Development of Machine Learning and Biostatistical Models for Cancer Pharmacogenomics Screens

Dissertation
Zur Erlangung des Doktorgrades der Naturwissenschaften
(Dr. rer. nat.)
Fakultät für Biologie
Ludwig-Maximilians-Universität München

vorgelegt von
Ana Cláudia Paulo Galhoz
aus
Lissabon, Portugal

München, Dezember 2024

Erster Gutachter:	A/Prof. Dr. habil. Michael P. Menden
Zweiter Gutachter:	Prof. Dr. Dirk Metzler
Tag der Abgabe:	16.12.2024
Tag der mündlichen Prüfung:	14.07.2025

Erklärung
Declaration

Hiermit erkläre ich, *
Hereby I declare

■ dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist.
that this work, complete or in parts, has not yet been submitted to another examination institution

■ dass ich mich anderweitig einer Doktorprüfung ohne Erfolg **nicht** unterzogen habe.
*that I did **not** undergo another doctoral examination without success*

□ dass ich mich mit Erfolg der Doktorprüfung im Hauptfach
that I successfully completed a doctoral examination in the main subject

und in den Nebenfächern
and in the minor subjects

bei der Fakultät für
at the faculty of ______________________

der
at ______________________
(Hochschule/*University*)
unterzogen habe.

□ dass ich ohne Erfolg versucht habe, eine Dissertation einzureichen oder mich der Doktorprüfung zu unterziehen.
that I submitted a thesis or did undergo a doctoral examination without success

Stuttgart, 22/07/2025
Ort, Datum/*place, date*

Ana Cláudia Paulo Galhoz
Unterschrift/*signature*

*) Nichtzutreffendes streichen/
delete where not applicable

Eigenständigkeitserklärung

Hiermit versichere ich an Eides statt, dass die vorliegende Dissertation mit dem Titel

Development of Machine Learning and Biostatistical Models for Cancer Pharmacogenomics Screens

von mir selbstständig verfasst wurde und dass keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden. Die Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinne nach entnommen sind, wurden in jedem Fall unter Angabe der Quellen (einschließlich des World Wide Web und anderer elektronischer Text- und Datensammlungen) kenntlich gemacht. Weiterhin wurden alle Teile der Arbeit, die mit Hilfe von Werkzeugen der künstlichen Intelligenz de novo generiert wurden, durch Fußnote/Anmerkung an den entsprechenden Stellen kenntlich gemacht und die verwendeten Werkzeuge der künstlichen Intelligenz gelistet. Die genutzten Prompts befinden sich im Anhang. Diese Erklärung gilt für alle in der Arbeit enthaltenen Texte, Graphiken, Zeichnungen, Kartenskizzen und bildliche Darstellungen.

Stuttgart, 22/07/2025

(Ort / Datum)

ANA CLÁUDIA PAULO GALHOZ

(Vor und Nachname in Druckbuchstaben)

Ana Cláudia Paulo Galhoz

(Unterschrift)

Affidavit

Herewith I certify under oath that I wrote the accompanying Dissertation myself.

Title: ______________________________

In the thesis no other sources and aids have been used than those indicated. The passages of the thesis that are taken in wording or meaning from other sources have been marked with an indication of the sources (including the World Wide Web and other electronic text and data collections). Furthermore, all parts of the thesis that were de novo generated with the help of artificial intelligence tools were identified by footnotes/annotations at the appropriate places and the artificial intelligence tools used were listed. The prompts used were listed in the appendix. This statement applies to all text, graphics, drawings, sketch maps, and pictorial representations contained in the Work.

(Location/date)

(First and last name in block letters)

(Signature)

Acknowledgements

Five years ago, I started a remarkable chapter in my life. This period would have not been possible without the existence and support of several individuals. To them, I dedicate the following appreciation.

First, I would like to thank my Doktorvater “Micha” for the opportunity to integrate such a diverse and talented team, and spark my interest in the application of mathematics to cancer pharmacology. Moreover, I appreciate his continuous support, endorsement and, mainly, his guidance to shape me to the researcher I am today. Will forever cherish your mentorship.

To my doctoral committee members, Dr. habil. Andreas Beyerlein and Prof. Dr. Dirk Metzler, I am extremely grateful for their valuable feedback on my research and progress during my TAC meetings.

Another fundamental pillar(s) in this journey were all the remarkable researchers with whom I had the chance to collaborate. These invaluable collaborations stimulated several research questions and helped me to evolve as a scientist. Namely, to Prof. Dr. Paul Lingor, Dr. Lucas Gomes, Dr. Laura Tzeplaeff, Iñigo Ayestaran, Dr. Kamyar Hadian, Prof. Dr. Daniel Krappmann and Dr. Denes Turei, with whom I worked more closely, I have the greatest admiration and appreciation to you and the work we developed together.

The academic journey would not be complete (and not as enjoyable) without a research-family. To the members of the Menden Lab, Ginte Kutkaite, Alexander Ohnmacht, Phong Nguyen, Christina Hillig, Göksu Avar, Jenny Riedel, Clara Meijs, Martin Meinel, Daniel Garger, Dr. Ines Assum, Dr. Diyuan Lu, Nikita Makarov, Maria Bordukova and Dr. Ali Farnoud, the biggest thank you, it was a pleasure to share this period with you. I’ll always treasure our funny (and surprisingly scientific) coffee breaks, projects feedback, encouragement and, of course, beer outings. In particular, I would like to thank Ginte for being an awesome colleague to collaborate with and a great friend behind the curtains and Alex for his constant interest and suggestions to my research.

Undoubtedly, I would have not been able to accomplish this endeavour without the strong support of my family and friends. I am especially grateful to my parents, uncles and cousin for their continuous recognition and moral encouragement. My greatest gratitude goes to my husband Daniele for his endless support and for being my greatest admirer, private cheerleader and source of laughter. Lastly, a special thank you to my sweet daughter Sofia for providing me the motivation, strength and sleepless nights needed to complete this dissertation.

Table of contents

Abstract

Cancer is a complex genetic disease emerging from the accumulation of somatic alterations that drive tumour growth. This disease is remarkably heterogeneous, comprising several subtypes driven by various distinct mutational events and with individual response mechanisms. Notably, its complexity renders this disease hard to research and contributes to be one of the top deadliest worldwide.

High-throughput drug screens have empowered numerous targeted and combination therapies for personalised patient treatment by revealing potentially relevant biomarkers. The application of large scale of genomic datasets, such as the Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Therapeutics Response Portal (CTRP), has sparked the need for suitable bioinformatic tools to properly mine, model and analyse cancer biomarkers in the data.

In this dissertation, I focused on three aims towards cancer biomarker discovery and developed distinct algorithms to analyse each task. **Aim 1**, analysing drug resistance mechanisms using statistical frameworks; **Aim 2**, investigating synergistic drug combinations in cells with uncontrolled proliferation markers using curve fitting methodologies; and **Aim 3**, identifying new cancer-specific driver genes based on a network-based approach.

Aim 1: To investigate acquired resistance to a treatment from initially responsive cell lines, I developed an outlier statistical model that identifies unexpectedly resistant cell lines from the GDSC and CCLE drug screens. This method not only reproduced known biomarkers in lung adenocarcinoma, but also outperformed a standardized outlier detection method. Furthermore, the proposed hierarchical statistical framework was also tested in terms of false discovery rate bounds.

Aim 2: Secondly, I looked into the modelling of drug responses with unexpectedly increase cell viability missed by standard methodologies, and proposed to leverage drug-induced uncontrolled proliferation as a new synergistic combination therapy with drugs that act on fast proliferating cells, e.g., DNA damaging agents. Building on this, I developed two mathematical frameworks based on Gaussian and linear models to capture cancer-type biomarkers of increased viability. Promising candidates in lung cancer were tested in additional drug screen experiments and potential synergistic drug combinations were hypothesised.

Aim 3: I proposed the weighted Protein-Protein Interaction (wPPI) tool based on PPI networks, combined with Gene Ontology and Human Phenotype Ontology datasets, to infer new tissue-specific genes closely related to cancer driver genes. Subsequently, the gene expression profiles of the top highest scoring candidates were used to develop drug response machine learning models in breast cancer. The performance of the built models was assessed and cross-compared with models created with several gene feature sets, namely unspecific tissue-specific genes and genes prioritised with other network-based methodology.

In summary, this dissertation introduces innovative and robust computational methodologies to advance tissue-specific cancer biomarker discovery. These approaches address multiple challenges associated with limited statistical power in precision oncology, including the investigation of rare phenomena and the insufficient understanding of key players of cancer progression. As an overarching goal, these methodologies are envisioned to not only enhance insights into the complex mechanisms underlying cancer, but also contribute to the design of refined targeted therapeutic strategies.

List of figures

List of tables

List of abbreviations

ANOVA	Analysis of Variance
AUC	Area Under the Curve
BH	Benjamini-Hochberg
BP	Biological Process
BRCA	Breast adenocarcinoma
CC	Cellular Component
CCLE	Cancer Cell Line Encyclopedia
cDNA	Complementary Deoxyribonucleic Acid
CFE	Cancer Functional Event
CIN	Chromosomal Instability
CML	Chronic Myeloid Leukaemia
CMS	Consensus Molecular Subtypes
CNV	Copy Number Variation
COREAD	Colorectal Adenocarcinoma
CTRP	Cancer Therapeutics Response Portal
CV	Cross-Validation
DGEx	Differential Gene Expression Analysis
DNA	Deoxyribonucleic Acid
EGFR	Epidermal Growth Factor Receptor
Emax	Maximal Effect at High Concentrations
FDR	False Discovery Rate
gCSI	Genentech Cell Line Screening Initiative
GDSC	Genomics of Drug Sensitivity in Cancer
GO	Gene Ontology
GP	Gaussian Process
GSEA	Gene Set Enrichment Analysis
HFDR	Hierarchical False Discovery Rate
HPO	Human Phenotype Ontology
HTS	High-throughput screen
IC_{50}	Half Maximal Inhibitory Concentration

ICV	Increasing Cell Viability
IntOGen	Integrative OncoGenomics
LAML	Acute Myeloid Leukemia
LOF	Loss-of-function
LOOCV	Leave-one-out Cross-Validation
LUAD	Lung adenocarcinoma
MDM2	Murine Double Minute 2
MF	Molecular Function
ML	Machine Learning
MoA	Mode of Action
mRNA	Messenger Ribonucleic Acid
MSE	Mean Squared Error
NCI	National Cancer Institute
NGS	Next-generation Sequencing
NLS	No-linear Least Squares
NP	Neyman-Pearson
NSCLC	Non-small-cell Lung Carcinoma
OCG	Oncogene
OMIM	Online Mendelian Inheritance in Man
PCA	Principal Component Analysis
PDX	Patient-Derived Xenograf
PPI	Protein-Protein Interaction
PTM	Post-Translational Modification
RMSE	Root-mean-squared Error
RNA	Ribonucleic Acid
RNA-seq	Ribonucleic Acid Sequencing
RSS	Residual Sum of Squares
SCC	Squamous Cell Carcinoma
SD	Standard Deviation
SNP	Single-nucleotide Polymorphism
SNV	Single-nucleotide Variant

SVM	Support Vector Machine
TKI	Tyrosine Kinase Inhibitors
TSG	Tumour Suppressor Gene
t-SNE	T-distributed Stochastic Neighbour Embedding
UNRES	UNexpectedly RESistant
UV	Ultraviolet
WHO	World Health Organization
wPPI	Weighted Protein-Protein Interaction

1. Introduction

1.1 Cancer biology

Cancer is a complex and heterogeneous disease driven by genetic, epigenetic, and environmental alterations. Its development is mainly characterized by dynamic somatic mutations in the genome that gradually transform normal cells into malignant clones with uncontrolled proliferation, apoptosis evasion and metastatic capability, i.e., cancer hallmarks (Hanahan & Weinberg, 2000). In 1838, pathologist Johannes Müller described for the first time that cancers are composed of cells (Müller, 1838). Later, his student Rudolf Virchow discovered that these cells are in fact derived from previous healthy cells (Virchow, 1858). These findings were fundamental to understand the disease origin and, together with technological advances, its several hallmarks (Hanahan & Weinberg, 2011).

The WHO agency states cancer is the second most frequent cause of death worldwide, with an estimation of nearly 10 million deaths across 185 countries in 2020 (Ferlay et al., 2021). The most common cancer types are breast (2.26 million cases) and lung (2.21 million cases). Notwithstanding, studies made in the United States (Siegel et al., 2022) and Europe (Dalmartello et al., 2022) noted and predicted a decline in the risk of developing cancer and mortality patterns for several cancer types (e.g., lung, breast and prostate cancer) due to the adoption of low risk factor habits, employment of early detection programs and continuous improvements in medical practices.

In summary, despite notable research advances, the inevitable accumulation of somatic alterations throughout a person's life combined with its multifaceted complexity, positions cancer as a challenging worldwide health concern. Therefore, a comprehensive response addressing its diverse mechanisms is crucial to enable early detection, effective treatment strategies and, ultimately, improve survival rates.

1.1.1 Genetic variations

Even though the human DNA is almost identical from one individual to another (around 99.5% shared base pairs), one can find subtle natural variation within each individual (Collins & Mansoura, 2001; Gonzaga-Jauregui et al., 2012; Levy et al., 2007). These alterations may be inherited through generations in the reproduction process from parental cells (germline variation), or from changes in any single dividing cell (excluding germ cells) by internal or external processes accumulated through a lifespan of an individual (somatic variation) (Figure 1).

Cancer is mainly driven by the accretion of somatic mutations derived from alterations in the nucleotide sequence during DNA replication or recombination. The former mechanism is one of the most essential processes the cell undergoes during the S-phase, DNA synthesis, in preparation for mitosis, where two full

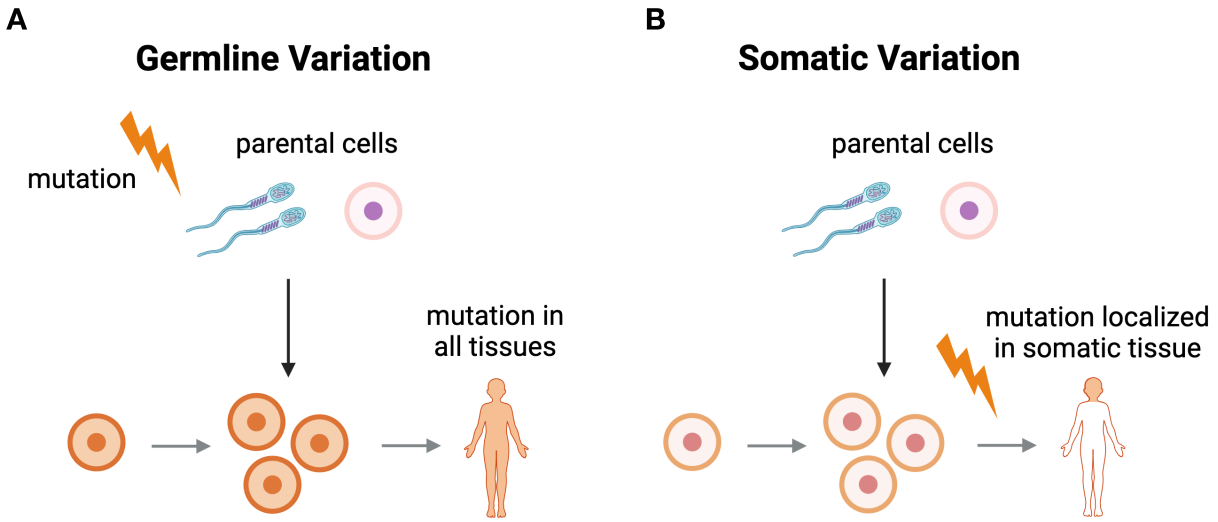

Figure 1: Germline and somatic variations.

(A) Germline variations originate from alterations in parental cells and culminate into mutations in all tissues. In contrast, **(B)** somatic variations can occur in a single dividing cell by internal or external processes and result in a somatic mutation in the locally affected tissue.
Figure created with BioRender.com.

copies of a cell are created from a single parental one. During this process, the original double helix is separated in two DNA strands, where one of the strands is read by a DNA polymerase enzyme and serves as a template to synthesize a new complementary DNA segment. Notwithstanding the high level of accuracy entangled in it, this process is error prone during the reading and writing stages. In brief, the human DNA consists of almost 6 billion base pairs and is estimated to have an average mistake rate of 1 in 10^{10} per replicated nucleotides (Kunkel & Bebenek, 2000; McCulloch & Kunkel, 2008; Pray, 2008). Leading to a potential ~120 thousand error rate per each cell division event.

Furthermore, DNA replication is conducted by polymerases enzymes and new mutations mainly manifest when mistakes are not corrected by these enzymes. Examples of these errors consist of single base insertions/deletions (indels; e.g., copy number variation (CNV)) or substitutions (e.g., single nucleotide polymorphisms (SNPs)) (Loeb et al., 1974). In addition, mutations can also be derived by DNA damage from exposure to environmental factors (induced mutations; e.g., ultraviolet radiation and smoking) or from internal natural reactions (spontaneous mutations).

Favourably, there are several intrinsic DNA repair mechanisms available throughout the cell cycle, namely DNA mismatch repair, nucleotide excision repair, base excision repair and double-strand break repair (Alhmoud et al., 2020). Moreover, in the event of persistent damage, repair processes such as apoptosis (controlled cell death) and senescence (cell cycle arrest) are activated by intrinsic or extrinsic pathways. The existence of these key safety repair protocols increases the accuracy of the DNA replication process and defend human cells against mutation effects.

1.1.2 Cancer functional events

1.1.2.1 Passenger and driver mutations

Somatic mutations can be grouped in two categories: passenger and driver mutations. The latter, albeit only represent a small portion of the total number of mutations, are responsible for driving the cancer progression and, therefore, have a selective advantage with regard to other cells (Stratton et al., 2009). Notably, cancer evolves by the cumulative collection of driver mutations which impacts key pathways and triggers cancer hallmarks (Hanahan & Weinberg, 2000). In contrast, passenger mutations are the most common type of variations, functionally neutral and with no direct impact on cancer development (Castro-Giner et al., 2015; McFarland et al., 2013).

Driver mutations take place in cancer driver genes and mainly manifest in the form of somatic mutations, with only a minority occurring in the germline. Most of the human genome comprises non-coding sequence regions (intergenic and introns), with only ~2% of gene-encoding genome (exons) (Elgar & Vavouri, 2008; Encyclopedia of Cancer, 2018). Most driver mutations are found in the exon's regions, but some are also present in non-coding regions (Rheinbay et al., 2020). Examples of the latter are mutations in the *TERT* (Fredriksson et al., 2014; Melton et al., 2015; Rheinbay et al., 2020; Weinhold et al., 2014), *PLEKHS1* (Fredriksson et al., 2014; Melton et al., 2015; Nik-Zainal et al., 2016; Weinhold et al., 2014) and *WDR74* (Nik-Zainal et al., 2016; Weinhold et al., 2014) promoter regions.

Point substitution mutations in the coding region can be subdivided into two categories: non-synonymous (alter protein sequence) and synonymous (coding silent) mutations. Non-synonymous mutations consist of a single nucleotide base modification that results in the alteration of the amino acid sequence of a protein. These can be manifested as missense, nonsense or splices site mutations, or as small insertions and deletions (indels) (Pleasance et al., 2010). In cancer, the most common type of protein-altering mutation are missense mutations and these are characterised by the change of one amino acid via a single base pair substitution (Vogelstein et al., 2013). A notable example of this category of mutations are single nucleotide substitutions in the p53 transcription factor during DNA binding activity (amino acid residues 102-292) (Baugh et al., 2018; Cho et al., 1994; Pavletich et al., 1993). These mutations prompt the disruption of the normal functional activity and stability of p53 during its response to DNA damage mechanisms, such as cell-cycle regulation and controlled cell death (Olivier et al., 2010). Alternatively, a nonsense mutation is a single nucleotide prematurely translated to a stop codon, consequently leading to protein truncation.

In this thesis I explore mechanisms of indirect resistance in cancer. In particular, I focus attention on lung adenocarcinoma and epidermal growth factor receptor (EGFR)-related mutations, where the gold-standard T790M mutation exhibits secondary resistance to EGFR tyrosine kinase inhibitors (TKI). Within this cancer type, two notable non-synonymous mutations are the replacement of leucine to arginine at codon 858 in exon 21 (L858R) and short deletions in exon 19, which provoke truncation of beta3-alphaC loop (Foster et al., 2016). These EGFR mutations are commonly linked to non-small cell lung cancer (NSCLC) (A. R. Li et al., 2008; Lynch et al., 2004; Mitsudomi & Yatabe, 2010; Paez et al., 2004; J. Su et al., 2017), and strongly sensitive to EGFR TKIs, such as gefitinib (Haber et al., 2005; Lynch et al., 2004; Paez et al., 2004). Tumours with these mutations are commonly referred to as "oncogene addicted" since tumour growth and survival is dependent on the presence of EGFR activating mutations (Soria et al., 2012).

Furthermore, non-synonymous mutations are intrinsically exposed to natural selection, and this is particularly essential in cases of alterations in cancer-related genes that could drive cancer biogenesis (Cartegni et al., 2002; Druillennec et al., 2012).

Distinctively, synonymous mutations are defined by nucleotide changes in protein coding regions that do not modify the translated protein sequence. In detail, a codon is composed of three nucleotides that encode a particular amino acid. However, due to redundancy in the genome, this sequence of nucleotides is not necessarily unique since multiple codons can be translated to the same amino acid. For instance, if the sequence GGT is altered in the third position to GGA, the resulting encoded amino acid would be glycine for both the non-altered and mutated codons (Pawlak et al., 2023). Therefore, typically, these mutations are interpreted as silent and invisible to natural selection. However, in contrast to its 'silent' perceptivity, synonymous mutations can alter splicing regulatory sites and mRNA stability or translation (Duan et al., 2003; Sauna & Kimchi-Sarfaty, 2011), thus leading to non-silent effects (Supek et al., 2014).

1.1.2.2 Copy number alterations

Genetic variations typically emerge in the synthesis (S-phase) and mitosis (in particular, during metaphase or anaphase) phases of the cell cycle throughout the course of cancer evolution. These alterations encompass a wide range of forms, namely copy number variations (CNVs), single-nucleotide variations (SNVs) and insertions or deletions (indels).

Taking into consideration CNVs frequently enclose genes and represent one of the most prevalent forms of genetic alteration, the study and comprehension of CNVs role is paramount in cancer diagnosis (Almal & Padh, 2012; Kuiper et al., 2010; Shlien & Malkin, 2009). CNVs are genomic structural regions where the number of copies of a gene has been modified in comparison to a reference genome. These alterations can be created due to amplifications or deletions, and lead to potential gain or loss of genomic DNA, respectively (Redon et al., 2006).

CNVs are encountered throughout several types of cancer with variant frequencies, and are associated with various key pathways in cancer (Andor et al., 2016). For example, the amplification of the chromosomal region 17q12-q21 in breast and gastric cancers (Cancer Genome Atlas Network, 2012; Iqbal & Iqbal, 2014; Liang et al., 2016; Slamon et al., 1987). This region contains the gene *HER2* (human epidermal growth factor receptor 2), often referred to by its alias *ERBB2* (Erb-B2 receptor tyrosine kinase 2). In detail, *HER2* encodes an 185-kDa tyrosine kinase receptor (p185HER-2) from the EGFR family (Doherty et al., 1999), which is responsible for the regulation of cell growth and proliferation via signalling pathways such as the PI3K-AKT-mTOR and MAPK (Yarden & Sliwkowski, 2001). Amplification of the *HER2* gene causes *HER2* overexpression and, consequently, enable activation via ligand-activated hetero-dimerization (heterodimers such as EGFR or HER3) or homo-dimerization. Notably, these dimerisation processes activate growth factor signalling and define compounds targeting *HER2* (e.g., Trastuzumab) to be commonly leveraged in clinics as an effective targeted therapy for breast cancer (Gajria & Chandarlapaty, 2011).

1.1.2.3 Other events

During the cell cycle, in particular through the nuclear division process in the M-phase, several genome rearrangements may occur due to error-prone mechanisms. Notably, during this phase, chromosome segregation takes place, where two sister chromatids are created, separated and posteriorly aligned. This process may lead to stalled or collapsed DNA replication forks (Hills and Diffley 2014), and if defectively repaired, result in structural and numerical genomic aberrations such as gain or loss of a chromosomal arm (aneuploidy) and/or more than one complete pair of chromosomes (polyploidy) (Storchova and Pellman 2004).

Chromosomal segregation error events occurring during the mitosis phase are referred to as chromosomal instability (CIN) (Lengauer et al., 1998), and represent a human cancer hallmark (Hanahan & Weinberg, 2011). Remarkably, the majority of human tumours manifest chromosomal rearrangements promoted by CIN (Santaguida and Amon 2015), and the CIN phenotype is enriched in metastatic and relapsing tumours (Turajlic and Swanton 2016; Sotillo et al. 2010). In addition, a burden of CIN phenomena is associated with cancer poor prognosis, therapeutic resistance and consequent low survival rate in several solid tumour types such as breast and colon cancer (Lee et al. 2011; Carter et al. 2006; Swanton et al. 2009; Walther et al. 2008). Despite its central role in cancer evolution, the measurement of CIN in cancer tumours is not easily testable and is only possible through indirect inference of chromosomal mis-segregation rates via next-generation sequencing (NGS) technologies, such as bulk DNA sequencing and single-cell sequencing (Bakhoum et al. 2018; Bakker et al. 2016).

Chromosomal translocations are another cancer driving genetic abnormality in the DNA. These are recurrent events where a chromosome segment breaks and is transferred to another chromosome or to a different site of the same chromosome (Rabbitts 1994). Although translocation phenomena contribute to genetic heterogeneity and evolution, typically these are linked to the production of oncogenes, and consequent misregulation of gene expression (Zheng 2013). Among various cancers, translocations are typical drivers of lymphomas and leukaemias. For instance, a famous translocation is derived via the fusion of genes *BCR* and *ABL* localised on chromosomes 22 and 9, respectively (Korsmeyer 1992). This fusion forms an abnormal chromosome, the Philadelphia chromosome (Ph), of chronic myelogenous leukaemia patients, and originates a new fusion gene *BCR-ABL* with enhanced cell growth, proliferation and resistance to cell death. Currently, the knowledge of this translocation is leveraged as a therapeutic marker for Ph-positive leukaemia patients (An et al. 2010).

In addition, tumour plasticity may render pre- or malignant cells more prone to a proto-oncogenic state through epigenetics' alterations (Flavahan et al. 2017). In particular, chromatin, transcriptional or proteomic aberrations can modify chromatin states, expression of genes and/or gene pathway mechanisms, that consequently lead to an oncogene activation and subsequent tumour development (Allis and Jenuwein 2016; Puisieux et al. 2014). For example, isocitrate dehydrogenase (*IDH*) gene mutations in gliomas lead to stochastic hypermethylation of *CTCF*, and subsequent insulation disruption of *PDGFRA* oncogene (Flavahan et al. 2016). Ultimately, this event induces the activation of *PDGFRA* and promotes proliferation of hypermethylated glioma cells.

1.1.3 Tumour evolution

In 1976, it was formalised that the creation and development of cancer is a dynamic process where subpopulations of cells with growth and survival advantage are favoured (Nowell, 1976). This process grants evolutionary advantage to cancer cells and triggers several cancer hallmarks, such as uncontrolled proliferation, evasion of programmed cell death and senescence.

As the tumour progresses, cancers evolve to a heterogeneous collection of tumour cells with distinct genetic and phenotypic signatures. This heterogeneity can be labelled as intratumour or intertumour (Burrell et al., 2013). The former reflects the genomic and biological changes that occur within the primary tumour and its metastases. Whilst the latter contemplates the heterogeneity between tumours of the same hispathological subtype occurring in different patients.

In addition, the dynamic nature of the tumour evolution can lead to spatial and temporal diversity (Hiley & Swanton, 2014), thus promoting a higher genomic instability (e.g., increased metastasis recurrence and outgrowth of cell clones due to drug resistance). Therefore, a comprehensive evaluation of tumour heterogeneity profile and its evolutionary outcomes are of great research relevance to improve clinical prognosis and therapies (Heppner & Miller, 1983; Lee et al., 2011; Seoane & De Mattos-Arruda, 2014).

1.1.4 Oncogenes and tumour suppressor genes

Cancer genes play a relevant role in cancer evolution and progression, and can be classified as tumour suppressors (TSGs) or oncogenes (OCGs) (Croce, 2008). The latter have origin in proto-oncogenes, which are genes that stimulate cell growth and viability. Once affected by activating mutations (e.g., amino acid amplifications or translocations), these proto-oncogenes can develop into its active status OCGs. This activation may ignite several cancer hallmarks such as cell cycle forward (i.e., commit the cell to pass from a G-phase to S-phase or mitosis), cell survival and evading apoptosis and cell senescence (Hanahan & Weinberg, 2011). Typical examples of oncogenic mutations are hotspot mutations of *KRAS* in codons at positions 12 and 13 (most frequent), and 18, 61, 117 and 146 (less frequent) (Bamford et al. 2004). These mutations are the result of natural selection and drive tumour evolution by activating PI3K-AKT and ERK signalling cascades (two of the main pathways involved in cancer events) (Janakiraman et al., 2010; Tan & Du, 2012).

In contrast, TSGs regulate normal cell division and replication processes, which cancer often impede. These can be inactivated by loss-of-function (LOF) events such as homozygous deletions, truncations, frameshifts and non-synonymous mutations. Notably, the inactivation of suppressor genes supports tumour evolution by impairing DNA damage response pathways and removing control mechanisms of cell growth, apoptosis and senescence (Kern & Winter, 2006; Vogelstein et al., 2013). In 1984, the first TSG (*RB*) was reported (Murphree & Benedict, 1984). This gene plays a fundamental regulator role in the cell cycle, by preventing the progression of the cell from the G1- to S-phase via inhibition of *E2F* transcription factors (Weinberg, 1995).

A tumour may develop due to the accumulation of driver alterations caused by cancer driver genes, which can be either TSGs or OCGs (Figure 2). However, depending on the context and tissue type, a gene may act as a TSG or OCG. For instance, *TP53* is commonly identified as a TSG in several cancer types, and acts as a driver gene when a mutation has altered both of its alleles. Nonetheless, TP53 can also attain gain-of-function alterations that grant oncogenic functions (Soussi & Wiman, 2015). Furthermore, several studies have reported that certain genes may share both functions depending on the mutational context (Bailey et al., 2018; Sanchez-Vega et al., 2018; Vogelstein & Kinzler, 2004). This versatile functionality illustrates how, in different environments, events such as loss-of-function of TSGs and gain-of-function in OCGs are naturally selected, in order to promote cancer adaption and evolution. Notably, a deeper investigation of these driver genes classes is a fundamental step towards improvement of precision medicine.

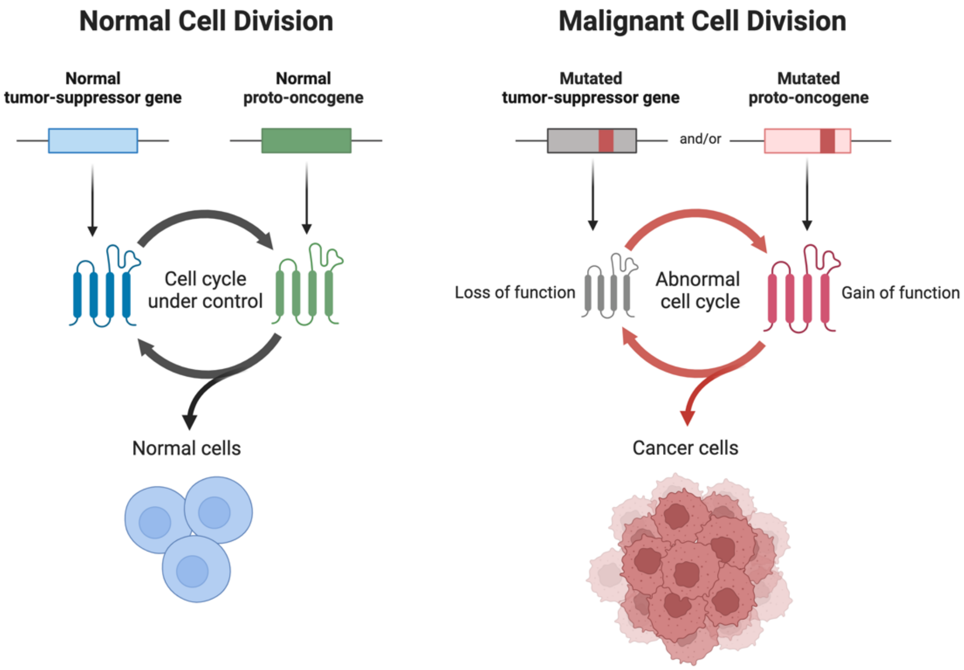

Figure 2: Role of normal and mutated tumour suppressor genes and proto-oncogenes in the cell cycle.

Normal genes regulate the cell cycle processes and originate normal cells. In contrast, mutated genes contribute to alterations in the cell cycle and the creation of cancer cells.
Figure created with BioRender.com using the template “Tumor Supressor Genes and Proto-oncogenes”.

1.2 Precision medicine in cancer

The rapid and increased development of NGS technologies in the past years, has enabled to unfold a deeper knowledge on the roles of genomics and the immune system in cancer research (Berger & Mardis, 2018; Malone et al., 2020; A. M. Tsimberidou et al., 2020). In parallel, the advancement in clinical trials empowered a shift from tumour type therapy to a gene targeting methodology by leveraging tumour molecular characterization, medical imaging and spatial profiling of immune tissues - thus, paving the way for more diverse and patient tailored biomarker analysis (Janiaud et al., 2019; Rodon et al., 2019; A.-M. Tsimberidou et al., 2012). The integration between these two pillars - tumour’s genomic alterations and clinical trials – constitutes the basis of precision oncology, and its continuous improvement is imperative to move from a “one size fits all” design to tailored and effective treatments (Manzari et al., 2021; Schilsky, 2010).

1.2.1 Biomarker discovery

Broadly, a cancer biomarker constitutes a set of molecules which may influence, predict or indicate the existence of a particular condition (Strimbu and Tavel 2010). In particular, biomarkers are key in stratifying between responders and non-responders tumours, and have a substantial number of applications in a clinical setting, including risk assessment, cancer progression monitoring, evaluation of response treatment and prognosis prediction (Henry and Hayes 2012). In my thesis, I will focus in distinguishing biomarkers of drug resistance, uncontrolled cell proliferation and drug response.

Cancer biomarkers comprise a wide set of disciplines, and can be partitioned into three main categories: diagnostic, prognostic and predictive (Sawyers 2008; Henry and Hayes 2012). The former is employed to detect if a patient holds a particular disease, condition or disease-subtype. For instance, the role of thyroid hormone receptor α1 (THRα1) as a strong diagnostic marker for lung squamous cell carcinoma (SCC) in non-small cell lung cancer (NSCLC) patients (Mohamed et al. 2021).

Within cancer patients, a prognostic biomarker indicates the likelihood of the disease's recurrence or progression regardless of therapy, and guides optimal clinical treatment procedures according to the likelihood assessment. Several significant examples of this type of biomarker have been established in breast cancer, such as the combined overexpression of c-erbB-2 oncogene and p53 tumour protein which alters cell proliferation mechanisms and is correlated with poor patient survival (Guerra et al. 2003; Sjögren et al. 1998; Nakopoulou et al. 1996; Beenken et al. 2001). As well as, the human epidermal growth factor receptor-2 (HER2) overexpression and amplification. HER2 plays a direct role in cell signalling deregulation, whilst driving the disruption of normal tissue growth and development, and is significantly linked to poor disease outcome in lymph node-positive breast cancer patients (Cooke et al. 2001; Hou et al. 2022).

On the other hand, predictive (or response) biomarkers can assess the response effectiveness to a particular therapy. In breast cancer, the overexpression of *HER2* is a response biomarker of HER2-targeting therapies such as trastuzumab (Piccart-Gebhart et al. 2005; Romond et al. 2005), lapatinib (Zardavas et al. 2013; Geyer et al. 2006) and pertuzumab (von Minckwitz et al. 2017; Swain et al. 2015). Similarly, in colorectal cancer, *KRAS* acts as a predictive biomarker where *KRAS*-activating mutations are linked with resistance and poor treatment response to EGFR inhibitors such as cetuximab (Van Cutsem et al. 2009; Allegra et al. 2009). Another notable example related to EGFR inhibition, is the case of patients with activating L858R mutation in non-small cell lung cancer (NSCLC) which are sensitive to treatments with afatinib and erlotinib (Chen et al. 2013; Jia et al. 2016).

In summary, cancer biomarkers are a notable tool in precision medicine in order to perform a more accurate diagnosis, develop effective personalised therapies, improve treatments' success rates, reduce side effects and contribute to a faster approval of cancer treatments to be used in clinics.

1.2.2 Therapies

1.2.2.1 DNA damage

DNA damage response mechanisms play a central role in several of the available cancer therapies used in clinics. This approach may render cells to compromise their damaged DNA responses, and therefore bypass relevant cell cycle checkpoints. Potentially, leading to an acceleration of the cell cycle and uncontrolled cell proliferation (Hoeijmakers 2009).

Although the act of purposefully create genomic instability seems controversial when discussing therapeutic alternatives, the accumulation of DNA mutations combined with impaired DNA damage response, can also trigger other cellular responses such as cell senescence and apoptosis, and consequently reduce tumour growth (Childs et al. 2014).

Unfortunately, due to the lack of specificity, all cells are affected during the DNA damage procedure. However, tumour cells are known to proliferate faster than healthy cells, and henceforth to be induced quicker to cellular control mechanisms. Hence, rendering them more susceptible to DNA damaging agents than healthy cells, and a suitable therapeutic choice for cancer treatment (Lord and Ashworth 2012). Currently, several DNA damaging agents are leveraged in clinics, such as ultraviolet (UV) rays (Rastogi et al. 2010), alkylating agents (Kondo et al. 2010) and cisplatin (Basu and Krishnamurthy 2010).

1.2.2.2 Conventional therapies

The most conventional cancer treatment procedures are surgery, radiotherapy and chemotherapy, and these can be employed unassisted or in combination, depending on the tumour status (Arruebo et al., 2011).

Typically, surgery performed in early stages of cancer is associated with an increased chance of success, since it is able to locally control the primary tumour, reduce inflammation, potentially remove all the existing cancerous cells and avoid additional treatments. Notably, this procedure is commonly employed in combination with chemotherapy, either prior to operation to shrink the tumour before resection, or post-operation to ensure the remaining cancer cell lines are destroyed. Taking into consideration the desired outcome, this procedure encompasses several types, such as curative surgery, preventive surgery and diagnostic surgery (King & Primrose, 2003). Although commonly used as a primary treatment, this treatment is not particularly successful within cancers where metastasis takes place in a later stage or not at all (e.g., head and neck, cervix and brain) and cancers whose infected tissue is not locally contained (e.g., leukaemia) (Tannock, 1998). Furthermore, surgery resection has been reported to contribute to metastatic seeding of cancer cells (Z. Chen et al., 2019; Tohme et al., 2017).

A widely distinguished modality used in clinics is chemotherapy, where cytotoxic drugs are administered as a way to induce cellular stress and destroy cancer cell lines. One type of chemotherapy is to apply alkylating agents, which provoke DNA damage by attacking cancer cells via attachment of alkyl groups to DNA base pairs or DNA crosslinking (Fu et al., 2012). In detail, this causes single and double DNA breaks which, if not repaired properly, lead to a potential replication fork collapse and, consequently, cellular death. During the course of this therapy, chemotherapeutic agents particularly target fast-proliferating cells, whether healthy or malignant. Thus, inducing several severe side-effects, which include weakened immune system, loss of hair and skin damages, in healthy cells such as blood and digestive tract cells (Schirrmacher, 2019). In addition, chronic phenomena like drug resistance and rapid drug metabolism represent some of the limitations of chemotherapy (Gutteridge, 1985).

Lastly, radiation therapy leverages high doses of ionising radiation to damage the genetic material of cancer cells, and consequently impair their division and proliferation capabilities. This method can ultimately lead to cell death or simply slow their growth, and is usually implemented with high-powered X-rays or proton beams via external or internal radiation (Baskar et al., 2012). Despite its remarkable efficiency in destroying locally a large quantity of cancer cells, radiation therapy can also damage surrounding healthy cells and tissues, and is associated with several side effects as a consequence of radiation administration (Bentzen, 2006).

1.2.2.3 Targeted therapies

Alterations in the normal regulation of cancer signalling networks, which are typically driven by a specific molecular target (e.g., a protein), are key instigators in cancer growth and progression. Targeted cancer therapies focus on the development of cancer drugs designed to interfere or target these specific cancer-driver nodes (Sawyers 2004).

The identification of molecular targets is based on a deep understanding of cancer biomarkers, and varies greatly from conventional approaches such as chemotherapies. Notably, targeted therapies aim for specific abnormally active proteins instead of targeting all fast-proliferating cells, as well as, present regularly a cytostatic instead of a cytotoxic approach (Breitbach et al. 2010).

Several promising examples of targets involving members of kinases (e.g., MAP kinase pathway) can be found being currently used in clinics. One of the most frequent mutations observed in melanoma patients is BRAF^{V600E}, encountered in up to 60% of all patients. This mutation is linked to several mechanisms related to melanoma evolution, such as metastasis, evasion of apoptosis, senescence and immune response (Maurer et al. 2011). Notably, BRAF^{V600E} is highly sensitive to treatments with FDA approved inhibitors, namely BRAF inhibitors such as vemurafenib and dabrafenib, and MEK inhibitors like trametinib and cobimetinib (Sosman et al. 2012; Ascierto et al. 2012).

Another notorious instance of targeted therapy is the tyrosine kinase inhibitor (TKI) imatinib in the treatment of chronic myeloid leukaemia (CML) patients (Cortes et al. 2005; Savage and Antman 2002). In CML, the *BCR-ABL1* fusion gene plays a vital driver role in the initiation and development of the malignancy, and is frequently detected in CML patients. During treatment, imatinib essentially targets the ATP binding site of ABL1 and prevents its and succeeding kinases activation (Quintás-Cardama and Cortes 2009).

Although targeted therapies are a valuable and revolutionary example of cancer therapy, unfortunately cancer cell lines can develop resistance to targeted therapy (Braun et al. 2020; Tsuruo et al. 2003). A possible driver is tumour evolution, which can induce drug resistance or introduce new ways for the cancer to develop, and, consequently, disable the possibility for the drug agent to successfully interact with the target biomarker.

A potential strategy to overcome these issues is to combine targeted therapies with other target agents or conventional cancer treatments through rational drug combinations.

1.2.3 Drug resistance

Drug resistance is a clinical response that can be intrinsic (primary) or acquired (secondary) to a specific treatment. While administering a target therapy, before all cancer cells are eliminated, due to evolutionary pressure and fast cell proliferation, a subpopulation of cells may gain mechanisms to cope (i.e., resistance). Subsequently, these resistant subclones can outgrow the cancer cells responding to the target agents, and lead to treatment failure (Vasan et al., 2019).

A noteworthy example of a resistance mechanisms involves the response of *TP53* mutants to the murine double minute 2 (MDM2) inhibitor nutlin-3a. The tumour suppressor p53 is a master regulator of several cancer key mechanisms such as apoptosis, senescence and cell cycle arrest, and is notoriously essential to combat cancer initiation and development (Lane 1992). Mutations in *TP53* are relatively common and occur in approximately 50% of all cancer types (Toledo and Wahl 2006; Hollstein et al. 1991). In cancers with wild-type *TP53*, its function is often suppressed by MDM2 overexpression. To restore *TP53* functionality, non-genotoxic drugs targeting MDM2, such as nutlin-3a, are employed (Vassilev et al. 2004). However, acquired resistance can arise when cancer cells develop other mutations not inhibited by MDM2, such as MDMX overexpression (Hu et al. 2006). As a consequence, *TP53* aberrations constitute an indirect example of cancer cells resistance to nutlin-3a treatment (Hientz et al. 2017).

Mutations of the epidermal growth factor receptor (EGFR) frequently occur in lung adenocarcinoma patients, and are commonly marked as drivers of non-small cell lung cancer (NSCLC) (Paez et al., 2004; Shigematsu & Gazdar, 2006). Typically, these mutations are manifested as small in-frame deletions in exon 19 or the point mutation L858R, and are strongly sensitive to TKIs, including gefitinib and erlotinib (Lynch et al., 2004; Pao et al., 2004). However, the efficacy of these TKIs can be impaired by the acquisition of an additional point mutation T790M, which confers resistance to the treatment (Pao et al., 2005). Notably, the T790M mutation changes the drug binding pocket of EGFR, resulting in steric interference of the TKIs binding and leading to EGFR-TKIs resistance (Kobayashi et al., 2005). The discovery of this example of secondary resistance became a clinical case of success, where gefitinib was ultimately approved as a first in line drug treatment for NSCLC patients without T790M mutation (Kazandjian et al., 2016).

1.2.4 Drug combinations

Combination therapy is a revolutionary treatment modality which combines multiple therapeutic methods. The principle behind it is to augment treatment efficiency by combining approved therapies with distinct target pathways in a synergistic or distinctive way (Bayat Mokhtari et al., 2017). Several combination screens are currently available (Menden et al., 2019; O'Neil et al., 2016) which enable the identification of new cancer combination biomarker candidates for an improved therapeutic outcome.

In addition to a time and cost-effective procedure, this process also aims to minimise the possibility for cancer cells to acquire drug resistance. Notably, due to genetic diversity within tumours, the likelihood of a small group of cells to develop resistance to a single drug is high when considering a monotherapy approach. However, when assuming a combinatorial modality, the risk of treatment failure due to drug resistance is reduced since multiple pathways are simultaneously targeted (Sharma et al., 2017).

Furthermore, due to its synergistic or additive approach, combination therapy provides the possibility of administering lower dosages for each drug involved in the process (Schmucker et al., 2021), as well as, prevents toxic effects on healthy cells whilst producing cytotoxicity to cancer cells via drug antagonism (Blagosklonny, 2005; Chou, 2006).

The potential of combination strategies is leveraged in various cancers, namely leukaemia (Ravandi et al., 2010), breast cancer (Ji et al., 2019), melanoma (Eroglu & Ribas, 2016). For instance, consider the overexpression of HER2, which is found approximately in one third of breast cancers, and is correlated with increased cancer growth and progression (Slamon et al., 1987). A standard therapy for these patients is the administration of the HER2 inhibitor trastuzumab (P. Carter et al., 1992). However, a significant fraction of patients expresses inherited or developed resistance mechanisms to this drug (Gajria & Chandarlapaty, 2011; Narayan et al., 2009). Commonly, a standard therapy is to administrate trastuzumab with chemotherapeutic agents, such as docetaxel and paclitaxel, as a first in line treatment regimen (Tolaney et al. 2015; Swain et al. 2015). Notwithstanding, for HER2-positive tumours, there are several synergistic drug combination protocols that do not involve chemotherapy. Namely, the dual HER2 blockage with trastuzumab and pertuzumab (Swain et al. 2015), which target different domains of HER2, or combinations of trastuzumab with TKIs lapatinib or neratinib (Blackwell et al. 2019; Yuan et al. 2022)

.

1.3 Pharmacogenomics landscape

A fundamental aspect when studying cancer is to understand the pillars behind its development and evolution. Cancer is driven by changes in cellular and molecular mechanisms that govern cell life. In the past, it was initially believed that cancer biology causality could be summarised by a small number of cancer genes which triggered tumorigenic cell signalling pathways (Land et al. 1983). Nowadays, with the development of NGS technologies coupled with advanced computational approaches, cancer research offers a more holistic and comprehensive overview of its hallmarks (Hanahan, 2022; Hanahan & Weinberg, 2011).

1.3.1 Multi-omics

1.3.1.1 Transcriptomics

Conceivably, apart from accumulated mutations, the same amount of DNA material is roughly present in all cells available within a particular organism. Nonetheless, without losing the available genetic material, cells may exhibit vastly different functionalities due to somatic mutations and differentiation processes (Alberts et al. 2014). This diversity is primarily reflected in the distinct genes expressed within these cells, and their expression is regulated in several key points. In detail, should an alteration in a gene in the DNA occur, this information will be stored in a temporary template and be posteriorly transcribed from the DNA into a single strand ribonucleic acid (RNA). This mechanism is called transcription and the resulting transcript includes all intronic regions and is often referred to as pre-mRNA. Following, this transcript is furtherly processed via splicing, where the introns are removed, and converted to a mature product called messenger RNA (mRNA). Posteriorly, through a process labelled translation, the ribosome reads the sequence and builds a specific string of amino acids, which is later incorporated to a functional protein (Crick, 1958) (Figure 3).

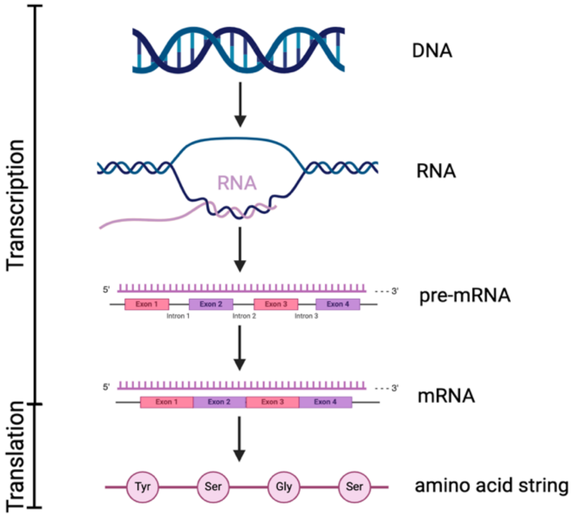

Figure 3: Illustration of transcription and translation processes.

A double helix DNA is unwound by an RNA polymerase and a segment of DNA is transcribed into a single RNA molecule. The resulting pre-mRNA transcript contains both intron and exon regions. Subsequently, the primary pre-mRNA is spliced, where the intron regions are removed, developing into mRNA. Finally, the mRNA is translated to an amino acid string.
Figure created with BioRender.com.

Several methods are available to quantify RNA transcripts, such as methodologies based on hybridization (e.g., microarray analysis) or sequence (e.g., tag-based sequencing). One of the most distinguishably used methods is RNA sequencing (RNA-seq) (Emrich et al., 2007; Lister et al., 2008; Wang et al., 2009). This technique leverages NGS technologies advances and aims to delineate and quantify transcriptome families - such as mRNA, small RNA, transfer RNA (tRNA) and ribosomal RNA (rRNA) – in a high-throughput manner over continuous modifications during cellular development.

The RNA-seq workflow starts with RNA extraction, subsequently it is copied by a method based on transcription and a complementary DNA (cDNA) is synthesised. Next, a sequencing library is prepared, where captured mRNA molecules are fragmented and reverse transcribed, adapters are connected to the cDNA molecules and the library is amplified by PCR. Following these protocols, quality control of the samples is assessed, reads are trimmed and filtered (e.g., low quality reads or duplicates are removed), reads are mapped to a reference genome or transcripts, and thereafter a quantification of how many reads overlap the given genome is made. These steps are complemented by normalisation and filtering of samples, followed by a statistical assessment of which genes are expressed and their individual level of expression (Wang et al., 2009).

An interesting application of the RNA-seq technology is how gene expression profiling can be leveraged to stratify tumours' subgroups. Typically, unsupervised clustering approaches such as hierarchical clustering and network clustering aid on the identification of molecular pattern signatures (Sørlie et al. 2003; Perou et al. 2000; Garber et al. 2001). For instance, in colorectal cancer (COREAD) the primary tumours have been classified into four consensus molecular subtypes (CMS; CMS1-4) (Guinney et al. 2015). In addition, the inclusion of clinical information complements these analyses and recognizes potential biomarkers linked to disease prognosis (Garber et al. 2001; Bramsen et al. 2017).

Furthermore, the output from RNA-seq experiments is frequently used to perform differential gene expression analysis (DGEx). Given the originated data is discrete and represented by count values, this allows the application of binomial based mathematical procedures to reveal significant markers between biological conditions. Several open-source software specialised in these analyses with pre-processing, statistical models and complementary visualisations are available to the research community, such as edgeR (Robinson et al. 2010), DESeq2 (Love et al., 2014) and Limma (Ritchie et al., 2015).

1.3.1.2 Proteomics

Inside a cell, besides the genome, there are functional proteins responsible for active roles, such as cell division, structure and organisation (Karplus & McCammon, 1983). Through a translation procedure, a segment of the DNA, which has been converted to RNA and mRNA (transcription, described in the previous section), is leveraged to produce proteins. Each piece of mRNA prescribes the order in which the sequence of amino acid blocks should be organised and assembled, and, thereafter, a protein is synthesised. This mechanism of converting genetic information, from DNA to protein, is acclaimed as "The Central Dogma" of molecular biology (Crick, 1958).

Proteomics is the field which studies the collection of proteins in an organism (Graves & Haystead, 2002). In particular, it investigates an extensive number of processes, namely protein expression profiling, protein-protein interactions (PPI), protein modifications and protein functions (Pandey & Mann, 2000). Proteomics is complementary to the Genomics and Transcriptomics research, and provides a comprehensive understanding of cancer mechanisms since it focuses on gene activity outcome products and proteins are typically more stable than RNA molecules. Despite its advantages, there are several challenges related to proteomics experiments and analyses. Explicitly, low reproducibility of data collection and processing across experiments, missing to detect low abundance proteins and lack of sensitivity and specificity of protein identification tools (Garbis et al., 2005; O. T. Schubert et al., 2017; Tabb et al., 2010).

Currently, several methods are available to analyse proteomics data, namely based on mass spectrometry (Aebersold & Mann, 2003; Domon & Aebersold, 2010; Medzihradszky et al., 2000). One of the advantages of proteomics in respect to transcriptomics, is the possibility to measure post-translational modifications (PTMs; e.g., phosphorylation and sulphation) of proteins, which reveals various proteins' functional status, e.g., protein activity, stability, localization and interaction with other proteins (F. Cheng et al., 2018; Himmelstein et al., 2017; Kong et al., 2020; Mann & Jensen, 2003; Zhou et al., 2014). For instance, methods such as CausalPath (Babur et al. 2021) have been successfully leveraged in research to predict pathway activities.

1.3.2 Phenotypic data

Investigation of phenotypic traits offers a translational bridge between the genotype of an organism and disease pathophysiology. Precisely, the phenotype of an organism is an observable manifestation of its genotype, such as height and hair colour. Nonetheless, these physical traits can be influenced by external factors, namely environment (e.g., sun exposure and diet) (Hunter, 2005). A particularly interesting phenomenon is the generation of genetic modifications that result in phenotypes with varying growth degrees and/or survival advantage, rending them relevant when exploring therapeutic mechanisms. For instance, tumour microenvironment may influence the autophapy process in cells, alter its phenotype and drive drug resistance (Amaravadi et al., 2016).

In order to dissect phenotypic profiles one can leverage human phenotype ontologies, such as Human Phenotype Ontology (HPO) (Köhler et al., 2021) and Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2015). These phenotypic resources provide a link between disease's biomedical information and phenotypic terms, and therefore facilitate the investigation of genetic diseases.

Various methods focus on the exploration of models based on phenotypic outcomes and multi-omics (i.e., genomics, transcriptomics, among others) due to their direct association and correlation (Chang et al., 1999; Chia et al., 2017; Mo et al., 2013; Sadanandam et al., 2013). These approaches are hypothesised to yield more thorough and effective models, with the identification of relevant cancer specific genes, pathways or other biomarkers and, ultimately, pave the way to a novel and more rigorous precision oncology.

Another pertinent phenotypic angle is the stratification between responders and non-responders in a pharmacogenomics context. Distinct genetic profiles may render patients to respond differently to a drug treatment, thus categorising populations in two or more distinctive subgroups depending on their drug-reaction profiles (Roses, 2000). This phenomenon raises the investigation of the genomics' variant profiles within the subpopulations, and the identification of possible markers which can improve treatment efficacy and minimise adverse drug effects (Evans & McLeod, 2003). Several notable examples of this classification can be found in oncology research with significant impact in precision medicine. For instance, the compound gefitinib was established as the first drug on the treatment of patients with advanced NSCLC (Herbst et al., 2004), after uncovering sensitive responses in subpopulations with EGFR mutations (Lynch et al., 2004; Paez et al., 2004).

1.3.3 Biological prior

Currently, with the advance of technology, there has been a substantial increase in the quantity and quality of multi-omics datasets generated to research cancer's cellular processes, ultimately driving more efficient therapies. Although this big data outcome presents an ideal scenario to further investigate the disease, these experiments often originate multiple potential candidate omics (dimensionality curse) and are convoluted with noise (Cantini et al., 2021; Jiang et al., 2022). As a consequence, this poses a challenge in the recognition of true positive candidate omics with functional relevance to cancer.

On this account, additional tools of biological prior are typically leveraged in order to

- Boost the statistical power of machine learning and statistical applications;
- Create more evidence across several omic layers;
- Validate analysis based on *gold standards*;
- Aid on the interpretation of results.

Ultimately, the integration of biological prior knowledge enables a more holistic interpretation of high-throughput results, as well as, underlines disease aetiology more effectively in comparison to only omics-based approaches (Jones et al., 2008; Martínez-Jiménez et al., 2020).

1.3.3.1 Biological Pathways

A biological pathway describes a set of biochemical reactions performed by a group of molecules within a cell, during cellular programs (e.g., cell division and cell death) (Hanahan & Weinberg, 2000; Vogelstein & Kinzler, 2004). These processes are activated when an initial molecule binds to a protein receptor and subsequently extends through a chain activation of another molecule. This procedure continues until all the molecules in the signalling pathway are activated and the cell function is finalised. If an issue occurs during this activation, aberrations may occur in the system and trigger disease development.

The analysis of biological pathways enables the identification of genes with a relevant and central role in a disease, as well as, to establish the signalling pathways associated with these genes (Graham & Xavier, 2020; Paczkowska et al., 2020). For instance, the overexpression of the E3 ligase, MDM2 provokes the suppression of *TP53*. Consequently, the inactivation of p53 induces an evasion of several cellular processes, such as cellular senescence and apoptosis, and leads to uncontrolled cell proliferation and survival (Ozaki and Nakagawara 2011; Aubrey et al. 2018). Therefore, several MDM2 inhibitors have been developed to avoid the interaction between MDM2 and p53, and assure the p53 activity of key cellular functionalities (Zhao et al. 2015).

Furthermore, biological pathways also empower the investigation of mutual exclusivity in cancer. It has been demonstrated that distinct tumours typically progress by activating the same oncogenic pathways (Sanchez-Vega et al. 2018), and key cancer driver oncogenes (e.g., *EGFR* and *KRAS*) frequently manifest mutually exclusive properties if mutated in the same signal pathway (Kandoth et al. 2013). Thence, several methods based on oncogenic network modules and statistical tests have been developed to identify mutually exclusive gene and pathway modules (Pulido-Tamayo et al. 2016; Ciriello et al. 2012; Babur et al. 2015).

Notably, the understanding of signalling pathway mechanisms ultimately enables a deeper and robust investigation of the disease, and is essential for the design of new targeted therapies. For example, in breast cancer the overexpression of HER2 leads to an abnormal activation of the PI3K/AKT and Ras/ERK downstream pathways, and subsequent uncontrolled tumour growth and poor clinical prognosis (Iqbal and Iqbal 2014). Therefore, drugs such as trastuzumab were developed to directly target HER2 and prevent its activation (Rimawi et al. 2015; Hudis 2007).

In cancer, the genetic alterations responsible for processes that may lead to tumorigenesis (e.g., apoptosis and cell growth) can be associated with dysregulations in several signalling pathways (Hanahan & Weinberg, 2000). Moreover, these perturbations co-occur at fluctuating frequencies and distinct pathways across different tumour and tissue types. Examples of distinguished pathways in cancer that are usually activated are the transforming growth factor beta (TGFβ) (Massagué, 2008) and the receptor tyrosine kinase (RTK)/Ras/MAP kinase (MAPK) (Santarpia et al., 2012).

Throughout the years, has been established several pathway databases that provide the users a free search tool of available and curated signalling pathways, such as Reactome (Gillespie et al., 2022), KEGG (Kanehisa et al., 2023) and WikiPathways (Martens et al., 2021). These repositories serve as a pillar in several bioinformatics studies to contextualise and interpret results in a functional matter (Ciriello et al., 2013; Reimand et al., 2019; Son et al., 2013), as well as, to empower sophisticated computational frameworks (Adam et al., 2020; Colaprico et al., 2020; Gao et al., 2019; Jiao et al., 2020).

1.3.3.2 Protein-Protein Interaction Networks

Biological networks can be leveraged to investigate alterations in cellular pathways derived by modifications in the protein activity, and enrich biological interpretability by integrating several biological layers. Furthermore, these networks offer a higher coverage of the disease and can be exploited for multi-omics data integration frameworks (Ma & Zhang, 2019; Yan et al., 2018) and on drug discovery (Altieri, 2008; Hopkins, 2008).

Within the biological networks umbrella, in this thesis I highlight the protein-protein interaction (PPI) networks (Stelzl et al., 2005). PPIs describe the biochemical interactions between proteins across different organisms during cellular processes (e.g., cell transcription and translation) under diverse environmental situations.

Typically, PPI networks are modelled as graph representations, where the nodes reflect the proteins and the edges denote the physical or functional of each pair of interacting proteins. These networks can be represented as directed or undirected graphs. The latter is the most common type of connection in PPIs and dictates if a protein A physically binds protein B. On the other hand, directed graphs provide information on the hierarchical direction of a reaction flow. This orientation can be applied, for example, to model metabolic reactions, and has been successfully leveraged in several biological segments (Cao et al., 2014; Vinayagam et al., 2011).

The graph representation of PPIs empowers the discovery of subtype-specific modules consisting of topological (e.g., locally highly connected interacting proteins) and functional (e.g., highlight proteins with imperative role in each module and pathways) properties of interest (Kim & Kim, 2018; Yin et al., 2021).

Several PPI resources are publicly available and these are mainly stratified between datasets with experimentally detected interactions from literature, such as BioGRID (Oughtred et al., 2021) and IntAct (Del Toro et al., 2022), and databases originated through computationally inferred interaction, such as STRING (Szklarczyk et al., 2018). In addition to these resources, there are available software's which compile several of these PPI datasets, and provide tools to leverage and model them within bioinformatics analysis, including Omnipath (Türei et al., 2021), HIPPIE (Alanis-Lobato et al., 2016) and Cytoscape (Shannon et al., 2003).

The integration of PPI networks with other genomic (e.g., gene networks and gene expression), phenotypic and/or drug information has been extensively applied in the investigation of disease biomarkers and potential drug targets in cancer and other diseases (Y. Cheng et al., 2019; Isik et al., 2015; Pu et al., 2022; Vinayagam et al., 2014). Furthermore, these network analyses can be further extended to model patient-specific systems and aid the design of personalised precision medicine (Li et al., 2017; Vaske et al., 2010).

1.3.3.3 Ontology Databases

Biologic ontologies serve as a complementary tool of the previously described systems when investigating disease mechanisms. These provide a comprehensive relationship between biological or phenotypic terms and genes/proteins within these terms, and are typically organised in a directed acyclic graph form. Accordingly, the use of ontologies is frequently recognized in data integration approaches to interpret findings systematically and identify potential disease genes/proteins of interest (Yang et al., 2015; Y. Zhang et al., 2022).

- **Gene Ontology**

Gene Ontology (GO) is a knowledge base resource with functional biological annotations of genes stratified in three different categories: biological process (BP), molecular function (MF) and cellular component (CC) (Ashburner et al., 2000). The database is organised in a graph form, where a node represents a GO term and edges depict functional relationships between terms. Moreover, the ontologies are arranged in a hierarchical structure, where the information becomes more specific as one navigates from the parent to the child nodes.

GO is estimated to currently encode more than 7.4 million biological concepts of several organisms and be extensively referenced in systems biology analysis published manuscripts. The database is mainly leveraged to perform enrichment analysis and integrate with other omics and databases (e.g., PPIs) (Tomczak et al., 2018).

- **Human Phenotype Ontology**

The study of phenotypic features and how these dynamically change as a consequence of genomic variation is essential to unveil the genes involved and their biological functions. In addition, the examination of diseases with shared phenotypic traits can reveal disease families and, consequently, disclose signalling modules involved (Brunner & van Driel, 2004; Rodriguez-Pinilla et al., 2007; Turner & Reis-Filho, 2006).

The Human Phenotype Ontology (HPO) database was developed to aid in the investigation of phenotypic data in a disease context (Köhler et al., 2021). Currently, HPO is composed of more than 13,000 terms based on clinical information annotation available in literature, and OMIM (Amberger et al., 2015), Orphanet (Weinreich et al., 2008) and DECIPHER (Firth et al., 2009) repositories. Each term represents a phenotypic anomaly, and these are modelled in an ontology structure similar to the one of GO. Furthermore, the information in the ontology is organised in six distinct sub-ontologies: phenotypic abnormality, mode of inheritance, medical history, disease frequency, clinical modifier and blood group.

1.3.4 Pharmacological screens

Effective and efficient biomarker discovery extensively relies on high-throughput screens (HTS). Although the most precise models would be achieved through *in vivo* models, these models' application is highly unethical and unfeasible in economic and safety terms (Workman et al., 2010). Therefore, drug screens assessed over a wide range of drugs against a substantial amount of cancer cell lines are leveraged since these enable to capture tumour heterogeneity and elaborate biomarker hypotheses to later guide clinical *in vivo* research (H. Gao et al., 2015; Kulasingam et al., 2010).

1.3.4.1 Cell models

Cancer cell lines are broadly leveraged as *in vitro* model systems of cancer tumours in cancer biomarker identification and validation, and drug response investigation.

The complex mechanism to produce a new cell line initiates by modifying and stabilising the function of the cells in order to elude cellular mechanisms within the cell cycle, such as apoptosis and senescence (Pascual et al., 2019). From this, a cell line capable to grow without limitations emerges, which is fundamental when taking into consideration the environmental differences between *in vivo* and in *vitro* conditions (Rockwell, 1980; van Staveren et al., 2009).

In parallel with the first discoveries on cancer cells, during the 20th century it was cultured the first cell line, called HeLa*,* from a cervical carcinoma patient (Gey, 1952). This discovery was later followed by the creation of the first large-scale high-throughput screen by the National Cancer Institute (NCI). The project was established with 59 unique tumour cell lines across nine diverse cancer types (e.g., breast, leukaemia and melanoma - pan-cancer screening), and labelled as NCI-60 (originally 60 cell lines were screened, however two cell lines were identical clones) (Shoemaker, 2006). The project was envisioned as a replacement of immunodeficient mice tumours for anti-cancer therapeutic screenings (Winograd et al., 2013), and can be leveraged to investigate the mode of action (MoA) of compounds through pattern recognition models.

Currently, the biggest HTS available are the Genomics of Drug Sensitivity in Cancer (GDSC) (Garnett et al., 2012; Iorio et al., 2016), Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) and the Cancer Therapeutics Response Portal (CTRP) (Basu et al., 2013; Seashore-Ludlow et al., 2015). These screens contain over 1,000 cancer cell lines treated with hundreds of compounds in several cancer tissue types, hence empowering the exploration of pan-cancer biomarkers.

Despite its simplistic nature, *in vitro* cell lines enable the investigation of experimental work in a cost effective, practical, fast and ethical manner (Kaur & Dufour, 2012). Furthermore, in order to overcome the lack of specificity of *in vitro* models, currently large-scale cell line panels (e.g., GDSC, CCLE and CTRP) are leveraged in research to model cancer heterogeneity and explore the diverse outcomes in clinics (Gillet et al., 2013).

In addition, it is possible to integrate the phenotypic information of cell models with omics, namely genomics, transcriptomics and proteomics, thus opening a large window of opportunity for researchers to study potential biomarkers in cancer and hypothesise treatment approaches (Huang & Vakoc, 2016).

Notwithstanding the extensive advantages in leveraging cell line screens, one should bear in mind the exhaustive and prolonged procedure of cell line creation, the possibility of cross-contamination (Buehring et al., 2004), the lack of ability to capture the full complexity and heterogeneity of tumours, and the non-matched outcomes between *in vitro* and *in vivo* (Weinstein, 2012).

1.3.4.2 GDSC, CCLE and CTRP screens

Precision oncology aims to optimise patient treatment based on robust and specific genetic and molecular characteristics. To this end, large-scale pharmacogenomic screens, based on cancer cell lines models and tested against a wide range of approved compounds, are leveraged to represent tumour heterogeneity and explore biomarkers of drug sensitivity, patient stratification and personalised treatments (S. V. Sharma et al., 2010).

After the development of the NCI-60 screen, efforts were allocated to expand the amount of cell lines and candidate drugs screened in pharmacological screens. From this, several noteworthy large HTS studies were generated, such as GDSC (Garnett et al., 2012; Iorio et al., 2016), CCLE (Barretina et al., 2012), CTRP (Basu et al., 2013; Seashore-Ludlow et al., 2015) and Genentech Cell Line Screening Initiative (gCSI) (Haverty et al., 2016).

The initial release of GDSC comprised 639 cancer cell lines and 130 drugs (Garnett et al., 2012). Subsequently, the screen was augmented to 1,001 cell lines and 265 compounds (Iorio et al., 2016). Currently, the project is frequently updated with new cell lines, compounds and drug response metrics, and its data is incorporated within two distinct datasets GDSC1 and GDSC2. On another hand, the CCLE project contains pharmacological screens of 24 drugs against 479 cancer cell lines and genomic data for 947 cell lines (Barretina et al., 2012). The CTRP first release included 185 compounds tested against 242 cancer cell lines (CTRP v1) (Basu et al., 2013), and was later expanded to 481 compounds and 860 cell lines (CTRP v2) (Seashore-Ludlow et al., 2015). The genetic information available in CTRP was integrated from the CCLE resource.

Moreover, all these pharmacological screens comprise a full molecular characterisation of the cancer cell lines. In detail, both GDSC and CCLE contain mutation profiles, copy number variations (with Affymetrix SNP6.0 arrays) (Freeberg et al., 2022) and gene expression information (with Affymetrix U219 and Affymetrix U133 Plus 2.0, for GDSC and CCLE, respectively) (Sarkans et al., 2018). In addition, GDSC also provides DNA methylation profiles (with IlluminaHumanMethylation450 BeadChip) (Edgar et al., 2002).

The experimental protocols of GDSC and CCLE vary in terms of the tested drug concentrations. Whilst GDSC tailors the concentration range to capture the sensitivity window, CCLE considers a fixed concentration range to all the compounds. This distinction resulted in several discrepancies in the drug response outcomes between the pharmacological screens (Haibe-Kains et al., 2013). Therefore, several studies explored the dissimilarities between the GDSC and CCLE screens (Bouhaddou et al., 2016; Cancer Cell Line Encyclopedia Consortium & Genomics of Drug Sensitivity in Cancer Consortium, 2015; Geeleher et al., 2016), and concluded that the observed divergences were mainly driven by distinct experimental and data analysis designs.

In addition, other studies also examined the drug response of GDSC, CCLE and CTRP screens, whilst having in consideration the variability intrinsic in the methodology and protocol designs (Pozdeyev et al., 2016; Xia et al., 2022). In essence, a satisfactory agreement between the findings across the three HTS was concluded, rendering these screens as valid tools for the exploration of cancer biomarkers.

The identification of robust and patient-tailored sensitive biomarkers is the main goal of HTS such as GDSC, CCLE and CTRP. Owing to the extensive molecular characterisation of cell lines and their large-scale pharmacological screening, these screens enable to define genomic, transcriptional and epigenetic profiles that describe the association between a compound and a cancer cell line (Ben-David et al., 2018). Furthermore, these screens include a wide landscape of cancer types, and provide information on compounds of clinical interest for targeted therapies, which facilitates the investigation of personalised and therapeutic biomarkers (Aldonza et al., 2020; Krug et al., 2020; McDermott et al., 2007; Safikhani et al., 2017).

Notwithstanding the various advantages of leveraging pharmacological screens, these remain an open research field with need of improvement. Notably, an expansion of current screens is of the utmost relevance to empower the investigation of further cancer biomarkers. However, not only this expansion is bounded by available labour and costs, but current screen designs suggest the majority of the screened cell lines do not respond to the tested compounds (Bouhaddou et al., 2016). In addition, current screens present difficulty to recognize resistance biomarkers due to cytotoxicity and limitation of tested concentrations (Figure 4), and rather identify potential candidates as non-responders (Ayestaran et al., 2020). Therefore, future pharmacological screens need to overcome these limitations by leveraging new methods of cell line screening and develop adequate cancer specific designs (Ling et al., 2018).

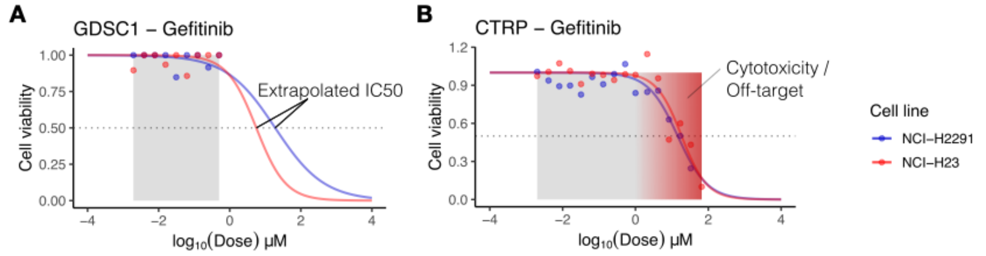

Figure 4: Challenges to identify resistance biomarkers in high-throughput screening.

Response curves of cell lines NCI-H2291 and NCI-H23 treated with gefitinib. **(A)** Response based on raw cell viability data from GDSC. Exemplification of a non-responder cell line where the IC$_{50}$ value was extrapolated beyond the tested drug concentrations. **(B)** Raw pharmacological data obtained from CTRP drug screens. Tested drug concentrations are higher and potentially cytotoxic. Consequently, this may lead to off-target effects in the cell line.
This image is a compilation of panels from the original supplementary figure 2 “Synopsis of pharmacology screens and response examples” by (Ayestaran et al., 2020) under a CC-BY-4.0 license (http://creativecommons.org/licenses/by/4.0/).

1.3.4.3 Drug response metrics

Cell line drug response is commonly assessed and represented by a cell viability assay succeeding multiple treatments of cancer cell lines over a wide range of drug concentrations. Henceforth, after a specific period of time, the number of viable cells is quantified in respect to a control substance and, subsequently, drug-cell line responses are derived to each given concentration (Figure 5).

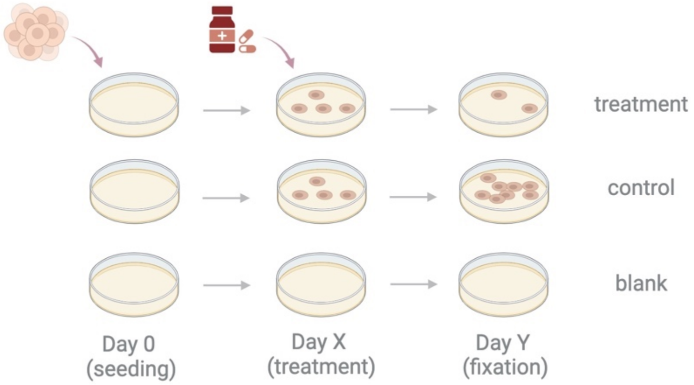

Figure 5: Typical cell viability assay protocol.

At day 0, treatment and control plates are seeded with cell lines. Posteriorly, treatment with a specific drug concentration and control compound are given to treatment and control cell lines, respectively. In the end, the intensities of the treatment, control and blank are quantified and cell viabilities are estimated.
Figure created with BioRender.com.

Drug-cell response values are typically represented by curve fitting processes (Figure 6 A-B), including sigmoid models (Vis et al., 2016) and Gaussian processes (GP) (D. Wang et al., 2020). These methods are summarised by curve measurements (Figure 6 C-D), such as the concentration necessary to reduce cell viability by half (IC_{50} – half maximal inhibitory concentration) or the area under the curve (AUC). For instance, the GDSC and CCLE screens represent the drug response via IC_{50} values (Barretina et al., 2012; Iorio et al., 2016), whilst CTRP uses AUC values (Seashore-Ludlow et al., 2015).

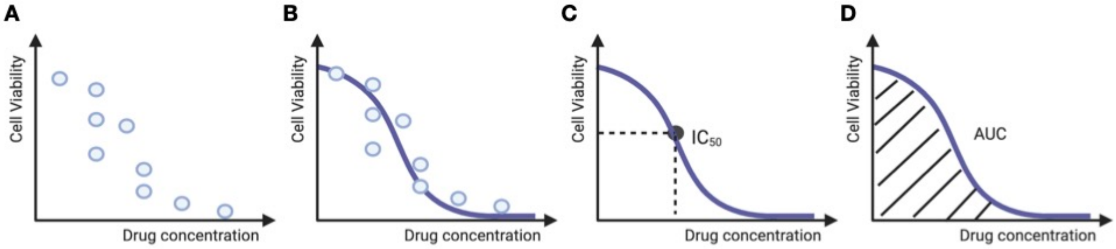

Figure 6: Raw drug response, curve fitting and drug response measures.

(A) Exemplification of initial cell viability values as a function of drug concentration. **(B)** Raw drug response is typically fitted using curve fitting methods such as sigmoid models and Gaussian Processes. In this figure it is illustrated the case of a responder cell line. **(C)** The half maximal inhibitory concentration (IC_{50}) and **(D)** the area under the curve (AUC) are commonly leveraged as metrics to characterize the drug response.
Figure created with BioRender.com.

Although it is conveniently manageable and interpretable to summarise the whole drug response in one representative metric, this also entails several limitations. For instance, AUC values are difficult to interpret and compare across different screens since it depends on the considered drug concentration ranges

(Pozdeyev et al., 2016). Moreover, IC_{50}'s are calculated by interpolation or extrapolation methods, depending if the IC_{50} value is within the tested concentration range or not, respectively. In the latter situation, the assigned IC_{50} value will not be an adequate representation of the drug response (Pozdeyev et al., 2016). For example, the CTRP project attempts to approach this problem by testing cells over high drug concentrations, however this is subsequently affected by the possibility of cytotoxicity.

In addition, metrics such as IC_{50} do not provide information regarding the curve behaviour, like the level of response sensitivity and the amount of noise in the data (Haibe-Kains et al., 2013; Haverty et al., 2016). Hence, there is no assessment on the reliability and specificity of the drug response outcomes. In order to overcome these issues, several studies (Bayer et al., 2023; Di Veroli et al., 2015; D. Wang et al., 2020) proposed to additionally leverage curve error and slope metrics, however these works are still not accepted as new gold-standards. Lastly, these measures do not take into consideration the cell growth rate when computing the drug response, neglecting the effects of cell proliferation rates or seeding densities of a cell line on the drug response (Hafner et al., 2016).

1.4 Computational methods in pharmacogenomics

Over the past years, biological research has collected a considerable amount of pharmacogenomic data thanks to new technologies such as DNA sequencing and HTS. This increase in data complexity has sparked the need to develop computational approaches that are able to deal with large datasets and investigate new hypotheses (Figure 7). Namely, various statistical and machine learning (ML) methods are leveraged to jump from the raw data into biological interpretable results in a systematic and reliable manner (Boniolo et al., 2021; Farnoud et al., 2022).

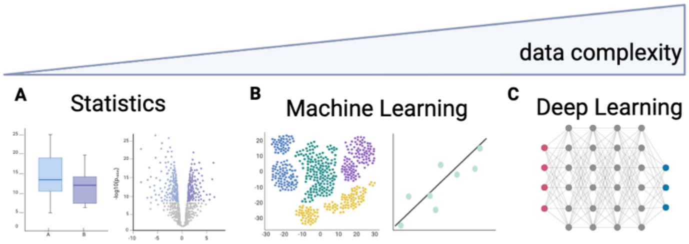

Figure 7: Data complexity augmentation requires advanced mathematical models.

(A) Simple statistical procedures such as hypothesis testing, provide important tools to investigate biomarkers in cancer pharmacogenomics. For example, it enables to stratify between sensitive and non-responders to a specific drug treatment. **(B)** Furthermore, machine learning methods, namely clustering, prediction and classification, can be leveraged for more complex tasks. Specifically, to perform drug response prediction based on multi-omics data and clustering of unprocessed data. **(C)** Notwithstanding, when handling significantly large and complex datasets, it becomes challenging to identify meaningful features and extract conclusions with standard approaches. Therefore, deep learning frameworks based on conventional neural networks are typically employed.
Figure created with BioRender.com.

In my thesis, I consider several computational concepts to integrate multi-omics data, investigate unexpected phenomena in drug response assays and prioritise the feature space in order to enhance ML predictive power. Ultimately, these approaches were created to identify new relevant biomarkers that can aid in the diagnosis and treatment of cancer patients.

1.4.1 Statistical methods for biomarker discovery

During the investigation of new cancer biomarkers, one requires statistical approaches to tackle the high-dimensional data for several purposes. Namely, to

- identify significant biomarkers;
- stratify between sensitive and non-responder cohorts;
- explore out-of-the-distribution observations, i.e., outliers;
- ensure non-biased conclusions.

One of the pillars of pharmacogenomics is the exploration of treatment efficacy and the variants involved with it. For this, typically one starts to identify distinct cell or patient cohorts and test their reaction to several treatments. Notably, these groups contain specific genetic variation, and therefore are defined by particular oncogenic alteration events, reflected on the oncogenes' status. In order to understand the relationship between these oncogenes and the compounds, it is conventional to leverage statistical tests such as t-test (STUDENT, 1908) or analysis of variance (ANOVA) (Girden, 1992). For instance, when exploring drug sensitivity, one stratifies between responding and non-responding cell lines. If this analysis is performed within a cancer type framework, a t-test is an adequate tool. However, within a pan-cancer analysis setting with several tissue types, an additional correction of the tissue type is required and therefore it is used ANOVA (Iorio et al., 2016). In my thesis I leverage ANOVA for two purposes: to identify drug sensitivity biomarkers across several cancer types, and to recognize increasing cell viability biomarkers within mutant subpopulations.

Another important factor while performing bioinformatic analysis, is the identification and validation of putative markers through hypothesis testing. Typical cases occur when performing differential expression analysis (Anders & Huber, 2010), testing the several combinations of drug-cell with a particular condition (e.g., sensitivity or resistance marker) (Iorio et al., 2016) or while performing gene set enrichment analysis (GSEA) (Subramanian et al., 2005). In these situations, a p-value will be assigned to each gene, drug-cell response, group of genes or pathway, and inform (under a specific threshold) of the probability of this particular event to happen if the null hypothesis is valid. Following this, it is possible to sort and recognize phenomena based on their significance (Panagiotakos, 2008).

However, frequently in this sort of analyses, thousands of hypotheses are tested, rendering the possibility of encountering false positive situations inevitable, and the assigned p-values misleading. Therefore, when performing statistical hypothesis testing, the p-values should be reported together with additional metrics independent of the sample size, such as confidence intervals or effect size (Halsey, 2019). For instance, while performing ANOVA analysis to identify sensitive drug response biomarkers, the p-values are complemented with Cohen's *D* effect size (Lakens, 2013).

Furthermore, it is recommended to correct the statistical tests taking into account the false positive rate inflation (Type I error). One of the most traditional methods is the conservative Bonferroni correction method (Bonferroni, 1936), where the raw p-values are divided by the number of total tests. An alternate classical technique is the Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995), which controls the false discovery rate (FDR) by ranking the p-values in an ascending order and multiplying them by the number of features over their corresponding rank.

In addition to the mentioned procedures, another group of FDR controlling techniques is the adjustment across multiple hypotheses testing families which are hierarchically structured, i.e. hierarchical false discovery rate (HFDR) controlling procedures. Several methods are possible to employ in this kind of hierarchical analysis, e.g. Yekutieli's (Yekutieli, 2008) and Benjamini-Yekutieli's (Benjamini & Yekutieli, 2001) methods. However, their application is dependent on several characteristics of the hierarchical framework, including the independence level within statistical families and which FDR correction procedure is implemented at each statistical level. In this thesis, I apply the Yekutieli HFDR control procedure in a two-level statistical framework where initially sensitive response biomarkers are identified, followed by the discovery of resistance biomarkers within the recognized sensitive ones.

Apart from these multiple testing correction procedures, re-sampling techniques such as bootstrapping or permutation methods are also commonly leveraged (Camargo et al., 2008).

Lastly, the exploration of novelties and outliers is of notable interest in bioinformatic research in the context of semi-supervised and unsupervised anomaly detection frameworks, respectively. The distinction between these two terms depends if the out of the distribution observation in question is something we want to investigate (novelty) or if its presence of damages in the raw data and/or subsequent analyses (outlier) (Hodge & Austin, 2004). Throughout my thesis I present two distinct examples of anomaly detection - the investigation of unexpected cell lines with resistance markers (novelty) and the identification of out of the distribution points whilst performing curve fitting of drug responses (outliers).

1.4.2 Dose-response curve fit

In cancer research, cell viability is typically assessed via HTS assays of several cell lines tested against multiple compounds over various concentration points. The resulting cell line drug response relationship is derived via a dose response curve. In the case of a full responder cell line, this curve typically resembles a sigmoidal curve with decreasing viability as the dosage increases (Figure 8 A). Furthermore, for the sake of simplicity, this information is commonly summarised by curve metrics such as IC_{50}'s or AUC values (Barretina et al., 2012; Garnett et al., 2012; Iorio et al., 2016; Seashore-Ludlow et al., 2015).

The modelling of drug responses can be performed using both parametric or non-parametric models and is typically conducted using nonlinear models, including the Non-linear Least Squares (NLS) algorithm (Becsey et al., 1968). Nonetheless, the modelling of these experimental designs is challenging due to the sparse sample size, small number of experimental replicates, existence of noise and nonlinear response behaviour. Therefore, most current curve fitting models are assumed to be parametric sigmoidal models with specific upper and lower boundaries (e.g., the cell viability is assumed to be reduced to zero for high drug concentrations) (Dawson et al., 2012; Iorio et al., 2016; Vis et al., 2016; Y. Wang et al., 2010).

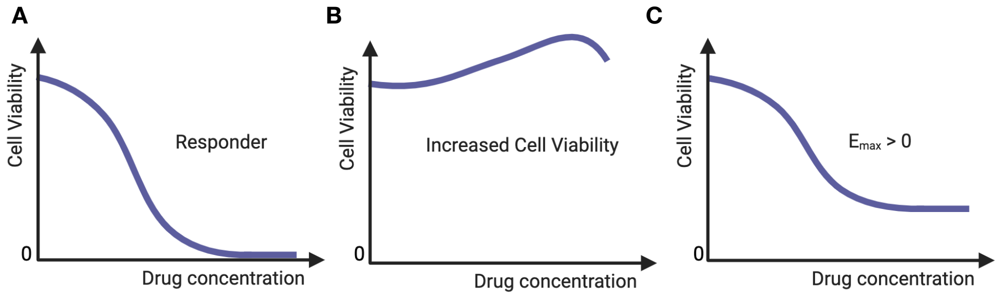

Figure 8: Types of drug responses in HTS.

(A) A cell line is called responder if its cell viability reduces as the drug concentration increases. Typically, the drug response is modelled as a sigmoid curve that equals to zero at the maximal drug concentration. Currently modelling of drug screen data neglects atypical behaviours, such as **(B)** increased cell viability and **(C)** E_{max} higher than zero.
Figure created with BioRender.com.

However, these restrictions often result in approximations where intrinsic noise of the experimental data and the true response of the curve fitting are ignored, leading to imprecise drug response estimations. In particular, several distinct curve patterns are currently unidentified or misclassified. Namely,

- *Increasing cell viability (ICV)*, i.e., cases where the cell lines exhibit an unexpected increase in the cell viability upon treatment (Figure 8 B). This phenomenon is counterintuitive in HTS, since one would expect a decrease in cell viability as drug dosage increases, eventually leading to cell death. In contrast, within an ICV scenario, a possible hypothesis is that the cancer drug facilitates cell growth by inhibiting cell cycle checkpoints, consequently accelerating the cell cycle. Such curve patterns are misclassified as non-responders by currently employed curve fitting methods.
- *Maximal effect at high concentrations (E_{max}) above zero* (Figure 8 C). Theoretically, the E_{max} value is modelled to converge to zero since high drug concentrations lead to cytotoxicity, and henceforth no cell viability. However, cancer cells may develop an intrinsic resistance to the compound driven by alterations in the gene expression, or other omics. Similarly to the previous situation, this type of responses are incorrectly modelled and wrongly classified as sensitive responses in HTS.

Both of these phenotypes hold significant potential to investigate in pharmacogenomics. For instance, HTS are typically tested over a short timeframe (3-6 days), rendering it difficult for a cell line to acquire resistance to a particular compound during the assay period. Hence, the possibility of an E_{max} to converge to a value greater than zero, likely elicits the existence of a resistance biomarker between the cell and the drug.

Furthermore, the case of ICV may conceal an extreme case of drug resistance or a prospective therapeutic scenario. Notably, uncontrolled tumour growth is a cancer hallmark and an event that should ideally be prevented from arising or actively combated. However, the existence of this uncontrolled proliferation between a drug and a cell line may pave the way for a novel and controversial (but effective) cancer therapy through synergistic drug combinations. For example, with drugs such as DNA damaging agents which favourably work with fast dividing tumours, and possibly reduce the tumour growth at a faster rate than its increase due to an efficient drug targeting scheme (O’Connor, 2015; Swift & Golsteyn, 2014).

In my thesis, I explore biomarkers of ICV in drug assays by developing a computational framework without parameterization of the cell viability upper and lower limits. This flexible approach allows to systematically investigate potential drugs which efficiently work with cell lines with fast proliferation based on their growth rates and hypothesise potent combination candidates of these two within cancer types.

In addition, a possible alternative to the commonly considered drug response fitting mechanisms based on NLS modelling of a sigmoid function, is to utilise Gaussian Processes (GP) which enable to measure and account for the uncertainty in the drug response data (D. Wang et al., 2020). This is of particular interest for HTS where there is a considerable amount of noise in the data and few replicates. Although this approach may entail higher complexity due to the need to specify a particular kernel, GPs can deal better with noisy HTS data and provide a more accurate assessment of the drug response (D. Wang et al., 2020). For this reason, in my thesis I propose the exploration of new drug response biomarkers through the usage of GP curve fit modelling with several kernels and report their overall accuracy in HTS screens typically assessed with a NLS sigmoidal framework (Iorio et al., 2016; Vis et al., 2016).

1.4.3 Networks for gene identification and prioritisation

An obstacle in cancer research is to efficiently tackle the vast amount of biological data and its quality in order to investigate underlying cellular processes.

A possible solution is to leverage biological networks, namely protein-protein interaction (PPI) networks, to stratify biological functional relationships between genes or proteins, infer relevant features to oncogenic changes and/or to perform knowledge guided investigations. For instance, these can be leveraged to identify drug response biomarkers (Garcia-Alonso et al., 2018; Kong et al., 2020), classify tumour subtypes (J. Cao et al., 2021; J. Su et al., 2010), predict patient outcome (Nuncia-Cantarero et al., 2018; Taylor et al., 2009) or as machine learning input tools (Costello et al., 2014; Kong et al., 2020; Schulte-Sasse et al., 2021). Several biological network databases with curated ensembles of several PPIs from literature and high-throughput experiments are currently available online, including STRING (Szklarczyk et al., 2018) and Omnipath (Türei et al., 2021).

Nonetheless, pairwise network tools such as PPIs potentially contain high false positive (i.e., incorrect interactions) and false negatives (i.e., missing interactions) rates, a significant amount of data to analyse and several non-relevant interactions (Liu et al., 2009; von Mering et al., 2002). Therefore, many network-based methodologies with functional inference have been proposed to create meaningful data analysis tools. In detail, these typically integrate several heterogeneous sources, such as physical binding, genetic and phenotypic interactions. This data integration can be conducted by combining the edges of several networks via mathematical inference (I. Lee et al., 2011; Peterson et al., 2015) or through edge weighting (Liu et al., 2009; Mostafavi et al., 2008; Z. Zhang et al., 2018). Subsequently, the built integrated network can be investigated by network-based inference methods, including graph-based clustering (J. Wang et al., 2019; H. Zhou et al., 2017) and label propagation (e.g., random walk with restart algorithm) (Lei et al., 2019; Mostafavi et al., 2008; Valdeolivas et al., 2019), to derive functional insights.

Several sophisticated methods focus on the integration of external functional annotations (Ashburner et al., 2000; Köhler et al., 2021) and topological network structure for the identification of relevant candidate genes (Ietswaart et al., 2021; Kumar et al., 2018; Mostafavi et al., 2008). In addition, through computational analyses, these studies can also be extended and customised to perform candidate's prioritisation via statistical inference (e.g., similarity probability or permutation-based with p-value ranking) for a specific task (Ietswaart et al., 2021; Kumar et al., 2018; Morrison et al., 2005). For instance, instead of leveraging all the unspecific GO term ontology annotations, one can use expert knowledge to guide the method and only focus on disease-relevant ontologies (e.g., annotations tailored to a particular disease or tissue). Notable graph-based methods of omics prioritisation include GeneFriends (Raina et al., 2023), pBRIT (Kumar et al., 2018) and PhenoRank (Cornish et al., 2018).

In this thesis I present a novel approach called weighted Protein-Protein Interaction (wPPI) (Galhoz et al., 2021) which integrates PPI data from Omnipath, and ontologies databases GO and HPO to identify and prioritise new genes based on a given genes of interest. The proposed framework is customised by tissue-type through a personalised filtering of the ontology databases and is ultimately used as a tool to systematically augment the set of features used in a posterior machine learning (ML) drug response model.

Despite the success of network-based approaches, current methods still face application challenges due to inconsistencies across data modalities (e.g., tissue-specific modules, different omics identifiers and levels of specification), intrinsic noise and structural disorder of PPI interactions, existence of directionality and hierarchy in both PPI networks and ontologies, as well as computational challenges such as data storage (Milano et al., 2022; Sevimoglu & Arga, 2014). Fortunately, these issues are starting to be addressed in emerging knowledge representation methodologies. For instance, the recently published BioCypher framework offers the integration of several biomedical resources harmonized by user-specific ontologies and fast query and extraction of knowledge graphs (Lobentanzer et al., 2023).

1.4.4 Machine learning for drug response prediction

Transitioning from knowledge guided inference of relevant features through statistical and network-based methods, here it is discussed another pillar of computational pharmacogenomics: the prediction of drug response using machine learning (ML) methods.

In particular, ML algorithms can be split in unsupervised and supervised learning, depending if the prediction is performed on unlabelled or labelled data, respectively (Vamathevan et al., 2019). The former is a data-driven approach which infers hidden or intrinsic patterns in the data by assessing similarity between the samples. These are commonly used for exploratory purposes, namely to stratify groups in the raw data, as a dimensionality reduction technique or for high-dimensional data visualisation. Typical examples of this modality are clustering algorithms, including k-means (Hartigan & Wong, 1979) and hierarchical clustering (Bridges, 1966), and segmentation techniques, such as Principal Component Analysis (PCA) (Pearson, 1901) and t-distributed stochastic neighbour embedding (t-SNE) (Gmail & Hinton, 2008).

In contrast, supervised ML methods aim to develop a mapping function between the known input labels and output values through training models. Subsequently, the built learning algorithms can be applied to predict future outputs on an unseen set of input data, such as an external independent dataset or part of the original input data which was not involved in model training. In addition, in order to achieve more accurate and robust predictions, one can also leverage prior knowledge (biological priors) or estimated subgroups from an unsupervised algorithm, as input to supervised approaches. For example, established clustering groups from gene expression profiles can be used to predict drug response (Majumdar et al., 2021).

Subject to the intended application goal, supervised learning methods can be partitioned into classification or regression tasks depending if the outcomes are discrete or continuous variables, respectively. Examples of a classification task are the stratification of cancer type classes from high-throughput data (Tyanova et al., 2016) or the categorization of cell lines as sensitive or resistant (Ahmadi Moughari & Eslahchi, 2021; Iorio et al., 2016). On the other hand, the prediction of the drug response of cancer cell lines is performed using regression methods (Costello et al., 2014; Menden et al., 2013).

Within each supervised learning technique, a broad spectrum of supervised linear and non-linear ML algorithms are available, including Support Vector Machine (SVM; C. Cortes & Vapnik, 1995), Ensemble (Z.-H. Zhou, 2011) and Kernel (Hofmann et al., 2008) methods for classification. Alternatively, for regression, Neural Networks (W. S. McCulloch & Pitts, 1943), Linear Regression (e.g., LASSO, Ridge and Elastic Net regularisation) (Bingham & Fry, 2010) and Decision Trees (Wu et al., 2008).

During model learning, a relevant aspect to have in consideration is model robustness in terms of over- and underfitting. These phenomena significantly affect the performance of ML approaches, and can be detected during model training or when assessing the accuracy of the trained model in a new independent test data. Particularly, overfitting is disclosed by a high accuracy in the training data, but fails to infer accurate predictions on the test data. Conversely, during underfitting the model is too simplistic and does not accurately predict the training data (Camacho et al., 2018).

To address undesirable fitting situations, adjustment of model's complexity, data augmentation, feature engineering and/or regularisation techniques (e.g., LASSO, Ridge or Elastic Net) should be employed. Several notable resampling methodologies are commonly leveraged, including cross-validation (CV) (Hastie et al., 2009; Picard & Cook, 1984) and bootstrapping (Mooney et al., 1993). For the former, the amount of k folds considered depends on the user and data structure, however a 10-fold CV or a leave-one-out cross-validation (LOOCV) are the most widely known structures. In addition, a repeated CV procedure can be leveraged for a more robust and accurate estimation of the model performance (J.-H. Kim, 2009).

Moreover, in order to assess the validity of predictions, model performance can be inferred by evaluation metrics depending if the predictive model is a classification or regression task. In the case of a classification ML model, metrics such as the area under the curve (AUC), accuracy and confusion matrix are widely considered (Lever, 2016). On the other hand, the performance of regression methods can be evaluated by variance-based metrics, including the Pearson Correlation Coefficient (Lee Rodgers & Nicewander, 1988) and root-mean-squared error (RMSE) (Hastie et al., 2009).

Throughout this thesis I focus on the application of linear models, namely ANOVA for the distinction between sensitive and resistant cell lines, and LASSO regression for the prediction of drug response. Although these are simplistic approaches, the usage of these linear methodologies is motivated by the balance between model simplicity and accuracy. These are expressed by the straightforward fitting and model's outcome interpretation, the possibility of taking into account various components (e.g., clinical and genetic features) and the consistent good results in literature for the designated tasks (Geeleher et al., 2014; E. W. Huang et al., 2020; Iorio et al., 2016). In addition, I propose to employ the LASSO model in combination with the previously mentioned wPPI network approach (Galhoz et al., 2021) for an appropriate knowledge-guided feature selection and, consequently, derive an informative final model. In order to ensure model robustness, the constructed predictive model will be trained and tuned using the repeated cross-validation technique, performance will be assessed by the Pearson correlation coefficient and additionally benchmarked against other state-of-the-art network-based feature selection procedures.

1.5 Aims of the thesis

Throughout the recent years, there has been a notable expansion of pharmacological screens and in the development of innovative approaches to analyse unexpected behaviours within these screens. These advancements grant unique opportunities to pave new pathways in precision medicine, by improving the current understanding of biomarkers in cancer research, as well as develop new therapy mechanisms.

Despite the continuous development, analyses built on drug screens are complex and statistically underpowered due to poor molecular characterisation and/or infrequent alteration events across cancer types. Moreover, standardised frameworks used to describe drug responses are rather simplistic, with parameterisation constraints which neglect unexpected phenotypes of potential interest in therapeutic settings.

In this thesis I focus on the development of new mathematical methodologies that tackle the low statistical power in HTS and contribute to the identification of novel cancer biomarkers through three different aims:

1. **Identify drug resistance in HTS –** Indirectly explore unpredicted resistance behaviours in HTS through an advanced and innovative statistical framework for outlier detection in sensitive cell lines. The proposed procedure outperforms existing state-of-the-art outlier identification methods, successfully recognizes established resistance markers in different cancer types and reveals several new biomarker hypotheses to be leveraged for drug combinations;
2. **Exploit increased cell viability induced by drugs for cancer treatment –** Creation of a sophisticated methodology to investigate the unexpected increasing cell viability phenotype, missed by standardised models. The proposed framework leverages both linear and non-linear flexible curve fitting strategies, and is complemented with a detailed exploration of the molecular characterisation, to recognize candidates for validation. In addition, I analyse cell proliferation rates to hypothesize potential drug combination with synergistic effects involving the validated candidates;
3. **Unveil novel cancer-specific genes via functional networks –** Design of a comprehensive network-based approach integrating signalling pathways, genomic and phenotypic data to recognize and prioritise new cancer-specific genes. The proposed framework is leveraged as a feature augmentation tool to enhance drug response prediction using gene expression profiles. The built ML models are benchmarked against another knowledge-based feature selection procedure and the identified cancer biomarkers are analysed in therapeutic contexts.

In the upcoming section I provide an overview of the publicly available datasets I used throughout my thesis, and detailed mathematical descriptions of the methods developed and leveraged for the investigation of the previously enumerated approaches. Subsequently, in chapter 3 I illustrate the relevant applications and outcomes of the applied methodologies. Lastly, in chapter 4, I summarise the main results of the three applications, discuss the relevance of the findings, debate current limitations and propose an outlook for possible future research.

2. Materials and Methods

2.1 Public resources

In this section I provide an overview of the public data resources, data processing tools and methods necessary to perform the bioinformatic analysis presented in this thesis.

Throughout this thesis I illustrate results for three distinct projects in pharmacogenomics (overview of the projects can be found in the **List of Publications** chapter). Therefore, several of the public materials are shared but were downloaded and used at different time points depending on the project at hand. For these situations, the relevant dates of access and indication of data versions used for each project are indicated.

2.1.1 GDSC and CTRP pharmacological screens

Raw pharmacology data from the GDSC (Garnett et al., 2012; Iorio et al., 2016) and CTRP (Basu et al., 2013; Seashore-Ludlow et al., 2015) projects are leveraged throughout this thesis. Two versions of the former are considered – one containing 495 compounds, 818 cell lines and 317,357 drug-cell combinations (Project 1, downloaded in 2019) and another consisting of 516 different compounds, 998 cell lines and 420,273 drug-cell combinations (Projects 2 and 3, downloaded in 2021; Figure 9).

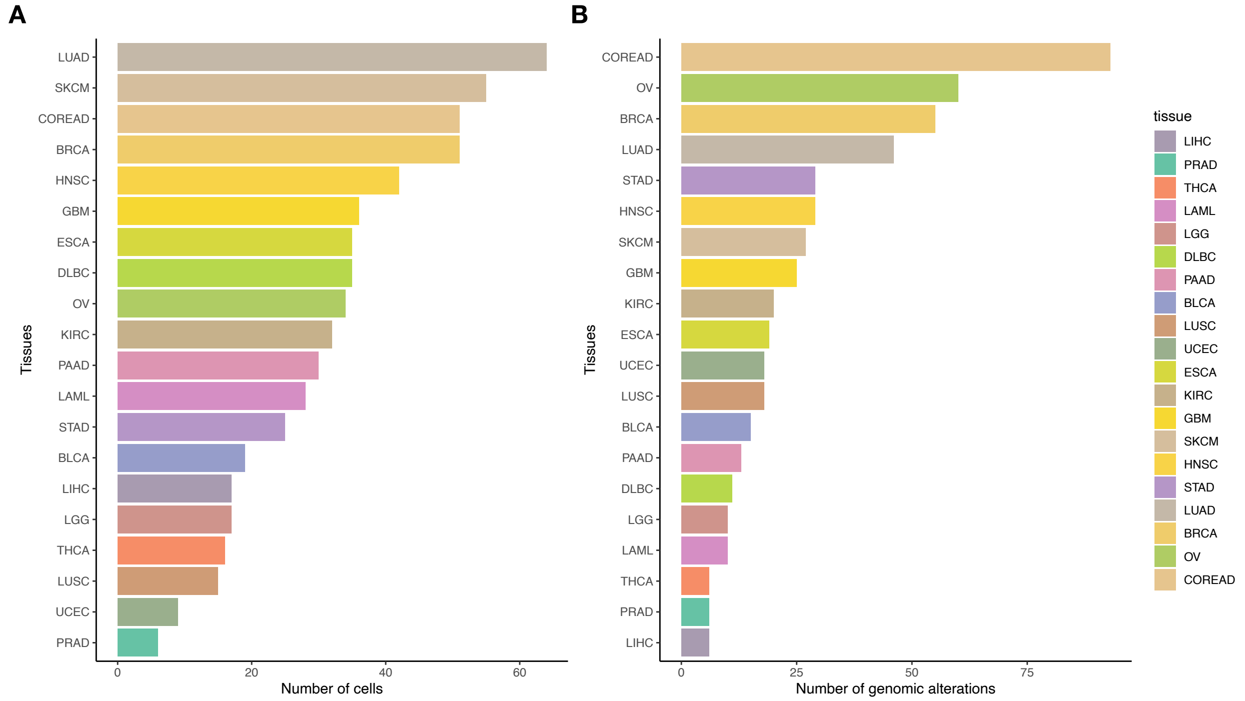

Figure 9: Overview of GDSC number of cells and functional events by cancer tissue.

Number of **(A)** cells and **(B)** genomic alterations per cancer type, for Projects 2 and 3. PANCAN tissue was not considered. Barplots coloured according to the cancer type.

For Project 1, only CTRP screening data (downloaded in 2019) with cell lines available in the GDSC dataset was considered, comprising 545 compounds, 504 cell lines and 220,461 drug-cell combinations. In total, the combination of GDSC and CTRP datasets consisted of 814 unique compounds and 816 unique cell lines to be investigated in Project 1.

In addition, drug response data was estimated for both pharmacological screens by the sigmoidal-based curve fitting method used in the GDSC project (Vis et al., 2016) and represented by the concentration required to reduce the cell viability by half (IC_{50}) metric. Moreover, the IC_{50} values were described in the natural logarithm form of the μM concentrations, where low values of IC_{50} state a higher sensitivity of the cell line to the tested compound and vice versa.

2.1.2 Experimental Design

For the GDSC and CTRP projects, the raw data was generated using a cell viability (CV) assay where cell lines were seeded in 384-well plates and treated with several compounds for 72 hours. In the GDSC project, for each drug, the cell lines were tested with 5 or 9 distinct concentrations in 4- or 2-fold dilution points, respectively. Whilst, in the CTRP project, the tested compounds were only plated into 2-fold concentration ranges. Afterwards, cells were stained using CellTiter-Glo and their fluorescence signal intensity *(I)* was assessed (Basu et al., 2013; Iorio et al., 2016).

Each cell assay consisted of three different types of wells: cells treated with a given drug at a specific concentration, cells treated with a control substance (typically DMSO) and blank wells (plate with no cells used for background correction) (Figure 5). Based on these measures, the cell viability of each drug is computed as:

$$CV_{drug} = \frac{I_{drug} - I_{blank}}{I_{control} - I_{blank}}$$

Equation 1

Where I_{drug} is the intensity of the well treated with the compound, I_{blank} and $I_{control}$ are the average intensities across the wells on that particular plate with no cells and treated with DMSO, respectively.

2.1.3 Genetic data

Cancer functional events (CFEs) of the screened cell lines were retrieved from the GDSC project (Iorio et al., 2016). These included copy number variations from Affymetrix Genome-Wide Human SNP Array 6.0, and mutation status of whole genome sequencing from Affymetrix GeneChip Human Genome HT-HGY122A Array. CFEs were described as binary events for each cell line across 21 different cancer types and 1,073 different mutations. In addition, gene expression profiles were detected for 1,018 cell lines across 17,418 genes.

2.1.4 Omnipath prior knowledge network

The Omnipath database was used to derive direct and indirect human Protein-Protein Interaction (PPI) networks around cancer genes of interest. The Omnipath combines over 100 resources of curated pathways with information of inter- and intracellular signalling interactions, including a collection of 61 protein network databases (Türei et al., 2021). For this thesis, the Omnipath network was downloaded using the R package OmnipathR (Valdeolivas et al., 2019), and consisted of a total of 46,285 protein interactions, with 4,663 and 4,654 unique protein and gene symbols, respectively (downloaded in 2022).

2.1.5 Ontology Databases

Gene Ontology (GO) (Ashburner et al., 2000) and Human Phenotype Ontology (HPO) (Köhler et al., 2021) resources were leveraged to build functional weighted PPI networks. The usage of these ontology databases in a bioinformatics' framework enables the addition of genomic and phenotypic knowledge to existing interacting PPI networks and inference of disease-specific interactions and genes. The GO and HPO ontologies were directly downloaded from the respective portals using the wPPI R package (Galhoz et al., 2021). The GO database encoded 606,840 gene-ontology combinations, with 18,346 unique GO IDs and 19,712 gene symbols; and the HPO database comprised 209,825 gene-ontology interactions, with 8,919 unique HPO IDs and 4,769 gene symbols (both ontology databases were accessed in 2022).

2.1.6 Cancer specific genes from IntOGen

The Integrative OncoGenomics (IntOGen) platform (Martínez-Jiménez et al., 2020) was used to derive cancer-specific driver genes. This bioinformatics tool includes a collection of seven computational driver identification methods, namely OncodriveFML (Mularoni et al., 2016) and oncodriveCLUSTL (Arnedo-Pac et al., 2019), and combines their outcome through a weighted scoring procedure to identify cancer specific drivers. The leveraged IntOGen dataset comprises 568 mutational cancer drivers (release of 02.02.2020), inferred from sequenced patient tumour samples from several sequencing projects such as ICGC (J. Zhang et al., 2019), TCGA (K. Tomczak et al., 2015) and PCAWG (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), with a total of 28,076 samples across 66 different cancer types.

2.2 Indirect drug resistance in pharmacological screens

In this chapter I outline a new statistical framework to investigate unexpected resistance biomarkers in publicly available HTS drug screens.

2.2.1 ANOVA models

The assessment of sensitivity in a pharmacological screen is a relative inference based on the domain of screened cell lines available. Notably, a cell line is sensitive to a compound if its drug response is below the maximum concentration tested for that drug, and significantly lower than the drug response seen in the remaining cell lines in the screen. Henceforth, cases where all cell lines respond to a compound do not represent a cancer biomarker but rather a high level of toxicity to the drug.

The stratification between sensitive and non-responding (i.e., *resistant*) cell lines is a classification problem, and is commonly solved using an analysis of variance (ANOVA) statistical framework across all cancer types available in the pharmacological screen (Garnett et al., 2012; Iorio et al., 2016).

Iñigo Ayestaran constructed ANOVA models based on the drug response (Y_i) and mutation status (M_{status}) of the CFEs across N cell lines ($Y_i = (y_{i1}, \dots, y_{iN})$). It is assumed the drug response of the tested cell lines are independent variables, with a normal distribution ($Y_i \sim N(\mu_i, {\sigma_i}^2)$). Furthermore, several factors such as tissue type (*tissue*), screening medium (*medium*), microsatellite instability status (*MSI*) and growth properties (*growth*) have a significant effect in the drug response (B.-J. Chen et al., 2015), and therefore were considered as covariates in the model.

Therefore, the following ANOVA model for each combination of drug, CFE and cancer type was considered:

$$Y_i \sim C(tissue) + C(medium) + C(MSI) + C(growth) + M_{status},$$

where Y_i is a vector representing the drug response IC$_{50}$ values of drug i across N cell lines, M_{status} is the CFE status binary coded as mutant (=1) or wild-type (=0) and $C(.)$ are cancer specific covariates.

To ensure statistical power and increase biological interpretability, CFEs without known driver genes and with less than four mutant cell lines were neglected.

In order to account for the effect of sample size (i.e., amount of tested cell lines with a specific CFE status), the statistical inference for each drug-gene combination was complemented with the signed Cohen's *D* statistic (Lakens, 2013). This effect size is given by the mean difference of two groups (one having a specific CFE feature and the other not) divided by the pooled standard deviation.

Ultimately, drug sensitivity biomarkers were selected based on an unadjusted p-value threshold of < 0,001 and Cohen's effect size < -1.

2.2.2 Outlier detection methods

- **Novel framework based on standard deviation**

The identification of sensitive markers in pharmacological screens is a well-recognized concept in cancer research. However, the exploration of resistant biomarkers remains a challenge since these are rare events and it is difficult to stratify between cytotoxicity and resistance.

In order to circumvent these challenges, in our work (Ayestaran et al., 2020), Iñigo Ayestaran proposed a new statistical framework to investigate, within the pool of sensitive cell lines of a specific cancer type and with a particular CFE (chapter 2.2.2), subpopulations of cell lines which present UNexpectedly RESistant (UNRES) markers.

Briefly, the mathematical workflow explores the standard deviation of the sensitive population's drug response and systematically identifies cell lines that significantly influence the upper bound of its distribution (i.e., the most likely resistant cell lines).

In detail, each subgroup of M sensitive cell lines ($Y_j, j = 1, \dots, M$) are firstly organized based on their IC50 values

$$y(IC_{50})_1 < \cdots < y(IC_{50})_M.$$

Next, the cell lines with the highest IC50 values are systematically removed from the collection of sensitive cell lines and the standard deviation (SD) of the new distribution is estimated. This procedure continues for a maximum of 5 iterations ($k = 1, \dots, M - 5$) or until half of the sensitive cell lines sample are reached ($k = 1, \dots, \frac{M}{2}$).

Finally, the standard deviation delta between the new collection of sensitive cell lines without the k most resistant cell lines (SD_k) and the original population (SD_0) is assessed

$$\Delta SD_k = SD_k - SD_0.$$

In order to identify UNRES cell lines, Iñigo employed a bootstrap resampling strategy where the $M - k$ drug responses values were permuted and randomly selected. It was considered $B = 10{,}000$ bootstrapping iterations to ensure a robust estimation, and computed bootstrap SD deltas for each loop $\Delta SD_{k(boot)} = SD_{k(boot)} - SD_0$. Subsequently, the generated bootstrap values were significantly tested under the null distribution by assessing the proportion of $\Delta SD_{k(boot)}$ with values higher or equal than the observed delta (ΔSD_k):

$$p^* = \frac{1}{B} \sum_{b=1}^{B} I\left(\Delta SD_{k(boot)} \geq \Delta SD_k\right),$$

where p^* indicates the statistics' p-value and $I(.)$ is an indicator function which is equal to 1 when the tested hypothesis is true, and 0 otherwise.

To ensure a reliable and significant identification of UNRES cell lines, the bootstrap p-values were adjusted for multiple hypothesis testing via the Benjamini-Hochberg FDR procedure with $\alpha = 15\%$.

Ultimately, due to the lack of statistical power, the candidate UNRES cell lines were defined as truly resistant cell lines if and only if the associated CFE was mutually exclusive (i.e., CFEs which were mutated in all UNRES cell lines and wild-type in the sensitive cell lines, and vice-versa) and if the CFE was enriched in the UNRES subpopulation (i.e., Fisher's enrichment test between sensitive and UNRES cell lines).

Moreover, to assess the quantitative effect of the UNRES cell lines in the SD distribution, the analysis was enhanced with an inference of the normalized variation between the UNRES cases and the remaining sensitive population:

$$SD_{normalized} = -\frac{\Delta SD_k - E[\Delta SD_k]}{SD_0 - E[\Delta SD_k]},$$

where the expected delta value $E[\Delta SD_k]$ is given by the median of the bootstrap distribution of ΔSD_k.

Lastly, Iñigo Ayestaran additionally quantified the probability of selection of UNRES cell lines purely by chance through a permutation-based analysis. In detail, for each combination of drug and CFE, it was applied a model where the drug response values were randomly permuted 100 times, and subsequently applied our sensitivity and UNRES identification frameworks. At the end, the total amount of selected sensitive and UNRES cell lines, and an upper limit of UNRES cell lines for each drug-CFE association was reported.

- Neyman-Pearson

The previously proposed UNRES cell lines identification workflow is in essence an outlier detection strategy which leverages the distribution of the sensitive cell line population and significantly explores out of the distribution markers in the upper bound.

In order to evaluate the performance of our suggested framework, I benchmarked it against the state-of-the-art Neyman-Pearson (NP) outlier detection approach (Neyman & Pearson, 1933). Notably, this strategy is based on multiple hypothesis testing, therefore we formulated the identification of UNRES cell lines as:

$$H_0: Y(IC_{50})_0 < x(IC_{50})_{critical}$$

$$H_1: Y(IC_{50})_0 \geq x(IC_{50})_{critical},$$

where the drug response of the original sensitive population $Y_0 = (y_{01}, \dots, y_{0M})$ follows a normal distribution $Y_0 \sim N(\mu_0, {\sigma_0}^2)$), $x(IC_{50})_{critical}$ is the statistics' critical value from which we define the significance area of interest, and we intend to identify cell lines that reject the null hypothesis H_0.

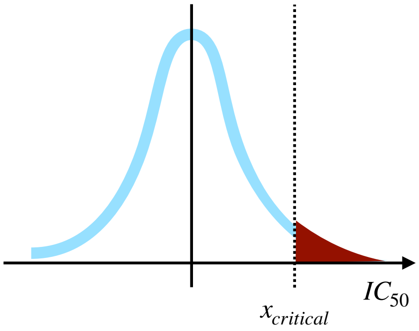

Figure 10: Neyman-Pearson outlier detection to identify resistance markers.

Given the drug response values of the sensitive population, it is expected to recognize cell lines with potential resistant markers as the ones with highest IC$_{50}$ values. Therefore, assuming a normal probability model and under the significance threshold $\alpha = 15\%$, the critical value $x_{critical}$ and rejection region (in red) are defined. Cell lines which fall in the rejection region are identified as resistant cell lines candidates.

To ensure similar conditions between the methods, I considered in the Neyman-Pearson framework the same significance level $\alpha = 15\%$ as in our proposed framework.

Ultimately, the UNRES cell lines detected by NP for each drug-CFE combination were reported, compared to our framework and investigated in regard to existing gold standards.

2.2.3 Hierarchical false discovery rate control

In the proposed pipeline, two statistical tests are performed – first, the selection of sensitive cell lines (chapter 2.2.1), followed by the identification of UNRES cell lines (chapter 2.2.2). Notably, these tests are hierarchically associated, since the exploration of UNRES cell lines occurs within populations of sensitive cell lines with a specific combination of drug and CFE.

Therefore, the tested multiple hypotheses are not independent, and the identification of significant markers is linked to a general control of the FDR that takes into consideration the hierarchical dependency within the framework.

Considering the hierarchical structure of our statistical pipeline and the possibility of arranging it based on multiple families of drug-CFE associations, the Hierarchical False Discovery Rate (HFDR) controlling procedure developed by Yekutieli (Yekutieli, 2008) was a suitable methodology to leverage and estimate a universal bound for the FDR of the overall procedure.

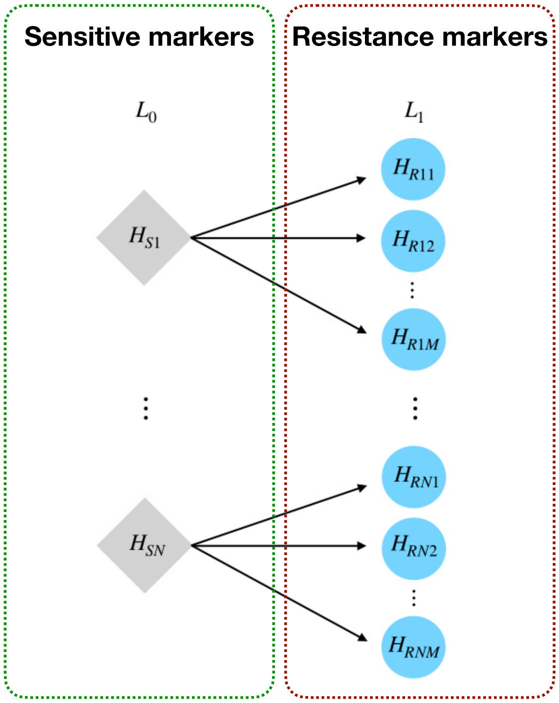

Figure 11: Hierarchical False Discovery Rate (HFDR) structure to test sensitive and resistance biomarkers in drug screens.

The procedure is arranged in a two-level hierarchical tree of statistical hypothesis. In the root level $\mathcal{L}_0$, each hypothesis $H_{St}, t = 1, \dots, N$ tests if the cell line contains a sensitive biomarker. For each rejected parent hypothesis H_{St}, the associated children nodes ($H_{Rtm}, m = 1, \dots, M$) are tested in level $\mathcal{L}_1$ in respect to resistance markers.
This figure was created by me and it is an adaptation of the original supplementary figure 5 “Hierarchical FDR procedure illustration” by (Ayestaran et al., 2020) under a CC-BY-4.0 license (http://creativecommons.org/licenses/by/4.0/).

First, I organized the statistical hypotheses in a 2-level tree of T disjoint drug-CFE families – where the first level tests if cell lines are sensitive biomarkers ($\mathcal{L}_0 = \{H_{S1}, \dots, H_{SN}\}$) and the second level if the cell lines of rejected parent hypotheses H_{St} present UNRES markers ($\mathcal{L}_1 = \{H_{Rt1}, \dots, H_{RtM};\ t = 1, \dots, N\}$). Notably, each hypothesis H_{Rtj} in level-1 is associated with a parent hypothesis H_{St}.

Moreover, several conditions are necessary to employ the HFDR procedure. In particular,

- hypotheses of each drug-CFE family are required to be tested simultaneously at each level;
- per definition, a set of hypotheses with the same parent node is called a family of hypotheses. Here, FDR is controlled within families, and the p-values are assumed to be independent across families;
- a family of hypotheses at level $\mathcal{L}_1$ is only tested if its parent hypothesis at level $\mathcal{L}_0$ was rejected, i.e., we investigate if a cell has UNRES markers only if it was previously identified as a sensitive cell line;
- at each level $\mathcal{L}$ the hypotheses are tested by the Benjamini-Hochberg (BH) procedure. However, it is not required to use the same α control rate across the tree levels.

Thoroughly, I formulated the HFDR procedure to our framework in the following way:

1. Test the root hypotheses $H_{St} = \{\text{sensitive cell line with drug-CFE } t\}$ at level-0 $\mathcal{L}_0$ under a significance rate $\alpha = 0{,}001$;
2. For each rejected parent hypothesis H_{St}, test child hypotheses $H_{Rtk} = \{\text{sensitive and UNRES cell line with drug-CFE } t\,;\ k = 1, \dots, M\}$ and employ the BH procedure to control the FDR at $\alpha = 0{,}15$ within each *t*-th family in $\mathcal{L}_1$.

The usage of different control rates α across the HFDR framework was motivated by different statistical requirements at each hierarchical level $\mathcal{L}$. Notably, a conservative threshold based on p-value and not on FDR at the first level $\mathcal{L}_0$ guarantees the identification of true sensitive biomarkers whilst enabling the possibility to subsequently investigate putative UNRES biomarkers. Moreover, a higher control rate at the final level ensures a low FDR in the identification of UNRES cell lines and of the overall HFDR procedure.

Finally, an estimation of the universal bound for the multiple family FDR was provided. I leveraged the following approximation deduced by Yekutieli (Yekutieli, 2008),

$$FDR \ \leq \alpha * \delta * \frac{\text{[no. discoveries + no. families]}}{\text{[no. discoveries + 1]}},$$

Equation 2

which depends on the number of significant markers in levels $\mathcal{L}_0$ and $\mathcal{L}_1$ (no. discoveries), number of drug-CFE combinations (no. families), control rate α and inflation rate δ (here assumed as $\delta = 1$, as theorized by Yekutieli) (Yekutieli, 2008).

2.3 Exploration of increasing cell viability in high-throughput screens

For this study I developed a new mathematical framework to explore unexpected drug response behaviours, in particular the phenomenon of increasing cell viability.

2.3.1 Principal component analysis

The dimensional reduction technique Principal Component Analysis (PCA) (Pearson, 1901) was used to visualise high-dimensional raw pharmacological data. This unsupervised tool enables to highlight intrinsic patterns in the drug response data prior to any post-processing analyses.

For each drug, cell viabilities (*Equation 1*) were computed based on drug, blank and control intensities for all available titration points $T_1, \ldots, T_N$ (where $N = 5$ or 9, depending on the drug in question). Following, PCA performs an orthogonal linear transformation and projects the N-dimensional cell viabilities into two principal components which maximize the feature's variance (i.e., thus minimizing the loss of statistical information). Notably, each principal component is defined by a linear combination of the features from the original data

$$PC_1 = a_1 CV_1 + \cdots + a_N CV_N$$

$$PC_2 = b_1 CV_1 + \cdots + b_N CV_N,$$

Where $Var(PC_1) > Var(PC_2)$, a_i and b_i are coefficients estimated by least squares optimization, and CV_i is the drug cell viability at the i-th titration point.

Ultimately, the application of PCA increases data interpretability by facilitating the visualisation of the drug response data in a 2D space, hence granting the possibility to easily investigate patterns in the data and identify clusters of drugs with similar response.

2.3.2 Curve fitting methods

The drug response of pharmacological high-throughput screens is typically modelled with the assumption of a sigmoidal behaviour, where the maximal cell viability is seen at the initial drug concentration, followed by a continuous decrease in viability as the drug concentration is increased.

As previously outlined (chapter 1.4.2), in HTS it is possible to encounter several other curve patterns. Therefore, an appropriate curve fitting that models all the possibilities is essential for an appropriate estimation of the drug response.

In this section, I outline the mathematical formulation of three different kinds of curve fit methodologies which were leveraged in this thesis.

- **Sigmoid curve fit**

A notable possibility to perform a sigmoid curve fitting of pharmacological data was proposed by the GDSC project (Vis et al., 2016) and it is implemented in the gdscIC50 R package (https://github.com/ CancerRxGene/gdscIC50). This tool was initially designed to fit data from GDSC, however it is also possible to leverage it on pharmacological data from other resources, such as data from CTRP, as we have performed in our work (Ayestaran et al., 2020).

As a way to tackle intrinsic noise in drug response data, this tool models under the assumption that the cell viability decreases with a sigmoidal shape within a scale from one to zero:

CV $= 1$ at initial concentration;

CV $\rightarrow 0$ at the maximum concentration.

Equation 3

Built on these simplifications, the authors described the sigmoid drug-response model $\boldsymbol{y}_{ij}$ as a function of the drug concentration $\boldsymbol{x}$, dependent on slope s_i and position p_{ij} parameters for cell line i and compound j:

$$\boldsymbol{y}(\boldsymbol{x}, s_i, p_{ij}) = \frac{1}{1 + e^{\frac{x - p_{ij}}{s_i}}},$$

with $s_i = \alpha + a_i$ and $p_{ij} = \beta + b_i + b_{ij}$

Equation 4

Notably, the shape parameter s_i only changes based on the cell line type, whilst the position p_{ij}varies across cell-drug combinations. Wherein, parameters a_i and b_i represent respectively the impact of slope and position in cell line i, and b_{ij} the joint effect of drug j and cell line i in the curve shape. In parallel, parameters α and β represent fixed population effects for the shape and position parameters, respectively.

The sigmoid formulation and its parameters in *Equation 4* were estimated using a Non-linear Mixed Effects model (Lindstrom & Bates, 1990), enabling the inference of cell line response coupled with nested random effects across all drugs. Moreover, the drug response IC_{50} value is given by p_{ij} and estimated by interpolation or extrapolation (in case the 50% viability is not achieved within the available dosage range).

- Linear curve fit

The previously described sigmoid formulation robustly models the most typical (and desirable) curve shape of drug response data. However, its intrinsic assumptions (*Equation 3*) prevent the identification of several drug response patterns.

For this reason, here I propose a relaxed and linear curve fitting approach where these assumptions are neglected. Due to its simplistic nature, a linear regression enables a fast, not computationally expensive and robust against overfitting modelling alternative.

Although a linear fit is not a perfect representation of the real curve behaviour, it is able to provide a rough and quick overall picture of the curve's growth. For these reasons, the linear formulation was leveraged in the pre-processing stage for the identification of noisy responses and putative outliers (see chapter 2.3.3), stratification between non-responders and responders, and as a basic estimator of responses with increasing cell viability (detailed in the chapter 2.3.5).

For each cell line-drug combination, the following linear formulation was considered:

$$\boldsymbol{y}(\boldsymbol{x}) = \beta_0 + \beta_1 \boldsymbol{x} + \epsilon$$

Equation 5

Where $\boldsymbol{x}$ is a vector with the drug concentration values, β_0 is the intercept at the initial drug concentration, the coefficient β_1 describes the curve slope and ϵ is a random disturbance term.

In contrast to the sigmoid curve fit, with the linear formulation the drug responses were summarised by the curve's slope β_1, instead of the IC_{50} value.

The linear fitting and slope information were computed using the lm function available in the stats R package (R Core Team, 2016).

- Gaussian curve fit

In addition to the linear formulation, I also considered a Gaussian-based curve fit methodology without any prior assumptions regarding the cell viability domain.

The usage of a Gaussian Process (GP) model provides a realistic and robust fitting of the drug response, where it is possible to include relevant parameter information (e.g., relations between variables and statistical distributions) and leverage several kernels (from linear to sigmoid-like curve shapes). Furthermore, GPs are known to perform interpolation with low uncertainty (Rasmussen & Williams, 2005; D. Wang et al., 2020).

Despite its numerous advantages, a Gaussian-based model is notable for its likelihood to overfit and is computationally more costly in comparison to the two previous formulations.

For the implementation of the Gaussian model, I leveraged the formulation available in the Kernlab R package (Karatzoglou et al., 2004) with its default parameters. This resource offers seven different kinds of kernel functions to test and considers the model's hyperparameters fixed (therefore saving some computational time in hyperparameter tuning).

In detail, the relationship f between the drug response $\boldsymbol{y}$ and the concentration points $\boldsymbol{x}$ is defined by the posterior distribution

$$p(\boldsymbol{y} = f(\boldsymbol{x})|x, \Psi) = N(f(\boldsymbol{x})|\boldsymbol{m}, K_{\Psi}(\boldsymbol{x}, \boldsymbol{x}')),$$

Equation 6

which follows a multivariate normal distribution, with mean drug response $\boldsymbol{m} = E[f(\boldsymbol{x})]$ and covariate function $K_{\Psi}(\boldsymbol{x}, \boldsymbol{x}'))$ dependent on hyperparameters Ψ at the points $\boldsymbol{x}$ and $\boldsymbol{x}'$.

The covariate function K_{Ψ} expresses how two points in the domain space statistically relate. Although this function can take several shapes, in this thesis I considered two kernel functions:

- Gaussian Radial Basis

$$K_{\Psi}(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma||\boldsymbol{x} - \boldsymbol{x}'||^2)$$

  Where hyperparameter σ denotes the variance. This kernel is the classical formulation and is typically applied for data without prior knowledge.

- Spline

$$K_{\Psi}(\boldsymbol{x}, \boldsymbol{x}') = 1 + \boldsymbol{x}\boldsymbol{x}'\big(1 + \min(\boldsymbol{x}, \boldsymbol{x}')\big) - \frac{\boldsymbol{x} + \boldsymbol{x}'}{2}\min(\boldsymbol{x}, \boldsymbol{x}')^2 + \frac{\min(\boldsymbol{x}, \boldsymbol{x}')^3}{3}$$

  Defined by a piecewise cubic polynomial, and with a sigmoid-like shape.

2.3.3 Noise and outlier removal

Pharmacological data contains a fair amount of intrinsic noise. This noisy behaviour may be conditioned by a large amount of meaningless data distribution (noisy responses), or by a small number of points in the drug-cell response (outlier points).

In order to tackle the corrupted data, I implemented the following mathematical approaches based on the original cell viabilities CV_i (*Equation 1*) and the linearly fitted drug response $\hat{\boldsymbol{y}}$ (*Equation 5*):

- **Noisy drug responses**

Noise in HTS experiments is common and unavoidable, and this is reflected in cell viability values. Prior to drug fitting, drug responses with high levels of noise were flagged and filtered out from the dataset. For this, noise was assessed based on the difference in cell viability between subsequent titration points (ΔCV_i) using the following index:

$$\eta_{drug} = \sum_{i=1}^{N-1} CV_i - CV_{i+1} - \text{range}(CV_1, \dots, CV_n)$$

Where the relative difference is given by $\Delta CV_i = CV_i - CV_{i+1}$, CV_i is the cell viability at the i-th titration point and N the amount of titration points for a specific drug-cell combination.

Notably, a low η_{drug} indicates a preservation of the curve's monotonic behaviour as the drug concentration increases (i.e., strictly increasing, decreasing or constant), whilst a high η_{drug} suggests a noisy curve with several fluctuations in the cell viability.

- **Outlier points**

The presence of a discrete amount of out-of-the-distribution cell viability points (i.e., outlier responses) can affect the curve fitting procedure and, in consequence, lead to a misclassification of the drug response.

Hence, the classic Cook's Distance outlier detection method (Cook & Dennis Cook, 1977) was used to identify and remove outlier points. In detail, the linear curve fit procedure (*Equation 5*) was employed to predict the drug response $\hat{\boldsymbol{y}}$, and the effect of removing the i-th $(i = 1, \dots, N)$ observation was assessed by:

$$\text{Cooks}_i = \frac{\sum_{j=1}^{N} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{\text{nr}_{\text{predictors}} * MSE}$$

Where N is the number of observations, $\hat{y}_j$ is the drug response at the j-th fitted response value, $\hat{y}_{j(i)}$ is the fitted response without the i-th observation, $\text{nr}_{\text{predictors}}$ is the number of predictors (i.e., coefficients) in the regression model and $MSE = \frac{\sum(\hat{y}_i - y_i)^2}{N}$ is the mean squared error (MSE).

Titration points were recognized as outlier points and filtered out from the analysis if they satisfied:

$$\text{Cooks}_i > 4 * \frac{\sum_{j=1}^{N} \text{Cooks}_j}{N}$$

That is, an observation was deemed an outlier point if its Cook's distance was higher than four times the mean of all the distances. In addition, to limit the filtering of observations across a single drug response, a maximum of 4 points were classified as outliers.

Following this post-processing stage, all the drug responses were estimated using the curve fitting methods presented in the chapter 2.3.2.

2.3.4 Drug response metrics

In order to describe the drug-cell line response, several metrics were considered. Specifically, measures based on the distribution of the cell viabilities (metrics prior to curve fitting), and metrics specific to the curve fitting (metrics after curve fitting).

- Metrics prior to curve fitting

The shape of the cell viability values distribution provides relevant information on the increasing, decreasing or constant behaviour of the drug response.

For instance, the indexes of Skewness and Kurtosis (Groeneveld & Meeden, 1984; Joanes & Gill, 1998) measure the symmetry and tail shape of the distribution with respect to a normal distribution:

- Skewness

Per definition, the Skewness of a dataset dictates how much the distribution deviates from a normal distribution. In particular, it separates the classification into symmetric or asymmetric distribution. Where the former indicates the dataset has an even distribution on both sides and resembles a normal distribution; while the latter suggests the data is skewed either to the left- (negative-skewed; Figure 12 A) or right-side (positive-skewed; Figure 12 C) of the distribution.

To infer this measure, the Fisher-Pearson coefficient of Skewness for each drug response $\boldsymbol{y} = (y_1, \dots, y_N)$ is computed:

$$S = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^3/N}{SD^3}$$

Where N is the number of observations, $\bar{y}$ the mean and SD the standard deviation.

Notably, $S = 0$ for a normal distribution (Figure 12 B), and approaches zero ($S \to 0$) for any symmetric data. Based on this knowledge, typically a dataset is labelled symmetric if $S \in [-0.5,0.5]$, negative-skewed if $S < -0.5$ (Figure 12 A) and positive-skewed if $S > 0.5$ (Figure 12 C).

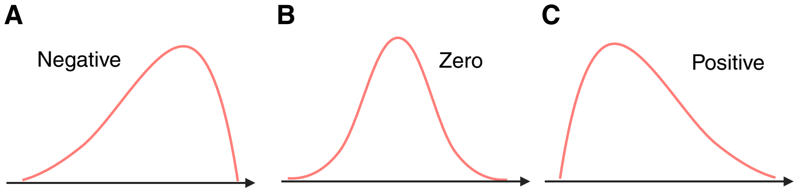

Figure 12: Examples of Skewness distributions.

(A) Negative-skewed distribution, with longer left tail and most of the distribution at the right side; **(B)** symmetric distribution with zero Skewness and data (approximately) equally distributed on both sides; **(C)** Positive-skewed with longer right tail and peak is towards the left side.

- Kurtosis

In parallel, Kurtosis measures the tailedness of a distribution with respect to a normal distribution. Commonly, this factor is employed to assess outliers in a distribution – light-tailed (Figure 13 A) indicate broader distribution peaks, shorter tails and less outlier-prone; while heavy-tailed (Figure 13 C) means a distribution with sharper peaks, longer tails and high likelihood of containing outliers spread across the dataset.

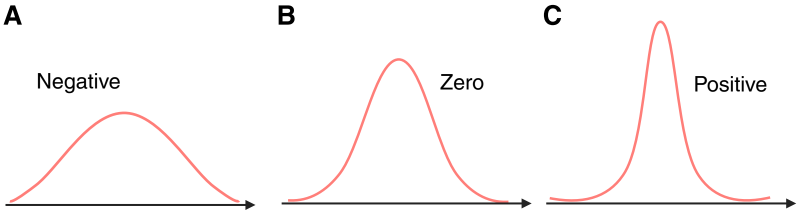

Figure 13: Examples of Kurtosis formulations.

(A) Negative Excess Kurtosis, with light-tails and flatter distribution; **(B)** normal-like distribution with zero Excess Kurtosis; **(C)** Positive Excess Kurtosis, with heavy-tails and sharper distribution.

The Kurtosis of a drug response $\boldsymbol{y} = (y_1, \dots, y_N)$ is given by:

$$K = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^4 / N}{SD^4}$$

For interpretation purposes, the measure Excess Kurtosis $\mathrm{K}_{\text{excess}} = \mathrm{K} - 3$ is frequently handled. For a dataset with normal distribution, the expected value of Excess Kurtosis is $\mathrm{K}_{\text{excess}} = 0$ (Figure 13 B). Moreover, a light-tailed dataset presents a negative Excess Kurtosis ($\mathrm{K}_{\text{excess}} < 0$; Figure 13 A) and a heavy-tailed one contains a positive value ($\mathrm{K}_{\text{excess}} > 0$; Figure 13 C).

Following the Cook's outlier detection procedure (chapter 2.3.3) application and based on their definitions, the Kurtosis and Skewness metrics were used to preliminary recognize probable increasing cell viability responses, i.e., negative-skewed with heavy-tails responses (points concentrated around high cell viability values).

- Metrics after curve fitting

Several summary statistics of the drug response curve fitting can be extracted from the constructed models. Briefly, both linear and Gaussian gradients were computed and used as a summary metric for these models. In addition, the Root Mean Squared Error (RMSE) was estimated to describe the models' fitness.

- Slope

Commonly, metrics like IC_{50} or area under the curve (AUC) are leveraged to characterise the drug response, as it is performed in the Sigmoid curve fit formulation by GDSC.

However, my research focus lies on the investigation and identification of unexpected drug responses, with emphasis on increasing cell viability cases. For this reason, a metric which explores the direction and steepness of the response, such as the curve gradient (i.e., slope), is a preferable measure.

In detail:

- a highly positive slope ($s \gg 0$) indicates a probable case of increasing cell behaviour;
- in contrast, a highly negative slope ($s \ll 0$) suggests a decreasing response;
- whereas, a close to zero slope ($s \cong 0$) is characteristic of non-responders.

Notably, the above formulations are conditioned by the slope magnitude and level of intrinsic noise in a particular drug response.

For the linear formulation (*Equation 5*) (Montgomery et al., 2021), the curve slope s is determined by the coefficient β_1 and defines the rate of change of the drug response $\boldsymbol{y}$ with respect to each unit increase of the drug concentration $\boldsymbol{x} = (x_1, \dots, x_N)$:

$$s = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Equation 7

Where $\bar{x}$ and $\bar{y}$ represent the average of the drug concentrations and cell viabilities over N observations, respectively

Alternatively, for the Gaussian model (*Equation 6*) (Rasmussen & Williams, 2005) the gradient of the relationship f is based on the differentiation of the covariate function K_Ψ. In detail, the expected gradient of the function f at concentration points $\boldsymbol{x}'$ (i.e., $f' = f(\boldsymbol{x}')$) can be reduced to the gradient of the covariate function:

$$E[\nabla f' | \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}'] = \nabla E[f' | \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x}'] = \nabla \sum_{i=1}^{N} \alpha_i K_\Psi(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{N} \alpha_i \nabla K_\Psi(\boldsymbol{x}, \boldsymbol{x}')$$

Equation 8

Where $\alpha = (\boldsymbol{K}_{\boldsymbol{\Psi}} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}$, with identity matrix I, hyperparameter σ and matrix $\boldsymbol{K}_{\boldsymbol{\Psi}}$ with covariate function values K_Ψ across all dosages.

The linear and Gaussian slopes (*Equation 7* and *Equation 8*) were estimated via the outputs of the lm() (R Core Team, 2016) and gausspr() (Karatzoglou et al., 2004) functions, respectively.

- RMSE

The Root Mean Squared Error (RMSE) was leveraged to evaluate the predictive power of the linear (*Equation 5*) and Gaussian (*Equation 6*) curve fitting methodologies. Intrinsically, this metric measures how accurately the drug prediction $\hat{\boldsymbol{y}}$ quantitatively varies with regard to the observed cell viabilities $\boldsymbol{y}$:

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

Notably, low RMSE values close to zero (RMSE → 0) indicate a suitable curve fit model with small deviations to the original response.

2.3.5 Association analysis

Linear and Gaussian drug response models (*Equation 5* and *Equation 6*, respectively) were fitted on the GDSC pharmacogenomic data, and slope information for each cell-drug combination was estimated. Following, these models were integrated with cancer-specific functional events' data (i.e., pan-cancer tissue was neglected) to investigate significant cases of increasing cell viability (ICV) across combinations of **A**lteration (somatic mutations, gene fusions or copy number alterations), **T**issue (TCGA labelled) and **D**rug types (ATD).

I considered the following distinct mathematical approaches for the exploration of significant ATD combinations containing ICV biomarkers:

- ANOVA

Under the assumptions of independency and normal distribution of the slope of the drug responses, ANOVA univariate models based on the mutational status (M_{status}) were built to test the slope distribution for each k-th ATD combination:

$$\text{slope} \sim M_{status},$$

where M_{status} is a binary variable defining mutant (=1) or wild-type (=0) oncogenes.

Following, the resulting p-values were corrected with the Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995) and the FDR threshold of $\alpha = 5\%$ was used to define significant associations.

In order to guarantee statistical power and focus on oncogenes with potential biomarkers of increasing cell viability, significant ATD candidates were selected for further investigation if the alterations contained at least 2 mutant/wild-type cell lines and positive delta mean slope between mutant and wild-type populations ($\Delta\text{slope} = \text{slope}_{\text{MT}} - \text{slope}_{\text{WT}} > 0$).

- Ranked hypergeometric test

An alternative approach is to investigate ATD combinations based on the enrichment of high valued slopes within the mutant population, which may indicate a driver mutation for a given drug. For this, the slopes of all N drug responses available in the dataset were initially ranked from the highest to the lowest fit slope ($\text{slope}_i > \text{slope}_j$, for $i < j$). Succeeding, information regarding wild-type populations was discarded and enrichment scores s_r for each k-th ATD combination were estimated:

$$s_r(\text{ATD}_k) = \text{ES}_{i^*}, \text{with } i^* = \underset{i}{\text{arg. max}}\, |\text{ES}_i|$$

$$\text{ES}_i = \begin{cases} 0, & \text{if } i = 0 \\ \text{ES}_{i-1} + \dfrac{1}{\sum_{j=1}^{K} \text{slope}_j} \left|\text{slope}_i\right|, & \text{if } 1 \le i \le N \text{ and } i \in \text{ATD}_k \\ \text{ES}_{i-1} + \dfrac{1}{N-K}, & \text{if } 1 \le i \le N \text{ and } i \notin \text{ATD}_k \end{cases}$$

Equation 9

Where K is the amount of drug responses with mutant status within the k-th ATD combination, N the total amount of drug responses available across all ATD combinations and $|.|$ represents the absolute value. Notably, in the formulation of *Equation 9* it is tested the statistical excess of mutants in the upper extreme of the cell line distribution (i.e., where the drug responses with the highest slopes are located).

Based on these formulations, p-values for positively enriched combinations $s_r(\text{ATD}_k) > 0$ (i.e., ATD combinations enriched with high slope values) were assessed:

$$p(\text{ATD}_k) = \frac{P(s_r(q) \geq s_r(\text{ATD}_k))}{P(s_r(q) \geq 0)}$$

Where q is a random selected set of slopes from the original pool of slopes ($\text{slope}_1, \dots, \text{slope}_N$) with the same size as the k-th ATD combination (i.e., $\text{size}(q) = \text{size}(\text{ATD}_k) = K$).

Similarly to the ANOVA approach, the p-values were corrected using the BH procedure and significant combinations were defined by $\alpha = 5\%$.

In addition to the significance inference, ATD combinations of interest were detected by the following criteria:

- Positive average slope of the mutant population $\text{avg slope}_{MT} > 0$ or positive delta mean slope $\Delta\text{slope} = \text{slope}_{\text{MT}} - \text{slope}_{\text{WT}} > 0$;
- Alterations with at least 2 mutant/wild-type cell lines;
- Positive enrichment score ($s_r(\text{ATD}_k) > 0$);

2.3.6 Correlation analysis

In order to suggest a novel and effective cancer therapy based on drug responses with increasing cell viability markers, it is needed to investigate compounds efficient with increased tumour growth (i.e., drugs able to reduce the viability of fast proliferating cell lines). Ultimately, the goal is to define potential compounds which synergistically combine with the formerly identified ATD combinations with ICV markers.

Accordingly, the cell growth rate of each cell line available in the GDSC dataset was assessed based on the measured intensities at days 0 (i.e., untreated cell lines) and 4 (i.e., cell lines treated with control compound DMSO) (Gonçalves et al., 2020):

$$\text{Rate}_{\text{cell growth}} = \frac{\text{mean}(I_{\text{post-treatment}})}{\text{mean}(I_{\text{pre-treatment}})}$$

Equation 10

Where $\text{mean}(I_{.})$ is the average of all recorded intensities pre- and post-treatment at days 0 and 4, respectively.

Following, a correlation analysis between the drug responses' IC_{50} value (assessed by the GDSC sigmoid formulation *Equation 4*) and the cell growth rate (*Equation 10*) was performed for each drug and cancer type combination:

$$R = \text{cor}(\text{IC}_{50}, \text{Rate}_{\text{cell growth}})$$

Intrinsically, the top negatively correlated drug-tissue specific associations were highlighted and investigated in the context of its drug targets and pathways to hypothesize synergistic drug combinations.

2.4 wPPI network for tissue specific drug response modelling

In the following sections I describe the methodology supporting the published Bioconductor package wPPI (Galhoz et al., 2021), and indicate how this tool was manipulated to identify new cancer-specific relevant genes and predict drug response of pharmacological data.

2.4.1 Systems biology network wPPI

Network-based methods enable feature-oriented and disease specific analyses with the possibility of integrating several data layers (e.g., omics, cell type, phenotypic) for a more context-aligned investigation. With this in mind, I developed the wPPI framework (Galhoz et al., 2021) which identifies and ranks a collection of new genes with potential relevance to a particular disease, based on known disease genes, PPI network topology, functional scores and a path search algorithm (Figure 14).

In detail, the wPPI pipeline can be subdivided in five distinct stages:

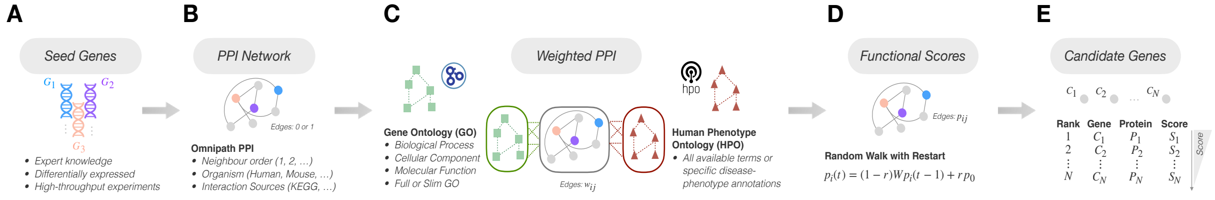

Figure 14: Schematic of the wPPI gene prioritisation tool.

(A) The framework starts by receiving a curated collection of established disease-specific genes of interest (seed genes), namely from expert knowledge, downstream analyses or high-throughput experiments. **(B)** A protein-protein interaction (PPI) network from Omnipath (Türei et al., 2021) is constructed around the given seed genes where each node represents a protein and edges denote an undirected relationship between interacting proteins. **(C)** In order to deduce functionally relevant interactions, the PPI network is integrated with ontology databases Gene Ontology (GO) (Ashburner et al., 2000) and Human Phenotype Ontology (HPO) (Köhler et al., 2021). Subsequently, protein interactions are weighted according to the network topological structure and functional similarities between the pair of proteins and ontology annotations. **(D)** Random Walk with Restart (RWR) search algorithm is applied to the weighted network to assess similarity probabilities between the given seed genes and genes in their vicinity (candidate genes). **(E)** At the final stage, the candidate genes available in the network are prioritised based on the inferred scores, and proposed as potential new disease-specific genes.

- Genes of interest

The wPPI algorithm prioritises candidate genes based on their similarity to known disease-relevant genes, i.e. seed genes. A curated list of seed genes can be derived from expert knowledge, transcriptomic analysis, downstream analysis (e.g., gene set and pathway enrichment analysis), genome-wide association studies (GWAS), drug targets or *gold standards* (Figure 14 A).

- Protein-protein interaction (PPI) network

The set of disease-specific genes of interest is expanded with *n-th* degree neighbors through a protein-protein interaction (PPI) network (Figure 14 B). The PPI data is acquired from the Omnipath database

(Türei et al., 2021), which comprises an extensive collection of over 100 signalling network resources with experimentally and predicted protein interactions, for several organisms (e.g., Human and mouse).

The built PPI network can be described using graph theory terminology. In detail, a graph $G = (V, E)$ is composed by a set of nodes $V = \{v_1, \ldots, v_M\}$, with M proteins, and a set of edges connecting pairs of proteins $E = \{e_{ij}, \forall(i, j) \in [1, M]\}$. An edge e_{ij} represents the undirected and binary association between proteins i and j, where an adjacent (or neighbour) pair of proteins equals 1, otherwise 0. These interactions can be represented by the adjacency correlation matrix

$$A = \begin{bmatrix} e_{11} & \cdots & e_{1M} \\ \vdots & \ddots & \vdots \\ e_{M1} & \cdots & e_{MM} \end{bmatrix},$$

$$\text{where } e_{ij} = \begin{Bmatrix} 1, \text{ if proteins } i \text{ and } j \text{ directly interact} \\ 0, \text{otherwise} \end{Bmatrix}.$$

- Functional similarity

In order to evaluate biologically relevant and context specific interactions in the network, each pair of interacting proteins of the constructed PPI graph network are weighted based on network topological features and functional annotations from public ontology databases (Figure 14 C).

Gene-phenotype relationships are inferred using the Human Phenotype Ontology (HPO) database (Köhler et al., 2021); whilst gene-gene functionalities are deduced from the Gene Ontology (GO) resource (Ashburner et al., 2000). In order to integrate the ontology annotation terms from HPO and GO with the PPI network, functional scores are calculated and assigned to each edge e_{ij}. Specifically, for each annotation term T, it is defined the ratio $G_T/_{G_{total}}$ between the number of genes annotated in T (G_T) and the total number of genes available in the GO/HPO resource (G_{total}). Building on this, shared annotation scores are calculated for each interacting pair of proteins i and j using the Fisher's combined probability score $s_{ij} = \sum_{\forall i,j \in T} -2 \log \left(G_T/_{G_{total}} \right)$ (R. A. Fisher, 1992).

In addition to the functional score assessment, the contribution of the network topology is taken into account through the number of common neighbours between connecting nodes $CN_{ij} = N_i \cap N_j$ (i.e., defined by the intersection of the neighbourhoods of genes i and j). This topological property essentially benefits interactions between proteins with similar local communities.

Succeeding, the original adjacency matrix (A) of the PPI graph G is converted to a weighted adjacency matrix normalized by column

$$A_w = \begin{bmatrix} w_{11} & \cdots & w_{1M} \\ \vdots & \ddots & \vdots \\ w_{M1} & \cdots & w_{MM} \end{bmatrix},$$

Equation 11

where each edge weight is defined according to its functional similarity scores from ontology databases and network geometry $w_{ij} = s_{ij}(\text{HPO}) + s_{ij}(\text{GO}) + CN_{ij}$.

In the interest of a more context specific analysis, the wPPI methodology can be customised to focus on a specific subset of HPO annotations (e.g., use only phenotypic annotations related to diabetes), one or more particular GO categories (i.e., Biological Process, Molecular Function and Cellular Component), and to consider the full or slim version of the GO database.

- Random walk with restart (RWR) algorithm

The Random Walk with Restart (RWR) search algorithm is employed to compute topological profiles between each pair of proteins, and therefore estimates how closely related the seed genes are to the candidate genes in the network (Figure 14 D) (Valdeolivas, Tichit, et al., 2019).

In detail, RWR is employed to the normalised weighted network A_w, and the seed genes are taken as starting points. Starting from a node v_i, the algorithm simulates a walker which, at every iteration, has the possibility to move to a direct neighbour with probability $p = 1 - r$ or return back to the starting position with probability $p = r$. The parameter r is the restart probability and can be fixed to a value between 0 and 1.

Formally, let $\boldsymbol{p}_0$ represent the initial probability vector with the i-th element equal to 1 and 0 in the other positions (i.e., the starting node is v_i), and $\boldsymbol{p}_t$ the moving probability vector of all nodes at iteration t. Hence, the probability vector at step $t + 1$ is given by

$$\boldsymbol{p}_{t+1}(v_i) = (1 - r)A_w\boldsymbol{p}_t + r\boldsymbol{p}_0$$

Where the transition matrix A_w is defined by the weighted adjacency matrix containing topological and functional information (*Equation 11*).

The RWR is applied over several iterations and the value of $\boldsymbol{p}_{t+1}$ is updated at each step until a steady state is reached. This is obtained when the condition $|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t|^2 < k$ is verified, where k is a predefined threshold.

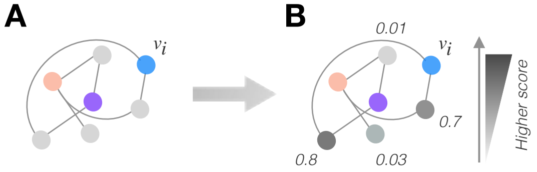

Figure 15: Schematic of Random Walk with Restart search algorithm.

(A) Initial PPI network with coloured seed genes (blue, purple and salmon) and candidate genes (in grey). For exemplification purposes, the node v_i coloured in blue is used as the starting point. **(B)** Application of RWR algorithm starting at node v_i until a steady state is reached. Mock resulting probabilities of transition to candidate genes are depicted and coloured accordingly (i.e., dark and light grey represent higher and lower scores, respectively). Notably, the candidate genes with higher scores reflect a higher functional and topological proximity to seed node v_i.

At the end of the algorithm, probabilities of transition to any node in the network are estimated and these values are used to rank the candidate genes in the PPI network (Figure 15). Notably, genes with higher probabilities are considered to be in close proximity to the seed genes, and consequently are more relevant to the disease than genes with lower probabilities.

- Gene prioritisation

Based on the prior that genes closely connected and functionally similar in PPI networks are potentially related to similar diseases, the candidate genes are prioritised based on their correlation to the given genes of interest (Figure 14 E).

Leveraging the transition probabilities estimated by the search algorithm RWR, gene scores are computed for all the candidate genes available in the network. Specifically, the final score for each candidate gene is described by the sum of its correlations in respect to the seed genes:

$$s_c = \sum_{i \in \{\text{seed genes}\}} p_\infty(i, c),$$

Equation 12

Where the steady state probability $p_\infty(i, c)$ defines the probability of the seed gene i reaching the candidate gene c.

According to this definition (*Equation 12*), the candidate genes are ranked and hypothesized as potential new genes of interest to the disease.

The wPPI gene prioritisation tool was ultimately consolidated to a R package, and published in the open-source project Bioconductor (Galhoz et al., 2021). In addition, a Shiny app was developed to enable users to interactively apply the wPPI and visualise the results (https://github.com/aGalhoz/wppi.shiny).

2.4.2 Lasso regression model

The network-based method wPPI can be leveraged to dissect new disease-specific features and subsequently build robust machine learning models for drug response prediction with the selected genes.

Let $\boldsymbol{y} = (y_1, \dots, y_N)$ be the vector of the drug response values of drug $d \in \boldsymbol{D}$, expressed by the logarithm of the half-maximal inhibitory concentration IC_{50}, and $\boldsymbol{g} = \{g_s, g_c\} = \{g_1, \dots, g_M\}$ a set of genes composed by input seed genes g_s and wPPI's candidate genes g_c.

I built Machine learning (ML) models based on LASSO regression (Tibshirani, 1996) to predict the drug response $\boldsymbol{y}$ using gene expression profiles from GDSC:

$$\boldsymbol{y}(d) = \beta_0 + \sum_{k=1}^{M} \beta_k X_k + \epsilon$$

Equation 13

Where the coefficients β_k are gene-specific estimators, $X_k = (x_{k1}, \dots, x_{kN})$ is the gene expression profiles vector for each k-th gene in the analysis, and ϵ is a random error term.

In order to select the best ML model, a hyperparameter tuning optimization step is performed, where the most optimal hyperparameter λ is evaluated. Specifically, the hyperparameter λ works as a penalty term on the model's features and defines the amount of shrinkage in the model, thence acting as a feature selection mechanism.

Let K_{top} define the set of selected features from the LASSO model (*Equation 13*). The selection of the parameter λ is achieved by estimating the coefficients $\hat{\beta}$ which minimize the error sum of squares (RSS):

$$L_{LASSO}(\hat{\beta}) = RSS + \lambda \sum_{k \in K_{top}} |\hat{\beta}_k| = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \sum_{k \in K_{top}} \hat{\beta}_k x_{ki} \right)^2 + \lambda \sum_{k \in K_{top}} |\hat{\beta}_k|,$$

Equation 14

Where y_i is the i-th drug response observation, x_{ki} the gene expression profile of the i-th cell line and k-th gene selected, N the amount of cell lines and $\hat{\beta}_k$ the LASSO coefficients.

Finally, the estimated regression coefficients (*Equation 14*) were fitted with the gene expression profiles of the selected genes to predict the drug response values $\hat{\boldsymbol{y}}$.

2.4.3 Performance metrics

I adopted the k-fold cross-validation procedure to perform hyperparameter optimisation and assess the performance of the built machine learning models.

To train and test the models, each group of cell lines treated by drug $d \in \boldsymbol{D}$ was randomly split into $k = 10$ (approximately) equally sized folds. Specifically, in each fold, $k - 1$ groups were used for model fitting (training dataset), and the model's performance was validated on the leaven out partition (test dataset). Next, this framework is repeated for every *k-th* fold, the performance across all folds is averaged and recorded as the predictive value of the drug response $\hat{\boldsymbol{y}}$ (Hastie et al., n.d.; Picard & Cook, 1984).

During the training phase, a 10-fold nested cross-validation procedure was employed for hyperparameter tuning. Accordingly, an additional validation set was created inside each training set fold and, through a grid search strategy, the optimal value of the LASSO's hyperparameter λ was evaluated.

Finally, the Pearson correlation coefficient (R) was used to assess the models' prediction accuracy by comparing the predicted ($\hat{\boldsymbol{y}}$) and observed ($\boldsymbol{y}$) IC_{50} drug response values (Lee Rodgers & Nicewander, 1988):

$$R_{pred-obs} = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}}$$

Where $\overline{(.)}$ represents the mean values of the variables.

2.4.4 Gene ontology enrichment analysis

Functional enrichment analysis of genes of interest prioritised by wPPI was conducted using ViSEAGO package (Brionne et al. 2019). This tool integrates a notable number of functional resources, such as gene annotations from Gene Ontology (GO) and pathway information from diverse sources (e.g., KEGG, MSigDB and REACTOME).

Using all the candidate genes as background, GO functional enrichment analysis was assessed through Fisher's exact test and reported in terms of Biological Process (BP). In detail, the over-representation of genes of interest in a specific GO functional annotation is tested by the following mathematical formulation:

$$P = \sum_{x=k}^{m} \frac{\binom{m}{x}\binom{n}{k+j-x}}{\binom{n+m}{k+j}}$$

Where:

- n is the total number of genes in the background (i.e., candidate genes)
- j is the total number of genes in a specific functional category
- m is the number of genes of interest
- k is the number of genes of interest in the functional category

The resulting p-values were adjusted for multiple hypothesis testing by the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), and significantly enriched pathways were defined with $\alpha = 5\%$.

3. Results

3.1 Indirect drug resistance in pharmacological screens

3.1.1 Overview of high-throughput drug screens

The drug response of the GDSC high-throughput screen (HTS) is expressed in terms of IC_{50} values. On the other hand, typically the data from CTRP is represented using AUC values. However, to ensure a common metric of representation across both pharmacological screens, IC_{50} values were computed for the CTRP dataset leveraging the sigmoid curve fit model used to estimate drug responses in the GDSC project (Vis et al., 2016).

In total, both drug screens comprised 814 unique compounds (397 from GDSC and 545 from CTRP), 820 unique cell lines (818 from GDSC and 490 from CTRP) across 19 different cancer types (Figure 16). Moreover, the cell lines and compounds were tested together and yielded 317,357 and 220,461 drug responses for GDSC and CTRP, respectively.

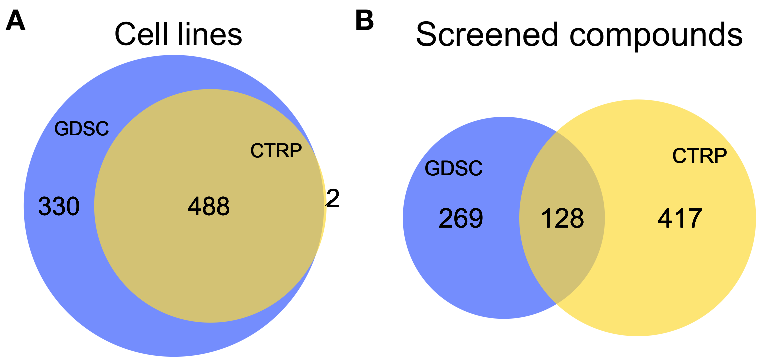

Figure 16: Overview of pharmacological screens represented in Venn diagrams.

(A) GDSC and CTRP datasets contain drug response metrics of 820 unique cell lines. From these, 488 cell lines are shared between the drug screens, 330 cell lines are only available in the GDSC and 2 only in CTRP. **(B)** The cell lines were tested against 814 unique drugs. Of which, 128 drugs are found in both datasets, 269 solely in GDSC and 417 only in CTRP. The amount of cell lines and compounds in the GDSC and CTRP datasets are coloured in blue and yellow, respectively.
This image was created by Iñigo Ayestaran and extracted from panel H of the original supplementary figure 2 “Synopsis of pharmacology screens and response examples” by (Ayestaran et al., 2020) under a CC-BY-4.0 license (http://creativecommons.org/licenses/by/4.0/).

In order to investigate sensitive and resistant biomarkers within these pharmacological screens, cancer functional events (CFE) data from GDSC were integrated into the analysis. Looking into the mutation frequency of cancer types with large amount of cell lines (Figure 17), is possible to note how rarely mutations occur across the panels. Specifically, majority of the mutations appear in less than half of the cell lines, which highlights one of the main limitations in the investigation of resistant markers using HTS.

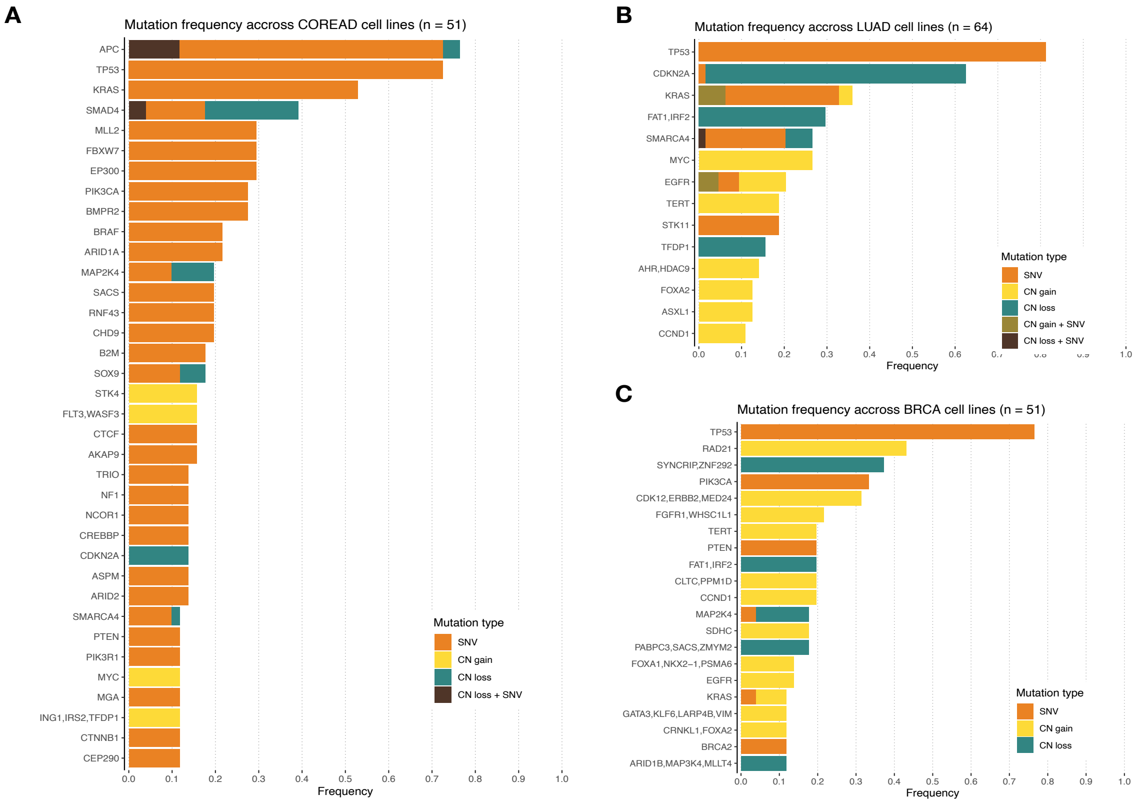

Figure 17: Mutational frequency in GDSC stratified by cancer type.

Cancer mutation events available from the GDSC project for **(A)** colorectal adenocarcinoma (COREAD), **(B)** lung adenocarcinoma (LUAD) and **(C)** breast cancer (BRCA). The mutations' frequencies are represented in ascending order of occurrence at the cancer cell lines of each cancer type. The mutations are divided into single nucleotide variant (SNV) and copy number (CN) gain/loss classes.
This image was produced by Iñigo Ayestaran and based on panels B and C of the original supplementary figure 2 "Synopsis of pharmacology screens and response examples" by (Ayestaran et al., 2020) under a CC-BY-4.0 license (http://creativecommons.org/licenses/by/4.0/).

3.1.2 Detection of sensitive cell lines

With the aim of building meaningful models with sufficient statistical power, the CFEs were reduced to cases where the mutations contain cancer driver genes, at least 4 mutated cell lines and majority of the cell line population with IC_{50} values bellow the maximal drug concentration. Based on these criteria, the GDSC and CTRP datasets were limited to 20,238 and 22,173 drug responses, respectively.

In order to identify sensitive biomarkers, an analysis of variance (ANOVA) with mutational status as input factor and adjusted to several medium-related covariates (i.e., tissue, medium, MSI and growth properties) was employed. The results were reported based on p-value significance and signed Cohen's effect size *D* (Figure 18). Notably, based on the definition of Cohen's effect size, a positive value indicates resistance markers, whilst a negative one indicates sensitivity. Thus, given the significance (p-value < 0,001) and focal (Cohen's effect size *D* < -1) criteria, the ANOVA models revealed 57 and 37 CFE-drug sensitive combinations for GDSC and CTRP, respectively.

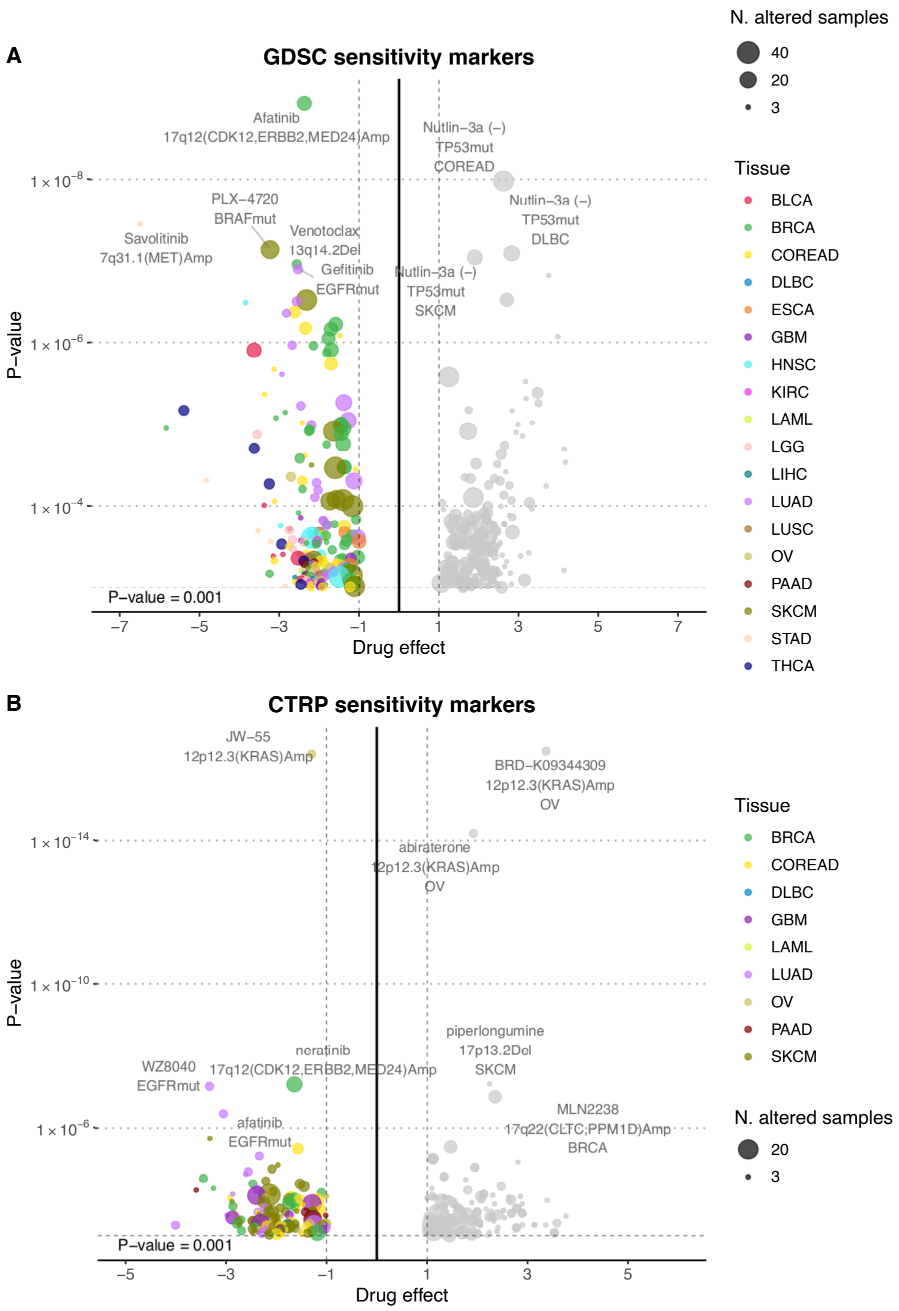

Figure 18: Volcano plots reveal sensitive markers based on drug response and ANOVA analysis.

Biomarkers obtained with **(A)** GDSC and **(B)** CTRP drug response, respectively. The x-axis denotes Cohen's drug effect size and the y-axis the significance of the CFE-drug association. The sensitive associations are located in left-hand size of the plot, where the drug effect size is negative. In total, 57 and 37 sensitive associations were discovered for GDSC and CTRP, respectively. The point size represents the amount of mutated cell lines, and is coloured by the cancer type.
This figure was created by Iñigo Ayestaran and extracted from the panels A and B of the supplementary figure 3 "Results for sensitivity biomarker discovery and CTRP UNRES identification" by (Ayestaran et al., 2020) under a CC-BY-4.0 license (http://creativecommons.org/licenses/by/4.0/).

3.1.3 Identification of UNexpectedly RESistant (UNRES) cell lines

The previously discovered CFE-drug associations with sensitivity markers are the starting point for the exploration of resistant biomarkers.

The UNRES discovery pipeline was leveraged to flag putative UNRES cell lines (see chapter 2.2.2). In detail, for each sensitive CFE-drug association, the framework analyses the effect in the standard deviation distribution when the cell lines with the highest IC_{50} values (up to 5 cell lines) are neglected.

Based on the significance threshold p-adjusted < 0.15, it was detected a total of 53 and 35 potential UNRES associations in the GDSC and CTRP screens, respectively (Figure 19, Appendix Table 11 and Table 12).

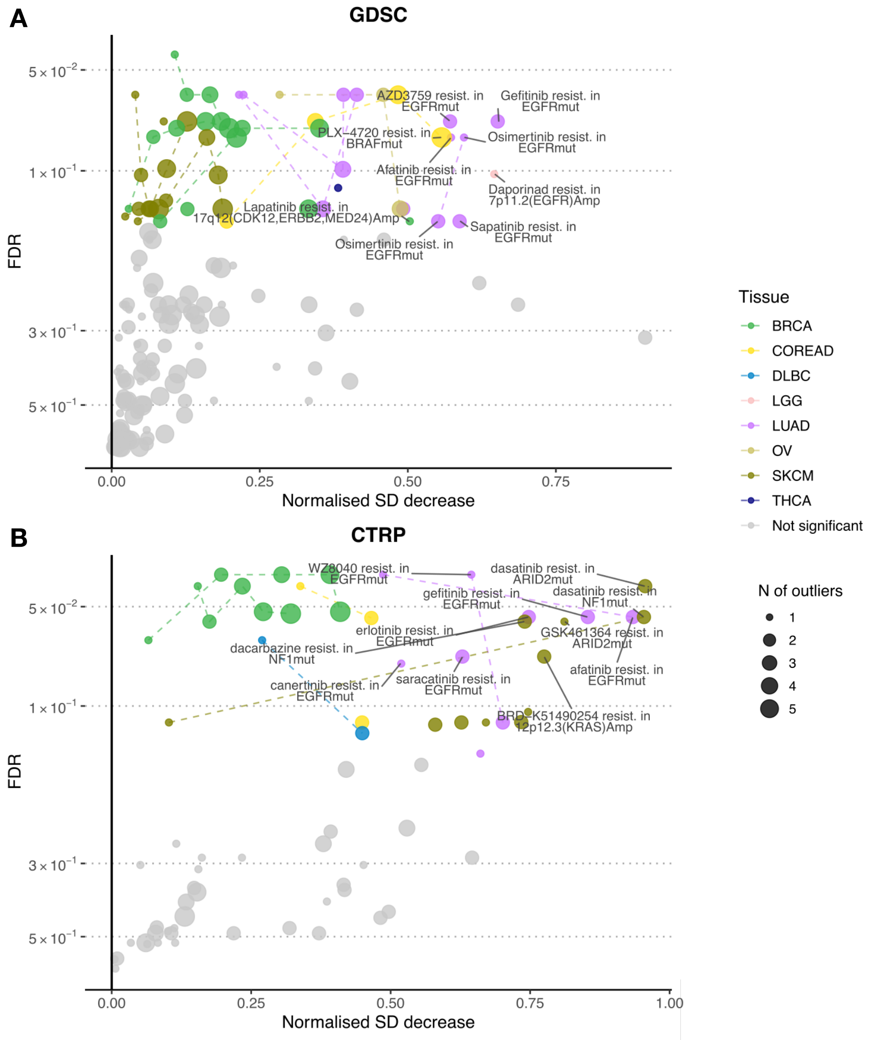

Figure 19: UNRES cell lines in high-throughput screens.

Identification of sensitive CFE-drug associations with resistance markers in **(A)** GDSC and **(B)** CTRP. The x-axis denotes de standard deviation (SD) delta when the cell lines with the i-th $(i = 1, \dots, 5)$ highest drug response are removed from the distribution. Significance was assessed through a bootstrap mechanism, and the adjusted p-values (FDR) were used to define significant bootstrap estimates via FDR < 15%. The significant UNRES associations are highlighted in the plot, coloured based on tissue type and circle size equal to the number of cell lines flagged as resistant. In addition, resistant cell lines from the same sensitive association were recognized through dashed lines.

This image was created by Iñigo Ayestaran and retrieved from panels B and C of the figure 1 “Identification of UNRES Cell Lines” by (Ayestaran et al., 2020) under a CC-BY-4.0 license (http://creativecommons.org/licenses/by/4.0/).

Looking into the resulting volcano plots (Figure 19), associations in lung adenocarcinoma are some of the most significant resistant markers with high standard deviation delta identified in both GDSC and CTRP.

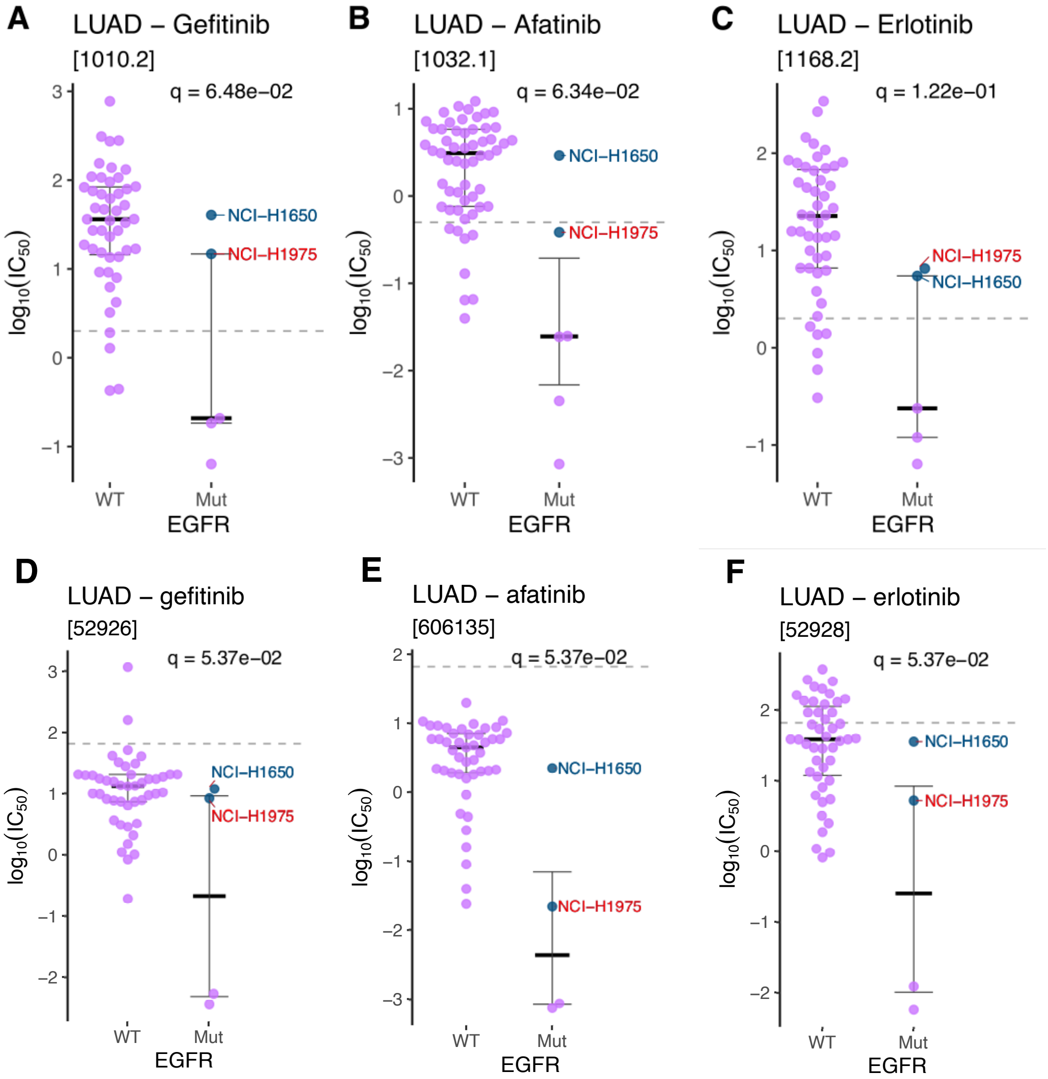

Figure 20: Boxplots of drug responses with resistant cell lines in lung adenocarcinoma.

Drug responses in lung adenocarcinoma (LUAD) for drugs gefitinib, afatinib and erlotinib in **(A)-(C)** GDSC and **(D)-(F)** CTRP, respectively. The highlighted mutant cell lines NCI-H1650 (PTEN mutation, coloured in blue) and NCI-H1975 (EGFR T790M mutation, coloured in red) were identified as resistance cell lines. The y-axis represents the drug response IC_{50} values in log form and the significance is denoted by the p-adjusted value (q value) from the bootstrap estimates.

This figure was created by Iñigo Ayestaran and compiled using panels A-C of figure 2 “Integration of Identified Hits with Public CRISPR Datasets” and panels C-E of supplementary figure 3 “Results for sensitivity biomarker discovery and CTRP UNRES identification” by (Ayestaran et al., 2020) under a CC-BY-4.0 license (http://creativecommons.org/licenses/by/4.0/).

Remarkably, the T790M mutated NCI-H1975 cell line reveals resistance to several first- and second-generation EGFR inhibitors, including afatinib, erlotinib and the gold-standard gefitinib (Figure 20). Furthermore, the NCI-H1650 cell line with PTEN deletion was also consistently identified as a resistant cell line across EGFR inhibitors (Figure 20). Noteworthy, this cell line disclosed resistance to compounds targeting the EGFR T790M mutation, namely the third-generation EGFR inhibitor osimertinib, and drugs WZ8040 and canertinib (Figure 21).

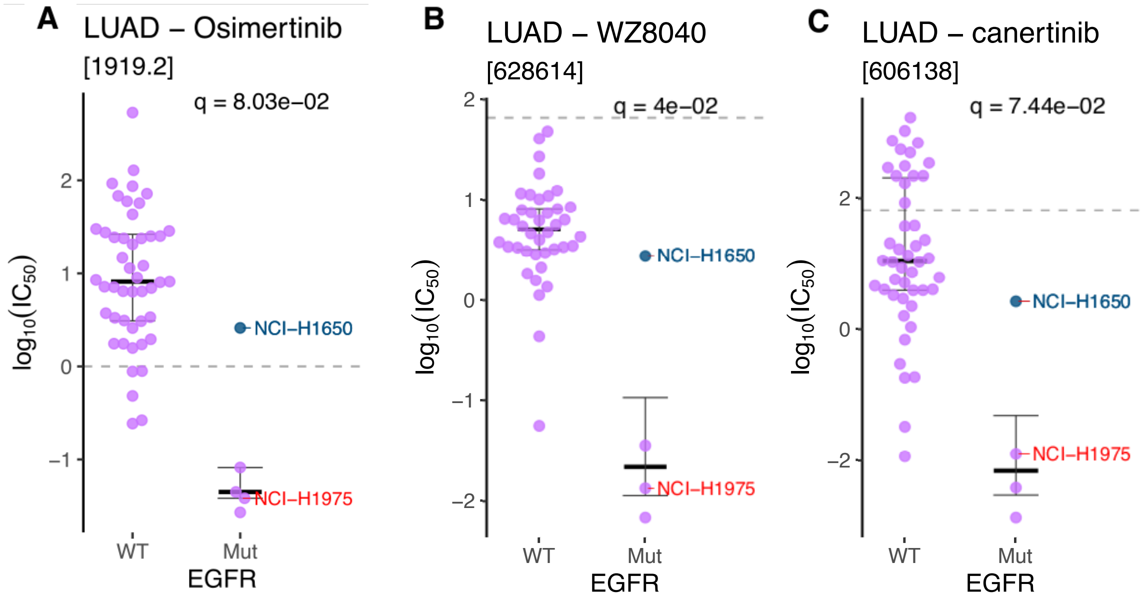

Figure 21: Boxplots depicting resistance of NCI-H1650 cell line in lung adenocarcinoma.

Drug response in lung adenocarcinoma (LUAD) for drugs osimertinib, WZ8040 and canertinib in **(A)** GDSC and **(B)-(C)** CTRP, respectively. The cell NCI-H1650 (in blue) reveals resistance to the drugs, whilst the T790M mutated cell NCI-H1975 (in red) is sensitive. Notably, this sensitive behaviour is not surprising since the tested compounds directly inhibit the EGFR T790M mutation. The x-axis stratifies the responses in wild-type (WT) and mutant (MT) cell lines and the y-axis delineates IC_{50} values in log form. Statistical differences between WT and MT groups were assessed through bootstrap and reported by the corrected p-value (q value).

This image was created by Iñigo Ayestaran and compiled using panel E of figure 2 “Integration of Identified Hits with Public CRISPR Datasets” and panels G-H of supplementary figure 3 “Results for sensitivity biomarker discovery and CTRP UNRES identification” by (Ayestaran et al., 2020) under a CC-BY-4.0 license (http://creativecommons.org/licenses/by/4.0/).

3.1.4 Benchmark with state-of-the-art outlier detection method

The presented UNRES framework was benchmarked against the established outlier detection method Neyman-Pearson (NP). Notably, per definition, here the concept of outlier is defined by a mutant cell line whose drug response is out of the distribution of the sensitive subpopulation with a specific CFE-drug combination.

In summary, under the significant threshold $\alpha = 15\%$, the NP approach identified a total of 26 and 8 potential resistant associations in the GDSC and CTRP drug screens (Table 1 and Table 2), respectively. From these, 12 and 5 associations were shared with the proposed UNRES framework for GDSC and CTRP, respectively.

It should be noted that, in comparison with the UNRES method, the NP results were distributed across a larger and a smaller amount of unique CFE-drug associations for GDSC (14 against 12) and CTRP (3 against 17), respectively. In addition, the NP approach failed to detect resistant gold-standards for lung adenocarcinoma, namely the EGFR-TKIs erlotinib and afatinib (in both screens, Figure 20 B-C, E-F), and gefitinib (in CTRP, Figure 20 D).

Tissue:drug:mutation	Method	resistant outliers UNRES	N. resistant outliers NP
BRCA-1799.2:gain:cnaBRCA26 (CDK12,ERBB2,MED24)	both	2 out	5
BRCA-255.1:gain:cnaBRCA26 (CDK12,ERBB2,MED24)	both	1,2,3,4,5 out	6
BRCA-255.1:gain:cnaBRCA27 (CLTC,PPM1D)	both	3,4 out	4
BRCA-293.1:gain:cnaBRCA27 (CLTC,PPM1D)	both	4 out	4
COREAD-1371.1:BRAF_mut	both	2,3,4,5 out	5
LUAD-1010.2:EGFR_mut	both	2 out	2
LUAD-1915.2:EGFR_mut	both	2 out	1
OV-326.1:PIK3CA_mut	both	1,2,3 out	2
SKCM-1036.1:BRAF_mut	both	4,5 out	5
SKCM-1061.1:BRAF_mut	both	1,2,3 out	5
SKCM-1373.1:BRAF_mut	both	1,2,3,4,5 out	3
SKCM-1047.1:loss:cnaSKCM4 (BNC2,CDKN2A,JAK2,PSIP1)	both	1,2,3,4,5 out	11
BRCA-381.1:gain:cnaBRCA26 (CDK12,ERBB2,MED24)	UNRES	1,2,3,4,5 out	-
BRCA-119.1:gain:cnaBRCA26 (CDK12,ERBB2,MED24)	UNRES	1 out	-
LGG-1248.1:gain:cnaLGG16 (EGFR)	UNRES	1 out	-
LUAD-1010.1:EGFR_mut	UNRES	1,2,3 out	-
LUAD-1032.1:EGFR_mut	UNRES	1,2,3 out	-
LUAD-1032.2:EGFR_mut	UNRES	1 out	-
LUAD-1168.2:EGFR_mut	UNRES	2 out	-
LUAD-1377.1:EGFR_mut	UNRES	1 out	-
LUAD-1549.2:EGFR_mut	UNRES	2 out	-
LUAD-1919.2:EGFR_mut	UNRES	1 out	-

SKCM-1003.2:BRAF_mut	UNRES	1 out	-
THCA-1373.1:BRAF_mut	UNRES	1 out	-
THCA-1036.1:BRAF_mut	NP	-	1
SKCM-1036.2:BRAF_mut	NP	-	8
SKCM-1371.1:BRAF_mut	NP	-	9
COREAD-1373.1:BRAF_mut	NP	-	3
COREAD-1373.2:BRAF_mut	NP	-	4
SKCM-1373.2:BRAF_mut	NP	-	6
HNSC-1377.1:TP53_mut	NP	-	7
LGG-1495.1:gain:cnaLGG16 (EGFR)	NP	-	1
BRCA-1549.2:gain:cnaBRCA26 (CDK12,ERBB2,MED24)	NP	-	3
BRCA-1560.2:PIK3CA_mut	NP	-	5
BRCA-1561.2:PIK3CA_mut	NP	-	1
SKCM-173.1:loss:cnaSKCM4 (BNC2,CDKN2A,JAK2,PSIP1)	NP	-	2
LUAD-221.1:EGFR_mut	NP	-	2
COREAD-326.1:PTEN_mut	NP	-	2

Table 1: Resistant associations identified with UNRES and Neyman-Pearson in GDSC.

Combinations of cancer type, drug ID and mutation event with potential resistant markers identified by the UNRES and Neyman-Pearson (NP) frameworks under $\alpha = 15\%$, in the GDSC screen. The results from the frameworks are represented in a distinct format – for UNRES it is specific which cell lines were selected as outliers, whilst for NP approach it is specified how many outliers.

Tissue:drug:mutation	**Method**	**N. resistant outliers UNRES**	**N. resistant outliers NP**
BRCA-606144:gain:cnaBRCA26 (CDK12,ERBB2,MED24)	both	1,2,3,4,5 out	1
LUAD-362063:loss:cnaLUAD34 (FAT1,IRF2)	both	1 out	1
SKCM-616355:gain:cnaSKCM12 (KRAS)	both	2 out	2
SKCM-32622:NF1_mut	both	2 out	2
SKCM-616355:NF1_mut	both	2 out	2
DLBC-410270:MLL2_mut	UNRES	1,2 out	1
BRCA-418038:gain:cnaBRCA26 (CDK12,ERBB2,MED24)	UNRES	1,2,3,4,5 out	-
COREAD-417416:AKAP9_mut	UNRES	1,2 out	-
COREAD-58339:AKAP9_mut	UNRES	2 out	-
LUAD-52926:EGFR_mut	UNRES	2 out	-
LUAD-52928:EGFR_mut	UNRES	2 out	-
LUAD-606035:EGFR_mut	UNRES	2 out	-
LUAD-606135:EGFR_mut	UNRES	1,2 out	-
LUAD-606138:EGFR_mut	UNRES	1 out	-
LUAD-628614:EGFR_mut	UNRES	1,2 out	-
SKCM-27894:ARID2_mut	UNRES	1 out	-
SKCM-347813:ARID2_mut	UNRES	1 out	-

SKCM-52882:ARID2_mut	UNRES	2 out	-
SKCM-649862:ARID2_mut	UNRES	1 out	-
SKCM-411809:NF1_mut	UNRES	2 out	-
SKCM-52882:NF1_mut	UNRES	1,2 out	-
SKCM-632907:NF1_mut	UNRES	2 out	-
COREAD-28801:FBXW7_mut	NP	-	9
BRCA-606138:gain:cnaBRCA26 (CDK12,ERBB2,MED24)	NP	-	1
LUAD-622913:EGFR_mut	NP	-	3

Table 2: Resistant associations identified with UNRES and Neyman-Pearson in CTRP.

Combinations of cancer type, drug ID and mutation event with potential resistant markers identified by the UNRES and Neyman-Pearson (NP) frameworks under $\alpha = 15\%$, in the CTRP screen. The results from the frameworks are represented in a distinct format – for UNRES it is specific which cell lines were selected as outliers, whilst for NP approach it is specified how many outliers.

3.1.5 Hierarchical statistical significance estimation

The UNRES framework comprises two statistical tests hierarchically dependent. First, CFE-drug associations with sensitive markers (under $\alpha_{\text{sensitivity}}$ = 0.001) are investigated, and, within the rejected hypothesis, cell lines as potential markers of unexpected resistance (under $\alpha_{\text{resistance}}$ = 0.15) are distinguished. As a result, the model raises two potential concerns related to FDR control:

1. What is the upper bound for FDR rate of the whole UNRES hierarchical framework ($\text{FDR}_{\text{UNRES}}$)? Are the results of each step bounded by the system's FDR?
2. How does the UNRES model compares to a baseline model?

To address the first part, the hierarchical FDR (HFDR) control procedure was employed. Given the formulation *Equation 2*, an upper bound was estimated based on the number of discoveries and CFE-drug combinations across the sensitivity and resistance assessments. The global bounds were $\text{FDR}_{\text{UNRES}} = 22.57\%$ and $\text{FDR}_{\text{UNRES}} = 22.40\%$ for GDSC and CTRP, respectively. Notably, these estimates are slightly higher than the leveraged thresholds, meaning the reported findings are controlled by the global $\text{FDR}_{\text{UNRES}}$ for each dataset.

In addition, permutation resampling of the drug responses within a tissue type in GDSC and CTRP enabled the creation of baseline models. For each cancer type, sensitivity and resistant hypotheses were generated, and an estimation of the number of false positives was performed. The estimates revealed 15.6% and 12.1% portions of the detected UNRES as potential false positives for GDSC and CTRP, respectively. Remarkably, these inferred values are in accordance with the significant threshold $\alpha_{\text{resistance}}$ = 0.15 used to identify resistant cell lines.

3.2 Exploration of increasing cell viability in high-throughput screens

3.2.1 Unsupervised visualization of raw drug response

Pharmacological data was acquired from the GDSC project, and consisted of 420,273 raw drug-cell combinations across 988 cell lines tested over 516 compounds. For preliminary visualisation purposes, I leveraged the drug response available in the GDSC project computed with a sigmoid curve fit model (Iorio et al., 2016; Vis et al., 2016). The drug response was separated into monotonic (i.e., $\forall x \leq y, \mathrm{IC}_{50}(x) \leq \mathrm{IC}_{50}(y)$ and vice-versa) and non-monotonic cases. This resulted in 29,078 and 391,195 monotonic and non-monotonic drug responses, respectively.

Unsupervised visualisation of all cell viabilities (described in terms of IC_{50} values), showed no clear pattern in the drug response (Figure 22 A). However, focusing on the monotonic responses, an explicit organization of monotonically increasing and decreasing responses is observed. In particular, it is unveiled a striking clustering of cases with unexpected increasing cell viability upon drug treatment (Figure 22 B).

Although this preliminary discovery was based on monotonic responses, when scanning non-monotonic responses, several potential cases of increasing cell viability are equally noticeable (Figure 23).

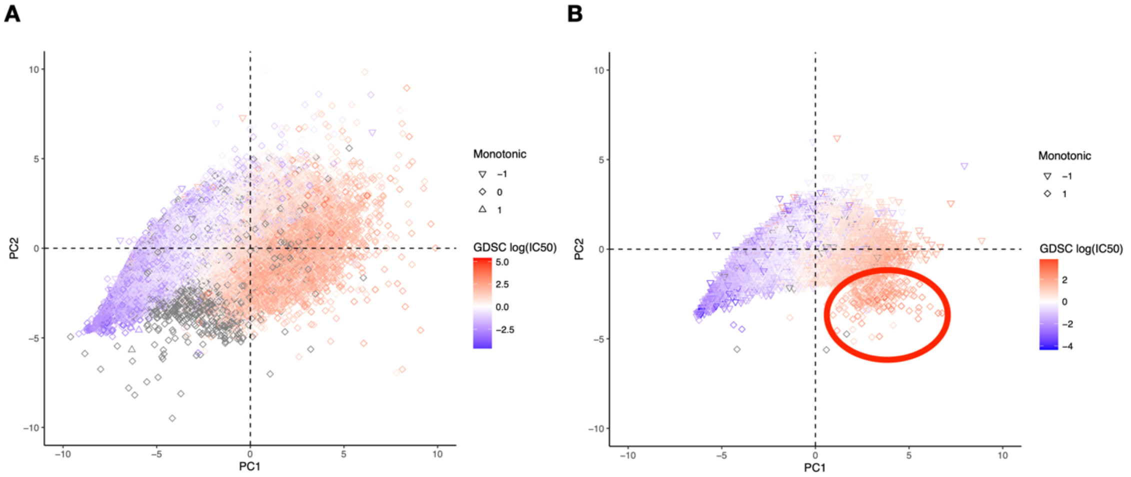

Figure 22: Principal Component Analysis (PCA) of cell viability values.

(A) All drug-cell combinations; **(B)** Monotonic increasing (+1) and decreasing (-1) responses. Region highlighted in red unfolds a notable clustering of increasing cell viability cases independent of IC_{50} values. The points are coloured based on IC_{50} values computed with the gdscIC50 package (Vis et al., 2016), and shaped according to non-monotonic (0), monotonic increasing (+1) and monotonic decreasing (-1) drug responses.

Briefly, these findings not only highlight the existence of unanticipated response behaviors across drug screens, but also suggest IC_{50} values are not an appropriate metric to describe drug response.

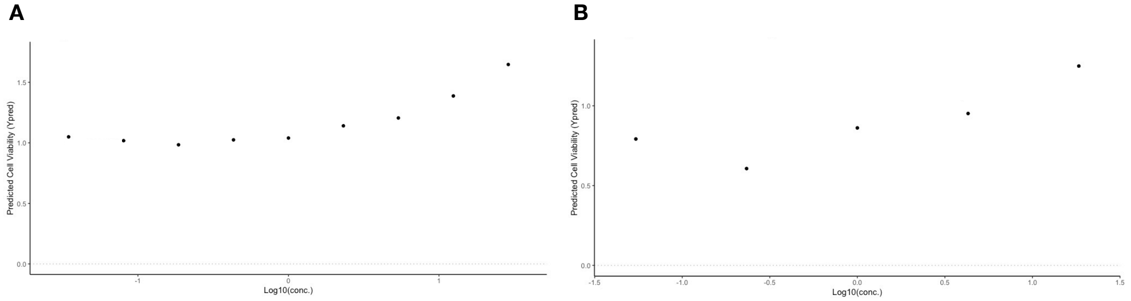

Figure 23: Examples of increasing cell viability in non-monotonic drug responses.

Raw drug responses with drug CHIR-99021 with cell lines **(A)** YT and **(B)** KP-4, respectively. In the x-axis it is represented the drug concentrations in log form, and in the y-axis the cell viability.

3.2.2 Curve measures prior to curve fitting

An initial attempt to characterise the drug responses based on the distribution of the cell viability values was performed using the Skewness and Kurtosis measures (see details in chapter 2.3.4). In detail, skewness provides information on the distribution direction (left, right or central) and kurtosis stratifies between peak and flat distributions. A normal distribution is characterised by zero skewness and zero kurtosis.

Notably, drug responses with increasing cell viability phenotypes would expectably present a distribution tilted to the right side with peak in the region of high cell viability $(\mathrm{CV} \gg 1)$ (Figure 24). Per definition, this would translate into negatively and positively valued skewness and kurtosis parameters, respectively.

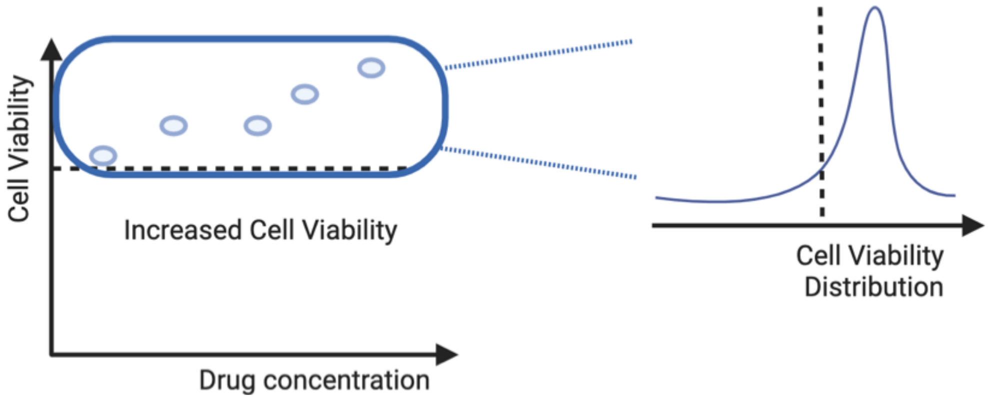

Figure 24: Distribution of drug response for increasing cell viability cases.

Increased cell viability responses present a non-normal distribution left skewed and with a sharp peak in the highlighted region.
Figure created with BioRender.com.

Looking into the distribution of the skewness and kurtosis parameters in monotonic and non-monotonic drug responses (Figure 25), no explicit patterns are detectable in the data. In particular, by focusing on the visualizations of monotonic responses (Figure 25 A-B), the previously highlighted region of interest (Figure 22 B) contains responses with no clear combination of skewness and kurtosis values.

Although the usage of these parameters bears the advantage of describing the drug responses in a computationally inexpensive manner, the results are highly influenced by the intrinsic noise in the responses. In addition, the outcomes of these parameters may be confounded with non-responders, and lead to inappropriate selection of phenotypes.

Ultimately, it is challenging to define an appropriate characterization of the drug responses solely based on the skewness and kurtosis parameters. In alternative, a comprehensive curve fitting methodology which incorporates the skewness and kurtosis parameters in its model and is robust against noise, is a more suitable solution to recognize “non-classical” phenotypes such as the increasing cell viability phenomenon.

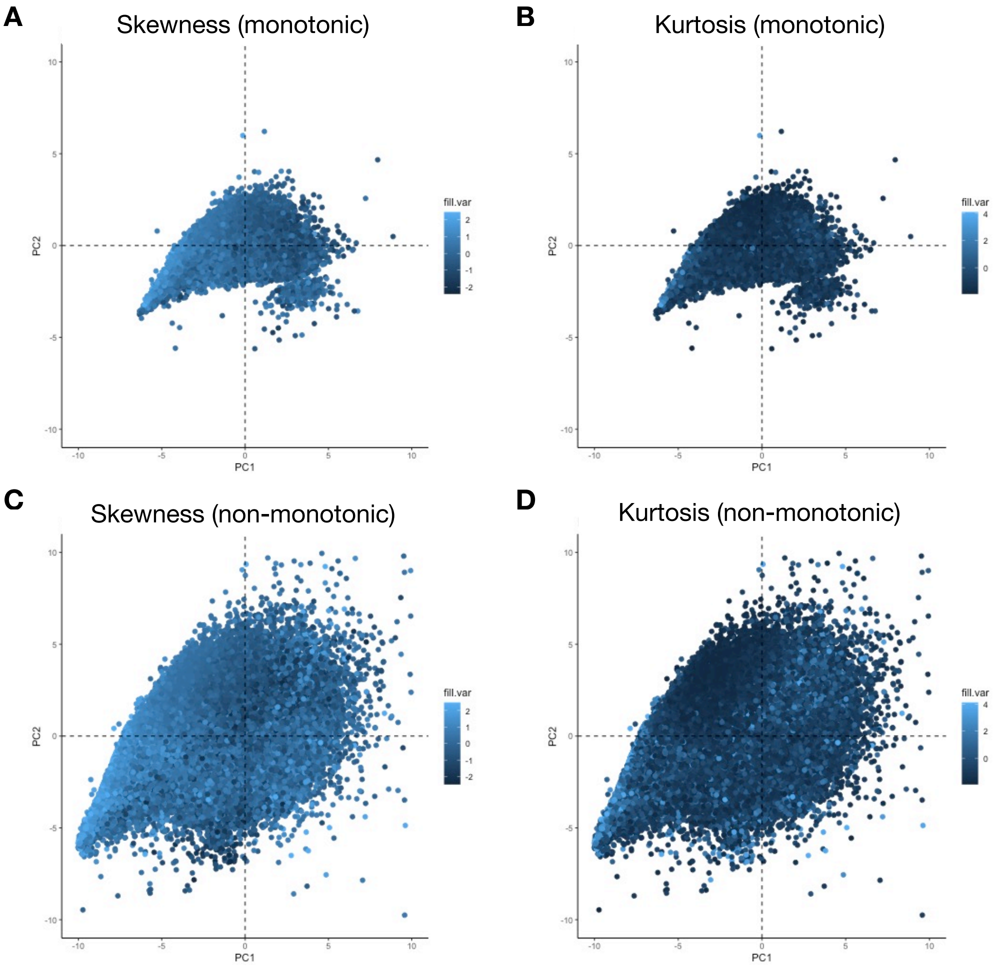

Figure 25: PCAs with Skewness and Kurtosis distributions.

PCAs of **(A)-(B)** monotonic and **(C)-(D)** non-monotonic drug responses. Plots coloured based on skewness and kurtosis values.

3.2.3 Pre-processing: noise and outlier detection

Drug response data is constitutionally noisy, and strongly influences the stratification between phenotypes in HTS (Figure 22 A). Sometimes this can be conditioned by a small number of points in the drug-cell response, or by a large amount of meaningless data distribution.

Noise was assessed based on the difference between the cell viability of subsequent concentration points (see details in chapter 2.3.3). According to its distribution (Figure 26), the threshold $\eta_{drug} > 2$ was employed to distinguish drug responses with high levels of noise where it is fundamentally not possible to estimate the true behaviour (Figure 27). This assumption identified 1,416 noisy responses and reduced the initial dataset to 418,857 drug-cell combinations.

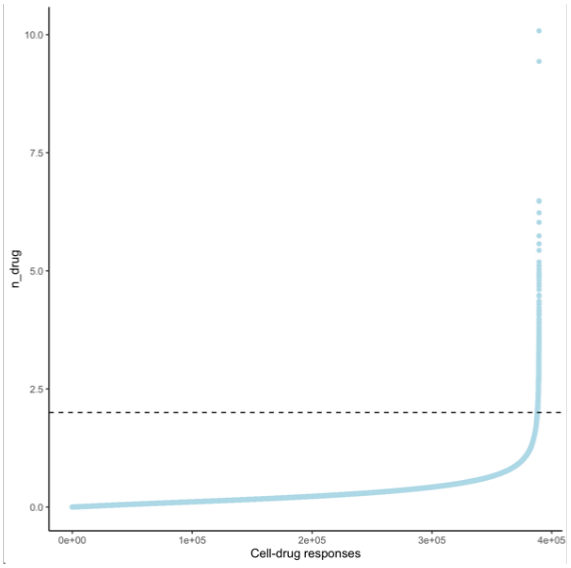

Figure 26: Distribution of noise across all combinations of cell and drug available in GDSC.

Noise was quantified based on the cell viability between subsequent titration points and represented by the index η_{drug}. A conservative filter $\eta_{drug} = 2$ was selected to remove unreliable responses (represented by the horizontal dashed line). The x-axis contains all the available drug-cell combinations and the y-axis the η_{drug} values.

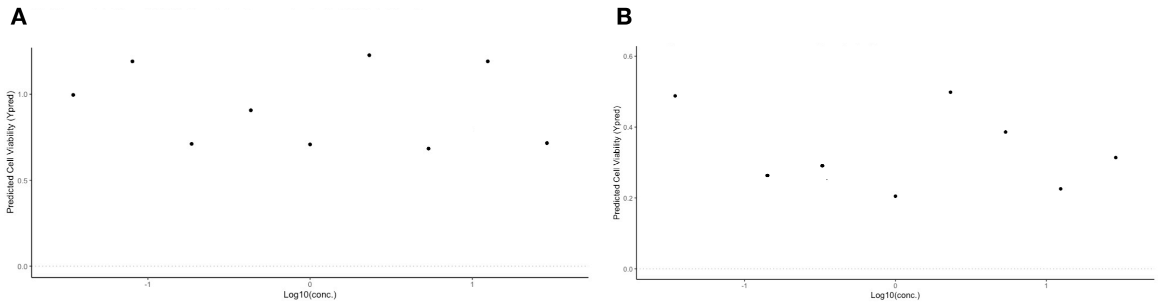

Figure 27: Examples of drug responses with high level of noise.

Drug responses of **(A)** cell line LAMA-84 with drug Doramapimod, and **(B)** cell line 22RV1 with drug Thapsigargin.

In order to avoid intrinsic changes on the real drug response due to the presence of out of the distribution points (Figure 28), I employed the classic Cook's distance outlier detection method to recognize and remove outlier points from the responses (details in chapter 2.3.3).

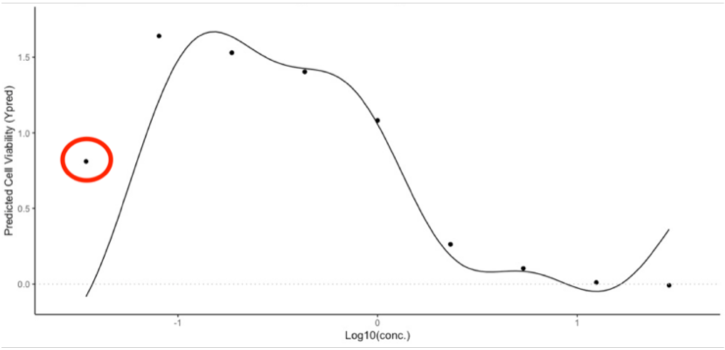

Figure 28: Example of drug response with outlier.

Drug response of cell line BB49-HNC and drug Obatoclax Mesylate with 9 titration points. The first observation is highlighted in red and was identified as an outlier point.

In summary, a total of 412,092 cell viability observations were identified as outliers and filtered out from the analysis. These outlier points were distributed across 322,366 unique drug-cell combinations and their quantity ranged between 1 and 4 points for each drug response (Table 3). Notably, majority of the drug responses included only one outlier point and 96,491 drug responses did not contain any detectable outlier.

Number of outliers	Number of drug-cell combinations
0	96,491
1	233,384
2	88,242
3	736
4	4

Table 3: Number of outliers across the drug-cell combinations in the dataset.

3.2.4 Curve Fitting: Linear and Gaussian Models

As previously discussed, currently available strategies to describe drug response in high-throughput screens are not suitable measures and neglect the identification of unexpected phenotypes, such as the increasing cell viability (details in chapters 1.4.2, 3.2.1).

In light of this, I explored two types of curve fitting methodologies, namely, Gaussian processes (GP) and linear models. Notably, these were modelled to enable the investigation of drug responses with increasing viability (see formulations of these models in chapter 2.3.2).

First, I investigated the curve fitting leveraging Gaussian Processes with Radial Basis and spline kernels. For this, I applied the gausspr function from the Kernlab R package with default parameters (Karatzoglou et al., 2004). Generally, the resulting curve fittings were successful in modelling the behaviour of the drug response and robust against non-linear alterations in the observations. In detail, owing to its sigmoidal like nature, the spline kernel revealed smoother fits in comparison with the Radial Basis kernel (Figure 29 B-C, E-F and Table 4). Despite the high fitting performance, the GP models bear high computational cost due to hyperparameter tuning and suggest potential overfitting profiles.

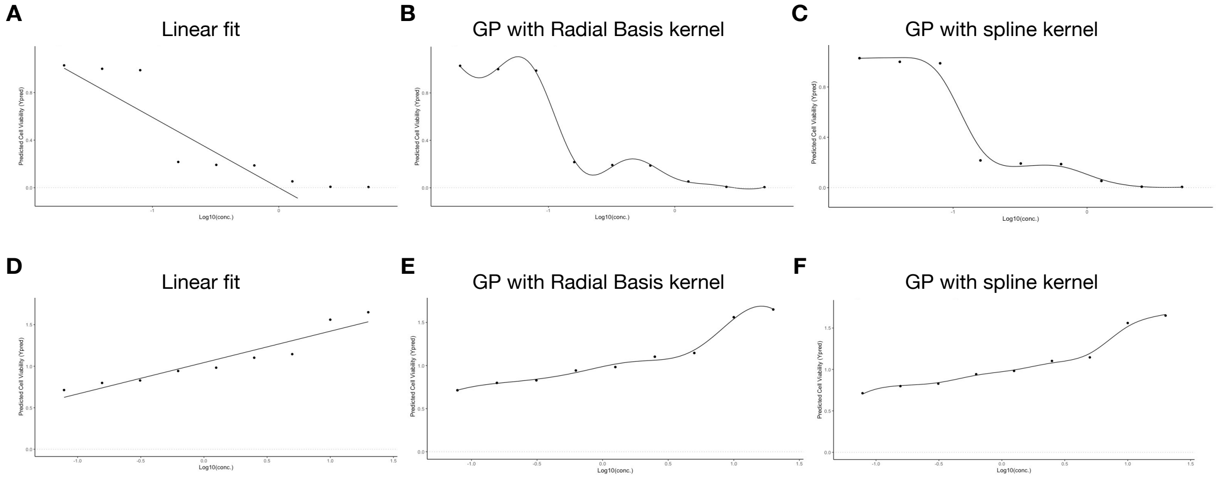

Figure 29: Examples of linear and Gaussian Processes curve fits.

Linear, GP with Radial Basis kernel and GP with spline kernel fittings for **(A)-(C)** NB14 cell line with drug Foretinib, and **(D)-(F)** CHP-212 cell line with drug Brivanib, respectively. The first horizontal panel illustrates a responder cell line and the second panel a case of increasing cell viability.

In addition, linear curve fits were also employed. In contrast to GP, linear models provide a less efficient curve fit and higher fitting errors (Figure 29 A, D and Table 4). However, these are rather simplistic and computational inexpensive models that can indicate the direction of the drug response, and, therefore, aid on the recognition of responses with increasing viability.

Type of curve fit model	Monotonic	Non-monotonic
Linear	$2{,}91 \times 10^{-1}$	$1{,}02 \times 10^{0}$
GP with Radial Basis kernel	$6{,}65 \times 10^{-3}$	$1{,}14 \times 10^{0}$
GP with spline kernel	$6{,}18 \times 10^{-3}$	$1{,}14 \times 10^{0}$

Table 4: Average RMSE values for linear and GP curve fits for monotonic and non-monotonic responses.

Furthermore, the gradient of the linear and Gaussian models was used to characterise the drug response. Based on the distribution of the slope of the linear and GP with spline kernel curve fitting models, I defined the subset of responses with high-slope based on the criteria slope$_{\text{linear}} > 0.2$ and slope$_{\text{Gaussian}} > 1.1$ (Figure 30).

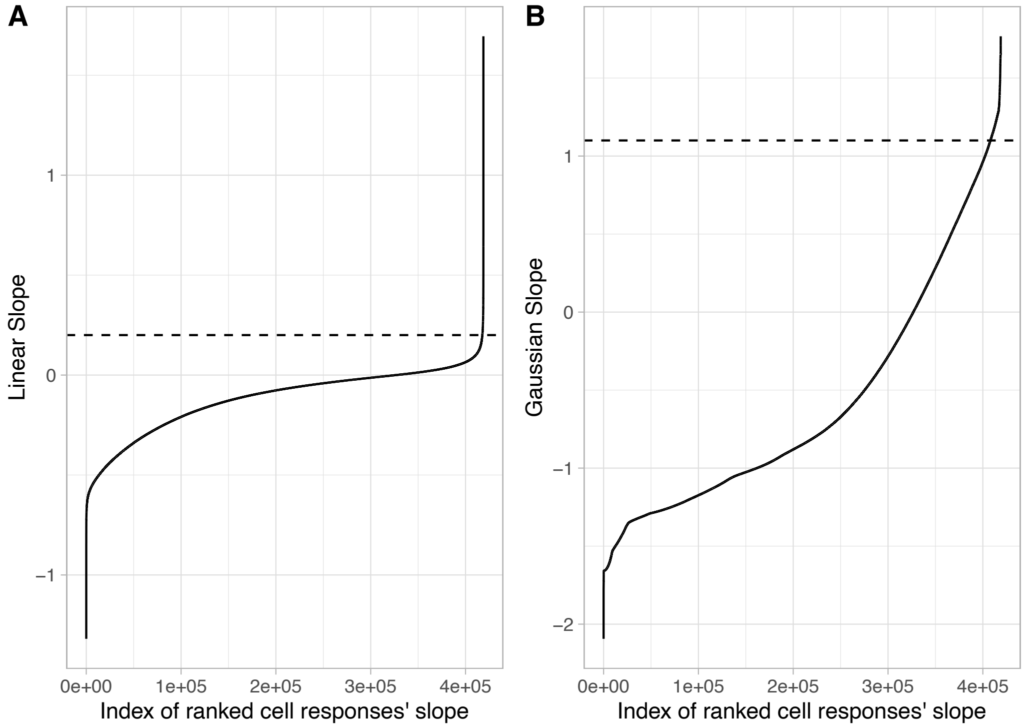

Figure 30: Distribution of slopes for drug responses in GDSC.

Slope distribution for drug-cell combinations using **(A)** Linear and **(B)** Gaussian with spline kernel curve fit models. Thresholds of slope$_{\text{linear}} = 0.2$ and slope$_{\text{Gaussian}} = 1.1$ defined on the elbow of the distribution of linear and Gaussian slopes, respectively. Thresholds represented by dashed horizontal lines.

3.2.5 Investigation of potential increasing cell viability markers

Based on the advantages given by both approaches, both linear and GP with spline kernel models were considered to investigate combinations of alteration, tissue and drug types (ATD) with significant markers of increasing cell viability in the mutant population.

Binary event matrices (BEM) were retrieved from the GDSC website (Iorio et al. 2016). These contained cancer specific information regarding cancer functional events (CFEs) of 998 unique cell lines. In particular, for the biomarker analysis it was considered cancer-type specific events BEMs, and therefore pan-cancer (PANCAN) BEMs were filtered out from the original dataset, thus reducing the original dataset to events regarding 988 cell lines.

Furthermore, only alterations with at least 2 cell lines in the mutant and wild-type populations were considered. In detail, these consisted of 40,286 combinations of alteration, tissue and drug types.

Subsequently, the reduced BEM information was merged with the linear and Gaussian drug response fittings. Given these datasets, two statistical frameworks based on ANOVA and hypergeometric tests were employed (details of each method in chapter 2.3.5).

For the ANOVA analysis, the aov() function (R Core Team, 2016) was applied to test the curve fit slope against the mutation status (slope $\sim \mathrm{M}_{\mathrm{status}}$). In parallel, for the hypergeometric enrichment analysis, the fgsea() function (Korotkevich et al. 2021) was employed to assess the enrichment of drug responses with high slope in the mutant population.

Based on the criteria of interest, positive delta mean slope $\Delta\text{slope} = \text{slope}_{\text{MT}} - \text{slope}_{\text{WT}} > 0$ and $\mathrm{FDR} = 5\%$, the ANOVA framework identified 48 and 125 significant ATD combinations using the linear and Gaussian curve fitting models, respectively (Figure 31, Appendix Table 13 and Table 14).

Similarly, leveraging the same criteria and focusing on positive enriched ATD combinations (i.e., with enrichment of high valued slopes), the hypergeometric enrichment framework distinguished 1,387 and 934 ATD combinations with the linear and Gaussian curve fitting models, respectively (Figure 32).

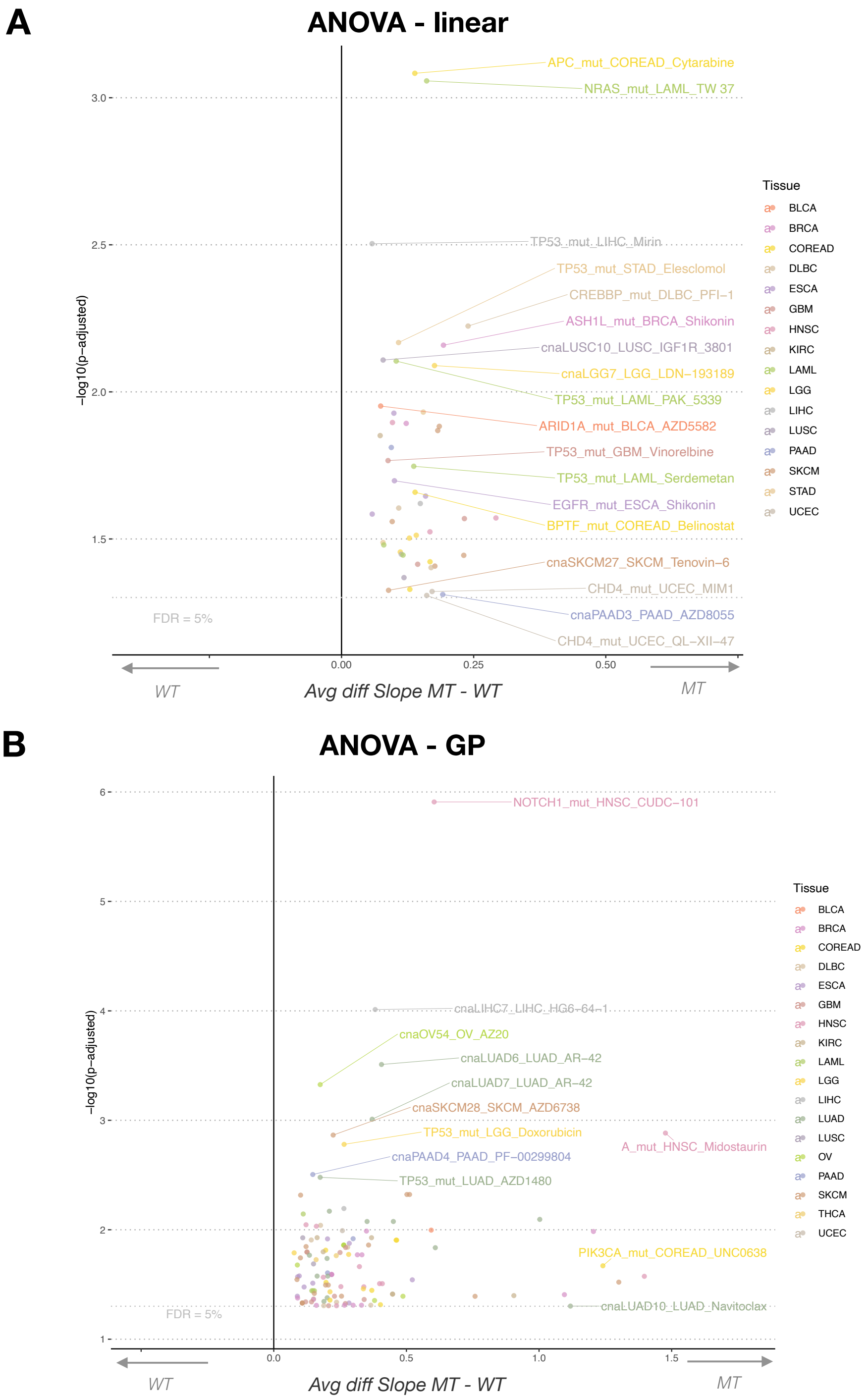

Figure 31: Volcano plots of ATD combinations with increasing cell viability markers from ANOVA pipeline.

ANOVA results using **(A)** linear and **(B)** Gaussian with spline kernel curve fitting models. Highlighted ATD combinations satisfy the significance criteria FDR $< 5\%$ and Δslope $=$ slope$_{\text{MT}}$ $-$ slope$_{\text{WT}}$ > 0, and are coloured according to the TCGA cancer abbreviation (48 and 125 ATD combinations for linear and Gaussian, respectively). The x- and y-axis represents the difference between the mean slope of the mutant and wild-type populations, and log p-adjusted value, respectively.

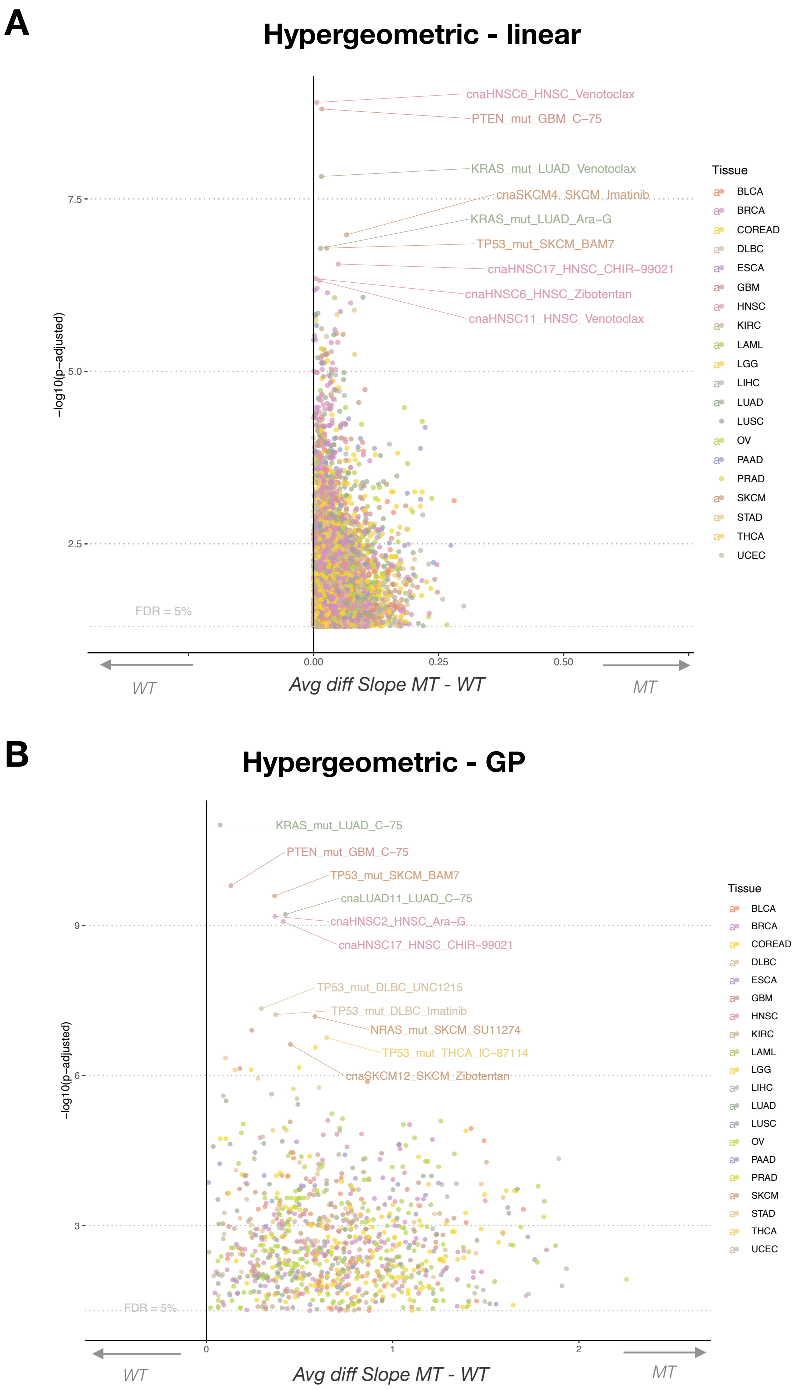

Figure 32: Volcano plots of ATD combinations with increasing cell viability markers from hypergeometric enrichment framework.

Enrichment results using **(A)** linear and **(B)** Gaussian with spline kernel curve fitting models. Highlighted ATD combinations satisfy the significance criteria FDR $< 5\%$, Δslope $=$ slope$_{MT}$ $-$ slope$_{WT}$ > 0 and positive enrichment score, and are coloured according to the TCGA cancer abbreviation (1,387 and 934 ATD combinations for linear and Gaussian, respectively). The x- and y-axis represents the difference between the mean slope of the mutant and wild-type populations, and log p-adjusted value, respectively.

3.2.6 Selection of increasing cell viability responses

The previously mentioned ANOVA and hypergeometric-based frameworks revealed thousands of potential ATD combinations with increasing cell viability markers. In order to assist on the prioritisation of these markers, for each ATD combination it was assessed the number of drug responses with high slope in the mutant population (i.e., $\text{slope}_{\text{linear}} > 0.2$ or $\text{slope}_{\text{Gaussian}} > 1.1$, Figure 30).

Briefly, for the ANOVA pipeline with both linear and Gaussian slopes, the majority of the ATD combinations did not include responses with high-slopes in the mutant population (Appendix Table 13 and Table 14). In fact, only 1 ATD combination with 2 high-slope responses and 3 ATD combinations with 1 high-slope responses were discovered. In contrast, with the hypergeometric enrichment framework, several candidate ATD combinations contained 2 up to 5 responses with high slope (Appendix Table 15 and Table 16).

In view of these results, I focused the investigation of increasing viability biomarkers to the findings obtained with the hypergeometric enrichment framework. Considering only ATD combinations with at least 2 responses with high slope, this unravelled a total 123 and 106 ATD combinations with increasing cell viability markers based on linear and Gaussian slopes, respectively (Figure 33, Appendix Table 15 and Table 16).

For each of these ATD combinations of interest, I meticulously selected candidates to be further analysed based on the linear and Gaussian curve fittings of the top 4 highest slopes, as well as, the slope distributions of the wild-type and mutant populations within each ATD combination. In detail, focus was drawn to cases satisfying:

- real increasing cell viability drug-cell responses – i.e., responses with full cell viability at the initial drug concentration ($\text{CV}_{\text{initial}} \cong 1$) that experience a significant escalation of the cell viability as the drug concentration is increased ($\text{CV}_i \to \infty,\ i \to \infty$);
- median slope of the mutant population noticeably higher than the wild-type one ($\text{median}(\text{slope}_{\text{MT}}) \gg \text{median}(\text{slope}_{\text{WT}})$).

Using these criteria, the set of ATD candidates was reduced to a refined subset of 19 ATD combinations (Table 5).

Together with my colleague Ginte Kutkaite, from these 19 candidates, we investigated cases where specific drugs have been noted to increase cell viability in several cancer types. Remarkably, only two drugs fell under these conditions: SB590885 in ESCA and KIRC and CHIR-99021 in HNSC, OV and LUAD.

Furthermore, based on publicly available literature resources and underlying alterations associated with increased cell viability in given cancer types, attention was drawn to the drug CHIR-99021, which is associated with two distinct amplifications in LUAD - cnaLUAD3 and cnaLUAD27 (Figure 34 and Figure 35). Specifically, the cnaLUAD3 segment contains a number of genes including known oncogenes *TERT* and *TRIP13*, while cnaLUAD27 includes another oncogene *MYC*. These findings are consistent with a well-established notion that oncogene amplification drives tumorigenesis by promoting genomic instability and ultimately uncontrolled cell proliferation (Hanahan and Weinberg 2000; Schwab 1998). Moreover, supporting our findings, several studies have reported synergistic effects between CHIR-99021 and chemotherapies in non-small cell lung cancer (NSCLC) and cholangiocarcinoma cells (Li et al. 2020; O’Flaherty et al. 2019).

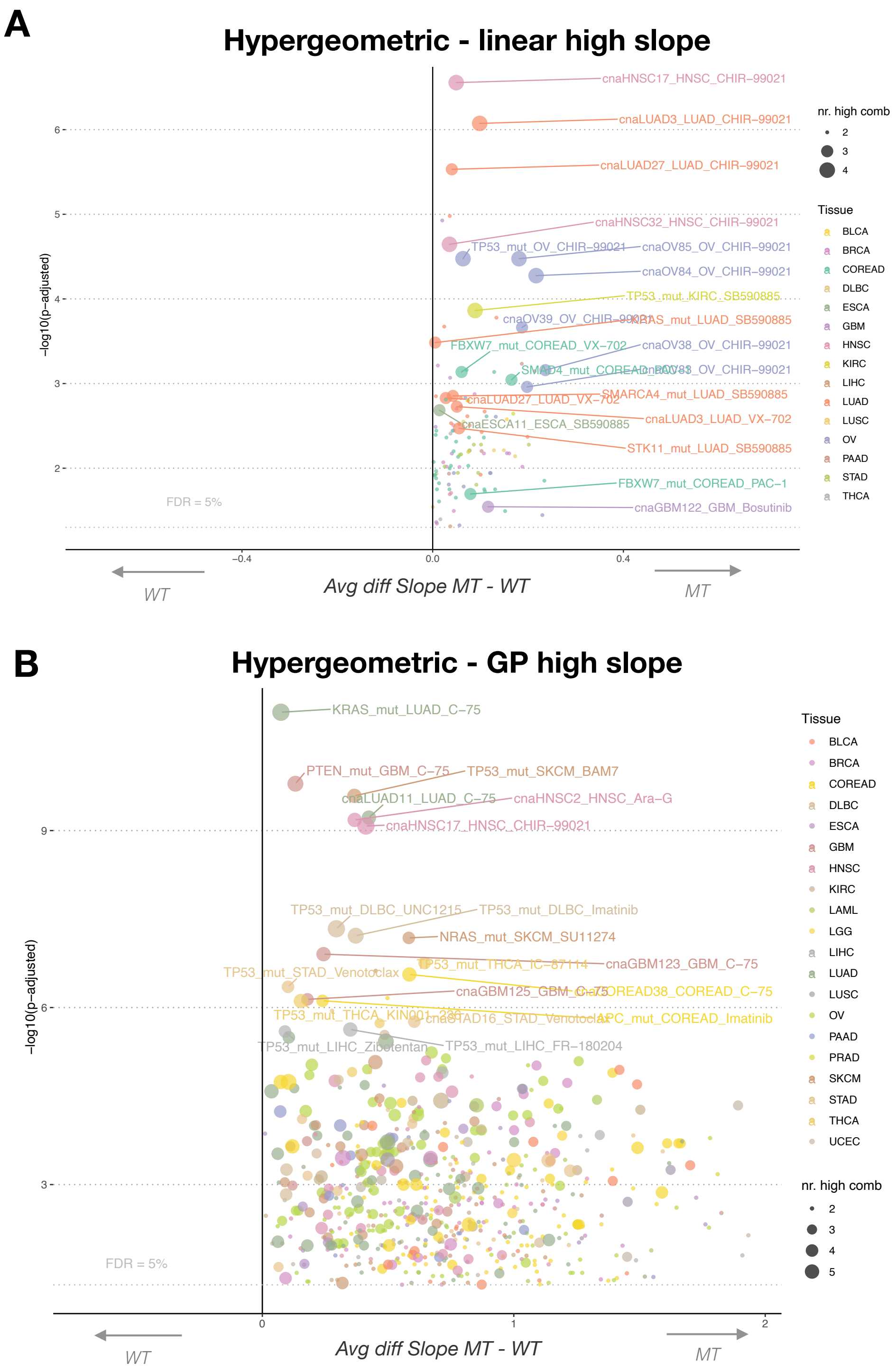

Figure 33: Volcano plots of ATD combinations with at least 2 drug responses with high slopes.

Hypergeometric enrichment analysis identified ATD combinations with increasing cell viability markers and at least 2 mutant responses with high slopes. Specifically, the framework revealed (**A)** 123 ATD combinations with linear slope and **(B)** 106 ATD combinations with Gaussian slope. Highlighted ATD combinations satisfied significance criteria and the points size is proportional to the number of responses with high slope in the mutant population.

Tissue	Drug	Drug Target	Target Pathway	Alteration	Diff. slope MT vs WT	p-adjusted value	Nr. high slope
OV	CHIR-99021	GSK3A, GSK3B	WNT signaling	cnaOV84	0,21699869	5,3084E-05	4
OV	CHIR-99021	GSK3A, GSK3B	WNT signaling	cnaOV85	0,18106223	3,3486E-05	4
LUAD	CHIR-99021	GSK3A, GSK3B	WNT signaling	cnaLUAD3	0,09866215	8,4423E-07	4
KIRC	SB590885	BRAF	ERK MAPK signaling	TP53_mut	0,08899524	0,00013713	4
HNSC	CHIR-99021	GSK3A, GSK3B	WNT signaling	cnaHNSC32	0,03523734	2,2635E-05	4
OV	CHIR-99021	GSK3A, GSK3B	WNT signaling	cnaOV38	0,2363685	0,00069351	3
OV	CHIR-99021	GSK3A, GSK3B	WNT signaling	cnaOV39	0,18749213	0,0002173	3
LUAD	CHIR-99021	GSK3A, GSK3B	WNT signaling	cnaLUAD27	0,0401948	2,9423E-06	3
LUAD	VX-702	p38	JNK and p38 signaling	cnaLUAD27	0,02756114	0,00148157	3
GBM	Bosutinib	SRC, ABL, TEC	Other, kinases	cnaGBM122	0,11619582	0,0285229	3
ESCA	SB590885	BRAF	ERK MAPK signaling	cnaESCA11	0,01421669	0,00206439	3
PAAD	IOX2	EGLN1	Other	ARID1A_mut	0,21400061	0,00071425	2
OV	AT7867	AKT	PI3K/MTOR signaling	cnaOV26	0,23012818	0,03581275	2
PAAD	IOX2	EGLN1	Other	CDKN2A_mut	0,18699293	0,00058396	2
OV	Doramapimod	p38, JNK2	JNK and p38 signaling	NF1_mut	0,13247077	0,0024625	2
LUAD	VNLG/124	HDAC,RAR	Chromatin histone acetylation	cnaLUAD14	0,13367648	0,00016655	2
OV	Doramapimod	p38, JNK2	JNK and p38 signaling	cnaOV94	0,05878087	0,0018564	2
OV	CHIR-99021	GSK3A, GSK3B	WNT signaling	cnaOV54	0,04438891	0,01140414	2
BRCA	NSC-87877	SHP-1 (PTPN6), SHP-2 (PTPN11)	Other	cnaBRCA18	0,00893763	0,04505482	2

Table 5: Top ATD combinations with increasing viability markers.

Highlighted combinations are the proposed candidates to be followed up. Information of drug target and target pathway were retrieved from the GDSC website.

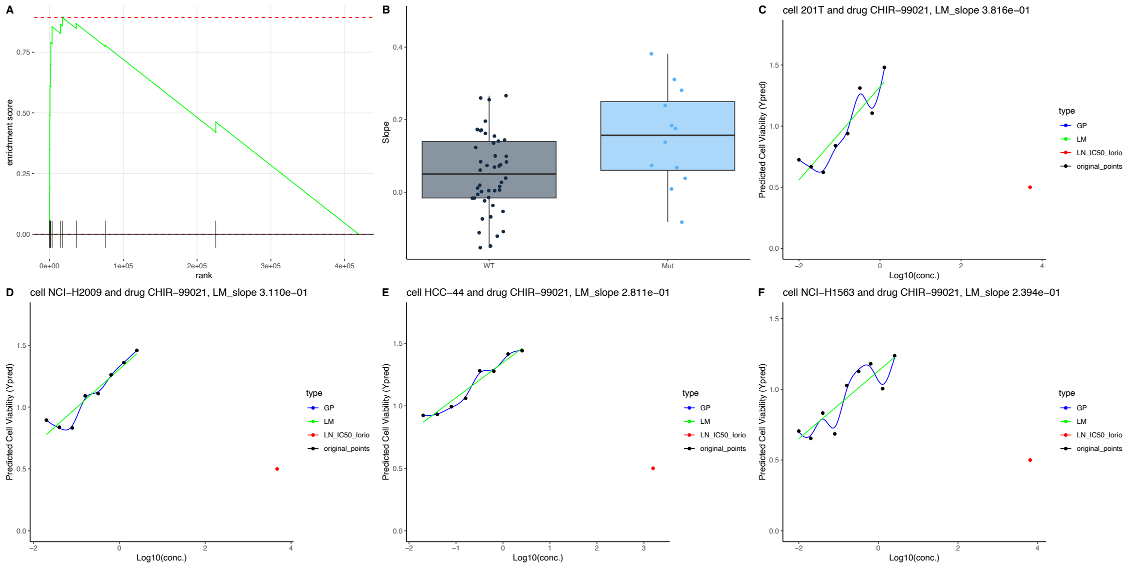

Figure 34: Increasing cell viability results for drug CHIR-99021 with cnaLUAD3 alteration in lung adenocarcinoma (LUAD).

(A) Hypergeometric enrichment score based on ranked slope; **(B)** boxplots of slope distribution of wild-type and mutant populations; **(C)-(D)** drug responses with the top 4 highest slopes, with cell lines 201T, NCI-H2009, HCC-44 and NCI-H1563.

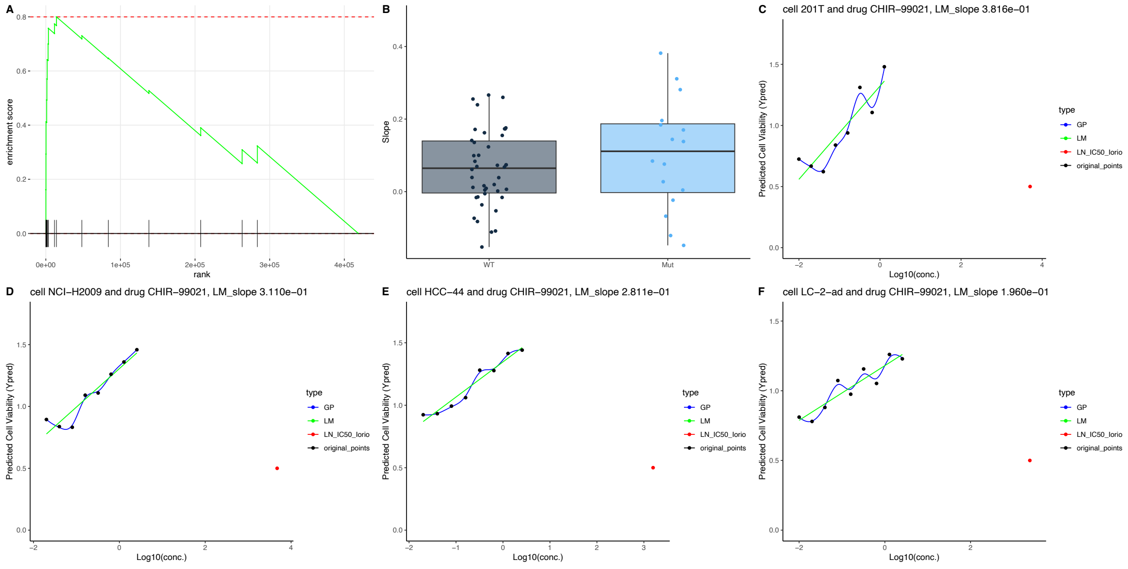

Figure 35: Increasing cell viability results for drug CHIR-99021 with cnaLUAD3 alteration in lung adenocarcinoma (LUAD).

(A) Hypergeometric enrichment score based on ranked slope; **(B)** boxplots of slope distribution of wild-type and mutant populations; **(C)-(D)** drug responses with the top 4 highest slopes, with cell lines 201T, NCI-H2009, HCC-44 and LC-2-ad.

3.2.7 Hypergeometric enrichment tests of cell lines and compounds

To extend and fine tune the list of putative candidates (Table 5), additional hypergeometric tests on the subset of high slopes (i.e., slope$_{\text{linear}} > 0.2$ or slope$_{\text{Gaussian}} > 1.1$, Figure 30) were performed to identify significantly enriched combinations of drug and tissue type. In contrast to the previous framework, this analysis was performed independent of cancer functional events.

With a threshold of p-adjusted value < 5%, I discovered a total of 29 significantly enriched drug-tissue types (Table 6). Remarkably, the top pairs include the candidates of increasing cell viability highlighted in the previous section, namely the compound CHIR-99021 in the tissues HNSC, OV and LUAD, and SB590885 in KIRC (Table 5).

Tissue	Drug	Drug Target	Target Pathway	p-adjusted value	Nr. high slope	Nr. responses
LUAD	CHIR-99021	GSK3A, GSK3B	WNT signaling	3,39E-11	7	58
HNSC	CHIR-99021	GSK3A, GSK3B	WNT signaling	1,24E-10	6	42
OV	CHIR-99021	GSK3A, GSK3B	WNT signaling	1,24E-10	5	22
PAAD	SB590885	BRAF	ERK MAPK signaling	4,66E-10	5	25
LUAD	VX-702	p38	JNK and p38 signaling	1,81E-09	6	57
LIHC	CHIR-99021	GSK3A, GSK3B	WNT signaling	3,27E-09	4	17
KIRC	SB590885	BRAF	ERK MAPK signaling	4,86E-08	4	28
LUAD	SB590885	BRAF	ERK MAPK signaling	4,86E-08	5	57
NSCLC	CHIR-99021	GSK3A, GSK3B	WNT signaling	6,27E-08	4	30
COREAD	PAC-1	Procaspase-3, Procaspase-7	Apoptosis regulation	3,43E-07	4	37
MESO	SB590885	BRAF	ERK MAPK signaling	5,91E-07	3	19
COREAD	VX-702	p38	JNK and p38 signaling	9,23E-07	4	46
COREAD	UNC0642	G9a(EHMT2), GLP(EHMT1)	Chromatin histone methylation	1,63E-06	4	46
ESCA	SB590885	BRAF	ERK MAPK signaling	5,05E-06	3	33
NSCLC	Axitinib	PDGFR, KIT, VEGFR	RTK signaling	5,75E-06	3	30
KIRC	Lenalidomide	CRBN	Protein stability and degradation	7,02E-06	3	30
NSCLC	Veliparib	PARP1, PARP2	Genome integrity	8,62E-06	3	28

GBM	UNC0642	G9a(EHMT2), GLP(EHMT1)	Chromatin histone methylation	1,07E-05	3	34
NSCLC	Vismodegib	SMO	Other	1,4E-05	3	30
GBM	Bosutinib	SRC, ABL, TEC	Other, kinases	1,69E-05	3	35
HNSC	IPA-3	PAK1	Cytoskeleton	2,03E-05	3	42
HNSC	PLX-4720	BRAF	ERK MAPK signaling	2,92E-05	3	40
HNSC	TL-2-105	not defined	Other	3,01E-05	3	39
COREAD	Vismodegib	SMO	Other	3,08E-05	3	43
BRCA	Dasatinib	ABL, SRC, Ephrins, PDGFR, KIT	Other, kinases	3,36E-05	3	45
BRCA	CHIR-99021	GSK3A, GSK3B	WNT signaling	4,01E-05	3	47
SKCM	Bexarotene	Retinioic X receptor (RXR) agonist	Other	4,82E-05	3	52
BRCA	NSC-87877	SHP-1 (PTPN6), SHP-2 (PTPN11)	Other	8,73E-05	3	48
LUAD	ZM447439	AURKA, AURKB	Mitosis	0,000114	3	61

Table 6: Enriched combinations of drug and tissue type in the subset of high slopes.

A total of 29 combinations were enriched based on FDR $= 5\%$. Only drug-tissue combinations with at least 2 drug responses were considered. The number of high-slopes and drug responses available in the drug-tissue combinations was assessed independent of mutation events.

3.2.8 Correlation between cell proliferation and drug response

In the previous sections, drugs with increasing cell viability markers were highlighted. Notably, the ultimate motivation of this work was to hypothesize a new cancer therapy based on drugs which provoke tumour growth and synergistically combine them with compounds that effectively control this fast proliferation.

For this, I assessed the correlation between the cell growth rate and the drug response (details in chapter 2.3.6), and identified negatively correlated pairs. Initially, preliminary analysis on all 499 compounds available at the GSDC screen revealed 471 negatively correlated pairs, in which several drugs are DNA damaging agents which target DNA replication pathways (Appendix Table 17). These findings are concordant with clinical chemotherapies where DNA damaging agents are used in combination with fast proliferating cells.

Subsequently, given the interest in investigating the drug CHIR-99021 and the alternations cnaLUAD3 and cnaLUAD27 in lung adenocarcinoma, I performed an additional correlation analysis stratified by tissue type and focused the results in lung adenocarcinoma (Table 7). Based on these findings and integrating them with the drug combination database DrugCombDB (Liu et al. 2021), my colleague Ginte Kutkaite selected the drugs Paclitaxel, 5-Fluorouracil and Docetaxel as potential candidates to screen in combination with CHIR-99021.

Tissue	Drug	Drug Target	Target Pathway	R
LUAD	Luminespib	HSP90	Protein stability and degradation	-0.668691
LUAD	Seliciclib	CDK2, CDK7, CDK9	Cell cycle	-0.639655
LUAD	ERK_6604	ERK1,ERK2	ERK MAPK signaling	-0.638096
LUAD	Paclitaxel	Microtubule stabiliser	Mitosis	-0.633072
LUAD	5-Fluorouracil	Antimetabolite (DNA & RNA)	Other	-0.623237
LUAD	Docetaxel	Microtubule stabiliser	Mitosis	-0.609848
LUAD	PRT062607	SYK	Other, kinases	-0.600892
LUAD	BMS-345541	IKK1, IKK2	Other, kinases	-0.595863
LUAD	BMS-345541	IKK-1, IKK-2	Other	-0.595863
LUAD	Pyridostatin	G-quadruplex stabiliser	DNA replication	-0.592316

Table 7: Correlation between cell growth rate and drug response in lung adenocarcinoma (LUAD).

Top 10 negatively correlated drug response and cell growth rate $R = \text{cor}(\text{IC}_{50}, \text{Rate}_{\text{cell growth}})$ in lung adenocarcinoma. Remarkably, interesting results are cases of negative IC$_{50}$ value combined with positive cell growth.

3.2.9 Wet lab experiments of potential ICV markers

Based on the previously described analysis, the compound CHIR-99021 systematically revealed enrichment of drug responses with increasing viability markers. Moreover, ATD combinations involving the amplifications cnaLUAD3 and cnaLUAD27 in lung adenocarcinoma revealed promising results to be furtherly investigated.

Together with the groups of Prof. Dr. Daniel Krappmann and Dr. Kamyar Hadian, validation experiments of specific compounds and cell lines were performed.

Specifically, we tested three GSK-3 inhibitors, namely CHIR-99021, SB216763 and 9-ING-41 (Duda et al. 2020), and eight cell lines. The cell lines were determined based on their drug response with CHIR-99021 (Figure 36). In detail, we selected two cell lines which presented increasing markers in both cnaLUAD3 and cnaLUAD27, only in cnaLUAD3 or cnaLUAD27 and two non-responders without biomarker (Table 8).

COSMIC ID	Cell line	Drug	Linear slope	alteration
908472	NCI-H1573	CHIR-99021	0.0017625	no biomarker
687807	NCI-H1838	CHIR-99021	-0.0162034	no biomarker
753600	NCI-H1563	CHIR-99021	0.2393604	LUAD3
909721	SK-LU-1	CHIR-99021	0.1756879	LUAD3
907786	LC-2-ad	CHIR-99021	0.1960237	LUAD27
908463	NCI-H1793	CHIR-99021	0.1699917	LUAD27
1240145	HCC-44	CHIR-99021	0.2810958	LUAD3 & LUAD27
687820	NCI-H2347	CHIR-99021	0.1838626	LUAD3 & LUAD27

Table 8: Overview of the drug response of the eight cell lines selected for validation against CHIR-99021.

COSMIC ID is the GDSC identifier of each cell line. Linear slope of the curve fitting model of the drug response of drug CHIR-99021 and specific cell line. In total 8 cell lines were selected, from which 2 had no biomarker and the rest contained the amplification cnaLUAD3, cnaLUAD27 or both.

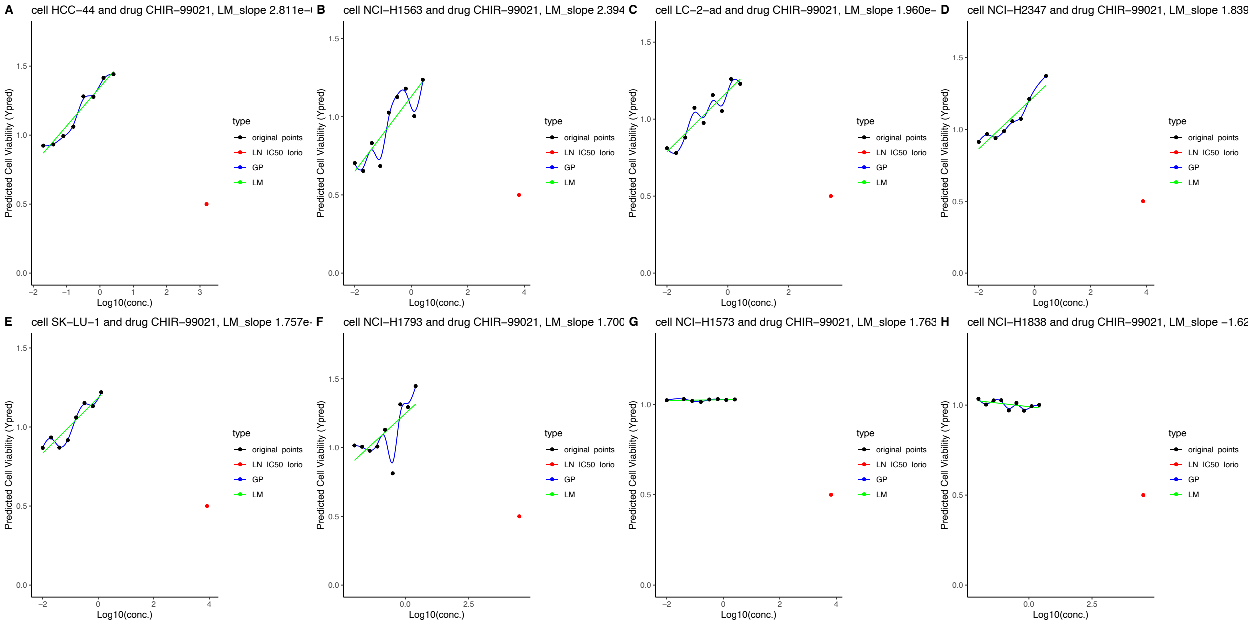

Figure 36: Drug responses of CHIR-99021 with 8 cell lines selected for validation screen.

(A)-(F) Drug responses with increased cell viability and containing amplifications cnaLUAD3 and/or cnaLUAD27. **(G)-(H)** In contrast, drug responses with approximate zero cell viability.

With the intention of replicating the results obtained by the GDSC screen, the experiments were designed with comparable drug concentrations (total of 20 concentrations, Appendix Table 18), 2 replicates of blanks, 3 replicates of each GSK-3 target and control (DMSO) compounds. Similarly to the GDSC screen, cells were treated for 72 hours and cell viabilities were determined using CellTiter-Glo.

Cell viability curves were inferred for all combinations of drug and cell lines. As expected, at high concentration ranges (around above 10 μM), toxic effects take part and the drug response is reduced to zero (Appendix Figure 48).

From all experimental curves, the drug responses involving the cell lines SK-LU-1 and NCI-H1793 showed the most promising responses. However, focusing on the drug concentration range where CHIR-99021 depicts increase proliferation in the GDSC screen (i.e., between 0,01 and 2,5 μM), only a marginal increase in the cell viability of these two cells was observed (Figure 36).

In addition to the validation experiments, the drug response between CHIR-99021 and the cell lines SK-LU-1 and NCI-H1793 was explored in public domain datasets. Namely, the CTRP version 2 drug screen (Seashore-Ludlow et al., 2015) and a recent study of cell cultured analysis of NSCLC cell lines clinics (Nair et al. 2023), built similar experimental designs and investigated the response of these two cells upon treatment of CHIR-99021. Remarkably, both works revealed an increase in the viability of both cells around the same concentration range as seen in the GDSC drug screen (Figure 38).

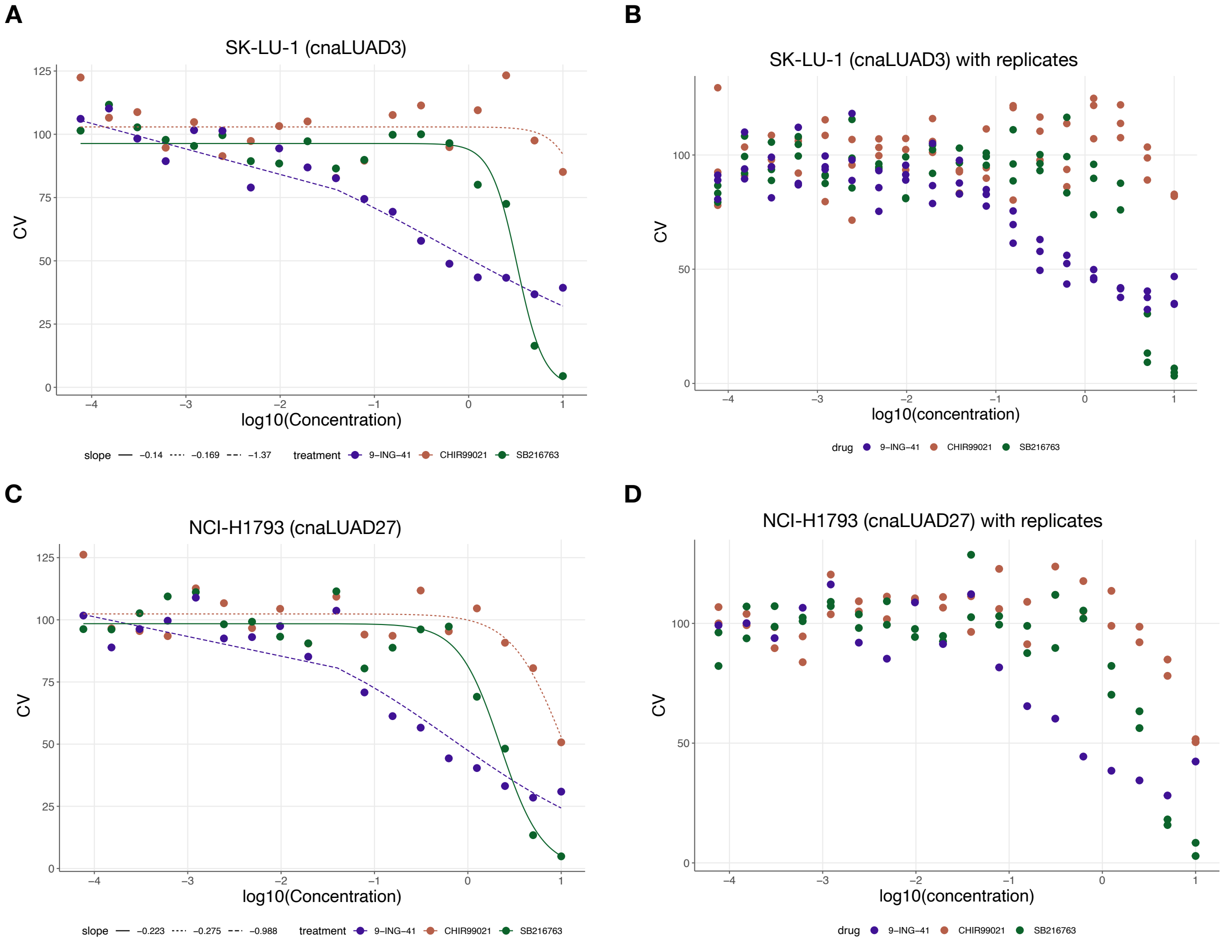

Figure 37: Most promising experimental cell viability results for validated cell lines and drugs.

Cell viabilities of GSK-3 inhibitors CHIR-99021, SB216762 and 9-ING-41 against cell lines **(A)-(B)** SK-LU-1 (with amplification cnaLUAD3) and **(C)-(D)** NCI-H1793 (with amplification cnaLUAD27). In the first column, the plots depict the mean cell viability values over all the replicates and a sigmoid curve fit using the nls() function from the stats R package. The second column illustrates the raw cell viability values of the three replicates at each titration point.

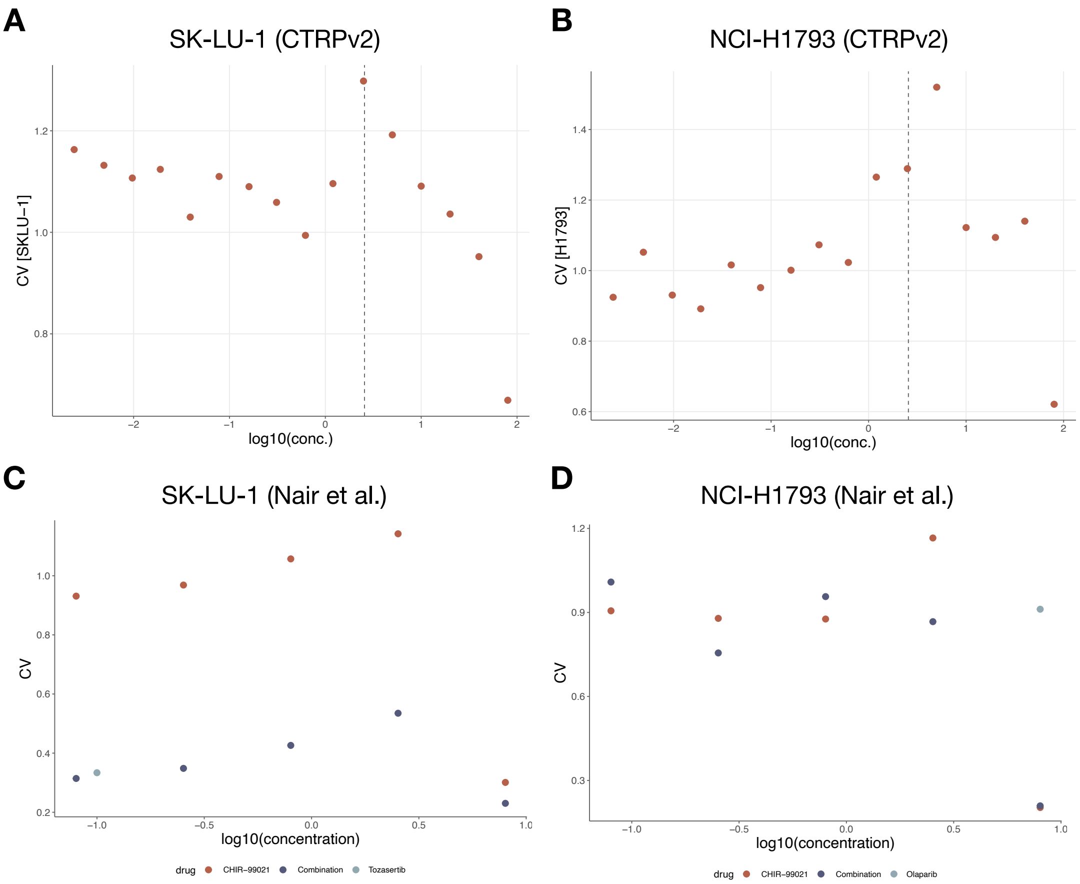

Figure 38: Evidence of increased viability of CHIR-99021 with SK-LU-1 and NCI-H1793 cell lines.

Cell viabilities observed in CTRP drug screen for cell lines **(A)** SK-LU-1 and **(B)** NCI-H1793. Here the vertical dot line indicates the maximum concentration tested in the GDSC screen and is clearly visible an increase in the viability of both cell lines around this concentration point. Furthermore, cell viabilities of CHIR-99021 in combination with **(C)** Tozasertib in cell SK-LU-1 and **(D)** Olaparib in cell NCI-H1793 from the (Nair et al. 2023) study.
The depicted visualizations were created by my colleague Ginte Kutkaite.

3.3 wPPI network for tissue specific drug response modelling

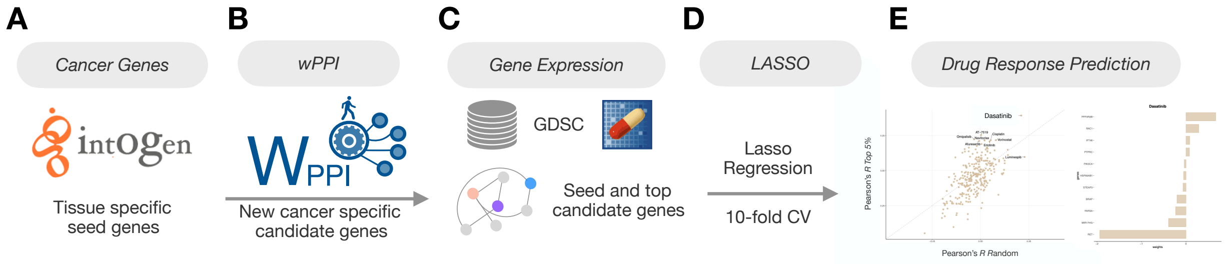

Figure 39: Drug response prediction pipeline using the wPPI package.

(A) Selection of tissue specific genes using IntOGen. These genes serve as seed genes for the wPPI framework. **(B)** Implementation of wPPI framework to the seed genes inferred by IntOGen. Identification of a set of new cancer specific candidate genes in close functional proximity to the seed genes. **(C)** Leverage of the gene expression profiles of the selected seed and wPPI candidate genes to predict drug response (represented by IC_{50} values) available in the GDSC project. **(D)** Application of LASSO regression to model drug response. Performance assessed by Pearson correlation coefficient and cross-validation. **(E)** Identification of drugs with high predictive power and analysis of genes and pathways involved.

3.3.1 Input features for gene prioritisation framework wPPI

IntOGen seed genes

The inference of new candidate cancer genes is correlated to the given seed genes (Figure 14 and Figure 39 A). Hence, in order to ensure an appropriate selection of cancer specific genes, the input seed genes were determined using the Integrative Onco Genomics (IntOGen) database which contains an extensive repository of cancer driver genes (Martínez-Jiménez et al., 2020). Specifically, this dataset comprised 66 different cancer types and 568 unique driver genes (repository downloaded in February 2020).

Given the availability of cells and number of mutations per cancer type (Figure 9), as well as, the phenotypic nature of the cancer tissues, I focused my analysis in breast adenocarcinoma (BRCA). This derived a total of 99 seed genes to be considered in the analysis (Appendix Table 19).

Ontology databases

The wPPI framework is a network-based gene prioritisation network which functionally scores candidate genes based on network topology and shared ontology annotations (Figure 14). In detail, phenotypic and genomic specific information is incorporated through the Human Phenotype Ontology (HPO) and Gene Ontology (GO) databases, respectively.

For the HPO database, I inferred together with my colleague Daniel Garger tissue-specific phenotypic annotations. In total, we defined 3 unique HPO annotations for BRCA. Furthermore, my colleague Phong Nguyen extracted genes involved in these tissue-specific HPO annotations. As a result, 63 unique gene symbols were found to be annotated in the BRCA-specific HPO annotations (Appendix Table 20).

When leveraging wPPI, it is possible to include both, one or none of these ontology databases in the prioritisation of new genes. In order to define the best model, I estimated the drug response with several combinations of HPO and GO databases:

- without HPO or GO annotations (scores estimated purely with topological information);
- with only BRCA-specific HPO annotations (Appendix Table 20) and all GO database;
- with all HPO and GO annotations.

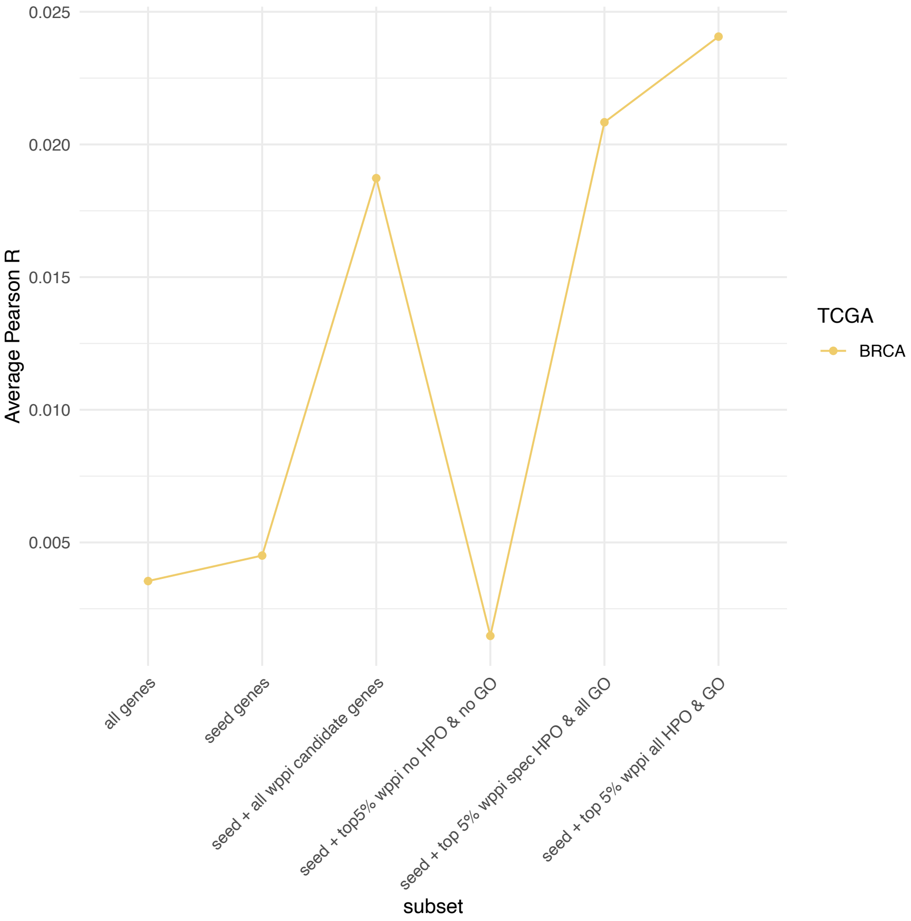

Figure 40: Average Pearson correlation for drug response based on gene sets with different combinations of HPO and GO annotations.

Six distinct models to predict drug response in BRCA: (1) all genes available in the gene expression; (2) seed genes from IntOGen (Appendix Table 19); (3) seed genes and all candidate genes inferred by wPPI with all HPO and GO annotations; (4) seed genes and top 5 % candidate genes from wPPI without HPO and GO annotations; (5) seed genes and top 5% candidate genes from wPPI with only BRCA specific HPO annotations (Appendix Table 20); (6) seed genes and top 5% candidate genes from wPPI with all HPO annotations. Results were averaged over all the drugs in the dataset (= 303 drugs).

Based on the performances (Figure 40), for the following analysis, I only considered models based on all HPO and GO annotations.

Parameters of the search algorithm

Once functional scores are calculated to all candidate genes, wPPI runs the Random Walk with Restart algorithm to search for functionally closely related genes in respect each seed gene (Figure 14). In order to establish "how far" the algorithm can investigate, this is modelled by the restart probability parameter r, which can take values between 0 and 1.

Similarly to the previous analysis, I investigated the most optimal value of r by modelling the drug response in BRCA over several values of r. Notably, the results suggest the restart probability parameter does not have a high influence on the model's performance (Figure 41). Notwithstanding, I considered $r = 0.2$ for the follow-up analysis since it presented the highest median across all drug responses.

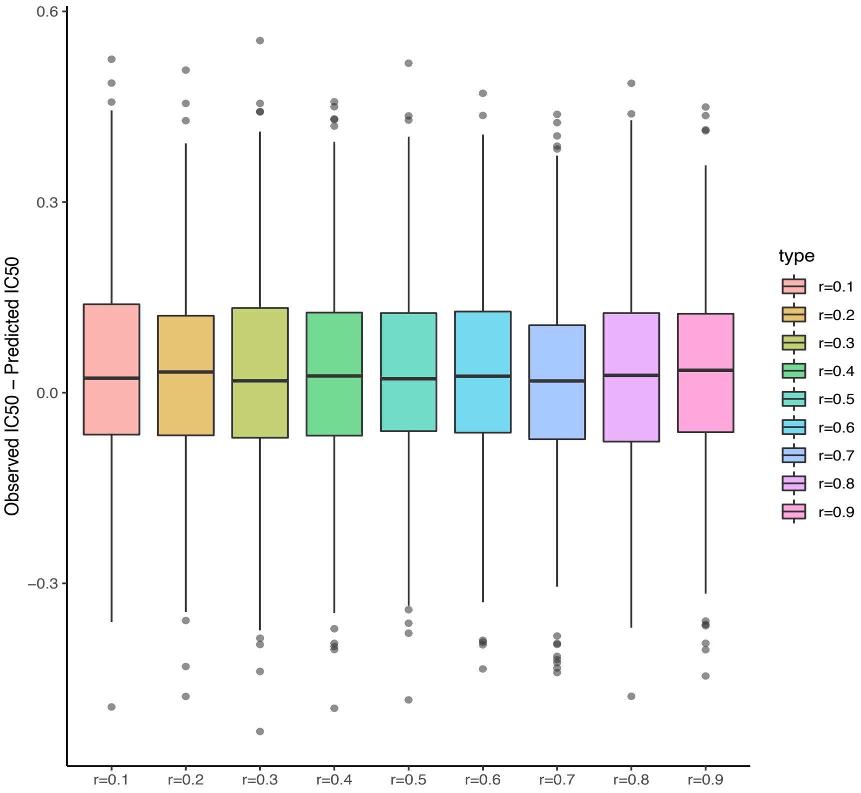

Figure 41: Model performance with several restart probabilities.

Pearson correlation between predicted and observed drug response (represented with IC_{50} values) for BRCA with restart probabilities $r = 0.1 - 0.9$.

3.3.2 Identification of new cancer specific candidate genes with wPPI

Taking the input seed genes (chapter 3.3.1), a PPI network was constructed with the seed genes and their direct interactors (i.e., first order degree neighbors, candidate genes) (Figure 14 B and Figure 39 B). For this, I used the Omnipath PPI network, which comprised 42,541 interactions and 7,463 unique proteins (Türei et al., 2021). Based on the availability of genes in the Omnipath PPI database (download version of 2020), the amount of candidate genes was 2,931 genes for BRCA.

Subsequently, given the specifics for the ontology databases and RWR algorithm previously discussed (chapter 3.3.1), a functional ranking of the cancer-specific candidate genes was estimated (Figure 14 C-E). Focusing on the top 5% ranked candidate genes, a total of 146 genes were discovered (Figure 42 and Appendix Table 21).

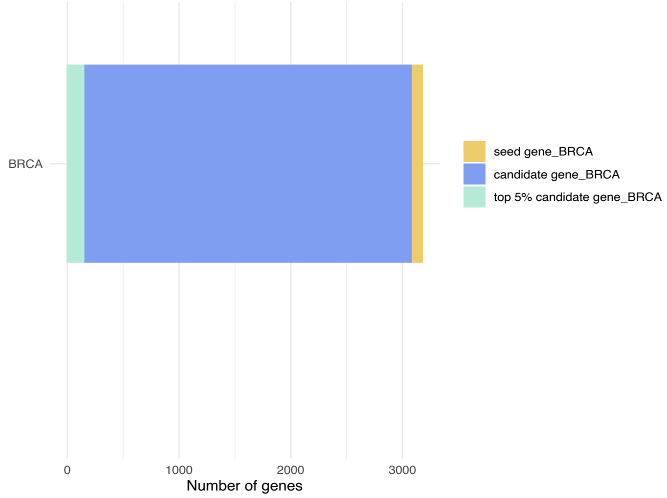

Figure 42: Overview of number of seed and candidate genes in BRCA.

In addition, I performed pathway enrichment analysis of the top 5% ranked genes in BRCA using the ViSEAGO package with default parameters (Brionne et al. 2019). Leveraging a significance threshold of $\alpha = 5\%$, several biological processes related with DNA damage response, cell development, proliferation and division were discovered (Figure 43). Interestingly, pathways involving ERRB4, which is typically implicated in breast cancer proliferation and survival of cancer cells, were significantly enriched. Notably, these findings highlight the performance of wPPI to robustly select new cancer-specific candidate genes.

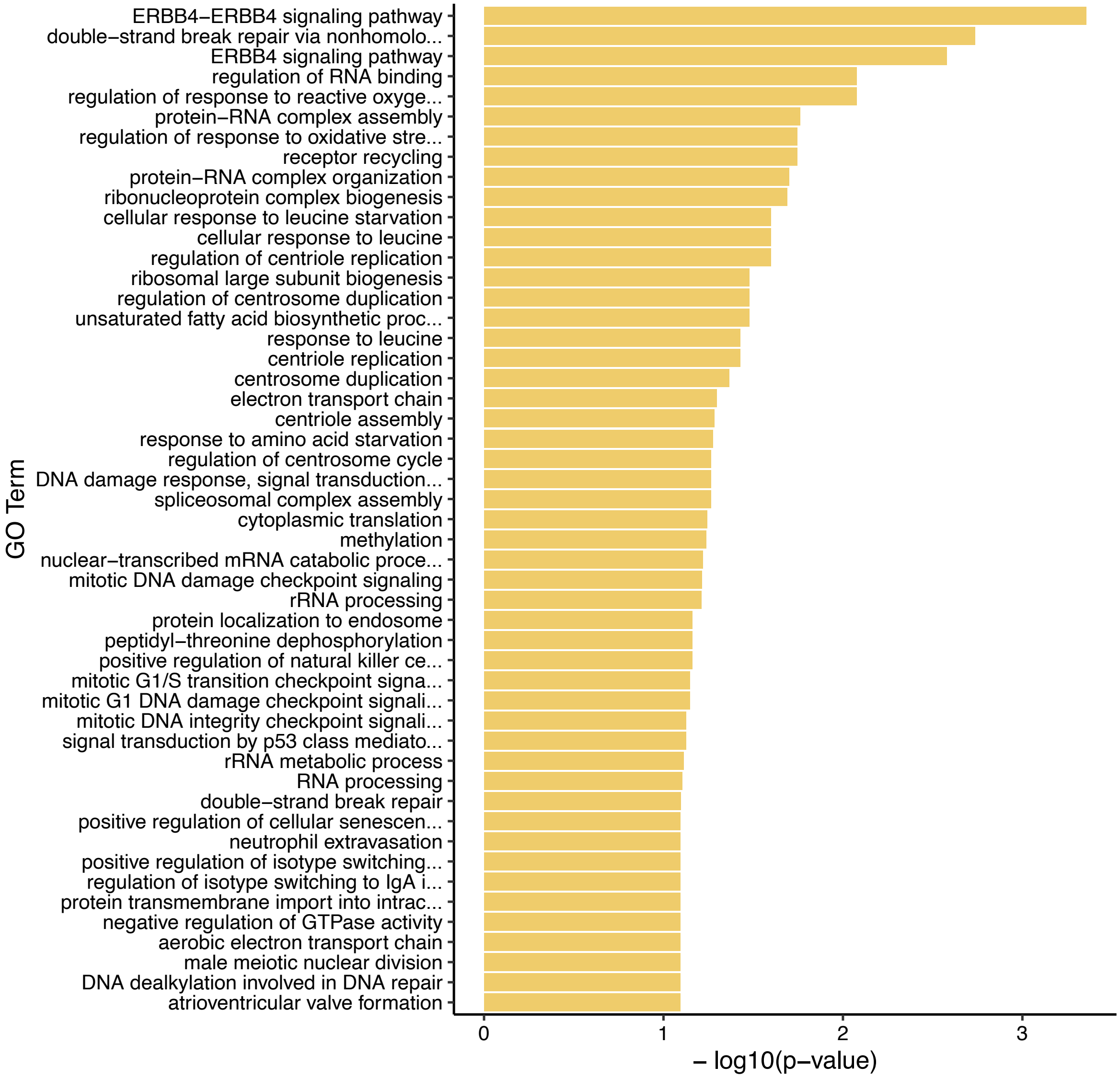

Figure 43: GO-term enrichment analysis of top 5% candidate genes from wPPI.

The top 50 enriched pathways for BRCA. The biological processes are ordered according to their p-value significance.

3.3.3 Drug response prediction based on seed and wPPI genes

Gene expression profiles of the seed genes and new candidate genes identified by wPPI were used as input features to train a LASSO machine learning (ML) model and predict the drug response in terms of IC_{50} values (Figure 39 D). For this task, I used datasets available in the GDSC project, namely gene expression of 17,419 genes and 1,018 cell lines and drug response measurements tested on 303 drugs. Since the analysis focused on breast tissue, the gene expression and drug responses were reduced to measurements in 49 cell lines (Figure 44).

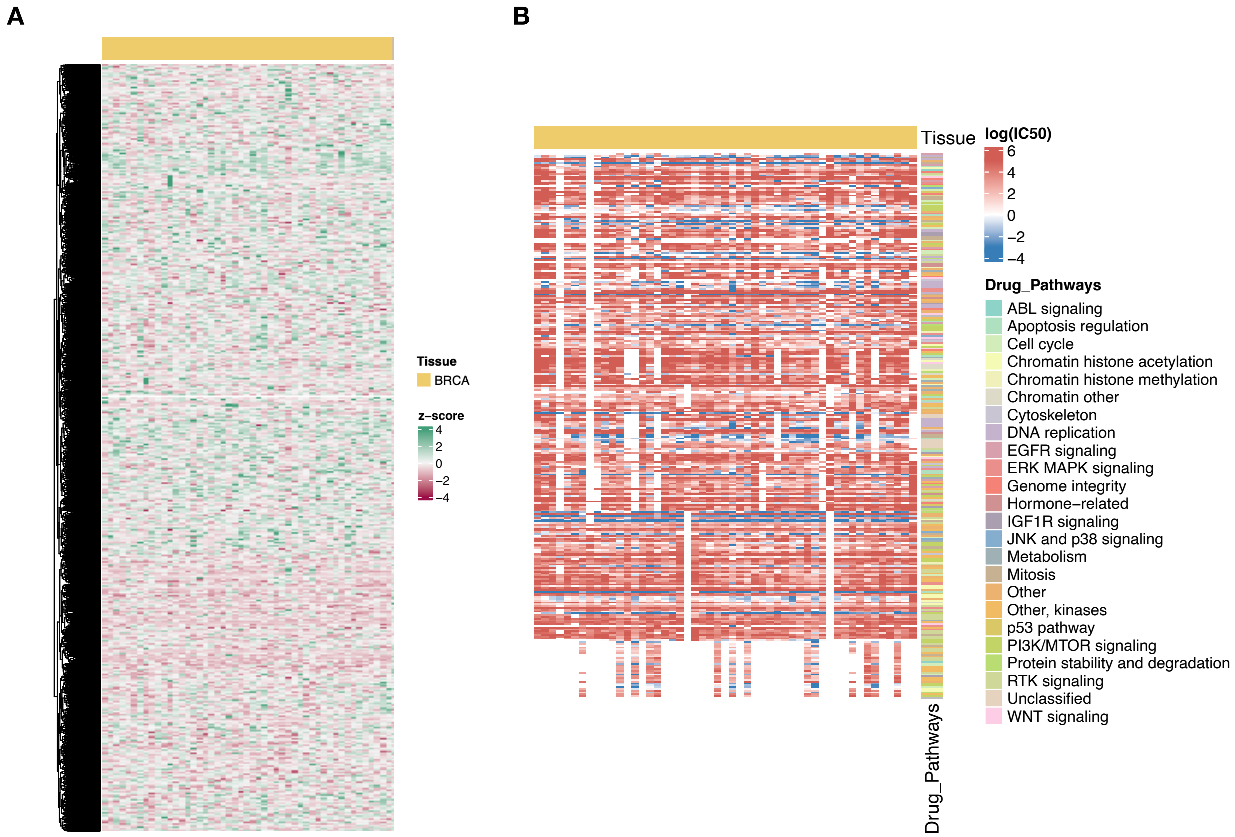

Figure 44: Heatmaps of gene expression and drug response in BRCA.

(A) Gene expression profiles of 17,419 genes and 49 cell lines. Expression values represented with a z-score transformation. **(B)** Drug response of 303 compounds and 49 cell lines. Responses represented by IC_{50} values in log form. Unknow drug response values were mapped to zero values.

In order to explore the performance of wPPI as a feature selection framework, I considered several sets of genes. Namely, the whole genome (all the genes in the gene expression, i.e., 17,419 genes), the tissue-specific seed genes from IntOGen (S_1; 96 genes), all the wPPI candidate genes (S_2; the first order degree neighbours of the seed genes, i.e., 2,931 genes) and the top 5% ranked wPPI genes (S_3; 146 genes).

First, I cross-compared the ML models of the gene subsets (S_1, S_2 and S_3) against the whole genome, and reported the average Pearson correlation R between the observed and predicted IC_{50} values over all drugs (Figure 40 and Table 9). As already observed in the previous chapter 3.3.1, the models' performances improve when leveraging a tissue-guided feature selection, with the highest averaged performance achieved with the top 5% ranked wPPI genes ($R_{S_3} = 2{,}99 \times 10^{-2}$).

Gene set	Whole Genome	Seed Genes (S1)	Seed Genes + All wPPI Genes (S2)	Seed Genes + Top 5% wPPI Genes (S3)
Pearson correlation *R*	$3{,}54 \times 10^{-3}$	$4{,}25 \times 10^{-3}$	$1{,}87 \times 10^{-2}$	$2{,}99 \times 10^{-2}$

Table 9: Averaged Pearson correlation between observed and predicted drug response of several gene sets.

Moreover, I created random models composed by sets of randomly selected genes with the same size as the gene subsets (S_1, S_2 and S_3). The performance of these random models was assessed over 500 iterations (Table 10) and cross-compared with the Pearson correlation obtained for the gene subsets S_1, S_2 and S_3 (Table 9).

Gene set	Random # S1	Random # S2	Random # S3
Pearson correlation *R*	$-1{,}30 \times 10^{-3}$	$4{,}02 \times 10^{-3}$	$-7{,}52 \times 10^{-4}$
Confidence Interval	$[-2{,}68 \times 10^{-3}, 7{,}77 \times 10^{-5}]$	$[2{,}57 \times 10^{-3}, 5{,}47 \times 10^{-3}]$	$[-2{,}12 \times 10^{-3}, 6{,}19 \times 10^{-4}]$

Table 10: Averaged bootstrapped Pearson correlation and confidence interval of random models with the same number of genes.

Briefly, the random models performed significantly worse in comparison with the models built with cancer-specific genes from IntOGen and wPPI (Table 9 and Table 10). Specifically, the random models build with a smaller input feature set (S_1 and S_3, with 96 and 146 genes, respectively), presented a negative Pearson correlation between the predicted and the observed drug response. This finding suggests inconsistency between the gene expression of the randomly selected genes with the observed IC_{50} values, and highlights that a knowledge-guided selection of a small subset of genes is required to provide accurate predictions.

In contrast, the random model formed with the biggest gene feature set (S_2, 2,931 genes; Table 10) revealed a comparable performance with the model constructed with all the first order degree PPI neighbours of the seed genes (Table 9). This result is probably driven by the fact that the set of genes is large enough to encompass genes which are drug targets.

As a next step, I comprehensively investigated the drug response prediction for each one of these gene subsets (S_1, S_2 and S_3). Specifically, I analysed the performance of each model against the random model and across all the compounds available (Figure 45).

Looking in detail to the models built with the seed genes and the top 5% ranked wPPI genes (Figure 45 C), and focusing on the drugs with the highest Pearson correlation (Figure 46 and Appendix Table 22) several drugs targeting the PI3K/Akt/mTOR, IGF1R and ERK MAPK signalling pathways were detected.

For instance, Tozasertib is an Aurora kinase inhibitor and regulates mitosis progression (Wang et al. 2024). This kinase has been linked to tumorigenesis in several cancer types and high levels of it have been reported in breast cancer patients (Yamamoto et al. 2013; Aradottir et al. 2015). Tozasertib revealed the highest performance across all tested compounds and gene subsets. Specifically, a Pearson correlation $R_{S_3} = 0{,}51$ was achieved in the ML model built with the seed and the top 5% wPPI genes (against $R_{\text{whole genome}} = 0{,}39$; $R_{S_1} = 0{,}33$; $R_{S_2} = 0{,}38$ and $R_{\text{random}} = 0{,}25$). The model was built based on seed genes ERBB2 and ERBB3 (notorious breast cancer drivers) and wPPI IGFBP4 gene (Figure 46 A). IGFBP4 is an insulin-like growth factor and responsible of transporting and regulating the hormone IGF-1. IGF-1 can activate PI3K/Akt/mTOR and MAPK/ERK pathways and overexpression of IGF-1 has been correlated with tumour proliferation in breast cancer (Christopoulos et al. 2015; Ekyalongo and Yee 2017; Wang et al. 2019).

AZD7762 targets checkpoint kinases 1 and 2 (CHEK1 and CHEK2), and has been described as a potential tool for combination therapies in breast cancer (Zabludoff et al. 2008; Park et al. 2016). Notably, this compound revealed high Pearson correlation across all the tested subsets ($R_{\text{whole genome}} = 0{,}46$; $R_{S_1} = 0{,}40$; $R_{S_2} = 0{,}49$ and $R_{S_3} = 0{,}43$), with the best model being with the seed and all the wPPI candidate genes. Deeping into the genes leveraged in prediction (Figure 46 B), the majority were seed genes and were also found in the ML model of Tozasertib, namely CLTC, ERBB3 and VRK2. In detail, CLTC expression has been identified as a marker for several cancer types, and linked with tumor growth and proliferation of breast cancer cells (Shijie et al. 2021; Jiao et al. 2013). Moreover, VRK2 expression was negatively correlated with the drug prediction and breast cancer drivers ERBB2 and ERBB3, which is consistent with literature findings (Fernández et al. 2010).

Other two compounds of interest were Vincristine and Phenformin (Figure 46 C-D), with ML predictive models mainly built with wPPI selected genes and significantly better performance with the top 5% wPPI genes (S_3) than the one obtained with the other subsets.

Vincristine is a chemotherapy drug currently used in clinics to advance-stage breast cancer patients (Byrne 1976; Velho 2012). Its prediction based on the top 5% wPPI was positively correlated with the observed drug response ($R_{S_3} = 0{,}21$), whilst the models built with the other gene subsets all revealed negative correlation with the real drug response ($R_{\text{whole genome}} = -0{,}071$; $R_{S_1} = -0{,}051$; $R_{S_2} = -0{,}015$ and $R_{\text{random}} = -0{,}039$). The wPPI genes used in the prediction with the highest absolute weights were AHSP, GML and ISCU (Figure 46 C). Currently, there is no clear evidence of a relation linking the genes AHSP and GML with breast cancer. However, AHSP gene has been reported to be directly regulated and activated by the seed gene GATA-1 (Gallagher et al. 2005; Lai et al. 2005), and GML is linked to apoptotic pathways (Kagawa et al. 1997; Kimura et al. 1997). On the other hand, the ISCU gene, which negatively contributed to the prediction, has been reported to be downregulated in a variety of cancer tissues, including breast. This gene is directly regulated by p53, surrounded by SNPs associated to higher risk of developing breast cancer and its suppression has been linked to a worst prognosis of cancer patients (Petronek et al. 2021; Favaro et al. 2010; Degtyareva et al. 2020; Funauchi et al. 2015).

Lastly, Phenformin inhibits melanoma cell growth, drives alteration in the cell cycle and has been recently suggested to be leveraged as an anticancer drug for breast cancer (García Rubiño et al. 2019; Guo et al. 2017; Liu et al. 2015). Similarly to Vincristine, this compound reported a significantly better prediction with the gene subset composed by the top 5% wPPI genes ($R_{S_3} = 0{,}32$), in comparison with the rest of the subsets ($R_{\text{whole genome}} = 0{,}001$; $R_{S_1} = 0{,}099$; $R_{S_2} = -0{,}015$ and $R_{\text{random}} = 0{,}038$). The top gene weights were derived by the wPPI genes B4GALT1 and ULK3 (Figure 46 D). B4GALT1 expression has been indicated to be directly estrogen-induced in breast cancer cells and to influence the proliferation of tumour cells (Villegas-Comonfort et al. 2012; Choi et al. 2012). In contrast, the expression of ULK3, which presented a negative weight in the prediction, has been negatively correlated with the formation of breast cancer and is directly involved in the activation of GLI1 and GLI2, which have been pointed as potential breast cancer prognostic markers (Goruppi et al. 2017; Zhang et al. 2023; Im et al. 2013).

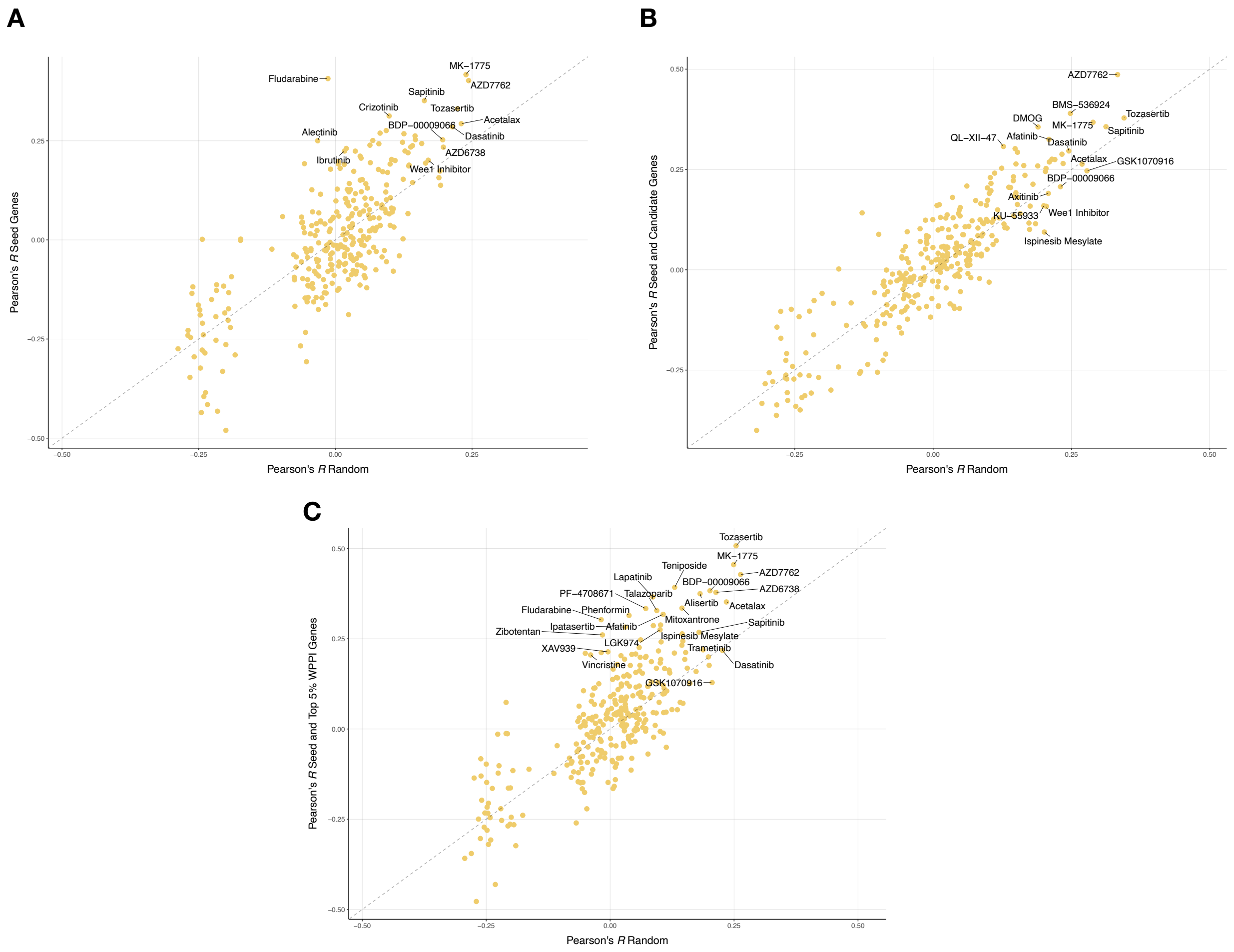

Figure 45: Performance of ML models with seed and wPPI candidate genes against random models across all drugs.

Performance of models built with specific combinations of seed and wPPI candidate genes against random models constructed with the same amount of input features. Models with **(A)** seed genes, **(B)** seed and all wPPI candidate genes and **(C)** seed and top 5% ranked wPPI candidate genes. A total of 303 drugs are illustrated and compounds with higher Pearson correlation than random are highlighted. The x-axis and y-axis represent the Pearson correlation of the seed plus wPPI candidate genes combination and of the randomly selected genes, respectively.

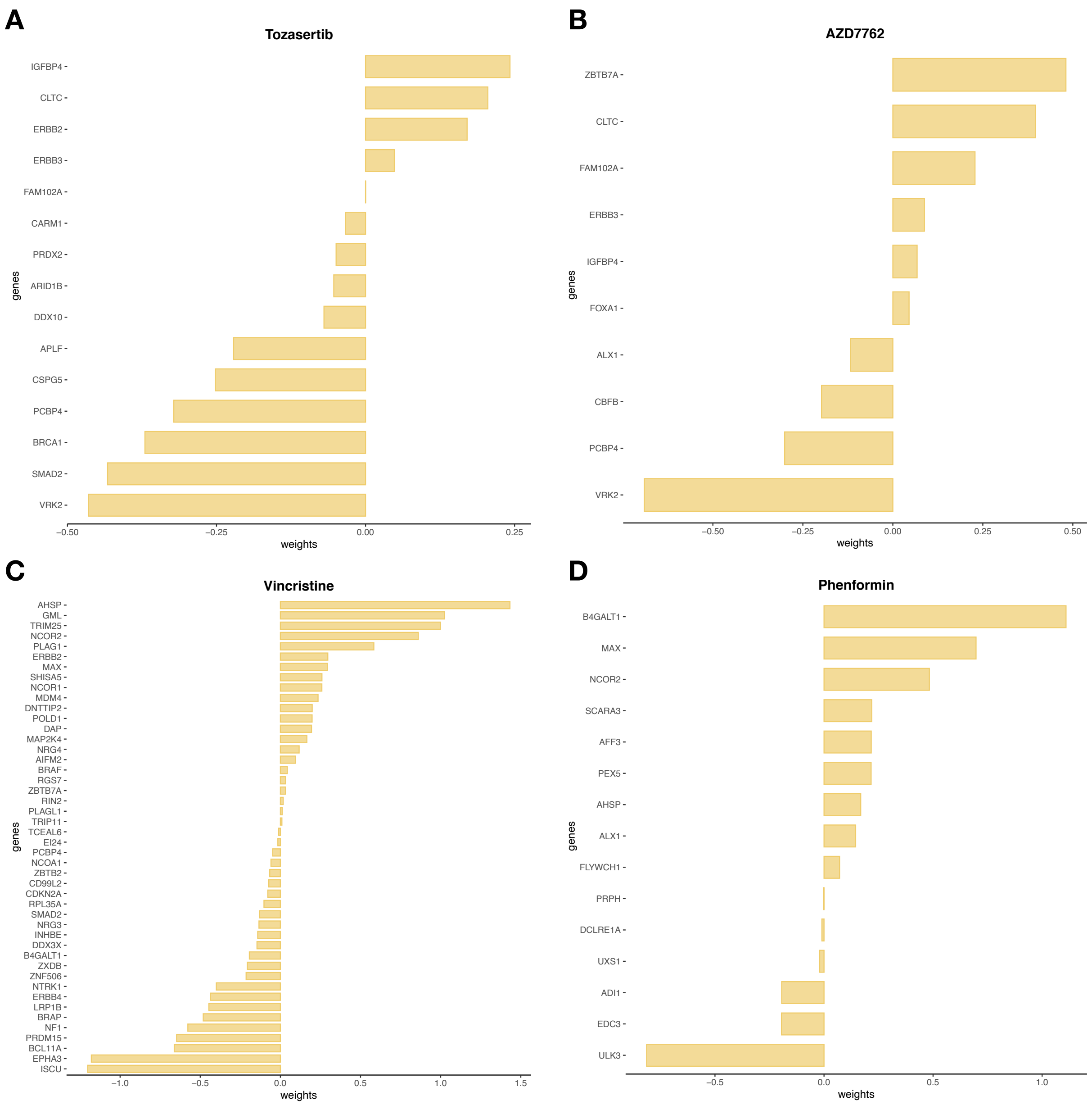

Figure 46: Illustration of drugs with the high Pearson correlation coefficient and their respective gene weights used in the modelling.

Weight distribution of the genes used in the drug response modelling of drugs **(A)** Tozasertib, **(B)** AZD7762, **(C)** Vinocristine and **(D)** Phenformin.

3.3.4 Comparison with existing gene prioritisation tool

In order to benchmark the wPPI as a suitable gene prioritisation tool of cancer specific genes, I leveraged the GeneFriends tool to select breast cancer specific genes using the same seed genes used for wPPI (Raina et al., 2023).

Employing the GeneFriends with the default parameters, a total of 474 candidate genes were discovered, where 103 of these genes were in common with the candidate genes from wPPI.

The Lasso regression model was built around the selected 474 genes, and a mean Pearson correlation of $R_{\text{GeneFriends}} = 3{,}59 \times 10^{-3}$ was estimated across all the drugs. Notably, this performance is significantly worse than the ones obtained with the sets of wPPI candidate genes (Table 9).

In addition, I also investigated the pathways involved with the candidate genes selected by GeneFriends (Figure 47). The resulting enriched pathways were mostly related with metabolic processes and regulation of gene expression, which are not specific to breast cancer.

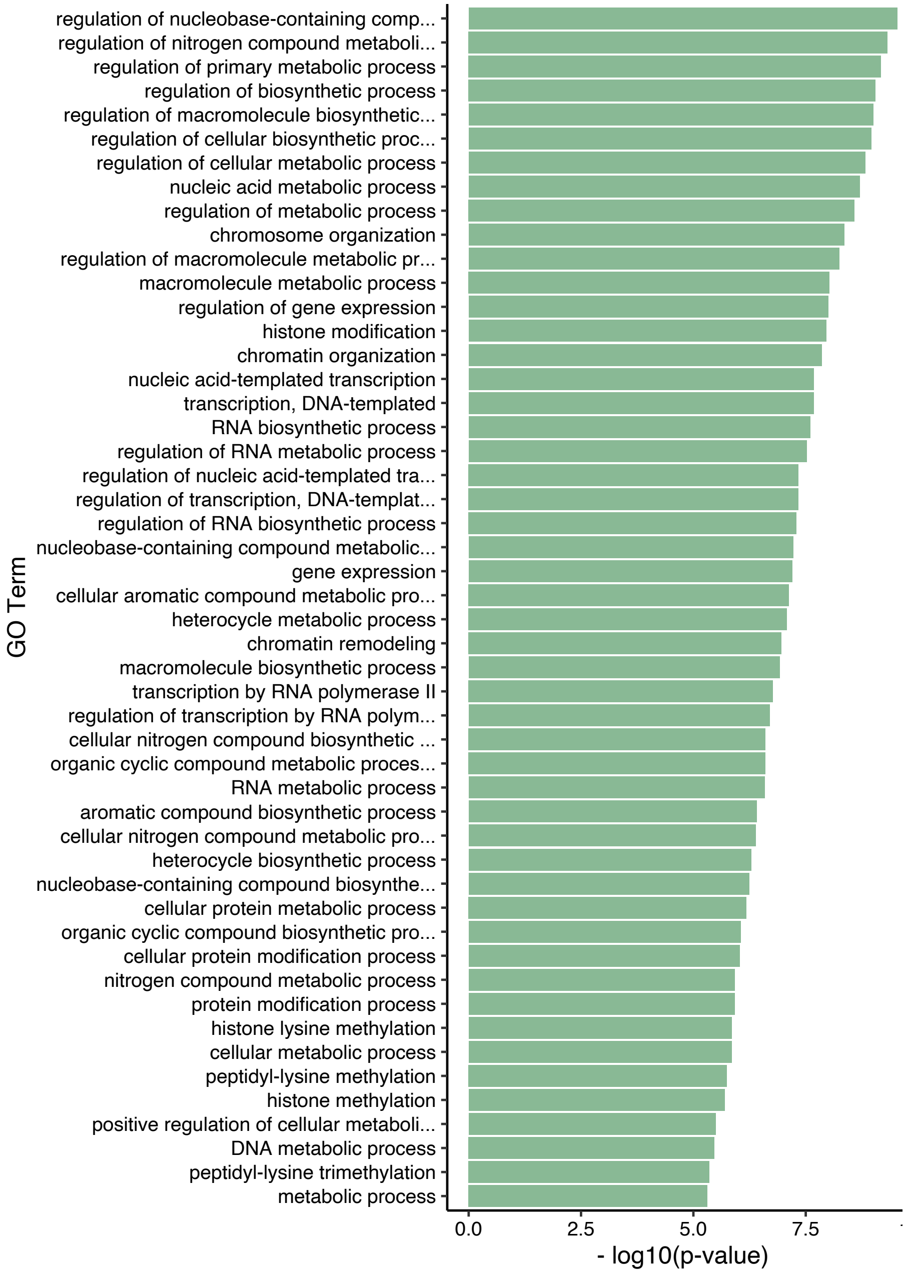

Figure 47: Gene Ontology enrichment analysis of candidate genes from GeneFriends.

4. Discussion

4.1 Summary

Over the last decade, cancer research has made significant improvements driven by developments in genomic sequencing techniques, formulation of synergistic combination therapies and creation of adaptive clinical trial designs (Jiang et al. 2022). These advances, in combination with the creation of specialized bioinformatic tools, have paved the way for the creation of personalised cancer treatment and improvements in the discovery of drugs and predictive biomarkers.

Cancer cells can be explored using large-scale high-throughput screens (HTS), such as the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) and the Genomics of Drug Sensitivity in Cancer (GDSC) (Garnett et al., 2012; Iorio et al., 2016) projects. These high-throughput technologies facilitate a substantial amount of data to be analysed and are a cost-effective alternative to real tumours. In addition, given that biological systems encompass interconnections between several components, the integration of various heterogeneous data modalities is often required. Together, this sparks challenges regarding data handing, dimensionality and complexity, and brings into light the need of advanced and specialised computational techniques to accurately investigate these datasets (Hawkins et al. 2010; Jia et al. 2022; Cortés-Ciriano et al. 2021; Nicora et al. 2020).

Currently, several algorithms are standardly leveraged to perform specific bioinformatic analysis tasks, such as ANOVA models to detect sensitive markers and regression models to predict drug response. However, the immeasurable quantity of available biological data enables innumerable unexplored opportunities to create new bioinformatic frameworks to investigate the data. As a result, creating applied and integrative computational methods to analyse and model large-scale datasets are a key step towards elucidating the complexity of cancer and accelerate personalised medicine.

In this thesis, I presented three distinct bioinformatic frameworks to comprehensively investigate and identify new tissue-specific cancer biomarkers from large-scale pharmacological data. Specifically, my frameworks are dedicated to addressing the following aims: (1) identification of resistance markers in HTS, (2) recognition of increasing viability in HTS and exploit for therapies, and (3) prediction of drug response based on cancer-specific gene prioritisation.

These studies were performed using cell line models and consolidated with multi-omic datasets or validations experiments. Together, the proposed research handles statistical vulnerabilities intrinsic in pharmacological data and delivers resources to systematically uncover actionable biomarkers across cancer types.

In brief, the presented work illustrates the use of mathematical models to distinguish biomarkers in multiple cancer subtypes and contributes to the development of novel precision medicine approaches.

4.2 Identifying indirect resistance in HTS

Drug screens, such as GDSC, were designed to maximize the investigation of drug sensitivity markers, where it is generally tested low concentrations in order to avoid cytotoxic effects (Garnett et al., 2012; Iorio et al., 2016). Consequently, in these screens, the majority of the cell lines end up not responding to the tested drug concentrations and the drug responses are summarised with extrapolated IC_{50} values. In contrast, the CTRP screen tests high concentration ranges, which can drive off-target effects in the drug response. Notably, these render difficult to distinguish between real drug resistance and non-responders.

Furthermore, direct approaches to investigate resistance may falsely recognize a non-responder cell line as a resistant one, and are limited to frequently mutated genes, e.g., *TP53* (Figure 17). Overall, these underline that while it is possible to statistically identify resistant lines, the biggest challenge is to recognize resistance biomarkers.

4.2.1 Recognition of indirect resistance with novel outlier detection method

To address the lack of statistical power to identify resistance biomarkers in HTS, together with Iñigo Ayestaran, we developed a systematic hierarchical framework which investigates populations of cell lines carrying sensitivity biomarkers, where a subset of cell lines exhibits unexpected resistance (UNRES) to the treatment.

In total, the framework revealed 57 and 37 significant sensitive markers in the GDSC and CTRP screens, respectively. These findings reflect the inherent design of these drug screens, where the GDSC screening efforts are focused on low concentration ranges to optimize the detection of sensitivity biomarkers.

For the investigation of resistant markers, Iñigo (1) first created a novel statistical outlier detection tool based on standard deviation changes when the highest drug responses values are excluded and I (2) then benchmarked the model against the state-of-the-art Neyman-Pearson (NP) outlier detection method. Both approaches identified multiple resistant markers, with several common associations in GDSC and CTRP.

The proposed UNRES framework recognized several established resistant biomarkers known in literature, such as gefitinib resistance with EGFR T790M mutation (Yun et al. 2008) and afatinib with *PTEN* mutation (Sos et al. 2009) in lung adenocarcinoma, whilst this was not consistently observed by the NP method. Furthermore, the UNRES method revealed a significant performance in the number of detected resistant markers with the CTRP database (17 with UNRES versus 3 with NP).

The considerable discrepancy in performance between these two outlier methods lies on the mathematical structure of the models: whilst the UNRES framework will unavoidably test the top highest IC_{50} values of the mutant population, the NP procedure focus its attention in a fixed critical region of the standard deviation distribution. Notably, if there are few mutant cell lines or if the distribution is approximately normal with small deviations, it is possible that no cell line exposed in the critical region and henceforth the specific CFE-drug association is not tested and recognized with a resistance marker by the NP procedure.

4.2.2 Statistical framework robust to hierarchical tests

The presence of two hierarchical statistical tests - testing of sensitive markers, followed by testing of resistant markers - also influences the statistical sensitivity of the proposed framework. In one hand, a strict significance threshold may filter out relevant sensitive associations to be explored in the second step, whilst, on the other hand, relaxed thresholds may induce a high number of false positives. In order to ensure a robust estimation of our hierarchical testing, I implemented a Hierarchical False Discovery Rate (HFDR) control procedure which provides an upper bound of the overall FDR, and controls the FDR both within and across CFE-drug associations. Specifically, we obtained an upper bound of 22.57% and 22.40% false positives rates in the overall analysis using the GDSC and CTRP screens, respectively. Remarkably, these rates were just slightly higher than the employed 15% threshold to infer UNRES cell lines, and probably influenced by small deviation differences between UNRES and sensitive cell lines.

4.3 Recognizing increasing viability in HTS and exploiting it for therapies

In addition to resistant responses, pharmacological drug screens can be leveraged to reveal other atypical and rare phenotypes of drug response.

Drug response derived from HTS data is typically modelled by curve fitting frameworks based on sigmoidal shapes with several parametrizations related to the expected responder behaviour of the drug response. Notably, unexpected shapes such as increased cell viability are missed by these approaches. This phenomenon is appealing to investigate since may indicate off-target effects, where these drugs inhibit cell cycle checkpoints, guiding to the acceleration of the cell cycle and consequent increased cell proliferation.

4.3.1 Robust curve fitting models

In this work, I have addressed the identification of responses with increased viability through a mathematical framework based on Gaussian Processes (GP) and linear curve fit approaches applied to responses in the GDSC drug screen.

The GP models dynamically adapted to data oscillations and were robust against the noisy nature of the drug responses, revealing small error rates of around 6×10^{-3} in monotonic responses. Notably, this was a phenomenon already observed in another GP-based modelling work I contributed to (D. Wang et al., 2020), where we showed the significant performance of a GP curve fit in comparison to a sigmoid one.

Despite the notorious performance of a GP curve fit, its accuracy highly dependents on the sampling of the cell viability points, both in terms of the quantity of points available and how these are spread across the domain. Specifically, in cases of responses where several outlier points are filtered out, or when the viability points are not evenly distributed, the predicted GP curve struggles to accurately assess the real shape of the curve in the regions with missing data. Furthermore, GP models are notoriously sensible to overfit when dealing with small sets of data. Whilst a small error rate is desired during curve fitting, this could be an indicator of model overfitting.

As an alternative to the GP fitting, I also explored a simpler approach based on linear predictions which was both used during the pre-processing stage and to recognize responses with increased cell viability. The linear gradient successfully distinguished between non-responders and responses with increasing markers, particularly in monotonic responses. Moreover, the model's simplicity made it robust against random fluctuations, provided easy to interpret coefficients and the hyperparameter estimation was a significantly less expensive computational task in comparison to GP.

The pitfall of the linear approach is naturally the poor modelling of the curve shape, as it was demonstrated by the error assessment of the curve fitting model. However, given in this work I was not interested in providing a perfect modelling of the original drug response, but rather a systematic approach that flags responses with a potential increasing behaviour, I found the linear gradient suitable for this task.

4.3.2 Statistical frameworks identify increased cell viability in HTS

Two distinct statistical frameworks based on ANOVA and hypergeometric tests were employed to the linear and GP curve fits to recognize significant increasing cell viability cases in mutant subgroups with respect to wild-type ones.

Initial analysis flagged hundreds and thousands of potential candidates using the ANOVA and hypergeometric enrichment analysis, respectively. Given the ANOVA framework incorporates the mutation status as a cofounding factor in the analysis, the noticeable difference in the amount of increased cell viability markers discovered with the two approaches, is probably driven by a similar slope distribution between the wild-type and mutant populations.

Majority of these candidates were encountered in the cancer types with the highest amount of cell lines, such as lung adenocarcinoma (LUAD), colorectal (COREAD), ovary (OV) and head and neck squamous cell carcinoma (HNSC). Nonetheless, this preliminary analysis did not reveal any striking drug or cancer event of particular interest. In order to short-list the potential candidates, I performed a data-driven approach based on the amount of drug responses with high-slopes in the mutant population of each alteration, tissue type and drug (ATD) association. Based on this strategy, the candidate pool was reduced to only a few hundred of ATD combinations discovered using the hypergeometric pipeline. Looking into these polished results, several compounds such as CHIR-99021 (GSK3α/β inhibitor), SB590885 (B-Raf inhibitor) and IOX2 (PHD2 inhibitor) were distinguished in more than one cancer type.

The systematic investigation of significant ATD associations was performed by cancer type in order to find tumour specific and exclusive markers. As a result, this reduced the available number of samples for each ATD combination and, subsequently, decreased the statistical power of the analyses. Therefore, shifting the focus of the analysis into cancer types with larger annotations of cells and drug, namely LUAD, OV and COREAD.

4.3.3 Evaluation of promising increased viability candidates

From all prospective results, the drug CHIR-99021 was unequivocally the most distinctive candidate, with top significance in LUAD, HNSC and OV for both linear and GP curve fitting models. This compound targets two isoforms of the glycogen synthase kinase-3 (GSK3α and GSK3β) of the Wnt signalling pathway, with GSK-3 being linked to several cancer specific mechanisms such as stem cell formation, and cell proliferation and differentiation.

Specifically, CHIR-99021 is commonly used in combination with other drugs to activate the canonical Wnt signalling pathway, which subsequently promotes cell proliferation and survival in several cancer types, such as LUAD (Fernandes et al. 2024; Li et al. 2013), acute myeloid leukemia (LAML) (Hu et al. 2010) and COREAD (Polakis 2005). Moreover, the function of GSK3 in cancer highly depends of the cancer type, with several reports of its role not only as an oncogene but also as a tumour suppressor (Zheng et al. 2007;

Dong et al. 2005). Notably, these results illustrate the strong influence of the drug CHIR-99021 in the regulation of tumorigenicity and how important a cancer-specific analysis is in order to design an efficient cancer therapy plan.

Significant increasing cell viability markers of CHIR-99021 were found in several combinations of cancer-type and mutation event. From these, focus was given to amplifications with known oncogenes, namely cnaLUAD3 with *TERT* and *TRIP13*, and cnaLUAD27 with *MYC*. Interestingly, *TERT* has been shown to stimulate Wnt pathway activation in stem cells (Park et al. 2009; Listerman et al. 2014), *TRIP13* promotes proliferation via the Wnt pathway in lung cells (Li et al. 2021) and *MYC* also regulates proliferation of colon cancer cells through the induction of LEF1, which interacts with β-catenin (Hao et al. 2019). Together, these studies suggest these oncogenes as possible biomarkers of increased cell viability with GSK3 inhibitors, such as CHIR-99021.

Ultimately, we sought to validate CHIR-99021 in an independent validation experiment that could mimic the same conditions used in the GDSC high-throughput screen. For this, an experimental protocol was designed, using similar media, three replicates, time assessment and approximate drug concentrations. Given the relevance of GSK3 inhibitors in cancer, two other GSK3-targeting compounds, SB216763 and 9-ING-41, were also tested.

Regrettably, the experimental results did not support our hypothesis of increased cell viability induced by CHIR-99021. Although some drug responses, namely the ones involving SK-LU-1 and NCI-H1793 cell lines, presented the indication of a possible start of increased cell viability around the concentration range of 0,5-1,5 $\mu\mathrm{M}$, the signal would be lost in the following cell viability points due to probable cytotoxic effects. Furthermore, the depicted drug response showed high variation between the three replicates, increasing the uncertainty in the experimental results. Nonetheless, from all the compounds tested, CHIR-99021 and SB216763 were the only compounds which manifested a marginal increase in the cell viability.

Several factors could have influenced the lack of validation of our experimental results. For instance, the validated cell lines were not the top candidates of increased cell viability derived from our statistical framework. Unfortunately, the cells 201T and NCI-H2009, which presented the highest drug response slope in GDSC, were not available for our analysis. Moreover, in order to explore the behaviour of the cells as the concentration increases, we extended the range of drug concentrations used in GDSC with high doses spread between large intervals. We hypothesize that at high drug concentrations, toxic effects start to play an active role and dead cells start to emerge and influence the viability signal of the overall response. Similarly, at low concentrations the data is too noisy, making it difficult to translate in follow-up analyses.

Despite the lack of validation of our results in the experimental screen, two sets of publicly available data supported our findings of increased viability with CHIR-99021 in the cell lines SK-LU-1 and NCI-H1793 (Figure 38) (Seashore-Ludlow et al. 2015; Nair et al. 2023). This not only highlights the robustness of the proposed methodology to capture candidates of increasing cell viability in drug screens, but also motivates a deeper examination of the experimental designs to recognize potential improvements.

4.4 Predicting drug response based on cancer-specific gene prioritisation

The identification of genes with biological significance to oncogenic pathways remains an open subject in cancer research. Although several literature resources have pinpointed numerous cancer-specific driver genes, an extension of these lists could provide a better understanding of the underlying mechanisms in cancer, as well as, boost the statistical power of frameworks dedicated to explore cancer biomarkers.

Within this context, computational analyses based on protein-protein interaction networks offer a remarkable opportunity to systematically integrate several biological layers, enabling to efficiently recognize new genes with functional and therapeutic relevance in cancer.

This approach is particularly useful for the previously discussed aims, where the identification of resistance and increasing viability biomarkers is limited due to rare mutation events. Notably, expanding the known mutation landscape may contribute to overcome vulnerabilities in HTS and advance biomarker discovery.

4.4.1 Unveiling new cancer-type specific genes

In this thesis, I presented wPPI, a new network-based systems biology framework which integrates multiple biological layers to identify novel cancer-specific genes. Specifically, wPPI leverages information from protein-protein interaction (PPI) networks from Omnipath, and genotype-phenotype relationships from Gene Ontology (GO) and Human Phenotype Ontology (HPO) ontology databases, respectively. The recognition of new cancer-specific genes is established by the pathway search algorithm Random Walk with Restart (RWR) which infers functional scores to candidate genes in the neighborhood of given cancer-type drivers.

Given breast cancer (BRCA) is one of the most prevalent cancer types worldwide and has a substantial amount of annotated both drugs and cell lines in GDSC, I used it to illustrate the capacity of wPPI as a cancer-specific gene identification and prioritisation tool. Leveraging a curated list of 99 BRCA driver genes from IntOGen, wPPI identified 2,931 new genes with potential relevance to BRCA.

The wPPI candidate genes with the highest functional scores were enriched in pathways with high relevance to breast cancer. Specifically, several pathways associated with DNA damage and repair, p53 signal transduction and ERBB4-related were depicted. The role of *ERBB4* in breast cancer is context-dependent, with the possibility of acting as a tumour driver when interacting with *HER2* or *HER3*, or tumour suppressor by inducing cell apoptosis or cell cycle arrest (El-Gamal et al. 2021; Sundvall et al. 2008). Moreover, although p53 acts as tumour suppressor in cancer, with vital roles in cellular response to DNA damage, a mutation in *TP53* can cancel its suppressive functions and shift its mechanisms to promote breast cancer progression (Marvalim et al. 2023).

For an easy and flexible application, wPPI enables users to input their own set of seed genes, set explorative parameters related to the RWR algorithm and define which ontology datasets to apply (just GO, HPO, both or none). In addition, within the GO context, users can also select which GO classes to leverage, namely Biological Process, Molecular Function and/or Cellular Component.

The influence of the ontology databases and different parameters of the RWR were investigated through a machine learning (ML) model to predict the drug response based on gene expression in the breast tissue. Looking into the models' performances, it is shown that the usage of GO annotations and cancer-driver genes as input are essential to infer tissue-relevant candidate genes. In contrast, the phenotypic information from HPO or the parameters of the RWR search algorithm did not significantly contribute to the performance of the model.

Regarding the first point, I believe that since I focused my investigation of candidate genes in the first order degree neighbors, any pair of genes would not be considerably spread in the PPI network and, therefore, the restart probability r parameter did not have enough power in the search of new proteins. The expansion to higher graph orders should reveal an influence of the restart probability in the selection of new genes.

Furthermore, the small effect of the phenotype information could be driven by the architecture of the HPO ontology database, where both the parental and child terms are included. Notably, in a tree structure, as one advances from the parental to the child annotations, the more specific are the terms encountered in the annotations. For instance, in the context of breast cancer, there were 63 genes associated to breast carcinoma (child annotation), but 92 genes for the neoplasm of the breast (parental annotation). Although wPPI is capable to work with small sets of annotations, the specificity of the child annotations renders lower chances to encounter pairs of genes which share the same HPO annotation, and, subsequently, has minimal functional impact in the prioritisation of genes.

On the other hand, the GO database enables to leverage a slim version where only parental broad terms with full coverage are used, ensuring a sufficient number of terms to calculate a robust functional score.

4.4.2 Better predictive performance based on network-driven genes

The incorporation of network analysis to augment the feature space and interpretability of ML models has become a common practice in biological contexts due to its robust predictive performance and potential translation to clinics. Specifically, one possible application is to leverage network analysis to identify drug response biomarkers in cancer (Kong et al. 2020; Lee et al. 2024; Topol 2015).

Within this framework, I leveraged wPPI to infer new genes with potential relevance to breast cancer, and used them to predict drug response using gene expression of the cell lines. In fact, I analysed the ML performance using several sets of feature space, namely all the genes available in the gene expression (whole genome, 17,419 genes), the breast cancer driver genes used as seed in wPPI (99 genes), and both all and the top 5% candidate genes from wPPI (2,931 and 146 genes, respectively).

Looking into the results, I observed a significant improvement in the prediction of the drug response when using models based on the wPPI genes together with the seed genes. Notably, this recognizes wPPI as a tool to prioritize genes with relevance to breast cancer, whilst reducing the noise in the predictive models.

In contrast, the predictions made with both the smallest (driver genes) and the biggest dataset (whole genome) revealed a limitation in the prediction performance. These results demonstrate two interesting factors: first, the driver genes alone are not robust enough and probably miss relevant markers of drug response; on the other hand, to use a large feature space without a proper selection of informative genes, probably leads to a redundant selection of features and predictions.

Furthermore, I explored the robustness of the models by considering random selections of feature sets of the same size. Expectedly, the random models based on a small set of features revealed an explicit worst prediction. In fact, the Pearson correlation factor disclosed the averaged predicted drug responses were inversely related to the original drug responses. Consistent with the previously discussed results, a knowledge-guided identification of the feature space is paramount to generate reliable predictive models.

The ML model prediction using wPPI outperformed another ML model built with a network-based framework with similar characteristics to wPPI. In detail, I leveraged GeneFriends since it offered a user-friendly tool and enabled to receive a set of input genes and return a set of genes with functional relevance based on a significance threshold. However, this comparison was perhaps imbalanced since GeneFriends solely depends on transcript data, and does not integrate genomics and phenotypic information such as wPPI. Other network-based approaches that include ontology annotation information (GO and/or HPO) are available in literature, but instead of delivering a ranked set of genes, these focus on different aspects such as the identification of disease modules (Buphamalai et al. 2021; Kong et al. 2020) or recognition of gene functions (Ietswaart et al. 2021).

4.5 Limitations and future outlook

This thesis focused on three aims targeting distinct objectives, but commonly addressing critical statistical vulnerabilities in pharmacological data analysis, including rare mutational events and insufficient molecular characterisation. Using robust and flexible computational methodologies, these approaches provided complementary insights into various aspects of drug response modelling, offering a holistic framework to investigate cancer mechanisms and advancing precision oncology.

Despite the significant advancements of the proposed work in cancer biomarker discovery, several limitations can be nominated, namely related with data quality, unexplored analytical optimisations and translational challenges. These issues motivate future developments in the current frameworks to ensure their potential applicability and successful translation to therapeutic strategies.

4.5.1 Data-based improvements

The performance of the computational frameworks was highly conditioned by the quantity and quality of the biological data available. Deficiencies in these two aspects were reflected by the lack of consistency in screening designs, insufficient molecular characterisation of the cell lines and low-frequency mutational events across cancer subtypes.

Challenges in high-throughput screens

The investigation of unexpected phenotypes in HTS, namely resistance and increasing cell viability, was statistically underpowered. Not only these two types of responses are rare events in the drug screens, but the annotated mutation data is often limited to larger cancer types (e.g., lung, colorectal and skin) and alterations involving key cancer drivers like *TP53* and *KRAS*. Notably, these constraints reduced the domain of exploration of these infrequent phenotypes, resulting in a small number of candidates verifying the significance thresholds.

Moreover, several signals within HTS data are obscured with incorrect, irrelevant, or incomplete biological annotations, which often contribute to noise in the predicted cell viabilities. This posed particular challenges to curve fitting methodologies, restricting their ability to accurately summarise the drug response and preventing subsequent association analyses to properly distinguish phenotypes. As a result, meaningful biomarkers could have been missed, underscoring one of the limitations of using drug screens to robustly identify biological signatures.

In this thesis, the effect of data incompleteness was also reflected in the estimation of drug response using gene expression and regression models (chapter 3.3). The average Pearson correlation between real and predicted values across all drugs was relatively low, driven by correlation values close to zero or negative in several drugs. Looking into the gene expression profiles of the BRCA cell lines (Figure 44), several columns depicted high sparsity patterns. This indicates that, for these drugs, only a limited number of cells contributed to the prediction of drug response, thus explaining the underwhelming performance.

In order to mitigate the effects of data incompleteness and/or low characterisation, a thorough literature curation procedure, consolidated with the integration of additional drug screening datasets, is essential. This combined approach could enhance the statistical power of the analyses and ensure the identification of relevant biomarkers, leading to an increase in the number of significant potential biomarkers of resistance, increased cell viability and of drug response.

Despite its potential benefits, the integration of drug screens can be challenging due to inconsistencies in experimental designs. For instance, the GDSC and CTRP pharmacological datasets employ varying ranges of drug concentrations, inconsistent metrics to summarise the drug response and distinct approaches to assess the response at high concentration levels. In addition, the lack of commonalities between the metadata of these screens further complicates this integration (Haibe-Kains et al. 2013). Together, these discrepancies render difficult to directly integrate or cross-compare results between these two screens.

During the investigation of UNRES cell lines, complexities involved of integrating drug response data from both GSDC and CTRP screen were highlighted (chapters 2.2 and 3.1). Whilst GDSC leverages IC_{50} values to define the drug responses, CTRP considers AUC values. Therefore, in order to guarantee a comparable measure, IC_{50} values were additionally computed for the drug responses in CTRP using the same curve fitting model applied for GDSC. Although such additional adjustments contribute to align the datasets, these increase the methodological complexity and constitute a challenge for systematic frameworks.

Currently available tools, such as CellMinerCDB (Rajapakse et al. 2018) and DepMap (Tsherniak et al. 2017), assist the exploration of meaningful biomarkers across various drug screens. However, these frameworks predominantly focus in providing exploratory tools to analyse discrepancies between pharmacological screens as a way to consolidate biomarker hypotheses. Although valuable, these frameworks present considerable limitations, such as the lack of additional analyses to harmonize metrics of different drug screens, or comprehensive characterisation of molecular data from distinct cell lines across the datasets. Additional refinements of these tools could significantly improve to the search of cancer biomarkers in HTS.

Tissue-specific or pan-cancer analysis

The analyses presented in this thesis were conducted across cancer subtypes, enabling to capture tissue-type specific vulnerabilities, and, consequently, hypothesize better tailored therapeutic strategies. In literature, several examples of cancer-specific biomarkers are available, such as EGFR mutations T790M and L858R in NSCLC (Wee and Wang 2017), or BRAF V600E, which is a therapeutic target in melanoma but in colorectal cancer is mainly associated with a poor prognosis (Ascierto et al. 2012; Ducreux et al. 2019).

Within this cancer-specific context, I systematically recognized the drug CHIR-99021 as a strong candidate of increased cell viability across several cancer types, including LUAD, OV and HNSC (section 3.2). A deeper look into the significant biomarkers revealed strong potential applicability in lung cancer, where the alteration events included known cancer drivers. Moreover, this result was further supported by additional enrichment analyses of drug and tissue-type combinations. Together, these findings suggested a potential influence of the tissue type in the increasing cell viability phenotype.

On the other hand, potential markers of interest can be confounded in cancer types with small wild-type and mutant populations, and therefore missed by a tissue-specific approach. In order to maximize the statistical power for HTS-based biomarker discovery, the computational frameworks presented in this thesis could be extended to a pan-cancer analysis. This strategy would probably reduce the number of statistical tests performed and, subsequently increase the pool of biomarker candidates.

Multi-omics integration

The combination of multiple molecular layers is crucial for a comprehensive investigation of biomarkers in HTS. For instance, in this thesis, I presented the network-based wPPI tool, which combines PPI networks, phenotypic and genomic ontologies, to model drug response data based on gene expression profiles (sections 2.4 and 3.3). Given its capability to infer meaningful genes and improve the prediction of the drug response in breast cancer, a future expansion of this multimodal approach to incorporate other data complexities, like mutation profiles, proteomics or pathway information could provide a more complete picture of underlying tissue-specific mechanisms, and further improve the predictive power of the machine learning models (Menden et al. 2018; Cortés-Ciriano et al. 2016; Timpe et al. 2015).

Given the diverse molecular alterations and interactions involved in cancer, the incorporation of multiple data layers (e.g., genomics, proteomics and phenotypic data) is an effective approach to capture the complexity of cancer and its interconnected mechanisms. In conclusion, an integrative multimodal strategy is paramount to guide the development of more effective and personalized therapies in precision oncology.

4.5.2 Extension of analytical tools

Robust and interpretable computational approaches are fundamental in the analysis of cancer datasets. Advanced techniques used in this thesis, namely statistical pipelines, machine learning models or graph networks, facilitated the systematic identification of actionable biomarkers in cancer through multimodal strategies. However, the effectiveness of these analytical tools was challenged by common limitations in biological data, such as data sparsity and intrinsic noise. The investigation of alternatives to mitigate these constraints in the proposed frameworks may enhance their predictive power in future applications.

Graph representations

In the course of this thesis, one of the main computational approaches was a network-based one. Specifically, I implemented a novel systems biology tool, wPPI, which leverages protein-protein interaction (PPI) networks from the Omnipath resource (method details in 2.4.1). Typically, for simplicity, network-driven methodologies do not take into account the directionality of proteins in the PPI network. In a similar way, since the analysis performed in this thesis was restricted to the first order degree neighbours, the scale of the search did not justify modelling this additional complexity in the gene prioritisation process of wPPI.

Some examples of a directional network propagation to distinguish therapies can be found in several studies in cancer (Wathieu et al. 2017; Mo et al. 2022). However, this is still not widely standardized in systems biology frameworks based on PPI data. Given its flexible design, the wPPI framework has potential to be applied in drug discovery applications, where understanding the activation or inhibition mechanisms of proteins is particularly relevant. Therefore, to properly investigate such cases, the integration of the directionality into wPPI would be recommended.

Moreover, within the wPPI pipeline, other graph representations were incorporated, such as the Gene Ontology (GO) and Human Phenotype Ontology (HPO) databases. Typically, these networks are hierarchically structured, where the parent nodes represent general categories and the child denote more specific terms. In contrast to GO, the HPO database delivers both the parental and child annotations regardless of its hierarchical organisation, which can lead to convoluted results. In the future, to increase the impact of HPO, a slimmed version of the HPO database composed by only broader categories could be considered. However, this integration would require a meticulous selection of relevant HPO annotations to prevent incorrect phenotypic associations.

Predictive machine learning models

The modelling of gene expression profiles to predict drug response was successfully performed using Lasso regression models. These models bear the advantage of delivering interpretable gene weights, facilitating the identification of potential biomarkers and future investigation of therapy designs. Despite these advantages, the performance of a Lasso model may be negatively affected by high intrinsic correlation between genes, skewed feature selection or existing nonlinear interactions across genes and/or between genes and drugs.

A possible extension of this work could involve to leverage more complex and advanced methodologies, such as deep learning, support vector machine or random forest models. These would be particularly interesting to apply in cases where the proposed wPPI framework is expanded to include cancers with higher amount of known cancer drivers (e.g., LUAD and COREAD), higher-order neighbour degrees in the PPI network, or with the integration of additional molecular layers beyond gene expression (e.g., proteomics).

Curve fitting

Drug screening data is notoriously noisy and represents a significant challenge in the study of HTS-driven biomarkers. Several pharmacological datasets focus their efforts in screening high volumes of compounds and cells, often at the expense of increasing the number of experimental replicates. For this reason, when curve fitting methodologies perform drug response predictions, high levels of uncertainty may potentially compromise the estimated drug response.

In order to effectively model the intrinsic noise in HTS, an alternative robust Bayesian curve-fitting approach combined with GP models, has been proposed (Wang et al. 2020). This approach revealed superior performances compared with the standardized curve fit assessment available for the GDSC dataset, due to its accountability of uncertainty in the mathematical model. This work served me as inspiration to develop a flexible non-parametric computational curve fitting method based on GP to robustly distinguish phenotypes in the drug screens. Notably, this advanced framework efficiently captured unexpected dose-response curve shapes, such as the increasing cell viability phenotype (results discussed in chapter 3.2.4).

Despite the numerous advantages of the proposed GP-based framework, several data-driven constraints still affected the predicted responses. For example, although I employed an outlier removal procedure during data preparation, remaining residual noise in the experimental data may have deviated the direction of the curve fitting due to potential model overfitting. In order to mitigate this effect, the application of more sophisticated outlier detection methods (Motulsky and Brown 2006) could be considered in a future work.

Another possibility to mitigate the noise-derived limitations, could involve data imputation with artificial viability values to ensure an even data sampling across the whole domain of cell viability. Systematic sampling techniques, like k-nearest neighbors or Bayesian methods, introduce new points based on the existing patterns in the data. Hence, in order to ensure that the true response signal is not overlooked, an efficient prior management of noise is necessary for an appropriate data imputation.

Additional unexpected phenotypes

In this thesis, I focused my attention to recognize responses with increasing cell viability using linear and Gaussian-based curve fitting methods (chapters 2.3 and 3.2). These approaches were built without specific parameterizations on the shape of the drug response, allowing to capture atypical drug responses. As a result, this flexible design enables to adapt these frameworks to further investigate other unexpected phenotypes in HTS, such as E_{max} value smaller or higher than zero (Figure 8).

In order to capture these additional patterns in HTS, I could extend and tailor my current work with GP curve fitting by dynamically exploring additional kernels, and integrate them with summary metrics specific to each unexpected response phenotype. For instance, the gradient of the curve fit model can indicate whether the response is increasing or decreasing. However, this measure alone does not inform regarding the starting and final viability of the cells, or about the distribution of the cell viability values.

The most challenging task would be to distinguish between responders with an E_{max} close to zero and responders with an E_{max} higher than zero. In both cases, the curve shape is decreasing but the threshold separating these classes is directly correlated with the distribution of the viability values. Therefore, I hypothesize that implementing a framework that dynamically takes into account the distribution of the cell viability, in combination with a GP curve fitting model, could be a possible alternative framework to discriminate unexpected drug response patterns in future work.

4.5.3 Translation of findings as novel cancer therapies

The findings unveiled throughout this thesis were functionally contextualised based on

- *gold-standards*, e.g., in the investigation of UNRES cell lines, known resistance biomarkers T790M mutation and PTEN deletion in LUAD were identified by the proposed outlier-detection pipeline (section 3.1.3);
- *in vitro* experiments, e.g., CHIR-99021 and other GSK3-inhibitor compounds were screened in an independent cell culture experiment against selected LUAD cell lines (section 3.2.9);
- other computational frameworks, e.g., the genes selected by the wPPI prioritisation tool were functionally analysed through pathway enrichment analyses and benchmarked with the GeneFriends prioritisation method (section 3.3.4).

The work developed in this thesis was based on cell line models, which are simplified representations of a real tumour. This, in combination with possible artifacts in the datasets or unsuitable computational analyses, means that HTS-driven biomarker candidates can only provide indications of biological signals.

In order to increase the translation impact of the presented pipelines, more realistic preclinical models based on CRISPR screens, organoids or patient-derived xenograft (PDX), could be considered. For instance, CRISPR resistance screens analysis could validate and clarify the resistance mechanisms of the identified UNRES cell lines and drugs (Ayestaran et al., 2020).

Additionally, as an extension of the work performed in chapter 3.3.3, CRISPR knockout screens of genes with drug response markers could help to validate the positive or negative correlation of these genes with the drug response. In parallel, gene expression profiles of organoids or *in vivo* patient-derived data, could more accurately capture tumour-specificity transcriptional activity, leading to more robust predictive models of drug response (Lee et al. 2018; Nguyen et al. 2021).

Furthermore, PDX models could be valuable complementary tools to preliminary investigations in cell lines, organoids or CRISPR screens, by verifying drug efficacy, and bridging preclinical findings to clinical applications (Liu et al. 2023).

A long-term goal of this thesis would involve to explore novel cancer therapies by designing synergistic drug combinations built on HTS-driven atypical drug responses biomarkers, namely resistance and increased cell viability biomarkers. Moreover, this investigation could be enriched by leveraging the wPPI tool to integrate additional molecular features (e.g., gene expression profiles, mutation events or CNVs) into the drug combinations prediction. In summary, I believe this integration strategy could suggest highly effective and targeted cancer therapies, ultimately driving significant advancements in precision oncology.

References

Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precision Oncology*, *4*, 19.

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, *422*(6928), 198–207.

Ahmadi Moughari, F., & Eslahchi, C. (2021). A computational method for drug sensitivity prediction of cancer cell lines based on various molecular information. *PloS One*, *16*(4), e0250620.

Alanis-Lobato, G., Andrade-Navarro, M. A., & Schaefer, M. H. (2016). HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, *45*(D1), D408–D414.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2014). *Molecular Biology of the Cell*. Garland Science.

Aldonza, M. B. D., Ku, J., Hong, J.-Y., Kim, D., Yu, S. J., Lee, M.-S., Prayogo, M. C., Tan, S., Kim, D., Han, J., Lee, S. K., Im, S. G., Ryu, H. S., & Kim, Y. (2020). Prior acquired resistance to paclitaxel relays diverse EGFR-targeted therapy persistence mechanisms. *Science Advances*, *6*(6), eaav7416.

Alhmoud, J. F., Woolley, J. F., Al Moustafa, A.-E., & Malki, M. I. (2020). DNA Damage/Repair Management in Cancers. *Cancers*, *12*(4). https://doi.org/10.3390/cancers12041050

Allegra, C. J., Jessup, J. M., Somerfield, M. R., Hamilton, S. R., Hammond, E. H., Hayes, D. F., McAllister, P. K., Morton, R. F., & Schilsky, R. L. (2009). American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *27*(12), 2091–2096.

Allis, C. D., & Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nature Reviews. Genetics*, *17*(8), 487–500.

Almal, S. H., & Padh, H. (2012). Implications of gene copy-number variation in health and diseases. *Journal of Human Genetics*, *57*(1), 6–13.

Altieri, D. C. (2008). Survivin, cancer networks and pathway-directed drug discovery. *Nature Reviews. Cancer*, *8*(1), 61–70.

Amaravadi, R., Kimmelman, A. C., & White, E. (2016). Recent insights into the function of autophagy in cancer. *Genes & Development*, *30*(17), 1913–1930.

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43*(Database issue), D789–D798.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106.

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P., & Maley, C. C. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, *22*(1), 105–113.

An, X., Tiwari, A. K., Sun, Y., Ding, P.-R., Ashby, C. R., Jr, & Chen, Z.-S. (2010). BCR-ABL tyrosine kinase inhibitors in the treatment of Philadelphia chromosome positive chronic myeloid leukemia: a review. *Leukemia Research*, *34*(10), 1255–1268.

Aradottir, M., Reynisdottir, S. T., Stefansson, O. A., Jonasson, J. G., Sverrisdottir, A., Tryggvadottir, L., Eyfjord, J. E., & Bodvarsdottir, S. K. (2015). Aurora A is a prognostic marker for breast cancer arising in BRCA2 mutation carriers. *Hip International: The Journal of Clinical and Experimental Research on Hip Pathology and Therapy*, *1*(1), 33–40.

Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A., & Lopez-Bigas, N. (2019). OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* , *35*(22), 4788–4790.

Arruebo, M., Vilaboa, N., Sáez-Gutierrez, B., Lambea, J., Tres, A., Valladares, M., & González-Fernández, A. (2011). Assessment of the evolution of cancer treatment therapies. *Cancers*, *3*(3), 3279–3330.

Ascierto, P. A., Kirkwood, J. M., Grob, J.-J., Simeone, E., Grimaldi, A. M., Maio, M., Palmieri, G., Testori, A., Marincola, F. M., & Mozzillo, N. (2012). The role of BRAF V600 mutation in melanoma. *Journal of Translational Medicine*, *10*(1), 1–9.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29.

Aubrey, B. J., Kelly, G. L., Janic, A., Herold, M. J., & Strasser, A. (2018). How does p53 induce apoptosis and how does this relate to p53-mediated tumour suppression? *Cell Death and Differentiation*, *25*(1), 104–113.

Ayestaran, I., Galhoz, A., Spiegel, E., Sidders, B., Dry, J. R., Dondelinger, F., Bender, A., McDermott, U., Iorio, F., & Menden, M. P. (2020). Identification of Intrinsic Drug Resistance and Its Biomarkers in High-Throughput Pharmacogenomic and CRISPR Screens. *Patterns (New York, N.Y.)*, *1*(5), 100065.

Babur, Ö., Gönen, M., Aksoy, B. A., Schultz, N., Ciriello, G., Sander, C., & Demir, E. (2015). Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biology*, *16*(1), 45.

Babur, Ö., Luna, A., Korkut, A., Durupinar, F., Siper, M. C., Dogrusoz, U., Vaca Jacome, A. S., Peckner, R., Christianson, K. E., Jaffe, J. D., Spellman, P. T., Aslan, J. E., Sander, C., & Demir, E. (2021). Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *Patterns (New York, N.Y.)*, *2*(6), 100257.

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M. C., Kim, J., Reardon, B., Ng, P. K.-S., Jeong, K. J., Cao, S., Wang, Z., Gao, J., Gao, Q., Wang, F., Liu, E. M., Mularoni, L., … Karchin, R. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *173*(2), 371–385.e18.

Bakhoum, S. F., Ngo, B., Laughney, A. M., Cavallo, J.-A., Murphy, C. J., Ly, P., Shah, P., Sriram, R. K., Watkins, T. B. K., Taunk, N. K., Duran, M., Pauli, C., Shaw, C., Chadalavada, K., Rajasekhar, V. K., Genovese, G., Venkatesan, S., Birkbak, N. J., McGranahan, N., … Cantley, L. C. (2018). Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature*, *553*(7689), 467–472.

Bakker, B., Taudt, A., Belderbos, M. E., Porubsky, D., Spierings, D. C. J., de Jong, T. V., Halsema, N., Kazemier, H. G., Hoekstra-Wakker, K., Bradley, A., de Bont, E. S. J. M., van den Berg, A., Guryev, V., Lansdorp, P. M., Colomé-Tatché, M., & Foijer, F. (2016). Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biology*, *17*(1), 115.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. R., & Wooster, R. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer*, *91*(2), 355–358.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603–607.

Baskar, R., Lee, K. A., Yeo, R., & Yeoh, K.-W. (2012). Cancer and radiation therapy: current advances and future directions. *International Journal of Medical Sciences*, *9*(3), 193–199.

Basu, A., Bodycombe, N. E., Cheah, J. H., Price, E. V., Liu, K., Schaefer, G. I., Ebright, R. Y., Stewart, M. L., Ito, D., Wang, S., Bracha, A. L., Liefeld, T., Wawer, M., Gilbert, J. C., Wilson, A. J., Stransky, N., Kryukov, G. V., Dancik, V., Barretina, J., … Schreiber, S. L. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, *154*(5), 1151–1161.

Basu, A., & Krishnamurthy, S. (2010). Cellular responses to Cisplatin-induced DNA damage. *Journal of Nucleic Acids*, *2010*. https://doi.org/10.4061/2010/201367

Baugh, E. H., Ke, H., Levine, A. J., Bonneau, R. A., & Chan, C. S. (2018). Why are there hotspot mutations in the TP53 gene in human cancers? *Cell Death and Differentiation*, *25*(1), 154–160.

Bayat Mokhtari, R., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., & Yeger, H. (2017). Combination therapy in combating cancer. *Oncotarget*, *8*(23), 38022–38043.

Bayer, F. P., Gander, M., Kuster, B., & The, M. (2023). CurveCurator: a recalibrated F-statistic to assess, classify, and explore significance of dose-response curves. *Nature Communications*, *14*(1), 7902.

Becsey, J. C., Berke, L., & Callan, J. R. (1968). Nonlinear least squares methods: A direct grid search approach. *Journal of Chemical Education*, *45*(11), 728.

Beenken, S. W., Grizzle, W. E., Crowe, D. R., Conner, M. G., Weiss, H. L., Sellers, M. T., Krontiras, H., Urist, M. M., & Bland, K. I. (2001). Molecular biomarkers for breast cancer prognosis: coexpression of c-erbB-2 and p53. *Annals of Surgery*, *233*(5), 630–638.

Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., Strathdee, C. A., Dempster, J., Lyons, N. J., Burns, R., Nag, A., Kugener, G., Cimini, B., Tsvetkov, P., Maruvka, Y. E., O'Rourke, R., Garrity, A., Tubelli, A. A., Bandopadhayay, P., … Golub, T. R. (2018). Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*, *560*(7718), 325–330.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *Annals of Statistics*, *29*(4), 1165–1188.

Bentzen, S. M. (2006). Preventing or reducing late side effects of radiation therapy: radiobiology meets molecular pathology. *Nature Reviews. Cancer*, *6*(9), 702–713.

Berger, M. F., & Mardis, E. R. (2018). The emerging clinical relevance of genomics in cancer medicine. *Nature Reviews. Clinical Oncology*, *15*(6), 353–365.

Bingham, N. H., & Fry, J. M. (2010). *Regression: Linear Models in Statistics*. Springer Science & Business Media.

Blackwell, K. L., Zaman, K., Qin, S., Tkaczuk, K. H. R., Campone, M., Hunt, D., Bryce, R., Goldstein, L. J., & 202 Study Group. (2019). Neratinib in Combination With Trastuzumab for the Treatment of Patients With Advanced HER2-positive Breast Cancer: A Phase I/II Study. *Clinical Breast Cancer*, *19*(2), 97–104.e4.

Blagosklonny, M. V. (2005). Overcoming limitations of natural anticancer drugs by combining with artificial agents. *Trends in Pharmacological Sciences*, *26*(2), 77–81.

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Seeber.

Boniolo, F., Dorigatti, E., Ohnmacht, A. J., Saur, D., Schubert, B., & Menden, M. P. (2021). Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opinion on Drug Discovery*, *16*(9), 991–1007.

Bouhaddou, M., DiStefano, M. S., Riesel, E. A., Carrasco, E., Holzapfel, H. Y., Jones, D. C., Smith, G. R., Stern, A. D., Somani, S. S., Thompson, T. V., & Birtwistle, M. R. (2016). Drug response consistency in CCLE and CGP. *Nature*, *540*(7631), E9–E10.

Bramsen, J. B., Rasmussen, M. H., Ongen, H., Mattesen, T. B., Ørntoft, M.-B. W., Árnadóttir, S. S., Sandoval, J., Laguna, T., Vang, S., Øster, B., Lamy, P., Madsen, M. R., Laurberg, S., Esteller, M., Dermitzakis, E. T., Ørntoft, T. F., & Andersen, C. L. (2017). Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. *Cell Reports*, *19*(6), 1268–1280.

Braun, T. P., Eide, C. A., & Druker, B. J. (2020). Response and Resistance to BCR-ABL1-Targeted Therapies. *Cancer Cell*, *37*(4), 530–542.

Breitbach, C. J., Reid, T., Burke, J., Bell, J. C., & Kirn, D. H. (2010). Navigating the clinical development landscape for oncolytic viruses and other cancer therapeutics: no shortcuts on the road to approval. *Cytokine & Growth Factor Reviews*, *21*(2-3), 85–89.

Bridges, C. C. (1966). Hierarchical Cluster Analysis. *Psychological Reports*, *18*(3), 851–854.

Brionne, A., Juanchich, A., & Hennequet-Antier, C. (2019). ViSEAGO: a Bioconductor package for clustering biological functions using Gene Ontology and semantic similarity. *BioData Mining*, *12*, 16.

Brunner, H. G., & van Driel, M. A. (2004). From syndrome families to functional genomics. *Nature Reviews. Genetics*, *5*(7), 545–551.

Buehring, G. C., Eby, E. A., & Eby, M. J. (2004). Cell line cross-contamination: how aware are Mammalian cell culturists of the problem and how to monitor it? *In Vitro Cellular & Developmental Biology. Animal*, *40*(7), 211–215.

Buphamalai, P., Kokotovic, T., Nagy, V., & Menche, J. (2021). Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nature Communications*, *12*(1), 1–15.

Burrell, R. A., McGranahan, N., Bartek, J., & Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, *501*(7467), 338–345.

Byrne, M. J. (1976). Cyclophosphamide, vincristine and procarbazine in the treatment of malignant melanoma. *Cancer*, *38*(5), 1922–1924.

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-Generation Machine Learning for Biological Networks. *Cell*, *173*(7), 1581–1592.

Camargo, A., Azuaje, F., Wang, H., & Zheng, H. (2008). Permutation - based statistical tests for multiple hypotheses. *Source Code for Biology and Medicine*, *3*, 15.

Cancer Cell Line Encyclopedia Consortium, & Genomics of Drug Sensitivity in Cancer Consortium. (2015). Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, *528*(7580), 84–87.

Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70.

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., & Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, *12*(1), 124.

Cao, J., Gong, J., Li, X., Hu, Z., Xu, Y., Shi, H., Li, D., Liu, G., Jie, Y., Hu, B., & Chong, Y. (2021). Unsupervised Hierarchical Clustering Identifies Immune Gene Subtypes in Gastric Cancer. *Frontiers in Pharmacology*, *12*, 692454.

Cao, M., Pietras, C. M., Feng, X., Doroschak, K. J., Schaffner, T., Park, J., Zhang, H., Cowen, L. J., & Hescott, B. J. (2014). New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* , *30*(12), i219–i227.

Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews. Genetics*, *3*(4), 285–298.

Carter, P., Presta, L., Gorman, C. M., Ridgway, J. B., Henner, D., Wong, W. L., Rowland, A. M., Kotts, C., Carver, M. E., & Shepard, H. M. (1992). Humanization of an anti-p185HER2 antibody for human cancer therapy. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(10), 4285–4289.

Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N., & Szallasi, Z. (2006). A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nature Genetics*, *38*(9), 1043–1048.

Castro-Giner, F., Ratcliffe, P., & Tomlinson, I. (2015). The mini-driver model of polygenic cancer evolution. *Nature Reviews. Cancer*, *15*(11), 680–685.

Chang, B. D., Broude, E. V., Dokmanovic, M., Zhu, H., Ruth, A., Xuan, Y., Kandel, E. S., Lausch, E., Christov, K., & Roninson, I. B. (1999). A senescence-like phenotype distinguishes tumor cells that undergo terminal proliferation arrest after exposure to anticancer agents. *Cancer Research*, *59*(15), 3761–3767.

Chen, B.-J., Litvin, O., Ungar, L., & Pe'er, D. (2015). Context Sensitive Modeling of Cancer Drug Sensitivity. *PloS One*, *10*(8), e0133850.

Cheng, F., Desai, R. J., Handy, D. E., Wang, R., Schneeweiss, S., Barabási, A.-L., & Loscalzo, J. (2018). Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nature Communications*, *9*(1), 2691.

Cheng, Y., Wang, K., Geng, L., Sun, J., Xu, W., Liu, D., Gong, S., & Zhu, Y. (2019). Identification of candidate diagnostic and prognostic biomarkers for pancreatic carcinoma. *EBioMedicine*, *40*, 382–393.

Chen, X., Zhu, Q., Zhu, L., Pei, D., Liu, Y., Yin, Y., Schuler, M., & Shu, Y. (2013). Clinical perspective of afatinib in non-small cell lung cancer. *Lung Cancer* , *81*(2), 155–161.

Chen, Z., Zhang, P., Xu, Y., Yan, J., Liu, Z., Lau, W. B., Lau, B., Li, Y., Zhao, X., Wei, Y., & Zhou, S. (2019). Surgical stress and cancer progression: the twisted tango. *Molecular Cancer*, *18*(1), 132.

Chia, S., Low, J.-L., Zhang, X., Kwang, X.-L., Chong, F.-T., Sharma, A., Bertrand, D., Toh, S. Y., Leong, H.-S., Thangavelu, M. T., Hwang, J. S. G., Lim, K.-H., Skanthakumar, T., Tan, H.-K., Su, Y., Hui Choo, S., Hentze, H., Tan, I. B. H., Lezhava, A., … DasGupta, R. (2017). Phenotype-driven precision oncology as a guide for clinical decisions one patient at a time. *Nature Communications*, *8*(1), 435.

Childs, B. G., Baker, D. J., Kirkland, J. L., Campisi, J., & van Deursen, J. M. (2014). Senescence and apoptosis: dueling or complementary cell fates? *EMBO Reports*, *15*(11), 1139–1153.

Choi, H.-J., Chung, T.-W., Kim, C.-H., Jeong, H.-S., Joo, M., Youn, B., & Ha, K.-T. (2012). Estrogen induced β-1,4-galactosyltransferase 1 expression regulates proliferation of human breast cancer MCF-7 cells. *Biochemical and Biophysical Research Communications*, *426*(4), 620–625.

Chou, T.-C. (2006). Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. *Pharmacological Reviews*, *58*(3), 621–681.

Cho, Y., Gorina, S., Jeffrey, P. D., & Pavletich, N. P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. *Science*, *265*(5170), 346–355.

Christopoulos, P. F., Msaouel, P., & Koutsilieris, M. (2015). The role of the insulin-like growth factor-1 system in breast cancer. *Molecular Cancer*, *14*, 43.

Ciriello, G., Cerami, E., Sander, C., & Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, *22*(2), 398–406.

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, *45*(10), 1127–1133.

Colaprico, A., Olsen, C., Bailey, M. H., Odom, G. J., Terkelsen, T., Silva, T. C., Olsen, A. V., Cantini, L., Zinovyev, A., Barillot, E., Noushmehr, H., Bertoli, G., Castiglioni, I., Cava, C., Bontempi, G., Chen, X. S., & Papaleo, E. (2020). Interpreting pathways to discover cancer driver genes with Moonlight. *Nature Communications*, *11*(1), 69.

Collins, F. S., & Mansoura, M. K. (2001). The Human Genome Project. *Cancer*, *91*(S1), 221–225.

Cooke, T., Reeves, J., Lanigan, A., & Stanton, P. (2001). HER2 as a prognostic and predictive marker for breast cancer. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, *12 Suppl 1*, S23–S28.

Cook, R. D., & Dennis Cook, R. (1977). Detection of Influential Observation in Linear Regression. In *Technometrics* (Vol. 19, Issue 1, p. 15). https://doi.org/10.2307/1268249

Cornish, A. J., David, A., & Sternberg, M. J. E. (2018). PhenoRank: reducing study bias in gene prioritization through simulation. *Bioinformatics* , *34*(12), 2087–2095.

Cortés-Ciriano, I., Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M., & Park, P. J. (2021). Computational analysis of cancer genome sequencing data. *Nature Reviews Genetics*, *23*(5), 298–314.

Cortés-Ciriano, I., van Westen, G. J. P., Bouvier, G., Nilges, M., Overington, J. P., Bender, A., & Malliavin, T. E. (2016). Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics (Oxford, England)*, *32*(1), 85–95.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cortes, J., Talpaz, M., O'Brien, S., Jones, D., Luthra, R., Shan, J., Giles, F., Faderl, S., Verstovsek, S., Garcia-Manero, G., Rios, M. B., & Kantarjian, H. (2005). Molecular responses in patients with chronic myelogenous leukemia in chronic phase treated with imatinib mesylate. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *11*(9), 3425–3432.

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S. A., Mpindi, J.-P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., NCI DREAM Community, Collins, J. J., Gallahan, D., Singer, D., … Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, *32*(12), 1202–1212.

Crick, F. H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, *12*, 138–163.

Croce, C. M. (2008). Oncogenes and cancer. *The New England Journal of Medicine*, *358*(5), 502–511.

Dalmartello, M., La Vecchia, C., Bertuccio, P., Boffetta, P., Levi, F., Negri, E., & Malvezzi, M. (2022). European cancer mortality predictions for the year 2022 with focus on ovarian cancer. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, *33*(3), 330–339.

Dawson, D. A., Genco, N., Bensinger, H. M., Guinn, D., Il'giovine, Z. J., Wayne Schultz, T., & Pöch, G. (2012). Evaluation of an asymmetry parameter for curve-fitting in single-chemical and mixture toxicity assessment. *Toxicology*, *292*(2-3), 156–161.

Degtyareva, A. O., Leberfarb, E. Y., Efimova, E. G., Brusentsov, I. I., Usova, A. V., Lushnikova, E. L., & Merkulova, T. I. (2020). rs2072580T>A Polymorphism in the Overlapping Promoter Regions of the SART3 and ISCU Genes Associated with the Risk of Breast Cancer. *Bulletin of Experimental Biology and Medicine*, *169*(1), 81–84.

Del Toro, N., Shrivastava, A., Ragueneau, E., Meldal, B., Combe, C., Barrera, E., Perfetto, L., How, K., Ratan, P., Shirodkar, G., Lu, O., Mészáros, B., Watkins, X., Pundir, S., Licata, L., Iannuccelli, M., Pellegrini, M., Martin, M. J., Panni, S., … Hermjakob, H. (2022). The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Research*, *50*(D1), D648–D653.

Di Veroli, G. Y., Fornari, C., Goldlust, I., Mills, G., Koh, S. B., Bramhall, J. L., Richards, F. M., & Jodrell, D. I. (2015). An automated fitting procedure and software for dose-response curves with multiphasic features. *Scientific Reports*, *5*, 14701.

Doherty, J. K., Bond, C., Jardim, A., Adelman, J. P., & Clinton, G. M. (1999). The HER-2/neu receptor tyrosine kinase gene encodes a secreted autoinhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(19), 10869–10874.

Domon, B., & Aebersold, R. (2010). Options and considerations when selecting a quantitative proteomics strategy. *Nature Biotechnology*, *28*(7), 710–721.

Dong, J., Peng, J., Zhang, H., Mondesire, W. H., Jian, W., Mills, G. B., Hung, M.-C., & Meric-Bernstam, F. (2005). Role of glycogen synthase kinase 3beta in rapamycin-mediated cell cycle regulation and chemosensitivity. *Cancer Research*, *65*(5), 1961–1972.

Druillennec, S., Dorard, C., & Eychène, A. (2012). Alternative splicing in oncogenic kinases: from physiological functions to cancer. *Journal of Nucleic Acids*, *2012*, 639062.

Duan, J., Wainwright, M. S., Comeron, J. M., Saitou, N., Sanders, A. R., Gelernter, J., & Gejman, P. V. (2003). Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Human Molecular Genetics*, *12*(3), 205–216.

Ducreux, M., Chamseddine, A., Laurent-Puig, P., Smolenschi, C., Hollebecque, A., Dartigues, P., Samallin, E., Boige, V., Malka, D., & Gelli, M. (2019). Molecular targeted therapy of BRAF-mutant colorectal cancer. *Therapeutic Advances in Medical Oncology*. https://doi.org/10.1177/1758835919856494

Duda, P., Akula, S. M., Abrams, S. L., Steelman, L. S., Martelli, A. M., Cocco, L., Ratti, S., Candido, S., Libra, M., Montalto, G., Cervello, M., Gizak, A., Rakus, D., & McCubrey, J. A. (2020). Targeting GSK3 and Associated Signaling Pathways Involved in Cancer. *Cells* , *9*(5). https://doi.org/10.3390/cells9051110

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210.

Ekyalongo, R. C., & Yee, D. (2017). Revisiting the IGF-1R as a breast cancer target. *NPJ Precision Oncology*, *1*. https://doi.org/10.1038/s41698-017-0017-y

El-Gamal, M. I., Mewafi, N. H., Abdelmotteleb, N. E., Emara, M. A., Tarazi, H., Sbenati, R. M., Madkour, M. M., Zaraei, S.-O., Shahin, A. I., & Anbar, H. S. (2021). A Review of HER4 (ErbB4) Kinase, Its Impact on Cancer, and Its Inhibitors. *Molecules*, *26*(23), 7376.

Elgar, G., & Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in Genetics: TIG*, *24*(7), 344–352.

Emrich, S. J., Barbazuk, W. B., Li, L., & Schnable, P. S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, *17*(1), 69–73.

Encyclopedia of Cancer. (2018). Academic Press.

Eroglu, Z., & Ribas, A. (2016). Combination therapy with BRAF and MEK inhibitors for melanoma: latest evidence and place in therapy. *Therapeutic Advances in Medical Oncology*, *8*(1), 48–56.

Evans, W. E., & McLeod, H. L. (2003). Pharmacogenomics--drug disposition, drug targets, and side effects. *The New England Journal of Medicine*, *348*(6), 538–549.

Farnoud, A., Ohnmacht, A. J., Meinel, M., & Menden, M. P. (2022). Can artificial intelligence accelerate preclinical drug discovery and precision medicine? *Expert Opinion on Drug Discovery*, *17*(7), 661–665.

Favaro, E., Ramachandran, A., McCormick, R., Gee, H., Blancher, C., Crosby, M., Devlin, C., Blick, C., Buffa, F., Li, J.-L., Vojnovic, B., Pires das Neves, R., Glazer, P., Iborra, F., Ivan, M., Ragoussis, J., & Harris, A. L. (2010). MicroRNA-210 regulates mitochondrial free radical response to hypoxia and krebs cycle in cancer cells by targeting iron sulfur cluster protein ISCU. *PloS One*, *5*(4), e10345.

Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. *International Journal of Cancer. Journal International Du Cancer*. https://doi.org/10.1002/ijc.33588

Fernandes, R., Barbosa-Matos, C., Borges-Pereira, C., Carvalho, A. L. R. T. de, & Costa, S. (2024). Glycogen Synthase Kinase-3 Inhibition by CHIR99021 Promotes Alveolar Epithelial

Cell Proliferation and Lung Regeneration in the Lipopolysaccharide-Induced Acute Lung Injury Mouse Model. *International Journal of Molecular Sciences*, *25*(2), 1279.

Fernández, I. F., Blanco, S., Lozano, J., & Lazo, P. A. (2010). VRK2 inhibits mitogen-activated protein kinase signaling and inversely correlates with ErbB2 in human breast cancer. *Molecular and Cellular Biology*, *30*(19), 4687–4697.

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R. M., & Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, *84*(4), 524–533.

Fisher, R. A. (1992). Statistical Methods for Research Workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in Statistics: Methodology and Distribution* (pp. 66–70). Springer New York.

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suvà, M. L., & Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, *529*(7584), 110–114.

Flavahan, W. A., Gaskell, E., & Bernstein, B. E. (2017). Epigenetic plasticity and the hallmarks of cancer. *Science*, *357*(6348). https://doi.org/10.1126/science.aal2380

Foster, S. A., Whalen, D. M., Özen, A., Wongchenko, M. J., Yin, J., Yen, I., Schaefer, G., Mayfield, J. D., Chmielecki, J., Stephens, P. J., Albacker, L. A., Yan, Y., Song, K., Hatzivassiliou, G., Eigenbrot, C., Yu, C., Shaw, A. S., Manning, G., Skelton, N. J., … Malek, S. (2016). Activation Mechanism of Oncogenic Deletion Mutations in BRAF, EGFR, and HER2. *Cancer Cell*, *29*(4), 477–493.

Fredriksson, N. J., Ny, L., Nilsson, J. A., & Larsson, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genetics*, *46*(12), 1258–1263.

Freeberg, M. A., Fromont, L. A., D’Altri, T., Romero, A. F., Ciges, J. I., Jene, A., Kerry, G., Moldes, M., Ariosa, R., Bahena, S., Barrowdale, D., Barbero, M. C., Fernandez-Orth, D., Garcia-Linares, C., Garcia-Rios, E., Haziza, F., Juhasz, B., Llobet, O. M., Milla, G., … Rambla, J. (2022). The European Genome-phenome Archive in 2021. *Nucleic Acids Research*, *50*(D1), D980–D987.

Fu, D., Calvo, J. A., & Samson, L. D. (2012). Balancing repair and tolerance of DNA damage caused by alkylating agents. *Nature Reviews. Cancer*, *12*(2), 104–120.

Funauchi, Y., Tanikawa, C., Yi Lo, P. H., Mori, J., Daigo, Y., Takano, A., Miyagi, Y., Okawa, A., Nakamura, Y., & Matsuda, K. (2015). Regulation of iron homeostasis by the p53-ISCU pathway. *Scientific Reports*, *5*, 16497.

Gajria, D., & Chandarlapaty, S. (2011). HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies. *Expert Review of Anticancer Therapy*, *11*(2), 263–275.

Galhoz, A., Turei, D., & P. Menden, M. (2021). *wppi*. Bioconductor. https://doi.org/10.18129/B9.BIOC.WPPI

Gallagher, P. G., Liem, R. I., Wong, E., Weiss, M. J., & Bodine, D. M. (2005). GATA-1 and Oct-1 are required for expression of the human alpha-hemoglobin-stabilizing protein gene. *The Journal of Biological Chemistry*, *280*(47), 39016–39023.

Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., Vermeulen, L., & Wang, X. (2019). DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis*, *8*(9), 44.

Gao, H., Korn, J. M., Ferretti, S., Monahan, J. E., Wang, Y., Singh, M., Zhang, C., Schnell, C., Yang, G., Zhang, Y., Balbin, O. A., Barbe, S., Cai, H., Casey, F., Chatterjee, S., Chiang, D. Y., Chuai, S., Cogan, S. M., Collins, S. D., … Sellers, W. R. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature Medicine*, *21*(11), 1318–1325.

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D., & Petersen, I. (2001). Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(24), 13784–13789.

Garbis, S., Lubec, G., & Fountoulakis, M. (2005). Limitations of current proteomics technologies. *Journal of Chromatography. A*, *1077*(1), 1–18.

Garcia-Alonso, L., Iorio, F., Matchan, A., Fonseca, N., Jaaks, P., Peat, G., Pignatelli, M., Falcone, F., Benes, C. H., Dunham, I., Bignell, G., McDade, S. S., Garnett, M. J., & Saez-Rodriguez, J. (2018). Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Research*, *78*(3), 769–780.

García Rubiño, M. E., Carrillo, E., Ruiz Alcalá, G., Domínguez-Martín, A., A Marchal, J., & Boulaiz, H. (2019). Phenformin as an Anticancer Agent: Challenges and Prospects. *International Journal of Molecular Sciences*, *20*(13). https://doi.org/10.3390/ijms20133316

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., Liu, Q., Iorio, F., Surdez, D., Chen, L., Milano, R. J., Bignell, G. R., Tam, A. T., Davies, H., Stevenson, J. A., … Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, *483*(7391), 570–575.

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

Geeleher, P., Gamazon, E. R., Seoighe, C., Cox, N. J., & Huang, R. S. (2016). Consistency in large pharmacogenomic studies. *Nature*, *540*(7631), E1–E2.

Geyer, C. E., Forster, J., Lindquist, D., Chan, S., Romieu, C. G., Pienkowski, T., Jagiello-Gruszfeld, A., Crown, J., Chan, A., Kaufman, B., Skarlos, D., Campone, M., Davidson, N., Berger, M., Oliva, C., Rubin, S. D., Stein, S., & Cameron, D. (2006). Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *The New England Journal of Medicine*, *355*(26), 2733–2743.

Gey, G. (1952). Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. *Cancer Research*, *12*, 264–265.

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., Deng, C., Varusai, T., Ragueneau, E., Haider, Y., May, B., Shamovsky, V., Weiser, J., Brunson, T., Sanati, N., … D'Eustachio, P. (2022). The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, *50*(D1), D687–D692.

Gillet, J.-P., Varma, S., & Gottesman, M. M. (2013). The clinical relevance of cancer cell lines. *Journal of the National Cancer Institute*, *105*(7), 452–458.

Girden, E. R. (1992). *ANOVA: Repeated Measures*. SAGE.

Gmail, L., & Hinton, G. (2008). *Visualizing Data using t-SNE*. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl

Gonçalves, E., Segura-Cabrera, A., Pacini, C., Picco, G., Behan, F. M., Jaaks, P., Coker, E. A., van der Meer, D., Barthorpe, A., Lightfoot, H., Mironenko, T., Beck, A., Richardson, L., Yang, W., Lleshi, E., Hall, J., Tolley, C., Hall, C., Mali, I., … Garnett, M. J. (2020). Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. *Molecular Systems Biology*, *16*(7), e9405.

Gonzaga-Jauregui, C., Lupski, J. R., & Gibbs, R. A. (2012). Human genome sequencing in health and disease. *Annual Review of Medicine*, *63*, 35–61.

Goruppi, S., Procopio, M.-G., Jo, S., Clocchiatti, A., Neel, V., & Dotto, G. P. (2017). The ULK3 Kinase Is Critical for Convergent Control of Cancer-Associated Fibroblast Activation by CSL and GLI. *Cell Reports*, *20*(10), 2468–2479.

Graham, D. B., & Xavier, R. J. (2020). Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature*, *578*(7796), 527–539.

Graves, P. R., & Haystead, T. A. J. (2002). Molecular biologist's guide to proteomics. *Microbiology and Molecular Biology Reviews: MMBR*, *66*(1), 39–63; table of contents.

Groeneveld, R. A., & Meeden, G. (1984). Measuring Skewness and Kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *33*(4), 391–399.

Guerra, I., Algorta, J., Díaz de Otazu, R., Pelayo, A., & Fariña, J. (2003). Immunohistochemical prognostic index for breast cancer in young women. *Molecular Pathology: MP*, *56*(6), 323–327.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B. M., Morris, J. S., Simon, I. M., Gerster, S., Fessler, E., De Sousa E Melo, F., Missiaglia, E., Ramay, H., Barras, D., … Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, *21*(11), 1350–1356.

Guo, Z., Zhao, M., Howard, E. W., Zhao, Q., Parris, A. B., Ma, Z., & Yang, X. (2017). Phenformin inhibits growth and epithelial-mesenchymal transition of ErbB2-overexpressing breast cancer cells through targeting the IGF1R pathway. *Oncotarget*, *8*(36), 60342–60357.

Gutteridge, W. E. (1985). Existing chemotherapy and its limitations. *British Medical Bulletin*, *41*(2), 162–168.

Haber, D. A., Bell, D. W., Sordella, R., Kwak, E. L., Godin-Heymann, N., Sharma, S. V., Lynch, T. J., & Settleman, J. (2005). Molecular targeted therapy of lung cancer: EGFR mutations and response to EGFR inhibitors. *Cold Spring Harbor Symposia on Quantitative Biology*, *70*, 419–426.

Hafner, M., Niepel, M., Chung, M., & Sorger, P. K. (2016). Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature Methods*, *13*(6), 521–527.

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J. W. L., & Quackenbush, J. (2013). Inconsistency in large pharmacogenomic studies. *Nature*, *504*(7480), 389–393.

Halsey, L. G. (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biology Letters*, *15*(5), 20190174.

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, *12*(1), 31–46.

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. In *Cell* (Vol. 100, Issue 1, pp. 57–70). https://doi.org/10.1016/s0092-8674(00)81683-9

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674.

Hao, Y.-H., Lafita-Navarro, M. C., Zacharias, L., Borenstein-Auerbach, N., Kim, M., Barnes, S., Kim, J., Shay, J., DeBerardinis, R. J., & Conacci-Sorrell, M. (2019). Induction of LEF1 by MYC activates the WNT pathway and maintains cell proliferation. *Cell Communication and Signaling*, *17*(1), 1–16.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, *28*(1), 100–108.

Hastie, T., Friedman, J., & Tibshirani, R. (n.d.). *The Elements of Statistical Learning*. Springer New York.

Haverty, P. M., Lin, E., Tan, J., Yu, Y., Lam, B., Lianoglou, S., Neve, R. M., Martin, S., Settleman, J., Yauch, R. L., & Bourgon, R. (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature*, *533*(7603), 333–337.

Hawkins, R. D., Hon, G. C., & Ren, B. (2010). Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, *11*(7), 476–486.

Henry, N. L., & Hayes, D. F. (2012). Cancer biomarkers. *Molecular Oncology*, *6*(2), 140–146.

Heppner, G. H., & Miller, B. E. (1983). Tumor heterogeneity: biological implications and therapeutic consequences. *Cancer Metastasis Reviews*, *2*(1), 5–23.

Herbst, R. S., Fukuoka, M., & Baselga, J. (2004). Gefitinib--a novel targeted approach to treating cancer. *Nature Reviews. Cancer*, *4*(12), 956–965.

Hientz, K., Mohr, A., Bhakta-Guha, D., & Efferth, T. (2017). The role of p53 in cancer drug resistance and targeted chemotherapy. *Oncotarget*, *8*(5), 8921–8946.

Hiley, C. T., & Swanton, C. (2014). Spatial and temporal cancer evolution: causes and consequences of tumour diversity. *Clinical Medicine*, *14 Suppl 6*, s33–s37.

Hills, S. A., & Diffley, J. F. X. (2014). DNA replication and oncogene-induced replicative stress. *Current Biology: CB*, *24*(10), R435–R444.

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, *6*. https://doi.org/10.7554/eLife.26726

Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Hoeijmakers, J. H. J. (2009). DNA damage, aging, and cancer. *The New England Journal of Medicine*, *361*(15), 1475–1485.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, *36*(3), 1171–1220.

Hollstein, M., Sidransky, D., Vogelstein, B., & Harris, C. C. (1991). p53 mutations in human cancers. *Science*, *253*(5015), 49–53.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, *4*(11), 682–690.

Hou, Y., Peng, Y., & Li, Z. (2022). Update on prognostic and predictive biomarkers of breast cancer. *Seminars in Diagnostic Pathology*, *39*(5), 322–332.

Huang, E. W., Bhope, A., Lim, J., Sinha, S., & Emad, A. (2020). Tissue-guided LASSO for prediction of clinical drug response using preclinical samples. *PLoS Computational Biology*, *16*(1), e1007607.

Huang, Y.-H., & Vakoc, C. R. (2016). A Biomarker Harvest from One Thousand Cancer Cell Lines. *Cell*, *166*(3), 536–537.

Hu, B., Gilkes, D. M., Farooqi, B., Sebti, S. M., & Chen, J. (2006). MDMX overexpression prevents p53 activation by the MDM2 inhibitor Nutlin. *The Journal of Biological Chemistry*, *281*(44), 33030–33035.

Hudis, C. A. (2007). Trastuzumab--mechanism of action and use in clinical practice. *The New England Journal of Medicine*, *357*(1), 39–51.

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nature Reviews. Genetics*, *6*(4), 287–298.

Hu, Y., Gu, X., Li, R., Luo, Q., & Xu, Y. (2010). Glycogen synthase kinase-3β inhibition induces nuclear factor-κB-mediated apoptosis in pediatric acute lymphocyte leukemia cells. *Journal of Experimental & Clinical Cancer Research*, *29*(1), 1–8.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93.

Ietswaart, R., Gyori, B. M., Bachman, J. A., Sorger, P. K., & Churchman, L. S. (2021). GeneWalk identifies relevant gene functions for a biological context using network representation learning. *Genome Biology*, *22*(1), 1–35.

Im, S., Choi, H. J., Yoo, C., Jung, J.-H., Jeon, Y.-W., Suh, Y. J., & Kang, C. S. (2013). Hedgehog related protein expression in breast cancer: gli-2 is associated with poor overall survival. *Korean Journal of Pathology*, *47*(2), 116–123.

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., van Dyk, E., Chang, H., de Silva, H., Heyn, H., Deng, X., Egan, R. K., Liu, Q., … Garnett, M. J. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, *166*(3), 740–754.

Iqbal, N., & Iqbal, N. (2014). Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications. *Molecular Biology International*, *2014*, 852748.

Isik, Z., Baldow, C., Cannistraci, C. V., & Schroeder, M. (2015). Drug target prioritization by perturbed gene expression and network information. *Scientific Reports*, *5*, 17417.

Janakiraman, M., Vakiani, E., Zeng, Z., Pratilas, C. A., Taylor, B. S., Chitale, D., Halilovic, E., Wilson, M., Huberman, K., Ricarte Filho, J. C., Persaud, Y., Levine, D. A., Fagin, J. A., Jhanwar, S. C., Mariadason, J. M., Lash, A., Ladanyi, M., Saltz, L. B., Heguy, A., … Solit, D. B. (2010). Genomic and biological characterization of exon 4 KRAS mutations in human cancer. *Cancer Research*, *70*(14), 5901–5911.

Janiaud, P., Serghiou, S., & Ioannidis, J. P. A. (2019). New clinical trial designs in the era of precision medicine: An overview of definitions, strengths, weaknesses, and current use in oncology. *Cancer Treatment Reviews*, *73*, 20–30.

Jiang, P., Sinha, S., Aldape, K., Hannenhalli, S., Sahinalp, C., & Ruppin, E. (2022). Big data in basic and translational cancer research. *Nature Reviews Cancer*, *22*(11), 625–639.

Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., PCAWG Tumor Subtypes and Clinical Translation Working Group, Danyi, A., de Ridder, J., van Herpen, C., Lolkema, M. P., Steeghs, N., Getz, G., Morris, Q. D., Stein, L. D., & PCAWG Consortium. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature Communications*, *11*(1), 728.

Jiao, X., Hooper, S. D., Djureinovic, T., Larsson, C., Wärnberg, F., Tellgren-Roth, C., Botling, J., & Sjöblom, T. (2013). Gene rearrangements in hormone receptor negative breast cancers revealed by mate pair sequencing. *BMC Genomics*, *14*, 165.

Jia, Q., Chu, H., Jin, Z., Long, H., & Zhu, B. (2022). High-throughput single-cell sequencing in cancer research. *Signal Transduction and Targeted Therapy*, *7*(1), 1–20.

Jia, Y., Yun, C.-H., Park, E., Ercan, D., Manuia, M., Juarez, J., Xu, C., Rhee, K., Chen, T., Zhang, H., Palakurthi, S., Jang, J., Lelais, G., DiDonato, M., Bursulaya, B., Michellys, P.-Y., Epple, R., Marsilje, T. H., McNeill, M., … Eck, M. J. (2016). Overcoming EGFR(T790M) and

EGFR(C797S) resistance with mutant-selective allosteric inhibitors. *Nature*, *534*(7605), 129–132.

Ji, X., Lu, Y., Tian, H., Meng, X., Wei, M., & Cho, W. C. (2019). Chemoresistance mechanisms of breast cancer and their countermeasures. *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie*, *114*, 108800.

Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(1), 183–189.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S.-M., Fu, B., Lin, M.-T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., … Kinzler, K. W. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, *321*(5897), 1801–1806.

Kagawa, K., Inoue, T., Tokino, T., Nakamura, Y., & Akiyama, T. (1997). Overexpression of GML promotes radiation-induced cell cycle arrest and apoptosis. *Biochemical and Biophysical Research Communications*, *241*(2), 481–485.

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M. D. M., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., & Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, *502*(7471), 333–339.

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., & Ishiguro-Watanabe, M. (2023). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, *51*(D1), D587–D592.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). An S4 package for kernel methods in R. *Reference Manual*.

Karplus, M., & McCammon, J. A. (1983). Dynamics of proteins: elements and function. *Annual Review of Biochemistry*, *52*, 263–300.

Kaur, G., & Dufour, J. M. (2012). Cell lines: Valuable tools or useless artifacts. *Spermatogenesis*, *2*(1), 1–5.

Kazandjian, D., Blumenthal, G. M., Yuan, W., He, K., Keegan, P., & Pazdur, R. (2016). FDA Approval of Gefitinib for the Treatment of Patients with Metastatic EGFR Mutation-Positive Non-Small Cell Lung Cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *22*(6), 1307–1312.

Kern, S. E., & Winter, J. M. (2006). Elegance, silence and nonsense in the mutations literature for solid tumors. *Cancer Biology & Therapy*, *5*(4), 349–359.

Kim, H., & Kim, Y.-M. (2018). Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types. *Scientific Reports*, *8*(1), 6041.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735–3745.

Kimura, Y., Furuhata, T., Shiratsuchi, T., Nishimori, H., Hirata, K., Nakamura, Y., & Tokino, T. (1997). GML sensitizes cancer cells to Taxol by induction of apoptosis. *Oncogene*, *15*(11), 1369–1374.

King, A. T., & Primrose, J. N. (2003). Principles of cancer treatment by surgery. *Surgery*, *21*(11), 284–288.

Kobayashi, S., Boggon, T. J., Dayaram, T., Jänne, P. A., Kocher, O., Meyerson, M., Johnson, B. E., Eck, M. J., Tenen, D. G., & Halmos, B. (2005). EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *The New England Journal of Medicine*, *352*(8), 786–792.

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., … Robinson, P. N. (2021). The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, *49*(D1), D1207–D1217.

Kondo, N., Takahashi, A., Ono, K., & Ohnishi, T. (2010). DNA damage induced by alkylating agents and repair pathways. *Journal of Nucleic Acids*, *2010*, 543531.

Kong, J., Lee, H., Kim, D., Han, S. K., Ha, D., Shin, K., & Kim, S. (2020). Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients. *Nature Communications*, *11*(1), 5485.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2021). Fast gene set enrichment analysis. In *bioRxiv* (p. 060012). https://doi.org/10.1101/060012

Korsmeyer, S. J. (1992). Chromosomal translocations in lymphoid malignancies reveal novel proto-oncogenes. *Annual Review of Immunology*, *10*, 785–807.

Krug, K., Jaehnig, E. J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., Miles, G., Mertins, P., Geffen, Y., Tang, L. C., Heiman, D. I., Cao, S., Maruvka, Y. E., Lei, J. T., Huang, C., Kothadia, R. B., Colaprico, A., Birger, C., Wang, J., … Clinical Proteomic Tumor Analysis Consortium. (2020). Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell*, *183*(5), 1436–1456.e31.

Kuiper, R. P., Ligtenberg, M. J. L., Hoogerbrugge, N., & Geurts van Kessel, A. (2010). Germline copy number variation and cancer risk. *Current Opinion in Genetics & Development*, *20*(3), 282–289.

Kulasingam, V., Pavlou, M. P., & Diamandis, E. P. (2010). Integrating high-throughput technologies in the quest for effective biomarkers for ovarian cancer. *Nature Reviews. Cancer*, *10*(5), 371–378.

Kumar, A. A., Van Laer, L., Alaerts, M., Ardeshirdavani, A., Moreau, Y., Laukens, K., Loeys, B., & Vandeweyer, G. (2018). pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion. *Bioinformatics* , *34*(13), 2254–2262.

Kunkel, T. A., & Bebenek, K. (2000). DNA replication fidelity. *Annual Review of Biochemistry*, *69*, 497–529.

Lai, M. I., Jiang, J., Silver, N., Best, S., Menzel, S., Garner, C., Weiss, M. J., & Thein, S. L. (2005). AHSP Is a Quantitative Trait Gene That Modifies the Phenotype of β Thalassemia. *Blood*, *106*(11), 3638–3638.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863.

Land, H., Parada, L. F., & Weinberg, R. A. (1983). Cellular oncogenes and multistep carcinogenesis. *Science (New York, N.Y.)*, *222*(4625), 771–778.

Lane, D. P. (1992). Cancer. p53, guardian of the genome. *Nature*, *358*(6381), 15–16.

Lee, A. J. X., Endesfelder, D., Rowan, A. J., Walther, A., Birkbak, N. J., Futreal, P. A., Downward, J., Szallasi, Z., Tomlinson, I. P. M., Howell, M., Kschischo, M., & Swanton, C. (2011). Chromosomal instability confers intrinsic multidrug resistance. *Cancer Research*, *71*(5), 1858–1870.

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., & Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, *21*(7), 1109–1121.

Lee, J., Kim, D., Kong, J., Ha, D., Kim, I., Park, M., Lee, K., Im, S.-H., & Kim, S. (2024). Cell-cell communication network-based interpretable machine learning predicts cancer patient response to immune checkpoint inhibitors. *Science Advances*, *10*(5), eadj0785.

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, *42*(1), 59–66.

Lee, S. H., Hu, W., Matulay, J. T., Silva, M. V., Owczarek, T. B., Kim, K., Chua, C. W., Barlow, L. J., Kandoth, C., Williams, A. B., Bergren, S. K., Pietzak, E. J., Anderson, C. B., Benson, M. C., Coleman, J. A., Taylor, B. S., Abate-Shen, C., McKiernan, J. M., Al-Ahmadie, H., … Shen, M. M. (2018). Tumor Evolution and Drug Response in Patient-Derived Organoid Models of Bladder Cancer. *Cell*, *173*(2), 515–528.e17.

Lei, X., Yang, X., & Fujita, H. (2019). Random walk based method to identify essential proteins by integrating network topology and biological characteristics. *Knowledge-Based Systems*, *167*, 53–67.

Lengauer, C., Kinzler, K. W., & Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature*, *396*(6712), 643–649.

Lever, J. (2016). Classification evaluation: it is important to understand both what a classification metric expresses and what it hides. *Nature Methods*, *13*, 603+.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., … Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS Biology*, *5*(10), e254.

Liang, L., Fang, J.-Y., & Xu, J. (2016). Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy. *Oncogene*, *35*(12), 1475–1482.

Li, A. R., Chitale, D., Riely, G. J., Pao, W., Miller, V. A., Zakowski, M. F., Rusch, V., Kris, M. G., & Ladanyi, M. (2008). EGFR mutations in lung adenocarcinomas: clinical testing experience and relationship to EGFR gene copy number and immunohistochemical expression. *The Journal of Molecular Diagnostics: JMD*, *10*(3), 242–248.

Li, C., Zhang, S., Lu, Y., Zhang, Y., Wang, E., & Cui, Z. (2013). The roles of Notch3 on the cell proliferation and apoptosis induced by CHIR99021 in NSCLC cell lines: a functional link between Wnt and Notch signaling pathways. *PloS One*, *8*(12), e84659.

Li, L., Xiang, Y., Zeng, Y., Xiao, B., Yu, W., Duan, C., Xia, X., Zhang, T., Zeng, Y., Liu, Y., & Dai, R. (2020). GSK-3β inhibition promotes doxorubicin-induced apoptosis in human cholangiocarcinoma cells via FAK/AKT inhibition. *Molecular Medicine Reports*, *22*(5), 4432–4441.

Lindstrom, M. L., & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, *46*(3), 673–687.

Ling, A., Gruener, R. F., Fessler, J., & Huang, R. S. (2018). More than fishing for a cure: The promises and pitfalls of high throughput cancer cell line screens. *Pharmacology & Therapeutics*, *191*, 178–189.

Listerman, I., Gazzaniga, F. S., & Blackburn, E. H. (2014). An investigation of the effects of the core protein telomerase reverse transcriptase on Wnt signaling in breast cancer cells. *Molecular and Cellular Biology*, *34*(2), 280–289.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), 523–536.

Liu, G., Wong, L., & Chua, H. N. (2009). Complex discovery from weighted PPI networks. *Bioinformatics* , *25*(15), 1891–1897.

Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., & Deng, L. (2021). Correction to "DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy." *Nucleic Acids Research*, *49*(18), 10801–10802.

Liu, Y., Wu, W., Cai, C., Zhang, H., Shen, H., & Han, Y. (2023). Patient-derived xenograft models in cancer therapy: technologies and applications. *Signal Transduction and Targeted Therapy*, *8*(1), 1–24.

Liu, Z., Ren, L., Liu, C., Xia, T., Zha, X., & Wang, S. (2015). Phenformin Induces Cell Cycle Change, Apoptosis, and Mesenchymal-Epithelial Transition and Regulates the AMPK/mTOR/p70s6k and MAPK/ERK Pathways in Breast Cancer Cells. *PloS One*, *10*(6), e0131207.

Li, Z.-H., Lei, L., Fei, L.-R., Huang, W.-J., Zheng, Y.-W., Yang, M.-Q., Wang, Z., Liu, C.-C., & Xu, H.-T. (2021). TRIP13 promotes the proliferation and invasion of lung cancer cells via the Wnt signaling pathway and epithelial-mesenchymal transition. *Journal of Molecular Histology*, *52*(1), 11–20.

Li, Z., Ivanov, A. A., Su, R., Gonzalez-Pecchi, V., Qi, Q., Liu, S., Webber, P., McMillan, E., Rusnak, L., Pham, C., Chen, X., Mo, X., Revennaugh, B., Zhou, W., Marcus, A., Harati, S., Chen, X., Johns, M. A., White, M. A., … Fu, H. (2017). The OncoPPi network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies. *Nature Communications*, *8*, 14356.

Lobentanzer, S., Aloy, P., Baumbach, J., Bohar, B., Carey, V. J., Charoentong, P., Danhauser, K., Doğan, T., Dreo, J., Dunham, I., Farr, E., Fernandez-Torras, A., Gyori, B. M., Hartung, M., Hoyt, C. T., Klein, C., Korcsmaros, T., Maier, A., Mann, M., … Saez-Rodriguez, J. (2023). Democratizing knowledge representation with BioCypher. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01848-y

Loeb, L. A., Springgate, C. F., & Battula, N. (1974). Errors in DNA replication as a basis of malignant changes. *Cancer Research*, *34*(9), 2311–2321.

Lord, C. J., & Ashworth, A. (2012). The DNA damage response and cancer therapy. *Nature*, *481*(7381), 287–294.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., Harris, P. L., Haserlat, S. M., Supko, J. G., Haluska, F. G., Louis, D. N., Christiani, D. C., Settleman, J., & Haber, D. A. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *The New England Journal of Medicine*, *350*(21), 2129–2139.

Majumdar, A., Liu, Y., Lu, Y., Wu, S., & Cheng, L. (2021). kESVR: An Ensemble Model for Drug Response Prediction in Precision Medicine Using Cancer Cell Lines Gene Expression. *Genes*, *12*(6). https://doi.org/10.3390/genes12060844

Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L., & Siu, L. L. (2020). Molecular profiling for precision cancer therapies. *Genome Medicine*, *12*(1), 8.

Mann, M., & Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology*, *21*(3), 255–261.

Manzari, M. T., Shamay, Y., Kiguchi, H., Rosen, N., Scaltriti, M., & Heller, D. A. (2021). Targeted drug delivery strategies for precision medicines. *Nature Reviews. Materials*, *6*(4), 351–370.

Martens, M., Ammar, A., Riutta, A., Waagmeester, A., Slenter, D. N., Hanspers, K., A Miller, R., Digles, D., Lopes, E. N., Ehrhart, F., Dupuis, L. J., Winckers, L. A., Coort, S. L., Willighagen, E. L., Evelo, C. T., Pico, A. R., & Kutmon, M. (2021). WikiPathways: connecting communities. *Nucleic Acids Research*, *49*(D1), D613–D621.

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., Gonzalez-Perez, A., & Lopez-Bigas, N. (2020). A compendium of mutational cancer driver genes. *Nature Reviews. Cancer*, *20*(10), 555–572.

Marvalim, C., Datta, A., & Lee, S. C. (2023). Role of p53 in breast cancer progression: An insight into p53 targeted therapy. *Theranostics*, *13*(4), 1421–1442.

Massagué, J. (2008). TGFβ in Cancer. *Cell*, *134*(2), 215–230.

Ma, T., & Zhang, A. (2019). Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). *BMC Genomics*, *20*(Suppl 11), 944.

Maurer, G., Tarkowski, B., & Baccarini, M. (2011). Raf kinases in cancer–roles and therapeutic opportunities. *Oncogene*, *30*(32), 3477–3488.

McCulloch, S. D., & Kunkel, T. A. (2008). The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Research*, *18*(1), 148–161.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

McDermott, U., Sharma, S. V., Dowell, L., Greninger, P., Montagut, C., Lamb, J., Archibald, H., Raudales, R., Tam, A., Lee, D., Rothenberg, S. M., Supko, J. G., Sordella, R., Ulkus, L. E., Iafrate, A. J., Maheswaran, S., Njauw, C. N., Tsao, H., Drew, L., … Settleman, J. (2007). Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(50), 19936–19941.

McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R., & Mirny, L. A. (2013). Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(8), 2910–2915.

Medzihradszky, K. F., Campbell, J. M., Baldwin, M. A., Falick, A. M., Juhasz, P., Vestal, M. L., & Burlingame, A. L. (2000). The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Analytical Chemistry*, *72*(3), 552–558.

Melton, C., Reuter, J. A., Spacek, D. V., & Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*, *47*(7), 710–716.

Menden, M. P., Casale, F. P., Stephan, J., Bignell, G. R., Iorio, F., McDermott, U., Garnett, M. J., Saez-Rodriguez, J., & Stegle, O. (2018). The germline genetic component of drug sensitivity in cancer cell lines. *Nature Communications*, *9*(1), 3385.

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

Menden, M. P., Wang, D., Mason, M. J., Szalai, B., Bulusu, K. C., Guan, Y., Yu, T., Kang, J., Jeon, M., Wolfinger, R., Nguyen, T., Zaslavskiy, M., AstraZeneca-Sanger Drug Combination DREAM Consortium, Jang, I. S., Ghazoui, Z., Ahsen, M. E., Vogel, R., Neto, E. C., Norman, T., … Saez-Rodriguez, J. (2019). Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nature Communications*, *10*(1), 2674.

Milano, M., Agapito, G., & Cannataro, M. (2022). Challenges and Limitations of Biological Network Analysis. *Biotech (Basel (Switzerland))*, *11*(3). https://doi.org/10.3390/biotech11030024

Mitsudomi, T., & Yatabe, Y. (2010). Epidermal growth factor receptor in relation to tumor development: EGFR gene and cancer. *The FEBS Journal*, *277*(2), 301–308.

Mohamed, F. E. Z. A., Abdelaziz, A. O., Kasem, A. H., Ellethy, T., & Gayyed, M. F. (2021). Thyroid hormone receptor α1 acts as a new squamous cell lung cancer diagnostic marker and poor prognosis predictor. *Scientific Reports*, *11*(1), 7944.

Montgomery, D. C., Peck, E. A., & Geoffrey Vining, G. (2021). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Mooney, C. Z., Duval, R. D., & Duvall, R. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference*. SAGE.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., & Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(11), 4245–4250.

Morrison, J. L., Breitling, R., Higham, D. J., & Gilbert, D. R. (2005). GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, *6*, 233.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., & Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, *9 Suppl 1*(Suppl 1), S4.

Motulsky, H. J., & Brown, R. E. (2006). Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate. *BMC Bioinformatics*, *7*(1), 1–20.

Mo, X., Niu, Q., Ivanov, A. A., Tsang, Y. H., Tang, C., Shu, C., Li, Q., Qian, K., Wahafu, A., Doyle, S. P., Cicka, D., Yang, X., Fan, D., Reyna, M. A., Cooper, L. A. D., Moreno, C. S., Zhou, W., Owonikoko, T. K., Lonial, S., … Fu, H. (2022). Systematic discovery of mutation-directed neo-protein-protein interactions in cancer. *Cell*, *185*(11), 1974–1985.e12.

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., & López-Bigas, N. (2016). OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biology*, *17*(1), 128.

Müller, J. (1838). *Ueber den feineren Bau und die Formen der krankhaften Geschwülste*.

Murphree, A. L., & Benedict, W. F. (1984). Retinoblastoma: clues to human oncogenesis. *Science*, *223*(4640), 1028–1033.

Nair, N. U., Greninger, P., Zhang, X., Friedman, A. A., Amzallag, A., Cortez, E., Sahu, A. D., Lee, J. S., Dastur, A., Egan, R. K., Murchie, E., Ceribelli, M., Crowther, G. S., Beck, E., McClanaghan, J., Klump-Thomas, C., Boisvert, J. L., Damon, L. J., Wilson, K. M., … Benes, C. H. (2023). A landscape of response to drug combinations in non-small cell lung cancer. *Nature Communications*, *14*(1), 1–19.

Nakopoulou, L. L., Alexiadou, A., Theodoropoulos, G. E., Lazaris, A. C., Tzonou, A., & Keramopoulos, A. (1996). Prognostic significance of the co-expression of p53 and c-erbB-2 proteins in breast cancer. *The Journal of Pathology*, *179*(1), 31–38.

Narayan, M., Wilken, J. A., Harris, L. N., Baron, A. T., Kimbler, K. D., & Maihle, N. J. (2009). Trastuzumab-Induced HER Reprogramming in "Resistant" Breast Carcinoma Cells. *Cancer Research*, *69*(6), 2191–2194.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, *231*(694-706), 289–337.

Nguyen, L. C., Naulaerts, S., Bruna, A., Ghislat, G., & Ballester, P. J. (2021). Predicting Cancer Drug Response In Vivo by Learning an Optimal Feature Selection of Tumour Molecular Profiles. *Biomedicines*, *9*(10), 1319.

Nicora, G., Vitali, F., Dagliati, A., Geifman, N., & Bellazzi, R. (2020). Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools. *Frontiers in Oncology*, *10*, 530980.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjærde, O. C., Langerød, A., Ringnér, M., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*(7605), 47–54.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, *194*(4260), 23–28.

Nuncia-Cantarero, M., Martinez-Canales, S., Andrés-Pretel, F., Santpere, G., Ocaña, A., & Galan-Moya, E. M. (2018). Functional transcriptomic annotation and protein-protein interaction network analysis identify NEK2, BIRC5, and TOP2A as potential targets in obese patients with luminal A breast cancer. *Breast Cancer Research and Treatment*, *168*(3), 613–623.

O'Connor, M. J. (2015). Targeting the DNA Damage Response in Cancer. *Molecular Cell*, *60*(4), 547–560.

O'Flaherty, L., Shnyder, S. D., Cooper, P. A., Cross, S. J., Wakefield, J. G., Pardo, O. E., Seckl, M. J., & Tavaré, J. M. (2019). Tumor growth suppression using a combination of taxol-based therapy and GSK3 inhibition in non-small cell lung cancer. *PloS One*, *14*(4), e0214610.

Olivier, M., Hollstein, M., & Hainaut, P. (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor Perspectives in Biology*, *2*(1), a001008.

O'Neil, J., Benita, Y., Feldman, I., Chenard, M., Roberts, B., Liu, Y., Li, J., Kral, A., Lejnine, S., Loboda, A., Arthur, W., Cristescu, R., Haines, B. B., Winter, C., Zhang, T., Bloecher, A., & Shumway, S. D. (2016). An Unbiased Oncology Compound Screen to Identify Novel Combination Strategies. *Molecular Cancer Therapeutics*, *15*(6), 1155–1162.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., & Tyers, M. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science: A Publication of the Protein Society*, *30*(1), 187–200.

Ozaki, T., & Nakagawara, A. (2011). Role of p53 in Cell Death and Human Cancers. *Cancers*, *3*(1), 994–1013.

Paczkowska, M., Barenboim, J., Sintupisut, N., Fox, N. S., Zhu, H., Abd-Rabbo, D., Mee, M. W., Boutros, P. C., PCAWG Drivers and Functional Interpretation Working Group, Reimand, J., & PCAWG Consortium. (2020). Integrative pathway enrichment analysis of multivariate omics data. *Nature Communications*, *11*(1), 735.

Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F. J., Lindeman, N., Boggon, T. J., Naoki, K., Sasaki, H., Fujii, Y., Eck, M. J., Sellers, W. R., Johnson, B. E., & Meyerson, M. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science*, *304*(5676), 1497–1500.

Panagiotakos, D. B. (2008). Value of p-value in biomedical research. *The Open Cardiovascular Medicine Journal*, *2*, 97–99.

Pandey, A., & Mann, M. (2000). Proteomics to study genes and genomes. *Nature*, *405*(6788), 837–846.

Pao, W., Miller, V. A., Politi, K. A., Riely, G. J., Somwar, R., Zakowski, M. F., Kris, M. G., & Varmus, H. (2005). Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Medicine*, *2*(3), e73.

Pao, W., Miller, V., Zakowski, M., Doherty, J., Politi, K., Sarkaria, I., Singh, B., Heelan, R., Rusch, V., Fulton, L., Mardis, E., Kupfer, D., Wilson, R., Kris, M., & Varmus, H. (2004). EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences*, *101*(36), 13306–13311.

Park, J.-I., Venteicher, A. S., Hong, J. Y., Choi, J., Jun, S., Shkreli, M., Chang, W., Meng, Z., Cheung, P., Ji, H., McLaughlin, M., Veenstra, T. D., Nusse, R., McCrea, P. D., & Artandi, S. E. (2009). Telomerase modulates Wnt signalling by association with target gene chromatin. *Nature*, *460*(7251), 66–72.

Park, J.-S., Lee, C., Kim, H.-K., Kim, D., Son, J. B., Ko, E., Cho, J.-H., Kim, N.-D., Nan, H.-Y., Kim, C.-Y., Yoon, S., Lee, S.-H., & Choi, H. G. (2016). Suppression of the metastatic spread of breast cancer by DN10764 (AZD7762)-mediated inhibition of AXL signaling. *Oncotarget*, *7*(50), 83308–83318.

Pascual, M., Mena-Varas, M., Robles, E. F., Garcia-Barchino, M.-J., Panizo, C., Hervas-Stubbs, S., Alignani, D., Sagardoy, A., Martinez-Ferrandis, J. I., Bunting, K. L., Meier, S., Sagaert, X., Bagnara, D., Guruceaga, E., Blanco, O., Celay, J., Martínez-Baztan, A., Casares, N., Lasarte, J. J., … Roa, S. (2019). PD-1/PD-L1 immune checkpoint and p53 loss facilitate tumor progression in activated B-cell diffuse large B-cell lymphomas. *Blood*, *133*(22), 2401–2412.

Pavletich, N. P., Chambers, K. A., & Pabo, C. O. (1993). The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots. *Genes & Development*, *7*(12B), 2556–2564.

Pawlak, K., Błażej, P., Mackiewicz, D., & Mackiewicz, P. (2023). The Influence of the Selection at the Amino Acid Level on Synonymous Codon Usage from the Viewpoint of Alternative Genetic Codes. *International Journal of Molecular Sciences*, *24*(2). https://doi.org/10.3390/ijms24021185

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., & Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747–752.

Peterson, C. B., Stingo, F. C., & Vannucci, M. (2015). Bayesian Inference of Multiple Gaussian Graphical Models. *Journal of the American Statistical Association*, *110*(509), 159–174.

Petronek, M. S., Spitz, D. R., & Allen, B. G. (2021). Iron-Sulfur Cluster Biogenesis as a Critical Target in Cancer. *Antioxidants (Basel, Switzerland)*, *10*(9). https://doi.org/10.3390/antiox10091458

Picard, R. R., & Cook, R. D. (1984). Cross-Validation of Regression Models. *Journal of the American Statistical Association*, *79*(387), 575–583.

Piccart-Gebhart, M. J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., Smith, I., Gianni, L., Baselga, J., Bell, R., Jackisch, C., Cameron, D., Dowsett, M., Barrios, C. H., Steger, G., Huang, C.-S., Andersson, M., Inbar, M., Lichinitser, M., Láng, I., … Herceptin Adjuvant (HERA) Trial Study Team. (2005). Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *The New England Journal of Medicine*, *353*(16), 1659–1672.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., … Stratton, M. R. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, *463*(7278), 191–196.

Polakis, P. (2005). Deregulated GSK3β activity in colorectal cancer: Its association with tumor cell survival and proliferation. *Biochemical and Biophysical Research Communications*, *334*(4), 1365–1373.

Pozdeyev, N., Yoo, M., Mackie, R., Schweppe, R. E., Tan, A. C., & Haugen, B. R. (2016). Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. *Oncotarget*, *7*(32), 51619–51625.

Pray, L. (2008). Discovery of DNA structure and function: Watson and Crick. *Nature Education*, *1*(1), 100.

Puisieux, A., Brabletz, T., & Caramel, J. (2014). Oncogenic roles of EMT-inducing transcription factors. *Nature Cell Biology*, *16*(6), 488–494.

Pulido-Tamayo, S., Weytjens, B., De Maeyer, D., & Marchal, K. (2016). SSA-ME Detection of cancer driver genes using mutual exclusivity by small subnetwork analysis. *Scientific Reports*, *6*, 36257.

Pu, L., Singha, M., Wu, H.-C., Busch, C., Ramanujam, J., & Brylinski, M. (2022). An integrated network representation of multiple cancer-specific data for graph-based machine learning. *NPJ Systems Biology and Applications*, *8*(1), 14.

Quintás-Cardama, A., & Cortes, J. (2009). Molecular biology of bcr-abl1-positive chronic myeloid leukemia. *Blood*, *113*(8), 1619–1630.

Rabbitts, T. H. (1994). Chromosomal translocations in human cancer. *Nature*, *372*(6502), 143–149.

Raina, P., Guinea, R., Chatsirisupachai, K., Lopes, I., Farooq, Z., Guinea, C., Solyom, C.-A., & de Magalhães, J. P. (2023). GeneFriends: gene co-expression databases and tools for humans and model organisms. *Nucleic Acids Research*, *51*(D1), D145–D158.

Rajapakse, V. N., Luna, A., Yamade, M., Loman, L., Varma, S., Sunshine, M., Iorio, F., Sousa, F. G., Elloumi, F., Aladjem, M. I., Thomas, A., Sander, C., Kohn, K. W., Benes, C. H., Garnett, M., Reinhold, W. C., & Pommier, Y. (2018). CellMinerCDB for Integrative Cross-Database Genomics and Pharmacogenomics Analyses of Cancer Cell Lines. *iScience*, *10*, 247–264.

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. MIT Press.

Rastogi, R. P., Richa, Kumar, A., Tyagi, M. B., & Sinha, R. P. (2010). Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair. *Journal of Nucleic Acids*, *2010*, 592980.

Ravandi, F., Cortes, J. E., Jones, D., Faderl, S., Garcia-Manero, G., Konopleva, M. Y., O'Brien, S., Estrov, Z., Borthakur, G., Thomas, D., Pierce, S. R., Brandt, M., Byrd, A., Bekele, B. N., Pratz, K., Luthra, R., Levis, M., Andreeff, M., & Kantarjian, H. M. (2010). Phase I/II study of combination therapy with sorafenib, idarubicin, and cytarabine in younger patients with acute myeloid leukemia. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *28*(11), 1856–1862.

R Core Team. (2016). R: A language and environment for statistical computing. Version 3.6. 0. Vienna, Austria. */tool/81287/ra-Language-and-Environment-for-Statistical …*.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D., & Bader, G. D. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, *14*(2), 482–517.

Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J. M., Juul, R. I., Lin, Z., Feuerbach, L., Sabarinathan, R., Madsen, T., Kim, J., Mularoni, L., Shuai, S., Lanzós, A., Herrmann, C., Maruvka, Y. E., … PCAWG Consortium. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, *578*(7793), 102–111.

Rimawi, M. F., Schiff, R., & Osborne, C. K. (2015). Targeting HER2 for the treatment of breast cancer. *Annual Review of Medicine*, *66*, 111–128.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140.

Rockwell, S. (1980). In vivo-in vitro tumour cell lines: characteristics and limitations as models for human cancer. *The British Journal of Cancer. Supplement*, *4*, 118–122.

Rodon, J., Soria, J.-C., Berger, R., Miller, W. H., Rubin, E., Kugel, A., Tsimberidou, A., Saintigny, P., Ackerstein, A., Braña, I., Loriot, Y., Afshar, M., Miller, V., Wunder, F., Bresson, C., Martini, J.-F., Raynaud, J., Mendelsohn, J., Batist, G., … Kurzrock, R. (2019). Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. *Nature Medicine*, *25*(5), 751–758.

Rodriguez-Pinilla, S. M., Sarrio, D., Moreno-Bueno, G., Rodriguez-Gil, Y., Martinez, M. A., Hernandez, L., Hardisson, D., Reis-Filho, J. S., & Palacios, J. (2007). Sox2: a possible driver of the basal-like phenotype in sporadic breast cancer. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *20*(4), 474–481.

Romond, E. H., Perez, E. A., Bryant, J., Suman, V. J., Geyer, C. E., Jr, Davidson, N. E., Tan-Chiu, E., Martino, S., Paik, S., Kaufman, P. A., Swain, S. M., Pisansky, T. M., Fehrenbacher, L., Kutteh, L. A., Vogel, V. G., Visscher, D. W., Yothers, G., Jenkins, R. B., Brown, A. M., … Wolmark, N. (2005). Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *The New England Journal of Medicine*, *353*(16), 1673–1684.

Roses, A. D. (2000). Pharmacogenetics and the practice of medicine. *Nature*, *405*(6788), 857–865.

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., Ostos, L. C. G., Lannon, W. A., Grotzinger, C., Del Rio, M., Lhermitte, B., Olshen, A. B., Wiedenmann, B., Cantley, L. C., Gray, J. W., & Hanahan, D. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine*, *19*(5), 619–625.

Safikhani, Z., Smirnov, P., Thu, K. L., Silvester, J., El-Hachem, N., Quevedo, R., Lupien, M., Mak, T. W., Cescon, D., & Haibe-Kains, B. (2017). Gene isoforms as expression-based biomarkers predictive of drug response in vitro. *Nature Communications*, *8*(1), 1126.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., Dimitriadoy, S., Liu, D. L., Kantheti, H. S., Saghafinia, S., Chakravarty, D., Daian, F., Gao, Q., Bailey, M. H., Liang, W.-W., Foltz, S. M., Shmulevich, I., Ding, L., Heins, Z., … Schultz, N. (2018). Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*, *173*(2), 321–337.e10.

Santaguida, S., & Amon, A. (2015). Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nature Reviews. Molecular Cell Biology*, *16*(8), 473–485.

Santarpia, L., Lippman, S. M., & El-Naggar, A. K. (2012). Targeting the MAPK–RAS–RAF signaling pathway in cancer therapy. *Expert Opinion on Therapeutic Targets*, *16*(1), 103–119.

Sarkans, U., Gostev, M., Athar, A., Behrangi, E., Melnichuk, O., Ali, A., Minguet, J., Rada, J. C., Snow, C., Tikhonov, A., Brazma, A., & McEntyre, J. (2018). The BioStudies database-one stop shop for all data supporting a life sciences study. *Nucleic Acids Research*, *46*(D1), D1266–D1270.

Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews. Genetics*, *12*(10), 683–691.

Savage, D. G., & Antman, K. H. (2002). Imatinib mesylate--a new oral targeted therapy. *The New England Journal of Medicine*, *346*(9), 683–693.

Sawyers, C. (2004). Targeted cancer therapy. *Nature*, *432*(7015), 294–297.

Sawyers, C. L. (2008). The cancer biomarker problem. *Nature*, *452*(7187), 548–552.

Schilsky, R. L. (2010). Personalized medicine in oncology: the future is now. *Nature Reviews. Drug Discovery*, *9*(5), 363–366.

Schirrmacher, V. (2019). From chemotherapy to biological therapy: A review of novel concepts to reduce the side effects of systemic cancer treatment (Review). *International Journal of Oncology*, *54*(2), 407–419.

Schmucker, R., Farina, G., Faeder, J., Fröhlich, F., Saglam, A. S., & Sandholm, T. (2021). Combination treatment optimization using a pan-cancer pathway model. *PLoS Computational Biology*, *17*(12), e1009689.

Schubert, O. T., Röst, H. L., Collins, B. C., Rosenberger, G., & Aebersold, R. (2017). Quantitative proteomics: challenges and opportunities in basic and applied research. *Nature Protocols*, *12*(7), 1289–1294.

Schulte-Sasse, R., Budach, S., Hnisz, D., & Marsico, A. (2021). Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence*, *3*(6), 513–526.

Schwab, M. (1998). Amplification of oncogenes in human cancer cells. In *BioEssays* (Vol. 20, Issue 6, pp. 473–479). https://doi.org/10.1002/(sici)1521-1878(199806)20:6<473::aid-bies5>3.0.co;2-n

Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., Cokol, M., Price, E. V., Coletti, M. E., Jones, V., Bodycombe, N. E., Soule, C. K., Gould, J., Alexander, B., Li, A., Montgomery, P., Wawer, M. J., Kuru, N., Kotz, J. D., Hon, C. S.-Y., Munoz, B., Liefeld, T., … Schreiber, S. L. (2015).

Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery*, *5*(11), 1210–1223.

Seoane, J., & De Mattos-Arruda, L. (2014). The challenge of intratumour heterogeneity in precision medicine. *Journal of Internal Medicine*, *276*(1), 41–51.

Sevimoglu, T., & Arga, K. Y. (2014). The role of protein interaction networks in systems biomedicine. *Computational and Structural Biotechnology Journal*, *11*(18), 22–27.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504.

Sharma, P., Hu-Lieskovan, S., Wargo, J. A., & Ribas, A. (2017). Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. *Cell*, *168*(4), 707–723.

Sharma, S. V., Haber, D. A., & Settleman, J. (2010). Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nature Reviews. Cancer*, *10*(4), 241–253.

Shigematsu, H., & Gazdar, A. F. (2006). Somatic mutations of epidermal growth factor receptor signaling pathway in lung cancers. *International Journal of Cancer. Journal International Du Cancer*, *118*(2), 257–262.

Shijie, L., Zhen, P., Kang, Q., Hua, G., Qingcheng, Y., & Dongdong, C. (2021). Deregulation of CLTC interacts with TFG, facilitating osteosarcoma via the TGF-beta and AKT/mTOR signaling pathways. *Clinical and Translational Medicine*, *11*(6), e377.

Shlien, A., & Malkin, D. (2009). Copy number variations and cancer. *Genome Medicine*, *1*(6), 62.

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews. Cancer*, *6*(10), 813–823.

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*(1), 7–33.

Sjögren, S., Inganäs, M., Lindgren, A., Holmberg, L., & Bergh, J. (1998). Prognostic and predictive value of c-erbB-2 overexpression in primary breast cancer, alone and in combination with other prognostic markers. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *16*(2), 462–469.

Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., & McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, *235*(4785), 177–182.

Son, J., Lyssiotis, C. A., Ying, H., Wang, X., Hua, S., Ligorio, M., Perera, R. M., Ferrone, C. R., Mullarky, E., Shyh-Chang, N., Kang, Y. ’an, Fleming, J. B., Bardeesy, N., Asara, J. M., Haigis, M. C., DePinho, R. A., Cantley, L. C., & Kimmelman, A. C. (2013). Glutamine supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. *Nature*, *496*(7443), 101–105.

Soria, J.-C., Mok, T. S., Cappuzzo, F., & Jänne, P. A. (2012). EGFR-mutated oncogene-addicted non-small cell lung cancer: current trends and future prospects. *Cancer Treatment Reviews*, *38*(5), 416–430.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A.-L., & Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, *100*(14), 8418–8423.

Sosman, J. A., Kim, K. B., Schuchter, L., Gonzalez, R., Pavlick, A. C., Weber, J. S., McArthur, G. A., Hutson, T. E., Moschos, S. J., Flaherty, K. T., Hersey, P., Kefford, R., Lawrence, D., Puzanov, I., Lewis, K. D., Amaravadi, R. K., Chmielowski, B., Lawrence, H. J., Shyr, Y., … Ribas, A. (2012). Survival in BRAF V600–Mutant Advanced Melanoma Treated with Vemurafenib. *The New England Journal of Medicine*, *366*(8), 707–714.

Sos, M. L., Koker, M., Weir, B. A., Heynck, S., Rabinovsky, R., Zander, T., Seeger, J. M., Weiss, J., Fischer, F., Frommolt, P., Michel, K., Peifer, M., Mermel, C., Girard, L., Peyton, M., Gazdar, A. F., Minna, J. D., Garraway, L. A., Kashkar, H., … Thomas, R. K. (2009). PTEN loss contributes to erlotinib resistance in EGFR-mutant lung cancer by activation of Akt and EGFR. *Cancer Research*, *69*(8), 3256–3261.

Sotillo, R., Schvartzman, J.-M., Socci, N. D., & Benezra, R. (2010). Mad2-induced chromosome instability leads to lung tumour relapse after oncogene withdrawal. *Nature*, *464*(7287), 436–440.

Soussi, T., & Wiman, K. G. (2015). TP53: an oncogene in disguise. *Cell Death and Differentiation*, *22*(8), 1239–1249.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., … Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, *122*(6), 957–968.

Storchova, Z., & Pellman, D. (2004). From polyploidy to aneuploidy, genome instability and cancer. *Nature Reviews. Molecular Cell Biology*, *5*(1), 45–54.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719–724.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, *5*(6), 463–466.

STUDENT. (1908). THE PROBABLE ERROR OF A MEAN. *Biometrika*, *6*(1), 1–25.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

Su, J., Yoon, B.-J., & Dougherty, E. R. (2010). Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network. *BMC Bioinformatics*, *11 Suppl 6*(Suppl 6), S8.

Su, J., Zhong, W., Zhang, X., Huang, Y., Yan, H., Yang, J., Dong, Z., Xie, Z., Zhou, Q., Huang, X., Lu, D., Yan, W., & Wu, Y.-L. (2017). Molecular characteristics and clinical outcomes of EGFR exon 19 indel subtypes to EGFR TKIs in NSCLC patients. *Oncotarget*, *8*(67), 111246–111257.

Sundvall, M., Iljin, K., Kilpinen, S., Sara, H., Kallioniemi, O.-P., & Elenius, K. (2008). Role of ErbB4 in Breast Cancer. *Journal of Mammary Gland Biology and Neoplasia*, *13*(2), 259–268.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., & Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, *156*(6), 1324–1335.

Swain, S. M., Baselga, J., Kim, S.-B., Ro, J., Semiglazov, V., Campone, M., Ciruelos, E., Ferrero, J.-M., Schneeweiss, A., Heeson, S., Clark, E., Ross, G., Benyunes, M. C., Cortés, J., & CLEOPATRA Study Group. (2015). Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. *The New England Journal of Medicine*, *372*(8), 724–734.

Swanton, C., Nicke, B., Schuett, M., Eklund, A. C., Ng, C., Li, Q., Hardcastle, T., Lee, A., Roy, R., East, P., Kschischo, M., Endesfelder, D., Wylie, P., Kim, S. N., Chen, J.-G., Howell, M., Ried, T., Habermann, J. K., Auer, G., … Downward, J. (2009). Chromosomal instability determines taxane response. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(21), 8671–8676.

Swift, L. H., & Golsteyn, R. M. (2014). Genotoxic anti-cancer agents and their relationship to DNA damage, mitosis, and checkpoint adaptation in proliferating cancer cells. *International Journal of Molecular Sciences*, *15*(3), 3403–3431.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. von. (2018). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613.

Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A.-J. L., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L., Carr, S. A., Clauser, K. R., Jaffe, J. D., Kowalski, K. A., Neubert, T. A., Regnier, F. E., Schilling, B., Tegeler, T. J., Wang, M., … Spiegelman, C. (2010). Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *Journal of Proteome Research*, *9*(2), 761–776.

Tan, C., & Du, X. (2012). KRAS mutation testing in metastatic colorectal cancer. *World Journal of Gastroenterology: WJG*, *18*(37), 5171–5180.

Tannock, I. F. (1998). Conventional cancer therapy: promise broken or promise delayed? *The Lancet*, *351 Suppl 2*, SII9–SII16.

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., & Wrana, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, *27*(2), 199–204.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *58*(1), 267–288.

Timpe, L. C., Li, D., Yen, T.-Y., Wong, J., Yen, R., Macher, B. A., & Piryatinska, A. (2015). Mining the Breast Cancer Proteome for Predictors of Drug Sensitivity. *Journal of Proteomics & Bioinformatics*, *8*(9), 204–211.

Tohme, S., Simmons, R. L., & Tsung, A. (2017). Surgery for Cancer: A Trigger for Metastases. *Cancer Research*, *77*(7), 1548–1552.

Tolaney, S. M., Barry, W. T., Dang, C. T., Yardley, D. A., Moy, B., Marcom, P. K., Albain, K. S., Rugo, H. S., Ellis, M., Shapira, I., Wolff, A. C., Carey, L. A., Overmoyer, B. A., Partridge, A. H., Guo, H., Hudis, C. A., Krop, I. E., Burstein, H. J., & Winer, E. P. (2015). Adjuvant paclitaxel and trastuzumab for node-negative, HER2-positive breast cancer. *The New England Journal of Medicine*, *372*(2), 134–141.

Toledo, F., & Wahl, G. M. (2006). Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nature Reviews. Cancer*, *6*(12), 909–923.

Tomczak, A., Mortensen, J. M., Winnenburg, R., Liu, C., Alessi, D. T., Swamy, V., Vallania, F., Lofgren, S., Haynes, W., Shah, N. H., Musen, M. A., & Khatri, P. (2018). Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Scientific Reports*, *8*(1), 5115.

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, *19*(1A), A68–A77.

Topol, E. J. (2015). Network-based approaches for drug response prediction and targeted therapy development in cancer. *Biochemical and Biophysical Research Communications*, *464*(2), 386–391.

Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., Meyers, R. M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W. F. J., Howell, S., Merkel, E., … Hahn, W. C. (2017). Defining a Cancer Dependency Map. *Cell*, *170*(3), 564–576.e16.

Tsimberidou, A. M., Fountzilas, E., Nikanjam, M., & Kurzrock, R. (2020). Review of precision cancer medicine: Evolution of the treatment paradigm. *Cancer Treatment Reviews*, *86*, 102019.

Tsimberidou, A.-M., Iskander, N. G., Hong, D. S., Wheler, J. J., Falchook, G. S., Fu, S., Piha-Paul, S., Naing, A., Janku, F., Luthra, R., Ye, Y., Wen, S., Berry, D., & Kurzrock, R. (2012). Personalized medicine in a phase I clinical trials program: the MD Anderson Cancer Center initiative. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *18*(22), 6373–6383.

Tsuruo, T., Naito, M., Tomida, A., Fujita, N., Mashima, T., Sakamoto, H., & Haga, N. (2003). Molecular targeting therapy of cancer: drug resistance, apoptosis and survival signal. *Cancer Science*, *94*(1), 15–21.

Turajlic, S., & Swanton, C. (2016). Metastasis as an evolutionary process. *Science*, *352*(6282), 169–175.

Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., Ölbei, M., Gábor, A., Theis, F., Módos, D., Korcsmáros, T., & Saez-Rodriguez, J. (2021). Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, *17*(3), e9923.

Turner, N. C., & Reis-Filho, J. S. (2006). Basal-like breast cancer and the BRCA1 phenotype. *Oncogene*, *25*(43), 5846–5853.

Tyanova, S., Albrechtsen, R., Kronqvist, P., Cox, J., Mann, M., & Geiger, T. (2016). Proteomic maps of breast cancer subtypes. *Nature Communications*, *7*, 10259.

Valdeolivas, A., Gabor, A., Turei, D., & Saez-Rodriguez, J. (2019). *OmnipathR: an R client for the OmniPath web service*. Bioconductor. https://saezlab.github.io/OmnipathR/articles/omnipath_intro.html

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E., & Baudot, A. (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* , *35*(3), 497–505.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews. Drug Discovery*, *18*(6), 463–477.

Van Cutsem, E., Köhne, C.-H., Hitre, E., Zaluski, J., Chang Chien, C.-R., Makhson, A., D'Haens, G., Pintér, T., Lim, R., Bodoky, G., Roh, J. K., Folprecht, G., Ruff, P., Stroh, C., Tejpar, S., Schlichting, M., Nippgen, J., & Rougier, P. (2009). Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *The New England Journal of Medicine*, *360*(14), 1408–1417.

van Staveren, W. C. G., Solís, D. Y. W., Hébrant, A., Detours, V., Dumont, J. E., & Maenhaut, C. (2009). Human cancer cell lines: Experimental models for cancer cells in situ? For cancer stem cells? *Biochimica et Biophysica Acta*, *1795*(2), 92–103.

Vasan, N., Baselga, J., & Hyman, D. M. (2019). A view on drug resistance in cancer. *Nature*, *575*(7782), 299–309.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., & Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* , *26*(12), i237–i245.

Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C., Fotouhi, N., & Liu, E. A. (2004). In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science*, *303*(5659), 844–848.

Velho, T. R. (2012). Metastatic melanoma - a review of current and future drugs. *Drugs in Context*, *2012*, 212242.

Villegas-Comonfort, S., Serna-Marquez, N., Galindo-Hernandez, O., Navarro-Tito, N., & Salazar, E. P. (2012). Arachidonic acid induces an increase of β-1,4-galactosyltransferase I expression in MDA-MB-231 breast cancer cells. *Journal of Cellular Biochemistry*, *113*(11), 3330–3341.

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., Assmus, H. E., Andrade-Navarro, M. A., & Wanker, E. E. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Science Signaling*, *4*(189), rs8.

Vinayagam, A., Zirin, J., Roesel, C., Hu, Y., Yilmazel, B., Samsonova, A. A., Neumüller, R. A., Mohr, S. E., & Perrimon, N. (2014). Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. *Nature Methods*, *11*(1), 94–99.

Virchow, R. (1858). *Die Cellularpathologie in ihrer Begründung auf physiologische und pathologische Gewebelehre: 20 Vorlesungen, gehalten während d. Monate Febr., März u. April 1858 im Patholog. Inst. zu Berlin*. Hirschwald.

Vis, D. J., Bombardelli, L., Lightfoot, H., Iorio, F., Garnett, M. J., & Wessels, L. F. A. (2016). Multilevel models improve precision and speed of IC_{50} estimates. In *Pharmacogenomics* (Vol. 17, Issue 7, pp. 691–700). https://doi.org/10.2217/pgs.16.15

Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine*, *10*(8), 789–799.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr, & Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, *339*(6127), 1546–1558.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, *417*(6887), 399–403.

von Minckwitz, G., Procter, M., de Azambuja, E., Zardavas, D., Benyunes, M., Viale, G., Suter, T., Arahmani, A., Rouchet, N., Clark, E., Knott, A., Lang, I., Levy, C., Yardley, D. A., Bines, J., Gelber, R. D., Piccart, M., Baselga, J., & APHINITY Steering Committee and Investigators. (2017). Adjuvant Pertuzumab and Trastuzumab in Early HER2-Positive Breast Cancer. *The New England Journal of Medicine*, *377*(2), 122–131.

Walther, A., Houlston, R., & Tomlinson, I. (2008). Association between chromosomal instability and prognosis in colorectal cancer: a meta-analysis. *Gut*, *57*(7), 941–950.

Wang, D., Hensman, J., Kutkaite, G., Toh, T. S., Galhoz, A., GDSC Screening Team, Dry, J. R., Saez-Rodriguez, J., Garnett, M. J., Menden, M. P., & Dondelinger, F. (2020). A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates. *eLife*, *9*. https://doi.org/10.7554/eLife.60352

Wang, J., Liang, J., Zheng, W., Zhao, X., & Mu, J. (2019). Protein complex detection algorithm based on multiple topological characteristics in PPI networks. *Information Sciences*, *489*, 78–92.

Wang, J., Luo, X.-X., Tang, Y.-L., Xu, J.-X., & Zeng, Z.-G. (2019). The prognostic values of insulin-like growth factor binding protein in breast cancer. *Medicine*, *98*(19), e15561.

Wang, Q., Liu, W., Zhou, H., Lai, W., Hu, C., Dai, Y., Li, G., Zhang, R., & Zhao, Y. (2024). Tozasertib activates anti-tumor immunity through decreasing regulatory T cells in melanoma. *Neoplasia* , *48*, 100966.

Wang, Y., Jadhav, A., Southal, N., Huang, R., & Nguyen, D.-T. (2010). A grid algorithm for high throughput fitting of dose-response curve data. *Current Chemical Genomics*, *4*, 57–66.

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

Wathieu, H., Issa, N. T., Fernandez, A. I., Mohandoss, M., Tiek, D. M., Franke, J. L., Byers, S. W., Riggins, R. B., & Dakshanamurthy, S. (2017). Differential prioritization of therapies to subtypes of triple negative breast cancer using a systems medicine method. *Oncotarget*, *8*(54), 92926–92942.

Wee, P., & Wang, Z. (2017). Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers*, *9*(5), 52.

Weinberg, R. A. (1995). The retinoblastoma protein and cell cycle control. *Cell*, *81*(3), 323–330.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., & Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genetics*, *46*(11), 1160–1165.

Weinreich, S. S., Mangon, R., Sikkens, J. J., Teeuw, M. E. en, & Cornel, M. C. (2008). Orphanet: a European database for rare diseases. *Nederlands tijdschrift voor geneeskunde*, *152*(9), 518–519.

Weinstein, J. N. (2012, March 28). *Cell lines battle cancer*. Nature Publishing Group UK. https://doi.org/10.1038/483544a

Winograd, B., Peckham, M., & Pinedo, H. M. (2013). *Human Tumour Xenografts in Anticancer Drug Development*. Springer Science & Business Media.

Workman, P., Aboagye, E. O., Balkwill, F., Balmain, A., Bruder, G., Chaplin, D. J., Double, J. A., Everitt, J., Farningham, D. A. H., Glennie, M. J., Kelland, L. R., Robinson, V., Stratford, I. J., Tozer, G. M., Watson, S., Wedge, S. R., Eccles, S. A., & Committee of the National Cancer Research Institute. (2010). Guidelines for the welfare and use of animals in cancer research. *British Journal of Cancer*, *102*(11), 1555–1577.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.

Xia, F., Allen, J., Balaprakash, P., Brettin, T., Garcia-Cardona, C., Clyde, A., Cohn, J., Doroshow, J., Duan, X., Dubinkina, V., Evrard, Y., Fan, Y. J., Gans, J., He, S., Lu, P., Maslov, S., Partin, A., Shukla, M., Stahlberg, E., … Stevens, R. (2022). A cross-study analysis of drug response prediction in cancer cell lines. *Briefings in Bioinformatics*, *23*(1). https://doi.org/10.1093/bib/bbab356

Yamamoto, S., Yamamoto-Ibusuki, M., Yamamoto, Y., Fujiwara, S., & Iwase, H. (2013). A comprehensive analysis of Aurora A; transcript levels are the most reliable in association with proliferation and prognosis in breast cancer. *BMC Cancer*, *13*, 217.

Yang, H., Robinson, P. N., & Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature Methods*. https://idp.nature.com/authorize/casa?redirect_uri=https://www.nature.com/articles/nmeth.3484&casa_token=DgoH9OyKcQYAAAAA:nNOnaE2TOUj2uT38wpg1NtI9DGP6LDPZqHGoQymuuTlh8mj3OYVVfTPjKsd1DnT424CtDvwtLG8eSO7Nxw

Yan, J., Risacher, S. L., Shen, L., & Saykin, A. J. (2018). Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Briefings in Bioinformatics*, *19*(6), 1370–1381.

Yarden, Y., & Sliwkowski, M. X. (2001). Untangling the ErbB signalling network. *Nature Reviews. Molecular Cell Biology*, *2*(2), 127–137.

Yekutieli, D. (2008). Hierarchical False Discovery Rate–Controlling Methodology. *Journal of the American Statistical Association*, *103*(481), 309–316.

Yin, J., Lin, C., Jiang, M., Tang, X., Xie, D., Chen, J., & Ke, R. (2021). CENPL, ISG20L2, LSM4, MRPL3 are four novel hub genes and may serve as diagnostic and prognostic markers in breast cancer. *Scientific Reports*, *11*(1), 15610.

Yuan, Y., Liu, X., Cai, Y., & Li, W. (2022). Lapatinib and lapatinib plus trastuzumab therapy versus trastuzumab therapy for HER2 positive breast cancer patients: an updated systematic review and meta-analysis. *Systematic Reviews*, *11*(1), 1–30.

Yun, C.-H., Mengwasser, K. E., Toms, A. V., Woo, M. S., Greulich, H., Wong, K.-K., Meyerson, M., & Eck, M. J. (2008). The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proceedings of the National Academy of Sciences*, *105*(6), 2070–2075.

Zabludoff, S. D., Deng, C., Grondine, M. R., Sheehy, A. M., Ashwell, S., Caleb, B. L., Green, S., Haye, H. R., Horn, C. L., Janetka, J. W., Liu, D., Mouchet, E., Ready, S., Rosenthal, J. L., Queva, C., Schwartz, G. K., Taylor, K. J., Tse, A. N., Walker, G. E., & White, A. M. (2008). AZD7762, a novel checkpoint kinase inhibitor, drives checkpoint abrogation and potentiates DNA-targeted therapies. *Molecular Cancer Therapeutics*, *7*(9), 2955–2966.

Zardavas, D., Cameron, D., Krop, I., & Piccart, M. (2013). Beyond trastuzumab and lapatinib: new options for HER2-positive breast cancer. *American Society of Clinical Oncology Educational Book. American Society of Clinical Oncology. Annual Meeting*. https://doi.org/10.14694/EdBook_AM.2013.33.e2

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L. D., & Ferretti, V. (2019). The International Cancer Genome Consortium Data Portal. *Nature Biotechnology*, *37*(4), 367–369.

Zhang, N., Li, Y., Sundquist, J., Sundquist, K., & Ji, J. (2023). Identifying actionable druggable targets for breast cancer: Mendelian randomization and population-based analyses. *EBio-Medicine*, *98*, 104859.

Zhang, Y., Chen, F., & Chandrashekar, D. S. (2022). Proteogenomic characterization of 2002 human cancers reveals pan-cancer molecular subtypes and associated pathways. *Nature*. https://www.nature.com/articles/s41467-022-30342-3

Zhang, Z., Song, J., Tang, J., Xu, X., & Guo, F. (2018). Detecting complexes from edge-weighted PPI networks via genes expression analysis. *BMC Systems Biology*, *12*(Suppl 4), 40.

Zhao, Y., Aguilar, A., Bernard, D., & Wang, S. (2015). Small-molecule inhibitors of the MDM2-p53 protein-protein interaction (MDM2 Inhibitors) in clinical trials for cancer treatment. *Journal of Medicinal Chemistry*, *58*(3), 1038–1052.

Zheng, H., Saito, H., Masuda, S., Yang, X., & Takano, Y. (2007). Phosphorylated GSK3beta-ser9 and EGFR are good prognostic factors for lung carcinomas. *Anticancer Research*, *27*(5B), 3561–3569.

Zheng, J. (2013). Oncogenic chromosomal translocations and human cancer (review). *Oncology Reports*, *30*(5), 2011–2019.

Zhou, H., Liu, J., Li, J., & Duan, W. (2017). A density-based approach for detecting complexes in weighted PPI networks by semantic similarity. *PloS One*, *12*(7), e0180570.

Zhou, X., Menche, J., Barabási, A.-L., & Sharma, A. (2014). Human symptoms-disease network. *Nature Communications*, *5*, 4212.

Zhou, Z.-H. (2011). *Ensemble methods: Foundations and algorithms*. Chapman & Hall/CRC. https://doi.org/10.1201/b12207

Appendix A: Supplementary information to chapters 2.2 and 3.1

A.1 Resistance markers in GDSC screen

Tissue	Drug	N. resistant outliers	UNRES p-value	UNRES FDR	UNRES cell lines	Putative resistance markers
BRCA	AST-1306	1	2.00e-04	4.50e-02	UACC-812	CHEK2 Mut, 12q15 Amplification (MDM2,NUP107), 1p12 Amplification (NOTCH2)
BRCA	AST-1306	2	1.80e-03	5.93e-02	UACC-812, HCC1569	NA
BRCA	AST-1306	3	2.90e-03	5.93e-02	UACC-812, HCC1569, MDA-MB-361	NA
COREAD	PLX-4720	4	2.80e-03	5.93e-02	KM12, LS-513, SW1417, RKO	BRAF_mut-p.A712T, BRAF_mut-p.A404fs*9, BRAF_mut-p.E204V, BRAF_mut-p.E204*
LUAD	Gefitinib	1	7.00e-04	5.93e-02	NCI-H1975	EGFR_mut-p.T790M, 11q13.3 Amplification (CCND1)
LUAD	Gefitinib	2	1.30e-03	5.93e-02	NCI-H1975, NCI-H1650	EGFR_mut-p.T790M
LUAD	Afatinib	1	2.30e-03	5.93e-02	NCI-H1650	NA
LUAD	Afatinib	2	1.90e-03	5.93e-02	NCI-H1650, NCI-H1975	EGFR_mut-p.T790M
OV	GSK690693	1	1.20e-03	5.93e-02	TOV-21G	PIK3CA_mut-p.H1047Y, KRAS Mut, LARP4B Mut, NF1 Mut, NSD1 Mut, PIK3R1 Mut, PTEN Mut, RNF43 Mut, STK11 Mut
OV	GSK690693	2	2.90e-03	5.93e-02	TOV-21G, OAW-42	PIK3CA_mut-p.H1047Y, PIK3CA_mut-p.H1047L
SKCM	SB590885	1	1.80e-03	5.93e-02	SK-MEL-31	NA
BRCA	CP724714	4	4.80e-03	7.12e-02	EFM-192A, JIMT-1, MDA-MB-330, HCC1569	NA
BRCA	AST-1306	4	5.30e-03	7.12e-02	UACC-812, HCC1569, MDA-MB-361, HCC1954	NA
COREAD	PLX-4720	3	4.90e-03	7.12e-02	KM12, LS-513, SW1417	BRAF_mut-p.A712T, BRAF_mut-p.A404fs*9, BRAF_mut-p.E204V, BRAF_mut-p.E204*
LUAD	Gefitinib	2	5.50e-03	7.12e-02	NCI-H1650, NCI-H1975	EGFR_mut-p.T790M, EGFR_mut-p.L858R
LUAD	AZD3759	2	5.70e-03	7.12e-02	NCI-H1975, NCI-H1650	EGFR_mut-p.T790M, EGFR_mut-p.L858R
SKCM	Camptothecin	1	4.60e-03	7.12e-02	UACC-257	NA
SKCM	Nutlin-3a (-)	5	5.70e-03	7.12e-02	Hs940-T, SK-MEL-2, CP66-MEL, Mewo, IGR-37	11q22.3 Deletion (CASP1)
BRCA	CP724714	3	6.90e-03	7.47e-02	EFM-192A, JIMT-1, MDA-MB-330	NA
BRCA	CP724714	5	7.30e-03	7.47e-02	EFM-192A, JIMT-1, MDA-MB-330, HCC1569, UACC-812	NA
BRCA	CP724714	3	6.50e-03	7.47e-02	EFM-192A, HCC1428, MCF7	NA
BRCA	CP724714	4	7.30e-03	7.47e-02	EFM-192A, HCC1428, MCF7, MDA-MB-361	NA
BRCA	CP724714	2	8.10e-03	7.92e-02	EFM-192A, JIMT-1	NA

BRCA	AST-1306	5	9.10e-03	7.95e-02	UACC-812, HCC1569, MDA-MB-361, HCC1954, MDA-MB-453	NA
COREAD	PLX-4720	5	9.90e-03	7.95e-02	KM12, LS-513, SW1417, RKO, SNU-C5	BRAF_mut-p.A712T, BRAF_mut-p.A404fs*9, BRAF_mut-p.E204V, BRAF_mut-p.E204*
LUAD	Afatinib	1	9.20e-03	7.95e-02	NCI-H1975	EGFR_mut-p.T790M, 11q13.3 Amplification (CCND1)
LUAD	Osimertinib	1	9.60e-03	7.95e-02	NCI-H1650	13q34 Deletion (TFDP1)
SKCM	Dabrafenib	3	8.60e-03	7.95e-02	SK-MEL-31, COLO-800, VMRC-MELG	NA
SKCM	Nutlin-3a (-)	4	1.27e-02	9.85e-02	Hs940-T, SK-MEL-2, CP66-MEL, Mewo	NA
LUAD	Gefitinib	3	1.32e-02	9.90e-02	NCI-H1975, NCI-H1650, PC-14	EGFR_mut-p.T790M
LGG	Daporinad	1	1.41e-02	1.02e-01	SW1783	PTEN Mut
SKCM	SB590885	2	1.50e-02	1.03e-01	SK-MEL-31, UACC-62	NA
SKCM	Dabrafenib	4	1.51e-02	1.03e-01	SK-MEL-31, COLO-800, VMRC-MELG, WM1552C	NA
THCA	Dabrafenib	1	1.70e-02	1.12e-01	8505C	NF2 Mut
SKCM	Dabrafenib	2	1.91e-02	1.23e-01	SK-MEL-31, COLO-800	NA
BRCA	CP724714	1	2.38e-02	1.30e-01	EFM-192A	NA
BRCA	Amuvatinib	4	2.54e-02	1.30e-01	EFM-192A, HCC1428, MDA-MB-361, HCC1419	NA
LUAD	Afatinib	3	2.36e-02	1.30e-01	NCI-H1650, NCI-H1975, PC-14	EGFR_mut-p.T790M
LUAD	Erlotinib	2	2.49e-02	1.30e-01	NCI-H1975, NCI-H1650	EGFR_mut-p.T790M, EGFR_mut-p.L858R
OV	GSK690693	3	2.17e-02	1.30e-01	TOV-21G, OAW-42, OC-314	PIK3CA_mut-p.H1047Y, PIK3CA_mut-p.H1047L, PIK3CA_mut-p.R108H
SKCM	SB590885	3	2.33e-02	1.30e-01	SK-MEL-31, UACC-62, RPMI-7951	NA
SKCM	Dabrafenib	5	2.49e-02	1.30e-01	SK-MEL-31, COLO-800, VMRC-MELG, WM1552C, HMV-II	NA
SKCM	Nutlin-3a (-)	2	2.38e-02	1.30e-01	Hs940-T, SK-MEL-2	NA
SKCM	Nutlin-3a (-)	3	2.22e-02	1.30e-01	Hs940-T, SK-MEL-2, CP66-MEL	NA
BRCA	Ibrutinib	2	2.66e-02	1.30e-01	MDA-MB-361, JIMT-1	NA
SKCM	PLX-4720	5	2.63e-02	1.30e-01	COLO-800, WM35, UACC-62, HMV-II, SK-MEL-31	NA
SKCM	Nutlin-3a (-)	1	2.86e-02	1.37e-01	Hs940-T	BAP1 Mut
BRCA	Lapatinib	1	3.24e-02	1.41e-01	UACC-812	CHEK2 Mut, Lack of TP53 Mut, 12q15 Amplification (MDM2,NUP107), Lack of 17q22 Amplification (CLTC,PPM1D), 20p12.1 Amplification (CRNKL1,FOXA2), 1p12 Amplification (NOTCH2)
BRCA	CP724714	2	3.25e-02	1.41e-01	EFM-192A, HCC1428	NA
LUAD	Sapatinib	2	3.27e-02	1.41e-01	NCI-H1650, NCI-H1975	EGFR_mut-p.T790M
LUAD	Osimertinib	2	3.26e-02	1.41e-01	NCI-H1650, PC-3 [JPC-3]	NA
SKCM	Dabrafenib	1	3.10e-02	1.41e-01	SK-MEL-31	NA
COREAD	PLX-4720	2	3.34e-02	1.42e-01	KM12, LS-513	BRAF_mut-p.A712T, BRAF_mut-p.A404fs*9, BRAF_mut-p.E204V, BRAF_mut-p.E204*

Table 11: UNRES cell lines discovered in GDSC drug screen.

A.2 Resistance markers in CTRP screen

Tissue	Drug	N. resistant outliers	UNRES p-value	UNRES FDR	UNRES cell lines	Putative resistance markers
BRCA	neratinib	2	1.40e-03	4.00e-02	JIMT-1, MDA-MB-361	NA
BRCA	neratinib	3	1.20e-03	4.00e-02	JIMT-1, MDA-MB-361, HCC1569	NA
BRCA	neratinib	4	1.80e-03	4.00e-02	JIMT-1, MDA-MB-361, HCC1569, MDA-MB-453	NA
LUAD	afatinib	1	1.90e-03	4.00e-02	NCI-H1650	13q34 Deletion (TFDP1)
LUAD	WZ8040	1	2.00e-03	4.00e-02	NCI-H1650	13q34 Deletion (TFDP1)
BRCA	ZSTK474	1	3.90e-03	4.33e-02	JIMT-1	4q34.1 Deletion (FAT1,IRF2)
BRCA	ZSTK474	3	3.80e-03	4.33e-02	JIMT-1, BT-474, HCC202	NA
COREAD	tipifarnib-P2	1	3.90e-03	4.33e-02	RKO	AKAP9_mut-p.K37E, AKAP9_mut-p.?, BRAF Mut, NUP98 Mut, ZNRF3 Mut
SKCM	dasatinib	2	3.90e-03	4.33e-02	COLO-792, MEL-JUSO	ARID2_mut-p.R274*, ARID2_mut-p.K119fs*31, ARID2_mut-p.L409*, Lack of ARID1A Mut, Lack of BRAF Mut, Lack of 12p12.3 Amplification (KRAS)
BRCA	neratinib	5	5.70e-03	5.18e-02	JIMT-1, MDA-MB-361, HCC1569, MDA-MB-453, HCC202	NA
BRCA	ZSTK474	4	5.20e-03	5.18e-02	JIMT-1, BT-474, HCC202, HCC2218	NA
BRCA	ZSTK474	5	6.30e-03	5.25e-02	JIMT-1, BT-474, HCC202, HCC2218, ZR-75-30	NA
LUAD	gefitinib	2	7.10e-03	5.37e-02	NCI-H1650, NCI-H1975	EGFR_mut-p.T790M, EGFR_mut-p.L858R, 9p21.3 Deletion (CDKN2A), Lack of 7p11.2 Amplification (EGFR)
LUAD	erlotinib	2	8.50e-03	5.37e-02	NCI-H1650, NCI-H1975	EGFR_mut-p.T790M, EGFR_mut-p.L858R, 9p21.3 Deletion (CDKN2A), Lack of 7p11.2 Amplification (EGFR)
LUAD	afatinib	2	8.40e-03	5.37e-02	NCI-H1650, NCI-H1975	EGFR_mut-p.T790M, EGFR_mut-p.L858R, 9p21.3 Deletion (CDKN2A), Lack of 7p11.2 Amplification (EGFR)
SKCM	dasatinib	2	8.60e-03	5.37e-02	COLO-792, Mewo	NF1_mut-p.W1236R, NF1_mut-p.?, NF1_mut-p.Q1336*, Lack of ARID1A Mut, Lack of BRAF Mut, Lack of 12p12.3 Amplification (KRAS)
COREAD	tipifarnib-P2	2	9.20e-03	5.41e-02	RKO, CW-2	AKAP9_mut-p.K37E, AKAP9_mut-p.?, AKAP9_mut-p.K2021fs*3, AKAP9_mut-p.N2045fs*3, AKAP9_mut-p.Y2858C,

						FAT1 Mut, Lack of FBXW7 Mut, FXR1 Mut
BRCA	ZSTK474	2	1.11e-02	5.55e-02	JIMT-1, BT-474	CNA: MED24, 13q11 Deletion (PABPC3,SACS,ZMYM2)
SKCM	GSK461364	1	1.07e-02	5.55e-02	COLO-792	ARID2_mut-p.R274*, ASPM Mut, ASXL2 Mut, PHLPP1 Mut
SKCM	dacarbazine	2	1.09e-02	5.55e-02	COLO-792, Mewo	NF1_mut-p.W1236R, NF1_mut-p.?, NF1_mut-p.Q1336*, Lack of ARID1A Mut, Lack of BRAF Mut, Lack of 12p12.3 Amplification (KRAS)
BRCA	neratinib	1	1.34e-02	6.32e-02	JIMT-1	4q34.1 Deletion (FAT1,IRF2)
DLBC	BRD-K13999467	1	1.39e-02	6.32e-02	OCI-LY-19	MLL2_mut-p.S2597fs*94, MLL2_mut-p.M1379fs*52, ARID1A Mut
LUAD	saracatinib	2	1.70e-02	7.08e-02	NCI-H1975, NCI-H1650	EGFR_mut-p.T790M, EGFR_mut-p.L858R, 9p21.3 Deletion (CDKN2A), Lack of 7p11.2 Amplification (EGFR)
SKCM	BRD-K51490254	2	1.64e-02	7.08e-02	SK-MEL-31, SK-MEL-28	Lack of ARID1A Mut, Lack of ARID2 Mut, Lack of NF1 Mut
LUAD	canertinib	1	1.86e-02	7.44e-02	NCI-H1650	13q34 Deletion (TFDP1)
SKCM	simvastatin	1	2.71e-02	1.04e-01	COLO-792	ARID2_mut-p.R274*, ASPM Mut, ASXL2 Mut, PHLPP1 Mut
SKCM	BRD-K51490254	2	3.02e-02	1.12e-01	Mewo, COLO-792	NF1_mut-p.Q1336*, NF1_mut-p.W1236R, NF1_mut-p.?, Lack of ARID1A Mut, Lack of BRAF Mut, Lack of 12p12.3 Amplification (KRAS)
COREAD	KU-55933	2	3.59e-02	1.12e-01	RKO, CW-2	AKAP9_mut-p.K37E, AKAP9_mut-p.?, AKAP9_mut-p.K2021fs*3, AKAP9_mut-p.N2045fs*3, AKAP9_mut-p.Y2858C, FAT1 Mut, Lack of FBXW7 Mut, FXR1 Mut
LUAD	WZ8040	2	3.16e-02	1.12e-01	NCI-H1650, PC-14	Lack of 8q24.21 Amplification (MYC)
SKCM	BI-2536	1	3.33e-02	1.12e-01	COLO-792	ARID2_mut-p.R274*, ASPM Mut, ASXL2 Mut, PHLPP1 Mut
SKCM	dasatinib	1	3.59e-02	1.12e-01	COLO-792	NF1_mut-p.W1236R, NF1_mut-p.?, ASPM Mut, ASXL2 Mut, PHLPP1 Mut
SKCM	ISOX	2	3.58e-02	1.12e-01	COLO-792, Mewo	NF1_mut-p.W1236R, NF1_mut-p.?, NF1_mut-p.Q1336*, Lack of ARID1A Mut, Lack of BRAF Mut, Lack of

						12p12.3 Amplification (KRAS)
SKCM	pevonedistat	2	3.76e-02	1.14e-01	COLO-792, Mewo	NF1_mut-p.W1236R, NF1_mut-p.?, NF1_mut-p.Q1336*, Lack of ARID1A Mut, Lack of BRAF Mut, Lack of 12p12.3 Amplification (KRAS)
DLBC	BRD-K13999467	2	4.11e-02	1.21e-01	OCI-LY-19, NU-DUL-1	MLL2_mut-p.S2597fs*94, MLL2_mut-p.M1379fs*52, MLL2_mut-p.D1215fs*115
LUAD	QW-BI-011	1	4.88e-02	1.39e-01	NCI-H1838	CLSPN Mut, LPHN2 Mut, NF1 Mut, 9p21.3 Deletion (CDKN2A), 7p11.2 Amplification (EGFR)
BRCA	neratinib	2	1.40e-03	4.00e-02	JIMT-1, MDA-MB-361	NA
BRCA	neratinib	3	1.20e-03	4.00e-02	JIMT-1, MDA-MB-361, HCC1569	NA
BRCA	neratinib	4	1.80e-03	4.00e-02	JIMT-1, MDA-MB-361, HCC1569, MDA-MB-453	NA
LUAD	afatinib	1	1.90e-03	4.00e-02	NCI-H1650	13q34 Deletion (TFDP1)
LUAD	WZ8040	1	2.00e-03	4.00e-02	NCI-H1650	13q34 Deletion (TFDP1)
BRCA	ZSTK474	1	3.90e-03	4.33e-02	JIMT-1	4q34.1 Deletion (FAT1,IRF2)
BRCA	ZSTK474	3	3.80e-03	4.33e-02	JIMT-1, BT-474, HCC202	NA
COREAD	tipifarnib-P2	1	3.90e-03	4.33e-02	RKO	AKAP9_mut-p.K37E, AKAP9_mut-p.?, BRAF Mut, NUP98 Mut, ZNRF3 Mut
SKCM	dasatinib	2	3.90e-03	4.33e-02	COLO-792, MEL-JUSO	ARID2_mut-p.R274*, ARID2_mut-p.K119fs*31, ARID2_mut-p.L409*, Lack of ARID1A Mut, Lack of BRAF Mut, Lack of 12p12.3 Amplification (KRAS)
BRCA	neratinib	5	5.70e-03	5.18e-02	JIMT-1, MDA-MB-361, HCC1569, MDA-MB-453, HCC202	NA
BRCA	ZSTK474	4	5.20e-03	5.18e-02	JIMT-1, BT-474, HCC202, HCC2218	NA
BRCA	ZSTK474	5	6.30e-03	5.25e-02	JIMT-1, BT-474, HCC202, HCC2218, ZR-75-30	NA
LUAD	gefitinib	2	7.10e-03	5.37e-02	NCI-H1650, NCI-H1975	EGFR_mut-p.T790M, EGFR_mut-p.L858R, 9p21.3 Deletion (CDKN2A), Lack of 7p11.2 Amplification (EGFR)
LUAD	erlotinib	2	8.50e-03	5.37e-02	NCI-H1650, NCI-H1975	EGFR_mut-p.T790M, EGFR_mut-p.L858R, 9p21.3 Deletion (CDKN2A), Lack of 7p11.2 Amplification (EGFR)
LUAD	afatinib	2	8.40e-03	5.37e-02	NCI-H1650, NCI-H1975	EGFR_mut-p.T790M, EGFR_mut-p.L858R, 9p21.3 Deletion (CDKN2A), Lack of 7p11.2 Amplification (EGFR)

SKCM	dasatinib	2	8.60e-03	5.37e-02	COLO-792, Mewo	NF1_mut-p.W1236R, NF1_mut-p.?, NF1_mut-p.Q1336*, Lack of ARID1A Mut, Lack of BRAF Mut, Lack of 12p12.3 Amplification (KRAS)
COREAD	tipifarnib-P2	2	9.20e-03	5.41e-02	RKO, CW-2	AKAP9_mut-p.K37E, AKAP9_mut-p.?, AKAP9_mut-p.K2021fs*3, AKAP9_mut-p.N2045fs*3, AKAP9_mut-p.Y2858C, FAT1 Mut, Lack of FBXW7 Mut, FXR1 Mut
BRCA	ZSTK474	2	1.11e-02	5.55e-02	JIMT-1, BT-474	CNA: MED24, 13q11 Deletion (PABPC3,SACS,ZMYM2)

Table 12: UNRES cell lines discovered in CTRP drug screen.

Appendix B: Supplementary information to chapters 2.3 and 3.2

B.1 Increasing viability markers using ANOVA with linear slope

Tissue	Drug	Drug ID	alteration	Diff. slope MT vs WT	Mean slope MT	p-adjusted value	Nr. high slope
COREAD	Cytarabine	1006	APC_mut	0,138654	-0,19895	0,000826	0
LAML	TW 37	1149	NRAS_mut	0,161032	-0,31581	0,000877	0
LIHC	Mirin	1048	TP53_mut	0,0578	-0,1728	0,003135	0
DLBC	PFI-1	1219	CREBBP_mut	0,239473	-0,08665	0,005974	0
STAD	Elesclomol	1031	TP53_mut	0,107866	-0,21114	0,006797	0
BRCA	Shikonin	170	ASH1L_mut	0,192817	-0,26716	0,006939	0
LUSC	IGF1R_3801	1738	cnaLUSC10	0,078662	-0,19333	0,007793	0
LAML	PAK_5339	1730	TP53_mut	0,103478	-0,27093	0,007857	0
LGG	LDN-193189	478	cnaLGG7	0,175882	-0,30592	0,00814	0
BLCA	AZD5582	1617	ARID1A_mut	0,074037	-0,17896	0,011173	0
STAD	BMS-536924	1091	cnaSTAD13	0,154696	-0,21662	0,011715	0
ESCA	CPI-613	415	cnaESCA12	0,099089	-0,25739	0,011814	0
HNSC	SNX-2112	328	cnaHNSC2	0,096384	-0,2426	0,012711	0
BRCA	Tanespimycin	1026	cnaBRCA26	0,122406	-0,19667	0,012798	0
SKCM	Mitoxantrone	1810	NRAS_mut	0,185041	-0,05122	0,013097	0
SKCM	Mitoxantrone	1810	cnaSKCM27	0,182066	-0,05364	0,013533	0
KIRC	Dyrk1b_0191	1407	TP53_mut	0,073097	-0,21935	0,014079	0
PAAD	(5Z)-7-Oxozeaenol	1242	SMAD4_mut	0,094195	-0,23221	0,01545	0
GBM	Vinorelbine	140	TP53_mut	0,08802	-0,16471	0,017123	0
LAML	Serdemetan	1133	TP53_mut	0,136631	-0,22271	0,017903	0
ESCA	Shikonin	170	EGFR_mut	0,100084	-0,39295	0,02005	0
COREAD	Belinostat	274	BPTF_mut	0,139139	-0,28304	0,021936	0
ESCA	CAY10603	276	NFE2L2_mut	0,159045	-0,27649	0,02261	0
LIHC	Vorinostat	1012	TP53_mut	0,149029	-0,19609	0,023962	0
DLBC	Entinostat	1593	EP300_mut	0,108243	-0,23744	0,024825	0
ESCA	NSC-207895	269	cnaESCA11	0,058029	-0,19473	0,026027	0
HNSC	Midostaurin	153	A_mut	0,292148	0,053966	0,026829	0
GBM	Doxorubicin	133	cnaGBM101	0,232435	-0,06616	0,026963	0
SKCM	Dinaciclib	1180	BRAF_mut	0,095915	-0,20792	0,027598	0
HNSC	Cytarabine	1006	PIK3CA_mut	0,167297	-0,18986	0,029903	0
LGG	Flavopiridol	432	cnaLGG14	0,141406	-0,26477	0,030723	0
LGG	Flavopiridol	432	NF1_mut	0,128396	-0,27499	0,031409	0
STAD	AZ960	1250	cnaSTAD3	0,078495	-0,17314	0,032522	0
LAML	AZD5582	1427	TP53_mut	0,079994	-0,24543	0,033135	0
LGG	(5Z)-7-Oxozeaenol	1242	cnaLGG14	0,111219	-0,24582	0,035027	0
LIHC	Tenovin-6	342	cnaLIHC10	0,113133	-0,27813	0,035654	0
LAML	JNK-9L	157	SACS_mut	0,116364	-0,37213	0,035874	0
SKCM	PLX-4720	1036	cnaSKCM28	0,231295	-0,04204	0,035944	0
COREAD	CAY10603	276	cnaCOREAD28	0,167429	-0,27299	0,037801	0
GBM	Tanespimycin	1026	PTEN_mut	0,144077	-0,21706	0,038547	0
SKCM	Topotecan	1808	cnaSKCM28	0,176617	-0,09896	0,03911	0

DLBC	Nutlin-3a (-)	1047	TP53_mut	0,169421	-0,17194	0,039581	0
LUSC	FH535	173	cnaLUSC54	0,118187	-0,25873	0,042794	0
COREAD	IMD-0354	442	TP53_mut	0,129258	-0,311	0,046909	0
SKCM	Tenovin-6	342	cnaSKCM27	0,088968	-0,29951	0,047266	0
UCEC	MIM1	446	CHD4_mut	0,171298	-0,22157	0,047708	0
PAAD	AZD8055	1059	cnaPAAD3	0,191562	-0,08989	0,048844	0
UCEC	QL-XII-47	235	CHD4_mut	0,161154	-0,2372	0,049204	0

Table 13: Combinations of alteration, tissue type and drug (ATD) with increasing viability markers identified by ANOVA model with linear slope, arranged according to significance values.

B.2 Increasing viability markers using ANOVA with Gaussian slope

Tissue	Drug	Drug ID	alteration	Diff. slope MT vs WT	Mean slope MT	p-adjusted value	Nr. high slope
HNSC	IPA-3	176	cnaHNSC19	1,396996671	0,49229497	0,02663	2
HNSC	Midostaurin	153	A_mut	1,476490887	0,435857881	0,001308	1
BRCA	BX-912	222	cnaBRCA27	1,205050423	0,127336753	0,01037	1
SKCM	RO-3306	1052	CDKN2A_mut	1,300699963	0,324957959	0,030105	1
HNSC	CUDC-101	273	NOTCH1_mut	0,604645207	-0,420596594	1,23E-06	0
LIHC	HG6-64-1	159	cnaLIHC7	0,382315989	-0,671169664	9,72E-05	0
LUAD	AR-42	272	cnaLUAD6	0,40610182	-0,625737697	0,000309	0
OV	AZ20	1184	cnaOV54	0,175065443	-0,84667261	0,000472	0
LUAD	AR-42	272	cnaLUAD7	0,370949106	-0,657285004	0,000977	0
SKCM	AZD6738	1394	cnaSKCM28	0,223983065	-0,692619097	0,001364	0
LGG	Doxorubicin	133	TP53_mut	0,265247246	-0,959799333	0,001658	0
PAAD	PF-00299804	363	cnaPAAD4	0,147047849	-1,034577129	0,003127	0
LUAD	AZD1480	1432	TP53_mut	0,174902356	-0,904102411	0,003319	0
SKCM	Apitolisib	382	cnaSKCM22	0,511948348	-0,548573807	0,004761	0
SKCM	Apitolisib	382	cnaSKCM24, cnaSKCM25	0,501992679	-0,557566024	0,004761	0
SKCM	PFI-1	1219	cnaSKCM28	0,101200654	-0,824516395	0,004826	0
LIHC	AZD8835	1445	cnaLIHC7	0,264470671	-0,698042055	0,006395	0
LUAD	Apitolisib	382	cnaLUAD10	0,209617633	-0,896557284	0,006766	0
LAML	AZD6738	1394	CREBBP_mut	0,111096017	-0,822664497	0,007175	0
LUAD	AKT inhibitor VIII	228	cnaLUAD22	1,002781795	-0,219513521	0,008039	0
LUAD	Tipifarnib	204	cnaLUAD13	0,45081383	-0,739272876	0,008403	0
LUAD	Tipifarnib	204	cnaLUAD29	0,35173439	-0,851456524	0,008403	0
HNSC	AZD5582	1617	cnaHNSC1	0,121631274	-0,771719015	0,009002	0
DLBC	VX-11e	2096	cnaDLBC2	0,176404735	-0,683599717	0,009113	0
HNSC	JNK-9L	157	cnaHNSC20	0,16070268	-0,764874771	0,009283	0
BLCA	CI-1040	1015	cnaBLCA5	0,593502139	-0,536027389	0,010077	0
UCEC	OSU-03012	167	LARP4B_mut	0,258652757	-0,878250668	0,010188	0
BRCA	Tipifarnib	204	cnaBRCA38	0,349468761	-0,880888622	0,010278	0
KIRC	LDN-193189	478	cnaKIRC12,cnaKIRC13,cnaKIRC14,cnaKIRC15	0,163433496	-0,953295778	0,011806	0
KIRC	BX-912	222	cnaKIRC24	0,368541083	-0,746269234	0,011819	0
LUSC	Podophyllotoxin bromide	1825	cnaLUSC41	0,109256049	-0,614671409	0,011838	0
PAAD	(5Z)-7-Oxozeaenol	1242	CDKN2A_mut	0,299117844	-0,675920455	0,01207	0
ESCA	Topotecan	1808	PIK3CA_mut	0,202638172	-0,677791441	0,012116	0
DLBC	Teniposide	1809	MLL2_mut	0,141328787	-0,671947086	0,012396	0
COREAD	Foretinib	308	KDM6A_mut	0,461295109	-0,688531154	0,01244	0
COREAD	Foretinib	308	SRGAP3_mut	0,463313418	-0,686685843	0,01244	0
ESCA	Shikonin	170	cnaESCA12	0,284770587	-0,562063994	0,013309	0
SKCM	Flavopiridol	432	ARID2_mut	0,357423346	-0,451671375	0,013795	0
OV	MPS-1-IN-1	294	cnaOV35	0,263915375	-0,980035407	0,013811	0
OV	MPS-1-IN-1	294	cnaOV84	0,263915375	-0,980035407	0,013811	0
OV	MPS-1-IN-1	294	cnaOV85	0,263915375	-0,980035407	0,013811	0

SKCM	Telomerase Inhibitor IX	1930	cnaSKCM3	0,121713854	-0,5905388	0,014246	0
DLBC	Oxaliplatin	1806	MLL2_mut	0,278611432	-0,520562397	0,014442	0
GBM	NG-25	260	NF1_mut	0,270868694	-1,028161321	0,01452	0
LUAD	GSK650394	177	SMARCA4_mut	0,609008413	-0,179319402	0,014583	0
COREAD	Refametinib	1526	FBXW7_mut	0,195748733	-0,770332347	0,015373	0
KIRC	PF-00299804	363	cnaKIRC12,cnaKIRC13,cnaKIRC14,cnaKIRC15	0,186317907	-0,909719853	0,015901	0
SKCM	MIM1	446	cnaSKCM4	0,125895494	-0,976857109	0,015971	0
GBM	FGFR_3831	1422	cnaGBM122	0,124767484	-0,774536636	0,015987	0
GBM	Obatoclax Mesylate	182	cnaGBM101	0,249898853	-0,844021032	0,01614	0
LGG	Cediranib	1922	cnaLGG16	0,0768098	-0,829696248	0,016255	0
LGG	AT-7519	219	cnaLGG14	0,282309865	-0,650272506	0,016804	0
BRCA	AZD8055	1059	cnaBRCA14	0,315244583	-0,300210543	0,016932	0
BRCA	AZD8055	1059	cnaBRCA16	0,332369212	-0,261096689	0,016932	0
LUAD	Piperlongumine	1243	cnaLUAD2	0,13446498	-0,841143231	0,017088	0
THCA	Embelin	172	NRAS_mut	0,23556058	-0,870309838	0,017235	0
SKCM	PFI-1	1219	cnaSKCM27	0,100483276	-0,820927407	0,018066	0
LUAD	AZD1332	1463	cnaLUAD34	0,200907963	-0,701267924	0,018283	0
THCA	Foretinib	2040	TP53_mut	0,15957413	-0,818213274	0,018767	0
LUSC	Telomerase Inhibitor IX	1930	cnaLUSC3	0,149850672	-0,630188252	0,020508	0
OV	FEN1_3940	1419	cnaOV51	0,088962838	-0,853201004	0,020993	0
COREAD	UNC0638	1236	PIK3CA_mut	1,240724488	0,334103348	0,021341	0
HNSC	CUDC-101	273	cnaHNSC14	0,322952567	-0,696253694	0,021708	0
PAAD	NG-25	260	CDKN2A_mut	0,202796489	-0,979913559	0,024771	0
BRCA	Eg5_9814	1712	cnaBRCA14	0,218253042	-0,480733159	0,025615	0
BRCA	Eg5_9814	1712	cnaBRCA15	0,218253042	-0,480733159	0,025615	0
BRCA	Eg5_9814	1712	cnaBRCA16	0,218253042	-0,480733159	0,025615	0
ESCA	CAY10603	276	cnaESCA7	0,096143841	-0,934103875	0,026428	0
LUAD	Flavopiridol	432	cnaLUAD2	0,163474889	-0,698041841	0,026504	0
LUAD	Flavopiridol	432	cnaLUAD7	0,203571448	-0,653248462	0,026504	0
ESCA	CAY10603	276	cnaESCA12	0,090040325	-0,938172885	0,026887	0
HNSC	PFI-1	1219	cnaHNSC29	0,186824891	-0,720733016	0,027163	0
SKCM	PFI-1	1219	cnaSKCM5	0,087660531	-0,832147309	0,028509	0
ESCA	Pazopanib	199	cnaESCA11	0,522260809	-0,703394836	0,028664	0
COREAD	Refametinib	1526	AKAP9_mut	0,196955303	-0,757193124	0,030257	0
ESCA	Lapatinib	1558	cnaESCA9	0,147196759	-0,766500806	0,030755	0
HNSC	CAP-232, TT-232, TLN-232	436	cnaHNSC19	0,398847533	-0,717586103	0,031062	0
HNSC	CAP-232, TT-232, TLN-232	436	cnaHNSC28	0,409256243	-0,708218264	0,031062	0
SKCM	AZD8931	1416	ARID2_mut	0,205518917	-0,717953178	0,031999	0
SKCM	AZD8931	1416	NF1_mut	0,195470383	-0,727021367	0,031999	0
HNSC	Mitomycin-C	136	NFE2L2_mut	0,253003745	-0,906991788	0,033076	0
HNSC	Refametinib	1014	CDKN2A_mut	0,339502157	-0,794037425	0,033198	0
ESCA	ACY-1215	264	cnaESCA7	0,114349154	-0,961296915	0,033337	0

LGG	BX795	1037	cnaLGG1	0,337919761	-0,854511439	0,034642	0
COREAD	Foretinib	308	CTNNB1_mut	0,369656242	-0,782876867	0,035777	0
LAML	ZSTK474	223	ASXL1_mut	0,153237712	-1,04266067	0,036071	0
LAML	ZSTK474	223	KRAS_mut	0,147884877	-1,046407655	0,036071	0
LAML	ZSTK474	223	TP53_mut	0,14330184	-1,063945965	0,036071	0
LGG	AZD8931	1416	cnaLGG16	0,211974389	-0,735521617	0,036877	0
SKCM	AZD6738	1394	cnaSKCM27	0,152898	-0,761757156	0,038159	0
KIRC	Bleomycin	190	TP53_mut	0,447867632	-0,653652355	0,03871	0
KIRC	Bleomycin	190	cnaKIRC22	0,447867632	-0,653652355	0,03871	0
BRCA	Midostaurin	153	PIK3CA_mut	1,095677871	0,070980763	0,039157	0
KIRC	IPA-3	176	SETD2_mut	0,905130635	-0,012315402	0,039966	0
SKCM	Dinaciclib	1180	cnaSKCM24, cnaSKCM25	0,255451148	-0,342301965	0,04043	0
ESCA	CDK9_5576	1708	cnaESCA7	0,088791623	-0,629728226	0,040468	0
GBM	ZM447439	1050	cnaGBM93,cnaGBM94,cnaGBM95	0,228940065	-0,877491018	0,040491	0
OV	FH535	173	cnaOV83	0,487051243	-0,50074131	0,040517	0
SKCM	PLX-4720	1036	cnaSKCM28	0,758757984	-0,066594611	0,040554	0
LUAD	Apitolisib	382	cnaLUAD19	0,201504567	-0,89567201	0,041814	0
BRCA	GNE-317	1926	cnaBRCA17	0,089919266	-0,639233821	0,042045	0
DLBC	TWS119	366	cnaDLBC2	0,150043697	-0,870748975	0,042947	0
BRCA	FTY-720	546	cnaBRCA32	0,149597932	-0,912330346	0,043269	0
GBM	OSI-027	299	PTEN_mut	0,34075665	-0,882344596	0,043368	0
COREAD	EHT-1864	1069	cnaCOREAD33	0,212006277	-1,000131314	0,04396	0
OV	Embelin	172	cnaOV38	0,379977504	-0,611077381	0,044274	0
LUAD	Flavopiridol	432	cnaLUAD31	0,189185738	-0,666653328	0,045171	0
GBM	UNC0638	1236	NF1_mut	0,12015983	-0,807388426	0,045597	0
GBM	UNC0638	1236	cnaGBM68	0,142291557	-0,771252685	0,045597	0
DLBC	NVP-BHG712	295	EP300_mut	0,236715223	-0,956054603	0,045898	0
ESCA	QL-XII-47	235	NOTCH1_mut	0,283639109	-0,759574968	0,046699	0
SKCM	NSC-207895	269	cnaSKCM22	0,107481374	-1,084630298	0,04672	0
SKCM	NSC-207895	269	cnaSKCM23	0,107481374	-1,084630298	0,04672	0
SKCM	NSC-207895	269	cnaSKCM24, cnaSKCM25	0,107481374	-1,084630298	0,04672	0
GBM	Obatoclax Mesylate	182	cnaGBM97,cnaGBM98,cnaGBM99,cnaGBM100,cnaGBM102,cnaGBM103,cnaGBM104,cnaGBM105,cnaGBM106	0,269796448	-0,808947406	0,047243	0
COREAD	Foretinib	308	ELF3_mut	0,402141001	-0,74261491	0,048388	0
DLBC	Vorinostat	1012	cnaDLBC2	0,365700794	-0,382365731	0,048899	0
BRCA	Lapatinib	1558	cnaBRCA26	0,26648522	-0,700709786	0,048915	0
BRCA	Lapatinib	1558	cnaBRCA27	0,299200766	-0,639090698	0,048915	0
BRCA	Lapatinib	1558	cnaBRCA30	0,330734394	-0,599690422	0,048915	0
BRCA	CD532	449	cnaBRCA33	0,160724818	-0,882146663	0,049058	0
DLBC	BIBF-1120	380	TP53_mut	0,188345783	-0,871747497	0,049381	0
GBM	Fedratinib	306	cnaGBM123	0,206704446	-1,039079143	0,049419	0

HNSC	IGF1R_3801	1738	SMAD4_mut	0,236621201	-0,583141086	0,049454	0
LUAD	Navitoclax	1011	cnaLUAD10	1,118362606	0,105253221	0,049831	0

Table 14: Combinations of alteration, tissue type and drug (ATD) with increasing viability markers identified by ANOVA model with Gaussian slope, arranged according to significance values and number of responses with high-slope (i.e., $\text{slope}_{\text{Gaussian}}$>1.1) in the mutant population.

B.3 Increasing viability markers using Hypergeometric enrichment with linear slope

Tissue	Drug	Drug ID	alteration	Enrichment Score	Diff. slope MT vs WT	Mean slope MT	p-adjusted value	Nr. high slope
OV	CHIR-99021	154	cnaOV84	0,962218	0,216999	0,231771	5,31E-05	4
OV	CHIR-99021	154	cnaOV85	0,949301	0,181062	0,231771	3,35E-05	4
LUAD	CHIR-99021	154	cnaLUAD3	0,892942	0,098662	0,205837	8,44E-07	4
HNSC	CHIR-99021	154	cnaHNSC17	0,828134	0,049332	0,149774	2,78E-07	4
KIRC	SB590885	1061	TP53_mut	0,825863	0,088995	0,184869	0,000137	4
OV	CHIR-99021	154	TP53_mut	0,805262	0,063644	0,170504	3,35E-05	4
HNSC	CHIR-99021	154	cnaHNSC32	0,773111	0,035237	0,165113	2,26E-05	4
OV	CHIR-99021	154	cnaOV38	0,958384	0,236368	0,258117	0,000694	3
OV	CHIR-99021	154	cnaOV83	0,952361	0,197978	0,227404	0,001095	3
OV	CHIR-99021	154	cnaOV39	0,943975	0,187492	0,258117	0,000217	3
COREAD	PAC-1	175	SMAD4_mut	0,90299	0,165299	0,199932	0,000903	3
LUAD	CHIR-99021	154	cnaLUAD27	0,799777	0,040195	0,196592	2,94E-06	3
LUAD	VX-702	1028	cnaLUAD3	0,724679	0,05075	0,142923	0,001874	3
COREAD	VX-702	1028	FBXW7_mut	0,717891	0,060616	0,132916	0,00073	3
LUAD	SB590885	1061	STK11_mut	0,69442	0,055461	0,218733	0,003374	3
LUAD	SB590885	1061	SMARCA4_mut	0,662497	0,042372	0,141069	0,0014	3
LUAD	VX-702	1028	cnaLUAD27	0,651763	0,027561	0,148128	0,001482	3
COREAD	PAC-1	175	FBXW7_mut	0,609987	0,079158	0,212148	0,020163	3
GBM	Bosutinib	1019	cnaGBM122	0,592087	0,116196	0,195961	0,028523	3
ESCA	SB590885	1061	cnaESCA11	0,589711	0,014217	0,24459	0,002064	3
LUAD	SB590885	1061	KRAS_mut	0,573302	0,005319	0,195751	0,000328	3
PAAD	IOX2	1230	ARID1A_mut	0,971096	0,214001	0,210681	0,000714	2
STAD	Doramapimod	1042	cnaSTAD51	0,957162	0,20162	0,177771	0,005267	2
COREAD	VX-702	1028	PBRM1_mut	0,956213	0,112968	0,164347	0,002442	2
OV	AT7867	356	cnaOV26	0,953078	0,230128	0,165439	0,035813	2
PAAD	IOX2	1230	CDKN2A_mut	0,950669	0,186993	0,246045	0,000584	2
LUSC	Lenalidomide	1020	cnaLUSC6,cnaLUSC7,cnaLUSC37	0,950292	0,172935	0,165074	0,007215	2
COREAD	VX-702	1028	FAM123B_mut	0,949635	0,129009	0,166072	0,00073	2
LUAD	TL-2-105	211	ARID1A_mut	0,946464	0,198925	0,2296	0,004255	2
OV	Doramapimod	1042	NF1_mut	0,94582	0,132471	0,229292	0,002463	2
LUSC	Lenalidomide	1020	cnaLUSC25	0,928981	0,128514	0,165074	0,006146	2
BRCA	NSC-87877	147	cnaBRCA29	0,926677	0,08796	0,176215	0,001333	2
KIRC	Lenalidomide	1020	cnaKIRC23	0,924315	0,169019	0,153241	0,00227	2
LUAD	VNLG/124	271	cnaLUAD14	0,921427	0,133676	0,310294	0,000167	2

OV	CHIR-99021	154	cnaOV35	0,920204	0,120343	0,171314	0,011404	2
COREAD	Vismodegib	1033	BRWD1_mut	0,912311	0,15194	0,249241	0,032124	2
LUAD	VX-702	1028	cnaLUAD12	0,910996	0,072853	0,154252	0,002433	2
LUAD	CHIR-99021	154	cnaLUAD10	0,909401	0,035948	0,154126	1,05E-05	2
GBM	UNC0642	1263	cnaGBM25	0,908084	0,15591	0,2357	0,006653	2
OV	CHIR-99021	154	cnaOV40	0,9047	0,121525	0,230221	0,011404	2
BLCA	QL-XII-61	1203	TP53_mut	0,901178	0,155601	0,194877	0,002638	2
LUSC	Lenalidomide	1020	cnaLUSC14	0,887209	0,149456	0,126363	0,006146	2
HNSC	CHIR-99021	154	cnaHNSC19	0,88097	0,043325	0,14171	0,003099	2
BRCA	IPA-3	176	cnaBRCA57	0,878559	0,188192	0,150477	0,009614	2
LUAD	JQ1	1218	cnaLUAD13	0,872514	0,102868	0,171232	0,003735	2
OV	CHIR-99021	154	cnaOV20	0,872381	0,065029	0,169958	0,016468	2
COREAD	AT7867	356	cnaCOREAD48	0,868113	0,234303	0,184547	0,014054	2
COREAD	VX-702	1028	MGA_mut	0,86597	0,076868	0,164474	0,00619	2
LUAD	VX-702	1028	cnaLUAD10	0,845373	0,059145	0,102656	0,001482	2
OV	Doramapimod	1042	cnaOV95	0,844996	0,068765	0,149072	0,001479	2
BRCA	Midostaurin	153	cnaBRCA39	0,842326	0,16051	0,133954	0,006695	2
COREAD	VX-702	1028	CTCF_mut	0,836008	0,107378	0,200389	0,004338	2
OV	Doramapimod	1042	cnaOV94	0,835077	0,058781	0,201843	0,001856	2
LUAD	VX-702	1028	cnaLUAD18	0,831054	0,055876	0,146819	0,003013	2
LUAD	CHIR-99021	154	cnaLUAD17	0,827104	0,077479	0,201773	0,003111	2
LUAD	C-75	435	cnaLUAD29	0,82158	0,024338	0,12895	0,000212	2
LIHC	CHIR-99021	154	cnaLIHC7	0,803436	0,036856	0,209196	0,048623	2
COREAD	VX-702	1028	CEP290_mut	0,799232	0,109652	0,225876	0,018768	2
COREAD	UNC0642	1263	NCOR1_mut	0,795855	0,130937	0,339461	0,017755	2
OV	CHIR-99021	154	cnaOV54	0,794608	0,044389	0,250515	0,011404	2
COREAD	UNC0642	1263	CTNNB1_mut	0,792157	0,108226	0,327945	0,017755	2
OV	CHIR-99021	154	cnaOV55	0,787009	0,056237	0,311382	0,047285	2
LUAD	CHIR-99021	154	STK11_mut	0,784156	0,00671	0,167347	0,00141	2
STAD	Doramapimod	1042	cnaSTAD47	0,777668	0,101424	0,141167	0,005267	2
OV	CHIR-99021	154	cnaOV52,cnaOV53	0,771591	0,03378	0,212834	0,015583	2
LUSC	PFI-3	1530	cnaLUSC10	0,767329	0,064508	0,145225	0,002819	2
COREAD	VX-702	1028	CHD9_mut	0,76587	0,077621	0,163638	0,00619	2
BRCA	AZD6094	1403	cnaBRCA47	0,762722	0,056173	0,149516	0,009675	2
COREAD	Vismodegib	1033	B2M_mut	0,757514	0,074769	0,177353	0,009368	2
LUAD	TL-2-105	211	cnaLUAD22	0,742463	0,100536	0,158473	0,011927	2
GBM	SGC0946	1264	cnaGBM125	0,741251	0,04027	0,119009	0,006708	2
HNSC	TL-2-105	211	cnaHNSC2	0,732983	0,050524	0,140552	0,001974	2
STAD	Doramapimod	1042	cnaSTAD16	0,728696	0,091649	0,220846	0,01568	2

COREAD	UNC0642	1263	cnaCOREAD49	0,72358	0,071139	0,168762	0,000691	2
BRCA	NSC-87877	147	cnaBRCA17	0,713667	0,044023	0,247119	0,003595	2
COREAD	VX-702	1028	SACS_mut	0,713229	0,076204	0,155106	0,004338	2
KIRC	Lenalidomide	1020	cnaKIRC24	0,710447	0,074879	0,131361	0,006415	2
LUAD	Axitinib	1021	SMARCA4_mut	0,709011	0,108949	0,156274	0,00214	2
STAD	Doramapimod	1042	cnaSTAD17	0,702646	0,095498	0,141167	0,011349	2
LUSC	Lenalidomide	1020	cnaLUSC10	0,69871	0,077544	0,112328	0,006146	2
GBM	PAC-1	175	PTEN_mut	0,689436	0,036565	0,123671	0,000854	2
KIRC	Lenalidomide	1020	cnaKIRC22	0,684645	0,073258	0,132288	0,001587	2
LUAD	VX-702	1028	STK11_mut	0,677764	0,038582	0,147294	0,002944	2
BRCA	AZD6094	1403	cnaBRCA17	0,675683	0,02226	0,133146	0,008511	2
KIRC	Lenalidomide	1020	cnaKIRC12,cnaKIRC13,cnaKIRC14,cnaKIRC15	0,666445	0,074744	0,128577	0,001587	2
COREAD	VX-702	1028	ARID1A_mut	0,659634	0,053363	0,173905	0,007339	2
THCA	VX-702	1028	TP53_mut	0,659328	0,037366	0,162773	0,003894	2
COREAD	UNC0642	1263	MLL2_mut	0,658156	0,049142	0,208432	0,004854	2
COREAD	VX-702	1028	B2M_mut	0,656197	0,048551	0,150983	0,022292	2
COREAD	JNK Inhibitor VIII	1043	PIK3CA_mut	0,653037	0,035746	0,178491	0,010043	2
DLBC	PFI3	1620	TP53_mut	0,651595	0,008114	0,114195	0,011626	2
STAD	Doramapimod	1042	cnaSTAD30	0,634323	0,059601	0,220846	0,006844	2
BRCA	AZD6094	1403	cnaBRCA18	0,633176	0,010931	0,118379	0,000612	2
COREAD	IAP_7638	1429	PIK3CA_mut	0,626907	0,043921	0,10846	0,006302	2
PAAD	SB590885	1061	SMAD4_mut	0,624828	0,034219	0,201394	0,011599	2
BRCA	NSC-87877	147	cnaBRCA22	0,623658	0,029318	0,115004	0,009591	2
COREAD	UNC0642	1263	RNF43_mut	0,607426	0,035179	0,267675	0,017755	2
COREAD	Vismodegib	1033	cnaCOREAD49	0,596759	0,041755	0,132623	0,003604	2
COREAD	VX-702	1028	PIK3CA_mut	0,589508	0,028824	0,160055	0,004338	2
THCA	IAP_7638	1429	BRAF_mut	0,580572	0,016298	0,146509	0,040054	2
COREAD	Vismodegib	1033	PIK3CA_mut	0,579758	0,038309	0,151816	0,009368	2
STAD	Doramapimod	1042	cnaSTAD41	0,575	0,086287	0,151409	0,005267	2
COREAD	VX-702	1028	cnaCOREAD49	0,574282	0,006342	0,138091	0,004037	2
BRCA	NSC-87877	147	cnaBRCA18	0,572611	0,008938	0,222921	0,045055	2
STAD	Doramapimod	1042	cnaSTAD13	0,566766	0,056558	0,13033	0,006978	2

COREAD	Vismodegib	1033	cnaCOREAD33	0,564205	0,027582	0,125329	0,001547	2
BRCA	Midostaurin	153	cnaBRCA26	0,551175	0,107518	0,100451	0,006695	2
PAAD	SB216763	1025	cnaPAAD4	0,538375	0,102613	0,137374	0,006266	2
COREAD	UNC0642	1263	EP300_mut	0,529069	0,009071	0,200803	0,017755	2
DLBC	GSK2606414	1618	TP53_mut	0,527546	0,000488	0,152111	0,028721	2
GBM	UNC0642	1263	cnaGBM125	0,524621	0,007535	0,149323	0,043333	2
COREAD	VX-702	1028	EP300_mut	0,522024	0,003023	0,174836	0,012205	2
COREAD	IAP_7638	1429	cnaCOREAD19	0,521494	0,013319	0,163593	0,015048	2
COREAD	MCT1_6447	1436	PIK3CA_mut	0,51746	0,007313	0,103562	0,013164	2
HNSC	GSK1904529A	202	cnaHNSC32	0,508083	0,035834	0,136844	0,020294	2
HNSC	GSK1904529A	202	cnaHNSC32	0,508083	0,035834	0,136844	0,020294	2
STAD	Doramapimod	1042	TP53_mut	0,501968	0,052837	0,151409	0,011349	2
COREAD	Vismodegib	1033	cnaCOREAD11	0,499988	0,0054	0,175732	0,011946	2
HNSC	PD173074	1049	cnaHNSC11	0,496794	0,038638	0,12689	0,01959	2
COREAD	Vismodegib	1033	KRAS_mut	0,490065	0,007688	0,117101	0,003604	2
COREAD	IAP_7638	1429	EP300_mut	0,484063	0,013398	0,107438	0,02472	2
BRCA	Dasatinib	1079	cnaBRCA31	0,44226	0,044286	0,12542	0,031964	2
BRCA	Dasatinib	1079	cnaBRCA31	0,44226	0,044286	0,12542	0,031964	2

Table 15: Combinations of alteration, tissue type and drug (ATD) with increasing viability markers identified by the hypergeometric model with linear slope, arranged according to significance values and number of responses with high-slope (i.e., $slope_{linear}$>0.2) in the mutant population. Only ATD combinations with at least two mutant responses with high-slope are depicted.

B.4 Increasing viability markers using Hypergeometric enrichment with Gaussian slope

Tissue	Drug	Drug ID	alteration	Enrichment Score	Diff. slope MT vs WT	Mean slope MT	p-adjusted value	Nr. high slope
STAD	XMD11-85h	1158	TP53_mut	0,694942	1,138687	0,12809	0,000387	5
COREAD	Erlotinib	1	EP300_mut	0,90587	1,49234	0,17312	0,000238	4
STAD	XMD11-85h	1158	cnaSTAD16	0,856011	1,224707	0,136431	0,000583	4
COREAD	Erlotinib	1	ARID1A_mut	0,831976	1,589181	0,140655	0,001358	4
LUAD	UNC0638	1236	cnaLUAD21	0,673723	1,139841	0,066189	0,008708	4
UCEC	FGFR_0939	1421	ARID1A_mut	0,98578	1,892104	0,07564	4,56E-05	3
COREAD	Erlotinib	1	SACS_mut	0,984591	1,664232	0,137077	0,0002	3
COREAD	Erlotinib	1	MAP2K4_mut	0,981977	1,654626	0,186106	0,0002	3
COREAD	Erlotinib	1	AKAP9_mut	0,981977	1,610047	0,174382	0,0002	3
KIRC	PFI-3	1530	PTEN_mut	0,979802	1,503329	0,113725	5,4E-05	3
BLCA	PLX-4720	1371	CDKN1A_mut	0,976013	1,701688	0,10108	0,000471	3
BLCA	KIN001-042	289	cnaBLCA18	0,975509	1,145038	0,058193	0,0004	3
LUSC	UNC0638	1236	cnaLUSC30,cnaLUSC32	0,974237	1,623793	0,079332	0,000564	3
COREAD	Linifanib	277	SOS2_mut	0,97348	1,138847	0,028733	4,22E-05	3
BLCA	MCT1_6447	1436	cnaBLCA9	0,969159	1,490512	0,074316	2,01E-05	3
LUAD	UNC0642	1263	cnaLUAD12	0,96812	1,208607	0,059727	0,000101	3
COREAD	JQ1	1218	NUP98_mut	0,961332	1,392523	0,081002	1,27E-05	3
OV	CHIR-99021	154	cnaOV84	0,961311	1,258175	0,231771	8,09E-06	3
OV	CHIR-99021	154	cnaOV38	0,961309	1,176964	0,258117	5,17E-05	3
LIHC	PHA-793887	301	cnaLIHC7	0,956858	1,342044	0,137468	0,000145	3
COREAD	AZD5582	1427	cnaCOREAD27	0,942312	1,143837	0,06497	0,000792	3
BLCA	IPA-3	176	cnaBLCA5	0,941676	1,42047	0,078233	1,14E-05	3
KIRC	Lenalidomide	1020	cnaKIRC23	0,931508	1,251987	0,153241	0,000571	3
OV	CHIR-99021	154	cnaOV85	0,928102	1,117152	0,195766	1,75E-05	3
LGG	Sunitinib	5	cnaLGG1	0,92336	1,192607	0,098974	0,000148	3
LUSC	UNC0638	1236	cnaLUSC25	0,907913	1,452893	0,062364	0,001959	3
OV	Veliparib	1018	cnaOV43	0,899182	1,183985	0,111806	0,001046	3
LAML	Dasatinib	51	NRAS_mut	0,898549	1,331125	0,053597	3,66E-05	3
COREAD	Erlotinib	1	cnaCOREAD19	0,886713	1,262343	0,137077	0,004748	3
COREAD	Erlotinib	1	CHD9_mut	0,883541	1,254483	0,186106	0,004748	3
COREAD	Erlotinib	1	PTEN_mut	0,883541	1,254483	0,186106	0,004748	3
COREAD	Erlotinib	1	BMPR2_mut	0,883463	1,253387	0,194914	0,004748	3
COREAD	Erlotinib	1	MLL2_mut	0,883463	1,253387	0,194914	0,004748	3
BRCA	UNC0642	1263	cnaBRCA12	0,828414	1,212999	0,046617	1,89E-05	3
BLCA	MPS-1-IN-1	294	cnaBLCA15	0,780797	1,393804	0,048508	0,000867	3
BRCA	UNC0642	1263	cnaBRCA64	0,753693	1,16442	0,05549	0,000292	3

DLBC	rTRAIL	1261	PTEN_mut	0,742834	1,101629	0,104614	0,009635	3
UCEC	FGFR_0939	1421	EP300_mut	0,737932	1,450967	0,058485	0,016784	3
LUAD	Bryostatin 1	197	RB1_mut	0,976571	1,163421	0,096963	0,000101	2
LIHC	PHA-793887	301	cnaLIHC10	0,971914	1,292134	0,052123	0,000263	2
LIHC	Bicalutamide	1502	cnaLIHC7	0,970527	1,264931	0,073834	0,000151	2
STAD	AZD4547	1497	BCOR_mut	0,969679	1,592553	0,050627	0,000274	2
LUAD	Bicalutamide	1502	U2AF1_mut	0,968247	1,669657	0,131842	0,000185	2
COREAD	JQ1	1218	TP53BP1_mut	0,968161	1,474772	0,090335	0,00116	2
STAD	Ruxolitinib	206	ARID2_mut	0,967024	1,113653	0,036436	0,000849	2
SKCM	TANK_1366	1461	cnaSKCM26	0,964198	1,269747	0,041974	0,001526	2
COREAD	FGFR_0939	1421	cnaCOREAD14	0,961949	1,154057	0,141782	0,000998	2
PRAD	MCT4_1422	1437	cnaPRAD7	0,96155	1,810884	0,117706	0,000668	2
OV	CHIR-99021	154	cnaOV83	0,961309	1,189109	0,227404	5,17E-05	2
LAML	Dasatinib	51	KRAS_mut	0,960798	1,204202	0,090169	0,000381	2
STAD	WHI-P97	288	cnaSTAD45	0,957023	1,475322	0,056081	0,000147	2
OV	MCT4_1422	1437	cnaOV50	0,952241	1,780874	0,069387	0,000163	2
LUAD	SB505124	1194	RBM10_mut	0,950534	1,234897	0,108705	0,001459	2
COREAD	JAK1_3715	1433	cnaCOREAD21	0,950329	1,568255	0,023352	0,002143	2
OV	AT7867	356	cnaOV102	0,948575	1,206786	0,0417	0,006821	2
COREAD	Tubastatin A	265	BRWD1_mut	0,948302	1,391116	0,037868	0,000555	2
HNSC	Ponatinib	155	A_mut	0,947013	1,291424	0,103037	0,000793	2
GBM	BMS-345541	203	cnaGBM134,cnaGBM135	0,943782	1,163311	0,070634	0,00254	2
HNSC	SGC0946	1264	cnaHNSC3	0,943195	1,114322	0,087709	0,002177	2
HNSC	SGC0946	1264	A_mut	0,941247	1,115209	0,083975	0,002177	2
SKCM	SGC0946	1264	cnaSKCM26	0,937842	1,116728	0,057478	0,001299	2
BRCA	UNC0642	1263	BRCA1_mut	0,937749	1,132761	0,048376	0,000292	2
LUAD	QL-XII-61	1203	cnaLUAD30	0,937644	1,618105	0,068006	0,000435	2
ESCA	UNC0638	1236	cnaESCA2,cnaESCA3	0,936381	1,725635	0,140194	0,001668	2
COREAD	PLX-4720	1371	cnaCOREAD23,cnaCOREAD53	0,93232	1,274254	0,09517	0,005122	2
OV	IAP_7638	1429	cnaOV61,cnaOV62,cnaOV63	0,931501	1,647349	0,055135	0,002457	2
ESCA	PFI-3	1530	NFE2L2_mut	0,930238	1,173527	0,056346	0,001884	2
COREAD	CHIR-99021	1241	cnaCOREAD28	0,925111	1,198795	0,151834	0,003329	2
GBM	JAK3_7406	1434	cnaGBM133	0,92234	1,825803	0,043519	0,003035	2

OV	AT7867	356	cnaOV26	0,91571	1,408289	0,165439	0,0158	2
LGG	CHIR-99021	1241	cnaLGG7	0,908751	1,348566	0,063622	0,004717	2
ESCA	AZD3514	1382	cnaESCA2,cnaESCA3	0,907794	1,762344	0,138821	0,004731	2
ESCA	Tubastatin A	265	cnaESCA2,cnaESCA3	0,907144	1,259352	0,02643	0,009578	2
LUAD	T0901317	333	cnaLUAD32	0,902914	1,180897	0,018668	0,008062	2
UCEC	Refametinib	1526	CHD4_mut	0,89891	1,933095	0,122935	0,007239	2
BRCA	QL-XII-61	1203	cnaBRCA30	0,8969	1,432752	0,058462	0,010841	2
COREAD	UNC0638	1236	KDM6A_mut	0,895613	1,115138	0,108913	0,007637	2
LUSC	UNC0638	1236	cnaLUSC6,cnaLUSC7,cnaLUSC37	0,895348	1,270226	0,048897	0,008125	2
BRCA	VX-702	1028	BRCA1_mut	0,895114	1,283577	0,131431	0,007354	2
COREAD	JQ1	1218	cnaCOREAD21	0,89255	1,146759	0,070519	0,01957	2
OV	Cisplatin	1496	SMAD4_mut	0,888427	1,776571	0,033225	0,005379	2
STAD	UNC0638	1236	BCOR_mut	0,885635	1,218922	0,094426	0,01059	2
LUAD	JAK3_7406	1434	U2AF1_mut	0,88238	1,21701	0,04833	0,022524	2
OV	JAK1_3715	1433	cnaOV87,cnaOV88	0,881508	1,272219	0,049873	0,031332	2
GBM	RO-3306	1052	cnaGBM124	0,87957	1,349715	0,071208	0,003353	2
COREAD	PLX-4720	1371	cnaCOREAD27	0,874312	1,179891	0,097373	0,005122	2
LIHC	Olaparib	1495	cnaLIHC10	0,87137	1,909235	0,037377	0,015854	2
OV	JAK1_3715	1433	cnaOV61,cnaOV62,cnaOV63	0,863284	1,234479	0,046686	0,031332	2
OV	JAK1_3715	1433	cnaOV18	0,859746	1,263849	0,04459	0,018692	2
OV	JAK1_3715	1433	cnaOV19	0,859743	1,225807	0,030517	0,031332	2
OV	FR-180204	263	cnaOV82	0,855189	1,138828	0,069213	0,04629	2
BRCA	JQ1	1218	cnaBRCA24	0,854893	1,346379	0,067912	0,040426	2
LAML	TANK_1366	1461	SACS_mut	0,852646	1,546972	0,01778	0,013053	2
COREAD	Piperlongumine	1243	CDH1_mut	0,846204	1,41016	0,015718	0,03848	2
BRCA	UNC0642	1263	cnaBRCA48	0,845387	1,12592	0,04517	0,001794	2
LIHC	AZD4547	1497	cnaLIHC7	0,843931	1,486166	0,054692	0,019199	2
COREAD	TANK_1366	1461	SOX9_mut	0,824442	1,286752	0,044442	0,00091	2
PAAD	FGFR_3831	1422	ARID1A_mut	0,82032	1,55538	0,027715	0,03685	2
COREAD	Piperlongumine	1243	BRWD1_mut	0,817048	1,376095	0,033485	0,042468	2
SKCM	IAP_7638	1429	cnaSKCM30	0,81208	1,125409	0,053952	0,017139	2
BLCA	IAP_7638	1429	MLL2_mut	0,797737	1,449961	0,021328	0,031936	2
COREAD	FGFR_3831	1422	SRGAP3_mut	0,7436	1,104326	0,138401	0,049119	2
OV	FEN1_3940	1419	PTEN_mut	0,715882	1,398655	0,053381	0,041698	2

LUAD	Genentech Cpd 10	225	cnaLUAD25	0,710741	1,521279	0,087733	0,036276	2
BLCA	NVP-BHG712	295	cnaBLCA9	0,680603	1,101755	0,049679	0,042106	2
HNSC	NPK76-II-72-1	257	cnaHNSC19	0,65233	1,138745	0,035398	0,022507	2

Table 16: Combinations of alteration, tissue type and drug (ATD) with increasing viability markers identified by the hypergeometric model with Gaussian slope, arranged according to significance values and number of responses with high-slope (i.e., $slope_{Gaussian}$>1.1) in the mutant population. Only ATD combinations with at least two mutant responses with high-slope are depicted.

B.5 Correlation between cell growth rate and drug response

Drug	Drug ID	Drug Target	Target Pathway	*R*
BI-2536	1086	PLK1, PLK2, PLK3	Cell cycle	-0,514
5-Fluorouracil	1073	Antimetabolite (DNA & RNA)	Other	-0,50601
Dactinomycin	1911	RNA polymerase	Other	-0,49023
Gemcitabine	1190	Pyrimidine antimetabolite	DNA replication	-0,48697
Oxaliplatin	1089	DNA alkylating agent	DNA replication	-0,47561
Epirubicin	1511	Anthracycline	DNA replication	-0,46112
Docetaxel	1819	Microtubule stabiliser	Mitosis	-0,45534
Luminespib	1559	HSP90	Protein stability and degradation	-0,45367
Paclitaxel	1080	Microtubule stabiliser	Mitosis	-0,45343
Vinorelbine	2048	Microtubule destabiliser	Mitosis	-0,43517
MK-1775	1179	WEE1, PLK1	Cell cycle	-0,43221
Buparlisib	1873	PI3Kalpha, PI3Kdelta, PI3Kbeta, PI3Kgamma	PI3K/MTOR signaling	-0,43021
Camptothecin	1003	TOP1	DNA replication	-0,43012
Telomerase Inhibitor IX	1930	Telomerase	Genome integrity	-0,42353
Oxaliplatin	1806	DNA alkylating agent	DNA replication	-0,41629
GNE-317	1926	PI3Kalpha	PI3K/MTOR signaling	-0,4158
AZD6738	1917	ATR	Genome integrity	-0,41505
BDP-00009066	1866	MRCKB_HUMAN	Cytoskeleton	-0,41462
PARP_0108	1459	PARP1, PARP2, PARP6	Genome integrity	-0,41019
Rapamycin	1084	MTORC1	PI3K/MTOR signaling	-0,4091
Obatoclax Mesylate	1068	BCL2, BCL-XL, BCL-W, MCL1	Apoptosis regulation	-0,40842
Gemcitabine	1393	Pyrimidine antimetabolite	DNA replication	-0,40682
BMS-345541	1249	IKK1, IKK2	Other, kinases	-0,40623
VE821	2111	ATR	Genome integrity	-0,40607
CDK9_5576	1708	CDK9	Cell cycle	-0,40422
Vinblastine	1004	Microtubule destabiliser	Mitosis	-0,40233
Ulixertinib	1908	ERK1, ERK2	ERK MAPK signaling	-0,39826
AZD5153	1706	BRD4	Chromatin other	-0,39257
PARP_9482	1460	PARP1, PARP2, PARP5a	Genome integrity	-0,39211
ICL1100013	1266	N-myristoyltransferase 1/2	Other	-0,39023
GSK591	2110	PMRT5	Chromatin histone methylation	-0,38953
AZD7969	1426	GSK3B	WNT signaling	-0,38949
VX-11e	2096	ERK2	ERK MAPK signaling	-0,38889
Dactolisib	1057	PI3K (class 1), MTORC1, MTORC2	PI3K/MTOR signaling	-0,38789
YK-4-279	1239	RNA helicase A	Other	-0,38662
Irinotecan	1088	TOP1	DNA replication	-0,38349
Docetaxel	1007	Microtubule stabiliser	Mitosis	-0,38282
Pevonedistat	1529	NAE	Other	-0,3791
OTX015	1626	BRD2, BRD3, BRD4	Chromatin other	-0,37639
AZ20	1184	ATR	Genome integrity	-0,37504
Pyridostatin	2044	G-quadruplex stabiliser	DNA replication	-0,3741

PRT062607	1631	SYK	Other, kinases	-0,37116
Cytarabine	1006	Antimetabolite	Other	-0,37059
Gemcitabine	135	Pyrimidine antimetabolite	DNA replication	-0,36526
Bleomycin (50 uM)	1378	dsDNA break induction	DNA replication	-0,36444
AZD6738	1394	ATR	Genome integrity	-0,36343
AZ6102	2109	TNKS1, TNKS2	WNT signaling	-0,36231
Palbociclib	1054	CDK4, CDK6	Cell cycle	-0,3588
TTK_3146	1464	TTK	Mitosis	-0,35843
Dinaciclib	1180	CDK1, CDK2, CDK5, CDK9	Cell cycle	-0,35759
Daporinad	1248	NAMPT	Metabolism	-0,35465
CDK9_5038	1709	CDK9	Cell cycle	-0,3512
Wee1 Inhibitor	1046	WEE1, CHEK1	Cell cycle	-0,3492
PLK_6522	1451	PLK1, PLK2, PLK3	Cell cycle	-0,34862
I-BET-762	1624	BRD2, BRD3, BRD4	Chromatin other	-0,34772
SN-38	1494	TOP1	DNA replication	-0,3475
CAY10566	416	Stearoyl-CoA desaturase	Other	-0,34227
CCT-018159	1170	HSP90	Protein stability and degradation	-0,34157
HG-5-113-01	1142	LOK, LTK, TRCB, ABL(T315I)	Other	-0,34056
ERK_2440	1713	ERK1,ERK2	ERK MAPK signaling	-0,34043
Doxorubicin	1386	Anthracycline	DNA replication	-0,34006
VE-822	1613	ATR	Genome integrity	-0,33998
KRAS (G12C) Inhibitor-12	1855	KRAS (G12C)	ERK MAPK signaling	-0,3367
VSP34_8731	1734	VSP34	Other	-0,33522
Talazoparib	1259	PARP1, PARP2	Genome integrity	-0,3346
Entospletinib	1630	SYK	Other, kinases	-0,33218
Bleomycin (10 uM)	1392	dsDNA break induction	DNA replication	-0,33165
Sabutoclax	1849	BCL2, BCL-XL, BFL1, MCL1	Apoptosis regulation	-0,33157
GSK650394	177	SGK2, SGK3	Other, kinases	-0,3306
Niraparib	1177	PARP1, PARP2	Genome integrity	-0,32933
I-BRD9	1928	BRD9	Chromatin other	-0,32723
AT13148	2170	AKT1	PI3K/MTOR signaling	-0,32542
OSU-03012	167	PDK1 (PDPK1)	Metabolism	-0,32429
Temozolomide	1375	DNA alkylating agent	DNA replication	-0,32354
MIM1	1996	MCL-1	Apoptosis regulation	-0,32315
WZ4003	1614	NUAK1, NUAK2	Other, kinases	-0,32239
Cisplatin	1005	DNA crosslinker	DNA replication	-0,32107
Alisertib	1051	AURKA	Mitosis	-0,32039
Lestaurtinib	1024	FLT3, JAK2, NTRK1, NTRK2, NTRK3	Other, kinases	-0,31775
Selisistat	341	SIRT1	Chromatin histone acetylation	-0,31458
Crizotinib	1083	MET, ALK, ROS1	RTK signaling	-0,31452
AZ960	1250	JAK2, JAK3	Other, kinases	-0,31399
GSK1904529A	1093	IGF1R, IR	RTK signaling &	-0,31343

			IGF1R signaling	
GSK2606414	1618	PERK	Metabolism	-0,31338
Pemetrexed	428	TYMS	DNA replication	-0,31299
ERK_6604	1714	ERK1,ERK2	ERK MAPK signaling	-0,31196
Epothilone B	201	Microtubule stabiliser	Mitosis	-0,3113
LJI308	2107	RSK2, RSK1, RSK3	PI3K/MTOR signaling	-0,31071
Methotrexate	1008	Antimetabolite	DNA replication	-0,31035
XMD11-85h	1158	BRSK2, FLT4, MARK4, PRKCD, RET, SRPK1	Other, kinases	-0,31023
SB216763	1025	GSK3A, GSK3B	WNT signaling	-0,31022
Olaparib	1495	PARP1, PARP2	Genome integrity	-0,30997
Eg5_9814	1712	KSP11	Other	-0,30919
Sphingosine Kinase 1 Inhibitor II	408	Sphingosine Kinase	Other, kinases	-0,30846
AZD5438	1401	CDK2	Cell cycle	-0,30832
5-Fluorouracil	179	Antimetabolite (DNA & RNA)	Other	-0,3069
Ulixertinib	2047	ERK1, ERK2	ERK MAPK signaling	-0,30609
Cediranib	1922	VEGFR, FLT1, FLT2, FLT3, FLT4, KIT, PDGFRB	RTK signaling	-0,30537
MIRA-1	1931	TP53	p53 pathway	-0,30482
QS11	151	ARFGAP1	Other	-0,30467

Table 17: Negative correlation between cell growth rate and drug response.

Results organised based on negative correlation, i.e. the most interesting results are cases of negative IC_{50} value combined with positive cell growth. Table depicts the correlation values for 100 unique drugs.

B.6 Drug concentrations in GDSC and validation screens

CHIR-99021	SB216763	9-ING-41	Validation
0.01	0.0390625		0.000076
0.02	0.0781250		0.000153
0.04	0.1562500		0.000305
0.08	0.3125000		0.000610
0.16	0.6250000		0.001221
0.32	1.2500000		0.002441
0.64	2.5000000		0.004883
1.28	5.0000000		0.009766
2.56	10		0.019531
			0.039063
			0.078125
			0.15625
			0.3125
			0.6250
			1.25
			2.5
			5
			10
			20
			40

Table 18: Drug concentrations in GDSC and validation experiments.

CHIR-99021 and SB216763 were available in the GDSC screen, and were tested with 9 distinct titration points. The drug 9-ING-41 was not available in GDSC, and therefore no drug concentration is reported. For comparison, in the validation screen experiments a wide range of concentration was employed with 20 titration points.

B.7 Cell viabilities of the validation screen

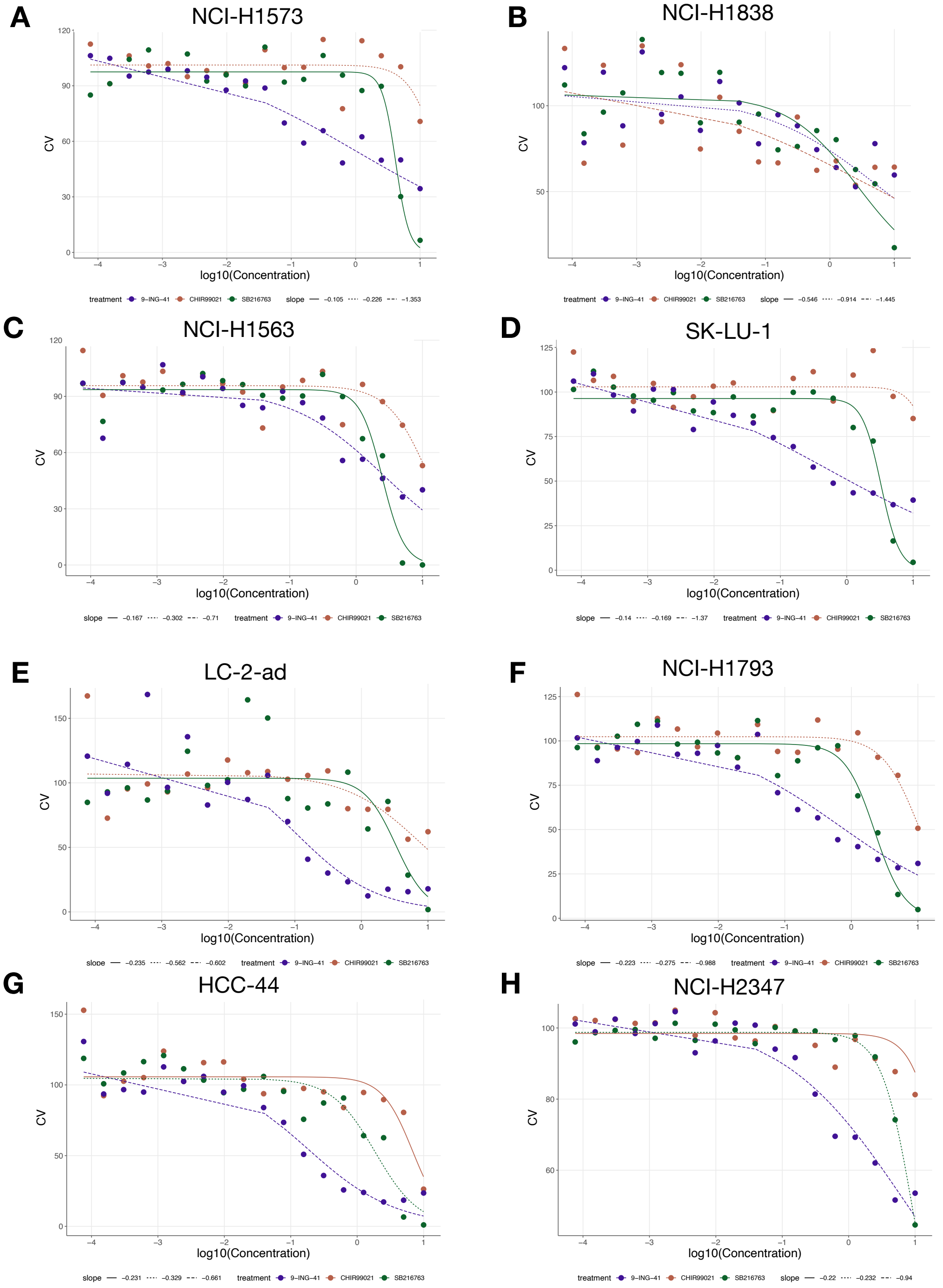

Figure 48: Cell viabilities of validated cell lines and drugs.

Cell viabilities of drugs CHIR-99021, SB216762 and 9-ING-41 with cell lines **(A)-(B)** NCI-H1573 and NCI-H1838 (non-responders); **(C)-(D)** NCI-H1563 and SK-LU-1 (cnaLUAD3); **(E)-(F)** LC-2-ad and NCI-H1793 (cnaLUAD27); **(G)-(H)** HCC-44 and NCI-H2347 (both cnaLUAD3 and cnaLUAD27). Viability values are the mean of all replicates, and curve fit was performed using the nls() function with a sigmoid formulation.

Appendix C: Supplementary information to chapters 2.4 and 3.3

C.1 Seed genes from IntOGen database

Seed gene	Cancer tissue
ABL2	BRCA
AFF3	BRCA
AKT1	BRCA
ALK	BRCA
ARHGEF12	BRCA
ARID1A	BRCA
ARID1B	BRCA
ASXL1	BRCA
ATM	BRCA
BAP1	BRCA
BRAF	BRCA
BRCA1	BRCA
BRCA2	BRCA
CACNA1D	BRCA
CASP8	BRCA
CBFB	BRCA
CDH1	BRCA
CDKN1B	BRCA
CDKN2A	BRCA
CLTC	BRCA
CREBBP	BRCA
CTCF	BRCA
CUX1	BRCA
DDX3X	BRCA
EGFR	BRCA
ELN	BRCA
EPAS1	BRCA
EPHA3	BRCA
ERBB2	BRCA
ERBB3	BRCA
ERBB4	BRCA
ESR1	BRCA
ETV5	BRCA
FAT1	BRCA
FAT3	BRCA
FAT4	BRCA
FBXW7	BRCA
FGFR2	BRCA

FOXA1	BRCA
GATA1	BRCA
GATA3	BRCA
GNAS	BRCA
GRIN2A	BRCA
HIST1H3B	BRCA
HOXC13	BRCA
HOXD13	BRCA
HRAS	BRCA
HSP90AA1	BRCA
JAK2	BRCA
KAT6B	BRCA
KDM6A	BRCA
KLF4	BRCA
KMT2C	BRCA
KMT2D	BRCA
KRAS	BRCA
LRP1B	BRCA
MAP2K4	BRCA
MAP3K1	BRCA
MAX	BRCA
MDM4	BRCA
MEN1	BRCA
MTOR	BRCA
MYH11	BRCA
MYH9	BRCA
MYO5A	BRCA
NCOA1	BRCA
NCOR1	BRCA
NCOR2	BRCA
NF1	BRCA
NIN	BRCA
NONO	BRCA
NOTCH2	BRCA
NTRK1	BRCA
NUMA1	BRCA
PDGFRB	BRCA
PIK3CA	BRCA
PIK3R1	BRCA
PLAG1	BRCA
POLD1	BRCA
PREX2	BRCA
PTEN	BRCA
PTPN13	BRCA
PTPRD	BRCA
RB1	BRCA

RGS7	BRCA
RHPN2	BRCA
RUNX1	BRCA
SALL4	BRCA
SF3B1	BRCA
SMAD2	BRCA
SMAD4	BRCA
SMARCD1	BRCA
SPEN	BRCA
TBX3	BRCA
TP53	BRCA
USP6	BRCA
ZBTB16	BRCA
ZFHX3	BRCA
ZXDB	BRCA

Table 19: Seed genes from IntOGen for breast adenocarcinoma.

C.2 HPO annotations and genes involved with BRCA

HPO Annotation	HPO ID	Gene ID	Gene Symbol
Breast carcinoma	HP:0003002	999	CDH1
Breast carcinoma	HP:0003002	2099	ESR1
Breast carcinoma	HP:0003002	3418	IDH2
Breast carcinoma	HP:0003002	545	ATR
Breast carcinoma	HP:0003002	11200	CHEK2
Breast carcinoma	HP:0003002	7517	XRCC3
Breast carcinoma	HP:0003002	4835	NQO2
Breast carcinoma	HP:0003002	5728	PTEN
Breast carcinoma	HP:0003002	7373	COL14A1
Breast carcinoma	HP:0003002	79728	PALB2
Breast carcinoma	HP:0003002	4436	MSH2
Breast carcinoma	HP:0003002	6794	STK11
Breast carcinoma	HP:0003002	5889	RAD51C
Breast carcinoma	HP:0003002	4210	MEFV
Breast carcinoma	HP:0003002	10483	SEC23B
Breast carcinoma	HP:0003002	672	BRCA1
Breast carcinoma	HP:0003002	675	BRCA2
Breast carcinoma	HP:0003002	2956	MSH6
Breast carcinoma	HP:0003002	9821	RB1CC1
Breast carcinoma	HP:0003002	388662	SLC6A17
Breast carcinoma	HP:0003002	8438	RAD54L
Breast carcinoma	HP:0003002	8493	PPM1D
Breast carcinoma	HP:0003002	5245	PHB
Breast carcinoma	HP:0003002	5892	RAD51D
Breast carcinoma	HP:0003002	4361	MRE11
Breast carcinoma	HP:0003002	5290	PIK3CA
Breast carcinoma	HP:0003002	5424	POLD1
Breast carcinoma	HP:0003002	4763	NF1
Breast carcinoma	HP:0003002	5888	RAD51
Breast carcinoma	HP:0003002	324	APC
Breast carcinoma	HP:0003002	7291	TWIST1
Breast carcinoma	HP:0003002	79719	AAGAB
Breast carcinoma	HP:0003002	6392	SDHD
Breast carcinoma	HP:0003002	472	ATM
Breast carcinoma	HP:0003002	2778	GNAS
Breast carcinoma	HP:0003002	4683	NBN
Breast carcinoma	HP:0003002	83990	BRIP1
Breast carcinoma	HP:0003002	1029	CDKN2A
Breast carcinoma	HP:0003002	4913	NTHL1
Breast carcinoma	HP:0003002	580	BARD1
Breast carcinoma	HP:0003002	6390	SDHB
Breast carcinoma	HP:0003002	5071	PRKN
Breast carcinoma	HP:0003002	3417	IDH1
Breast carcinoma	HP:0003002	4193	MDM2

Breast carcinoma	HP:0003002	3845	KRAS
Breast carcinoma	HP:0003002	7486	WRN
Breast carcinoma	HP:0003002	207	AKT1
Breast carcinoma	HP:0003002	2263	FGFR2
Breast carcinoma	HP:0003002	3161	HMMR
Breast carcinoma	HP:0003002	54894	RNF43
Breast carcinoma	HP:0003002	1499	CTNNB1
Breast carcinoma	HP:0003002	4978	OPCML
Breast carcinoma	HP:0003002	5002	SLC22A18
Breast carcinoma	HP:0003002	23022	PALLD
Breast carcinoma	HP:0003002	10111	RAD50
Breast carcinoma	HP:0003002	4089	SMAD4
Breast carcinoma	HP:0003002	6391	SDHC
Breast carcinoma	HP:0003002	4292	MLH1
Breast carcinoma	HP:0003002	5426	POLE
Breast carcinoma	HP:0003002	7157	TP53
Breast carcinoma	HP:0003002	100144748	KLLN
Breast carcinoma	HP:0003002	205717	USF3
Breast carcinoma	HP:0003002	841	CASP8
Hereditary Breast And Ovarian Cancer Syndrome			
Multifocal breast carcinoma			

Table 20: HPO annotations and genes involved with breast adenocarcinoma.

C.3 Top 5% wPPI candidate genes

Seed gene	Cancer tissue
ABCA3	BRCA
ABRAXAS1	BRCA
ACTR8	BRCA
ADGRB1	BRCA
ADI1	BRCA
AHSP	BRCA
AIFM2	BRCA
ALB	BRCA
ALX1	BRCA
ANKRD17	BRCA
ANKS1B	BRCA
AP3B2	BRCA
APLF	BRCA
ARHGAP22	BRCA
ARHGAP4	BRCA
ATMIN	BRCA
ATP5PF	BRCA
B4GALT1	BRCA
BCL11A	BRCA
BEX1	BRCA
BRAP	BRCA
C4BPB	BRCA
CARM1	BRCA
CCM2	BRCA
CCNG2	BRCA
CD99L2	BRCA
CGAS	BRCA
CHAMP1	BRCA
CSH2	BRCA
CSPG5	BRCA
CTSD	BRCA
DAP	BRCA
DCHS1	BRCA
DCLRE1A	BRCA
DDX10	BRCA
DNTTIP2	BRCA
DOCK6	BRCA
EBP	BRCA
EDC3	BRCA
EDN2	BRCA
EI24	BRCA
EIF4EBP3	BRCA
FABP5	BRCA

FAM102A	BRCA
FBXO31	BRCA
FLYWCH1	BRCA
GCG	BRCA
GML	BRCA
GREB1	BRCA
GRIP1	BRCA
HJURP	BRCA
HOXC6	BRCA
HOXC9	BRCA
HSPBP1	BRCA
IFT46	BRCA
IGFBP4	BRCA
ING2	BRCA
INHBE	BRCA
INVS	BRCA
ISCU	BRCA
L3MBTL2	BRCA
LTB4R2	BRCA
LUC7L2	BRCA
MAF1	BRCA
MAP3K9	BRCA
MLLT1	BRCA
MMP16	BRCA
MPG	BRCA
MTA3	BRCA
MUC4	BRCA
NDUFV3	BRCA
NOP2	BRCA
NOP53	BRCA
NRG2	BRCA
NRG3	BRCA
NRG4	BRCA
NSD2	BRCA
P4HA2	BRCA
PACS2	BRCA
PADI4	BRCA
PATL1	BRCA
PCBP3	BRCA
PCBP4	BRCA
PEX5	BRCA
PIDD1	BRCA
PISD	BRCA
PLAGL1	BRCA
PLEKHA4	BRCA
PLK4	BRCA

PPARGC1B	BRCA
PPM1G	BRCA
PRDM15	BRCA
PRDX2	BRCA
PRPF19	BRCA
PRPH	BRCA
PYGO2	BRCA
RABGGTA	BRCA
RASA4	BRCA
RELL2	BRCA
RGS4	BRCA
RIN2	BRCA
RNF144B	BRCA
RNF168	BRCA
RPL35A	BRCA
RPL37A	BRCA
RPRM	BRCA
S100A4	BRCA
SCARA3	BRCA
SCN3B	BRCA
SESN1	BRCA
SESN2	BRCA
SESN3	BRCA
SH3BGRL3	BRCA
SH3BP4	BRCA
SHISA5	BRCA
SLC38A2	BRCA
SLX4	BRCA
SSR4	BRCA
SYTL1	BRCA
TCEAL6	BRCA
THEM4	BRCA
TIGAR	BRCA
TIPARP	BRCA
TMIGD2	BRCA
TRIAP1	BRCA
TRIM22	BRCA
TRIM25	BRCA
TRIP11	BRCA
TSC1_TSC2	BRCA
TWF1	BRCA
UCHL3	BRCA
ULK3	BRCA
USP28	BRCA
USP4	BRCA
USP9X	BRCA

UXS1	BRCA
VRK2	BRCA
WASHC2A	BRCA
WRAP53	BRCA
ZBTB2	BRCA
ZBTB7A	BRCA
ZMAT3	BRCA
ZNF420	BRCA
ZNF506	BRCA
ZNF768	BRCA
ZNRF2	BRCA

Table 21: Top 5% candidate genes obtained with wPPI for BRCA.

C.4 Drugs with highest performance in drug response prediction

Drug Name	Drug Target	Drug Target Pathway	*R* Whole Genome	*R* Seed Genes	*R* All wPPI genes	*R* top 5% wPPI genes
Tozasertib	AURKA, AURKB, AURKC, others	Mitosis	0,390886	0,330962	0,378349	0,507688
MK-1775	WEE1, PLK1	Cell cycle	0,234541	0,416909	0,3676	0,455271
AZD7762	CHEK1, CHEK2	Cell cycle	0,459132	0,401807	0,486379	0,428097
Teniposide		DNA replication	0,136477	0,2678	0,292931	0,392465
BDP-00009066	MRCKB_HU-MAN	Cytoskeleton	0,225545	0,252586	0,20693	0,383094
AZD6738	ATR	Genome integrity	0,264725	0,233921	0,288025	0,378632
Alisertib	AURKA	Mitosis	0,199881	0,243214	0,26493	0,374932
Lapatinib	EGFR, ERBB2	EGFR signaling	0,23795	0,24258	0,235349	0,365944
Acetalax		Unclassified	0,233922	0,293219	0,263342	0,352011
Mitoxantrone		DNA replication	0,226403	0,262313	0,252516	0,334954
PF-4708671	S6K1	PI3K/MTOR signaling	0,096943	0,250757	0,229969	0,333649
Talazoparib	PARP1, PARP2	Genome integrity	0,250895	0,269377	0,204998	0,328013
Afatinib	ERBB2, EGFR	EGFR signaling	0,374082	0,27585	0,324955	0,317749
Phenformin	Biguanide agent	Other	0,001037	0,099849	-0,01549	0,314599
Fludarabine		Unclassified	-0,04923	0,406979	0,102287	0,302846
Linsitinib	IGF1R	IGF1R signaling	0,046492	0,080315	0,192044	0,288236
Midostaurin	PKC, PPK, FLT1, c-FGR, others	Other	0,100154	0,036091	0,215258	0,286401
Ipatasertib	AKT1, AKT, AKT3	PI3K/MTOR signaling	-0,00205	0,22391	0,058412	0,282306
LGK974	PORCN	WNT signaling	0,208171	0,138621	0,231873	0,275009
Sapitinib	EGFR, ERBB2, ERBB3	EGFR signaling	0,322194	0,351216	0,356562	0,267933
Ispinesib Mesylate	KSP	Mitosis	0,185426	0,234234	0,093978	0,263703
Zibotentan	Endothelin-1 receptor (EDNRA)	Other	-0,05233	-0,04796	-0,02467	0,260703
Cytarabine	Antimetabolite	Other	0,113954	0,130744	0,084259	0,246953
Docetaxel	Microtubule stabiliser	Mitosis	0,157346	0,002614	0,244906	0,243585
Crizotinib	MET, ALK, ROS1	RTK signaling	0,157913	0,312602	0,103864	0,241391
CP724714	ERBB2	RTK signaling	0,188894	0,252839	0,274771	0,232046
AZD5363	AKT1, AKT2, AKT3, ROCK2	Other, kinases	0,02474	0,105357	0,035511	0,225739
Trametinib	MEK1, MEK2	ERK MAPK signaling	0,237925	0,19375	0,275602	0,219384
Dasatinib	ABL, SRC, Ephrins, PDGFR, KIT	Other, kinases	0,234148	0,285279	0,29615	0,217422

Gemcitabine	Pyrimidine anti-metabolite	DNA replication	0,104439	0,120705	0,085044	0,216272
XAV939	TNKS1, TNKS2	WNT signaling	0,034207	0,09833	-0,02107	0,213758
Ponatinib	ABL, PDGFRA, VEGFR2, FGFR1, SRC, TIE2, FLT3	Other, kinases	0,168379	0,188671	0,100644	0,211884
Avagacestat	Amyloid beta20, Amyloid beta40	Other	-0,00808	-0,02752	-0,02076	0,211646
BX-912	PDK1 (PDPK1)	Metabolism	0,145743	0,246326	0,139803	0,210417
Podophyllotoxin bromide		Unclassified	-0,12984	0,036123	-0,02832	0,209728
Nutlin-3a (-)	MDM2	p53 pathway	0,0765	-0,12067	0,134132	0,207812
GW-2580	CSF1R	RTK signaling	0,091003	0,079199	0,04054	0,206578
Vincristine		Mitosis	-0,0706	-0,05127	-0,01498	0,205185
Epothilone B	Microtubule stabiliser	Mitosis	-0,03722	0,230779	0,017455	0,203053
WHI-P97	JAK3	Other, kinases	0,091391	0,238935	0,116409	0,200884
BMS-536924	IGF1R, IR	IGF1R signaling	0,24459	0,13779	0,389783	0,199746
Dihydrorotenone		Unclassified	0,052686	0,141683	0,064662	0,198684
Cisplatin	DNA crosslinker	DNA replication	0,106475	0,029424	0,039025	0,198253
JNK-9L	JNK2, JNK3	JNK and p38 signaling	0,008019	0,211268	-0,01674	0,193442
PD0325901	MEK1, MEK2	ERK MAPK signaling	0,187155	0,186555	0,114792	0,192811
Amuvatinib	KIT, PDGFRA, FLT3	RTK signaling	0,053505	0,17025	0,051041	0,19268
ZM447439	AURKA, AURKB	Mitosis	0,029657	0,091725	0,115051	0,186553
Telomerase Inhibitor IX	Telomerase	Genome integrity	0,220479	0,101308	0,259824	0,18295
Ibrutinib	BTK	Other, kinases	0,150802	0,224171	0,087624	0,178849
Bortezomib	Proteasome	Protein stability and degradation	0,083515	0,156998	0,268667	0,176349

Table 22: Top 50 drugs with the highest performance in machine learning models build with the top 5% ranked wPPI genes.

Appendix D: Curriculum vitae

WORK EXPERIENCE

Scientific Researcher *(June 2024 – Current)*

Computational Health Center, Helmholtz Center Munich, Germany

- Statistical analysis of clinical questionnaire data
- Multi-omics (metabolomics, proteomics, lipidomics) analysis and integration
- Focus on Amyotrophic Lateral Sclerosis
- Manuscript preparation, scientific presentations and writing scientific reports
- Work developed within the premodiALS international consortium

Ph.D. Candidate *(May 2019 – May 2024)*

Computational Health Center, Helmholtz Center Munich, Germany

- Development and application of statistical and ML tools for biomarker discovery
- Implementation of curve fitting methods based on sigmoid, GP and linear models
- Creation of a network-based approach to enhance feature space in drug response prediction
- Applications in cancer pharmacogenomics, neurodegenerative diseases and COVID-19
- Manuscript preparation, scientific presentations and support in grant proposals
- Collaborations with international research partners and supervision of students

Engineer at Toyota Motor Europe via AKKA Technologies *(April 2018 – March 2019)*

Toyota Motor Europe NV/SA, Belgium

- Creation of a Surrogate Modelling toolbox for drivetrain applications based on GP
- Application of multi-objective optimization methods
- Presentations for cross-disciplinary internal discussions

EDUCATION

Ph.D. fellowship in Biology *(May 2019 – Current)*

Ludwig Maximilian University of Munich, Germany

M.Sc.: Mathematics and Applications *(September 2015 – December 2017)*

Technical University of Lisbon, Portugal
Eindhoven University of Technology, Netherlands

B.Sc.: Applied Mathematics and Computation *(September 2011 - July 2016)*

Technical University of Lisbon, Portugal

SKILLS

Languages

Portuguese (Native), English (Fluent), Italian & Spanish (Conversational), German (Basic)

Programming

R (Advanced), MatLab, Python, Shell Scripting, SQL and Mathematica (Intermediate)

Developer tools

Git, CI, HPC, RMarkdown, BioConductor packages, Jupyter Notebook, LaTeX

ADDITIONAL INFORMATION

Awards

- Top 10% most downloaded paper at Clinical and Translational Medicine Journal, 2022
- ECMI certificate for a Master in Mathematics for Industry, 2017
- Academic Merit Diploma at Technical University of Lisbon, 2016/2017

Extras

- Online presentation of bioinformatic analysis at ENCALS meeting in Stockholm, July 2024
- Creation & Publication of the wPPI BioConductor package, 2021
- Poster presentation in the DZD Advisory Meeting in Hohenkammer, December 2019
- Poster presentation in the DZD Diabetes Research School at IRB Barcelona, September 2019
- Participation in the ECMI Modelling Week at Lappeenranta Univ. of Technology, July 2017
- Member of the SCIFI committee at Eindhoven Univ. of Technology, February-July 2017

Appendix E: List of publications

The work discussed throughout this dissertation is specific to the following combination of published and unpublished works.

Published manuscript:

- **(Project 1, chapters 2.2 and 3.1)** Ayestaran, I.*, **Galhoz, A**.*, Spiegel, E., Sidders, B., Dry, J.R., Dondelinger, F., Bender, A., McDermott, U., Iorio, F. and Menden, M.P., (2020). Identification of intrinsic drug resistance and its biomarkers in High-throughput pharmacogenomic and CRISPR screens. *Patterns*, *1*(5);

  Declaration of contribution: The described methods and results are an adaptation of the original publication (Ayestaran et al., 2020), and have been rewritten by me.

  Illustrations related to this work are either referenced to the original publication or produced by me.

  Data download and processing, implementation of frameworks to identify sensitive and resistant cell lines, and biological interpretation were performed by Iñigo Ayestaran.

  My contribution to this work was to ensure statistical rigorousness by developing a hierarchical false discovery rate (HFDR) control of the proposed methodology and benchmark it against the state-of-the-art statistical outlier detection method Neyman-Pearson.

In preparation:

- **(Project 2, chapters 2.3 and 3.2) Galhoz, A.** *, Kutkaite, G. *, … & Menden, M. P.; Exploration of increasing cell viability behaviour in high-throughput screens as novel cancer therapy.

  Declaration of contribution: Data processing from raw pharmacological data into cell viabilities, drug response estimations using several curve fit methods, post-processing bioinformatic analysis and development of a statistical framework to identify increasing cell viability were implemented by me.

  My colleague Ginte Kutkaite jointly assisted in the final selection of increasing cell viability candidates to be followed up. In addition, she led the development of validation protocols and analyses.

  The collaboration partners from the labs Prof. Dr. Daniel Krappmann and Dr. Kamyar Hadian, facilitated and performed the validation experiments of selected candidates.

- **(Project 3, chapters 2.4 and 3.3) Galhoz, A.**, … & Menden, M. P.; wPPI network approach for tissue specific drug response modelling.

  Declaration of contribution: Data handling and integration (drug screens, protein-protein interaction and ontology databases), creation of the wPPI gene prioritisation tool, implementation of the LASSO model to predict drug response and any presented post-analysis were carried out by me.

  The Bioconductor package wPPI was performed in collaboration with Dr. Denes Turei, who supported the connection to the OmnipathR package (Türei et al., 2021) and refined the wPPI's source code according to Bioconductor guidelines. In addition, to configure wPPI to a cancer-specific analysis, Phong Nguyen created a collection of HPO terms linked to TCGA cancer annotations and Daniel Garger assisted on the curation of relevant HPO terms associated with several cancer types.

In addition, during my PhD, I co-authored in the following publications which are not discussed in this thesis:

- Tschuck, J., Padmanabhan Nair, V., **Galhoz, A.**, Zaratiegui, C., Tai, H.M., Ciceri, G., Rothenaigner, I., Tchieu, J., Stockwell, B.R., Studer, L., Cabianca, D.S., Menden, M.P., Vincendeau, M. and Hadian, K., (2024). Suppression of ferroptosis by vitamin A or radial-trapping antioxidants is essential for neuronal development. *Nature Communications, 15*(1), 7611.
- Ohnmacht, A.J., Stahler, A., Stintzing, S., Modest, D.P., Holch, J.W., Westphalen, C.B., Hölzel, L., Schübel, M.K., **Galhoz, A**., Farnoud, A., Ud-Dean, M., Vehling-Kaiser, U., Decker, T., Moehler, M., Heinig, M., Heinemann, V. and Menden, M. P., (2023). The oncology biomarker discovery framework reveals cetuximab and bevacizumab response patterns in metastatic colorectal cancer. *Nature Communications*, *14*(1), 5391;
- Erber, J., Kappler, V., Haller, B., Mijočević, H., **Galhoz, A.**, da Costa, C.P., Gebhardt, F., Graf, N., Hoffmann, D., Thaler, M., Lorenz, E., Roggendorf, H., Kohlmayer, F., Henkel, A., Menden, M.P., Ruland, J., Spinner, C.D., Protzer, U., Knolle, P. and Lingor, P., (2022). Infection control measures and prevalence of SARS-CoV-2 IgG among 4,554 university hospital employees, Munich, Germany. *Emerging Infectious Diseases*, *28*(3), 572;
- Caldi Gomes, L.*, **Galhoz, A.***, Jain, G., Roser, A.E., Maass, F., Carboni, E., Barski, E., Lenz, C., Lohmann, K., Klein, C., Bähr, M., Fischer, A., Menden, M.P., Lingor, P., (2022). Multi-omic landscaping of human midbrains identifies disease-relevant molecular targets and pathways in advanced-stage Parkinson's disease. *Clinical and translational medicine*, *12*(1), e692;
- Nguyen, P., Ohnmacht, A.J., **Galhoz, A.**, Büttner, M., Theis, F. and Menden, M.P., (2021). Artificial intelligence and machine learning in diabetes research. *Der Diabetologe*, 1-11;
- Wang, D., Hensman, J., Kutkaite, G., Toh, T.S., **Galhoz, A**., GDSC Screening Team Lightfoot Howard Yang Wanjuan Soleimani Maryam Barthorpe Syd Mironenko Tatiana Beck Alexandra Richardson Laura Lleshi Ermira Hall James Tolley Charlotte Barendt William, Dry, J.R., Saez-Rodriguez, J., Garnett, M.J., Menden, M.P. and Dondelinger, F., (2020). A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates. *Elife*, *9*, e60352;

* authors contributed equally and are shared first-authors.