**University of Warsaw**
Faculty of Mathematics, Informatics and Mechanics

**Krzysztof Koras**
Student no. 321779

# Computational methods for anti-cancer drug sensitivity prediction

**PhD dissertation**
**in COMPUTER SCIENCE**

Supervisor:
**Dr hab. Ewa Szczurek**
Institute of Informatics

Warsaw, June 2022

## Abstract

Computational models for drug sensitivity prediction have the potential to significantly improve personalized cancer medicine. Drug sensitivity assays, combined with molecular profiling of cancer cell lines and drugs become increasingly available for training such models. Existing models largely differ in terms of the modeling framework, utilized data and modeling objectives. This thesis is devoted to comprehensive modeling of drug sensitivity data and builds upon three projects. In the first one, we comprehensively developed and evaluated several feature selection strategies for per-drug sensitivity prediction. In the second, we developed a deep recommender system for prediction of kinase inhibitors efficacy across cancer cell lines, with a tailored model interpretability approach. The third project established a methodology for clustering of the latent data representations within a variational autoencoder framework, with an application to drug sensitivity prediction and new compounds generation. The thesis highlights crucial challenges regarding the problem of drug sensitivity prediction problem and provides several means to address them. Specifically, research topics include feature selection, multi-task learning, model intepretability, representation learning and generative modeling. The research presented in the thesis naturally evolved from using well-established machine learning models with more emphasis put on data exploratory side, to developing custom methods based on neural networks and generative modeling, introducing novel technical solutions.

## Keywords

machine learning, deep learning, generative modeling, personalized medicine, drug sensitivity prediction, kinase inhibitors, cancer cell lines, recommender system, autoencoder, interpretability, variational autoencoder, gaussian mixture model


## Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatics, Computer Science

## Subject classification

Applied computing → Life and medical sciences → Computational biology
Computing methodologies → Machine learning → Supervised learning → Supervised learning by regression

## Tytuł pracy w języku polskim

Obliczeniowe metody przewidywania wrażliwości na leki przeciwnowotworowe

## Acknowledgments

First, I would like to thank my supervisor, dr hab. Ewa Szczurek. This thesis, as well as my other research during PhD studies would not be possible without her. Above all, I am grateful for Ewa's continuous support as well as her incredibly competent and professional scientific guidance and supervision.

Second, I would like to acknowledge dr Eike Staub from Merck KGaA. I am very grateful for being part of a scientific collaboration between our Faculty and Merck, in which Eike's supervision from the Merck side was a huge factor in making it such an engaging and interesting experience.

I would also like to thank all of my collaborators with whom I worked on my projects: Ewa Kizling, Marcin Możejko and Paulina Szymczak from the University of Warsaw, as well as Johanna Mazur, Dilafruz Juraeva and Julian Kreis from Merck KGaA.

Finally, I am also grateful to all my other colleagues from Ewa Szczurek's lab for creating such an enjoyable work environment. Without them, my PhD research would definitely not be as pleasant experience as it was.

This study was supported by the research grant
2019/33/B/NZ2/00956
from the National Science Centre, Poland.

The work presented in this dissertation has been published in the following research papers and preprints:

Krzysztof Koras, Dilafruz Juraeva, Julian Kreis, Johanna Mazur, Eike Staub, and Ewa Szczurek. Feature selection strategies for drug sensitivity prediction. *Scientific Reports* 10, 9377 (2020). https://doi.org/10.1038/s41598-020-65927-9

Krzysztof Koras, Ewa Kizling, Dilafruz Juraeva, Eike Staub, and Ewa Szczurek. Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines. *Scientific reports* 11, 15993 (2021). https://doi.org/10.1038/s41598-021-94564-z

Krzysztof Koras, Marcin Możejko, Paulina Szymczak, Eike Staub, and Ewa Szczurek. A generative recommender system with GMM prior for cancer drug generation and sensitivity prediction. *arXiv* (2022). https://doi.org/10.48550/arxiv.2206.03555

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Cancer is one of the leading causes of death worldwide. According to the International Agency for Research on Cancer, in 2020 there were nearly 20 million new cases and nearly 10 million cancer-related deaths worldwide [1]. Approximately 39.5% of people will be diagnosed with cancer at some point during their lifetimes [2]. The national costs for cancer care in the United States were estimated at 208.9 billion USD in 2020, and are projected to grow in the next years. Notably, this expected growth of costs is not only due to the projected increase of cancer incidents, but also to the fact that new, more effective, but also more expensive forms of treatment are being adopted as a standard of care.

Cancer is a disease which occurs when some of the body's cells become abnormal and start to divide uncontrollably [3]. Cancer cells arise from the accumulation of genetic alterations leading to the abnormal expression of mRNA and proteins which disrupt the cellular mechanisms of growth and division control. These cells can form lumps of tissue referred to as tumors. Cancerous (or malignant) tumors can invade nearby tissues or spread to other parts of a body through the blood or lymph system in a process of metastasis [4]. Metastasis is responsible for a majority of cancer-related deaths [5], with some sources estimating the death rate for metastatic cancers as high as 90% [6, 7].

One of the biggest challenges in cancer treatment is tumor heterogeneity. The inter-patient and inter-tumor heterogeneity refers to the fact that cancers that are similar at the macroscopic level can be very different at the molecular level, e.g. having different mutation or gene expression profiles. These differences can account for significant variation in treatment outcomes and disease prognosis. In order to mitigate this problem, there is a growing focus on precision, or personalized, medicine in cancer [8, 9, 10]. National Cancer Institute (NCI) defines precision medicine as "a form of medicine that uses information about a person's own genes or proteins to prevent, diagnose, or treat disease" [11]. In cancer treatment, that means utilizing the molecular description of a specific tumor in order to tailor the most optimal available treatment.

While modern sequencing technologies, including next-generation sequencing (NGS) [12], have enabled in-depth profiling of cancer cells and tumors, the existing approved biomarkers for treatment are mostly limited to a single gene or a combination of few genes [8, 9]. NGS produces inherently high-dimensional data which is hard to interpret for humans. Therefore, there is a need for computational models for drug sensitivity prediction (DSP), which are able to predict treatment outcomes based on high-dimensional characterizations of given cancer cells. Drug sensitivity is also often referred to as drug response.

The thesis is devoted to the problem of computational drug sensitivity prediction. It builds upon three research projects that I worked on during my PhD studies together with

collaborators. Although projects deal with the same general problem, they largely differ in terms of modeling framework, utilized data, and machine learning (ML) methodologies used.

The thesis is organized as follows. In the rest of Chapter 1, I describe the biggest research challenges present in the field and how research presented in the thesis addresses these challenges. Chapters 2 and 3 provide preliminaries helpful in comprehending the presented work, considering cancer biology, data characterizations and machine learning aspects. In chapters 4, 5 and 6, I present the projects which make up for the thesis. Besides description of methods and results, each of those chapters provides a more project-specific background and discussion. Finally, Chapter 7 comprises a summary and discussion of all presented work in a broader context.

## 1.1. Research topics covered in the thesis

### 1.1.1. Challenges

Traditional statistical models and more complex machine learning approaches can be utilized to predict drug sensitivity from existing data. However, as complexity of these models increases, they require more observations to train them [13]. Although this is especially the case for deep learning models, traditional ML methods also require sufficient number of observations for training, especially in high-dimensional settings. Clinical outcomes of patients combined with molecular profiles of their tumors would constitute the most robust dataset for clinically relevant models for DSP, however, these data sources are usually limited in size due to factors such as high costs, problems with data sharing and anonymization, or other legal regulations. In addition, testing multiple therapeutics on the same patient is infeasible in clinical practice. Consequently, most of the methods for computational DSP resort to *in vivo* cancer models, or *in vitro* cancer models in the form of immortalized cancer cell lines (CCL) or cell lines derived from patient biopsies, with cancer cell lines being the most common data source for training of DSP models. Although the drawbacks of cell line data have been raised and extensively studied [14, 15, 16, 17], these resources remain a vital tool for development of DSP models, due to the large amounts of drug screening data and the depth of molecular profiling of cell lines [18, 19].

The problem of drug sensitivity prediction in *in vitro* cancer models can be stated as follows: given molecular features of cancer cell lines and/or features of drugs $X$ and quantified drug sensitivity $y$, the goal is to find a function $f(X)$ which estimates $y$, where $f$ can be approximated by a machine learning model (Fig. 1.1). Computational DSP can be therefore seen as a common machine learning prediction task, with quantified drug sensitivity as a target variable. If $y$ is continuous, this takes a form of a classical regression task.

Nevertheless, although DSP in essence is a standard ML prediction task, there are number of domain-specific challenges to be considered. These challenges arise from several factors, such as high dimensionality of data, relatively small data abundance, lack of well-defined approaches to evaluate the models, or model interpretability. Research presented in this thesis addresses some of these challenges. Specifically, the thesis concerns the following research topics:

**Feature selection** The problem of choosing appropriate features which are most suitable for modeling drug sensitivity has two major aspects. In the high-level view, this corresponds to choosing appropriate type of data, such as different omics measurements of cell lines and different molecular descriptions of drugs. While, in the case of cell lines, gene expression and mutations emerge as an important data type suitable for many task including DSP, in the

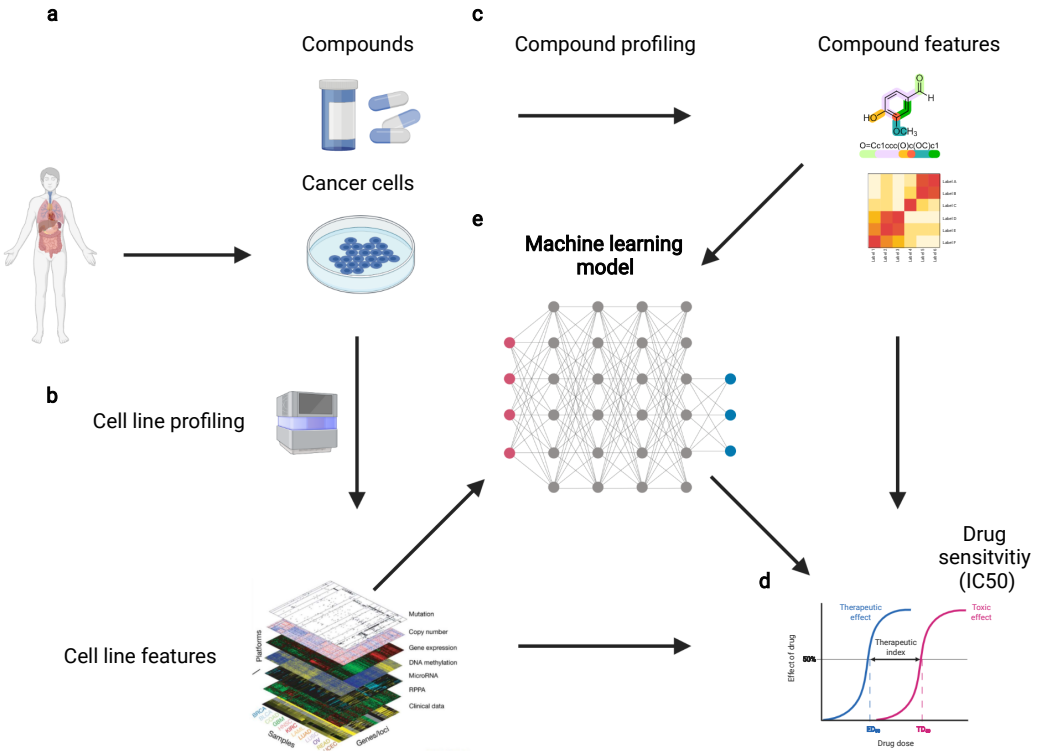


Figure 1.1: **Drug sensitivity prediction modeling workflow.** (**a**) Cancer cell lines (or cancer cells, or other cancer models) originating from human tumor samples are screened against a set of compounds. (**b**) Cell lines are profiled, for example using NGS, in order to extract multi-omics features. Those features serve as input to ML model. (**c**) Similarly, compounds can be profiled, for example by describing their chemical formula, fingerprints, or other, higher-level data. Those features can also serve as as input to ML model. (**d**) Drug sensitivity is a function of cell line's and compound's features. True values are obtained by lab screening, but can be predicted by the trained ML model provided with cell line's and compound's features. Figure created with `BioRender.com`.

case of drugs this problem is more nuanced with different data types having their advantages and disadvantages. The second, lower-level aspect concerns which features to choose when general data type is already fixed, e.g. expression or mutation of which genes are the most indicative of the drug sensitivity.

**Multi-task learning** One important choice regarding the modeling framework in DSP is between building independent per-drug models or building a single, multi-task model, which can predict drug sensitivity for multiple drugs. While, in principle, per-drug models can be more accurate by utilizing specific drug's properties, their major disadvantage is the lack of possibility to model relationships between drugs' features and their action, which can lead to important biological insights. On the other hand, multi-task models can capture action of multiple drugs on multiple cancer cell lines via single model, which makes it more universal as well as provide more means to intepret the model.

**Interpretability** One of the main concerns generally in ML, and specifically in the DSP field, is the interpretability of ML models. It is hard to define interpretability mathematically, however, it can be defined non-mathematically in several ways. One of the suitable definitions of interpretability is "the degree to which an observer can understand the cause of a decision" [20, 21]. In general, the higher the model's interpretability, the easier it is for someone to comprehend why certain decisions or predictions have been made [22]. Interpretablity of ML models is especially important in the field of computational oncology, or, more broadly, computational medicine. In the problem of DSP specifically, one of the most important questions that could be answered by interpretability analysis is: what are the underlying chemical and biological mechanisms which drive the drug sensitivity and are captured by the model?

**Representation learning** Manual feature engineering can be infeasible, especially in the high-dimensional setting and complex problems like DSP. One of the reasons behind the success of neural networks in many applications is their ability to automatically learn representations of input data that are most suitable for a given task. Although this is especially transparent in the fields such as computer vision and natural language processing, representation learning is very beneficial also in the field of computational biology. Specifically, it can be used to find the most informative, lower-dimensional representations of drugs and cell lines. Those representations extract the most important information from the large sets of initial input features, can be passed as an input to the subsequent models performing predictions, or be visualized and analyzed to increase model's interpretability.

**Generative modeling** Recently, there is a growing interest in building models which are not only able to predict certain variables, but can also generate new objects. In the field of computational medicine, a particular interest concerns generation of new drug candidates. Generative modeling requires that objects such as drugs or cell lines are not modeled as a single data points, but rather probability distributions over some variables. The main challenge in this context is to generate objects which have some desired properties and are valid, e.g. drugs which are effective against some type of cancer and are possible to synthesize.

### 1.1.2. How the research presented in this thesis addresses the challenges

The thesis builds upon three research projects generally concerning the problem of computational drug sensitivity prediction. Each of those projects is addressing the challenges listed above, but with differing methodologies and different points of emphasis:

**Per-drug feature selection strategies for drug sensitivity prediction [23].** The focal point of the first project was feature selection. As large-scale pharmacogenomics studies provide a wide range of biological measurements, the high dimensionality of such data makes it difficult to develop effective predictive models and to identify the features which are the real drivers for cell lines' response to treatment. Here, we compared standard, data-driven feature selection methods to feature selection based on prior knowledge regarding drug targets, target pathways, and gene expression signatures. We assessed these methodologies on a pharmacogenomic dataset, evaluating a total of 2484 ML models. Such drugs for which smaller feature sets yielded better predictive performance were identified and described.

**Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines [24].** The second work considered aspects of high-level feature selection, representation learning and model interpretability. Drug sensitivity prediction problem can also be defined as a recommendation problem, where the goal is to recommend the best drug for a given cancer cell line. Here, rather than building a separate model for each compound, we developed DEERS, a multi-task deep recommender system which achieves this objective for multiple drugs and cell lines. The model utilizes both molecular features of the cancer cell lines as well as kinase inhibition profiles of the drugs. We introduced a novel model interpretability approach, which in addition to the set of modeled features considers also the genes and processes outside of the features set used for training. Predictive performance of DEERS was evaluated against simpler matrix factorization approaches and common ML models. The novel interpretability approach was utilized to identify a group of biological processes that are most responsible for driving cancer cell lines' response to treatment.

**A generative recommender system with GMM prior for cancer drug generation and sensitivity prediction [25].** Finally, in the third work, the focus was switched to deep generative modeling, specifically variational autoencoders (VAE), which also involved representation learning and model interpretablity. In particular, we established a methodology for distribution-based clustering of of data representations in the variational autoencoder's latent space using external guiding data through gaussian mixture models (GMM). In principle, the developed method can be applied to a broad spectrum of vector-represented data. However, to put it in the context of DSP, this new model was applied on drugs' SMILES vector representations as input data. SMILES, standing for Simplified Molecular Input Line Entry System, is a line notation system used for describing the structure of chemical species using short strings [26, 27, 28]. In this case, the guiding data were the drugs' inhibition profiles (which were also utilized in the second project but using standard autoencoders). Here, we developed a drug variational autoencoder, which takes drugs' SMILES vector representations as input, and outputs reconstructed SMILES representations along with predicted inhibition profiles across panel of protein kinases. This VAE uses GMM with learnable parameters as a latent space prior distribution, where component assignments are observed for a subset of drugs. This drugs VAE model was incorporated into larger modeling framework along with cancer cell lines' data, creating VADEERS, an extension to the DEERS model capable of predicting cell lines response to drugs and drugs' inhibition profiles. The variational autoencoder part of the model can be utilized to generate new drugs with properties defined by the guiding data used for clustering.

# Chapter 2

# Cancer biology

## 2.1. Cancer development and properties

Since cancer can result from abnormal proliferation of any kind of cells, there are many ways to classify cancers. One important distinction is between benign and malignant tumors. A benign tumor remains constrained to its original location, not posing a threat to other organs and making it susceptible to localized treatment such as surgical intervention. On the other hand, a malignant tumor is able to invade surrounding normal tissues and metastize, which makes it dangerous and resistant to localized forms of treatment. Only malignant tumors are properly reffered to as cancers. Tumors can then be classified according to the type of cells from which they arise, falling mostly into one of three groups: carcinomas, sarcomas, and leukemias or lymphomas [29]. Further classification involves tumor's or cancer's site of origin, e.g. lung cancer or prostate cancers.

One of the most fundamental characteristics of cancer is tumor clonality, i.e. the development of tumors from single, abnormal cells [29]. However, this tumor clonality does not mean that the abnormal cell of origin has initially possessed all of the attributes of a cancer cell. On the contrary, the development of cancer is a multi-step process, in which cells become more malignant through a progressive series of alterations (mutations) in DNA sequence. These successive rounds of mutation and selective expansion of these abnormal cells eventually results in the formation of a tumor mass [30]. Therefore, tumor progression creates clones, some of which acquire mutations giving them selective advantage in the overall tumor population. This process of clonal evolution [31] gives rise to tumor heterogeneity and significantly affects treatment. For example, application of a general cytotoxic drug may initially kill majority of cancer cells, but it also may create an evolutionary niche for remaining, treatment-resistant population of cells and allow them to become dominant within the tumor.

The uncontrolled proliferation of cancer cells *in vivo* can be mimicked *in vitro* through cell cultures. Cells do not exist in isolation, their behavior is dependent on signals coming from the surrounding environment, such as growth factors which trigger cell division. These external growth factors (or ligands) bind to membrane-bound glycoprotein receptors that transmit the message via a series of intracellular signals that promote or inhibit the expression of specific genes [30]. In a cell culture, normal cells will cease to proliferate when reaching a finite cell concentration, which is dependent on the availability of growth factors in a culture medium. Cancer cells, on the other hand, can continue to grow and divide without dependence on external growth factors, increasing the chances of acquiring further mutations in a genome. There are three main mechanisms through which cancer cells achieve the independence from external growth factors [30]. Firstly, cancer cells can produce their own growth factors,

stimulating their own proliferation. Secondly, they can induce changes in, or change the number of cell surface protein receptors which transmit growth-stimulatory signals into the cell interior, making the cell more responsive to external growth factors. Finally, they can induce changes in intracellular growth factor signaling pathways, leading to the undesired signals triggering cell proliferation. Many of the aforementioned cancer mechanisms are caused by unfavorable mutations in proto-oncogenes. A proto-oncogene is a gene involved in normal cell growth [32]. However, upon activation, a proto-oncogene can turn into oncogene [33], producing oncoproteins, which in turn predispose cell to be cancerous. Oncoproteins, as well as other proteins involved in downstream and upstream of key signaling pathways are often molecular targets of anti-cancer therapeutics.

Another fundamental feature of cancer cells is their ability to bypass anti-growth signals. Besides growth factors, normal cells are also subjected to signals which are responsible for cell quiescence, which serve as brakes against proliferation signals [30]. Similar to growth factor signaling pathways, signals that normally suppress cell division are also received by cell-surface receptors that are coupled to intracellular signaling pathways. The genes that encode this class of proteins involved in restraining normal cell division are termed tumour suppressor genes [34]. Mutations in these genes can lead to loss-of-function and increased cell division. Probably the most known tumor supressor gene is *TP53*. About 50% of cancers have mutations in *TP53* [35, 36]. Protein product of *TP53* called p53 is responsible for regulating cell cycle progression and programmed cell death, or apoptosis. Cells exhibiting loss-of-function of the p53 protein fail to undergo apoptosis in response to DNA damage caused by treatment such as radiation and chemotherapy drugs. This mechanism largely contributes to the resistance of many tumors to chemotherapy.

Another important characteristics of cancer cells concern their interactions with other tissue elements. One of the most important and harmful steps in tumor development is the process of angiogenesis, i.e. formation of new blood vessels. After reaching some size, tumors require these new vessels to provide nutrients and oxygen to cancer cells. The new blood vessels are created in response to growth factors secreted by cancer cells themselves, which promote proliferation of endothelial cells in the walls of capillaries in the neighboring tissue, resulting in the growth of new capillaries into the tumor [29]. One transparent example of such pro-angiogenic growth factor is vascular endothelial growth factor (VEGF) [37, 38]. Angiogenesis is crucial from the standpoint of drug design, since it is a key process in transition between relatively small, constrained tumor, to an agressive tumor able to invade other tissues. In addition, new capillaries formed during angiogenesis are easily penetrable by tumor cells, enabling them to spread via circulatory system, laying the foundation for the process of metastasis, which is the main cause of death in cancer patients [5, 6, 7].

## 2.2. Drug action

Depending on how it works, the pharmacological treatment of cancer can be broadly classified into four major types: chemotherapy, hormone therapy, immunotherapy and targeted therapy [39, 40].

The term "chemotherapy" is usually reserved for compounds which are generally cytotoxic and are therefore used to kill cancer cells [41]. Chemotherapeutics target the cell cycle, interfering with cell proliferation by damaging its DNA and RNA and their metabolism [42]. Because of its general cytotoxicity and non-specificity to cancer cells, chemotherapy causes major undesired side effects [42, 43]. Hormone therapy acts upon cancers which rely on hormones to grow, by stopping the body from producing particular hormone, blocking the

hormone from ataching to cancer cells, or by altering hormone's function [44]. This type of therapy is commonly used in treatment of breast and prostate cancers, which rely on estrogen and testosterone, respectively. Immunotherapy is a type of biological therapy [42] which helps patient's own immune system to identify and attack cancer cells, which normally can avoid body's immunological response [45].

Targeted therapy is considered to be a cornerstone of precision medicine [46, 47]. Targeted therapy is based on drugs and compounds which suppress the growth and spread of cancer by targeting specific molecules involved in cancer development. In contrast to standard chemotherapy, targeted therapy is cancer-specific, i.e. it acts upon specific molecular targets affecting cancer cells, while leaving normal cells mostly unaffected [48]. In addition, targeted agents are usually cytostatic, blocking tumor proliferation, while standard chemotherapy drugs are cytotoxic, i.e. they kill existing cells [46].

The majority of targeted therapeutics belong to two categories: monoclonal antibodies and small-molecule drugs [49]. Monoclonal antibodies are proteins designed in the lab to attach to specific molecules present on cancer cells' surface [47]. On the other hand, because of their smaller size, small-molecule agents [50] can penetrate to cells' interior and are therefore used to target intracellular signaling molecules [47]. The type and molecular action of targeted therapeutics is often indicated in their names; small molecules with inhibitory properties have "-ib" as a suffix (e.g. erlotinib, imatinib), while monoclonal antibodies have "-mab" as a suffix (e.g. cetuximab, trastuzumab) [51]. There are several general biological mechanisms through which targeted drugs attempt to fight cancer. Signal transduction inhibitors suppress activity of molecules involved in signal transduction, or cell signaling, i.e. the process by which a cell responds to environmental signals through a cascade of biochemical reactions [46, 52]. Angiogenesis inhibitors stop formation of new blood vessels interfering with the action of related growth factors such as VEGF. Proteasome inhibitors [53] disrupt cancer cells' function causing the cells to die.

One particular, important category of targeted drugs are protein kinase inhibitors [54, 55]. Protein kinases are a class of enzymes which act by adding a phosphate group to a protein, modulating the protein's function and often making it active. They are often classified based upon amino acid which they phosphorylate: serine, threonine or tyrosine [56]. There are over 500 protein kinases in humans [54], involved in numerous cellular mechanisms and signaling pathways, including cell growth and proliferation. Some protein kinases are cell surface receptors which initiate and intracellular pathway of activation after the receptor binds with its ligand (e.g. growth factor), while others are intracellular and participate in signal transduction within a cell [56]. Mutations of protein kinases often lead to uncontrolled growth and proliferation, making these mutated variants potentially alluring drug targets. Inhibitors of cell surface receptors are called protein kinase receptor inhibitors, intracellular receptors are targeted by non-receptor kinase inhibitors, while some kinase inhibitors have specificity for multiple kinases and are referred to as multi-kinase inhibitors [56]. It is estimated that protein kinases constitute the second most targeted group of drug targets [54] and 20–33% of drug discovery efforts worldwide concentrate on the protein kinase family [57]. As of 2020, there were 52 small molecule protein kinase inhibitors approved by the US Food and Drug Administration (FDA), from which 11 target serine/threonine protein kinases, 2 are directed against dual specificity protein kinases, 11 inhibit non-receptor protein-tyrosine kinases, and 28 block receptor protein-tyrosine kinases [57].

Despite its advantage over conventional chemotherapy, targeted therapy is not without drawbacks. Some targeted therapeutics can cause undesired side effects, including skin problems, cardio-vascular problems, autoimmune responses, and other, more mild conditions [58]. Still, intra- and inter-tumor heterogeneity poses a great obstacle in treatment with targeted

drugs. Because of increased selectivity of targeted therapeutics, only a subset of patients with specific biomarkers can be effectively treated. Moreover, these therapeutics often prolong patents survival but fail to cure the disease due to the acquired resistance to of some tumor subclones [51]. Finally, the development of such drugs is not straightforward since some attractive biological targets are often not druggable, i.e. can't be inhibited by known chemical structures. Overall, these issues highlight the importance of approaches for appropriate matching of biological makeups of tumors to treatment, including computational models.

## 2.3. Data characterization

### 2.3.1. Cell viability studies

In recent years, experimental efforts by different research consortia have produced several public pharmacogenomic databases containing molecular and drug sensitivity profiles across a large number of cancer cell lines [59]. The drug screenings are usually conducted using a robotic system in which compounds are delivered to wells containing cell cultures of interest [60]. Following an incubation period lasting usually 72 hours upon drug delivery, a cell viability readout is conducted. This process is performed several times (commonly 5-10) with multiple drug concentrations, producing several dose-response data points, which can be fitted to produce a dose-response curve. Such dose-response curves are generalized in order to get a single, real-valued numerical indicator of drug sensitivity. Two commonly used univariate metrics include half maximal inhibitory concentration (IC50), which is a drug concentration required to reduce cell viability by 50% [61], and area under the dose-response curve (AUC). Therefore, in the case of both metrics, lower values indicate better drug efficacy. Overall, the protocol described above results in the table, or a matrix, containing drug sensitivity numerical indicators for each of the tested drug-cell line pair.

One of the first public pharmacogenomic resources was the NCI-60 dataset released by National Cancer Institute, containing screening data of thousands of compounds 60 cell lines spanning nine cancers [62, 63]. Notably, NCI-60 facilitated some important drug discoveries, including 26S proteasome inhibitor bortezomib used in multiple myeloma treatment [13]. Since then, larger-scale databases have been publicly released, including pan-cancer assays containing hundreds of cell lines from multiple tumor sites [13, 64]. One of the most prominent of such databases, also exploited the most in this thesis, is the Genomics od Drug Sensitivity in Cancer (GDSC) [65] database developed by Sanger Institute, currently containing drug screening measurements of hundreds of compounds across 1000 human cancer cell lines.

### 2.3.2. Genomics

Connecting drug sensitivity phenotypes with biological description of cell lines requires molecular profiling of the latter. These cells' characterizations can be extracted from different molecular levels such as genome, transcriptome, proteome, or metabolome. Two most exploited omics data types in this thesis are genomics, corresponding to DNA level and transcriptomics, corresponding to mRNA level or gene expression.

At the DNA level, two important variation types are single nucleotide polymorphisms and chromosomal rearrangements [60]. SNPs refer to variants in a single nucleotide block, e.g replacement of one nucleotide base with another. Such alterations can lead to the subsequent production of abnormal protein, making the cell cancerous. These variants can be detected through direct DNA sequencing, or, specifically, through whole exome sequencing (WES) [66, 67], measuring variations in the coding regions of the genome. This can result in a per-gene

binary information indicating presence or absence of a mutation in a given gene in a given cell line. Overall, from a technical perspective, mutation data used for prediction tasks including DSP is binary, and usually sparse containing mostly zeros, since for many considered genes only a small subset of cell lines exhibit point mutations in them.

Another type of genetic alterations, or mutations, may come from changes on the chromosomal level. In particular, copy number variations (CNV) indicate duplications of deletions of whole fragments of DNA, denoting number of copies of a given fragment or a segment [60]. For the purpose of DSP and other prediction tasks, it is beneficial to convert segment-level information to gene-level information, indicating number of copies of a given gene, which makes it more feasible to compare these features across cell lines. CNV are important in a context of cancer, since some cancer cells exhibit over-amplification of oncogenes or deletions of tumor-suppressor genes. Again, such preprocessed data produces gene-level discrete features for each cell line, with binary values, indicating presence or absence of abnormal number of copies, or values indicating number of copies itself. For many tasks, including work presented in this thesis, point mutation and CNV data is included for only a subset of highly-variate or cancer related genes, as opposed to considering all $\sim 20,000$ of human genes.

### 2.3.3. Transcriptomics

Gene expression is the process in which cells convert the information stored in DNA to functional proteins. The first step of this process involves transcription, in which DNA of a gene is copied into complementary fragment of RNA, producing a messenger RNA (mRNA) transcript. The message carried by mRNA is then used to synthesize a protein carrying out specific cellular functions. Therefore, measurement of mRNA can indicate which genes are active and describe the cell's type, state, or a condition.

Gene expression, or transcriptomics profiling is usually carried out using two common technologies used to derive gene expression, or transcriptomics profiles, are microarrays and RNA sequencing [68], with the latter being more modern and gaining more popularity [69]. Both tools measure the levels of mRNA transcripts across multiple genes and samples. Following post-processing and normalization, the end result are continuous numerical indicators of expression of considered genes, which can be used to differentiate between samples. In pharmacogenomic datasets, transcriptomics profiles of cancer cell lines are commonly whole genome-wide, i.e. they span $\sim 20,000$ coding genes [65, 18]. Such abundance of features naturally generates a problem of feature selection. Another technical difficulties associated with this datatype are correlation between features and noisiness of the data. Still, gene expression constitutes perhaps the most ubiquitous data type for representing cell lines in DSP and other fields of computational biology [70, 59].

### 2.3.4. Drug characterizations

Independent of their sensitivity profiles across cell lines, there are several ways to represent drugs in order to utilize prior knowledge about them or feed those representations into a machine learning model. One category of such characterizations concerns drugs' function and mechanism of action. For example, drugs can be described in terms of their target signaling pathways or, more specifically, gene targets for which they were designed for. Such meta-annotations can often be found in pharmacogenomic databases themselves or in the external data repositories dedicated solely to drugs and compounds annotations [71, 72]. Utilizing binary information about drugs' putative targets can be beneficial, e.g. it can cause the ML model to pay more attention to cell lines' molecular features corresponding to those targets,

however, such information also has its drawbacks. Firstly, from a technical standpoint, if such targets are fed into a model as binary-encoder features, this data would be sparse, since most of the compounds have only one or a few putative targets. Secondly, from a drug function perspective, such description does not capture the whole picture of drug action, since it does not account for drug off-targeting which influences how they interact with the cells [73]. For the latter reason, it might be more beneficial to represent drugs by their continuous inhibition profiles across a panel of targets, i.e. a vector containing a numerical indicators of inhibition levels for a given set of targets. On the other hand, such data is less abundant and available for only a subset of drugs, since it requires a relatively significant experimental effort.

Another, qualitatively different, approach to drug or molecule characterization concerns their chemical structure. One common representation of substances' chemical structure are SMILES strings [26, 27, 28]. SMILES stands for "Simplified Molecular-Input Line-Entry System" and describes the atoms and bonds of a molecule [74]. As raw SMILES are strings, i.e. sequences of characters, they require an additional featurization step which converts them to numerical representations which can be fed into ML algotithm. These molecular featurizers can range from relatively simple ones, e.g. molecular fingerprints which are binary vectors of that indicate the presence or absence of specific features in a molecule [74], to featurizers which are complex models themselves, e.g. neural networks used in natural language processing [75, 76] or based on graph convolutions [77, 74]. One major advantage of using SMILES is that it is a raw data type, i.e. it is available for almost every drug or compound, which enables to gather more data for training. On the other hand, this characterization is on much lower-level compared to e.g. inhibition profiles or drug function, which, in principle, should be the most important for prediction of cellular responses. These functional characteristics have to be derived explicitly or implicitly from SMILES by the model itself. Furthermore, a need for featurizer adds an extra degree of freedom and important choice when building a model for DSP since different featurizers represent different aspects of SMILES string.

Overall, these considerations make choosing a proper drug representation not a straightforward problem. Notably, all of the afore-mentioned types of drug information have been utilized in works presented in the thesis, providing an overview of them in different models and applications.

# Chapter 3

# Machine learning models

## 3.1. Conventional machine learning

Machine learning can be described as a branch of artificial intelligence dedicated to models which can learn from data without being explicitly programmed. ML models adjust their parameters during the training phase, learning from training data, in order to make predictions or decisions on a new, unseen data. In a supervised learning framework [78, 79], observations or examples come in a form of a pair containing an input (vector of features) and a desired output (which can be a single-valued label, or a target variable). The goal of a ML algorithm is then to use these labeled training examples to learn a function mapping from input variables (features) to target variable, which can be used to determine correct labels for new examples. If the target variable is discrete, this is reffered to as a classification problem, whereas continous target variable poses a regression problem.

A typical machine learning workflow involves four major steps: data extraction, data splitting and preprocessing, model selection, and model evaluation. Following data extraction, data is splitted into training, validation, and test (held-out) set. Data preprocessing involves preparation of the data so they can be fed into and ML model and additional steps such as scaling or normalization. Model selection involves training different models on the training data and validating them on the unseen validation set. Instead of performing this process on a single training-validation data split, one common technique is a k-fold cross-validation in which available training data is split into k folds, and in each iteration model is trained on k-1 folds and evaluated on the remaining one, where the end evaluation metrics are averaged over k iterations. Model selection often involves hyperparameter-tuning, where different combinations of the same algorithm's hyperparameters are tested and evaluated. After the best overall model is chosen, it is applied to the held-out set in order to get a final estimate of its performance. The particular steps of the above-described problems may take a different form depending on the data and a task at hand.

One important consideration when comparing the models and evaluating their performance are evaluation metrics. For a regression problems, a common metric is root mean squared error (RMSE) between actual and predicted labels:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}, \tag{3.1}$$

where $m$ is a number of observations, $y_i$ denotes actual value for an $i$th example, and $\hat{y}_i$ denotes a predicted value for an $i$th example. Therefore, lower values of RMSE indicate better

model performance. The mean squared error (MSE) is commonly used as an loss function for regression algorithms during the training phase, in the evaluation phase the square root is taken to have an error on the same scale as the actual labels. Although most common, RMSE may not always be the most intuitive metric. For that reason, regression models are often evaluated using Pearson correlation coefficient which measures how closely predicted and actual labels co-vary:

$$r = \frac{\sum_{i=1}^{m}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{m}(y_i - \bar{y})^2(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{3.2}$$

where $\bar{y}$ denotes the mean of actual labels and $\bar{\hat{y}}$ denotes the mean of predicted labels across $m$ observations. The Pearson $r$ coefficient is in the $[-1, 1]$ range and values closer to 1 indicate bigger positive correlation between two variables, hence better predictive performance.

One important concept in ML is the bias-variance tradeoff [80, 81]. Models with too high bias are too simple to accurately fit the data, and, as such, exhibit high training and test errors. On the other hand, models with too high variance are too complex, resulting in overfitting to the training data. As such, these models will have a low training error, but will fail to generalize to unseen observations, ending up modeling the random noise in the training data and exhibiting high test error. Finding the optimal spot for model's bias and variance is at the core of ML model development. Model regularization aims to reduce variance of the model. Common regularization techniques involve reducing the number of parameters or shrinkage methods, i.e. reducing the values of parameters, however, many ML algorithms have their own specific hyper-parameters controlling model complexity.

## 3.2. Deep learning

Deep learning refers to a subfield of machine learning concerning deep neural networks (DNN). Neural networks are comprised of layers of nodes, or neurons, which are connected to neurons in a neighboring layer with associated weights. In a common, fully-connected feed forward network, the initial input features vector is propagated through network's hidden layers until the output layer, which, in case of supervised learning, outputs the target variable estimate. This forward propagation is essentially a series of linear transformations performed by matrices with learnable parameters, followed by an application of a non-linear, element-wise activation function in each hidden layer. This allows the neural network to represent complex, non-linear functions mapping from input features to the desired output. During the training phase, DNN adjust its parameters in order to minimize a loss function defined in terms of model predicted outputs and the actual outputs. The term "deep" refers to the number of hidden layers, however, there is no clear definition of how many hidden layers make a given neural network "deep".

From a functional perspective, perhaps the most important distinction between conventional ML models and deep neural networks is the ability of deep learning system to automatically extract important features from the raw input data [79]. In this context, deep neural networks essentially perform a hierarchical feature learning, exploiting the unknown structure in the input data in order to come up with subsequently better representations, with higher-level learned features defined in terms of lower-level features [82]. This makes DNNs an essential class of algorithms for representation learning.

The training of DNNs is performed in an iterative fashion using gradient descent-based optimization [79, 83]. At each iteration, a batch of training input data is fed to the model

and the forward pass is performed. Then, a specified loss function is calculated based on the predicted and actual values. Next, the gradients of the loss function w.r.t. model's parameters are calculated using the backpropagation algorithm, in which prediction error is propagated to individual parameters according to the chain rule. After gradients are calculated, the parameters are simultaneously updated in the direction of negative gradient, adjusting the parameters so that the loss function is closer to the minimum. The size of this update step is controlled by a learning rate hyper-parameter. This process is repeated for the subsequent batches. A single passage through entire training data is referred to as an epoch. The training is performed for a specified number of epochs or after a desired stopping criterion is met.

A typical loss function for a straightforward regression problem is MSE, which maximizes the likelihood of the data. However, loss functions can be much more complex and their custom specification is a main way to force a model to achieve a desired task or incorporate some prior knowledge into the modeling problem. As DNN contain a large number of parameters, they are also susceptible to overfitting. Common regularization techniques in deep learning include adding a L2 loss, shrinking the parameters, dropout layers, in which some proportion of nodes in the layer are randomly ignored, or early stopping, in which model training is stopped when the performance on validation data starts to decrease [82].

## 3.3. Generative modeling

One distinction that can be made for ML models is between discriminative and generative models. Informally, generative models have the ability to generate new observations or data instances, e.g. new images, words, or compounds resembling the real ones, while discriminative models can only distinguish between different kinds of observations. More formally, generative models learn a model of the joint probability $p(\mathbf{x}, y)$ of the input features $\mathbf{x}$ and label $y$, or just $p(\mathbf{x})$ in a unsupervised setting with no labels. In contrast, discriminative classifiers either learn conditional probability $p(y|\mathbf{x})$ directly, or learn an explicit mapping from inputs $\mathbf{x}$ to label $y$ [84]. In the context of generative modeling, given data points can be viewed as realisations of the underlying random phenomenon [85]. The goal then is to use this observed data to learn a probability density model which resembles the true data generating distribution as closely as possible. Notably, this requires from such model to properly capture the underlying relationships and patterns in the data, hence generative modeling is often highly intertwined with representation learning.

Consider a dataset $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ comprising of data points ($N$-dimensional vectors of input features, where $N$ is the number of features) $\mathbf{x} \in \mathbb{R}^N$. In a generative modeling framework, samples (observations) in $D$ are seen as realisations of an underlying random variable $X$ in an $N$-dimensional space $\mathcal{D}$. The goal is then to learn the distribution $p^*$ underlying samples in $D$, where density at particular $\mathbf{x}$ denoted as $p^*(X = \mathbf{x})$, or, more concisely, just $p^*(\mathbf{x})$ [85]. In a parametric approach to this problem, consider a family of parameterized density functions $\mathcal{P}_\Theta = \{p_\theta | \theta \in \Theta\}$, where $\Theta$ denotes a set of possible parameters' values $\theta$. Next, define a performance metric given by a function $\mathcal{L}(p^*, p_\theta)$ which measures how well $p_\theta$ captures the true distribution $p^*$. The goal of the generative model is to find such $\theta^* \in \Theta$ which, depending on the form of a performance metric, minimizes or maximizes $\mathcal{L}$, hence optimizing for the appropriate recovery of $p^*$. Choice of $\mathcal{L}$ depends on the type of the model and its desired properties.

Often, generating samples from $p_\theta$ directly in a data space $\mathcal{D}$ may be infeasible, especially in a high-dimensional setting and in a presence of dependencies between features. One common way to reformulate $p_\theta$ is through latent variables. In the latent variable models, the

assumption is that most of the variability in the data can be explained by some number of factors of variation [85], which are captured by latent variables. Latent variables are called latent (or hidden), because they are not directly observed and need to be inferred from observed variables (whose realisations are datapoints $\mathbf{x}$). Latent variables can represent types or classes of objects being modeled, and can also be interpreted informally as a decision that the model makes prior to generating a data sample. For example, in a case of images of hand-written digits, it might be beneficial for a model to first decide which digit to generate, and the produce the corresponding pixel-level values based on this prior decision [86]. Therefore, in the latent generative models, a latent representation $\mathbf{z}$ is chosen first, and an actual sample $\tilde{\mathbf{x}}$ is generated given $\mathbf{z}$. Importantly, in this setting, it is assumed that most of the variability and relations between features of $\mathbf{x}$ is captured by these latent variables, and individual feature variables are independent given $\mathbf{z}$ (conditional independence) [85]. Given a model with latent variable $Z$ in space $\mathcal{Z}$, the corresponding distribution of the observed variable $X$ is obtained through marginalization of the joint distribution of latent and observed variables:

$$p(\mathbf{x}) = \int_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x},\mathbf{z})d\mathbf{z} = \int_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \tag{3.3}$$

In principle, a deterministic function $f_\theta : \mathcal{Z} \mapsto \mathbb{R}^N$ can be defined to map latent representations to data points. In practice, $f_\theta$ is usually used to parametrize the probability distribution of the output random variable. For example, $f_\theta(\mathbf{z})$ might output the mean of the Gaussian distribution with unit isotropic standard deviation, corresponding to conditional probability density of $\mathbf{x}$ given $\mathbf{z}$:

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|f_\theta(\mathbf{z}), \mathbb{I}), \tag{3.4}$$

where $\mathbb{I}$ is the identity matrix. When integrated over $\mathbf{z}$, this allows for representing complex probability distributions in the data space.

One popular example of a latent variable model is a Gaussian mixture model (GMM) or a mixture of Gaussians. GMM distribution can be simply stated as a linear superposition of Gaussian components [80]:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.5}$$

where $\pi_k$ is a weight of a $k$th component, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are mean vector and covariance matrix corresponding to a $k$th Gaussian component, respectively. This model can be formulated in terms of a discrete latent variable. Consider a $K$-dimensional binary random variable $\mathbf{z}$ in which a particular, $k$th element $z_k$ is equal to 1 and rest of elements are 0, with the probability density described in terms of weights $\pi_k$, such that:

$$z_k \in \{0, 1\} \tag{3.6}$$

$$\sum_{k=1}^{K} z_k = 1 \tag{3.7}$$

$$p(z_k = 1) = \pi_k \tag{3.8}$$

$$0 \leq \pi_k \leq 1 \text{ and } \sum_{k=1}^{K} \pi_k = 1. \tag{3.9}$$

In a GMM, the conditional probability distribution of $\mathbf{x}$ given a particular value of $\mathbf{z}$ is a Gaussian distribution parametrized by means and covariance matrix corresponding to a

particular component:

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{3.10}$$

Therefore, the joint distribution $p(\mathbf{x}, \mathbf{z})$ is given by $p(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, and the marginal distribution of $x$ is obtained by summing over all possible values ($K$ components) of $\mathbf{z}$:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.11}$$

arriving at the linear superposition formulation in Eq. (3.5). Using the joint probability and the Bayes rule we can also perform inference and find conditional probability $p(\mathbf{z}|\mathbf{x})$, indicating which components were most likely to produce a particular $\mathbf{x}$.

It turns out that applying the standard maximum likelihood framework in order to learn the parameters of GMM is not straightforward [80]. GMMs are commonly learned using the expectation-maximization (EM) algorithm, but can also be trained using variational inference and gradient based optimization [80].

## 3.4. Variational autoencoders

Often, the complexity of the data requires a complicated, expressive function mapping from the latent variable to data space. Such functions can be computed with deep neural networks. Models which utilize DNNs in a generative process are referred to as deep generative models. One particular category of deep generative models are variational auto-encoders (VAEs).

VAEs can be simplistically introduced as an extension of deterministic autoencoders with incorporated randomness. In that setting, a single datapoint $\mathbf{x}$ is mapped to a probability distribution over latent space, rather than a single point. The corresponding latent representation is then sampled from this mapped distribution and fed into a deterministic decoder to obtain the output. Although sometimes useful and simple, this formulation does not capture the main idea behind VAE theory, as the encoder-decoder architecture is only introduced as the means to optimize a training objective derived from a probabilistic model.

In fact, VAEs can be seen as latent variable models, in which the main idea is to transform some simple distribution $p(\mathbf{z})$ using a non-linear function to obtain a more complex distribution in a data space. As in latent variable models in general, the goal is to maximize the marginal probability over $x$:

$$p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{x}|f_\theta(\mathbf{z}))p(\mathbf{z}), \tag{3.12}$$

where $p(\mathbf{z})$ is referred to as a prior distribution over latent variable $\mathbf{z}$. In standard VAEs, prior distribution is a standard normal (with $\mathbf{0}$ mean and unit isotropic standard deviation). The function $f_\theta$ is computed by a neural network referred to as a decoder and is used to output the parameters of a Gaussian distribution in the data space:

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|f_\theta^\mu(\mathbf{z}), f_\theta^\Sigma(\mathbf{z})). \tag{3.13}$$

In principle, given a computable and differentiable w.r.t. model's parameters formula for $p(\mathbf{x})$, one could use gradient descent to optimize it. However, due to the neural net computing $f_\theta$, the integral in Eq. (3.12) does not have a closed form and is intractable. A natural way to solve this issue would be to approximate $p(\mathbf{x})$ using Monte Carlo estimation; sample large number of $\mathbf{z}$ values from prior $p(\mathbf{z})$ and compute $p(\mathbf{x}) \approx \frac{1}{k}\sum_{i=1}^{k} p(\mathbf{x}|\mathbf{z}_i)$, where

$k$ is a number of samples and $\mathbf{z}_i$ indicates a particular sample. A problem with this is that with a standard normal multi-dimensional Gaussian $p(\mathbf{z})$ and higher-dimensional data space, most of the samples $\mathbf{z}_i$ would produce $p(\mathbf{x}|\mathbf{z}_i)$ which is very close to zero, hence the accurate approximation would require very large number of samples. This problem can be mitigated with weighted sampling, i.e. first identifying the region of latent space which is most likely to produce a given $\mathbf{x}$ and sample $\mathbf{z}$s from there. In VAEs, this posterior distribution of $\mathbf{z}$ given $\mathbf{x}$ is computed by an inference network $q_\phi$, referred to as an encoder. A usual choice for the form which this approximate posterior takes is a Gaussian distribution, this time parameterized with the means and covariance matrices outputted by the encoder:

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|g_\phi^\mu(\mathbf{x}), g_\phi^\Sigma(\mathbf{x})). \tag{3.14}$$

The approximated posterior can then be used to re-weight the objective integral [85]:

$$p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})d\mathbf{z} = \int_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}d\mathbf{z} \tag{3.15}$$

As typical in probabilistic models, in VAEs we are interested in computing log-probability of the evidence ($\ln p(\mathbf{x})$). By combining weighted sampling from Eq. (3.15) with Jensen inequality, we can derive the variational lower bound of the evidence:

$$\ln p_\theta(\mathbf{x}) = \ln\Big(\int_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x})p_\theta(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}d\mathbf{z}\Big) \tag{3.16}$$

$$\geq \int_{\mathbf{z}} q_\phi(\mathbf{z}|\mathbf{x}) \ln\Big(p_\theta(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\Big)d\mathbf{z} \tag{3.17}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{x}|\mathbf{z})||p(\mathbf{z})), \tag{3.18}$$

where $D_{KL}(q_\phi(\mathbf{x}|\mathbf{z})||p(\mathbf{z}))$ is the Kullback-Leibler divergence [87, 80] between learned posterior ($q_\phi(\mathbf{x}|\mathbf{z})$ and prior $p(\mathbf{z})$. This yields the final maximization objective of VAE, referred to as evidence lower bound (ELBO), expressed in terms of encoder's and decoder's parameters:

$$\mathcal{L}_{\phi,\theta}^{ELBO} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{x}|\mathbf{z})||p(\mathbf{z})), \tag{3.19}$$

where, in practice, $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}$ is computed by sampling from the learned posterior $q_\phi(\mathbf{z}|\mathbf{x})$. This form of ELBO explains the auto-encoder interpretation of VAE model; learning $q_\phi(\mathbf{z}|\mathbf{x})$ corresponds to encoding known $\mathbf{x}$ into latent representation $\mathbf{z}$ via the encoder, while learning $p_\theta(\mathbf{x}|\mathbf{z})$ corresponds to reconstructing $\mathbf{x}$ from $\mathbf{z}$ via the decoder. Hence, the first term in Eq. (3.19) can be seen as a reconstruction error, while the second as the regularization term [86]. Given Gaussian distributions as parametric forms of both $q_\phi(\mathbf{x}|\mathbf{z})$ and $p(\mathbf{z})$, the $D_{KL}$ term can be computed analytically.

While ELBO from Eq. (3.19) is tractable and can be computed in practice, it does require sampling from the learned posterior, which makes the ELBO non-differentiable and hence unsuitable for the gradient descent optimization. This obstacle can be bypassed using the so called reparametrization trick, where randomness is injected as an additional input independent from model's parameters [88]. Specifically, for a given datapoint, a random sample $\epsilon$ from standard Gaussian is drawn, and then transformed into posterior $q_\phi(\mathbf{z}|\mathbf{x})$ using means and covariances outputted by the decoder:

$$\tilde{\mathbf{z}} = g_\phi^\mu(\mathbf{x}) + g_\phi^\Sigma(\mathbf{x}) \odot \epsilon, \tag{3.20}$$

where $\tilde{\mathbf{z}}$ is denotes a single sample from a distribution, $\odot$ denotes an element-wise product, and $\epsilon \sim \mathcal{N}(0, 1)$. Since sampling is now formulated in terms of addition and multiplication w.r.t. to model's parameters, gradient can flow through a hidden layer.

## 3.5. Computational drug sensitivity prediction

As stated above, the amount of clinical data available for DSP models in cancer is insufficient, mainly due to the infeasibility of testing multiple compounds on cancer patients. Therefore, most of the databases for DSP concern pre-clinical models, such as cell lines, tumor organoids [89], and patient-derived xenographs in mice [90]. The important advantage of the latter two is that they also mimick the tumor microenvironment, which is not a feature of cell lines. Another potential drawback of cell lines is that they may diverge from the original tumor [91]. However, cancer cell lines are still a vital tool in cancer research in general, and are particularly suitable for development of DSP models due to the relatively high availability of large and systematic databases [64, 18]. The common feature of such existing public resources is that they contain panels of responses of cell lines across multiple approved drugs and other compounds along with molecular features which can be used to describe the cell lines. However, particular databases may differ from each other in terms of the number of screened drugs and cell lines, available omics data, drug annotations, and dose-response curve summary metrics. An overview of existing pharmacogenomic databases and their depiction in terms of factors listed above can be found in [59]. Notably, there are also databases and models concerning the responses for combination therapy, i.e. combination of two or more compounds. However, this thesis is devoted to the prediction of sensitivity to monotherapy, i.e. a single compound.

In principle, the development of ML models for DSP does not significantly differ from the standard ML development [62], in that it consists of essentially the same steps: data extraction, data splitting and preprocessing, model selection, and model evaluation. Nonetheless, when developing such models, there are several concerns and choices to make which are specific to computational DSP. Some of these concerns are closely related to challenges listed in Chapter 1. The differences in existing approaches to DSP can be depicted in terms of several major aspects.

One of such major aspects is the overall modeling framework, namely per-drug or multi-task, or multi-drug approach. In the former, a separate, independent ML model is built for every compound, while in the latter, a single model able to predict responses to multiple drugs is developed. Notably, these two settings significantly differ in terms of the form of the data, data preprocessing and model evaluation. In the per-drug setting, a single datapoint corresponds to a cell line, while in a multi-task framework a single datapoint is a drug-cell line pair, therefore theoretically in the latter approach there is more data available, however, with repeating drugs and cell lines. This has consequences in the choice of ML algrithms; in the per-drug approach common, traditional ML models such as linear regression, K-nearest neighbors, SVMs are used, while multi-task framework utilizes more complex, customizable models, in particular neural networks [13, 59]. Multi-task models have become increasingly popular in recent years, replacing pre-existing per-drug approaches. Still, new variants of models in the latter framework continue being developed and reported as successful [59].

Another important factor in terms of which DSP models are distinguished is the choice of input data. In terms of omics data used to describe the cell lines, gene expression is the most ubiquitous one, especially when a model uses single cell lines' modality, or data type [62, 70, 59]. Adding additional data types such as point mutations or CNVs can improve predictive performance for some target-specific drugs. On the other hand, it adds more initial features and therefore make feature selection harder. The second essential part of the data aspect is whether or not a given approach utilizes drug information, and if so, what is the form of this information. Again, incorporation of drug features into DSP modeling has become more popular in recent years [59].

Besides two major differences described above, there are several, less significant factors differentiating between DSP models. One of them is the form of the target variable, i.e. the drug sensitivity metric. In the pharmacogenomic databases, it comes in a continuous form, naturally posing a regression problem, although it can be discretisized (e.g. as sensitive vs. non-sensitive) [92], turning the task into a classification problem. While this approach has some advantages, it also loses some information about the sensitivity profiles and is volatile w.r.t. discretization procedure. Therefore, all of the works depicted in this thesis concern DSP in a form of regression problem. Another factor is a model interpretability approach. This is largely influenced by the model itself; some standard ML algorithms, such as linear regression and K-nearest neighbors have some properties which make them more interpretable, while others, e.g. SVMs, are not as straightforward to analyze. While neural networks have the opinion of not being easily interpretable, because of their flexibility they also allow for customizable, tailored approaches to interpretability.

From a strictly technical perspective of ML algorithms used, DSP has been approached by a whole variety of methods coming from many different ML subfields, including common, well-established ML algorithms, ensemble methods, network-based methods, deep neural networks, recommender systems and more. As mentioned above, the choice of the type of algorithm depends on which challenges associated with DSP one chooses to address. Because of the abundance of existing methods, it is infeasible to list them all in this section; a more comprehensive list and classification of existing methods can be found in [59, 93], while background sections of subsequent chapters contain context-specific literature overview relevant for the given project.

# Chapter 4

# Per-drug feature selection strategies for drug sensitivity prediction

## 4.1. Background

The ability to predict a response of a specific cancer type to a therapy is one of the main goals of precision medicine. Considering molecular features of cancer cells is crucial for mitigating heterogeneity and for tailoring the therapy to specific patients [94]. The emergence of large scale high-throughput screening studies [95, 65, 96, 97, 98] have allowed researchers to develop computational models for drug sensitivity prediction from molecular profiles of human cancer cell lines or drug properties [62, 99]. Although the inconsistencies and limitations of cell line data have been raised and extensively studied [14, 15, 16, 17], these resources remain a vital tool for the development of such models.

Arguably, the desired quality of computational models of drug sensitivity is not only their predictive performance, but also interpretability. To evaluate candidate drug efficacy on a specific patient's tumor, many approaches apply black-box algorithms with a set of highly dimensional features as input. In clinical practice, the capability of extracting such high-volume data from patient's material is limited. Thus, there is a growing need of proper identification of concise, limited subset of features, or biomarkers, that are most informative of drug sensitivity. Therefore, strong emphasis should be put on feature selection approaches for drug sensitivity prediction. Despite its paramount importance, no systematic assessment of feature selection strategies in the task of drug response prediction was so far performed.

The problem of drug response prediction has been approached by a wide spectrum of linear and non-linear machine learning algorithms, including regularized linear regression, k-nearest neighbors (KNN), support vector machines and random forests [70, 92, 100, 101, 102, 103]. Multitask learning was proposed to improve drug sensitivity prediction by pooling information learned for different drugs [104, 100]. Finally, a number of kernel-based multi-view and multi-task models were introduced for drug sensitivity [105, 106, 107]. Although these approaches show very good predictive performance, they suffer from low interpretability. As a remedy, a multi-task learning approach based on a Bayesian model for collaborative filtering was proposed [108], which allows for identifying general interactions between features of the drugs with features of the cell lines. For example, it gives insights in the form of "activation of pathway Y will confer sensitivity to any drug targeting protein X". This approach, however, does not directly address the crucial need of identifying biomarkers for specific drugs.

For that aim, data-driven, automatic techniques of feature selection were applied [107, 109, 102]. Generally, the problem of identifying the optimal subset of features is intractable [110].

Data-driven feature selection thus proceeds either as a heuristic search over the space of feature combinations, or is embedded directly in the learning algorithm by imposing sparsity of parameters associated with the features via regularization. Although these methods can achieve good predictive performance and deal with the curse of high data dimensionality [102], feature importance estimates and selection might not always be accurate and stable, especially in vastly high-dimensional data and in the presence of correlation between features [111]. Stability selection was proposed to mitigate this problem when regularized regression is applied [112], but it still comes without the guarantee to choose the most biologically relevant predictive features.

Drug prediction approaches largely differ with respect to the type of features that they model. Among the molecular data feature types which characterize the cancer cell lines, gene expression was assessed as the most informative, with remaining types such as mutation or copy number data bringing limited predictive power [92, 70]. Accordingly, genome-wide gene expression is the most common choice in the case of models utilizing single data type [62, 113, 102, 107, 108]. Other studies reported that in some cases gene expression alone might not be sufficient, especially in a cancer- or drug-specific setting [114, 115]. Importantly, expanding the feature space related only to cancer cell lines' biology with drug-related properties was shown to improve predictive performance [100, 108, 115, 106, 107]. The predictive drug-specific features may be related to their chemical properties, such as compound structure [100, 115], their known primary targets or pathway activation [106, 107]. Recently, multiple methods based on deep learning have emerged, showing promising results in the application to drug sensitivity prediction [116]. The published neural network architectures range from common stacks of fully connected layers [117] to more sophisticated architectures involving residual and convolutional networks [118, 119, 120]. Furthermore, methods employing autoencoders [121, 122] and variational autoencoders [123] have been proposed. Due to their complicated, non-linear structure, neural networks may suffer from the lack of interpretablity, including difficulties in assessment of feature importance. However, methods from the growing field of explainable artificial intelligence can help to mitigate this problem [124].

Here, we utilize the knowledge regarding drug targets and their mode of action to select plausible features describing the cancer cell lines. This drug-related prior knowledge is thus used to directly limit the initial feature space, rather than first expanding it and next using data-driven selection techniques to narrow it down. We argue that this approach for feature selection in combination with common regression techniques can provide a simple and highly interpretable model without losing the predictive performance characteristic for models starting from high-dimensional data. In fact, the direct utilization of prior knowledge is the number one strategy recommended for feature selection according to the classics in machine learning [110]. It was however, never exploited in the task of drug response prediction. We assess this methodology in a systematic fashion for a broad spectrum of anti-cancer compounds, integrating multiple data types and comparing the results to the baseline models utilizing genome-wide gene expression data and data-driven feature selection techniques. On top of that, we evaluate gene expression signatures as the means of dimensionality reduction of the transcriptomics data and evaluate their predictive power in this context. This comprehensive analysis pin-points a set of drugs for which easily interpretable, informative, small sets of features can be identified.

## 4.2. Methods

### Analyzed dataset

The analyzed dataset was acquired from the Genomics of Drug Sensitivity in Cancer (GDSC) [65] database. A total of 251 compounds were included in the analysis. Each was assigned one of 24 classes of target pathways, defined by the GDSC.

The total set of samples consisted of 983 cancer cell lines originated from 13 tissue sites. The available data types for describing the cell lines included: gene expression (17737 features), coding variants (310 features), copy number variants (CNV, 425 features) and tissue type (13 features). Coding variants and copy number variants were represented as binary calls determining the presence or absence of a variant in a given gene or segment, respectively. We have dummy encoded the tissue types resulting in 13 distinct binary features for every cell line. All biological input data were acquired directly from the GDSC resource.

GDSC provides two types of metrics representing the drug efficacy: half maximal inhibitory concentration ($IC_{50}$) and area under the dose-response curve (AUC). Since in our analysis we did not observe significant differences in predictive performance when using one metric in favor of the other, we picked AUC as our single target variable.

### Predictive algorithms

We employed two common machine learning algorithms in order to predict the AUC values: elastic net linear regression and random forest regression. We implemented both methods using Python3 scikit-learn 0.19.2 library [125]. See Supplementary Methods for descriptions of the algorithms and implementation details.

### Feature selection

With a total of 18485 biological features that can be used to describe the cancer cell lines, the analyzed dataset is very high-dimensional. In contrast, the number of samples is in the order of hundreds, which poses the danger of overfitting. This might especially be the case when considering all available genome-wide information regardless of the drug being modeled. Here, we investigate different feature selection methods to mitigate this problem. These approaches can be divided into two groups: biologically driven and automatic, data-driven selection methods.

#### Biologically driven feature selection

**Features based only on drug targets and tissue type, shortly only targets (OT).** In the most restricted feature space, we included only predictors corresponding to the direct targets of the drugs, as well as tissue type. Drug targets information was derived directly from GDSC. As an additional resource, we used DrugBank [71] database, assigning targets for 88 matched compounds. For each drug target, we included features representing the target gene's expression, coding variant and copy number variation. In the case of copy number data, a given genetic feature was incorporated if the corresponding segment included at least one of the drug target genes. We only considered drugs with explicit gene targets annotation in GDSC or DrugBank and for which at least one feature in addition to the tissue type was available in the data. These conditions were met for 184 compounds. Applying two regression algorithms for each drug resulted in 368 separate models.

**Set of features based on drug targets, tissue type, and target pathways, shortly pathway genes (PG).** In this approach, we included features related to genes that belonged to the same signaling pathway as the set of target genes. Pathways information was derived from Reactome [126, 127] database (version 66 accessed on October 2018). For each compound, first its target set was derived, followed by finding all pathways which included at least one of the given targets. The total set of considered genes was then computed as the union of all members of the found pathways. Lastly, corresponding gene expressions, coding variants, copy number variants and tissue types were extracted to create the final feature set. The drug targets and pathway information was available for 186 drugs, producing 372 models.

**Sets of features resulting from addition of gene expression signatures, shortly OT + S or PG + S.** Gene expression signatures can explain the activation level of complex biological phenomena in the investigated cell lines. Here, we refer to a gene signature as a set of genes related to a certain known biological phenomenon that can be deduced from cancer gene expression data. For each signature $S$ with $i$ genes, we calculated two scores. The first characterizes the coherent expression and the second estimates the activation level of $S$. Given a gene expression matrix for $S$ in $n$ samples ($X^{i \times n}$), the previously described coherence score (CS) [128], is calculated as the mean pairwise Pearson correlation between all columns of $X$. Therefore, a strong negative or positive correlation between all genes in $S$ is indicated by CS values close to $-1$ and 1, respectively. The activity of $S$ (i.e. the signature score) for each sample is calculated by first $z$-scoring the gene expression values across samples, followed by averaging the resulting $z$-scores across genes. Here, we calculated the signature scores using the cancer cell line expression data provided by GDSC. We set the threshold for a significantly coherent activation of $S$ to $\mathrm{CS}(S) \geq 0.1$, resulting in 128 signature features. The OT + S set contains features based on target genes, signature scores and tissue type. The PG + S set contains target genes, pathway genes, signature scores and tissue type. Applying two regression algorithms for each drug resulted in 740 separate models.

**Set of features based on genome-wide gene expression, shortly genome-wide (GW).** Finally, we constructed a feature set based exclusively on the expression of 17737 genes as features. We evaluated this feature set for 251 drugs in total, resulting in 502 different models.

**Data-driven feature selection, shortly GW SEL.**

In addition to feature pre-selection based on drug properties and biological relevance, we also evaluated automated feature selection algorithms in application to genome-wide expression data. We used two techniques, based on linear and non-linear methods. First, stability selection, which uses lasso regression on multiple bootstrap samples in order to choose robust features [112]. Such selected features were next passed as input for elastic net models (further referred to as *GW SEL EN*). For the second technique, feature importance estimates derived directly from random forest were used. These features were then used for random forest regression models (*GW SEL RF*). For more detailed description of both techniques, see Supplementary Methods.

## Model evaluation

During cross-validation tuning, we used *Mean Squared Error* (MSE) as a scoring metric for best hyperparameters search. Although MSE is suitable for evaluation of different models within one compound, it is not reliable when comparing results across diverse drugs because

of differences in corresponding AUC distributions. Furthermore, when a given target variable distribution has little variation, one can achieve a reasonably low MSE just by predicting the mean of a target variable. In order to avoid this problem and identify the models which performed well, we used *Relative Root Mean Squared Error* (RelRMSE), which is normalized in such a way that the score of 1 corresponds to a dummy model which always predicts the mean of target variable in the training data. RelRMSE is defined as the fraction of the dummy model's RMSE on the test data and the analyzed model's RMSE on the test data:

$$\text{RelRMSE} = \frac{\text{RMSE}_{dummy}}{\text{RMSE}_{model}}, \tag{4.1}$$

i.e. better performance corresponds to bigger RelRMSE metric, with a baseline score of 1.

The use of RelRMSE allowed us to distinguish drugs for which predictive algorithms could not outperform the dummy model, meaning that for those compounds no actual learning occurred.

In order to make further assessments and comparisons between compounds, we used Pearson correlation coefficient with the response AUC in the test set as a performance metric. As stated in a previous section, the recorded results for each method were averaged over five modeling procedures that were performed with different data splits.

## 4.3. Results

### Modeling workflow

In order to comprehensively evaluate different feature selection strategies, we devised the following workflow (Fig. 4.1). We first extracted the sensitivity data for each particular drug and corresponding screened cell lines along with their biological features: gene expression, coding variants, copy number variation (CNV) and tissue type (see Methods for the details of the analyzed dataset). We then employed each of the feature selection approaches, which can be divided into two categories: biologically driven and automatic, data-driven selection methods. We considered different biologically driven feature selection strategies, depending on the type of prior knowledge used to define them. In the first approach, we narrowed the initial feature set by including only the features corresponding to drug's direct gene targets (shortly only targets, OT feature set). In the second, we considered the union of the direct target genes and the drug's target pathway genes (pathway genes, PG feature set). Finally, we additionally extended the only targets features and the pathway genes features with gene expression signatures, resulting in two more feature sets (OT + S and PG + S). For a baseline model we considered all available, 17737 gene expression features, referred to as the genome-wide model (GW). For data-driven feature selection we applied two techniques to the baseline gene expression feature set: stability selection (GW SEL EN) and random forest feature importance estimation (GW SEL RF). See Methods for more detailed description of the feature selection approaches. After the feature selection step, we fed the resulting data into elastic net (EN) or random forest (RF) algorithms and evaluated the predictive performance on the test set (Fig. 4.1). This modeling process was performed independently for each drug.

### Models with genome-wide features have larger feature sets and more samples than the models with biologically-driven features.

The median numbers of input features are 3 and 387 for only targets and pathway genes feature sets, respectively (Fig. 4.2a). The input features are further expanded by including

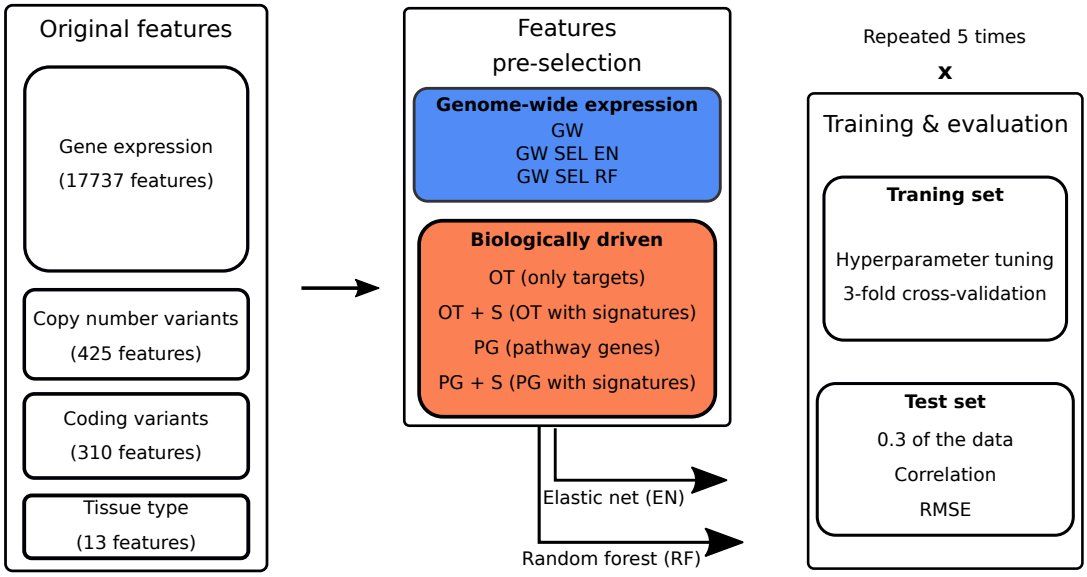


Figure 4.1: **Flowchart describing the modeling framework for a single compound.** Abbreviations: GW – genome-wide, PG – pathway genes, OT – only targets, EN – elastic net, RF – random forest, SEL – automated feature selection, S – gene expression signatures. For every feature space, we performed modeling separately for each drug. We randomly split the corresponding data into training and test set, with 0.3 of the data included in the test set. We used 3-fold cross-validation on the training data for hyperparameter tuning and evaluated the best model on the test set. The whole modeling process was repeated five times with different training/test set data splits.

128 gene expression signatures. In the case of methods based on automated feature selection, the optimal number of features, $k$, is shown. The median $k$ values are 70 and 1155 for random forests and for stability selection, respectively. All foregoing values constitute a drastic decrease in comparison to the number of 17737 genome-wide input features.

The number of samples for each drug also slightly differs for only targets and pathway genes feature sets, since for some cell lines the coding variants or CNV information are not available (Fig. 4.2b). This results in a lower number of samples for models with biologically driven features, with the median of 849 for only targets and 818 for pathway genes feature sets, compared to 876 for genome-wide expression features.

## Drug response distributions are different across compounds, tend to have low variance for drugs targeting specific genes and pathways and high variance for drugs targeting general cellular mechanisms.

The area under the dose-response curve (AUC; Methods) measures the overall drug efficacy, with lower values corresponding to stronger efficacy. The distribution of this metric varies significantly among compounds with different target pathways (Fig. 4.2c). The median AUC value per target pathway ranges from 0.98 for hormone-related drugs to 0.73 for compounds targeting metabolism pathways. The smallest variation of AUC is observed for drugs targeting the hormone-related pathways. The largest AUC variation is observed for the apoptosis regulation pathway. The AUC for drugs targeting general mechanisms, such as DNA replication or metabolism, tends to have larger variance, which means their sensitivity is easier to model.

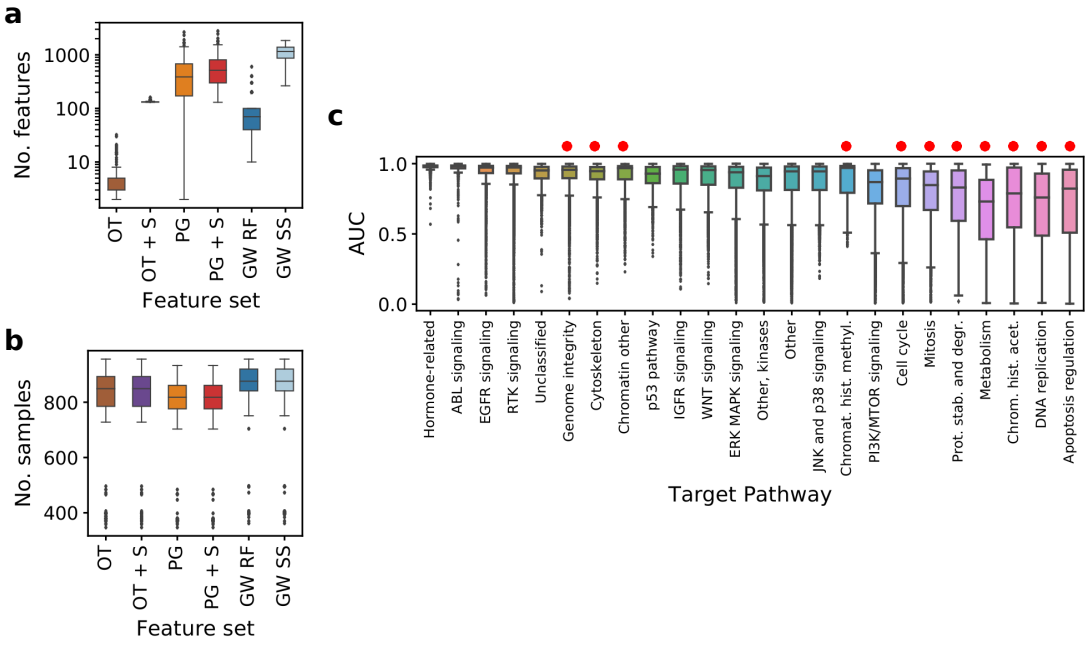


Figure 4.2: **Models' properties and response variable grouped by target pathways.** (**a**) Number of input features across compounds in different methods. For genome-wide models, number of features was 17737 for each drug. Vertical axis uses log scale. (**b**) Number of samples across compounds in different methods. Abbreviation SS refers to stability selection (Methods). (**c**) AUC values grouped by target pathway of the drug, raw data from GDSC. Target pathways are sorted by interquartile range of the AUC values. Pathways corresponding to more general cell mechanisms are marked with red dots. See Fig. 4.1 for abbreviations.

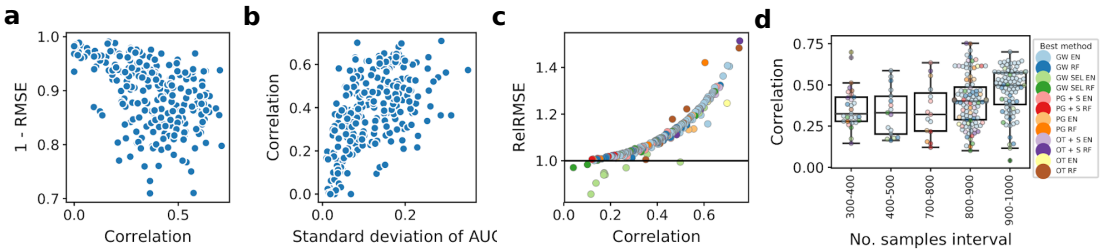


Figure 4.3: **Predictive performance for all of the analyzed drugs.** (**a**) 1 - RMSE versus correlation per drug, obtained by elastic net using genome-wide gene expression data as predictors. For 1-RMSE, higher values correspond to better performance. (**b**) Correlation versus standard deviation of true AUC for all cell lines screened for a given drug, correlation obtained by genome-wide elastic net. (**c**) RelRMSE versus correlation obtained by the best model for a given drug. Higher values of RelRMSE correspond to better performance and improvement over a dummy model, which predicts average AUC. Each point represents a single drug. For each of them, corresponding best performance was determined using correlation as a metric. Colors represent models with feature set that obtained the best performance for a given drug. Horizontal line at 1 represents the baseline RelRMSE score. Most of these correlations are statistically significant (test based on Student's t-distribution at 0.05 significance level, Fig. A1). (**d**) Distribution of per-drug predictive performance grouped by per-drug number of available samples. Colors represent models with feature set that obtained the best performance for a given drug. See Fig. 4.1 for model abbreviations.

### The per-drug results show the importance of comparing to a dummy model and that different feature selection strategies are best suited for different drugs.

Since root mean squared error (RMSE) measures the level of model error, and correlation measures the model agreement with the test set, both large (1-RMSE) and high correlation should coherently indicate a high model performance. The negative relation between (1 - RMSE) quantity and correlation, however, confirms the fact that raw RMSE is not a good metric for performance comparison between compounds (Fig. 4.3a; Methods). Instead, correlation achieved by the model increases with the modeled AUC variance (Fig. 4.3b).

Both these facts support that relative root mean squared error (RelRMSE; ratio of the RMSE obtained by a dummy model to the RMSE obtained by the analyzed model; see Methods) is a better performance measure than raw RMSE (Fig. 4.3c). Indeed, RelRMSE grows with the correlation. Importantly, for some drugs, the best performing models fail to achieve the baseline RelRMSE score of 1 or are very close to 1 (Fig. 4.3c). Further inspection of these models reveals that they can capture only the mean AUC, since the modeled AUC distribution does not have enough variation. In total, there were 19 of such compounds and these were excluded from further analysis.

It is apparent from Fig. 4.3c, that for most of the drugs, the best suited method is modeling using genome-wide features and elastic net. However, this is not the case for compounds with the top corresponding modeling performances, as the two best correlation scores are achieved by models with biologically driven feature space. These two compounds are Dabrafenib and Linifanib, both with correlation of 0.75, for models with feature spaces: only targets genes with gene expression signatures and only targets genes, respectively. In terms of performance, they are followed by Trametinib (correlation 0.71) and Alectinib (correlation 0.70), both scores being achieved by genome-wide methods. In general, as we consider more top performances, the frequency of genome-wide methods among them increases, although they are not as highly represented when looking at the small group of absolute best scores.

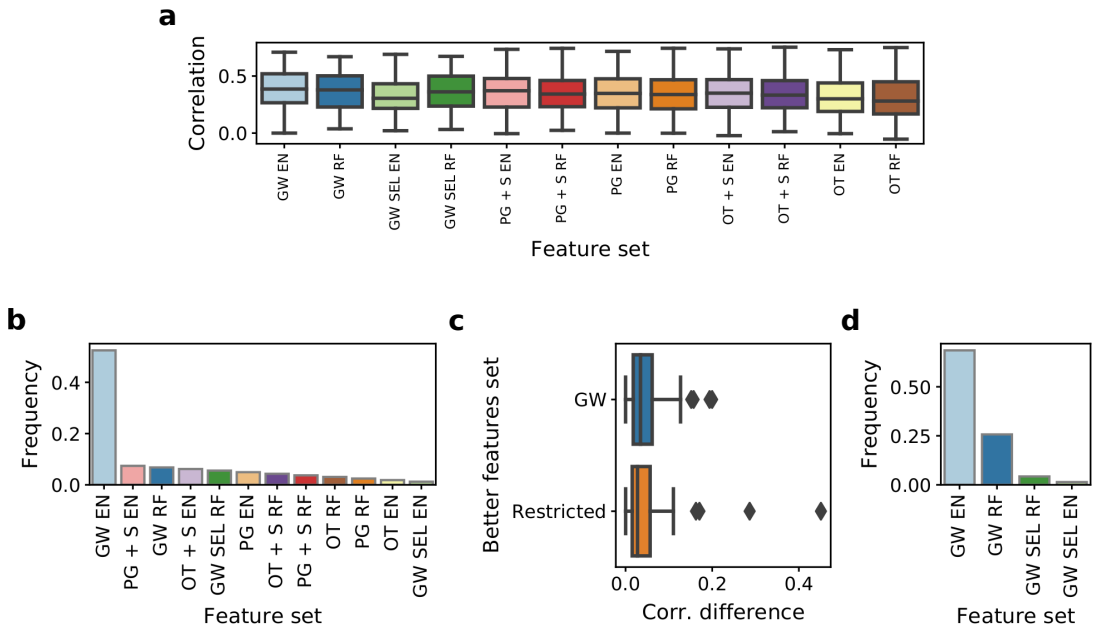


Figure 4.4: **Frequencies of all applied methods among best models per drug.** (**a**) Correlation of AUC predictions with the true AUC values in the test set across compounds in methods with different feature spaces. Results are shown for 175 drugs which were common across all applied models. (**b**) Model frequencies for compounds for which all methods were applied. (**c**) Differences in correlation between best model per drug overall and best model from the other class. Two cases are shown – genome-wide and biologically driven feature sets. (**d**) Model frequencies among best models for compounds where models with biologically driven could not have been applied. See Fig. 4.1 for abbreviations.

The considered set of drugs is diversified in terms of available data (Fig. 4.3d). The bigger number of samples leads to better predictive performance, as more training data mitigates the overfitting effect, especially in high-dimensional setting. However, there is a significant spread in performance among drugs with similar number of samples, implicating that available data is not a single factor explaining the differences in performance.

### The difference in predictive performance of biologically driven versus genome-wide models is small, despite using significantly less input features.

In general, genome-wide feature set combined with elastic net (GW EN) emerges as the best model with the median correlation of 0.39 (Fig. 4.4a). However, models with biologically driven feature spaces perform very similarly, (excluding only targets (OT) approaches), with the best median correlation of 0.37 produced by models employing target pathway genes features combined with gene expression signatures and elastic net (PG + S EN). Furthermore, the difference in median performance was negligible between genome-wide random forest (GW RF, with 17737 features) and genome-wide random forest with automated selection (GW SEL RF, with 70 features on average). This suggests that for many compounds, most gene expression features do not have significant power in predicting drug response. The spread in performance (defined as the difference between the maximum and the minimum value) reaches over 0.6 for all of the methods, suggesting that each drug should be approached individually in terms of modeling.

The standard, genome-wide model achieves the best performance for over half of considered drugs (Fig. 4.4b). However, for many of these cases the correlation difference between the best genome-wide model and the best model with biologically driven features is not significantly large, with the median of only 0.034 (Fig. 4.4c). The reverse is also true, with median correlation difference between the best biologically driven model and the worse genome-wide model 0.028. Despite a drastic reduction in feature space, the biologically driven models based either on only targets or pathways yield the best modeling performance for 23 drugs, outperforming all other models including the genome-wide approach. For further 60 drugs, the best models have feature space expanded with expression signatures. Noticeably, there are also 15 cases where data-driven feature selection helps to produce better performance with much smaller subset of the original feature set (Fig 4.4b, d).

### Predictive performance using different feature selection strategies depends on drugs' target pathways.

Next, we investigate the general tendencies concerning which feature selection is particularly better suited for modeling drugs targeting specific pathways. To this end, we compare the overall performance of biologically driven feature selection as one group to the baseline of genome-wide features and the genome-wide features with automatic selection as another (Fig. 4.5a), for different target pathways. Genome-wide models achieve better performance in 15 out of 24 pathways in total, however, the difference is statistically significant in only four of them (at 0.05 significance level): DNA replication, metabolism, apoptosis regulation pathways and a group of pathways referred to as "other". This indicates that these models capture a broad mechanism of action of the corresponding drugs. Conversely, the target pathways for which the models with biologically driven features most notably outperform models with genome-wide features include ABL, IGFR and EGFR signaling pathways, although these results are not statistically significant due to small sample sizes. The models with biologically-driven features perform also better for the hormone-related pathway, but overall the modeling performance is bad in this case and we do not consider this result reliable. In summary, compounds with specific signaling target pathways seem to benefit more from the initially restricted feature space. Notably, the median number of available sample sizes for drugs targeting specific pathways is similar between the pathways (Fig. A2a) and does not affect the modeling performance (Fig. A2b). Although the number of drugs per target pathway does differ between the pathways, these differences should not affect the comparison outcome as the comparisons of model performance are made within a given pathway.

We next inspect in detail the results for distinct drugs coming from DNA replication and RTK signaling pathways, respectively (Fig. 4.5b and c). Among the drugs targeting the DNA replication pathway, Bleomycin, Methotrexate and SN-38 exhibit good modeling ability with the genome-wide features. However, in case of Methotrexate similar performance is achieved also by methods with biologically driven feature space, contrary to SN-38. Conversely to DNA replication pathway, among the drugs targeting the RTK signaling pathway the best result is more often produced by biologically driven features, with most noticeable cases of Linifanib and Quizartinib. In contrast, Alectinib exhibits good modeling performance exclusively with genome-wide approaches. In general, although the above described general tendencies apply, information about drug's target pathway alone seems to be insufficient to clearly tell which feature space is the most suitable for predicting its response, with the potential exception of the DNA replication pathway.

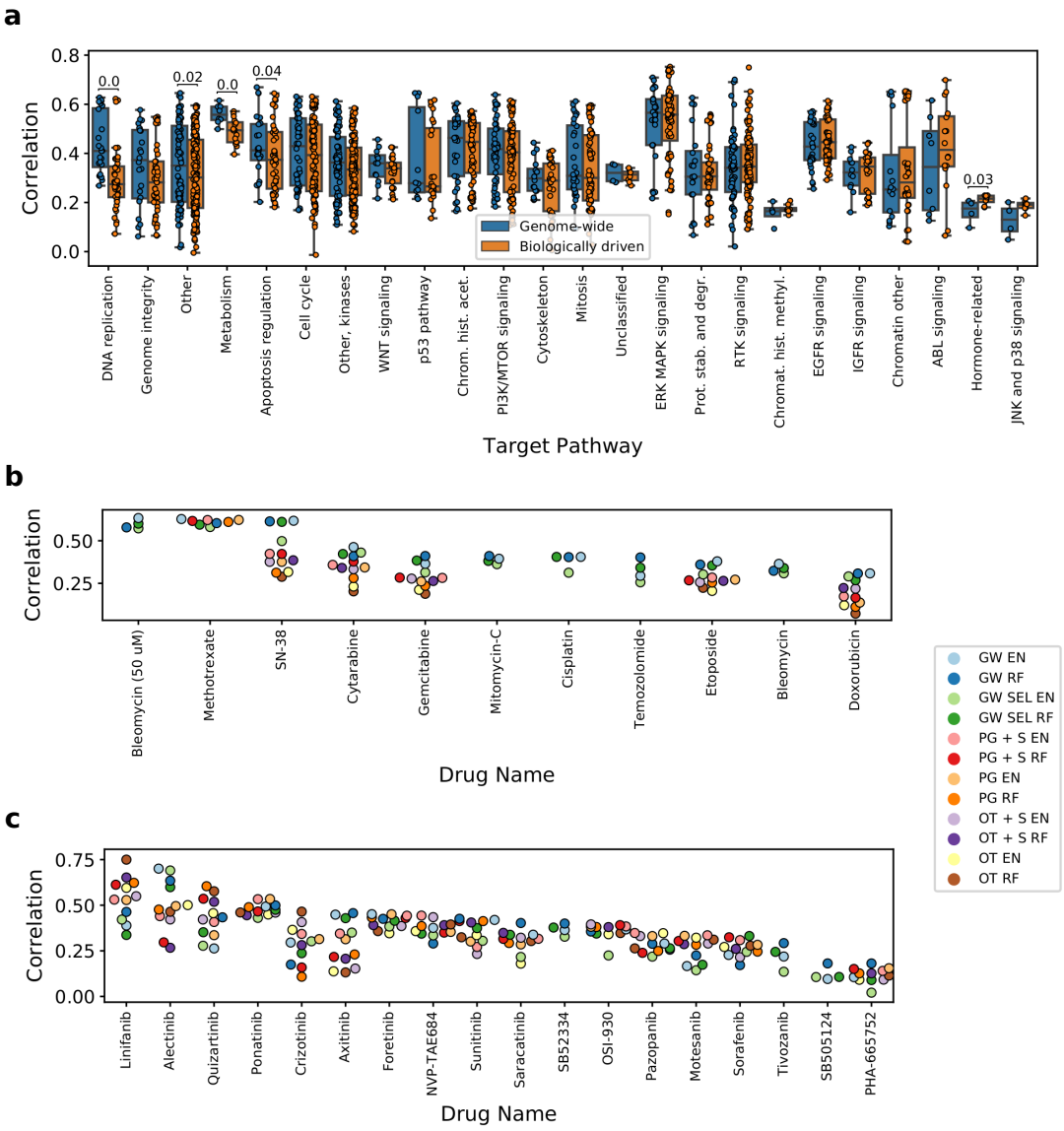


Figure 4.5: **Predictive performance in relation to compounds' target pathway.** (**a**) Correlation with the test set grouped by pathways. Methods were classified into two groups – one that uses genome-wide feature space, and one with biologically driven feature space. Numbers displayed represent p-values for the one-sided Mann-Whitney-Wilcoxon test. Lack of number means no statistical significance at 0.05 significance level. (**b**) Predictive performance for drugs with DNA replication target pathway. (**c**) Predictive performance for drugs with RTK signaling pathway. See Fig. 4.1 for model abbreviations.

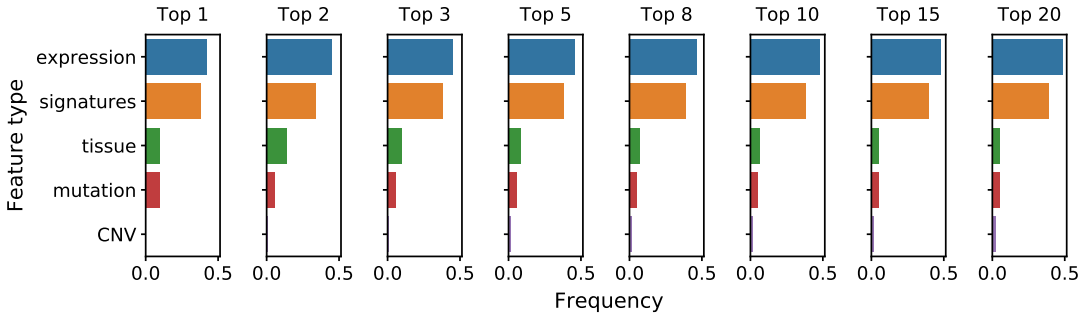


Figure 4.6: **Frequencies of considered feature types among top $k$ most predictive features.** Feature importance coefficients were extracted from top 50 drugs in terms of modeling performance using methods with biologically driven feature space.

### Gene expression and mutations constitute the most predictive feature types.

In order to assess, which feature types are most informative of drug response, we consider such models with biologically driven feature space, which use all five available data types (Fig. 4.6). To make results more robust, we consider only top 50 drugs in terms of corresponding modeling performance achieved by the biologically driven feature sets, resulting in worst considered model's correlation of 0.47. Next, we extract top $k$ most predictive features in each model and record the frequencies of particular data classes among them. Results confirm the fact that gene expression is the most predictive feature type, although mutation (coding variant) and tissue type are also important, especially for drugs designed to target specific cancer type with a particular mutation. In contrast, copy number variants seem not to incorporate much useful information. The relative effect of gene expression data increases with number of considered most predictive features, but this is expected given that this category is the most frequent of all available data types overall. Finally, the high frequency of gene expression signatures among the top predictive features implies that the signatures can act as good representatives of genome-wide information.

### Feature selection enables interpretation of the mode of action and pinpointing biomarkers for the best modeled drugs.

We further focus the analysis on ten drugs of most interest (Fig. 4.7), based on two simple criteria: top modeling performance achieved by all of the feature selection methods, or distinctly better performance achieved by one of the methods' class (genome-wide or biologically driven) in comparison to another. In five of those compounds the best result is produced by models with the genome-wide features, whereas another five are better modeled with biologically driven features.

From all analyzed drugs, Dabrafenib emerges as the compound which is the easiest to model. The highest correlation of 0.75 is achieved by the model combining only targets features with gene expression signatures and random forest (OT + S RF), and performance of other approaches is only slightly worse (Fig. 4.7). This good modeling ability with the OT + S RF features could be explained by two factors. First, the AUC distribution corresponding to Dabrafenib is well-diversified, with relatively many cell lines sensitive to treatment (Fig. 4.8a), which leads to better modeling performance (compare Fig. 4.3b). Second, the relative effects of the selected features are in excellent concordance with the Dabrafenib's pharmaceutical properties. The most predictive feature – mutation in BRAF oncogene (Fig. 4.8a) – and

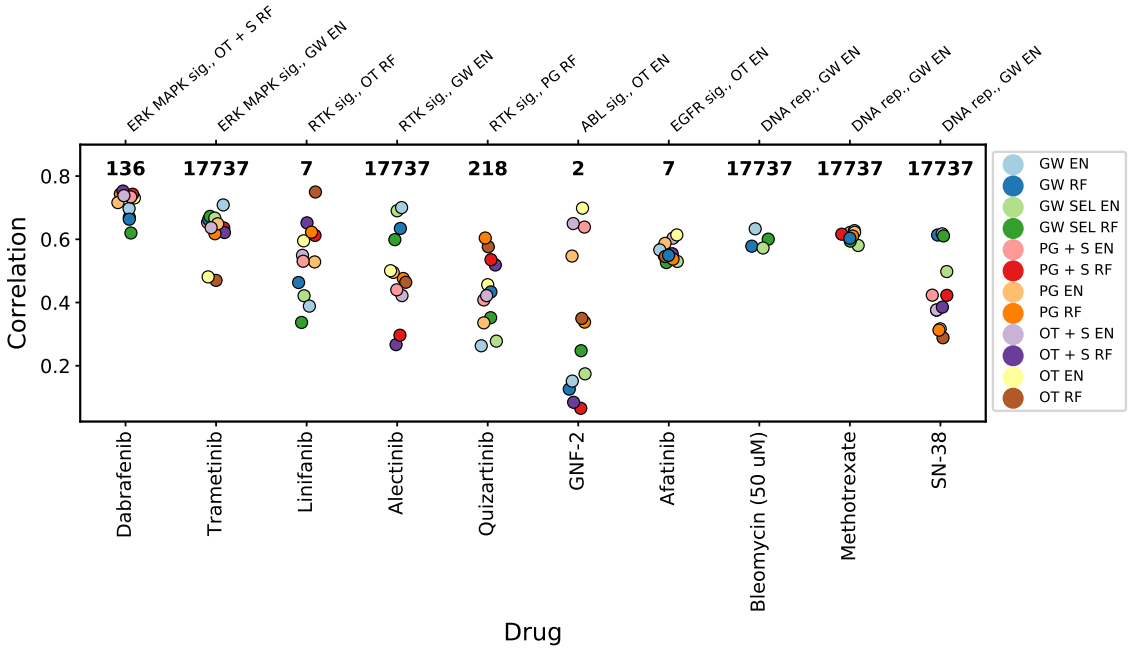


Figure 4.7: **Results for specific compounds exhibiting good ability to model with one or all of the methods.** Displayed numbers represent number of features which was used by best performing model for a particular drug. Top horizontal axis show compounds' target pathways along with model which achieved the best modeling result. See Fig. 4.1 for model abbreviations.

the second most predictive feature – the BRAF gene expression signature – well agree with the design of Dabrafenib as the BRAF inhibitor. Interestingly, the feature corresponding to BRAF gene expression alone ranks lower, 28 among 136 features for the best OT + S RF model and as low as 15817 among 17737 features for the GW EN model in terms of predictive power. Finally, in concordance with Dabrafenib's intended use in treatment of BRAF mutation-positive melanomas and lung cancers [129, 130], the skin tissue feature is the third most predictive one for the best OT + S RF model.

In the case of Linifanib, the best result (0.75 correlation) is accomplished by using only 7 features related to the drug's targets (only targets and random forest, OT RF model), which significantly outperforms the genome-wide models (Fig. 4.7). Linifanib is an inhibitor of FMS-like tyrosine kinase 3 (FLT3) and vascular endothelial growth factor receptor (VEGF) tyrosine kinases involved in clinical trials concerning non-small cell lung cancer (NSCLC), breast, liver, and colorectal cancer as well as leukemia [131, 132, 133]. Contrary to the Dabrafenib's example, Linifanib is one of the rare examples where good modeling results are achievable despite low standard deviation of the AUC distribution (Fig. 4.8b). The high correlation achieved by the OT RF model mainly comes from its ability to accurately predict lowered AUC for three outlying, sensitive cell lines. The most decisive predictive feature in this model is the expression of FLT3 gene, which exhibits high over expression in these cell lines, with much higher mean 11.53 expression than the mean 3.30 for all cell lines in the training set. The expression of FLT3 ranks lower (11th) among features of the genome-wide model.

Similarly to Linifanib, Quizartinib is also characterized by low variation in the treatment response (Fig. 4.8c), and is also an FLT3 inhibitor. Quizartinib is tested in clinical trials for acute myeloid leukemia (AML) [134]. The best biologically-driven model (pathway genes and

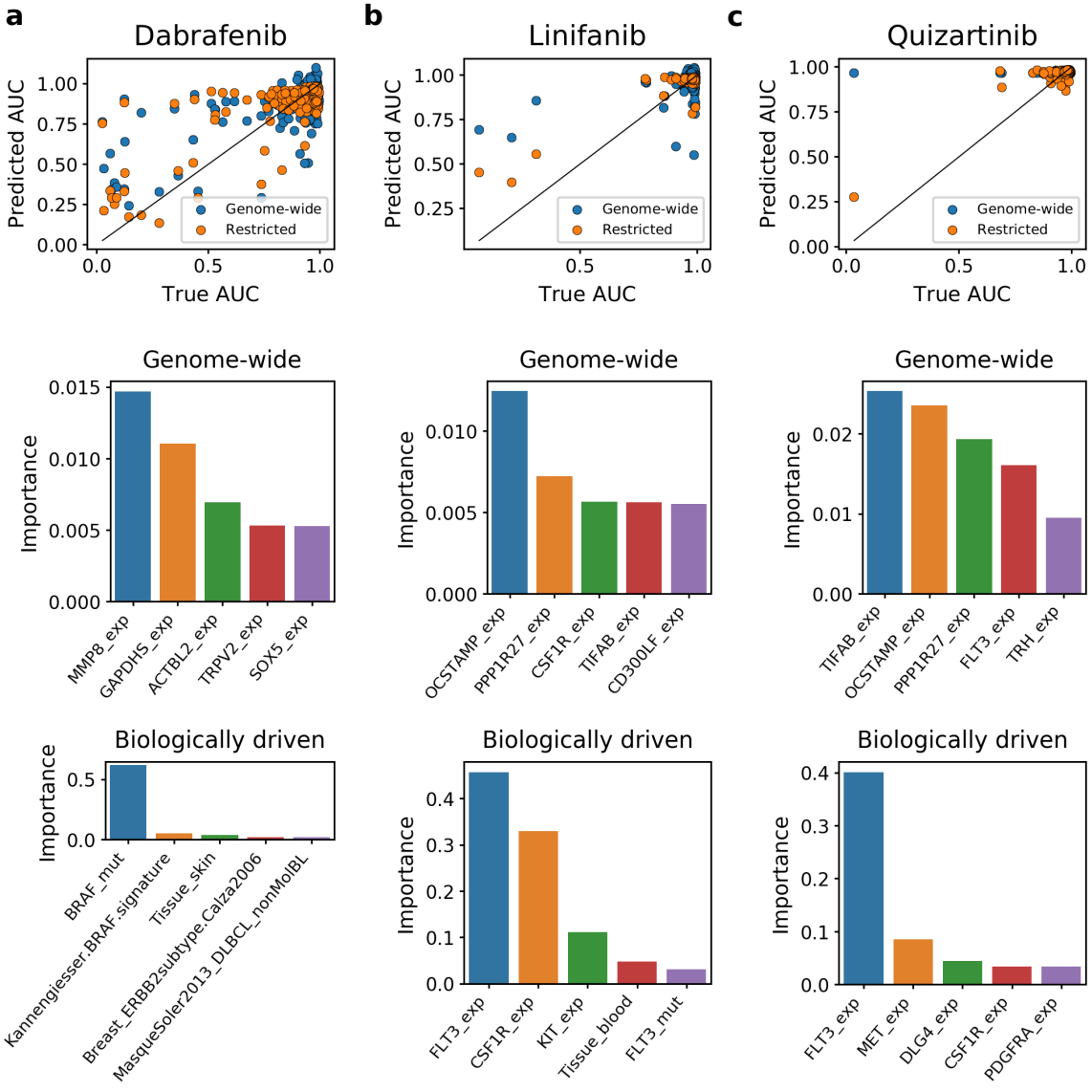


Figure 4.8: **Predicted versus actual AUC values and most predictive features for (a) Dabrafenib, (b) Linifanib and (c) Quizartinib.** Top panels show predicted versus actual AUC values when both biologically driven and genome-wide models were trained and tested on the same sets of samples. The biologically driven models correspond to best suited feature set for each drug: OT + S RF for Dabrafenib, OT RF for Linifanib and PG RF for Quizartinib. Middle and bottom panels present top 5 most informative features when fitting the model with genome-wide data (middle) and biologically driven feature space (bottom).

random forest, PG RF) uses features related to genes present in drug's target pathway (218 features), and the most important feature is expression of FLT3. The accurate prediction done by PG RF model for the single outlying, responsive sample (Fig. 4.8c) probably arises from the over-expression of FLT3 in that cell line (11.20 value for that feature in this sample versus the mean of 3.26 for all training samples). Although expression of FLT3 also appears as the fourth most important feature in the genome-wide model, it is unable to correctly predict AUC for the responsive cell line, since the relative impact of FLT3 is much smaller. Overall, these three examples well show that feature selection can facilitate derivation of interpretable insights.

## 4.4. Discussion

This work, constituting the first project of the thesis, is, to our knowledge, the first comprehensive analysis of feature selection strategies for drug sensitivity prediction. Previous systematic assessments [92, 70] compared different modeling techniques and data types describing the cell lines, but did not comprehensively evaluate feature selection approaches. Similarly, although numerous modeling methods were developed specifically for the task of drug sensitivity prediction [62, 99], they were solely optimized for predictive power and not interpretability. If feature selection was applied at all, it was not driven by pre-existing biological knowledge, but performed using standard and often not robust selection techniques such as regularization [107].

Such comprehensive feature selection assessment is needed for several reasons. First of all, both feature selection driven by pre-existing biological knowledge and data driven selection have their advantages and disadvantages. Intuitively, selecting the features using *a priori* knowledge of the drug mode of action as a guideline should improve modeling. On the other hand, it is also restricting the available information for the model, and if the prior knowledge is wrong, may result in missing important dependencies. Given the vast number of features compared to the number of samples, the models with genome-wide data as features or ones with automated feature selection are badly ill-posed and prone to over-fitting. On the other hand, they are given the advantage of a larger number of samples (resulting in higher power) and access to more information, compared to the models with biologically driven features (Fig. 4.2). Second, as there is no obvious recipe for choosing the feature set for a particular drug, the in-depth comparative analysis of different feature selection strategies may suggest indications for the recommended type of features for drugs depending on their mode of action or knowledge of their target pathway. Finally, if the best performing feature set is small, each particular feature can be inspected and further evaluated as a potential biomarker for the drug.

Here, different feature selection strategies driven by prior knowledge were compared to using genome-wide feature sets and the data-driven, automatic feature selection techniques across all analyzed drugs. We identified the best suited feature set for each drug and investigated them in the context of drugs' target pathways. Finally, we evaluated the predictive power of different features types and inspected example drug-specific models in more detail. The entire assessment workflow aimed at the identification of such strategies that could deliver highly predictive, but also highly interpretable models, bringing insights about specific drugs that are informative for their application in precision medicine.

Both Jang *et al.* [92] and the DREAM challenge [70] assessments indicated that adding the features representing mutation and copy number status on top of genome-wide expression features did not improve the overall performance of modeling drug sensitivity [92, 70]. This

is likely due to the fact that gene expression is sometimes already reflecting genomic changes or tissue type. In contrast, our analysis shows that additional features corresponding to mutations are often significant predictors when they are evaluated as part of smaller feature set and are not vastly outnumbered by the gene expression features (for example, in the cases of Dabrafenib, PLX-4720, Nutlin-3a, SB590885 and Pelitinib).

Our results bring important conclusions about feature selection strategies for drug sensitivity prediction. In general, the baseline genome-wide set of features or data-driven feature selection yields higher median predictive performance than biologically driven features. There are, however, multiple individual drugs, for which the feature selection driven by biological knowledge gives the best results, including models for the drugs with the top two performance scores. Moreover, feature selection driven by prior knowledge drastically reduces the number of features. At the same time, if the drop of performance in comparison with genome-wide models occurs, it is often only slight.

In addition, the presented analysis illuminates the mechanisms behind the sensitivity of different cancer cell lines to different types of drugs, suggesting which types of features should be used to model different classes of drugs. Drugs that are generally toxic or target general cellular mechanisms such as DNA replication or metabolism affect a relatively large proportion of cancer cell lines and thus have a wide response distribution. These compounds tend to be better modeled using genome-wide features, indicating that their effect on the cancer cells depends on a large spectrum of different cellular features. Conversely, for drugs targeting specific pathways, sensitivity distribution tends to be narrow, with most cells not responding at all and only a few interesting outliers of sensitive cells. For these compounds, high-level drug properties such as direct targets or target pathways allow to build highly predictive models with small numbers of interpretable features, such as Dabrafenib, Linifanib or Quizartinib. In particular, highly predictive models with an extremely low number of input features can be obtained, as in the cases of Linifanib, Afatinib, and GNF-2. Overall, this analysis shows the importance of using adequate feature selection strategies for each individual drug.

Overall, this work can be seen as a good entry point into a problem of drug sensitivity prediction, while providing guidelines regarding feature selection helpful for further research in this field.

# Chapter 5

# Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines

## 5.1. Background

Matching the optimal drugs for individual cancer patients remains a crucial problem of precision medicine [135]. Drug sensitivity data from cancer models are frequently generated to provide the basis for the discovery of molecular markers to predict drug efficacy. To predict the response of a specific cell line to a specific drug, there is a need of computational models that can leverage the abundance of information about drugs and cancer cell lines.

Kinase inhibitors are a class of anticancer drugs that target specific mutated kinases and disregulated biological processes in tumor cells [55]. As such, they constitute flagship examples of personalized cancer treatments [54, 57]. The set of kinase inhibitors is deeply investigated experimentally. First, they are commonly characterized by their *inhibition profiles*, measuring their strength of inhibition of a vector of kinases [136, 137]. Second, large-scale experiments were performed, measuring the sensitivity of cancer cell lines to these and other cancer compounds [138, 95, 139]. Third, the molecular features of the cancer cell lines, such as gene mutations and gene expression were measured [138, 95, 140]. Despite their limitations [14, 15, 141, 17], cancer cell lines commonly act as laboratory proxies for patients' tumors and it is known that their molecular features are key determinants of their response to anticancer drugs [95, 16]. Arguably, the kinase inhibitor drugs are best characterized by their kinase inhibition profiles, which, apart from the intended on-targets, manifest also off-target effects. Despite their frequent use during the early phases of drug development, when inhibitory profiles of kinase inhibitors are optimized, to our knowledge such data has not been used for modelling of drug response.

Computational drug sensitivity prediction has been approached by many machine learning methodologies [62, 70, 142], ranging from traditional algorithms [92, 102, 103, 113] to models based on neural networks and deep learning [116, 122, 121, 118, 143, 119]. Recently, the problem has also been addressed using generative modeling, including variational autoencoders [123, 144, 145], as well as using the reinforcement learning framework [146].

The problem of drug sensitivity prediction can be stated as a recommendation problem, where cancer cell lines and drugs are analogous to users and items, respectively. The goal

is to recommend the best drug for a given cell line. One of the most popular recommender system techniques is matrix factorization (MF), where the user-item interaction matrix is decomposed into a product of two lower-dimensional rectangular matrices. The problem of so called matrix factorization with side information incorporates features of users and items in the factorization process. The simplest approach to such MF problems involves linear projection of the features to lower-dimensional hidden space, followed by computing the dot product between corresponding user and item hidden representations in order to obtain user-item interaction prediction [147, 108, 148]. Recently, this approach has been modified by introducing non-linearity in the projection step, where the projections are computed by neural networks or autoencoders, but the corresponding hidden representations are still connected via a dot product in the linear fashion. Dot product, however, as a simple linear function, has a limited ability to capture the complex user-item interactions in the hidden space. To address this issue, deep neural networks have been proposed to replace the dot product for modeling the user-item interactions in the latent space [149, 150]. Since neural networks are known as the universal approximators [151], they are expected to be more suitable to learn complex relationships between the hidden representations of the users and items and the response variable.

While the neural-network based models are more expressive, previous analyses point out that the deep learning models do not necessarily outperform simpler models when the latter are finely tuned, and that some published neural network model results are hard to reproduce [152]. Moreover, deep neural networks have a reputation of being difficult to interpret due to their non-linearity and complex structure. The majority of so called explainable artificial intelligence methods focus on finding attributions between specific neurons in the network by analyzing the underlying gradient flow [153, 154, 155, 156]. Although useful, these methods provide rather standard utilities (e.g. feature importances), often available also for traditional machine learning models. Moreover, the insights derived from such interpretability approaches are limited by the features chosen for training the model.

We argue that a desired recommender system for the problem of drug sensitivity prediction should satisfy several objectives. First, it should solve a multi-task learning problem, i.e. model multiple drugs and cell lines simultaneously. This allows to capture general mechanisms driving the drug-cell lines interactions. Second, it should achieve state-of-the art predictive performance, especially in the task of predicting drug sensitivities for new cell lines. This is due to the fact that in this setting, the new cell line mimics a new patient, and the recommendation problem corresponds to identifying the best therapy for that patient. Finally, the model should be interpretable. Specifically, the model should explain the rationale behind its predictions and provide biological and pharmacological insights regarding the mechanism underlying the drugs-cell lines interactions. The emphasis on model interpretability is crucial in the context of its potential clinical applications.

To address these objectives, we develop a recommender system model for drug sensitivity prediction, called DEERS (Drug Efficacy Estimation Recommender System). DEERS incorporates two autoencoders to project the drug and cell line features, respectively, into lower dimensional representations, and uses a feed forward network to predict the sensitivities of the cell lines to the drugs based on their hidden representations. The proposed framework brings several advantages. First, the model solves a multi-drug and multi-cell line sensitivity learning problem and utilizes cell lines biological data and drugs inhibition profiles as side information (Fig. 5.1a,b). Second, the model is highly predictive. In a comparative analysis, DEERS outperforms two other MF-based recommender system models, and achieves similarly good results to the best performing XGBoost algorithm. Third, we provide an approach for model interpretability, on two levels: i) meaningful drug and cell line feature representa-

tion learning, and ii) explaining the cell line sensitivities to drugs in terms of the underlying biological processes.

The crucial aspect of the proposed interpretability approach is that it offers the widest possible assessment of the specific genes and biological processes that underlie the action of the drugs on the cell lines. The novelty of this approach stems from the fact that it considers also such genes and processes that were not included in the set of modeled features. Using the interpretablity approach, we demonstrate that the low-dimensional representations of the model capture the high dimensional features of drugs or cell lines, specifically the molecular patterns of cell lines and drug inhibition profiles that govern the response of distinct cell lines to drugs (Fig. 5.1c). Finally, we find the relationships between drug response and biological processes of cell lines (Fig. 5.1d).

## 5.2. Methods

### Analyzed data

The analyzed dataset comprised measurements of drug sensitivity of cell lines using viability assays for a total of 922 cell lines and 74 drugs, corresponding to 52,730 drug-cell line pairs. Both sensitivity metrics provided by the GDSC (AUC and IC50) were used to train and assess the performance of the presented models. Drug sensitivity of a cell line is the prediction target of our modeling approach.

The group of 74 drugs selected for modeling consisted exclusively of kinase inhibitors. The drugs in this group differ from other cancer drugs by their mode of action. Data to characterize the 74 kinase inhibitors were extracted from the HMS LINCS KINOMEscan data resource [157]. The features set of these drugs consisted of binding strength across a panel of 294 protein kinases (Fig. 5.1a). The value for a given compound-kinase pair represents a percent of control, where a 100% result means no inhibition of kinase binding to the ligand in the presence of the compound, and where low percent results mean strong inhibition [158, 159]. The data was acquired for those 74 drugs which were also present in the GDSC database, yielding a final drug characterization matrix for 74 drugs and 294 protein kinases.

Data to characterize the 922 cell lines were downloaded from the GDSC. For the molecular features of the cell lines, we considered only the genes coding for kinases present in KINOMEscan dataset, as well as any putative gene targets of all considered compounds. This resulted in the set of 202 genes, for which mRNA expression levels (202 features) and binary mutation calls (21 features) were extracted for all cell lines. Furthermore, the dummy-encoded tissue type was added, producing additional 18 binary features, yielding the final set of 241 biological features for 922 cell lines (Fig. 5.1a).

### DEERS: a deep neural network model of drug sensitivity accounting for inhibition of protein kinases by drugs and cancer cell line features

The goal of the proposed model is to predict a response of a given cell line to a given drug, i.e. estimate the corresponding AUC or IC50 value, given the drug and cell line feature representations (Fig. 5.1a). The final prediction is computed in two steps: first, we compute lower-dimensional representations of the considered drug and cell line, and second, the representations are combined, in order to make the sensitivity estimation. This problem can be viewed as a matrix factorization task, where every element of the target matrix $y^{(i,j)}$ is

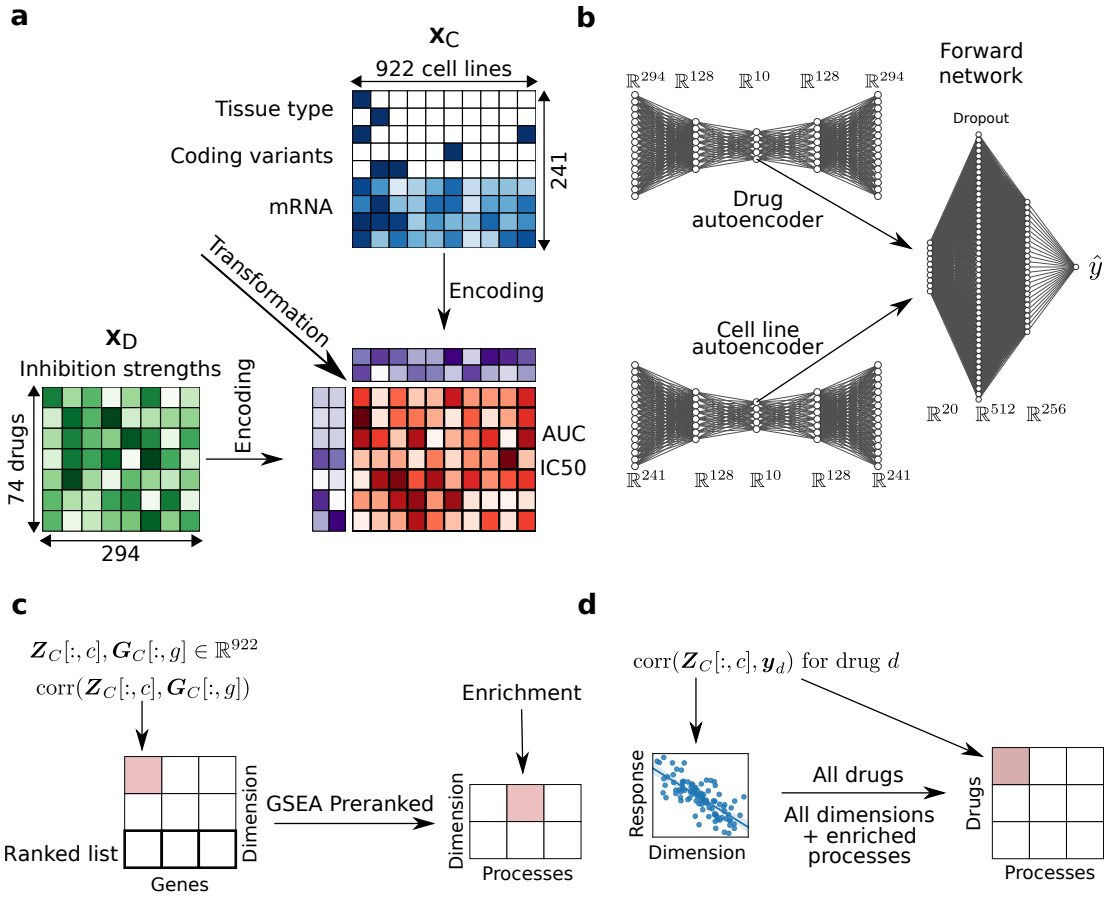


Figure 5.1: **Overview of the data and the modeling process.** (**a**) Recommender system framework for drug sensitivity prediction from drug and cell line features. The drugs are described by their inhibition profiles on a panel of 294 kinases. The biological features of the cell lines include continuous mRNA expressions, binary indicators of coding variants, and dummy-encoded tissue type. Two drug response metrics are considered: AUC and IC50. The recommender system first independently encodes drugs and cell lines input data into lower-dimensional representations. The two hidden representations are then transformed in order to compute the drug response estimation. (**b**) Architecture of the DEERS model. First, the drugs and cell lines inputs are passed into corresponding autoencoders which output the 10-dimensional representations and reconstructed data. Next, the hidden representations are concatenated and used as an input to the two-layered, feed-forward network which outputs the drug response estimate. (**c**) Method for relating biological meaning to hidden dimensions of cell lines. First, the hidden dimensions of the cell line autoencoder are correlated with gene expression data. Here, $\boldsymbol{Z}_C$ denotes the matrix with cell line hidden representations stacked in rows, $\boldsymbol{Z}_c[:,c]$ denotes a column of $\boldsymbol{Z}_C$, $\boldsymbol{G}_C$ denotes the gene expression data for cell lines and $\boldsymbol{G}_C[:,g]$ denotes a column of $\boldsymbol{G}_C$. The resulting ranked lists, one per each dimension, are then passed as an input to GSEA Preranked analysis, obtaining biological processes enriched in every hidden dimension (see Methods). (**d**) Method for relating the drug action directly to biological processes. For a given drug $d$, the cell line response is correlated with a given cell line hidden dimension $c$. The obtained correlation coefficient is then mapped to the biological processes enriched in hidden dimension $c$. This procedure is performed for every drug and every hidden dimension, obtaining the matrix relating drugs to the biological processes (see Methods).

modeled as some form of a transformation of the corresponding hidden representations of the drug and cell line (Fig. 5.1a).

DEERS is a deep neural network-based recommender system. It consists of three major parts: drug autoencoder, cell line autoencoder and the subsequent feed-forward neural network. (Fig. 5.1b) [82, 160]. The two autoencoder networks have the same architecture, with one 128-dimensional hidden layer in both encoder and the decoder with the rectified linear unit (ReLU) activation function, and the 10-dimensional hidden representation layer. The subsequent feed-forward network consists of a 20-dimensional input layer, followed by two hidden layers of length 512 and 256 with the ReLU activation. The regularization of the system is incorporated via the dropout with 0.5 probability, applied in the first, 512-dimensional hidden layer of the feed-forward network.

Consider a training data point consisting of original drug $i$ and cell line $j$ feature vector representations along with the corresponding response value, $(\mathbf{x}_D{}^{(i)}, \mathbf{x}_C{}^{(j)}, y^{(i,j)})$. The input training data vectors are first passed into drug and cell line autoencoders, producing reduced, 10-dimensional vector representations (the hidden representations) $(\mathbf{z}_D{}^{(i)}, \mathbf{z}_C{}^{(j)})$ and reconstructed inputs $(\mathbf{x}_D^{'\,(i)}, \mathbf{x}_C^{'\,(j)})$ (Fig. 5.1b). The hidden representations $\mathbf{z}_D{}^{(i)}$ and $\mathbf{z}_C{}^{(j)}$ are then concatenated, forming a 20-dimensional vector, which serves as an input for the subsequent feed-forward neural network, which in turn computes the final response estimate $\hat{y}^{(i,j)}$ (Fig. 5.1b).

DEERS has three outputs and three main optimization goals: minimizing the differences between $\mathbf{x}_D{}^{(i)}$ and $\mathbf{x}_D^{'\,(i)}$, minimizing the differences between $\mathbf{x}_C{}^{(j)}$ and $\mathbf{x}_C^{'\,(j)}$, and minimizing the errors between $y^{(i,j)}$ and $\hat{y}^{(i,j)}$. The incorporation of reconstruction errors causes the network to find informative representations of the input drug and cell line features. In addition, it is desired for the hidden dimensions to be independent. This enables the hidden representations to capture more information about the full input data and facilitates easier interpretations of the hidden dimensions. In the proposed model, it is achieved by minimizing the squared values in the off-diagonal entries of the drugs and cell lines covariance matrices in the latent space. All of the described optimization tasks are captured by a single cost function, which is iteratively minimized for each training batch to train the model:

$$\begin{aligned} J(\boldsymbol{W}) = {} & \mathrm{MSE}(\mathbf{y} - \hat{\mathbf{y}}) \\ & + r_D \cdot \mathrm{MSE}(\boldsymbol{X}_D - \boldsymbol{X}'_D) + r_C \cdot \mathrm{MSE}(\boldsymbol{X}_C - \boldsymbol{X}'_C) \\ & + d \cdot \sum_{m,n,m \neq n} (\boldsymbol{K}_D[m,n])^2 + d \cdot \sum_{m,n,m \neq n} (\boldsymbol{K}_C[m,n])^2, \end{aligned} \tag{5.1}$$

where $J$ is the cost function, MSE denotes mean squared error, $\boldsymbol{W}$ is a set of the model parameters (weights), $r_D$ is the real-valued weight of the drugs reconstruction error, $\boldsymbol{X}_D$ is the drugs' data matrix in the training batch, $\boldsymbol{X}'_D$ is the drugs data reconstruction matrix in the batch, $r_C$ is a real-valued weight of the cell lines reconstruction error, $\boldsymbol{X}_C$ is the cell lines data matrix in the batch, $\boldsymbol{X}'_C$ is the cell lines data reconstruction matrix in the batch, $d$ is a weight of the dependence penalty, $\boldsymbol{K}_D$ is the covariance matrix of drugs hidden representations in the batch, and $\boldsymbol{K}_C$ is the covariance matrix of cell lines hidden representations in the batch, and $\boldsymbol{K}[m,n]$ denotes the $m,n$-th entry of matrix $\boldsymbol{K}$.

Intuitively, the cost function weights $r_D$, $r_C$ and $d$ control the contribution of the particular optimization task in the general optimization goal of the system. Setting all of these weights to zero would result in a network without decoding tasks and no dependence restrictions on the hidden dimensions of the drugs and cell lines.

## Compared models

We compare the proposed model to four other methods; two of which are based on traditional machine learning algorithms, while the other two are forms of matrix factorization.

In order to evaluate the traditional methods in a multi-task setting, where the data for all drugs and all cell lines are modeled at once, the traditional methods are used to predict drug response for the union of drugs and cell lines features. To this end, for every data point $(\mathbf{x}_D{}^{(i)}, \mathbf{x}_C{}^{(j)}, y^{(i,j)})$, we first concatenate vectors $\mathbf{x_D}{}^{(i)}$ and $\mathbf{x_C}{}^{(j)}$, forming one 535-dimensional vector per drug-cell line pair. Applying this to all available data points produces a $52730 \times 535$ input data matrix $\boldsymbol{X}$ and the corresponding 52730-dimensional vector with true response values $\mathbf{y}$. This data is used to train and evaluate two common machine learning algorithms: Elastic net [161] and XGBoost [162]. The former is a linear model and the latter is a more complex, nonlinear model.

The compared matrix factorization models aim at solving a similar matrix-factorization type of problem (Fig. 5.1a) and can be seen as simpler or reduced versions of the proposed model. The first is a basic matrix factorization with side information method, reducing the dimension of the additional information about both drugs and cell lines using linear projections, and applying a dot product to produce the prediction of the response variable (here, the sensitivity of cell lines to drugs). We refer to this model as Lin MF (Fig. S3a). The basic architecture of this model is the same as the model applied by Yang *et al.* [108].

The second of the compared matrix factorization-based models is an non-linear extension of the basic model, where the dimensionality reduction is performed via one-layered autoencoders and data reconstruction is also taken into consideration (Fig. S3b). Similarly as in Lin MF, the final prediction is obtained by taking the dot product of the corresponding hidden representations, in contrast to the proposed DEERS model, where a separate feed-forward network is used to obtain the response estimate (Fig. 5.1b). We refer to this model as Autoen MF. To estimate the parameters of both Lin MF and Autoen MF we use gradient descent optimization implemented in Adam optimizer [163].

## Experimental setup and model training

In order to assess the performance of the considered models on the unseen cell lines, we construct the validation and test sets by first randomly selecting two sets of 100 unique cell lines each. We then extract the data points containing selected cell lines, producing the validation and test sets with $\sim$ 5000 drug-cell line pairs each. The rest of the pairs corresponding to the remaining 722 unique cell lines (with $\sim 42,000$ pairs) constitute the training set.

Before the training, the input cell line data were preprocessed by standard scaling of the continuous gene expression data so that every feature has zero mean and unit standard deviation, while binary coding variants and dummy encoded tissue types were unmodified. For the input drug data, all features were standardized in the same way as the gene expression. Since the GDSC AUC values are in the range of [0, 1], they were not scaled, while the log IC50 values were linearly preprocessed with min-max scaler to the [0, 1] range. Notably, all values necessary to perform each of the applied preprocessing schemes were calculated only on the training set and applied to the validation and test sets.

We use the training and validation sets in order to find the optimal set of hyperparameters, consisting of: network architecture, cost function weights $r_D$, $r_C$ and $d$, regularization type and learning rate. We establish the DEERS architecture as consisting of two-layer autoencoders, with 10-dimensional hidden representations (Fig. 5.1b). The subsequent feed-forward network

has two hidden layers of size 512 and 256. The optimal cost function weights were set to $r_d = 0.1$, $r_C = 0.25$ and $d = 0.1$. As a regularization type, we use combination of dropout applied in the first hidden layer of the feed-forward network (Fig. 5.1b) and early stopping. With these hyperparameters fixed, for every split of the data (into the training, validation and test sets) we tune the learning rate, dropout rate and number of epochs for early stopping.

After all parameters are found, we use them to train the model using the union of training and validation sets, and apply the resulting model to the test set in order to assess the performance. We repeat this procedure ten times with different cell lines in training, validation and test sets in order to improve the robustness of the results.

We adopt the similar methodology for the compared models, where we first tune the hyperparameters using training and validation sets, and then apply the final retrained model to the test set, using the same data splits for training, validation and testing for all models.For the compared models, we perform this experimental procedure five times. In addition, we incorporate a simple data augmentation scheme, where we add a random gaussian noise with zero mean to the cell lines gene expression data and the corresponding AUC or IC50 values. The standard deviations of cell lines and response noise were 0.6 and 0.15, respectively. The augmentation was performed iteratively in every batch during training, tripling the original batch size. This data augmentation scheme was added for the two models involving autoencoders, i.e. both the Autoen MF and the DEERS model.

## Interpretation of hidden dimensions in DEERS

This analysis aims at an explanation of the model predictions from the biological standpoint. In order to incorporate all available data for model interpretation, we first re-train the model with all available 922 cell lines and 74 drugs, without excluding any cell lines, and using IC50 as a drug response metric.

The interpretation of the hidden dimensions concerns assigning a biological meaning to the individual dimensions of the hidden space. To this end, we first pass the input drugs and cell lines input representations into their corresponding, already trained autoencoders, producing a 10-dimensional representation for each 294-dimensional input data vector corresponding to a drug and a 10-dimensional representation for each 241-dimensional input data vector corresponding to a cell line, respectively.

### Associating input features with hidden dimensions

To compute the association of each input feature with each hidden dimension, we utilize the Integrated Gradients method [156], by computing the attributions between input features and the ten neurons constituting the hidden representation layers. This is performed separately for the drug and the cell line autoencoders, and the attributions are averaged across the drugs and cell lines, respectively. As a result, we obtain drugs and cell lines feature-representation attribution matrices of size $294 \times 10$ and $241 \times 10$, respectively, where each entry is a score reflecting how much a given feature impacts the given variable in the hidden space. We then perform the row-wise hierarchical clustering on the resulting attribution matrices, grouping features associated with the same dimension together. The clustering was performed after normalizing the rows to unit norm, using the Ward linkage method and the Euclidean distance metric. This interpretability approach is applied separately for the 10 dimensions encoding the drugs and for the 10 dimensions encoding the cell lines.

### Associating biological processes with hidden dimensions encoding the cell lines data

In this interpretability analysis, we exploit the fact that the cell line autoencoder in DEERS is trained to reconstruct the data and to find low-dimensional representations that reflect the true properties of the analyzed cell lines. The produced hidden representations of the cell lines are organized into a $922 \times 10$ matrix $\boldsymbol{Z}_C$, where every row $j$ corresponds to a single cell line hidden representation, and every column $c$ represents the values of a given hidden variable across all cell lines. Next, we examine the full genome-wide gene expression data of the full set 17,419 genes extracted from GDSC. In this way, this analysis goes beyond the restricted set of the modeled input 241 cell line features. Using this data, we construct a $922 \times 17419$ matrix $\boldsymbol{G}_C$, where every row $j$ corresponds to a single cell line gene expression profile, and every column $g$ represents the expression values of a given gene across the examined cell lines. We then compute a $17,419 \times 10$ correlation matrix $\boldsymbol{C}$, where every entry $\boldsymbol{C}[g, c]$ corresponds to Spearman correlation coefficient between $g^{th}$ column of $\boldsymbol{G}_C$ and $c^{th}$ column of $\boldsymbol{Z}_C$, i.e. the correlation between the expression of a given gene and a value of a given hidden dimension across 922 considered cell lines (Fig. 5.1c) .

Given such correlation matrix $\boldsymbol{C}$, we create a ranked list of genes for every hidden dimension, where the ranking metric is the correlation coefficient of the genes with that dimension. The genes at the top and bottom of the ten resulting ranked lists are the ones that are most positively or negatively correlated with the corresponding dimensions, respectively. We then take the first and the last 1000 genes with corresponding correlation coefficients for every hidden dimension and run the GSEA Preranked analysis [164] against gene sets that are involved in specific biological processes as defined by the Biological Process GO Terms (Fig. 5.1c). The GSEA Preranked is performed using the gseapy Python package [164, 165, 166]. We then extract the top 15 enriched terms with the smallest FDR value for every hidden dimension, which indicates the general biological mechanisms are most related to that dimension. Finally, we eliminate the redundant gene ontology terms using the Revigo tool [167], assigning the set of biological mechanisms to every dimension of the cell lines hidden space.

## 5.3. Results

DEERS was developed with two aims in mind. One, to achieve state-of-the art predictive performance in predicting the response of cancer cell lines to kinase inhibitor drugs. Second, to identify the biological mechanisms that drive this response. Below, we evaluate the performance of DEERS in comparison to other models and conduct its interpretability analysis.

### Evaluation of the predictive performance of DEERS in comparison to other models

The drug sensitivity measurements were acquired from the Genomics of Drug Sensitivity in Cancer (GDSC) [138] database. GDSC provides two sensitivity measurements, summarizing the dose-response curve: area under the curve (AUC) and log half maximal inhibitory concentration (IC50), defined as a drug concentration needed to reduce cell viability by 50%.

The predictive performance of DEERS is compared with four other methods. Two of those, Elastic net and XGBoost, are traditional, frequently used machine learning algorithms. Remaining two, referred to as Lin MF and Autoen MF, are versions of matrix factorization with side information (see Methods for a description of the compared models). In order to evaluate the performance of DEERS and other models on a test set containing responses of

unseen cell lines, we first pass the drugs and cell lines input data to the model and obtain a table of predicted responses for each drug and cell line pair. Given such a table, we calculate the Pearson correlation and RMSE (root mean squared error) of the true to predicted responses across all drug-cell line pairs. In addition to such metrics calculated globally, we also group the previously described table, and calculate correlation (abbreviated corr.) and RMSE of true and predicted responses across pairs per given drug or cell line. To aggregate the per-drug and the per-cell line results, we take the median across the cell lines and drugs, respectively. The per cell line results mimic an envisioned clinical application of the model, where prediction of drug efficacy will be made for a new patient with specific tumor features, enabling a personalized medicine approach. This evaluation scheme yields six performance metrics per model (referred to as "Pairs RMSE", "Pairs corr.", "Per-drug RMSE", "Per-drug corr.", "Per-cl RMSE" and "Per-cl corr."). These metrics are evaluated both for IC50 (Tab. 5.1) and AUC (Tab. 5.2). The metrics are computed for several experiments with different random data splits into training, validation and test sets (ten experiments for DEERS and five for each of the compared methods). In order to obtain a more robust comparison between DEERS and the simpler approach of matrix factorization with side information, we group the results for the five experiments of the Lin MF and Autoen MF models, yielding two ten-element groups of results per each evaluation metric (ten for DEERS and ten for the matrix factorization with side information). We then perform the one-sided Wilcoxon rank-sum tests, testing whether DEERS obtains statistically significantly better performance in a given evaluation metric.

In general, IC50 as a prediction target is easier to learn than AUC. Indeed, in terms of correlation between predicted and true response values, better results are obtained by all models for IC50 than for AUC.

| | Alg. type | Pairs RMSE | Pairs corr. | Per-drug RMSE | Per-drug corr. | Per-cl RMSE | Per-cl corr. |
|---|---|---|---|---|---|---|---|
| Elastic net | T | 0.09 ±0.002 | 0.80 ±0.007 | **0.08 ±0.019** | 0.31 ±0.155 | **0.08 ±0.002** | 0.84 ±0.003 |
| XGBoost | T | **0.08 ±0.002** | **0.83 ±0.009** | **0.08 ±0.017** | **0.40 ±0.131** | **0.08 ±0.001** | **0.86 ±0.006** |
| Lin MF | RS | 0.09 ±0.003 | 0.78 ±0.012 | 0.09 ±0.003 | 0.30 ±0.045 | **0.08 ±0.002** | 0.85 ±0.008 |
| Autoen MF | RS | 0.09 ±0.002 | 0.80 ±0.009 | 0.09 ±0.004 | 0.31 ±0.024 | 0.08 ±0.003 | 0.84 ±0.006 |
| DEERS w/o inhib. profs. | RS | 0.09 ±0.002 | 0.80 ±0.012 | **0.08 ±0.002** | 0.38 ±0.047 | **0.08 ±0.002** | 0.84 ±0.003 |
| DEERS | RS | **0.08** *** **±0.002** | **0.82** *** **±0.006** | **0.08** *** **±0.002** | **0.41** *** **±0.035** | **0.08** *** **±0.002** | **0.86** ** **±0.010** |

Table 5.1: **Predictive performance of DEERS and compared models when using IC50 as a drug response metric.** The presented values are averages of metrics taken across several experiments (ten for DEERS and five for each other method), with different data splits, along with the corresponding standard deviations. The presented per-drug and per-cell line results are medians taken across all considered drugs and cell lines, respectively. The evaluated models are split into two categories: frequently used, traditional machine learning algorithms (T) and recommender system class (RS). Best results within a model category are highlighted with bold font, while the best results overall are underlined. Asterisks indicate the intervals containing the p-values of the one-sided Wilcoxon rank-sum tests of the better performance of DEERS over the other two RS models: no asterisks – [0.05, 1), * – [0.01, 0.1), ** – [0.001, 0.01), *** – (0, 0.001). Abbreviations: alg. – algorithm, corr. – correlation, cl – cell line, w/o inhib. profs. – without inhibition profiles.

With IC50 as the response variable, the DEERS model mostly outperforms or at least performs similarly well as the other two matrix factorization-based models with regard to

all of the six performance metrics (indicated by bolded values in Tab. 5.1). For IC50, the XGBoost outperforms the other traditional method, Elastic net, in all performance measures. This indicates that nonlinear models are needed to capture the dependence of IC50 on drug and cell line features. DEERS and XGBoost achieve comparable evaluation results (with the best model according to each evaluation metric underlined in Tab. 5.1). In particular, DEERS obtains a high Pearson correlation coefficient r=0.82, calculated on all drug-cell line pairs in the test set. Moreover, the median per cell line correlation of r=0.86 indicates that DEERS achieves the state-of-the-art performance in predicting cell line responses to drugs, which most closely resembles the hypothetical clinical setup. Notably, compared to per-cell line correlation, all models obtain relatively poor results in terms of per-drug correlation. This may be due to the fact that our input data is asymmetric as it covers much fewer drugs (74) than cell lines (922).

| | Alg. type | Pairs RMSE | Pairs corr. | Per-drug RMSE | Per-drug corr. | Per-cl RMSE | Per-cl corr. |
|---|---|---|---|---|---|---|---|
| Elastic net | T | 0.13 ±0.002 | 0.71 ±0.011 | 0.11 ±0.050 | 0.23 ±0.188 | 0.12 ±0.003 | 0.77 ±0.005 |
| XGBoost | T | **0.12 ±0.002** | **0.77 ±0.013** | **0.10 ±0.050** | **0.34 ±0.176** | **0.11 ±0.002** | **0.81 ±0.012** |
| Lin MF | RS | 0.13 ±0.004 | 0.73 ±0.012 | **0.11 ±0.005** | 0.34 ±0.044 | 0.12 ±0.004 | 0.80 ±0.011 |
| Autoen MF | RS | 0.13 ±0.005 | 0.75 ±0.008 | **0.11 ±0.005** | 0.27 ±0.044 | 0.12 ±0.006 | 0.80 ±0.003 |
| DEERS | RS | **0.12** * **±0.004** | **0.76** ** **±0.013** | **0.11 ±0.005** | **0.35** * **±0.027** | **0.11 ±0.005** * | **0.81 ±0.014** |

Table 5.2: **Predictive performance of DEERS and compared models when using AUC as a drug response metric.** Table columns and formatting the same as in Tab. 5.1

In the case of AUC as the response variable, the comparison of model performance yields similar results as the IC50. Here again DEERS outperforms the other two matrix factorization-based methods, while from the two traditional methods XGBoost performs better than Elastic net (Tab. 5.2). Overall, the performance of DEERS is very similar to XGBoost. For AUC, the DEERS achieves r=0.76 Pearson correlation coefficient calculated on all drug-cell line pairs in the test set. For the per-cell line results, the median correlation across the unseen cell lines is r=0.81, constituting the best result along with XGBoost.

### Evaluation of the added value of inhibition profiles and putative targets

In order to quantify the benefit of incorporating inhibition profiles of the drugs, we performed an ablation study and estimated the performance of DEERS with drug putative targets as drug input data. This model is referred to as "DEERS without inhibition profiles" in Tab. 5.1. To this end, the reduced drug features were defined by a binary matrix with 74 rows corresponding to kinase inhibitors and 92 drug targets, and entries 1 if the drug has the gene as target and 0 otherwise. With this alternative drug input data and IC50 as a target variable, we evaluated DEERS using the same procedure as previously (with five experimental iterations), with all hyperparameters besides learning and dropout rates unchanged. Learning and dropouts rates were tuned using validation set in the same manner as before. DEERS with inhibition profiles outperforms DEERS with binary targets in 3 evaluation metrics (Pairs RMSE, Pairs corr., Per-cl corr.), achieves the same results in 2 metrics (Per drug RMSE and Per-cl RMSE) and slightly underperforms in Per-drug corr. metric. The improvement in Pairs RMSE, Pairs corr., Per-cl corr. metrics constitutes 11.1%, 2.5%, and 2.4% relative increase, respectively.

We also performed the ablation study in the opposite direction, and removed the set of putative targets of the analyzed 74 drugs from the kinases panel used to describe the inhibition profiles. This resulted in removing of 45 kinases, leaving 249 features to describe the drugs. This change did not affect the predictive performance, as there was no significant results difference in any evaluation metric compared to DEERS trained with the whole panel of 294 kinases. This analysis underlines the added value of accounting for the inhibition profiles of the drugs.

### Evaluation on an independent dataset

In order to estimate the performance of DEERS on other data than cell line sensitivities from GDSC, we extracted drug sensitivity data from the Cancer Cell Line Encyclopedia (CCLE) [95] project. Next, we constructed a dataset consisting of an intersection between the our analyzed dataset (containing data for 74 drugs derived from GDSC for kinase inhibitors), and the CCLE dataset in terms of cell lines and drugs, along with corresponding, min-max-scaled CCLE IC50 values. The data regarding the intersection between GDSC and CCLE, as well as CCLE IC50 values were extracted using the PharmacoDB package [168, 64]. The resulting dataset contained 351 common cell lines and 5 common drugs (Crizotinib, Lapatinib, PD0325901, PLX-4720 and Sorafenib), constituting 1747 pairs in total. The cell lines and drugs were described by the same features as in the original GDSC dataset. We next used the GDSC data corresponding to the remaining 571 cell lines that are not present in the CCLE-GDSC intersection dataset and all 74 drugs to train DEERS. From those 571 cell lines of GDSC, 50 were randomly chosen to construct the validation dataset for tuning the learning and dropout rates (see Methods). We then re-trained the model with the best hyperparameters on all 571 cell lines and applied it to the CCLE-GDSC intersection dataset, obtaining IC50 predictions for unseen cell lines. It is important to note that the maximum obtainable correlation between the model predictions and the true IC50 values in the intersection dataset in this experiment is 0.53, defined by the correlation between the true IC50 values in the CCLE dataset and the true IC50 values in the GDSC for these cell line-drug pairs. Given this upper bound, the obtained correlation result of 0.40 is relatively high. In comparison, for the XGBoost evaluated in the same scheme as described above, the obtained correlation is 0.39.

Taken together, these results demonstrate that thanks to its deep neural network-based recommender system architecture and utilization of informative drug features, DEERS obtains state-of-the art performance in predicting cell lines sensitivity to drugs in a multitask setup. In contrast to the other well performing model, XGBoost, however, DEERS obtains highly informative reduced-dimension representations of the cell line and drug features, respectively. This aspect of the model is discussed below.

### Attributions between input features and hidden dimensions using neural network analysis

As the first step of the DEERS model interpretability analysis, we computed the attributions between the input features and the hidden dimensions using Integrated Gradients (see Methods). Next, we performed hierarchical clustering of the resulting attribution matrix, in which the rows were the features, and columns were the hidden dimensions. The clustering identifies well-defined groups of features associated with each specific hidden dimension (Fig. B1). There is very little overlap between feature groups for both drugs and cell lines, indicating that hidden dimensions are independent in terms of which features affect them the most.

This independence effect is also evident when we compare the covariance matrices of drugs and cell lines represented in a hidden space, when dependence penalty was incorporated and not incorporated into the overall cost function (Fig. B2). The number of drug input features associated with a single hidden dimension ranges from 20 for dimensions 2 and 8, to 44 for dimension 1. For the cell lines, this number ranges from 11 for hidden dimension 7 and 44 for hidden dimension 1.

### Linking hidden dimensions to the general biological mechanisms

In the next step of the interpretability analysis, we associate each hidden dimension of the cell line autoencoder with a biological process. To this end, for each hidden dimension and each gene, we correlate the values of the hidden dimension with the expression values of the gene across cell lines. For a given hidden dimension, the obtained correlations are then ranked and we apply gene set enrichment analysis (GSEA) to identify biological processes positively or negatively correlated with that dimension (Fig. 5.1c). Importantly, this analysis links the dimensions to all genes measured in the cell lines, that is, also to the genes outside of the cell line features used in the model (see Methods for a full description of this analysis). Here, we run the GSEA considering the gene-sets included in the Gene Ontology Biological Processes. The analysis and subsequent filtering of redundant terms yield a final set of GO terms for each dimension of the hidden space of the cell line autoencoder (Fig 5.2). We identify 67 GO terms in total, many of which are related to cancer (e.g. DNA replication, regulation of cell cycle process, regulation of angiogenesis). The number of enriched terms per dimension varies from 6 to 13. The majority of enrichment scores (67%) are positive, which indicates that they are positively correlated with that dimension. Conversely, the negatively signed FDR value implies that the given term is negatively correlated. Markedly, the sets of enriched terms almost do not overlap between the dimensions, indicating the independence of the dimensions in terms of their associated biological mechanisms. Out of 67 terms, only 12 are associated with more than one hidden dimension, from which 10 are associated with two dimensions.

When inspecting the heatmap (Fig. 5.2), we identify groups of biological mechanisms associated with specific hidden dimensions. For example, hidden dimension 2 is mainly linked with DNA replication and cell cycle, as terms enriched in it include: DNA replication, DNA-dependent DNA replication, G1/S transition of mitotic cell cycle and regulation of cell cycle process. Dimension 4 is related to protein metabolism (post-translational protein modification, cellular protein metabolic process, cellular protein modification process), while dimension 3 is connected with DNA and RNA metabolism (DNA metabolic process, RNA metabolic process, rRNA metabolic process) and known cancer-related processes like regulation of MAPK cascade and regulation of angiogenesis. Other such terms include regulation of extrinsic apoptotic signaling pathway (dimension 9), cellular response to DNA damage stimulus (dimension 6), cellular response to tumor necrosis factor (dimension 0) and DNA damage response, signal transduction by p53 class mediator (dimension 1). Interestingly, some of the terms are not commonly linked to cell cycle or other processes related to oncogenesis, e.g. for dimension 8 the set of enriched terms includes central nervous system development, nervous system development and axonogenesis. This analysis provides a form of interpretation of hidden dimensions from the biological standpoint and facilitates a better understanding of the model prediction based on cell lines hidden representations. Overall, the obtained list of biological processes reflects the repertoire of common biological mechanisms that are affected by the analyzed kinase inhibitors in the set of analyzed cell lines, and as a general summary can only be obtained from such a multitask learning model as DEERS.

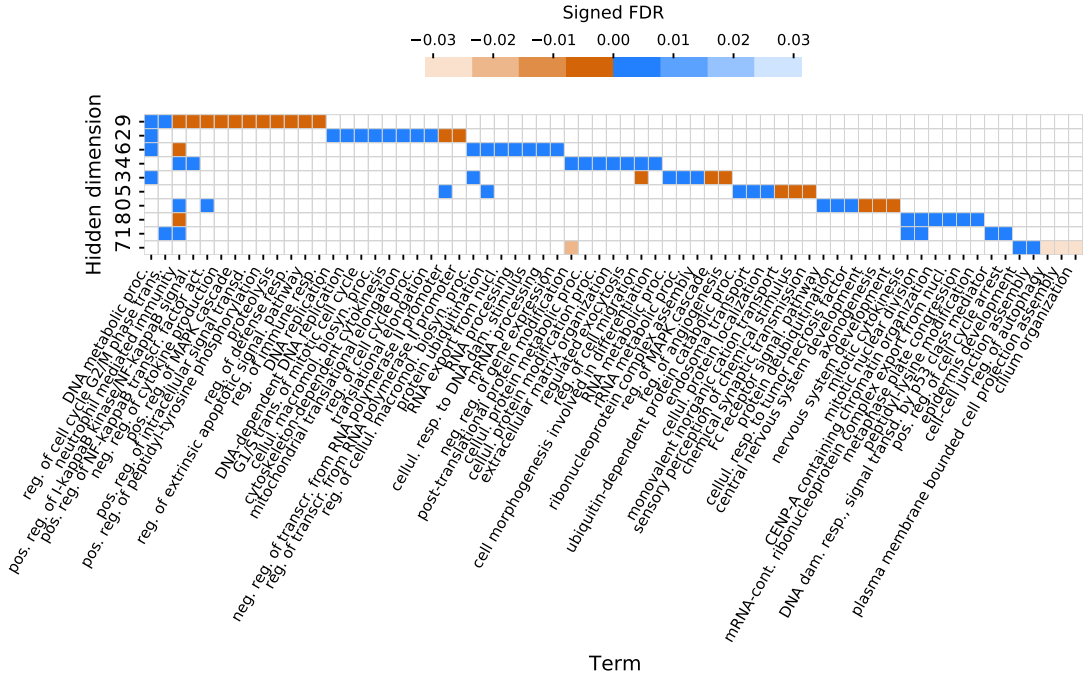


Figure 5.2: **Heatmap reflecting Biological Process GO terms enriched in hidden dimensions of the cell line autoencoder.** Negatively and positively signed FDR values correspond to negative and positive enrichment scores, i.e. negative and positive side of the ranked list, respectively. Hidden dimensions are sorted by the number of enriched terms. Abbreviations: proc. – process, reg. – regulation, resp. – response, neg. – negative, pos. – positive, transcr. – transcription, transd. – transduction, mRNA-cont – mRNA-containing, dam. – damage, nucl. – nucleus, macromol. – macromolecule, biosyn. – biosynthetic, cellul. – cellular, signal. – signaling, act. – activity, trans. – transition.

### Case studies

We further focus the analysis on three case studies, showing how the model predictions and true responses can be explained and interpreted for individual drugs and features. For this purpose, we examine three specific compounds: the pan-CDK inhibitor PHA-793887, the ALK/CDK7 inhibitor XMD14-99 and the BRAF inhibitor Dabrafenib (Fig. 5.3). First, we establish which features are most important for the model prediction given the input data for a particular compound. To this end, we calculate the attributions between input features and the final output layer of the model using the Integrated Gradients method [156]. The attributions are first computed separately for each cell line and IC50 as the response variable, and next summarized by averaging over all cell lines. Second, for each compound we display the cell lines in two chosen dimensions of the hidden space of the cell line autoencoder, and color them by their IC50 response to the compound. In this way, we identify such regions in this space that are correlated with sensitivity to the compound. Finally, we explore in detail how well one chosen hidden dimension correlates with the true response and we list the biological processes that are associated with that dimension (as per analysis in Fig. 5.2). Altogether, the case studies identify such features and hidden dimensions that are important for modeling the response, and such biological processes that are important for the action of the three analyzed drugs.

PHA-793887 is an inhibitor of multiple cyclin dependent kinases (CDK) with activity against CDK2, CDK1 and CDK4 [169]. According to the attribution analysis, the activity against CDK is reflected in the most informative drug features, where CDK2 kinase is one of the most important drug features for prediction (Fig. 5.3a, top row panel). However, the most important feature of that drug is MK03. According to Uniprot [170], the gene coding for MK03 is *MAPK3*, also known as ERK1 [171]. Other CDK inhibitors have been shown to inhibit not only the CDKs, but also ERK1 [172]. Moreover, there is an evidence within the KINOMEscan data, stating that several other MAP-kinases (but not including MAPK3) are inhibited by PHA-793887 [173]. Interestingly, cell line features corresponding to the CDK family do not obtain top attribution values. Instead, the highest average attributions are associated with the expression of *BTK, PIM2* and *TEC* genes, suggesting that their activity in the cell lines is important for PHA-793887 action (Fig. 5.3a, second row panel). Again, there is some evidence for another CDK inhibitor, abemaciclib, targeting one of the listed genes, namely PIM kinase [174]. Representing cell lines in two dimensions (by hidden dimensions 3 and 0) identifies a region corresponding to a good response of PHA-793887 (Fig. 5.3a, third row panel). This validates that that in general the hidden dimensions well represent the cell line data and that in particular these two hidden dimensions well capture the cell line response to PHA-793887. However, most of the cell lines response variance can be explained using hidden dimension 3 alone, which is negatively correlated with the true response (Pearson correlation r = -0.40; Fig. 5.3a, bottom row panel). The biological process terms enriched for different dimensions, visualized in Fig. 5.2, can provide the meaning behind these dimensions. Analysing the processes associated with the most informative dimension 3 can shed the light on the way the response to PHA-793887 is conveyed in the cell lines. The hidden dimension 3 is associated with eight biological processes, five of them positively (DNA metabolic process, regulation of cellular macromolecule biosynthetic process, RNA metabolic process, rRNA metabolic process, ribonucleoprotein complex assembly) and three of them negatively (regulation of cell migration, regulation of MAPK cascade, regulation of angiogenesis). Since the GSEA analysis is performed using gene expression data, large values of hidden variable 3 (which correspond to a better response) implicitly indicate the over-expression of genes associated with the five positively enriched terms, whereas the over-

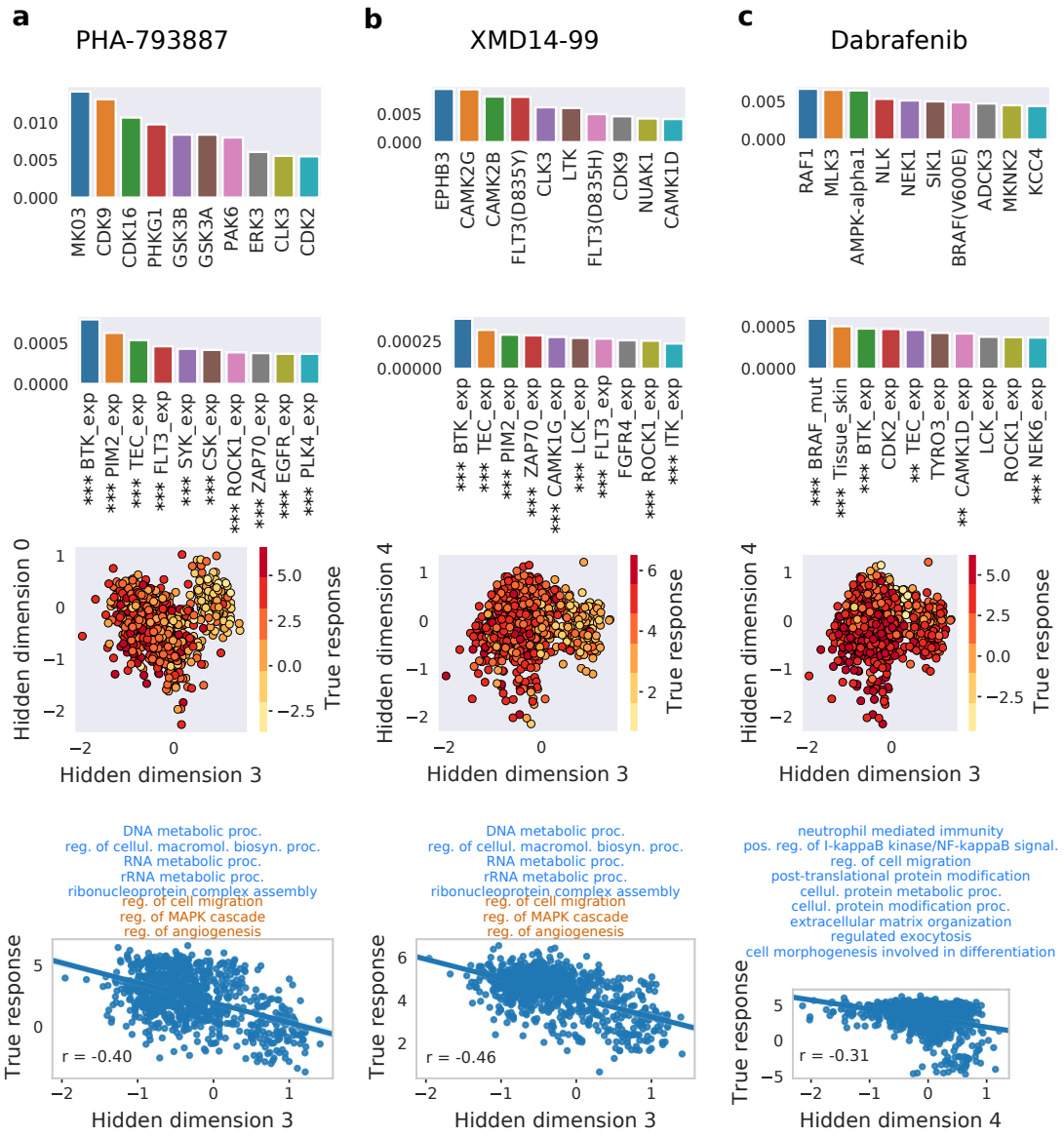


Figure 5.3: **Case studies corresponding to compounds: (a) PHA-793887, (b) XMD14-99 and (c) Dabrafenib.** Top row panels: top ten most important drug features for a response prediction by the model, derived using Integrated Gradients. Second row panels: top ten most important cell lines features for a model's response prediction. Feature names abbreviations: exp. – expression, mut. – mutation. Asterisks indicate the intervals containing the p-values of the Spearman correlation coefficients $r$ between a given feature and log IC50 values for a given drug across screened cell lines: no asterisks – [0.05, 1), * – [0.01, 0.1), ** – [0.001, 0.01), *** – (0, 0.001). Third row panels: cell lines plotted using two chosen hidden dimensions of the cell line hidden space, colored by the true log IC50 values (shown for those cell lines that were screened against the presented drug). Bottom row panels: scatter plots of true log IC50 values w.r.t. hidden dimension most correlated with the response for a given drug. The presented r values are the Spearman correlation coefficients. Text on top shows which GO terms are enriched in a considered hidden dimension, following Fig. 5.2, where blue and brown colors correspond to positive and negative enrichment, respectively. See Fig. 5.2 for term names abbreviations.

expression of genes related to three negatively enriched terms can indicate poor response.

According to GDSC annotations, XMD14-19 targets include ALK, CDK7, LTK and others. The known target LTK is listed as one of the most important drug features (i.e., with a large attribution; Fig. 5.3b, top row panel). Cell line features with top attributions for XMD14-19 strongly overlap with those related to PHA-793887 (Fig. 5.3b, second row panel). In particular, the top three to features are exactly the same (expression of *BTK, PIM2* and *TEC* genes), indicating some similarity between these two drugs. Hidden dimensions 3 and 4 allow to visualize regions in the cell line hidden space with distinctive responses (Fig. 5.3b, third row panel). Similarly to PHA-793887, hidden dimension 3 carries the most information about cell lines response to XMD14-99 and thus we can conclude that the same biological processes may be associated with the response to these two drugs (Fig. 5.3b, bottom row panel).

In the case of Dabrafenib, both drug and cell line feature sensitivities are consistent with its design and clinical usage. Dabrafenib is a selective inhibitor of mutant BRAF kinase, approved by the FDA for the treatment of metastatic melanoma with mutant BRAF(V600) [175, 176]. Accordingly, the two most important cell line features are *BRAF* mutation and skin tissue indicator (Fig. 5.3c, second row panel). The inhibition of BRAF also emerges among the most informative drug features, though being preceded by the inhibition of RAF1, MLK and AMPK (Fig. 5.3c, top row panel). The hidden dimensions 3 and 4 capture significant information about cell lines response (Fig. 5.3c, third row panel), with the hidden dimension 4 being the sole good indicator of the Dabrafenib efficacy (Fig. 5.3c, fourth panel). The hidden dimension 4 has nine positively enriched biological process terms associated with it (Figures 5.2; 5.3c, bottom row panel). Thus, we conclude that a better response to Dabrafenib corresponds to the over-expression of genes involved in: neutrophil mediated immunity, positive regulation of I-kappaB kinase/NF-kappaB signaling, regulation of cell migration, post-translational protein modification, cellular protein metabolic process, cellular protein modification process, extracellular matrix organization, regulated exocytosis, and cell morphogenesis involved in differentiation.

### Associating biological processes to all of the analyzed drugs

In the final step of the interpretability analysis, we associate biological processes to all of the analyzed drugs. This analysis is based on the idea behind the bottom panels of Fig. 5.3. Just like for PHA-793887, XMD14-99 and Dabrafenib, we can calculate the correlation coefficient between the response profile for a given drug and a given hidden dimension across cell lines, for each of 74 drugs and each of 10 hidden dimensions (Fig. 5.1d). This calculation yields a $74 \times 10$ matrix, in which each entry represents a Spearman correlation coefficient for a given compound and hidden dimension. We then utilize the associations between hidden dimensions and biological processes presented in Fig. 5.2 in order to connect drugs to biological processes. For a given drug and process, we first establish which hidden dimension is enriched for that process. Next, we assign a corresponding correlation coefficient between the drug response and the dimension to the drug and process pair. If more than one hidden dimension is enriched for the process, we take the average of the corresponding correlations. This analysis produces a $74 \times 67$ drug-process matrix, where each entry is a correlation coefficient indicating how important a given biological process is for driving the response of a cell line to a given drug. We divide this matrix into five sub-matrices by the main target pathways of the drugs: RTK signaling, PI3K/MTOR signaling, ERK MAPK signaling, Cell cycle, and Others. Finally, we perform the row-wise hierarchical clustering of each such drug-process sub-matrix in order to group drugs by the similarity of processes that drive their efficacy (Fig. 5.4). The obtained

clustermaps clearly indicate such processes that are shared among drugs targeting the same pathways, as well as point at their differences, some of which are related to the particular gene targets.

Five of the drugs targeting the RTK signaling pathway (Fig. 5.4a) are positively correlated with a large group of processes related to DNA replication and cell cycle (DNA replication, DNA-dependent DNA replication, G1/S transition of mitotic cell cycle, cellular macromolecule biosynthetic process, cytoskeleton-dependent cytokinesis, mitochondrial translational elongation, regulation of cell cycle process, translational elongation, negative regulation of transcription from RNA polymerase II promoter, regulation of transcription from RNA polymerase II promoter), from which four drugs are positvely correlated with processes related to transport and sensory perception (ubiquitin-dependent protein catabolic process, processes endosomal transport, cellular protein localization, monovalent inorganic cation transport, sensory perception of chemical stimulus and chemical synaptic transmission). These drugs, however, visibly divide into two distinct groups with respect to a group of processes related to RNA metabolism and regulation of MAPK cascade and angiogenesis (RNA metabolic process, rRNA metabolic process, ribonucleoprotein complex assembly, regulation of MAPK cascade, regulation of angiogenesis). This difference is the reflection of putative targets of the drugs; drugs which have ERBB2 or EGFR as the putative targets are not or are only slightly correlated with these processes, while remaining drugs are strongly negatively correlated with them. In general, the RTK signaling drugs with shared target genes have similar associated processes and cluster together.

Drugs targeting the PI3K/MTOR signaling pathways generally share very similar profiles of association with biological processes (Fig. 5.4b). All but three of the drugs in this category have consistent positive correlation with five processes (epidermis development, cell-cell junction assembly, regulation of autophagy, plasma membrane bounded cell projection assembly and cilium organization), which distinguishes PI3K/MTOR signaling from other drug categories (e.g. ERK MAPK signaling drugs have noticeably different association profiles, Fig. 5.4c). There are, however, still some processes which tend to be correlated only with a subset of drugs. For example, seven drugs (AKT inhibitor VIII, ZSTK474, Omipalisib, Buparlisib, GSK1059615, WYE-125132 and Torin 2) exhibit stronger negative correlation with five previously listed processes related to RNA metabolism and regulation of MAPK cascade and angiogenesis, while others do not. The associations with the foregoing processes seem to be more prevalent in drugs which have the mammalian target of rapamycin (mTOR) kinase among their putative targets in addition to phosphoinositide 3-kinases (PI3Ks), with the exception of Dactolisib. Notably, two drugs presumably targeting solely mTOR (WYE-125132, Torin 2) have very similar association profiles across all 67 processes. Another considerable group of processes with relatively high correlation in the PI3K/MTOR signaling category consists of eight processes: negative regulation of cytokine production, positive regulation of MAPK cascade, positive regulation of intracellular signal transduction, positive regulation of peptidyl-tyrosine phosphorylation, proteolysis, regulation of defense response, regulation of extrinsic apoptotic signaling pathway and regulation of immune response. Two out of three drugs associated with these processes (TGX221 and AZD6482) target solely PI3Kbeta. Associations with the listed eight processes are also noticeable in the cell cycle category (Fig. 5.4d), specifically for CGP-082996, CGP-60474, Seliciclib and Palbociclib, as well as for WZ-1-84 and XMD8-85 in the others category (Fig. 5.4e).

In general, these results can serve as a validation and explanation of the model predictions, as well as provide insights regarding the drugs mechanisms of action and drivers of the cell lines response. Considering drug-biological process associations within a certain drug category enables insights into drugs action on a more general level than putative targets or target

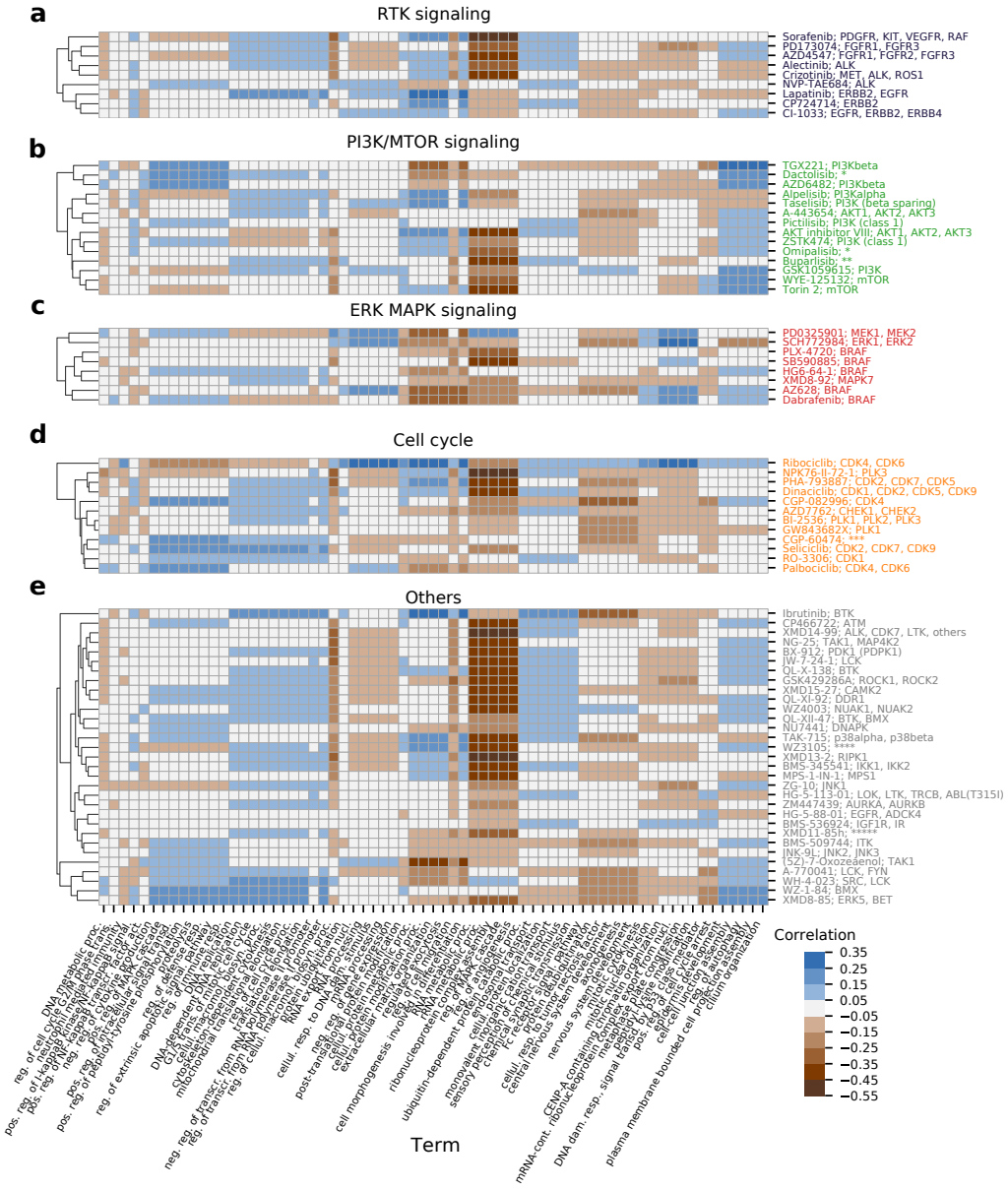


Figure 5.4: **Associations between cell lines biological processes (horizontal axes) and all of the 74 analyzed drugs (vertical axes), plotted separately for (a) RTK signaling, (b) PI3K/MTOR signaling, (c) ERK MAPK signaling, (d) Cell cycle and (e) Others target pathways.** First, for every drug, the Spearman correlation coefficient between every cell line hidden dimension and the response is computed, as shown in Fig. 5.3. These correlations are then assigned to biological processes associated with a given hidden dimension (see Fig. 5.2). If more than one hidden dimension is related to a process, an average of correlation is taken and assigned to the process. Drugs were hierarchically clustered using the Euclidean distance and average linkage within a given target pathway category. The horizontal axis is shared for all panels. The vertical axis tick label is formatted as: Drug Name; Putative Targets. Drugs target pathways and putative targets are taken from GDSC annotations. The labels are color-coded by the target pathway. For some drugs, putative targets have not been listed for readability. Those targets are: * – PI3K (class 1), MTORC1, MTORC2, ** – PI3Kalpha, PI3Kdelta, PI3Kbeta, PI3Kgamma, *** – CDK1,CDK2,CDK5,CDK7,CDK9, PKC, **** – RC, ROCK2, NTRK2, FLT3, IRAK1, others, ***** – BRSK2, FLT4, MARK4, PRKCD, RET, SRPK1. The color scale of correlation heatmaps is the same for all categories. See Fig. 5.2 for term names abbreviations.

pathway information alone.

## 5.4. Discussion

In this work, constituting the second project of this thesis, we propose a deep neural network recommender system-based approach to the problem of kinase inhibitor sensitivity prediction based on side information about drugs and cancer cell lines. The proposed model, DEERS, combines dimensionality reduction of the cell line and drug features using autoencoders and neural network-based prediction based on the obtained hidden representations. The modeled drug features are the strengths of inhibition of kinases by the drugs. The cell line features include expression and mutation calls for the same kinases in cancer cell lines, complemented by primary tissue type of origin for the cell lines. To our knowledge, this type of modeling using these types of input data has not been applied before to predict sensitivity to kinase inhibitors.

Our focus on modeling kinase inhibitors is motivated by the fact that binding profiles across kinases represent exquisite data to characterize such drugs. Alternative information about drugs could be the list of specific known drug targets. In contrast to continuous and rich data about kinase inhibition, however, annotations of known targets are relatively incomplete. The quality of the kinase binding data that we used is assured by a standardized assay platform interrogating a large number of kinases. Therefore, off-target inhibition effects are most likely captured completely. An alternative could be to use information on which signaling pathways are affected by a drug since this information is often provided in drug databases. However, clearly the information about target pathways is only high-level, less detailed than using kinase binding data, and suffers from incomplete understanding about the complete set of pathways that a drug effects in different cellular contexts.

Despite its advantages, the data used to describe the drugs can also be seen as a source of limitation of this study. While usage of continuous inhibition profiles provides more information regarding drugs' action, these types of data are available only for a subset of drugs. In comparison, utilizing raw data (e.g. SMILES) would lead to more labelled samples for training, which could improve the predictive performance. However, using continuous inhibition profiles sheds a light on drugs' mechanisms of action w.r.t. their response on cell lines and provide more interpretable drugs' representations. The ablation study regarding inhibition profiles and putative targets shows that drugs' interactions across a large enough panel of kinases can provide more informative profile of drugs, even in the absence of their putative targets. This provides valuable insight on which properties of the drugs are useful to model in these kind of prediction tasks.

The DEERs model aims at two goals: 1) high predictive performance and 2) outstanding model interpretability. Our analysis constitutes a thorough comparative assessment of model performance, evaluating both traditional and variants of matrix factorization-based methods. Out of the two traditional models, XGBoost achieves better results than Elastic net, indicating that accounting for non-linear interactions among features is crucial for prediction performance. DEERS outperforms the other two matrix factorization-based approaches, Lin MF (basic matrix factorization model) and Autoen MF (a model using autoencoders for dimensionality reduction and a dot-product for combining the reduced data to make prediction). We observe little difference in performance between the Lin MF model and Autoen MF (Tables 2, 1), although, importantly, the latter has a more difficult optimization goal. Indeed, similar to DEERS, Autoen MF reconstructs the input features from the reduced representations. The advantage of DEERS over these two MF-based models is most likely caused by the

incorporation of the feed-forward network which combines the hidden representations, instead of the simple dot product. Compared to the dot product, which only considers element-wise product, these additional feed-forward network layers allow the system to adjust the weights after the data encoding step and to estimate a more complex function that maps from the hidden representations to the response. Importantly, despite the more complex mapping, the hidden dimensions in the DEERS model are still clearly indicative of the true response in some cases. Across all compared models, both DEERS and XGBoost show top and very similar performance. In contrast to XGBoost, however, DEERS is easier to interpret, as it provides highly informative 10-dimensional representations of the input cell lines molecular setup and the drug features.

A model by Manica *et al.* [177] aims at similar objectives as DEERS, and accounts for information about both drugs and cell lines, but is not directly comparable to DEERS, as it bases on different input data and uses a distinct interpretability approach. The best model of Manica *et al.* [177] achieves 0.104 RMSE between scaled predicted and true IC50 for the strict data split, compared to 0.08 obtained by DEERS. Importantly, this result cannot be interpreted in favor of DEERS, since in the strict data split in Manica *et al.* [177] subsets of both drugs and cell lines are left out in the validation set, therefore posing a more challenging problem. The interpretability analysis in the Manica *et al.*[177] model is of the "ante-hoc"-type as it utilizes the neural attention mechanisms . In this way, the interpretability is intrinsically build into the system [178]. Some other examples of attention-based methods and transformers [179] have been successfully applied in related fields [180, 181]. Regarding the drugs this model identifies molecular substructures that are the most responsible for making a prediction for a given drug. This differs from DEERS, since we used the higher-level drugs' features in the form of the inhibition profiles, which are not explicitly connected with the drugs' chemical structure. Both for the Manica *et al.* [177] model and DEERS, external evidence validating the interpretability results can be found.

Extensive interpretability analysis demonstrates that the 10 hidden dimensions of the drug and cell line autoencoders seem to capture the majority of important information for both drugs and cell lines. The results imply mutual independence of hidden dimensions (Fig. 5.2, B1, B2) and also suggest that the hidden representations are representative of the drug and cell line input data. In particular, hidden dimensions 3 and 4 show as most relevant for driving the cell lines drug response for the majority of drugs (Fig. 5.2, 5.4), as demonstrated by the presented examples (Fig. 5.3, bottom panels). The correlation analysis of genome-wide cell lines features and hidden representations, combined with GSEA, helps to provide biological meaning to the hidden dimensions (Fig. 5.2). The same analysis performed using the restricted set of kinase- and tissue type-related 241 cell lines features that are used to train the models would have resulted in the bias towards GO terms or pathways related to protein kinases in general. Instead, using genome-wide gene expression helps to identify the enriched terms which are not influenced by the choice of features in the training data and spanning a broader range of biological processes. Moreover, this methodology is potentially very versatile, as different drugs and cell lines properties outside of the training data can be correlated, and different gene set libraries can be queried for enrichments. We show that combining the influence of distinct hidden dimensions for drug response, and biological processes associated with the hidden dimensions constitutes a framework which can directly explain drug response by concrete biological mechanisms. Such a map facilitates the easier explanation of drugs' mechanisms of action and can potentially identify the new, unexpected ones (Fig. 5.4). Overall, this study shows how data encoding combined with the series of analyses can help to increase the interpretability even in the case of deep neural network recommender models, while maintaining the complex nature of such systems.

This work differs from the first one in several major ways. First, the research problem; while feature selection is related to model's interpretability, in this work we tackle the intepretability problem more directly. Secondly, the whole modeling paradigm; previously we built separate model for each compound and here we performed multi-task learning. Thirdly, the machine learning methodology; while previous work utilized standard ML methods, here we entered a deep learning paradigm and developed our custom model. The choice of compounds' inpute features, namely their inhibition profiles, stemmed from the first work. We have observed that drug targets constitute an important information for drug sensitivity prediction, however, we were also aware of drugs' off-targeting and that putative targets might not be a comprehensive description, hence we decided to use the inhibition profiles and focus the modeling on kinase inhibitors. Overall, this study is rather not an extension of the first one, more a qualitatively different work within the same field, but with very different emphasis, modeling paradigm, and level of technical advancement.

# Chapter 6

# A generative recommender system with GMM prior for cancer drug generation and sensitivity prediction

## 6.1. Background

Kinase inhibitors constitute a very important class of targeted anticancer therapeutics. They are commonly characterized by their *inhibition profiles*, measuring their strength of inhibition across a panel of protein kinases. Despite being a specific category of anticancer compounds, kinase inhibitors can be further grouped into different clusters, for example by their target pathways, or kinase inhibitory activity. Specifically, it has been observed that they differ by the numbers of so called off-targets, i.e. unintentionally inhibited kinases, which is reflected in their inhibitory profiles. While a number of kinase inhibitor drugs is already successfully applied in the clinic, there is a pressing need for novel drug discovery, due to the mechanism of resistance to existing drugs and the large variety of mutations that could be targeted in individual patients' tumors.

Unfortunately, the current pre-clinical process of proposing novel compounds proves inefficient, as the proposed drugs fail further stages of clinical trials, yielding the entire process of novel drug discovery a daunting, time and money consuming task [182, 183, 184]. Deep generative models transform the field of molecule discovery, providing promising synthetic molecules such as proteins or drugs with desired chemical properties [185, 186, 187, 188, 189, 190, 191]. However, these approaches are not directly applicable to kinase inhibitors or other anticancer therapeutics. First, they require large amounts of compounds for training, while the number of known kinase inhibitors is scarce. Second, they do not account for the molecular features of tumors that the drugs are supposed to act on. Specifically, drug sensitivity is a function of both compound's and tumor's features, and it is the relationship between these two features sets that determines the treatment outcome. Although multiple machine learning models were proposed for the prediction of the sensitivity of cancer cell lines to the drugs [62, 13, 59, 92, 70], including recommender system-based approaches [192, 24, 193], these methods lack a generative ability. Therefore, a new class of generative models for kinase inhibitor discovery and simultaneous sensitivity prediction is needed, which would restrict the vast space of potential generative model hypotheses by accounting for and drawing from the wealth of additional experimental data such as kinase inhibitor profiles and their clustering, cell line sensitivity screens, as well as the molecular profiling of the cell lines.

In this work, we propose a novel variational autoencoder (VAE) model for generation of

specific types of anticancer compounds, guided by the clustering of their inhibitory profiles within the GMM prior. The model infers representations of reduced dimensionality of the drugs' SMILES [194] (a string-based representation of molecule's chemical structure), and leverages drug sensitivity screening data, as well as molecular features of the cancer cell lines. The proposed model offers the following key functionalities:

- **clustering of the drugs in the latent space and generation of novel drugs from specific clusters**, here having specific types of inhibitory profiles,

- **inference of latent representations** of the drugs' SMILES and the molecular features of the cell lines,

- **prediction of sensitivity of the cancer cell lines to the drugs**.

On the most general level, the proposed model can be thought of as an extension of a recommender system with side information [195, 196, 197, 198, 199] with a generative model. In a recommender system with side information, objects and users are characterized by vectors of their specific features (the side information), and users assign scores to objects, yielding an object per user matrix. The task of a recommender system is to predict the scores given the side information for the objects and users. Our contributed extension is twofold. First, we assume the objects can be grouped into clusters, and we generate new objects with features that are characteristic of their corresponding clusters. Second, we aim to predict the scores for these synthetic objects and for existing users. In our particular application, in the generative recommender system the objects correspond to drugs from the family of kinase inhibitors, users to cancer cell lines, while the scores correspond to the sensitivity of the cell lines to the drugs, i.e. the ability of the drugs to kill cancer cells. Hence the name of the model, i.e, Variational Autoencoder-based Drug Efficacy Estimation Recommender System (VADEERS).

The majority of existing methods for compound generation focus on generating compounds with specified chemical properties, without taking into account the broader biological context, e.g. efficacy of compounds against cancer cell lines or other cancer models. Recently, Born *et al.* [200] proposed a model aiming at generating compounds which target specific gene expression profiles via a hybrid variational autoencoder acting as compounds generator. In the proposed reinforcement learning paradigm, the compound generator serves as an agent, while the output of a drug sensitivity prediction model serves as a reward function, which allows to train the agent to produce more and more effective compound against a given gene expression profile. Joo *et al.* [201] proposed a conditional VAE, in which generation of new molecular fingerprints is conditioned on drug sensitivity. Related problem was also approached without resolving to probabilistic generative models by efficient Monte Carlo tree search [202]. VAEs have also appeared in the context of drug sensitivity modeling focusing more on the prediction aspect [203], or applied to cancer models, rather than compounds [123, 204].

Recommender systems were traditionally applied in the field of e-commerce, where user decisions are recorded online. Generative adversarial network-based models proved useful in dealing with the lack of explicitly negative samples in such applications [205, 206, 207, 208]. GANs were also used to imitate user behavior dynamics in reinforcement learning-based recommender systems to better approximate the reward function and simulate the environment in this setting [209, 210]. We are not aware of recommender systems equipped with VAE that accounts of a given clustering of the objects.

## 6.2. Methods

### Variational Autoencoders

Recall from Section 3.4 that variational autoencoders [88] can be viewed as neural network-based probabilistic graphical models designed to learn complexed probability distributions. They are latent variable models, where sampling from a prior distribution $p(\mathbf{z})$ in the latent space $Z$ is followed by a sampling from $p_\theta(\mathbf{x}|\mathbf{z})$ in the data space $X$. The probability distribution $p_\theta(\mathbf{x}|\mathbf{z})$ is modeled using a neural network parametrized by $\theta$ called a decoder. Since usually the log marginal likelihood $p(\mathbf{x})$ is intractable, its following evidence lower bound (ELBO) is maximized:

$$\mathcal{L}^{ELBO}_{\phi,\theta}(X) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \tag{6.1}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is the variational approximation of a posterior distribution $p(\mathbf{z}|\mathbf{x})$, and $D_{KL}$ stands for Kullback–Leibler divergence. $q_\phi(\mathbf{z}|\mathbf{x})$ is usually modeled as a Gaussian distribution $\mathcal{N}(\mu(\mathbf{x}), diag(\sigma(\mathbf{x})))$ where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are outputs from a neural network parametrized by $\phi$, called encoder. The first term in Eq. (6.1) is often referred to as the reconstruction term, while the second as the regularization term, as it forces the posterior towards the prior. In case when both $q_\phi(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$, are Gaussian distributions, the Kullback-Leibler divergence has an easy to optimize analytical form [88], which in turn enables an efficient optimization of Eq. (6.1). In a more general case, Eq. (6.1) can be further decomposed into [211]:

$$\mathcal{L}^{ELBO}_{\phi,\theta} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z}) + \ln p(\mathbf{z}) - \ln q_\phi(\mathbf{z}|\mathbf{x})]. \tag{6.2}$$

Since $\ln q_\phi(\mathbf{z}|\mathbf{x})$ is under the expectation over $q_\phi(\mathbf{z}|\mathbf{x})$, this equals:

$$\mathcal{L}^{ELBO}_{\phi,\theta} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\ln p_\theta(\mathbf{x}|\mathbf{z}) + \ln p(\mathbf{z})] + \mathbb{H}[q_\phi(\mathbf{z}|\mathbf{x})], \tag{6.3}$$

where $\mathbb{H}[q_\phi(z|x)]$ denotes the entropy of the posterior. Typically, the expected value in Eq. (6.2) or (6.3) is approximated by sampling $L$ point(s) from $q_\phi(\mathbf{z}|\mathbf{x})$ [211]. In VAEs, the goal is to maximize the ELBO, while when training neural networks in general commonly the goal is to minimize a cost function. Therefore, with $L = 1$, the loss function takes the form of

$$\mathcal{L}_{\phi,\theta}(\mathbf{x}^{(i)}) = -\ln p_\theta(\mathbf{x}^{(\mathbf{i})}|\mathbf{z}^{(i)}) - \ln p(\mathbf{z}^{(i)}) - \mathbb{H}[q_\phi(\mathbf{z}|\mathbf{x}^{(i)})], \tag{6.4}$$

where $\mathbf{x}^{(i)}$ indicates the $i$-th data point and $z^{(i)}$ is a sample from a $q(\mathbf{z}|\mathbf{x}^{(i)})$. In case when a variational posterior approximation is $\mathcal{N}(\mu(\mathbf{x}), diag(\sigma(\mathbf{x})))$, the entropy has an easy to optimize analytical form. In such a case, optimization of Eq. (6.4) is feasible for $\ln p(\mathbf{z}^{(i)})$ that are easy to evaluate.

### GMM VAEs

In VAE, different choices of prior distributions $p(\mathbf{z})$ impose different trade-offs between the optimization simplicity and the complexity of modeled distributions. E.g., the classical choice of the prior to be the standard normal distribution, i.e. $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ carries the simplicity of optimization of $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$, at the cost of not imposing any particular structure on the latent space, nor utilizing any prior knowledge regarding the data. Another popular choice of a prior $p(\mathbf{z})$ is a GMM [212, 213, 214, 215, 216]. In this model, for a given point $\mathbf{z}$ there is a categorical hidden variable $C \sim Cat(\pi_1, \ldots \pi_K)$, defining the component of the mixture for that point. The conditional probability of $\mathbf{z}$ given the component $C = k$, is then

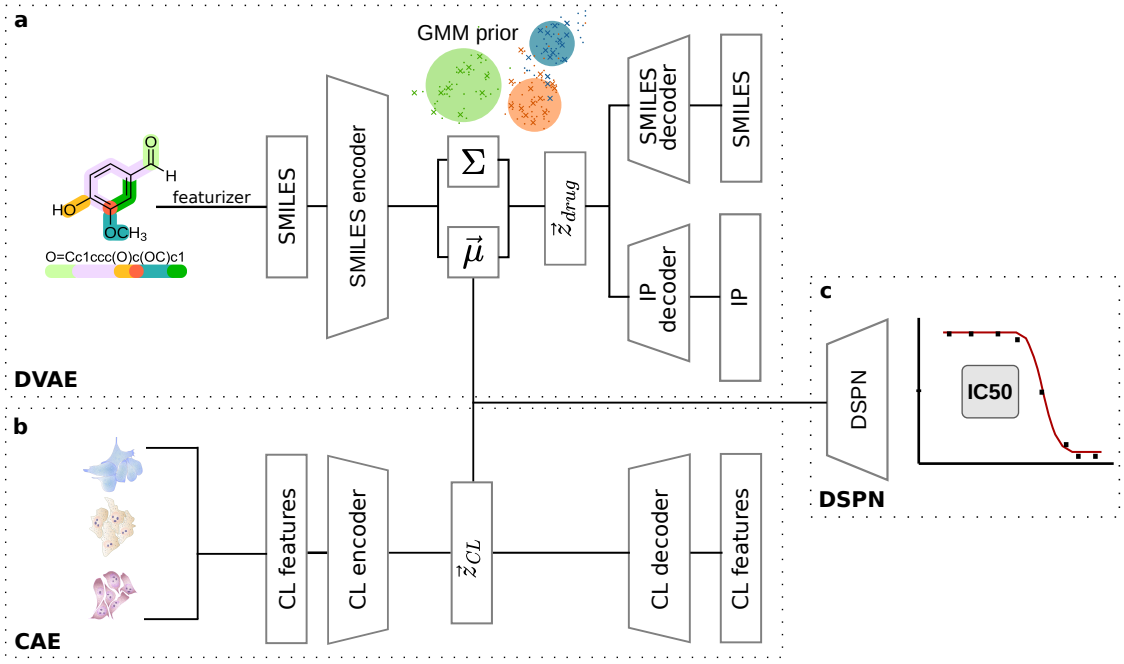


Figure 6.1: **Model's overview.** (**a**) DVAE module. (**b**) CAE module. (**c**) DSPN module. DSPN takes concatenation of DVAE's encoder output, i.e. mean vector, and CAE's latent vector as input.

defined by a Gaussian distribution $\mathcal{N}(\mu_k, \Sigma_k)$, for $k = 1, \ldots, K$. Therefore, in a GMM VAE, the prior for $z$ in Eq. (6.3) is obtained by marginalizing over the values of $C$:

$$p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k, \Sigma_k). \tag{6.5}$$

The closed, analytical form of Eq. (6.5) makes computation of the $\ln p(\mathbf{z})$ term in Eq. (6.3) tractable and enables an efficient optimization of ELBO. Such a GMM prior naturally corresponds to a clustering, where the points from the same component $k$ come from the same Gaussian distribution and thus group together in the latent space.

### Generative recommender system overview

The proposed VADEERS model is a neural network consisting of three major modules: drug variational autoencoder (DVAE) (Fig. 6.1a), cell lines autoencoder (CAE) (Fig. 6.1b), and drug sensitivity prediction network (DSPN) (Fig. 6.1c). The whole model has two inputs, one to DVAE and another to CAE, and four outputs in total. The input to DVAE is a vector representation of a drug's chemical formula, expressed as its SMILES string. DVAE consists of an encoder, which projects the data into a lower-dimensional latent space, and two decoders: the first one outputs the input's reconstruction, while the second predicts the drug's inhibition profile (IP), i.e. the vector of inhibition strengths across a panel of protein kinases. We assume that the drugs can be grouped into clusters by some guiding data that specifies the grouping. Here, the guiding data used are the inhibitory profiles, i.e., drugs with similar inhibitory profiles form distinct clusters (see Section 6.2 and Supplementary Methods). CAE is an autoencoder taking vector representation of a cell line's biological features as input to an encoder, and returning its reconstruction from a decoder.

The use of DVAE and CAE allows to find lower-dimensional, informative data representations of drugs and cell lines, respectively. During the forward pass, the latent representations of a given cell line and a given drug are extracted, concatenated and passed as an input to DSPN, which predicts the numerical value indicating the sensitivity of that cell line to that drug. Here, this value is represented by log half maximal inhibitory concentration (IC50) [61], defined as a drug concentration needed to reduce cell viability by 50%.

### Extension of a GMM VAE model to unknown guiding data and unknown components

We propose an extension to the classical GMM VAE by allowing a semi-supervision of the GMM prior. Specifically, we consider that the categorical variables defining the mixture components for each point in the latent space are partially observed for some of the training data. More formally, the training input data $\mathcal{D}$ with $N$ samples can be divided into two disjoint sets $\mathcal{D}_o$ and $\mathcal{D}_h$, with $\mathcal{D} = \mathcal{D}_o \cup \mathcal{D}_h$. The set $\mathcal{D}_o = \left\{x^{(1)}, \ldots, x^{(n)}\right\}$, where $0 \leq n \leq N$, is a set of samples $x^{(i)}$ for which the component variable $C^{(i)}$ is observed and $C^{(i)} = k^{(i)}$, where $k^{(i)} \in \{1, \ldots, G\}$, for $G \leq K$. We further refer to these observed component values as the *guiding labels*. Note that with $G < K$ some of the assumed components will not appear as a guiding label for any training sample. Such components correspond to additional clusters of samples that truly exist but are never observed. By allowing additional components in the latent space, we still are able to model these clusters in the latent space. In contrast to $\mathcal{D}_o$, the set $\mathcal{D}_h = \left\{x^{(n+1)}, \ldots, x^{(N)}\right\}$, is a set of samples, for which the components are hidden. In such a setting, the latent prior $p(\mathbf{z})$ in Eq. (6.5) is replaced by

$$p(\mathbf{z}) = \begin{cases} \mathcal{N}(\mathbf{z}|\mu_{k^*}, \Sigma_{k^*}) & \text{if } \mathbf{z} \text{ is a sample for input } x^{(i)} \in \mathcal{D}_o \text{ and } C^{(i)} = k^* \\ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k, \Sigma_k) & \text{otherwise,} \end{cases} \tag{6.6}$$

where $k^*$ indicates the particular Gaussian component (cluster), $K$ is the total number of components, and $\pi_k$ is the weight of component $k$. Note that the prior defined by Eq. (6.6) is a generalization of the classical GMM prior defined by Eq. (6.5). Indeed, Eq. (6.6) reduces to Eq. (6.5) in the case when $\mathcal{D}_o = \emptyset$.

The guiding labels can be defined by an independent, given clustering of the samples in some external data space. In such a case, the external data is referred to as *guiding data*.

The described model with the GMM prior is used to implement the DVAE in VADEERS (Fig. 6.1a). Here, the input samples $x^{(i)}$ correspond to SMILES, the guiding data is defined by the inhibitory profiles of the drugs, and the guiding labels by the clusters of these inhibitory profiles. In this way, we transfer a clustering of inhibition profiles to the latent space, as the latent representation of drugs sharing $k^*$ are made to follow the same Gaussian distribution in $Z$.

Importantly, the parameters of GMM, i.e. $(\pi_k, \boldsymbol{\mu}_k, \Sigma_k)$ for $k$ in $\{1, \ldots, K\}$, can be learned via gradient descent together with the remaining parameters of VADEERS, yielding a complete model of the data, including the GMM prior. The use of GMM as a prior then allows to sample $\mathbf{z}$ from a particular of component of $p(\mathbf{z})$ e.g. corresponding to a particular guiding label. At the same time, for the drug $\mathbf{x}$ for which the guiding label is unknown, the cluster assignment is obtained using a posterior inference over $C$ based on the values sampled from $q(\mathbf{z}|\mathbf{x})$. This enables the inference of the guiding label (i.e., the category of inhibition profiles) for every drug.

Recall from section 6.2 that DVAE part of VADEERS has two decoders corresponding to reconstructed compounds' input and compounds' inhibition profiles (IP). Combining Eq. (6.4)

with the fact that decoders' outputs are continuous, we define the DVAE's loss function for a single compound as:

$$\mathcal{L}_{DVAE}(\mathbf{x}_S{}^{(i)}, \mathbf{x}_S{}^{(i)'}, \mathbf{x}_I{}^{(i)}, \mathbf{x}_I{}^{(i)'}, \mathbf{z}^{(i)}) = r_S \cdot \mathrm{MSE}(\mathbf{x}_S{}^{(i)}, \mathbf{x}_S{}^{(i)'}) + r_I \cdot \mathrm{MSE}(\mathbf{x}_I{}^{(i)}, \mathbf{x}_I{}^{(i)'}) \quad (6.7)$$

$$- r_P \cdot \ln p_\lambda(\mathbf{z}^{(i)}) - r_E \cdot \mathbb{H}[q_\phi(\mathbf{z}|\mathbf{x}_S{}^{(i)})], \quad (6.8)$$

where $\mathbf{x}_S{}^{(i)}$ and $\mathbf{x}_S{}^{(i)'}$ are the $i$-th compound's SMILES representation and it's reconstruction, respectively, $\mathbf{x}_I{}^{(i)}$ and $\mathbf{x}_I{}^{(i)'}$ are true and predicted inhibition profiles, respectively, $\mathbf{z}^{(i)}$ is a $\mathbf{z}$ sample corresponding to the $i$-th compound, $r_S$ is the positive real-valued weight corresponding to the compounds' input reconstruction error, MSE denotes mean squared error, $r_I$ is the weight of the IP prediction error, $r_P$ is the weight corresponding to the prior likelihood, $r_E$ is the weight of encoder's entropy, and the last term corresponds to the entropy of latent variables returned by the encoder.

In the case of CAE, the loss function is given simply by the reconstruction error between the cell lines' input and its reconstruction:

$$\mathcal{L}_{CAE}(\mathbf{x}_B{}^{(j)}, \mathbf{x}_B{}^{(j)'}) = \mathrm{MSE}(\mathbf{x}_B{}^{(j)}, \mathbf{x}_B{}^{(j)'}), \quad (6.9)$$

where $\mathbf{x}_B{}^{(i)}$ and $\mathbf{x}_B{}^{(i)'}$ are cell line's input features and their reconstruction, respectively. Finally, the loss corresponding to DSPN is the error between continuous true and predicted IC50 values defined for compound-cell line pair:

$$\mathcal{L}_{DSPN}(y^{(i,j)}, \hat{y}^{(i,j)}) = MSE(y^{(i,j)}, \hat{y}^{(i,j)}), \quad (6.10)$$

where $y^{(i,j)}$ and $\hat{y}^{(i,j)}$ is a true and predicted IC50 corresponding to $i$th compound and $j$th cell line, respectively. The loss for the whole model is the weighted sum of above expressions:

$$\mathcal{L}_{Model}(\cdot) = \mathcal{L}_{DVAE}(\cdot) + r_{CAE} \cdot \mathcal{L}_{CAE}(\cdot) + r_{DSPN} \cdot \mathcal{L}_{DSPN}(\cdot), \quad (6.11)$$

where $r_{CAE}$ and $r_{DSPN}$ are weights corresponding to CAE and DSPN errors, respectively (arguments are replaced by $\cdot$ for simplicity). The formulation with the vector $r$ of weights allows to change model's emphasis by controlling these hyperparameters.

## Dataset

The analyzed dataset $\mathcal{D} = \{\mathbf{X}_S, \mathbf{X}_I, \mathbf{X}_B, \mathbf{Y}_R, \mathbf{y}_G\}$ consists of five parts, where $\mathbf{X}_S \in \mathbb{R}^{304\times300}$ denotes drugs' SMILES vector representations, $\mathbf{X}_I \in \mathbb{R}^{117\times294}$ denotes drugs' inihibition profiles across a panel of protein kinases, $\mathbf{X}_B \in \mathbb{R}^{922\times241}$ denotes a matrix of cell lines biological features, $\mathbf{Y}_R \in \mathbb{R}^{922\times304}$ denotes a matrix with drug response indicators for a given cell line $c$ and drug $d$, and $\mathbf{y}_G \in \mathbb{R}^{117}$ denotes a vector of guiding labels for a subset of considered drugs (see below).

The primary source of drug sensitivity data for cell lines was the Genomics of Drug Sensitivity in Cancer (GDSC) database [138, 140]. The set of 304 compounds in $\mathbf{X}_S$ extracted from GDSC database were represented by their chemical structure indicated by corresponding SMILES strings. In order to convert SMILES strings into numerical vector representations, we used the pre-trained Mol2vec model [75] treating it as SMILEs featurizer which produces 300-dimensonial vectors of continuous values. These representations served as an input to DVAE.

Another considered features of compounds were their inhibition profiles, i.e., their binding strengths across a panel of 294 protein kinases. Such inhibitory profiles were available and

extracted for 117 compounds from the HMS LINCS KINOMEscan data resource [157]. The value for a given compound-kinase pair represents a percent of control, where a value of 100% means no inhibition of kinase binding to its ligand in the presence of the compound, and where low value means a strong inhibition [158, 159].

Data to characterize the 922 cell lines were downloaded from the GDSC. For the molecular features of the cell lines, we considered only the genes coding for kinases present in the KINOMEscan dataset, as well as subset of putative gene targets of considered compounds. This resulted in a set of 202 genes, for which mRNA expression levels (202 features) and binary mutation calls (21 features) were extracted for all cell lines. Furthermore, the dummy-encoded tissue type was added, producing additional 18 binary features, yielding the final set of 241 biological features for 922 cell lines.

For the drug response indicators in $\mathbf{Y}_R$ we used the log half maximal inhibitory concentration (IC50) values from GDSC. For a given compound-cell line pair, IC50 is defined as a drug concentration needed to reduce cell viability by 50%. Note that some values in $\mathbf{Y}_R$ were missing since not every cell line is screened against every available compound.

In principle, guiding labels in $\mathbf{y}_G$ could be any discrete class assignments for compounds. In this case, we utilized inhibition profiles from $\mathbf{X}_I$ to assign compounds to their functional categories. To this end, compounds were clustered according to their inhibition profiles using K-means, with the number of clusters set to $G = 3$. Cluster assignments resulting from this approach were then used as the guiding labels for the 117 compounds with known inhibition profiles.

### Experimental setup, VADEERS model architecture, training and implementation

The validation and test sets were constructed by randomly selecting two sets of 100 unique cell lines each. We then extracted the compound-cell line datapoints containing selected cell lines and used them to construct validation and test sets, while the rest of the pairs corresponding to the remaining 722 unique cell lines constituted the training set.

The hyperparameters of the model were empirically determined using the validation set. The encoders for both DVAE and CAE were fully-connected forward networks with two hidden layers with 128 and 64 neurons, respectively. All of the decoders followed a similar architecture, but with 64 neurons in a first hidden layer and 128 in a second. The latent space dimensionality in both DVAE and CAE was set to 10.

The DSPN was a fully-connected network with three hidden layers with 512, 256 and 128 neurons, respectively, and an output layer outputting an IC50 prediction. Dropout with $p = 0.5$ was applied at the first and second layer. ReLu activation function was used throughout the whole model.

Model training was performed on 200 epochs using the Adam optimizer [83] with a learning rate of 0.005 and batch size of 128. The whole model was trained together for the first 150 epochs, after which, DVAE and CAE were frozen and DSPN alone was trained for another 50 epochs with a newly set learning rate of 0.001, decreasing by a factor of 0.1 with every 10 epochs. In addition, every 1000 training steps there was a break devoted to only training DVAE part. During each break, DVAE was trained for 100 epochs using only compounds with known inhibition profiles, with the batch size of 8. For the purpose of experiments, the $r$ loss function weights were all set to 1. Both the number of unique guiding labels and components in GMM prior were set to 3, i.e, $G = K = 3$.

The neural networks related code was implemented using Python 3.8.8, PyTorch 1.10.0 and PyTorch Lightning 1.5.0. K-means clustering for the guiding data was implemented using

scikit-learn 1.0.1 [125].

## 6.3. Results

We evaluated three versions of the proposed model, differing by the way the DVAE module was implemented: i) a classical VAE with the standard normal prior ("Vanilla VAE"), ii), the DVAE as described in Section 6.2 (with GMM prior and loss function given by Eq. (6.7) and (6.11), however, only weights $\pi_k$'s and components' means $\mu_k$'s were the trainable parameters of the GMM prior (Eq. (6.6)), while components' covariance matrices $\Sigma_k$'s set to be identity matrices ("GMM VAE constrained"), iii) the DVAE was as described in Section 3.2., in its least constrained version, where all parameters of the GMM, including $\Sigma_k$'s, were trainable ("GMM VAE unconstrained").

| DVAE version | IC50 RMSE | IC50 Pearson | IP RMSE |
|---|---|---|---|
| Vanilla VAE | $1.33 \pm 0.022$ | $0.87 \pm 0.006$ | $1.13 \pm 0.109$ |
| GMM VAE constrained | $1.33 \pm 0.023$ | $0.87 \pm 0.006$ | $1.09 \pm 0.062$ |
| GMM VAE unconstrained | $1.34 \pm 0.012$ | $0.87 \pm 0.004$ | $1.04 \pm 0.030$ |

Table 6.1: **IC50 and IP prediction performance for VADEERS with different versions of the DVAE module.**

### Predictive performance

The predictive performance of the IC50 estimation was assessed by calculating the root mean squared error (RMSE) and Pearson correlation between the true and predicted IC50 values. In addition, we computed the RMSE between the true and predicted inhibition profiles, corresponding to the second decoder in DVAE (Table 6.1). This procedure was repeated five times with different random data splits (see Methods).

Despite the large differences in the complexity of the prior distribution, the three model versions perform almost equally well in terms of IC50 prediction. This suggest that models achieved the limit of predictive performance for this particular dataset. Although the optimization of the IC50 prediction was not the main goal of this study, the low RMSE and high correlation indicate that VADEERS correctly captures drug and cell line features and reliably predicts the sensitivity of unseen cancer cell lines to kinase inhibitor drugs.

In contrast to IC50, the three model versions do differ in terms of the ability to reconstruct inhibition profiles (Table 1), measured by RMSE between true and reconstructed IPs. The best result in that regard is achieved by the GMM VAE unconstrained model (RMSE = 1.04). Such a result is expected, as lower constraints on the latent space representations make it easier for the model to optimize this metric. However, note that in the described setup the IP RMSE metric was computed for the training data, therefore it should rather be interpreted as model's ability to converge w.r.t. this particular metric than model's predictive performance. Still, this result suggests that the nature of the latent space is important with regard to the decoding, in the sense that this task is not entirely dependent on the decoder alone and is improved by imposing a latent space's structure.

### Latent space structure

Figure 6.2 compares the structures of the latent spaces of the three considered models. It is clear that in both GMM VAE models the clustering defined by the guiding labels is preserved

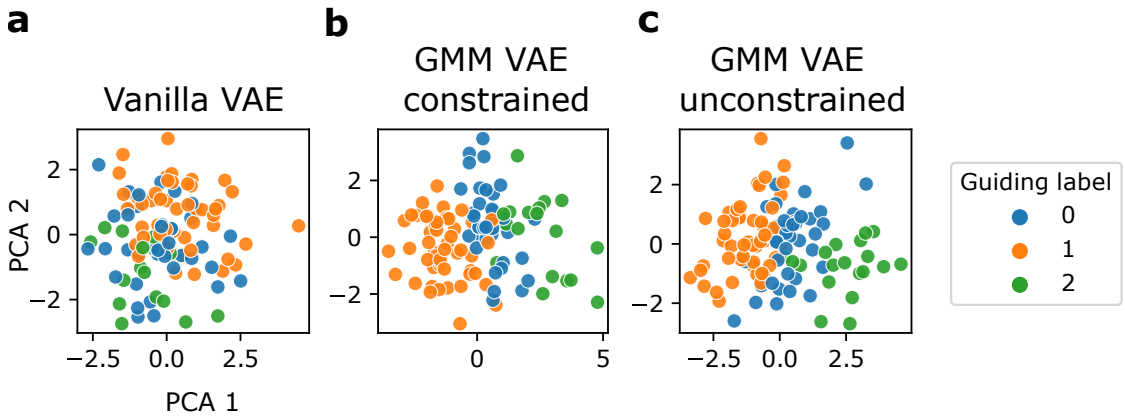


Figure 6.2: **Latent spaces of the three VADEERS model versions, differing by the DVAE subnetwork.** For each model, the latent representations of the 117 compounds with guiding labels (see Section 6.2) are obtained by passing their SMILES representations to the model's DVAE's encoder. The encoder's output for a given model is then plotted in 2D using principal components analysis (PCA) and colored with the corresponding guiding labels. Results for the first of five random data splits.

in the latent space (Fig. 6.2 b, c), i.e. points with the same guiding label are grouped together. By visual assessment, the clusters are the most clearly separated for the GMM VAE unconstrained model (Fig. 6.2c). This is also reflected by the corresponding Silhoutte scores that are much higher for GMM models than for Vanilla VAE, with the highest one obtained by GMM VAE constrained (Fig. 6.3a). Interestingly, the latent clustering is to some extent preserved also for the Vanilla VAE model version (Fig. 6.2a). This suggests that the sole presence of the IP decoder encourages compounds with similar IPs to group together. However, the use of the GMM prior imposes that explicitly. Most importantly, the GMM prior defines the clusters of latent drug representations by associating them to the GMM components, with each guiding label obtaining its separate cluster. In this way, first, we are able to assign a label to a new drug by finding its latent representation and component. Second, we are able to generate new drugs with a pre-specified guiding label.

### Generative performance

In a VAE, new data points can be generated by first sampling from the latent prior and then passing each sample $\mathbf{z}$ to a decoder. The use of a GMM instead of a normal prior allows to perform this process more precisely; the sample $\mathbf{z}$ can be obtained from a given, $k$th Gaussian component, which should reflect the actual properties of the compounds in the guiding data space. This is the case in this study, where the guiding labels stem from the clustering in the space of the drugs' IPs. In Fig. 6.4, we verify this hypothesis by visualizing the IPs of the generated samples.

The IPs generated by the Vanilla VAE model do not form any particular clusters (Fig. 6.4b). In contrast, the samples generated from both GMM VAEs clearly confirm that the above assumption is correct; samples generated from different components are also distinguishable after decoding, i.e. the information regarding the latent component is preserved in the true data space (Fig. 6.4c, d). Interestingly, not only the grouping of the data points into clusters, but even the actual mutual spatial arrangement of those clusters is preserved between the true IPs (Fig. 6.4a), latent space (Fig. 6.2 b, c), and the generated IPs (Fig.

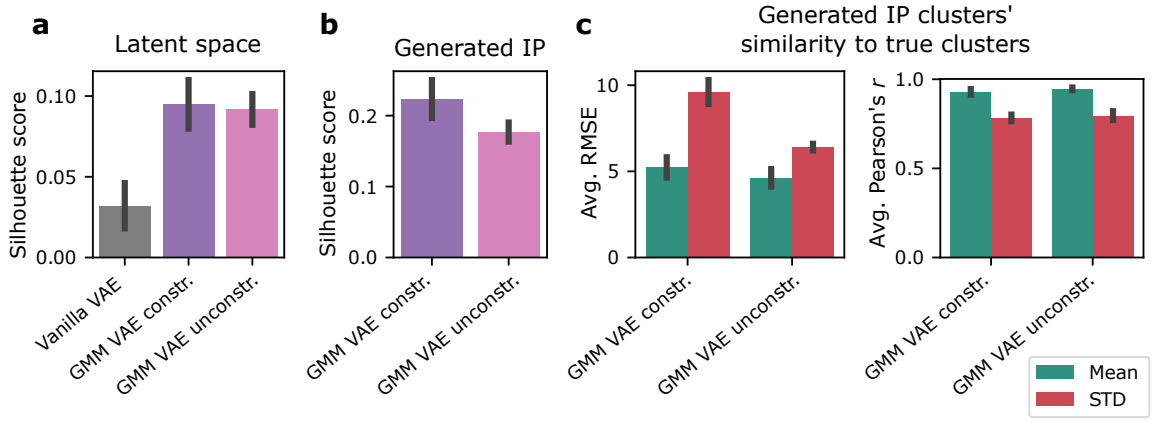


Figure 6.3: **Numerical assessment of models' generative performance.** All presented metrics are averaged over five experimental runs, with error bars corresponding to standard deviations across experimental runs. (**a**) Silhouette score of latent representations, with compounds' guiding labels as compounds' true clusters. (**b**) Silhouette scores of generated inhibition profiles, with GMM components from which samples are drawn as true clusters. (**c**) Average RMSE (left panel) and Pearson correlation (right panel) between true and generated feature-wise, within-cluster IPs' statistics, shown for cluster means (centroids) and STDs. Average metrics are taken by first averaging over all three clusters and next over the experimental runs.

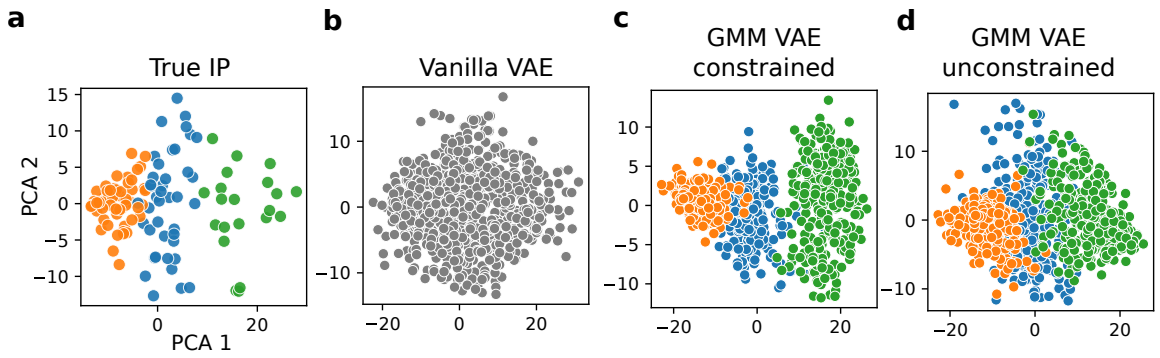


Figure 6.4: **True and generated inhibition profiles visualized in 2D.** (**a**) The true IPs for the 117 available drugs. (**b**) 900 IPs generated from the Vanilla VAE. (**c**) IPs generated from the GMM VAE constrained model. 300 samples are drawn. (**d**) IPs generated from the GMM VAE unconstrained model. Again, 300 samples are drawn per-component. Colors correspond to guiding label or a corresponding GMM component, see Fig. 6.2 for legend.

6.4c, d). However, note that the points are visualized after PCA. Fixing GMM components' $\Sigma_k$ to $I$ impacts the generated IPs; GMM VAE constrained produces more concise and better-separated clusters than unconstrained (Fig. 6.4c, d), which is also reflected by corresponding Silhouette scores (Fig. 6.3).

The similarities between the true and generated IPs can also be shown without resolving to dimensionality reduction methods by computing per-cluster, feature-wise statistics such as mean or standard deviation (STD) (Fig. 6.5), where feature-wise means can be thought of as clusters' centroids. The IPs generated by both variants of GMM VAE exhibit an excellent concordance with true data in terms of cluster means (Fig. 6.5a). This is apparent on both absolute values level and correlation across features within a given cluster. For example, for a true data, cluster 2 in general corresponds to relatively high inhibition, but for some kinases (features) inhibition is low, and low inhibition of exact same kinases is observed for the generated data. While differences between GMM VAE constrained and unconstrained in terms of generated IP centroids are hard to assess visually, this is not the case for the within-cluster, feature-wise STDs (Fig. 6.5b). Again, the effect of fixing GMM components' STD is visible; for the constrained model, within-cluster STDs are much smaller compared to the true ones. This also demonstrates that IP decoder has relatively low variance; namely, it is unable to compensate for the low variance of samples from $p(\mathbf{z})$ in order to bring the generated data's STD closer to the true one. In case of the unconstrained model, the STD is higher and closer to the true one. This clearly demonstrates that in the absence of constraints, the learned components' STDs are larger, and more closely resemble the true data. Indeed, for this particular model, the average value in the covariance matrices $\Sigma_k$ is 9.25. These differences are also clear when assessed by computing the RMSE between true and generated IPs' STD averaged across all three components (Fig. 6.3c).

In essence, numerical results support all the observations made based on visual assessment. The quality of clustering in the latent space is the highest for GMM VAE constrained model (mean Silhouette score across five experimental runs 0.095), followed by GMM VAE unconstrained and the lowest value obtained by Vanilla VAE (Fig. 6.3a). Similarly, GMM VAE constrained obtains better clustering quality of generated IPs than the unconstrained version (Silhouette scores 0.223 and 0.177, respectively, Fig. 6.3b).

In contrast, GMM VAE constrained obtains slightly worse results than unconstrained in terms of within-cluster statistics similarity between true and generated IPs (6.3c). Both constrained and unconstrained models are similarly close to the original data in terms of cluster centroids (average RMSE between true and generated centroids across three clusters of 5.228 and 4.627, respectively), especially in terms of correlation (average Pearson correlation 0.928 and 0.947, respectively). As in Fig. 6.5b), the differences are more noticeable when considering within-cluster STDs; both models achieve similar correlation (0.783 and 0.796, respectively), but GMM VAE constrained is worse than unconstrained w.r.t. RMSE (9.602 for constrained and 6.407 for unconstrained).

## 6.4. Discussion

In this work, we propose VADEERS, a multi-task generative recommender system for drug sensitivity prediction. The generative part of the model, DVAE, is a variational autoencoder with two decoders and a GMM latent prior. The latent GMM is optimized using guiding labels in order to reflect a given external clustering. This allows to sample data points from a cluster of interest, i.e., having specific desired features. Hence DVAE, along with other modules, forms a comprehensive model of drugs' and cell lines' properties and interactions,

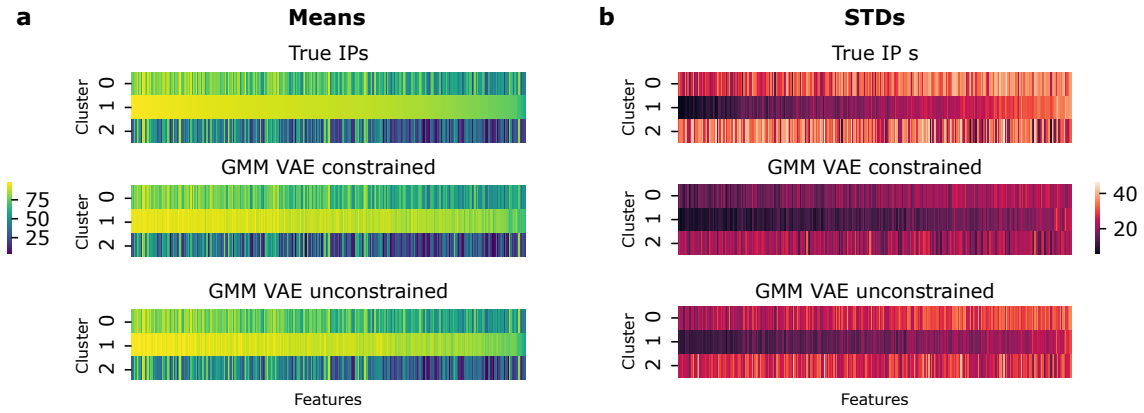


Figure 6.5: **True and generated IPs' feature-wise, within-cluster (a) means and (b) STDs.**

with a guided generative ability.

One of the limitations of the proposed model is its inability to generate data points with totally arbitrary features. Namely, the model allows to generate new data points with properties that strictly reflect the clustering observed in the training data. In principle, this could be bypassed by performing various operations on multiple generated data points, however, testing this hypothesis was not in the scope of this analysis. Another important limitation corresponds to the analyzed data; a different choice of data for drugs' representations (e.g. representing SMILES strings as graphs) and guiding data might be more suitable for generating molecule candidates, which, at least in theory, could be synthesized. Both above aspects are directions of future work regarding this study.

This work introduces several key concepts important for drug sensitivity modeling and compound generation. Still, the proposed model, and more specifically, GMM VAE with semi-supervised clustering with guiding labels, is generic and not limited only to modeling compounds. The notion of optimizing latent space with guiding labels can potentially be beneficial and improve the performance of generative models also in other applications. Moreover, the proposed model offers additional functionality not exploited in this study. For example, setting the number of Gaussian components $K$ greater than number of unique labels $G$ might lead to identification of novel subgroups of samples, not limited to the original choice of guiding labels.

Since VADEERS integrates several sources of data on compounds, cell lines, and drug sensitivity, it can provide important insights regarding these modalities. For example, the usage of the guiding labels and subsequent sampling from a given Gaussian component helps to identify the relationship between variables used as guiding labels and the variables which are decoded by DVAE, i.e, the SMILES embeddings and IPs. The model comes with the added bonus of the drug sensitivity estimation for these samples for a given cell line or a panel of cell lines.

In principle, any work on models involving generation of compounds with given properties, including the one presented here, can potentially be used to generate or help to generate harmful chemical agents, such as highly addictive or toxic compounds. In addition, it should be stated that any compound candidate or drug sensitivity estimation originating from an *in silico* model, including the one presented here, is not properly validated in experimental nor clinical setting and should not be directly acted upon without such a validation.

The crucial difference between this work and two previous ones is the new field it enters, namely the generative modeling, here in the context of chemical compounds. As mentioned in

the introduction to this chapter, we felt that methods for compounds' generation in the biological context has not been extensively studied, which was the motivation to enter the topic of generative modeling. From the technical standpoint, VADEERS is the most natural extension of the previous DEERS model with the generative component; essentially drug autoencoder was replaced by drug variational autoencoder with two decoders. However, there are other major changes; VADEERS model aims also at predicting drugs' inhibition profiles from raw data, namely SMILES strings, rather than receiving them as input. This adds another dimension to model's multi-task nature; multi-tasking not only refers to multiple compounds themselves but also multiple compounds' properties. Although the model is formulated in a general way and is adaptable, the motivation to employ GMM and latent space clustering was driven by application; previous works showed high variability among compounds and the primary goal was to capture that variability in a structured way via clustering. Overall; VADEERS constitutes a major extension to a DEERS model with several new modeling aspects. It is also the most comprehensive and technically advanced model from all presented in the thesis.

# Chapter 7

# Summary

The thesis builds upon three research projects in a broad topic of computational drug sensitivity prediction. They consider several major research questions at the intersection of general machine learning and this specific application. Specifically, the topics undertaken in the thesis include multi-task learning, model intepretability, representation learning and generative modeling.

The work presented in Chapter 4 focuses on choosing appropriate feature selection approach for modeling drug sensitivity. In this work, a separate model with different feature selection was developed for each compound, resulting in 2484 distinct models. Although using existing, standard machine learning algorithms, the study gives several important insights about drug sensitivity modeling, not limited only to feature selection. In the context of the thesis, this work can be viewed as the one focusing on exploring and understanding of the nature of the analyzed data, rather than technical modeling advancement, yet still providing useful guidelines for anti-cancer drug sensitivity modeling. Some of the observations committed in the study also directly influenced the subsequent research directions.

Chapter 5 describes the Drug Efficacy Estimation Recommender System, or DEERS model. This work differs from the one from Chapter 4 in several major ways. Rather than building separate model for every drug, DEERS is a multi-task (in terms of compounds) model, which makes the sensitivity prediction as a function of both compound's and cell line's features. The choice of compound's features for this study, namely inhibition profiles across panel of protein kinases, was partially influenced by observations committed in Chapter 4. Furthermore, kinase inhibitors constitute perhaps the most prominent class of anti-cancer targeted therapeutics, which makes them an interesting object of research, but, to our knowledge were not a part of any systematic assessment before. Another major difference is a shift from conventional machine learning to deep learning. Besides the model itself, DEERS work also introduces a novel interpretability approach which enables connecting cell lines' latent space dimensions to biological processes, which is not limited only to this particular network architecture. Subsequently, it allowed to connect drug action directly to biological mechanisms in the cell lines, creating a map which can provide clues about why particular compounds work against particular cell line. The model is also more technically advanced, as several custom methodologies have been introduced. In general, this work constitutes a coherent continuation of the drug sensitivity prediction problem, but with qualitatively different set of tools and important shifts in emphasis.

Finally, VADEERS, the model depicted in Chapter 6, can be viewed as a natural extension of the DEERS model into a topic of generative modeling. The motivation for VADEERS was to build a drug sensitivity prediction model which is not only a discriminator. Namely,

it operates on probability distributions rather than single data points and has an ability to generate new objects. Given the potential application, it is more natural to apply generative modeling to compounds, rather than cell lines. From the technical standpoint, the most natural way to extend DEERS by a generative component is to replace the drug autoencoder with a variational autoencoder. However, two other major changes are made. First, the input compound's features are replaced by their SMILES strings and inhibition profiles become the target for prediction by one of two decoders. Secondly, a GMM prior with semi-supervised guiding is introduced, which enables clustering of the data representations in the latent space according to external clustering in an independent data space, referred to as a guiding data space. This not only imposes a structure on the latent space, increasing the model's interpretability, but, more importantly, enables to sample from a particular Gaussian component of a prior mixture and generate new samples with defined features that correspond to that component. Results show that the relationships between data points in a guiding data space are indeed preserved in the latent space. Moreover, the features of generated, new data points show very good concordance with the real, observed ones. In summary, VADEERS constitutes the most comprehensive, multitask model presented here, which accounts for multiple data modalities and which possesses both discriminative and generative abilities.

In summary, this thesis presents a comprehensive and cohesive research related to a single, but wide topic. Each of the presented works answer a particular subset of the posed research challenges. Overall, this thesis can be seen as a comprehensive work on computational drug sensitivity prediction, showing multiple research problems associated with the topic and the means to address them.

# Appendix A

# Chapter 4 Supplementary Material

## A.1. Supplementary Methods

### A.1.1. Elastic net regression

Elastic net regression belongs to the family of regularized linear regression models where the target value is expected to be a noisy linear combination of input features. The model introduces the regularization through adding a combination of $\ell_1$ and $\ell_2$ norms of it's coefficients to the loss function. Therefore, elastic net's optimization problem can be represented as:

$$\min_w \frac{1}{n}\|Xw - y\|_2^2 + \alpha\rho\|w\|_1 + \alpha(1-\rho)\|w\|_2^2$$

where $n$ is the number of samples, $X$ and $w$ represent the input data and coefficients vector, respectively. The amount and type of regularization are controlled by hyperparameters $\alpha$ and $\rho$, corresponding to *alpha* and *l1_ratio* arguments in scikit-learn implementation [125]. These two parameters where tuned during cross-validation.

### A.1.2. Random forest regression

Random forest is an ensemble method, which works by combining the outputs of several decision trees in order to make final predictions. In contrast to linear regression, decision trees is a non-parametric method which learns simple decision rules inferred from the data features. It's goal is to partition the input feature space such that samples with similar labels are grouped together. At each node $m$, corresponding data $Q$ is split into two subsets:

$$\begin{aligned} Q_{left}(\theta) &= (x,y)|x_j < t_m \\ Q_{right}(\theta) &= Q \setminus Q_{left} \end{aligned}$$

where $\theta$ is a candidate split consisting of a feature $j$ and corresponding threshold $t_m$ and $(x, y)$ represents training samples. Decision trees select parameters $j$ and $t_m$ which minimize the impurity of resulting subsets. The choice of a specific impurity function depends on application. In our analysis, we used mean squared error, which is common for regression tasks.

Decision trees have many advantages, but are also prone to create over-complex graphs which tend to overfit the data. In random forest, each tree is built from the bootstrap sample from the training set. Furthermore, the best split at each node is picked from a random subset of features. Such randomness, combined with averaging the predictions of single trees, helps to decrease the variance of the overall model. The hyperparameters we tuned when using

random forests included (following scikit-learn notation): *n_ estimators* – number of trees in the ensemble, *max_ features* – maximum number of features considered when splitting the data, *max_ depth* – maximum depth of the trees, *min_ samples_ split* – minimum number of samples required to perform the split and *min_ samples_ leaf* – minimum number of samples allowed in a leaf node.

### A.1.3. Stability selection with lasso regression.

Stability selection [112] works by generating $N$ bootstrap samples of available data and using an underlying feature selection algorithm (in this case lasso regression) to determine which features are relevant for a given sample. For every generated sample, it fits the selection algorithm with a specified value of regularization parameter $\lambda$, which produces a selection set $\hat{S}_i^{\lambda}$ indicating which features to choose. Given selection sets from each sample, the empirical probability of choosing a particular feature $k$ can be computed as:

$$\hat{\Pi}_k^{\lambda} = \Pr[k \in \hat{S}^{\lambda}] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\{k \in \hat{S}_i^{\lambda}\}}$$

i.e. counting the number of times $k$ occurred as an important component for in the samples. This process is then repeated for several values of $\lambda$. The final stable set of relevant features can be then constructed as follows:

$$\hat{S}^{\text{stable}} = \{k : \max_{\lambda \in \Lambda} \hat{\Pi}_k^{\lambda} \geq \pi_{\text{thr}}\}$$

where $\Lambda$ is a set of all $\lambda$ values and $\pi_{\text{thr}}$ is a predefined probability threshold. In our work, we used scikit-learn compatible implementation of stability selection [217] combined with lasso regression, fitting for $N = 100$ samples with five different values of $\lambda$: $10^{-5}, 10^{-4}, 5 \cdot 10^{-4}, 10^{-3}$ and $10^{-2}$.

When applying automated stability selection, we first fitted the model using five different values of $\lambda$ and 100 bootstrap samples, which resulted in stability scores corresponding to every feature. We then iterated over predefined range (0, 1) of stability thresholds $\pi_{\text{thr}}$ with 0.025 increment, performing the whole modeling process with a corresponding number of features at each iteration using elastic net regression. This procedure was repeated for five random data splits. In order to establish the single best stability threshold for every compound, we averaged the results over data splits. The performance metrics used to evaluate the model were then the averages of metrics achieved with the chosen best threshold for every data split.

### A.1.4. Feature importance derived from random forest

In a single decision tree, the depth of a feature used as a decision node represents the relative importance of that feature when predicting the target variable. Features present at the top of a tree contribute to the final prediction result for a bigger fraction of samples. The importance of a particular feature is also associated with the decrease of impurity when splitting the data using that feature (i.e. the bigger the importance, the bigger decrease in impurity measure). Therefore, the corresponding impurity decrease can be used to estimate the feature importance in a single tree [218]. In random forests, this predictive ability of a given feature can be averaged over several trees to define a new metric, *Mean Decrease Impurity* (MDI) [219], which provides the feature importance estimate with reduced variance.

When using random forest for feature selection, after data extraction and hyperparameter tuning steps, we first trained the algorithm on the whole training set and extracted a vector

with values representing the importance of all features. We then ranged over a grid of values $k$, each time performing the whole modeling process using random forests regression with $k$ most important features and recording the corresponding results. Similarly, as in the stability selection setting, one best value of $k$ was chosen by averaging the results over five data splits, and the corresponding performance metrics for best found $k$ were averaged in order to evaluate the model.

## A.2. Supplementary Figures

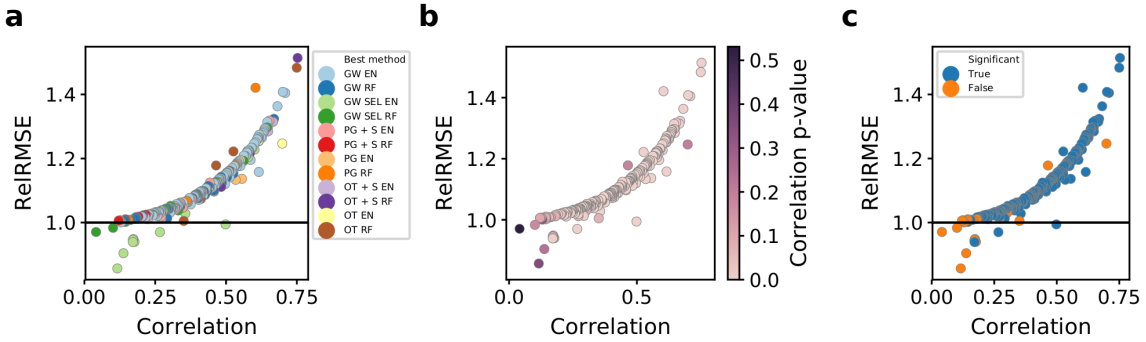


Figure A.1: **P-values of achieved correlations with the test set, calculated based on Student's t-distribution.** (**a**) RelRMSE vs. correlation obtained by the best model for a given drug (copy of Fig. 3c from the main manuscript for reference). (**b**) Same plot as in panel **a**, colored by the corresponding correlation p-value. (**c**) Same plot as in panel **a**, with corresponding correlation p-values classified into significant and non-significant categories at 0.05 confidence level. See Fig. 1 in the main manuscript for model abbreviations.

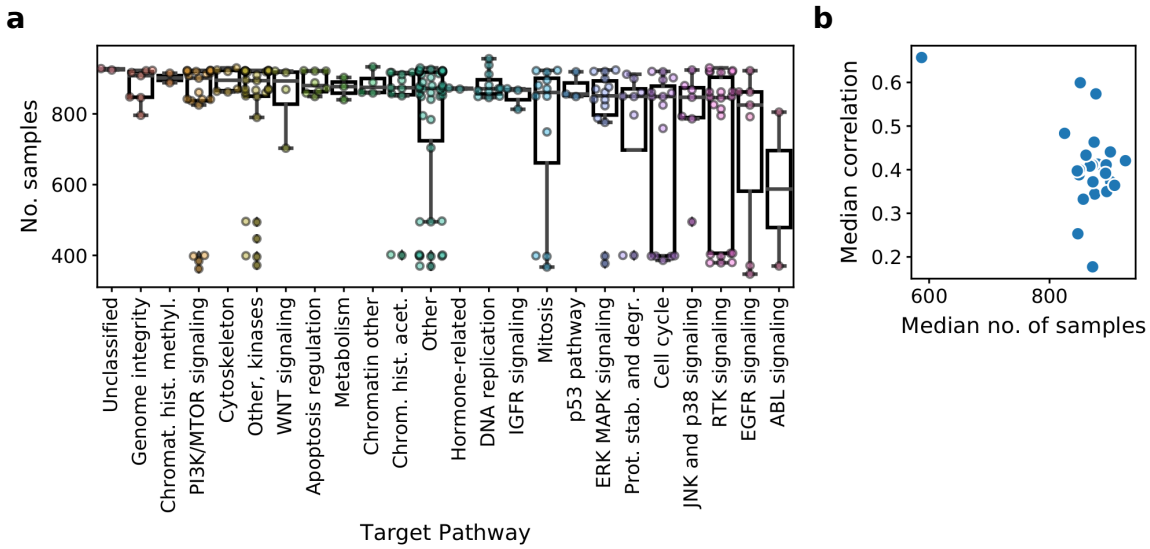


Figure A.2: **Data availability and modeling perfromance grouped by target pathways of the drugs.** (**a**) Number of available samples per drugs belonging to a specific target pathway. Target pathways are sorted by median of per-drug samples. The median values are similar across target pathways (excluding ABL signaling pathway), however, with some pathways exhibiting significant spread. (**b**) Median modeling performance versus median number of per-drug available samples across drugs belonging to given pathways (each point represents a specific target pathway).

# Appendix B

# Chapter 5 Supplementary Material

## B.1. Supplementary Figures

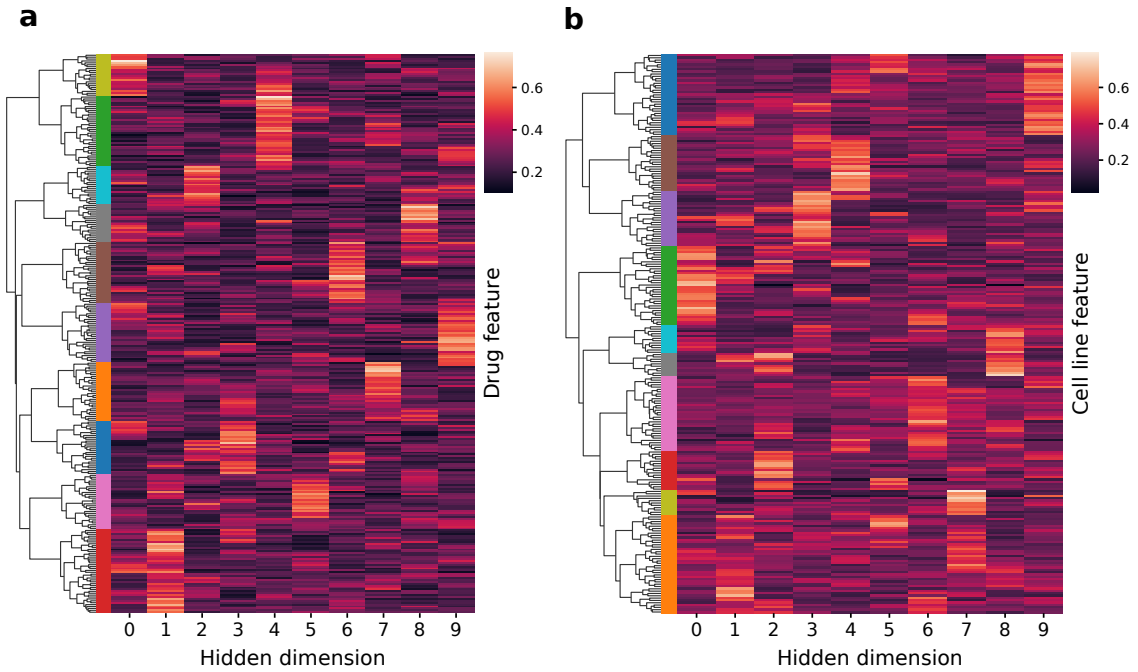


Figure B.1: **Clustermaps of attribution scores between input features and hidden dimensions for (a) drug autoencoder and (b) cell line autoencoder.** Color reflects the importance of a given feature for a given hidden dimension. The vertical color bars next to the dendrograms represent the cluster assignment of groups of features.

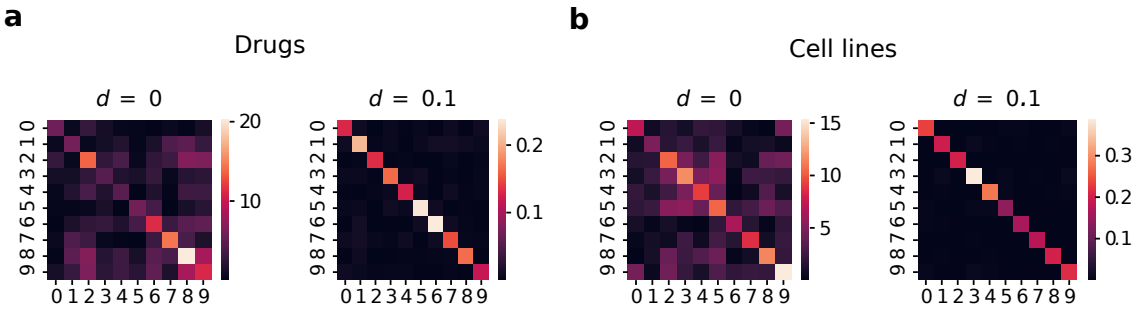


Figure B.2: **Effect of dependence penalty $d$ shown for (a) drugs and (b) cell lines.** First, the drug and cell line data are passed to drug and cell line autoencoders of the DEERS model, respectively, obtaining 10-dimensional hidden representations of all drugs and cell lines used in the analysis. The covariance matrices are calculated across drugs and cell lines in the hidden space. The displayed covariance matrices correspond to the case where DEERS was trained without depencence penalty ($d = 0$), and with dependence penalty term $d = 0.1$.

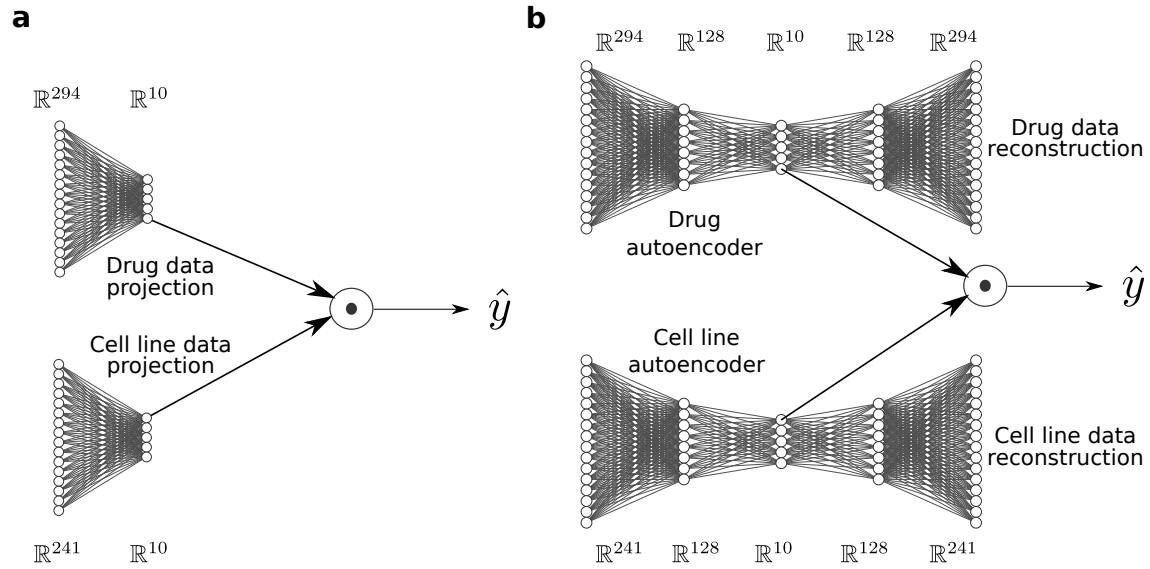


Figure B.3: **Architecture of the models used for comparison. (a)** Linear matrix factorization model (Lin MF). First, drug and cell line data are linearly projected to 10-dimensional hidden representations. Prediction of the response of a cell line to a drug $\hat{y}$ is obtained by applying the dot product to the corresponding hidden representations. **(b)** Non-linear extension of the basic linear model (Autoen MF). The dimensionality reduction is performed via autoencoders with one hidden layer. Prediction of the response of a cell line to a drug is again obtained by taking the dot product of the corresponding, 10-dimensional hidden representations. Drug and cell line data reconstruction errors are also included in the optimization goal.

# Bibliography


[1] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray. Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. `https://gco.iarc.fr/today`. Accessed October 2021.

[2] National Cancer Institute, Cancer Statistics. `https://www.cancer.gov/about-cancer/understanding/statistics`. Accessed October 2021.

[3] National Cancer Institute, The Definition of Cancer. `https://www.cancer.gov/about-cancer/understanding/what-is-cancer`. Accessed October 2021.

[4] National Cancer Institute, Definition of metastasis. `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/metastasis`. Accessed October 2021.

[5] Hanna Dillekås, Michael Rogers, and Oddbjørn Straume. Are 90% of deaths from cancer caused by metastases? *Cancer Medicine*, 8, 08 2019.

[6] Thomas N. Seyfried and Leanne C. Huysentruyt. On the origin of cancer metastasis. *Critical reviews in oncogenesis*, 18 1-2:43–73, 2013.

[7] Xiangming Guan. Cancer metastases: challenges and opportunities. *Acta Pharmaceutica Sinica B*, 5(5):402–418, 2015.

[8] Eoghan Malone, Marc Oliva Bernal, Peter Sabatini, Tracy Stockley, and Lillian Siu. Molecular profiling for precision cancer therapies. *Genome Medicine*, 12, 01 2020.

[9] Seung Shin, Ann Bode, and Zigang Dong. Precision medicine: the foundation of future cancer therapeutics. *npj Precision Oncology*, 1, 12 2017.

[10] qi Zhang, Qihan Fu, Xueli Bai, and Tingbo Liang. Molecular profiling–based precision medicine in cancer: A review of current evidence and challenges. *Frontiers in Oncology*, 10:532403, 10 2020.

[11] National Cancer Institute, Definition of precision medicine. `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/precision-medicine`. Accessed October 2021.

[12] National Cancer Institute, Definition of next-generation sequencing. `https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/next-generation-sequencing`. Accessed October 2021.

[13] George Adam, Ladislav Rampášek, Zhaleh Safikhani, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Machine learning approaches to drug response prediction: challenges and recent progress. *npj Precision Oncology*, 4, 12 2020.

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

[15] Nicolas Stransky, Mahmoud Ghandi, Gregory V. Kryukov, Levi Garraway, Joseph Lehar, Manway Liu, Dmitriy Sonkin, Audrey Kauffmann, Kavitha Venkatesan, Elena J. Edelman, Markus Riester, Jordi Barretina, Giordano Caponigro, Robert Schlegel, William Sellers, Frank Stegmeier, Michael Morrissey, Arnaud Amzallag, Iulian Pruteanu-Malinici, and Julio Saez-Rodriguez. Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, 528, 11 2015.

[16] Jean-Pierre Gillet, Sudhir Varma, and Michael M. Gottesman. The Clinical Relevance of Cancer Cell Lines. *JNCI: Journal of the National Cancer Institute*, 105(7):452–458, 02 2013.

[17] Jean-Pierre Gillet, Anna Maria Calcagno, Sudhir Varma, Miguel Marino, Lisa J. Green, Meena I. Vora, Chirayu Patel, Josiah N. Orina, Tatiana A. Eliseeva, Vineet Singal, Raji Padmanabhan, Ben Davidson, Ram Ganapathi, Anil K. Sood, Bo R. Rueda, Suresh V. Ambudkar, and Michael M. Gottesman. Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *Proceedings of the National Academy of Sciences*, 108(46):18708–18713, 2011.

[18] 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, Kevin Hu, Alexander Y. Andreev-Drakhlin, Jaegil Kim, Julian M. Hess, Brian J. Haas, François Aguet, Barbara A. Weir, Michael V Rothberg, Brenton R. Paolella, Michael S. Lawrence, Rehan Akbani, Yiling Lu, Hong L. Tiv, Prafulla C. Gokhale, Antoine de Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph Lehár, Joshua M. Korn, Dale A. Porter, Michael D. Jones, Javad Golji, Giordano Caponigro, Jordan E. Taylor, Caitlin M. Dunning, Amanda L. Creech, Allison Warren, James M. McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imieliński, Yosef E. Maruvka, Andrew D. Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D. Jaffe, Andrew A. Lane, David M. Weinstock, Cory M. Johannessen, Michael P. Morrissey, Frank P. Stegmeier, Robert Schlegel, William C. Hahn, Gad Getz, Gordon B. Mills, Jesse S. Boehm, Todd R. Golub, Levi A. Garraway, and William R. Sellers. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569:503–508, 2019.

[19] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 10 2018.

[20] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017.

[21] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2018.

[22] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[23] Krzysztof Koras, Dilafruz Juraeva, Julian Kreis, Johanna Mazur, Eike Staub, and Ewa Szczurek. Feature selection strategies for drug sensitivity prediction. *Scientific Reports*, 10:9377, 2020.

[24] Krzysztof Koras, Ewa Kizling, Dilafruz Juraeva, Eike Staub, and Ewa Szczurek. Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines. *Scientific reports*, 11:15993, 2021.

[25] Krzysztof Koras, Marcin Możejko, Paulina Szymczak, Eike Staub, and Ewa Szczurek. A generative recommender system with gmm prior for cancer drug generation and sensitivity prediction, 2022.

[26] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[27] David Weininger, Arthur Weininger, and Joseph L. Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.*, 29:97–101, 1989.

[28] David Weininger. Smiles, 3. depict. graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.*, 30:237–243, 1990.

[29] Cooper GM. The cell: A molecular approach. 2nd edition. the development and causes of cancer. `https://www.ncbi.nlm.nih.gov/books/NBK9963/`. Accessed December 2021.

[30] Momna Hejmadi. *Introduction to Cancer Biology*. Ventus Publishing, 2 edition, 2013.

[31] Mel Greaves and Carlo Maley. Clonal evolution in cancer. *Nature*, 481:306–13, 01 2012.

[32] National Cancer Institute, Definition of proto-oncogene. `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/proto-oncogene`. Accessed October 2021.

[33] National Cancer Institute, Definition of oncogene. `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/oncogene`. Accessed October 2021.

[34] National Cancer Institute, Definition of tumor suppressor gene. `https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/tumor-suppressor-gene`. Accessed October 2021.

[35] Evan Baugh, Hua Ke, Arnold Levine, Richard Bonneau, and Chang Chan. Why are there hotspot mutations in the tp53 gene in human cancers? *Cell Death & Differentiation*, 25, 11 2017.

[36] Gaoyang Zhu, Chaoyun Pan, Jin-Xin Bei, Bo Li, Chen Liang, Yang Xu, and Xuemei Fu. Mutant p53 in cancer progression and targeted therapies. *Frontiers in Oncology*, 10, 2020.

[37] Napoleone Ferrara. Vascular Endothelial Growth Factor: Basic Science and Clinical Progress. *Endocrine Reviews*, 25(4):581–611, 08 2004.

[38] Peter Carmeliet. Carmeliet, p. vegf as a key mediator of angiogenesis in cancer. oncology 69, 4-10. *Oncology*, 69 Suppl 3:4–10, 02 2005.

[39] National cancer institute, types of cancer treatment. `https://www.cancer.gov/about-cancer/treatment/types`. Accessed December 2021.

[40] American cancer society, treatment types. `https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types.html`. Accessed December 2021.

[41] American cancer society, chemotherapy. `https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/chemotherapy.html`. Accessed December 2021.

[42] Volker Schirrmacher. From chemotherapy to biological therapy: A review of novel concepts to reduce the side effects of systemic cancer treatment (review). *International journal of oncology*, 54(2):407–419, 2018.

[43] American cancer society, how chemotherapy drugs work. `https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/chemotherapy/how-chemotherapy-drugs-work.html`. Accessed December 2021.

[44] American cancer society, hormone therapy. `https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/hormone-therapy.html`. Accessed December 2021.

[45] American cancer society, how immunotherapy is used to treat cancer. `https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/immunotherapy/what-is-immunotherapy.html`. Accessed December 2021.

[46] National cancer institute, targeted cancer therapies. `https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies/targeted-therapies-fact-sheetl`. Accessed December 2021.

[47] Wabel AL-Busairi and Maitham Khajah. The principles behind targeted therapy for cancer treatment. In Ahmed Lasfar and Karine Cohen-Solal, editors, *Tumor Progression and Metastasis*, chapter 8. IntechOpen, Rijeka, 2020.

[48] American cancer society, how targeted therapies are used to treat cancer. `https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/targeted-therapy/what-is.html`. Accessed December 2021.

[49] National cancer institute, targeted therapy to treat cancer. `https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies`. Accessed December 2021.

[50] Lei Zhong, Yueshan Li, Liang Xiong, Wenjing Wang, Ming Wu, Ting Yuan, Wei Yang, Chenyu Tian, Zhuang Miao, Wang Tianqi, and Shengyong Yang. Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduction and Targeted Therapy*, 6:201, 05 2021.

[51] Won Duk Joo, Irene Visintin, and Gil Mor. Targeted cancer therapy – are the days of systemic chemotherapy numbered? *Maturitas*, 76(4):308–314, 2013.

[52] Alexander Levitzki and Shoshana Klein. Signal transduction therapy of cancer. *Molecular Aspects of Medicine*, 31(4):287–329, 2010. Signal transduction therapy of cancer.

[53] Ana Nunes and Christina Annunziata. Proteasome inhibitors: structure and function. *Seminars in Oncology*, 44, 04 2018.

[54] K. B., Naiara Orrego-Lagarón, Eileen Mcgowan, Indu Parmar, Amitabh Jha, Basil Hubbard, and H P Vasantha Rupasinghe. Kinase-targeted cancer therapies: Progress, challenges and future directions. *Molecular Cancer*, 17, 02 2018.

[55] Radhamani Kannaiyan and Daruka Mahadevan. A comprehensive review of protein kinase inhibitors for cancer therapy. *Expert Review of Anticancer Therapy*, 18(12):1249–1270, 2018.

[56] LiverTox: Clinical and Research Information on Drug-Induced Liver Injury. Protein Kinase Inhibitors. `https://www.ncbi.nlm.nih.gov/books/NBK548591/`, 2012. Accessed December 2021.

[57] Robert Roskoski. Properties of fda-approved small molecule protein kinase inhibitors: A 2020 update. *Pharmacological Research*, 152:104609, 2020.

[58] American cancer society, targeted therapy side effects. `https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/targeted-therapy/side-effects.html`. Accessed December 2021.

[59] Farzaneh Firoozbakht, Behnam Yousefi, and Benno Schwikowski. An overview of machine learning methods for monotherapy drug response prediction. *Briefings in Bioinformatics*, 10 2021.

[60] Ranadip Pal. Chapter 2: Data characterization. In Ranadip Pal, editor, *Predictive Modeling of Drug Sensitivity*, pages 15–43. Academic Press, 2017.

[61] Daniel Vis, Lorenzo Bombardelli, Howard Lightfoot, Francesco Iorio, Mathew Garnett, and Lodewyk Wessels. Multilevel models improve precision and speed of ic 50 estimates. *Pharmacogenomics*, 17, 05 2016.

[62] Francisco Azuaje. Computational models for predicting drug responses in cancer research. *Briefings in Bioinformatics*, 18(5):820–829, 07 2016.

[63] Robert Shoemaker. Robert, h. s. the nci60 human tumour cell line anticancer drug screen. nat. rev. cancer 6, 813-823. *Nature reviews. Cancer*, 6:813–23, 11 2006.

[64] 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, 10 2017.

[65] Cyril Benes, Daniel A. Haber, Dave Beare, Elena J. Edelman, Howard Lightfoot, I. Richard Thompson, James A. Smith, Jorge Soares, Michael R. Stratton, Nidhi Bindal, P. Andrew Futreal, Patricia Greninger, Simon Forbes, Sridhar Ramaswamy, Wanjuan Yang, 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, 11 2012.

[66] National Cancer Institute, Definition of whole exome sequencing. https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/whole-exome-sequencing. Accessed October 2021.

[67] Eleanor G. Seaby, Reuben J. Pengelly, and Sarah Ennis. Exome sequencing explained: a practical guide to its clinical application. *Briefings in Functional Genomics*, 15(5):374–384, 12 2015.

[68] Wenqian Zhang, Ying Yu, Falk Hertwig, Jean Thierry-Mieg, Wenwei Zhang, Danielle Thierry-Mieg, Jian Wang, Cesare Furlanello, Viswanath Devanarayan, Jie Cheng, Youping Deng, Barbara Hero, Huixiao Hong, Meiwen Jia, Li Li, Simon Lin, Yuri Nikolsky, André Oberthuer, Tao Qing, and Matthias Fischer. Comparison of rna-seq and microarray-based models for clinical endpoint prediction. *Genome Biology*, 16:133, 06 2015.

[69] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20, 07 2019.

[70] James C. Costello, Laura 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, 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*, 32:1202–1212, 2014.

[71] David S Wishart, Yannick Djoumbou, An Chi Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, and Michael Wilson. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic acids research*, 46, 11 2017.

[72] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, 11 2020.

[73] Daniel Rudmann. On-target and off-target-based toxicologic effects. *Toxicologic pathology*, 41, 10 2012.

[74] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

[75] Sabrina Jaeger-Honz, Simone Fulle, and Samo Turk. Mol2vec: Unsupervised machine learning approach with chemical intuition. *Journal of Chemical Information and Modeling*, 58, 12 2017.

[76] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020.

[77] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[78] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Always learning. Pearson, 2016.

[79] Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[80] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[81] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[82] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[83] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[84] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Adv. Neural Inf. Process. Sys*, 2, 04 2002.

[85] Thomas Lucas. *Deep generative models : over-generalisation and mode-dropping*. PhD thesis, 09 2020.

[86] Carl Doersch. Tutorial on variational autoencoders, 2021.

[87] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.

[88] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[89] Fleur Weeber, Salo N. Ooft, Krijn K. Dijkstra, and Emile E. Voest. Tumor organoids as a pre-clinical cancer model for drug discovery. *Cell Chemical Biology*, 24(9):1092–1100, 2017.

[90] E. Izumchenko, K. Paz, D. Ciznadija, I. Sloma, A. Katz, D. Vasquez-Dunddel, I. Ben-Zvi, J. Stebbing, W. McGuire, W. Harris, R. Maki, A. Gaya, A. Bedi, S. Zacharoulis, R. Ravi, L.H. Wexler, M.O. Hoque, C. Rodriguez-Galindo, H. Pass, N. Peled, A. Davies, R. Morris, M. Hidalgo, and D. Sidransky. Patient-derived xenografts effectively capture responses to oncology therapy in a heterogeneous cohort of patients with solid tumors. *Annals of Oncology*, 28(10):2595–2605, 2017. Vemurafenib in patients with BRAFV600 mutation–positive metastatic melanoma.

[91] Marina Salvadores, Francisco Fuster-Tormo, and Fran Supek. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Science Advances*, 6(27):eaba1862, 2020.

[92] In Sock Jang, Elias Chaibub Neto, Juistin Guinney, Stephen Friend, and Adam Margolin. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 19:63–74, 01 2014.

[93] Jinyu Chen and Louxin Zhang. A survey and systematic assessment of computational methods for drug response prediction. *Briefings in Bioinformatics*, 22(1):232–246, 01 2020.

[94] P. L. Bedard, A. R. Hansen, M. J. Ratain, and L. L. Siu. Tumour heterogeneity in the clinic. *Nature*, 501(7467):355–364, Sep 2013.

[95] J Barretina, Giordano Caponigro, N Stransky, Kavitha Venkatesan, Adam Margolin, Sunghyok Kim, C.J. Wilson, Joseph Lehar, G.V. Kryukov, D Sonkin, A Reddy, M Liu, L Murray, M.F. Berger, J.E. Monahan, Paula Keskula, J Meltzer, A Korejwa, J Jane-Valbuena, and M de Silva. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity (vol 483, pg 603, 2012). *Nature*, 492:290–290, 01 2012.

[96] Matthew Rees, Brinton Seashore-Ludlow, Jaime Cheah, Drew J Adams, Edmund V Price, Shubhroz Gill, Sarah Javaid, Matthew E Coletti, Victor Jones, Nicole Bodycombe, Christian Soule, Benjamin Alexander, Ava Li, Philip Montgomery, Joanne D Kotz, Cindy Hon, Benito Munoz, Ted Liefeld, Vlado Dančík, and Stuart L Schreiber. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature chemical biology*, 12, 12 2015.

[97] Brinton Seashore-Ludlow, Matthew G. Rees, Jaime H. Cheah, Murat Cokol, Edmund V. Price, Matthew E. Coletti, Victor Jones, Nicole E. Bodycombe, Christian K. Soule, Joshua Gould, Benjamin Alexander, Ava Li, Philip Montgomery, Mathias J. Wawer, Nurdan Kuru, Joanne D. Kotz, C. Suk-Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dančík, Joshua A. Bittker, Michelle Palmer, James E. Bradner, Alykhan F. Shamji, Paul A. Clemons, and Stuart L. Schreiber. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery*, 5(11):1210–1223, 2015.

[98] Amrita Basu, Nicole E. Bodycombe, Jaime H. Cheah, Edmund V. Price, Ke Su Liu, Giannina Ines Schaefer, Richard Y Ebright, Michelle Lynn Stewart, Daisuke Ito, Stephanie W. Wang, Abigail L. Bracha, Ted Liefeld, Mathias Wawer, Joshua C. Gilbert, Andrew J Wilson, Nicolas Stransky, Gregory V. Kryukov, Vlado Dančík, Jordi G Barretina, Levi A. Garraway, Chung chau. Hon, B Robin Muñoz, Joshua A. Bittker, Brent R. Stockwell, Dineo Khabele, Andrew M Stern, Paul A. Clemons, Alykhan F. Shamji, and Stuart L. Schreiber. An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules. *Cell*, 154:1151–1161, 2013.

[99] Mehreen Ali and Tero Aittokallio. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical Reviews*, 11(1):31–39, February 2019.

[100] Michael Menden, Francesco Iorio, Mathew Garnett, Ultan Mcdermott, Cyril Benes, Pedro Ballester, and Julio Saez-Rodriguez. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PloS one*, 8:e61318, 04 2013.

[101] Trish P. Tran, Edison Ong, Andrew P. Hodges, Giovanni Paternostro, and Carlo Piermarocchi. Prediction of kinase inhibitor response using activity profiling, in vitro screening, and elastic net regression. *BMC Systems Biology*, 8(1):74, Jun 2014.

[102] Zuoli Dong, Naiqian Zhang, Chun Li, Haiyun Wang, Yun Fang, Jun Wang, and Xiaoqi Zheng. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC cancer*, 15:489, 06 2015.

[103] Gregory Riddick, Hua Song, Susie Ahn, Jennifer Walling, Diego Borges-Rivera, Wenjun Zhang, and Howard A. Fine. Predicting in vitro drug sensitivity using Random Forests. *Bioinformatics*, 27 2:220–4, 2011.

[104] H. Yuan, I. Paskov, H. Paskov, A. J. Gonzalez, and C. S. Leslie. Multitask learning improves prediction of cancer drug sensitivity. *Sci Rep*, 6:31619, 08 2016.

[105] Anna Cichonska, Tapio Pahikkala, Sandor Szedmak, Heli Julkunen, Antti Airola, Markus Heinonen, Tero Aittokallio, and Juho Rousu. Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics*, 34(13):i509–i518, 06 2018.

[106] Muhammad Ammad-ud din, Suleiman A. Khan, Disha Malani, Astrid Murumägi, Olli Kallioniemi, Tero Aittokallio, and Samuel Kaski. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics*, 32(17):i455–i463, 08 2016.

[107] Muhammad Ammad-ud din, Suleiman Khan, Krister Wennerberg, and Tero Aittokallio. Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. *Bioinformatics (Oxford, England)*, 33:i359–i368, 07 2017.

[108] Mi Yang, Jaak Simm, Chi Chung Lam, Pooya Zakeri, Gerard J. P. van Westen, Yves Moreau, and Julio Saez-Rodriguez. Linking drug target and pathway activation for effective therapy using multi-task learning. *Scientific Reports*, 8, 12 2018.

[109] Xiaolu Xu, Hong Gu, Yang Wang, Jia Wang, and Pan Qin. Autoencoder Based Feature Selection Method for Classification of Anticancer Drug Response. *Frontiers in Genetics*, 10:233, 03 2019.

[110] Isabelle Guyon and Andre Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

[111] Utkarsh Mahadeo Khaire and R. Dhanalakshmi. Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences*, 2019.

[112] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[113] P. Geeleher, N. J. Cox, and R. S. Huang. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.*, 15(3):R47, Mar 2014.

[114] Samir Amin, Wai-Ki Yip, Stephane Minvielle, A Broijl, Yi Li, Bret Hanlon, David Swanson, Parantu K. Shah, Philippe Moreau, Bronno van der Holt, Mark van Duin, Florence Magrangeas, P Sonneveld P, Kenneth C Anderson, Cheng Li, Herve Avet-Loiseau, and Nikhil C Munshi. Gene Expression Profile Alone Is Inadequate In Predicting Complete Response In Multiple Myeloma. *Leukemia*, 28, 04 2014.

[115] Isidro Cortes, G.J.P. Van Westen, Guillaume Bouvier, Michael Nilges, John Overington, Andreas Bender, and Thérèse Malliavin. Improved Large-Scale Prediction of Growth Inhibition Patterns on the NCI60 Cancer Cell-Line Panel. *Bioinformatics*, pages 1–11, 01 2015.

[116] Delora Baptista, Pedro G Ferreira, and Miguel Rocha. Deep learning for drug response prediction in cancer. *Briefings in Bioinformatics*, 01 2020. bbz171.

[117] Theodore Sakellaropoulos, Konstantinos Vougas, Sonali Narang, Filippos Koinis, Athanassios Kotsinas, Alexander Polyzos, Tyler J. Moss, Sarina Piha-Paul, Hua Zhou, Eleni Kardala, Eleni Damianidou, Leonidas G. Alexopoulos, Iannis Aifantis, Paul A. Townsend, Mihalis I. Panayiotidis, Petros Sfikakis, Jiri Bartek, Rebecca C. Fitzgerald, Dimitris Thanos, Kenna R. [Mills Shaw], Russell Petty, Aristotelis Tsirigos, and Vassilis G. Gorgoulis. A deep learning framework for predicting response to therapy in cancer. *Cell Reports*, 29(11):3367 – 3373.e4, 2019.

[118] Fangfang Xia, Maulik Shukla, Thomas Brettin, Cristina Garcia-Cardona, Judith Cohn, Jonathan Allen, Sergei Maslov, Susan Holbeck, James Doroshow, Yvonne Evrard, Eric Stahlberg, and Rick Stevens. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics*, 19, 12 2018.

[119] Yoosup Chang, Hyejin Park, Hyun-Jin Yang, Seungju Lee, Kwee Yum Lee, Tae 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. *Scientific Reports*, 8, 12 2018.

[120] Ali Oskooei, Jannis Born, Matteo Manica, Vigneshwari Subramanian, Julio Sáez-Rodríguez, and María Rodríguez Martínez. PaccMann: Prediction of anticancer compound sensitivity with multi-modal attention-based neural networks, 2018.

[121] Yu-Chiao Chiu, Hung-I 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. *BMC Medical Genomics*, 12, 01 2019.

[122] M. Li, Y. Wang, R. Zheng, X. Shi, y. li, F. Wu, and J. Wang. DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1, 2019.

[123] 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*, 35(19):3743–3751, 03 2019.

[124] Wojciech Samek and Klaus-Robert Müller. Towards Explainable ArtifFicial Intelligence. *Lecture Notes in Computer Science*, page 5–22, 2019.

[125] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[126] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, 11 2017.

[127] Antonio Fabregat, Konstantinos Sidiropoulos, Guilherme Viteri, Oscar Forner, Pablo Marin-Garcia, Vicente Arnau, Peter D'Eustachio, Lincoln Stein, and Henning Hermjakob. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics*, 18(1):142, Mar 2017.

[128] Eike Staub. An Interferon Response Gene Expression Signature Is Activated in a Subset of Medulloblastomas. *Translational Oncology*, 5(4):297 – IN6, 2012.

[129] Axel Hauschild, Jean-Jacques Grob, Lev V Demidov, Thomas Jouary, Ralf Gutzmer, Michael Millward, Piotr Rutkowski, CU Blank, Wilson H Miller, Eckhart Kaempgen, Salvador Martín-Algarra, Boguslawa Karaszewska, Cornelia Mauch, Vanna Chiarion-Sileni, Anne-Marie Martin, Suzanne Swann, Patricia Haney, Beloo Mirakhur, Mary E Guckert, and Paul B Chapman. Dabrafenib in BRAF-mutated metastatic melanoma: A multicentre, open-label, phase 3 randomised controlled trial. *Lancet*, 380:358–65, 06 2012.

[130] Arjun Khunger, Monica Khunger, and Vamsidhar Velcheti. Dabrafenib in combination with trametinib in the treatment of patients with BRAF V600-positive advanced or metastatic non-small cell lung cancer: clinical evidence and experience. *Therapeutic Advances in Respiratory Disease*, 12:175346661876761, 03 2018.

[131] Linifanib. *Drugs R D*, 10(2):111–122, 2010.

[132] Eng-Huat Tan, Glenwood D. Goss, Ravi Salgia, Benjamin Besse, David R. Gandara, Nasser H. Hanna, James Chih-Hsin Yang, Raymond Thertulien, Michael Wertheim, Julien Mazieres, Thomas Hensing, Christa Lee, Neeraj Gupta, Rajendra Pradhan, Jiang Qian, Qin Qin, Frank A. Scappaticci, Justin L. Ricker, Dawn M. Carlson, and Ross A. Soo. Phase 2 Trial of Linifanib (ABT-869) in Patients with Advanced Non-small Cell Lung Cancer. *Journal of Thoracic Oncology*, 6(8):1418 – 1425, 2011.

[133] Eunice S. Wang, Karen Yee, Liang Piu Koh, Donna Hogge, Sari Enschede, Dawn M. Carlson, Matthew Dudley, Keith Glaser, Evelyn McKeegan, Daniel H. Albert, Xiaohui Li, Rajendra Pradhan, and Wendy Stock. Phase 1 trial of linifanib (ABT-869) in patients with refractory or relapsed acute myeloid leukemia. *Leukemia & Lymphoma*, 53(8):1543–1551, 2012. PMID: 22280537.

[134] M. Levis. Quizartinib for the treatment of FLT3/ITD acute myeloid leukemia. *Future Oncol*, 10(9):1571–1579, 2014.

[135] Holger Fröhlich, Rudi Balling, Niko Beerenwinkel, Oliver Kohlbacher, Santosh Kumar, Thomas Lengauer, Marloes Maathuis, Yves Moreau, Susan Murphy, Teresa Przytycka, Michael Rebhan, Hannes Röst, Andreas Schuppert, Matthias Schwab, Rainer Spang, Daniel Stekhoven, Jimeng Sun, Andreas Weber, Daniel Ziemek, and Blaz Zupan. From hype to reality: Data science enabling personalized medicine. *BMC Medicine*, 16, 08 2018.

[136] Krisna C. Duong-Ly, Karthik Devarajan, Shuguang Liang, Kurumi Y. Horiuchi, Yuren Wang, Haiching Ma, and Jeffrey R. Peterson. Kinase inhibitor profiling reveals unexpected opportunities to inhibit disease-associated mutant kinases. *Cell Reports*, 14(4):772–781, 2016.

[137] Chandrasekhar V. Miduturu, Xianming Deng, Nicholas Kwiatkowski, Wannian Yang, Laurent Brault, Panagis Filippakopoulos, Eunah Chung, Qingkai Yang, Juerg Schwaller, Stefan Knapp, Randall W. King, Jiing-Dwan Lee, Sanna Herrgard, Patrick Zarrinkar, and Nathanael S. Gray. High-throughput kinase profiling: A more efficient approach toward the discovery of new kinase inhibitors. *Chemistry & Biology*, 18(7):868–879, 2011.

[138] Cyril Benes, Daniel A. Haber, Dave Beare, Elena J. Edelman, Howard Lightfoot, I. Richard Thompson, James A. Smith, Jorge Soares, Michael R. Stratton, Nidhi Bindal, P. Andrew Futreal, Patricia Greninger, Simon Forbes, Sridhar Ramaswamy, Wanjuan Yang, 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, 11 2012.

[139] Brinton Seashore-Ludlow, Matthew G. Rees, Jaime H. Cheah, Murat Cokol, Edmund V. Price, Matthew E. Coletti, Victor Jones, Nicole E. Bodycombe, Christian K. Soule, Joshua Gould, Benjamin Alexander, Ava Li, Philip Montgomery, Mathias J. Wawer, Nurdan Kuru, Joanne D. Kotz, C. Suk-Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dančík, Joshua A. Bittker, Michelle Palmer, James E. Bradner, Alykhan F. Shamji, Paul A. Clemons, and Stuart L. Schreiber. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery*, 5(11):1210–1223, 2015.

[140] 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, 2016.

[141] John Mpindi, Bhagwan Yadav, Päivi Ostling, Prson Gautam, Disha Malani, Astrid Murumägi, Akira Hirasawa, Sara Kangaspeska, Krister Wennerberg, Olli Kallioniemi, and Tero Aittokallio. Consistency in drug response profiling. *Nature*, 540:E5–E6, 11 2016.

[142] Mehreen Ali and Tero Aittokallio. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical Reviews*, 11, 08 2018.

[143] Matteo Manica, Ali Oskooei, Jannis Born, Vigneshwari Subramanian, Julio Saez-Rodriguez, and Maria Rodriguez Martinez. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics*, 2019, 10 2019.

[144] Hongyuan Dong, Jiaqing Xie, Zhi Jing, and Dexin Ren. Variational autoencoder for anti-cancer drug response prediction, 2021.

[145] Qi Wei and Stephen A. Ramsey. Predicting chemotherapy response using a variational autoencoder approach. *bioRxiv*, 2021.

[146] Jannis Born, Matteo Manica, Ali Oskooei, Joris Cadow, Greta Markert, and Maria Rodriguez Martinez. Paccmannrl: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience*, 24:102269, 03 2021.

[147] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[148] J. Simm, A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, and Y. Moreau. Macau: Scalable bayesian factorization with high-dimensional side information using mcmc. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2017.

[149] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering, 2017.

[150] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system. *ACM Computing Surveys*, 52(1):1–38, jan 2020.

[151] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.

[152] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, page 101–109, New York, NY, USA, 2019. Association for Computing Machinery.

[153] Wojciech Samek and Klaus-Robert Müller. Towards Explainable Artificial Intelligence. *Lecture Notes in Computer Science*, page 5–22, 2019.

[154] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks, 2018.

[155] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[156] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

[157] HMS LINCS KINOMEscan data. `http://lincs.hms.harvard.edu/kinomescan/`.

[158] Miles Fabian, William Biggs, Daniel Treiber, Corey Atteridge, Mihai Azimioara, Michael Benedetti, Todd Carter, Pietro Ciceri, Philip Edeen, Mark Floyd, Julia Ford, Margaret Galvin, Jay Gerlach, Robert Grotzfeld, Sanna Herrgard, Darren Insko, Michael Insko, Andiliy Lai, Jean-Michel Lélias, and David Lockhart. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nature biotechnology*, 23:329–36, 04 2005.

[159] HMS LINCS KINOMEscan Overview and Assay Principle. `https://www.discoverx.com/technologies-platforms/competitive-binding-technology/kinomescan-technology-platform`.

[160] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, Jan 2015.

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

[162] Tianqi Chen and Carlos Guestrin. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016.

[163] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[164] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[165] Edward Chen, Christopher Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Meirelles, Neil Clark, and Avi Ma'ayan. Enrichr: Interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14:128, 04 2013.

[166] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 05 2016.

[167] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLOS ONE*, 6(7):1–9, 07 2011.

[168] Petr Smirnov, Zhaleh Safikhani, Nehme El-Hachem, Dong Wang, Adrian She, Catharina Olsen, Mark Freeman, Heather Selby, Deena M.A. Gendoo, Patrick Grossmann, Andrew H. Beck, Hugo J.W.L. Aerts, Mathieu Lupien, Anna Goldenberg, and Benjamin Haibe-Kains. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics*, 32(8):1244–1246, 12 2015.

[169] Christophe Massard, Jean-Charles Soria, D. Alan Anthoney, A. Proctor, Angela Scaburri, Maria Adele Pacciarini, Bernard Laffranchi, Cinzia Pellizzoni, Guido Kroemer, Jean-Pierre Armand, Rastilav Balheda, and Christopher J. Twelves. A first in man, phase i dose-escalation study of pha-793887, an inhibitor of multiple cyclin-dependent kinases (cdk2, 1 and 4) reveals unexpected hepatotoxicity in patients with solid tumors. *Cell Cycle*, 10(6):963–970, 2011.

[170] The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 11 2020.

[171] Uniprot entry. `https://www.uniprot.org/uniprot/P27361#function`. Accessed: 2021-06-21.

[172] Jitka Holcakova, Peter Tomasec, Joachim Bugert, Eddie Wang, Gavin Wilkinson, Vladimir Krystof, Miroslav Strnad, and Borivoj Vojtesek. The inhibitor of cyclin-dependent kinases, olomoucine ii, exhibits potent antiviral properties. *Antiviral chemistry & chemotherapy*, 20:133–42, 01 2010.

[173] PHA-793887 targets. https://maayanlab.cloud/Harmonizome/gene_set/PHA-793887/LINCS+KinomeScan+Kinase+Inhibitor+Targets. Accessed: 2021-06-21.

[174] Lacey M Litchfield, Karsten Boehnke, Manisha Brahmachary, Cecilia Mur, Chen Bi, Jennifer R Stephens, J Michael Sauder, Sonia M Gutiérrez, Ann M McNulty, Xiang S Ye, Wenjuan Wu, María José Lallena, Xueqian Gong, Farhana F Merzoug, Valerie M Jansen, and Sean G Buchanan. Combined inhibition of pim and cdk4/6 suppresses both mtor signaling and rb phosphorylation and potentiates pi3k inhibition in cancer cells. *Oncotarget*, 11(17):1478–1492, 2020.

[175] Axel Hauschild, Jean-Jacques Grob, Lev Demidov, Thomas Jouary, Ralf Gutzmer, Michael Millward, Piotr Rutkowski, CU Blank, Wilson Miller, Eckhart Kaempgen, Salvador Martín-Algarra, Boguslawa Karaszewska, Cornelia Mauch, Vanna Chiarion-Sileni, Anne-Marie Martin, Suzanne Swann, Patricia Haney, Beloo Mirakhur, Mary Guckert, and Paul Chapman. Dabrafenib in braf-mutated metastatic melanoma: A multicentre, open-label, phase 3 randomised controlled trial. *Lancet*, 380:358–65, 06 2012.

[176] Bilgen Gencler and Müzeyyen Gönül. Cutaneous Side Effects of BRAF Inhibitors in Advanced Melanoma: Review of the Literature. *Dermatology Research and Practice*, 2016:1–6, 03 2016.

[177] Matteo Manica, Ali Oskooei, Jannis Born, Vigneshwari Subramanian, Julio Saez-Rodriguez, and Maria Rodriguez Martinez. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics*, 2019, 10 2019.

[178] Kacper Sokol and Peter Flach. Explainability fact sheets. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan 2020.

[179] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[180] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2:573–584, 10 2020.

[181] Pavel Karpov, Guillaume Godin, and Igor Tetko. Transformer-cnn: Swiss knife for qsar modeling and interpretation. *Journal of Cheminformatics*, 12, 03 2020.

[182] H.C. Stephen Chan, Hanbin Shan, Thamani Dahoun, Horst Vogel, and Shuguang Yuan. Advancing drug discovery via artificial intelligence. *Trends in Pharmacological Sciences*, 40(8):592–604, 2019. Special Issue: Rise of Machines in Medicine.

[183] Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen. Innovation in the pharmaceutical industry: New estimates of r&d costs. *Journal of Health Economics*, 47:20–33, 2016.

[184] Steven M Paul, Daniel S Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R Lindborg, and Aaron Leigh Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9:203–214, 2010.

[185] Joe Greener, Lewis Moffat, and David Jones. Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific Reports*, 8, 11 2018.

[186] Brian L. Hie and Kevin K. Yang. Adaptive machine learning for protein engineering. *Current Opinion in Structural Biology*, 72:145–152, 2022.

[187] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2323–2332. PMLR, 10–15 Jul 2018.

[188] Wengong Jin, Dr.Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4839–4848. PMLR, 13–18 Jul 2020.

[189] Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, Sandra Viknander, Wissam Abuajwa, Otto Savolainen, Rolandas Meškys, Martin Engqvist, and Aleksej Zelezniak. Expanding functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, 3:1–10, 04 2021.

[190] Paulina Szymczak, Marcin Możejko, Tomasz Grzegorzek, Marta Bauer, Damian Neubauer, Michał Michalski, Jacek Sroka, Piotr Setny, Wojciech Kamysz, and Ewa Szczurek. Hydramp: a deep generative model for antimicrobial peptide discovery. 2022.

[191] Joshua Meyers, Benedek Fabian, and Nathan Brown. De novo molecular design and generative models. *Drug Discovery Today*, 26(11):2707–2715, 2021.

[192] Liliana Brandão, Fernando Paulo Belfo, and Alexandre Silva. Wavelet-based cancer drug recommender system. *Procedia Computer Science*, 181:487–494, 2021. CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020.

[193] Chayaporn Suphavilai, Denis Bertrand, and Niranjan Nagarajan. Predicting Cancer Drug Response using a Recommender System. *Bioinformatics*, 34(22):3907–3914, 06 2018.

[194] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[195] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[196] Mi Yang, Jaak Simm, Chi Chung Lam, Pooya Zakeri, Gerard J. P. van Westen, Yves Moreau, and Julio Saez-Rodriguez. Linking drug target and pathway activation for effective therapy using multi-task learning. *Scientific Reports*, 8, 12 2018.

[197] J. Simm, A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, and Y. Moreau. Macau: Scalable bayesian factorization with high-dimensional side information using mcmc. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2017.

[198] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering, 2017.

[199] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 52(1), feb 2019.

[200] Jannis Born, Matteo Manica, Ali Oskooei, Joris Cadow, Greta Markert, and María Rodríguez Martínez. Paccmannrl: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience*, 24(4):102269, 2021.

[201] Sunghoon Joo, Min Kim, Jaeho Yang, and Jeahyun Park. Generative model for proposing drug candidates satisfying anticancer properties using a conditional variational autoencoder. *ACS Omega*, XXXX, 07 2020.

[202] Sejin Park and Hyunju Lee. A molecular generative model with genetic algorithm and tree search for cancer samples. *CoRR*, abs/2112.08959, 2021.

[203] Hongyuan Dong, Jiaqing Xie, Zhi Jing, and Dexin Ren. Variational autoencoder for anti-cancer drug response prediction, 2020.

[204] Peilin Jia, Ruifeng Hu, Guangsheng Pei, Yulin Dai, Yin-Ying Wang, and Zhongming Zhao. Deep generative neural network for accurate drug response imputation. *Nature Communications*, 12(1), mar 2021.

[205] Dong-Kyu Chae, Jin-Soo Kang, Sang-Wook Kim, and Jung-Tae Lee. Cfgan: A generic collaborative filtering framework based on generative adversarial networks. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 137–146, New York, NY, USA, 2018. Association for Computing Machinery.

[206] Haoyu Wang, Nan Shao, and Defu Lian. Adversarial binary collaborative filtering for implicit feedback. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[207] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. IRGAN. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, aug 2017.

[208] Binbin Jin, Defu Lian, Zheng Liu, Qi Liu, Jianhui Ma, Xing Xie, and Enhong Chen. Sampling-decomposable generative adversarial recommender. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22629–22639. Curran Associates, Inc., 2020.

[209] Xinshi Chen, Shuang Li, Hui Li, Shaohua Jiang, Yuan Qi, and Le Song. Generative adversarial user model for reinforcement learning based recommendation system. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1052–1061. PMLR, 09–15 Jun 2019.

[210] Xueying Bai, Jian Guan, and Hongning Wang. A model-based reinforcement learning with adversarial training for online recommendation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[211] Jakub M. Tomczak and Max Welling. Vae with a vampprior, 2017.

[212] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering, 2016.

[213] Fabian Falck, Haoting Zhang, Matthew Willetts, George Nicholson, Christopher Yau, and Chris Holmes. Multi-facet clustering variational autoencoders, 2021.

[214] Chunsheng Guo, Jialuo Zhou, Huahua Chen, Na Ying, Jianwu Zhang, and Di Zhou. Variational autoencoder with optimizing gaussian mixture model priors. *IEEE Access*, 8:43992–44005, 2020.

[215] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *CoRR*, abs/1611.02648, 2016.

[216] Wenxian Shi, Hao Zhou, Ning Miao, and Lei Li. Dispersed exponential family mixture vaes for interpretable text generation, 2019.

[217] stability_selection module. `https://thuijskens.github.io/stability-selection/docs/index.html`.

[218] Gilles Louppe. *Understanding Random Forests: From Theory to Practice*. PhD thesis, University of Liège, Faculty of Applied Sciences, 10 2014.

[219] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 431–439. Curran Associates, Inc., 2013.