Dissertation

submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by: Roman Kurilov, M.Sc.
born in: Kirovsk, Russia
Oral examination: 15th February 2019

Assessment of modeling strategies for drug response prediction in cell lines and xenografts

Referees:
Prof. Dr. Benedikt Brors
Prof. Dr. Thorsten Zenz

TABLE OF CONTENTS

SUMMARY

Despite significant progress in cancer research, effective cancer treatment is still a challenge. Cancer treatment approaches are shifting from standard cytotoxic chemotherapy regimens towards a precision oncology paradigm, where a choice of treatment is personalized, i.e. based on a tumor's molecular features. In order to match tumor molecular features with therapeutics we need to identify biomarkers of response and build predictive models. Recent growth of large-scale pharmacogenomics resources which combine drug sensitivity and multi-omics information on a large number of samples provides necessary data for biomarker identification and drug response modelling. However, although many efforts of using this information for drug response prediction have been made, our ability to accurately predict drug response using genetic data remains limited.

In this work we used pharmacogenomics data from the largest publicly available studies in order to systematically assess various aspects of the drug response model-building process with the ultimate goal of improving prediction accuracy. We applied several machine learning methods (regularized regression, support vector machines, random forest) for predicting response to a number of drugs. We found that while accuracy of response prediction varies across drugs (in most of the cases R^2 values vary between 0.1 and 0.3), different machine learning algorithms applied for the the same drug have similar prediction performance. Experiments with a range of different training sets for the same drug showed that predictive power of a model depends on the type of molecular data, the selected drug response metric, and the size of the training set. It depends less on number of features selected for modelling and on class imbalance in training set. We also implemented and tested two methods for improving consistency for pharmacogenomics data coming from different datasets.

We tested our ability to correctly predict response in xenografts and patients using models trained on cell lines. Only in a fraction of the tested cases we managed to get reasonably accurate predictions, particularly in case of response to erlotinib in the NSCLC xenograft cohort, and in cases of responses to erlotinib and docetaxel in the NSCLC and BRCA patient cohorts respectively.

This work also includes two applied pharmacogenomics analyses. The first is an analysis of a drug-sensitivity screen performed on a panel of Burkitt cell lines. This combines unsupervised data exploration with supervised modelling. The second is an analysis of drug-sensitivity data for the DKFZ-608 compound and the generation of the corresponding response prediction model.

In summary, we applied machine learning techniques to available high-throughput pharmacogenomics data to study the determinants of accurate drug response prediction. Our results can help to draft guidelines for building accurate models for personalized drug response prediction and therefore contribute to advancing of precision oncology.

ZUSAMMENFASSUNG

Trotz erheblicher Fortschritte in der Krebsforschung bleibt die effektive Behandlung von Krebs eine Herausforderung. Die Behandlungsansätze verschieben sich von der üblichen zytotoxischen Chemotherapie hin zu einem präzisionsonkologischen Modell, in dem die Behandlungswahl personalisiert ist und auf den molekularen Eigenschaften des Tumors basiert. Um passende Therapeutika für die molekularen Krebseigenschaften zu finden, müssen Biomarker für das Therapieansprechen identifiziert und prädiktive Modelle erstellt werden. Das jüngste Wachstum an umfassenden pharmacogenomischen Ressourcen, die Wirkstoffsensitivität und multi-omics Informationen einer großen Anzahl an Proben vereinen, liefern die nötigen Daten für Biomarker-Identifizierung und Erstellung von Modellen zum Wirkstoffansprechen. Trotz vieler Bemühungen diese Informationen zur Vorhersage von Therapieansprechen zu nutzen, bleiben die Möglichkeiten, Wirkstoffansprechen präzise aus genetischen Daten vorherzusagen, begrenzt.

In der vorliegenden Arbeit wurden pharmacogenomische Daten der größten öffentlich verfügbaren Studien genutzt, um systematisch verschiedene Aspekte der Erstellungsprozesse von Wirkstoff-Ansprech-Modellen einzuschätzen, mit dem ultimativen Ziel die Vorhersagegenauigkeit zu verbessern. Mehrere maschinelle Lernverfahren (regularisierte Regression, Support Vector Machinen, Random Forest) wurden auf eine Vielzahl von Wirkstoffen angewandt, um das Ansprechen vorherzusagen. Dabei wurde herausgefunden, dass die Genauigkeit der Ansprechvorhersage von Wirkstoff zu Wirkstoff variiert (in dem meisten Fällen liegen die R^2-Werte zwischen 0.1 und 0.3). Die verschiedenen Algorithmen für maschinelles Lernen weisen aber ähnliche Prognosefähigkeiten auf, wenn sie auf den gleichen Wirkstoff angewandt werden. Experimente mit einer Reihe verschiedener Trainingsdatensätze für den gleichen Wirkstoff haben gezeigt, dass die Vorhersagekraft eines Modells von der Art der molekularen Daten, der gewählten Metrik für Wirkstoffansprechen und der Größe des Trainingsdatensatzes abhängt. Es hängt dagegen weniger von der Anzahl der Merkmale, die für die Modellierung gewählt wurden, oder dem Ungleichgewicht der Klassen im Trainingsdatensatz ab. Außerdem wurden zwei Methoden implementiert und getestet, die die Konsistenz von Pharmacogenomicsdaten aus verschiedenen Datensätzen verbessert.

Desweiteren wurde evaluiert, ob das Ansprechen in Xenotransplantaten und Patienten mit Hilfe von Modellen, die auf Zelllinien trainiert wurden, vorhergesagt werden kann. Hinreichend genaue Prognosen konnten nur in einem Bruchteil der getesteten Fälle erreicht werden, vor allem in Bezug auf Erlotinib in der NSCLC Xenotransplantat Kohorte beziehungsweise Erlotinib und Docetaxel in den NSCLC und BRCA Patientenkohorten.

Diese Arbeit beinhaltet auch zwei angewandte pharmakogenomische Analysen. Die erste ist eine Analyse eines Wirkstoffempfindlichkeitscreenings, welches auf einer Reihe von Burkitt Zelllinien basiert. Dabei wurde unüberwachte Datenerkundung mit überwachter Modell-Erstellung kombiniert. Die zweite ist eine Analyse der Wirkstoffempfindlichkeitsdaten für den DKFZ-608 Wirkstoff und die Erstellung des zugehörigen Modells zur Ansprechensvorhersage.

Zusammengefasst wurden maschinelle Lernverfahren auf verfügbare Hochdurchsatz-Pharmacogenomicsdaten angewandt, um die Einflussfaktoren auf präzise Vorhersagen über Wirkstoffansprechen zu untersuchen. Die Ergebnisse können das Konzipieren von Richtlinien zur Erstellung genauer Modelle für das personalisierte Vorhersagen von Wirkstoffansprechen unterstützen und somit einen Beitrag für den Fortschritt der Präzisionsonkologie leisten.

LIST OF ABBREVIATIONS

AUC	Area under the curve (drug response metric)
AUROC	Area under the receiver operating characteristic (ROC) curve (classification accuracy metric)
BGP	Binary gene pairs
CCLE	Cancer Cell Line Encyclopedia
CI	Combination index
CTRP	Cancer Therapeutics Response Portal
DREAM	Dialogue for Reverse Engineering Assessments and Methods (consortium)
FDA	Food and Drug Administration (US federal agency)
FS	Feature selection
gCSI	Genentech Cell Line Screening Initiative
GDSC	Genomics of Drug Sensitivity in Cancer
GLDS	General level of drug sensitivity
IC_{50}	Half-maximal inhibitory concentration
ML	Machine learning
NIBR PDXE	Novartis Institutes of Biomedical Research patient-derived tumor xenograft encyclopedia
PC	Principal component
PCA	Principal component analysis
PDX	Patient-derived xenograft
R^2	R-squared, coefficient of determination
RF	Random Forest
RMSE	Root of mean squared error
SVM	Support vector machines

1 INTRODUCTION

Personalized oncology is an approach to cancer treatment that seeks to identify effective therapeutic strategies for every patient. This identification is possible via integration of genomic and drug-sensitivity data and subsequent generation of drug-response associations. While personalized approach is not yet a part of routine care for most cancer patients, its abundance is continuing to grow due to the progress in areas of multi-omics tumor characterization, drug-sensitivity testing and data integration. In this thesis we examine applicability of pharmacogenomics data (genomics + drug response) available up to date for accurate drug response prediction using machine learning models.

In this introductory chapter I'll start with giving an overview of general principles of cancer treatment, then we'll discuss drug therapy with an emphasis on targeted therapies. After that I'll describe experimental drug sensitivity testing and large pharmacogenomics projects that generated data we used in our analyses, also we'll discuss a problem with (in)consistency between pharmacogenomics data coming from different projects. In the end of the chapter I'll introduce basic machine learning concepts and describe machine learning methods we used in our analyses for building predictive drug response models.

1.1 General principles of cancer treatment

The standard treatment modalities for patients with cancer include surgery, radiotherapy and drug therapy.[1] Surgery aims to physically cut out the tumor. Radiotherapy uses ionizing radiation (e.g. X-rays) to kill cancer cells and shrink tumors. Drug therapy is the treatment of cancer with single drugs or drug combinations.

The choice of treatment approaches depends on 3 groups of factors: tumor factors, treatment factors and patient factors. Tumor factors include type of cancer (characterized by histological and molecular information) spread of cancer and its stage. Treatment factors include availability of treatment, evidence on its efficacy for given disease and side effects.
Patient factors include patient performance status and patient preferences.[1]

With respect to goal of cancer treatment, the are two types of therapy: curative and palliative. The goal of curative therapy is to cure the patient of cancer. In cases where cure is not feasible (e.g. when cancer is metastatic) the therapy is palliative, i.e. the goal is to improve symptoms, quality of life and prolong survival through tumor stabilisation or shrinkage.[2]

There is a special terminology for describing the outcome of treatment. Complete disappearance of all tumor would constitute a complete clinical remission. If tumor have been reduced by 50% or more it would be described as partial clinical

remission. In case when tumor is unchanged by treatment it can be described as stable disease. If the tumors grew during the treatment, this would be considered as progressive disease. There is a difference between “complete remission” and “cure”, since the latter term means not only that there is no traces of cancer after treatment but also that cancer will never come back therefore term “cure” can be truly applied only in retrospect. Indeed, some patients who achieve even a complete remission may have a regrowth of cancer after a disease-free period. Such regrowth is referred to as a recurrence or relapse of the tumor.[1]

1.2 Drug therapy. Targeted/non-targeted therapy dichotomy

Drug therapy (or systemic therapy) is a mainstay of treatment for most types of cancer.[2] Drug therapies work in various ways to destroy cancer cells, stop them from spreading or slow down their growth. Usually cancer drug treatments fall into four categories: conventional cytotoxic chemotherapy, hormonal therapy, targeted therapy and immunotherapy.[3] Hormonal therapy exploits dependence of some cancers on hormones and stops tumor growth by blocking certain hormones e.g. estrogen in breast cancer or testosterone in prostate cancer. Immunotherapy is a diverse set of therapeutic strategies designed to induce patient’s own immune system to fight the tumor; these approaches exploit the fact that cancer cells often have molecules on their surface that can be detected by the immune system, known as tumor antigens, they are often proteins or other macromolecules.[4]

Our work focuses on drug response prediction for cytotoxic chemotherapeutics and targeted therapies. These two categories form a dichotomy with respect to drug’s selectivity. Most cytotoxic chemotherapy drugs are agents designed to attack actively dividing cells, based on the fact that cancer cells divide more rapidly than normal cells. However, cytotoxic chemotherapeutics are unspecific and also destroy some normal cells causing unwanted toxic effects. On the other hand a targeted therapy has a specific molecular target which is restricted to and critical for cancer cell growth, therefore targeted therapy shouldn’t be toxic for patient’s normal tissue.[2] Let’s review both chemotherapy and targeted therapy in greater detail.

1.2.1 Cytotoxic chemotherapy

Chemotherapy acts by interfering with basic properties of cancer cell, such as growth and proliferation, DNA synthesis, metabolism and other essential cellular functions. Regardless of the specific mechanism it generally kills a cancer by activating the apoptosis.[1]

The era of chemotherapy began in the 1940s when nitrogen mustards and antifolate drugs were first used for cancer treatment.[5] Since then many chemicals have been studied and tested for their effects on cancer cells, and a large number of compounds now play a role in cancer treatment. They can be broadly separated into categories based on mechanism of action (Table 1).

Table 1. Chemotherapy agents and their mechanisms of action. Adapted from Pardee & Stein.[1]

Category	Mechanism of action	Drugs
Alkylating agents; Platinum-based agents	Cause breaks or mutations in DNA	cyclophosphamide, temozolomide; cisplatin, oxaliplatin, carboplatin
Anti-metabolites	Block DNA synthesis	methotrexate, cytarabine, 5-fluorouracil, capecitabine, gemcitabine, 6-mercaptopurine
Anti-tumor antibiotics	Bind to DNA and prevent RNA synthesis leading to cell death	doxorubicin, epirubicin, bleomycin, mitoxantrone
Topoisomerase inhibitors	Block unwinding of DNA and therefore interfere with replication and transcription	etoposide, irinotecan, topotecan
Microtubule inhibitors	Prevent microtubules in cells from supporting cell division	paclitaxel, docetaxel, vincristine, vinblastine

As the majority of chemotherapeutic are cytotoxic i.e. their function is to kill rapidly dividing cells, they are invariably associated with a range of toxic side effects due to lack of specificity to cancer cells. These side effects are especially prevalent in cells/tissues with a high turnover including skin, gastrointestinal tract and bone marrow and lead to some adverse reactions. Hair loss, mucositis (inflammation of the lining of the digestive tract), diarrhoea, vomiting, myelosuppression (decreased production of blood cells) are all common side effects that can have devastating morbidity and can be fatal.[2]

1.2.2 Targeted therapy

Recently the ability to characterize specific gene mutations in different cancers, and a greater biological understanding of the cellular events and pathways driving carcinogenesis, has led to new approaches to cancer treatment. Such approaches, called targeted therapy, are aimed at specific molecular alterations that contribute to the growth of cancer cells and therefore they deliver growth inhibitory or cytotoxic effects in a much more cell-specific manner.[3] Targeted therapies can be divided into two categories -- small-molecule inhibitors and monoclonal antibodies (see Table 2). Small molecule inhibitors can pass through cell membrane and inhibit mutated/overexpressed proteins critical for cancer growth. Monoclonal antibodies act on the surface of cancer cell by binding to specific cell surface proteins which prevents growth signal transmission or induces the immune response.[1]

Table 2. Examples of targeted therapies. Adapted from Pardee & Stein.[1]

Drug	Major targets	Disease
Small molecule		
Sorafenib	BRAF, VEGFR; EGFR	Renal cell cancer; liver cancer
Sunitinib	VEGFR, c-kit, FLT3	Renal cell cancer; GI stromal tumor
Erlotinib	EGFR	Non-small cell lung cancer; pancreatic cancer
Gefitinib	EGFR	Non-small cell lung cancer
Bortezomib	Proteasome	Myeloma
Lapatinib	Her-2/Neu, EGFR	Breast cancer
Imatinib	Bcr-Abl; c-kit	Chronic myelocitic leukemia; acute lymphoblastic leukemia; GI stromal tumor, mastocytosis
Dasatinib	Bcr-Abl; c-kit	Chronic myelocitic leukemia; acute lymphoblastic leukemia
Monoclonal antibodies		
Bevacizumab	VEGF	Colorectal cancer; non-small cell lung cancer
Rituximab	CD20	B-cell lymphoma
Cetuximab	EGFR	Colorectal cancer; head and neck cancer
Gemtuzumab	CD33	Acute myelogenous leukemia
Alemtuzumab	CD52	Chronic lymphocytic leukemia
Y-ibritumomab	CD20	Non-Hodgkin's lymphoma

During the last decade efforts of international consortia (with TCGA[6] and ICGC[7] being most prominent ones) produced detailed characterization of the common somatic genetic alterations in a variety of different tumor types. Theoretically those driver alterations (i.e. alterations that cause cancer phenotype) which are druggable (i.e. can blocked/inhibited by therapeutics in a specific manner) present potential treatment opportunities. Druggable alterations have been identified in a substantial proportion of several major tumor types (Fig. 1). Many of these mutations encode targets of already approved drugs.[8]

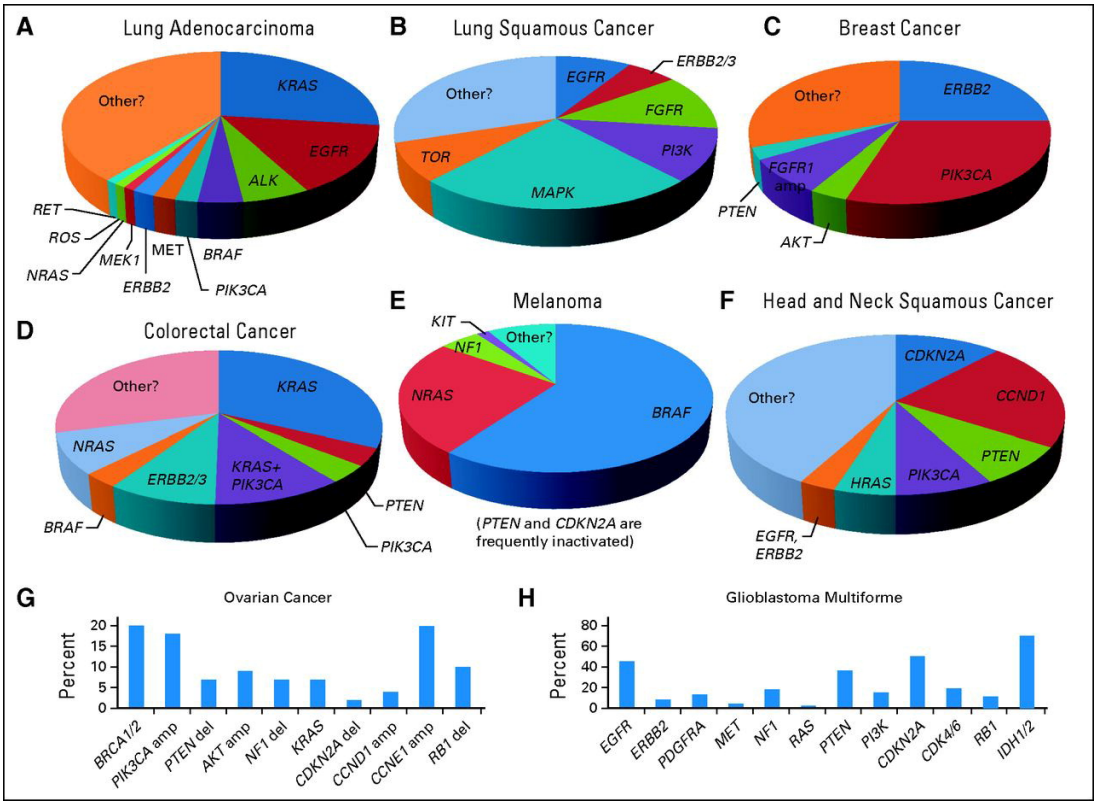


Figure 1. Genomic alterations affecting actionable signaling pathways in common solid tumors. "Pie charts for (A) lung adenocarcinoma, (B) lung squamous cancer, (C) breast cancer, (D) colorectal cancer, (E) melanoma, and (F) head and neck squamous cancer show the distribution of known recurrent driver cancer gene mutations, with emphasis on those genes/pathways targeted by ≥ one anticancer agent that is either FDA approved or in clinical development; other denotes proportion of tumors containing undruggable drivers or where driver gene has not yet been conclusively delineated. Bar graphs are shown for (G) ovarian cancer and (H) glioblastoma multiforme, where plausibly actionable cancer gene mutations are frequent but not mutually exclusive; in these cases, driver genes are commonly dysregulated by chromosomal copy number alterations as well as base mutations" (Figure taken from Garraway[8]).

Indeed, due to unprecedented advances in cancer drug development since the beginning of 21st century a broad spectrum of therapeutics directed against multiple effector proteins spanning most cancer signaling pathways has entered clinical trials and, in some cases, clinical practice[8] (Fig 2). The recent analyses showed that by 2017 there were around 90 FDA approved targeted therapies with more than 100 associated molecular targets[9,10] (The updated list of FDA approved drugs is available at mycancergenome.org web resource[11]).

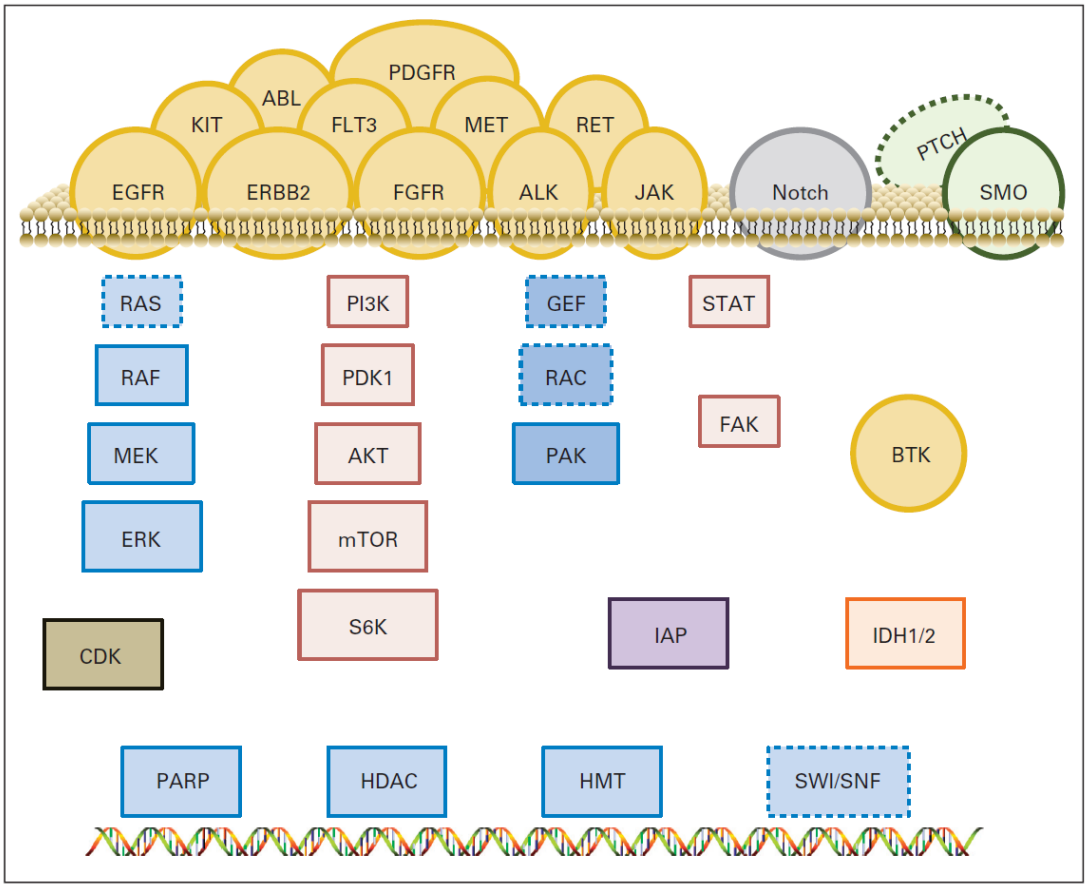


Figure 2. Spectrum of targeted anticancer agents in clinical development. "Exemplary oncoproteins and cancer pathways targeted by at least one US Food and Drug Administration–approved or developmental drug are indicated. Examples include receptor tyrosine kinases (orange); mitogen-activated protein kinase pathway proteins (RAS, RAF, MEK, and ERK; light blue); phosphoinositol-3 kinase (PI3K) pathway components (PI3K, phosphoinositide-dependent protein kinase 1 [PDK1], AKT, mammalian target of rapamycin [mTOR], and ribosomal S6 kinase [S6K]; pink); the RAC/PAK pathway (dark blue), the Janus kinase (JAK) –signal transducer and activator of transcription pathway (STAT) pathway; Notch (gray); the sonic hedgehog pathway, including patched (PTCH) and smoothened (SMO) proteins (green); cyclin-dependent kinases (CDKs; olive), inhibitor of apoptosis (IAP) proteins (purple); and isocitrate dehydrogenases (IDH1/2; peach). Other targets of developmental drugs whose efficacy may be governed by driver genetic alterations (bottom) include poly (ADP-ribose) polymerase (PARP), histone deacetylases (HDACs), and histone methyltransferases (HMTs). Proteins with dashed borders (RAS, guanine exchange factors [GEFs], RAC, PTCH, SWI/SNF) represent key pathway or epigenetic effectors not yet directly targeted by drugs in development. ALK, anaplastic lymphoma kinase; BTK, Bruton's tyrosine kinase; EGFR, epidermal growth factor receptor; FAK, focal adhesion kinase; FGFR, fibroblast growth factor receptor; FLT3, FMS-related tyrosine kinase 3; PDGFR, platelet-derived growth factor receptor" (Figure taken from Garraway[8]).

Analysis of subcellular location of targets of these drugs showed that a bit less than half of the targets are located in the plasma membrane, one quarter is located in cytoplasm, one quarter is located in nucleus and just 7% is in extracellular space (Fig 3a). Analysis of targets' functional annotation revealed that two largest groups of targets are enzymes (57%) and receptors (26%). Within the enzymes the largest

subfamily is tyrosine kinases, other prominent subfamilies include serine/threonine kinases, peptidases and epigenetic modulators. Within receptor subfamilies there are transmembrane receptors, ligand-dependent nuclear receptors and G-protein coupled receptors (Fig. 3b,c).[9]

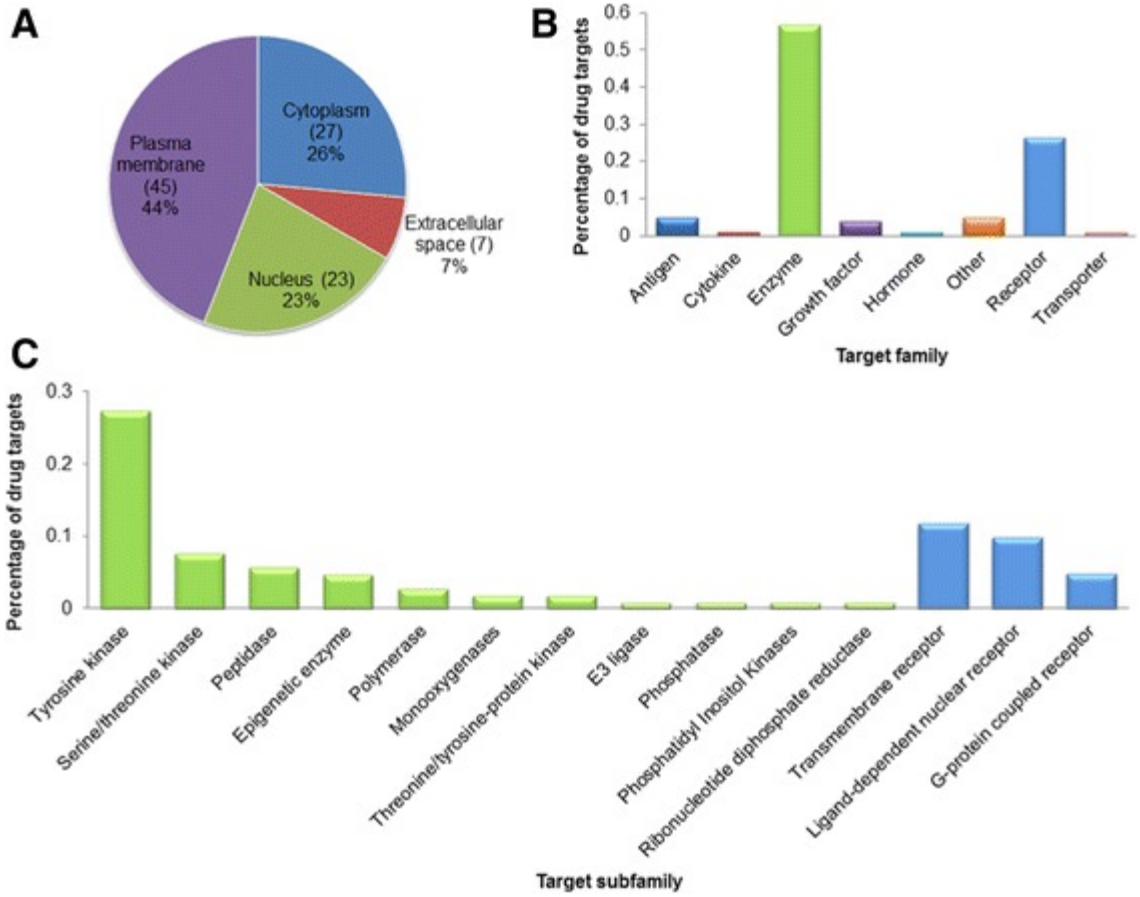


Figure 3. Subcellular location and function of drug targets. (**a**) Percentage of anticancer drug targets belonging different subcellular locations (**b**) drug targets function families breakdown (**c**) detailed functional classification for enzymes and receptors families (Figure taken from Sun et al.[9])

Comparison of list of consensus 33 cancer driver genes from Cancer Gene Census[12] with list of 109 protein targets of FDA approved drugs showed that there are only 30 proteins in overlap between the two sets.[10,13] Therefore there is a huge potential in extending the spectrum of targeted drugs covering the whole range of cancer vulnerabilities including oncogene and non-oncogene dependencies.[14]

Most tumors usually have defects in more than one signalling pathway. Therefore, a dual-targeting or multitargeting might be a rational strategy to eliminate cancer cells efficiently[1] (Figure 4 shows different hallmarks of tumor progression and classes of targeted therapies that interfere with these hallmarks[15]). Also combining drugs that interfere with "parallel" pathways can limit the emergence of drug resistance which is critical since drug resistance results in cancer relapse. Let's review resistance to treatment in the next subsection.

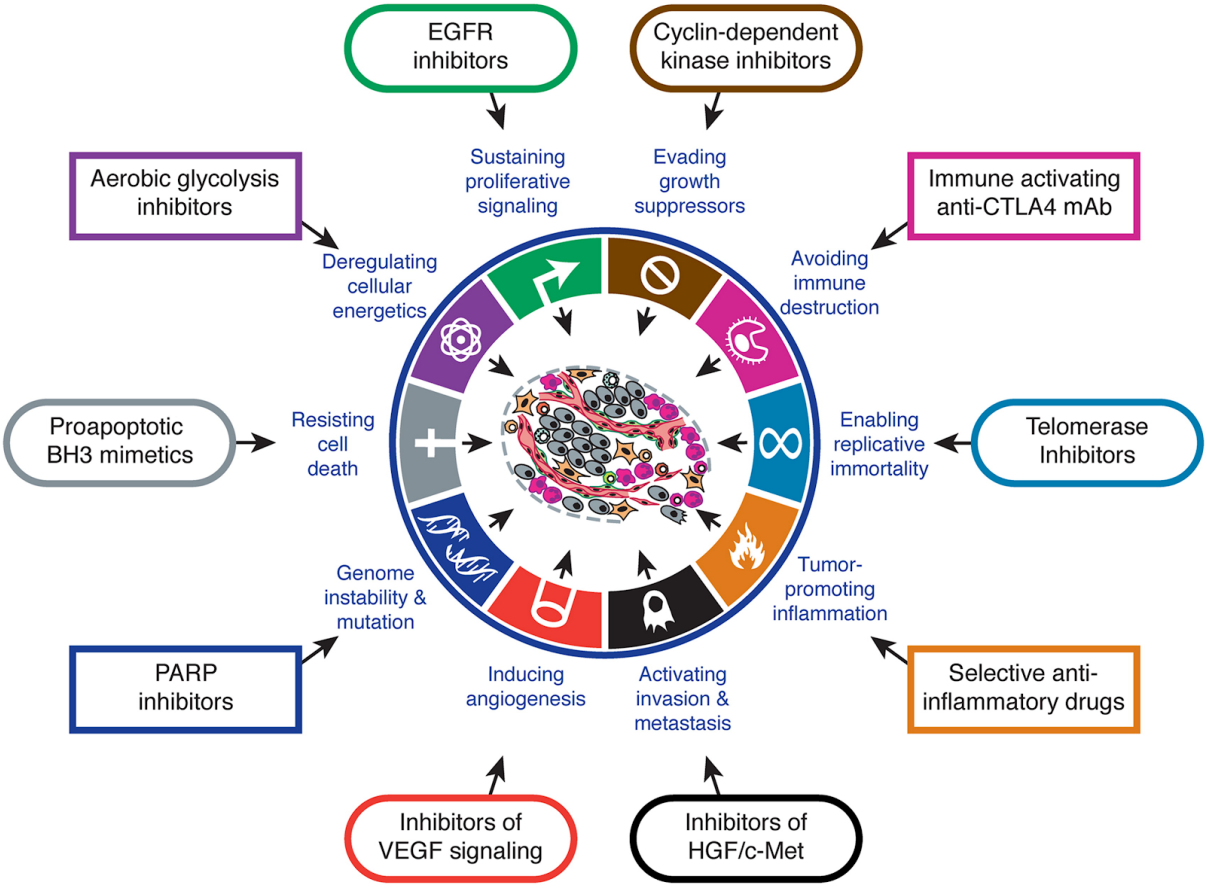


Figure 4. Therapeutic Targeting of the Hallmarks of Cancer (Figure taken from Hanahan & Weinberg[15]) Drug classes are matched with hallmarks they interfere with.

1.2.3 Resistance to treatment

Resistance represents major obstacles to successful cancer therapy. Resistance occurs to both chemotherapies and to targeted therapies.[16,17]

Drug resistance can be divided into two categories -- intrinsic or acquired. Intrinsic resistance indicates that before receiving drug therapy, resistance-mediating factors pre-exist in the tumour cells which makes the therapy ineffective. Acquired drug resistance can develop during treatment of tumours that were initially sensitive and can be caused by mutations arising during treatment, or through various other adaptive responses, such as increased expression of the drug target and activation of alternative compensatory signalling pathways.[16]

Mechanisms of resistance to cytotoxic chemotherapies and targeted drugs largely overlap.

Holohan et al.[16] classifies all mechanisms implicated in resistance into 6 groups: (1) drug transport and metabolism which includes drug efflux and drug inactivation/lack of activation, (2) alterations in drug targets, which is a major cause of resistance for targeted therapies; alteration can be in a form of mutation that alters drug-target binding site in the protein or in a form of protein overexpression, (3) DNA damage repair which is a typical mechanism of resistance to DNA-damaging agents, (4) downstream resistance mechanisms such as deregulation of apoptosis or

autophagy, (5) resistance-promoting adaptive responses, this group of mechanisms include activation of prosurvival signalling, oncogenic bypass and pathway redundancy and epithelial-mesenchymal transition, (6) tumor microenvironment protection via integrins, cytokines and growth factors.

In addition to these mechanisms intra-tumour heterogeneity can also contribute to development of resistance -- since tumor can have different cell subpopulations with distinct genomic alterations (i.e. some fraction of cells can be sensitive to treatment while another fraction can be resistant) drug resistance can arise through therapy-induced selection of a resistant minor subpopulation of cells that were present in the original tumor.[16,18]

The use of high-throughput sensitivity and molecular screening techniques can help to identify resistance mechanisms and allow patient stratification with respect to the treatment response i.e. predict events of sensitivity or resistance. Let's review different types of drug sensitivity testing in the next section.

1.3 Drug sensitivity testing

Clinical response to anticancer therapeutics is heterogeneous, which is a major barrier to effective cancer care. An ability to more accurately predict response before choice of treatment would improve patient response rates and reduce unnecessary treatments.
In order to predict treatment response it's necessary to combine two kinds of data -- tumor molecular information and drug response information. Since it's difficult to generate this data for a large number of patient, alternative model systems are used for this end (Fig. 5).

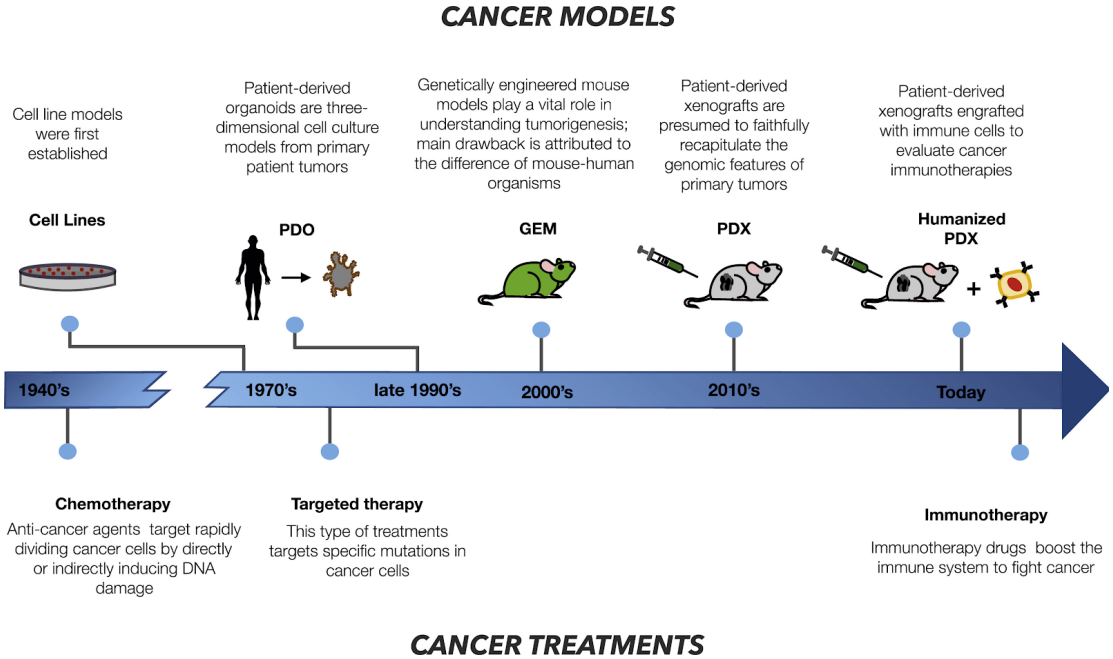


Figure 5. Cancer models. "Figure shows schematically the timeline of development of cancer models and cancer therapies to address the challenges in personalized medicine" (Figure taken from Kalamara et al.[19])

1.3.1 Cancer preclinical models

Model systems for studying drug response can be broadly divided into *in vitro*, *in vivo* and *ex vivo* types.

In vitro. Cancer cell lines are the most popular amongst all preclinical cancer models. They can be relatively easily grown in the laboratory and therefore can be subject to multiple experiments, including multi-omics molecular characterization as well as treatment with many drugs or drug combinations. Connection of molecular characterization and treatment provide data for studying pharmacogenomic associations. A major drawback of cell line models is the lack of tumor microenvironment and of intrinsic heterogeneity compared to the original tumor.[20,19]

Patient-derived organoids (PDO) or organotypic cultures, three-dimensional cell cultures derived from a patient's tumor are considered to be a better *in vitro* model.[21,22] They more accurately represent the intrinsic environment of the original tumor since they can include multiple cell types and self-organize into tissue-like structures. Drug response-genomic associations in organoids are more similar to those of real tumors compared to cell lines. A main drawback of PDO models is the difficulty to maintain them in long-term cultures.[19]

In vivo. Mouse-based *in vivo* models are a valuable tool for preclinical evaluation of novel therapeutic strategies in cancer. Patient-derived tumor xenograft (PDX) models[9,23] are obtained by direct implants of patient's tumor cells or tissue fragments in immunodeficient mice. PDX models can recapitulate the heterogeneity and intrinsic drug sensitivity of the primary tumor. However, they are a limited model of tumors *in vivo*, in particular of the interaction of the tumor with the immune system.[19]

Another type of mouse models is genetically engineered mouse (GEM) model. GEM is a mouse whose genome has been edited by genetic engineering techniques to initiate tumorigenesis. GEM models harbor significant genetic heterogeneity although they do not reflect the complex heterogeneity of a human tumor. The difference between mouse and human organism is the main obstacle for immuno-oncology drug discovery studies. This gap is filled by “humanized” PDX models which are additionally engrafted with human immune cells.[19] Main challenges for both PDX and GEM are time required to generate tumor material and test the treatment regimen which can take 4-8 month, variability in engraftment rate, and higher costs compared to the other model systems.[22,24]

Ex vivo. Ex vivo model systems involve taking a sample out of the organism or patient and studying it under more controlled conditions than in vivo. Obtained samples are typically not cultured for long periods, so they still retain the original characteristics.[19] This approach is becoming routine in investigations of basic biological mechanisms in haematological malignancies and in drug discovery.[25,26]

1.3.2 Cell line drug screening

The majority of pharmacogenomics studies performed up to date is based on cancer cell line drug screens. Cell line drug screening is the process of screening anti-cancer compounds against a panel of cell lines. Typically a single experiment (one drug-one cell line) includes several probes for testing different concentrations of a drug, in each probe cell line is incubated with a drug and then cell viability (i.e. amount of survived cells) is assessed via viability assays that measure either metabolic (i.e. number of ATP molecules) or DNA content.[27]

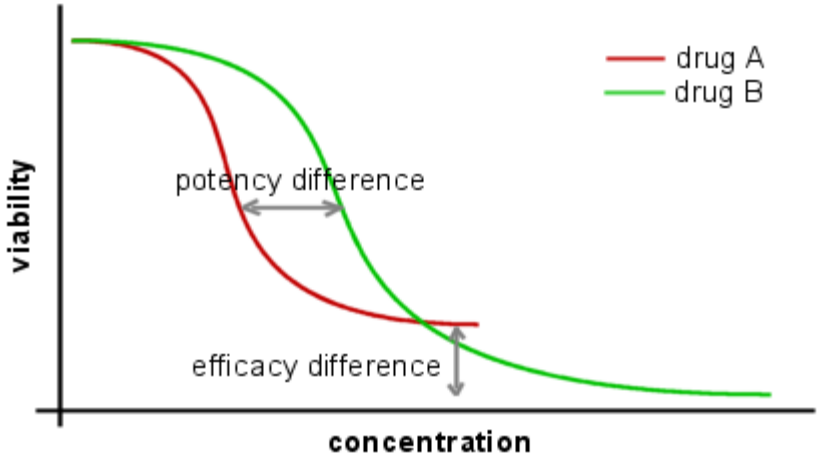


Figure 6. Exemplary dose-response curves from two drugs that have different potency and efficacy.

Information from dose-response curves (Fig. 6) can be summarised using standard drug response metrics as half-maximal inhibitory concentration (IC_{50}) which is a measure of drug potency and area under the dose–response curve (AUC) that takes into account drug potency and efficacy.

Although cell lines is the most widely used model in pharmacogenomics it's important to acknowledge that there are differences between cell lines and real tumors. Cell lines acquire molecular changes in the culture (genetic, expression changes), they no longer retain tumor heterogeneity, they have higher growth rates, unlike real tumors cell lines are 2D and lack tumor microenvironment.[22,28]

1.4 Main pharmacogenomics datasets

Several large-scale datasets have been generated in order to link genomic and pharmacologic profiles of cell lines (Fig. 7). The first one was NCI-60 panel,[29] established in the late 1980s, which utilized 60 cancer cell lines and aimed to identify and characterize novel compounds with tumor-killing properties. These cell lines were molecularly profiled to identify biomarkers of response, providing the first resource for cancer pharmacogenomics. Since then, cell line screening has become a popular platform for cancer research and screens with larger number of cell lines followed.

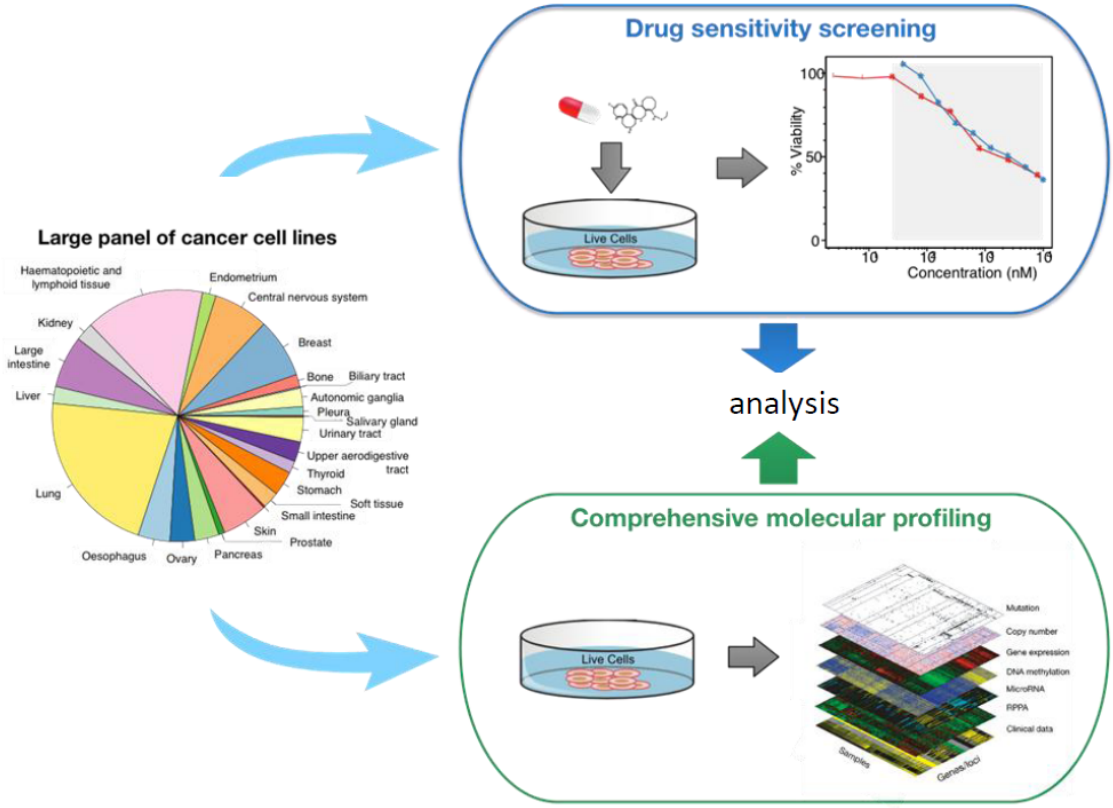


Figure 7. Organisation of large scale pharmacogenomics profiling. Typical pharmacogenomics study consists of two parts -- cell line molecular profiling and drug sensitivity profiling. (Figure taken from Haibe-Kains[30])

Recently, data on several large-scale pharmacogenomics became publicly available (see Table 3).

Table 3. List of pharmacogenomics datasets with corresponding cell lines and drug numbers.

Dataset	# of cell lines	# of drugs
CCLE[31]	505	24
CTRP[32]	860	481
GDSC[34]	1000	265
gCSI[36]	410	16
NIBR PDXE[37]	23-38 xenografts per drug	62

In the Cancer Cell Line Encyclopedia (CCLE)[31] project ~1000 human cancer cell lines were collected and molecularly characterized, including gene expression, copy number alteration, and somatic mutation profiling. The project released the

drug sensitivity profiles of 24 anticancer drugs across 504 cell lines. Analogously to the CCLE project, the Cancer Therapeutics Response Portal (CTRP)[32] screened around 500 compounds and some of their combinations on the CCLE cell lines.[33]
The Genomics of Drug Sensitivity in Cancer (GDSC)[34] project profiled the sensitivity of about 1000 cell lines to 250 compounds and identified many genetic alterations associated with drug efficacy. The project investigators also compared genomic information on 11 289 tumors from TCGA with profiled cancer cell lines and identified shared 'cancer functional events' (CFEs), key molecular aberrations, shared between cell line and tumor data. Drug response information from cell lines with similar CFEs to the tumor of interest were used to predict the response to a given drug therapy. Mutation and copy number information were the most predictive of drug response in specific tissue types while gene expression was most informative for pan-cancer predictions.[35]
The Genentech Cell Line Screening Initiative (gCSI)[36] screened 16 compounds on 410 cell lines, these compounds and some of the cell lines were previously tested by CCLE and GDSC projects.
As we discussed, mouse models can better recapitulate the drug response of real tumors. However due to higher costs and more laborious procedures large-scale pharmacogenomics studies performed on mice are less abundant. Novartis Institutes of Biomedical Research PDX Encyclopedia (NIBR PDXE)[37] established ~1000 patient-derived tumor xenograft (PDX) models representing 16 different cancer types with a diverse set of mutation profiles. Around 200 of these PDX models were screened *in vivo* to assess the responses to 62 compounds.

Altogether these pharmacogenomics studies provided rich resources for improving our understanding of cellular response to drugs, and generated data allowed to develop prediction algorithms that match drug response with genomic features. Let's review various modelling approaches that have been applied to pharmacogenomics data in the next section.

1.5 Existing modelling approaches and DREAM challenge

1.5.1 Existing modelling approaches

Various machine learning approaches can be used for drug response prediction, ranging from linear models that have advantage of good interpretability to non-linear models that usually show better performance but worse interpretability.

Depending on outcome variable all models can be divided into two classes -- regression and classification models. Regression models predict continuous outcome e.g. IC_{50} or AUC numerical values, classification models predict categorical outcome, which is often binary in drug response prediction problems e.g. "sensitivity" vs. "resistance".

Typically drug-response prediction problems have a "bottleneck of data dimensionality"[33] -- molecular data have high-dimensionality (just one molecular layer gives ~20000 of features), while number of samples that have drug response information is relatively small (100-1000 samples). Therefore quite often modelling approaches include certain feature selection (dimensionality reduction) strategies. Feature selection can be done prior to model fitting e.g. by selecting features that correlated with outcome, selecting features with high variance or selecting features using prior literature knowledge. Also feature selection can be a part of model fitting process e.g. as in case of regularized methods (see section 1.7).

A number of different machine learning methods has been applied to drug response prediction problem (see Table 4), the most widely used were regularized regression methods (lasso, elastic net, ridge regression), kernel-based methods (e.g. SVM) and ensemble methods (e.g. Random Forest).

Table 4. Spectrum of machine learning methods applied to drug response prediction problem. Adapted from Ali & Aittokallio.[38]

Class	Method	Example applications
Regularized linear regression	Lasso regression	Fang et al.[39] applied lasso regression to CCLE data using iterative approach for feature selection.
	Ridge regression	"Geeleher et al. (2014)[40] and Geeleher et al. (2017)[41] applied ridge regression model to predict drug responses in GDSC cell lines, and inferred marker panels for predicting comprehensive drug response profiles in patient tumors in the TCGA dataset."
	Elastic net regression	"Jang et al.[42] found elastic net regression as one of the best-performing modeling strategies for drug response prediction in CCLE and GDSC cancer cell lines." Falgreen et al.[43] used elastic net to predict resistance in diffuse large B-cell lymphoma patients treated with cyclophosphamide, doxorubicin and vincristine. Elastic net regression was used in a number of other studies -- Barretina et al.[31], Iorio et al.[34], Aben at al.[44]
Kernel-based	SVM (support vector machines)	Dong et al.[45] used SVM classification model to predict drug sensitivity for several drugs using baseline gene expression of cell line panels from CCLE and GDSC studies. Other applications include Jang et al.[42] and Hejase & Chan.[46]
	BEMKL (Bayesian	Kernelized regression model for drug response prediction based on data integration across

	efficient multiple kernel learning)	multiple omics profiles, through multi-task, multiple kernel learning was the best performing method in the DREAM challenge (Costello et al.[47]) In a follow-up work by Ali et al.[48] this method was applied to NCI-60 cell line panel.
	cwKBMF (component-wise kernelized Bayesian matrix factorization)	Ammad-ud-din et al.[49,50] proposed a model that utilizes cell line information along with the drug chemical properties as an additional information source through selective data integration. Model was applied to GDSC and CTRP cancer cell line data and to in-house AML cell lines data.[38]
Ensemble	Random Forest	"Riddick et al.[51] built an ensemble regression model using random forest (RF) for drug sensitivity prediction in NCI-60 cell line panel. The model was also used to create drug-specific gene expression signatures and identify core cell lines associated with each drug's response." Other applications of RF include, e.g., Iorio et al.[34], Nguyen et al.[52], and Rahman et al.[53]
Neural networks	"Classical" Neural Networks	Menden et al.[54] used a neural networks algorithm to train models that co-utilise cell line genomic and drug physicochemical and structural features.
	Deep learning	Recently a number of approaches based on deep learning was proposed e.g. Ding et al.[55], Chang et al.[56]

1.5.2 DREAM challenge

Despite the big number of various approaches applied to drug response prediction problem, different approaches are rarely being compared in terms of prediction accuracy in a systematic manner. The DREAM challenge[47] organized by consortium of Dialogue on Reverse Engineering Assessment and Methods (DREAM) provided an opportunity for such systematic comparison.

The task of the challenge was to predict drug response in the panel of breast cancer cell lines (18 lines were in the test set, and drug response data for 35 lines was available for training) profiled with 28 drugs. Available molecular data included genomics, transcriptomics, methylation and protein information. In this challenge 44 modelling approaches were evaluated on their performance.

This DREAM challenge reported a number of useful observations[33]:
(1) all top solutions utilized nonlinear methods
(2) predictive performance benefited from prior knowledge of biological pathways
(3) gene expression data provided the highest predictive performance among all molecular data types; also performance could be further improved by including other data types
(4) integrating predictions from independent methods (via an ensemble model consisted of different independent models) produced the most robust results since different methods had complementary advantages in examining different aspects of the data.

1.6 Problem of inconsistency between pharmacogenomics datasets

One of the common strategies for evaluating accuracy of drug response models is to fit a model using data from one dataset, and then assess model accuracy on another dataset. When Papillon-Cavanagh and his colleagues[57] tried to use this strategy, namely training models on GDSC and then validating them on CCLE, they realized that the validation in terms of model accuracy worked only for a small subset of drugs that are in common between two datasets.

Subsequent investigation performed by Haibe-Kains et al.[58] discovered inconsistencies between drug response data published in GDSC and CCLE studies. These inconsistencies had negative impact on the development of drug response models.[59] The study triggered a number of subsequent efforts to assess the consistency between the two pharmacogenomics datasets which resulted in the number of papers arguing either pro or against consistency.[60-64] These studies suggest different methodological strategies one can use to assess the consistency and discussed possible sources of discrepancies in the data.[59,65] Also this discussion facilitated the development of resources that provide access to many different cell line pharmacogenomics datasets in a unified manner.[66,67]

Meanwhile Genentech published a comparative study[36] using their own data (gCSI dataset) as reference and observed that their drug sensitivity data was more similar to CCLE (which used the same pharmacological assay) than to GDSC. This study evaluated different aspects of the screening protocols that are relevant for measuring drug response, including the readout of cell viability, seeding density, and cell culture media conditions. The authors discovered that differences in media conditions and seeding density contributed to inconsistencies between studies, and that metabolic and DNA-content drug sensitivity assays exhibit different levels of noise.[59]

Currently the consensus is that molecular profiles (e.g. expression profiles) of CCLE and GDSC show consistency while inconsistency is observed for drug sensitivity measures between two datasets. The main sources of this inconsistency

are intrinsic noise of pharmacological profiling, differences in experimental protocols and differences in curve fitting. The standardisation on the level of experimental procedures as well as in data analysis will help to reduce inconsistency in the future studies which should improve identification of robust drug response biomarkers and developing accurate drug response models.[59,65]

1.7 Basic machine learning concepts, Feature Selection and description of used methods (elastic net, SVM, Random Forest)

1.7.1 Basics

Machine learning, a collection of data-analytical methods aimed at building predictive models from multi-dimensional datasets, plays an increasingly important role in modern biological research.[68] Let's now review basic concepts and principles of machine learning.

All machine learning methods can be broadly divided into two classes -- supervised and unsupervised methods. The main distinction is that supervised methods deal with labelled samples in a way that they learn how to predict correct label from the sample's features (Fig. 8) while unsupervised methods analyse samples without taking label information into account. Although we mainly focus here on supervised methods (since our main interest to predict a certain kind of labels -- drug response) unsupervised learning, including standard clustering techniques, can provide some basis for generating prediction models, particularly unsupervised methods are helpful in data selection and visualization.[69]

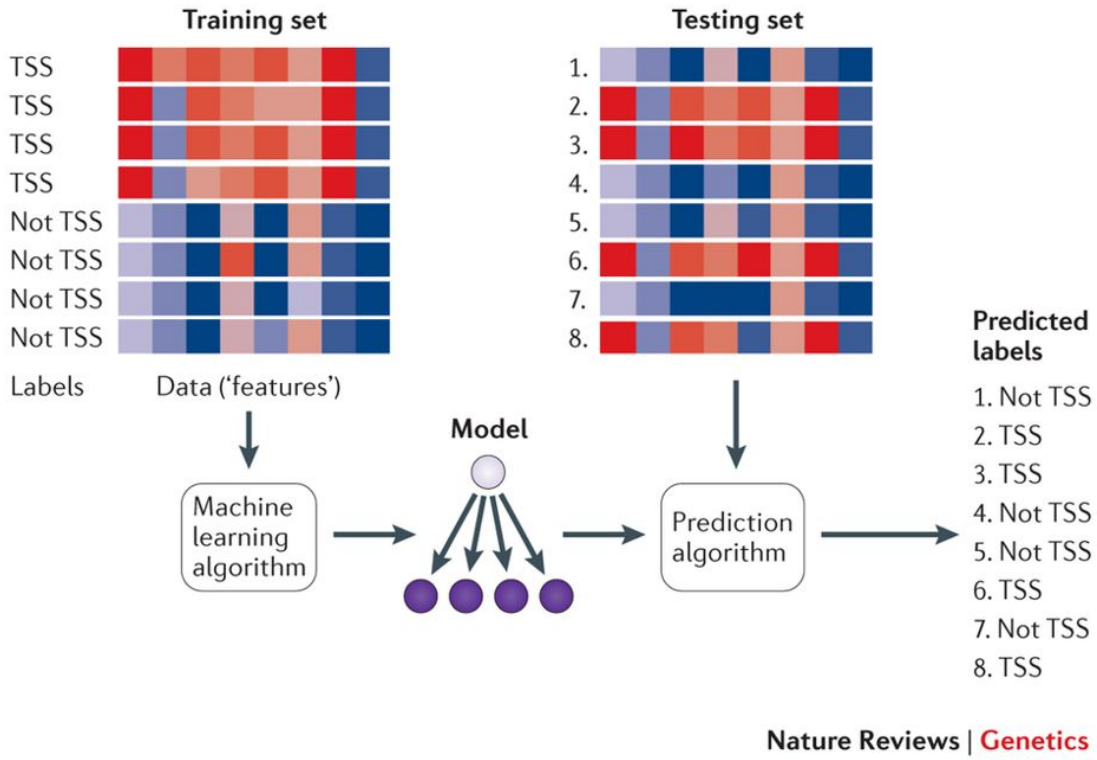


Figure 8. A canonical example of a machine learning application. "A training set of DNA sequences is provided as input to a learning procedure, along with binary labels indicating whether each sequence is centred on a transcription start site (TSS) or not. The learning algorithm produces a model that can then be subsequently used, in conjunction with a prediction algorithm, to assign predicted labels (such as 'TSS' or 'not TSS') to unlabelled test sequences. In the figure, the red–blue gradient might represent, for example, the scores of various motif models (one per column) against the DNA sequence." (Figure taken from Libbrecht & Noble[70])

Supervised problems can be separated into classification and regression problems. When labels are categorical the problem is a classification one, labels on continuous scale constitute a regression problem.

Let's consider typical steps of supervised machine learning (ML) workflow.

1. Getting data and preprocessing. We usually get data as a matrix (see Fig. 8) where each row contains data for certain sample and each column contains data for certain feature (or vice versa). In biological problems features can include one or several types of molecular data e.g. expression data, copy number data, methylation, mutation statuses for a number of genes. One of the columns typically contain labels. Preprocessing can include necessary transformation of features and dealing with missing data (e.g imputation of missing values or exclusion of samples that contain missing values). Then a subset of data, training set, is used in the next step.
2. Training a model. During training (or "fitting") process ML algorithm is finding the optimal set of model parameters that translate features in the training set into accurate predictions of the labels.

3. Testing a model. When model is ready we can apply it to test data (part of data that was not used for training) and then by comparing predicted labels with true labels we can estimate accuracy of our model.

An essential goal of model training is to build a model that is generalizable beyond the data used for fitting the model, i.e. a model that can make accurate predictions on a new input data. That's why it's important to divide the full data available for training into two sets -- training and test sets (some strategies involve division into three sets -- training, test, and validation). This division allows to assess generalizability of the model fitted on training set by checking its accuracy on test set. In cases when model explains the data in training set accurately but accuracy on test set is substantially lower one may conclude that overfitting has occurred, which means that model is "over-fitted" to the data in training set and is not generalizable. Therefore division into 2 or 3 sets gives a way to select more accurate model and estimate its accuracy in unbiased way.[69]

1.7.2 Feature selection

As we have already discussed in the section 1.5, data used for modelling often has a number of features that by several orders of magnitude higher that number of samples. High dimensionality of the data can contribute to the problem of overfitting and often increases computational time for model fitting. There are several methods to perform a feature selection from high-dimensional data which can be divided into three classes: filter, wrapper and embedded methods.

Filter methods evaluate the relevance of the predictors prior to model fitting procedure. Features are evaluated individually i.e. in a univariate manner on the basis of association with outcome. For example in regression problems, one can assess correlation between individual features and outcome and select for modelling only those features that have a correlation coefficient higher than certain threshold.[71]

Wrapper methods compare multiple models using procedures that add and/or remove predictors to find the optimal combination that results in maximal model performance. Essentially, wrapper methods are search algorithms that use different sets of predictors as the inputs and utilize model performance as the output to be optimized. Examples of wrapper methods include recursive feature elimination, genetic algorithms, and simulated annealing.[71]

Embedded methods are group of techniques that perform model selection as a part model construction process. Thus in these methods predictor search algorithm is coupled with a parameter estimation and they are optimized using a single objective function.[71] We will consider one of the most common type of embedded methods, regularized regression, in the next subsection.

Throughout the work described in the thesis we mainly used filter-based feature selection i.e. we assessed relationship between an outcome and each feature independently. For this assessment two functions from caret package[75] were used: anovaScores function for ranking features in classification tasks, and gamScores

function for ranking features in regression tasks. Function anovaScores treats the outcome as the independent variable and the predictor as the outcome. In this way, the null hypothesis is that the mean predictor values are equal across the different classes. For regression, gamScores fits a smoothing spline in the predictor to the outcome using a generalized additive model and tests to see if there is any functional relationship between the two. In each function the p-value is used as the score.[76]

1.7.3 ML Methods

Here we will describe the main ideas behind the three groups of machine learning algorithms that we used throughout the study: regularized regression, support vector machines and random forest.

Regularized regression

First let's introduce linear regression which provides a foundation for regularised regression. In linear regression approach we model the relationship between the variable we want to explain/predict (dependent variable), and the variables we want to use for prediction (independent variables also called covariates) in the following way:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \; ; \; \varepsilon \sim N(0, \sigma^2) \qquad (1)$$

y_i is the i^{th} value of outcome variable, x_{ij} is the i^{th} value of j^{th} predictor (x_j), β_j is the regression coefficient for predictor x_j, n is the total number of samples and p is the total number of predictors. ε is error which is normally distributed with the 0 mean and variance σ^2.

In this way we would like to find such values for coefficients β_j that our outcome variable y_i is equal to the linear combination of independent variables $\sum_{j=1}^{p} \beta_j x_{ij}$. In other words we would like to optimize (minimize) our objective function, the sum of squared errors (SSE):

$$SSE = \sum_{i=i}^{n} \left| y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right|^2 \qquad (2)$$

Regularized regression optimizes a sum of the linear regression objective function (sum of the squared errors, SSE, Equation 2) and convex penalty terms on coefficients. These penalties help to find coefficients of the optimal solution in high-dimensional space while preventing the regression procedure from overfitting the training data.[72]

One common penalty, called L1 or LASSO (least absolute shrinkage and selection operator) shrinkage, limits the sum of absolute values of all coefficients (Equation 3). LASSO regression achieves feature selection by setting most coefficients to zero and leaving the coefficients of essential variables as the only nonzero coefficients.[72]

$$SSE_{L1} = \sum_{i=i}^{n} \left| y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right|^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \quad (3)$$

here λ is a parameter that controls strength of the L1 penalty.

Another common penalty, called L2 or ridge shrinkage, limits the sum of squares of all coefficients (Equation 4). Ridge regression assigns nonzero coefficients to most variables and therefore does not perform feature selection. Ridge regression can achieve better prediction accuracy than LASSO when the features are highly collinear.[72]

$$SSE_{L2} = \sum_{i=i}^{n} \left| y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right|^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (4)$$

here λ is a parameter that controls strength of the L2 penalty.

Elastic net regression combines the advantage of LASSO and ridge regressions. Elastic net optimizes the sum of the objective function and the two penalties (Equation 5). The penalty weights (λ_1 and λ_2) can be selected via cross-validation procedure.[72]

$$SSE_{Enet} = \sum_{i=i}^{n} \left| y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right|^2 + \lambda_1 \sum_{j=1}^{p} \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \quad (5)$$

here λ_1 and λ_2 are parameters that control strength of the L1 and L2 penalties respectively.

Support Vector Machines (SVM)

The SVM algorithm is based on finding the hyperplane in high-dimensional space that maximizes the margin between classes in the training data (Fig. 9). Selecting hyperplane with the largest margin between classes maximizes the SVM's ability to predict the correct classification of previously unseen examples. The training examples that are closest to the hyperplane are called support vectors since they are supporting the margin.[73,74]

Computing SVM classifier means solving the following optimization problem ("soft margin" formulation):

$$min_{w,b}\left(\frac{1}{2}|w|^2 + C\sum_i \xi_i\right) \quad (6)$$

$$\text{where } \xi_i = (1 - f(x_i)y_i)_+$$

Here **w** is a weight vector (normal vector to hyperplane), $f(x_i)$ is the class prediction for a data point i and y_i is a class of point i (1 or -1). ξ is a "hinge loss" function which penalizes points whose functional margin, $f(x_i)y_i$, is smaller than 1. ξ is always non-negative. Since a distance between a point and hyperplane is inversely proportional to |w|, $d = \frac{f(x)}{|w|}$, by minimizing |w| we maximize separation between points from two classes; term $C\sum_i \xi_i$ penalizes those data points that are close to a separating hyperplane.

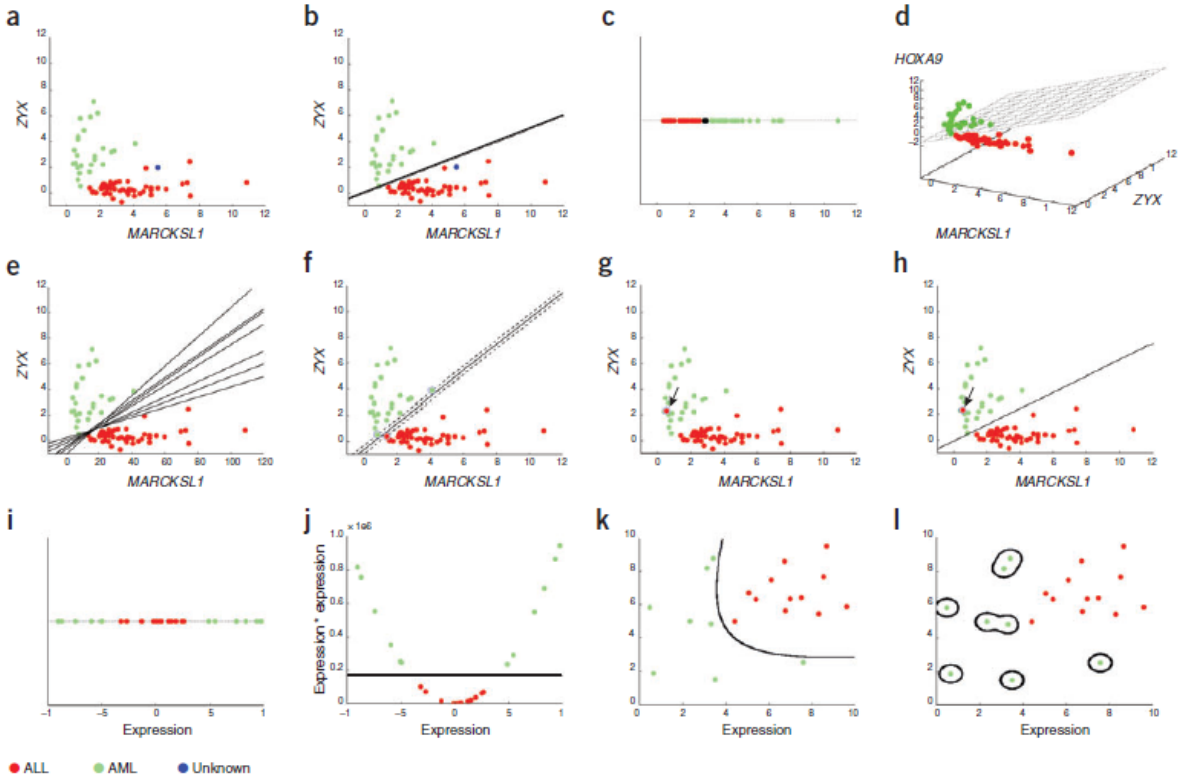


Figure 9. Main concepts of SVMs "(**a**) Two-dimensional expression profiles of lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) samples. Each dimension corresponds to the measured mRNA expression level of a given gene. The SVM's task is to assign a label to the gene expression profile labeled 'Unknown'. (**b**) A separating hyperplane. Based upon this hyperplane, the inferred label of the 'Unknown' expression profile is 'ALL'. (**c**) A hyperplane in one dimension. The hyperplane is shown as a single black point. (**d**) A hyperplane in three dimensions. (**e**) Many possible separating hyperplanes. (**f**) The maximum-margin hyperplane. The three support vectors are circled. (**g**) A data set containing one error, indicated by arrow. (**h**) A separating hyperplane with a soft margin. Error is indicated by arrow. (**i**) A nonseparable one-dimensional data set. (**j**) Separating previously nonseparable data. (**k**) A linearly nonseparable two-dimensional data set, which is linearly separable in four dimensions. (**l**) An SVM that has overfit a two-dimensional data set." (Figure taken from Noble[74])

An important concept to mention when discussing SVM method is kernel function (or kernel). Essentially kernel function can be seen as a similarity function that allows the SVM to perform classification in the two-dimensional space even when the data is one-dimensional (see Fig. 9 i,j). In general, kernel function projects the data from a low-dimensional space to a space of higher dimension. If one pick a good kernel function, then the data can become separable in the resulting higher dimensional space, even if it wasn't separable in the lower dimensional space.[74]

Random Forest (RF)

Random forest is an ensemble method which utilizes many decision trees as individual learners. The idea of RF is based on bagging, an approach in which each learner is trained on a different bootstrap to increase their variation.[72]

Let's consider the steps of random forest algorithm (given a dataset containing N samples and M features; see Fig. 10):

1. Create n bootstrap samples (subsamples) from the original data. Typically n can range from 100 to several thousands.
2. For each bootstrap sample, train a decision tree using m features (m is typically much smaller than M) at each node of the tree. The m features are selected randomly from the M features in the dataset and the decision tree will select the best split among the m features.
3. A new test sample is classified by all the trees and the final decision is done by majority vote/averaging.

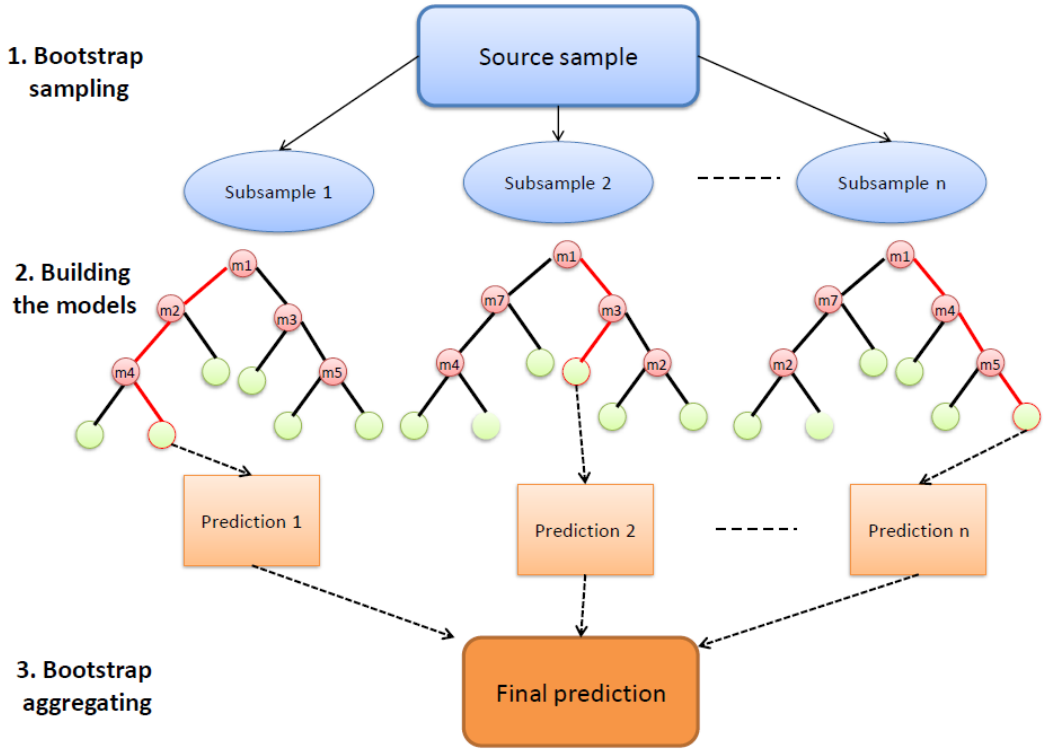


Figure 10. Steps of Random Forest algorithm. 1. Random sampling. 2.Building the models. 3. Bootstrap aggregating / majority vote.

1.8 Aims of thesis and thesis' structure

Computational prediction of drug response in cancer is a challenging problem. Despite the biological challenge of predicting complex cellular phenotype (which drug response is) using typically static multi-omics data (i.e. data that acquired before drug exposure) there are also data analysis challenges arising from complexity of the available data in terms of volume, noise and heterogeneity.[69] Growing amount of pharmacogenomics data together with challenges gives us an opportunity to learn principles of creating accurate drug response prediction models, knowledge which is important in the emerging era of personalized medicine.

The main goal of this thesis was to analyse the data from largest public pharmacogenomics screens in order to learn factors that determine accuracy of drug response prediction. This broad goal can be formulated as a set of more concrete study aims:

1. Assess ways for improving consistency between independent pharmacogenomics datasets.
2. Assess ways for increasing accuracy of drug prediction using some modifications over standard ML approaches and feature engineering.
3. Analyse influence of ML algorithm selection on resulting prediction accuracy.
4. Study how properties of training set influence prediction accuracy. Properties include: type of molecular data, number of features selected for modelling, number of samples, type of drug response metric.
5. Compare accuracy of drug response prediction between cell lines and xenografts.
6. Assess the feasibility of using models trained on cell line data for drug response prediction in xenografts and patients.

Also in the process of working on main aims a number of auxiliary, more practical aims arose through collaborations:

1. Analyse a drug-sensitivity screen performed on a panel of Burkitt cell lines.
2. Analyse drug-sensitivity data on DKFZ-608 compound and create a model of drug response to the compound.
3. Create an interactive visualisation that allow to see a group of samples from CTRP/GDSC screens on 2-dimensional plane, where x and y axes display drug response values for 2 different drugs of choice.

The following chapters of the thesis present results accordingly to described study aims. Each chapter/analysis focuses on a certain problem or aspect of drug response modelling, see Fig. 11.

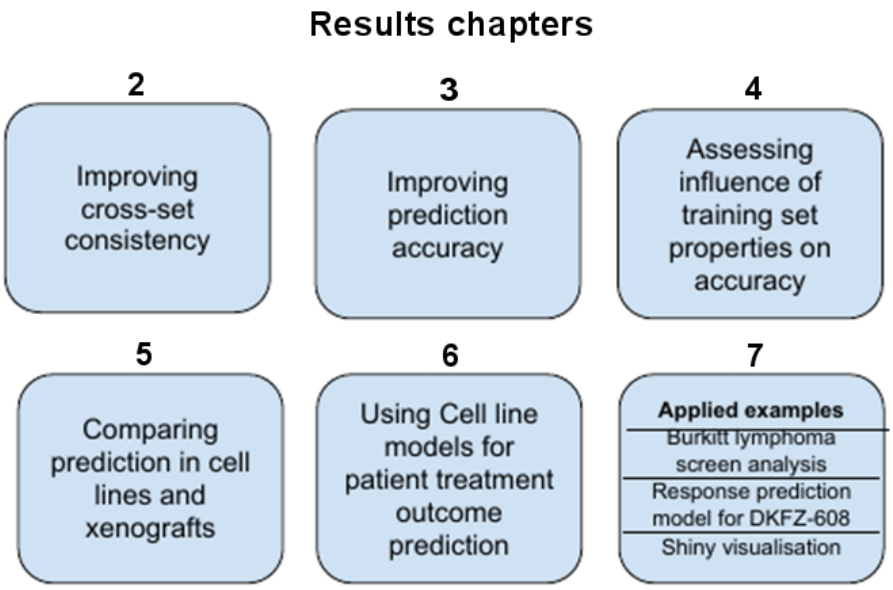


Figure 11. Sections of the Results chapter.

In the first section I touch the problem of cross-set inconsistency and try to improve consistency of drug response data coming from different cell line datasets. In the second section we assess a number of methodological improvements (multi-task modelling instead of general single-task, combining training data for drugs that share the same target, addressing class imbalance) with respect to their ability to improve accuracy of prediction. In the third section we again try to learn which aspects of modelling affect the accuracy of prediction but this time we rather focus on the properties of training data i.e. sample size, type of molecular data, drug response metric etc. In the fourth section in addition to cell line we take into account xenograft data and compare models trained for the same drugs on cell line and on xenograft data. In the fifth section we assess our ability to predict treatment outcome in a number of patient cohorts using classification models trained on cell line data. Finally in the sixth section I will present two collaboration-studies in which we utilized drug response data analysis or modelling, and one interactive drug-response data visualisation example that we created.

Results described in the chapters 4 and 5 are part of our manuscript “Drug response prediction in cell lines and xenografts” by Kurilov R, Haibe-Kains B and Brors B.[77] Results described in the section 7.3.1 of the chapter 7 are part of the paper “Drug-based perturbation screen uncovers synergistic drug combinations in Burkitt lymphoma” by Tomska K, Kurilov R. et al.[78]

2 METHODS TO IMPROVE CROSS-SET CONSISTENCY

2.1 Introduction

As it was discussed in the general introduction (section 1.6) there is an observed inconsistency between pharmacogenomic data from different screens which presents an obstacle on the way of producing accurate models and/or identifying biomarkers of response using this data. Here we make an attempt to improve an agreement between pharmacogenomic data coming from different datasets. In the first part or our analysis (sections 2.2.1 and 2.3.1, drug response consistency) we focus on the agreement between drug sensitivity data alone and in the second part (sections 2.2.2 and 2.3.2, biomarkers' consistency) we focus on the agreement between genomic-drug sensitivity data associations.

2.2 Data and Methods

2.2.1 Drug response consistency

We focused on agreement between CTRP and GDSC datastets. Particularly we tried to improve consistency between two datasets for 19 drugs belonging to 6 classes defined by drug's target molecule (Table 5). Cell line genomics and drug response data were obtained via the PharmacoGx package (version 1.8.3).[66] As a drug response metric we used area under the dose-response curve, AUC (see Fig. 20).

Table 5. Drug targets with associated drugs.

Drug target	Common drugs (GDSC-CTRP)
HDAC	Tubastatin A, Belinostat, Vorinostat, MS-275 (entinostat)
EGFR	Lapatinib, Erlotinib, Afatinib, Gefitinib
MEK1, MEK2 (MAP2K1, MAP2K2)	Trametinib, Selumetinib
BRAF	PLX4720, Dabrafenib
HSP90	SNX-2112, 17-AAG (tanespimycin)
mTOR	OSI-027, Rapamycin (sirolimus), BEZ235 (NVP-BEZ235), Temsirolimus, AZD803

We tested two approaches for improving the cross-set agreement:
1) **Cell line filtering.** We Identify cell lines whose drug response for certain drug is different from average drug response for drugs from the same class (drugs with the same target) within dataset, remove these cell lines within each dataset, and then compare drug response consistency for this drug between subsetted datasets.

2) **GLDS correction.** We correct drug response for general level of drug sensitivity (GLDS) using for correction drug response data from group of unrelated drugs (specific for a certain drug) and then compare drug response consistency for this drug between corrected datasets. The idea and motivation for GLDS correction is described in the study Geeleher et al.[79]

The algorithms for both methods:
1) Cell line filtering
Let's assume that we have 3 drugs that inhibit the same molecular target, drug "1", drug "2" and drug "3", and we have drug response values for these drugs in two datasets, GDSC and CTRP (see Fig. 12). We also assume that the cell lines tested in both datasets are the same.

1. calculate Spearman correlation between each pair of drugs (i.e. between drug "1" from GDSC and drug "1" from CTRP, between drug "2" from GDSC and drug "2" from CTRP etc.) before filtering
2. scale each column (i.e. divide by column's standard deviation)
3. calculate a column with mean values (across drugs "1", "2", "3") for each dataset
4. calculate difference between each drug column and mean column (in both datasets)
5. from each column remove 10% of cell lines with highest difference
6. calculate Spearman correlation between each pair of drugs after removal of cell lines (i.e. after filtering)
7. calculate average correlation after 10 random removal of 10% (from original set)

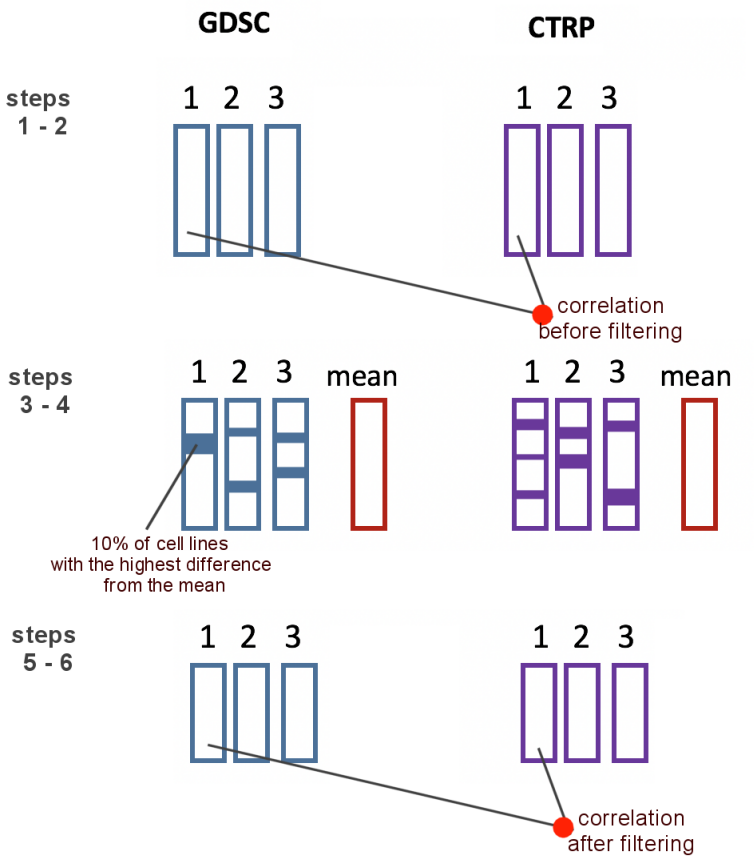


Figure 12. Steps of the cell line filtering method. Rectangle bars represent vectors of cell lines' drug response values.

2) GLDS correction (Fig. 13a)

1. for each drug within each dataset we defined a set of unrelated drugs based on drug response correlation (Spearman correlation <0.15) and drug class annotation.
2. calculate correlation between a pair of drugs before correction
3. subtract from each drug's column mean of all columns from unrelated drugs
4. calculate correlation after correction

2.2.2 Biomarkers' consistency

Here we focused on biomarkers' consistency in two groups: 1) between GDSC, CTRP and NIBR PDXE[37] and 2) between GDSC, CTRP and gCSI, before and after GLDS correction.

In the first group there were 6 drugs in overlap between 3 studies: 5-Fluorouracil (DNA), erlotinib (EGFR), gemcitabine (cytoskeleton), paclitaxel (β- tubulin), tamoxifen (estrogen receptor), trametinib (MEK).

In the second group there were 12 drugs in overlap between 3 studies: bortezomib (proteasome inhibitor), crizotinib (ALK, ROS1), docetaxel (microtubules), doxorubicin (DNA intercalating agent), erlotinib (EGFR), GDC-0941 (PI3K),

gemcitabine (DNA synthesis inhibitor), lapatinib (EGFR), entinostat (HDAC1, HDAC3), paclitaxel (β - tubulin), sirolimus (mTOR), vorinostat (HDAC).

Cell line genomics and drug response data were obtained via the PharmacoGx package (version 1.8.3).[66] As a drug response metric we used area under curve (AUC). Xenograft genomics and raw drug response data for NIBR PDXE were taken from papers' supplementary data.[37]

We applied GLDS correction method to improve biomarkers' consistency in the following way. After defining the lists of unrelated drugs for each drug within each dataset, we calculated 2 models for each feature from each dataset (for each of 6 drugs) - simple model and model with 10 additional predictors which are 10 principal components calculated from the matrix of drug responses of unrelated drugs (Fig. 13b).

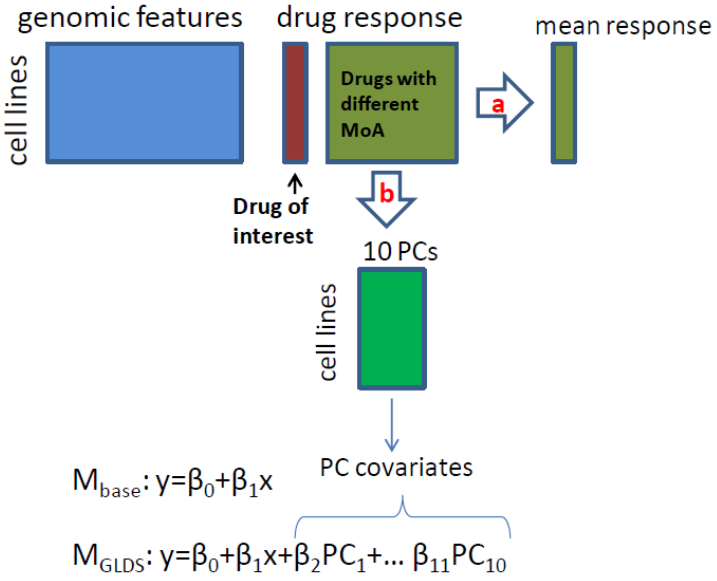


Figure 13. Schematic overview of GLDS method. (**a**) simple variant of GLDS where we just use a vector of mean responses calculated from responses of unrelated drugs. (**b**) GLDS version for biomarker discovery which uses 10 PC calculated from responses of unrelated drugs as covariates in linear model.

In order to assess consistency of biomarkers before and after GLDS we analysed regression coefficients (associated with biomarker).
In each dataset we selected top N features with highest absolute value of regression coefficient and then looked at the number of overlapped selected features between 3 datasets before and after GLDS correction.
Also we took a union of these features (from 3 datasets) and calculated Pearson correlation of regression coefficient vectors for features within this union between each pair of datasets (GDSC-CTRP, GDSC-NIBR or GDSC-gCSI, CTRP-NIBR or CTRP-gCSI) before and after GLDS correction.

2.3 Results

2.3.1 Drug response consistency

After applying the first method, simple cell line filtering, Spearman correlation between drug responses from GDSC and CTRP increased at least slightly almost in all cases (we did the test for 6 groups of drugs with the following targets: HDAC, EGFR, MEK1/MEK2, BRAF, HSP90, mTOR, see Table 5).
After applying the second method (GLDS correction, implemented as simple as subtracting average drug response, calculated on unrelated drugs, from the drug response of the drug of interest) cross-set correlation only decreased in all tested cases. Detailed results are shown in the Table 6.

Table 6. Results of cell line filtering and GLDS correction. Cases where cross-set correlation increased after filtering are marked in red.

target	drug	**initial correlation**	**Spearman corr. after filtering**	average correlation after 10 random subsettings	# of common lines	# of lines after filtering	**Spearman corr. after GLDS correction**
	MS-275	0.36	0.5	0.33	165	136	0.23
	Belinostat	0.5	0.59	0.54	199	161	0.25
	Tubastatin A	0.31	0.37	0.3	222	176	0.25
HDAC	Vorinostat	0.57	0.64	0.59	390	314	0.23
	Lapatinib	0.1	0.16	0.05	146	114	-0.04
	Erlotinib	0.27	0.22	0.26	148	113	0.02
	Afatinib	0.05	0.09	0.03	383	322	-0.02
EGFR	Gefitinib	0.07	0.1	0.08	372	303	0.06
	Trametinib	0.51	0.48	0.53	192	158	0.09
MEK1,2	selumetinib	0.25	0.33	0.26	429	346	0.06
	PLX4720	0.35	0.38	0.34	433	353	0.32
BRAF	Dabrafenib	0.3	0.27	0.29	187	153	-0.06
	SNX-2112	0.56	0.58	0.3	430	349	0.23
HSP90	17-AAG	0.11	0.2	0.08	392	314	-0.16
	OSI-027	0.39	0.38	0.39	424	344	0.3
	Rapamycin	0.35	0.45	0.39	162	124	0.12
	BEZ235	0.21	0.29	0.19	279	228	-0.04
	Temsirolimus	0.13	0.22	0.17	201	158	0.12
mTOR	AZD8055	0.34	0.43	0.33	390	313	0.08
	Average	0.30	0.35	0.3	292	236	0.11

2.3.2 Biomarkers' consistency

First, we analysed biomarkers' consistency before and after GLDS correction between GDSC, CTRP and NIBR datasets**.** Results are shown in the Table 7. (cases where consistency after GLDS correction improved are marked in red)
Only for one drug out of 6, Trametinib, we improved consistency between each pair of sets. For two drugs that target cytoskeleton – Gemcitabine and Paclitaxel, only NIBR-GDSC and NIBR-CTRP consistencies were improved (but not GDSC-CTRP). For Tamoxifen consistency between GDSC and CTRP seriously decreased after GLDS (marked in purple in the table).

Table 7. Results of GLDS correction, for GDSC-CTRP-NIBR data, top 500 biomarkers. Cases where cross-set correlation between vectors of regression coefficient increased are marked in red. Cases where the number of common biomarkers (out of total 500) between 3 datasets increased are marked in red too.

GDSC - CTRP - NIBR (10 princ. comp.)		# common biomarkers	cor GDSC-CTRP	cor GDSC-NIBR	cor CTRP-NIBR
fluorouracil	before	21	0.71	0.10	0.013
	after	13	0.64	0.04	0.018
erlotinib	before	5	0.27	-0.27	-0.07
	after	2	0.28	-0.27	-0.08
gemcitabine	before	19	0.62	0.13	0.22
	after	17	0.57	0.19	0.25
paclitaxel	before	3	0.02	0.06	0.002
	after	4	-0.15	0.11	0.06
tamoxifen	before	6	0.18	0.07	-0.11
	after	4	0.014	0.14	-0.07
trametinib	before	2	0.60	-0.22	-0.07
	after	20	0.84	0.24	0.26

Also we performed the same analysis using another cell line dataset (gCSI[36]) as a third set i.e. GDSC-CTRP-gCSI, corresponding results are presented in the Table 8. Consistency improved at least slightly for 6 (out of 12 common) drugs between CTRP and gCSI set, only for two drugs between GDSC and and gCSI and only for one drug between GDSC and CTRP.

Table 8. Results of GLDS correction, GDSC-CTRP-gCSI data, for top 500 biomarkers. Cases where cross-set correlation between vectors of regression coefficient increased are marked in red. Cases where the number of common biomarkers (out of total 500) between 3 datasets increased are marked in red too.

GDSC - CTRP - gCSI (10 princ. comp.)		# common biomarkers	cor GDSC-CTRP	cor GDSC-gCSI	cor CTRP-gCSI
bortezomib	before	2	-0.40	0.07	0.23
	after	1	-0.14	0.11	0.12
crizotinib	before	2	-0.34	0.24	-0.11
	after	3	-0.39	0.31	-0.28
docetaxel	before	7	-0.48	-0.38	0.73
	after	10	-0.47	-0.36	0.76
doxorubicin	before	6	0.41	0.49	0.70
	after	8	0.34	0.35	0.72
erlotinib	before	2	0.40	0.66	0.65
	after	4	0.40	0.53	0.74
GDC-0941	before	6	0.54	0.47	0.44
	after	4	0.45	0.41	0.37
gemcitabine	before	18	0.67	0.68	0.71
	after	19	0.61	0.63	0.71
lapatinib	before	3	0.20	0.71	0.32
	after	4	0.48	0.70	0.56
entinostat	before	4	0.72	-0.55	-0.56
	after	4	0.54	-0.41	-0.61
paclitaxel	before	4	0.01	0.03	0.74
	after	3	-0.19	-0.18	0.77
sirolimus (rapamycin)	before	0	-0.31	-0.05	0.61
	after	0	-0.24	-0.11	0.62
vorinostat	before	13	0.77	0.75	0.75
	after	13	0.77	0.75	0.74

3 ATTEMPTS TO IMPROVE PREDICTION ACCURACY

3.1 Introduction

In this chapter we applied a number of various model training approaches with the common goal of improving the accuracy of predictions. In an attempt to co-utilise information from drugs that share the same target we tried to use regularised linear regression multi-task models (multi-task means that we train a group of drug response models jointly rather learn training each model independently) as well standard (single-task) models built on aggregated data (Fig. 19). In order to take into account interactions between genomic features we tried modelling with binary gene pairs (BGP) and multiplied features. Also we tried modelling with sample weighting giving higher weights to (usually) under-represented sensitive cell line samples. Finally we checked the effects of class imbalance and cross-set consistency on accuracy.

3.2 Data and Methods

In this group of analyses we used genomic and drug response data from two large datasets, CTRP and GDSC obtained via the PharmacoGx package (version 1.8.3).[66] Particularly we focused on 19 drugs belonging 6 classes defined by drug's target molecule (Table 5). As a drug response metric we used area under the dose-response curve, AUC (see Fig. 20).

As modelling techniques we used regularized regression (lasso/ridge/elastic net) via glmnet package and Random Forest via caret + randomForest packages. Feature selection, model fitting and accuracy evaluation were performed using the following procedure:

1. We select one dataset as a training set, and the other one as a test set. (e.g. GDSC as a training set and CTRP as a test set).
2. We perform feature selection using the gamScores function from caret package (see details in the section 1.7.2) on the training set.
3. Then we fit the model with N selected features (with lowest p-values) on the training set data. Model's hyperparameters are selected using cross-validation testing.
4. We apply model to the test set, and calculate R^2 (explained variance, calculated as a square of correlation between predicted and observed outcomes) and RMSE (root mean squared error).

3.3 Results

3.3.1 Multi-task glmnet models

Multi-task learning is a class of machine learning approaches in which multiple learning tasks are solved at the same time. We hypothesized that we can gain some accuracy if we use multi-task approach in order to predict drug response for drugs that share the same molecular target.

For testing this approach we used the functionality of glmnet package. Glmnet allows users to fit regularized regression models, i.e. it penalizes a number of variables with non-zero coefficients. In multi-task case information sharing between the tasks "involves which variables are selected, since when a variable is selected, a coefficient is fit for each response".[80] We compared single-task (standard) and multi-task lasso models using all cell lines (pan-cancer modelling) and cell lines for each major tissue type separately (tissue-specific modelling). In both cases we haven't found a consistent accuracy gain in multi-task models' results (Fig. 14 and Fig. 15).

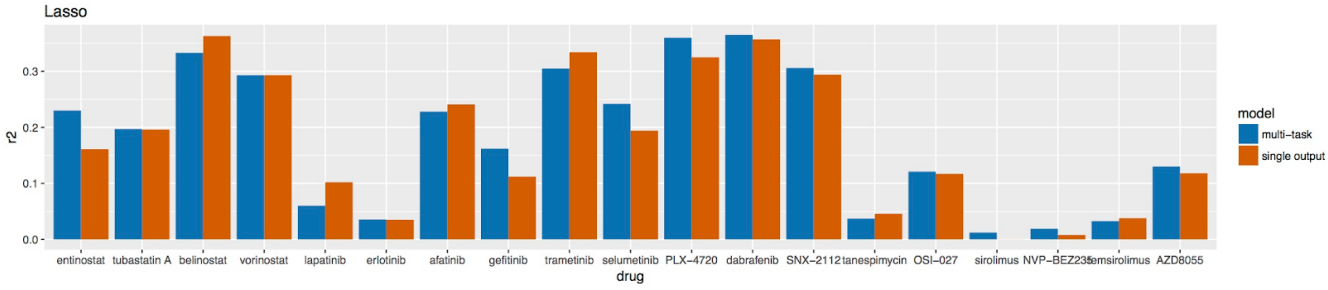


Figure 14. Pan-cancer results of multi-task modelling in terms of R^2. Lasso regression was used for modelling. Training on GDSC dataset, validation on CTRP dataset.

Figure 15. Tissue-specific results of multi-task modelling for 6 tissues with the largest number of corresponding samples, R^2. Training on GDSC dataset, validation on CTRP dataset. Red bars represent multi-task models, blue bars represent single-task (standard) models.

3.3.2 Modelling on aggregated data

Here we combined training samples for drugs sharing the same molecular targets (see Fig. 16) and compared performance of models trained on the aggregated datasets with performance of models trained on single drug datasets.

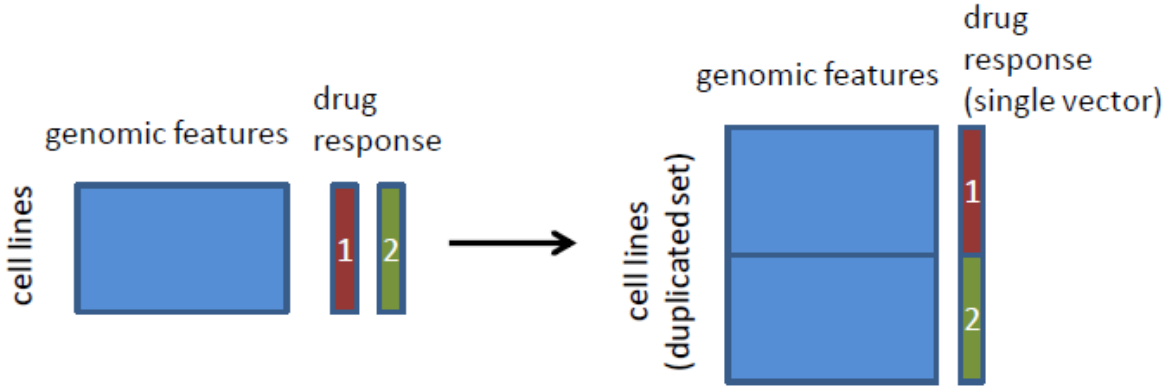


Figure 16. Visualization for the idea of modelling on aggregated data. In this example drug 1 and drug 2 are drugs which share the same target, and we combine their drug responses in a single vector in order to build a common aggregated model.

For modelling we used Random Forest modelling method. Results in terms of R^2 are plotted in the Fig. 17. Similarly to multi-task modelling results here we don't observe advantage of aggregated models over simple ones.

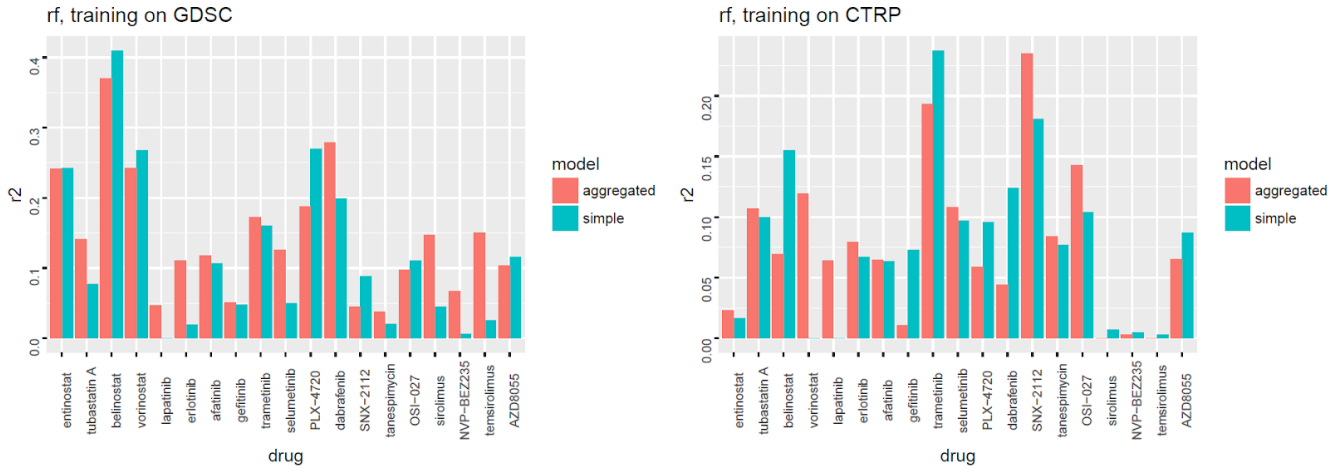


Figure 17. R^2 results for aggregated and standard approaches. Left picture -- training is done on GDSC set, testing is on CTRP set, right picture -- training is done on CTRP set, testing is on GDSC set.

3.3.3 Modelling with feature interactions

Here we tried to improve model accuracy using feature engineering approach. We tested two types of engineered features based on combining information from a pair of expression features:

1) Binary gene pairs (BGP):

$$BGP(A,B)=\begin{cases}1, \text{ if } expA>expB\\0, \text{ if } expA<expB\end{cases}$$

2) Multiplications of gene pairs: expA*expB

In order to reduce total number of resulting features we used a number of feature sets filtered with different feature selection approaches:
Total list of tested feature sets (“200 top cor” and “200 top var” denote here 200 features with highest correlation with outcome or 200 most variant features) is presented in the Table 9:

Table 9. List of tested feature sets

Feature set	# of features
1. All features	20000
2. 200 features with the highest correlation (with drug response)	200
3. 200 features with the highest variance (across all samples)	200
4. BGP features constructed from 200 features with the highest correlation	20000 Comment: 200*200=40000 but since BGP(A,B) and BGP (B,A) contain

	essentially the same information [BGP(A,B) =1- BGP (B,A)] the total number of used features here is 40000/2=20000
5. 200 BGP features with highest correlation (with drug response) constructed from 200 original features with the highest correlation	200
6. Multiplicated features constructed from 200 features with the highest correlation	20000 Comment: 200*200=40000 but since exp(A)*exp(B) = exp(B)*exp(A) the total number of used features here is 40000/2=20000
7. 200 multiplicated features with highest correlation (with drug response) constructed from 200 original features with the highest correlation	200
8. BGP features constructed from 200 features with the highest variance	20000
9. 200 BGP features with highest correlation (with drug response) constructed from 200 original features with the highest variance	200
10. Multiplicated features constructed from 200 features with the highest variance	20000
11. 200 multiplicated features with highest correlation (with drug response) constructed from 200 original features with the highest variance	200

We used lasso regression for modelling, R^2 results averaged across 19 tested drugs are showed in the Fig. 18.

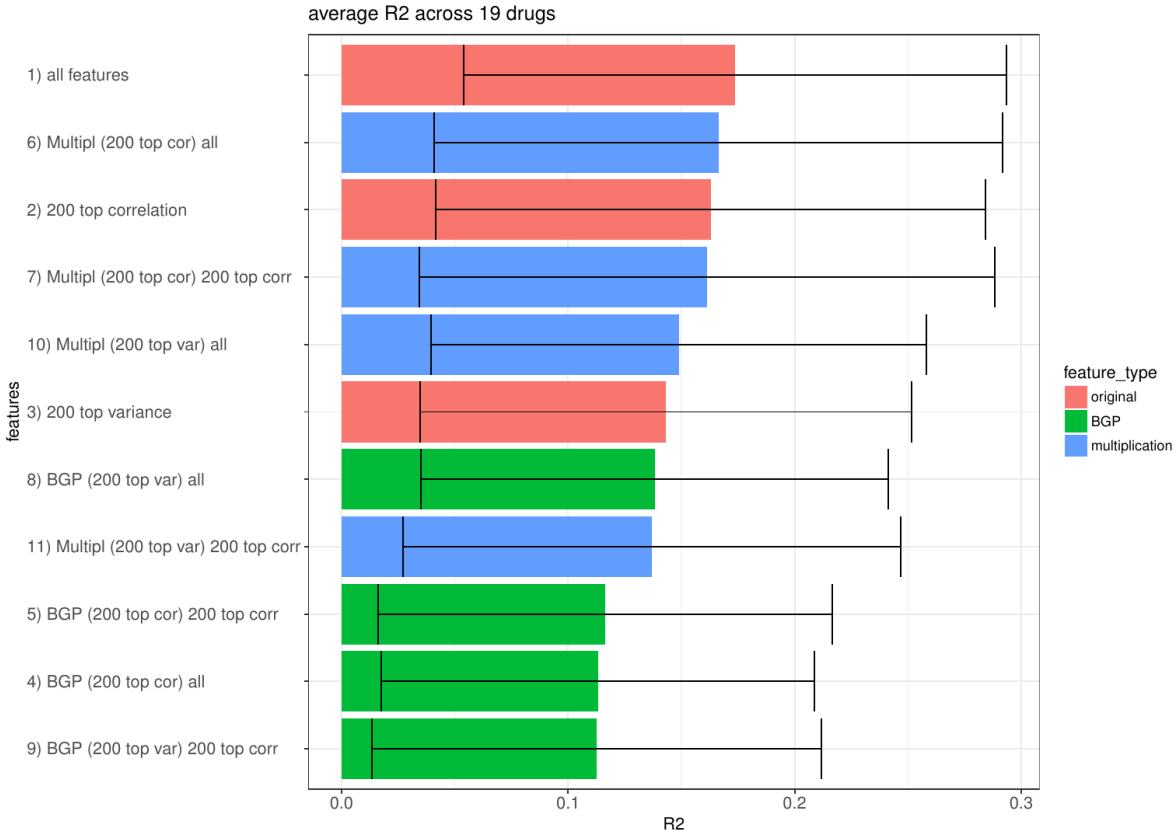


Figure 18. R^2 results averaged across 19 tested drugs for each of the feature sets tested. Red bars correspond to non-engineered features, green bars correspond to BGP features, blue bars correspond to multiplicated features. Error bars depict ± one standard deviation.

We can see that "multiplicated" features perform better than BGP, although just all features without feature selection or top 200 non-engineered (i.e. original) features selected by correlation perform as good as the best "multiplication" feature set.

Also we performed the same analysis using Random Forest instead of lasso as a modelling method, R^2 results are shown in the Table 10.

Table 10. R^2 results for random forest model based on different feature sets.

Feature selection	average R^2 across 19 drugs (GDSC → CTRP)	average R^2 (CTRP → GDSC)
200 top correlation	0.126	0.092
200 top variance	0.119	0.035
BGP (200 top cor) 200 top corr	0.132	0.129
Multipl (200 top cor) 200 top corr	0.103	0.083

In this case BGP performed better than multiplication features and even slightly better than top 200 highly correlated original features.

3.3.4 Modelling with weights

There is a common pattern in the distribution of drug response/drug sensitivity values for a drug among the panel of tested cell lines -- usually sensitive samples are less represented compared to the resistant ones (this is especially the case for targeted therapies). Here, in order to improve accuracy of prediction, we tried to assign higher weights to under-represented sensitive samples using a number of different weighting schemes.

We used lasso regression for modelling, results for different weighting schemes used are shown in the Table 11. As we see weighting doesn't produce visible increase in accuracy.

Table 11. Different weighting schemes tested and corresponding average R^2 values.

weighting	average R^2 across 19 drugs (GDSC → CTRP)	average R^2 (CTRP → GDSC)
no weighting	0.175	0.167
w=1/AUC	0.175	0.163
w=1 (AUC>0.5), w=2 (AUC<0.5)	0.176	0.168
w=1 (AUC>0.5), w=10 (AUC<0.5)	0.178	0.161

3.3.5 How class imbalance and cross-set inconsistency affect prediction accuracy

We checked the effects of class imbalance and cross-set consistency (measured as cross-set AUC pearson correlation) on model performance (R^2 from lasso single-task models) for 19 drugs:

- Class imbalance in GDSC set is negatively correlated with R^2 of models trained on GDSC set (-0.08) but surprisingly class imbalance in CTRP set is positively correlated with R^2 of models trained on CTRP set (0.21).
- For both types of models, expectedly, there is a strong correlation between a cross-set AUC correlation and R^2, 0.85 for models trained on GDSC set (and tested on CTRP set) and 0.91 for models trained on CTRP set (and tested on GDSC set).

4 TESTING INFLUENCE OF DIFFERENT ASPECTS OF MODEL TRAINING ON PREDICTION ACCURACY

4.1 Introduction

In this analysis we applied established machine learning methods to characterize how different model training strategies influence resulting performance. For this end we used data from largest comprehensive cell line drug screens performed up to date -- the Cancer Cell Line Encyclopedia[31] (CCLE), Cancer Therapeutic Portal[32] (CTRP) and Genomics of Drug Sensitivity in Cancer[34] (GDSC). In the analysis we particularly focus on the properties of training set -- feature type, response metric and number of features used in the model (Fig. 19).

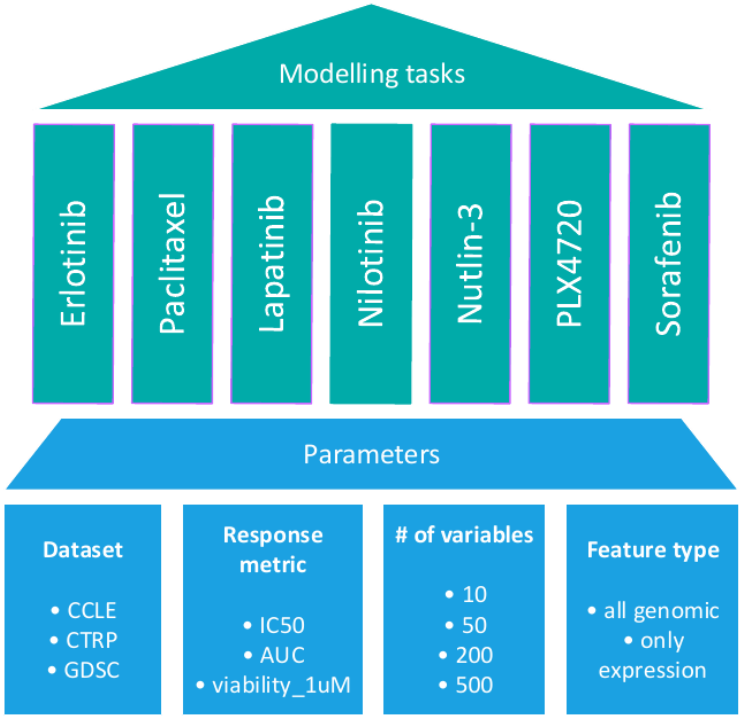


Figure 19. Overview of the analysis in the chapter 4. Modelling tasks (drug responses) and tested parameters (dataset, response metric, # of variables, feature type) are shown.

4.2 Data and Methods

4.2.1 Data

For modelling we used molecular (microarray expression, copy number, mutation information) and drug response data from 3 large cell line sensitivity screenings -- CCLE[31], CTRP[32] and GDSC[34].

Cell line genomics and drug response data were obtained via the PharmacoGx package (version 1.8.3)[66]. Particularly for IC_{50} (half maximal inhibitory concentration) and AUC (area under the drug response curve) data we used values recomputed by the package from the raw data -- "ic50_recomputed" and "auc_recomputed" (Fig. 20). In order to handle outlier values in IC_{50} data we truncated the distribution at the 85th percentile. Also in this analysis in addition to AUC and IC_{50} we used another metric -- viability at 1µM (µmol) extracted from the raw drug response data in each dataset (Fig. 20).

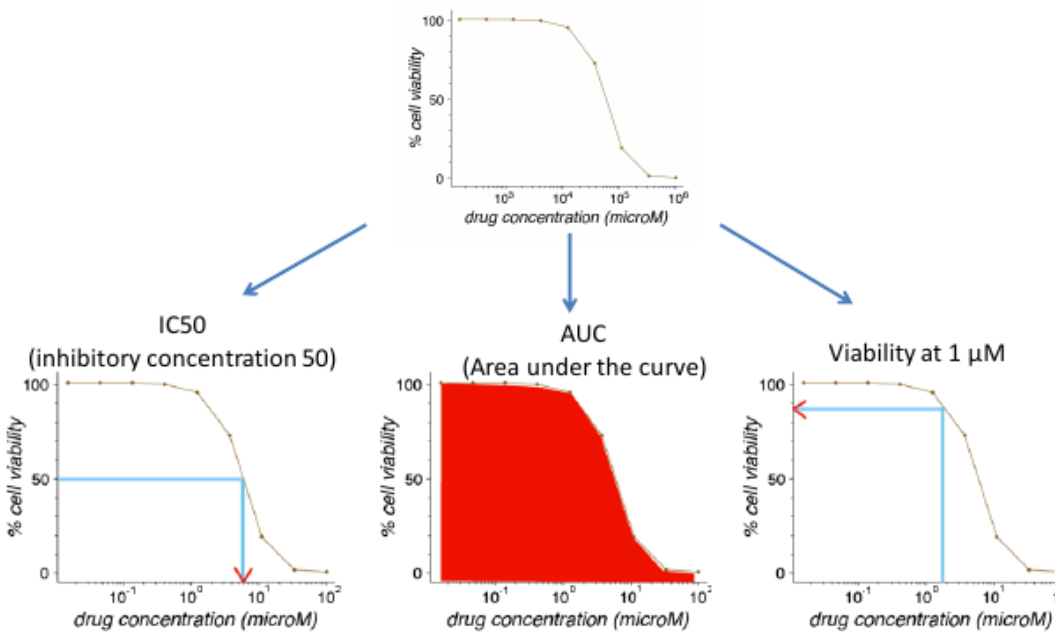


Figure 20. Different cell line drug response metrics tested in the chapter 4. From the raw viability data we can calculate 3 drug response metrics -- IC_{50}, AUC and viability at 1µM.

4.2.2 Modelling

For feature selection we employed filter-type feature selection function gamScores from caret package (see details in the section 1.7.2). After feature selection we fit the model with N selected features (with lowest p-values) on the training set data.

As a modelling method we used SVM with Radial base function (svmRadial). In order to select hyperparameters (sigma and C(cost)), 30 different combinations of them are tested on training data using 10 fold cross-validation, and then the

combination that provides the lowest RMSE is used for fitting the final model. Hyperparameters ranges: sigma ⊂ [0.001;0.01], cost ⊂ [0.03;10000].
As the accuracy measures we used R^2 (explained variance) and RMSE (root of mean squared error).

Feature selection, model fitting and accuracy evaluation were performed using the following procedure (Fig. 21):

1. We randomly split the data into training (70%) and test (30%) sets.
2. We perform feature selection on the training set.
3. Then we fit the model with N selected features (with lowest p-values) on the training set data. Model's hyperparameters are selected using cross-validation testing.
4. We apply model to the test set, and calculate R^2 (calculated as a square of correlation between predicted and observed outcomes) and RMSE.
5. We repeat steps (1-4) ten times and get average R^2 and RMSE.

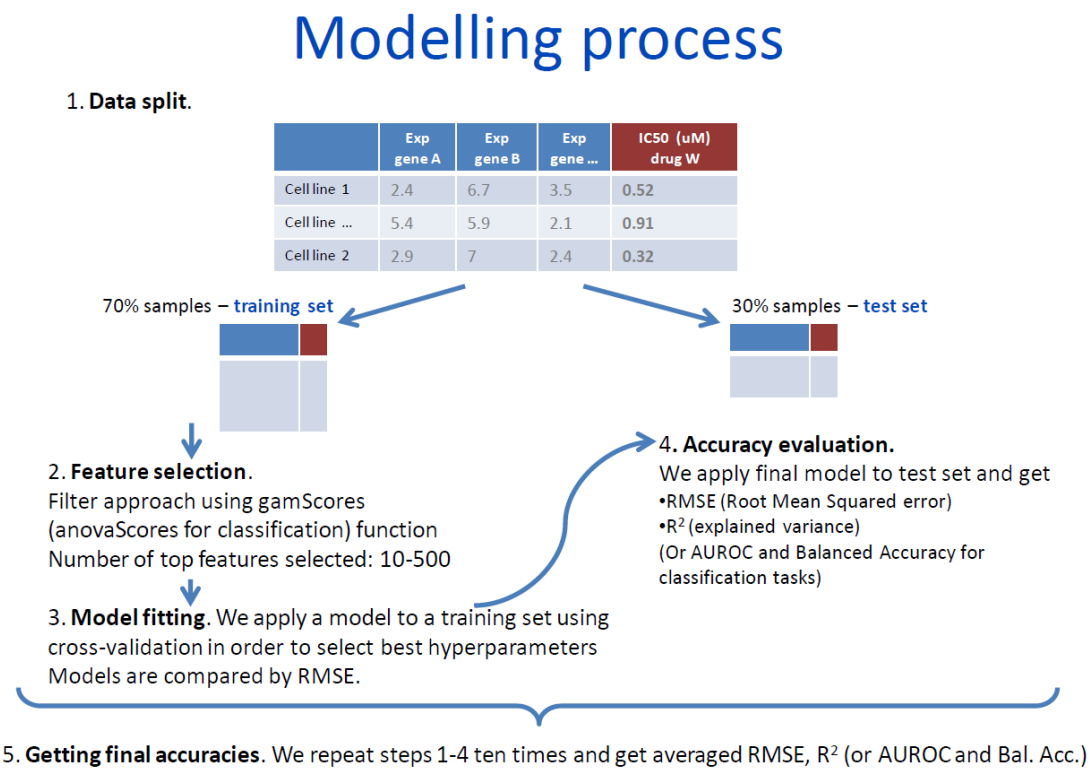


Figure 21. Overview of the modelling process. Main steps of modelling process are shown schematically: 1) Data split, 2) Feature selection, 3) Model fitting, 4) Accuracy evaluation.

4.3 Results

We trained a number of models for seven drugs which are common between the 3 largest cell line datasets, CCLE, CTRP and GDSC: erlotinib (EGFR), paclitaxel (β-tubulin), lapatinib (EGFR), PLX4720 (RAF), sorafenib (RAF), nutlin-3 (MDM2) and nilotinib (ABL/BCR-ABL). Our drug-specific models differed in terms of:

- Molecular feature types -- only expression vs. expression+copy number+mutation
- Drug response metrics -- IC_{50}, AUC, Viability_1µM
- Number of variables in the model: 10, 50, 200, 500

Results (for tests with all genomic features) in terms of R^2 are plotted in Fig. 22a. Each plot shows results for certain response metric combination, and within each plot there are results for each drug in each data set for the tests with 500 variables.

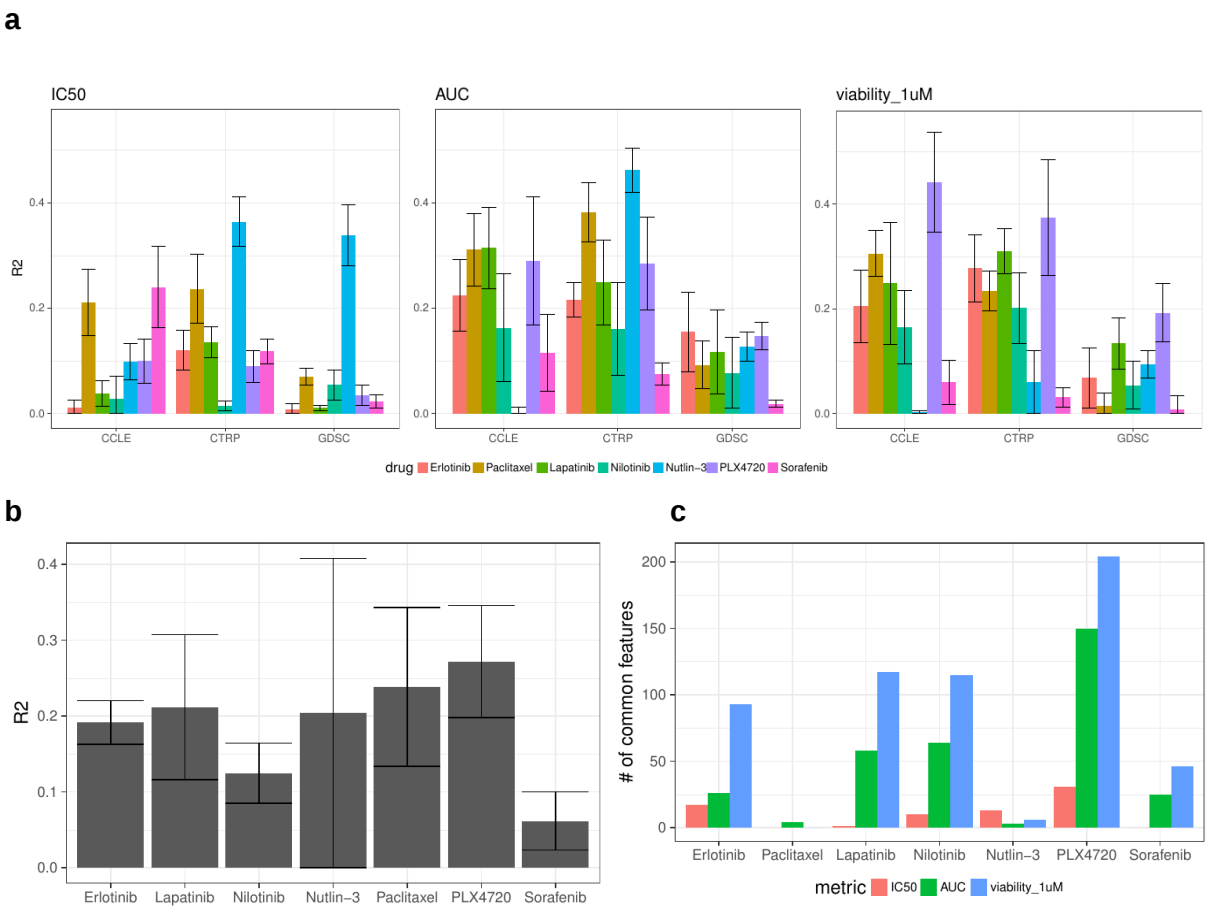


Figure 22. Accuracy results and analysis of common features (**a**) R^2 for 7 drugs across multiple testing conditions, number of variables=500. (**b**) Average (across three datasets) R^2 values for each drug separately (for models with AUC metric). Error bars depict ± one standard deviation. (**c**) Number of common features out of top 500 features between 3 datasets (CCLE, CTRP and GDSC) for each drug within each drug response metric.

We plotted predictions against observed values for each drug response metric for some drug/dataset combinations. In each plot the data points correspond to a test set from one particular (random) train/test sets split (Fig. 23). A dotted line shows where data points should lie in the ideal case of 100% correct predictions. In cases where the accuracy of prediction is satisfactory, the data points are grouped around the dotted line e.g. IC_{50}, paclitaxel, CTRP (1st row, 1st column); AUC, nutlin-3, CTRP (2nd row, 2nd column); viability_1uM, paclitaxel, CCLE (3rd row, 1st column). In cases where prediction accuracy is low predicted and observed values are not correlated and therefore data points are not grouped around the dotted line.

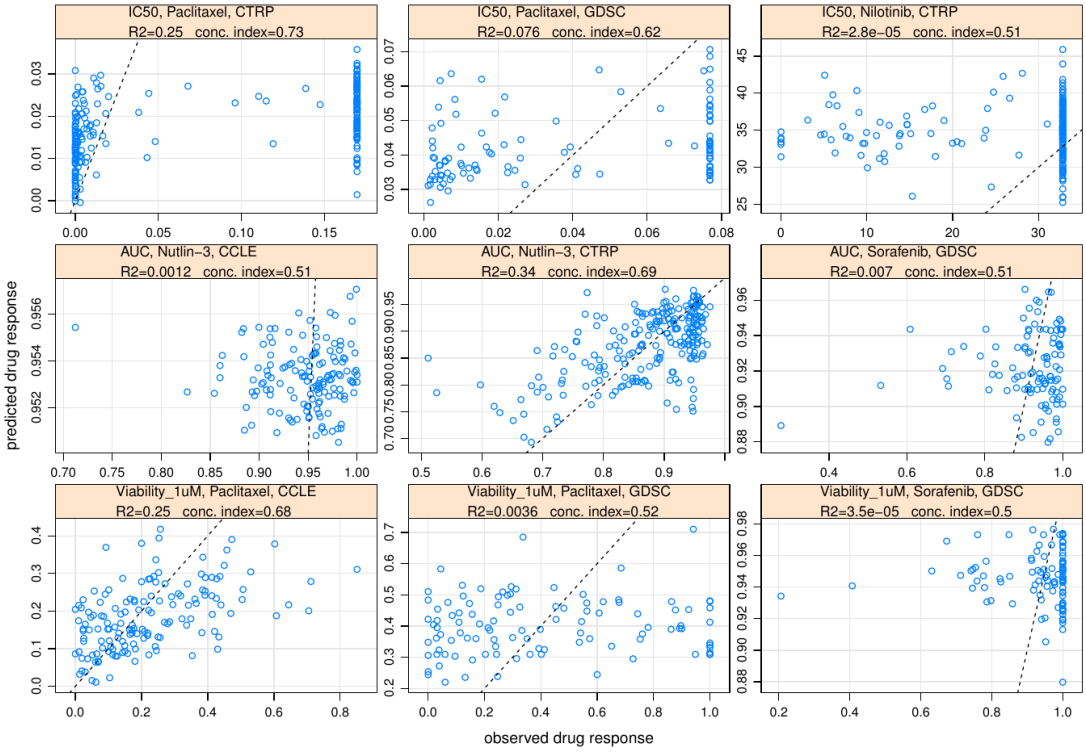


Figure 23. Observed vs. predicted values for drugs/datasets combinations that show either relatively high or extremely low R^2 accuracies within each drug response metric (IC_{50}, AUC, viability at 1μM).

We tested how prediction performance depends on the number of top variables (variables that have a strong correlation with outcome) selected for modelling and surprisingly found almost no correlation between R2 and number of variables, Pearson correlation across all tests = 0.02. Correlation coefficients for each drug individually vary between -0.08 and 0.11

In these tests the average value of R^2 for modelling with all genomics data (0.153) was just slightly higher than for modelling with only expression data (0.145). While these differences are small for most of the drugs, for nutlin-3 and PLX4720 they are a bit more pronounced, which means that mutation status information contributes to the explained variance of drug response for these two drugs (Fig. 24). Below, we will discuss only models that are based on all genomic features (expression, copy number and mutation values).

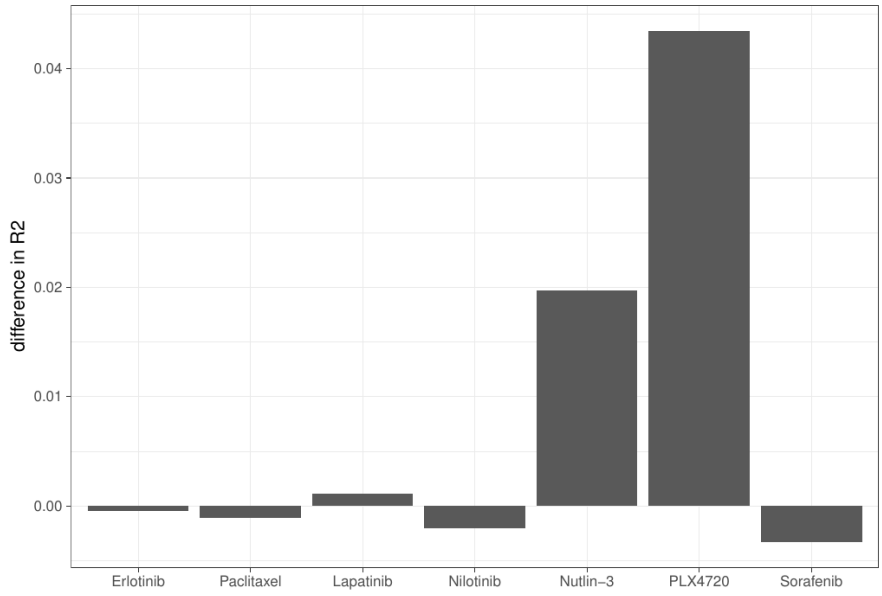


Figure 24. Difference in average R^2 between models that use all genomic features and models that use only expression features for 7 drugs.

We checked the number of common features out of selected top 500 features for each drug between 3 datasets (CCLE, CTRP and GDSC) within each drug response metric. The number of common features is relatively small for IC_{50} metric (average = 10) and higher for AUC and viability at 1µM metrics (average = 47 and 83 correspondingly, see Fig. 22c). Independently of response metric used there is almost no common features across paclitaxel models and across nutlin-3 models.

Average R^2 values for each drug (for AUC models) are shown in the Fig. 22b. Five drugs -- PLX4720, paclitaxel, lapatinib, nutlin-3 and erlotinib had average R^2 between 0.2 and 0.3 while Nilotinib and Sorafenib showed the lowest average predictability (R^2=0.12 and 0.06 correspondingly). Average R^2 for each tissue separately are shown in the Fig. 25.

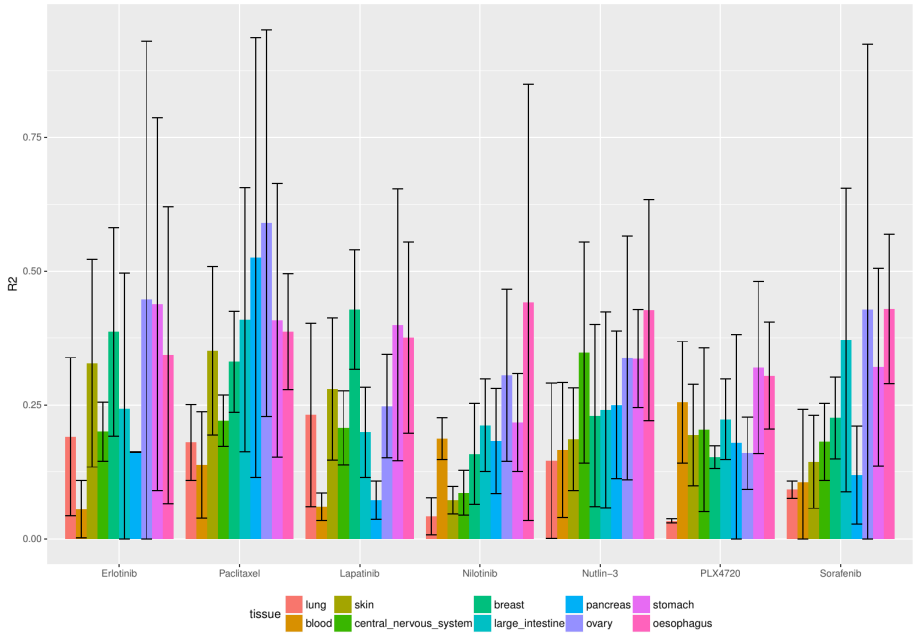


Figure 25. Average (across three datasets) R^2 values for each tissue and each drug separately (for models with AUC metric and 500 variables). Error bars depict ± one standard deviation.

We compared our results for CCLE dataset with elastic net modelling performance from the original CCLE study[31] and performance from integrated (combined) random forest method (CRF)[8] which was the second top performing method in the DREAM drug response prediction challenge[47] (We are comparing our results with the second top performing method instead of the first one simply because both methods have quite similar accuracy score in the original paper, wpc-index equals to 0.583 and 0.577 correspondingly, but the second method to our convenience was already tested on CCLE dataset with essentially the same accuracy metric that we are using in our analysis). Corresponding R^2 values are shown in the Table 12.

Table 12. Comparison between prediction results from different methods in the form of R^2 values. Dataset: CCLE. Response metric: AUC. Elastic net denotes the approach used in [31]. CRF-400 and CRF-20000 denote the approach used in [81]. SVM-500 denotes our results. Highest R^2 value for each drug is highlighted in boldface.

Drug	Elastic Net	CRF-400	CRF-20000	SVM-500 (our results)
erlotinib	0.09	0.16	0.18	**0.22**
paclitaxel	**0.36**	0.30	0.30	0.31
lapatinib	0.20	0.30	0.28	**0.31**
nilotinib	**0.58**	0.30	0.30	0.16
nutlin-3	0.01	0.08	**0.10**	0.003
PLX4720	**0.30**	0.20	0.23	0.29
sorafenib	0.07	0.17	**0.22**	0.12

5 TISSUE TYPE, DOUBLING TIME AND DRUG RESPONSE PREDICTION IN CELL LINES AND XENOGRAFTS

5.1 Introduction

In this analysis we compared the task of drug response prediction with the presumably easier prediction tasks, tissue type prediction and doubling time prediction, in terms of accuracy in order to see whether (and to which extent) potentially easier phenotypes than drug response can be predicted from genomic data. Additionally we compared the level of consistency for these results between cell lines and xenografts (Fig. 26).

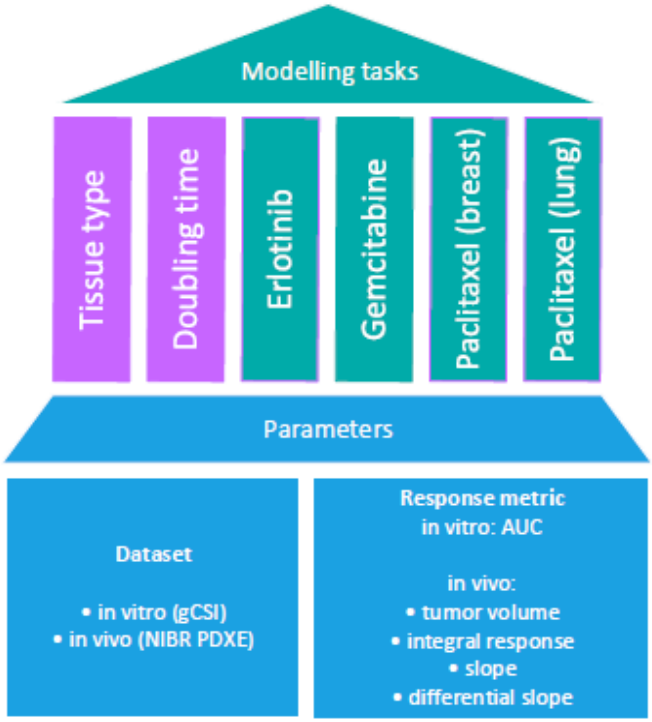


Figure 26. Overview of the analysis in the chapter 5. Modelling tasks (tissue type, doubling time and drug responses) and tested parameters (dataset, response metrics) are shown.

5.2 Data and Methods

5.2.1 Data

For modelling we used molecular (microarray expression, copy number, mutation information) and drug response data from cell line sensitivity screen gCSI[36] and xenograft screen NIBR PDXE[37]. Cell line genomics data for gCSI dataset were taken from paper's supplementary data[36], drug response data for

gCSI was obtained via the PharmacoGx package (version 1.8.3).[66] Xenograft genomics and raw drug response data for NIBR PDXE were taken from papers' supplementary data.[37]

As a cell line drug response metric we used Area under drug response curve (AUC). For characterizing xenograft drug response we tested 4 different xenograft drug response metrics derived from raw drug response data (volume of the tumour during the treatment course at different days between day 0 and day 21): tumour volume (at the day 21), integral response, slope of the tumor growth curve and differential slope (Fig. 27).

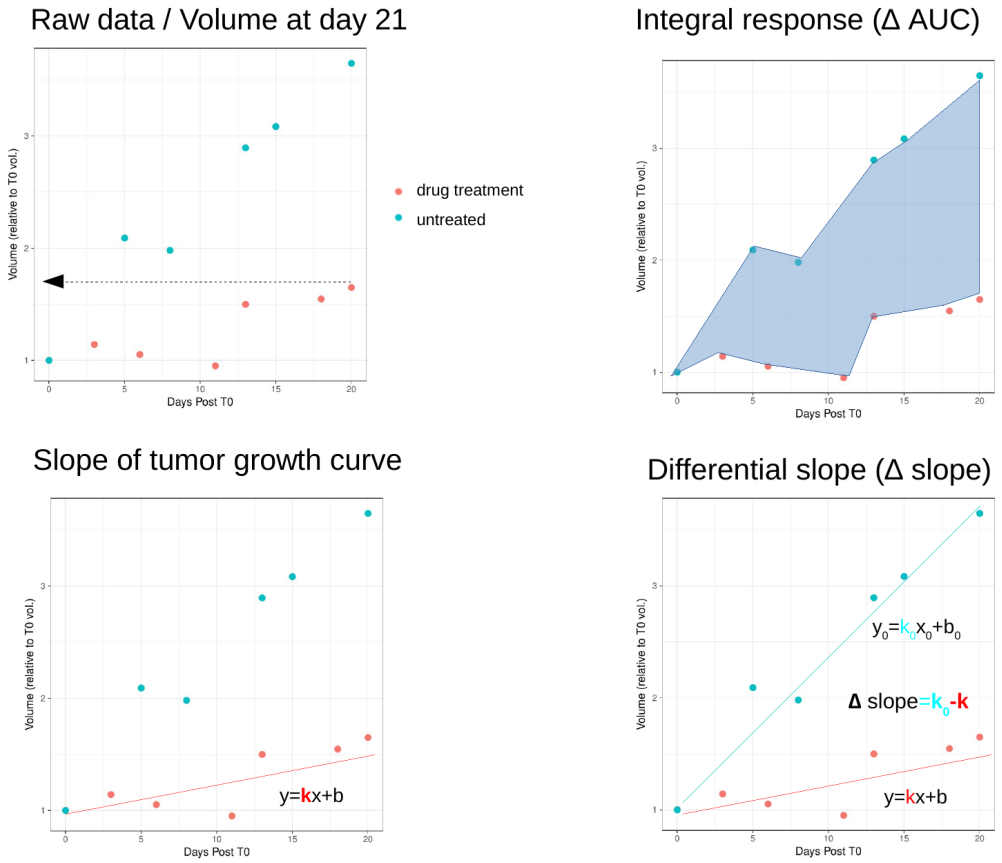


Figure 27. Xenograft's drug response metrics. Top: volume of the tumor, integral response; bottom: slope of the tumor growth curve, differential slope.

5.2.2 Modelling

For feature selection we employed filter-type feature selection functions gamScores (for regression tasks) and anovaScores (for classification tasks) from caret package (see details in the section 1.7.2). After feature selection we fit the model with N selected features (with lowest p-values) on the training set data.

As a modelling method we used Random Forest for all modelling tasks except tissue classification for which xgBoost method was used. In order to select hyperparameters (e.g. mtry, number of predictors at each split, for RF), 30 different values for a hyperparameter are tested on training data using 10 fold

cross-validation, and then the value that provides the lowest RMSE (or AUROC for classification tasks) is used for fitting the final model. As the accuracy measures we used R^2 (explained variance) and RMSE (root of mean squared error) for regression tasks and percentage of correctly predicted samples for tissue classification tasks.

Feature selection, model fitting and accuracy evaluation were performed using the following procedure (see Fig. 21):

1. We randomly split the data into training (70%) and test (30%) sets.
2. We perform feature selection on the training set.
3. Then we fit the model with N selected features (with lowest p-values) on the training set data. Model's hyperparameters are selected using cross-validation testing.
4. We apply model to the test set, and calculate R^2 (explained variance, calculated as a square of correlation between predicted and observed outcomes) and RMSE. Alternatively for tissue classification the percentage of correctly predicted samples was calculated.
5. We repeat steps (1-4) ten times and get average R^2 and RMSE (or percentage of correctly predicted samples for tissue classification tasks).

5.3 Results

We used the gCSI study[36] as in vitro training set, it is a high quality pharmacogenomic dataset reasonably consistent with CCLE and GDSC datasets. We used the NIBR PDXE[37] as in vivo validation set set since this is the only publicly available xenograft screen. We assessed 6 modelling tasks in each set -- tissue type prediction, doubling time/slope of the tumor growth curve, erlotinib response (lung samples), gemcitabine response (pancreas samples), paclitaxel response (breast samples) and paclitaxel response (lung samples). Erlotinib, gemcitabine and paclitaxel were selected since these three drugs were tested in both gCSI and NIBR PDXE study. Tissue type and doubling time prediction tasks serve as a positive controls -- we assume that these phenotypes should be explained by genomics data better than drug response.

For tissue prediction and doubling time/slope of the untreated tumor growth curve we used the top 400 features with lowest p-values. For drug response prediction (since sample sizes were much lower) we used just the top 100 features. In xenograft tissue prediction we had 5 tissue-classes with 27-50 samples per tissue. In cell line tissue prediction we tried modelling with 13 tissue-classes with 10-68 samples per tissue and with 6 largest tissue-classes (each tissue had at least 23 samples).

Prediction accuracy results for all prediction tasks are collected in the Table 13. For tissue type prediction we report percentage of correctly predicted samples, for doubling time/slope and drug response prediction we report R^2 and concordance index values.

Table 13. Prediction accuracy for all prediction tasks. We report here accuracy (percentage of correctly predicted samples) for tissue prediction, R^2 and concordance index for doubling time/slope and drug response prediction.

	Tissue (**Accuracy**)	Doubling time (cell lines) / slope of the growth curve (xenografts) (**R^2**, concordance index)	Drug response (**R^2**, concordance index)					
			drug resp. metric	erlotinib (EGFR) Lung 68 lines 25 xen.	gemcitabine (DNA synth.) Pancreas 26 lines 32 xen.	paclitaxel (β-tubulin) Breast 29 lines 38 xen.	paclitaxel (β-tubulin) Lung 68 lines 23 xen.	Average across 4 drugs
Cell lines (gCSI) 329 samples	$Acc_{6tissues}$=**0.79** $Acc_{13tissues}$=**0.64**	**0.17** (0.64)	AUC	**0.06** (0.57)	**0.13** (0.61)	**0.14** (0.66)	**0.08** (0.57)	**0.10** (0.60)
Xenografts (NIBR) 191 samples, 23-38 samples per drug	Acc=**0.89**	**0.19** (0.60)	Tumor volume	**0.34** (0.69)	**0.04** (0.49)	**0.08** (0.53)	**0.46** (0.74)	**0.23** (0.61)
			Integral response	**0.18** (0.59)	**0.03** (0.50)	**0.08** (0.57)	**0.09** (0.47)	**0.10** (0.53)
			slope	**0.31** (0.65)	**0.15** (0.54)	**0.11** (0.54)	**0.44** (0.63)	**0.25** (0.59)
			Differential slope	**0.12** (0.50)	**0.09** (0.53)	**0.10** (0.54)	**0.27** (0.35)	**0.15** (0.46)

In addition to tissue prediction tests performed using all available cell line and xenografts we made a test with equal number of samples per tissue (16 samples in training set and 7 samples in test set). A heatmap for confusion table for the case of cell line predictions is shown in the Fig. 28.

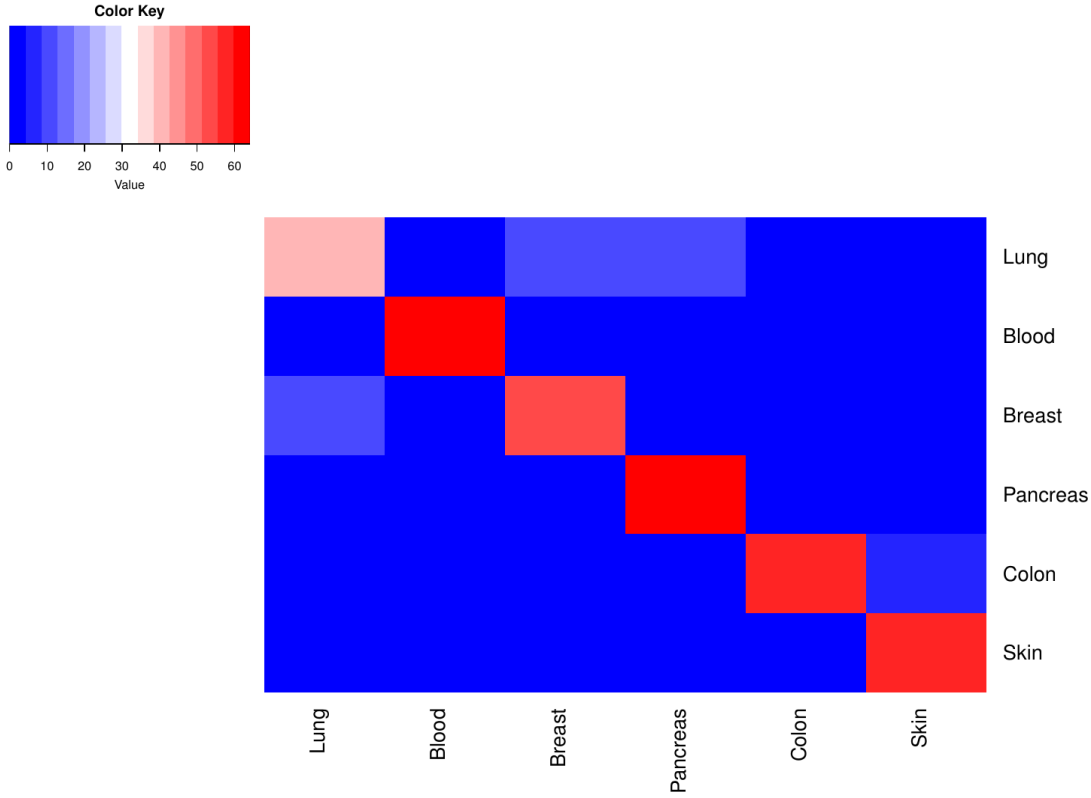
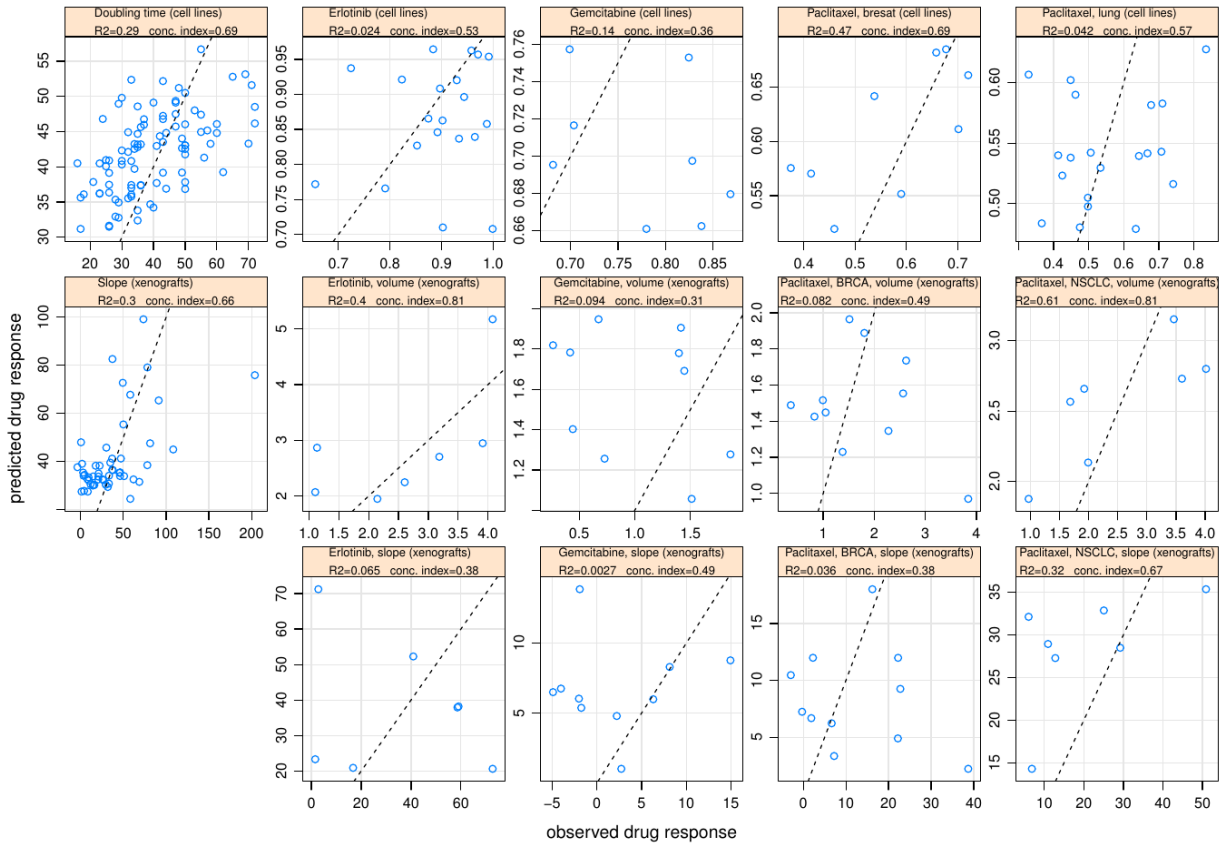


Figure 28. Confusion table heatmap for results of tissue classification in cell lines. Row labels depict true classes, column labels depict predicted classes. Test set contained 7 samples for each tissue, and modelling procedure was repeated 10 times, so for each tissue class 70 predictions were made. Color shows the number of predicted classes per each true class.

While the differences in accuracies between tissue type prediction and doubling time/slope prediction are consistent for cell lines and xenografts, drug response prediction accuracies for the same drugs are not consistent between cell lines and xenografts. Particularly best performing drugs are gemcitabine and paclitaxel (breast) for cell lines and erlotinib, paclitaxel (lung) for xenografts. The level of consistency is also illustrated by the number of common features (between cell lines and xenografts) out of top features pre-selected for modelling. There are 31 common features out of top 400 for tissue prediction between cell lines and xenografts, only 4 common features out of top 400 for doubling time/slope prediction and almost no common features out of top 100 for drug response prediction.

In order to make the quality of prediction across different regression prediction tasks visually accessible, we plotted observed versus predicted values for different prediction tasks. In each task the data points correspond to a test set from one particular (random) train/test sets split (Fig. 29a).

a



b

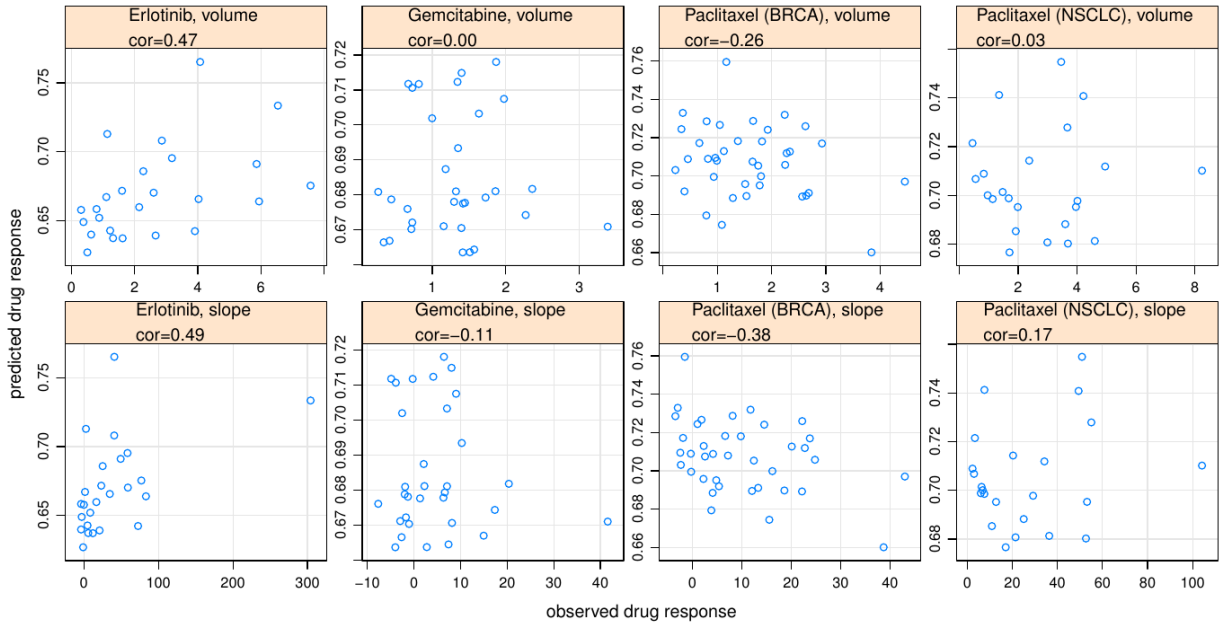


Figure 29. Observed vs. predicted plots (**a**) Observed vs. predicted values for different regression prediction tasks. (**b**) Observed vs. predicted values for [cell lines → xenografts] type of prediction and corresponding correlation coefficients, only data volume and slope response metrics are shown

Manual inspection of outlier cases from observed vs. predicted plots (Fig. 29a) shows that often a model's inability to provide an accurate prediction for certain samples (outliers) is driven by under-representation of samples with similar molecular characteristics to these outliers in the training set.

Also we tested how well the models trained on cell line data can explain drug response in xenografts. For that we 1) trained drug response models using cell line genomics and drug response data (AUC), 2) got model's predictions using xenografts molecular data as inputs, 3) assessed the correlation of resulting predictions (in cell line AUC units) with actual xenografts drug response. We tried 2 strategies with respect to training set composition -- training using all cell lines and training using only cell lines that match tissue type of corresponding set of xenograft samples. Results in terms of correlation coefficients as well as predicted vs. observed plots for this analysis (for the case where we used all cell lines for training) are shown in the Fig. 29b.
Among four xenograft response metrics used in this study volume and slope are expected to be positively correlated with cell line AUC while integral response (ΔAUC) and differential slope are expected to be negatively correlated. For both training strategies (all cell lines or cell lines from one relevant tissue type) we managed to get predictions with the right sign of correlation coefficient (between predictions and observed drug response) and substantial absolute value for all drug response metrics only for erlotinib. Volume and slope metrics worked especially good in case of erlotinib with correlation about 0.5 in both cases.

6 PATIENT TREATMENT OUTCOME PREDICTION USING CLASSIFICATION MODELS TRAINED ON CELL LINES

6.1 Introduction

The next logical step after predicting drug response in xenografts is to try to predict drug response/treatment outcome in patients. Following the study design proposed in the Zhao et al.[82] we compared prediction performance of models trained on cell line data and applied to patient expression data. We used the same 3 patients set that were used in the paper[82] -- myeloma cohort treated with bortezomib (proteasome inhibitor), NSCLC cohort treated with erlotinib (EGFR inhibitor) and BRCA cohort treated with docetaxel (inhibits cell division by binding to microtubules), see Fig. 30.

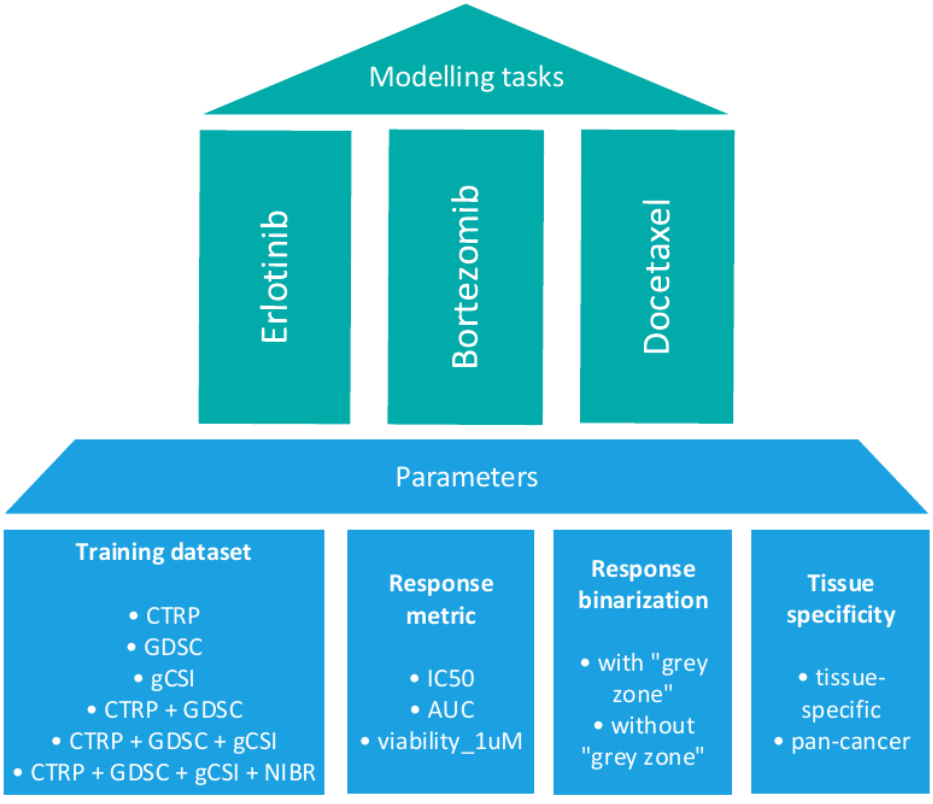


Figure 30. Overview of the analysis in the chapter 6. Modelling tasks (drug responses) and tested parameters (training dataset, response metric, response binarization strategy, tissue specificity) are shown.

6.2 Data and Methods

6.2.1 Data

For modelling we used molecular (microarray expression, copy number, mutation information) and drug response data from 4 large cell line sensitivity screenings -- CCLE[31], CTRP[32], GDSC[34], gCSI[36] and one xenograft screen -- NIBR PDXE[37]. Genomics and cell line data were obtained either from projects web portals (CCLE, CTRP, GDSC) or from corresponding papers' supplementary data (gCSI, NIBR PDXE).

For cell lines' drug response characterization we used IC_{50}, AUC and viability at 1uM metrics. For xenograft's response characterization we used xenograft response metrics from original publication[37] -- BestResponse and BestAverageResponse:

"The response was determined by comparing tumor volume change at time t to its baseline: % tumor volume change = ΔVolt = 100% × ((Vt – Vinitial) / Vinitial). The BestResponse was the minimum value of ΔVolt for t ≥ 10 d. For each time t, the average of ΔVolt from t = 0 to t was also calculated. We defined the BestAvgResponse as the minimum value of this average for t ≥ 10 d."

For each [cell line dataset + patient cohort dataset] pair we combined two corresponding expression sets, and for each of such combined sets we trained a classification model on cell line data and tested predictive performance on patient data.

In order to homogenise z-transformed expression data between cell line and patient sets we used ComBat method from "sva" package[83]. Cell line's continuous response (IC_{50}, AUC, viability at 1uM) and xenograft response (BestResponse, BestAverageResponse) were binarized to "sens" and "resist" labels. Treatment response labels for patient samples were obtained from paper's[82] RData (http://compbio.cs.toronto.edu/cp2p/) and converted to "sens" and "resist" values, criteria for patient response binarization are listed in the Table 14.

Table 14. Criteria for patient response binarization for each drug (from Zhao et al.[82]).

bortezomib	"In the original paper, response was measured in terms of change in paraprotein: patients were classified as achieving complete response (CR, with 100% decrease in paraprotein), partial response (PR, 50% decrease), minimal response (MR, 25% decrease), no change (NC - absence of response, 2 measures of stable disease), or PD (25% increase in paraprotein), using European Group for Bone Marrow Transplantation criteria. We grouped MR, PR and CR patients into a response category; NC and PD patients into nonresponse."
erlotinib	"Therapy response was defined as progression-free survival time of 2 or more months"
docetaxel	"Response to docetaxel neoadjuvant treatment was based on whether 25% of the tumor remained after four cycles of docetaxel"

6.2.2 Modelling

For feature selection we employed filter-type feature selection function anovaScores from caret package (see details in the section 1.7.2). After feature selection we fit the model with N selected features (with lowest p-values) on the training set data.

As a modelling method we used SVM with Radial base function (svmRadial). In order to select hyperparameters (sigma and C(cost)), 30 different combinations of them are tested on training data using 10 fold cross-validation, and then the combination that provides the lowest Balanced Accuracy (half sum of sensitivity and specificity) is used for fitting the final model. As the accuracy measures we used Balanced Accuracy and AUROC.

Feature selection, model fitting and accuracy evaluation were performed using the following procedure :

1. In each combined (cell-line + patient) dataset cell line samples are used as a training set (according to the state of tissue-specificity option either all or only cell lines from certain tissue are taken), patient samples -- as a test set.
2. We perform feature selection using the anovaScores function on the training set.
3. Then we fit the model with 200 selected features (with lowest p-values) on the training set data. Model's hyperparameters are selected using cross-validation testing.
4. We apply model to the test (patient) set, and calculate Balanced Accuracy and AUROC.

6.3 Results

For each of three drugs we tested several modelling options:

1) Training dataset -- single: CTRP, GDSC, gCSI, NIBR, and combined: CTRP+ GDSC, CTRP+GDSC+gCSI and CTRP+GDSC+gCSI+NIBR
2) Response metric -- IC50, AUC, Viability_1uM (binarized)
3) 2 types of binarization -- with and without intermediate class (grey zone)
4) 2 types of sample selection -- pan-cancer (all cell lines) and tissue-specific (only cell lines that match tissue of corresponding patient set)

Accuracy results for bortezomib turned out to be the lowest, with the balanced accuracy around 0.5 almost for all tested options, for erlotinib and docetaxel accuracy is generally higher. In the following tables we report average results for all tested options (Tables 15 -- 19).

Table 15. A combination of training options that produces the most accurate predictions for each of three drugs.

drug	dataset	Response metric	binarization	Tissue specificity	Balanced Accuracy	AUROC
bortezomib	ctrp_gdsc_gcsi	auc	With grey zone	pan-cancer	0.53	0.455
erlotinib	gdsc	auc	With grey zone	tissue-specific	0.80	0.786
docetaxel	ctrp_gdsc	auc	With grey zone	pan-cancer	0.83	0.87

Table 16. Average Balanced Accuracy for all datasets for each drug.

drug	CTRP	GDSC	gCSI	CTRP+ GDSC	CTRP+ GDSC+gCSI	CTRP+GDSC +gCSI+NIBR
bortezomib	0.48	0.47	0.47	0.49	0.49	--
erlotinib	0.52	0.52	0.41	0.505	0.53	0.53
docetaxel	0.46	0.55	0.46	0.67	0.59	--

Table 17. Average Balanced Accuracy for all response for each drug:

drug	IC50	AUC	viability
bortezomib	0.48	0.48	0.47
erlotinib	0.47	0.57	0.47
docetaxel	0.53	0.55	0.55

Table 18. Average Balanced Accuracy for two binarization options for each drug.

drug	Without “grey zone”	With “grey zone”
bortezomib	0.485	0.47
erlotinib	0.47	0.52
docetaxel	0.57	0.52

Table 19. Average Balanced Accuracy for two tissue-specificity options for each drug.

drug	pan-cancer	tissue-specific
bortezomib	0.48	0.47
erlotinib	0.53	0.46
docetaxel	0.57	0.51

We also compared AUROC values between our best models (judged by AUROC) and best models from Zhao et al. study[82] (they used only GDSC data for training):

Table 20. Area Under the ROC curve (AUROC) comparison between our models and models from Zhao et al.

drug	Our best models		Best models from Zhao et al.	
	IC50	AUC	IC50	AUC
bortezomib	0.60	0.57	0.62	0.68
erlotinib	0.68	0.83	0.73	0.73
docetaxel	0.88	0.89	0.91	0.91

Some conclusions:

- For different drugs different modelling options are beneficial.
- Datasets. Best single datasets for Erlotinib are CTRP and GDSC, and for Docetaxel it is just GDSC, performances of combined datasets are similar to best single ones for Erlotinib but it's significantly higher than performance of single best set for Docetaxel (CTRP+GDSC).
- Response Metrics. AUC works better than two other metrics for Erlotinib, for Docetaxel AUC and viability show just marginally higher accuracy than IC50.
- Grey zone. Average balanced accuracy for binarization with "grey zone" is higher than without for Erlotinib, but it is the opposite for Docetaxel.
- Tissue specificity. For both erlotinib and docetaxel pan-cancer models show better performance than tissue-specific.

7 APPLIED EXAMPLES OF DRUG RESPONSE DATA ANALYSIS

7.1 Introduction

In this chapter we show some examples of modelling or other kinds of drug response association analyses performed in collaboration frameworks. In the first result's subsection we discuss the analysis of the data from drug sensitivity screen of 42 blood cancer cell lines performed in the lab of Thorsten Zenz at NCT (National Center for Tumor Diseases, Heidelberg), the analysis consists of a descriptive part (drug-drug correlations, cell line-drug clustering, assessing synergisity of drug combinations) as well as a modelling part where we build and compare drug response models for a subset of the drugs used in the screen.
In the second result's subsection we describe the process of developing a drug response model for a TRXR1 inhibitor DKFZ-608 which was recently characterized in the lab of our collaborator Nikolas Gunkel.
In the third result's subsection we present a Shiny applications that visualizes samples' drug response data for a selected pair of drugs, the idea of this interactive visualisation was proposed by Thorsten Zenz.

7.2 Data and Methods

7.2.1 Data

In subsection 7.3.1 for modelling tests we use genomics and drug response data (AUC) from GDSC dataset.
For training drug response model fro DKFZ-608 described in the section 7.3.2 we use genomics data from GDSC dataset.
Shiny application described in the section 7.3.3 visualizes drug response (viability at individual concentrations) and genomic data from GDSC and CTRP datasets.

7.2.2 Modelling (in the section 7.3.1)

For feature selection we employed filter-type feature selection function gamScores from caret package (see details in the section 1.7.2). After feature selection we fit the model with N selected features (with lowest p-values) on the training set data.
As a modelling method we used SVM with Radial base function (svmRadial) and elastic net regression. As accuracy measures we used R^2 (explained variance) and RMSE (root of mean squared error).

Feature selection, model fitting and accuracy evaluation were performed using the following procedure:

1. We sort all samples by the values of drug response and then we take each third sample in the test set, therefore we get balanced (with respect to sensitive/resistant samples ratio) training (66%) and test (33%) sets.
2. We perform feature selection on the training set.
3. Then we fit the model with N selected features (with lowest p-values) on the training set data. Model's hyperparameters are selected using cross-validation testing.
4. We apply model to the test set, and calculate R^2 (calculated as a square of correlation between predicted and observed outcomes) and RMSE (root mean square error).

7.3 Results

7.3.1 Burkitt lymphoma drug sensitivity screen analysis

This section presents the analysis of drug sensitivity screen of 42 blood cancer cell lines which was done in collaboration with Katarzyna Tomska and Thorsten Zenz. This work is described in the paper Tomska et al.[78]

18 out of 42 cell lines were Burkitt lymphoma (BL) lines which is highly aggressive B-cell lymphoma associated with MYC translocation. Cell lines were profiled with 32 drugs, profiling included single-drug screen and drug-combination screen. I describe the analysis of this data below.

We summarised drug sensitivity data from multiple concentrations of each drug using IC_{50} and AUC metrics. (Fig. 20) In the subsequent analysis we used these two metrics as well as viability at individual concentrations.

First thing we assessed was correlations between drug responses across all drugs (using viability data across all concentrations and AUC data) This analysis confirmed screening platform consistency (individual concentrations from the same drug clustered together) and identified groups of drugs with similar drug response: gefitinib and lapatinib -- EGFR pathway inhibitors, saracatinib and dasatinib -- BCR-ABL/SRC inhibitors, idelalisib (PI3K), everolimus(mTOR) and MK-2206 (AKT) -- PI3K pathway, ibrutinib(BTK) and PRT062607 (SYK), idelalisib (PI3K) and PRT062607 (SYK), CHK inhibitor AZD7762 and DNA-PK inhibitor NU7441 (Fig. 31).

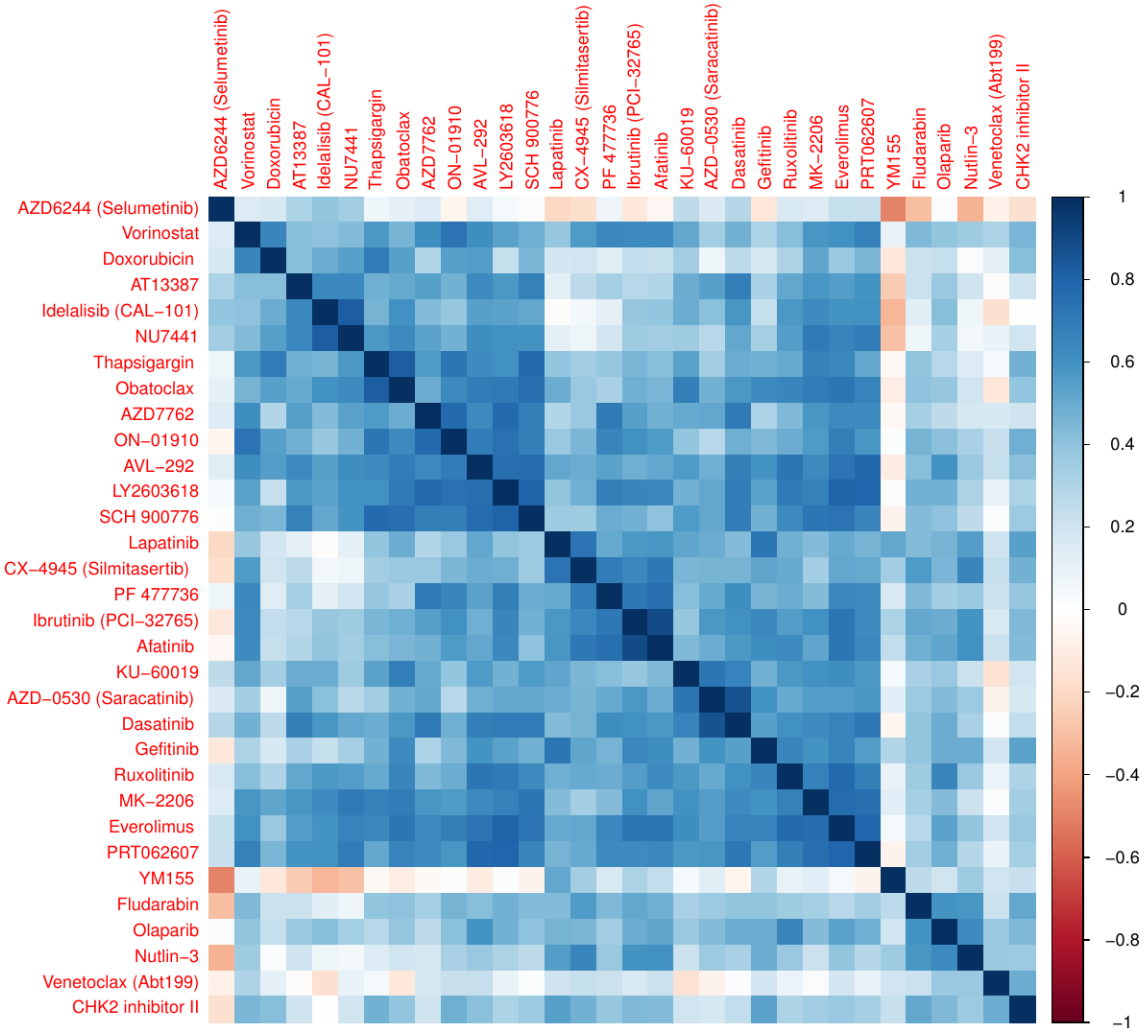


Figure 31. Correlation between drug response vectors across all drugs, based on AUC response values.

We were also interested in identifying subgroups of cell lines which show distinct drug sensitivity/resistance patterns. To this end we employed cell line-drug clustering (based on AUC and raw drug response values). We identified three clusters of response: Cluster I, Cluster II and intermediate group. (Fig. 32) Cluster I contained cell lines resistant to multiple drugs including inhibitors of PI3K and BCR pathways (ibrutinib, AVL-292, idelalisib, MK-2206, PRT062607 and everolimus). The cluster consisted of resistant BL and myeloma (MM) cell lines. Cluster II showed the strongest response to BCR and PI3K inhibitors and to a number of other inhibitors including SRC inhibitors (dasatinib, saracatinib). The intermediate group showed heterogeneous response driven by cell line-specific genotype context.

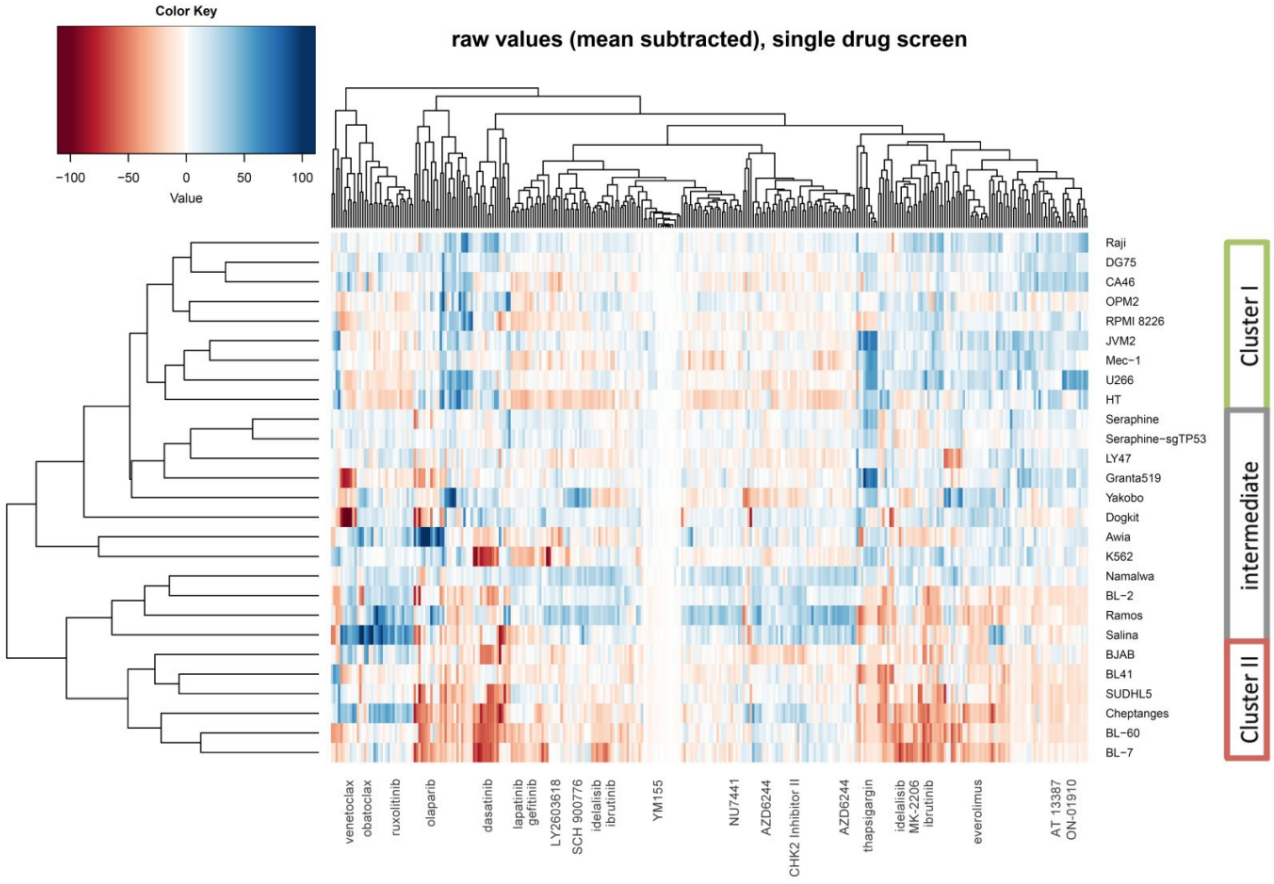


Figure 32. Heatmap with individual concentration raw viabilities across all drugs and all cell lines. Values plotted are raw viabilities-mean viability(across all cell lines for certain drug), range is [-100,100]. Cluster labels are shown next to cell line names.

In order to analyze results from combinatorial screen for each tested drug pair we calculated Combination Index (CI) which is measure of drug synergy/antagonism:

$$CI = \frac{C_{A,50}}{IC_{50,A}} + \frac{C_{B,50}}{IC_{50,B}},$$

where $IC_{50,A}$ and $IC_{50,B}$ are IC50 values for individual library drugs; $C_{A,50}$ and $C_{B,50}$ are concentrations of drug A and B at which they were used in combination which lead to 50% viability.[84]

We defined combinations with CI < 0.85 as synergistic, those with 0.85 < CI < 1.15 as additive and combinations with CI > 1.15 as antagonistic. To get an overview of combination effects in BL we clustered CI values across all drug pairs and cell lines (Fig. 33).

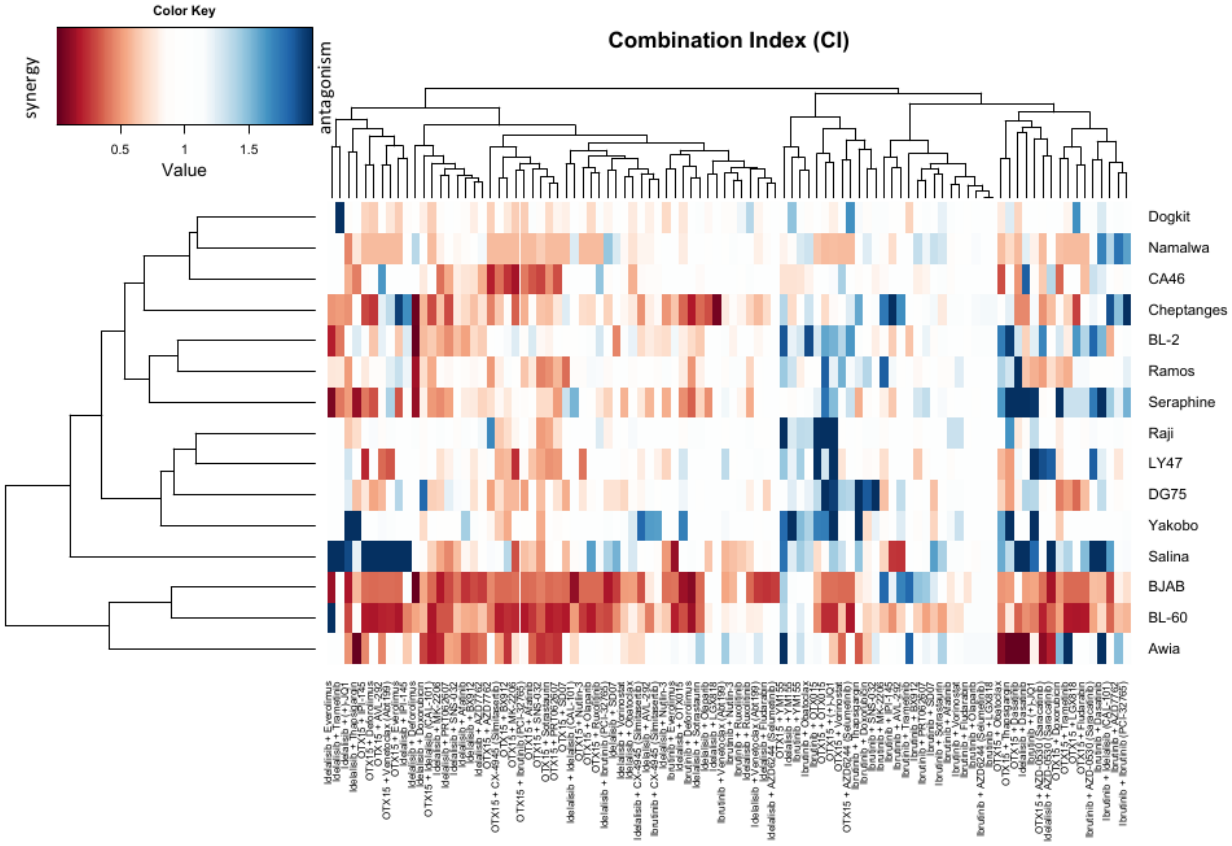


Figure 33. Combination Index values (CI) heatmap across all tested drug pairs and all cell lines. Blue color denotes synergistic pairs, red color denotes antagonistic pairs.

Combinatorial screen consisted of three types of drug combinations: combinations with ibrutinib (BTK inhibitor), with idelalisib (PI3K inhibitor) and with OTX015 (BET inhibitor).

The majority of drug combinations with ibrutinib were additive, lowest CIs were observed for its combinations with MK2206 (AKT inhibitor) and doxorubicin (DNA intercalating agent).

Idelalisib was synergistic in combinations with JQ1 (BET inhibitor), chemotherapeutics, MK2206 (AKT inhibitor), PRT062607 (SYK inhibitor).

Multiple strong synergistic effects were observed for OTX015. The lowest CI scores were found for SNS-032 (CDK2/7/9 inhibitor), inhibitors of PI3K/AKT/mTOR and BCR pathway inhibitors. Combinations of OTX015 with thapsigargin and YM155 were antagonistic. Also the important finding was that synergy for OTX015 with idelalisib, MK-2206, everolimus and SNS-032 were observed at concentrations *in vitro* that can be safely administered in the clinical setting in *vivo*.

We also tested drug response models for 20 drugs (see Figure 34 and Table 21) that were tested both in this screen and in GDSC study. For these tests we used molecular and drug sensitivity data from GDSC[34]. For training we used data either from all cell lines (pan-cancer training, ~1000 samples) or from lymphoma cell lines (lymphoma training, ~50 samples). In order to test prediction accuracy we applied

models to genomic data from test panel of 16 Burkitt lymphoma cell lines (those cell lines we tested in our screen and were profiled in GDSC study).

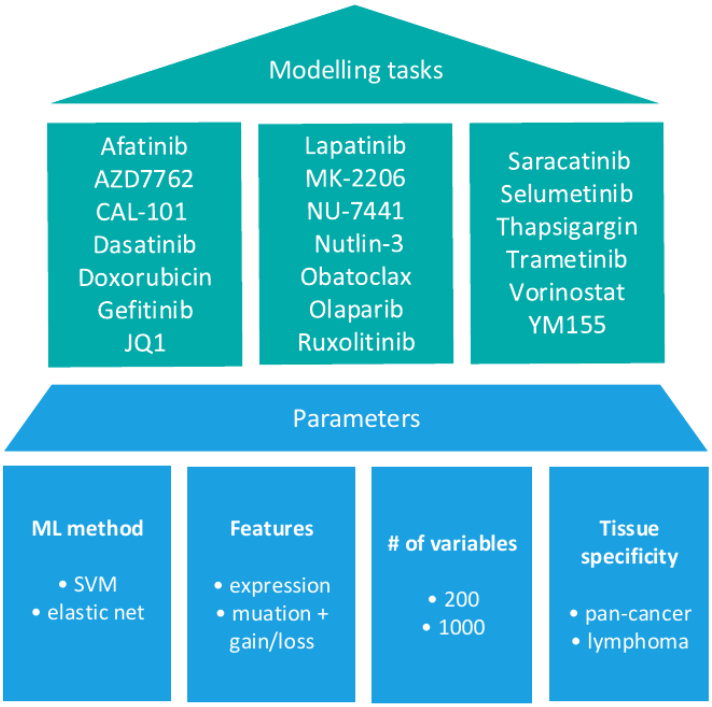


Figure 34. Overview of the model analysis in section 7.3.1. Modelling tasks (drug responses) and tested parameters (modelling method, feature type, # of variables, tissue specificity) are shown.

Table 21. List of drugs (with their molecular targets) used in the screen for which we built drug response models.

Drug	Molecular target	Drug	Molecular target
Afatinib	EGFR, ERBB2	Nutlin-3	MDM2
AZD7762	CHK	Obatoclax	BCL-2
CAL-101 (Idelalisib)	PI3K	Olaparib	PARP
Dasatinib	BCR-ABL	Ruxolitinib	JAK
Doxorubicin	Topo II	Saracatinib	SRC/BCR-ABL
Gefitinib	EGFR	Selumetinib	MEK
JQ1	BET	Thapsigargin	ATPase, Ca++ transporting
Lapatinib	EGFR, ERBB2	Trametinib	MEK
MK-2206	AKT	Vorinostat	HDAC
NU-7441	DNA-PK	YM155	survivin

We made two series of tests -- one using all the molecular features (expression, mutation, gain/loss and methylation information) and the other using only mutation and gain/loss features. In both series we tested SVM and Elastic net models with either 200 or 1000 variables with highest correlation with drug response. Average R^2 for tests with all molecular features and tests with only mutation + gain/loss features are present in the Table 22. Results for all drugs from tests with all molecular features are shown in the Fig. 35.

Table 22. Average R^2 for tests with all molecular features and tests with only mutation + gain/loss information

	ML method	# of variables	Pan-cancer training	Lymphoma training
All molecular features	SVM	200	0.092	0.114
		1000	0.099	0.118
	Elastic net	200	0.124	0.126
		1000	0.098	0.165
Only mutation + gain/loss features	SVM	200	0.084	0.084
	Elastic net	200	0.103	0.169

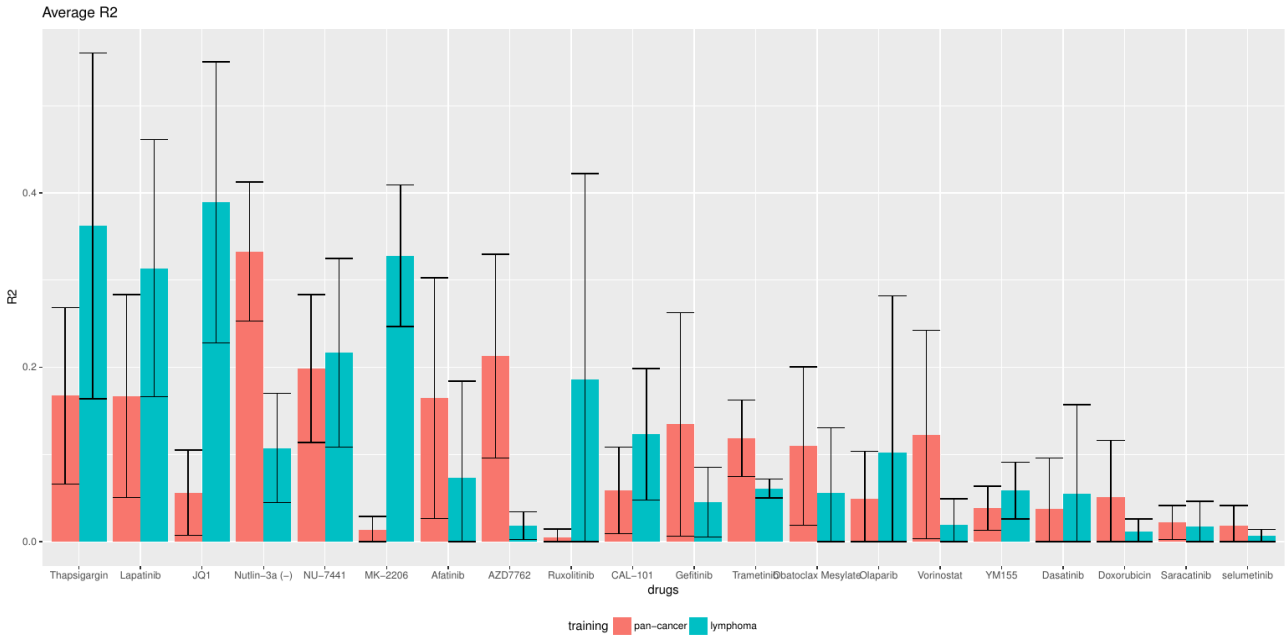


Figure 35. Average R^2 values for all drugs (models with all genomic features). Bar color reflects the training type: red -- pan-cancer training, blue -- lymphoma training.

We also assessed an influence of training set imbalance on model accuracy by correlating class imbalance for each drug with average R^2 for a drug. We calculated class imbalance using the formula: class imbalance= $|0.5 - \frac{\#\ sensitive}{\#sensitive + \#\ resistant}|$, where # of sensitive and # of resistant were calculated as number of samples with AUC below and above the median AUC value respectively. Results are shown in the Table 23.

Table 23. Spearman correlation between class imbalance and R^2

	Pan-cancer	Lymphoma
All genomic features (mostly expression features due to FS)	-0.06	-0.26
Mutation + gain/loss features	-0.14	0.16

Summary of the results:

- Lymphoma models almost in all cases outperform pan-cancer models
- Elastic net models outperform SVM models
- Models trained on mutation + gain/loss features only perform just slightly worse than models trained on all features (exp+mut+gain/loss)
- Depending on molecular data type used for training, sets of most explainable drugs are different:

  – using all(mostly expression) features: Thapsigargin, Lapatinib, JQ1, Nutlin-3, NU-7441, MK-2206, Afatinib, AZD7762

  – using mutation and gain/loss features: Nutlin-3, Doxorubicin, Selumetinib, Obatoclax, CAL-101, Thapsigargin
- For most of the drugs lymphoma-trained models give more accurate predictions than pan-cancer models. However for some drugs it's the opposite especially for Nutlin-3 , AZD7762, Afatinib, Trametinib.
- Class imbalance has a slightly negative influence on accuracy.

7.3.2 Drug response prediction model for DKFZ-608 compound

This section describes the work performed in collaboration with Nikolas Gunkel (Department of Drug Discovery, German Cancer Research Center) on characterization of DKFZ-608 chemical compound. DKFZ-608, a homoleptic dithiocarbamate gold complex, was discovered in Drug Discovery Department. It was characterized as TRXR1 (thioredoxin reductase) inhibitor with selective anti-SCLC (small cell lung cancer) activity[85].

Drug response to DKFZ-608 was measured in 75 cell lines (in the Department of Drug Discovery) and IC_{50} values were obtained. 43 out of these 75 cell lines were also genomically profilled in GDSC dataset, this allowed us to perform drug response-genomics associations analysis on these 43 cell lines. We assessed correlation between each individual genomic feature vector (expression/ copy

number/ mutation) and the vector drug response (i.e. vector of IC_{50} values). After correcting for multiple testing using “Benjamini-Hochberg” method we ended up with 307 gene expression features with adjusted p-values below 0.05.

We analysed this 307-gene list using DAVID annotation tool.[86] Identified functional clusters with highest enrichment scores included “Nucleus and transcription regulation”, “DNA damage/ DNA repair”, “cell cycle and mitosis”, “mRNA processing” and “zinc finger” (Table 24).

Table 24. Summarised results of DAVID Functional Annotation clustering.

Cluster	Annotation Terms	Enrichment score
1	Nucleus, transcription, DNA-binding	9.56
2	DNA damage/repair	3.05
3	Cell cycle	2.42
4	RNA binding	2.42
5	PAS domain	2.31
6	negative regulation of gene expression, epigenetic	2.25
7	Zinc fingers	2.22

Using this 307 gene expression features we built an ensemble drug response prediction model consisted of one Support Vector Machine (with Radial base function) and one Random Forest model and predicted IC_{50} values for remaining 937 GDSC cell lines (i.e. for those cell lines which were not tested with DKFZ-608 compound). Later 14 of these 937 cell lines were tested with DKFZ-608 and ranking of observed IC_{50} values was consistent with ranking of corresponding predicted IC_{50} values, R^2=0.36. When we looked at PCA plots for these newly tested 14 NSCLC (non-small cell lung cancer) cell lines based on either all genomic features or selected 307 gene expression features (see Fig. 36) we realised that PCA plot based on all genomic features achieves better separation between samples with high and low IC_{50} values compared to the PCA plot based on 307 expression features which suggests that a refined model trained on combined set of cell lines (43 + 14) will provide more accurate predictions.

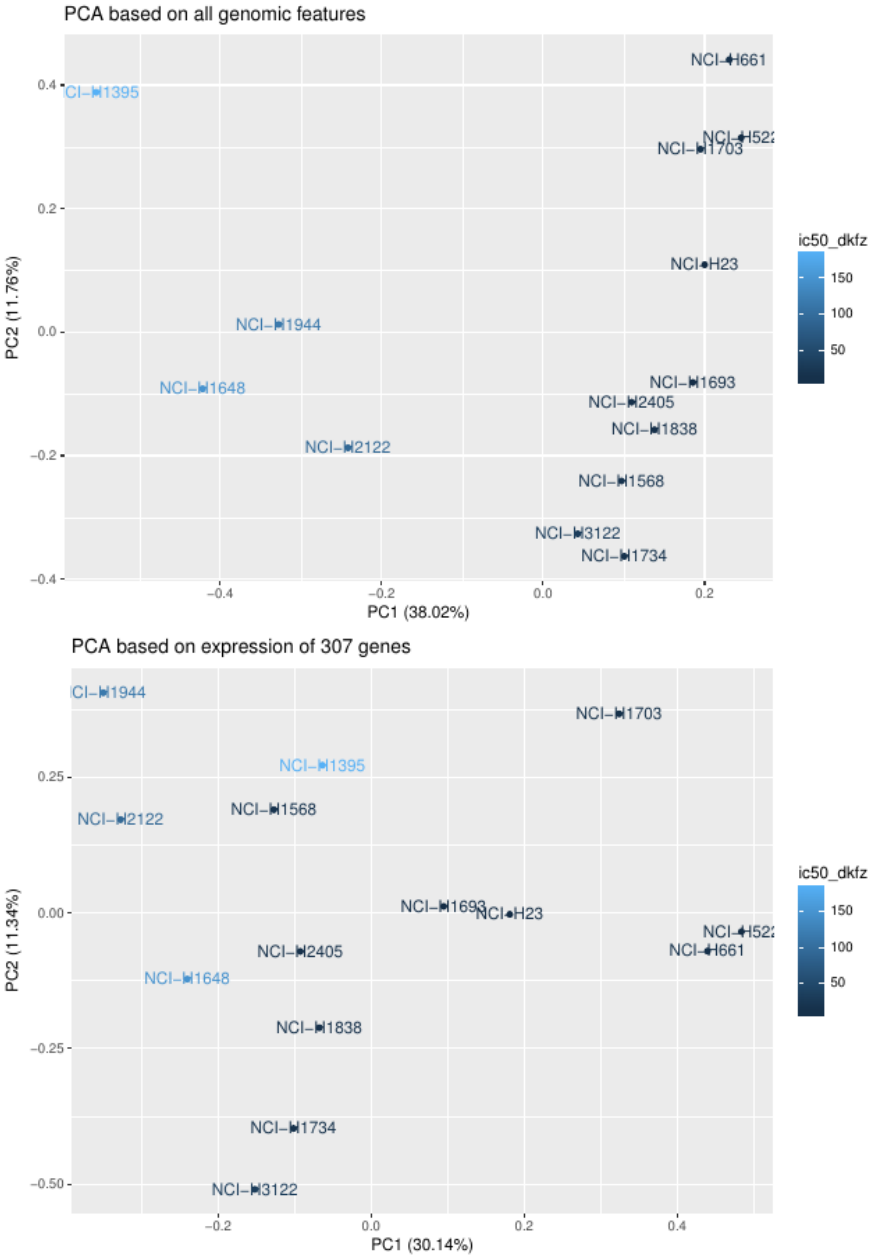


Figure 36. PCA plots showing NSCLC cell lines clustering. Color of the cell line label shows the level of corresponding IC_{50} value. Upper PCA plot based on all genomic features, lower PCA plot based on 307 expression features previously selected for the drug response model.

In addition we also identified drugs that produce similar response profiles by assessing the correlation between DKFZ-608 IC_{50} vector (for 43 cell lines) and IC_{50} vectors from other drugs from GDSC dataset. The highest correlation (r=0.66) was found for Navitoclax, inhibitor of Bcl-2 family proteins. Also using expression data before and after treatment with DKFZ-608 in a number of cell lines (H209, Jurkat, Raji) we calculated expression signatures of the drug, i.e. obtained lists of genes that become up- and down-regulated upon treatment. We queried LINCS database[87] (clue.io) using these lists and obtained information about compounds that produce similar expression changes. Classes of identified the most similar compounds included proteasome inhibitors, heat shock proteins inhibitors and NFkB inhibitors.

7.3.3 Shiny application for complex drug response visualization

In order to analyse the difference between cell lines' drug responses for a given pair of drugs (drug A vs. drug B) I was working on the corresponding drug response data visualization using Shiny technology. In the resulting shiny application user can define a dataset (GDSC or CTRP), a pair of drugs of interest (with an option of filtering drugs by molecular target) and cell lines with tissue filtering.

Additionally the app provides an option of assessing molecular differences between cell lines that might drive observed difference in drug response. One can manually define two clusters on the plot and obtain a list of molecular features with the most significant differences between the two clusters ranked by t-test p-value.

The application is available at shinyapps.io repository: https://drugs.shinyapps.io/drug_pair_ind_conc/

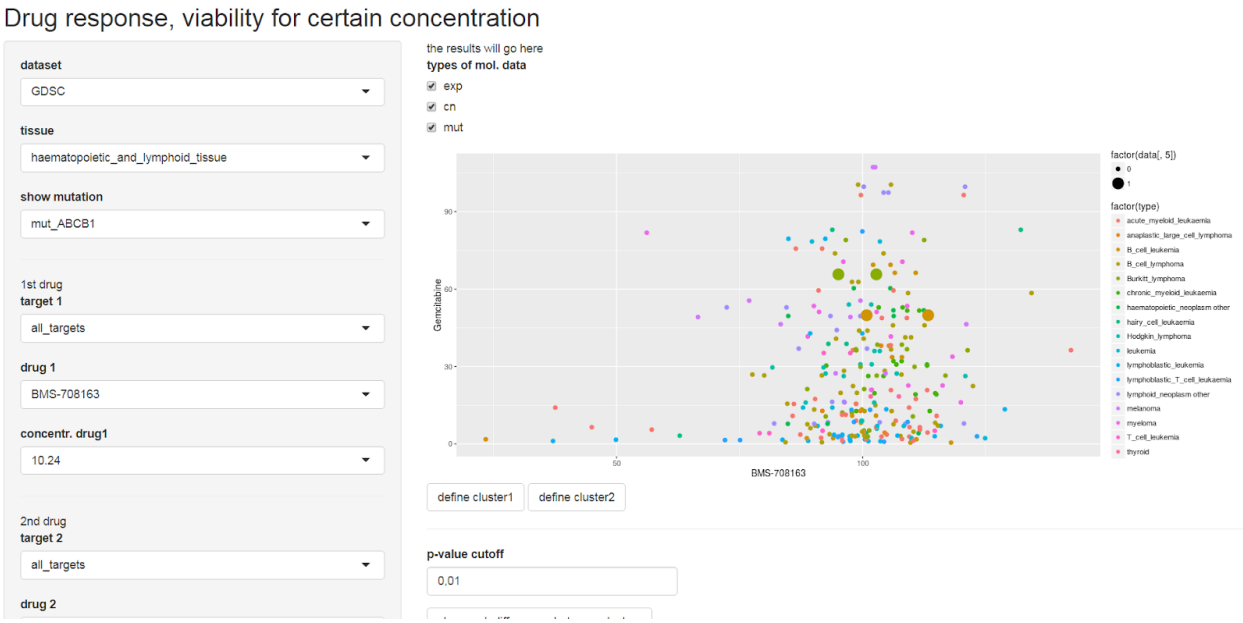


Figure 37. Interface of the Shiny application. Left panel allows a user to select the required data, plot on the right side is updated interactively.

8 DISCUSSION

8.1 Improving accuracy of drug response prediction in cell lines

8.1.1 Machine learning methods

With the aim of improving accuracy of drug response predictions we tested a number of various model training approaches. I review and discuss the results here.

Utilising information from drugs that share the same target via multi-task learning or learning on aggregated data. We applied multi-task learning (using glmnet models) and learning on aggregated data (using Random Forest models) in order to co-utilize information on genomic-drug response associations for groups of drugs that share the same molecular targets (Table 5). Comparing performance of both approaches with a standard single-task approach we haven't observed an improvement in predictive accuracy for multi-task models or models built on aggregated data. One possible explanation for this lack of improvement is that drugs can have the same target, but at the same time may exhibit differences in their inhibition profile[88] (i.e. have a different spectrum of additional targets). Therefore, just combining data from a group of drugs for model training should not necessarily help to identify more accurate associations and may not result in accuracy improvement. Simply getting more samples tested with the same drug would be more productive in terms of accuracy improvement (see discussion on training set sizes in the next subsection "Training set properties").

Modelling with feature interactions. Feature engineering is an important machine learning concept. Essentially feature engineering is the process of constructing new features from original features (i.e. from original data) with the idea that new features will provide better model performance. Here we tried two types of engineered features based on interactions between gene expression features – binary gene pairs (BGP) and gene multiplications. Due to combinatorial complexity we were able to test only a small subset of all possible interactions, i.e. we focused on interactions between only 200 gene expression features with the highest correlation with outcome (drug response). Even with a limited initial set of features, BGP features showed a modest advantage in accuracy over original features when we used the random forest method for modelling (with elastic net method neither BGP nor gene multiplication features showed an advantage in accuracy). Typically, feature engineering approach can bring an advantage when it is used for incorporating domain knowledge, for example one could preselect genes for constructing new features not only by taking features with the highest association

with drug response but also taking all genes associated with cancer, e.g. from the Cancer Gene Census.[89]

Class imbalance. We examined how class imbalance (between sensitive and resistant samples) in training data influences the resulting accuracy of drug response prediction, using lasso models built for the group of 19 drugs (Table 5), and elastic net and SVM models built for the group of 20 drugs (Table 20). In both cases the correlation between a measure of class imbalance and R^2 were negative but relatively small (less than -0.1). This may explain why no accuracy improvement was achieved when we tried modelling with weights where higher weights were assigned to under-represented sensitive samples. Therefore we would conclude that the level of drug-sensitivity data imbalance observed in the studied large pharmacogenomics datasets doesn't constitute a problem for predictive performance.

Choice of machine learning algorithm. We compared accuracy of predictions (generated for seven drugs from CCLE dataset) between our method based on SVM, method from CCLE paper[31] based on elastic net and the second top-performing method from the DREAM challenge[81] that utilises random forest (see Table 12). The comparison shows that there is no a single method that outperforms the others for all drugs.
Also we observed that in cases when all preprocessing, feature selection and model evaluating steps are kept the same and only modelling method (i.e. machine learning algorithm) changes, the resulting accuracy of drug response prediction shows a relatively modest variation. If we look at the models for erlotinib response in cell lines, the average R^2 for random forest models (tested on gCSI data; section 3.4) is comparable with average R^2 for Support Vector Machines models (tested on data from CCLE, CTRP and GDSC; section 3.3): R^2_{RF}=0.23, R^2_{SVM}=0.19.

8.1.2 Training set properties

In the section 3.3 we described the analysis performed on the data from three largest pharmacogenomics datasets where we assessed the influence of different properties of training set on resulting model's accuracy, now let's review the main findings.

Number of features. According to our results the number of top features selected for modelling (via filter-based feature selection), within the tested range of 10-500 features, doesn't influence the resulting accuracy of predictions, which rather depends on the strength of correlation between top feature(s) and the outcome. Also we observed similar outcome when we compared models with 200 and 1000 features that were tested on Burkitt lymphoma lines (Table 21, section 3.6.1).

Type of molecular data. We found that expression information has the highest predictive power compared to other data types (i.e. mutation and copy number information). This was also confirmed when we compared models with only

expression information and models with only mutation + gain/loss information on Burkitt lymphoma lines (Table 21, section 3.6.1). The same finding was observed in the analysis of DREAM challenge methods.[47] Probably expression information can explain drug response better than mutation and copy number information because it's functionally closer to phenotypic level.

Drug response metric. The choice of a drug response characterization metric has a serious impact on the accuracy of predictions. Having compared three drug response metrics for cell lines we found that the area under the drug response curve (AUC) provides the highest predictive performance (R^2_{IC50}=0.111, R^2_{AUC}=0.186, $R^2_{viability\_1uM}$=0.162). AUC combines information about drug efficacy and potency into a single value, and it was reported to be the robust metric previously.[90]

Size of training set. We also showed that the size of the training set is an important determinant for the accuracy of a model. In our tests based on gCSI data the average R^2 for models trained on cell lines from all tissues (n=329) was 0.267, while for models that used only cell lines from a certain tissue for each drug (n=26-68), the average R^2 was 0.102. Also results from the section 3.6.2 where we describe the process of model development for DKFZ-608 compound suggest that additional samples used for training lead to better model performance.

Accuracy across drug panel. We observe that the accuracy of drug response prediction varies across the drugs, e.g. in our tests based on CCLE, CTRP and GDSC datasets and the AUC metric average R^2 ranges from 0.06 for Sorafenib to 0.27 for PLX4720. Accuracy is quite heterogeneous across the drugs in the results in sections 3.2 and 3.6.1 as well.

8.1.3 Comparing drug response prediction with other prediction tasks

In the section 3.4 we described the analysis where we compared the task of drug response prediction with "positive control" tasks, namely tissue type prediction and prediction of cell doubling time. Now let's review the results.

We found that tissue type classification can be achieved with relatively high accuracy. Percentage of correctly predicted samples is 0.79 for cell line set, and 0.89 for xenograft set. Accuracy of prediction depends substantially on the size of training set for each tissue class. When we additionally included tissues which have from 10 to 20 samples to our cell line modelling set, the average accuracy dropped from 0.79 to 0.64.

While accuracy of tissue type classification is high on average, it varies across tissues. In a series of tests where we use equal number of samples per tissue (16 samples in training set and 7 samples in test set), we found that we have the lowest accuracy for lung samples, higher accuracy for breast samples, and the highest accuracy for pancreas, colon, skin and blood samples (Fig. 28). Interestingly this

ranking holds for both cell line and xenografts predictions. A fraction of lung samples is often misclassified as breast samples (and to a lesser extent the other way around) which results in comparatively lower accuracies for lung samples. This can be explained by the partial overlap between features that separate lung and breast samples from other samples, particularly expression of some transcription regulators genes and membrane protein genes.

The second cellular phenotype we tried to predict was the cell line doubling time or slope of untreated tumor growth curve in case of xenografts. Here we got lower accuracy compared to the tissue type prediction. Average accuracy is quite consistent between cell lines and xenografts: $R^2_{cell\ lines}$=0.17, $R^2_{xenografts}$=0.19. The lower accuracy of prediction shows that unlike for tissue type there is less information about speed of cell division in the static expression data, which can be due to the post-translational regulation of cell cycle proteins activity.

Thus, depending on cellular phenotype we want to predict with genomic data, the accuracy of prediction varies substantially. While expression data contain enough information to predict tissue of cell line with high accuracy, the accuracy for prediction of more complex dynamic phenotypes like doubling time or response to treatment is substantially lower.

8.1.4 Cross-set consistency

As we discussed in the introduction (section 1.6) the problem of consistency between pharmacogenomics dataset is crucial for biomarker discovery and consequently for drug response modelling. Indeed in the setting when we build a model using one dataset and assess model's accuracy on another one the level of consistency between the two sets determines the accuracy of prediction, correlation between vector of cross-set AUC correlation values and vector of R^2 values for 19 drugs was ~0.9 (section 3.2.5).

In an attempt to battle cross-set inconsistency we tested two methods, simple cell line filtering and correction for general level of drug sensitivity (GLDS), GLDS method was proposed in the study Geeleher et al.[79] The idea of the method is to separate a drug-specific part of drug response signal from a general part which reflects intrinsic properties of cell line like tissue of origin, division rate, whether cell is primed to apoptosis, drug accumulation/afflux properties. The separation of drug-specific signal is achieved by taking into account drug responses from drugs that have unrelated mechanism of action to the drug in question.
Firstly we tried to improve consistency between drug response data (AUC) alone. Cell line filtering improved consistency for almost all tested drugs at the obvious cost of reducing sample size (because of the cell lines that were filtered out). GLDS, in the simple form of subtracting mean of AUC values of unrelated drugs from AUC of the drug in question, didn't improve the consistency.
Secondly we applied GLDS to improve biomarkers' (i.e. genomic feature - AUC association) consistency. Here we used GLDS version based on regression model

where outcome is a drug response and in addition to our main covariate, a biomarker, we add 10 additional covariates which are 10 principal components calculated on matrix composed of drug response vectors from unrelated drugs. We compared biomarkers' association cross-set consistency across GDSC-CTRP-NIBR PDXE sets and across GDSC-CTRP-gCSI sets for common drugs between these groups of studies. Consistency was improved but only for some subsets of the drugs and datasets (e.g. for two out of three datasets). The fact that we have less improved cases between GDSC and either CTRP or gCSI compared to CTRP - gCSI indicates that it's difficult to overcome the inconsistency that originates from difference in experimental techniques -- in GDSC study viability assays based on DNA content were used, in CTRP and gSCI viability assays were based on metabolic content.[36]

To conclude, inconsistency between pharmacogenomics datasets remain to be an important issue, and it seems that there is no universal computational solution for it. Users of pharmacogenomics data should be aware that due to the difference in experimental assays and data analysis techniques associations between genomic data and drug response may exhibit certain level of inconsistency.[59]

8.2 Using models trained on cell line data for drug response prediction in xenografts and patients

The ultimate goal of pharmacogenomics research is to learn how to accurately predict drug response for patients. In the sections 3.4 and 3.5 we described the analyses of applicability of models trained on cell line data for drug response predictions in xenografts and patients respectively. Let's now discuss these results.

8.2.1 Xenografts

As in the case of cell lines we found that the way of drug response quantification matters for resulting model performance. We compared different xenograft drug response metrics (Fig. 27) and found that simple metrics like "tumor volume (day 21)" and "slope" perform better (average R^2=0.23 and 0.25) than those which additionally take into account data from untreated controls -- "Integral response" and "Differential slope", average R^2=0.095 and 0.145 respectively (in this comparison xenograft data was used for model training and testing).

Armed with these four response metrics we assessed our ability to predict drug response in xenografts using models trained on cell line data. We tested our predictions in four cohorts -- NSCLC treated with erlotinib, PDAC treated with gemcitabine, BRCA and NSCLC treated with paclitaxel. Only for the NSCLC cohort treated with Erlotinib our predictions were moderately accurate, i.e. positively correlated (r=0.5) with tumor volume and slope of the tumor growth curve (Fig. 29b). Performance of the models built and tested on xenograft and cell line data separately can't explain why this type of prediction worked only for erlotinib-treated cohort. Indirectly it shows that pharmacogenomic associations were consistent

between cell line and xenografts dataset only in the case of erlotinib, which has only one target-molecule EGFR, but not in the cases of more pleiotropic gemcitabine and paclitaxel which block DNA synthesis and cell division respectively.
It's important to note a data preprocessing aspect of this analysis, since cell expression data was profiled with microarrays, and xenograft data -- using RNA-seq (and available as fpkm values) we had to harmonise expression values between the two platforms. For that we applied z-score normalization (i.e subtracting the mean and dividing by the standard deviation of the samples) for each gene in cell line and xenograft datasets. We also tested a quantile normalisation with a different number of bins but overall prediction results didn't improve.

So far NIBR PDXE dataset is the only publicly available large scale pharmacogenomic xenograft study. With more xenografts and patient material screens available in the future, it will be possible to understand in which cases drug response associations are transferable between cell lines and xenografts/patients and in which cases they are not.

8.2.2 Patients

In the section 3.5 we described the process of applying classification models built on cell line data for prediction of treatment outcome (sensitive to treatment vs. resistant to treatment) in three different patient cohorts. By testing a range of models we identified a model with highest prediction accuracy in each cohort. While results for bortezomib were quite poor (balanced accuracy of the best model=0.53), best models for erlotinib and docetaxel showed reasonably good accuracy (balanced accuracy ~0.8 in both cases)

It's important to note that classification results depend on the way we binarize patient's therapy response into two classes. This binarization is based on arbitrary criteria which are different for each cohort (see Table 14). Also there are multiple choices with respect to cell line drug response variable that is used for training the model. Similarly to our tests based on cell lines, AUC values selected for cell line response provided better classification on patients than IC50 values. We tested two type of binarization for cell line response -- without and with a "grey zone", i.e. whether we binarize the whole spectrum of IC_{50} or AUC values, or we take for model training only samples with high and low IC_{50}/AUC values; the models that provided highest prediction accuracy were trained with "grey zone" option.

Comparing pan-cancer models with tissue-specific models haven't lead to a conclusive results, best model for erlotinib was trained on tissue-specific samples (lung samples) while best model for docetaxel was trained on samples from all tissues. Tissue-specific samples on one hand provide more relevant associations, but on the other hand samples from different tissues show more genetic variance which allows to identify genomic-response associations that are not seen in a smaller less heterogeneous group of samples.

At least for Docetaxel combining training data from different datasets results in a serious accuracy improvement (Table 16), which confirms our observation that training size is a determinant of model's accuracy. Since training cell line data and patient expression data were coming from different Affymetrix microarray platforms, we used ComBat method from sva package to reduce the batch effect. According to PCA plots before and after batch effect correction expression values changed just to a very little extent, therefore we hypothesise that applying different approach for data homogenization might result in improved classification accuracy.
Overall we see that applying cell line data for treatment response prediction in patients include many nuances with respect to data preprocessing and model training but in the end at least for some drugs it's possible to come up with models that predict binary treatment outcome with reasonable accuracy.

8.3 Conclusions

With a growing amount of pharmacogenomics data from model organisms (as well as patient genomic data) there is an increasing need in understanding how to integrate and extract meaningful information from this data. Concretely pharmacogenomics data can be used for generating machine learning models of drug response which, assuming the certain level of accuracy, can help in stratifying patient cohorts for clinical trials and in selecting efficient treatment strategy for individual patient. Our work provide some guidance for training and testing strategies for such models.

We found that generally model's predictive power depends on the type of molecular data, selected drug response metric, and the size of training set. It depends less on number of features (more important is the strength of correlation between top feature(s) and the outcome) and on class imbalance in training set. While predictive power varies across drugs (i.e. across models for different drugs) models built for the same drug using different machine learning methods accuracy usually show similar level of accuracy. We also found that unlike drug response tissue type can be predicted in cell lines and xenografts with quite high accuracy. Testing our ability to correctly predict response in xenografts and patients using models trained on cell lines produced positive results only in a fraction of tested cases, one of the positive examples was response to Erlotinib which was predicted with reasonable accuracy in xenograft and in patient cohort (two independent cohorts).

Possible reasons for low predictive power of drug response models were concisely summarised in the recent review from Kalamara et al.[19]:
"(i) noise in the data, and the aforementioned (ii) relative low number of samples when compared to the features, (iii) incomplete omics characterization, in particular in terms of proteomic and metabolomics, and (iv) limited readouts. Noise in the data can be either biological or technical and while methods to correct for technical variation are developed allowing us to bring datasets of different platforms together, the discussion on what is the correct way to apply them is not settled, adding uncertainty to downstream results. The low amount of samples toughens the discrimination of the true signal from noise. Incomplete "omics" characterization leaves open the possibility that the answer to our questions lies in

the things that have not been measured. Finally, the number of readouts is very relevant as subsequently the data used for models is vastly static (prior to treatment), while the effect of the drug is a dynamic process, whereby the drug modulates molecular components of the cell, that responds to this as an integrated system, as the target of the drug is often embedded in a complex molecular network that includes multiple pathways, crosstalks among them, and feedbacks."

Some of the outlined problems actually reflect the directions for data generation which is already happening and probably will continue in the future:

More functional and genomics data. More cell line screening data will be generated,[91] including drug combination screens (as the screen we describe in the section 3.6.1). Also data from other models is becoming available i.e. ex vivo profiling of patient material[25,26], patient-derived organoid and patient derived xenograft screens[9,23,92] (PDO and PDX). In addition the amount of patient sequencing data will grow, and its integration with clinical data (NGS-EHR integration) will provide invaluable resource for biomarker discovery and validation.[93]

Richer multi-omics characterization. Methylation, proteomics and metabolomics profiling are already being used in combination with drug screens and will become more abundant.

Richer drug response characterization. High-throughput imagine assays will provide a compliment to viability assays giving more detailed information on cellular changes upon the drug exposure.[94] Also expression changes characterization will continue to play an important role in describing cellular response to treatment.[87]

To conclude, in this study we used data from largest publicly available cell lines and xenograft pharmacogenomics screens to elucidate accuracy determinants of drug response prediction models. The amount of various multi-omics and drug response data in model organisms and humans will continue to grow and so will the opportunities for data integration as well as corresponding challenges. Gaining the ability to extract value out of this information in order to build models of drug response capable of accurate treatment outcome prediction in humans will facilitate translation of research findings into clinical practice and bring closer the era of personalised cancer medicine.

REFERENCES

1. Pardee, A. B., & Stein, G. S. *The biology and treatment of cancer: Understanding cancer.* John Wiley & Sons. (2011)
2. Newman, W.G. ed. *Pharmacogenetics: Making cancer treatment safer and more effective*. Heidelberg, Germany:: Springer. (2010)
3. Palumbo, M.O. et al. Systemic cancer therapy: achievements and challenges that lie ahead. *Frontiers in pharmacology*, *4*, p.57. (2013)
4. Syn, N.L., Teng, M.W., Mok, T.S. and Soo, R.A. De-novo and acquired resistance to immune checkpoint targeting. *The Lancet Oncology*, *18*(12), pp.e731-e741. (2017)
5. Chabner, B. A., & Roberts Jr, T. G. Chemotherapy and the war on cancer. *Nature Reviews Cancer*, *5*(1), 65. (2005)
6. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339 (2013)
7. Hudson, T. J. *et al.* International Cancer Genome Consortium. International network of cancer genome projects. *Nature* 464, 993–998 (2010).
8. Garraway, L. A. Genomics-driven oncology: framework for an emerging paradigm. *Journal of Clinical Oncology*, *31*(15), 1806-1814. (2013)
9. Sun, J. et al. A systematic analysis of FDA-approved anticancer drugs. *BMC systems biology*, *11*(5), 87. (2017).
10. Santos, R., et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, *16*(1), 19. (2017).
11. Web resource: https://www.mycancergenome.org/content/molecular-medicine/overview-of-targeted-therapies-for-cancer/
12. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* 4, 177–183 (2004).
13. Workman, P. & Al-Lazikani, B. Drugging cancer genomes. *Nat. Rev. Drug Discov.* 12, 889–890 (2013).
14. Luo, J., Solimini, N. L. & Elledge, S. J. Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell* 136, 823–837 (2009).
15. Hanahan, D., & Weinberg, R. A. Hallmarks of cancer: the next generation. *cell*, *144*(5), 646-674. (2011)
16. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nature reviews. Cancer* 13, 714–726, doi: 10.1038/nrc3599 (2013).
17. Masui, K. et al. A tale of two approaches: complementary mechanisms of cytotoxic and targeted therapy resistance may inform next-generation cancer treatments. *Carcinogenesis*, *34*(4), pp.725-738. (2013)
18. Fisher, R., Pusztai, L., & Swanton, C. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, *108*(3), 479. (2013)
19. Kalamara, A., Tobalina, L. and Rodriguez, J.S. How to find the right drug for each patient? Advances and challenges in pharmacogenomics. *Current Opinion in Systems Biology*.(2018)
20. Goodspeed, A., Heiser, L.M., Gray, J.W. and Costello, J.C. Tumor-derived cell lines as molecular models of cancer pharmacogenomics. *Molecular Cancer Research*, pp.molcanres-0189. (2015)

21. Sachs, N. & Clevers, H. Organoid cultures for the analysis of cancer phenotypes. *Curr. Opin. Genet. Dev.* 24, 68–73 (2014).
22. Friedman, A.A., Letai, A., Fisher, D.E. and Flaherty, K.T. Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer*, *15*(12), p.747. (2015)
23. Bruna, A. et al. A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. *Cell*, *167*(1), 260-274. (2016)
24. Letai, A. Functional precision cancer medicine—moving beyond pure genomics. *Nature medicine*, *23*(9), p.1028. (2017)
25. Dietrich, S.,et al.. Drug-perturbation-based stratification of blood cancer. *The Journal of clinical investigation*, *128*(1), pp.427-445. (2018)
26. Pemovska, T.,et al.. Individualized Systems Medicine (ISM) strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer discovery*, pp.CD-13. (2013)
27. Chan, G. K. Y., Kleinheinz, T. L., Peterson, D., & Moffat, J. G. A simple high-content cell cycle assay reveals frequent discrepancies between cell number and ATP and MTS proliferation assays. *PloS one*, *8*(5), e63583. (2013)
28. Gillet, J.P., Varma, S. and Gottesman, M.M. The clinical relevance of cancer cell lines. *Journal of the National Cancer Institute*, *105*(7), pp.452-458. (2013)
29. Shoemaker, R.H. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, *6*(10), p.813. (2006)
30. Haibe-Kains B, Large-scale in vitro Drug Screening and Applications, *talk* at Distinguished Scientist Lecture Series, Institute for Research in Immunology and Cancer, Université de Montréal, 2017 March 6, slides: https://www.pmgenomics.ca/bhklab/research/presentations
31. Barretina, J., *et al*. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483.7391**,603 (2012).
32. Seashore-Ludlow, B., *et al*. Harnessing connectivity in a large-scale small-molecule sensitivity dataset." *Cancer discovery* **5.11**, 1210-1223 (2015).
33. Jiang, P., Sellers, W.R. and Liu, X.S., Big Data Approaches for Modeling Response and Resistance to Cancer Drugs. *Annual Review of Biomedical Data Science,* 1, pp1-27 (2018)
34. Iorio, F., *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166.3**, 740-754 (2016).
35. Lavertu, A. et al. Pharmacogenomics and big genomic data: from lab to clinic and back again. *Human molecular genetics* 27.R1, R72-R78 (2018)
36. Haverty, P. M., *et al*. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533.7603**, 333 (2016).
37. Gao, H., *et al*. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature medicine* **21.11**, 1318 (2015).
38. Ali, M. and Aittokallio, T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical reviews*, pp.1-9. (2018)
39. Fang, Y., et al. DISIS: prediction of drug response through an iterative sure independence screening. *PloS one* **10.3**, e0120408. (2015)
40. Geeleher, P., Cox, N.J. and Huang, R.S. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome biology*, ***15*(3)**, p.R47. (2014)
41. Geeleher, P. et al. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome research* (2017).

42. Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H., & Margolin, A. A. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Biocomputing 2014*, pp. 63-74 (2014).
43. Falgreen, S., et al. Predicting response to multidrug regimens in cancer patients using cell line experiments and regularised regression models. *BMC cancer* **15.1**, 235. (2015)
44. Aben, N., Vis, D. J., Michaut, M., & Wessels, L. F. TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, ***32*(17)**, i413-i420 (2016)
45. Dong, Z., et al. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC cancer* **15.1**, 489 (2015)
46. Hejase, H.A. and Chan, C. Improving drug sensitivity prediction using different types of data. *CPT: pharmacometrics & systems pharmacology*, *4*(2), pp.98-105. (2015)
47. Costello, J. C., *et al*. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology* **32.12**, 1202 (2014).
48. Ali, M., Khan, S.A., Wennerberg, K. and Aittokallio, T., Global proteomics profiling improves drug sensitivity prediction: results from a multi-omics, pan-cancer modeling approach. *Bioinformatics*, *34*(8), pp.1353-1362. (2017)
49. Ammad-Ud-Din, M., et al. Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *Journal of chemical information and modeling* **54.8**, 2347-2359 (2014)
50. Ammad-Ud-Din, M., et al. Drug response prediction by inferring pathway-response associations with kernelized Bayesian matrix factorization. *Bioinformatics* **32.17**, i455-i463 (2016)
51. Riddick, G. et al. Predicting in vitro drug sensitivity using Random Forests. Bioinformatics, 27(2), pp.220-224. (2010)
52. Nguyen, L., Dang, C.C. and Ballester, P. Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. *F1000Research*, *5*. (2016)
53. Rahman, R., Matlock, K., Ghosh, S., & Pal, R. Heterogeneity aware random forest for drug sensitivity prediction. *Scientific Reports*, *7*(1), 11347. (2017)
54. Menden, M.P., et al. Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one* **8.4**, e61318 (2013)
55. Ding, M. Q., Chen, L., Cooper, G. F., Young, J. D., & Lu, X. 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 (2018)
56. Chang, Y., et al. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. *Scientific reports*, *8*(1), p.8857. (2018)
57. Papillon-Cavanagh, S., *et al*. Comparison and validation of genomic predictors for anticancer drug sensitivity. *Journal of the American Medical Informatics Association* **20.4**, 597-602 (2013).
58. Haibe-Kains, B., et al. Inconsistency in large pharmacogenomic studies. *Nature* **504.7480**, 389 (2013)
59. Haibe-Kains B. Addressing the (in)consistency of pharmacogenomic datasets. *F1000 Research Blog:* https://blog.f1000.com/2017/02/13/addressing-the-inconsistency-of-pharmacogenomic-datasets/(2018)
60. Safikhani, Z., et al. Revisiting inconsistency in large pharmacogenomic studies. *F1000Research* 5 (2016)

61. Cancer Cell Line Encyclopedia Consortium, and Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528.7580**, 84 (2015)
62. Geeleher, P., Gamazon, E. R., Seoighe, C., Cox, N. J., & Huang, R. S. Consistency in large pharmacogenomic studies. *Nature*, ***540*.7631,** E1. (2016)
63. Bouhaddou, M., et al. Drug response consistency in CCLE and CGP. *Nature* **540.7631**, E9 (2016)
64. Mpindi, J. P., et al. Consistency in drug response profiling. *Nature* **540.7631,** E5 (2016)
65. Rahman, R. et al. Evaluating the consistency of large-scale pharmacogenomic studies. *Briefings in Bioinformatics*. (2018)
66. Smirnov, P., *et al*. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* **32.8**, 1244-1246 (2015).
67. Smirnov, P. et al. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic acids research*, *46*(D1), pp.D994-D1002. (2017)
68. Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C. and Collins, J.J. Next-Generation Machine Learning for Biological Networks. *Cell*. (2018)
69. Azuaje, F. Computational models for predicting drug responses in cancer research. *Briefings in bioinformatics*, *18*(5), pp.820-829. (2016)
70. Libbrecht, M.W. and Noble, W.S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), p.321. (2015)
71. Kuhn, M., and Kjell J. *Applied predictive modeling*. Vol. 26. New York: Springer. (2013)
72. Friedman, J., Hastie, T., & Tibshirani, R. *The elements of statistical learning* (Vol. 1, No. 10). New York, NY, USA:: Springer series in statistics. (2001)
73. Boser, B.E., Guyon, I.M. & Vapnik, V.N. A training algorithm for optimal margin classifiers. in *5th Annual ACM Workshop on COLT* (ed. Haussler, D.) 144–152, ACM Press, Pittsburgh, PA. (1992).
74. Noble, W.S. What is a support vector machine?. *Nature biotechnology*, *24*(12), p.1565. (2006)
75. Kuhn, M. Caret package. *Journal of statistical software*, *28*(5), 1-26. (2008)
76. Kuhn, M. Variable selection using the caret package. *http://cran.cermin.lipi.go.id/web/packages/caret/vignettes/caretSelection.pdf* (2012).
77. Kurilov R, Haibe-Kains B. & Brors B. Drug response prediction in cell lines and xenografts. *Manuscript submitted for publication* (2018)
78. Tomska K., et al. Drug-based perturbation screen uncovers synergistic drug combinations in Burkitt lymphoma. *Scientific Reports* (2018)
79. Geeleher, P., Cox, N. J., & Huang, R. S. Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models. *Genome Biology*, ***17*(1)**, 190 (2016)
80. Friedman J., Hastie T., and Tibshirani R.. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version* 1.4 (2009).
81. Wan, Q., & Pal, R. An ensemble based top performing approach for NCI-DREAM drug sensitivity prediction challenge.*PloS one*, ***9*.6**, e101183 (2014)
82. Zhao, C., Li, Y., Safikhani, Z., Haibe-Kains, B., & Goldenberg, A. Using Cell line and Patient samples to improve Drug Response Prediction. *bioRxiv*, 026534 (2015)
83. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882-883. (2012).

84. Chou, T. C. & Talalay, P. Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors. *Advances in enzyme regulation* 22, 27–55 (1984).
85. Amtmann E et al. DKFZ-608, a novel TRXR1 inhibitor with potent and selective anti-SCLC activity, allows long term maintenance of first line therapy effects.*Manuscript in preparation* (2018)
86. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc.* **4(1)**:44-57 (2009)
87. Subramanian, A. A et al. next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, ***171*(6)**, 1437-1452. (2017)
88. Klaeger S. et al. The target landscape of clinical kinase drugs. *Science* **358.6367** eaan4368 (2017)
89. Forbes S.A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic acids research* **45, no. D1**, D777-D783 (2016)
90. Fallahi-Sichani, M., Honarnejad, S., Heiser, L. M., Gray, J. W., & Sorger, P. K. Metrics other than potency reveal systematic variation in responses to cancer drugs. *Nature chemical biology*, ***9*(11)**, 708 (2013).
91. Boehm, J.S. and Golub, T.R., An ecosystem of cancer cell line factories to support a cancer dependency map. *Nature Reviews Genetics*, *16*(7), p.373. (2015)
92. Witkiewicz, A. K. et al. Integrated patient-derived models delineate individualized therapeutic vulnerabilities of pancreatic cancer. *Cell reports*, *16*(7), 2017-2031.(2016)
93. Singal G et al. Development and validation of a real-world clinico-genomic database. *Am. Soc. Clin. Oncol.* 35:2514 (2017)
94. Simm, J. et al. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell chemical biology*, *25*(5), pp.611-618. (2018)

AUTHOR'S PUBLICATIONS

1. Main manuscript which includes the work presented in the Chapter 3 "Testing influence of different aspects of model training on prediction accuracy" and the Chapter 4 "Tissue type, doubling time and drug response prediction in cell lines and xenografts":

Kurilov R, Haibe-Kains B. & Brors B. Drug response prediction in cell lines and xenografts. *Manuscript submitted for publication* (2018)

2. Paper which includes the work presented in the section 3.6.1 "Burkitt lymphoma drug sensitivity screen analysis":

Tomska K, Kurilov R, Lee KS, Hüllein J, Lukas M, Sellner L, Walther T, Wagner L, Oleś M, Brors B, Huber W. Drug-based perturbation screen uncovers synergistic drug combinations in Burkitt lymphoma. *Scientific reports.* 8(1):12046.(2018)

ACKNOWLEDGMENTS

First of all I would like to thank my supervisor Benedikt Brors for his support and supervision.
I would like to thank Misha Kapushesky for his co-supervision in the initial stage of the project and for help on the subsequent stages. I want to thank Benjamin Haibe-Kains for giving me the opportunity to work in his lab during my internship in Toronto and for supervising me.
I would like to thank Dilafruz Juraeva for advices and support. I would also like to thank David Weese for discussions and advices.
I want to thank my collaborators Kasia Tomska, Thorsten Zenz and Nikolas Gunkel.
I would also like to thank a member of my Thesis Advisory Committee Wolfgang Huber.
I want to thank all my friends and colleagues from ABI, CO, CRG, DMG groups. I would like to thank Birgit Vey for help with administrative matters and support. I would also like to thank Sadaf Mughal for her help and advices. Also I want to thank Lina Sieverling for translating the summary.
Also I would like to thank the Graduate Program office of DKFZ for providing great conditions and environment for graduate studies.

APPENDIX: REPRODUCIBILITY

Chapters 2, 3, 6, 7

Scripts and the instructions for the reproduction of the analyses from Chapters 2, 3, 6 and 7 are available via github repository: https://github.com/RomaHD/thesis_code

In all scripts, the user needs to change the variable path to the repository root folder. Prior to running scripts from the chapters data.R script from the root folder should be executed in order produce the necessary data files for the Chapters 2,3,7.

Chapter 2

	name	**thesis's section**
1.	drug_resp_consistency.R	"Drug response consistency"
2.	biomarkers_consistency.R	"Biomarkers' consistency"

Chapter 3

	name	**thesis's section**
1.	multi_task.R	"Multi-task glmnet models"
2.	aggregated.R	"Modelling on aggregated data"
3.	feature_interactions.R	"Modelling with feature interactions"
4.	weights.R	"Modelling with weights"
5.	imbalance_and_consistency.R	"How class imbalance and cross-set inconsistency affect prediction accuracy"

Chapter 6

Data: patient expression data files “bortezomib.patient.RData”, “erlotinib_data.RData”, and “pp.RData” should be downloaded from http://compbio.cs.toronto.edu/cp2p/ and extracted in the folder /data
Scripts should be executed in the following order:

	name	**description**
1.	data_prep.R	preprocess all data necessary for the modelling
2.	main_analysis.R	performs the main anlysis

Chapter 7

	name	thesis's section/description
1.	bl_unsupervised_analysis.R	"Burkitt lymphoma drug sensitivity screen analysis" -- usupervised part
2.	bl_modelling.R	"Burkitt lymphoma drug sensitivity screen analysis" -- modelling
3.	dkfz608_modelling.R	"Drug response prediction model for DKFZ-608 compound"
4.	shiny_data.R	"Shiny application for complex drug response visualization" -- data preparation
5.	app.R	"Shiny application for complex drug response visualization" -- app.R file of the application

SessionInfo

Session environment information

#sessionInfo()

R version 3.5.0 (2018-04-23) Platform: x86_64-redhat-linux-gnu (64-bit) Running under: CentOS Linux 7 (Core)

Matrix products: default BLAS/LAPACK: /usr/lib64/R/lib/libRblas.so

locale: [1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8

[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8 LC_PAPER=en_US.UTF-8 LC_NAME=C

[9] LC_ADDRESS=C LC_TELEPHONE=C LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages: [1] parallel stats graphics grDevices utils datasets methods base

other attached packages: [1] PharmacoGx_1.8.3 e1071_1.7-0 doMC_1.3.5 iterators_1.0.9 foreach_1.4.4 caret_6.0-80 ggplot2_2.2.1

[8] lattice_0.20-35

loaded via a namespace (and not attached): [1] fgsea_1.4.1 colorspace_1.3-2 class_7.3-14 rprojroot_1.3-2 lsa_0.73.1 pls_2.6-0
[7] DRR_0.0.3 SnowballC_0.5.1 prodlim_2018.04.18 lubridate_1.7.4 codetools_0.2-15 splines_3.5.0
[13] mnormt_1.5-5 robustbase_0.93-0 knitr_1.20 RcppRoll_0.3.0 magicaxis_2.0.3 broom_0.4.4
[19] ddalpha_1.3.3 cluster_2.0.7-1 kernlab_0.9-26 sfsmisc_1.1-2 mapproj_1.2.6 compiler_3.5.0
[25] backports_1.1.2 assertthat_0.2.0 Matrix_1.2-14 lazyeval_0.2.1 limma_3.34.9 htmltools_0.3.6
[31] tools_3.5.0 bindrcpp_0.2.2 igraph_1.2.1 gtable_0.2.0 glue_1.2.0 RANN_2.5.1
[37] reshape2_1.4.3 dplyr_0.7.5 maps_3.3.0 fastmatch_1.1-0 Rcpp_0.12.17 slam_0.1-43
[43] Biobase_2.38.0 gdata_2.18.0 nlme_3.1-137 psych_1.8.4 timeDate_3043.102 gower_0.1.2
[49] stringr_1.3.1 gtools_3.5.0 DEoptimR_1.0-8 MASS_7.3-50 scales_0.5.0 ipred_0.9-6
[55] relations_0.6-8 RColorBrewer_1.1-2 sets_1.0-18 yaml_2.1.19 gridExtra_2.3 downloader_0.4
[61] rpart_4.1-13 stringi_1.2.3 NISTunits_1.0.1 plotrix_3.7-2 randomForest_4.6-14 caTools_1.17.1
[67] BiocGenerics_0.24.0 BiocParallel_1.12.0 lava_1.6.1 geometry_0.3-6 rlang_0.2.1 pkgconfig_2.0.1
[73] bitops_1.0-6 evaluate_0.10.1 pracma_2.1.4 purrr_0.2.5 bindr_0.1.1 recipes_0.1.3
[79] labeling_0.3 CVST_0.2-2 tidyselect_0.2.4 plyr_1.8.4 magrittr_1.5 R6_2.2.2
[85] gplots_3.0.1 dimRed_0.1.0 sm_2.2-5.5 pillar_1.2.3 foreign_0.8-70 withr_2.1.2
[91] survival_2.42-3 abind_1.4-5 nnet_7.3-12 tibble_1.4.2 KernSmooth_2.23-15 rmarkdown_1.10
[97] grid_3.5.0 data.table_1.11.4 marray_1.56.0 piano_1.18.1 ModelMetrics_1.1.0 digest_0.6.15
[103] tidyr_0.8.1 stats4_3.5.0 munsell_0.5.0 celestial_1.4.1 magic_1.5-8 tcltk_3.5.0

Chapters 4, 5

Scripts and the instructions for the reproduction of the analyses from Chapters 4 and 5 are available via github repository:
https://github.com/RomaHD/DrugRespPrediction
(subfolders “analysis1” and “analysis2” respectively)

In all scripts, the user needs to change the variable path to the repository root folder.

Analysis I (Chapter 4)

Scripts should be executed in the following order:

	name	**description**
1.	data_for_analysis1.R	obtains data from PharmacoGx and preprocess it for a subsequent analysis
2.	analysis.R	performs the analysis and saves results into results_table.RData
3.	plotting_R2.R	produces figures 2a (Fig. 22a in this thesis), 2b (Fig. 22b) and supplementary figures s3, s4, s6 (Fig. 24), s7, s8 (Fig. 25), s9
4.	top_features.R	calculates and saves molecular features associated with drug response vectors and produces figure 2c (Fig. 22c)
5.	modelling_obs_pred.R	produces predictions for each modelling task and saves them into raw_predictions2.RData
6.	plotting_obs_vs_pred.R	produces figure 3 (Fig. 23) and supplementary figures s5-1, s5-2, s5-3

Analysis II (Chapter 5)

Data: gCSI molecular data, particularly files “gcsi.genomics.rda”, “gcsi.genomics.feature.info.rda”, and “gcsi.line.info.rda” should be downloaded from http://research-pub.gene.com/gCSI-cellline-data/compareDrugScreens_current.tar.gz and extracted in the folder analysis2/data

Scripts should be executed in the following order:

	name	**description**
1.	nibr_preprocessing.R	preprocess molecular and drug response xenograft data
2.	main_analysis.R	performs the analysis, saves results and produces supplementary figures s11-1 and s11-2
3.	tissue_classification_equal_groups.R	performs tissue classification for the case with equal number of samples per each tissue and produces figure 4 (Fig. 28 in this thesis)
4.	plotting_obs_vs_pred.R	produces figure 5a (Fig. 29a)
5.	drug_resp_gcsi_to_nibr.R	creates and tests model for xenograft predictions trained on cell line data, produces figure 5b (Fig. 29b) and supplementary figure s10

SessionInfo

Analysis has been done in the following session environment.

#sessionInfo()

R version 3.4.4 (2018-03-15) Platform: x86_64-redhat-linux-gnu (64-bit) Running under: CentOS Linux 7 (Core)

Matrix products: default BLAS/LAPACK: /usr/lib64/R/lib/libRblas.so

locale: [1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C LC_TIME=en_US.UTF-8

[4] LC_COLLATE=en_US.UTF-8 LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8

[7] LC_PAPER=en_US.UTF-8 LC_NAME=C LC_ADDRESS=C

[10] LC_TELEPHONE=C LC_MEASUREMENT=en_US.UTF-8
LC_IDENTIFICATION=C

attached base packages: [1] splines stats graphics grDevices utils datasets
methods base

other attached packages: [1] PharmacoGx_1.8.3 ggfortify_0.4.5 gam_1.15
foreach_1.4.4 caret_6.0-78

[6] ggplot2_2.2.1 lattice_0.20-35 survcomp_1.28.5 prodlim_1.6.1 survival_2.42-3

loaded via a namespace (and not attached): [1] nlme_3.1-137 lsa_0.73.1
survivalROC_1.0.3 bitops_1.0-6
[5] lubridate_1.7.3 dimRed_0.1.0 RColorBrewer_1.1-2 SnowballC_0.5.1
[9] tools_3.4.4 R6_2.2.2 rpart_4.1-13 KernSmooth_2.23-15 [13] sm_2.2-5.4
lazyeval_0.2.1 BiocGenerics_0.24.0 colorspace_1.3-2
[17] rmeta_3.0 nnet_7.3-12 withr_2.1.1 tidyselect_0.2.4
[21] gridExtra_2.3 mnormt_1.5-5 compiler_3.4.4 Biobase_2.38.0
[25] slam_0.1-42 caTools_1.17.1 scales_0.5.0 sfsmisc_1.1-2
[29] DEoptimR_1.0-8 psych_1.7.8 robustbase_0.92-8 randomForest_4.6-12 [33]
relations_0.6-7 stringr_1.3.0 digest_0.6.15 foreign_0.8-70
[37] pkgconfig_2.0.1 plotrix_3.7 limma_3.34.9 maps_3.3.0
[41] rlang_0.2.0 ddalpha_1.3.1.1 MLmetrics_1.1.1 SuppDists_1.1-9.4
[45] bindr_0.1 BiocParallel_1.12.0 gtools_3.5.0 dplyr_0.7.4
[49] ModelMetrics_1.1.0 magrittr_1.5 Matrix_1.2-14 Rcpp_0.12.15
[53] celestial_1.4.1 munsell_0.4.3 piano_1.18.1 stringi_1.1.6
[57] MASS_7.3-50 gplots_3.0.1 plyr_1.8.4 recipes_0.1.2
[61] grid_3.4.4 gdata_2.18.0 parallel_3.4.4 mapproj_1.2.6
[65] pillar_1.2.1 fgsea_1.4.1 igraph_1.2.1 xgboost_0.6.4.1
[69] marray_1.56.0 reshape2_1.4.3 codetools_0.2-15 stats4_3.4.4
[73] fastmatch_1.1-0 CVST_0.2-1 NISTunits_1.0.1 glue_1.2.0
[77] downloader_0.4 data.table_1.10.4-3 bootstrap_2017.2 gtable_0.2.0
[81] RANN_2.5.1 purrr_0.2.4 tidyr_0.8.0 kernlab_0.9-25
[85] assertthat_0.2.0 DRR_0.0.3 gower_0.1.2 broom_0.4.3
[89] pracma_2.1.4 e1071_1.6-8 class_7.3-14 timeDate_3043.102
[93] RcppRoll_0.2.2 tibble_1.4.2 iterators_1.0.9 cluster_2.0.7-1
[97] sets_1.0-18 bindrcpp_0.2 lava_1.6 ROCR_1.0-7
[101] magicaxis_2.0.3 ipred_0.9-6