Pontifícia Universidade Católica
do Rio de Janeiro

Sofia Pontes de Miranda

Predicting drug sensitivity of cancer cells based on genomic data

Dissertação de Mestrado

Dissertation presented to the Programa de Pós-graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção.

Advisor: Prof. Julia Lima Fleck
Co-Advisor: Prof. Stephen Piccolo

Rio de Janeiro
January 2021

Pontifícia Universidade Católica
do Rio de Janeiro

Sofia Pontes de Miranda

Predicting drug sensitivity of cancer cells based on genomic data

Dissertation presented to the Programa de Pós–graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção. Approved by the Examination Committee.

Prof. Julia Lima Fleck
Advisor
Departamento de Engenharia Industrial - PUC-Rio

Prof. Stephen R. Piccolo
Co-Advisor
BYU

Prof. Fernanda Araujo Baião Amorim
Departamento de Engenharia Industrial - PUC-Rio

Prof. Vincent Augusto
EMSE

Rio de Janeiro, January 22nd 2021

Sofia Pontes de Miranda

The author has graduated Industrial Engineering from Pontifical Catholic University of Rio de Janeiro (2018).

Bibliographic data

Miranda, Sofia Pontes de

Predicting drug sensitivity of cancer cells based on genomic data / Sofia Pontes de Miranda ; advisor: Julia Lima Fleck ; co-advisor: Stephen Piccolo. – 2021.
152 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2021.
Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Aprendizado de máquina. 3. Genômica. 4. Predição da eficácia a droga. 5. Aprendizado semi-supervisionado. 6. Aprendizado supervisionado. I. Fleck, Julia Lima. II. Piccolo, Stephen. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

To my parents, for their support
and encouragement.

Acknowledgements

To Professor Julia Fleck, for her inspiring teachings, stimulating discussions and precious suggestions.

To Professor Stephen Piccolo, for the dedicated support and substantial contributions in the development of this research.

To my family, for the continuous support, essential encouragement and unconditional love.

To the Examining Committee, for the availability and assessment of the results of this research.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Pontes de Miranda, Sofia; Lima Fleck, Julia (Advisor); Piccolo, Stephen (Co-Advisor). **Predicting drug sensitivity of cancer cells based on genomic data**. Rio de Janeiro 2020. 152p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Accurately predicting drug responses for a given sample based on molecular features may help to optimize drug-development pipelines and explain mechanisms behind treatment responses. In this dissertation, two case studies were generated, each applying different genomic data to predict drug response. Case study 1 evaluated DNA methylation profile data as one type of molecular feature that is known to drive tumorigenesis and modulate treatment responses. Using genome-wide, DNA methylation profiles from 987 cell lines in the Genomics of Drug Sensitivity in Cancer (GDSC) database, we used machine-learning algorithms to evaluate the potential to predict cytotoxic responses for eight anti-cancer drugs. We compared the performance of five classification algorithms and four regression algorithms representing diverse methodologies, including tree-, probability-, kernel-, ensemble- and distance-based approaches. By applying artificial subsampling in varying degrees, this research aims to understand whether training based on relatively extreme outcomes would yield improved performance. When using classification or regression algorithms to predict discrete or continuous responses, respectively, we consistently observed excellent predictive performance when the training and test sets consisted of cell-line data. Classification algorithms performed best when we trained the models using cell lines with relatively extreme drug-response values, attaining area-under-the-receiver-operating-characteristic-curve values as high as 0.97. The regression algorithms performed best when we trained the models using the full range of drug-response values, although this depended on the performance metrics we used. Case study 2 evaluated RNA-seq data as one of the most popular molecular data used to study drug efficacy. By applying a semi-supervised learning approach, this research aimed to understand the impact of combining labeled and unlabeled data to improve model prediction. Using genome-wide RNA-seq labeled data from an average of 125 AML

tumor samples in the Beat AML database (varying by drug type) and 151 unlabeled AML tumor samples in The Cancer Genome Atlas (TCGA) database, we used a semi-supervised model structure to predict cytotoxic responses for four anti-cancer drugs. Semi-supervised models were generated, while assessing several parameter combinations and were compared against supervised classification algorithms.

Keywords

Machine Learning; Genomics; Methylation; RNA-seq; Classification Models; Regression Models; Drug Response Prediction; Semi-supervised learning; Supervised learning; Cancer.

Resumo

Pontes de Miranda, Sofia; Lima Fleck, Julia (Advisor); Piccolo, Stephen (Co-Advisor). **Prevendo a eficácia de drogas a partir de células cancerosas baseado em dados genômicos.** Rio de Janeiro 2020. 152p. Dissertação de Mestrado – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Prever com precisão a resposta a drogas para uma dada amostra baseado em características moleculares pode ajudar a otimizar o desenvolvimento de drogas e explicar mecanismos por trás das respostas aos tratamentos. Nessa dissertação, dois estudos de caso foram gerados, cada um aplicando diferentes dados genômicos para a previsão de resposta a drogas. O estudo de caso 1 avaliou dados de perfis de metilação de DNA como um tipo de característica molecular que se sabe ser responsável por causar tumorigênese e modular a resposta a tratamentos. Usando perfis de metilação de 987 linhagens celulares do genoma completo na base de dados Genomics of Drug Sensitivity in Cancer (GDSC), utilizamos algoritmos de aprendizado de máquina para avaliar o potencial preditivo de respostas citotóxicas para oito drogas contra o câncer. Nós comparamos a performance de cinco algoritmos de classificação e quatro algoritmos de regressão representando metodologias diversas, incluindo abordagens *tree-*, *probability-*, *kernel-*, *ensemble-* e *distance-based*. Aplicando sub-amostragem artificial em graus variados, essa pesquisa procura avaliar se o treinamento baseado em resultados relativamente extremos geraria melhoria no desempenho. Ao utilizar algoritmos de classificação e de regressão para prever respostas discretas ou contínuas, respectivamente, nós observamos consistentemente excelente desempenho na predição quando os conjuntos de treinamento e teste consistiam em dados de linhagens celulares. Algoritmos de classificação apresentaram melhor desempenho quando nós treinamos os modelos utilizando linhagens celulares com valores de resposta a drogas relativamente extremos, obtendo valores de *area-under-the-receiver-operating-characteristic-curve* de até 0,97. Os algoritmos de regressão tiveram melhor desempenho quando treinamos os modelos utilizado o intervalo completo de valores de resposta às drogas, apesar da dependência das métricas de desempenho utilizadas.

O estudo de caso 2 avaliou dados de RNA-seq, dados estes comumente utilizados no estudo da eficácia de drogas. Aplicando uma abordagem de aprendizado semi-supervisionado, essa pesquisa busca avaliar o impacto da combinação de dados rotulados e não-rotulados para melhorar a predição do modelo. Usando dados rotulados de RNA-seq do genoma completo de uma média de 125 amostras de tumor AML rotuladas da base de dados Beat AML (separados por tipos de droga) e 151 amostras de tumor AML não-rotuladas na base de dados The Cancer Genome Atlas (TCGA), utilizamos uma estrutura de modelo semi-supervisionado para prever respostas citotóxicas para quatro drogas contra câncer. Modelos semi-supervisionados foram gerados, avaliando várias combinações de parâmetros e foram comparados com os algoritmos supervisionados de classificação.

Palavras-chave

Aprendizado de máquina; Genômica; Metilação; Sequenciamento de RNA; Modelos de classificação; Modelos de regressão; Predição da eficácia a droga; Aprendizado semi-supervisionado; Aprendizado supervisionado; Câncer.

Table of Contents

Abbreviations

ACC – Accuracy

AML – Acute myeloid leukemia

AUC – Area under the receiver operating characteristic curve

DNA – deoxyribonucleic acid

E – Entropy

FDR – Benjamini-Hochberg False Discovery Rate

FS – Feature selection

GBM – Gradient Boosting Machines

GDC – Genomic Data Commons

GDSC – Genomics of Drug Sensitivity in Cancer

IG – Information gain

KNN – K-nearest neighbors

MAE – Mean absolute error

MCC – Matthews correlation coefficient

MMCE – Mean misclassification error

MSE – Mean squared error

NB – Naïve Bayes

PCC – Pearson correlation coefficient

RF – Random forests

RMSE – Root-mean-square deviation

RNA – ribonucleic acid

RNA-seq –RNA-sequencing

R^2 – R-squared

SCC – Spearman's rank correlation coefficient

SL – Supervised learning

SSL – Semi-supervised learning

SVM – Support vector machines

TCGA – The Cancer Genome Atlas

UL – Unsupervised learning

XGBoost – Extreme Gradient Boosting

"Caminante, son tus huellas el camino y nada más; caminante, no hay camino, se hace camino al andar."

Antonio Machado, *Cantares*

Traveler, your footprints are the road, nothing else; traveler, there is no road; you make your own path as you tread.

Caminhante, são tuas pegadas o caminho e nada mais; caminhante, não há caminho, faz-se o caminho ao caminhar.

1
Introduction

Cancers are complex, dynamic diseases characterized by aberrant cellular processes such as excessive proliferation, resistance to apoptosis, and genomic instability (Hanahan and Weinberg, 2011). Tumors are caused by somatic variations, which can affect individual nucleotides or larger segments of DNA (Yao and Dai, 2014). Dysregulation of cellular function can also be caused by epigenetic modifications, including aberrant DNA methylation (Esteller et al., 2001). One goal of cancer research is to advance precision medicine through identifying genomic and epigenomic features that influence treatment outcomes in individuals (McLeod, 2013). In this context, therapeutic decisions have the potential to be guided by molecular signatures.

Molecular features offer insights to patients' traits that may impact drug response. By studying genomic information, scientists are able to infer patient characteristics, understand how they will influence response to treatment and optimize treatment decision. Such information can be used to classify cancer patients into groups that will most likely benefit from a certain treatment, thus generating a tool to aid in clinical decision making.

When studying cancer, researchers may use patient or cell line samples. Cancer cell lines are cell cultures derived from tumor samples. They represent one of the least expensive and most studied preclinical models (Masters, 2000). Drug screening in cell lines can be used to prioritize candidate drugs for testing in humans. In performing a screen, researchers calculate IC_{50} values, which quantify the amount of drug necessary to induce a biological response in half of the cells tested for a given experiment (Sebaugh, 2011). Drugs with a relatively high potency (corresponding to low log-transformed IC_{50} values) are generally considered to be the strongest candidates for use in humans, although patient safety must also be evaluated. After a candidate drug has

been identified, researchers may seek to identify molecular markers associated with those responses, comparing cell lines that respond to the drug against those that do not. Such markers might be useful for elucidating drug mechanisms or eventually predicting clinical responses in patients (Iorio et al., 2016).

Over the past decade, researchers have catalogued the molecular profiles of more than a thousand cancer cell lines and their responses to hundreds of drugs (Barretina et al., 2012; Yang et al., 2013; Rees et al., 2016). In addition, recent efforts to catalog molecular profiles in human tumors have resulted in massive collections of publicly available molecular data for tumor samples (ICGC, 2010; Tomczak et al., 2015; Forbes et al., 2017). These resources have been made publicly available, thus providing an opportunity for researchers to identify molecular signatures that predict drug responses in preclinical and clinical settings.

1.1 Dissertation Objectives

The objective of this research is to develop new methodologies to accurately predict drug response based on molecular features. By exploring data analytics and machine learning approaches, this research aims to contribute to clinical decision making and the optimization of drug-development pipelines. In particular, this dissertation aims to:

- Evaluate the performance of methylation data to effectively predict drug response;
- Understand the behavior of two learning strategies (supervised and semi-supervised) when applied to molecular data;
- Assess different feature selection methods that would be adequate to molecular features;
- Evaluate the impact of subsampling strategies on drug response prediction performance.

1.2 Dissertation Structure

This work is organized in the following structure:

- **Chapter 2 – Methodology.** Presents fundamental concepts of biological data, data analytics techniques and processes, machine learning fundamentals, algorithms and evaluation metrics. The chapter concludes with a review of existing literature.
- **Chapter 3 – Case Studies.** Details two case studies containing different machine learning strategies and applications. Each case study has the following structure:
    - **Introduction.** Introduces each case study, including database details, study motivation and brief research description.
    - **Methods.** Presents machine learning processes and models unique to each case study. This topic encompasses the description of preprocessing methods, hyperparameters description, model development and performance evaluation.
    - **Results.** Presents main findings of each case study.
    - **Discussion.** Analyzes obtained results and links it to existent literature.
    - **Conclusion.** Addresses main conclusions of each case study.

2
Methodology

This section presents the general methodology that may be applied to both research modules. Descriptions of used biological data as well as main machine learning processes are presented.

2.1
Biological Data

Genomics is an interdisciplinary field in Biology that studies the structure, function, evolution, mapping and changes in the complete DNA sequence, including all genes (Ginsburg et al., 2009). In this chapter we present the fundamental concepts of molecular biology in the field of genomics.

2.1.1
Genomic Data Fundamentals

Each cell in the human body holds a person's unique deoxyribonucleic acid, also known as DNA. The DNA is composed of two polynucleotide chains that coil around each other, forming a double helix; it is responsible for carrying the genetic instructions for the development, operation, growth and reproduction of all cells, tasks and processes in the human body. Genes are small segments of information within the DNA, where each gene is responsible for a specific instruction or trait for that one individual. Through the process of transcription, ribonucleic acid (RNA) is created based on the DNA. This RNA is then "translated" into proteins, which then carry out all functions that are necessary to create and maintain life (Gunder et al., 2011).

Technologies for profiling gene-expression levels are widely available and reflect the downstream effects of genomic and epigenomic aberrations. However, gene-expression profiles may be difficult to apply in the clinic because of the RNA instability

(Geeleher et al., 2017). Moreover, gene-expression data are generated using a wide range of technologies (e.g., different types of oligonucleotide microarrays and RNA - sequencing), and are preprocessed using diverse algorithms. Thus, it is often difficult to combine datasets from multiple sources (e.g., preclinical and tumor data).

DNA methylation is an epigenetic mechanism that controls gene-expression levels. The addition of a methyl group to DNA may lead to changes in DNA stability, chromatin structure and DNA-protein interactions. Hypermethylation of CpG islands in promoter regions of DNA has been acknowledged as an important means of gene inactivation and its occurrence has been detected in almost all types of human tumors (Esteller, 2002). Similar to genetic alterations, methylation changes to DNA may alter a gene's behavior. However, hypermethylation can be reversed with the use of targeted therapy (Szyf, 2008), making it an attractive target for anticancer therapy (Szyf, 1994; Arechederra et al., 2018).

In some cases, DNA methylation levels for a single gene may control cellular responses for a given drug. For example, MGMT hypermethylation predicts temozolomide responses in glioblastomas (Hegi et al., 2005), and BRCA1 hypermethylation predicts responses to poly ADP ribose polymerase inhibitors in breast carcinomas (Island, 2010). However, in many cases, drug responses are likely influenced by the combined effects of many genes interacting in the context of signaling pathways (Faivre et al., 2006). Accordingly, to maximize our ability to predict drug responses, it is critical to account for this complexity.

RNA-sequencing (RNA-seq) is a technique used to examine RNA sequences in a given sample. It analyses the gene expression transcriptome patterns encoded in the RNA (Wang et al., 2009). Transcriptome refers to the complete set of all RNA transcripts in a population of cells. Understanding these patterns allows researchers to connect the genome information with its functional protein expression behavior. RNA-seq allows us to know which genes are activated in a given cell, their expression level, and when this activation occurs (Ozsolak et al., 2011). By understanding this behavior, researchers are able to assess if these changes may indicate disease development. Since some of these observations would not be detected by DNA sequencing, RNA-seq analysis is a key process to identify genomic features that may influence cancer.

2.1.2
Batch Effect Removal

Gene expression technologies are able to measure the expression of several thousand genes at a time. Using multiple probes, they are able to assess transcriptome patterns and identify changes in response to any perturbation. However, researchers are concerned about the reliability of this technology and hence, its utility (Sims, 2009; Kathleen Kerr, 2003). These microarray technologies are very sensitive to external stimuli and can be affected by a number of different variables, such as reagents from different lots, time of the analysis, position of the culture dish and different technicians using the machine (Lander, 1999; Müller et al., 2016).

Due to this reason, we can observe differences in microarray values depending on the batch of a given sample. The term "batch effects" refers to the moment that a given sample was analyzed and the variables that may have influenced its results and generated possible errors. Batch effects are almost inevitable as gene expression microarray technologies can assay few samples per batch; thus, generating small differences between each group. Since technologies may be able to analyze up to 96 samples in each batch and hundreds of samples are needed for any type of analysis, batch effects unavoidably impact results.

To address this issue, scientists have developed ways to adjust data for batch effects. There are several processes that have been developed to adjust samples. Ideally, they would yield the same results but since they are based on different statistical models, their overall effectiveness may vary. A popular process to remove batch effects is called "Combating Batch Effects When Combining Batches of Gene Expression Microarray Data" (ComBat) (Johnson et al., 2007). It is an empirical based method that estimates parameters for location and scale adjustment of each batch for each gene independently (Müller et al., 2016). This framework follows an empirical Bayes framework for adjusting data that is robust to outliers in small sample sizes (Johnson et al., 2007). ComBat is considered one of the most efficient ways to remove batch effects and can robustly manage high dimensional data that contains a small number of samples (Chen et al., 2011).

2.1.3 Database Description

In this section, the three databases used in this research are briefly described. Access information as well as a short database description are given.

Obtaining molecular profiles from tumor samples and generating drug responses from clinical trials are complex and expensive tasks. For this reason, pre-clinical biological models that are able to capture the molecular features of cancer and the impact of diverse therapeutic treatment options are necessary (Iorio et al., 2016). Human cell lines are an important tool for drug development, enabling experimental modeling and fostering the understanding of drug development. The Genomics of Drug Sensitivity in Cancer (GDSC) database contains data for human cell lines derived from common and rare types of adult and childhood cancers. GDSC provides pharmacological, genomic, transcriptomic, and epigenetic characterization of over 1,001 human cancer cell lines (Iorio et al., 2016). Two databases are available, GDSC1 (curated between 2010 and 2015) and GDSC2 (curated from 2015 onwards), including 987 and 809 cell line samples, respectively. These datasets can be accessed through the GDSC portal (http://www.cancerrxgene.org).

Understanding the difficult access to cancer patients' molecular information, the National Institute of Health (NIH) launched the The Cancer Genome Atlas (TCGA) pilot project in 2005. TCGA aims to accelerate comprehensive understanding of cancer genetics, new anti-cancer treatment strategies, diagnosis methods and preventive approaches (Tomczak, 2015). The TCGA database is an open source project that was developed to become a catalogue of cancer genomic profiles. It aims to gather and discover cancer genome alterations in large cohorts of over 30 human tumor samples (Hutter and Zenklusen, 2018). To encourage integrated multi-dimensional analyses, TCGA offers several genetic data structures, including RNA sequencing, MicroRNA sequencing, DNA sequencing, SNP-based platforms, Array-based DNA methylation sequencing and Reverse-phase protein array for over 30 types of cancer. The TCGA platform can be accessed through the Genomic Data Commons (GDC) Portal (https://portal.gdc.cancer.gov/).

Another database covering tumor molecular information is the Beat AML project. This database focuses on acute myeloid leukemia (AML), which is a cancer of

the bone marrow and blood that quickly advances if not treated. In AML, the bone marrow produces abnormal platelets and white and red cells. Approximately 20,000 new cases of AML and 11,000 deaths are expected in 2020 (Surveillance, Epidemiology and End Result Program (SEER), 2020). The Beat AML database reports a cohort of 672 tumor samples collected from 562 patients (Tyner et al., 2018). Samples were assessed using whole-exome sequencing, RNA sequencing and ex-vivo drug sensitivity analyses for 122 anti-cancer drugs. Data availability includes clinical, genomic and transcriptomic information from patients. Beat AML can be accessed through the Vizome Platform (http://vizome.org/aml/).

2.2 Data Analytics

To analyze the biological data, several machine learning analyses are proposed to aid the data interpretation. In this procedure, specific processes and methods are applied and are described below.

2.2.1 Supervised vs. Unsupervised vs. Semi-Supervised Learning

In the machine learning field, there are two main learning approaches: supervised and unsupervised learning. Both approaches use statistics and mathematical algorithms to find patterns in data. Thus, by studying previous data, learning its patterns and applying these newly found patterns, machine learning models are able to make predictions for new data points.

When we want to make an assessment, we know that we should think about the variables (also known as features) that may impact our final decision. For example, if we were to answer the question “Should I take my jacket with me when I leave the house?”, we know that the answer depends on a couple of variables, such as: how long I am staying out, how cold it will be today, whether it is raining, where I am going and so on. The answers to these questions are called our input data. They are information that will aid us to make a decision.

In supervised learning (SL), the machine learning algorithm learns patterns based on examples. Thus, by feeding past samples and their final outcomes, the model is able to understand patterns that influenced and generated that specific decision. In our example, we want to predict if we should take our coat (y) by analyzing our input variables (x). We feed both information, x and y, to the model and it will generate a mathematical mapping function that will connect each input data combination to the correct output data. A key assumption of supervised learning is that our dataset contains samples (where each sample is a set of x and y) that represent all possible outcomes. This learning process is called "supervised" because it behaves as a teacher, who supervises the learning process of his students. As we know the correct output to each sample, the algorithm iteratively makes predictions based on our input data and is corrected if it makes a mistake. This learning process repeats itself until the algorithm has achieved an acceptable learning performance (Friedman et al., 2001).

Although supervised learning is a very popular strategy, we may not always have easy access to a labeled dataset. Unsupervised learning (UL) is where you only have input data (x), but no corresponding outputs (y) are available. This learning strategy received the name "unsupervised" as we do not know the correct outputs of our training set in advance. Therefore, the model is unable to recreate the environment of "a teacher supervising the learning process of his students." Machine learning algorithms are left to uncover any existent patterns and structures on their own. The model's goal is to find any underlying data structure, pattern and/or distribution that yields more information about the dataset behavior (Friedman et al., 2001).

Semi-supervised learning (SSL) is a machine learning strategy that comprises the previous two strategies, supervised and unsupervised learning. SSL uses a combination of a small labeled dataset and a larger unlabeled dataset (Van Engelen et al., 2020). The first step is to train a model with the labeled samples (supervised learning process). Then, this initial model is used to predict the unlabeled datasets pseudo-labels (unsupervised process). This pseudo-labeling approach is called "self-training method" (also known as "self-learning"), which is one of the most basic pseudo-labeling strategies (Triguero et al., 2015). Lastly, we concatenate both datasets, now both with outputs, into a new training set and generate a new model. This strategy

aims to reduce error and improve accuracy of models generated with a small sample size, but it is impossible to guarantee that the introduction of unlabeled data will improve performance consistently (Van Engelen et al., 2020).

2.2.2 The High Dimensionality of Biological Data

In biomedical research, the need for classification in high dimensional data can be a constant challenge. This challenge arises from several factors such as the exponential relationship between the computational complexity and the number of dimensions (Yu et al., 2003), the exponential increase of the required number of labeled samples (Maimon et al., 2002), data sparsity in higher dimensions hampering the learning process (Caruana et al., 2008) and high computational cost, leading to long training periods and possibly worsen predictions.

The phrase “curse of dimensionality”, attributed to Richard Bellman (Bellman, 2015), refers to the various difficulties of using brute force to optimize a mathematical function with too many input features. The reduction of dimensionality of a dataset is a preprocessing step to solve the problems of a high dimensional dataset. It comprises the reduction of unnecessary or redundant data characteristics (features).

There are two main strategies to reduce dimensionality: feature construction and feature selection (FS), where both focus on improving learning performance, decreasing computational costs and generating a sturdier model.

Feature construction assigns the original high dimensional attributes into a lower dimensional space, while retaining as much data information as possible. This space is normally generated by applying linear and non-linear embedding methods (Guyon et al., 2008). However, since feature construction generates a lower dimensional space by combining features together, we lose particular feature significance. Thus, we do not know the specific impact of a feature in the final problem.

Due to this problem, feature selection is preferred in bioinformatics (Li et al., 2017). Feature selection is performed to select relevant and informative features to a specific problem, thus generating a feature subset to be used in model creation (Guyon et al., 2008).

2.2.3
Feature Selection

Feature selection (FS) consists of identifying a feature subgroup that will provide the data information as the original feature set by excluding redundant and/or irrelevant features. Features can be defined as irrelevant and redundant if they do not contribute to data processing or if they hold similar information as another feature (Gnana et al., 2016). The key objective of feature selection is to generate more comprehensible models and improve prediction performance, decreasing training time and improving evaluation metrics (Gandhi et al., 2017).

Genomic data includes a large number of features for each sample, each having its own biological relevance in healthcare research and patient treatment. Feature selection is an efficient way to identify relevant features and reduce dimensionality. A FS analysis either outputs a feature rank, where each feature is assigned a relevance score (higher score implies higher feature relevance) or outputs a feature subset of previously determined size (Guyon et al., 2008).

FS learning follows a similar structure as the machine learning approaches described in section 2.2.1. Learning can be supervised, semi-supervised or unsupervised depending on label availability. If the input data contains output labels, supervised strategies can be used; if not, unsupervised strategies are applied. If only part of the input data contains an output value, semi-supervised strategies can be explored (Li et al., 2017).

Feature selection algorithms have four main approaches: filter, wrapper, embedded and hybrid methods (which comprises a combination of the other three methods). Filter methods are generally used as a preprocessing step in supervised learning, where a feature subset is chosen based on their relationship with the target label. This relationship is evaluated based on different univariate statistical techniques and feature importance methods. Filter methods are faster and less computationally expensive than other methods, being a popular choice when dealing with high dimensional data.

Wrapper methods search for an optimal feature subsample. This search iteratively tries new feature subsets based on inferences made on previous iterations. This method presents high accuracy, but usually demands high computer resources as it resembles a greedy search problem. Embedded methods portray the benefits of filter and wrapper approaches, combining the feature selection and algorithm learning process. It is implemented on algorithms that include an intrinsic FS method, such as decision trees and neural networks. They are less computationally demanding than wrapper methods but have their application limited to specific learning algorithms.

Information gain (IG) algorithm is a popular information-based filter method. It measures the entropy (E) of a given dataset and selects the subset with lower entropy. Entropy measures the amount of variance, or uncertainty, in the given data and can be defined as (Li et al., 2017):

$$E = -\sum_{i}^{N} p_i \times \log_2 p_i$$

Where N represents the total number of classes in the dataset and p_i represents the probability of randomly picking a sample of class i.

Therefore, the amount of entropy of a given dataset represents how unpredictable (or how pure) that same dataset is (Guyon et al., 2008). Information gain measures how much variance (entropy) is removed when selecting a specific feature subset. In other words, it measures the reduction of uncertainty for one variable (output) given a known value for another variable (MacKay et al., 2003) and can be formally stated as:

$$IG = I(X,Y) = E(X) - E(X|Y)$$

Where, I(X,Y) is the mutual information (information gain), E(X) is the entropy for X and E(X|Y) is the conditional entropy for X given Y.

Another approach to select an optimal feature subset is by applying similarity-based methods. ReliefF (Kira et al, 1992) assesses feature importance by evaluating

the ability of features to preserve data similarity by building an affinity matrix and then obtaining attribute scores. A disadvantage of this method is that it is not capable of dealing with feature redundancy (Li et al., 2017). This method is able to estimate feature quality in classification problems that have strong dependency between attributes. (Urbanowicz et al., 2018). It applies a nearest neighbor-based function to identify feature statistics (relevance). Assuming that *S* instances are randomly selected from the total *N* instances, then the *ReliefF* feature score can be described as (Li et al., 2017):

$$ReliefF = \frac{1}{c}\sum_{j=1}^{S}\left(-\frac{1}{t_j}\times\sum_{x_r\in NH(j)} d\big(X(j,i)-X(r,i)\big)+\sum_{y\neq y_i}\frac{1}{v_{jy}}\times\frac{P(y)}{1-P(y)}\times\sum_{x_r\in NM(j,y)} d(X(j,i)-X(r,i))\right)$$

Where, *NH(j)* and *NM(j,y)* are the nearest instances of x_j in the same class and in class *y*. *NH(j)* size is t_j and *NM(j,y)* size is v_{jy}. P(y) represents the probability of a sample being of class *y* and *c* represents the total number of classes.

Another feature selection evaluation approach is statistic-based evaluations. The Kruskal-Wallis H test (Kruskal et al., 1952) is a rank-based non-parametric statistical test that evaluates whether two or more independent groups are different in a variable of interest (output). When applied as a feature selection method, it acts as a filter algorithm, testing each feature and generating scores for attributes. It is an alternative evaluation to the One Way ANOVA (Ostertagova et al., 2014). The H test determines whether the medians of the evaluated groups are different, thus it will evaluate if there is a significant difference between groups. It is defined as (Kruskal et al., 1952):

$$H\,Test = \frac{12}{M(M+1)}\times\sum_{i=1}^{S}\frac{R_i^2}{n_i}-3(M+1)$$

Where, S represents the number of samples, n_i represents the number of observations in the i^{th} sample, $M=\sum n_i$ and R represents the sum of the ranks in the i^{th} sample.

Another filter method to select features is the statistic-based Pearson correlation coefficient (PCC). It measures the linear correlation between variables X and Y, ranging from -1 to 1; where 1 represents a perfect positive linear correlation, 0 represents no correlation and -1 represents perfect negative correlation. Pearson can be formally defined as:

$$PCC = \frac{cov(X,Y)}{\sigma_X \times \sigma_Y}$$

Where, *cov(X,Y)* represent the covariance between the variables *X* and *Y*. σ_X, σ_Y represent the standard deviation for *X* and *Y*, respectively.

2.2.4 Data Subsampling

Generating machine learning models that require large datasets may be computationally expensive. In each model iteration, computer resources requirements may be high, thus hampering the learning process. Instead of using the full dataset to compute model inferences, one may apply only a subsample set of the full training set. This subsampling strategy may be done either randomly or in a specific order.

While random subsampling will select N arbitrary samples from the dataset, a more specific subsample strategy will target explicit subsample groups. This strategy can be used to evaluate performance patterns arising from the use of samples with a specific characteristic.

2.2.5 Classification and Regression Analyses

In supervised learning, there are two main types of analyses. Classification and regression strategies follow the same basic supervised concept, which trains a model based on a known dataset and its known outputs. The main difference between both strategies is the predicted output type.

In a classification problem, the output variable will be a categorical value. The classifier generates different output labels based on the input data. These classifiers can be further separated into two different types, binary and multiclass classifiers. In a binary classification, there are only two possible output categories such as 0/1, yes/no

or high/low. In a multiclass classifier, the output value is one of at least three possible categories (such as small, medium and large) and, thus, generating more complex models (Harrington, 2012).

In general, a classification model can be defined by considering S to be the training set comprised of x_i and y_i pairs, where x_i is equal to the input data and y_i is equal to the output label. The classifier then generates a function $f(x)$ that maps each input data x_i to its associated label y_i.

In a regression problem, the model is trained to generate a continuous numerical outcome. Since the predictive model generates a numerical output, model performance must be reported as an error, calculated as the difference between observed and expected values. Regression models can be easily defined using the same concept as a classification task, with the main difference being that y_i assumes a numeric value instead of a discrete label (Harrington, 2012).

2.2.6 Machine Learning Algorithms

The following algorithms were used in this work: (i) Random Forest, (ii) Support vector Machine, (iii) Gradient Boosting Machine, (iv) Naïve Bayes and (v) K-Nearest Neighbors. These techniques were selected due to their frequent use as machine learning algorithms in the literature and are described below.

2.2.6.1 Random Forest (RF)

Random forest (RF) is a supervised learning algorithm where the generated "forest" is a combination of decision trees (ensemble strategy) built in a way that each tree depends on the values of a random vector each time a split occurs. The RF algorithm builds several decision trees and merges them together to obtain a more accurate and stable prediction, either a class or a discrete value.

Every time the branching occurs, the algorithm is not allowed to consider the majority of the predictors. Breiman (2001) defines a random forest as a classifier that consists of a collection of tree-structured classifiers, where each single structure is an independent identically distributed random vector. Several splits will not consider the strongest predictors, giving a greater chance to other predictors. This way, the tree is

decorrelated, thus reducing the variance over a single tree and consequently making it more reliable. (James et al., 2013). When a large number of uncorrelated models are generated (decision trees), they function as a group, protecting each other from their individual errors and thus, outperforming any of the individual decision tree models. In this study, the random forest algorithm uses randomly selected features or a combination of features at each node to grow a tree. Samples are classified by taking the most popular voted class from all the tree predictors in the forest (Breiman, 2001). For feature selection and decision tree pruning, the Gini Index criteria (Breiman et al., 1984) is used to assess the impurity of the evaluated feature regarding its class.

2.2.6.2
Support Vector Machines (SVM)

Originally introduced by Vapnik (1998), SVM method has a robust performance with sparse and noisy data, making it one of the most popular techniques for regression and classification of datasets (Chang et al., 2011). SVM method allows for a nonlinear mapping of an N-dimensional input vector to a high dimensional space separated by hyperplanes. The SVM separates a given set of binary training data with a hyperplane that is maximally distant from the closest value from the training data (Chapelle., Vapnik, Bousquet and Mukherjee, 2002), as observed in Figure 1.

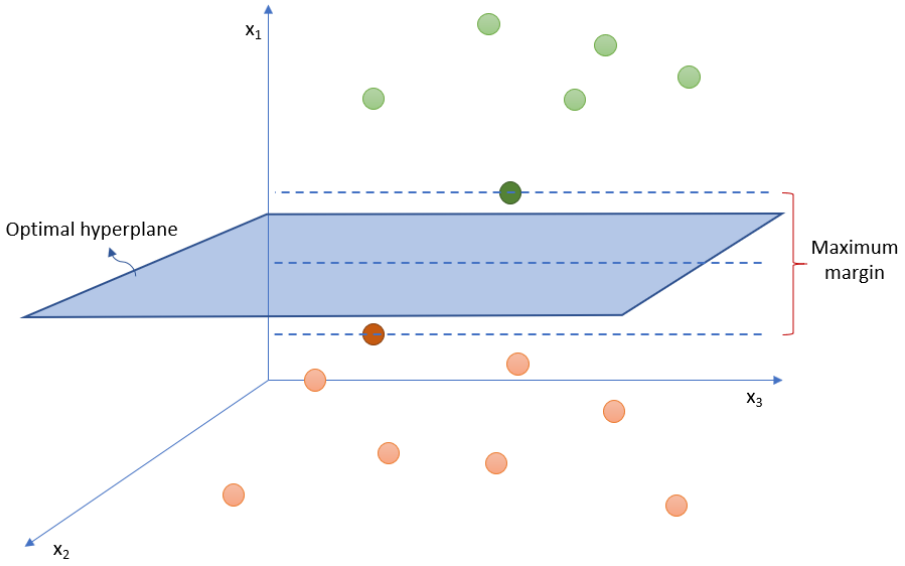


Figure 1 - SVM in R^3.

In a classification, the hyperplanes represent the decision boundaries that help classify the data points. Depending on which side of the hyperplane a data point is situated, it will be attributed to a specific class. The dimension of the hyperplane will depend on the number of existing features. In regression, the goal is to find a function (hyperplane) that approximately predicts discrete values from an input domain while allowing a tolerance error (epsilon) in the prediction.

2.2.6.3 Gradient Boosting Machine (GBM)

Decision trees method is a simple yet efficient tool that consists of dividing the input parameter space into distinct regions that follow a set of if-then rules. However, despite several advantages, decision trees have several limitations that cause them to be less accurate than other methods (Touzani, Granderson and Fernandes, 2017). To overcome these limitations, several methods have been introduced, including gradient boosting machines (GBM). The GBM's main approach is to interactively combine "weaker learners" (simpler models) to create a stronger learner that allows for improved performance (boosting strategy). This is done in a gradual and sequential manner. This method was originally introduced for classification problems by various authors (Schapire, 1990; Freund, 1995; Freund et al., 1996). A connection between boosting algorithms and loss functions was later added by Friedman, Hastie and Tibshirani (2000). Friedman (2001) extended the problem to include regression models, where the GBM is treated as an optimization algorithm that aims to minimize the loss function. This way, GBM iteratively adds a new weak learner (represented by a decision tree) at each step, aiming to continuously decrease the loss function until an optimal value is found or a stopping criterion is met.

The loss function represents a measure that indicates how well the model's coefficients are fitting the data. In a classification task the loss function evaluates the prediction performance of the classification algorithm, while in a regression, the loss function is based on the error between predicted and real values.

In this work, gradient boosting machine is applied by using the Extreme Gradient Boosting (xgboost) framework, which is similar yet more efficient (Friedman et al., 2000; Friedman, 2001). It contains both a linear model solver and tree learning

algorithms. It is capable to perform parallel computation on a single machine, thus making this framework quicker than the traditional GBM framework (Chen et al., 2015).

2.2.6.4
Naïve Bayes Classifier (NB)

A Bayes classifier is a simple, yet efficient, method to predict the different classes of a data set. This method is based on Bayes theorem and is the simplest form of a Bayesian network, in which all attributes are considered to be independent (Zhang, 2004). Thus, the effect of an attribute value on a specific class has no dependence on the values of the other attributes, also known as conditional independence. Leung (2007) explains the idea of how a naïve Bayesian classifier works:

1) Consider S to be the training set. There are m classes (C_1, C_2, …, C_m), each represented by an N-dimensional vector portraying the N values of attributes in that class.

2) Consider a sample X; the Bayes classifier will estimate which class X belongs to, according to the class with the highest posterior probability conditioned on X. Therefore, we find the class that maximizes $P(C_i|X)$. According to Bayes theorem, we have:

$$P(C_i|X) = \frac{P(X|C_i) \times P(C_i)}{P(X)}$$

where:

- $P(C_i)$ – probability of C_i occurring
- $P(X)$ – probability of X occurring
- $P(C_i|X)$ – probability of C_i occurring given that X has already occurred
- $P(X|C_i)$ – probability of X occurring given that C_i has already occurred

2.2.6.5
K-Nearest Neighbors (KNN)

KNN is a classifier method that is commonly used in decision procedures that classify a variable into a category of the nearest sample in the training set (Cover T & Hart P, 1967). Considering a test observation x_i and a positive integer K, the classifier identifies the K-neighbors in the training data closest to x_i. It then estimates the fraction

of points of this group whose response values are equal to j (James et al., 2013). This idea is also applicable to regressions. A classification example based on Figure 2 is provided as follows.

Assuming two existing categories (blue star and red triangle), an individual would like to classify a new variable (black square). We may consider the following rules:

- K=1 → 1-NN rule decides that the new variable belongs to the category of the nearest neighbor and ignores all others.
- K=2 or more → K-NN rule decides that the new variable belongs to the category of the majority of votes of the k neighbors.

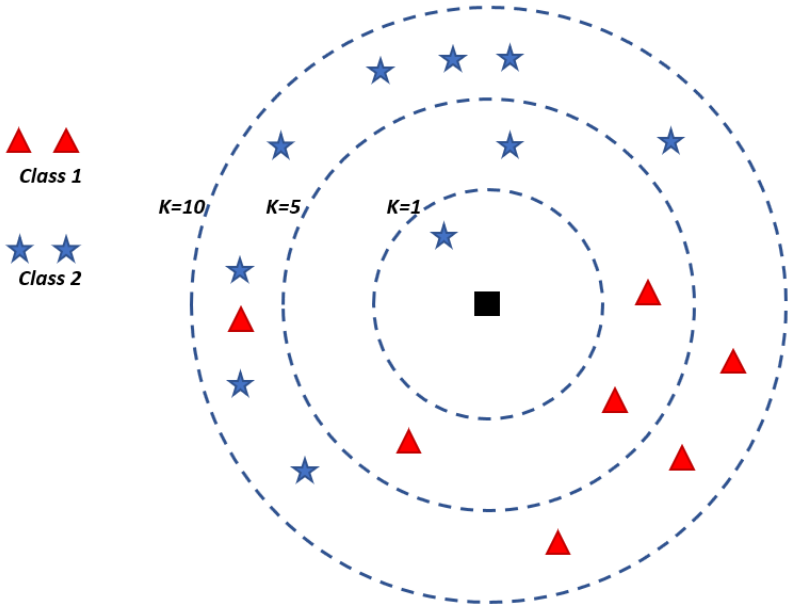


Figure 2 - KNN classification technique. Adapted from (Cover and Hart, 1967).
1-NN: The square will be classified as a blue star. 5-NN: The square will be classified as a red triangle. 10-NN: The square will be classified as a blue star.

2.2.7 Evaluation Metrics

When evaluating a model, evaluating algorithm performance is an essential step in an effective analysis. A key aspect of evaluation metrics is their ability to

discriminate between different model results, aiding our quest in the creation and selection of a model which provides high accuracy when predicting future data.

2.2.7.1 Classification

When assessing the outputs from a classification model, we can generate a confusion matrix. It is a M x M matrix, where M corresponds to the number of existent classes being predicted. In this research, M=2, thus, obtaining a 2 x 2 matrix (Figure 3).

		Real Values		
		Positive	Negative	
Predicted Values	Positive	TP	FP	Positive Predictive Value
	Negative	FN	TN	Negative Predictive Value
		Recall	Specificity	

Figure 3 – Confusion Matrix.

Where:

- $TP = True\ Positive$
- $TN = True\ Negative$
- $FP = False\ Positive$
- $FN = False\ Negative$

From this matrix, we are able to obtain specific metrics defined below.

- ***Accuracy (ACC) :*** the proportion of the total number of predictions that were correct.

$$ACC = \frac{TP + TN}{TP + TN + FN + FP}$$

- ***Positive Predictive Value (PPV):*** the proportion of positive cases that were correctly identified.

$$PPV = \frac{TP}{TP + FP}$$

- ***Negative Predictive Value (NPV):*** the proportion of negative cases that were correctly identified.

$$NPV = \frac{TN}{TN + FN}$$

- ***Recall:*** the proportion of actual positive cases which are correctly identified.

$$Recall = \frac{TP}{TP + FN}$$

- ***Specificity:*** the proportion of actual negative cases which are correctly identified.

$$Specificity = \frac{TN}{TN + FP}$$

When we aim for the best precision and recall at the same time, we evaluate the F1 score, which is the harmonic mean of the precision and recall values for the given classification task (Forman, 2003).

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

When predicting a binary (2 classes) problem, another evaluation strategy is also possible. The Matthews Correlation Coefficient (MCC) consists of regarding the predicted and real values as two different variables and then compute their correlation coefficient. The higher the correlation, better the prediction. MCC ranges from -1 to 1, where 1 corresponds to a perfect positive correlation, thus the model is predicting well (Baldi et al., 2000).

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

Lastly, the area under the receiver operating characteristic curve (AUC) metric is one of the most popular metrics when evaluating a binary classification problem. The receiver operating characteristic (ROC) curve represents a plot between the false positive rate (1-Specificity) and the True Positive rate (also known as Sensitivity, portraying the performance of a classification model at all possible classification probability thresholds. AUC then measures the area under the ROC curve, ranging in values from 0 to 1. An AUC value of 1 represents a model whose predictions are 100% correct and an AUC value of 0 represents a model with 100% wrong predictions (Fan et al., 2006).

2.2.7.2 Regression

When designing a regression model, we are attempting to reduce the error between predicted and expected values. We use a function, called the cost function, to measure this error.

The Mean Absolute Error (MAE) is the average of the absolute differences between the expected value and the model predicted value. It measures how different the predicted value was from the actual value (Hyndman, 2006).

$$MAE = \frac{\sum_{i=1}^{N} |o - e|}{N}$$

Where, $e = expected\ value$, $o = observed\ value$ and $N = sample\ size$.

The Root Mean Squared Error (RMSE) is the root-squared of the average difference between the real value and the predicted value. The root square ensures robust results, preventing positive and negative differences from cancelling each other out (Hyndman, 2006).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(e - o)^2}{N}}$$

Where, $e = expected\ value$, $o = observed\ value$ and $N = sample\ size$.

Another evaluation metric is R-squared, which is a statistical measure that represents the proportion of variance in the dependent variable (predicted value) that is predictable from the independent variables (input data). In other words, it represents how well fitted is the regression model. The closer the r-squared value is to 1, the better fitted is the model (Cameron and Windmeijer, 1997).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(e - o)^2}{\sum_{i=1}^{N}(e - a)^2}$$

Where, $e = expected\ value$, $N = sample\ size$, $o = observed\ value$ and $a = average\ expected\ value$

The last evaluation regression metric used in this research is Spearman's rank correlation coefficient (SCC). SCC is a nonparametric metric which measures the association strength between two variables on an ordinal scale. This statistical measure measures how much ranked variables are associated based on an increasing or decreasing relationship (monotonic function). Spearman's rho, which is also denoted, ranges between -1 to 1, where a 1 value indicates a perfect association of ranks, a 0

value indicates no association and a -1 value indicates a perfect negative association of ranks (Spearman, 1961).

$$SCC = \frac{6 \times \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}$$

Where, $d = difference\ in\ paired\ ranks$ and $N = sample\ size$.

2.3 Literature Review

Many computational methods have been proposed to predict anticancer drug sensitivity based on genetic, genomic or epigenomic features of cancer samples. The most common approach is to generate a drug-specific model, which is independently trained using molecular observations and drug-response data from cancerous samples tested with each drug individually. Linear-regression based, drug-specific models have been developed using gene expression data (Barretina et al., 2012; Geeleher et al., 2014; Iorio et al., 2016) or a combination of gene expression data and other genomic data types, such as copy number alterations and DNA methylation (Chen and Sun, 2017). Non-linear models using a single data type or multiple data types have also been proposed, including artificial neural networks, random forests, support vector machines (SVM), kernel regression, latent and Bayesian approaches, attractor landscape analysis of network dynamics, unsupervised pathway activity models, and recommender systems (Costello et al., 2014; Dong et al., 2015; Zhang et al., 2015; Ammad-ud-din et al., 2016; Corte's-Ciriano et al., 2016; Gupta et al., 2016; Ammad-ud-din et al., 2017; Choi et al., 2017; Rahman et al., 2017; Ali et al., 2018; Chang et al., 2018; Dhruba et al., 2018; Ding et al. 2018; Huang et al., 2018; Suphavilai et al., 2018; Wang et al., 2019; Xu et al., 2019; Emdadi et al., 2020). Transfer-learning techniques have also been proposed to improve drug-response prediction performance for one type of cancer by incorporating data from other types of cancer (Turki et al., 2017; Zhu et al., 2020). Drug response information has also been modeled in combination with chemical drug properties using elastic net regression, support vector machines, regularized matrix factorization and manifold Learning (Menden et al., 2013; Yuan et al., 2016; Wang et al., 2017; Moughari et al., 2020; Su et al., 2020).

Most recent cell-line studies have emphasized the potential to predict drug responses based on gene-expression profiles (Costello et al., 2014; Yuan et al., 2014; Zhao et al., 2015; Chiu et al., 2019; Parca et al., 2019).

This dissertation aims to expand existing research to include new prediction strategies and processes. By exploring additional machine learning and data analytics methods, we aim to contribute to new state of the art approaches when predicting drug efficacy.

In this research, we explore the effect of artificial subsampling the data in varying proportions. We show the impact on predictive models and its performance when training on relatively extreme outcomes. This subsampling approach had yet to be tested and yielded great results. Classification models portrayed improved outcomes when applying this subsampling technique.

This research also explores the effect of applying a semi-supervised model in genomic data. We evaluate possible changes in several hyperparameters within model creation, further understanding algorithm patterns and behavior within this new strategy. As a semi-supervised learning approach becomes more popular in healthcare data analytics, we hope to contribute to general modeling knowledge when dealing with this new prediction model setup.

3
Case Studies

3.1
Case Study 1 - Predicting drug sensitivity of cancer cells based on DNA methylation levels

3.1.1
Introduction

In this study, we use DNA methylation profiles from preclinical samples to model drug responses for eight anti-cancer drugs. We compare the performance of five classification algorithms and four regression algorithms that encompass a diverse range of methodologies, including tree-based, probability-based, kernel-based, ensemble-based and distance-based approaches. We use classical algorithms as a way to establish a performance baseline against which other algorithms might be compared when working with DNA methylation profiles. For regression, we predict IC_{50} values directly. For classification, we use discretized IC_{50} values. For both types of algorithm, we artificially subsample the data to varying degrees to evaluate whether training models based on relatively extreme outcomes would yield improved performance; we assess our ability to predict drug responses using as few as 10% of the cell lines (those with the most extreme IC_{50} values). An underlying motivation of this approach was to decrease data-generation costs. For example, if it could be shown that generating data for relatively few (extreme) responders performs as well as or better than generating data for responders across the full range of response values, cost savings may result. Perhaps surprisingly, the classification algorithms performed best when only 10-20% of the cell lines were used. The regression algorithms performed best when we trained the models using the full range of drug-response values, although this depended on the performance metrics we used. Finally, we derived classification models from the cell-line data and predicted drug responses for TCGA patients. In most cases, the models failed to generalize effectively; however, predictions by the Random Forests algorithm were significantly correlated with Temozolomide responses for low-grade gliomas.

3.1.2 Methods

The GDSC database contains data for human cell lines derived from common and rare types of adult and childhood cancers. GDSC provides multiple types of molecular data for these cell lines in addition to response values for 265 anti-cancer drugs. In this work, we used database version GDSC1, which includes data for 987 cell lines curated between 2010 and 2015 (Iorio et al., 2016). Drug responses were measured as the natural log of the fitted IC_{50} value. The more sensitive the cell line, the lower the IC_{50} value for any given drug. We developed machine-learning models of drug response using DNA methylation data from GDSC1 that had been preprocessed and summarized as gene-level *beta* values (Iorio et al., 2016); these values ranged between 0 and 1 (higher values indicated relatively high methylation for a given gene). We used all available methylation regions, represented by gene-level summarized values, as input to the classification and regression algorithms.

For external validation, we used DNA methylation data and clinical drug-response values from TCGA. We selected eight drugs that were administered to TCGA patients and present in GDSC: Gefitinib, Cisplatin, Docetaxel, Doxorubicin, Etoposide, Gemcitabine, Paclitaxel and Temozolomide. These drugs represent a variety of molecular mechanisms, including DNA crosslinking, microtubule stabilization and pyrimidine anti-metabolization. Aside from Gefitinib, which we used for model optimization on GDSC data, these drugs were associated with the largest number of patient drug-response values in TCGA (Huang et al., 2020). GDSC provides DNA methylation values for 6,035 TCGA samples that had been preprocessed using the same pipeline as the GDSC samples. We obtained drug-response data for TCGA patients from (Ding et al., 2016).

Cell lines with missing IC_{50} values were excluded on a per-drug basis; thus, sample sizes differed across the drugs. We applied Z-score normalization on a per-gene basis across all samples in GDSC and TCGA. Next, we used ComBat (Leek et al., 2020) to adjust for systematic differences between the two datasets (GDSC and TCGA); we also specified cell type as a covariate to adjust for methylation patterns associated with this factor.

We started with a classification analysis. Classification algorithms are widely available, and their predictions are intuitive to interpret—they assign probabilities to each sample for each class. To enable classification for the GDSC cell lines, we discretized the IC_{50} values into "low" and "high" values. However, the choice of a threshold for distinguishing low and high values was

necessarily arbitrary. Initially, we used the median IC_{50} value across all cell lines as a threshold. However, cell lines with an IC_{50} just above or below this threshold naturally showed very little difference in their drug responses, even though they were assigned to different classes. In contrast, cell lines with extreme IC_{50} values (far from the threshold) had much more distinct drug responses. To investigate the effects of using a threshold to discretize the IC_{50} values for classification, we used subsampling. We created 10 different scenarios that included increasing percentages of the overall data. First, we sorted the samples by IC_{50} value in ascending order. For the first scenario, we evaluated cell lines with the 5% lowest and 5% highest IC_{50} values (10% of the total data). In the next scenario, we evaluated cell lines with the 10% lowest and 10% highest IC_{50} values (20% of the total data), and so on. The last scenario included all the data, where the lowest 50% were considered to have low IC_{50} values and the highest 50% were considered to have high values (Figure S1). For the regression analysis, we followed a similar process for subsampling, but retained the continuous nature of the IC_{50} values.

For both classification and regression, we used the Random Forests (tree-based) (Breiman, 2001), Support Vector Machines (kernel-based) (Vapnik, 1998), Gradient Boosting Machines (ensemble-based) (Breiman, 1997) and k-Nearest Neighbors (distance-based) (Cover and Hart, 1967) algorithms. We used the Naïve Bayes (probability-based) (Maron, 1961) algorithm for classification, but not for regression, because this algorithm is only designed for classification analyses. We performed the analyses using the R programming language (R Core Team, 2019) and Rstudio (https://rstudio.com). The machine-learning algorithms were implemented in the following R packages: mlr (Bischl et al., 2016), e1071 (Meyer et al., 2019), xgboost (Chen et al., 2015), randomForest (Liaw and Wiener, 2002), and kknn (Schliep and Hechenbichler, 2016).

Using the GDSC cell-line data, we sought to select the best hyperparameters for each algorithm via nested cross validation. We used the *mlr* package (Bischl et al., 2016) to randomly assign the cell lines to 10 outer folds and 5 inner folds (per outer fold). For each combination of algorithm and data-subsampling scenario, we evaluated the performance of all hyperparameter combinations (Table 1) using the inner folds; we used MMCE (Mean Misclassification Error) (Schiffner et al., 2016) for classification and MSE (Mean Squared Error) (Hyndman, 2006) for regression as evaluation metrics in the inner folds (defaults in *mlr*). For the outer-fold predictions, we assessed performance for predicting drug responses using several performance metrics. This enabled us to evaluate how consistently the algorithms performed. For the classification analysis,

we used accuracy (1 - MMCE), area under the receiver operating characteristic curve (AUC) (Fan et al., 2006), F1 measure (Forman, 2003), Matthews correlation coefficient (MCC) (Baldi et al., 2000), recall and specificity. For the regression analysis, we used Mean Absolute Error (MAE), Root Mean Square Error (RMSE) (Hyndman, 2006), R-squared coefficient of determination (Cameron and Windmeijer, 1997) and Spearman's rank correlation coefficient (SCC) (Spearman, 1961).

Table 1: Descriptions of the algorithms we tested and hyperparameters that we evaluated via nested cross validation. Hyperparameter optimization was performed for all tested algorithms. All parameter combinations for each algorithm were evaluated via nested cross validation; optimal combinations were then used for outer-fold predictions.

Algorithm	Hyperparameters	Definition	Tested Values
classif.svm and regr.svm	1. Kernel	The kernel function used to transform data to higher-dimensional spaces and then become linearly separable.	Linear; Radial; Polynomial; Sigmoid
	2. Cost	The regularization parameter in the cost function, to penalize missing classifications.	0.1; 1; 10; 100
	3. Scale	Whether the variables should be scaled.	True; False
classif.randomForest and regr.randomForest	1. Ntree	The number of trees to grow.	100; 500; 1000
	2. Nodesize	Minimum size of terminal nodes.	1; 3; 5; 7
	3. Importance	Whether the importance of predictors should be assessed.	True; False
classif.kknn and regr.kknn	1. K	The number of neighbors considered.	3; 7; 10
	2. Scale	Whether to scale variables to have equal standard deviation.	True; False
classif.naiveBayes	1. Laplace	The amount of Laplace (additive) smoothing.	0; 1; 5; 10

classif.xgboost	1. Nround	The maximum number of boosting iterations.	100; 250; 500
	2. Max_depth	The maximum depth of a tree.	1; 5; 10
	3. Eta	How much the contribution of each tree is scaled to the overall approximation, to control the learning rate.	0.1; 0.3; 0.5
regr.xgboost	1. Nround	The maximum number of boosting iterations.	100; 250; 500
	2. Eta	How much the contribution of each tree is scaled to the overall approximation, to control the learning rate.	0.1; 0.3; 0.5

After assessing the algorithms separately for the classification and regression approaches, we evaluated the predictive ability of these two types of tasks against one another. We calculated the Spearman correlation coefficient as a nonparametric measure of the concordance between the predicted probabilities (classification algorithms) and predicted IC_{50} values (regression algorithms).

For the classification and regression analyses, we used feature selection to identify genes deemed to be most informative. We performed an information-gain analysis, assigning an importance score to each feature (gene). More specifically, we estimated the relative importance of each gene based on the conditional entropy of the class variable with respect to that gene. Entropy measures the amount of randomness in the information. Thus, higher information gain implies lower entropy. This analysis was implemented using the FSelectorRcpp package (Zawadzki and Kosinski, 2020). To assess the functional relevance of the top-ranked genes, we used a gene-set overlap technique implemented in the Molecular Signatures Database 3.0 (Liberzon et al., 2011). As candidate gene sets, we included the *C2 (curated gene sets)*, *C4 (computational genes sets)*, and *C6 (oncogenic signature gene sets*). We used a False Discovery Rate q-value threshold of 0.05.

For additional validation, we trained classification models based on discretized drug responses in the GDSC cell lines and then predicted patient drug responses using tumor data from TCGA. These patient responses were based on clinical data, having no direct relation to IC_{50}

values. Because the patient-response values were categorical in nature, we only performed classification for these data. We used nested cross validation to perform hyperparameter optimization using the GDSC (training) data. To evaluate the relationship between the predicted labels and actual clinical responses, we calculated Spearman's rank correlation coefficient and a corresponding p-value for each combination of algorithm and data-subsampling scenario; then we used the Benjamini-Hochberg False Discovery Rate to adjust for multiple tests (Benjamini and Hochberg, 1995).

3.1.3
Results

Using data from 987 cell lines, we used machine-learning algorithms to evaluate the potential to predict cytotoxic responses based on genome-wide, DNA methylation profiles. Second, we examined which genes were most predictive of these responses. Finally, we evaluated the feasibility of predicting clinical responses in humans based on models derived from cell-line data.

3.1.3.1
Classification analysis using cell-line data

We collected DNA methylation data and IC_{50} response values for eight drugs from the GDSC repository. In our initial analysis, we aimed to predict categories (classes) of drug sensitivity. These categories represented whether each cell line exhibited a "low" or "high" response to each drug, corresponding to relatively low or high IC_{50} values, respectively. This categorization facilitated a simplified yet intuitive interpretation of the treatment outcomes and enabled us to use classification algorithms, which have been implemented for a broader range of algorithmic methodologies than regression algorithms.

Before performing classification, we categorized each cell line on a per-drug basis, according to whether its IC_{50} value was greater than the median across all cell lines. One limitation of categorizing the cell lines in this way was that cell lines just above or below the median threshold showed a relatively small difference in IC_{50} values, even though they were assigned to different classes. Generally, IC_{50} values did not follow a multimodal distribution (Figure 4). Therefore, we evaluated whether classification performance could be improved by excluding cell lines with an IC_{50} value relatively close to the median, even though this would reduce the amount of data

available for training and testing. We evaluated ten scenarios that varied the number of cell lines used. In the most extreme scenario, we used methylation data for cell lines with the 5% lowest and 5% highest IC_{50} values. In describing these subsampling scenarios, we use a notation that indicates the percentage of samples on each side of the distribution as well as the algorithm type. For example, when we analyzed the samples with the 5% highest and 5% lowest IC_{50} values and employed a classification algorithm, we indicate this using "+-5%c". The equivalent scenario for regression was represented as +-5%r.

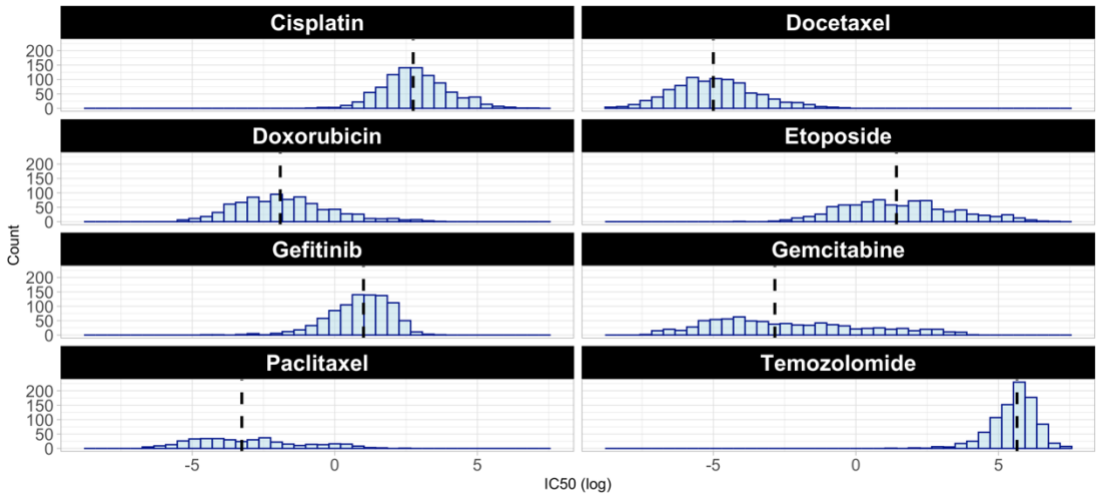


Figure 4: Histograms for each drug based on drug response (IC_{50} values) for the GDSC dataset. The black line represents the median value for each subsample across all available cell lines for each drug.

We evaluated the performance of five classification algorithms using six performance metrics (see Methods). In addition, we optimized hyperparameters via nested cross validation; Table 1 lists the hyperparameters we evaluated. Initially, we evaluated Gefitinib, an EGFR inhibitor. Overall, the algorithms performed best when relatively few cell lines (+-5%c and +-10%c) were used to train and test the models, attaining area-under-the-receiver-operating-characteristic curve (AUC) and classification-accuracy values as high as 0.93 and 0.84 (Table 2). This pattern was consistent across all five algorithms and all six metrics that we evaluated (Figure 5). However, the SVM algorithm consistently achieved higher classification performance than the other algorithms for this drug.

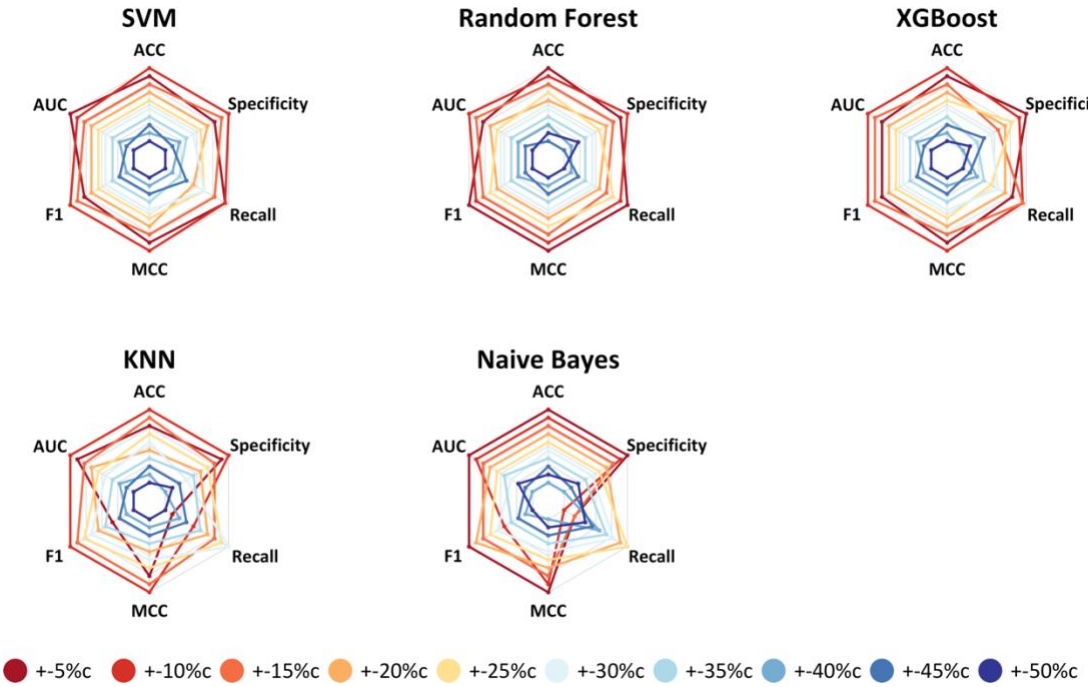


Figure 5: Gefitinib classification results across six metrics. These "spider" graphs illustrate how each classification algorithm performed in each subsampling scenario via cross validation on the GDSC cell-line data. Results that are further away from the center represent higher metric values (relatively better performance) than results closer to it. These metrics are accuracy (ACC), specificity, recall, Matthews correlation coefficient (MCC), F1 score (F1) and area under the receiver operating characteristic curve (AUC). Scenarios that used relatively few cell lines—but those with the most extreme IC_{50} values— performed best for all algorithms. Specific metric values may be found in Table 2.

Table 2: Classification results for all subsampling scenarios and algorithms for Gefitinib. Bold font indicates the best-performing combination for each metric.

Scenario	Method	ACC	AUC	F1	MCC	Recall	Specificity
+-5%c	SVM	0.82	**0.93**	0.80	0.65	0.85	0.78
+-5%c	Random Forest	0.82	0.82	**0.82**	0.66	**0.89**	0.74
+-5%c	KNN	0.72	0.84	0.67	0.45	0.63	0.80
+-5%c	XGBoost	0.77	0.83	0.75	0.54	0.76	0.78
+-5%c	Naïve Bayes	0.73	0.74	0.73	0.45	0.76	0.70

+-10%c	SVM	**0.84**	0.92	**0.82**	**0.69**	0.85	**0.83**
+-10%c	Random Forest	0.80	0.89	0.79	0.61	0.84	0.77
+-10%c	KNN	0.75	0.86	0.71	0.49	0.68	**0.83**
+-10%c	XGBoost	0.78	0.88	0.77	0.56	0.80	0.75
+-10%c	Naïve Bayes	0.68	0.69	0.66	0.35	0.68	0.67
+-15%c	SVM	0.81	0.86	0.81	0.63	0.83	0.79
+-15%c	Random Forest	0.75	0.84	0.75	0.50	0.78	0.71
+-15%c	KNN	0.72	0.79	0.71	0.45	0.71	0.73
+-15%c	XGBoost	0.74	0.83	0.75	0.51	0.80	0.68
+-15%c	Naïve Bayes	0.66	0.66	0.68	0.32	0.76	0.56
+-20%c	SVM	0.75	0.83	0.75	0.51	0.77	0.73
+-20%c	Random Forest	0.72	0.80	0.73	0.44	0.76	0.69
+-20%c	KNN	0.68	0.78	0.69	0.37	0.71	0.66
+-20%c	XGBoost	0.72	0.80	0.73	0.44	0.76	0.69
+-20%c	Naïve Bayes	0.64	0.64	0.68	0.28	0.79	0.48
+-25%c	SVM	0.74	0.81	0.75	0.48	0.78	0.70
+-25%c	Random Forest	0.72	0.79	0.74	0.45	0.79	0.66
+-25%c	KNN	0.70	0.77	0.71	0.41	0.73	0.68
+-25%c	XGBoost	0.72	0.79	0.72	0.43	0.74	0.70
+-25%c	Naïve Bayes	0.60	0.62	0.67	0.23	0.80	0.41
+-30%c	SVM	0.72	0.78	0.74	0.45	0.78	0.66
+-30%c	Random Forest	0.69	0.75	0.70	0.38	0.74	0.63
+-30%c	KNN	0.68	0.75	0.70	0.37	0.74	0.63
+-30%c	XGBoost	0.69	0.77	0.70	0.38	0.74	0.63
+-30%c	Naïve Bayes	0.60	0.60	0.66	0.21	0.79	0.41
+-35%c	SVM	0.68	0.76	0.70	0.37	0.72	0.64
+-35%c	Random Forest	0.67	0.73	0.69	0.34	0.73	0.60
+-35%c	KNN	0.67	0.71	0.68	0.34	0.70	0.64
+-35%c	XGBoost	0.66	0.70	0.67	0.32	0.69	0.62
+-35%c	Naïve Bayes	0.59	0.60	0.66	0.20	0.79	0.40
+-40%c	SVM	0.67	0.73	0.68	0.35	0.71	0.63
+-40%c	Random Forest	0.65	0.71	0.67	0.30	0.71	0.58
+-40%c	KNN	0.60	0.66	0.61	0.21	0.64	0.57
+-40%c	XGBoost	0.65	0.70	0.65	0.29	0.68	0.61
+-40%c	Naïve Bayes	0.57	0.58	0.64	0.16	0.78	0.36
+-45%c	SVM	0.67	0.72	0.69	0.35	0.72	0.62
+-45%c	Random Forest	0.64	0.70	0.66	0.30	0.71	0.57
+-45%c	KNN	0.63	0.66	0.64	0.26	0.66	0.60
+-45%c	XGBoost	0.65	0.69	0.65	0.31	0.67	0.62
+-45%c	Naïve Bayes	0.58	0.59	0.65	0.18	0.78	0.39

+-50%c	SVM	0.65	0.70	0.66	0.30	0.70	0.60
+-50%c	Random Forest	0.64	0.69	0.66	0.29	0.70	0.59
+-50%c	KNN	0.60	0.65	0.60	0.20	0.61	0.59
+-50%c	XGBoost	0.63	0.68	0.64	0.27	0.65	0.62
+-50%c	Naïve Bayes	0.58	0.59	0.64	0.17	0.77	0.39

When evaluating the seven remaining drugs, we continued to see a trend in which using a relatively small proportion of the data resulted in better classification performance. For Cisplatin, Docetaxel, Doxorubicin, and Etoposide, the best performance was attained for +-5%c and +-10%c, and the best-performing algorithms were always SVM or Random Forests (RF) (Tables S1-S7). In contrast, for Gemcitabine, the highest AUC value (0.82) was obtained for +-15%c (SVM algorithm). For Paclitaxel, the Random Forests algorithm performed best for +-10%c (AUC = 0.75). The overall highest AUC value was attained for Docetaxel (0.97, +-10%c, Random Forests and SVM). Figures S2-S8 illustrate these results across all algorithms, metrics, and drugs and show that generally the top-performing algorithms were consistent across all metrics, although these patterns were less consistent in scenarios where the highest AUC values were lower than 0.80.

To further analyze combinations of subsampling scenarios and classification algorithms, we ranked the AUC values for all combinations and for each drug (where the lowest rank was considered best and represented the highest AUC value). Subsequently, we calculated the average AUC rank across all drugs. The best performance was attained for +-10%c (SVM) and +-10%c (Random Forests), achieving average ranks of 4.75 and 5.13, respectively (Table 3). When we evaluated the minimum, mean, and maximum AUC values for each combination of drug and algorithm, Docetaxel attained the best overall performance (Table 4).

Table 3: Summary of AUC values across all combinations of subsampling scenario and algorithm. We ranked the AUC values for each combination and then calculated the average rank across the combinations (lower ranks imply better performance). In addition, this table lists the minimum, maximum, and standard deviation AUC value across the combinations.

Scenario	Method	Average AUC Rank	Min AUC Value	Max AUC Value	Standard Deviation AUC Value

+-10%c	Random Forest	4.75	0.72	0.97	0.08
+-10%c	SVM	5.13	0.65	0.97	0.10
+-5%c	SVM	5.14	0.74	0.95	0.08
+-15%c	XGBoost	7.50	0.68	0.94	0.09
+-15%c	SVM	7.63	0.66	0.93	0.10
+-5%c	Random Forest	7.71	0.77	0.93	0.06
+-5%c	XGBoost	7.86	0.69	0.96	0.09
+-15%c	Random Forest	9.13	0.70	0.92	0.08
+-10%c	XGBoost	10.13	0.58	0.94	0.12
+-20%c	SVM	10.75	0.66	0.92	0.09
+-25%c	SVM	11.25	0.70	0.90	0.07
+-10%c	KNN	12.75	0.67	0.91	0.09
+-5%c	KNN	13.14	0.69	0.92	0.07
+-25%c	XGBoost	15.38	0.67	0.89	0.07
+-20%c	XGBoost	15.88	0.65	0.91	0.09
+-20%c	Random Forest	16.00	0.64	0.91	0.08
+-30%c	SVM	16.25	0.68	0.86	0.06
+-25%c	Random Forest	16.50	0.70	0.88	0.07
+-35%c	SVM	19.00	0.68	0.84	0.05
+-30%c	XGBoost	20.50	0.62	0.87	0.08
+-30%c	Random Forest	20.63	0.65	0.85	0.07
+-15%c	KNN	21.25	0.61	0.87	0.10
+-20%c	KNN	23.38	0.63	0.88	0.09
+-35%c	Random Forest	24.13	0.65	0.82	0.06
+-35%c	XGBoost	25.25	0.61	0.83	0.07
+-40%c	SVM	26.00	0.66	0.81	0.05
+-25%c	KNN	26.63	0.62	0.85	0.08
+-30%c	KNN	26.63	0.64	0.83	0.07
+-40%c	XGBoost	26.88	0.62	0.79	0.05

+-45%c	SVM	28.25	0.65	0.77	0.04
+-5%c	Naïve Bayes	28.57	0.64	0.79	0.05
+-40%c	Random Forest	28.63	0.65	0.79	0.05
+-35%c	KNN	32.25	0.62	0.78	0.06
+-50%c	SVM	32.38	0.64	0.76	0.04
+-45%c	XGBoost	32.63	0.61	0.76	0.05
+-50%c	XGBoost	32.63	0.59	0.78	0.06
+-45%c	Random Forest	33.00	0.62	0.77	0.05
+-10%c	Naïve Bayes	34.75	0.57	0.81	0.09
+-50%c	Random Forest	36.38	0.62	0.76	0.05
+-40%c	KNN	37.50	0.62	0.75	0.05
+-45%c	KNN	39.00	0.60	0.72	0.04
+-15%c	Naïve Bayes	41.13	0.57	0.75	0.07
+-50%c	KNN	41.88	0.59	0.71	0.04
+-20%c	Naïve Bayes	43.38	0.54	0.76	0.07
+-25%c	Naïve Bayes	44.13	0.57	0.72	0.06
+-30%c	Naïve Bayes	44.25	0.57	0.71	0.05
+-35%c	Naïve Bayes	45.50	0.57	0.68	0.04
+-40%c	Naïve Bayes	47.13	0.57	0.67	0.04
+-45%c	Naïve Bayes	47.63	0.56	0.66	0.04
+-50%c	Naïve Bayes	48.75	0.55	0.64	0.04

Table 4: Minimum, mean and maximum AUC value for each combination of drug and algorithm, averaged across all subsampling scenarios.

Drug	Method	Min	Mean	Max
Gefitinib	SVM	0.70	0.80	0.93
Gefitinib	Random Forest	0.69	0.77	0.89
Gefitinib	Naïve Bayes	0.58	0.63	0.74
Gefitinib	KNN	0.65	0.75	0.86
Gefitinib	XGBoost	0.68	0.77	0.88

Cisplatin	SVM	0.66	0.78	0.88
Cisplatin	Random Forest	0.65	0.76	0.86
Cisplatin	Naïve Bayes	0.59	0.63	0.73
Cisplatin	KNN	0.60	0.72	0.84
Cisplatin	XGBoost	0.69	0.78	0.87
Paclitaxel	SVM	0.65	0.68	0.72
Paclitaxel	Random Forest	0.64	0.69	0.72
Paclitaxel	Naïve Bayes	0.54	0.58	0.61
Paclitaxel	KNN	0.61	0.65	0.68
Paclitaxel	XGBoost	0.58	0.67	0.73
Temozolomide	SVM	0.74	0.84	0.95
Temozolomide	Random Forest	0.73	0.82	0.90
Temozolomide	Naïve Bayes	0.63	0.69	0.76
Temozolomide	KNN	0.68	0.79	0.92
Temozolomide	XGBoost	0.74	0.83	0.93
Etoposide	SVM	0.66	0.75	0.88
Etoposide	Random Forest	0.63	0.71	0.89
Etoposide	Naïve Bayes	0.56	0.61	0.71
Etoposide	KNN	0.59	0.68	0.84
Etoposide	XGBoost	0.66	0.74	0.86
Gemcitabine	SVM	0.65	0.74	0.82
Gemcitabine	Random Forest	0.66	0.72	0.78
Gemcitabine	Naïve Bayes	0.56	0.59	0.73
Gemcitabine	KNN	0.62	0.66	0.69
Gemcitabine	XGBoost	0.67	0.73	0.79
Docetaxel	SVM	0.76	0.87	0.97
Docetaxel	Random Forest	0.76	0.86	0.97
Docetaxel	Naïve Bayes	0.64	0.72	0.81
Docetaxel	KNN	0.71	0.81	0.91
Docetaxel	XGBoost	0.76	0.87	0.96
Doxorubicin	SVM	0.64	0.70	0.80
Doxorubicin	Random Forest	0.62	0.68	0.78
Doxorubicin	Naïve Bayes	0.56	0.58	0.64
Doxorubicin	KNN	0.59	0.65	0.79
Doxorubicin	XGBoost	0.59	0.65	0.71

3.1.3.2 Regression analysis using cell-line data

We performed a regression analysis using the same DNA methylation data but with continuous IC_{50} response values for the same eight drugs. For this analysis, we applied four regression algorithms and evaluated their performance using nested cross validation and four performance metrics (RMSE, MAE, R-squared and SCC). As with the classification analysis, we performed data subsampling to evaluate the effects of using relatively extreme IC_{50} values. For Gefitinib and the MAE and RMSE metrics, all algorithms performed best when all cell lines were used to train and test the models, attaining RMSE values as low as 0.95 (lower is better, see Table 5). However, for the R-squared and SCC metrics, the +-5%r subsampling scenario resulted in the best performance in some cases. Typically, the magnitude of the differences between the original and predicted IC_{50} values was larger toward the extremes, resulting in relatively high MAE and RMSE values when middle values were excluded. In contrast, SCC is a rank-based metric, and the algorithms struggled most to differentiate between IC_{50} values toward the middle of the distribution. We observed similar patterns for the other seven drugs (Tables S8-S14).

Across all drugs and metrics, the SVM and Random Forests algorithms performed best for every combination of drug and performance metric (Figure 6). Furthermore, predictive performance was highly consistent for all metrics (Figures S9-S15). When evaluating the mean RMSE ranked values (where the lowest rank was considered best and represented the lowest RMSE value), the RF and SVM algorithms and the +-50%r scenarios performed best (Table 6), and predictions for Temozolomide were more accurate overall than those for other drugs (Table 7).

Table 5: Regression results for all combinations of subsampling scenarios and algorithms for Gefitinib. Bold font indicates the best-performing combination for each metric.

Scenario	Method	MAE	RMSE	R^2	Spearman
+-5%r	SVM	1.28	1.54	**0.50**	**0.63**
+-5%r	Random Forest	1.61	1.83	0.31	0.51
+-5%r	KNN	1.54	1.96	0.18	0.46
+-5%r	XGBoost	1.36	1.84	0.36	0.48
+-10%r	SVM	1.08	1.36	0.46	0.60
+-10%r	Random Forest	1.26	1.53	0.34	0.53

+-10%r	KNN	1.27	1.65	0.21	0.47
+-10%r	XGBoost	1.17	1.56	0.31	0.50
+-15%r	SVM	1.11	1.37	0.35	0.57
+-15%r	Random Forest	1.18	1.41	0.33	0.53
+-15%r	KNN	1.18	1.52	0.20	0.47
+-15%r	XGBoost	1.16	1.48	0.25	0.50
+-20%r	SVM	1.04	1.27	0.35	0.59
+-20%r	Random Forest	1.11	1.32	0.30	0.53
+-20%r	KNN	1.10	1.42	0.18	0.48
+-20%r	XGBoost	1.13	1.42	0.18	0.44
+-25%r	SVM	0.99	1.21	0.31	0.54
+-25%r	Random Forest	1.04	1.24	0.28	0.52
+-25%r	KNN	1.02	1.32	0.18	0.47
+-25%r	XGBoost	1.03	1.26	0.26	0.51
+-30%r	SVM	0.92	1.14	0.31	0.54
+-30%r	Random Forest	0.97	1.18	0.26	0.49
+-30%r	KNN	0.96	1.25	0.17	0.45
+-30%r	XGBoost	0.97	1.20	0.23	0.47
+-35%r	SVM	0.88	1.10	0.25	0.52
+-35%r	Random Forest	0.93	1.14	0.21	0.45
+-35%r	KNN	0.92	1.20	0.10	0.40
+-35%r	XGBoost	0.92	1.15	0.18	0.42
+-40%r	SVM	0.84	1.06	0.22	0.44
+-40%r	Random Forest	0.86	1.06	0.21	0.43
+-40%r	KNN	0.88	1.14	0.10	0.36
+-40%r	XGBoost	0.88	1.10	0.16	0.39
+-45%r	SVM	0.79	1.01	0.21	0.44
+-45%r	Random Forest	0.80	1.02	0.21	0.42
+-45%r	KNN	0.84	1.10	0.06	0.35
+-45%r	XGBoost	0.81	1.04	0.18	0.40
+-50%r	SVM	**0.73**	**0.95**	0.23	0.45
+-50%r	Random Forest	0.74	**0.95**	0.22	0.43
+-50%r	KNN	0.78	1.02	0.10	0.36
+-50%r	XGBoost	0.75	**0.95**	0.22	0.41

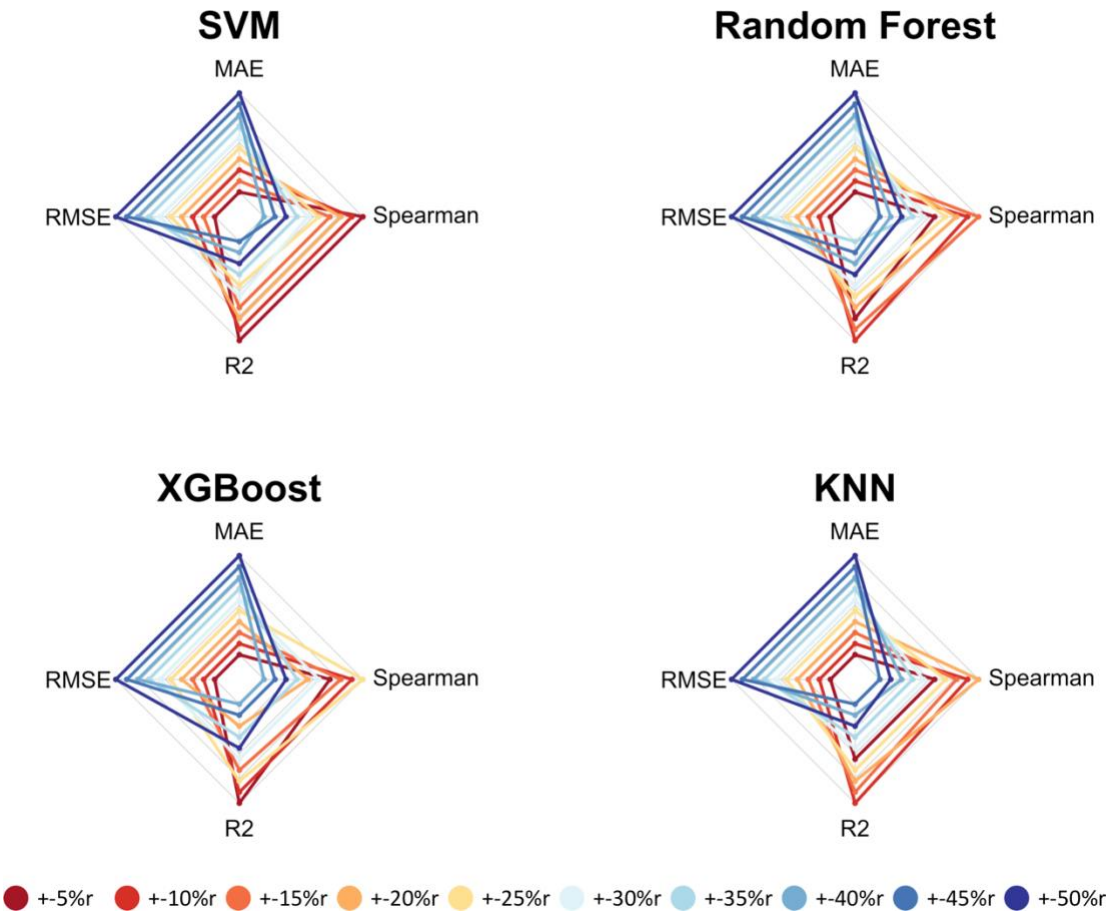


Figure 6: Gefitinib regression results across four metrics. These "spider" graphs illustrate how each regression algorithm performed in each subsampling scenario via cross validation on the GDSC cell-line data. Results that are further away from the center represent higher metric values (relatively better performance) than results closer to it. These metrics are RMSE (Root Mean Square Error), MAE (Mean Absolute Error), R-squared and Spearman correlation coefficient. Scenarios that used all cell lines performed best for all algorithms. Specific metric values may be found in Table 5.

Table 6: Average RMSE rank for all combinations of subsampling scenarios and algorithms. RMSE values were ranked for each drug and were then averaged. Lower ranks imply a better result. We also include standard deviation and the minimum and maximum RMSE values. Bold font indicates the best-performing combination for each metric.

Scenario	Method	Average RMSE Rank	Min RMSE Value	Max RMSE Value	Standard Deviation RMSE Value
+-50%r	Random Forest	**1.50**	**0.67**	**2.53**	**0.61**
+-50%r	SVM	1.75	0.68	2.56	**0.61**
+-50%r	XGBoost	2.88	0.69	2.54	**0.61**
+-45%r	SVM	4.38	0.69	2.66	0.64
+-45%r	Random Forest	4.75	0.70	2.65	0.63
+-50%r	KNN	6.75	0.73	2.70	0.64
+-45%r	XGBoost	6.88	0.73	2.67	0.64
+-40%r	SVM	8.00	0.72	2.77	0.67
+-40%r	Random Forest	8.88	0.73	2.78	0.67
+-45%r	KNN	10.50	0.78	2.82	0.66
+-40%r	XGBoost	11.00	0.78	2.82	0.67
+-35%r	SVM	11.75	0.76	2.92	0.71
+-35%r	Random Forest	13.00	0.76	2.94	0.71
+-40%r	KNN	13.88	0.81	2.94	0.69
+-30%r	SVM	15.38	0.80	3.07	0.75
+-35%r	XGBoost	15.88	0.81	3.02	0.74
+-30%r	Random Forest	16.63	0.79	3.09	0.75
+-35%r	KNN	18.75	0.84	3.09	0.73
+-30%r	XGBoost	19.25	0.84	3.17	0.77
+-25%r	SVM	19.88	0.80	3.25	0.81
+-25%r	Random Forest	21.25	0.84	3.33	0.82
+-30%r	KNN	21.63	0.88	3.28	0.79
+-20%r	SVM	23.13	0.82	3.50	0.89
+-25%r	XGBoost	23.25	0.88	3.40	0.82
+-25%r	KNN	25.13	0.92	3.48	0.84
+-20%r	Random Forest	25.75	0.90	3.55	0.89
+-15%r	SVM	26.88	0.86	3.57	0.91
+-20%r	XGBoost	28.75	0.93	3.71	0.92
+-20%r	KNN	29.50	0.97	3.82	0.94
+-15%r	Random Forest	29.63	0.95	3.71	0.92
+-10%r	SVM	30.25	0.93	3.94	1.02
+-15%r	KNN	32.50	1.03	4.07	1.01
+-15%r	XGBoost	33.13	1.06	4.00	1.02
+-10%r	Random Forest	33.63	1.02	4.14	1.04
+-10%r	XGBoost	35.38	1.11	4.37	1.13
+-5%r	SVM	36.25	1.16	4.15	1.04

+-10%r	KNN	36.25	1.16	4.51	1.11
+-5%r	Random Forest	37.50	1.28	4.30	1.01
+-5%r	KNN	38.88	1.35	4.47	1.01
+-5%r	XGBoost	39.75	1.49	4.79	1.28

Table 7: Minimum, mean and maximum RMSE value for each drug and algorithm combination, averaged across all subsampling scenarios.

Drug	Method	Min	Mean	Max
Gefitinib	SVM	0.95	1.20	1.54
Gefitinib	Random Forest	0.95	1.27	1.83
Gefitinib	KNN	1.02	1.36	1.96
Gefitinib	XGBoost	0.95	1.30	1.84
Cisplatin	SVM	1.04	1.36	2.14
Cisplatin	Random Forest	1.04	1.38	2.11
Cisplatin	KNN	1.10	1.44	2.16
Cisplatin	XGBoost	1.05	1.43	2.16
Paclitaxel	SVM	1.87	2.50	3.56
Paclitaxel	Random Forest	1.84	2.50	3.58
Paclitaxel	KNN	1.95	2.64	3.74
Paclitaxel	XGBoost	1.91	2.75	4.74
Temozolomide	SVM	0.68	0.82	1.16
Temozolomide	Random Forest	0.67	0.86	1.28
Temozolomide	KNN	0.73	0.95	1.35
Temozolomide	XGBoost	0.69	0.93	1.49
Etoposide	SVM	1.80	2.30	2.93
Etoposide	Random Forest	1.84	2.36	2.93
Etoposide	KNN	1.94	2.49	3.03
Etoposide	XGBoost	1.89	2.48	3.28
Gemcitabine	SVM	2.56	3.24	4.15
Gemcitabine	Random Forest	2.53	3.30	4.30
Gemcitabine	KNN	2.70	3.52	4.51
Gemcitabine	XGBoost	2.54	3.45	4.79
Docetaxel	SVM	1.22	1.47	1.99
Docetaxel	Random Forest	1.23	1.52	2.14
Docetaxel	KNN	1.34	1.69	2.74
Docetaxel	XGBoost	1.25	1.55	2.23
Doxorubicin	SVM	1.59	2.14	3.17

Doxorubicin	Random Forest	1.58	2.16	3.28
Doxorubicin	KNN	1.69	2.24	3.21
Doxorubicin	XGBoost	1.61	2.25	3.51

3.1.3.3 Classification and regression evaluation

As a way to compare the predictions of the classification versus regression algorithms, we used SCC as a nonparametric measure. For the classification algorithms, we calculated the SCC between the probabilistic predictions that these algorithms produced and the original IC_{50} values. For the regression algorithms we used the SCC values that quantified the correlation between the predicted and actual IC_{50} values. Then for each combination of subsampling scenario and drug, we compared the SCC for the same algorithm types against each other (Figure 7). These coefficients were strongly correlated with each other, illustrating that the classification and regression algorithms typically ranked the patients similarly in relation to the original IC_{50} values.

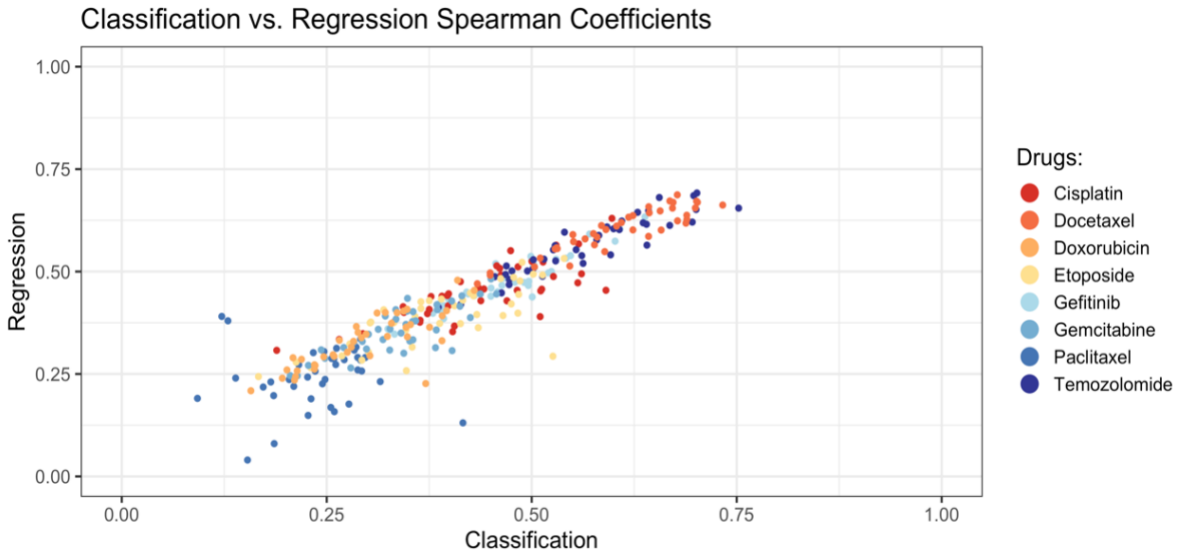


Figure 7: Spearman correlation coefficient results for classification algorithms (predicted probabilities) and regression algorithms (predicted IC_{50} values). For the classification analyses, we calculated the Spearman correlation coefficient between the predicted probabilities and the original IC_{50} values. These are represented on the x-axis. The y-axis represents the Spearman coefficients from the regression analyses. Each dot reflects results for a particular combination of drug, subsampling scenario and algorithm.

3.1.3.4 Informative genes for predicting cell-line responses

The DNA methylation assays target CpG islands associated with genes across the genome. After identifying analysis scenarios that resulted in optimal performance for classification and regression, we used feature ranking to identify genes that were most informative in these scenarios. For the classification analysis, we focused on the +-5%c scenario. For the regression task, we focused on the +-50%r scenario. Table 8 lists the 20 top-ranked genes for Gefitinib. The *CTGF* gene was ranked 1st for the classification analysis and 13th for the regression analysis. The *CTGF* protein plays important roles in signaling pathways that control tissue remodeling via cellular adhesion, extracellular matrix deposition, and myofibroblast activation (Lipson, 2012); these processes are known to influence tumorigenesis and may alter drug responses (Hirohashi and Kanai, 2003). For example, EGFR is expressed in many head and neck squamous cell carcinomas and non-small cell lung carcinomas, yet many of these patients do not respond to Gefitinib treatment (Frederick et al., 2007). This lack of response has been associated with a loss of cell-cell adhesion, elongation of cells and tumor-cell invasion of the extracellular matrix (Yauch et al., 2005; Thomson et al., 2005; Witta et al., 2006). F11R was ranked second in importance for the classification analysis and seventeenth for the regression analysis. The protein encoded by this gene is a junctional adhesion molecule that regulates the integrity of tight junctions and permeability (Naik and Eckfeld, 2003). Although these associations provide some support for our feature-ranking results and that adhesion processes are important to Gefitinib responses, none of the other top-20 genes overlapped between the classification and regression analysis. The lack of agreement between the classification and regression results is not surprising. For example, even though the Random Forests algorithm uses a similar methodology for classification and regression, it is not unlikely that different genes would be selected for classification versus regression. We used data for thousands of genes, and different genes may exhibit similar methylation patterns, so the algorithms may choose different (correlated) genes by random chance. Secondly, the algorithms optimized against different objective functions for classification versus regression; even small differences in how the algorithms prioritized genes could lead to large differences in the gene ranks. However, the SVM and RF models represent multivariate patterns; thus, known cancer genes may alter drug responses in combination with the genes identified via our univariate feature-selection approach, even if they are not among the top-ranked genes.

Table 8: Most informative genes for predicting cell-line responses for Gefitinib. We used an information-gain analysis to rank genes based on their association with Gefitinib drug response. Genomic coordinates are based on build 37 of the human genome. We used information gain to rank the genes; higher scores indicate more informativeness.

Classification			Regression		
Gene	*Coordinates*	*Score*	*Gene*	*Coordinates*	*Score*
CTGF	chr6:132271356-132271658	0.272	SNAI2	chr8:49835987-49836231	0.060
F11R	chr1:160990718-160991225	0.248	CARD10	chr22:37914768-37915883	0.055
MUM1	chr19:1354420-1355350	0.228	PTGFRN	chr1:117452203-117453452	0.053
RXRB, SLC39A7	chr6:33167885-33168715	0.220	PNMAL1	chr19:46974557-46975073	0.053
DUSP7	chr3:52089652-52090845	0.204	A2M, LOC144571	chr12:9217328-9217715	0.052
TFAP2A	chr6:10419399-10420323	0.203	DGKZ	chr11:46366876-46367101	0.052
C20orf56	chr20:22559553-22560001	0.201	SDCBP2	chr20:1305899-1306554	0.052
RAB38	chr11:87908243-87908614	0.201	ACAP1, KCTD11, TMEM95	chr17:7254622-7255808	0.052
RAB34	chr17:27044168-27045049	0.196	ANKRD57, SEPT10	chr2:110370906-110373301	0.051
VIM	chr10:17270430-17272617	0.192	SLC44A2	chr19:10735999-10736396	0.050
PAK6	chr15:40531244-40531589	0.192	ALOX12	chr17:6898820-6900427	0.049
GATA2	chr3:128215212-128216905	0.190	ZNF625	chr19:12266998-12267686	0.048
SLC9A2	chr2:103235376-103236554	0.188	CTGF	chr6:132271356-132271658	0.048
C20orf56	chr20:22557517-22559240	0.187	KLF5	chr13:73632860-73634370	0.048
FERMT1	chr20:6103436-6103970	0.186	NCOR2	chr12:125003217-125003482	0.048
RBM4B	chr11:66444997-66445471	0.185	TBCD, ZNF750	chr17:80790368-80790581	0.047
ORAI2	chr7:102073605-102074334	0.183	F11R	chr1:160990718-160991225	0.046
LOC338799, SETD1B	chr12:122240899-122243390	0.181	OR10H1	chr19:15918423-15918704	0.045
ABHD5	chr3:43731998-43733108	0.181	PLEK2	chr14:67878534-67879167	0.044
MAZ	chr16:29818681-29819554	0.176	DGUOK	chr2:74153853-74154281	0.043

Tables S15-S21 indicate the top-20 ranked genes for the other 7 drugs. To gain insight regarding the roles that these genes might play in drug responses, we identified gene sets (e.g., pathways, oncogenic signatures) that significantly overlapped with these genes (Tables S22-S23). For the classification analysis, we identified significant gene sets for 5 drugs (Gefitinib, Cisplatin, Docetaxel, Doxorubicin, Etoposide). Many of these gene sets are associated with cell

differentiation, cell-cell communication and drug resistance; however, these mechanisms did not always align with the respective drugs or target proteins that we expected based on the drugs' known mechanisms. We observed similar patterns for the regression analysis. Two perhaps notable findings are that 1) a gene set associated with EGFR overexpression was associated with Gefitinib responses (this drug targets EGFR) and 2) a gene set associated with Gefitinib resistance was associated with Cisplatin responses, and it has been shown that Cisplatin's ability to induce cell death is dependent in part on EGFR signaling in some cases (Arany et al., 2004).

3.1.3.5
Using methylation profiles from cell lines to predict tumor/patient drug responses

The above analyses used methylation profiles to predict drug responses in cell lines. Via cross validation, we showed that high levels of predictive accuracy are attainable using this approach. We also found that subsampled datasets with more extreme IC_{50} values yielded the best classification results and that the SVM and Random Forests algorithms typically produced the most accurate results. Next we evaluated whether this performance would hold true in a translational-medicine context. The GDSC repository provides methylation profiles for 6,035 tumors from TCGA; these data had been preprocessed using the same methodology as the GDSC samples, thus enabling easier integration and reducing technical biases. For 1,638 TCGA patients, clinical drug-response information was available. These data indicate clinical outcomes over the course of the patients' treatment by physicians (not as part of clinical trials). In many cases, drug-response values for multiple drugs were recorded for a given patient. Each response value was categorized as "clinical progressive disease," "stable disease," "partial response," or "complete response". These respective categories represent increasing levels of response to a given drug.

We trained the SVM and Random Forests classification algorithms on the full GDSC dataset and predicted drug-response categories for each TCGA patient for which methylation and drug-response data were available. Based on our cross-validation results from the GDSC analysis, we focused on the +-5%c and +-10%c scenarios. For each TCGA test sample, our models generated a probabilistic prediction indicating whether that patient would respond to a given drug. We compared these predictions against the ordinal clinical responses for each combination of subsampling scenario (+-5%c and +-10%c), drug, and algorithm (SVM and RF); we calculated the SCC and a corresponding p-value for each comparison and adjusted for multiple tests. Generally,

the predictions exhibited low correlation with clinical responses (Table 9); However, the predictions for lower-grade glioma patients who had been treated with Temozolomide were relatively strongly correlated with clinical responses (rho = 0.372; FDR = 0.014), though this result was specific to the Random Forests algorithm and the +-5%c scenario (Figure 8). Temozolomide is an oral alkylating agent, is used commonly to treat lower-grade glioma patients, and may reduce seizures and improve prognosis (Rees, 2015).

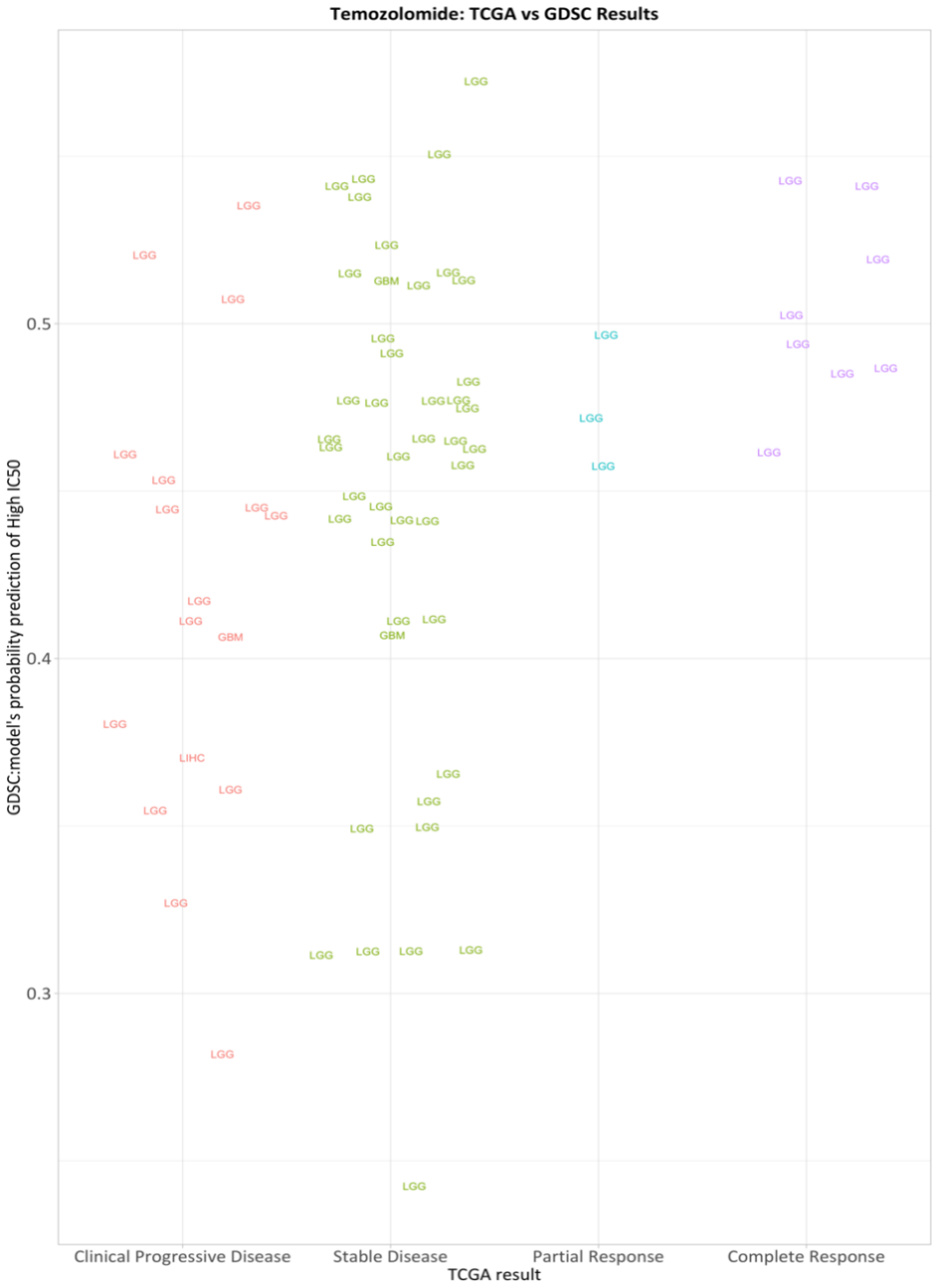


Figure 8: Predicting patient drug response from cell-line methylation profiles for Temozolomide (n=85). For each TCGA test sample, we used classification models from the GDSC data (+-5%c Random Forest) to generate probabilistic predictions of drug response.

Table 9: Correlation between predicted drug responses based on GDSC cell lines and recorded clinical responses in TCGA patients for selected combinations of subsampling scenarios and algorithms across all drugs. We treated the clinical drug responses as an ordinal variable and used the Spearman rank correlation coefficient to assess the extent to which the predicted responses correlated with the clinical responses. FDR = Benjamini-Hochberg False Discovery Rate.

Drug	Scenario	Algorithm	# Samples	Spearman	P-value	FDR
Gefitinib	+-5%c	SVM	2	1.000	1.00E+00	1.000
Gefitinib	+-5%c	Random Forest	2	1.000	1.00E+00	1.000
Gefitinib	+-10%c	SVM	2	1.000	1.00E+00	1.000
Gefitinib	+-10%c	Random Forest	2	-1.000	1.00E+00	1.000
Cisplatin	+-5%c	SVM	189	-0.127	8.11E-02	0.331
Cisplatin	+-5%c	Random Forest	189	0.041	5.72E-01	0.721
Cisplatin	+-10%c	SVM	189	-0.051	4.82E-01	0.697
Cisplatin	+-10%c	Random Forest	189	0.100	1.72E-01	0.424
Paclitaxel	+-5%c	SVM	110	0.234	1.40E-02	0.149
Paclitaxel	+-5%c	Random Forest	110	-0.163	8.84E-02	0.331
Paclitaxel	+-10%c	SVM	110	0.104	2.80E-01	0.498
Paclitaxel	+-10%c	Random Forest	110	-0.073	4.48E-01	0.697
Temozolomide	+-5%c	SVM	85	-0.217	4.65E-02	0.331
Temozolomide	+-5%c	Random Forest	85	0.372	4.53E-04	0.014
Temozolomide	+-10%c	SVM	85	-0.060	5.86E-01	0.721
Temozolomide	+-10%c	Random Forest	85	0.176	1.07E-01	0.343
Etoposide	+-5%c	SVM	31	0.125	5.01E-01	0.697
Etoposide	+-5%c	Random Forest	31	-0.260	1.58E-01	0.422
Etoposide	+-10%c	SVM	31	0.083	6.58E-01	0.753
Etoposide	+-10%c	Random Forest	31	-0.223	2.29E-01	0.440

Gemcitabine	+-5%c	SVM	56	-0.235	8.11E-02	0.331
Gemcitabine	+-5%c	Random Forest	56	0.227	9.30E-02	0.331
Gemcitabine	+-10%c	SVM	56	-0.170	2.10E-01	0.440
Gemcitabine	+-10%c	Random Forest	56	0.207	1.25E-01	0.364
Docetaxel	+-5%c	SVM	61	0.132	3.09E-01	0.521
Docetaxel	+-5%c	Random Forest	61	-0.158	2.25E-01	0.440
Docetaxel	+-10%c	SVM	61	0.096	4.60E-01	0.697
Docetaxel	+-10%c	Random Forest	61	-0.155	2.34E-01	0.440
Doxorubicin	+-5%c	SVM	61	-0.237	6.56E-02	0.331
Doxorubicin	+-5%c	Random Forest	61	0.338	7.78E-03	0.125
Doxorubicin	+-10%c	SVM	61	-0.063	6.31E-01	0.748
Doxorubicin	+-10%c	Random Forest	61	0.075	5.67E-01	0.721

3.1.4 Discussion

In an ideal setting, patient data would be used to train predictive models for clinical drug responses directly, as these data may accurately reflect tumor behavior in patients. Environmental factors, the tumor microenvironment, co-existing conditions, and a variety of other factors can affect a tumor's behavior in ways that may not be accounted for in preclinical studies. However, acquiring drug-response data directly from human patients may require conducting many experimental tests on a given patient, which could be unethical, harmful, and subject to many confounding factors. In addition, patients are typically assigned standard-of-care protocols based on their specific cancer type. As a result, experimental drug-response data for large patient cohorts are scarcely available. An alternative approach is to use preclinical samples to identify molecular signatures of drug response and later use those signatures to predict clinical drug responses in patients.

Cell lines serve as preclinical models for drug development. Being able to accurately predict drug responses for a given cell line based on molecular features may help in optimizing drug-development pipelines and explain mechanisms behind treatment responses. We focused on DNA methylation profiles as one type of molecular feature that is known to drive tumorigenesis

and modulate treatment responses (Esteller, 2002). When using classification or regression algorithms to predict discrete or continuous responses, respectively, we consistently observed excellent predictive performance when the training and test sets both consisted of cell-line data. Although conventional wisdom advises against discretizing a continuous response variable, where possible, due to loss of information, we wished to evaluate the potential to make effective predictions in this scenario, in part because clinical treatment responses are sometimes represented as discrete values.

Of note, this study focuses primarily on evaluating the effect of subsampling on model performance rather than on introducing new algorithms. Using subsampling, we observed that classification performance generally improved as more extreme examples were used for training and testing, whereas the opposite was often true for the regression analyses. This suggests that during regression, the algorithms benefitted from seeing examples across a diverse range of IC_{50} values for a given drug, whereas the classification algorithms were confounded by seeing examples with relatively similar drug responses, even though sample sizes were smaller. However, again we note that the regression results often differed depending on the evaluation metric used. These results have potential financial implications: if researchers can identify cell lines that are extreme responders for a particular drug, they may only need to generate costly molecular profiles for those cell lines. Future research may elucidate whether this finding generalizes to other types of molecular data and other drugs.

Previous efforts to associate DNA methylation levels with drug responses include work from Shen et al. (2007) who quantified methylation for 32 CpG islands in the NCI-60 cell lines, creating a sensitivity database for ~30k drugs and identifying biomarkers that predict drug sensitivity. Instead, our work uses microarray data to quantify methylation levels for thousands of genes across 987 cell lines but for fewer drugs. Rather than searching for individual genes that predict drug sensitivity, we constructed predictive models that represent patterns spanning as many as thousands of genes. Such an approach may better represent complex interactions among genes and thus yield improved predictive power, but a tradeoff is reduced model interpretability. We sought to shed some insight into the biological mechanisms that influence drug responses via feature selection, but methods for deriving such insights from genome-wide data are still in their infancy. Recent work using mathematical optimization models shows promise as a way to integrate

molecular data from cell lines with drug-sensitivity information to infer resistance mechanisms (Fleck et al., 2016; Fleck et al., 2019).

A variety of computational methods have been proposed to predict drug responses for cell lines based on molecular data. Classical algorithms like decision trees and support vector machines have been used to predict the clinical efficiency of anti-cancer drugs and to classify drug responses (Stetson et al., 2015; Borisov et al., 2018; Oskooei et al., 2018; Webber et al., 2018; Parca et al., 2019; Su et al., 2019). Neural networks (Menden et al., 2013) and deep neural networks (Chiu et al., 2019) have been used to predict drug response based on genomic profiles from cell lines. Other techniques have included elastic net regression (Basu et al., 2013; Webber et al., 2018; Parca et al., 2019), linear ridge regression (Geeleher et al., 2017), and LASSO regression (Huang et al., 2020). Alternative approaches based on computational linear algebra or network structures have also been applied to infer drug response in cell lines; these include matrix factorization (Guan et al., 2019), matrix completion (Nguyen and Le, 2018), and link prediction (Stanfield et al., 2017) methods. Finally, a community-based competition assessed the ability to predict therapeutic responses in cell lines using 44 regression-based algorithms (Costello et al., 2014). In our study we used diverse algorithms, but our primary focus was data subsampling and evaluating the potential to make accurate predictions of drug response in cell lines using relatively extreme responders, rather than to introduce new algorithms.

We attempted to predict clinical responses for patients from TCGA, but the accuracy of these predictions was typically poor. Integrating datasets can introduce batch effects (Leek et al., 2010) and other systematic biases; we attempted to mitigate these biases using data that had been preprocessed identically for GDSC and TCGA and using an empirical Bayesian method. However, subtle differences in the way biological samples are handled and processed in the lab can make generalization difficult to achieve. Furthermore, inherent differences between cell lines and tumors may confound such predictions. Cell lines are grown in a controlled environment, and the cells are relatively homogeneous, whereas tumor samples are a heterogeneous milieu of cells. In addition, TCGA tumor responses were based on clinical observations, so there was no direct mapping between these measurements and IC_{50} values for the cell lines. Furthermore, our approach to quantifying predictive performance was different for the GDSC cross-validation analysis compared to the TCGA training/testing analysis. In the former, the class variable represented two possible outcomes (response and non-response). In the latter, the class variable was ordinal. Yet

another challenge was that we used cell lines from all available cell types in GDSC. Better accuracy might be attained when training and testing on a single cell type; however, larger sample sizes would be necessary.

Our study has additional limitations that could be addressed in future research. For one, we focused on DNA methylation profiles in isolation, but other types of molecular features likely modulate treatment responses. A number of cell-line studies have used gene-expression profiles to predict drug responses, and future studies could evaluate the potential benefits of incorporating more than one type of molecular feature into response-prediction models. The treatment-response data were often imbalanced, meaning that not all response classes included similar numbers of patients. Hence, additional work could analyze the effect of class imbalance on model performance. Finally, we adjusted the methylation data for dataset and cell type using an empirical Bayesian framework. However, as few as 2-3 samples were available for some of the cell types, so the correction method may have had difficulty adjusting based on such small numbers of examples.

3.1.5 Conclusion

We applied machine-learning algorithms to predict cytotoxic responses for eight anti-cancer drugs using genome-wide, DNA methylation profiles from 987 cell lines from the Genomics of Drug Sensitivity in Cancer (GDSC) database. We then compared the performance of the classification and regression algorithms and evaluated the effect of sample size on model performance by artificially subsampling the data to varying degrees. The classification algorithms performed best when relatively few cell lines were used to train and test the models, attaining AUC values as high as 0.97. In contrast, the regression algorithms typically performed best when all cell lines were used to train and test the models, though this result depended on the evaluation metric used. For additional validation, we evaluated our ability to train a model based on drug responses in the GDSC cell lines and then accurately predict patient drug responses using data from The Cancer Genome Atlas (TCGA). Because patient-response values are categorical in nature, we only performed classification for these data. In most cases, classification algorithms trained on the full GDSC dataset to predict drug-response categories for TCGA patients were unable to identify patterns in the cell-line methylation data that translated to patient responses.

3.2
Case Study 2 - Predicting drug sensitivity of cancer cells based on RNA sequencing data

3.2.1
Introduction

This study uses RNA-sequencing data from AML tumor samples to model drug efficacy from four drugs. We evaluate the performance of five classification algorithms in a semi-supervised learning setting. Tree-based, probability-based, kernel-based, ensemble-based and distance-based methodologies are studied in this research. Since obtaining molecular profiles from tumor samples and generating drug responses in clinical trials can be complex and expensive (Iorio et al., 2016), an underlying motivation for this approach is to find new strategies to expand the size of the training dataset when developing models based on tumor data.

In a semi-supervised learning environment, these classical algorithms were applied to understand their behavior and patterns in self-training procedure. We use Beat AML labeled data and TCGA AML unlabeled data as input values and discretized IC_{50} values as output labels. An initial classifier is generated using labeled samples to predict pseudo-labels for unlabeled data. Then, a new training set is generated using original labeled data and unlabeled data with predicted pseudo-labels. Subsequently, a new classifier is generated using this new training set.

Specific algorithms yielded best performance when applying a semi-supervised learning strategy; support vector machines and naïve Bayes algorithms presented best results in most scenarios. Feature selection analysis showed that models performed best when selecting features over the 0.45 threshold score. Finally, we also observed that probabilistic prediction acceptance threshold did not impact models.

3.2.2 Methods

The BeatAML database contains data for human tumors derived from Acute Myeloid Leukemia (AML) cancer. It provides multiple types of clinical and molecular data for these samples as well as response values for 122 anti-cancer drugs. The TCGA database contains tumor data derived from over 30 types of cancer, including AML. In this research, we use Beat AML transcriptome profiling data and drug response values, also known as our labeled data. We also use TCGA AML transcriptome profiling data, known as our unlabeled data. Drug responses were measured as the IC_{50} value; the more sensitive the cell line, the lower the IC_{50} value for any given drug. We developed semi-supervised machine learning models of drug response using RNA-seq data from both Beat AML and TCGA, both following the HTSeq-Counts pipeline. We selected 4 drugs commonly used as AML anti-cancer treatment: Gilteritinib, Midostaurin, Quizartinib and Venetoclax. The Genomic Data Commons (GDC) Data Portal provides Beat AML RNA-seq values for 288 samples and for 302 TCGA AML samples. Both data had been preprocessed using the same pipeline. Drug response data for Beat AML patients were obtained from the Vizome portal (Tyner et al., 2018).

For both Beat AML and TCGA databases, RNA-Seq samples were filtered to include only first diagnosis patients over 18 years old. Only one sample per-patient was included. For Beat AML, samples with missing IC_{50} values were excluded on a per-drug basis; thus, sample sizes differed across the drugs. We applied Z-score normalization on a per-gene basis across all samples in Beat AML and TCGA AML. Next, we used ComBat (Leek et al., 2020) to adjust for systematic differences between the two datasets (Beat AML and TCGA AML). We opted for a classification analysis as their predictions are intuitive to interpret as they assign probabilities to each class as a classification threshold. To enable classification for the Beat AML patient samples, we discretized the IC_{50} values into "low" and "high" values. The used threshold was the median IC_{50} value across all samples.

We performed a feature selection analysis to reduce data dimensionality by identifying most informative genes. This evaluation is performed for each drug, on their Beat AML training set. Training set represents 80% of total samples of each drug and testing set represents the other 20%. We applied three feature selection methods independently: the Kruskal-Wallis test, Pearson correlation and ReliefF. We chose these three algorithms as they showed the top performance

when evaluated against other methods in a genomic environment; this result was obtained as preliminary data from research done by Piccolo Lab (Brigham Young University, USA). The Kruskal-Wallis algorithm was implemented using R software and the *kruskal.test* function, which is already pre-installed in R. Pearson correlation and ReliefF algorithms were implemented using the Weka software (Eibe et al., 2016) and the functions *CorrelationAttributeEval* and *ReliefFAttributeEval* respectively. All three strategies are filter tasks, each generating an importance score to each feature (gene) as a final output. For each output list, we normalize score values (min-max normalization), ensuring that each gene will always have an importance score between 0 and 1. We then average the three scores obtained by each gene, generating an average score for each feature. Subsequently, we generated 11 feature selection scenarios to be evaluated; this was performed by ordering genes based on their average scores and then, creating several cuts on the data based on different threshold values. For the first scenario, if a gene had an average score greater than 0.3, the gene was included in the scenario "Cut30". In the next scenario, if the gene had an average score value greater than 0.35, it would be included in the "Cut35" scenario, and so on. The last scenario ("Cut80") included genes that had an average rank score above 0.8. We then generated training and testing sets for each feature selection scenario, where each scenario set would contain accepted features only. After finalizing feature selection evaluation and scenario creation, we then started model development.

A semi-supervised self-training model encompasses supervised and unsupervised learning (Van Engelen et al., 2020). We use a labelled dataset (Beat AML); in other words, we possess both the input information (RNA-seq) and output labels (IC_{50} values) for these samples. We also utilize an unlabeled dataset (TCGA AML), for which we only have access to the RNA-seq data. First, we generated a supervised classifier using the labeled dataset. Using this classifier, we predicted pseudo-labels to TCGA samples. Each predicted label comes with a probabilistic prediction, indicating whether that patient would respond to a given drug. In other words, this probability indicates how certain the model is that a specific patient would respond well (or not) to a particular treatment. After each TCGA sample had received a pseudo-label (and a probabilistic prediction), we defined if we would want to include this new labeled data (TCGA AML) to our original labeled dataset (Beat AML). To make this decision, we looked at the probabilistic prediction of each sample. If one sample had a probabilistic prediction above a certain acceptance threshold, we

would accept the pseudo-label and concatenate this new sample with the BeatAML data, thus generating a new training set.

We generated 10 probabilistic prediction acceptance thresholds to be evaluated, varying acceptance values from 0.5 to 0.9. For the first scenario, if an output label had a probabilistic prediction above 0.5, the sample would be included in the training set. In the next scenario, if an output label had a probabilistic prediction above 0.55, the sample would be included in the training set, and so on. The last scenario accepted samples in which the probabilistic predictions were above 0.9.

By performing this evaluation, we generated a new supervised training set built from the Beat AML dataset and approved TCGA AML samples. We then trained a new classification model using this new training set. This final model was then evaluated using our Beat AML test set. We used the Random Forests (tree-based) (Breiman, 2001), Support Vector Machines (kernel-based) (Vapnik, 1998), Gradient Boosting Machines (ensemble-based) (Breiman, 1997), k-Nearest Neighbors (distance-based) (Cover and Hart, 1967) and Naïve Bayes (probability-based) (Maron, 1961) algorithms for classification. We performed the analyses using the R programming language (R Core Team, 2019) and Rstudio (https://rstudio.com). The machine-learning algorithms were implemented in the following R packages: mlr (Bischl et al., 2016), e1071 (Meyer et al., 2019), xgboost (Chen et al., 2015), randomForest (Liaw and Wiener, 2002) and kknn (Schliep and Hechenbichler, 2016). During the training of all classifiers, we sought to select the best hyperparameters for each algorithm via 5-fold cross validation of the training set.

For each combination of drug, algorithm, feature selection scenario and probabilistic prediction acceptance threshold, we evaluated the performance of all hyperparameter combinations (Table 10) and assessed performance for predicting drug responses using several evaluation metrics. We used accuracy, area under the receiver operating characteristic curve (AUC) (Fan et al., 2006), F1 measure (Forman, 2003), Matthews correlation coefficient (MCC) (Baldi et al., 2000), recall and specificity. We also generated supervised classifiers using only Beat AML data and compared performance across learning strategies.

Table 10: Descriptions of the algorithms we tested and hyperparameters that we evaluated via nested cross validation. Hyperparameter optimization was performed for all tested algorithms. All parameter combinations for each algorithm were evaluated via nested cross validation; optimal combinations were then used for outer-fold predictions.

Algorithm	Hyperparameters	Definition	Tested Values
classif.svm	1. Kernel	The kernel function used to transform data to higher-dimensional spaces and then become linearly separable.	Linear; Radial; Polynomial; Sigmoid
	2. Cost	The regularization parameter in the cost function, to penalize missing classifications.	0.1; 1; 10; 100
	3. Scale	Whether the variables should be scaled.	True; False
classif.randomForest	1. Ntree	The number of trees to grow.	100; 500; 1000
	2. Nodesize	Minimum size of terminal nodes.	1; 3; 5; 7
	3. Importance	Whether the importance of predictors should be assessed.	True; False
classif.kknn	1. K	The number of neighbors considered.	3; 7; 10
	2. Scale	Whether to scale variables to have equal standard deviation.	True; False
classif.naiveBayes	1. Laplace	The amount of Laplace (additive) smoothing.	0; 1; 5; 10
classif.xgboost	1. Nround	The maximum number of boosting iterations.	100; 250; 500
	2. Max_depth	The maximum depth of a tree.	1; 5; 10
	3. Eta	How much the contribution of each tree is scaled to the overall approximation, to control the learning rate.	0.1; 0.3; 0.5

3.2.3 Results

Using data from Beat AML and TCGA AML, we applied a semi-supervised learning approach to evaluate the potential to predict drug response based on genome-wide RNA sequencing data. We assess several hyperparameters within the model to understand their impact on model behavior.

We collected RNA-seq data and IC_{50} response values for four drugs from the Beat AML database and RNA-seq data from TCGA AML. We aimed to predict categories (classes) of drug sensitivity, where each category represented whether each cell line exhibited a "low" or "high" response to each drug. Each class corresponded to relatively low or high IC_{50} values. We categorized each cell line on a per-drug basis, according to whether its IC_{50} value was greater than the median across all cell lines. This categorization promotes a simplified yet intuitive interpretation of the treatment outcomes while enabling the use of diverse classification methods.

To reduce dimensionality, we performed a feature selection analysis. We applied three FS methods independently and used their average score list to generate several feature selection scenarios to be assessed. Subsequently, during the development of the semi-supervised model, we also wanted to evaluate the impact of different probabilistic prediction acceptance thresholds. For each combination of drug, algorithm, feature selection scenario and probabilistic prediction acceptance threshold, we optimized algorithms' hyperparameters via 5-fold cross validation and assessed the semi-supervised model performance for predicting drug responses. This generates a total of 550 scenarios for each drug (11 FS scenarios x 5 algorithms x 10 Probability Thresholds).

To evaluate the performance of the semi-supervised learning strategy, we compared obtained AUC values to the ones resulting from a supervised model. Initially, we evaluated the impact of the probabilistic prediction acceptance threshold. Overall, the classification algorithms were not greatly impacted by this parameter. We expected that as the threshold became higher, performance would improve as a result of more certainty in pseudo-labels precision. However, by analyzing Figure 9, it is clear that is it not possible to pinpoint a single threshold value which would improve the results for the different drug and algorithm combinations, with each of them responding differently to the distinct thresholds. Venetoclax shows improvement in relation to the

supervised model with any combination of algorithm and threshold, while Quizartinib presents deteriorated results. Naïve Bayes algorithm exhibits no correlation between the delta AUC and the threshold.

Figure 9 – Probabilistic prediction acceptance threshold analysis. Bar graphs illustrate the performance of each threshold in each combination of algorithm and drug scenario. Y-axis represents the Delta AUC value (Semi-supervised model AUC value minus Supervised model AUC value). Higher AUC values (higher bars) represent relatively better performance. Expected results were that as threshold values increased, semi-supervised performance would improve. However, no single threshold value portrays significant improvement in Delta AUC. Also, Naïve Bayes algorithm was not impacted by different thresholds.

Since it was not feasible to choose a single threshold to be applied constantly to the different scenarios, we decided to use the average Delta AUC value found for the different thresholds. By doing so, we reduce the impact of any outlier, and it is possible to compare the different scenarios without having to take the effect of the different thresholds into account.

Having ruled out the impact of the threshold on the results, it was possible to analyze the impact of each algorithm in the analysis of the different drugs. Figure 10 illustrates what we had already observed in Figure 9. The SVM algorithm shows the best results all around, also being very consistent in terms of the absolute value found for the different drugs, apart from Quizartinib. The spider graph also shows that the drug Venetoclax presents good results for all methods, having very consistent and positive results. Also, Gilteritinib has inconsistent results, varying greatly based on the applied algorithm, as well as having the lowest overall value from all the Method-Drug combinations. Quartizinib showed very consistent, but also very poor results, having mainly negative ones, meaning it was consistently outperformed by the supervised learning strategy.

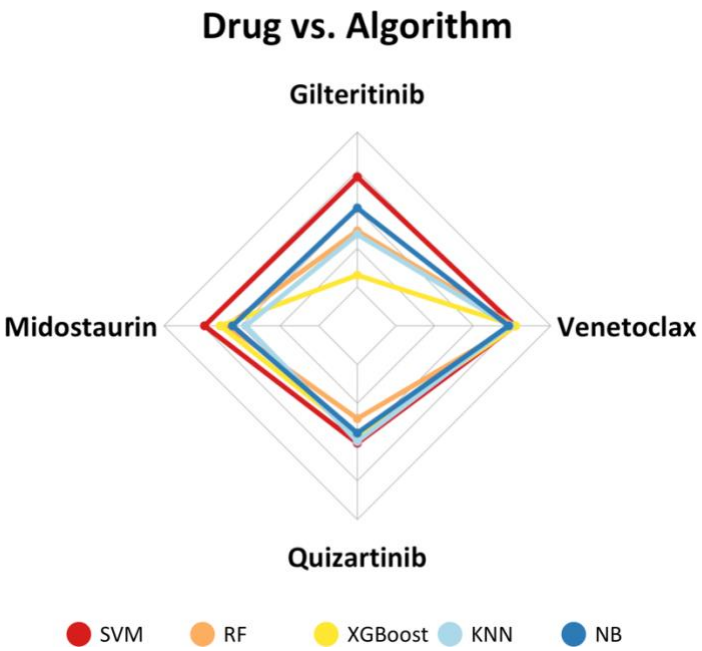


Figure 10 – Classification algorithm results across the four analyzed drugs. The "spider" graph illustrates drug performance across different algorithms in a semi-supervised environment. Results were averaged between all probabilistic prediction acceptance thresholds and feature selection scenarios. Spider graph presents the Delta AUC (Semi-supervised model AUC value minus Supervised model AUC value), ranging from -0.3 to 0.3. Results that are further away from the center represent higher AUC values (relatively better performance) than results closer to it. SVM algorithm presents best performance for Gilteritinib, Midostaurin and Quizartinib. For Venetoclax, XGBoost and SVM present best and second-best performance respectively. Venetoclax also presents constant higher results across the four evaluated drugs.

Table 11 complements the conclusions that were drawn from Figure 10 by presenting with exact values what we could perceive in the previous graph. Midostaurin and Venetoclax portrayed most consistent positive results, beating the supervised learning the most times.

Table 11: Minimum, mean and maximum AUC value for each combination of drug and algorithm, averaged across all probabilistic prediction thresholds and feature selection scenarios. Bold font indicates the best-performing combinations.

Drug	Method	Min	Mean	Max
Gilteritinib	SVM	-0.222	0.126	0.290
Gilteritinib	RF	-0.332	-0.082	0.036
Gilteritinib	NB	-0.222	0.006	0.240
Gilteritinib	KNN	-0.333	-0.096	0.093
Gilteritinib	XGBoost	-0.387	-0.254	-0.006
Midostaurin	SVM	**0.024**	**0.142**	**0.362**
Midostaurin	RF	-0.165	0.054	0.217
Midostaurin	NB	-0.124	0.034	0.143
Midostaurin	KNN	-0.231	-0.017	0.129
Midostaurin	XGBoost	-0.109	0.081	0.180
Quizartinib	SVM	-0.101	0.004	0.098
Quizartinib	RF	-0.138	-0.091	0.051
Quizartinib	NB	-0.180	-0.035	0.056
Quizartinib	KNN	-0.087	-0.003	0.054
Quizartinib	XGBoost	-0.127	-0.028	0.078
Venetoclax	SVM	0.088	0.162	0.253
Venetoclax	RF	0.126	0.159	0.218
Venetoclax	NB	0.052	0.137	0.220
Venetoclax	KNN	0.077	0.141	0.206
Venetoclax	XGBoost	**0.108**	**0.164**	**0.294**

Table 12 also confirms that the SVM algorithm showed best performance in a semi-supervised learning environment. Second-best classification method was Naïve Bayes, portraying higher average AUC values when compared to the other three algorithms. The table also shows that the Average AUC values were mostly positive, meaning that, overall, the semi-supervised classifier presented better results than the supervised learning method.

Table 12: Summary of AUC values across all combinations of FS scenario and algorithm, averaged across all probabilistic prediction thresholds and drugs.

Scenario	Method	Average AUC Value	Min AUC Value	Max AUC Value	Standard Deviation AUC Value
50	SVM	0.178	0.037	0.362	0.144
70	SVM	0.150	0.034	0.252	0.096
55	SVM	0.148	-0.003	0.291	0.123
45	SVM	0.142	0.065	0.243	0.083
75	NB	0.124	0.012	0.220	0.093
60	NB	0.116	0.049	0.194	0.062
75	SVM	0.115	-0.017	0.251	0.121
30	SVM	0.105	0.063	0.156	0.042
55	NB	0.104	-0.035	0.240	0.113
70	NB	0.104	0.006	0.216	0.094
60	SVM	0.100	-0.057	0.195	0.122
65	SVM	0.087	-0.101	0.253	0.154
35	SVM	0.082	-0.039	0.187	0.105
70	XGBoost	0.079	-0.030	0.187	0.112
75	XGBoost	0.078	-0.015	0.211	0.098
40	SVM	0.069	-0.041	0.156	0.084
50	KNN	0.065	-0.060	0.147	0.094
75	KNN	0.063	0.013	0.142	0.058
70	RF	0.060	-0.099	0.217	0.159
65	NB	0.043	-0.073	0.174	0.106
75	RF	0.039	-0.097	0.185	0.134
55	RF	0.035	-0.102	0.148	0.108
30	KNN	0.029	-0.087	0.206	0.129
70	KNN	0.029	-0.111	0.106	0.098
50	NB	0.026	-0.059	0.111	0.070
80	SVM	0.022	-0.222	0.186	0.177
80	NB	0.020	-0.222	0.208	0.179
45	RF	0.020	-0.104	0.153	0.112
55	KNN	0.018	-0.059	0.119	0.076
35	KNN	0.018	-0.132	0.204	0.147
40	RF	0.015	-0.138	0.126	0.114
30	RF	0.014	-0.132	0.186	0.137
40	NB	0.007	-0.110	0.103	0.088

65	RF	0.005	-0.188	0.167	0.179
35	RF	0.004	-0.127	0.142	0.115
60	KNN	0.004	-0.104	0.120	0.111
45	NB	0.003	-0.089	0.073	0.068
55	XGBoost	0.003	-0.262	0.187	0.198
50	RF	-0.001	-0.121	0.127	0.126
60	XGBoost	-0.001	-0.249	0.140	0.173
50	XGBoost	-0.004	-0.265	0.139	0.185
65	KNN	-0.019	-0.231	0.165	0.176
45	KNN	-0.020	-0.151	0.099	0.119
80	XGBoost	-0.021	-0.347	0.294	0.274
60	RF	-0.022	-0.215	0.145	0.157
65	XGBoost	-0.035	-0.286	0.166	0.194
35	XGBoost	-0.044	-0.387	0.211	0.263
45	XGBoost	-0.045	-0.310	0.125	0.198
30	XGBoost	-0.050	-0.342	0.152	0.226
80	RF	-0.057	-0.332	0.218	0.241
40	KNN	-0.058	-0.333	0.150	0.203
80	KNN	-0.059	-0.285	0.180	0.202
40	XGBoost	-0.063	-0.330	0.172	0.219
35	NB	-0.073	-0.158	0.052	0.092
30	NB	-0.082	-0.180	0.090	0.119

Table 12, in conjunction with the heatmaps below (Figure 11), aids in the identification of the best feature selection scenarios for the semi-supervised classifier. It becomes explicit that there is no evident winning FS scenario when evaluating all possible drug-algorithm combinations. Top performer combination varies depending on which drug and algorithm is being evaluated. However, it was possible to identify that the low scenarios, from 30 to 45, had the worst results in general.

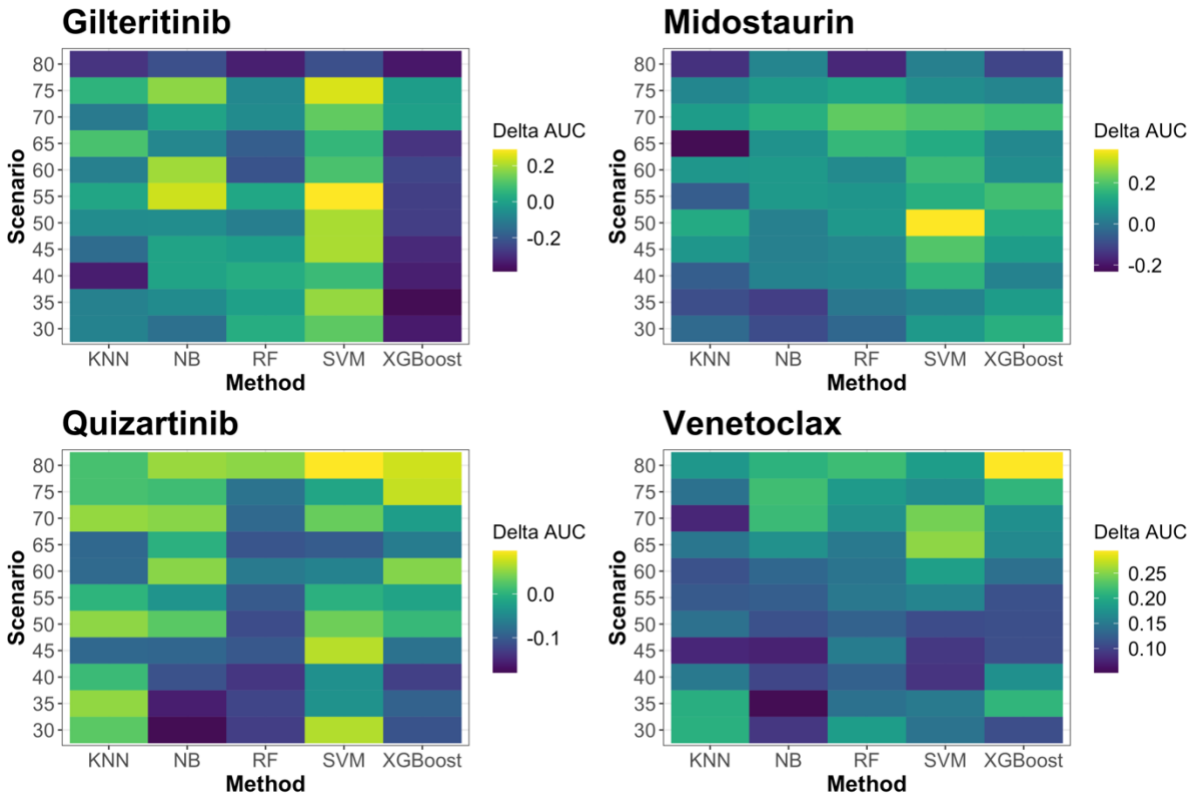


Figure 11 – Delta AUC variation across all drug, scenario and classification method combination. Heatmaps illustrate semi-supervised versus supervised AUC performance across all combinations of scenario and classification methods. For each combination, results were averaged across all probabilistic prediction thresholds. Top performing scenario varies according to drug and algorithm being evaluated.

3.2.4 Discussion

When generating predictive models for drug response, tumor data should be used directly as they reflect the full patient characteristics. Environmental factors, the tumor microenvironment, co-existing conditions and a variety of other factors can affect a tumor's behavior in ways that may not be accounted for in preclinical studies. However, acquiring patient molecular and drug response data can be an expensive and complex process.

As a result, experimental drug response data for tumor databases are limited. Patient cohorts tend to have small sample sizes, hampering the training phase of model development. To overcome this difficulty, data analytics processes and machine learning strategies are pursued when generating predictive models.

Most popular predictive models' strategies follow a supervised learning approach, requiring large amounts of molecular features and cytotoxic responses. Since large tumor cohorts are scarcely available, other learning strategies may adapt better to this molecular data. We focused on a semi-supervised learning approach to train a classification model based on RNA-sequencing data. When comparing the semi-supervised model performance against a more traditional supervised approach, we observed prediction improvements in several tested scenarios.

The current study focuses primarily on evaluating the impact of a semi-supervised self-training approach on molecular data. In this procedure, a supervised classification model is trained using labelled data. This classifier is then used to predict pseudo-labels to unlabeled samples. Then, most confident predictions are added to our labeled data, generating a new training set. The classifier is re-trained using this new training set, composed of original labeled data and pseudo-labelled data. We evaluated the influence of several model parameters on prediction performance, including the number of used features, the probabilistic prediction acceptance threshold and the used base learner. Rather than introducing new prediction algorithms, we focused on understanding the impact of this learning strategy on molecular features. We observed that specific algorithms (SVM and Naïve Bayes) and that feature selection scenarios containing features with scores over 0.45 performed best in this semi-supervised strategy. The probabilistic prediction acceptance threshold did not seem to impact prediction performance. The results from this work require future validation before they can be extended to other types of molecular data and other drugs.

Research efforts dedicated to the application of semi-supervised learning in healthcare include work in breast cancer diagnosis (Zemmal et al., 2016; Peng et al 2016), lung cancer diagnosis (Khosravan et al., 2018), skin cancer diagnosis (Masood et al., 2015), lymph node metastases diagnosis (Jaiswal et al., 2019), cancer recurrence (Park et al., 2014; Shi et al., 2011), cancer sub-type detection (Bair et al., 2004; Koestler et al., 2010; Steinfeld et al., 2008), cancer patient clustering (Ma et al., 2018), protein classification (Weston et al., 2005), cancer survival analysis (Chai et al., 2017) and phenotype prediction (Smith et al., 2020). Rampášek et al. (2019) generated a Drug Response Variational Autoencoder (Dr. Vae) that applies latent representation of underlying gene states before and after drug use. In this study, we focused on expanding the current semi-supervised knowledge when predicting drug response.

Some limitations of our study need to be further investigated in future research. Our efforts have been directed to RNA-sequencing data, but other types of molecular features may be important in conjunction to RNA-seq to assess treatment responses. Second, features were selected based on machine learning algorithms but were not evaluated from a biological perspective. Hence, additional work could analyze the effect of choosing features based on biological knowledge versus from a machine learning perspective. Another analysis to be explored would be to test different pseudo-labeled techniques; we applied the self-training semi-supervised method, which is considered the most basic pseudo-labeling approach existent. Diverse semi-supervised methodologies have been created and applied to other research areas and we analyze one possible approach; other strategies could also be tested in the molecular biology environment.

3.2.5 Conclusion

Using a semi-supervised learning strategy, machine learning algorithms were applied to predict drug sensitivity to 4 anti-cancer drugs. Acute myeloid leukemia tumor samples derived from two databases were used to generate these models. We obtained Beat AML genome wide RNA-seq data and drug response values and attained AML RNA-seq data from The Cancer Genome Atlas (TCGA). Several parameters were assessed during model development to further understand their impact on model's behavior. For each combination of drug, algorithm, feature selection scenario and probabilistic prediction acceptance threshold, we evaluated the efficiency of the generated classifiers and compared their performance against a supervised classification model. Support vector machines (SVM) presented best performance in a semi-supervised setting, attaining AUC improvements over the supervised model in most scenarios. When evaluating feature selection scenarios, a greater reduction in original features yielded best results. Probabilistic prediction acceptance threshold did not impact the semi-supervised model.

4
References

Ali, M., Khan, S. A., Wennerberg, K., & Aittokallio, T. (2018). Global proteomics profiling improves drug sensitivity prediction: results from a multi-omics, pan-cancer modeling approach. *Bioinformatics*, *34*(8), 1353-1362.

Ammad-Ud-Din, M., Khan, S. A., Malani, D., Murumägi, A., Kallioniemi, O., Aittokallio, T., & Kaski, S. (2016). Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics*, *32*(17), i455-i463.

Ammad-Ud-Din, M., Khan, S. A., Wennerberg, K., & Aittokallio, T. (2017). Systematic identification of feature combinations for predicting drug response with Bayesian multi-view multi-task linear regression. *Bioinformatics*, *33*(14), i359-i368.

Arany, I., Megyesi, J. K., Kaneto, H., Price, P. M., & Safirstein, R. L. (2004). Cisplatin-induced cell death is EGFR/src/ERK signaling dependent in mouse proximal tubule cells. *American Journal of Physiology-Renal Physiology*, *287*(3), F543-F549.

Arechederra, M., Daian, F., Yim, A., Bazai, S. K., Richelme, S., Dono, R., ... & Maina, F. (2018). Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. *Nature Communications*, 9(1), 3164.

Azuaje F. (2017). Computational models for predicting drug responses in cancer research. *Briefings in Bioinformatics*, *18*(5), 820–829. doi:10.1093/bib/bbw065.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, *16*(5), 412-424.

Bair, E., & Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, *2*(4), e108.

References

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483:603-607, 2012.

Basu, A., Bodycombe, N. E., Cheah, J. H., Price, E. V., Liu, K., Schaefer, G. I., ... & Bracha, A. L. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154(5), 1151-1161.

Bellman, R. E. (2015). *Adaptive control processes: a guided tour* (Vol. 2045). Princeton university press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., ... & Jones, Z. M. (2016). mlr: Machine Learning in R. *The Journal of Machine Learning Research*, *17*(1), 5938-5942.

Borisov, N., Tkachev, V., Suntsova, M., Kovalchuk, O., Zhavoronkov, A., Muchnik, I., & Buzdin, A. (2018). A method of gene expression data transfer from cell lines to cancer patients for machine-learning prediction of drug efficiency. *Cell Cycle*, 17(4), 486-491.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cameron, A. C., & Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics*, *77*(2), 329-342.

References

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008, July). An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning* (pp. 96-103).

Chai, H., Li, Z. N., Meng, D. Y., Xia, L. Y., & Liang, Y. (2017). A new semi-supervised learning model combined with cox and sp-aft models in cancer survival analysis. *Scientific reports*, *7*(1), 1-12.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 1-27.

Chang, Yoosup, Hyejin Park, Hyun-Jin Yang, Seungju Lee, Kwee-Yum Lee, Tae Soon Kim, Jongsun Jung, and Jae-Min Shin. (2018).

Chang, Y., Park, H., Yang, H. J., Lee, S., Lee, K. Y., Kim, T. S., ... & Shin, J. M. (2018). Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Scientific reports*, *8*(1), 1-11.

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. Machine learning, 46(1-3), 131-159.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, *6*(2), e17238.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1-4.

Chen, T. and Sun, W. (2017). Prediction of cancer drug sensitivity using high-dimensional omic features. *Biostatistics*, 18:1.

Chiu YC, Chen HH, Zhang T, Zhang S, Gorthi A, Wang LJ, Huang Y, Chen Y. (2019). Predicting drug response of tumors from integrated profiles by deep neural networks. *BMC Medical Genomics* 12(Suppl 1): 18, DOI:10.1186/s12920-018-0460-9.

Choi, M., Shi, J., Zhu, Y., Yang, R., & Cho, K. H. (2017). Network dynamics-based cancer panel stratification for systemic prediction of anticancer drug response. *Nature communications*, *8*(1), 1-12.

Corte's-Ciriano, I. et al. (2016). Improved large-scale prediction of growth in- hibition patterns using the NCI60 cancer cell line panel. *Bioinformatics* 32: 85–95.

Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology* 32(12):1202-1212.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Dhruba, S. R., Rahman, R., Matlock, K., Ghosh, S., & Pal, R. (2018). Application of transfer learning for cancer drug sensitivity prediction. *BMC bioinformatics*, *19*(17), 497.

Ding, Z., Zu, S., & Gu, J. (2016). Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics*, 32(19), 2891-2895.

Ding, M. Q., Chen, L., Cooper, G. F., Young, J. D., & Lu, X. (2018). Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Molecular Cancer Research*, *16*(2), 269-278.

Dong, Z., Zhang, N., Li, C., Wang, H., Fang, Y., Wang, J., & Zheng, X. (2015). Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC cancer*, *15*(1), 1-12.

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Emdadi, A., and Changiz E. "DSPLMF: A Method for Cancer Drug Sensitivity Prediction Using a Novel Regularization Approach in Logistic Matrix Factorization." *Frontiers in Genetics* 11 (2020): 75.

Esteller, M., Corn, P. G., Baylin, S. B., & Herman, J. G. (2001). A gene hypermethylation profile of human cancer. *Cancer research, 61*(8), 3225-3229.

Esteller, M. (2002). CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* 21, 5427–5440 doi:10.1038/sj.onc.1205600.

Faivre, S., Djelloul, S., & Raymond, E. (2006, August). New paradigms in anticancer therapy: targeting multiple signaling pathways with kinase inhibitors. In *Seminars in oncology* (Vol. 33, No. 4, pp. 407-420). WB Saunders.

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. *Canadian Journal of Emergency Medicine*, *8*(1), 19-20.

Fleck JL, Pavel AB, Cassandras, CG. Integrating mutation and gene expression cross-sectional data to infer cancer progression. *BMC Systems Biology* 10:12. DOI: 10.1186/s12918-016-0255-6, 2016.

Fleck JL, Pavel AB, Cassandras, CG. A pan-cancer analysis of progression mechanisms and drug sensitivity in cancer cell lines. *Molecular Omics* 15:399-405. DOI: 10.1039/c9mo00119k, 2019.

Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, et al. (2017). COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research* 45:D777-783.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, *3*(Mar), 1289-1305.

Frederick, B. A., Helfrich, B. A., Coldren, C. D., Zheng, D., Chan, D., Bunn, P. A., & Raben, D. (2007). Epithelial to mesenchymal transition predicts gefitinib resistance in cell lines of head and neck squamous cell carcinoma and non–small cell lung carcinoma. *Molecular cancer therapeutics*, *6*(6), 1683-1691.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, *28*(2), 337-407.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Gandhi, S. S., & Prabhune, S. S. (2017, January). Overview of feature subset selection algorithm for high dimensional data. In *2017 International Conference on Inventive Systems and Control (ICISC)* (pp. 1-6). IEEE.

Geeleher, P. et al. (2014). Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol.* 15: R47.

Geeleher P, Zhang Z, Wang F, Gruener RF, Nath A, Morrison G, Bhutra S, Grossman RL, Huang RS. (2017). Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomic studies. *Genome Research* 27:1743-1751.

Ginsburg, G. S., & Willard, H. F. (Eds.). (2009). *Essentials of genomic and personalized medicine*. Academic Press.

Gnana, D. A. A., Balamurugan, S. A. A., & Leavline, E. J. (2016). Literature review on feature selection methods for high-dimensional data. *International Journal of Computer Applications*, *975*, 8887.

Guan, N. N., Zhao, Y., Wang, C. C., Li, J. Q., Chen, X., & Piao, X. (2019). Anticancer drug response prediction in cell lines using weighted graph regularized matrix factorization. *Molecular Therapy-Nucleic Acids*, 17, 164-174.

Gunder, L., McClary, L. M. G., & Martin, S. (2011). *Essentials of medical genetics for health professionals*. Jones & Bartlett Learning.

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

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (Eds.). (2008). *Feature extraction: foundations and applications* (Vol. 207). Springer.

Hanahan D, Weinberg RA. (2011). Hallmarks of cancer: the next generation. *Cell* 144:646-674.

Harrington, P. (2012). *Machine learning in action*. Manning Publications Co.

Hegi, M. E., Diserens, A. C., Gorlia, T., Hamou, M. F., De Tribolet, N., Weller, M., ... & Bromberg, J. E. (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. *New England Journal of Medicine*, *352*(10), 997-1003.

Hirohashi, S., & Kanai, Y. (2003). Cell adhesion system and human cancer morphogenesis. *Cancer science*, *94*(7), 575-581.

Huang, C., Clayton, E. A., Matyunina, L. V., McDonald, L. D., Benigno, B. B., Vannberg, F., & McDonald, J. F. (2018). Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. *Scientific reports*, *8*(1), 1-8.

Huang EW, Bhope A, Lim J, Sinha S, Emad A. Tissue-guided LASSO for prediction of clinical drug response using preclinical samples. *PLoS Computational Biology* 16(1):e1007607. DOI:10.1371/journal.pcbi.1007607, 2020.

Hutter, C., & Zenklusen, J. C. (2018). The cancer genome atlas: creating lasting value beyond its data. Cell, 173(2), 283-285.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679-688.

ICGC (International Cancer Genome Consortium). International network of cancer genome projects. *Nature* 464:993-998, 2010.

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., ... & Cokelaer, T. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, *166*(3), 740-754.

Island, B. C. (2010). BRCA1 CpG island hypermethylation predicts sensitivity to poly (adenosine diphosphate)-ribose polymerase inhibitors. *J. Clin. Oncol*, *28*, e563-e564.

Jaiswal, A. K., Panshin, I., Shulkin, D., Aneja, N., & Abramov, S. (2019). Semi-supervised learning for cancer detection of lymph node metastases. *arXiv preprint arXiv:1906.09587*.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118-127.

References

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *28*(1), 27-30.

Kathleen Kerr, M. (2003). Design considerations for efficient and effective microarray studies. *Biometrics*, *59*(4), 822-828.

Khosravan, N., & Bagci, U. (2018). Semi-supervised multi-task learning for lung cancer diagnosis. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 710-713). IEEE.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine Learning Proceedings 1992* (pp. 249-256). Morgan Kaufmann.

Koestler, D. C., Marsit, C. J., Christensen, B. C., Karagas, M. R., Bueno, R., Sugarbaker, D. J., ... & Houseman, E. A. (2010). Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics*, *26*(20), 2578-2585.

Lander, E. S. (1999). Array of hope. *Nature genetics*, *21*(1), 3-4.

Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, Torres LC (2020). sva: Surrogate Variable Analysis. R package version 3.38.0.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, *11*(10), 733-739.

Li, C., & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, *98*(1), 31-36.

References

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, *50*(6), 1-45.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12), 1739-1740.

Lipson, K. E., Wong, C., Teng, Y., & Spong, S. (2012, December). CTGF is a central mediator of tissue remodeling and fibrosis and its inhibition can reverse the process of fibrosis. In *Fibrogenesis & tissue repair* (Vol. 5, No. S1, p. S24). BioMed Central.

Ma, T., & Zhang, A. (2018). Affinity network fusion and semi-supervised learning for cancer patient clustering. *Methods*, *145*, 16-24.

MacKay, D. J., & Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Maimon, O., & Rokach, L. (2002, February). Improving supervised learning by feature decomposition. In *International Symposium on Foundations of Information and Knowledge Systems* (pp. 178-196). Springer, Berlin, Heidelberg.

Masood, A., Al-Jumaily, A., & Anam, K. (2015). Self-supervised learning model for skin cancer diagnosis. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 1012-1015). IEEE.

Masters, J. R. (2000). Human cancer cell lines: fact and fantasy. *Nature reviews Molecular cell biology*, *1*(3), 233-236.

References

Maron, M. E. (1961). Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, *8*(3), 404-417.

McLeod, H. L. (2013). Cancer pharmacogenomics: early promise, but concerted effort needed. *Science* 339, 1563–1566.

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, 8(4), e61318.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1. 7–1.

Moughari, Fatemeh Ahmadi, and Changiz Eslahchi. (2020). ADRML: anticancer drug response prediction using manifold learning. *Scientific Reports*, 10(1), 1-18.

Müller, C., Schillert, A., Röthemeier, C., Trégouët, D. A., Proust, C., Binder, H., ... & Tiret, L. (2016). Removing batch effects from longitudinal gene expression-quantile normalization plus combat as best approach for microarray transcriptome data. *PloS one*, *11*(6), e0156594.

Naik, U. P., & Eckfeld, K. (2003). Junctional adhesion molecule 1 (JAM-1). *Journal of biological regulators and homeostatic agents*, *17*(4), 341-347.

Negrini, S., Gorgoulis, V. G., & Halazonetis, T. D. (2010). Genomic instability—an evolving hallmark of cancer. *Nature reviews Molecular cell biology*, 11(3), 220-228.

Nguyen, G. T., & Le, D. H. (2018). A matrix completion method for drug response prediction in personalized medicine. *In Proceedings of the Ninth International Symposium on Information and Communication Technology* (pp. 410-415). ACM.

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, *8*(8), 1551.

Oskooei, A., Manica, M., Mathis, R., & Martínez, M. R. (2019). Network-based biased tree ensembles (NetBiTE) for drug sensitivity prediction and drug sensitivity biomarker identification in cancer. *Scientific reports*, 9(1), 1-13.

Ostertagova, E., Ostertag, O., & Kováč, J. (2014). Methodology and application of the Kruskal-Wallis test. In *Applied Mechanics and Materials* (Vol. 611, pp. 115-120). Trans Tech Publications Ltd.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics*, *12*(2), 87-98.

Parca, L., Pepe, G., Pietrosanto, M., Galvan, G., Galli, L., Palmeri, A., ... & Helmer-Citterich, M. (2019). Modeling cancer drug response through drug-specific informative genes. *Scientific Reports*, 9(1), 1-11.

Park, C., Ahn, J., Kim, H., & Park, S. (2014). Integrative gene network construction to analyze cancer recurrence using semi-supervised learning. *PloS one*, *9*(1), e86309.

Peng, L., Chen, W., Zhou, W., Li, F., Yang, J., & Zhang, J. (2016). An immune-inspired semi-supervised algorithm for breast cancer diagnosis. *Computer methods and programs in biomedicine*, *134*, 259-265.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahman, R., Matlock, K., Ghosh, S., & Pal, R. (2017). Heterogeneity aware random forest for drug sensitivity prediction. *Scientific reports*, *7*(1), 1-11.

References

Rampášek, L., Hidru, D., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2019). Dr. VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics*, *35*(19), 3743-3751.

Rees, J. (2015). Temozolomide in low-grade gliomas: living longer and better.

Schiffner, J., Bischl, B., Lang, M., Richter, J., Jones, Z. M., Probst, P., ... & Thomas, J. (2016). mlr Tutorial. *arXiv preprint arXiv:1609.06146*.

Schliep, K., Hechenbichler, K., & Lizee, A. (2016). kknn: Weighted k-nearest neighbors. *R package version*, *1*(1).

Sebaugh, J. L. (2011). Guidelines for accurate EC50/IC50 estimation. *Pharmaceutical statistics*, *10*(2), 128-134.

Shen, L., Kondo, Y., Ahmed, S., Boumber, Y., Konishi, K., Guo, Y., ... & Issa, J. P. J. (2007). Drug sensitivity prediction by CpG island methylation profile in the NCI-60 cancer cell line panel. *Cancer Research*, 67(23), 11335-11343.

Shi, M., & Zhang, B. (2011). Semi-supervised learning improves gene expression-based prediction of cancer recurrence. *Bioinformatics*, *27*(21), 3017-3023.

Sims, A. H. (2009). Bioinformatics and breast cancer: what can high-throughput genomic approaches actually tell us?. *Journal of clinical pathology*, *62*(10), 879-885.

Smith, A. M., Walsh, J. R., Long, J., Davis, C. B., Henstock, P., Hodge, M. R., ... & Fisher, C. K. (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. BMC bioinformatics, 21(1), 1-18.

Spearman, C. (1961). The proof and measurement of association between two things.

Stanfield, Z., Coşkun, M., & Koyutürk, M. (2017). Drug response prediction as a link prediction problem. *Scientific reports*, 7, 40321.

Steinfeld, I., Navon, R., Ardigò, D., Zavaroni, I., & Yakhini, Z. (2008). Clinically driven semi-supervised class discovery in gene expression data. *Bioinformatics*, *24*(16), i90-i97.

Stetson, L. C., Pearl, T., Chen, Y., & Barnholtz-Sloan, J. S. (2014). Computational identification of multi-omic correlates of anticancer therapeutic response. *BMC genomics*, *15*(7), S2.

Su, R., Liu, X., Wei, L., & Zou, Q. (2019). Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods*.

Su, R., Liu, X., Xiao, G., & Wei, L. (2020). Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction. *Briefings in Bioinformatics*, *21*(3), 996-1005.

Suphavilai, C., Bertrand, D., & Nagarajan, N. (2018). Predicting cancer drug response using a recommender system. *Bioinformatics*, *34*(22), 3907-3914.

Surveillance, Epidemiology and End Result Program (SEER). Cancer stat facts: leukemia — Acute Myeloid Leukemia (AML). National Cancer Institute. https://seer.cancer.gov/statfacts/html/amyl.html (2020).

Szyf, M. (1994). DNA methylation properties: consequences for pharmacology. *Trends in Pharmacological Sciences*, *15*(7), 233-238.

Szyf, M. (2008). The role of DNA hypermethylation and demethylation in cancer and cancer therapy. *Current Oncology*, *15*(2), 72.

Thomson S, Buck E, Petti F, et al. Epithelial to mesenchymal transition is a determinant of sensitivity of non-small-cell lung carcinoma cell lines and xenografts to epidermal growth factor receptor inhibition. *Cancer Res* 2005;65:9455 – 62.

Tomczak K, Czerwinska P, Wiznerowicz M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology* 19:A68-77.

Triguero, I., García, S., & Herrera, F. (2015). Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, *42*(2), 245-284.

Turki, T., Wei, Z., & Wang, J. T. (2017). Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access*, *5*, 7381-7393.

Tyner, J. W., Tognon, C. E., Bottomly, D., Wilmot, B., Kurtz, S. E., Savage, S. L., ... & Agarwal, A. (2018). Functional genomic landscape of acute myeloid leukaemia. *Nature*, *562*(7728), 526-531.

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, *85*, 189-203.

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373-440.

Vapnik, V. (1998). The support vector method of function estimation. In *Nonlinear Modeling* (pp. 55-85). Springer, Boston, MA.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57-63.

Wang, L., Li, X., Zhang, L., & Gao, Q. (2017). Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC cancer*, *17*(1), 1-12.

Wang, X., Sun, Z., Zimmermann, M. T., Bugrim, A., & Kocher, J. P. (2019). Predict drug sensitivity of cancer cells with pathway activity inference. *BMC medical genomics*, *12*(1), 5-13.

Webber JT, Kaushik S, Bandyopadhyay S. Integration of tumor genomic data with cell lines using multi-dimensional network modules improves cancer pharmacogenomics. *Cell Systems* 7:526-536, 2018.

Weston, J., Leslie, C., Ie, E., Zhou, D., Elisseeff, A., & Noble, W. S. (2005). Semi-supervised protein classification using cluster kernels. *Bioinformatics*, *21*(15), 3241-3247.

Witta SE, Gemmill RM, Hirsch FR, et al. Restoring E-cadherin expression increases sensitivity to epidermal growth factor receptor inhibitors in lung cancer cell lines. *Cancer Res* 2006;66:944 – 50.

Xu, Xiaolu, Hong Gu, Yang Wang, Jia Wang, and Pan Qin. (2019). Autoencoder based feature selection method for classification of anticancer drug response. *Frontiers in Genetics*, 10, 233.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., … Garnett, M. J. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(Database issue), D955–D961. doi:10.1093/nar/gks1111.

Yao, Y., & Dai, W. (2014). Genomic instability and cancer. *Journal of carcinogenesis & mutagenesis*, *5*.

Yauch RL, Januario T, Eberhard DA, et al. (2005). Epithelial versus mesenchymal phenotype determines in vitro sensitivity and predicts clinical activity of erlotinib in lung cancer patients. *Clin Cancer Res*; 11:8686 – 98.

Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 856-863).

References

Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, et al. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature Biotechnology* 32(7):644-652.

Yuan, H., Paskov, I., Paskov, H., González, A. J., & Leslie, C. S. (2016). Multitask learning improves prediction of cancer drug sensitivity. *Scientific reports*, *6*, 31619.

Zemmal, N., Azizi, N., Dey, N., & Sellami, M. (2016). Adaptive semi supervised support vector machine semi supervised learning with features cooperation for breast cancer classification. *Journal of Medical Imaging and Health Informatics*, *6*(1), 53-62.

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., & Liu, X. S. (2015). Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Comput Biol*, *11*(9), e1004498.

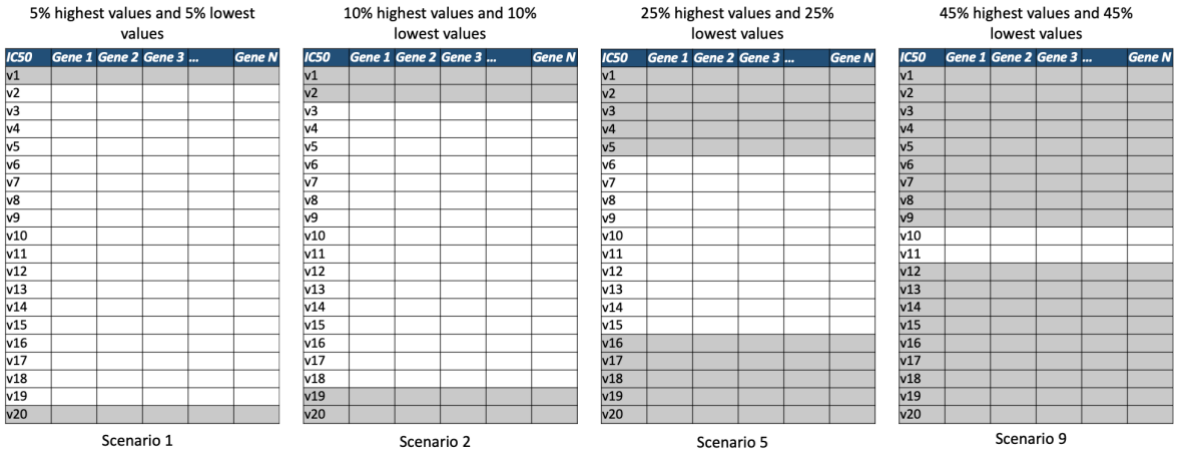
Zhao Q, Shi X, Xie Y, Huang J, Shia B, Ma S. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Briefings in Bioinformatics* 16(2):291-303, 2014.

Zhu, Y., Brettin, T., Evrard, Y. A., Partin, A., Xia, F., Shukla, M., ... & Stevens, R. L. (2020). Ensemble transfer learning for the prediction of anti-cancer drug response. *Scientific reports*, *10*(1), 1-11.

Zygmunt Zawadzki and Marcin Kosinski (2020). FSelectorRcpp: 'Rcpp' Implementation of 'FSelector' Entropy-Based Feature Selection Algorithms with a Sparse Matrix Support. R package version 0.3.3. https://CRAN.R-project.org/package=FSelectorRcpp

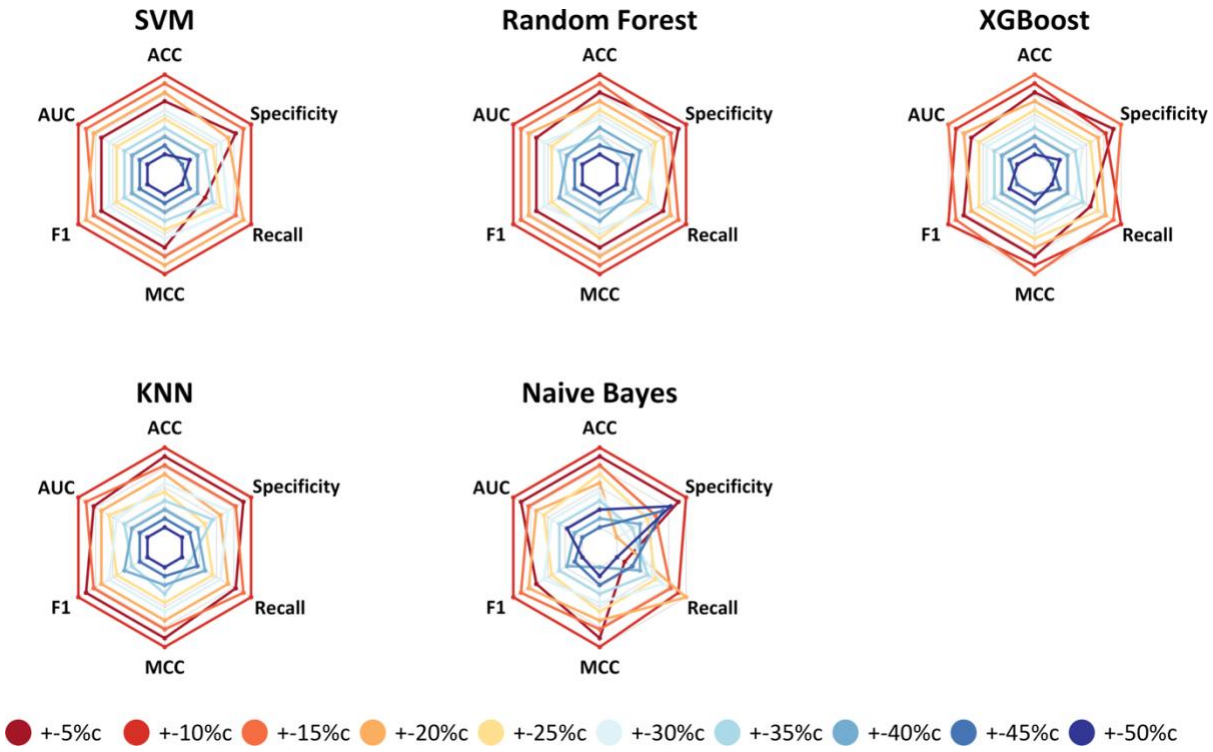
5
Appendix

5.1
Supplementary Figures

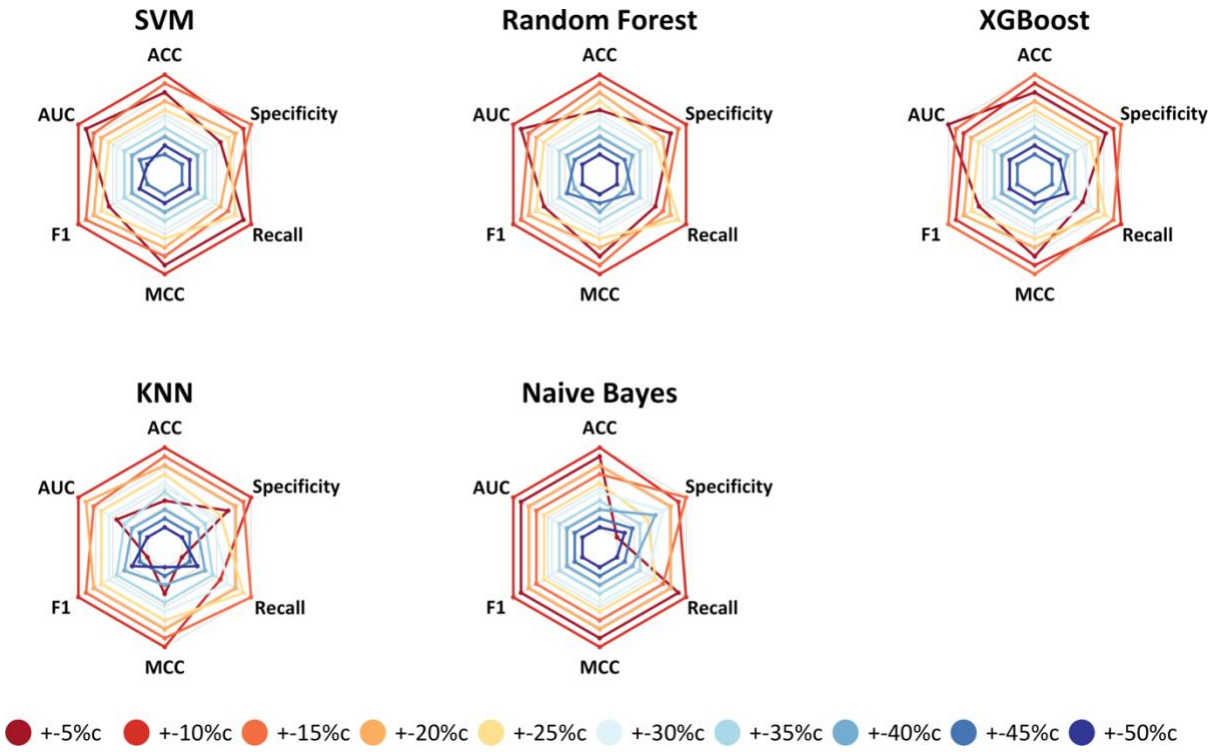


(Supplementary Figure) Figure S1: Example of subsampling process. When performing classification, we discretized drug-response (IC_{50}) values. To evaluate alternative thresholds for discretization, we performed a subsampling analysis. In Scenario 1 illustrated above, we considered the cell lines with the lowest and highest 5% of IC_{50} values. In Scenario 2, we considered the cell lines with the lowest and highest 10% of IC_{50} values. Each scenario used 10% more data than the previous scenario (5% on each side). This pattern continues until all data were considered in the analysis.

Appendix

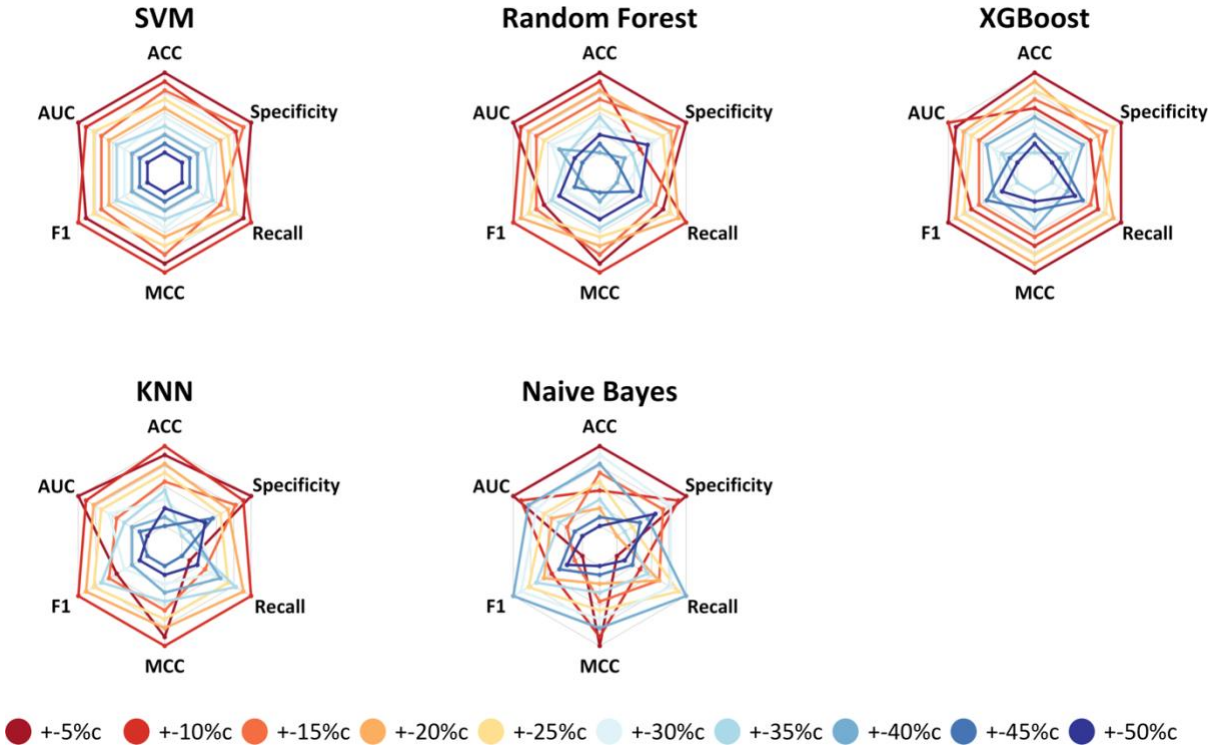


(Supplementary Figure) Figure S2: Graphs for Cisplatin classification analysis. The graphs compare different scenarios ranked in order of best result. GDSC cell-line data were used to generate ten subsampling scenarios, which we then tested via nested cross validation. Scenarios that are further away from the center represent higher metric values than scenarios closer to it. The evaluated metrics for each algorithm are accuracy (ACC), specificity, recall, Matthews correlation coefficient (MCC), F1 score (F1) and area under the receiver operating characteristic curve (AUC).



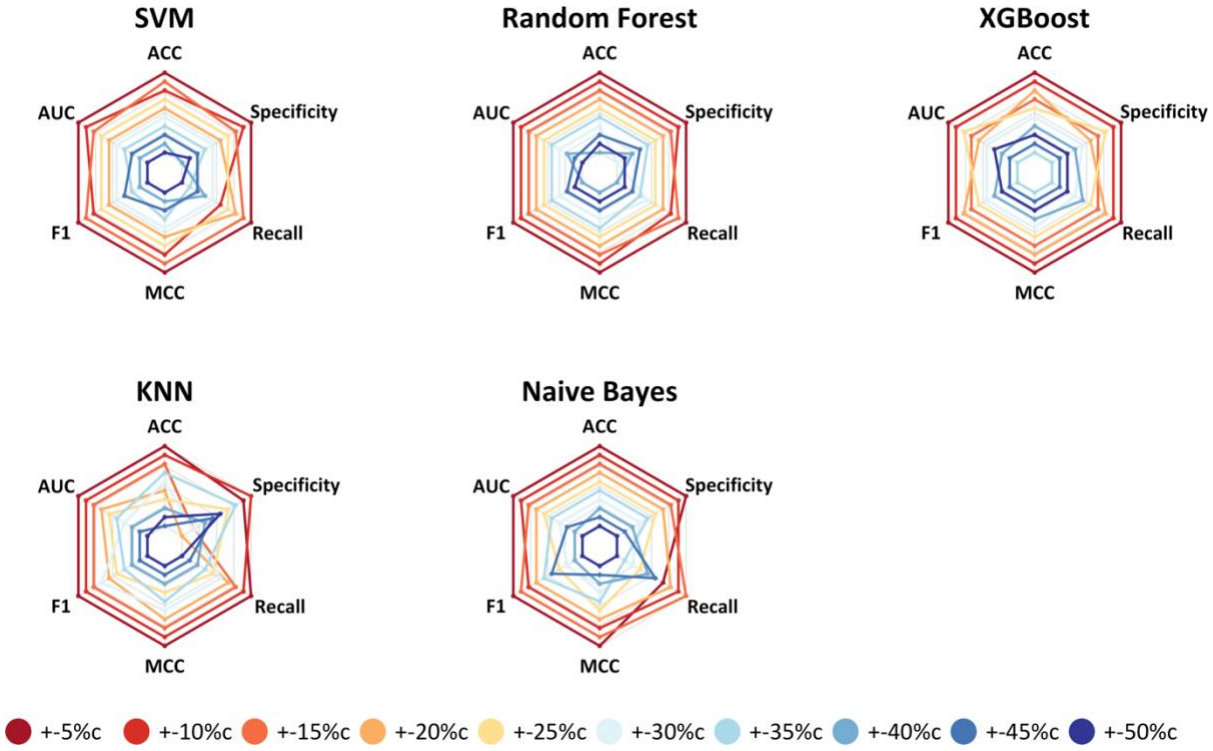
(Supplementary Figure) Figure S3: Graphs for Docetaxel classification analysis. The graphs compare different scenarios ranked in order of best result. GDSC cell-line data were used to generate ten subsampling scenarios, which we then tested via nested cross validation. Scenarios that are further away from the center represent higher metric values than scenarios closer to it. The evaluated metrics for each algorithm are accuracy (ACC), specificity, recall, Matthews correlation coefficient (MCC), F1 score (F1) and area under the receiver operating characteristic curve (AUC).

Appendix



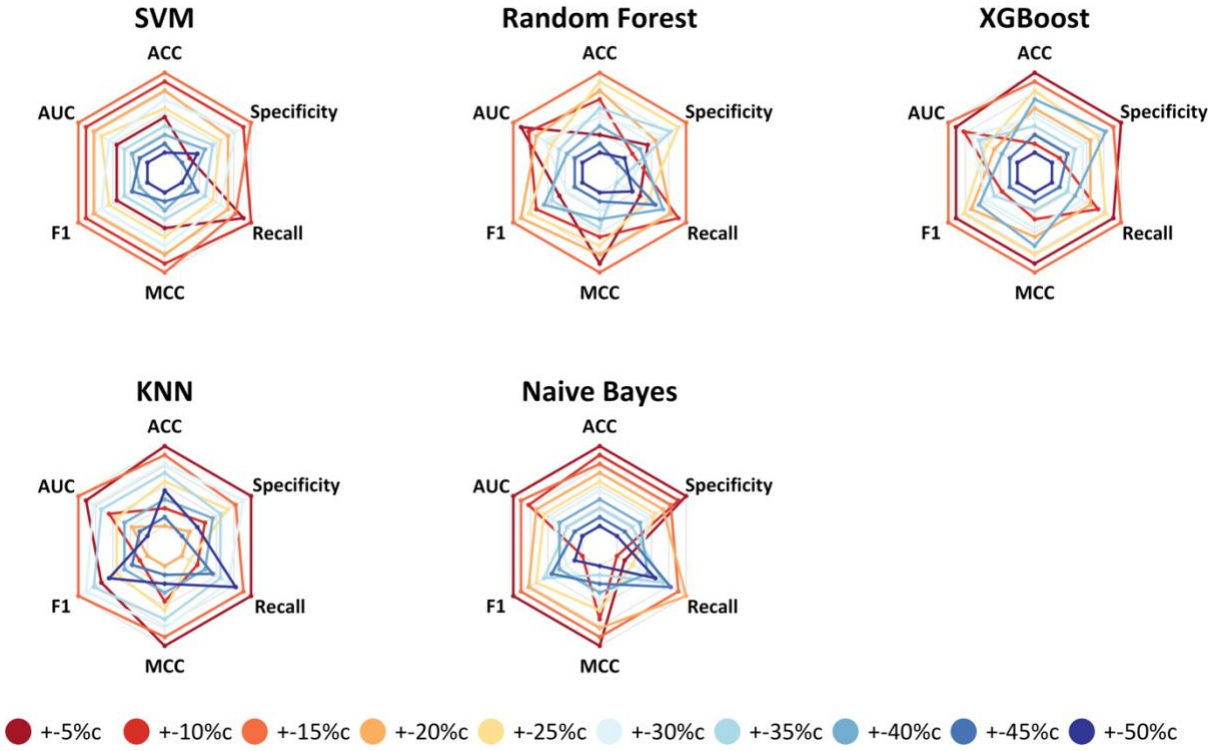
(Supplementary Figure) Figure S4: Graphs for Doxorubicin classification analysis. The graphs compare different scenarios ranked in order of best result. GDSC cell-line data were used to generate ten subsampling scenarios, which we then tested via nested cross validation. Scenarios that are further away from the center represent higher metric values than scenarios closer to it. The evaluated metrics for each algorithm are accuracy (ACC), specificity, recall, Matthews correlation coefficient (MCC), F1 score (F1) and area under the receiver operating characteristic curve (AUC).

Appendix



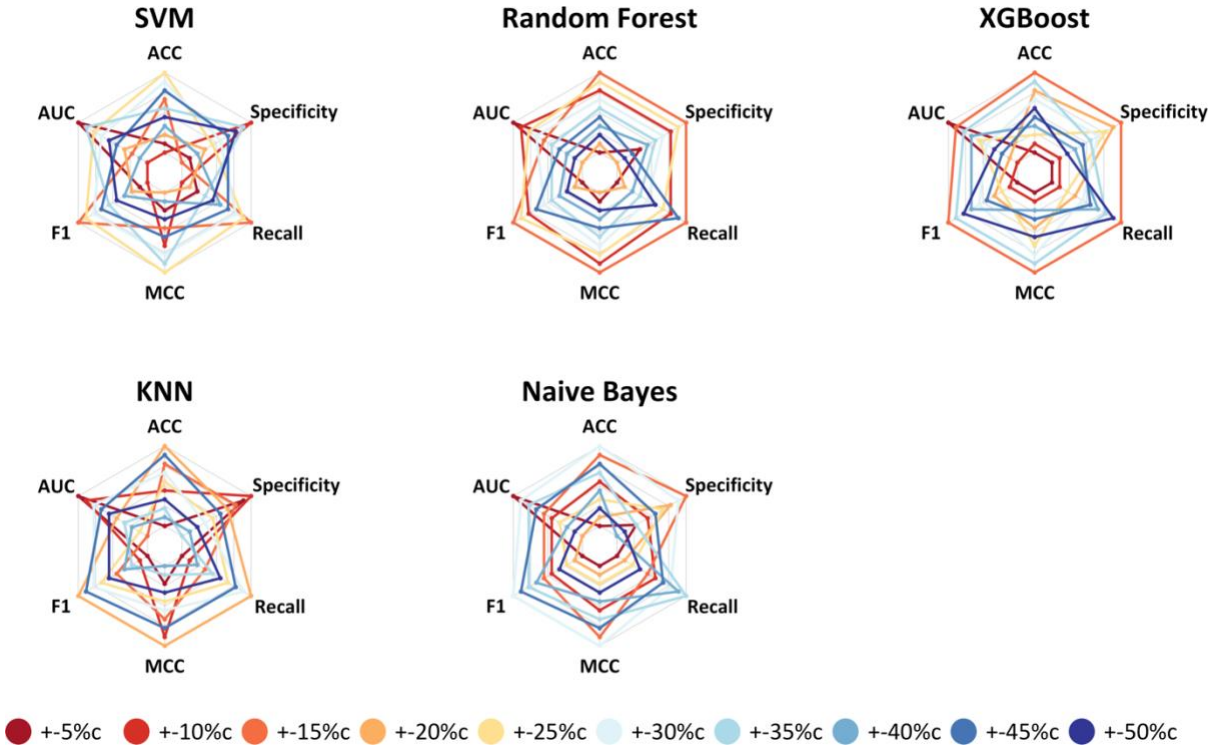
(Supplementary Figure) Figure S5: Graphs for Etoposide classification analysis. The graphs compare different scenarios ranked in order of best result. GDSC cell-line data were used to generate ten subsampling scenarios, which we then tested via nested cross validation. Scenarios that are further away from the center represent higher metric values than scenarios closer to it. The evaluated metrics for each algorithm are accuracy (ACC), specificity, recall, Matthews correlation coefficient (MCC), F1 score (F1) and area under the receiver operating characteristic curve (AUC).

Appendix



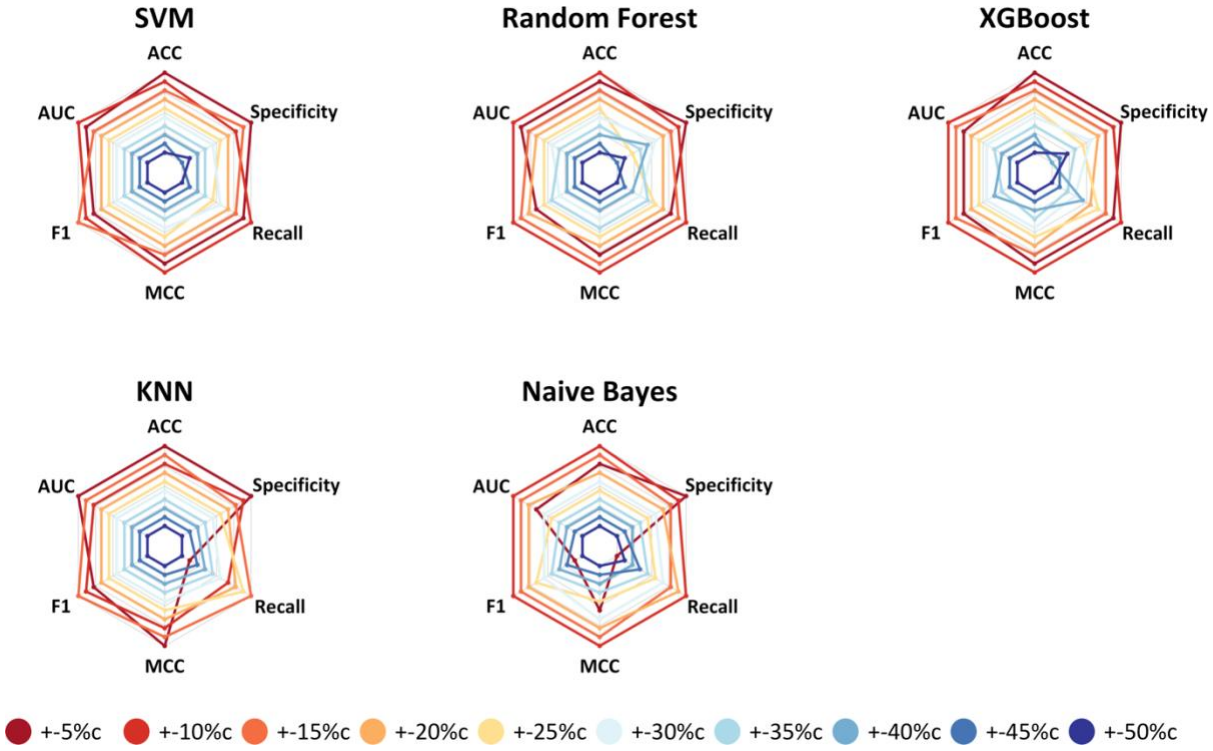
(Supplementary Figure) Figure S6: Graphs for Gemcitabine classification analysis. The graphs compare different scenarios ranked in order of best result. GDSC cell-line data were used to generate ten subsampling scenarios, which we then tested via nested cross validation. Scenarios that are further away from the center represent higher metric values than scenarios closer to it. The evaluated metrics for each algorithm are accuracy (ACC), specificity, recall, Matthews correlation coefficient (MCC), F1 score (F1) and area under the receiver operating characteristic curve (AUC).

Appendix



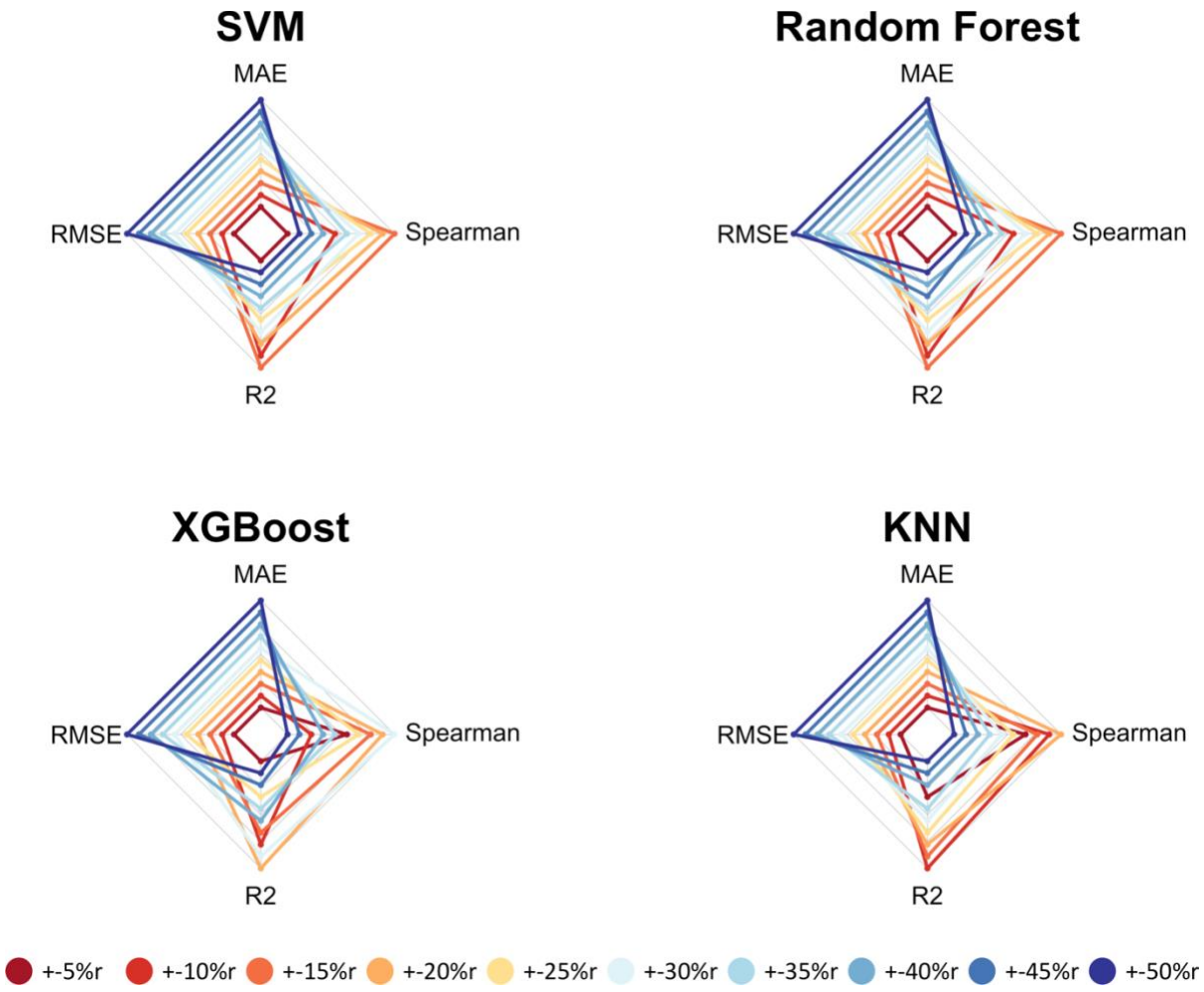
(Supplementary Figure) Figure S7: Graphs for Paclitaxel classification analysis. The graphs compare different scenarios ranked in order of best result. GDSC cell-line data were used to generate ten subsampling scenarios, which we then tested via nested cross validation. Scenarios that are further away from the center represent higher metric values than scenarios closer to it. The evaluated metrics for each algorithm are accuracy (ACC), specificity, recall, Matthews correlation coefficient (MCC), F1 score (F1) and area under the receiver operating characteristic curve (AUC).

Appendix



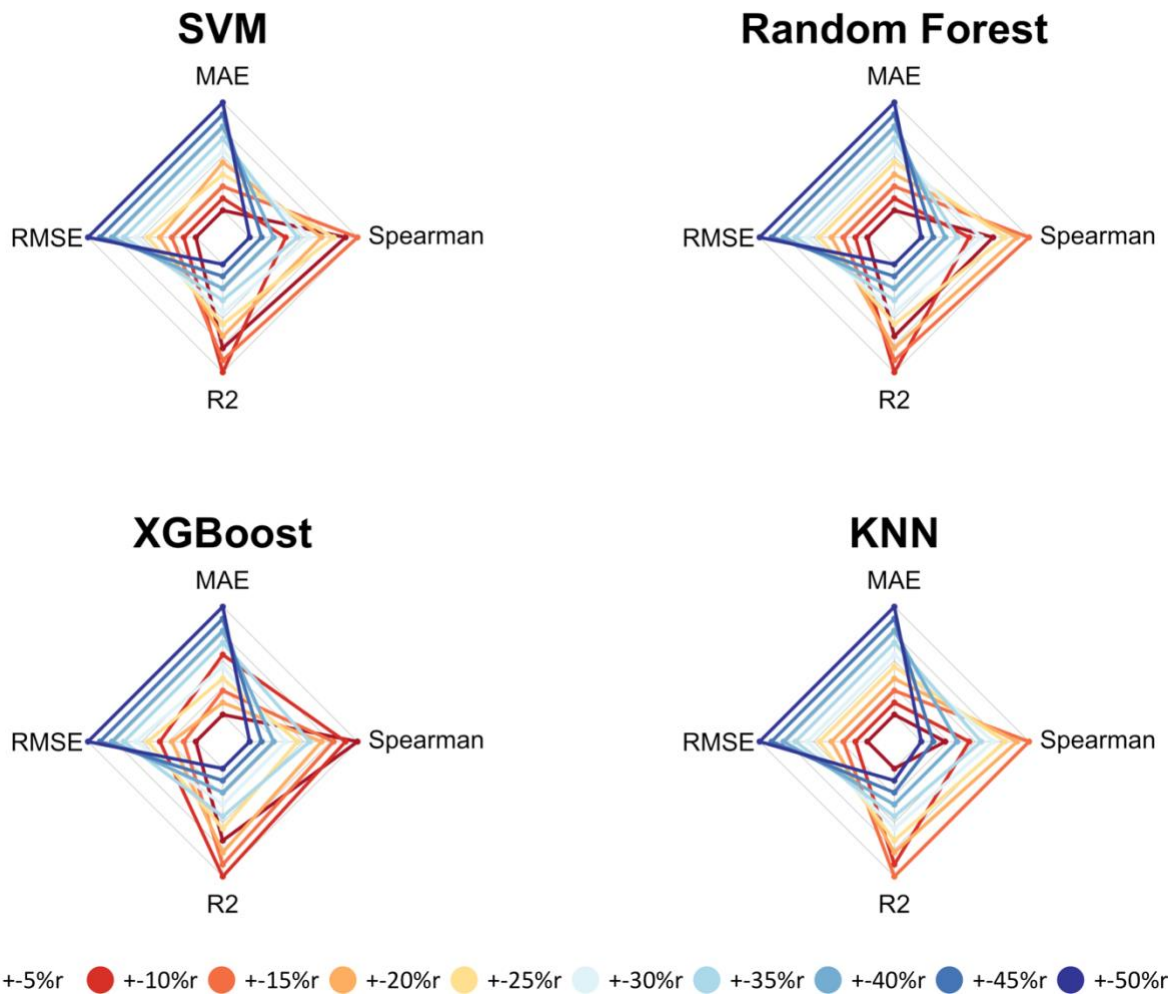
(Supplementary Figure) Figure S8: Graphs for Temozolomide classification analysis. The graphs compare different scenarios ranked in order of best result. GDSC cell-line data were used to generate ten subsampling scenarios, which we then tested via nested cross validation. Scenarios that are further away from the center represent higher metric values than scenarios closer to it. The evaluated metrics for each algorithm are accuracy (ACC), specificity, recall, Matthews correlation coefficient (MCC), F1 score (F1) and area under the receiver operating characteristic curve (AUC).

Appendix



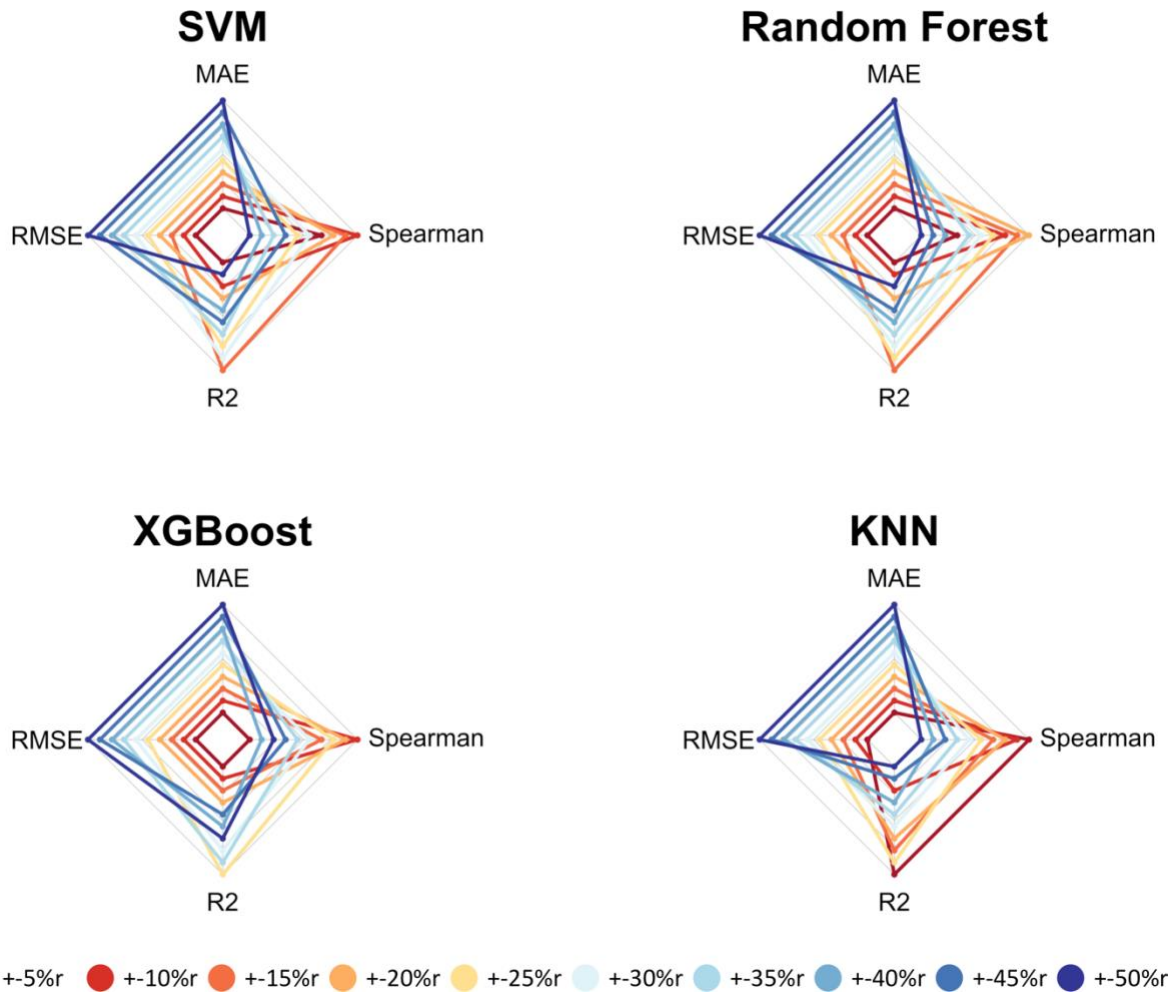
(Supplementary Figure) Figure S9: Graphs for Cisplatin regression analysis. We used DNA methylation data from cell lines to predict continuous IC_{50} response values using four regression algorithms. We evaluated the algorithms' performance via nested cross validation for ten subsampling scenarios. Graphs illustrate performance for these scenarios, ranked in order of relative performance for four metrics: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), R-squared and Spearman correlation coefficient. Scenarios further away from the center represent relatively low metric values (and thus better performance). Scenarios that used all cell lines performed best for all algorithms.

Appendix

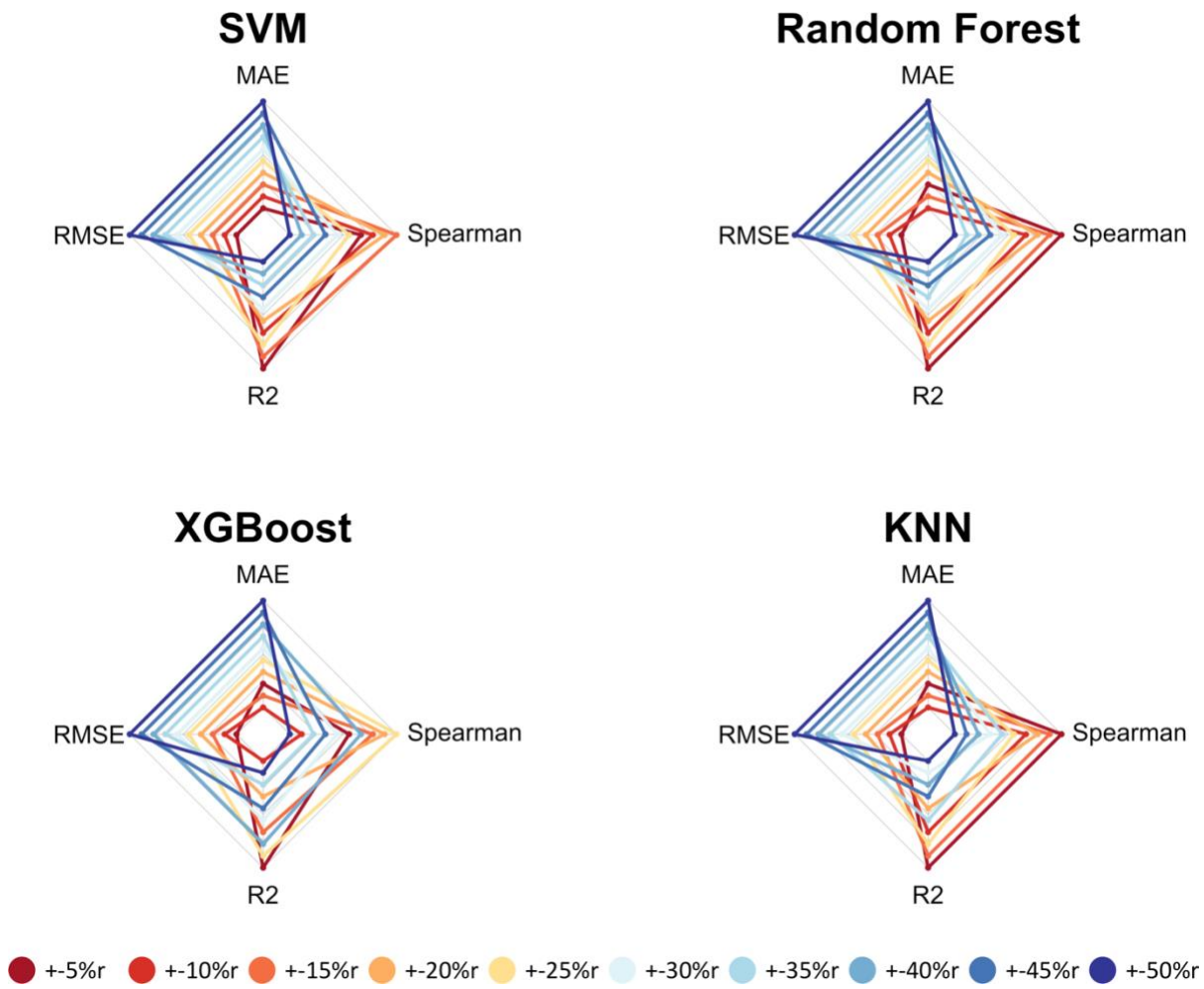


(Supplementary Figure) Figure S10: Graphs for Docetaxel regression analysis. We used DNA methylation data from cell lines to predict continuous IC_{50} response values using four regression algorithms. We evaluated the algorithms' performance via nested cross validation for ten subsampling scenarios. Graphs illustrate performance for these scenarios, ranked in order of relative performance for four metrics: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), R-squared and Spearman correlation coefficient. Scenarios further away from the center represent relatively low metric values (and thus better performance). Scenarios that used all cell lines performed best for all algorithms.

Appendix

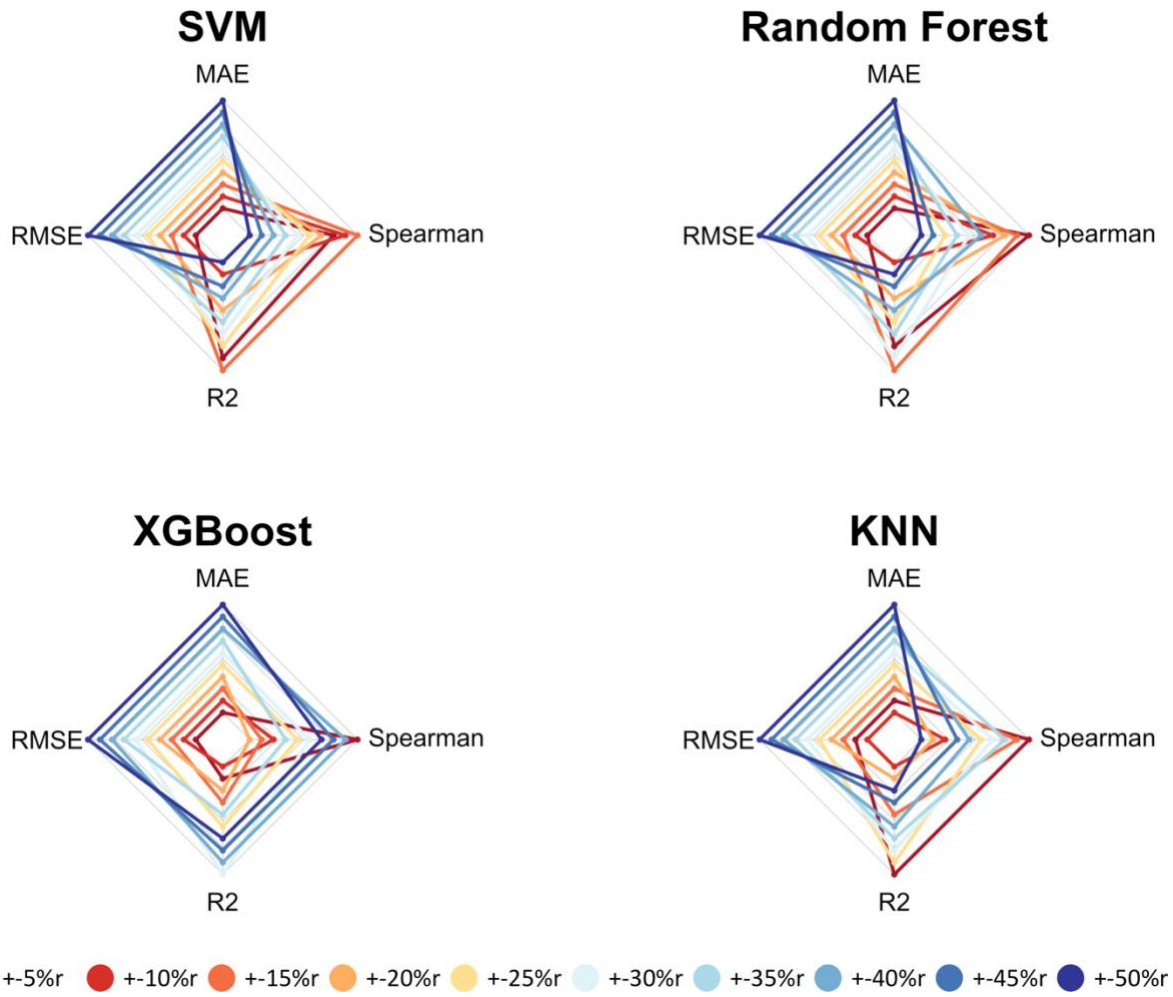


(Supplementary Figure) Figure S11: Graphs for Doxorubicin regression analysis. We used DNA methylation data from cell lines to predict continuous IC_{50} response values using four regression algorithms. We evaluated the algorithms' performance via nested cross validation for ten subsampling scenarios. Graphs illustrate performance for these scenarios, ranked in order of relative performance for four metrics: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), R-squared and Spearman correlation coefficient. Scenarios further away from the center represent relatively low metric values (and thus better performance). Scenarios that used all cell lines performed best for all algorithms.

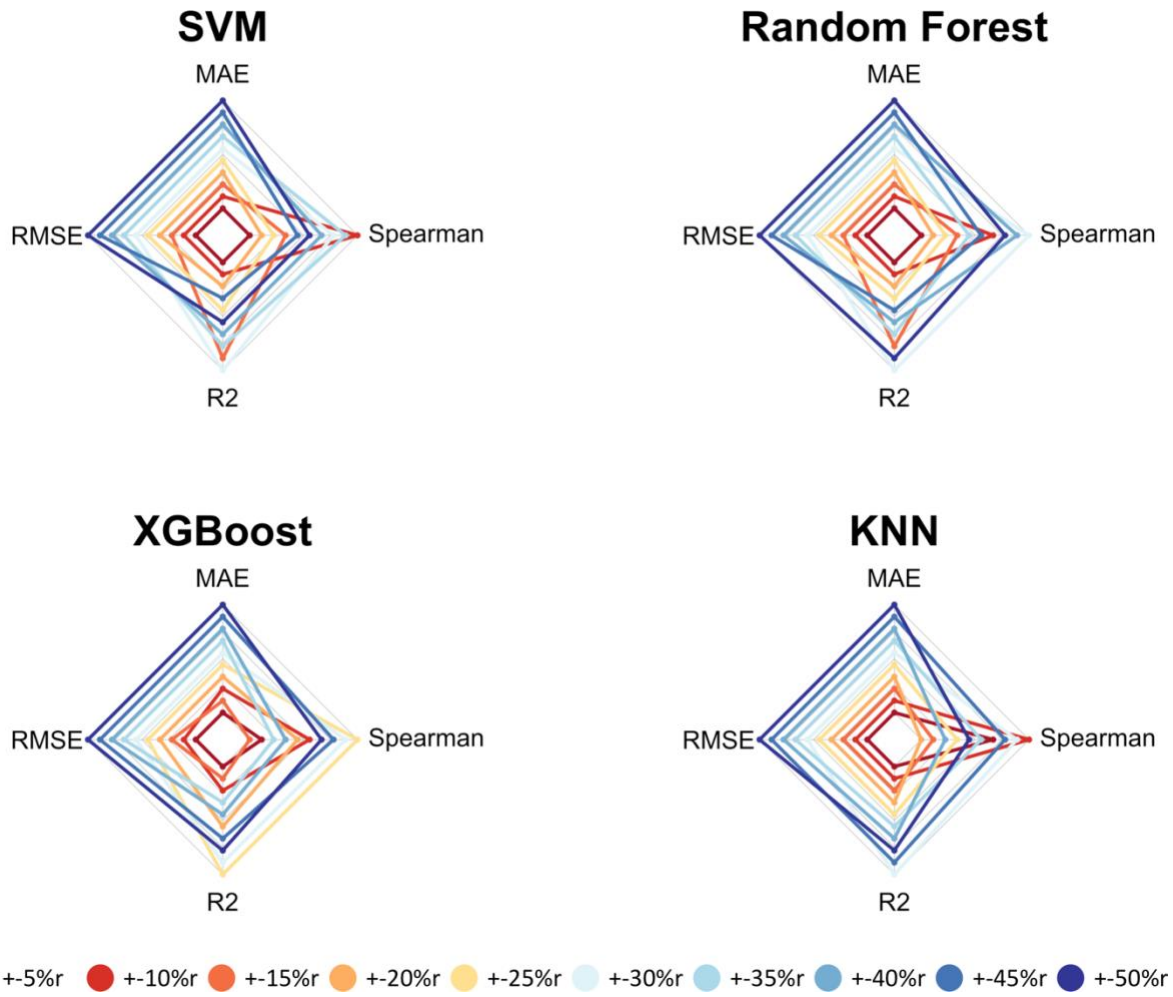


(Supplementary Figure) Figure S12: Graphs for Etoposide regression analysis. We used DNA methylation data from cell lines to predict continuous IC_{50} response values using four regression algorithms. We evaluated the algorithms' performance via nested cross validation for ten subsampling scenarios. Graphs illustrate performance for these scenarios, ranked in order of relative performance for four metrics: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), R-squared and Spearman correlation coefficient. Scenarios further away from the center represent relatively low metric values (and thus better performance). Scenarios that used all cell lines performed best for all algorithms.

Appendix

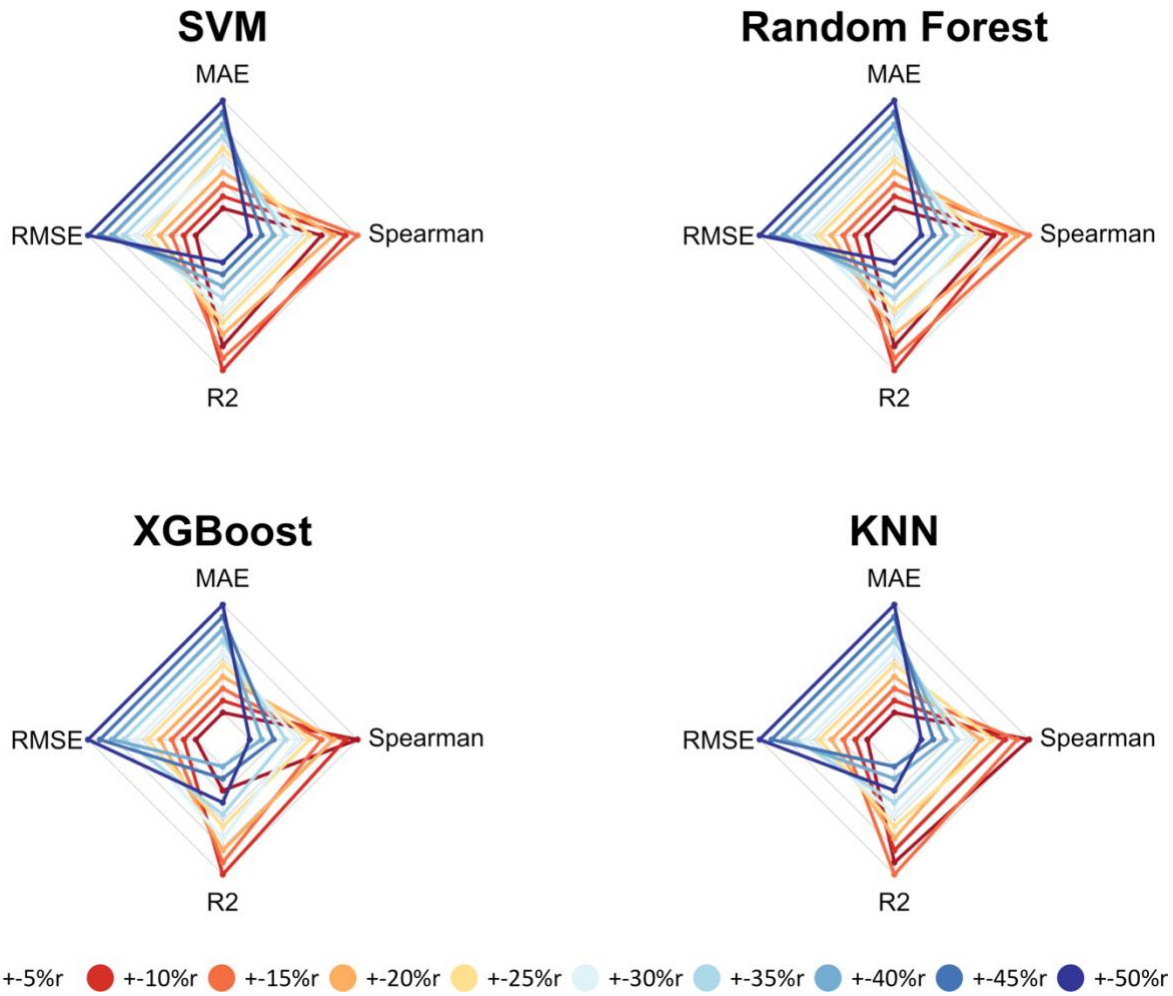


(Supplementary Figure) Figure S13: Graphs for Gemcitabine regression analysis. We used DNA methylation data from cell lines to predict continuous IC_{50} response values using four regression algorithms. We evaluated the algorithms' performance via nested cross validation for ten subsampling scenarios. Graphs illustrate performance for these scenarios, ranked in order of relative performance for four metrics: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), R-squared and Spearman correlation coefficient. Scenarios further away from the center represent relatively low metric values (and thus better performance). Scenarios that used all cell lines performed best for all algorithms.



(Supplementary Figure) Figure S14: Graphs for Paclitaxel regression analysis. We used DNA methylation data from cell lines to predict continuous IC_{50} response values using four regression algorithms. We evaluated the algorithms' performance via nested cross validation for ten subsampling scenarios. Graphs illustrate performance for these scenarios, ranked in order of relative performance for four metrics: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), R-squared and Spearman correlation coefficient. Scenarios further away from the center represent relatively low metric values (and thus better performance). Scenarios that used all cell lines performed best for all algorithms.

Appendix



(Supplementary Figure) Figure S15: Graphs for Temozolomide regression analysis. We used DNA methylation data from cell lines to predict continuous IC_{50} response values using four regression algorithms. We evaluated the algorithms' performance via nested cross validation for ten subsampling scenarios. Graphs illustrate performance for these scenarios, ranked in order of relative performance for four metrics: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), R-squared and Spearman correlation coefficient. Scenarios further away from the center represent relatively low metric values (and thus better performance). Scenarios that used all cell lines performed best for all algorithms.

5.2
Supplementary Tables

(Supplementary Table) Table S1: Classification results for all combinations of subsampling scenarios and algorithms for Cisplatin. Bold font indicates the best-performing combination for each metric.

Scenario	Method	ACC	AUC	F1	MCC	Recall	Specificity
+-5%c	SVM	0.74	0.80	0.72	0.46	0.70	0.78
+-5%c	Random Forest	0.72	0.79	0.70	0.42	0.67	0.76
+-5%c	KNN	0.74	0.78	0.74	0.43	0.72	0.76
+-5%c	XGBoost	0.76	0.81	0.72	0.53	0.70	0.83
+-5%c	Naïve Bayes	0.67	0.67	0.66	0.32	0.65	0.67
+-10%c	SVM	**0.84**	**0.88**	**0.83**	**0.67**	**0.84**	**0.85**
+-10%c	Random Forest	0.79	0.86	0.79	0.58	0.77	0.81
+-10%c	KNN	0.81	0.84	0.79	0.60	0.78	0.84
+-10%c	XGBoost	0.77	0.85	0.77	0.55	0.77	0.77
+-10%c	Naïve Bayes	0.74	0.73	0.75	0.46	0.80	0.68
+-15%c	SVM	0.77	**0.88**	0.75	0.54	0.75	0.78
+-15%c	Random Forest	0.74	0.84	0.73	0.49	0.73	0.76
+-15%c	KNN	0.72	0.80	0.71	0.43	0.72	0.71
+-15%c	XGBoost	0.79	0.87	0.77	0.59	0.75	0.83
+-15%c	Naïve Bayes	0.64	0.64	0.68	0.31	0.79	0.50
+-20%c	SVM	0.76	0.84	0.76	0.55	0.77	0.76
+-20%c	Random Forest	0.71	0.79	0.70	0.45	0.69	0.74
+-20%c	KNN	0.69	0.77	0.68	0.38	0.69	0.69
+-20%c	XGBoost	0.73	0.84	0.72	0.49	0.71	0.76
+-20%c	Naïve Bayes	0.62	0.62	0.67	0.26	0.80	0.43
+-25%c	SVM	0.70	0.77	0.69	0.39	0.70	0.69
+-25%c	Random Forest	0.68	0.75	0.67	0.37	0.65	0.72
+-25%c	KNN	0.65	0.72	0.64	0.29	0.63	0.67
+-25%c	XGBoost	0.72	0.80	0.71	0.45	0.70	0.75
+-25%c	Naïve Bayes	0.62	0.62	0.66	0.25	0.76	0.48
+-30%c	SVM	0.71	0.79	0.71	0.42	0.71	0.71
+-30%c	Random Forest	0.68	0.76	0.67	0.37	0.65	0.72
+-30%c	KNN	0.66	0.72	0.65	0.32	0.63	0.69
+-30%c	XGBoost	0.69	0.77	0.68	0.39	0.65	0.73
+-30%c	Naïve Bayes	0.61	0.62	0.66	0.24	0.77	0.45
+-35%c	SVM	0.69	0.76	0.69	0.38	0.70	0.67
+-35%c	Random Forest	0.64	0.72	0.63	0.29	0.62	0.67

+-35%c	KNN	0.63	0.68	0.61	0.27	0.59	0.68
+-35%c	XGBoost	0.68	0.75	0.67	0.37	0.67	0.69
+-35%c	Naïve Bayes	0.59	0.60	0.63	0.18	0.71	0.47
+-40%c	SVM	0.67	0.72	0.67	0.34	0.68	0.66
+-40%c	Random Forest	0.65	0.71	0.64	0.30	0.61	0.69
+-40%c	KNN	0.63	0.66	0.62	0.26	0.62	0.64
+-40%c	XGBoost	0.65	0.72	0.64	0.30	0.62	0.68
+-40%c	Naïve Bayes	0.58	0.60	0.62	0.16	0.69	0.48
+-45%c	SVM	0.63	0.70	0.63	0.26	0.62	0.63
+-45%c	Random Forest	0.63	0.69	0.62	0.27	0.59	0.68
+-45%c	KNN	0.60	0.64	0.60	0.21	0.59	0.61
+-45%c	XGBoost	0.64	0.70	0.63	0.29	0.61	0.67
+-45%c	Naïve Bayes	0.58	0.59	0.61	0.17	0.66	0.50
+-50%c	SVM	0.62	0.66	0.61	0.24	0.60	0.64
+-50%c	Random Forest	0.61	0.65	0.60	0.22	0.59	0.63
+-50%c	KNN	0.57	0.60	0.57	0.14	0.57	0.57
+-50%c	XGBoost	0.64	0.69	0.63	0.29	0.61	0.68
+-50%c	Naïve Bayes	0.58	0.60	0.60	0.16	0.65	0.51

(Supplementary Table) Table S2: Classification results for all combinations of subsampling scenarios and algorithms for Docetaxel. Bold font indicates the best-performing combination for each metric.

Scenario	Method	ACC	AUC	F1	MCC	Recall	Specificity
+-5%c	SVM	0.85	0.95	0.84	0.73	0.89	0.80
+-5%c	Random Forest	0.82	0.93	0.81	0.67	0.80	0.83
+-5%c	KNN	0.72	0.81	0.66	0.47	0.63	0.80
+-5%c	XGBoost	0.84	0.96	0.82	0.69	0.83	0.85
+-5%c	Naïve Bayes	0.74	0.79	0.73	0.54	0.72	0.76
+-10%c	SVM	**0.89**	**0.97**	**0.89**	**0.78**	**0.91**	0.87
+-10%c	Random Forest	0.88	**0.97**	0.87	0.76	0.89	0.87
+-10%c	KNN	0.83	0.91	0.80	0.66	0.74	**0.91**
+-10%c	XGBoost	0.86	0.94	0.85	0.71	0.86	0.86
+-10%c	Naïve Bayes	0.79	0.81	0.78	0.58	0.73	0.86
+-15%c	SVM	0.86	0.93	0.86	0.72	0.85	0.88
+-15%c	Random Forest	0.84	0.92	0.83	0.69	0.83	0.85
+-15%c	KNN	0.82	0.87	0.80	0.64	0.77	0.86
+-15%c	XGBoost	0.86	0.94	0.85	0.72	0.86	0.86
+-15%c	Naïve Bayes	0.69	0.75	0.62	0.41	0.53	0.86
+-20%c	SVM	0.85	0.92	0.84	0.69	0.84	0.85
+-20%c	Random Forest	0.83	0.91	0.83	0.66	0.84	0.82
+-20%c	KNN	0.79	0.88	0.78	0.60	0.75	0.84
+-20%c	XGBoost	0.84	0.91	0.84	0.67	0.84	0.83
+-20%c	Naïve Bayes	0.71	0.76	0.65	0.43	0.57	0.85
+-25%c	SVM	0.84	0.90	0.84	0.68	0.85	0.83
+-25%c	Random Forest	0.82	0.88	0.83	0.65	0.84	0.81
+-25%c	KNN	0.77	0.85	0.77	0.56	0.75	0.80
+-25%c	XGBoost	0.83	0.89	0.83	0.66	0.84	0.81
+-25%c	Naïve Bayes	0.66	0.72	0.59	0.35	0.49	0.83
+-30%c	SVM	0.79	0.86	0.79	0.58	0.81	0.77
+-30%c	Random Forest	0.80	0.85	0.80	0.59	0.80	0.79
+-30%c	KNN	0.77	0.83	0.76	0.54	0.74	0.79
+-30%c	XGBoost	0.81	0.87	0.82	0.63	0.83	0.80
+-30%c	Naïve Bayes	0.65	0.71	0.57	0.33	0.47	0.83
+-35%c	SVM	0.77	0.84	0.77	0.55	0.78	0.77
+-35%c	Random Forest	0.75	0.82	0.75	0.51	0.75	0.76
+-35%c	KNN	0.73	0.78	0.73	0.47	0.73	0.74
+-35%c	XGBoost	0.77	0.83	0.77	0.55	0.77	0.78
+-35%c	Naïve Bayes	0.64	0.68	0.56	0.31	0.47	0.82
+-40%c	SVM	0.73	0.81	0.73	0.46	0.75	0.71

Appendix

+-40%c	Random Forest	0.72	0.79	0.70	0.43	0.69	0.74
+-40%c	KNN	0.71	0.75	0.70	0.42	0.69	0.73
+-40%c	XGBoost	0.73	0.79	0.72	0.45	0.71	0.74
+-40%c	Naïve Bayes	0.64	0.67	0.55	0.29	0.45	0.83
+-45%c	SVM	0.70	0.77	0.70	0.40	0.71	0.68
+-45%c	Random Forest	0.71	0.77	0.71	0.43	0.71	0.71
+-45%c	KNN	0.67	0.72	0.66	0.34	0.67	0.66
+-45%c	XGBoost	0.69	0.76	0.68	0.38	0.68	0.70
+-45%c	Naïve Bayes	0.62	0.65	0.53	0.26	0.43	0.81
+-50%c	SVM	0.70	0.76	0.70	0.40	0.71	0.69
+-50%c	Random Forest	0.71	0.76	0.71	0.41	0.71	0.70
+-50%c	KNN	0.66	0.71	0.67	0.32	0.68	0.64
+-50%c	XGBoost	0.71	0.78	0.71	0.42	0.71	0.71
+-50%c	Naïve Bayes	0.60	0.64	0.48	0.21	0.38	0.81

(Supplementary Table) Table S3: Classification results for all combinations of subsampling scenarios and algorithms for Doxorubicin. Bold font indicates the best-performing combination for each metric.

Scenario	Method	ACC	AUC	F1	MCC	Recall	Specificity
+-5%c	SVM	**0.71**	**0.80**	0.69	**0.45**	0.70	**0.72**
+-5%c	Random Forest	0.66	0.78	0.61	0.36	0.63	0.70
+-5%c	KNN	0.65	0.79	0.57	0.36	0.51	0.79
+-5%c	XGBoost	0.70	0.69	0.67	0.45	0.65	0.74
+-5%c	Naïve Bayes	0.61	0.64	0.56	0.25	0.53	0.67
+-10%c	SVM	**0.71**	**0.80**	**0.71**	**0.45**	**0.72**	0.70
+-10%c	Random Forest	0.65	0.76	0.67	0.37	0.71	0.59
+-10%c	KNN	0.67	0.72	0.67	0.39	0.66	0.67
+-10%c	XGBoost	0.63	0.71	0.62	0.28	0.62	0.64
+-10%c	Naïve Bayes	0.57	0.59	0.62	0.20	0.72	0.42
+-15%c	SVM	0.69	0.72	0.67	0.39	0.67	0.71
+-15%c	Random Forest	0.64	0.70	0.62	0.29	0.62	0.66
+-15%c	KNN	0.60	0.63	0.57	0.21	0.54	0.66
+-15%c	XGBoost	0.63	0.68	0.61	0.26	0.60	0.66
+-15%c	Naïve Bayes	0.57	0.57	0.61	0.14	0.74	0.40
+-20%c	SVM	0.65	0.70	0.65	0.32	0.67	0.64
+-20%c	Random Forest	0.64	0.70	0.64	0.28	0.63	0.65
+-20%c	KNN	0.63	0.67	0.61	0.25	0.60	0.66
+-20%c	XGBoost	0.65	0.68	0.64	0.29	0.65	0.65
+-20%c	Naïve Bayes	0.56	0.57	0.62	0.12	0.74	0.37
+-25%c	SVM	0.68	0.73	0.68	0.36	0.69	0.67
+-25%c	Random Forest	0.63	0.70	0.63	0.27	0.63	0.63
+-25%c	KNN	0.61	0.65	0.59	0.22	0.57	0.65
+-25%c	XGBoost	0.64	0.68	0.63	0.29	0.62	0.67
+-25%c	Naïve Bayes	0.57	0.57	0.63	0.15	0.75	0.39
+-30%c	SVM	0.63	0.68	0.63	0.27	0.63	0.63
+-30%c	Random Forest	0.59	0.65	0.59	0.18	0.59	0.60
+-30%c	KNN	0.59	0.64	0.57	0.17	0.55	0.63
+-30%c	XGBoost	0.58	0.62	0.57	0.16	0.56	0.60
+-30%c	Naïve Bayes	0.58	0.57	0.64	0.17	0.75	0.41
+-35%c	SVM	0.63	0.68	0.63	0.26	0.63	0.62
+-35%c	Random Forest	0.60	0.65	0.60	0.19	0.60	0.59
+-35%c	KNN	0.59	0.62	0.58	0.18	0.57	0.61
+-35%c	XGBoost	0.57	0.61	0.56	0.14	0.55	0.58
+-35%c	Naïve Bayes	0.56	0.57	0.63	0.13	0.74	0.38
+-40%c	SVM	0.61	0.66	0.60	0.24	0.60	0.62

+-40%c	Random Forest	0.57	0.65	0.57	0.16	0.56	0.58
+-40%c	KNN	0.58	0.62	0.57	0.18	0.55	0.62
+-40%c	XGBoost	0.59	0.62	0.57	0.18	0.56	0.62
+-40%c	Naïve Bayes	0.58	0.58	0.64	0.17	0.76	0.40
+-45%c	SVM	0.60	0.65	0.60	0.20	0.60	0.60
+-45%c	Random Forest	0.58	0.62	0.58	0.16	0.59	0.56
+-45%c	KNN	0.56	0.60	0.52	0.13	0.49	0.63
+-45%c	XGBoost	0.58	0.61	0.59	0.16	0.59	0.56
+-45%c	Naïve Bayes	0.55	0.56	0.61	0.11	0.71	0.40
+-50%c	SVM	0.59	0.64	0.59	0.19	0.59	0.59
+-50%c	Random Forest	0.59	0.62	0.59	0.19	0.59	0.60
+-50%c	KNN	0.59	0.59	0.56	0.17	0.54	0.63
+-50%c	XGBoost	0.57	0.59	0.57	0.15	0.58	0.56
+-50%c	Naïve Bayes	0.55	0.56	0.61	0.11	0.70	0.40

(Supplementary Table) Table S4: Classification results for all combinations of subsampling scenarios and algorithms for Etoposide. Bold font indicates the best-performing combination for each metric.

Scenario	Method	ACC	AUC	F1	MCC	Recall	Specificity
+-5%c	SVM	**0.80**	0.88	0.80	**0.61**	0.81	**0.79**
+-5%c	Random Forest	0.79	**0.89**	0.79	0.58	0.79	**0.79**
+-5%c	KNN	0.78	0.84	**0.81**	0.54	**0.88**	0.67
+-5%c	XGBoost	0.76	0.86	0.77	0.56	0.77	0.77
+-5%c	Naïve Bayes	0.71	0.71	0.73	0.40	0.79	0.63
+-10%c	SVM	0.72	0.84	0.71	0.44	0.70	0.74
+-10%c	Random Forest	0.72	0.82	0.71	0.44	0.72	0.71
+-10%c	KNN	0.68	0.77	0.67	0.37	0.69	0.68
+-10%c	XGBoost	0.74	0.82	0.73	0.47	0.75	0.74
+-10%c	Naïve Bayes	0.66	0.66	0.69	0.30	0.79	0.53
+-15%c	SVM	0.73	0.78	0.74	0.47	0.75	0.72
+-15%c	Random Forest	0.69	0.75	0.69	0.38	0.73	0.65
+-15%c	KNN	0.63	0.70	0.63	0.25	0.68	0.58
+-15%c	XGBoost	0.69	0.76	0.69	0.38	0.72	0.66
+-15%c	Naïve Bayes	0.65	0.65	0.70	0.30	0.82	0.47
+-20%c	SVM	0.69	0.77	0.69	0.40	0.71	0.68
+-20%c	Random Forest	0.67	0.73	0.67	0.35	0.72	0.62
+-20%c	KNN	0.61	0.69	0.62	0.24	0.65	0.57
+-20%c	XGBoost	0.69	0.75	0.70	0.39	0.73	0.65
+-20%c	Naïve Bayes	0.61	0.61	0.66	0.23	0.79	0.43
+-25%c	SVM	0.71	0.77	0.69	0.42	0.71	0.71
+-25%c	Random Forest	0.64	0.70	0.65	0.28	0.70	0.57
+-25%c	KNN	0.61	0.67	0.60	0.22	0.60	0.61
+-25%c	XGBoost	0.69	0.77	0.68	0.37	0.70	0.67
+-25%c	Naïve Bayes	0.59	0.60	0.65	0.19	0.78	0.39
+-30%c	SVM	0.68	0.74	0.67	0.36	0.68	0.67
+-30%c	Random Forest	0.62	0.67	0.64	0.23	0.70	0.53
+-30%c	KNN	0.61	0.65	0.62	0.23	0.64	0.58
+-30%c	XGBoost	0.66	0.72	0.66	0.32	0.67	0.65
+-30%c	Naïve Bayes	0.58	0.58	0.64	0.16	0.78	0.37
+-35%c	SVM	0.64	0.71	0.64	0.29	0.63	0.65
+-35%c	Random Forest	0.63	0.67	0.65	0.26	0.69	0.57
+-35%c	KNN	0.61	0.65	0.60	0.22	0.58	0.64
+-35%c	XGBoost	0.61	0.66	0.61	0.22	0.62	0.59
+-35%c	Naïve Bayes	0.58	0.58	0.65	0.17	0.78	0.38
+-40%c	SVM	0.63	0.67	0.63	0.25	0.65	0.61

+-40%c	Random Forest	0.59	0.65	0.60	0.18	0.62	0.55
+-40%c	KNN	0.57	0.62	0.56	0.14	0.55	0.59
+-40%c	XGBoost	0.65	0.68	0.65	0.29	0.68	0.62
+-40%c	Naïve Bayes	0.57	0.58	0.64	0.16	0.78	0.36
+-45%c	SVM	0.64	0.68	0.64	0.28	0.64	0.64
+-45%c	Random Forest	0.60	0.64	0.62	0.21	0.65	0.56
+-45%c	KNN	0.57	0.61	0.56	0.14	0.55	0.59
+-45%c	XGBoost	0.63	0.68	0.63	0.26	0.64	0.61
+-45%c	Naïve Bayes	0.57	0.58	0.65	0.16	0.79	0.36
+-50%c	SVM	0.62	0.66	0.62	0.24	0.63	0.61
+-50%c	Random Forest	0.59	0.63	0.60	0.19	0.64	0.55
+-50%c	KNN	0.57	0.59	0.55	0.13	0.54	0.59
+-50%c	XGBoost	0.64	0.68	0.65	0.28	0.66	0.61
+-50%c	Naïve Bayes	0.55	0.56	0.62	0.11	0.76	0.34

(Supplementary Table) Table S5: Classification results for all combinations of subsampling scenarios and algorithms for Gemcitabine. Bold font indicates the best-performing combination for each metric.

Scenario	Method	ACC	AUC	F1	MCC	Recall	Specificity
+-5%c	SVM	0.68	0.74	0.67	0.33	0.74	0.62
+-5%c	Random Forest	0.65	0.77	0.64	0.34	0.64	0.64
+-5%c	KNN	0.69	0.69	0.59	0.32	0.64	0.74
+-5%c	XGBoost	0.72	0.76	0.70	0.44	0.69	0.74
+-5%c	Naïve Bayes	0.67	0.73	0.67	0.40	0.74	0.60
+-10%c	SVM	0.74	0.80	0.73	0.47	**0.75**	0.72
+-10%c	Random Forest	0.65	0.76	0.65	0.31	0.67	0.64
+-10%c	KNN	0.59	0.67	0.56	0.20	0.54	0.65
+-10%c	XGBoost	0.64	0.75	0.63	0.32	0.65	0.62
+-10%c	Naïve Bayes	0.59	0.59	0.63	0.18	0.73	0.46
+-15%c	SVM	**0.75**	**0.82**	**0.75**	**0.53**	0.72	**0.78**
+-15%c	Random Forest	0.71	0.78	0.71	0.44	0.71	0.70
+-15%c	KNN	0.65	0.69	0.63	0.32	0.61	0.69
+-15%c	XGBoost	0.71	0.79	0.71	0.46	0.69	0.73
+-15%c	Naïve Bayes	0.59	0.60	0.66	0.20	0.80	0.39
+-20%c	SVM	0.70	0.77	0.69	0.40	0.69	0.71
+-20%c	Random Forest	0.66	0.73	0.66	0.33	0.66	0.66
+-20%c	KNN	0.56	0.63	0.53	0.12	0.50	0.62
+-20%c	XGBoost	0.66	0.72	0.65	0.33	0.64	0.69
+-20%c	Naïve Bayes	0.59	0.59	0.66	0.19	0.80	0.37
+-25%c	SVM	0.68	0.74	0.68	0.36	0.67	0.69
+-25%c	Random Forest	0.66	0.72	0.66	0.32	0.65	0.67
+-25%c	KNN	0.61	0.66	0.58	0.21	0.54	0.68
+-25%c	XGBoost	0.68	0.72	0.67	0.36	0.67	0.69
+-25%c	Naïve Bayes	0.57	0.58	0.64	0.16	0.77	0.37
+-30%c	SVM	0.70	0.74	0.69	0.40	0.69	0.71
+-30%c	Random Forest	0.65	0.72	0.64	0.30	0.64	0.67
+-30%c	KNN	0.63	0.69	0.60	0.26	0.57	0.69
+-30%c	XGBoost	0.66	0.73	0.64	0.32	0.64	0.68
+-30%c	Naïve Bayes	0.57	0.58	0.64	0.15	0.77	0.37
+-35%c	SVM	0.66	0.71	0.66	0.33	0.66	0.67
+-35%c	Random Forest	0.65	0.70	0.65	0.31	0.64	0.67
+-35%c	KNN	0.61	0.67	0.59	0.23	0.56	0.66
+-35%c	XGBoost	0.65	0.73	0.64	0.30	0.63	0.67
+-35%c	Naïve Bayes	0.56	0.57	0.64	0.14	0.78	0.34
+-40%c	SVM	0.63	0.70	0.63	0.26	0.62	0.64

Appendix

+-40%c	Random Forest	0.65	0.69	0.65	0.30	0.65	0.64
+-40%c	KNN	0.60	0.65	0.57	0.20	0.54	0.65
+-40%c	XGBoost	0.66	0.72	0.65	0.33	0.63	0.69
+-40%c	Naïve Bayes	0.56	0.57	0.64	0.14	0.77	0.36
+-45%c	SVM	0.63	0.67	0.63	0.25	0.64	0.61
+-45%c	Random Forest	0.62	0.66	0.63	0.25	0.65	0.60
+-45%c	KNN	0.58	0.63	0.57	0.17	0.55	0.61
+-45%c	XGBoost	0.64	0.67	0.63	0.27	0.62	0.65
+-45%c	Naïve Bayes	0.56	0.57	0.64	0.14	0.78	0.34
+-50%c	SVM	0.61	0.65	0.61	0.23	0.61	0.62
+-50%c	Random Forest	0.62	0.66	0.63	0.25	0.64	0.61
+-50%c	KNN	0.60	0.62	0.58	0.20	0.57	0.63
+-50%c	XGBoost	0.61	0.67	0.62	0.23	0.62	0.60
+-50%c	Naïve Bayes	0.56	0.56	0.63	0.12	0.78	0.34

(Supplementary Table) Table S6: Classification results for all combinations of subsampling scenarios and algorithms for Paclitaxel. Bold font indicates the best-performing combination for each metric.

Scenario	Method	ACC	AUC	F1	MCC	Recall	Specificity
+-5%c	SVM	0.58	NA	0.50	0.22	0.55	0.60
+-5%c	Random Forest	0.55	NA	0.39	0.12	0.45	0.65
+-5%c	KNN	0.55	NA	0.29	0.15	0.30	0.80
+-5%c	XGBoost	0.40	NA	0.30	-0.18	0.35	0.45
+-5%c	Naïve Bayes	0.50	NA	0.40	-0.02	0.50	0.50
+-10%c	SVM	0.56	0.65	0.43	0.27	0.44	0.69
+-10%c	Random Forest	0.63	**0.72**	0.59	0.33	0.56	0.69
+-10%c	KNN	0.61	0.68	0.48	0.30	0.38	**0.82**
+-10%c	XGBoost	0.55	0.58	0.52	0.18	0.49	0.62
+-10%c	Naïve Bayes	0.56	0.57	0.57	0.17	0.62	0.51
+-15%c	SVM	0.63	0.66	0.64	0.25	**0.69**	0.56
+-15%c	Random Forest	**0.70**	0.70	**0.68**	**0.41**	0.64	0.76
+-15%c	KNN	0.62	0.61	0.56	0.24	0.51	0.75
+-15%c	XGBoost	0.66	0.73	0.63	0.33	0.61	0.71
+-15%c	Naïve Bayes	0.59	0.59	0.59	0.18	0.61	0.58
+-20%c	SVM	0.59	0.66	0.55	0.19	0.54	0.63
+-20%c	Random Forest	0.55	0.64	0.52	0.12	0.51	0.59
+-20%c	KNN	0.65	0.65	0.60	0.31	0.57	0.73
+-20%c	XGBoost	0.63	0.65	0.58	0.24	0.56	0.70
+-20%c	Naïve Bayes	0.55	0.54	0.53	0.07	0.54	0.54
+-25%c	SVM	0.64	0.70	0.63	0.31	0.64	0.64
+-25%c	Random Forest	0.64	0.71	0.60	0.29	0.56	0.71
+-25%c	KNN	0.61	0.62	0.57	0.23	0.53	0.68
+-25%c	XGBoost	0.62	0.67	0.57	0.25	0.54	0.69
+-25%c	Naïve Bayes	0.55	0.57	0.55	0.09	0.58	0.52
+-30%c	SVM	0.63	0.69	0.62	0.28	0.62	0.64
+-30%c	Random Forest	0.62	0.70	0.59	0.24	0.55	0.68
+-30%c	KNN	0.62	0.68	0.59	0.23	0.57	0.68
+-30%c	XGBoost	0.62	0.67	0.59	0.26	0.56	0.69
+-30%c	Naïve Bayes	0.60	0.61	0.61	0.19	0.65	0.55
+-35%c	SVM	0.62	0.72	0.60	0.28	0.56	0.69
+-35%c	Random Forest	0.61	0.69	0.58	0.23	0.54	0.67
+-35%c	KNN	0.56	0.63	0.53	0.13	0.51	0.61
+-35%c	XGBoost	0.64	0.69	0.62	0.28	0.60	0.68
+-35%c	Naïve Bayes	0.58	0.60	0.60	0.17	0.66	0.50
+-40%c	SVM	0.60	0.66	0.60	0.20	0.58	0.63

Appendix

+-40%c	Random Forest	0.60	0.67	0.57	0.21	0.54	0.66
+-40%c	KNN	0.56	0.63	0.53	0.12	0.51	0.61
+-40%c	XGBoost	0.62	0.68	0.60	0.24	0.57	0.66
+-40%c	Naïve Bayes	0.55	0.57	0.59	0.11	0.66	0.45
+-45%c	SVM	0.63	0.67	0.61	0.25	0.59	0.67
+-45%c	Random Forest	0.61	0.68	0.59	0.22	0.56	0.65
+-45%c	KNN	0.63	0.66	0.60	0.25	0.56	0.69
+-45%c	XGBoost	0.62	0.66	0.59	0.24	0.56	0.67
+-45%c	Naïve Bayes	0.58	0.59	0.60	0.18	0.64	0.52
+-50%c	SVM	0.62	0.68	0.60	0.24	0.56	0.68
+-50%c	Random Forest	0.58	0.66	0.57	0.16	0.55	0.60
+-50%c	KNN	0.60	0.65	0.56	0.20	0.53	0.67
+-50%c	XGBoost	0.62	0.67	0.62	0.25	0.61	0.63
+-50%c	Naïve Bayes	0.55	0.55	0.57	0.10	0.60	0.50

(Supplementary Table) Table S7: Classification results for all combinations of subsampling scenarios and algorithms for Temozolomide. Bold font indicates the best-performing combination for each metric.

Scenario	Method	ACC	AUC	F1	MCC	Recall	Specificity
+-5%c	SVM	**0.88**	0.93	0.84	0.73	0.89	0.86
+-5%c	Random Forest	0.80	0.89	0.77	0.59	0.84	0.77
+-5%c	KNN	0.80	0.92	0.73	0.59	0.70	**0.89**
+-5%c	XGBoost	0.87	0.89	0.83	0.72	0.91	0.84
+-5%c	Naïve Bayes	0.70	0.69	0.69	0.39	0.77	0.64
+-10%c	SVM	0.87	**0.95**	**0.86**	**0.74**	**0.92**	0.81
+-10%c	Random Forest	0.82	0.90	0.82	0.64	0.88	0.76
+-10%c	KNN	0.76	0.84	0.75	0.53	0.76	0.76
+-10%c	XGBoost	0.85	0.93	0.85	0.73	0.91	0.80
+-10%c	Naïve Bayes	0.76	0.76	0.77	0.52	0.91	0.61
+-15%c	SVM	0.86	0.91	**0.86**	0.72	0.87	0.84
+-15%c	Random Forest	0.80	0.88	0.80	0.60	0.84	0.75
+-15%c	KNN	0.78	0.85	0.79	0.57	0.83	0.73
+-15%c	XGBoost	0.83	0.90	0.84	0.65	0.87	0.78
+-15%c	Naïve Bayes	0.71	0.73	0.75	0.45	0.87	0.56
+-20%c	SVM	0.81	0.87	0.81	0.63	0.84	0.79
+-20%c	Random Forest	0.76	0.84	0.78	0.53	0.84	0.69
+-20%c	KNN	0.73	0.81	0.73	0.46	0.77	0.68
+-20%c	XGBoost	0.76	0.84	0.76	0.51	0.78	0.73
+-20%c	Naïve Bayes	0.70	0.71	0.75	0.42	0.88	0.52
+-25%c	SVM	0.76	0.86	0.77	0.53	0.78	0.74
+-25%c	Random Forest	0.73	0.83	0.75	0.47	0.82	0.64
+-25%c	KNN	0.71	0.81	0.72	0.43	0.77	0.65
+-25%c	XGBoost	0.74	0.83	0.75	0.50	0.80	0.69
+-25%c	Naïve Bayes	0.68	0.69	0.73	0.38	0.86	0.49
+-30%c	SVM	0.76	0.83	0.76	0.52	0.79	0.73
+-30%c	Random Forest	0.71	0.81	0.73	0.43	0.77	0.65
+-30%c	KNN	0.70	0.78	0.72	0.41	0.76	0.65
+-30%c	XGBoost	0.73	0.81	0.74	0.47	0.77	0.70
+-30%c	Naïve Bayes	0.68	0.69	0.73	0.39	0.87	0.49
+-35%c	SVM	0.73	0.80	0.73	0.46	0.74	0.72
+-35%c	Random Forest	0.71	0.79	0.73	0.43	0.78	0.64
+-35%c	KNN	0.69	0.77	0.70	0.38	0.74	0.63
+-35%c	XGBoost	0.71	0.78	0.72	0.42	0.76	0.66
+-35%c	Naïve Bayes	0.67	0.68	0.72	0.36	0.86	0.47
+-40%c	SVM	0.70	0.77	0.70	0.39	0.72	0.67

Appendix

+-40%c	Random Forest	0.70	0.77	0.71	0.41	0.76	0.65
+-40%c	KNN	0.67	0.74	0.69	0.35	0.73	0.62
+-40%c	XGBoost	0.71	0.77	0.72	0.42	0.77	0.65
+-40%c	Naïve Bayes	0.65	0.66	0.71	0.33	0.85	0.45
+-45%c	SVM	0.68	0.76	0.69	0.37	0.71	0.65
+-45%c	Random Forest	0.66	0.74	0.69	0.34	0.74	0.59
+-45%c	KNN	0.64	0.71	0.66	0.28	0.71	0.57
+-45%c	XGBoost	0.69	0.75	0.69	0.38	0.72	0.65
+-45%c	Naïve Bayes	0.64	0.66	0.70	0.32	0.85	0.43
+-50%c	SVM	0.68	0.74	0.69	0.36	0.71	0.66
+-50%c	Random Forest	0.66	0.73	0.68	0.32	0.72	0.59
+-50%c	KNN	0.63	0.68	0.65	0.26	0.69	0.56
+-50%c	XGBoost	0.68	0.74	0.69	0.36	0.70	0.66
+-50%c	Naïve Bayes	0.62	0.63	0.69	0.26	0.83	0.41

(Supplementary Table) Table S8: Regression results for all combinations of subsampling scenarios and algorithms for Cisplatin. Bold font indicates the best-performing combination for each metric.

Scenario	Method	MAE	RMSE	R^2	Spearman
+-5%r	SVM	1.82	2.14	0.11	0.35
+-5%r	Random Forest	1.81	2.11	0.13	0.37
+-5%r	KNN	1.74	2.16	0.10	0.44
+-5%r	XGBoost	1.60	2.16	0.07	0.43
+-10%r	SVM	1.26	1.59	0.37	0.49
+-10%r	Random Forest	1.41	1.67	0.30	0.49
+-10%r	KNN	1.31	1.67	0.31	0.46
+-10%r	XGBoost	1.36	1.78	0.20	0.39
+-15%r	SVM	1.17	1.44	**0.39**	0.63
+-15%r	Random Forest	1.29	1.52	0.32	0.56
+-15%r	KNN	1.30	1.62	0.23	0.45
+-15%r	XGBoost	1.30	1.65	0.20	0.45
+-20%r	SVM	1.11	1.40	0.31	**0.57**
+-20%r	Random Forest	1.21	1.43	0.28	0.55
+-20%r	KNN	1.23	1.51	0.18	0.46
+-20%r	XGBoost	1.21	1.48	0.23	0.47
+-25%r	SVM	1.08	1.33	0.24	0.52
+-25%r	Random Forest	1.14	1.34	0.23	0.51
+-25%r	KNN	1.17	1.43	0.13	0.41
+-25%r	XGBoost	1.14	1.40	0.16	0.45
+-30%r	SVM	0.99	1.23	0.26	0.52
+-30%r	Random Forest	1.06	1.26	0.24	0.51
+-30%r	KNN	1.10	1.34	0.13	0.40
+-30%r	XGBoost	1.04	1.28	0.21	0.49
+-35%r	SVM	0.96	1.19	0.23	0.51
+-35%r	Random Forest	1.00	1.20	0.21	0.48
+-35%r	KNN	1.03	1.28	0.10	0.38
+-35%r	XGBoost	1.01	1.24	0.16	0.43
+-40%r	SVM	0.92	1.15	0.17	0.46
+-40%r	Random Forest	0.94	1.14	0.19	0.44
+-40%r	KNN	0.97	1.20	0.10	0.35
+-40%r	XGBoost	0.94	1.16	0.16	0.42
+-45%r	SVM	0.86	1.09	0.17	0.45
+-45%r	Random Forest	0.87	1.08	0.19	0.44
+-45%r	KNN	0.91	1.14	0.09	0.33
+-45%r	XGBoost	0.89	1.11	0.15	0.38

+-50%r	SVM	**0.82**	**1.04**	0.16	0.41
+-50%r	Random Forest	**0.82**	**1.04**	0.17	0.40
+-50%r	KNN	0.86	1.10	0.07	0.31
+-50%r	XGBoost	**0.82**	1.05	0.15	0.38

(Supplementary Table) Table S9: Regression results for all combinations of subsampling scenarios and algorithms for Docetaxel. Bold font indicates the best-performing combination for each metric.

Scenario	Method	MAE	RMSE	R^2	Spearman
+-5%r	SVM	1.60	1.99	0.51	0.67
+-5%r	Random Forest	1.86	2.14	0.44	0.65
+-5%r	KNN	2.05	2.74	0.10	0.50
+-5%r	XGBoost	1.55	2.23	0.39	0.66
+-10%r	SVM	1.23	1.62	**0.59**	0.62
+-10%r	Random Forest	1.35	1.69	0.55	0.63
+-10%r	KNN	1.34	1.94	0.43	0.55
+-10%r	XGBoost	1.14	1.62	0.58	0.65
+-15%r	SVM	1.20	1.54	0.56	0.67
+-15%r	Random Forest	1.31	1.64	0.50	**0.69**
+-15%r	KNN	1.26	1.76	0.43	0.61
+-15%r	XGBoost	1.26	1.70	0.46	0.64
+-20%r	SVM	1.14	1.47	0.50	0.66
+-20%r	Random Forest	1.23	1.54	0.47	0.67
+-20%r	KNN	1.25	1.68	0.35	0.60
+-20%r	XGBoost	1.26	1.63	0.39	0.62
+-25%r	SVM	1.15	1.46	0.44	0.67
+-25%r	Random Forest	1.19	1.48	0.43	0.66
+-25%r	KNN	1.19	1.58	0.34	0.59
+-25%r	XGBoost	1.20	1.55	0.38	0.60
+-30%r	SVM	1.12	1.43	0.39	0.64
+-30%r	Random Forest	1.17	1.44	0.38	0.64
+-30%r	KNN	1.17	1.52	0.30	0.58
+-30%r	XGBoost	1.18	1.49	0.34	0.59
+-35%r	SVM	1.09	1.38	0.35	0.63
+-35%r	Random Forest	1.13	1.39	0.34	0.62
+-35%r	KNN	1.16	1.50	0.23	0.56
+-35%r	XGBoost	1.12	1.39	0.34	0.60
+-40%r	SVM	1.04	1.30	0.35	0.61
+-40%r	Random Forest	1.09	1.34	0.32	0.58
+-40%r	KNN	1.13	1.45	0.19	0.51

+-40%r	XGBoost	1.09	1.36	0.29	0.57
+-45%r	SVM	1.00	1.26	0.32	0.59
+-45%r	Random Forest	1.03	1.28	0.30	0.58
+-45%r	KNN	1.07	1.39	0.17	0.49
+-45%r	XGBoost	1.04	1.30	0.27	0.53
+-50%r	SVM	**0.95**	**1.22**	0.30	0.56
+-50%r	Random Forest	0.98	1.24	0.28	0.54
+-50%r	KNN	1.03	1.34	0.15	0.47
+-50%r	XGBoost	0.99	1.25	0.26	0.51

(Supplementary Table) Table S10: Regression results for all combinations of subsampling scenarios and algorithms for Doxorubicin. Bold font indicates the best-performing combination for each metric.

Scenario	Method	MAE	RMSE	R^2	Spearman
+-5%r	SVM	2.83	3.17	0.06	0.39
+-5%r	Random Forest	3.02	3.28	-0.01	0.33
+-5%r	KNN	2.71	3.21	0.03	0.45
+-5%r	XGBoost	2.82	3.51	-0.17	0.23
+-10%r	SVM	2.34	2.63	0.06	**0.48**
+-10%r	Random Forest	2.53	2.75	0.00	0.37
+-10%r	KNN	2.33	2.74	-0.02	0.41
+-10%r	XGBoost	2.40	2.87	-0.11	0.40
+-15%r	SVM	2.18	2.45	**0.14**	0.40
+-15%r	Random Forest	2.23	2.46	0.13	0.40
+-15%r	KNN	2.23	2.60	0.02	0.32
+-15%r	XGBoost	2.25	2.61	0.01	0.29
+-20%r	SVM	1.96	2.25	0.08	0.40
+-20%r	Random Forest	2.04	2.27	0.07	0.40
+-20%r	KNN	1.96	2.33	0.01	0.37
+-20%r	XGBoost	2.00	2.34	0.02	0.34
+-25%r	SVM	1.79	2.10	0.12	0.36
+-25%r	Random Forest	1.84	2.08	0.13	0.36
+-25%r	KNN	1.82	2.21	0.03	0.29
+-25%r	XGBoost	1.82	2.14	0.08	0.34
+-30%r	SVM	1.66	1.94	0.13	0.37
+-30%r	Random Forest	1.71	1.96	0.11	0.35
+-30%r	KNN	1.71	2.07	0.01	0.27
+-30%r	XGBoost	1.72	2.04	0.05	0.29
+-35%r	SVM	1.54	1.83	0.11	0.34
+-35%r	Random Forest	1.57	1.84	0.11	0.33
+-35%r	KNN	1.59	1.94	0.01	0.26
+-35%r	XGBoost	1.57	1.89	0.06	0.29
+-40%r	SVM	1.44	1.75	0.09	0.33
+-40%r	Random Forest	1.47	1.76	0.08	0.30
+-40%r	KNN	1.49	1.83	-0.01	0.26
+-40%r	XGBoost	1.48	1.80	0.04	0.24
+-45%r	SVM	1.33	1.64	0.09	0.35
+-45%r	Random Forest	1.35	1.66	0.08	0.29
+-45%r	KNN	1.41	1.75	-0.03	0.26
+-45%r	XGBoost	1.37	1.70	0.02	0.24

+-50%r	SVM	**1.24**	1.59	0.06	0.30
+-50%r	Random Forest	1.25	**1.58**	0.07	0.27
+-50%r	KNN	1.33	1.69	-0.05	0.21
+-50%r	XGBoost	1.28	1.61	0.04	0.24

(Supplementary Table) Table S11: Regression results for all combinations of subsampling scenarios and algorithms for Etoposide. Bold font indicates the best-performing combination for each metric.

Scenario	Method	MAE	RMSE	R^2	Spearman
+-5%r	SVM	2.32	2.93	**0.36**	0.49
+-5%r	Random Forest	2.36	2.93	**0.36**	0.49
+-5%r	KNN	2.29	3.03	0.32	0.43
+-5%r	XGBoost	2.33	3.28	0.23	0.37
+-10%r	SVM	2.27	2.77	0.21	0.49
+-10%r	Random Forest	2.44	2.88	0.15	0.44
+-10%r	KNN	2.40	2.99	0.05	0.40
+-10%r	XGBoost	2.50	3.24	-0.09	0.29
+-15%r	SVM	2.21	2.62	0.27	**0.53**
+-15%r	Random Forest	2.37	2.70	0.22	0.48
+-15%r	KNN	2.36	2.86	0.13	0.43
+-15%r	XGBoost	2.37	2.89	0.11	0.39
+-20%r	SVM	2.13	2.48	0.20	0.52
+-20%r	Random Forest	2.28	2.57	0.14	0.45
+-20%r	KNN	2.28	2.71	0.03	0.41
+-20%r	XGBoost	2.23	2.66	0.08	0.40
+-25%r	SVM	1.98	2.31	0.22	0.48
+-25%r	Random Forest	2.13	2.39	0.15	0.43
+-25%r	KNN	2.11	2.51	0.07	0.36
+-25%r	XGBoost	2.04	2.39	0.15	0.42
+-30%r	SVM	1.90	2.20	0.17	0.44
+-30%r	Random Forest	2.00	2.26	0.13	0.39
+-30%r	KNN	2.02	2.41	0.00	0.28
+-30%r	XGBoost	1.96	2.29	0.10	0.36
+-35%r	SVM	1.78	2.10	0.16	0.42
+-35%r	Random Forest	1.86	2.14	0.13	0.36
+-35%r	KNN	1.87	2.25	0.04	0.29
+-35%r	XGBoost	1.86	2.20	0.08	0.30
+-40%r	SVM	1.65	1.96	0.16	0.41
+-40%r	Random Forest	1.72	2.02	0.11	0.37
+-40%r	KNN	1.76	2.14	0.01	0.28
+-40%r	XGBoost	1.69	2.01	0.12	0.37
+-45%r	SVM	1.52	1.86	0.17	0.43
+-45%r	Random Forest	1.59	1.91	0.13	0.38
+-45%r	KNN	1.64	2.03	0.01	0.27
+-45%r	XGBoost	1.60	1.94	0.10	0.32

+-50%r	SVM	**1.46**	**1.80**	0.14	0.38
+-50%r	Random Forest	1.49	1.84	0.09	0.32
+-50%r	KNN	1.56	1.94	-0.01	0.24
+-50%r	XGBoost	1.53	1.89	0.04	0.26

(Supplementary Table) Table S12: Regression results for all combinations of subsampling scenarios and algorithms for Gemcitabine. Bold font indicates the best-performing combination for each metric.

Scenario	Method	MAE	RMSE	R^2	Spearman
+-5%r	SVM	3.52	4.15	**0.20**	0.45
+-5%r	Random Forest	3.98	4.30	0.15	0.43
+-5%r	KNN	3.71	4.47	0.07	0.39
+-5%r	XGBoost	3.79	4.79	-0.04	0.41
+-10%r	SVM	3.35	3.94	0.09	0.45
+-10%r	Random Forest	3.87	4.14	0.01	0.39
+-10%r	KNN	3.93	4.51	-0.19	0.26
+-10%r	XGBoost	3.72	4.37	-0.12	0.31
+-15%r	SVM	3.02	3.57	0.24	**0.48**
+-15%r	Random Forest	3.39	3.71	0.18	0.42
+-15%r	KNN	3.45	4.07	0.01	0.33
+-15%r	XGBoost	3.39	4.00	0.04	0.31
+-20%r	SVM	3.02	3.50	0.15	0.44
+-20%r	Random Forest	3.26	3.55	0.12	0.40
+-20%r	KNN	3.28	3.82	-0.03	0.25
+-20%r	XGBoost	3.18	3.71	0.03	0.30
+-25%r	SVM	2.81	3.25	0.17	0.43
+-25%r	Random Forest	3.01	3.33	0.13	0.37
+-25%r	KNN	2.92	3.48	0.05	0.31
+-25%r	XGBoost	2.91	3.40	0.09	0.33
+-30%r	SVM	2.63	3.07	0.16	0.42
+-30%r	Random Forest	2.73	3.09	0.15	0.38
+-30%r	KNN	2.72	3.28	0.05	0.31
+-30%r	XGBoost	2.69	3.17	0.11	0.33
+-35%r	SVM	2.50	2.92	0.15	0.41
+-35%r	Random Forest	2.58	2.94	0.14	0.38
+-35%r	KNN	2.57	3.09	0.05	0.31
+-35%r	XGBoost	2.57	3.02	0.08	0.31
+-40%r	SVM	2.36	2.77	0.13	0.40
+-40%r	Random Forest	2.18	2.66	0.10	0.33
+-40%r	KNN	2.41	2.94	0.03	0.29
+-40%r	XGBoost	2.36	2.82	0.11	0.36

+-45%r	SVM	2.23	2.66	0.11	0.36
+-45%r	Random Forest	2.17	2.61	0.10	0.34
+-45%r	KNN	2.28	2.82	0.00	0.27
+-45%r	XGBoost	2.23	2.67	0.10	0.35
+-50%r	SVM	2.10	2.56	0.08	0.33
+-50%r	Random Forest	2.11	**2.54**	0.10	0.34
+-50%r	KNN	2.16	2.70	-0.02	0.25
+-50%r	XGBoost	**2.08**	**2.54**	0.09	0.33

(Supplementary Table) Table S13: Regression results for all combinations of subsampling scenarios and algorithms for Paclitaxel. Bold font indicates the best-performing combination for each metric.

Scenario	Method	MAE	RMSE	R^2	Spearman
+-5%r	SVM	3.27	3.56	-151.71	0.08
+-5%r	Random Forest	3.39	3.58	-134.10	0.04
+-5%r	KNN	3.18	3.74	-111.03	0.24
+-5%r	XGBoost	3.96	4.74	-166.85	-0.14
+-10%r	SVM	2.85	3.10	-0.01	0.38
+-10%r	Random Forest	2.95	3.11	-0.02	0.30
+-10%r	KNN	2.84	3.36	-0.18	**0.39**
+-10%r	XGBoost	2.66	3.35	-0.18	0.19
+-15%r	SVM	2.57	2.80	0.07	0.27
+-15%r	Random Forest	2.66	2.82	0.06	0.28
+-15%r	KNN	2.60	3.07	-0.12	0.20
+-15%r	XGBoost	2.71	3.26	-0.26	0.13
+-20%r	SVM	2.40	2.67	0.01	0.24
+-20%r	Random Forest	2.52	2.71	-0.01	0.15
+-20%r	KNN	2.36	2.79	-0.08	0.17
+-20%r	XGBoost	2.41	2.79	-0.08	0.19
+-25%r	SVM	2.18	2.48	0.04	0.26
+-25%r	Random Forest	2.25	2.47	0.04	0.26
+-25%r	KNN	2.18	2.60	-0.06	0.22
+-25%r	XGBoost	2.12	2.48	0.03	0.29
+-30%r	SVM	2.02	2.29	**0.09**	0.32
+-30%r	Random Forest	2.04	2.28	**0.09**	0.32
+-30%r	KNN	1.96	2.36	0.03	0.29
+-30%r	XGBoost	2.05	2.40	0.00	0.24
+-35%r	SVM	1.85	2.18	0.05	0.34
+-35%r	Random Forest	1.89	2.17	0.06	0.28
+-35%r	KNN	1.90	2.29	-0.06	0.23
+-35%r	XGBoost	1.97	2.35	-0.11	0.16
+-40%r	SVM	1.73	2.05	0.05	0.30
+-40%r	Random Forest	1.75	2.04	0.05	0.31
+-40%r	KNN	1.76	2.15	-0.04	0.22
+-40%r	XGBoost	1.82	2.18	-0.09	0.18
+-45%r	SVM	1.63	1.98	0.01	0.27
+-45%r	Random Forest	1.62	1.95	0.04	0.29
+-45%r	KNN	1.65	2.03	-0.04	0.26
+-45%r	XGBoost	1.65	2.03	-0.02	0.24

+-50%r	SVM	**1.49**	1.87	0.04	0.29
+-50%r	Random Forest	1.50	**1.84**	0.07	0.31
+-50%r	KNN	1.54	1.95	-0.04	0.23
+-50%r	XGBoost	1.56	1.91	-0.02	0.23

(Supplementary Table) Table S14: Regression results for all combinations of subsampling scenarios and algorithms for Temozolomide. Bold font indicates the best-performing combination for each metric.

Scenario	Method	MAE	RMSE	R^2	Spearman
+-5%r	SVM	0.88	1.16	0.50	0.65
+-5%r	Random Forest	1.06	1.28	0.40	0.61
+-5%r	KNN	1.03	1.35	0.33	0.65
+-5%r	XGBoost	1.04	1.49	0.21	0.62
+-10%r	SVM	0.72	0.93	**0.57**	**0.69**
+-10%r	Random Forest	0.83	1.02	0.47	0.62
+-10%r	KNN	0.87	1.16	0.32	0.58
+-10%r	XGBoost	0.82	1.11	0.37	0.62
+-15%r	SVM	0.66	0.86	0.56	**0.69**
+-15%r	Random Forest	0.76	0.95	0.46	0.65
+-15%r	KNN	0.75	1.03	0.36	0.60
+-15%r	XGBoost	0.81	1.06	0.33	0.56
+-20%r	SVM	0.64	0.82	0.47	0.68
+-20%r	Random Forest	0.73	0.90	0.37	0.62
+-20%r	KNN	0.73	0.97	0.27	0.52
+-20%r	XGBoost	0.72	0.93	0.32	0.60
+-25%r	SVM	0.63	0.80	0.41	0.64
+-25%r	Random Forest	0.68	0.84	0.36	0.61
+-25%r	KNN	0.69	0.92	0.24	0.53
+-25%r	XGBoost	0.70	0.88	0.29	0.54
+-30%r	SVM	0.63	0.80	0.34	0.59
+-30%r	Random Forest	0.63	0.79	0.37	0.60
+-30%r	KNN	0.67	0.88	0.22	0.50
+-30%r	XGBoost	0.67	0.84	0.29	0.54
+-35%r	SVM	0.60	0.76	0.32	0.55
+-35%r	Random Forest	0.61	0.76	0.31	0.56
+-35%r	KNN	0.65	0.84	0.15	0.47
+-35%r	XGBoost	0.63	0.81	0.22	0.51
+-40%r	SVM	0.57	0.72	0.31	0.55
+-40%r	Random Forest	0.57	0.73	0.29	0.53
+-40%r	KNN	0.63	0.81	0.08	0.45
+-40%r	XGBoost	0.60	0.78	0.17	0.49

+-45%r	SVM	0.54	0.69	0.29	0.53
+-45%r	Random Forest	0.54	0.70	0.27	0.51
+-45%r	KNN	0.59	0.78	0.07	0.42
+-45%r	XGBoost	0.56	0.73	0.19	0.49
+-50%r	SVM	0.52	0.68	0.22	0.50
+-50%r	Random Forest	**0.51**	**0.67**	0.26	0.49
+-50%r	KNN	0.55	0.73	0.10	0.41
+-50%r	XGBoost	0.53	0.69	0.21	0.48

Appendix

(Supplementary Table) Table S15: Informative genes for predicting cell-line responses for Cisplatin. We used the feature selection to identify informative genes for Cisplatin drug-response prediction. Genomic coordinates are based on build 37 of the human genome. We used information gain to rank the genes; a higher score indicates a more informative gene.

Classification			Regression		
Gene	*Coordinates*	*Score*	*Gene*	*Coordinates*	*Score*
CGN	chr1:151483573-151483902	0.251	C17orf81, CLDN7	chr17:7164285-7166245	0.069
CLDN4, WBSCR27	chr7:73245434-73246045	0.207	CLDN4, WBSCR27	chr7:73245434-73246045	0.066
JMJD6, MXRA7	chr17:74706465-74707067	0.199	CLDN3	chr7:73183379-73185115	0.059
MYO5C	chr15:52587353-52588172	0.196	ESRP1	chr8:95652455-95652873	0.056
C17orf81, CLDN7	chr17:7164285-7166245	0.195	CDH1	chr16:68771034-68772344	0.055
FUT2	chr19:49206443-49206818	0.188	TUBGCP2, ZNF511	chr10:135123238-135123448	0.051
ID3	chr1:23885682-23886212	0.187	IFT172, KRTCAP3, NRBP1	chr2:27664939-27665151	0.051
EFR3A	chr8:132916322-132917060	0.186	LAD1	chr1:201368560-201369032	0.049
LOC100129354, NBEAL2	chr3:47050486-47051609	0.185	TUBGCP2, ZNF511	chr10:135122851-135123109	0.047
AKR1B1	chr7:134143115-134144063	0.183	SPINT1	chr15:41135719-41137210	0.046
TUBGCP2, ZNF511	chr10:135122851-135123109	0.177	HRC	chr19:49655102-49655395	0.045
BRD3	chr9:136919143-136919376	0.175	SYK	chr9:93563775-93564546	0.044
KIRREL2, NPHS1	chr19:36347044-36348101	0.174	C1orf172	chr1:27286065-27287101	0.043
FKBP2, VEGFB	chr11:64008283-64009487	0.173	CGN	chr1:151483573-151483902	0.043
BASP1	chr5:17275369-17275638	0.171	ESRP1	chr8:95653898-95654733	0.043
CMTM3	chr16:66638254-66639561	0.171	LLGL2, TSEN54	chr17:73520956-73522540	0.043
CCDC19	chr1:159869901-159870143	0.170	RAB4A	chr1:229406646-229407129	0.042
ITGA5	chr12:54811981-54812202	0.170	MYO5C	chr15:52587353-52588172	0.042
VIM	chr10:17270430-17272617	0.169	MAP7	chr6:136870826-136872145	0.041
EPS8L2, TMEM80	chr11:705794-706534	0.169	CDC42BPG	chr11:64611714-64612634	0.041

Appendix

(Supplementary Table) Table S16: Informative genes for predicting cell-line responses for Docetaxel. We used the feature selection to identify informative genes for Docetaxel drug-response prediction. Genomic coordinates are based on build 37 of the human genome. We used information gain to rank the genes; a higher score indicates a more informative gene.

Classification			Regression		
Gene	*Coordinates*	*Score*	*Gene*	*Coordinates*	*Score*
ELK3	chr12:96588665-96589145	0.357	NFATC2	chr20:50158904-50159509	0.086
DAPK3	chr19:3970536-3970746	0.327	VGLL4	chr3:11610137-11610370	0.083
SNAI2	chr8:49835987-49836231	0.302	CSNK1E	chr22:38712684-38713333	0.073
EXT1	chr8:119123974-119124432	0.296	COL7A1, UQCRC1	chr3:48631882-48632901	0.069
VGLL4	chr3:11610137-11610370	0.288	FLRT2	chr14:85996494-85996958	0.068
MMP14, MRPL52	chr14:23305893-23307013	0.268	C8orf58, PDLIM2	chr8:22456091-22456508	0.068
NCOR2	chr12:125003217-125003482	0.264	DAPK3	chr19:3970536-3970746	0.067
CMAH	chr6:25139920-25140246	0.261	PLEKHG5	chr1:6545143-6545559	0.063
PRNP	chr20:4666827-4667874	0.253	EMP3	chr19:48833394-48833720	0.063
PLEKHG5	chr1:6550083-6551115	0.250	RAB34	chr17:27044168-27045049	0.062
DUSP5	chr10:112257163-112258684	0.248	C22orf9, MIR1249	chr22:45598721-45599080	0.060
CBR3	chr21:37507198-37508259	0.240	ELK3	chr12:96588665-96589145	0.060
TNK2	chr3:195622187-195623033	0.240	PIK3CG	chr7:106508057-106508733	0.059
GADD45A	chr1:68150913-68152270	0.236	PTRF	chr17:40573740-40575526	0.059
FLRT2	chr14:85996494-85996958	0.232	EIF3G	chr19:10230162-10230682	0.059
ZC3H7B	chr22:41697388-41698601	0.232	ERBB2	chr17:37856448-37856891	0.059
EIF3G	chr19:10230162-10230682	0.231	HCFC1R1, THOC6, TNFRSF12A	chr16:3073686-3074443	0.058
GPR176	chr15:40211961-40213444	0.228	SOLH	chr16:587567-588172	0.056
COL7A1, UQCRC1	chr3:48631882-48632901	0.228	INPP5D	chr2:233925091-233925318	0.056
PTK2	chr8:142010440-142011907	0.220	COG5, DUS4L	chr7:107204114-107204797	0.056

Appendix

(Supplementary Table) Table S17: Informative genes for predicting cell-line responses for Doxorubicin. We used the feature selection to identify informative genes for Doxorubicin drug-response prediction. Genomic coordinates are based on build 37 of the human genome. We used information gain to rank the genes; a higher score indicates a more informative gene.

Classification			**Regression**		
Gene	*Coordinates*	*Score*	*Gene*	*Coordinates*	*Score*
SKAP1	chr17:46507344-46507778	0.195	TMEM177	chr2:120436530-120437010	0.033
SLC27A2	chr15:50474322-50475186	0.194	NEK10	chr3:27410612-27411066	0.031
PYGM, RASGRP2	chr11:64509433-64513826	0.188	ZFP3	chr17:4981357-4981979	0.028
CGN, MIR554, TUFT1	chr1:151512661-151513199	0.184	WDYHV1	chr8:124428605-124429425	0.027
LACTB2, XKR9	chr8:71581050-71581650	0.181	PPM1H	chr12:63328143-63329135	0.027
MXRA8	chr1:1289707-1291126	0.173	NCRNA00029	chr20:61665780-61666555	0.027
CAMK2N1	chr1:20810462-20813511	0.173	MATN2	chr8:98881311-98881843	0.026
OSTC	chr4:109571693-109572039	0.173	RIMKLA	chr1:42845978-42846988	0.026
PTK2	chr8:142010440-142011907	0.169	INHBB	chr2:121101800-121104534	0.026
CGN	chr1:151483573-151483902	0.167	GDA	chr9:74764241-74764903	0.026
SCIN	chr7:12610165-12610834	0.166	CMAS	chr12:22199062-22199589	0.025
C2orf43	chr2:21022564-21022934	0.162	ATP1B2	chr17:7554139-7555338	0.025
TMEM45B	chr11:129685737-129686211	0.155	C3orf57	chr3:161089626-161090649	0.025
TMEM177	chr2:120436530-120437010	0.153	C8orf84	chr8:74005021-74005856	0.024
RG9MTD3	chr9:37753655-37753949	0.152	STYXL1, TMEM120A	chr7:75623357-75624164	0.024
CLDN4, WBSCR27	chr7:73245434-73246045	0.151	ICA1	chr7:8301031-8302252	0.024
GAL	chr11:68451359-68452846	0.150	YBX2	chr17:7197431-7198417	0.024
CGB7	chr19:49559222-49560497	0.150	TUBGCP2, ZNF511	chr10:135123238-135123448	0.024
COMMD2	chr3:149469909-149470388	0.150	CP110, GDE1	chr16:19535074-19535635	0.024
PODXL	chr7:131242693-131243006	0.150	TMEM219	chr16:29973023-29973570	0.024

Appendix

(Supplementary Table) Table S18: Informative genes for predicting cell-line responses for Etoposide. We used the feature selection to identify informative genes for Etoposide drug-response prediction. Genomic coordinates are based on build 37 of the human genome. We used information gain to rank the genes; a higher score indicates a more informative gene.

Classification			Regression		
Gene	*Coordinates*	*Score*	*Gene*	*Coordinates*	*Score*
GNMT, PEX6	chr6:42928218-42928810	0.367	C8orf84	chr8:74005021-74005856	0.041
TRUB1	chr10:116697893-116698376	0.314	RBP4	chr10:95360389-95361387	0.039
RHPN2	chr19:33555246-33556431	0.308	AQP11	chr11:77300360-77301391	0.036
USP43, WDR16	chr17:9548389-9549616	0.303	TBR1	chr2:162270888-162271413	0.035
ARHGAP21	chr10:25011963-25013816	0.301	ANKRD37, UFSP2	chr4:186317143-186318255	0.035
HOOK1	chr1:60280624-60281048	0.300	LACTB2, XKR9	chr8:71581050-71581650	0.034
ANKRD13D, SSH3	chr11:67070807-67071801	0.299	PCCA	chr13:100740956-100741805	0.034
SLC44A2	chr19:10735999-10736396	0.290	FAM111B	chr11:58873889-58874486	0.034
SPRY4	chr5:141705391-141705688	0.290	CGN, MIR554, TUFT1	chr1:151512661-151513199	0.033
DDAH1	chr1:85929940-85931168	0.288	C5orf49	chr5:7850957-7851413	0.033
RHOU	chr1:228870810-228872297	0.288	TJP1	chr15:30114110-30115215	0.033
MARVELD3	chr16:71659829-71660747	0.282	DNAJC5, TPD52L2	chr20:62525796-62526638	0.033
MOSC2	chr1:220921411-220922176	0.279	RNF20	chr9:104295917-104296232	0.032
IFT88	chr13:21140951-21141719	0.279	EPB41L4B	chr9:112083333-112083549	0.032
KRT18	chr12:53342805-53343162	0.278	EPS8	chr12:15941718-15942740	0.032
CYB5A	chr18:71958141-71959770	0.276	GNMT, PEX6	chr6:42928218-42928810	0.032
CRB3, DENND1C	chr19:6463991-6464780	0.276	CYP39A1, SLC25A27	chr6:46620541-46621189	0.031
EPS8	chr12:15941718-15942740	0.275	LOC646762	chr7:29724188-29725436	0.031
TMEM171	chr5:72415611-72416766	0.274	RHEB	chr7:151216068-151217901	0.031
ADCY6	chr12:49183049-49183282	0.271	TMEM45B	chr11:129685737-129686211	0.031

Appendix

(Supplementary Table) Table S19: Informative genes for predicting cell-line responses for Gemcitabine. We used the feature selection to identify informative genes for Gemcitabine drug-response prediction. Genomic coordinates are based on build 37 of the human genome. We used information gain to rank the genes; a higher score indicates a more informative gene.

Classification			**Regression**		
Gene	*Coordinates*	*Score*	*Gene*	*Coordinates*	*Score*
IFFO1, NOP2	chr12:6664425-6665336	0.220	YBX2	chr17:7197431-7198417	0.033
LOC100287834	chr7:62858468-62858826	0.216	TMEM177	chr2:120436530-120437010	0.032
TNK1	chr17:7284223-7284687	0.203	MACROD2	chr20:13976700-13977068	0.032
RNF39	chr6:30042918-30043500	0.199	ZNF793	chr19:37997790-37998125	0.032
C1orf229	chr1:247274585-247275757	0.190	CCDC64	chr12:120426547-120428066	0.030
SLC44A2	chr19:10735999-10736396	0.185	DUSP8, HCCA2, LOC338651	chr11:1593550-1594378	0.030
FAM174B	chr15:93198374-93199181	0.184	NEK10	chr3:27410612-27411066	0.029
EFNA1	chr1:155098434-155100451	0.183	CHN1	chr2:175869574-175870289	0.029
LAD1	chr1:201368560-201369032	0.183	ATP6V1C2	chr2:10861206-10862382	0.029
BIRC8	chr19:53794411-53794732	0.182	TLR2	chr4:154605086-154606052	0.029
YBX2	chr17:7197431-7198417	0.179	ZNF514	chr2:95824802-95825721	0.028
CRB3, DENND1C	chr19:6463991-6464780	0.179	CA8	chr8:61193312-61194195	0.028
KIAA0284	chr14:105332408-105332651	0.178	C17orf81, CLDN7	chr17:7164285-7166245	0.027
LOC100287704, LOC100287834	chr7:62809609-62809812	0.178	SCAI	chr9:127905675-127905947	0.027
CFDP1	chr16:75466850-75467527	0.176	MANSC1	chr12:12502942-12503465	0.027
C11orf90	chr11:93583374-93583717	0.172	SHC2	chr19:457800-462256	0.027
CYR61, DDAH1	chr1:86046362-86047240	0.166	STK25	chr2:242447017-242448558	0.027
TEAD4	chr12:3067960-3069444	0.166	ACPL2	chr3:140951193-140951451	0.027
CAMK2G	chr10:75633600-75634796	0.162	ZFP3	chr17:4981357-4981979	0.027
HM13	chr20:30102057-30102856	0.162	RUFY1	chr5:178986513-178986999	0.027

Appendix

(Supplementary Table) Table S20: Informative genes for predicting cell-line responses for Paclitaxel. We used the feature selection to identify informative genes for Paclitaxel drug-response prediction. Genomic coordinates are based on build 37 of the human genome. We used information gain to rank the genes; a higher score indicates a more informative gene.

Classification			Regression		
Gene	*Coordinates*	*Score*	*Gene*	*Coordinates*	*Score*
C2orf29	chr2:101869023-101869876	0.347	TUBB2B	chr6:3227025-3229688	0.062
LMO2	chr11:33890357-33891495	0.342	PRDM6	chr5:122424905-122425958	0.059
BLOC1S1, ITGA7, RDH5	chr12:56109798-56110298	0.315	ONECUT3	chr19:1753216-1755606	0.059
CTNNA2	chr2:79739696-79740243	0.307	DGKA	chr12:56325774-56326223	0.056
DTNA	chr18:32073444-32074292	0.274	PNMAL1	chr19:46974557-46975073	0.056
CDC40, WASF1	chr6:110500025-110500966	0.274	AP1S1, VGF	chr7:100806279-100809064	0.054
HMGA1	chr6:34202567-34206193	0.274	CNTNAP2	chr7:145813030-145814084	0.054
PHYHIPL	chr10:60935827-60937049	0.244	ZFP36	chr19:39897241-39898942	0.054
GALNTL6	chr4:172733734-172735118	0.244	SLC10A4	chr4:48485362-48486473	0.054
ZNF625	chr19:12266998-12267686	0.244	PAPOLB, RADIL	chr7:4901336-4901753	0.053
ATP5J, GABPA	chr21:27106814-27108211	0.238	TMEM25, TTC36	chr11:118401235-118402069	0.053
psiTPTE22	chr22:17083384-17083628	0.238	ABCB1	chr7:87230059-87230260	0.053
CACNG8	chr19:54466357-54466725	0.230	NOL4	chr18:31802358-31803792	0.053
FBXO36, TRIP12	chr2:230785912-230787665	0.224	DLL3	chr19:39989397-39990140	0.052
EYA4	chr6:133562086-133563586	0.224	MCFD2, TTC7A	chr2:47167858-47168978	0.052
EBPL	chr13:50265224-50265598	0.224	B3GAT1	chr11:134257428-134257631	0.052
DPRXP4, RNF135	chr17:29298046-29298606	0.224	ADAMTS3	chr4:73434855-73435321	0.051
AIM1	chr6:106959764-106960985	0.224	MAP3K12	chr12:53886562-53887101	0.051
IMMP2L	chr7:111202079-111202683	0.224	C6orf97	chr6:151814980-151815527	0.051
SFPQ	chr1:35657467-35658811	0.216	PIK3R3	chr1:46598126-46599129	0.051

Appendix

(Supplementary Table) Table S21: Informative genes for predicting cell-line responses for Temozolomide. We used the feature selection to identify informative genes for Temozolomide drug-response prediction. Genomic coordinates are based on build 37 of the human genome. We used information gain to rank the genes; a higher score indicates a more informative gene.

Classification			**Regression**		
Gene	*Coordinates*	*Score*	*Gene*	*Coordinates*	*Score*
TJP1	chr15:30114110-30115215	0.345	NELF, PNPLA7	chr9:140356314-140356987	0.087
AGAP2, LOC100130776	chr12:58119909-58121551	0.329	TJP1	chr15:30114110-30115215	0.085
ARHGAP29	chr1:94702690-94703344	0.316	MGAT1	chr5:180229375-180230147	0.082
C4orf14, POLR2B	chr4:57842634-57843893	0.307	PLEKHA1	chr10:124134088-124134933	0.082
RARA	chr17:38472958-38473201	0.298	DDAH1	chr1:85929940-85931168	0.081
ZNF280D	chr15:57025347-57026150	0.295	TEAD1	chr11:12695414-12696981	0.081
SYDE1	chr19:15217951-15218617	0.283	DSTN	chr20:17549628-17550051	0.079
TUBGCP2, ZNF511	chr10:135122851-135123109	0.281	ICAM3, RAVER1	chr19:10443688-10446022	0.077
ACP1, SH3YL1	chr2:263400-265238	0.277	SLC44A2	chr19:10735999-10736396	0.076
TBC1D12	chr10:96162023-96163327	0.275	CHST12	chr7:2442792-2444011	0.075
CTU1	chr19:51607207-51607840	0.274	FERMT3, STIP1	chr11:63974829-63975048	0.075
LARGE	chr22:34315841-34318637	0.272	GAS2L3	chr12:100967293-100967845	0.075
UTRN	chr6:144605926-144608280	0.269	CASZ1	chr1:10853894-10856964	0.074
AK1	chr9:130639738-130640143	0.269	SPN	chr16:29675845-29676120	0.074
DOCK1	chr10:128593609-128595048	0.269	PTPN14	chr1:214724104-214725056	0.074
NSUN7	chr4:40751842-40752497	0.267	LOC100133985	chr2:70352204-70352531	0.073
PARD6G	chr18:78004028-78005438	0.264	ERRFI1	chr1:8085554-8086854	0.073
RRN3P2	chr16:29086220-29086434	0.262	TMEM149, U2AF1L4	chr19:36231186-36232219	0.073
PKN1	chr19:14551998-14552255	0.259	FAT1	chr4:187644319-187648253	0.072
AGAP2	chr12:58132478-58132734	0.259	GNG7	chr19:2578956-2579746	0.072

Appendix

(Supplementary Table) Table S22: Gene-set analysis for the classification analysis. We used a statistical overrepresentation test to identify protein classes associated with the top-20 ranked genes in the feature-selection analysis.

Gefitinib				
Gene Set Name [# Genes (K)]	*Description*	*# Genes in Overlap (k)*	*p-value*	*FDR q-value*
KOINUMA_TARGETS_OF_SMAD2_OR_SMAD3 [843]	Genes with promoters occupied by SMAD2 or SMAD3 [GeneID=4087, 4088] in HaCaT cells (keratinocyte) according to a ChIP-chip analysis.	9	1.81 e-10	1.32 e-6
GENTILE_UV_RESPONSE_CLUSTER_D4 [54]	Cluster d4: genes progressively down-regulated in WS1 cells (fibroblast) through 12 h after irradiation with high dose UV-C.	3	3.02 e-6	1.1 e-2
FRIDMAN_SENESCENCE_UP [77]	Genes up-regulated in senescent cells.	3	8.85 e-6	2.15 e-2
SHEDDEN_LUNG_CANCER_GOOD_SURVIVAL_A12 [320]	Cluster 12 of method A: up-regulation of these genes in patients with non-small cell lung cancer (NSCLC) predicts good survival outcome.	4	2.15 e-5	3.9 e-2
TSUNODA_CISPLATIN_RESISTANCE_UP [15]	Genes up-regulated in bladder cancer cells resistant to cisplatin [PubChem=2767] compared to the parental cells sensitive to the drug.	2	2.74 e-5	3.98 e-2
Cisplatin				
Gene Set Name [# Genes (K)]	*Description*	*# Genes in Overlap (k)*	*p-value*	*FDR q-value*
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESEN_SENCHYMAL_DN [465]	Genes down-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones.	6	7.17 e-7	4.94 e-3
KOINUMA_TARGETS_OF_SMAD2_OR_SMAD3 [843]	Genes with promoters occupied by SMAD2 or SMAD3 [GeneID=4087, 4088] in HaCaT cells (keratinocyte) according to a ChIP-chip analysis.	7	1.43 e-6	4.94 e-3
REACTOME_CELL_CELL_COMMUNICATION [130]	Cell-Cell communication	4	2.04 e-6	4.94 e-3
SENGUPTA_NASOPHARYNGEAL_CARCINOMA_DN [358]	Genes down-regulated in nsopharyngeal carcinoma relative to the normal tissue.	5	4.59 e-6	8.35 e-3
HUPER_BREAST_BASAL_VS_LUMINAL_DN [58]	Genes down-regulated in basal mammary epithelial cells compared to the luminal ones.	3	9.19 e-6	1.34 e-2
GU_PDEF_TARGETS_UP [71]	Integrin, VEGF, Wnt and TGFbeta signaling pathway genes up-regulated in PC-3 cells (prostate cancer) after	3	1.69 e-5	1.55 e-2

	knockdown of PDEF [GeneID=25803] by RNAi.			
ONDER_CDH1_TARGETS_2_DN [473]	Genes down-regulated in HMLE cells (immortalized nontransformed mammary epithelium) after E-cadhedrin (CDH1) [GeneID=999] knockdown by RNAi.	5	1.76 e-5	1.55 e-2
COLDREN_GEFITINIB_RESISTANCE_DN [228]	Genes down-regulated in NSCLC (non-small cell lung carcinoma) cell lines resistant to gefitinib [PubChem=123631] compared to the sensitive ones.	4	1.88 e-5	1.55 e-2
WP_PRIMARY_FOCAL_SEGMENTAL_GLOMERULOSC_OSCLEROSIS_FSGS [74]	Primary Focal Segmental Glomerulosclerosis FSGS	3	1.92 e-5	1.55 e-2
FERRANDO_T_ALL_WITH_MLL_ENL_FUSION_UP [89]	Top 100 genes positively associated with T-cell acute lymphoblastic leukemia MLL T-ALL) expressing MLL-ENL fusion [GeneID=4297;4298].	3	3.33 e-5	2.23 e-2

Docetaxel

Gene Set Name [# Genes (K)]	*Description*	*# Genes in Overlap (k)*	*p-value*	*FDR q-value*
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESEN_SENCHYMAL_DN [465]	Genes down-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones.	7	3.97 e-9	1.82 e-5
LIM_MAMMARY_STEM_CELL_UP [481]	Genes consistently up-regulated in mammary stem cells both in mouse and human species.	7	5.02 e-9	1.82 e-5
HUANG_DASATINIB_RESISTANCE_UP [80]	Genes whose expression positively correlated with sensitivity of breast cancer cell lines to dasatinib [PubChem=3062316].	4	1.05 e-7	2.54 e-4
KOINUMA_TARGETS_OF_SMAD2_OR_SMAD3 [843]	Genes with promoters occupied by SMAD2 or SMAD3 [GeneID=4087, 4088] in HaCaT cells (keratinocyte) according to a ChIP-chip analysis.	7	2.3 e-7	4.19 e-4
PETROVA_ENDOTHELIUM_LYMPHATIC_VS_BLOOD_OOD_DN [162]	Genes down-regulated in BEC (blood endothelial cells) compared to LEC (lymphatic endothelial cells).	4	1.78 e-6	2.59 e-3
SESTO_RESPONSE_TO_UV_C5 [46]	Cluster 5: genes changed in primary keratinocytes by UVB irradiation.	3	2.15 e-6	2.6 e-3
EGFR_UP.V1_UP [192]	Genes up-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] and engineered to express	4	3.49 e-6	3.35 e-3

	ligand-activatable EGFR [Gene ID=1956].			
MITSIADES_RESPONSE_TO_APLIDIN_UP [446]	Genes up-regulated in the MM1S cells (multiple myeloma) after treatment with aplidin [PubChem=44152164], a marine-derived compound with potential anti-cancer properties.	5	3.76 e-6	3.35 e-3
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_SAL_DN [455]	Genes down-regulated in luminal-like breast cancer cell lines compared to the basal-like ones.	5	4.15 e-6	3.35 e-3
ENK_UV_RESPONSE_EPIDERMIS_DN [513]	Genes down-regulated in epidermis after to UVB irradiation.	5	7.42 e-6	5.4 e-3

Doxorubicin

Gene Set Name [# Genes (K)]	*Description*	*# Genes in Overlap (k)*	*p-value*	*FDR q-value*
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESEN_SENCHYMAL_UP [453]	Genes up-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones.	5	6.43 e-6	2.61 e-2
PILON_KLF1_TARGETS_UP [501]	Genes up-regulated in erythroid progenitor cells from fetal livers of E13.5 embryos with KLF1 [GeneID=10661] knockout compared to those from the wild type embryos.	5	1.05 e-5	2.61 e-2
DUTERTRE_ESTRADIOL_RESPONSE_24HR_DN [504]	Genes down-regulated in MCF7 cells (breast cancer) at 24 h of estradiol [PubChem=5757] treatment.	5	1.08 e-5	2.61 e-2
MIKKELSEN_MEF_HCP_WITH_H3K27ME3 [590]	Genes with high-CpG-density promoters (HCP) bearing histone H3 trimethylation mark at K27 (H3K27me3) in MEF cells (embryonic fibroblast).	5	2.29 e-5	4.17 e-2

Etoposide

Gene Set Name [# Genes (K)]	*Description*	*# Genes in Overlap (k)*	*p-value*	*FDR q-value*
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESEN_SENCHYMAL_UP [453]	Genes up-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones.	6	2.29 e-7	1.66 e-3
MODULE_180 [119]	Genes in the cancer module 180.	4	7.5 e-7	2.73 e-3
LIM_MAMMARY_STEM_CELL_DN [416]	Genes consistently down-regulated in mammary stem cells both in mouse and human species.	5	4.25 e-6	1.03 e-2
MODULE_342 [213]	Genes in the cancer module 342.	4	7.59 e-6	1.38 e-2

MEISSNER_NPC_HCP_WITH_H3_UNMETHYLATED [542]	Genes with high-CpG-density promoters (HCP) that have no histone H3 methylation marks in neural precursor cells (NPC).	5	1.53 e-5	1.96 e-2
BOYLAN_MULTIPLE_MYELOMA_D_DN [82]	Genes down-regulated in group D of tumors arising from overexpression of BCL2L1 and MYC [GeneID=598;4609] in plasma cells.	3	1.62 e-5	1.96 e-2
MEISSNER_BRAIN_HCP_WITH_H3K27ME3 [271]	Genes with high-CpG-density promoters (HCP) bearing the H3K27 tri-methylation (H3K27me3) mark in brain.	4	1.95 e-5	2.03 e-2

Gemcitabine
No overlaps found.

Paclitaxel
No overlaps found.

Temozolomide
No overlaps found.

Appendix

(Supplementary Table) Table S23: Gene-set evaluation using GSEA for the regression analysis. We used a statistical overrepresentation test to identify protein classes associated with the top-20 ranked genes in the feature-selection analysis.

Gefitinib				
Gene Set Name [# Genes (K)]	*Description*	*# Genes in Overlap (k)*	*p-value*	*FDR q-value*
RICKMAN_TUMOR_ DIFFERENTIATED_WELL_ VS_PS_POORLY_DN [380]	Down-regulated genes that vary between HNSCC (head and neck squamous cell carcinoma) groups formed on the basis of their level of pathological differentiation: well vs poorly differentiated tumors.	5	3.4 e-6	1.9 e-2
EGFR_UP.V1_UP [192]	Genes up-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] and engineered to express ligand-activatable EGFR [Gene ID=1956].	4	5.97 e-6	1.9 e-2
COLLER_MYC_ TARGETS_DN [7]	Genes down-regulated in 293T (transformed fetal renal cell) upon expression of MYC [GeneID=4609].	2	7.83 e-6	1.9 e-2
Cisplatin				
Gene Set Name [# Genes (K)]	*Description*	*# Genes in Overlap (k)*	*p-value*	*FDR q-value*
HOLLERN_EMT_BREAST_ TUMOR_DN [123]	Genes that that have low expression in mammary tumors of epithelial-mesenchymal transition (EMT) histology.	9	2.26 e-17	1.64 e-13
CHARAFE_BREAST_ CANCER_LUMINAL_VS_ MESEN_SENCHYMAL_ UP [453]	Genes up-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones.	11	7.47 e-16	2.72 e-12
ONDER_CDH1_TARGETS_ 2_DN [473]	Genes down-regulated in HMLE cells (immortalized nontransformed mammary epithelium) after E-cadhedrin (CDH1) [GeneID=999] knockdown by RNAi.	9	4.61 e-12	1.12 e-8
COLDREN_GEFITINIB_ RESISTANCE_DN [228]	Genes down-regulated in NSCLC (non-small cell lung carcinoma) cell lines resistant to gefitinib [PubChem=123631] compared to the sensitive ones.	7	5.61 e-11	1.02 e-7
MCBRYAN_PUBERTAL_ BREAST_4_5WK_UP [270]	Genes up-regulated during pubertal mammary gland development between week 4 and 5.	7	1.83 e-10	2.66 e-7
WU_CELL_MIGRATION [183]	Genes associated with migration rate of 40 human bladder cancer cells.	6	1.05 e-9	1.27 e-6

LIM_MAMMARY_STEM_ CELL_DN [416]	Genes consistently down-regulated in mammary stem cells both in mouse and human species.	7	3.67 e-9	3.82 e-6
BOYAULT_LIVER_CANCER_ SUBCLASS_G1_UP [116]	Up-regulated genes in hepatocellular carcinoma (HCC) subclass G1, defined by unsupervised clustering.	5	7.58 e-9	6.89 e-6
MODULE_180 [119]	Genes in the cancer module 180.	5	8.62 e-9	6.97 e-6
KEGG_TIGHT_ JUNCTION [132]	Tight junction.	5	1.45 e-8	1.06 e-5

Docetaxel

Gene Set Name [# Genes (K)]	*Description*	*# Genes in Overlap (k)*	*p-value*	*FDR q-value*
NIKOLSKY_BREAST_ CANCER_16P13_ AMPLICON [119]	Genes within amplicon 16p13 identified in a study of 191 breast tumor samples.	4	1.05 e-6	7.64 e-3
CHARAFE_BREAST_ CANCER_LUMINAL_VS_ MESEN_SENCHYMAL_ DN [465]	Genes down-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones.	5	1.11 e-5	2.89 e-2
KOINUMA_TARGETS_ OF_ SMAD2_OR_SMAD3 [843]	Genes with promoters occupied by SMAD2 or SMAD3 [GeneID=4087, 4088] in HaCaT cells (keratinocyte) according to a ChIP-chip analysis.	6	1.37 e-5	2.89 e-2
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY [75]	B cell receptor signaling pathway	3	1.59 e-5	2.89 e-2

Doxorubicin

No overlaps found.

Etoposide

Gene Set Name [# Genes (K)]	*Description*	*# Genes in Overlap (k)*	*p-value*	*FDR q-value*
SENGUPTA_NASOPHARYNGEAL_CARCINOMA_DN [358]	Genes down-regulated in nsopharyngeal carcinoma relative to the normal tissue.	5	3.12 e-6	2.27 e-2

Gemcitabine

No overlaps found.

Paclitaxel

No overlaps found.

Temozolomide

No overlaps found.