Latent-variable models for drug response prediction and genetic testing

by

Ladislav Rampášek

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Latent-variable models for drug response prediction and genetic testing

Ladislav Rampášek
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2020

High-throughput DNA sequencing and related biotechnologies revolutionized our understanding of human genomics and diseases with genetic component, particularly of cancer – one of the leading causes of death world-wide. Despite the progress in cancer research and availability of over 150 FDA-approved anti-cancer drugs [193], the cancer treatment is often unsuccessful. Identifying the best cancer treatment using computational models to personalize drug response prediction holds great promise to improve patient's chances of successful recovery. Unfortunately, the computational task of predicting drug response remains very challenging.

In this thesis I develop a deep latent-variable machine learning model with amortized variational inference that improves accuracy of drug response prediction over the currently used models. Besides increased expressiveness of this model thanks to parameterization by neural networks, the achieved improvement stems from integration of drug-induced perturbation profiles, a resource not fully utilized before.

Clinical trial datasets of cancer treatments which also include genomic characterization of the tumours are small and scarce, therefore for the vast majority of drugs only responses in pre-clinical biological models are available. To this end, I assess applicability of popular domain adaptation approach, based on domain-invariant representation learning, to the drug response prediction task. I conclude that necessary conditions of successful domain adaptation are often not satisfied in the available datasets and as such many current methods are misguided.

Last but not least, in this thesis I also contribute to the area of non-invasive prenatal testing. Using a hidden Markov model I propose a method for analysis of cell-free DNA fragments isolated from maternal plasma that also contain admixture of DNA fragments derived from the fetal genome. Here presented method is a first proof-of-concept for non-invasive sub-chromosomal CNV detection.

To those who suffer from cancer or genetic disorders,
your hardship gave me the strength to persevere.

Acknowledgements

I sincerely thank to the many who have helped me complete my work:

- Anna Goldenberg and Mike Brudno for their supervision, guidance and provided opportunities.
- Friends and collaborators for their help and support, especially Yulia Rubanova, Aryan Arbabi, Orion Buske, Kingsley Chang, Aziz Mezlini, Lauren Erdman, Alex Adam, Petr Smirnov, Zhaleh Safikhani, and Benjamin Haibe-Kains.
- My supervisory committee members, Alan Moses and David Duvenaud, for their guidance, constructive feedback and support.

Contents

List of Tables

List of Figures

Glossary

AUPR area under precision-recall curve.

AUROC area under receiver operation characteristic curve.

biomarker (of drug response) a measurable indicator (in any genomic modality) that can be used to predict whether a drug-treatment will be effective in a specific patient.

CCL cancer cell line; derived from a human cancer cells, it is a pre-clinical biological model of cancer; Section 2.4.1.

cfDNA cell-free DNA; short DNA fragments present in blood plasma, typically a result of cell apoptosis.

CMap-L1000v1 Connectivity Map L1000 v1; NIH LINCS dataset of reagent-induced transcriptional perturbations in a cell line dataset [192].

CNV copy-number variation; duplication or deletion of a genomic region.

CTRPv2 Cancer Therapeutics Response Portal v2; a dataset of drug sensitivity in CCLs [170].

Dr.VAE drug response variational autoencoder [166]; Chapter 3.

ELBO evidence lower bound; lower bound of data likelihood $P(X)$ in a probabilistic model, we refer to it in context of training VAE-based models [115, 174]; Section 2.5.2.

HMM hidden Markov model; Section 2.5.1.

MMD maximum mean discrepancy; Section 2.6.3.

organoid an *in vitro* pre-clinical model system that shows a realistic micro-anatomy in three dimensions; provides more realistic model of cancer biology than cancer cell lines while still allows for high-throughput screens.

PDX patient-derived tumor xenograft; an immunodeficient mouse with implanted human cancer tumor tissue, provides an accurate pre-clinical model of human cancer; Section 2.4.1.

S2S learning mode in which a model is trained and evaluated on a source domain dataset.

S2T unsupervised domain adaptation learning mode; a model is trained on a source domain dataset and evaluated on a target domain dataset.

SNP single nucleotide polymorphism; a position in DNA where the present nucleotide commonly varies within a population.

SSVAE semi-supervised variational autoencoder [117]; Section 2.5.3.

ST2T semi-supervised domain adaptation learning mode; a model is trained on data from both source and target domain but evaluated on a target domain.

T2T learning mode in which a model is trained and evaluated on a target domain dataset.

TCGA The Cancer Genome Atlas, a compendium of human cancers.

VAE variational autoencoder [115, 174]; Section 2.5.2.

VFAE variational fair autoencoder [139]; Section 2.6.3.

Chapter 1

Introduction

1.1 Motivation

Advent of affordable high-throughput DNA sequencing and related biotechnologies has opened a window into human genomics that has vastly expanded our understanding of cancers and congenital disorders. These data have been extensively used to discover and study biological processes and to identify causes and mechanisms of diseases with genomic component. Deeper understanding of disease mechanisms and stratification to common subtypes has led to development of new targeted treatments, therapies and diagnostic tests. But this effort is far from over. With decreasing cost and increasing availability there is need for computational methods to move the frontier of utility of the data provided by these biotechnologies in clinical applications, for early diagnoses and successful treatments in addition to the primary research.

Particularly in oncology, stratification of patients to common subtypes is often too coarse. Identification of biomarkers based on which chemotherapeutic drugs could be reliably prescribed with close to perfect success rate remains an open problem not only for many existing cytotoxic drugs, but also for targeted therapies as the drug targets alone are generally poor therapeutic indicators [42, 46]. Precision medicine decisions based on genomic makeup of patient's cancer has tremendous potential to improve treatment outcomes [69, 48, 96, 46] by tailoring a treatment to the individual patient.

To facilitate improved outcomes for a wide range of patients in clinical practice, genetic testing and diagnosis need to be safe, timely, and affordable. For solid tumour cancers a biopsy of the tumor is often required which limits monitoring and early detection of the cancer development. In prenatal testing for genetic abnormalities, traditional invasive prenatal diagnostic methods, such as chorionic villus sampling or amniocentesis carry a small but non-negligible risk of miscarriage [49]. The discovery of fetal cell-free DNA in maternal plasma [135] enabled development of non-invasive prenatal testing [17] and later also led to development of liquid biopsy of cancers [155]. In both cases, only a blood sample of the patient is required to sequence the cell-free DNA present in blood plasma, no other invasive procedure is necessary. However, a new set of computational methods is needed to make most of this new data.

Current high-throughput sequencing technologies can produce large amounts of data for a spectrum of applications. But given the complexity of underlying biological processes and target applications, exact algorithms and statistical tests are often not enough to fully utilize them. Developments in the area of machine learning can provide tools and methodologies well fit for this challenge.

1.2 Goals and achievements

In this thesis I develop machine learning methods that contribute to clinical translation of the advancements in biotechnology and primary research it has enabled. The impact of the work presented here is in developing advanced machine learning algorithms for two main application areas: i) precision medicine in cancer, and ii) early detection of genetic aberrations.

- I pioneered use of the variational autoencoder framework [115, 174] for gene expression representation learning, modeling of drug-induced perturbation effects [167], and semi-supervised drug response prediction [166]. I developed, Dr.VAE, the first method that successfully improved drug response prediction by generative modelling of drug perturbation profiles. This deep latent-variable model outperformed currently most used drug response prediction methods for majority of the 26 tested FDA-approved drugs.

- I analyzed and assessed necessary conditions for deployability of domain adaptation methods to improve clinical drug response prediction for cancer patients from pre-clinical studies. My findings are that currently considered methods in this application area, based on domain-invariant representations, are misguided. Instead, the field should focus on semi-supervised transfer learning approaches, efficient learning from small clinical datasets and mindful incorporation of biological priors.

- In the area of non-invasive prenatal testing, I developed the first method for detection of sub-chromosomal copy number variations in a fetal genome by probabilistic analysis of cell-free DNA fragments isolated from maternal blood plasma [165]. Experiments based on *in silico* introduction of novel CNVs into plasma samples, with 13% fetal DNA concentration, demonstrated a sensitivity of 90% for CNVs >400 kilobases with only 13 calls in unaffected genome. This work was also featured in GenomeWeb [198].

1.3 Organization

This thesis is organized into six chapters as follows:

- Chapter 1 briefly introduces motivation and overview of the thesis.
- Chapter 2 provides background and relevant literature review in the application domains and of the employed machine learning methodologies. Particularly, introduced is pharmacogenomics, drug response prediction in cancers, selected latent-variable generative models and domain adaptation methods.
- Chapter 3 presents Drug Response Variational Autoencoder [166], a novel probabilistic model with amortized variational inference for drug response prediction in cancer cell lines that is the first method to jointly model drug-induced transcriptome perturbations [167] together with efficacy of the treatment in killing of the cancer cells in order to improving the latter. Shown in ablation studies, these aspects of Dr.VAE contribute to improved classification performance over currently most used drug response prediction methods.
- Chapter 4 presents assessment of domain adaptation approach for learning clinically applicable drug response prediction classifier from pre-clinical, particularly cancer cell line, drug sensitivity datasets. First, in a series of simulation experiments, the importance of necessary conditions for successful domain adaptation through learning of domain-invariant data representation is shown. Next, in a pilot study and experiments with a clinical response data set, practical validity of these assumptions are assessed.
- Chapter 5 presents fCNV, a probabilistic method for detection of copy number variations (CNV) in a fetal genome from maternal plasma cell free DNA sequencing [165]. The fCNV is the first computational method for non-invasive sub-chromosomal copy number variation detection. Using a Hidden Markov Model, fCNV detects aberrant regions by combining signals from SNP variant allele frequencies and relative sequencing coverage.
- Chapter 6 summarizes the thesis and outlines future directions in the studied areas.

Chapter 2

Background and literature review

This thesis touches upon several major areas of research in machine learning, computational biology and cancer treatment. In this chapter we first provide a brief introduction to pharmacogenomics in general and provide examples also outside cancer treatment. Pharmacogenomics studies may include genome-wide analysis of multiple genomic modalities, such as mutations, copy number variations, mRNA expression, methylation, and others. We briefly describe the most common data types later in this chapter. We then turn the attention to pharmacogenomics in cancer, specifically to the drug response prediction in cancers. In the last parts of this chapter we provide background to machine learning methodologies used and built upon in this thesis.

2.1 Pharmacogenomics

Pharmacogenomics studies impact of genomic variation in an individual on their response to drug treatments [133]. Many modern drugs are targeted drugs, that target a specific cellular function or activity, and the effectiveness of the drug can be very dependent on how effectively the body handles the chemical compound. Therefore, the effect of a particular drug and dose will differ between individual patients. Currently, most of the drugs are prescribed based on normatives that are derived from simple clinical variables like gender, age, body mass, prescription of other drugs or a particular disease state. However, with progressively more and more targeted therapies the outcome can be significantly influenced by the genomic background of the patient. This is especially the case in cancers, that is the main focus of this thesis.

The field of pharmacogenomics is still in early stages, but the decreasing cost of genetic testing is making its translation to clinical practice possible [211], while the increasing complexity and targeted nature of modern drug treatments is making it a necessity. It is predicted that in the future pharmacogenomics will allow the development of drug therapies tailored to the patient, i.e. personalized medicine, for a variety of diseases, e.g. Alzheimer disease, cancer, HIV/AIDS, and more [172]. In this section we describe the major research areas in pharmacogenomics, but first we start by reviewing the necessary terminology related to drugs and their activity.

A drug undergoes several stages after the administration. Firstly, the drug needs to absorb into the body, typically through the digestion system. Then it is distributed around the body and needs to reach the site of its action. There, the intended biological target interaction takes place, e.g. binding to an enzyme, ion channels, or various type of receptors. The drug is then metabolically processed and excreted from the body. Theoretically, genes that influence any stage of this process could affect the overall drug response.

There are two main groups of genes that are important when studying variation in drug responses. The first are those influencing the *pharmacokinetic* properties of drugs, such as drug metabolizing enzymes and drug transporters, that affect how the drug is handled by the body. The second are those influencing *pharmacodynamic* properties of drugs, including drug targets such as enzymes, receptors and ion channels, and their associated pathways, which determine the drug's effect on the body.

In clinical application, two aspects of a drug are most important: the *efficacy* and *toxicity* of the drug. Efficacy refers to the best therapeutic response that a drug can produce, i.e. it is a measure of clinical effectiveness. Efficacy can be expressed in terms of the percentage of patients who show positive response given a standard dose. Toxicity then refers to the extent to which a drug causes unwanted or harmful effects, and may be expressed as the percentage of patients who show adverse side effects when administered a standard dose. The optimal dose range for a drug is that which maximizes efficacy while minimizing toxicity. The optimal dose can vary between individuals.

2.1.1 Research areas in pharmacogenomics

Drug response prediction A drug that is effective in one person may have little to no therapeutic effect in another. Some patients may show partial response, while some may develop undesirable side effects. Thus predicting effectiveness of a particular drug treatment in a specific patient or dose level at which the treatment is effective but save for the patient is crucial.

One of the most common pharmacogenetic tests today is *HLA-B*57:01* genotype testing before Abacavir therapy for HIV/AIDS [143], which may cause a potentially lethal hypersensitivity reaction because of the specific rare mutation.

Another example is that of Warfarin, an anticoagulant normally used in thrombosis prevention, whose biological effects are complex and involve many genes [108]. There is a clear association shown between the dose requirement of Warfarin and variants of particular metabolic and pharmacodynamic enzymes. Thus screening for mutations in genes coding for these enzymes is important for appropriate dose prescription.

Biomarker identification Identification of biomarkers, such as mutations or particular gene expression levels, that are predictive of drug efficacy is important for designing affordable clinical tests for clinical outcome or optimal dose prediction. Additionally, studying the genomic influences on a drug action can also lead to better understanding of the disease mechanisms.

Drug repurposing Also referred to as drug repositioning, is an application of an already approved drug in treatment of another condition [95]. Probably the most widely known example is that of Pfizer's Viagra, originally developed for pulmonary arterial hypertension and then approved for erectile dysfunction treatment.

Drug combination/synergy It has been showcased that combination of different drugs can have synergistic effect resulting in increased efficacy of the treatment, while reducing the toxicity thanks to reduced overall drug dosage and thus reduction of medication side effects. Development of methods for drug synergy prediction is an increasingly active research area [62].

Drug design Advances in genomics and pharmacogenomics, among others, have propelled a new paradigm in drug discovery. The old approach was a brute-force approach based on screening of large library of chemical compounds for activity with target cells. The initial pool of around a million of compounds would then be progressively filtered for desired activity, toxicity, production feasibility, etc., and after around a decade, the clinical trials would start with a handful of candidate compounds without precise knowledge about their cause of action. The contemporary approach is to design a drug for particular action, based on genomic studies of the disease. Pharmacogenomic studies are then a part of such targeted drug development, replacing the previously dominant trial and error approach [176].

2.2 Overview of human genomics

Here we briefly summarize background of human genomics [203] that is most relevant for application areas studied in this thesis.

Firstly, the genetic information is encoded in a deoxyribonucleic acid (DNA) molecule as a sequence of nucleotides with either cytosine (C), guanine (G), adenine (A), or thymine (T) nucleobase. Nucleotides A with T, and C with G form hydrogen bonds, thanks to which the structure of the DNA molecule is a complementary double helix. A packaged and organized DNA molecule is called a *chromosome*.

The human genome is composed of 22 pairs of autosomal chromosomes (one set inherited from the mother and the other from the father) and two sex chromosomes, XX in females and XY in males. This is the nuclear DNA, as it is stored in the nucleus of a cell. Additionally, we recognize also mitochondrial DNA (sometimes referred to as the 25^{th} chromosome) that is located in mitochondria which are always inherited from maternal side.

The most well studied, and perhaps the most important, regions of chromosomes are the *genes*. It is currently estimated [180] that there are around 43,000 genes in the human genome, approximately 21,000 of which encode the primary structure of all proteins a cell can produce. Those parts of genes that can be translated into amino acids (proteins) are so-called coding regions, or all together called the *exome*. The exome forms only a fraction of the whole genome,

less than 2%. The rest of the DNA has mostly regulatory and supportive function, subject to ongoing research.

Now we give an overview of the *central dogma of molecular biology* [35] that in short stands as "DNA makes RNA and RNA makes protein". That is, genes from the DNA are first *transcribed* to precursor messenger RNA (pre-mRNA). This primary transcript is then processed, most notably, the non-coding regions that do not encode for amino acids (introns) are spliced out. The coding parts, called the exons, are then joint together to form messenger RNA (mRNA). Alternative splicing of mRNA can occur when appropriate, which significantly increases the diversity of proteins a single mRNA can produce. Notably, the abundance of alternative splicing is what sets humans apart from species with similar number of genes, as thanks to alternative splicing humans can produce many times more proteins. The mRNA is then *translated* to a protein. This is done in a sequential fashion. Consecutive triplets of nucleotides in mRNA, denoted as codons, code for one of 20 possible amino acids. As the mRNA is read, a protein composed of a sequence of amino acids defined by the codons is progressively created. The proteins fold into complex 3D structures and interact with each other to create protein complexes. It is estimated that a human body can produce up to a million of unique protein complexes with a wide variety of functions. Proteins are building blocks of cells and also most of the regulation and activity in a cell is carried out by proteins.

2.2.1 Point mutations

Point mutations, or *single nucleotide variations* (SNVs), are substitution changes in the DNA such that a single nucleotide of a genome is altered compared to the reference. There are more than 600 million SNVs in the human genome [112]. Those SNVs that are known to commonly vary in a population (more than 1% of the population differs from the reference) are denoted as *single nucleotide polymorphisms* (SNPs) [184]. Most of them are harmless and on average a person has around 4 to 5 million SNPs in their genome that differ from the reference, however this number is population specific and some populations (e.g. African) differ from the reference more. Further, novel mutations do occur as well and slowly accumulate during one's life span. Depending on whether these mutations happen in germ cells, which means they are heritable, they are denoted as germline mutations or somatic mutations if they happen in other tissues.

The harmfulness of a mutation depends on what kind of alternation it is and in what genomic context it occurs in [126, 197]. If it occurs in a protein coding region (exon) it can result into a change of the amino acid code. If the new triplet (codon) of nucleotides cannot be translated into an amino acid anymore, it is a *nonsense mutation*. If the codon changes meaning, i.e. codes for a different amino acid, it is a *missense mutation*. In case the new amino acid is chemically similar to the original one, such a mutation can also be called *neutral*. The amino acid coding is redundant, so it can happen that the meaning of a codon is not changed. In that case, such mutation is called a *silent mutation*, however it still can have an impact on the folding of the protein and thus also a functional effect. A single nucleotide deletion/insertion in a coding region

leads to *frameshift*, that changes the offset with which the protein coding mRNA is read (and interpreted into codons) during translation and thus resulting in a very different protein.

Mutations outside coding regions are less studied. Mutations in transcription factor sites can change efficiency with which a transcription factor (a protein) binds at that locus, and so influence regulatory mechanisms. Loci in genome that do influence expression levels are called expression quantitative trait loci (eQTL).

Point mutations and short indels can be measured by DNA sequencing, either whole-genome or exome-only for cost reduction. With sequencing it is possible to detect novel or rare variants. In case it is desired to genotype the sample only on a set of pre-selected positions, e.g. the known polymorphic sites, the so-called SNP microarrays are a very common and cheap option.

2.2.2 Structural variations

Structural variation is variation in how genome or a chromosome is structured [56]. Such variations are typically translocations, inversions, or copy-number variations in parts of the chromosomes. Most commonly studied are copy-number variations (CNVs) as duplication or deletion of a genomic region can influence transcription levels of the impacted set of genes both directly and indirectly. In cancer, catastrophic structural variations can occur, e.g. chromothripsis, leading to chromosomal and gene aberrations such as creation of pseudo chromosomes, gene fusions and other [53].

Copy number variations are also the cause of many congenital genetic disorders, either inherited from a parent or occurring *de novo*. Common disorders due to a medium- to large-sized CNV include: DiGeorge syndrome, caused by deletion of ~3Mb region of chromosome 22 (22q11.2); Prader-Willi syndrome, the deletion subtype of which is caused by a ~4Mb deletion in paternal chromosome 15; or Down syndrome that is caused by trisomy of chromosome 21. It is important to identify these aberrations as soon as possible in order to provide appropriate counseling and care. One of the contributions of this thesis, described in Chapter 5, is a probabilistic method fCNV for prenatal detection of copy number variations. The presented approach is a non-invasive approach in a sense, that it does not require a direct fetal DNA sample but instead makes use of fetal DNA fragments that are freely circulating in maternal plasma (cfDNA).

2.2.3 Gene expression

So far we have discussed DNA alterations, such as mutations and structural variants. However a living cell is a dynamic system with highly complex regulatory mechanisms that are far from understood. As mentioned earlier, proteins are the building blocks of a cell as well as they carry out most of the activity ongoing in a cell. Therefore it is crucial for a cell to produce the right proteins at the right time. The current technology of mass spectrometry for measuring protein levels and their compositions are expensive and do not achieve necessary throughput. Instead, as

a surrogate, the mRNA transcript levels are measured. Gene expression level is then the amount of mRNA in the cell that is derived from that particular gene.

RNA can be sequenced similarly to how DNA is sequenced. Today, RNAseq using 2nd generation high-throughput sequencing from Illumina is most common. The cheaper option are mRNA microarrays, designed to measure abundance of particular RNAs, typically the RNA that corresponds to the exome.

Gene expression levels are inherently noisy, as they depend on the tissue type, environment, external stresses or current state of the cell cycle. In cancer, the cell activity is altered to a very significant degree, and expression levels of some genes are changed manyfold. Some of these dysregulated or otherwise aberrant genes are frequently indicated in multiple cancers, the COSMIC Cancer Gene Census details 719 cancer-driving genes [189]. However there is generally a considerable amount of tissue and individual-specific bias in cancer tumour gene expression levels, e.g. because of intra-tumoural heterogeneity.

2.3 Hallmarks of cancer

In 2000 Hanahan and Weinberg [90] published a list of six biological capabilities that cells acquire during their development into cancerous cells. These six hallmarks of cancer were later extended to eight [91]. These hallmarks are fundamental principles that underlie cancer and show how very diverse cancers can be. One of the original assumptions in cancer research has been that normal cells evolve in a multistep process, progressively acquiring these hallmark capabilities. However this may not be always the case, for example in chromothripsis the chromosomes are believed to be massively rearranged during a single catastrophic event. Nevertheless these eight hallmarks summarize well the fundamental processes leading to cancerous tumours:

1. *Unlimited proliferation*; cancer cells modify growth signalling and proliferate at high rate.

2. *Evading growth suppressors*; means decreased sensitivity to inhibitory signals that regulate excessive growth.

3. *Resistance to programmed cell death* (apoptosis); cells normally possess regulatory mechanisms that in case of excessive damage to the cell trigger its self-destruction. TP53 and RB proteins are the most well-known apoptosis-inducing proteins that are often not functional in cancers.

4. *Enabling replicative immortality*; repeated cycles of cell division lead to biological ageing (senescence). Naturally, only stem cells and developing cells are immortal in this sense. In normal (non-immortalized) cells, telomeres that protect the ends of chromosomes shorten with each cycle, creating a sort of counting mechanism. However cancer cells develop a mechanism to preserve the telomere length, making them replicatively immortal.

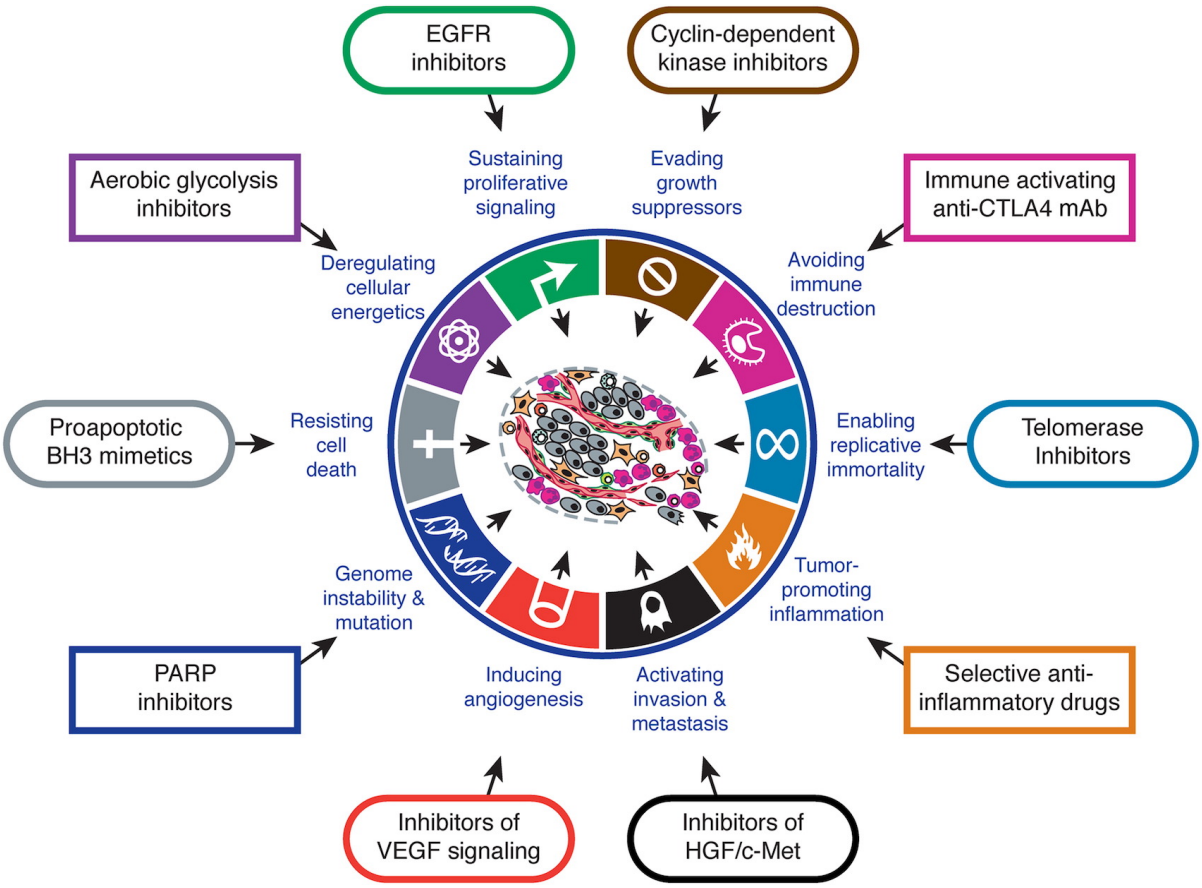


Figure 2.1: Therapeutic targeting of the hallmarks of cancer. Figure reprinted from Hanahan and Weinberg [91] with permission from Elsevier.

5. *Induced angiogenesis*; like normal tissues, tumours require nutrients, oxygen and a way to remove metabolic wastes. The tumour tissue finds a way how to induce creation of new vasculature (angiogenesis) in the tumour to address these needs.

6. *Invasiveness and metastasis*; in late stages, cancer cells invade into nearby blood and lymphatic vessels, enabling them to invade distant tissues (metastasis) forming new nodules that grow into tumours.

7. *Reprogramming of the cellular energetics*; in order to sustain the proliferation, cancer cells modify metabolism to support such activity.

8. *Avoiding immune destruction*; cancer cells adapt to avoid destruction by T and B lymphocytes, macrophages, and natural killer cells.

2.3.1 Therapeutic targeting of hallmark capabilities

The decades of cancer mechanisms research enabled introduction of mechanism-based targeted therapies to treat human cancers. Hanahan and Weinberg [91] categorized the rapidly growing

arsenal of targeted therapeutics according to their effects on one or more hallmark capabilities, Figure 2.1. The fact that the current targeted therapeutics can be aligned with the hallmark capabilities can be viewed as a confirmation of the hallmarks. Positive efficacy of a drug targeted to inhibit or impair one of the hallmarks shows that the hallmark capability is truly important for the biology of a tumour.

However, many resulting clinical responses of such targeted therapies have been only transitory, frequently followed by relapses. This is caused by redundancy present in many biological mechanisms. Therefore inhibiting only one key component is often not enough for complete disruption of a hallmark capability, as the residual cancer cells grow to adapt to such an intervention. Frequently, the adapted form of cancer that is formed under a selective pressure imposed by the therapy is even more difficult to treat. Therefore drug synergy targeting multiple targets and hallmarks is what is believe to be the future of cancer treatment. In short, it is believed that the cancerous cells need to be exterminated quickly and completely in the early stages of the cancer progression, before resistant subclones with large genetic diversity can evolve.

2.4 Computational approaches to drug response prediction in cancer

Cancer is a leading cause of death worldwide and the most important impediment to increasing life expectancy in every country of the world in the 21st century [22]. Fortunately, from 2011 to 2015, there has been a small but prominent decrease in death rates for all races/ethnicities combined for 11 out of 18 most common cancers among men and 14 of the 20 most common cancers among women. The continued decreases in death rates for colorectal cancer, prostate cancer and female breast cancer are largely due to advances in early detection and more effective treatments [36]. In this section we will focus on the computational challenges of identifying the best treatment that improves chances of successful recovery.

Until recently, treatments were chosen based on the type of cancer in a one-size-fits-all manner. We are now witnessing the advent of precision oncology [69, 48, 96] that takes into account patients' genomic makeup for treatment decisions [205, 107, 69], illustrated in Figure 2.2. Treatment approval based on tumour-site agnostic molecular aberration biomarkers has become reality. The year 2017 marked the first FDA approval of such a treatment [161]. Based on clinical trials in 15 types of cancer, pembrolizumab was approved for treatment of solid tumours with mismatch repair deficiency or high microsatellite instability [127]. Larotrectinib is another promising treatment, targeting the tropomyosin receptor kinase gene fusion in a variety of cancers [50]. Unfortunately, there are no established biomarkers for majority of the anticancer drug compounds. Identification of reliable biomarkers is a challenge not only for the most commonly used cytotoxic drugs, but also in the case of targeted therapies as the drug targets alone are generally poor therapeutic indicators [42, 46].

Discovery of biomarkers predictive of drug response and development of multivariate

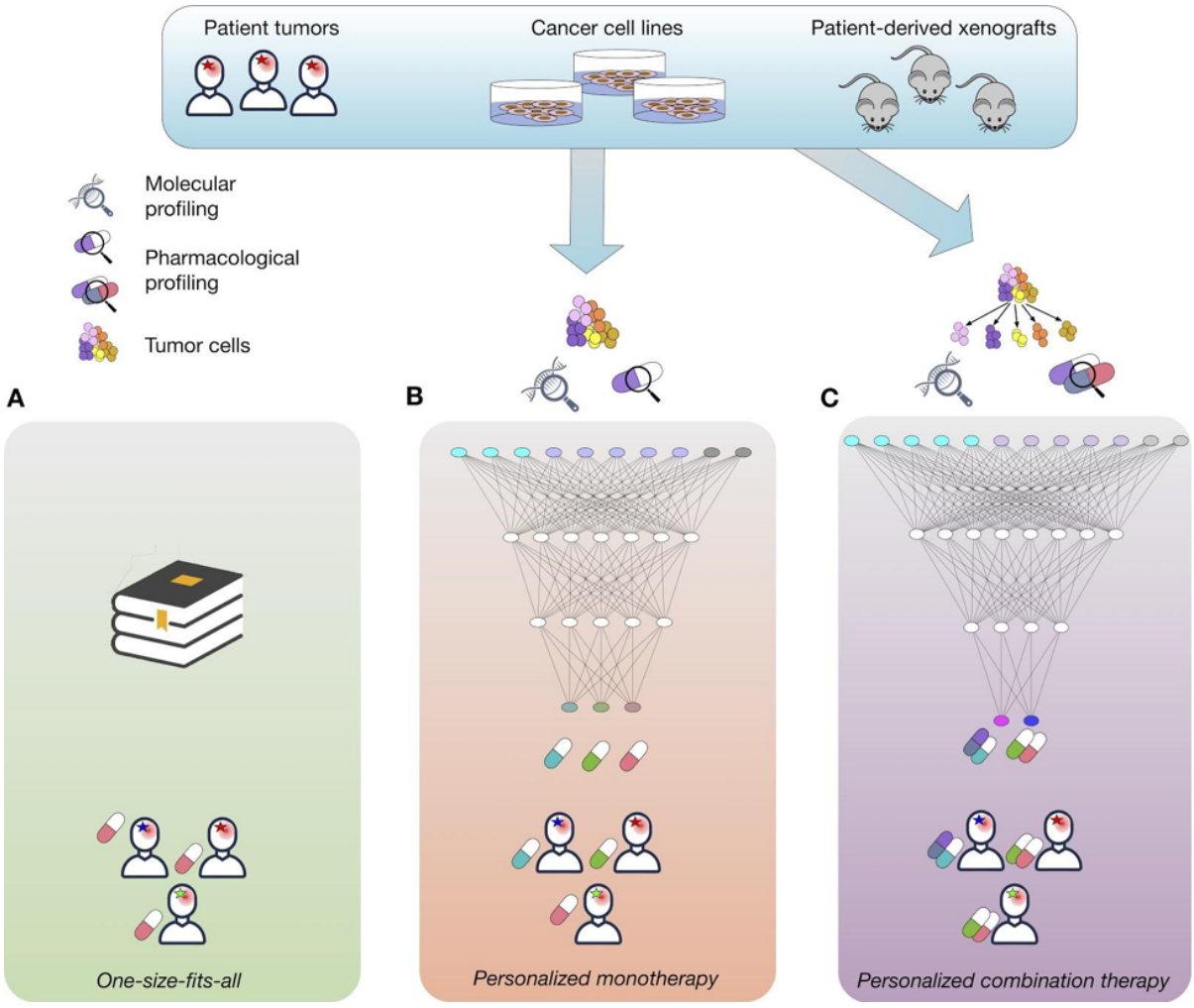


Figure 2.2: **Graphical abstract of drug response prediction.** Patient data is limited, so to predict drug response, much of the existing literature uses model system data, e.g. immortalized cell lines and PDX. (A) Currently most patients in cancer are still treated in a one-size-fits-all manner according to the type (or subtype) of cancer they have. (B) There is a growing number of examples of personalizing monotherapy in practice, where depending on the mutations in the tumour, the patient can be prescribed a targeted drug. This approach is applicable to fewer than 20% of the patients. The computational contribution is to take a large number of model systems and patients, when available and construct a predictive model to identify the best drug for majority of the patients. (C) Due to tumour heterogeneity and acquired drug resistance, monotherapies may not be effective, there is currently a growing body of work predicting drug synergy and effective drug combinations. Originally these models were trained using bulk data, but more recently, single cell data-based approaches are starting to show promise.

companion diagnostics require efficient computational tools and substantial number of samples. Traditional statistical models and more sophisticated machine learning approaches have been used to build predictors of drug response and resistance both in the clinical [160] and preclinical [43] settings. As predictive models increase in complexity, the number of observations required to train these models increases as well. While -omic profiles and clinical outcomes of patients are the most relevant data sources for the development of clinically-relevant predictors, these datasets are often limited in size due to many factors including high costs, limited accrual rates and complex regulatory landscape. In addition, by the nature of the experiment, unbiased testing of multiple therapeutic strategies for the same patient in the patient itself is practically infeasible. Cancer models provide access to patient tumours in preclinical models, both in vivo and in vitro, allowing researchers to test multiple drugs and combinations in parallel [43]. Although these preclinical models recapitulate patient therapy response to varying degrees, they provide massive amounts of pharmacogenomic data for drug response prediction. Here we review these preclinical models and the recent applications of machine learning to prediction of response.

2.4.1 Prediction of response to monotherapies

Large-scale efforts to associate molecular profiles with drug response phenotypes in preclinical models date back to the late 90s when the National Cancer Institute Developmental Therapeutics Program released large-scale pharmacogenomic data of 60 cancer cell lines (NCI60) screened with tens of thousands of chemical compounds, including a large panel of FDA-approved drugs [185]. NCI60 facilitated several drug discoveries, notably a 26S proteasome inhibitor bortezomib that is now used in multiple myeloma treatment [185]. Since then, high-throughput in vitro drug screens of cancer cell lines (CCLs) derived by immortalization of human cancer cells became popular experimental bases for discovery of multi-omic underpinnings of drug sensitivity and resistance [141]. Since this seminal study, multiple large-scale databases have been publicly released to the cancer research community [188, 134]. More recently, advances in growing tumours in animal models enabled the generation of large collection of patient-derived xenografts (PDX) to monitor tumour growth with and without drug treatment in mice [5]. Novartis published the largest PDX-based pharmacogenomic dataset to date, referred to as the PDX Encyclopedia [67]. The NCI recently announced the Patient-Derived Models Repository (PDMR) with comprehensive molecular profiling and commitment to release pharmacological profiles in the future. A series of databases and tools have been developed recently to harmonize and make easily available multiple pharmacogenomic studies investigating anticancer monotherapies, Table 2.1.

Cell line pharmacogenomic data sets

Cell lines derived from human cancers are the most common model used in pharmacogenomic studies. The original donor cancer cells were intentionally mutated to induce replicative immortality as not all cancer cells would proliferate indefinitely. The ability to grow these

Table 2.1: Platforms harmonizing preclinical pharmacogenomic datasets and providing basic processing functions for biomarker discovery.

Platforms	Cancer models	# Models	# Drugs	Reference
PharmacoGx, PharmacoDB	Cell lines	1691	759	[187, 188]
GDSCTools	Cell lines	1001	265	[32]
CellminerCDB	Cell lines	~1000	~50,000	[164]
CancerDP	Cell lines	1061	24	[88]
PDXFinder	PDX	567	33	Unpublished
Xeva	PDX	277	61	[150]
Cancer-Drug eXplorer	2D cell cultures	462	60	[128]

Table 2.2: Available pan-cancer cell line screen datasets of monotherapy drug response.

Dataset	Institution	# CCLs	# Tissues	# Drugs	Ref.
CCLE	Broad Institute	1061	24	24	[9, 74]
CTRP v2	Broad Institute	888	25	544	[170]
FIMM	Institute for Molecular Medicine Finland	50	4	52	[152]
gCSI	Genentech	754	30	16	[92]
GDSC1000	Wellcome Trust Sanger Institute and Massachusetts General Hospital Cancer Center	1109	36	250	[102]
GSK	GlaxoSmithKline	310	25	19	[84]
NCI60	National Cancer Institute	74	9	49278	[185]

cultures indefinitely is the major advantage of this model. It enables for standardized cancer cell line models that are used for decades, can be tested to response to any drug or drug combination at various drug concentration levels. It is thus possible to compare efficacy of multiple different drugs on (at least partially) common set of cell lines.

Use of cell lines has facilitated high-throughput screening of *in vitro* efficacy of many approved and experimental drugs [134]. As earlier mentioned, the first major data collection was NCI60 [185], which development started in late 1980' and continued for over two decades. Several other followed after NCI60, the current most widely used data sets are the Cancer Cell Line Encyclopedia (CCLE) [9, 74] initiated by Novartis/Broad Institute; the Genomics of Drug Sensitivity in Cancer project (GDSC1000) [102] by the Wellcome Trust Sanger Institute, previously known as Cancer Genome Project (CGP) [68]; and the Cancer Therapeutics Response Portal (CTRP v2) [170]. The most relevant pan-cancer datasets of cell line sensitivity to monotherapy treatments are summarized in Table 2.2.

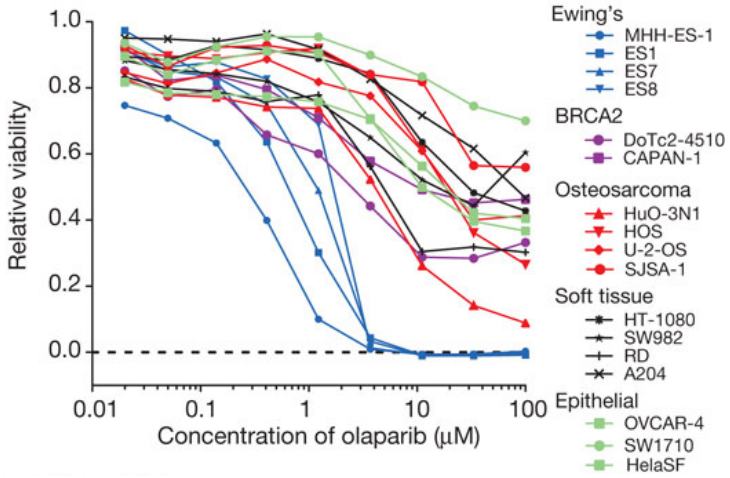


Figure 2.3: Dose-response curves of various cell lines to olaparib from the CGP data set. Figure reprinted from Garnett et al. [68] with permission of Springer Nature, ©2012.

Cell lines are typically tested for response sensitivity under varying drug concentrations. Usually up to 10 dose concentrations are tested. Then so-called dose-response curves are analysed. These dose-response curves show percentage of original cell line culture that survived under respective drug concentration. Figure 2.3 shows several dose-response curves as measured in the CGP.

There are several ways one can summarize a dose-response curve into a single real number representing the cell line's sensitivity. Originally, the half-maximal inhibitory concentration (IC_{50}) was commonly used. This is the drug concentration at which half of the original cells die. However this IC_{50} summary statistic does not work that well in practice, and other statistics are replacing it. Most commonly it is AAC statistic that is used, which is based on the area above the dose-response curve. Alternatively, statistics based on the slope of the curve are used as well.

Even though these *in vitro* cell line experiments are the simplest models, it was showed that the reproducibility of the reported results is not perfect [89], likely caused by mislabelling or contamination of the cancer cell lines. Additionally, also choice of the sensitivity measure will significantly affect result of downstream analysis [59, 104].

The major systematic issue of cell lines is that they undercome the immortalization process that significantly modifies gene expression of the cells. Additionally the cell lines accumulate spontaneous mutations over time. Further, cell line experiments are *in vitro* experiments without presence of vascular system and lack immune system response. Therefore the measured response and putative biomarkers do not necessarily translate to human patients.

Patient-derived tumour xenografts

Patient-derived tumour xenografts (PDXs) are mouse models with implanted human cancer tumour tissue. Within an immunodeficient mouse a human tumour can grow in close to original environmental conditions, like tissue structure, blood circulation, oxygen and nutrients access,

or hormone levels. Furthermore, the tumour tissue mostly maintains its genetic and epigenetic abnormalities as found in the patient. Thanks to that the PDX models exhibit drug response very similar to that of the donor patient. Recently a high throughput response screening of 62 drugs in 1075 PDXs has been published [67], providing a compelling data source.

2.4.2 Methods for monotherapy prediction

The most typical computational approaches to drug response prediction, specifically in pre-clinical models, consist of (1) quantification of drug response; (2) molecular feature selection or dimensionality reduction of the cellular measurements; (3) machine learning model fitting to predict drug response; and (4) model evaluation [7, 41]. Multiple studies explored which genomic modalities harbour the most predictive signal of drug response by analyzing performance of predictive models. The most commonly utilized modalities include single nucleotide variations, copy number variations, RNA expression, methylation, and proteomics. Despite their widespread use in clinical settings, mutations and copy number variations have been shown to account for only a small subset of candidate biomarkers, while gene expression, methylation and protein abundance are regarded as the most predictive modalities [102, 179, 104], each can be complemented by the multi-omic view of the cancer [191, 34, 149]. Perhaps the main obstacle in effectively leveraging all data modalities is the dimensionality and the correlation structure among the features. A combined set of measurements can reach hundreds of thousands of features, while the number of available patients or cell lines remains in the hundreds. This prohibitive ratio of measurements to samples limits the class of applicable predictive models requiring feature selection or learning of reduced representations. Papillon-Cavanagh et al. [156] identified univariate feature selection as a robust selection approach, later improved by mRMR Ensemble feature selection [40]. Jang et al. [104] performed extensive comparative analyses of machine learning methods for drug response prediction in cancer cell lines, recommending using elastic net or ridge regression with input features from all genomic profiling platforms. Costello et al. [34] summarized a crowdsourced DREAM drug prediction challenge, revealing two leading trends among the most successful methods. First, the importance of the ability to model nonlinear relationships between data and outcomes, and second, the incorporation of prior knowledge, e.g. biological pathways. The challenge winning model, Bayesian multitask multiple kernel learning method [34, 79], incorporated both of these approaches together with multi-drug learning [4].

The integration of prior biomedical knowledge has since been recognized as a promising approach for drug response prediction. Lee et al. [129] developed a method that integrates disease relevant multi-omic prior information to prioritize gene-drug associations. Most recently, Zhang et al. [219] and Wang et al. [208] introduced methods based on similarity network fusion and similarity-regularized matrix factorization respectively that take into account similarity among cell lines, drugs and targets. Drug chemical features and similarities were shown to be a promising additional information that can improve drug response prediction performance.

Deep learning methods for monotherapy prediction

The use of neural networks for drug response prediction dates back to the 90s. El-Deredy et al. [57] showed that a neural network trained on tumour nuclear magnetic resonance (NMR) spectra data has potential as a drug response predictor in gliomas, and may be used to provide information about the metabolic pathways involved in drug response. Neural networks, however, did not become a method of choice for monotherapy prediction yet. In fact, despite the recent prevalence of deep neural network (DNN) methods across many areas and industries, including related fields, such as computational chemistry [37, 206, 3, 75, 77, 148], DNNs have only fairly recently found their way into the drug response prediction. The reason for this is the typically low ratio of the number of samples to the number of measurements per sample that does not favour traditional feedforward neural architectures. Overparameterization in these models easily leads to overfitting and poor generalization to new datasets. However, in recent years, more public data has become available and newly developed deep neural network models are showing promise. For example, Chang et al. [26] developed the CDRscan model, featuring a convolutional neural network architecture trained on a dataset of ~1000 drug response experiments per compound. Their model achieved significantly improved performance compared to other classical machine learning approaches, Random Forests and SVM.

Table 2.3: **Computational tools for monotherapy prediction.** A non-exhaustive summary of the most recent monotherapy prediction methods with an available web service or source code. (*) A web application has been promised by the authors, but is not available yet as of July 2019.

Name	Availability	Purpose	Methodology and Features	Reference
HNMDRP	Matlab and R code	Drug response prediction in CCLs	Genomic and compound features combined with drug-target interaction and PPI	Zhang et al. [219]
KRL	Python code	Drug prioritization (ranking) in CCLs transferable to patients	Kernelized rank learning using genomic features, (predominantly gene expression)	He et al. [94]
CDRscan	*Web Application	Drug response prediction in CCLs	Deep neural network trained on somatic mutations and drug compound fingerprints	Chang et al. [26]
CancerDP	Web Application	Drug response prediction in CCLs	SVM models using (combination of) genomic features (mutations, CNVs, expression levels)	Gupta et al. [88]
BMTMKL	Matlab and R code	Drug response prediction in CCLs	Bayesian multiview (original genomic modalities + aggregated views) multitask model	Costello et al. [34]

Another promising direction is autoencoders that are able to learn from smaller datasets. A major contribution of this thesis is a pioneering work on use of variational autoencoders for drug response prediction, first published in conference workshop contributions [167, 168] then in the final journal publication [166]. We evaluated semi-supervised variational autoencoders on monotherapy response prediction and developed a joint drug response prediction model, Dr.VAE, that leveraged pre- and post-treatment gene expression in cell lines, showing improved performance in drug response prediction on a variety of FDA approved drugs compared comprehensively to many classical machine learning approaches. Dr.VAE project is the subject of Chapter 3. In parallel with our work, Way and Greene [210] explored the use of variational autoencoders for unsupervised low-dimensional gene expression representation. Dincer et al. [45] developed DeepProfile, a method that combined variational autoencoders; Section 2.5.2; to learn 8-dimensional representation of gene expression in AML patients and then used this representation to fit a Lasso linear model for drug response prediction. Similarly, Chiu et al. [28] pretrained autoencoders on mutation data and expression features on TCGA dataset and subsequently trained a deep drug response predictor. The brief summary of methods is available in Table 2.3. The trend of model development shows that as more data become available and deep learning methods become better adapted to high dimensional / low sample size data, there is hope for convergence and creation of sophisticated models that will likely push the field of computational drug response prediction forward to eventually become clinically relevant.

Predicting clinical drug response from pre-clinical experiments

Clinical trial datasets of cancer treatments which include genomic characterization of the tumours are small and scarce, therefore for the vast majority of drugs it is not possible to train a predictive model using only clinical datapoints. Many have tried to use, perhaps without explicitly realizing it, approaches that can be categorized into domain adaptation or transfer learning. They fit their models using pre-clinical datasets and then adapt them to or directly evaluate them on the limited clinical data.

The first methods in this area [71, 72, 221] are based on aligning the source and target domains and thus can be characterized as domain adaptation approaches. They involve a dataset preprocessing step to remove the difference in marginal distribution of the gene expression between cell lines and a target patient cohort by methods typically used for dataset homogenization or batch effect removal such as ComBat [110], SVA [130, 131] or removal of first few Principal Components. Using so preprocessed datasets they fit and evaluated standard prediction models such as logistic regression, support vector machines or random forests. A step further is PRECISE method by Mourragui et al. [151] that explicitly aligns principal subspaces between the domains.

Recent transfer learning approaches are neural network-based autoencoding models, that learn a low-dimensional representation of gene expression and then fit a response predictor on such embedding. Dincer et al. [45], Chiu et al. [28] refit a patient-specific predictor, while Sharifi-Noghabi et al. [183] directly apply the cell-line-trained predictor to clinical evaluation

dataset without retraining.

The Chapter 4 is dedicated to in-depth assessment of domain adaptation for drug response prediction, while the machine learning background is given in Section 2.6.

2.4.3 Limitations of monotherapies

While drug response prediction can help pick an optimal therapy given the current molecular characteristics of the cancer cells, tumours often exhibit drug resistance over the course of the treatment. Consequently, patients that respond initially to therapy regress as their cancer either adapts to overcome the chosen treatment, or an existing resistant subclone repopulates the tumour [97]. For therapies inhibiting the activity or signalling of their target, a common mechanism towards resistance is feedback selecting for upregulated expression of the target protein. For example, resistance to 5-FU has been demonstrated to arise from the amplification of its target thymidylate synthase (TS) [105], with corresponding overproduction of TS enzyme and mRNA transcripts [16]. Furthermore, especially for tyrosine kinase inhibitors, tumours will evolve to re-activate pathways downstream of the targeted protein. A classical example is the resistance to the EGFR inhibitor gefitinib which can often be explained by an acquired T790M mutation reducing drug binding affinity [121]. Other mechanisms of resistance include modifications to enzymes involved in drug metabolism to either reduce conversion of drugs to active forms or deactivate the compound [146, 99], and more recently, the intra-tumour heterogeneity [195].

Intra-tumour heterogeneity and clonality is intimately linked with the problem of resistance to treatment. The approaches reviewed above make drug response predictions using the molecular state from bulk profiling at a single point in time. However, the heterogeneity within tumours and the evolution of the cell population during treatment means that these predictions may not extrapolate to every cell within the tumour throughout the course of treatment. Intratumoural heterogeneity can act as the fuel behind clonal evolution during treatment, with the treatment acting as selective pressure selecting for a resistant subclone [213].

Drug combinations are crucial next step for addressing the issue of drug resistance and preventing recurrence caused by a negligible amount of remaining cancer cells. Synergistic combinations can also reduce toxicity by allowing for lower doses of either drug to be used. By enabling reduced doses, drug combinations can further increase the feasibility of drug repurposing by increasing the potency of compounds that are only effective at clinically irrelevant doses [194]. Thus a promising future direction in cancer treatment is the development of computational methods that account for tumour clonality and / or can predict anti-cancer drug combination synergies as compared to only monotherapy response. However this research is outside of the scope of this thesis and in what follows we specialize in monotherapy drug response prediction.

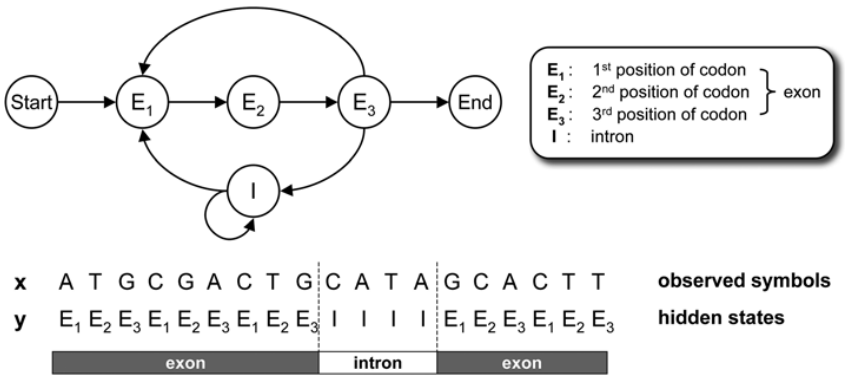


Figure 2.4: A simplified hidden Markov model of eukaryotic genes. Figure from Yoon [217] reproduced with permission of Bentham Science Publishers LTD. in the format thesis/dissertation via Copyright Clearance Center.
In top right is shown a set of four possible hidden states. In top left is shown the state space transition model, albeit without transition probabilities. In bottom, shown is an example of observed DNA sequence and the corresponding hidden states. In this example, the HMM models a gene simply as a sequence of nucleotide triplets (codons). No further gene structure is modeled nor are the triplets guaranteed to encode one of 20 amino acids.

2.5 Introduction to selected latent variable models

This section briefly introduces probabilistic latent variable models used and built upon in this thesis. First is described hidden Markov model with exact polynomial time inference and learning algorithms, used for probabilistic sequence classification in Chapter 5. Next follows gentle introduction to variational inference and variational autoencoders, a probabilistic framework for deep latent variable models with approximate amortized inference [115, 174, 116]. Finally, we describe semi-supervised variational autoencoder [117], an extension that additionally enables semi-supervised classification, which is also the basis for variational fair autoencoder [139] discussed in the next section. The approach of amortized variational inference and the described models are utilized and evaluated in Chapters 3 and 4.

2.5.1 Hidden Markov Model

A hidden Markov model (HMM) is a probabilistic model well suited for modeling of sequences of observable events that depend on unobservable (hidden) states. The hidden states are assumed to form a Markov chain, where the probability distribution over a hidden state at step i depends only on the previous state $i-1$; the *transition* probability. While the probability distribution of the observed symbol at step i then depends on the underlying state of the step i; the *emission* probability.

HMMs have been extensively used in many fields for decades, notably in speech processing [111] and analysis of biological sequences [52, 217]. In bioinformatics, extensions like pair-HMM or profile-HMM [54], were developed for various tasks such as pairwise and multiple

sequence alignment, genome annotation (Figure 2.4), protein secondary structure prediction, RNA structural alignment, and many more, such as recently for chromatin state discovery and characterization [58] or for base-calling in nanopore DNA sequencing [199, 136, 39] (now replaced by recurrent neural networks [21]).

First, we introduce the necessary notation for components of an HMM $\lambda = (A, B)$:

$\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$	set of N possible states
A	$N \times N$ transition probability matrix where each $a_{i,j}$ represents the probability of moving from state s_i to state s_j
$\mathcal{Z} = z_1, z_2, \ldots, z_T$	sequence of T (hidden) states, $\forall t : z_t \in \mathcal{S}$
$\mathcal{X} = x_1, x_2, \ldots, x_T$	sequence of T observations (emissions) from a vocabulary $\mathcal{V} = \{w_k\}_1^M$ of M possible tokens
B	$N \times M$ emission probability matrix where $b_{i,j}$ represents the probability of generating token w_j in state s_i; we shall denote $P(x_t = w_j \mid z_t = s_i)$ also as $b_i(x_t)$
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	initial probability distribution over states $\mathcal{S}$

Hidden Markov models can be characterized by three fundamental problems, as introduced by Rabiner [163], for which we then describe efficient algorithms:

Likelihood: Given an HMM $\lambda = (A, B)$ and a sequence of observations $\mathcal{X}$, determine $P(\mathcal{X}|\lambda)$ (forward algorithm, described below).

Inference: Given an HMM λ and a sequence of observations $\mathcal{X}$, find the most probable hidden state path in λ (Viterbi algorithm, described below).

Learning: Given an observation sequence and the set of possible states $\mathcal{S}$ in the HMM, find HMM model parameters A, B that maximize its probability (maximum likelihood learning).

In case the path of hidden states is known for the training observation sequences, the forward-backward algorithm can be used to estimate parameters by a variant of simple maximum likelihood estimation (MLE). Otherwise, if the ground-truth hidden states of training sequences are not known, an iterative expectation-maximization (EM) algorithm of Baum-Welsh [11] is needed. Details of these training algorithms are out of scope of this background chapter.

Forward and backward algorithm

To compute the likelihood $P(\mathcal{X}|\lambda)$ of a sequence $\mathcal{X}$ in an HMM $\lambda = (A, B)$ we need to sum over all possible hidden state paths. Note, from now on we leave out conditioning on λ to reduce

clutter in the notation when possible.

$$P(\mathcal{X}) = \sum_{\mathcal{Z}} P(\mathcal{X}, \mathcal{Z}) = \sum_{\mathcal{Z}} P(\mathcal{X}|\mathcal{Z})P(\mathcal{Z}) \tag{2.1}$$

This marginalization over $\mathcal{Z}$ can be efficiently computed in $O(N^2T)$ time by a dynamic programming (DP) algorithm called the forward algorithm or α-pass. During computation of the forward algorithm we keep an $T \times N$ table α, where

$$\alpha_t(i) = P(x_1, x_2, \ldots, x_t, z_t = s_i) \qquad (\alpha\text{-pass definition, 2.2})$$

is the probability of all possible hidden state paths that could generate the t prefix of observation sequence $\mathcal{X}_{\leq t} = x_1, x_2, \ldots, x_t$ and end in state s_i, i.e. $z_t = s_i$. The $\alpha_t(i)$ can be computed recursively by summing over all paths of length $t-1$ extended by transition to state s_i and emission of x_t in this state s_i:

$$\alpha_t(i) = \sum_{j=1}^{N} \alpha_{t-1}(j) \cdot a_{j,i} \cdot b_i(x_t) \qquad (\alpha\text{-pass recursion, 2.3})$$

Given we correctly initialize the dynamic programming table:

$$\alpha_1(i) = \pi_i b_i(x_1) \qquad (\alpha\text{-pass initialization, 2.4})$$

the likelihood of the observation sequence $\mathcal{X}$ is then:

$$P(\mathcal{X}) = \sum_{i=1}^{N} \alpha_T(i) \qquad (\alpha\text{-pass result, 2.5})$$

The dynamic programming recursion that the forward algorithm computes can be rewritten in backward fashion, creating an equivalent algorithm that computes

$$\beta_t(i) = P(x_{t+1}, x_{t+2}, \ldots, x_T, z_t = s_i) \qquad (\beta\text{-pass definition, 2.6})$$

This algorithm is called the backward algorithm or β-pass. We can use it to compute likelihood $P(\mathcal{X})$ as well, but its main utility is in the so-called forward-backward algorithm for maximum likelihood estimation of the model parameters $\lambda = (A, B)$.

Viterbi algorithm

Viterbi algorithm is an algorithm for finding an optimal sequence of hidden states $\mathcal{Z}^*$ for a given sequence of observations $\mathcal{X}$, an inference task also called decoding. Viterbi algorithm is a dynamic programming algorithm similar to the forward algorithm with two main modifications. Firstly, to compute likelihood of $\mathcal{X}$ in an optimal sequence of hidden states instead of marginalized

over all possible paths, as done by the forward algorithm above, we need to only consider the best extension of previous prefix paths instead of summing over all of the possible extensions. This requires simply replacing the summation by maximum operation in the recursion formula. Secondly, in order to reconstruct the optimal $\mathcal{Z}^*$, we need to keep track of optimal path extensions in form of backpointers $\tau_t(i)$ during the computation.

Viterbi dynamic programming algorithm:

$$v_t(i) = \max_{z_1,\ldots,z_{t-1}} P(z_1,\ldots,z_{t-1},x_1,\ldots,x_t,z_t=s_i) \qquad \text{(Viterbi definition, 2.7)}$$

$$v_1(i) = \pi_i b_i(x_1) \qquad \text{(Viterbi initialization, 2.8)}$$

$$\tau_1(i) = \perp \qquad (2.9)$$

$$v_t(i) = \max_{j=1}^{N} v_{t-1}(j)\cdot a_{j,i}\cdot b_i(x_t) \qquad \text{(Viterbi recursion, 2.10)}$$

$$\tau_t(i) = \arg\max_{j=1}^{N} v_{t-1}(j)\cdot a_{j,i}\cdot b_i(x_t) \qquad (2.11)$$

Finally, the probability of the observation sequence $\mathcal{X}$ given the most likely hidden state path $\mathcal{Z}^*$ in the HMM λ is:

$$P(\mathcal{X}|\mathcal{Z}^*) = \max_{i=1}^{N} v_T(i) \qquad \text{(Viterbi result, 2.12)}$$

We can then recover the entire most likely hidden state path $\mathcal{Z}^*$ by following backtraces $\tau_t(i)$ from the last state z_T^* of the optimal path:

$$z_T^* = \arg\max_{i=1}^{N} v_T(i) \qquad \text{(Viterbi backtrace start, 2.13)}$$

The final backtracing step takes additional $O(T)$ time, which does not change the overall time complexity $O(N^2T)$ of the algorithm.

2.5.2 Variational Autoencoder

Generative modeling is a paradigm in machine learning that aims to approximate how real data is generated, which can lead to learning of meaningful data representations, reusable also for other down-stream tasks beyond realistic sample generation. The variational autoencoders (VAEs) [115, 174] are a principled framework for learning expressive deep latent variable models that have been successfully applied to a variety of problems ranging from aforementioned generative modeling [25, 109, 118, 119], to semi-supervised learning [117, 139, 140, 162, 18], representation learning [20, 55, 169, 216, 78, 210], transfer learning [45], one-shot learning [175], and more [124].

In this section we provide a brief introduction to the framework of variational autoencoders,

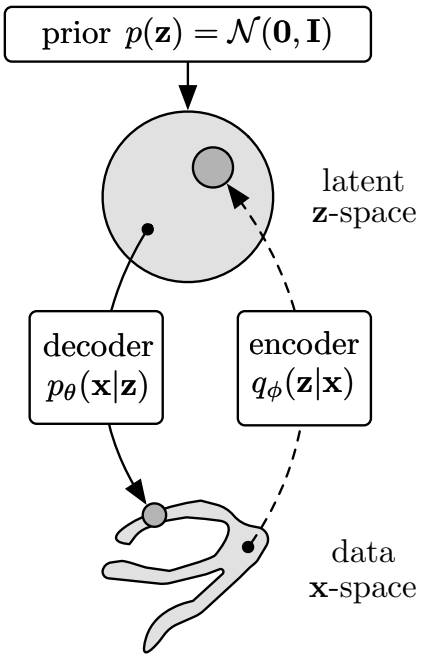


Figure 2.5: An illustration of variational autoencoder.

for an in-depth introduction please refer to Kingma and Welling [116], and to Koller and Friedman [122] for probabilistic latent variable models at large.

The major advantage of the VAE framework is that it enables modeling of highly complex conditional distributions between random variables of a probabilistic model. Each conditional distribution is modeled as a parametric distribution whose parameters are computed by a neural network. While there is no restriction on architecture of these neural networks, the used parametric distributions have to be *reparameterizable*. The reparameterization, explained later, allows stochastic gradient descent (SGD) to be used to conveniently learn the generative model jointly with its corresponding inference model.

We will demonstrate this variational framework on the simplest latent variable model $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, called the variational autoencoder (VAE), after which the framework bears its name. For simplicity, in this section we assume all distributions are multivariate Normals. The generative conditional distribution $p(\mathbf{x}|\mathbf{z})$, also called the decoder, is thus a Normal distribution parameterized by a deep neural network:

$$(\boldsymbol{\mu}_\mathbf{x}, \log \boldsymbol{\sigma}_\mathbf{x}) = \text{DecoderNN}_\theta(\mathbf{z}) \tag{2.14}$$

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\mathbf{x}, \text{diag}(\boldsymbol{\sigma}_\mathbf{x})) \tag{2.15}$$

where θ denotes the weights and biases of the neural network. Next, the prior over latent variable $\mathbf{z}$ is a standard Normal:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2.16}$$

The exact posterior inference and learning in VAE are in general intractable problems. Therefore, using the approach of variational inference, a parametric inference model $q_\phi(\mathbf{z}|\mathbf{x})$ is introduced and trained such that $q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x})$. This approximate posterior model is often referred to as the encoder or recognition model. Here ϕ denotes the variational parameters of this model. Analogously to p_θ above, the $q_\phi(\mathbf{z}|\mathbf{x})$ can be parameterized by a neural network with parameters ϕ:

$$(\boldsymbol{\mu}_\mathbf{z}, \log \boldsymbol{\sigma}_\mathbf{z}) = \text{EncoderNN}_\phi(\mathbf{x}) \tag{2.17}$$

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\mathbf{z}, \text{diag}(\boldsymbol{\sigma}_\mathbf{z})) \tag{2.18}$$

Note that VAEs use what is called the amortized variational inference [73], as a single inference model is learned that shares variational parameters across all datapoints, no per-datapoint optimization loop is required for the inference.

Evidence Lower BOund. Now that VAE model is defined, we need an efficient way to train it to maximize the *data likelihood*, also called *evidence*, w.r.t. the model parameters:

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \qquad \text{(per-datapoint likelihood, 2.19)}$$

As introduced above, VAE is a variational method. Instead of optimizing $p_\theta(\mathbf{x})$ exactly we aim to optimize alternative objective, the evidence lower bound (ELBO), that will lead to joint training of both model θ and variational ϕ parameters. The lower bound can be derived as follows [116]:

$$\log p_\theta(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x})] \tag{2.20}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p_\theta(\mathbf{z}|\mathbf{x})}\right] \tag{2.21}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right] \tag{2.22}$$

$$= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right]}_{=\mathcal{L}(\mathbf{x};\theta,\phi) \text{ (ELBO)}} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{z}|\mathbf{x})}\right]}_{=D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right]} \tag{2.23}$$

The emergent Kullback-Leibler (KL) divergence represents divergence of the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ from the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$, and also the tightness of the ELBO $\mathcal{L}(\mathbf{x}; \theta, \phi)$ w.r.t. the exact evidence $\log p_\theta(\mathbf{x})$:

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \log p_\theta(\mathbf{x}) - \underbrace{D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})\right]}_{\geq 0} \tag{2.24}$$

$$\leq \log p_\theta(\mathbf{x}) \tag{2.25}$$

Maximization of $\mathcal{L}(\mathbf{x}; \theta, \phi)$ thus leads to increase of the model likelihood and decrease of the

variational gap between the approximate and exact posterior.

Reparameterization trick. In order to train the VAE using SGD, we need to be able to backpropagate through ELBO to get gradients w.r.t. θ and ϕ parameters.

$$\begin{aligned}
\mathcal{L}(\mathbf{x};\theta,\phi) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] && (2.26)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}) + \log p(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\right] && (2.27)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}) + \log \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] && (2.28)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})}\right] && (2.29)\\
&= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{expected reconstruction likelihood}} \quad \underbrace{- D_{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]}_{\text{posterior regularization}} && (2.30)
\end{aligned}$$

ELBO written in the form above can be easily differentiated w.r.t. generative parameters θ, however for ϕ we cannot backpropagate through the expectation of the reconstruction likelihood term as it depends on ϕ (for Normal distributions the KL term can be evaluated analytically). Here comes in the reparameterization trick [115, 174], a change of variables such that the random variable $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ is expressed as a differentiable function g of another random variable $\boldsymbol{\epsilon}$ given $\mathbf{x}$ and ϕ:

$$\mathbf{z} = g(\boldsymbol{\epsilon}, \phi, \mathbf{x}) \tag{2.31}$$

Thanks to the fact that any Normal distribution can be reparameterized as a linear combination of a standard Normal:

$$\begin{aligned}
A &\sim \mathcal{N}(\mu, \sigma^2) && (2.32)\\
\epsilon &\sim \mathcal{N}(0, 1) && (2.33)\\
A &\overset{d}{=} \sigma\epsilon + \mu && \text{(change of variables, 2.34)}
\end{aligned}$$

we can define g as follows:

$$\begin{aligned}
(\boldsymbol{\mu}_\mathbf{z}, \log \boldsymbol{\sigma}_\mathbf{z}) &= \text{EncoderNN}_\phi(\mathbf{x}) && (2.35)\\
\boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) && (2.36)\\
\mathbf{z} &= \underbrace{\boldsymbol{\sigma}_\mathbf{z}\boldsymbol{\epsilon} + \boldsymbol{\mu}_\mathbf{z}}_{g(\boldsymbol{\epsilon},\phi,\mathbf{x})} && \text{(reparameterization trick, 2.37)}
\end{aligned}$$

It is then possible to rewrite the expectation of reconstruction likelihood formula such that it does not depend on the variational parameters ϕ:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] = \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\log p_\theta(\mathbf{x}|g(\boldsymbol{\epsilon}, \phi, \mathbf{x})\right] \tag{2.38}$$

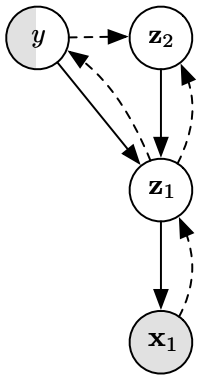


Figure 2.6: An illustration of semi-supervised variational autoencoder. Solid lines represent generative distributions p_θ and dashed lines represent variational posteriors q_ϕ.

This expectation now can be approximated by a simple Monte Carlo estimator, gradients of which are straightforward to compute w.r.t. both θ and ϕ, allowing the ELBO to be optimized by SGD.

In addition to differentiable encoders and decoders, the VAE framework as presented relies on using latent variables with parametric distributions that are reparameterizable. Many continuous distributions like Normal, Laplace or Cauchy are reparameterizable, but e.g. Beta is not. For several discrete distributions continuous approximations with gradient estimators were developed, e.g. for Bernoulli and categorical variables [142, 103, 204, 83] or for permutations [147]. Next, KL divergence for these distributions may not be analytical, in that case we need to use a Monte Carlo estimator or rewrite the ELBO into a different form. Last but not least, the restriction of using parametric distributions can be overcome by using flow-based extensions of the VAE framework, that can potentially model any continuous distribution [173, 118, 201]. Review of these models is outside of the scope.

2.5.3 Semi-supervised Variational Autoencoder

Semi-supervised Variational Autoencoder (SSVAE) is a deep latent variable model introduced by Kingma et al. [117] for semi-supervised learning joint with generative modeling. SSVAE is an extension of VAE by a second layer of hidden variables that further disentangle the VAE latent embedding into a latent class variable y and the other factors of variation $\mathbf{z}_2$, Figure 2.6. Kingma et al. [117] originally proposed SSVAE as a "stacked" model of two independently trained models, however the model can be trained jointly as shown by Louizos et al. [139].

Using the VAE framework of amortized variational inference the evidence lower bounds can

be derived for both labelled and unlabelled datapoints, denoted $\mathcal{L}_L$ and $\mathcal{L}_U$, respectively:

$$\mathcal{L}_L(\mathbf{x}_1, \mathbf{y};\ \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x}_1,\mathbf{y})}\big[\log p_\theta(\mathbf{x}_1, \mathbf{z}_1, \mathbf{z}_2, \mathbf{y}) - \log q_\phi(\mathbf{z}_1, \mathbf{z}_2|\mathbf{x}_1, \mathbf{y})\big] \tag{2.39}$$

$$\begin{aligned} &= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[\log p_\theta(\mathbf{x}_1|\mathbf{z}_1) - D_{KL}\left[q_\phi(\mathbf{z}_2|\mathbf{z}_1, \mathbf{y})||p(\mathbf{z}_2)\right]\big] \\ &\quad + \mathbb{E}_{q_\phi(\mathbf{z}_2|\mathbf{z}_1,\mathbf{y})}\big[-D_{KL}\left[q_\phi(\mathbf{z}_1|\mathbf{x}_1)||p_\theta(\mathbf{z}_1|\mathbf{z}_2, \mathbf{y})\right]\big] \\ &\quad + \log p(\mathbf{y}) \end{aligned} \tag{2.40}$$

$$\mathcal{L}_U(\mathbf{x}_1;\ \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2,\mathbf{y}|\mathbf{x}_1)}\big[\log p_\theta(\mathbf{x}_1, \mathbf{z}_1, \mathbf{z}_2, \mathbf{y}) - \log q_\phi(\mathbf{z}_1, \mathbf{z}_2, \mathbf{y}|\mathbf{x}_1)\big] \tag{2.41}$$

$$\begin{aligned} &= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[\log p_\theta(\mathbf{x}_1|\mathbf{z}_1) - D_{KL}\left[q_\phi(\mathbf{y}|\mathbf{z}_1)||p(\mathbf{y})\right]\big] \\ &\quad + \mathbb{E}_{q_\phi(\mathbf{y}|\mathbf{z}_1)q_\phi(\mathbf{z}_2|\mathbf{z}_1,\mathbf{y})}\big[-D_{KL}\left[q_\phi(\mathbf{z}_1|\mathbf{x}_1)||p_\theta(\mathbf{z}_1|\mathbf{z}_2, \mathbf{y})\right]\big] \\ &\quad + \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)q_\phi(\mathbf{y}|\mathbf{z}_1)}\big[-D_{KL}\left[q_\phi(\mathbf{z}_2|\mathbf{z}_1, \mathbf{y})||p(\mathbf{z}_2)\right]\big] \end{aligned} \tag{2.42}$$

The model parameters θ and variational parameters ϕ can be jointly optimized by SGD to maximize the ELBOs, analogously to VAE. The final training objective $\mathcal{J}_{\text{SSVAE}}$ is the sum of all per-datapoint ELBOs and of an additional prediction loss computed on labelled datapoints as otherwise the predictive posterior $q_\phi(\mathbf{y}|\mathbf{z}_1)$ would not get trained on the labelled data:

$$\begin{aligned} \mathcal{J}_{\text{SSVAE}} = &\sum_{(\mathbf{x}_1,\mathbf{y})\in L} \mathcal{L}_L\left(\mathbf{x}_1, \mathbf{y}; \theta, \phi\right) + \sum_{\mathbf{x}_1 \in U} \mathcal{L}_U\left(\mathbf{x}_1; \theta, \phi\right) \\ &+ w_y \sum_{(\mathbf{x}_1,\mathbf{y})\in L} \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[\log q_\phi(\mathbf{y} = \mathbf{t}|\mathbf{z}_1)\big] \end{aligned} \tag{2.43}$$

2.6 Introduction to domain adaptation

Domain adaptation is a methodology for learning a classifier in a setting, when the training data distribution; the source domain; is shifted from the test data distribution; the target domain. Typically, many labelled datapoints are available in the source domain, while in the target domain only a few or none. Note that this is the case in the problem of patient drug response prediction based on labelled dataset of cell line responses, the topic of Chapter 4. This section summarizes a popular approach to unsupervised domain adaptation based on learning of domain-invariant representations [123, 209, 222], such that the latent representation of the input is discriminative of the primary classification task while simultaneously invariant to the domain. This approach became popular with uptake of neural networks as they are particularly powerful feature extractors. We will describe in further detail two particular methods, domain-adversarial neural networks [66, 65, 2] and variational fair autoencoders [139, 18], that, respectively, represent two most common types of methodologies: i) adversarial training for domain invariance based on paradigm of generative adversarial networks [81], and ii) latent domain matching by discrepancy minimization or sub-space alignment [61, 19, 218, 151]. We conclude by summarizing most recent developments [186, 87, 101] and emergent criticism of the domain-invariant representation

learning approach [222, 214, 106].

2.6.1 Theoretical background

Here we summarize the theoretical motivation behind domain adaptation via learning of domain-invariant representations given Ben-David et al. [13, 14].

Formally, denote $\mathcal{X}$ the input feature space of the classification task and $\mathcal{Y} = \{0, 1\}$ to be the set of possible labels; assuming here a binary classification task for simplicity. Then in case of domain shift, there exist two different distributions over the joint space $\mathcal{X} \times \mathcal{Y}$ of features and labels, denoted as the *source* domain D_S and the *target* domain D_T. Additionally, let us denote $D_T^{\mathcal{X}}$ as the marginal distribution over features $\mathcal{X}$ in the target domain, which is equivalent to the D_T with labels $\mathcal{Y}$ marginalized out.

In case of unsupervised domain adaptation task the training set with $N = n + n'$ samples is composed of n labelled samples S drawn i.i.d. from D_S and of n' unlabelled samples T drawn i.i.d. from the target input domain $D_T^{\mathcal{X}}$:

$$S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \sim (D_S)^n \tag{2.44}$$

$$T = \{\mathbf{x}_i\}_{i=n+1}^{N} \sim \left(D_T^{\mathcal{X}}\right)^{n'} \tag{2.45}$$

$$\text{training set } = S \cup T \tag{2.46}$$

Then the goal of unsupervised domain adaptation is to learn a classifier $\eta : \mathcal{X} \to \mathcal{Y}$, such that it maximizes the probability of correct classification of samples from D_T, without seeing labelled target domain samples in the training. An equivalent alternative is to minimize the so-called target domain risk:

$$R_{D_T}(\eta) = 1 - \Pr_{(\mathbf{x},y)\in D_T}[\eta(\mathbf{x}) = y] \tag{2.47}$$

$$= \Pr_{(\mathbf{x},y)\in D_T}[\eta(\mathbf{x}) \neq y] \tag{2.48}$$

The general approach is to bound the target risk, i.e. the probability of target domain prediction error, by the sum of the source domain risk and of a distance between the source and target domain distributions. In this case Ben-David et al. [14] utilize the notion of $\mathcal{H}$-divergence [113] that is defined as follows.

Definition of $\mathcal{H}$-divergence: First, define a hypothesis class $\mathcal{H}$ as a set of binary classifiers $\eta : \mathcal{X} \to \{0, 1\}$ on an input space $\mathcal{X}$. Then given two domain distributions $D_S^{\mathcal{X}}$ and $D_T^{\mathcal{X}}$ over $\mathcal{X}$, and a hypothesis class $\mathcal{H}$, the $\mathcal{H}$-divergence between $D_S^{\mathcal{X}}$ and $D_T^{\mathcal{X}}$ is:

$$d_{\mathcal{H}}(D_S^{\mathcal{X}}, D_T^{\mathcal{X}}) = 2 \sup_{\eta\in\mathcal{H}} \left| \Pr_{\mathbf{x}\sim D_S^{\mathcal{X}}}[\eta(\mathbf{x}) = 1] - \Pr_{\mathbf{x}\sim D_T^{\mathcal{X}}}[\eta(\mathbf{x}) = 1] \right| \tag{2.49}$$

Intuitively, $\mathcal{H}$-divergence then corresponds to performance of a classifier from $\mathcal{H}$ that best

distinguishes samples drawn from the two domains $D_S^{\mathcal{X}}$ and $D_T^{\mathcal{X}}$. The empirical $\mathcal{H}$-divergence between two samples $S\sim (D_S^{\mathcal{X}})^m$ and $T\sim (D_T^{\mathcal{X}})^{m'}$ then can be computed as:

$$\hat{d}_{\mathcal{H}}(S,T) = 2\left(1 - \min_{\eta\in\mathcal{H}}\left[\frac{1}{n}\sum_{i=1}^{n} I[\eta(\mathbf{x}_i)=0] - \frac{1}{n'}\sum_{i=n+1}^{N} I[\eta(\mathbf{x}_i)=1]\right]\right) \tag{2.50}$$

Now coming back to the estimation of an upper bound on the target risk $R_{D_T}(\eta)$. The $R_{D_T}(\eta)$ can be assumed to be bounded by the sum of source risk and the domain $\mathcal{H}$-distance, i.e. formally for every classifier $\eta \in \mathcal{H}$:

$$R_{D_T}(\eta) \leq R_{D_S}(\eta) + d_{\mathcal{H}}(D_S^{\mathcal{X}}, D_T^{\mathcal{X}}) \tag{2.51}$$

Ben-David et al. [14] (Theorem 2) proved that for sample sets $S \sim (D_S)^n$ and $T \sim (D_T^{\mathcal{X}})^n$, with probability at least $1-\delta$ over the choice of S and T, it holds that for every classifier $\eta \in \mathcal{H}$:

$$R_{D_T}(\eta) \leq R_S(\eta) + \hat{d}_{\mathcal{H}}(S,T) + \beta + \gamma \tag{2.52}$$

where

$$R_S(\eta) = \frac{1}{n}\sum_{i=1}^{n} I[\eta(\mathbf{x}_i)=y_i], \qquad \text{(empirical source risk)} \tag{2.53}$$

$$\beta = \inf_{\eta^*\in\mathcal{H}}[R_{D_S}(\eta^*) + R_{D_T}(\eta^*)], \qquad \text{(best combined risk achievable in } \mathcal{H}) \tag{2.54}$$

and γ being a term dependent on the sample size, δ, and VC dimension of $\mathcal{H}$. This result shows that $R_{D_T}(\eta)$ will be small when:

1. the term β is small, that is, there exists a classifier that achieves low risk on both source and target distributions,

2. the learning procedure that is searching through $\mathcal{H}$ to find a good η should minimize a trade-off between the empirical source risk $R_S(\eta)$ and the empirical $\mathcal{H}$-divergence $\hat{d}_{\mathcal{H}}(S,T)$.

Ben-David et al. [14] further suggest that a good strategy to control the empirical $\mathcal{H}$-divergence $\hat{d}_{\mathcal{H}}(S,T)$ is by finding representation of the source and target domain samples such that the two sets are indistinguishable. Then the empirical $\mathcal{H}$-divergence is minimized, which implies that a classifier with small risk $R_S(\eta)$ on the source domain will also have low target domain risk $R_{D_T}(\eta)$.

Ganin et al. [66] follow this path and show how to train a deep neural network to represent the source and target domain samples to be indistinguishable, i.e. to keep $\hat{d}_{\mathcal{H}}(S,T)$ low, while simultaneously minimize the source domain prediction error and thus obtaining a classifier that performs well also on target domain.

2.6.2 Domain-Adversarial Neural Network

In this part we briefly present the concept of Domain-Adversarial Neural Network (DANN) proposed by Ganin et al. [66] for unsupervised domain adaptation. The DANN architecture is depicted in Figure 2.7.

In standard deep neural network it is possible to interpret one of the hidden layers as the latent representation, denoted as features $\mathbf{f}(\mathbf{x})$, of the input $\mathbf{x}$. That is, $\mathbf{f}(\mathbf{x}) = G_f(\mathbf{x}, \theta_f)$, where G_f is a feed-forward neural network with parameters θ_f. Then the other part of the network can be seen as the task classifier, here denoted as label predictor, $G_y(\mathbf{f}(\mathbf{x}), \theta_y)$. As motivated in previous section, it is desirable for these features to be predictive of the class labels in the original source domain and at the same time invariant w.r.t. discrimination between the source and target domain. Ganin et al. [66] proposed to achieve this by employing one more classifier $G_d(\mathbf{f}(\mathbf{x}), \theta_d)$ that operates on the features $\mathbf{f}$, the domain classifier. In such network, θ_f, θ_y are trained to *minimize* the class label prediction loss L_y, and θ_d is trained to *minimize* the domain prediction loss L_d, while at the same time θ_f also needs to be trained to *maximize* the domain prediction loss L_d. Authors showed that this domain-adversarial approach optimizes the domain $\mathcal{H}$-divergence. The learning procedure can be expressed as the following stochastic updates process:

$$\theta_f \leftarrow \theta_f - \mu \left(\frac{\partial L_y^i}{\partial \theta_f} - \lambda \frac{\partial L_d^i}{\partial \theta_f} \right) \tag{2.55}$$

$$\theta_y \leftarrow \theta_y - \mu \frac{\partial L_y^i}{\partial \theta_y} \tag{2.56}$$

$$\theta_d \leftarrow \theta_d - \mu \frac{\partial L_d^i}{\partial \theta_d} \tag{2.57}$$

where λ is a parameter that controls the trade-off between minimizing L_y and maximizing L_d.

Ganin et al. [66] reduced the proposed stochastic update process to standard backpropagation training of neural networks by introducing so-called *gradient reversal layer* (GRL). A GRL layer $\mathcal{R}$ has only one parameter and that is λ. In forward pass GRL is an identity function $\mathcal{R}_\lambda(\mathbf{x}) = \mathbf{x}$, but in backward pass, the gradient of GRL is negative λ, i.e. $\frac{\partial \mathcal{R}_\lambda(\mathbf{x})}{\partial \mathbf{x}} = -\lambda \mathbf{I}$. That is, GRL is not influencing the forward pass, but in backward pass it reverses and scales the gradient propagated through it. Plugging such gradient reversal layer between G_f and G_d, that is between the feature extractor and the domain classifier makes the standard backpropagation learning that minimizes error $E(\theta_f, \theta_y, \theta_d)$ equivalent to the updates above, where:

$$E(\theta_f, \theta_y, \theta_d) = \sum_{\mathbf{x}_i \in S} L_y(G_y(G_f(\mathbf{x}_i, \theta_f), \theta_y), y_i) + \sum_{\mathbf{x}_i \in S \cup T} L_d(G_d(\mathcal{R}_\lambda(G_f(\mathbf{x}_i, \theta_f)), \theta_d), d_i) \tag{2.58}$$

$$S = \{(\mathbf{x}_i, d_i = 0, y_i)\}_{i=1}^{n} \quad \text{(labelled source domain samples)} \tag{2.59}$$

$$T = \{(\mathbf{x}_i, d_i = 1)\}_{i=n+1}^{N} \quad \text{(unlabelled target domain samples)} \tag{2.60}$$

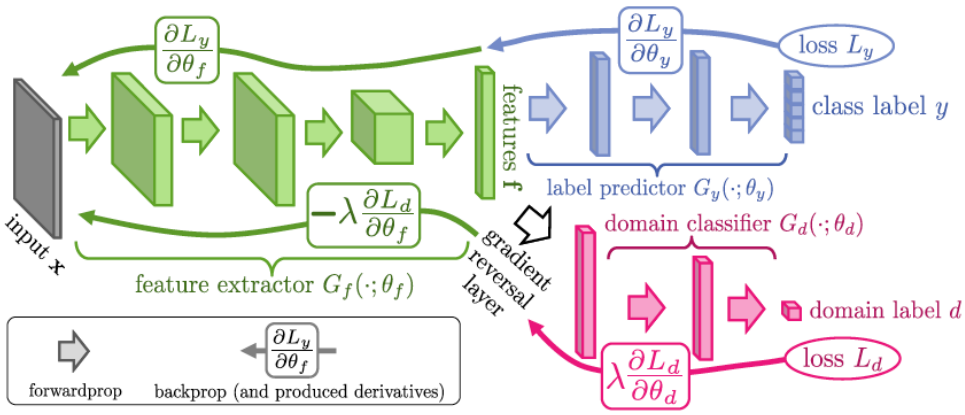


Figure 2.7: General architecture of domain-adversarial neural network; figure from Ganin et al. [66] licensed under CC BY 4.0. The first part of the network, here denoted as the feature extractor (in green) is trained to provide the most representative features for the primary classification task, while not being discriminative of the source and target domain. This is achieved by backpropagating the gradient of the task classifier (here denoted as label predictor, in blue) to the feature extractor. At the same time, the gradient of the domain classifier (in magenta) is reversed before backpropagating to the feature extractor, thus promoting domain invariant features $\mathbf{f}$.

To maintain theoretical bounds from Ben-David et al. [14] Theorem 2 (eq. 2.52) the label classification hypothesis class $\mathcal{H}_y$ that corresponds to G_y has to be a subset of domain hypothesis class $\mathcal{H}_d$ of G_d, i.e. the domain classifier part of the network has to be at least as expressive as the label classification sub-network.

2.6.3 Variational Fair Autoencoder

Variational Fair Autoencoder (VFAE) is an extension of the SSVAE model [139, 18]. The data encoder $q_\phi(\mathbf{z}_1|\mathbf{x}_1)$ and decoder $p_\theta(\mathbf{x}_1|\mathbf{z}_1)$ are extended to be conditioned on an observed variable s that is a domain indicator variable, Figure 2.8. Here we assume that $s \in \{0, 1\}$.

VFAE was proposed in context of learning fair representations, when the variable s is

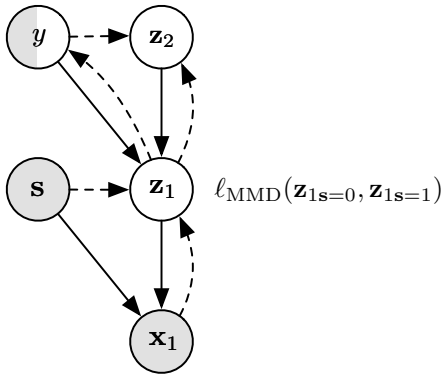


Figure 2.8: An illustration of variational fair autoencoder.

considered sensitive information that may be correlated with the dependent variable $\mathbf{y}$ because of a dataset bias, and the goal is to remove this 'unfair' influence of s on prediction of $\mathbf{y}$. In case of domain adaptation, s as considered a nuisance variable and the goal is to remove the domain s from the latent representations $\mathbf{z}_1$ in order to improve cross-domain prediction performance on $\mathbf{y}$. In addition to conditioning of the input data $\mathbf{x}_1$ encoder and decoder on s, Louizos et al. [139] incorporated an additional penalty term based on the Maximum Mean Discrepancy (MMD) [85] measure to remove remaining dependencies of $\mathbf{z}_1$ on s, thus achieving domain-invariant representation.

The VFAE training objective $\mathcal{J}_{\text{VFAE}}$ is similar to $\mathcal{J}_{\text{SSVAE}}$ with two differences. Firstly, the ELBOs of labelled and unlabelled datapoints are trivially extended by conditioning on the domain variable s. Secondly, added is the MMD distance between $\mathbf{z}_1$ embeddings of the datapoints of $s = 0$ domain and $s = 1$ domain, denoted $\mathbf{Z}_{1,s=0}$ and $\mathbf{Z}_{1,s=1}$, respectively.

$$\begin{aligned}\mathcal{J}_{\text{VFAE}} = &\sum_{(\mathbf{x}_1,\mathbf{y},s)\in L} \mathcal{L}_L\left(\mathbf{x}_1,\mathbf{y},s;\theta,\phi\right) + \sum_{(\mathbf{x}_1,s)\in U} \mathcal{L}_U\left(\mathbf{x}_1,s;\theta,\phi\right) \\ &+ w_y \sum_{(\mathbf{x}_1,\mathbf{y},s)\in L} \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1,s)}\big[\log q_\phi(\mathbf{y}=\mathbf{t}|\mathbf{z}_1)\big] \\ &- w_{\text{MMD}}\ \ell_{\text{MMD}}(\mathbf{Z}_{1,s=0},\mathbf{Z}_{1,s=1})\end{aligned} \tag{2.61}$$

Maximum Mean Discrepancy is a kernel-based statistic developed for comparing samples from two probability distributions and testing the null hypothesis that these distributions are equal [85, 86]:

$$\text{MMD}^2(x,y) = \frac{1}{N_x^2}\sum_i\sum_j k(x_i,x_j) + \frac{1}{N_y^2}\sum_i\sum_j k(y_i,y_j) - 2\sum_i\sum_j k(x,y) \tag{2.62}$$

where $k(a,b) = e^{-\frac{\|a-b\|^2}{2\sigma^2}}$ is an RBF (Gaussian) kernel with bandwidth σ. Exact computation of the MMD takes time quadratic in the number of samples, however there exists a linear time approximation using *random kitchen sinks* approach, FastMMD, developed by Zhao and Meng [223]:

$$\text{MMD}^2(x,y) = \|\phi(X) - \phi(Y)\|^2 \tag{2.63}$$

where,

$$\phi(X) = \frac{1}{N_x}\sum_i \sqrt{\frac{2}{D}}\cos\left(\frac{W}{\sigma}x_i + b\right) \tag{2.64}$$

$$W \sim \mathcal{N}(0,1) \in R^{n_{basis}\times D} \tag{2.65}$$

$$b \sim U(0,2\pi) \tag{2.66}$$

In our implementation of VFAE used in experiments in Chapter 4, we use the FastMMD linear approximation method. We set the number of random basis $n_{basis} = 500$, while we set the RBF

kernel bandwidth based on "median heuristic" [70]:

$$\sigma = \sqrt{\frac{\gamma}{2}} \tag{2.67}$$

where γ is an estimate of the median square distance in the sample set. Note that Botros and Tomczak [18, eq. 33] suggest to use $\gamma = 2M$, where M is the dimensionality of latent representation $\mathbf{z}_1$. This corresponds to the median heuristic when assuming the two sets of samples, X and Y, being compared are drawn independently from the standard Normal distribution. In that case the random variable $X - Y$ is distributed as a Normal with mean 0 and variance 2. Therefore the expectation of its square equals its variance + square of its mean, which is 2, resulting in the expected square distance of $2M$ for samples from M-dimensional standard Normal distribution. This would be a reasonable assumption if we imposed $\mathcal{N}(\mathbf{0}, \mathbf{I})$ prior over $\mathbf{z}_1$. However that is only the case in a VAE, in VFAE $q_\phi(\mathbf{z}_1|\mathbf{x}_1)$ is a diagonal Normal, but not necessarily with zero mean and unit variance, thus we use an empirical estimate of the median square distance to set γ, as mentioned earlier.

In a follow-up work, Botros and Tomczak [18] slightly improved on the VFAE model by using a different variational posterior factorization coupled with learned mixture priors (VampPrior [200]). They also explored use of Mutual Information for domain matching, however in case of fully observed domain indicator s, the MMD regularizer yielded better results.

2.6.4 Limitations of learning domain-invariant representations

Transfer learning and domain adaptation by learning of domain-invariant representations has been theoretically motivated, as discussed in Section 2.6.1, and frequently used in the literature. In addition to the aforementioned methods, many other successful neural network models encourage similarity between their latent representations w.r.t. the domains. In [137, 138] Long et al. used MMD or its extension to align layer activations of convolutional neural networks. Zhuang et al. [224] used KL-divergence to regularize representation learned by a supervised autoencoder such that the marginal domain distributions are matched. Bousmalis et al. [19] and Hou et al. [98] developed autoencoding models that split their latent representation into a domain-invariant and a domain-specific part, recognizing that in some applications domain-specific features are important for the prediction task. However this approach is only feasible for semi-supervised domain adaptation.

A common pattern to these methods is that the domain similarity is often enforced by minimizing a certain distance between the domain-specific data representations. The often used KL-divergence is an asymmetric measure of how one probability distribution p is different from another distribution q in terms of information loss. It can be viewed as the difference between the cross entropy of p and q and the entropy of p. In case q is Normal distribution, the KL-divergence is minimized when the first two moments of the q density match the first two moments of p [125]. Next, particularly popular has been the use of Maximum Mean Discrepancy (MMD) [86];

described above in Section 2.6.3. Through the Taylor expansion of MMD with a universal kernel, such as the Gaussian kernel, the MMD minimization can be considered a minimization of a distance between weighted sums of all raw moments of the two distributions [132]. In addition, several other discrepancy metrics have been proposed. Zellinger et al. [218] introduced Central Moment Discrepancy, a differentiable distance function that directly measures also differences of higher order central moments. Last but not least, Ben-David et al. [13] proposed Proxy $\mathcal{A}$-distance, which essentially measures how difficult it is for a classifier to discriminate between the domains. While Proxy $\mathcal{A}$-distance was used to quantify domain discrepancy [76], it was however not until Ganin et al. [66] that directly minimized this distance.

Another common approach has been dimensionality reduction followed by manifold alignment [82, 51, 80, 61]. Particularly interesting is recent work of Mourragui et al. [151], who introduced PRECISE, a PCA sub-space alignment method applied to domain adaptation for drug response prediction.

All the reviewed methods and approaches for unsupervised domain adaptation rely on over-simplified conclusion that better domain-invariance leads to better domain-adaptation. However that is not universally true, as *sufficient* and *necessary* conditions are not well enough understood, which has very recently been noted [214, 106, 222]. A common necessary assumption for unsupervised domain adaptation is *covariate shift*, which states that the conditional distribution of class labels given input features $P(Y|X)$ does not change between domains, while the marginal distribution $P(X)$ of the features (covariates) is allowed to differ. Ben-David et al. [15] already showed that covariate shift on its own is not a sufficient condition. While Ben-David and Urner [12] showed that *sufficient support* assumption, which states that source domain feature density has to at least ϵ-overlap all regions with any target domain support, is sufficient. Unfortunately, as noted by D'Amour et al. [38] and Johansson et al. [106], the higher dimensional the input space is, the less likely it is that such an overlap would hold. Thus sufficient support assumption is typically invalid in many high-dimensional datasets.

Johansson et al. [106] pointed out that the 3rd term of Ben-David's bound β, eq. 2.54, is ignored by DANN [66] and other methods, which can lead to its explosion, making the upper bound of the target domain risk uninformative. This scenario happens when enforcing domain-invariant representation causes significant loss of class-predictive features, which means, that the hypothesis class is significantly worse in the learned representation space compared to the original input space. Johansson et al. [106] suggest that adding autoencoding objective may help to maintain the original hypothesis class. Therefore autoencoding models such as VFAE or Domain Separation Networks [19] can be more robust, while it also points to existence of a trade-off between domain-invariance and richness of hypothesis class in a learned representation. Their analysis concludes that support overlap of target and source domain feature distributions is more important than achieving their matching.

Wu et al. [214] recognized that unequal ratio of classes in source and target domain, i.e. *label shift*, leads to a failure of domain-invariant representation learning. Under these circumstances,

enforcing domain-invariance necessary leads to representation that partially mismatches classes between the domains. Wu et al. [214] suggest several asymmetrically relaxed distribution alignment distances that mitigate this issue. Secondly, in case there is no class-preserving overlap between the domains, domain-invariant representation can lead to domain alignment that arbitrarily permutes target domain classes. This is because in unsupervised domain adaptation all such alignments are indistinguishable from each other and achieve the same training objective. In this case additional assumptions are necessary, like the *cluster assumption*, that datapoints from different domains that are close together belong in the same class. DIRT-T and VADA method of Shu et al. [186] rely on this cluster assumption.

Zhao et al. [222] also came to the conclusion that domain-invariance is not a sufficient condition for successful domain adaptation, particularly in case of *conditional shift*, that is, when class-conditional distributions of input features are not stationary across domains. They show that learning invariant representations can in fact be harmful as it can lead to breaking of originally favourable underlying problem structure. Their theoretical analysis uncovered that when marginal distribution of class labels differs between the domains, then there is a fundamental trade-off between achieving small combined classification error in both domains and learning domain-invariant representations. This is a conclusion similar to that of Wu et al. [214].

Therefore the aforementioned conditions of successful unsupervised domain adaptation via domain-invariant representation learning need to be assessed in the application dataset. We conduct an empirical analysis for the problem of drug response prediction from gene expression features in Chapter 4.

Chapter 3

Dr.VAE: improving drug response prediction via modeling of drug perturbation effects[†]

3.1 Introduction

Personalized drug response prediction promises to improve the therapy response rate in life-threatening diseases, such as cancer. There are two main impediments that make the task of drug response prediction highly challenging. First, the space of all possible treatments and their combinations for a given condition is prohibitively large to be explored exhaustively in clinical settings, drastically limiting the sample size for many therapies and tissues of interest. Second, cancer heterogeneity among patients is very high, reducing the statistical power of biomarker detection. These two conditions make it hard to characterize the genotype-to-phenotype landscape comprehensively making it difficult to accurately stratify drug treatment options for a particular cancer patient. To fulfill the promise of precision medicine, we need predictive models that can take advantage of heterogeneous, sparsely sampled data and data generated from pre-clinical model systems, such as cancer cell lines, to improve our prediction ability.

In the last decade there have been several public releases of large-scale drug screens in cancer cell lines. The greatest advantage of cell lines is their potential for high-throughput experiments as it is possible to screen cell lines against thousands of chemical compounds, both clinically-approved and experimental. This screening task was undertaken by several large consortia and pharmaceutical companies resulting in large, valuable public data resources [171, 68, 9, 215, 170, 92]. The availability of these large cancer cell line datasets spurred the development of predictive models [156, 7, 220, 208, 219, 129, 8, 179, 196] and computational challenge-based competitions [34, 149].

Particularly influential has been the NCI-DREAM drug prediction challenge [34]. This

[†]Work published in Oxford *Bioinformatics* (2019) [166] and earlier workshop contributions [167, 168].

challenge had 44 competing methodological submissions, categorized into six major methodological types. Their post-competition analysis revealed two particular trends among the most successful methods, the ability to model non-linear relationships between data and outcomes, and incorporating prior knowledge such as biological pathways. The winner of this challenge incorporated these approaches together with multi-drug learning by developing Bayesian multitask multiple kernel learning method [34].

Complementary to large-scale cell line viability screens, the National Institutes of Health Library of Integrated Network-based Cellular Signatures (NIH LINCS) Connectivity Map (CMap) [192] project measured the transcriptional perturbations induced by over 20000 chemical compounds by profiling 1000 landmark genes in a set of 77 human cell lines before and after short-term drug treatment. These case-control matched experiments show how the expression of these genes changed in response to drug treatment at various concentration levels, typically after 6 or 24 h treatment duration. The set of drug-induced up- and down-regulation signatures is referred to as a drug perturbation signature [192, 187]. Combining response and perturbation data is expected to ultimately yield a better and more biologically relevant model of drug response [192, 154].

Previous work by [154] studied transcriptomic perturbations of six breast cancer cell lines, from an initial CMap release, in combination with phenotypic drug response measurements to determine whether cell lines that have similar phenotypic drug response also share common patterns in drug-induced gene expression perturbation. Their analysis concluded that this is the case for some drugs (inhibitors of cell-cycle kinases), but for other drugs the molecular response was cell-type specific, and for some drug-cell line combinations a significant transcription perturbation had no measurable impact on cell growth. These results motivated us to develop a unified method that could identify more complex associations of molecular perturbations and phenotypic responses that are potentially unique to a cell line subgroup.

The drug response prediction problem suffers from a high feature-to-sample ratio, where only a limited number of samples are available compared to the large number of measured molecular features (e.g. gene expression levels for thousands of genes). One way to alleviate this hindrance is to find a reduced representation of the original data that captures the essential information needed for the prediction task. Here, we take the approach of semi-supervised generative modeling based on variational autoencoders (VAE) [115] that present a way to model complex conditional distributions. Way and Greene [210] have shown that VAE can extract biologically meaningful representation of cancer transcriptomic profiles, while Dincer et al. [45] combined a pre-trained VAE and a separately trained linear model in a drug response prediction method named DeepProfile. Contrary to Dincer et al. [45] we aim to jointly learn a latent embedding that improves our ability to predict drug response (phenotypic outcome), while leveraging the originally unsupervised (unknown phenotypic outcome) drug perturbation experiments to aid in the learning of such embedding.

We introduce Drug Response Variational Autoencoder (Dr.VAE), a deep generative model to

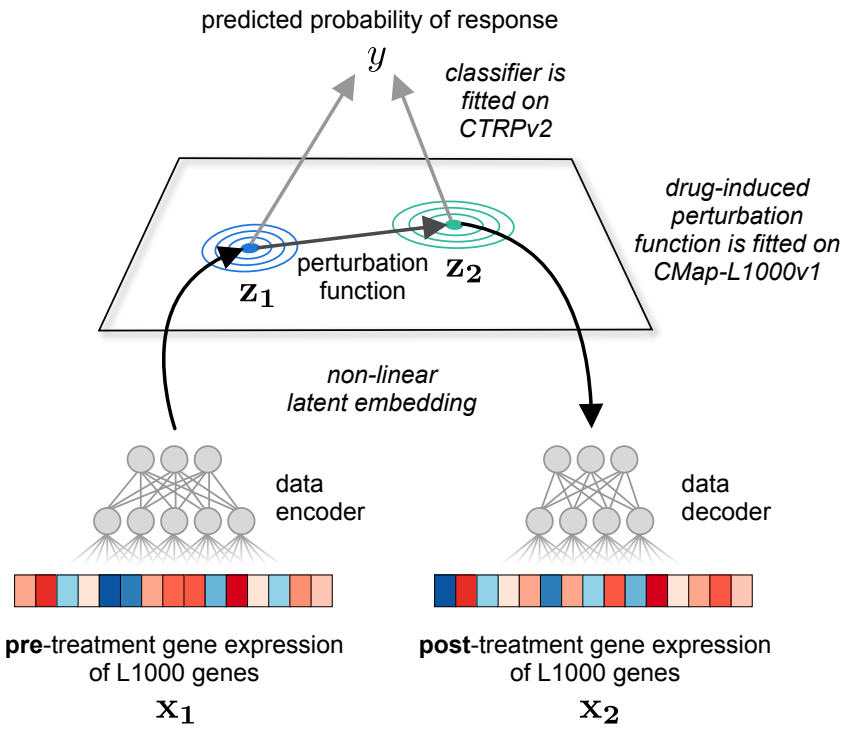


Figure 3.1: **An overview of Dr.VAE prediction process.** In training, Dr.VAE learns a drug response classifier jointly with a latent representation of pre-treatment gene expression and its drug-induced change. To make a prediction, we first embed the pre-treatment gene expression $\mathbf{x}_1$, and then, from this latent representation $\mathbf{z}_1$ we predict latent representation of post-treatment state $\mathbf{z}_2$. Based on both $\mathbf{z}_1$ and $\mathbf{z}_2$, a logistic regression classifier predicts the probability of positive response. Additionally, we can decode the predicted post-treatment latent representation $\mathbf{z}_2$ to the gene expression data space, but this is not required for drug response classification.

predict drug response from transcriptomic perturbation signatures. Dr.VAE is a probabilistic graphical model where each conditional distribution is computed by a deep neural network. The model jointly learns a drug response predictor and a generative model of drug perturbation effects in a low-dimensional latent representation of gene expression. This latent space is defined by an encoder and decoder, both parameterized by a neural network, that, respectively, translate to and from this latent space. The entire model, together with neural networks for approximate inference, is optimized jointly end-to-end to maximize evidence (marginal likelihood) of the observed training data. An overview of Dr.VAE is illustrated in Figure 3.1.

In our results, Dr.VAE significantly outperformed classification models typically used in the field in more than half of the tested drugs and performed on par for most of the other drugs. We show that the achieved improvement of Dr.VAE in drug response prediction is indeed due to the joint modeling of drug response and drug-induced perturbation effects. This result is further confirmed by observing that even unsupervised generative modeling of gene expression and drug-induced perturbations yields a low-dimensional representation that is better suited for subsequent training of standard classification models than the original data representation or

representation obtained by principal component analysis (PCA).

3.2 Materials and methods

3.2.1 Pharmacogenomics high-throughput cell line datasets

We harness two principally different types of pharmacogenomics datasets, both retrieved via PharmacoGx R package [187] and PharmacoDB [188]. First is a database of sensitivity of cancer cell lines to drug treatment, the Cancer Therapeutic Response Portal (CTRPv2) [170], that provides relative viability of cell lines at various drug concentration levels for combination of up to 860 cell lines and 481 drug compounds. Sensitivity of the cell lines to a drug treatment is quantified by the area above the dose-response curve (AAC), which was recomputed by PharmacoGx from raw CTRPv2 experimental results. We further binarized the continuous AAC by the waterfall method [9, 89], turning the sensitivity prediction task into a discrete classification task.

Secondly, we utilized the NIH LINCS Consortium CMap project. The recently extended CMap, termed CMap-L1000v1 [192], screened perturbation effects of 19811 drug compounds on gene expression of L1000 landmark genes in up to 77 cell lines. Experiments in CMap-L1000v1 do not measure the drug treatment sensitivity, however some of the cell lines were independently tested in CTRPv2 as well. We cross-referenced these cell lines and assigned the corresponding label to their perturbation measurements.

From the CMap-L1000v1 dataset, we used the level 3 data, i.e. the quantile normalized gene expression of 978 landmark genes measured on Luminex based L1000 platform shown to be consistent with gene expression measured by RNAseq [192, 178]. From the available set of experimental conditions, we selected perturbation experiments with duration of 6 h conducted at the most common concentration level for each particular drug. That is, a concentration level that most cell lines were measured at for that drug. In case a cell line was not tested at the chosen concentration, we used the closest tested concentration. Next, we matched controls (DMSO vehicle) experiments to the drug perturbation experiments by the batch ID and bead ID, to minimize batch effects between the cases and controls. Further, we filtered the selected case-control pairs by correlation (>0.75 Pearson ρ) to filter out possibly mislabeled experiments or outliers.

CTRPv2 and CMap-L1000v1 datasets had 973 common genes. We standardized the expression values to zero mean and unit variance within each gene. For further homogenization, including batch effect removal and differences between two incorporated data sources, we also removed the first principal component (explaining 12.8% of variation) from the pooled dataset.

We selected 26 drugs tested in both CTRPv2 and CMap-L1000v1 datasets based on two simple criteria: (i) for each selected drug at least eight distinct cell lines were tested in CMap-L1000v1 perturbation experiments; and (ii) at least 20% of screened cell lines in CTRPv2 were sensitive to the drug after binarization of dose-response AAC. The dataset summary is detailed

in Appendix A.6.

3.2.2 Dr.VAE

We present Dr.VAE, a new machine learning model based on a semi-supervised generative model. Dr.VAE learns a latent embedding of the gene expression. The latent embedding takes advantage of both cell line viability experiments that measure drug response outcome directly and, at the same time, the drug-induced transcription change, which in our case is modeled as a linear function in this latent space. This is achieved via joint training of the model on (i) 'perturbation pairs' $[\mathbf{x}_1, \mathbf{x}_2]$ of pre-treatment (control) and post-treatment gene expression (outcome label $\mathbf{y}$ is only observed for some pairs) and (ii) 'singletons' of pre-treatment gene expression with no known post-treatment expression. Most of the outcome $\mathbf{y}$ labeled data are in the latter category. We model the drug perturbation effects with a single step latent time series model, similar to Deep Kalman Filter [124] and structured graphical models with amortized inference [109]. The graphical representation of Dr.VAE model is shown in Figure 3.2.

Formally, Drug Response VAE models a joint distribution $p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y})$ of pre-treatment and post-treatment gene expression $\mathbf{x}_1, \mathbf{x}_2$, their latent embedding $\mathbf{z}_1, \mathbf{z}_2$, response class $\mathbf{y}$, and class-independent latent representation of the pre-treatment expression $\mathbf{z}_3$. Factorization of this joint probability distribution is depicted in Figure 3.2(a) (solid edges) and is as follows:

$$p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}) = p(\mathbf{x}_1|\mathbf{z}_1) \cdot p(\mathbf{x}_2|\mathbf{z}_2) \cdot p(\mathbf{z}_2|\mathbf{z}_1) \cdot p(\mathbf{z}_1|\mathbf{z}_3, \mathbf{y}) \cdot p(\mathbf{z}_3) \cdot p(\mathbf{y}) \tag{3.1}$$

Individual conditional generative distributions $p(\cdot)$ of Dr.VAE take the form of diagonal multivariate Gaussian distributions, while $p(\mathbf{y})$ is a uniform categorical prior over the binary response $\mathbf{y}$ and prior $p(\mathbf{z}_3)$ is a unit Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The conditional distributions are parameterized by neural networks with a set of parameters θ, analogously to a VAE [115, 174]. We want to model all gene expression measurements in a single latent space, thus the pre- and post-treatment gene expression have to be embedded into a common latent space. This is achieved by sharing the 'data decoder' $p_\theta(\mathbf{x}_k|\mathbf{z}_k)$ for both $k \in \{1, 2\}$. Additionally, we constrain the mean function of the perturbation $p_\theta(\mathbf{z}_2|\mathbf{z}_1)$ to be a linear function $\mathbf{z}_1 + \mathbf{W}\mathbf{z}_1 + \mathbf{b}$. Here, the $\mathbf{W}$ and $\mathbf{b}$ are initialized close to zero, such that $p_\theta(\mathbf{z}_2|\mathbf{z}_1)$ starts as an identity function in the beginning of optimization process.

In order to train and use our model, we need to be able to perform efficient inference of the hidden variables from the observed variables. We turn to stochastic variational inference and introduce an approximation q to the true posterior. We assume this approximate posterior q to factorize as shown in Figure 3.2(a) (dashed edges). Akin to generative distributions p introduced above, the variational distributions are diagonal multivariate Gaussian distributions, with exception of $q_\phi(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2)$, parameterized by neural networks with a set of parameters ϕ. The 'data encoder' $q_\phi(\mathbf{z}_k|\mathbf{x}_k)$, detailed in Figure 3.2(d), is shared between pre- and post-treatment for the same reason the data decoder is shared. The classification posterior $q_\phi(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2)$ is a

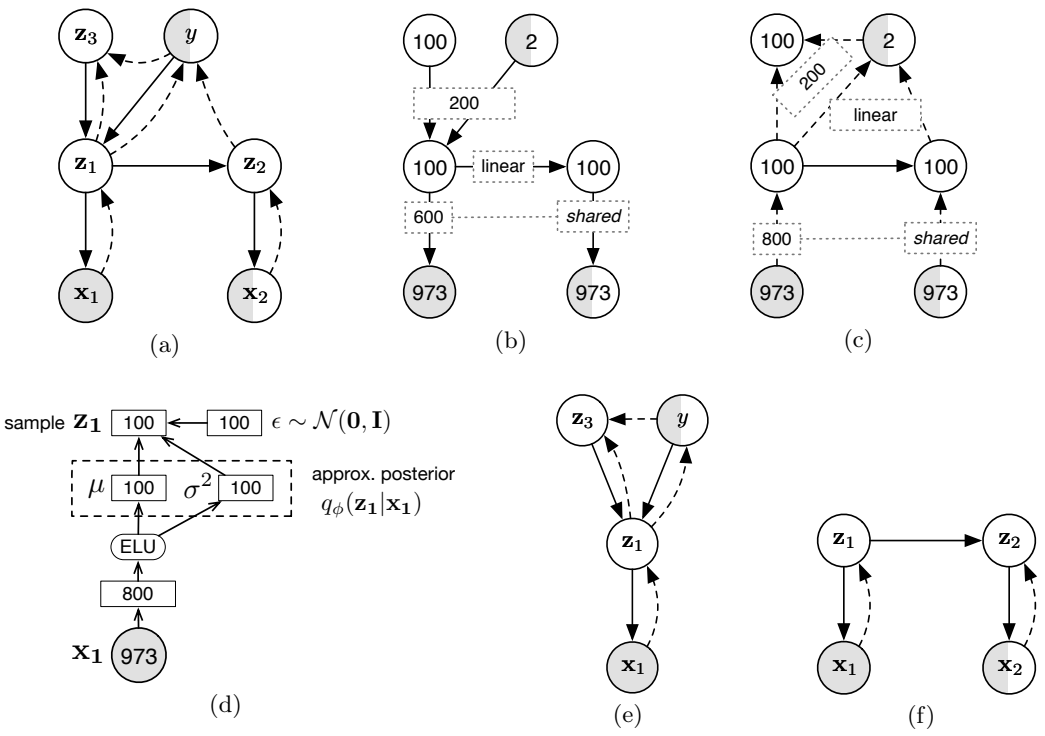


Figure 3.2: **Dr.VAE model and its derivatives.** (a) Factorization of the generative distribution p (solid edges) and of the approximate posterior q (dashed edges). In case the post-treatment gene expression $\mathbf{x}_2$ is not observed, we use the expected posterior $\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\left[p_\theta(\mathbf{z}_2|\mathbf{z}_1)\right]$ for $\mathbf{z}_2$ instead. (b,c) Hyperparameters of the generative and inference model, respectively. Node labels show dimensionality of the corresponding random variables, while edge labels show architecture of the encoders/decoders between the respective random variables. Note, that the 'data decoder' $p_\theta(\mathbf{x}_k|\mathbf{z}_k)$ is shared for both $k \in \{1, 2\}$ and so is the 'data encoder' $q_\phi(\mathbf{z}_k|\mathbf{x}_k)$. (d) Detailed depiction of data-to-latent-space encoder $q_\phi(\mathbf{z}_k|\mathbf{x}_k)$ and of the reparameterization trick. (e) Factorization of SSVAE model [117], we set the hyperparameters of generative and inference distributions equivalently to the analogous distributions in Dr.VAE as shown in (b,c,d). (f) Factorization of PertVAE model, we set the hyperparameters of generative and inference distributions equivalently to the analogous distributions in Dr.VAE (b,c,d).

categorical distribution parameterized by a linear function with soft-max activation over two output units. In our implementation, we use the latent embedding of pre-treatment state and the predicted perturbation difference $[\mathbf{z}_1, \mathbf{z}_2 - \mathbf{z}_1]$ instead of $[\mathbf{z}_1, \mathbf{z}_2]$ as the classifier input. We found that this slightly improves the performance.

Ideally we would want to fit the θ and ϕ parameters to maximize the evidence (marginal likelihood) of the observed data, which is a difficult task and subject to active research in the area of stochastic inference. However, following [115, 139, 117] we can derive a lower bound on the evidence of each set of observed variables. We have four different sets of observed variables that correspond to four different types of data we want to fit Dr.VAE to. Therefore there are

four different evidence lower bounds for us to optimize:

$$\text{labeled perturbation pairs } LP: \quad \sum \mathcal{L}_{LP}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}; \theta, \phi) \tag{3.2}$$

$$\text{unlabeled perturbation pairs } UP: \quad \sum \mathcal{L}_{UP}(\mathbf{x}_1, \mathbf{x}_2; \theta, \phi) \tag{3.3}$$

$$\text{labeled pre-treatment singletons } LS: \quad \sum \mathcal{L}_{LS}(\mathbf{x}_1, \mathbf{y}; \theta, \phi) \tag{3.4}$$

$$\text{unlabeled pre-treatment singletons } US: \quad \sum \mathcal{L}_{US}(\mathbf{x}_1; \theta, \phi) \tag{3.5}$$

The sum of these four specific evidence lower bounds, $\text{ELBO}_{\text{DrVAE}}$, is the evidence lower bound we need to maximize. Moreover, we need to explicitly introduce cross-entropy loss of the predictive posterior $\log q_\phi(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2)$ so that it is trained on labeled data as well. Analogous to semi-supervised variational autoencoder (SSVAE) [117], this explicit loss is required since in the labeled data the random variable $\mathbf{y}$ is observed and therefore the lower bounds $\mathcal{L}_{LP}$ and $\mathcal{L}_{LS}$ are conditioned on $\mathbf{y}$ and do not contribute to fitting of $q_\phi(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2)$. Using the reparameterization trick [115] it is possible to backpropagate through the final objective and jointly optimize parameters of all p_θ and q_ϕ distributions by gradient decent. In our implementation, we compute the parameter updates by Adam [114] for both θ and ϕ parameters. Derivation of the final objective function is presented in Section 3.5.2.

Detailed Dr.VAE architecture is shown in Figure 3.2(b-d). Throughout the model, we used ELU activation function [31] as the non-linearity of our choice.

3.2.3 Perturbation variational autoencoder

We specifically denote the part of Dr.VAE that models drug-induced gene expression perturbations as the Perturbation Variational Autoencoder (PertVAE). PertVAE is an unsupervised model, depicted in Figure 3.2(f), which we use to study the contribution of drug effect modeling on learned latent gene expression representation. We parameterize the PertVAE the same way as analogous parts in Dr.VAE. Detailed derivation of PertVAE is presented in Section 3.5.1.

3.3 Results

We evaluated our drug response prediction method, Dr.VAE, on 26 Food and Drug Administration-approved drug compounds selected from the intersection of two independent in vitro drug screening studies: (i) the CTRPv2 [170] where viability of up to 855 cell lines was measured in response to drug treatment, and (ii) drug-induced transcriptomic perturbations, assayed by NIH LINCS CMap project (CMap-L1000v1) [192], in up to 60 different cell lines for the selected set of drugs.

We compared Dr.VAE to ridge logistic regression (RidgeLR), random forest (RForest) with 100 trees, and support vector machine with a radial basis function kernel (SVMrbf) applied directly to gene expression and also transformed through dimensionality reduction. We used the implementation of these methods as available in the scikit-learn library [158]. For each drug, the best regularization parameter of RidgeLR was found in cross-validation. To assess the impact

of drug-induced perturbations on the drug response prediction task we also compared Dr.VAE to SSVAE [117] where the focus is on classification using solely pre-treatment gene expression. SSVAE does not include any information of drug-induced transcriptomic perturbations. All evaluated models were fit independently to each of the 26 drugs, reusing the same deep learning architecture. We assessed the performance of the classifiers using the area under the ROC curve (AUROC) and the precision recall curve (AUPR) (presented in Appendix A).

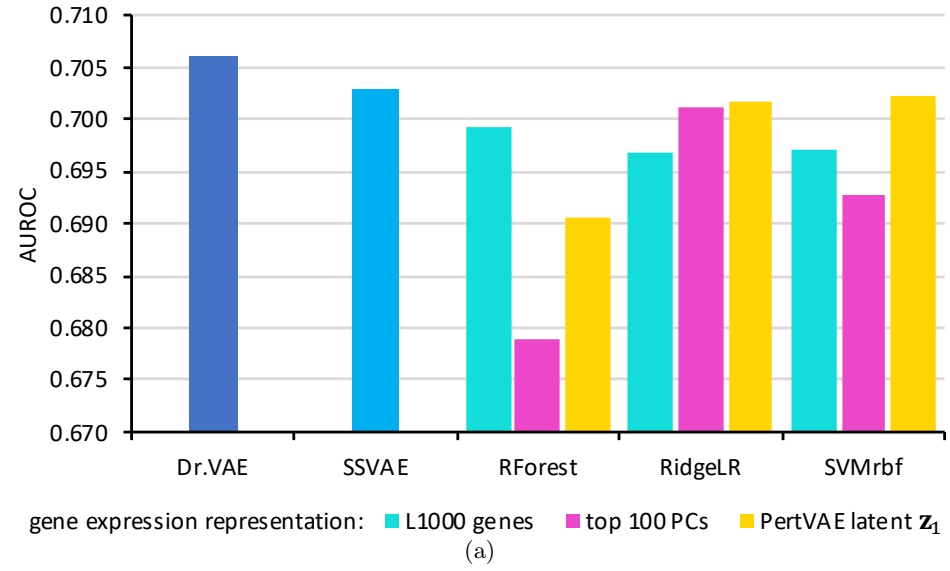
We generated 100 train-validation-test data splits by performing repeated 5-fold cross-validation 20-times. The perturbation data from CMap-L1000v1 were split based on cell line identifiers so that all measurements pertaining to one cell line were assigned to one fold. The CTRPv2 sensitivity data were split such that the ratio of responders/non-responders was approximately equal in each fold, except cell lines that are in the intersection of CTRPv2 and CMap-L1000v1, which were assigned to their corresponding CMap-L1000v1 folds. The CMap-L1000v1 folds were pooled into training and validation splits only, as for some drugs the availability of perturbation experiments was limited to only as few as eight cell lines. Therefore test splits consisted exclusively of data from CTRPv2 that had no known post-treatment gene expression. This way Dr.VAE is most fairly evaluated against methods that cannot model perturbation effects, which is the typical scenario when response prediction has to be made solely based on pre-treatment features. During training of Dr.VAE and SSVAE models, a validation fold was used for early stopping and selection of classification loss weight. All compared methods were trained and evaluated on the same 100 train-validation-test data splits.

3.3.1 Drug response prediction from expression of L1000 genes

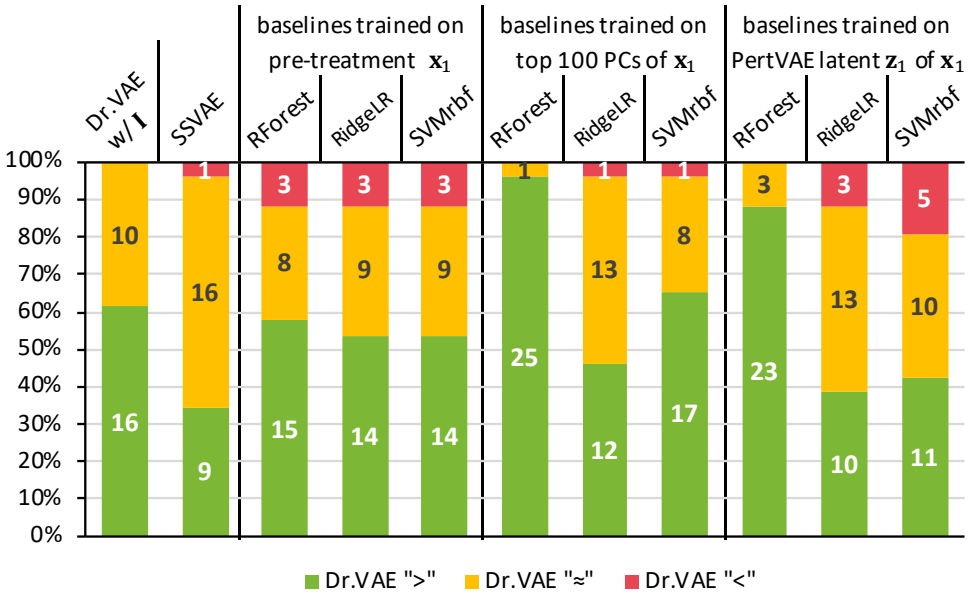
We jointly trained Dr.VAE on both CTRPv2 cell line sensitivity dataset and CMap-L1000v1 6 h-long perturbations and compared the performance to three established baseline classification models. Each model was trained on the expression of 973 genes that form the intersection of genes measured by the L1000 platform in CMap and RNAseq in CTRPv2. For a fair comparison, the baseline classifiers were trained on the very same data splits as Dr.VAE, consisting of CTRPv2 and CMap pre-treatment (control) experiments. Following the random variable notation from our Dr.VAE model, Figure 3.1 and 3.2, these data correspond to $\mathbf{x}_1$.

Dr.VAE outperforms all three baseline classifiers for at least 14 out of 26 (53.8%) tested drugs, and performs with no statistically significant difference on nine drugs. On only 3 out of 26 (11.5%) drugs the baseline models performed better than Dr.VAE, Figure 3.3 and 3.4. The presented comparisons are based on one-sided Wilcoxon Signed-Rank Test (P-value < 0.05) over 100 data splits. Detailed performance of all models applied to each individual drug is presented in Appendix A.1, the corresponding P-values are shown in Appendix A.2. Results in terms of the AUPR follow a similar pattern (Appendix A).

For bortezomib, niclosamide, paclitaxel, decitabine and clofarabine, cancer drugs with no established univariate biomarkers of response, Dr.VAE improved response prediction over every standard classification method by at least 1% and up to 4.4% of AUROC, while AUPR improved



(a)



(b)

Figure 3.3: **Summarized classification results.** (a) AUROC of Dr.VAE and baseline methods. Shown is average over 26 drugs, each evaluated in 100 train-validation-test splits. (b) Dr.VAE is comparable or better than any other baseline for >80% of the drugs (P-value <0.05 Wilcoxon test).

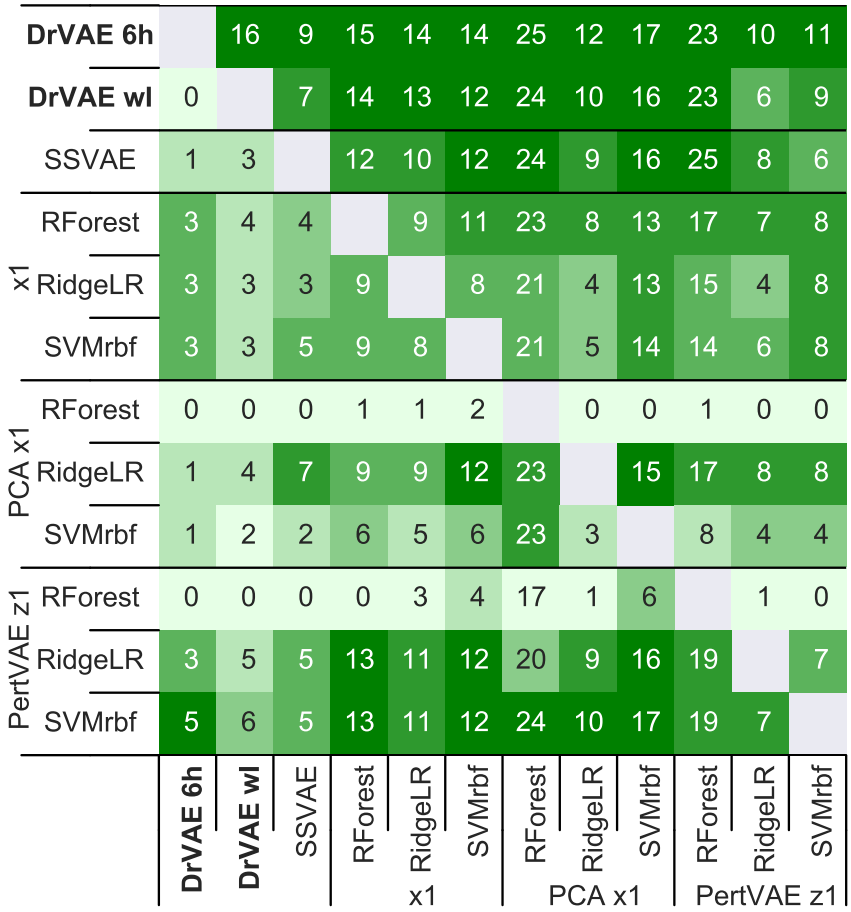


Figure 3.4: **All to all comparison of tested methods.** For each method, there is a row showing the count of 26 drugs for which this method significantly outperforms the other methods corresponding to individual columns. The comparison is based on test AUROC performance in 100 train-validation-test splits. Statistical significance of observed differences in test performance for any two methods was tested by one-sided Wilcoxon Signed-Rank Test (P-value <0.05). The heatmap color is normalized within each column, emphasizing methods that are the best contenders compared to the method corresponding to that column.

by at least 0.7% and up to 4.8%. We have observed the best improvement over RidgeLR for mitomycin and sirolimus with 5.7% and 3.9% AUROC improvement, respectively. Sirolimus inhibits the activation of a key regulatory kinase, the mammalian Target Of Rapamycin (mTOR). As showed in [154], perturbation effects induced by PI3K/Akt/mTOR kinases are typically cell-type specific, which possibly hampers response prediction for these drugs. In this case, Dr.VAE was able to better stratify the response classes, improving the response prediction, particularly over RidgeLR and SVMrbf. Mitomycin, an antibiotic that causes cross-linking of DNA and inhibition of DNA synthesis, is used as a chemotherapy drug in the treatment of various malignant neoplasms. Prediction of sensitivity to mitomycin treatment appears to benefit from employing non-linear prediction models such as RForest and SVMrbf. Dr.VAE can model non-linear relationships and performs on par with the RForest and SVMrbf, considerably outperforming RidgeLR.

Contrarily, in the case of fluvastatin and bosutinib, Dr.VAE trails RidgeLR by 1.5% and 0.9% in test AUROC, repsectively. Fluvastatin belongs to a class of drugs called statins. Statin inhibitors are used to control hypercholesterolemia but have been indicated to have a potential as anticancer agents as well. Sensitivity to statins is highly dependent on strength of a feedback mechanism, the activation of which has been reported to peak at time points >8 h post-treatment [30]. Modeling of 6 h-long perturbations is insufficient in this case and as such Dr.VAE did not improve sensitivity prediction. Reduced performance of Dr.VAE in the case of bosutinib is likely due to modeling of perturbations at only the most common drug concentration level. Bosutinib is a tyrosine kinase inhibitor, used in chronic myelogenous leukemia therapy, primarily targeting Bcr-Abl kinase. [154] observed that such inhibitors of extracellular matrix receptors and receptor tyrosine kinases, exhibited considerably more variance in perturbation signatures with changing drug dose than other drugs. Since we selected perturbation experiments at only one drug concentration level, that with largest number of experiments, it is possible that modeling perturbation effects at only this one concentration level is not sufficiently informing the treatment sensitivity prediction.

3.3.2 Perturbation experiments improve drug response prediction

We investigated the contribution of drug perturbation experiments to response classification via two ablation studies. First, we compared Dr.VAE to semi-supervised VAE [117]. SSVAE was fit to the pre-treatment gene expression in cell lines from CMap-L1000v1 and CTRPv2 without observing post-treatment gene expression and without modeling the drug effects. Since SSVAE is conceptually a subset of Dr.VAE's architecture, we used the same hyperparameters for the corresponding encoders/decoders as in Dr.VAE, Figure 3.2(e). SSVAE outperforms baseline methods according to AUROC but is not as good as Dr.VAE. Dr.VAE achieves significantly better test AUROC than SSVAE on 9 out of 26 (34.6%) drugs (P-value <0.05) with no statistically significant difference on 16 drugs (61.5%) and only for one drug (vincristine) SSVAE outperforms Dr.VAE, Figure 3.3.

To evaluate the contribution of the perturbation function to the classification performance, we modified each trained Dr.VAE instance by replacing the learned drug perturbation function with an identity function (denoted as 'Dr.VAE w/I') without retraining the rest of the model. The modified 'Dr.VAE w/I' achieves AUROC close to Dr.VAE, however slightly worse in absolute value over the 26 drugs. For 16 drugs Dr.VAE has significantly better performance than Dr.VAE w/I and for 10 drugs there was no significant difference, showing that while functions more complex than identity may be able to learn from the perturbation data, more drug perturbation data are required to substantially improve response prediction for many drugs.

Our results show that Dr.VAE improves drug response classification performance thanks to modeling of drug perturbation pairs. As our second set of experiments show, the learned perturbation function contributes to better classification. However, most of the observed improvement appears to stem from more informative latent gene expression representation, that, compared to SSVAE, is learned by joint modeling of drug perturbations as well as sensitivity response. The superior performance of Dr.VAE w/I compared to SSVAE is a testament to that effect.

3.3.3 The importance of dimensionality reduction

Dr.VAE and SSVAE learn a lower dimensional latent representation of the data and the classifier jointly. To understand the importance of the joint optimization, we also explored a learning paradigm where we first optimize the latent representation in an unsupervised fashion and only then train a classifier using the already learned embedding. To this end we performed two sets of experiments. First, we evaluated dimensionality reduction by PCA. PCA projects the data into a space given by orthogonal vectors called principal components that are selected in the order of largest possible variance they account for in the data. We chose to represent the CTRPv2 and CMap-L1000v1 pre-treatment gene expression of L1000 genes in terms of their first 100 principal components that we estimated on each training data fold. Second, we trained just the perturbation part of Dr.VAE, which we denote as PertVAE, to assess dimensionality reduction using a deep generative model. PertVAE is an unsupervised model that does not model drug response outcomes. Instead it learns to model drug perturbation effects from the perturbation pairs, Figure 3.2(f). We then used the mean of the 100-dimensional latent embedding $\mathbf{z}_1$ of the pre-treatment gene expression as the reduced representation for subsequent fitting of standard classifiers.

Both PCA and PertVAE were fit on each training data fold and the learned projections then applied to test data fold. We used the same 100 train-validation-test splits as in the previous experiments, thus the classification test results can be mutually compared by Wilcoxon Signed-Rank Test with the above mentioned Dr.VAE and multiple baseline results, Figure 3.3(b) and Figure 3.4. In terms of mean AUROC, Figure 3.3(a), and mean AUPR, Appendix A.1, all three standard classifiers perform better when fit on the PertVAE embedding $\mathbf{z}_1$ than when fit on the PCA projection onto the first 100 principal components. In the case of both of these reduced

representations, notable is the improvement of the RidgeLR classifier that performs better than when trained directly on expression of the L1000 genes. These two methods, together with SVMrbf trained on the PertVAE $\mathbf{z}_1$ embedding, achieve the most competitive results, nearly equal to SSVAE. However, our Dr.VAE model that combines PertVAE and a drug response classifier in an end-to-end fashion delivers the best overall classification performance, accomplishing statistically better or equivalent AUROC for at least 21 out of 26 drugs (80.8%) than any other evaluated method.

3.3.4 Modeling of drug perturbation effects

We have shown that Dr.VAE can distill useful information from drug perturbation experiments to improve cell line response classification. We seek to investigate how well Dr.VAE model can predict the actual post-treatment gene expression levels. In the following set of experiments we assessed how well Dr.VAE can predict the post-treatment expression in the latent space, corresponding to random variable $\mathbf{z}_2$, as well as in the gene space, which corresponds to $\mathbf{x}_2$. Particularly, we computed the expected root mean square error (RMSE) of Dr.VAE predictions over $\mathbf{z}_2$ and $\mathbf{x}_2$ when computed from pre-treatment $\mathbf{x}_1$ compared to the expected embedding $\mathbf{z}_2$ computed from post-treatment $\mathbf{x}_2$ and the true observed $\mathbf{x}_2$, respectively. Furthermore, we compared how the RMSE of Dr.VAE predictions improved over the 'Dr.VAE w/I' baseline model where we replaced the learned perturbation function by an identity function (as introduced previously). On training data, Dr.VAE predicted the mean of $\mathbf{z}_2$ with RMSE 10.5% lower compared to Dr.VAE w/I, yet on validation data it was 9.6% worse on average across all 26 drugs. This result shows that Dr.VAE, while being primarily optimized for drug response classification, learns to partially model drug perturbation effects, but on average, suffers from data limitations and overfitting.

To elucidate the connection between Dr.VAE performance and limitations of available perturbation experiments, we computed the correlation of Dr.VAE $\mathbf{z}_2$ prediction improvement over Dr.VAE w/I across the set of 26 drugs with three data statistics: (i) effect-to-replicate variance ratio (ERVR) in CMap-L1000v1 perturbation experiments, (ii) number of unique cell lines tested for a given drug in CMap-L1000v1 and (iii) the product of the previous two. The computed Pearson correlations are shown in Table 3.1. The ability of Dr.VAE to generalize from the training to validation set correlates with both the strength of the perturbation signal in the data (quantified as ERVR) and the dataset size, yet the strongest is correlation with the product of these two variables, $\rho = 0.814$ (P-value 4.35×10^{-7}). The computation of ERVR measure is described in Section 3.5.3.

For prediction of post-treatment gene expression $\mathbf{x}_2$ we observed an analogous conclusion to prediction of its latent representation $\mathbf{z}_2$. The detailed results are shown in Appendix A.7. We conclude that there are presently data limitations (number and noise/signal resolution of drug perturbation experiments) for generalizable post-treatment gene expression prediction yet, as shown above, we can still distill information that improves drug response classification.

Table 3.1: The ability of Dr.VAE to model post-treatment gene expression correlates with signal/noise ratio and quantity of perturbation experiments. We computed Δ RMSE improvement of Dr.VAE in post-treatment expression prediction over Dr.VAE w/I, averaged over validation data splits, and correlated it to overall CMap-L1000v1 dataset statistics. The Pearson correlation was computed for prediction Δ improvement of both post-treatment gene expression $\mathbf{x}_2$ and its latent representation $\mathbf{z}_2$. Additionally we include correlation with difference in Dr.VAE and SSVAE classification performance.

Δ RMSE evaluated on	Dataset property correlated to	ρ	P-value
$\mathbf{z}_2$	Effect/rep. variance ratio (ERVR)	0.66	2.4×10^{-4}
$\mathbf{x}_2$	ERVR	0.72	4.0×10^{-5}
$\mathbf{z}_2$	Num. unique CLs in CMap (NCL)	0.71	4.2×10^{-5}
$\mathbf{x}_2$	NCL	0.52	6.4×10^{-3}
$\mathbf{z}_2$	ERVR * NCL	0.81	4.4×10^{-7}
$\mathbf{x}_2$	ERVR * NCL	0.73	2.6×10^{-5}
$\mathbf{x}_2$	Dr.VAE - SSVAE [AUROC]	0.29	0.15
$\mathbf{x}_2$	Dr.VAE - SSVAE [AUPR]	0.20	0.33

Lastly, we investigated whether there is a correlation between classification performance improvement of Dr.VAE over SSVAE, which does not model perturbation effects, and the ability of Dr.VAE to generalize post-treatment gene expression prediction to validation set. We found weak correlation between the classification improvement in terms of both AUROC (Pearson $\rho = 0.293$; P-value 0.147), and AUPR (Pearson $\rho = 0.199$; P-value 0.329). These results suggest that Dr.VAE tends to improve over SSVAE for the drugs Dr.VAE manages to model the transcriptomic perturbations induced by the drug compound.

3.4 Discussion

We developed Dr.VAE, the first unified machine learning method for drug response prediction that enables semi-supervised learning and successfully leverages prior information in the form of drug-induced transcriptomic perturbations. Our approach follows several previously identified trends for improved drug response prediction [34], as we can model non-linearities in the data and incorporate prior knowledge.

Typical discriminative feedforward neural networks do not fare well in drug response prediction, most likely because of the data limitation (number of features versus number of samples). We showed that joint generative modeling of drug response and perturbation effects alleviates this to a significant extent, possibly acting as an effective regularization and robust feature extraction that does not overfit the way discriminative neural networks do.

We tested 26 Food and Drug Administration-approved drug compounds for which both perturbation and drug response experimental data were available. Our experiments showed that

for those drugs that have sufficient data to capture the variation and effect on gene expression, incorporating those effects yields a significant improvement over logistic regression, random forest and support vector machines. Dr.VAE significantly outperformed these models in more than half of the tested drugs and performed on par in other cases. Through a series of experiments, we showed that the observed improvement of Dr.VAE in drug response prediction can be credited to its joint modeling of both response and drug-induced perturbation effects.

Our study has several potential limitations. First, we considered only the gene expression modality, as it has been consistently shown to provide the most predictive power in multiple previous studies on drug response [104, 34]. There is accumulating evidence, however, that multi-omic predictors that additionally integrate methylation, copy number variation, mutational status or proteomic data can achieve improved prediction performance. It is relatively straightforward to extend Dr.VAE, thanks to the stochastic variational inference approach we adopted. Categorical or Poisson likelihood functions can be used to model discrete (mutational status) or count (CNVs) data types, respectively, in addition to the Gaussian likelihood we used to model continuous gene expression. We caution however, that inclusion of additional features accentuates the already unfavorable ratio of the number of features to the number of available training examples, which could prove, and indeed has been, problematic for any method, including ours.

Second, we modeled CMap-L1000v1 perturbations after 6 h of treatment duration at the most common concentration level for each drug. That allowed us to pool the largest possible number of experiments tested under consistent experimental settings. It can be argued that 6 h is too short for many feedback regulatory mechanisms to manifest themselves and as such these experiments alone do not provide complete picture of the transcriptomic response. Notably, drug-cell line viability assays are typically done with longer treatment duration, such as 72 h. This is the case for a statin inhibitor fluvastatin, as we observed in out experiments. Thus we also trained our Dr.VAE with 24 h perturbation experiments, however, potentially because of the limited number of such experiments, this did not improve our prediction performance. A potential future improvement to our method could be an extension which leverages all available perturbation experiments of various durations and drug concentrations.

Every conditional distribution that Dr.VAE is composed of is parameterized by a neural network. The ability to adjust hyperparameters to match complexity of the data makes Dr.VAE a very flexible model. Since we opted for simplicity, most of our neural networks have one hidden layer, while the classification posterior and perturbation function are linear. As more data become available we will be able to take full advantage of the new methodological developments in the generative deep learning field, further improving the performance of Dr.VAE and other drug response prediction models. However so far our attempts to use deeper networks or utilize normalizing flows to approximate posteriors by more complex distributions [173, 118] have not significantly improved the performance to justify the added complexity.

In conclusion, we have demonstrated deep generative modeling to be a promising methodological approach for method development in the field of drug response prediction. In

particular, this approach has two major benefits. First, the flexibility of this paradigm allowed us to integrate transcriptional perturbation effects into the drug response prediction framework in a unique way. Second, all conditional distributions that form our Dr.VAE model, as well as variational posteriors used for approximate inference in Dr.VAE, are parameterized by neural networks that can model complex non-linear relationships. We have shown that both aspects compounded in our Dr.VAE, which outperformed the most used methods in the field for the majority of the evaluated drug compounds.

Processed data and software implementation using PyTorch [157] are available at: https://github.com/rampasek/DrVAE.

3.5 Supplementary Methods and Results

3.5.1 Perturbation variational autoencoder

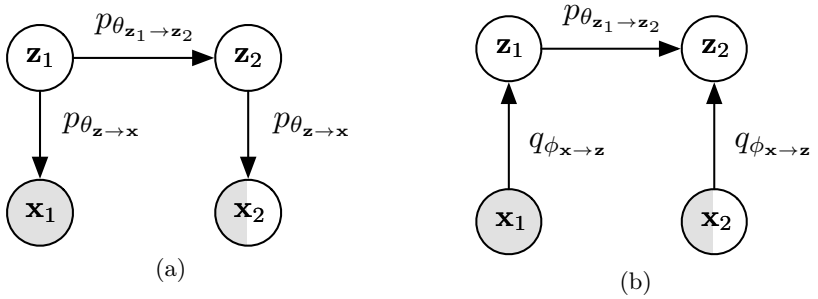


Figure 3.5: **Perturbation VAE.** (a) Factorization of the generative distribution p, (b) Factorization of the approximate posterior distribution q. Note, we use the generative $p_{\theta_{\mathbf{z}_1 \to \mathbf{z}_2}}$ in case $\mathbf{x}_2$ is not observed.

Perturbation Variational Autoencoder (PertVAE) is an unsupervised model for drug-induced gene expression perturbations, that embeds the data space (gene expression) in a lower dimensional latent space. In the latent space we model the drug-induced effect as a linear function, which is trained jointly with the embedding encoder and decoder.

We fit PertVAE on "perturbation pairs" $[\mathbf{x}_1, \mathbf{x}_2]$ of pre-treatment and post-treatment gene expression with shared stochastic embedding encoder $q_{\phi_{\mathbf{x} \to \mathbf{z}}}$ and decoder $p_{\theta_{\mathbf{z} \to \mathbf{x}}}$. The original dimension of each vector $\mathbf{x}$ is 973 landmark genes. Additionally we use unpaired pre-treatment data (with no know post-treatment state) to improve learning of the latent representation. The graphical representation of PertVAE model is shown in Figure 3.5.

Joint distribution. PertVAE models joint $p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2)$, which is assumed to factorize as:

$$p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2) = p(\mathbf{x}_1|\mathbf{z}_1) \cdot p(\mathbf{x}_2|\mathbf{z}_2) \cdot p(\mathbf{z}_2|\mathbf{z}_1) \cdot p(\mathbf{z}_1) \tag{3.6}$$

Generative distribution p_θ. PertVAE's generative process is as follows:

$$p(\mathbf{z}_1) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3.7}$$

$$p_{\theta_{\mathbf{z}_1 \to \mathbf{z}_2}}(\mathbf{z}_2|\mathbf{z}_1) = \mathcal{N}\left(\mathbf{z}_2|\boldsymbol{\mu}_{\mathbf{z}_2} = f_\theta(\mathbf{z}_1), \boldsymbol{\sigma}^2_{\mathbf{z}_2} = \exp^{f_\theta(\mathbf{z}_1)}\right) \tag{3.8}$$

$$k \in \{1, 2\} : p_{\theta_{\mathbf{z} \to \mathbf{x}}}(\mathbf{x}_k|\mathbf{z}_k) = \mathcal{N}\left(\mathbf{x}_k|\boldsymbol{\mu}_{\mathbf{x}_k} = f_\theta(\mathbf{z}_k), \boldsymbol{\sigma}^2_{\mathbf{x}_k} = \exp^{f_\theta(\mathbf{z}_k)}\right) \tag{3.9}$$

The parameters of these distributions are computed by functions f_θ, which are neural networks with a total set of parameters θ. For brevity we refer to these parameters as θ instead of more specific subsets $\theta_{\mathbf{z} \to \mathbf{x}}$ or $\theta_{\mathbf{z}_1 \to \mathbf{z}_2}$ when such level of detail unnecessarily clutters the notation.

We constrain the mean function in $p_{\theta_{\mathbf{z}_1 \to \mathbf{z}_2}}$ to be a linear function $f_{\theta_{\mathbf{z}_1 \to \mathbf{z}_2}}(\mathbf{z}_1)$ of the following

form:

$$f_{\theta_{\mathbf{z}_1 \to \mathbf{z}_2}}(\mathbf{z}_1) \equiv \mathbf{z}_1 + \mathbf{W}\mathbf{z}_1 + \mathbf{b} \tag{3.10}$$

with $\mathbf{W}$ and $\mathbf{b}$ initialized close to zero such that $f_{\theta_{\mathbf{z}_1 \to \mathbf{z}_2}}(\mathbf{z}_1)$ starts as an identity function. We found that together with L2 penalization this formulation improves stability and generalization of the model.

Approximate posterior q_ϕ. Depending on the type of the data, we assume the approximate posterior q with a set of parameters ϕ factorizes as:

$$\text{perturbation pairs:} \quad q_\phi(\mathbf{z}_1, \mathbf{z}_2 | \mathbf{x}_1, \mathbf{x}_2) = q_{\phi_{\mathbf{x} \to \mathbf{z}}}(\mathbf{z}_1|\mathbf{x}_1) \cdot q_{\phi_{\mathbf{x} \to \mathbf{z}}}(\mathbf{z}_2|\mathbf{x}_2) \tag{3.11}$$

$$\text{pre-treatment singleton:} \quad q_\phi(\mathbf{z}_1, \mathbf{z}_2, \mathbf{x}_2 | \mathbf{x}_1) = q_{\phi_{\mathbf{x} \to \mathbf{z}}}(\mathbf{z}_1|\mathbf{x}_1) \cdot p_{\theta_{\mathbf{z}_1 \to \mathbf{z}_2}}(\mathbf{z}_2|\mathbf{z}_1) \cdot p_{\theta_{\mathbf{z} \to \mathbf{x}}}(\mathbf{x}_2|\mathbf{z}_2) \tag{3.12}$$

Analogously to the shared generative $p_{\theta_{\mathbf{z} \to \mathbf{x}}}$ distribution, also $q_{\phi_{\mathbf{x} \to \mathbf{z}}}(\mathbf{z}_k|\mathbf{x}_k)$ is shared for both $k \in \{1, 2\}$ and takes from of a diagonal Gaussian:

$$k \in \{1, 2\} : q_{\phi_{\mathbf{x} \to \mathbf{z}}}(\mathbf{z}_k|\mathbf{x}_k) = \mathcal{N}\left(\mathbf{z}_k | \boldsymbol{\mu}_{\mathbf{z}_k} = f_\phi(\mathbf{x}_k), \boldsymbol{\sigma}^2_{\mathbf{z}_k} = \exp^{f_\phi(\mathbf{x}_k)}\right) \tag{3.13}$$

Fitting θ and ϕ parameters. We jointly optimize the generative model θ and variational ϕ parameters to maximize evidence lower bound, $\text{ELBO}_{\text{PertVAE}}$, of training data. The training data consists of a set of perturbation pairs P and unpaired "singleton" examples S that we leverage to train the latent space variational autoencoder as well.

$$\sum_{(\mathbf{x}_1, \mathbf{x}_2) \in P} \log p\left(\mathbf{x}_1, \mathbf{x}_2\right) + \sum_{\mathbf{x}_1 \in S} \log p\left(\mathbf{x}_1\right) \geq \text{ELBO}_{\text{PertVAE}} \tag{3.14}$$

$$\text{ELBO}_{\text{PertVAE}} = \sum_{(\mathbf{x}_1, \mathbf{x}_2) \in P} \mathcal{L}_P\left(\mathbf{x}_1, \mathbf{x}_2; \theta, \phi\right) + \sum_{\mathbf{x}_1 \in S} \mathcal{L}_S\left(\mathbf{x}_1; \theta, \phi\right) \tag{3.15}$$

The individual per-example lower bounds $\mathcal{L}_P$ and $\mathcal{L}_S$ take the following form:

$$\mathcal{L}_P(\mathbf{x}_1, \mathbf{x}_2;\ \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}_1, \mathbf{z}_2 | \mathbf{x}_1, \mathbf{x}_2)}\big[\log p_\theta(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2) - \log q_\phi(\mathbf{z}_1, \mathbf{z}_2 | \mathbf{x}_1, \mathbf{x}_2)\big] \tag{3.16}$$

$$\begin{aligned} &= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[\log p_\theta(\mathbf{x}_1|\mathbf{z}_1) - D_{KL}\left[q_\phi(\mathbf{z}_2|\mathbf{x}_2)||p_\theta(\mathbf{z}_2|\mathbf{z}_1)\right]\big] \\ &\quad + \mathbb{E}_{q_\phi(\mathbf{z}_2|\mathbf{x}_2)}\left[\log p_\theta(\mathbf{x}_2|\mathbf{z}_2)\right] \\ &\quad - D_{KL}\left[q_\phi(\mathbf{z}_1|\mathbf{x}_1)||p(\mathbf{z}_1)\right] \end{aligned} \tag{3.17}$$

$$\mathcal{L}_S(\mathbf{x}_1;\ \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[\log p_\theta(\mathbf{x}_1, \mathbf{z}_1) - \log q_\phi(\mathbf{z}_1|\mathbf{x}_1)\big] \tag{3.18}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\left[\log p_\theta(\mathbf{x}_1|\mathbf{z}_1)\right] - D_{KL}\left[q_\phi(\mathbf{z}_1|\mathbf{x}_1)||p(\mathbf{z}_1)\right]$$

The expectations that are part of our evidence lower bounds are evaluated approximately, by Monte Carlo sampling. In practice we use two MC samples. Thanks to so-called

reparameterization trick it is possible to backpropagate through such approximation of the expectations, yielding an unbiased gradient estimator of the distribution parameters known as Stochastic Gradient Variational Bayes (SGVB) [115]. In our case, no approximation is necessary for evaluation of the Kullback-Leibler divergences present in the ELBO, as there is a close form solution for D_{KL} between two normal distributions. Now we have all the tools to evaluate $\text{ELBO}_{\text{PertVAE}}$ and compute its approximate gradients w.r.t. both θ and ϕ parameters. We then use Adam [114] to compute the parameter updates during the optimization process.

3.5.2 Drug response variational autoencoder

Analogously to semi-supervised variational autoencoder, we extended the unsupervised Perturbation VAE to a semi-supervised model by incorporating a modified "M2 model" [117]. The extended model, drug response variational autoencoder (Dr.VAE), enables us to model both drug-induced perturbation effects as well as treatment response outcome at the same time. We train Dr.VAE jointly and not as a stack of two models PertVAE + M2 model, similarly to how semi-supervised VAE can be trained jointly [139].

We use similar type of data to train Dr.VAE as we use for PertVAE, however some of the perturbation pairs and pre-treatment singletons now can have a binary outcome label $\mathbf{y}$ associated with them, denoting if the drug treatment was successful or not. Schema of Dr.VAE model is shown in main text, Fig 1.

Joint distribution. Drug response VAE extends PertVAE to model a joint distribution $p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y})$ factorized as:

$$p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}) = p(\mathbf{x}_1|\mathbf{z}_1) \cdot p(\mathbf{x}_2|\mathbf{z}_2) \cdot p(\mathbf{z}_2|\mathbf{z}_1) \cdot p(\mathbf{z}_1|\mathbf{z}_3, \mathbf{y}) \cdot p(\mathbf{z}_3) \cdot p(\mathbf{y}) \tag{3.19}$$

Generative distributions p_θ. The individual generative distributions, Dr.VAE factorizes to, have the following form:

$$p(\mathbf{y}) = \text{Cat}(\mathbf{y}|\boldsymbol{\pi} = \text{Unif}) \tag{3.20}$$

$$p(\mathbf{z}_3) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3.21}$$

$$p_\theta(\mathbf{z}_1|\mathbf{z}_3, \mathbf{y}) = \mathcal{N}\left(\mathbf{z}_1|\boldsymbol{\mu}_{\mathbf{z}_1} = f_\theta(\mathbf{z}_3, \mathbf{y}), \boldsymbol{\sigma}^2_{\mathbf{z}_1} = \exp^{f_\theta(\mathbf{z}_3, \mathbf{y})}\right) \tag{3.22}$$

$$p_\theta(\mathbf{z}_2|\mathbf{z}_1) = \mathcal{N}\left(\mathbf{z}_2|\boldsymbol{\mu}_{\mathbf{z}_2} = f_\theta(\mathbf{z}_1), \boldsymbol{\sigma}^2_{\mathbf{z}_2} = \exp^{f_\theta(\mathbf{z}_1)}\right) \tag{3.23}$$

$$k \in \{1, 2\} : p_\theta(\mathbf{x}_k|\mathbf{z}_k) = \mathcal{N}\left(\mathbf{x}_k|\boldsymbol{\mu}_{\mathbf{x}_k} = f_\theta(\mathbf{z}_k), \boldsymbol{\sigma}^2_{\mathbf{x}_k} = \exp^{f_\theta(\mathbf{z}_k)}\right) \tag{3.24}$$

Same way as in PertVAE, we share the "data decoder" $p_\theta(\mathbf{x}_k|\mathbf{z}_k)$ among both $k \in \{1, 2\}$.

Approximate posterior q_ϕ. Depending on the type of the data, we assume the approximate posterior q to factorize as:

$$\text{labeled pair: } q_\phi(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3|\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) = q_\phi(\mathbf{z}_1|\mathbf{x}_1) \cdot q_\phi(\mathbf{z}_2|\mathbf{x}_2) \cdot q_\phi(\mathbf{z}_3|\mathbf{z}_1, \mathbf{y}) \tag{3.25}$$

$$\text{unlabeled pair: } q_\phi(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}|\mathbf{x}_1, \mathbf{x}_2) = q_\phi(\mathbf{z}_1|\mathbf{x}_1) \cdot q_\phi(\mathbf{z}_2|\mathbf{x}_2) \cdot q_\phi(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2) \cdot q_\phi(\mathbf{z}_3|\mathbf{z}_1, \mathbf{y}) \tag{3.26}$$

$$\text{labeled singleton: } q_\phi(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{x}_2|\mathbf{x}_1, \mathbf{y}) = q_\phi(\mathbf{z}_1|\mathbf{x}_1) \cdot p_\theta(\mathbf{z}_2|\mathbf{z}_1) \cdot p_\theta(\mathbf{x}_2|\mathbf{z}_2) \cdot q_\phi(\mathbf{z}_3|\mathbf{z}_1, \mathbf{y}) \tag{3.27}$$

$$\begin{aligned}\text{unlab. singleton: } q_\phi(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{x}_2, \mathbf{y}|\mathbf{x}_1) = q_\phi(\mathbf{z}_1|\mathbf{x}_1) \cdot p_\theta(\mathbf{z}_2|\mathbf{z}_1) \cdot p_\theta(\mathbf{x}_2|\mathbf{z}_2) \cdot \\ \cdot q_\phi(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2) \cdot q_\phi(\mathbf{z}_3|\mathbf{z}_1, \mathbf{y})\end{aligned} \tag{3.28}$$

The "data encoder" $k \in \{1, 2\} : q_\phi(\mathbf{z}_k|\mathbf{x}_k)$ is shared and parameterized the same way as in PertVAE. The additional approximate posterior distributions then take the following form:

$$q_\phi(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2) = \text{Cat}\big(\mathbf{y}|\boldsymbol{\pi} = \text{softmax}(f_\phi(\mathbf{z}_1, \mathbf{z}_2 - \mathbf{z}_1))\big) \tag{3.29}$$

$$q_\phi(\mathbf{z}_3|\mathbf{z}_1, \mathbf{y}) = \mathcal{N}\left(\mathbf{z}_3|\boldsymbol{\mu}_{\mathbf{z}_3} = f_\phi(\mathbf{z}_1, \mathbf{y}), \boldsymbol{\sigma}^2_{\mathbf{z}_3} = \exp^{f_\phi(\mathbf{z}_1, \mathbf{y})}\right) \tag{3.30}$$

The afford mentioned factorizations of the joint and of the posteriors also provide a recipe for sampling and inference in the model by Monte Carlo sampling.

Fitting θ and ϕ parameters. We have 4 different sets of partially observed variables, which correspond to different types of data. Therefore there are 4 different evidence lower bounds to optimize:

$$\text{labeled perturbation pairs } LP\text{:} \quad \sum \mathcal{L}_{LP}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}; \theta, \phi) \tag{3.31}$$

$$\text{unlabeled perturbation pairs } UP\text{:} \quad \sum \mathcal{L}_{UP}(\mathbf{x}_1, \mathbf{x}_2; \theta, \phi) \tag{3.32}$$

$$\text{labeled pre-treatment singletons } LS\text{:} \quad \sum \mathcal{L}_{LS}(\mathbf{x}_1, \mathbf{y}; \theta, \phi) \tag{3.33}$$

$$\text{unlabeled pre-treatment singletons } US\text{:} \quad \sum \mathcal{L}_{US}(\mathbf{x}_1; \theta, \phi) \tag{3.34}$$

The sum of these 4 specific evidence lower bounds, $\text{ELBO}_{\text{DrVAE}}$, is the evidence lower bound we need to maximize. The derivation of these specific lower bounds follows the same principles as shown above for PertVAE and as shown in the derivation of semi-supervised VAE [117, 139]. Particularly, for unlabeled and labeled perturbation pairs, denoted UP and LP respectively, we

obtain the following bounds:

$$\mathcal{L}_{UP}(\mathbf{x}_1, \mathbf{x}_2;\ \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3,\mathbf{y}|\mathbf{x}_1,\mathbf{x}_2)}\big[\log p_\theta(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}) - \\ - \log q_\phi(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}|\mathbf{x}_1, \mathbf{x}_2)\big] \tag{3.35}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[\log p_\theta(\mathbf{x}_1|\mathbf{z}_1) - D_{KL}\left[q_\phi(\mathbf{z}_2|\mathbf{x}_2)||p_\theta(\mathbf{z}_2|\mathbf{z}_1)\right]\big] \tag{3.36}$$
$$+ \mathbb{E}_{q_\phi(\mathbf{z}_2|\mathbf{x}_2)}\left[\log p_\theta(\mathbf{x}_2|\mathbf{z}_2)\right]$$
$$+ \mathbb{E}_{q_\phi(\mathbf{y}|\mathbf{z}_1,\mathbf{z}_2)q_\phi(\mathbf{z}_2|\mathbf{x}_2)q_\phi(\mathbf{z}_3|\mathbf{z}_1,\mathbf{y})}\big[-D_{KL}\left[q_\phi(\mathbf{z}_1|\mathbf{x}_1)||p_\theta(\mathbf{z}_1|\mathbf{z}_3,\mathbf{y})\right]\big]$$
$$+ \mathbb{E}_{q_\phi(\mathbf{y}|\mathbf{z}_1,\mathbf{z}_2)q_\phi(\mathbf{z}_1|\mathbf{x}_1)q_\phi(\mathbf{z}_2|\mathbf{x}_2)}\big[-D_{KL}\left[q_\phi(\mathbf{z}_3|\mathbf{z}_1,\mathbf{y})||p(\mathbf{z}_3)\right]\big]$$
$$+ \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)q_\phi(\mathbf{z}_2|\mathbf{x}_2)}\big[-D_{KL}\left[q_\phi(\mathbf{y}|\mathbf{z}_1,\mathbf{z}_2)||p(\mathbf{y})\right]\big]$$

$$\mathcal{L}_{LP}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y};\ \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})}\big[\log p_\theta(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}) - \\ - \log q_\phi(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3|\mathbf{x}_1, \mathbf{x}_2, \mathbf{y})\big] \tag{3.37}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[\log p_\theta(\mathbf{x}_1|\mathbf{z}_1) - D_{KL}\left[q_\phi(\mathbf{z}_2|\mathbf{x}_2)||p_\theta(\mathbf{z}_2|\mathbf{z}_1)\right]\big] \tag{3.38}$$
$$+ \mathbb{E}_{q_\phi(\mathbf{z}_2|\mathbf{x}_2)}\left[\log p_\theta(\mathbf{x}_2|\mathbf{z}_2)\right]$$
$$+ \mathbb{E}_{q_\phi(\mathbf{z}_3|\mathbf{z}_1,\mathbf{y})}\big[-D_{KL}\left[q_\phi(\mathbf{z}_1|\mathbf{x}_1)||p_\theta(\mathbf{z}_1|\mathbf{z}_3,\mathbf{y})\right]\big]$$
$$+ \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[-D_{KL}\left[q_\phi(\mathbf{z}_3|\mathbf{z}_1,\mathbf{y})||p(\mathbf{z}_3)\right]\big]$$
$$+ \log p(\mathbf{y})$$

Evidence lower bounds for unlabeled singletons (US) and labeled singletons (LS):

$$\mathcal{L}_{US}(\mathbf{x}_1;\ \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3,\mathbf{y}|\mathbf{x}_1)}\big[\log p_\theta(\mathbf{x}_1, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}) - \\ - \log q_\phi(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}|\mathbf{x}_1)\big] \tag{3.39}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}[\log p_\theta(\mathbf{x}_1|\mathbf{z}_1)] \tag{3.40}$$
$$+ \mathbb{E}_{q_\phi(\mathbf{y}|\mathbf{z}_1,\mathbf{z}_2)p_\theta(\mathbf{z}_2|\mathbf{z}_1)q_\phi(\mathbf{z}_1|\mathbf{x}_1)q_\phi(\mathbf{z}_3|\mathbf{z}_1,\mathbf{y})}\big[-D_{KL}\left[q_\phi(\mathbf{z}_1|\mathbf{x}_1)||p_\theta(\mathbf{z}_1|\mathbf{z}_3,\mathbf{y})\right]\big]$$
$$+ \mathbb{E}_{q_\phi(\mathbf{y}|\mathbf{z}_1,\mathbf{z}_2)q_\phi(\mathbf{z}_1|\mathbf{x}_1)p_\theta(\mathbf{z}_2|\mathbf{z}_1)}\big[-D_{KL}\left[q_\phi(\mathbf{z}_3|\mathbf{z}_1,\mathbf{y})||p(\mathbf{z}_3)\right]\big]$$
$$+ \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)p_\theta(\mathbf{z}_2|\mathbf{z}_1)}\big[-D_{KL}\left[q_\phi(\mathbf{y}|\mathbf{z}_1,\mathbf{z}_2)||p(\mathbf{y})\right]\big]$$

$$\mathcal{L}_{LS}(\mathbf{x}_1, \mathbf{y};\ \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3|\mathbf{x}_1,\mathbf{y})}\big[\log p_\theta(\mathbf{x}_1, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{y}) - \\ - \log q_\phi(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3|\mathbf{x}_1, \mathbf{y})\big] \tag{3.41}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}[\log p_\theta(\mathbf{x}_1|\mathbf{z}_1)] \tag{3.42}$$
$$+ \mathbb{E}_{q_\phi(\mathbf{z}_3|\mathbf{z}_1,\mathbf{y})q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[-D_{KL}\left[q_\phi(\mathbf{z}_1|\mathbf{x}_1)||p_\theta(\mathbf{z}_1|\mathbf{z}_3,\mathbf{y})\right]\big]$$
$$+ \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[-D_{KL}\left[q_\phi(\mathbf{z}_3|\mathbf{z}_1,\mathbf{y})||p(\mathbf{z}_3)\right]\big]$$
$$+ \log p(\mathbf{y})$$

In these lower bounds, we compute expectation w.r.t. $q_\phi(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2)$ exactly by summation as $\mathbf{y}$ is in our case a binary random variable:

$$\mathbb{E}_{q_\phi(\mathbf{y}|\mathbf{z}_1,\mathbf{z}_2)}[f(\mathbf{y}, \cdot)] = \sum_{t\in\{0,1\}} q_\phi(\mathbf{y} = t|\mathbf{z}_1, \mathbf{z}_2) f(\mathbf{y}, \cdot) \tag{3.43}$$

Note, that the Kullback-Leibler divergence of two categorical distributions over $\mathbf{y}$ can be computed analytically. Moreover, to mitigate the problem of overly strong prior causing the optimization to get stuck in bad local optima, we follow [118] and allow "free bits" in D_{KL} of approximate posterior and prior over $\mathbf{y}$ and $\mathbf{z}_3$ variables. Now, using the approach as described previously for PertVAE, we can evaluate $\text{ELBO}_{\text{DrVAE}}$ and its gradients w.r.t. both θ and ϕ parameters.

Next, akin to semi-supervised VAE [117], we need to explicitly introduce loss of the predictive posterior $\log q_\phi(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2)$ in order for it to be trained on labeled data as well. This is required since in the labeled data the random variable $\mathbf{y}$ is an observed variable and therefore the lower bounds $\mathcal{L}_{LP}$ and $\mathcal{L}_{LS}$ are conditioned on $\mathbf{y}$ and do not contribute to training of $q_\phi(\mathbf{y}|\mathbf{z}_1, \mathbf{z}_2)$.

Additionally, we found beneficial to include explicit perturbation prediction loss $\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)p_\theta(\mathbf{z}_2|\mathbf{z}_1)}[\log p_\theta(\mathbf{x}_2|\mathbf{z}_2)]$ in addition to minimization of KL divergence between the approximate posterior and predicted distribution over $\mathbf{z}_2$, $\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)}\big[D_{KL}\left[q_\phi(\mathbf{z}_2|\mathbf{x}_2)||p_\theta(\mathbf{z}_2|\mathbf{z}_1)\right]\big]$, that is a part of $\mathcal{L}_{UP}$ and $\mathcal{L}_{LP}$.

Eventually, the final objective $\mathcal{J}_{\text{DrVAE}}$ we seek to maximize is

$$\begin{aligned}\mathcal{J}_{\text{DrVAE}} = \text{ELBO}_{\text{DrVAE}} &+ \\ &+ \alpha \sum_{LP\cup LS} \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)p_\theta(\mathbf{z}_2|\mathbf{z}_1)}\left[\log q_\phi(\mathbf{y} = \mathbf{t}|\mathbf{z}_1, \mathbf{z}_2)\right] + \\ &+ \beta \sum_{LP\cup UP} \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x}_1)p_\theta(\mathbf{z}_2|\mathbf{z}_1)}\left[\log p_\theta(\mathbf{x}_2|\mathbf{z}_2)\right]\end{aligned} \tag{3.44}$$

where α and β are hyperparameters weighting the classification and perturbation loss, respectively, relative to the $\text{ELBO}_{\text{DrVAE}}$ and are selected based on the classification performance on a validation set distinct from a test set. We found $\beta = 0.05$ to be well suited for all drugs, while α is found for each drug individually by cross-validation in grid search over $\{1, 3, 5, 10\}$.

3.5.3 Supplementary Results

Effect-to-replicate variance ratio in perturbation experiments

The tested set of drugs manifests considerable diversity in the type of drugs (cytotoxic or targeted) and number of available perturbation experiments ranging from 32 to 417 in as few as 8 and up to 60 distinct cell lines. The magnitude of the perturbation effect together with the number of available experiments and variance in biological replicates have paramount impact on how well our Dr.VAE can model these drug perturbations. To quantify this connection, we first need to quantify the drug perturbation effect.

We computed the effect-to-replicate variance ratio (ERVR) for perturbation experiments of each cell line with at least two biological replicates as the ratio of variance between clusters to the variance within clusters where we denote biological replicates of the control (pre-treatment) gene expression as one cluster and the matched post-treatment gene expression replicates as the second cluster. For a drug d with perturbation experiments on a set of cell lines C_d its effect-to-replicate variance ratio is

$$\text{ERVR}_d = \frac{1}{|C_d|} \sum_{c \in C_d} \frac{Var \text{ between } S_c^{pre} \text{ and } S_c^{post}}{Var(S_c^{pre}) + Var(S_c^{post})} \tag{3.45}$$

$$= \frac{1}{|C_d|} \sum_{c \in C_d} \frac{Var(S_c) - Var(S_c^{pre}) - Var(S_c^{post})}{Var(S_c^{pre}) + Var(S_c^{post})} \tag{3.46}$$

where $S_c = S_c^{pre} \cup S_c^{post}$ is the set of gene expression measurements of cell line c composed of replicates of its pre- and post-treatment experiments for drug d, S_c^{pre} and S_c^{post}, respectively.

The computed ERVR for each drug is listed in Supplementary Table A.6. We analyzed the impact of ERVR in the results section of the main paper.

Reconstruction of gene expression from latent representation

Dr.VAE learns a non-linear representation of expression of the 1000 landmark genes in a reduced 100-dimensional latent space. In the above we studied how this embedding fares in drug response classification and drug perturbation prediction. Lastly, we evaluated how well Dr.VAE can reconstruct the original gene expression from this reduced representation for a held out set of cell lines. We computed RMSE and Pearson correlation of the reconstructed gene expression on the test set of CTRPv2 cell lines and compared it to PCA reconstruction from the first 100 principal components. The data splitting and training of Dr.VAE and PCA were performed the same way as in the other experiments described in the main paper.

Dr.VAE accomplished average gene expression reconstruction RMSE of 0.380 per gene, while PCA managed 0.329. In terms of Pearson correlation, the reconstructed gene expression from Dr.VAE and PCA correlated with the original expression levels with ρ equal 0.767 and 0.829, respectively. While Dr.VAE was trained for three tasks concurrently with model selection focused on the classification task, it still achieved good reconstruction accuracy. This shows that Dr.VAE does indeed learn latent representation of gene expression.

Chapter 4

Assessing domain adaptation for improvement of clinical drug response prediction

4.1 Introduction

Precision medicine in oncology aims to find the most effective treatment for the patients. Currently many of the treatments are selected by standard of care guidelines mostly based on primary site of the cancer, clinical variables or tumor biopsy, e.g. cancer cell shapes and tumor composition, which are not personalized and work for some patients, but not for most. In recent years increasingly more molecular biomarkers are used to stratify the genetically diverse cancers. These include univariate genomic markers of drug response, such as copy number or mutational status of particular cancer genes, e.g. ERBB2 (HER2) copy number amplification is predictive of response to lapatinib, or presence of $BRAF^{V600E}$ mutation is associated with response to MEK inhibitors in skin melanomas. However for many drugs, there are no known strong univariate biomarkers. Thus there is a need for more complex predictive models. Unfortunately clinical response datasets that include molecular characterization of the tumors (gene expression, CNVs, mutations or methylation) are not large enough, or not available at all, to train reliable machine learning models with large number of genomic features.

Therefore many have sought to improve clinical drug response prediction by utilizing drug response datasets of pre-clinical models (cancer cell lines, organoids or patient-derived tumour xenografts) beyond univariate biomarker discovery. Despite early excitement, transfer learning approaches have so far not yielded much more than proofs of concept [71, 221, 72], demonstrating that there can be response signal in cell line datasets that is transferable to patients.

Most recently, Mourragui et al. presented PRECISE [151], a domain-adaptation approach that uses linear alignment of source and target sub-spaces by aligning Principal Components of the two domains and interpolating between them to obtain domain-invariant bases to project both

domains onto. Mourragui et al. [151] rely on the "imputed drug-wide association study" (IDWAS) approach of Geeleher et al. [72] to evaluate their method on a clinical dataset. IDWAS was proposed to facilitate discovery of new drug-response biomarkers in originally unlabeled clinical cancer datasets, by using response predictors trained on cell line sensitivity datasets to "impute" the clinical response. These predicted responses, which Geeleher et al. [72] refer to as "imputed drug response", are then correlated with genetic variants measured in the clinical cohort, e.g. CNVs or mutations in the exome, to evaluate whether known clinically actionable associations of drugs and somatic alterations were recapitulated by the cell-line-trained response predictor. An example of a successfully recapitulated association was that of ERBB2 over-expression in breast cancers responsive to lapatinib treatment. Geeleher et al. [72] showed that in TCGA clinical cohort the predicted response to lapatinib, based on gene expression data alone, correlated with the patients' ERBB2 copy number status reported using immunohistochemistry. Geeleher et al. [72] suggested to use IDWAS as a hypothesis generator in identification of novel drug response biomarkers. However this approach is not suited for method comparison. This evaluation does not provide clear classification performance measure, rather just correlation with stratification by a known clinical biomarker, such as the ERBB2 copy number for lapatinib. Additionally note, that while ERBB2 over-expression is correlated with positive response to lapatinib and thus a clinically relevant biomarker, it is not perfect. In NeoALTTO clinical trial of lapatinib [10], the treatment was successful for less than half of the patients with over-expressed ERBB2 [64]. Due to problems with evaluation approach of Geeleher et al. [72] deployed in evaluation of PRECISE [151], it is unknown whether their models outperform simpler models trained only on a small patient-domain dataset, or whether their models are sensitive enough to pick up response signal beyond strong univariate biomarkers.

4.1.1 Assessment of domain adaptation assumptions

The domain difference between cancer cell lines and original tumours is caused by the process by which cell lines are derived from patient tumour cells, immortalization, that introduces mutations and systematic as well as essentially random gene expression changes into the cells. Next, cell lines grow in 2-D *in vitro* environment that does not correspond well to the *in vivo* tumour microenvironment, and further, cell lines continue to accumulate sporadic mutations. These aspects result in difference in marginal distribution $P(X)$ of gene expression between the domains. Previous methods mentioned above focused on mitigating the $P(X)$ domain shift by dataset homogenization using ComBat [110], surrogate variable analysis [130, 131], or more sophisticated sub-space alignment methods. However this marginal $P(X)$ shift is not the only difference between the two domains.

There are likely two principal causes for a lack of consistent success in domain adaptation attempts for drug response prediction. Firstly, a crucial assumption for unsupervised domain adaptation is the *covariate shift* assumption, that the conditional $P(Y|X)$ of the response outcome Y given the gene expression X does not change between the domains. However it is

not exactly valid in case of drug response in pre-clinical datasets and clinical trial datasets [212]. There are known cases when a treatment was found effective in cell lines but failed in patients and vice versa. Unfortunately there are no large public datasets of matched clinical tumour samples and from them derived cell lines. Therefore it is not possible to use matched data pairs to estimate the domain impact on $P(Y|X)$. One of the reasons for this $P(Y|X)$ inconsistency is that the causes influencing $P(X)$ shift between the domains, e.g. tumour microenvironment or the lack of, can also be influencing the outcome Y. Next, the response is quantified differently in different domains. In CCLs it is typically a summary statistic derived from a dose-response curve, where one point on the curve corresponds to relative amount of cells surviving after typically 72h of treatment by a particular drug dose. For patients, the positive clinical response can be defined in multiple ways and in varying time horizons, such as tumour shrinkage under a threshold after a cycle of treatment using the Response Evaluation Criteria in Solid Tumors (RECIST) [182] score, or using pathological complete response (pCR) as a surrogate for long-term outcomes. But there is not always a universal consensus on the exact evaluation criteria and thus the outcome evaluation can vary from trial to trial, as noted by the FDA guidance for use of pCR in breast cancer clinical trials [33]. Penault-Llorca et al. [159] clearly demonstrated that differences in pCR definition do matter as to what the computed response rates in a trial are. Therefore the genotype/gene expression X to outcome relation $P(Y|X)$ in pre-clinical models and in clinical trials can be considerably different.

Secondly, pre-clinical models have a selection bias for cancers that survive the immortalization process or manage to engraft in mouse models. Typically it is the more invasive tumours that pre-clinical models can be derived from. Similarly, there are often recruitment criteria in clinical trials of new treatments. Therefore datasets in these two domains are not i.i.d., but rather irregularly sampled. Cumulative effect of these and other factors, e.g. treatment dose and schedule, prior treatment, or environment, on learnable functional relationship of X and treatment outcome Y is then virtually impossible to quantify.

Despite the discussed issues, magnitude of which is difficult to quantify, it is worth evaluating performance of domain adaptation in drug response prediction, albeit with the limitations in mind. In few other applications, such as sentiment prediction in Amazon reviews among different shopping departments, domain adaptation methods that learn invariant representations were demonstrated to improve the target domain classification performance [66, 139, 218, 18] despite a lack of strict theoretical guarantees [106]. Thus given proof-of-concept results of Geeleher et al. [71] and others [221, 151, 183] for clinical drug response prediction from pre-clinical models, here we closely investigate domain adaptation for this task.

Based on previous successes in modeling of gene expression by deep generative models with approximate inference, presented in Chapter 3 and [210, 166], we decided to evaluate variational fair autoencoder (VFAE) of Louizos et al. [139], a domain-adaptation approach by latent space domain matching, for the drug response prediction. Description of VFAE and domain-adaptation background are presented in Section 2.6.

4.2 General experimental setup

In the following sections we focus on evaluation of VFAE as a domain adaptation method in three applications: (i) synthetic experiments designed to showcase success and failure cases of domain adaptation under several domain-shift scenarios; (ii) pilot study using gene expression datasets with real patient-cell-line domain shift and a substitute classification task with a strong signal in the data; and (iii) drug response prediction in real cancer datasets of fluorouracil and paclitaxel treatments. Our experimental setup is mostly shared across the three aforementioned applications; thus we first proceed with its detailed description.

4.2.1 Learning modes and data splitting procedure

We considered both unsupervised and semi-supervised domain adaptation (DA). In unsupervised DA, the typical DA scenario, a model is fit on labeled data from source domain and unlabeled data from the target domain, while the models' classification performance is evaluated on a held out set of labeled target domain examples. In short, we will refer to this unsupervised DA scenario as the source-to-target learning mode, denoted as S2T. In semi-supervised DA, additionally, we have also labeled target domain examples available for training, however only in a small number compared to the number of available labeled source domain data points. We shall denote semi-supervised DA as ST2T (source-and-target-to-target) learning mode. Note, ST2T is a flavour of transfer learning.

In real dataset applications we further evaluated classification methods on source and target domains alone, denoted as S2S and T2T, respectively. Comparison of achievable performance in these four learning modes provides us with empirical insight into difficulty of the prediction task and the present domain shift in real dataset applications.

Classification performance achieved on source domain, S2S, provides empirical generalization upper bound for baseline classification methods as well as for our selected VFAE architectures; here tested in SSVAE regime without domain conditioning. The gap between S2S and S2T shows empirical domain generalization gap (assuming no label distribution shift) that could be closed by a perfect DA method. The performance in T2T mode is a benchmark for practicality of a DA method deployment in practice. It is not practical to deploy a DA model trained in S2T mode if it does not provide improved performance compared to T2T models. Next, if ST2T is lower than T2T we can speak of negative transfer [202], that indicates that the domain shift is perhaps too large, or the domain adaptation or transfer learning method is misspecified. Finally, comparison of S2T and ST2T results directly elucidates value of even limited count of labeled target domain examples in model fitting on generalization of the model.

Dataset splitting into training, validation and testing set follows stratified 5-fold cross-validation scheme as is illustrated in Figure 4.1. In unsupervised domain adaptation mode, S2T, the training set contains unlabeled and also delabeled target domain examples. This way any difference between S2T and ST2T results can be fully attributed to utilization of the target

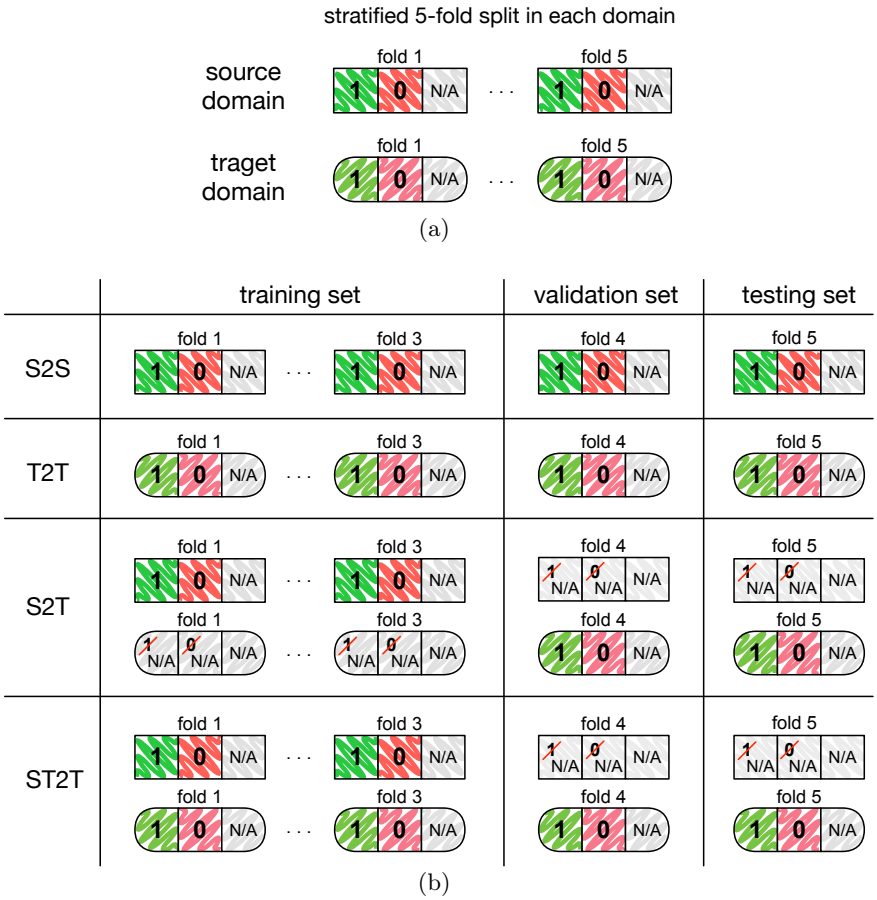
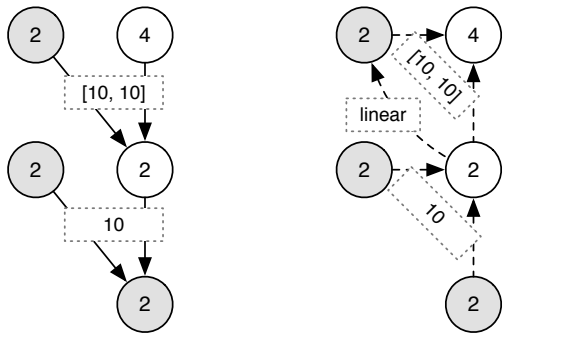


Figure 4.1: **Dataset splitting procedure.** (a) Both source and target domain datasets are first divided into 5 folds stratified by the target label $\mathbf{y}$. Green indicates positive examples, red negative examples, and grey is for unlabeled examples. (b) Folds are assigned to training, validation and test sets according to the depicted scheme. To achieve S2T and ST2T learning modes, labels of examples from applicable domains are censored and used as unlabeled examples. The fold assignment is then moved round-robin as in standard cross-validation.

domain labels in model fitting, and not to the dataset size. Similarly, source domain data are delabeled in validation and testing sets of S2T and ST2T, as such they only contribute to evaluation of between-domain MMD and reconstruction loss.

4.2.2 Baseline methods

Similarly to experiments in Chapter 3, we evaluated three established classification methods: ridge logistic regression (RidgeLR), random forest (RForest) with 100 trees, and support vector machine with a radial basis function kernel (SVMrbf). We used scikit-learn library [158] implementation of these classification methods. Each of them was evaluated on up to three different representations of the input $\mathbf{x}_1$ data: standardized $\mathbf{x}_1$, $\mathbf{x}_1$ projected into vector space of the first 100 principal components, and 100 dimensional latent encoding $\mathbf{z}_1$ learned by a VFAE or SSVAE model.



(a) parameterization of p (b) parameterization of q

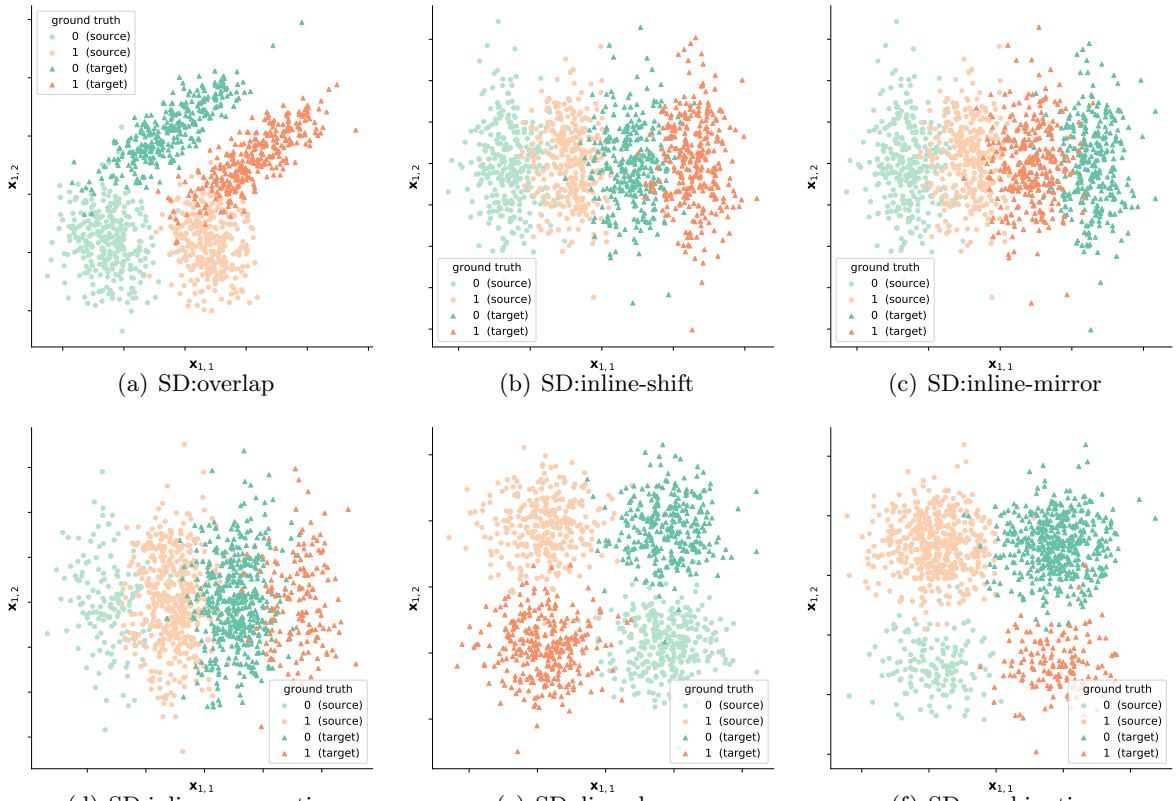
Figure 4.2: **VFAE architecture used in experiments with simulated datasets.** Depicted are the dimensions of VFAE random variables and architecture of feed-forward neural networks that parameterize (a) the generative distributions p, and (b) the approximate posterior distributions q of the model.

4.3 Experiments with simulated datasets

We simulated six different 2-D datasets to study VFAE for unsupervised and semi-supervised domain adaptation. These datasets include scenarios in which an unsupervised domain adaptation method that learns a domain-invariant representation, such as VFAE, is expected to perform well, and to fail in others [214, 106]. Next, we investigate whether VFAE in semi-supervised learning mode (ST2T) can succeed where an unsupervised learning mode fails, and whether learning of a domain invariant representation is still of any benefit in ST2T compared to baseline linear and non-linear classifiers that have to learn a more complicated classification boundary that applies to both source and target domain examples.

In each data scenario we sampled 500 positive and 500 negative examples in both source and target domain, i.e. 2000 data points in total. A random half of the dataset was then stripped off class labels to create unlabeled source and target examples. In following experiments we follow experimental setup as described in Section 4.2. However, since domains and classes in our simulations are clearly defined, it suffices that we evaluate just one random training-validation-testing split per each dataset. The training set of each of our six 2-D datasets is plotted in Figure 4.3.

We used a small VFAE architecture, depicted in Figure 4.2, that we kept unchanged across all the conducted experiments. Summary of test results from all experiments is shown in Figure 4.4, where Figure 4.4(a) is a table of classification results and Figure 4.4(b) presents domain discrepancy evaluation. Additional figures that illustrate VFAE training progression in each experiment can be found in Appendix B.



(a) SD:overlap (b) SD:inline-shift (c) SD:inline-mirror

(d) SD:inline-uneq-ratio (e) SD:diag-classes (f) SD:combination

Figure 4.3: **Simulated datasets.** Six simulated datasets we used for synthetic experiments of VFAE behaviour.

dataset (SD:x)	mode	VFAE	RForest	RidgeLR	SVMrbf
overlap	S2T	0.902	0.750	0.764	0.761
inline-shift	S2T	0.911	0.663	0.663	0.663
inline-mirror	S2T	0.095	0.670	0.670	0.670
inline-mirror	ST2T	0.952	0.952	0.488	0.944
inline-uneq-ratio	S2T	0.492	0.387	0.387	0.387
inline-uneq-ratio	ST2T	0.947	0.762	0.000	0.781
diag-classes	S2T	0.000	0.592	0.327	0.468
diag-classes	ST2T	0.968	0.952	0.976	0.976
combination	S2T	0.048	0.033	0.033	0.064
combination	ST2T	0.983	0.900	0.000	0.918

(a) Class $\mathbf{y}$ prediction performance quantified as F_1 score.

dataset (SD:x)	mode	RForest F_1 score		RForest AUROC		$\ell_{\text{MMD}}(\mathbf{z}_{1\mathbf{s}=0}, \mathbf{z}_{1\mathbf{s}=1})$	
		$\mathbf{x}_1 \rightarrow \mathbf{s}$	$\mathbf{z}_1 \rightarrow \mathbf{s}$	$\mathbf{x}_1 \rightarrow \mathbf{s}$	$\mathbf{z}_1 \rightarrow \mathbf{s}$	initial	final
overlap	S2T	0.929	0.667	0.980	0.741	0.407	0.114
inline-shift	S2T	0.970	0.513	0.995	0.506	0.212	0.089
inline-mirror	S2T	0.973	0.550	0.994	0.524	0.211	0.099
inline-mirror	ST2T	0.973	0.579	0.994	0.577	0.211	0.113
inline-uneq-ratio	S2T	0.955	0.670	0.990	0.720	0.200	0.198
inline-uneq-ratio	ST2T	0.955	0.739	0.990	0.761	0.200	0.550
diag-classes	S2T	0.965	0.627	0.993	0.685	0.246	0.061
diag-classes	ST2T	0.965	0.737	0.993	0.792	0.246	0.051
combination	S2T	0.990	0.707	0.994	0.761	0.189	0.123
combination	ST2T	0.990	0.988	0.994	0.995	0.189	0.487

(b) Domain $\mathbf{s}$ discrepancy evaluation.

Figure 4.4: **Results on simulated datasets.** (a) Test classification performance on $\mathbf{y}$, measured as F_1 score, of VFAE and baseline classifiers in all simulated scenarios. (b) Assessment of domain invariance of the learned VFAE latent representation $\mathbf{z}_1$ by: (i) comparison of domain $\mathbf{s}$ classification performance from original data representation $\mathbf{x}_1$ and latent $\mathbf{z}_1$ using a random forest classifier; (ii) comparison of domain discrepancy, measured as MMD, in VFAE latent representation before and after training. All presented results are evaluated on the test set. Note that a test set consists of held-out labeled target domain examples and of unlabeled examples from both domains for domain discrepancy evaluation.

4.3.1 Domain shift as mean and covariance shift with overlap (SD:overlap)

The first 2-D dataset we generated consists of source domain with two classes represented by two non-overlapping Normal distributions aligned next to each other along the first data space dimension. The domain shift is introduced as a mean shift predominantly along the second dimension and by a change in the covariance matrix of the class-representing Normal distributions resulting in their PDFs being shaped as a diagonal ellipse. The domain shift is set-up such that there is an overlap between negative class of source and target domain, and analogously for the positive class. We refer to this dataset as 'SD:overlap', and show its scatter plot in Figure 4.3(a).

In this case we expected unsupervised domain adaptation by a VFAE to work well, since satisfactory conditions set by [214] are met while drawbacks of MMD loss are not exploited. Indeed, VFAE managed to match the domains in its latent representation $\mathbf{z}_1$, illustrated in Figure B.1, and thus a linear classification component of VFAE, that is trained jointly with the rest of the model, successfully generalizes from source to the target domain. VFAE achieved 0.902 F_1 score on held-out target domain examples compared to 0.761 of SVMrbf, the best baseline with no domain adaptation. The domain invariance of the latent embedding learned by VFAE is quantified by significant decrease in ability to predict the original domain from this embedding and also by reduction of domain MMD from 0.407 at the beginning of training to 0.114 at the end of training, Figure 4.4.

4.3.2 Domain shift as in-line mean shift (SD:inline-shift)

In this dataset we simulated the domain shift to be a mean shift that moves the domains far apart enough for a cloud of examples from the same class to be disjoint between the two domains. Like in the previous dataset, we first sampled examples of two classes in the source domain from two Normal distributions, respectively, aligned next to each other along the first data space dimension. However in this dataset the domain shift is a mean shift along this first data space dimension. As such we denote this dataset as 'SD:inline-shift'.

In our experiment, VFAE together with domain MMD loss successfully aligned the domains, and learned a domain-invariant representation, which translates to practically no domain signal in the latent embedding $\mathbf{z}_1$, Figure 4.4(b). Additionally this embedding also correctly matches on class $\mathbf{y}$, Figure B.2, and thus gave rise to an accurate domain-invariant classifier, Figure 4.4(a).

Since the respective classes do not overlap between the source and target domain, it is not guaranteed that an *unsupervised* domain adaptation would always work without any other assumption. In this case we could assume that the domain shift is only a mean shift, thus mapping of the domains onto each other is likely to also correctly match on the classes. However, if the correct class mapping between the domains does not follow the steepest decent on the domain matching loss, an unsupervised domain matching will lead to mismatching of the classes and in turn to a poor target classification performance. We demonstrate this failures on the next two datasets, SD:inline-mirror and SD:inline-uneq-ratio, each showing a different failure.

Mirror class reversal (SD:inline-mirror)

Here we extend the domain shift from the above SD:inline-shift by swapping of the class order, making the domain shift to correspond to a mirror reversal of the source domain along the first data dimension. In unsupervised S2T setting, this domain shift appears exactly like the SD:inline-shift. Therefore, without any additional assumption, *every* DA approach that learns a domain-invariant representation is bound to fail in one or the other case. This impossibility results from no class-preserving overlap between the source and target domain.

As expected, VFAE mismatches the classes when projecting source and target domains onto a domain-invariant representation in unsupervised S2T setting, Figure 4.4. Next, we were interested to evaluate VFAE in semi-supervised ST2T setting. We observe, Figure B.3, that with labeled target domain examples in training set the VFAE can learn a domain-invariant embedding that also correctly matches the classes.

Unequal class ratios (SD:inline-uneq-ratio)

We recreated a failure mode described by Wu et al. [214], when marginal class distribution is shifted between the domains. That is, the ratio of classes is not equal between source and target domains. Such class distribution shift is generally an issue when we optimize for a symmetrical domain alignment, as is the case when using MMD loss or domain adversarial learning. Here, training VFAE in S2T mode led to domain-invariant embedding that matched many of the negative target domain examples to positive source domain examples, see Figure B.4.

Unsupervised domain adaptation can be fixed by asymmetrical relaxation of the domain adaptation loss as proposed in [214]. In case of VFAE with MMD loss, we need labeled target-domain examples to recover from this issue. Then VFAE can correctly match the domains and classes, Figure B.5, and accomplish better classification performance than baseline classifiers. However, in case of unequal class ratios between the domains, successful class matching between domains does not coincide with minimization of the domain MMD. Therefore the resulting latent embedding has higher domain MMD, Figure 4.4.

4.3.3 Diagonally opposed classes with 90° rotation (SD:diag-classes)

In this experiment we consider a variation on example by Johannson et al. [106] that further illustrates that just the fact that an optimal domain-invariant representation exists, even though a necessary condition, is not alone a sufficient condition. This dataset consists of two classes, represented by two non-overlapping Gaussians, that lie on a diagonal of a 2-D data space. Domain shift is a clock-wise rotation by 90° about the mean of the source domain. Similarly to previous examples SD:inline-shift and SD:inline-mirror, it is impossible, in unsupervised S2T setting, to distinguish a failure case that matches the domains but mismatches classes from a successful case yielding accurate target classification performance.

Our experimental results confirm the expectations. In unsupervised domain adaptation mode

VFAE slipped into the failure case, Figure B.6, or successful case based on random initialization of the dataset sampling and model initialization. While in semi-supervised ST2T mode, labeled target domain examples steered domain-invariant representation learning towards the success case, Figure B.7 and 4.4.

4.3.4 Domain shift as cross inversion with unequal class ratios (SD:combination)

The last experiment combines both aforementioned pitfalls for unsupervised domain adaptation. Firstly it is lacking class-consistent overlap between domains, and secondly, unequal class ratio that we saw to be a cause for a failure when optimizing for symmetric domain matching in SD:inline-uneq-ratio experiments.

This is an impossible case without any labeled target examples. Unsurprisingly unsupervised domain adaptation fails to successfully match classes between the domains, Figure B.8. Yet we were interested to study whether VFAE with MMD loss can recover correct domain-invariant embedding given access to some labeled target domain examples. A small hyperparameter adjustment of weights of classification loss $w_y = 18$ and MMD loss $w_{\text{MMD}} = 6$ led to successful domain alignment, Figure 4.4(a) and B.9, outperforming all baseline classification methods. However the final VFAE embedding $\mathbf{z}_1$ did not end up adversarially domain-invariant, Figure 4.4(b), as a random forest classifier can still distinguish the two domains. Note that adversarial domain invariance is not one of the VFAE training objectives, in our previously discussed experiments it is a result of domain matching by MMD. Unlike in SD:inline-uneq-ratio, we observed that the MMD loss had plateaued before the domains were aligned in a way a classifier cannot distinguish them. At that point; the last epoch is shown in Figure B.9(d); most of the MMD gradient points in direction of symmetrical domain alignment, and very little in direction leading to adversarial domain invariance.

4.3.5 Summary of simulation experiments

We have shown that unsupervised domain adaptation by VFAE can work well when a set of conditions is satisfied. However in many real world applications (particularly in high-dimensional datasets) class-consistent overlap between domains is unlikely. Thus S2T learning is not guarantee to work. Next, we have experimentally explored impact of marginal class distribution shift between domains, that an MMD loss is sensitive to. The good news is that in semi-supervised mode VFAE was able to learn correct domain matching that facilitated learning of a linear classifier, in such latent $\mathbf{z}_1$ space, that maximizes target domain classification performance often better than even non-linear baselines trained using the original data representation. Therefore demonstrating that promoting domain-invariant representation can improve classification performance even in ST2T setting.

Table 4.1: **Gene expression datasets summary.**

dataset	*total count*	*lung tissue*		*fluorouracil*			*paclitaxel*		
		pos	neg	pos	neg	unk	pos	neg	unk
TCGA	9359	1011	8348	87	54	9218	95	60	9204
CTRPv2	933	173	760	161	597	175	488	257	188

4.4 Experiments with real datasets

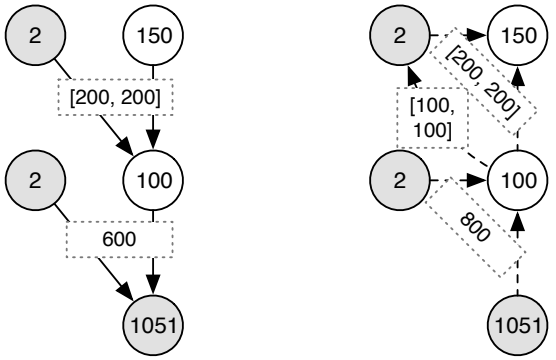
In this section we present domain adaptation experiments on real gene expression datasets, attempting DA between a patient cohort generated by The Cancer Genome Atlas (TCGA) Research Network: https://www.cancer.gov/tcga and cancer cell lines from Cancer Therapeutics Response Portal v2 (CTRPv2) [170] dataset. First we attempt to generalize lung tissue type classification trained using labeled patient examples to accurate prediction in the cell line dataset. Second, we evaluate DA for drug response prediction in a real scenario, when available are several hundred response-labeled cell lines and only much fewer response-labeled patients. The used datasets are summarized in Table 4.1.

In both TCGA and CTRPv2 datasets the gene expression was quantified from RNA-Seq using Kallisto [23] tool and normalized for sequencing depth and transcript length as TPM (Transcripts Per Million). Next, transcript expression levels were pooled into gene expression levels using GENCODE v23 human genome reference annotation [63]. Finally, we selected all landmark L1000 genes, resulting in 1051 genes, as our input feature set. We used per-gene standardized logarithm of TPM, $\log_2(\text{TPM} + 0.001)$, as the input gene expression levels. In the following text, we occasionally refer to the input features also as $\mathbf{x}_1$ as it corresponds to the observed random variable $\mathbf{x}_1$ of the VFAE model.

4.4.1 VFAE and SSVAE architecture hyperparameters

We experimented with two VFAE architectures selected based on the size of the labeled part of a training set. In tissue type prediction experiments, when primarily training on labeled TCGA patients, we trained a larger model denoted as "archTCGA". In all drug response prediction experiments and tissue type prediction experiment on a cell line dataset we employed a smaller "archCCL" architecture analogous to that of SSVAE presented in Chapter 3 and [166]. The two "archTCGA" and "archCCL" model architectures are illustrated in Figure 4.5.

In the following domain adaptation experiments we also use an SSVAE model as an ablation baseline, since it has comparable expressiveness but lacks conditioning on domain indicator variable $\mathbf{s}$ and domain-invariance regularization by MMD loss on the latent embedding $\mathbf{z}_1$. Each time we compare SSVAE and VFAE in one scenario, they share the common architectural hyperparameters. Note that when evaluating a VFAE model in S2S or T2T learning mode it reduces to an SSVAE model.



(a) archTCGA, p (b) archTCGA, q

(c) archCCL, p (d) archCCL, q

Figure 4.5: **VFAE architectures used in experiments with gene expression datasets.** Shown are hyperparameters of neural networks that parameterize the generative distributions p, and the approximate posterior distributions q of VFAE model. We used a larger "archTCGA" model architecture (a, b) or a smaller "archCCL" architecture (c, d) depending on the number of labeled examples in a training set.

4.4.2 Evaluation procedure

All experimental results presented in this section are based on evaluation in 10-times randomized 5-fold cross-validation following data-splitting procedure described in Section 4.2. In training of our deep generative models we used early stopping based on validation set performance with 20 epoch patience, and if the validation performance did not increase in 8 epochs the learning rate was halved. Classification performance was quantified by the area under the ROC curve (AUROC) and the precision-recall curve (AUPR). Where applicable we show results with their 95% confidence interval.

4.4.3 Tissue type prediction pilot study

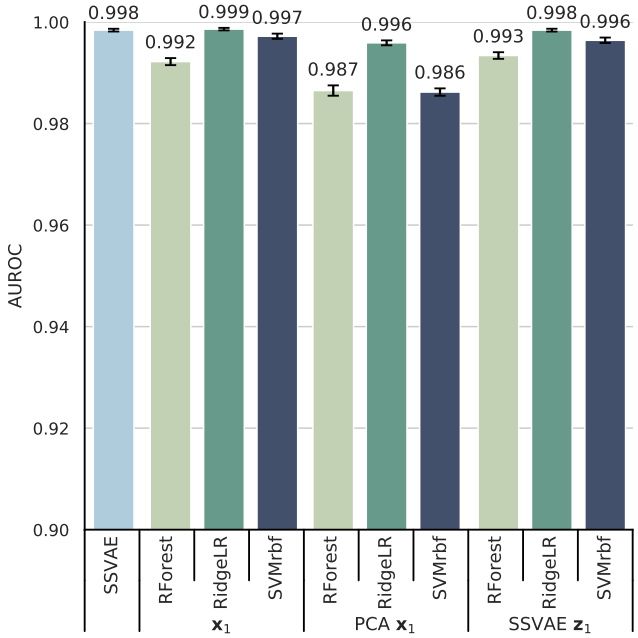
The first real gene expression dataset application we consider is a classification of the primary site tissue type of cancer tumour samples. Particularly, we simplify this task to a binary classification problem of distinguishing lung-derived samples from the others. We selected lung cancer samples as the positive class because lung is the most abundant primary site among TCGA and CTRPv2 samples. The source domain is 9359 TCGA patients (1011 lung cancers) and the target domain is 933 cancer cell lines (173 derived from lung cancer). In TCGA cohort we pooled LUAD (lung adenocarcinoma) and LUSC (lung squamous cell carcinoma) cancers together, each with 513 and 498 patients, respectively.

We designed the tissue prediction pilot study with three main reasons in mind. Firstly, this classification task is expected to have strong and broad (genome-wide) signal in both cell lines and patients. Secondly, even though cell line creation process introduces considerable perturbation of the transcriptome, it is reasonable to expect that the tissue signal is sufficiently conserved. Thus the covariate shift assumption is likely to hold. Finally, this pilot problem definition allows us to quantitatively evaluate VFAE on a large labeled datasets of real cancer gene expression, and conduct power analysis in this idealized setting. Note that here we attempt domain adaptation from patients to cell lines, contrarily to drug response prediction application, since in this case the patient domain contains 10x more labeled samples than the cell line domain.

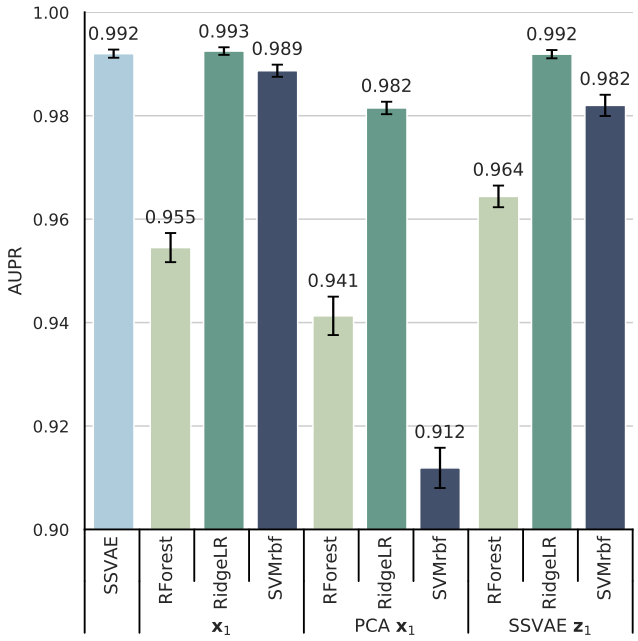
TCGA patients domain (S2S)

First, we set out to experimentally quantify strength of the tissue type signal in the source domain dataset. We trained SSVAE and a set of baselines to classify lung cancer patients in TCGA cohort. Note that here we trained an SSVAE model as this is a standard classification scenario in a single domain. Most of the evaluated models were able to achieve close to perfect generalization, scoring near 1 test AUROC, see Figure 4.6. SSVAE performed nearly identically to SVMrbf and Ridge logistic regression, with Random forest ensemble model falling slightly behind. We can conclude that lung tissue type is a property with a strong signal in the TCGA dataset.

A weight of SSVAE classification loss w_y used during training did not have much influence



(a) Area under ROC curve



(b) Area under P-R curve

Figure 4.6: **Source domain (patients) lung tissue type classification (S2S).** All classification methods can generalize and accurately predict lung tissue type in held out patient samples when there is no domain shift present and dataset is of sufficient size (9359 examples). Note, VFAE is not evaluated as it reduces to SSVAE in the case of a single domain.

on the result of this experiment. In a grid search the best validation results were achieved with $w_y \in \{30, \ldots, 50\}$. Next, it is worth mentioning SSVAE $\mathbf{x}_1$ reconstruction performance: Pearson $\rho = 0.877$, $R^2 = 0.822$; was nearly identical to that of PCA: $\rho = 0.878$, $R^2 = 0.826$. The obtained SSVAE performance also confirms that the deep generative model architecture "archTCGA" we selected is well suited to the task.

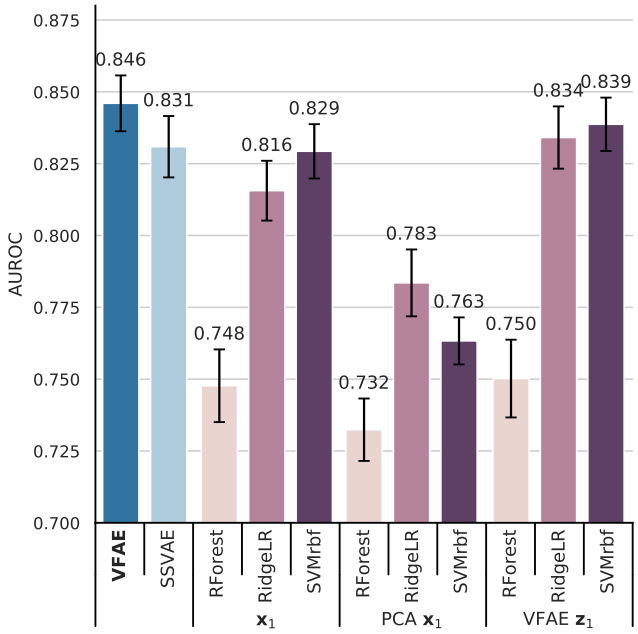
Adapting tissue type prediction from patients to cell lines (S2T)

We evaluated VFAE in unsupervised domain adaptation (S2T) learning mode against baselines that do not perform any domain adaptation. Consequently we conducted an ablation study to quantify contribution of features that set VFAE apart from SSVAE on the final classification performance. Those two features are: domain conditioning of $\mathbf{z}_1$ encoder and $\mathbf{x}_1$ decoder, and additional domain-invariance regularization of $\mathbf{z}_1$ embedding by domain MMD loss. Last but not least we performed power analysis by reducing the number of labeled examples in a training set.

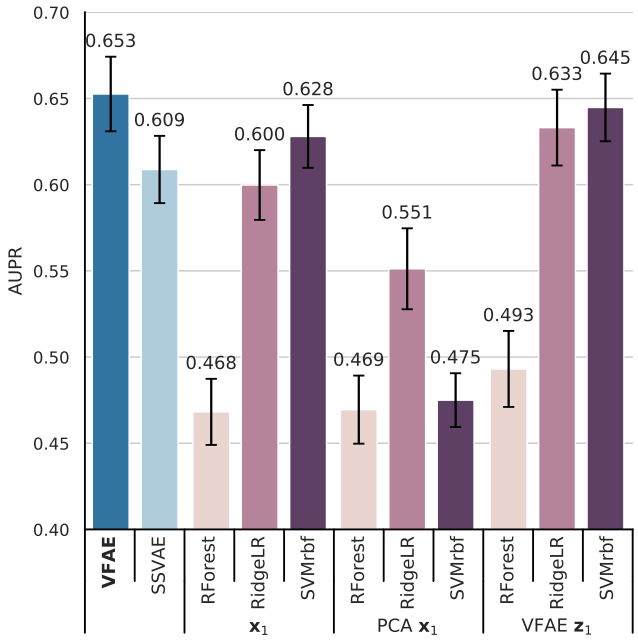
We run a grid search for weights of the classification loss w_y and domain MMD loss w_{MMD} used in VFAE training. The best weights, $w_y = 15$ and $w_{\text{MMD}} = 1200$, were selected based on validation split AUROC performance. VFAE achieved the best domain adaptation performance, eclipsing random forest, L2 logistic regression and SVM baselines, Figure 4.7. Note that baseline classifiers trained on VFAE latent embedding $\mathbf{z}_1$ perform better than when trained on the original input $\mathbf{x}_1$ of L1000 genes, while using PCA representation yields the poorest performance. This goes to show that VFAE learns embedding with distinctly better domain-invariant class features. At the same time the VFAE is good for a competitive reconstruction accuracy of the input gene expression; Pearson $\rho = 0.836$ and $R^2 = 0.759$; even though behind the PCA; $\rho = 0.867$ and $R^2 = 0.806$.

In Figure 4.8 we show the effect of w_{MMD} on VFAE results by modulating w_{MMD} from zero to 1800 in 300 increments. Usage of MMD regularization clearly aids in learning of domain-invariant $\mathbf{z}_1$ embedding, that, to a point, leads to better target domain classification performance. The VFAE performance is equal or better than the baselines for any $w_{\text{MMD}} \geq 300$.

Next, we conducted a power analysis by censoring the TCGA patient cancer type class in a training set. The number of labeled training examples was reduced to $1000, 2000, 3000, 4000$ and 5000 examples, respectively; the censored examples were used as unlabeled examples. As expected, the performance of all models increases with the number of labeled examples, while VFAE performance is higher than that of any other method, Figure 4.9. The increasing performance shows that the models' capacity had not been saturated even when all (approximately 5615) examples are used in training. Next, since the gap between VFAE and SVMrbf, the second-best model, tends to widen with the increasing number of labeled examples, the VFAE would likely benefit from further increased training set size more than the baseline. Most consequentially, in the lowest data regime there is only a marginal difference between VFAE and SVMrbf, which is concerning in case of intended application to limited-sized drug response datasets.

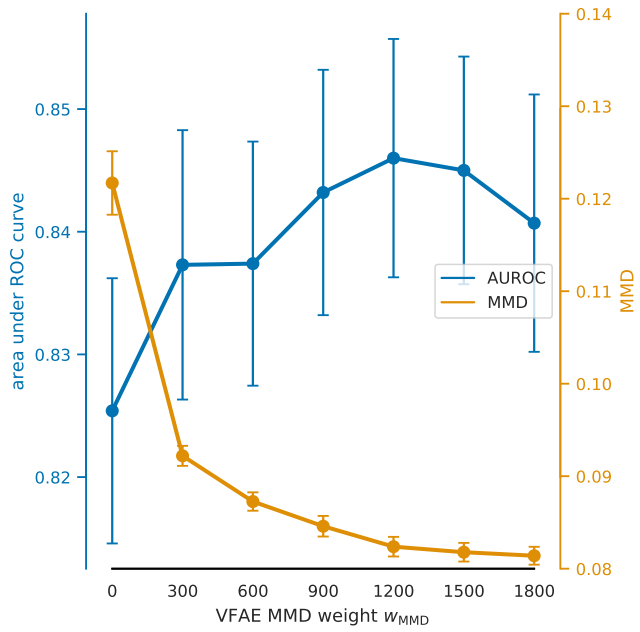


(a) Area under ROC curve

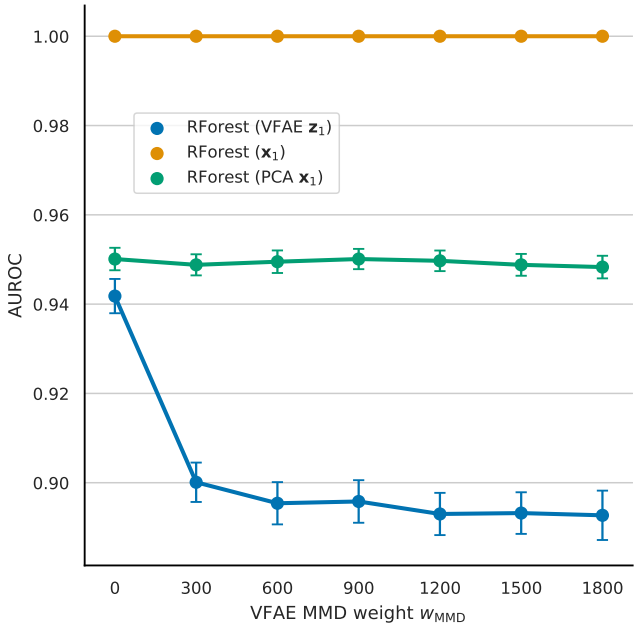


(b) Area under P-R curve

Figure 4.7: **Unsupervised domain adaptation (S2T) for predicting lung tissue type.** Classification performance of predicting lung tissue type in cancer cell lines by adaptation from patients to cell lines domain. VFAE was trained with loss weights $w_y = 15$ and $w_{\text{MMD}} = 1200$. Furthermore, we evaluated SSVAE and baseline classifiers trained on: (i) original L1000 genes, denoted as $\mathbf{x}_1$; (ii) top 100 Principal Components of $\mathbf{x}_1$; and (iii) VFAE's learned latent representation $\mathbf{z}_1$. Shown are test set (cell lines) results in terms of: (a) AUROC, and (b) AUPR. Error bars indicate 95% confidence interval.

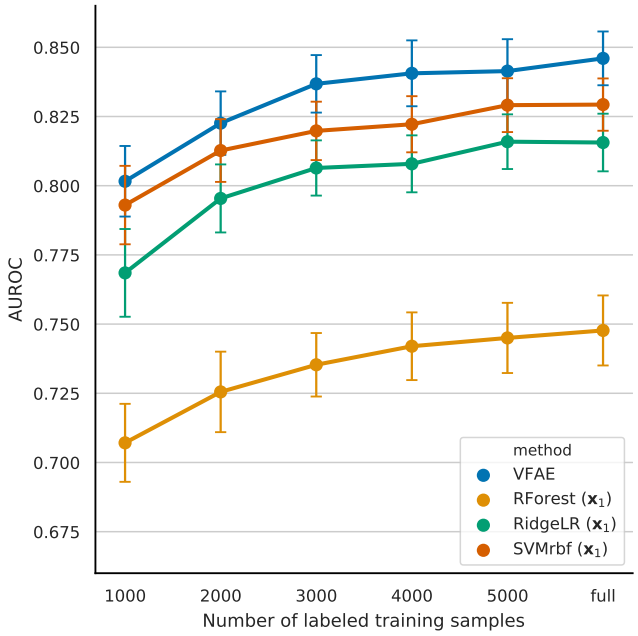


(a)

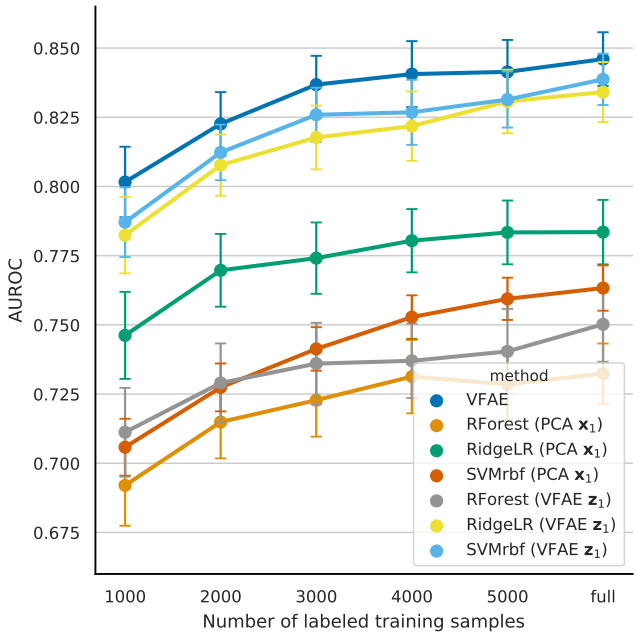


(b)

Figure 4.8: **Effect of MMD regularization in VFAE unsupervised domain adaptation (S2T).** Results are shown for VFAE trained with loss weights $w_y = 15$ and varying $w_{\text{MMD}} \in \{0..1800..300\}$. (a) Relationship of VFAE's test AUROC in tissue prediction task $\mathbf{y}$ and the domain discrepancy in VFAE's latent representation $\mathbf{z}_1$. (b) Shows classification performance of predicting the domain label $\mathbf{s}$ from VFAE's $\mathbf{z}_1$ compared to original L1000 genes and their first 100 PCs representation. Presented is test AUROC of a random forest classifier. Error bars indicate 95% confidence interval.



(a)



(b)

Figure 4.9: **Impact of labeled training set size in S2T setting.** Lung tissue classification performance of VFAE and baselines for varying number of labeled patients (source domain) in the training set. (a) VFAE compared to baselines trained on original data. (b) VFAE compared to baselines trained on 100 PC representation and 100-dimensional VFAE latent representation of the original data. Error bars indicate 95% confidence interval.

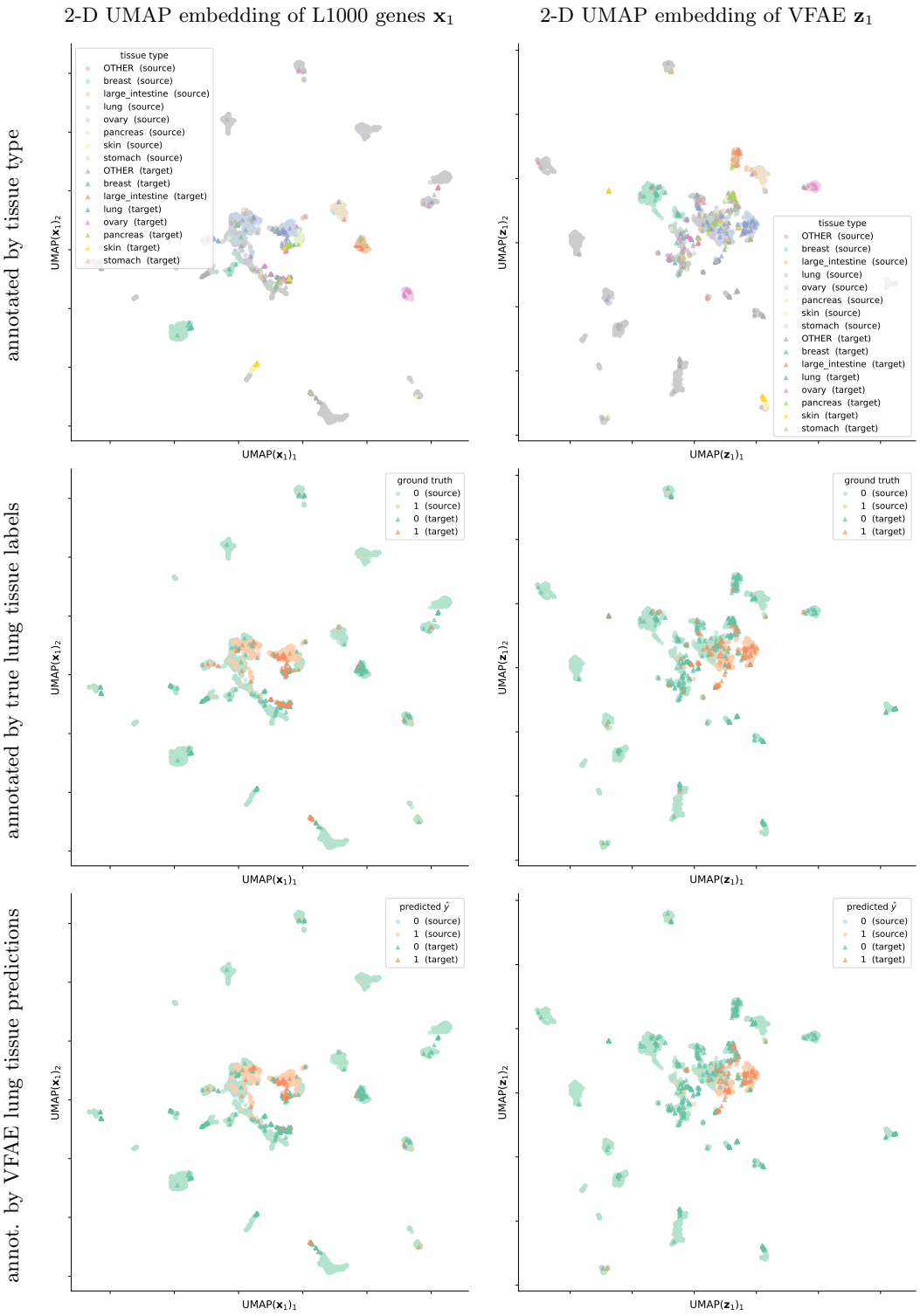


Figure 4.10: **Pilot study dataset visualization.** 2-D UMAP embedding of a TCGA (source domain) and CTRPv2 (target domain) training split. Shown is embedding of the input L1000 genes expression (left) and of the VFAE $\mathbf{z}_1$ embedding (right) annotated by tissue type (top row), true lung tissue labels (middle row), and VFAE predictions when trained in S2T mode (bottom row). VFAE embedding improves alignment of cell lines and patients, which improves domain adaptation for lung tissue classification.

Finally, we qualitatively investigated distribution of the dataset and its latent representation by VFAE. We used Uniform Manifold Approximation and Projection (UMAP) [144] to project the L1000 genes expression $\mathbf{x}_1$ and its corresponding VFAE embedding $\mathbf{z}_1$ onto a 2-D plane, shown in Figure 4.10. Annotation by the most abundant tissue types reveals that in several tissue types the cell lines and patients are rather close or overlap in the UMAP projection of L1000 gene expressions. While this supports our expectation that VFAE should be able to align cell lines and patients of the same tissue type, we can see in the plot of VFAE latent embedding of this data that even though the alignment of the domains is improved; for qualitative evaluation refer to Figure 4.8; the domains are not precisely matched on all tissue types. The domain alignment in VFAE embedding seems to have successfully aligned tissue types for which patients and cell lines are close or overlapping already in the original data input. Further, the unequal ratio of tissue type samples between our two domains in connection with use of MMD loss, an issue investigated in synthetic experiments earlier in Section 4.3.2, may too be contributing to the occasional tissue mismatch in the VFAE latent embedding. Importantly, the alignment of lung-cancer-derived samples is rather successful, as illustrated in the middle row of the Figure 4.8. This may be aided by optimization for the lung tissue classification task in training of the VFAE as well. Interestingly, lung-cancer-derived cell line samples that are outside of the patient cohort support; see the mid-right pane of Figure 4.8; are typically those incorrectly classified by the VFAE model; the bottom-right pane.

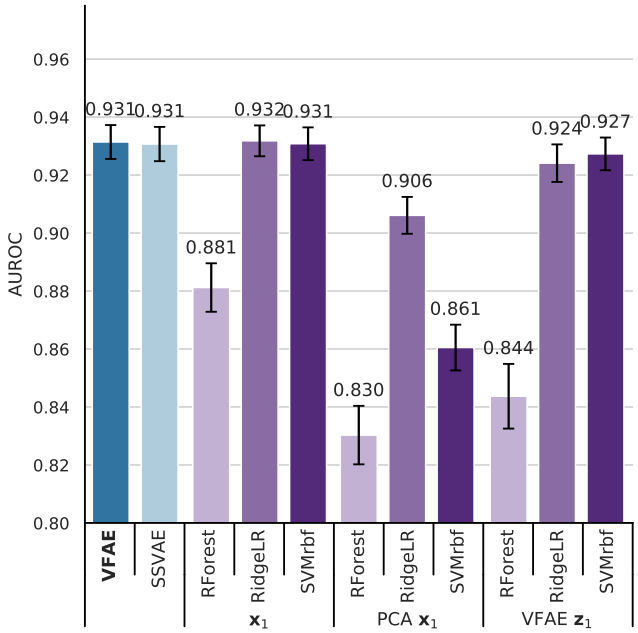
Semi-supervised domain adaptation (ST2T)

In this semi-supervised DA scenario we investigated how much a relatively small amount of labeled target domain examples can improve generalization of the prediction models. In our experiments we obtained significantly higher test performance across all models, Figure 4.11. However, unlike in our previous simulation experiments, training for a domain-invariant representation in VFAE does not lead to improved target domain classification. SSVAE performs almost equally to VFAE for any setting of MMD weight. Both evaluated deep generative models are comparable to the best baselines, while SVMrbf edges out the other models in terms of AUPR.

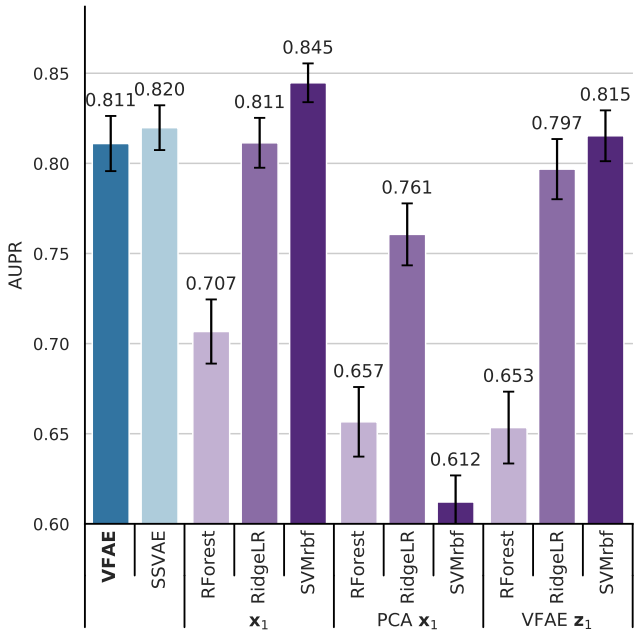
Cancer cell line domain (T2T)

The final set of tissue type prediction experiments assays how well can the considered methods generalize from a small training set in the target domain alone. Achievable performance in T2T mode sets a benchmark for assessment of domain adaptation practicality; whether DA methods can surpass training on a limited target domain training set.

The ‘archTCGA’ SSVAE architecture is too large for cell line dataset of this small size, therefore we used the smaller ‘archCCL’ instead. Still, considering previous S2S results, it seems that the dataset is too small for SSVAE to outperform baseline classification methods. Even though the difference is small, both logistic regression and SVM with the RBF kernel score

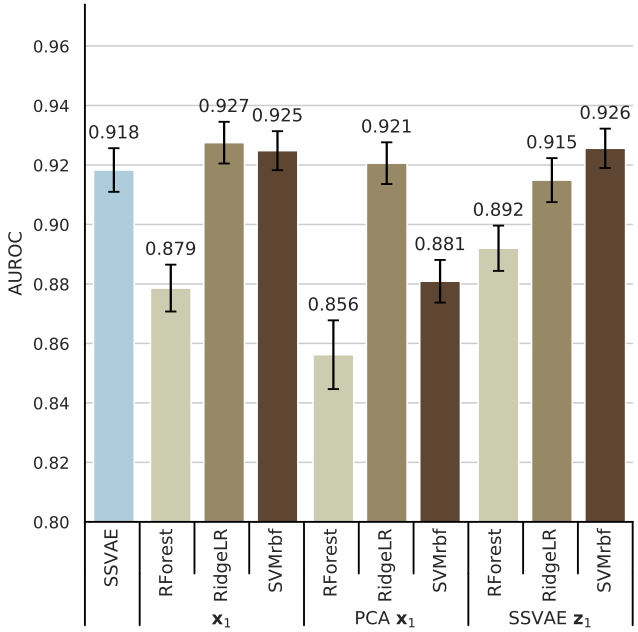


(a) Area under ROC curve

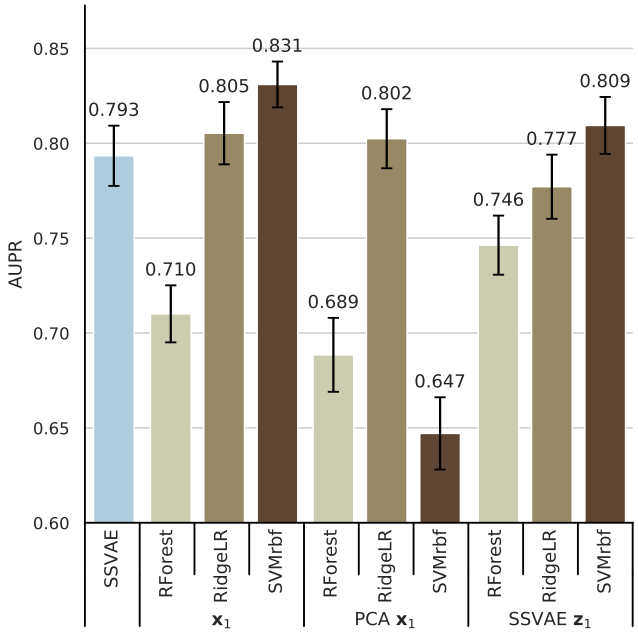


(b) Area under P-R curve

Figure 4.11: **Semi-supervised domain adaptation (ST2T) for predicting lung tissue type.** Classification performance of VFAE was nearly identical for any setting of w_{MMD}, yielding results comparable with SSVAE and the best baseline models.



(a) Area under ROC curve



(b) Area under P-R curve

Figure 4.12: **Target domain (cell lines) lung tissue type classification (T2T).** Most classification methods can achieve high prediction performance when trained and tested solely on cell line dataset of limited size (933 examples in total). Note, VFAE is not evaluated as it reduces to SSVAE in the case of a single domain.

higher than SSVAE, Figure 4.12. In terms of reconstruction of the $\mathbf{x}_1$ gene expression SSVAE performed well, yet behind PCA; Pearson ρ 0.756 of SSVAE vs 0.787 of PCA on the test set.

From the aforementioned practicality perspective, models trained in unsupervised S2T mode are worse than in T2T, while semi-supervised ST2T training does yield slightly improved performance over the T2T models. When comparing T2T to ST2T learning modes, we see that addition of a large amount of labeled data from the source domain actually improved prediction performance over training solely on a limited target domain dataset and lead to the best performing models overall.

Conclusion of the pilot study

Our pilot study with real gene expression datasets showed that domain adaptation by VFAE is feasible in case of broad class signal and sufficient similarity of the source and target domains. VFAE trained in unsupervised domain adaptation mode outperformed every other evaluated method that utilized labeled samples only from the source domain. Our ablation study also showed the domain adaptation features of VFAE model contribute to this improved target domain performance. On the other hand, we also observed signs of limitations of the unsupervised domain-invariant representation learning with Maximum Mean Discrepancy, that we had studied in the simulation experiments. Yet these limitations were not critically detrimental in this lung tissue prediction pilot study.

Taking a pragmatic perspective, the negative observation is that S2T mode lead to worse performance than T2T. That means training standard classification models on the limited target domain dataset is better than attempting unsupervised domain adaptation by VFAE in this case. Next, the power analysis in S2T mode rises concerns about domain adaptation applied to drug response prediction because of the small number of labeled examples in those datasets. However ST2T learning mode lead to training of the best performing models overall, proving that the patients and cell lines indeed share some common signal at least in relation to the tissue type. How these aspects pan out in the case of drug response task we analyze in the next section.

4.4.4 Drug response prediction from cell lines to clinical datasets

Finally, we evaluate domain adaptation for response prediction of fluorouracil and paclitaxel treatments. We aimed to leverage CTRPv2 dataset [170] that provides treatment sensitivity in cancer cell lines (source domain), to train a response classifier applicable to clinical patients (target domain).

In the following experiments we used TCGA patient cohort, with limited number of patients with known response to fluorouracil or paclitaxel treatments, where the clinical outcomes were curated by Ding et al. [47]. We bifurcated the patients' RECIST response scores such that patients with "Stable Disease", "Clinical Progressive Disease" or "Partial Response" score were considered non-responders and only patients with "Complete Response" score were labeled as positive responders. For exact CTRPv2 and TCGA dataset sizes refer to Table 4.1. In case of

cancer cell lines, the drug sensitivity in CTRPv2 experiments was first quantified by the area above dose-response curve as computed by PharmacoGx R package [187] and then bifurcated by the waterfall method [9, 89] into responders and non-responders.

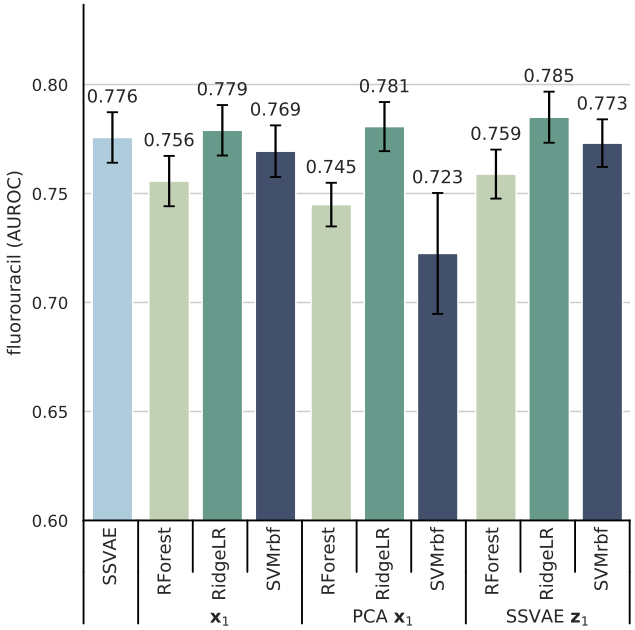
We follow the same experimental procedure as described earlier in this section. Next, due to limited labeled training set size in either learning scenario, we use VFAE/SSVAE architecture "archCCL", Figure 4.5, in all following experiments.

Results

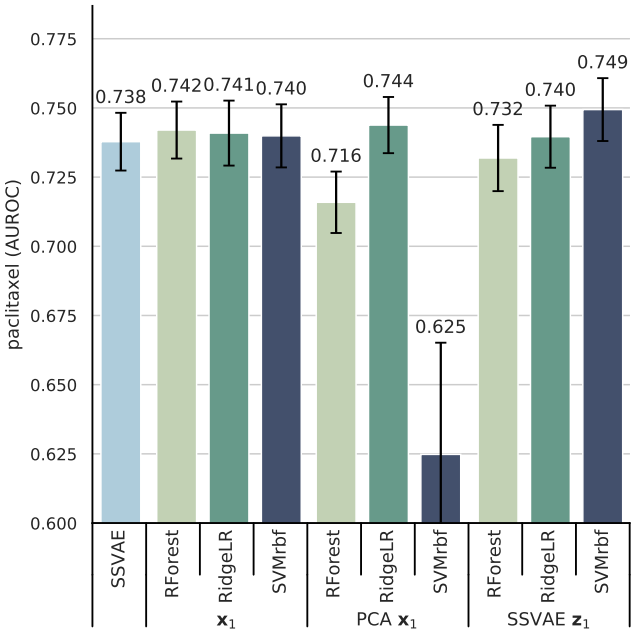
Domain adaptation by VFAE led to improved classification performance over SSVAE in both unsupervised (S2T) and semi-supervised (ST2T) domain adaptation mode, Figure 4.14 and 4.15. However the achieved performance does not surpass baseline methods. Overall, the best approach to predict patient response is SSVAE trained on patient data only (T2T), Figure 4.16.

The drug response classification signal is fairly strong in the cancer cell line dataset, where for both fluorouracil and paclitaxel all baseline models and SSVAE achieve similar performance well above 0.7 AUROC, Figure 4.13. However, considering the poor S2T results, the classifiers learned on cell lines do not transfer well to the patient dataset. Next, there is a sizeable gap between patient classification performance in ST2T and T2T mode as well. Additionally, in S2T and ST2T the best results were achieved with $w_{\text{MMD}} = 0$, which shows that forcing CCL and patient invariance by MMD is not helping to generalize classification of the drug response from CCL to patients. This may point to invalidity of covariate shift assumption. Perhaps the drift between response in CCLs and patients is too large for practically successful domain adaptation or limitations of VFAE come into play too.

Supplementary to the above quantitative evaluation we qualitatively assessed VFAE fit in S2T mode for fluorouracil response. Figure 4.17 shows 2-D projection by UMAP of the input L1000 gene expression $\mathbf{x}_1$ and VFAE embedding $\mathbf{z}_1$ annotated by either tissue type of the samples or by the fluorouracil response. The subclustering of the dataset is dominated by tissue type, while the response labeled samples are unequally distributed among different tissue (cancer) types in cell lines compared to patients. Here we can observe several necessary assumptions of domain-invariant representation learning for domain adaptation, earlier investigated in our simulation experiments, be violated: i) responders in one domain are typically not close to responders in the other domain; ii) unsupervised domain-invariant representation tends to match domains on tissue type as this component of variance dominants the response signal; and iii) unequal ratio of outcome classes between domains and tissue type hinders unsupervised DA with MMD loss, as studied in Section 4.3.2 and confirmed by selection of $w_{\text{MMD}} = 0$ in our hyperparameter grid-search. As such our experiments point to impossibility of successful *unsupervised* domain adaptation for drug response without further assumptions in general.

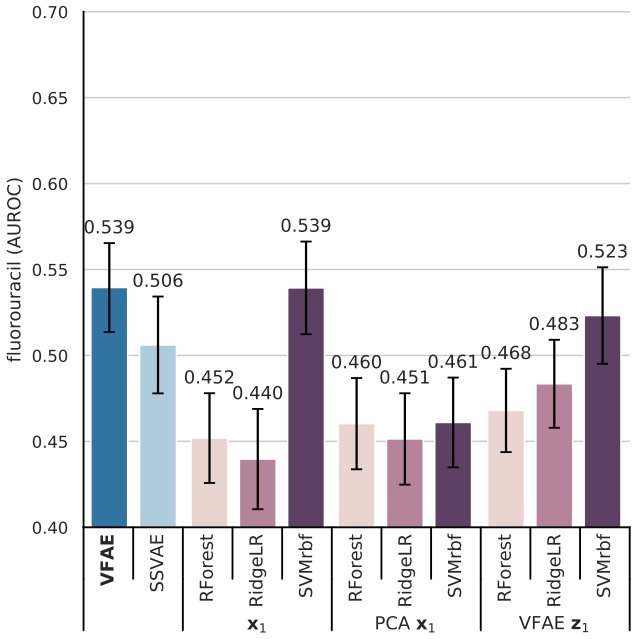


(a) fluorouracil

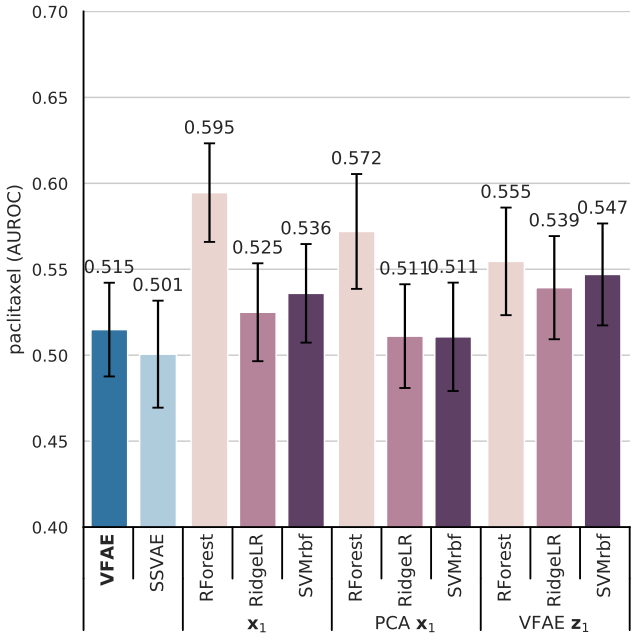


(b) paclitaxel

Figure 4.13: **Cancer cell lines only (S2S)** drug response prediction.

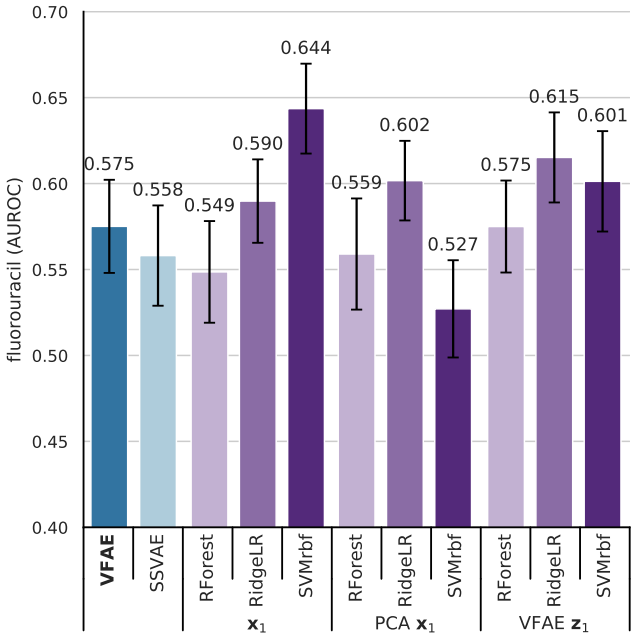


(a) fluorouracil

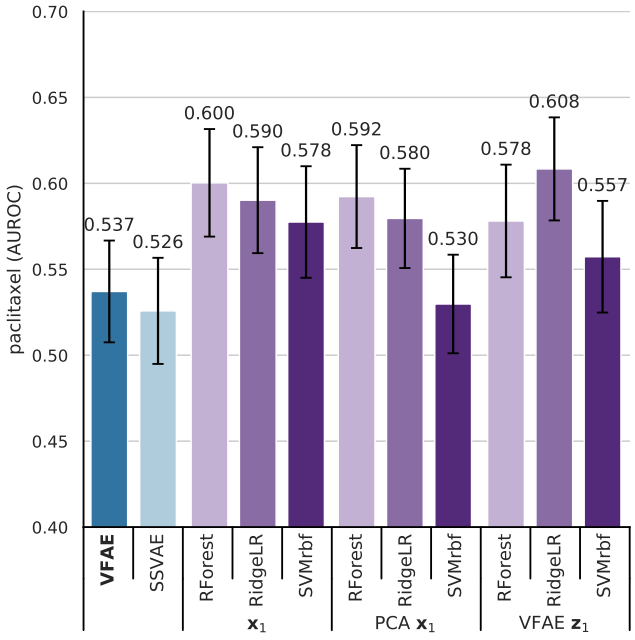


(b) paclitaxel

Figure 4.14: **Unsupervised domain adaptation (S2T)** for drug response prediction trained on cell lines and evaluated on TCGA patients.

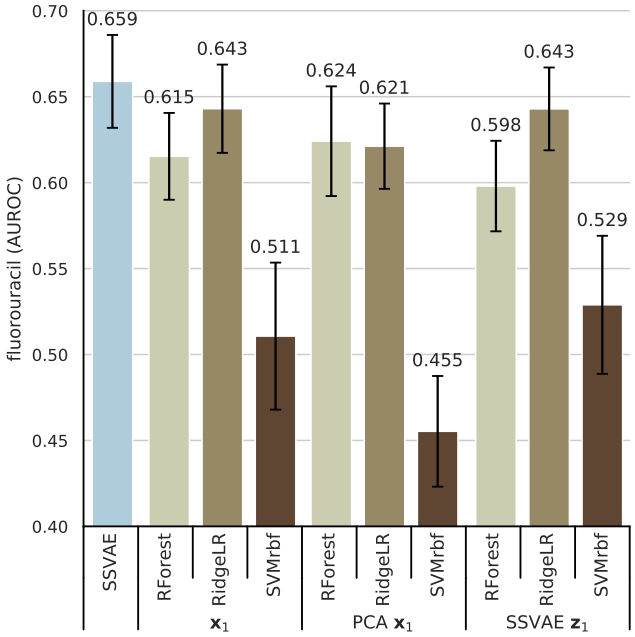


(a) fluorouracil

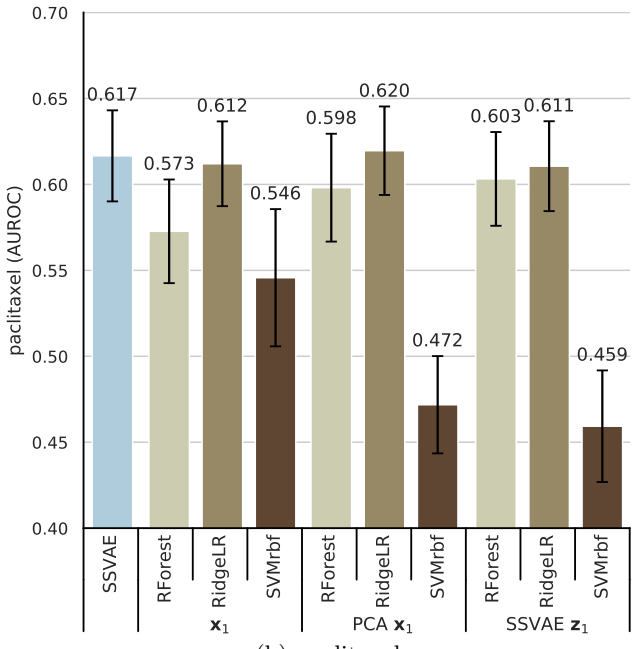


(b) paclitaxel

Figure 4.15: **Semi-supervised domain adaptation (ST2T)** for drug response prediction trained on both cell lines and TCGA patients, evaluated on held out patients.



(a) fluorouracil



(b) paclitaxel

Figure 4.16: **TCGA patients domain only (T2T)** drug response prediction.

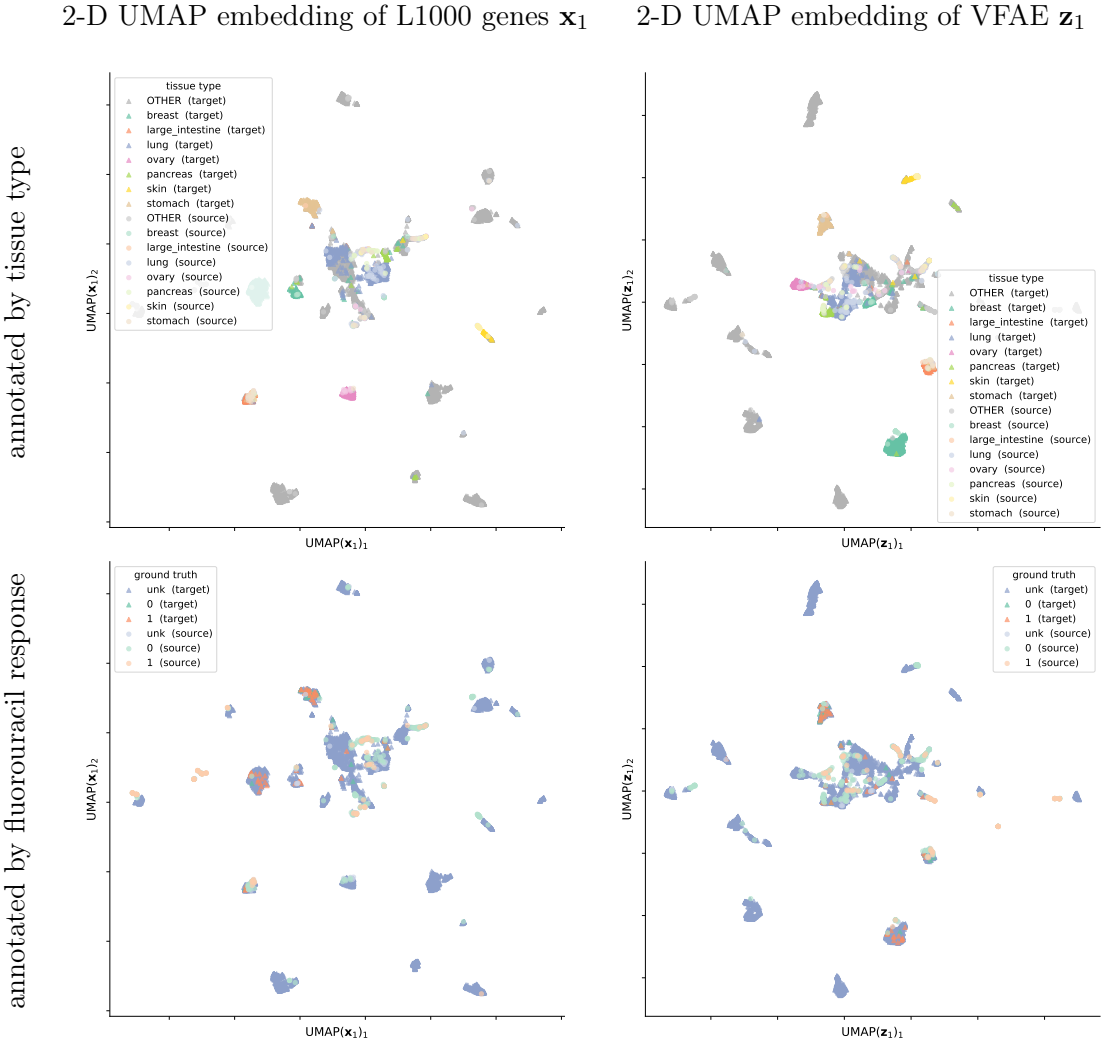


Figure 4.17: **Fluorouracil response in CTRPv2 and TCGA visualization.** 2-D UMAP embedding of a CTRPv2 (source domain) and TCGA (target domain) training split. Shown is embedding of the input L1000 genes expression (left) and of the S2T VFAE $\mathbf{z}_1$ embedding (right) annotated by tissue type (top) and ground truth fluorouracil response (bottom). Penalization of VFAE latent embedding to be more domain invariant does not lead to accurate transfer of fluorouracil response classification from cell lines to patients. Such representation matches marginal distribution of cell lines and patients that is dominated by tissue type, which does not benefit the drug response prediction task.

4.5 Discussion

In this chapter we investigated how drug response measured in pre-clinical models, particularly cancer cell lines, can be of use to improve clinical drug response prediction using domain adaptation methodology. We summarized theoretical assumptions for unsupervised domain adaptation and experimentally illustrated them in a series of simulations studies. In our experiments we have focused on a particular domain adaptation model, the Variational Fair Autoencoder [139], but most of our conclusions are not specific to this model, and rather apply to the family of domain adaptation methods that aim to learn domain-invariant data representations.

In a pilot study, where we aimed to learn a domain-invariant classifier of lung-cancer-derived samples, we saw that variational autoencoder-based models like SSVAE and VFAE can model gene expression comparably to PCA in terms of reconstruction quality, while their representation is co-optimized and thus considerably better suited for down-stream classification tasks. Next, our pilot study confirmed that when domain adaptation assumptions hold, or at very least are not strongly violated, domain adaptation features of VFAE undoubtedly contribute to improved domain-generalization of the task classifier.

Despite the success in our pilot study, application of VFAE to response prediction to fluorouracil and paclitaxel treatments turned out unsuccessful. We tracked down two principal causes of the lack of success: i) the drug response datasets do not conform well with the assumptions for successful domain adaptation, ii) even if the DA assumptions held as well as in our pilot study, the number of response-labeled samples is too low, which we showed in a down-sampling experiment, Figure 4.9.

Concerning the necessary assumptions we have the following takeaways. Firstly, the response signal, i.e. $P(Y|X)$, has to be mostly conserved between domains. Unfortunately in some cases response in cell lines differs from clinical response. Thus the covariate shift assumption may not hold and unsupervised domain adaptation is not possible without incorporating a suitable prior knowledge. Utilization of more faithful pre-clinical models, like organoids or PDX mouse models could alleviate this issue as these biological models are closer to the biology of cancer in humans. The practical limitation of these models is that with their increasing realism typically increases cost and decreases scalability of high-throughput pharmacogenomic studies, more so for PDX than organoids. Last but not least, after the biology of cancer can be accurately replicated in pre-clinical models, also the treatment response outcome has to be quantified in a way directly comparable to the outcome in a clinical trial. Without a matching response quantification the covariate shift assumption cannot hold.

Next, different class ratios between domains is detrimental in case symmetrical domain matching is enforced [214]. However in drug response datasets this class ratio asymmetry is a common occurrence and most importantly, the extend of it cannot be assessed from unlabeled target domain data. Furthermore, other heterogeneity in the datasets, such as ratio of cancer types or subtypes, between the domains may have similar effect. Here, a correct prior knowledge has to be incorporated, as some drugs are expected to have response that is tissue or cancer

subtype specific, while other drugs target a specific genomic aberration that is a common response marker across many cancer types, e.g. across all solid tumours.

Based on our series of simulation experiments, culminating in a combination of the aforementioned issues in the SD:combination experiment, as well as experiments with real gene expression and drug response datasets, we conclude that unsupervised domain adaptation by exact domain-matching is not generally applicable to drug response prediction from cell lines to patients. Not without thorough assessment and addressing of the points discussed here. Going forward, we will need either to develop domain adaptation methods with inductive bias that leads to successful domain matching for drug response, or rather instead of unsupervised domain adaptation focus on semi-supervised domain adaptation. Perhaps methods such as DIVA [101] that do not force domain-invariance and rather try to disentangle domain-specific and class-specific factors of variations may be more suitable. But perhaps most likely avenue of further research is utilization of transfer learning or multitask learning methods where response in cell lines is considered a correlated task that may help to learn a more robust prediction model given limited number of labeled patient data, ideally together with incorporation of biological priors. In case negative transfer is observed, i.e. when treatment response in cell lines is very different from patients, we best focus on patient-only models with treatment-relevant feature preselection, e.g. gene sets implicated in given cancer type or pathways related to the drug mechanism of action, or other prior incorporation.

Chapter 5

fCNV: probabilistic method for detecting copy number variation in a fetal genome using maternal plasma sequencing[†]

5.1 Introduction

Until recently, the prenatal analysis of a fetal genome required samples directly obtained from the fetus by invasive procedures like chorionic villus sampling or amniocentesis, where amniotic fluid is sampled from around the developing fetus. Amniocentesis, however, has several important disadvantages. Foremost, it carries a non-trivial risk of miscarriage (estimated procedure-related fetal loss rate is 0.6% to 1% [49]), and hence is refused by a fraction of patients. Secondly, amniocentesis cannot be performed too early, as the risk of miscarriage rises significantly, and is typically indicated for the 15th week of pregnancy, outside of the time-frame for the safest abortion options (<12 weeks) and leaving only limited time for follow-up analysis. Finally, amniocentesis is a complex and expensive medical procedure ($1,500–$3,000). Consequently, amniocentesis is typically performed only to confirm or reject a diagnosis if a genetic disease is suspected, e.g. high likelihood of Down syndrome based on prenatal ultrasound.

The several years has seen the initial development of alternative, non-invasive methods for prenatal genetic testing. Prominent among these are methods that are based on analysis (arrays or sequencing) of cell-free DNA (cfDNA) extracted from maternal blood plasma, which contains an admixture of fetal and maternal DNA. The fraction of fetal DNA in such an admixture varies depending on multiple factors, including maternal weight and size of the fetus, but typically builds up from ~5-7% early in the pregnancy to 10% at week 10 [207] to as much as 50% before

[†]Work published in Oxford *Bioinformatics* (2014) [165], presented at conferences RECOMB 2014 and ISMB 2014, and featured in GenomeWeb [198].

delivery [207, 60]. In experiments conducted by Kitzman et al. [120] (and utilized in this paper) the estimated admixture in samples obtained at 8 weeks and 18.5 weeks of gestation was 7% and 13%, respectively.

The decreasing cost of DNA sequencing has made it practical to directly sequence cfDNA extracted from maternal blood to identify likely genetic disorders present in the fetus. Non-invasive methods are becoming more commonly used to directly identify aneuploidies (abnormal chromosome counts) and are also enabling preventive screening for heritable genetic diseases, resulting in better prenatal health care [181]. While most non-invasive genetic diagnostics aim to test for a particular previously known biomarker, Kitzman et al. [120] demonstrated the possibility of the reconstruction of the whole genome of the fetus by combining whole-genome sequencing of parental genomes with deep sequencing of cfDNA from maternal plasma (78x coverage). The key intuition in this method is the comparison of allelic ratios at individual SNP loci, as the inheritance of a particular paternal allele affects the percentage of reads with that allele at the particular position in the genome. This method heavily relies on the availability of phased parental genotypes, as these allow for the inference of likely co-inherited SNPs, leading to an improvement in the signal-to-noise ratio. It consequently provides for high accuracy identification of inherited (98% accuracy) but not *de novo* single nucleotide variants (17 correct calls out of 44 true *de novo* sites, with 3884 called positions).

While most efforts to detect Copy Number Variants (CNVs) from cfDNA sequencing have so far concentrated on whole-chromosome events (e.g. Chu et al. [29]), the past year has seen the first few attempts at methods for genome-wide identification of sub-chromosomal *de novo* CNVs. Such methods are desired to enable non-invasive prenatal screening for diseases like DiGeorge syndrome (~3Mb deletion), Prader-Willi syndrome (the deletion subtype caused by a ~4Mb deletion), and other disorders associated with a mid to large sized CNV. So far two publications address the problem of detecting sub-chromosomal CNVs [27, 190]. While the exact methods used in both of these approaches differ, both rely on depth of coverage: they map the reads to the genome, divide the genome into bins, and identify the CNVs by comparing the number of reads mapped to each bin. The key idea in these methods is that deletions/duplications will result in more/fewer fetal reads within a window, and this difference can be identified using statistical methods. Srinivasan et al. [190] use depth-of-coverage computed in 1Mb windows across the genome to identify CNVs that are typically >1MB, though they do report discovery of a 300kb CNV. 9 of the 22 discovered CNVs in 11 patients were concordant with karyotyping results, with most discrepancies being short (<1Mb) CNVs. Importantly, they use extremely short (25bp) reads, allowing for larger number of fragments at equal coverage depth. Chen et al. [27] use even larger 10Mb windows, again considering only the number of fragments mapped and are able to successfully identify variants 9-29Mb with only one false positive among 6 true positives in 1311 patients. Both methods utilize low coverage WGS of the plasma cfDNA, while leveraging the large number of samples.

In this chapter we introduce a novel model for non-invasive prenatal identification of de

novo CNVs with increased sensitivity compared to methods published so far. Our method combines three types of information within a unified probabilistic model. First, our method takes advantage of the imbalance of allelic ratios at SNP positions that are introduced by various types of paternally and maternally inherited CNVs. Secondly, following the work of Kitzman et al. [120], we use parental genotypes to phase nearby SNPs, modelling their co-inheritance (or recombination) and thus improving the signal-to-noise ratio. Finally, we observed that allelic ratios poorly differentiate between certain types of CNVs: for example, as further described below, a duplication of a paternally inherited allele results in extremely similar allelic ratios to deletion of a maternally inherited one. We thus combine the allelic ratios with the depth-of-coverage signal to better differentiate between such cases. Our simulation results, based on *in silico* introduction of novel CNVs into plasma samples with 13% fetal DNA concentration, demonstrate a sensitivity of 90% for CNVs >400 kilobases (with 13 calls in an unaffected genome), and 40% for 50-400kb CNVs (with 108 calls in an unaffected genome).

5.2 Methods

Our method models two types of signal from the data: (i) imbalance of the allele distributions at SNP loci (discussed in Section 5.2.1), and (ii) number of fragments sequenced from ~1kb genomic regions (discussed in Section 5.2.2). Though each of these is noisy, the two are (nearly) independent (modulo number of reads overlapping the SNP position) variables and can be combined into a single generative model. For this purpose we use a Hidden Markov Model (HMM), where we interpret the allele counts at SNP loci as emissions, while the coverage is used as a prior probability for each state (see Section 5.2.3).

For our method we assume that we have phased haplotypes of both parents, and deep sequencing data of cfDNA from maternal plasma. In practice we used whole genome sequencing (WGS) data for the parents, with phasing based on 1000 Genomes data (see Section 5.3.1). All *de novo* CNVs thus correspond to a particular parental haplotype duplication or deletion event. Labelling the two maternal and paternal haplotypes as M_A, M_B, P_A, P_B. For each inheritance pattern – normal inheritance, maternal duplication, paternal duplication, maternal deletion, paternal deletion – we introduce a set of *phased inheritance patterns* that enumerates all the possible configurations of fetal haplotypes corresponding to the respective inheritance pattern. For example a duplication in the maternal gamete will consist of one (or more) of six phased inheritance patterns:

$$M_AM_AP_A,\quad M_AM_BP_A,\quad M_BM_BP_A,$$
$$M_AM_AP_B,\quad M_AM_BP_B,\quad M_BM_BP_B$$

There are a total of 20 phased inheritance patterns (PP): 6 each for maternal/paternal duplication, 2 each for maternal/paternal deletion, and 4 for normal inheritance). We refer to the number of alleles (copy count) inherited by the fetus as $|PP|$. We use r to refer to the percentage of cfDNA

that is fetus-derived; this parameter is estimated from positions in the genome where the parents are homozygous for alternate alleles.

5.2.1 SNP Allele Distribution

For every SNP locus we observe a distribution of nucleotides in maternal plasma reads. In this section we focus on calculating the probability of the observation with respect to a phased inheritance pattern. Formally, we observe the counts of the 4 nucleotides $\{k_{\mathtt{A}}, k_{\mathtt{C}}, k_{\mathtt{G}}, k_{\mathtt{T}}\}$ and compute the probability of observing each of these from a particular phased inheritance pattern PP. Ideally, these counts should follow multinomial distribution with the parameter vector $(p_{\mathtt{A}}, p_{\mathtt{C}}, p_{\mathtt{G}}, p_{\mathtt{T}})$. However we have found that modelling them as independent Gaussians with variance equal to the mean (as an approximation of the Poisson), makes the inference of the inheritance pattern more robust to noise.

More formally,

$$Pr[k_x \mid M_A, M_B; P_A, P_B; r; PP] \sim \mathcal{N}(\mu_x, \mu_x) \tag{5.1}$$

To compute the expected support μ_x for $x \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, we first adjust the mixture ratio r based on the expected number of fetal haplotypes $|PP|$, as absence/presence of an additional fetal copy in the plasma sample influences the local fetal mixture ratio. We accommodate this influence of $|PP|$ expected fetal haplotypes instead of regular two as follows:

$$r' = \frac{|PP| \cdot r/2}{|PP| \cdot r/2 \ + \ (1-r)} \tag{5.2}$$

Then for each nucleotide x we compute the probability p_x of observing a read supporting x. Such a read might have originated from multiple haplotypes, including two maternal haplotypes and $|PP|$ fetal haplotypes. We can individually evaluate this probability for each haplotype and subsequently sum them to obtain p_x:

$$\begin{aligned} p_x = & \sum_{i \in \{A,B\}} [M_i \text{ equals } x] \cdot m_i (1 - r') \\ & + \sum_{y \in PP} [y \text{ equals } x] \cdot \frac{r'}{|PP|} \end{aligned} \tag{5.3}$$

For reads putatively coming from maternal portion of the cfDNA sample, we correct for maternal CNVs by using the allele ratios m_i as observed in maternal-only sequencing data. Additionally, in order to mitigate noise we add pseudocount α (proportional to the genome-wide coverage) to these counts.

$$m_i = \frac{\alpha + \#\text{reads supporting } M_i \text{ in maternal sample}}{2\alpha + \sum_{j \in \{A,B\}} \#\text{reads supporting } M_j \text{ in maternal sample}} \tag{5.4}$$

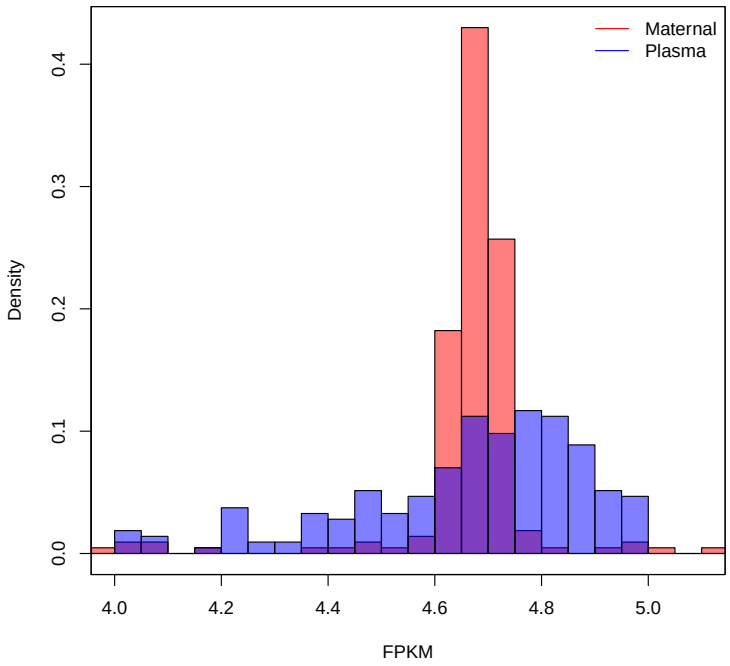


Figure 5.1: Distribution of fragments per kilobase of chromosome 1 per million fragments (FPKM) in 1 megabase segments for plasma sample (blue) and maternal sample (red) of the I1 trio.

We thus obtain the expected probability distribution for each nucleotide observed at this SNP locus.

In order to obtain the expected number of reads μ_x supporting particular variant at this SNP locus, we have to multiply p_x by the number of reads mapped,

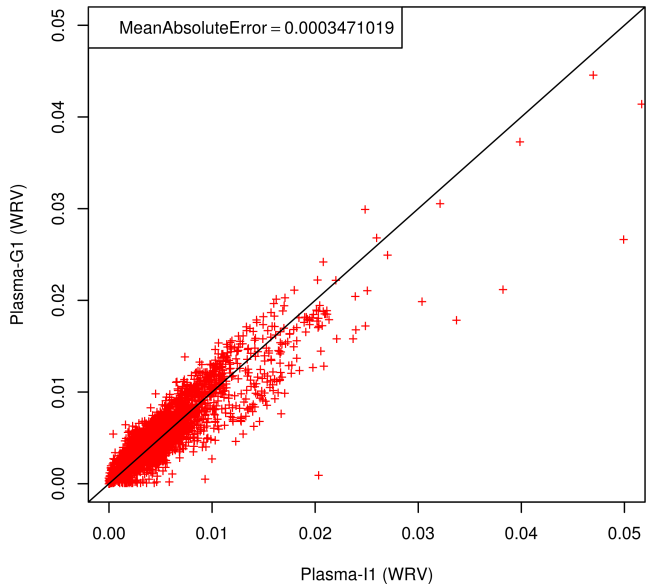
$$\mu_x = p_x \cdot \#\text{mapped reads} \tag{5.5}$$

As we describe later, we use this probability distribution $\mathcal{N}(\mu_x, \mu_x)$ that is conditional on phased pattern PP as the emission distribution for each nucleotide in our HMM.

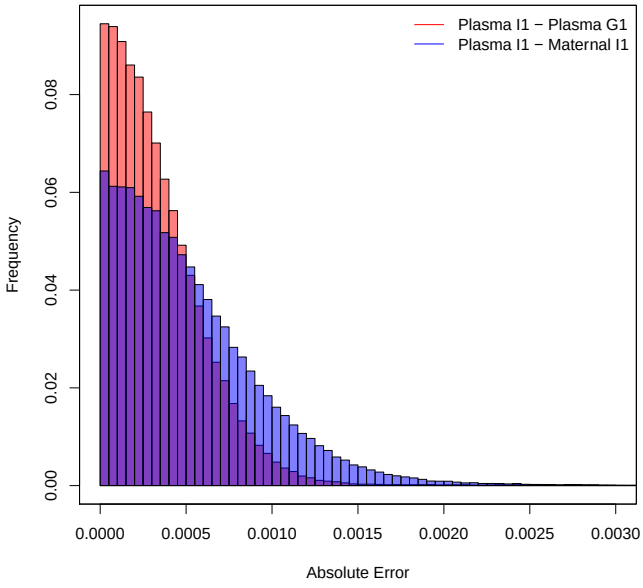
5.2.2 CNVs and Depth of Coverage

Variations in number of fragments sequenced per a region is a standard measure used for detection of mid to large sized CNVs (see Medvedev et al. [145] for a review), and has also been used for CNV detection from maternal plasma [190, 27]. However the relatively low admixture of fetal DNA in the maternal plasma together with cfDNA sequencing biases considerably limit potential of methods relying on coverage signal from a single sample. Furthermore, the high variability of the coverage derived from blood plasma (Figure 5.1) makes it difficult to identify shorter CNVs. Thus methods of Srinivasan et al. [190], Chen et al. [27] use large bins and require multiple datasets to establish a baseline for CNV calling.

Simultaneously, the coverage forms an important complementary signal to the allelic distributions described above: certain ratios have very similar probability under different phased



(a)



(b)

Figure 5.2: (a) A scatterplot demonstrating the correlation of window ratio values (WRV) between plasma samples of I1 and G1 trios. The shown WRVs were computed for windows of size 1kb in chromosome 1. (b) Histogram of absolute errors between WRVs from different samples; comparing distribution of absolute error between plasma samples of I1 and G1 trios (red), and between plasma sample and maternal sample of I1 trio (blue). There is a notably heavier tail in case of plasma to maternal sample error distribution, composed of windows with weak WRV correspondence – an artifact of wider coverage distribution in plasma cfDNA sample compared to standard WGS maternal sample (Figure 5.1). This artifact causes plasma to maternal sample WRV comparison to have higher mean absolute error (0.000521, compared to 0.000347 for plasma I1 to plasma G1) even though they are from the same trio.

patterns, e.g. a deletion of a maternally inherited allele may yield distributions similar to a paternally inherited duplication. Incorporating the coverage signal helps to discriminate such states. In our method, we use the coverage information as a noisy predictor to complement the signal we obtain from SNP loci.

As a measure of coverage in a genomic region we use *window ratio value* (WRV) analogous to the *bin ratio value* measure used by Srinivasan et al. [190], which is essentially the number of fragments mapped to the region and normalized by the number of fragments mapped to other regions with similar GC content. Note that window ratio values are independent of GC content and depth of sequencing of the sample.

For the purpose of our model, we split the genome to non-overlapping windows, each containing a single SNP, with breakpoints being in the middle between two adjacent SNPs. For each SNP i the corresponding WRV_i for the window W_i containing the i-th SNP position is then computed as the ratio of number of fragments N_{W_i} mapped to W_i to the sum of fragments mapped to 200 windows of the same size with GC content closest to W_i:

$$WRV_i = \frac{N_{W_i}}{\sum_{W \in neigh^{200}_{\mathtt{GC}}(W_i)} N_W} \tag{5.6}$$

However, the variable length of the windows makes such computations expensive as computation of $neigh^{200}_{\mathtt{GC}}(W_i)$ is linear in number of windows. To make the WRV computations practical, we scale N_{W_i} to correspond to expected number of fragments as if $|W_i| = 1$kb by multiplying N_{W_i} by $1000/|W_i|$ (for clarity, not shown in our equations). Then WRVs in 1kb bins can be precomputed, enabling us to find $neigh^{200}_{\mathtt{GC}}(W_i)$ in time logarithmic from the number of bins. Using 1kb bins is a good approximation as the mean distance between two adjacent SNP loci is expected to be 1kb.

Overall, our goal is to estimate the probability of observing WRV^S_i in the studied plasma sample conditional on the number of fetal haplotypes ($|PP|$), which is either three for duplication, one for deletion, or two for normal inheritance. To do so, we use a reference sample to obtain WRV^R_i for comparison (computed in the same genomic window W_i). Further we need to compute two more reference WRV^R_is, each scaled to reflect one CNV type. For duplication, we would expect to see $(1 + r/2)$ times more fragments while for deletion $(1 - r/2)$ times less fragments, thus the scaled $WRV^{R,|PP|}_i$ is estimated as

$$WRV^{R,|PP|}_i = \frac{N_{W^R_i} \cdot \big(1 + (|PP| - 2) \cdot r/2\big)}{\sum_{W \in neigh^{200}_{\mathtt{GC}}(W^R_i)} N_{W^R}} \tag{5.7}$$

Finally, we can compute the probability of WRV^S_i being generated from an event with fetal

allele copy count $|PP|$ as:

$$\mathcal{N}(WRV_i^{R,|PP|} - WRV_i^S; \mu = 0, \sigma_{\text{noise}}) \tag{5.8}$$

where we model the difference between WRV_i^S and WRV_i^R as a Gaussian noise with zero mean and empirically estimated variance σ_{noise}.

By normalizing the probabilities of WRV_i^S w.r.t. all phased patterns, we obtain priors for each phased pattern that are used in the HMM described in the next section.

As a reference plasma sequencing coverage we use plasma sample of the G1 trio of Kitzman et al. [120] dataset, as the overall coverages observed in corresponding bins between the two samples correlate well (mean absolute error of WRVs being 0.000347, see Figure 5.2(a)). Since coverage variation of cfDNA from plasma has much wider distribution than standard WGS, a sample from other plasma is more suitable than the same trio maternal sample (see Figure 5.2(b)) for purpose of coverage distribution reference in our model. Availability of additional plasma datasets would enable us to further improve the accuracy of the reference bins.

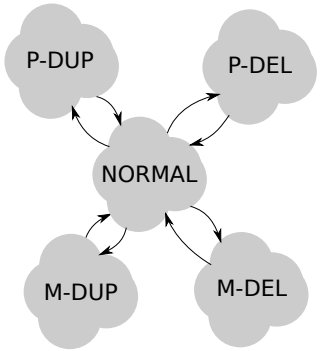
Note that compared to previous methods we use significantly smaller windows: ~1kb versus 100kb-1Mb used by other methods [27, 190]. As mentioned earlier, our goal here is not to detect CNVs immediately, but to rather compute a probability distribution over the number of haplotypes the fetus has inherited, which are used as priors in the more complex model. Due to the independence assumptions inherent in the HMM we want these priors, applied at each state, to be (approximately) independent, and hence we picked non-overlapping windows each containing one SNP locus.

5.2.3 Hidden Markov Model for CNV Inference

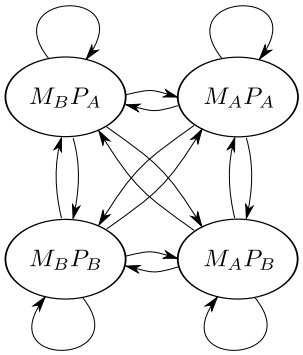
To combine the signals from individual SNP positions, we use an HMM with 20 states corresponding to modelled phased inheritance patterns (Figure 5.3). That means each sate represents a possible set of parental haplotypes inherited by the fetus. States representing normal inheritance are central to the model assuming that two CNVs cannot be immediately subsequent. Between states of the same inheritance pattern, we allow for transitions reflecting either recombinations or errors in phasing. For each state, the emissions are the counts of individual alleles in reads mapped to that particular SNP position. The probability of the observed emission is the probability of such allele counts in the expected allele distribution conditional on phased inheritance pattern as described above in Section 5.2.1.

To incorporate the coverage information, for each SNP position we multiply the transition probabilities into the state by the copy number priors obtained in Section 5.2.2. Specifically, each edge incoming to a state is multiplied by the corresponding prior of inheriting that many haplotypes, which are then normalized so that the sum of the probabilities leaving each state is one.

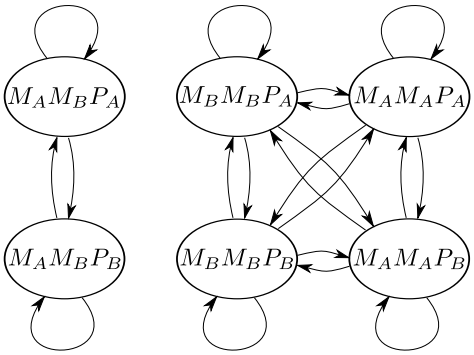
The transition probabilities within an event type (e.g. maternal duplication) were set to 0.01,



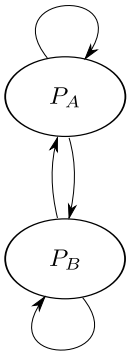
(a) Overall HMM architecture



(b) States for normal inheritance



(c) States for maternal duplication



(d) States for maternal deletion

Figure 5.3: Hidden Markov model used for CNV inference. (a) High-level architecture of the HMM with 5 sets of states corresponding to 5 types of fetal inheritance. Note, we do not allow two CNVs to be adjacent, thus switching between two CNVs always has to go through a normal inheritance state. Edges in (a) represent edges coming in/out of all states between two sets of states. (b-d) Correspond to the diagram of states of the HMM within the normal inheritance, maternal duplication, and maternal deletion states of (a). Paternal duplications/deletions are analogous to (c) and (d). Inner edges in (b-d) serve to model errors in phasing or recombination events.

to reflect expected haplotype block lengths of several hundred SNPs. Further, the transition probability for starting a CNV was set to one in ten thousand SNP loci (0.0001) with length expected to span approximately one thousand SNPs (i.e. transition probability back to normal inheritance was set to 0.001).

5.2.4 CNV Simulation *in silico*

To evaluate the accuracy of our CNV discovery algorithm we created simulated datasets with CNVs of various sizes inserted into the sequenced plasma. While previous approaches have used simple Poisson modelling of the coverage of cfDNA for simulation purposes [27], we propose a more elaborate model to more accurately model the extremely uneven coverage that we observe in

cfDNA samples (Figure 5.1). Our simulation performs the deletion or duplication of a particular fetal allele. We need to resolve the haplotypes of every individual in the trio, to correctly add or remove reads originating from a target haplotype of the CNV event. Similarly to our detection method (described in Results, below), we used Beagle 4 [24] with 1000 Genomes Project reference haplotypes, however we also use the fetal genome sequenced after delivery, and utilize pedigree information to phase each individual in the trio.

In order to simulate a duplication, of either maternal or paternal origin, we used the parental DNA sequencing data from the family trio data set. First, we filtered for reads mapping to the intended region of duplication that also match the target haplotype of the parent according to the parental phasing. In case of reads not uniquely mapping to either of the two parental haplotypes, i.e. the read mapped to a region without any heterozygous SNP locus, the read was selected randomly with probability 0.5. Subsequently, the filtered reads were uniformly down-sampled according to fetal DNA mixture ratio and the original plasma DOC in this region to match the expected number of reads derived from a single fetal haplotype in plasma sequencing. Resulting reads were then mixed together with original plasma reads to create a plasma sample containing the desired duplication in the fetal genome.

To simulate a deletion, we first identified a fetal haplotype inherited from the parent of choice, which was to be deleted. We filtered the plasma sample removing reads coming from this target fetal haplotype. That is, each read mapped to the intended deletion region was removed with probability of belonging to the fetus and also being inherited from the intended parent. In order to find this probability we used the phasing to check which maternal and fetal haplotypes match the SNPs in the read. If none of the four haplotypes matched the read, we removed the read with probability $r/2$ where r is the fetal DNA admixture ratio. If the fetal target haplotype matched the read, it was removed with probability

$$\frac{r/2}{N_{\mathrm{m}} \cdot (1-r)/2 + N_{\mathrm{f}} \cdot r/2} \tag{5.9}$$

where $0 < N_{\mathrm{f}} \leq 2$ and $0 \leq N_{\mathrm{m}} \leq 2$ are respectively the number of fetal and maternal haplotypes that matched the read.

We also simulated plasma data sets with decreased fetal DNA mixture ratio. In order to achieve a desired down-rated admixture ratio r' in our plasma sample, we had to remove appropriate number of reads coming from the fetal DNA. First, we have computed the appropriate fraction of fetal-origin reads, w.r.t. original admixture ratio r, to be removed from the plasma as

$$r_{del} = 1 - \frac{1-r}{r} \cdot \frac{r'}{1-r'} \tag{5.10}$$

Similarly to simulation of a deletion, we have then filtered the plasma reads for reads originating from the fetal genome. Since this cannot be decided without ambiguity, we estimated the

Table 5.1: Summary of mother-father-child trio I1 sequencing data [120].

Individual	Sample	DOC
Mother (I1-M)	Plasma (5 ml, gestational age 18.5 weeks)	78
	Whole blood (< 1 ml)	32
Father (I1-P)	Saliva	39
Child (I1-C)	Cord blood at delivery	40

corresponding probability p_{f}:

$$p_{\mathrm{f}}(seq) = \begin{cases} \dfrac{N_{\mathrm{f}} \cdot r/2}{N_{\mathrm{m}} \cdot (1-r)/2 + N_{\mathrm{f}} \cdot r/2} & \text{iff } N_{\mathrm{m}} + N_{\mathrm{f}} > 0 \\ r & \text{iff } N_{\mathrm{m}} + N_{\mathrm{f}} = 0 \end{cases} \tag{5.11}$$

where N_{f} and N_{m}, as above, are the number of fetal and maternal haplotypes that match SNP alleles of the read. Thus a read was then removed with probability equal to

$$r_{del} \cdot p_{\mathrm{f}}(seq) \tag{5.12}$$

5.3 Results

5.3.1 Datasets and Processing

In our experiments, we used whole genome sequencing data of two mother-father-child trios I1 (Table 5.1), and G1, obtained and published by Kitzman et al. [120]. In our experiments we mainly used the first trio I1 with 13% fetal admixture in obtained plasma. For maternal, paternal, and plasma datasets the reads were aligned to the hg19 genome using BWA. We genotyped both the parents using Samtools and Vcftools. To improve the precision of genotyping we only consider variants at positions previously identified as variable within the 1000 Genomes Project. Subsequently we phased the haplotypes using Beagle 4 [24] with reference haplotype panels from 1000 Genomes Project.

5.3.2 Evaluation

We simulated 360 CNVs in I1 plasma to evaluate our method's recall, while G1 plasma sample served as a reference in DOC-based CNV estimation as described in Section 5.2.2. For each test case, we picked a random position in chromosome 1, outside known centromere and telomeres regions, to place the simulated CNV. Our simulation methods are described in detail in Section 5.2.4. We then ran our algorithm on a genomic window starting 20Mb before the simulated CNV and ending 20Mb after the CNV. The results are shown in Table 5.2. We acknowledge a CNV as correctly called if CNV predictions of the same type span at least 50% of it, while precision is computed as the fraction of correct CNV calls over all calls of that category.

Table 5.2: Summary of recall on test set composed of 360 *in silico* simulated CNVs in I1 maternal plasma samples with 13% and 10% fetal admixture ratio. The 'ratios only' column corresponds to the method that only uses allelic ratios, but not the coverage prior. In such cases both the precision and recall are mostly dominated by the model combining both signals. (We write 'NA' in a precision field if no call of such CNV category was predicted by the model).

mixture ratio	length		Paternal Del (20)		Paternal Dup (40)		Maternal Del (20)		Maternal Dup (40)	
			ratios only	combined	ratios only	combined	ratios only	combined	ratios only	combined
13%	50k - 400k	recall	55	60	55	60	10	15	25	22
		precision	73	52	25	75	67	100	2	2
	400k - 3M	recall	100	100	98	98	30	40	73	78
		precision	100	100	100	100	86	100	23	89
	>3M	recall	95	100	93	95	95	100	100	100
		precision	100	100	100	100	100	100	100	100
10%	50k - 400k	recall	50	45	48	48	0	0	15	15
		precision	71	69	23	30	NA	NA	2	2
	400k - 3M	recall	100	100	90	92	5	20	38	45
		precision	100	100	95	100	100	80	10	16
	>3M	recall	95	95	100	100	45	40	93	88
		precision	100	100	100	100	100	100	97	97

To evaluate the effect the admixture has on accuracy, we repeated this experiment not only with the original plasma dataset, but also once down-sampled to only contain 10% admixture.

The results indicate that our method can achieve nearly perfect recall and precision for variants > 3 megabases, and promising results down to CNVs of 400 kilobases. Maternally inherited events are typically more difficult to identify than paternally inherited ones, and deletions more difficult than duplications, possibly due to complete dropout of fetal alleles due to reduced admixture.

To evaluate power of individual signals utilized by our unified model, we also tested models that take into consideration only either the allelic ratios or coverage information. The allelic ratios only model is as described above in Section 5.2.3 but without multiplying of copy number prior in the transition probabilities. Obtained results are shown together with the results of the unified model in Table 5.2.

For predicting fetal CNVs based solely on coverage information we split the sample to bins of uniform size and computed WRVs for each, following the work of Srinivasan et al. [190]. We then ran a simple HMM with 3 states corresponding to normal inheritance, duplication, and deletion, respectively. The WRVs in bins were interpreted as emissions and the emission distributions were computed as described in Section 5.2.2, Equation 5.8. We tested the HMM with bin sizes of 100kb and 300kb, and the results are summarized in Table 5.3. Using larger bins limit resolution of the method, e.g. in case of 300kb bins the obtained recall on $<$ 400kb CNVs is (close to) zero. On the other hand for large CNVs $>$ 3Mb using 300kb bin size mostly improves upon 100kb bins in terms of both recall and precision.

Note, that a direct comparison to the methods of Chen et al. [27] and Srinivasan et al. [190] is not possible, as they are tailored to low coverage plasma sequencing data and require a large

Table 5.3: Summary of results obtained by an HMM using only WRV signal. The same test set composed of 360 *in silico* simulated CNVs was used as in Table 5.2. We ran the model with 100kb, and 300kb bin sizes. (We write 'NA' in a precision field if no call of such CNV category was predicted by the model).

mixture ratio	length	bin size ->	Paternal Del (20) 100kb	Paternal Del (20) 300kb	Paternal Dup (40) 100kb	Paternal Dup (40) 300kb	Maternal Del (20) 100kb	Maternal Del (20) 300kb	Maternal Dup (40) 100kb	Maternal Dup (40) 300kb
13%	50k - 400k	recall	5	0	22	0	20	0	8	0
		precision	5	0	100	NA	2	0	100	NA
	400k - 3M	recall	75	25	50	20	75	25	30	18
		precision	35	21	100	100	11	5	100	100
	>3M	recall	75	75	50	55	89	80	32	55
		precision	37	94	100	100	81	48	100	100
10%	50k - 400k	recall	5	0	18	0	10	0	8	2
		precision	8	0	100	NA	2	0	100	100
	400k - 3M	recall	60	20	32	18	60	15	18	8
		precision	26	9	100	100	6	3	100	100
	>3M	recall	75	60	45	45	85	70	22	45
		precision	25	86	100	100	40	24	100	100

Table 5.4: In silico recall and number of CNVs of various sizes generated in a genome-wide run. For each CNV size we also show (in parenthesis) the number of calls that are from at least 50% overlapped by CNVnator [1] calls on the fetal, maternal, and paternal genomes, respectively.

Combined Model		50-200k	200-400k	400-750K	750k-3M	3M-7.5M	10M+
in silico CNV recall	Maternal origin	0%	40%	57%	73%	100%	100%
	Paternal origin	43%	77%	100%	97%	93%	100%
WG calls and their (F, M, P) overlap		82 (7, 8, 4)	26 (2, 3, 2)	9 (1, 1, 0)	4 (2, 1, 2)	0 (-)	0 (-)

number of control plasma samples to evaluate significance of observed coverage variation in the studied plasma sample for CNV calling.

To further test precision of our combined method, we ran our combined model on the whole plasma dataset (expected to contain no large de-novo variants) and observed the number of CNV calls for each size. These numbers are shown in Table 5.4, with *in silico* accuracy for each length shown for comparison. Notably, a large fraction of the larger false positive calls correspond to CNVs already present in parents (and hence inherited, rather than de novo).

5.3.3 Implementation Note

Our model is implemented in the Python programming language with the PyPy interpreter. When ran on a whole genome dataset our implementation required up to 20GB of system memory and took less than 4 hours of single thread CPU time to finish. The implementation is available at http://github.com/compbio-UofT/fCNV.

5.4 Discussion

In this chapter we introduced a novel probabilistic method for the identification of *de novo* Copy Number Variants from maternal blood plasma sequencing with largely increased sensitivity compared to methods published so far. Our method combines three types of data: allelic ratios, reflecting the changes in the expected observations of various alleles at SNP positions in the presence of the CNV; phasing information, allowing for the combining of allelic ratios across multiple SNP positions, thus improving the signal-to-noise ratio; and depth of coverage information reflecting the change in expected sequencing depth in the presence of the CNV. We applied the resulting method to simulated sequencing data, demonstrating promising results for CNVs > 400 kilobases in length, and especially for CNVs of paternal origin.

Simultaneously, we believe our method can be further improved in several ways. First, our approach of modelling the depth of coverage prior using small windows is likely suboptimal. Especially because the method is searching for larger CNVs, using larger windows would be advantageous; however in this case the observations of coverage at adjacent SNPs would no longer be independent, and thus not properly modelled as an HMM. We believe a more expressive model that is able to model such interactions between coverage terms would improve upon the current results. Secondly, our method does not directly model potential inherited CNVs in the father (maternally inherited CNVs are modelled through the use of maternal priors at each position). Explicitly pre-computing and utilizing information about these inherited CNVs is likely to reduce the false positive rate of ours and related methods. Thirdly, we incorporated the coverage signal in our model by comparing the observed WRV with the corresponding WRV in a reference plasma sample (G1 in our experiments). Using multiple plasma references would reduce individual-specific biases, thus improve the overall performance.

The main limitation of our method in practice is the need for deep maternal plasma cfDNA sequencing in order to exploit the allelic ratios signal. Note that the parental genome WGS could be replaced by genotyping using SNP arrays, however the need for a paternal sample is a limitation for broad clinical use.

Chapter 6

Conclusion

Identifying the best treatment for cancer patients that maximizes their chances of successful recovery remains a largely unresolved problem for many cancer types and both cytotoxic and targeted treatments. In this thesis I developed a machine learning model, Dr.VAE, that improves accuracy of drug response prediction over the currently used models. The drug response variational autoencoder, described in Chapter 3, leverages advancements of deep learning recently introduced in the area of latent-variable generative modeling. Neural network aspect of the model facilitated greatly increased expressiveness and effective feature extraction in the model. On the other hand, the latent-variable aspect enabled integration of drug-induced perturbation profiles, a resource not fully utilized before, and further enabled semi-supervised training of the model.

However my Dr.VAE model is limited to training and prediction in pre-clinical datasets of cancer cell line drug sensitivity. Clinical datasets mapping genomic profiles to oncological treatment response are currently of insufficient size for training of such deep learning models. Therefore approaches of domain adaptation or transfer learning are required to make the pre-clinical models clinically valuable. To this end, I assessed applicability of domain-invariant representation learning, a popular domain adaptation approach, to the drug response prediction task. In my analysis, presented in Chapter 4, I came to the conclusion that necessary conditions of successful domain adaptation via the means of learning domain-invariant representation are often not satisfied in the available datasets. This analysis brings insight into why homogenization of cell lines and patient cohorts does not necessary lead to improved clinical prediction performance, a result also empirically observed by Zhao et al. [221].

To improve chances of successful unsupervised domain adaptation or transfer learning, the research community likely needs to move towards high-throughput screening of biologically more accurate pre-clinical cancer models. While cell lines enabled discoveries of several important univariate biomarkers and facilitated successful drug repositioning (Section 2.4.1), it is possible that there this biological model reaches limits of its utility. A biologically more accurate model of cancer is necessary to discover more complex biomarkers and to train accurate powerful machine learning models. We need to study anti-cancer drug sensitivity and there-by induced phenotype

perturbations in an environment as close to *in vivo* as possible yet feasible to scale towards acquisition of massive datasets. Organoids are a promising step in this direction. Patient derived xenografts, while currently the most accurate pre-clinical models, have inherent drawback of cost, scalability and consistency. The high-throughput aspect is crucial for application and development of powerful machine learning methods. Last but not least, special care has to be taken to adequately match treatment outcome measures between a pre-clinical (source) and clinical (target) domain, as well as cancer type and subtype composition of the datasets.

Pre-clinical models however cannot completely substitute clinical trial data. Labeled clinical datasets will always be needed to evaluate transferability of a model into the clinic. Clinical data should, to largest extent possible, be also included in training of the models. As I observed in my experiments in Chapter 4, training in source + target domain (ST2T) regime, that is using both cell line and patient datapoints in training, leads to the best target domain performance if the source domain dataset is at all helpful. Therefore approaches of transfer learning (semi-supervised domain adaptation) will likely yield the best clinical predictive models.

In case of limited dataset sizes, it can be beneficial to incorporate biological priors, as I did in Chapter 3. Many other priors can be used, perhaps most straightforward is feature (genes, variants, etc.) preselection to increase power of a prediction method. In my studies I selected the L1000 landmark genes [192], but based on mechanism of action of a particular drug, a more relevant, yet smaller, set of genes could be selected. Another promising avenue for improving drug response prediction in limited datasets is to leverage similarity of anti-cancer drugs in their mechanism of action. Here, machine learning approaches for transfer learning and multi-task learning could deliver a further improvement in prediction accuracy as already shown in several prior studies [34, 4, 183].

The tissue type of cancer's primary site is often correlated with the response outcome. That shows that for many existing treatments, the response is considerably tissue-specific. Therefore it can be promising to train tissue-specific prediction models if the dataset sizes allow it. Unfortunately current cell line datasets provide more than 200 cancer samples in only two tissue types: lung tissue and blood cancers (combined haematopoietic and lymphoid tissue).

Next, as I alluded in Section 2.4.3, efficient monotherapies are unlikely to exist for all cancers and all patients. Combination therapies can target multiple aberrant cancer processes at once by combination of several synergistic drugs and achieve higher efficacy and/or lower toxicity. In fact many of currently used cancer treatments are predetermined "drug cocktails" [100]. Many expect that future progress in cancer treatment will come from discoveries of new combination therapies [153] and ultimately personalized predictions of treatment combinations that account for intra-tumoural heterogeneity and attack all cancer subclones at the same time.

In this thesis I also contributed to the area of non-invasive prenatal testing based on sequencing of circulating cell-free DNA shed from the placenta into the maternal circulation during pregnancy. Here presented method was the first proof-of-concept for sub-chromosomal CNV detection. It was later followed by Arbabi et al. [6] who leveraged characteristics of placenta-derived cfDNA

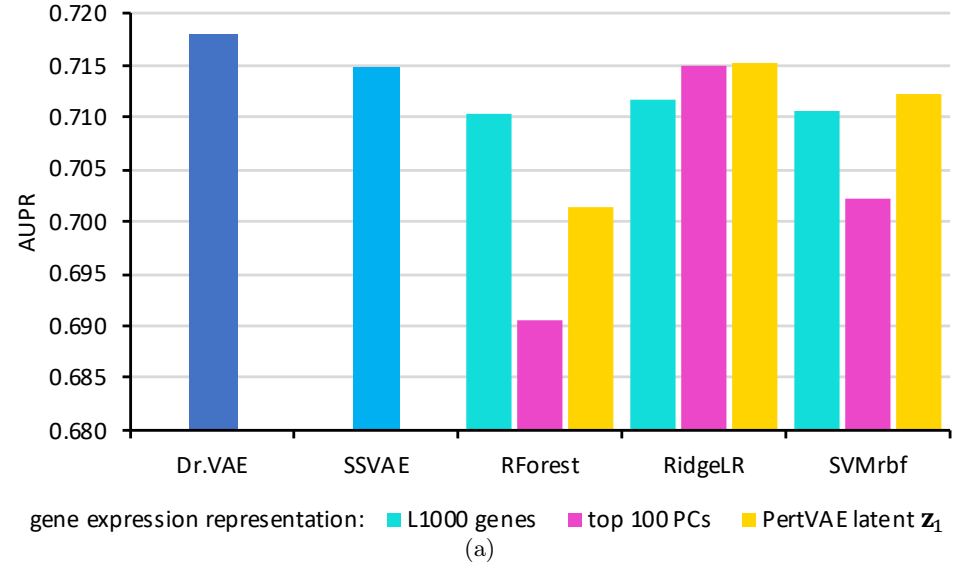
fragment size distribution to detect even shorter CNVs. Progress in this area has led to routinely available commercial screening tests for aneuploidy and major sub-chromosomal aberrations. Despite the advancements in underlying biotechnology, all commercially available tests remain *screening* tests and to establish a diagnosis, an invasive test is still necessary [17].

Exciting is the future development in a related area – liquid biopsy of cancer [177, 155], which consists of detection and analysis of circulating tumour cells and DNA fragments released by tumours into the patient's circulation. Feasibility of this approach is dependent on development of reliable and cost-effective biotechnology, which has been a hurdle until very recently [44, 93]. Emergence of this technology will continually enable a wide range of clinical applications in cancer screening, diagnosis, prognosis, and personalized treatment selection [155]. I hope that biotechnologies, such as liquid biopsy, will lead to generation of massive-scale clinical datasets of genomics in cancer, including progressive monitoring of the tumour development in response to treatment, and will eventually facilitate development of accurate machine learning models for precision medicine in cancer that are fully personalized and adaptive.

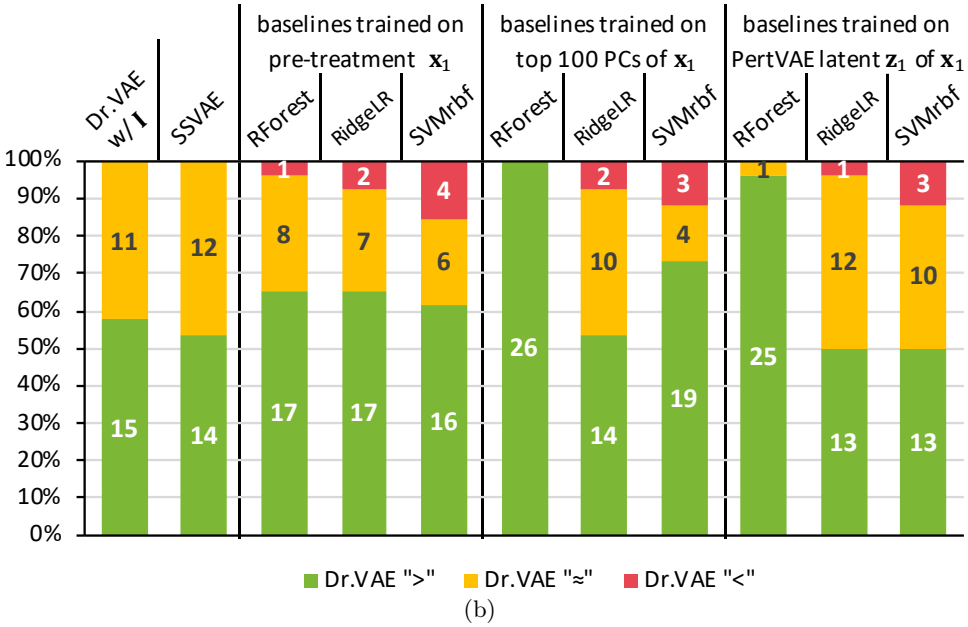
Appendices

Appendix A

Supplementary figures and tables for Chapter 3 (Dr.VAE)



(a)



(b)

Figure A.1: **Summarized classification results (AUPR).** (a) Area under the precision-recall curve of Dr.VAE and baseline methods. Shown is average over 26 drugs, each evaluated in 100 train-validation-test splits. (b) Dr.VAE is comparable or better than any other baseline for >85% of the drugs (p-val < 0.05 Wilcoxon test).

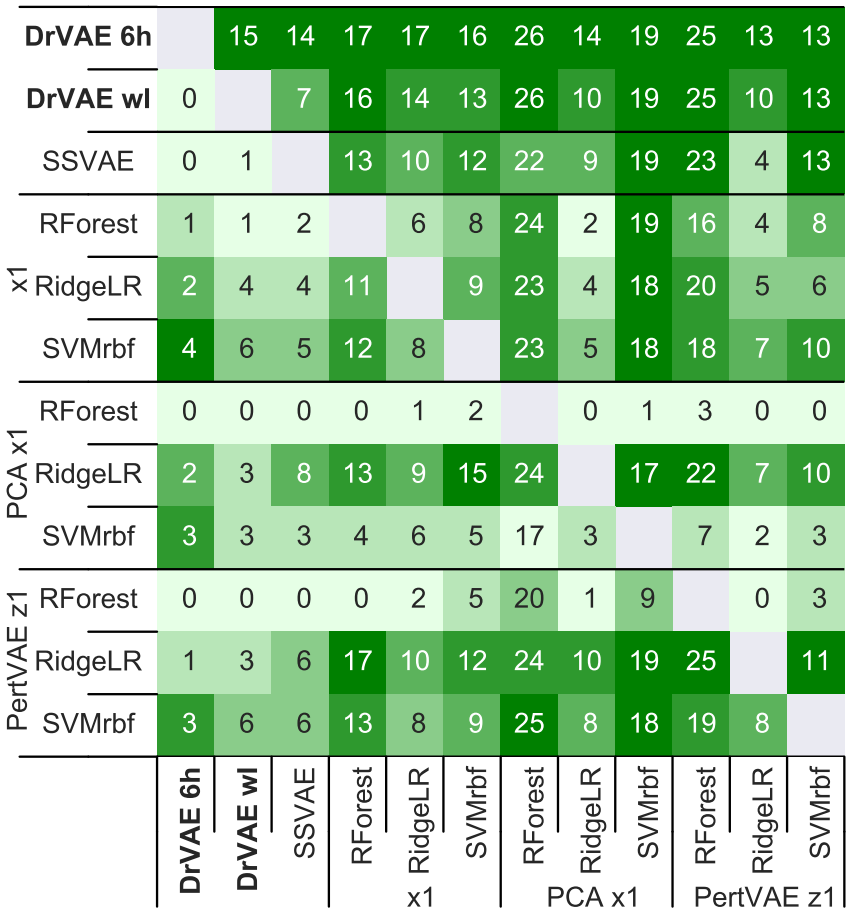


Figure A.2: **All to all comparison of tested methods (AUPR).** Comparison of all tested methods by one-sided Wilcoxon Signed-Rank Test (p-val < 0.05) based on test area under the precision-recall curve performance in 100 train-validation-test splits. Cell at (i, j) position shows the number of drugs for which a method in row i outperforms the method corresponding to j-th column. Analogous to Figure 3.4 in the main text that presents comparison by test AUROC.

Table A.1: **Per-drug AUROC classification results.** Cross-validated test AUROC (area under the ROC curve) of our Dr.VAE to SSVAE and other classification models detailed for each evaluated drug. Methods including PCA and PertVAE are 2-step methods: (i) fit the unsupervised model, (ii) use latent representation to fit a standard classifier.

AUROC				baselines trained on pre-treatment $\mathbf{x}_1$			baselines trained on top 100 PCs of $\mathbf{x}_1$			baselines trained on PertVAE latent $\mathbf{z}_1$ of $\mathbf{x}_1$		
drug	**Dr.VAE (6h)**	**Dr.VAE w/ I**	SSVAE	RForest	RidgeLR	SVMrbf	RForest	RidgeLR	SVMrbf	RForest	RidgeLR	SVMrbf
bortezomib	0.670	0.669	0.663	0.659	0.649	0.641	0.629	0.648	0.644	0.649	0.666	0.669
bosutinib	0.743	0.743	0.736	0.737	0.750	0.735	0.714	0.758	0.749	0.729	0.747	0.735
ciclosporin	0.638	0.637	0.642	0.632	0.638	0.637	0.622	0.636	0.640	0.619	0.631	0.633
clofarabine	0.757	0.755	0.754	0.746	0.745	0.749	0.723	0.746	0.738	0.742	0.754	0.754
dasatinib	0.775	0.774	0.773	0.764	0.771	0.778	0.768	0.776	0.784	0.756	0.772	0.775
decitabine	0.840	0.838	0.838	0.827	0.832	0.822	0.803	0.838	0.810	0.818	0.839	0.820
docetaxel	0.733	0.729	0.724	0.735	0.719	0.711	0.669	0.725	0.696	0.712	0.725	0.724
etoposide	0.866	0.863	0.868	0.847	0.869	0.874	0.832	0.867	0.843	0.846	0.864	0.866
fluvastatin	0.650	0.648	0.649	0.657	0.660	0.672	0.643	0.654	0.651	0.642	0.647	0.652
fulvestrant	0.561	0.559	0.549	0.578	0.551	0.549	0.550	0.554	0.552	0.551	0.545	0.547
gemcitabine	0.734	0.734	0.728	0.726	0.721	0.732	0.715	0.734	0.721	0.723	0.726	0.729
lovastatin	0.651	0.651	0.645	0.647	0.646	0.650	0.619	0.645	0.631	0.639	0.655	0.651
mitomycin	0.734	0.731	0.732	0.729	0.694	0.733	0.716	0.723	0.728	0.722	0.715	0.738
niclosamide	0.687	0.688	0.683	0.678	0.670	0.669	0.659	0.675	0.677	0.666	0.669	0.685
omacetaxine me	0.735	0.730	0.725	0.709	0.729	0.726	0.673	0.732	0.718	0.708	0.737	0.738
paclitaxel	0.753	0.751	0.749	0.744	0.729	0.735	0.723	0.739	0.732	0.739	0.746	0.753
PLX-4032	0.583	0.583	0.585	0.577	0.582	0.559	0.570	0.582	0.579	0.574	0.591	0.592
prochlorperazine	0.609	0.608	0.603	0.593	0.610	0.609	0.577	0.614	0.603	0.595	0.605	0.594
sirolimus	0.670	0.670	0.672	0.668	0.645	0.654	0.652	0.656	0.665	0.665	0.670	0.683
sitagliptin	0.582	0.581	0.586	0.556	0.561	0.564	0.554	0.566	0.572	0.565	0.569	0.594
teniposide	0.739	0.737	0.737	0.744	0.729	0.737	0.708	0.732	0.710	0.725	0.726	0.725
topotecan	0.752	0.748	0.752	0.738	0.743	0.748	0.726	0.742	0.734	0.737	0.746	0.743
trifluoperazine	0.659	0.657	0.651	0.676	0.658	0.648	0.635	0.661	0.647	0.640	0.666	0.647
valdecoxib	0.708	0.707	0.704	0.703	0.708	0.674	0.689	0.707	0.689	0.690	0.706	0.688
vincristine	0.788	0.786	0.792	0.777	0.773	0.786	0.765	0.779	0.777	0.778	0.786	0.794
vorinostat	0.743	0.742	0.734	0.732	0.738	0.731	0.714	0.741	0.725	0.724	0.739	0.726
MEAN	0.706	0.704	0.703	0.699	0.697	0.697	0.679	0.701	0.693	0.691	0.702	0.702

Table A.2: **Per-drug statistical comparison of Dr.VAE to other evaluated methods by AUROC**. Statistical comparison of Dr.VAE to a set of evaluated baseline models on basis of their test AUROC on 100 data splits. Shown is p-value of one-sided Wilcoxon Signed-Rank Test rejecting null hypothesis of "Dr.VAE performance is worse or no different from the compared model performance" in favor of alternative hypothesis "Dr.VAE mean performance is higher than the compared baseline model". In Dr.VAE column, the mean test AUROC is shown.

AUROC				baselines trained on pre-treatment $\mathbf{x}_1$			baselines trained on top 100 PCs of $\mathbf{x}_1$			baselines trained on PertVAE latent $\mathbf{z}_1$ of $\mathbf{x}_1$		
drug	**Dr.VAE (6h)**	**Dr.VAE w/ I**	SSVAE	RForest	RidgeLR	SVMrbf	RForest	RidgeLR	SVMrbf	RForest	RidgeLR	SVMrbf
bortezomib	0.670	6.2E-03	1.9E-03	1.3E-03	4.5E-09	4.3E-12	5.5E-13	1.7E-10	2.9E-10	5.1E-07	1.5E-01	4.3E-01
bosutinib	0.743	1.2E-02	3.4E-04	3.4E-02	1.0E+00	4.2E-03	8.4E-11	1.0E+00	9.4E-01	2.0E-05	9.0E-01	2.7E-03
ciclosporin	0.638	1.5E-01	8.9E-01	1.2E-01	4.5E-01	4.8E-01	4.9E-04	3.8E-01	8.5E-01	4.2E-04	9.1E-02	2.5E-01
clofarabine	0.757	2.1E-06	8.8E-02	2.9E-05	6.1E-04	4.8E-05	1.7E-15	2.7E-06	3.0E-11	2.5E-08	1.9E-01	6.3E-02
dasatinib	0.775	1.9E-04	9.4E-02	1.9E-06	5.4E-02	9.7E-01	1.3E-03	7.2E-01	1.0E+00	4.9E-11	1.3E-02	1.4E-01
decitabine	0.840	6.0E-07	2.8E-01	1.0E-07	6.9E-05	2.6E-11	3.1E-15	4.9E-02	4.9E-16	1.6E-12	3.6E-01	2.9E-14
docetaxel	0.733	8.5E-06	2.1E-03	5.2E-01	2.0E-03	9.2E-08	3.3E-15	2.5E-02	1.3E-11	1.4E-04	2.1E-02	7.7E-03
etoposide	0.866	6.1E-09	7.7E-01	3.6E-12	9.9E-01	1.0E+00	1.9E-16	6.1E-01	4.1E-14	5.8E-13	1.2E-01	4.5E-01
fluvastatin	0.650	7.7E-09	3.6E-01	9.8E-01	1.0E+00	1.0E+00	2.1E-02	8.5E-01	5.3E-01	3.2E-03	7.1E-02	7.1E-01
fulvestrant	0.561	1.3E-01	9.4E-02	9.8E-01	5.8E-02	8.4E-02	1.7E-01	1.7E-01	2.1E-01	8.9E-02	3.9E-02	4.1E-02
gemcitabine	0.734	4.1E-01	5.4E-03	8.0E-03	9.5E-04	3.0E-01	5.1E-07	6.8E-01	2.0E-06	1.2E-04	1.1E-03	1.2E-02
lovastatin	0.651	3.4E-01	6.4E-02	1.1E-01	1.3E-02	2.6E-01	3.4E-12	1.8E-03	3.0E-09	1.2E-03	9.6E-01	5.4E-01
mitomycin	0.734	5.1E-09	2.6E-01	8.4E-02	8.4E-13	5.5E-01	5.6E-06	2.8E-05	2.9E-02	5.0E-05	5.7E-10	9.8E-01
niclosamide	0.687	6.9E-01	1.6E-02	1.8E-03	5.2E-08	3.8E-07	7.8E-11	1.2E-04	4.9E-04	5.7E-08	6.4E-09	8.5E-02
omacetaxine me	0.735	5.3E-08	2.2E-04	1.5E-08	1.3E-01	5.5E-03	8.6E-16	5.2E-01	1.2E-06	1.4E-07	8.8E-01	8.8E-01
paclitaxel	0.753	2.9E-08	8.6E-02	6.2E-05	1.4E-09	1.1E-10	2.9E-13	4.3E-08	2.9E-10	1.2E-05	1.1E-03	5.7E-01
PLX-4032	0.583	5.3E-01	8.8E-01	8.7E-02	3.7E-01	8.8E-08	6.8E-03	4.5E-01	1.8E-01	2.8E-02	9.8E-01	9.9E-01
prochlorperazine	0.609	2.6E-02	3.3E-02	8.4E-07	6.7E-01	7.5E-01	1.0E-10	9.5E-01	1.4E-01	2.3E-04	8.5E-02	5.7E-07
sirolimus	0.670	5.3E-01	6.0E-01	3.5E-01	1.8E-08	1.8E-07	1.6E-05	4.6E-05	6.5E-02	7.9E-02	4.1E-01	1.0E+00
sitagliptin	0.582	3.6E-01	6.8E-01	2.3E-02	5.7E-03	4.4E-02	4.5E-03	3.6E-02	3.7E-01	1.4E-01	4.5E-02	9.7E-01
teniposide	0.739	1.3E-03	2.1E-01	9.0E-01	3.6E-02	4.2E-01	6.1E-07	1.5E-01	1.3E-09	3.3E-03	8.1E-04	5.5E-06
topotecan	0.752	4.0E-14	7.0E-01	1.2E-05	2.6E-04	9.3E-02	3.3E-13	5.0E-04	2.5E-09	2.7E-06	3.5E-03	3.9E-04
trifluoperazine	0.659	7.2E-03	6.8E-03	1.0E+00	5.1E-01	5.2E-03	2.0E-05	7.0E-01	4.1E-03	3.4E-05	9.7E-01	5.4E-03
valdecoxib	0.708	1.5E-01	1.0E-01	5.3E-02	3.2E-01	8.9E-15	3.9E-06	1.3E-01	3.9E-07	9.8E-08	2.1E-01	6.3E-10
vincristine	0.788	1.6E-08	9.8E-01	3.5E-05	1.6E-08	1.7E-01	4.0E-11	1.6E-05	6.2E-06	5.7E-05	8.9E-02	1.0E+00
vorinostat	0.743	4.7E-01	4.2E-04	1.2E-04	1.1E-01	7.6E-05	5.2E-13	3.8E-01	6.7E-08	6.0E-08	1.3E-01	5.6E-08
% Dr.VAE ">"		61.5%	34.6%	57.7%	53.8%	53.8%	96.2%	46.2%	65.4%	88.5%	38.5%	42.3%
% Dr.VAE "≈"		38.5%	61.5%	30.8%	34.6%	34.6%	3.8%	50.0%	30.8%	11.5%	50.0%	38.5%
% Dr.VAE "<"		0.0%	3.8%	11.5%	11.5%	11.5%	0.0%	3.8%	3.8%	0.0%	11.5%	19.2%

Table A.3: **Per-drug AUPR classification results.** Cross-validated test AUPR (area under PR curve) of our Dr.VAE to SSVAE and other classification models detailed for each evaluated drug. Methods including PCA and PertVAE are 2-step methods: (i) fit the unsupervised model, (ii) use latent representation to fit a standard classifier.

AUPR				baselines trained on pre-treatment $\mathbf{x}_1$			baselines trained on top 100 PCs of $\mathbf{x}_1$			baselines trained on PertVAE latent $\mathbf{z}_1$ of $\mathbf{x}_1$		
drug	**Dr.VAE (6h)**	**Dr.VAE w/ I**	SSVAE	RForest	RidgeLR	SVMrbf	RForest	RidgeLR	SVMrbf	Rforest	RidgeLR	SVMrbf
bortezomib	0.810	0.809	0.804	0.802	0.798	0.773	0.774	0.803	0.787	0.789	0.805	0.795
bosutinib	0.538	0.537	0.532	0.533	0.542	0.546	0.510	0.555	0.552	0.530	0.547	0.547
ciclosporin	0.380	0.378	0.384	0.361	0.371	0.389	0.351	0.363	0.375	0.342	0.377	0.373
clofarabine	0.823	0.822	0.820	0.811	0.813	0.814	0.795	0.816	0.803	0.806	0.820	0.815
dasatinib	0.830	0.829	0.827	0.820	0.823	0.832	0.823	0.829	0.841	0.816	0.825	0.828
decitabine	0.761	0.759	0.756	0.740	0.755	0.737	0.721	0.757	0.729	0.738	0.763	0.731
docetaxel	0.847	0.844	0.841	0.842	0.837	0.826	0.791	0.842	0.798	0.827	0.841	0.834
etoposide	0.758	0.753	0.758	0.742	0.759	0.768	0.734	0.755	0.736	0.741	0.753	0.762
fluvastatin	0.727	0.726	0.725	0.729	0.736	0.751	0.719	0.725	0.726	0.718	0.726	0.733
fulvestrant	0.371	0.370	0.366	0.399	0.387	0.379	0.349	0.394	0.370	0.363	0.380	0.379
gemcitabine	0.825	0.825	0.821	0.813	0.815	0.818	0.802	0.824	0.804	0.811	0.821	0.817
lovastatin	0.734	0.734	0.732	0.728	0.730	0.736	0.704	0.725	0.721	0.719	0.738	0.736
mitomycin	0.838	0.836	0.836	0.829	0.799	0.830	0.813	0.821	0.818	0.821	0.822	0.829
niclosamide	0.786	0.787	0.784	0.772	0.777	0.771	0.756	0.778	0.766	0.760	0.777	0.776
omacetaxine me	0.885	0.882	0.878	0.874	0.879	0.879	0.849	0.885	0.867	0.870	0.886	0.887
paclitaxel	0.848	0.847	0.844	0.842	0.833	0.823	0.819	0.839	0.809	0.833	0.844	0.840
PLX-4032	0.427	0.427	0.417	0.421	0.415	0.406	0.401	0.415	0.425	0.415	0.433	0.437
prochlorperazine	0.709	0.709	0.703	0.687	0.706	0.704	0.673	0.711	0.696	0.693	0.702	0.695
sirolimus	0.761	0.761	0.761	0.756	0.737	0.734	0.740	0.745	0.747	0.753	0.757	0.763
sitagliptin	0.396	0.395	0.400	0.388	0.400	0.404	0.354	0.402	0.364	0.376	0.392	0.404
teniposide	0.848	0.847	0.848	0.848	0.839	0.847	0.816	0.844	0.826	0.832	0.837	0.832
topotecan	0.859	0.856	0.857	0.847	0.852	0.849	0.835	0.852	0.833	0.842	0.855	0.842
trifluoperazine	0.351	0.348	0.342	0.357	0.355	0.357	0.338	0.357	0.366	0.327	0.344	0.350
valdecoxib	0.837	0.837	0.833	0.824	0.837	0.799	0.806	0.836	0.805	0.815	0.833	0.808
vincristine	0.856	0.853	0.855	0.848	0.844	0.853	0.840	0.851	0.843	0.848	0.854	0.859
vorinostat	0.865	0.865	0.860	0.859	0.863	0.849	0.840	0.864	0.849	0.850	0.862	0.844
MEAN	0.718	0.717	0.715	0.710	0.712	0.711	0.690	0.715	0.702	0.701	0.715	0.712

Table A.4: **Per-drug statistical comparison of Dr.VAE to other evaluated methods by AUPR**. Statistical comparison of Dr.VAE to a set of evaluated baseline models on basis of their test area under precision-recall curve on 100 data splits. Shown is p-value of one-sided Wilcoxon Signed-Rank Test rejecting null hypothesis of "Dr.VAE performance is worse or no different from the compared model performance" in favor of alternative hypothesis "Dr.VAE mean performance is higher than the compared model". In Dr.VAE column, the mean test AUPR is shown.

AUPR				baselines trained on pre-treatment $\mathbf{x}_1$			baselines trained on top 100 PCs of $\mathbf{x}_1$			baselines trained on PertVAE latent $\mathbf{z}_1$ of $\mathbf{x}_1$		
drug	**Dr.VAE (6h)**	**Dr.VAE w/ I**	SSVAE	RForest	RidgeLR	SVMrbf	RForest	RidgeLR	SVMrbf	RForest	RidgeLR	SVMrbf
bortezomib	0.810	1.5E-03	9.4E-05	1.1E-04	6.1E-07	1.8E-16	1.7E-15	3.7E-06	7.6E-12	1.8E-12	3.0E-02	1.8E-07
bosutinib	0.538	1.8E-01	2.8E-02	1.4E-01	9.4E-01	9.9E-01	2.1E-07	1.0E+00	1.0E+00	1.2E-02	9.9E-01	9.9E-01
ciclosporin	0.380	7.0E-04	8.2E-01	2.6E-04	3.0E-02	9.5E-01	2.5E-06	4.9E-04	2.8E-01	2.1E-10	3.3E-01	7.6E-02
clofarabine	0.823	3.1E-05	2.9E-02	1.1E-09	1.0E-04	8.6E-08	2.1E-16	6.0E-05	4.4E-14	5.0E-14	5.3E-02	2.3E-06
dasatinib	0.830	6.5E-03	3.5E-02	6.1E-05	2.4E-03	9.4E-01	2.0E-02	5.5E-01	1.0E+00	2.3E-07	6.2E-03	1.5E-01
decitabine	0.761	4.3E-04	1.3E-02	1.7E-11	1.4E-02	3.9E-11	9.8E-17	6.6E-03	4.9E-15	7.1E-12	8.3E-01	1.7E-15
docetaxel	0.847	2.4E-07	3.3E-03	1.1E-01	3.6E-03	8.9E-10	2.2E-16	4.6E-02	1.0E-16	1.3E-07	1.2E-02	1.6E-04
etoposide	0.758	6.8E-08	2.8E-01	1.1E-05	5.9E-01	1.0E+00	6.7E-09	5.4E-02	4.7E-08	2.7E-06	7.7E-03	8.0E-01
fluvastatin	0.727	2.3E-03	1.4E-01	5.2E-01	1.0E+00	1.0E+00	5.7E-03	2.8E-01	3.5E-01	5.4E-03	3.8E-01	9.3E-01
fulvestrant	0.371	8.8E-02	2.3E-01	1.0E+00	9.7E-01	9.3E-01	2.1E-02	9.9E-01	7.5E-01	2.3E-01	8.5E-01	9.3E-01
gemcitabine	0.825	3.8E-01	8.9E-03	3.4E-08	1.7E-03	4.4E-04	1.6E-12	6.3E-01	2.5E-11	4.2E-09	8.4E-03	1.4E-04
lovastatin	0.734	3.3E-01	2.7E-01	4.4E-02	3.0E-02	6.0E-01	4.2E-11	6.9E-05	1.2E-04	9.8E-06	9.4E-01	6.8E-01
mitomycin	0.838	5.3E-09	8.2E-02	7.5E-06	1.1E-15	3.3E-04	1.1E-13	1.9E-12	7.1E-12	9.9E-12	2.9E-12	1.8E-05
niclosamide	0.786	9.1E-01	1.6E-01	1.2E-05	1.2E-03	6.5E-06	1.0E-12	6.3E-03	3.4E-09	4.1E-11	5.4E-04	1.3E-04
omacetaxine me	0.885	6.2E-07	9.3E-05	8.6E-06	9.4E-03	2.8E-03	1.5E-15	6.8E-01	1.5E-09	8.3E-08	9.0E-01	8.6E-01
paclitaxel	0.848	4.2E-04	2.5E-02	2.9E-04	1.1E-08	5.1E-14	6.9E-15	2.4E-07	2.1E-17	3.2E-08	1.3E-02	8.5E-04
PLX-4032	0.427	2.2E-01	6.9E-03	6.9E-02	2.0E-03	1.8E-06	1.8E-07	5.7E-04	3.7E-01	2.2E-03	9.2E-01	9.9E-01
prochlorperazine	0.709	1.3E-01	1.7E-02	4.7E-10	2.6E-01	4.8E-02	3.9E-14	8.0E-01	2.2E-05	1.6E-06	6.8E-03	8.6E-05
sirolimus	0.761	2.4E-01	4.2E-01	5.3E-02	7.8E-11	4.3E-13	1.1E-07	1.6E-07	2.8E-05	1.2E-03	2.6E-02	7.2E-01
sitagliptin	0.396	2.3E-01	4.9E-01	9.6E-02	4.7E-01	7.8E-01	1.9E-05	6.0E-01	4.4E-03	2.4E-02	1.7E-01	9.2E-01
teniposide	0.848	1.3E-04	3.4E-01	4.9E-01	2.3E-03	2.6E-01	5.5E-11	4.5E-02	3.3E-11	2.0E-07	2.2E-06	1.1E-08
topotecan	0.859	5.6E-12	1.1E-01	3.1E-09	7.6E-05	1.6E-06	7.7E-15	1.7E-04	4.1E-16	2.0E-11	3.7E-03	1.2E-11
trifluoperazine	0.351	7.2E-03	2.8E-02	8.7E-01	6.5E-01	8.9E-01	1.5E-02	7.8E-01	1.0E+00	1.5E-04	1.4E-01	4.8E-01
valdecoxib	0.837	5.1E-01	1.2E-02	7.7E-10	3.3E-01	1.9E-17	5.6E-15	1.8E-01	4.9E-15	1.7E-13	1.0E-02	8.5E-17
vincristine	0.856	3.6E-12	4.6E-01	1.9E-05	2.5E-07	3.5E-02	2.0E-10	1.7E-02	2.8E-07	1.8E-05	1.1E-01	9.6E-01
vorinostat	0.865	7.3E-01	1.8E-03	7.2E-04	1.8E-01	7.9E-09	3.2E-15	4.7E-01	1.8E-10	2.1E-10	6.5E-02	2.0E-12
% Dr.VAE ">"		57.7%	53.8%	65.4%	65.4%	61.5%	100.0%	53.8%	73.1%	96.2%	50.0%	50.0%
% Dr.VAE "≈"		42.3%	46.2%	30.8%	26.9%	23.1%	0.0%	38.5%	15.4%	3.8%	46.2%	38.5%
% Dr.VAE "<"		0.0%	0.0%	3.8%	7.7%	15.4%	0.0%	7.7%	11.5%	0.0%	3.8%	11.5%

Table A.5: **Overall statistical comparison of Dr.VAE to other evaluated methods.** Wilcoxon signed-rank test p-values that the performance of Dr.VAE is overall better (greater) than that of a compared method in terms of their test AUROC and AUPR, respectively. The Wilcoxon paired test is conducted on the methods' average per-drug performance on the set of tested 26 drugs, i.e. comparing corresponding columns in Table A.1 and Table A.3, for AUROC and AUPR measure, respectively.

compared method		p-value that Dr.VAE performance is greater			
		uncorrected p-values		Bonferroni corrected p-vals	
		AUROC	AUPR	AUROC	AUPR
	DrVAE w/ I	9.36e-06	2.05e-05	1.03e-04	2.26e-04
	SSVAE	1.97e-03	1.48e-04	2.17e-02	1.63e-03
baselines trained on pre-treatment $\mathbf{x}_1$	RForest	1.97e-03	1.80e-04	2.17e-02	1.98e-03
	RidgeLR	2.18e-04	1.42e-03	2.40e-03	1.56e-02
	SVMrbf	4.19e-04	1.92e-02	4.61e-03	2.11e-01
baselines trained on top 100 PCs of $\mathbf{x}_1$	RForest	4.15e-06	4.15e-06	4.57e-05	4.57e-05
	RidgeLR	1.54e-03	1.08e-02	1.69e-02	1.19e-01
	SVMrbf	2.55e-05	7.32e-05	2.81e-04	8.05e-04
baselines trained on PertVAE latent $\mathbf{z}_1$ of $\mathbf{x}_1$	RForest	4.15e-06	4.15e-06	4.57e-05	4.57e-05
	RidgeLR	2.71e-03	8.20e-03	2.98e-02	9.02e-02
	SVMrbf	2.31e-02	2.17e-02	2.54e-01	2.39e-01

Table A.6: **Dataset summarization.** For each from the selected set of 26 FDA-approved drugs, this table shows the number of cell lines tested in CMap-L1000v1 for drug-induced perturbation effects on gene expression, the total number of extracted control-perturbation pairs including biological replicates, as well as number of these cell lines for which drug sensitivity was matched and retrieved from CTRPv2. Next, shown is effect-to-replicate variance ratio that quantifies signal-to-noise strength in the perturbation experiments. The last two columns show the number of drug-response-labeled samples (out of 927 total cell lines) and the ratio of positive responders in CTRPv2 drug sensitivity data set.

	CMap-L1000v1 perturbation data set								CTRPv2 sensitivity d.s.	
drug	labeled pairs	labeled unique CLs	unlabeled pairs	unlabeled unique CLs	total pairs	total unique CLs	effect/rep. variance ratio	effect/rep. silhouette score	number of labeled CLs	responder %
bortezomib	97	39	29	12	126	51	0.512	0.046	824	67.60%
bosutinib	26	7	14	6	40	13	0.209	-0.023	823	24.54%
ciclosporin	219	36	49	13	268	49	0.226	-0.026	808	20.67%
clofarabine	27	7	5	2	32	9	0.248	-0.004	854	58.43%
dasatinib	43	11	10	3	53	14	0.266	0.011	845	59.05%
decitabine	23	7	26	6	49	13	0.142	-0.044	849	25.91%
docetaxel	22	2	34	7	56	9	0.245	-0.046	422	64.69%
etoposide	25	6	13	5	38	11	0.356	0.018	840	26.90%
fluvastatin	44	7	16	6	60	13	0.215	-0.060	820	59.02%
fulvestrant	23	2	33	7	56	9	0.141	-0.022	206	25.24%
gemcitabine	74	28	50	23	124	51	0.366	-0.024	777	59.07%
lovastatin	66	9	33	7	99	16	0.180	-0.008	849	59.95%
mitomycin	64	7	11	2	75	9	0.237	0.008	838	60.14%
niclosamide	146	39	48	14	194	53	0.415	0.015	830	63.73%
omacetaxine mep	19	4	22	6	41	10	0.645	0.041	625	74.88%
paclitaxel	90	9	29	3	119	12	0.175	-0.017	827	64.81%
PLX-4032	138	41	37	13	175	54	0.356	-0.011	820	25.37%
prochlorperazine	55	7	6	2	61	9	0.049	-0.035	823	60.63%
sirolimus	330	44	87	16	417	60	0.262	-0.024	852	58.33%
sitagliptin	13	4	19	5	32	9	0.254	0.003	212	26.42%
teniposide	67	27	54	23	121	50	0.415	-0.017	410	63.17%
topotecan	42	6	8	2	50	8	0.494	0.118	855	64.56%
trifluoperazine	164	35	55	20	219	55	0.339	-0.035	782	21.10%
valdecoxib	91	36	28	13	119	49	0.373	-0.019	803	63.51%
vincristine	42	7	14	2	56	9	0.198	-0.007	845	60.83%
vorinostat	98	40	47	15	145	55	0.647	0.105	825	67.03%
MEAN	78.77	17.96	29.88	8.96	108.65	26.92	0.306	-0.002	740.92	50.98%

Table A.7: **Post-treatment expression prediction results.** Shown is prediction RMSE of full Dr.VAE model on post-treatment latent representation $\mathbf{z}_2$ and post-treatment gene expression $\mathbf{x}_2$ computed on training and validation sets, and the Δ improvement of full Dr.VAE over Dr.VAE with an identity function instead of learned perturbation function (denoted "Dr.VAE w/ I" in the main text) in these measures. Pearson correlation of these Δ improvements to data set statistics are shown in Table 3.1 of the main text.

RMSE	*training set*				*validation set*				CMap-L1000v1 stats	
drug	RMSE of predicted $\mathbf{z}_2$	Δ RMSE over $\mathbf{z}_2$ w.r.t. $\mathbf{I}$	RMSE of predicted $\mathbf{x}_2$	Δ RMSE over $\mathbf{x}_2$ w.r.t. $\mathbf{I}$	RMSE of predicted $\mathbf{z}_2$	Δ RMSE over $\mathbf{z}_2$ w.r.t. $\mathbf{I}$	RMSE of predicted $\mathbf{x}_2$	Δ RMSE over $\mathbf{x}_2$ w.r.t. $\mathbf{I}$	effect/rep. variance ratio	number of unique CLs in CMap
bortezomib	0.248	0.090	0.511	0.014	0.443	0.025	0.608	0.008	0.512	51
bosutinib	0.297	0.026	0.513	0.011	0.529	-0.075	0.675	-0.004	0.209	13
ciclosporin	0.326	0.039	0.525	0.005	0.455	-0.023	0.592	0.001	0.226	49
clofarabine	0.316	0.025	0.467	0.011	0.490	-0.040	0.622	-0.003	0.248	9
dasatinib	0.289	0.030	0.492	0.010	0.495	-0.036	0.640	0.001	0.266	14
decitabine	0.281	0.012	0.485	0.007	0.462	-0.045	0.613	-0.003	0.142	13
docetaxel	0.316	-0.013	0.518	0.005	0.503	-0.114	0.618	-0.003	0.245	9
etoposide	0.305	0.019	0.524	0.010	0.530	-0.077	0.687	-0.003	0.356	11
fluvastatin	0.288	0.004	0.530	0.005	0.448	-0.057	0.641	-0.002	0.215	13
fulvestrant	0.302	-0.007	0.507	0.004	0.406	-0.044	0.603	-0.002	0.141	9
gemcitabine	0.248	0.068	0.514	0.008	0.478	-0.009	0.618	-0.003	0.366	51
lovastatin	0.304	0.013	0.515	0.004	0.489	-0.065	0.634	-0.001	0.180	16
mitomycin	0.342	0.005	0.500	0.008	0.500	-0.070	0.604	-0.001	0.237	9
niclosamide	0.287	0.056	0.550	0.009	0.436	-0.009	0.636	0.004	0.415	53
omacetaxine mep	0.283	0.089	0.531	0.031	0.522	-0.001	0.753	0.008	0.645	10
paclitaxel	0.316	0.000	0.509	0.004	0.505	-0.118	0.625	-0.002	0.175	12
PLX-4032	0.270	0.051	0.538	0.007	0.429	-0.013	0.619	0.002	0.356	54
prochlorperazine	0.303	-0.004	0.501	0.005	0.417	-0.065	0.593	-0.001	0.049	9
sirolimus	0.321	0.040	0.559	0.005	0.463	-0.031	0.621	0.002	0.262	60
sitagliptin	0.306	0.006	0.448	0.007	0.498	-0.060	0.581	-0.004	0.254	9
teniposide	0.240	0.083	0.506	0.012	0.439	0.011	0.610	0.002	0.415	50
topotecan	0.363	0.059	0.570	0.031	0.663	-0.072	0.774	-0.004	0.494	8
trifluoperazine	0.299	0.032	0.535	0.005	0.467	-0.046	0.611	-0.002	0.339	55
valdecoxib	0.239	0.065	0.514	0.007	0.447	-0.004	0.611	-0.001	0.373	49
vincristine	0.347	-0.001	0.505	0.008	0.519	-0.095	0.625	-0.003	0.198	9
vorinostat	0.256	0.111	0.511	0.018	0.468	0.040	0.617	0.013	0.647	55
MEAN	0.296	0.035	0.514	0.010	0.481	-0.042	0.632	0.000	0.306	26.92

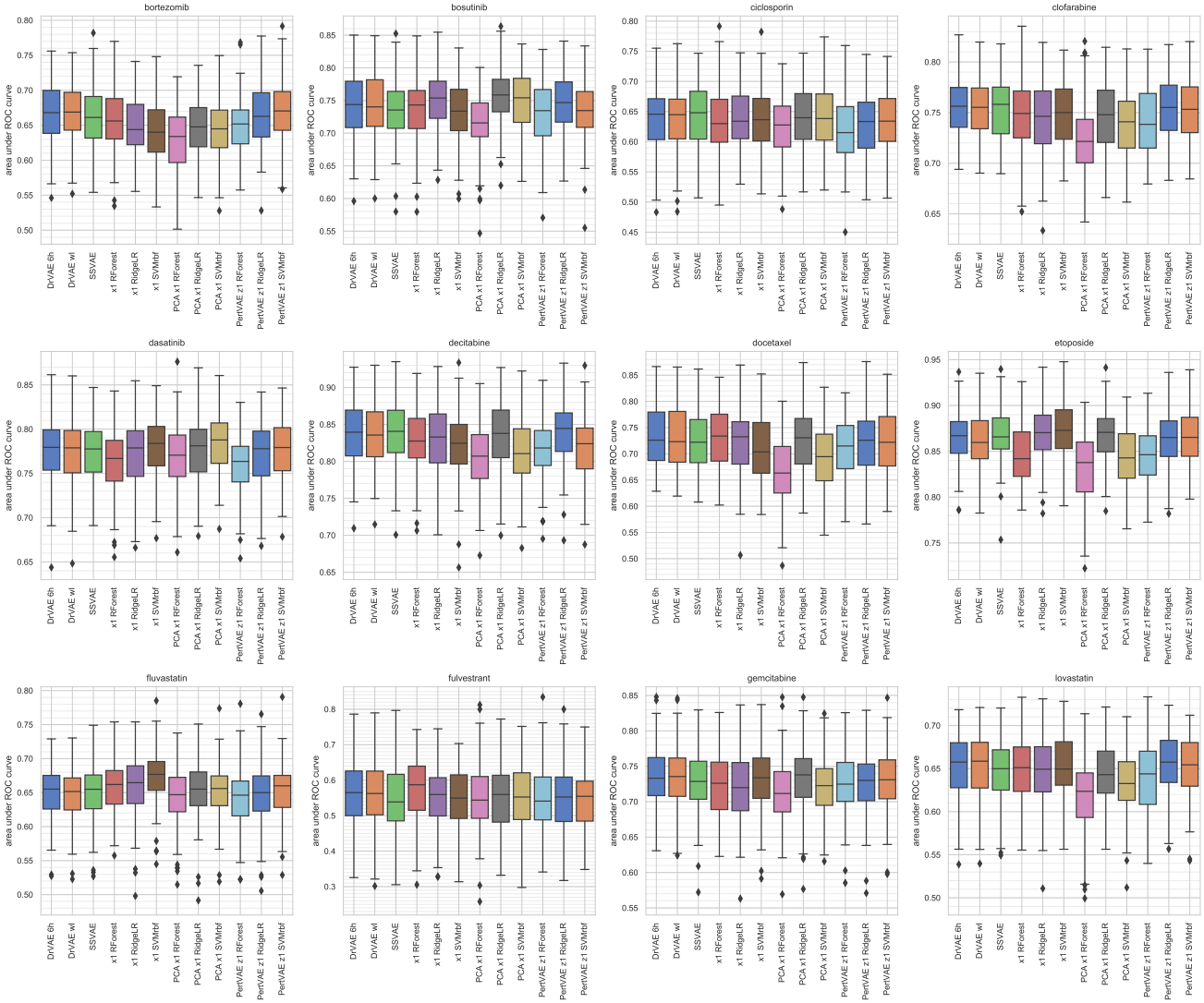


Figure A.3: **Boxplots of test AUROC in 100 data splits.** For each of 26 tested drugs we show the distribution of test set area under ROC curve of 12 evaluated methods in 20 times repeated 5-fold CV. (continues on the next page)

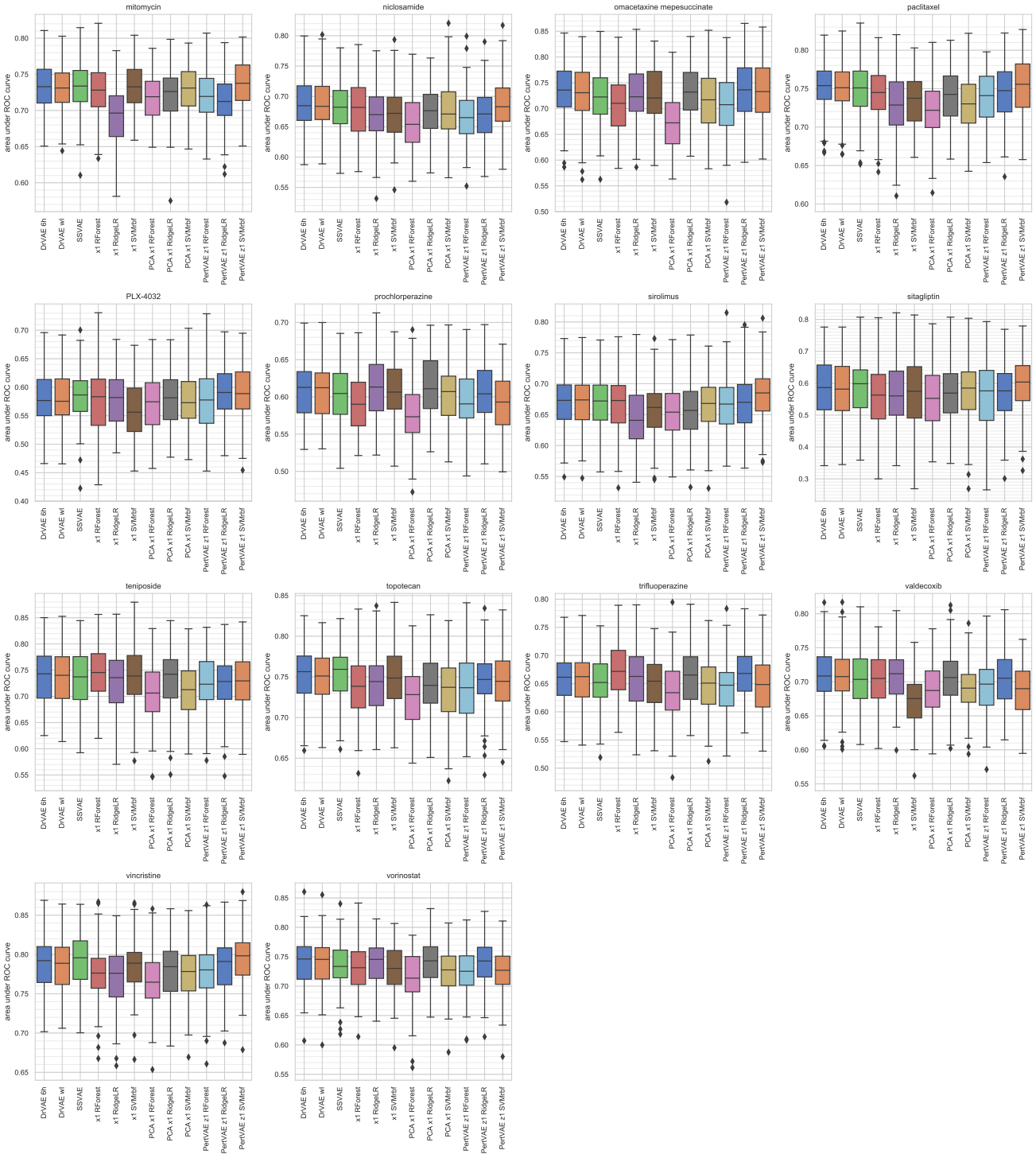


Figure A.3: **Boxplots of test AUROC in 100 data splits.** (continued)

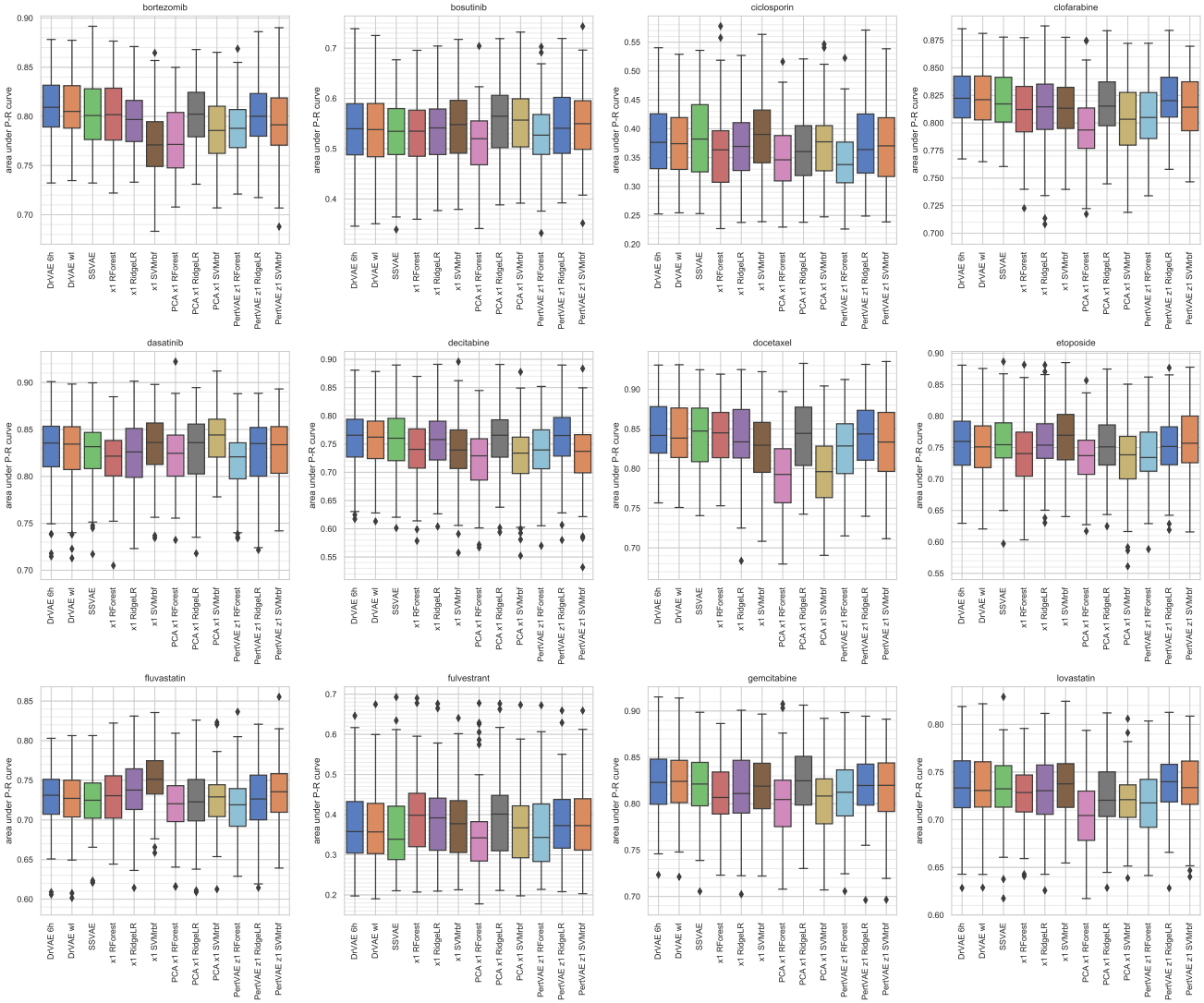


Figure A.4: **Boxplots of test AUPR in 100 data splits.** For each of 26 tested drugs we show the distribution of test set area under Precision-Recall curve of 12 evaluated methods in 20 times repeated 5-fold CV. (continues on the next page)

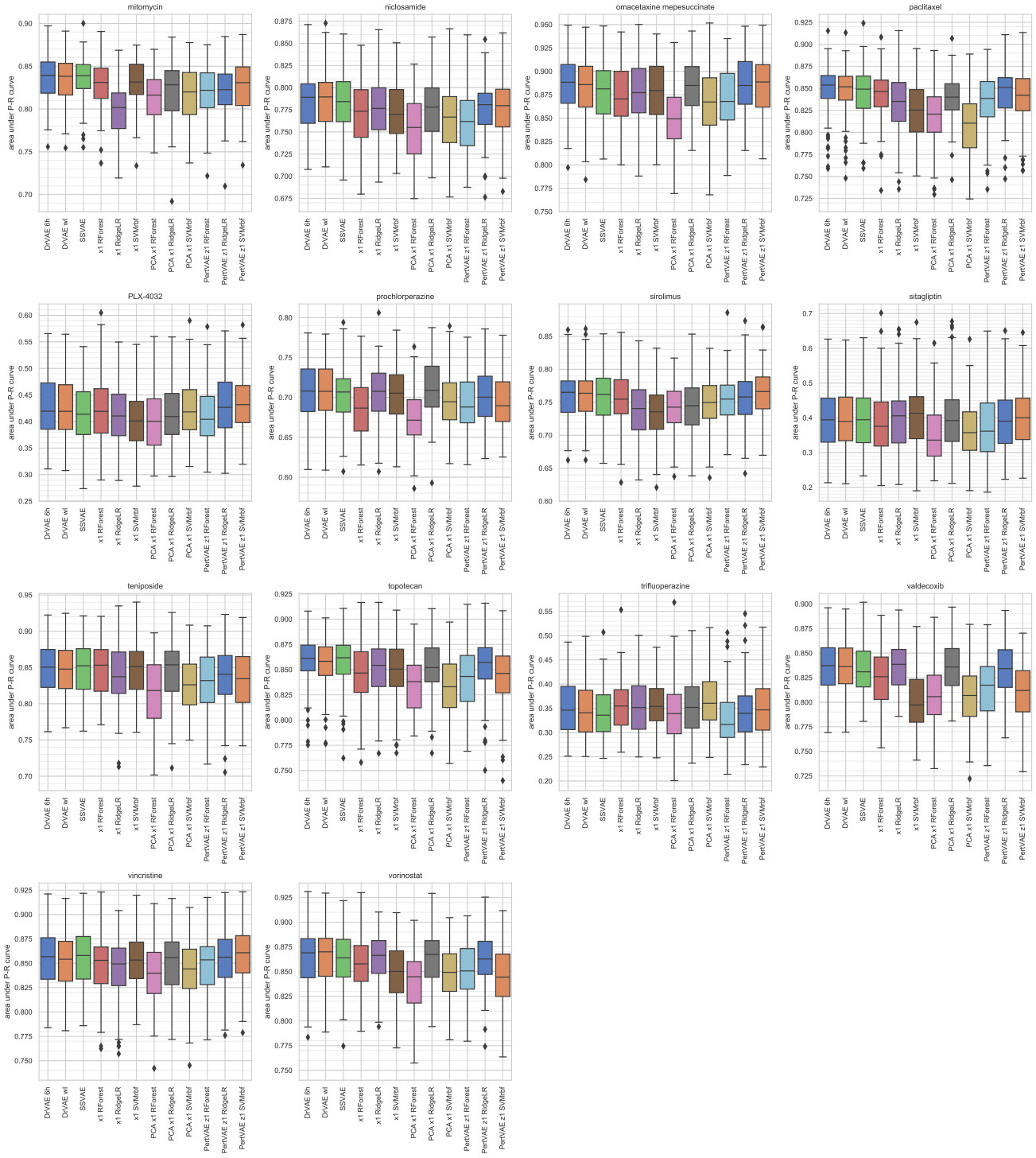
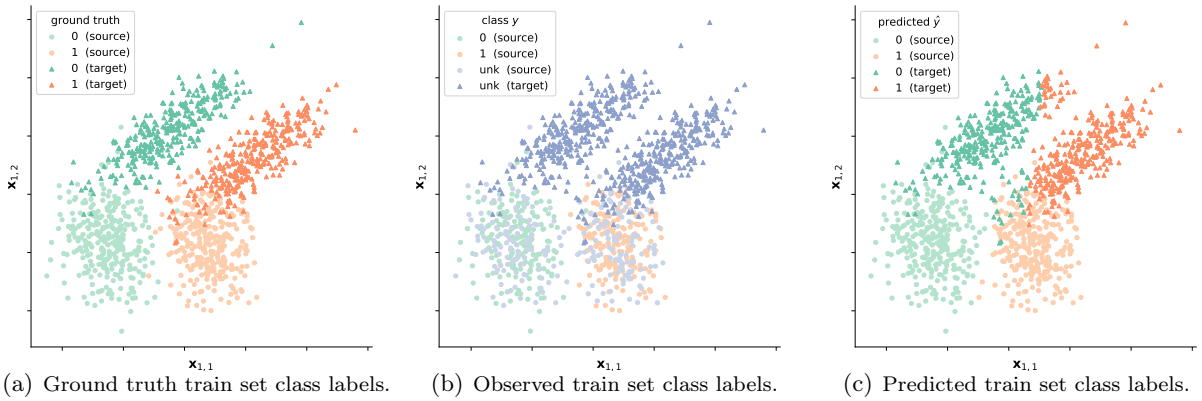


Figure A.4: **Boxplots of test AUPR in 100 data splits.** (continued)

Appendix B

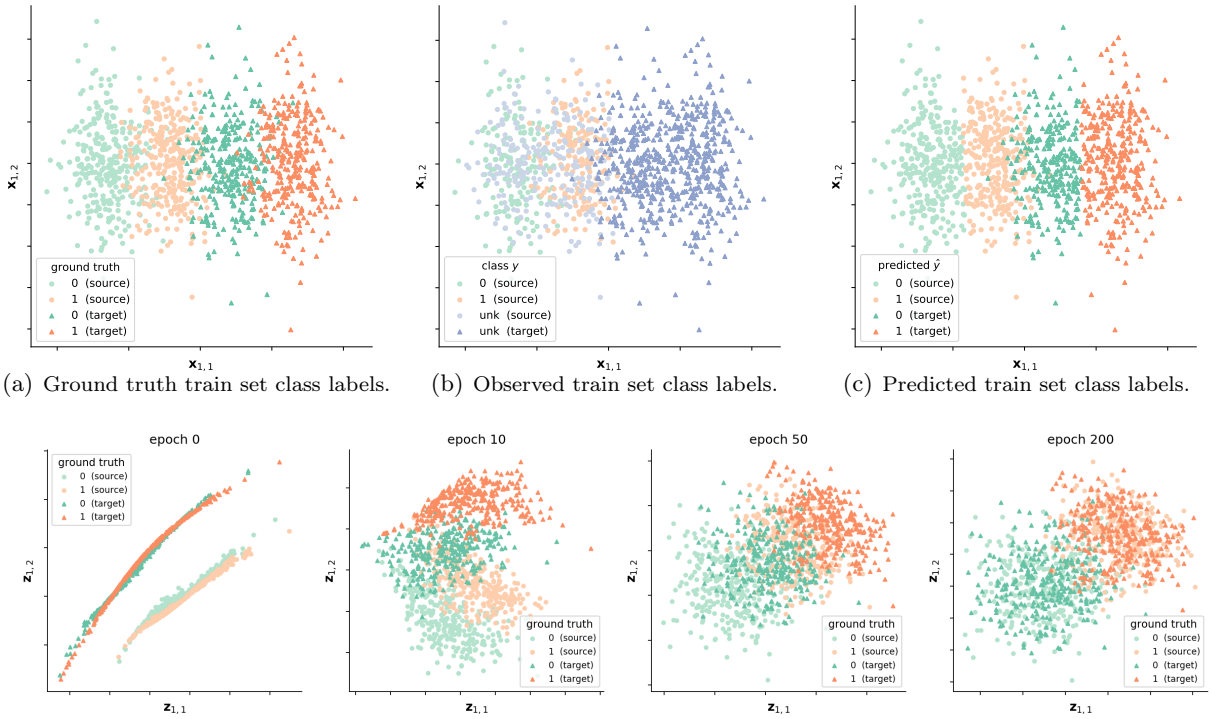
Supplementary figures for Chapter 4 (domain adaptation)



(a) Ground truth train set class labels. (b) Observed train set class labels. (c) Predicted train set class labels.

(d) VFAE latent embedding $\mathbf{z}_1$ during training progression; annotated by ground truth class labels.

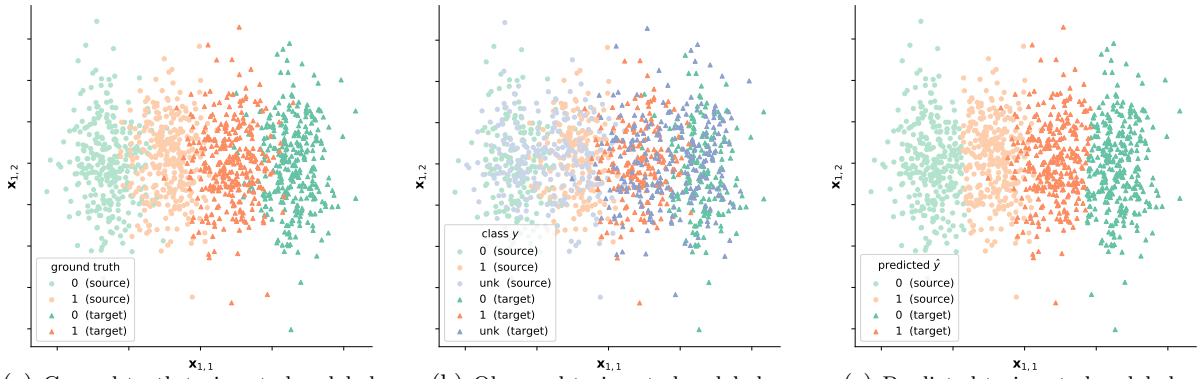
Figure B.1: **SD:overlap S2T.** A successful unsupervised domain adaptation case when necessary conditions are met.



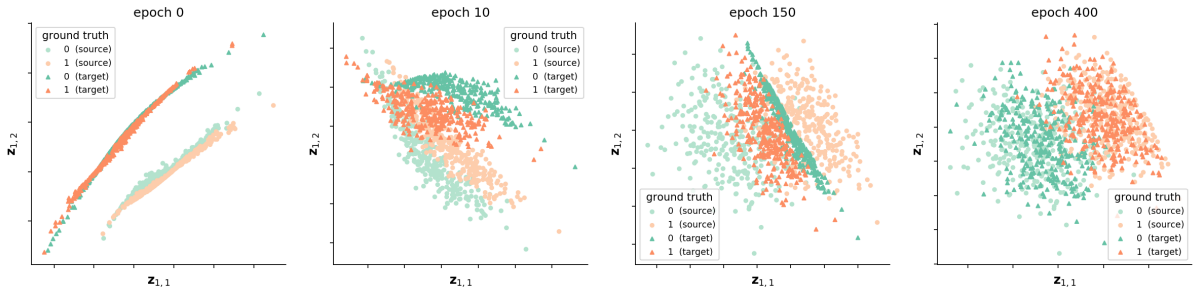
(a) Ground truth train set class labels. (b) Observed train set class labels. (c) Predicted train set class labels.

(d) VFAE latent embedding $\mathbf{z}_1$ during training progression; annotated by ground truth class labels.

Figure B.2: **SD:inline-shift S2T.** In this case of a simple mean shift, unsupervised DA is successful despite lack of class-preserving overlap between the domains. This is thanks to the successful domain matching coinciding with the matching imposed by the steepest descend on the domain-matching objective.

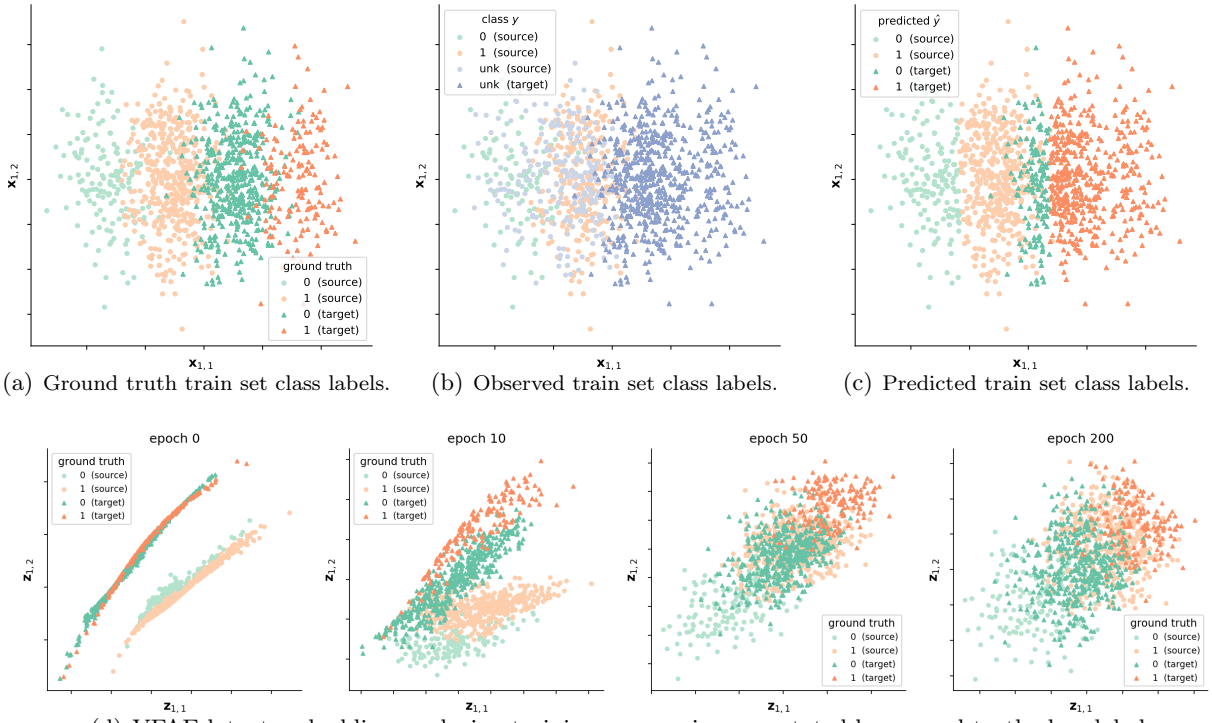


(a) Ground truth train set class labels. (b) Observed train set class labels. (c) Predicted train set class labels.



(d) VFAE latent embedding $\mathbf{z}_1$ during training progression; annotated by ground truth class labels.

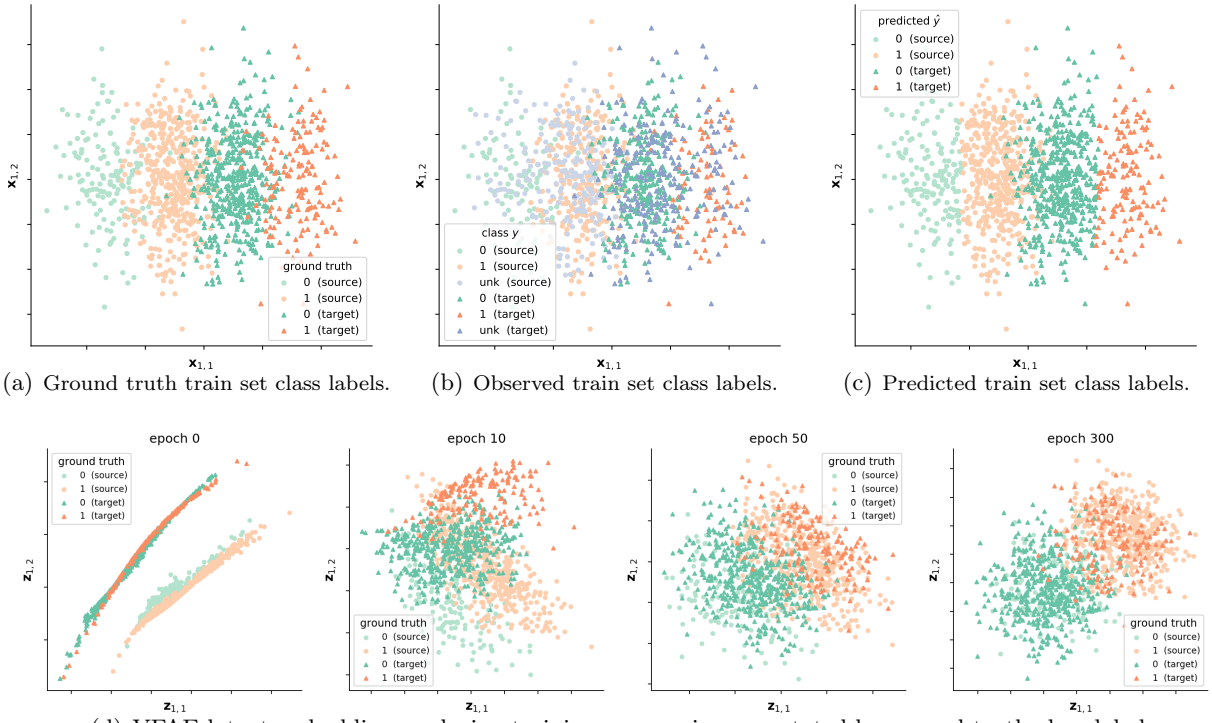
Figure B.3: **SD:inline-mirror ST2T.** Unsupervised domain matching (S2T) fails in this case, as it is essentially an adversarial modification of SD:inline-shift case: nothing else changed but the class association in the target domain is flipped, which has no impact on the VFAE training loss, and as such it is indistinguishable from SD:inline-shift in S2T mode. Labeled target domain samples are needed in the training to learn a successful representation that correctly matches classes between the two domains.



(a) Ground truth train set class labels. (b) Observed train set class labels. (c) Predicted train set class labels.

(d) VFAE latent embedding $\mathbf{z}_1$ during training progression; annotated by ground truth class labels.

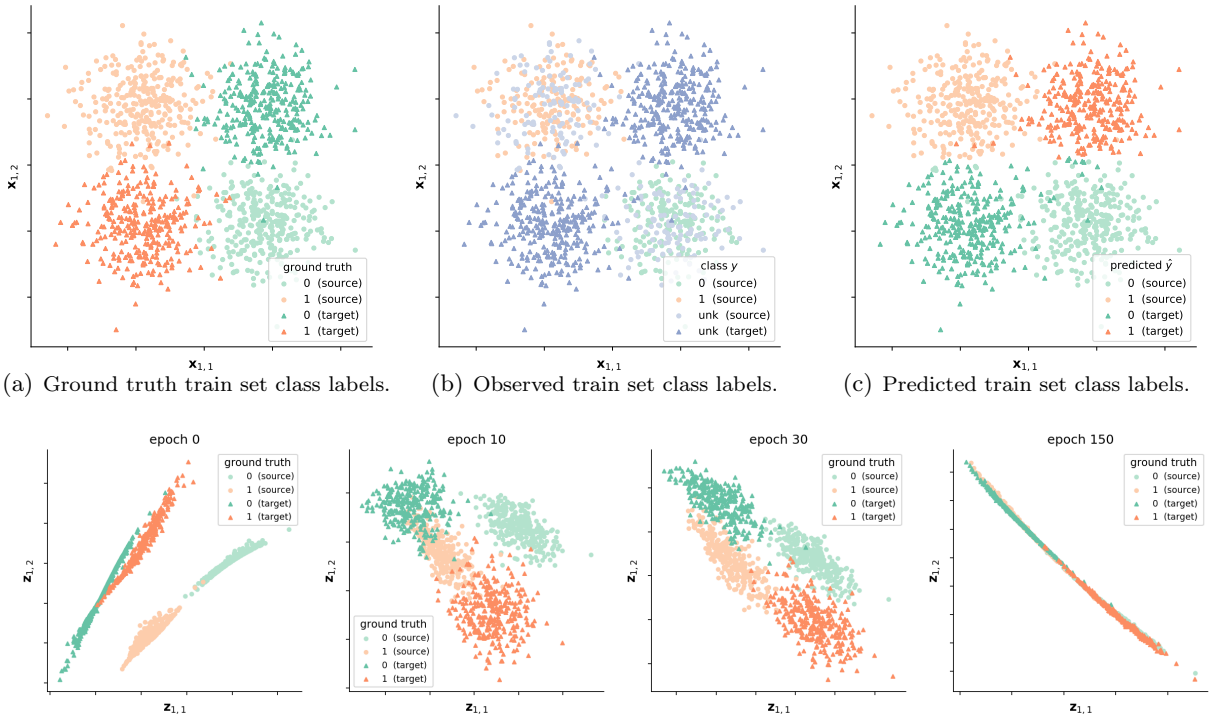
Figure B.4: **SD:inline-uneq-ratio S2T.** Unequal class ratios between the domains causes alignment of some target domain negatives to source domain positives in order to minimize domain discrepancy.



(a) Ground truth train set class labels. (b) Observed train set class labels. (c) Predicted train set class labels.

(d) VFAE latent embedding $\mathbf{z}_1$ during training progression; annotated by ground truth class labels.

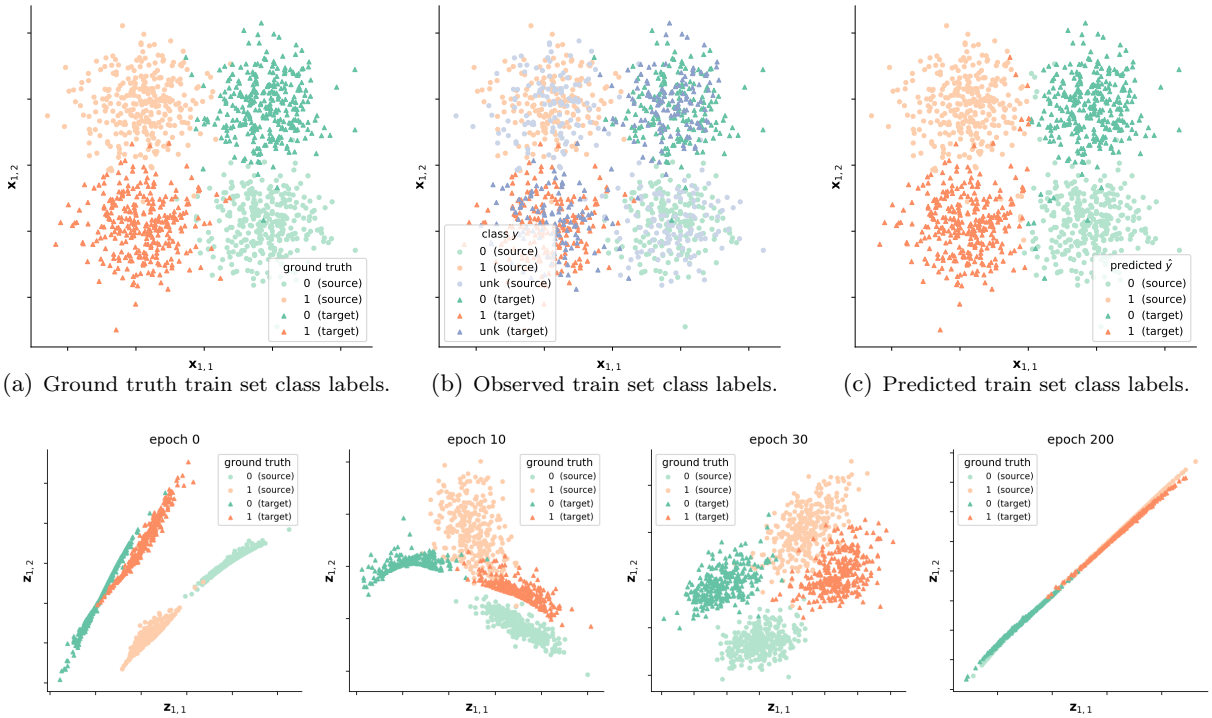
Figure B.5: **SD:inline-uneq-ratio ST2T.** Labeled target domain samples are needed to learn a successful representation that overcomes symmetric domain matching failure in case of unequal class ratio between the domains.



(a) Ground truth train set class labels. (b) Observed train set class labels. (c) Predicted train set class labels.

(d) VFAE latent embedding $\mathbf{z}_1$ during training progression; annotated by ground truth class labels.

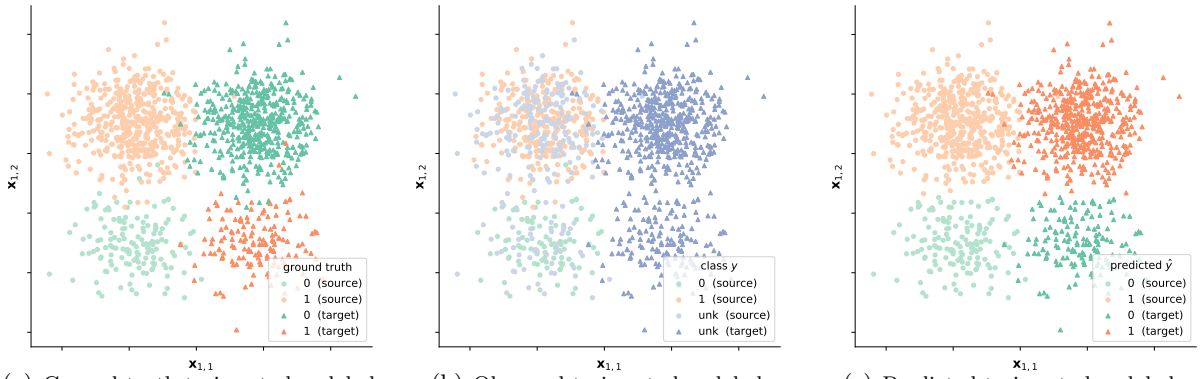
Figure B.6: **SD:diag-classes S2T.** Without further assumptions this data case is in general impossible for unsupervised domain adaptation. Whether VFAE manages to correctly align domains depends on random initialization. In this run, VFAE mismatched the classes in the two domains.



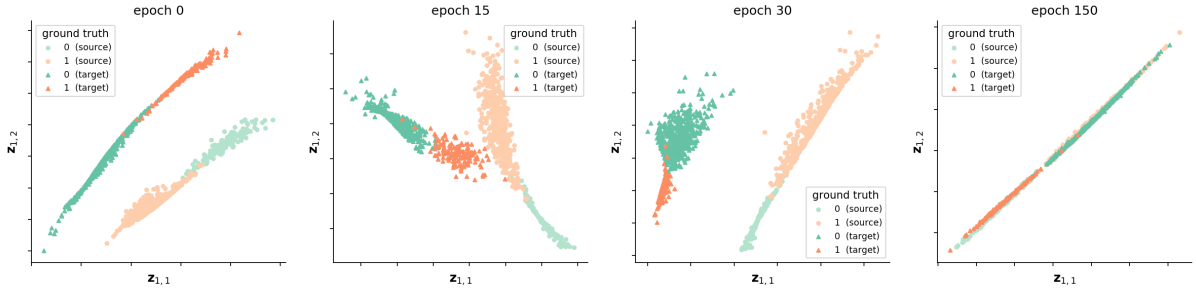
(a) Ground truth train set class labels. (b) Observed train set class labels. (c) Predicted train set class labels.

(d) VFAE latent embedding $\mathbf{z}_1$ during training progression; annotated by ground truth class labels.

Figure B.7: **SD:diag-classes ST2T.** Do reliably match the domains successfully, i.e. positives to positives and negatives to negatives, use of labeled target domain samples is required.

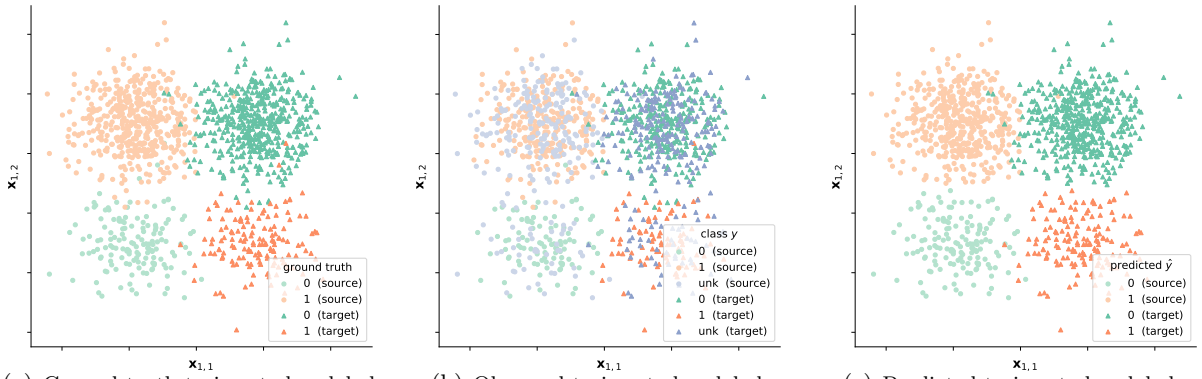


(a) Ground truth train set class labels. (b) Observed train set class labels. (c) Predicted train set class labels.



(d) VFAE latent embedding $\mathbf{z}_1$ during training progression; annotated by ground truth class labels.

Figure B.8: **SD:combination S2T.** This experiment is a combination of SD:diag-classes and SD:inline-uneq-ratio experiments. As such it is in general impossible for an unsupervised domain adaptation method to correctly match the two domains. As expected, VFAE fails in S2T learning mode.



(a) Ground truth train set class labels. (b) Observed train set class labels. (c) Predicted train set class labels.

(d) VFAE latent embedding $\mathbf{z}_1$ during training progression; annotated by ground truth class labels.

Figure B.9: **SD:combination ST2T.** Labeled target domain samples are needed to learn a representation in which samples from one class are projected close to each other, effectively removing the domain shift, even despite unequal class ratio between the domains. This shows how powerful a VFAE can be for semi-supervised domain adaptation.

Bibliography

[1] Alexej Abyzov, Alexander E Urban, Michael Snyder, and Mark Gerstein. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, 21(6):974–984, 2011. 104

[2] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014. 28

[3] Alexander Aliper, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, and Alex Zhavoronkov. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.*, 13(7):2524–2530, July 2016. 17

[4] Muhammad Ammad-Ud-Din, Suleiman A Khan, Krister Wennerberg, and Tero Aittokallio. Systematic identification of feature combinations for predicting drug response with bayesian multi-view multi-task linear regression. *Bioinformatics*, 33(14):i359–i368, July 2017. 16, 107

[5] Samuel Aparicio, Manuel Hidalgo, and Andrew L Kung. Examining the utility of patient-derived xenograft mouse models. *Nature Reviews Cancer*, 15(5):311–316, May 2015. 13

[6] Aryan Arbabi, Ladislav Rampášek, and Michael Brudno. Cell-free DNA fragment-size distribution analysis for non-invasive prenatal CNV prediction. *Bioinformatics*, 32(11):1662–1669, 2016. 107

[7] Francisco Azuaje. Computational models for predicting drug responses in cancer research. *Briefings in bioinformatics*, 18(5):820–829, September 2017. 16, 37

[8] Francisco Azuaje, Tony Kaoma, Céline Jeanty, Petr V Nazarov, Arnaud Muller, Sang-Yoon Kim, Anna Golebiewska, Gunnar Dittmar, and Simone P Niclou. Dr. Paso: Drug response prediction and analysis system for oncology research. *bioRxiv*, page 237727, 2017. 37

[9] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012. 14, 37, 40, 84

[10] José Baselga, Ian Bradbury, Holger Eidtmann, Serena Di Cosimo, Evandro De Azambuja, Claudia Aura, Henry Gómez, Phuong Dinh, Karine Fauria, Veerle Van Dooren, et al. Lapatinib with trastuzumab for her2-positive early breast cancer (neoaltto): a randomised, open-label, multicentre, phase 3 trial. *The Lancet*, 379(9816):633–640, 2012. 61

[11] Leonard Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3:1–8, 1972. 21

[12] Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pages 139–153. Springer, 2012. 35

[13] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007. 29, 35

[14] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010. 29, 30, 32

[15] Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010. 35

[16] S H Berger, C H Jenh, L F Johnson, and F G Berger. Thymidylate synthase overproduction and gene amplification in fluorodeoxyuridine-resistant human cells. *Mol. Pharmacol.*, 28(5):461–467, November 1985. 19

[17] Diana W Bianchi and Rossa WK Chiu. Sequencing of circulating cell-free dna during pregnancy. *New England Journal of Medicine*, 379(5):464–473, 2018. 1, 108

[18] Philip Botros and Jakub M Tomczak. Hierarchical VampPrior variational fair auto-encoder. *arXiv preprint arXiv:1806.09918*, 2018. 23, 28, 32, 34, 62

[19] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, pages 343–351, 2016. 28, 34, 35

[20] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015. 23

[21] Vladimír Boža, Broňa Brejová, and Tomáš Vinař. Deepnano: deep recurrent neural networks for base calling in minion nanopore reads. *PloS one*, 12(6):e0178751, 2017. 21

[22] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*, September 2018. 11

[23] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525, 2016. 71

[24] Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013. 101, 102

[25] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015. 23

[26] Yoosup Chang, Hyejin Park, Hyun-Jin Yang, Seungju Lee, Kwee-Yum Lee, Tae Soon Kim, Jongsun Jung, and Jae-Min Shin. Cancer drug response profile scan (CDRscan): A deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci. Rep.*, 8(1):8857, June 2018. 17

[27] Shengpei Chen, Tze Kin Lau, Chunlei Zhang, Chenming Xu, Zhengfeng Xu, Ping Hu, Jian Xu, Hefeng Huang, Ling Pan, Fuman Jiang, et al. A method for noninvasive detection of fetal large deletions/duplications by low coverage massively parallel sequencing. *Prenatal diagnosis*, 33(6): 584–590, 2013. 93, 96, 99, 100, 103

[28] Yu-Chiao Chiu, Hung-I Harry Chen, Tinghe Zhang, Songyao Zhang, Aparna Gorthi, Li-Ju Wang, Yufei Huang, and Yidong Chen. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *arXiv preprint*, May 2018. 18

[29] T. Chu, K. Bunce, W.A. Hogge, and D.G. Peters. Statistical model for whole genome sequencing and its application to minimally invasive diagnosis of fetal genetic disease. *Bioinformatics*, 25(10): 1244–1250, 2009. 93

[30] James W Clendening, Aleksandra Pandyra, Zhihua Li, Paul C Boutros, Anna Martirosyan, Richard Lehner, Igor Jurisica, Suzanne Trudel, and Linda Z Penn. Exploiting the mevalonate pathway to distinguish statin-sensitive multiple myeloma. *Blood*, 115(23):4787–4797, 2010. 47

[31] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *International Conference on Learning Representations, arXiv:1511.07289*, 2016. 43

[32] Thomas Cokelaer, Elisabeth Chen, Francesco Iorio, Michael P Menden, Howard Lightfoot, Julio Saez-Rodriguez, and Mathew J Garnett. GDSCTools for mining pharmacogenomic interactions in cancer. *Bioinformatics*, November 2017. 14

[33] Patricia Cortazar and Charles E Geyer. Pathological complete response in neoadjuvant treatment of breast cancer. *Annals of surgical oncology*, 22(5):1441–1446, 2015. 62

[34] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Muhammad Ammad-ud din, Petteri Hintsanen, Suleiman A Khan, John-Patrick Mpindi, Olli Kallioniemi, Antti Honkela, Tero Aittokallio, Krister Wennerberg, NCI DREAM Community, James J Collins, Dan Gallahan, Dinah Singer, Julio Saez-Rodriguez, Samuel Kaski, Joe W Gray, and Gustavo Stolovitzky. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202–1212, December 2014. 16, 17, 37, 38, 50, 51, 107

[35] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561, 1970. 7

[36] Kathleen A Cronin, Andrew J Lake, Susan Scott, Recinda L Sherman, Anne-Michelle Noone, Nadia Howlader, S Jane Henley, Robert N Anderson, Albert U Firth, Jiemin Ma, Betsy A Kohler, and Ahmedin Jemal. Annual report to the nation on the status of cancer, part i: National cancer statistics. *Cancer*, 124(13):2785–2800, July 2018. 11

[37] George E Dahl, Navdeep Jaitly, and Ruslan Salakhutdinov. Multi-task neural networks for QSAR predictions. *arXiv preprint arXiv:1406.1231*, June 2014. 17

[38] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*, 2017. 35

[39] Matei David, Lewis Jonathan Dursi, Delia Yao, Paul C Boutros, and Jared T Simpson. Nanocall: an open source basecaller for oxford nanopore sequencing data. *Bioinformatics*, 33(1):49–55, 2016. 21

[40] Nicolas De Jay, Simon Papillon-Cavanagh, Catharina Olsen, Nehme El-Hachem, Gianluca Bontempi, and Benjamin Haibe-Kains. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics*, 29(18):2365–2368, September 2013. 16

[41] Carlos De Niz, Raziur Rahman, Xiangyuan Zhao, and Ranadip Pal. Algorithms for drug sensitivity prediction. *Algorithms*, 9(4):77, November 2016. 16

[42] Wendy De Roock, Veerle De Vriendt, Nicola Normanno, Fortunato Ciardiello, and Sabine Tejpar. KRAS, BRAF, PIK3CA, and PTEN mutations: implications for targeted therapies in metastatic colorectal cancer. *Lancet Oncol.*, 12(6):594–603, June 2011. 1, 11

[43] Muthu Dhandapani and Aaron Goldman. Preclinical cancer models and biomarkers for drug development: New technologies and emerging tools. *J. Mol. Biomark. Diagn.*, 8(5), September 2017. 13

[44] Russell J Diefenbach, Jenny H Lee, Richard F Kefford, and Helen Rizos. Evaluation of commercial kits for purification of circulating free dna. *Cancer genetics*, 228:21–27, 2018. 108

[45] Ayse Berceste Dincer, Safiye Celik, Naozumi Hiranuma, and Su-In Lee. DeepProfile: Deep learning of cancer molecular profiles for precision medicine. *bioRxiv*, page 278739, May 2018. 18, 23, 38

[46] Michael Q Ding, Lujia Chen, Gregory F Cooper, Jonathan D Young, and Xinghua Lu. Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Mol. Cancer Res.*, 16(2):269–278, February 2018. 1, 11

[47] Zijian Ding, Songpeng Zu, and Jin Gu. Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics*, 32(19):2891–2895, 2016. 83

[48] M Doherty, T Metcalfe, E Guardino, E Peters, and L Ramage. Precision medicine and oncology: an overview of the opportunities presented by next-generation sequencing and big data and the challenges posed to conventional drug development and regulatory approval pathways. *Ann. Oncol.*, 27(8):1644–1646, August 2016. 1, 11

[49] Wilson R Douglas, Sylvie Langlois, and Jo-Ann Johnson. Mid-trimester amniocentesis fetal loss rate (committee opinion). *Journal of Obstetrics and Gynaecology Canada*, 29(7):586–590, 2007. 1, 92

[50] Alexander Drilon, Theodore W Laetsch, Shivaani Kummar, Steven G DuBois, Ulrik N Lassen, George D Demetri, Michael Nathenson, Robert C Doebele, Anna F Farago, Alberto S Pappo, Brian Turpin, Afshin Dowlati, Marcia S Brose, Leo Mascarenhas, Noah Federman, Jordan Berlin, Wafik S El-Deiry, Christina Baik, John Deeken, Valentina Boni, Ramamoorthy Nagasubramanian,

Matthew Taylor, Erin R Rudzinski, Funda Meric-Bernstam, Davendra P S Sohal, Patrick C Ma, Luis E Raez, Jaclyn F Hechtman, Ryma Benayed, Marc Ladanyi, Brian B Tuch, Kevin Ebata, Scott Cruickshank, Nora C Ku, Michael C Cox, Douglas S Hawkins, David S Hong, and David M Hyman. Efficacy of larotrectinib in TRK Fusion–Positive cancers in adults and children. *N. Engl. J. Med.*, 378(8):731–739, February 2018. 11

[51] Lixin Duan, Ivor W Tsang, and Dong Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):465–479, 2012. 35

[52] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998. 20

[53] Misko Dzamba, Arun K Ramani, Pawel Buczkowicz, Yue Jiang, Man Yu, Cynthia Hawkins, and Michael Brudno. Identification of complex genomic rearrangements in cancers using cougar. *Genome research*, 27(1):107–117, 2017. 8

[54] Sean R. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998. 20

[55] Harrison Edwards and Amos Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016. 23

[56] Evan E Eichler. Genetic variation, comparative genomics, and the diagnosis of disease. *New England Journal of Medicine*, 381(1):64–74, 2019. 8

[57] W El-Deredy, S M Ashmore, N M Branston, J L Darling, S R Williams, and D G Thomas. Pretreatment prediction of the chemotherapeutic response of human glioma cell cultures using nuclear magnetic resonance spectroscopy and artificial neural networks. *Cancer Res.*, 57(19): 4196–4199, October 1997. 17

[58] Jason Ernst and Manolis Kellis. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215, 2012. 21

[59] Mohammad Fallahi-Sichani, Saman Honarnejad, Laura M Heiser, Joe W Gray, and Peter K Sorger. Metrics other than potency reveal systematic variation in responses to cancer drugs. *Nature chemical biology*, 9(11):708–714, 2013. 15

[60] H Christina Fan, Wei Gu, Jianbin Wang, Yair J Blumenfeld, Yasser Y El-Sayed, and Stephen R Quake. Non-invasive prenatal measurement of the fetal genome. *Nature*, 487(7407):320–324, 2012. 93

[61] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013. 28, 35

[62] Julie Foucquier and Mickael Guedj. Analysis of drug combinations: current methodological landscape. *Pharmacology research & perspectives*, 3(3):e00149, 2015. 6

[63] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 2018. 71

[64] Debora Fumagalli, David Venet, Michail Ignatiadis, Hatem A Azim, Marion Maetens, Françoise Rothé, Roberto Salgado, Ian Bradbury, Lajos Pusztai, Nadia Harbeck, et al. Rna sequencing to predict response to neoadjuvant anti-her2 therapy: a secondary analysis of the neoaltto randomized clinical trial. *JAMA oncology*, 3(2):227–234, 2017. 61

[65] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014. 28

[66] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. viii, 28, 30, 31, 32, 35, 62

[67] Hui Gao, Joshua M Korn, Stéphane Ferretti, John E Monahan, Youzhen Wang, Mallika Singh, Chao Zhang, Christian Schnell, Guizhi Yang, Yun Zhang, O Alejandro Balbin, Stéphanie Barbe, Hongbo Cai, Fergal Casey, Susmita Chatterjee, Derek Y Chiang, Shannon Chuai, Shawn M Cogan, Scott D Collins, Ernesta Dammassa, Nicolas Ebel, Millicent Embry, John Green, Audrey Kauffmann, Colleen Kowal, Rebecca J Leary, Joseph Lehar, Ying Liang, Alice Loo, Edward Lorenzana, E Robert McDonald, 3rd, Margaret E McLaughlin, Jason Merkin, Ronald Meyer, Tara L Naylor, Montesa Patawaran, Anupama Reddy, Claudia Röelli, David A Ruddy, Fernando Salangsang, Francesca Santacroce, Angad P Singh, Yan Tang, Walter Tinetto, Sonja Tobler, Roberto Velazquez, Kavitha Venkatesan, Fabian Von Arx, Hui Qin Wang, Zongyao Wang, Marion Wiesmann, Daniel Wyss, Fiona Xu, Hans Bitter, Peter Atadja, Emma Lees, Francesco Hofmann, En Li, Nicholas Keen, Robert Cozens, Michael Rugaard Jensen, Nancy K Pryer, Juliet A Williams, and William R Sellers. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature medicine*, 21(11):1318–1325, November 2015. 13, 16

[68] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, Qingsong Liu, Francesco Iorio, Didier Surdez, Li Chen, Randy J Milano, Graham R Bignell, Ah T Tam, Helen Davies, Jesse A Stevenson, Syd Barthorpe, Stephen R Lutz, Fiona Kogera, Karl Lawrence, Anne McLaren-Douglas, Xeni Mitropoulos, Tatiana Mironenko, Helen Thi, Laura Richardson, Wenjun Zhou, Frances Jewitt, Tinghu Zhang, Patrick O'Brien, Jessica L Boisvert, Stacey Price, Wooyoung Hur, Wanjuan Yang, Xianming Deng, Adam Butler, Hwan Geun Choi, Jae Won Chang, Jose Baselga, Ivan Stamenkovic, Jeffrey A Engelman, Sreenath V Sharma, Olivier Delattre, Julio Saez-Rodriguez, Nathanael S Gray, Jeffrey Settleman, P Andrew Futreal, Daniel A Haber, Michael R Stratton, Sridhar Ramaswamy, Ultan McDermott, and Cyril H Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, March 2012. viii, 14, 15, 37

[69] Levi A Garraway, Jaap Verweij, and Karla V Ballman. Precision oncology: an overview. *J. Clin. Oncol.*, 31(15):1803–1805, May 2013. 1, 11

[70] Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017. 34

[71] Paul Geeleher, Nancy J Cox, and R Stephanie Huang. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome biology*, 15 (3):R47, 2014. 18, 60, 62

[72] Paul Geeleher, Zhenyu Zhang, Fan Wang, Robert F Gruener, Aritro Nath, Gladys Morrison, Steven Bhutra, Robert L Grossman, and R Stephanie Huang. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome research*, 27(10):1743–1751, 2017. 18, 60, 61

[73] Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. *Proceedings of the annual meeting of the cognitive science society*, 36, 2014. 25

[74] Mahmoud Ghandi, Franklin W Huang, Judit Jané-Valbuena, Gregory V Kryukov, Christopher C Lo, E Robert McDonald, Jordi Barretina, Ellen T Gelfand, Craig M Bielski, Haoxin Li, et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503, 2019. 14

[75] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. *arXiv preprint*, April 2017. 17

[76] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011. 35

[77] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a Data-Driven continuous representation of molecules. *ACS Cent Sci*, 4(2):268–276, February 2018. 17

[78] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018. 23

[79] Mehmet Gönen and Adam A Margolin. Drug susceptibility prediction against a panel of drugs using kernelized bayesian multitask learning. *Bioinformatics*, 30(17):i556–i563, September 2014. 16

[80] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2012. 35

[81] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 28

[82] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, pages 999–1006. IEEE, 2011. 35

[83] Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. *arXiv preprint arXiv:1711.00123*, 2017. 27

[84] Joel Greshock, Kurtis E Bachman, Yan Y Degenhardt, Junping Jing, Yuan H Wen, Stephen Eastman, Elizabeth McNeil, Christopher Moy, Ronald Wegrzyn, Kurt Auger, et al. Molecular target class is predictive of in vitro response profile. *Cancer research*, 70(9):3677–3686, 2010. 14

[85] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007. 33

[86] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012. 33, 34

[87] Aditya Grover, Christopher Chute, Rui Shu, Zhangjie Cao, and Stefano Ermon. Alignflow: Cycle consistent learning from multiple domains via normalizing flows. *arXiv preprint arXiv:1905.12892*, 2019. 28

[88] Sudheer Gupta, Kumardeep Chaudhary, Rahul Kumar, Ankur Gautam, Jagpreet Singh Nanda, Sandeep Kumar Dhanda, Samir Kumar Brahmachari, and Gajendra P S Raghava. Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine. *Sci. Rep.*, 6:23857, March 2016. 14, 17

[89] Benjamin Haibe-Kains, Nehme El-Hachem, Nicolai Juul Birkbak, Andrew C Jin, Andrew H Beck, Hugo JWL Aerts, and John Quackenbush. Inconsistency in large pharmacogenomic studies. *Nature*, 504(7480):389–393, 2013. 15, 40, 84

[90] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000. 9

[91] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5): 646–674, March 2011. viii, 9, 10

[92] Peter M Haverty, Eva Lin, Jenille Tan, Yihong Yu, Billy Lam, Steve Lianoglou, Richard M Neve, Scott Martin, Jeff Settleman, Robert L Yauch, et al. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature*, 533(7603):333, 2016. 14, 37

[93] Hua-Jun He, Erica V Stein, Yves Konigshofer, Thomas Forbes, Farol L Tomson, Russell Garlick, Emiko Yamada, Tony Godfrey, Toshiya Abe, Koji Tamura, et al. Multilaboratory assessment of a new reference material for quality assurance of cell-free tumor dna measurements. *The Journal of Molecular Diagnostics*, 2019. 108

[94] Xiao He, Lukas Folkman, and Karsten Borgwardt. Kernelized rank learning for personalized drug recommendation. *Bioinformatics*, 34(16):2808–2816, August 2018. 17

[95] J Javier Hernandez, Michael Pryszlak, Lindsay Smith, Connor Yanchus, Naheed Kurji, Vijay M Shahani, and Steven V Molinski. Giving drugs a second chance: overcoming regulatory and financial hurdles in repurposing approved drugs as cancer therapeutics. *Frontiers in oncology*, 7:273, 2017. 6

[96] John Heymach, Lada Krilov, Anthony Alberg, Nancy Baxter, Susan Marina Chang, Ryan Corcoran, William Dale, Angela DeMichele, Catherine S Magid Diefenbach, Robert Dreicer, Andrew S Epstein, Maura L Gillison, David L Graham, Joshua Jones, Andrew H Ko, Ana Maria Lopez, Robert G Maki, Carlos Rodriguez-Galindo, Richard L Schilsky, Mario Sznol, Shannon Neville Westin, and Harold Burstein. Clinical cancer advances 2018: Annual report on progress against cancer from the american society of clinical oncology. *J. Clin. Oncol.*, 36(10):1020–1044, April 2018. 1, 11

[97] Caitriona Holohan, Sandra Van Schaeybroeck, Daniel B Longley, and Patrick G Johnston. Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer*, 13(10):714–726, October 2013. 19

[98] Haodi Hou, Jing Huo, and Yang Gao. Cross-domain adversarial auto-encoder. *arXiv preprint arXiv:1804.06078*, 2018. 34

[99] Genevieve Housman, Shannon Byler, Sarah Heerboth, Karolina Lapinska, Mckenna Longacre, Nicole Snyder, and Sibaji Sarkar. Drug resistance in cancer: an overview. *Cancers*, 6(3):1769–1792, September 2014. 19

[100] Quanyin Hu, Wujin Sun, Chao Wang, and Zhen Gu. Recent advances of cocktail chemotherapy by combination drug delivery systems. *Advanced drug delivery reviews*, 98:19–34, 2016. 107

[101] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoder. *Workshop at International Conference on Learning Representations*, 2019. 28, 91

[102] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K Egan, Qingsong Liu, Tatiana Mironenko, Xeni Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S Gray, Daniel A Haber, Michael R Stratton, Cyril H Benes, Lodewyk F A Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J Garnett. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, July 2016. 14, 16

[103] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 27

[104] In Sock Jang, Elias Chaibub Neto, Justin Guinney, Stephen H Friend, and Adam A Margolin. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. In *Pac. Symp. Biocomput*, volume 19, pages 63–74. World Scientific, 2014. 15, 16, 51

[105] C H Jenh, P K Geyer, F Baskin, and L F Johnson. Thymidylate synthase gene amplification in fluorodeoxyuridine-resistant mouse cell lines. *Mol. Pharmacol.*, 28(1):80–85, July 1985. 19

[106] Fredrik Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536, 2019. 29, 35, 62, 65, 69

[107] Amber Johnson, Jia Zeng, Ann M Bailey, Vijaykumar Holla, Beate Litzenburger, Humberto Lara-Guerra, Gordon B Mills, John Mendelsohn, Kenna R Shaw, and Funda Meric-Bernstam. The right drugs at the right time for the right patient: the MD anderson precision oncology decision support platform. *Drug Discovery Today*, 20(12):1433–1438, December 2015. 11

[108] Julie A Johnson and Larisa H Cavallari. Warfarin pharmacogenetics. *Trends in cardiovascular medicine*, 25(1):33–41, 2015. 5

[109] Matthew Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016. 23, 41

[110] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007. 18, 61

[111] Dan Jurafsky and James H Martin. Speech and language processing. vol. 3, 2014. 20

[112] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, page 531210, 2019. 7

[113] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 180–191. VLDB Endowment, 2004. 29

[114] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations, arXiv:1412.6980*, 2015. 43, 55

[115] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *International Conference on Learning Representations, arXiv:1312.6114*, December 2014. x, xi, 2, 20, 23, 26, 38, 41, 42, 43, 55

[116] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019. 20, 24, 25

[117] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014. xi, 20, 23, 27, 42, 43, 44, 47, 55, 56, 58

[118] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in Neural Information Processing Systems 29*, pages 4743–4751, 2016. 23, 27, 51, 58

[119] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016. 23

[120] J.O. Kitzman, M.W. Snyder, M. Ventura, A.P. Lewis, R. Qiu, L.V.E. Simmons, H.S. Gammill, et al. Noninvasive whole-genome sequencing of a human fetus. *Science Translational Medicine*, 4 (137):137ra76–137ra76, 2012. 93, 94, 99, 102

[121] Susumu Kobayashi, Titus J Boggon, Tajhal Dayaram, Pasi A Jänne, Olivier Kocher, Matthew Meyerson, Bruce E Johnson, Michael J Eck, Daniel G Tenen, and Balázs Halmos. EGFR mutation and resistance of Non–Small-Cell lung cancer to gefitinib. *N. Engl. J. Med.*, 352(8):786–792, February 2005. 19

[122] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 24

[123] Wouter M. Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018. 28

[124] Rahul G Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2101–2109. AAAI Press, 2017. 23, 41

[125] Gerhard Kurz, Florian Pfaff, and Uwe D Hanebeck. Kullback-leibler divergence and moment matching for hyperspherical probability distributions. In *2016 19th International Conference on Information Fusion (FUSION)*, pages 2087–2094. IEEE, 2016. 34

[126] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1):D1062–D1067, 2017. 7

[127] Dung T Le, Jennifer N Durham, Kellie N Smith, Hao Wang, Bjarne R Bartlett, Laveet K Aulakh, Steve Lu, Holly Kemberling, Cara Wilt, Brandon S Luber, Fay Wong, Nilofer S Azad, Agnieszka A Rucki, Dan Laheru, Ross Donehower, Atif Zaheer, George A Fisher, Todd S Crocenzi, James J Lee, Tim F Greten, Austin G Duffy, Kristen K Ciombor, Aleksandra D Eyring, Bao H Lam, Andrew Joe, S Peter Kang, Matthias Holdhoff, Ludmila Danilova, Leslie Cope, Christian Meyer, Shibin Zhou, Richard M Goldberg, Deborah K Armstrong, Katherine M Bever, Amanda N Fader, Janis Taube, Franck Housseau, David Spetzler, Nianqing Xiao, Drew M Pardoll, Nickolas Papadopoulos, Kenneth W Kinzler, James R Eshleman, Bert Vogelstein, Robert A Anders, and Luis A Diaz, Jr. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science*, 357(6349): 409–413, July 2017. 11

[128] Jin-Ku Lee, Zhaoqi Liu, Jason K Sa, Sang Shin, Jiguang Wang, Mykola Bordyuh, Hee Jin Cho, Oliver Elliott, Timothy Chu, Seung Won Choi, Daniel I S Rosenbloom, In-Hee Lee, Yong Jae Shin, Hyun Ju Kang, Donggeon Kim, Sun Young Kim, Moon-Hee Sim, Jusun Kim, Taehyang Lee, Yun Jee Seo, Hyemi Shin, Mijeong Lee, Sung Heon Kim, Yong-Jun Kwon, Jeong-Woo Oh, Minsuk Song, Misuk Kim, Doo-Sik Kong, Jung Won Choi, Ho Jun Seol, Jung-Il Lee, Seung Tae Kim, Joon Oh Park, Kyoung-Mee Kim, Sang-Yong Song, Jeong-Won Lee, Hee-Cheol Kim, Jeong Eon Lee, Min Gew Choi, Sung Wook Seo, Young Mog Shim, Jae Ill Zo, Byong Chang Jeong, Yeup Yoon, Gyu Ha Ryu, Nayoung K D Kim, Joon Seol Bae, Woong-Yang Park, Jeongwu Lee, Roel G W Verhaak, Antonio Iavarone, Jeeyun Lee, Raul Rabadan, and Do-Hyun Nam. Pharmacogenomic

landscape of patient-derived tumor cells informs precision oncology therapy. *Nat. Genet.*, 50(10): 1399–1411, October 2018. 14

[129] Su-In Lee, Safiye Celik, Benjamin A Logsdon, Scott M Lundberg, Timothy J Martins, Vivian G Oehler, Elihu H Estey, Chris P Miller, Sylvia Chien, Jin Dai, Akanksha Saxena, C Anthony Blau, and Pamela S Becker. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nature communications*, 9(1):42, January 2018. 16, 37

[130] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007. 18, 61

[131] Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008. 18, 61

[132] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. *International Conference on Machine Learning*, pages 1718–1727, 2015. 35

[133] Julio Licinio and Ma-Li Wong. *Pharmacogenomics: the search for individualized therapies*. John Wiley & Sons, 2009. 4

[134] Alexander Ling, Robert F Gruener, Jessica Fessler, and R Stephanie Huang. More than fishing for a cure: The promises and pitfalls of high throughput cancer cell line screens. *Pharmacol. Ther.*, June 2018. 13, 14

[135] YM Lo, Noemi Corbetta, Paul F Chamberlain, Vik Rai, Ian L Sargent, Christopher WG Redman, and James S Wainscoat. Presence of fetal DNA in maternal plasma and serum. *The Lancet*, 350 (9076):485–487, 1997. 1

[136] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature methods*, 12(8):733, 2015. 21

[137] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. *International Conference on Machine Learning*, pages 97–105, 2015. 34

[138] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217, 2017. 34

[139] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *International Conference on Learning Representations, arXiv:1511.00830*, 2016. xi, 20, 23, 27, 28, 32, 33, 42, 55, 56, 62, 90

[140] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*, 2016. 23

[141] Ricardo Macarron, Martyn N Banks, Dejan Bojanic, David J Burns, Dragan A Cirovic, Tina Garyantes, Darren V S Green, Robert P Hertzberg, William P Janzen, Jeff W Paslay, Ulrich Schopfer, and G Sitta Sittampalam. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.*, 10(3):188–195, March 2011. 13

[142] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 27

[143] Michael A Martin and Deanna L Kroetz. Abacavir pharmacogenetics–from initial reports to standard of care. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 33 (7):765–775, 2013. 5

[144] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 80

[145] P. Medvedev, M. Stanciu, and M. Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nature methods*, 6:S13–S20, 2009. 96

[146] C Meijer, N H Mulder, H Timmer-Bosscha, W J Sluiter, G J Meersma, and E G de Vries. Relationship of cellular glutathione to the cytotoxicity and resistance of seven platinum compounds. *Cancer Res.*, 52(24):6885–6889, December 1992. 19

[147] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. *arXiv preprint arXiv:1802.08665*, 2018. 27

[148] Michael P Menden, Francesco Iorio, Mathew Garnett, Ultan McDermott, Cyril H Benes, Pedro J Ballester, and Julio Saez-Rodriguez. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One*, 8(4):e61318, April 2013. 17

[149] Michael Patrick Menden, Dennis Wang, Yuanfang Guan, Michael Mason, Bence Szalai, Krishna C Bulusu, Thomas Yu, Jaewoo Kang, Minji Jeon, Russ Wolfinger, Tin Nguyen, Mikhail Zaslavskiy, Astrazeneca-Sanger Drug Combination, In Sock Jang, Zara Ghazoui, Mehmet Eren Ahsen, Robert Vogel, Elias Chaibub Neto, Thea Norman, Eric K Y Tang, Mathew J Garnett, Giovanni Di Veroli, Stephen Fawell, Gustavo Stolovitzky, Justin Guinney, Jonathan R Dry, and Julio Saez-Rodriguez. A cancer pharmacogenomic screen powering crowd-sourced advancement of drug combination prediction. *bioRxiv*, page 200451, February 2018. 16, 37

[150] Arvind Singh Mer, Wail Ba-alawi, Petr Smirnov, Yi Xiao Wang, Ben Brew, Janosch Ortmann, Ming-Sound Tsao, David Cescon, Anna Goldenberg, and Benjamin Haibe-Kains. Integrative pharmacogenomics analysis of patient derived xenografts. *Accepted in Cancer Research*, page 471227, 2019. 14

[151] Soufiane Mourragui, Marco Loog, Mark A van de Wiel, Marcel JT Reinders, and Lodewyk FA Wessels. Precise: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics*, 35(14):i510–i519, 2019. 18, 28, 35, 60, 61, 62

[152] John Patrick Mpindi, Bhagwan Yadav, Päivi Östling, Prson Gautam, Disha Malani, Astrid Murumägi, Akira Hirasawa, Sara Kangaspeska, Krister Wennerberg, Olli Kallioniemi, et al. Consistency in drug response profiling. *Nature*, 540(7631):E5, 2016. 14

[153] National Cancer Institute. Identifying novel drug combinations to overcome treatment resistance. *NCI webportal*, Dec 2016. URL `www.cancer.gov/about-cancer/treatment/research/drug-combo-resistance`. 107

[154] Mario Niepel, Marc Hafner, Qiaonan Duan, Zichen Wang, Evan O Paull, Mirra Chung, Xiaodong Lu, Joshua M Stuart, Todd R Golub, Aravind Subramanian, et al. Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. *Nature Communications*, 8(1):1186, 2017. 38, 47

[155] Raffaele Palmirotta, Domenica Lovero, Paola Cafforio, Claudia Felici, Francesco Mannavola, Eleonora Pellè, Davide Quaresmini, Marco Tucci, and Franco Silvestris. Liquid biopsy of cancer: a multimodal diagnostic tool in clinical oncology. *Therapeutic advances in medical oncology*, 10: 1758835918794630, 2018. 1, 108

[156] Simon Papillon-Cavanagh, Nicolas De Jay, Nehme Hachem, Catharina Olsen, Gianluca Bontempi, Hugo J W L Aerts, John Quackenbush, and Benjamin Haibe-Kains. Comparison and validation of genomic predictors for anticancer drug sensitivity. *Journal of the American Medical Informatics Association*, 20(4):597–602, July 2013. 16, 37

[157] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Autodiff Workshop*, 2017. 52

[158] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011. 43, 64

[159] Frederique Penault-Llorca, Catherine Abrial, Ines Raoelfils, Anne Cayre, Marie-Ange Mouret-Reynier, Marianne Leheurteur, Xavier Durando, Jean-Louis Achard, Pierre Gimbergues, and Philippe Chollet. Comparison of the prognostic significance of chevallier and sataloff's pathologic classifications after neoadjuvant chemotherapy of operable breast cancer. *Human pathology*, 39(8): 1221–1228, 2008. 62

[160] Jose Luis Perez-Gracia, Miguel F Sanmamed, Ana Bosch, Ana Patino-Garcia, Kurt A Schalper, Victor Segura, Joaquim Bellmunt, Josep Tabernero, Christopher J Sweeney, Toni K Choueiri, Miguel Martín, Juan Pablo Fusco, Maria Esperanza Rodriguez-Ruiz, Alfonso Calvo, Celia Prior, Luis Paz-Ares, Ruben Pio, Enrique Gonzalez-Billalabeitia, Alvaro Gonzalez Hernandez, David Páez, Jose María Piulats, Alfonso Gurpide, Mapi Andueza, Guillermo de Velasco, Roberto Pazo, Enrique Grande, Pilar Nicolas, Francisco Abad-Santos, Jesus Garcia-Donas, Daniel Castellano, María J Pajares, Cristina Suarez, Ramon Colomer, Luis M Montuenga, and Ignacio Melero. Strategies to design clinical studies to identify predictive biomarkers in cancer research. *Cancer Treat. Rev.*, 53: 79–97, February 2017. 13

[161] Vinay Prasad, Victoria Kaestner, and Sham Mailankody. Cancer drugs approved based on biomarkers and not tumor Type—FDA approval of pembrolizumab for mismatch Repair-Deficient solid cancers. *JAMA Oncol*, 4(2):157–158, February 2018. 11

[162] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, pages 2352–2360, 2016. 23

[163] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. 21

[164] V N Rajapakse, A Luna, M Yamade, L Loman, S Varma, and others. Integrative analysis of pharmacogenomics in major cancer cell line databases using CellMinerCDB. *bioRxiv*, 2018. 14

[165] Ladislav Rampášek, Aryan Arbabi, and Michael Brudno. Probabilistic method for detecting copy number variation in a fetal genome using maternal plasma sequencing. *Bioinformatics*, 30(12): i212–i218, 2014. 2, 3, 92

[166] Ladislav Rampášek, Daniel Hidru, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics*, March 2019. URL https://doi.org/10.1093/bioinformatics/btz158. x, 2, 3, 18, 37, 62, 71

[167] Ladislav Rampášek, Daniel Hidru, Peter Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Modeling post-treatment gene expression change with a deep generative model. *11th International Workshop on Machine Learning in Systems Biology at ISMB/ECCB*, 2017. 2, 3, 18, 37

[168] Ladislav Rampášek, Daniel Hidru, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Drug response variational autoencoder. *MLCB workshop at Conference on Neural Information Processing Systems (NIPS)*, 2017. 18, 37

[169] Siamak Ravanbakhsh, Francois Lanusse, Rachel Mandelbaum, Jeff Schneider, and Barnabas Poczos. Enabling dark energy science with deep generative models of galaxy images. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 23

[170] Matthew G Rees, Brinton Seashore-Ludlow, Jaime H Cheah, Drew J Adams, Edmund V Price, Shubhroz Gill, Sarah Javaid, Matthew E Coletti, Victor L Jones, Nicole E Bodycombe, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature chemical biology*, 12(2):109, 2016. x, 14, 37, 40, 43, 71, 83

[171] William C Reinhold, Margot Sunshine, Hongfang Liu, Sudhir Varma, Kurt W Kohn, Joel Morris, James Doroshow, and Yves Pommier. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer research*, 72(14): 3499–3511, 2012. 37

[172] Mary V Relling and William E Evans. Pharmacogenomics in the clinic. *Nature*, 526(7573):343, 2015. 4

[173] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015. 27, 51

[174] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014. x, xi, 2, 20, 23, 26, 41

[175] Danilo Jimenez Rezende, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra. One-shot generalization in deep generative models. *arXiv preprint arXiv:1603.05106*, 2016. 23

[176] Allen D Roses. Pharmacogenetics in drug discovery and development: a translational perspective. *Nature reviews Drug discovery*, 7(10):807, 2008. 6

[177] Giovanna Rossi and Michail Ignatiadis. Promises and pitfalls of using liquid biopsy for precision medicine. *Cancer research*, 79(11):2798–2804, 2019. 108

[178] Zhaleh Safikhani, Petr Smirnov, Mark Freeman, Nehme El-Hachem, Adrian She, Quevedo Rene, Anna Goldenberg, Nicolai J Birkbak, Christos Hatzis, Leming Shi, et al. Revisiting inconsistency in large pharmacogenomic studies. *F1000Research*, 5(2333), 2016. 40

[179] Zhaleh Safikhani, Petr Smirnov, Kelsie L Thu, Jennifer Silvester, Nehme El-Hachem, Rene Quevedo, Mathieu Lupien, Tak W Mak, David Cescon, and Benjamin Haibe-Kains. Gene isoforms as expression-based biomarkers predictive of drug response in vitro. *Nature Communications*, 8(1): 1126, October 2017. 16, 37

[180] Steven L Salzberg. Open questions: How many genes do we have? *BMC biology*, 16(1):94, 2018. 6

[181] C.J. Saunders, N.A. Miller, S.E. Soden, D.L. Dinwiddie, S.F. Kingsmore, et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Science Translational Medicine*, 4(154):154ra135, 2012. 93

[182] Lawrence H Schwartz, Saskia Litière, Elisabeth de Vries, Robert Ford, Stephen Gwyther, Sumithra Mandrekar, Lalitha Shankar, Jan Bogaerts, Alice Chen, Janet Dancey, et al. Recist 1.1—update and clarification: From the recist committee. *European journal of cancer*, 62:132–137, 2016. 62

[183] Hossein Sharifi-Noghabi, Olga Zolotareva, Colin C Collins, and Martin Ester. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14): i501–i509, 07 2019. 18, 62, 107

[184] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001. 7

[185] Robert H Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, 6(10):813–823, October 2006. 13, 14

[186] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018. 28, 36

[187] Petr Smirnov, Zhaleh Safikhani, Nehme El-Hachem, Dong Wang, Adrian She, Catharina Olsen, Mark Freeman, Heather Selby, Deena MA Gendoo, Patrick Grossmann, et al. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics*, 32(8):1244–1246, April 2015. 14, 38, 40, 84

[188] Petr Smirnov, Victor Kofia, Alexander Maru, Mark Freeman, Chantal Ho, Nehme El-Hachem, George-Alexandru Adam, Wail Ba-Alawi, Zhaleh Safikhani, and Benjamin Haibe-Kains. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Research*, 46(D1):D994–D1002, January 2018. 13, 14, 40

[189] Zbyslaw Sondka, Sally Bamford, Charlotte G Cole, Sari A Ward, Ian Dunham, and Simon A Forbes. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, page 1, 2018. 9

[190] Anupama Srinivasan, Diana W Bianchi, Hui Huang, Amy J Sehnert, and Richard P Rava. Noninvasive detection of fetal subchromosome abnormalities via deep sequencing of maternal plasma. *The American Journal of Human Genetics*, 92(2):167–176, 2013. 93, 96, 98, 99, 103

[191] Lindsay C Stetson, Taylor Pearl, Yanwen Chen, and Jill S Barnholtz-Sloan. Computational identification of multi-omic correlates of anticancer therapeutic response. *BMC genomics*, 15(7):S2, October 2014. 16

[192] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, David L Lahr, Jodi E Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M Enache, Federica Piccioni, Sarah A Johnson, Nicholas J Lyons, Alice H Berger, Alykhan F Shamji, Angela N Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S Gray, Paul A Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J Haggarty, Lucienne V Ronco, Jesse S Boehm, Stuart L Schreiber, John G Doench, Joshua A Bittker, David E Root, Bang Wong, and Todd R Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, November 2017. x, 38, 40, 43, 107

[193] Jingchun Sun, Qiang Wei, Yubo Zhou, Jingqi Wang, Qi Liu, and Hua Xu. A systematic analysis of fda-approved anticancer drugs. *BMC systems biology*, 11(5):87, 2017. ii

[194] Wei Sun, Philip E Sanderson, and Wei Zheng. Drug combination therapy increases successful drug repositioning. *Drug Discov. Today*, 21(7):1189–1195, July 2016. 19

[195] Xiao-Xiao Sun and Qiang Yu. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacol. Sin.*, 36(10):1219–1227, October 2015. 19

[196] Mehmet Tan, Ozan Fırat Özgül, Batuhan Bardak, Işıksu Ekşioğlu, and Suna Sabuncuoğlu. Drug response prediction by ensemble learning and drug-induced gene expression signatures. *Genomics*, 2018. ISSN 0888-7543. doi: 10.1016/j.ygeno.2018.07.002. 37

[197] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research*, 47(D1):D941–D947, 2018. 7

[198] Uduak Grace Thomas. Researchers develop more precise statistical tools for identifying novel CNVs in fetal DNA. *GenomeWeb*, July 2014. URL www.genomeweb.com/informatics/researchers-develop-more-precise-statistical-tools-identifying-novel-cnvs-fetal. 2, 92

[199] Winston Timp, Jeffrey Comer, and Aleksei Aksimentiev. Dna base-calling from a nanopore using a viterbi algorithm. *Biophysical journal*, 102(10):L37–L39, 2012. 21

[200] Jakub Tomczak and Max Welling. Vae with a vampprior. *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223, 2018. 34

[201] Jakub M Tomczak and Max Welling. Improving variational auto-encoders using householder flow. *arXiv preprint arXiv:1611.09630*, 2016. 27

[202] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010. 63

[203] Kenneth Y Tsai and Hensin Tsao. Primer on the human genome. *Journal of the American Academy of Dermatology*, 56(5):719–735, 2007. 6

[204] George Tucker, Andriy Mnih, Chris J Maddison, John Lawson, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. *Advances in Neural Information Processing Systems*, pages 2627–2636, 2017. 27

[205] Julianne D Twomey, Nina N Brahme, and Baolin Zhang. Drug-biomarker co-development in oncology - 20 years and counting. *Drug Resist. Updat.*, 30:48–62, January 2017. 11

[206] Thomas Unterthiner, Andreas Mayr, Günter Klambauer, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, and Sepp Hochreiter. Deep learning as an opportunity in virtual screening. In *Proceedings of the deep learning workshop at NIPS*, volume 27, pages 1–9, 2014. 17

[207] Eric Wang, Annette Batey, Craig Struble, Thomas Musci, Ken Song, and Arnold Oliphant. Gestational age and maternal weight effects on fetal cell-free dna in maternal plasma. *Prenatal diagnosis*, pages 1–5, 2013. 92, 93

[208] Lin Wang, Xiaozhong Li, Louxin Zhang, and Qiang Gao. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC cancer*, 17(1): 513, August 2017. 16, 37

[209] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153, 2018. 28

[210] Gregory P Way and Casey S Greene. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *Pacific Symposium on Biocomputing*, volume 23, page 80. NIH Public Access, 2018. 18, 23, 38, 62

[211] Michelle Whirl-Carrillo, Ellen M McDonagh, JM Hebert, Li Gong, K Sangkuhl, CF Thorn, Russ B Altman, and Teri E Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, 2012. 4

[212] Jennifer L Wilding and Walter F Bodmer. Cancer cell lines for drug discovery and development. *Cancer research*, 74(9):2377–2384, 2014. 62

[213] Catherine M Worsley, Elizabeth S Mayne, and Rob B Veale. Clone wars: the evolution of therapeutic resistance in cancer. *Evol Med Public Health*, 2016(1):180–181, June 2016. 19

[214] Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. *International Conference on Machine Learning*, pages 6872–6881, 2019. 29, 35, 36, 65, 68, 69, 90

[215] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, Sridhar Ramaswamy, P Andrew Futreal, Daniel A Haber, Michael R Stratton, Cyril Benes, Ultan McDermott, and Mathew J Garnett. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, January 2013. 37

[216] Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3881–3890. JMLR.org, 2017. 23

[217] Byung-Jun Yoon. Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10(6):402–415, 2009. viii, 20

[218] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (CMD) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017. 28, 35, 62

[219] Fei Zhang, Minghui Wang, Jianing Xi, Jianghong Yang, and Ao Li. A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Scientific reports*, 8(1):3355, February 2018. 16, 17, 37

[220] Naiqian Zhang, Haiyun Wang, Yun Fang, Jun Wang, Xiaoqi Zheng, and X Shirley Liu. Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS computational biology*, 11(9):e1004498, 2015. 37

[221] Cheng Zhao, Ying Li, Zhaleh Safikhani, Benjamin Haibe-Kains, and Anna Goldenberg. Using cell line and patient samples to improve drug response prediction. *bioRxiv*, page 026534, 2015. 18, 60, 62, 106

[222] Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019. 28, 29, 35, 36

[223] Ji Zhao and Deyu Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372, 2015. 33

[224] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised representation learning: Transfer learning with deep autoencoders. *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. 34