changes [158], complex diseases are generally characterized by polymorphisms in noncoding regions that control gene regulation [159, 160]. This evidence supports the idea that complex diseases are driven by the cumulative effect of a large number of relatively small individual effects on key genes and pathways [161].

This long tail effect has been documented in many cancer types, including pancreatic, prostate, and breast [162-165]. This phenomenon has significant implications for the accurate prediction of the therapeutic effectiveness of targeted medications. Considering only the genomic status of a drug target or a drug target and a small number of related genes does not work. This was demonstrated in a study examining the effectiveness of certain PI3K inhibitors in restricting the growth of an array of PDX models. The findings showed a wide range of drug effectiveness irrespective of the genomic status of *PIK3CA* and *PTEN* [8]. Proper modeling of tumors and their response to treatments thus needs to consider interactions between key driver genes or pathways and a host of less common but still impactful polymorphisms that impart unique variation to individual tumors.

Modeling complex diseases in this manner requires genome-wide technologies capable of monitoring many cellular transcripts in parallel. Initially, this took the form of microarrays, with specialized applications ranging from SNP genotyping to RNA expression profiling. More recently, focus has shifted to next-generation sequencing technologies, but microarrays remain in

widespread use. Gene expression analysis was found to be especially useful for cancer classification. Combined with unsupervised machine learning methods such as hierarchical clustering, microarray expression data were used to identify a new subtype of acute lymphoblastic leukemia [166], stratify diffuse large B-cell lymphoma into three prognostic types [167, 168], and separate breast cancer patients into subgroups with significant differences in treatment outcome [169]. The addition of supervised learning techniques enabled the creation of clinically relevant gene signatures containing dozens of genes that combined enabled prediction of disease outcome and potential benefit from specific therapies [170]. Such gene signatures were the precursors to modern-day gene panels [171-173].

Even before the existence of microarrays, clinical data were being incorporated into successful models for predicting susceptibility, recurrence, and survivability in a variety of cancer types using many different types of models, including naïve Bayesian classifiers [174], support vector machines [175], decision trees [176], and artificial neural networks [177]. As genomic data started becoming available, it was rapidly incorporated into models alongside clinical data [178-180], and then eventually used in the creation of genomic-only models [181-183]. The practice of machine learning in cancer pharmacogenomics came as a natural extension of the application of these techniques in classification and prognosis.

However, predicting the effectiveness of specific drugs in individual tumors presents some unique challenges. Most of the patient datasets used to create clinical models enrolled only a few dozen participants, although there were a few exceptions [184]. For models based on a small number of clinical features, these sample sizes are adequate. However, more complex models seeking to incorporate high dimensional genomic data require a larger training set. Of larger concern was the availability of drug response data. Any particular patient generally only attempted

a handful of different treatment regimens. It was both ethically and logistically impossible to evaluate the effects of dozens of potential therapies on every patient for the sole purpose of collecting data for training predictive models. For these reasons, pharmacogenomics investigators turned to cell line screening experiments for increased throughput, just as their drug discovery counterparts did before them.

The earliest machine-learning-focused cell line pharmacogenomics studies generally concentrated on predicting the effectiveness of treatment compounds on specific, individual types of cancer, especially breast cancer [105, 185]. In 2012, the NCI, along with the Dialogue on Reverse Engineering Assessment and Methods project, held a challenge to assess the performance of drug sensitivity prediction algorithms [186]. 44 different prediction methods were evaluated on a dataset of 53 breast cancer cell lines and 28 therapeutic compounds. The best-performing method was found to be Bayesian multitask multiple kernel learning; a random forest approach took second place.

As large pharmacogenomics studies such as CCLE and CGP began publishing datasets containing cell lines from many different tissue types, the search turned towards developing unified models capable of predicting a sample's drug response regardless of tissue of origin. Algorithms used in this endeavor include ridge regression [187], elastic net [188], random forest [102, 189], perceptron [190], and support vector machine [191]. In virtually every study, separate models are trained to predict the IC50 of individual drugs from gene expression microarray data. To improve on the current state of the art, new approaches need to integrate other data types made available by large pharmacogenomics studies, such as mutation and copy number alteration information. The proven ability of deep learning methodologies to model complex interactions makes them strong candidates for this task [192, 193].

2.4 Preliminary Work

This section describes preliminary work conducted in 2016 that explored the possibility of predicting drug effectiveness from integrated omics data using deep learning. It involved preprocessing omics data from large pharmacogenomics experiments, developing a latent representation of this data, using the latent representations to build drug sensitivity models, and evaluating the performance of those models. This work is presented in an abridged form here – for additional details and experiments, please refer to the associated publication [194].

2.4.1 Methods

2.4.1.1 Data Retrieval and Feature Engineering

I retrieved data from the two largest available pharmacogenomics studies, the Genomics of Drug Sensitivity in Cancer (GDSC) [7], and the Cancer Cell Line Encyclopedia (CCLE) [6]. GDSC data were chosen for use in model training despite its lower cell line count. This is because GDSC has significantly more drug data. 140 compounds were evaluated in GDSC, compared to 24 in CCLE. CCLE data were set aside for later use in model evaluation.

Microarray gene expression data from GDSC was obtained in the form of raw Affymetrix CEL files. Batch effects were removed using Robust Multi-Array Averaging [195]. Expression values from replicate experiments were averaged, and data corresponding to Affymetrix spike control probes were manually removed. This procedure generated an array of 22,215 probe-level expression measurements in 727 cell lines.

To reduce computational load, feature selection was performed on GDSC gene expression data. Reducing the cardinality of the feature set resulted in an order of magnitude reduction in the

number of parameters requiring tuning in downstream deep learning. Feature selection was performed using three variance metrics, based on the assumption that features with low variance across the dataset have limited predictive utility [196]. Hartigans' dip test for unimodality was used to select for features with multimodal distributions [197]. The outlier sum method was used to select for features characterized by largely unimodal distributions but with significant outlier populations [198]. Median absolute deviation was used to select for genes with a high variance across samples regardless of distribution shape. Selection via these three methods yielded 3,080 gene expression features. A mixture of two normal distributions was fitted to each feature's expression profile using the expectation-maximization algorithm [199]. A Bonferroni-corrected t-test was performed to verify statistical significance between the two groups for each gene. These two groups were then used to determine a cutoff for discretizing the expression levels of each gene into low and high values. After this procedure, the preprocessed GDSC gene expression dataset contains discretized gene expression data of 3080 probes.

Copy number estimates for 426 genes were obtained in a preprocessed form from GDSC. These values were determined by processing Affymetrix SNP 6.0 microarray data using the PICNIC algorithm for copy number prediction [200]. Copy-number estimates ranging from 0 to 10 were normalized to real values between 0 and 1. Genes with copy-number estimates greater than 10 were set to the maximum normalized value of 1.

Mutation annotations for 71 genes were obtained in a preprocessed form from GDSC. SNPs were identified via capillary sequencing, and chromosomal rearrangements were determined with the use of specialized primers. Genes with a rearrangement event or other non-silent mutation were encoded with a value of 1. Unmutated genes or genes with silent mutations were assigned a value of 0.

Cell line annotations were used to combine the gene expression, copy number, and mutation datasets into a single array for deep learning. Cell lines without all three data types available were discarded, leaving a final dataset of 3577 features in 624 cell lines.

2.4.1.2 Developing Latent Representations of Omics Data Using Deep Learning

Matlab code for training a deep autoencoder was obtained from Geoffrey Hinton's website and modified to use the feature selected GDSC dataset for unsupervised representation learning [201]. To train the autoencoder, the 624 cell lines in the dataset were randomly split into training and validation sets of 520 and 104 samples, respectively. The training set contains the data actually used to learn the model, while the validation set is used to provide an unbiased evaluation of the model for hyperparameter tuning. Specifically, the validation set was used to enabling application of the early stopping rule to determine an appropriate number of backpropagation cycles. An architecture with seven hidden layers of size 1300, 552, 235, 100, 235, 552, and 1300 was used[4]. The model weights were initialized by pretraining a stacked restricted Boltzmann machine for 50 epochs. Using a minibatch size of 26, conjugate gradient descent was used to optimize the autoencoder weights. The early stopping rule was invoked to stop backpropagation after 400 epochs of training when reconstruction error on the training set started to diverge from reconstruction error on the testing set.

After training of the autoencoder was completed, the GDSC final dataset of 3577 features in 624 cell lines was propagated through the neural network. Values for the hidden nodes in the

[4] The selection of this structure was relatively arbitrary. The only real guideline was that the first layer probably needed to be larger than 1000 and the middle layer needed to be larger than 10. The intervening layers are a geometric progression.

first four layers were extracted to produce four latent representations of the GDSC data with 1300, 552, 235, and 100 features, respectively.

2.4.1.3 Predicting Drug Sensitivity with Latent Representations

GDSC drug sensitivity measurements in the form of normalized activity area values were obtained and discretized into sensitive and resistant categories by applying the waterfall method described by CCLE [6]. This procedure is briefly summarized as follows: drug sensitivity measurements for a single drug across all cell lines are sorted in increasing order to generate a waterfall distribution. A linear regression is fitted to this distribution, and a Pearson correlation is calculated to determine goodness of fit for the linear regression solution. If the Pearson correlation is less than 0.95, indicating a poor fit, the major inflection point is estimated to be the point on the waterfall distribution with the maximal distance to a line drawn between the start and endpoints of the distribution. Otherwise, the median value is used as the inflection point instead. The inflection point is used as the cutoff for designating cell lines as sensitive or resistant to the drug relative to other cell lines in the experiment.

This drug sensitivity metric was chosen over traditional targets such as IC50 or -log(IC50) due to technical and biological variance associated with its measurement and estimation [202] and to avoid the additional future complexity of extrapolating from in vitro nominal concentrations to make conclusions in vivo [203]. In contrast, predicting a binary outcome variable that describes relative sensitivity allows for more straightforward adjustments later if required.

I used elastic net regression to generate logistic models for drug sensitivity prediction [204]. This is a form of logistic regression that combines lasso and ridge regularization, which enables it to perform feature selection while maintaining stability in cases of multicollinearity. The elastic net contains two hyperparameters, alpha and lambda. Alpha defines the relative weight of

the lasso and ridge penalization terms. Lambda determines the overall size of the regularization penalty. Alpha was fixed at 0.5, and predictive performance was optimized over a range of lambdas. Regression was performed with 25-fold cross-validation.

For each drug, six models were built to predict the same target consisting of the discretized sensitivity data for that drug across the cell lines in which it was evaluated. The input vectors for the six models were a combined set of original unprocessed omics features, the feature selected data set used to train the deep autoencoder, and each of the four layers of latent representations generated from the trained autoencoder.

2.4.2 Results

The elastic net models trained with all omics features achieved an average area under the receiver operating characteristic of 0.81. Applying variance-based feature selection before training the elastic net did not enhance overall predictive performance, resulting in an average AUROC of 0.78. However, some individual drugs are better predicted with feature selection than without. Aggregate predictive performance of models using latent variables as predictive features was an average AUROC of 0.79. Interestingly, some drugs modeled poorly using the original or selected omics features are significantly better predicted using hidden layer models. As a group, the best model for every drug has an average AUROC of 0.87 (Figure 1A).

For any given drug, the performance often varies when latent variables from different layers are used as predictive features. 57 drugs are best predicted using the original omics features, 30 are best predicted using selected features, and 53 drugs are best predicted using latent representation features. These findings suggest that complex relationships useful in the prediction of drug sensitivity are uncovered by deep learning (Figure 1B).

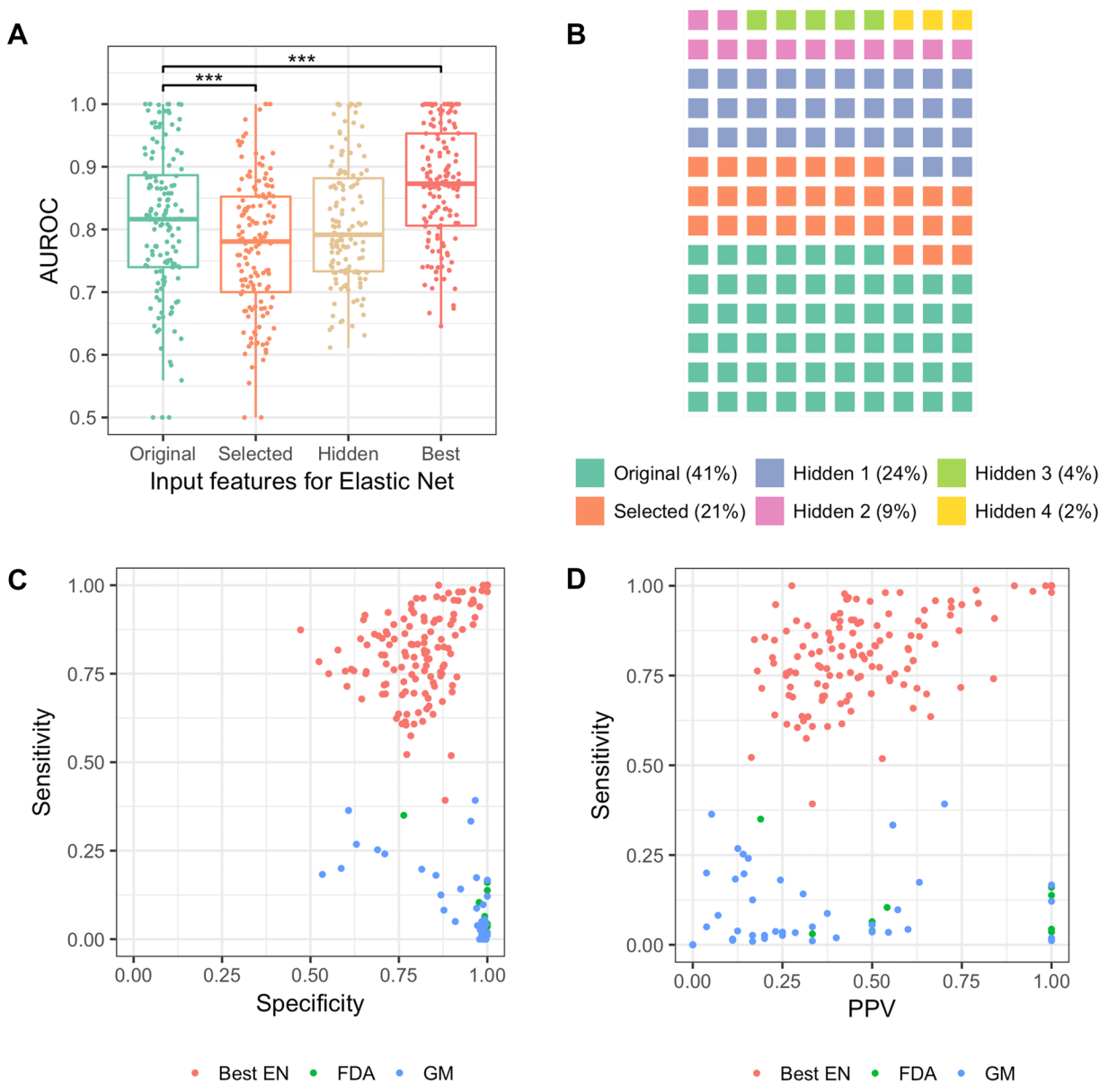

Figure 1: Learning cellular states using deep learning. A, Predictive performance of elastic net models relative to predictive features used as inputs. B, Proportion of best models from each category of input feature. C, Sensitivity and specificity of 140 best elastic net models (Best EN) compared with 43 genomic marker rule-based models (GM) and 10 FDA genomic guideline clinical indications (FDA). D, Sensitivity and positive predictive value of 140 best elastic net models (Best EN) compared with 43 genomic marker rule-based models (GM) and 10 FDA genomic guideline clinical indications (FDA). ***, $P < 10e{-}3$.

Average sensitivity, specificity, and positive predictive value for the best models using variable thresholds are 0.82, 0.82, and 0.51, respectively (Figure 1C, 1D). Exceptional performance was achieved for 15 drugs, in which sensitivity and specificity values were greater than 0.98, positive predictive value was greater than 0.94, and AUROC values were greater than 0.99.

Out of the 140 drugs in the dataset, 29 have unknown or nonspecific mechanisms of action, leaving 111 molecularly targeted specific therapies. Of these 111 drugs, some combination of mutation or copy number information is available for 53 drug targets, whereas the genomic status of the target genes of the remaining 58 drugs was not measured. For each of the 53 drugs, a rule-based classifier was created to predict drug sensitivity. 10 of these drugs have an FDA-approved genomic testing indication for their clinical use, and the rule-based classifier mirrors these predefined indications. These 10 compounds are comprised of poly(ADP-ribose) polymerase and tyrosine kinase inhibitors, with genetic tests for *BRCA1/2*, *EGFR*, *ERBB2*, *ALK*, and *BCR-ABL* mutations to approve their use. For the remaining 43 drugs, the rule-based classifier makes its decision based on the genomic status of the target protein. If it is either mutated or copy-number amplified, the cell line is predicted to be sensitive to the drug.

The average sensitivity, specificity, and positive predictive value of the rule-based models are 0.10, 0.93, and 0.38, respectively (Figure 1C, 1D). Most cell lines are insensitive to molecularly targeted therapies, and the majority of cell lines were predicted by the rule-based models to be insensitive, resulting in generally high specificity. However, the low sensitivity and positive predictive values indicate that significant numbers of sensitive cell lines do not have the relevant genomic markers and are not identified, and the majority of cell lines hosting a genomic marker are actually not sensitive to the drugs. Generally, the best elastic net models significantly

outperform the rule-based models, indicating the existence of potential opportunities for therapeutic improvement in these areas.

To investigate the external validity of the machine learning predictive models, they were evaluated using experimental data from CCLE. 15 of the drugs studied in GDSC were also investigated in CCLE. The best models for these 15 drugs were re-evaluated using data from CCLE.

2.4.3 External Validation

Microarray gene expression data from CCLE was obtained in the form of raw Affymetrix CEL files. Batch effects were removed using Robust Multi-Array Averaging [195]. Expression values from replicate experiments were averaged, and data corresponding to Affymetrix spike control probes were manually removed. This procedure generated an array of 54,675 probe-level expression measurements in 1067 cell lines. A mixture of two normal distributions was fitted to each gene's expression profile using the expectation-maximization algorithm [199], and these groups were used to determine a cutoff to discretize the expression levels of each gene into low and high values.

HapMap normalized CCLE copy number alteration data were obtained and used to estimate raw copy number by assuming a base frequency of two copies per gene. Copy-number estimates ranging from 0 to 10 were normalized to real values between 0 and 1. Genes with copy-number estimates greater than 10 were set to the maximum normalized value of 1.

CCLE mutation data were collected using two methods, Oncomap 3.0 and hybrid capture analysis. These data were combined using cell line annotations, and reclassified as mutated or not based on The Cancer Genome Atlas specification for the Mutation Annotation Format.

Following preprocessing of CCLE omics data, values for the 3,577 features used to train the GDSC autoencoder were extracted and used to create a CCLE selected dataset. All features were available in both experiments with the exception of mutation data on four genes. These missing values were filled in with uninformative average values derived from the GDSC dataset. After this procedure, the CCLE selected dataset is an array of 3,577 features measured in 1067 cell lines. CCLE drug sensitivity data in the form of normalized activity area values was obtained and discretized into relative sensitive and resistant categories by applying the waterfall method as described previously. The CCLE selected dataset was propagated through the autoencoder trained on the GDSC selected dataset to create a latent encoding of the CCLE data. This latent representation was then used as inputs to the GDSC elastic net models to make drug sensitivity predictions for the 15 drugs shared by the two experiments. These predictions were evaluated against the waterfall discretized experimentally measured drug response data from CCLE.

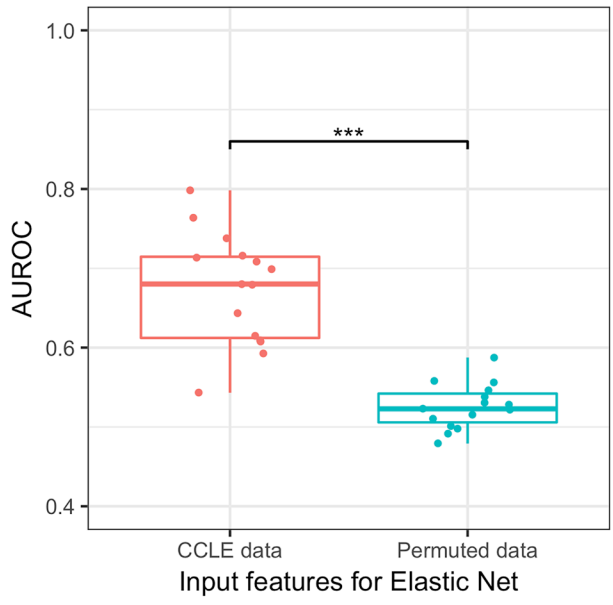

Figure 2: External validity of predictive models. AUROC values for elastic net models developed using GDSC omics data evaluated using CCLE omics data or randomly permuted CCLE omics data. *, *P* < 10e-6.**

The 15 models achieved an average AUROC of 0.67 (Figure 2). This is significantly higher than prediction results obtained using randomly permuted input data, indicating that the relationships modeled by deep learning persist even under different experimental conditions.

2.4.4 Conclusions

In this preliminary work, I combined genome-scale omics data with machine learning techniques to accurately predict the performance of a wide range of targeted and untargeted therapies on cancer cell lines. The findings indicate that data-driven approaches may significantly outperform rule-based methods using the genomic status of drug targets as therapeutic indicators. Completing these experiments provided support for the viability of the project and created a precursory experimental framework for organizing subsequent work.

3.0 Research Design

In this section, I present the motivation behind his research in the form of specific research questions. Further, I provide an overview of the individual studies that comprise this dissertation as guided by four specific aims.

3.1 Research Questions

1. Is it possible to build computational models with some level of accuracy for drug sensitivity prediction using omics features and experimental response data from large pharmacogenomics experiments?
2. What technologies can be employed to build these models and how can they be optimized for this task?
3. Is there potential for these models to improve the state of clinical care? If so, by how much?
4. Can either these models or the overall modeling framework generalize to clinical data in order to provide such a benefit?

3.2 Dissertation Overview

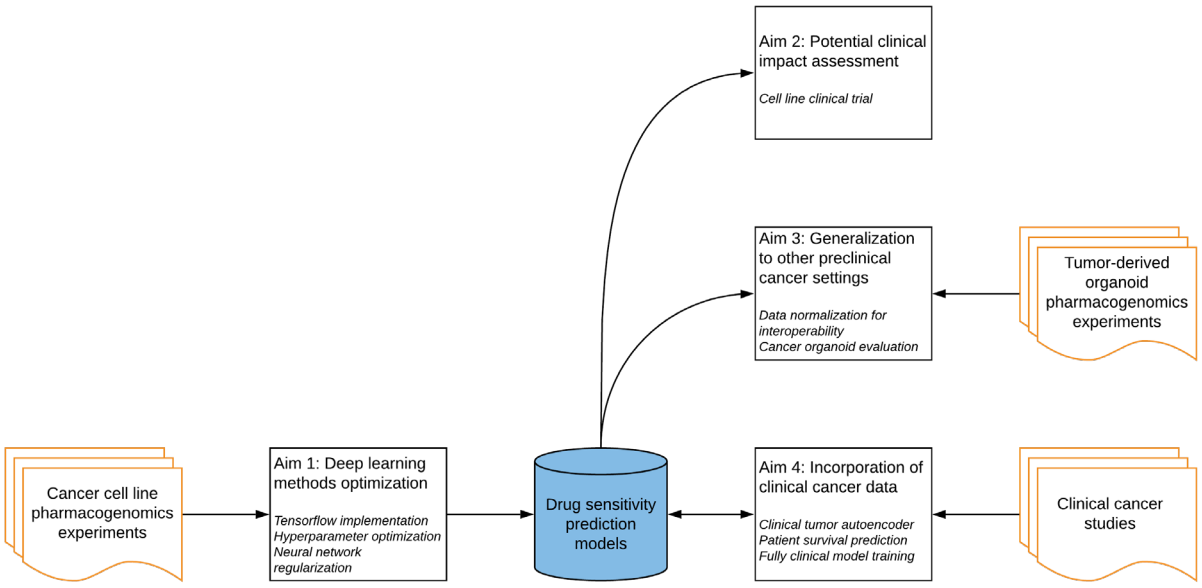

Figure 3: Dissertation overview.

Figure 3 provides an overview of this dissertation research that consists of four specific aims that are associated with the research questions defined above.

Aim 1: To optimize deep learning methodology.

- Incorporate newly available data from the Genomics of Drug Sensitivity in Cancer large pharmacogenomics experiment into predictive models.
- Optimize the machine learning model developed in preliminary work using a modern machine learning framework.
- Optimize the architecture hyperparameters of the deep neural network autoencoder.
- Evaluate potential benefits from including various regularization methods in the training of the deep neural network autoencoder.

Aim 2: To determine the potential for clinical impact.

- Simulate a cell line clinical trial by using a group of tumor cell lines to represent a cohort of patients.
- Compare outcomes achieved by following the current standard of care versus using artificial intelligence-supported decision making.

Aim 3: To generalize predictive models from cancer cell lines to cancer organoids.

- Explore data normalization techniques to enable interoperability of predictive models with omics data from different sources.
- Evaluate the ability of computational models trained using cancer cell lines to predict drug sensitivity in a variety of different cancer organoids.

Aim 4: To incorporate available clinical data via transfer learning.

- Evaluate the potential to improve model performance or generalizability by training the deep autoencoder on data from clinical tumor samples.
- Further generalize predictive models from cancer cell lines to predict clinical outcomes or survival of patients using omics data collected in the process of care.
- Evaluate the potential to create predictive models utilizing only clinical data.

4.0 Aim 1: Deep Learning Optimization

Aim 1 focuses on improving the machine learning model used in preliminary work. The autoencoder described in **2.4.1.2** was adopted with relatively few modifications from code originally written for the task of recognizing images of integers 0 through 9 from the MNIST handwritten digits dataset [201]. As a result, several significant optimizations are possible.

First, preliminary work was conducted on GDSC release 5.0, published in June 2014 [205]. The next major version, release 6.0, was made available two years later in June 2016 [102]. The predictive models were updated to take advantage of this new data. Second, the scientific computing code was developed and used before the creation of modern open source machine learning frameworks. The autoencoder was reimplemented using Google's TensorFlow machine learning platform. Third, a very small number of neural network architectures were investigated in the preliminary work. A more rigorous search was conducted to optimize these hyperparameters. Fourth, recent advances in deep learning include the application of various regularization techniques to latent layers. These modifications to the training process are intended to increase generalizability and reduce training time. The potential benefit of these regularization methods was examined. I discuss each of these optimization components in the following sections.

4.1 Data Retrieval and Feature Engineering

The GDSC project's version 6.0 release increased the coverage from 140 compounds evaluated in 624 cell lines to 265 compounds evaluated in 1001 cell lines. I decided to make this

updated dataset the basis for the following work. Although the original download location no longer exists, the data used in the following steps can currently be accessed from the GDSC archives in the directory for release 6.0. The dataset was processed in a manner similar to that used in preliminary work, described in **2.4.1.1**.

4.1.1 Methods

4.1.1.1 Predictive Feature Data

Microarray gene expression data from GDSC was obtained in the form a pre-processed Robust Multi-Array Averaged dataset [102]. This dataset consists of an array of 17,738 gene-level expression measurements in 1018 cell lines. As before, feature selection was performed on the GDSC gene expression data to reduce computational load using Hartigan's dip test for unimodality [197], the outlier sum method [198], and median absolute deviation. Selection via these three methods yielded 3,108 gene expression features. Also as before, a mixture of two normal distributions was fitted to each feature's expression profile using the expectation-maximization algorithm [199]. A Bonferroni-corrected t-test was performed to verify statistical significance between the two groups for each gene. These two groups were then used to determine a cutoff for discretizing the expression levels of each gene into low and high values. After this procedure, the preprocessed GDSC gene expression dataset contains discretized expression data on 3,108 genes.

Copy number estimates for 585 genes were obtained in a preprocessed form from GDSC. These values were determined by processing Affymetrix SNP 6.0 microarray data using the PICNIC algorithm for copy number prediction [200]. Copy-number estimates ranging from 0 to 8 were normalized to real values between 0 and 1. Genes with copy-number estimates greater than

8 were set to the maximum normalized value of 1. Genes for which data were incomplete were discarded, yielding a dataset containing copy number estimates for 565 genes.

Mutation annotations for 587 genes were obtained in a preprocessed form from GDSC. SNPs were identified from whole exome sequencing data via the CaVEMan and Pindel algorithms for identifying substitutions and small insertions/deletions, respectively. Genes with a rearrangement event or other non-silent mutation were encoded with a value of 1. Unmutated genes or genes with silent mutations were assigned a value of 0.

Cell line annotations were used to combine the gene expression, copy number, and mutation datasets into a single array for deep learning. Cell lines without all three data types available were discarded, leaving a final dataset of 4260 features in 963 cell lines.

4.1.1.2 Drug Sensitivity Data

GDSC drug sensitivity measurements for 1074 cell lines in the form of normalized activity area values were obtained and discretized into sensitive and resistant categories by applying the waterfall method used in CCLE [6] and described in **2.4.1.3**. Cell lines for which no drug sensitivity experiments were conducted were discarded, yielding a drug sensitivity target dataset of 265 drugs evaluated in 963 cell lines.

4.2 TensorFlow Implementation

Modern machine learning platforms such as TensorFlow and Pytorch have become standard for the development and deployment of machine learning models in industry and academia [206]. The most significant advantage of utilizing a modern framework for our

application is the ability to utilize GPU computing on Nvidia Corporation's CUDA parallel computing platform. In this section, I discuss my re-implementation of the deep autoencoder in TensorFlow. Furthermore, I evaluate the performance of predictive models built using data representations from the TensorFlow autoencoder and compare them against corresponding models created using the Matlab autoencoder.

4.2.1 Methods

4.2.1.1 Developing Latent Representations of Omics Data Using Deep Learning

Code for training a deep autoencoder using TensorFlow 1.0 was written to use the new feature selected GDSC dataset for unsupervised representation learning. I closely matched the methodology used in preliminary work, as described in **2.4.1.2**. An architecture with seven hidden layers of size 1300, 552, 235, 100, 235, 552, and 1300 was used. The layers were fully connected with sigmoid activation functions. The output layer is linear, and the cost function is a binary cross entropy. The starting model weights were set using Xavier initialization [207]. The network was trained using a gradient descent optimizer. The early stopping rule was invoked to stop backpropagation after 250 epochs of training when reconstruction error on the training set started to diverge from reconstruction error on the testing set. As a control, the Matlab autoencoder used in preliminary work, was trained and evaluated alongside the TensorFlow autoencoder.

To train the autoencoders, the 963 cell lines in the dataset were randomly split into training and testing sets of 858 and 105 samples, respectively. After training of the autoencoders was completed, the GDSC final dataset of 4260 features in 963 cell lines was propagated through the neural networks. Values for the hidden nodes in the first four layers of each autoencoder were

extracted to produce latent representations of the GDSC data with 1300, 552, 235, and 100 features.

4.2.1.2 Predicting Drug Sensitivity with Latent Representations

Elastic net regression was used to generate logistic models for drug sensitivity prediction [204] as described previously in **2.4.1.3**. For each drug, ten models were built to predict the same target consisting of discretized sensitivity data for that drug across the cell lines in which it was tested. The input vectors for the ten models were a combined set of original unprocessed omics features, the feature selected dataset used to train the deep autoencoders, and the four layers of latent representations generated from both trained autoencoders.

4.2.2 Results

The elastic net models trained with original unprocessed features achieved an average area under the receiver operating curve of 0.73 (Figure 4). Applying variance-based feature selection and discretization before training the elastic net did not enhance overall predictive performance, although some individual drugs are better predicted with feature selection than without.

Aggregate predictive performance of the best models using hidden latent variables as predictive features was not higher than models trained with original unprocessed features. This finding applied to both autoencoder implementations tested in this experiment. In fact, the average AUROC of the best models based on hidden latent variables was 0.73 for both autoencoders. A paired t-test comparing the performance of these two groups of models failed to reject the null hypothesis.

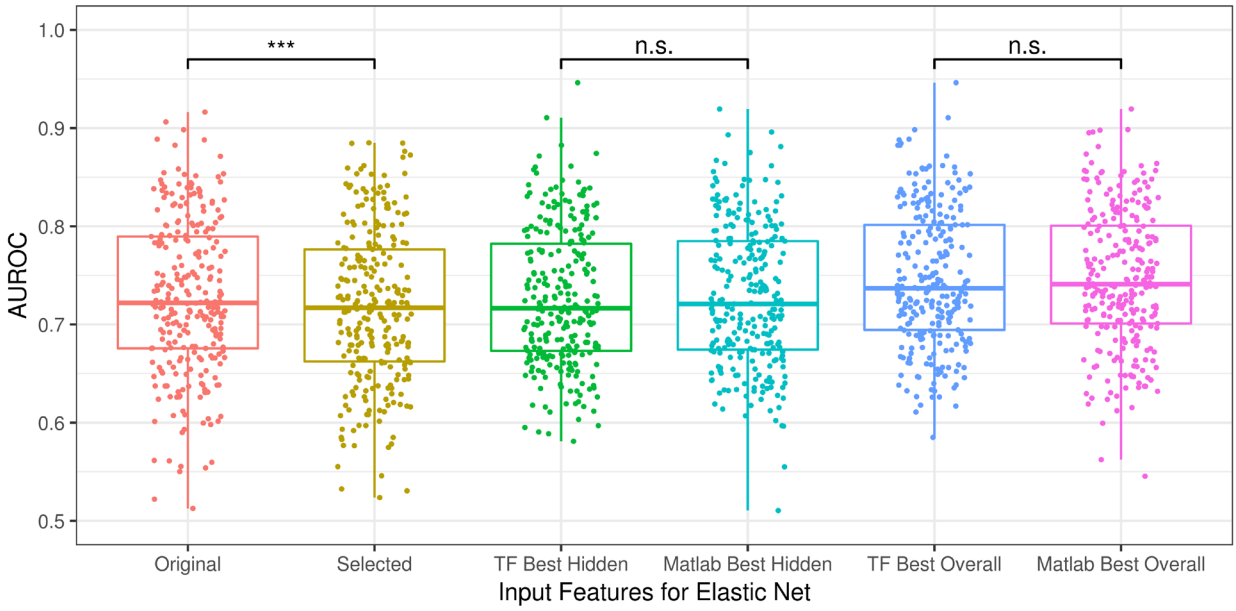

Figure 4: Learning cellular states with different autoencoder implementations. Predictive performance of elastic net models relative to predictive features used as inputs, derived from TensorFlow (TF) or Matlab implementationso of the same autoencoder. ***, $P < 0.001$

4.2.3 Discussion

The purpose of this experiment was to monitor the transfer of the deep learning methodology from one code platform to another. In the process, this experiment recapitulated many of the original findings found in preliminary work.

As before, selecting and discretizing a smaller subset of features from the original integrated omics dataset causes a statistically significant ($P < 0.001$) decrease in predictive performance, presumably due to information loss. This decrease is then rescued by encoding the selected, discretized dataset and using the resulting hidden latent variable representations for predictive modeling. In the end, the best available predictive models are a mix of original, selected, and hidden layer models.

In moving from the Matlab autoencoder to the TensorFlow autoencoder, predictive performance did not deteriorate. This finding supports the use of the TensorFlow autoencoder in subsequent optimization and modeling work. Due to the change in code base, time to train an autoencoder has been reduced from over an hour to less than two minutes. This has enabled the rapid prototyping necessary for proper optimization of the neural network hyperparameters.

4.2.4 Limitations

The results of this experiment are unfortunately not directly comparable to those achieved in preliminary work, reported in **2.4.2**. Despite consistent methodology, the greatly expanded dataset resulted in empirical differences in feature selection and discretization. In addition, AUROC values reported in preliminary work were obtained by generating ROC curves on the training data using fully trained prediction models. In the current and all subsequent experiments, performance metrics were calculated by computing the average AUROC of individual component models during cross validation. This removes a bias by ensuring that the data being used for evaluation has not been seen by the model during training. It is for these reasons that the Matlab autoencoder was re-trained and re-evaluated alongside the TensorFlow autoencoder for this experiment.

4.3 Deep Neural Network Architecture Optimization

The neural network architecture used in preliminary work consisted of seven hidden layers of size 1300, 552, 235, 100, 235, 552, and 1300. With the speed increase from GPU computing, it

became possible to rigorously evaluate competing potential architectures for the autoencoder. This process is complicated by the fact that efficient algorithms do not exist for hyperparameter optimization. As a result, random search and grid search are both popular methods [208]. I chose a hybrid method, where the boundaries of the search space are defined by a grid, and the parameters within are selected at random.

In order to evaluate candidate deep learning architectures, it was necessary to select an evaluation metric other than predictive modeling performance. This is because although we can now perform deep learning very fast, elastic net regression remains computationally intensive, especially for larger feature sets. Test set reconstruction error was chosen for evaluation, as this metric represents the ability of the autoencoder to learn relationships between input features that generalize to unseen data. In this section, I discuss the search for an optimal autoencoder structure.

4.3.1 Methods

4.3.1.1 Candidate Autoencoder Structure Construction

An autoencoder structure factory was created to produce candidate architectures for evaluation. Three hyperparameters were defined by hand. These were the size of the first hidden layer, the size of the smallest hidden layer in the middle, and the number of layers in the autoencoder. In the literature, the smallest middle hidden layer is often referred to as the code layer. The allowable sizes for the first hidden layer were 1000, 1100, 1200, 1300, 1400, and 1500. The allowable sizes for the code layer were 50, 100, 200, 300, 400, and 500. The allowable number of layers in the autoencoder was five, seven, and nine, which corresponds to three, four, and five layers to produce hidden representations, respectively. The remaining hyperparameters, which

were the number of hidden nodes in any undefined layers, were determined randomly according to the following procedure.

Given a first hidden layer size and code layer size, a five-layer autoencoder can be constructed by randomly picking a number between the size of the first hidden layer and the size of the code layer, and making that the size of the layer in between. For example, if the first hidden layer has 1000 nodes and the code layer has 300 nodes, then a randomly selected number from the interval (300, 1000), such as 735, could be used to create a five-layer autoencoder with dimensions of 1000, 735, 300, 735, 1000. For every combination of first hidden layer size and code layer size, 20 random intermediate numbers were selected, resulting in 720 different candidate architectures with five layers.

Given an autoencoder structure with five layers, a nine-layer autoencoder can similarly be constructed by picking two random numbers. The first random number is selected from between the sizes of the first and second layer in the five-layer structure, and the second random number is selected between the sizes of the second and third layers. For example, given the previously used five-layer autoencoder structure of 1000, 735, 300, 735, 1000, then a random number is selected from the interval (735, 1000), such as 800, and another number is selected from the interval (300, 735), such as 512. This creates a nine-layer autoencoder with dimensions of 1000, 800, 735, 512, 300, 512, 735, 800, 1000. In this manner, the 720 candidate architectures with five layers were used to create 720 candidate architectures with nine layers. The search space for architectures with nine layers is much larger than the search space for five-layer architectures, so an additional 720 nine-layer architectures were randomly generated, creating a total of 1440 candidate architectures with nine layers.

Given an autoencoder structure with nine layers, a seven-layer autoencoder can be constructed by removing the third and corresponding seventh layer from the nine-layer structure. For example, given the previously used nine-layer autoencoder structure of 1000, 800, 735, 512, 300, 512, 735, 800, 1000, removing the two layers of size 735 leaves a seven-layer autoencoder with dimensions of 1000, 800, 512, 300, 512, 800, 1000. In this manner, the 1440 candidate architectures with nine layers were used to create 1440 candidate architectures with seven layers.

In total, 3600 candidate architectures were generated. 11 were duplicates, and were discarded. The architecture used in preliminary work was manually added. 3590 architectures were evaluated.

4.3.1.2 Candidate Autoencoder Structure Evaluation

The TensorFlow autoencoder implementation described in **4.2.1.1** was used with some minor modifications. The 963 cell lines in the dataset were randomly split into training and testing sets of 750 and 213, respectively. The training set of 750 samples was then randomly split into five folds, each consisting of a training set of 600 samples, and a validation set of 150 samples. All autoencoders were trained for 1,000 epochs. At each step, the reconstruction error on the validation set and test set were recorded. After training was complete, the epoch with the lowest reconstruction error on the validation set was determined, and the corresponding reconstruction error on the test set at that time was recorded. This training and evaluation process was repeated for each of the five folds, and the resulting five test set reconstruction errors were averaged to create the target evaluation metric.

For evaluation, autoencoder performance was visualized against the complexity of the underlying neural network. Mirroring the deep learning optimization cost function, performance was defined as a cross entropy reconstruction error in the form of a negative log-likelihood.

Complexity was represented as the number of hyperparameters in each neural network on a log scale.

During analysis, an additional metric was created to describe the magnitude of the information bottleneck present in the autoencoder. Since the encoding component of the autoencoder consists of successively smaller layers, it is possible to calculate a shrinkage ratio by dividing the number of nodes in a layer by the number of nodes in the next layer. For example, if the current layer has 500 nodes, and the next layer has 150, then this represents a shrinkage of 3.33. On the natural log scale, the bottleneck for this particular transition of layers is thus 1.20. This metric is calculated for each transition between layers, and the maximum value obtained is defined as the bottleneck for the entire autoencoder. The bottleneck metric represents the largest layer-to-layer drop in node size present in the deep neural network.

4.3.1.3 Optimal Autoencoder Structure Design and Evaluation

To aid in the selection of an optimal autoencoder structure from the configurations tested, a high-dimensional Bayesian information criterion (BIC) was calculated to quantify the tradeoff between reconstruction performance and complexity [209, 210]. Performance is defined as the log likelihood of the model given the data, and complexity is a log-scaled count of the number of parameters, or weights, in the deep neural network. The top 20 autoencoder structures by this criterion were examined, and in combination with the other evaluation metrics, were used to inform the design of a single new autoencoder structure for subsequent experiments.

The new deep autoencoder structure was implemented and trained on the GDSC integrated omics dataset as described in **4.2.1.1**. The only modification made was to the autoencoder structure.

To determine the impact of the new structure on predictive performance, the new autoencoder was used to generate predictive features for elastic net regression as described in **2.4.1.3** and **4.2.1.2**. For each drug, six models were built to predict the same target consisting of discretized sensitivity data for that drug across the cell lines in which it was tested. The input vectors for the six models were a combined set of original unprocessed omics features, the feature selected dataset used to train the deep autoencoder, and the four layers of latent representations generated from the autoencoder. Predictive performance was compared against that of models trained on representations derived from the autoencoder structure used in preliminary work, originally reported in **4.2.2**.

4.3.2 Results

4.3.2.1 Candidate Autoencoder Structure Evaluation

For the autoencoder architectures tested, performance exhibits a classic relationship with complexity. As the number of parameters in the neural network increases, the reconstruction performance generally improves. This effect, however, is subject to diminishing returns (Figure 5A).

The clearest visible pattern in the results is the impact of the size of the first hidden layer on model performance. As the size of the first hidden layer increases from 1000 to 1500, the performance of the autoencoder generally increases, even as overall complexity remains the same. This effect is also subject to diminishing returns (Figure 5B). The impact of the number of hidden layers on model performance is also visible. The number of hidden layers is closely related to the overall complexity of the model. In general, an autoencoder with fewer layers is less complex and performs worse than an autoencoder with more layers. However, there do exist five-layer and

seven-layer autoencoders that perform just as well as all but the most complex nine-layer autoencoders (Figure 5C).

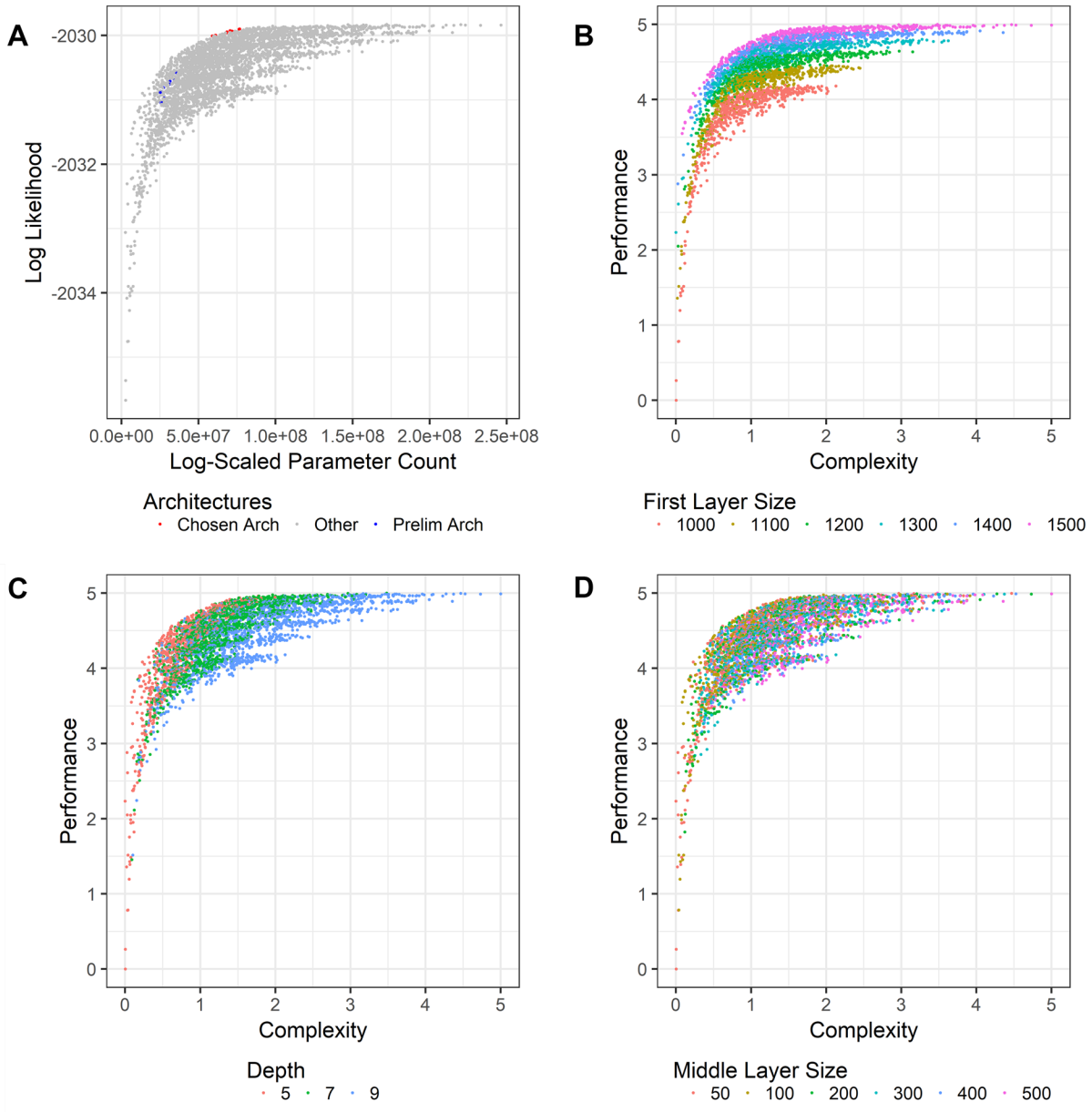

Figure 5: Autoencoder network architecture optimization. Reconstruction performance of deep neural network autoencoders relative to the number of parameters in the neural network. All panels depict the same data, with different highlights. A, Generalized location of finalized architecture vs. preliminary architecture. B, Impact of first hidden layer size. C, Impact of autoencoder depth. D, Impact of middle layer size.

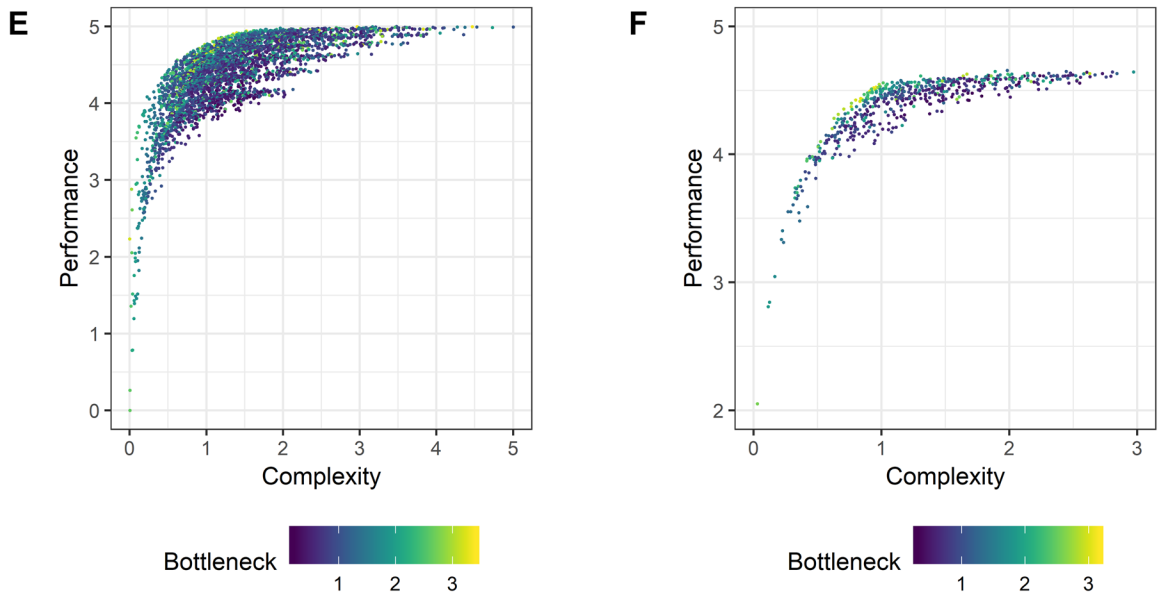

Figure 5: Autoencoder network architecture optimization (continued). Reconstruction performance of deep neural network autoencoders relative to the number of parameters in the neural network. All panels depict the same data, with different highlights. E, Impact of information bottleneck. F, Impact of information bottleneck, zoomed, 1300 first hidden layer size only.

The impact of the size of the middle hidden layer, or code layer is more difficult to discern (Figure 5D). A clear relationship is visible wherein smaller code layers are generally found in lower-complexity models. It also seems possible that at a given first layer size and constant complexity level, having a smaller code layer is favored in terms of performance. For a given first hidden layer size, autoencoders with large bottlenecks occupy the leading edge of the performance-complexity optimization curve (Figure 5E, 5F).

4.3.2.2 Optimal Autoencoder Structure Design and Evaluation

The top 20 architectures sorted by BIC are listed in Table 2. There are several commonalities visible in these high-performing architectures. First, they are all either five or seven

layers deep, suggesting that nine-layer architectures are too complex. Second, the size of the first hidden layer is large. In fact, it is 1500 for all of them. Third, the size of the code layer is small – either 50, 100, or 200. Fourth, regardless of the number of layers in the autoencoder structure, there is an extreme bottleneck between the second and third hidden layers.

Table 2: Autoencoder network architecture optimization.

Architecture Structure	Layer Depth	Scaled BIC
1500 1357 50 1357 1500	5	4.655
1500 1317 50 1317 1500	5	4.651
1500 1479 50 1479 1500	5	4.646
1500 1401 116 100 116 1401 1500	7	4.646
1500 1104 133 50 133 1104 1500	7	4.645
1500 1135 50 1135 1500	5	4.645
1500 1366 50 1366 1500	5	4.645
1500 1156 73 50 73 1156 1500	7	4.644
1500 1325 111 100 111 1325 1500	7	4.644
1500 1282 82 50 82 1282 1500	7	4.643
1500 1260 100 1260 1500	5	4.643
1500 1129 100 1129 1500	5	4.641
1500 1148 200 1148 1500	5	4.638
1500 1177 110 100 110 1177 1500	7	4.638
1500 1228 100 1228 1500	5	4.637
1500 1150 50 1150 1500	5	4.637
1500 1373 100 50 100 1373 1500	7	4.637
1500 1314 100 1314 1500	5	4.635
1500 1270 71 50 71 1270 1500	7	4.634
1500 1496 112 100 112 1496 1500	7	4.633

These results generally agree with the overall trends observed in the previous section. Based on this information, a seven-layer autoencoder structure with layers 1500, 1200, 110, 100, 110, 1200, 1500 was created for use in subsequent experimentation. Aggregate predictive performance of the best models using hidden latent variables from the optimized autoencoder as predictive features did not outperform the corresponding models from the preliminary autoencoder. As before, the average AUROC of the best models based on hidden latent variables was 0.73 for both autoencoders. A paired t-test comparing the performance of these two groups of models failed to reject the null hypothesis.

However, an examination of the predictive performance of individual hidden layers reveals statistically significant differences between the predictive performance of features derived from the second ($P < 0.01$) and third ($P < 0.05$) hidden layers (Figure 6).

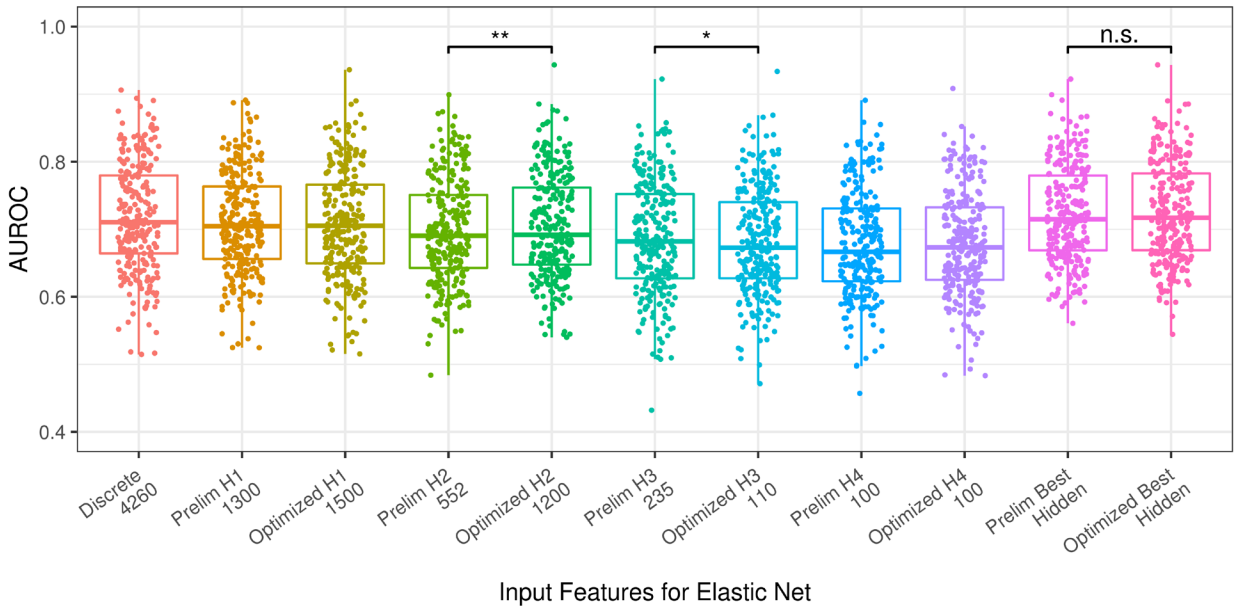

Figure 6: Learning cellular states using an optimized autoencoder. Predictive performance of elastic net models relative to predictive features used as inputs, derived from differing autoencoder structures. *, $P < 0.05$. **, $P < 0.01$.

4.3.3 Discussion

Through the evaluation of thousands of different autoencoder structures, it became possible to gain an understanding of the complex relationship between an autoencoder's reconstruction ability and the complexity of the underlying neural network. Specifically, we determined that at a given complexity level, the best-performing models favor large first hidden layers and small code layers. This structure enables the creation of an extreme bottleneck between the second and third hidden layers.

The use of the optimized autoencoder structure in the drug sensitivity predictive modeling workflow did not improve predictive performance in the aggregate compared to using the preliminary autoencoder. This is likely because the absolute reconstruction performance of the two autoencoders was very similar. It is reasonable that minor differences in encoding efficacy do not translate into noticeable differences in predictive ability when using a logistic regression model with robust regularization such as elastic net.

Interestingly, some of the hidden layers exhibited differential predictive potential when moving from one autoencoder to the other. This is likely due to the significant differences in the number of nodes in those layers. For the second hidden layer, the preliminary autoencoder has only 552 nodes while the optimized autoencoder has 1200. The third hidden layers exhibit this in reverse – the preliminary autoencoder has 235 layers while the optimized autoencoder has 110. In both cases, the predictor with the larger number of nodes exhibits better aggregate performance. On the surface, this is surprising because a higher dimension feature set generally presents a more difficult regression or classification problem. However, the process of representing a complex high-dimensional dataset with a smaller-dimensional one is generally inclined to incur some loss of information, so follows that predictive performance may suffer as a result. This finding mirrors

the general within-autoencoder trend that the smaller hidden layers generally do not perform as well in aggregate as the first hidden layer, which is the largest.

The optimized autoencoder structure does not outperform the structure used in preliminary work for drug sensitivity prediction. This is likely because although the architecture structures are different, their reconstruction performance is not that far apart in absolute terms. The optimized autoencoder structure of 1500, 1200, 110, 100, 110, 1200, 1500 is the standard structure for all subsequent experiments.

4.3.4 Limitations

The optimized autoencoder structure designed as a result of the random grid search actually lies on the edge of the grid. The first hidden layer size of 1500 is the largest value explored for that particular hyperparameter. Given the convincing pattern depicted in Figure 5B, it is likely that further increasing this value could have continued to improve the reconstruction performance of the autoencoder models. Without actually exploring those values, we do not know whether they would yield meaningful increases in reconstruction ability or if the resulting increased complexity would make for a bad tradeoff.

It is difficult to attribute the behavior of the predictive modeling results to changes made to the autoencoder alone. Although these experiments are technically controlled in that the same elastic net procedure is used to build models for drug sensitivity after encoding the data with different neural networks, there is a possibility that differential interactions between the regression method and the feature engineering methods could impact the results. Evaluating a different regression or classification algorithm, perhaps one without regularization, could yield some insight

here, even if, based on preliminary work, such an algorithm is very likely to perform worse than the elastic net.

4.4 Deep Neural Network Regularization

After determination of the optimal autoencoder structure, recent advances in deep learning were evaluated for potential to improve the autoencoder. Regularization methods such as sparsification [211], dropout [212], and batch normalization of weights [213] have gained popularity at conferences and in modeling competitions due to their demonstrated ability to reduce overfitting and improve generalizability of neural network models.

Originally, the sparsification method proposed for evaluation was L1 regularization. After some consideration, this was replaced with a Kullback-Leibler (KL) divergence penalty. The KL-divergence, sometimes referred to as entropy or relative entropy, expresses a measure of disagreement or discrepancy between two probability distributions [214]. This seemed more meaningful for our purposes than the L1 norm, which simply penalizes coefficients that grow too large in absolute terms.

Dropout normalization is perhaps the simplest normalization method. By randomly excluding weight parameters in training cycles, dropout simulates the process of training many different neural networks at once. In addition, it increases training speed because a smaller number of nodes are updated during each training cycle [212].

Batch normalization is a process that seeks to coordinate the update of multiple layers in a deep neural network at once. In traditional backpropagation, each layer is updated by itself. However, in practice, oftentimes all layers are updated at once, potentially significantly changing

the distribution of inputs for a given layer at the same time the weights for that layer are being modified. Batch normalization is the process of standardizing these intermediate values, so a change in inputs during a weight update does not change the distribution [215].

In this section, I discuss the evaluation of the potential benefits of these modifications on the optimized deep neural network structure.

4.4.1 Methods

The optimized autoencoder structure determined in **4.3.2.2** was used to encode integrated omics features from the GDSC for elastic net regression as described in **2.4.1.3**, **4.2.1.2**, and **4.3.1.3**. For each drug, six models were built to predict the same target consisting of discretized sensitivity data for that drug across the cell lines in which it was tested. The input vectors for the six models were a combined set of original unprocessed omics features, the feature selected dataset used to train the deep autoencoder, and the four layers of latent representations generated from the autoencoder. The only modification from previous methods was the addition of one of three regularization methods – KL-divergence, dropout, or batch normalization, to the first hidden layer of the autoencoder. In the implementation of these three methods, there was only one hyperparameter to specify. The dropout rate was set at 0.5, or 50%. Applying dropout is effectively training and sampling from a probability distribution of network architectures. A dropout rate of 0.5 creates the highest variance for this distribution, creating the strongest possible regularization effect.

4.4.2 Results

Predictive performance of elastic net models built using regularized representation layers was compared against the same models built using an autoencoder without regularization, originally reported in **4.3.2.2**. No statistically significant differences in performance was found using any of the three regularization methods (Figure 7).

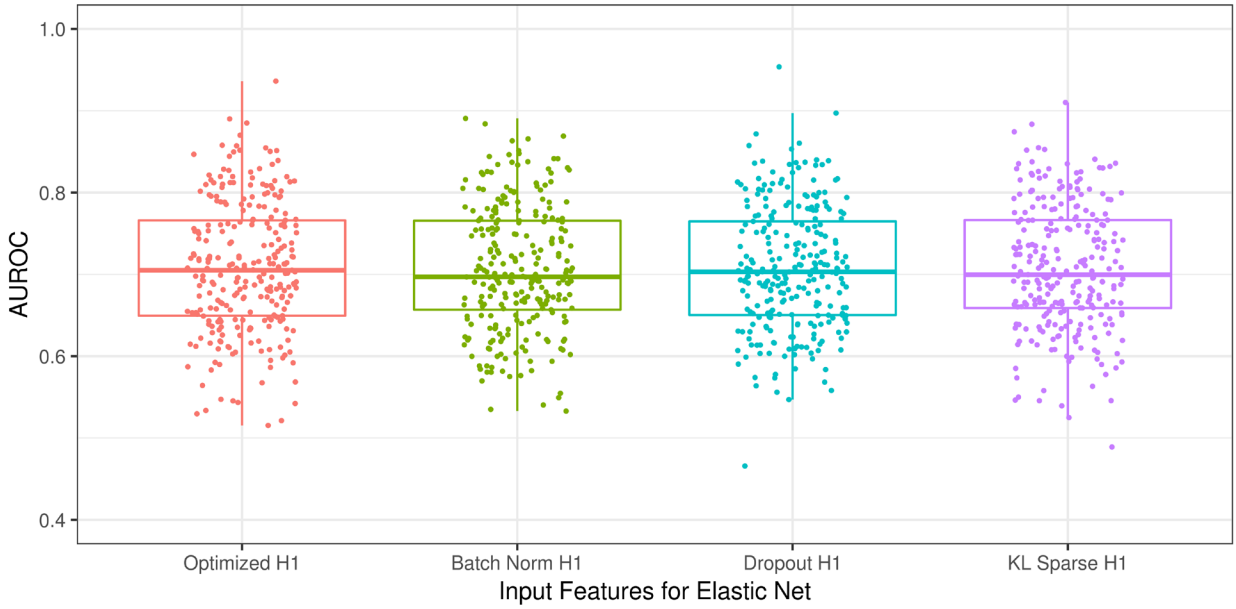

Figure 7: Learning cellular states using regularized autoencoders. Predictive performance of elastic net models relative to predictive features used as inputs, derived from a single autoencoder structure with various forms of regularization. H1 = hidden layer 1.

4.4.3 Discussion

There are several reasons why regularization did not significantly improve the predictive. First, the autoencoder is trained in an unsupervised fashion. As there is no exposure to the target

result data, there can be no overfitting to those results. Second, we employ early stopping when training the autoencoder. This is a very conservative method for preventing a highly complex model, such as a neural network, from learning spurious relationships that may randomly exist in the training data [216]. Third, our predictive modeling utilizes elastic net regression, which is a logistic regression with both L1 and L2 regularization. Taken together, these three characteristics put our model at an extremely low risk for overfitting. It is not surprising then, that additional regularization techniques designed specifically to combat overfitting did not improve predictive performance.

4.4.4 Limitations

This investigation of the impact of regularization was conducted after determining an optimized deep neural network structure. This was done for efficiency – if regularization was to be studied before settling on a single architecture, then each of the three regularization methods would need to have been tested on multiple architectures.

However, this raises the possibility that the reason regularization caused no improvement in performance is because we had already tailored the autoencoder to our application. We did not seem to have overfit, since we would have likely seen an improvement with regularization. Instead, by selecting an autoencoder architecture with the right complexity, we had already gained whatever benefit potentially existed in applying regularization to the autoencoder. It is conceivable that with an unoptimized and excessively large autoencoder, there may be a benefit to employing regularization, especially if the number of training cycles is not restricted by early stopping. There is evidence in machine learning literature that large networks can perform very well in practical scenarios if regularized appropriately [217]. If that is indeed the case, then a more efficient method

of arriving at a working model would be to simply take an overly complicated autoencoder and train it with regularization. This would undoubtedly be faster than testing 3600 different autoencoders. However, such size compromises interpretability, which is undesirable for biomedical applications.

5.0 Aim 2: Determine Potential for Clinical Impact

Aim 2 focuses on exploring the potential therapeutic benefit from a successful translation of the models created and evaluated in preliminary work and refined in Aim 1. Although deep learning based predictive models performed favorably when compared to rule-based genomic biomarker models in preliminary work, a collection of such rule-based models does not create an accurate representation of clinical practice. The proportion of patients who are approved to attempt molecularly targeted therapy is small, and the rest are prescribed a series of nonspecific cytotoxic medications according to national or institutional guidelines [5].

A true evaluation of the potential clinical impact of accurate drug prediction models must compare the ability to select targeted and nonspecific medications against the standard of care. Using a group of tumor cell lines derived from the same cancer sub-type to simulate a cohort of patients, it is possible to run a cell line "clinical trial" to accomplish this task. I discuss the details of this experiment in the following sections.

5.1 Cancer Cell Line Trial

In order to run this experiment, the cancer type selected must satisfy two conditions. First, there must be a substantial number of cell lines of that cancer type in the available pharmacogenomics data. Second, there must be adequate representation of the targeted and cytotoxic medications used to treat this type of cancer in the compounds tested by the pharmacogenomics experiment. An analysis of the GDSC data reveals that two groups of cancer

cell lines satisfy both conditions. Thus, two separate trials will be run on the non-small-cell lung carcinoma (NSCLC) and upper aero-digestive tract (ADT) tumor cell lines.

Using current National Comprehensive Cancer Network (NCCN) standard of care guidelines for a particular cancer, it is possible to simulate attempted treatment regimens for each cell line as it progresses through the standard of care. The effectiveness of this standard of care model can then be compared against a theoretical clinical decision support system that predicts the chance of success before attempting a therapy. If the computational predictive models developed from deep learning are to be clinically useful, they should consistently find an effective therapy in fewer attempts than the current standard of care. Such models would also be useful in repurposing unapproved medications for the cell lines for which the standard of care did not include an effective therapy.

5.1.1 Methods

5.1.1.1 Standard of Care Modeling

Non-small cell lung cancer

The NCCN 2018 clinical guidelines for treating advanced NSCLC with chemotherapy were cross referenced against available pharmacogenomic data from GDSC. The resulting simplified clinical guideline is outlined in Table 3. Candidate therapies are listed in general preferred order of administration.

Table 3: A simplified standard of care for NSCLC.

Therapy Type	Drug Name	Indication/Notes
Targeted Therapy	Erlotinib	*EGFR* activating mutation
	Alectinib	*EML4-ALK* rearrangement
	Crizotinib	*ROS1* rearrangement
	Dabrafenib	*BRAF* V600E
Nonspecific Chemotherapy	Cisplatin	
	Paclitaxel	
	Gemcitabine	Squamous only
	Docetaxel	
Notable Omissions	Pembrolizumab	No PD-L1 or sensitivity data
	Osimertinib	No *EGFR* T790M data
	Ceritinib	No drug sensitivity data
	Pemetrexed	No drug sensitivity data

Table 4: A simplified standard of care for ADT cancers.

Therapy Type	Drug Name	Indication/Notes
Targeted Therapy	Erlotinib	*EGFR* activating mutation
	Gefitinib	*EGFR* activating mutation
Nonspecific Chemotherapy	Cisplatin	
	5-Fluorouracil	
	Docetaxel	
	Paclitaxel	
Notable Omissions	Irinotecan	No drug sensitivity data
	Oxaliplatin	No drug sensitivity data

Upper aero-digestive tract cancer

The NCCN 2018 clinical guidelines for treating advanced esophageal or head and neck cancers were cross referenced against available pharmacogenomic data from the GDSC. The resulting simplified clinical guideline is outlined in Table 4. As before, candidate therapies are listed in general preferred order of administration, favoring targeted therapies where available.

Standard of care modeling

Using these simplified guidelines, cancer cell lines of relevant lineages were modeled as individual patients progressing through the standard of care. First, their molecular genotyping information was used to determine eligibility for targeted therapy. Cell lines carrying relevant single-gene biomarkers were evaluated on targeted therapy first, by checking the GDSC drug sensitivity data described in **4.1.1.2** for a sensitivity response. If targeted therapies did not work, the patient progressed to nonspecific chemotherapy with the remainder of the cohort. Patients progressed along the standard of care until they had been successfully treated once or had been unsuccessfully treated three times. Success is defined as the administration of a therapy that was found to be effective in the discretized drug sensitivity experimental results from the GDSC. The results were recorded to compare against an alternative course of action.

5.1.1.2 AI-Supported Decision Modeling

A simple implementation of a clinical decision support system into the care workflow does not suggest new drugs from outside the existing armamentarium. Rather, it simply reorders already approved medications to create an individualized personal "standard" of care. Drugs are evaluated in order of model performance – that is, the better-performing models are used first, because they are less likely to make mistakes. Instead of paying attention to single-gene biomarkers for the administration of targeted therapy, the computational model's predictive recommendation is

followed instead. For nonspecific chemotherapy without markers, the drug is not universally administered, but instead, the computational model's predictive recommendation is followed as well.

Predictive model generation

Up until this point, the autoencoders used to create low-dimensional representations of cancer cell line omics data from the GDSC have been trained on the entire dataset. Although this is the correct course of action to produce the strongest possible model by utilizing all of the available training data, there does exist potential for leakage, as the entire predictive modeling pipeline cannot be said to have been evaluated on previously unseen data. Although this may not necessarily result in overfitting for reasons discussed in **4.4.3**, a modification was made to the predictive modeling workflow that has been used thus far as described in **2.4.1.3**, **4.2.1.2**, **4.3.1.3**, and **4.4.1**.

Given a particular cell line to be evaluated, that cell line's descriptive omics feature information and experimental drug response information is removed from the GDSC. A complete autoencoder and set of elastic net regression models are trained on the remaining data using the same methodology as before. From the various regression models trained for a particular drug, the best one is selected on the basis of cross-validated AUROC. In order to generate a drug-cell line effectiveness call, the cell line's descriptive features are encoded by the autoencoder, and evaluated using the selected best model. This leave-one-out process is repeated for every cell line to be evaluated. Given the 109 NSCLC cell lines and 79 ADT cell lines in the GDSC, this process is computationally extremely intensive, although it is far more tractable than performing the process on the entire dataset. For this reason, this experimental design modification was adopted for this experiment only.

Drug sensitivity predictions generated in this manner were used to simulate recommendations from an AI clinical decision support system following patients through the standard of care. As with the standard of care modeling described in the previous section, patients were followed until they had been successfully treated once or unsuccessfully treated three times. As before, success is defined as the administration of a therapy that was found to be effective in the discretized drug sensitivity experimental results from the GDSC. Results were recorded for comparison to the standard of care.

5.1.2 Results

Non-small cell lung cancer

Sankey diagrams for visualizing the path of 109 NSCLC cell lines through the standard of care and model-assisted decision making can be found in Figure 8. For 19 of these cell lines, there is no effective treatment in the list of approved medications. The remaining 90 pass through the different care modalities with varying results.

Following the standard of care, 22 patients are successfully treated on the first attempt. 37 are successfully treated on the second attempt, bringing the total to 59. One patient is successfully treated on the third attempt, resulting in a total of 60 patients receiving an available, effective therapy within three regimens, an overall success rate of 67%.

Following the AI-supported standard of care, 70 patients are successfully treated on the first attempt. 13 are successfully treated on the second attempt, bringing the total to 83. Four patients are successfully treated on the third attempt, resulting in a total of 87 patients receiving an available, effective therapy within three regimens, an overall success rate of 97%.

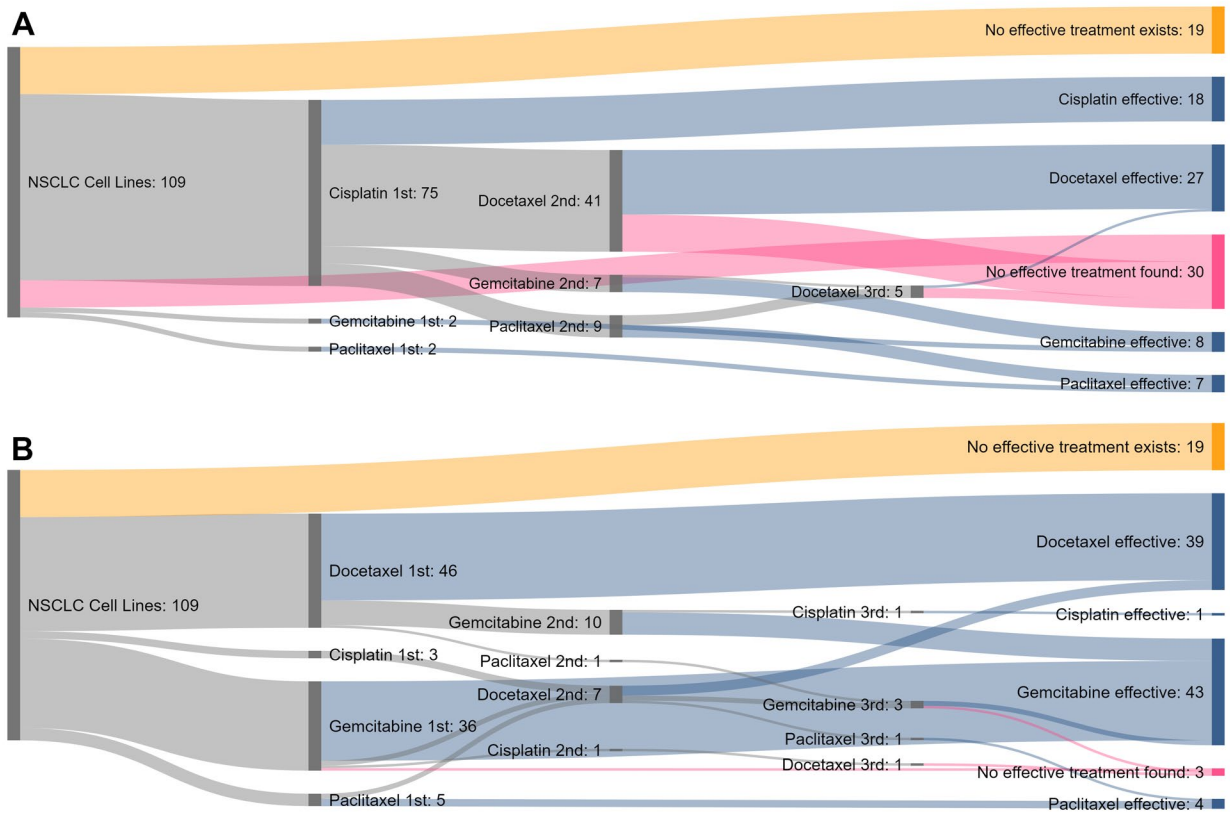

Figure 8: Evaluating the NSCLC standard of care with a cell line trial. Sankey diagrams depicting the therapeutic fate of NSCLC cell lines in A, Standard of care, and B, AI-assisted standard of care.

Following the standard of care, an average of 2.70 therapies are attempted before either finding an effective treatment or running out of options. In contrast, the AI-supported standard of care averages 1.22 therapies.

Upper aero-digestive tract cancer

Sankey diagrams for visualizing the path of 79 ADT cell lines through the standard of care and model-assisted standard of care can be found in Figure 9. For 15 of these cell lines, there is no effective treatment in the list of approved medications. The remaining 64 pass through the different care modalities, again with varying results.

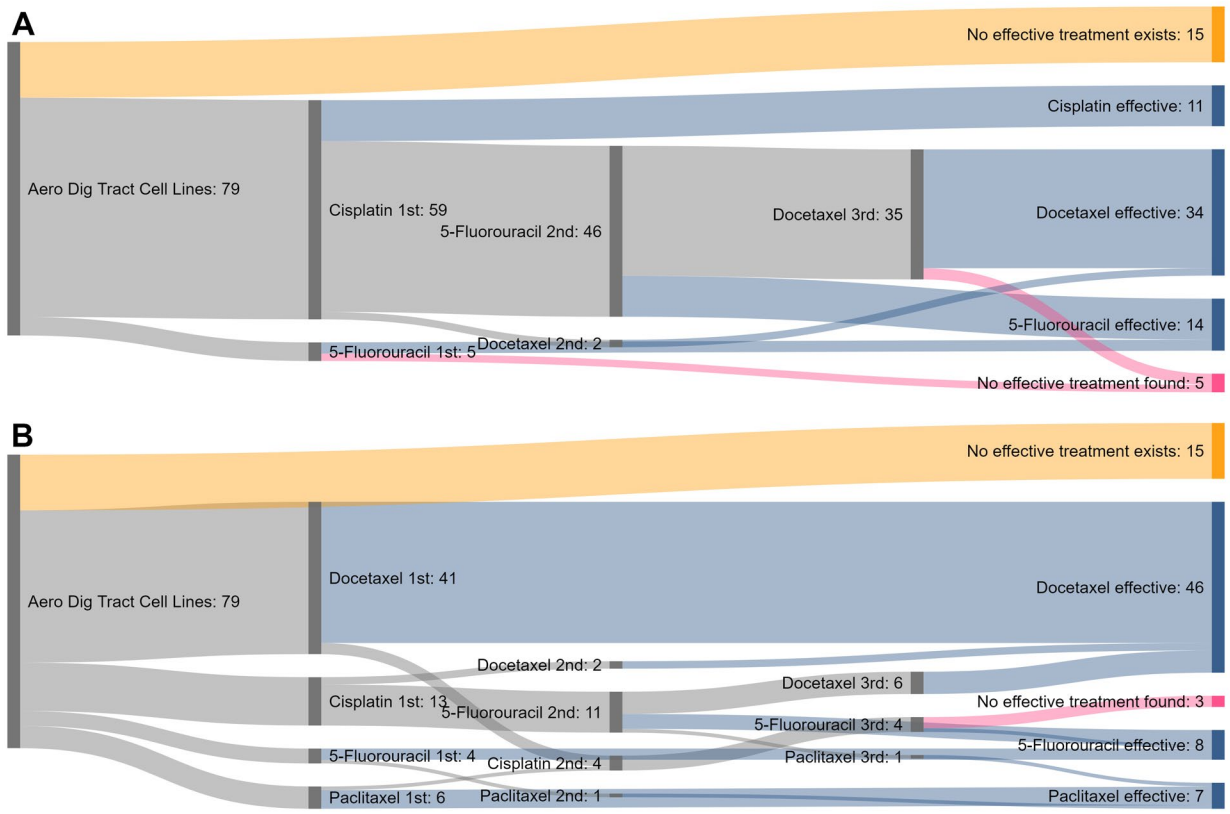

Figure 9: Evaluating the aero-digestive tract cancer standard of care with a cell line trial. Sankey diagrams depicting the therapeutic fate of aero-digestive tract cancer cell lines in A, Standard of care, and B, AI-assisted standard of care.

Following the standard of care, 14 patients are successfully treated on the first attempt. 13 are successfully treated on the second attempt, bringing the total to 27. 32 patients respond well to the third therapy regimen, resulting in a total of 59 patients receiving an available, effective therapy within three regimens, an overall success rate of 92%.

Following the AI-supported standard of care, 46 patients are successfully treated on the first attempt. 13 more are successfully treated on the second attempt, bringing the total to 53. Finally, 8 patients are treated on the third try, resulting in a total of 61 patients receiving an available, effective therapy within three regimens, an overall success rate of 95%. Although the

final success percentage is similar, the AI-supported method treats more patients effectively sooner than following the standard of care.

Following the standard of care, an average of 2.30 therapies are attempted before either finding an effective treatment or running out of options. In contrast, the AI-supported standard of care averages 1.78 attempts.

5.1.3 Discussion

Some patterns can be seen in the results. A particularly strong AI model for Docetaxel causes it to be the first evaluated therapy for both cohorts of patients. Since it is an accurate model, it has a high success rate when assigning subjects to be treated. This strategy significantly outperforms the current standard of using Cisplatin as first-line therapy for all patients. In the NSCLC results, a significant portion of the cell line population is successfully treated with Gemcitabine by the AI. The standard of care is unable to match this performance because current guidelines do not indicate Gemcitabine for non-squamous tumors.

Despite several targeted medications being approved for either type of cancer, these therapies were not attempted in either treatment system. In the standard of care, this is due to no patients having the relevant biomarkers to indicate administration of a targeted therapy. This does reflect reality in real patient tumors, where positive biomarkers for targeted therapy are rare. In the AI-supported system, the models for targeted therapies were simply lower in priority than the best-performing nonspecific models, and patients were successfully treated before being evaluated for targeted therapies.

These findings reinforce the preceding work by supporting the conclusion that it is possible to accurately predict drug sensitivity from integrated omics data using artificial intelligence

models. These models have demonstrated superior performance when compared to the current standard of care guidelines for NSCLC and ADT tumor cell lines. The advantage is two-fold. First, a greater percentage of cell lines are successfully assigned an effective therapy in the AI-assisted standard of care than the non-assisted standard of care. Second, the AI-assisted method successfully treats more cell lines earlier in the process. Translated to the clinic, these advantages would lead to reduced costs, side effects, and lost time through the avoidance of ineffective therapies.

5.1.4 Limitations

Although these findings are promising, it is unlikely that a random assortment of cancer cell lines can really effectively approximate a population of real-world patients. This is because the distribution of cancer subtypes in the cell lines is likely to be different from the corresponding distribution in a human patient population. While we have demonstrated that effective predictive models could potentially provide a lot of value to clinical care, the delivery of that value relies on the models actually working in the clinical environment. Without some indication of generalizability, it is difficult to progress cell line based models past proof of concept. The next chapter addresses this limitation.

6.0 Aim 3: Explore Generalizability to Organoids

Aim 3 focuses on determining whether the models being trained so far will properly generalize to new data. Although the deep learning drug prediction models successfully generalized to another large pharmacogenomics experiment in **2.4.3**, the ultimate intention of these models is not to predict the effectiveness of drugs on cell lines, but to predict drug sensitivity of clinical cancers. However, due to the same lack of clinical data that precludes the training of these predictive models using clinical samples in the first place, it is difficult to evaluate their performance on real clinical tumors. An intermediate step is to determine whether computational models trained on cell line data can predict drug sensitivity of patient-derived organoids, which more closely resemble clinical tumors than immortalized cell lines.

As organoid culture technique is still a relatively new technology [218], there is not an abundance of data available for this purpose. However, three recently published studies contain tissue-specific pharmacogenomic datasets for bladder [114], colorectal [115], and liver [112] cancer organoids. While these datasets are not large enough to be used to train predictive models, they can be used to test them.

There are only two complications. First, the organoid datasets are small. The bladder, colorectal, and liver datasets have 11, 19, and 5 samples, respectively. Discretizing the gene expression data via mixture fitting as described in **4.1.1.1** will not be possible. Models will need to be built using continuous data. Second, as the organoid experiments are very recent, they did not quantify gene expression using microarrays, instead opting for RNA-Seq. This difference must be reconciled before models trained using microarray features can be tested using data collected using RNA-Seq.

A variety of techniques exist for this propose. Some, such as probe region expression estimation [219], and variance modeling at the observational level [220] require sample matching, which is not possible in this case. Other methods include training distribution matching [221], quantile normalization [222], feature specific quantile normalization [223], and nonparanormal transformation [224]. The ideal method will preserve internal data dependencies while normalizing the overall distribution of data, stripping experiment-specific, methodology-dependent, batch effects.

Once a suitable data normalization method has been selected, it may be used to normalize the organoid datasets, enabling the external validation experiment that is the main goal of this aim. I discuss the evaluation of normalization methods and subsequent organoid generalizability experiment in the following sections.

6.1 Gene Expression Normalization

A straightforward method to determine whether two datasets are normalized relative to each other is to combine the datasets into one, and then perform an unsupervised clustering on the combined dataset. Data from two different sources will tend to separate into distinct clusters if batch or platform effects are not addressed beforehand. In contrast, after successful normalization, the combined dataset should behave like a single source, and will cluster according to some other criteria or feature. Thus, the ability for competing data normalization procedures to achieve this effect can be a useful selection criterion.

Although the purpose of finding an appropriate data normalization method is to enable application of predictive models trained on cell lines to be evaluated in organoids, the organoid

datasets themselves are too small to run this experiment. However, publicly accessible RNA-Seq data from The Cancer Genome Atlas (TCGA) can be used instead [225].

In this experiment, TCGA clinical samples characterized by RNA-Seq are combined with GDSC cell line samples characterized using microarrays, and this combined set of samples is clustered according to gene expression profile. Without normalization, the methodology and data source effects will separate the TCGA samples from the GDSC samples. Different normalization methods can be applied to one or both datasets before they are combined and clustered once again. If the normalization is effective, then samples from the two different experiments should not appear different due to that reason alone, and the samples should separate according to tissue type, that is, TCGA breast cancers should cluster with GDSC breast cancer cell lines.

6.1.1 Methods

6.1.1.1 Data Retrieval and Feature Engineering

Log2 normalized RNA-Seq data describing the TCGA was downloaded from the UCSC Xena browser [225]. For simplicity, the seven most abundant tissue types within the dataset were extracted, and the rest of the dataset was discarded. The retained dataset contained cancers of the breast, skin, kidney, colon, rectum, lung, and head/neck.

Microarray gene expression data from the GDSC was obtained in the form of a pre-processed Robust Multi-Array Averaged dataset [102], as in **4.1.1.1**. However, instead of conducting feature selection, cancer cell lines corresponding to the seven TCGA tissue types selected for the experiment were extracted, and the rest of the dataset was discarded.

Gene names for the TCGA dataset were renamed from the Human Genome Organisation (HUGO) nomenclature to Ensembl gene IDs to match the GDSC data. Genes that were not

measured in both datasets were discarded from both datasets. At the end of this procedure, the GDSC clustering dataset contained 278 cell lines, and the TCGA clustering dataset was composed of 3527 clinical samples. All samples in both datasets had complete gene expression measurements on 14,511 genes.

6.1.1.2 Clustering

The combined dataset of GDSC cell lines and TCGA samples were clustered according to the gene expression features using k-means. Naturally, k was set at 7, since a successful clustering operation would be expected to separate the cancers by tissue of origin. In addition to this control experiment without normalization, five clustering experiments were completed after normalization by one of the following methods: quantile Gaussian normalization, feature-specific quantile normalization, feature-specific linear transformation, nonparanormal transformation, and training distribution matching. For normalization methods that require the designation of a reference dataset and a transform dataset, the TCGA data were transformed using the GDSC data as the reference.

6.1.2 Results

Clustering the data without any normalization was not a success (Figure 10A). All of the GDSC cell lines clustered into a single group, while the TCGA clinical samples separated neatly by tissue type. Quantile normalization and training distribution matching achieved virtually the same results as the control (Figure 10B, 10F). Linear transformation of the clustering features was counterproductive, preventing proper clustering of the TCGA samples while not helping differentiation of the GDSC cell lines (Figure 10D).

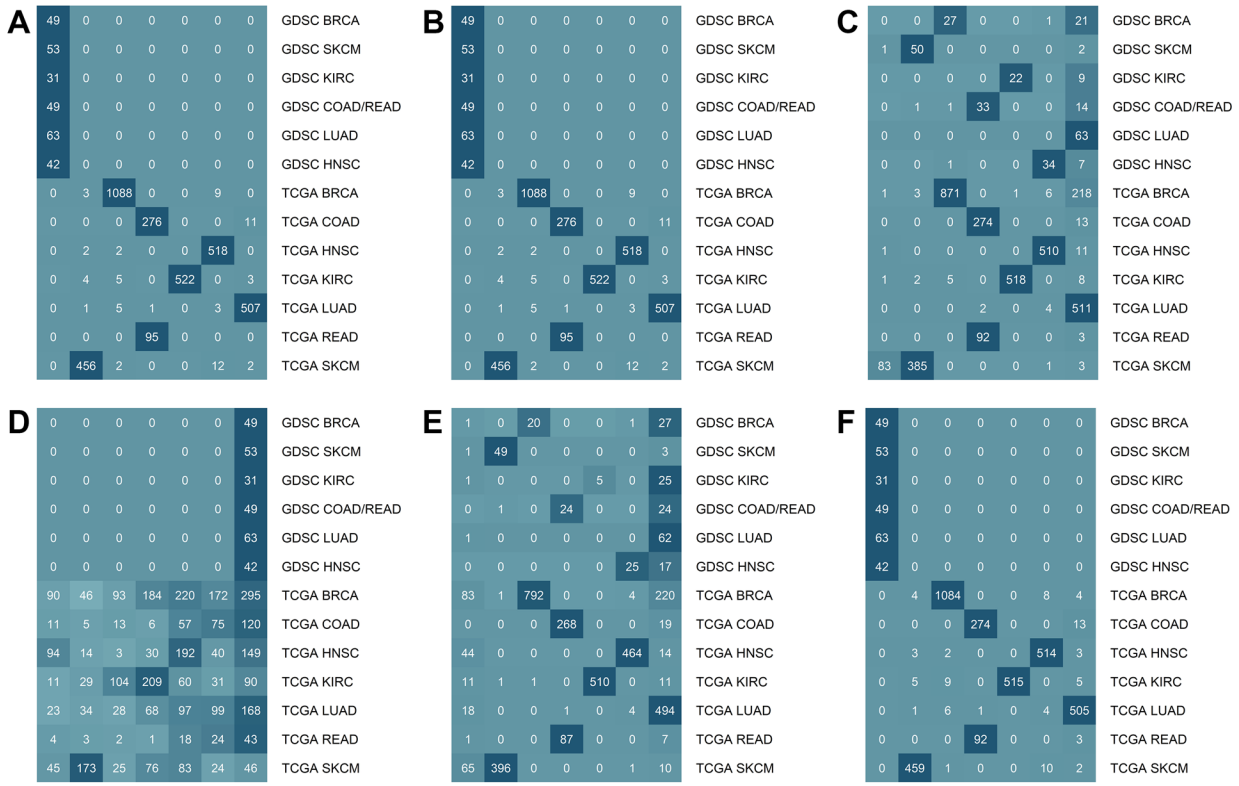

Figure 10: K-means clustering of a combined GDSC and TCGA dataset. Before clustering, the datasets were preprocessed with A, No normalization, B, Quantile gaussian normalization, C, Feature-specific quantile normalization, D, Feature-specific linear transformation, E, Nonparanormal transformation, and F, Training distribution matching.

The two normalization methods that saw success were feature-specific quantile normalization, and nonparanormal transformation (Figure 10C, 10E). Both of these methods failed to distinguish a cluster containing the TCGA lung samples, but otherwise managed to generate mono-origination clusters with data from both TCGA and GDSC.

6.1.3 Discussion

Normalization of these samples for clustering is a difficult task, because the normalization process must overcome two challenges. First, we are attempting to combine data from cell lines with data from clinical tissue. Second, the cell lines were assayed by microarray while the clinical tissue were profiled using RNA-Seq. The normalization method must reconcile both of these differences while maintaining the underlying gene expression information that enables successful clustering of the cancer samples.

It is not surprising then, that clustering the data without normalization does not distinguish between cancer types. Similarly, quantile normalization of the entire dataset probably changed very little, as expression levels of all genes in all samples is likely already close to normally distributed. What is surprising is that training distribution matching (TDM) failed to work at all. TDM was developed specifically to enable the transfer of machine learning models built on microarrays to be applied to data generated from RNA-seq [221]. TDM accomplishes this task by transforming test data to approximate the same distribution of expression values as the reference data without impacting the rank correlation of the datasets. This approach is probably too conservative in an application where we are also attempting to normalize across cell lines and clinical tissue.

Although the two successful methods produce similar results, they function differently. Feature specific quantile normalization operates by assuming the distribution of an individual gene across one dataset should be similar to the distribution of that same gene across the other dataset [223]. It then normalizes these values. Nonparanormal transformation, however, does not require a reference dataset to normalize against. That is because it transforms features into a multivariate Gaussian distribution, the exact form of which is estimated from the underlying data. In this

manner, the nonparanormal actually seeks to represent independence relationships found in the data. Using this method to individually transform both datasets relies on an assumption that similar gene expression distributions will give rise to similar multivariate Gaussian transformations, which can then be effectively merged [224].

The nonparanormal transformation seems to have a significant advantage in not requiring a reference dataset. This offers flexibility because it allows the creation of models utilizing one dataset before the nature of the evaluation dataset is known. Based on this feature, the nonparanormal transformation was selected as the preferred normalization method for subsequent experiments.

6.1.4 Limitations

Although we explored a selection of parametric and nonparametric normalization methods in this experiment, the evaluation was not exhaustive. There exist numerous other algorithms and tools for normalizing datasets for interoperability. In addition, individual methods we did evaluate can often be implemented in multiple ways to varying effect. In the end, practical concerns prevented us from exhaustively optimizing this component of the research. However, as it is a critical piece, we did ensure that we found a working method and a promising backup strategy before moving on.

6.2 Nonparanormal Transformation

Before utilizing the nonparanormal transformation in an experimental setting, we must first verify that it does not negatively impact the existing modeling workflow. Using the nonparanormal transformation on a gene expression dataset generates a collection of predictive features with a multivariate Gaussian distribution centered around 0. This can be integrated smoothly into the existing modeling workflow by centering the distribution around 0.5 and linearly scaling the transformed dataset into the range [0,1].

6.2.1 Methods

Microarray gene expression data from GDSC was obtained in the form of a pre-processed Robust Multi-Array Averaged dataset [102], as in **4.1.1.1**. However, instead of performing feature selection, the nonparanormal transformation was applied. The resulting expression dataset was then linearly transformed. As it is generally not the best idea to do feature selection on transformed data, the 3,108 genes selected in **4.1.1.1** were carried over and extracted from the nonparanormal transformed GDSC gene expression dataset. This dataset was then combined with the copy number and mutation data from **4.1.1.1**, creating a deep learning dataset of 4260 features in 963 cell lines. This dataset is directly comparable with the dataset generated in **4.1.1.1**, with the only difference being that the gene expression data is transformed and continuous, rather than discrete.

This data was then used to train a deep autoencoder as described in **4.2.1.1**. Linearly scaling the transformed expression data allows the use of the same autoencoder configuration. No changes to activation or cost functions are necessary. After training, the autoencoder was used to encode the input GDSC dataset, and elastic net was used to generate logistic models for drug sensitivity

prediction as in **2.4.1.3**, **4.2.1.2**, and **4.3.1.3**. For each drug, six models were built to predict the same target consisting of discretized sensitivity data for that drug across the cell lines in which it was tested. The input vectors for the six models were a combined set of original omics features including nonparanormal transformed expression data, the feature selected dataset used to train the deep autoencoder, and the four layers of latent representations generated from the autoencoder.

6.2.2 Results

The elastic net models trained with all omics features including transformed gene expression data achieved an average area under the receiver operating characteristic of 0.73 (Figure 11).

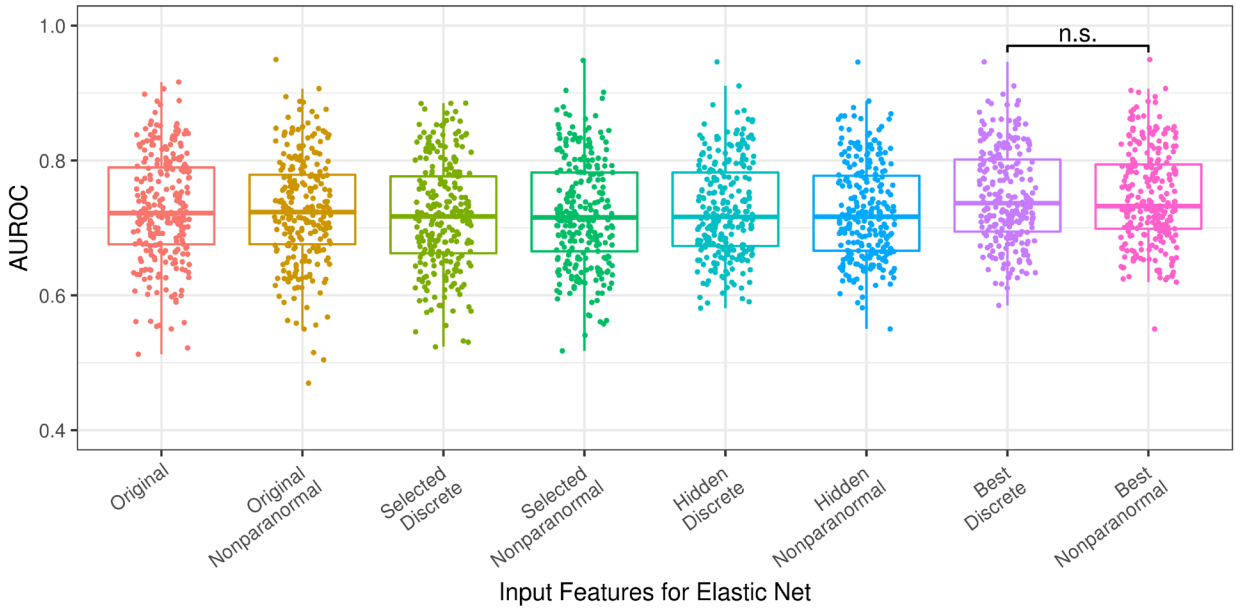

Figure 11: Learning cellular states using transformed expression data. Predictive performance of elastic net models relative to predictive features used as inputs to the same autoencoder structure, including discretized or continuous gene expression data.

Applying feature selection to this dataset did not improve overall predictive performance. Aggregate predictive performance of the best models using hidden latent variables as predictive features was not higher than models trained with all omics features. When compared against predictive models trained on the corresponding dataset including discrete gene expression data from **4.2.2**, no statistically significant differences were observed.

6.2.3 Discussion

Applying the nonparanormal transformation to the gene expression data did not significantly affect the performance of drug sensitivity models trained directly on the data or on encoded features from an autoencoder. Based on these findings, it was deemed acceptable to use the nonparanormal transformation to harmonize datasets for model training and evaluation.

An additional finding here is that the decrease in performance between models trained on the original unprocessed dataset and the selected discrete dataset persists when the gene expression data are nonparanormal transformed, and thus still continuous. This indicates that the decrease in performance is a result of the feature selection, and not of the discretization.

6.2.4 Limitations

The primary purpose of this experiment was to evaluate the impact of using the nonparanormal transformation on gene expression data prior to deep learning and predictive model generation. The experimental design used was partially driven by convenience and partially by a desire for realism – the control models to compare against had already been trained and evaluated, and we wanted to see how the nonparanormal transformation would perform in the eventual

intended usage scenario alongside other omics features. However, a different approach would have excluded the copy number and mutation data from the experiment. This would remove any potential influence from those omics data classes, ensuring that any deviation seen in the predictive modeling results would be directly related to changes made in the preprocessing of the gene expression data.

6.3 Organoid Drug Sensitivity Prediction

The three organoid datasets available for evaluating predictive models trained on the GDSC consist of a bladder dataset in which 50 drugs had been evaluated on 11 organoids [114], a colorectal dataset in which 83 drugs had been evaluated on 19 organoids [115], and a liver dataset in which 29 drugs had been evaluated on 5 organoids [112]. Additionally, a pancreatic cancer cell line dataset was obtained[5], in which 302 drugs had been evaluated on 23 cell lines [109]. We decided to also evaluate model performance on this dataset for an additional test of external validity.

6.3.1 Methods

6.3.1.1 Data Retrieval and Feature Engineering

Gene expression data for the bladder (GSE103990) and colorectal (GSE65253) datasets were obtained in a pre-processed normalized counts format from the Gene Expression Omnibus.

[5] This dataset was not part of the original plan. It was actually first downloaded by mistake.

Gene expression data for the liver dataset was obtained in a pre-processed RPKM format from supplementary dataset 1 in the original publication. Gene names for the colorectal dataset were renamed from HUGO nomenclature to Ensembl gene IDs. Log2 normalization was applied to data from all three organoid experiments.

Gene expression data for the pancreatic cancer cell line dataset was obtained in a pre-processed log-normalized counts format from the Gene Expression Omnibus (GSE84023). Gene names for the pancreatic cancer cell line dataset were renamed from HUGO nomenclature to Ensembl gene IDs. Any replicate records were averaged. Gene expression data from all four studies was normalized via nonparanormal transformation.

Having been generated on different sequencing platforms and thus pre-processed through different pipelines, there existed slight variation in gene expression coverage from experiment to experiment. When cross-referenced against the 3,108 genes selected from the GDSC dataset in **4.1.1.1**, the set of genes present in all four studies plus the GDSC was found to be 2785, or 89.6%. Expression values for these genes were retained for use in evaluation, and the rest of the data discarded.

6.3.1.2 Developing Latent Representations with Deep Learning

Nonparanormal transformed, feature-selected gene expression data from the GDSC was repurposed from the previous experiment described in **6.2.1**. Gene expression data for the 2785 genes common to the four evaluation datasets were retained, and the rest of the data were discarded, leaving a deep learning dataset of 2785 features in 963 cell lines. This dataset was used to train an optimized deep learning autoencoder as described in **4.3.1.3**.

6.3.1.3 Predicting Drug Sensitivity with Latent Representations

Drug Sensitivity Data Preprocessing

Drug sensitivity data for all four evaluation datasets were obtained in the form of normalized area under dose-response curves as supplemental data from their respective publications. As with the gene expression data, replicate experiments were averaged. Each dataset was individually discretized using the waterfall method as in **4.1.1.2**.

In order to create an evaluation dataset large enough to generate meaningful performance metrics, the organoid datasets were merged at this time. Drug compounds for which fewer than 19 organoids had been evaluated or were not tested in the GDSC were discarded. This process created an organoid evaluation dataset of 69 drugs, each tested in some subset of 35 organoids.

The pancreatic cancer cell line dataset was kept separate. Cross referencing drugs tested in this dataset with drugs tested in the GDSC, target data on 82 drugs were kept, creating an evaluation dataset of 82 drugs tested in 302 pancreatic cancer cell lines.

Predictive Modeling

After training, the autoencoder from the previous section was used to encode the input GDSC dataset, and elastic net was used to generate logistic models for drug sensitivity prediction as in **2.4.1.3**, **4.2.1.2**, **4.3.1.3**, and **6.2.1**. For each drug, five models were built to predict the same target consisting of discretized sensitivity data for that drug across the cell lines in which it was tested. The input vectors for the five models were the feature selected dataset used to train the deep autoencoder, and the four layers of latent representations generated from the deep autoencoder. This process creates five predictive models for each drug. From these five, the best one is selected on the basis of cross-validated AUROC on the GDSC sensitivity training data.

In order to generate a drug-sample effectiveness call for an organoid or cell line, the target sample's preprocessed gene expression data from **6.3.1.1** is encoded by the autoencoder, and predictions are made using the chosen best model. These predictions are then evaluated against actual sensitivity calls from the relevant organoid or cancer cell line experiment. This was completed for the pancreatic cell line dataset, and the combined organoid dataset.

6.3.2 Results

The expression-only elastic net models achieved an average area under the receiver operating characteristic of 0.73 on the pancreatic cell lines dataset. On the organoid dataset, an average area under the receiver operating characteristic of 0.69 was achieved (Figure 12).

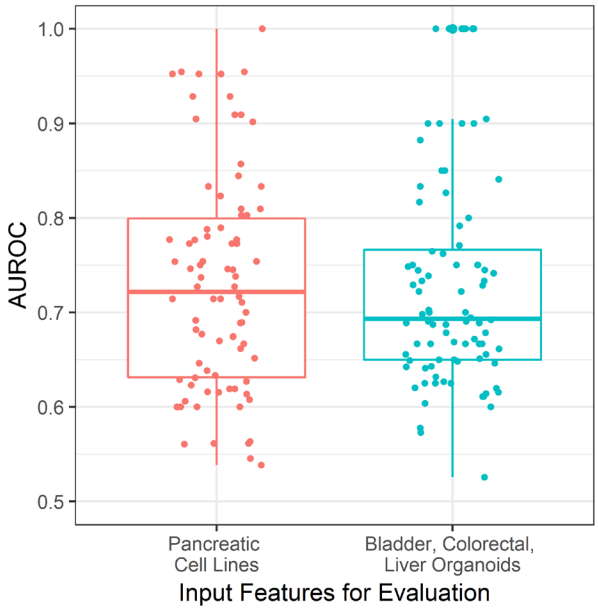

Figure 12: External validation on pancreatic cancer cell lines and bladder, colorectal, and liver cancer organoids. Predictive performance of elastic net models trained on gene expression and drug sensitivity data from the GDSC, evaluated on drug sensitivity experimental results from four external studies.

6.3.3 Discussion

These findings demonstrate that drug sensitivity models trained using data on the GDSC can accurately predict drug response in pancreatic cancer cell lines and in bladder, colorectal, and liver organoids. This is extremely promising, because these results indicate that the complex covariance relationships learned by training the deep autoencoder on the GDSC generalize not only to other cancer cell lines, but to an entirely different cancer modeling system in organoids. This is a positive indication, although not a guarantee, that this methodology may be successfully applied to clinical tumors.

Performance in the organoids was generally worse than in the pancreatic cancer cell lines. This may be a result of organoid results being more difficult to predict due to differences between organoids and cell lines, or some particular characteristic of the studies themselves. The first of these options is likely, but in reality, it could be a combination of many factors.

6.3.4 Limitations

In an ideal version of this experiment, it would have been possible to swap the training and evaluation datasets, and repeat the findings. Unfortunately, the organoid and pancreatic cancer cell line datasets are too small to use for training predictive models. Larger evaluation datasets would be extremely useful once they become available.

Although physiologically closer to clinical tumors than immortalized cell lines, organoids do have their own shortcomings. First, not all primary tumors can successfully inoculate an organoid culture, so a population of organoids should not be considered a representative sample of clinical tumors. Second, drug response can be difficult to assess experimentally, as an organoid

culture contains many different cell types, some of which are not cancerous, and it is not always clear whether a therapeutic compound is successfully destroying tumor cells or noncancerous tissue in the organoid. From an experimental point of view, these two different outcomes are difficult to distinguish, as both are characterized by a reduction in organoid size and cell density.

7.0 Aim 4: Incorporate Clinical Data with Transfer Learning

A lack of clinical drug response data does not necessitate that deep learning drug sensitivity prediction models be trained entirely on information from cell line pharmacogenomics experiments. Since the modeling design consists of an unsupervised autoencoder paired with a supervised regression, it is possible to incorporate uncategorized clinical samples in the unsupervised portion of the modeling process.

Recent work in the Lu laboratory indicates that unsupervised deep learning techniques, when applied to gene expression data, can learn a representation of the internal cellular state [226]. This suggests that the autoencoders we have been training are learning a representation of the internal state of immortalized cancer cell lines. It is possible instead to train the autoencoder on a different set of data, thereby learning a representation of a different internal state. In the interest of generating clinically applicable models, it is best that clinical samples be used for this purpose. This can be accomplished by using gene expression data from TCGA [225]. When trained on this data, the autoencoder should then capture covariance relationships that are significant in the internal cellular state of clinical tumors.

Cell line drug sensitivity data can then be used to train the supervised regression component of the drug predictive model, thus creating the best possible version of deep learning based drug sensitivity prediction models. In the following sections, I describe the training of this model and explore its clinical utility. Further, I explore the possibility of training both the unsupervised deep learning component and the supervised regression component on clinical data.

7.1 Transfer Learning

Transfer learning is a problem modality in machine learning that focuses on gaining experience in one task and using that experience to improve performance on a different but related problem. In separating our deep learning and regression steps, we have already been incorporating transfer learning in our predictive modeling process- we are using knowledge gained from encoding and reconstructing descriptive omics data to help predict drug sensitivity. In this experiment, we incorporate an additional form of transfer. By training the autoencoder on TCGA, we are transferring knowledge gained from encoding clinical tumors to help build drug sensitivity models based on experimental cell line data.

7.1.1 Methods

7.1.1.1 Predictive Feature Data

Log2 normalized RNA-Seq data describing the entire TCGA dataset was downloaded from the UCSC Xena browser [225], as in **6.1.1.1**. Before normalization, variance-based feature selection was conducted on the expression data as described in **4.1.1.1**. This yielded a list of 2831 genes exhibiting non-unimodal expression in the TCGA data. A nonparanormal transformation was applied to the entire dataset, and then gene expression values for the selected genes were extracted afterwards, yielding a dataset containing gene expression values for 2831 genes in 9649 clinical samples. Genes for which expression was not measured in the GDSC were discarded, leaving a final TCGA gene expression dataset of 2758 genes in 9659 clinical samples.

Copy number estimates for 24,776 genes measured in 10845 samples in the TCGA were obtained in a preprocessed form from the UCSC Xena browser. These values were determined by

processing Affymetrix SNP 6.0 microarray data using the GISTIC2 algorithm to produce segmented copy number variation data, which was then mapped to genes to produce gene-level estimates. These values were converted log2 form to estimated counts. Estimates ranging from 0 to 8 were linearly scaled to real values between 0 and 1. Genes with copy number estimates greater than 8 were set to the maximum normalized value of 1. Genes for which copy number estimates were incomplete or not available in the GDSC were discarded, leaving a final TCGA copy number dataset of 562 genes in 10845 clinical samples.

Mutation annotations for 22,052 genes measured in 10182 samples in the TCGA were obtained in a preprocessed form from the UCSC Xena browser. These values were obtained by generating aligned reads from whole exome sequence data using an implementation of Burrows wheeler aligner (BWA), and then calling variants with the MuTect2 pipeline. Genes with a rearrangement event or other non-silent mutation in a coding region were encoded with a value of 1. Unmutated genes or genes with silent mutations were assigned a value of 0. Genes for which mutation data were incomplete or not available in the GDSC were discarded, leaving a final TCGA mutation dataset of 434 genes in 10182 samples.

TCGA sample annotations were used to combine the gene expression, copy number, and mutation datasets into a single array for deep learning. Samples without all three data types available were discarded, leaving a final dataset of 2754 features in 8812 clinical samples.

7.1.1.2 Developing Latent Representations with Deep Learning

The TCGA dataset processed in the previous section was used to train an optimized deep autoencoder as described in **4.3.1.3**.

Microarray gene expression data from GDSC was obtained in the form of a pre-processed Robust Multi-Array Averaged dataset [102]. A nonparanormal transformation was applied to the

entire dataset. Following the transformation, this dataset was cross referenced with the feature selected gene expression dataset from TCGA. Extraneous features were discarded, leaving a GDSC gene expression dataset of 2758 genes in 1014 cell lines.

Copy number estimates and mutation annotations for GDSC were repurposed from **4.1.1.1**. These were cross referenced with their TCGA counterparts, and then combined with the GDSC gene expression data, ultimately yielding a GDSC feature dataset of 3754 data points measured in 963 cell lines.

To create a control, the GDSC autoencoder training set from **6.2.1** was repurposed and used to train an autoencoder alongside the TCGA dataset. This autoencoder differs from the one trained in that section because it uses the optimized autoencoder structure determined in Aim 1 instead of the structure used in preliminary work.

Two sets of latent representations were produced by encoding the corresponding GDSC feature sets with the two deep autoencoders, one trained on GDSC data, and the other trained on TCGA data.

7.1.1.3 Predicting Drug Sensitivity with Latent Representations

Elastic net regression was used to generate logistic models for drug sensitivity prediction using these latent representations of GDSC data as in **2.4.1.3**, **4.2.1.2**, **4.3.1.3**, **6.2.1**, and **6.3.1.3**. For each drug, nine models were built to predict the same target consisting of discretized sensitivity data for that drug across the cell lines in which it was tested. The input vectors for these nine models were the feature selected dataset used to train the deep autoencoders, and each of the four layers of latent representations generated from both deep autoencoders.

7.1.2 Results

Aggregate performance of the best models using hidden latent variables from the TCGA-trained autoencoder as predictive features did not outperform the corresponding models that used data representations from the GDSC autoencoder. As before, the average AUROC of the best models based on hidden latent variables was 0.73 for both autoencoders. A paired t-test comparing the performance of these two groups failed to reject the null hypothesis (Figure 13).

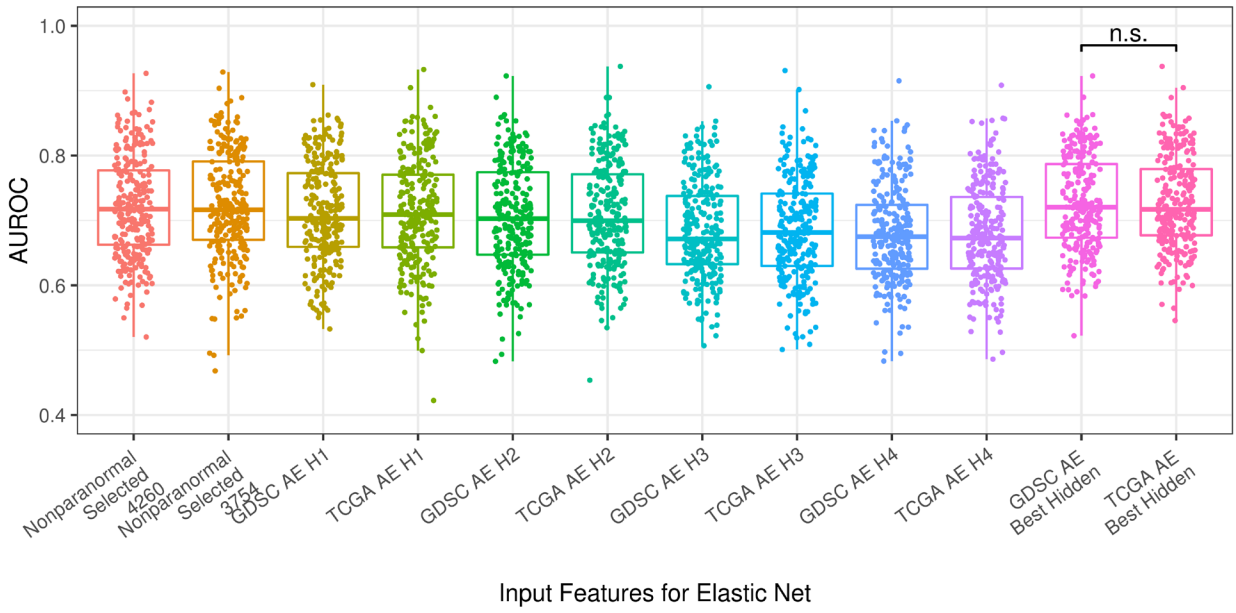

Figure 13: Learning cellular states using a TCGA autoencoder. Predictive performance of elastic net models relative to predictive features used as inputs, derived from autoencoders trained on either GDSC (GDSC AE) or TCGA (TCGA AE) integrated omics data. Results from individual layers are shown. H1-H4 = hidden layers 1-4.

7.1.3 Discussion

Predictive models trained on the two different-sized feature selected datasets have similar performance, indicating that differential feature selection on TCGA vs. GDSC most likely did not impact the other results. Further, the relationships learned by feature selection and deep learning on the TCGA clinical samples are informative when building predictive models for drug sensitivity in GDSC cell lines. This result implies that the autoencoder learns relationships between the input data features that are characteristic of cancer in general, not specifically cancer cell lines, or clinical cancer samples. It is an indication that models trained on these data representations may be widely generalizable, irrespective of their original source.

The expected outcome for this experiment was that switching the autoencoder to train on a different dataset would generally cause predictive models utilizing deep learning features to perform worse. However, there was no detectable loss of performance when using the TCGA autoencoder. In addition to the points mentioned above, this suggests that when training the autoencoder on the GDSC cell line data in this and previous experiments, we did not overfit the encoder, and thus the hidden latent representations, to the training data. Nevertheless, to completely avoid the possibility of this happening in the future, the TCGA autoencoder was designated as the standard to use for any subsequent experiments involving building models from GDSC cell lines.

7.1.4 Limitations

Ideally, this experiment is conducted with the TCGA and GDSC autoencoders being trained on the exact same set of features using data from the two different pharmacogenomics

experiments. This would ensure that differences in results would be solely due to training the autoencoder on clinical samples versus cell line data. However, this is difficult to execute in practice, because feature selection must be employed on one dataset or the other. It was ultimately determined that the actual training of the autoencoder and the selection of samples to train it are two components of a single unsupervised learning problem, and should be treated as such. This actually biases the experiment towards the GDSC autoencoder – if the TCGA autoencoder had unexpectedly performed better than the GDSC autoencoder, a case could not be made that its outperformance was the result of having encountered the descriptive information during training.

7.2 Clinical Patient Survival Prediction with Cell Line Based Models

The creation and validation of a set of predictive models incorporating integrated omics data from the TCGA in the previous section opens the door to using the rich TCGA dataset for model building or model validation. Unfortunately, while the TCGA is a great source for omics data characterizing tumors, the phenotypic data is much more limited. Therapeutic records for patients are sparse, and validated follow-up response information, such as response evaluation criteria in solid tumors (RECIST) scores, are even rarer.

However, the TCGA does have a significant amount of patient survival data for certain cancer subtypes. One of these subtypes is lung cancer. As noted in Aim 2, the therapeutic options evaluated in the GDSC contain most therapies administered to patients in the current standard of care for lung cancer. This presents a unique opportunity to validate predictive models trained from cell line dose-response experiments on real clinical tumors in a retrospective manner.

7.2.1 Methods

Clinical phenotype data for 877 lung adenocarcinoma patients and 765 lung squamous cell carcinoma patients were downloaded from the TCGA using the UCSC Xena browser [225]. These records were largely compiled from clinical questionnaires completed at time of enrollment and on subsequent physician visits. Patient records that did not include therapeutic information were discarded. This was a majority of records – 177 lung adenocarcinoma patients and 135 lung squamous cell carcinoma patients remained. These were combined into a dataset of 312 lung cancer patients. These were cross referenced against the TCGA molecular profiling data generated in **7.1.1.1**. Nine records belonged to patients without genomic sequencing data on file. These were discarded.

Next, survival data for TCGA lung adenocarcinoma patients and lung squamous cell carcinoma patients were downloaded in separate files from the UCSC Xena browser. These datasets were combined and cross referenced against the patients for which therapy information and molecular profiling features were available. Three more records were removed, bringing the dataset to 300.

Finally, the therapeutic regimens recorded were matched to specific drug prediction models created in the previous section. This matching process was done last because it had to be completed manually, as the therapeutic regimen data contains many misspellings. After this process, patient records with medications that did not match to an available predictive model were discarded. 96 of the remaining samples were lost at this step, leaving 204. The set of predictive models chosen for the experiment were built using GDSC feature data encoded by an autoencoder trained on TCGA data because this set of models was believed to be at the lowest risk of accidentally overfitting the GDSC cell line dataset.

Finally, the clinical staging of the tumors was examined. The vast majority of tumors were found to be in stages Ia through IIIa. A decision was made to remove a small number of late-stage (IIIb or IV) samples from the experiment, as these are highly advanced diseases that are unlikely to survive regardless of choice of chemotherapy. 19 of these records were removed, bringing the final number to 185.

Fully normalized, feature selected, processed integrated omics data for these 185 patients were repurposed from the previous experiment. Latent representations were generated using the TCGA-trained deep autoencoder from the previous experiment, and drug sensitivity calls were made using the relevant models. As in previous sections, the model to be used for any particular drug was selected on the basis of cross-validated AUC performance on the GDSC training set.

Any patient whose lung cancer was predicted to be sensitive to a drug they were given was labeled as a chemotherapy responder. In contrast, patients whose tumors were predicted to be insensitive to their therapy were designated as non-responders. In this manner, the population of 185 patients were sorted into two groups. A survival function was estimated for each group, and the results were compared.

7.2.2 Results

Patients classified as responders had better survival outcomes than patients classified as non-responders (Figure 14A). Median survival in the non-responder group was three years, while median survival in the responder group was greater than five years. The difference between two groups is statistically significant ($P < 0.05$). When broken down into clinical stage, responders and non-responders were evenly distributed across the varying progression levels of disease (Figure 14B).

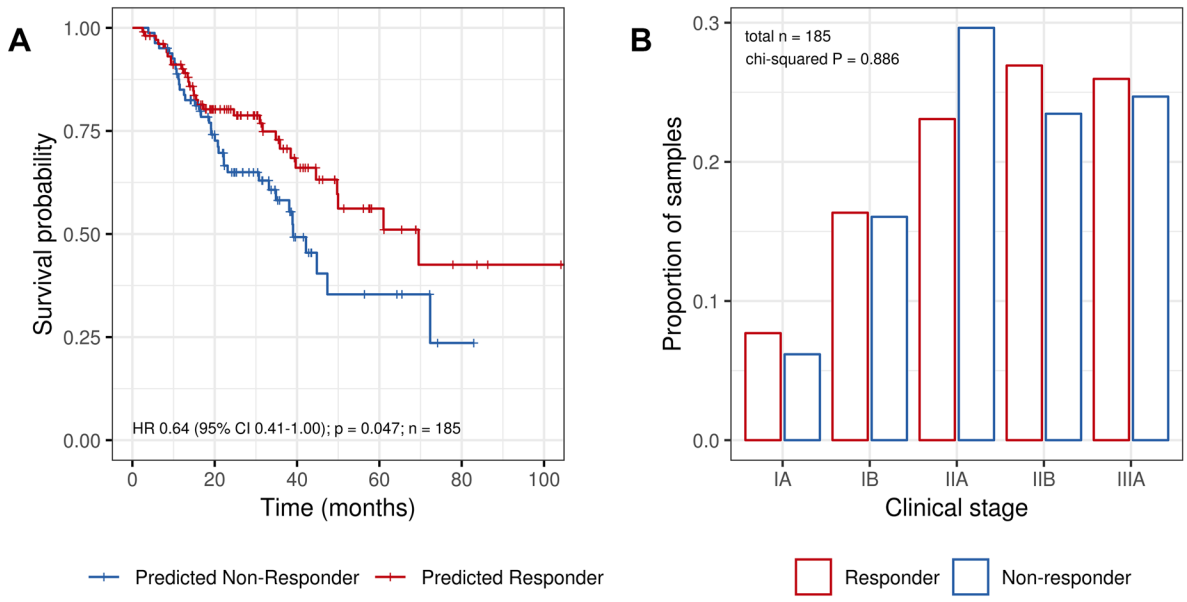

Figure 14: Predicting overall survival of TCGA lung cancer patients. A, Kaplan-Meier plot for predicted responders and non-responders. B, Clinical stage distribution of actual responders and non-responders.

7.2.3 Conclusions

In this experiment, we applied drug sensitivity models originally trained on cancer cell line pharmacogenomics data to the classification of clinical cancer samples. We found that they were able to accurately classify lung cancer patients from TCGA as chemotherapy responders and non-responders.

These results echo the findings from Aim 2 in suggesting that there exists significant opportunity to improve the current standard of care using data-driven computational modeling. These results reinforce findings from Aim 3 in suggesting that models developed on preclinical data can generalize to new tissue types, new cancer disease models, and now to clinical tumors.

This study provides the strongest external validation so far for the idea that omics data contain information that are important and useful for the prediction of cancer drug response, and that this information can be learned and encoded from large pharmacogenomics studies.

7.2.4 Limitations

This is an extremely positive result. However, this experiment investigated modeling generalizability for a relatively small number of medications in the context of a single cancer type. Although it is reasonable to believe the general conclusion that transfer learning from large pharmacogenomics studies is possible and may be achieved for other medications in other cancer types, difficulties remain (**Appendix A**). Physiological differences between immortalized cancer cell lines and real clinical tumors may cause unexpected behavior, and the complexity of modern clinical chemotherapy regimens may limit the usefulness of models learned using single-therapy experimental data.

7.3 Clinical Patient Survival Prediction with Clinical Tumor Based Models

The use of cell line based predictive models to predict the survival of clinical patients in the TCGA in the previous section was a powerful demonstration of the utility of omics data and our ability to extract useful information via deep learning. Although our decision to use cell line based large pharmacogenomics experiments to create our models is largely due to the scarcity of suitable clinical data at the time, such data are increasing in availability as patients, clinicians,

researchers, and other healthcare stakeholders continue to realize the benefits of molecular profiling.

One example of this increasing availability is the recent publication of a French dataset of 143 human colorectal tumors with molecular phenotyping data and extensive clinical annotations, including overall survival data and RECIST scores [227]. In this section, we use this dataset to train a predictive model for a common colorectal chemotherapy regimen, and evaluate it on colorectal tumors from the TCGA.

7.3.1 Methods

7.3.1.1 Predictive Feature Data

Log2 normalized RNA-Seq data describing the TCGA was downloaded from the UCSC Xena browser [225], as in **7.1.1.1**. A nonparanormal transformation was applied to the entire dataset, and then gene expression values for 3091 genes chosen via variance-based feature selection were extracted, leaving a final dataset of 3091 features in 8812 clinical samples.

Microarray gene expression data for 143 colorectal tumors from the French dataset were obtained in the form of raw Affymetrix CEL files from the Gene Expression Omnibus (GSE62050, GSE72970). Batch effects were removed using Robust Multi-Array Averaging [195]. Data corresponding to Affymetrix spike control probes were manually removed. This procedure generated an array of 19,939 gene-level expression measurements in 143 clinical tumors. This dataset was normalized against colon and rectal cancer samples extracted from the nonparanormal transformed TCGA dataset via feature-specific quantile normalization. Afterwards, the same 3091 genes were extracted, leaving a normalized descriptive dataset of 3091 features in 143 colorectal tumors. Alongside the gene expression information, chemotherapy response in the form of

RECIST scores was obtained from the Gene Expression Omnibus. Out of the 143 samples in the dataset, 32 patients had received the FOLFOX combination therapy of Folinic acid, 5-Fluorouracil, and Oxaliplatin. 81 had received the FOLFIRI combination therapy of Folinic acid, 5-Fluoruracil, and Irinotecan. No further preprocessing of the chemotherapy response data was required.

Expanded versions of Log2 normalized RNA-Seq data describing 456 colon and 166 rectal cancers from the TCGA were downloaded from the UCSC Xena browser [225]. These data were normalized against colon and rectal adenocarcinoma samples already contained in the nonparanormal transformed TCGA dataset via feature-specific quantile normalization. Afterwards, the same 3091 genes were extracted, creating a TCGA colorectal dataset of 3091 features in 622 tumors. This represents a significant expansion in data from the 382 colorectal cancers originally included in the TCGA. Clinical phenotype data describing a subset of these samples were downloaded from the National Cancer Institute's GDC Data Portal [228]. These records were manually parsed to determine which patients had received FOLFOX or FOLFIRI combination therapy. Overall survival data describing a different subset of these samples were downloaded from the UCSC Xena browser. These three datasets (gene expression, chemotherapy regimen, and survival) were cross-referenced against each other. Patient samples for which all three pieces of information were not available were discarded. This yielded an evaluation dataset of 92 colorectal tumors, all of which had received FOLFOX.

7.3.1.2 Developing Latent Representations with Deep Learning

The TCGA dataset consisting of 3091 gene expression features measured in 8812 pan-cancer clinical samples was used to train an optimized deep autoencoder as described in **4.3.1.3**.

This autoencoder was then used to create latent representations of the 32 patients who had received FOLFOX combination therapy in the French dataset, and of the 92 colorectal cancer patients who had received FOLFOX combination therapy in the TCGA.

7.3.1.3 Predicting Drug Sensitivity with Latent Representations

Elastic net regression was used to generate a logistic model for drug sensitivity prediction using the latent representation of the French data. In total, five models were built to predict the same target consisting of discrete RECIST designations for the 32 patients treated with FOLFOX in the French dataset. The input vectors for these five models consisted of the feature selected dataset used as input to the autoencoder, and the resulting four layers of latent representations generated from this input. Although the small size of the training dataset presents difficulties for cross-validation, the model to be evaluated was chosen based on performance on the training dataset. This was the fifth model, created using the fourth, smallest latent representation.

This model was used to classify the 92 colorectal cancer patients from the TCGA who had received FOLFOX combination therapy. Any patient whose cancer was predicted to be sensitive to the therapy was labeled as a chemotherapy responder. In contrast, patients whose tumors were predicted to be insensitive to FOLFOX were designated as non-responders. In this manner, the population was sorted into two groups. A survival function was estimated for each group, and the results were compared.

7.3.2 Results

Patients classified as responders had better survival outcomes than patients classified as non-responders (Figure 15). Overall survival for colorectal cancer patients is better than that of

lung patients, and the length of follow-up for this data is not long enough to estimate differences in median survival rates between the two groups. However, the difference in survival curves between the two groups is statistically significant ($P < 0.05$).

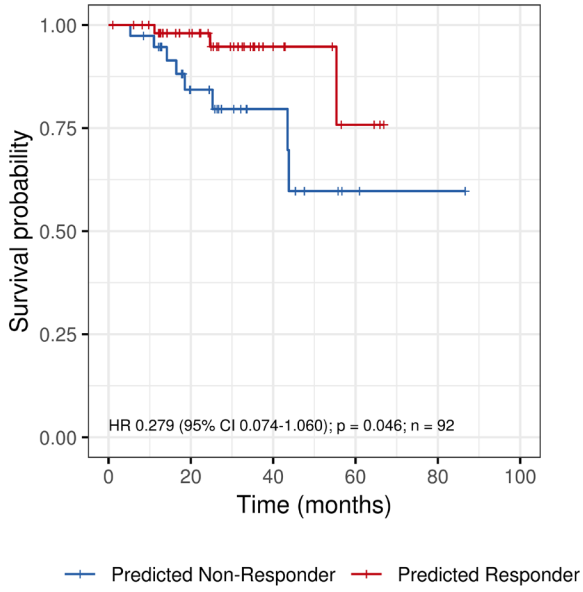

Figure 15: Predicting overall survival of TCGA colorectal cancer patients. Kaplan-Meier plot for predicted responders and non-responders.

7.3.3 Conclusions

In this experiment, we applied a chemotherapy response model originally trained on information from one clinical dataset to the classification of samples from a different clinical dataset. We found that the model was able to accurately classify colorectal cancer patients from TCGA as FOLFOX responders and non-responders.

These results indicate that the predictive methodology developed using cell line pharmacogenomics data generalizes to other training datasets. In doing so, it enables the possibility of training models completely on clinical information, as long as the necessary data is available.

Such an approach promises to create more clinically applicable models, as it is not vulnerable to spurious effects arising from differences between real tumors and preclinical cancer models. Thus, these findings advocate the expanded collection of molecular profiling data in the course of care, as well as the importance of clinical follow-up to determine the effectiveness of administered therapy.

7.3.4 Limitations

Improving the effectiveness of chemotherapy administration requires two valuable components – not prescribing an ineffective therapy, and correctly prescribing an effective one. By accurately identifying a group of chemotherapy non-responders, the models suggest that it may be worthwhile to consider alternative treatment options for that group of patients. Thus, these patient survival prediction studies demonstrate that deep learning based artificial intelligence models can accomplish the first task. Addressing the second component requires an interventional study that is outside the scope of this dissertation.

8.0 Final Conclusions and Future Work

This dissertation describes the process used to create and refine a novel data-driven approach to precision medicine. We have trained and validated our predictive models for drug sensitivity on a wide range of preclinical and clinical data. While the models themselves are an important contribution, the framework that was developed to create them is invaluable. Effectively encoding molecular features with deep learning unlocks an almost limitless number of potential applications, of which drug sensitivity prediction is just one. We have demonstrated that it is possible. We were able to accomplish the four aims we set out to achieve in **3.2**:

Aim 1: To optimize deep learning methodology.

- We incorporated newly available data from the Genomics of Drug Sensitivity in Cancer large pharmacogenomics study to improve the performance and expand the coverage of our drug sensitivity models.
- We optimized the deep learning autoencoder using Google's TensorFlow machine learning framework.
- We determined how the characteristics of an autoencoder's structure impact its reconstructive performance. We used this understanding to design a new autoencoder structure.
- We evaluated three regularization methods for deep neural networks and decided they were not needed at this time.

Aim 2: To determine the potential for clinical impact.

- We simulated two cell line clinical trials by using groups of tumor cell lines to represent a cohort of patients.

- We demonstrated that incorporating AI-supported decision making into the standard of care both increases the chance of a patient receiving an effective therapy and reduces the number of treatment regimens attempted before they receive it.

Aim 3: To generalize predictive models from cancer cell lines to cancer organoids.

- We explored five techniques for normalizing gene expression data from different experiments for interoperability.
- Using one of those methods, we evaluated our predictive models using data from three different cancer organoid studies and one external cancer cell line study.

Aim 4: To incorporate available clinical data via transfer learning.

- We trained a deep neural network autoencoder on molecular data from clinical cancer samples collected by The Cancer Genome Atlas.
- We used this autoencoder to apply our predictive models for drug sensitivity to predict survival outcomes of clinical patients.
- Using a small amount of high quality clinical data, we created predictive models trained completely on clinical data and used them to predict survival outcomes of clinical patients.

8.1 Deep Learning Autoencoders Enable Powerful Predictive Modeling

We began by building a deep learning autoencoder and using it to encode gene expression, copy number variation, and mutation annotation data from large pharmacogenomics experiments. We re-implemented this autoencoder in the TensorFlow framework to greatly increase training speed. Using this increased speed, we evaluated 3600 candidate architectures to inform the design

of an autoencoder that achieves good reconstructive performance without excessive complexity. We learned that the key features are the size of the first hidden layer and the magnitude and location of the information bottleneck. We then evaluated regularization methods designed to reduce training time and improve generalizability and determined that these techniques were not needed.

As we found in preliminary work, deep learning latent features did not generally provide the best performance when compared layer-by-layer. However, for select groups of drugs, predictive models trained on deep learning latent features performed at a level not obtainable by using raw, unprocessed predictive features. More investigation is required to determine what causes a drug to be better predicted using deep learning latent features, but the answer must be related to the statistical covariance relationships learned during unsupervised representation learning. This is one of the primary strengths of autoencoders – their utility in dimensionality reduction of large, complex datasets allows the effective application of other machine learning methods. The other advantage is that they are trained in an unsupervised manner with unlabeled data, which is plentiful, instead of labeled data, which is scarce. As more clinical data continues to be collected in the process of care, this latter advantage may fade but the former will not.

8.2 Accurate Predictive Models Exhibit Significant Potential Impact

We modeled two sets of cancer cell lines as groups of patients and conducted two simulated clinical trials. These studies were done on non-small cell lung carcinomas and upper aero-digestive tract cancers. Following the current standard of care, we found that most of the subjects in our studies do not receive an effective chemotherapy on the first or second try. Even after the third attempt, a significant group of patients are not successfully treated, even when an effective therapy

exists and is approved for their tumor type. In contrast, when using our predictive models to aid in the decision-making process, a large majority of patients were correctly treated on the first attempt. We demonstrated that effective predictive models can reduce rates of ineffective administration, leading to better outcomes and reduced costs.

8.3 Generalizable Predictive Models Function in Other Preclinical Settings

We explored a variety of data normalization techniques to unlock interoperability of molecular data from different sources. This enabled us to separate our predictive models from the dataset they were trained on, and generalize them to applications in other settings. Our models successfully predicted drug sensitivity in a different cell line dataset, and in three cancer organoid datasets. Generalization to the organoid setting is a positive indication that these predictive models may be useful in clinical tumors, because organoids are phenotypically and genotypically similar to the primary tumors from which they are derived.

8.4 Transfer Learning Unlocks Translation to Clinical Setting

Using clinical data to train heavily computational models for most purposes is impossible because the data that are required typically does not exist in the proper form or in sufficient quantity. For example, useful therapeutic data are not often available in most clinical datasets, and recorded responses to specific therapies are even less common. However, molecular profiling data

are readily available and rapidly increasing in supply. This data is useful for unsupervised learning, such as training deep learning autoencoders.

We used The Cancer Genome Atlas, a clinical cancer dataset eight times larger than the largest cell line pharmacogenomics study, to train a deep learning autoencoder to learn the cellular state of cancers. We used this autoencoder to transfer drug sensitivity prediction models trained on cell lines or clinical cancer samples to successfully predict chemotherapy outcomes in other clinical patients. This result was a revalidation of all preceding experiments, demonstrating that our modeling approach is powerful, accurate, and generalizable.

8.5 Future Work

There are very many potential areas of inquiry that future work could explore.

In unsupervised learning, a fully connected deep autoencoder is just the beginning. There exist countless variations and iterations upon the same basic idea. Like most neural networks exceeding a certain size, autoencoders are exceedingly powerful but their capacity comes at the cost of being relatively opaque models. Exploration of different autoencoder implementations that may offer enhanced transparency and explainability would be worthwhile, especially if the eventual goal is clinical translation. Examples include sparse and variational autoencoders.

In supervised learning, there are many options available other than the logistic regression method used throughout most of this research. Multitask learning, especially, holds significant potential for the drug sensitivity prediction task. Just as testing a drug typically used in bladder

cancer on breast cancer cell lines improves our predictive modeling[6], multitask learning enables insights from predicting one class of drugs, such as *EGFR* inhibitors, to be used in improving the prediction of other medications, such as *MEK* inhibitors.

In pharmacogenomics, there is an ever-expanding supply of data. The continued application of these methodologies to new data to create novel models is worthwhile. Meanwhile, the continued validation of existing models, especially using clinical data or in clinical trials is an important next step in translating this research into clinical impact.

Future work need not be restricted by the scope of past work. This technology is not limited to our current task of predicting chemotherapy outcomes in cancer, which is an important, specific application. This dissertation research has demonstrated that the general strategy of learning new representations of existing data can be very powerful. Future extensions of this work appear promising.

[6] We have tried training predictive models using only cell lines from a specific tissue type. The sample count is too small. However, adding additional samples, even from a different tissue type, can fix the problem.

Appendix A Clinical Tumor Response Prediction with Cell Line Based Models

Although high-quality clinical data that evaluates the effectiveness of a therapeutic decision is difficult to find, when it does exist it can present a rare opportunity for model evaluation. In the absence of such data, heuristic measures such as survival are used to approximate therapeutic response. Such measures can suffer from imperfect correlation [229] and troublesome effect sizes [230]. In contrast, clinical follow-up for the explicit purpose of determining the effectiveness of therapy provides validated metrics that are well-suited for model evaluation [231]. Specifically, Response Evaluation Criteria in Solid Tumors (RECIST) scores, which classify a tumor as sensitive or resistant to a therapy, are more directly comparable to the sensitive or not sensitive designations applied to pharmacogenomics experimental results in the process of creating our training datasets in **4.1.1.2** and **2.4.1.3**.

This appendix details an experiment that was not part of originally proposed work but became possible due to data availability. In this study, we apply cell line based deep learning models in an attempt to predict clinical RECIST scores. The results have interesting implications for the training and application of such models in the future.

Appendix A.1 Methods

Log2 normalized RNA-Seq data describing the TCGA was downloaded from the UCSC Xena browser [225], preprocessed, and used to train an optimized deep learning autoencoder as described in **7.3.1.2**.

Microarray gene expression, chemotherapy regimen, and therapy response data for 143 colorectal tumors from the French dataset were obtained from the Gene Expression Omnibus (GSE62050, GSE72970) and preprocessed as described in **7.3.1.1**.

Microarray gene expression data from GDSC was obtained in the form of a pre-processed Robust Multi-Array Averaged dataset [102], as in **4.1.1.1**. However, instead of performing feature selection, the nonparanormal transformation was applied. Following the transformation, this dataset was cross referenced with the feature selected gene expression datasets from the TCGA and the French dataset. Extraneous features were discarded, leaving a GDSC gene expression dataset of 3091 data points measured in 963 cell lines.

The TCGA autoencoder was then used to encode the French colorectal and GDSC gene expression datasets to create latent representations.

Elastic net regression was used to generate logistic models for predicting the sensitivity of cancers to 5-Fluorouracil and Irinotecan using these latent representations of the GDSC data. In total, five models were built to predict the same target consisting of discretized sensitivity data for each drug across the cell lines in which it was tested. The input vectors for these five models were the feature selected GDSC dataset and each of the four layers of latent representations of the GDSC dataset generated from the autoencoder.

As in previous experiments, the model to be used for evaluation for each drug was selected on the basis of cross-validated AUC performance on the GDSC training set. These models were then used to predict the therapeutic response of clinical tumors from the French colorectal dataset that had been administered combination FOLFIRI therapy. The predictions were evaluated against the actual RECIST scores recorded for those patients.

Appendix A.2 Results

Cell-line based models for predicting response to individual administrations of 5-Fluorouracil and Irinotecan were unable to predict the response of clinical tumors to combination FOLFIRI (Figure 16). The fitted probabilities provided by the models were grouped closely together in the output space. The 5-Fluorouracil model in particular appeared to perform poorly, assigning relatively lower probabilities of response to many samples that turned out to be sensitive to FOLFIRI treatment.

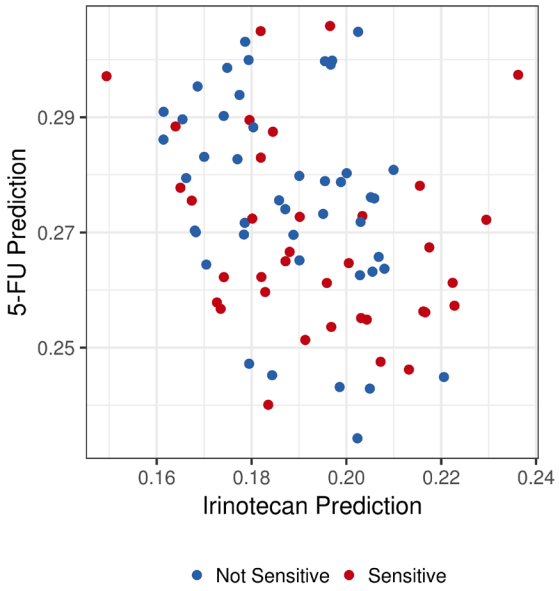

Figure 16: Predicting clinical FOLFIRI sensitivity from cell line based deep learning models for individual component medications.

Appendix A.3 Conclusions

The closeness of the fitted probabilities provided by the models suggests there may be utility in recalibrating deep learning predictive models when applying them to new data. However, there are other potential explanations for this phenomenon. The clinical samples used for evaluation are close to each other in the input space, since we are providing only colorectal samples to pan-cancer models. Logically, this would cause the outputs to be close to each other in the output space. Further investigation is required.

As for 5-Fluorouracil, there exists evidence in the literature to suggest that the metabolic effects of the compound in cell cultures may be different from its effects when administered in vivo [232]. Specifically, high *ABCC11* expression in cancer cell lines is an indicator for therapeutic resistance [233], while the high expression of the same *ABCC11* protein in clinical cancer patients is associated with better therapeutic response and longer disease-free survival [234]. It is possible that for 5-Fluorouracil, the predictive model learned a similar relationship that is valid in cancer cell lines, but is reversed in clinical tumors.

Taken together, these findings indicate that transferring information learned from cell line pharmacogenomics studies to clinical tumors is not always straightforward, especially in the context of complex combination therapies. This experiment is a reminder to exercise caution in the clinical application of pre-clinical models, and hints that the training of models on exclusively clinical data as described in in **7.3** may be the way forward.

Appendix A.4 Limitations

The models used in this study were trained exclusively on gene expression information, because mutation and copy number data were not available in the evaluation dataset. As a result, the models evaluated were weaker than they would have been if they had access to additional omics data types.

This study attempted to use two single-therapy models to predict the effectiveness of a combination therapy that consists of three medications. No single-therapy model was available for the third medication because there was no training data available to create one. The absence of this third model makes it very difficult to account for the synergistic effects that are responsible for making the combination therapy effective in the first place. Without this critical component, it is impossible to determine for certain whether the puzzle is simply incomplete or if the other two components actually do not work.

Bibliography

[1] V. Prasad, T. Fojo, and M. Brada, "Precision oncology: origins, optimism, and potential," *The Lancet Oncology,* vol. 17, no. 2, pp. e81-e86, 2016.

[2] V. Prasad, and R. P. Gale, "What precisely is precision oncology-and will it work," *ASCO post*, 2017.

[3] T. Fojo, "Precision oncology: a strategy we were not ready to deploy." p. 9.

[4] C. Rubio-Perez, D. Tamborero, M. P. Schroeder, A. A. Antolín, J. Deu-Pons, C. Perez-Llamas, J. Mestres, A. Gonzalez-Perez, and N. Lopez-Bigas, "In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities," *Cancer cell,* vol. 27, no. 3, pp. 382-396, 2015.

[5] I. F. Tannock, and J. A. Hickman, "Limits to personalized cancer medicine," *N Engl J Med,* vol. 375, no. 13, pp. 1289-1294, 2016.

[6] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, and D. Sonkin, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature,* vol. 483, no. 7391, pp. 603-607, 2012.

[7] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, and J. Soares, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature,* vol. 483, no. 7391, pp. 570-575, 2012.

[8] H. Gao, J. M. Korn, S. Ferretti, J. E. Monahan, Y. Wang, M. Singh, C. Zhang, C. Schnell, G. Yang, and Y. Zhang, "High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response," *Nature medicine,* vol. 21, no. 11, pp. 1318, 2015.

[9] J. T. Jørgensen, and M. Hersom, "Companion diagnostics—a tool to improve pharmacotherapy," *Annals of translational medicine,* vol. 4, no. 24, 2016.

[10] J. H. Breasted, *The Edwin Smith Surgical Papyrus: published in facsimile and hieroglyphic transliteration with translation and commentary in two volumes*: Chic. UP, 1930.

[11] A. Castiglioni, *A history of medicine*: Routledge, 2019.

[12] W. G. Spencer, "Celsus' De Medicina—a learned and experienced practitioner upon what the art of medicine could then accomplish," *Proceedings of the Royal Society of Medicine,* vol. 19, no. Sect_Hist_Med, pp. 129-139, 1926.

[13] J. D. Haller, "Guy de Chauliac and his chirurgia magna," *Surgery,* vol. 55, no. 2, pp. 337-343, 1964.

[14] S. W. Jackson, "Melancholia and the waning of the humoral theory," *Journal of the history of medicine and allied sciences,* vol. 33, no. 3, pp. 367-376, 1978.

[15] R. A. Kyle, "Amyloidosis: a convoluted story," *British journal of haematology,* vol. 114, no. 3, pp. 529-538, 2001.

[16] T. J. Eberlein, "Current management of carcinoma of the breast," *Annals of surgery,* vol. 220, no. 2, pp. 121, 1994.

[17] S. I. Hajdu, "A note from history: landmarks in history of cancer, part 2," *Cancer,* vol. 117, no. 12, pp. 2811-2820, 2011.

[18] R. Pearl, "Tobias Venner and his Via Recta," *Human Biology,* vol. 4, no. 4, pp. 558-583, 1932.

[19] P. Pott, "Chirurgical observations relative to… cancer of the scrotum…," *CA: A Cancer Journal for Clinicians,* vol. 24, no. 2, pp. 110-116, 1974.

[20] S. I. Hajdu, "The first tumor pathologist," *Annals of Clinical & Laboratory Science,* vol. 34, no. 3, pp. 355-356, 2004.

[21] J. T. Miller, S. Y. Rahimi, and M. Lee, "History of infection control and its contributions to the development and success of brain tumor operations," *Neurosurgical focus,* vol. 18, no. 4, pp. 1-5, 2005.

[22] W. J. Morton, *The X-ray; or, Photography of the invisible and its value in surgery*: American Technical Book Company, 1896.

[23] R. Abbe, *Radium and radioactivity*: Dorman Lithographing Company, 1904.

[24] K. Strebhardt, and A. Ullrich, "Paul Ehrlich's magic bullet concept: 100 years of progress," *Nature Reviews Cancer,* vol. 8, no. 6, pp. 473-480, 2008.

[25] F. A. Barrett, "Alfred Haviland's nineteenth-century map analysis of the geographical distribution of diseases in England and Wales," *Social Science & Medicine,* vol. 46, no. 6, pp. 767-781, 1998.

[26] R. Harrison, "SPECIMENS OF BILHARZIA AFFECTING THE URINARY ORGANS," *The Lancet,* vol. 134, no. 3439, pp. 163, 1889.

[27] P. Rous, "A transmissible avian neoplasm.(sarcoma of the common fowl.)," *The Journal of experimental medicine,* vol. 12, no. 5, pp. 696-705, 1910.

[28] L. Loeb, "Estrogenic hormones and carcinogenesis," *Journal of the American Medical Association,* vol. 104, no. 18, pp. 1597-1601, 1935.

[29] J. Cook, G. Haslewood, C. Hewett, I. Hieger, E. Kennaway, and W. Mayneord, "Chemical compounds as carcinogenic agents," *The American Journal of Cancer,* vol. 29, no. 2, pp. 219-259, 1937.

[30] J. Ewing, *Neoplastic diseases: a treatise on tumors*: WB Saunders Company, 1922.

[31] F. W. Stewart, "The diagnosis of tumors by aspiration," *The American journal of pathology,* vol. 9, no. Suppl, pp. 801, 1933.

[32] L. S. Dudgeon, and C. Wrigley, "On the demonstration of particles of malignant growth in the sputum by means of the wet-film method," *The Journal of Laryngology & Otology,* vol. 50, no. 10, pp. 752-763, 1935.

[33] I. M. Modlin, "Surgical triumvirate of Theodor Kocher, Harvey Cushing, and William Halsted," *World journal of surgery,* vol. 22, no. 1, pp. 103-113, 1998.

[34] M. Lederman, "The early history of radiotherapy: 1895–1939," *International Journal of Radiation Oncology* Biology* Physics,* vol. 7, no. 5, pp. 639-648, 1981.

[35] F. E. Adair, and H. J. Bagg, "Experimental and clinical studies on the treatment of cancer by dichlorethylsulphide (mustard gas," *Annals of surgery,* vol. 93, no. 1, pp. 190, 1931.

[36] H. C. March, "Leukemia in radiologists," *Radiology,* vol. 43, no. 3, pp. 275-278, 1944.

[37] A. Ochsner, and M. DeBakey, "Carcinoma of the lung," *Archives of Surgery,* vol. 42, no. 2, pp. 209-258, 1941.

[38] K. M. Lynch, and W. A. Smith, "Pulmonary asbestosis III: Carcinoma of lung in asbesto-silicosis," *The American Journal of Cancer,* vol. 24, no. 1, pp. 56-64, 1935.

[39] A. B. Hill, E. L. Faning, K. Perry, R. Bowler, H. Buckell, H. Druett, and R. Schilling, "Studies in the incidence of cancer in a factory handling inorganic compounds of arsenic," *British journal of industrial medicine*, pp. 1-15, 1948.

[40] D. Baltimore, "Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses," *Nature,* vol. 226, no. 5252, pp. 1209-1211, 1970.

[41] H. M. Temin, and S. Mizutami, "RNA-dependent DNA polymerase in virions of Rous sarcoma virus," *nature,* vol. 226, pp. 1211-1213, 1970.

[42] M. Vogt, and R. Dulbecco, "Virus-cell interaction with a tumor-producing virus," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 46, no. 3, pp. 365, 1960.

[43] F. P. Li, and J. F. Fraumeni Jr, "Rhabdomyosarcoma in children: epidemiologic study and identification of a familial caneer syndrome," *Journal of the National Cancer Institute,* vol. 43, no. 6, pp. 1365-1373, 1969.

[44] H. T. Lynch, M. Shaw, C. Magnuson, A. Larsen, and A. J. Krush, "Hereditary factors in cancer: study of two large midwestern kindreds," *Archives of internal medicine,* vol. 117, no. 2, pp. 206-212, 1966.

[45] G. N. Papanicolaou, "Atlas of exfoliative cytology," *Atlas of exfoliative cytology.*, 1954.

[46] W. Boyd, "Text. Surgical Pathology," 1947.

[47] S. Farber, L. K. Diamond, R. D. Mercer, R. F. Sylvester Jr, and J. A. Wolff, "Temporary remissions in acute leukemia in children produced by folic acid antagonist, 4-aminopteroyl-glutamic acid (aminopterin)," *New England Journal of Medicine,* vol. 238, no. 23, pp. 787-793, 1948.

[48] C. Huggins, and C. V. Hodges, "Studies on prostatic cancer," *Cancer Res,* vol. 1, pp. 297, 1941.

[49] M. C. Li, R. Hertz, and D. B. Spencer, "Effect of methotrexate therapy upon choriocarcinoma and chorioadenoma," *Proceedings of the Society for Experimental Biology and Medicine,* vol. 93, no. 2, pp. 361-366, 1956.

[50] G. H. Fletcher, "Regaud lecture perspectives on the history of radiotherapy," *Radiotherapy and Oncology,* vol. 12, no. 4, pp. 253-271, 1988.

[51] R. McWhirter, "Treatment of cancer of breast by simple mastectomy and roentgenotherapy," *Archives of Surgery,* vol. 59, no. 4, pp. 830-842, 1949.

[52] L. J. Old, E. A. Boyse, D. A. Clarke, and E. A. Carswell, "Antigenic properties of chemically induced tumors," *Annals of the New York Academy of Sciences,* vol. 101, no. 1, pp. 80-106, 1962.

[53] D. Stehelin, H. E. Varmus, J. M. Bishop, and P. K. Vogt, "DNA related to the transforming gene (s) of avian sarcoma viruses is present in normal avian DNA," *Nature,* vol. 260, no. 5547, pp. 170-173, 1976.

[54] H. Harris, O. Miller, G. Klein, P. Worst, and T. Tachibana, "Suppression of malignancy by cell fusion," *Nature,* vol. 223, no. 5204, pp. 363-368, 1969.

[55] S. Surget, M. P. Khoury, and J.-C. Bourdon, "Uncovering the role of p53 splice variants in human malignancy: a clinical perspective," *OncoTargets and therapy,* vol. 7, pp. 57, 2014.

[56] J. Folkman, "Tumor angiogenesis: therapeutic implications," *New england journal of medicine,* vol. 285, no. 21, pp. 1182-1186, 1971.

[57] R. J. D'Amato, M. S. Loughnan, E. Flynn, and J. Folkman, "Thalidomide is an inhibitor of angiogenesis," *Proceedings of the National Academy of Sciences,* vol. 91, no. 9, pp. 4082-4085, 1994.

[58] S. I. Hajdu, M. Vadmal, and P. Tang, “A note from history: Landmarks in history of cancer, part 7,” *Cancer,* vol. 121, no. 15, pp. 2480-2513, 2015.

[59] S. J. Pocock, *Clinical trials: a practical approach*: John Wiley & Sons, 2013.

[60] J. Goldie, "Scientific basis for adjuvant and primary (neoadjuvant) chemotherapy." pp. 1-7.

[61] G. Bonadonna, R. Zucali, S. Monfardini, M. de Lena, and C. Uslenghi, “Combination chemotherapy of Hodgkin's disease with adriamycin, bleomycin, vinblastine, and imidazole carboxamide versus MOPP,” *Cancer,* vol. 36, no. 1, pp. 252-259, 1975.

[62] S. A. Rosenberg, J. C. Yang, S. L. Topalian, D. J. Schwartzentruber, J. S. Weber, D. R. Parkinson, C. A. Seipp, J. H. Einhorn, and D. E. White, “Treatment of 283 consecutive patients with metastatic melanoma or renal cell cancer using high-dose bolus interleukin 2,” *Jama,* vol. 271, no. 12, pp. 907-913, 1994.

[63] D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, and M. Pegram, “Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2,” *New England journal of medicine,* vol. 344, no. 11, pp. 783-792, 2001.

[64] G. D. Demetri, M. Von Mehren, C. D. Blanke, A. D. Van den Abbeele, B. Eisenberg, P. J. Roberts, M. C. Heinrich, D. A. Tuveson, S. Singer, and M. Janicek, “Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors,” *New England Journal of Medicine,* vol. 347, no. 7, pp. 472-480, 2002.

[65] S. S. Neelapu, F. L. Locke, N. L. Bartlett, L. J. Lekakis, D. B. Miklos, C. A. Jacobson, I. Braunschweig, O. O. Oluwole, T. Siddiqi, and Y. Lin, “Axicabtagene ciloleucel CAR T-cell therapy in refractory large B-cell lymphoma,” *New England Journal of Medicine,* vol. 377, no. 26, pp. 2531-2544, 2017.

[66] J. Bellmunt, R. De Wit, D. J. Vaughn, Y. Fradet, J.-L. Lee, L. Fong, N. J. Vogelzang, M. A. Climent, D. P. Petrylak, and T. K. Choueiri, “Pembrolizumab as second-line therapy for advanced urothelial carcinoma,” *New England Journal of Medicine,* vol. 376, no. 11, pp. 1015-1026, 2017.

[67] S. P. Kang, K. Gergich, G. M. Lubiniecki, D. P. de Alwis, C. Chen, M. A. Tice, and E. H. Rubin, “Pembrolizumab KEYNOTE-001: an adaptive study leading to accelerated approval for two indications and a companion diagnostic,” *Annals of Oncology,* vol. 28, no. 6, pp. 1388-1398, 2017.

[68] S. H. Kaufmann, “Paul Ehrlich: founder of chemotherapy,” *Nature Reviews Drug Discovery,* vol. 7, no. 5, pp. 373-373, 2008.

[69] E. Boyland, “Experiments on the chemotherapy of cancer: The effect of certain antibacterial substances and related compounds,” *Biochemical Journal,* vol. 32, no. 7, pp. 1207, 1938.

[70] J. J. Bittner, "A genetic study of the transplantation of tumors arising in hybrid mice," *The American Journal of Cancer,* vol. 15, no. 3, pp. 2202-2247, 1931.

[71] M. Shear, *Some aspects of a joint institutional research program on chemotherapy of cancer: current laboratory and clinical experiments with bacterial polysaccharide and with synthetic organic compounds*, 1947.

[72] L. Law, T. B. Dunn, P. J. Boyle, and J. Miller, "Observations on the effect of a folic-acid antagonist on transplantable lymphoid leukemias in mice," *Journal of the National Cancer Institute,* vol. 10, pp. 179-192, 1949.

[73] B. A. Chabner, "NCI-60 cell line screening: a radical departure in its time," *JNCI: Journal of the National Cancer Institute,* vol. 108, no. 5, 2016.

[74] J. Rygaard, and C. O. Poulsen, "Heterotransplantation of a human malignant tumour to "Nude" mice," *Acta Pathologica Microbiologica Scandinavica,* vol. 77, no. 4, pp. 758-760, 1969.

[75] R. H. Shoemaker, "The NCI60 human tumour cell line anticancer drug screen," *Nature Reviews Cancer,* vol. 6, no. 10, pp. 813-823, 2006.

[76] V. T. DeVita, and E. Chu, "A history of cancer chemotherapy," *Cancer research,* vol. 68, no. 21, pp. 8643-8653, 2008.

[77] J. G. Moffat, J. Rudolph, and D. Bailey, "Phenotypic screening in cancer drug discovery—past, present and future," *Nature reviews Drug discovery,* vol. 13, no. 8, pp. 588-602, 2014.

[78] D. Swinney, "Phenotypic vs. target-based drug discovery for first-in-class medicines," *Clinical Pharmacology & Therapeutics,* vol. 93, no. 4, pp. 299-301, 2013.

[79] J. Zimmermann, G. Caravatti, H. Mett, T. Meyer, M. Müller, N. B. Lydon, and D. Fabbro, "Phenylamino-pyrimidine (PAP) derivatives: a new class of potent and selective inhibitors of protein kinase C (PKC)," *Archiv der Pharmazie,* vol. 329, no. 7, pp. 371-376, 1996.

[80] J. Zimmermann, E. Buchdunger, H. Mett, T. Meyer, and N. B. Lydon, "Potent and selective inhibitors of the Abl-kinase: phenylamino-pyrimidine (PAP) derivatives," *Bioorganic & Medicinal Chemistry Letters,* vol. 7, no. 2, pp. 187-192, 1997.

[81] R. Capdeville, E. Buchdunger, J. Zimmermann, and A. Matter, "Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug," *Nature reviews Drug discovery,* vol. 1, no. 7, pp. 493-502, 2002.

[82] M. Deininger, E. Buchdunger, and B. J. Druker, "The development of imatinib as a therapeutic agent for chronic myeloid leukemia," *Blood,* vol. 105, no. 7, pp. 2640-2653, 2005.

[83] R. Lalonde, K. Kowalski, M. Hutmacher, W. Ewy, D. Nichols, P. Milligan, B. Corrigan, P. Lockwood, S. Marshall, and L. Benincosa, "Model-based drug development," *Clinical Pharmacology & Therapeutics,* vol. 82, no. 1, pp. 21-32, 2007.

[84] F. Sams-Dodd, "Target-based drug discovery: is something wrong?," *Drug discovery today,* vol. 10, no. 2, pp. 139-147, 2005.

[85] N. P. Shah, J. M. Nicoll, B. Nagar, M. E. Gorre, R. L. Paquette, J. Kuriyan, and C. L. Sawyers, "Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia," *Cancer cell,* vol. 2, no. 2, pp. 117-125, 2002.

[86] D. C. Swinney, and J. Anthony, "How were new medicines discovered?," *Nature reviews Drug discovery,* vol. 10, no. 7, pp. 507-519, 2011.

[87] N. P. Coussens, J. C. Braisted, T. Peryea, G. S. Sittampalam, A. Simeonov, and M. D. Hall, "Small-Molecule Screens: A Gateway to Cancer Therapeutic Agents with Case Studies of Food and Drug Administration–Approved Drugs," *Pharmacological reviews,* vol. 69, no. 4, pp. 479-496, 2017.

[88] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, and M. Waltham, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature genetics,* vol. 24, no. 3, pp. 227-235, 2000.

[89] L. A. Garraway, H. R. Widlund, M. A. Rubin, G. Getz, A. J. Berger, S. Ramaswamy, R. Beroukhim, D. A. Milner, S. R. Granter, and J. Du, "Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma," *Nature,* vol. 436, no. 7047, pp. 117-122, 2005.

[90] W. E. Evans, and M. V. Relling, "Pharmacogenomics: translating functional genomics into rational therapeutics," *science,* vol. 286, no. 5439, pp. 487-491, 1999.

[91] J. Greshock, K. E. Bachman, Y. Y. Degenhardt, J. Jing, Y. H. Wen, S. Eastman, E. McNeil, C. Moy, R. Wegrzyn, and K. Auger, "Molecular target class is predictive of in vitro response profile," *Cancer research,* vol. 70, no. 9, pp. 3677-3686, 2010.

[92] B. Haibe-Kains, N. El-Hachem, N. J. Birkbak, A. C. Jin, A. H. Beck, H. J. Aerts, and J. Quackenbush, "Inconsistency in large pharmacogenomic studies," *Nature,* vol. 504, no. 7480, pp. 389-393, 2013.

[93] C. Klijn, S. Durinck, E. W. Stawiski, P. M. Haverty, Z. Jiang, H. Liu, J. Degenhardt, O. Mayba, F. Gnad, and J. Liu, "A comprehensive transcriptional portrait of human cancer cell lines," *Nature biotechnology,* vol. 33, no. 3, pp. 306, 2015.

[94] P. M. Haverty, E. Lin, J. Tan, Y. Yu, B. Lam, S. Lianoglou, R. M. Neve, S. Martin, J. Settleman, and R. L. Yauch, "Reproducible pharmacogenomic profiling of cancer cell line panels," *Nature,* vol. 533, no. 7603, pp. 333-337, 2016.

[95] C. C. L. E. Consortium, and G. o. D. S. i. C. Consortium, “Pharmacogenomic agreement between two cancer cell line data sets,” *Nature,* vol. 528, no. 7580, pp. 84, 2015.

[96] N. Pozdeyev, M. Yoo, R. Mackie, R. E. Schweppe, A. C. Tan, and B. R. Haugen, “Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies,” *Oncotarget,* vol. 7, no. 32, pp. 51619, 2016.

[97] P. Geeleher, E. R. Gamazon, C. Seoighe, N. J. Cox, and R. S. Huang, “Consistency in large pharmacogenomic studies,” *Nature,* vol. 540, no. 7631, pp. E1-E2, 2016.

[98] J. P. Mpindi, B. Yadav, P. Östling, P. Gautam, D. Malani, A. Murumägi, A. Hirasawa, S. Kangaspeska, K. Wennerberg, and O. Kallioniemi, “Consistency in drug response profiling,” *Nature,* vol. 540, no. 7631, pp. E5-E6, 2016.

[99] A. Gupta, P. Gautam, K. Wennerberg, and T. Aittokallio, “A normalized drug response metric improves accuracy and consistency of anticancer drug sensitivity quantification in cell-based screening,” *Communications biology,* vol. 3, no. 1, pp. 1-12, 2020.

[100] Z. T. Hu, Y. Ye, P. A. Newbury, H. Huang, and B. Chen, "AICM: A Genuine Framework for Correcting Inconsistency Between Large Pharmacogenomics Datasets." pp. 248-259.

[101] Z. Safikhani, P. Smirnov, M. Freeman, N. El-Hachem, A. She, Q. Rene, A. Goldenberg, N. J. Birkbak, C. Hatzis, and L. Shi, “Revisiting inconsistency in large pharmacogenomic studies,” *F1000Research,* vol. 5, 2016.

[102] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, and H. Lightfoot, “A landscape of pharmacogenomic interactions in cancer,” *Cell,* vol. 166, no. 3, pp. 740-754, 2016.

[103] B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, and J. Gould, “Harnessing connectivity in a large-scale small-molecule sensitivity dataset,” *Cancer discovery,* vol. 5, no. 11, pp. 1210-1223, 2015.

[104] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, and J. K. Asiedu, “A next generation connectivity map: L1000 platform and the first 1,000,000 profiles,” *Cell,* vol. 171, no. 6, pp. 1437-1452. e17, 2017.

[105] L. M. Heiser, A. Sadanandam, W.-L. Kuo, S. C. Benz, T. C. Goldstein, S. Ng, W. J. Gibb, N. J. Wang, S. Ziyad, and F. Tong, “Subtype and pathway specific responses to anticancer compounds in breast cancer,” *Proceedings of the National Academy of Sciences,* vol. 109, no. 8, pp. 2724-2729, 2012.

[106] R. Marcotte, A. Sayad, K. R. Brown, F. Sanchez-Garcia, J. Reimand, M. Haider, C. Virtanen, J. E. Bradner, G. D. Bader, and G. B. Mills, “Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance,” *Cell,* vol. 164, no. 1-2, pp. 293-309, 2016.

[107] J. L. Wilding, and W. F. Bodmer, "Cancer cell lines for drug discovery and development," *Cancer research,* vol. 74, no. 9, pp. 2377-2384, 2014.

[108] P. H. Cashin, H. Mahteme, W. Graf, H. Karlsson, R. Larsson, and P. Nygren, "Activity ex vivo of cytotoxic drugs in patient samples of peritoneal carcinomatosis with special focus on colorectal cancer," *BMC cancer,* vol. 13, no. 1, pp. 435, 2013.

[109] A. K. Witkiewicz, U. Balaji, C. Eslinger, E. McMillan, W. Conway, B. Posner, G. B. Mills, E. M. O'Reilly, and E. S. Knudsen, "Integrated patient-derived models delineate individualized therapeutic vulnerabilities of pancreatic cancer," *Cell reports,* vol. 16, no. 7, pp. 2017-2031, 2016.

[110] S. Y. C. Choi, D. Lin, P. W. Gout, C. C. Collins, Y. Xu, and Y. Wang, "Lessons from patient-derived xenografts for better in vitro modeling of human cancer," *Advanced drug delivery reviews,* vol. 79, pp. 222-237, 2014.

[111] M. A. Lancaster, and J. A. Knoblich, "Organogenesis in a dish: modeling development and disease using organoid technologies," *Science,* vol. 345, no. 6194, pp. 1247125, 2014.

[112] L. Broutier, G. Mastrogiovanni, M. M. Verstegen, H. E. Francies, L. M. Gavarró, C. R. Bradshaw, G. E. Allen, R. Arnes-Benito, O. Sidorova, and M. P. Gaspersz, "Human primary liver cancer–derived organoid cultures for disease modeling and drug screening," *Nature medicine,* vol. 23, no. 12, pp. 1424, 2017.

[113] D. Gao, I. Vela, A. Sboner, P. J. Iaquinta, W. R. Karthaus, A. Gopalan, C. Dowling, J. N. Wanjala, E. A. Undvall, and V. K. Arora, "Organoid cultures derived from patients with advanced prostate cancer," *Cell,* vol. 159, no. 1, pp. 176-187, 2014.

[114] S. H. Lee, W. Hu, J. T. Matulay, M. V. Silva, T. B. Owczarek, K. Kim, C. W. Chua, L. J. Barlow, C. Kandoth, and A. B. Williams, "Tumor evolution and drug response in patient-derived organoid models of bladder cancer," *Cell,* vol. 173, no. 2, pp. 515-528. e17, 2018.

[115] M. van de Wetering, H. E. Francies, J. M. Francis, G. Bounova, F. Iorio, A. Pronk, W. van Houdt, J. van Gorp, A. Taylor-Weiner, and L. Kester, "Prospective derivation of a living organoid biobank of colorectal cancer patients," *Cell,* vol. 161, no. 4, pp. 933-945, 2015.

[116] D. W. Nebert, "Pharmacogenetics and pharmacogenomics: why is this relevant to the clinical geneticist?," *Clinical genetics,* vol. 56, no. 4, pp. 247-258, 1999.

[117] L. Luzzatto, and P. Arese, "Favism and glucose-6-phosphate dehydrogenase deficiency," *New England Journal of Medicine,* vol. 378, no. 1, pp. 60-71, 2018.

[118] G. Mendel, "Experiments on plant hybrids," *English trans.), in C. Stern and ER Sherwood, The Origin of Genetics (San-Francisco: Freeman, 1906)*, pp. 2, 1965.

[119] P. E. Carson, C. L. Flanagan, C. Ickes, and A. S. Alving, "Enzymatic deficiency in primaquine-sensitive erythrocytes," *Science,* vol. 124, no. 3220, pp. 484-485, 1956.

[120] A. G. Motulsky, "Drug reactions, enzymes, and biochemical genetics," *Journal of the American Medical Association,* vol. 165, no. 7, pp. 835-837, 1957.

[121] O. Vesterberg, "History of electrophoretic methods," *Journal of Chromatography A,* vol. 480, pp. 3-19, 1989.

[122] M. Pirmohamed, "Pharmacogenetics and pharmacogenomics," *British journal of clinical pharmacology,* vol. 52, no. 4, pp. 345, 2001.

[123] S. Ball, and N. Borman, "Pharmacogenetics and drug metabolism," *Nature Biotechnology,* vol. 16, no. 2, pp. 4-5, 1998.

[124] A. Marshall, "Getting the right drug into the right patient," *Nature biotechnology,* vol. 15, no. 12, pp. 1249-1252, 1997.

[125] E. H. Corder, A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. Small, A. Roses, J. Haines, and M. A. Pericak-Vance, "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families," *Science,* vol. 261, no. 5123, pp. 921-923, 1993.

[126] M. R. Farlow, D. K. Lahiri, J. Poirier, J. Davignon, and S. Hui, "Apolipoprotein E Genotype and Gender Influence Response to Tacrine Therapy a," *Annals of the New York Academy of Sciences,* vol. 802, no. 1, pp. 101-110, 1996.

[127] J. Poirier, M.-C. Delisle, R. Quirion, I. Aubert, M. Farlow, D. Lahiri, S. Hui, P. Bertrand, J. Nalbantoglu, and B. M. Gilfix, "Apolipoprotein E4 allele as a predictor of cholinergic deficits and treatment outcome in Alzheimer disease," *Proceedings of the National Academy of Sciences,* vol. 92, no. 26, pp. 12260-12264, 1995.

[128] E. Lai, J. Riley, I. Purvis, and A. Roses, "A 4-Mb high-density single nucleotide polymorphism-based map around human APOE," *Genomics,* vol. 54, no. 1, pp. 31-38, 1998.

[129] S. Ikegawa, "A short history of the genome-wide association study: where we were and where we are going," *Genomics & informatics,* vol. 10, no. 4, pp. 220, 2012.

[130] J. Emery, and S. Hayflick, "The challenge of integrating genetic medicine into primary care," *Bmj,* vol. 322, no. 7293, pp. 1027-1030, 2001.

[131] N. A. Holtzman, and T. M. Marteau, "Will Genetics Revolutionize Medicine?," *New England Journal of Medicine,* vol. 343, no. 2, pp. 141-144, 2000.

[132] X. Chen, and P. Sullivan, "Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput," *The pharmacogenomics journal,* vol. 3, no. 2, pp. 77-96, 2003.

[133] F. S. Collins, and A. E. Guttmacher, "Genetics moves into the medical mainstream," *Jama,* vol. 286, no. 18, pp. 2322-2324, 2001.

[134] N. Ghasemi, B. Bernhardt, B. Biesecker, C. Mastromarino, M. Boulton, R. Williamson, M. Boulton, C. Cummings, R. Williamson, and A. Clarke, "The new genetics and primary care: General practitioners views on their role and educational needs," *Journal of Medical Sciences,* vol. 7, no. 5, pp. 189-197, 2000.

[135] J. Emery, R. Walton, M. Murphy, J. Austoker, P. Yudkin, C. Chapman, A. Coulson, D. Glasspool, and J. Fox, "Computer support for interpreting family histories of breast and ovarian cancer in primary care: comparative study with simulated cases," *Bmj,* vol. 321, no. 7252, pp. 28-32, 2000.

[136] P. S. Harper, H. E. Hughes, and J. Raeburn, "Clinical Genetics Services into the 21st Century: Summary of a Report of the Clinical Genetics Committee of the Royal College of Physicians," *Journal of the Royal College of Physicians of London,* vol. 30, no. 4, pp. 296, 1996.

[137] J. W. Watters, and H. L. McLeod, "Cancer pharmacogenomics: current and future applications," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer,* vol. 1603, no. 2, pp. 99-111, 2003.

[138] H. McLeod, E. Krynetski, M. Relling, and W. Evans, "Genetic polymorphism of thiopurine methyltransferase and its clinical relevance for childhood acute lymphoblastic leukemia," *Leukemia,* vol. 14, no. 4, pp. 567-572, 2000.

[139] C. R. Yates, E. Y. Krynetski, T. Loennechen, M. Y. Fessing, H.-L. Tai, C.-H. Pui, M. V. Relling, and W. E. Evans, "Molecular diagnosis of thiopurine S-methyltransferase deficiency: genetic basis for azathioprine and mercaptopurine intolerance," *Annals of internal medicine,* vol. 126, no. 8, pp. 608-614, 1997.

[140] M. Etienne, J. Lagrange, O. Dassonville, R. Fleming, A. Thyss, N. Renee, M. Schneider, F. Demard, and G. Milano, "Population study of dihydropyrimidine dehydrogenase in cancer patients," *Journal of Clinical Oncology,* vol. 12, no. 11, pp. 2248-2253, 1994.

[141] Z. Lu, R. Zhang, and R. B. Diasio, "Dihydropyrimidine dehydrogenase activity in human peripheral blood mononuclear cells and liver: population characteristics, newly identified deficient patients, and clinical implication in 5-fluorouracil chemotherapy," *Cancer research,* vol. 53, no. 22, pp. 5433-5438, 1993.

[142] D. J. Park, J. Stoehlmacher, W. Zhang, D. D. Tsao-Wei, S. Groshen, and H.-J. Lenz, "A Xeroderma pigmentosum group D gene polymorphism predicts clinical outcome to platinum-based chemotherapy in patients with advanced colorectal cancer," *Cancer research,* vol. 61, no. 24, pp. 8654-8658, 2001.

[143] J. Stoehlmacher, V. Ghaderi, S. Iobal, S. Groshen, D. Tsao-Wei, D. Park, and H.-J. Lenz, "A polymorphism of the XRCC1 gene predicts for response to platinum based treatment in advanced colorectal cancer," *Anticancer research,* vol. 21, no. 4B, pp. 3075-3079, 2001.

[144] K. Zhang, P. Mack, and K. Wong, "Glutathione-related mechanisms in cellular resistance to anticancer drugs," *International journal of oncology,* vol. 12, no. 4, pp. 871-953, 1998.

[145] R. M. Weinshilboum, and S. L. Sladek, "Mercaptopurine pharmacogenetics: monogenic inheritance of erythrocyte thiopurine methyltransferase activity," *American journal of human genetics,* vol. 32, no. 5, pp. 651, 1980.

[146] R. M. Lunn, K. J. Helzlsouer, R. Parshad, D. M. Umbach, E. L. Harris, K. K. Sanford, and D. A. Bell, "XPD polymorphisms: effects on DNA repair proficiency," *carcinogenesis,* vol. 21, no. 4, pp. 551-555, 2000.

[147] M. R. Spitz, X. Wu, Y. Wang, L.-E. Wang, S. Shete, C. I. Amos, Z. Guo, L. Lei, H. Mohrenweiser, and Q. Wei, "Modulation of nucleotide excision repair capacity by XPD polymorphisms in lung cancer patients," *Cancer research,* vol. 61, no. 4, pp. 1354-1357, 2001.

[148] J. Stoehlmacher, D. J. Park, W. Zhang, S. Groshen, D. D. Tsao-Wei, M. C. Yu, and H.-J. Lenz, "Association between glutathione S-transferase P1, T1, and M1 genetic polymorphism and survival of patients with metastatic colorectal cancer," *Journal of the National Cancer Institute,* vol. 94, no. 12, pp. 936-942, 2002.

[149] G. Cooper, C. Cai, and X. Lu, "Tumor-specific causal inference (tci): A bayesian method for identifying causative genome alterations within individual tumors," *bioRxiv*, pp. 225631, 2017.

[150] C. Massard, S. Michiels, C. Ferté, M.-C. Le Deley, L. Lacroix, A. Hollebecque, L. Verlingue, E. Ileana, S. Rosellini, and S. Ammari, "High-throughput genomics and clinical outcome in hard-to-treat advanced cancers: results of the MOSCATO 01 trial," *Cancer discovery,* vol. 7, no. 6, pp. 586-595, 2017.

[151] V. Prasad, "Perspective: The precision-oncology illusion," *Nature,* vol. 537, no. 7619, pp. S63-S63, 2016.

[152] E. Piñeiro-Yáñez, M. Reboiro-Jato, G. Gómez-López, J. Perales-Patón, K. Troulé, J. M. Rodríguez, H. Tejero, T. Shimamura, P. P. López-Casas, and J. Carretero, "PanDrugs: a novel method to prioritize anticancer drug treatments according to individual genomic data," *Genome medicine,* vol. 10, no. 1, pp. 1-11, 2018.

[153] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An expanded view of complex traits: from polygenic to omnigenic," *Cell,* vol. 169, no. 7, pp. 1177-1186, 2017.

[154] A. D. Roses, "Pharmacogenetics and the practice of medicine," *Nature,* vol. 405, no. 6788, pp. 857-865, 2000.

[155] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, and A. Chakravarti, "Finding the missing heritability of complex diseases," *Nature,* vol. 461, no. 7265, pp. 747-753, 2009.

[156] S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'dushlaine, K. Chambert, S. E. Bergen, and A. Kähler, "A polygenic burden of rare disruptive mutations in schizophrenia," *Nature,* vol. 506, no. 7487, pp. 185-190, 2014.

[157] H. Shi, G. Kichaev, and B. Pasaniuc, "Contrasting the genetic architecture of 30 complex traits from summary association data," *The American Journal of Human Genetics,* vol. 99, no. 1, pp. 139-153, 2016.

[158] D. Botstein, and N. Risch, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease," *Nature genetics,* vol. 33, no. 3, pp. 228-237, 2003.

[159] Y. I. Li, B. Van De Geijn, A. Raj, D. A. Knowles, A. A. Petti, D. Golan, Y. Gilad, and J. K. Pritchard, "RNA splicing is a primary link between genetic variation and disease," *Science,* vol. 352, no. 6285, pp. 600-604, 2016.

[160] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, and L. Hindorff, "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations," *Nucleic acids research,* vol. 42, no. D1, pp. D1001-D1006, 2014.

[161] L. I. Furlong, "Human diseases through the lens of network biology," *Trends in genetics,* vol. 29, no. 3, pp. 150-159, 2013.

[162] J. Armenia, S. A. Wankowicz, D. Liu, J. Gao, R. Kundra, E. Reznik, W. K. Chatila, D. Chakravarty, G. C. Han, and I. Coleman, "The long tail of oncogenic drivers in prostate cancer," *Nature genetics,* vol. 50, no. 5, pp. 645-651, 2018.

[163] M. T. Chang, S. Asthana, S. P. Gao, B. H. Lee, J. S. Chapman, C. Kandoth, J. Gao, N. D. Socci, D. B. Solit, and A. B. Olshen, "Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity," *Nature biotechnology,* vol. 34, no. 2, pp. 155, 2016.

[164] L. Santarpia, G. Bottai, C. M. Kelly, B. Győrffy, B. Székely, and L. Pusztai, "Deciphering and targeting oncogenic mutations and pathways in breast cancer," *The oncologist,* vol. 21, no. 9, pp. 1063, 2016.

[165] N. Waddell, M. Pajic, A.-M. Patch, D. K. Chang, K. S. Kassahn, P. Bailey, A. L. Johns, D. Miller, K. Nones, and K. Quek, "Whole genomes redefine the mutational landscape of pancreatic cancer," *Nature,* vol. 518, no. 7540, pp. 495-501, 2015.

[166] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, and A. Patel, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer cell,* vol. 1, no. 2, pp. 133-143, 2002.

[167] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, and X. Yu, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature,* vol. 403, no. 6769, pp. 503-511, 2000.

[168] A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, and J. M. Giltnane, "The use of

molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma," *New England Journal of Medicine,* vol. 346, no. 25, pp. 1937-1947, 2002.

[169] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, and S. S. Jeffrey, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proceedings of the National Academy of Sciences,* vol. 98, no. 19, pp. 10869-10874, 2001.

[170] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Van Der Kooy, M. J. Marton, and A. T. Witteveen, "Gene expression profiling predicts clinical outcome of breast cancer," *nature,* vol. 415, no. 6871, pp. 530-536, 2002.

[171] J. M. Bartlett, J. Thomas, D. T. Ross, R. S. Seitz, B. Z. Ring, R. A. Beck, H. C. Pedersen, A. Munro, I. H. Kunkler, and F. M. Campbell, "Mammostrat® as a tool to stratify breast cancer patients at risk of recurrence during endocrine therapy," *Breast cancer research,* vol. 12, no. 4, pp. R47, 2010.

[172] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, and T. Park, "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," *New England Journal of Medicine,* vol. 351, no. 27, pp. 2817-2826, 2004.

[173] B. Wallden, J. Storhoff, T. Nielsen, N. Dowidar, C. Schaper, S. Ferree, S. Liu, S. Leung, G. Geiss, and J. Snider, "Development and verification of the PAM50-based Prosigna breast cancer gene signature assay," *BMC medical genomics,* vol. 8, no. 1, pp. 54, 2015.

[174] B. Zupan, J. DemšAr, M. W. Kattan, J. R. Beck, and I. Bratko, "Machine learning for survival analysis: a case study on recurrence of prostate cancer," *Artificial intelligence in medicine,* vol. 20, no. 1, pp. 59-75, 2000.

[175] Y.-J. Mangasarian, and W. Wolberg, "Breast cancer survival and chemotherapy: a support vector machine analysis," *Discret Math Probl with Med Appl DIMACS Work Discret Math Probl with Med Appl December 8–10, 1999, Volume 55*, p. 1, 2000.

[176] E. D. Crawford, J. T. Batuello, P. Snow, E. J. Gamito, D. G. McLeod, A. W. Partin, N. Stone, J. Montie, R. Stock, and J. Lynch, "The use of artificial intelligence technology to predict lymph node spread in men with clinically localized prostate carcinoma," *Cancer: Interdisciplinary International Journal of the American Cancer Society,* vol. 88, no. 9, pp. 2105-2109, 2000.

[177] G. P. Drago, E. Setti, L. Licitra, and D. Liberati, "Forecasting the performance status of head and neck cancer patient treatment by an interval arithmetic pruned perceptron," *IEEE transactions on biomedical engineering,* vol. 49, no. 8, pp. 782-787, 2002.

[178] J. W. Catto, D. A. Linkens, M. F. Abbod, M. Chen, J. L. Burton, K. M. Feeley, and F. C. Hamdy, "Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro-fuzzy modeling and artificial neural networks," *Clinical Cancer Research,* vol. 9, no. 11, pp. 4172-4177, 2003.

[179] A. Michael, G. Ball, N. Quatan, F. Wushishi, N. Russell, J. Whelan, P. Chakraborty, D. Leader, M. Whelan, and H. Pandha, "Delayed disease progression after allogeneic cell vaccination in hormone-resistant prostate cancer and correlation with immunologic variables," *Clinical cancer research,* vol. 11, no. 12, pp. 4469-4478, 2005.

[180] T. Ochi, K. Murase, T. Fujii, M. Kawamura, and J. Ikezoe, "Survival prediction using artificial neural networks in patients with uterine cervical cancer treated by radiation therapy alone," *International journal of clinical oncology,* vol. 7, no. 5, pp. 0294-0300, 2002.

[181] T.-K. Man, M. Chintagumpala, J. Visvanathan, J. Shen, L. Perlaky, J. Hicks, M. Johnson, N. Davino, J. Murray, and L. Helman, "Expression profiles of osteosarcoma that can predict response to chemotherapy," *Cancer research,* vol. 65, no. 18, pp. 8142-8150, 2005.

[182] H. Rodriguez-Luna, H. E. Vargas, T. Byrne, and J. Rakela, "Artificial neural network and tissue genotyping of hepatocellular carcinoma in liver-transplant recipients: prediction of recurrence," *Transplantation,* vol. 79, no. 12, pp. 1737-1740, 2005.

[183] J. S. Wei, B. T. Greer, F. Westermann, S. M. Steinberg, C.-G. Son, Q.-R. Chen, C. C. Whiteford, S. Bilke, A. L. Krasnoselsky, and N. Cenacchi, "Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma," *Cancer research,* vol. 64, no. 19, pp. 6883-6891, 2004.

[184] F. Sato, Y. Shimada, F. M. Selaru, D. Shibata, M. Maeda, G. Watanabe, Y. Mori, S. A. Stass, M. Imamura, and S. J. Meltzer, "Prediction of survival in patients with esophageal carcinoma using artificial neural networks," *Cancer: Interdisciplinary International Journal of the American Cancer Society,* vol. 103, no. 8, pp. 1596-1605, 2005.

[185] A. Daemen, O. L. Griffith, L. M. Heiser, N. J. Wang, O. M. Enache, Z. Sanborn, F. Pepin, S. Durinck, J. E. Korkola, and M. Griffith, "Modeling precision treatment of breast cancer," *Genome biology,* vol. 14, no. 10, pp. R110, 2013.

[186] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, P. Hintsanen, S. A. Khan, and J.-P. Mpindi, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature biotechnology,* vol. 32, no. 12, pp. 1202, 2014.

[187] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines," *Genome biology,* vol. 15, no. 3, pp. R47, 2014.

[188] I. S. Jang, E. C. Neto, J. Guinney, S. H. Friend, and A. A. Margolin, "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data," *Biocomputing 2014*, pp. 63-74: World Scientific, 2014.

[189] R. Rahman, K. Matlock, S. Ghosh, and R. Pal, "Heterogeneity aware random forest for drug sensitivity prediction," *Scientific reports,* vol. 7, no. 1, pp. 1-11, 2017.

[190] M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester, and J. Saez-Rodriguez, "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties," *PLoS one,* vol. 8, no. 4, 2013.

[191] S. Gupta, K. Chaudhary, R. Kumar, A. Gautam, J. S. Nanda, S. K. Dhanda, S. K. Brahmachari, and G. P. Raghava, "Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine," *Scientific reports,* vol. 6, pp. 23857, 2016.

[192] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, and M. M. Hoffman, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface,* vol. 15, no. 141, pp. 20170387, 2018.

[193] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics,* vol. 19, no. 6, pp. 1236-1246, 2018.

[194] M. Q. Ding, L. Chen, G. F. Cooper, J. D. Young, and X. Lu, "Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics," *Molecular Cancer Research,* vol. 16, no. 2, pp. 269-278, 2018.

[195] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics,* vol. 4, no. 2, pp. 249-264, 2003.

[196] B. Hellwig, J. G. Hengstler, M. Schmidt, M. C. Gehrmann, W. Schormann, and J. Rahnenführer, "Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes," *BMC bioinformatics,* vol. 11, no. 1, pp. 276, 2010.

[197] J. A. Hartigan, and P. M. Hartigan, "The dip test of unimodality," *The annals of Statistics,* vol. 13, no. 1, pp. 70-84, 1985.

[198] R. Tibshirani, and T. Hastie, "Outlier sums for differential gene expression analysis," *Biostatistics,* vol. 8, no. 1, pp. 2-8, 2007.

[199] T. Benaglia, D. Chauveau, D. Hunter, and D. Young, "mixtools: An R package for analyzing finite mixture models," 2009.

[200] C. D. Greenman, G. Bignell, A. Butler, S. Edkins, J. Hinton, D. Beare, S. Swamy, T. Santarius, L. Chen, and S. Widaa, "PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data," *Biostatistics,* vol. 11, no. 1, pp. 164-175, 2010.

[201] G. E. Hinton, and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science,* vol. 313, no. 5786, pp. 504-507, 2006.

[202] Y. He, Q. Zhu, M. Chen, Q. Huang, W. Wang, Q. Li, Y. Huang, and W. Di, “The changing 50% inhibitory concentration (IC50) of cisplatin: a pilot study on the artifacts of the MTT assay and the precise measurement of density-dependent chemoresistance in ovarian cancer,” *Oncotarget,* vol. 7, no. 43, pp. 70803, 2016.

[203] M. Gülden, and H. Seibert, “In vitro–in vivo extrapolation: estimation of human serum concentrations of chemicals equivalent to cytotoxic concentrations in vitro,” *Toxicology,* vol. 189, no. 3, pp. 211-222, 2003.

[204] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” *Journal of statistical software,* vol. 33, no. 1, pp. 1, 2010.

[205] Y. Yang, X. Dong, B. Xie, N. Ding, J. Chen, Y. Li, Q. Zhang, H. Qu, and X. Fang, “Databases and web tools for cancer genomics study,” *Genomics, proteomics & bioinformatics,* vol. 13, no. 1, pp. 46-50, 2015.

[206] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard, "Tensorflow: A system for large-scale machine learning." pp. 265-283.

[207] X. Glorot, and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." pp. 249-256.

[208] J. Bergstra, and Y. Bengio, “Random search for hyper-parameter optimization,” *Journal of machine learning research,* vol. 13, no. Feb, pp. 281-305, 2012.

[209] J. Chen, and Z. Chen, “Extended BIC for small-n-large-P sparse GLM,” *Statistica Sinica*, pp. 555-574, 2012.

[210] X. Gao, and P. X.-K. Song, “Composite likelihood Bayesian information criteria for model selection in high-dimensional data,” *Journal of the American Statistical Association,* vol. 105, no. 492, pp. 1531-1540, 2010.

[211] D. L. Donoho, “For most large underdetermined systems of linear equations the minimal ?1-norm solution is also the sparsest solution,” *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences,* vol. 59, no. 6, pp. 797-829, 2006.

[212] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research,* vol. 15, no. 1, pp. 1929-1958, 2014.

[213] S. Ioffe, and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, 2015.

[214] C. M. Bishop, “Machine learning and pattern recognition,” *Information science and statistics. Springer, Heidelberg*, 2006.

[215] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*: MIT press, 2016.

[216] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P.-Y. Glorennec, H. Hjalmarsson, and A. Juditsky, *Nonlinear black-box modeling in system identification: a unified overview*: Linköping University, 1995.

[217] E. A. Smirnov, D. M. Timoshenko, and S. N. Andrianov, "Comparison of regularization methods for imagenet classification with deep convolutional neural networks," *Aasri Procedia,* vol. 6, pp. 89-94, 2014.

[218] T. Sato, D. E. Stange, M. Ferrante, R. G. Vries, J. H. Van Es, S. Van Den Brink, W. J. Van Houdt, A. Pronk, J. Van Gorp, and P. D. Siersema, "Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium," *Gastroenterology,* vol. 141, no. 5, pp. 1762-1772, 2011.

[219] K. Uziela, and A. Honkela, "Probe region expression estimation for RNA-seq data for improved microarray comparability," *PloS one,* vol. 10, no. 5, 2015.

[220] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts," *Genome biology,* vol. 15, no. 2, pp. R29, 2014.

[221] J. A. Thompson, J. Tan, and C. S. Greene, "Cross-platform normalization of microarray and RNA-seq data for machine learning applications," *PeerJ,* vol. 4, pp. e1621, 2016.

[222] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics,* vol. 19, no. 2, pp. 185-193, 2003.

[223] J. M. Franks, G. Cai, and M. L. Whitfield, "Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data," *Bioinformatics,* vol. 34, no. 11, pp. 1868-1874, 2018.

[224] H. Liu, J. Lafferty, and L. Wasserman, "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs," *Journal of Machine Learning Research,* vol. 10, no. Oct, pp. 2295-2328, 2009.

[225] M. Goldman, B. Craft, A. Brooks, J. Zhu, and D. Haussler, "The UCSC Xena Platform for cancer genomics data visualization and interpretation," *BioRxiv*, pp. 326470, 2018.

[226] L. Chen, C. Cai, V. Chen, and X. Lu, "Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model." p. S9.

[227] M. Del Rio, C. Mollevi, F. Bibeau, N. Vie, J. Selves, J.-F. Emile, P. Roger, C. Gongora, J. Robert, and N. Tubiana-Mathieu, "Molecular subtypes of metastatic colorectal cancer are associated with patient response to irinotecan-based therapies," *European Journal of Cancer,* vol. 76, pp. 68-75, 2017.

[228] M. A. Jensen, V. Ferretti, R. L. Grossman, and L. M. Staudt, "The NCI Genomic Data Commons as an engine for precision medicine," *Blood, The Journal of the American Society of Hematology,* vol. 130, no. 4, pp. 453-459, 2017.

[229] E. Amir, B. Seruga, R. Kwong, I. F. Tannock, and A. Ocaña, "Poor correlation between progression-free and overall survival in modern clinical trials: are composite endpoints the answer?," *European Journal of Cancer,* vol. 48, no. 3, pp. 385-388, 2012.

[230] A. Tan, R. Porcher, P. Crequit, P. Ravaud, and A. Dechartres, "Differences in treatment effect size between overall survival and progression-free survival in immunotherapy trials: a meta-epidemiologic study of trials with results posted at ClinicalTrials. gov," *Journal of Clinical Oncology,* vol. 35, no. 15, pp. 1686-1694, 2017.

[231] J. Edeline, E. Boucher, Y. Rolland, E. Vauléon, M. Pracht, C. Perrin, C. Le Roux, and J. L. Raoul, "Comparison of tumor response by Response Evaluation Criteria in Solid Tumors (RECIST) and modified RECIST in patients treated with sorafenib for hepatocellular carcinoma," *Cancer,* vol. 118, no. 1, pp. 147-156, 2012.

[232] B. Mohelnikova-Duchonova, B. Melichar, and P. Soucek, "FOLFOX/FOLFIRI pharmacogenetics: the call for a personalized approach in colorectal cancer therapy," *World journal of gastroenterology: WJG,* vol. 20, no. 30, pp. 10316, 2014.

[233] T. Oguri, Y. Bessho, H. Achiwa, H. Ozasa, K. Maeno, H. Maeda, S. Sato, and R. Ueda, "MRP8/ABCC11 directly confers resistance to 5-fluorouracil," *Molecular cancer therapeutics,* vol. 6, no. 1, pp. 122-127, 2007.

[234] I. Hlavata, B. Mohelnikova-Duchonova, R. Vaclavikova, V. Liska, P. Pitule, P. Novak, J. Bruha, O. Vycital, L. Holubec, and V. Treska, "The role of ABC transporters in progression and clinical outcome of colorectal cancer," *Mutagenesis,* vol. 27, no. 2, pp. 187-196, 2012.